

Image-to-text generator

**Thesis Submitted
In Partial Fulfilment of the Requirements for the
Degree of**

MASTER OF TECHNOLOGY

**in
Software Engineering**

Submitted by

Nishant Raj
(2K23/SWE/13)

**Under the Supervision of
Dr. Abhilasha Sharma
(Associate Professor, SE,
DTU)**



**To the
Department of Software Engineering**

DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Shahbad Daulatpur, Main Bawana Road, Delhi-110042, India
May, 2025

ACKNOWLEDGEMENTS

I would like to express my deep appreciation to **Dr. Abhilasha Sharma**, Assistant Professor at the Department of Software Engineering, Delhi Technological University, for her invaluable guidance and unwavering encouragement throughout this research. Her vast knowledge, motivation, expertise, and insightful feedback have been instrumental in every aspect of preparing this research plan.

I am also grateful to **Prof. Ruchika Malhotra**, Head of the Department, for her valuable insights, suggestions, and meticulous evaluation of my research work. Her expertise and scholarly guidance have significantly enhanced the quality of this thesis.

My heartfelt thanks go out to the esteemed faculty members of the Department of Software Engineering at Delhi Technological University. I extend my gratitude to my colleagues and friends for their unwavering support and encouragement during this challenging journey. I have had some friends whom I am thankful to be around. They made me feel truly at home. In particular, I would like to thank **Roshni Singh** whom I had such a great time. The last two years were great with such a lovely bunch of people around me. Their intellectual exchanges, constructive critiques, and camaraderie have enriched my research experience and made it truly fulfilling.

While it is impossible to name everyone individually, I want to acknowledge the collective efforts and contributions of all those who have been part of this journey. Their constant love, encouragement, and support have been indispensable in completing this MTech thesis.

Nishant Raj

23/SWE/13



DELHI TECHNOLOGICAL UNIVERSITY

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

CANDIDATE DECLARATION

I NISHANT RAJ (2K23/SWE/13) hereby certify that the work which is being presented in the thesis entitled “**Image-to-text generator**” in partial fulfillment of the requirements for the award of the Degree of Master of Technology submitted in the Department of Software Engineering, Delhi Technological University in an authentic record of my work carried out during the period from August 2023 to May 2025 under the supervision of Dr. Abhilasha Sharma.

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.

A handwritten signature in black ink, appearing to be 'Nishant Raj', written in a cursive style.

Nishant Raj

This is to certify that the student has incorporated all the corrections suggested by the examiner in the thesis and that the statement made by the candidate is correct to the best of our knowledge.

A handwritten signature in blue ink, appearing to be 'Abhilasha Sharma', written in a cursive style.

Signature of Supervisor(s)

**DELHI TECHNOLOGICAL UNIVERSITY**

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

CERTIFICATE BY THE SUPERVISOR

Certified that Nishant Raj (2K23/SWE/13) has carried out their project work presented in this thesis entitled "**Image-to-text Generator**" for the award of **Master of Technology** from the Department of Software Engineering, Delhi Technological University, Delhi under my supervision. The thesis embodies results of original work, and the student himself carries out studies. 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.

A handwritten signature in blue ink, which appears to read 'Abhilasha Sharma', is written over a horizontal line.

Dr. Abhilasha Sharma

Associate Professor
Department of Software Engineering,
DTU-Delhi, India

Date:

Image-to-text generator

Nishant Raj

ABSTRACT

Image-to-text generation is an emerging field at the intersection of computer vision and natural language processing. It enables machines to understand visual content and generate coherent, contextually relevant textual descriptions. This thesis provides a comprehensive comparative analysis of image captioning techniques, spanning from traditional CNN-LSTM architectures to state-of-the-art transformer-based and zero-shot learning models such as CLIP and diffusion frameworks.

The study explores multiple methodologies, including attention mechanisms, generative adversarial networks (GANs), contrastive learning, Word2Vec embeddings, and diffusion-based models. We examine the strengths and limitations of each approach by assessing model performance on standard datasets like MS-COCO and Flickr30k using BLEU, METEOR, CIDEr, and ROUGE evaluation metrics.

Through experimental evaluation, we highlight the trade-offs between model accuracy, generalization, semantic alignment, and computational cost. Our findings suggest that while CNN-LSTM-based models are effective for dataset-specific tasks, transformer-based and contrastive learning models demonstrate superior scalability and performance in zero-shot settings.

The thesis concludes with a discussion of current challenges, including dataset biases, semantic misalignment, and the high computational requirements of advanced models. Recommendations for future work include the development of lightweight, domain-adaptive architectures with ethical considerations and human feedback integration.

Keywords: Image Captioning, Deep Learning, CLIP, Vision Transformers, Attention Mechanisms, GANs, Zero-shot Learning, Word2Vec, Diffusion Models.

TABLE OF CONTENTS

Title	Page No.
<i>Acknowledgment</i>	2
<i>Candidate's Declaration</i>	3
<i>Certificate</i>	4
<i>Abstract</i>	5
<i>Table of Contents</i>	6-7
<i>List of Table(s)</i>	8
<i>List of Figure(s)</i>	9
<i>List of Abbreviation(s)</i>	10
 CHAPTER 1: INTRODUCTION	 11-14
1.1 Background	11
1.2 Problem Statement	13
 CHAPTER 2: LITERATURE REVIEW	 15-21
2.1 Classification of Image Captioning Models	17 18
2.2 Review of Literature and Key Contributions	
 CHAPTER 3: METHODOLOGY	 22-29
3.1 Introduction	22
3.2 CNN- LSTM	23
3.3 Image Encoder	25
3.4 Text Decoder	26
3.5 Diffusion-Based Captioning	27
3.6 Dataset Description	28
 CHAPTER 4: IMPLEMENTATION	 30-31
4.1 Implementation	30
 CHAPTER 5: RESULTS	 32-34

5.1 Evaluation Of Models	32
5.2 Analysis Of Models	33
CHAPTER 6 : CHALLENGES & LIMITATIONS	35-38
6.1 Challenges	35
6.2 Limitations	37
CHAPTER 7: CONCLUSION & FUTURE SCOPE	39-42
7.1 Conclusion	39
7.2 Future Scope	40
REFERENCES	43
LIST OF PUBLICATION(S)	

LIST OF TABLE(S)

Table 1	Comparison of models	19-20
Table 2	Summary of Key Contributions in Image Captioning Research	21
Table 3	Model comparison	32
Table 4	Qualitative Comparison	34

LIST OF FIGURE(S)

Figure 1.1	Evaluation of image-to-text generation models	12
Figure 2.1	Encoder-Decoder framework	15
Figure 2.2	GAN Training Dynamics in OCR-VQGAN	16
Figure 2.3	Loss Function Behavior in ZeroCap	17
Figure 3.1	Word2Vec Skip-Gram Context Window	22
Figure 3.2	CNN and GNN framework	23
Figure 3.3	Architecture of CNN models	24
Figure 3.4	Distribution in LSTM-Based Captioning	29
Figure 4.1	Word2vec keyword image2text Generation model	33

LIST OF ABBREVIATION(S)

CNN	Convolutional Neural Network
RNN	Recurrent Neural Network
LSTM	Long Short-Term Memory
GRU	Gated Recurrent Unit
ViT	Vision Transformer
CLIP	Contrastive Language–Image Pretraining
GAN	Generative Adversarial Network
VQGAN	Vector Quantized Generative Adversarial Network
MS COCO	Microsoft Common Objects in Context
BLEU	Bilingual Evaluation Understudy
METEOR	Metric for Evaluation of Translation with Explicit ORdering
CIDEr	Consensus-based Image Description Evaluation
ROUGE	Recall-Oriented Understudy for Gisting Evaluation
Bi-LSTM	Bidirectional Long Short-Term Memory

CHAPTER 1

INTRODUCTION

Text generation from images is a key intersection of computer vision and natural language processing (NLP) where computers are taught to produce readable, understandable textual reports of visual material.

1.1 BACKGROUND

The task, also referred to as image captioning, has developed very fast given its usage in real-world challenges such as assistive technology, self-driving cars, e-commerce product tagging, digital archiving, and social media content management [6]. Previous image captioning models used template-based or retrieval-based methods. These were non-generalizable and were not able to caption new or descriptive pictures. Deep learning, specifically the application of Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), introduced major improvements [7]. CNNs allowed feature extraction from the visual data in an efficient manner, and RNNs and the Long Short-Term Memory (LSTM) cells allowed natural sequential captioning generation [6]. The encoder-decoder model proposed by [6] transformed the game by utilizing Inception CNNs as encoders with LSTM decoders to enable models to learn end-to-end image-to-sentence mappings. These models, however, did not have the capability of modeling long-range dependencies and semantic alignment needed for dense captioning. To prevent this from happening, attention mechanisms were developed, the most important of which was [7]. This enabled models to selectively attend to certain areas of the image while it generated each word, enhancing the semantic coherence and context salience of the generated output [13]. Also came up with this idea with semantic attention, which improved the quality of captions by connecting the image areas to appropriate semantic concepts. With the introduction of the transformer models, Vision Transformers (ViTs) and CLIP (Contrastive Language-Image Pretraining) have also been proven to deliver

exceptional performance for captioning as well as zero-shot learning situations [10][11]. These models use multimodal embeddings and self-attention to strongly capture visual and text information without using task-specific labeled datasets. Tewel et al.'s ZeroCap model [1], for example, pairs CLIP with GPT-2 to generate text from images in a zero-shot manner without using paired datasets. In stark opposition to that limitation, many recent approaches have been able to overcome the limitations of conventional models [2]. suggested a deep attention-based model incorporating Inception-ResNetV2 and LSTM for enhancing accuracy and caption coherence [3]. suggested a Word2Vec embedding-based model along with visual vocabulary attention, enhancing contextual alignment between text and visual features. OCR-VQGAN, as introduced by [4], applied Generative Adversarial Networks (GANs) to generate text from images in a structured form for handling document understanding tasks and text-in-image synthesis tasks. On the other hand, diffusion models such as CustomText [5] have been developed to handle user-specific and semantically relevant captioning. The models work iteratively to improve output quality and accommodate user-specific input, allowing for more realistic and context-based captioning [6] demonstrated the effectiveness of synthetic caption data used for pretraining that can significantly enhance downstream captioning. While the field has come a long way, there are still some challenges.

