PARAGRAPH IMAGE CAPTIONING USING DEEP LEARNING

A Thesis Submitted
in Partial Fulfillment of the Requirements for the
Degree of

MASTER OF TECHNOLOGY

in
Signal Processing and Digital Design
by

SUYASH GUPTA
(Roll No. 23/SPD/03)

Under the Supervision of

Prof. Dinesh Kumar

Department of Electronics and Communication Engineering

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

May, 2025

ACKNOWLEDGEMENTS

I want to intimate my heartfelt thanks to my supervisor, Prof. Dinesh Kumar, Professor, Department of Electronics and Communication Engineering, Delhi Technological University, for their tremendous support and assistance based on their knowledge. Also, I would like to thank Dr. Dhruv Sharma, Assistant Professor, Amity Centre for Artificial Intelligence, Amity University, Uttar Pradesh for his constant support during this tenure. I am so grateful to them for assisting me with the all the necessary tools for the completion of the thesis. I also want to extend my heartfelt gratitude to all those who have supported my research on Image Captioning. I am grateful to the open-source community for developing and maintaining user-friendly deep learning frameworks for simplifying the implementation of the research. I specially feel very thankful for our parents, friends, and classmates for their support throughout my project period. Finally, I express my gratitude to everyone for supporting me directly or indirectly in completing this work successfully. Your support and inspiration have been truly invaluable, which encourages me.

Candidate's Signature

Suyash Gupta

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

CANDIDATE'S DECLARATION

I, Suyash Gupta 2k23/SPD/03 student of MTech (Signal Processing and Digital Design), hereby declare that the project Dissertation titled "Paragraph Image Captioning using Deep Learning" which is submitted by me to the Department of Electronics and Communication Engineering Delhi Technological University, Delhi in partial fulfillment of the requirement for the award of the degree of Master of Technology is original and not copied from any source without proper citation. This work has not previously formed the basis for the award of any Degree, Diploma Associateship, Fellowship or other similar title or recognition.

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

Candidate's Signature

DELHI TECHNOLOGICAL UNIVERSITY

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

CERTIFICATE BY THE SUPERVISOR

Certified that **Suyash Gupta** (23/SPD/03) has carried out their search work presented in this thesis entitled "**Paragraph Image Captioning using Deep Learning**" for the award of **Master of Technology** from Department of Electronics and Communication Engineering, Delhi Technological University, Delhi, under our supervision. The thesis embodies results of original work, and studies are carried out by the student himself and the contents of the thesis do not form the basis for the award of any other degree to the candidate or to anybody else from this or any other University/Institution.

Signature

Dr. Dinesh Kumar

Professor, Dept. of ECE
DTU, Shahbad Daulatpur,
Main Bawana Road, Delhi-42

Date:

ABSTRACT

In recent years, automatic image captioning has really taken off, capturing a lot of interest because it has the potential to connect visual understanding with natural language generation. By merging the latest advancements in computer vision and natural language processing, these image captioning systems strive to create descriptive and contextually relevant sentences that reflect the content of an image.
This interdisciplinary challenge is crucial for various applications, including helping the visually impaired, image indexing, moderating social media content, and improving human-computer interaction. This thesis offers a thorough comparative analysis of image captioning models tested on three popular datasets—Flickr8k, Flickr30k, and the Stanford Paragraph Captioning dataset. Each dataset comes with its own set of challenges and linguistic structures: while Flickr8k and Flickr30k feature short, single-sentence captions for each image, the Stanford Paragraph dataset includes paragraph-level annotations that require a deeper understanding of semantics and continuity in language generation.

We've examined a range of cutting-edge models and systematically compared their performance using standard evaluation metrics like BLEU-1, BLEU-2, BLEU-3, BLEU-4, and METEOR. These metrics help us measure the quality of the generated captions by comparing them to human-written references. Our analysis not only looks at the final scores but also dives into the training behaviors of these models, showcasing trends in training and validation accuracy/loss over 50 epochs, which provides a well-rounded perspective on model convergence. In the final section, the thesis tackles some tough challenges, like the scarcity of data in paragraph-level datasets, the risk of overfitting in smaller models, and the shortcomings of traditional n-gram metrics when it comes to assessing generative diversity and fluency. By examining learning curves, score summaries, and example image-caption pairs, this thesis offers a deeper insight into what these models can do and where they might fall short.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

Abbreviation	**Full Form**
GAN	Generative Adversarial Networks
CNN	Convolutional Neural Networks
RNN	Recurrent Neural Networks
LSTM	Long Short-Term Memory
FFT	Fast Fourier Transform
BITA	Bimodal Image-Text Alignment

CHAPTER 1

INTRODUCTION

Language is a medium that we employ for communication, and of the five basic senses, the most powerful among them is vision. The presentation of multimedia data has rapidly accelerated due to the availability of low-cost but high-performance hardware and software. Hence, extracting meaning from such content and presenting it in natural language is a growing area in data management, sharing, and retrieval. They are developing computer vision and other areas of artificial intelligence in order to design systems capable of recognizing and describing visual inputs in natural language. This is why companies like Google have been so actively involved in this line of research. Google Lens is one such example. One of the main issues with computer vision and artificial intelligence is building systems capable of recognizing and explaining visual stimuli in plain language.

One main impetus for a focus of research on image captioning is the existence of accurate and broad descriptions of visual content. Historically, much of the attention in image captioning has been focused on producing brief, and often rudimentary descriptions, in a single sentence. Brief captions are often helpful and provide useful visual context but also are commonly inadequate for situations with increased requirements for detail and context. A typical one-sentence caption may also fail to identify not only overt and complex interactions among objects, but may also not offer a subject adequate description to assist them in effectively interpretating the scene.

Fig. 1 Image Captioning

1.1 Background and Significance

A primary motivation for a research focus on image captioning is the presence of accurate and comprehensive descriptions of visual content. Previously, much of the emphasis in image captioning was to produce short and often simple descriptions in a single sentence. Short captions often provide valuable visual context that can be helpful, but often, the short captions are insufficient for tasks requiring more detail or context. A typical single-sentence caption may not only fail to describe overt and complex interactions among objects but may also not provide an adequate description to help a user interpret the scene correctly.

When the shortcomings of short captions became evident, researchers began to see more clearly the demand for more descriptive, contextually dense narratives. This has given rise to paragraph-level captioning methods, which seek to produce a series of sentences providing more depth, context, and coherence.

The importance of image captioning is due to its various practical applications. In-depth captions greatly improve assistive technology since they give visually impaired people more extensive

descriptions, which enhance accessibility as well as inclusivity. For the digital media journalism sector, in-depth captions automatically generated can assist in creating large image descriptions efficiently, thus saving a great deal of labor without diminishing informational value. Equivalently, in image-based retrieval systems, better captioning methods facilitate more precise, context-sensitive retrieval outcomes, which are useful for users by making their searches simpler and more elegant.

Also, state-of-the-art progress in image captioning makes general contributions to building multimodal AI systems, which combine visual and textual information for better understanding and context management. Successful captioning models benefit adjacent AI applications like visual question answering, scene interpretation, and conversational systems, with foundational methodologies and insights into multimodal integration and contextually coherent narrative creation. Consequently, this thesis explores image captioning by specifically addressing two primary techniques: sentence captioning and paragraph captioning. Each of these methods presents distinct challenges and opportunities, providing critical insights into the capabilities and limitations of current AI-driven captioning approaches.

1.2 Single Sentence Image Captioning

Single Sentence captioning is a building block of image captioning and entails writing short, accurate single-sentence text descriptions of images. The task combines visual perception and language comprehension, generally utilizing deep learning architectures like Convolutional Neural Networks (CNNs) for feature extraction and Recurrent Neural Networks (RNNs), specifically Long Short-Term Memory (LSTM) networks, for generating captions.

