DESIGN A FRAMEWORK FOR GENERATION OF IMAGE DESCRIPTION USING DEEP LEARNING

Thesis submitted to Delhi Technological University
in partial fulfillment of the requirements
for the award of the degree of

DOCTOR OF PHILOSOPHY
In
INFORMATION TECHNOLOGY

By
LAKSHITA AGARWAL
(2K21/PHDIT/05)

Under the Supervision of
DR. BINDU VERMA

DEPARTMENT OF INFORMATION TECHNOLOGY
DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
NEW DELHI-110042, INDIA

MAY-2025

Acknowledgements

I have an abundance of gratitude for all of the people who contributed to helping me finish my Doctor of Philosophy journey. Their constant encouragement, direction, and support have been crucial to this endeavour. I would first and foremost like to thank my supervisor, **Dr. Bindu Verma**. She has consistently inspired me with her guidance, commitment, and enthusiasm for the advancement of knowledge. Her expertise and thoughts influenced my research path and helped me gain a deeper comprehension of the topic. I am so grateful that I was able to work under her guidance. My research was greatly influenced by the invaluable advice and assistance of my DRC members, **Prof. Dinesh Kumar Vishwakarma, Prof. Anil Singh Parihar**, and **Prof. Seba Susan**. I am also very thankful to the distinguished faculty members of the IT Department, DTU, including **Prof. Kapil Sharma, Dr. Virender Ranga, Dr. Ritu Agarwal, Dr. Varsha Sisaudia, Dr. Priyanka Meel, Dr. Anamika Chauhan, Dr. Rahul Gupta**, and **Ms. Geetanjali Bhola**.

I would like to dedicate a special part of my acknowledgment to my beloved grandparents, **Shri. Dharampal Agarwal** and **Late Mrs. Susheela Devi Agarwal**, whose wisdom, love, and values continue to inspire and guide me every day. Their unwavering principles have shaped my journey in countless ways. My heartfelt gratitude goes to my parents, **Mr. Vijay Kumar Agarwal** and **Mrs. Anita Agarwal**, for their unconditional love, sacrifices, and constant belief in my abilities. Their unwavering encouragement has been the foundation of my academic pursuits, motivating me to strive for excellence. I also extend my deepest appreciation to my younger brother **Tushar Agarwal**, for his constant support, encouragement, and faith in me. Their belief in my potential has been a source of immense strength and motivation. Each of you has played an indispensable role throughout this journey, and I am profoundly grateful for your presence, love, and support.

To my dear friends, for their unwavering support, encouragement, and companionship. Their presence has not only made the challenging times more bearable but has also added joy to every moment of success. I am also sincerely grateful to everyone whose names may not be mentioned here but who has played a significant role in my academic journey. Their support, patience, and kindness have been invaluable to both my personal and professional growth. This journey has been made possible through the collective support of all these remarkable individuals, and for that, I am profoundly thankful.

Lakshita Agarwal

(Lakshita Agarwal)

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

Certificate

This is to certify that the thesis entitled **"Design a Framework for Generation of Image Description using Deep Learning"**, being submitted by **Lakshita Agarwal (2K21/PHDIT/05)** to the **Department of Information Technology, Delhi Technological University, India,** in partial fulfiLlment of the requirements for the award of the Degree of **Doctor of Philosophy** in **Information Technology,** is an authentic record of work carried out by her under the guidance and supervision of **Dr. Bindu Verma**.

This research work has not been submitted, in part or full, to any other University or Institution for the award of any degree or diploma.

Date: 09/05/2025
Place: New Delhi

Lakshita Agarwal

Lakshita Agarwal
(Ph.D. Student)

Bindu
14/08/25
(Supervisor)
Dr. Bindu Verma
(Assistant Professor)
Department of Information Technology
Delhi Technological University, Delhi- 110042

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

Declaration

I hereby declare that the Ph.D. thesis entitled **"Design a Framework for Generation of Image Description using Deep Learning"** being submitted to the **Delhi Technological University, Delhi,** in partial fulfillment of the requirements for the award of the degree of **Doctor of Philosophy** in **Department of Information Technology,** is an authentic record of work carried out by me under the guidance and supervision of **Dr. Bindu Verma**.

I also mention that the research work is original and has not been submitted by me, in part or full, to any other University or Institution for the award of any degree or diploma.

Lakshita Agarwal

Lakshita Agarwal
(Ph.D. Student)
Department of Information Technology,
Delhi Technological University, New Delhi-110042 India

Dedicated to:

My Parents, Brother, Friends, and Family: For their eternal love, endless support, and encouragement, as well as for their constant consideration and understanding during this journey...

Abstract

Image description generation, an intricate cross-disciplinary work between computer vision and natural language processing (NLP), is intended to produce contextually precise and semantically rich textual descriptions of visual information. The proposed work is dedicated to filling significant research gaps in computerised image captioning by suggesting sophisticated deep-learning architectures that promote contextual knowledge, semantic density, and generality across multimedia applications. The research is organised into three main tasks: (1) creating an automatic system for producing contextually and semantically rich image descriptions; (2) constructing a deep learning system to enhance description accuracy and prediction scores; and (3) designing image description models specific to multimedia uses. For the purpose of fulfilling these objectives, the thesis proposes several novel models. The VGG16-SceneGraph-BiGRU model integrates VGG16 for visual feature extraction, scene graphs for object relationship capture, and a BiGRU network for sequential language modelling, resulting in coherent and contextually enriched descriptions. Additionally, the Tri-FusionNet model combines a Vision Transformer (ViT) encoder, two attention mechanisms, a RoBERTa decoder, and a CLIP module to support improved feature extraction and multimodal alignment, enhancing description accuracy. Domain-specific use cases, such as medical imaging and autonomous driving, are also examined in the thesis with models designed specifically for the application, such as a ViT-GPT4 framework for chest X-ray analysis and a ResNet50 with a GPT2-based system to describe video-based behaviour. The proposed models are tested on benchmark data like MS COCO, Flickr8k, Flickr30k, IU Chest X-ray, NIH Chest X-ray, MSVD, and BDD-X Vehicular dataset on metrics like BLEU (1-4), CIDEr, METEOR, and ROUGE-L. The results show significant gains in description quality, semantic completeness, and contextual accuracy, setting new state-of-the-art image description generation benchmarks.

Author Research Publications

Journal Papers:

1. **Lakshita Agarwal**, and Bindu Verma. "From methods to datasets: A survey on Image-Caption Generators". *Multimedia Tools and Applications* (2023): 28077–28123. **(SCIE Indexed, IF: 3)**
   DOI: 10.1007/s11042-023-16560-x ***(Published)***.

2. **Lakshita Agarwal**, and Bindu Verma. "Enriching image description generation through multi-modal fusion of VGG16, scene graphs and BiGRU". *The Visual Computer* (2024): 1–21. **(SCIE Indexed, IF: 3)**
   DOI: 10.1007/s00371-024-03790-9 ***(Published)***.

3. **Lakshita Agarwal**, and Bindu Verma. "Tri-FusionNet: Enhancing Image Description Generation with Transformer-based Fusion Network and Dual Attention Mechanism".
   Archived at: arxiv.org/abs/2504.16761. Communicated in *IEEE Transactions on Human-Machine Systems* **(IF: 3.5)** ***(Communicated: 1st Major Revision Submitted)***.

4. **Lakshita Agarwal**, and Bindu Verma. "Advanced Chest X-Ray Analysis via Transformer-Based Image Descriptors and Cross-Model Attention Mechanism".
   Archived at: arxiv.org/abs/2504.16774. Communicated in *Computational Intelligence, Wiley* **(SCIE Indexed, IF: 1.8)** ***(Communicated)***.

5. **Lakshita Agarwal**, and Bindu Verma. "Towards Explainable AI: Multi-Modal Transformer for Video-based Image Description Generation".
   Archived at: arxiv.org/abs/2504.16788. Communicated in *Signal, Image and Video Processing, Springer* **(SCIE Indexed, IF: 2.0)** ***(1st Major Revision)***.

International Conferences:

1. **Lakshita Agarwal**, and Bindu Verma. "Comparison of Deep Learning Models for Automatic Image Descriptors". *In 2023 IEEE 20th India Council International Conference (INDICON)* (pp. 914–919). IEEE, 2023.
   DOI: 10.1109/INDICON59947.2023.10440731 *(Published)*.

2. **Lakshita Agarwal**, and Bindu Verma. "Utilizing Transformer-Based Image Descriptors for Improving Chest X-Ray Analysis". In: Nanda, S.J., Yadav, R.P., Gandomi, A.H., Saraswat, M. (eds) *Data Science and Applications*. ICDSA 2024. Lecture Notes in Networks and Systems, vol 1266. Springer, Singapore.
   DOI: 10.1007/978-981-96-2647-2_40 *(Published)*.

Contents

Acknowledgements **ii**

Certificate **iv**

Declaration **v**

Dedication **vi**

Abstract **vii**

Author Research Publications **viii**

List of tables **xvii**

List of Figures **xx**

List of Abbreviations **xxi**

1 Introduction **1**

1.1 Applications of Image Description Generation 3
1.2 Challenges in Image Description Generation 4
1.3 Research Gaps and Motivation 5
1.4 Problem Definition 7
  1.4.1 Research Objectives 8
1.5 Contributions in the Thesis 8
1.6 Outlines of the Thesis 11

2 Literature Survey 15
2.1 Literature Survey on Traditional Approaches for Image Description Generation . . . 16
2.2 Literature Survey on Deep Learning Methods for Image Description Generation . . . 17
2.3 Literature Survey based on Applications of Image Description Generation 20
2.4 In Thesis Prospective: . . . 25

3 Proposed a Unified Framework for Contextual and Semantic Image Description Generation 28
3.1 Introduction . . . 28
3.2 Literature Survey . . . 30
3.3 Proposed Architecture . . . 31
3.3.1 VGG16 Feature Extraction: . . . 32
3.3.2 Scene Graphs: . . . 33
3.3.3 BiGRU for Temporal Context: . . . 35
3.3.4 Description Generation Process: . . . 38
3.4 Experimental Analysis . . . 39
3.4.1 Implementation Details: . . . 41
3.4.2 Performance Measures and Evaluation . . . 42
3.4.3 Ablation Study: . . . 44
3.4.4 Result and Analysis: . . . 50
3.4.5 Comparison With Other State-of-the-art Methods: . . . 54
3.5 Conclusion . . . 57

4 Deep Learning Framework for Improved Image Description Accuracy 58
4.1 Introduction . . . 58
4.2 Literature Survey . . . 59
4.3 Proposed Architecture . . . 61
4.3.1 Overview of the Proposed Architecture: . . . 62

4.4 Experimental Analysis 71
  4.4.1 Implementation Details: 74
  4.4.2 Ablation Study: 75
  4.4.3 Results and Analysis: 78
  4.4.4 Comparison with Other State-of-the-art Methods: 83
4.5 Conclusion 86

5 Image Description Models for Multimedia Application: Chest X-Ray Analysis 88
5.1 Introduction 89
5.2 Literature Survey 90
5.3 Proposed Architecture 91
  5.3.1 Data Pre-processing 92
  5.3.2 Encoder Module of the Proposed Architecture 93
  5.3.3 Cross-model attention mechanism 95
  5.3.4 Decoder Module of the Proposed Architecture 95
5.4 Experimental Analysis 98
  5.4.1 Implementation Details: 99
  5.4.2 Ablation Study: 99
  5.4.3 Results and Analysis: 101
  5.4.4 Comparison With Other state-of-the-art Methods: 102
5.5 Conclusion 103

6 Image Description Models for Multimedia Application: Video-based Image Description Generation 104
6.1 Introduction 104
6.2 Literature Survey 105
6.3 Proposed Architecture 107
  6.3.1 Data Pre-processing 107
  6.3.2 Model Architecture 108
6.4 Experimental Analysis 111

6.4.1 Ablation Study 111
6.4.2 Results and Analysis: 114
6.4.3 Comparison With State-of-the-Art Methods: 115
6.5 Conclusion 117

7 Conclusion 118

7.1 Summary and Contribution of the Thesis 119
7.2 Future Directions 121

Bibliography 122

Appendix A Vision Transformer (ViT) 139

A.1 Overview 139
A.1.1 Image Patching and Embedding 139
A.1.2 Positional Encoding 140
A.1.3 Transformer Encoder 140
A.1.4 Classification Token (CLS Token) 141
A.1.5 Classification Head (MLP Head) 141

Appendix B RoBERTa (Robustly Optimized BERT Approach) 142

B.1 Overview 142
B.1.1 Architecture 142
B.1.2 Position-wise Feed-Forward Network (FFN) 143
B.1.3 Token Embedding 144
B.1.4 Final Output 144
B.1.5 Applications 144

Appendix C Bidirectional Gated Recurrent Unit (BiGRU) Model 145

C.1 Overview 145
C.1.1 Architecture 145
C.1.2 Bidirectional GRU 146
C.1.3 Applications 147

Appendix D Scene Graphs **148**

D.1 Overview 148

D.1.1 Scene Graph Construction 148

D.1.2 Scene Graphs in Image Captioning 149

D.1.3 Integration with Deep Learning Models 149

D.1.4 Applications 149

Appendix E GPT-2 and GPT-4 Transformers **151**

E.1 Overview 151

E.1.1 GPT-2 Architecture 151

E.1.2 GPT-4 Architecture 152

E.1.3 Key Differences Between GPT-2 and GPT-4 153

List of Publications and their Proofs **154**

Plagiarism Report **160**

Curriculum Vitae **162**

List of Tables

2.1 Descriptive Summarization of Image Description Generation Approaches 21

3.1 Hyper-Parameters used for the Experimental Analysis 41
3.2 Ablation Study: Performance comparison of different components on MS COCO dataset 44
3.3 Ablation Study: Performance comparison of different components on Flickr8k dataset 45
3.4 Ablation Study: Performance comparison of different components on Flickr30k dataset 46
3.5 Quantitative Results Obtained on Flickr8k Dataset 49
3.6 Results Obtained on MS COCO Dataset 51
3.7 Results Obtained for Flickr8k Dataset 52
3.8 Results Obtained for Flickr30k Dataset 53
3.9 Results obtained from the proposed model 53
3.10 Comparative Analysis of Image Description Generation Models on MS COCO Dataset 55
3.11 Comparative Analysis of Image Description Generation Models on Flickr8k Dataset 56
3.12 Comparative Analysis of Image Description Generation Models on Flickr30k Dataset 56

4.1 Comparison of Tri-FusionNet with state-of-the-art models on scalability and computational cost. 73
4.2 Architectural Details for the Proposed Model 74

4.3 Hyperparameters and Evaluation Metrics for the Tri-FusionNet Framework . . . 75
4.4 Performance Metrics of Image Description Generation Models in the MSCOCO Dataset . . . 76
4.5 Performance Metrics of Image Description Generation Models in the Flickr30k Dataset . . . 76
4.6 Performance Metrics of Image Description Generation Models in the Flickr8k dataset . . . 77
4.7 Quantitative Results Obtained on MSCOCO Dataset . . . 78
4.8 Quantitative Results Obtained on Flickr30k Dataset . . . 79
4.9 Quantitative Results Obtained on Flickr8K Dataset . . . 80
4.10 Overall Results obtained from the proposed model . . . 81
4.11 Qualitative Analysis of Generated Image Descriptions . . . 82
4.12 Comparative Analysis of Image Description Generation Models on MSCOCO Dataset . . . 84
4.13 Comparative Analysis of Image Description Generation Models on online MSCOCO Test Server . . . 85
4.14 Comparative Analysis of Image Description Generation Models on Flickr30k Datasets . . . 85
4.15 Comparative Analysis of Image Description Generation Models on Flickr8k Dataset . . . 86

5.1 Architectural Details for the Proposed Model . . . 99
5.2 Hyperparameters and Evaluation Metrics for the CrossViT-GPT4 Framework . . . 100
5.3 Evaluation of Models on the IU and NIH Chest X-Ray Datasets . . . 100
5.4 Quantitative Results Obtained on IU Chest X-Ray Dataset . . . 101
5.5 Quantitative Results Obtained on NIH Chest X-Ray Dataset . . . 101
5.6 Comparative Analysis on IU and NIH Chest X-Ray Datasets . . . 102

6.1 Experimental Setup and Performance Metrics . . . 112

6.2 Ablation Study: Comparing the Impact of Model Components on Different Datasets 112
6.3 Results and Analysis of the Proposed Framework 114
6.4 Comparison of State-of-the-Art Models for Vehicle Action Understanding116

List of Figures

1-1 Graphical abstract of the overall chapter-wise distribution of the thesis. 14

2-1 Distribution of approaches used in image description generation [55]. 25

3-1 Structural representation of the proposed framework: The architecture consists of three phases. In the initial phase, the VGG16 model extracts the high-level visual characteristics from the pre-processed images. A scene graph creation model for extracting visual attributes in the next step. Finally, feeding this combined information into a Bidirectional Gated Recurrent Unit (BiGRU) for generating the image descriptions using the dense layer of network. 32
3-2 Basic architecture of VGG16 [67]. 33
3-3 Basic architecture of scene graph generation model: The model consists of the pre-trained Faster R-CNN which is used for identifying the objects contained in images and to generate scene graphs that include data on the labels and bounding boxes of the identified objects. . . . 34
3-4 Architecture of BiGRU Module for Description Generation. 35
3-5 A sample image from Flickr8k dataset 40
3-6 A sample image from the Flickr30k dataset 40
3-7 Comparative graphical depiction of the results obtained for various models evaluated on MS-COCO dataset 45
3-8 Comparative graphical depiction of the results obtained for various models evaluated on Flickr8k dataset 46

3-9 Comparative graphical depiction of the results obtained for various models evaluated on Flickr30k dataset 47
3-10 Representation of the Correlation Heatmap of Evaluation Metrics for MSCOCO, Flickr8k and Flickr30 dataset. 48
3-11 Graphical representation of the evaluation scores obtained from MSCOCO, Flickr8k and Flickr30k dataset for the proposed framework. 54

4-1 Structural representation of Tri-FusionNet framework: The architecture consists of three phases: firstly, high-level visual features are first extracted from pre-processed images using a Vision transformer encoder with dual-attention mechanism; next, words from the input caption file are tokenized by a RoBERTa decoder; and last, the combined data is fed into a CLIP-integrating module to create image descriptions using dense network layers. 62
4-2 Model architecture for vision transformer encoder with dual attention mechanism [85]. 64
4-3 Architecture of RoBERTa Decoder Module. 67
4-4 Model architecture for CLIP integration module. 69
4-5 A sample image from the Flickr30k dataset 72
4-6 Example of obtained heat maps based on dual-attention mechanism. . 76
4-7 Graphical representation of the results obtained from MSCOCO, Flickr30k and Flickr8k dataset for the proposed Tri-FusionNet framework. . . . 82

5-1 A representation of the proposed framework's structure: CrossViT-GPT4- During the initial stage, the ViT encoder utilizes cross-model attention to extract high-level visual features from the pre-processed images. GPT-4 decoder uses tokenization to extract individual words from the provided input caption file. It then utilizes the dense layers of the network to produce image descriptions. 91
5-2 Data Pre-Processing Step 93
5-3 Illustration of ViT- Encoder module with heat maps 94

5-4 Cross-Model Attention Mechanism. 95
5-5 GPT-4 Decoder Module. 96

6-1 Framework for the Proposed Model (ResNet50-GPT2): The system incorporates ResNet50 for image feature extraction and a GPT-2 encoder-decoder model to generate context-aware video-based image descriptions. 108
6-2 Demonstration of Ablation Study for the Proposed Work 113
6-3 Graphical Representation of the Results Obtained 115
6-4 Graphical Representation of the SOTA Methods 116

A-1 Basic Architecture of Vision Transformer (ViT) [135] 140

B-1 Basic Architecture of RoBERTa 143

C-1 Basic architecture of Bi-GRU Model 146

E-1 Basic Architecture of GPT-2 Transformer Model 152

List of Abbreviations

NLP	Natural Language Processing
AI	Artificial Intelligence
SIFT	Scale-Invariant Feature Transform
HOG	Histogram of Oriented Gradients
CNN	Convolutional Neural Network
VGG16	Visual Geometry Group 16
ResNet	Residual Neural Network
RNN	Recurrent Neural Network
LSTM	Long Short-Term Memory
ViT	Vision Transformer
GPT	Generative Pre-trained Transformer
BERT	Bidirectional Encoder Representations from Transformers
CLIP	Contrastive Language-Image Pre-training
CCTV	Closed-Circuit Television
GRU	Gated Recurrent Unit
BiGRU	Bidirectional Gated Recurrent Unit
MS COCO	Microsoft Common Objects in Context
BLEU	Bilingual Evaluation Understudy
CIDEr	Consensus-based Image Description Evaluation
METEOR	Metric for Evaluation of Translation with Explicit Ordering
Rouge-L	Recall-Oriented Understudy for Gisting Evaluation - Longest Common Subsequence
B	BLEU Score
M	METEOR
C	CIDEr
R-L	Rouge-L

RoBERTa	Robustly Optimized BERT Approach
IU	Indiana University
NIH	National Institutes of Health
BDD-X	Berkeley Deep Drive-X (eXplanation)
2D MHMMs	Two-dimensional multi-resolution hidden Markov models
KNN	K-Nearest Neighbors
SVMs	Support Vector Machine
M	Million
OSCAR	Object-Semantics Aligned Pre-training
VIVO	Visual Vocabulary Pre-Training
GAN	Generative Adversarial Networks
R-CNN	Region-based Convolutional Neural Network
M2	Meshed-Memory Transformer
MT	Modal Transformer
MTTSNet	Multi-Task Triple-Stream Network
FCLN	Fully Convolutional Localization Network
ETDC	Enhanced Transformer Dense Captioner
TDC	Transformer-based Dense Captioner
RoIs	Regions of Interests
Corr	Correlation
SvDCorr	Singular Value Decomposition-based correlation
Cos	Cosine Similarity
SvdCos	Singular Value Decomposition-based cosine similarity
RL	Reinforcement Learning
CGO	Captions with Guiding Object
m-RNN	Multi-modal Recurrent Neural Network
LSTM-A	Long Short-Term Memory with Attributes
MIL	Multiple-Instance Learning
GCNs	Graph-Convolutional Networks
X-Lan	X-Linear Attention

CAAG	Context-Aware Auxiliary Guidance
SCST	Self-critical sequence training
SPICE	Semantic Propositional Image Caption Evaluation
SPIDEr	combination of SPICE and CIDEr
VSU	Visual Semantic Units
GPU	Graphics Processing Unit
et al.	and others
SGAE	Scene Graph Auto-Encoder
MCA	Multi-Level Cross-Modal Alignment
max	Maximum
GB	Gigabyte
RAM	Random Access Memory
JPEG	Joint Photographic Experts Group
.json	JavaScript Object Notation
exp	Exponential
BP	Brevity Penalty
P	Precision
R	Recall
PE	Positional Encoding
pos	Position
dim	Dimension
LCS	Longest Common Sub-sequence
VG-Cap	Visual-Grounding Captioning
TLGSA	Transformer-based local graph semantic attention
ETA	EnTangled Attention
SAMT	Semantic Attribute Multi-Tagging
LSM	Learnable Sparse Mechanism
LFE	Local Feature Enhancement
PMA-Net	Prototypical Memory Attention Network
SSL	Self-Supervised Learning

SPT	Spatial Pyramid Transformer
DLCT	Dual-Level Collaborative Transformer
concat	Concatinate
Q	Queries
K	Keys
V	Values
VRAM	Video Random Access Memory
CPU	Central Processing Unit
BLIP	Bootstrapping Language-Image Pre-training
NIC	Natural Image Captioning
HAAV	Hierarchical Aggregation of Augmented Views
SCN	Semantic Compositional Networks
RFNet	Recurrent Fusion Network
AoANet	Attention on Attention Network
TMA	Topic-based multi-channel attention
FA	Feature-augmented
FF	Feed-forward
YOLO	You Only Look Once

Chapter 1

Introduction

An image description is a brief textual phrase summarising an image's scene, objects, or situation. A cutline is a more in-depth explanation accompanying the image, providing additional context or details. By offering contextual information and relating visual content to pertinent topics, descriptions are essential for improving image understanding. Automated image description generation is increasingly popular, driven by different pre-trained models and other internet-based tools in different research areas [1].

Image caption generation is a sophisticated task combining computer vision and natural language processing (NLP) to understand visual information and transform it into coherent textual descriptions. As a fundamental field of artificial intelligence, it tries to bridge the gap between language generation and image understanding by producing meaningful descriptions. While object recognition is primarily concerned with detecting objects, image captioning necessitates a deeper level of abstraction to describe scenes' interactions and relationships. It requires a solid knowledge of visual semantics and language structures, thus being especially vital for assistive technology to understand images. While AI-based captioning models have come a long way, generating correctly and contextually descriptive captions remains problematic, especially when object interactions are complex or visual objects are occluded. Effective captioning models take place through a review of predictive methods and the construction of strong problem-solving frameworks to produce descriptive words for images of all types [2].

The ultimate objective is to create effective algorithms that process and encode image content and form coherent relationships between image and text features. This means detecting objects with their interactions and subtle details in a scene and generating semantically fluent and dense captions. Image captioning can be understood as an image-to-sequence learning task in which a model maps pixel-based visual information into a structured word sequence from a fixed vocabulary [3]. The encoder-decoder model is followed in the image description generation task. A deep learning-based vision model—typically a Convolutional Neural Network (CNN)—initially encodes the image into a feature representation. A language model, usually a Transformer or Recurrent Neural Network (RNN), receives this feature representation along-with the ground-truth annotations and decodes them into an understandable written description.

Traditional approaches for image description generation were based on unrefined feature extraction and rule-based sentence formation, which mainly focused on object detection and templates. Techniques like SIFT [4] and HOG [5] could detect shapes and textures but were usually lacking in context, hence descriptions were not very accurate. Rule-based systems used templates with object labels that ensured grammatical correctness but resulted in rigid, generic descriptions [6]. Although probabilistic models, such as n-grams and Hidden Markov Models, included some variations, their rigid structure and lack of flexibility made it difficult to capture complex object interactions [7]. Despite these advancements, traditional methods had some serious drawbacks: reliance on the accuracy of object detection, poor generalization, and limited contextual awareness. As a result, deep learning-based approaches were adopted to provide more flexibility and richer context-aware descriptions [8].

The emergence of deep learning made it possible to learn hierarchical representations of images directly from large datasets. [9] Earlier models used the feature extraction properties of CNNs like VGG16 and ResNet, processing those features by using RNNs or LSTMs to generate descriptions [10]. Such models couldn't focus on any specific regions in the images; hence, attention mechanisms were brought in to enhance the contextual accuracy of the generated descriptions. While generating every phrase in the description, image captioning models are able to dynamically focus on important areas of an image by using attention mechanisms. The model enhances contextual correctness and captures finer details by giving distinct attention weights to different visual components at each decoding stage, resulting in more relevant and detailed descriptions [11]. Recently, models such as Vision Transformer (ViT) combined with Generative Pre-Trained Transformer (GPT) and Bidirectional Encoder Representations from Transformers (BERT) surpassed the performance of the traditional CNN-based model. Such transformers handle long-range dependencies greatly with self-attention applications and hence, increase effectiveness for both visual feature extraction and language generation [12]. Models such as ViT-GPT2 and Contrastive Language-Image Pre-training (CLIP) better align image and text representations, generating very highly context-aware descriptions [13] [14].

Deep learning methods overcome traditional limitations by avoiding unrefined features and automatically learning complex relationships. Such models generate more coherent and meaningful descriptions and take advantage of attention mechanisms and transformer architectures for improved fluency and contextual understanding. However, they require significant amounts of training data, computational resources, and fine-tuning for optimal performance. Transformer-based models use self-attention procedures to better understand contextual linkages and long-range dependencies, outperforming conventional CNN-RNN architectures in image captioning. They make parallel processing possible, which speeds up training and produces descriptions that are more logical and contextually aware.

1.1 Applications of Image Description Generation

Image description generators have many practical applications in the real world and across different fields. In autonomous driving, for example, self-driving cars depend on computer vision to navigate and make decisions. Image captioning can help improve situational awareness by giving real-time descriptions of objects, road conditions, and potential hazards, enhancing overall safety and system reliability. NLP and image captioning are used by assistive technology such as screen readers, object recognition apps (Seeing AI, Be My Eyes), and wearables with AI (Google Glass, smart glasses) to give visually impaired people speech input in real-time. By providing insightful descriptions of objects, locations, and environmental obstacles, these systems translate visual input into descriptive text and audio, facilitating autonomous navigation. By producing thorough descriptions of X-rays, MRIs, and CT scans, image captioning helps in diagnostic imaging in healthcare. AI models improve workflow efficiency and decision-making for radiologists by assisting them in identifying anomalies, monitoring the course of diseases, and minimising errors—particularly in places with limited resources.

Image description generation also plays a very significant role in video surveillance and security. With wide penetration of CCTV cameras, automated captioning can bring real-time alerts by detecting unusual activities and triggering alarms in cases of security breaches or accidents, thus significantly enhancing the crime prevention and response mechanisms. Moreover, it enhances information retrieval systems, such as Google Image Search. Since the search engine can now generate textual descriptions of images, it can better categorize and retrieve images for more accurate results. Similarly, captioning is widely used in social media for automated image annotation, content moderation, and accessibility enhancements, thereby allowing a better understanding of visual content across platforms.

1.2 Challenges in Image Description Generation

Despite advancements in image description generation, there are still a number of issues with natural language processing and computer vision. Developing a trustworthy assessment metric that compares machine-generated and human-generated descriptions is one of the main problems [15] [16]. The richness of human language is frequently not captured by current measurements, making objective evaluation difficult. It is still challenging to guarantee contextual relevance and grammatical quality in descriptions since models may have trouble interpreting complex language. Another challenging problem is producing human-like descriptions while maintaining semantic consistency, which calls for striking a compromise between grammatical correctness and insightful, understandable explanations. The process of creating descriptions is further complicated by managing occluded or unclear visual aspects and preserving complex object relationships.

Furthermore, captioning models frequently have trouble producing descriptions that appropriately capture the meaning or emotional tone of an image— a critical skill in fields like social media and journalism. Computational efficiency is still a problem since real-time applications are challenging because of the high processing power requirements for producing high-quality descriptions. The development process is further complicated by the need for domain-specific training datasets and a greater comprehension of specialized vocabulary when adapting captioning models to specialized domains, such as scientific or medical imaging.

1.3 Research Gaps and Motivation

We reviewed the state-of-the-art methods in image description generation and identified the key research gaps. Significant progress has been made, but issues in contextual accuracy, generalization across datasets, and real-time performance still persist as major difficulties for developing reliable and effective description generation systems for real-world applications.

- Existing models often struggle to establish an effective alignment between visual content and textual descriptions, primarily due to their reliance on large-scale architectures that require significant computational resources and memory. This limits their practical applicability, especially when capturing nuanced relationships between objects, spatial arrangements, and contextual interactions within a scene.

- Many state-of-the-art models are capable of detecting individual objects; however, they often fall short in capturing fine-grained details and complex semantic relationships among those objects. Moreover, achieving this level of detailed understanding typically demands resource-intensive architectures, which makes it challenging to deploy such models in real-time or memory-constrained environments.

- Another significant gap lies in the limited generalization capability of current models across diverse and domain-specific datasets. While many models perform well on standard benchmark datasets, they often exhibit poor adaptability when exposed to unseen images or specialized domains such as medical imaging. This limitation is further compounded by these models' high computational and memory demands, which hinder their scalability and practical deployment in real-world applications.

- Models based on deep learning usually require large datasets labelled appropriately to be trained correctly. However, special domains like chest X-ray analysis and video-based image description generation usually lack enough data, which creates challenges in developing models that can do well in real-time applications or multimedia content analysis.

- Although existing models only rely on visual data, they are limited in their capacity to capture intent, emotions, and deeper context. Therefore, integrating multimodal learning with audio, depth perception, or external knowledge might improve description accuracy.

- Since models frequently inherit biases from training data, addressing bias in image captioning is a significant challenge. This leads to skewed or conventional descriptions that inaccurately represent a variety of demographics and cultures.

- Another significant challenge is ensuring real-time description generation. Latest deep learning models, especially transformer-based architecture, require a huge number of computational resources. Thus, it becomes hard to deploy these models on resource-constrained environments and for real-time applications

The significance for precise, contextually appropriate, and semantically rich image descriptions in practical applications is what motivates research to fill these research gaps. Enhancing visual-semantic alignment, obtaining fine-grained image details, and guaranteeing model generalisation across various datasets are some of the main problems. Incomplete descriptions result from many models' inability to capture subtle object relationships, spatial arrangements, and contextual interactions. Furthermore, deep learning models need large, labelled datasets, yet there is frequently insufficient data in specialised domains like medical imaging and autonomous driving. The necessity for multimodal learning integration is highlighted by the fact that existing models only use visual data, which limits their capacity to capture intent, emotions, and deeper context. In addition to producing skewed or stereotypical descriptions, bias in training data calls for mitigating techniques to guarantee equitable representation. Real-time performance is another major issue because transformer-based designs require a lot of processing power, which makes deployment challenging in contexts with limited resources. By addressing these issues, this study seeks to improve image description models' precision, effectiveness, and applicability in a variety of crucial fields.

1.4 Problem Definition

The task of image description generation requires coherent, contextually accurate, and semantically rich textual descriptions from visual content. Existing models fail to adequately align visual information with textual descriptions, especially when dealing with complex scenes containing multiple objects or intricate relationships. Generalisation across various datasets or strong performance in domain-specific applications like autonomous driving or medical imaging are also not achieved by existing models due to a lack of specialisation in data for such applications. Further, while deep learning-based models have made tremendous strides, they need huge, labelled datasets for proper training and substantial computational resources, making them difficult to deploy in real-time applications, especially in resource-constrained environments. The proposed research will address the challenges by designing an automatic framework for generating contextually and semantically complete image descriptions, a deep learning framework to enhance the accuracy of descriptions, and adaptable models for real-time, domain-specific applications. This work attempts to advance the state of the art in image description generation towards greater reliability, efficiency, and applicability in a wider variety of real-world scenarios.

1.4.1 Research Objectives

OBJECTIVE 1: To design an automatic framework for generating semantically rich image descriptions by aligning visual content with text and capturing object-scene relationships.

OBJECTIVE 2: To develop a transformer-based deep learning framework that improves image descriptions by effectively capturing visual details, contextual relationships, and multimodal alignment.

OBJECTIVE 3: To develop image description models for medical imaging and video-based image description generation in multimedia applications.

1.5 Contributions in the Thesis

In this thesis, we focus on image description generation and identify several key research gaps in this field. We are focusing on the whole area of generating contextually accurate and semantically complete descriptions for images, which involves huge challenges. In Section 1.4.1, three research objectives are outlined. We then sequentially address each of these objectives as given in the contributions of the thesis:

(I) **OBJECTIVE 1:** ***To design an automatic framework for generating semantically rich image descriptions by aligning visual content with text and capturing object-scene relationships.***

The difficulty of capturing semantic meaning, spatial context, and fine-grained object relationships in image descriptions motivated us to suggest a model that combines visual feature extraction, object relationships, and sequence learning for increased accuracy. We proposed an image description generation method that utilised VGG16 for extracting visual features from the images, scene graphs to understand the object relationships, and a BiGRU model to learn sequence-to-sequence representations. The scene graph enhances the contextual understanding since it captures both spatial and semantic relationships between the objects in the image. Visual features extracted using VGG16 are combined with scene graph embeddings and processed within a BiGRU network within the proposed model. The model is able to generate coherent and contextually meaningful descriptions due to the bidirectional processing capability in BiGRU. It is trained on MS COCO, Flickr8k, and Flickr30k datasets, and performance is evaluated in terms of BLEU (1-4), CIDEr, METEOR, and ROUGE-L metrics. Experimental results confirm that the proposed VGG16-SceneGraph-BiGRU model outperforms competitive approaches, achieving better accuracy and semantic enrichment for the descriptions of the image.

(II) **OBJECTIVE 2:** ***To develop a transformer-based deep learning framework that improves image descriptions by effectively capturing visual details, contextual relationships, and multimodal alignment.***

By adding transformer-based fusion and dual attention techniques, we aimed to improve image description generation further than what we achieved with our previous objective. Tri-FusionNet, a revolutionary deep-learning framework, had been developed as a result, greatly improving the generated descriptions' accuracy and prediction performance. Tri-FusionNet combines a dual attention mechanism to enhance emphasis on spatial and contextual variables, a Vision Transformer (ViT) encoder to extract fine-grained visual features, and a RoBERTa decoder to produce fluid and semantically rich textual descriptions. Using contrastive learning, the addition of a CLIP-integrating module also makes it easier to match textual and visual representations. These elements work together to give the model the ability to produce descriptions that are more accurate and contextually rich. Tri-FusionNet demonstrated its supremacy in image description generation by achieving state-of-the-art performance across BLEU (1-4), CIDEr, METEOR, and ROUGE-L metrics after being trained on the MS COCO, Flickr8k, and Flickr30k datasets.

(III) **OBJECTIVE 3:** ***To develop image description models for medical imaging and video-based image description generation in multimedia applications.***

In order to broaden the scope of our research, we developed transformer-based frameworks specifically designed for multimedia applications, particularly for autonomous driving and medical imaging. These domain-specific models demonstrate the versatility and applicability of the image description generation technology. We proposed two advanced transformer-based frameworks for image description generation in multimedia applications, including medical imaging and autonomous driving. The first model is for advanced Chest X-Ray Analysis, where a Vision Transformer (ViT) is utilised for visual feature extraction and a GPT-4-based decoder with cross-model attention is added to improve the contextual accuracy of descriptions. Therefore, cross-model attention effectively combines both visual and textual elements, enabling more precise as well as understandable chest X-ray image descriptions. It was applied on the National Institutes of Health (NIH) and the Indiana University Chest X-ray dataset with state-of-the-art performances across multiple metrics to demonstrate how the model would assist radiologists in diagnosis as well as the development of suitable treatment plans. Understanding and analyzing video actions are essential for producing insightful and contextualized descriptions, especially for video-based applications like intelligent monitoring and autonomous systems. This work introduces a novel framework for producing natural language descriptions from images and videos by combining textual and visual modalities. The suggested architecture makes use of ResNet50 to extract visual features from video frames that are taken from the Microsoft

Research Video Description Corpus (MSVD), and Berkeley DeepDrive eXplanation (BDD-X). The extracted visual characteristics are converted into patch embeddings and then run through an encoder-decoder model based on transformers and GPT-2. In order to align textual and visual representations and guarantee high-quality description production, the system uses multi-head self-attention and cross-attention techniques. The model's efficacy is demonstrated by performance evaluation using BLEU (1-4), CIDEr, METEOR, and ROUGE-L.

1.6 Outlines of the Thesis

This thesis is divided into seven chapters and five appendices.

1. **Chapter 1 Introduction**

   The objective of this chapter is to overview image description generation, emphasizing that multi-modal fusion enhances the description quality. Motivation for conducting the research has been identified in terms of describing key research gaps and challenges followed by defined objectives of the study. Finally, a brief outline of the structure of the thesis is presented.

2. **Chapter 2 Literature Survey**

   This chapter discusses different approaches to image captioning, from traditional rule-based techniques to deep learning and transformer-based models. It emphasizes the importance of multi-modal learning in improving caption generation and explores its applications in medical imaging and autonomous driving. The chapter also discusses commonly used evaluation metrics to measure model performance and effectiveness.

3. **Chapter 3 Proposed a Unified Framework for Contextual and Semantic Image Description Generation**

   This chapter introduces the unified model that combines VGG16, Scene Graphs, and BiGRU model for the generation of semantically rich and contextually complete descriptions. It discusses the use of VGG16 for feature extraction, scene graphs to capture object relationships, and BiGRU for sequential learning. The data details, training methodology, and performance evaluation are also discussed.

4. **Chapter 4 Deep Learning Framework for Improved Image Description Accuracy**

   The proposed chapter presents a transformer-based framework called Tri-FusionNet designed to improve description generation accuracy. This focuses on the fusion of the ViT, RoBERTa, and CLIP modules. It also involves the designing of a dual attention mechanism. The dataset selection, modelling training process, and evaluation metric are then discussed. Finally, it compares the architecture with state-of-the-art models.

5. **Chapter 5 Image Description Models for Multimedia Application: Chest X-Ray Analysis**

   This chapter covers one domain-specific application of image description generation. This part discusses the Advanced Chest X-Ray Analysis model based on ViT-GPT4, using cross-model attention for medical image description generation.

6. **Chapter 6 Image Description Models for Multimedia Application: Video-based Image Description Generation**

   This chapter presents a transformer-based GPT-2 model that integrates visual and textual modalities to generate action-based descriptions for video datasets leveraging ResNet50 and evaluated on standard metrics.

7. **Chapter 7 Conclusion and Future Directions**

   This chapter summarises key contributions of the thesis, improvements realised in terms of image description generation, significant findings, discussions regarding the effect of the models proposed, and potential directions for further work, which might include advances on real-time captioning as well as better domain-specific applications.

8. **Appendix A Vision Transformer (ViT)**
   In this appendix we describe the detailed architecture of ViT, a self-attention mechanism, as well as using it for image description generation based on feature extraction.

9. **Appendix B RoBERTa (Robustly Optimized BERT Approach)**
   In this appendix, we describe how RoBERTa is constituted, its architecture based on a transformer, as well as how it significantly improves the description generation capability to be much more context-rich and semantically accurate.

10. **Appendix C Bidirectional Gated Recurrent Unit (BiGRU) Model**
    In this appendix, we give a detailed description of the Bi-GRU model, including its bidirectional processing, gating mechanisms, and its role in sequential learning for image description generation.

11. **Appendix D Scene Graphs**
    In this appendix, we discuss scene graphs, how they are created, and what they add to the interpretation of object relationships by image captioning models.

12. **Appendix E GPT-2 and GPT-4 Transformers**
    In this appendix, we expand on the GPT-2 and GPT-4 transformers, language modeling with these transformers, and how these can be used for generating natural language descriptions of images.

The Figure 1-1 represents an overall graphical abstract of the chapter-wise distribution of the thesis, illustrating the research problem, objectives, proposed models, and key outcomes on image description generation using deep learning.

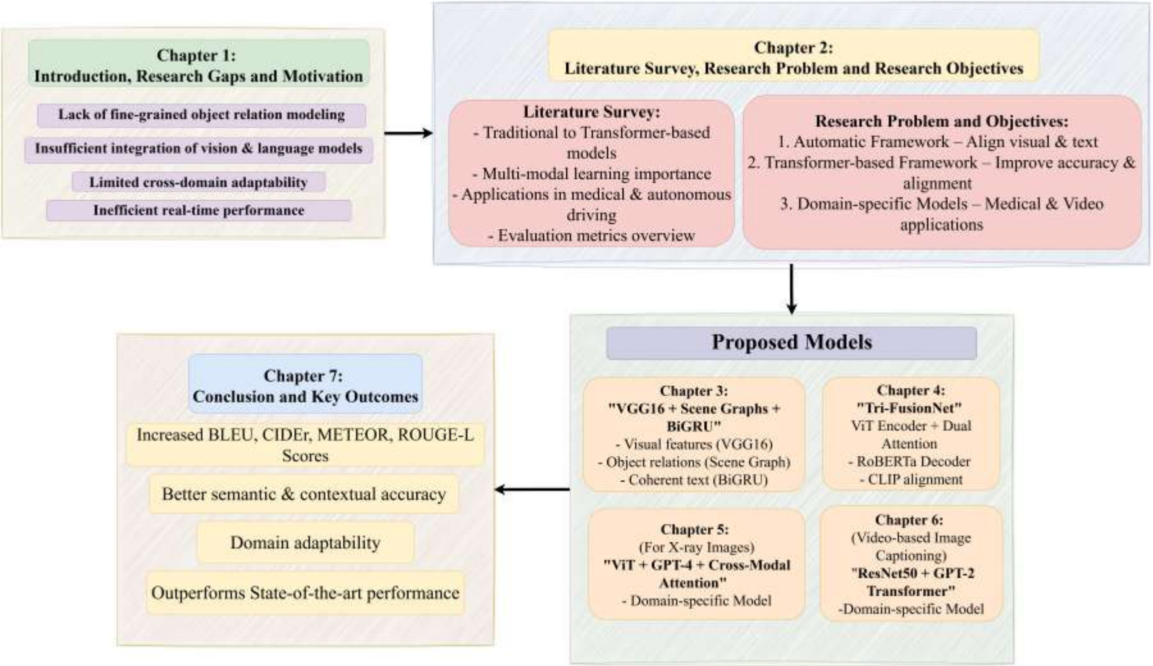

Chapter 1:
Introduction, Research Gaps and Motivation
Lack of fine-grained object relation modeling
Insufficient integration of vision & language models
Limited cross-domain adaptability
Inefficient real-time performance
Chapter 2:
Literature Survey, Research Problem and Research Objectives
Literature Survey:
- Traditional to Transformer-based models
- Multi-modal learning importance
- Applications in medical & autonomous driving
- Evaluation metrics overview
Research Problem and Objectives:
1. Automatic Framework – Align visual & text
2. Transformer-based Framework – Improve accuracy & alignment
3. Domain-specific Models – Medical & Video applications
Proposed Models
Chapter 3:
"VGG16 + Scene Graphs + BiGRU"
- Visual features (VGG16)
- Object relations (Scene Graph)
- Coherent text (BiGRU)
Chapter 4:
"Tri-FusionNet"
ViT Encoder + Dual Attention
- RoBERTa Decoder
- CLIP alignment
Chapter 5:
(For X-ray Images)
"ViT + GPT-4 + Cross-Modal Attention"
- Domain-specific Model
Chapter 6:
(Video-based Image Captioning)
"ResNet50 + GPT-2 Transformer"
-Domain-specific Model
Chapter 7:
Conclusion and Key Outcomes
Increased BLEU, CIDEr, METEOR, ROUGE-L Scores
Better semantic & contextual accuracy
Domain adaptability
Outperforms State-of-the-art performance

Figure 1-1: Graphical abstract of the overall chapter-wise distribution of the thesis.