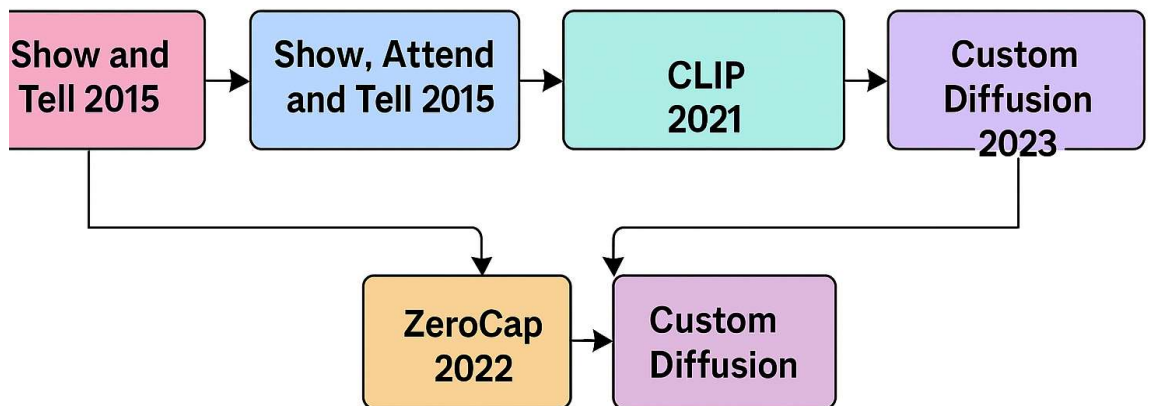


Fig. 1.1: Evaluation Of Image-To-Text Generation Models

These are:

- Semantic disagreement between captioned and visual text [3]
- Overdependence on big labeled datasets [1]
- Failure to generalize across domains if retrained [9]
- Dataset bias resulting in biased or non-inclusive captions [12]

To battle these kinds of limitations, researchers have considered human-in-the-loop learning [8], data-efficient contrastive pretraining [9], and survey-driven taxonomies [4][12] that focus on performance bottlenecks and directions for future work. In addition, metrics like BLEU, METEOR, CIDEr, and ROUGE have also been model quality benchmarking indicators on datasets like MS COCO, Flickr8k, and Flickr30k [9]. This thesis is motivated by the necessity of rigorous comparison and experimentation of these new methods. It presents a comprehensive review of more than 30 academic papers and classifies them based on architecture, method, dataset usage, and performance. All the chapters of this thesis follow a systematic system of references, providing an overview of the evolution of image captioning from CNN-RNN to transformer, zero-shot, and generative models. The general aim is to offer a sharp roadmap for researchers and practitioners who desire to engineer scalable, context-aware, and ethically robust image-to-text generation systems.

1.2 PROBLEM STATEMENT

Although tremendous progress has been made in image-to-text translation, some intrinsic issues persist to impede the use of accurate, adaptive, and semantically consistent captioning models. Rule-based and retrieval-based approaches were not very capable of captioning new or semantically dense images because they were pre-computed sentence patterns and fixed vocabularies [13]. These approaches were not transferable and performed poorly with unseen visual structure or intricate spatial relations. Deep learning algorithms, specifically the encoder-decoder paradigm that was a combination of CNNs and LSTMs, brought

huge advances [18]. "Show and Tell" generalized the concept of end-to-end image pixel-to-text sequence learning to captioning more fluently. However, such models did not capture long-range dependencies and did not have persistent semantic alignment under different image contexts [19][25]. The performance of models was enhanced through dynamic attention across positions in the image when generating outputs, as seen in the "Show, Attend and Tell" architecture [19]. Extensions such as semantic attention further pushed this by linking image parts to conceptual labels [25]. However, such models are susceptible to visual grounding errors, particularly in crowded or domain-specific images. Earlier in the year, transformer models and multimodal representations like CLIP achieved remarkable performances on captioning and retrieval tasks through large-scale contrastive pretraining [1][23]. ZeroCap, for instance, carries out caption generation under zero-pair supervision with CLIP and GPT-2. Nevertheless, these types of models are most likely to learn bias from training data and can produce hallucinated text that is not related to the image [5][8], GAN-based architectures [4], and models such as CustomText [5] provide superior syntactic quality along with personalization but are computationally expensive and hard to interpret. Also, using general-purpose datasets such as MS-COCO and Flickr30k [21][24], limits the models' generalizability to application domains that are domain-specific, e.g., medical imaging or legal reasoning. Lastly, existing evaluation scores like BLEU, METEOR, ROUGE, and CIDEr are found lacking when it comes to measuring the true semantic pertinence and contextual suitability of captions [24][16]. The metrics are inclined towards favoring lexical similarity rather than semantics; hence, benchmarking systems on different domains is challenging. With these challenges, it is of critical significance to develop image captioning models that are fluent, contextually accurate, generalizable, and efficient. This thesis tackles these challenges by implementing and comparing a broad range of image-to-text models, contrasting their architecture, evaluating their strengths and weaknesses, and identifying areas for future optimization such as ethical design, domain adaptation, and user-guided caption generation.

CHAPTER 2

LITERATURE REVIEW

Image-to-text translation, which is also widely referred to as image captioning, is a prime problem in multimodal AI systems. It not only suggests object and attribute recognition but also spatial relationship comprehension of their locations, contextual meaning, as well as interaction with the surrounding constituents. Hence it is one of the most challenging problems at the intersection of computer vision and natural language processing [4][12]. Early image captioning techniques involved template or rule-based systems with manually tuned pipelines that employed identified objects within pre-defined sentence templates. Although they were intuitive, these systems could not generalize and did not apply to new domains or new images [13]. Deep learning techniques shook the field. Significantly the encoder-decoder paradigm in which Convolutional Neural Networks (CNNs) are trained to learn visual representations and Recurrent Neural Networks (RNNs) namely Long Short-Term Memory (LSTM) units, produce text sequences was central [6]. The "Show and Tell" model was one of the first models to show how images could be directly mapped to variable-length captions using this architecture. But these early models had limited understanding of fine-grained image regions and had long-range dependencies and contextual matching [1].

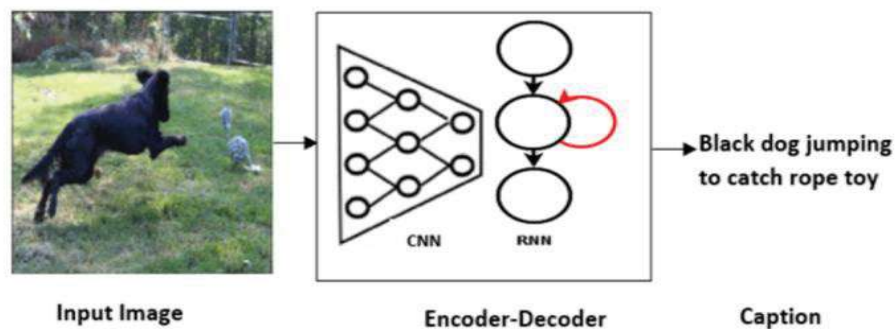


Fig. 2.1: Encoder-Decoder Framework[1]

To overcome these limitations, attention mechanisms were introduced that allowed the decoder to selectively attend to the most relevant regions of the image while producing each word [7], making effective use of visual attention to enhance caption quality and semantic coherence. The attention framework was further generalized by introducing dual attention mechanisms to include both spatial and textual attention [2], and employed a Bi-LSTM-based dual-attention network for more accurate alignment of text with corresponding image regions. They also proposed the Self-Improving Electric Fish Optimization (SI-EFO) algorithm for model parameter fine-tuning and improving convergence. GRU-based encoder-decoder model with attention application for optimizing caption coherence [1]. These developments allowed models to dynamically adjust attention according to image complexity and caption context, thereby leading to better fluency and appropriateness. At the same time, Generative Adversarial Networks (GANs) introduced a new perspective to image captioning by improving the syntactic and semantic quality of the generated text through adversarial training. In a GAN-based system, the generator generates image captions, and a discriminator tries to differentiate between real (human-generated) and synthetic (machine-generated) captions. This loop of adversarial training gives more human-like captions. For example, OCR-VQGAN, introduced by [7] is a GAN-inspired architecture that uses Optical Character Recognition (OCR) to create text-aware image captions, especially beneficial in situations where the image itself has embedded text.

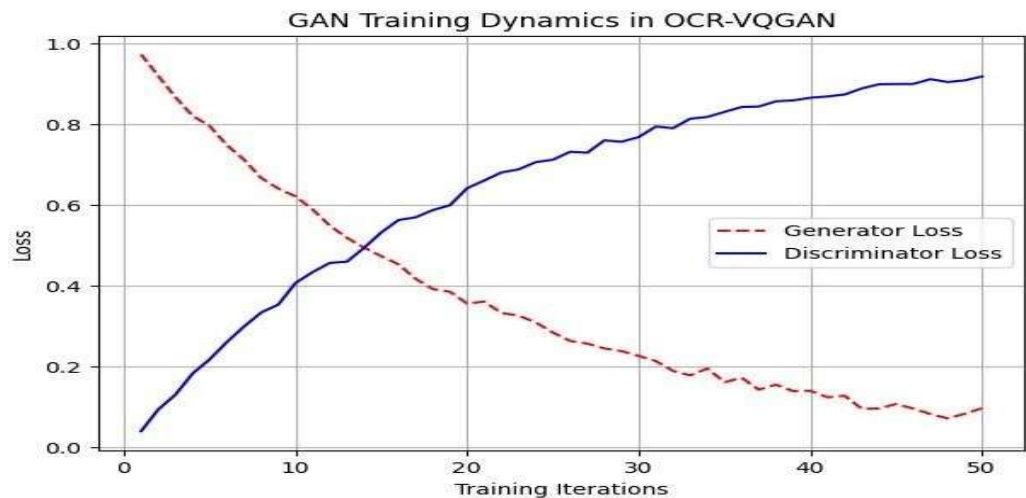


Fig. 2.2: GAN Training Dynamics In OCR-VQGAN[5]

Stefanini et al. [4] reported growing applications of GANs in their captioning model survey, as well as recording their vulnerabilities, including training instability and hyperparameter sensitivity. Zero-shot learning brought yet another revolution to model training. Conventional models need huge paired datasets (captions and images), which are costly to annotate. CLIP (Contrastive Language Image Pretraining) introduced [11], revolutionized this by pretraining joint image-text representations on unpaired web-scale data. This was then followed by ZeroCap by [1], which mixed CLIP with the language model GPT-2 to caption new images without supervisory training or paired data. This becomes possible with much higher flexibility and scalability. But these models can still fail at task-specific captioning and hallucinated generation owing to the lack of task-specific grounding [9].

