IMAGE CAPTIONING USING DEEP-LEARNING TECHNIQUES

A Thesis Submitted
In Fulfillment of the Requirements
for the Degree of

DOCTOR OF PHILOSOPHY

by

DHRUV SHARMA

(2K20/PHDEC/02)

Under the Supervision of

Dr. Dinesh Kumar
Professor, Delhi Technological University

&

Dr. Chhavi Dhiman
Asst. Prof., Delhi Technological University

Department of Electronics & Communication Engineering

DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Shahbad Daulatpur, Main Bawana Road, Delhi-110042. India

December, 2024

DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Shahbad Daulatpur, Main Bawana Road, Delhi-42

CANDIDATE'S DECLARATION

I **Dhruv Sharma (2K20/PHDEC/02)** hereby certify that the work which is being presented in the thesis entitled "**Image Captioning using Deep-Learning Techniques"** in partial fulfillment of the requirements for the award of the Degree of Doctor of Philosophy, submitted in the Department of Electronics & Communication Engineering, Delhi Technological University is an authentic record of my own work carried out during the period from August 2020 to May 2024 under the supervision of **Prof. Dinesh Kumar**, Professor in Department of Electronics and Communication Engineering, Delhi Technological University, Delhi, India, and **Dr. Chhavi Dhiman**, Assistant Professor in Department of Electronics and Communication Engineering, Delhi Technological University, Delhi, India.

The matter presented in the thesis has not been submitted by me for the award of any other degree of this or any other Institute.

Candidate's Signature

DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Shahbad Daulatpur, Main Bawana Road, Delhi-42

CERTIFICATE BY THE SUPERVISOR(s)

Certified that **DHRUV SHARMA (**2K20/PHDEC/02) has carried out their research work presented in this thesis entitled **"Image Captioning using Deep-Learning Techniques"** for the award of **Doctor of Philosophy** from Department of Electronics & Communication Engineering, Delhi Technological University, Delhi, under our supervision. The thesis embodies results of original work, and studies are carried out by the student himself and the contents of the thesis do not form the basis for the award of any other degree to the candidate or to anybody else from this or any other University/Institution.

Date:

Prof. Dinesh Kumar
Department of ECE
Delhi Technological University
Delhi-110042, India

Dr. Chhavi Dhiman
Department of ECE
Delhi Technological University
Delhi-110042, India

Prof. C.S. Rai
University School of Information and Communication Technology
Guru Gobind Singh Indraprastha University
Delhi-110078

ACKNOWLEDGEMENT

I owe tremendous debt and would like to express deep feelings of gratitude for the support and guidance of several people who have helped me to accomplish the research program with the support and direction of several persons. This challenging and rewarding experience has definitely helped me grow in character as well as academically. It gives me a great pleasure to now have the opportunity to express my gratitude towards them.

Words cannot express my gratitude towards my supervisor Prof. Dinesh Kumar, Department of Electronics and Communication Engineering, Delhi Technological University, Delhi, India and Dr. Chhavi Dhiman, Assistant Professor, Department of Electronics and Communication Engineering, Delhi Technological University, Delhi, India. I greatly acknowledge their invaluable patience, feedback, guidance, and continuous motivation throughout my journey. I extend my sincere regards to HOD sir for his constant support. I wish to express my gratitude towards the DRC chairperson, and the esteemed faculty members for their valuable time and efforts throughout this journey.

I express my heartfelt gratitude to my mother Ms. Suman Sharma, father Mr. S.P. Sharma and also to my elder brother Mr. Xitij Sharma, and sister-in-law Ms. Mahak Bhati Sharma for their unconditional love, encouragement and blessings. Also, I would like to thank my sister Dr. Vrinda Sharma for her support during this tenure. They have been a guiding force all their life and tried to measure up to their expectations.

Further, I would like to express my heartfelt gratitude to Gagan Abbot and Aishvary Varyani for helping me during this tenure and constantly helping me during this journey. I am also grateful to all my colleagues of the Electronics and Communication department, especially Roli Kushwaha, Aakanksha Gupta Kamakshi Rautela, Ishaan Sharma, Shikha Singhal, Kavita Bhatt, Monika Singh, Snehlata Yadav, Bhawana Rawat. I also express my thanks to all the staff members of the department for their continuous support in our academic activities.

I am also thankful to my friends Yatin Kataria, Gagan Singh, Ritika Gupta, Aakriti Vishnoi, Nomica Kapoor, and Manas Bajpai for their guidance and motivation. With this, I also want to thank some of my students Kush Agarwal, Amit Swami, Avi Gupta, Ansh Rathod, Abhishek Sharma, Akshat Yadav, Shwaas, and Rishabh for their valuable support during this entire journey.

DHRUV SHARMA

Dedicated to my loving parents

Ms. Suman Sharma & Mr. Surya Prakash Sharma

ABSTRACT

Image Caption generation is a description of the contents of an image in the form of natural language sentences. It is evolving as an active research area of Computer Vision (CV) and Natural Language Processing (NLP). It generates syntactically and semantically correct sentences by describing important objects, attributes, and their relationships with each other. The very nature of it makes it suitable for applications such as image retrieval [1] [2], human-robot interaction [3] [4], aid to the blind [5], and visual question answering [6]. With the advent of technology and proliferating demand of society, automatic and intelligent image captioning based systems have become the need of the hour.

Many Convolutional Neural Network (CNN)-based architectures are utilized at the encoder for efficient extraction of image features while Long Short-Term Memory (LSTM)-based decoder is further utilized for the generation of captions. Different variants of LSTM and CNNs with attention mechanism are utilized by the traditional methods for generation of meaningful and accurate descriptions of images. Though the captions generated by the traditional methods are simple yet they sometimes have some limitations which is mainly due to repetitive words or inaccurate descriptions of the scene, which usually may not reflect in natural language usage. Also, the traditional image captioning techniques fails to capture the relationship between objects and surroundings, thereby, neglecting the fine-grained details and diverse scenes, hence, leading to the ambiguities in generated language.

To overcome the challenges faced by the traditional captioning models, this

work focusses on generation of image captions using advanced deep-learning techniques. These techniques help to better understand the visual content and context of images by providing contextually relevant information rather than just listing objects. Therefore, leading to the model's capability to generate meaningful, nuanced, and detailed descriptions for different real-world applications. The work in this thesis thus investigates different deep-learning based models or frameworks for factual, stylized and paragraph-based description of images.

For generation of factual-based description of images, this thesis first discusses about Lightweight Transformer with GRU integrated decoder for image captioning. The proposed Lightweight Transformer exploits a single encoder-decoder based transformer model for generation of factual captions. Extensive experiments on MSCOCO dataset demonstrated that the proposed approach achieves a competitive score on all the evaluation metrics. For efficient description of the captions, it becomes necessary to learn higher order interactions between detected objects and the relationship among them. Most of the existing models take into account the first order interactions while ignoring the higher order ones. It is challenging to extract discriminant higher order semantics visual features in images with highly populated objects for caption generation. In this direction, an efficient higher order interaction learning framework is proposed in this study using encoder-decoder based image captioning. To leverage higher order interactions among multiple objects, an efficient XGL Transformer (XGL-T) model is introduced that exploits both spatial and channel-wise attention. The proposed XGL-T model captures rich semantic concepts from objects, attributes, and their relationships. Extensive experiments are conducted on publicly available MSCOCO Karapathy test split and the best performance of the work

is observed as 81.5 BLEU@1, 67.1 BLEU@2, 51.6 BLEU@3, 39.9 BLEU@4, 134 CIDEr, 59.9 ROUGE-L, 29.8 METEOR, 23.8 SPICE using CIDEr-D Score Optimization Strategy.

Methods developed in the recent past focused mainly on the description of factual contents in images thereby ignoring the different emotions and styles (romantic, humorous, angry, etc.) associated with the image. To overcome this, few works incorporated style-based caption generation that captures the variability in the generated descriptions. This thesis presents a Style Embedding-based Variational Autoencoder for Controlled Stylized Caption Generation Framework (RFCG+SE-VAE-CSCG). It generates controlled text-based stylized descriptions of images. It works in two phases i.e., (i) Refined Factual Caption Generation (RFCG), and (ii) SE-VAE-CSCG. The former defines an encoder-decoder model for the generation of refined factual captions whereas, the latter presents a style embedding-based variational autoencoder for controlled stylized caption generation. More so, with the use of a controlled text generation model, the proposed work efficiently learns disentangled representations and generates realistic stylized descriptions of images. Experiments on MSCOCO, Flickr30K, and FlickrStyle10K provide state-of-the-art results for both refined and style-based caption generation.

Further, multi-level Variational Autoencoder Transformer (VAT)-based framework, MrA^2VAT, is also proposed in this work for the generation of descriptions of images in the form of a paragraph. The proposed framework utilizes a combination of visual and spatial features which are further attended by the proposed multi-resolution multi-head attention (M^2A) to capture the relationships between the query

representation and different attention granularities. To increase the language diversity and to remove the redundant sentences from the generated paragraph, the proposed framework also leverages a language discriminator. Extensive experiments on the Stanford Paragraph Dataset are conducted that provide superior results on all evaluation metrics with or without a language discriminator.

Different deep-learning techniques are devised for the development of factual and stylized image captioning models. Previous models focused more on the generation of factual and stylized captions separately providing more than one caption for a single image. To address this issue, a novel Unified Attention and Multi-Head Attention-driven Caption Summarization Transformer ($UnMA$-CapSumT) based Captioning Framework is discussed in this thesis which integrates different captioning methods to describe the contents of an image with factual and stylized (romantic and humorous) elements. The proposed framework exploits both factual captions and stylized captions generated by the Modified Adaptive Attention-based factual image captioning model (MAA-FIC) and Style Factored Bi-LSTM with attention (SF-Bi-ALSTM) driven stylized image captioning model respectively. Further, summarization transformer $UnMHA - ST$ combines both factual and stylized descriptions of an input image to generate styled rich coherent summarized captions. Extensive experiments are conducted on Flickr8K and a subset of FlickrStyle10K with supporting ablation studies to prove the efficiency and efficacy of the proposed framework.

Also, this work presents two main application areas of the image captioning namely: medical image captioning and aid-to-the-blind. For medical image captioning, $FDT - Dr^2T$ framework is proposed which leverages the fusion of texture features

and deep features in the first stage by incorporating ISCM-LBP + PCA-HOG feature extraction algorithm and Convolutional Triple Attention-based Efficient XceptionNet ($C - TaXNet$). Further, fused features from the FDT module are utilized by the Dense Radiology Report Generation Transformer (Dr^2T) model with modified multi-head attention generating dense radiology reports by highlighting specific crucial abnormalities. For aid to the blind, an adaptive attention mechanism and Bi-LSTM-based automated Image caption generation framework is proposed in this study. The proposed model exploits Inception-V3 to extract various global spatial features and the adaptive attention module helps to decide whether to attend to the image (and if so, to which regions) or to the visual sentinel maps. Further, at the decoding end, a Bi-LSTM network refines the text description. The proposed model performance is evaluated in terms of BLEU scores on Flickr8K and Visual Assistance Dataset.

LIST OF PUBLICATIONS

- **<u>SCI/SCIE Indexed Journal Papers Published</u>**

1. Dhruv Sharma, Chhavi Dhiman, Dinesh Kumar, "Control with Style: Style Embedding-based Variational Autoencoder for Controlled Stylized Caption Generation Framework," ***IEEE Transactions on Cognitive and Developmental Systems***, DOI: 10.1109/TCDS.2024.3405573, 2024 **(I.F. – 5, IEEE).**

2. Dhruv Sharma, Chhavi Dhiman, Dinesh Kumar, FDT-Dr2T: A Unified Dense Radiology Report Generation Transformer Framework for X-Ray Images," ***Machine Vision and Applications***, **35**, 68 (2024). https://doi.org/10.1007/s00138-024-01544-0, 2024. **(I.F. – 3.3, Springer).**

3. Dhruv Sharma, Chhavi Dhiman, Dinesh Kumar, "Evolution of visual data captioning Methods, Datasets, and evaluation Metrics: A comprehensive survey," ***Expert Systems with Applications,*** 221 (2023): 119773. **(I.F. – 8.5, Elsevier)**

4. Dhruv Sharma, Chhavi Dhiman, Dinesh Kumar, "XGL-T Transformer Model for Intelligent Image Captioning," ***Multimedia Tools and Applications***, **83**, 4219–4240 (2024). https://doi.org/10.1007/s11042-023-15291-3. **(I.F. – 3.6, Springer).**

- **<u>Papers Submitted/to be Communicated in SCI/SCIE Indexed Journals</u>**

1. Dhruv Sharma, Chhavi Dhiman, Dinesh Kumar, “UnMA -CapSumT: Unified and Multi-Head Attention-driven Caption Summarization Transformer” ***ACM Transactions on Multimedia Computing, Communications, and Applications. (Under Review) (I.F. – 5.1, SCIE Indexed).***

2. Dhruv Sharma, Chhavi Dhiman, Dinesh Kumar, “MrA2VAT: Variational Autoencoder Transformer for Dense Paragraph Image Captioning” ***ACM Transactions on Multimedia Computing, Communications, and Applications. (Under Review) (I.F. – 5.1, SCIE Indexed).***

- **<u>Scopus Indexed International Conference Papers</u>**

1. Dhruv Sharma, Chhavi Dhiman and Dinesh Kumar, "Automated Image Caption Generation Framework using Adaptive Attention and Bi-LSTM," ***2022 IEEE Delhi Section Conference (DELCON)***, New Delhi, India, 2022, pp. 1-5, doi: 10.1109/DELCON54057.2022.9752859.

2. Dhruv Sharma, Chhavi Dhiman and Dinesh Kumar, "A Review of Stylized Image Captioning Techniques, Evaluation Parameters, and Datasets," *2022* ***4th International Conference on Artificial Intelligence and Speech Technology (AIST)***, Delhi, India, 2022, pp. 1-5, doi: 10.1109/AIST55798.2022.10064842.

3. Dhruv Sharma, Rishabh Dingliwal, Chhavi Dhiman and Dinesh Kumar, "Lightweight Transformer with GRU Integrated Decoder for Image Captioning," *2022* ***16th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS)***, Dijon, France, 2022, pp. 434-438, doi: 10.1109/SITIS57111.2022.00072.

TABLE OF CONTENTS

.

List of Figures

List of Tables

List of Abbreviations and Symbols

The list of abbreviations and symbols used in this thesis is given below. Some other abbreviations / symbols, which are not mentioned here are described locally.

NLP:	Natural Language Processing
CNN:	Convolutional Neural Network
RNN:	Recurrent Neural networks
CV:	Computer Vision
LSTM:	Long Short-Term Memory
AI:	Artificial Intelligence
RCM:	Retrieval-Based Caption Model
TCM:	Template-Based Caption Model
DLCM:	Deep-Learning-Based Caption Model
CRF:	Conditional Random Field
DCC:	Deep Compositional Captioner
NOC:	Novel-Object Captioner
DNOC:	Decoupled Novel Object Captioner
CRN:	Cascaded Revision Network
VIVO:	VIsual VOcabulary pretraining
MoRE:	Mixture of Recurrent Experts
RPN:	Region Proposal Network
FCLN:	Fully Convolutional Localization Network
DSE-LSTM:	Dense Semantic Embedding Network-LSTM
MTTSNet:	Multi-Task Triple-Stream Network
PFE:	Precise Feature Extraction
DAM:	Depth-aware Attention model
CAVP:	Context-Aware Visual Policy
VTCM:	Visual-Textual Coupling Model

HSGED:	Hierarchical Scene Graph Encoder-Decoder
CIC:	Change Image Captioning
DUDA:	Dual Dynamic Attention Model
M-VAM:	Mirrored Viewpoint Adapted Matching
VACC:	Viewpoint-Agnostic with Cycle Consistency
R^3Net:	Relation-embedded Representation Reconstruction Network
SC-NLM:	Structure-Content Natural Language Model
sgl-LSTM:	Self-guiding multimodal LSTM
m-LSTM:	Multimodal LSTM
NIC:	Neural Image Caption
g-LSTM:	guided LSTM
LRCN:	Long-term Recurrent Convolutional Network
DGDN:	Deep Generative Deconvolutional Network
CGAN:	Conditional Generative Adversarial Networks
SCST:	Self-Critical Sequence Training
GRU:	Gated Recurrent Unit
CSMN:	Context Sequence Memory Network
AoA:	Attention on Attention
LSTM-A:	LSTM with attributes
EE-LSTM:	Element Embedding LSTM
BLEU:	Bilingual Evaluation Understudy
ROUGE:	Recall Oriented Understudy for Gisting Evaluation
LCS:	Longest Common Subsequence)
TF-IDF:	Term Frequency-Inverse Document Frequency
SPICE:	Semantic Propositional Image Captioning Evaluation
RoI:	Region-of-Interest
FFN:	Feed-Forward Network
GRU:	Gated Recurrent Unit
XGL-T:	XGL Transformer

RLF: Region-Level Features

ILF: Image-Level Features

CAM: Conventional Attention Module

SSE: Skip-Squeeze and Excitation

SE: Squeeze and Excitation

ELU: Exponential Linear Unit

GLU: Gated Linear Unit

$B@N$: $BLEU@N$

C: $CIDEr - D$

M: $METEOR$

R: $ROUGE - L$

S: $SPICE$

SE-VAE: Style Embedding-based Variation Autoencoder

F: Factual

R: Romantic

H: Humorous

RFCG: Refined Factual Caption Generation

RFC: Refined Factual Captions

BoC: Bag of Captions

RFCG: Refined Factual Image Caption Generation

SMU: Smooth Maximum Unit

SENet: Squeeze and Excitation Network

GT: Ground Truth

RFC: Refined Factual Captions

RC: Romantic captions

HC: Humorous captions

VAE: Variational Auto-Encoder

CSCG: Controlled Stylized Caption Generation

RFCG: Refined Factual Caption Generation

cls: Style Transfer Accuracy

ppl: Perplexity

Mr^2VAT: Multi Resolution and Adaptive Attention driven Variational Autoencoder Transformer

M^2A: Multi-resolution Multi-head Attention

AA: Adaptive Attention

$pos-CWE$: Fused positional character and word embedding

AdBL-LD: Attention-based Dual Bi-LSTM Language Discriminator

FIC: Factual Image Captioning

$UnMHA-ST$: Unified and Multi-head Attention-based Caption Summarization Transformer

Un-A: Unified Attention

OOV: Out of vocabulary

MAA-FIC: Modified Adaptive Attention-based Factual Image Captioning

CBOW: Continuous Bag of Words

MAA: Modified Adaptive based Attention

fTA: fastText with Attention

fTA-WE: fTA-Word-Embedding

AVC: Automatic Visual Captioning

MIC: Medical Image Captioning

Dr^2T: Dense Radiology Report Generation Transformer

RFDB: Residual Feature Distillation Block

SRB: Shallow Residual Block

$C-TaXNet$: Convolutional Triple Attention-based Efficient XceptionNet

CTAM: Convolutional Triple Attention Module

GLSEP: Global Log-Sum-Exp Pooling

MHA: Multi-Head Attention

SPA: Scaled-dot Product Attention

RB: Residual Block

fT-WE:	fastText word embeddings
VTI:	Variational Topic Inference
BPTT:	Back Propagation Through Time
VQA:	Visual Question Answering

Chapter-1

Introduction

Nowadays, it is easy to generate and collect visual data which possess copious information for addressing real-world problems such as healthcare [1], public surveillance [2], sports analysis [3], anomaly detection [4], crowd analysis [2] [3]. It has led to easy accessibility of images and videos. Hence, an automatic and intelligent visual understanding and content summarization technique have emerged as a paramount interest [5] [6]. The research community [7] [8] is working towards a smart visual understanding of the images and videos. However, there exists a large semantic gap between low-level and high-level abstract knowledge of the visual data. Caption Generation can serve as a good solution to bridge the semantic gaps between low-level and high-level abstract knowledge of the visual data and can serve various real-world applications i.e., video surveillance systems [9] visual recognition [10], visual assistive systems [11], health care systems [12], scene understanding [13]. Though many traditional computer-vision techniques [14] [15] for object classification or detection have shown promising results, still they usually generate partial and unstructured outputs, such as bounding boxes and object labels in a video frame. The obtained semantic primitives can be utilized for caption generation for images/videos. Whereas, Natural Language Processing (NLP) can be used to describe these visual observations as sentences, which are much easier for understanding. This chapter introduces the fundamental building blocks of visual image captioning, its application and the

challenges involved therein. In the last section, the major research contributions of the thesis are discussed, followed by thesis organisation.

1.1 Image Captioning

Image captioning is a challenging task that describes the visual content of an image into a natural language and provides automated insights of images giving answers to questions like where you are? (beach, cafe, etc.), what do you wear? (color), and more importantly what you are doing? (playing, walking, etc.). It recognizes the objects, their attributes, and their relationships in an image and generates syntactically and semantically correct sentences. With the advancements of neural networks, image captioning has gained immense popularity which helps in generating human-like descriptions according to the input image. It has a promising future to facilitate the intelligence of mankind and can serve as a helpful tool for visually impaired people. Many other applications can be developed in this direction such as finding the expiry date of a specific food item or knowing about the weather by taking a picture. The process of image captioning can be defined in three fundamental steps as depicted in Fig. 1.1, i) Object Detection ii) Attributes/ Feature Extraction, followed by iii) Sentence Generation. Initially, after having detected the object/image, its features/attributes of the given input image i.e., color, objects, boundaries, and texture details are extracted, encrypted, and translated as an appropriate description. There exist various challenges in the field of caption generation of visual data. Researchers in today's world have designed computer vision-enabled captioning models which can describe *"what"* (e.g., classification [16], segmentation [14]) and *"where"* (e.g., detection [15], tracking [17]). However, it is bad at knowing *"why"*, e.g., why is it a

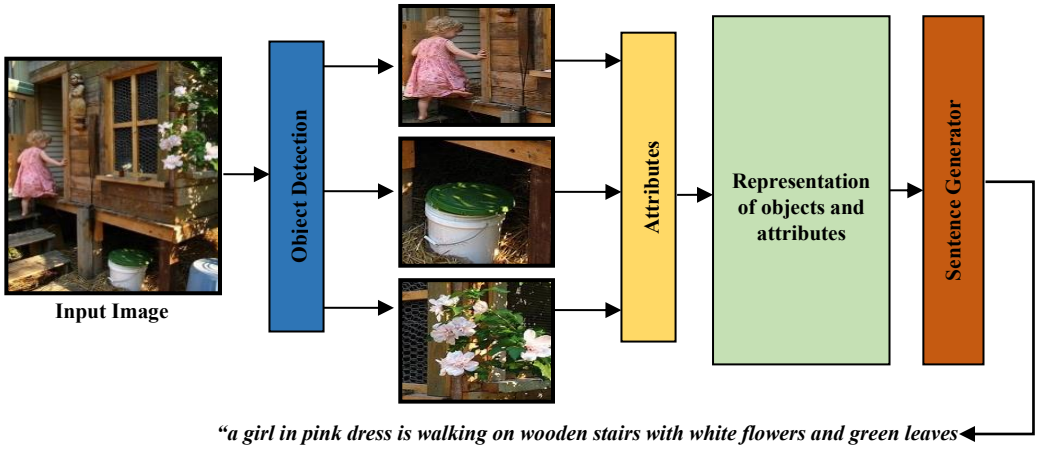


Fig. 1.1: Fundamental Steps involved in an Image Captioning Model

girl? Note that the *"why"* here does not merely mean by asking for visual reasons — attributes like two hands, two legs, hair, that are already well-addressed by machines; further, it also lacks high-level common-sense reasons — such as girl is climbing wooden stairs — that are still elusive, even for human philosophers, not to mention for machines. Further image captioning models should be designed such that the model can define the caption semantics as clearly as possible by describing multiple target objects-"bucket with green lid", "white flowers", instead of just describing a single target object-"girl in a pink dress". To sum up, in its current art, image captioning has gained steady headway and produced brusque and generic vivid captions, with the introduction of Convolutional Neural Network (CNN) [18] and Recurrent Neural networks (RNN) [19] since 2015. For this to grow fully and become an assistive technology, an archetype is required that shifts towards goal-oriented captions; where the caption not only describes a scene from day-to-day life but also answers a specific need that is helpful in many applications.

1.2 Challenges to Visual Image Captioning

Human beings can easily recognize their surroundings and can describe any image or video scene in their natural language but in the case of machines, it is very difficult to generate human-like descriptions of images and videos. However, machines can recognize various human activities from video frames and images to a certain extent, but the automatic description of visual scenes for complex and long-term human activities, is still a challenging task. From a linguistic perspective, activity recognition is all about extracting semantic similarity among human actions represented by verb phrases and transforming visual information into semantic text. It is analogous to grounding words in perception and action.

It has been observed that more attention is required on the generation of attractive and detailed descriptions for images automatically. In the field of CV, we are interested in constructing and learning models which can characterize images or videos by recognizing their categories or other high-level features. In the field of NLP, we usually encounter the inverse challenges to parse a language description by identifying connotations and denotation of the sequence of words. These challenges arise because languages are directly related concepts rather than the lossless recording of objects or activities in the real world. The major challenges to visual image captioning are as follows:

a) ***Compositionality and naturalness of natural language and visual scenes:*** Traditional techniques do not acknowledge and recognize minute details of images and videos. Therefore, the interaction of objects is an onerous task thus making the traditional techniques suffer from a lack of compositionality and naturalness. The biggest challenge here is the subtleness of the action units. Sometimes they are either

not visible, or are hard for vision techniques to detect. For instance, unclear unit boundaries and occlusions of interactive objects present other difficulties to accurately decode the intention of the human activities in an image. In some of the works, attention-based models [20] [21] are designed to address this issue.

b) Intermediate representation learning: Learning mid-level representations between visual domain and natural language domain is a key problem in visual to text techniques. Therefore, high-level visual features are the need of art to represent the visual data completely.

c) Recounting of visual contents: There exist state-of-the-art [19] [22] that recognize semantic elements in the visual data, still, fail to rank in order in accordance with the theme of the image or video that may guide to generate more relevant textual descriptions. Also, we need to find out how much detail we are looking to recount and what type of language complexity is to be applied.

d) Benchmark datasets with moneyed text: To automatically evaluate language descriptions for visual contents, we need standard datasets [23] [24] for evaluating new methods and algorithms. Sentence-level annotations that are aligned to the image and video are basic requirements. For corpus descriptions of different languages, a general image or video description system capable of handling multiple languages should be developed.

e) Evaluation of quality of captions generated: With the use of automated metrics [25] [26], we can partially evaluate the quality of the captions generated. In some cases, this evaluation remains inadequate and sometimes even misleading. The best way to evaluate the quality of automatically generated texts is a subjective assessment by linguists, which is hard to achieve. To improve system performance, the evaluation

indicators should be optimized to make them more in line with human experts' assessments.

The long-standing problem in CV and AI is the semantic gap between low-level visual data and high-level abstract knowledge. Therefore, semantic gap in bridging language and vision points to the need for incorporating common sense and reasoning into scene understanding. Also, the techniques for visual captioning should be able to leverage more flexible semantic units, i.e., they should have various combination of nouns, verbs, and other language units. Further, the progress on visual captioning scene understanding will make CV system more reliable for the use of aid-to-blind, visual question answering, google image search, etc.

1.3 Problem Statement

The challenges involved in generation of semantically and syntactically correct descriptions of images, motivate to develop efficient algorithms to fulfil the purpose of developing robust captioning models and or frameworks. In order to achieve this, models and frameworks are to de designed that generate factual, stylized and paragraph-based descriptions of images. The proposed frameworks extract discriminant higher order semantics visual features in images with highly populated objects. Therefore, describing the image contents factually and emotionally.

1.4 Theoretical Formulation

- ✓ The need for an appropriate and efficient caption generation model is identified which can generate syntactically and semantically relevant and diverse description of images.
- ✓ Issues involved in recognizing small size objects in images and include these objects in generation of descriptions of images are highlighted. Small sized

objects should be included in captions for better interpretability of the scenes. A captioning framework should be developed that extracts discriminant higher order semantics visual features in images with highly populated objects for caption generation.

- ✓ Stylized Captioning is still an open issue due to limited available data. Therefore, style-based captioning framework should be developed by incorporating unpaired images and captions. This makes the caption generation acceptable in wider range of applications by not being restricted with the availability limited data.
- ✓ Dense Visual Caption generation methods closely look at the minute details in the frame by correlating text at multiple level of abstraction. The issues involved in the generation of dense or paragraph-based captions are highlighted. A paragraph-based captioning framework with language discriminator is developed that overcomes the issues related to language diversity and removal of redundant information from the generated paragraph.
- ✓ Recent applications of transformers to computer-vision field have attracted extensive attention. Therefore, very limited work is available that leveraged the strength of transformer model for visual data captioning. Transformer-based models are developed that benefit from the scalability and versatility that contemporary deep learning models provide, and they excel at processing complicated data and producing high-quality, context-aware captions.
- ✓ Recently developed deep-learning-based captioning models either focussed on factual content representation or on the stylized part of the image. There is a need for development of captioning model that describes the contents of image

by highlighting both factual and stylized contents and provides more coherent description of an image.

1.5 Experimental Validation

The developed models are experimentally validated on publically available datasets. These datasets have numerous real-life challenges. The quality and variety of the data required to train and assess the models is one of the primary obstacles to image captioning and retrieval. Most of the datasets currently available for these tasks are synthetic, have size restrictions, or are skewed towards genres, styles, or domains.

- ✓ Higher-order interaction between visual content and natural sentence are learned by defining a novel XGL attention module that extracts high level image features and relationships between objects, thereby, handling diverse and complex scenes, generating more contextually relevant captions.
- ✓ Due to limited available data for stylized image captioning, it is difficult to preserve the correlations between images and captions. Therefore, to overcome this, unpaired set of images and captions are utilized to make the model learn both unstructured (semantics) and structured (style) feature distributions jointly to generate controlled and plausible stylized captions.
- ✓ Also, there is no such data available that is skewed for both factual and stylized captions. In this direction, a summarization transformer framework is developed that provides a summarized caption by integrating factual and stylized image captioning tasks. This integration makes the captioning model to capture the variety and richness of visual and linguistic information in the real world.

- ✓ It is observed that with the exploitation of language discriminator and dissimilarity scores in the paragraph generation model, the model is able to generate more coherent and diverse descriptions of images with no redundant information.

1.6 Motivation to Visual Image Captioning

Image captioning is a fascinating field which is an amalgamation of computer vision and natural language processing that describes the contents of an image in the form of natural language. Image captioning is driven by a number of pragmatic and social requirements in addition to the intellectual challenge it poses. Image captioning can help the visually impaired individuals by describing the content in a meaningful way, thereby enabling them to understand the real-world scenes in a better way. For example, there are many applications like, TapTapSee, SeeingAI, etc. developed for visually impaired persons. These applications use the device camera to identify people and objects, and then the app audibly describes those objects. Medical image captioning highlights the relationships between image objects and clinical findings, which makes it a very challenging task. The generation of medical reports not only reduces the doctor's workload and accelerates clinical workflow but also aids in the efficient exploitation of medical content. Therefore, this produces faster and more accurate interpretations of findings, and offers important assistance to doctors.

Autonomous vehicles use visual captioning technology to describe the surroundings for efficient vehicle-to-passenger communication. This helps in identifying pedestrians, other vehicles, or obstacles thereby, allowing for more natural communication between humans and machines in collaborative settings. Furthermore, Semantic segmentation aids in object recognition within the framework of image

analytics, but it is unable to provide an explanation of the objects' relationships with one another through the use of verbs or contextual data. For instance, a security camera might identify a person and an automobile, but it might not disclose that the person is breaking into a car. In addition to alerting people to see the images and videos and take action, automatic caption generation can assist in identifying these occurrences.

Image captioning can also benefit the search engines and content management systems can more readily index and retrieve images based on their content when text descriptions are included for photos. This is especially important for huge datasets where users need to search through thousands of visuals to identify a specific image. Vegetable diseases may now be identified from leaf images thanks to recent developments in deep learning technology. In terms of accuracy, stability, and portability, the current computer vision-based approaches for disease recognition have demonstrated impressive accomplishments. Nevertheless, these approaches lack a textual foundation to back up the users' judgment and are unable to offer a basis for decision-making for the outcome. Therefore, with the help of image captioning system, the disease recognition model is able to diagnose the type of disease by using its physical characteristics and provides a tool to generate description sentences by analyzing the visual features in the image.

Visual Storytelling is yet another application and research area of image captioning which has the ability to capture the nuances of complex scenes involving multiple objects, interactions, and contextual elements. This enables the generation of detailed and contextually rich textual descriptions that effectively convey the essence of the visual content.

Such applications of visual image captioning motivate us to develop intelligent, robust and efficient captioning frameworks which helps to generate image description in the form of natural language sentences. Therefore, in this thesis, novel factual, stylized and dense captioning frameworks are established which describes the images either factually, or by highlighting a particular style or emotion (romantic or humorous). Also, a novel framework is developed that generates a paragraph-based description of images by describing more minute details of an image and generating coherent and diverse sentences.

1.7 Significance of Study

Image captioning is an amalgamation of Natural Language Processing (NLP) and Computer Vision (CV). This process is an automatic description of contents of an image in the form of natural language sentences. The very nature of it makes it suitable for applications such as image retrieval, human-robot interaction, aid to the blind, radiology report generation, agriculture, visual storytelling, and visual question answering and the like.

The study's main conclusions encourage a broader framework for image captioning in numerous real-world applications. Such applications will support the daily routine that most users share. Another important significance of this study is the establishment of a state-of-the-art that will enable the research community to delve deeper into this field.

1.8 Thesis Overview

Chapter-2 provides a detailed study conducted on different state-of-the-art and their analysis, i.e., merits and demerits for visual image captioning. Also, the study

conducted presents different evaluation parameters utilized to evaluate the performance of image captioning models. Furthermore, the detailed study conducted helped to draw an outline of research gaps in the concerned area. The final research objectives are defined which are addressed in the thesis later.

In Chapter-3, two intelligent factual image captioning models are presented. The first model is a lightweight transformer model with GRU integrated decoder. The proposed lightweight transformer utilizes a single encoder-decoder architecture for generation of captions. The second model is the XGL-Transformer model for image captioning that generates factual-based descriptions of images by capturing higher-order interactions thereby, improving the caption generation model.

Furthermore, in Chapter-4 a novel stylized-based image captioning framework is presented. The proposed framework that learns both unstructured (semantics) and structured (style) feature distributions jointly to generate controlled and plausible stylized captions. In Chapter-5, paragraph-based image captioning framework is discussed. The proposed framework leverages a multi-level variational autoencoder based transformer that captures consistent long-term structure and provides correlation between visual and long-text embeddings for the generation of intermediate paragraph-based descriptions.

Chapter-6 presents a novel caption summarization transformer that provides a summarized caption for a given image. The summarization transformer integrates two image captioning tasks, factual and stylized image captioning to output a caption that describes the content of image by highlighting the factual and the stylized content. Also, the thesis discusses about two main applications of image captioning in Chapter-7. The chapter focusses on aid to blind and radiology report generation applications of

image captioning. The proposed models generates syntactically and semantically correct sentences by describing important objects, attributes, and their relationships with each other.

Finally, Chapter 8 highlights the important conclusions drawn from these methods and gives the details of future scope of work.

Chapter-2

Literature Review

The state-of-the-arts for coherent and reliable captioning for images are covered in this chapter. In order to comprehend how different captioning tasks (like stylized image captioning, paragraph-based image captioning, etc.) and captioning methods have evolved over time and improve descriptions of images, the literature can be generally divided into two categories: traditional techniques and deep-learning-based techniques. Further, the chapter also discusses about different evaluation parameters which are utilized to test the performance of the captioning models.

2.1 Image Captioning Techniques

Image captioning, a popular area of research in the field of Artificial Intelligence (AI), deals with mainly two domains for analysis of image: (i) image understanding (ii) language description. Image understanding deals with the detection and recognition of objects that help extract semantic features whereas language description carries out the representation of sentences using both linguistic and semantic understanding of the language. The challenge is to design an image captioning model that can generate more human-like rich descriptions of images with the understanding of objects or scene recognition in an image and the relationship among them. The image captioning problem is categorized into two main categories as shown in Fig 2.1: (i) Traditional Techniques (Retrieval-Based and Template-Based methods) and (ii) Deep Learning- Based methods. Retrieval-based techniques [7] [27] retrieve the closest matching images and generate descriptions as a caption of the query images.

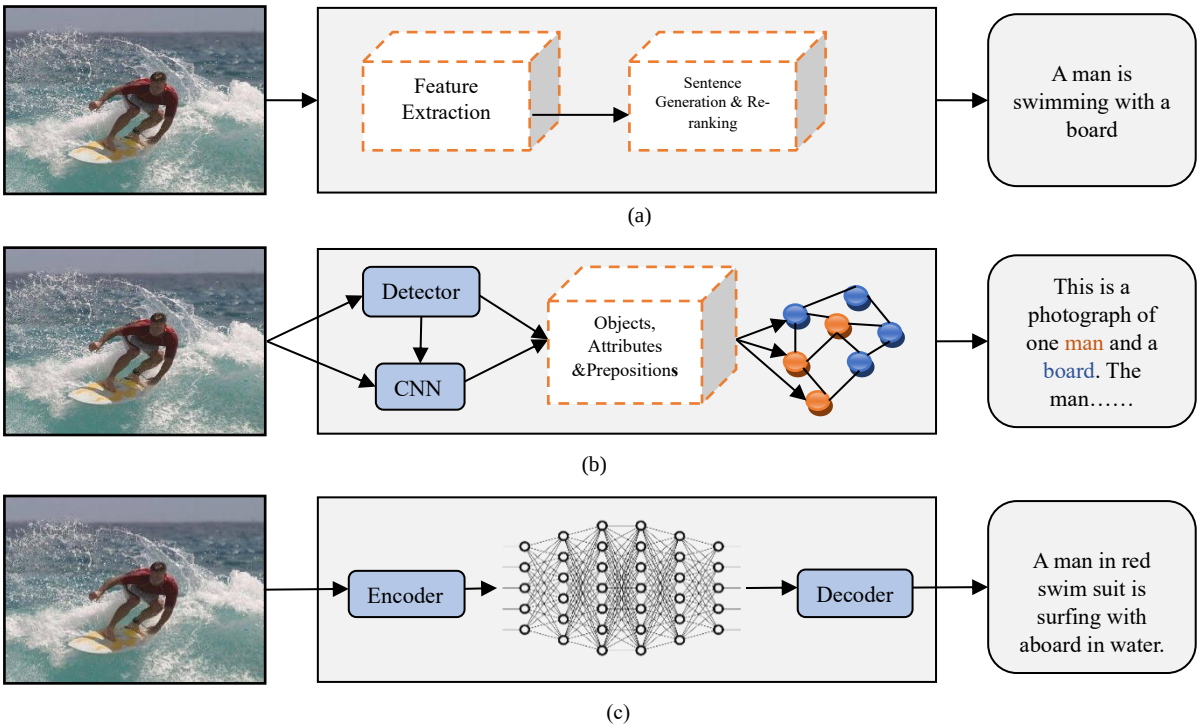


Fig 2.1: Image Captioning Techniques (a) Reterival-Based Caption Model (RCM), (b) Template-Based Caption Model (TCM), (c) Deep-Learning-Based Caption Model (DLCM)

These methods use re-ranking to produce correct sentences but fail to adjust descriptions for new images. Template-based image captioning models [28] [29] generate descriptions with predefined syntactic rules. Such methods cannot generate meaningful sentences as they cannot express visual content correctly. Nonetheless, the field of image captioning has gained popularity in the recent past owing to the introduction of deep-learning architectures [30] [31]. These architectures utilize encoder-decoder structures to understand images. Further detailed categorization of deep-learning-based image captioning tasks and methods is shown in Fig 2.2.

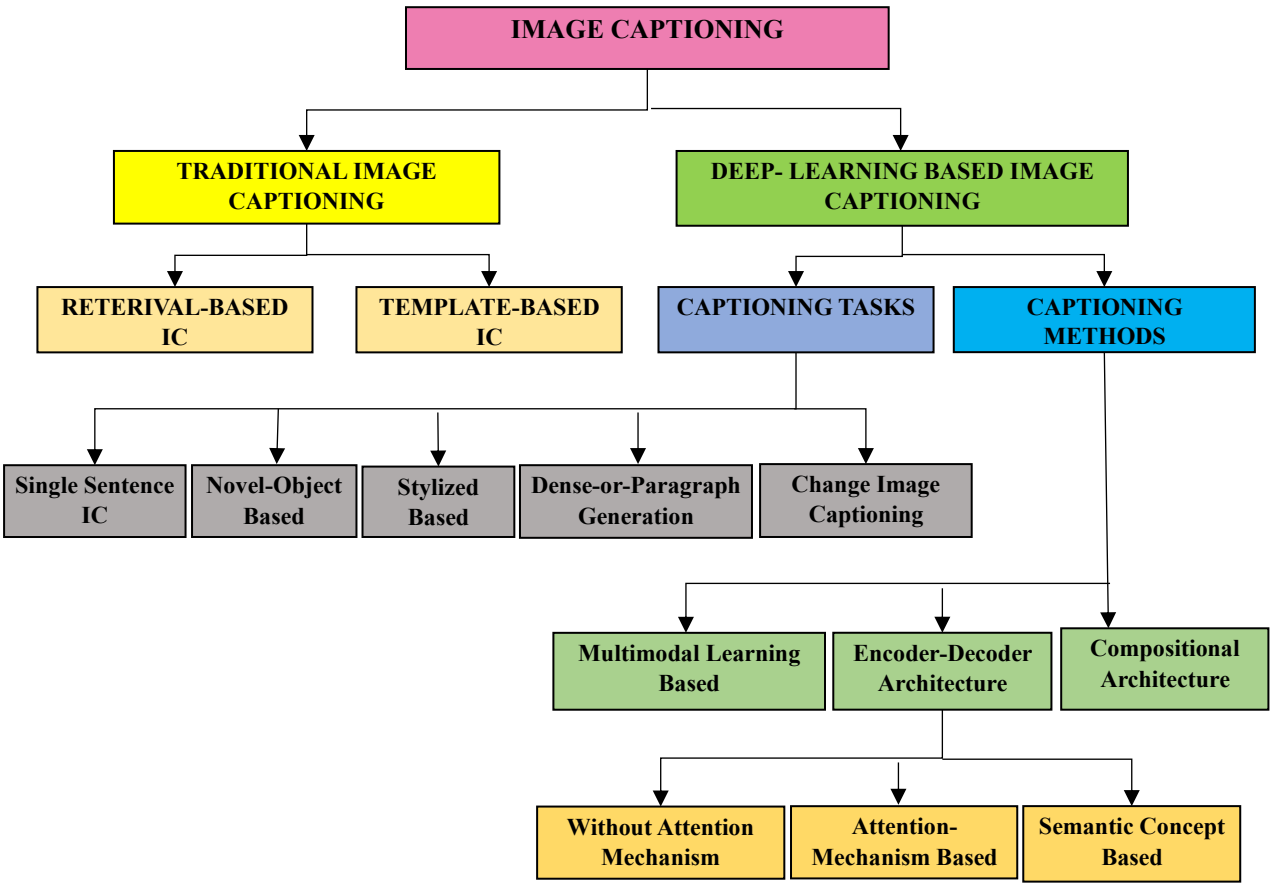


Fig.2.2: Taxonomy of Classification of Image Captioning Techniques

2.1.1 Traditional Image Captioning Techniques

Traditional techniques for image captioning can be categorized as Retrieval-based and Template-based techniques. Different methods related to traditional techniques with their pros and cons are discussed in the following sections.

2.1.1.1 Retrieval-Based Image Captioning Techniques

The most traditional form of image captioning technique is retrieval-based image captioning. It uses a query image and produces a caption for the given input image by retrieving a sentence or a set of sentences for a pre-specified pool of sentences. The caption is generated either as a single sentence or a combination of those retrieved sentences. [7] established a meaning space $< object, action, scene >$which links

sentences to images. It created a system that provides rich and subtle representations of information by computing a score and linking an image to a sentence. The score is later used to attach the descriptive sentence to a given image. The score closest to the query image is used to select the final description of the image as a caption. The work [32] employed global image descriptors to retrieve a set of images from a web-scale collection of captioned photographs and utilized details of the retrieved images to perform re-ranking according to the similarity in contents of the query image. [27] introduced a dataset for sentence-based image descriptions and evaluated using a ranking concept. It assumed that for a given query image there always exists a sentence that is appropriate for it [27]. This assumption is not always true. Therefore, instead of using retrieved sentences as descriptions of query images directly in the other line of retrieval-based research, retrieved sentences are utilized to compose a new description for a query image. To project image and text items into a common space, Canonical Correlation Technique [33] [34] correlated different captions generated for each training sample. Further, it measured cosine similarity to determine the similarity in documents for text analysis in new common space by and selects the top-ranked sentences which act as a description for a query image. [35] proposed a method for the generation of a description for a particular query image. It extracted global features to retrieve a set of query images which were further trained for the selection of phrases from the ones associated with retrieved images and finally a description of images is generated based on the selected relevant phrase. [36] proposed a tree-based method similar to [35], which utilized web captioned images. The disadvantages of retrieval-based image captioning methods are evident. These methods generate a description for query images in well-formed human-written sentences. The generated descriptions are

grammatically correct and fluent which means they can easily extract semantic information but requires training datasets that contain all types of attributes that adapt to new combinations of objects or novel scenes. Under certain conditions, generated descriptions may even be irrelevant to image contents as this method is not good in discovering words outside the training data. Retrieval-based methods have large limitations to their capability to describe images.

2.1.1.2 Template-Based Image Captioning Techniques

Template-Based caption generation technique generates captions both syntactically and semantically via a very constrained process. For a given image, this technique first detects a set of visual concepts (objects, attributes, information from images) and uses a specified grammar rule which combines the information and describes the images by filling the obtained data or information into the pre-defined blanks of sentence template. [29] proposed a method where nouns, verbs, scenes, and prepositions (known as quadruplet) were used to describe a sentence template. Description of images was done by using a detection algorithm [37] [38] that provided an estimate of objects and scenes in the image and further, the method employed a language model [39] to predict words, scenes, and prepositions for the formation of captions of the input image. Such techniques uses triplet of a scene of an object or an action that fills the gaps of templates in the format <<adj1, obj1>, prep, <adj2, obj2>> for encoding recognition. [40] proposed an algorithm that first used an image recognizer the purpose of which was to obtain visual information from the image, including objects, their attributes, and the spatial relationships between different objects. Furthermore, Conditional Random Field (CRF) helped to render the image contents and generates image description as a tree generating process based on visual

recognition results and represented images by using <objects, actions, spatial relationships> triplets [28] [41].

Methods discussed above use visual models which predicate individual words from a query image in a piece-wise manner. To generate more descriptive sentences under template-based learning, phrases are used in the generation of sentences [42]. Therefore, many methods have been proposed utilizing phrases under template-based image captioning. The captions generated by template-based image captioning methods are syntactically correct, and the descriptions yielded by such methods are usually more relevant to image contents than retrieval-based ones.

However, there are some disadvantages of these methods also. Since template-based description generations are strictly constrained to image contents recognized by visual models, there are limitations to coverage, creativity, and complexity of generated sentences due to the availability of a small number of visual words. Moreover, compared to human-written captions, using rigid templates as the main structure of sentences, generated descriptions less natural.

2.1.2 Deep-Learning based Image Captioning Techniques

Retrieval-based and template-based captioning of images were adopted mainly in the early work. With recent advancements in deep neural networks, researchers are embracing deep-learning-based image captioning techniques. Deep neural networks are extensively embraced for tackling the image captioning task. Therefore, the classification of deep neural network-based image captioning is outlined based on the two subcategories namely: i) image captioning task, and ii) image captioning methods. Detailed classification for different tasks and methods for deep-learning-based image captioning is shown in Fig 2.2. In this section, all the state-of-the-art in Fig. 2.2 are

reviewed in detail and a further overview of all state-of-the-art deep-learning-based image captioning techniques is presented in Tables 1-9.

2.1.2.1 Image Captioning Task

Image descriptions generated by the captioning model can be in the form of single sentence or in the form of paragraph, or the description generated may enhanced by the type of style or sentiment or may include novel object description, or the caption generated may incorporate the changes in the before and after images of a particular scene. Therefore, deep-learning based image captioning techniques can be classified into various image captioning tasks i.e., (i) Single Sentence IC, (ii) Novel-Object-Based IC, (iii) Stylized IC, (iv) Dense or Paragraph-based IC and, (v) Change Image Captioning. The detailed overview of various image captioning tasks with their state-of-the-art techniques is discussed and analyzed in *sub-sections 2.1.2.1.1 to 2.1.2.1.5* sub-section focusses on the above-mentioned image captioning tasks.

2.1.2.1.1 Single Sentence Image Captioning

An automated Single Sentence Image captioning generates the contents of the whole image in the form of a single sentence. The caption generated at the output may adopt any type of deep-learning-based captioning method like encoder-decoder based method, semantic concept-based methods, attention mechanism-based methods, etc. Further, detailed discussion of single sentence IC task is provided in sub-section 2.1.2.2

2.1.2.1.2 Novel-Object-Based Image Captioning Techniques

Recent deep-learning-based image captioning techniques have achieved very favourable results, but these techniques depend mainly on the sentence captions and

their paired image datasets. These methods generate captions for objects within the context because requiring a large set of training image-sentence pairs. The novel-object-based image captioning technique is capable of describing novel objects which are not present in paired image caption datasets. This technique is based on the following three steps: (1) unpaired data of image and text are trained by a separate language model and a classifier. (2) a deep caption generation model is usually trained on data (paired-image). (3) the models trained in (1) and (2) are combined and trained which generated the descriptions for the novel objects. Novel object-based captioning of images is not only trained for image-text paired sets but is also trained for unpaired ones. [43] presented a Deep Compositional Captioner (DCC) which generated the captions for unseen objects present in the images. This model described novel objects and their interactions with other objects. [44] described a copying mechanism known as LSTM-C for caption generation for novel objects. In this method, a classifier was developed for novel objects by using a separate dataset for object recognition. It integrated appropriate words for output captions using an RNN-decoder with a copying mechanism. With the increase in complexity of the caption generation for novel objects, a Novel-Object Captioner (NOC) [45] has been introduced to generate captions for unseen objects in images. It learned semantic knowledge and various external sources to recognize various unseen objects. This model exploited semantic information to generate captions in the ImageNet dataset for hundreds of object categories that are not observed in MSCOCO. [46] introduced a concept of zero-shot novel object caption generation using Decoupled Novel Object Captioner (DNOC). It generated novel objects descriptions without extra training sentences. The zero-shot learning technique bridged the gap between visual and textual semantics. [47]

discussed a new pointing mechanism-based framework and is also known as LSTM-P or LSTM with pointing that facilitated vocabulary expansion and encouraged global coverage of objects in the sentences generated. LSTM-P provided superior results on COCO and ImageNet datasets when compared with other state-of-the-art methods [45] [44]. [48] encouraged the development of captioning models that could learn visual concepts from other object detection datasets. However, when these models have been applied in the wild a much larger variety of visual concepts are to be learned. [49] presented a novel network structure known as Cascaded Revision Network (CRN) that described an image using existing vocabulary from in-domain knowledge. With lesser out-of-domain knowledge, generated captions may contain ambiguous words for images with novel objects. Re-edit is done for primary captioning sentences by a series of cascaded operations after which external knowledge is utilized which selects more accurate words for the novel objects and hence generates accurate captions for the unseen or novel objects. Caption generation for novel objects is a highly desirable yet challenging task. Therefore, [50] described VIsual VOcabulary pretraining (VIVO) that trained a multi-layer transformer model to generate fluent captions for novel objects along with the locations of these objects.

2.1.2.1.3 Stylized Image Captioning Techniques

Based on the contents of the images, existing models generated descriptions of attributed due to factual descriptions and did not consider the stylized part of the image from other patterns. Stylized captions are considered to be more expressive and attractive in comparison to the flat description of the generated images. Fig. 2.3 shows the block diagram representation of the stylized technique. This method generates the information from images using a CNN-based encoder. A text corpus is prepared which

extracts various stylized concepts such as romance, sentiments, etc. From the information generated from the CNN-based encoder and generated corpus, the language generation block generates attractive captions. This technique has become very popular as this is mainly used in many real-time applications. For example, people nowadays upload many photographs on many social media platforms, which need attractive and stylized captions for them. [51] proposed a novel image captioning technique known as StyleNet that generated attractive captions with the addition of various styles. The factual and style factors are separate from the captions generated with the use of CNN and a factored LSTM architecture and outperform the existing approaches with the FlickerStyle10K dataset which contains 10K Flicker images with humorous and romantic captions. Attractive and stylized captions can be generated with the use of multitasking sequence-to-sequence training by identifying style factors. In our day-to-day conversations, decision making and interpersonal relationships, various nonfactual expressions such as shame, pride, etc. are used. [52] described a method known as SentiCap which can generate image captions using positive and negative sentiments. This method combined two CNN + RNNs running in parallel in which one is responsible for the generation of non-factual words while the other generates the words with sentiments. This technique produces emotional image captions using only 2000+ training sentences which consisted of sentiments and produced 86.4% positive captions. Stylized image captioning has gained popularity in the past few years and therefore advancements are being made in this technique by multi-style image captioning. [53] claimed multi-style image captioning, known as MSCap with a standard factual image caption dataset and a multi-stylized language corpus with unpaired images. This model contains four modules namely style

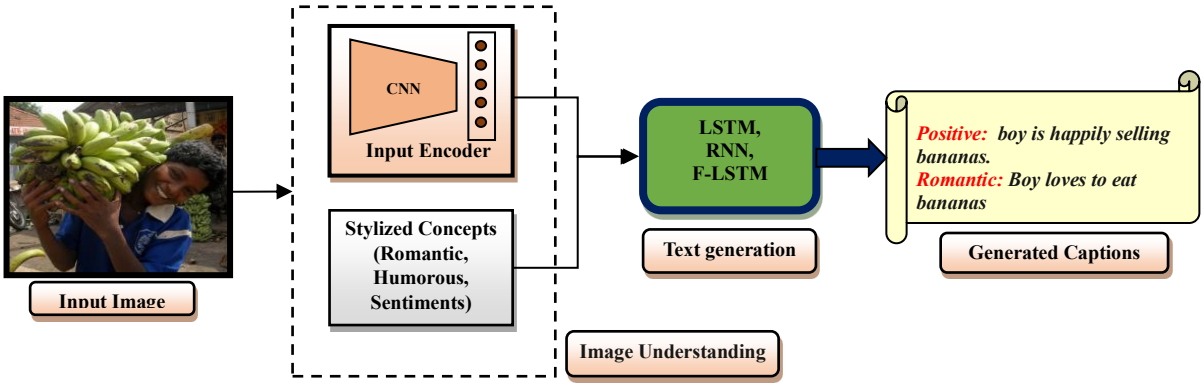


Fig 2.3: Basic Block Diagram representation of Stylised Image Captioning

dependent caption generator, caption discriminator, a style classifier, and at last experiments are conducted which demonstrates the outstanding performance of the work in [53]. [54] proposed a new variant of LSTM named style-factual LSTM and was used for the generation of captions that had a specific style. Without the use of extra ground-truth supervision, the method proposed in [54] outperformed the various state-of-the-art approaches by using factual and stylized knowledge. Stylized-based captioning suffers from limited style variation and content regression. [55] described a controllable stylish image description model which generated various stylist captions by plugging in style-specific parameters. [56] proposed a method MemCap which that explicitly encoded the knowledge about linguistic styles with memory mechanism. MemCap first extracts content relevant style knowledge from memory module with attention mechanism and further, the extracted knowledge was incorporated into a language model. The effectiveness of [56] is demonstrated by StyleNet and FlickerStyle10K datasets. To increase the diversity of captions [57] presented a framework named Mixture of Recurrent Experts (MoRE) that derived SVD from weighting matrices of RNN. This model generated diverse and stylized descriptions of images also in terms of content accuracy. [58] described stylized image captioning

with paired stylized data that extracted style phrases from small-scale stylized sentences and graft them to large-scale factual captions.

2.1.2.1.4 Dense or Paragraph Image Captioning

The tasks and methods mentioned in sub-sections from *2.1.2.1.1 to 2.1.2.1.3* and 2.1.2.2, can generate one sentence to describe an image. These models extract features by considering entire all images but some regions like background and objects from the images cannot be extracted and these regions of the image do not have caption generated. This mechanism fails to retain and process effective information for caption generation. Moreover, the detected attributes cannot be combined by the models to generate a sentence that is limited by specific grammar. Procedural steps involved in dense image captioning techniques are as follows: (1) Region proposals are generated for the different regions of the given image. (2) CNN is used to obtain the region-based image features. (3) The outputs of Step 2 are used by a language model to generate captions for every region. Fig. 2.4 (a) exhibits the basic principle involved in the dense captioning technique. An example of image paragraph generation is shown in Fig 2.4 (b) which supports that dense captioning generates the description for each region and each description is independent, while image paragraph can generate the paragraph with related sentences. [59] defined DenseCap, a dense captioning model that jointly addressed localization and description tasks with the help of Fully Convolutional Localization Network (FCLN) architecture. The FCLN architecture is composed of CNN, RNN, and novel dense localization layer. DenseCap [59] network was evaluated on the Visual Genome dataset which provided improvements in speed and accuracy. [15] used Region Proposal Network (RPN), trained to generate high-quality region proposals. RPN is a fully convolutional network, merged

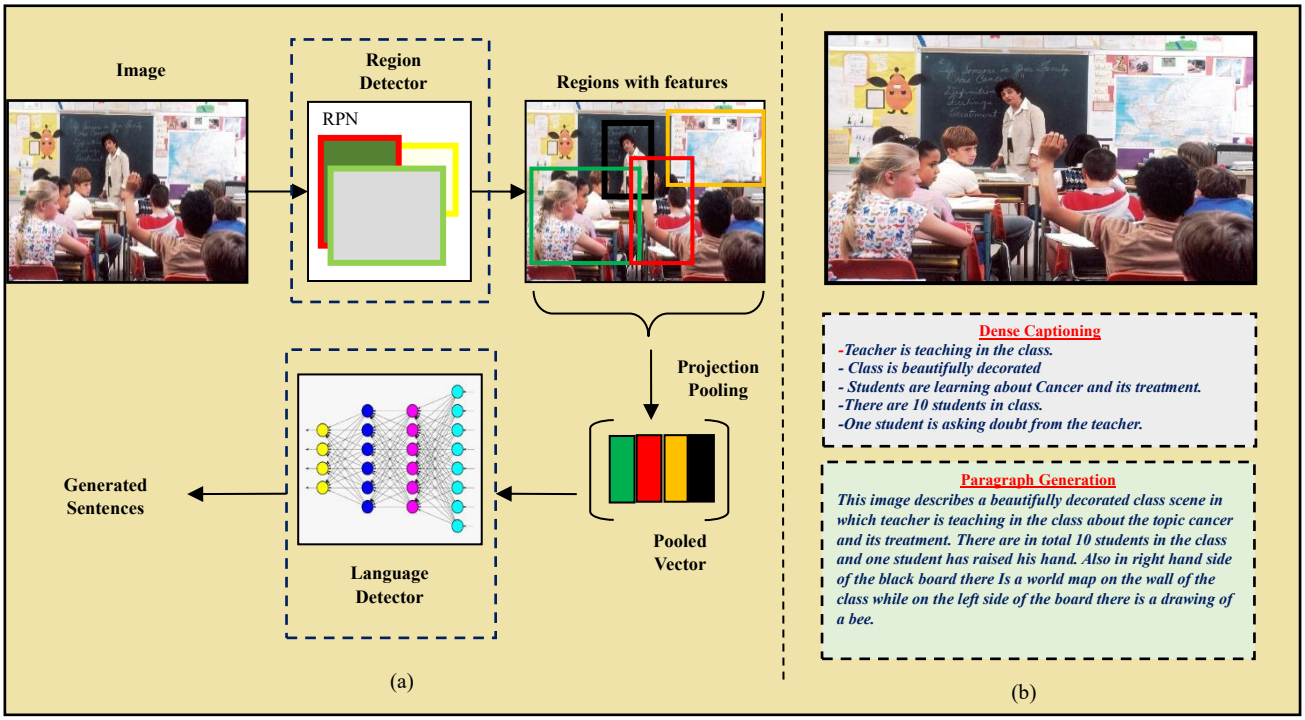


Fig 2.4: (a) Fundamental Steps Involved in Dense Image Captioning (b) An example to illustrate Dense Captioning & Paragraph Generation

with Fast R-CNN for detection. [60] presented a unified method that was based on two novel concepts: (1) joint inference; which jointly depends on visual features and predicted captions of regions and (2) context fusion; which combines context features and visual features for a description of rich captions for a particular image to generate dense captions. Dense Semantic Embedding Network-LSTM (DSE-LSTM) [61] preferred to extract dense semantic embeddings from DSE-network and predicted a word for every semantic feature at each step. [62] introduced relational captioning that generated multiple captions with the help of relational information between objects in an image. The work [63] [62], presented a multi-task triple-stream network (MTTSNet) to generate diverse and rich captions on large-scale datasets.