Chapter 2

Literature Survey

Lakshita Agarwal and Bindu Verma. "From methods to datasets: A survey on Image-Caption Generators", Multimedia Tools and Applications (2024): Volume 83, pages 28077–28123. Impact Factor-3.0 *(Published).*

Lakshita Agarwal and Bindu Verma. "Comparison of Deep Learning Models for Automatic Image Descriptors", in *2023 IEEE 20th India Council International Conference (INDICON): 914-919,* IEEE (December, 2023) *(Published).*

This chapter gives an overview of the state-of-the-art approaches in image description generation. These approaches range from traditional methods to deep learning-based methods. Methods in this area are critically reviewed. Traditional image description generation often involves stages, such as the extraction of image features, detection of objects, and template-based generation of descriptions. The feature extraction process of these methods involves unrefined features and predefined rules. In contrast, deep learning-based methods automate the process of feature extraction and description generation using models like CNNs for visual feature extraction and RNNs for text generation. The recent advancements with transformer-based models like Vision Transformer (ViT) and GPT have further improved the quality of generated descriptions. The traditional and deep learning-based approaches have been given a lot of attention in the literature, with each method contributing to the evolution of image description generation systems.

2.1 Literature Survey on Traditional Approaches for Image Description Generation

Early stages of image description generation relied heavily on unrefined features, rule-based models, and template-based description generation systems. These methods focused on identifying objects within an image and used predefined templates to generate simple descriptions. However, these systems were limited by their inability to handle complex interactions between objects or generate diverse, meaningful descriptions. Jeon et al. [17] proposed an intuitive approach for annotating and retrieving images using a small vocabulary of blobs to characterize image regions. They demonstrated the effectiveness of Cross-Media Relevance Models for ranked retrieval and suggested that formal information retrieval models could enhance this research area. The authors highlighted the need for labeled training data for algorithm evaluation and suggested improved feature extraction or continuous features would improve results. They also presented the possibility of using real descriptions instead of keywords for image annotation. Li et al. [18] proposed a statistical modeling approach to automatic linguistic indexing of images based on two-dimensional multi-resolution hidden Markov models (2D MHMMs). Their technique provided independent concept learning, scalability, and maintained spatial relationships between pixels. The experiments were promising and accurate, yet they had problems with learning hard concepts from a 2D image and high visual diversity within some categories. The authors themselves suggested that future work was essential to overcome such limitations.

Rennie et al. [19] developed self-critical sequence training: a reinforcement learning-based approach toward improving image description generation using the Reinforce algorithm. A bi-linear model was introduced by Lebret et al. [6] to generate image descriptions by linking image representations with descriptions. Their phrase-based language model, which used syntactic statistics, offered a simpler yet effective alternative to RNN-based models on MS-COCO; further improvements could refine image-description ranking and increase its adaptability to other datasets. Another traditional method, retrieval-based methods [20], matches input images with similar pre-annotated ones and chooses the best description using similarity metrics like K-Nearest Neighbours (KNN) [21] and Support Vector Machines (SVMs) [22]. These methods worked well for structured datasets, but they had trouble with diverse, complex images and failed to capture long-range dependencies, resulting in lower-quality, less contextual captions.

Expanding on these conventional techniques, it is clear that they have a number of significant limitations that reduce their capacity to produce detailed and contextualised image descriptions. Despite being uncomplicated, template-based methods are rigid and result in generic, repeating captions that fail to convey the individuality of various images. Likewise, retrieval-based techniques are ineffective for new or complicated scenes with numerous interacting objects since they depend on the presence of comparable images in a pre-annotated dataset. These methods' capacity to generalise beyond the training data is constrained by their inability to capture the complex relationships between objects and their difficulty adapting to a variety of datasets. Furthermore, they fail to take semantic understanding into account, which results in descriptions that are shallow and inaccurately contextualised.

2.2 Literature Survey on Deep Learning Methods for Image Description Generation

In order to overcome the drawbacks of conventional techniques, deep learning-based strategies become an effective solution. The most common approach for visual feature extraction was Convolutional Neural Networks (CNNs), whereas sequential descriptions were produced by Recurrent Neural Networks (RNNs), especially Long

Short-Term Memory (LSTM) networks. In contrast to traditional techniques, deep learning models could generate varied and cohesive descriptions by learning patterns from enormous datasets. By enabling individuals to concentrate on important areas of an image when creating descriptions, attention mechanisms significantly improved these models and increased the usefulness of the descriptions. By refining a description quality depending on evaluation metrics, reinforcement learning approaches like self-critical sequence training [19] improved the performance of these deep learning models. A number of deep learning-based techniques showed significant advances in image description generation. The hierarchical LSTM-based method (phi-LSTM), which was presented by Chan et al. [23], successfully organized natural language descriptions. In order to compete with supervised methods, Fung et al. [24] created an unsupervised model that simply needed a visual concept detector, a description corpus, and an image dataset. The LSTM-A model was introduced by Yao et al. [25] to improve description production by combining CNN-RNN architectures with attribute-based learning. SentiCap, a sentiment-aware model that enhanced the capacity to produce emotion-driven descriptions, was also presented by Mathews et al. [26]. These developments demonstrated deep learning's superiority over traditional approaches. Still, they also exposed important drawbacks, including poor adaptation to previously unseen data, trouble identifying long-range relationships, and inefficiencies in handling complex image relationships. Even with the advancements in deep learning, attention-based RNN models have their own drawbacks. They found it difficult to express intricate scene compositions and vague object interactions. The training was inefficient due to RNNs' sequential nature, which limited their capacity to manage long-term dependencies by processing input separately. As a result, transformer-based architectures gained popularity, bringing with them self-attention mechanisms capable of modelling whole input sequences at once.

By enhancing visual and textual information alignment, transformer models like Vision Transformer (ViT) and GPT-based language models greatly improved image captioning. This alignment was further improved using CLIP (Contrastive Language-Image Pre-training), which allowed models to produce accurate and fluent descriptions. In order to outperform recurrent models on benchmark datasets, Cornia et al. [27] introduced the M2 Meshed Transformer, which incorporates memory mechanisms. The EnTangled Attention (ETA) Transformer, developed by Li et al. [28], successfully closed the semantic gap between language and vision. Furthermore, Yu et al. [29] suggested a multi-modal transformer that enhanced the precision of image descriptions by capturing intra- and inter-modal interactions. Transformer-based model developments in recent years have kept pushing the limits of image description generation. MTTSNet, which creates object-based descriptions while capturing object relationships for improved recall, was first presented by Kim et al. [30]. FCLN, as proposed by Johnson et al. [31], facilitates effective region localization and description creation. Nevertheless, many of these methods still have trouble expanding their vocabulary and dealing with the constraints of textual context. This problem was solved by Shao et al. [32] using the Enhanced Transformer Dense Captioner (ETDC), which expanded vocabulary usage and included contextual text. Furthermore, by treating regions of interest (RoIs) differently, their Transformer-based Dense Captioner (TDC) [33] enhanced adaptability and allowed for better domain-specific applications.

These developments demonstrate how conventional rule-based and retrieval techniques led way to deep learning strategies and, eventually, transformer-based systems. Each phase of development has improved the generated image descriptions' accuracy, fluency, and contextual comprehension. Problems still exist even though transformer-based models have raised the bar for image captioning by successfully capturing long-range dependencies and coordinating textual and visual information. Research on topics including domain adaptability, multimodal comprehension, and effective real-time processing is still going on. Resolving these issues is essential to expanding the range of disciplines in which image description generation can be used and laying the foundation for understanding its numerous practical applications.

2.3 Literature Survey based on Applications of Image Description Generation

Early efforts to create associations between written descriptions and visual attributes marked the beginning of the path toward generating accurate descriptions. The automatic generation of image descriptions by Pan et al. [34], which linked image features with keywords in a captioned training set, was one of the pioneering initiatives. In large and varied data sets, their approach obtained an accuracy of about 45%. It may be used in medical image information and approaches like Singular Value Decomposition-based correlation (SvdCorr), Cosine similarity (Cos), Singular Value Decomposition-based cosine similarity (SvdCos), and Correlation (Corr). By adding blob-tokens, they improved performance even further. Researchers switched to increasingly complex models as the field advanced. Xiong et al. [35] presented a hierarchical Transformer-based model in medical imaging that performed better than alternative methods for generating medical reports. As a more sophisticated method, dense description generation created several detailed phrases to explain different areas of an image [36]. Automatic captioning, which provides real-time scene descriptions to improve accessibility, has proved crucial in assistive technology for those with visual impairments outside medical imaging [37]. Vision-language models aid in scene comprehension in autonomous driving by describing traffic situations, road conditions, and obstructions [38]. This helps self-driving systems make better decisions. E-commerce systems also use automated captioning to generate product descriptions, which improves user engagement and searchability.

However, there are still great obstacles to overcome. Deep learning models frequently have trouble generalising across datasets and domains like medical imaging, remote sensing, and industrial automation, and visual-semantic alignment is especially challenging in complex images with several interacting elements [39]. In environmental analysis, land-use classification, and disaster monitoring, where accurate textual descriptions can facilitate quick decision-making, satellite imagery captioning is essential [40] [41]. Automatic image captioning helps with threat detection and anomaly reporting in security and surveillance by instantly summarising video feeds. Furthermore, real-time performance issues arise due to transformer-based architectures' processing demands, and conventional evaluation measures fall short in capturing the semantic depth of output descriptions.

The following Table 2.1 is a comprehensive summary of some key image description generation methods and their pros and cons.

Table 2.1: Descriptive Summarization of Image Description Generation Approaches

Model Proposed	**Advantages**	**Disadvantages**
Cross-Media Relevance Models [17]	Uses a small vocabulary of blobs for image region characterization. Useful for annotating and retrieving photos.	Does not extract image features. Cannot be applied to larger datasets.
2D MHMMs [18]	Utilizes stochastic processes for image captioning. A statistical modeling approach.	Difficult to teach concepts with a small number of images. Time-consuming and complex to handle.
Automatic image annotation method [42]	Combines picture indexing and NLP techniques. Lays a solid foundation for complex images.	Requires extensive research in NLP and lacks high accuracy.
Corr, SvdCorr, SvdCos and Cos models [34]	Links keywords with image features. Achieves 45% accuracy on larger datasets.	Time-consuming, as it uses blobs instead of full sentences.
Domain-specific image captioning model [43]	Deletes erroneous terms while maintaining high detail. Uses both automatic and human evaluations.	Data-driven and limited to specific fields. Difficult to understand and implement.

Model Proposed	**Advantages**	**Disadvantages**
Neural captioning system: Attention model with GRU [44]	Generates meaningful descriptions for images. Outperforms previous models.	Focuses on specific parts of the image, limiting broader scene understanding.
Midge Framework [45]	Generates the most natural descriptions of images.	Needs refinement to capture more linguistic phenomena.
Reinforcement learning (RL) technique [46]	Leverages RL for caption generation and considers human feedback.	Needs improvement in accuracy.
Captions with guiding objects (CGO) [47]	Provides fluent and accurate captions.	Focuses on one object, limiting multi-object scene description. Struggles with multiple objects in images.
Bi-directional model [12]	Learns long-term interactions and generates descriptive captions.	Requires further study in RNN-based approaches.
Multi-modal RNN model (m-RNN) [48]	Connects images and sentences for complex representations.	Embedded layers can become large with a bigger dictionary.
SentiCap RNN model [26]	Generates sentiment-based captions. 88% accuracy for positive captions.	Limited in handling emotions like pride or wrath.
LSTM-A (Long Short-Term Memory with Attributes) model [25]	Integrates attributes into CNNs and RNNs for image captioning.	Needs improvements in model accuracy.
Multiple Instance Learning (MIL) framework [49]	Uses attention processes for image captioning.	Validation accuracy is not high across multiple datasets.

Model Proposed	**Advantages**	**Disadvantages**
Convolution image captioning [50]	Uses CNN for image captioning. Performs better than LSTM+Attn baseline.	Lacks sequential elements found in RNN-based models.
Graph Convolutional Networks (GCNs) models [51]	Aligns linguistic words and visual semantic units for captions. Uses structured graphs and GCNs for contextual embedding.	Complex to implement without access to VSUs in datasets.
GCN-LSTM architecture [52]	Integrates spatial and semantic relationships in image captions. Increases CIDEr-D performance by 8.6% on MS-COCO.	Attention mechanism still needs thorough validation. Slow on larger datasets.
X-Linear Attention (X-LAN) block architecture [53]	Focuses on bi-linear pools for attention. CIDEr performance improves by 11%.	Can focus on irrelevant areas of the image.
Context-Aware Auxiliary Guidance (CAAG) model [8]	Uses global context for improved captioning. Enhances existing models in reinforcement learning.	CIDEr-D performance of 128.8% is lower than X-LAN's 132.0%.
Self-critical sequence training (SCST) model [19]	Uses Reinforce algorithm for policy-gradient captioning.	Did not produce appreciable gains.
SPIDEr model [54]	Combines SPICE and CIDEr scores for caption evaluation. Qualitatively improves human evaluation.	Often produces illegible sentences and repetitions. Insensitive to syntactic quality.

Model Proposed	Advantages	Disadvantages
Bi-linear model [6]	Links image representations with sentences. Useful for phrase identification.	Needs further research on image-sentence ranking. Needs language model improvement.
phi-LSTM Architecture [23]	Describes salient features of objects in an image.	Does not produce relevant captions for images.
Retrieval-based methods [13]	Solves ranking problems using multi-modality techniques. Uses recursive dependency trees for feature extraction.	Needs further improvements in accuracy.
Unsupervised Image captioning model [24]	Uses an encoder, sentence generator, and discriminator. First unsupervised approach using Shutterstock.	Limited accuracy on labeled image-sentence pairs.

In this work, we analyze the distribution of various image captioning approaches explored in the literature. Figure 2-1 presents a pie chart illustrating the proportion of different methodologies employed for generating image descriptions.

Most of the existing works adopted RNN-based architectures (45%), and then attention-based models at 30%, followed by transformer-based models at 15%, and other techniques accounted for 10%. This shows that, in image description generation, most research is still relying on pure RNN-based structures because they are robust to sequential dependencies. Deep learning has been the key area of research for modern image description generation efforts, primarily for its ability to merge computer vision and natural language processing. Robust architectures are needed to generate accurate and context-aware descriptions. Attention mechanisms, adversarial learning, and deep reinforcement learning have gained much attention in recent years, as shown in this study. Long Short-Term Memory (LSTM) networks, along with Faster

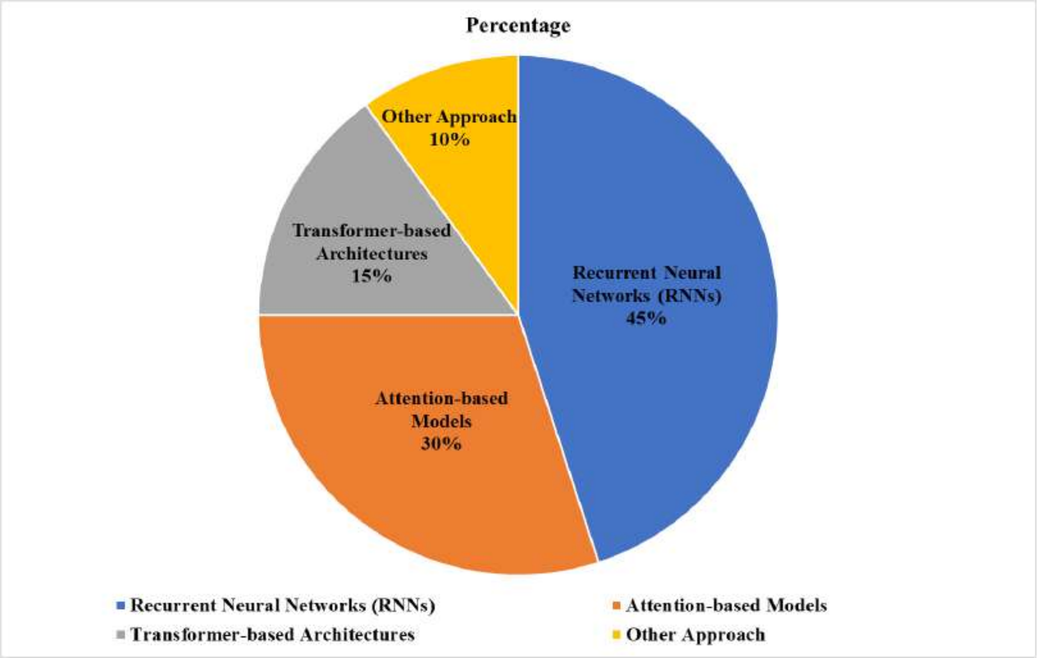


Figure 2-1: Distribution of approaches used in image description generation [55].

R-CNN for object detection, are still being used widely, which further strengthens the need for structured visual and textual representations to be integrated for effective description generation.

2.4 In Thesis Prospective:

In computer vision and natural language processing, creating semantically rich visual descriptions has consistently been a challenge. While deep learning models employing CNNs and RNNs enhanced captions, they still had problems with alignment and long-range dependencies, while traditional approaches had trouble with diversity and object interactions. Although Transformers solved a lot of issues, problems with real-time processing and domain-specific applications still exist.

Traditional image captioning approaches frequently produced fragmented or too simplified descriptions because they had trouble appropriately capturing object relationships, spatial interdependence, and contextual significance. Despite advancements, many deep learning models still lacked a systematic method for simulating object interactions and contextual dependencies inside an image since they only used

CNNs for feature extraction and RNNs for sequence creation. Due to these constraints, a more reliable framework that could produce comprehensive and contextually appropriate descriptions had to be created. The first suggested framework combines scene graphs, VGG16, and a BiGRU-based sequence model to improve contextual understanding and visual feature extraction to overcome these difficulties. BiGRU creates organized textual descriptions that guarantee coherence and contextual accuracy; VGG16 captures crucial visual components; and the scene graph module enhances understanding of the spatial and semantic relationships between objects. This method achieves high accuracy levels across BLEU, CIDEr, METEOR, and ROUGE-L evaluation parameters and substantially enhances description fluency and semantic relevance. It was trained on the MS COCO, Flickr8k, and Flickr30k datasets.

Although the initial framework was successful in capturing spatial and semantic linkages, more enhancements were required to enable more dense descriptions and improved global context processing. We improved efficiency by introducing a transformer-based model that uses advanced attention mechanisms to improve contextual correctness and visual-textual alignment. The second framework presents Tri-FusionNet, a deep-learning model that combines Vision Transformer (ViT), RoBERTa, and CLIP for better image description generation to further increase performance. ViT improves image representation by capturing fine-grained features and ensuring improved spatial focus on important objects through its dual attention methods. While CLIP improves visual-textual coherence by refining cross-modal alignment through contrastive learning, RoBERTa strengthens textual encoding for more accurate and contextually aware description generation. Tri-FusionNet achieves state-of-the-art performance, greatly improving the semantic richness, contextual relevance, and description fluency after being trained on the MS COCO, Flickr8k, and Flickr30k datasets.

To extend image description generation applications to specialized domains, two domain-specific frameworks are proposed. The first framework, CrossViT-GPT4, is designed for medical image captioning, utilizing ViT for feature extraction and GPT-4 with cross-modal attention to generating radiology reports. Evaluated on the NIH and IU Chest X-ray datasets, it demonstrates high diagnostic accuracy and aids radiologists in decision-making. The second framework fine-tunes GPT-2 on the Flickr8k and BDD-X datasets for generating action-justification pairs for vehicle behavior descriptions to improve the interpretability of autonomous driving systems by enhancing the transparency of decisions and explainability. These models are effective for multimedia applications. It is ensured through the incorporation of attention mechanisms and transformer-based architectures, providing robust performance in complex visual scenes involving multiple interacting objects. In addition, efficiency improvements make real-time deployment possible, which means that autonomous vehicles and AI-driven medical diagnosis systems are feasible. These proposed approaches yield substantial improvements in description accuracy, fluency, and semantic relevance, paving the way for more reliable and interpretable AI-driven multimodal applications.

Chapter 3

Proposed a Unified Framework for Contextual and Semantic Image Description Generation

Lakshita Agarwal and Bindu Verma. "Enriching Image Description Generation through Multi-modal Fusion of VGG16, Scene Graphs and BiGRU", The Visual Computer (2025): 1-21. Impact Factor-3.0 ***(Published).***

3.1 Introduction

In this chapter, we discuss the challenges involved in producing precise and contextually rich image descriptions, which is an essential problem for applications like autonomous navigation, information retrieval, accessibility for the blind, and human-computer interaction. Traditional approaches frequently have trouble guaranteeing fluency in the generated language, preserving contextual coherence, and capturing fine-grained visual information. We suggest a framework that combines scene graphs for contextual comprehension, VGG16 for visual feature extraction, and Bi-directional Gated Recurrent Unit (BiGRU) for sequence modelling to address these problems. While VGG16 collects deep visual features and BiGRU processes them in both directions to improve sequence learning, scene graphs improve the representation of object relationships. Standard evaluation measures, including BLEU, CIDEr, METEOR, and ROUGE-L, are used to train and assess the model on benchmark datasets, including MS COCO, Flickr8k, and Flickr30k. Results from experiments show that our method performs better than current approaches, confirming the usefulness of combining deep visual characteristics, scene graphs, and BiGRU for reliable image description generation.

The key contributions of the chapter are mentioned below:

- The work suggests a novel approach that integrates scene graph, VGG16 and BiGRU for image description generation. The strengths of each component are combined in a novel way to improve contextual awareness and produce descriptions for complex visual images that are more accurate and relevant.

- The combination of BiGRU, scene graph and VGG16 is useful to produce descriptions that are more insightful and accurate as well as which closely match the content of the images.

- The model performs well when dealing with complicated visual scenarios that contain several objects, a variety of properties and sophisticated interactions. It can capture fine-grained details and interactions between objects due to the structured representation of the scene graph and bidirectional context understanding of BiGRU.

- On benchmark datasets including MS COCO, Flickr8k and Flickr30K, the performance of the proposed framework is thoroughly evaluated, proving its superiority to other state-of-the-art image description generation techniques. The research offers brief performance comparisons and analysis, further supporting the validation of the model.

- The model has an improved contextual comprehension of the descriptions, improved description generation quality and is used for handling challenging situations appropriately. It represent a significant achievement in the field and opens the way for more complex and context-aware image description generation systems.

3.2 Literature Survey

Image description generation is an interdisciplinary research area combining computer vision and Natural Language Processing (NLP). It involves recognizing objects and representing their relationships through accurate and semantically meaningful descriptions [16]. Jin Dai et al. [56] introduced a multi-modal attention-based model using ResNet-101 for feature extraction and Faster R-CNN for object recognition. A multi-head attention mechanism enhanced linguistic learning, while GPU parallel computing accelerated training. Comparative studies demonstrated improved captioning accuracy.

Deep learning frameworks are widely used in computer vision and human-computer interaction. Chen et al. [12] proposed a bi-directional mapping approach for images and captions, integrating GRU with Bahdanau's attention model. Attention mechanisms, crucial for image description, enable models to focus on specific image regions and capture long-range dependencies [57]. Yucong et al. [58] combined BiGRU with attention for image description, showing promising results but requiring further optimization. Transformer models incorporating local graph semantic attention [59] improved captioning by fusing semantic and spatial data, while Li et al. [60] enhanced content and structural relations using geometric and semantic graphs.

Recent works integrate scene graphs for better object attribute representation [61]. Li et al. [62] developed a model combining object detection, scene graph construction, and region-based description generation. Scene graphs are often paired with LSTMs, as seen in Xu et al. [63], who proposed a framework for structured representations. Yang et al. [64] introduced the Scene Graph Auto-Encoder (SGAE) to generate more human-like descriptions. To enhance performance on MSCOCO, Zhao et al. [65] developed a Multi-Level Cross-Modal Alignment (MCA) module for aligning description and image scene graphs while reducing noise. A transformer-based approach integrating scene graphs [66] leveraged a Graph Convolutional Network (GCN) for improved captioning, achieving state-of-the-art results on MSCOCO and Flickr30k.

This work extends existing research by developing a more effective image description model capable of handling complex visual scenes and improving structured information extraction. The objective is to enhance description quality and adaptability for real-world applications, pushing the boundaries of traditional approaches for context-aware image understanding.

3.3 Proposed Architecture

The proposed research suggests a hybrid framework that combines the capabilities of VGG16, scene graphs and BiGRU in order to address the limitations. Scene graphs and bidirectional processing are used in the suggested model to better capture object relationships and context. The approach also focuses on solving the difficulties in creating thorough and semantically meaningful descriptions for images. The aim is to optimize the model parameters in order to maximize the evaluation metrics that measure the similarity between the descriptions that are generated and the descriptions provided for reference. Figure 3-1 represents the proposed model framework.

In the proposed work, the model receives an input image as the initial stage. To maintain uniformity in pixel values and dimensions, the image is firstly pre-processed. The visual features of a high-level present inside the pre-processed images are then extracted using a VGG16 model. The output of the features are then stored in a fixed-size vector of dimensional features, which is used for capturing significant visual patterns and representations in the images. The model creates a scene graph for the images after receiving the visual attributes. Then, the scene graph and VGG16 features are integrated and a Bidirectional Gated Recurrent Unit (BiGRU) is fed the output information. After going through the VGG16, scene graphs and BiGRU layers to process the input image, the model generates a description for the image. The overview of the components present in the framework are discussed.

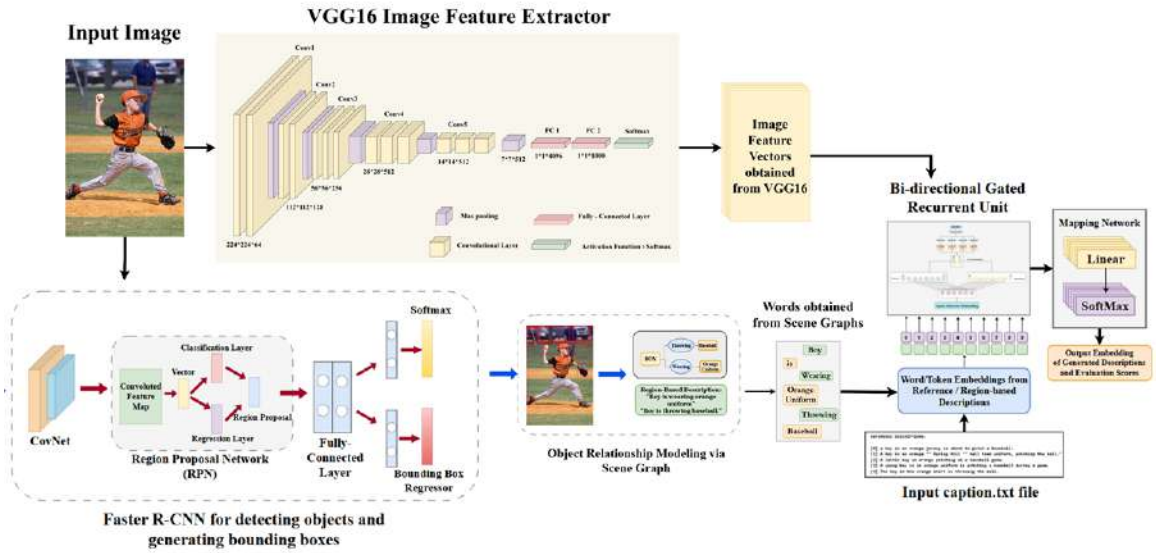

Input Image
VGG16 Image Feature Extractor
Conv1
Conv2
Conv3
Conv4
Conv5
FC 1
FC 2
Softmax
224*224*64
112*112*128
56*56*256
28*28*512
14*14*512
7*7*512
1*1*4096
Max pooling
Fully - Connected Layer
Convolutional Layer
Activation Function : Softmax
Image Feature Vectors obtained from VGG16
Bi-directional Gated Recurrent Unit
Mapping Network
Linear
SoftMax
Output Embedding of Generated Descriptions and Evaluation Scores
CovNet
Classification Layer
Vector
Convoluted Feature Map
Region Proposal
Regression Layer
Region Proposal Network (RPN)
Fully-Connected Layer
Softmax
Bounding Box Regressor
Faster R-CNN for detecting objects and generating bounding boxes
Object Relationship Modeling via Scene Graph
Words obtained from Scene Graphs
Boy
Is
Wearing
Orange Uniform
Throwing
Baseball
Word/Token Embeddings from Reference / Region-based Descriptions
Input caption.txt file

Figure 3-1: Structural representation of the proposed framework: The architecture consists of three phases. In the initial phase, the VGG16 model extracts the high-level visual characteristics from the pre-processed images. A scene graph creation model for extracting visual attributes in the next step. Finally, feeding this combined information into a Bidirectional Gated Recurrent Unit (BiGRU) for generating the image descriptions using the dense layer of network.

3.3.1 VGG16 Feature Extraction:

The first part of the proposed work comprises of a VGG16 model of deep CNN approach. This model has been previously trained for extracting detailed visual information from the input images. We can get high-level representations that capture significant visual information by utilising the hierarchical framework and learning filters of VGG16. Figure 3-2 represents the basic architecture of VGG16-CNN model.

Initially, the VGG16 model of deep learning, is trained on ImageNet, and then it is been loaded. The model is restructured by removing the last classification layer, leaving the last layer as the output. This layer serves as the feature extractor for the images. Each image is pre-processed to meet input requirements of the VGG16 model. They are resized to $(224, 224)$ pixels and then converted into a numpy array. Then, using the image IDs as the key, the VGG16 model is implemented on the images for representing their features. These extracted features are then kept

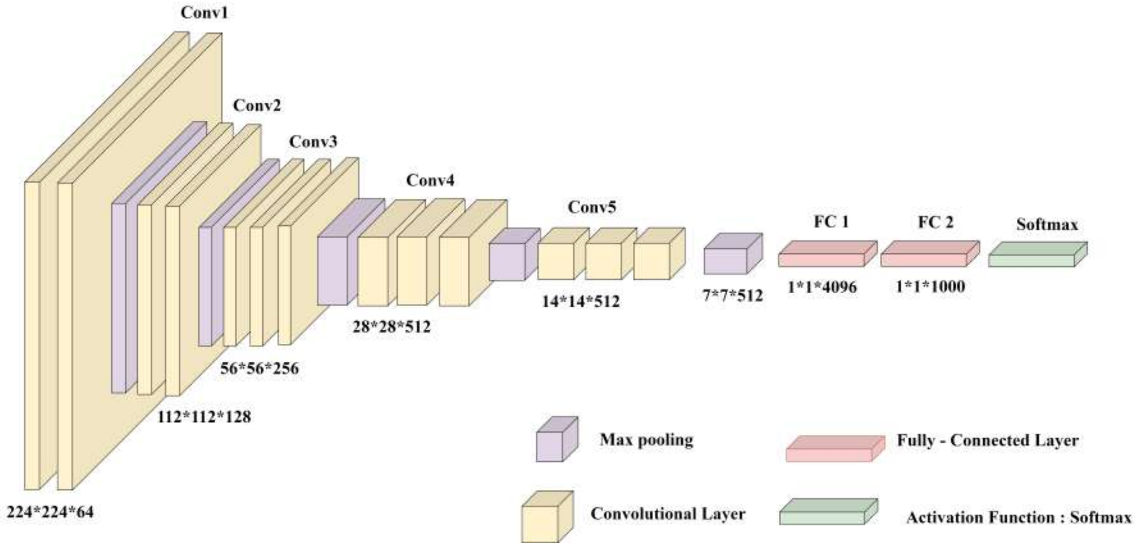


Figure 3-2: Basic architecture of VGG16 [67].

in a pickle file for further usage, preventing the need to recompute image features during consecutive executions. The second-to-last layer of the VGG16 model is the ultimate fully-connected layer, which is used for feature extraction. To decrease the spatial dimensions of a feature map while preserving the essential information, the layer of max-pooling is applied after the convolutional layers. The mathematical representation for max-pooling layer is denoted by Equation (3.1).

$$Y_i = \max(\text{Region of Interest}(X_i)) \tag{3.1}$$

where, X_i represents the output obtained from the previous layer or input data and Y_i reflects the result of the ith layer utilized in the max-pooling operation. The visual components of the image are extracted using the VGG16 model, which also serves as an encoder and are fed into the BiGRU model, which generates the descriptions, word per word.

3.3.2 Scene Graphs:

The scene graph extractor is the second component of the proposed framework. In scene graphs, objects are represented as the nodes and connections between them are represented as the edges. It employs a Faster R-CNN model which is already pre-trained for identifying the objects contained in images and extract scene graphs that include data on the labels and bounding boxes of the identified objects. A PyTorch tensor is then created from the images. This tensor is further used to detect objects, class labels and bounding boxes, by running it through the Faster R-CNN model. By tracing bounding boxes and adding labels to the related images, we can visually interpret the scene graphs. Figure 3-3 illustrates the structure of scene graph generation from an input image.

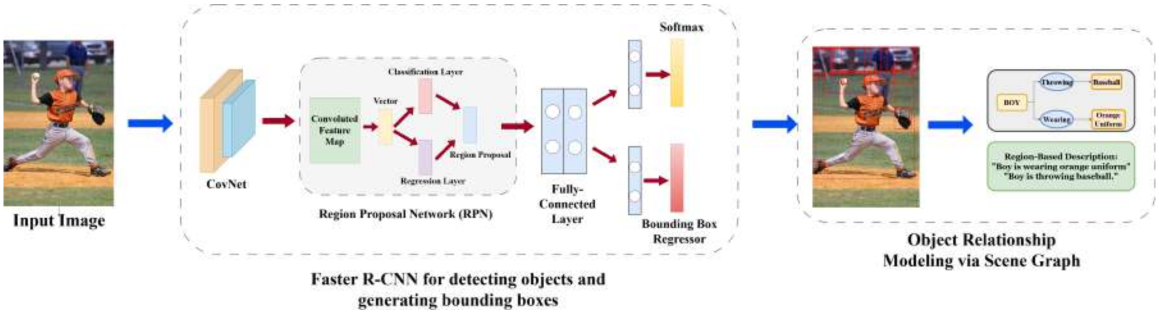

Input Image
CovNet
Convoluted Feature Map
Vector
Classification Layer
Region Proposal
Regression Layer
Region Proposal Network (RPN)
Fully-Connected Layer
Softmax
Bounding Box Regressor
Faster R-CNN for detecting objects and generating bounding boxes
BOY
Throwing
Baseball
Wearing
Orange Uniform
Region-Based Description:
"Boy is wearing orange uniform"
"Boy is throwing baseball."
Object Relationship Modeling via Scene Graph

Figure 3-3: Basic architecture of scene graph generation model: The model consists of the pre-trained Faster R-CNN which is used for identifying the objects contained in images and to generate scene graphs that include data on the labels and bounding boxes of the identified objects.

From the architecture, it can be observed that, the scene graph represents the relationships between different elements in the scene. First, faster R-CNN is used to detect objects. Then, it uses a relationship prediction process, which helps to determine the visual relationships between the objects that were discovered. By considering visual characteristics and spatial arrangements, this method enables us to deduce the relationships between objects within the scene. Here, the vertices (V) for the image are: "Boy", "Baseball" and "Orange Uniform", whereas the attributes (A) are "Throwing" and "Wearing". The final representation between these vertices and attributes can be presented in the form of descriptions. The approach obtains the use of specific semantic data by adding scene graphs, enabling more precise and context-aware production of the descriptions. Equation (3.2) describes the general structure of a scene graph:

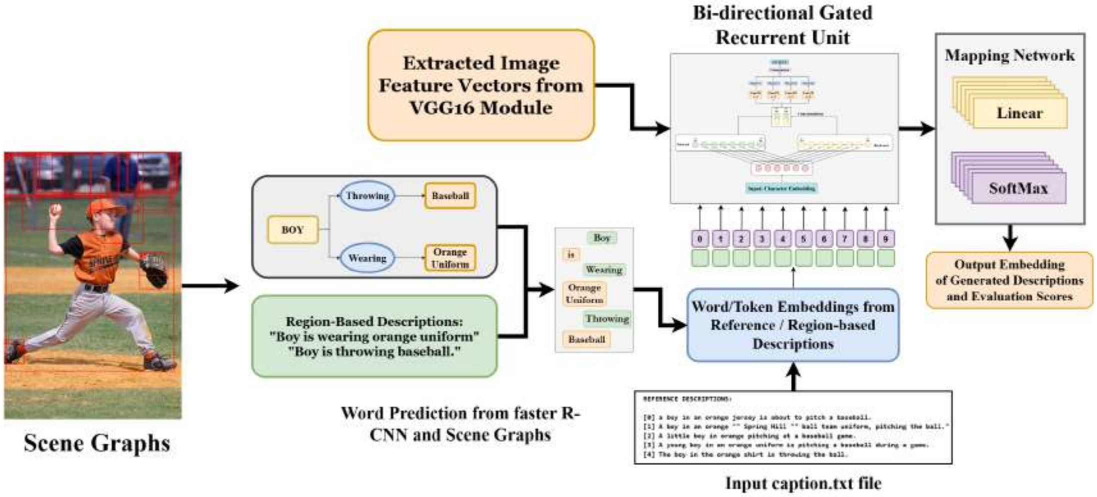
$$G = (V, E, A) \tag{3.2}$$

where, $V = v_1, v_2, ..., v_n$ is the collection of nodes that represent the scene's objects, $E = (v_i, v_j) \mid v_i, v_j \in V$ is the collection of edges and $A = a_1, a_2, ..., a_n$ is the group of attributes connected to each node $v_i \in V$.

In order to improve upon this even further, we use a BiGRU network that examines the identified objects and the interactions between them in a sequential manner, gathering contextual data and strengthening the scene graph. This approach enables us to create a more detailed and context-aware representation of the scene.

3.3.3 BiGRU for Temporal Context:

After the extraction of the objects and features from the scene graphs as well as from the feature vector of VGG16, they are passed through a Bidirectional Gated Recurrent Unit (BiGRU). It is introduced as the third and the final component to capture temporal dependencies in the generated descriptions of the proposed framework. The model successfully learns the context-relevant details necessary for producing logical and meaningful descriptions by analysing the sequence of it in both forward direction as well as in the backward directions. BiGRU may consider the contextual data from both sides of the input sequence due to its bidirectional processing. Figure 3-4 illustrates the architecture of BiGRU module used for description generation.


Bi-directional Gated Recurrent Unit
Extracted Image Feature Vectors from VGG16 Module
Mapping Network
Linear
SoftMax
Output Embedding of Generated Descriptions and Evaluation Scores
BOY
Throwing
Baseball
Wearing
Orange Uniform
Boy
is
Wearing
Orange Uniform
Throwing
Baseball
0
1
2
3
4
5
6
7
8
9
Region-Based Descriptions: "Boy is wearing orange uniform" "Boy is throwing baseball."
Word/Token Embeddings from Reference / Region-based Descriptions
Scene Graphs
Word Prediction from faster R-CNN and Scene Graphs
Input caption.txt file

```
REFERENCE DESCRIPTIONS:

[0] a boy in an orange jersey is about to pitch a baseball.
[1] A boy in an orange "" Spring Hill "" ball team uniform, pitching the ball."
[2] A little boy in orange pitching at a baseball game.
[3] A young boy in an orange uniform is pitching a baseball during a game.
[4] The boy in the orange shirt is throwing the ball.
```

Figure 3-4: Architecture of BiGRU Module for Description Generation.

As observed, the BiGRU module computes the words predicted from the scene graphs and the words/token embeddings obtained from the reference file. Also, the visual features are fed into BiGRU from the VGG16 module. Therefore, BiGRU is considered to have two Gated Recurrent Unit (GRU) layers and processes the input sequences in two different directions: one layer is used for processing the sequences of features from an image, and the second layer processes the sequences obtained from the words/token embeddings.

The working principle of the BiGRU module is explained as follows:

Forward Pass: During this pass of the BiGRU model, the input sequence is processed forwardly from beginning to end. Equation (3.3) represents the working of the forward pass.

$$h_{ft} = (1 - z_{ft)} \odot h_{f(t-1)} + z_{ft} \odot \tilde{h}_{ft} \tag{3.3}$$

where, h_{ft} denotes the forward hidden state at the time step t which is a vector used to represent the model's internal memory. $\odot$ denotes the multiplication which is done element-wise and $h_{f(t-1)}$ represents the previous hidden state. z_{ft} is the update gate which is used for finding how much information the candidate activation produces and it can be calculated by Equation (3.4):

$$z_{ft} = \sigma(W_{fz} \cdot [h_{f(t-1)}], x_t]) \tag{3.4}$$

where, W_{fz} is the weight matrix applied to the concatenated input for update gate, the Sigmoid activation function is abbreviated as σ and x_t stands for the input frame at time step t. $\tilde{h}_{ft}$ denotes the candidate activation which is a dynamic value that is used to calculate the new hidden state. It can further be evaluated by Equation (3.5):

$$\tilde{h}_{ft} = \tanh(W_{fh} \cdot [r_{ft} \odot h_{f(t-1)}, x_t]) \tag{3.5}$$

where, W_{fh} represents weight matrices for activation function in forward pass and the non-linearity of the activation is provided by *tanh* function, which has the values between -1 and 1. r_{ft} is used for determining what information from the previous hidden state $h_{f(t-1)}$ must be discarded or forgotten and it can further be calculated by Equation (3.6):

$$r_{ft} = \sigma(W_{fr} \cdot [h_{f(t-1)}, x_t]) \tag{3.6}$$

where, W_{fr} is the weight matrices for the reset gate.

Backward Pass: The input sequence is processed in the reverse order, that is, from the end to the beginning, during the BiGRU model's backward pass. Following Equation (3.7) represents the working of the backward pass.

$$h_{bt} = (1 - z_{bt}) \odot h_{b(t+1)} + z_{bt} \odot \tilde{h}_{bt} \tag{3.7}$$

where, h_{bt} represents the hidden state of backward direction at time step t, $\odot$ denotes the multiplication which is done element-wise and $h_{b(t+1)}$ represents the next hidden state. z_{bt} is the update gate which finds how much information the candidate activation yields and it can be calculated by Equation (3.8):

$$z_{bt} = \sigma(W_{bz} \cdot [h_{b(t+1)}, x_t]) \tag{3.8}$$

where, W_{bz} is the weight matrix applied to the concatenated input for update gate, σ is an abbreviation for the Sigmoid activation function and x_t represents the input frame at time step t. $\tilde{h}_{bt}$ denotes the candidate activation which is a dynamic value that is used to calculate the new hidden state and it can further be evaluated by Equation (3.9):

$$\tilde{h}_{bt} = \tanh(W_{bh} \cdot [r_{bt} \odot h_{b(t+1)}, x_t]) \tag{3.9}$$

where, W_{bh} represents weight matrices for activation function in backward pass and the non-linearity of the activation is provided by $tanh$ function, which has the values between -1 and 1. r_{bt} determines which information from the next hidden state $h_{b(t+1)}$ should be discarded or forgotten and it can further be calculated by Equation (3.10):

$$r_{bt} = \sigma(W_{br} \cdot [h_{b(t+1)}, x_t]) \tag{3.10}$$

where, W_{br} is the weight matrices for the reset gate.

At each time step, the final hidden state is obtained by concatenating the forward and backward hidden states: $h_t = [h_{ft}, h_{bt}]$. This enables the model to incorporate data from both the present and the future, which is advantageous for the procedure of creating image descriptions.