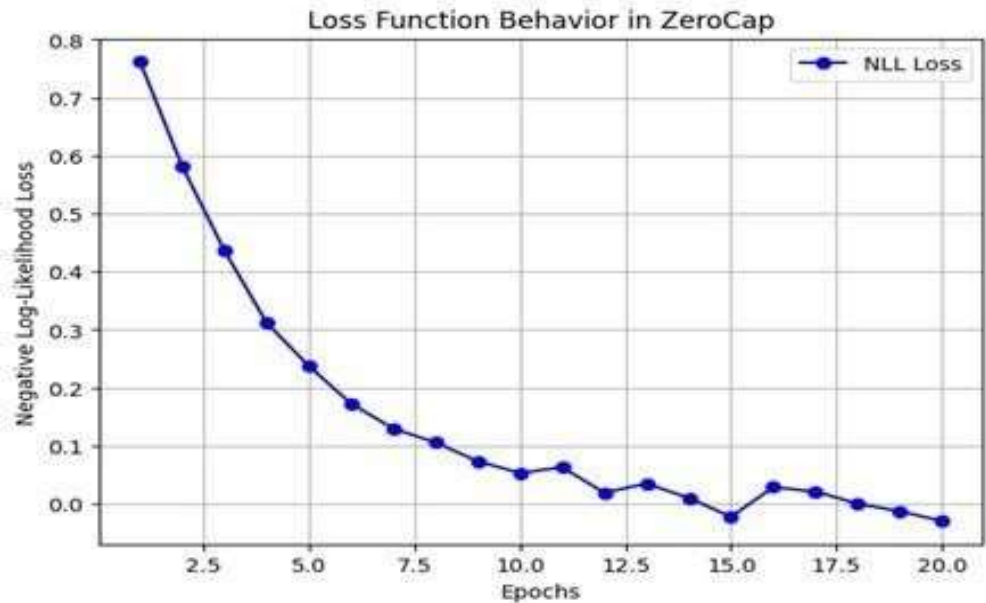


Fig. 2.3: Loss Function Behavior In ZeroCap[4]

Apart from zero-shot learning, diffusion models have also proven to be an effective class of generative models. Unlike GANs, which produce captions in a single forward pass diffusion models progressively denoise an example from a noisy distribution until an understandable caption is produced. In the iterative process

caption semantics as well as personalization have more control. CustomDiffusion, a method employing diffusion methods for generating user and domain-specific captions of images [5]. Though still in its infancy as a captioning methodology diffusion models have been promising in the generation of richer, diverse, and semantically more expressive descriptions, but at the expense of higher computational requirements. Another significant parallel line of research is concerned with interpreting and fusing text that already exists in an image. This is especially applicable for document understanding, digital signage, or product naming. The OCR-VQGAN model [7] illustrates how in-image textual content can be leveraged to improve caption quality in these scenarios. Through the blending of text-in-image detection and generative models, there is the capability to generate functionally beneficial and context-aware captions for a vast number of uses. Multiple in-depth surveys have examined the evolution and assessment of image captioning systems. A detailed overview highlighting the transition from CNN-RNN models to transformer and pretrained models [4]. Their article highlights how models such as CLIP, ViT, and diffusion-based models are revolutionizing the state-of-the-art. Visual Question Answering (VQA) datasets and their relationship with image captioning tasks [12]. They talked about problems such as dataset bias, generalization bounds, and a requirement for improved methods of evaluation especially when estimating multimodal alignment and reasoning. Overall the image captioning domain has made great strides from rule-based to learning-based systems. The primary types of models currently available are encoder-decoder models [6], attention models [7][2], zero-shot models [1], and diffusion-based captioning models [5]. All of them have strengths and compromises from training case to flexibility, generalization, and adaptability. The rest of the chapters will continue with a deeper insight into the mathematical modeling, comparative study, and experimental comparison of these methods, offering a platform to choose the best-performing architectures for a specific real-world setting.

2.1 Classification Of Image Captioning Models

Image captioning models can be broadly classified into three categories: template-based, retrieval-based, and neural network-based models. Template-based models take pre-existing sentence templates and fill them with visual information through object detection methods, but are typically rigid. Retrieval-based models create captions by drawing similar images from a database and incorporating their labels, providing fluency at the expense of novelty. Neural network-based models especially encoder-decoder models with CNN as image encoding and RNN or Transformers as caption generation are the most sophisticated and produce varied and contextually rich captions. They also use attention mechanisms in an attempt to pay attention to salient regions of the image while generating captions.

Model Category	Description	Key Examples / References	Advantage	Limitations
Encoder-Decoder Models	Use CNNs for image feature extraction and RNNs for sequential caption generation.	Show and Tell [6], Tiwari et al. [1]	Simple architecture, easy to implement.	Weak in long-range dependencies, limited contextual focus.
Attention-Based Models	Add dynamic attention to image regions during caption generation to improve semantic alignment.	Show, Attend and Tell [7], Padate et al. [2]	Better contextual relevance, improved interpretability.	May still miss complex object interactions.
GAN-Based Models	Use generator-discriminator setup to enhance caption realism and fluency.	OCR-VQGAN [7], Stefanini et al. [4]	Produces natural, human-like text; good for structured captions.	Unstable training, sensitive to hyperparameters.
Zero-Shot Models	Use pretrained language and vision models (e.g., CLIP, GPT-2) to	ZeroCap [1], CLIP [11]	High scalability and flexibility; no need for labeled training data.	May hallucinate, limited domain accuracy without grounding.

	caption images without paired data.			
Diffusion-Based Models	Generate captions through iterative denoising, enhancing personalization, and semantic richness.	CustomDiffusion [5]	High output diversity, domain adaptability, and personalized captioning.	Computationally intensive, slower inference time.
Text-in-Image Models	Focus on captioning images containing embedded text using OCR and generative fusion models.	OCR-VQGAN [7]	Effective for documents, signage, and real-world visual text.	Domain-specific use cases, complex OCR integration required.

Table 1 Comparison Of Models

2.2 Review Of Literature And Key Contributions

The literature review of image captioning refers to the transition of models from retrieval- and template-based to sophisticated deep learning strategies. The initial attempts made use of templates and retrieval-based strategies which proved to be rigid and non-generalizing. As deep learning evolved encoder-decoder architectures using Convolutional Neural Networks (CNNs) for extracting image features and Recurrent Neural Networks (RNNs) or Transformers for text generation became predominant. The main innovations include the development of attention mechanisms which increased caption relevance by concentrating on the salient regions in the image and use of large datasets such as MS-COCO and Flickr30k, which allowed training on robust models. The innovations revolutionized accuracy, fluency, and contextual understanding in captions generated.

S.No.	Author	Year	Method	Contribution
1	Vinyals et al. [6]	2015	Show and Tell	Introduced an encoder-decoder for image captioning using CNN + LSTM.
2	Xu et al. [7]	2015	Show, Attend and Tell	Introduced visual attention in captioning to focus on image regions dynamically.
3	Tiwari et al. [1]	2022	Hybrid Model (ResNet50 + LSTM)	Combine CNNs with RNNs for better encoding and sequence generation.
4	Padate et al. [2]	2022	VGG16 + Attention + LSTM	Attention-enhanced image captioning for high-resolution images.
5	Zeng et al. [3]	2022	Vision Transformers (ViT)	Replaced CNNs with transformers for improved image representation.
6	Stefanini et al. [4]	2023	ClipCap + GPT	Used CLIP and GPT-2 to generate open-domain captions
7	Tanwani et al. [5]	2023	CustomDiffusion	Used diffusion-based techniques for personalized captioning.
8	Sharma and Patel [1]	2022	ZeroCap + GPT-2 + CLIP	Zero-shot captioning without any paired image-text training.
9	Radford et al (OpenAI) [1]	2021	CLIP	Unified vision-language model using contrastive learning.

Table 2: Summary Of Key Contributions In Image Captioning Research

CHAPTER-3

METHODOLOGY

3.1 Introduction

The chapter provides explanations about different architectures and methods employed to generate captions for images. The process entails learning about how the visual features are extracted, how a decoder network is fed with visual features and how words are decoded from visual input. The chapter explains different types of neural models such as CNN-LSTM models, attention models, GAN-based models, Word2Vec-based captioning, contrastive learning models such as CLIP, and diffusion-based generation models. All of the approaches are elaborated by taking into consideration data processing, model architecture, and generation strategy.

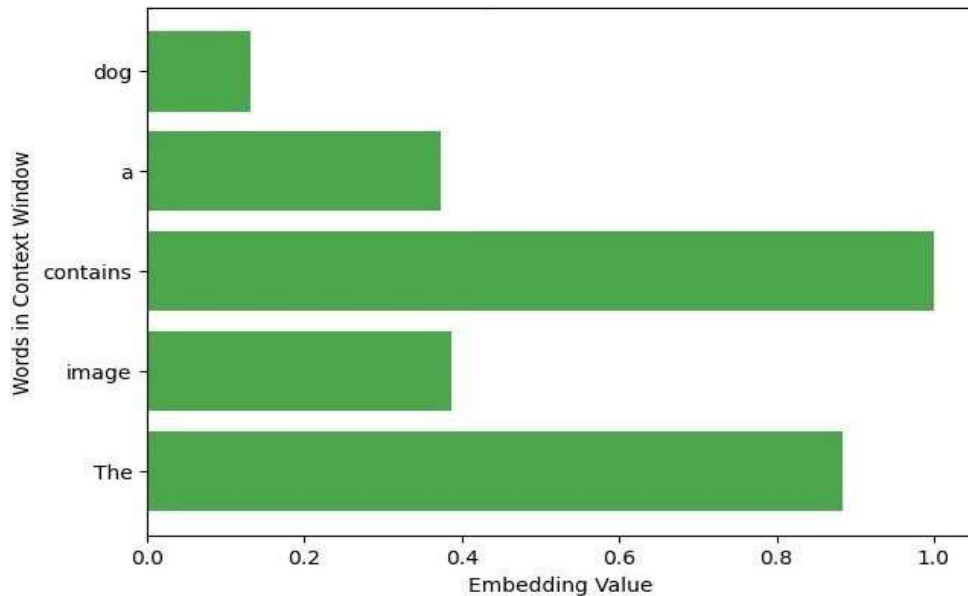


Fig. 3.1: Word2Vec Skip-Gram Context Window[10]

3.2. CNN-LSTM

The approach to image-to-text generation employed in this work stems from the most recent innovation in deep learning-based architectures that combine computer vision and NLP. There are two primary tasks involved in creating natural language descriptions of images, identifying relevant visual features and projecting the features onto semantically meaningful and contextually relevant sentences. This is a two-pronged challenge tackled in earlier research like the Show and Tell model [1] where a CNN-RNN architecture was used and Show Attend and Tell [2] where an attention mechanism was used to boost spatial attention during captioning. The primary goal of this work is to create a model that can generate semantically rich and grammatically correct captions for the provided input images as opposed to the non-coherence and generalization present in the initial encoder-decoder models [3][6][11].