Recent advancements in dense image captioning are related to precise feature extraction that realizes a complete understanding of an image by localizing and

describing multiple salient regions. [64] defined precise feature extraction (PFE) to provide enhanced dense captions of images. The dense relational captioning method [62], generated multiple captions that provide relational information between objects and explicit descriptions for a different combination of objects. This method was advantageous in terms of diversity and amount of information. Dense Image captioning models are relatively independent and generate unrelated captions. Paragraph generation is based on finding the relationship between objects and co-referencing while generating sentences. [65] attempted to generate dense paragraphs to describe an image in the form of detailed and unified stories by detecting semantic regions in images and using a hierarchical neural network. It helped to generate coherent sentences which increased the complexity of the model. [66] presented a Depth- image with details. The effectiveness of image paragraph generation was reduced due to a lack of diversity between sentences. [67] considered the sequence level training for image paragraph generation and produced diverse paragraphs with an integrated penalty on trigram repetition.

Description of an image by a full paragraph involves organizing sentences orderly, coherently, and Depth-aware Attention Model (DAM) for paragraph generation. In DAM [66] the depths of image areas were estimated and spatial relationships between objects were revealed by a linguistic decoder. The model generated the paragraph description logically and coherently. [68] managed to generate longer, richer, and more fine-graded descriptions of an image in a paragraph by presenting a Context-Aware Visual Policy (CAVP) network. [69] also attempted a method for coherent paragraph generation using CNN. Dual-CNN decoder helped produce a semantically coherent description of images and provided state-of-the-art

results on the Stanford Image paragraph dataset. [70] presented a transformer-based method known as meshed memory transformer for caption generation. The method devised mesh-like connectivity at the decoder to exploit low-level and high-level features. [71] presented a hierarchical topic-guided image paragraph generation framework also known as the Visual-Textual Coupling Model (VTCM) that coupled a visual extractor with a deep topic model. LSTM and transformer are jointly optimized to guide the generation of paragraphs and provided improved results over [70] for the Stanford Image Paragraph dataset. The work [72], described visual scenes with long sentences. It included perceptual and semantic information and described what is in the image. [73] presented Hierarchical Scene Graph Encoder-Decoder (HSGED) that generated coherent and distinctive paragraphs and achieved state-of-the-art results on the Stanford Image Paragraph dataset. The image paragraph captioning method should generate consistent sentences rather than contradictory ones. To overcome this, [74] presented a method that incorporated incorporate objects' spatial coherence into a language-generating model. This method achieved promising results by extracting effective object features for image paragraph captioning.

2.1.2.1.5 Change Image Captioning (CIC)

The captioning tasks discussed in sub-sections 2.1.2.1.1 to 2.1.2.1.4, deal with the detection and recognition of objects that help extract semantic features and describe these features in the form of natural language. In this section, we will elucidate about Change Image Captioning (CIC) which identifies a change and describes it in an incisive manner. CIC differentiates two images of a changing scene, captured at

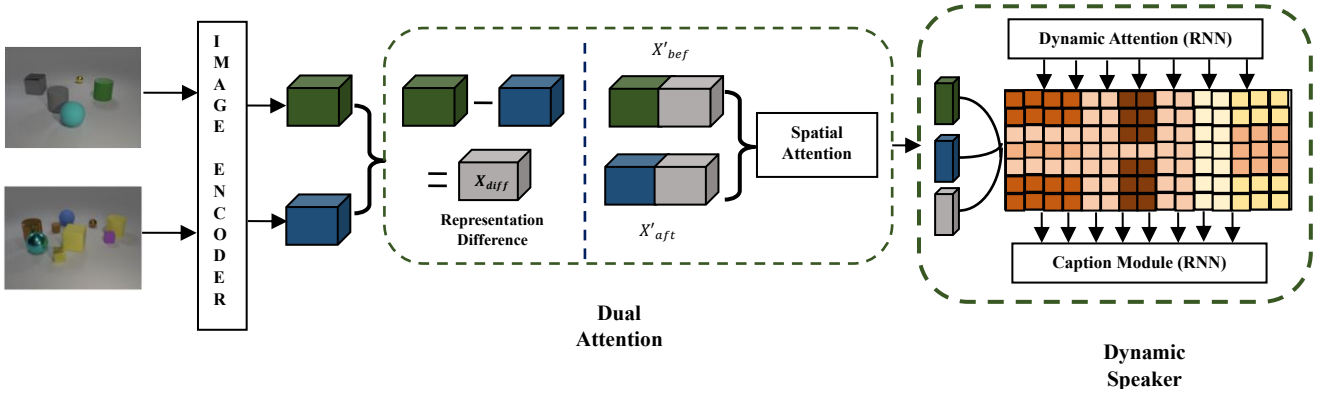


Fig. 2.5: Architecture of DUDA [75]

different time steps- before and after scene. This helps to observe analytically possible changes that occurred in the scene. This view of image caption generation gives an active and attentive eye on the attributes and the change in attributes with time steps in an image. This process requires an additional reference image to identify the change over time and generate the captions adaptively. Hence, we can say that the CIC method utilizes the time dimension along with the spatial dimension of an image. It is highly useful in real-time applications such as aerial imagery [76] [77] analysis for disaster response systems [78] and monitoring of land cover dynamics [79], and street scenes surveillance [80] .

CIC is a very challenging task that aims to describe the subtle difference between two similar images in the form of natural language. Images acquired from different camera perspectives and different illumination exposure make CIC more challenging. Therefore, change detection in an image becomes an essential step to generate a caption. Earlier methods [76] [77] [79] utilized an unsupervised approach to detect changes due to the high cost involved in labelling the large ground-truth samples. Further, the semi-supervised approach [78] discussed, relies on hierarchical shape representation. Another work [81] detected changes based on dense optical flow to

address the difference in viewpoints. The works [82] [83] addressed more subtle, fine-grained change detection, where an object may change its appearance over time. To tackle this problem, [84] estimated a dense flow field between images to address viewpoint differences. In this direction, Park et al. [75] defined a Dual Dynamic Attention Model (DUDA) based on attention mechanism rather than pixel-level difference or flow. The model distinguished relevant scene changes from illumination/viewpoint variations with the help of a dynamic attention scheme. The DUDA framework, Fig. 2.5, consists of two main components: (i) Dual attention, and (ii) Dynamic Speaker. Dual Attention processes multiple visual inputs separately and addresses Change Captioning in the presence of distractors. However, Dynamic Speaker predicts attention over the visual features at each time step and obtains the dynamically attended feature. Another work [85] presented Mirrored Viewpoint-Adapted Matching (M-VAM) encoder to distinguish viewpoint changes from semantic changes. M-VAM self-predicted the difference (changed) region maps by feature-level matching without any explicit supervision. Further, [86] worked upon a semantic relation-aware difference representation learning network. This network explicitly learned the difference representation in the existence of distractors. [87] focused on a training scheme that used an auxiliary task to improve the training of the network. This auxiliary network helped in the generation of precise and detailed captions and provided state-of-the-art results on benchmark datasets for change captioning. Viewpoint-Agnostic change captioning network with Cycle Consistency (VACC) [88] method explicitly distinguish between the real change from a large amount of clutter and irrelevant changes [89] worked upon Relation-embedded Representation Reconstruction Network (R^3Net) and provided state-of-art results on Spot-the-Diff

and CLEVR-change dataset. The existing scene changes captioning approaches recognize and generate change captions from single-view images. Those methods have limited ability to deal with camera movement, object occlusion, which are common in real-world settings. To resolve these issues, [90] [91] presented a framework that described changes from multiple viewpoints (or 3-D vision) in form of natural language.

Table 2.1: Overview of different Deep-Learning-based Captioning Tasks

Ref.	CT	Citations	Method	Dataset	Advantages	Limitations
[43]	Novel Object-Based Image Captioning	259	DCC	MSCOCO ImageNet	Describes new objects which are not present in current caption corpora, provides rich descriptions of images	Some sentences generated are grammatically incorrect but they do incorporate new words.
[44]		119	LSTM-C	MSCOCO ImageNet	Improvement in performance can be observed	Not suitable for large scale image benchmarks YFCC100M
[45]		146	NOC	MSCOCO ImageNet	The model can describe many more novel objects and provide state-of-art results	This model fails to describe new objects.
[46]		43	DNOC	MSCOCO	Evaluation examples contain unseen objects and no additional sentence data is available.	The complex structure of the model
[47]		41	LSTM-P	MSCOCO ImageNet	Covers more objects in the generation of captions and thus improves the captioning mechanism. Improved F1 scores	Placements and moments of copying novel objects in sentences are not fully yet understood in the literature.
[48]		11	nocaps	nocaps MSCOCO	Dataset provides better object detection and improvements in the captioning mechanism are obtained.	Analysis shows that there is significant room for improvement for the image captioning task.
[49]		22	Cascaded Revision Network	MSCOCO ImageNet	Can efficiently describe images with unseen objects.	METEOR score is lower than LSTM-C
[50]		1	VIVO	MSCOCO nocaps	Achieved new state of art results on nocaps and surpassed the human CIDEr score	Needs to leverage a large amount of vision data to provide

						improvement in visual vocabulary
[52]	Stylized Image Captioning	137	SentiCap	MSCOCO	Able to generate emotional captions for over 90% of the images and are evaluated using crowdsource and automatic evaluations.	More or less inappropriate in content description or sentiments or both for some images
[51]		197	StyleNet	FlickerStyle10K	Able to learn styles from a monolingual textual corpus. Can generate visually attractive and stylish captions	Romantic and humorous styles are combined but no improvements were there.
[53]		40	MSCap	COCO, SentiCap, FlickerStyle10K	Better fluency than StyleNet. Captions generated are fluent and relevant and are correctly stylized	MSCap rates can be improved which can further improve the efficacy
[54]		39	Styled Factual-LSTM	MSCOCO FlickerStyle10K	Captures both factual and stylized information	For negative caption generation, the performance is competitive with SentiCap
[55]		9	LSTM with Domain Layer Norm	MSCOCO	This model progressively includes new styles which are more preferred by human subjects.	Transfer accuracy of the source to humor is low and for lyrics style, it is also low
[56]		14	MemCap	SentiCap FlickerStyle8K	Generates sentences that describe the content of the image accurately and reflect the desired linguistic style appropriately	-
[58]		-	SAN	SentiCap FlickerStyle8K	The framework generates corresponding stylized captions for images in the largescale factual corpus.	Noise may be introduced in each process
[57]		1	MoRE	MSCOCO	Provides improvements in terms of accuracy, diversity, and styled captions	-
[15]	Dense or Paragraph Image Captioning	24897	RPN	PASCAL VOC 2007	Efficient and accurate region proposal framework. Provides object detection accuracy	Complexity increases with Fast R-CNN
[59]		1001	CNN-RNN	Visual Genome	Supports end-to-end training and efficient test-time performance and provides visually pleasing results.	Failure occurs in some cases where interaction between objects can be seen.
[65]		245	HRN	MSCOCO	This model interpretability generates descriptive paragraphs using the only subset of image regions and with	-

					the use of a wider vocabulary.	
[60]		97	LSTM	Visual Genome	This novel model incorporates joint inference and context fusion and achieves state-of-art performance on Visual Genome	Sequential modeling needs enhancements for this framework.
[68]		59	CAVP	Stanford Paragraph Dataset MSCOCO	Superior to RL-based methods and provides top-ranking performances on MSCOCO and Stanford image captioning datasets.	the models optimized by CIDEr or BLEU would be more superior over that by cross-entropy
[66]		21	CNN-LSTM + Attention	Visual Genome	Strengthen image paragraph captioning by enriching raw data with extra geometric information which also improves diversity	Takes a much longer time to generate paragraphs with diversity.
[67]		29	SCST + Repetition Penalty	Visual Genome	This work increases diversity in paragraph generation and provides substantial improvement in state-of-art techniques.	Many language issues arise for paragraph generation.
[64]		-	LSTM	Visual Genome	Provides better regional features which promote better implementation of region positioning and descriptions.	Complexity arises in the calculation of mAP (mean Average Precision)
[61]		23	DSE-LSTMS	MSCOCO Flickr30K	A bidirectional LSTM structure is used which captures previous and future contexts. TReLU can improve distinctness in the captions	MSCOCO is more challenging than Flickr30K in captioning and retrieval tasks.
[70]		173	M^2-Transformer	MSCOCO	Provide object details and small detections which achieve new state-of-art results on COCO.	This model is slightly worse on the ROUGE evaluation metric.
[69]		6	Dual-CNN	Stanford Paragraph Dataset	The model provides more efficiency by giving less training time and is effective with high CIDEr score obtained.	For the encoder side, more semantic information like positional relationships must be considered.
[62]		2	MTTSNet	Visual Genome	This model facilitates POS-aware relational captioning. This new framework can open new applications.	Suffers from visual ambiguity, geometric ambiguity, and illumination.
[71]			VTCM-Transformer	Stanford Paragraph Dataset	This method captures the correlation between image and text at multiple levels of abstraction and	The complexity of the model increases as this model also generates the topic

					learns semantic topics from images.	for the paragraph generated.
[72]		4		Stanford Paragraph Dataset	Generates both accurate and diverse image paragraphs by utilizing both visual and linguistic information	-
[74]		-	ORA	Stanford Paragraph Dataset	The proposed method generates slightly higher METEOR scores.	-
[73]		-	HSGED	Visual Genome Stanford Paragraph Dataset	Semantic and hierarchical knowledge of an image can be transferred into the language domain easily	-
[63]		37	MTTSNet	Relational Captioning Dataset	MTTSNet facilitates POS aware relational captioning	-
[75]	Change Image Captioning	43	DUDA	CLEVR-Change Spot-the-diff	The model is robust to distractors in the sense that it can distinguish relevant scene changes from illumination/viewpoint changes	The spot-the-Diff dataset is not the definitive test for robust change captioning as it does not consider the presence of distractors
[88]		-	VACC	CLEVR-DC CLEVR-Change Spot-the-diff	The cycle consistency module that evaluates the quality of caption	
[85]		6	M-VAM	CLEVR-Change Spot-the-diff	M-VAM encoder can accurately filter out the viewpoint influence and figure out the semantic changes from the images	Negative samples are seen due to resizing operations; as a result, objects become too small to be recognized.
[87]		2		CLEVR-Change Spot-the-diff	This method composed query image retrieval as an auxiliary task to improve the primary task of image change captioning	
[86]		-	SRDRL + AVL	CLEVR-Change Spot-the-diff	This method achieves state-of-the-art performances.	This method needs improvement to learn more fine-grained difference representation
[91]		4	Indoor CIC	3D-Dataset	First attempt for indoor scene change captioning.	There is still room for improvement, especially for object attribute understanding
[90]		7	3D-CIC	3D-Dataset	Three syntactic datasets are created.	The dataset should contain more

						complex scenes, object models with higher diversity, and placing object models at various locations in the scenes.
[89]		-	R3: Net	CLEVR-Change Spot-the-diff	This method can explicitly distinguish semantic changes from viewpoint changes	This model does not consider very slight movements as the decoder receives the weak information of change

2.1.2.2 Image Captioning Methods

Deep-learning based image captioning methods can be broadly categorized into three main categories namely: *(i) Multimodal-learning-based IC, (ii) Encoder-Decoder architecture-based methods, (iii) Compositional Architecture-based methods.* Multimodal-learning based image captioning methods deals with the learning of image and text jointly in multimodal-space. Further, encoder-decoder architecture-based methods follows encoder-based image understanding and decoder-based text generation whereas compositional architecture-based methods generate multiple captions at the output and re-rank those generated captions in order to generate high-quality image captions. Different models related to the above-mentioned image captioning methods with their pros and cons are discussed in the following sections.

2.1.2.2.1 Multimodal Learning-based Image Captioning Methods

Multimodal learning-based image captioning models learn both image and text jointly in multimodal space. The basic steps of such models include an image encoder, language encoder, projection of encoded vectors in multimodal space, followed by language decoder. Projection of image and text encoded vectors into multimodal space

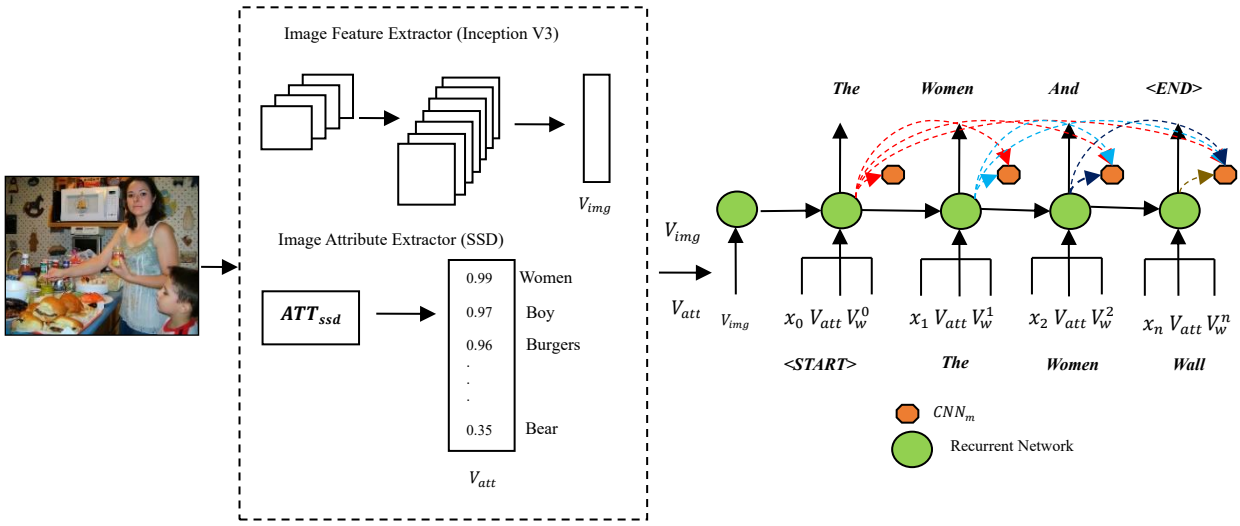


Fig.2.6: Illustration of Fusion Approach for Multimodal Image Captioning [92]

maps the image features into a common space with the word features. It helps these models to learn richer discriminant features for each sample and yield improved captions. [93] were the first to propose a multimodal learning-based image captioning method. This technique has an advantage over the traditional ones as this method describes images without the use of templates, syntactic trees, and structured prediction. This concept can also be extended to other modalities like audio. [94] proposed an extension of [93] which learns a joint image sentence embedding with the use of LSTM for sentence encoding. It used a new language model named as Structure-Content Natural Language Model (SC-NLM) for caption generation and reported results superior to [93]. [95] presented a bidirectional retrieval of images and sentences with the use of deep, multimodal embeddings of visual and natural language data. This model worked on a finer level and embedded fragments of objects and sentences to interpret predictions for the image-sentence retrieval task. Multimodal-RNN or m-RNN method [96] generated the probability distribution of a word given a previous word and an image. The effectiveness of m-RNN is evaluated on four benchmark

datasets namely IAPR TC-12, Flickr8K, Flick30K, and MSCOCO. [97] restored visual features from the given description. It can generate sentences and retrieve both images and sentences. This method defined an additional recurrent visual hidden layer with RNN that made a reverse projection which provided better results when compared with other state-of-the-art. [30] presented a hierarchical multimodal learning model with an attention mechanism. The model included a CNN network for image encoding, an RNN network for identification of objects in images in a sequential manner, and a multimodal learning-based RNN with an attention mechanism for caption generation with intermediate semantic objects and the global visual contents. [98] defined a sequence-to-sequence RNN model. This model extracted the object's features in an image and arranged them in order using a CNN-based structure that generates the corresponding words in the sentences.

Recent techniques for multimodal-based captioning of images are based on CNN and RNN models which have achieved excellent performance on captioning of images. [92] proposed a multimodal fusion method for the description of images as depicted in Fig. 2.6. This model generated captions in four parts i.e. a CNN for extraction of features, an attribute extraction model, RNN for prediction of words, and a CNN-based model for the generation of language. Extensive experiments were conducted on Flickr8K, MSCOCO, and Flickr30K which proves that the model proposed provides impressive results for the generation of captions for images. [99] presented a self-guiding multimodal LSTM (sgl-LSTM) model that handled an uncontrollable imbalanced real-world image-sentence dataset. It is based on multimodal LSTM (m-LSTM) that deals with noisy samples and can fully explore the dataset itself. The model outperformed the traditional RNN model-based technique [97] [30] in

describing the key components of the input images. [100] touched upon news image captioning application. News captioning is different from generic captions as news images contain more detailed information. To understand and learn news images, a multimodal attention mechanism is defined. It incorporates a multimodal pointer generation network to extract visual information. Experiments on the Dailymail test dataset and BBC test dataset provide improvements in results in terms of BLEU, METEOR, and ROUGE-L evaluation metrics.

2.1.2.2.2 Encoder-Decoder Architecture Based Image Captioning Methods

Encoder-decoder-based image captioning methods are based on an encoder-based image understanding module and a decoder-based text generation module. These methods are further categorized into three main categories namely: *(i) Encoder-decoder Architecture Without Attention Mechanism (ii) Attention mechanism-based architectures, and (iii) Semantic Concept-based architectures*. The detailed analysis with pros and cons of different encoder-decoder architectures-based methods is discussed in the following section.

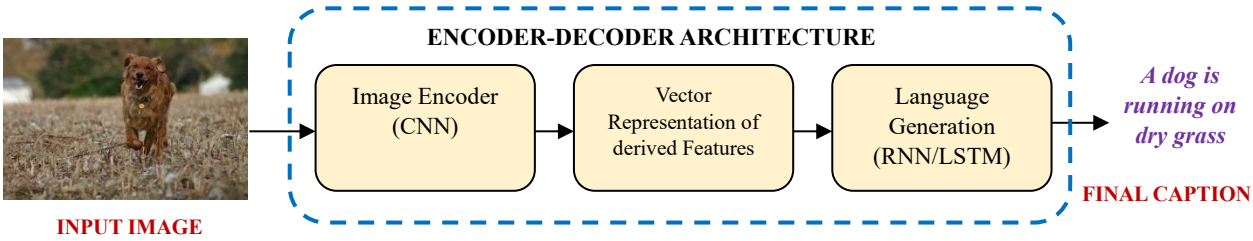


Fig.2.7: Basic Principle Involved in Encoder-Decoder Architecture Based Image Captioning

2.1.2.2.2.1 Encoder-Decoder Architecture Without Attention Mechanism

The neural-network-based image captioning methods work on the principle of the end-to-end framework. Encoder-Decoder architecture [101] [102] was originally designed to translate sentences between different languages. The idea behind adopting this method is to see the image captioning technique as a translation problem but with different modalities. Machine translation architectures extract features from hidden activation of CNN passes to an LSTM to generate a sequence of words. Encoder-Decoder architecture without attention is depicted in Fig.2.7. From this figure, two vital observations can be stated as follows:

i) Relationships between the detected objects and the scenes can be derived by using a CNN model.
ii) Image captions can be generated by using the output obtained in (1) to a language model which gets converted into words and combined phrases.

Kiros et al. [93] presented image captions generation word by word. [18] defined Neural Image Caption (NIC) generator that is similar to [93]. It is based on maximum likelihood estimation and proposes a CNN and an LSTM for the representation of images and the generation of captions respectively. This method is suspected to vanishing gradient problem because the information related to images is fed only at the beginning of the process and words are generated based on the previous hidden state and the current time step which continues until this process gets the end token of the sentence. As the process continues the role of initial words becomes weaker thereby degrading the quality of sentences generated. Hence, to overcome the challenges of LSTM for long-length sentence generation [103] presented an extension of LSTM known as guided LSTM or g-LSTM which successfully generated long-length

sentences [104] [105]. While designing, different length normalization strategies were considered to control the length of sentences. Various methods (multimodal embedding space) are being adopted to extract the semantic information from the generated captions. [106] designed another variation for encoder-decoder architecture that stacks multiple LSTM, also known as Long-term Recurrent Convolutional Network (LRCN) generate captions for static as well as dynamic images as well. In [107], a special type of text generation method was defined to generate a specific object or region description known as referring expression using unidirectional LSTM. The generated referring expressions help infer ambiguity in object/scene representation. Recent techniques for the detection and classification of objects [108] [109] exhibited that the deep hierarchal method performs better than shallower ones. [110] defined a deeper Bi-directional LSTM framework to generate semantically rich sentences. It utilized both past and future context information that helped in learning long-term visual language interactions. [111] incorporated high-level semantics concepts into encoder-decoder architecture through the combination of CNN and RNN in the form of attribute probabilities. This method achieved a significant improvement in the state-of-the-art for both image captioning and visual question answering. [112] opted for a semi-supervised learning method to train Deep Generative Deconvolutional Network (DGDN) [113] as a decoder and deep-CNN as an encoder. It can model the image in the absence of associated captions, thus known as semi-supervised. [114] discussed attention-based encoder-decoder framework, unified Graph Convolutional Network with LSTM to define the semantics and spatial object relationships in image encoder. Extensive experiments were carried out on the COCO dataset and better results were obtained which remarkably increased CIDEr-D from 120.1 to 128.7.

[115] introduced a novel decision-making framework for image captioning. This model utilized a “policy network” and a “value network” to collaboratively generate captions. [116] presented Conditional Generative Adversarial Networks (CGAN) that aim to improve the naturalness and diversity in generated captions. This model jointly learns a generator to produce descriptions conditioned on images and is an evaluator to assess how well a description fits with the visual content. [117] presented an extension of traditional reinforcement learning-based architecture. It dealt with the inconsistent evaluation problem and distinguished whether generated captions are human-generated or machine-generated with the use of a discriminator (CNN or RNN based structures). [118] considered the problem of optimizing image captioning systems using reinforcement learning. They worked upon a form of REINFORCE algorithm and presented Self-Critical Sequence Training (SCST) for image captioning. This method provided an improvement in the CIDEr score from 104.9 to 114.7.

Herdade et al. [119] used an object relation transformer with an abstract feature vector to provide a spatial relationship between input detected objects. This approach exhibited improvements on all captioning metrics for the MSCOCO dataset. [120] combined CNN with an attention-based Gated Recurrent Unit (GRU) to generate a better description for a given image. It is less complex and easy to train for about 100 epochs. [121] presented a CNN-RNN based encoder-decoder network for image captioning in the Hindi Language by manually translating the popular MSCOCO dataset from English to Hindi. It provided state-of-art results with a BLEU-1 score of 62.9, BLEU-2 score of 43.3, BLEU-3 score of 29.1, and BLEU4 score of 19.0. Image captioning is used in several applications as mentioned earlier, this technique can also be used to capture eating episodes of the subject and further record rich visual

information. [122] defined a framework for passive dietary intake monitoring. It is known as egocentric image captioning which unifies food recognition, volume estimation, and scene understanding. This work [122] which experimented to generate captions for egocentric images is the first-ever work in egocentric image captioning.

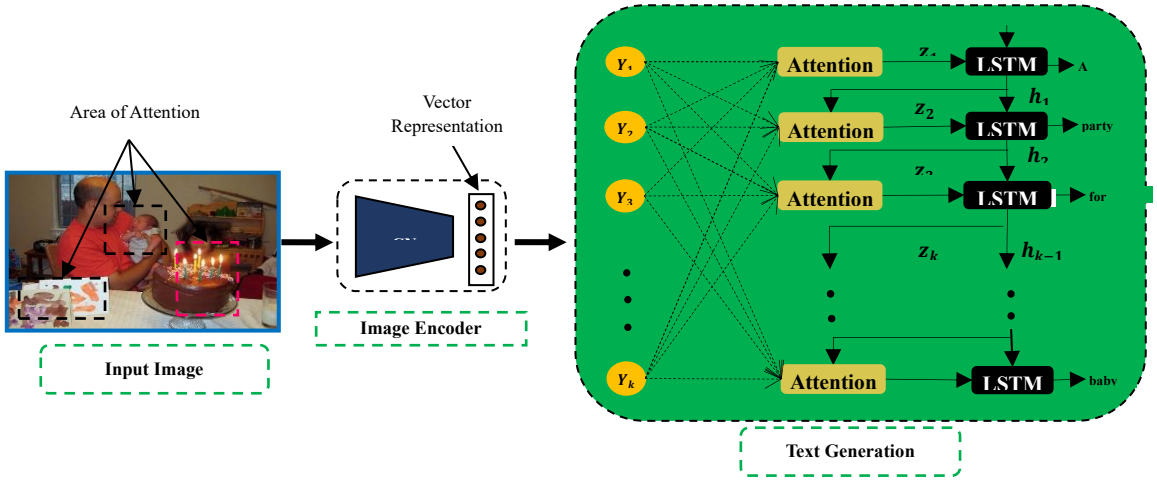


Fig 2.8: Basic Structure for Attention-Based Image Captioning

2.1.2.2.2.2 Encoder-Decoder Architecture with Attention Mechanism

For Image captioning, CNN-based encoders [93] have been used to extract visual features for the image and an RNN-based decoder to convert the extracted visual features into natural language. Encoder-decoder-based methods [117] [119] are unable to analyze the image over time. Also, such methods do not consider the spatial aspects of an image that are relevant for the captioning of images instead these methods generate captions for the scene as a whole. To overcome these limitations, the Attention mechanism-based captioning of images came into existence. Attention mechanisms are now being widely applied which yields significant improvements for various tasks. The essential function of the attention mechanism is to map the text description of the image to different regions of the image. For the attention-based captioning model, C_t is the feature map extracted from the region after CNN which is defined by:

$$c_t = g(V, h_t) \tag{2.1}$$

Where,g is the attention mechanism, $V = [v_1, v_2, \ldots \ldots \ldots v_k]$ is the vector representing the image features corresponding to k region of the image and h_t is the hidden state of RNN at time t. Attention distribution of k regions $b_t$is:

$$\widetilde{b_t} = w_h^T \tanh\left(W_v V + \left(W_g h_t\right) I^T\right) \tag{2.2}$$

$$b_t = softmax\left(\widetilde{b_t}\right) \tag{2.3}$$

From the equations (2.2) and (2.3) final c_t is given by:

$$c_t = \sum_{i=1}^{k} b_{ti} v_{ti} \tag{2.4}$$

These methods are becoming increasingly popular in deep learning for image captioning. Various attention mechanism methods discussed in this paper include: *(1) Soft-Attention* [123]*, (2) Spatial and channel-wise attention* [124]*, (3) adaptive* [31] [125] *(4) multi-head and self-attention* [126] *(5) Semantic Attention* [127]. The basic model for the attention-based mechanism is shown in Fig. 2.8 wherein CNN is used to extract the information from the input image. The information so extracted is fed to the language generation part. It generates i words or phrases based on the information. The hidden state of LSTM h_i, used to select the relevant part of the image. z_i ,The output of the attention model is used as an input to LSTM for extraction of salient features of the image focused in each time step of the language generation model. The generated captions are updated dynamically until the end of the language generation model.

Xu et al. [128] were the first to introduce the concept of attention-based captioning of images. It described the salient contents of an image and generated corresponding words for the salient parts at the same time. The method is based on

stochastic hard attention and deterministic soft attention for generating captions. The use of an attention-based mechanism in captioning of images provided improvement in BLEU and METEOR metrics. [129] discussed another method in the category of attention-based mechanism which extracted the flow of abstract meaning based on the semantic relationship between visual and textual information. This model focused on scene-specific content for the extraction of high-level semantic information. The novelty of the model [129] lies in the fact that this method introduced multiple visual regions of the image at multiple scales. Extensive experiments were carried out on three benchmark datasets: MSCOCO, Flickr8K, and Flickr30K that justified the superiority of this technique. [130] presented a review-based attention mechanism technique. The work performed multiple review steps with attention to various CNN hidden states to define the output vector that provided global facts of the image. For example, a reviewer module can first review: What sort of objects is present in the image? Then it can review the location or position of objects and subsequent review can extract all the important information of an image. The information obtained can further be passed to a decoder for caption generation. [127] presented a semantic attention-based technique that combined both top-down and bottom-up approaches to selectively extract semantic features followed by conversion into captions. [20] proposed an area-based attention mechanism for caption generation. This method was directly associated with caption words and image regions and predicted the next word as well as the corresponding image region in each time step. The work [20] when combined with the spatial transformer network produced high-quality image captions for the MSCOCO dataset.

Most attention mechanism-based methods forced visual attention to be active for each generated word. There are some words in the generated captions for which visual attention is not required like *"a, the, of"* do not require any visual attention as this could affect the generation process and can also degrade the overall efficiency of the process. [31] talked about an adaptive attention model with a visual sentinel. It employed a spatial attention mechanism to extract spatial features followed by adaptive attention for visual sentinel. In the work [125] a combination of DenseNet and adaptive attention with visual sentinel was presented. In this model, DenseNet was used to extract the global features from an image while at the same time sentinel gate was set by an adaptive attention mechanism that decided whether the image feature information should be used for word generation or not. LSTM network helped in the decoding phase for the generation of captions. [131] presented an idea of neural image captioning which evaluated and corrected the attention map at time steps. This method made a consistent map between image regions and the words generated which can be made possible with the introduction of a quantitative evaluation metric. Experiments were carried out on Flickr30K and MSCOCO datasets showed prominent improvements in both attention correctness and quality of captions generated. CNN dubbed SCN-CNN [124] considered spatial and channel-wise attention for computation of attention map. This method modulates the sentence generation context in multi-layer feature maps and visual attention. SCN-CNN architecture is evaluated on three benchmark datasets Flickr8K, Flickr30K, and MSCOCO which significantly outperformed their counterparts.

Bottom-up saliency-based attention methods [123] [132] are beneficial in reducing the gap between human-generated and machine-generated descriptions. [132]

worked on a bottom-up saliency-based attention mechanism. This method proved that the better a captioning model performs. the better an attention agreement it has with human descriptions. Also, this method provided better results on unseen data. [123] proposed a method that used both top-down and bottom-up approaches which can attend both object-level regions and salient image regions. It provided better results than [132] as the bottom-up mechanism in [123] used faster R-CNN which provides better results on the MSCOCO dataset. Context Sequence Memory Network (CSMN) [133] is different from previous discussed techniques [123] [132]. It generated captions for images by extracting context features with the use of "hashtag prediction" and "post generation".

Attention on Attention (AoA) [134] framework extended the conventional attention mechanism to determine relevance between attention results and queries. AoA is applied to both encoder and decoder of the image captioning model called AoANet that provided a new state-of-the-art performance on the MSCOCO dataset with a CIDEr-D score of 129.8. [135] utilized the concept of dual attention. It combined visual and textual attention for image caption generation. [126] explored visual relationships between regions in an implicit way that helped provide alignment between caption words and visual regions. Another work [136] defined the task adaptive attention concept for image captioning. It generated non-visual words by introducing diversity regularization and enhanced the expression ability of the proposed module. Improvement in performance for the MSCOCO dataset was observed by plugging the task-adaptive module into a vanilla transformer-based image captioning model.

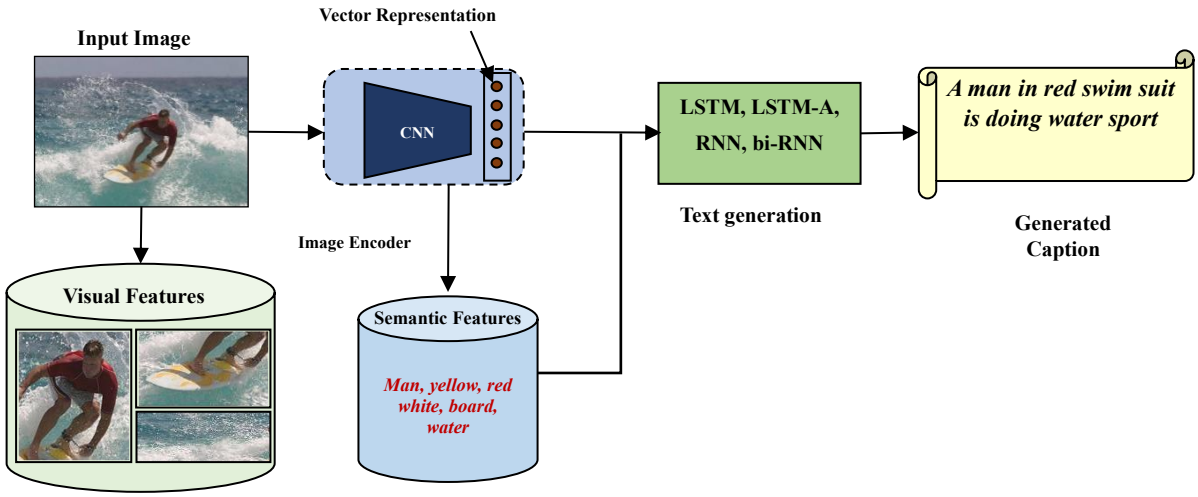


Fig 2.9: Basic Block representation of Semantic-Concept based Image Captioning

2.1.2.2.2.3 Encoder-Decoder Architecture with Semantic-Concept-Based

Semantic-concept-based image captioning is the ability to provide a detailed and coherent description of semantically important objects. The basic structure of semantic concept-based captioning of images is presented in Fig 2.9. Generally, CNN based encoder is used to extract features and semantic concepts. The extracted features are fed into a language generation model and the semantic concepts are fed to different hidden states of the language model which further produces a description of images with semantic concepts.

Karapathy et al. [19] combined CNN and bi-directional RNN to generate captions. Automatic description of an image with natural language is a challenging task in the field of computer vision and natural language processing. Attributes of an image are considered rich semantic cues. [22] proposed LSTM with attributes (LSTM-A) that integrated attributes into CNN and RNN image captioning framework by training them in an end-to-end manner. The architecture was tested on the MSCOCO dataset and thus obtained METEOR and CIDEr-D scores of 25.2% and 98.6% respectively. [127] provided detailed and coherent object descriptions by using top-

down and bottom-up approaches for generating captions. [128] further worked on fixed and predefined spatial location. This method can work on any resolution and on any location of an image with the use of a feedback process that accelerates to generate better captions for images. The earlier discussed methods [22] [127] do not include high-level semantic concepts. [137] discussed a high-level semantic-based captioning model. It extracted attributes using a CNN-based classifier and the extracted attributes were used as high-level semantic objects to generate semantically rich captions. An analysis is carried out on MSCOCO and large-scale Stock3M datasets that provided consistent improvements, especially on the SPICE metric. [138] presented a novel scene-graph-based semantic representation of an image. It built a vocabulary of semantic concepts and the CNN-RNN-SVM framework was used to generate the scene-graph-based sequence. Another work [139] defined SG2Caps which utilizes the scene-graph labels for competitive image captioning performance with the basic idea to reduce the semantic gap between the graphs obtained from an input image and its caption. The framework proposed in [139] outperformed [138] by a large amount and indicates scenes as promising representations for captioning of images. High-level semantic information provided abstractedness and generality of an image which is beneficial to improve the performance. [140] generated logical and rich descriptions of images by fusing the image features and high-level semantics followed by a language generation model. [141] presented a novel architecture for caption generation to better explore semantics available in captions. The model constructed caption-guided visual relationship graphs and it further incorporated a visual relationship to predict the word and object/predicate tag sequences. The Element Embedding LSTM (EE-LSTM) [142] generated rich descriptions of semantic features.

2.1.2.2.3 Compositional-Architecture Based Image Captioning Methods

Compositional architecture is an alternative means to build an image captioning model, that connects independent and loosely coupled components through a pipeline. It involves the following steps: (i) Visual features extraction using a CNN (ii) Visual concepts (i.e, attributes) encoding from visual features. (iii) Multiple captions generation from visual concepts (iv) re-ranking of generated captions using a deep multimodal similarity model to select high-quality image captions. Fig. 2.10 illustrates an example of a compositional architecture-based method that provides data from each component until a final result is obtained. [143] used visual detectors, a language model, and a multimodal similarity model to generate captions including different parts of speech like nouns, verbs, and adjectives. The model provides better results on the MSCOCO dataset and produced a BLUE-4 score of 29.1%. [144] attempted to generate captions for open-domain images (Instagram images). This model generated captions for landmarks and celebrities by detecting a diverse set of visual concepts. Description images with adaptive adjunct words generate more informative sentences. In this direction, [145] proposed a compositional network-based method to generate captions in the form of structured words <object, attribute, activity, scene>. It used multi-task and multi-layer optimization methods to generate semantically meaningful sentences with structural words [146] combined the advantages of RNN and LSTM and defined parallel fusion RNN-LSTM architecture which performed better than the other state-of-the-art [143] [144]. [147] defined the Text2Scene concept which generated descriptions for a compositional scene in form of natural language. This model is not based on GAN's and provides better and superior results when compared with GAN-based methods for the generation of captions for images. It can

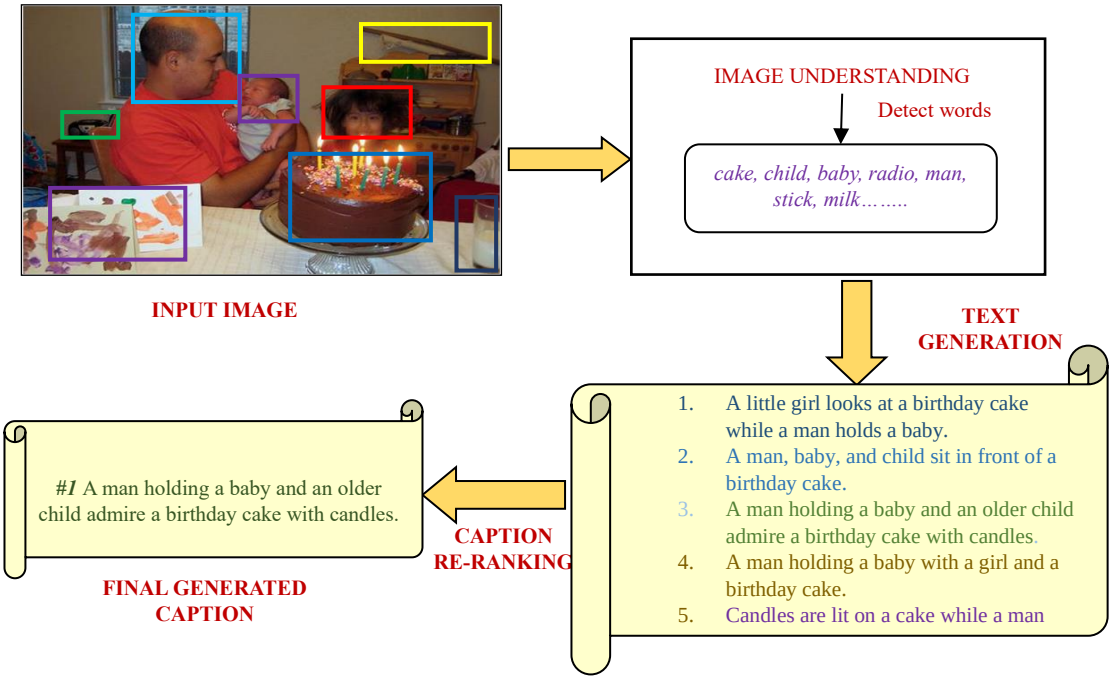


Fig. 2.10: An illustrative example of compositional architecture-based image captioning

handle different types of images i.e. cartoon-like scenes, object-layouts corresponding to real images, and synthetic images for caption generation. Compositional-architecture-based models [148] [149] are studied to measure how well a model composes unseen combinations of concepts while describing images. [148] combined caption generation and image-sentence ranking in this direction. The model [148] used a decoding mechanism that re-ranked the captions according to their similarity to the image. The model performed better when compared with other state-of-the-art methods.

Image captioning based on the compositional-architecture model has gained popularity because of its fluency which is an important factor for evaluation. [149] presented a hierarchical framework that generated accurate and detailed captions of images by exploring both compositionality and sequentially of natural language to

produce detail-rich sentences with specific descriptions of objects such as color, count, etc. Captioning of images focuses only on generalizing to images from the same distribution as the training set and not on generalizing to the different distribution of images. [150] investigated different methods to improve the compositional generalization for the syntactic structure of a caption and provide performance improvements.

Table 2.2: Overview of Different Deep-Learning-based Captioning Methods

Ref.	CM	#Citations	Method	Dataset	Pros	Cons
Kiros et al. [93]	Multi-modal Learning	672	Multimodal Learning	IAPR TC-12	First step towards multimodal learning and provides an improvement in BLEU scores.	Problem of text retrieval with extraneous descriptions that do not exist in the image
Kiros et al. [94]		1162	CNN-LSTM	Flickr8K, Flickr30K	Provides explicit embedding between images and sentences.	Complexity increases because of SC-NLM.
Karpathy et al. [95]		802	R-CNN	Pascal 1K Flickr8K Flickr30K	Improves the performance of the image sentence retrieval task.	Does not incorporate spatial reasoning and is not a better sentence fragment representation.
[96]		1131	m-RNN	Flickr8K Flickr30K IAPR TC-12 MSCOCO	Model incorporates more complex image representations with more sophisticated language models.	m-RNN with AlexNet requires modifications.
[97]		522	LSTM + Bi-RNN	Pascal 1K Flickr8K Flickr30K MSCOCO	This model is capable of learning long-term interactions.	Small datasets lead to overfitting
[30]		14	CNN + RNN	Flickr30K MSCOCO	More consistent with expressing the process of humans with the generation of a good description of images.	Complex analysis.
[98]		55	CNN-RNN	MSCOCO	The effectiveness of the model is measured quantitatively and qualitatively which provides the state-of-art result	-
[92]		18	GRU, LSTM	Flickr8K, Flickr30K, MSCOCO	The model uses image attributes information to enhance the image	Reduces captioning in real scenes, and cannot cover the rich underlying semantics

					representation. Can be used for dense captioning	
[99]		11	Multimodal LSTM	FlickrNYC MSCOCO	sg-LSTM model generates more accurate descriptions which are more meaningful than those provided by Flickr users	For certain case, the image description generated is not accurate.
[100]		2	Multimodal + Visual Attention	BBC News Dataset DailyMail Test	Visual attention and multimodal coverage mechanisms improve the model.	-
[107]	Encoder-Decoder w/o Attention	528	CNN-LSTM	MSCOCO	The model task allows for easy objective evaluation	A particular kind of object is too small to detect. Lacks enough training data.
[103]		374	LSTM + g-LSTM	Flickr30K MSCOCO Flickr8K	-	-
[18]		5094	LSTM (CNN-RNN)	Flickr30K MSCOCO Pascal	Robust model	Improvement in descriptions can be obtained with the use of unsupervised data
[112]		558	CNN + DGDN	Flickr8K, Flickr30K MSCOCO	Model is learned using variational auto-encoder with semi-supervised learning	Complex model due to DGDN
[111]		403	CNN-RNN	Flickr8K, Flickr30K MSCOCO	Enhancements in evaluation metrics can be observed.	There is s big gap between this model and human performance, low accuracy
[110]		207	LSTM + bi-LSTM	Flickr8K, Flickr30K MSCOCO	The effectiveness, generality, and robustness of proposed models were evaluated on numerous datasets	A better result can be obtained with multimodal and attention mechanism
[106]		5505	CNN-LSTM + LRCN	Flickr30K MSCOCO	Learning sequential dynamics with a deep sequence model shows improvement when compared with other methods.	Complex Model
[115]		269	CNN-RNN	MSCOCO	The proposed framework is modular w.r.t. the network design	This model fails to understand some important visual contents that only take small portions of the images
[116]		383	CGAN	Flickr30K MSCOCO	This framework provides an evaluator that is more consistent with human evaluation.	Major errors are the inclusion of incorrect details. e.g. colors (red/yellow hat), and counts (three/four people)

[118]		1161	SCST	MSCOCO	Provides an improvement in CIDEr score	-
[114]		422	GCN-LSTM	MSCOCO	Explores visual relationships which enhance the image captioning by increasing the CIDEr-D score.	The complexity of the model increases for the determination of spatial object relationship
[117]		43	CNN + RNN with GAN	MSCOCO	Provides optimization in evaluation metrics like CIDEr, BLEU, SPICE.	Generation of duplicate words or phrases is an issue with this model
[122]		-	Egocentric Image Captioning	EgoShots	Provides minute details from the images	-
[121]		-	CNN + RNN	MSCOCO	Outperforms other methods of captioning from English to Hindi.	Errors are there in the recognition of objects in images.
[120]		-	CNN + Attention based GRU	MSCOCO	Generates better descriptive text for images	-
[119]		113	Object Relation Transformer	MSCOCO	Modification in conventional transformer technique which can encode 2D position and size of objects.	62 errors were observed which are grouped in 4 categories objects, relations, attributes, and syntax
[127]	Encoder-Decoder-w-Attention Mechanism	1342	CNN-RNN (Semantic Attention)	MSCOCO Flickr30K	Combination of top-down and bottom-up strategies extracts rich information from images	An incorrect visual attribute may disrupt the model to attend incorrect concepts.
[133]		126	CMSN	Instagram Dataset	First personalized image captioning approach with hashtag prediction and post generation	Absolute metric values for the Instagram caption is low.
[132]		60	Deep CNN + LSTM	MSCOCO	-	-
[124]		1043	SCA-CNN	MSCOCO Flickr8K Flickr30K	Provides improvements in description of images	Improvements in results can be seen with temporal attention mechanism
[131]		191	CNN-LSTM	Flickr30K MSCOCO	Attention maps provide a positive correlation between attention correctness and captioning quality.	The quantitative results show that there is room for improvement to improve the captioning performance.
[31]		998	LSTM + Spatial Attention	MSCOCO Flickr30K	This model provides a fallback option for the decoder which make this model be used in many other applications excluding image captioning	Models give poor results for smaller objects like “surf-board”, “clock” etc.
[20]		157	CNN-RNN	MSCOCO	This model is the first step towards weakly supervised learning	Background elements are missing in some captions

[123]		2370	R-CNN Soft Attention	MSCOCO	This method enables attention to be calculated more naturally at the level of objects and other salient regions,	Feature binding problem arises which can be further resolved using attention
[134]		250	AoANet	MSCOCO	Experiments are conducted on MSCOCO which demonstrates this model is superior and effective when compared with human evaluation	Complexity increases with two attentions on attention.
[135]		15	CNN-RNN + FCN	AIC-ICC	The model is effective and feasible in image caption generation and the model is merged with FCN.	-
[125]		6	LSTM Adaptive Attention	Flickr30K MSCOCO	Significant improvement can be observed in BLEU and METEOR scores which improves the quality of image captioning	-
[136]		13	Adaptive Attention + Vanilla Transformer	MSCOCO	This model is useful for the generation of non-visual words.	-
[126]		4	Parallel Attention Mechanism	MSCOCO	The model can generate high-quality captions by capturing related visual relationships for generating accurate interaction descriptions	This model can cause an inaccurate description of image captioning.
[19]	**Encoder-Decoder Semantic Concept**	4764	CNN + bi-RNN	Flickr8K, Flickr30K MSCOCO	Provides image sentence rankings that provide rich descriptions of images.	Generates descriptions at a fixed resolution and additive bias interactions in RNN are less expressive.
[22]		506	LSTM-A	MSCOCO	Provides improvements in high-level attribute representation of images	-
[128]		7942	CNN-LSTM	Flickr8K Flickr30K MSCOCO	Provides state of art results for BLEU and METEOR for all three datasets as this model can attend non-objects salient features.	More extensive visualization is required.
[137]		255	CNN-RNN	Flickr8K Flickr30K MSCOCO	This method provides more accuracy and outperforms other state-of-art methods	Knowledge-based queries cannot be handled by this model.
[138]		13	CNN-RNN-SVM	MSCOCO	Experimental results on the provide superior and competitive results as the scene-graph improves the performance of the model.	Difficult to train images with the scene graph

[142]		11	LSTM and EE-LSTM	MSCOCO Flickr8K Flickr30K	EE-LSTM language model generates sentences with more details and outperforms LSTM with a significant margin.	There are some negative examples of the proposed method as some images have the only object due to which the advantage of EE-LSTM cannot be fully exploited.
[140]		-	CNN +LSTM + Attention	MSCOCO Flickr30K	The model generates a more comprehensive and smooth natural language description.	The model has a strong dependence on the accuracy of high-level semantic acquisition.
[141]		12	CGVRG	MSCOCO Visual Genome	Caption-guided visual-representation graphs provide enhancement in text and visual features	A complex model that can provide better results if applied to several other languages, visual modeling tasks
[139]		1	SG2Caps (GCN-LSTM)	MSCOCO Visual Genome V-COCO	Generates high-quality captions without using visual features.	More research is needed to reach human-level accuracy and diversity.
[143]	Compositional Architecture Based	1280	CNN-LSTM	MSCOCO	Can extract nouns, verbs, and adjectives from all regions of image, captions generated are better than human-generated captions 34% of the time.	BLEU and METEOR metric is very low.
[144]		125	ConvNet+ DMSM	MSCOCO, Adobe-MITFiveK Instagram images	Detects a broad range of visual concepts and generates rich captions.	Instagram images are filtered images or handcrafted abstract pictures which are difficult to process.
[145]		21	LSTM	UIUC Pascal Dataset Flickr8K	Structural words are generated which provides a semantically meaningful description of images.	-
[146]		39	RNN-LSTM	Flickr8K	RNN-LSTM provides better results than dominated architectures with improvement in efficiency	Parallel threads lead to many complexities in the model
[147]		43	CNN-RNN	MSCOCO Abstract Scenes	Model capture finer semantic concepts from visually descriptive text and generate complex scenes	Inception score does not evaluate correspondence between text and images
[148]		14	LSTM Embedding + Attention	MSCOCO	Produce captions with include combinations of unseen objects. Improvements in generalization performance	Model is better at generalizing to transitive verbs than intransitive verbs
[149]		3	LSTM + Attention	MSCOCO	The framework is easily expandable to include	-

					additional functional modules of more sophisticated designs	
[150]		1	RNN + Transformer	MSCOCO	Shows consistent improvements especially for inanimate color-noun combinations.	Model complex multi-task model

2.2 Image Captioning Performance Evaluation Parameters or Metrics

The performance of architecture greatly depends on the selection of the right evaluation metrics. In this section, the commonly used evaluation metrics are discussed in detail.

2.2.1 BLEU metric

BLEU stands for Bilingual Evaluation Understudy. It is the most popular metric for evaluation that evaluates the performance of machine translation systems. This metric was proposed by IBM [151] The central idea behind BLEU is "the better the BLEU score is if the closer is the machine translation to the professional human translation, yields a higher translation quality". It is used for the comparison and counting of the number of co-occurrences and depends on $n-gram$ precision that computes per-corpus $n-gram$co-occurrence, $n \in [1,4]$. The BLEU score is calculated by:

$$BLEU = BP.\exp\left(\sum_{n=1}^{N} w_n \log p_n\right); BLEU \in [0,1] \quad (2.5)$$

where N is usually set to 4 w_n is weight and is set to $1/N$ and BP is brevity penalty and p_n is modified precision given by the eqns. (2.6) and (2.7):

$$BP = \begin{cases} 1; & l_c > l_s \\ e^{(1-l_c/l_s)}; & l_c \leq l_s \end{cases} \quad (2.6)$$

where $l_c and l_s$ represents the length of the candidate sentence (c_i) and reference sentence(s_{ij}).

$$p_n = \frac{\sum_{C\in(Candidates)} \sum_{n-gram\in C} Count_{clip}(n-gram)}{\sum_{C\in Candidates} \sum_{n-gram\in C} Count(n-gram)} \quad (2.7a)$$

$$Count_{clip}(n-gram) = \min\{Count(n-gram), MaxRefcount(n-gram)\} \quad (2.7b)$$

The main advantage of the BLEU metric is that it has been adopted across the translation industry as a measure of MT quality that is designed for different language groups. Also, it considers n-gram instead of words. However, the BLEU metric considers each matched n-gram equally, which is one of its major drawbacks. Also, BLEU scores do not handle morphological-rich languages well and do not map well to human judgments.

2.2.2 METEOR Metric

METEOR or Metric for Evaluation of Translation with Explicit Ordering is the evaluation parameter that evaluates the machine translation output and is a better correlation with human judgment. METEOR [26] was proposed in 2005 after finding the significance of the recall rate. This evaluation parameter somehow addresses some of the flaws inherited by BLEU and is a measure that is based on a single-precision weighted average. Meteor metric calculates the accuracy, recall rate, and weighted F-score or F_{mean} for all cases of matching word, stem and synonym based on the matching unigrams and a penalty function for incorrect word order. Let us consider m is the number of mapped between two texts than precision (P) and recall (R) can be given as $m/c \ and \ {}^{m}/_{r}$ where c and r are the candidates and the reference lengths. Therefore,

$$F_{mean} = \frac{PR}{\alpha P+(1-\alpha)R} \quad (2.8)$$

$$P = \frac{|m|}{\sum_k h_k(c_i)} \quad (2.9)$$

$$R = \frac{|m|}{\sum_k h_k(s_{ij})} \tag{2.10}$$

To account for word order in candidate a penalty function is given by the relation:

$$P_{function} = \gamma(\frac{c}{m})^{\emptyset} \tag{2.11}$$

Where, c is the number of matching chunks and m is the total number of matches and also $\gamma, \alpha \; and \; \emptyset$ are the evaluation parameters whose default values are 0.5, 3 and 3. $METEOR$ evaluation considers recall and accuracy based on entire corpus and also includes some features which are not included in other metrics like synonym matching.

$$METEOR = (1 - P_{function})\, F_{mean} \tag{2.12}$$

2.2.3 ROUGE Metric

$ROUGE$ stands for Recall Oriented Understudy for Gisting Evaluation, $ROUGE$ is an automatic summarization method that came into existence in 2004 and was proposed by Chin-Yew Lin [152] This is an n-gram recall-rate method that evaluates abstracts based on the co-occurrences information of n-grams. The basic idea for $ROUGE$ includes n-grams, word-pairs and word-sequences which are counted to evaluate the quality of abstracts. With the use of summaries generated by the experts', one can measure the robustness and stability of the system. This metric consists of $ROUGE - NN \in [\,1,4], ROUGE - L, ROUGE - W$ and skip-bigram co-occurrence statistics $(ROUGE - S)$ and $ROUGE - SU$ which is an extension of $ROUGE - S$.

a) $ROUGE - N$ is a method that is mainly used for short-summary assessment or single-document and is based on n-gram co-occurrence statistics. For any n, the total number of n-grams are counted across all reference summaries and are compared with the candidate summary. This method has many advantages as it is concise and spontaneous. However, for a large value of N the degree of discrimination is not high

but is very low. ROUGE-L is based on LCS (longest common subsequence). Let $A\ and\ B$ be the two given sentences, if C is a subsequence of $A\ and\ B$, then sequence C is a common subsequence of $A\ and\ B$. LCS of $A\ and\ B$ is the subsequence that maximizes the length of C. Recall, precision and $F-score$ is evaluated in terms of LCS as :

$$R = \frac{LCS(A,B)}{m}\ , P = \frac{LCS(A,B)}{n}\ and\ F = \frac{(1+\mu^2)RP}{R+\mu^2 P} \tag{2.13}$$

Where m is the length of reference summary $A\ and\ B$ be the candidate summary with length n. F measures the similarity between$A\ and\ B$.