3.3.4 Description Generation Process:

After the layers of VGG16, scene graph and BiGRU comes the next phases in the proposed model of generating image descriptions, i.e., to predict the descriptions for given input images. The reference file for descriptions is an essential part of a description generation system for analysing the performance of the proposed model. The ground-truth descriptions for each input image in the dataset are contained in this file. A fundamental method utilised in NLP tasks, such as image description generation, is word embedding. It captures the semantic connections present between the words by characterizing them as dense, continuous-valued vectors in a space of high-dimension. This deep layer of network is used in the model to transform the input sequence of word indices into dense word embedding which are transmitted into the description generation model of BiGRU, allowing it to continuously and contextually comprehend the meaning of words. The model creates predicted descriptions during training and their quality is evaluated by comparing them with the descriptions given in the reference file. Evaluation metrics of BLEU 1-4, CIDEr, METEOR and ROUGE-L are employed to determine how closely the predicted descriptions match the actual descriptions. These metrics make it possible to measure how effectively the model incorporates the semantics and context of the images into the descriptions it generates.

The Algorithm 3.1 for the proposed model with its score evaluation is described below:

Algorithm 3.1 Image Description Generation

```
procedure GenerateDescription($I, SG, D_{ref}$)
    Input: Image $I$, Scene Graphs $SG$, Reference Description $D_{ref}$
    Output: Description $D$
    Preprocess the input image $I$ and extract visual features with the help of VGG16 model which is pre-trained.
    Extract object features and their relationship using the scene graphs $SG$ utilising a faster R-CNN model.
    Set the BiGRU model's hidden states to their initial values to provide image descriptions.
    Set the initial input of BiGRU as a start token which is obtained from the reference descriptions $D_{ref}$ provided in the dataset.
    Perform Word Embedding which is used for representing the words in the form of vectors that are dense in a continuous semantic space.
    $D \leftarrow$ empty description
    while not maximum length or end token reached do
        Pass visual features, object features, relationship features obtained from VGG16 and $SG$ along with the current input to the BiGRU model and update the hidden states present inside it using the current input.
        Generate the next token using the BiGRU and dense layer of network.
        Append the generated token to the description $D$.
        Set the current input to the generated token.
    end while
    Process the generated description $D$ by removing special tokens and converting tokens to words.
    Calculate evaluation metrics of: BLEU 1-4, CIDEr, METEOR and ROUGE-L for the generated description $D$ and reference description $D_{ref}$.
    Return the final generated description $D$ and the calculated scores.
end procedure
```

3.4 Experimental Analysis

For the evaluation of the proposed model, it requires a 64-bit version of Windows 11, an Intel Core i7 processor, 16 GB of RAM and an NVIDIA TITAN RTX graphics card with 24 GB of RAM. The suggested framework has been implemented using Keras and TensorFlow 2.12. The proposed model is evaluated on three benchmark datasets such as MS COCO, Flickr8k and Flickr30k.

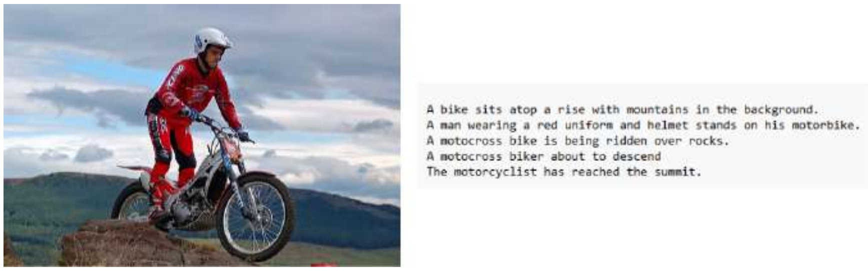
Microsoft Common Objects in Context (MS COCO) dataset[1]: This dataset is a well-known benchmark for image description generation tasks in the

---

[1] https://github.com/cocodataset/cocoapi

domains of computer vision and NLP. It dataset involves 82,783 JPEG images and has about 5 human-generated descriptions per image.

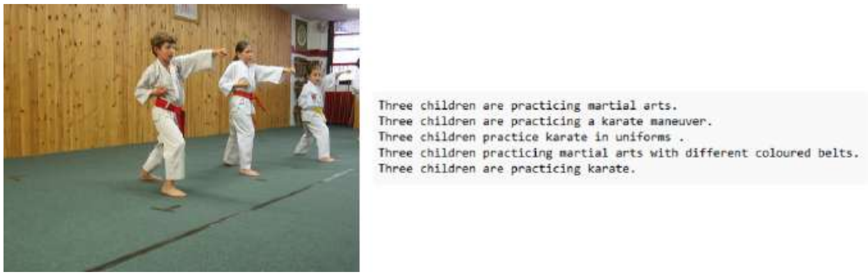
Flickr8k Dataset[2]: The Flickr8k dataset consists of 8092 JPEG images in total, which come in various sizes and shapes. The remaining 1000 photos are for development, with the remaining 6000 being used for training and testing. Figure 3-5 illustrates an image with five different reference descriptions.



```
A bike sits atop a rise with mountains in the background.
A man wearing a red uniform and helmet stands on his motorbike.
A motocross bike is being ridden over rocks.
A motocross biker about to descend
The motorcyclist has reached the summit.
```

Figure 3-5: A sample image from Flickr8k dataset

Flickr30k Dataset[3]: The Flickr30k dataset includes 5 human-annotated reference descriptions along with 31,783 images that are obtained from Flickr. Figure 3-6 is an example of the dataset.



```
Three children are practicing martial arts.
Three children are practicing a karate maneuver.
Three children practice karate in uniforms .
Three children practicing martial arts with different coloured belts.
Three children are practicing karate.
```

Figure 3-6: A sample image from the Flickr30k dataset

The availability of multiple descriptions per image in the datasets allows for the capture of inherent diversity and subjectivity in describing images. This enables the evaluation of models in generating diverse and contextually relevant descriptions. These datasets serve as valuable resources for training and evaluating our hybrid model, allowing us to leverage their large-scale image-description pairs to learn robust

---

[2] https://www.kaggle.com/adityajn105/flickr8k

[3] https://www.kaggle.com/hsankesara/flickr-image-dataset

visual representations and help improve the accuracy and quality of the predicted descriptions.

3.4.1 Implementation Details:

The model of VGG16 has been used to load and pre-process the images in order to extract visual features. Descriptions are pre-processed by converting them into lowercase, removing digits and special characters. Tokenization has been performed using the tokenizer object. Scene graphs are produced to further discover the objects and their interactions present in the images. In the context of scene graphs, a JavaScript Object Notation (.json) file is commonly used as a data format to store and represent the structure and attributes of a scene graph. It is a communication format that is easy to read and write for people as well as easy for programmers to understand and generate.

The proposed architecture consists of a bidirectional GRU layer to capture contextual information, with image features and word embedding concatenated to combine visual and textual information. The layers of the model are fully connected, pooling and convolutional layers. The input images have a shape of (batch_size, 224, 224, 3) and the output shape varies for each layer. Evaluation is conducted on a separate test set, and predictions have been made by feeding images through the trained model. Different evaluation scores are obtained to assess the standard of the produced descriptions. Table 3.1 provides the details of hyper-parameters used for the model that is suggested in the proposed work.

Table 3.1: Hyper-Parameters used for the Experimental Analysis

Batch Size	32
Number of Epochs	75
Dropout	0.5
Optimizer	Adam
Loss Function	Categorical Cross-Entropy
Datasets Used	MS COCO, Flick8k and Flickr30k
Evaluation Metrics	BLEU, METEOR, CIDEr and Rouge-L

3.4.2 Performance Measures and Evaluation

After the generation of descriptions for the given image dataset, the main parameters used in the image description generator are the evaluation metrics: BLEU 1-4, CIDEr, METEOR, and ROUGE-L. The following scores are primarily utilized to evaluate the quality of the descriptions generated by the model. They are also used to predict the highest correlations between the descriptions and to analyse their accuracy.

Bilingual Evaluation Understudy Score (BLEU): By evaluating n-gram overlap, the BLEU score calculates how similar the created descriptions are to the reference phrases. It can be calculated using the following Equation (3.11):

$$BLEUscore = BP * exp(1/N * sum(log(P_n))) \tag{3.11}$$

where, N is the total number of accurately matched values of n-grams between the reference as well as the generated descriptions is divided by the total number of n-grams present in the generated descriptions which are used for determining P_n. The Brevity Penalty (BP) term adjusts the score by contrasting the length of the predicted candidate descriptions with the typical length of the given reference descriptions. Following Equation (3.12) can be used to compute it:

$$BP = \begin{cases} \exp\left(1 - \frac{r}{c}\right) & \text{if } c > r \\ 1 & \text{if } c \leq r \end{cases} \tag{3.12}$$

where, r is the length of the given reference descriptions, and c is the length of candidate description. This score is obtained in the range of 0 to 1. It gives us a way to determine how well a reference description matches the set of generated descriptions from a particular model.

Consensus-based Image Description Evaluation (CIDEr): The CIDEr automatic evaluation metric is created especially for activities involving the creation of image descriptions. The n-gram weights employed by CIDEr are inversely connected to their frequency in a corpus between the predicted description and reference descriptions and they are utilised to determine a weighted total of similarity scores.

The following Equation (3.13) is used to determine the CIDEr score:

$$CIDEr = \sum_{i} \frac{w_i \times s_i}{N} \tag{3.13}$$

where i stands for each n-gram, w_i is its given weight, s_i is its similarity score and N is the total overall number of n-grams taken into account.

Metric for Evaluation of Translation with Explicit ORdering (METEOR): By matching up the generated description and the reference description at the word level, METEOR takes into account both precision and recall. In the case of precision and recall, the harmonic mean is modified, and the final score is computed using a penalty term for unaligned words. The expected and reference descriptions must exactly match the METEOR score, which ranges from the value of 0 to 1.

The METEOR score value is calculated by following Equation (3.14):

$$METEOR = (1 - \alpha) \times P + \alpha \times R \tag{3.14}$$

where the factor α balances the contribution of recall and precision. Following Equation (3.15), is used for calculating the accuracy (P) and recall (R):

$$Precision(P) = \frac{\text{matching words}}{\text{predicted words}}, Recall(R) = \frac{\text{matching words}}{\text{reference words}} \tag{3.15}$$

Recall-Oriented Understudy for Gisting Evaluation (ROUGE): he longest common sub-sequence, skip-bigram statistics, and n-gram co-occurrence statistics are just a few of the approaches used for measuring the overlap occurring between the predicted phrases and the descriptions of reference in the Rouge-L score. This score ranges from the value of 0 to 1, with 1 denoting an exact match between the reference and predicted phrases for the relevant n-gram or sequence.

Between the predicted description and the reference descriptions, the Rouge-L score is mainly used for calculating the Longest Common Sub-sequence (LCS) in any description and is denoted by Equation (3.16):

$$Rouge - L = \frac{\text{LCS}}{\text{reference words}} \tag{3.16}$$

These performance measures provide different perspectives on the quality of generated descriptions, including linguistic similarity, consensus with reference descriptions, and content overlap. They are frequently used to assess and compare models for creating visual descriptions.

3.4.3 Ablation Study:

In order to assess the respective contributions of scene graphs, the BiGRU layer, and the VGG16 visual features to the hybrid model for the generation of image descriptions, we conducted an ablation study. This study set out to systematically assess effect of each component by evaluating how it affects the performance measures. A thorough study of the outcomes from the MS COCO dataset within the framework of the work proposed is shown in Table 3.2. It presents various evaluation metrics for different models applied to the input image. The table presents an ablation study evaluating different model configurations on the MS COCO dataset for image description generation. Performance is measured using standard metrics including BLEU, CIDEr, METEOR, and ROUGE-L. The proposed model, VGG16 + Scene Graph + BiGRU, outperforms all others, achieving the highest scores across BLEU-1 to BLEU-4 (0.816, 0.785, 0.727, 0.561), CIDEr (1.281), METEOR (0.293), and ROUGE-L (0.599), indicating its strong alignment with human-annotated captions. Intermediate configurations such as VGG16 + BiGRU without scene graph and

Table 3.2: Ablation Study: Performance comparison of different components on MS COCO dataset

Model Configuration	**B-1**	**B-2**	**B-3**	**B-4**	**C**	**M**	**R-L**
VGG16 + GRU without Scene Graphs	0.598	0.471	0.426	0.291	0.834	0.171	0.337
VGG16 + BiGRU without Scene Graphs	0.714	0.522	0.489	0.316	0.947	0.213	0.415
VGG16 + Scene Graph + GRU	0.765	0.692	0.653	0.507	1.106	0.254	0.532
VGG16 + Scene Graph + BiGRU (Proposed)	**0.816**	**0.785**	**0.727**	**0.561**	**1.281**	**0.293**	**0.599**

VGG16 + Scene Graph + GRU also demonstrate improved performance, highlighting the individual contributions of scene understanding and bidirectional sequence modelling. The baseline VGG16 + GRU without scene graph model records the lowest scores, emphasizing the benefit of integrating scene graphs and a BiGRU decoder for generating more contextually rich and semantically accurate image descriptions. The graphical representation of the comparison results from the MS COCO dataset is shown in Figure 3-7.

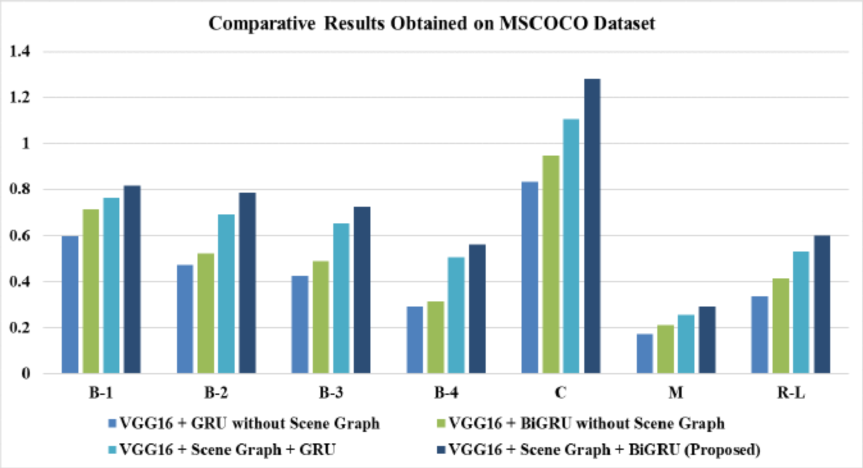


Figure 3-7: Comparative graphical depiction of the results obtained for various models evaluated on MS-COCO dataset

Table 3.3 displays a comprehensive analysis of the results obtained from the Flickr8k dataset in the context of the proposed work. It presents various evaluation metrics for different models applied to the input image. The table summarizes the performance of various models for image description generation on the Flickr8k dataset. VGG16 + Scene Graphs + BiGRU achieved the highest BLEU-1 to BLEU-4 scores (0.683, 0.587, 0.483, 0.397), demonstrating its superior ability to produce descriptions closely aligned with human references. While VGG16 + BiGRU without

Table 3.3: Ablation Study: Performance comparison of different components on Flickr8k dataset

Model Configuration	**B-1**	**B-2**	**B-3**	**B-4**	**C**	**M**	**R-L**
VGG16 + GRU without Scene Graph	0.554	0.322	0.306	0.278	0.948	0.203	0.398
VGG16 + BiGRU without Scene Graph	0.554	0.322	0.306	0.278	1.005	0.237	0.435
VGG16 + Scene Graph + GRU	0.635	0.487	0.439	0.356	1.084	0.268	0.507
VGG16 + Scene Graph + BiGRU (Proposed)	**0.683**	**0.587**	**0.483**	**0.397**	**1.118**	**0.298**	**0.545**

scene graph also performed well, it did not surpass the proposed model, achieving competitive scores, particularly in BLEU-1 (0.554) and BLEU-2 (0.322). The incorporation of scene graphs and BiGRU layers in the proposed model contributed significantly to its performance. In contrast, models such as VGG16 + Attention, VGG16 + LSTM, and ResNet50 + Bi-LSTM exhibited weaker performance across most metrics. Overall, the proposed model excels in generating accurate, relevant, and well-structured descriptions on the Flickr8k dataset. Figure 3-8 represents the graphical demonstration of the comparative results obtained on the Flickr8k dataset.

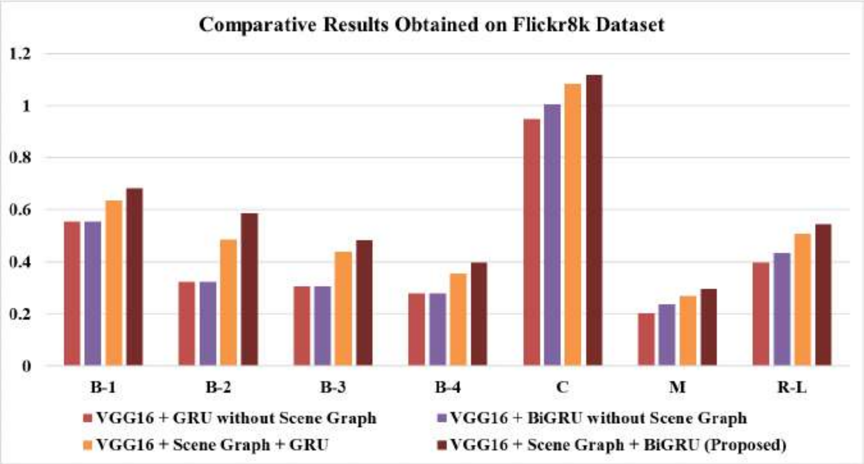


Figure 3-8: Comparative graphical depiction of the results obtained for various models evaluated on Flickr8k dataset

The analysis of the results from the Flickr30k dataset using various metrics is shown in Table 3.4. The table presents the results of an ablation study on the Flickr30k dataset for image description generation. The proposed model, VGG16 + Scene Graph + BiGRU, outperformed all other configurations, achieving the highest BLEU-1 to BLEU-4 scores (0.675, 0.546, 0.465, 0.339), CIDEr (1.006), METEOR

Table 3.4: Ablation Study: Performance comparison of different components on Flickr30k dataset

Model Configuration	**B-1**	**B-2**	**B-3**	**B-4**	**C**	**M**	**R-L**
VGG16 + GRU without Scene Graph	0.572	0.338	0.316	0.214	0.894	0.253	0.245
VGG16 + BiGRU without Scene Graph	0.572	0.338	0.316	0.214	0.894	0.253	0.245
VGG16 + Scene Graph + GRU	0.640	0.450	0.408	0.293	0.957	0.264	0.300
VGG16 + Scene Graph + BiGRU (Proposed)	**0.675**	**0.546**	**0.465**	**0.339**	**1.006**	**0.287**	**0.345**

(0.287), and Rouge-L (0.345). This highlights the effectiveness of integrating scene graphs with BiGRU layers in generating high-quality, contextually accurate descriptions. In contrast, VGG16 + GRU and VGG16 + BiGRU, both without scene graphs, showed identical performance with lower scores across all metrics. VGG16 + Scene Graph + GRU also showed improvement over these models, with better performance in BLEU-1, BLEU-2, BLEU-3, and CIDEr, though still falling short of the proposed model. Overall, the inclusion of scene graphs alongside BiGRU significantly enhances performance, as demonstrated by the results of the proposed approach. Figure 3-9 represents the graphical demonstration of the comparative results obtained on Flickr30k dataset.

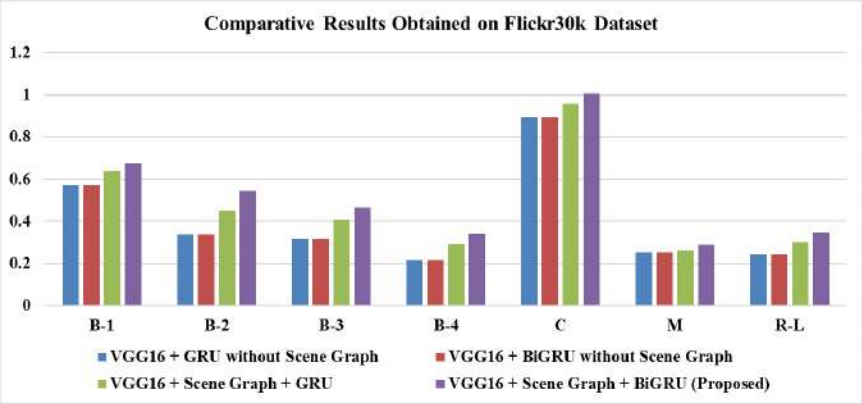
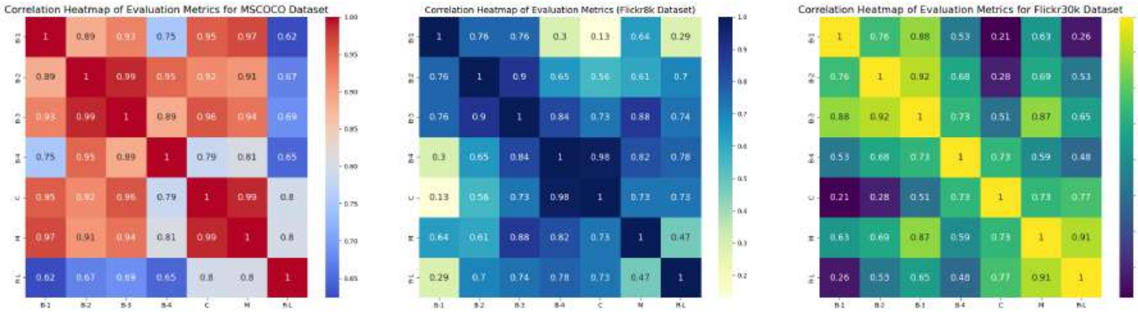


Figure 3-9: Comparative graphical depiction of the results obtained for various models evaluated on Flickr30k dataset

The following Figure 3-10 depicts the Correlation Heatmap of Evaluation Metrics for MSCOCO, Flickr8k and Flickr30 dataset in 3-10(a), 3-10(b) and 3-10(c), respectively. Across the MS COCO, Flickr30k, and Flickr8k datasets, the correlation heatmaps for the evaluation metrics provide valuable insight on the connections between different performance metrics that are utilised to assess the models. The heatmaps show how different metrics are related to each other, including ROUGE-L (R-L), METEOR (M), CIDEr (C), and BLEU scores (B-1 to B-4). Strong positive correlations between the BLEU scores for all three datasets show that models that do well on one BLEU metric also typically perform well on the others. This is to be expected because BLEU-1 to BLEU-4 measure accumulated n-gram precision. CIDEr



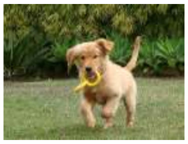
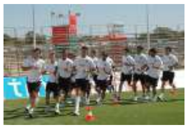
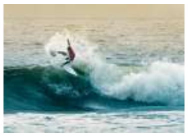
(a) (b) (c)

Figure 3-10: Representation of the Correlation Heatmap of Evaluation Metrics for MSCOCO, Flickr8k and Flickr30 dataset.

has a moderate to strong connection across datasets, with higher-order BLEU scores (B-3, B-4), indicating that models with more thorough context capture are likely to score higher on CIDEr. Also, the CIDEr and BLEU metrics show moderately good correlations with METEOR and ROUGE-L, showing their complementary roles in assessing model performance. Overall, the heatmaps show how the model evaluation is consistent across different metrics and datasets, validating the robustness and ability to generalize the performance of the suggested model over a range of evaluation criteria.

The qualitative word-by-word comparison of predictions of several models for the Flickr8k dataset is presented in Table 3.5. The significance of the suggested VGG16 + Scene Graphs + BiGRU model is demonstrated by a word-by-word comparison of the predicted descriptions generated by different models on the Flickr8k dataset. While VGG16 + Attention and VGG16 + LSTM can both predict a generic description such as "Dog in grass" and "Dog running in grass," respectively, for the first image, they are not very detailed when it comes to important details. A more accurate prediction, "Puppy running across grass," is produced by VGG16 + BiGRU and is more in line with the actual description. But with "Puppy running across the grass with yellow toy," the suggested model accurately and thoroughly captures all the important elements. Likewise, with regard to the second image, the models VGG16 + Attention and VGG16 + LSTM predict, respectively, "Men in White Shirt" and "Group of Men in White Shirt," despite the absence of any contextual information

Table 3.5: Quantitative Results Obtained on Flickr8k Dataset

Test Image	Ground Truth Descriptions	Model: Predicted Description
	1. A cute puppy fetches a yellow ring chew toy in the yard. 2. A puppy is running across the grass with a yellow toy in its mouth. 3. A tan dog is running through the grass with a yellow toy in its mouth. 4. A tan dog with a red collar runs with a yellow toy in its mouth. 5. Fluffy golden puppy holding yellow rings in mouth while running through grass.	**1. VGG16 + Attention:** Dog in grass. **2. VGG16 + LSTM:** Dog running in grass. **3. VGG16 + BiGRU:** Puppy running across grass. **4. VGG16 + Scene Graphs + BiGRU (Proposed):** Puppy running across the grass with yellow toy.
	1. A group of men in white shirts and dark shorts are running on an athletic field. 2. A group of people in matching uniforms jogging around a track. 3. A group of soccer players run a lap. 4. A team of men jog around orange cones. 5. A team of soccer players in white strips are running around cones on a sports field.	**1. VGG16 + Attention:** Men in white shirt. **2. VGG16 + LSTM:** Group of men in white shirt. **3. VGG16 + BiGRU:** Group of men in white running. **4. VGG16 + Scene Graphs + BiGRU (Proposed):** Group of players in white shirt running on field.
	1. A man surfing in the ocean. 2. A surfer catches a wave and tries to hold on as the surf collapses around him. 3. A surfer is riding a surfboard on top of a breaking wave. 4. A surfer is riding a wave in a large body of water. 5. A surfer on a wave.	**1. VGG16 + Attention:** Man in ocean. **2. VGG16 + LSTM:** Surfer riding ocean. **3. VGG16 + BiGRU:** Man riding surfboard. **4. VGG16 + Scene Graphs + BiGRU (Proposed):** Surfer riding surfboard in wave of water.

regarding the activity and location. The "Group of men in white running" from VGG16 + BiGRU is still lacking in context, but it is more pertinent. On the other hand, the "Group of players in white shirt running on pitch" in the suggested model accurately depicts the surroundings, the individuals, and their actions. In the third image, VGG16 + Attention yields the ambiguous prediction "Man in ocean," but VGG16 + LSTM and VGG16 + BiGRU provide some specificity but exclude crucial meanings like "wave." With a thorough comprehension of the image content, the predicted image, "Surfer riding surfboard in wave of water," accurately combines all necessary components.

Through efficient object relationships, activity detection, and scene context capture, the suggested VGG16 + Scene Graphs + BiGRU model outperforms the baseline models, consistently yielding more detailed and contextually accurate descriptions.

3.4.4 Result and Analysis:

We use MS COCO, Flickr8k and Flickr30k as the benchmark datasets for the proposed architecture. The following Table 3.6 represents the results obtained on the sample input images, their scene graphs, obtained descriptions along with reference description and finally their predicted scores of BLEU 1-4 (B-1 to B-4), CIDEr (C), METEOR (M) and ROUGE-L (R-L), for MS COCO dataset.

Table 3.7 denotes the analysis of results obtained for Flickr8k dataset. It shows the predicted descriptions and the obtained metrics of evaluation.

Table 3.8 depicts the outputs obtained for the image samples from Flickr30k dataset. It shows the predicted descriptions and the obtained scores of evaluation.

The final outcomes from a suggested framework for producing image descriptions for the three datasets are shown in Table 3.9. As it is already noted, several metrics are employed to evaluate the model's effectiveness, that includes BLEU 1-4 (B-1 to B-4), CIDEr (C), METEOR (M) and ROUGE-L (R-L). The following average overall results are obtained using the MS COCO dataset: BLEU 1-4: 0.816, 0.785,0.727 and 0.561; CIDER: 1.281; METEOR: 0.293 and ROUGE-L: 0.599. On Flickr8k dataset, the average overall scores are obtained to be: CIDER: 1.118; METEOR:

Table 3.6: Results Obtained on MS COCO Dataset

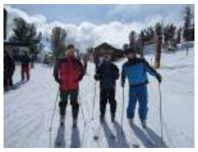
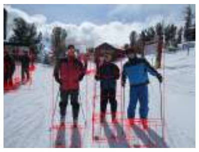
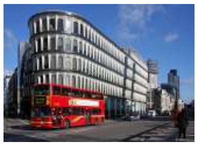
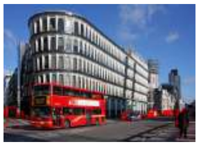
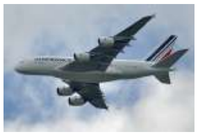
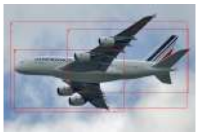
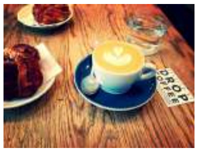
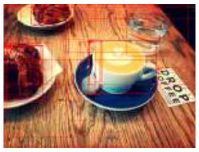
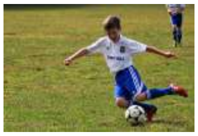
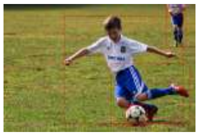
Input Image	**Obtained Scene Graph**	**Descriptions**	**Scores Obtained**
		**Reference:** A group of three people skiing in a snow-covered place. **Predicted:** A group of people skiing.	**B-1:** 0.789 **B-2:** 0.652 **B-3:** 0.623 **B-4:** 0.435 **C:** 0.786 **M:** 0.187 **R-L:** 0.532
		**Reference:** A red double-decker bus driving down a street. **Predicted:** A red bus on the street.	**B-1:** 0.563 **B-2:** 0.543 **B-3:** 0.452 **B-4:** 0.394 **C:** 0.678 **M:** 0.127 **R-L:** 0.389
		**Reference:** A large passenger jet flying through the sky. **Predicted:** A jet flying in the sky.	**B-1:** 0.798 **B-2:** 0.697 **B-3:** 0.602 **B-4:** 0.523 **C:** 0.923 **M:** 0.211 **R-L:** 0.378
		**Reference:** Two plates of bread, a cup of coffee, and a glass of water is kept on the brown table. **Predicted:** A cup of coffee on the table.	**B-1:** 0.559 **B-2:** 0.536 **B-3:** 0.443 **B-4:** 0.397 **C:** 0.875 **M:** 0.128 **R-L:** 0.352
		**Reference:** A boy in a white t-shirt and blue shorts is kicking a soccer ball on the ground. **Predicted:** A boy in a white t-shirt is kicking a soccer ball.	**B-1:** 0.801 **B-2:** 0.752 **B-3:** 0.694 **B-4:** 0.512 **C:** 1.006 **M:** 0.264 **R-L:** 0.547

Table 3.7: Results Obtained for Flickr8k Dataset

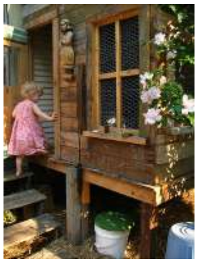
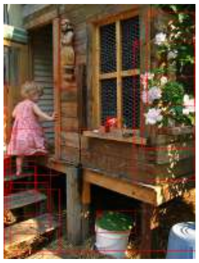
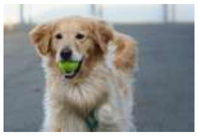
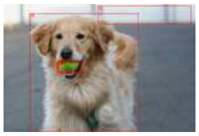
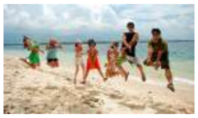
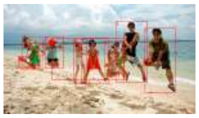
Input Image	Obtained Scene Graph	Descriptions	Scores Obtained
		**Reference:** 1. A child in a pink dress is climbing up a set of stairs in an entryway. 2. A girl going into a wooden building. 3. A little girl climbing into a wooden playhouse. 4. A little girl climbing the stairs to her playhouse. 5. A little girl in a pink dress going into a wooden cabin. **Predicted:** A little girl climbing wooden stairs.	**B-1:** 0.628 **B-2:** 0.534 **B-3:** 0.436 **B-4:** 0.388 **C:** 1.003 **M:** 0.287 **R-L:** 0.523
		**Reference:** 1. A brown and white dog holds a tennis ball in his mouth. 2. A dog has a tennis ball in its mouth. 3. A golden-colored dog, with his eyes alert, holds a brightly colored tennis ball in his mouth. 4. A tan dog is playing with the green ball. 5. Dog running with a tennis ball in its mouth. **Predicted:** A dog with a tennis ball.	**B-1:** 0.676 **B-2:** 0.567 **B-3:** 0.412 **B-4:** 0.348 **C:** 0.972 **M:** 0.241 **R-L:** 0.457
		**Reference:** 1. A group of people jump in the sand at the beach. 2. A group of teenagers is jumping in the air on the beach. 3. A group of young people jump up in the air while on the beach. 4. A group of young people posing in the air on a sandy beach. 5. Seven people are jumping in the air along the shore. **Predicted:** A group of people on the beach.	**B-1:** 0.568 **B-2:** 0.437 **B-3:** 0.264 **B-4:** 0.156 **C:** 0.785 **M:** 0.142 **R-L:** 0.295

Table 3.8: Results Obtained for Flickr30k Dataset

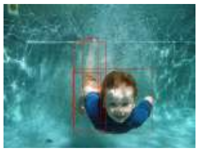
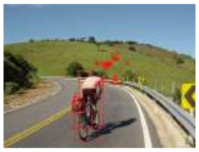
Input Image	**Obtained Scene Graph**	**Descriptions**	**Scores Obtained**
		**Reference:** 1. A child in a blue shirt and orange swim trunks is underwater. 2. A child smiling at the camera while swimming underwater. 3. A child swims underwater in a pool. 4. A red-haired girl in a blue T-shirt is swimming underwater in a pool. 5. Child dressed in blue is smiling underwater. **Predicted:** A child swims in a pool.	**B-1:** 0.582 **B-2:** 0.543 **B-3:** 0.432 **B-4:** 0.335 **C:** 0.997 **M:** 0.237 **R-L:** 0.312
		**Reference:** 1. A cyclist is riding a bicycle on a curved road up a hill. 2. A man in aerodynamic gear riding a bicycle down a road around a sharp curve. 3. A man on a mountain bike is pedaling up a hill. 4. Man bicycles up a road, while cows graze on a hill nearby. 5. The biker is riding around a curve in the road. **Predicted:** A man is riding a bicycle on a hill.	**B-1:** 0.342 **B-2:** 0.235 **B-3:** 0.154 **B-4:** 0.116 **C:** 0.775 **M:** 0.158 **R-L:** 0.165

Table 3.9: Results obtained from the proposed model

Dataset	**B-1**	**B-2**	**B-3**	**B-4**	**C**	**M**	**R-L**
MS COCO	0.816	0.785	0.727	0.561	1.281	0.293	0.599
Flickr8k	0.683	0.587	0.483	0.397	1.118	0.298	0.545
Flickr30k	0.675	0.546	0.465	0.339	1.006	0.287	0.345

0.298; ROUGE-L: 0.545; BLEU-1: 0.683; BLEU-2: 0.587; BLEU-3: 0.483; BLEU-4: 0.397. The recommended model also did well on the Flickr30k dataset, obtaining the following average overall scores: CIDER: 1.006, METEOR: 0.287, ROUGE-L: 0.345, BLEU-1: 0.675, BLEU-2: 0.546, BLEU-3: 0.465, BLEU-4: 0.339. The graphical depiction of the outcomes is shown in Figure 3-11.

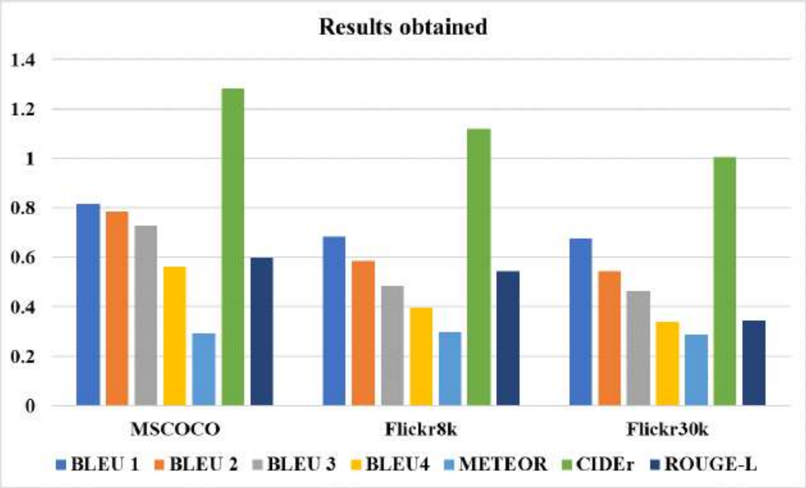


Figure 3-11: Graphical representation of the evaluation scores obtained from MSCOCO, Flickr8k and Flickr30k dataset for the proposed framework.

These scores evaluate the model's performance in generating descriptions that align well with the reference descriptions. Higher scores indicate a higher degree of similarity and quality in the generated descriptions.

3.4.5 Comparison With Other State-of-the-art Methods:

A comprehensive generalization study has been conducted to evaluate the proposed hybrid model's capacity to generate image descriptions on the three datasets: MSCOCO, Flickr8k and FLickr30k. The primary objective of the suggested study is to assess the model's performance throughout a range of visual domains and datasets, focusing on its ability to generate accurate and considerate descriptions outside of the training set. Table 3.10 presents a comparative analysis of various image description generation models on the MS COCO dataset. It includes evaluation metrics such as BLEU, METEOR, CIDEr and Rouge-L to measure the performance of different models.

Table 3.10: Comparative Analysis of Image Description Generation Models on MS COCO Dataset

Model	**BLEU**	**METEOR**	**CIDEr**	**Rouge-L**
Neural Image Caption (NIC) + Attention [68]	0.625	0.195	0.660	-
LSTM-A [25]	0.787	0.270	1.160	0.564
Graph Convolutional Networks(GCN) [52]	0.799	0.282	1.231	0.579
GCN + LSTM [69]	0.774	0.281	1.170	0.572
GCN + LSTM + Ruminant Decoder [70]	0.343	0.264	1.061	0.552
LSTM + BiGRU [11]				
1. Soft Attention	0.707	0.239	-	-
2. Hard Attention	0.718	0.203	-	-
Scene Graph Auto-Encoder (SGAE) [64]	0.808	0.284	1.278	0.586
Transformer Model [27]	0.810	0.291	1.274	0.592
Multi-level Cross-modal with scene graphs [65]	0.785	0.282	0.117	0.576
Scene graphs with Transformer [66]	0.802	0.291	1.216	0.599
VG-Cap [60]	0.792	0.290	1.255	0.591
VGG16 + Scene Graphs + BiGRU (Proposed)	**0.816**	**0.293**	**1.281**	**0.599**

Using BLEU, METEOR, CIDEr, and Rouge-L metrics, the table provides a comparative examination of several image description-generating methods on the MS COCO dataset. With the highest scores for all metrics—BLEU (0.816), METEOR (0.293), CIDEr (1.281), and Rouge-L (0.599)—the proposed model stands out as having the best ability to produce descriptions that closely match the reference texts in terms of both language diversity and content relevance. Several other models, such as Transformer Model [27] and Scene Graph Auto-Encoder (SGAE) [64], also perform well and rank among the best models. But other models, such as LSTM + BiGRU with Soft Attention and Hard Attention [11] and GCN + LSTM + Ruminant Decoder [70], perform comparably worse on several measures, suggesting room for improvement. Also, multi-level cross-modal with scene graph [65], scene graphs with transformer model [66] and VG-Cap [60] highlighted comparable scores. As a result, the proposed model is the most effective image description generation model on the MS COCO dataset, outperforming other models across all examined metrics. This examination offers insightful information about the performance and capabilities.

Table 3.11 presents a comparative analysis of various image description generation models on the Flickr8k dataset. It offers evaluation metrics for evaluating how well various models perform, including BLEU, METEOR, CIDEr and Rouge-L. Using BLEU, METEOR, CIDEr, and Rouge-L as assessment metrics, the table compares

Table 3.11: Comparative Analysis of Image Description Generation Models on Flickr8k Dataset

Model	**BLEU**	**METEOR**	**CIDEr**	**Rouge-L**
VGG16 + Attention [71]	0.630	-	-	-
Neural Image Caption (NIC) + Attention [68]	0.579	-	-	-
LSTM + BiGRU [11]				
1. Soft Attention	0.670	0.189	-	-
2. Hard Attention	0.670	0.203	-	-
g-LSTM [72]	0.635	0.203	-	-
Visual enhanced gLSTM [73]	0.650	0.205	0.546	0.518
Transformer based local graph semantic attention (TLGSA) [59]	0.659	-	0.471	0.465
VGG16 + Scene Graphs + BiGRU (Proposed)	**0.683**	**0.298**	**1.118**	**0.545**

several image description generation methods on the Flickr8k dataset. With the highest scores for BLEU (0.683), METEOR (0.298), CIDEr (1.118), and Rouge-L (0.545) among all metrics, the proposed model outperforms the others in producing descriptions that are both linguistically varied and pertinent to the content. Other models, such as Transformer based local graph semantic attention (TLGSA) [59] and Visual enhanced gLSTM [73], perform competitively, especially in BLEU and METEOR, but not well enough in CIDEr and Rouge-L scores. LSTM + BiGRU [11] is one model that uses both soft and hard attention processes. It performs moderately, suggesting that there is an opportunity for improvement in terms of producing detailed and precise visual descriptions. This analysis offers important information on the performance and capabilities of alternative models, assisting researchers in selecting the optimum model for image description generation tasks on the Flickr8k dataset.

Table 3.12 compares various image description generation models using the Flickr30k dataset. The analysis includes metrics for evaluating the models' performance, such as BLEU, METEOR, CIDEr, and Rouge-L.

Table 3.12: Comparative Analysis of Image Description Generation Models on Flickr30k Dataset

Model	**BLEU**	**METEOR**	**CIDEr**	**Rouge-L**
Neural Image Caption (NIC) + Attention [68]	0.573	-	-	-
LSTM + BiGRU [11]				
1. Soft Attention	0.667	0.185	-	-
2. Hard Attention	0.669	0.184	-	-
Adaptive Attention [74]	0.667	0.204	0.531	-
Topic Oriented Captioning [75]	0.646	0.192	0.396	0.322
Transformer based local graph semantic attention (TLGSA) [59]	0.643	-	0.450	0.289
VGG16 + Scene Graphs + BiGRU (Proposed)	**0.675**	**0.287**	**1.006**	**0.345**

Utilizing the Flickr30k dataset, the table presents a comparative examination of many image description generation models with the help of the BLEU, METEOR, CIDEr, and Rouge-L metrics. With the greatest scores in BLEU (0.675), METEOR (0.287), CIDEr (1.006), and Rouge-L (0.345), the proposed model stands out and demonstrates its strong capacity to produce meaningful and accurate descriptions that closely match the reference texts. While they do not outperform the proposed model in all metrics, other models such as Transformer-based local graph semantic attention (TLGSA) [59], also perform well. Meanwhile, models like Topic Orientated Captioning [75] and Adaptive Attention [74] perform competitively, especially in BLEU and METEOR, but fall short in CIDEr and Rouge-L scores for Flickr30k dataset. Consequently, the proposed model outperforms other state-of-the-art models in each evaluation metric, demonstrating that it is the best model for creating image descriptions for all three datasets.

3.5 Conclusion

In this chapter, we introduced a unified hybrid framework that combines scene graphs, BiGRU, and VGG16 to generate image descriptions. This approach makes extensive use of scene graphs to capture semantic links, BiGRU to produce coherent and contextually rich descriptions, and VGG16 for robust visual feature extraction. The algorithm we used performs exceptionally well in generating accurate and comprehensive descriptions for a variety of photos, according to the evaluation. Even with these encouraging outcomes, there are still a number of directions that could be explored. Incorporating attention techniques to improve the model's emphasis on particular image regions or objects could be beneficial for future studies. This would improve the quality of output descriptions by giving distinct visual components varied weights. Furthermore, using transformer-based models like BERT, GPT, or Transformer may strengthen the model's comprehension of context and its capacity to capture long-range dependencies.

Chapter 4

Deep Learning Framework for Improved Image Description Accuracy

Lakshita Agarwal and Bindu Verma. "Tri-FusionNet: Enhancing Image Description Generation with Transformer-based Fusion Network and Dual Attention Mechanism".
Archived at: http://arxiv.org/abs/2504.16761. Communicated in: IEEE Transactions on Human-Machine Systems, Impact Factor-3.5. ***(1st Major Revision Submitted)***

4.1 Introduction

While the previous approach showed significant improvements in identifying object relationships and producing logical descriptions of images, some drawbacks remained. The model's capacity to generalise across complex and varied images was limited by its dependence on specified scene graphs and successive processing. Furthermore, the depth of generated descriptions was limited by the constraints of conventional CNN-RNN architectures in capturing global context and long-range dependencies. This chapter introduces Tri-FusionNet, a transformer-based architecture that combines a

Vision Transformer (ViT) encoder with dual attention, a RoBERTa decoder, and a CLIP integration module to enhance image description creation. By using contrastive learning to improve vision-language alignment and self-attention processes to improve contextual comprehension, Tri-FusionNet produces more accurate, detailed, and semantically relevant captions. Each component's contribution is examined using an ablation research, which shows how dual attention, CLIP, and ViT affect the overall performance of the model. This chapter discusses Tri-FusionNet's architecture, training procedure, and evaluation on MS COCO, Flickr8k, and Flickr30k while showcasing the model's competitive performance versus state-of-the-art models using metrics including BLEU, CIDEr, METEOR, and ROUGE-L.