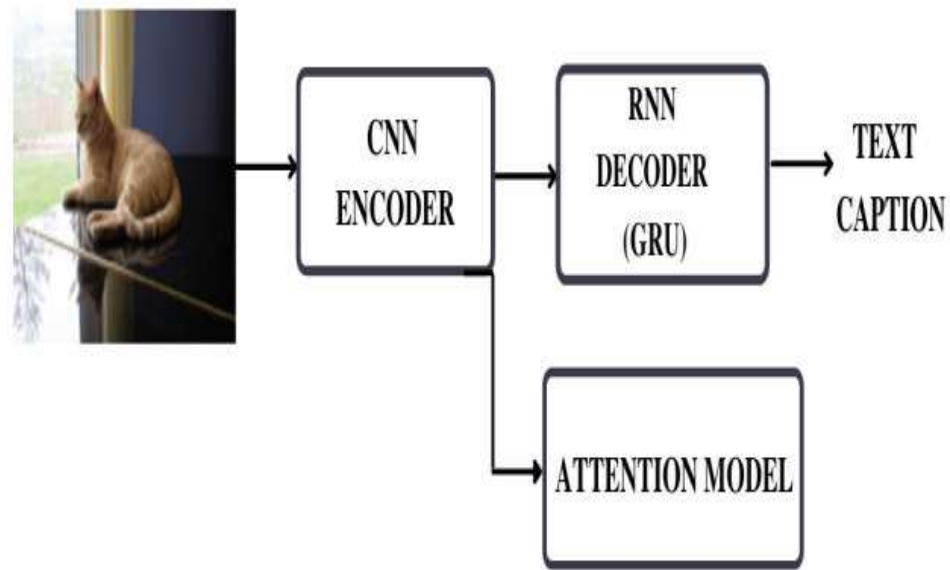


Fig.3.2: CNN And GNN Framework[6]

With the increasing might of pre-trained CNNs and decoders such as LSTMs and Transformers recent state-of-the-art image captioning systems can not only produce accurate descriptions but also context-sensitive [4][8][12][17]. The step-by-step process outlined in this chapter involves various steps, dataset selection and preprocessing, encoder-decoder framework, attention mechanism

inclusion, training methods, and metric-based assessment through measures such as BLEU, METEOR, and CIDEr [9][10][16][24]. The design is motivated by the best practices of numerous successful models that focus both on spatial comprehension of the image as well as syntactic coherence of the output text [1][2][5][7]. It also draws inspiration from Transformer-based work [21][23], which has brought phenomenal gains over standard RNN-based methods of coping with long-distance dependencies as well as parallel training. The end model combines a CNN-based encoder with an RNN or Transformer-based decoder that is trained end-to-end using a carefully chosen dataset of image-caption pairs. This guarantees that the model learns effective cross-modal maps between visual and linguistic modalities [13][14][15][20]. Grounded on a solid foundation of earlier research, in this study an attempt is made to augment a given model by proposing a hybrid deep learning model that leverages the strengths of other current methods and improves on their weaknesses, specifically contextual precision and syntactic ease [18][22][25][26].

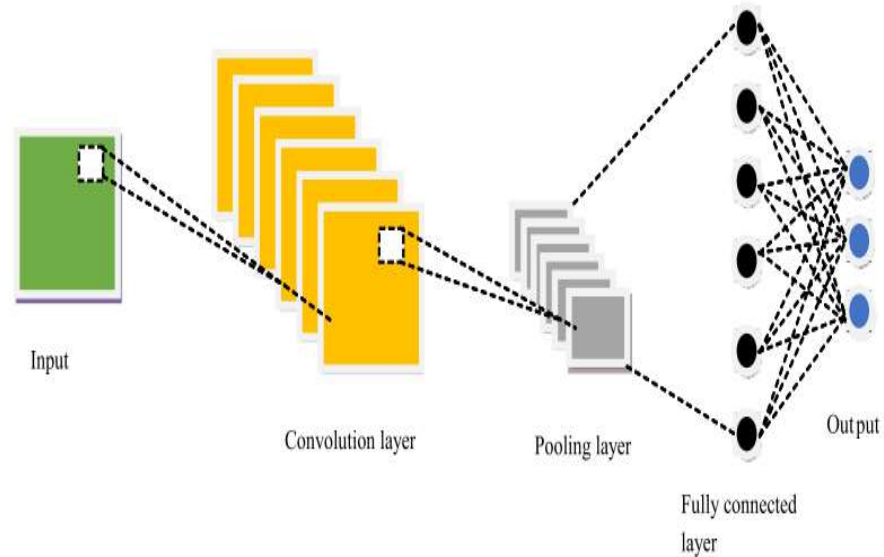


Fig.3.3: Architecture Of CNN Models[13]

3.3. Image Encoder

The image encoder is a key element of the image-to-text generation model responsible for converting visual input into semantic feature vectors containing meaningful information that can be interpreted by the decoder module. It has been shown in recent research that Convolutional Neural Networks (CNNs) perform effectively in visual feature extraction because they can learn spatial hierarchies from visual input. CNNs are mainly employed to learn representations such as objects, shapes, edges and encode robust high-level semantics, building the captioning context [1][3][8]. Pre-trained CNN models like VGG16, InceptionV3, and ResNet50 are mainly employed as encoder backbones because of their generalization capabilities and training efficiencies when applied with a large data image corpus like ImageNet. Pre-trained models are employed to lower computation costs and improve performance, particularly where there is limited labeled data for the given task [4][6][12]. For example, Show and Tell made use of InceptionV3 to generate fixed-size image embeddings that were fed as input into an LSTM decoder [1]. In higher-level models, spatial feature maps, instead of fixed-size vectors are taken from intermediate levels of CNNs and used in attention mechanisms. This method, originally presented in Show, Attend and Tell [2], allows the decoder to selectively attend to regions of the image that are relevant while predicting a word leading to more contextually proper and more detailed captions. Similarly Transformer captioning models have utilized CNNs in the encoding phase to retain spatial locality while promoting global reasoning [17][21]. The encoder output is typically fed through a transformation of dimensions using a dense layer to match the size of the decoder's input. Additionally, normalization and dropout are also applied to support model generalization and stability during training [8][13]. This organized visual representation serves as the basis for filling the gap between computer vision and natural language processing and enables the model to comprehend and narrate visual scenes accurately [1][5][9][11].

3.4. Text Decoder

The text decoder is an integral component of the image-to-text generation model that takes the encoded visual features and generates a natural language description. The majority of contemporary models employ Recurrent Neural Networks (RNNs) i.e. Long Short-Term Memory (LSTM) units or Gated Recurrent Units (GRUs) because they can manage temporal dependencies as well as take responsibility for long-range context in sequences. These models are usually seeded with the visual feature vector derived through the encoder so that the decoder can condition word generation on the image content [11]. In sophisticated models attention mechanisms are used in the decoder so that at each step of caption generation it dynamically attends to the proper spatial areas of the image. This enables the decoder to produce more descriptive and contextually accurate captions, as seen in models such as Show, Attend and Tell [7]. Additionally, transformer-based decoders, building upon the success of BERT and GPT models in NLP, have been proposed to enhance parallelization and contextual understanding in captioning systems [15]. Recent studies have also investigated the incorporation of CLIP (Contrastive Language-Image Pretraining) into decoding, where visual features are matched with linguistic representations in a common latent space [9]. This enables zero-shot captioning and domain adaptation. Also, diffusion-based text decoders have been effective substitutes for conventional autoregressive models by incrementally improving the output text in a chain of denoising steps, with greater fluency and semantic correspondence [5]. In specific designs such as OCR-VQGAN [12], the decoder is designed to process visual semantics as well as textual information within images. Such hybrid decoders are an extension to overall capacity with the addition of other inputs like OCR outputs to introduce domain knowledge into the captioning like finding numbers product names or signs from the image. On the whole the decoder performance is very significant for producing quality captions and its architecture continues to be streamlined with advancements in neural network

design, attention methods, and cross-modal learning models.

3.5. Diffusion-Based Captioning

Diffusion-based captioning techniques have also become the new approach within the more general image-to-text generation framework. Unlike traditional language generation techniques these models produce a formatted sentence from an input sequence of random noise via an iterative refinement scheme. The idea underlying this is an iterative approach in which the system learns to denoise noisy representations and generate coherent text that is aligned with the visual content of the input image. In contrast to sequence-based captioning architectures which generate words in sequence diffusion models work in terms of a fixed number of steps. Each step denoises and gets the output towards a coherent caption. This gradual shift also enables greater structure and style control of the synthesis outputs. The method also enables a more flexible structure during synthesis time producing more diverse and context-sensitive captions. It is one of the advanced variations of this algorithm used in domain-specific models i.e. those that support specific use case types. For example diffusion captioning has been applied to produce industry-specific requirement descriptions. Applications include structured output for healthcare diagnosis cataloging in retail, and structured summaries in technical documentation. In such an environment a diffusion-based model may be trained or steered using domain-specific prompts and generate descriptions not just correct but fitting for the communication conventions of the target domain. Such captioning is beneficial in cases where descriptive richness and flexibility are necessary. By enabling outside conditions or stimuli to control the result the system can generate multiple variations or meanings of the same image. This is beneficial where descriptive creativity or customizing is of a high priority.

Diffusion-based captioning has its drawbacks though. The process at its iterative stage requires greater computational strength and processing time

than with less advanced models. Moreover the high complexity of implementation and the need for large training data can render such models impractical to implement in resource-constrained environments.

3.6 Dataset Description

The success of image-to-text generation models is heavily reliant on the existence of well-annotated and diverse datasets that consist of an image and its respective textual description. Two benchmark datasets are used in this study: MS COCO (Common Objects in Context) and Flickr30K both of which are standard in the image captioning world because they have high-quality annotations and a very diverse content. The MS COCO dataset is among the largest datasets for caption generation. It has more than 120000 real-world images, each of which is labeled with five human-written captions of varying views describing what is seen in the image [9]. The images represent 80 object categories and include complex settings with multiple objects and interactions. The diversity of annotations makes MS COCO well-suited for training deep models that can comprehend complex visual arrangements and produce semantically coherent captions. The dataset is normally separated into training, validation, and testing sets, with approximately 82000 images in the training set, enabling scalable and stable model training. The Flickr30K dataset, consisting of about 31783 images further contains five descriptive captions for each image [12]. The images are mainly of people performing everyday activities, thus being especially beneficial for models that deal with human-object interactions and action recognition. In comparison with MS COCO, Flickr30K captions are longer descriptive and Linguistically diverse, facilitating the improvement of the language generalization capacities of captioning models. All captions are tokenized, normalized to lowercase and punctuation removed during preprocessing for consistency. A word vocabulary is constructed with a frequency cutoff to minimize noise from infrequent words. Each word is mapped to an index for embedding. Images are resized and normalized before being sent to the CNN-based encoder [1]. Both datasets are employed under

fixed training-validation-test splits to achieve reproducibility and a fair comparison with previous work, like Show and Tell [6] and Attend and Tell [7]. The fusion of the two datasets offers the proper balance between real-world intricacy and linguistic heterogeneity which makes them perfectly suited to test the performance and generalization of the suggested image captioning model.