This method requires only a simple matching in accordance with the occurrence of words.

(b) $ROUGE-W$ is weighted LCS which introduces a weighted coefficient W which represents the length of the longest continuously matching common subsequence.

(c) $ROUGE-W$ is more distinguishable than LCS method as the sentences with more consecutive matches are weighted in comparison to those with fewer matches.

(d) $ROUGE-S$ is a new concept introduced that is based on the concept of skip-bigram. Skip-bigram is any pair of words with random gaps in sentence orders. The recall and precision, in this case, are evaluated as the ratio of the total number of possible bigrams $C(n,2)$, where C is known as the combination function. The main disadvantage of this method is that under unlimited jumping distance many meaningless words occur which can be reduced by setting a particular limit for the maximum jump distance.

(e) $ROUGE-SU$ is skip-bigram plus unigram-based co-occurrence statistics. If a sentence does not contain any bigram overlap it will not provide weight to any

sentence. Thus ROUGE-SU is considered as the place of $ROUGE-S$ which is an extension to $ROUGE-S$. $ROUGE$ Metric is appropriate for single-document evaluation and provides good performance in short summaries but has a problem in the evaluation of multi-document text summaries.

Though ROUGE is a significant evaluation metric it does not carter for words that have the same meaning as it measures syntactical matches rather than semantics.

2.2.4 CIDEr Metric

Consensus-based Image Description Evaluation [25] is the metric that measures the similarity of the generated captions against a set of sentences written by humans which means they provide a strong correlation with the human evaluation. Evaluation is carried out in terms of saliency, grammar, and accuracy. This parameter is based on a protocol that is consensus-based and measures the similarity between the sentences to be evaluated and a set of consensus image descriptions. CIDEr considers each sentence as a "document" made up of a set of n-grams. This metric encodes the frequencies of candidate sentence n-gram that are present in reference sentences. TF-IDF (Term frequency-inverse document frequency) is used to calculate the weight for every n-gram which further evaluates the similarity between each reference caption and the caption generated by the model by calculating the average cosine distance of the TF-IDF vectors. Mathematically, CIDEr is expressed as:

$$g_k(s_{ij}) = \frac{h_k(s_{ij})}{\sum_{w_l \in \xi} h_l(s_{ij})} \log\left(\frac{|I|}{\sum_{I_p \in I} \min\left(1, \sum_q h_k(s_{pq})\right)}\right) \tag{2.14}$$

$$CIDEr_n(c_i, s_{ij}) = \frac{1}{m}\sum_j \frac{g^n(c_i).g^n(s_{ij})}{\|g^n(c_i)\|\|g^n(s_{ij})\|} \tag{2.15}$$

Where $g_k(s_{ij})$ represents TF-IDF weight for each n-gram. Also, $h_k(s_{ij})$ represent the number of occurrences of an n-grams$s_{ij}$and $h_k(c_i)$ represents the number of

occurrences of an n-gram ω_k in c_i. $S_i = \{s_{i1}s_{i2}s_{i3} \dots \dots s_{im}\}$ indicates the set of reference captions. ξ is a list of all n-grams and $|I|$ represents the number of images in the dataset. Equation (13) represents the CIDEr score for n-grams of length n.

When using multiple lengths of n-grams to catch grammatical features and richer semantic information, the CIDEr score can be calculated as:

$$CIDEr(c_i, s_{ij}) = \sum_{n=1}^{N} \omega_n CIDEr_n(c_i, s_{ij}) \quad (2.16)$$

In comparison to BLEU, CIDEr does not treat each matching word equally but has a focused treatment which helps in the improvement of the accuracy of existing measures. The most popular version of CIDEr in image and video description evaluation is CIDEr-D, which incorporates a few modifications in the originally proposed CIDEr to prevent higher scores for the captions that badly fail in human judgments.

2.2.5 SPICE Metric

Semantic Propositional Image Captioning Evaluation (SPICE) [153] was introduced in 2016 and is the latest evaluation metric for image and video descriptions which is capable of measuring similarity between the scene graph tuples generated from the descriptions by machine and the ground truth. The semantic scene graph encodes objects, their attributes, and relationships through a dependency parse tree. A scene graph tuple of a caption k contains attributes$A(k)$, relations $R(k)$ and object classes $0(k)$ are represented as:

$$G(k) =< A(k), R(k), O(k) > \quad (2.17)$$

Like METEOR, SPICE also uses WordNet to find and treat synonyms as positive matches... SPICE can also tell us about ‘which caption-generator best understands colors?’ and ‘can caption generators count? Moreover, SPICE captures human

judgments better when compared to other evaluation metrics. For instance, in the sentence ‘‘black cat swimming through river”, the failure case could be the word “swimming” being parsed as “object” and the word “cat” parsed as “attribute” resulting in a very bad score.

2.3 Research Gaps

On the basis of the outlines of literature review of the earlier state-of-the-arts for different image captioning tasks and methods research gaps are identified and a layout of solutions for the identified research gaps are listed, which are as follows:

1. Most of the earlier visual caption generation methods [1] [59] are based on traditional architectures which generate words in a sequential manner. This leads to syntactically correct, but semantically irrelevant language structures that further amount to a lack of diversity in generated captions.

2. As reported in literature, spatial attention-based captioning models fail to generate captions for small size objects in the frame. Small sized objects should be included in captions for better interpretability of the scenes.

3. The recent application of transformers to the computer-vision field has attracted extensive attention. Therefore, very limited work is available that leveraged the strength of the transformer model for visual data captioning.

4. Generation of captions with styles help reflect various human emotions (romance, humour, etc.) in the captions. Stylized captioning has not been much explored due to limited available data. Therefore, it is still an open issue.

5. Few works attempted to generate captions using change image captioning style. However, these works need more fine-grained feature difference representation and better object attribute understanding for better captions.

6. Real-time caption generation for images remains the most intimidating challenge to deal with. Therefore, unsupervised learning and reinforcement-based learning may prove to be more realistic ways of caption generation in real time. However limited work has been observed for unsupervised based and reinforcement-based caption generation for captioning.

7. Dense Visual Caption generation methods closely look at the minute details in the images by correlating text at multiple level of abstraction. The entire process takes longer time to generate well-defined description of images. Further, Dense captioning for short videos has been attempted however, it will be relevant to extend dense captioning for long videos by including multiple events and speech or audio modalities

2.4 Research Objectives

The challenges involved in the identification of captioning of images, motivate to develop efficient models and frameworks to fulfil the purpose of developing a robust captioning system. In order to achieve this, five research objectives are formulated here to handle the practical challenges involved mentioned above, which are as follows:

- ✓ Objective 1: To review different image captioning techniques.
- ✓ Objective 2: To propose an efficient image captioning model.

- ✓ Objective 3: To design a style-transfer based image captioning model for effective caption generation. Objective 4: To generate refined image captions using the integration of different image captioning techniques.
- ✓ Objective 5: To propose an intelligent dense image captioning model.

2.5 Research Gaps and Objectives Mapping

- ✓ Objective-1, comprehensive review is the systematic analysis of existing methods and frameworks for different image captioning tasks that identified different research gaps, limitations, and trends. This further help in the formation of other objectives.
- ✓ Objective 2 proposes transformer-based image captioning model that included small-sized objects in the generation of captions to better interpret the scenes.
- ✓ Further, a style-transfer-based image captioning framework (objective-3) is proposed. Due to limited available data, this framework generated stylized image captions by utilizing the knowledge of factual image captioning model. Therefore, the proposed framework tackles the problem of limited available stylized captioning data by generating coherent and diverse stylized descriptions of image.
- ✓ Objective-4, incorporates the captions generated from factual and stylized image captioning model with a text-summarization transformer thereby leveraging the strength of transformer model for visual data captioning to generate captions with more fine-grained feature difference representation and better object attribute understanding for better captions.

- ✓ Objective-5, utilizes transformer-based architectures that generates dense or paragraph-based descriptions of images that closely look at the minute details in the images by correlating text at multiple level of abstraction.

Chapter-3

Intelligent Factual Image Captioning Based Models

The objective of this chapter is to highlight different transformer-based image captioning models. The key components of this chapter include the study for different transformer-based architectures for single-sentence image captioning with extraction of higher order interactions between detected objects and the relationship among them. The proposed transformer-based deep-learning models are supported by experimental validation, results discussions and comparative analysis of results with the similar state-of-the-arts.

3.1 Lightweight Transformer with GRU Integrated Decoder for Image Captioning

Most traditional image captioning systems use an encoder-decoder structure, in which an input image is encoded into an intermediate representation of the information contained within the image, and then decoded into a descriptive text sequence, which is inspired by neural machine translation. This encoding can consist of a single feature vector output of a CNN [128], or multiple visual features obtained from different regions within the image. In the latter case, the regions can be consistently sampled [18], or directed by an object detector [123] which has been shown to yield enhanced performance. Transformer-based architectures characterize the state-of-the-art in sequence modelling tasks like machine translation and language comprehension. Their pertinency to multi-modal contexts like image captioning is also being tinkered

actively. With the aim of building a lightweight and production deployment friendly model, we present the Lightweight Transformer with a GRU integrated decoder for Image Captioning. Following this premise, we investigate the design of a self-attentive lightweight approach with more evolved decoder.

3.1.1 Proposed Methodology

Our architecture takes inspiration from the Transformer model [154] for machine translation and Object Relation Transformer for image feature extraction incorporating two key novelties: (i) image regions are encoded in a multi-level fashion, in which low-level and high-level relations are taken into account. When modeling these relationships, our model can learn and encode a priori knowledge by using Squeeze-and-Excitation Networks. (ii) The generation of the sentence, done with a multi-headed self-attention-based mechanism, uses a GRU to further enhance the language and word mapping with extracted image features. The proposed model first extract appearance features using multi-level Inception-V3 architecture, as described in sub-section 3.1.1.1. Thereafter, proposed Lightweight Transformer is exploited to generate the final captions. Sub-section 3.1.1.2 describes how we use the Transformer [154] architecture in general for image captioning. Sub-section 3.1.1.3 explains novel addition of reduced number of encoders and decoders for transformer model and GRU integrated decoder.

3.1.1.1 Object Detection

For detection of objects and feature extraction, InceptionV3 [155] model pre-trained on “ImageNet” dataset, is incorporated. The basic architecture of Inception-V3 module for object detection is shown in Fig. 3.1. Features are extracted from the images by using the standard pre-processing on images for InceptionV3 model. The

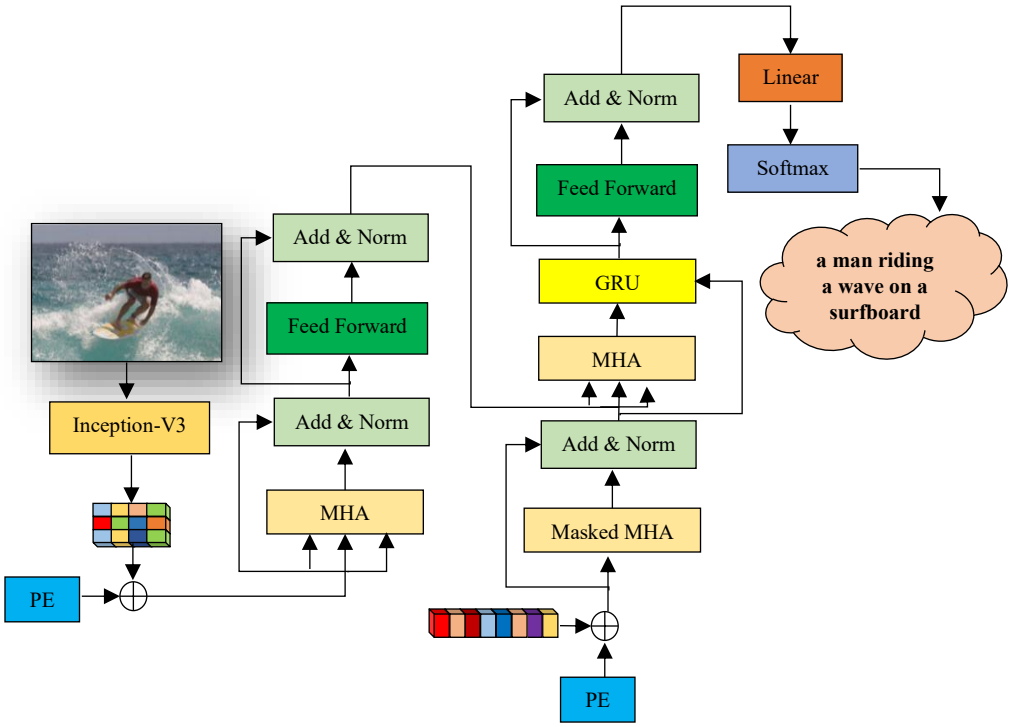


Fig. 3.1: Proposed Lightweight Transformer with GRU Integrated Decoder

final step is to scale all remaining bounding boxes to the same spatial size using a Region-of-Interest (RoI) pooling layer. For each i^{th} object bounding box, we also employ mean-pooling over the spatial dimension to create a 2048-dimensional feature vector called ROI_Fv. The above-mentioned steps are applied at two different layers of the InceptionV3 network to extract multiple levels of features. These feature vectors, i.e., $Multi\_level\_Fv$, as defined in eqn. 3.2 are then combined and reduced to a 2048-dimensional feature vector and then used as input to the Transformer model.

$$ROI_{Fv} = \left[ROI_{Fv}; ROI_{i,1\times 2048}\right],\ i = number\ of\ objects\ detected \quad (3.1)$$

$$Multi\_Level\_Fv = Concat(ROI_{Fv}, ROI_{Fv_{end}}) \quad (3.2)$$

3.1.1.2 Standard Transformer Model

The standard transformer model is divided into two main sections namely: (i) the encoder unit, and (ii) the decoder unit. Both, encoder and decoder consist of stack

of layers as depicted in Figure 3.1. The proposed Lightweight Transformer uses the image features as an input at the encoder and use these features at the decoder to generates the output in the form of natural language sentence. The feature vector from the Inception-V3 model is processed through the input embedding layer of the transformer which contains an FC layer followed by ReLU activation and dropout layer. It decreases the dimension of image feature from 2048 to 512 (d_{model}). The output feature vector from the embedding layer is further used as an input to the first encoder transformer layer. Also, x_n here is assumed to be the n-th token from a set of N tokens. Also, the input to the remaining encoder layer is usually taken from the tokens generated from the preceding layer. The entire encoder section consists of multi-head attention layer and a feed-forward network (FFN). There are 8 identical heads in the self-attention layer itself, where every attention head is calculated mathematically as:

$$Q = XW_Q, K = XW_K, V = XW_V \tag{3.3}$$

where X is the input vector and is represented by $x_1, x_2 \ldots x_N$ and is stacked into a matrix and $W_Q, W_K, and\ W_V$ are learned projection matrices. Further, attention weights for each appearance features are calculated by using eqn. (3.4)

$$\Omega_A = \frac{QK^T}{\sqrt{d_k}} \tag{3.4}$$

where A represents an $N\ x\ N$ attention weight matrix. Furthermore, the output of the head is represented by Eq. (3.5)

$$head(X) = self - att(Q, K, V) = softmax(\Omega_A V) \tag{3.5}$$

The eqns. (3.3) to (3.5) are intended for every head independently. Also, 8 heads are concatenated to produce a single output as represented by eqn. (3.6)

$$MulitHead(Q, K, V) = concat(head_1, \ldots\ldots, head_h)W_O \tag{3.6}$$

After this section FFN is utilized at every output of the attention layer and is represented mathematically as:

$$FFN(x) = \max(0, xW_1 + b_1)W_2 + b_2 \tag{3.7}$$

where $W_1, W_2, and\ b_1, b_2$ represent the weights and biases of two FC layers. Furthermore, Layer-Norm with skip connections is added to produce the final output of the encoder layer. The decoder employs the generated tokens from the final encoder layer as input to generate the captions.

3.1.1.3 Lightweight Transformer

The proposed model incorporates only a single encoder and a single decoder structure, thereby reducing the model complexity and the total number of learnable parameters by a factor of six. The encoder takes inspiration from a vision encoder, taking input as the extracted appearance features for every image, after certain pre-processing applied on them. Positional encoding is applied on top of the input embeddings (appearance features), and then passed through a multi-headed attention layer. The dumped features are then passed through a fully connected layer, which generates the output of the encoder block.

The decoder follows the standard structure of the standard transformer model, with one addition of a Gated Recurrent Unit (GRU) layer [156]. The input caption embeddings are first passed through positional encoding to capture positional

relationship between words and then passed through the GRU layer with the Glorot Uniform Initializer. The dumped results are then fed through a series of fully connected layers and finally a linear activation function followed by a softmax layer, generating the output probabilities.

3.1.2 Experimental Work and Results

Experiments are conducted on MSCOCO-2014 Captions dataset. This dataset contains 113K, 5K and 5K training validation, and test images with 5 human annotated captions for each image. Further, the experiment is conducted on a subset of the entire dataset consisting of combination of training and testing samples of 35K images.

3.1.2.1 Implementation Details

The proposed Lightweight Transformer was implemented on TensorFlow 2.8.0 using the MS-COCO Dataset 2014 for Image Captioning [157]. The experiments were run on Google Colab using NVIDIA Tesla T4 GPU. A sub-set of the dataset was taken to train the model consisting of 28,000 training examples and 7,000 testing examples, following an 80:20 split. Our best performing model was trained for 30 epochs using a sparse categorical cross-entropy loss with Adam optimizer, and a batch size of 64. The training took approximately 1 hour and 55 minutes with sparse categorical cross-entropy on single GPU.

The effectiveness of the proposed model is evaluated in terms of the following evaluation metrics: CIDEr [25], BLEU-n [151], METEOR [26] and ROUGE-L [152] metrics. Although it has been demonstrated experimentally that BLEU and ROUGE show very less correlation with human judgments as compared to other metrics, however still all the metrics are evaluated for all image captioning tasks.

3.1.2.2 Quantitative Results

The quantiative results comparision for proposed Lightweight Transformer with the other state-of-the-art is presented in Table 3.1. The proposed model performs better than SCST [118] and Up-Down [123], which utlises attentions over regions and over grid of features. Further, our model achieves state-of-the-art results when compared to RFNet [158] that merge different CNN features by utilizing recurrent network. and GCN-LSTM [114], which exploits pairwise relationships between image regions through graph CNN. Further, proposed Lightweight Transformer achieves good results when compared with SGAE [159], and AoANet [134]. The compariosn is also made with respect to [160] that exploits standard transformer model to generate captions. Finally, when compared with the M2 Transformer [70], our proposed model does not perform better as M2 transformer utlizes memory-augmented encoder and a meshed decoder to capture feature from all layers of the stacked encoders and decoders therby increasing the trainable parameteres and computational complexity.

Table 3.1: Comparison of the Quantitate Results Obtained for the Proposed Lightweight Transformer; (*B-1, B-2, B-3, B-4: BLEU-n; M: METEOR; R: ROUGE; C-CIDEr)*

Method	B-1	B-2	B-3	B-4	M	R	C
SCST [118]	78.1	61.9	47.0	35.2	27.0	56.3	114.7
Up-Down [123]	80.2	64.1	49.1	36.9	27.6	57.1	117.9
RFNet [158]	80.4	64.9	64.9	38.0	28.2	58.2	122.9
GCN-LSTM [114]	80.8	65.5	65.5	38.7	28.5	58.5	125.3
SGAE [159]	81.0	65.6	65.6	38.5	28.2	58.6	123.8
ETA [161]	81.2	65.5	65.5	38.9	28.6	58.6	122.1
AoANet [134]	81.0	65.8	65.8	39.4	29.1	58.9	126.9
GCN-LSTM+HP [114]	81.6	66.2	66.2	39.3	28.2	59.0	127.9
M2 Transformer [70]	81.6	66.4	66.4	39.7	29.4	59.2	129.3
Proposed Lightweight Transformer	**81.0**	**65.2**	**65.2**	**37.8**	**27.9**	**58.0**	**123.1**

3.1.2.3 Qualitative Results

The qualitative results obtained for the proposed Lightweight Transformer is depicted in Fig. 3.2. It represents captions generated by the proposed transformer with attention plots to support the generated captions. These plots demonstrate how the generated sentences are closely related to the objects, scenes, and their attributes. The notable achievement of the proposed transformer is that it provides a lighter model due to single encoder-decoder transformer layer with a faster inference time and lower training time, while maintaining an acceptable level of accuracy.

Fig. 3.2: A sample (or examples) for the proposed Lightweight Transformer Model

3.2 XGL-T Transformer for Intelligent Image Captioning

With the advancements in deep-learning, the transformer-based models [154] are being widely used for image captioning task. Yu et al. [162] proposed multimodal transformer that performed multimodal reasoning and led to more accurate caption generation. He et al. [163] proposed an image transformer that widened the original transformer layer's inner architecture to adapt to the structure of images. This model captured relative spatial relationship between image regions and provided a better computational complexity when compared with [154] [162]. Furthermore, [161] was designed to extract more discriminative features by considering novel objects. Pan et al. [164] proposed X-linear attention network for image captioning that leveraged 2^{nd} and higher order interactions by exploiting the low-rank bilinear pooling [165]. [166] learned the object-relationship dependencies for visual understanding and generation of captions. Furthermore, Yang et al. [167] introduced a novel architecture ReFormer that expressed pair-wise relationships between objects in an image. Most of the above-mentioned works provided only the first order interaction between image contents and sentences thereby limiting the capacity of multi-modal reasoning. Very few works facilitated higher order interactions. Whereas, in the proposed work as depicted in Fig. 3.3, the use of low-rank bilinear pooling with scaled version of GELU activation (i.e., x-GELU) function provides improvements in fine-grained visual recognition. This also strengthens the higher order scene understanding and enhanced semantic concepts of objects, attributes and relationships between encoder and decoder that provides discriminative cues for generation of sentences.

3.2.1 Proposed Methodology

The proposed work exploits higher-order interactions between visual content and natural sentence are learned by defining an efficient XGL-Transformer (XGL-T) image captioning learning model. XGL attention modules' stack in encoding phase is deployed to encode Region-Level Features (RLF) along with higher-order interactions. This provides a set of enhanced RLF and Image-Level Features (ILF). Further, in decoding

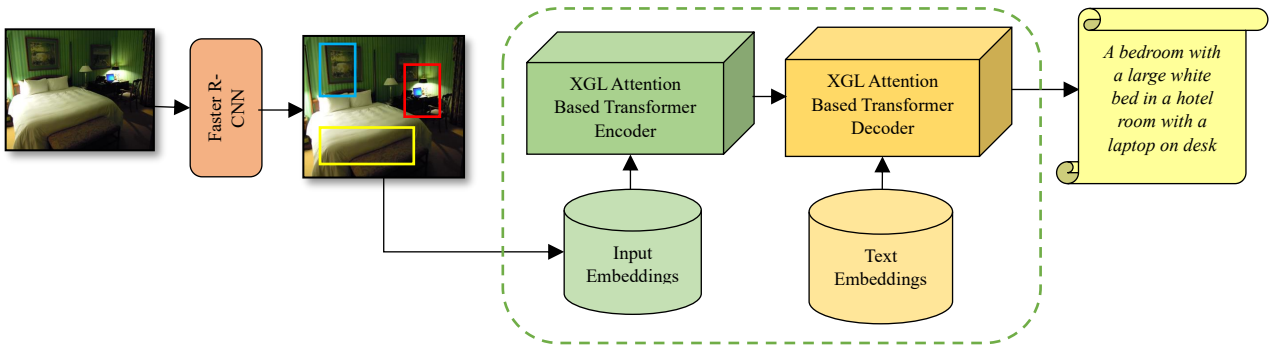


Fig. 3.3: Basic Block Diagram Representation of Proposed Methodology

phase, XGL attention module equipped with Bi-LSTM and GEGLU [168] based learning provides higher-order interaction reasoning, and improves the performance of sentence generation system. Experimental results indicate the promising evidence that the proposed model can attain a great improvement in the task of image captioning. The main contributions of the proposed work for the generation of factual captions can be summed as follows:

- ✓ A novel encoder decoder based XGL-Transformer model is proposed for efficient image caption generation.
- ✓ In the proposed model, XGL attention module with low-rank bilinear

pooling and skip-squeeze and excitation introduces attention to higher-order feature interactions, and helps generating effective image captions.

- ✓ A scaled version of GELU activation function, x-GELU activation, is defined. It normalizes the independent features, makes the gradients smaller, and provides a better solution to vanishing gradient problem.

3.2.1.1 $x-$ GELU Activation Function

Activations [169] like ReLU, SELU and ELU enable network, become faster and converge better but they fail to cover dropout regularization [170]. Further, GELU [171] activation provides both the advantages i.e., faster convergence and better dropout regularization. In this section, an extension to GELU activation function known as x-GELU is introduced that normalizes the independent features and solves the problem to vanishing gradients by making the gradients smaller. The x-GELU is mathematically defined as:

$$\varkappa(x) = x \times \{GELU(x)\} \tag{3.8}$$

$$\varkappa(x) = x \times \{x[P(X \leq x)]\} = x \times \{x\emptyset(x)\} \tag{3.9}$$

Where, $\emptyset(x)$ is the standard Gaussian cumulative distribution function. Further, if $X \sim \mathcal{N}(0,1)$, eqn. (3.9) can be approximated as:

$$\varkappa(x) = x \times \{0.5 \times x(1 + \tanh[\sqrt{2}/\pi\,(x + 0.044715x^3)]\} \tag{3.10}$$

3.2.1.2 XGL Attention Mechanism

Conventional Attention Module (CAM) [128], as shown in Fig. $3.4(a)$, learns to selectively attend to salient features for the generation of sentences. Though, the interaction between distinct modalities is nicely triggered by the conventional

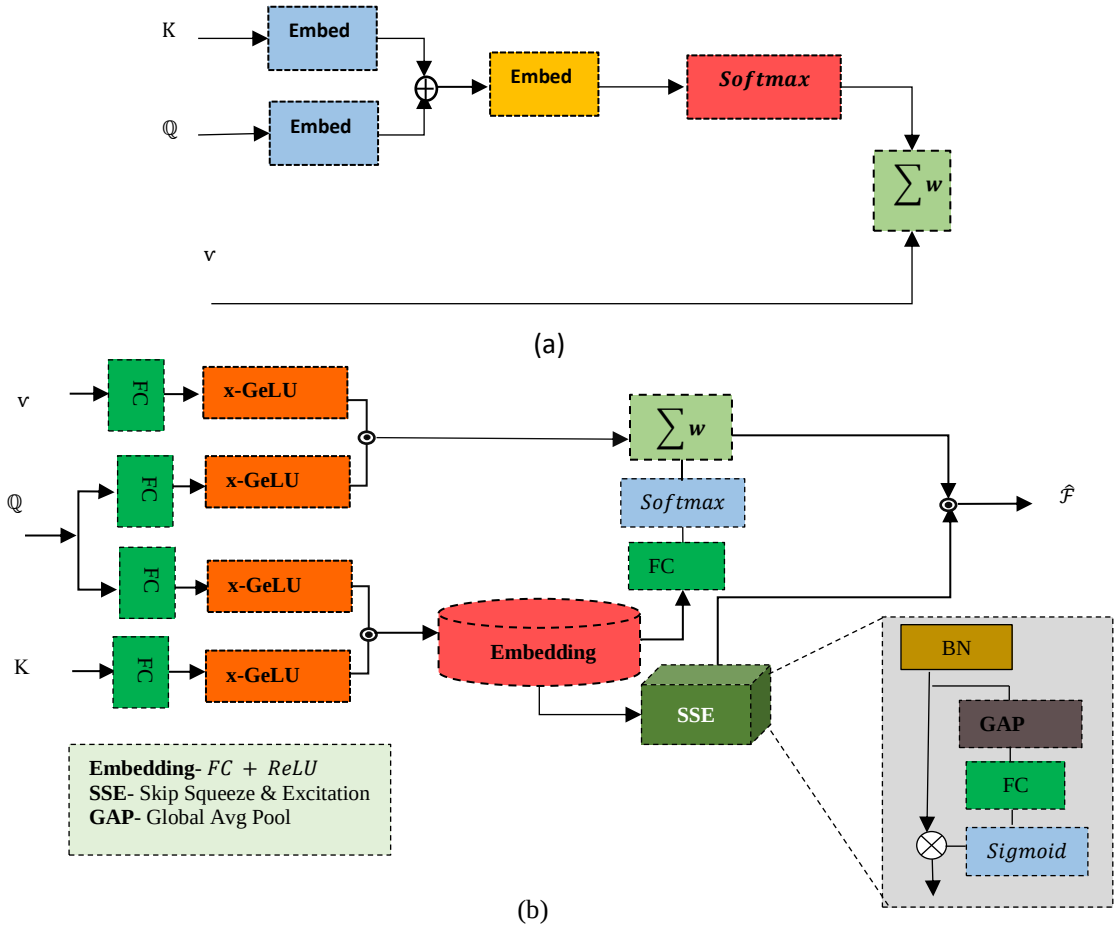


Fig. 3.4: (a) Conventional Attention Module (CAM), (b) proposed XGL Attention Module

attention module, but only the first order interaction of feature is exploited thus restricting it to the limited capacity of complicated multimodal reasoning for image captioning. Therefore, XGL attention module is introduced to strengthen the automatic learning of features, through exploitation of higher order interactions, and by increasing the representative capacity of the output attended features. Fig. 3.4(b) depicts an introduced attention module named as XGL attention module. This module aims to (i) enhance the visual information, and (ii) how consecutive

words are correlated to generate sentences. It captures the higher order feature interactions by exploiting low-level bilinear pooling [165].

Suppose, the query $\mathbb{Q} \in \mathbb{R}^{D_q}$, a set of keys K, and values ν where $\mathrm{K} = \{k_i\}_{i=1}^{N}$ and $\nu = \{v_i\}_{i=1}^{N}$. Also, $k_i \in \mathbb{R}^{D_k}$ and $v_i \in \mathbb{R}^{D_v}$. For each query $\mathbb{Q}$ and key $\{k_i\}$, XGL attention module first produces a joint bilinear query-key representation $\mathbb{P}_i^k \in \mathbb{R}^{D_{\mathbb{P}}}$ by performing low-rank bilinear pooling.

$$\mathbb{P}_i^k = \varkappa(\omega_k k_i) \odot \varkappa(\omega_q^k \, \mathbb{Q}) \tag{3.11}$$

where, $\omega_k \in \mathbb{R}^{D_{\mathbb{P}} \times D_k}$ and $\omega_{\mathbb{Q}}^k \in \mathbb{R}^{D_{\mathbb{P}} \times D_{\mathbb{Q}}}$ are embedding matrices, $\odot$ represents element wise multiplication, and $\varkappa$ represents x-GELU activation. Also, $\mathbb{P}_i^k$ represents higher order query-key interactions. Further, two types of attention distributions are received as spatial and channel-wise information from all bilinear query-key representations $\{\mathbb{P}_i^k\}_{i=1}^{N}$. For spatial attention distribution, each query-key representation is introduced into the embedding layer (fully connected layer with ReLU activation) with its corresponding attention weights.

$$\mathcal{S} = \sigma\left(\omega_{\mathbb{P}}^k \mathbb{P}_i^k\right) \tag{3.12}$$

where, $\omega_{\mathbb{P}}^k \in \mathbb{R}^{D_c \times D_{\mathbb{P}}}$ is the embedding matrix, σ represents ReLU activation function, and $\mathcal{S}$ is transformed query-key representation. The transformed query-key representation, $\bar{\mathcal{S}}$, is normalized using a fully connected layer followed by a softmax layer.

$$\bar{\mathcal{S}} = softmax(\omega_b \mathcal{S}) \tag{3.13}$$

where, $\omega_b \in \mathbb{R}^{1 \times D_c}$ is embedding matrix.

For channel-wise attention measurement, Skip-Squeeze and Excitation (SSE) [172] operation, which is based on Squeeze and Excitation (SE) [173] design, is performed over all query-key transformed representations. This block helps increase in the performance for the channel-wise attention measurement. SSE block aggregates all transformed bilinear query-key representation with batch normalized query-key representations followed by global average pooling, fully-connected layer and sigmoid layer. Further, the batch normalized query-key representation and the sigmoid layer output are multiplied. The SSE block execution can be mathematically represented as follows:

$$\mathbb{B} = batchnorm(\mathcal{S}) \tag{3.14}$$

$$\overline{\mathbb{B}} = \frac{1}{N}\sum_{i=1}^{N} \mathbb{B}_i \tag{3.15}$$

$$\widetilde{\mathbb{B}} = \omega_{\mathcal{B}}\overline{\mathbb{B}} \text{ and } \widehat{\mathbb{B}} = sigmoid(\widetilde{\mathbb{B}}) \tag{3.16}$$

$$\mathbb{B}_{SSE} = \mathbb{B} \otimes \widehat{\mathbb{B}} \tag{3.17}$$

The query-value representation is also obtained by low-rank bilinear pooling for each query $\mathbb{Q}$ and value $\{v_i\}$

$$\overline{\mathbb{P}}_i^k = \varkappa(\omega_v v_i) \odot \varkappa(\omega_q^k\, \mathbb{Q}) \tag{3.18}$$

where, $\omega_v \in \mathbb{R}^{D_{\mathbb{P}} \times D_v}$ represents the embedding matrix and $\overline{\mathbb{P}}_i^k \in \mathbb{R}^{D_{\mathbb{P}}}$ represents the transformed higher order key-value interactions. Weighted sum of the features, $\mathcal{F}$, from the spatial attention distribution and the transformed query-value representation is calculated as:

$$\mathcal{F} = \sum_{i=1}^{N} \overline{\mathcal{S}}_i \overline{\mathbb{P}}_i^k \tag{3.19}$$

Finally, the proposed XGL attention module provides the attended value features $\hat{\mathcal{F}}$ using the weighted features $\mathcal{F}$ and the channel-wise attention measurement of the transformed query-key representation:

$$\hat{\mathcal{F}} = \mathbb{F}_{\varkappa}(\mathrm{K}, \mathbb{Q}, \mathrm{v}) = \mathbb{B}_{SSE} \odot \mathcal{F} \tag{3.20}$$

This attention module produces more representative attended features, which can further be leveraged to encoder-decoder framework for image captioning tasks. This may be attributed to the fact that the higher-order interactions are exploited by using bilinear low-rank pooling with x-GELU activation.

3.2.1.3 XGL Transformer (XGL-T)

A unified attention module, as presented in section 3.2.1.2, is plugged into encoder-decoder framework to capture higher order interactions (intra-or inter-modal) for image caption generation. This subsection talks about a deep end-to-end encoder-decoder architecture as portrayed in Fig. 3.5 with the obvious aim to describe an image in the form of sentence. This architecture stacks XGL attention blocks for retrieval of deep image features and uses them for the generation of textual captions.

3.2.1.3.1 XGL-T Image Encoder

Image encoder module transforms a set of visual features from the input image into a series of encodings. Image encoder block is designed by employing the proposed XGL attention mechanism. This module provides strengthening to the encoded RLF and/or ILF through higher-order correlations. The image encoder contains stacks of $(1 + \ell)$ similar layers with ℓ =3. Further, each identical layer consists of two main components; XGL attention module and key-values updating

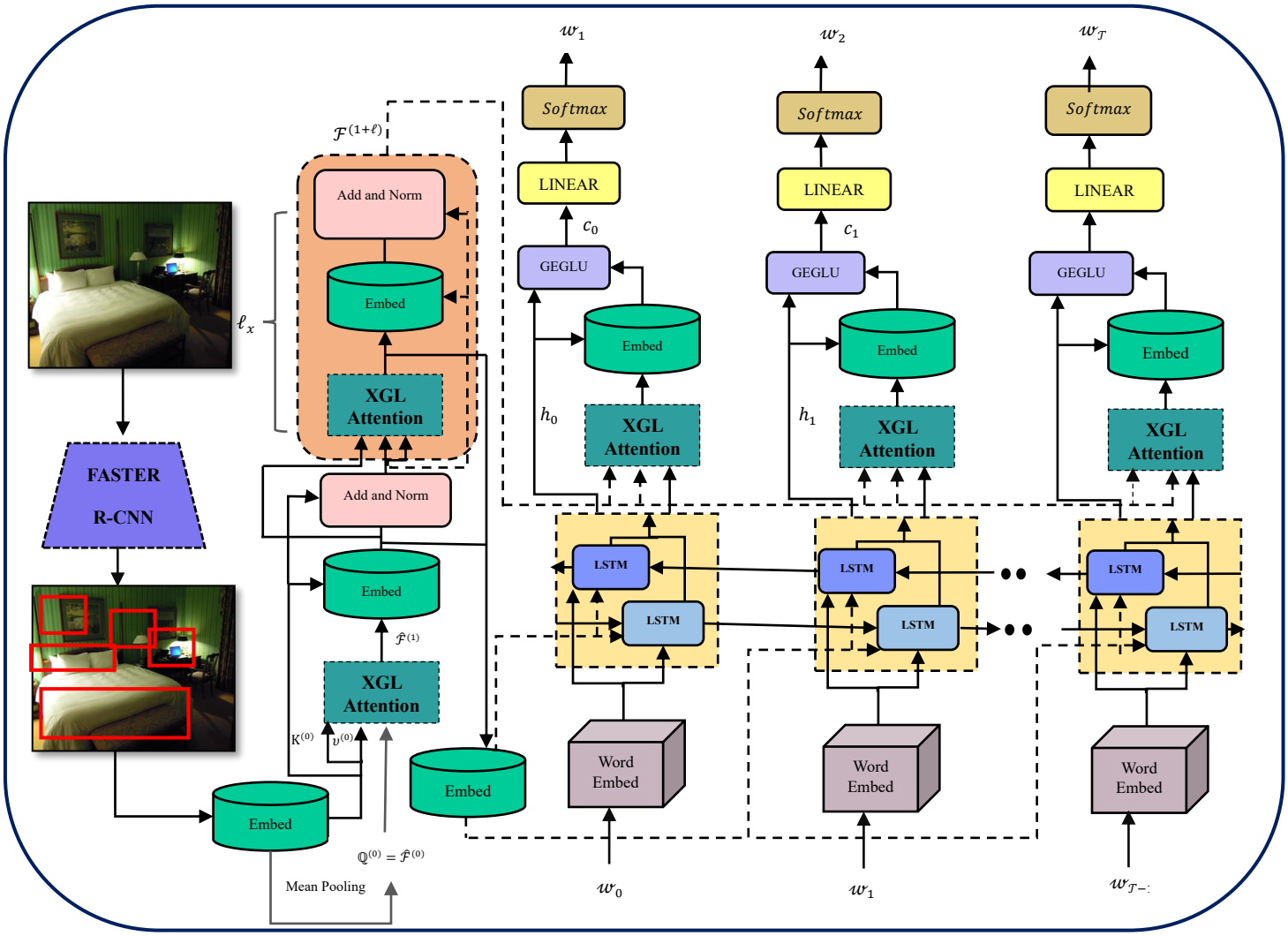


Fig. 3.5: Proposed XGL-Transformer (XGLT)

module. Initially, for the first XGL attention block, a mean pooled region features ($\hat{\mathcal{F}}^{(0)}$) is taken as the query $\mathbb{Q}$ and is coupled with initial keys ($\mathrm{K}^{(0)}$) and values ($\mathrm{v}^{(0)}$). The output to this XGL block is the attended image feature ($\hat{\mathcal{F}}^{(1)}$). This feature acts as an input query $\mathbb{Q}$ to the next x-GELU attention block. The keys ($\mathrm{K}^{(0)}$) and values ($\mathrm{v}^{(0)}$) are updated by feeding the $\hat{\mathcal{F}}^{(1)}$ to the embedding layer (fully connected layer with ReLU activation) and to Add-and-Norm layer. The updating process is repeated ℓ times through a subsequence of ℓ stacked identical

layers. For the ℓ^{th} order XGL attention block, the query $\mathbb{Q} = \hat{\mathcal{F}}^{(\ell-1)}$, key $\mathrm{K}^{\ell-1} = \{k_i^{\ell-1}\}_{i=1}^{N}$, and values $\mathrm{v}^{(\ell-1)} = \{v_i^{\ell-1}\}_{i=1}^{N}$:

$$\hat{\mathcal{F}}^{(\ell)} = \mathbb{F}_{\varkappa}(\mathrm{K}^{\ell-1}, \hat{\mathcal{F}}^{(\ell-1)}, \mathrm{v}^{(\ell-1)}) \tag{3.21}$$

where, $\hat{\mathcal{F}}^{(\ell)}$ is output attention feature.

Further, the keys and values are updated, conditioned on the output attention feature $\hat{\mathcal{F}}^{(\ell)}$ followed with regularization and layer normalization as in [154].

$$k_i^{\ell} = LayerNorm(\sigma\left(\omega_{\ell}^{k}\left[\hat{\mathcal{F}}^{(\ell)} k_i^{\ell-1}\right]\right) + k_i^{\ell-1}) \tag{3.22}$$

$$v_i^{\ell} = LayerNorm(\sigma\left(\omega_{\ell}^{v}\left[\hat{\mathcal{F}}^{(\ell)} v_i^{\ell-1}\right]\right) + v_i^{\ell-1}) \tag{3.23}$$

Eventually the last attention module outputs the updated values $\hat{\mathcal{F}}^{(1+\ell)}$ as the enhanced region level features which provide higher order feature interactions in between. Also, the process represented by eqns. (3.21) to (3.23) is repeated three times ($\ell = 3$) by stacking $(1 + \ell)$ XGL attention modules, to capture higher order interactions of $2(1 + \ell)^{th}$ order i.e., 8th order. To exploit even more higher order feature interactions, multiple XGL attention blocks need to be stacked. Consequently, it may lead to a huge rise in memory demand and computational cost. X-LAN [164] model attempted to represent infinity order interactions by developing X-LAN attention block with Exponential Linear Unit (ELU) activation. Instead, this work proposes, XGL attention module with x-GELU activation that outperforms X-LAN [164] while capturing feature interactions up to 8th order only and hence reducing the memory demand and the computational cost.

3.2.1.3.2 XGL-T Language Decoder

For a sentence, $\mathbb{S}_{1:\mathcal{T}} = \{w_1, w_2, \ldots, w_{\mathcal{T}}\}$ of $\mathcal{T}$ words, where $w_{\mathcal{T}}$ is the textual feature of $\mathcal{T}^{th}$ word. The captioning decoder aims to generate the output sentence using transformed ILF that are induced via image encoder. To further leverage the higher order inter-modal interactions between visual content and natural sentence, the proposed XGL attention module is integrated with Bi-LSTM to perform multi-modal reasoning to yield high-level feature representation. Bi-LSTM is, of course, an extension to standard LSTM wherein learning algorithm is fed with the input sentence $\mathbb{S}_{1:\mathcal{T}}$ once from the beginning to the end and vice-versa, Bi-LSTM based language learning model supports detailed encryption of visual data. At each decoding step, the mean-pooled RLF and attended ILF are concatenated to make this as an input to the embedding layer so as to obtain transformed ILF.

$$\tilde{\mathcal{F}} = \omega_{\mathcal{g}}\{\hat{\mathcal{F}}^{(0)}, \hat{\mathcal{F}}^{(1)}, \ldots, \hat{\mathcal{F}}^{(1+\ell)}\} \tag{3.24}$$

where, $\omega_{\mathcal{g}}$ is the embedding matrix. Further, the input to the Bi-LSTM is taken as the set of concatenation of current input word $w_{\mathcal{T}}$, the ILF $\tilde{\mathcal{F}}$, the previous hidden state h_{t-1} and the previous context vector c_{t-1}. Further, the output from the Bi-LSTM is fed as the query input to the XGL attention module. The keys and values for the module are set as the enhanced region level features $\hat{\mathcal{F}}$ thereby making the output of this attention module more meaningful as it can now capture higher order interactions between image features and hidden state. The context vector (symbol) is finally derived by concatenating Bi-LSTM current hidden state and the output attention feature followed by an embedding layer and GEGLU [168]. GEGLU is a

variant to Gated Linear Unit (GLU) [174] that improves the performance of the transformer by producing better perplexities for the de-noising objective. This variant of GLU is simple to implement with less computational drawbacks and provides better results on various language understanding tasks. Finally, next word $w_{\mathcal{T}+1}$ is predicted using a softmax layer and context vector c_t.

3.2.1. Experimental Work and Results

The experiments are conducted on the most widely used image captioning dataset MSCOCO [157]. The total number of images in the MSCOCO dataset is 123,287, inclusive of 82,783 training images, 40,504 validation images, and 40,775 test images. Each image encompasses 5 human annotated captions. It is worth mentioning that the official testing set does not come with annotations, and thus the only way to evaluate it is to use an online testing server. Further, for offline evaluation, the well-known Karpathy split [19] is used. It includes 113,287 training images, 5,000 validation images, and 5,000 test images. All training phrases are pre-processed by converting them to lower case and eliminating the words that appear less than 6 times, resulting in a vocabulary of 9,488 distinct words.

3.2.2.1 Implementation Details

To detect the objects and extract the image region features from these objects, a Faster-RCNN network, pre-trained on ImageNet [175] and Visual Genome [176], is incorporated. Input feature vector of dimension $[1 \times 2048]$ is transformed $[1 \times 1024]$ vector. For XGL attention module, the dimensionality of the query-key representation is set as $\mathcal{D}_q = 1024$ while for the transformed bilinear feature $\mathcal{D}_f = 512$. In the proposed XGL-T, 4 XGL attention modules are stacked in image

encoder and the decoder is equipped with only one XGL attention module. The implementation is done in python with PyTorch. The experimental setup makes use of optimizer [37], adopts the training schedule as in [154], and the whole architecture is optimized using cross-entropy loss.

The proposed model is first trained by minimizing the cross-entropy loss:

$$\mathcal{L}_{XE}(\theta) = -\sum_{t=1}^{\mathcal{T}} \log\left(p_\theta(w_t^* \mid \mathbb{S}_{1:t-1}^*)\right) \tag{3.25}$$

where, $\mathcal{T}$ is the number of words in sentence; θ denotes all the parameters in the model; $\mathbb{S}_{1:t-1}^*$ is the ground truth. Further, a Reinforcement mechanism [118] is used to optimize the CIDEr-D [25] metric.

$$\mathcal{L}_{XE}(\theta) = \mathbb{E}_{\mathbb{S}_{1:\mathcal{T}}^s \sim p_\theta}[\mathfrak{r}(\mathbb{S}_{1:\mathcal{T}}^s; \mathbb{S}_{1:\mathcal{T}}^*] \tag{3.26}$$

where the reward $\mathfrak{r}(.)$ indicates the CIDEr-D metric for the sampled sentence $\mathbb{S}_{1:\mathcal{T}}^s$ and the ground truth $\mathbb{S}_{1:\mathcal{T}}^*$

The mini batch size for the same is 40 with $10{,}000$ warmup steps. The number of iterations, set as 50 epochs, avoids slow convergence with low rank bilinear pooling. The captioning model is further optimized with CIDEr-D reward with learning rate 2×10^{-5} and epochs are set as 35. The inference stage uses beam search strategy with beam size of 3. Performance of the proposed work is reported on the basis of: (i) $BLEU@N$ $(B@N)$ [177] which is used for the comparison and counting of the number of co-occurrences and depends on $n-gram$ precision that computes per-corpus $n-gram$ co-occurrence, $n \in [1,4]$. (ii) $CIDEr-D$ (C) [25] provides evaluation in terms of saliency, grammar, and accuracy. (iii) $METEOR$ (M) [26] evaluates the scores for matching word, stem

and synonyms. (iv) $ROUGE-L$ (R) [152] measures syntactical matches rather than semantics, and (v) $SPICE$ (S) [153] measures the similarity between the machine generated scene graph tuples and the ground truth.

3.2.2.2 Results Obtained for the proposed x-GELU Activation

To understand the effect of proposed x-GELU activation function, the performance of a simple MLP based neural network with 3 hidden layers [169], is observed. Where, the ReLU activation function at each hidden layer is replaced with ELU, GELU, and proposed x-GELU activation function. The corresponding observed results for MSCOCO dataset, Karpathy split [19], are reported in Table 3.2.

Table 3.2: Performance comparison of Sigmoid, ReLU, ELU, GELU and x-GELU activation

Activation Function	Test_Acc (%)	Val_Acc (%)	Test_Loss	Val_Loss
Sigmoid	97.28	97.39	0.3262	0.3265
ReLU	98.34	98.21	0.2324	0.265
ELU	99.34	98.51	0.0076	0.0419
GELU	99.73	99.12	0.0035	0.0367
x-GELU	**99.85**	**99.13**	**0.0016**	**0.0082**

It is evident that x-GELU activation provides the best validation and testing accuracy of 99.13% and 99.85% respectively. Further, validation and test losses are obtained as 0.0082 and 0.0016 respectively. From Fig.'s $3.6(a)$, $3.6(b)$, $3.7(a)$, and $3.7(b)$ it is evident that the losses for Sigmoid and ReLU activation functions are high in comparison to ELU, GELU and x-GELU. Also, x-GELU activation function provides highest accuracy with minimum losses which proves the superiority of x-GELU activation function over Sigmoid, ReLU, ELU and GELU respectively.

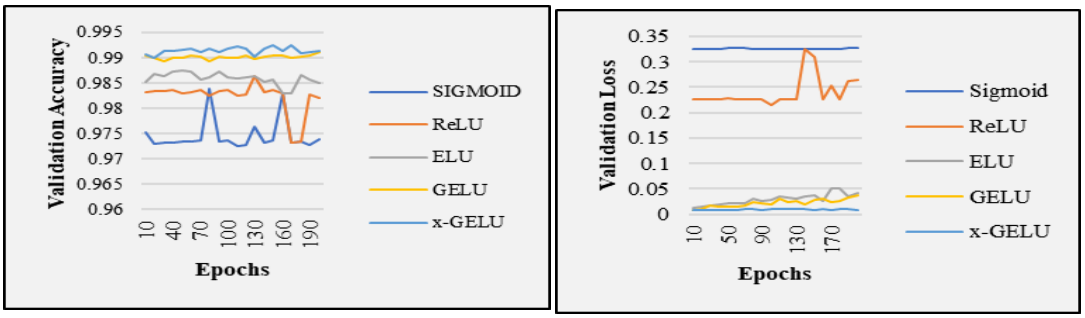


Fig. 3.6: (a) Validation Accuracy and (b) Validation Loss Plots for SIGMOID, ReLU, ELU, GELU and x-GELU activations

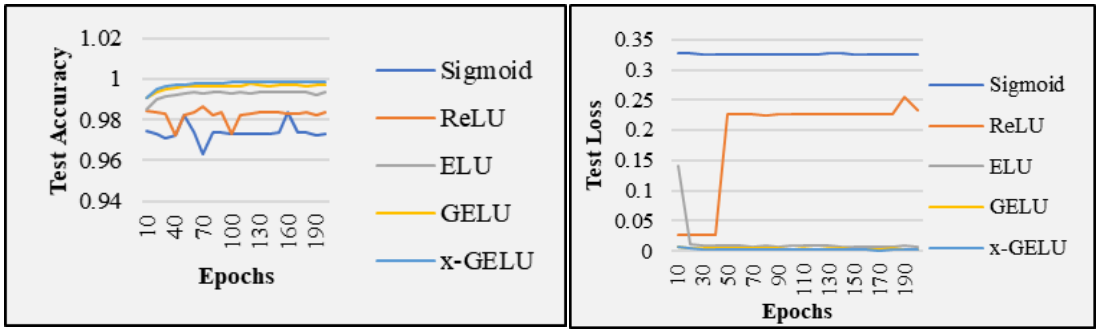


Fig. 3.7: (a) Test Accuracy and (b) Test Loss Plots for SIGMOID, ReLU, ELU, GELU and x-GELU activations

3.2.2.3 Quantitative Results

Table 3.3 represents the results of proposed XGL-T transformer on MSCOCO Karapathy test split. The results are evaluated by minimizing cross entropy loss and CIDEr-D score optimization which proves the efficiency of the proposed model. Better results are obtained for CIDEr-D score optimization by avoiding overfitting and thus it can better utilize the potential of large models.

A comparison of the proposed XGLT model with other state-of-the-art models is provided in Tables 3.4 and 3.5, for MSCOCO Karpathy test split for Cross Entropy

Table 3.3: Results of the proposed XGL-T on MSCOCO "Karapathy" test split

Score	Cross Entropy Loss	CIDEr-D Score Optimization
B-1	78.4	81.5
B-2	63.2	67.1
B-3	48.4	51.6
B-4	37.8	39.9
S	22.1	23.8
C	122.4	134.0
M	29.9	29.8
R-L	58.4	59.9

Loss and for CIDEr-D Score Optimization. The results prove the efficiency of the proposed image captioning XGL-T model. It is quite evident that the proposed model successfully makes the higher-order interactions between the objects and hence results in better caption generations in comparison to earlier state-of-the-art. Further, by incorporating an attention mechanism that learns to recognize specific spatial regions for sentence generation, RFNet [158] and Up-Down [123] significantly improve the performance when compared with [22] [178]. Additionally, GCN-LSTM [114] and SGAE [159] perform better than [123] while using rich semantic information from images (such as visual relations between objects or scene graph) for sentence generation. Intuitively AoANet [134] enhanced conventional visual attention by providing relevance between the query and the attention results. This supports the idea that enhancing attention mechanism is an effective way to improve the interaction between visual content and natural sentences, and hence improves the image captioning. Therefore, incorporation of the proposed XGL attention module in the encoder-decoder based transformer architecture has proved its worth so far as leveraging of higher order interactions is concerned.

Further, the proposed transformer model provides results, on all five evaluation parameters, better than the baseline transformer-based encoder-decoder structure [179]. It is also observed that XGL-T model helps boost the performance of the captioning module when compared with [161] [164] [180]. One of the key reasons to outperform X-Transformer [164] results is Bi-LSTM based decoding that takes advantage of semantic concepts of objects, attributes and relationships. The improvements in SPICE metric observations validate better correlation of the proposed XGL-T model with human assessment ability irrespective of word orders. The results shown in Fig. 3.8, further, verify that the sentences so generated provide a strong correlation with the human evaluation by including n-grams, word-pairs, and word-sequences.

Table 3.4: Performance of the proposed model and other state-of-the-art methods on MSCOCO "Karapathy" test split for Cross Entropy Loss

Method	B-1	B-2	B-3	B-4	C	R	M	S
CNN-LSTM [13]	-	-	-	29.6	94.0	52.6	25.2	-
SCST [118]	-	-	-	30.0	99.4	53.4	25.9	-
LSTM-A [22]	75.4	-	-	35.2	108.8	55.8	26.9	20.0
VS-LSTM [178]	76.3	-		34.3	110.2	-	26.9	-
RFNet [158]	76.4	60.4	46.6	35.8	112.5	56.5	27.4	20.5
Up-Down [123]	77.2	-	-	36.2	113.5	56.4	27.0	20.3
GCN-LSTM [114]	77.3	-	-	36.8	116.3	57.0	27.9	20.9
LBPF [180]	77.8	-	-	37.4	116.4	57.5	28.1	21.2
SGAE [159]	77.6	-	-	36.9	116.7	57.2	27.7	20.9
AoANet [134]	77.4	-	-	37.2	119.8	57.5	28.4	21.3
Transformer [179]	76.1	59.9	45.2	34.0	113.3	56.2	27.6	21.0
ETA [161]	77.3	-	-	37.1	117.9	57.1	28.2	21.4
X-Transformer [164]	77.3	61.5	47.8	37.0	120.0	57.5	28.8	21.8
X-LAN [164]	78.0	62.3	**48.9**	**38.2**	122.0	58.0	28.8	21.9
XGL-T (Proposed)	**78.4**	**63.2**	48.4	37.8	**122.4**	**58.4**	**29.9**	**22.1**

Table 3.5: Performance of the proposed model and other state-of-the-art methods on MSCOCO "Karapathy" test split for CIDEr-D score optimization

Method	B-1	B-2	B-3	B-4	C	R	M	S
CNN-LSTM [13]	-	-	-	31.9	106.3	54.3	25.5	-
SCST [118]	-	-	-	34.2	114.0	55.7	26.7	-
LSTM-A [22]	78.6	-	-	35.5	118.3	56.8	27.3	20.8
VS-LSTM [178]	78.9	-		36.3	120.8	-	27.3	-
RFNet [158]	79.1	63.1	48.4	36.5	121.9	57.3	27.7	21.2
Up-Down [123]	79.8	-	-	36.3	120.1	56.9	27.7	21.4
GCN-LSTM [114]	80.5	-	-	38.2	127.6	58.3	28.5	22.0
LBPF [180]	80.5	-	-	38.3	127.6	58.4	28.5	22.0
SGAE [159]	80.8	-	-	38.4	127.8	58.6	28.4	22.1
AoANet [134]	80.2	-	-	38.9	129.8	58.8	29.2	22.4
Transformer [179]	80.2	64.8	50.5	38.6	128.3	58.5	28.8	22.6
ETA [161]	81.5	-	-	39.3	126.6	58.9	28.8	22.7
X-Transformer [164]	80.9	65.8	51.5	39.7	132.8	59.1	29.5	23.4
X-LAN [164]	80.8	65.6	51.4	39.5	132.0	59.2	29.5	23.4
XGL-T (Proposed)	**81.5**	**67.1**	**51.6**	**39.9**	**134.0**	**59.9**	**29.8**	**23.8**

3.2.2.4 Qualitative Results

The qualitative performance analysis is carried out to validate the observed results discussed in sub section 3.2.2.3 above. An example, as shown in Fig. 3.8, showcases the captions are generated by the proposed XGL-T model and X-LAN [164] and their ground truth. A close look at the examples reveals that the proposed XGL-T method produces more descriptive captions by focusing more on the salient object regions. Further, it provides better correlation between the objects by dynamically integrating higher-order interactions using Bi-LSTM based decoding. For example, in the first case X-LAN [164] model is not able to detect '*laptop on the desk'* but the proposed XGL-T model recognizes it correctly. It has also been noticed that the XGL-T method generates more correct description of objects that are absolutely missing in the G.T. For example, *'remote'* is correctly predicted as *'cell-phone'*. This highlights the significance of obtaining high-order interactions.