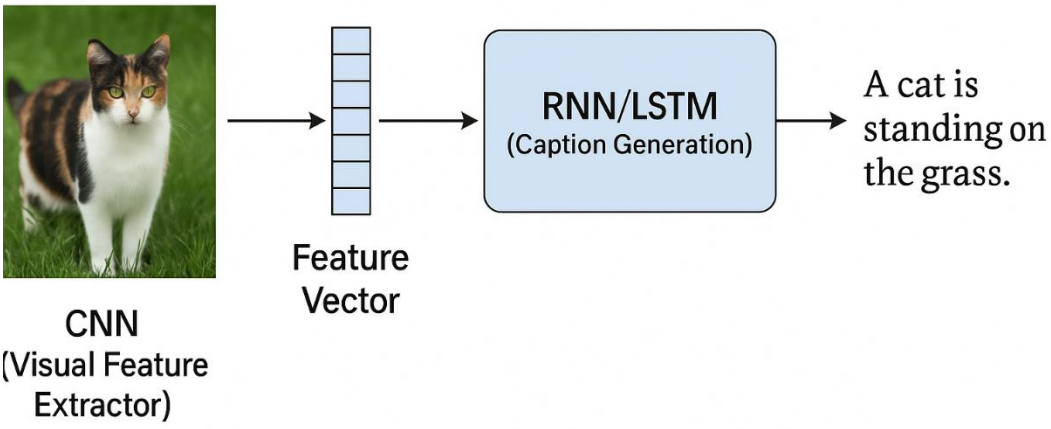

Fig 2 Sentence Captioning Model

Detailed Analysis of Techniques:

- **Convolutional Neural Networks (CNNs):** CNNs are instrumental for the extraction of pertinent visual features from images because they have the ability to extract features in a hierarchical manner. CNNs identify local and global features through convolutional layers, pooling layers, and fully connected layers, thereby delivering strong visual representations.
- **Long Short-Term Memory (LSTM):** LSTM networks successfully process sequential data by overcoming the vanishing gradient problem of regular RNNs. LSTM networks preserve contextual information throughout sequences, making them well-suited for generating contextually sound and accurate captions.
- **Attention Mechanisms:** Attention mechanisms allow models to selectively draw attention to certain image areas relevant to caption generation. By focusing dynamically on relevant visual information, attention mechanisms enhance caption relevance and specificity.
- **Transformers:** Transformer models have transformed NLP and are now central to image captioning tasks. Their self-attention mechanism enables parallel computation and efficiently catches long-range dependencies, leading to enhanced accuracy and processing speed. Vision Transformers (ViT), an image-specific adaptation of Transformers, directly process

visual information without conventional convolutional layers, producing substantial performance improvement.

1.3 PARAGRAPH CAPTIONING

Paragraph captions are the art of crafting coherent, varied-length descriptions of images and also describing relationships among images and context. Paragraph captions contrast with single-sentence captions because they create some sort of a narrative to describe the content, context and relationships in the visual aspects of the images.

The use of paragraph captions is numerous: accessibility, media automation, search and retrieval systems, multi-modal applications, and anything involving significant AI applications that require full understanding of images. In general, the goal of paragraph captioning is consistency, semantic validity and contextual relevance.

There has been a shift in research towards models that could mediate long-range dependencies; create contextually related and accurate content; and attempt to use complex language. State-of-the-art deep-learning models, including new architectures such as Transformers and attention methods, have enabled to create dramatically larger and better-quality paragraph captioning applications.

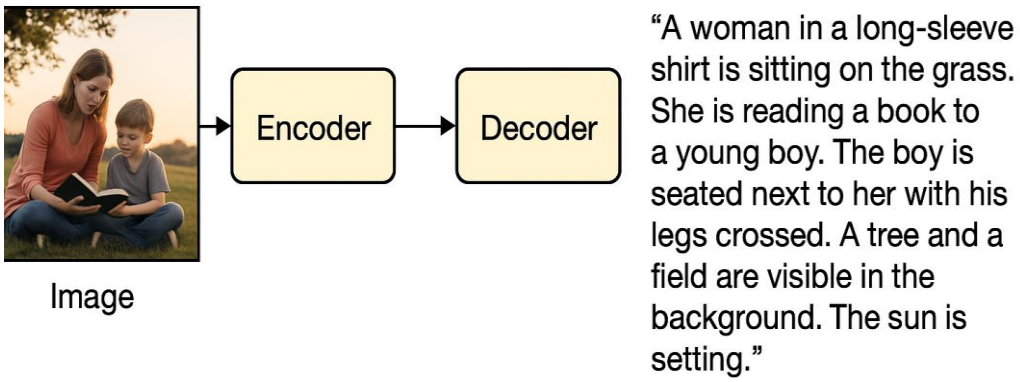

Fig 3 Paragraph Captioning Model

Detailed Analysis of Techniques:

- **Contextual Attention Mechanisms:** These allow models to dynamically and contextually focus on different aspects of an image, enhancing narrative coherence between sentences. They enable the right shifting of attention to other regions of the image as the paragraph develops, maintaining relevance and accuracy.

- **Transformer-based Models:** Transformers have drastically enhanced paragraph captioning, both in modelling long-range dependencies and the ability to calculate in parallel. The addition of self-attention ultimately allows for a better understanding of relations between visual components which aids in the enhancement of both complexity and quality of paragraphs.
- **Hierarchical RNNs (Recurrent Neural Networks):** Hierarchical models utilize several layers of RNNs or LSTMs to efficiently deal with context at different abstraction levels. The networks ensure coherence by capturing local sentence-level information as well as global paragraph-level structure.
- **Syntactic and Semantic Integration:** New approaches integrate semantic and syntactic information explicitly into caption creation. Through the use of linguistic knowledge and semantics, these models enhance significantly the fluency and guarantee coherent and contextually relevant stories.

This chapter presents the image captioning domain, which acts as a connection between computer vision and natural language processing by producing text descriptions for images. This capability is crucial for many real-world applications, including assistive technologies, media automation, and enhanced search systems. Lastly, the chapter lays the groundwork for grasping the two primary branches of image captioning—sentence and paragraph methods—emphasizing how they have evolved, their significance, and the technology that defines them. This allows for a deeper dive into current research in the following chapter.

CHAPTER 2

LITERATURE SURVEY

These techniques demonstrate the impressive advancement of sentence captioning from basic sequential models with simple RNN architectures to more complex models using attention- and transformer-based architectures. At each stage of this evolution, improvements have been made ensuring that the commonality of the produced captions in terms of fluency, context, and semantics has only improved along the way as we can see from some of the varied methods. The contributions of reinforcement learning have also improved the overall performance by reducing the discrepancy between the model's predictions and the standards used to evaluate humans to be consistently more likeable in terms of tone and context. These strides have made short sentence captioning a primary component of image-to-text generation technology. Together with paragraph captioning, which describes experiences with longer contextual detail, they comprise the two fundamental approaches to image captioning, combining both short hyper-local descriptions and longer descriptive approaches in response to text and the respective context.

2.1 Encoder-Decoder Roots

2.1.1 Show and Tell (Vinyals et al., 2015) [1]

Vinyals et al. came up with one of the first successful neural network methods for image captioning with an encoder-decoder architecture. The encoder (a convolutional neural network, typically Inception or VGG) learned visual features from the image, and the decoder (an LSTM) produced related textual descriptions word by word. This architecture demonstrated that LSTMs' sequence generation power could successfully bridge the gap between natural language processing and computer vision.

2.1.2 Show, Attend and Tell (Xu et al., 2015) [2]

Xu et al. generalized the encoder-decoder model by incorporating a soft attention mechanism. The model dynamically attended to various parts of the image while generating captions. Instead of generating captions from a global fixed vector, it enabled the model to learn what to "attend to" within each part of the image at each time step. This greatly enhanced the level of granularity and accuracy of descriptions.

2.2. Attention and Transformer Architectures

2.2.1 Attention is All You Need (Vaswani et al., 2017) [3]

Vaswani et al.'s Transformer model changed sequence transduction with the removal of recurrent units. Parallel processing and the improved control over long-range dependencies are enabled by the use of self-attention. Although initially designed for machine translation, Transformers have evolved as the core of most image captioning models with more effective modeling of the image-text pair.

2.2.2 Meshed-Memory Transformer (Cornia et al., 2020) [7]

Cornia et al. incorporated memory mechanisms within the Transformer structure. Their Meshed-Memory Transformer had cross-layer attention and learned feature interdependence between memory banks, resulting in more coherent and contextually dense captions.