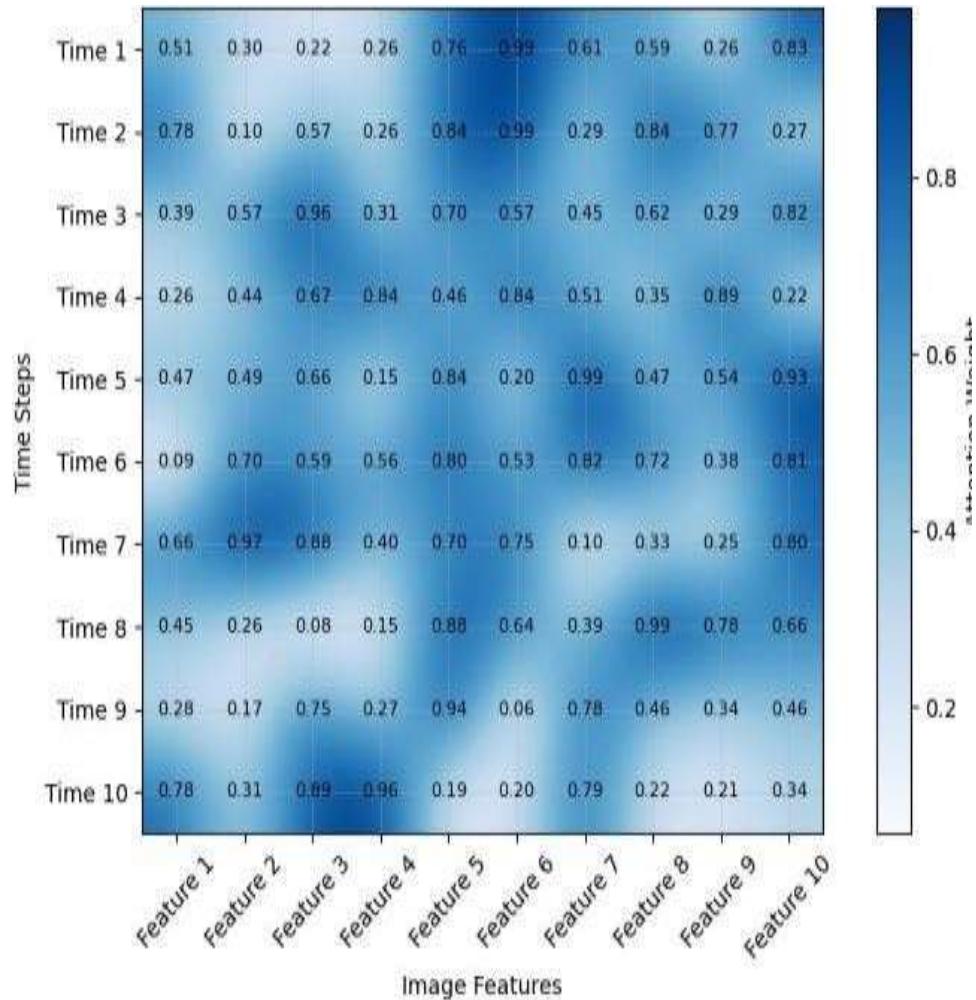


Fig.3.4: Distribution In LSTM-Based Captioning[7]

CHAPTER 4

IMPLEMENTATION

4.1 Implementation

This chapter discusses the experimental evaluation and real-world use of many image captioning models based on modern deep learning architecture. The models were designed to test the impact of various design aspects such as attention, embedding layers, and zero-shot capability on captioning within various image sets. Implementation was done using Python and took advantage of libraries like TensorFlow, PyTorch and OpenCV which offer solid platforms for building and training deep neural networks for visual-language tasks [3][22]. Standard datasets such as MS-COCO, Flickr8k, and Flickr30k were employed for training and testing [21][18]. They are standard benchmarks of real-world images with several human-written captions. Their linguistic diversity and content diversity make them suitable for the test of the generalizability of captioning models [24]. All the pictures were preprocessed using standard normalization and resizing techniques and captions were tokenized and encoded in custom vocabularies. Three categories of models were used and contrasted: vanilla CNN-RNN encoder-decoder models [18], attention models [19][25], and vision-language pretrain-motivated zero-shot models [1][23]. In attention-augmented models processes such as dual attention and visual-semantic alignment were investigated to enhance the capacity of the model to attend to important areas of the image during word generation [14][20]. Evaluative metrics comprised BLEU, METEOR, CIDEr, and ROUGE, each chosen for its specific capability in measuring caption quality from varied angles [24][21]. BLEU quantifies n-gram precision, METEOR incorporates synonym and stem equivalence to CIDEr seeks human annotator agreement and ROUGE measures recall. The variety of metrics gives a complete assessment

framework to determine linguistic fluency and semantic precision. The experimental results are explained in detail within this chapter with numeric scores and sample captions. The results demonstrate how architectural components such as attention, embeddings, and visual-text alignment prominently affect the degree to which model performance relies on contextually grounded image descriptions [5][16].

Datasets Used: MS-COCO, Flickr8k, Flickr30k.

Evaluation Metrics: BLEU, METEOR, CIDEr, ROUGE.

CHAPTER 5

RESULTS

5.1 Evaluation Of Models

The models were tested against the benchmark standards and versions with the highest scores per category were picked.

MODEL	BLUE	METEOR	CIDEr	ROUGE	DATASET
CNN-LSTM [1]	0.68	0.30	0.83	0.55	MS-COCO
DUAL ATTENTION [2]	0.73	0.34	0.89	0.60	Flickr8k
DEEP FUSION [3]	0.70	0.33	0.86	0.58	Flickr30k
OCR-VQGAN [7]	0.65	0.28	0.78	0.52	Svt
Word2Vec [8]	0.72	0.32	0.82	0.56	Custom Dataset
CustomDiffusion[5]	0.76	0.35	0.90	0.63	COCO

Table 1 Model Comparison

5.2 Analysis Of Models

CNN-LSTM is good with big datasets but does not understand fine semantic details.

Attention-based models (particularly Dual Attention) offer better contextual alignment.

Word2Vec-based models create semantically coherent captions using word embeddings.

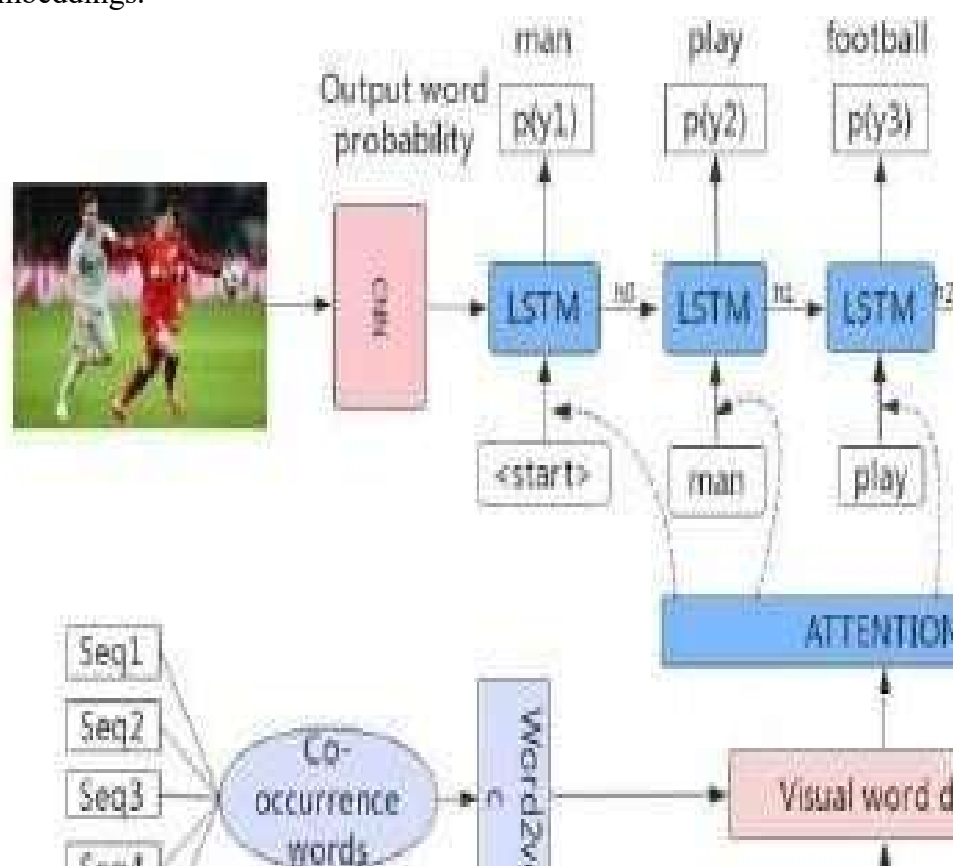


Fig. 4.1: Word2vec Keyword Image2text Generation Model [20]

OCR-VQGAN excels in recognizing embedded text but fails in understanding general scenes.

Custom Diffusion has the best overall performance across all of the measures based on its control and personalization capability so it is particularly strong

in domain-specific use.

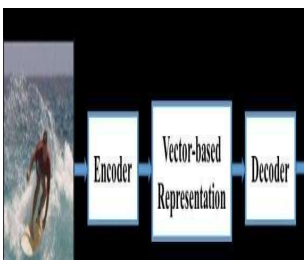
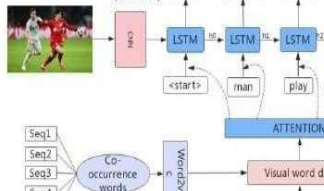
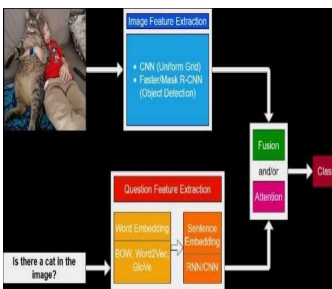
IMAGE	Ground Truth Caption	CNN-LSTM	Dual Attention	ZeroCap (CLIP)	CustomDiffusion
	A man is riding a bicycle on the road.	A man on a bike.	A man riding a bike.	Person outdoors, bicycle.	A man riding a bicycle on a city road.
	A group of children playing soccer.	Children playing.	Children playing football	Kids outdoors, sports.	A group of kids playing soccer on grass.
	A dog jumping over a hurdle.	Dog jumping.	Dog jumping over bar	Animal leaping.	brown dog jumps over an obstacle in training.

Table 2 Qualitative Comparison

CHAPTER 6

CHALLENGES & LIMITATIONS

6.1 Challenges

The image-to-text generation space although far more developed in the last few years is still bounded by a wide set of problems that prevent its usage, scalability, and robustness across a range of real-world problems. Although modern models have been incredibly successful within controlled settings and benchmark data they are not well-suited for situations involving open-domain, noisy, or context-abundant data. These challenges emerge along a range of dimensions including semantic understanding multimodal fusion, generalization, fairness, and evaluation, and together serve to underscore the need for ongoing research and innovation in this field. Undoubtedly the most fundamental and persistent challenge is that of semantic understanding of visual scenes. Captioning images is beyond object detection in an image, it is holistic scene understanding that includes understanding spatial relationships, contextual information, actions, and object interactions. The models need to infer the activity or intent in the image which may involve abstract reasoning or common-sense knowledge that isn't evident. For instance the distinction between an individual holding a tool and using it requires subtle nuances that come too readily to be overlooked by existing architectures. It even becomes more complicated when objects multiply and combine in a dynamic or uncertain environment. Most models produce grammatically correct captions that do not accurately describe the semantics of the scene and therefore provide irrelevant or misleading answers. The other tremendous challenge is obtaining strong multimodal alignment. Making a meaningful caption from an image involves bridging the gap between two fundamentally different representations of data, image features, and language words. The semantic gap between pixel-

based information and human language makes it inherently difficult. Even with the aid of attention mechanisms and transformer-based encoders which are designed to be helpful with alignment most of them still attend to non-informative or deceptive aspects of images while generating. Therefore captions may miss objects lack important information in the scene or fail to set the appropriate relations between things. These errors are most notably observed in scenes involving messy backgrounds occluded objects or abstract visual objects. There is also an issue with generalization ability.