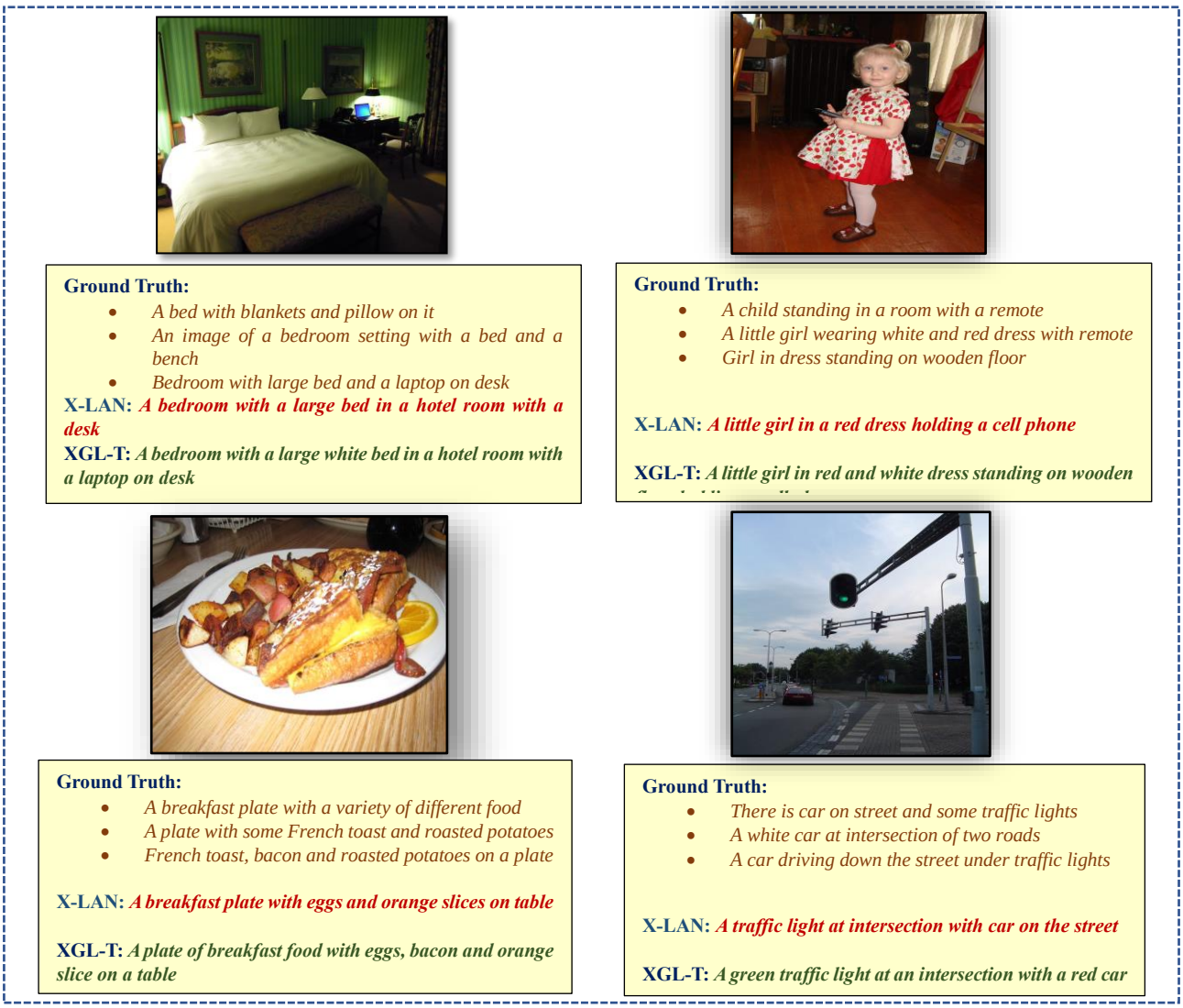


Fig. 3.8: Examples of our XGL-T captioning results compared with X-LAN with corresponding Ground Truth's

However, there are a few instances where the XGL-T model could detect the objects other than the actual one such as '*eggs'* were detected instead of '*French toast*'.

3.2.2.5 Ablation Study

An ablation study is carried out to exhibit the influence of proposed XGL attention module, Bi-LSTM based language learning model and GEGLU activation

defined in the proposed XGL-T model. Table 3.6 shows a step-by-step evolution of the proposed X-GLT model, while the captions generated by each variant, are presented in Fig. 3.9. We start with a very basic design, Model 1, that employs Faster R-CNN at the encoding phase and LSTM with CAM for generation of sentences. This ablated base model produces results similar to Up-Down approach [123]. It does not provide better correlation between different objects. Next, the base model is extended as Model 2, by replacing the base Model 1 with the transformer-based architecture which uses Faster R-CNN at the encoding phase with four stacked layers of XGL attention module and at the decoding phase, conventional attention is replaced with XGL attention module and LSTM with GLU. Model 2 exploited the higher-order interactions to some extent by focusing on the salient object. The results obtained for Model 2 provides enhancement in the capacity of multi-modal reasoning. This further validates the effectiveness of modelling high order interactions between image regions in encoder. This also provides more refined description of images with improvements in the evaluation parameters, w.r.t Model 1. Furthermore, in proposed Model 3, at the decoding end, LSTM module is replaced with the Bi-LSTM followed by GEGLU in place of GLU which provides better language understanding by providing more descriptive information of images. In terms of evaluation metrics there is substantial improvement observed in terms of B@N scores but a large performance gain is achieved in terms of CIDEr, METEOR, ROGUE and SPICE evaluation metrics. Further from Fig. 3.9, it is evident that the proposed model provides more precise description of images. This further demonstrates the benefit of using Model 3 to capture high order interactions between visual regions while simultaneously

Faster R-CNN + CAM + LSTM: *a street with building and cars*
Faster R-CNN + XGL Attention + LSTM + GLU: *People are sitting in a city street with lots of cars and stores*
Faster R-CNN + XGL Attention + Bi-LSTM + GEGLU: *People are sitting near variety of shops and stores on a busy city street with cars*

Faster R-CNN + CAM + LSTM: *a dog sitting on the floor looking at the camera*
Faster R-CNN + XGL Attention + LSTM + GLU: *a brown dog with collar sitting on the floor next to a couch*
Faster R-CNN + XGL Attention + Bi-LSTM + GEGLU: *a brown dog with collar sitting on the floor next to a table and chair*

Fig. 3.9: Ablation Studies Results for XGL-T

triggering high order interactions across other modalities for multi-modal reasoning.

Table 3.6 An ablation study for proposed XGL-T transformer

Model	Encoder	Decoder	B-1	B-2	B-3	B-4	C	R-L	M	S
M-1	Faster R-CNN	CAM + LSTM	76.4	60.3	46.7	36.1	114.1	56.7	27.9	20.9
M- 2	Faster R-CNN + 4 × XGL Attention	XGL Attention + LSTM + GLU	78.3	62.3	48.9	37.9	122.4	57.8	28.8	21.8
M-3 (XGL-T)	Faster R-CNN + 4 × XGL Attention	XGL Attention + Bi-LSTM + GEGLU	**78.4**	**62.5**	**49.0**	**37.9**	**124.1**	**58.4**	**29.9**	**22.1**

3.3 Significant Outcomes

This chapter presents two transformer-based image captioning models for the generation of single sentence description of images. The key highlights of the chapter include:

1. The chapter first introduces a Lightweight Transformer with GRU for image captioning with minimum number of encoding and decoding transformer structure. The proposed Lightweight Transformer incorporates a feature

encoding approach that extracts a priori knowledge through multiple high- and low-level appearance features. Also, it integrates a GRU layer in the decoder structure to enhance language model, with using just a single encoder and a single decoder overall. Extensive experiments on MSCOCO dataset demonstrated that our approach achieves a competitive score on all the evaluation metrics. Moreover, the results were obtained by a model trained with minimal computational resources. Further, qualitative analysis proves that proposed model can yield captioning results demonstrating better appearance awareness with a better language model. Furthermore, this approach can be implemented in any transformer variant to make it parameter-efficient without decreasing the performance.

2. This chapter also discusses an efficient XGL-Transformer model for image captioning. The proposed work defined a novel x-GELU activation driven XGL attention mechanism to generate captions whereby, XGL attention-based encoding and decoding technique were used reduce the vanishing gradient problem and further capture higher-order interactions. Further, superiority of proposed model was established through the experimental results in terms of CIDEr, SPICE. BLEU@$n, n \in [1,4]$, METEOR, ROUGE-L evaluation parameters. The ablation study results proved the improvement in the performance of the overall system showing the impact of proposed XGL attention modules by introducing them in encoder and decoder phases. Also, Bi-LSTM followed by GEGLU is employed in decoding phase that further demonstrates the improvement in the results by capturing higher order interactions between visual regions while simultaneously triggering higher

order interactions across other modalities for multi-modal reasoning. More so, the proposed model can further be utilized to capture higher-order interactions to generate paragraph-based captions where more than one caption can be generated for each image. Also, the addition of sentiments or emotions in the generated captions may further help generate stylized captions. Besides that, the concept of knowledge graph can also be introduced with the proposed transformer model that can improve the performance of the model.

Chapter- 4

Style-Transfer based Image Captioning

This chapter introduces a novel framework for the style-based caption generation using Style Embedding-based Variation Autoencoder (SE-VAE). The proposed framework learns both the unstructured (semantics) and structured (style) feature distributions jointly and generates controlled and plausible stylized captions by updating the style weights.

4.1 Control with Style: Style Embedding-based Variational Autoencoder for Controlled Stylized Caption Generation Framework

This chapter presents a novel stylized image captioning framework. Style-based description of an image provides a way to imitate the language expressing the behaviour of human beings. It learns about linguistic styles and utilizes this knowledge to generate style-based image descriptions. Fig. 4.1, distinguishes between the Factual (F) and style-based (romantic (R) and humorous (H)) caption generation. This example shows that style rich captions provide more reasonable and sentimental descriptions with different opinions or feelings. These artistic captions, so generated, can be further used in numerous applications like storytelling, for visually impaired solutions, visual question answering etc. and for providing better understanding of human expressions.

Ideally, a stylized image captioning model should fulfill two conditions, (i) provide the correct description of images, and *(ii)* generate correct stylized words or phrases in appropriate positions of descriptions. However, the generated sentences do not possess sufficient style-related information which is important to describe

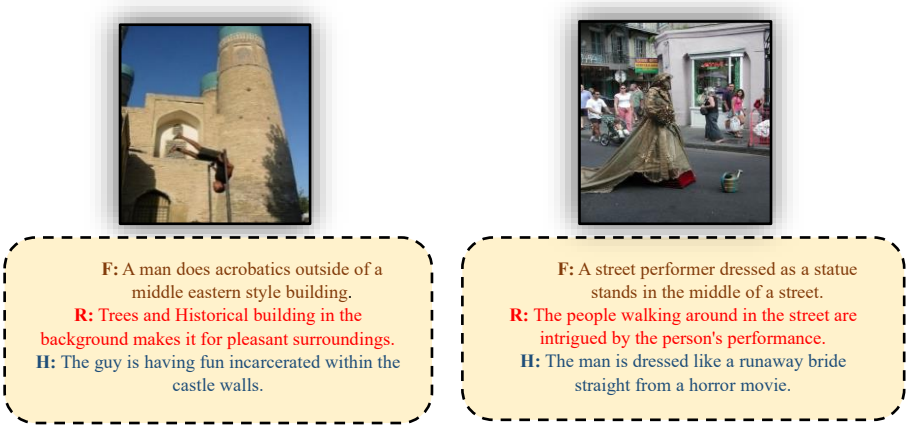


Fig. 4.1: Difference between Factual Image Captions and Stylized Image Captions

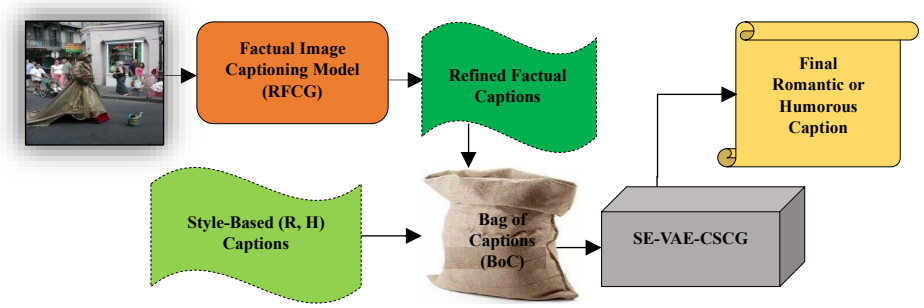


Fig. 4.2: Block Diagram Representation of the proposed framework

the content of an image. This may happen due to the small size of stylized dataset due to which it is difficult to preserve the correlations between the images and captions. This makes the generation of stylized descriptions very difficult. To overcome this challenge and capture the variability in style rich captions, the frameworks should utilize the unpaired set of images and captions, i.e., image is not available with any styled caption. This makes the caption generation more meaningful and acceptable in a wider range of applications by not being restricted to the availability of factual and styled captions. The proposed work defines Style Embedding-based Variational Autoencoder for Controlled Stylized Caption Generation Framework (SE-VAE-CSCG). It works in two phases. Initially, it generates refined factual captions and forms

a Bag-of-Captions (BoC) by combining both paired and unpaired sets of samples. In the second phase, the BoC feeds the captions to the SE-VAE Controlled Caption Generator. Fig. 4.2 represents a block diagram of the proposed method.

4.1.1 Proposed Methodology

The proposed framework is divided into two phases namely: (1) Refined Factual Caption Generation (RFCG), and (2) Stylized Image Captioning using SE-VAE, and modified controlled text generation module. The two-phase proposed framework is depicted in Fig. 4.3. To generate a style-based description of images, first the factual descriptions are extracted and the generated factual captions are fed to the SE-VAE-CSCG. This model generates stylized descriptions with an unpaired style transfer i.e., for a given image there is no need for corresponding factual and stylized captions. Also, this model does not require any image features for the generation of style-based descriptions.

4.1.1.1 Refined Factual Caption Generation

Refined Factual Image Caption Generation (RFCG) utilizes an encoder-decoder structure to generate refined factual captions of images using the available factual GT's per image. It helps the model produce human-independent (unbiased) refined captions by merging both visual features (global and object-level local features) with finetuned GloVe embeddings-based text encodings in the encoder. The decoder using Bi-LSTM and LSTM layers, learns the combined visual and textual encodings to generate the most likely refined factual descriptions unbiased from human perception.

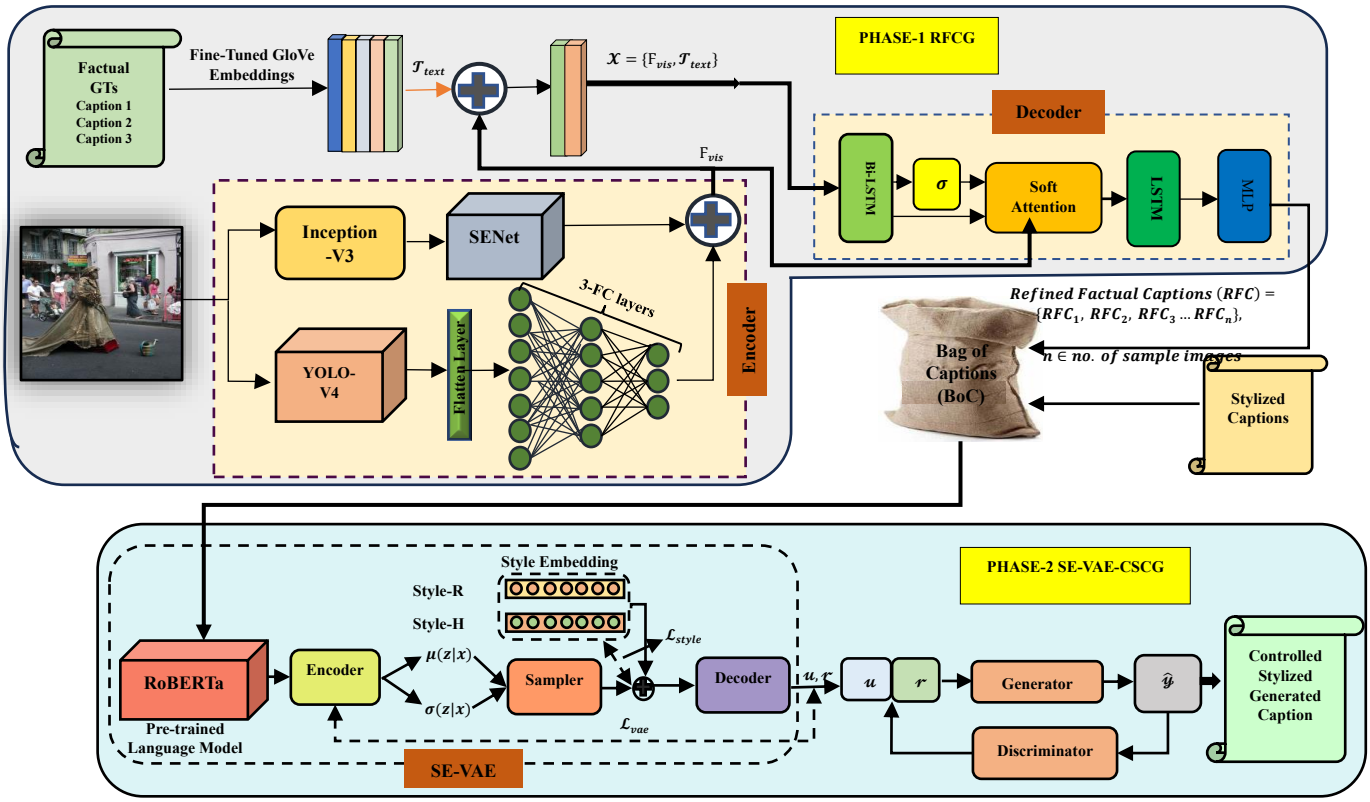


Fig. 4.3: Illustration of the proposed Control with style Framework: Phase-I RFCG Module, and Phase-II SE-VAE-CSCG Module. In Phase-I, RFCG module encoder generates visual embeddings as F_{vis} for the input image. Whereas RFCG module decoder receives combination of text embeddings $\mathcal{T}_{text}$ and the visual embeddings F_{vis} given by $\mathcal{X} = \{F_{vis}, \mathcal{T}_{text}\}$, generating refined captions as Refined Factual Captions (RFC) = $\{RFC_1, RFC_2, RFC_3 \ldots RFC_n\}$, n ∈ no. of sample images. From Phase I, a Bag of Captions (BoC) is defined that leverages Style Embedding based Variational Auto-Encoder-CSCG (SE-VAE-CSCG) in Phase-2 for generation of controlled stylized captions.

4.1.1.1.1 RFCG- Encoder

In the RFCG-Encoder, Inception-V3 [155] extracts global spatial features of images. Symmetric and asymmetric modules in 42 layered architecture of Inception V3 [155] provide low error rates with high efficiency when compared with its previous versions and

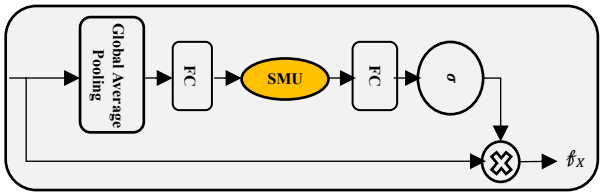


Fig. 4.4: SMU-activated SENet

its contemporaries. A feature vector of dimensions $(10 \times 10 \times 2048)$ extracted from the Inception-V3 is passed through a Smooth Maximum Unit (SMU) activated Squeeze and Excitation Network (SENet) [173] to improve the channel interdependencies with no added computational cost. It performs channel scaling with learned weighted features as f_{vec}. SMU [181] activated features are extracted to map with original channels, that has the potential to provide performance improvements when compared with traditional activation functions i.e., ReLU. SMU can be realized using the smooth activation of the maximum function which can smoothly approximate the ReLU activation. The internal architecture is provided in Fig. 4.4. Mathematically, squeeze and excitation operations are represented by:

$$z_{se} = f_{sq} = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} m_{se}(i,j) \tag{4.1}$$

$$s = f_{ex}(z, W) = s_{ex} \tag{4.2}$$

The output is scaled and is given by:

$$f_X = f_{scale}(s_{ex}, m_{se}) \tag{4.3}$$

To incorporate local descriptions of the visual semantics, local objects in the scene are detected using YOLO-V4 [182]. DropBlock regularization in YOLO-V4 supports better detection speed, accuracy, and mean average precision. The detected object features are flattened and passed through 3-fully connected layers resulting in a feature vector of

dimension $[1 \times 256]$. Now, both the global and local visual encodings are concatenated as $\mathrm{F}_{vis} = \{\mathbb{f}_1, \mathbb{f}_2, \ldots . \mathbb{f}_N\}, \mathbb{f}_i \in \mathbb{R}^{\mathcal{D}}, i \in (1, N)$ and are further merged with fine-tuned GloVe [183] embedding-based text encodings given by $\mathcal{T}_{text}$. Unlike Word2Vec [184], GloVe embeddings highlight word co-occurrences to obtain the word embeddings. Now, the text and visual encodings, are given as $\mathcal{X} = \{\mathrm{F}_{vis}, \mathcal{T}_{text}\}$, are combined and passed on to the RFCG decoder.

4.1.1.1.2 RFCG-Decoder

The RFCG language decoder is based on the Bi-LSTM network. The Bi-LSTM network very well deals with the problem of fixed sequence-to-sequence prediction. The combined text and visual embeddings $\mathcal{X} = \{\mathrm{F}_{vis}, \mathcal{T}_{text}\}$ are fed to the Bi-LSTM layer. The output generated from the Bi-LSTM network is activated with sigmoid activation. Also, the sigmoid-activated (σ) output is attended to using a soft attention mechanism. The soft attended features obtained are learned with the help of F_{vis}. Also, by multiplying the related segmentation map with low weight, soft attention discredits unimportant parts. Therefore, high attention zones retain their original worth, whereas low attention areas approach zero. The process can be described mathematically as:

$$\mathbb{S}_i^t = \mathcal{W}_{\mathbb{S}} \tanh(\mathcal{W}_{\mathcal{X}} \mathcal{X}_i + \mathcal{W}_h h_{t-1} + \mathscr{b}_s) + \mathcal{W}_f \mathbb{f}_i \tag{4.4}$$

where, $\mathcal{W}_{\mathbb{S}}$, $\mathcal{W}_{\mathcal{X}}$, $\mathcal{W}_{\mathbb{f}}$ $\mathcal{W}_h$ are the weights and $\mathscr{b}_s$ is the bias factor. Also, $\mathbb{S}_i^t$ is the importance score and

$$\mathbb{S}_i^t = (\mathbb{S}_1^t, \mathbb{S}_2^t \ldots . \mathbb{S}_n^t)^T \tag{4.5}$$

Furthermore, the attention at time t is given by:

$$a_i^t = \frac{\exp(\mathbb{S}_i^t)}{\sum_{i=1}^{n} \exp(\mathbb{S}_i^t)} \tag{4.6}$$

For improved feature learning, the soft-attended and sigmoid-activated features are learned using LSTM layers. Further, the resulting output generates a word on the next time node using the MLP function that incorporates backpropagation through a time algorithm to update the parameters of the LSTM network. The objective function defined for the optimization of the proposed RFCG is given by:

$$\vartheta = arg \max_{\vartheta} \sum_{m,y} \sum_{t=0}^{N} \log p(y_t | m, \vartheta, y_1, y_2 \dots y_{t-1}) \tag{4.7}$$

ϑ is the learnable parameter with m feature maps whose weight is $m \in \{m_1, m_2 \dots m_t\}$. y is the sentence that describes the image well. Further, the cross-entropy loss is incorporated to minimize the loss function to maximize the probability of each correct word appearing. The cross entropy-loss function is given by:

$$\mathcal{L}(\vartheta) = -\sum_{t=1}^{N} \log\left(p_\theta(w_t^* | \mathbb{G}_{1:t-1}^*)\right) \tag{4.8}$$

where N is the number of words in a sentence; ϑ denotes all the parameters in the model; $\mathbb{G}_{1:t-1}^*$is the Ground Truth (GT).

4.1.1.1.3 Bag-of-Captions (BoC)

The Refined Factual Captions (RFC) generated from the factual image captioning model, given as $RFC = \{RFC_1, RFC_2, RFC_3 \dots RFC_n\}, n \in no.of\ sample\ images$ and the stylized captions- Romantic captions (RC) and Humorous captions (HC) combinedly define BoC, as shown in Fig. 4.5. It consists of 7K paired sample set as $\{RFC, RC, HC\}$ while the remaining samples include only RFC obtained from Flickr30K and MSCOCO datasets. Therefore, the BoC contains a

set of paired captions and unpaired captions. All the samples are combined and shuffled which is further split into training, validation, and test sets. Table 4.1 presents the details of training, testing, and validation splits. The contents of BoC are fed to SE-VAE-CSCG which are utilized for the generation of diverse and stylized descriptions of an image without directly depending on the image features.

Table 4.1: Test, Train, and Validation Splits for BoC

Datasets	Captions in BoC	Training Samples	Testing Samples	Validation Samples
Flickr30K FlickrStyle10K	RFC- 31,783 RC-7000 HC- 7000	32,048	6868	6867
MSCOCO FlickrStyle10K	RFC-1,23,287	1,15,287	11,000	11,000

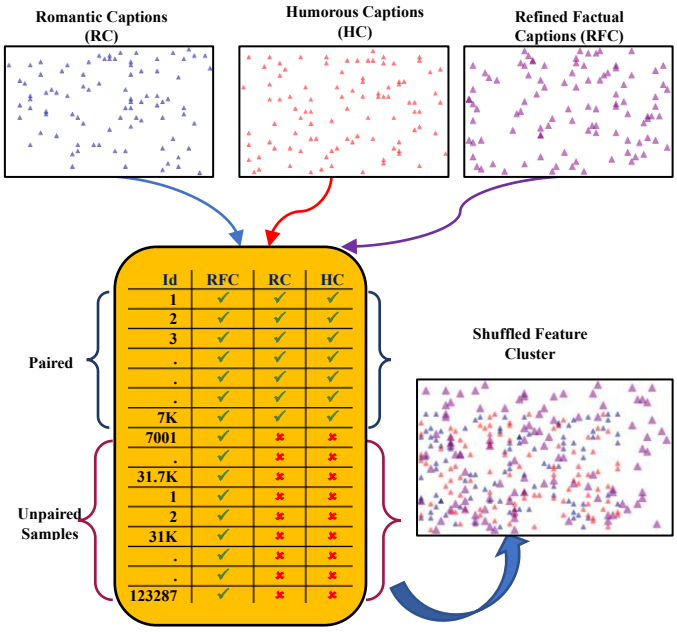


Fig. 4.5: Structure of Bag-of-Captions (BoC)

4.1.1.2 SE-VAE-CSCG Model

It presents the Phase 2 of the proposed framework which is defined in two parts: (i) SE-VAE module, and (ii) CSCG module. The proposed modules generate stylized

captions of an image using stylized embeddings based on variational autoencoder (SE-VAE) and controlled stylized caption generation (as depicted in Fig. 4.3).

4.1.1.2.1 SE-VAE Module

The recent controlled text generation-based works [185] [186] [187] have explored the controlled text learning ability of Variational Auto-Encoder (VAE). In view of these works, the proposed work, attempts to introduce style-based caption representation for controlled text generations by defining modified Style-Embedding based Variational Autoencoder (SE-VAE), as depicted in Fig. 4.3. It receives paired and unpaired sample sets from BoC in the form of contextualized vector representation, derived from RoBERTa [188] by inferencing from both sentence and word embeddings. The VAE encoder constructs a latent distribution $\mathbb{N}(\mu, \sigma)$ using a mean and a variance vector. The latent distribution, z, is assumed to be a normal distribution whose loss function is defined as:

$$\mathcal{L}_{vae} = -\mathbb{E}_{q_{(Z|X)}}[\log p(x|z)] + \varphi.\, \mathbb{KL}(q(z|x)||p(z)) \tag{4.9}$$

Where the first term in the above equation represents the likelihood of the reconstruction of the original text x, while the second term is the KL-divergence between the latent distribution and standard normal distribution. Further, $p(z)$ is the prior for standard normal distribution, and $q(z|x)$ is the posterior distribution. Also, φ is the balancing parameter that learns the capacity between self-reconstruction and style features.

Previous works [189] [190] incorporated disentangling style attributes whereas the proposed method utilizes the style learnt from structured samples to generate target style

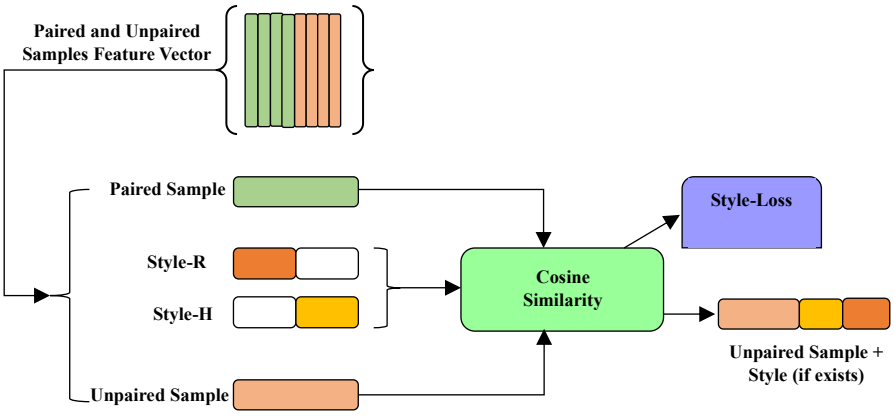


Fig. 4.6: Style-loss Calculation and Representation

representations for both unstructured and structured representations. It adjusts the style strength by simply adjusting the style weights. The embeddings are differentiated for two different styles (romantic, humorous) using one hot vector encodings as Style-R and Style-H. Also, it enhances the latent feature representations for both structured and unstructured latent vectors. Further, the style embeddings are represented by: $\mathbb{S} = \{s_1, s_2, \dots, s_k\}, s_i \in \mathbb{R}^{k \times d}$, where $k = 2$ is the number of styles (Style-R and Style-H). These embeddings are randomly initialized and updated by minimizing the similarity between the style embeddings and latent fusion. The structural representation is depicted in Fig. 4.6, where the similarity is minimized using the cosine similarity with the assumption that the style embedding is highly related to the latent features. The style loss for the same is defined as:

$$\mathcal{L}_{style} = -\sum_{i=1}^{k} \mathcal{b}_i \log\left(sigmoid\left(\cos\left(s_i, d(z)\right)\right)\right) \tag{4.10}$$

In eqn. (4.10), the sigmoid function ensures the range of cosine similarity. $\mathcal{b}_i$=1 if the style given represents the style of the input sentence, otherwise $\mathcal{b}_i$=0. $d(z)$ represents the stop gradient which is used to compute the style loss. Therefore, eqn. (4.9) is modified as:

$$\mathcal{L}_{\grave{vae}} = -\mathbb{E}_{q_{(Z|X)}}[\log p(x|z) + d(s_x)\,] + \varphi.\,\mathbb{KL}(q(z|x)||p(z)) \tag{4.11}$$

For sentence x, s_x represents its style embedding which is used as a constant vector. Hence, the total loss function defined for the proposed SE-VAE is:

$$\mathcal{L}'_{SE-VAE} = \delta_{vae}\mathcal{L}_{\grave{vae}} + \delta_{style}\mathcal{L}_{style} \tag{4.12}$$

where, δ_{vae} and δ_{style} are the hyperparameters that balance the weights between VAE-loss and style loss.

4.1.1.2.2 CSCG Module

Controlled Stylized Caption Generation (CSCG) module, leverages a generator and a discriminator structure. The output generated from the SE-VAE is represented in the form of unstructured representation u and structured representation r. It trains the generator $\mathcal{G}$ that reconstructs the captions for generating plausible text. Further, the discriminator $\mathcal{D}$ enforces the generator to produce coherent attributes. The generator and the discriminator form a pair of collaborative learners and provide feedback signals to each other. Also, the collaborative optimization represents the wake-sleep algorithm [191].

The generator $\mathcal{G}$ is a Deep LSTM-RNN architecture that generates the final caption sequence in the form of tokens $\hat{x} = \{\hat{x}_1, \ldots, \hat{x}_T\}$ conditioned on (u, r), mathematically represented as:

$$\hat{x} \sim \mathcal{G}(u, r) = p_{\mathcal{G}}(\hat{x}|u, r) = \prod_t p(\hat{x}|\hat{x}^{<t}, u, r) \tag{4.13}$$

where, $\widehat{x^{<t}}$ are the tokens preceding $\hat{x}$. Also, $\hat{x}$ is parametrized using the softmax function for time step t.

$$\hat{x} \sim softmax(o_t/\mathcal{T}) \tag{4.14}$$

$\mathcal{T} > 0$ is the temperature normally set to 1 and o_t is the logit vector.

The unstructured part incorporates standard Gaussian prior modelling for continuous variables whereas, the structured representation contains both continuous and discrete variables that encode different styles efficiently. Further, the SE-VAE is modified by combining the parameters of the encoder and generator. This modified SE-VAE loss is minimized to optimize the reconstructed captions, which is represented mathematically as:

$$\mathcal{L}_{SE-VAE} = -\mathbb{E}_{q_E(\mathcal{u}|x)q_{\mathcal{D}}(\mathcal{r}|x)}\left[\log p_{\mathcal{G}}(x|\mathcal{u},\mathcal{r}) + d(\mathcal{s}_x)\right] + \varphi.\,\mathbb{KL}(q_E(\mathcal{u}|x)||p(\mathcal{u})) \tag{4.15}$$

The distribution defined over $\mathcal{u}$ and $\mathcal{r}$ for the encoder and discriminator is mathematically given as:

$$\mathcal{u}{\sim}E(x) = q_E(\mathcal{u}|x) \tag{4.16}$$

$$\mathcal{D}(x) = q_{\mathcal{D}}(\mathcal{r}|x) \tag{4.17}$$

Further, these distributions are utilized as the output at the current step and the input to the next step along the sequence of decision-making. The resulting caption denoted by $\widetilde{\mathcal{G}_{\mathcal{T}}}(\mathcal{u},\mathcal{r})$, is given as input to the discriminator that measures the fitness to the target attributes for structured and unstructured representation. The loss for the same is calculated as:

$$\mathcal{L}_{attr,\mathcal{r}} = \mathbb{E}_{p(u)p(r)}\left[\log q_{\mathcal{D}}(\mathcal{r}|\,\widetilde{\mathcal{G}_{\mathcal{T}}}(\mathcal{u},\mathcal{r}))\right] \tag{4.18}$$

$$\mathcal{L}_{attr,\mathcal{u}} = \mathbb{E}_{p(u)p(r)}\left[\log q_E(\mathcal{u}|\,\widetilde{\mathcal{G}_{\mathcal{T}}}(\mathcal{u},\mathcal{r}))\right] \tag{4.19}$$

Combining the equations, the final generator objective function is given as:

$$min(\mathcal{L}oss) = \mathcal{L}_{SE-VAE} + \alpha\mathcal{L}_{attr,u} + \beta\mathcal{L}_{attr,r} \tag{4.20}$$

where α and β are the balancing parameters.

The discriminator is also a Deep-LSTM-based architecture that learned differently as compared to the SE-VAE encoder. The unstructured representation u is learned in an unsupervised manner while the structured representation r uses labelled examples to entail a designated style (romantic or humorous). Therefore, efficient semi-supervised learning is defined for the discriminator. To learn the specified semantic meaning and style, a set of labelled examples ($\mathcal{X}_L = \{(x_L, r_L)\}$) is used represented by:

$$\mathcal{L}_{disc} = \mathbb{E}_{\mathcal{X}_L}[-\log q_D(r_L|x_L)] \tag{4.21}$$

Besides the conditional generator, $\mathcal{G}$ synthesizes noisy style-attribute pairs which used semi-supervised learning. To alleviate this issue minimum entropy regularization [192] is incorporated to provide robust model optimization. The resulting objective is given by:

$$\mathcal{L}_{\mathcal{D}-\mathcal{O}} = \mathbb{E}_{p\mathcal{G}(\hat{x}|u,r)p(u)p(r)}[log q_{\mathcal{D}}(r|\hat{x}) + \rho\mathcal{H}\big(q_{\mathcal{D}}(r'|\hat{x})\big)] \tag{4.22}$$

where, $\mathcal{H}\big(q_{\mathcal{D}}(r'|\hat{x})\big)$ is the empirical Shannon entropy of the distribution $q_{\mathcal{D}}$ and ρ is the balancing parameter. The final training discriminator objective is given by the following equation:

$$\min\left(\mathcal{L}_{\mathcal{D}}\right) = \mathcal{L}_{disc} + \theta\mathcal{L}_{\mathcal{D}-\mathcal{O}} \tag{4.23}$$

4.2 Experimental Work and Results

To evaluate the performance of the proposed framework MSCOCO, Flickr30K, and FlickrStyle10K datasets are used in the experimentations. The Flickr30K and

MSOCOCO datasets are used for refined Factual Caption generation. To generate text-controlled stylized captions for any given image, style annotations for two different styles namely romantic and humorous styles are utilized from FlickrStyle10K [51]. For this dataset, only 7K style annotations are publicly available. This results in the collection of 7K paired samples and 24,783 from Flickr30K and 1,23,287 MSCOCO which are used for Phase II experiments.

4.2.1 Implementation Details

In phase-1, the Refined Factual Caption Generation (RFCG) is trained with Adam [193] optimizer with a learning rate of $1e^{-5}$ for the encoder module whereas for the language decoder, the learning rate is set as $4e^{-4}$. Further, the batch size is set as 64 and RFCG is trained for 60 epochs. Also, to extract the text features, fine-tuned GloVe embeddings are utilized with dimensions for embeddings as 300. To evaluate the performance of the proposed RFCG, BLEU@N [177] and METEOR (M) [26] scores are used.

For the generation of stylized captions, in Phase-2, using the proposed SE-VAE-based controlled text generation, the Adam optimizer is utilized with a learning rate of 0.0005. The input to the SE-VAE is the embeddings extracted from RoBERTa with a batch size of 512 tokens. The dimensions for the latent features and style embeddings are $[1 \times 768]$. The model can generate sentences with a sentence length limit of ≤ 25. To evaluate the performance of generated stylized captions, the Style Transfer Accuracy (cls) and Perplexity (ppl) of the proposed model are evaluated. Also, the relevancy of the captions generated is evaluated in terms of BLEU@N (B-1, B-3) [177] and METEOR (M) [26].

4.2.2 Experimental Results for Refined Factual Captioning Generation (RFCG)

This section of the chapter summarizes the performance of the proposed RFCG architecture on the Flickr30K and MSCOCO Karapathy split datasets. Table II and Table III present the comparison of the proposed RFCG based generated refined captions with other state-of-the-art on Flickr30K and MSCOCO datasets. Compared with the other state-of-the-art in Table 4.2, the proposed method provides a significant improvement in terms of B-1, B-2, B-3, B-4, and METEOR. The key reason for the consistently improved performance of the proposed RFCG model is that it leverages the benefit of rich Yolo-V4 based region specific local features and Inception-V3 based global visual features while merging with GloVe embeddings-based text features in comparison to raw VGG 16 based encoders [125], and CNN based encoders [127] visual features. In addition to this the proposed RFCG module incorporates Bi-LSTM and rich visual features enabled soft attention-based language decoder which provides further enhancements in the generating refined captions when compared with hard and soft attention [123]. Also, From Table 4.3 it is evident that the our proposed RFCG module provided significant results when compared with many recent state-of-the-art [134] [164] [194, 178, 114]. Also, [195] LLM-based fusion reports very low scores as fusion of captions collapse into a caption that does not offer more detail. To overcome this the proposed framework generates more detailed factual captions which are further fused with style-based captions for the generation of stylized image captions.

Further, when compared with [70] [196] the model provides improvements in terms of B-2, B-3, B-4, and M only. Figs. 4.7(a) and 4.7(b) present the validation and test accuracy and loss curves respectively for the proposed RFCG model. From this, it is evident that with the increase in the number of epochs, the accuracy of the proposed

RFCG model increases while the losses decrease. It is evident from the figures that after 60 epochs, the performance of the proposed RFCG module does not provide significant. improvements in the results. Hence, early stopping criteria is used to stop the training after receiving consistent results for next 10 epochs. Fig. 4.8 presents the qualitative results of the proposed RFCG. It is observed that the proposed RFCG expresses the relationships between objects, attributes, and scenes with more refined syntactic and semantic descriptions over base LSTM and other models [123, 125, 164].

Table 4.2: Comparison Results obtained for the proposed RFCG Module on Flickr30K Dataset

Model	B-1	B-2	B-3	B-4	M
LSTM [128]	66.3	42.3	27.7	18.3	-
G-LSTM [103]	64.6	44.6	30.5	20.6	17.9
Soft Attention [123]	66.7	43.4	28.8	19.1	18.5
Hard Attention [123]	66.9	43.9	29.6	19.9	18.5
Semantic Attention [127]	64.7	46.0	32.4	23.0	18.9
D-Ada [125]	66.7	48.6	32.1	22.4	21.4
Proposed RFCG	**69.6**	**49.4**	**32.1**	**23.3**	**21.6**

Table 4.3: Comparison Results obtained for the proposed RFCG Module on Flickr30K Dataset

Model	B-1	B-2	B-3	B-4	M
LSTM-A [22]	75.4	-	-	35.2	26.9
VS-LSTM [178]	76.3	-	-	34.3	26.9
Up-Down [123]	77.2	-	-	36.2	27.0
RDN [196]	77.5	61.8	47.9	36.8	27.2
GCN-LSTM [114]	77.3	-	-	36.8	27.9
AoANet [134]	77.4	-	-	37.2	28.4
X-Transformer [164]	77.3	61.5	47.8	37.0	28.8
Fusion-LLM [195]	-	-	-	29.0	28.7
M2 [70]	**80.3**	-	**-**	39.1	29.2
Proposed RFCG	80.2	**62.6**	**48.4**	**39.4**	**29.4**

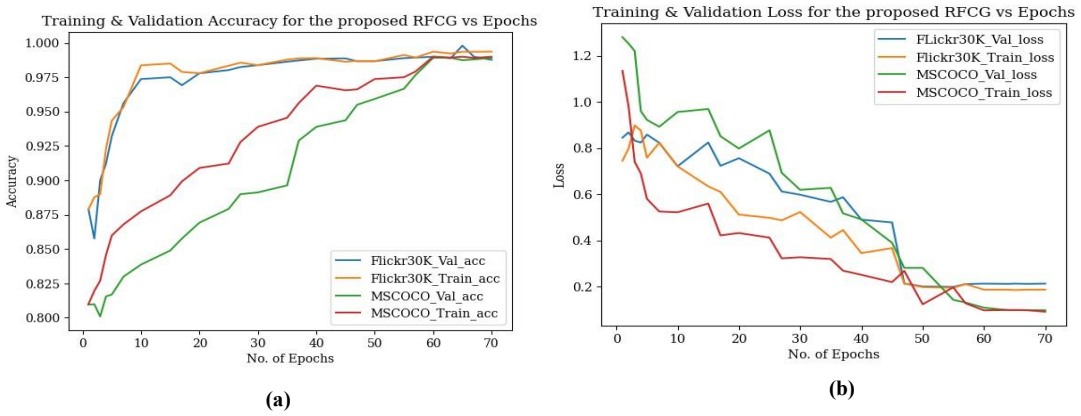


Fig. 4.7: Training & Validation (a) Loss, (b) Accuracy for the proposed RFCG

4.2.4 Experimental Results for SE-VAE-CSCG

For this phase, we define two test cases for our experiment: i) Test Case 1 and ii) Test Case 2. Test Case 1 utilizes 24,783 unpaired samples from Flickr 30K whereas Test Case 2 utilizes 1,23,287 unpaired samples from MSCOCO. Also, both the test cases contain 7K paired samples. Table 4.4 presents the comparison of the proposed SE- VAE CSCG model for both the test cases with other state-of-the-art in terms of BLEU-1, BLEU-3, and METEOR (M) scores with style accuracy (cls) and perplexity (ppl). StyleNet [54] and SF-LSTM [51] incorporate factored LSTM and style-factual LSTM decoders for end-to-end learning of stylized romantic and humorous captions. The works [58] [53] [197] [198] attempted multi-style-based caption learning using unpaired data but the proposed SE-VAE-CSCG captures the unstructured and stylized disentangled representations better than [58] [53]. Furthermore, using the concept of controlled styled SE-VAE, a significant rise of B-1 score value by an amount of 3.9, 15.8, 8.5 and 7.4 respectively is observed. From Table 4.4 we can infer that test case 2 provides significant results for all the evaluation metrics. Hence, increasing the number of unpaired samples confirms efficient learning of linguistic styles in the final

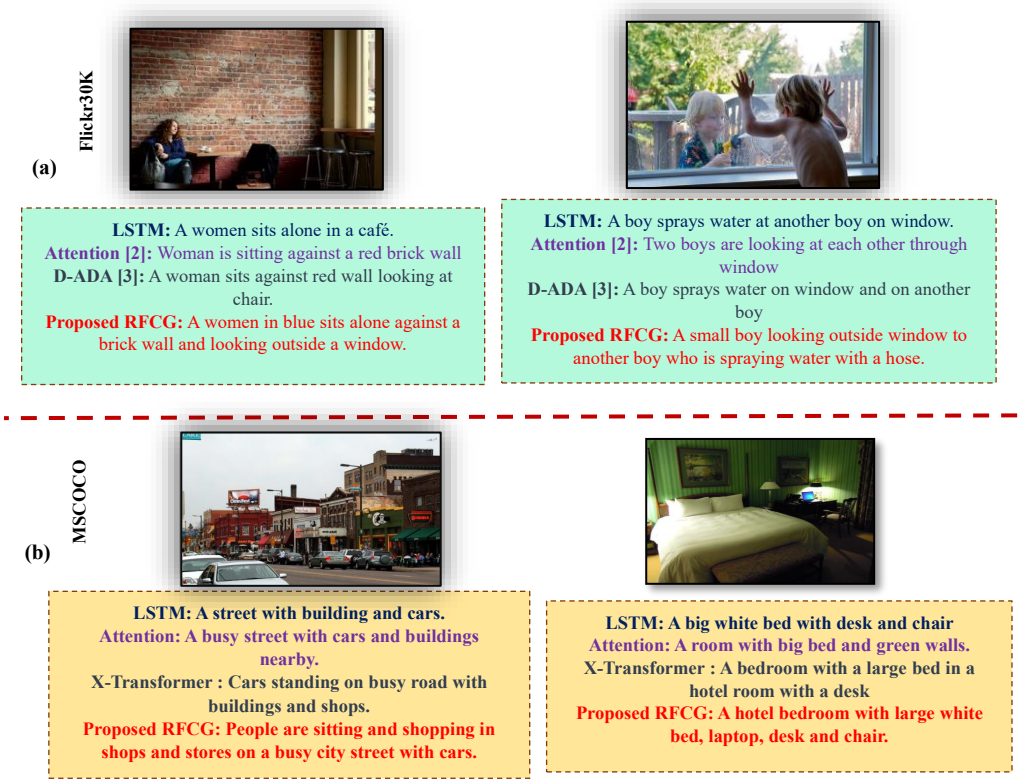


Fig. 4.8: (a) Refined Captions Generated by Proposed RFCG Model on Flickr30K and (b) MSCOCO dataset

encodings. Furthermore, it can be observed from the qualitative results of the proposed SE-VAE-CSCG, as reported in Fig. 4.9, that the image content is described with the usage of the correct style i.e., romantic and humorous with more realism, fluency, grammatical correctness and diversity.

Table 4.4: Comparison Results on FlickrStyle10K Dataset *(Test case 1: 24,783 unpaired samples from Flickr 30K; Test case 2: 1,23,287 unpaired samples from MSCOCO)*

Style	Model	B-1	B-3	M	cls	ppl
Romantic	StyleNet [51]	13.3	1.5	4.5	57.1	6.9
	MSCap [53]	17.0	2.0	5.4	91.3	-
	SF-LSTM [54]	27.8	8.2	11.2	-	-
	Detach and attach [197]	24.3	-	-	82.4	-
	Wu et al. [198]	25.4	5.7	9.2	-	-
	SAN [58]	28.3	8.7	11.5	90.9	9.1
	SE-VAE-CSCG (Test Case-1)	**30.2**	**9.4**	**12.5**	**91.6**	**9.7**
	SE-VAE-CSCG (Test Case-2)	**32.8**	**12.2**	**13.1**	**94.1**	**12.1**
Humorous	StyleNet [51]	13.4	0.9	4.3	42.5	7.3
	MSCap [53]	16.3	1.9	5.3	88.7	-
	SF-LSTM [54]	27.4	8.5	11.0	-	-
	Detach and attach [197]	23.0	-	-	89.2	-
	Wu et al. [198]	27.2	5.9	9.0	-	-
	SAN [58]	27.6	8.1	11.2	87.8	8.4
	SE-VAE-CSCG (Test Case-1)	**29.7**	**9.0**	**11.6**	**89.7**	**9.3**
	SE-VAE-CSCG (Test Case-2)	**31.2**	**10.6**	**12.0**	**92.4**	**11.6**

Fig. 4.9: Stylized Romantic and Humorous Captions Generated Using Controlled Stylized Caption Generation (CSCG) (a) With Flickr30K unpaired samples (24,783) and (b) With MSCOCO unpaired samples (1,23,287 samples).

4.2.5 Ablation Study

An ablation study is carried out to exhibit the influence of (i) RFCG module and (ii) SE-VAE-CSCG module (iii) number of unpaired captions fed to the BoC from MSCOCO dataset on the performance of the proposed framework.

1. In RFCG module, the encoder utilizes efficient Inception-V3 + Yolo-V4 visual features. The key reason to use Inception-V3 + Yolo-V4 based features, is highlighted in the ablation study reported in Table 4.5, for Flickr30K dataset. It confirms that RFCG module performs superior for Inception-V3 + Yolo-V4 based visual features over VGG-16, Inception V3, R-CNN, Yolo-V4, with a prominent rise in B-1, B-2, B-3, B-4 and M scores.
2. The ablation study results for proposed SE-VAE-CSCG module are shown in Table 4.6. It can be observed that the style-based captions are generated using three different models Basic-VAE + CSCG, VAE + CSCG, and SE-VAE + CSCG respectively. It is observed that there is a significant increase in the B-1, B-2, M, cls, and ppl scores for the proposed SE-VAE-CSCG module. The scores obtained for both Test Cases 1 and 2 make it evident that the proposed SE-VAE-CSCG module generates stylized descriptions with favourable accuracy, and perplexity trade-offs by efficiently lifting the word-level knowledge to sentence-level knowledge and learning disentangled representations.
3. Further, ablation is also carried out to study the influence of number of unpaired captions fed to the BoC from MSCOCO dataset to exhibit the generalizability and independence of the proposed framework over limited no of structured captions. Initially, for MSCOCO and FlickrStyle10K dataset we

utilized 7K paired samples and 20% unpaired samples from MSCOCO. With increase in the number of samples from 20% to 50%, 80%, and 100% samples a prominent increase in values of B-1 and cls is observed and is as shown in Fig. 4.10.

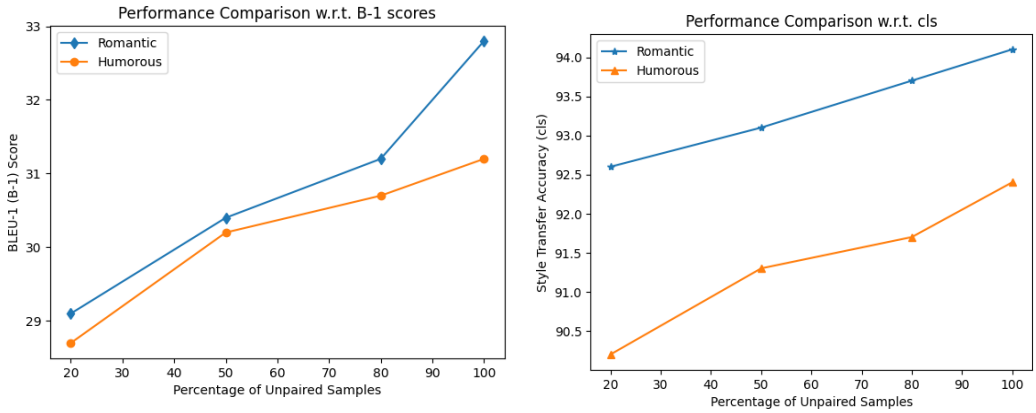


Fig. 4.10: Ablation study comparison results for percentage of number of samples utilized for generation of romantic and humorous captions.

Table 4.5: Ablation Study Results for Proposed RFCG Module for Flickr30K Dataset

Encoder Feature Representation	B-1	B-2	B-3	B-4	M
VGG-16	62.0	43.4	26.1	19.2	17.8
Inception-V3	64.7	45.2	26.2	19.5	18.9
R-CNN	65.8	45.7	27.3	20.1	19.2
Yolo-V4	65.9	45.7	28.1	20.7	19.3
Inception-V3 + Yolo-V4	**69.6**	**49.4**	**32.1**	**23.3**	**21.6**

4.3 Significant Outcomes

This chapter presents a novel style-based caption generation framework that generates style-controlled descriptions of images in two phases: (i) Refined Factual Caption Generation (RFCG), and (ii) Controlled Style Caption Generation, SE-VAE-CSCG. The proposed framework generates realistic and stylized descriptions of images with controlled text generation. The proposed RFCG generates meaningful

Table 4.6: Ablation Study Results for Proposed SE-VAE-CSCG Module *(Test case 1: 24,783 unpaired samples from Flickr 30K; Test case 2: 1,23,287 unpaired samples from MSCOCO)*

Style	Case	Basic -VAE	VAE	SE	CS CG	B-1	B-3	M	cls	ppl
Romantic	Test Case 1	✓	×	×	✓	18.7	5.4	6.2	66.7	6.5
		×	✓	×	✓	27.2	8.2	10.5	86.4	7.9
		×	✓	✓	✓	**30.2**	**9.4**	**12.5**	**91.6**	**9.7**
	Test Case 2	✓	×	×	✓	18.9	6.1	6.4	68.3	6.8
		×	✓	×	✓	28.1	8.6	10.8	87.9	7.9
		×	✓	✓	✓	**31.7**	**11.1**	**14.2**	**93.7**	**11.3**
Humorous	Test Case 1	✓	×	×	✓	18.2	5.1	6.1	62.1	6.2
		×	✓	×	✓	26.9	8.1	10.2	85.7	7.2
		×	✓	✓	✓	**29.7**	**9.0**	**11.6**	**89.7**	**9.3**
	Test Case 2	✓	×	×	✓	18.8	5.4	6.3	67.1	7.1
		×	✓	×	✓	27.4	8.7	10.9	87.0	7.5
		×	✓	✓	✓	**30.3**	**10.1**	**11.9**	**91.4**	**10.6**

human-independent (unbiased) refined captions by merging both visual features (global features and object-level local features) and textual features. The presented SE-VAE-CSCG module exhibits independence on the number of structured samples, as it delivers refined stylized captions even for majority of unstructured samples projected in terms of improved style accuracy.

This chapter focuses on another novel concept of Bag of Captions (BoC) which is a collection of both paired and unpaired samples of captions. This helps learn disentangled representations by lifting the word-level knowledge to sentence-level knowledge. Further, an ablation study is also conducted to support the experiments. Future work may involve large-scale stylized datasets incorporating different styles.

Chapter-5

Paragraph or Dense Image Captioning Model

This chapter focusses on the image captioning model that describes an image in the form of long narrative and unified story describing the semantic details of an image. This generates human-like detailed descriptions in the form of multiple sentences rather than a single sentence. Therefore, describing an image by covering more fine-grained entities that lead to dependability in semantic rather than wording.

5.1 MrA^2VAT: Multi Resolution and Adaptive Attention driven Variational Autoencoder Transformer for Dense Paragraph Image Captioning

While image paragraph generation has improved, still the existing methods fail to maintain coherence and consistency in describing the image contents and thus may lead to the possibility of misleading information [67]. This is mainly due to the fact that current approaches only take into account visual attention during the current time step, thereby lacking compositional reasoning such as object relationships and comparison. Also, models were developed that focussed on language policy rather than visual policy resulting in poor or irrelevant image description, as a consequence of which a lack of link between sentences within the paragraph was focussed. Furthermore, language diversity and the generation of redundant information are also serious issues when generating paragraph-based image descriptions.

5.1.1 Proposed Methodology

This chapter presents a novel framework MrA^2VAT **M**ulti-**R**esolution and **A**daptive **A**ttention driven **V**ariational **A**utoencoder **T**ransformer for paragraph image

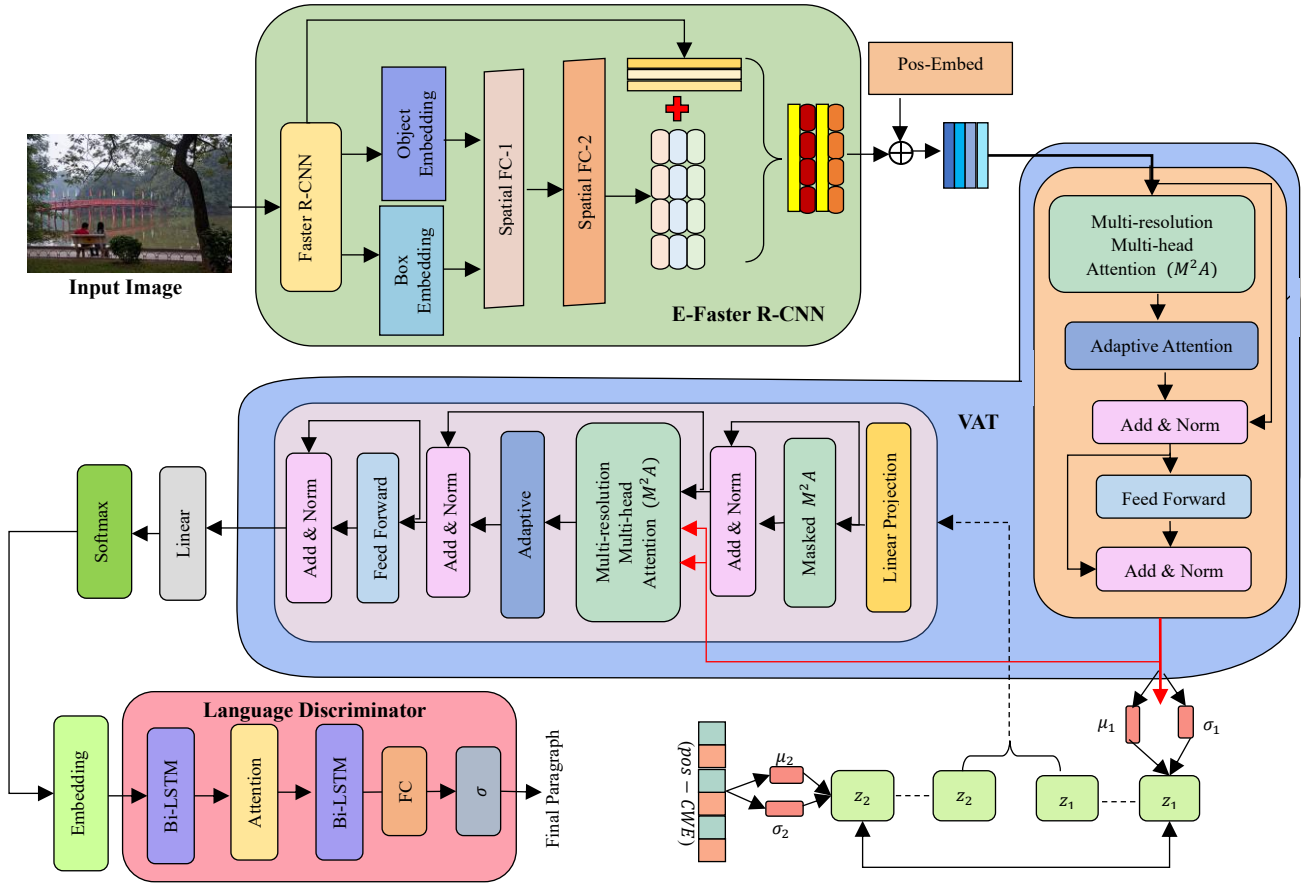


Fig. 5.1: Structure of the proposed MrA^2VAT with Language Discriminator: Input image features are extracted from the E-Faster-R-CNN which further processed by the proposed MrA^2VAT to generate paragraph-based image captions. Language discriminator further enhances the performance of the proposed framework by generating dense and coherent paragraph-based descriptions.

captioning. The proposed framework leverages a multi-level variational autoencoder based transformer that captures consistent long-term structure and provides correlation between visual and long-text embeddings for the generation of intermediate paragraph-based descriptions. To the best of my knowledge, the work is a maiden attempt in the field of paragraph image captioning that utilizes VAT. The proposed MrA^2VAT framework as depicted in Fig. 5.1, aims to generate diverse paragraphs with

reduced redundancy by the utilization of a language discriminator and dissimilarity score. The main contributions of the work are as follows:

(1) This Chapter presents a novel multi-level variational autoencoder transformer driven by Multi-resolution Multi-head Attention (M^2A) and Adaptive Attention (AA) for generation of dense and coherent paragraph-based descriptions.