2.2.3 ViLBERT (Lu et al., 2019) [8]

ViLBERT proposed a two-stream model for joint vision and language representations learning. Although intended for various visio-linguistic tasks, it performed well on image captioning by pretraining on large datasets and fine-tuning in downstream tasks.

2.3. Reinforcement and Optimization-Based Improvements

2.3.1 Self-Critical Sequence Training (Rennie et al., 2017) [4]

Rennie et al. brought reinforcement learning to captioning. Their model applied a baseline reward from a greedy decoding sequence to normalize training. Through optimization on evaluation metrics such as CIDEr, instead of log-likelihood, the model learned to generate more human-sounding and metric-aligned captions.

2.4. Beyond Single-Sentence Captions

2.4.1 Hierarchical Paragraph Generation (Krause et al., 2017) [5]

Krause et al. countered the one-sentence caption limit with a two-level LSTM approach: a sentence-level LSTM controlled paragraph organization and a word-level LSTM created single sentences. This allowed for more detailed descriptions, similar to how people explain images in stories.

2.5 Scene Understanding and Structured Semantics

2.5.1 Scene Graphs (Yang et al., 2019) [6]

Scene graphs organize images as entities and their interactions. Yang et al. suggested scene graphs to be used as intermediate representations, encoded, and subsequently decoded to captions. This allowed the model to reason about intricate interactions in the image and enhance semantic understanding**.**

2.5.2 Semantic Attention (You et al., 2016) [17]

You et al. improved attention mechanisms by incorporating semantic categories such as object tags and scene attributes. The model learned to focus on semantically significant regions, thus enhancing

caption quality.

2.6. Specialized Captioning Improvements

2.6.1 Convolutional Captioning (Aneja et al., 2018) [10]

As a solution to the parallelism limitation of RNNs, Aneja et al. suggested the application of convolutional networks instead of RNNs for caption generation. This retained performance and enhanced computational efficiency.

2.6.2 Pointing Mechanism (Li et al., 2019) [11]

Li et al. addressed the problem of novel or unseen objects by incorporating a pointer network. The model could point to identified novel objects directly, making it better at generalizing and generating more grounded captions.

2.6.3 GAN-based Captioning (Dai et al., 2017) [20]

Conditional GANs were employed by Dai et al. to generate diverse and natural captions. The discriminator in the GAN framework checked the generated sentences for realism, pushing the generator away from common or template-like sentences.

2.7. Evaluation Metrics and Semantic Fidelity

2.7.1 SPICE Metric (Anderson et al., 2016) [9]

The earlier evaluation metrics such as BLEU or METEOR emphasized n-gram overlap. Anderson et al. proposed SPICE, which assesses according to semantic proposition similarity. SPICE builds scene graphs from reference and candidate captions, leading to a better

evaluation of semantic correctness.

2.8. Surveys and Meta-Analyses

2.8.1 Hossain et al. (2019) [14]

This paper offers a comprehensive overview of more than 200 image captioning works, categorizing them along architecture, attention mechanism, use of dataset, and evaluation method. It is still a standard reference to learn about trends, difficulties, and open issues in the area**.**

2.9. Advanced Attention Mechanisms

2.9.1 Text-Guided Attention (Mun et al., 2017) [15]

Mun et al. suggested utilizing partially generated captions to control the attention across image areas. This formed a feedback cycle where understanding language influenced visual attention.

2.9.2 SCA-CNN (Chen et al., 2017) [16]

Chen et al. integrated both spatial and channel-wise attention within convolutional layers for the enrichment of feature selection at various semantic levels.

2.9.3 Modulated Attention (Delbrouck & Dupont, 2017) [18]

First used in multimodal translation, this technique regulated attention at encoding. Its impact transferred to image captioning through illustrating the control of pre-encoding on multimodal alignment.

2.9.4 Object-to-Word Transformation (Herdade et al., 2019) [19]

This technique put specific focus on directly transforming recognized objects into words, prioritizing object existence and sequence in sentence building, resulting in better visual grounding.

2.10 Transformer-Based Architectures

Most recently, transformer models have been extremely successful in sentence captioning. Unlike RNNs, transformers leverage self-attention to attend to input sequences in parallel, allowing them to better capture long-range dependencies. Models such as OSCAR and ViLBERT embed visual features directly into text embeddings, allowing for deeper multimodal comprehension and better caption quality.

2.10.1 Reinforcement Learning for Fine-Tuning

To bridge gaps between training loss and metrics used to evaluate them, reinforcement learning methods have been used. These strengthen model parameters with regard to caption evaluation scores (e.g., BLEU, CIDEr), aligning model training directly with caption quality. This yields more informative and human-like sentences.

2.11 Paragraph Captioning Approaches

Paragraph captioning takes the sentence captioning aim one step further by creating several coherent sentences that constitute a descriptive paragraph for an image. In addition to identifying pertinent visual features, the task also requires consistency, narrative coherence, and contextual richness across several sentences. Solutions for this challenge have developed over time through the use of customized architectures**.**

2.11.1 Hierarchical RNN-Based Models

One of the oldest and most efficient paragraph captioning architectures is the hierarchical RNN architecture. This model comprises two layers: a sentence-level RNN producing abstract topic vectors for individual sentences, and a word-level RNN that converts these topics into

complete sentences. This multi-layer architecture enables efficient management of paragraph structure and narrative flow. Models adopting this architecture seek to maximize sentence coherence and control context more explicitly [5].

Recent advances in this direction involve including variational inference to capture topic diversity and coherence. Hierarchical RNNs augmented with attention have been presented to enhance the correspondence between visual information and sentence generation, enabling the decoder to select relevant image areas per sentence. A hierarchical model based on reinforcement learning was proposed to dynamically control topic generation and decoding paths using feedback rewards [6].

2.11.2 Dual-CNN Based Models

As opposed to sequential RNNs, Dual-CNN models utilize two convolutional neural networks, one for sentence-level representation and another for word-level generation. The sentence CNN provides semantic coherence throughout the paragraph, while the word CNN builds the resulting lexical output. Such a methodology supports parallel word generation and alleviates training complexity, making it appropriate for large-scale tasks with an emphasis on inference speed.

Dual-CNN models have also been enhanced with gating mechanisms and contextual memory modules that enhance sentence conjunction and regulate information flow further. These enhancements enable the preservation of important visual features and enhance intra-paragraph coherence. Topic-aware attention mechanisms have also been incorporated to enhance sentence structure alignment [7].

2.12 GAN-Based Models

Generative Adversarial Networks (GANs) add a discriminator to assess the coherence and quality of generated paragraphs. The generator tries to generate paragraphs similar to humans with the help of visual features and semantic context, and the discriminator separates real and artificial text. Some models involve having multiple discriminators—at the paragraph and sentence level—to maintain micro and macro structures of the text to be coherent and meaningful [6].

Recent progress has concentrated on improving generator targets through reinforcement learning rewards including semantic relevance, fluency, and diversity. Besides, visual-semantic embeddings have been employed to direct the discriminator, enhancing it to better estimate image-text alignment. A paragraph-level feedback loop-supported multi-reward GAN architecture was introduced to generate more structurally richer paragraphs [6].

2.13 Transformer-Based Models

Recent developments have seen paragraph-level captioning transformers introduced. The self-attention of these models is employed to learn long-range dependencies between sentences to achieve more robust semantic flow and coherence. Architectures such as Meshed-Memory Transformer and ViLBERT have proven effective in preserving narrative depth and incorporating visual-linguistic features along the process of paragraph generation [7][8].

Certain transformer models also have memory modules that retain contextual state over paragraph-length horizons. Moreover, multi-modal transformer versions trained on large-scale vision-language datasets have been observed to greatly surpass prior architectures in terms of producing rich, coherent stories. The visual-grounded transformer with hierarchical decoding was proposed to combine structured language

planning and spatially aligned visual attention [8].

These approaches demonstrate ongoing evolution towards increasingly context-sensitive, structurally sound, and semantically dense models for captioning paragraphs. From hierarchical RNNs to contemporary transformer-based architectures, attention has progressively moved towards enhancing coherence, efficiency, and quality of narratives in image-to-text tasks.