The major contribution of the proposed chapter are as follows:

- Development of Tri-FusionNet, a novel image description generation model that advances multi-modal approaches for producing precise and contextually rich descriptions.
- Integration of multiple components to enhance the model's fine-tuning capabilities, improving the overall efficacy of image description generation.
- Combination of a ViT encoder with dual attention, a RoBERTa decoder, and a CLIP integration module to enable efficient interaction between different modalities for more accurate and contextually appropriate descriptions.
- Effective handling of challenges posed by various benchmark datasets, demonstrating adaptability to dataset-specific characteristics and the potential to achieve new state-of-the-art results.

4.2 Literature Survey

Autonomous image description systems have been studied using deep learning frameworks. The domains of computer vision and human-computer interaction make substantial use of these frameworks. Because the transformer is efficient at gathering long-range dependencies and modelling sequential data, it shows potential for multi-modal tasks like creating visual descriptions. Cornia et al. [27] introduced M2, a Meshed Transformer with Memory, for image description generation, while Li et al. [28] created the EnTangled Attention (ETA) Transformer, demonstrating state-of-the-art performance. Additionally, Yu et al. [29] presented a Multi-modal Transformer (MT) model that allows for complex multi-modal reasoning and accurate description development by storing intra-modal and inter-modal interactions in a single attention block.

In order to achieve competitive performance without task-specific training, Radford et al. [76] showed that pre-training a model to predict image-caption pairs using a dataset of 400 million (image, text) pairs allows zero-shot transfer to diverse downstream tasks. Using distinct task IDs, MiniGPT-v2 is a unified model created to manage a variety of vision-language tasks, including as visual question responding, visual grounding, and image description, to enhance learning effectiveness and performance on several benchmarks [77]. SAMT-Generator, a multi-stage transformer feature augmentation network with second attention and Maxout decoding, was proposed by Yang et al. [78]. State-of-the-art results were obtained by integrating spatially aware pseudo-supervised and scale-wise reinforcement modules in a unique approach to the S^2 transformer. [79]. Using HAPE positional encoding, LSM, RNorm function, and LFE, Yang et al. [80] introduced CA-Captioner, a fully Transformer-based image captioning model that improves performance, particularly in terms of BLEU4 and CIDEr measures.

The double-attention architecture that Parvin et al. [81] presented performed better than the most advanced models. PMA-Net [82] incorporates prototypical memory vectors into Transformer-based image captioning, achieving a good CIDEr boost, by using prior activations to enhance semantic attention and performance. HAAV [83] introduces a novel approach to image captioning by treating heterogeneous encodings (such visual and textual) as enhanced views, employing a hierarchical decoder to adaptively weigh views, and a shared encoder with contrastive loss to enhance representation quality. Notable CIDEr improvements were achieved. Yao et al. [84]

developed DualVision Transformer (Dual-ViT), which offers improved accuracy with efficient token vector compression. In order to achieve state-of-the-art performance for picture description generation, Spatial Pyramid Transformer (SPT) was primarily utilized for adaptive semantic interaction across grid resolutions, maintaining spatial and fine-grained information.

Despite advances in medical imaging and natural language processing (NLP) with models such as the ETA Transformer, Multimodal Transformer (MT), and Dual Level Collaborative Transformer (DLCT), the difficulties in describing images still exist. These difficulties include the lack of global information that is essential for scene understanding and the semantic gap between language and vision. To solve these problems, the proposed Tri-FusionNet combines vision transformer(ViT) with dual-attention, RoBERTa, and CLIP transformers. It is feasible to improve comprehension of visual content using CLIP for visual-textual fusion, RoBERTa for textual interpretation, and ViT for picture embedding.

4.3 Proposed Architecture

The proposed work presents Tri-FusionNet, which combines a dual attention mechanism with Vision Transformer (ViT), CLIP, and RoBERTa decoder to improve spatial and channel-wise information extraction from images. Better descriptive words are generated over a range of images, advancing image description generation. The model is optimized with Adam optimizer and Cross-Entropy loss over epochs. Its performance is assessed using metrics including BLEU, CIDEr, ROUGE-L, and METEOR scores on benchmark datasets. Figure 4-1 represents the framework of the proposed approach.

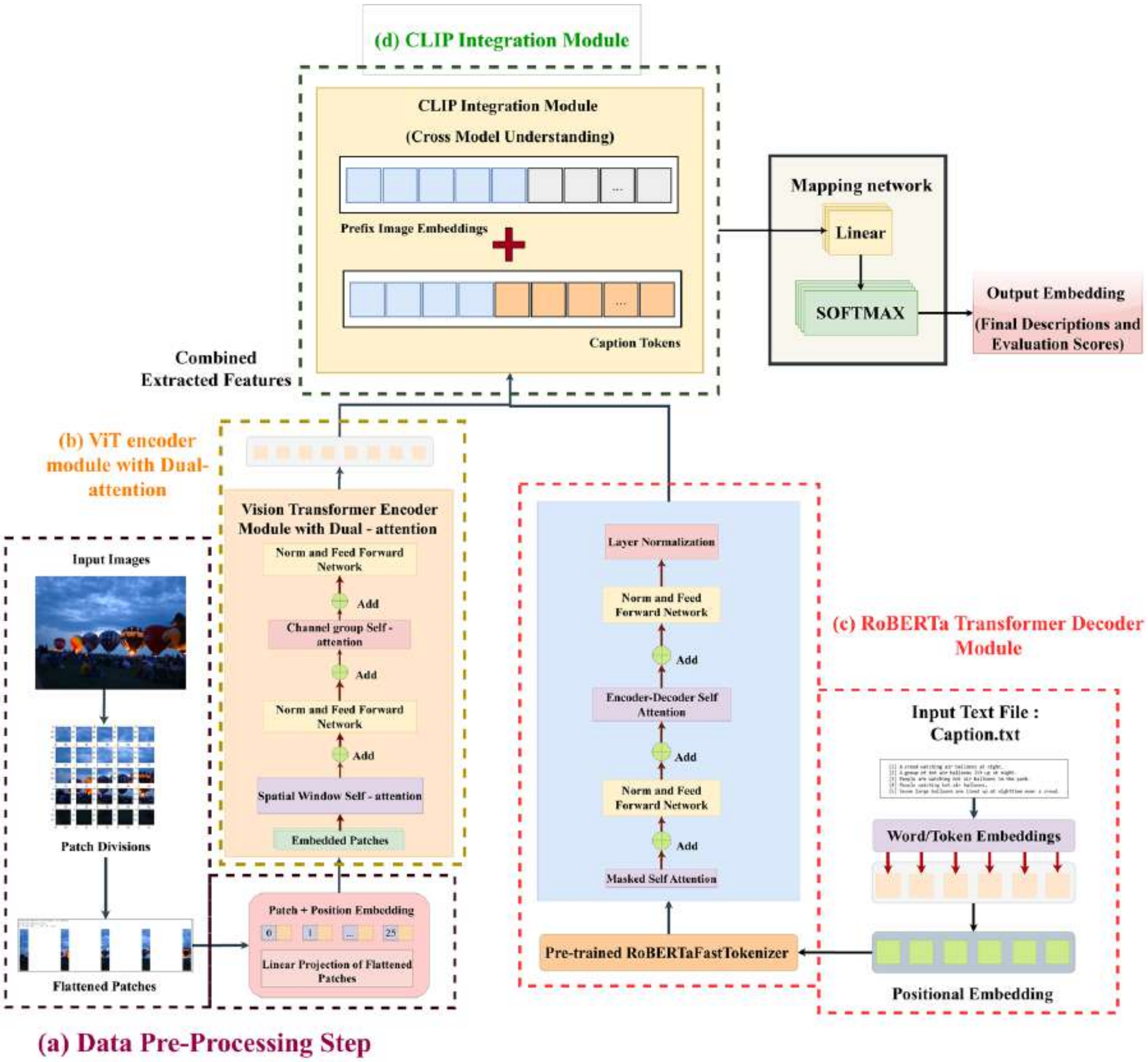

(d) CLIP Integration Module
CLIP Integration Module
(Cross Model Understanding)
Prefix Image Embeddings
Caption Tokens
Combined Extracted Features
Mapping network
Linear
SOFTMAX
Output Embedding
(Final Descriptions and Evaluation Scores)
(b) ViT encoder module with Dual-attention
Vision Transformer Encoder Module with Dual - attention
Norm and Feed Forward Network
Add
Channel group Self - attention
Add
Norm and Feed Forward Network
Add
Spatial Window Self - attention
Embedded Patches
Input Images
Patch Divisions
Flattened Patches
Patch + Position Embedding
Linear Projection of Flattened Patches
(a) Data Pre-Processing Step
Layer Normalization
Norm and Feed Forward Network
Add
Encoder-Decoder Self Attention
Add
Norm and Feed Forward Network
Add
Masked Self Attention
Pre-trained RoBERTaFastTokenizer
(c) RoBERTa Transformer Decoder Module
Input Text File : Caption.txt
Word/Token Embeddings
Positional Embedding

Figure 4-1: Structural representation of Tri-FusionNet framework: The architecture consists of three phases: firstly, high-level visual features are first extracted from preprocessed images using a Vision transformer encoder with dual-attention mechanism; next, words from the input caption file are tokenized by a RoBERTa decoder; and last, the combined data is fed into a CLIP-integrating module to create image descriptions using dense network layers.

4.3.1 Overview of the Proposed Architecture:

Data Pre-processing

The data preprocessing step in the proposed Tri-FusionNet framework involves normalizing the input images' pixel values to enhance the model's performance. The mean and standard deviation of the pixel values throughout the dataset are computed to accomplish this. The distribution of the pixel values is then altered to have a standard deviation (or variance) of one and a mean of zero. This normalization process reduces the effect of varied lighting or contrasts between images by standardizing the range of pixel values. The model can learn more from the data by focusing the values around zero and scaling them to have unit variance, guaranteeing that each feature contributes equally to the training process. Next, to prepare the data for the Vision Transformer (ViT) encoder module in image description generation, input images are divided into fixed-size patches. To construct embeddings, each patch is flattened and projected linearly. The process of patch extraction, where I represents the input image and P denotes the patch size, is illustrated in Equation (4.1). This step divides the image into non-overlapping portions, enabling the ViT to process the image efficiently.

$$\text{Patches} = \text{Reshape}\left(I, \left(\frac{\text{Height}}{P} \times \frac{\text{Width}}{P}\right), P \times P \times \text{Channels}\right) \tag{4.1}$$

Positional encoding is introduced to express spatial information by adding additional vectors (usually sine and cosine functions) to the original element embedding that highlights sequence positions. This improves the model's comprehension of sequence order, which is important for tasks like image description generation. Equation (4.2) defines positional encoding for position pos and dimension dim, where i denotes the index.

$$\text{PE}(\text{pos}, 2i) = \sin\left(\frac{\text{pos}}{10000^{\frac{2i}{\text{dim}}}}\right), \tag{4.2}$$

$$\text{PE}(\text{pos}, 2i+1) = \cos\left(\frac{\text{pos}}{10000^{\frac{2i}{\text{dim}}}}\right) \tag{4.3}$$

With the help of positional encoding, token embedding is further enhanced. Figure 4-1(a) represents the data pre-processing step for the work. The ViT encoder receives sequences of tokenized patches that have been enhanced with positional data. Through this procedure, the model is able to efficiently collect the image's spatial and visual properties, which sets the stage for the next tasks, like labeling the image.

Vision Transformer Encoder Module with Dual Attention Mechanism

The input image is divided into fixed-size patches using the Vision Transformer's encoder-based architecture, which is then processed through transformer layers and linearly embedded. Its ability to retrieve relevant data from both spatial and channel-wise dimensions is enhanced by the dual-attention process, which makes accurate descriptions possible. Figure 4-2 illustrates the complete architectural structure.

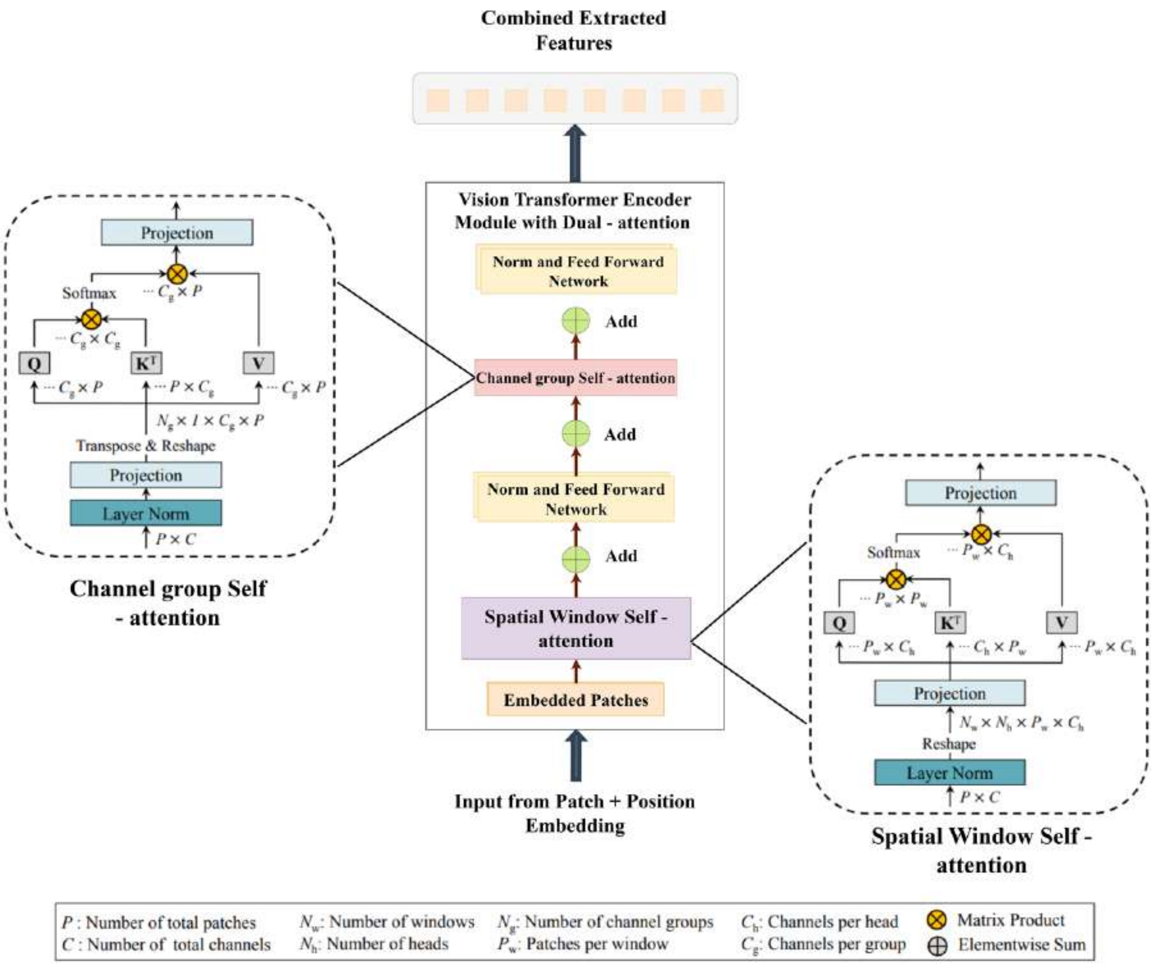

Combined Extracted Features
Vision Transformer Encoder Module with Dual - attention
Norm and Feed Forward Network
Add
Channel group Self - attention
Add
Norm and Feed Forward Network
Add
Spatial Window Self - attention
Embedded Patches
Input from Patch + Position Embedding
Projection
Softmax
$C_g \times P$
$C_g \times C_g$
Q
K^T
V
$C_g \times P$
$P \times C_g$
$C_g \times P$
$N_g \times 1 \times C_g \times P$
Transpose & Reshape
Projection
Layer Norm
$P \times C$
Channel group Self - attention
Projection
Softmax
$P_w \times C_h$
$P_w \times P_w$
Q
K^T
V
$P_w \times C_h$
$C_h \times P_w$
$P_w \times C_h$
Projection
$N_w \times N_h \times P_w \times C_h$
Reshape
Layer Norm
$P \times C$
Spatial Window Self - attention
P : Number of total patches
C : Number of total channels
N_w: Number of windows
N_h: Number of heads
N_g: Number of channel groups
P_w: Patches per window
C_h: Channels per head
C_g: Channels per group
⊗ Matrix Product
⊕ Elementwise Sum

Figure 4-2: Model architecture for vision transformer encoder with dual attention mechanism [85].

In the vision transformer encoder with a dual attention mechanism, we adopt a hierarchical layout. The encoder combines spatial window attention with channel-group attention and is structured into four phases: (a) insertion of the patch embedding layer, (b) application of spatial window attention, (c) utilization of channel-group attention, and (d) incorporation of the jointly extracted features. In the initialization phase, the input is in the form of patches along with positional embedding obtained from the data pre-processing stage. Assuming a visual feature R with dimensions $R^{(P\times C)}$, where P represents the total number of patches and C is the total number of channels, the standard global self-attention is represented by Equation (4.4):

$$S_A(Q,K,V)=\text{Concat}(\text{head}1,\ldots,\text{head}N_h), \tag{4.4}$$

where,

$$\text{head}_i=\text{Attention}(Q_i,K_i,V_i)=\text{Softmax}\left(\frac{Q_iK_i^T}{\sqrt{C_h}}\right)V_i \tag{4.5}$$

here, Q$_i = X_i\times(W_i)^Q$, K$_i = X_i\times(W_i)^K$ and V$_i = X_i\times(W_i)^V$. Consider $R^{P\times C_h}$ dimensional visual features with N_h heads, where X_i represents the ith head of the input feature and W_i denotes the projection weights for the ith head in the context of Q, K, V, with $C=C_h\times N_h$.

The model concurrently arranges spatial window attention and channel group attention in the vision transformer with a dual attention mechanism to obtain both local and global data but with a linear complication to the spatial dimension. The spatial window attention algorithm calculates self-attention within local windows, which are positioned to equally and non-overlapping segments in the field of vision. Assuming the presence of N_w distinct windows, each comprised of P_w patches, the total number of patches, denoted as P, can be expressed as $P=P_w\times N_w$. The representation of spatial window attention is given by Equation (4.6):

$$\text{Attention}w(Q,K,V)=\left(\text{Attention}(Q_i,K_i,V_i)\right)^{N_w}i=0 \tag{4.6}$$

where each $Q_i,K_i,V_i\in\mathbb{R}^{P_w\times C_h}$ are local window queries, keys, and values, respectively, and P is the spatial size presenting the linear complexity.

An alternative viewpoint on self-attention is provided by channel-wise attention, which ensures thorough spatial domain coverage by concentrating on tokens at the patch level as opposed to pixels. Each transposed token, with the number of heads set to 1, engages with global data in a linear spatial complexity along the channel dimension. $C = N_g \times C_g$ is the outcome of letting C_g represent the number of channels in each group and N_g the number of groups. As a result, Equation (4.7) defines global channel group attention, which allows tokens at the image level to interact across channels.

$$\text{Attention}c(Q, K, V) = \Big((\text{Attentiongroup}(Q_i, K_i, V_i))^T\Big)_{i=0}^{N_g} \tag{4.7}$$

where, every $Q_i, K_i, V_i \in \mathbb{R}^{P \times C_g}$ represents grouped channel-wise image-level queries, keys and values.

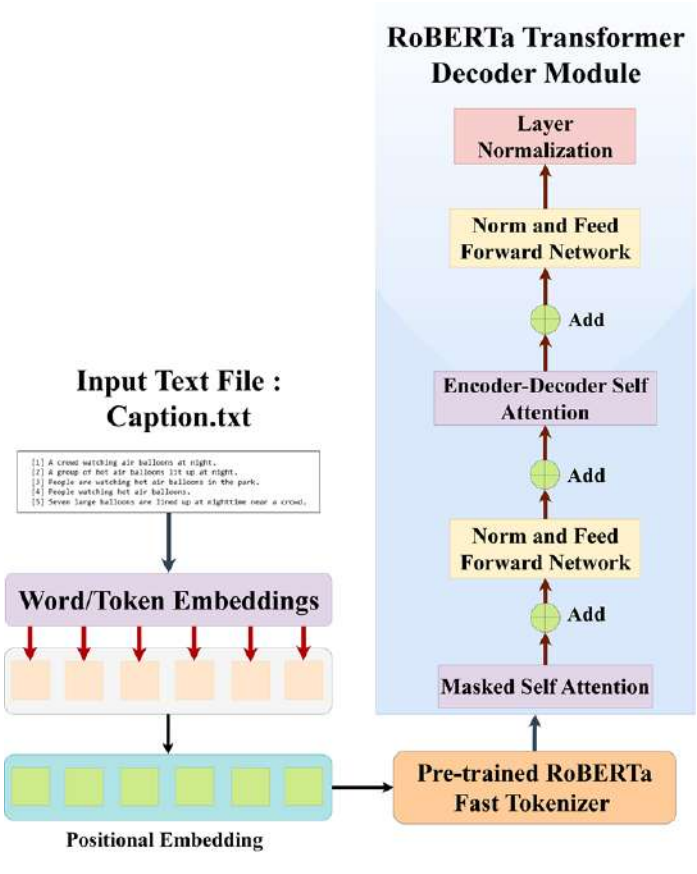
The final encoder output of the Vision Transformer using the dual attention mechanism can be represented by Equation (4.8):

$$E_{\text{output}} = \text{Concat}(\text{Attention}_w(Q, K, V), \text{Attention}_c(Q, K, V)) \tag{4.8}$$

$\text{Attention}_w(Q, K, V)$ represents the spatial window attention and $\text{Attention}_c(Q, K, V)$ represents the channel group attention, as defined in Equations (4.6) and (4.7), respectively. In order to achieve linear complexity in both dimensions for computational efficiency, the final encoder output combines channel group attention and spatial window attention. In contrast to spatial-wise global attention, channel attention functions globally, aggregating information rather than locally. It is a complement to spatial window attention. By distributing weights according to how relevant they are to the task, the model uses its attention mechanism to prioritize particular image patches. By using these weights, heat maps are produced that show the areas of the image where the model focuses. On the heat map, places that are important for accurate and contextually rich descriptions are indicated by intense areas.

RoBERTa Decoder

The study presents a novel approach to image description generation that combines natural language processing (NLP) and computer vision. A Vision Transformer (ViT) with dual attention is employed to extract visual information. Simultaneously, textual data from a "caption.txt" file is used during training to build word embeddings and guide the model in learning semantic relationships between images and their descriptions. This process helps optimize the model and align visual and textual features effectively. During inference, the model does not require an input description or text file. Instead, it generates a new description based solely on the visual features extracted from the input image. This distinction ensures the model's flexibility and generalization, enabling it to produce unique and contextually rich descriptions for each image. These embeddings are seamlessly integrated into a decoder model based on RoBERTa, an enhanced version of BERT, along with ViT-extracted image features. The suggested method's decoder module is shown in Figure 4-3.


RoBERTa Transformer Decoder Module
Layer Normalization
Norm and Feed Forward Network
Add
Encoder-Decoder Self Attention
Add
Norm and Feed Forward Network
Add
Masked Self Attention
Input Text File : Caption.txt
Word/Token Embeddings
Positional Embedding
Pre-trained RoBERTa Fast Tokenizer

[1] A crowd watching air balloons at night.
[2] A group of hot air balloons lit up at night.
[3] People are watching hot air balloons in the park.
[4] People watching hot air balloons.
[5] Seven large balloons are lined up at nighttime near a crowd.

Figure 4-3: Architecture of RoBERTa Decoder Module.

The decoder uses RoBERTa's pre-trained language understanding abilities to contextualize the data from the text and image embedding, allowing it to produce descriptions that generate better explanations and are relevant to the context. To decode and generate image descriptions, the RoBERTa decoder module is obtained from the following Equation (4.9).

$$D_{\text{Output}} = \text{RoBERTaDecoder}\left(W_{\text{emb}}, C\right)\right), \tag{4.9}$$

where, W_{emb} represents the word embeddings generated from the caption data, C is the additional contextual information derived from the original caption file, and D_{Output} is the output generated by the RoBERTa decoder, which consists of the final description of the image. Through multiple layers of feed-forward neural networks and self-attention, multiple word embeddings are processed by the RoBERTa decoder, which enables it to efficiently contextualize the textual information.

CLIP Integration Module

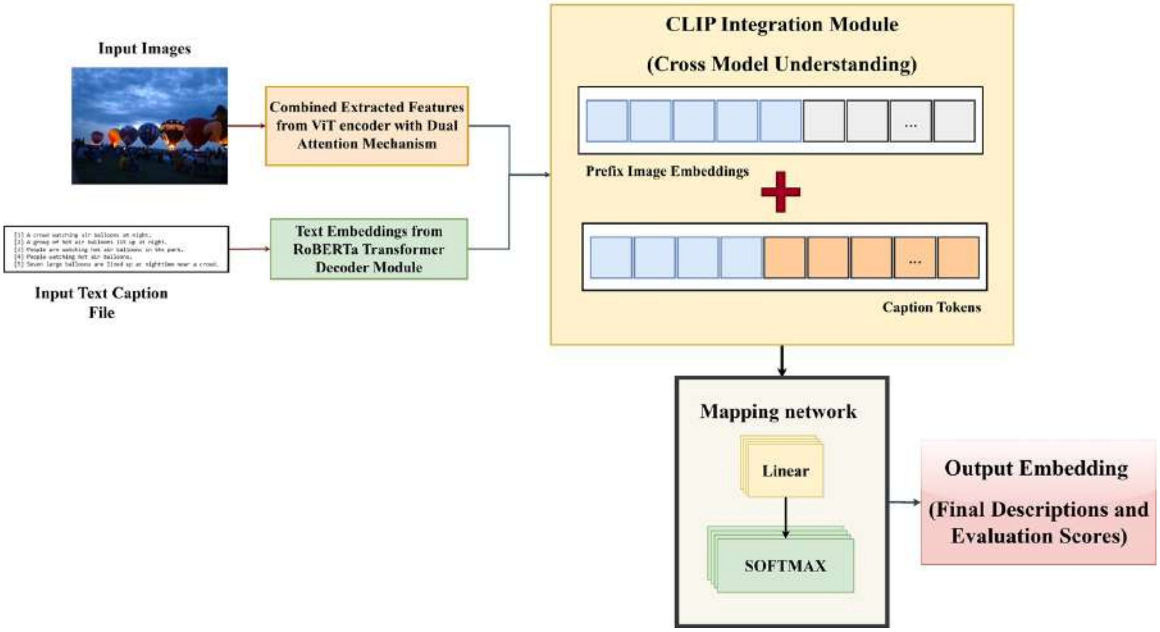
The final step of the proposed approach is the CLIP Integration module, which combines image embeddings from a Vision Transformer (ViT) with a dual attention mechanism and text embeddings from the RoBERTa decoder. This integration aims to combine the ViT encoder, RoBERTa decoder, and CLIP for effective image description generation. During the embedding generation process, the ViT encoder is used to obtain image embeddings from the image dataset, while the RoBERTa decoder generates text embeddings from the caption data. The joint representation of the image and its corresponding text is then obtained by aligning these embeddings in the subsequent phase. Specifically, CLIP's encoder is used to align the visual and textual embeddings through contrastive learning, ensuring that the visual and textual features are mapped into a shared space for accurate description generation.

The embeddings obtained from both the ViT encoder and the RoBERTa decoder are concatenated to form a unified representation, as shown in Equation (4.10):

$$CLIP_{\text{integration}} = \text{Concatenate}(E_{\text{output}}, D_{\text{output}}), \tag{4.10}$$

where E_{output} is the image embedding from the ViT encoder, D_{output} is the text embedding from the RoBERTa decoder, and $CLIP_{\text{integration}}$ represents the combined embedding used by the CLIP model. This integrated representation helps generate coherent and accurate image descriptions by aligning the visual and textual information effectively. The model uses a contrastive loss function in the embedding space to distinguish between positive and negative pairs, ensuring the alignment between visual and textual features.

The final integration module, after encoding and decoding the image-description pair, generates the output descriptions. This process is guided by the CLIP model, as depicted in Figure 4-4.


[1] A crowd watching air balloons at night.
[2] A group of hot air balloons lit up at night.
[3] People are watching hot air balloons in the park.
[4] People watching hot air balloons.
[5] Seven large balloons are lined up at nighttime near a crowd.

Input Images
Combined Extracted Features from ViT encoder with Dual Attention Mechanism
Text Embeddings from RoBERTa Transformer Decoder Module
Input Text Caption File
CLIP Integration Module
(Cross Model Understanding)
...
Prefix Image Embeddings
...
Caption Tokens
Mapping network
Linear
SOFTMAX
Output Embedding
(Final Descriptions and Evaluation Scores)

Figure 4-4: Model architecture for CLIP integration module.

The reference file for descriptions is crucial in assessing how well the Tri-FusionNet model generates descriptions. Visual features extracted from images are projected into a higher-dimensional space using a linear layer, allowing the model to capture intricate relationships among input features. These features are then normalized into a probability distribution across the lexicon of possible words or tokens using a Softmax layer. During training and evaluation, the model generates predicted descriptions, which are compared to reference descriptions using evaluation metrics such as CIDEr, ROUGE-L, BLEU 1-4, and METEOR to assess the model's ability to capture and translate context and image semantics into accurate descriptions.

The Tri-FusionNet model operates through a synergistic pipeline. The Vision Transformer (ViT) encoder with dual attention, RoBERTa decoder, and CLIP integration module collaborate to generate accurate and contextually rich image captions.

- To handle input images, the Vision Transformer (ViT) encoder flattens, divides them into patches, and embeds positional encodings. Its dual attention method combines Spatial Window Self-Attention, which concentrates on local spatial relationships to maintain contextual importance, with Channel Group Self-Attention, which catches fine-grained features inside image channels. As a result, both local details and global context are successfully represented in the rich visual feature embeddings.

- To process the input text file with ground truth descriptions, the RoBERTa decoder tokenises the text, embeds the tokens, then decodes them into contextual embeddings. It uses its pre-trained language understanding to ensure fluency and coherence and Masked Self-Attention to concentrate on sequentially relevant tokens. This produces textual embeddings that effectively express the descriptions' semantic meaning and linguistic structure.

- Using a contrastive learning technique, the CLIP integration module maps both modalities into a common latent space by aligning textual embeddings from the RoBERTa decoder and visual features from the ViT encoder. By facilitating cross-modal fusion and preserving semantic coherence, this alignment makes it possible to produce linguistically correct and visually justified descriptions.

The model optimizes for metrics such as BLEU, CIDEr, METEOR, and ROUGE-L and provides accurate and contextually rich descriptions through parallel processing, dual attention, and contrastive learning. By utilizing each module's unique capabilities to convert raw image and text inputs, the approach guarantees high-quality description generation.

The Algorithm 4.1 for the proposed model with its score evaluation is described below:

Algorithm 4.1 Algorithm for Tri-FusionNet Framework

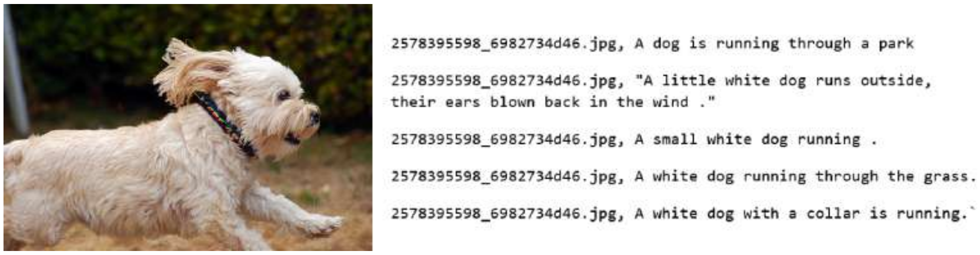
```
procedure GenerateDescription($I, T_{ref}$)
    Input: Image $I$, Reference Description $T_{ref}$
    Output: Generated Description $T$, Evaluation Scores
    Step 1: Pre-processing
    $I_{proc}, T_{enc} \leftarrow$ Preprocess Image and Tokenize Text($I, T_{ref}$)
    Step 2: Feature Extraction using Vision Transformer
    $V_{features} \leftarrow$ Extract Visual Features using ViT($I_{proc}$)
    $V_{enhanced} \leftarrow$ Apply Dual Attention($V_{features}$)
    Step 3: Text Encoding with RoBERTa
    $T_{features} \leftarrow$ Encode Text using RoBERTa($T_{enc}$)
    Step 4: Cross-Modal Feature Fusion with CLIP
    $F_{combined} \leftarrow$ Align Visual and Textual Features using CLIP($V_{enhanced}, T_{features}$)
    Step 5: Description Generation
    $T_{logits} \leftarrow$ Forward Pass through CLIP Decoder($F_{combined}$)
    $T_{gen} \leftarrow$ Decode Logits using CLIP Tokenizer($T_{logits}$)
    $T_{final} \leftarrow$ Process Generated Description($T_{gen}$)
    Step 6: Evaluation
    $Scores \leftarrow$ Compute BLEU, CIDEr, METEOR, ROUGE-L($T_{final}, T_{ref}$)
    Return $T_{final}, Scores$
end procedure
```

4.4 Experimental Analysis

The evaluation of the proposed model requires a system with specific hardware capabilities. Experiments were conducted using Google Colab Pro+, which provides access to high-performance resources. The system utilized includes up to 52 GB of RAM, an NVIDIA A100 Tensor Core GPU with 40 GB of VRAM, and a virtual CPU equivalent to an Intel Xeon processor. The framework was implemented using Keras and TensorFlow 2.12. Tri-FusionNet generates comprehensive and contextually rich image descriptions by integrating three transformer modules, which inherently increases computational demands. Optimization techniques such as empirical analysis, parallelization, mixed precision training, and effective resource management were employed to address these demands. These strategies reduce memory consumption and computational overhead, ensuring efficient training and inference. Despite its complexity, Tri-FusionNet remains viable on hardware with limited resources through model parallelism and other optimizations. The model's scalability was assessed by testing on larger datasets and higher-resolution images, with training times increasing linearly with dataset size demonstrating efficient scaling. Furthermore, its optimized attention mechanisms and parameter-sharing strategies enable competitive computational efficiency compared to other state-of-the-art architectures. These advancements allow Tri-FusionNet to balance high performance with resource constraints, making it adaptable for deployment in diverse environments. The proposed model is evaluated on three benchmark datasets: MSCOCO, Flickr8k, and Flickr30k.

Microsoft Common Objects in Context (MSCOCO) dataset [1]**:** MSCOCO dataset serves as a widely recognized benchmark for tasks related to image description generation within the fields of computer vision and natural language processing (NLP). It plays a crucial role in extensive studies on image interpretation and the generation of pertinent descriptions. The MS COCO 2014 dataset comprises 82,783 JPEG images, each accompanied by approximately 5 human-generated descriptions per image.

Flickr30k Dataset [2]**:** The Flickr30k dataset includes 5 human-annotated reference descriptions along with 31,783 images that are obtained from Flickr. It serves as a common baseline for methods for creating visual descriptions and is primarily used for understanding the visual representation of an image that matches its description. Figure 4-5 is an example of the dataset.



```
2578395598_6982734d46.jpg, A dog is running through a park
2578395598_6982734d46.jpg, "A little white dog runs outside,
their ears blown back in the wind ."
2578395598_6982734d46.jpg, A small white dog running .
2578395598_6982734d46.jpg, A white dog running through the grass.
2578395598_6982734d46.jpg, A white dog with a collar is running.`
```

Figure 4-5: A sample image from the Flickr30k dataset

---

[1] https://github.com/cocodataset/cocoapi

[2] https://www.kaggle.com/hsankesara/flickr-image-dataset

Flickr8k Dataset [3]**:** The Flickr8k dataset consists of 8092 JPEG images in total, which come in various sizes and shapes. The remaining 1000 photos are for development, with the remaining 6000 being used for training and testing. There are five different descriptions for every image. These datasets serve as valuable resources for training and evaluating the proposed model, allowing us to leverage their large-scale image-description pairs to learn robust visual representations and helps in improving the accuracy and quality of the predicted sentences.

The efficiency, computational cost, and scalability of Tri-FusionNet have been compared with the characteristics of some of the most advanced image description models in Table 4.1. Key differences are shown in the table, including performance and resource needs for expanding to larger datasets and higher-resolution images.

Table 4.1: Comparison of Tri-FusionNet with state-of-the-art models on scalability and computational cost.

Model	**Architecture**	**Scalability**	**Computational Cost**	**Efficiency**
NIC [71]	CNN with Attention Mechanism	Limited scalability	Low cost and less resource-intensive	Good for smaller tasks but lacks richness
BLIP Transformer [86]	Vision-Language Pre-training	Scales well with efficient pretraining	Moderate cost and compact architecture	Balanced performance and cost
M2 Transformer [27]	Memory-Augmented Transformer	Effective for mid-sized datasets	Moderate cost and memory-efficient	Competitive with moderate resources
Tri-FusionNet	**ViT, RoBERTa, CLIP**	**Handles large datasets**	**Moderate cost and memory efficient**	**Superior image-text alignment**

Because of its transformer and CLIP components, the proposed Tri-FusionNet model has a moderate processing cost but excels at handling huge datasets. On the other hand, BLIP Transformer [86] provides a fair trade-off between cost and performance, whereas models such as NIC [71] and M2 Transformer [27] are more resource-efficient but less scalable.

---

[3] https://www.kaggle.com/adityajn105/flickr8k

4.4.1 Implementation Details:

In the proposed Tri-FusionNet framework, the model integrates Dual Attention with a Vision Transformer (ViT) to process visual data and employs a RoBERTa decoder for generating textual descriptions. The ViT encoder extracts image features, which are aligned with textual features in a shared embedding space using the CLIP Integrator. This alignment leverages contrastive learning to improve the coherence and precision of the generated descriptions. A novel loss function is employed during training, balancing the objectives of description generation and CLIP alignment. The model is trained over 75 epochs using the Adam optimizer. A batch size of 32 was chosen to balance efficient training, stable gradient updates, and manageable computational resources, ensuring effective learning without excessive memory usage. The input images are pre-processed into patches of fixed size for the ViT encoder, while textual data is tokenized and embedded using RoBERTa. The evaluation of the generated descriptions is carried out using established metrics, including BLEU (1-4), METEOR, ROUGE-L, and CIDEr, which measure linguistic accuracy, contextual relevance, and semantic fidelity. These metrics provide a comprehensive assessment of the model's ability to translate visual content into meaningful and contextually rich textual descriptions.

Table 4.2 outlines the architectural details of the Tri-FusionNet model, highlighting the output shapes and parameter counts for each layer, providing a clear understanding of the structure of the framework.

Table 4.2: Architectural Details for the Proposed Model

Layer Type	**Output Shape**	**Parameters**
ViT Dual Attention	(batch_size, d_model)	2M
RoBERTa Decoder	(batch_size, seq_len)	100M
CLIP Integrator	(batch_size, joint_dim)	10M
Fully Connected Layer	(batch_size, 512)	262,656
Pooling Layer	(batch_size, 256)	0
Convolutional Layer	(batch_size, 128)	295,040
Total Number of Parameters	-	**112.56M**

Table 4.3 provides a summary of the Tri-FusionNet framework's hyperparameters and evaluation metrics.

Table 4.3: Hyperparameters and Evaluation Metrics for the Tri-FusionNet Framework

Hyperparameter	**Value/Description**
Model Name	Tri-FusionNet
Encoder	Vision Transformer (ViT) with Dual Attention
Decoder	RoBERTa
Integration Module	CLIP Integrator with contrastive learning
Training Epochs	75
Batch Size	32
Optimizer	Adam
Loss Function	Categorical Cross-Entropy Loss
Evaluation Metrics	BLEU (1-4), METEOR, ROUGE-L, CIDEr

4.4.2 Ablation Study:

An ablation study was carried out in order to fully comprehend the contributions of each element in the suggested image description generation model. The objective of this study is to assess the impact of the proposed Tri-FusionNet model, which integrates the CLIP model, RoBERTa, and Vision Transformer (ViT) with dual attention, on the overall performance. Through heat map analysis, the model can learn more about how it interprets visual cues and focuses on various areas of the image. This comprehension can enhance the model's architecture, boost its functionality, and make the model's decision-making process more interpretable, as shown in Figure 4-6.

A thorough study of the outcomes within the framework of the work proposed for MSCOCO dataset is shown in Table 4.4. It presents various results of evaluation metrics for different models applied to the input image dataset. The Tri-FusionNet model establishes a new performance benchmark for image description generation on the MSCOCO dataset, outperforming other baseline models. It achieves the highest scores in all metrics, including BLEU-1 (0.893), BLEU-2 (0.821), BLEU-3 (0.794), BLEU-4 (0.725), CIDEr (1.88), METEOR (0.78), and ROUGE-L (0.689). These

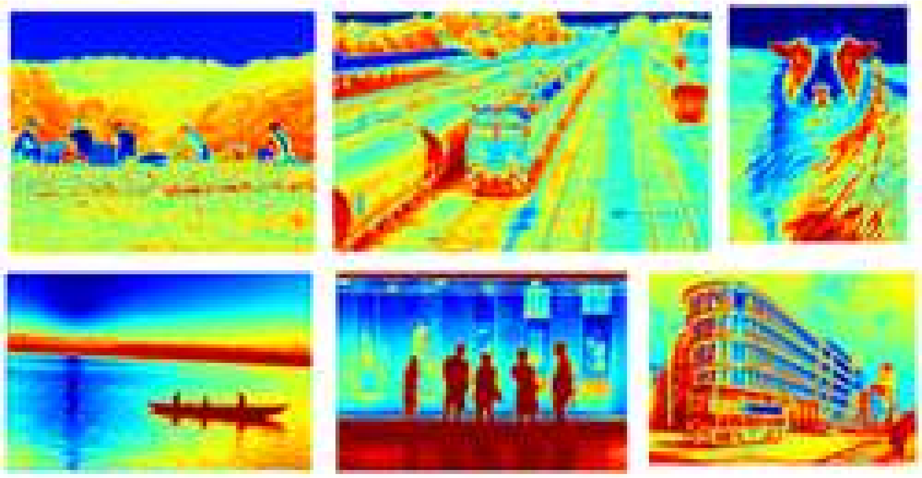


Figure 4-6: Example of obtained heat maps based on dual-attention mechanism.

Table 4.4: Performance Metrics of Image Description Generation Models in the MSCOCO Dataset

Model	**B-1**	**B-2**	**B-3**	**B-4**	**C**	**M**	**R-L**
ViT with Self Attention	0.57	0.45	0.34	0.29	1.05	0.40	0.39
ViT with Dual Attention	0.61	0.50	0.38	0.27	1.52	0.43	0.31
ViT without RoBERTa and CLIP	0.62	0.56	0.42	0.34	0.785	0.44	0.39
RoBERTa without ViT and CLIP	0.73	0.68	0.51	0.43	1.54	0.62	0.50
CLIP without ViT and RoBERTa	0.66	0.54	0.43	0.32	1.09	0.51	0.38
ViT + RoBERTa without CLIP	0.74	0.63	0.55	0.46	1.80	0.68	0.56
ViT + RoBERTa + CLIP without Dual Attention	0.78	0.69	0.65	0.58	1.83	0.73	0.62
Tri-FusionNet (Proposed)	**0.893**	**0.821**	**0.794**	**0.725**	**1.88**	**0.78**	**0.689**

results highlight its effectiveness in generating accurate, diverse, and high-quality image descriptions.

Table 4.5 displays a comprehensive analysis of the results obtained from the Flickr30k dataset in the context of the proposed work.