(2) Further, a fused positional character and word embedding ($pos - CWE$) that utilized absolute and relative position information by encoding the position of each word.

(3) To avoid the generation of repetitive and monotonous descriptions, the proposed framework uses language discriminator with a dissimilarity score that improves the performance of the generated paragraphs especially in terms of BLEU-1 and METEOR scores.

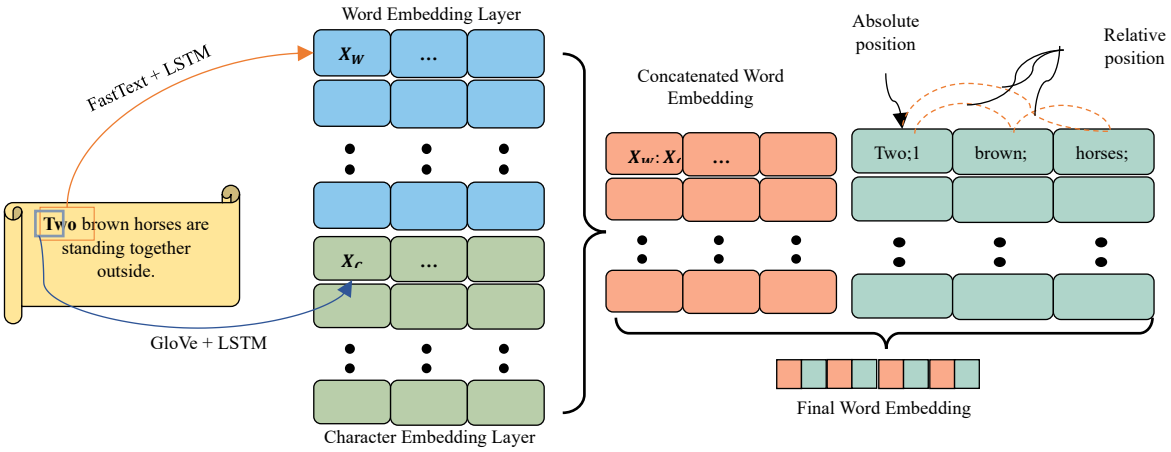


Fig. 5.2: Final Word Embedding representation from Fused Position Character and Word Embedding ($pos - CWE$).

5.1.1.1 Fused Positional Word and Character Embedding ($pos - CWE$)

As shown in Fig. 5.2, the proposed $pos - CWE$ converts the input paragraph

information into a corresponding relationship matrix with the utilization of character and word embedding matrix. The character embedding layer exploits a pre-trained GloVe [183] embedding with LSTM whereas the character embedding uses FastText [199] word embedding with LSTM. Let the dimensions of the word vector ($\mathcal{v}_{word}$) and character embedding vector ($\mathcal{v}_{char}$) be $e^{\mathcal{w}}$ and e^{c} respectively. Also, consider the length of each word be ℓ such that the word $\mathcal{w}$ is represented with the combination of character c as $e^{\mathcal{w}} * \ell$. Hence the final word vector for word $\mathcal{w}$ is $[\mathcal{v}_{word}; \mathcal{v}_{char}] \in \mathbb{R}^{e^{\mathcal{w}}+e^{c}}$. The proposed $pos - CWE$ also incorporates the concept of positional information to obtain the final input embedding. Locational information is considered important for the generation of paragraphs. In the case the words: *"Two boys are playing football"* and *"Two football are playing balls"* are the same for the embedding model but their meanings are entirely different. Therefore, position is an important concept which is to be added while designing the text embedding. Hence, to overcome these types of issues, this chapter presents the concept of positional (absolute and relative positions) encoding that numbers the position of each word. Also, with this mechanism one can easily distinguish words at different positions.

Further, the absolute and relative position embedding vectors are expressed as:

$$PE_{(pos,2i)} = \sin(pos/10000^{2i/d}) \tag{5.1}$$

$$PE_{(pos,2i+1)} = \cos(pos/10000^{2i/d}) \tag{5.2}$$

where pos is the position of each word, i represents the dimension of i^{th} word, and d is the dimension of the word vector. In addition to absolute position, equations (5.1) and (2) can express relative position relationships. same can be explained as:

$$\sin(\alpha + \beta) = sin\alpha * cos\beta + cos\alpha * sin\beta \quad (5.3)$$

$$\cos(\alpha + \beta) = cos\alpha * cos\beta - sin\alpha * sin\beta \quad (5.4)$$

Let us assume position vectors m and n, such that $n = m + \ell$, where ℓ is the distance between vectors m and n. Also, as per eqn. (5.3), $\sin(n) = \sin(m + \ell)$. Hence, position vector n can be expressed as the linear change of the position vector, m thereby representing the relative position information.

5.1.1.2 Multi-Resolution and Adaptive Attention driven Variational Auto Encoder Transformer MrA^2VAT

The proposed framework is two folds: (i) the former presents Multi-resolution Multi-head attention and Adaptive attention driven multi-level VAT whereas, (ii) the latter incorporates a language discriminator and calculates a dissimilarity score with length penalty for the generation of enhanced paragraph-based description of images. With the application of knowledge of multi-level VAE in transformer structure, the inference network encodes each latent variable upstream to determine its posterior distribution, whereas the generative network samples downward to obtain the distributions across the latent variables. The latent variable distribution at the bottom is inferred from the top-layer latent codes, rather than fixed (as in a standard VAE model).

5.1.1.2.1 Image Feature Extraction using (E-Faster R-CNN)

The input to the proposed MrA^2VAT encoder is the visual embedding generated by detecting $\mathcal{N}$ objects from image $\mathbb{I} = \{i_1, i_2, \dots, i_{\mathcal{N}}\}$ by leveraging Faster R- CNN that generated visual features $\mathcal{F} = \{f_1, f_2, \dots f_{\mathcal{N}}\}, f_k \in \mathbb{R}^{2048}$. Let the bounding boxes

with width, height, and its center coordinates (w, h, x, y) be represented by $\mathcal{B} = \{\mathscr{b}_1, \mathscr{b}_2, \ldots, \mathscr{b}_{\mathcal{N}}\}, \mathscr{b}_k \in \mathbb{R}^D$. Predefined geometry patterns are utilized to represent spatial relation embeddings $\mathcal{Q} = \{\mathcal{Q}_{ij} : \mathcal{Q}_{ij} \in \mathbb{R}^D\}$. Also, vector $\delta_{ij} \in \mathbb{R}^4$ provides geometric following relation of two bounding boxes [119].

$$\delta_{ij} = log\left(\frac{|x_i - x_j|}{w_i}\right), log\left(\frac{|y_i - y_j|}{h_i}\right), log\left(\frac{|w_i|}{w_j}\right), log\left(\frac{|h_i|}{h_j}\right) \tag{5.5}$$

Finally, δ_{ij} is projected into a high-dimensional space as $\Theta_b(i, j)$. Therefore, the spatial embedding $\mathcal{Q}_{ij}$ is given by the relation:

$$f_k'' = ReLU(W_p f_k + b_p) \tag{5.6}$$

$$\mathcal{Q}_{ij} = \mathscr{r}(Concat(\Theta_b(i, j), f_i'', f_j'') \tag{5.7}$$

where, $W_p \in \mathbb{R}^{D \times 4D}$ and $b_p \in \mathbb{R}^D$. Also, f_k'' is the object feature vector projection and $\mathscr{r}(\cdot)$ is the two-layer MLP.

5.1.1.2.2 MrA²VAT for generation of paragraphs

The visual features generated are the linear projections of the input which are split into three matrices query $\mathscr{q}$, key $\mathscr{k}$, and values $\mathscr{v}$ such that:

$$\mathscr{q} = \mathcal{X}\mathscr{w}^{\mathscr{q}}, \mathscr{k} = \mathcal{X}\mathscr{w}^{\mathscr{k}}, \mathscr{v} = \mathcal{X}\mathscr{w}^{\mathscr{v}} \tag{5.8}$$

where, $\mathscr{q}$, $\mathscr{k}$, $\mathscr{v} \in \mathbb{R}^{n \times d}$ and $\mathscr{w}^{\mathscr{q}}, \mathscr{w}^{\mathscr{k}}, \mathscr{w}^{\mathscr{v}} \in \mathbb{R}^{d \times d}$. Before the application of attention layer, we employ segment means [200] to compress the dimensions of keys and values and by selecting the best three heads. The best three heads available are known as fine-grained, coarse-grained, and medium-grained feature representations of keys and values respectively. Similarly, to explore the relationship between query representation and attention granularity, we allow the query to select the appropriate resolution based on the information encoded in its representation. This is done by incorporating a router [201] before the attention layer which helps in the selection of

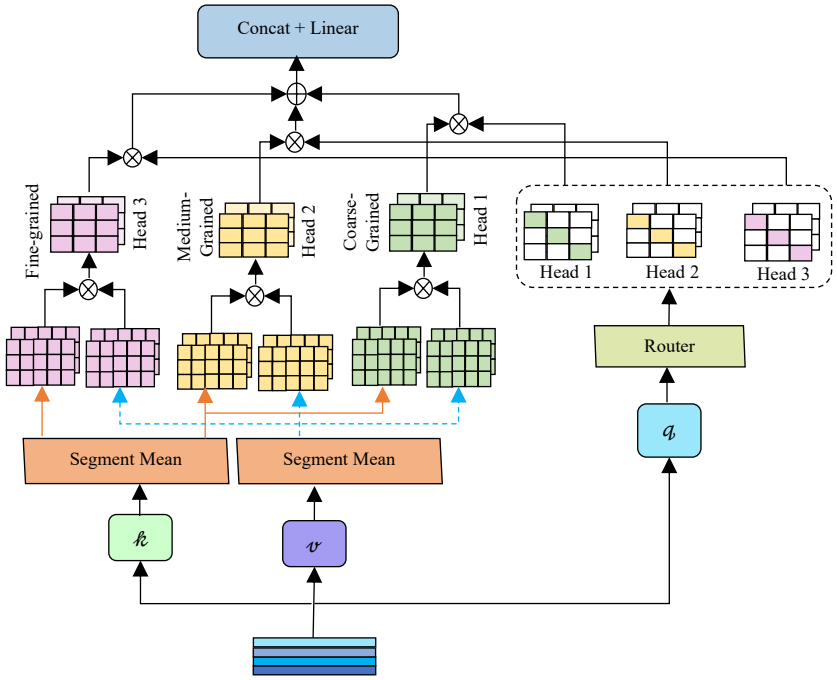


Fig. 5.3: Proposed Multi-Resolution Multi-Head Attention

best three attention heads. We also adopt a parametrized function which is normalized via a SoftMax layer. This projects query from $d - dimension$ to $H - dimension$ generating the head with higher router probability:

$$\mathcal{P} = SoftMax(\Gamma(q)) \tag{5.9}$$

$$\Gamma(q) = qw \tag{5.10}$$

Further, the proposed Multi-Resolution Multi-Head Attention (M^2A) adopts Kernal attention [202]. The same is depicted in Fig. 5.3 and can be expressed mathematically as:

$$Attention\left(q_i^h, \tilde{k}^h, \tilde{v}^h\right) = \frac{\sum_{j=1}^{N} sim(q_i^h, \tilde{k}^h{}_j)\tilde{v}^h{}_j}{\sum_{j=1}^{N} sim(q_i^h, \tilde{k}^h{}_j)} \tag{5.11}$$

$$Attention\left(q_i^h, \tilde{k}^h, \tilde{v}^h\right) = \frac{\Psi(q_i^h)^T \sum_{j=1}^{N} \Psi(\tilde{k}^h{}_j)(\tilde{v}^h{}_j)^T}{\Psi(q_i^h)^T \sum_{j=1}^{N} \Psi(\tilde{k}^h{}_j)} \tag{5.12}$$

After this, the best three heads are divided into multiple sub-heads (of same number) whose resolution is same as that of the original head. This allows the attention

model to jointly learn the information at different positions from different representational subspaces. Finally, the M^2A attention is expressed mathematically as:

$$M^2A(q, v, k) = (\sum_{h=1}^{H=3} Head_h)w^0 \quad (5.13)$$

$$Head_h = Concat(sh_0, sh_1, \dots, sh_s) \quad (5.14)$$

$$sh_s = attention\left(q^h W_s^q, \tilde{k}^h W_s^k, \tilde{v}^h W_s^v\right) \quad (5.15)$$

As the layers deepen, they contain higher-dimensional aspects of objects, such as those of individual sections or the entire entity. Therefore, to preserve the quality of the features extracted and to preserve the importance of the information proposed, VAT also utilizes adaptive attention mechanism. This enhances the quality of distribution and extraction of more discriminant features. Hence, the adaptive attention block provides refinement in the features obtained from the M^2A attention module. The structure of the adaptive attention is shown in Fig. 5.4.

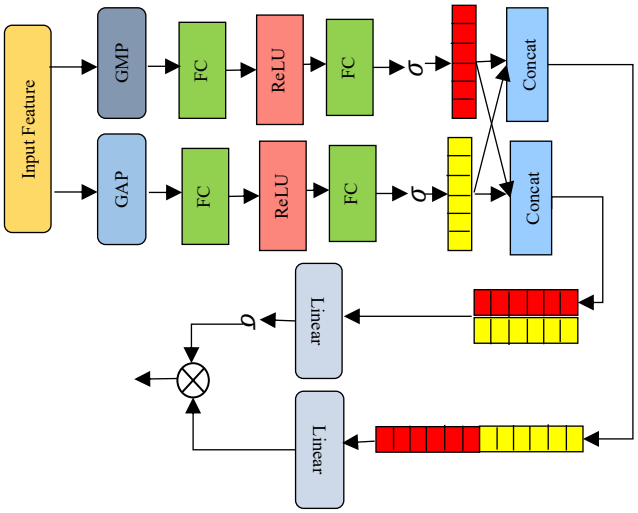


Fig. 5.4: Structure of the proposed Adaptive Attention

To extract useful high-level feature information without loss of generality, the proposed VAT exploits a two-layer hierarchy of latent variables z_1 and z_2 obtained from MrA^2VAT encoder and the textual features obtained from $pos - CWE$. The

latent variables z_1 and z_2 are sampled from a Gaussian Distribution with mean values μ_1, μ_2 and covariances as σ_1, σ_2 respectively. z_1 and z_2 are sampled stochastically from the posterior distribution, and text sequences are generated on z_1 and z_2 via a generative decoder. It is pertinent to mention that earlier VAE-based models do suffer from posterior collapse issue. But, with the utilization of hierarchical latent variable, we are able to mitigate this issue. The posterior distribution over the latent variables is assumed to be conditionally dependent on the input x. The joint posterior distribution for two latent variables is given by:

$$q_\phi(z_1, z_2|x) = q_\phi(z_2|x)q_\phi(z_1|x) \tag{5.16}$$

Considering the generative network, the latent variable at the bottom is sampled conditioned on the one at the top. Thus,

$$p_\theta(z_1, z_2) = p_\theta(z_2)p_\theta(z_1|z_2) \tag{5.17}$$

Further, the loss is given using the equation (5.18) considering the effect of equations (5.16) and (5.17) and abbreviating p_θ and q_ϕ as p and q respectively:

$$\mathcal{L}'_{ml-vae} = \mathbb{E}_{q\phi(z_1|x)}[logp(x|z_1)] - \mathcal{D}_{KL}(q(z_1, z_2|x)||p(z_1, z_2)) \tag{5.18}$$

$$\mathcal{D}_{KL}(q(z_1, z_2|x)\big|\big|p(z_1, z_2)\big) = \int q(z_2|x)\, q(z_1|x) log \frac{q(z_2|x)q(z_1|x)}{p(z_2)p(z_1|z_2)} dz_1 dz_2 \tag{5.19}$$

$$= \int_{z_1,z_2} [q_\phi(z_2|x)q_\phi(z_1|x) log \frac{q_\phi(z_1|x)}{p_\theta(z_1|z_2)} + \int q(z_2|x)\, q(z_1|x) log \frac{q(z_2|x)}{p(z_2)} dz_1 dz_2] \tag{5.20}$$

$$\mathcal{D}_{KL}(q(z_1, z_2|x)||p(z_1, z_2)) = \mathbb{E}_{q(z_2|x)}[\mathcal{D}_{KL}\big(q(z_1|x)\big|\big|p(z_1, z_2)\big)] + \mathcal{D}_{KL}(q(z_2|x)||p(z_2)) \tag{5.21}$$

Given the Gaussian assumptions for both the prior and posterior distributions, both KL-divergence terms can be written in closed-form. Furthermore, the latent vector is projected to the input space generating $(s_1, s_2, \dots, s_n)$. The decoder network

for the proposed MrA^2VAT generates the output probability that generates the reconstruction result via LSTM-based language model. This improves the performance of the generated paragraphs. Also, to deal with the zero KL term, the loss function is further modified adding the effect of KL annealing algorithm [203].

$$KL_w = \frac{1}{1+e^{-kn+b}} \tag{5.22}$$

where, k and b are the coefficients and n is the training step.

$$\mathcal{L}_{ml-vae} = \mathbb{E}_{q\phi(z_1|x)}[logp(x|z_1)] - KL_w * \mathcal{D}_{KL}(q(z_1, z_2|x)||p(z_1, z_2)) \tag{5.23}$$

5.1.1.2.3 Attention-based Dual Bi-LSTM Language Discriminator ($\boldsymbol{AdBL-LD}$)

The language discriminator encodes each input sentence by utilizing an attention-based dual Bi-LSTM module ($AdBL$). The first layer of $AdBL$ is the Bi-LSTM that learns the mappings from the input to the Bi-LSTM and the hidden state. The output generated is utilized by the attention layer that uses feature-based attention mechanism as:

$$e_t^i = V^T \tanh\left(Wh_t + UK_i + b\right) \tag{5.24}$$

$$att_t^i = \frac{\exp\left(e_t^i\right)}{\sum_{j=1}^{N} \exp\left(e_t^j\right)} \tag{5.25}$$

After the attention mechanism, another Bi-LSTM layer is utilized followed by a fully connected layer and sigmoid activation function. Let the input sentence to the language discriminator be $\mathbb{S}$, such that $\mathcal{LD}(\mathbb{S})$ is the output, where $\mathcal{LD}(\mathbb{S})$ is the score such that $\mathcal{LD}(\mathbb{S}) \in [0,1]$. The value of 1 indicates the grammatically most accurate and diverse sentence.

In order to improve the diversity and reduce the redundancy issues in the generated image descriptions, the proposed work calculates a dissimilarity score by

calculating the Word Movers Distance [204] score. Let $\not{f}c_i$ be the i^{th} sentence of the final caption for time t. Therefore, the dissimilarity score Υ for sentence s with respect to final caption $\not{f}c$ is

$$\Upsilon(s, \not{f}c) = \frac{\sum_{i=1}^{t} WMD(s, \not{f}c_i)}{t}, t > 0 \tag{5.26}$$

With this, we also exploit the concept of length penalty to the short sentences so that they do not get selected for inclusion in the final paragraph generation. For reference sentence $\mathscr{r}$ with length $|\mathscr{r}|$ and median length ϑ, to neglect the short-truncated sentences from being selected in the final paragraph, we select minimum medium length ϑ_{min} to calculate the length penalty which is given by:

$$LP(s, \mathscr{r}) = \left(1, \frac{|\mathscr{r}|}{\max(\vartheta_{min}, \vartheta(\mathscr{r}))}\right) \tag{5.27}$$

The first caption of the final paragraph is the sentence with maximum language score from the reference paragraphs (ground truth and the paragraphs generated from MrA^2VAT). For subsequent sentences, similarity of reference sentences with the final caption at that instant of time is calculated which helps in selection of finest sentence. Mathematically:

$$fs = arg\max_{s} \mathcal{LD}(\mathbb{S}) \tag{5.28}$$

$$fs = arg\max_{s} \mathcal{LD}(\mathbb{S}) + \Upsilon(s, \not{f}c) * LP(s, \mathscr{r}) \tag{5.29}$$

Eq. (5.28) calculates the first sentence for the final paragraph followed by selection of more sentences using Eqn. (5.29).

5.2 Experimental Work and Results

To validate the effectiveness of the proposed framework, extensive experiments are conducted on Stanford Paragraph Dataset [65]. This dataset contains 19,561

images with one human-generated paragraph for each image. The images are split into training (14575), validation (2487), and test set (2489).

5.2.1 Implementation Details

For $pos - CWE$ embeddings, we set the embedding dimension of 300 with dropout of 0.1. For training the embedding model, Adam [193] optimization is preferred with a learning rate of 0.001. For extraction of image features pretrained Faster R-CNN [15] is utilized to detect $\mathcal{N}$ objects. Further, the encoder and decoder structures of the proposed MrA^2VAT is a stack of 6 identical transformer layers with hidden dimension of 256. The dimension of the latent variable is set to 16. Again, Adam optimizer is utilized for training with 0.9 and 0.999 as the momentum parameters. The model is trained for 75 epochs with learning rate of 0.00005 and $\epsilon = 10^8$. We insert batch normalization [205] layer in the middle of every two adjacent transformer layers. Language discriminator encodes each input sentence with Bi-LSTM having hidden dimension of 512. Further, it is trained for 30 epochs with binary cross entropy loss. To evaluate the performance of the proposed framework with or without language discriminator, we evaluate BLEU [151], METEOR [26], and CIDEr [25] metrics.

5.2.2 Results for $pos - CWE$ Word Embedding

The proposed $pos - CWE$ word embedding t-SNE plot is presented in Fig. 5.5 The proposed embedding is able to create word embeddings capable of deriving several kinds of analogies by computing the similarity. Each datapoint in the t-SNE plot represents a word. Also, the plot shows that comparable words cluster together without prior knowledge, hence demonstrating that our word embedding preserves meaning or semantics. Also, Table 5.1 presents a similarity score generated between different

synonyms and antonyms from the proposed word embedding thereby depicting the different word analogies.

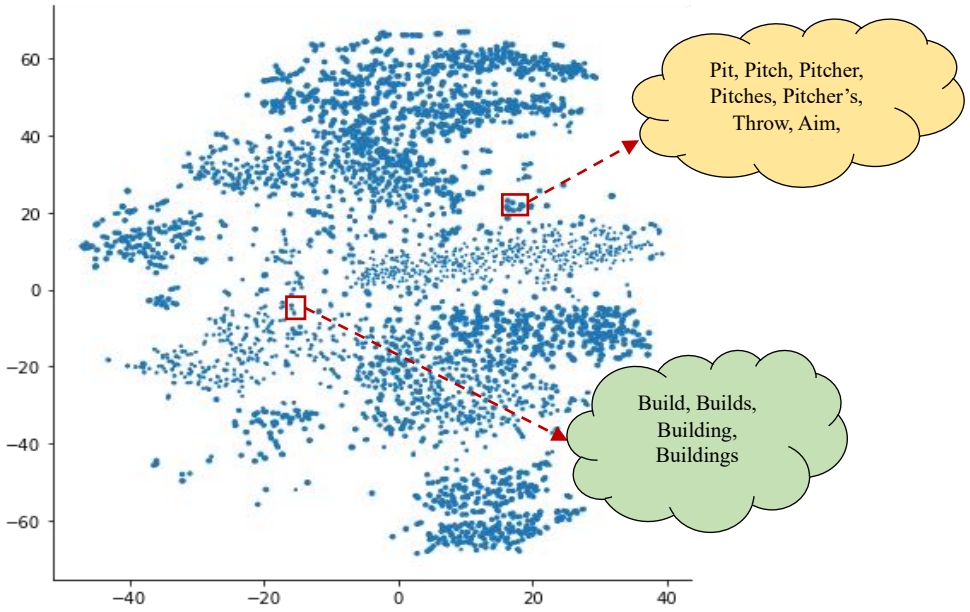


Fig. 5.5: t-SNE plot for the proposed $pos - CWE$

Table 5.1: Similarity Scores of word embeddings for synonyms and antonyms

Synonyms			Antonyms		
Word-1	**Word-2**	**Similarity**	**Word-1**	**Word-2**	**Similarity**
new	recent	0.946	back	front	-0.247
leash	chain	0.903	right	left	-0.073
picture	image	0.999	hazy	clear	-0.289
stadium	ground	0.968	small	big	-0.157

5.2.3 Attention Visualization Results

To visualize the multi-resolution multi-head attention attended features, high-level feature maps are generated as presented in Fig. 5.6. These maps give the best explanation of extracted features by focusing on most of the relevant parts of the object in the image. Thereby, the visualization shows that the proposed attention model generates maps that represent the fine-grained, medium-grained, and coarse-grained

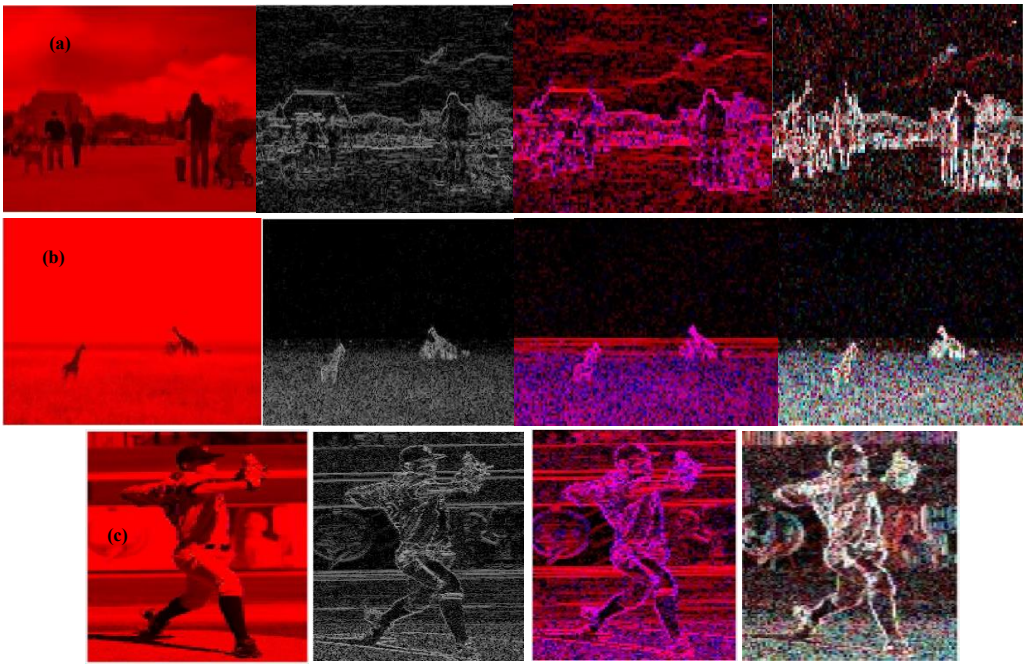


Fig. 5.6: Visualization of Multi-Resolution Multi-Head Attention to capture the potential relations between query representation and clues of different attention granularities.

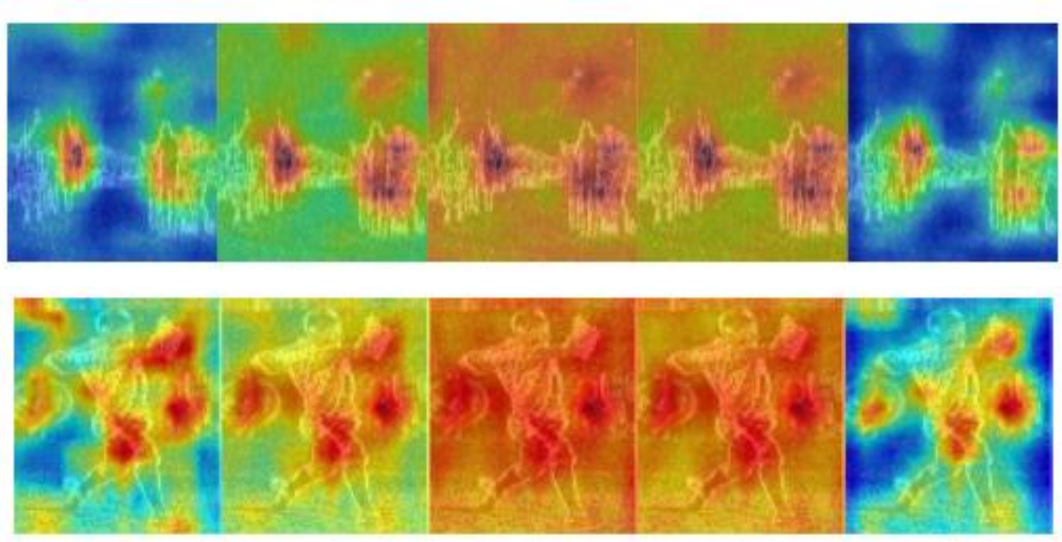


Fig. 5.7: Visualization of image attention maps generated from the proposed adaptive attention for extraction of more discriminant image features.

features well. Furthermore, Fig. 5.7 depicts the attention maps generated by leveraging the proposed adaptive attention module. The attention visualization in Figs. 5.6 and 5.7 shows

that the proposed attentions can gradually filter out noises and pinpoint the regions that are of great importance.

5.2.4 Quantitative Results

Table 5.2 reports the quantitative results for the proposed MrA^2VAT without and with the language discriminator. The results are compared with the different state-of-the-art methods for generation of paragraph on the Stanford Paragraph Dataset. Earlier models [65] [206] [207] reported minimum values of scores for all the parameters. Further, an improvement in the scores was observed in the CRL [208], DHPV [209], DAM [66], VRD [210], Para-CNN [211], S2DT [212], CAVP [68], and Dual-CNN [69]. Also, few methods [74] [213] [214] [215] [216] utilized spatial and semantic relationship concept to provide further improvement. Further, our framework utilizes a combination of visual and spatial features extracted from E-Faster R-CNN which are further interacted with the proposed multi-level VAT.

In comparison to these methods the proposed framework utilizes M^2A and Adaptive attention driven multi-level VAT for generation of coherent and meaningful paragraphs. To further increase the diversity and reduce the redundancy in the generated paragraphs, the proposed framework utilizes a language discriminator. This also provides further enhancement in terms of all evaluation parameters particularly for METEOR.

5.2.5 Qualitative Results

Table 5.3 presents the qualitative results obtained for the proposed framework, with or without $AdBL - LD$. From these generated paragraphs, it is evident that the proposed MrA^{2}VAT generated paragraph with no redundant sentences. Also, we can

Table 5.2: Comparison of the proposed MrA^2VAT with and without $AdBL - LD$ with other state-of-the-art on Stanford Paragraph Dataset

Model	B-1	B-2	B-3	B-4	M	C
Regions-Hierarchical [65]	41.90	24.11	14.23	8.69	15.95	13.52
RTT-GAN [206]	42.06	25.35	14.92	9.21	18.39	20.36
TOMS [207]	43.10	25.80	14.30	8.40	18.60	20.80
CAPG-VAE [215]	42.38	25.52	15.15	9.43	18.62	20.93
CAE-LSTM [216]	-	-	-	9.67	18.82	25.15
SCST [67]	43.54	27.44	17.33	10.58	17.86	30.63
CRL [208]	43.12	27.03	16.72	9.95	17.42	31.47
DHPV [209]	43.35	26.73	16.92	10.99	17.02	22.47
DAM [66]	35.00	20.20	11.70	6.60	13.90	17.30
VRD [210]	41.74	24.94	14.94	9.34	17.32	14.55
Para-CNN [211]	43.30	25.80	15.60	9.50	17.20	20.60
CAVP [68]	42.01	25.86	15.33	9.26	16.83	21.10
Dual-CNN [69]	41.60	24.40	14.30	8.60	15.60	17.40
S2DT [212]	44.47	27.38	16.87	10.17	17.64	24.33
OR-ATT [136]	44.55	28.54	18.19	11.18	17.97	33.12
DualRel [213]	45.30	28.91	18.46	11.30	17.86	34.02
PaG-MEG-SCST [214]	46.96	29.57	18.61	11.51	18.24	29.43
$\boldsymbol{MrA^2VAT}$**-w/o-** $\boldsymbol{AdBL - LD}$	**48.39**	**31.24**	**20.17**	**12.10**	**19.38**	**36.02**
$\boldsymbol{MrA^2VAT}$**-w-** $\boldsymbol{AdBL - LD}$	**51.16**	**32.80**	**21.63**	**12.99**	**23.26**	**36.53**

infer that the proposed model is able to generate descriptions by including more minute details from the images for example in the first case the proposed model with language discriminator predicted *"There are also tress in the background".* Also, the proposed model is able to concatenate two sentences from the ground truth to generate a more meaningful sentence for the paragraph by preserving the semantics. This can be observed from Case IV where the model generated "The front of the train is black with white letters on the front." From the ground truth *"The train is olive green with a black front and The letters on the front are white".* It is evident from the qualitative results that the proposed framework generated superior results with no redundant sentences and more diversity. Also, the model is able to detect and describe more minute details from images.

Table 5.3: Paragraphs Generated by using the proposed framework with and without Language Discriminator *(red highlighted text represents the redundant captions being generated, blue highlighted text generated more accurate description)*

Input Image	Ground Truth	MrA^2VAT-w/o- $AdBL-LD$	MrA^2VAT-w- $AdBL-LD$
	There are people gathered walking on the snow. Some people are flying kites. There's a man and a child and a stroller beside the man. The stroller is grey and black. There is a man to the left that is walking his dog. The dog is on a leash. There's a building in the background. The sky is partly cloudy. One of the kites flying is yellow.	A man in is walking on a snow with a child, stroller, and a child. The stroller is gray and black. People are flying kites in the sky. The sky is partly cloudy. People are flying kite in the sky. A man is walking with a dog. The man is walking with a dog is tied with leash. There is a brown building at the background with more people.	A man wearing black jacket is walking with a child and grey and black stroller on the snow ground. Another man is walking with a brown dog. The dog is tied with a leash. Some people are gathered near a brown building. Some other people are flying yellow kite in partly cloudy sky. There are also tress in the background.
	Giraffes are standing in a field. There is one giraffe standing by himself facing the right and two more giraffes standing together further away looking towards the left. The two giraffes are standing next to a small dry tree. The field that the giraffes are in has dry yellow grass. The sky is hazy and gray	The sky is gray. The sky is hazy and giraffes are standing on field. The sky is gray. The field is dry. One giraffe is standing alone. Two giraffes are standing looking towards left and one giraffe standing alone. Two giraffes are standing near trees. One giraffe is looking in right.	The sky is gray and hazy and giraffes are standing is a dry field. The field has dry yellow grass all around. One giraffe is standing alone and facing towards the right. Two giraffes are standing at the back near small dry tress. The two giraffes are looking towards left on dry filed.
	An intense young man is throwing the ball from a pitcher's mound. He is a pitcher in a baseball game that is being played in a stadium. The outer edges of the stadium have billboards advertising products such as G2 on them. The pitcher has a uniform on that is grey and blue. His cap is black with a fork shaped symbol on the front of it. A brown leather glove is being worn on the man's right hand. The stadium is filled with well-manicured green grass and brown dirt that forms the area for the players to run on.	A young man in gray and blue unifrom is throwing ball from a pitcher's mould. They are playing baseball match. The young man is throwing ball from pitcher's mould. The stadium outer edge shows G2 product advertisement. The man is wearing gray uniform with black hat. His right hand has brown leather gloves. The stadium is covered with green grass. The stadium is covered with green grass.	A baseball match is being played in a stadium. A powerful young man in gray and blue uniform is throwing the ball from pitcher's mound. The pitcher's cap is black with fork symbol on its front. He is wearing brown leather glove in his right hand. The stadium is covered with well-manicured green grass and brown pitch. The brown pitch in stadium helps the players to run on. The bill boards in outer edges of the stadium have advertising products G2.

	This is a picture of a train yard. The train is stopped in the yard. The train is olive green with a black front. The train has a large white headlight on it. The letters on the front are white. There are many wires above the yard. There are people looking at the train. There is a boy taking a picture of the train. The boy has a small shoulder bag.	This is a picture of train yard. The olive train is in the picture. The olive train is stopped in the yard with white headlight. The front of train is black. The train has white color written on front. People are looking and taking pictures of the train. A boy is standing with shoulder bad. People are taking pictures of the train. There are many wires on the yard.	The picture is of a train yard. The olive colored train with white headlight on it is standing in the yard. The front of the train is black with white letters on the front. People are looking and taking pictures of the train. A boy in blue is taking picture of the train. The boy has a shoulder bag. Two men are standing in front of wooden yard. There are many wires above the train yard.

5.2.6 Ablation Study

(1) To study the impact of length penalty on the language discriminator, we vary minimum median length (ϑ) and the minimum threshold score (ranging between 0.5 to 4.5) for selection of final sentence that is to be included in final paragraph. Fig. 5.8 shows the variation of length penalty and length of sentences generated for median lengths 3, 4, 5, and 6 respectively. Also, to effectively measure the performance of the language discriminator we first used only the METEOR scores as this score performs stemming and synonym matching. Fig. 5.9 presents the METEOR scores for five different runs. From Fig. 5.9, we can infer that for a score of 2.5 the proposed MrA^2VAT-w-LD provides better results with respect to METEOR score.

(2) Also, an ablation is conducted to study the influence of feature extraction module, paragraph generation module and baseline transformer for paragraph generation with and without the proposed language discriminator ($AdBL - LD$). Table 5.4 reports the results for the study conducted. We started our experiment with models M-1 and M-2 that utilized Faster R-CNN for image feature extraction and baseline transformer for generation of descriptive paragraphs. From the scores reported in Table 5.4, it is inferred that there is an improvement with respect to all the scores when we generated

paragraphs with language discriminator. Further, for M-3 we utilized the proposed framework $MrA^2VAT - wo - AdBL$. With the utilization of VAT, we are able to achieve significant rise in scores for all parameters. For, M-4 we utilized the proposed E-Faster R-CNN with $MrA^2VAT - wo - AdBL$ which helps in extraction of more minute details from images, thereby improving the generation of paragraphs.

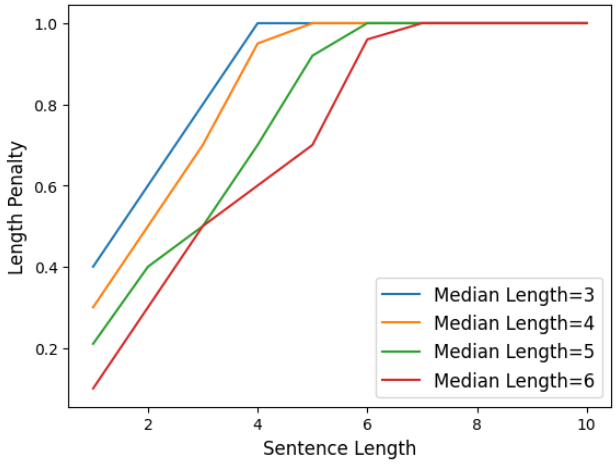


Fig. 5.8: Length penalty given for different values of ϑ

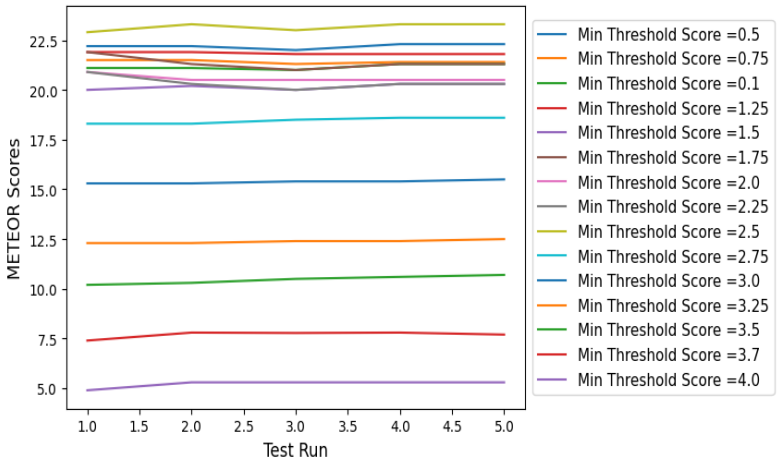


Fig. 5.9: Variation of METEOR scores on different test run to study the influence of Minimum Threshold score

We extend our experiment; by examining the effect of language score, dissimilarity score and length penalty on the proposed language discriminator with E-Faster R-CNN and the proposed VAT. Model, M-5 takes into account the influence of language scores whereas M-6 utilizes language scores and dissimilarity scores. For M-5 and M-6 a significant improvement is observed with respect to METEOR scores only as METEOR performs stemming and synonym matching. Further, to enhance the generated paragraph and the scores, we utilized the concept of length penalty with language and dissimilarity scores in M-7 and hence we obtain significant improvements in terms of all the parameters.

Faster R-CNN +Baseline Transformer-w- $AdBL - LD$

A man is walking on snow ground with a stroller and his child. There is a brown dog moving. Another man is moving with dog. There is a building in the background. Kites are in the sky. The sky is cloudy.

Faster R-CNN + Baseline Transformer w/o- $AdBL - LD$

A man is walking on snow with his child. A man is walking with a stroller and his child. There is a brown dog moving. A brown dog is moving in the field. There is a building at back. The sky is cloudy. The sky is cloudy.

E-Faster R-CNN +MrA^2VAT**-wo-**$AdBL - LD$

A man in is walking on a snow with a child, stroller, and a child. The stroller is gray and black. People are flying kites in the sky. The sky is partly cloudy. People are flying kite in the sky. A man is walking with a dog. The man is walking with a dog is tied with leash. There is a brown building at the background with more people.

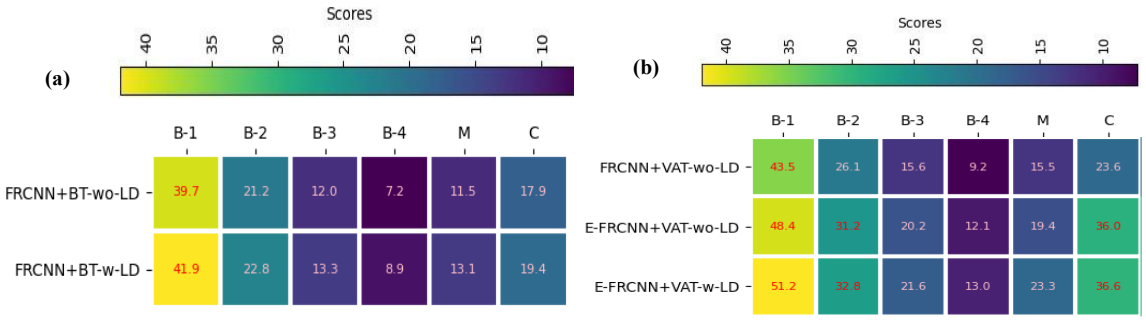
E-Faster R-CNN +MrA^2VAT**-w-**$AdBL - LD$

A man wearing black jacket is walking with a child and grey and black stroller on the snow ground. Another man is walking with a brown dog. The dog is tied with a leash. Some people are gathered near a brown building. Some other people are flying yellow kite in partly cloudy sky. There are also tress in the background.

Fig. 5.10: Qualitative Results obtained for different Ablated Models (red highlighted text represents the redundant captions being generated, blue highlighted text generated more accurate description).

Figs. 5.10 and 5.11 provide the qualitative results and heat maps for the ablation study conducted. We present the paragraphs generated by M-1, M-2, M-4, and M-7 respectively. From these results, we can easily infer that the proposed framework i.e., MrA^2VAT-w-$AdBL - LD$ decreases the redundancy and generated diverse and coherent paragraph-based descriptions of images. Also, heat maps are plotted to

visualize the influence of the proposed MrA^2VAT framework with the baseline transformer.



FRCNN: Faster R-CNN, E-FRCNN: Efficient Faster R-CNN

BT: Baseline Transformer, VAT: Variational Autoencoder Transformer, LD: Language Discriminator

Fig. 5.11: Heat Maps obtained for different Ablated Models to study the influence of (a) baseline transformer and the (b) proposed framework.

Table 5.4: Ablation Study Results for Proposed Framework to study the influence of Baseline Transformer and the Variational Autoencoder Transformer Module with and without Language Discriminator

Model	Feature Extraction	Paragraph Generation	Language Discriminator			B-1	B-2	B-3	B-4	M	C
			Language Score	Dissimilarity Score	Length Penalty						
M-1	**Faster R-CNN**	**Baseline Transformer**	✖	✖	✖	39.71	21.20	12.00	7.24	11.51	17.90
M-2	**Faster R-CNN**	**Baseline Transformer**	✓	✓	✓	41.90	22.78	13.32	8.92	13.08	19.35
M-3	**Faster R-CNN**	MrA^2VAT	✖	✖	✖	43.54	26.10	15.63	9.21	15.54	23.59
M-4	**E-Faster R-CNN**	MrA^2VAT	✖	✖	✖	48.39	31.24	20.17	12.10	19.38	36.02
M-5	**E-Faster R-CNN**	MrA^2VAT	✓	✖	✖	48.38	31.26	20.17	12.19	19.90	36.19
M-6	**E-Faster R-CNN**	MrA^2VAT	✓	✓	✖	48.38	31.26	20.17	12.34	20.88	36.25
M-7	**E-Faster R-CNN**	MrA^2VAT	✓	✓	✓	**51.16**	**32.80**	**21.63**	**12.99**	**23.26**	**36.53**

Table 5.5: A comparison of unique words generated for the proposed framework with and without language discriminator

Model	words per para	Total unique words	Unique words per para
$\boldsymbol{MrA^2VAT}$**-wo-**$\boldsymbol{AdBL-LD}$	112.03	1988	50.20
$\boldsymbol{MrA^2VAT}$**-w-** $\boldsymbol{AdBL-LD}$	130.52	2983	61.43

(3) To examine the diversity in generated paragraphs, total number of unique words which are used to generate the final paragraph is calculated. Table 5.5 reports number of unique words per paragraph with and without language discriminator. This illustrates the diverse nature of the model in terms of paragraph generated. The proposed model with LD utilizes more unique words per paragraph for the generation of paragraphs in comparison to MrA^2VAT-w/o-LD.

5.3 Significant Outcomes

This chapter presents multi-resolution multi-head attention and adaptive attention driven variational autoencoder-based transformer framework MrA^2VAT. The proposed framework generates diverse, coherent and meaningful paragraph descriptions by exploiting an attention-based dual Bi-LSTM language discriminator. Also, with KL-term vanishing employed in the proposed multi-level VAT, we are able to achieve an overall improvement in terms of all evaluation parameters when compared with other baseline state-of-the-art.

Further, by leveraging the concept of language score, dissimilarity scores, and length penalty on the proposed language discriminator, we are able to generate a more diverse paragraph with no redundant sentences. Furthermore, a significant rise in terms of BLEU-1 (around 5.7%) and METEOR scores (around 27.5%) is observed for the

proposed framework. The Chapter also presents a novel positional embedding ($pos - CWE$) that utilizes absolute and relative position information from embedding to enhance the relationship of each word generated.

Chapter-6

Summarization Caption Generation using Factual and Stylized Captioning Tasks for Refined Image Captioning

The objective of this chapter is to generate refined summarized image captions by combining Factual Image Captioning (FIC) and Stylized Image Captioning techniques. The key components of this chapter include the Unified and Multi-head Attention-based Caption Summarization Transformer $UnMHA-ST$ to capture the intra- and intermodal interactions of multimodal information and generate refined attended representations. The proposed caption summarization framework is supported by experimental validation, results discussions, and analysis of results with similar state-of-the-art and ablation studies.

6.1 UnMA-CapSumT: Unified and Multi-Head Attention-driven Caption Summarization Transformer

With the advancements in deep-learning technology, different models for factual [125] [166] and stylized image captioning [54] [53] are developed. These works generate separate sentences for factual and stylized (romantic and humorous) descriptions of an image. Therefore, these methods sometimes may give erroneous descriptions by either describing the factual content in dull language or may lead to the incorrect representation of the style. This may be due to poor learning knowledge of factual content and its associated linguistic styles. To the best of my knowledge, there are no such works in literature that provide single-sentence summarized descriptions incorporating factual, romantic, and humorous contents. To address the

abovementioned issues and to improve the learning of knowledge of factual content and its associated linguistic styles, this chapter extends the task of image captioning to abstractive text summarization. The task is to automatically summarize the source article into an accurate short version that reflects its central content comprehensively.

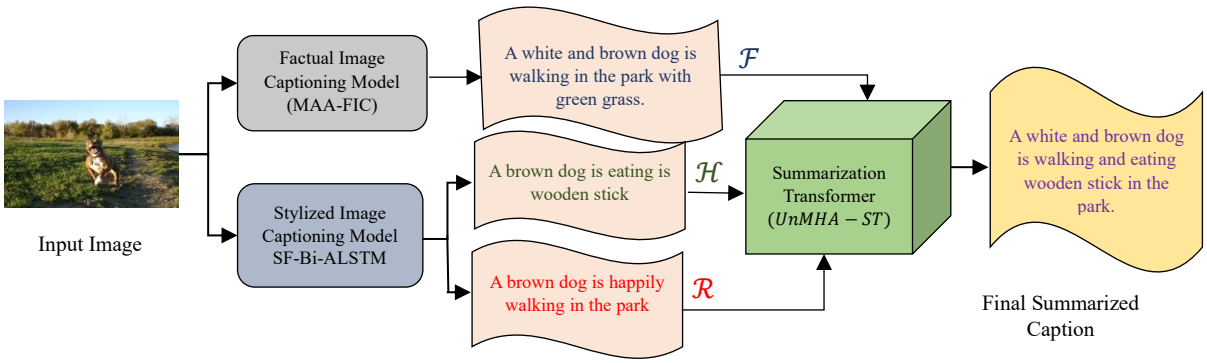


Fig. 6.1: Block Diagram Representation of the proposed $UnMA$ −CapSum Transformer based Captioning Framework

6.1.1 Proposed Methodology

This chapter presents a novel Unified Attention (Un-A) and Multi-Head Attention-driven Caption Summarization Transformer ($UnMA$-CapSumT) based Captioning Framework. An overview of the proposed framework is depicted in Fig. 6.1, the framework is twofold: (i) it integrates Modified Adaptive Attention-based (MAA-FIC) factual image captioning and Style Factored Bi-LSTM with attention (SF-Bi-ALSTM) based stylized image captioning techniques for the generation of factual, stylized – {romantic and humorous} description of an image, and (ii) the proposed summarizer $UnMHA-ST$ combines both factual and stylized descriptions of input image to generate styled rich coherent summarized captions. The key highlights of this chapter are:

i. This chapter proposes a Unified and Multi-head Attention-based Caption

Summarization Transformer $UnMHA - ST$ to capture the intra- and intermodal interactions of multimodal information and generate refined attended representations. Also, it utilizes the concept of a pointer generator network and coverage mechanism to solve the problems of rare words which arise due to out-of-vocabulary (OOV) and repetition issues.

ii. The proposed framework integrates MAA-FIC and SF-Bi-ALSTM-based factual and stylized image captioning models for the generation of factual and stylized descriptions of images which are utilized by $UnMHA - ST$ for generation of styled rich coherent summarized captions.

iii. The chapter also presents an efficient Attention enabled fastText word embedding, fTA-WE, that integrates the attention mechanism into the Continuous Bag of words (CBOW) model of fastText that efficiently learns vector representation of words and enhances the performance of the proposed $UnMHA - ST$.

6.2 Modified Adaptive Attention-based Factual Image Captioning Model (MAA-FIC)

For the generation of factual image description, the proposed framework utilizes the MAA-FIC factual image captioning model. The diagram for the proposed MAA-FIC model is presented in Fig. 6.2. The input to the model is an image $\mathcal{I}$ and the output is the descriptive factual sentence $\mathcal{F}$ with $\mathcal{k}$ encoded words ($\mathcal{w}$): $\mathcal{F} = \{\mathcal{w}_1, \mathcal{w}_2, \ldots\ldots \mathcal{w}_{\mathcal{k}}\}$.

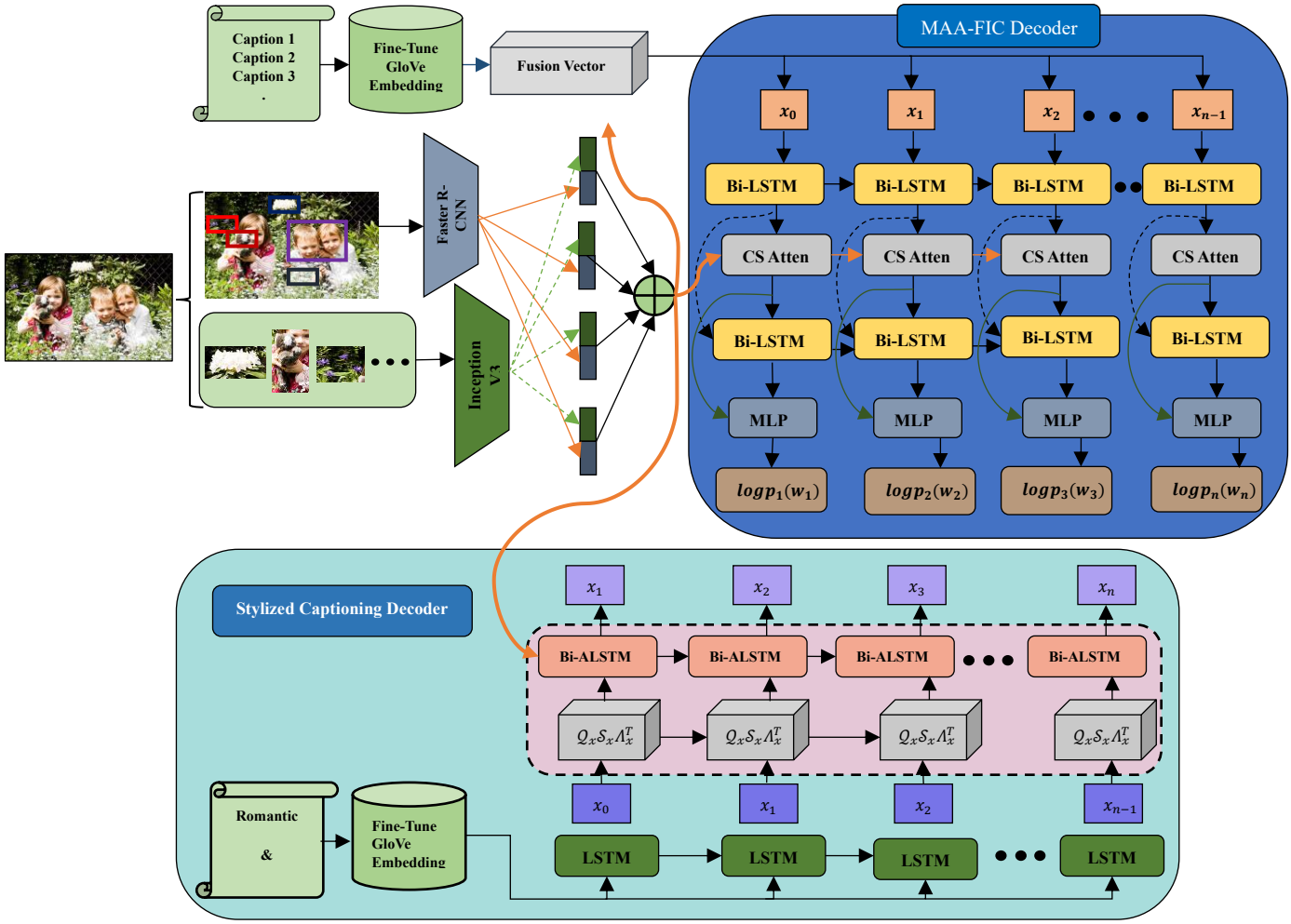


Fig.6.2: Proposed MAA-FIC Model and SF-Bi-ALSTM-based Stylized Image Captioning Model.

6.2.1 Feature Extraction

For object detection, the proposed MAA-FIC model incorporates Faster R-CNN for the detection of objects. After the detection of objects, each detected object is mapped to a feature vector. For each image $\mathcal{I}$, n-objects are detected, given by, $\{\Phi_1, \Phi_2, \ldots, \Phi_n\}$; $\Phi_i \in \mathbb{D}^d$ where, $\mathbb{D}^d$ is the $\mathbb{D}-$ dimensional vector of top n-boxes as the region of objects.