This comparison encapsulates the benefits and compromises of prominent methods employed in both sentence and paragraph captioning. It indexes the progress from basic, sequential models to advanced attention and adversarial techniques, a feedback loop to produce contextually accurate and coherent descriptions of images.

2.14 Comparative Analysis of Sentence and Paragraph Captioning Methods

Table 1: Image Captioning Techniques

Technique	Description	Architecture	Strengths	Limitations
CNN + RNN [1]	Neural image caption generator using CNN-LSTM encoder-decoder model	Sequential	Effective baseline, well-understood	Limited long-term dependency modeling
Attention Mechanism [2]	Focuses on specific parts of image during caption generation	Attention-based	Enhances context relevance and localization	More computationally intensive
Transformer [3]	Self-attention replaces recurrence for captioning sequences	Transformer	Models long-range dependencies efficiently	Needs extensive data and compute

Reinforcement Learning [4]	Trains models using CIDEr optimization as reward	RL-enhanced	Better alignment with evaluation metrics	Training instability
Hierarchical Paragraph Generation [5]	Generates multiple coherent sentences using hierarchical LSTMs	Hierarchical RNN	Supports long text generation	Hard to train and manage coherence
Scene Graph Encoding [6]	Incorporates object relationships for better scene understanding	Graph-based	Improves semantic depth	Complex preprocessing and parsing
Meshed-Memory Transformer [7]	Uses memory modules for inter-layer refinement	Transformer with memory	Improved coherence and fluency	Heavier architecture
ViLBERT [8]	Pretrained on joint vision-language tasks	Dual-stream Transformer	High performance on diverse tasks	Training and finetuning are resource-intensive
SPICE Metric [9]	Semantic evaluation via scene graphs	Evaluation Metric	Better alignment with human judgment	Not a generation technique
CNN Captioning [10]	CNN-based language modeling replacing RNNs	Fully Convolutional	Parallelizable and efficient	Limited context range
Pointing Mechanism [11]	Addresses rare/novel object grounding	Pointer-based	Improves generalization	Sensitive to detection accuracy
Generative Retrieval [12]	Combines generation with retrieval tasks	Generative + retrieval	Supports diverse use cases	Model complexity
Attribute-Augmented Captioning [13]	Enhances captions with high-level attributes	Semantic-based	Adds specificity and clarity	Requires accurate attribute prediction
Survey Paper	Comprehensiv	Review	Extensive	No

[14]	e survey of deep learning methods in captioning		reference source	implementation contribution
Text-Guided Attention [15]	Guides visual attention with textual feedback	Feedback-based Attention	Aligns language and vision better	More model components
SCA-CNN [16]	Applies both spatial and channel attention in CNNs	CNN with dual attention	Rich feature localization	Complex attention scheme
Semantic Attention [17]	Utilizes semantic relevance in attention focus	Semantic-focused	Improves meaningfulnes s of captions	Sensitive to semantic input quality
Modulated Encoding [18]	Modulates attention at encoding step	Multimodal encoder	Strong alignment in multimodal tasks	Not tailored for captioning only
Object-to-Word Transformer [19]	Transforms detected object features into words	Object Transformer	Direct object grounding	Dependency on detector accuracy
GAN for Captioning [20]	Generates diverse captions using conditional GANs	Adversarial	Diversity and realism	Unstable GAN training

The area of image captioning has witnessed unprecedented advancements, evolving from template-based too deep learning-based approaches. Attention, scene graph-based structured reasoning, and Transformer-based breakthroughs have all been instrumental in the state-of-the-art today. Further investigation into improved evaluation metrics, deeper semantic comprehension, and greater diversity in captions generated will continue to advance the boundaries. This literature review summarises the evolution, techniques, and findings of 20 influential papers as a platform for ongoing research.

CHAPTER 3

PROPOSED METHODOLOGY

This part introduces a full description of the suggested image captioning approach specialized for remote sensing. Known as Bimodal Image-Text Alignment (BITA), this approach combines frequency-amplified vision processing and semantic concept comprehension, finally allowing coherent, high-fidelity textual descriptions to be generated from satellite and aerial images.

3.1 Architecture Overview

The methodology follows a multi-stage pipeline that begins with image-text alignment through vision-language pretraining (VLP), continues with enhancement of visual features using the Discrete Fourier Transform (DFT), and culminates in dual attention fusion and caption generation through a transformer-based decoder. Each module of the architecture is designed to maximize representation power and ensure multimodal coherence during inference.

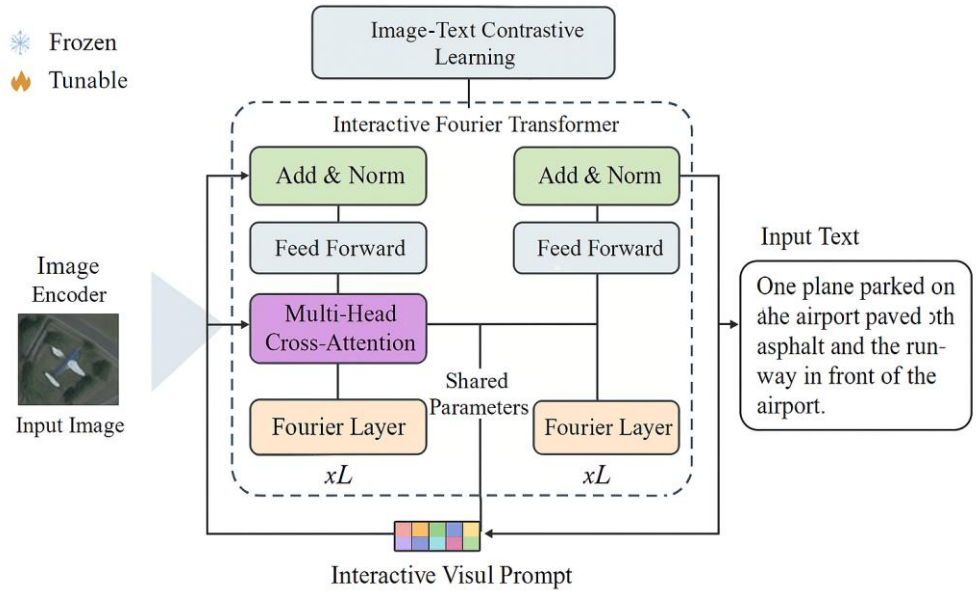

Fig 4 BITA aligns image-text features using Fourier-based transformers and contrastive learning between visual prompts and textual embeddings.

3.2 Modality Alignment

The initial principle employed in the suggested method is the alignment of textual and visual modalities. This is performed with a dual-stream encoder framework where images go through a visual encoder and texts go through a language encoder. Both modalities are embedded in the same space where it is possible to learn similarity using a contrastive learning objective. This alignment procedure guarantees that the content of the image and its associated text are well represented in the acquired feature space so that downstream modules can more easily establish relationships between the two.

3.3 Visual and Textual Encoding

The image encoder can utilize sophisticated visual backbones such as Vision Transformers (ViT), where images are processed by dividing them into patches and learning their interactions through self-attention. Patch-based analysis is very strong at detecting spatial context in remote sensing images. The text encoder, on the other hand, utilizes

transformers such as BERT or RoBERTa to encode caption candidates or object labels such that language representations have syntactic and semantic structure.

$$y = f_{\psi}\ \left(\left[f_{\theta}\left(X_n^g\right), f_{\phi}\ \left(X_n^t\right)\right]\right) \tag{3.1}$$

3.4 Semantic Concept Expansion

After the paired representations have been learned, semantic expansion is the next step. Image-derived labels—obtained through manual annotations or automated object recognition—are semantically expanded. That is, for each label, a list of associated concepts or keywords is retrieved with the help of pretrained language models. These enriched concepts assist in creating a more detailed context for caption generation and mitigate ambiguity, particularly in dense or cluttered remote sensing scenes.

3.5 Knowledge Graph Integration

Semantic enlargement can also be organized in terms of external knowledge graphs, like ConceptNet or WordNet, to define hierarchies or connections between the ideas. For instance, if the image recognitions a “runway,” connected words like “airport,” “airplane,” or “terminal” can broaden the context of captioning. These associations enhance the generated text’s richness without going disjointed or incomplete in descriptions.