Table 4.5: Performance Metrics of Image Description Generation Models in the Flickr30k Dataset

Model	**B-1**	**B-2**	**B-3**	**B-4**	**C**	**M**	**R-L**
ViT with Self Attention	0.47	0.33	0.23	0.17	0.907	0.34	0.331
ViT with Dual Attention	0.43	0.36	0.275	0.242	1.02	0.373	0.21
ViT without RoBERTa and CLIP	0.614	0.586	0.542	0.443	0.855	0.544	0.393
RoBERTa without ViT and CLIP	0.653	0.589	0.431	0.367	1.14	0.512	0.460
CLIP without ViT and RoBERTa	0.516	0.424	0.343	0.232	1.256	0.534	0.489
ViT + RoBERTa without CLIP	0.724	0.613	0.455	0.346	1.250	0.368	0.256
ViT+ RoBERTa + CLIP without Dual Attention	0.741	0.621	0.573	0.428	1.092	0.389	0.432
Tri-FusionNet (Proposed)	**0.767**	**0.654**	**0.647**	**0.456**	**1.679**	**0.478**	**0.567**

The Flickr30k dataset was used to evaluate several image description generation models, including ViT with RoBERTa, ViT with Dual Attention, and Tri-FusionNet. The Tri-FusionNet model significantly outperforms the others, achieving the highest BLEU-1 (0.767), BLEU-2 (0.654), BLEU-3 (0.647), and BLEU-4 (0.456) scores. It also excels in CIDEr (1.679), METEOR (0.478), and ROUGE-L (0.567), showcasing its ability to generate accurate, diverse, and context-aware descriptions. Tri-FusionNet establishes a new standard for image captioning on the Flickr30k dataset.

Table 4.6 displays a comprehensive analysis of the results obtained from the Flickr8k dataset in the context of the proposed work. It presents various evaluation metrics for different models applied to the input image. Using the Flickr8k dataset, several image description models were compared. While ViT with Self Attention and ViT with Dual Attention show improved performance, the proposed Tri-FusionNet model outperforms all other models. It achieves the highest scores in BLEU-1 (0.784), BLEU-2 (0.678), BLEU-3 (0.538), and BLEU-4 (0.479), as well as leading in CIDEr (1.381), METEOR (0.389), and ROUGE-L (0.654). This demonstrates its effectiveness in generating accurate, diverse, and structurally rich image descriptions. By combining language and vision transformer components, the Tri-FusionNet consistently outperforms other models in the MSCOCO, Flickr30k and Flickr8k datasets, demonstrating its superior architecture for image description generation and producing better image descriptions. It also yields higher scores in BLEU-1 to BLEU-4 metrics and CIDEr, ROUGE-L and other metrics.

Table 4.6: Performance Metrics of Image Description Generation Models in the Flickr8k dataset

Model	**B-1**	**B-2**	**B-3**	**B-4**	**C**	**M**	**R-L**
ViT with Self Attention	0.547	0.343	0.323	0.217	0.607	0.234	0.131
ViT with Dual Attention	0.543	0.436	0.344	0.246	1.002	0.253	0.121
ViT without RoBERTa and CLIP	0.542	0.466	0.342	0.266	0.675	0.344	0.343
RoBERTa without ViT and CLIP	0.553	0.459	0.413	0.337	0.983	0.312	0.156
CLIP without ViT and RoBERTa	0.525	0.324	0.244	0.132	1.065	0.234	0.289
ViT + RoBERTa without CLIP	0.745	0.513	0.456	0.312	1.189	0.298	0.457
ViT+ RoBERTa + CLIP without Dual Attention	0.756	0.652	0.518	0.460	1.231	0.321	0.528
Tri-FusionNet (Proposed)	**0.784**	**0.678**	**0.538**	**0.479**	**1.381**	**0.389**	**0.654**

4.4.3 Results and Analysis:

The approach was evaluated on three benchmark dataset: MSCOCO, Flickr30k and Flickr8k for the proposed architecture. The following Table 4.7 represents the results obtained for the MSCOCO dataset. Using three sets of test images from the MSCOCO dataset, the table presents a thorough comparison of ground-truth descriptions and predicted descriptions generated by a model. This is an effective way of assessing and determining how well the model can provide meaningful and accurate descriptions of images.

Table 4.7: Quantitative Results Obtained on MSCOCO Dataset

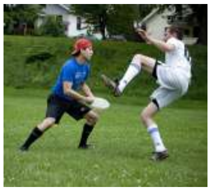
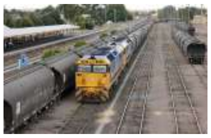
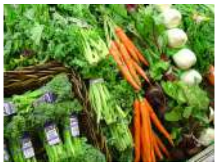
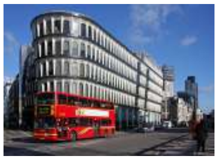
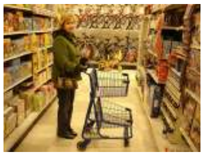
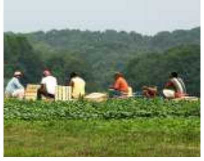
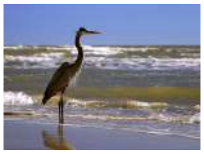
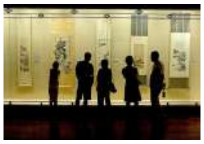
Test Image	**Ground Truth**	**Predicted Description**
	Two men playing frisbee on grass surrounded by trees.	Two men playing frisbee on grass.
	Trains on railway tracks, with trees and blue sky.	Trains on tracks with trees.
	Vegetables including carrot, radish, and turnip on a table.	Green and red vegetables on a table.
	A red bus in front of a white building with blue sky.	A white building with blue sky.

Table 4.8 aims to showcase how well the model aligns for Flickr30k dataset. Four sets of test images are listed in the table, together with the accompanying ground truth descriptions and the model's predicted descriptions for each. This table is as an assessment tool, demonstrating the model's capacity to produce accurate and relevant descriptions for a range of images found in the Flickr30k dataset.

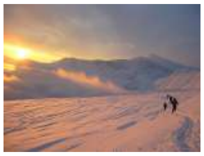
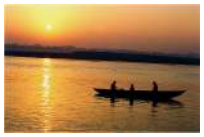
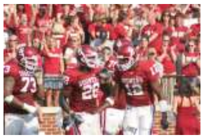
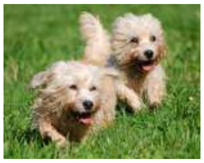
Table 4.9 is designed to demonstrate the alignment of the model with the actual

Table 4.8: Quantitative Results Obtained on Flickr30k Dataset

Test Image	**Ground Truth Descriptions**	**Predicted Description**
	1. A white woman standing in a grocery store not-so-candidly posing for the camera while examining the items on a shelf. 2. The lady with the shopping cart is surrounded by toys galore and various children's bicycles. 3. A woman in a green winter coat stands with a cart in the middle of a department store aisle. 4. A woman in a green jacket is in the toy aisle with a shopping cart and her purse. 5. The woman in the green coat is pushing a cart through the toy aisle.	A woman in a green coat shops in the toy aisle with a cart.
	1. Six men are sitting or laying on a patch of earth in a wooded area. 2. A group of workers sitting in a field take a break from work. 3. A group of men are sitting in the farm fields taking a break. 4. Six men sit in a field of crops containing wooden crates. 5. Pickers working out on a farm.	Six men take a break in a field.
	1. A gray bird stands majestically on a beach while waves roll in. 2. A white crane stands tall as it looks out upon the ocean. 3. A tall bird is standing on the sand beside the ocean. 4. A large bird stands in the water on the beach. 5. A water bird standing at the ocean's edge.	A bird stands tall on the beach.
	1. A group of people stares at a wall that is filled with drawings in a building. 2. There are five people here looking at some pictures on the wall. 3. Five people are taking in an exhibit of Japanese art. 4. People watching the arts in an exhibition. 5. Five people looking at artwork.	Five people looking at art in a building.

content of the images, as indicated by the provided ground truth descriptions for the Flickr8k dataset.

Table 4.9: Quantitative Results Obtained on Flickr8K Dataset

Test Image	**Ground Truth Descriptions**	**Predicted Description**
	1. A winter landscape with four people walking in the snow. 2. Beautiful snowy landscape with people treading through the snow. 3. Cross-country skiers are traveling towards the mountains at sunset. 4. Four people walking across thick snow during a sunset. 5. The sun is almost behind the snowy mountains.	Four people walk through a snowy mountain.
	1. A beautiful sunset with three people in a boat on the lake. 2. As the sun sets, three people are on a small boat enjoying the view. 3. Three people are in a canoe on a calm lake with the sun reflecting yellow. 4. Three people are on a boat in the middle of the water while the sun is in the back. 5. Three people in a boat float on the water at sunset.	Three people enjoying a beautiful sunset from a boat.
	1. A crowd wearing red cheers on the red football team. 2. Football players in red congratulate each other as crowds in red cheer behind. 3. The Oklahoma Sooners football team discuss their game while fans cheer. 4. Two football players talk during a game. 5. Two Oklahoma Sooner football players talk on the sideline.	A crowd in red cheers on the football team.
	1. The two small dogs run through the grass. 2. Two fluffy white dogs running in green grass. 3. Two small dogs run through the grass. 4. Two small dogs that look almost identical are playing in the grass. 5. Two yellow dogs run together in green grass.	Two small dogs run through the green grass.

Two sets of test images are included in the table, each with the corresponding ground truth descriptions and the predicted descriptions from the model. This table serves as a means of assessment, showcasing the model's capacity to provide accurate and perceptive descriptions for a range of images from the Flickr8k dataset.

When compared to reference descriptions, the model produces predicted descriptions during the training and evaluation phases. The model's efficacy in capturing visual semantics and context is evaluated by comparing the predicted and real descriptions using metrics such as BLEU 1-4, CIDEr, METEOR and ROUGE-L. Table 4.10 provides an overview of a framework that has been suggested for the development of image descriptions from the three datasets.

Table 4.10: Overall Results obtained from the proposed model

Dataset	**B-1**	**B-2**	**B-3**	**B-4**	**C**	**M**	**R-L**
MSCOCO	0.893	0.821	0.794	0.725	1.88	0.78	0.689
Flickr30k	0.767	0.654	0.647	0.456	1.679	0.478	0.567
Flickr8k	0.784	0.678	0.538	0.479	1.381	0.389	0.654

The performance metrics of the proposed model are shown in the table for the MSCOCO, Flickr30k and Flickr8k datasets. These metrics include BLEU-1 to BLEU-4, CIDEr, METEOR and ROUGE-L scores. The suggested framework obtained BLEU scores for MSCOCO of 0.893 (B-1), 0.821 (B-2), 0.794 (B-3) and 0.725 (B-4) and for CIDEr, METEOR and ROUGE-L, 1.483, 0.358 and 0.789, respectively. With BLEU scores ranging from 0.767 (B-1) to 0.456 (B-4), bolstered by a CIDEr score of 1.679 and METEOR (0.478) and ROUGE-L (0.567) ratings suggesting sufficient matching and overlap, competitive performance was observed on the Flickr30k dataset. A CIDEr score of 1.381, favorable matches, and overlap was shown by BLEU scores on the Flickr8k dataset, which varied from 0.784 (B-1) to 0.479 (B-4). METEOR (0.389) and ROUGE-L (0.654) scores also showed favorable matches and overlap. Across a range of benchmark datasets, the proposed approach performs well overall in producing image descriptions. The graphical depiction of the outcomes is shown in Figure 4-7. These scores serve as evaluations of the model's performance in generating descriptions that align well with the reference descriptions. Higher scores indicate a higher degree of similarity and quality in the generated descriptions.

Table 4.11 presents a qualitative comparison between the ground truth descriptions and the descriptions generated by the proposed model, highlighting successful

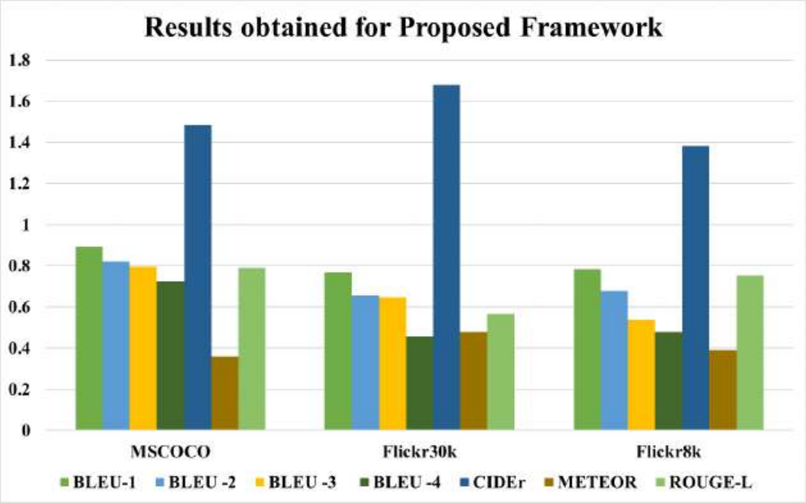
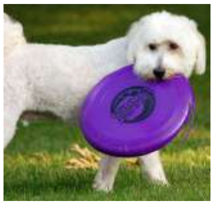
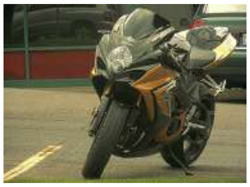
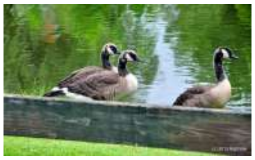
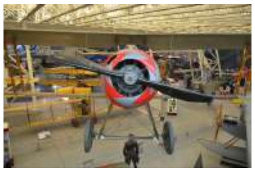


Figure 4-7: Graphical representation of the results obtained from MSCOCO, Flickr30k and Flickr8k dataset for the proposed Tri-FusionNet framework.

and unsuccessful predictions and error analysis. The following table compares the

Table 4.11: Qualitative Analysis of Generated Image Descriptions

Test Image	**Ground Truth Description**	**Generated Description**	**Remarks/Error Analysis**
	A white dog is holding a purple frisbee in its mouth on the green grass.	White dog holding purple frisbee on grass.	**Success:** Correct identification of objects and actions.
	A gold and black motorcycle parked on a paved surface road.	A car parked on road.	**Failure:** Incorrect object recognition.
	Three ducks stand by a calm pond with a wooden fence in front of them.	Three ducks stand by pond.	**Success:** Correct identification of objects and actions.
	A red and white vintage plane on display in a museum.	Spaceship preparing to launch.	**Failure:** Incorrect object recognition.

generated descriptions of four sample images with ground truth annotations to evaluate the performance of an image captioning model. Successful examples include correctly identifying "three ducks on a peaceful pond" and "a white dog clutching a purple frisbee on grass." These cases demonstrate the model's ability to recognize objects and actions in straightforward scenarios. However, the model exhibits notable failures, such as mistaking a gold and black motorcycle for "a car parked on a road" and, in row 4, misidentifying a red and white vintage plane as "a spaceship preparing to launch." These errors highlight the model's difficulty in accurately identifying complex objects or understanding the context of the scene. This analysis provides valuable insights into the model's strengths in simple situations and its limitations when faced with more intricate or ambiguous visuals, offering guidance for improving the model's training and recognition capabilities.

4.4.4 Comparison with Other State-of-the-art Methods:

A comprehensive generalisation analysis highlights the performance of the proposed model outside of the training set by assessing its capacity to produce image descriptions over a wide range of visual domains and datasets. A comparison study of image description generation models using the MSCOCO dataset is presented in Table 4.12, which evaluates model performance using metrics such as BLEU, METEOR, CIDEr and Rouge-L.

The performance of different models on the MSCOCO dataset indicates strong progress in image description generation. Models like Meshed Transformer [27], Global Enhanced Transformer [89], and Multimodal Transformer [29] show strong overall performance in terms of high scores obtained in BLEU, METEOR, CIDEr, and Rouge-L metrics. In contrast, models like Topic-based multi-channel attention (TMA) [87], Transformer-based local graph semantic attention (TLGSA) [59], and Dynamic-balanced double-attention [88] have relatively weaker performance, especially in terms of BLEU and CIDEr scores. Geometry Attention Transformer [91] and S^2 Transformer [79] obtain the best CIDEr scores, while Local-global guidance for transformer [94] reaches the best Rouge-L score of 0.651. Among the best-performing

Table 4.12: Comparative Analysis of Image Description Generation Models on MSCOCO Dataset

Model	**BLEU**	**METEOR**	**CIDEr**	**Rouge-L**
Topic-based multi-channel attention (TMA) model [87]	0.658	-	0.800	0.534
Transformer based local graph semantic attention (TLGSA) [59]	0.724	-	1.003	0.534
Dynamic-balanced double-attention [88]	0.741	0.254	1.107	0.537
Meshed Transformer [27]	0.816	0.294	1.293	0.592
Global Enhanced Transformer [89]	0.816	0.284	1.301	0.591
Multimodal Transformer [29]	0.817	0.294	1.30	0.596
S^2 Transformer [79]	0.811	0.296	1.335	0.591
PMA-Net [82]	0.847	0.305	1.414	0.613
HAAV [83]	0.810	0.302	1.415	-
SPT Transformer [90]	0.812	0.296	1.344	0.592
Geometry attention transformer [91]	0.811	0.384	1.27	0.591
Vision-enhanced and Consensus-aware Transformer [92]	0.822	0.296	1.345	0.596
X-transformer + Faster RCNN [93]	0.821	0.296	1.334	0.598
Local-global guidance for transformer [94]	0.861	0.392	1.329	0.651
SAMT [78]	0.774	0.284	1.205	0.572
Tri-FusionNet (Proposed)	**0.893**	**0.780**	**1.880**	**0.689**

models, PMA-Net [82] and HAAV [83] obtain CIDEr scores higher than 1.4. In addition, PMA-Net obtains the best BLEU score of 0.847 among all the models. However, Tri-FusionNet breaks all other records as the new baseline, obtaining state-of-the-art performance scores for both BLEU 0.893, METEOR 0.780, CIDEr 1.880, and also Rouge-L with a 0.689 score in all the tested metrics, which highlights its capacity in producing better descriptions with semantic accuracy of an image for all the existing models.

To further validate the performance of the proposed Tri-FusionNet model, we evaluated it on the online MSCOCO test server, which is widely used as a benchmark for image description generation. The comparative results are presented in Table 4.13, highlighting the superior performance of Tri-FusionNet in all the metrics, demonstrating its efficiency in generating contextually rich and accurate image descriptions. The performance comparison of different image captioning models on the MSCOCO online test server is displayed in the table. The effectiveness of the suggested Tri-FusionNet in generating precise and varied image descriptions is demonstrated by the fact that it outperforms all current techniques across BLEU, METEOR, ROUGE-L, and CIDEr measures.

Table 4.14 compares various image description generation models using the Flickr30k

Table 4.13: Comparative Analysis of Image Description Generation Models on online MSCOCO Test Server

Model	**BLEU**	**METEOR**	**Rouge-L**	**CIDEr**
NIC [71]	0.714	0.302	0.521	0.945
m-RNN [48]	0.753	0.301	0.593	0.926
ReviewNet [95]	0.810	0.301	0.609	0.967
SCN [96]	0.828	0.309	0.619	1.013
Adaptive [74]	0.834	0.311	0.628	1.051
Att2all [19]	0.859	0.312	0.635	1.157
GateCap [97]	0.863	0.319	0.644	1.190
LSTM-A [25]	0.862	0.312	0.635	1.170
Up-Down [98]	0.877	0.322	0.648	1.192
RFNet [99]	0.877	0.327	0.657	1.240
GCN-LSTM [52]	-	0.330	0.660	1.259
SGAE [64]	0.882	0.327	0.661	1.252
AoANet [100]	0.880	0.338	0.667	1.282
Tri-FusionNet (Proposed)	**0.885**	**0.750**	**0.678**	**1.580**

Table 4.14: Comparative Analysis of Image Description Generation Models on Flickr30k Datasets

Model	**BLEU**	**METEOR**	**CIDEr**	**Rouge-L**
Topic-based multi-channel attention (TMA) model [87]	0.650	-	0.334	0.436
Transformer based local graph semantic attention (TLGSA) [59]	0.643	-	0.450	0.489
Dynamic-balanced double-attention [88]	0.678	0.209	0.517	0.500
HAAV [83]	0.743	0.251	0.856	-
Multimodal Transformer [29]	0.744	0.236	-	-
Local-global guidance for transformer [94]	0.758	0.263	0.708	0.560
X-transformer + Faster RCNN [93]	0.753	0.253	0.707	0.543
Tri-FusionNet (Proposed)	**0.767**	**0.478**	**1.679**	**0.567**

dataset. The analysis includes metrics for evaluating the models' performance, such as BLEU, METEOR, CIDEr, and Rouge-L. Models like Transformer-based local graph semantic attention (TLGSA) [59], Topic-based multi-channel attention (TMA) [87], and Dynamic-balanced double- attention [88] show lower performance, particularly in BLEU and CIDEr, indicating limitations in generating high-quality descriptions. While the Multimodal Transformer [29] performs well in BLEU and METEOR, it lacks comprehensive metric coverage. Local-global guidance for Transformer [94] and X-transformer + Faster RCNN [93] demonstrate strong BLEU and CIDEr scores but fall slightly behind in Rouge-L. HAAV [83] achieves notable performance, especially with a CIDEr score of 0.856. However, the proposed Tri-FusionNet outperforms all models, achieving the highest scores across all metrics: 0.767 for BLEU, 0.478 for METEOR, 1.679 for CIDEr, and 0.567 for Rouge-L, establishing it as the most effective model for generating accurate and contextually rich descriptions on the Flickr30k dataset.

Table 4.15 presents a comparative analysis of various image description generation models on the Flickr8k dataset. It offers evaluation metrics for evaluating how well various models perform, including BLEU, METEOR, CIDEr and Rouge-L. Models like TLGSA based on transformer [59], Vision encoder-decoder [101], Topic-based multi-channel attention [87], and Optimal transformers with beam search [102] have performance relatively lower and are bad at BLEU and CIDEr score, due to the lesser ability for producing high quality and relevant descriptions. Specifically, even though SAMT [78] has scores of BLEU and METEOR comparatively better values, it also lacks other competitive values of CIDEr and Rouge-L. On the other hand, proposed Tri-FusionNet obtains excellent results with setting new state-of-the-art for all metrics: BLEU-0.784, METEOR-0.389, CIDEr-1.381, and Rouge-L - 0.654. Such results emphasize that the proposed Tri-FusionNet model is more likely to produce exceptional, meaningful, and expressive image descriptions compared to existing models for Flickr8k.

Table 4.15: Comparative Analysis of Image Description Generation Models on Flickr8k Dataset

Model	**BLEU**	**METEOR**	**CIDEr**	**Rouge-L**
Topic-based multi-channel attention (TMA) model [87]	0.630	-	0.472	0.465
Transformer based local graph semantic attention (TLGSA) [59]	0.659	-	0.471	0.565
Vision encoder decoder [101]	0.395	0.177	0.380	0.297
Optimal transformers with Beam Search [102]	0.634	0.1987	0.520	-
SAMT [78]	0.682	0.212	-	0.448
Tri-FusionNet (Proposed)	**0.784**	**0.389**	**1.381**	**0.654**

4.5 Conclusion

In this chapter, a new model, Tri-FusionNet, was presented to generate image descriptions that combine the CLIP transformer, RoBERTa, and the Vision Transformer with dual attention processes. The proposed model outperforms state-of-the-art models as a result of extensive trials on the MSCOCO, Flickr30k, and Flickr8k datasets. It achieved notable gains across important evaluation criteria, such as BLEU, METEOR, CIDEr, and ROUGE. The model is successful because it can capture both local and global image data, use the CLIP transformer to effectively align textual and visual modalities, and employ RoBERTa for improved language understanding. The suggested model has useful applications in real-time image captioning systems, including autonomous car systems for scene detection, assistive technologies for the blind, and content management systems for automatic image tagging. Deploying the model still presents difficulties, though, such as managing edge devices' demanding processing requirements, ensuring the system is resilient to threats and noisy inputs, and resolving ethical challenges like bias in generated descriptions. Furthermore, the model's performance may be impacted by the unpredictability introduced by real-world settings, such as occlusions, dim lighting, computational cost, and other social circumstances. To enable more extensive real-world applications, future research will concentrate on improving fine-tuning procedures, tackling deployment issues, and extending the model's capabilities through sophisticated multimodal fusion techniques.

Chapter 5

Image Description Models for Multimedia Application: Chest X-Ray Analysis

Lakshita Agarwal and Bindu Verma. "Advanced Chest X-Ray Analysis via Transformer-Based Image Descriptors and Cross-Model Attention Mechanism."
Archived at: http://arxiv.org/abs/2504.16774. Communicated in: Computational Intelligence (Wiley), SCIE Indexed, Impact Factor 1.8. ***(Communicated).***

Lakshita Agarwal and Bindu Verma. "Utilizing Transformer-Based Image Descriptors for Improving Chest X-Ray Analysis." In: Nanda, S.J., Yadav, R.P., Gandomi, A.H., Saraswat, M. (eds) Data Science and Applications. ICDSA 2024. Lecture Notes in Networks and Systems, vol 1266. Springer, Singapore. ***(Published).***

5.1 Introduction

Building on the advances of transformer-based models in image description generation discussed in the previous chapter, this chapter focuses on a specialized application in the medical domain: chest X-ray analysis. Although models like Tri-FusionNet have demonstrated their effectiveness in generating detailed and contextually rich descriptions, medical imaging presents unique challenges, such as domain-specific language, subtle abnormalities, and the need for high interpretability. This chapter introduces a novel framework that integrates a Vision Transformer (ViT) encoder with cross-modal attention and a GPT-4-based transformer decoder to address these challenges. ViT extracts high-quality visual features that are fused with textual data through cross-modal attention to enhance contextual relevance and precision. The GPT-4 decoder then generates accurate and detailed descriptions. The model was evaluated on the IU and NIH chest X-ray data sets, achieving high BLEU, CIDEr, METEOR, and ROUGE-L scores. This approach improves chest radiograph interpretation, helping radiologists in efficient diagnosis and treatment.

The key contributions of this chapter are mentioned below:

- The proposed work uses ViT to capture spatial relationships and medical terminology, improving the model's ability to interpret complex radiographic features.
- Using GPT-4, it processes pixel-level information and generates precise and contextually relevant descriptions for medical imaging.
- The model achieves superior performance compared to existing methods on the IU and NIH Chest X-ray datasets, demonstrating improved accuracy and reliability.
- The chapter highlight the potential of the model to enhance clinical workflows, assist in diagnosis, and support treatment planning through automated medical image analysis.
- It provides a comprehensive evaluation of the strengths and limitations of the model, showcasing its effectiveness in improving automated radiology reporting.

5.2 Literature Survey

The primary objective of the image description model is to identify different objects and represent their relationships through accurate, semantically correct sentences. Various methods have been adapted to medical imaging tasks [103]. Sun et al. [104] proposed a feature-augmented (FA) module validated on datasets such as MS-COCO, which uses the multi-modal pre-trained CLIP model and channels attention within the encoder to enhance image description and captioning performance. Yao et al. [84] proposed a CNN-RNN framework for anatomical structure and abnormality identification in radiological imaging reports. Li et al. [105] used BERT-based models to generate radiological reports from chest X-ray images, showing the significance of contextual language understanding in medical image analysis. Shaikh et al. [106] proposed an encoder-decoder transformer model combined with a pre-trained CheXNet model, evaluated on the IU X-ray dataset, for chest X-ray report generation. The CheXReport model [107] obtains the best-reported performance on the MIMIC-CXR dataset, given the use of Swin Transformer blocks. Retrieval-based methods that incorporate deep neural networks are also presented [108]. In another work, Conditional Self Attention Memory-Driven Transformer was proposed, which outperformed all existing state-of-the-art approaches with a high BLEU score value for radiological report production by taking ResNet152 v2 for feature extraction and the self-attention memory-driven transformer for text generation [109]. Despite these developments, current models fail to integrate textual and visual information effectively and are not able to handle medical language well, which makes the descriptions less understandable and accurate. Moreover, because these models are highly sensitive to errors, misunderstandings in medical contexts could potentially affect diagnoses.

The suggested approach, CrossViT-GPT4, enhances previous research by integrating the benefits of GPT-4, which excels in contextual language modeling, with ViT and a cross-model attention technique for spatial feature extraction. This combination offers a comprehensive solution for automatically generating image descriptions in chest X-ray analysis.

5.3 Proposed Architecture

The complicated process of automatically generating descriptions for medical images has drawn a lot of interest, necessitating the combination of computer vision and natural language processing (NLP). The goal of the research is to improve the extraction of spatial and semantic information from chest X-rays by utilising a hybrid framework that integrates GPT-4, cross-modal attention, and Vision Transformer (ViT). ViT efficiently records spatial relationships and fine-grained medical features, whereas GPT-4 generates accurate and contextually rich descriptions by processing pixel-level data. The suggested framework enhances automated chest X-ray interpretation, supporting clinical diagnosis and decision-making by tackling issues with producing precise and semantically meaningful radiological reports. Figure 5-1 represents the overall framework of the proposed approach.

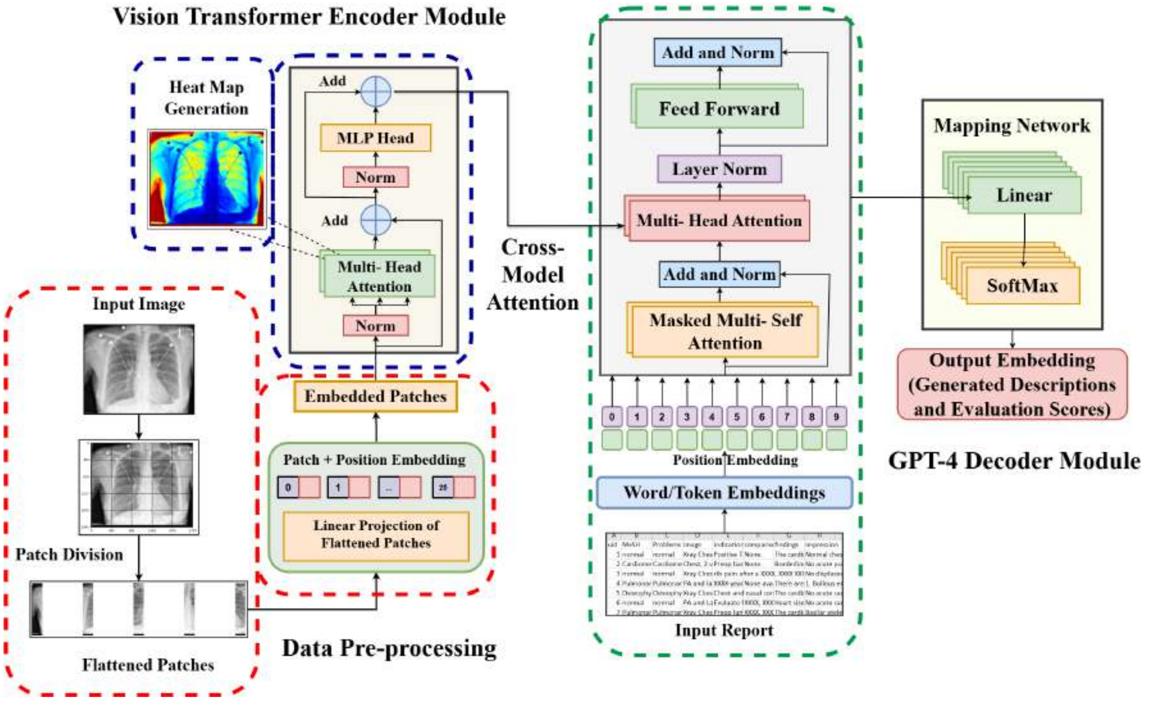


Figure 5-1: A representation of the proposed framework's structure: CrossViT-GPT4- During the initial stage, the ViT encoder utilizes cross-model attention to extract high-level visual features from the pre-processed images. GPT-4 decoder uses tokenization to extract individual words from the provided input caption file. It then utilizes the dense layers of the network to produce image descriptions.

Each step of the proposed model is discussed below:

5.3.1 Data Pre-processing

The input images (I) are initially partitioned into patches (P) of a predetermined size to prepare the input data for the encoder. Each image of size ($H \times W \times C$) is divided into smaller patches of size ($P_h \times P_w$), where H and W are the image height and width, C is the number of channels, and P_h, P_w define the patch dimensions. The total number of patches (N) is given by:

$$N = \frac{HW}{P_h P_w} \tag{5.1}$$

Each patch is then reshaped into a flattened vector:

$$P = \text{Reshape}(I) \in \mathbb{R}^{N \times (P_h P_w C)} \tag{5.2}$$

where Reshape(I) denotes the operation that reorganizes the image into patch vectors. These flattened patches undergo a linear projection to embed them into a lower-dimensional space:

$$E_{\text{patch}} = \text{Linear}(\text{Flatten}(I_{\text{patch}})) + \text{PE} \tag{5.3}$$

where I_{patch} represents the image patches, Linear denotes the linear transformation matrix used for embedding, and Flatten is the operation that flattens each patch into a vector. The positional encoding (PE) uses sine and cosine functions to incorporate spatial information, ensuring the model understands the positional arrangement of patches. This is defined as:

$$\text{PE}(\text{pos}, 2i) = \sin\left(\frac{\text{pos}}{10000^{(2i/\text{dim})}}\right) \tag{5.4}$$

$$\text{PE}(\text{pos}, 2i+1) = \cos\left(\frac{\text{pos}}{10000^{(2i/\text{dim})}}\right) \tag{5.5}$$

The overall data pre-processing workflow is illustrated in Figure 5-2.

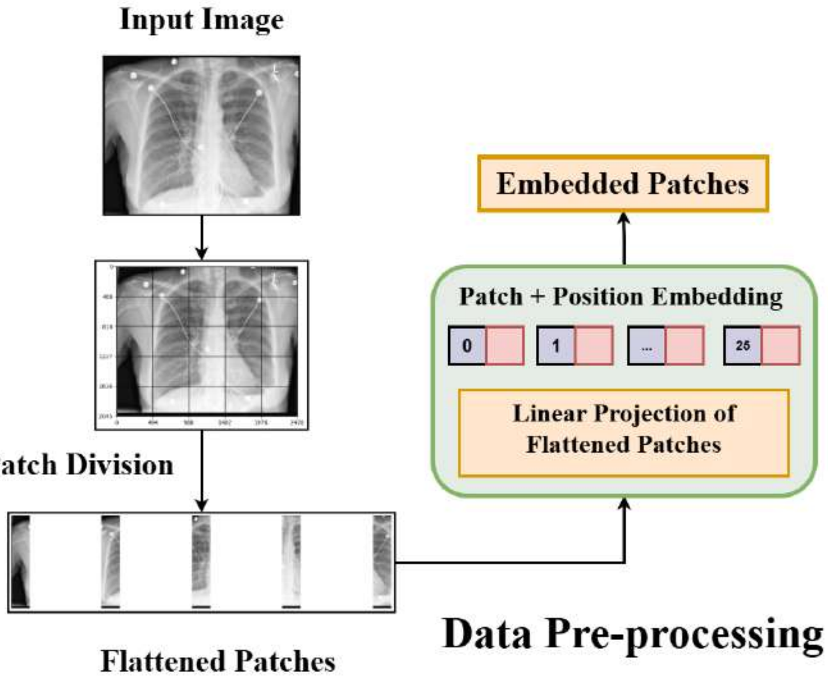


Figure 5-2: Data Pre-Processing Step

This pre-processing step ensures that image patches retain their spatial information while being effectively embedded into a sequence of vectors for further processing.

5.3.2 Encoder Module of the Proposed Architecture

The encoder part efficiently collects the spatial and visual characteristics of the image, thereby preparing the model for subsequent tasks such as image labeling. The vision transformer (ViT) encoder is used to receive the sequences of tokenized patches that have been enhanced with positional data. The architectural structure for the Vision transformer encoder is illustrated in Figure 5-3. For multi-head self-attention, the attention is computed in parallel across h heads and concatenated, as denoted in Equations (5.6), (5.7) and (5.8):

$$\text{Attention}(Q, K, V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{5.6}$$

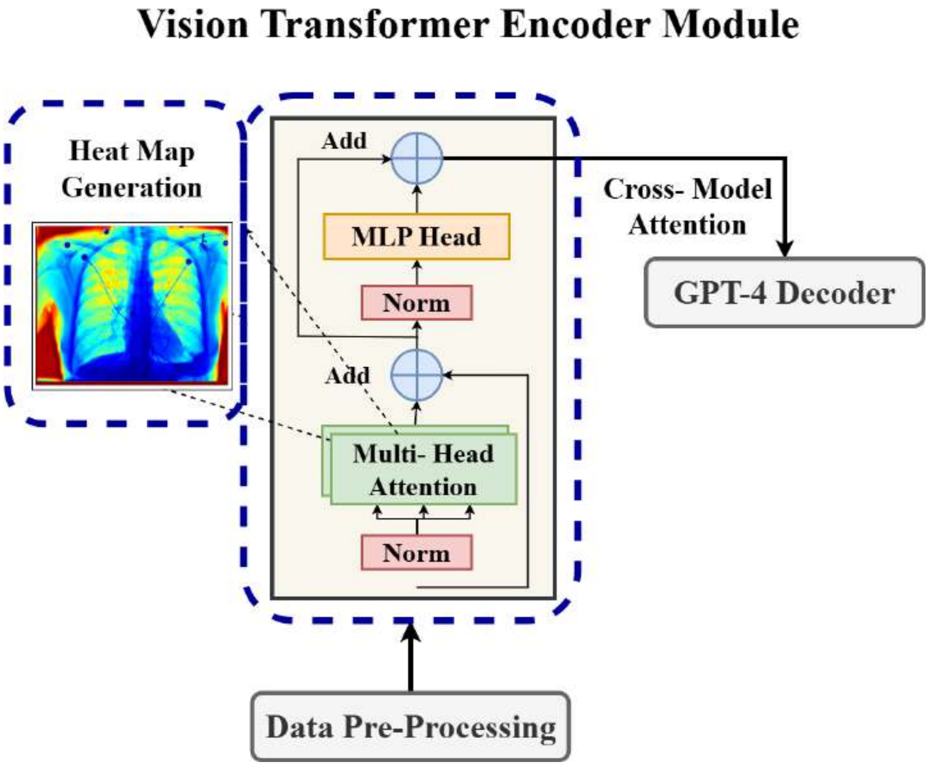


Figure 5-3: Illustration of ViT- Encoder module with heat maps

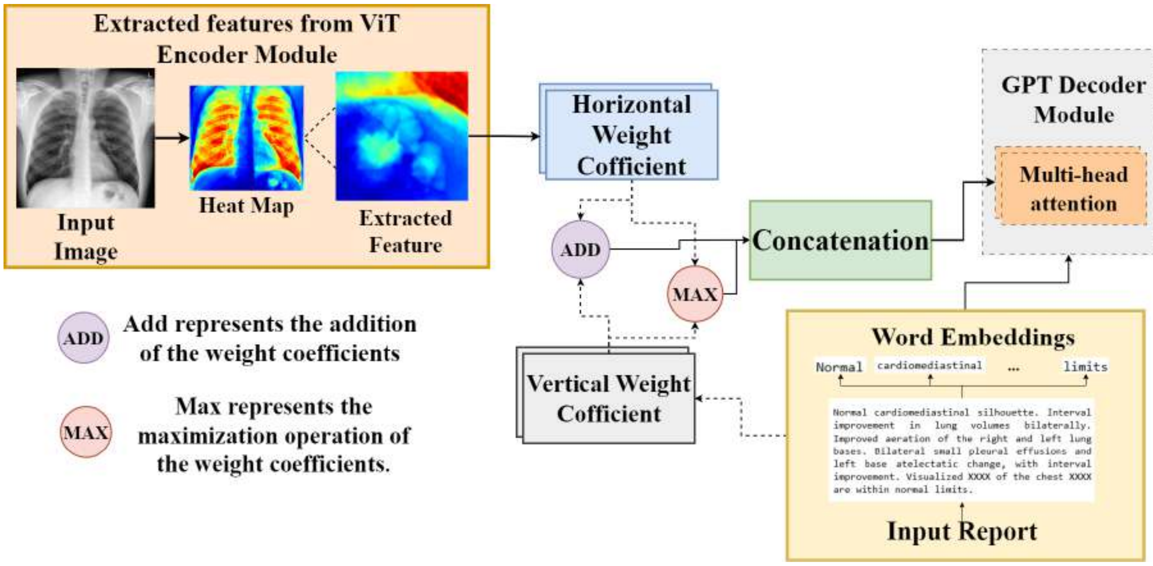
$$\text{MultiHead}(Q,K,V) = \text{Concat}\left(\text{Head}_1,\ldots,\text{Head}_h\right)W^O \tag{5.7}$$

$$\text{Head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V) \tag{5.8}$$

where, Q, K, and V represent the queries, keys, and values matrices, respectively. d_k is the dimension of the keys, W_i^Q, W_i^K, and W_i^V are projection matrices for each head and W^O is the output projection matrix. The multi-head attention module further generates heat maps to offer insights into the attention mechanism of the model, as illustrated in Figure 5-3. These heat maps aid in identifying the specific areas of the input image that impact key description elements. Additionally, the cross-model attention mechanism is used to transmit these extracted features to the GPT-4 decoder module for final mapping.

5.3.3 Cross-model attention mechanism

Through cross-modal attention, which dynamically focuses on pertinent features from each modality, the proposed model is able to match visual details with textual descriptions. In image description generation, the model creates sentences by focusing on various aspects of an image and lining them up with the context that the text provides. Cross-model attention mechanism can be represented by the following Figure 5-4. The alignment enhances the model's capacity to generate precise and contextually appropriate descriptions, eventually leading to a richer understanding and more coherent outputs. The cross-modal attention mechanism is defined as Equation (5.9):


Extracted features from ViT Encoder Module
Input Image
Heat Map
Extracted Feature
Horizontal Weight Cofficient
ADD
MAX
Concatenation
GPT Decoder Module
Multi-head attention
Vertical Weight Cofficient
Word Embeddings
Normal
cardiomediastinal
...
limits
Input Report
ADD
Add represents the addition of the weight coefficients
MAX
Max represents the maximization operation of the weight coefficients.

Normal cardiomediastinal silhouette. Interval improvement in lung volumes bilaterally. Improved aeration of the right and left lung bases. Bilateral small pleural effusions and left base atelectatic change, with interval improvement. Visualized XXXX of the chest XXXX are within normal limits.

Figure 5-4: Cross-Model Attention Mechanism.

$$\mathrm{CrossAttention}(Q_{\mathrm{t}}, K_{\mathrm{i}}, V_{\mathrm{i}}) = \mathrm{Softmax}\left(\frac{Q_{\mathrm{t}} K_{\mathrm{i}}^{T}}{\sqrt{d_{\mathrm{t}}}}\right) V_{\mathrm{i}} \tag{5.9}$$

where, Q_{t} represents text queries, K_{i} and V_{i} are the keys and values from the image features and d_{t} is the dimension of the text queries.

5.3.4 Decoder Module of the Proposed Architecture

The primary architecture of the decoder is based on the GPT-4 transformer model, which is specifically designed for context-based tasks. The structure consists of layers of feed-forward and multiple self-attention neural networks. The diagram in Figure 5-5 illustrates the structural framework of GPT-4.

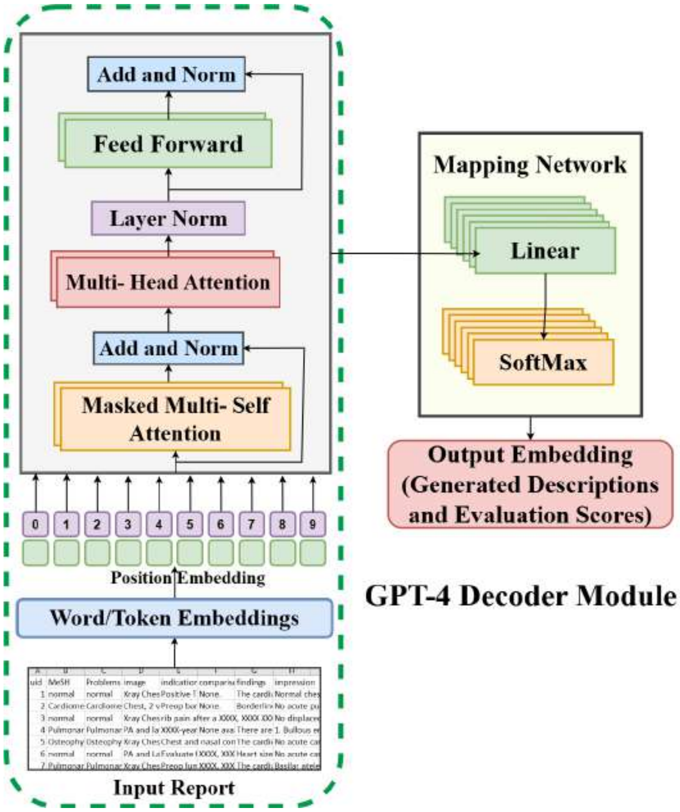


Figure 5-5: GPT-4 Decoder Module.