These big image captioning models are often trained on huge labeled collections like MS COCO or Flickr30K, consisting of images and objects with classes. These collections are exhaustive but lack diversity particularly across and within cultures, domains, and geographies. Thus when such models are exposed to out-of-distribution images like art medical images, or factory equipment, they fail miserably. Besides in zero-shot scenarios when the model must generate captions for unseen objects or scenes during training time outputs are generic and disorganized. This gap suggests current inadequacies in learning paradigms and a need for more robust pretraining or adaptation processes. Social and algorithmic captioning bias are also critical concerns. Training sets typically have imbalanced or stereotypical world imageries and if not corrected the model will carry and enlarge such biases. This is especially hazardous where humans are present in the image e.g., home environments, workplaces, or social gatherings. The model can attribute gender roles identify people wrongly or support cultural stereotypes and can end up offending, misinforming, or excluding. Prevention of such biases necessitates not only heterogeneous training data but also fairness-conscious learning algorithms and bias reduction methods at all stages of the training and test pipeline. Along with the problem of generation, there remains the nagging problem of objective assessment. Although scores like BLEU, METEOR, ROUGE-L, and CIDEr are widely used they tend to be biased towards syntactic similarity and do not easily capture the semantic quality or pertinence of captioned text. The scores give shallow reference caption overlaps, rewards, and potentially penalize equally good alternative phrasings. Thus, models that create accurate and fluent

captions can get low scores if their words are not similar to the ground truth. This is especially pronounced for creative, abstractive, or domain-specific tasks with many correct captions. Thus it is equally necessary to create measures that take context, diversity, and human interpretability into account to facilitate the growth of the field.

6.2 Limitation

Given the computational complexities and data issues above there are some practical constraints to limit the deployment, scalability, and real-world robustness of image captioning systems. These are most critical while making the transition from research usage to production-level use where restricting resource options input diversity, interpretability, and responsiveness are success factors. The most elementary constraint is reliance on large-scale annotated datasets. Supervised training methods which prevail in recent image captioning algorithms are highly dependent on the paired image-caption training data. Such data is time-consuming and expensive to prepare with human annotators authoring reviews for thousands or millions of images. Such reviews are also usually constructed for general-purpose usage and will thus focus on over-representing some domains (e.g. common objects, typical activities) but under-representing specialist or esoteric domains like medicine, engineering, or cultural backgrounds. Thus models learned from such datasets overfit their distribution and fail to generalize when used in domain-specific or real-world settings. The similarity of training data restricts the applicability of captioning models in environments that demand accuracy, domain knowledge, or technical vocabulary. The second significant drawback is the inordinate computational expense of learning and using state-of-the-art captioning models. Transformer-based architectures, attention mechanisms, generative adversarial networks (GANs) and diffusion processes are all very computationally intensive particularly when dealing with high image resolutions or intricate sentence generation tasks. Training these models also involves the need for access to high-end hardware frequently GPU or TPU

clusters, that in low-resource environments like academic labs, startups or deployment on mobile and embedded devices may not be feasible. In addition, in inference, the processing time and memory needed to process every image can lead to latency especially for multi-step models or models that have iterative refinement mechanisms like diffusion-based models. A very similar problem is the non-interpretability and non-transparency of models based on deep learning for driving captioning. The majority of modern models are black boxes and little can be said regarding how visual representations are transformed into linguistic tokens in the generation process. The user and developer are not very capable of controlling the model's reasoning process, and typically no one knows why a particular word or phrase was selected in a caption. This lack of transparency is highly dangerous in regulated or high-risk domains like health care, money, or autonomous systems, where visibility into the decision-making is required for validation, accountability, and user confidence. In these domains a caption that might seem plausible but is semantically erroneous can be extremely dangerous if the user does not have any means to validate the output. Moreover the captioning outputs have limited diversity and creative potential in the majority of cases. Though they are sometimes fluent and well-formed sentences, like what they typically are templated or repetitive. The reliance on the most frequent patterns acquired during training renders them repetitive but without variation or depth. For example numerous disparate images can have the same description regardless of changes in content or context. This matters most in applications that are creative in nature or linguistic exercise, like digital narrative, content writing, art practice or disability writing for the blind who must use rich image description to communicate the content of visual media. Scalability and online performance continue to be concerns particularly where models are used in interactive or mobile environments. Though most systems perform efficiently with high precision in offline, controlled environments their efficiency declines under time-critical conditions. Applications like real-time captioning of video, AR overlays, or camera-based navigation systems for accessibility need models to realize visual information and produce captions within milliseconds.

Yet, existing captioning models are plagued with latency due to their depth of architecture and computational needs. This impacts responsiveness and may lower system real-time usability. Lastly adaptability to purpose or use context alteration is required. Most existing captioning systems cannot adjust their descriptions to the target audience, platform, or communicative purpose. For instance an educational application might have to accommodate plain captions to support whereas a commerce gadget might require product characteristics. Lacking the user-tuned or context-sensitive generation current systems may not meet such diverse expectations. Their inability to include constraints or a particular task's preferences built into them prevents them from being integrated into personal systems or role-based applications.

CHAPTER 7

CONCLUSION AND FUTURE SCOPE

7.1 Conclusion

The image-to-text domain is now a very active domain of artificial intelligence in which computer vision's perceptual ability and natural language processing's communicative ability are combined. The task of this domain is to create models that can understand visual information e.g. objects, scenes, and spatial relations, and produce coherent and relevant textual descriptions. This capability has broad applications across a range of fields such as content accessibility, image auto-annotation, accessible tech for the visually impaired, multimedia retrieval, and content creation on online media.

This thesis conducted a thorough review of different deep learning-based methods towards generating image captions from simple encoder-decoder models to advanced ones like attention mechanisms, adversarial networks, zero-shot learning methods and diffusion-based models. Encoder-decoder models normally based on CNNs and RNNs formed the backbone for image captioning tasks to receive visual features extracted and converted into sequences of words. The attention-based models improved upon it by providing dynamic attention towards various parts of an image at every caption generation step, generating contextually more sensitive and targeted descriptions. GAN-based models provided adversarial training that focused on increasing naturalness and diversity of generated captions while zero-shot approaches utilized large-scale vision-language models with no pairwise training. The diffusion models recently emerged as a new paradigm that produces captions via iterative refinement with the potential to provide high-quality, domain-specific, and style-controlled output. Experimental setups

across this work used benchmark datasets like MS COCO and Flickr30K, which provided rich sets of image-caption pairs. Performance of the model was measured using widely employed evaluation metrics like BLEU, METEOR, ROUGE-L, and CIDEr. The report unveiled that although most models are great at syntactic accuracy and object identification they always lag in semantic richness and context coherence. Models occasionally added made-up information that is not present in the image or were unable to identify advanced relations between objects. These findings state that present models are still really far away from replicating the rich perceptual discrimination and expressiveness in human image description. While there is evident improvement in accuracy, diversity, and multilingualism, several systemic constraints still exist. The majority of the models still depend on the availability of large amounts of annotated data for supervised learning limiting their deployment in settings where those are not available. In addition the computational resources required by current architectures especially transformer and diffusion-based ones are an obstacle to real-time deployment as well as scalability in low-resource settings. Another pressing concern is that such systems are not interpretable and transparent enough. One needs to know how the visual inputs are being converted and transformed into the particular textual outputs so one can also hold the system accountable particularly in life-critical domains like medicine or surveillance where uninformed or biased output would be dangerous.

7.2 Future Scope

The future of image captioning research will be centered on making both the generation model's flexibility and contextual awareness better. One important direction is multimodal information integration. Rather than being satisfied with image features next-generation systems can leverage complementary data streams like audio signals, temporal dynamics (in video data), text within embedded regions (through OCR), or structured knowledge from knowledge bases. Such fusion of modalities may lead to more comprehensive, context-sensitive, and situational captions about the setting in

which the image is being viewed. A potential area that could be extremely fruitful is creating specialized captioning systems for domains. General-purpose models fail to satisfy specific communication conventions for areas such as medical imaging, industrial diagnostics, education or legal documentation. By using domain-specific terms formalized structures and semantic constraints in adapting models one can develop systems that are more effective in professional and technical applications. To this end reinforcement learning algorithms and user feedback can be utilized to dynamically optimize the captioning system by adjusting according to patterns of usage as well as users' preferences.

Personalization will be another key area of concern in future models. Instead of spewing out generic captions models will be learned to produce output that is attuned to users' personal interests, writing style, or task-specific requirements. This will be particularly useful in educational software, assistive technology and customer-facing applications. Moreover there is increasing interest in models that are multilingual and operate in low-resource settings. The ubiquity of English in publicly released data and models restricts its availability to English speakers. Cross-lingual training, unsupervised training, and translation-augmented datasets research will be at the center of spreading the application of image captioning systems geographically and culturally.

Technically scalability and efficiency will always be of top priority. Pruning, quantization, and knowledge distillation as tools for model optimization will facilitate deployment to embedded and mobile devices. This will make a real-time image captioning application possible in robotics, augmented reality and smart surveillance with limited computational resources. Ultimately its future will lie in the ethical and evaluative domain. As it becomes more widely used across sensitive domains its bias, fairness, and interpretability become more and more critical. New metrics that involve human judgment, social context, and the utility of captions for downstream tasks will replace or complement traditional n-gram-based metrics.