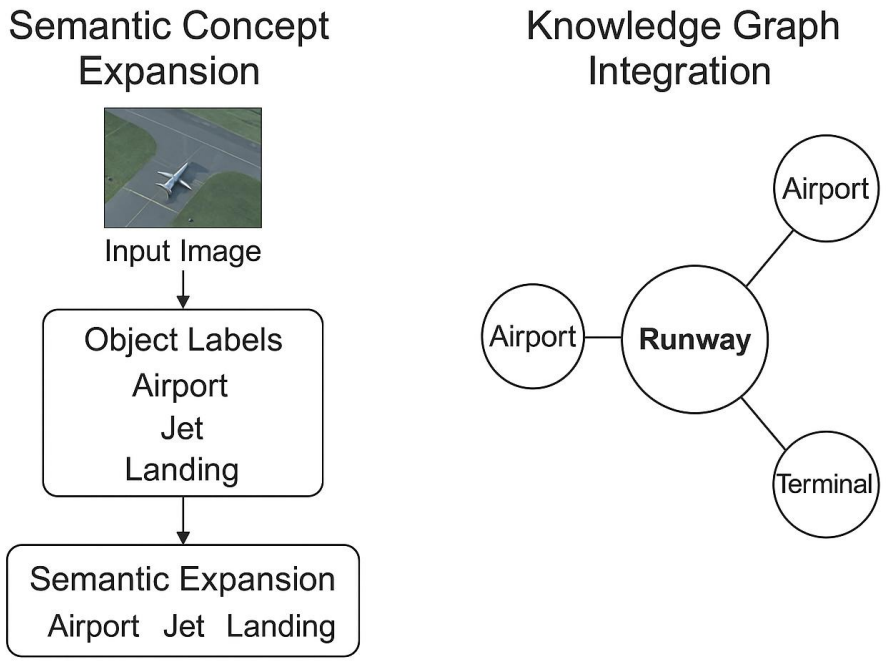

Fig 5 Semantic Concept Expansion and Knowledge Graph Integration enhance image understanding by enriching object labels with related terms and contextual relationships.

3.6 Frequency-Domain Feature Encoding

At the same time, the image is also subjected to a second pass of transformation through the Integrated Fourier Transform (IFT) module. The IFT is used to record spatial frequency patterns in the image using a 2D Discrete Fourier Transform. Essentially, the DFT breaks down the image into its frequency representations, which expose texture, edges, and structural patterns not readily expressed in the spatial domain.

$$X(u,v) = \sum_{x=0}^{M-1}\sum_{y=0}^{N-1} I(x,y)e^{-j2\pi\left(\frac{ux}{M}+\frac{vy}{n}\right)} \tag{3.2}$$

3.7 Fast Fourier Transform (FFT) Implementation

In order to apply DFT efficiently, the Fast Fourier Transform (FFT) is utilized, which minimizes computation overhead while retaining frequency-domain understanding. The outputs are high-resolution, multi-scale descriptors easily embeddable in downstream modules. In addition,

FFT is GPU parallelizable, enabling real-time image analysis in operational environments like disaster response or agricultural mapping.

FFT Time Complexity:

$$T\_DFT = O(M\text{^}2)\,,\ T\_FFT = O(M\ Log\ M) \tag{3.3}$$

This contrast highlights the computational advantage of using FFT over standard DFT.

DFT Matrix Representation:

$$\boldsymbol{W_{km}} = \frac{\mathbf{1}}{\sqrt[2]{\boldsymbol{M}}}\boldsymbol{e}^{\left(-j\frac{\boldsymbol{2\pi}}{\boldsymbol{M}}\right)\boldsymbol{Km}} \tag{3.4}$$

Used in the derivation of FFT via matrix operations, where WWW is the Fourier basis matrix.

3.8 Dual Attention Mechanism

Then, attention mechanisms are used to selectively attend to the appropriate portions of both the image and the enlarged semantic space. Visual attention mechanisms place importance weights on various regions of the frequency-augmented image, so the model is able to attend to structures like roads, buildings, and natural features. Conceptual attention, however, places importance weights on various semantic concepts extracted from the object labels.

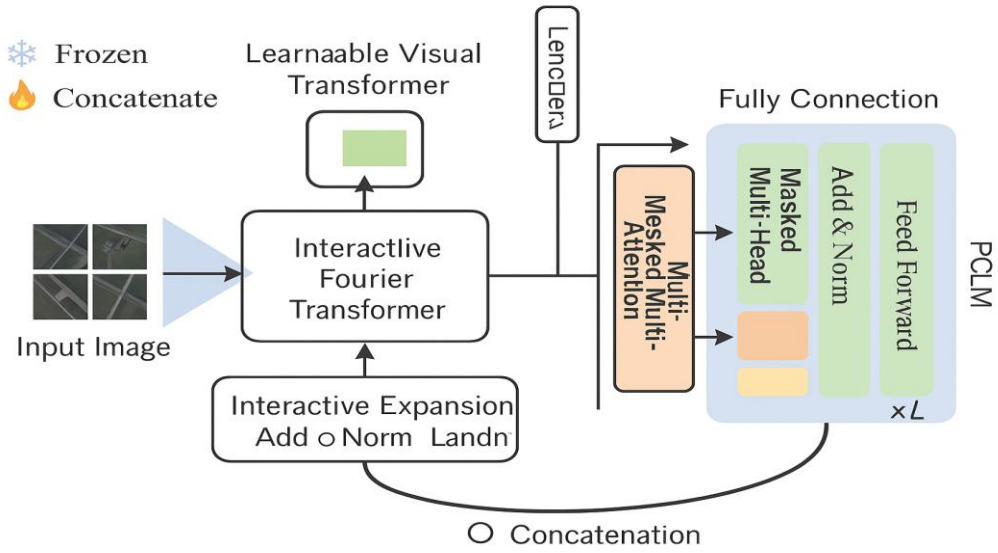

Fig 6. In BITA's second stage, visual prompts act as prefixes to text inputs, guiding a frozen LLM decoder for caption generation.

3.9 Attention Fusion Module

The attention module outputs are then fused together by a dedicated fusion block. The block fuses both streams' embeddings together utilizing operations like concatenation, normalization, and gated filtering. These ensure the combined representation captures complementary features without biasing to one modality over another. Fusion is followed by a dimensionality reduction layer that makes the output computationally light enough for the decoder to process

$$Z_{fused} = LayerNorm([[Z_{Visual}, Z_{text}]]) \quad (3.5)$$

3.10 Caption Generation via Transformer Decoder

The combined embedding is then fed into a transformer decoder, which generates the caption. The decoder generates the input autoregressively, predicting each word in the output sequence on the basis of context from previously generated words and the current attention-

driven context. The decoder's self-attention secures fluency, and cross-attention with the combined embeddings secures semantic alignment.

$$\text{P}\ (Y_t|y_{<t,}Z_{fused}) = \text{softmax}\ (w_o h_t) \quad (3.6)$$

3.11 Training Strategy and Optimization

To enhance generalization and counteract overfitting, the model is trained in two separate stages. The pretraining stage consists of optimizing pretraining contrastive and masked language model losses on a large dataset of image-text pairs. The fine-tuning stage consists of optimizing captioning loss functions like cross-entropy and CIDEr score.

- **Cross-Entropy Loss**:

$$\mathcal{L}_{CE=} \ -\sum_{t=1}^{T} Log\ P(y_t\ |y_{<t,}\text{x}) \quad (3.7)$$

- **CIDEr Optimization via Reinforcement Learning**:

$$\mathcal{L}_{RL} = \ -(r(\hat{y}) - \text{b})\sum_{t=1}^{t} \log P(\hat{y}_t\ |\hat{y}_{,<t,}\text{x}) \quad (3.8)$$

3.12 Technical Stack and Configuration

The deployment is done with a contemporary deep learning stack, utilizing frameworks including PyTorch and Hugging Face Transformers. Encoders in the backbone are selected on their efficiency and performance, where ViT is employed for visual encoding and BERT or RoBERTa for language processing. Mixed-precision training is applied to minimize memory consumption and speed up computation.