Using self-attention mechanisms, every layer progressively enhances input embedding, enabling the model to discern distant relationships and contextual information. The self-attention within the GPT-4 decoder is given by Equation (5.10):

$$\text{SelfAttention}(Q_{\text{t}}, K_{\text{t}}, V_{\text{t}}) = \text{Softmax}\left(\frac{Q_{\text{t}} K_{\text{t}}^{T}}{\sqrt{d_{\text{t}}}}\right) V_{\text{t}} \tag{5.10}$$

where, Q_{t}, K_{t}, and V_{t} represent the text queries, keys, and values respectively and d_{t} is the dimension of the text queries. To create the next word in the sequence, the output layer predicts the probability distribution across the vocabulary using word/-token embedding, representing words as vectors. Due to this methodology, GPT-4 can generate cohesive and contextually suitable information by extensively working on more substantial sentences. The final textual output is obtained by passing the combined attention outputs through a feed-forward (FF) network and then generating Logits for each token in the vocabulary as represented in Equation (5.11):

$$\text{Logits} = \text{FF}\left(\text{CrossAttention} + \text{SelfAttention}\right) \tag{5.11}$$

Linear and Softmax layers in network mapping are essential for producing the end result and predicting evaluation scores when generating image descriptions. The model can represent intricate relationships with them appropriately. Equation (5.12) represents the expression denoted as $\text{Linear}(W, X)$.

$$\text{Linear}(W, X) = WX + b \tag{5.12}$$

The weight matrix is symbolized by W, the input vector is indicated by X, and the bias vector is designated by b. Next, the updated characteristics are normalized into a probability distribution throughout the lexicon of potential words or tokens using the Softmax layer, as represented by Equation 5.13.

$$\text{Softmax}(\mathbf{z})_i = \frac{e^{z_i}}{\sum_j e^{z_j}}, \tag{5.13}$$

where, the input vector is denoted as z. The accuracy of these predictions is evaluated by comparing them to the descriptions in the reference file. The evaluation metrics used are ROUGE-L, CIDEr, METEOR, and BLEU 1-4, which assess the degree of similarities between the predicted and actual descriptions. These metrics enable the evaluation of the model's ability to effectively integrate the context and semantics of the images into the generated descriptions.

The following describes Algorithm 5.1 for the suggested model along with its score evaluation:

Algorithm 5.1 Algorithm for the proposed model, CrossViT-GPT4.

```
Input: Image $I$
Output: Generated Description $\text{Output}_{\text{text}}$, Evaluation Scores
Step 1: Data Pre-processing
$P \leftarrow$ Partition image $I$ into patches
$E_{\text{patch}} \leftarrow$ Flatten and linearly project each patch
$E_{\text{patch}} \leftarrow$ Apply positional encoding
Step 2: Vision Transformer Encoding
Visual_features $\leftarrow$ ViT Encoder($E_{\text{patch}}$)
Enhanced_features $\leftarrow$ MultiHeadAttention($Q, K, V$)
Step 3: Cross-Modal Attention
Cross_modal_features $\leftarrow$ CrossAttention($Q_{\text{t}}, K_{\text{i}}, V_{\text{i}}$)
Step 4: GPT-4 Decoding
for each time step $t$ in decoding do
    Text_features $\leftarrow$ SelfAttention($Q_{\text{t}}, K_{\text{t}}, V_{\text{t}}$)
    Logits $\leftarrow$ Generate next token logits
    $\text{Output}_{\text{text}} \leftarrow$ Softmax(Logits)
end for
Step 5: Evaluation
Evaluation_scores $\leftarrow$ Calculate BLEU, CIDEr, METEOR, and ROUGE-L
Output: $\text{Output}_{\text{text}}$, Evaluation_scores
```

5.4 Experimental Analysis

The experiments were performed using Google Colab Pro+, which provides high-performance resources, including 52 GB of RAM, an NVIDIA A100 GPU with 40 GB of VRAM, and a virtual CPU equivalent to an Intel Xeon processor. The framework was implemented using Keras and TensorFlow 2.12. The ViT model has 86.7 million parameters and GPT-4 has 1.76 trillion. Input images are (batch_size, 224, 224, 3), and output shapes vary by layer. The developed framework was evaluated on two benchmark datasets: the Indiana University Chest X-Ray dataset (IU X-Ray) and the NIH Chest X-ray dataset. The IU X-ray dataset consists of 3,955 XML radiologist reports and 7,471 PNG images while the NIH dataset consists of 112,120 frontal X-rays with annotations for 13 thoracic illnesses across 30,805 individuals.

5.4.1 Implementation Details:

To improve medical image captioning, the CrossViT-GPT4 architecture combines a GPT-4 decoder and a ViT encoder with cross-modal attention. While cross-modal attention aligns textual and visual input for improved contextual understanding, ViT removes spatial and semantic elements. Reports produced by GPT-4 are logical and clinically significant. Using the Adam optimiser, a batch size of 32, and categorical cross-entropy loss, the model is trained for 75 epochs on the IU and NIH Chest X-ray datasets. For medical terminology, a domain method of adaptation refines the language model. BLEU (1-4), METEOR, ROUGE-L, and CIDEr assess performance, guaranteeing linguistic correctness and clinical significance. Table 5.1describes the CrossViT-GPT4 framework's components, emphasizing the output dimensions and number of parameters for each layer.

Table 5.1: Architectural Details for the Proposed Model

Layer Type	**Output Shape**	**Parameters**
ViT Encoder with Cross-Attention	(batch_size, d_model)	3M
GPT-4 Decoder	(batch_size, seq_len)	125M
Cross-Modal Attention Layer	(batch_size, joint_dim)	5M
Fully Connected Layer	(batch_size, 512)	262,144
Pooling Layer	(batch_size, 256)	0
Convolutional Layer	(batch_size, 128)	280,960
Total Number of Parameters	-	**133.54M**

Table 5.2 describes the evaluation metrics and hyperparameters used to train and evaluate the model. By ensuring that CrossViT-GPT4 efficiently captures the fine-grained medical data required for precise and clinically meaningful image descriptions, this implementation helps radiologists diagnose and arrange treatments more successfully.

5.4.2 Ablation Study:

The suggested model's image description components were clarified using ablation analysis. This study investigates the capabilities of the proposed architecture, CrossViT-

Table 5.2: Hyperparameters and Evaluation Metrics for the CrossViT-GPT4 Framework

Hyperparameter	**Value/Description**
Model Name	CrossViT-GPT4
Encoder	Vision Transformer (ViT) with Cross-Modal Attention
Decoder	GPT-4
Training Datasets	IU Chest X-ray, NIH Chest X-ray
Training Epochs	75
Batch Size	32
Optimizer	Adam
Loss Function	Categorical Cross-Entropy Loss
Evaluation Metrics	BLEU (1-4), METEOR, ROUGE-L, CIDEr

GPT4. In Table 5.3, we compare the performance of various models on two major chest X-ray datasets: the IU Chest X-Ray dataset and the NIH Chest X-Ray dataset. The results presented in the following table summarize the performance of each model across these metrics. The table presents the performance comparison of image description generation models on the IU and NIH Chest X-ray datasets using BLEU scores (B-1 to B-4), CIDEr (C), METEOR (M), and ROUGE-L (R-L). The proposed CrossViT-GPT4 model outperforms all other models across all metrics for both datasets. On the IU Chest X-ray dataset, it achieves BLEU-1 of 0.854, BLEU-2 of 0.817, BLEU-3 of 0.804, and BLEU-4 of 0.785, along with a CIDEr score of 0.883, METEOR of 0.759, and ROUGE-L of 0.712. Similarly, on the NIH Chest X-ray dataset, it attains BLEU-1 of 0.825, BLEU-2 of 0.806, BLEU-3 of 0.795, and BLEU-4 of 0.772, with a CIDEr of 0.857, METEOR of 0.726, and ROUGE-L of 0.705. These results highlight the superior performance of CrossViT-GPT4 in generating high-quality medical image descriptions.

Table 5.3: Evaluation of Models on the IU and NIH Chest X-Ray Datasets

Dataset	**Model**	**B-1**	**B-2**	**B-3**	**B-4**	**CIDEr**	**METEOR**	**ROUGE-L**
IU Chest X-Ray	ViT without GPT-4	0.398	0.373	0.354	0.343	0.658	0.286	0.316
	ViT + Self Attention	0.283	0.258	0.236	0.205	0.574	0.194	0.188
	ViT + Self Attention + LSTM	0.601	0.572	0.554	0.498	0.587	0.436	0.342
	ViT + GPT-2	0.704	0.663	0.634	0.626	0.792	0.564	0.474
	ViT + GPT-4	0.782	0.764	0.712	0.701	0.762	0.735	0.695
	CrossViT-GPT4 (Proposed)	**0.854**	**0.817**	**0.804**	**0.785**	**0.883**	**0.759**	**0.712**
NIH Chest X-Ray	ViT without GPT-4	0.383	0.363	0.334	0.312	0.578	0.265	0.287
	ViT + Self Attention	0.284	0.268	0.253	0.248	0.534	0.187	0.176
	ViT + Self Attention + LSTM	0.438	0.425	0.409	0.387	0.589	0.436	0.386
	ViT + GPT-2	0.687	0.658	0.644	0.612	0.763	0.582	0.604
	ViT + GPT-4	0.789	0.767	0.758	0.716	0.825	0.714	0.674
	CrossViT-GPT4 (Proposed)	**0.825**	**0.806**	**0.795**	**0.772**	**0.857**	**0.726**	**0.705**

5.4.3 Results and Analysis:

The suggested model, CrossViT-GPT4, provides an excellent approach for illustrating images in the IU and NIH chest X-ray datasets. It combines the ViT encoder and the GPT-4 decoder modules. The results for the chest X-ray datasets from IU and NIH are shown in able 5.4 and 5.5, respectively.

Table 5.4: Quantitative Results Obtained on IU Chest X-Ray Dataset

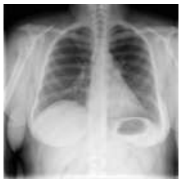
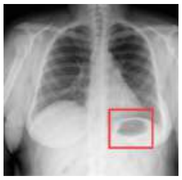
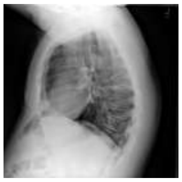
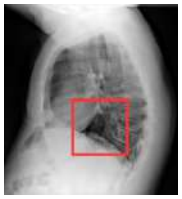
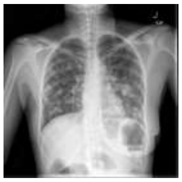
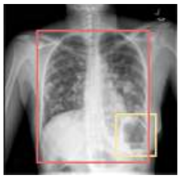
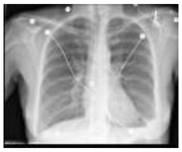
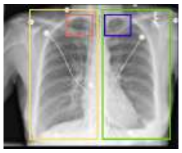
Test Image	**Ground Truth Report**	**Disease Prediction**	**Predicted Report**
	The heart is normal in size. The mediastinal contours are within normal limits. There is mild prominence of the superior mediastinum which is somewhat lucent and reflects mediastinal and vascular structures. No focal consolidation is seen. There is no pleural effusion.		Mediastinal contours within normal limits. Mild mediastinum somewhat lucent.
	Lateral view of the chest shows an unchanged cardiomediastinal silhouette. The cardiac silhouette remains moderately enlarged, exaggerated by epicardial fat pads. Interstitium is prominent. No focal airspace consolidation or pleural effusion. There is spine spondylosis.		Unchanged cardiomediastinal silhouette. Interstitium prominent. Spine spondylosis.

Table 5.5: Quantitative Results Obtained on NIH Chest X-Ray Dataset

Test Image	**Ground Truth Report**	**Disease Prediction**	**Predicted Report**
	1. The heart is normal size. 2. The mediastinum is unremarkable. 3. There is no pleural effusion, pneumothorax, or focal airspace disease. 4. There is stable irregularity of the posterior left 6th rib which represents an old fracture.		No pleural effusion, pneumothorax, or focal airspace disease.
	1. The heart size and cardiomediastinal silhouette are normal. 2. There is hyperexpansion of the lungs with flattening of the hemidiaphragms. 3. There is no focal airspace opacity, pleural effusion, or pneumothorax. 4. There are multilevel degenerative changes of the thoracic spine.		Cardiovascular silhouette normal. No focal airspace opacity, pleural effusion, or pneumothorax.

The tables provide a quantitative comparison between the actual data and the reports generated by the model for different test images. The results evaluate the model's capacity to forecast diseases and produce descriptions that showcase the model's precision by utilizing X-ray images for both datasets.

5.4.4 Comparison With Other state-of-the-art Methods:

Compared to the most advanced methodologies currently available, the proposed method demonstrates a substantial enhancement in disease prediction accuracy and report production for X-ray images. The method, CrossViT-GPT4, effectively performs on both the IU and NIH chest X-ray datasets by utilizing a fusion model that incorporates the ViT encoder and GPT-4.0 decoder.

The study presented in Table 5.6 thoroughly analyzes various models used to generate image descriptions on the IU X-ray and NIH datasets. The table compares various models for generating image descriptions on the IU Chest X-Ray and NIH Chest X-Ray datasets, evaluating their performance using BLEU, METEOR, CIDEr, and Rouge-L metrics. On the IU Chest X-Ray dataset, R2Gen [110] demonstrates intermediate performance, while R2Gen + ChexNet [111] improves BLEU and METEOR scores but slightly reduces Rouge-L. Both XPRONET [112] and Knowledge-injected U-Transformer [105] achieve a BLEU score of 0.525, with the latter excelling in METEOR and Rouge-L. Contrastive Attention exhibits moderate performance, whereas AERMNet [114] provides balanced results across metrics. The proposed CrossViT-GPT4 model outperforms all these approaches, achieving significantly higher scores

Table 5.6: Comparative Analysis on IU and NIH Chest X-Ray Datasets

Model	**BLEU**	**METEOR**	**CIDEr**	**Rouge-L**
IU Chest X-Ray Dataset				
R2Gen [110]	0.470	0.187	-	0.371
R2Gen + ChexNet [111]	0.508	0.222	-	0.365
Cross-modal PROtotype driven NETwork (XPRONET) [112]	0.525	0.220	-	0.411
Contrastive attention [105]	0.492	0.193	-	0.381
Knowledge-injected U-Transformer [113]	0.525	0.242	-	0.409
AERMNet [114]	0.486	0.219	0.560	0.398
CrossViT-GPT4 (Proposed)	**0.854**	**0.759**	**0.883**	**0.712**
NIH Chest X-Ray Dataset				
Semantic Attention [115]	0.467	0.192	0.560	0.204
Co-Attention [116]	0.756	0.597	0.755	0.675
Clinical-BERT [117]	0.383	0.144	-	0.275
ChestBioX-Gen [118]	0.668	0.189	0.416	0.674
CrossViT-GPT4 (Proposed)	**0.825**	**0.726**	**0.857**	**0.705**

across all metrics: BLEU (0.854), METEOR (0.759), CIDEr (0.883), and Rouge-L (0.712). On the NIH Chest X-Ray dataset, the CrossViT-GPT4 model similarly leads with the highest scores across all metrics: BLEU (0.825), METEOR (0.726), CIDEr (0.857), and Rouge-L (0.705). In comparison, Semantic Attention [115], Co-Attention [116], ChestBioX-Gen [118], and Clinical-BERT [117] exhibit less balanced or lower performance. These results underscore the superior capability of the proposed model in generating accurate and semantically rich image descriptions across both datasets.

5.5 Conclusion

In this chapter, combining the Vision Transformer encoder module (ViT) with cross-model attention and the Generative Pre-trained Transformers 4 (GPT 4.0) decoder module results in an efficient framework, CrossViT-GPT4, for producing visual descriptions. The model uses vision-based feature extraction and language modeling to analyze and characterize complex medical images thoroughly. When textual and visual components are combined, chest X-ray pathology reports can be described and explained more precisely. Transformer-based designs are scalable and adaptable, promoting medical image analysis research and innovation. Their adaptability enables them to efficiently manage diverse datasets and tasks. By enhancing medical image processing, the suggested approach may increase diagnostic precision and assist clinical decision-making platforms. Poor image quality can seriously affect model accuracy by restricting the ability to extract features and generate accurate descriptions.

Chapter 6

Image Description Models for Multimedia Application: Video-based Image Description Generation

Lakshita Agarwal and Bindu Verma. "Towards Explainable AI: Multi-Modal Transformer for Video-based Image Description Generation". Archived at: http://arxiv.org/abs/2504.16788. Communicated in: Signal, Image and Video Processing (Springer) SCIE Indexed, Impact Factor- 2.0. ***(1st Major Revision)***

6.1 Introduction

In this chapter, we introduce a method for understanding and analyzing video actions, which is essential for producing insightful and contextualized descriptions, particularly in video-based applications like intelligent monitoring and autonomous systems. This work presents a novel framework for generating natural language descriptions from images and videos by integrating textual and visual modalities. The proposed architecture utilizes ResNet50 to extract visual features from video frames sourced from the Microsoft Research Video Description Corpus (MSVD), the Berkeley DeepDrive eXplanation (BDD-X) dataset, and filtered image-caption pairs from Flickr8k.

The extracted visual features are transformed into patch embeddings and processed through an encoder-decoder model based on transformers and GPT-2. To ensure high-quality description generation, the system employs multi-head self-attention and cross-attention mechanisms for aligning textual and visual representations. The model's effectiveness is validated through performance evaluation using BLEU (1-4), CIDEr, METEOR, and ROUGE-L.

The major contributions of the proposed work are summarized below:

- To provide natural language descriptions of video sequences, the following work proposes a transformer-based architecture that integrates a GPT-2-based language model with the visual features retrieved by ResNet50.
- The method combines video datasets from various sources (MSVD and BDD-X), allowing the model to generalize across domains. High-quality training samples are guaranteed via sophisticated preprocessing.
- Gradient accumulation and mixed precision training are used to optimise the model, increasing computing efficiency without sacrificing output quality.
- Fluency, contextual relevance, and coherence are ensured by thoroughly evaluating the generated descriptions using BLEU (1–4), CIDEr, METEOR, and ROUGE-L standards.
- Although this method is extremely beneficial for intelligent transportation and autonomous driving, it can also be used in other fields, including robotics, assistive technology, and surveillance.

6.2 Literature Survey

In computer vision and natural language processing, video-based description creation is an essential task that allows systems to produce textual summaries of visual content. The creation of video descriptions must consider temporal relationships, object interactions, and dynamic scene changes across frames, in contrast to static image captioning. Early video description generation approaches included recurrent neural networks (RNNs) or long short-term memory (LSTM) networks for text synthesis and convolutional neural networks (CNNs) for feature extraction.

A Sequence-to-Sequence (Seq2Seq) model trained on MSVD [119] was presented by Venugopalan et al. [120] to produce video descriptions. However, because of the limits of LSTMs, these models have trouble handling contextual inconsistencies and long-range dependencies. Later, to make generated descriptions more relevant, attention techniques were added [121]. Transformer-based models have led to a considerable improvement in video captioning systems. End-to-end dense Video Captioning was proposed by Zhou et al. [122], who used spatiotemporal attention mechanisms to grasp multiple frames. Compared to conventional RNN-based models, MART (Memory-Augmented Recurrent Transformer), which was introduced by Lei et al. [123], performs more effectively at capturing long-term video dependencies. More recently, multimodal pretraining for video-language understanding has been investigated by VideoBERT [124] and ClipBERT [125]. These methods achieve state-of-the-art outcomes on datasets such as MSVD [119] and MSR-VTT [126] by aligning textual and visual representations. Datasets such as Berkeley DeepDrive eXplanation (BDD-X) [127] have been useful for producing explainable AI-driven annotations for vehicle-based video description generation. Studies like Shoman et al. [128] integrated explainability processes into vision-language models and concentrated on justification-based video descriptions for autonomous driving. An attention-guided transformer was developed by Cui et al. [129] to improve contextual reasoning in captioning models based on BDD-X.

Recent advancements in multimodal learning combine text, audio, and visual data. Hori et al. [130] introduced an audio-visual attention model using speech and scene context, while Yu et al. [29] enhanced captioning on datasets like Flickr8k and MSVD using spatiotemporal object graphs. These methods improve contextual awareness, particularly in dynamic scenarios like vehicle interactions. Despite progress, generating accurate, human-like video descriptions remains challenging. The proposed approach addresses this by combining a GPT-2 encoder-decoder with ResNet50-based feature extraction, outperforming prior methods on BLEU (1–4), CIDEr, METEOR, and ROUGE-L. It refines multimodal embeddings for context-aware captions and boosts efficiency with gradient accumulation and mixed precision training. By bridging explainable vehicular (BDD-X) and general (MSVD) video captioning, the model delivers interpretable, context-rich descriptions suitable for real-world use.

6.3 Proposed Architecture

The proposed study introduces a context-sensitive and transformer-based reasoning framework to generate logical and relevant video descriptions. Using the BDD-X and MSVD datasets, the system learns from both action-justified driving videos and diverse scene representations. It employs an optimised GPT-2 encoder-decoder model with multihead attention and ResNet50-based visual feature extraction to enhance contextual understanding. Training is guided by action-justification pairs from BDD-X and captioned sequences from MSVD, promoting coherent human-like output. Gradient accumulation and mixed precision training boost computational efficiency without compromising accuracy. Evaluations using BLEU (1–4), CIDEr, METEOR, and ROUGE-L show superior performance over traditional methods. The basic architecture of the proposed framework is shown in Figure 6-1, which further demonstrates the pipeline for data pre-processing, model training, and description generation.

6.3.1 Data Pre-processing

The input data includes video frames from the BDD-X and MSVD datasets. Video frames are extracted, pertinent images are filtered, and then the frames are transformed into structured input for the model as part of the pre-processing pipeline. ResNet50, which converts spatial information from images into high-dimensional feature vectors, is used for feature extraction. After that, a linear projection and position encoding method are used to patch and embed these feature vectors. The GPT-2 encoder-decoder module then receives the processed embeddings and aligns the word/token embeddings with the appropriate visual features. To further improve

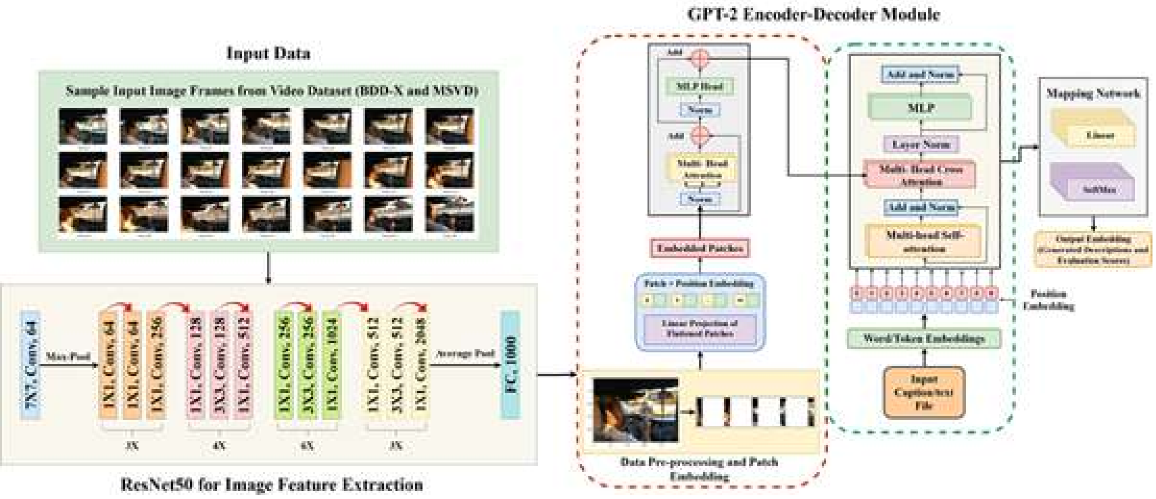


Figure 6-1: Framework for the Proposed Model (ResNet50-GPT2): The system incorporates ResNet50 for image feature extraction and a GPT-2 encoder-decoder model to generate context-aware video-based image descriptions.

model resilience, tokenization, data augmentation, and train-test splits are used.

6.3.2 Model Architecture

The proposed method combines a GPT-2 encoder-decoder model to generate context-aware descriptions of vehicle actions with ResNet50 to extract image features. The model is fine-tuned on the BDD-X and MSVD datasets to become optimal in dynamic scenarios for generating meaningful descriptions. The training process employs gradient accumulation and mixed precision training to achieve optimal computing efficiency.

ResNet50 acts as a feature extractor by passing each input frame $x \in \mathbb{R}^{3\times H\times W}$ through a series of convolutional, batch normalization, ReLU, and residual layers. The output feature vector f is computed as:

$$f = \text{ResNet50}(x) = \text{AvgPool}(F_{\text{res}}(x)) \tag{6.1}$$

where, F_{res} denotes the output of the final convolutional block, and AvgPool is the global average pooling operation applied to extract the final feature representation $f \in \mathbb{R}^{2048}$.

Using position embeddings $P \in \mathbb{R}^{N\times d}$ and a learnable linear projection $W_p \in \mathbb{R}^{2048\times d}$, the visual characteristics retrieved by ResNet50 are transformed into embedded visual tokens:

$$E_v = fW_p + P \tag{6.2}$$

The GPT-2 model uses these embedded patches as input tokens to produce descriptions that are logical and sensitive to context. The model utilizes self-attention mechanisms to enhance contextual understanding, formulated as:

$$\text{Self-Attention}(Q, K, V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{6.3}$$

where d_k is the key vector dimension, and Q, K, V represent the query, key, and value matrices.

To maintain the semantic relevance of the generated descriptions to video-based activities, the multi-head cross-attention method aligns image embeddings with textual descriptions. The decoder uses attention-weighted contextual embeddings to improve the textual output:

$$\text{Decoder-Attention}(Q, K, V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{6.4}$$

The final word prediction probability is computed as:

$$y_t = \text{Softmax}(W_o h_t + b_o) \tag{6.5}$$

where W_o and b_o are the output weight matrix and bias, respectively, and h_t represents the hidden state at time step t.

To enhance generalization, the model employs a combined loss function incorporating cross-entropy loss and L2 regularization:

$$\mathcal{L} = -\sum_{t=1}^{T} \log P(y_t \mid y_{<t}) + \lambda\|\theta\|^2 \tag{6.6}$$

where λ is the regularization parameter and $P(y_t \mid y_{<t})$ denotes the probability of predicting y_t given prior words.

In BLEU-4, CIDEr, METEOR, and ROUGE-L metrics, the proposed framework outperforms traditional methods, ensuring that the generated descriptions are sensible, logical, and context-aware. These metrics are utilized to analyze the contextual validity and linguistic quality of the descriptions generated by the GPT-2 model for unseen videos upon training. The method achieves performance improvement over traditional methods by contrasting the generated descriptions with ground-truth annotations. Ultimately, the framework generates natural language narratives that effectively describe dynamic visual scenes, promoting explainability in a range of applications.

The following summarizes the Algorithm 6.1 for the proposed model and its evaluation:

Algorithm 6.1 Context-Aware Description Generation

```
Input: Video frames (BDD-X, MSVD), image-caption pairs (Flickr8k)
Output: Generated descriptions, Evaluation Scores
1. Preprocessing and Embedding
Extract and embed image frames using patch encoding
Tokenize and format textual descriptions
2. Dataset Split
Create balanced train-test sets with captions and action-justification pairs
3. Feature Extraction
Visual_features ← ResNet50($I$)
Project features and tokenize text using GPT-2
4. Model Training
Initialize GPT-2 with multi-head attention and cross-modal fusion
Train with AdamW, gradient accumulation, and mixed precision
5. Inference and Evaluation
for each test sample do
    Generate and decode descriptions
end for
Compute BLEU-4, CIDEr, METEOR, ROUGE-L
6. Output
Present evaluation scores and visualize results
```

6.4 Experimental Analysis

The proposed model was implemented in PyTorch and evaluated on Google Colab Pro+ using 52 GB RAM, an NVIDIA A100 GPU (40 GB VRAM), and a virtual Intel Xeon CPU. It generates context-aware descriptions using a GPT-2 encoder-decoder (approximately 66.7M parameters) and ResNet50 for visual feature extraction. Input frames were resized to `(batch_size, 224, 224, 3)`. Training was conducted over 75 epochs with the AdamW optimizer (learning rate 1×10^{-5}, weight decay 0.01), using a batch size of 32 and gradient accumulation. Mixed precision training and gradient clipping were applied to improve computational efficiency and training stability. The model was trained using the MSVD dataset[1], which contains short descriptive captions for diverse video clips, and the BDD-X dataset[2], which includes over 26K annotated actions across 8.4M frames and 6,970 videos. During preprocessing, linguistic descriptions were tokenized and structured into action-justification pairs, while image frames were embedded via ResNet50 and projected linearly into patch embeddings. These embeddings were then fed into the multi-head self-attention module of the GPT-2 encoder-decoder to align image and text contexts effectively. Performance was evaluated on a 20% test split using BLEU-4, CIDEr, METEOR, and ROUGE-L. The model consistently outperformed conventional captioning techniques, demonstrating its ability to generate logical, contextually accurate descriptions across diverse visual inputs. The architecture's use of self- and cross-attention mechanisms ensured semantic accuracy and narrative coherence, reinforcing its applicability in real-world visual captioning tasks. The following Table 6.1 depicts the overall architectural structure of the proposed framework.

6.4.1 Ablation Study

Utilising the BDD-X and the MSVD datasets, an ablation research was carried out to evaluate the contribution of various components in the suggested model. The re-

---

[1] `https://www.kaggle.com/datasets/vtrnanh/msvd-dataset-corpus`

[2] `https://github.com/JinkyuKimUCB/BDD-X-dataset`

Table 6.1: Experimental Setup and Performance Metrics

Component	**Details**
Framework Used	PyTorch
Image Feature Extractor	ResNet-50
Model	GPT-2 Transformer-based Encoder-Decoder
Input Size	(224, 224, 3)
Batch size	32
Number of Epochs	75
Learning Rate	1×10^{-5}
Optimizer	AdamW
Training Strategy	Gradient Accumulation, Mixed-Precision Training
Datasets Used	BDD-X, MSVD, Filtered Flickr8k
Evaluation Metrics	BLEU 1-4, CIDEr, METEOR, ROUGE-L

search methodically eliminated or altered important architectural components, such as multi-head attention, gradient accumulation and mixed precision training optimisations, and visual feature extraction (ResNet-50), in order to assess each component's effect on model performance. Model modifications were compared using the BLEU-1 to BLEU-4, CIDEr, METEOR, and ROUGE-L metrics, which provided a thorough assessment of the effects of each component on the quality of description production. As can be seen from Table 6.2, the results demonstrate the importance of each element in producing textual descriptions for video frames that are both logical and contextually rich. The model's performance on the MSVD and BDD-X datasets

Table 6.2: Ablation Study: Comparing the Impact of Model Components on Different Datasets

Dataset	**Model Variant**	**BLEU-1**	**BLEU-2**	**BLEU-3**	**BLEU-4**	**CIDEr**	**METEOR**	**ROUGE-L**
BDD-X	GPT-2 without ResNet50	0.693	0.668	0.638	0.601	0.921	0.234	0.640
	GPT-2 + Single-Head Attention	0.731	0.705	0.673	0.634	1.032	0.257	0.672
	GPT-2 + Without Gradient Accumulation	0.785	0.758	0.723	0.681	1.142	0.278	0.710
	ResNet50 + GPT-2 (Proposed)	**0.860**	**0.835**	**0.800**	**0.755**	**1.235**	**0.312**	**0.782**
MSVD	GPT-2 without ResNet50	0.710	0.685	0.654	0.612	0.980	0.242	0.651
	GPT-2 + Single-Head Attention	0.742	0.715	0.683	0.641	1.105	0.265	0.685
	GPT-2 + Without Gradient Accumulation	0.798	0.772	0.740	0.698	1.231	0.293	0.724
	ResNet50 + GPT-2 (Proposed)	**0.880**	**0.855**	**0.823**	**0.778**	**1.315**	**0.329**	**0.795**

emphasizes the critical role of each architectural component. The most substantial performance drop occurred when ResNet-50 was removed, resulting in BLEU-4 and CIDEr declines of 15.4% and 25.4% on BDD-X, and 21.3% and 25.5% on MSVD, respectively—underscoring the importance of strong visual feature extraction. Replacing multi-head attention with a single-head variant led to moderate performance degradation, with a 16.0% drop in CIDEr on the MSVD dataset, highlighting its contribution to capturing fine-grained cross-modal interactions. Omitting gradient accumulation reduced training stability and effectiveness, leading to average drops of 9.8% in BLEU-4 and 7.5% in CIDEr across the two datasets. Figure 6-2 illustrates these results. The proposed framework obtained the best scores on all evaluation measures, outperforming the ablated versions on all the datasets consistently. The improvements are especially clear in MSVD and BDD-X, where temporal dependency modelling and interpreting intricate visual scenes are critical. This better performance is a result of the synergy among multiple important elements: ResNet50 for feature extraction of high-level spatial features from images, GPT-2 as a strong language model with the ability to produce fluent and coherent text sequences, multi-head attention for modelling fine-grained alignments between visual and linguistic modalities, and gradient accumulation with mixed-precision training for efficient and stable optimization. The use of GPT-2, specifically, adds to the model's capacity to generate contextually accurate and linguistically advanced descriptions. Collectively, these fea-

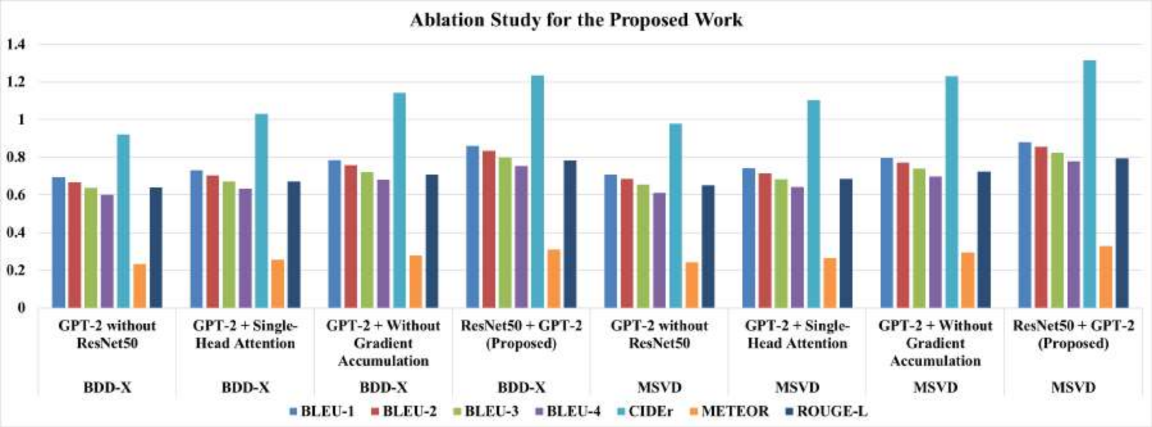


Figure 6-2: Demonstration of Ablation Study for the Proposed Work

tures make the proposed framework capable of generating high-quality, context-aware descriptions, with each module contributing significantly to the overall performance improvement.

6.4.2 Results and Analysis:

To assess the efficacy of the suggested architecture, we carried out a comprehensive performance analysis on two benchmark datasets: MSVD and BDD-X. The suggested framework was tested against many model ablations, such as excluding gradient accumulation, using single-head attention, and removing ResNet-50. The evaluation's findings are compiled in Table 6.3. The outcomes shown in the table demonstrate how well our suggested methodology performs across datasets, consistently achieving top scores in all evaluation metrics. While high BLEU scores reflect fluency and grammatical accuracy, the superior performance in CIDEr and METEOR highlights strong semantic alignment and high-quality descriptions. With a BLEU-4 score of 0.778 and a CIDEr score of 1.315, the MSVD dataset showcases the model's strength in producing diverse and precise captions for short video content. Similarly, the BDD-X dataset results—0.755 BLEU-4 and 1.235 CIDEr—illustrate the model's robustness in capturing intricate driving scenes and generating contextually appropriate, descriptive narratives. Consistently high ROUGE-L scores (0.795 for MSVD and 0.782 for BDD-X) further validate the framework's ability to generate outputs that closely align with human references. Notably, elevated CIDEr values emphasize the model's effectiveness in capturing contextual and domain-specific nuances, while higher METEOR scores reflect improved alignment and paraphrasing capabilities. Figure 6-3 demonstrates the graphical representation of the results obtained for the proposed framework. In general, the findings demonstrate that the suggested approach greatly

Table 6.3: Results and Analysis of the Proposed Framework

Dataset	**BLEU-1**	**BLEU-2**	**BLEU-3**	**BLEU-4**	**CIDEr**	**METEOR**	**ROUGE-L**
BDD-X	0.860	0.835	0.800	0.755	1.235	0.312	0.782
MSVD	0.880	0.855	0.823	0.778	1.315	0.329	0.795

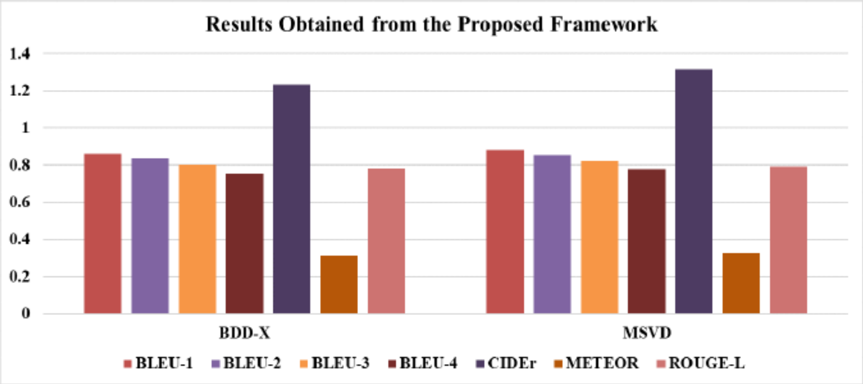


Figure 6-3: Graphical Representation of the Results Obtained

improves the ability of video-based descriptions, resulting in syntactically fluid, semantically rich, and contextually relevant descriptions.

6.4.3 Comparison With State-of-the-Art Methods:

We compare the performance of the suggested model with current state-of-the-art (SOTA) captioning methods on two benchmark datasets—BDD-X and MSVD—in order to assess its efficacy. A thorough evaluation employing several evaluation metrics, such as BLEU-1 to BLEU-4, CIDEr, METEOR, and ROUGE-L, is given by the results in Table 6.4. On both datasets, the proposed framework significantly outperforms earlier approaches, proving its capacity to produce more semantically precise and contextually rich descriptions for behaviours involving vehicles. Specifically, the suggested model leads in terms of BLEU-4 and CIDEr scores, demonstrating enhanced description coherence and relevance. The table's comparative analysis demonstrates how the BDD-X and MSVD datasets have advanced regarding vehicle action comprehension and general video-based description generation. Previous models like OSCAR [132], ViT-GPT2 [131], X-LAN [53], and AoANet [100] showed consistent improvements in descriptive quality on the BDD-X dataset, with AoANet receiving the highest scores. However, with a BLEU-4 score of 0.755 and a CIDEr score of 1.235, the proposed model outperforms all of these approaches, significantly advancing over prior research. Similarly, on the MSVD dataset, transformer-based models such as M^2 Transformer [27] and GRU-EVE [134] performed well, with GRU-

Table 6.4: Comparison of State-of-the-Art Models for Vehicle Action Understanding

Dataset	**Model**	**BLEU-1**	**BLEU-2**	**BLEU-3**	**BLEU-4**	**CIDEr**	**METEOR**	**ROUGE-L**
BDD-X	GPT-2 (Baseline)	0.312	0.285	0.210	0.152	0.653	0.185	0.356
	X-LAN [53]	0.451	0.412	0.356	0.293	0.998	0.265	0.541
	ViT-GPT2 [131]	0.482	0.445	0.398	0.324	1.086	0.258	0.523
	OSCAR [132]	0.510	0.472	0.423	0.358	1.124	0.276	0.548
	AoANet [100]	0.532	0.495	0.442	0.372	1.189	0.289	0.573
	ResNet50-GPT2 (Proposed)	**0.860**	**0.835**	**0.800**	**0.755**	**1.235**	**0.312**	**0.782**
MSVD	GPT-2 (Baseline)	0.334	0.305	0.243	0.178	0.705	0.192	0.372
	Show, Attend and Tell [71]	0.480	0.443	0.395	0.304	0.943	0.252	0.520
	M^2 Transformer [27]	0.520	0.486	0.438	0.391	1.270	0.292	0.586
	Dense Video Captioning [133]	0.498	0.463	0.412	0.357	1.135	0.268	0.552
	GRU-EVE [134]	0.535	0.502	0.456	0.370	1.246	0.285	0.594
	ResNet50-GPT2 (Proposed)	**0.880**	**0.855**	**0.823**	**0.778**	**1.315**	**0.329**	**0.795**

EVE achieving a BLEU-4 score of 0.370 and CIDEr of 1.246. The proposed model demonstrates clear improvements, especially in BLEU-4 (0.778) and CIDEr (1.315), showcasing its ability to generate contextually rich and semantically aligned descriptions. Figure 6-4 depicts a graphical representation of SOTA methods. Overall, the suggested approach continuously outperforms comparison on all datasets, setting new benchmarks and proving its ability to capture fine-grained contextual information for interpreting video-based actions.

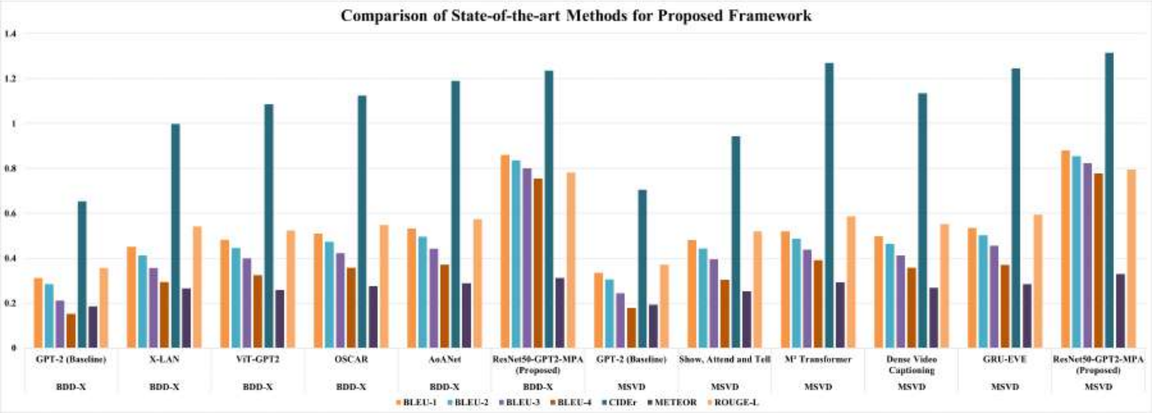


Figure 6-4: Graphical Representation of the SOTA Methods

6.5 Conclusion

The proposed approach successfully combines ResNet50 for feature extraction with an improved GPT-2 model, allowing for the production of contextually aware, high-quality descriptions of video footage from the BDD-X and MSVD datasets. Through the use of transformer-based language modelling, the method guarantees descriptions that are consistent with ground-truth annotations and logical. In comparison to traditional approaches, the combination of ResNet50 for visual feature extraction and GPT-2 for text generation, as well as dataset augmentation, gradient accumulation, and mixed precision training, improves training efficiency and performance. Although transformer-based architectures are effective when used for large datasets, the study emphasizes that domain-specific fine-tuning is necessary to produce accurate and comprehensible descriptions. By enhancing interpretability, transparency, and decision-making clarity, this research advances explainable AI and increases the adaptability and dependability of automated video comprehension for various applications.

Chapter 7

Conclusion

The thesis fully investigates automated image description generation, covering the vital issues of producing contextually precise and semantically rich descriptions in various domains. In order to achieve three main research objectives, we present and defend four advanced deep learning architectures in this thesis. The VGG16-SceneGraph-BiGRU model improves contextual comprehension by combining visual information with object relationships. In contrast, the Tri-FusionNet architecture attains state-of-the-art description generation using transformer-based fusion and dual attention. The thesis emphasizes the versatility of image description generation models in multimedia applications. The ViT-GPT4-based framework improves chest radiograph diagnosis with accurate descriptions, while the ResNet50 model with GPT 2-based model improves the transparency of context-aware description generation in video-based datasets. Experiments on different datasets verify the efficacy of the models, pushing image description generation further toward unifying visual comprehension and natural language synthesis. Finally, this work provides a great platform for future research on image description generation, indicating directions such as adding attention mechanisms, investigating transformer-based architectures, and developing multimodal learning strategies. These developments can further enhance real-time description generation, domain-based applications, video-based image description generation, and accessibility technology, making image description generation systems more impactful and useful.

7.1 Summary and Contribution of the Thesis

In this thesis, we proposed an extensive study on automatic image description generation, solving the major challenges of producing contextually correct and semantically rich descriptions. The study fills the gap between natural language processing and computer vision through the creation of sophisticated deep learning architectures specific to different applications. The primary contributions of the thesis are as follows:

- The incapacity of existing models to accurately depict object relationships within an image frequently results in the generation of descriptions that lack semantic significance. Therefore, we present the unified hybrid VGG16-SceneGraph-BiGRU system to address this problem. The proposed work combines a BiGRU network for sequential learning, scene graphs to model object interactions, and VGG16 for visual feature extraction. Extensive tests conducted on the MS COCO, Flickr8k, and Flickr30k datasets show that the suggested model dramatically improves contextual coherence and performs better than current methods. Standard metrics such as BLEU (1-4), CIDEr, METEOR, and ROUGE-L were used to thoroughly test the models.