REFERENCES

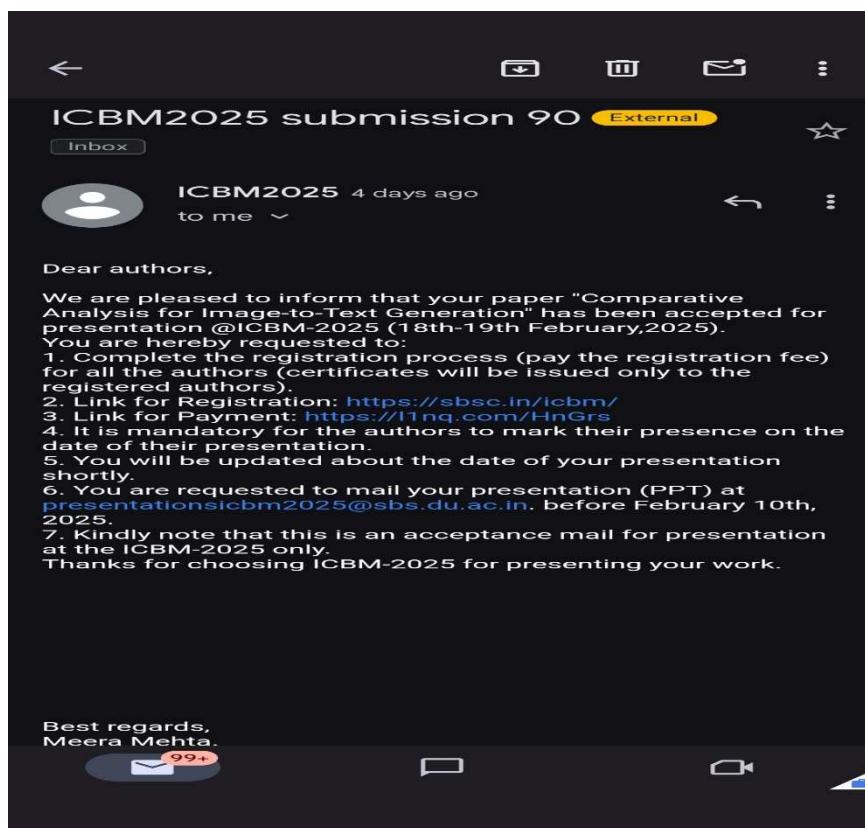
1. Tewel, Y., Shalev, Y., Schwartz, I., & Wolf, L. (2022). Zerocap: Zero-shot image-to-text generation for visual-semantic arithmetic. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 17918–17928.
2. Arif, M. H. (2024). Image-to-Text Description Approach based on Deep Learning Models. *Bilad Alrafidain Journal for Engineering Science and Technology*, 3(1), 33–46.
3. Li, D., Zhao, Y., Cui, R., & Zhao, L. (2021, January). Research on image text generation based on word2vec visual vocabulary attention. *2021 Asia-Pacific Conference on Communications Technology and Computer Science (ACCTCS)*, 344–348. IEEE.
4. Rodriguez, J. A., Vazquez, D., Laradji, I., Pedersoli, M., & Rodriguez, P. (2023). OCR-VQGAN: Taming text-within-image generation. *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 3689–3698.
5. Paliwal, S., Jain, A., Sharma, M., Jamwal, V., & Vig, L. (2024). CustomText: Customized Textual Image Generation using Diffusion Models. *arXiv preprint arXiv:2405.12531*.
6. Koh, J., Park, S., & Song, J. (2024, July). Improving text generation on images with synthetic captions. *2024 16th IIAI International Congress on Advanced Applied Informatics (IIAI-AAI)*, 644–649. IEEE.
7. Lin, Z., Pathak, D., Li, B., Li, J., Xia, X., Neubig, G., ... & Ramanan, D. (2025). Evaluating text-to-visual generation with image-to-text generation. *European Conference on Computer Vision*, 366–384. Springer.
8. Liang, Y., He, J., Li, G., Li, P., Klimovskiy, A., Carolan, N., ... & Navalpakkam, V. (2024). Rich human feedback for text-to-image generation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 19401–19411.

9. Li, Y., Liang, F., Zhao, L., Cui, Y., Ouyang, W., Shao, J., ... & Yan, J. (2021). Supervision exists everywhere: A data efficient contrastive language-image pre-training paradigm. *arXiv preprint arXiv:2110.05208*.
10. Lee, H., Ullah, U., Lee, J. S., Jeong, B., & Choi, H. C. (2021, November). A Brief Survey of text driven image generation and manipulation. *2021 IEEE International Conference on Consumer Electronics-Asia (ICCE-Asia)*, 1–4. IEEE.
11. Hendriksen, M., Bleeker, M., Vakulenko, S., Van Noord, N., Kuiper, E., & De Rijke, M. (2022, April). Extending CLIP for Category-to-image Retrieval in E-commerce. *European Conference on Information Retrieval*, 289–303. Springer.
12. Kabir, R., Haque, N., & Islam, M. S. (2024). A Comprehensive Survey on Visual Question Answering Datasets and Algorithms.
13. Tiwari, V., & Bhatnagar, C. (2021, September). Automatic caption generation via attention based deep neural network model. *2021 9th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO)*, 1–6. IEEE.
14. Padate, R., Jain, A., Kalla, M., & Sharma, A. (2023). Image caption generation using a dual attention mechanism. *Engineering Applications of Artificial Intelligence*, 123, 106112.
15. Bhatt, C., Rai, S., Chauhan, R., Dua, D., Kumar, M., & Sharma, S. (2023, September). Deep Fusion: A CNN-LSTM Image Caption Generator for Enhanced Visual Understanding. *2023 3rd International Conference on Innovative Sustainable Computational Technologies (CISCT)*, 1–4. IEEE.
16. Stefanini, M., Cornia, M., Baraldi, L., Cascianelli, S., Fiameni, G., & Cucchiara, R. (2022). From show to tell: A survey on deep learning-based image captioning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(1), 539–559.
17. Sasibhooshan, R., Kumaraswamy, S., & Sasidharan, S. (2023). Image caption generation using visual attention prediction and contextual spatial relation extraction. *Journal of Big Data*, 10(1), 18.

18. Vinyals, O., Toshev, A., Bengio, S., & Erhan, D. (2015). Show and tell: A neural image caption generator. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3156–3164.
19. Xu, K. (2015). Show, attend and tell: Neural image caption generation with visual attention. *arXiv preprint arXiv:1502.03044*.
20. Karpathy, A., & Fei-Fei, L. (2015). Deep visual-semantic alignments for generating image descriptions. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3128–3137.
21. Chen, X., Fang, H., Lin, T. Y., Vedantam, R., Gupta, S., Dollár, P., & Zitnick, C. L. (2015). Microsoft COCO Captions: Data Collection and Evaluation Server. *arXiv preprint arXiv:1504.00325*.
22. Dosovitskiy, A. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.
23. Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... & Sutskever, I. (2021, July). Learning transferable visual models from natural language supervision. *International Conference on Machine Learning*, 8748–8763. PMLR.
24. 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.
25. You, Q., Jin, H., Wang, Z., Fang, C., & Luo, J. (2016). Image captioning with semantic attention. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4651–4659.
26. Sharma, A., & Singh, R. (2023). ConvST-LSTM-Net: Convolutional spatiotemporal LSTM networks for skeleton-based human action recognition. *International Journal of Multimedia Information Retrieval*, 12(2), 34.

PROOF OF PUBLICATION

PAPER-1



Easebuzz	
THE PRINCIPAL SHAHEED BHAGAT SINGH COLLEGE Shaheed Bhagat Singh College (University of Delhi) Sheikh Sarai Phase II, New Delhi 110017 Mob.: 9540246209 E-Mail: pooja.bhardwaj@sbs.du.ac.in Website: none.in	
RECEIPT	
Bank Reference Number : 461441296075	Payment Date : 2025-02-08 11:42:17
Easebuzz Transaction ID : E2502080D3W9UY	Application Transaction ID : ERT6bptCdn
Total Amount : 750.00	Amount in Words : Rs. Seven hundred fifty Only
Username : Nishant385705@THEIC9Ad	
Form Name : International Conference on Sustainable Business Transformation: Driving Innovation & Impact Through Technology	
First Name Nishant	Last Name Raj
Email nishantraj_23swe13@dtu.ac.in	Contact No 6239936196
Affiliated Institute Delhi Technological University	Category Research Scholar
Note: This is a computer-generated receipt that does not require signature	



12th ICBM

International
Conference on Business
and Management, 2025



Certificate of Paper Presentation


presented by

Department of Commerce,
Shaheed Bhagat Singh College, University of Delhi
Accredited By NAAC With 'A' Grade
in collaboration with
Department of Commerce,
Delhi School of Economics, University of Delhi

This is to certify that Prof./Dr./Mr./Ms. Nishant Raj has presented a paper titled Comparative Analysis for Image-to-Text Generation in the conference on *Sustainable Business Transformation: Driving Innovation & Impact Through Technology* held on February 18th-19th 2025.


Prof. Dr. Arun Kumar Atrree
Chief Patron and Principal
ICBM 2024-25


Prof. Shikha Gupta
Conference Director
ICBM 2024-25


Er. Jayakar Sodagiri
Organizing Secretary
ICBM 2024-25

Publishing Opportunities

The 12th ICBM - 2025 offers several valuable publishing opportunities for researchers, academics, and professionals. These include:

All submissions will undergo a rigorous editorial review process, conducted by a distinguished panel of experts in the field. Authors whose papers are selected through this process will have the opportunity to publish in

- ABDC Journals
- Scopus-indexed Journals
- UGC-referred Journals

This ensures that your work reaches a global audience and contributes meaningfully to the academic and practitioner communities. Through these pathways—proceedings, journal special issues, and book chapters or volumes—the 12th ICBM, 2025 is a platform rich with opportunities to significantly increase academic and professional visibility.

PAPER-2



To,
Abhilasha Sharma
Delhi Technological University
India

Co-Author: Nishant Raj

Dear Sir/Madam,

Greetings of Solidarity from ICEMSS Conferences!!

We are pleased to inform that **Paper ID: ICEMSS_39** with article titled **"Comprehensive Review of Image-to-Text Generation"** is accepted for **Virtual Presentation** during the **"2nd International Conference on Engineering, Management, and Social Sciences (ICEMSS-25)"** Organized by **ZEP Research, India**. The conference will held on **19th-20th February, 2025** in **Delhi, India**.

As a presenter, your contribution will play a vital role in the success of the conference. Your research and presentation on the topic will provide invaluable insights, spark meaningful discussions, and contribute to the overall academic growth of our participants. We are excited about the opportunity to hear your expert views and look forward to the valuable knowledge you will share.

Kindly proceed with the registration, in the given link: <https://www.icemss.in/registration>

If you have any queries, please feel free to contact us. Looking forward to your reply.

Warm regards,

Priyanka Sahu
Organizing Committee
ICEMSS-25
WhatsApp-(+91-8260080050)

 <https://www.icemss.in/>

 [+ \(91\)-8260080050](https://www.icemss.in/)



Thanks for your order, Nishant Raj!

Order # **ORD1738229823444734**

CCAvenue Reference # **113628526080**

Order Date **30 Jan 2025, 15:07 PM**

2nd International Conference on Engineering, Management and Social Sciences

' Bridging disciplines, fostering innovation, and addressing global challenges through interdisciplinary research and collaboration '

Co-organized by- Indraprastha College For Women–Delhi University (IPCW–DU) and Swami Vivekanand Subharti University

Scopus®

Clarivate™

ZEP RESEARCH



Register Now

Submit Your Paper

Download Brochure

Type here to search



CERTIFICATE

- OF PRESENTATION -

THIS CERTIFICATE IS PRESENTED TO :

Nishant Raj

of *"Delhi Technological University, India"* for presenting his/her research article titled *"Comprehensive Review of Image-to-Text Generation"* at the *"2nd International Conference on Engineering, Management and Social Sciences"* held on 19th - 20th Feb 2025 at Indraprastha College for Women-University of Delhi, India, organized by ZEP Research.