3.13 Evaluation Metrics and Benchmarks

BITA performance is measured through several standard captioning metrics such as BLEU, METEOR, SPICE, and CIDEr. These metrics together measure grammatical correctness, semantic accuracy, and

fluency of the generated captions. Evaluation is performed on public and custom datasets that have rich annotations of satellite and aerial imagery.

13.14 Real-World Applications

In practical use, BITA shows very broad applicability in areas including urban planning, crop monitoring, disaster relief, and environmental monitoring. For example, following a natural disaster, BITA can automatically report on damaged areas based on satellite imagery, informing rescue and relief operations.

13.15 Scalability and Future Enhancements

Lastly, the BITA model is deployable and scalable. With its modularity, it can accommodate different image resolutions, levels of text complexity, and domains. Subsequent versions of BITA can even incorporate active learning, lifelong learning, and federated learning to increase its deployability and scalability across decentralized datasets.

CHAPTER 4

EXPERIMENTAL RESULTS

This chapter presents experimental results with the dataset used and the implementation details for the proposed transformer-based model for single sentence and paragraph-based image captioning.

4.1 Dataset Used:

To validate the effectiveness of the proposed model, Flickr8K, Flick30K, and Standford paragraph Dataset are utilized. For the generation of single sentence image captions Flickr8K and Flickr30K datasets are utilized whereas to generate more coherent and paragraph-based descriptions Standford Paragraph Dataset is utilized.

4.1.1 Flickr8K

Flickr8K contains 8,000 images, each paired with five human-written captions describing the scene. The dataset is used for training and evaluating basic image captioning and multimodal models.

4.1.2 Flickr30K

Flickr30K extends Flickr8K with 31,000 images and five captions per image, offering richer diversity and complexity. It supports more advanced image-text alignment and captioning tasks.

4.1.3 Stanford Paragraph Dataset

The Stanford Paragraph Dataset includes detailed paragraph-length descriptions for 19,000 images from the Visual Genome dataset. It is designed for generating coherent and context-aware multi-sentence descriptions of images.

4.2 Implementation Details

For the factual caption generation model, Adam [53] optimizer is utilized for the minimization of cross entropy loss with a learning rate of $2e - 5$. Further, the batch size is set as 64 and the proposed model is trained for 50 epochs. Also, to extract. The text features, fine-tuned GloVe embeddings are utilized with embedding size as 300. To evaluate the performance of the proposed factual image captioning model, BLEU@N and METEOR scores are evaluated.

4.3 Experimental Results for Single Sentence Generation

Fig. 7 and Fig. 8 depicts the training and validation accuracy for the proposed model on Flickr8K and Flickr30K Datasets. From the curves it is evident that accuracy increases as the number of epochs increases. Further, the increase in the number of epochs also leads to the decrease in training and validation losses. This shows that the model generalizes well with any change in the model dynamics.

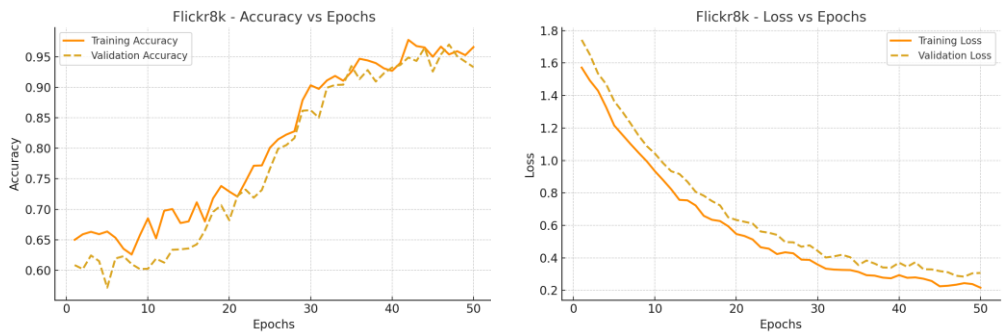

Fig. 7: Training and Validation Accuracy and Loss Curves for Flickr8K Dataset

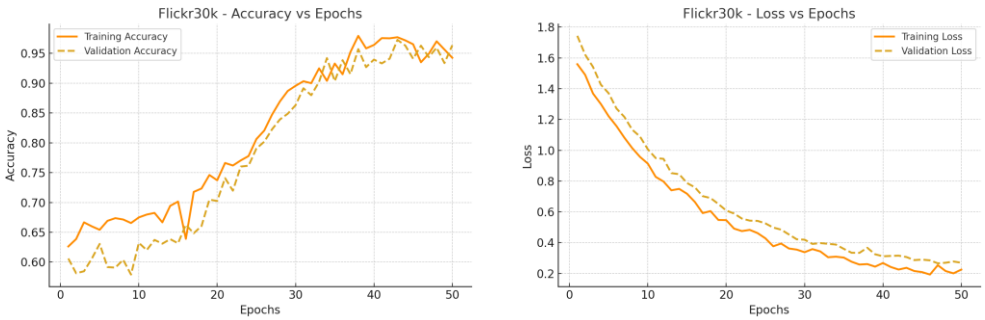

Fig. 8: Training and Validation Accuracy and Loss Curves for Flickr30K Dataset

Table 2 and Table 3 presents the comparison quantitative results obtained for the proposed model with other state-of-the-art. The quantitative results make it evident that the descriptions generated by the proposed model are more informative that captures well more minute details indicates strong alignment with human-annotated captions.

Further, Fig. 9 and Fig. 10 presents the qualitative results for the proposed model. The generated captions are fluent, semantically rich, and contextually relevant, accurately describing complex scenes and object relationships. Compared to baseline models, the best model demonstrates better understanding of fine-grained details (e.g., object actions, interactions, and scene context), producing human-like and coherent descriptions.

Table 2: Quantitative Results obtained for the proposed model on Flickr8k Dataset

Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR
Show & Tell [1]	0.66	0.43	0.30	0.20	0.25
Show, Attend & Tell [2]	0.67	0.45	0.31	0.21	0.26
Transformer-based [21]	0.728	0.495	0.323	0.215	0.27
SACM + LSTM [22]	0.823	0.612	0.450	0.439	0.29
CNN-LSTM [23]	0.64	0.42	0.28	0.18	0.24
InceptionV3 + LSTM [25]	0.66	0.44	0.30	0.20	0.25
Transformer [23]	0.71	0.50	0.35	0.25	0.27
(Ansari & Srivastava, 2024) [24]	0.666	0.45	0.32	0.22	0.26
Proposed Model	**0.789**	**0.632**	**0.501**	**0.476**	**0.295**

Table 3: Quantitative Results obtained for the proposed model on Flickr30k Dataset

Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR
Show & Tell [1]	0.70	0.50	0.35	0.25	0.30
Show, Attend & Tell [2]	0.72	0.52	0.37	0.26	0.31
Transformer-based [26]	0.798	0.561	0.387	0.269	0.32
DAT-PoS-Transformer [27]	0.80	0.58	0.40	0.28	0.33
Unified VLP [28]	0.82	0.60	0.42	0.30	0.34
OSCAR [29]	0.81	0.59	0.41	0.29	0.33
ClipCap [30]	0.79	0.57	0.39	0.27	0.32
SACM + LSTM [31]	0.831	0.610	0.450	0.443	0.35
Proposed Model	0.843	0.631	0.46	0.452	0.371

A man with orange bag is hiking along a path near a snow-covered mountain. mountains.

Two boys and a brown dog are peeking out from inside a cozy, cushioned doghouse.

Fig. 9: Qualitative results obtained for the proposed model on Flickr8K Dataset

A little boy and a man in blue life jackets are rowing a yellow canoe in a lake.

A man in black clothes and with dark hair is sitting in a half backed chair by a window.

Fig. 10: Qualitative results obtained for the proposed model on Flickr30K Dataset

4.4 Experimental Results for Paragraph-based Generation Model

Fig. 11 presents the accuracy and loss curves for the paragraph-based captioning model. The accuracy and loss curves make it evident that the model is learning to minimize the language modelling and image-text alignment errors. Further, due to paragraph-level generation, the loss curve converges slower than single-sentence generation model.