The localization of objects is performed to extract spatial relationships between objects. The n-objects detected using Faster R-CNN [115] for each image are fed to

Inception-V3 and output the feature vector that represents the spatial location of each object, represented as $\{\theta_1, \theta_2, \dots, \theta_n\}$; $\theta_i \in \mathbb{D}^{\mathbb{T}}$, where, $\mathbb{D}^{\mathbb{T}}$ is the $t - dimensional$ vector of the spatial location of each object. The features extracted from the Faster R-CNN module and Inception-V3 module are concatenated. Therefore, each concatenated feature vector consists of a detected object feature vector (Φ_i) and the feature vector obtained from the localization of objects (θ_i). Mathematically, the concatenated image feature vector is represented as:

$$\mathbb{F}_i = [\Phi_i; \theta_i], \mathbb{F}_i \in d^D, D = d + \mathbb{T} \tag{6.1}$$

6.2.2 Text Generation Network

This section discusses about the proposed Modified Adaptive based Attention (MAA) and Bi-LSTM-based factual language decoder module. The MAA utilizes both channel and spatial attention mechanisms with bidirectional sequence learning. This technique combines the weighted combination of all the encoded input vectors, with the most relevant vectors being attributed with the highest weights. For the model with attention mechanism:

$$a_t = \mathfrak{f}(\mathbb{F}, \hbar_t) \tag{6.2}$$

where $\mathfrak{f}$ is the attention mechanism and $\mathbb{F} = [\mathbb{F}_1, \mathbb{F}_2 \dots, \mathbb{F}_{\hbar}]$ represents the concatenated image features and $\hbar_t$ is the hidden state of the RNN at time t. Channel-wise attention distribution for $\hbar$ regions is defined as:

$$\tilde{\mathcal{C}} = \tilde{\mathcal{C}}_{\mathcal{W}} = \mathcal{W}_{hc}^T \tanh\left((\mathcal{W}_{c\mathbb{F}} \otimes \mathbb{F} + \beta_c) \oplus \left(\mathcal{W}_{\mathfrak{f}c}\hbar_t\right)\mathcal{I}^T\right) \tag{6.3}$$

$$\psi = softmax(\tilde{\mathcal{C}}) \tag{6.4}$$

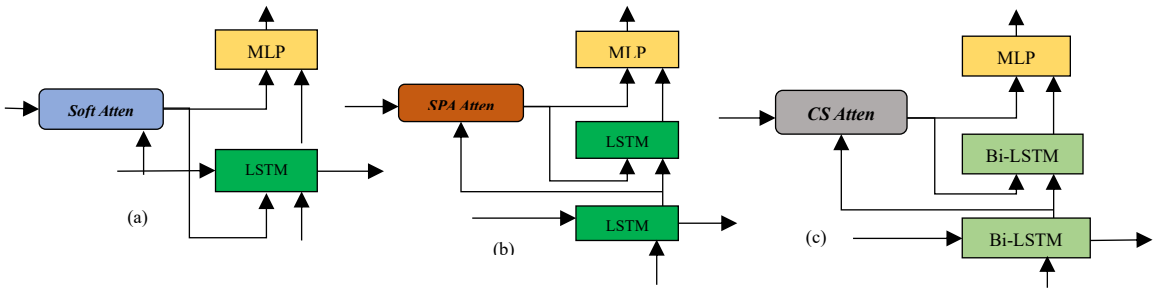


Fig. 6.3: (a) Soft-attention Mechanism, (b) Adaptive Attention Mechanism, and (c) Proposed Modified Adaptive Attention

Further, spatial attention weights of the $\hbar$ are regions mathematically expressed as:

$$\tilde{\mathcal{S}} = \tilde{\mathcal{S}}_{\mathcal{W}} = \mathcal{W}_{hs}^{T} \tanh\left(\left(\mathcal{W}_{s\mathbb{F}}\mathbb{F} + \beta_s\right) \oplus \left(\mathcal{W}_{\hbar s}\hbar_t\right)\mathcal{I}^{T}\right) \tag{6.5}$$

$$\varphi = softmax(\tilde{\mathcal{S}}) \tag{6.6}$$

where, β_s and β_c are bias terms. After attaining the channel and spatial attention weights, channel-spatial weights are thus combined to obtain a modulated feature map. This provides the modified attentional distributions for the $\hbar$ regions as:

$$\psi = \phi_c(h_t, \mathbb{F}) \tag{6.7}$$

$$\varphi = \phi_s(h_t, \hbar(\mathbb{F}, \psi)\,) \tag{6.8}$$

$$\tilde{\mathcal{X}} = \hbar(\mathbb{F}, \psi, \varphi) \tag{6.9}$$

Therefore, the final modified attention weights are defined mathematically as:

$$a_t = \sum_{i=1}^{\hbar}(\mathcal{X}_{it}\mathbb{F}_{it}) \tag{6.10}$$

Based on the channel and spatial attention mechanisms, modified adaptive attention is proposed with visual sentinel and Bi-LSTM layers. Fig. 6.3 presents the basic architecture of MAA. The visual sentinel helps the model to focus more on wither

image information or on the language rules. The memory unit in Bi-LSTM stores the information of both the previous partial visual information and the language rules. Therefore, visual sentinel $\mathcal{s}_t$ formula based on the Bi-LSTM memory unit is:

$$m_t = \theta(\mathcal{W}_x \mathcal{X}_t + \mathcal{W}_h h_{t-1}) \tag{6.11}$$

$$\mathcal{s}_t = m_t \odot \tanh(\tilde{h}_t) \tag{6.12}$$

where, $\mathcal{X}_t$ is the input of the Bi-LSTM, m_t is the sentinel gate, is the output of the last time node (h_{t-1}), and $\tilde{h}_t$ is the memory cell. Also, the modified adaptive attention (MAA) mechanism is obtained following the sentinel gate $\mathcal{s}_t$

$$\hat{a}_t = \delta_t \mathcal{s}_t + (1 - \delta_t) a_t \tag{6.13}$$

$\delta_t \in [0,1]$ and is the new sentinel gate at time t. The value of δ_t decides whether the model focuses on image information or language rules. If the value of δ_t is equal to zero, it denotes focusing on image information whereas the value of 1 indicates that the focus is on the language rules. The channel-spatial attention distribution of $\mathcal{k}$ regions obtained by splicing an element and is given by:

$$\widehat{\mathcal{X}} = softmax([\widetilde{\mathcal{X}}; \mathcal{W}_h^T \tanh(\mathcal{W}_s \mathcal{s}_t + \mathcal{W}_m h_t))]) \tag{6.14}$$

To generate the natural language description of an input image, fine-tuned GloVe text embeddings are employed and are represented as $\mathbb{T}$. The input to the text generation module is the weighted fused vector obtained after the fusion of image features $\mathbb{F}$ and text features $\mathbb{T}$.

$$\mathcal{Q}_t = \{\mathbb{F}_t \mathbb{T}_t\} \tag{6.15}$$

The fused feature vector $\mathcal{Q}_t$ is incorporated for the generation of accurate description of an image. Further, the resulting output generates a word on the next time node using the MLP function that incorporates backpropagation through a time algorithm to update the parameters of the LSTM network. The objective function defined for the optimization of the proposed model is given by:

$$\Delta = arg\max_{\vartheta} \sum_{r,y} \sum_{t=0}^{N} \log p(\mathcal{F}_t | r, \Delta, \mathcal{F}_1, \mathcal{F}_2 \ldots \mathcal{F}_{t-1}) \tag{6.16}$$

ϑ is the learnable parameter with m feature maps whose weight is $r \in \{r_1, r_2 \ldots r_t\}$. The cross-entropy loss is incorporated to minimize the loss function and to maximize the probability of each correct word appearing. The cross entropy-loss function is given by:

$$\mathcal{L}(\Delta) = -\sum_{t=1}^{N} \log\left(p_t(w_t^* | \mathbb{G}_{1:t-1}^*)\right) \tag{6.17}$$

where N represents total words in a sentence; Δ denotes all the parameters in the model; $\mathbb{G}_{1:t-1}^*$defines the ground truth.

6.3 SF-Bi-ALSTM Based Stylized Image Captioning

The proposed SF-Bi-ALSTM module as depicted in Fig. 6.2 serves as the building block for the generation of style-based descriptions of images. The input to the proposed style-based caption generation model is image $\mathcal{I}$ and the output is the descriptive romantic and humorous sentences $\mathcal{R}$ and $\mathcal{H}$ with n and m encoded words: $\mathcal{R} = \{w_1, w_2, \ldots\ldots w_n\}$, $\mathcal{H} = \{w_1, w_2, \ldots\ldots w_m\}$. The proposed stylized captioning model is based on encoder-decoder-based architecture. For the encoder, this model employs the same strategy as employed by the MAA-FIC model as discussed in *Section 6.2*, whereas the SF-Bi-ALSTM model is utilized for the generation of style-based captions.

6.3.1 SF-Bi-ALSTM Module

This section presents a new and modified variant of LSTM and Factored LSTM namely, Modified Style Factored Bi-LSTM. Traditional LSTM captures the long-term dependencies among words in sentences but fails to capture the styles in sentences. Also, Factored LSTM incorporates style factors into the visual caption generation model but fails to generate more attractive and coherent style-based descriptions. To overcome these limitations SF-Bi-ALSTM Module is designed that produces more meaningful stylized descriptions, combining LSTM with attention layers from both directions. Further, the proposed Bi-LSTM learns to refine the input vector from network hidden states and sequential context information.

Consider, $\mathcal{M}_x \in \mathbb{R}^{m \times n}$, then $\mathcal{Q}_x \in \mathbb{R}^{m \times r}, \mathcal{S}_x \in \mathbb{R}^{r \times r}$, and $\Lambda_x \in \mathbb{R}^{r \times n}$. $\mathcal{M}_x$ can be represented in the form of three matrices:

$$\mathcal{M}_x = \mathcal{Q}_x \mathcal{S}_x \Lambda_x^T \tag{6.18}$$

Further, the style-specific matrix $\{\mathcal{S}_x\}$ can be represented as:

$$\mathcal{S}_x = \frac{1}{\mathcal{N}} |\mathcal{S}\mathcal{S}^T| \tag{6.19}$$

The memory gates and cells in the traditional Bi-LSTM are modified with respect to $\mathcal{M}_x$ and attention mechanism:

$$\tilde{x}_t = sigmoid(\mathcal{Q}_{x_t} \mathcal{S}_{x_t} \Lambda_{x_t} x_t + W_h \tilde{h}_{t-1}) \odot x_t \tag{6.20}$$

$$\tilde{\imath}_t = sigmoid(\mathcal{Q}_{ix} \mathcal{S}_{ix} \Lambda_{ix} \tilde{x}_t + W_{ih} \tilde{h}_{t-1} + \tilde{b}_i) \tag{6.21}$$

$$\tilde{f}_t = sigmoid(\mathcal{Q}_{fx} \mathcal{S}_{fx} \Lambda_{fx} \tilde{x}_t + W_{fh} \tilde{h}_{t-1} + \tilde{b}_f) \tag{6.22}$$

$$\tilde{o}_t = sigmoid(\mathcal{Q}_{ox} \mathcal{S}_{ox} \Lambda_{ox} \tilde{x}_t + W_{oh} \tilde{h}_{t-1} + \tilde{b}_o) \tag{6.23}$$

$$\tilde{g}_t = sigmoid(\mathcal{Q}_{gx} \mathcal{S}_{gx} \Lambda_{gx} \tilde{x}_t + W_{gh} \tilde{h}_{t-1} + \tilde{b}_g) \tag{6.24}$$

$$\tilde{c}_t = \tilde{f}_t \odot c_{t-1} + \tilde{x}_t \odot \tilde{g}_t \tag{6.25}$$

$$h_t = \tilde{o}_t \odot tanh\tilde{c}_t \tag{6.26}$$

$$\mathscr{p}_{t+1} = (\mathbb{C}h_t) \tag{6.27}$$

where, x_t is the input and an input update gate is also defined as $\tilde{x}_t$ according to input and the hidden state h_{t-1}. Also, Λ_{ix}, $\Lambda_{fx}, \Lambda_{ox}$, and Λ_{gx} are the input weight matrices, and W_{ix}, W_{fx}, W_{ox}, and W_{gx} are the weight matrices which are applied to recurrently update the matrices of hidden states. Also $\{\mathcal{M}\}, \{\mathcal{Q}\}$, and $\{\Lambda\}$ are the matrices shared by two styles of romantic $\mathcal{S}_{\mathcal{R}}$ and humorous $\mathcal{S}_{\mathcal{H}}$ respectively. The style factored Bi-ALSTM model is implemented as:

$$h_t^f = \tilde{o}_t^f \odot tanh\tilde{c}_t^f \tag{6.28}$$

$$h_t^b = \tilde{o}_t^b \odot tanh\tilde{c}_t^b \tag{6.29}$$

$$y_t = W_{hy}^f h_t^f + W_{hy}^b h_t^b + b_y \tag{6.30}$$

6.3.2 Style-based Language Generation

To generate the style-based descriptions of images, the decoder of the proposed model leverages multi-task learning for sequence tagging [217]. This helps to learn to disentangle the style factors from the text corpus. The inputs to the SFA-Bi-LSTM module are the image feature matrix $\mathcal{I}_{\mathscr{f}}$ extracted from the image encoder and the text feature embedding matrix $\mathcal{T}_{\mathscr{f}}$ for romantic and humorous text corpus. The matrix $\mathcal{T}_{\mathscr{f}}$ is obtained via fine-tuned GloVe embedding. The extracted features are utilized by the SFA-Bi-ALSTM module according to eqn. (6.20) to (6.30). The output generated from the proposed Bi-LSTM is soft-attended and generates either romantic or humorous descriptions of images. Also, at time step $\mathscr{t}$, negative log-likelihood loss is incorporated to minimize the loss function and to maximize the probability of each word appearing for both romantic and humorous styles.

$$\mathcal{L}(\Theta) = -\sum_{i=1}^{n} \mathcal{s}_{ij} log \hat{\mathcal{s}}_{\Theta,ij} + (1 - \mathcal{s}_{ij}) \log (1 - \hat{\mathcal{s}}_{\Theta,ij}) \quad (6.31)$$

6.4 Caption Summarization Module

The image descriptions generated from the MAA-FIC and SF-Bi-ALSTM based stylized captioning model are collected and summarized to generate a caption that contains factual, romantic, and humorous content in one caption. This section discusses about the proposed novel transformer-based summarization model, UnMHA-ST with fTA-Word-Embeddings, to generate a summarized caption for an image.

6.4.1 fTA-Word-Embedding (fTA-WE)

To improve the performance and interpretability of the word embedding models, this work proposes a fTA (fastText with Attention) word embedding as depicted in Fig. 6.4. The proposed embedding module utilizes fastText [199] word embeddings with soft-attention [104] to learn vector representation of words. This helps to keep the words closer to each other which are semantically related. Also, fTA-WE leverage sub-words to enrich the vector space for rare and unseen words. The context vector c_v for vocabulary $\mathcal{V} = [v_1, \dots, v_{\mathcal{N}}]^T \in \mathbb{R}^{\mathcal{N} \times \mathcal{D}}$ of size $\mathcal{N}$ and word vector size $\mathcal{D}$ for fastText is represented by:

$$c_v = \sum_{i \in [-b,b]-\{0\}} u_{e_j} \quad (6.32)$$

where, $\mathcal{U} = [u_1, \dots, u_{\mathcal{N}}] \in \mathbb{R}^{\mathcal{N} \times \mathcal{D}}$ represents the sub-words and is used in the calculations of context vectors to efficiently learn a representation for each word. Taking the word $'white'$ as an example the sub words formed for $n = 3$ $grams$ are $< wh, whi, hit, \dots >$. Further, e_j and b represent the index of each word and the size

of the context window respectively. If w_0 is the masked word index with vector v_{w_0}, the probability of w_0 to occur in the context of $\{w_{-b}, \dots, w_{-1}, w_1, \dots, w_b\}$

$$p(w_0|W_{[-b,b]-\{0\}}) \propto exp v_{w_0}^T c_v \tag{6.33}$$

The embedding of each word is represented by the embeddings of all sub-words which can be expressed mathematically as:

$$\tilde{u}_w = \sum_{e_j \in \mathbb{S}_w} u_{e_j} \tag{6.34}$$

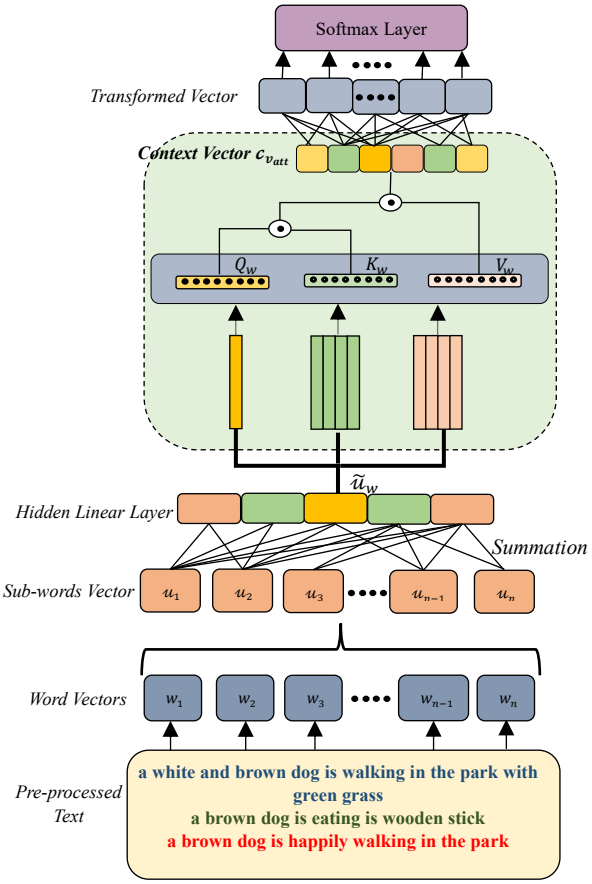


Fig. 6.4: Proposed fTA-WE

The attention mechanism of the fTA-WE embeddings, utilizes key matrix $\mathcal{K}$, value V, and query matrix Q. The attention weight and the context vector with attention is:

$$a_{w_i} = \exp\left(v_{w_0}^T\, c_v\right) \tag{6.35}$$

$$c_{v_{att}} = \sum_{i\in[-b,b]-\{0\}} a_{w_i} \left(\sum_{e_j\in\mathbb{S}_w} \tilde{u}_{e_j}\right) \tag{6.36}$$

The probability of the masked word is given by:

$$p(w_0|W_{[-b,b]-\{0\}}) \propto \exp\left(\tilde{u}_{w_0}^T c_{v_{att}}\right) \tag{6.37}$$

The attended vector is linearly transformed to obtain the transformed vector representation of the input text. Furthermore, the obtained transformed vector representation is used as an input to the text summarization transformer to generate the summarized captions for the images.

6.4.2 $UnMHA-ST$ Text Summarization Transformer

The proposed $UnMHA-ST$ Text Summarization Transformer takes input from two captioning mechanisms discussed in *Sections 6.2 and 6.3* respectively. It generates a summarized caption of an image that reflects the factual, romantic, and humorous contents in a single caption. The proposed $UnMHA-ST$ as depicted in Fig. 6.5, utilizes a unified attention mechanism in addition to multi-head attention. This helps in concurrently capturing the intra-modal and intermodal interactions of multimodal information which further helps in the generation of their related attended representations. With a unified attention module, the proposed $UnMHA-ST$ also leverages a pointer-generator network and coverage mechanism to solve the problems of rare words which arise due to out-of-vocabulary (OOV) and repetition issues. The input to the transformer network is the text embedding that is generated from the fTA-WE model. These embeddings are added with the positional embedding and are further split into key k, values v, and query q, respectively. To jointly attend information from

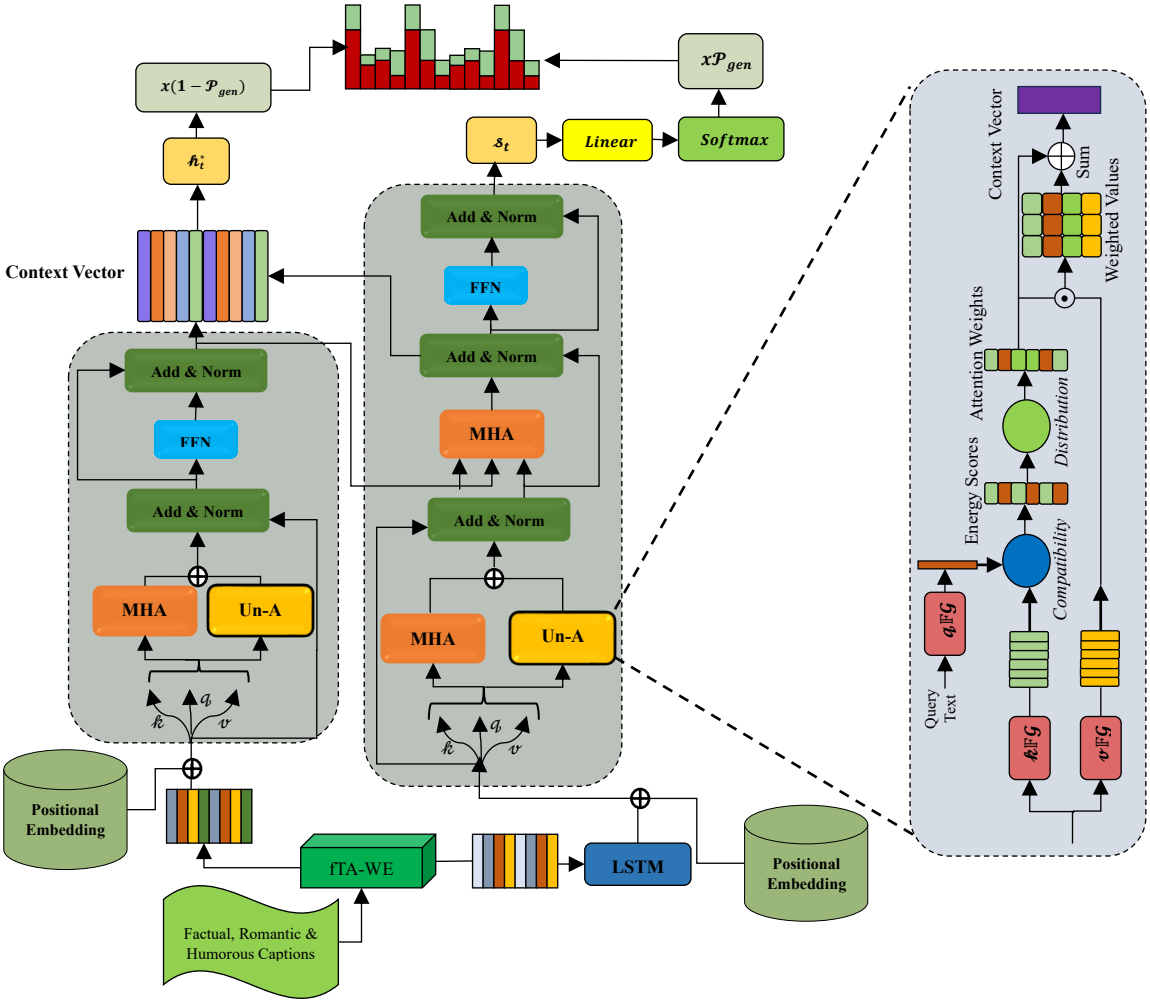


Fig. 6.5: Proposed $UnMHA-ST$ Model for Summarized Caption Generation

different representation subspaces at different positions, MHA is employed which is mathematically represented as:

$$Attention(k, q, v) = softmax(\frac{qk^T}{\sqrt{d_k}})v \tag{6.38}$$

$$MHA(k, q, v) = Concat(h_1, h_2, \dots, h_i)W^0 \tag{6.39}$$

where, $h_j = Attention(kW_j^k, qW_j^q, vW_j^v)$ (6.40)

The unified attention (Un-A) module as depicted in Fig. 6.5 utilizes an iterative alternating attention mechanism [218] that allows a fine-grained exploration of both the query and the document. Un-A does not collapse the query into a single vector instead it evaluates keys and query by using a compatibility function $e = \mathbb{F}(q, \hbar)$. The function $\mathbb{F}$ is the energy function. The energy scores thus obtained are transferred into the attention weights using a distribution function $\mathcal{G}$, $\mathcal{E} = \mathcal{G}(e)$. Therefore, the vector obtained from the compatibility and distribution function is combined with the transformer values v, which are merged to represent the context vector in a more compact form.

$$\mathcal{Z}_i = \mathcal{E}_i \mathcal{v}_i \tag{6.41}$$

$$C = \sum_{i=1}^{d_k} \mathcal{Z}_i \tag{6.42}$$

The vectors obtained from both the attention mechanisms are further added and passed through the subsequent section of the encoder layer FFN.

$$\Omega_t = f(MHA(\hbar, q, \mathcal{v}), C) \tag{6.43}$$

Also, the decoder structure follows the traditional transformer structure [154] with a unified attention mechanism. The embeddings input to the decoder network are shifted right version of the embeddings at the encoder. These embeddings are made to pass through the multi-head attention network and the unified attention network. The dumped results are then fed through a series of FC layers and finally a linear activation function followed by a softmax layer, generates the output probabilities. In addition to this, the whole model incorporates a pointer generator network [219] and a coverage mechanism [219]. The context vector is obtained from the encoder output and the

output state of the hidden layer inside the decoder and predicts new words for the dictionary. The pointer network's attention provides information about the most significant words at any given time, which is helpful for prediction. The current moment's attention distribution and the sum of weights for the encoder's hidden layer are represented as:

$$b^t = softmax(\Omega_t) \tag{6.44}$$

$$n_t = \sum_i b^t \, \mathrm{h}_i \tag{6.45}$$

Further, vocabulary probability distribution and the probability of distribution of words is mathematically expressed as:

$$\mathcal{P}_{vocab} = softmax(t'(t[\Omega_t, n_t] + u) + u') \tag{6.46}$$

where, t, t', u, u' are the learnable parameters. Also, the probability distribution of words is:

$$\mathcal{P}(\mathcal{W}) = \mathcal{P}_{vocab}(\mathcal{W}) \tag{6.47}$$

Hence, the model, through a pointer network, must either directly copy words from the source or create new terms to address the out-of-vocabulary issues. Therefore, the pointer-generator network is employed with probability to copy a word from the source text or to generate the word:

$$\mathcal{P}_{gen} = sigmoid\left(w_h^T h_t^* + w_n^T n_t + w_x^T x_t + r_{ptr}\right) \tag{6.48}$$

where, w_n^T, w_h^T, w_x^T, and r_{ptr} are learnable parameters and $\mathcal{P}_{gen} \epsilon [0,1]$.

Further, $\mathcal{P}_{gen}$ acts as a switch that either generates a word from the vocabulary using sampling from $\mathcal{P}_{vocab}$, or copies a word from the input sequence by sampling from the

attention distribution. Also, there exists an extended vocabulary for each document which is union of the vocabulary, and all words appearing in the source document.

$$\mathcal{P}(\mathcal{W}) = \mathcal{P}_{gen}\mathcal{P}_{vocab}(\mathcal{W}) + (1 - \mathcal{P}_{gen}) \sum_{j:w_j} a_{ij} \tag{6.49}$$

Note, if w is an OOV word, then $\mathcal{P}_{vocab}(\mathcal{W}) = 0$. Also, if w does not appear in the source document, then $\sum_{j:w_j} a_{ij} = 0$

For time-step i, negative log-likelihood loss is evaluated for the target word w_i and is given by:

$$\mathcal{L}_i = -log\mathcal{P}(\mathcal{W}_i) \tag{6.50}$$

$$\mathcal{L} = \frac{1}{T}\sum_{i=0}^{T} \mathcal{L}_i \tag{6.51}$$

where T is the target sequence length. To overcome the issues related to the repetition of words, a coverage mechanism is employed to direct the attention version to non-repeating words. The attention distributions generated using all previous prediction steps are added to create the coverage vector χ_i^v.

$$\chi_t = \sum_{t'=0}^{t-1} \Omega_{t'} \tag{6.52}$$

Therefore, eqn. (6.43) is modified by considering the effect of the coverage mechanism. Hence, this makes it easier for the attention mechanism to stop attending to the same places repeatedly and stop producing repetitive text as a result.

$$\Omega_{t'} = f(MHA(\mathcal{k}, q, \mathcal{v}), C, \chi_t) \tag{6.53}$$

The coverage loss is also defined to penalize repeatedly attending to the same locations.

$$\mathcal{L}_c = \lambda \sum_i \min\left(\chi_t, \Omega_{t'}\right) \tag{6.54}$$

where λ is the balancing parameter. Hence, the total loss for the proposed $UnMHA - ST$ is the combination of eqns. (6.53) and (6.54) respectively,

$$\mathcal{L}_{total} = \mathcal{L} + \mathcal{L}_c \tag{6.55}$$

6.5 Experimental Work and Results

To validate the effectiveness of the proposed MAA-FIC, experiments are conducted on the Flickr8K dataset. This dataset contains around 8K images paired with five descriptions for each image. To generate a style-based description of an image FlickrStyle10K dataset is utilized. This dataset contains textual annotations for romantic and humorous styles respectively. For this dataset, only 7K annotations are publicly available. For the task of summarization, the text data obtained from the two captioning methods (discussed in Sections 6.2 and 6.3) is collected and utilized for the generation of summarized description of an image reflecting factual, romantic, and humorous elements. The data used for summarization contains around 7000 paragraphs with three sentences (one each for factual, romantic, and humorous caption).

6.5.1 Implementation Details

For the factual caption generation model, Adam optimizer is utilized for the minimization of cross-entropy loss with a learning rate of $2e-5$. Further, the batch size is set as 64 and the proposed model is trained for 70 epochs. Also, to extract. the text features, fine-tuned GloVe embeddings are utilized with embedding size as 300. To evaluate the performance of the proposed factual image captioning model, BLEU@N [177] and METEOR [26] scores are evaluated.

To generate romantic and humorous captions, ADAM [193] optimizer is utilized at the encoder end with a learning rate of $2e-5$. For language decoder embedding size is set as 300 and the dimension of the hidden layer of the proposed SF-Bi-ALSTM module is set as 512. The model is trained for 60 epochs with a batch size of 96. To evaluate the performance of generated stylized captions, style transfer accuracy and perplexity of the proposed model are evaluated. Also, the relevancy of the captions generated is evaluated in terms of BLEU@N [177] and METEOR [26].

For the summarization module, a single-layer $UnMHA-ST$ transformer with a pointer generator network and coverage mechanism is incorporated. The model is trained for 100 epochs with a dropout rate of 0.1 and a batch size of 4. Further, the pointer-generator network leverages 300-dimensional GloVe embeddings with a vocabulary size of 400K. To evaluate the performance of the proposed model, ROUGE-1, ROUGE-2, and ROUGE-L scores are evaluated.

6.5.2 MAA-FIC Module

Table 6.1 reports the comparison of the proposed MAA-FIC model with other state-of-the-art on the Flickr8K dataset. The comparison of the proposed model is made in terms of $BLEU@N$ and $METEOR$ scores. MAA-FIC utilized fusion of Inception-V3 and Faster R-CNN at the encoder and provides state-of-the-art results when compared with ImageNet-CNN [128], VGGNet-CNN [96] based feature extraction method. Also, Jia et al. [103] utilized semantic information from the image as an extra input to each unit of the LSTM block. Very few works [123] [124] focused on attention mechanism-based architectures and the proposed MAA-FIC utilized modified adaptive attention (MAA). MAA combines channel and spatial attention

mechanisms to focus more on most of the important image information as well as the position of the image regions. Furthermore, the MAA provided a significant increase in the BLEU@N and METEOR scores when compared with the [93] [110] [220] [221].

Table 6.1: Results for the proposed MAA-FIC Module on Flickr8K Dataset

Model	B-1	B-2	B-3	B-4	M
LSTM [128]	66.0	42.0	27.0	18.0	-
g-LSTM [103]	64.7	45.9	31.8	21.6	20.60
Log Bilinear [93]	65.6	42.4	27.7	17.7	17.31
SCA-CNN [124]	68.2	49.6	35.9	25.8	-
Hard Attention [123]	67.0	45.7	31.4	21.3	20.30
m-RNN [96]	48.2	35.7	26.9	20.8	-
Deep-Bi-LSTM [110]	65.5	46.8	32.0	21.5	-
SDCD [220]	67.2	45.1	30.5	21.5	-
JRAN [221]	67.7	47.4	33.2	22.7	20.9
Proposed (MAA-FIC)	**74.8**	**52.3**	**38.7**	**26.8**	**22.6**

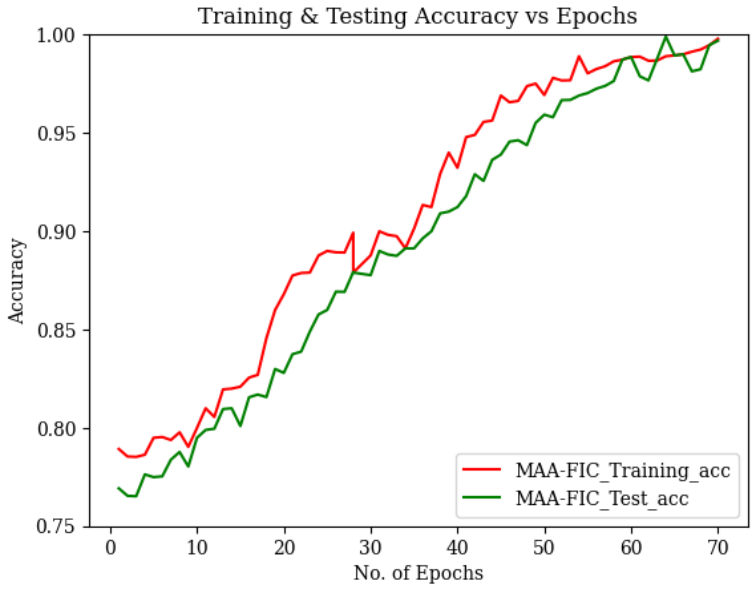
Figs. 6.6 (a) and (b) depict the accuracy and loss curves for the proposed MAA-FIC. The curves make it evident that with the increase in the number of epochs, the accuracy of the proposed model increases while the losses decrease. Fig. 6.7 represents the qualitative results obtained for the proposed MAA-FIC model. The factual captions obtained for Flickr8K produces syntactically and semantically correct descriptions by focusing more on the salient object regions. Also, by leveraging the modified adaptive attention the proposed model provides a better correlation between the objects.

6.5.3 SF-Bi-ALSTM Based Stylized Image Captioning

Table 6.2 reports the results using both, the romantic and the humorous references for the proposed SF-Bi-ALSTM-based stylized image captioning module with other state-of-the-art. The comparison is carried out with respect to $BLEU-1$, $BLEU-3$, and $METEOR$ (M) scores with style accuracy (cls) and perplexity (ppl).The results reported in Table 6.2 make

it evident that (1) given a specified style, the proposed model is tailored to that style outperforming the baseline approaches across different automatic evaluation metrics; (2) the relative performance variation shows that the proposed model can successfully simulate the style factors in caption generation. Further, StyleNet [54] and SF-LSTM [51] incorporate factored LSTM and style-factual LSTM decoders for the generation

(a)



(b)

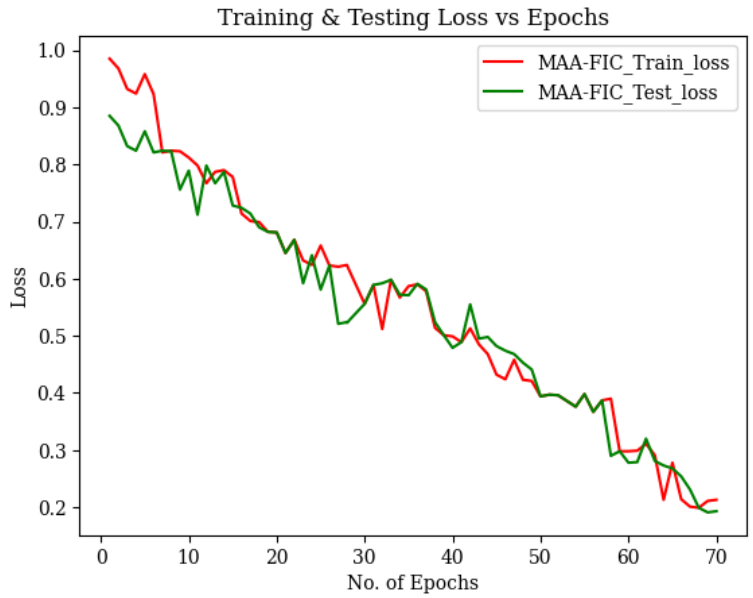


Fig. 6.6: (a) Accuracy and (b) Loss curves for the proposed MAA-FIC

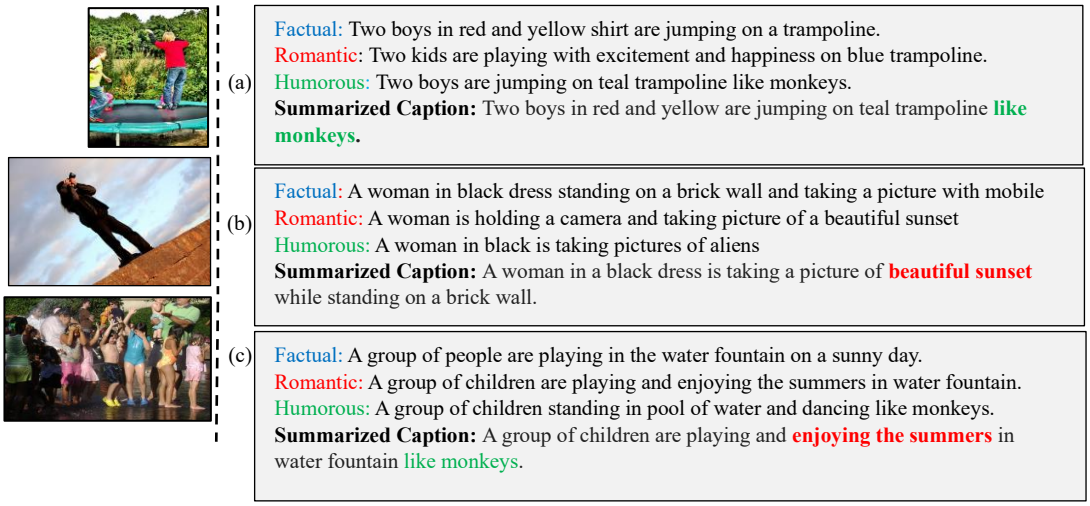


Fig. 6.7: Qualitative results obtained for the proposed $UnMA$ −CapSumT (a) UnMA-CapSumT Factual + Humorous (b) UnMA-CapSumT Factual + Romantic (c) UnMA-CapSumT Factual + Romantic + Humorous

Table 6.2: Results for the proposed SF-Bi-ALSTM Module on FlickrStyle10K Dataset

Style	Model	B-1	B-3	M	cls	ppl
Romantic	StyleNet [51]	13.3	1.5	4.5	57.1	6.9
	MSCap [53]	17.0	2.0	5.4	91.3	-
	SF-LSTM [54]	27.8	8.2	11.2	-	-
	SAN [58]	28.3	8.7	11.5	90.9	9.1
	Detach and attach [197]	24.3	-	-	82.4	-
	Wu et al. [198]	25.4	5.7	9.2	-	-
	Proposed	**29.8**	**9.2**	**11.9**	**92.0**	**9.3**
Humorous	StyleNet [51]	13.4	0.9	4.3	42.5	7.3
	MSCap [53]	16.3	1.9	5.3	88.7	-
	SF-LSTM [54]	27.4	8.5	11.0	-	-
	SAN [58]	27.6	8.1	11.2	87.8	8.4
	Detach and attach [197]	23.0	-	-	89.2	-
	Wu et al. [198]	27.2	5.9	9.0	-	-
	Proposed	**29.2**	**9.0**	**11.8**	**89.8**	**8.7**

of stylized captions. These methods proposed end-to-end learning framework for the generation of factual and style-based descriptions. On the other hand, the proposed model defines a modified style factored Bi-LSTM with captions. Further, the works [58] [53] implemented the task of multi-style-based caption learning using unpaired

data but these methods when compared with the proposed stylized captioning model generate more attractive and coherent style-based descriptions.

The qualitative results presented in Fig. 6.7 for the generation of romantic and humorous elements proves that the generated descriptions describe the content of the image well. The generated captions express the content in a romantic *(A little girl enjoying the joys of childhood with her brother in background)* or humorous *(A group of children standing in pool of water and dancing like monkeys).* More intriguingly, the descriptions generated are not only romantic or humorous but they also suit the visual content of the image coherently, making the caption visually appealing and relevant. Also, qualitative results were obtained to verify that the sentences so generated provide a strong correlation with the human evaluation

6.5.4 UnMHA-ST Text Summarization Transformer

The summarized captions generated from the proposed UnMHA-ST is presented in Fig. 6.7. From Fig. 6.7 it is evident that there exist some cases as shown in Fig. 6.7(a), according to the context of the image the summarized captions depict only factual and humorous elements and for a few images as shown in Fig. 6.7(b) romantic style has dominated with the factual element. Further, there are cases (Fig. 6.7(c)) in which the proposed summarization framework can successfully include romantic and humorous styles with factual elements.

Table 6.3 reports the ROUGE-1, ROUGE-2, and ROUGE-L scores for the baseline transformer model [154] with Doc2Vec, GloVe, and fastText word embeddings. From the results reported, it is evident that the fastText word embeddings are superior as they can derive word vectors for unknown words or out of vocabulary

words. Further, the baseline transformer model is replaced with the proposed $UnMHA-ST$ and fTA- WE word embedding which results in significant improvement of R-1, R-2, and R-L scores. To overcome the OOV issues the proposed $UnMHA-ST$ transformer is utilized with pointer- generator. Also, to enhance the performance of the proposed framework and to avoid the repetition issues $UnMHA-ST$ is equipped with coverage mechanism. This provides significant improvements by generating fluent summarized captions. Also, box plot-based model performance comparison in terms of R-1, R-2, and R-L is shown in Fig. 6.8.

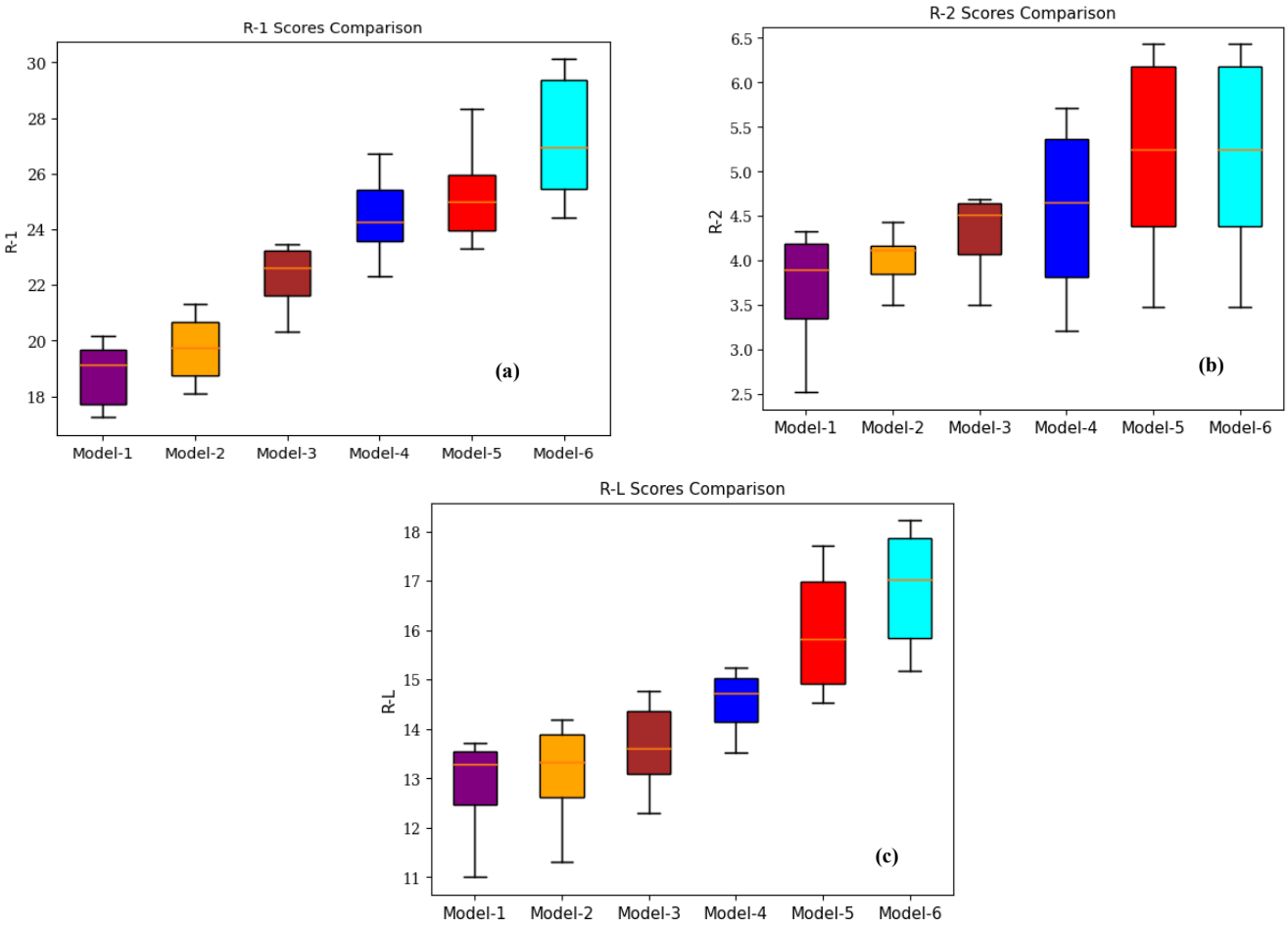


Fig. 6.8: Comparison for Quantitate Results Obtained for the proposed Framework (a) R-1, (b) R-2, and (c) R-L

GT: A group of people are playing in the water fountain on a sunny day. A group of children are playing and enjoying the summers in water fountain. A group of children standing in pool of water and dancing like monkeys.
Model-1: *A children are playing with water in the the summers in water pool*
Model-2: Children are playing in water and dancing in water enjoying in water in the the summers
Model-3: Children are dancing in water like monkeys enjoying in water fountain in the the summers
Model-4: *A group of people are playing in water fountain and dancing in pool of the summers in water fountain.*
Model-5: *A group of people are enjoying and enjoying in water fountain like monkeys.*
Model-6: *A group of children are playing and enjoying the summers in water fountain like monkeys.*

GT: A women in blue clothes sits alone against a brick wall and looking outside a window. Woman sits against brick wall in coffee shop having sweet coffee waiting for her lover. A woman sits in a café looking at a lizard on the brick red wall
Model-1: *A women sitting against red wall alone in café drinking coffee in café waiting for lover in coffee shop.*
Model-2: A women sitting near red wall in cafeteria wait for lizards at red wall and waiting for lover in cafeteria.
Model-3: A women in blue is looking at lizards on red wall and sitting against red wall in cafeteria.
Model-4: *A women in blue sitting in café against red wall waiting for lover in café while looking at lizards in wall.*
Model-5: *A women sitting in cafeteria against red wall waiting for her coffee lover and looking at lizards on red wall in cafeteria.*
Model-6: *A women in blue clothes sitting alone against red wall in café having a sweet coffee and waiting for her lover while looking at lizards*

Fig. 6.9: Observed Qualitative Results for the proposed *UnMA*-CapSumT *(Text in Red represents the repeated words that appeared in the summarized captions)*: Model -6 (proposed model) provides a summarized caption for an image without repetition of words and out-of-vocabulary issues

Further, Fig. 6.9 presents the qualitative ablated summarized captions for Model 1 to Model-6. Model-1, the baseline transformer model with Doc2Vec word embedding generated a poor-quality summarized caption with redundant and repeated information. Further, fastText embedding with the baseline transformer model (Model-3) provided a slight improvement in the generated summarized captions. With the use of the proposed $UnMHA - ST$ and fTA-WE in Model-4 provided an improvement in results but still there exists repetition of words like "in water fountain" in first case and "café" and "wall" in second case. In Model-5, the pointer-generator network adaptively points to the contextual words with appropriate styles hence resulting in generation of romantic and humorous elements in the summarized captions. For the second case, Model-5 generates a word "cafeteria" that is synonym to the word "café"

or “coffee shop”. Further, Model-6 generated a summarized caption that provided a strong correlation with the human evaluation by generating syntactically and semantically correct descriptions by highlighting both factual and stylized elements. Also, the summarized caption generated is free from OOV and repetition problems with the use of a pointer-generator network and coverage mechanism.

Table 6.3: Ablation Results obtained to study the influence of the Baseline Transformer and the proposed $UnMHA - ST$

Model	Transformer	Embedding	R-1	R-2	R-L
Model-1	Baseline [154]	Doc2Vec	20.18	4.33	13.71
Model-2	Baseline [154]	GloVe	21.3	4.43	14.19
Model-3	**Baseline** [154]	**fastText**	**23.46**	**4.65**	**14.77**
Model-4	$UnMHA - ST$	fTA-WE	26.72	5.71	15.24
Model-5	$UnMHA - ST$ + Pointer Generator Network	fTA-WE	28.32	6.43	17.72
Model-6	$\boldsymbol{UnMHA - ST}$ **+ Pointer Generator Network + Coverage Mechanism**	**fTA-WE**	**30.11**	**6.73**	**18.22**

6.6 Significant Outcomes

This chapter presents a novel caption summarization framework, $UnMA-$CapSumT to generate summarized captions highlighting the factual, romantic, and humorous elements in a single caption. The proposed framework is divided into two stages: (i) generation of factual, romantic, and humorous descriptions of images using the MAA-FIC model and SF-Bi-ALSTM-based stylized image captioning model, (ii) using the descriptions generated in (i) to produce a summarized single line caption for the images by incorporating multi-head attention and unified attention driven $UnMHA - ST$ transformer.

Also, the proposed $UnMHA - ST$ transformer utilizes a pointer-generator network and coverage mechanism to avoid the issues related to OOV and repetition

problems. The summarized caption generated provides an improvement in learning knowledge of factual content and its associated linguistic styles. Also, the observed qualitative results make it evident that the proposed framework provided a strong correlation with human evaluation by generating semantically and syntactically correct descriptions.

Further, the improvements in the performance and interpretability of the proposed framework are enhanced with the utilization of an efficient word embedding fTA-WE. Experimental results prove that the proposed MAA-FIC, SF-Bi-ALSTM-based image captioning model, and $UnMHA - ST$ summarization transformer model provide state-of-the-art results and generate visually and grammatically correct factual, romantic, humorous, and summarized captions for a given image.

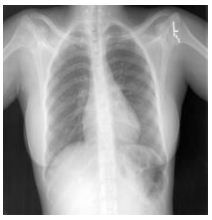
Chapter-7

Applications to Visual Image Captioning

Automatic Visual Captioning (AVC) generates syntactically and semantically correct sentences by describing important objects, attributes, and their relationships with each other. The very nature of it makes it suitable for applications such as image retrieval [222] [223], human-robot interaction [224] [225], aid to the blind [226], visual question answering [227] and the like. In this chapter we will discuss about two main application areas of image captioning namely: medical image captioning and aid to the blind. Medical image captioning highlights the relationships between image objects and clinical findings. Further, Image captioning can help visually impaired people understand their environment. For example, the Intelligent Eye app combines image and text processing to help the blind navigate themselves.

7.1 $FDT - Dr^2T$: A Unified Dense Radiology Report Generation Transformer Framework for X-Ray Images

Medical Image Captioning (MIC) [228] can assist doctors by accelerating the reporting process and reducing their workload. Fig. 7.1 presents an example of the captioning of medical images. Medical image captioning highlights the relationships between image objects and clinical findings, which makes it a very challenging task. The generation of medical reports not only reduces the doctor's workload and accelerates clinical workflow but also aids in the efficient exploitation of medical content. Therefore, this produces faster and more accurate interpretations of findings and offers important assistance to doctors.



Normal Chest with no pulmonary enema. The cardiac silhouette and mediastinum size are within normal limits. There is no focal consolidation. There are no traces of pleural effusion. There is no evidence of pneumothorax.

Fig 7.1: An Example of Radiology Report Generated

7.1.1 Proposed Methodology

With the increase in the advancements in technology, different methods [229, 230, 231] are developed to automatically generate medical reports based on the data available. These methods generated incorrect reports for some cases with erroneous and repeated sentences. Some models [229, 232, 233] generated incoherent sentences with incorrect order of words, which were not clinically acceptable. Also, these methods failed to report some rare but important abnormalities due to ambiguities and incorrect detection of objects from radiological images. To overcome these challenges, this paper presents an efficient framework for the generation of paragraph-based reports sometimes also known as dense MIC. The proposed framework is split into two stages (as depicted in Fig. 7.2): (i) FDT-Image feature extraction module and (ii) Dense Radiology Report Generation Transformer (Dr^2T) model for coherent medical report generation.

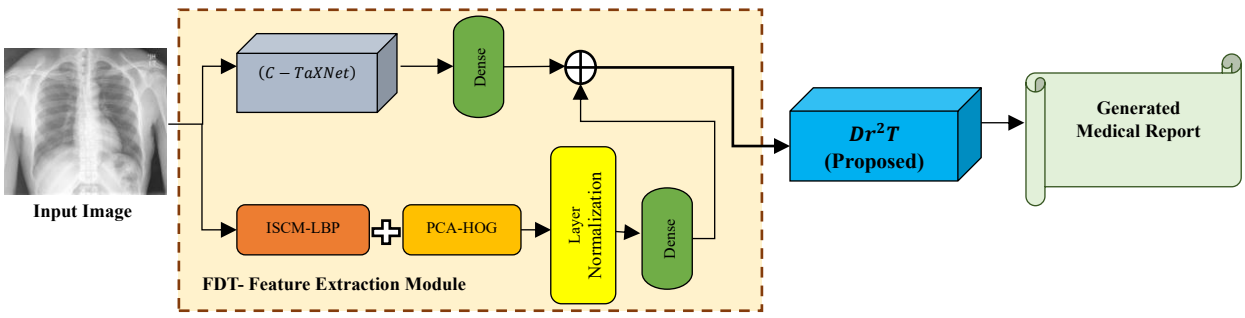


Fig. 7.2: Block Diagram of the proposed FDT- Dr^2T framework

The proposed FDT module efficiently fuses or concatenates deep features and texture features to extract effective features from the data available. The fused features obtained from the first stage are further leveraged by the transformer-based module for efficient and coherent report generation. Furthermore, the architecture of the proposed framework is presented in Fig. 7.3

7.1.1.1 FDT-Image Feature Extraction

Medical image feature extraction aims to find the most compact and informative set of features. Deep-feature extraction incorporates a modified XceptionNet [234] module with a Residual Feature Distillation Block (RFDB) [235] and a Shallow Residual Block (SRB). Further, the proposed modified XceptionNet module also leverages a triplet attention module for the extraction of multi-level rich deep -features. Further, for the extraction of texture features, the ISCM-LBP [236] algorithm is incorporated with reduced-dimensional HOG features [237]. This helps preserve a trade-off between model performance and parameters and ensures the appropriate level selection for the extraction of pertinent texture features.

7.1.1.1.1 Convolutional Triple Attention-based Efficient XceptionNet ($C-TaXNet$)

The proposed $C-TaXNet$ model (Fig. 7.3) leverages a modified version XceptionNet module with RFDB and SRB as depicted in Fig. 7.4. RFDB utilizes multiple features distillation connections to learn more discriminative feature representation. For more fine-grained residual learning SRB is utilized without introducing additional parameters [16]. Further, the SRB is equipped with SELU activation function [238] described mathematically as:

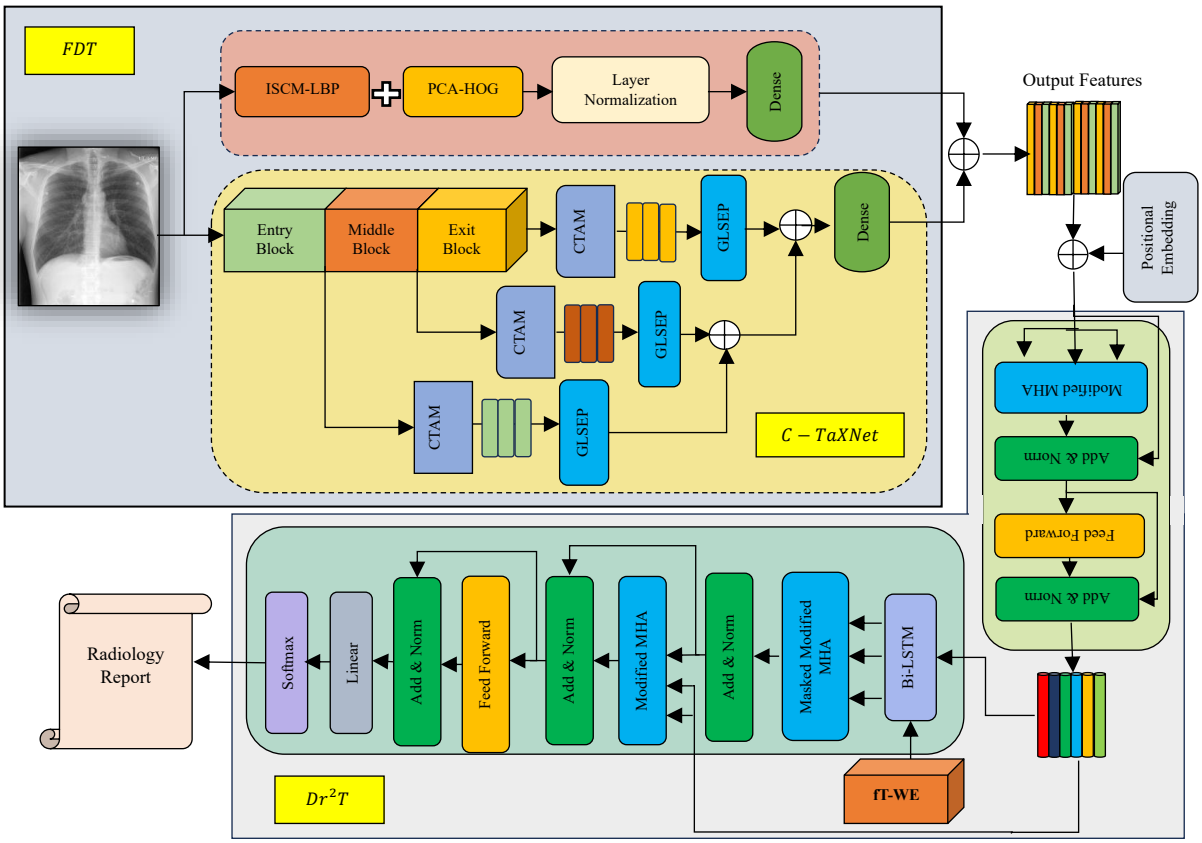


Fig. 7.3: Proposed FDT- Dr^2T Framework

$$g(x) = \begin{cases} \lambda x, x > 0 \\ \alpha\lambda(e^x - 1), x \leq 0 \end{cases} \tag{7.1}$$

where α and λ are the activation constants with values of $\alpha = 1.6732$ and $\lambda = 1.0507$. Also, the proposed $C - TaXNet$ utilizes the concept of Convolutional Triple Attention Module (CTAM) [239] (almost parameter-free) that model channel attention and spatial attention [240]. The internal structure of the CTAM module is presented in Fig. 7.5. Consider $f_{input} \in \mathbb{R}^{C \times H \times W}$ be the feature vector of the corresponding input image which serves as an input to the Entry block of the $C - TaXNet$. The features extracted from each Xception block are made to pass through CTAM which is a three-stage architecture. For the first stage, height and channel dimension interactions are

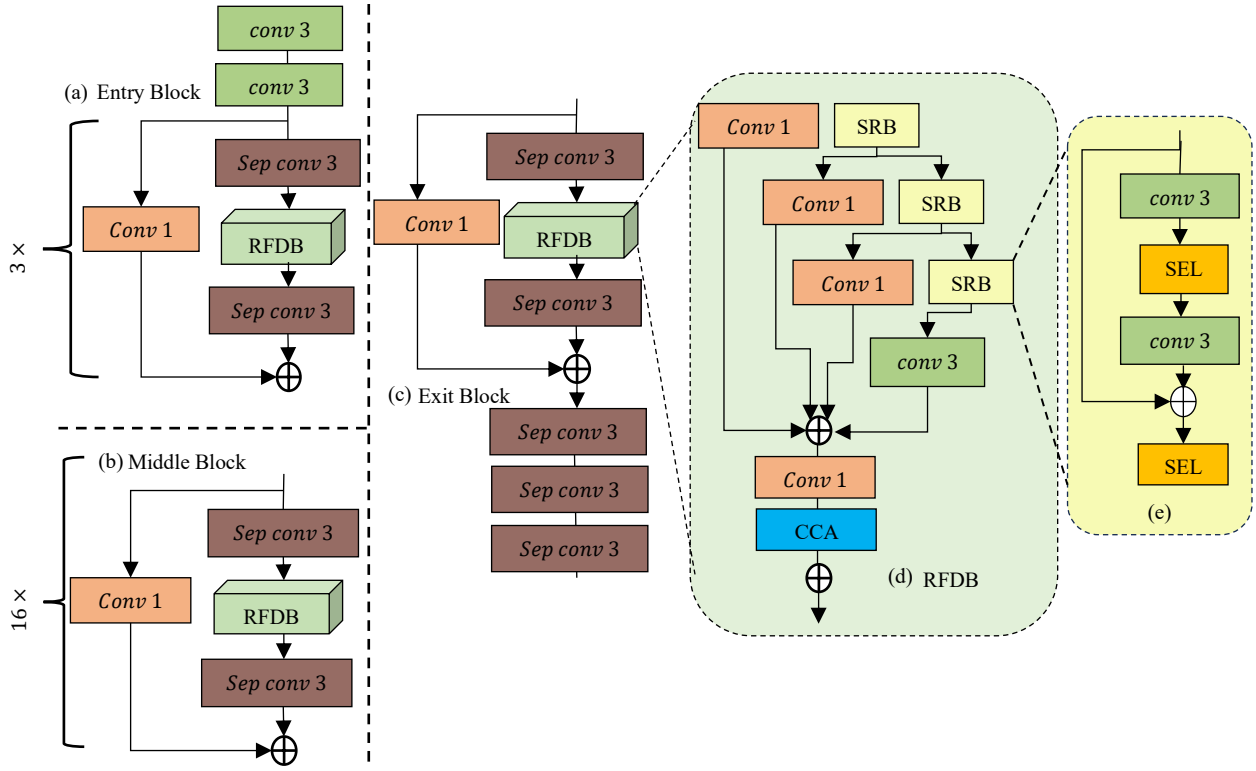


Fig. 7.4: Modified XceptionNet (a) Entry Block, (b) Middle Block, and (c) Exit Block, (d) RFDB Block, (e) SRB Block

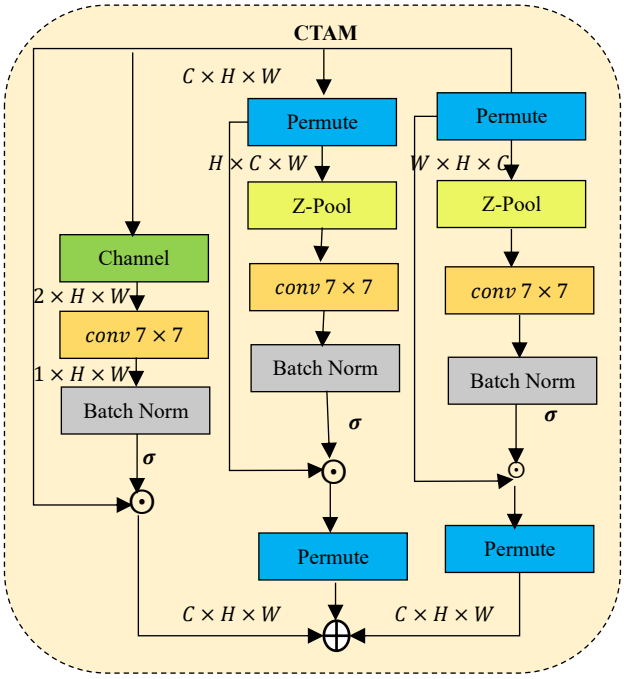


Fig. 7.5: Conventional Triple Attention Module

calculated by rotating the input 90° along the height dimension. The next stage calculates the interaction by rotating the input tensor to 90° anti-clockwise along W-axis. For the final stage, the channels of the input tensor are reduced by using Z-pool. The Z-pool concatenates the AveragePool and MaxPool across that dimension. Mathematically, Z-pool can be represented as:

$$z - pool(x) = [MaxPool_{0d}(x) AvgPool_{0d}(x)] \tag{7.2}$$

where $0d$ is the zeroth dimension across which the AveragePool and MaxPool operation takes place. Further, the refined triple attended output is represented by:

$$y = \frac{1}{3}(\overline{\hat{x}_1 \sigma(\varphi_1(\hat{x}_1^*))} + \overline{\hat{x}_2 \sigma(\varphi_2(\hat{x}_2^*))} + x\sigma(\varphi_3(\hat{x}_3^*))) \tag{7.3}$$

In the above equation, σ represents the sigmoid activation function; $\varphi_1, \varphi_2,$ and φ_3 represent the standard 2-D convolutional layers in three branches of triplet attention. Also, after the application of CTAM at each block of the Efficient XceptionNet we apply Global Log-Sum-Exp Pooling (GLSEP) [241] individually and concatenate the reduced features into a single vector.