- Although image captioning has advanced, traditional systems still have trouble with multimodal alignment, which results in descriptions that are not consistent. In order to address this problem, we present a deep learning-based framework known as Tri-FusionNet, which consists of a CLIP module for better vision-language alignment, a RoBERTa decoder for enhanced linguistic fluency, two attention mechanisms to improve multimodal learning, and a Vision Transformer (ViT) encoder for robust feature extraction. Significant improvements in prediction scores and caption quality have been established by experimental evaluations conducted on several benchmark datasets like MSCOCO, Flickr30k and Flickr8k. The models consistently showed improved performance and established new benchmarks in image description generation after being thoroughly tested using standard evaluation metrics such as BLEU (1-4), CIDEr, METEOR, and ROUGE-L.

- To help medical practitioners, medical image captioning needs to be accurate and comprehensible. However, domain-specific expertise is frequently absent from current models. In order to tackle this, we suggest a ViT-GPT4 architecture that uses a GPT-4-based decoder with cross-modal attention to produce medically appropriate descriptions and Vision Transformers (ViT) for high-level feature extraction. The model's efficacy in producing comprehensible and trustworthy radiology reports is demonstrated by its validation on the NIH Chest X-rays and Indiana University Chest X-ray datasets. The models consistently shown improved performance and established new benchmarks in image description generation after being thoroughly tested using standard evaluation metrics such as BLEU (1-4), CIDEr, METEOR, and ROUGE-L.

- Finally, natural language descriptions are needed to increase the interpretability and transparency of autonomous systems for video-based tasks. However, current models are not good at producing accurate and context-aware descriptions of changing scenes in rich environments. In order to tackle this issue, we created a ResNet50-based GPT-2 framework that combines textual and visual modalities, utilizing video frames from MSVD and BDD-X as well as filtered image-caption pairs from Flickr8k. Through aligning visual characteristics with textual accounts with multi-head self-attention and cross-attention mechanisms, the model creates structured and explainable descriptions of various video contexts. Performance tests through BLEU (1-4), CIDEr, METEOR, and ROUGE-L affirm the effectiveness of the model in creating coherent and context-aware descriptions in video understanding and autonomous decision-making.

Compared with current state-of-the-art models, the experimental findings consistently show greater performance. In addition to these quantitative enhancements, our research contributes to multimodal learning by expanding real-world uses in autonomous systems and medical imaging. The suggested frameworks significantly advance deep learning research by improving the interpretability, accuracy, and contextual richness of image description generation.

7.2 Future Directions

- Future research can concentrate on creating real-time image description generation systems capable of producing accurate and contextually sound descriptions in real-time, allowing applications for live video analysis and assistive technology for the blind.
- Enhancing multimodal fusion techniques—using deeper architectures like Graph Convolutional Networks (GCNs) and dynamic attention—to better capture interactions between textual and visual data is a crucial area for future research.
- Models should also be extended to handle 3D and multiview image data for better contextual understanding in applications like medical imaging, augmented reality, and robotics.
- In the future, self-supervised and unsupervised learning methods will be investigated to decrease reliance on large annotated datasets and make image description generation more scalable and transferable to new domains.
- Future research can investigate incorporating image description generation systems with AI-enabled IoT devices, facilitating real-time environmental comprehension and description creation for applications such as smart surveillance and home automation.
- In the future, as one possible direction, we would like to explore combining emotional and affective context in image descriptions for more human and empathetic-like description generation.

References

[1] Wikipedia contributors, "Photo caption — Wikipedia, the free encyclopedia," 2022, [Online; accessed 28-February-2022].

[2] F. Chen, X. Li, J. Tang, S. Li, and T. Wang, "A survey on recent advances in image captioning," in *Journal of Physics: Conference Series*, vol. 1914, no. 1. IOP Publishing, 2021, p. 012053.

[3] H. Wang, Y. Zhang, and X. Yu, "An overview of image caption generation methods," *Computational intelligence and neuroscience*, vol. 2020, 2020.

[4] J. Wu, Z. Cui, V. S. Sheng, P. Zhao, D. Su, and S. Gong, "A comparative study of sift and its variants," *Measurement science review*, vol. 13, no. 3, pp. 122–131, 2013.

[5] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, vol. 1. Ieee, 2005, pp. 886–893.

[6] R. Lebret, P. O. Pinheiro, and R. Collobert, "Simple image description generator via a linear phrase-based approach," *arXiv preprint arXiv:1412.8419*, 2014.

[7] A. Ghoshal, P. Ircing, and S. Khudanpur, "Hidden markov models for automatic annotation and content-based retrieval of images and video," in *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, 2005, pp. 544–551.

[8] Z. Song, X. Zhou, Z. Mao, and J. Tan, "Image captioning with context-aware auxiliary guidance," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 3, 2021, pp. 2584–2592.

[9] G. Sharma, P. Kalena, N. Malde, A. Nair, and S. Parkar, "Visual image caption generator using deep learning," in *2nd International Conference on Advances in Science & Technology (ICAST)*, 2019.

[10] S. He, W. Liao, H. R. Tavakoli, M. Yang, B. Rosenhahn, and N. Pugeault, "Image captioning through image transformer," in *Proceedings of the Asian conference on computer vision*, 2020.

[11] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *International conference on machine learning.* PMLR, 2015, pp. 2048–2057.

[12] X. Chen and C. L. Zitnick, "Learning a recurrent visual representation for image caption generation," *arXiv preprint arXiv:1411.5654*, 2014.

[13] R. Socher, A. Karpathy, Q. V. Le, C. D. Manning, and A. Y. Ng, "Grounded compositional semantics for finding and describing images with sentences," *Transactions of the Association for Computational Linguistics*, vol. 2, pp. 207–218, 2014.

[14] S. Bai and S. An, "A survey on automatic image caption generation," *Neurocomputing*, vol. 311, pp. 291–304, 2018.

[15] A. Elhagry and K. Kadaoui, "A thorough review on recent deep learning methodologies for image captioning," *arXiv e-prints*, pp. arXiv–2107, 2021.

[16] M. Stefanini, M. Cornia, L. Baraldi, S. Cascianelli, G. Fiameni, and R. Cucchiara, "From show to tell: A survey on image captioning," *arXiv e-prints*, pp. arXiv–2107, 2021.

[17] J. Jeon, V. Lavrenko, and R. Manmatha, “Automatic image annotation and retrieval using cross-media relevance models,” in *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, 2003, pp. 119–126.

[18] J. Li and J. Z. Wang, “Automatic linguistic indexing of pictures by a statistical modeling approach,” *IEEE Transactions on pattern analysis and machine intelligence*, vol. 25, no. 9, pp. 1075–1088, 2003.

[19] S. J. Rennie, E. Marcheret, Y. Mroueh, J. Ross, and V. Goel, “Self-critical sequence training for image captioning,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 7008–7024.

[20] M. Yang, J. Liu, Y. Shen, Z. Zhao, X. Chen, Q. Wu, and C. Li, “An ensemble of generation-and retrieval-based image captioning with dual generator generative adversarial network,” *IEEE Transactions on Image Processing*, vol. 29, pp. 9627–9640, 2020.

[21] G. Guo, H. Wang, D. Bell, Y. Bi, and K. Greer, “Knn model-based approach in classification,” in *On The Move to Meaningful Internet Systems 2003: CoopIS, DOA, and ODBASE: OTM Confederated International Conferences, CoopIS, DOA, and ODBASE 2003, Catania, Sicily, Italy, November 3-7, 2003. Proceedings.* Springer, 2003, pp. 986–996.

[22] M. A. Chandra and S. Bedi, “Survey on svm and their application in image classification,” *International Journal of Information Technology*, vol. 13, no. 5, pp. 1–11, 2021.

[23] Y. H. Tan and C. S. Chan, “Phrase-based image caption generator with hierarchical lstm network,” *Neurocomputing*, vol. 333, pp. 86–100, 2019.

[24] Y. Feng, L. Ma, W. Liu, and J. Luo, “Unsupervised image captioning,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 4125–4134.

[25] T. Yao, Y. Pan, Y. Li, Z. Qiu, and T. Mei, “Boosting image captioning with attributes,” in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 4894–4902.

[26] A. Mathews, L. Xie, and X. He, “Senticap: Generating image descriptions with sentiments,” in *Proceedings of the AAAI conference on artificial intelligence*, vol. 30, no. 1, 2016.

[27] M. Cornia, M. Stefanini, L. Baraldi, and R. Cucchiara, “Meshed-memory transformer for image captioning,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 10 578–10 587.

[28] G. Li, L. Zhu, P. Liu, and Y. Yang, “Entangled transformer for image captioning,” in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 8928–8937.

[29] J. Yu, J. Li, Z. Yu, and Q. Huang, “Multimodal transformer with multi-view visual representation for image captioning,” *IEEE transactions on circuits and systems for video technology*, vol. 30, no. 12, pp. 4467–4480, 2019.

[30] D.-J. Kim, J. Choi, T.-H. Oh, and I. S. Kweon, “Dense relational captioning: Triple-stream networks for relationship-based captioning,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 6271–6280.

[31] J. Johnson, A. Karpathy, and L. Fei-Fei, “Densecap: Fully convolutional localization networks for dense captioning,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 4565–4574.

[32] Z. Shao, J. Han, K. Debattista, and Y. Pang, “Textual context-aware dense captioning with diverse words,” *IEEE Transactions on Multimedia*, 2023.

[33] Z. Shao, J. Han, D. Marnerides, and K. Debattista, “Region-object relation-aware dense captioning via transformer,” *IEEE Transactions on Neural Networks and Learning Systems*, 2022.

[34] J.-Y. Pan, H.-J. Yang, P. Duygulu, and C. Faloutsos, "Automatic image captioning," in *2004 IEEE International Conference on Multimedia and Expo (ICME)(IEEE Cat. No. 04TH8763)*, vol. 3. IEEE, 2004, pp. 1987–1990.

[35] Y. Xiong, B. Du, and P. Yan, "Reinforced transformer for medical image captioning," in *Machine Learning in Medical Imaging: 10th International Workshop, MLMI 2019, Held in Conjunction with MICCAI 2019, Shenzhen, China, October 13, 2019, Proceedings 10*. Springer, 2019, pp. 673–680.

[36] 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, 2019.

[37] P. Dognin, I. Melnyk, Y. Mroueh, I. Padhi, M. Rigotti, J. Ross, Y. Schiff, R. A. Young, and B. Belgodere, "Image captioning as an assistive technology: Lessons learned from vizwiz 2020 challenge," *Journal of Artificial Intelligence Research*, vol. 73, pp. 437–459, 2022.

[38] L. Liu, S. Lu, R. Zhong, B. Wu, Y. Yao, Q. Zhang, and W. Shi, "Computing systems for autonomous driving: State of the art and challenges," *IEEE Internet of Things Journal*, vol. 8, no. 8, pp. 6469–6486, 2020.

[39] Q. Wang, Z. Yang, W. Ni, J. Wu, and Q. Li, "Semantic-spatial collaborative perception network for remote sensing image captioning," *IEEE Transactions on Geoscience and Remote Sensing*, 2024.

[40] S. Das and R. Sharma, "A textgcn-based decoding approach for improving remote sensing image captioning," *IEEE Geoscience and Remote Sensing Letters*, 2024.

[41] T. Yuan, X. Zhang, B. Liu, K. Liu, J. Jin, and Z. Jiao, "Surveillance video-and-language understanding: from small to large multimodal models," *IEEE Transactions on Circuits and Systems for Video Technology*, 2024.

[42] P. Hède, P.-A. Moëllic, J. Bourgeoys, M. Joint, and C. Thomas, "Automatic generation of natural language description for images." in *RIAO*. Citeseer, 2004, pp. 306–313.

[43] R. Mason and E. Charniak, "Domain-specific image captioning," in *Proceedings of the Eighteenth Conference on Computational Natural Language Learning*, 2014, pp. 11–20.

[44] R. Khan, M. S. Islam, K. Kanwal, M. Iqbal, M. Hossain, Z. Ye *et al.*, "A deep neural framework for image caption generation using gru-based attention mechanism," *arXiv preprint arXiv:2203.01594*, 2022.

[45] M. Mitchell, J. Dodge, A. Goyal, K. Yamaguchi, K. Stratos, X. Han, A. Mensch, A. Berg, T. Berg, and H. Daumé III, "Midge: Generating image descriptions from computer vision detections," in *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, 2012, pp. 747–756.

[46] P. H. Seo, P. Sharma, T. Levinboim, B. Han, and R. Soricut, "Reinforcing an image caption generator using off-line human feedback," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 03, 2020, pp. 2693–2700.

[47] Y. Zheng, Y. Li, and S. Wang, "Intention oriented image captions with guiding objects," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 8395–8404.

[48] J. Mao, W. Xu, Y. Yang, J. Wang, Z. Huang, and A. Yuille, "Deep captioning with multimodal recurrent neural networks (m-rnn)," *arXiv preprint arXiv:1412.6632*, 2014.

[49] M. Ilse, J. Tomczak, and M. Welling, "Attention-based deep multiple instance learning," in *International conference on machine learning*. PMLR, 2018, pp. 2127–2136.

[50] J. Aneja, A. Deshpande, and A. G. Schwing, "Convolutional image captioning," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 5561–5570.

[51] L. Guo, J. Liu, J. Tang, J. Li, W. Luo, and H. Lu, "Aligning linguistic words and visual semantic units for image captioning," in *Proceedings of the 27th ACM International Conference on Multimedia*, 2019, pp. 765–773.

[52] T. Yao, Y. Pan, Y. Li, and T. Mei, "Exploring visual relationship for image captioning," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 684–699.

[53] Y. Pan, T. Yao, Y. Li, and T. Mei, "X-linear attention networks for image captioning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 10 971–10 980.

[54] S. Liu, Z. Zhu, N. Ye, S. Guadarrama, and K. Murphy, "Improved image captioning via policy gradient optimization of spider," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 873–881.

[55] L. Agarwal and B. Verma, "From methods to datasets: A survey on image-caption generators," *Multimedia Tools and Applications*, vol. 83, no. 9, pp. 28 077–28 123, 2024.

[56] J. Dai and X. Zhang, "Automatic image caption generation using deep learning and multimodal attention," *Computer Animation and Virtual Worlds*, vol. 33, no. 3-4, p. e2072, 2022.

[57] S. Ding, S. Qu, Y. Xi, A. K. Sangaiah, and S. Wan, "Image caption generation with high-level image features," *Pattern Recognition Letters*, vol. 123, pp. 89–95, 2019.

[58] Q. Yucong and M. Li, "Image caption based on bigru and attention hybrid model," in *Proceedings of the 2021 4th International Conference on Artificial Intelligence and Pattern Recognition*, 2021, pp. 128–136.

[59] K. Qian, Y. Pan, H. Xu, and L. Tian, “Transformer model incorporating local graph semantic attention for image caption,” *The Visual Computer*, pp. 1–12, 2023.

[60] J. Li, Z. Mao, H. Li, W. Chen, and Y. Zhang, “Exploring visual relationships via transformer-based graphs for enhanced image captioning,” *ACM Transactions on Multimedia Computing, Communications and Applications*, vol. 20, no. 5, pp. 1–23, 2024.

[61] Y. Zhang, Y. Pan, T. Yao, R. Huang, T. Mei, and C.-W. Chen, “Boosting scene graph generation with visual relation saliency,” *ACM Transactions on Multimedia Computing, Communications and Applications*, vol. 19, no. 1, pp. 1–17, 2023.

[62] Y. Li, W. Ouyang, B. Zhou, K. Wang, and X. Wang, “Scene graph generation from objects, phrases and region captions,” in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 1261–1270.

[63] N. Xu, A.-A. Liu, J. Liu, W. Nie, and Y. Su, “Scene graph captioner: Image captioning based on structural visual representation,” *Journal of Visual Communication and Image Representation*, vol. 58, pp. 477–485, 2019.

[64] X. Yang, K. Tang, H. Zhang, and J. Cai, “Auto-encoding scene graphs for image captioning,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 10 685–10 694.

[65] S. Zhao, L. Li, and H. Peng, “Aligned visual semantic scene graph for image captioning,” *Displays*, vol. 74, p. 102210, 2022.

[66] Z. Li, J. Wei, F. Huang, and H. Ma, “Modeling graph-structured contexts for image captioning,” *Image and Vision Computing*, vol. 129, p. 104591, 2023.

[67] S. Ayoub, Y. Gulzar, F. A. Reegu, and S. Turaev, “Generating image captions using bahdanau attention mechanism and transfer learning,” *Symmetry*, vol. 14, no. 12, p. 2681, 2022.

[68] A. Karpathy and L. Fei-Fei, “Deep visual-semantic alignments for generating image descriptions,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3128–3137.

[69] Z. Shi, X. Zhou, X. Qiu, and X. Zhu, “Improving image captioning with better use of captions,” *arXiv e-print arXiv:2006.11807*, 2020.

[70] L. Guo, J. Liu, S. Lu, and H. Lu, “Show, tell, and polish: Ruminant decoding for image captioning,” *IEEE Transactions on Multimedia*, vol. 22, no. 8, pp. 2149–2162, 2019.

[71] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, “Show and tell: A neural image caption generator,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3156–3164.

[72] X. Jia, E. Gavves, B. Fernando, and T. Tuytelaars, “Guiding the long-short term memory model for image caption generation,” in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 2407–2415.

[73] J. Zhang, K. Li, Z. Wang, X. Zhao, and Z. Wang, “Visual enhanced glstm for image captioning,” *Expert Systems with Applications*, vol. 184, p. 115462, 2021.

[74] J. Lu, C. Xiong, D. Parikh, and R. Socher, “Knowing when to look: Adaptive attention via a visual sentinel for image captioning,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 375–383.

[75] N. Yu, X. Hu, B. Song, J. Yang, and J. Zhang, “Topic-oriented image captioning based on order-embedding,” *IEEE Transactions on Image Processing*, vol. 28, no. 6, pp. 2743–2754, 2018.

[76] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, “Learning transferable visual models from natural language supervision,” in *International conference on machine learning.* PMLR, 2021, pp. 8748–8763.

[77] J. Chen, D. Zhu, X. Shen, X. Li, Z. Liu, P. Zhang, R. Krishnamoorthi, V. Chandra, Y. Xiong, and M. Elhoseiny, "Minigpt-v2: large language model as a unified interface for vision-language multi-task learning," *arXiv preprint arXiv:2310.09478*, 2023.

[78] X. Yang, Y. Yang, S. Ma, Z. Li, W. Dong, and M. Woźniak, "Samt-generator: A second-attention for image captioning based on multi-stage transformer network," *Neurocomputing*, vol. 593, p. 127823, 2024.

[79] P. Zeng, H. Zhang, J. Song, and L. Gao, "S2 transformer for image captioning." in *IJCAI*, 2022, pp. 1608–1614.

[80] X. Yang, Y. Yang, J. Wu, W. Sun, S. Ma, and Z. Hou, "Ca-captioner: A novel concentrated attention for image captioning," *Expert Systems with Applications*, vol. 250, p. 123847, 2024.

[81] H. Parvin, A. R. Naghsh-Nilchi, and H. M. Mohammadi, "Image captioning using transformer-based double attention network," *Engineering Applications of Artificial Intelligence*, vol. 125, p. 106545, 2023.

[82] M. Barraco, S. Sarto, M. Cornia, L. Baraldi, and R. Cucchiara, "With a little help from your own past: Prototypical memory networks for image captioning," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 3021–3031.

[83] C.-W. Kuo and Z. Kira, "Haav: Hierarchical aggregation of augmented views for image captioning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 11 039–11 049.

[84] T. Yao, Y. Li, Y. Pan, Y. Wang, X.-P. Zhang, and T. Mei, "Dual vision transformer," *IEEE transactions on pattern analysis and machine intelligence*, 2023.

[85] M. Ding, B. Xiao, N. Codella, P. Luo, J. Wang, and L. Yuan, "Davit: Dual attention vision transformers," in *European Conference on Computer Vision*. Springer, 2022, pp. 74–92.

[86] J. Li, D. Li, C. Xiong, and S. Hoi, “Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation,” in *International conference on machine learning.* PMLR, 2022, pp. 12 888–12 900.

[87] K. Qian and L. Tian, “A topic-based multi-channel attention model under hybrid mode for image caption,” *Neural Computing and Applications*, vol. 34, no. 3, pp. 2207–2216, 2022.

[88] C. Wang and X. Gu, “Dynamic-balanced double-attention fusion for image captioning,” *Engineering Applications of Artificial Intelligence*, vol. 114, p. 105194, 2022.

[89] J. Ji, Y. Luo, X. Sun, F. Chen, G. Luo, Y. Wu, Y. Gao, and R. Ji, “Improving image captioning by leveraging intra-and inter-layer global representation in transformer network,” in *Proceedings of the AAAI conference on artificial intelligence*, vol. 35, no. 2, 2021, pp. 1655–1663.

[90] H. Zhang, P. Zeng, L. Gao, X. Lyu, J. Song, and H. T. Shen, “Spt: Spatial pyramid transformer for image captioning,” *IEEE Transactions on Circuits and Systems for Video Technology*, 2023.

[91] C. Wang, Y. Shen, and L. Ji, “Geometry attention transformer with position-aware lstms for image captioning,” *Expert systems with applications*, vol. 201, p. 117174, 2022.

[92] S. Cao, G. An, Z. Zheng, and Z. Wang, “Vision-enhanced and consensus-aware transformer for image captioning,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 10, pp. 7005–7018, 2022.

[93] L. Lou, K. Lu, and J. Xue, “A novel cross-fusion method of different types of features for image captioning,” in *2023 International Joint Conference on Neural Networks (IJCNN).* IEEE, 2023, pp. 1–8.

[94] H. Parvin, A. R. Naghsh-Nilchi, and H. M. Mohammadi, "Transformer-based local-global guidance for image captioning," *Expert Systems with Applications*, vol. 223, p. 119774, 2023.

[95] Z. Yang, Y. Yuan, Y. Wu, W. W. Cohen, and R. R. Salakhutdinov, "Review networks for caption generation," *Advances in neural information processing systems*, vol. 29, 2016.

[96] Z. Gan, C. Gan, X. He, Y. Pu, K. Tran, J. Gao, L. Carin, and L. Deng, "Semantic compositional networks for visual captioning," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 5630–5639.

[97] S. Wang, L. Lan, X. Zhang, and Z. Luo, "Gatecap: Gated spatial and semantic attention model for image captioning," *Multimedia Tools and Applications*, vol. 79, pp. 11 531–11 549, 2020.

[98] 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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 6077–6086.

[99] W. Jiang, L. Ma, Y.-G. Jiang, W. Liu, and T. Zhang, "Recurrent fusion network for image captioning," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 499–515.

[100] L. Huang, W. Wang, J. Chen, and X.-Y. Wei, "Attention on attention for image captioning," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 4634–4643.

[101] A. Abdelaal, N. F. ELshafey, N. W. Abdalah, N. H. Shaaban, S. A. Okasha, T. Yasser, M. Fathi, K. M. Fouad, and I. Abdelbaky, "Image captioning using vision encoder decoder model," in *2024 International Conference on Machine Intelligence and Smart Innovation (ICMISI)*. IEEE, 2024, pp. 101–106.

[102] A. Shetty, Y. Kale, Y. Patil, R. Patil, and S. Sharma, “Optimal transformers based image captioning using beam search,” *Multimedia Tools and Applications*, vol. 83, no. 16, pp. 47 963–47 977, 2024.

[103] X. Pan, T. Ye, D. Han, S. Song, and G. Huang, “Contrastive language-image pre-training with knowledge graphs,” *Advances in Neural Information Processing Systems*, vol. 35, pp. 22 895–22 910, 2022.

[104] J.-T. Sun and X. Min, “Research on image caption generation method based on multi-modal pre-training model and text mixup optimization,” *Signal, Image and Video Processing*, pp. 1–19, 2024.

[105] F. Liu, C. Yin, X. Wu, S. Ge, Y. Zou, P. Zhang, and X. Sun, “Contrastive attention for automatic chest x-ray report generation,” *arXiv e-print arXiv:2106.06965*, 2021.

[106] Z. Shaikh and J. Bharti, “Transformer-based chest x-ray report generation model,” in *International Conference on Soft Computing and Signal Processing*. Springer, 2023, pp. 227–236.

[107] F. A. Zeiser, C. A. da Costa, G. de Oliveira Ramos, A. Maier, and R. da Rosa Righi, “Chexreport: A transformer-based architecture to generate chest x-ray reports suggestions,” *Expert Systems with Applications*, p. 124644, 2024.

[108] G. Reale-Nosei, E. Amador-Domínguez, and E. Serrano, “From vision to text: A comprehensive review of natural image captioning in medical diagnosis and radiology report generation,” *Medical Image Analysis*, p. 103264, 2024.

[109] I. Shahzadi, T. M. Madni, U. I. Janjua, G. Batool, B. Naz, and M. Q. Ali, “Csamdt: Conditional self attention memory-driven transformers for radiology report generation from chest x-ray,” *Journal of Imaging Informatics in Medicine*, pp. 1–13, 2024.

[110] Z. Chen, Y. Song, T.-H. Chang, and X. Wan, “Generating radiology reports via memory-driven transformer,” *arXiv e-print arXiv:2010.16056*, 2020.

[111] V. Wijerathna, H. Raveen, S. Abeygunawardhana, and T. D. Ambegoda, “Chest x-ray caption generation with chexnet,” in *2022 Moratuwa Engineering Research Conference (MERCon)*. IEEE, 2022, pp. 1–6.

[112] J. Wang, A. Bhalerao, and Y. He, “Cross-modal prototype driven network for radiology report generation,” in *European Conference on Computer Vision*. Springer, 2022, pp. 563–579.

[113] Z. Huang, X. Zhang, and S. Zhang, “Kiut: Knowledge-injected u-transformer for radiology report generation,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 19 809–19 818.

[114] X. Zeng, T. Liao, L. Xu, and Z. Wang, “Aermnet: Attention-enhanced relational memory network for medical image report generation,” *Computer Methods and Programs in Biomedicine*, vol. 244, p. 107979, 2024.

[115] Q. You, H. Jin, Z. Wang, C. Fang, and J. Luo, “Image captioning with semantic attention,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 4651–4659.

[116] B. Jing, P. Xie, and E. Xing, “On the automatic generation of medical imaging reports,” *arXiv e-print arXiv:1711.08195*, 2017.

[117] B. Yan and M. Pei, “Clinical-bert: Vision-language pre-training for radiograph diagnosis and reports generation,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 3, 2022, pp. 2982–2990.

[118] M. Y. Ouis and M. A. Akhloufi, “Chestbiox-gen: contextual biomedical report generation from chest x-ray images using biogpt and co-attention mechanism,” *Frontiers in Imaging*, vol. 3, p. 1373420, 2024.

[119] D. Chen and W. B. Dolan, “Collecting highly parallel data for paraphrase evaluation,” in *Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies*, 2011, pp. 190–200.

[120] S. Venugopalan, M. Rohrbach, J. Donahue, R. Mooney, T. Darrell, and K. Saenko, “Sequence to sequence-video to text,” in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 4534–4542.

[121] L. Yao, A. Torabi, K. Cho, N. Ballas, C. Pal, H. Larochelle, and A. Courville, “Describing videos by exploiting temporal structure,” in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 4507–4515.

[122] L. Zhou, Y. Zhou, J. J. Corso, R. Socher, and C. Xiong, “End-to-end dense video captioning with masked transformer,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 8739–8748.

[123] J. Lei, L. Wang, Y. Shen, D. Yu, T. L. Berg, and M. Bansal, “Mart: Memory-augmented recurrent transformer for coherent video paragraph captioning,” *arXiv preprint arXiv:2005.05402*, 2020.

[124] C. Sun, A. Myers, C. Vondrick, K. Murphy, and C. Schmid, “Videobert: A joint model for video and language representation learning,” in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 7464–7473.

[125] J. Lei, L. Li, L. Zhou, Z. Gan, T. L. Berg, M. Bansal, and J. Liu, “Less is more: Clipbert for video-and-language learning via sparse sampling,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 7331–7341.

[126] J. Xu, T. Mei, T. Yao, and Y. Rui, “Msr-vtt: A large video description dataset for bridging video and language,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 5288–5296.

[127] J. Kim, A. Rohrbach, T. Darrell, J. Canny, and Z. Akata, "Textual explanations for self-driving vehicles," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 563–578.

[128] M. Shoman, D. Wang, A. Aboah, and M. Abdel-Aty, "Enhancing traffic safety with parallel dense video captioning for end-to-end event analysis," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 7125–7133.

[129] C. Cui, Y. Ma, X. Cao, W. Ye, Y. Zhou, K. Liang, J. Chen, J. Lu, Z. Yang, K.-D. Liao *et al.*, "A survey on multimodal large language models for autonomous driving," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2024, pp. 958–979.

[130] C. Hori, T. Hori, T.-Y. Lee, Z. Zhang, B. Harsham, J. R. Hershey, T. K. Marks, and K. Sumi, "Attention-based multimodal fusion for video description," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 4193–4202.

[131] I. Vasireddy, G. HimaBindu, and B. Ratnamala, "Transformative fusion: Vision transformers and gpt-2 unleashing new frontiers in image captioning within image processing," *International Journal of Innovative Research in Engineering & Management*, vol. 10, no. 6, pp. 55–59, 2023.

[132] X. Li, X. Yin, C. Li, P. Zhang, X. Hu, L. Zhang, L. Wang, H. Hu, L. Dong, F. Wei *et al.*, "Oscar: Object-semantics aligned pre-training for vision-language tasks," in *European Conference on Computer Vision.* Springer, 2020, pp. 121–137.

[133] R. Krishna, K. Hata, F. Ren, L. Fei-Fei, and J. Carlos Niebles, "Dense-captioning events in videos," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 706–715.

[134] N. Aafaq, N. Akhtar, W. Liu, S. Z. Gilani, and A. Mian, "Spatio-temporal dynamics and semantic attribute enriched visual encoding for video captioning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 12 487–12 496.

[135] D. Alexey, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv: 2010.11929*, 2020.

Appendix A

Vision Transformer (ViT)

A thorough yet brief explanation of Vision Transformer (ViT), including its architecture, mathematical formulation, benefits, and applications, is discussed in this appendix.

A.1 Overview

The Vision Transformer (ViT) adapts the transformer architecture, initially designed for NLP tasks, to process visual data. Unlike CNNs, which capture local spatial features, ViTs use self-attention to model global relationships between image patches. ViTs have shown state-of-the-art performance in various computer vision tasks such as classification, object detection, and segmentation. Vision Transformer architecture comprises several key stages as depicted in Figure A-1:

A.1.1 Image Patching and Embedding

The image is divided into fixed-size patches, for example, a 224x224 image split into 16x16 patches, resulting in 196 patches. Each patch is flattened into a 1D vector, then projected into a higher-dimensional space using a learnable linear projection.

$$N = \frac{H}{P} \times \frac{W}{P}$$

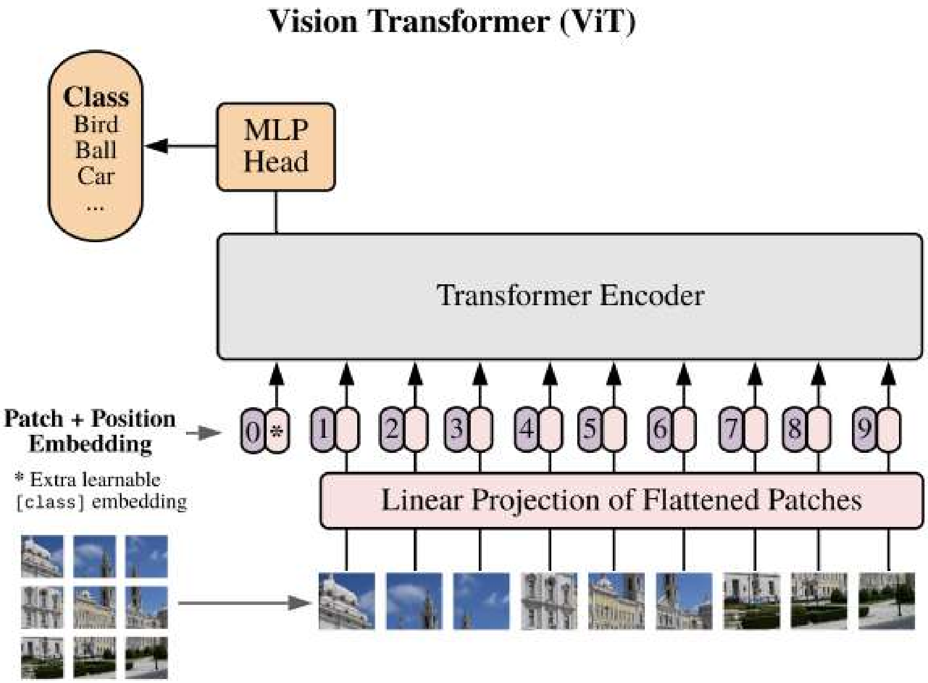

Vision Transformer (ViT)
Class
Bird
Ball
Car
...
MLP Head
Transformer Encoder
Patch + Position Embedding
* Extra learnable [class] embedding
0
*
1
2
3
4
5
6
7
8
9
Linear Projection of Flattened Patches

Figure A-1: Basic Architecture of Vision Transformer (ViT) [135]

where N is the number of patches.

A.1.2 Positional Encoding

Since transformers do not inherently preserve spatial order, positional encodings are added to each patch embedding. These encodings provide spatial context, allowing the model to understand patch relationships. Positional encodings can be either fixed or learned, with most ViTs using learnable encodings.

A.1.3 Transformer Encoder

The patch embeddings with positional information pass through transformer encoder layers, which include Multi-Head Self-Attention (MSA) and a Feed-Forward Network (FFN).

Multi-Head Self-Attention: Self-attention allows each patch to attend to every other patch, modeling long-range dependencies. The attention mechanism is defined as:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

Feed-Forward Network: Each transformer layer passes the patches through a feed-forward network with two fully connected layers and a non-linear activation.

Multiple transformer encoder layers are stacked, refining the patch embeddings and producing more abstract representations of the image.

A.1.4 Classification Token (CLS Token)

A CLS token is introduced at the beginning of the sequence. It aggregates information from all patches, learning to represent the entire image. After the transformer layers, the CLS token is used for classification.

A.1.5 Classification Head (MLP Head)

The output of the CLS token is passed through a Multi-Layer Perceptron (MLP) with a softmax layer to predict the image label.

The Vision Transformer (ViT) revolutionizes image processing by capturing global relationships between image patches, offering a powerful alternative to CNNs. While flexible and scalable, ViTs require large datasets and computational resources, and are expected to play an increasingly important role in future computer vision tasks.

Appendix B

RoBERTa (Robustly Optimized BERT Approach)

This appendix provides a brief overview of the RoBERTa decoder, highlighting its architecture, functionality, and applications in natural language processing.

B.1 Overview

RoBERTa (Robustly Optimized BERT Approach) is an improved variant of the BERT model, designed to enhance pretraining by optimizing key hyperparameters and removing certain limitations. It utilizes the transformer architecture as depicted in Figure B-1, originally introduced for NLP tasks, and has been shown to achieve superior performance on a wide range of natural language understanding tasks.

B.1.1 Architecture

RoBERTa builds upon the BERT architecture by training with larger batches, more data, and longer training times. It differs from BERT in that it removes the Next Sentence Prediction (NSP) objective and trains on longer sequences. RoBERTa processes input text using multiple transformer layers, each consisting of Multi-Head Self-Attention (MSA) and Feed-Forward Networks (FFN).

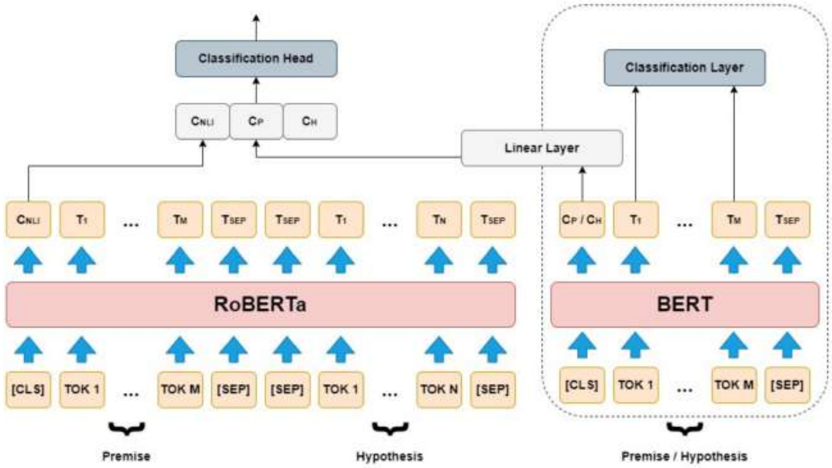


Figure B-1: Basic Architecture of RoBERTa

Multi-Head Self-Attention: The self-attention mechanism allows RoBERTa to model dependencies across the entire sequence, focusing on important contextual relationships between words. This is achieved by calculating attention scores using the dot product between queries (Q) and keys (K), followed by a softmax normalization:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

Feed-Forward Network: After the attention mechanism, the data is passed through a feed-forward network consisting of two fully connected layers, followed by a non-linear activation function (typically GELU).

B.1.2 Position-wise Feed-Forward Network (FFN)

Each transformer layer includes a position-wise FFN, consisting of two linear transformations with a non-linear activation function in between. The FFN allows the model to process each token independently, enhancing the learning of contextual information.

B.1.3 Token Embedding

RoBERTa tokenizes the input text into subword units, which are then embedded into continuous vectors using a shared embedding matrix. These embeddings are passed through the transformer layers to learn richer representations.

B.1.4 Final Output

The output of the transformer layers is processed by a classification head (MLP) to make predictions. The hidden states corresponding to each token are passed through the MLP, with the final output representing the predicted sequence or token.

B.1.5 Applications

RoBERTa has been used successfully in various NLP tasks such as question answering, sentiment analysis, and language modeling. Its robust pretraining allows it to generalize well to a wide range of text-based tasks, making it a popular choice for downstream NLP applications.

RoBERTa, through its improved architecture and optimization techniques, has become a leading model for natural language understanding, demonstrating significant improvements over BERT in multiple benchmarks.

Appendix C

Bidirectional Gated Recurrent Unit (BiGRU) Model

This appendix provides a brief overview of the Bi-GRU (Bidirectional Gated Recurrent Unit) model, covering its architecture, working principle, and applications in sequence modeling.

C.1 Overview

The Bi-GRU model is a type of Recurrent Neural Network (RNN) that uses GRU cells for sequence modeling. It enhances traditional GRU by processing input sequences in both forward and backward directions, allowing the model to capture contextual information from both past and future time steps. This bidirectional processing makes it more effective for tasks where context from both directions is important, such as language modeling, speech recognition, and machine translation.

C.1.1 Architecture

Bi-GRU consists of two GRU layers: one processes the sequence in a forward direction, while the other processes it in reverse. The outputs from both directions are concatenated, enabling the model to learn richer context from the entire sequence.

The basic architecture of Bi-GRU Model is depicted in Figure C-1.

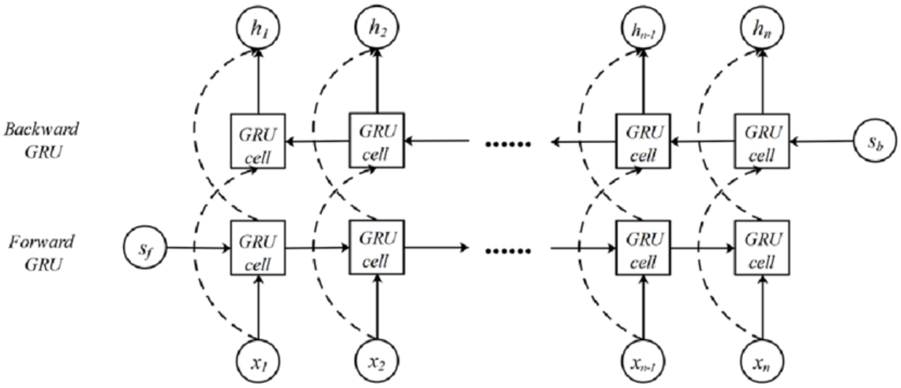


Figure C-1: Basic architecture of Bi-GRU Model

GRU Cell: The GRU unit is a variant of LSTM (Long Short-Term Memory) that simplifies the gating mechanism to improve efficiency. It consists of two main gates: the update gate and the reset gate.

$$\text{Update Gate: } z_t = \sigma(W_z x_t + U_z h_{t-1})$$

$$\text{Reset Gate: } r_t = \sigma(W_r x_t + U_r h_{t-1})$$

$$\text{Hidden State: } h_t = (1 - z_t) \odot h_{t-1} + z_t \odot \tanh(W_h x_t + U_h(r_t \odot h_{t-1}))$$

where z_t is the update gate, r_t is the reset gate, and h_t is the hidden state.

C.1.2 Bidirectional GRU

In the Bi-GRU model, two GRU layers are applied to the input sequence: one processes the sequence from left to right (forward) and the other from right to left (backward). The outputs of both directions are concatenated to form a richer representation of the sequence. This bidirectional approach helps the model capture dependencies from both past and future contexts, which is especially useful for tasks such as sequence classification, named entity recognition, and sentiment analysis.

C.1.3 Applications

Bi-GRU models are widely used in various natural language processing (NLP) tasks, including:

- **Machine Translation:** Capturing context from both directions of the input sentence improves translation quality.
- **Speech Recognition:** Bi-GRU helps in understanding both previous and future speech context, leading to better transcription accuracy.
- **Text Classification:** Bidirectional processing allows the model to better understand the context of words and sentences.

Bi-GRU models have become a standard in sequence modeling tasks, offering an effective method for learning contextual relationships in sequential data.

Appendix D

Scene Graphs

This appendix provides a concise overview of scene graphs and their role in image description generation.

D.1 Overview

Scene graphs are structured representations of visual scenes where objects are nodes, and relationships between them are edges. Scene graphs capture the spatial and semantic relationships between objects, providing a higher-level understanding of the image. They are used in image captioning to enhance the generation of contextually rich and accurate descriptions by incorporating relationships between visual entities.

D.1.1 Scene Graph Construction

To construct a scene graph, objects in an image are detected using object detection models (e.g., Faster R-CNN or YOLO). Each detected object is represented as a node in the graph. The relationships between these objects, such as "on," "in front of," or "next to," are identified through methods such as region-based convolutional neural networks (R-CNNs) or graph neural networks (GNNs). These relationships are represented as edges connecting the nodes.

Example: In an image of a dog sitting next to a tree, the scene graph might have nodes for "dog" and "tree" and an edge labeled "next to" connecting the two.

D.1.2 Scene Graphs in Image Captioning

In image captioning, scene graphs provide additional context that helps the model generate more accurate and semantically rich captions. Instead of focusing solely on detecting objects, scene graphs enable the model to understand how objects interact with one another. This contextual information allows the model to produce captions that are not just object-centric but also include relationships between objects.

Example: "A dog is sitting next to a tree" is a more descriptive caption than simply "A dog is in the image." Scene graphs help generate such detailed descriptions by providing information about the spatial relationships between objects.

D.1.3 Integration with Deep Learning Models

Scene graphs are often integrated with deep learning models, especially transformer-based models like Vision Transformers (ViT) and attention-based mechanisms. By combining scene graph representations with visual features extracted from the image, models can generate captions that reflect both the objects in the scene and their relationships.

Scene Graphs and Attention Mechanisms: Scene graphs can be used as inputs to attention-based models, where attention is focused not only on individual objects but also on the relationships between them, further enriching the generated descriptions.

D.1.4 Applications

Scene graphs in image description generation are particularly useful for tasks requiring rich contextual understanding, such as:

- **Complex Scene Understanding:** Scene graphs help in understanding complex relationships in images, such as interactions between multiple objects.

- **Detailed Caption Generation:** By capturing spatial and semantic relationships, scene graphs enable the generation of detailed, human-like captions.
- **Visual Question Answering (VQA):** Scene graphs can be used to answer questions that require an understanding of object relationships, such as "What is the dog doing near the tree?"

Scene graphs enhance image captioning by adding a structured and relational layer of information, improving both the accuracy and context of generated descriptions.

Appendix E

GPT-2 and GPT-4 Transformers

This appendix provides an overview of the GPT-2 and GPT-4 transformers, highlighting their architecture, working principles, and applications.