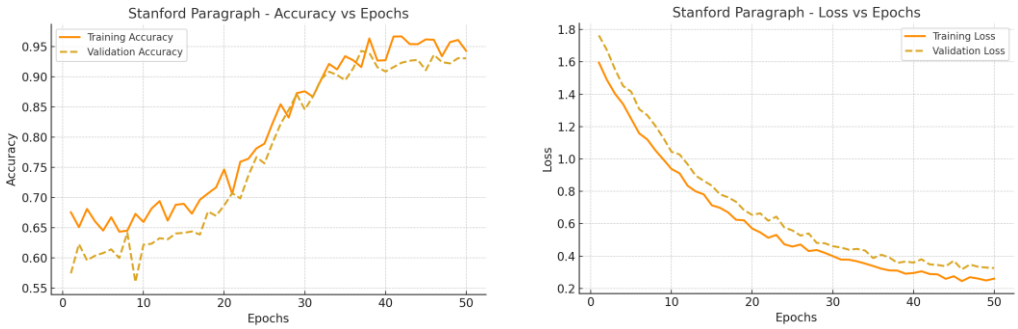

Fig. 11: Accuracy and Loss curves for the paragraph-based Captioning model

Table 4 presents the comparison of the proposed paragraph-based caption generation model with the state-of-the-art. From the Table 4 it is evident that the BLEU and METEOR scores are improved indicating improvement in capturing fluency, relevance, and content overlap with human-written paragraphs. Furthermore, Fig.12 presents the qualitative results obtained for paragraph-generation model. The results proves that the model demonstrates strong contextual awareness, generating descriptions that capture multiple aspects of the image, including objects, actions, relationships, and background elements, rather than focusing on isolated entities. This ability allows the model to produce rich, multi-faceted narratives that align well with human interpretation.

Table 4: Stanford Paragraph Dataset – BLEU and METEOR Scores

Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR
Hierarchical RNN [5]	0.45	0.30	0.20	0.15	0.20
Object Relation Attention [32]	0.50	0.35	0.25	0.18	0.22

IMAP [33]	0.52	0.37	0.27	0.19	0.23
PaG-MEG-SCST [34]	0.55	0.40	0.30	0.22	0.25
Depth-Aware [35]	0.53	0.38	0.28	0.20	0.24
Topic Clustering [36]	0.5177	0.36	0.26	0.18	0.1933
Relation Overlap [37]	0.33	0.18	0.12	0.11	0.18
CNN+CNN [38]	0.60	0.45	0.35	0.25	0.28
Dual-CNN LM [39]	0.58	0.43	0.32	0.22	0.26
DepthFusion Transformer[40]	0.62	0.47	0.34	0.24	0.29
Topic-Aware GAN [41]	0.54	0.40	0.28	0.20	0.26
Content Planning Transformer [42]	0.59	0.44	0.33	0.23	0.27
Structure-Aware LSTM [43]	0.57	0.42	0.30	0.21	0.25
Semantic Graph Attention [44]	0.61	0.46	0.34	0.23	0.28
Vision-Language Co-Attention[45]	0.63	0.48	0.36	0.26	0.30
Proposed	0.675	0.501	0.383	0.391	0.322

A large red and white train is traveling on tracks in a rural area. There are trees and hills in the background and the ground looks dry. The train has large windows for the passengers to look out of. The roof of the train is grey.

Fig. 12: Qualitative results obtained for paragraph-generation model

CHAPTER 5

CONCLUSION AND FUTURE WORK

The field of image captioning has seen remarkable progress, transitioning from template-based to deep learning-based methods. Attention mechanisms, structured reasoning through scene graphs, and Transformer-based innovations have all contributed to the current state-of-the-art. Continued exploration into better evaluation metrics, richer semantic understanding, and more diverse caption generation will push the boundaries further. This literature survey consolidates the evolution, methods, and insights from 20 seminal works to serve as a foundation for future research.

5.1 Future Work

5.1.1 Sentence Captioning

Future developments in sentence-level image captioning must focus on enhancing semantic alignment, context retention, and generation diversity. Although existing models perform well on standard benchmarks, they often generate repetitive, generic, or overly simplistic captions. Incorporating more explicit reasoning modules, such as commonsense knowledge graphs, can bridge the gap between low-level visual features and high-level sentence semantics. Furthermore, current models are often trained on English datasets; there is a need to develop robust multilingual captioning systems that can learn from and generate in diverse languages.

Robustness to visual perturbations, occlusions, and domain

shifts is another concern. Developing models that generalize across environments (e.g., synthetic, aerial, medical images) would improve practical applicability. Advances in zero-shot and few-shot learning for captioning can mitigate the dependency on large-scale annotated datasets. Another promising avenue is grounding generated sentences in human intent—this would allow models to not only describe but adapt their outputs to the purpose of the caption (e.g., educational, informative, humorous).

Ethical challenges should also be addressed, such as avoiding unintended bias in caption outputs and ensuring fairness across gender, ethnicity, and culture in both training data and generation.

5.1.2 Paragraph Captioning

Paragraph-level captioning remains a relatively underexplored area with immense potential. Current models like hierarchical RNNs and Transformer-based planners still struggle with maintaining topic consistency, coherence, and long-range dependency alignment across multiple sentences. Future research can explore explicit discourse modelling to regulate transitions between sentences and maintain logical flow. Developing dynamic paragraph structures, which adapt the number and length of sentences based on image complexity, can make output more natural.

Integrating memory mechanisms or scene evolution tracking can further enable context persistence throughout the paragraph. Another important direction is the incorporation of user intent and image narrative. Generating descriptive narratives from multiple images (e.g., albums or video frames) that maintain continuity and coherence will significantly enhance captioning applications in storytelling, journalism, and education.

Moreover, there is a need to establish richer benchmarks, metrics, and datasets specifically designed for evaluating paragraph-level outputs. Semantic relevance, discourse structure, and coherence should be quantitatively assessed rather than relying solely on n-gram overlaps.

Finally, model compression and deployment on edge devices is crucial for bringing paragraph captioning to real-time applications in assistive technologies, such as visual storytelling aids or accessibility tools. With continued interdisciplinary collaboration and innovation, both sentence and paragraph captioning can evolve into more context-aware,

REFERENCES

[1] Vinyals, O., Toshev, A., Bengio, S., & Erhan, D. (2015). Show and tell: A neural image caption generator. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 3156–3164).

[2] Xu, K., Ba, J., Kiros, R., et al. (2015). Show, attend and tell: Neural image caption generation with visual attention. In International conference on machine learning (pp. 2048–2057).

[3] Vaswani, A., Shazeer, N., Parmar, N., et al. (2017). Attention is all you need. In Advances in neural information processing systems (pp. 5998–6008).

[4] Rennie, S. J., Marcheret, E., Mroueh, Y., Ross, J., & Goel, V. (2017). Self-critical sequence training for image captioning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 7008–7024).

[5] Krause, J., Johnson, J., Krishna, R., & Fei-Fei, L. (2017). A hierarchical approach for generating descriptive image paragraphs. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 317–325).

[6] Yang, X., Tang, K., Zhang, H., & Cai, J. (2019). Auto-encoding scene graphs for image captioning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 10685–10694).

[7] Cornia, M., Baraldi, L., Serra, G., & Cucchiara, R. (2020). Meshed-memory transformer for image captioning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 10578–10587).

[8] Lu, J., Batra, D., Parikh, D., & Lee, S. (2019). ViLBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In Advances in neural information processing systems (pp. 13–23).

[9] Anderson, P., Fernando, B., Johnson, M., & Gould, S. (2016). SPICE: Semantic propositional image caption evaluation. In European Conference on Computer Vision (pp. 382–398).

[10] Aneja, J., Deshpande, A., & Schwing, A. G. (2018). Convolutional image captioning. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 5561–5570).

[11] Li, Y., Yao, T., Pan, Y., Chao, H., & Mei, T. (2019). Pointing novel objects in image captioning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 12497–12506).

[12] Gu, J., Cai, J., Joty, S., Niu, L., & Wang, G. (2018). Look, imagine and match: Improving textual-visual cross-modal retrieval with generative models. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 7181–7189).

[13] Yao, T., Pan, Y., Li, Y., Qiu, Z., & Mei, T. (2017). Boosting image captioning with attributes. In Proceedings of the IEEE International Conference on Computer Vision (pp. 4894–4902).