7.1.1.1.2 ISCM-LBP + PCA-HOG Feature Extraction Algorithm

This section provides the details for the hand-crafted feature extraction algorithm. Traditional LBP features [242] are sensitive to noise and incorporate centre window pixels as the threshold, without any enhancement in the centre pixel and considering the relationship between the centre pixel and the neighbourhood pixel [243]. When the central pixel intensity value is too large or too small, there might be a possibility that the detailed features are missed. Therefore, ISCM-LBP algorithm is incorporated that utilizes standard deviation to improve the resilience of the centre

pixel and the function of the neighbourhood pixel. This further enhances the strength of the centre pixel and the role of the neighbourhood pixel. The specific operations are as follows:

- The 3 × 3 window is subdivided into sub-windows of size 3 × 3 to calculate the average gray value of each window Finally, standard deviation σ is calculated for 9 windows.

$$\sigma = \sqrt{\frac{\sum_{j=1}^{9}(\widetilde{X}_j - \widetilde{X})^2}{9}} \tag{7.4}$$

  where $\widetilde{X}_j$ average gray value of each window and $\widetilde{X}$ is the average value of the entire window.

- For threshold value, for $\not{f} \leq \sigma$, the threshold q_t is the median of the nine pixels in the 3 × 3 window and LBP is evaluated by the formula

$$LBP(X_c, Y_c) = \sum_{P=0}^{7} 2^P \, \mathbb{S}(q_p - q_c) \tag{7.5}$$

$$\mathbb{S}(N) = \begin{cases} 1, N \geq 0 \\ 0, N < 0 \end{cases} \tag{7.6}$$

Further, if $\not{f} > \sigma$, let m, and n be the maximum and minimum value from the 9-pixel values. The threshold is evaluated using:

$$q_t = \frac{m+n+\tilde{q}_c}{3} \tag{7.7}$$

- Finally, for median M of nine pixels, ISCM-LBP features are evaluated using the formula:

$$ISCM - LBP = \sum_{P=0}^{7} 2^P \, \mathbb{S}(\tilde{q}_p - q_t) \tag{7.8}$$

$$q_t = \begin{cases} M, \not{f} \leq \sigma \\ \frac{m+n+\tilde{q}_c}{3}, \not{f} > \sigma \end{cases} \tag{7.9}$$

$$\mathbb{S}(N) = \begin{cases} 1, \tilde{q}_p - q_t \geq 0 \\ 0, \tilde{q}_p - q_t < 0 \end{cases} \tag{7.10}$$

The ISCM-LBP only extracts the texture information features and somehow neglects some edge density features from images. Therefore, we incorporate dimensionality reduction of histogram of gradient orientation i.e., PCA-HOG. This helps in the extraction of large dimensional HOG features which are reduced with the utilization of PCA mapping. Finally, the ISCM-LBP and HOG-PCA features are serially concatenated that help to retain all the texture feature details.

Flowchart depicting the extraction process for the ISCM-LBP + PCA-HOG algorithm is presented in Fig. 7.6. Furthermore, the features obtained from $C - TaXNet$ and ISCM-LBP + PCA-HOG are fed to dense blocks which are the learnable feature encodings that convert the deep-features and hand-crafted features into learnable vectors $\mathbb{f}_{deep}$ and $\mathbb{f}_{hcf}$.

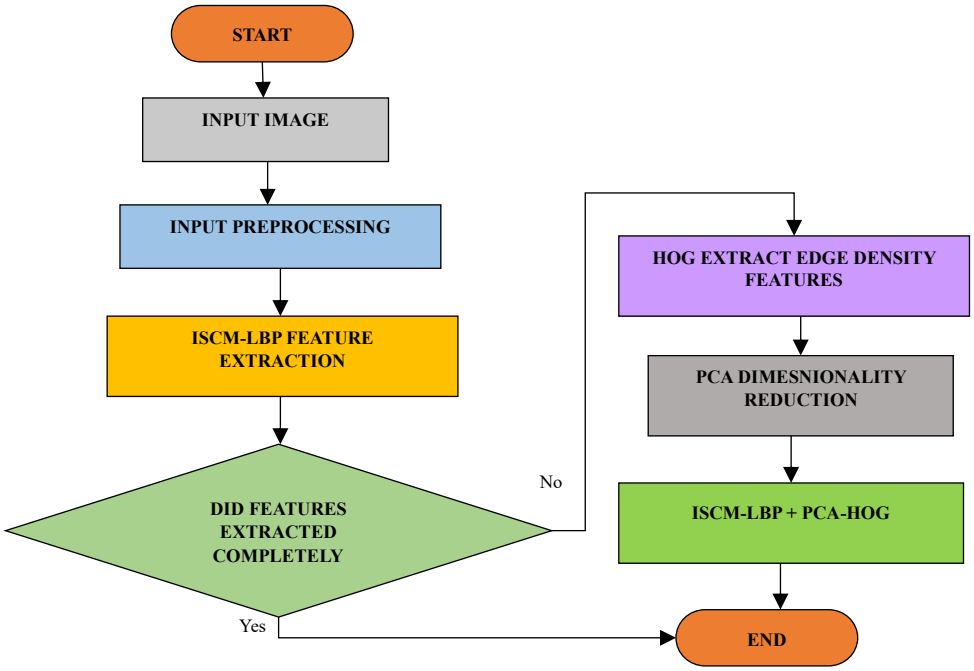


Fig. 7.6: Flowchart for ISCM-LBP + PCA-HOG Algorithm

7.1.1.2 Dense Radiology Report Generation Transformer (Dr^2T)

The image features obtained from the FDT-Image feature extraction module are fed to the transformer-based architecture for the generation of dense medical reports. Fig. 7.3 presents the detailed structure of the proposed Dr^2T transformer model. The traditional transformer model incorporates Multi-Head Attention (MHA) and calculates a weighted average of values based on the similarity between queries and keys. The attention scores assigned to each value determines its importance. To extract more minute details in the form of deep local features, this chapter presents a modified MHA as shown in Fig. 7.7, which selects the best-head from the structure and concatenates the best-head and the multi-head representations to explore more correlations between the higher-order feature representations.

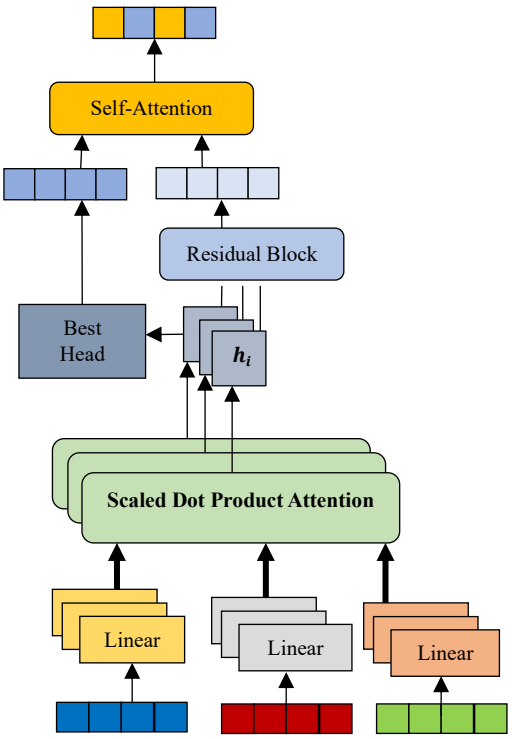


Fig. 7.7: Proposed Modified MHA

7.1.1.2.1 Dr^2T Encoder

The encoder structure of the Dr^2T converts a set of visual attributes from the input image into a number of encodings. The encodings or embeddings obtained are represented in the form of a linear transformation of local features as keys $\mathcal{k}$, query q, and values $\mathcal{v}$ respectively. With this, we first calculate the multi-head self-attention scores $m\mathcal{h}_i$ using scaled-dot product attention (SPA). Mathematically,

$$m\mathcal{h}_i = [\mathcal{h}_1, \mathcal{h}_2 \dots \dots, \mathcal{h}_i] \tag{7.11}$$

$$SPA = softmax\left(\frac{q_i \mathcal{k}_i^T}{\sqrt{d_i}}\right)\mathcal{v}_i \tag{7.12}$$

Further, we select the best head ($m\mathcal{h}_b$) out of the calculated multi-head self-attention scores. Also, these heads $m\mathcal{h}_i$ are made to pass through to a residual block (RB) in order to avoid gradient explosion.

$$S_t = \sum RB(m\mathcal{h}_i) \tag{7.13}$$

The output generated from the RB and the best-selected head is fed to the self-attention network to provide a better correlation among the features obtained. The features obtained after the self-attention network is represented by:

$$\mathcal{S} = S - A(m\mathcal{h}_b, S_t) \tag{7.14}$$

The Modified-MHA features are added, normalized and fed to an FFN to get higher order visual features f_v as output which acts as input to Dr^2T Decoder module.

7.1.1.2.2 Dr^2T Decoder

The Dr^2T decoder aims for the generation of dense medical reports either in the form of single or multiple sentences with the use of the high-level features obtained from the Dr^2T encoder. A combination of visual features f_v and textual embeddings f_t are used to provide higher-order intra and inter-modal interactions between the

image and text pairs. Also, the textual embeddings are extracted through fastText word embeddings (fT-WE) [199]. Therefore, textual and visual embeddings are concatenated via Bi-LSTM to perform multi-modal reasoning and high-level feature representation. This whole fT-WE with visual embedding (from Dr^2T encoder) and Bi-LSTM layer supports detailed encryption of visual data and provides inter-and intra-model interactions between the text and visual features.

At the decoding stage, we further integrate the Masked Modified-MHA and Modified-MHA with add and norm layers and feed-forward network layer. Finally, the medical report is generated by the generation of words through FC and softmax layers. The proposed Dr^2T model is trained by minimizing the cross-entropy loss:

$$\mathcal{L}_{XE}(\theta) = -\sum_{t=1}^{\mathcal{T}} \log\left(p_\theta(w_t^* \mid \mathbb{S}_{1:t-1}^*)\right) \tag{4.15}$$

where, $\mathcal{T}$ is the number of words in sentence; θ denotes all the parameters in the model; $\mathbb{S}_{1:t-1}^*$ is the ground truth.

7.1.2 Experimental Details and Results

To evaluate the performance of the proposed framework, publically available dataset IU X-Ray [244] is used for the generation of paragraph-based reports. IU Chest X-Ray contains about 7470 pairs of images and reports. Each report consists of the following sections: impressions, findings, tags, comparison and indication. The report for the image includes the impression and findings only. Following standard procedure [245], images are normalized and resized to 224×224, making them appropriate for extracting visual features. The text data was pre-processed to produce 572 unique tags and 1915 unique words by changing all tokens to lowercase and deleting any non-alpha tokens. On average, there are 2.2 tags, 5.7 sentences, and 6.5 words per sentence

attached to each image. Additionally, we discovered that the top 1,000 terms account for 99.0% of all word occurrences in the dataset; as a result, we only used the top 1,000 terms listed in dictionaries. Finally, we chose 500 images at random for validation and 500 images as a test. Further, to validate the effectiveness of the proposed Dr^2T we evaluated BLUE-n [177], METEOR [26], CIDEr [25], and ROUGE-L [152] scores.

7.1.2.1 Implementation Details

For the first phase of the experiment, the input images are resized into 480×480 resolution and Adam [193] optimizer is used for training of $C - TaXNet$ and texture feature extraction module. The feature extraction FDT module is trained for 100 epochs with a learning rate of $10e - 4$ and batch size of 32. For second phase, the Dr^2T model is trained for further 70 epochs with a learning rate of $1e - 4$ for the encoder and decoder. We apply a linear decay scheduler with a warmup ratio of 0.01 to control the learning rate. For regularization, we set dropout to 0.1 and use a weight decay of 0.01. For the generation of paragraph, we set the dimensions of all hidden states and word embeddings as 512 and the maximum output length is set as 45. For both FDT and Dr^2T modules, we adopt a softmax cross-entropy loss.

7.1.2.2 Quantitate comparison with state-of-the-art

For the deep-feature extraction module, the proposed $C - TaXNet$ is compared with different generic networks (ResNet [16], VGGNet [108], AlexNet [109], DenseNet [246], Inception-V3 [155], and XceptionNet [234]) in terms of accuracy and loss. Table 7.1 presents the results obtained for the proposed $C - TaXNet$. It is evident that the proposed network provides significant improvements in terms of accuracy

with the addition of triple attention module. Also, Fig. 7.8 presents the attention maps generated from the proposed $C - TaXNet$. The proposed model attends to the input image efficiently by highlighting the minute image regions that are of importance.

Table 7.1: Comparison of the performance of the proposed $C - TaXNet$ and other Deep-feature Extraction Framework

Model	Accuracy (%)	Loss
ResNet [16]	89.64	3.547
VGGNet [108]	90.32	3.10
AlexNet [109]	89.10	3.09
Inception-V3 [155]	94.35	2.341
DenseNet [246]	94.13	2.61
XceptionNet [234]	96.80	1.928
Proposed ($C - TaXNet$)	**98.74**	**1.29**

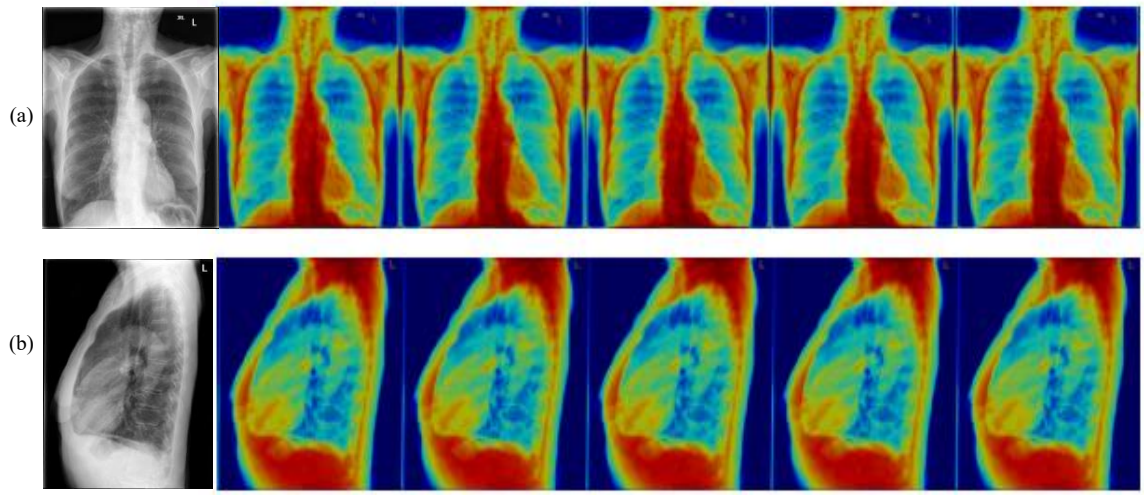


Fig.7.8: Attention Visualization generated by the proposed $C - TaXNet$

Table 7.2 presents the dimension of extracted features and feature extraction speed for the proposed ISCM-LBP+HOG- PCA algorithm. From Table 7.2 it is evident that with the improvement in speed of feature extraction, a reduction in data redundancy is observed. The results prove that more accurate image features with

Table 7.2: Comparison of the performance of different texture feature extraction algorithms.

Method	Feature Dimension	Feature Extraction Time (s)
LBP	1982	0.0121
ISCM-LBP	2770	0.0068
HOG	3221	0.0039
HOG-PCA	1026	0.0070
ISCM-LBP + HOG-PCA	2951	0.0105

strong anti-interference capabilities and minimal computing complexity are extracted by the ISCM-LBP approach. Additionally, PCA-HOG extracts the image edge features and ensures that, even at accelerated extraction speeds, the feature dimension is successfully lowered without compromising the edge features that have already been retrieved. With serial fusion of ISCM-LBP and HOG-PCA relatively complete image features with edge and texture features are obtained. Figs. 7.9 and 7.10 compare different texture feature extraction algorithms on the basis of the feature dimension and time taken (s) for the extraction of features.

Table 7.3 reports the quantitative results obtained on the proposed FDT-DMICT. We compare our proposed approach with a wide range of state-of-the-art medical report generation methods, i.e., Variational topic inference (VTI) [245], a graph-based method [247], cross-modal memory (CMR) [248], CMAS [249], Eddie-Transformer

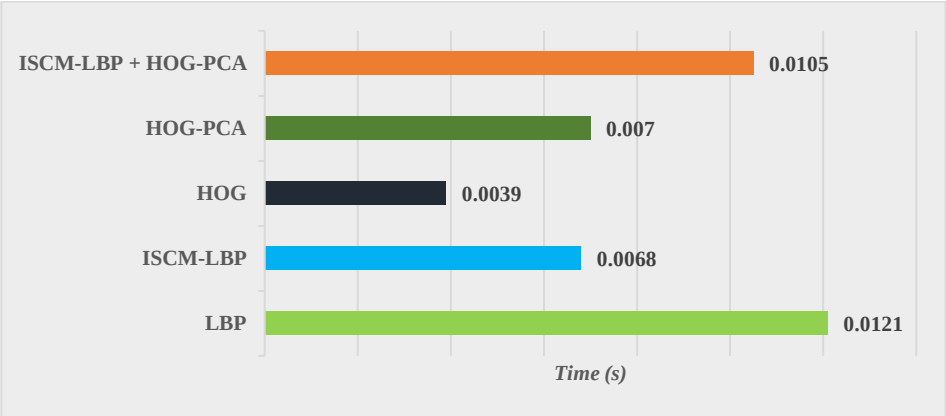


Fig. 7.9: Feature Extraction time taken by different texture feature extraction algorithms

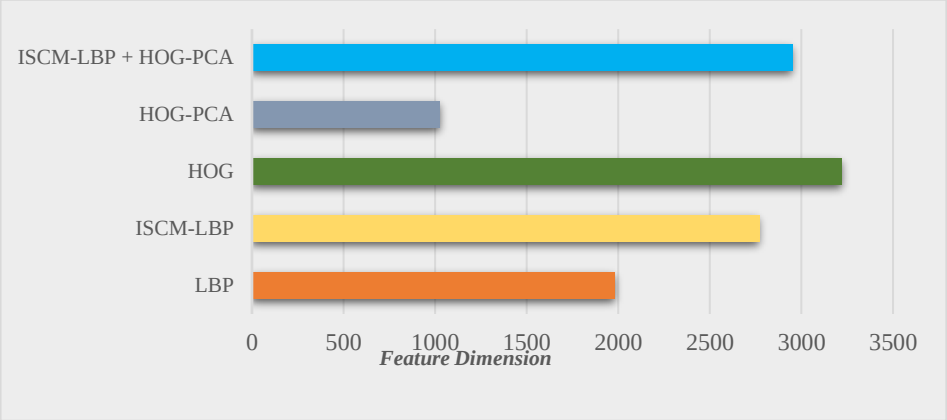


Fig. 7.10: Comparison of performance of different texture feature extraction algorithm.

Table 7.3: Comparison of quantitative results obtained for the proposed $FDT - Dr^2T$ framework with state-of-the-art.

Method	B-1	B-2	B-3	B-4	M	CIDEr	R-L
VTI [245]	0.493	0.360	0.291	0.154	0.218	-	0.375
Wang et al. [247]	0.450	0.301	0.213	0.158	-	-	0.384
CMR [248]	0.475	0.309	0.222	0.170	0.191	-	0.375
TieNet [230]	0.330	0.194	0.124	0.081	-	-	-
RTMIC [254]	0.350	0.234	0.143	0.096	-	-	-
Li et al. [231]	0.438	0.298	0.208	0.151	-	-	0.322
R2Gen [100]	0.470	0.304	0.219	0.165	0.187	-	0.371
Eddie-Transformer [250]	0.466	0.307	0.218	0.158	-	-	0.358
CMAS [249]	0.464	0.301	0.210	0.154	-	-	0.362
Delta-Net [251]	0.485	0.324	0.238	0.184	-	-	0.379
IIHT [252]	0.513	0.375	0.297	0.245	0.264	-	0.492
PPKED [253]	0.483	0.315	0.224	0.168	-	0.351	0.376
Co-Att [229]	0.517	0.386	0.306	0.247	0.217	0.327	0.447
LXMERT [255]	0.498	0.32	0.229	0.169	0.205	-	0.379
ITHN [256]	0.491	0.328	0.231	0.183	0.210	-	0.387
CMCA [257]	0.496	0.349	0.268	0.215	0.209		0.392
ORGAN [258]	0.510	0.346	0.255	0.195	0.205		0.399
Proposed ($FDT - Dr^2T$)	**0.531**	**0.398**	**0.322**	**0.251**	**0.277**	**0.384**	**0.506**

[250], DeltaNet [251], IIHT [252], PPKED [253], Co-Att [229], RTMIC [254]. From the results reported it is evident that the $FDT - Dr^2T$ framework outperforms the recent LXMERT [255], ITHN [256], CMCA [257], ORGAN [258] state-of-the-art as well by a large margin across all evaluation metrics. Furthermore, the proposed work outperforms different hierarchical-based approaches [7] [8] [18] [20]. The proposed

$FDT - Dr^2T$ reports abnormalities by detecting opacity by observing the contrast, detecting hyperexpanded lungs by determining the change in size, and identifying the location of airspace disease.

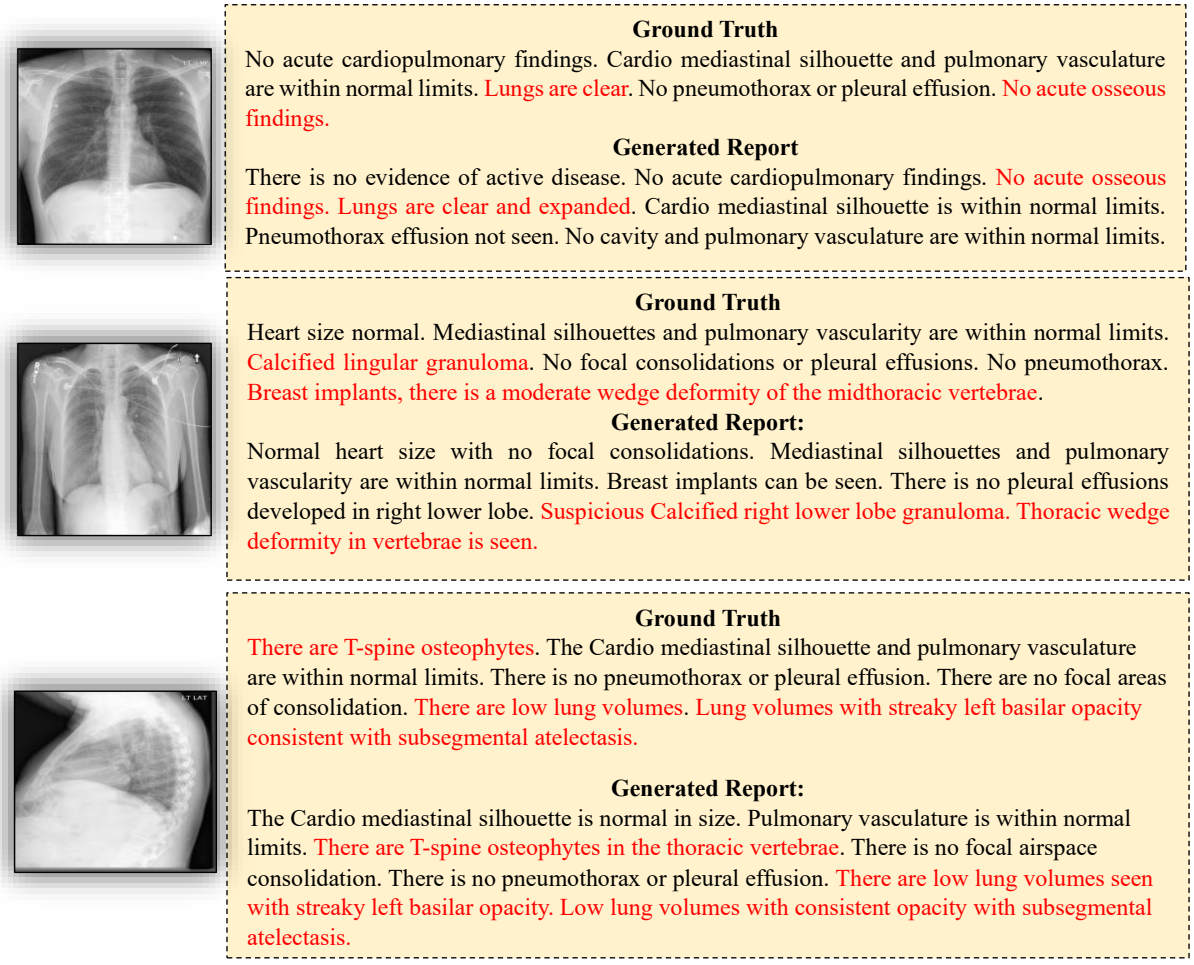


Fig. 7.11: Qualitative Results for the proposed $FDT - Dr^2T$, *(red marks highlight different abnormalities)*

7.1.2.3 Qualitative Results for the Proposed $FDT - Dr^2T$

Fig. 7.11 presents the reports generated by the proposed $FDT - Dr^2T$. The figure describes the report generated corresponding to the input image and the ground truth. From the qualitative results portrayed in Fig. 7.11, we can infer that the proposed $FDT - Dr^2T$ framework is better at capturing abnormalities. For the third image, our

framework easily detects and reports "*T-spine osteophytes in the thoracic vertebrae".* Therefore, the proposed framework has a remarkable ability to accurately generate comprehensive reports for abnormal cases. Further, with this, the proposed framework provides accurate locations of the abnormalities with a well balance of normal sentences and abnormal sentences. Also, we can infer from these results that the $FDT - Dr^2T$ can efficiently alleviate the visual data deviation problem by efficiently detecting changes in the contrast, size, and disease location, which helps in the coherent generation of medical reports. Further, there were very few cases observed where the proposed framework does not generate descriptions as per the grammar rules like in Fig. 7.11, the generated sentence is "*Pneumothorax effusion not seen".*

7.1.2.4 Ablation Studies

An ablation study is conducted which exhibits the effectiveness of different modules for the efficient generation of paragraph-based medical reports. This study effectively highlights the influence of baseline Transformer and Dr^2T. This study also helps to gain a better understanding of the overall behaviour of the ablated models on the performance of the system. Table 7.4 reports the results obtained for different models i.e., (M-1 to M-6), and Fig. 7.12 provides the qualitative results obtained for all the ablated models. We start with the baseline transformer model [56] with Inception-V3 and LBP+HOG features for the FDT module (M-1). The results obtained with M-1 do not provide significant results. In second model i.e., M-2, we replace Inception-V3 with XceptionNet, a slight improvement is observed in B-1, B-2, METEOR, and ROUGE-L scores. Further, in M-3, the baseline transformer model is replaced with the proposed Dr^2T, and significant qualitative and quantitative results are obtained in

comparison with M-1 and M-2. With baseline transformer and XceptionNet with ISCM-LBP + HOG-PCA deep and texture features module (M-4), there is no significant improvement observed with respect to M-3 in the evaluation scores. Therefore, to improve the qualitative and quantitative results, we exploited the proposed $C - TaXNet$ for deep-image feature extraction with ISCM-LBP + HOG-PCA and baseline transformer i.e., M-5. To further obtain coherent medical reports with correct abnormalities we employed the proposed FDT module ($C - TaXNet$ and ISCM-LBP + HOG-PCA) and the proposed Dr^2T transformer. M-6 provides better correlations for higher-order feature representations from medical images and generates coherent reports with rare and important abnormalities. The qualitative results obtained for different model makes it evident that a coherent report is generated with the use of proposed $FDT - Dr^2T$ highlighting the abnormalities in the report generated. Furthermore, Fig. 7.13 presents the influence of the proposed Dr^2T and the baseline transformer with respect to evaluation parameters.

Table 7.4: Ablation Study Results to study the influence of different FDT networks/algorithms with the baseline and the proposed Dr^2T transformer.

FDT									
Deep-Features	**Texture Features**	**Transformer**	**B-1**	**B-2**	**B-3**	**B-4**	**M**	**C**	**R-L**
Inception-V3	LBP+HOG	Baseline [154]	31.1	18.4	12.6	9.2	16.4	30.2	38.4
XceptionNet	LBP+HOG	Baseline [154]	34.2	20.7	15.4	11.2	16.9	30.4	38.4
XceptionNet	LBP+HOG	$\boldsymbol{Dr^2T}$	**42.7**	**30.3**	**21.6**	**17.2**	**18.1**	**31.3**	**39.1**
XceptionNet	ISCM-LBP+HOG-PCA	Baseline [154]	42.9	29.6	21.6	17.2	18.3	31.5	39.0
$\boldsymbol{C - TaXNet}$	ISCM-LBP+HOG-PCA	Baseline [154]	46.1	32.7	24.9	21.4	21.8	32.1	41.4
$\boldsymbol{C - TaXNet}$	ISCM-LBP+HOG-PCA	$\boldsymbol{Dr^2T}$	**53.1**	**39.8**	**32.2**	**25.1**	**27.7**	**38.4**	**50.6**

M-1: No acute cardiopulmonary abnormality. Lungs are clear. Pulmonary vasculature is within normal limits. No pneumothorax or pleural effusion.
M-2: No acute cardiopulmonary findings. Lungs are clear. No evidence of pleural effusion. Pulmonary vasculature is within normal limits. Lungs are clear. No pneumothorax or pleural effusion.
M-3: No acute cardiopulmonary findings. Pulmonary vasculature is within normal limits. No evidence of focal air space. There is no pneumothorax or pleural effusion. The heart is not enlarged.
M-4: No acute cardiopulmonary findings. Cardio mediastinal silhouette is unremarkable. Pulmonary vasculature is within normal limits. Lungs are clear. No pneumothorax effusion. No acute bony abnormality.
M-5: No acute cardiopulmonary abnormality. No acute bony abnormality. Pulmonary vasculature is within normal limits. Lungs are clear. No pneumothorax effusion. The heart is not enlarged.
M-6 ($FDT - Dr^2T$): There is no evidence of active disease. No acute cardiopulmonary findings. No acute osseous findings. Lungs are clear and expanded. Cardio mediastinal silhouette is within normal limits. Pneumothorax effusion not seen. No cavity and pulmonary vasculature are within normal limits.

Fig. 7.12: Qualitative Results obtained for different ablated models.

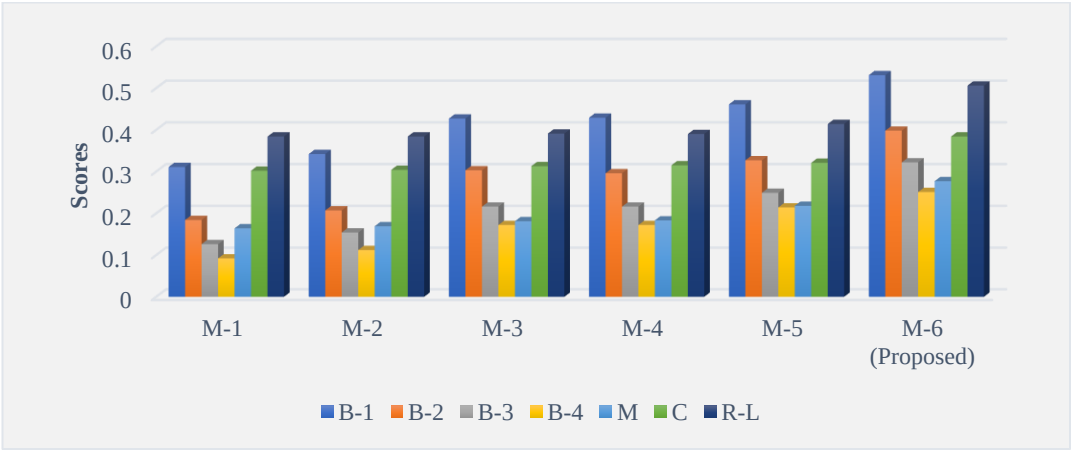


Fig. 7.13: Comparison highlighting the influence of different ablated models with respect to different evaluation parameters.

7.2 Automated Image Caption Generation Framework using Adaptive Attention and Bi-LSTM

Though image captioning has shown a great improvement with deep-learning techniques, the existing approaches are paying more attention to the enhancement of the RNN while ignoring the dominant extracted features. The attention mechanism describes each word located in the image region with a specific weight but ignores the

words such as ‘‘on’’ / ‘‘is’’ in the text description that do not correlate with the features of the image region. To overcome these issues, this section of the chapter proposes Inception V3, adaptive attention, and Bi-LSTM-based model for generating the caption of images for day-to-day life images and for aid-to-the-blind.

7.2.1 Proposed Methodology

The work discussed proposes an encoder-decoder framework for image captioning. CNN-based encoder compresses the input information to form a single fixed-length vector that is passed to the RNN based decoder. The proposed architecture, as depicted in Fig. 7.14, can be described as, (i) Feature extraction using InceptionV3, (ii) Adaptive Attention Mechanism, and (iii) Generation of the caption. The proposed work deals with forcing the text description to locate in an image region with improved accuracy and flexibility. InceptionV3 [] is a CNN architecture from the inception family. This architecture provides an improved version of CNN that includes label smoothing and 7 X 7 factorized convolutions. It also uses an auxiliary classifier that propagates label information lower down the network with batch normalization for the layers. LSTM is an RNN that captures context messages in long sequences enabling it to workout gradient disappearance and explosion problems. Further, to resolve ‘‘long-term dependence’’, LSTM is designed in such a way that it ensures the availability of memory cells and gates at each time these are required. Attention module structure makes use of LSTM for feature extraction. It focuses on image features by embedding the attention mechanism. The LSTM model structure (see Fig. 7.15) and the related mathematical equations are defined as:

$$i_t = \sigma(x_t \times W_{xi} + h_{t-1} \times W_{hi}) \tag{7.16}$$

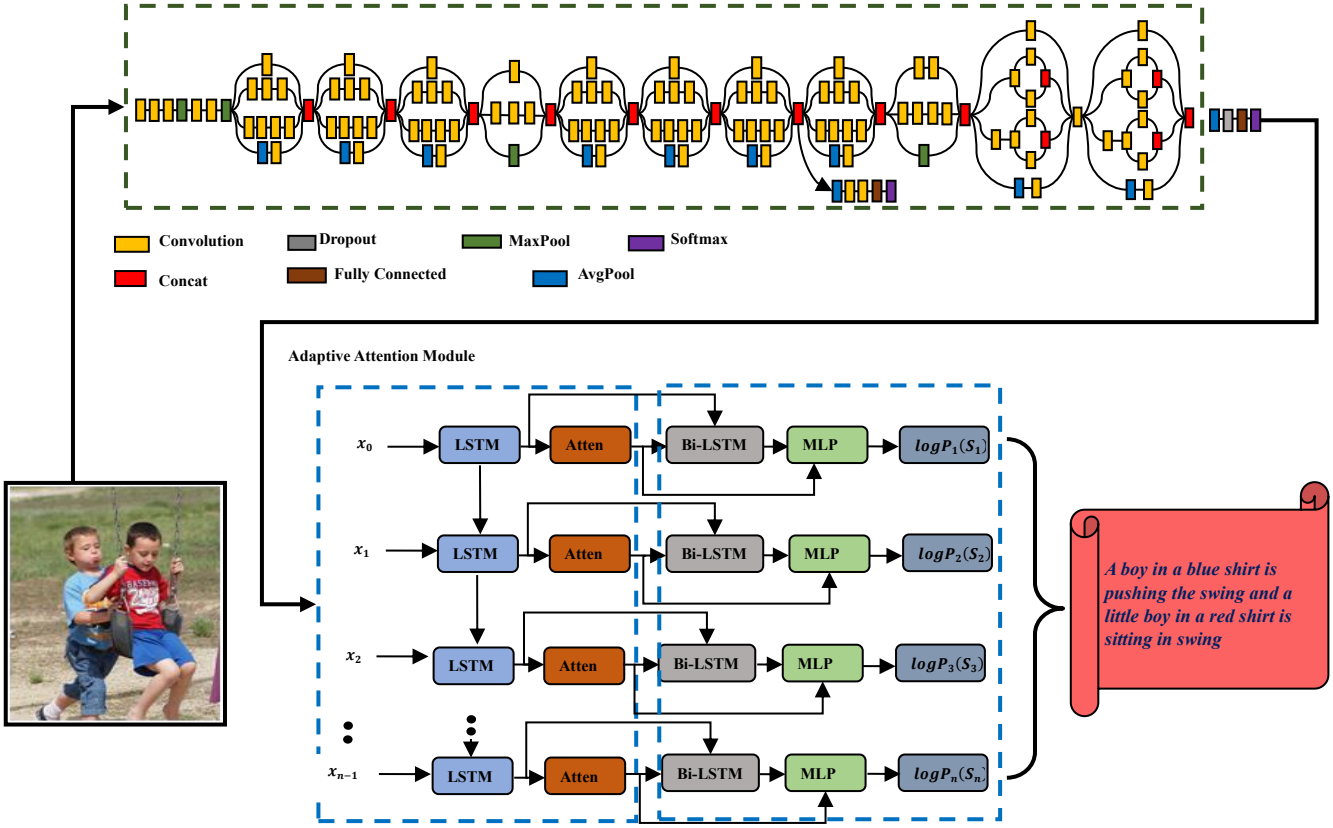


Fig. 7.14: Proposed Model Architecture

$$f_t = \sigma(x_t \times W_{xf} + h_{t-1} \times W_{hf}) \tag{7.17}$$

$$\bar{c_t} = tanh(x_t \times W_{xc} + h_{t-1} \times W_{hc}) \tag{7.18}$$

$$o_t = \sigma(x_t \times W_{xo} + h_{t-1} \times W_{ho}) \tag{7.19}$$

$$c_t = f_t \odot c_{t-1} + i_t \odot \bar{c_t} \tag{7.20}$$

$$h_t = o_t \odot \tanh(c_t) \tag{7.21}$$

x_t is the cell input, c_t is the input activation, h_t and $h_{t-1}$are the hidden state and the previous hidden state respectively, f_t is forget gate o_t is output gate, c_t and c_{t-1} are the current cell state and previous cell state and W_x and W_h are the weights of input and hidden state.

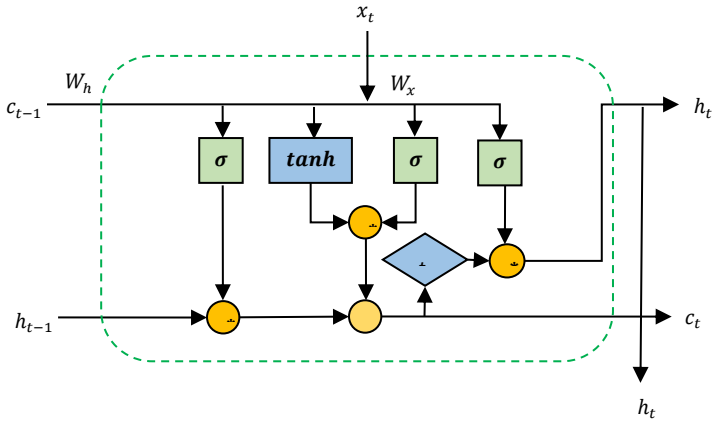


Fig.7.15: Structure of LSTM

Bidirectional LSTM, an extension of LSTM, puts two independent RNNs together that make the input flow in both directions to preserve the future and the past information. The basic structure for Bi-LSTM is depicted in Fig. 7.16. Bi-LSTMs train with two LSTMs namely the forward pass (past layer) and a backward pass (future layer). Both the activations (forward and backward) are considered to calculate the output. The usage of Bi-LSTM provides significant improvements in a network as it understands the context better way.

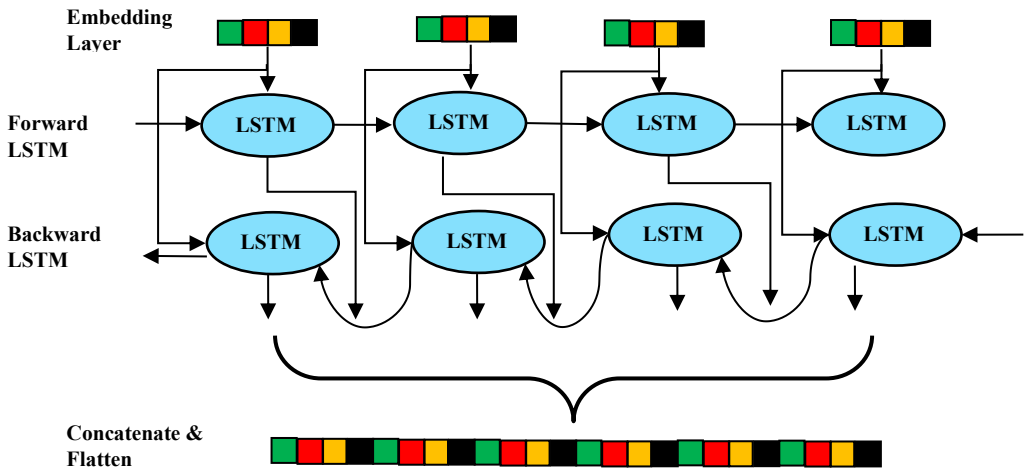


Fig. 7.16: Structure of Bi-LSTM

The attention mechanism maps distinct features of an image with the corresponding text description. Traditional models used h_{t-1} for the calculation of the

area of region at the current moment. In the proposed work, a feature map extracted from the region $c_t$is calculated with the use of h_t. Attention in common language means concentrating on one or a few things while ignoring others, therefore, the attention mechanism maps better the text descriptions with the corresponding image regions. For the model with the attention mechanism,c_t is defined by:

$$c_t = g(V, h_t) \tag{7.22}$$

where g is the attention mechanism, $V = [v_1, v_2, \ldots\ldots\ldots v_k]$are the features of image for k region, and $h_t$is the hidden state of RNN at time t. Attention distribution of k regions $b_t$is:

$$\widetilde{b_t} = w_h^T \tanh\left(W_v V + \left(W_g h_t\right) I^T\right) \tag{7.23}$$

$$b_t = softmax\left(\widetilde{b_t}\right) \tag{7.24}$$

From Equations (2) and (3), the finalc_t is given by:

$$c_t = \sum_{i=1}^{k} b_{ti} v_{ti} \tag{7.25}$$

Further, the concept of visual sentinel that is based on the spatial attention mechanism. Visual sentinel helps the model decide when to put more emphasis on image information or the language rules. Further, LSTM stores both visual information and language rules. The visual sentinel formula is given by:

$$g_t = \theta(x_t \times W_x + h_{t-1} \times W_h) \tag{7.26}$$

$$s_t = g_t \odot \tanh\left(\widetilde{h_t}\right) \tag{7.27}$$

where s_t is the visual sentinel, x_t is LSTM input, h_{t-1} is last node output, h_t and g_t

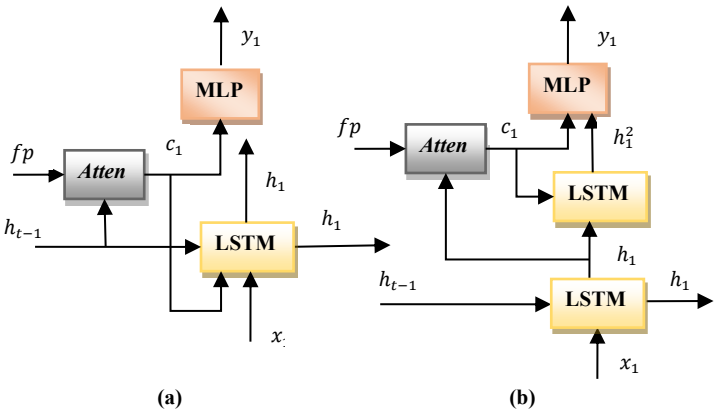


Fig. 7.17: (a) Soft-Attention model, (b) improved spatial attention model structure

are the memory cell and the sentinel gate. Sentinel gate makes the model emphasize on image or visual sentinel. The spatial adaptive attention $\widehat{c_t}$ is obtained according to the following equation:

$$\widehat{c_t} = \beta_t s_t + (1 - \beta_t) C_t \tag{7.28}$$

where, C_t is visual sentinel information and $\beta_t \in [0,1]$ and is a new sentinel gate at time t. Also,

$$\beta_t = \begin{cases} 0, means\ focusing\ only\ on\ image\ regions \\ 1, means\ focusing\ more\ on\ language\ rules \end{cases} \tag{7.29}$$

To calculate β_t, the attention score $\widetilde{b_t}$ is used. Also, b_t (attention distribution) for$k$ regions are expanded by splicing an element after $\widetilde{b_t}$:

$$\widehat{b_t} = softmax([\widetilde{b_t}; w_h^T \tanh\,(W_s s_t (W_g h_t)))] \tag{7.30}$$

$$\widehat{b_t} \in R^{k+1} \tag{7.31}$$

The expanded $\widehat{b_t}$ has $k + 1$ elements. The last bit of the new $\widehat{b_t}$calculated is assigned to β_t

$$b_t = \widehat{b_t}[k+1] \tag{7.32}$$

The probability distribution of words:

$$h_t^2 = LSTM(\widehat{c_t}, \widetilde{b_t}) \tag{7.33}$$

$$P_t = softmax(W_p(\widehat{c_t} + h_t^2) \tag{7.34}$$

Generation of captions makes use of the words that are generated from the current time node, features extracted from Inception V3 with weights, and the correct captions for the images that provide the words for the next time node. C and S input to the Bi-LSTM unit given in a forward and backward which generates an output word with of help of $softmax$ function. Further, the loss is calculated with the use of a cross-entropy function with the word corresponding to the next time node. Back Propagation Through Time (BPTT) algorithm is applied that updates different parameters of LSTM network. The model optimization objective function is given by Eq. (7.35):

$$\theta^* = arg\max_{\theta} \sum_{c,s} \sum_{t=0}^{N} \log\, p(s_t/c\,, \theta, s_0, s_1, \ldots\ldots, s_{t-1}) \tag{7.35}$$

where parameter θ is to be learned by model and n is the sequence length. C, the feature map with weight $C \in \{c_1, c_2, \ldots\ldots, c_t\}$. S is the correct description of an image, and s_i is every word vector contained in the sentence. Further, equation (7.36) defines the loss function:

$$L(c, s) = -\sum_{t=1}^{N} log\; p_t\,(s_t) \tag{7.36}$$

LSTM parameters are updated to maximize the probability of each correct word to decrease the loss function.

7.2.2 Experimental Work and Results

The experimentation for the proposed model is done on two datasets: Flickr8K and Visually Assistance Dataset [259]. Flickr8K contains 8000 images with 1000 images each for both testing and validation and 6000 images for training. Each image in the dataset contains 5 captions provided by human annotators. Visual Assistance dataset consists of 1600 different images belonging to 21 different categories. This dataset is mainly focused on the several use cases to assist visually impaired people in different ways: streets, staircase, currency detection, trash cans, bus stop, washrooms, ATM queue, etc. The proposed model performance is validated with the most common BLEU metric for the evaluation of the proposed model. The model is evaluated with respect to five different categories on two datasets namely Flickr8K and Visual Assistance.

Table 7.5: Results of the proposed Model on Visual Assistance Dataset

Method	Encoder	B-1	B-2	B-3	B-4
LSTM	VGG-16	46.9	24.4	15.2	10.1
LSTM+LA	VGG-16	57.9	32.1	26.1	15.1
LSTM+LA	Inception-V3	60.7	44.6	32.8	25.2
LSTM+AA	Inception-V3	67.2	48.6	35.1	26.2
Bi-LSTM+AA	Inception-V3	70.2	49.1	35.9	26.6

Table 7.6: Comparison Results of the proposed model on Flickr8K Dataset

Method	B-1	B-2	B-3	B-4
Xu et.al [128]	67.0	45.7	31.4	45.7
Jia et.al [103]	67.0	49.1	35.8	26.4
Vinyals et.al [18]	63.0	-	-	-
Chen et. al [124]	68.2	49.6	35.9	25.8
Zhao et.al [92]	64.5	46.2	32.7	22.7
Wang et.al [110]	65.5	46.8	32.0	21.5
InceptionV3 + AA +Bi-LSTM	71.2	51.0	36.0	20.3

BLEU-1,2,3, and 4 scores, for the visual assistance dataset, with respect to the five categories in Table 7.5 reveal that the proposed model (Inception V3+ AA + Bi-LSTM) provides improvements in the BLEU scores over other techniques. Further, the results obtained using the proposed model are compared with [18] [128] [103] [92] [124] [110] as tabulated in Table 7.6. The proposed model provides improvements in BLEU-1, BLEU-2, and BLEU-3 scores.

Fig. 7.18 (a) and 7.18 (b) represent the text descriptions generated by different categories for Flickr8K and Visual Assistance Dataset. These figures establish that the sentences generated are well correlated with objects, scenes, and their attributes though there are some complicated relationships between scenes and objects.

Inception V3 + LA + LSTM: A child with a yellow cup and milk all over his face

Inception V3 + AA + LSTM: A little boy spilled his milk from the green cup

Inception V3 + AA + Bi-LSTM: A little boy drinks milk from green cup and gets milk all over table

(a)

Inception V3 + LA + LSTM: Group of people playing musical instruments on street.

Inception V3 + AA + LSTM: Group of people playing guitar on the street.

Inception V3 + AA + Bi-LSTM: Group of people is singing songs and playing instruments on the street.

(b)

Fig. 7.18: Example of text generated (a) Flickr8K dataset (b) Visual Assistance Dataset; LA: Local Attention, AA: Adaptive Attention

7.3 Significant Outcomes

This chapter discusses applications of the visual image captioning in medical domain and for aid-to-blind. Medical image captioning highlights the relationships between image objects and clinical findings. Further, Image captioning can help visually impaired people understand their environment.

This chapter first presents an efficient two-stage framework, $FDT - Dr^2T$ for generation of dense and coherent radiology or medical reports. This efficiently learns image features from the first stage to select the highest scoring head from the structure and concatenates best-head and the multi-head representations to explore more correlations and provide better higher-order feature representations. Also, the proposed framework is able to detect changes in the contrast, size, and disease location of medical images which help in the coherent generation of medical reports. Furthermore, the proposed $FDT - Dr^2T$ framework imitates the working patterns of radiologists by highlighting the specific crucial abnormalities in reports.

The effectiveness of the proposed framework is analyzed on the IU-Chest X-ray dataset and provided state-of-the-art results in terms of $BLEU@N, METEOR, CIDEr, ROUGE - L$ evaluation metrics. To support the experimental analysis an ablation study is also carried out that exhibits the effect of different feature representation methods and generates medical reports using the baseline transformer and the proposed Dr^2T. Further, there were very few cases (around 2%-3%) observed where the proposed framework does not generate descriptions to the extent expected. These cases generated sentences with missing words and some grammatical errors. This could be alleviated by utilizing fused

character or word embedding models or by incorporating attention-based embedding models.

In future, the proposed framework can be utilized for other fatal diseased parts like breast, brain, etc. subject to the availability of their captioning datasets. Also, the proposed model can be utilized for the generation of radiology reports for 3D imaging data i.e., brain MRI dataset etc. We can incorporate different modalities of medical images for the generation of more detailed radiology reports. Furthermore, explainable AI solutions paired with examination by qualified physicians appear to be quite useful in understanding the outcomes of complex deep learning models and facilitating the evaluation process.

Furthermore, an encoder-decoder-based image captioning with an adaptive attention mechanism is discussed in this chapter. The proposed architecture used Inception-V3 for extraction of global features of an image and adaptive attention mechanism with visual sentinel to better correlate words with their corresponding image regions.

Further, Bi-LSTM at the decoder end is introduced for the generation of the required language. We have performed extensive attention evaluation to analyse our adaptive attention on Flickr8K and Visual Assistance datasets. Our model achieves state-of-the-art performance and improves the quality of caption generated with improvements in BLEU scores. The proposed model can also have many useful applications in other domains also.

Chapter-8

Conclusions and Future Scope

This chapter summarizes the research findings based on theoretical and/or experimental contributions, future research directions, and the work's potential societal and technical implications.

8.1 Conclusions

This study mainly highlights three captioning tasks namely, factual image captioning, stylized image captioning, and paragraph-based image captioning. Based upon these captioning tasks, deep-learning-based models are developed that describe the contents of the images highlighting the factual and stylized contents of images. Further, the model is developed that describes the contents of the image in the form of a paragraph. Also, the study presents the application to visual image captioning in aid-to-blind and medical domains. The models and approaches developed are as follows:

- A lightweight transformer is presented with minimal encoder and decoder transformer structure. The proposed transformer extracts multiple high-level and low-level appearance features at the encoding phase, while the decoding phase integrates a GRU layer to further enhance the language generation. Extensive experiments on MSCOCO dataset demonstrated that the proposed approach achieves significant scores as 81.0, 65.2, 65.2, 37.8, 27.9, 58.0, and 123.1 for B-1, B-2, B-3, B-4, METEOR, ROUGE-L, and CIDEr respectively. Further, qualitative analysis proves that proposed model yields captioning

results demonstrating better appearance awareness with a better language model.

- To capture higher-order interactions between objects, attributes and relationships between encoder and decoder that provide discriminative cues for generation of sentences, a novel XGL attention module is proposed which utilizes x-GELU activation. The proposed attention mechanism is further incorporated by using the XGL-Transformer model for the generation of factual image descriptions. Extensive experiments on the MSCOCO dataset validate the superiority of the proposed model in terms of CIDEr, SPICE. BLEU@n, $n \in$ [1,4], METEOR, ROUGE-L evaluation parameters. Bi-LSTM followed by GEGLU is employed in decoder that further demonstrates the improvement in the results by capturing higher order interactions between visual regions while simultaneously triggering higher order interactions across other modalities for multi-modal reasoning. Also, an ablation study is conducted that proved the improvement in the performance of the overall system showing the impact of proposed XGL attention modules by introducing them in the encoder and decoder phases.
- For the generation of style-based (romantic and humorous) description of images, a novel style-based caption generation framework, Control with Style is proposed. The proposed framework generates realistic and stylized descriptions of images with controlled text generation in two phases: (i) Refined Factual Caption Generation (RFCG), and (ii) Controlled Style Caption Generation, SE-VAE-CSCG. The framework delivers refined stylized captions even for majority of unstructured samples projected in terms of improved style

accuracy. Also, the two-phase framework helps to learn disentangled representations by lifting the word-level knowledge to sentence-level knowledge. Further, an ablation study is also conducted to support the experiments.

- To generate diverse, coherent, and meaningful paragraphs for images a multi-resolution multi-head attention and adaptive attention driven variational autoencoder-based transformer framework MrA^2VAT is proposed. The framework exploits an attention-based dual Bi-LSTM language discriminator which further utilizes the concept of language score, dissimilarity scores, and length penalty. This enables the framework to generate more diverse paragraphs with no redundant sentences. Extensive experiments are conducted on the Standford Paragraph Dataset that validates the effectiveness of the proposed framework. Furthermore, a significant rise in terms of BLEU-1 (around 5.7%) and METEOR scores (around 27.5%) is observed.
- A Transformer-based summarization framework, *UnMA*-CapSumT is also designed which integrated Factual Image Captioning and Stylized Image Captioning models to generate the summarized image captions. The proposed framework is divided into two stages: (i) generation of factual, romantic, and humorous descriptions of images using the MAA-FIC model and SF-Bi-ALSTM-based stylized image captioning model, (ii) using the descriptions generated in (i) to produce a summarized single line caption that highlights factual, romantic, and humorous contents. With the exploitation of pointer-generator network and coverage mechanism in the proposed summarization transformer, the summarized caption is free from issues related to OOV and

repetition problems thereby providing an improvement in the learning knowledge of factual content and its associated linguistic styles. Further, the improvements in the performance and interpretability of the proposed framework are enhanced with the utilization of an efficient word embedding fTA-WE. Experimental results prove that the proposed MAA-FIC, SF-Bi-ALSTM-based image captioning model, and $UnMHA - ST$ summarization transformer model provided state-of-the-art results and generates visually and grammatically correct factual, romantic, humorous, and summarized captions for a given image.

- As Automatic Visual Captioning (AVC) generates syntactically and semantically correct sentences by describing important objects, attributes, and their relationships with each other. The very nature of it makes it suitable for applications such as image retrieval [222] [223], human-robot interaction [224] [225], aid to the blind [226], and visual question answering [227] and the like. This study therefore discusses about two key application areas of the image captioning namely: medical image captioning and aid to the blind.
  - ✓ An efficient two-stage framework, $FDT - Dr^2T$ for the generation of dense and coherent radiology or medical reports. The proposed framework efficiently learns image features by detecting changes in the contrast, size, and disease location of medical images. This helps to explore more correlations and provide better higher-order feature representations which help in the coherent generation of medical reports. Furthermore, the proposed $FDT - Dr^2T$ framework imitates the working patterns of radiologists by highlighting the specific crucial abnormalities in reports. The effectiveness of

the proposed framework is analyzed on the IU-Chest X-ray dataset and provided state-of-the-art results in terms of BLEU-n, METEOR, CIDEr, ROUGE-L evaluation metrics.