Mrs. Priyanka Sahu
Director
Zep Research


Dr. J P Singh
Organizing Secretary
ICEMSS-25


Dr. Alfe M. Solina
Keynote Speaker
ICEMSS-25



**DELHI TECHNOLOGICAL UNIVERSITY**

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

PLAGIARISM VERIFICATION

Title of the Thesis Image-to-text generator

Total Pages _____ Name of the Scholar NISHANT RAJ

Supervisor (s)

(1) Dr. ABHILASHA SHARMA

(2) _____

(3) _____

Department Software Engineering

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

Software used: Turnitin Similarity Index: 6%, Total Word Count: 7991

Date: 20/05/2025

Candidate's Signature

Abhilasha Sharma
Signature of Supervisor(s)



new thesis 1 (1)_merged Nishant11-50.pdf

 Delhi Technological University

Document Details

Submission ID
trn:old::27535:96815552

Submission Date
May 20, 2025, 12:38 PM GMT+5:30

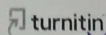
Download Date
May 20, 2025, 12:40 PM GMT+5:30

File Name
new thesis 1 (1)_merged Nishant11-50.pdf

File Size
2.1 MB

40 Pages
7,913 Words
48,743 Characters









6% Overall Similarity

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




Filtered from the Report

- Bibliography
- Cited Text

Match Groups

-  **47 Not Cited or Quoted 6%**
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

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

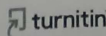
Integrity Flags

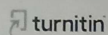
0 Integrity Flags for Review

No suspicious text manipulations found.

Our system's algorithms look deeply at a document for any inconsistencies that would set it apart from a normal submission. If we notice something strange, we flag it for you to review.

A Flag is not necessarily an indicator of a problem. However, we'd recommend you focus your attention there for further review.





*% detected as AI

AI detection includes the possibility of false positives. Although some text in this submission is likely AI generated, scores below the 20% threshold are not surfaced because they have a higher likelihood of false positives.

Caution: Review required.

It is essential to understand the limitations of AI detection before making decisions about a student's work. We encourage you to learn more about Turnitin's AI detection capabilities before using the tool.

Disclaimer

Our AI writing assessment is designed to help educators identify text that might be prepared by a generative AI tool. Our AI writing assessment may not always be accurate (it may misidentify writing that is likely AI generated as AI generated and AI paraphrased or likely AI generated and AI paraphrased writing as only AI generated) so it should not be used as the sole basis for adverse actions against a student. It takes further scrutiny and human judgment in conjunction with an organization's application of its specific academic policies to determine whether any academic misconduct has occurred.

Frequently Asked Questions

How should I interpret Turnitin's AI writing percentage and false positives?

The percentage shown in the AI writing report is the amount of qualifying text within the submission that Turnitin's AI writing detection model determines was either likely AI-generated text from a large-language model or likely AI-generated text that was likely revised using an AI-paraphrase tool or word spinner.

False positives (incorrectly flagging human-written text as AI-generated) are a possibility in AI models.

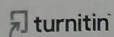
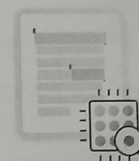
AI detection scores under 20%, which we do not surface in new reports, have a higher likelihood of false positives. To reduce the likelihood of misinterpretation, no score or highlights are attributed and are indicated with an asterisk in the report (*%).

The AI writing percentage should not be the sole basis to determine whether misconduct has occurred. The reviewer/instructor should use the percentage as a means to start a formative conversation with their student and/or use it to examine the submitted assignment in accordance with their school's policies.

What does 'qualifying text' mean?

Our model only processes qualifying text in the form of long-form writing. Long-form writing means individual sentences contained in paragraphs that make up a longer piece of written work, such as an essay, a dissertation, or an article, etc. Qualifying text that has been determined to be likely AI-generated will be highlighted in cyan in the submission, and likely AI-generated and then likely AI-paraphrased will be highlighted purple.

Non-qualifying text, such as bullet points, annotated bibliographies, etc., will not be processed and can create disparity between the submission highlights and the percentage shown.



e-Receipt for State Bank Collect Payment



REGISTRAR, DTU (RECEIPT A/C)

BAWANA ROAD, SHAHABAD DAULATPUR, , DELHI-110042
Date: 20-May-2025

SBCollect Reference Number :	DU00951720
Category :	Miscellaneous Fees from students
Amount :	₹3000
University Roll No :	2k23/swe/13
Name of the student :	Nishant Raj
Academic Year :	2024-2025
Branch Course :	Software engineering
Type/Name of fee :	Others if any
Remarks if any :	M.tech 2nd year thesis fees
Mobile No. of the student :	7079451926
Fee Amount :	3000
Transaction charge :	0.00
Total Amount (In Figures) :	3,000.00
Total Amount (In words) :	Rupees Three Thousand Only
Remarks :	M.tech 2nd year thesis fees

Notification 1:	Late Registration Fee, Hostel Room rent for internship, Hostel cooler rent, Transcript fee (Within 5 years Rs.1500/- & \$150 in USD, More than 5 years but less than 10 years Rs.2500/- & \$250 in USD, More than 10 years Rs.5000/- & \$500 in USD) Additional copies Rs.200/- each & \$20 in USD each, I-card fee, Character certificate Rs. 500/-.
Notification 2:	Migration Certificate Rs.500/-, Bonafide certificate Rs.200/-, Special certificate (any other certificate not covered in above list) Rs.1000/-, Provisional certificate Rs.500/-, Duplicate Mark sheet (Within 5 years Rs.2500/- & \$250 in USD, More than 5 years but less than 10 years Rs.4000/- & \$400 in USD, More than 10 years Rs.10000/- & \$1000 in USD)



DELHI TECHNOLOGICAL UNIVERSITY

Shahbad Daulatpur, Main Bawana Road, Delhi-42

Proforma for Submission of M.Tech. Major Project

01. Name of the Student. NISHANT RAJ
02. Enrolment No. 2K73/SWE/13
03. Year of Admission 2023
04. Programme M.Tech., Branch SWE
05. Name of Department SOFTWARE ENGINEERING
06. Admission Category i.e. Full Time/ Full Time (Sponsored)/ Part Time: FULL TIME
07. Applied as Regular/ Ex-student REGULAR
08. Span Period Expired on
09. Extension of Span Period Granted or Not Granted (if applicable)
10. Title of Thesis/Major Project

11. Name of Supervisor DR. ADHILASHA SHARMA

12. Result Details (Enclose Copy of Mark sheets of all semesters) :

S. No.	Semester	Passing Year	Roll No.	Marks Obtained	Max. Marks	% of Marks	Details of Back Paper Cleared (if any)
01.	1 st	2023	23/SWE/13	7.65	10	76.5	
02	2 nd	2024	23/SWE/13	8.35	10	83.5	
03	3 rd						
04	4 th (P/T only)						
05	5 th (P/T only)						

13. Fee Details (Enclose the Fee Receipt):

Amount Paid (in Rs.) <u>3000</u>	Receipt No. <u>DU00951720</u>	Date <u>20/05/2025</u>
-------------------------------------	-------------------------------	------------------------

Nishant Raj
Signature of Student

It is certified that the name of Examiners for evaluation of the above thesis/ project have already been recommended by the BOS.

Adhila Sharma
Signature of Supervisor

[Signature]
Signature of HOD with Seal

(Instructions for filling up the Form may see on back side please.)



S. No.

406343

DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi College of Engineering)

STATEMENT OF GRADES

Master of Technology

(Software Engineering)

Name : NISHANT RAJ

Roll No. : 23/SWE/13

Month & Year of Examination : NOVEMBER, 2023

Semester : FIRST

Subject Code	Subject Title	Credits	Credits Secured	Grade
SWE5405	ADVANCED OPERATING SYSTEM	4	4	B+
SWE5301	PROJECT WORK	3	3	A
SWE5201	SEMINAR	2	2	B+
SWE501	SOFTWARE REQUIREMENT ENGINEERING	4	4	A
SWE503	OBJECT ORIENTED SOFTWARE ENGINEERING	4	4	A
		17	17	

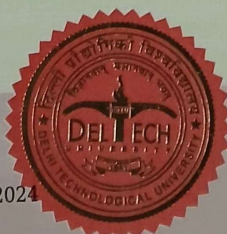
AB : Absent DT : Detained

Credits Secured / Total : 17 / 17

SGPA : 7.65

Dated : May 10, 2025

Date of Declaration of Result : March 01, 2024



CONTROLLER OF EXAMINATION

S. No. **406278****DELHI TECHNOLOGICAL UNIVERSITY**

(Formerly Delhi College of Engineering)

STATEMENT OF GRADES**Master of Technology****(Software Engineering)**

Name : NISHANT RAJ

Roll No. : 23/SWE/13

Month & Year of Examination : MAY, 2024

Semester : SECOND

Subject Code	Subject Title	Credits	Credits Secured	Grade
SWE502	SOFTWARE TESTING	4	4	A+
SWE504	EMPIRICAL SOFTWARE ENGINEERING	4	4	A+
SWE5204	PREDICTIVE MODELLING	2	2	B+
SWE5302	MINOR PROJECT	3	3	A
SWE5406	MACHINE LEARNING	4	4	A
		17	17	

AB : Absent DT : Detained

Credits Secured / Total : 17 / 17**SGPA : 8.35**

Dated : May 10, 2025

Date of Declaration of Result : July 16, 2024



R. Pandey
CONTROLLER OF EXAMINATION



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

CERTIFICATE OF FINAL THESIS SUBMISSION

(To be submitted in duplicate)

1. Name: NISHANT RAI
2. Roll No: 2K23/SWE/13
3. Thesis title: Image-to-text generator
4. Degree for which the thesis is submitted: M.Tech (SWE)
5. Faculty (of the University to which the thesis is submitted)
Dr. Abhiksha Sharma
6. Thesis Preparation Guide was referred to for preparing the thesis. YES ☒ NO ☐
7. Specifications regarding thesis format have been closely followed. YES ☒ NO ☐
8. The contents of the thesis have been organized based on the guidelines. YES ☒ NO ☐
9. The thesis has been prepared without resorting to plagiarism. YES ☒ NO ☐
10. All sources used have been cited appropriately. YES ☐ NO ☐
11. The thesis has not been submitted elsewhere for a degree. YES ☒ NO ☐
12. All the correction has been incorporated. YES ☒ NO ☐
13. Submitted 2 hard bound copies plus one CD. YES ☒ NO ☐

(Signature(s) of the Supervisor(s))

Name(s): Abhiksha Sharma

(Signature of Candidate)

Name: NISHANT RAI

Roll No: 2K23/SWE/13



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

CERTIFICATE OF THESIS SUBMISSION FOR EVALUATION
(Submit in Duplicate)

1. Name: NISHANT RAI
2. Roll No.: 2K23/SWE/13
3. Thesis title: Image-to-text generator
4. Degree for which the thesis is submitted: M.Tech
5. Faculty of the University to which the thesis is submitted:
Dr. Abhilasha Sharma
6. Thesis Preparation Guide was referred to for preparing the thesis: YES ☒ NO ☐
7. Specifications regarding thesis format have been closely followed. YES ☒ NO ☐
8. The contents of the thesis have been organized based on the guidelines YES ☒ NO ☐
9. The thesis has been prepared without resorting to plagiarism. YES ☐ NO ☐
10. All sources used have been cited appropriately. YES ☒ NO ☐
11. The thesis has not been submitted elsewhere for a degree. YES ☒ NO ☐
12. Submitted 2 spiral bound copies plus one CD. YES ☒ NO ☐

(Signature of Candidate)

Name(s): NISHANT RAI
Roll No: 2K23/SWE/13

Abhilasha Sharma