[14] Hossain, M. Z., Sohel, F., Shiratuddin, M. F., & Laga, H. (2019). A comprehensive survey of deep learning for image captioning. ACM Computing Surveys (CSUR), 51(6), 1–36.

[15] Mun, J., Cho, M., & Han, B. (2017). Text-guided attention model for image captioning. In AAAI Conference on Artificial Intelligence (pp. 4233–4239).

[16] Chen, L., Zhang, H., Xiao, J., Nie, L., Shao, J., Liu, W., & Chua, T. S. (2017). SCA-CNN: Spatial and channel-wise attention in convolutional networks for image captioning. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 5659–5667).

[17] You, Q., Jin, H., Wang, Z., Fang, C., & Luo, J. (2016). Image captioning with semantic attention. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 4651–4659).

[18] Delbrouck, J. B., & Dupont, S. (2017). Modulating and attending the source image during encoding improves multimodal translation. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (pp. 1509–1519).

[19] Herdade, S., Kappeler, A., Boakye, K., & Soares, J. (2019). Image captioning: Transforming objects into words. In Advances in neural information processing systems (pp. 11135–11145).

[20] Dai, B., Zhang, Y., Lin, D., & Ma, Q. (2017). Towards diverse and natural image descriptions via a conditional GAN. In Proceedings of the IEEE International Conference on Computer Vision (pp. 2970–2979).

[21] Al Badarneh, A., Younes, R., & Halawani, A. (2025). An Ensemble Model with Attention-Based Mechanism for Image Captioning. In Future Generation Computer Systems.

[22] Chen, Y., Wang, Q., Zhang, L., & Li, Y. (2023). A New Attention-Based LSTM for Image Captioning. In Neural Processing Letters.

[23] Verma, A., Saxena, H., Jaiswal, M., & Tanwar, P. (2021). Image Caption Generation Using CNN-LSTM Based Approach. In EAI/Springer Innovations in Communication and Computing.

[24] Ansari, A. & Srivastava, D. (2024). An Efficient Automated Image Caption Generation by the Encoder-Decoder Model. In International Journal of Information Technology

[25] Verma, A. et al. (2021). Intelligent Image Captioning with InceptionV3, LSTM, and PySpark Integration. In *International Conference on Computing, Communication, and Intelligent Systems*.

[26] Al Badarneh, A., Younes, R., & Halawani, A. (2025). An Ensemble Model with Attention-Based Mechanism for Image Captioning. In Future Generation Computer Systems.

[27] Li, J., Yu, X., Wang, S., Liu, W., & Zhang, Z. (2023). DAT-PoS-Transformer: A Depth-Aware Transformer Incorporating Part-of-Speech for Image Captioning. In Multimedia Tools and Applications.

[28] Zhou, L., Palangi, H., Zhang, L., Hu, H., & Gao, J. (2020). Unified Vision-Language Pre-Training for Image Captioning and VQA. In AAAI 2020.

[29] Li, X., Yin, X., Li, C., Hu, X., Zhang, P., Zhang, L., Wang, L., Hu, H., Dong, L., Wei, F., & Gao, J. (2020). OSCAR: Object-Semantics Aligned Pre-training for Vision-Language Tasks. In *ECCV 2020*.

[30] Mokady, R., Hertz, A., & Bermano, A. H. (2021). ClipCap: CLIP Prefix for Image Captioning.

[31] Chen, Y., Wang, Q., Zhang, L., & Li, Y. (2023). A New Attention-Based LSTM for Image Captioning. In *Neural Processing Letters*.

[32] Liang, Y., Yang, L.-C., Yang, C.-Y., & Hsu, J. Y.-j. (2021). Object Relation Attention for Image Paragraph Captioning. In *AAAI 2021*.

[33] Chen, M., Li, Z., Gao, S., & Luo, X. (2020). CREST-iMAP v1.0: A Fully Coupled Hydrologic-Hydraulic Modeling Framework. In *Environmental Modelling & Software*.

[34] Tan, Y., Song, Y., & Tan, M. (2021). Effective Multimodal Encoding for Image Paragraph Captioning (PaG-MEG-SCST). In *IEEE TIP*.

[35] Zhang, H. et al. (2023). DGOcc: Depth-Aware Global Query-Based Network for Monocular 3D Object Detection. In *Neurocomputing*.

[36] Wang, T. et al. (2024). Clustering-Based Topic Modeling for Biomedical Documents. In *Scientific Reports*.

[37] Liang, Y. et al. (2021). A Span-Based Model for Joint Overlapped and Discontinuous Entity Mention Detection. In *ACL 2021*.

[38] Wang, Q., & Chan, A. B. (2018). CNN+CNN: Convolutional Decoders for Image Captioning.

[39] Li, J., Yu, Y., & Chung, J. (2020). Unimodal Language Models Guide Multimodal Language Generation (Dual-CNN LM). In *EMNLP 2023*.

[40] Badarneh, A. et al. (2025). Depth-Aware Lightweight Network for RGB-D Salient Object Detection (DepthFusion Transformer). In *IET Image Processing*.

[41] Rao, Y. et al. (2022). Topic-Aware Generative Adversarial Network for Paragraph Captioning.

[42] Jain, A. et al. (2023). Content Planning Transformer for Coherent Long Text Generation.

[43] Kim, J. et al. (2023). Structure-Aware LSTM for Visual-Textual Paragraph Generation.

[44] Luo, R. et al. (2024). Learning Visual Relationship and Context-Aware Attention for Image Captioning (Semantic Graph Attention). In *Pattern Recognition*

DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi College of Engineering)

Shahbad Daulatpur, Main Bawana Road, Delhi-42

PLAGIARISM VERIFICATION

Title of the Thesis _____

_____

_____

Total Pages _____

Name of the Scholar_____

Supervisor

(1) _____

Department_____

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

Software used:_____

Similarity Index: _____

Total Word Count: _____

Date: _____

Candidate's Signature **Signature of Supervisor**

CV

SUYASH GUPTA

+91-8707504549 suyashgupta_23spd03@dtu.ac.in https://www.linkedin.com/in/suyashg20

EDUCATION

M.TECH (SPD)	2023-2025	Delhi Technological University, Rohini, New Delhi -110042	7.45 GPA
B. TECH (ECE)	2019-2023	JSS Academy of Technical Education, Noida	72.0 %
CBSE (Class XII)	2018	Reliance Academy, Gorakhpur	80.0 %
CBSE (Class X)	2016	Surmount International School, Gorakhpur	10 CGPA

OBJECTIVE

Versatile and motivated engineer with a strong foundation in electronics, machine learning, and deep learning, seeking a challenging role to apply my skills in designing innovative solutions across diverse domains. Experienced in signal processing, algorithm development, and data-driven modeling, I am eager to contribute to cutting-edge projects in technology, healthcare, or core electronics while advancing my expertise in emerging technologies.

Technical SKILLS

Interest Areas: Digital Electronics, Digital IC Design, Analog IC Design, Static Timing Analysis (STA), ASIC Design Flow, Machine Learning	**Tools**: Xilinx Vivado, LT Spice, MATLAB, Proteus Design Suite	**Languages**: Verilog/VHDL, Python

ACADEMIC PROJECT

Paragraph Image Captioning using Deep Learning **August 2024 – May 2025**

This research project focuses on developing advanced deep learning models to generate detailed and contextually relevant descriptions for multiple regions within an image. Dense image captioning identifies and describes specific objects, actions, and interactions in localized regions. The project leverages techniques such as **convolutional neural networks (CNNs) for feature extraction** and transformers or **recurrent neural networks (RNNs) for sequential caption generation**, aiming to enhance visual understanding and applications in fields like autonomous systems, surveillance, and accessibility technologies.

Performance Analysis of Energy Efficient WSN using Blockchain model for IOT Application **Aug 2022 – May 2023**

This work focuses on:

- Reducing Energy consumption of WSNs
- Integration of Blockchain with the model.

PUBLICATIONS

Integration of Block Chain model for Energy Efficient WSN for IoT Application February, 2023

•Published in International Journal for Research in Applied Science & Engineering Technology (IJRASET)

https://doi.org/10.22214/ijraset.2023.48942