- ✓ An encoder-decoder-based image captioning with an adaptive attention mechanism for aid to blind application is also presented. The proposed model exploits Inception-V3 for extraction of global features of an image and adaptive attention mechanism with visual sentinel to better correlate words with their corresponding image regions. Further, Bi-LSTM at the decoder end is introduced for the generation of the required language. The proposed model achieves state-of-the-art performance and improves the quality of caption generated with improvements in BLEU scores. The proposed model can also have many useful applications in other domains also.

8.2 Future Research Scope

Although deep-learning-based techniques have achieved remarkable progress in recent years, a robust and efficient captioning model which can generate high-quality captions for images and videos is yet to be achieved in terms of visual recognition and computational linguistics. In the future, image captioning can be improved in three aspects. The first aspect is to enhance the adaptability of the network, so that the model can pay attention to the specific situations and concerns, and generate targeted descriptions according to different situations and concerns. The second aspect is to optimize the evaluation algorithm to evaluate the quality of the output sequence of the model more accurately. The third aspect is to enhance the robustness of the model and avoid the influence of interference characteristics on the model output. Future research will probably concentrate on enhancing model accuracy, producing tailored and

context-aware captions, guaranteeing equity and openness, and expanding the system to do more difficult jobs like real-time applications or video captioning. Various promising research directions are yet to be explored fully in the future.

- More novel deep-learning based architectures (Vision transformers, graph neural networks, etc.) may be developed in future that can capture well the interactions between the objects and their relationships. thus, generating the descriptions subtle aspects of the image like specific attributes, emotions, or actions.
- Multimodal Learning-based models are yet to be explored in future. Transformer-based multimodal or fusion models may be developed that primarily focus on key parts of an image and how that visual focus maps to meaningful language. Further, joint embedding spaces are encouraged that share representations for vision and language to improve their alignment and coherence.
- Knowledge-aware captioning models may be developed in the near future that incorporate the external knowledge (e.g., knowledge graphs, cultural information, scene context) to provide captions that are contextually accurate. This may also include memory-augmented networks that can retain past information to generate more meaningful captions in dynamic contexts.
- Zero-shot and few-shot learning mechanisms develop models that can generate captions for entirely new objects or scenes without explicit training data and with only few labelled data. Therefore, to improve the performance of the captioning models by describing the images objects and situations with limited

amount of data, meta-learning, transfer learning, or self-supervised learning approaches should be encouraged.

- Current captioning models generates captions for static images, hence ignoring the dynamic interactions between human and AI. Therefore, interactive systems may be developed where human and AI collaborate to provide necessary user feedback (e.g., "Describe the background in more detail") or Systems asking clarifying questions before generating a caption (e.g., “Should I focus on people or objects?”). This human-AI feedback is necessary to refine or modify the generated captions for accuracy or specific needs.
- Multilingual captioning is still an open issue as training models that can generate captions in low-resource languages where large annotated datasets may not be available. To overcome this, cross-lingual transfer learning, multilingual models, and neural machine translation techniques should be encouraged that allow models to generalize across languages with minimal retraining.
- Explainable AI-based models may be developed that can not only generate captions but also explain why certain features in an image led to specific descriptions. Medical Image captioning and captioning of autonomous systems are the two main sensitive applications that need rationale behind the description of a particular scene for safety or validation purposes or medical diagnostic.
- Dense or Paragraph-based image captioning is one of such promising direction as it generates more elaborated descriptions of a given image. Attention mechanism has shown a prominent impact on the generation of the description

of visual contents. Models may be developed in this direction to improve the performance of existing state-of-the-art. Although success has been achieved in recent years, there is always still room for improvements for the generation of richer descriptions of images.

- Real-time caption generation for images remains the most intimidating challenge to deal with. Therefore, unsupervised learning and reinforcement-based learning may prove to be more realistic way of caption generation in real time.

- Existing Visual Captioning techniques focus on the visual description problem. It would be more interesting to think one step forward, and develop a visual understanding system such as Visual Question Answering (VQA) and Visual Reasoning. These have the potential to perform much better foreseeable future.

8.3 Future Applications

In the future, by utilizing these approaches, one can develop a variety of real-life application systems such as:

- By producing thorough descriptions of objects, environments, and activities, image captioning can aid those with visual impairments in understanding their surroundings. Furthermore, by combining object detection technologies with image captioning, devices may identify common objects, people, and even emotions, thereby enhancing the quality of life for individuals with visual impairments. A mobile app, for instance, might provide autonomous shopping, navigation, and other functions by describing scenes in real time.

- Image captioning can be used in e-commerce to automatically create product descriptions, which helps customers find things more easily through search engines by providing them with thorough visual descriptions.
- Image captioning can help social networking platforms by automatically tagging and labelling information, eliminating the need for manual annotation. Also, AI-generated captions can help identify inappropriate or harmful content, and can be helpful in removal of dangerous or inappropriate images from social media.
- Image captioning models can also be used in generating preliminary pathology reports helping doctors identify the abnormalities and suggesting potential diagnosis.
- The models can be useful for children with learning difficulties, AI could generate descriptive explanations of images to simplify complex visual information, facilitating better understanding.
- Deep-learning based image captioning models can be used by autonomous vehicles to better understand the surrounding environment and describe the same in the form of natural language sentences. Models are developed that can help in identifying pedestrians, other vehicles, or obstacles.
- In gaming or VR, image captioning could generate real-time descriptions of game environments or events, making games more accessible to players with disabilities. Also, in narrative-driven games, AI-generated captions can enhance storytelling by providing vivid descriptions that change dynamically with the game's progress.

REFERENCES

[1] F. Liu, Y. Peng and M. Rosen, "An effective deep transfer learning and information fusion framework for medical visual question answering," in *Cross-Language Evaluation Forum for European Languages*, Lugano, Switzerland, 2019.

[2] Z. Xu, L. Mei, Z. Lv, C. Hu, X. Luo, H. Zhang and Y. Liu, "Multi-Modal Description of Public Safety Events Using Surveillance and Social Media," *IEEE Transaction on Big Data,* vol. 5, no. 4, pp. 529-539, January 2017.

[3] W. Yang, "Analysis of sports image detection technology based on machine learning," *EURASIP Journal on Image and Video Processing,* vol. 17, January 2019.

[4] L. Bergman and Y. Hoshen, "Classification-based Anomaly detection for general data," in *arXiv:2005.02359*, 2020.

[5] R. Bernardi, R. Cakici, D. Elliott, A. Erdem, E. Erdem, N. Ikizler-Cinbis, F. Keller, A. Muscat and B. Plank, "Automatic Description Generation from Images: A Survey of Models, Datasets, and Evaluation Measures," *Journal of Artificial Intelligence Research,* vol. 55, pp. 409-442, April 2016.

[6] A. Singh, T. Doren and S. Bandyo, "A Comprehensive Review on Recent Methods and Challenges of Video Description," *arXiv:2011.14752v1,* November 2020.

[7] A. Farhadi, M. Hejrati, M. Sadeghi, P. Young, C. Rashtchian, J. Hockenmaier and D. Forsyth, "Every picture tells a story: Generating sentences from images," in *Proceedings of the European Conference on Computer Vision*, Crete, Greece , 2010.

[8] A. Kojima, M. Izumi, T. Tamura and K. Fukunaga, “Generating natural language description of human behavior from video images,” in *Proceedings 15th International Conference on Pattern Recognition.*, Barcelona, Spain, 2000.

[9] M. Nivedita, P. Chandrashekar, S. Mahapatra and A. Phamila, “Image Captioning for Video Surveillance System using Neural Networks,” *International Journal of Image and Graphics,* vol. 21, no. 4, March 2021.

[10] J. Redmon and A. Farahadi, “‘YOLOv3: An incremental improvement’’,” in *arXiv:1804.02767*, 2018.

[11] M. Weiss, S. Chamorro, R. Girgis, M. Luck, S. Kahou, J. Cohen, D. Nowrouzezahrai, D. Precup, F. Golemo and C. Pal, “Navigation agents for the visually impaired: A sidewalk simulator and experiments,” in *arXiv:1910.13249*, 2019.

[12] J. Pavlopoulos, V. Kougia and I. Androutsopo, “A Survey on Biomedical Image Captioning,” in *Association for Computational Linguistics*, Minneapolis, Minnesota, 2019.

[13] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth and B. Schiele, “The cityscapes dataset for semantic urban scene understanding,” in *IEEE Conference on Computer Vision Pattern Recognition (CVPR)*, Caesars Palace, 2016.

[14] J. Long, E. Shelhamer and T. Darrell., “Fully convolutional networks for semantic segmentation,” in *arXiv:1411.4038* , 2015.

[15] S. Ren, K. He, R. Girshick and J. Sun, “Faster R-CNN: Towards real-time object detection with region proposal networks,” *Advances in Neural Information Processing Systems,* vol. 28, pp. 91-99, December 2015.

[16] K. He, X. Zhang, S. Ren and J. Sun, “Deep residual learning for image recognition,” in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, 2016.

[17] M. Kristan, J. Matas, A. Leonardis, M. Felsberg, L. Cehovin, G. Fernandez, T. Vojir, G. Hager, G. Nebehay and R. Pflugfelder, “The visual object tracking vot2015 challenge results,” in *International Conference on Computer Vision Workshops (ICCV Workshops)*, Santiago, Chile, 2015.

[18] O. Vinyals, A. Toshev, S. Bengio and D. Erhan, “Show and Tell: A Neural Image Caption Generator,” in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* , Boston, USA, 2015.

[19] A. Karpathy and L. Fei-Fei, “Deep visual-semantic alignments for generating image descriptions,” in *IEEE Conference on Computer Vision and Pattern Recognition*, Boston, 2015.

[20] M. Pedersoli, T. Lucas, C. Schmid and J. Verbeek, “Areas of Attention for Image Captioning,” in *arXiv:1612.01033v2*, 2017.

[21] X. Long, C. Gan and G. d. Melo, “Video Captioning with Multi-Faceted Attention,” *Transactions of the Association for Computational Linguistics,* vol. 6, no. 1, p. 173–184, December 2016.

[22] T. Yao, Y. Pan, Y. Li, Z. Qiu and T. Mei, “Boosting image captioning with attributes,” in *International Conference on Computer Vision (ICCV)*, Venice, 2017.

[23] M. Rohrbach, S. Amin, M. Andriluka and B. Schiele, “A database for fine grained activity detection of cooking activities,” in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Providence, RI, USA, 2012.

[24] A. Miech, D. Zhukov, J.-B. Alayrac, M. Tapaswi, I. Laptev and J. Sivic, “HowTo100M: Learning a Text-Video Embedding by Watching Hundred Million Narrated Video Clips,” in *arXiv:1906.03327v2*, 2019.

[25] R. Vedantam, C. L. Zitnick and D. Parikh, “Cider: Consensus-based image description evaluation,” in *IEEE Conference on Computer Vision and Pattern Recognition*, Boston, 2015.

[26] S. Banerjee and A. Lavie, “METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments,” in *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, Ann Arbor, Michigan, 2005.

[27] M. Hodosh, P. Young and J. Hockenmaier, “Framing image description as a ranking task: Data, models and evaluation metrics,” *Journal of Artificial Intelligence Research,* vol. 47, pp. 853-899, August 2013.

[28] M. Mitchell, J. Dodge, A. Goyal, K. Yamaguchi, K. Stratos, X. Han, A. Mensch, A. Berg, T. Berg and H. Daume, “Generating image descriptions from computer vision detections,” in *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, Avignon , 2012.

[29] Y. Yang, C. L. Teo, H. Daume and Y. Aloimono, “Corpus-guided sentence generation of natural images,” in *Empirical Methods in Natural Language Processing*, Edinburgh United Kingdom, 2011.

[30] Y. Cheng, F. Huang, L. Zhou, C. Jin, Y. Zhang and T. Zhang, “A Hierarchical Multimodal Attention-based Neural Network for Image Captioning,” in *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information*, Shinjuku, Tokyo, Japan, 2017.

[31] J. Lu, C. Xiong, D. Parikh and R. Socher, “Knowing When to Look: Adaptive Attention via A Visual Sentinel for Image Captioning,” in *arXiv:1612.01887v2*, 2017.

[32] V. Ordonez, G. Kulkarni and T. L. Berg, “Im2Text: describing images using 1 million,” in *Proceedings of the Advances in Neural Information Processing Systems*, 2011.

[33] F. R. Bach and M. I. Jordan, “Kernel independent component analysis,” *Journal of Machine Learning,* vol. 3, pp. 1-48, July 2002.

[34] D. R. Hardoon, S. R. Szedmak and J. R. Shawe-Taylor, “Canonical correlation analysis: An overview with application to learning methods,” *Neural Computation,* vol. 16, no. 12, pp. 2639-2664, December 2004.

[35] A. Gupta, Y. Verma and C. V. Jawahar, “Choosing linguistics over vision to describe images,” in *Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence*, Toronto, Ontario, Canada, 2012.

[36] P. Kuznetsova, V. Ordonez, T. Berg and Y. Choi, “Treetalk: Composition and compression of trees for image descriptions,” *Transaction of Association for Computational Linguistics,* vol. 10, no. 2, pp. 351-362, 2014.

[37] P. F. Felzenszwalb, R. B. Girshick, D. McAllester and D. Ramanan, “Object detection with discriminatively trained part based models,” *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 32, no. 9, p. 1627–1645, September 2010.

[38] A. Oliva and A. Torralba, “Modeling the shape of the scene: a holistic representation of the spatial envelope,” *International Journal of Computer Vision,* vol. 42, no. 3, pp. 145-175, May 2001.

[39] T. Dunning, “Accurate methods for the statistics of surprise and coincidence,” *Computational Linguistics,* vol. 19, no. 1, pp. 61-74, March 1993.

[40] S. Li, G. Kulkarni, T. L. Berg, A. C. Berg and Y. Cho, “Composing simple image descriptions using web-scale n-gram,” in *Fifteenth Conference on Computational Natural Language Learning*, Portland, Oregon, USA, 2011.

[41] G. Kulkarni, V. Premraj, V. Ordonez, S. Dhar, S. Li, Y. Choi, A. Berg and T. Berg, “Babytalk: Understanding and generating simple image descriptions,” *IEEE Transactions on Pattern Analysis and Machine Intellegence,* vol. 35, no. 12, pp. 2891-2903, June 2013.

[42] Y. Ushiku, T. Harada and Y. Kuniyoshi, “Efficient Image Annotation for Automatic Sentence Generation,” in *Proceedings of the 20th ACM International Conference on Multimedia*, Nara, Japan, 2012.

[43] L. A. Hendricks, S. Venugopalan, M. Rohrbach, R. Mooney, K. Saenko and T. Darrel, “Deep compositional captioning: Describing novel object categories,” in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Caesars Palace, 2016.

[44] T. Yao, Y. Pan, Y. Li and T. Mei, “Incorporating copying mechanism in image captioning for learning novel objects,” in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, Hawaii, 2017.

[45] S. Venugopalan, L. Hendricks, . M. Rohrbach, . R. Mooney, T. Darrell and K. Saenko, “Captioning images with diverse objects,” in *arXiv preprint arXiv:1606.07770*, 2016.

[46] Y. Wu, L. Zhu, L. Jiang and Y. Yang, “Decoupled Novel Object Captioner,” in *MM '18: Proceedings of the 26th ACM international conference on Multimedia*, Seoul, Korea, 2018.

[47] Y. Li, T. Yao, Y. Pan, H. Chao and T. Mei, “Pointing Novel Objects in Image Captioning,” in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Long Beach, CA, 2019.

[48] H. Agrawal, K. Desai, Y. Wang, X. Chen, R. Jain, M. Johnson, D. Batra, D. Parikh, S. Lee and P. Anderson, “nocaps: novel object captioning at scale,” in *IEEE/CVF International Conference on Computer Vision (ICCV)*, Seoul, Korea, 2019.

[49] Q. Feng, Y. Wu, H. Fan, C. Yan, M. Xu and Y. Yang, "Cascaded Revision Network for Novel Object Captioning," *arXiv:1908.02726,* August 2019.

[50] X. Hu, X. Yin, K. Lin, L. Wang, L. Zhang, J. Gao and Z. Liu, "VIVO: Visual Vocabulary Pre-Training for Novel Object Captioning," in *arXiv:2009.13682v2*, 2021.

[51] C. Gan, Z. Gan, X. He and J. Gao, "Stylenet: Generating attractive visual captions with styles," in *IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, Hawaii, 2017.

[52] A. P. Mathews, L. Xie and X. He, "SentiCap: Generating Image Descriptions with Sentiments," in *AAAI'16: Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, Phoenix, Arizona, 2016.

[53] L. Guo, J. Liu, P. Yao, J. Li and H. Lu, "MSCap: Multi-Style Image Captioning with Unpaired Stylized Text," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Long Beach Convention & Entertainment Center, 2019.

[54] T. Chen, Z. Zhang, Q. You, C. Fang, Z. Wang, H. Jin and J. Luo, ""Factual" or "Emotional": Stylized Image Captioning with Adaptive Learning and Attention," in *arXiv:1807.03871*, 2018.

[55] C.-K. Chen, Z. F. Pan, M. Sun and M.-Y. Liu, "Unsupervised Stylish Image Description Generation via Domain Layer Norm," in *arXiv:1809.06214v1*, 2018.

[56] W. Zhao, X. Wu and X. Zhang, "MemCap: Memorizing Style Knowledge for Image Captioning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, California USA, 2020.

[57] M. Heidari, M. Ghatee, A. Nickabadi and A. P. Nezhad, “DIVERSE AND STYLED IMAGE CAPTIONING USING SVD-BASED MIXTURE OF RECURRENT EXPERTS,” in *arXiv:2007.03338v1*, 2020.

[58] G. Li, Y. Zhai, Z. Lin and Y. Zhang, “Similar Scenes arouse Similar Emotions: Parallel Data Augmentation for Stylized Image Captioning,” in *MM '21: Proceedings of the 29th ACM International Conference on Multimedia*, Virtual Event, China, 2021.

[59] J. Johnson, A. Karpathy and L. Fei-Fei., “Densecap: Fully convolutional localization networks for dense captioning,” in *IEEE Conference on Computer Vision and Pattern Recognition*, Caesars Palace, 2016.

[60] L. Yang, K. Tang, J. Yang and L.-J. Li, “Dense Captioning with Joint Inference and Visual Context,” in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Caesars Palace, 2016.

[61] X. Xiao, L. Wang, K. Ding, S. Xiang and C. Pan, “Dense semantic embedding network for image captioning,” *Pattern Recognition,* vol. 90, pp. 285-296, June 2019.

[62] D.-J. Kim, T.-H. Oh, J. Choi and I. S. Kweon, “Dense Relational Image Captioning via Multi-task Triple-Stream Networks,” in *arXiv:2010.03855v2*, 2020.

[63] D.-J. Kim, J. Choi, T.-H. Oh and I. S. Kweon, “Dense Relational Captioning: Triple-Stream Networks for Relationship-Based Captioning,” in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[64] Z. Zhang, Y. Zhang, Y. Shi, W. Yu, L. Nie, G. He, Y. Fan and Z. Yang, “Dense Image Captioning Based on Precise Feature Extraction,” in *International Conference on Neural Information Processing*, Sydney, Australia, 2019.

[65] J. Krause, J. Johnson, R. Krishna and F. Li, “A Hierarchical Approach for Generating Descriptive Image Paragraphs,” in *arXiv:1611.06607*, 2016.

[66] Z. Wang, Y. Luo, Y. Li, Z. Huang and H. Yin, “Look Deeper See Richer:Depth-aware Image Paragraph Captioning,” in *Proceedings of the 26th ACM international conference on Multimedia*, Seoul, Korea, 2018.

[67] L. M. Kyriazi, G. Han and A. M. Rush, “Training for Diversity in Image Paragraph Captioning,” in *Conference on Empirical Methods in Natural Language Processing*, Brussels, Belgium, 2018.

[68] Z.-J. Zha, D. Liu, H. Zhang, Y. Zhang and F. Wu, “Context-Aware Visual Policy Network for Fine-Grained Image Captioning,” *IEEE Transactions on Pattern Analysis and Machine Intellegence,* October 2018.

[69] R. Li, H. Liang, Y. Shi, F. Feng and X. Wang, “Dual-CNN: A Convolutional language decoder for paragraph image captioning,” *Neurocomputing,* vol. 396, pp. 92-101, July 2020.

[70] M. Cornia, M. Stefanini, L. Baraldi and R. Cucchiara, “Meshed-Memory Transformer for Image Captioning,” in *arXiv:1912.08226v2*, 2020.

[71] D. Guo, R. Lu, B. Chen and Z. Zeng, “Matching Visual Features to Hierarchical Semantic Topics for Image Paragraph Captioning,” in *arXiv:2105.04143v1*, 2021.

[72] N. Ilinykh and S. Dobnik, “When an Image Tells a Story: The Role of Visual and Semantic Information for Generating Paragraph Descriptions,” in *13th International Conference on Natural Language Generation*, Helix, Dublin City University, DCU, 2020.

[73] X. Yang, C. Gao, H. Zhang and J. Cai, “Hierarchical Scene Graph Encoder-Decoder for Image Paragraph Captioning,” in *MM '20: Proceedings of the 28th ACM International Conference on Multimedia*, Seattle WA USA, 2020.

[74] L.-C. Yang, C.-Y. Yang and J. Y.-j. Hsu, “Object Relation Attention for Image Paragraph Captioning,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, Virtual Confrence, 2021.

[75] D. H. Park, T. Darrell and A. Rohrbach, “Robust Change Captioning,” in *arXiv:1901.02527v2*, 2019.

[76] Z. Liu, G. Li, G. Mercier, Y. He and Q. Pan, “Change detection in heterogenous remote sensing images via homogeneous pixel transformation,” *IEEE Transactions on Image Processing,* vol. 27, no. 4, p. 1822–1834, December 2018.

[77] J. Tian, S. Cui and P. Reinartz, “Building change detection based on satellite stereo imagery and digital surface models,” *IEEE Transactions on Geoscience and Remote Sensing,* vol. 52, pp. 406-417, January 2014.

[78] L. Gueguen and R. Hamid, “Large-scale damage detection using satellite imagery,” in *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Boston, USA, 2015.

[79] S. H. Khan, X. He, F. Porikli and M. Bennamoun, “Forest change detection in incomplete satellite images with deep neural networks,” *IEEE Transactions on Geoscience and Remote Sensing,* vol. 55, p. 5407–5423, June 2017.

[80] P. F. Alcantarilla, S. Stent, G. Ros, R. Arroyo and R. Gherardi, “ Street-view change detection with deconvolutional networks,” *Autonomous Robots,* vol. 42, no. 7, pp. 1301-1322, 2018.

[81] K. Sakurada, W. Wang, N. Kawaguchi and R. Nakamur, “Dense optical flow based change detection network robust to difference of camera viewpoints,” in *arXiv:1712.02941*, 2017.

[82] W. Feng, F.-P. Tian, Q. Zhang, N. Zhang, L. Wan and J. Sun, “Fine-grained change detection of misaligned scenes with varied illuminations,” in *International Conference on Computer Vision (ICCV)*, Santiago, Chile, 2015.

[83] R. Huang, W. Feng, Z. Wang, M. Fan, L. Wan and J. Sun, “Learning to detect fine-grained change under variant imaging conditions,” in *International Conference on Computer Vision Workshops (ICCV Workshops)*, Venice, Italy, 2017.

[84] S. Stent, R. Gherardi, B. Stenger and R. Cipolla, “Precise deterministic change detection for smooth surfaces,” in *In Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV)*, Lake Placid, USA, 2016.

[85] X. Shi, X. Yang, J. Gu, S. Joty and J. Cai, “Finding It at Another Side: A Viewpoint-Adapted Matching Encoder for Change Captioning,” in *arXiv:2009.14352v1*, 2020.

[86] Y. Tu, T. Yao, L. Li, J. Lou, S. Gao, Z. Yu and C. Yan, “Semantic Relation-aware Difference Representation Learning for Change Captioning,” in *Findings of the Association for Computational Linguistics: ACL-IJCNLP*, Online, 2021.

[87] M. Hosseinzadeh and Y. Wang, “Image Change Captioning by Learning from an Auxiliary Task,” in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Nashville, TN, USA, 2021.

[88] H. Kim, J. Kim, H. Lee, H. Park and G. Kim, “Viewpoint-Agnostic Change Captioning with Cycle Consistency,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, Montreal, 2021.

[89] Y. Tu, L. Li, C. Yan, S. Gao and Z. Yu, “R3Net:Relation-embedded Representation Reconstruction Network for Change Captioning,” in *arXiv:2110.10328v1*, 2021.

[90] Y. Qiu, Y. Satoh, R. Suzuki, K. Iwata and H. Kataoka, “3D-Aware Scene Change Captioning From Multiview Images,” *IEEE Robotics and Automation Letters,* pp. 2377-3766, 2020.

[91] Y. Qiu, Y. Satoh, R. Suzuki, K. Iwata and H. Kataoka, “Indoor Scene Change Captioning Based on Multimodality Data,” *Sensor Signal and Information Processing III,* vol. 20, no. 17, pp. 1-18, August 2020.

[92] D. Zhao, Z. Chang and S. Guo, “A multimodal fusion approach for image captioning,” *Neurocomputing,* vol. 329, pp. 476-485, Feb 2019.

[93] R. Kiros, R. Salakhutdinov and R. Zemel, “Multimodal Neural Language Models,” in *Proceedings of the 31st International Conference on Machine Learning (PMLR)*, Bejing, China, 2014.

[94] R. Kiros, R. Salakhutdinov and R. S. Zemel, “Unifying visual-semantic embeddings with multimodal neural language models,” *arXiv:1411.2539v1,* pp. 1-13, 2014.

[95] A. Karpathy, A. Joulin and F.-F. Li, “Deep fragment embeddings for bidirectional image sentence mapping,” in *Advances in neural information processing systems*, Montreal, Canada, 2014.

[96] J. Mao, W. Xu, Y. Yang, J. Wang, Z. Huang and A. Yuille, “Deep captioning with multimodal recurrent neural networks (m-rnn),” in *arXiv:1412.6632*, 2015.

[97] X. Chen and C. L. Zitnick, “Mind’s eye: A recurrent visual representation for image caption generation,” in *IEEE conference on computer vision and pattern recognition*, Boston, USA, 2015.

[98] C. Liu, F. Sun, C. Wang, F. Wang and A. Yuille, “MAT: A Multimodal Attentive Translator for Image Captioning,” in *arXiv:1702.05658v3*, 2017.

[99] X. Xian and Y. Tian, “Self-Guiding Multimodal LSTM—When We Do Not Have a Perfect Training Dataset for Image Captioning,” *IEEE Transactions on Image Processing,* vol. 28, no. 11, pp. 5241 - 5252, May 2019.

[100] J. Chen and H. Zhuge, “A News Image Captioning Approach Based on Multi-Modal Pointer-Generator Network,” *Concurrency and Computation Practice and Experience,* 2020.

[101] N. Kalchbrenner and P. Blunsom, “Recurrent Continuous Translation Models,” in *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, Seattle, Washington, USA, 2013.

[102] K. Cho, D. Bahdanau, F. Bougares, H. Schwenk and Y. Bengio, “Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation,” in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Doha Qatar, 2014.

[103] X. Jia, E. Gavves, B. Fernando and T. Tuytelaars, “Guiding the long-short term memory model for image caption generation,” in *IEEE International Conference on Computer Vision*, Santiago, Chile, 2015.

[104] D. Bahdanau, K. Cho and Y. Bengio, “Neural machine translation by jointly learning to align and translate,” in *arXiv:1409.0473*, 2015.

[105] K. Cho, B. V. Merriënboer, D. Bahdanau and Y. Bengio, “On the properties of neural machine translation: Encoder-decoder approaches,” in *In Association for Computational Linguistics*, Doha, Qatar, 2014.

[106] J. Donahue, L. Hendricks, M. Rohrbach, . S. Venugopalan, . S. Guadarrama, K. Saenko and . T. Darrell, “ Long-Term Recurrent Convolutional Networks for Visual Recognition and Description,” *IEEE Transactions on Pattern Analysis and Machine Intelligence,* pp. 677-691, 2015.

[107] J. Mao, J. Huang, A. Toshev, O. Camburu, A. L. Yuille and K. Murphy, "Generation and comprehension of unambiguous object descriptions," in *IEEE conference on computer vision and pattern recognition*, Caesars Palace, 2016.

[108] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," in *arXiv:1409.1556*, 2014.

[109] A. Krizhevsky, I. Sutskever and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Netwroks," *Communications of the ACM,* vol. 60, no. 6, pp. 84-90, June 2017.

[110] C. Wang, H. Yang and C. Meinel, "Image Captioning with Deep Bidirectional LSTMs," *ACM Transactions on Multimedia Computing, Communications, and Applications,* vol. 14, no. 2s, pp. 1-20, 2018.

[111] Q. Wu, C. Shen, L. Liu, A. Dick and A. v. d. Hengel, "What Value Do Explicit High Level Concepts Have in Vision to Language Problems?," in *arXiv:1506.01144v6*, 2016.

[112] Y. Pu, Z. Gan, R. Henao, X. Yuan, C. Li, A. Stevens and L. Carin, "Variational Autoencoder for Deep Learning of Images, Labels and Captions," in *NIPS'16: Proceedings of the 30th International Conference on Neural Information Processing Systems*, Barcelona, Spain, 2016.

[113] Y. Pu, X. Yuan, A. Stevens, C. Li and L. Carin, "A deep generative deconvolutional image model," in *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics (AISTATS)*, Cadiz, Spain, 2016.

[114] T. Yao, Y. Pan, Y. Li and T. Mei, "Exploring visual relationship for image captioning," in *ECCV*, 2018.

[115] Z. Ren, X. Wang, N. Zhang, X. Lv and L.-J. Li, “Deep Reinforcement Learning-based Image Captioning with Embedding Reward,” in *arXiv:1704.03899v1*, 2017.

[116] B. Dai, S. Fidler, R. Urtasun and D. Lin, “Towards Diverse and Natural Image Descriptions via a Conditional GAN,” in *arXiv:1703.06029v3*, 2017.

[117] C. Chen, S. Mu, W. Xiao, Z. Ye, L. Wu and Q. Ju, “Improving Image Captioning with Conditional Generative Adversarial Nets,” in *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence*, Hawaii, USA, 2019.

[118] S. J. Rennie, E. Marcheret, Y. Mroueh, J. Ross and V. Goel, “Self-critical sequence training for image captioning,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[119] S. Herdade, A. Kappeler, K. Boakye and J. Soares, “Image Captioning: Transforming Objects into Words,” in *arXiv:1906.05963v2*, 2020.

[120] N. Patwari and D. Naik, “En-De-Cap: An Encoder Decoder model for Image Captioning,” in *International Conference on Computing Methodologies and Communication (ICCMC)*, Tamil Naddu, India, 2021.

[121] S. K. Mishra, R. Dhir, S. Saha, P. Bhattacharyya and A. K. Singh, “Image captioning in Hindi language using transformer networks,” *Computers & Electrical Engineering,* vol. 92, June 2021.

[122] J. Qiu, F. P.-W. Lo, X. Gu, M. L. Jobarteh, W. Jia and T. Baranowski, “Egocentric Image Captioning for Privacy-Preserved Passive Dietary Intake Monitoring,” in *arXiv:2107.00372v1*, 2021.

[123] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould and L. Zhang, “Bottom-up and top-down attention for image captioning and visual question answering,” in *IEEE Conference on Computer Vision and Pattern Recognition*, Salt Lake City, 2018.

[124] L. Chen, H. Zhang, J. Xiao, L. Nie, J. Shao, W. Liu and T.-S. Chua, “SCA-CNN: Spatial and Channel-wise Attention in Convolutional Networks for Image Captioning,” in *arXiv:1611.05594v2*, 2017.

[125] Z. Deng, Z. Jiang, R. Lan, W. Huang and X. Luo, “Image captioning using DenseNet network and Adaptive Attention,” *Signal Processing: Image Communication,* vol. 85, p. 115836, 2020.

[126] Z. Zhang, Q. Wu, Y. Wang and F. Chen, “Exploring region relationships implicitly: Image captioning with visual relationship attention,” *Image and Vision Computing,* vol. 109, May 2021.

[127] Q. You, H. Jin, Z. Wang, C. Fang and J. Luo, “Image captioning with semantic attention,” in *IEEE Conference on Computer Vision and Pattern Recognition*, Caesars Palace, 2016.

[128] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudino, R. Zemel and Y. Bengio, “Show, attend and tell: Neural image caption generation with visual attention,” in *Proceedings of the 32nd International Conference on Machine Learning (PMLR)*, Lille, France, 2015.

[129] J. Jin, K. Fu, R. Cui, F. Sha and C. Zhang, “ Aligning where to see and what to tell: image caption with region-based attention and scene factorization,” in *arXiv preprint arXiv:1506.06272.*, 2015.

[130] Z. Yang, Y. Yuan, Y. Wu, R. Salakhutdinov and W. W. Cohen, “Encode, Review, and Decode: Reviewer Module for Caption Generation,” in *arXiv:1605.07912*, 2016.

[131] C. Liu, J. MAo, F. Sha and A. Yuille, “Attention Correctness in Neural Image Captioning,” in *AAAI'17: Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, California, USA, 2017.

[132] H. R. Tavakoli, R. Shetty, A. Borji and J. Laaksonen, “Paying Attention to Descriptions Generated by Image Captioning Models,” in *arXiv:1704.07434v3*, 2017.

[133] C. C. Park, B. Kim and G. Kim, “Attend to You: Personalized Image Captioning with Context Sequence Memory Networks,” in *arXiv:1704.06485v2*, 2017.

[134] L. Huang, W. Wang, J. Chen and X.-Y. Wei, “Attention on Attention for Image Captioning,” in *IEEE International Conference on Computer Vision*, Coex, 2019.

[135] M. Liu, L. Li, H. Hu, W. Guan and J. Tian, “Image caption generation with dual attention mechanism,” *Image Processing and Management,* vol. 57, no. 2, p. 102178, March 2020.

[136] C. Yan, Y. Hao, L. Li, J. Yin, A. Liu, Z. Mao, Z. Chen and X. Gao, “Task-Adaptive Attention for Image Captioning,” *IEEE Transactions on Circuits and Systems for Video Technology,* March 2021.

[137] Q. Wu, C. Shen, P. Wang, A. Dick and A. v. d. Hengel, “Image captioning and visual question answering based on attributes and external knowledge,” *IEEE transactions on pattern analysis and machine intelligence,* vol. 40, no. 6, pp. 1367-1381, March 2018.

[138] L. Gao, B. Wang and W. Wang, “Image Captioning with Scene-graph Based Semantic Concepts,” in *ICMLC 2018: Proceedings of the 2018 10th International Conference on Machine Learning and Computing*, Macau, China, 2018.

[139] S. Tripathi, K. Nguyen, T. Guha, B. Du and T. Q. Nguyen, “SG2Caps: Revisiting Scene Graphs for Image Captioning,” in *arXiv:2102.04990v1*, 2021.

[140] X. Liu and Q. Xu, “Adaptive Attention-based High-level Semantic Introduction for Image Caption,” *ACM Transactions on Multimedia Computing, Communications, and Applications,* vol. 16, no. 4, pp. 1-22, Dec 2020.

[141] Z. Shi, X. Zhou, X. Qiu and X. Zhu, “Improving Image Captioning with Better Use of Captions,” in *arXiv:2006.11807v1*, 2020.

[142] X. Zhang, S. He, X. Song, R. W. Lau, J. Jiao and Q. Ye, “Image captioning via semantic element embedding,” *Neurocomputing,* vol. 395, pp. 212-221, June 2020.

[143] H. Fang, S. Gupta, F. Iandola, R. K. Srivastava, L. Deng, P. Dollar, J. Gao, X. He, M. Mitchell, J. C. Platt, C. L. Zitnick and G. Zweig, “From Captions to Visual Concepts and Back,” in *arXiv:1411.4952v3*, 2016.

[144] K. Tran, X. He, L. Zhang, J. Sun, C. Carapcea, C. Thrasher, C. Buehler and C. Sienkiewicz, “Rich Image Captioning in the Wild,” in *arXiv:1603.09016v2*, 2016.

[145] S. Ma and Y. Han, “Describing images by feeding LSTM with structural words,” in *International Conference on Multimedia and Expo (ICME)*, Hamburg, 2016.

[146] M. Wang, L. Song, X. Yang and C. Luo, “A parallel-fusion RNN-LSTM architecture for image caption generation,” in *International Conference on Image Processing (ICIP)*, Phoenix, Arizona, 2016.

[147] F. Tan, S. Feng and V. Ordonez, “Text2Scene: Generating Compositional Scenes from Textual Descriptions,” in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Long Beach, CA, 2019.

[148] M. Nikolaus, M. Abdou, M. Lamm, R. Aralikatte and D. Elliott, “Compositional Generalization in Image Captioning,” in *arXiv:1909.04402v2*, 2019.

[149] J. Tian and J. Oh, “Image Captioning with Compositional Neural Module Networks,” in *arXiv:2007.05608v1*, 2020.

[150] E. Bugliarello and D. Elliott, “The Role of Syntactic Planning in Compositional Image Captioning,” in *arXiv:2101.11911v1*, 2021.

[151] P. K, R. S, W. T and Z. W.J, “IBM Research Report Bleu: a method for automatic evaluation of machine translation,” *ACL Proceedings of Annual Meeting of the Association for Computational Linguistics,* vol. 30, pp. 311-318, 2002.

[152] C.-Y. Lin, “ROUGE: A Package for Automatic Evaluation of Summaries,” *Association for Computational Linguistics,* vol. Text Summarization Branches Out, pp. 74-81, 2004.

[153] P. Anderson, B. Fernando, M. Johnson and S. Gould, “SPICE: Semantic Propositional Image Caption Evaluation,” in *arXiv:1607.08822*, 2016.

[154] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser and . I. Polosukhin, “Attention is all you need,” in *NIPS*, 2017.

[155] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens and Z. Wojna, “Rethinking the Inception Architecture for Computer Vision,” in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, 2016.

[156] C. Junyoung, C. Gulcehre, K. Cho and Y. Bengio, “Empirical evaluation of gated recurrent neural networks on sequence modeling.,” in *arXiv preprint arXiv:1412.3555*, 2014.

[157] T. Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollar and C. Lawrence, “Microsoft coco: Common objects in context,” in *European Confrence on Computer Vision*, Zurich, Switzerland, 2014.

[158] W. Jiang, L. Ma, Y.-G. Jiang, W. Liu and T. Zhang, “Reccurent fusion network for image captioning,” in *arXiv:1807.09986*, 2018.

[159] X. Yang, K. Tang, H. Zhang and J. Cai, “Auto-encoding scene graphs for image captioning,” in *IEEE Conference on Computer Vision and Pattern Recognition*, Long Beach, CA, 2019.

[160] Z. Yang, Y. Zhang, S. Rehman and Y. Huang, “Image Captioning with Object Detection and Localization,” in *arXiv:1706.02430*, 2017.

[161] G. Li, L. Zhu, P. Liu and Y. Yang, “Entangled transformer for image captioning,” in *IEEE Conference on Computer Vision and Pattern Recognition*, Long Beach, CA, 2019.

[162] J. Yu, J. Li, Z. Yu and Q. Huang, “Multimodal transformer with multiview visual representation for image captioning,” *IEEE Transactions on Circuits and Systems for Video Technology,* vol. 30, no. 12, pp. 4467-4480, 2020.

[163] S. He, W. Liao, H. R. Tavakoli, M. Yang, B. Rosenhahn and N. Pugeault, “Image Captioning through Image Transformer,” in *Asian Conference on Computer Vision*, 2020.

[164] Y. Pan, T. Yao, Y. Li and T. Mei, “X-Linear Attention Networks for Image Captioning,” in *arXiv:2003.14080v1*, 2020.

[165] J. H. Kim, J. Jun and B. T. Zhang, “Bilinear attention networks,” in *NIPS*, 2018.

[166] Y. Wang, N. Xu, A.-A. Liu, W. Li and Y. Zhang, “High-Order Interaction Learning for Image Captioning,” *IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY,* 2021.

[167] X. Yang, Y. Liu and X. Wang, “ReFormer: The Relational Transformer for Image Captioning,” *arXiv:2107.14178v2,* pp. 1-9, 2022.

[168] N. Shazeer, “GLU Variants Improve Transformers,” in *arxiv:2002.05202v1*, 2020.

[169] A. Nguyen, K. Pham, D. Ngo, T. Ngo and L. Pham, “An analysis of State-of-the-art activation functions for supervised deep-learning network,” in *arXiv:2104.02523v1*, 2021.

[170] A. Nguyen, K. Pham, D. Ngo, T. Ngo and L. Pham, “An Analysis of State-of-the-art Activation Functions For Supervised Deep Neural Network,” in *arXiv:2104.02523v1*, 2021.

[171] D. Hendrycks and K. Gimpel, “Gaussian error linear units,” in *arXiv:1606.08415*, 2016.

[172] A. Goyal, . A. Bochkovskiy, J. Deng and V. Koltun, “Non-deep Networks,” in *arXiv:2110.07641v1* , 2021.

[173] J. Hu, L. Shen, S. Albanie, G. Sun and E. Wu, “Squeeze-and-Excitation Networks,” in *arXiv:1709.01507*, 2019.

[174] Y. N. Dauphin, A. Fan, M. Auli and D. Grangier, “Language modeling with gated convolutional networks,” in *ICML*, 2017.

[175] J. Deng, W. Dong, R. Socher, L. J. Li, K. Li and L. Fei-Fei, “Imagenet: A large-scale hierarchical image database,” in *IEEE Conference on Computer Vision and Pattern Recognition*, Miami, 2009.

[176] R. Krishna, Y. Zhu, O. Groth and J. Johnson, “Visual Genome: Connecting language and vision using crowdsourced dense image annotations,” *International Journal of Computer Vision,* vol. 123, no. 1, pp. 32-73, 2017.

[177] K. Papineni, S. Roukos, T. Ward and W.-J. Zhu, “"Bleu: a method for automatic evaluation of machine,” in *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, 2002.

[178] N. Li and Z. Chen, “Image cationing with visual-semantic LSTM,” in *Proceedings of the 27th International Joint Conference on Artificial Intellegence*, 2018.

[179] P. Sharma, N. Ding, S. Goodmam and R. Soricut, “Conceptual Captions: A Cleaned, Hypernymed, Image Alt-text Dataset For Automatic Image Captioning,” *Association for Computational Linguistics,* vol. Proceedings of the

56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), p. 2556–2565, 2018.

[180] Y. Qin, J. Du, Y. Zhang and H. Lu, “Look Back and Predict Forward in Image Captioning,” in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[181] K. Biswas, S. Kumar, S. Banerjee and A. K. Pandey, “SMU: SMOOTH ACTIVATION FUNCTION FOR DEEP NETWORKS USING SMOOTHING MAXIMUM TECHNIQUE,” in *arXiv:2111.04682v2*, 2022.

[182] A. Bochkovskiy, C.-Y. Wang and H.-Y. M. Liao, “YOLOv4: Optimal Speed and Accuracy of Object Detection,” in *arXiv:2004.10934v1*, 2020.

[183] J. Pennington, R. Socher and C. Manning, “GloVe: Global Vectors for Word Representation,” in *Conference on Empirical Methods in Natural Language Processing*, Doha, Qatar, 2014.

[184] T. Mikolov, K. Chen, G. Corrado and J. Dean, “Efficient Estimation of Word Representations in Vector Space,” in *arXiv:1301.3781*, 2013.

[185] S. R. Bowman, L. Vilnis, O. Vinyals, A. Dai, R. Jozefowicz and S. Bengio, “Generating Sentences from a Continuous Space,” in *SIGNLL Conference on Computational Natural Language Learning*, Berlin, Germany, 2016.

[186] V. John, L. Mou, H. Bahuleyan and O. Vechtomova, “Disentangled Representation Learning for Non-Parallel Text Style Transfer,” in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy, 2019.

[187] H. Xu, S. Lu, Z. Sun, C. Ma and C. Guo, “VAE based Text Style Transfer with Pivot Words Enhancement Learning,” in *arXiv:2112.03154v1*, 2021.

[188] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer and V. Stoyanov, “RoBERTa: A Robustly Optimized BERT Pretraining Approach,” in *arXiv:1907.11692*, 2019.

[189] Z. Hu, Z. Yang, X. Liang, R. Salakhutdinov and E. P. Xing, “Toward Controlled Generation of Text,” in *International Conference on Machine Learning (PMLR)*, Sydney, Australia, 2017.

[190] F. Betti, G. Ramponi and M. Piccardi, “Controlled Text Generation with Adversarial Learning,” in *13th International Conference on Natural Language Generation*, Dublin, Ireland, 2020.

[191] G. E. Hinton, P. Dayan, B. J. Frey and R. M. Neal, “The "Wake-Sleep" Algorithm for Unsupervised Neural Networks,” 1995.

[192] R. Scott, H. Lee, D. Anguelov, C. Szegedy, D. Erhan and A. Rabinovich, “Training deep neural networks on noisy labels with bootstrapping,” in *arXiv:1412.6596*, 2014.

[193] D. P. Kingma and J. Ba, “Adam: A method for stochastic,” in *arXiv:1412.6980*, 2014.

[194] T. Yao, Y. Pan, Y. Li, Z. Qiu and T. Mei, “Boosting Image Captioning with Attributes,” in *IEEE International Conference on Computer Vision (ICCV)*, 2017.

[195] S. Bianco, L. Celona, M. Donzella and P. Napoletano, “Improving Image Captioning Descriptiveness by Ranking and LLM-based Fusion,” in *arXiv:2306.11593*, 2023.

[196] L. Ke, W. Pei, R. Li, X. Shen and Y.-W. Tai, Reflective Decoding Network for Image Captioning, arXiv:1908.11824, 2019.

[197] Y. Tan, Z. Lin, P. Fu, L. Wang, Y. Cao and W. Wang, “Detach and Attach: Stylized Image Captioning without Paired Stylized Dataset,” in *Proceedings of the 30th ACM International Conference on Multimedia*, 2022.

[198] X. Wu, W. Zhao and J. Luo, “Learning Cooperative Neural Modules for Stylized Image Captioning,” *International Journal of Computer Vision,* vol. 130, p. 2305–2320, 2022.

[199] P. Bojanowski, E. Grave, A. Joulin and T. Mikolov, “Enriching Word Vectors with Subword Information,” in *arXiv:1607.04606*, 2017.

[200] Y. Xiong, Z. Zeng, R. Chakraborty, M. Tan, G. Fung, Y. Li and V. Singh, “Nyströmformer: A Nyström-Based Algorithm for Approximating Self-Attention,” in *arXiv:2102.03902*, 2021.

[201] N. Shazeer, A. Mirhoseini, K. Maziarz, A. Davis, Q. Le, G. Hinton and J. Dean, “Outrageously Large Neural Networks: The Sparsely-Gated Mixture-of-Experts Layer,” in *arXiv:1701.06538*, 2017.

[202] A. Katharopoulos, A. Vyas, N. Pappas and F. Fleuret, “Transformers are RNNs: Fast Autoregressive Transformers with Linear Attention,” in *arXiv:2006.16236*, 2020.

[203] S. Semeniuta, A. Severyn and E. Barth, “A Hybrid Convolutional Variational Autoencoder for Text Generation,” in *arXiv:1702.02390*, 2017.

[204] M. Kusner, Y. Sun, N. Kolkin and K. Weinberger, “From word embeddings to document distances,” in *32nd International Conference on International Conference on Machine Learning*, 2015.

[205] S. Ioffe and C. Szegedy, “Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift,” in *arXiv:1502.03167*, 2015.

[206] X. Liang, Z. Hu, H. Zhang, C. Gan and E. P. Xing, “Recurrent Topic-Transition GAN for Visual Paragraph Generation,” in *arXiv:1703.07022*, 2017.

[207] Y. Mao, C. Zhou, X. Wang and R. Li, “Show and tell more: Topicoriented multi-sentence image captioning,” in *Twenty-Seventh International Joint Conference on Artificial Intelligence*, 2018.

[208] Y. Luo, Z. Huang, Z. Zhang, Z. Wang, J. Li and Y. Yang, “Curiosity-driven Reinforcement Learning for Diverse Visual Paragraph Generation,” in *arXiv:1908.00169*, 2019.

[209] S. Wu, Z.-J. Zha, Z. Wang, H. Li and F. Wu, “Densely supervised hierarchical policy value network for image paragraph generation,” in *Twenty-Eighth International Joint Conference on Artificial Intelligence*, 2019.

[210] W. Che, X. Fan, R. Xiong and D. Zhao, “Paragraph Generation Network with Visual Relationship Detection,” in *Proceedings of the 26th ACM international conference on Multimedia*, 2018.

[211] S. Yan, Y. Hua and N. Robertson, “ParaCNN: Visual Paragraph Generation via Adversarial Twin Contextual CNNs,” in *arXiv:2004.10258*, 2020.

[212] Y. Shi, Y. Liu, F. Feng, R. Li, Z. Ma and X. Wang, “S2TD: A Tree-Structured Decoder for Image Paragraph Captioning,” in *Proceedings of the 3rd ACM International Conference on Multimedia in Asia*, 2021.

[213] Y. Liu, Y. Shi, F. Feng, R. Li, Z. Ma and X. Wang, “Improving Image Paragraph Captioning with Dual Relations,” in *IEEE International Conference on Multimedia and Expo (ICME)*, 2022.

[214] T.-S. Nguyen and B. Fernando, “Effective Multimodal Encoding for Image Paragraph Captioning,” *IEEE Transactions on Image Processing,* vol. 31, pp. 6381-6395, 2022.

[215] M. Chatterjee and A. G. Schwing, “Diverse and Coherent Paragraph Generation from Images,” in *European Conference on Computer Vision (ECCV)*, 2018.

[216] J. Wang, Y. Pan, T. Yao, J. Tang and T. Mei, “Convolutional Auto-encoding of Sentence Topics for Image Paragraph Generation,” in *Twenty-Eighth International Joint Conference on Artificial Intelligence*, 2019.

[217] M.-T. Luong, Q. V. Le, I. Sutskever, O. Vinyals and L. KAiser, “Multi-task Sequence to Sequence Learning,” in *arXiv:1511.06114v4*, 2015.

[218] M. Daniluk, T. Rocktäschel, J. Welbl and S. Riedel, “Frustratingly Short Attention Spans in Neural Language Modeling,” in *arXiv:1702.04521*, 2017.

[219] A. See, P. J. Liu and C. D. Manning, “Get To The Point: Summarization with Pointer-Generator Networks,” in *arXiv preprint arXiv:1704.04368*, 2017.

[220] S. Ding, S. Qu, Y. Xi and S. Wan, “Stimulus-driven and concept-driven analysis for image caption generation,” *Neurocomputing,* vol. 398, pp. 520-530, 2020.

[221] C. Wang and X. Gu, “Learning joint relationship attention network for image captioning,” *Expert Systems with Applications,* vol. 211, p. 118474, 2023.

[222] S. Karaoglu, R. Tao, T. Gevers and A. Smeulders, “Words matter: Scene text for image classification and retrieval,” *IEEE Transactions on Multimedia,* vol. 19, no. 5, pp. 1063-1076, 2017.

[223] A. Gordo and D. Larlus, “Beyond instance-level image retrieval: Leveraging captions to learn a global visual representation for semantic retrieval,” in *Conference on Computer Vision and PAttern Recognition*, Hawai‘i Convention Center, 2017.

[224] A. Das, S. Kottur, K. Gupta, A. Singh, D. Yadav, J. Moura, D. Parikh and D. Batra, “Visual dialog,” in *IEEE Conerence on Computer Vision and Pattern Recognition*, Hawai‘i Convention Center, 2017.

[225] X. Lin and D. Parikh, “Leveraging visual question answering for imagecaption ranking,” *ECCV, ser. Lecture Notes in Computer Science,* vol. 9906, pp. 261-277, 2016.

[226] S. Wu, J. Wieland, O. Farivar and J. Schiller, “Automatic alt-text: Computer-generated image descriptions for blind users on a social network service,” in *ACM Conference on Computer-Supported Cooperative Work and Social Computing*, Portland, Oregon, 2017.

[227] A. Fukui, D. H. Park, D. Yang, A. Rohrbach, T. Darrell and M. Rohrbach, “Multimodal compact bilinear pooling for visual question answering and visual grounding,” in *arXiv:1606.01847*, 2016.

[228] D.-R. Beddiar, M. Oussalah and T. Seppänen , “Automatic captioning for medical imaging (MIC): a rapid review of literature,” *Artificial Intellegence Review,* vol. 56, p. 4019–4076, 2023.

[229] B. Jing, P. Xie and E. Xing, “On the Automatic Generation of Medical Imaging Reports,” in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Melbourne, Australia, 2018.

[230] X. Wang, Y. Peng, L. Lu, Z. Lu and R. M. Summers, “TieNet: Text-Image Embedding Network for Common Thorax Disease Classification and Reporting in Chest X-Rays,” in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, USA, 2018.

[231] Y. Li, X. Liang, Z. Hu and E. P. Xing, “ Hybrid retrieval-generation reinforced agent for medical image,” in *Conference on Neural Information Processing Systems*, Montreal Convention Centre, 2018.

[232] C. Y. Li, X. Liang, Z. Hu and E. P. Xing, “Knowledge-driven Encode, Retrieve, Paraphrase for Medical Image Report Generation,” in *arXiv:1903.10122*, 2019.

[233] M. Li, F. Wang, X. Chang and X. Liang, “Auxiliary Signal-Guided Knowledge Encoder-Decoder for Medical Report Generation,” in *arXiv:2006.03744*, 2020.

[234] F. Chollet, “Xception: Deep Learning with Depthwise Separable Convolutions,” in *arXiv:1610.02357*, 2016.

[235] J. Liu, J. Tang and G. Wu, “Residual Feature Distillation Network for Lightweight Image Super-Resolution,” in *arXiv:2009.11551*, 2020.

[236] Q. Liu, Y. Song, Q. Tang, X. Bu and N. Hanajima, “Wire rope defect identification based on ISCM-LBP and GLCM features,” *The Visual Computer,* 2023.

[237] N. Dalal and B. Triggs, “Histograms of oriented gradients for human detection,” in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, San Diego, CA, USA, 2005.

[238] G. Klambauer, T. Unterthiner, A. Mayr and S. Hochreiter, “Self-Normalizing Neural Networks,” in *arXiv:1706.02515*, 2017.

[239] D. Misra, T. Nalamada, A. U. Arasanipalai and Q. Hou, “Rotate to Attend: Convolutional Triplet Attention Module,” in *arXiv:2010.03045*, 2020.

[240] S. Woo, J. Park, J.-Y. Lee and I. S. Kweon, “CBAM: Convolutional Block Attention Module,” in *arXiv:1807.06521*, 2018.

[241] M. Korschens, P. Bodesheim and J. Denzler, “Beyond Global Average Pooling: Alternative Feature Aggregations for Weakly Supervised Localization,” in *Proceedings of the 17th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*, Online, 2022.

[242] A. Hadid, “The Local Binary Pattern Approach and its Applications to Face Analysis,” in *First Workshops on Image Processing Theory, Tools and Applications*, Sousse, Tunisia, 2008.

[243] Z. Pan, S. Hu, X. Wu and P. Wang, “Adaptive center pixel selection strategy in Local Binary Pattern for texture classification,” *Expert Systems with Applications,* vol. 180, p. 115123, 2021.

[244] D. D. Fushman , M. D. Kohli, M. B. Rosenman, S. E. Shooshan, L. Rodriguez, S. Antani, G. R. Thoma and C. J. McDonals, “Preparing a collection of radiology examinations for distribution and retrieval,” *Journal of the American Medical Informatics Association,* vol. 23, no. 2, pp. 304-310, 2016.

[245] I. Najdenkoska, X. Zhen, M. Worring and L. Shao, Variational Topic Inference for Chest X-Ray Report Generation, arXiv:2107.07314, 2021.

[246] G. Huang, Z. Liu, L. v. d. Maaten and K. Q. Weinberger, “Densely Connected Convolutional Networks,” in *arXiv:1608.06993*, 2018.

[247] S. Wang, L. Tang, M. Lin, G. Shih, Y. Ding and Y. Peng, Prior knowledge enhances Radiology Report Generation, arXiv:2201.03761, 2022.

[248] Z. Chen, Y. Shen, Y. Song and X. Wan, Cross-modal Memory Networks for Radiology Report Generation, arXiv:2204.13258 , 2022.

[249] B. Jing, Z. Wang and E. Xing, “Show, Describe and Conclude: On Exploiting the Structure Information of Chest X-ray Reports,” in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy, 2019.

[250] H. T. Nguyen, D. Nie, T. Badamdorj, Y. Liu, L. Hong, J. Truong and L. Cheng, EDDIE-Transformer: Enriched Disease Embedding Transformer for X-Ray Report Generation, Kolkata, India: IEEE 19th International Symposium on Biomedical Imaging (ISBI), 2022.

[251] X. Wu, S. Yang, Z. Qiu, S. Ge, Y. Yan, X. Wu, Y. Zheng, S. K. Zhou and L. Xiao, DeltaNet: Conditional Medical Report Generation for COVID-19 Diagnosis, Gyeongju, Republic of Korea: Proceedings of the 29th International Conference on Computational Linguistics, 2022.

[252] K. Fan, X. Cai and M. Niranjan, IIHT: Medical Report Generation with Image-to-Indicator Hierarchical Transformer, : arXiv:2308.05633 , 2023.

[253] F. Liu, X. Wu, S. Ge, W. Fan and Y. Zou, Exploring and Distilling Posterior and Prior Knowledge for Radiology Report Generation, Online: EEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021.

[254] Y. Xiong, B. Du and P. Yan, Reinforced Transformer for Medical Image Captioning, Machine Learning in Medical Imaging- Springer, 2019.

[255] V. Wijerathna, H. Raveen, S. Abeygunawardhana and T. D. Ambegoda, "Chest X-Ray Caption Generation with CheXNet," in *Moratuwa Engineering Research Conference (MERCon)*, Moratuwa, Sri Lanka, 2022.

[256] B. P. Voutharoja, L. Wang and L. Zhou, "Automatic Radiology Report Generation by Learning with Increasingly Hard Negatives," in *arXiv:2305.07176*, 2023.

[257] X. Song, X. Zhang, J. Ji, Y. Liu and P. Wei, "Cross-modal Contrastive Attention Model for Medical Report Generation," in *Proceedings of the 29th International Conference on Computational Linguistics*, Gyeongju, Republic of Korea, 2022.

[258] W. Hou, K. Xu, Y. Cheng, W. Li and J. Liu, "ORGAN: Observation-Guided Radiology Report Generation via Tree Reasoning," in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, Toronto, Canada, 2023.

[259] https://www.kaggle.com/aishrules25/automatic-image-captioning-for-visually-impaired.

AUTHOR BIOGRAPHY

Dhruv Sharma
2K20/PHDEC/02

Department of Electronics and Communication Engineering

Delhi Technological University, Delhi, India

Email: dhruv.0906@yahoo.in

Dhruv Sharma received the B.Tech. degree in electrical and electronics engineering from GGSIPU University, New Delhi, India, in 2015, and M. Tech degree in signal processing from Ambedkar Institute of Advanced Communication Technology & Research, New Delhi, India, in 2017. He is currently working as a Research Scholar with the Department of Electronics and Communication Engineering, Delhi Technological University, New Delhi, India. His research interests include machine learning, deep learning, computer vision, natural language processing, and image captioning and visual question answering. He is also a reviewer for various Journals/Transactions of IEEE, Elsevier, and Springer.