E.1 Overview

GPT (Generative Pre-trained Transformer) models are a family of transformer-based architectures designed for natural language generation (NLG). GPT-2 and GPT-4 are two popular models in this family, each offering unique capabilities and improvements in language generation tasks.

E.1.1 GPT-2 Architecture

GPT-2 is a large-scale, unsupervised language model developed by OpenAI. It utilizes the transformer architecture and is pre-trained on a vast corpus of text data using a causal (autoregressive) language modeling objective. In GPT-2, the model generates text by predicting the next token in a sequence based on the preceding context. The basic architecture is shown in Figure E-1.

Architecture: GPT-2 is built on a stack of transformer decoder layers, with each layer consisting of Multi-Head Self-Attention (MSA) and Feed-Forward Networks (FFN). GPT-2 processes sequences of tokens one at a time, predicting the next token

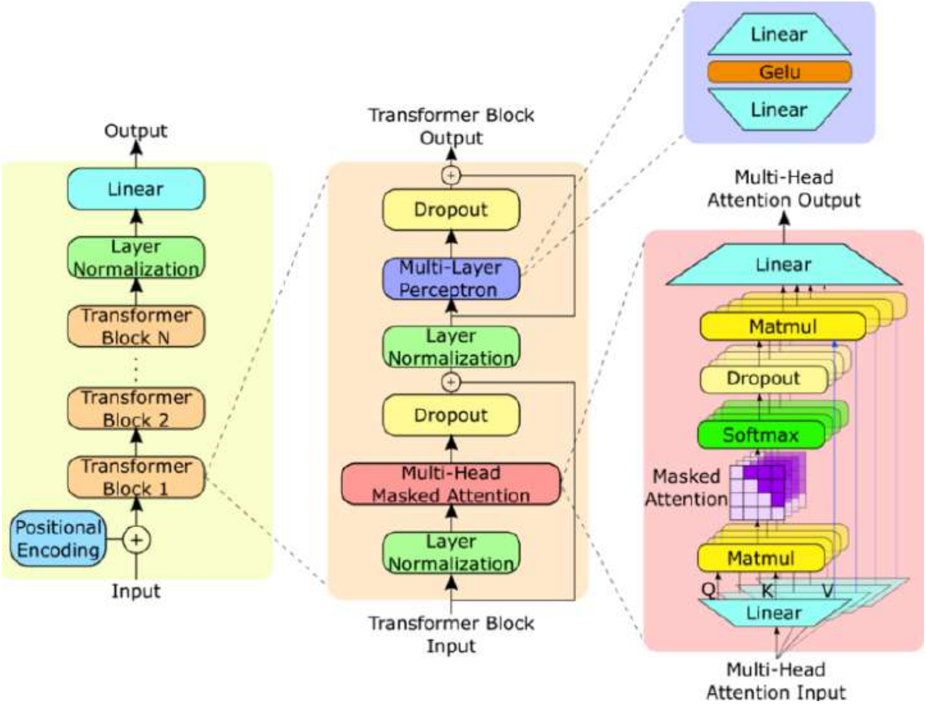


Figure E-1: Basic Architecture of GPT-2 Transformer Model

based on prior tokens.

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

where Q (query), K (key), and V (value) are the learned representations of input tokens.

Applications: GPT-2 is used for a variety of tasks, including text generation, summarization, translation, and dialogue systems.

E.1.2 GPT-4 Architecture

GPT-4 is a more advanced version of the GPT series, with significantly more parameters and improved capabilities over GPT-2. It introduces refinements in both training techniques and model architecture to enhance performance, especially on more complex tasks.

Architecture: GPT-4 uses a larger transformer architecture with more layers and parameters compared to GPT-2. It improves upon GPT-2's autoregressive generation by incorporating better handling of context, leading to more coherent and contextually accurate responses. GPT-4 also benefits from multimodal capabilities, being able to process both text and image inputs.

Capabilities: GPT-4 achieves better performance on tasks requiring deep understanding, such as logical reasoning, multi-step problem solving, and generating creative content. It also demonstrates improved performance on nuanced NLP tasks like sentiment analysis, translation, and summarization.

E.1.3 Key Differences Between GPT-2 and GPT-4

- **Size:** GPT-4 is significantly larger, with billions more parameters compared to GPT-2, allowing it to model more complex relationships and generate more accurate outputs.
- **Multimodal:** GPT-4 can handle both text and image inputs, whereas GPT-2 is limited to text-based inputs.
- **Contextual Understanding:** GPT-4 exhibits superior contextual understanding, allowing it to generate more coherent and contextually relevant responses.
- **Performance:** GPT-4 excels at tasks involving complex reasoning, logic, and multi-step problem solving, whereas GPT-2 performs well on simpler tasks.

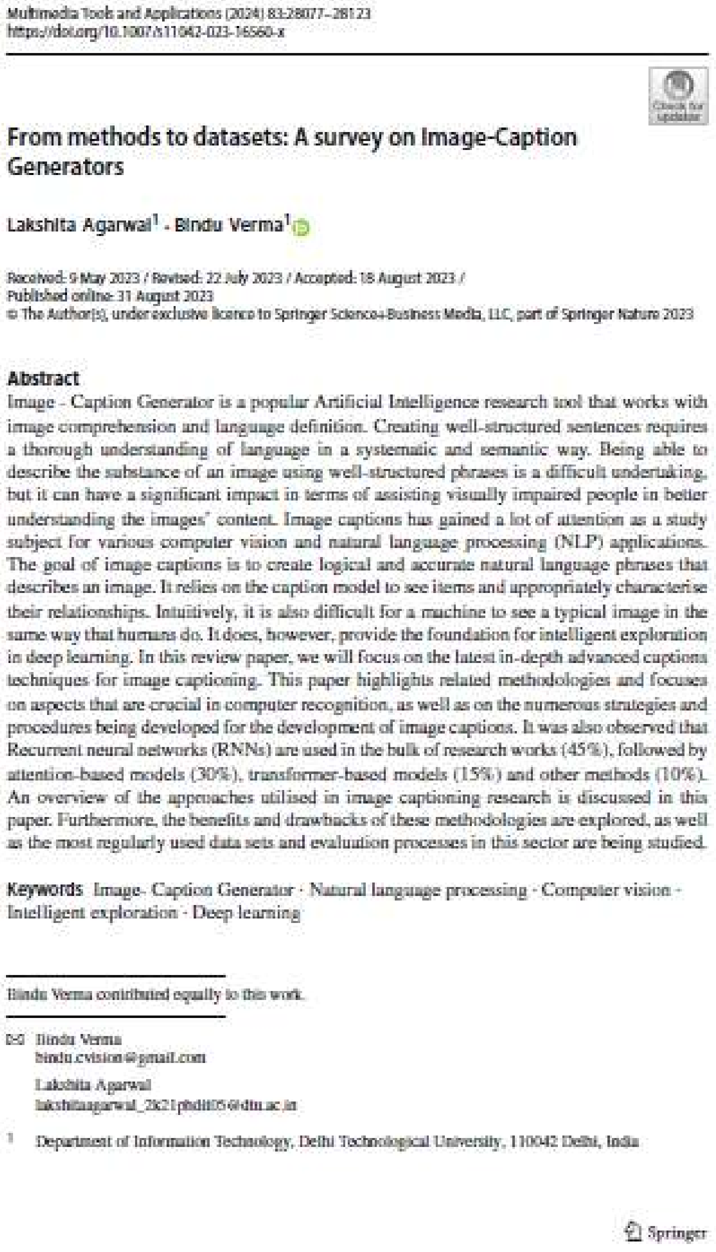
GPT-2 and GPT-4 transform the field of natural language processing by enabling advanced language understanding and generation, with GPT-4 representing the state-of-the-art in language modeling and multimodal processing.

LIST OF PUBLICATION AND THEIR PROOFS

LIST OF JOURNALS:

Journal Paper: 1 - SCIE Indexed

Agarwal, L., Verma, B. From methods to datasets: A survey on Image-Caption Generators. *Multimedia Tools Appl* **83**, 28077–28123 (2024). **DOI:** https://doi.org/10.1007/s11042-023-16560-x ***(Published, SCIE Indexed, I.F. = 3)***



Multimedia Tools and Applications (2024) 83:28077–28123
https://doi.org/10.1007/s11042-023-16560-x

From methods to datasets: A survey on Image-Caption Generators

Lakshita Agarwal[1] · Bindu Verma[1]

Received: 9 May 2023 / Revised: 22 July 2023 / Accepted: 18 August 2023 /
Published online: 31 August 2023
© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2023

Abstract

Image - Caption Generator is a popular Artificial Intelligence research tool that works with image comprehension and language definition. Creating well-structured sentences requires a thorough understanding of language in a systematic and semantic way. Being able to describe the substance of an image using well-structured phrases is a difficult undertaking, but it can have a significant impact in terms of assisting visually impaired people in better understanding the images' content. Image captions has gained a lot of attention as a study subject for various computer vision and natural language processing (NLP) applications. The goal of image captions is to create logical and accurate natural language phrases that describes an image. It relies on the caption model to see items and appropriately characterise their relationships. Intuitively, it is also difficult for a machine to see a typical image in the same way that humans do. It does, however, provide the foundation for intelligent exploration in deep learning. In this review paper, we will focus on the latest in-depth advanced captions techniques for image captioning. This paper highlights related methodologies and focuses on aspects that are crucial in computer recognition, as well as on the numerous strategies and procedures being developed for the development of image captions. It was also observed that Recurrent neural networks (RNNs) are used in the bulk of research works (45%), followed by attention-based models (30%), transformer-based models (15%) and other methods (10%). An overview of the approaches utilised in image captioning research is discussed in this paper. Furthermore, the benefits and drawbacks of these methodologies are explored, as well as the most regularly used data sets and evaluation processes in this sector are being studied.

Keywords Image- Caption Generator · Natural language processing · Computer vision · Intelligent exploration · Deep learning

---

Bindu Verma contributed equally to this work.

✉ Bindu Verma
bindu.cvision@gmail.com

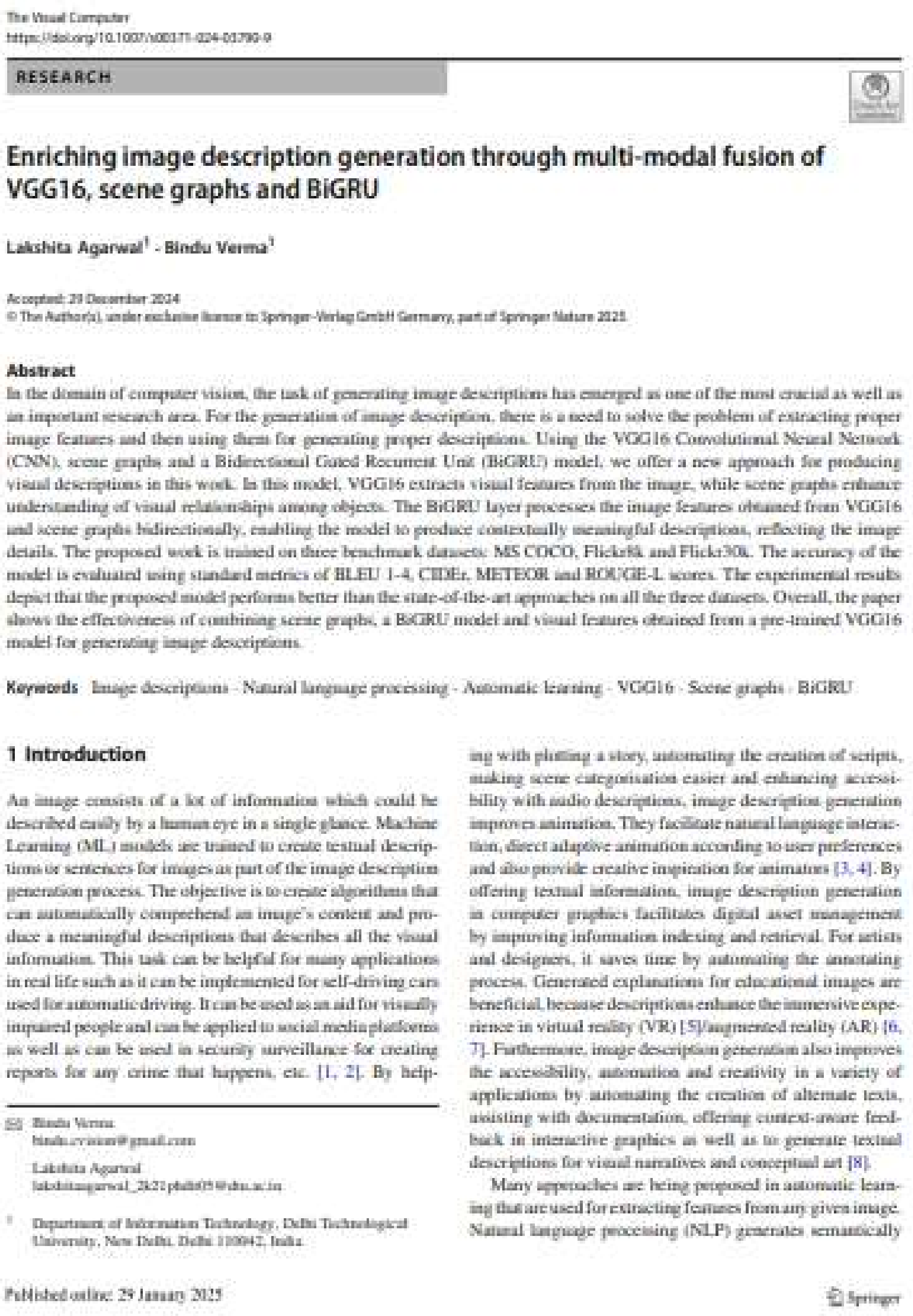
Lakshita Agarwal
lakshitaagarwal_2k21phdit05@dtu.ac.in

[1] Department of Information Technology, Delhi Technological University, 110042 Delhi, India

Journal Paper: 2 - SCIE Indexed

Agarwal, L., Verma, B. Enriching image description generation through multi-modal fusion of VGG16, scene graphs and BiGRU. *Visual Computer* (2025).

DOI: https://doi.org/10.1007/s00371-024-03790-9 ***(Published, SCIE Indexed, I.F. = 3.0)***



The Visual Computer
https://doi.org/10.1007/s00371-024-03790-9

RESEARCH

Enriching image description generation through multi-modal fusion of VGG16, scene graphs and BiGRU

Lakshita Agarwal[1] · Bindu Verma[1]

Accepted: 29 December 2024
© The Author(s), under exclusive licence to Springer-Verlag GmbH Germany, part of Springer Nature 2025

Abstract
In the domain of computer vision, the task of generating image descriptions has emerged as one of the most crucial as well as an important research area. For the generation of image description, there is a need to solve the problem of extracting proper image features and then using them for generating proper descriptions. Using the VGG16 Convolutional Neural Network (CNN), scene graphs and a Bidirectional Gated Recurrent Unit (BiGRU) model, we offer a new approach for producing visual descriptions in this work. In this model, VGG16 extracts visual features from the image, while scene graphs enhance understanding of visual relationships among objects. The BiGRU layer processes the image features obtained from VGG16 and scene graphs bidirectionally, enabling the model to produce contextually meaningful descriptions, reflecting the image details. The proposed work is trained on three benchmark datasets: MS COCO, Flickr8k and Flickr30k. The accuracy of the model is evaluated using standard metrics of BLEU 1-4, CIDEr, METEOR and ROUGE-L scores. The experimental results depict that the proposed model performs better than the state-of-the-art approaches on all the three datasets. Overall, the paper shows the effectiveness of combining scene graphs, a BiGRU model and visual features obtained from a pre-trained VGG16 model for generating image descriptions.

Keywords Image descriptions · Natural language processing · Automatic learning · VGG16 · Scene graphs · BiGRU

1 Introduction

An image consists of a lot of information which could be described easily by a human eye in a single glance. Machine Learning (ML) models are trained to create textual descriptions or sentences for images as part of the image description generation process. The objective is to create algorithms that can automatically comprehend an image's content and produce a meaningful descriptions that describes all the visual information. This task can be helpful for many applications in real life such as it can be implemented for self-driving cars used for automatic driving. It can be used as an aid for visually impaired people and can be applied to social media platforms as well as can be used in security surveillance for creating reports for any crime that happens, etc. [1, 2]. By helping with plotting a story, automating the creation of scripts, making scene categorisation easier and enhancing accessibility with audio descriptions, image description generation improves animation. They facilitate natural language interaction, direct adaptive animation according to user preferences and also provide creative inspiration for animators [3, 4]. By offering textual information, image description generation in computer graphics facilitates digital asset management by improving information indexing and retrieval. For artists and designers, it saves time by automating the annotating process. Generated explanations for educational images are beneficial, because descriptions enhance the immersive experience in virtual reality (VR) [5]/augmented reality (AR) [6, 7]. Furthermore, image description generation also improves the accessibility, automation and creativity in a variety of applications by automating the creation of alternate texts, assisting with documentation, offering context-aware feedback in interactive graphics as well as to generate textual descriptions for visual narratives and conceptual art [8].

Many approaches are being proposed in automatic learning that are used for extracting features from any given image. Natural language processing (NLP) generates semantically

---

✉ Bindu Verma
bindu.cvision@gmail.com

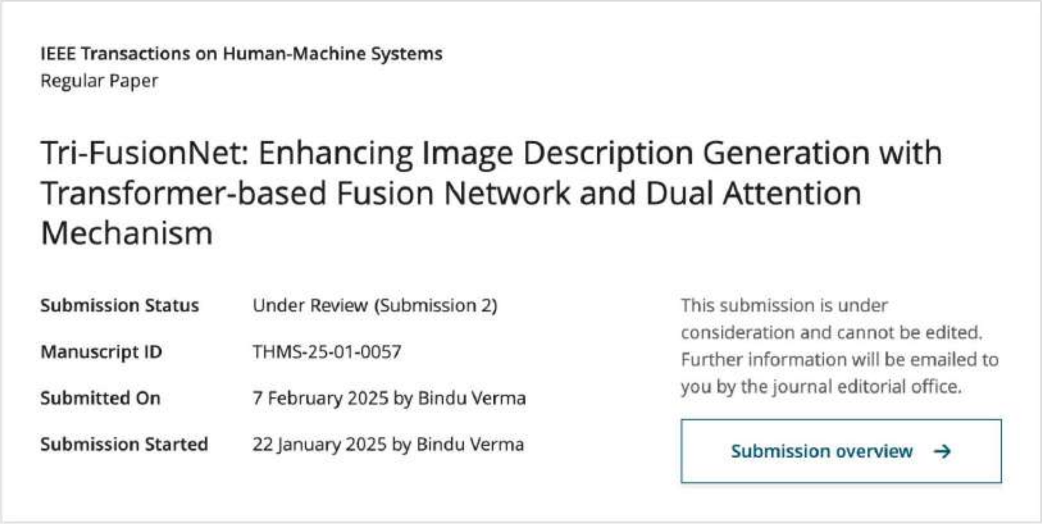
Lakshita Agarwal
lakshitaagarwal_2k21phdit05@dtu.ac.in

[1] Department of Information Technology, Delhi Technological University, New Delhi, Delhi 110042, India

Published online: 29 January 2025

Journal Paper: 3 - IEEE Transaction

Lakshita Agarwal, and Bindu Verma. "Tri-FusionNet: Enhancing Image Description Generation with Transformer-based Fusion Network and Dual Attention Mechanism" is communicated in IEEE Transactions on Human-Machine Systems ***(Communicated- 1st Major Revision Submitted, I.F. = 3.5). Archived at: http://arxiv.org/abs/2504.16761.***



IEEE Transactions on Human-Machine Systems
Regular Paper

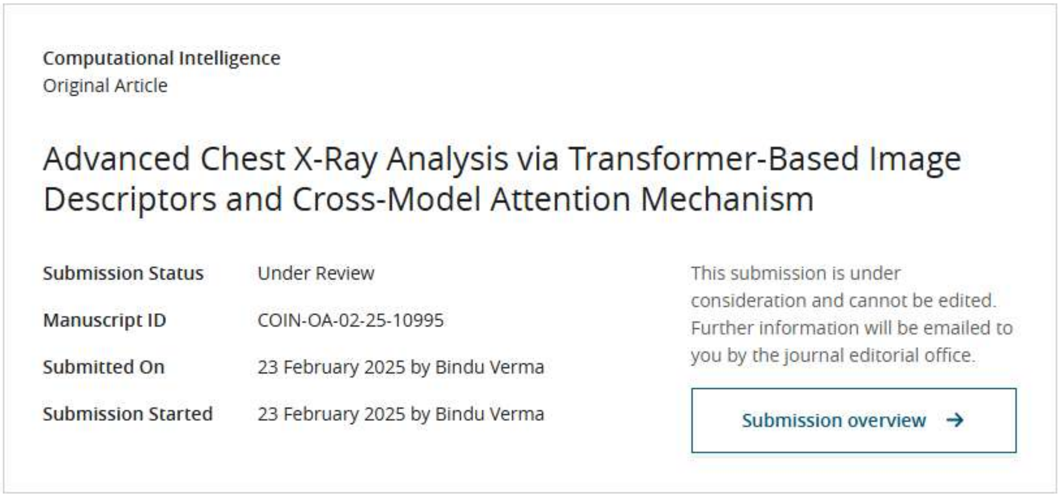
Tri-FusionNet: Enhancing Image Description Generation with Transformer-based Fusion Network and Dual Attention Mechanism

Submission Status	Under Review (Submission 2)
Manuscript ID	THMS-25-01-0057
Submitted On	7 February 2025 by Bindu Verma
Submission Started	22 January 2025 by Bindu Verma

This submission is under consideration and cannot be edited. Further information will be emailed to you by the journal editorial office.

Journal Paper: 4 – SCIE Indexed

Lakshita Agarwal, and Bindu Verma. "Advanced Chest X-Ray Analysis via Transformer-Based Image Descriptors and Cross-Model Attention Mechanism" is communicated in Computational Intelligence (Wiley) ***(Communicated, SCIE Indexed, I.F. = 1.8). Archived at: http://arxiv.org/abs/2504.16774.***



Computational Intelligence
Original Article

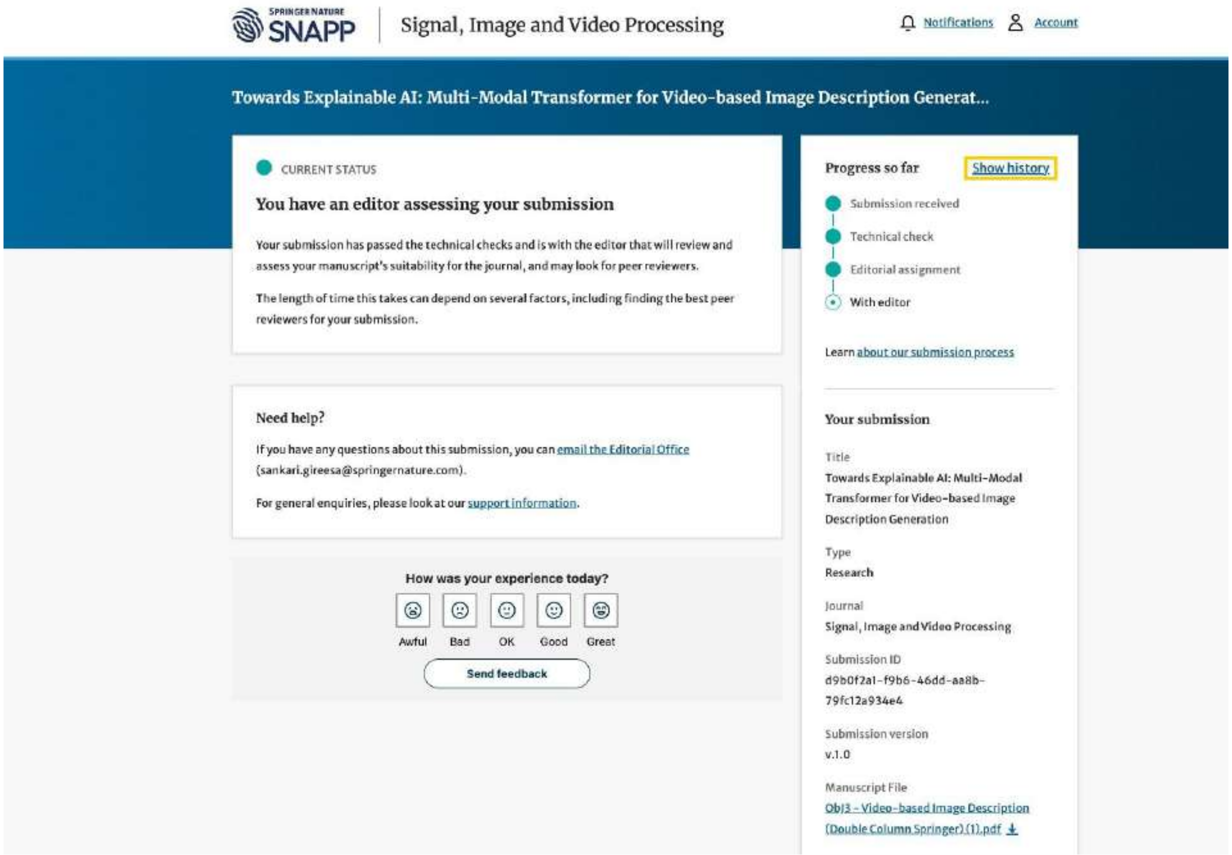
Advanced Chest X-Ray Analysis via Transformer-Based Image Descriptors and Cross-Model Attention Mechanism

Submission Status	Under Review
Manuscript ID	COIN-OA-02-25-10995
Submitted On	23 February 2025 by Bindu Verma
Submission Started	23 February 2025 by Bindu Verma

This submission is under consideration and cannot be edited. Further information will be emailed to you by the journal editorial office.

Journal Paper: 5 – SCIE Indexed

Lakshita Agarwal, and Bindu Verma. "Towards Explainable AI: Multi-Modal Transformer for Video-based Image Description Generation" is communicated in Signal, Image and Video Processing (Springer). ***(Communicated, SCIE Indexed, I.F. = 2.0). Archived at: http://arxiv.org/abs/2504.16788***



Towards Explainable AI: Multi-Modal Transformer for Video-based Image Description Generat...

CURRENT STATUS

You have an editor assessing your submission

Your submission has passed the technical checks and is with the editor that will review and assess your manuscript's suitability for the journal, and may look for peer reviewers.

The length of time this takes can depend on several factors, including finding the best peer reviewers for your submission.

Need help?

If you have any questions about this submission, you can email the Editorial Office (sankari.gireesa@springernature.com).

For general enquiries, please look at our support information.

How was your experience today?

Progress so far

- Submission received
- Technical check
- Editorial assignment
- With editor

Learn about our submission process

Your submission

Title
Towards Explainable AI: Multi-Modal Transformer for Video-based Image Description Generation

Type
Research

Journal
Signal, Image and Video Processing

Submission ID
d9b0f2a1-f9b6-46dd-aa8b-79fc12a934e4

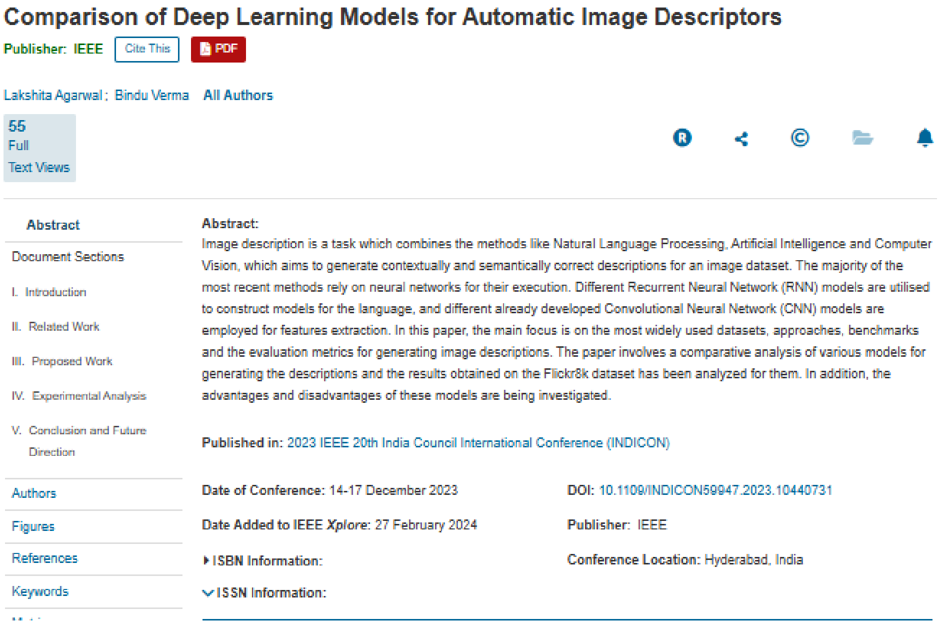
Submission version
v.1.0

Manuscript File
ObJ3 - Video-based Image Description (Double Column Springer) (1).pdf

LIST OF CONFERENCES

International Conference: 1

L. Agarwal and B. Verma, "Comparison of Deep Learning Models for Automatic Image Descriptors," *2023 IEEE 20th India Council International Conference (INDICON)*, Hyderabad, India, 2023, pp. 914-919. **DOI**: 10.1109/INDICON59947.2023.10440731 ***(Published)***



Comparison of Deep Learning Models for Automatic Image Descriptors

Publisher: IEEE

Lakshita Agarwal ; Bindu Verma All Authors

55 Full Text Views

Abstract:
Image description is a task which combines the methods like Natural Language Processing, Artificial Intelligence and Computer Vision, which aims to generate contextually and semantically correct descriptions for an image dataset. The majority of the most recent methods rely on neural networks for their execution. Different Recurrent Neural Network (RNN) models are utilised to construct models for the language, and different already developed Convolutional Neural Network (CNN) models are employed for features extraction. In this paper, the main focus is on the most widely used datasets, approaches, benchmarks and the evaluation metrics for generating image descriptions. The paper involves a comparative analysis of various models for generating the descriptions and the results obtained on the Flickr8k dataset has been analyzed for them. In addition, the advantages and disadvantages of these models are being investigated.

Published in: 2023 IEEE 20th India Council International Conference (INDICON)

Date of Conference: 14-17 December 2023

DOI: 10.1109/INDICON59947.2023.10440731

Date Added to IEEE *Xplore*: 27 February 2024

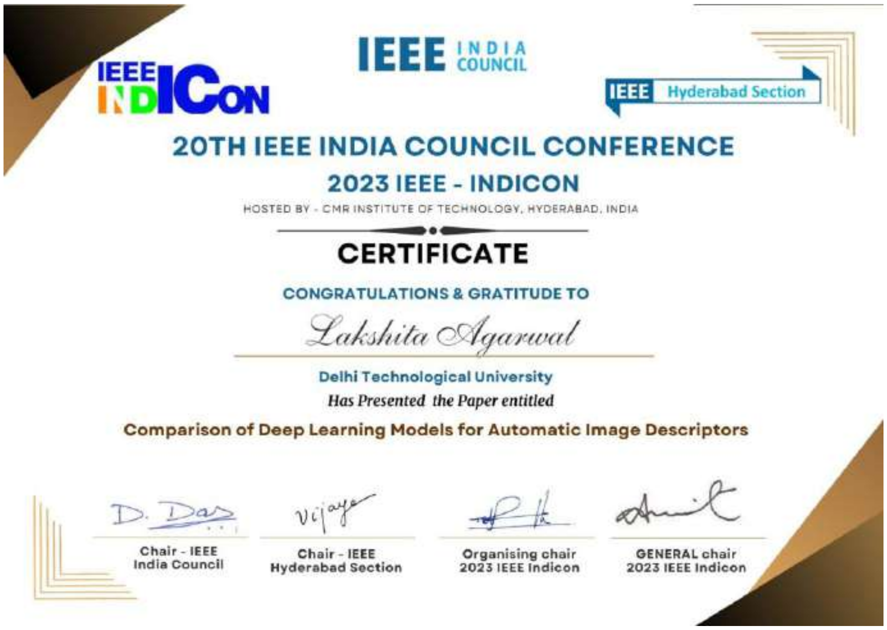
Publisher: IEEE

▸ **ISBN Information:**

Conference Location: Hyderabad, India

⌄ **ISSN Information:**

International Conference 1: Certificate



IEEE INDICON

IEEE INDIA COUNCIL

IEEE Hyderabad Section

20TH IEEE INDIA COUNCIL CONFERENCE

2023 IEEE - INDICON

HOSTED BY - CMR INSTITUTE OF TECHNOLOGY, HYDERABAD, INDIA

CERTIFICATE

CONGRATULATIONS & GRATITUDE TO

Lakshita Agarwal

Delhi Technological University

Has Presented the Paper entitled

Comparison of Deep Learning Models for Automatic Image Descriptors

D. Das
Chair - IEEE India Council

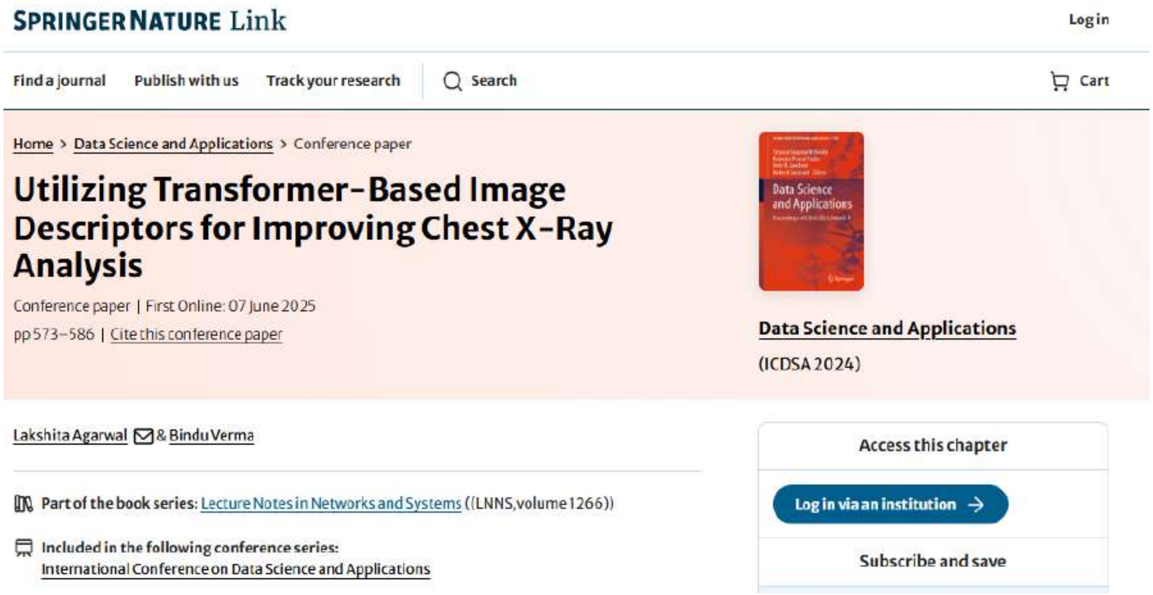
Chair - IEEE Hyderabad Section

Organising chair 2023 IEEE Indicon

GENERAL chair 2023 IEEE Indicon

International Conference: 2

Lakshita Agarwal, and Bindu Verma. "Utilizing Transformer-Based Image Descriptors for Improving Chest X-Ray Analysis" presented in *5th International Conference on Data Science and Applications (ICDSA 2024), Springer*. **(Published)**



SPRINGER NATURE Link

Log in

Find a journal   Publish with us   Track your research   Search   Cart

Home > Data Science and Applications > Conference paper

Utilizing Transformer-Based Image Descriptors for Improving Chest X-Ray Analysis

Conference paper | First Online: 07 June 2025

pp 573–586 | Cite this conference paper

Data Science and Applications

(ICDSA 2024)

Lakshita Agarwal & Bindu Verma

Part of the book series: Lecture Notes in Networks and Systems ((LNNS,volume 1266))

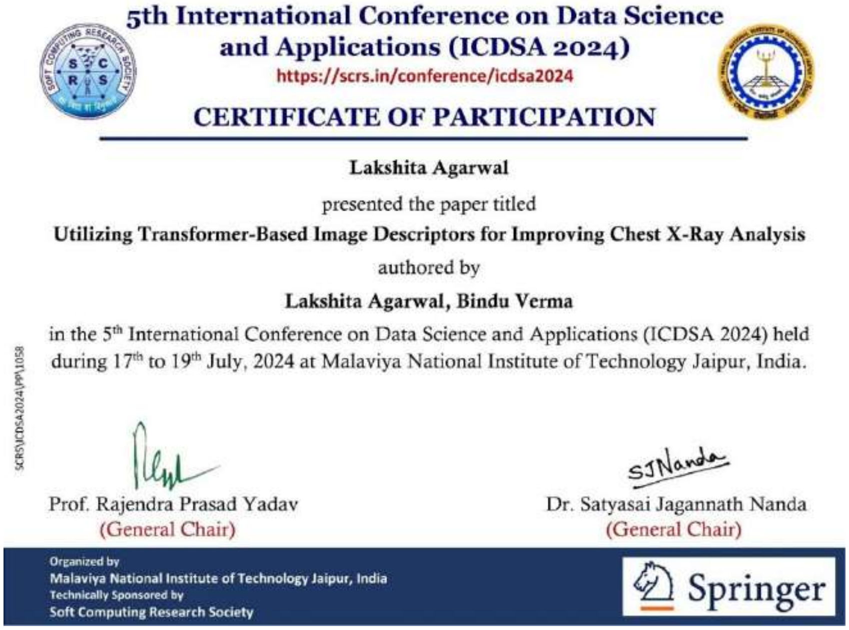
Included in the following conference series:
International Conference on Data Science and Applications

Access this chapter

Log in via an institution →

Subscribe and save

International Conference 2: Certificate



5th International Conference on Data Science and Applications (ICDSA 2024)

https://scrs.in/conference/icdsa2024

CERTIFICATE OF PARTICIPATION

Lakshita Agarwal

presented the paper titled

Utilizing Transformer-Based Image Descriptors for Improving Chest X-Ray Analysis

authored by

Lakshita Agarwal, Bindu Verma

in the 5th International Conference on Data Science and Applications (ICDSA 2024) held during 17th to 19th July, 2024 at Malaviya National Institute of Technology Jaipur, India.

SCRS\ICDSA2024\PP\1058

SJNanda

Prof. Rajendra Prasad Yadav
(General Chair)

Dr. Satyasai Jagannath Nanda
(General Chair)

Organized by
Malaviya National Institute of Technology Jaipur, India
Technically Sponsored by
Soft Computing Research Society

Springer

DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi College of Engineering)
Shahbad Daulatpur, Main Bawana Road, Delhi-110042

PLAGIARISM VERIFICATION

Title of the Thesis: Design a Framework for Generation of Image Description using Deep Learning

Total Pages: 171 Pages

Name of the Scholar: Lakshita Agarwal

Supervisor: Dr. Bindu Verma

Department: Information Technology

This is to report that the above thesis was scanned for similarity detection. Process and outcome are given below:

Software used: Turnitin **Similarity Index:** 7% **Word Count:** 40,741 Words

Date: May 8, 2025

Lakshita Agarwal
Candidate's Signature

Signature of Supervisor

Lakshita Agarwal

Lakshita Agarwal PHD Thesis.pdf

Delhi Technological University

Document Details

Submission ID

trn:oid:::27535:94895527

Submission Date

May 8, 2025, 3:30 PM GMT+5:30

Download Date

May 8, 2025, 3:34 PM GMT+5:30

File Name

Lakshita Agarwal PHD Thesis.pdf

File Size

25.3 MB

170 Pages

40,741 Words

234,194 Characters

7% Overall Similarity

The combined total of all matches, including overlapping sources, for each database.

Filtered from the Report

- Bibliography
- Quoted Text
- Cited Text
- Small Matches (less than 9 words)

Exclusions

- 2 Excluded Sources

Match Groups

241 Not Cited or Quoted 7%
Matches with neither in-text citation nor quotation marks

0 Missing Quotations 0%
Matches that are still very similar to source material

0 Missing Citation 0%
Matches that have quotation marks, but no in-text citation

0 Cited and Quoted 0%
Matches with in-text citation present, but no quotation marks

Top Sources

2%	Internet sources
5%	Publications
3%	Submitted works (Student Papers)

Lakshita Agarwal

Lakshita Agarwal
(Ph.D. Student)

Dr. Bindu Verma
(Supervisor)

Integrity Flags

0 Integrity Flags for Review

No suspicious text manipulations found.

Lakshita Agarwal

Ph.D. Scholar (Department of IT, Delhi Technological University, Delhi)

Email IDs: lakshitaa3@gmail.com, lakshitaagarwal_2k21phdit05@dtu.ac.in.

Contact Nos.: 9837786375, 7983277026

Date of Birth: 03/03/1998

EDUCATION:

- ***Doctorate of Philosophy (Ph.D.- Information Technology)- August 2021- Present (Pursuing):*** Delhi Technological University (DTU), Delhi. **CGPA- 9.60 CGPA** (Course Work).
- ***Masters of Technology (M. Tech.-Computer Engineering)- 2019 to 2021:*** College of Technology- G.B. Pant University of Agriculture & Technology, Pantnagar, Uttarakhand. **CGPA- 8.246 CGPA.**
- ***Bachelors of Technology (B.Tech.- Computer Science and Engineering)- 2015 to 2019:*** College of Engineering Roorkee, Affiliated To- Uttarakhand Technical University, Dehradun, Uttarakhand. **Aggregate- 81.72%**.
- ***Class 12th (2014- 2015):*** Maria Assumpta Convent School, Kashipur, Uttarakhand (CBSE). **Aggregate- 78.2%.**
- ***Class 10th (2012- 2013):*** Maria Assumpta Convent School, Kashipur, Uttarakhand (CBSE). **CGPA- 9.2 CGPA.**

CAREER OBJECTIVE:

My goal is to get associated with an organization/institution where I can utilize my skills and gain further experience while enhancing the productivity and reputation of the organization/institution.

LIST OF PUBLICATIONS:

Journal Papers:

1. **Lakshita Agarwal**, and Bindu Verma. "From methods to datasets: A survey on Image-Caption Generators" Multimedia Tools and Applications (2023): 28077–28123.
2. **Lakshita Agarwal**, and Bindu Verma. "Enriching image description generation through multi-modal fusion of VGG16, scene graphs and BiGRU" The Visual Computer (2024): 1-21.
3. **Lakshita Agarwal**, Chetan Singh Negi, Jalaj Sharma, and Sunita Jalal. "A survey on the controller placement problem in SDN." International Journal of Advanced Networking and Applications 13, no. 2 (2021): 4896-4914.

International Conferences:

1. **Lakshita Agarwal**, and Bindu Verma. "Comparison of Deep Learning Models for Automatic Image Descriptors". In 2023 IEEE 20th India Council International Conference (INDICON) (pp.

914-919) IEEE, 2023.

2. **Lakshita Agarwal**, and Bindu Verma. "Utilizing Transformer-Based Image Descriptors for Improving Chest X-Ray Analysis" presented in 5th International Conference on Data Science and Applications (ICDSA 2024), Springer.

PROFESSIONAL SKILLS AND PROFICIENCY:

- **Domains of Interest:** Computer Vision, Natural Language Processing, Deep Learning, Artificial Intelligence, Digital Image Processing, Autonomous Driving, Software Defined Networks.
- **Coding Languages-** Python, C, C++, Java and HTML.
- **Microsoft Office Skills-** Word, PowerPoint, Excel.
- **Operating System-** Windows, UNIX/Linux
- **Tools-** Google Colab, Kaggle, GitHub, Anaconda Navigator, NS2 and NS3 Simulators, Visual Studio, Orange.

ACHIEVEMENTS & CERTIFICATIONS:

1. Attended One-week Short Term Course on 'Recent Trends in Machine Learning and Deep Learning for AI Applications organized by Department of Information Technology, Delhi Technological University, Delhi from 5th June, 2023 - 9th June, 2023.
2. Attended the ‘Delhi Technological University Seminar on Awareness of Engineering Village’ at Delhi Technological University, on 08 March, 2022.
3. Certificate of participation in the online course on “Overview of Geoprocessing Using Python” conducted by IIRS, Dehradun from 18.01.2021- 29.01.2021 (duration: 13 hours and 30 minutes).
4. Certification in The Fundamentals of Digital Marketing by Google Digital Unlocked on 24.05.2020 (duration: 40 hours).
5. Participated in a workshop on Android App Development and Big Data Analytics on 20.03.2018 at COER, Roorkee conducted by DUCAT, Noida.
6. Certification in Life Skills training conducted by GTT and NASSCOM Foundation at COER, Roorkee on 13.09.2017 (Access to Employment certificate).

DECLARATION:

I hereby declare that all the above-mentioned information is true and correct to the best of my knowledge.

Lakshita Agarwal

- *Lakshita Agarwal*