

ADVANCING IMAGE CAPTIONING WITH TRANSFORMER BASED TECHNIQUES

A DISSERTATION

SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE AWARD OF THE DEGREE
OF

MASTER OF TECHNOLOGY
IN
DATA SCIENCE

Submitted by

AMAN RAWAT
(23/DSC/03)

Under the supervision of
Mrs. PRIYA SINGH



SOFTWARE ENGINEERING
DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Bawana Road, Delhi 110042

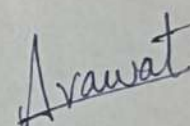
MAY, 2025

DEPARTMENT OF SOFTWARE ENGINEERING
DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Bawana Road, Delhi-110042

CANDIDATE'S DECLARATION

I, AMAN RAWAT, Roll No's – 23/DSC/03 students of M.Tech (DATA SCIENCE), hereby declare that the Dissertation titled “**ADVANCING IMAGE CAPTIONING WITH TRANSFORMER BASED TECHNIQUES**” which is submitted by me to the SOFTWARE ENGINEERING, Delhi Technological University, Delhi in partial fulfilment of the requirement for the award of 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.

Place: Delhi


AMAN RAWAT

Date: 19/5/23

DEPARTMENT OF SOFTWARE ENGINEERING
DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Bawana Road, Delhi-110042

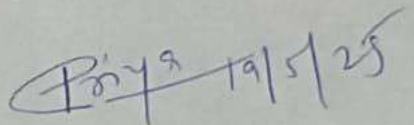
CERTIFICATE

I hereby certify that the Project Dissertation titled “**ADVANCING IMAGE CAPTIONING WITH TRANSFORMER BASED TECHNIQUES**” which is submitted by AMAN RAWAT, Roll No's – 23/DSC/03, SOFTWARE ENGINEERING ,Delhi Technological University, Delhi in partial fulfilment of the requirement for the award of the degree of Bachelor of Technology, is a record of the project work carried out by the students under my supervision. To the best of my knowledge this work has not been submitted in part or full for any Degree or Diploma to this University or elsewhere.

Place: Delhi

Date:

19/5/25

 19/5/25

Mrs. PRIYA SINGH

SUPERVISOR

DEPARTMENT OF SOFTWARE ENGINEERING
DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Bawana Road, Delhi-110042

ACKNOWLEDGEMENT

I wish to express our sincerest gratitude to Mrs. PRIYA SINGH for his continuous guidance and mentorship that he provided me during the project. He showed me the path to achieve our targets by explaining all the tasks to be done and explained to me the importance of this project as well as its industrial relevance. He was always ready to help me and clear our doubts regarding any hurdles in this project. Without his constant support and motivation, this project would not have been successful.

Place: Delhi

AMAN RAWAT

Date:

Abstract

Image captioning relates to the automatic generation of natural language descriptions for visual content, and It has seen major progress through the acceptance of deep learning methods.. This thesis critically explores the transformation of image captioning methods, with a particular focus on the transformative impact of Vision Transformers (ViTs) . While common methods employing CNNs and RNNs had provided initial advancements their basis, they are generally poor at understanding global context and relationships within the entire image. Vision Transformers overcome this deficit by employing self-attention and allowing thorough understanding of fine detail as much as overall context of the image. This study compares ViT-based models with traditional techniques across a variety of architectures and benchmark datasets, particularly MS COCO. The findings indicate that ViT-based approaches significantly outperform conventional models in generating semantically rich and contextually accurate captions. Additionally, this thesis introduces a novel image captioning framework ViBERT, which merges advantages of both Vision transformer and Bidirectional Encoder Representations from Transformers in an encoder-decoder architecture. Sometimes traditional models often fail in capturing the long range semantic dependencies and global visual setting, ViBERT effectively leverages ViT's visual attention and BERT's deep contextual understanding to generate more strong and semantic correct description. The performance of the proposed model is calculate using standard performance measures.

Contents

Candidate's Declaration	i
Certificate	ii
Acknowledgement	iii
Abstract	iv
Content	vi
List of Tables	vii
List of Figures	viii
List of Symbols, Abbreviations	ix
1 INTRODUCTION	1
1.1 Motivation	1
1.2 Problem Statement	1
1.3 Overview	2
1.4 Objectives	3
2 LITERATURE REVIEW	4
2.1 Image Feature Extraction	4
2.1.1 Convolutional Neural Network	5
2.1.2 Vision Transformer	5
2.2 NLP in Image Captioning	6
2.2.1 Recurrent Neural Networks	6
2.2.2 Gated Recurrent Units	7
2.2.3 Long Short Term Memory	8
2.2.4 Bidirectional Encoder Representations from Transformers	9
2.3 Image Caption Generators	10
3 METHODOLOGY	12
3.1 An Overview of Image Captioning Model	12
3.2 Architecture of Proposed Model	13
3.3 Baseline Model	15
3.4 Dataset Description	16
3.4.1 Flickr 8k Dataset	16
3.4.2 MSCOCO dataset	16
3.5 Parameter Settings	17

3.6	Performance Metrics	17
4	RESULTS and DISCUSSION	18
4.1	Result of Review of ViT Based Image captioning Technique	18
4.1.1	Performance Comparison of ViT-based and Conventional Image Captioning Technique	18
4.1.2	Limitation of Conventional and ViTs-based Technique	19
4.2	Result of Proposed ViBERT Model	20
4.2.1	Performance Evaluation of ViBERT Compared to Other Leading Captioning Models	20
4.3	Overall Findings	21
5	CONCLUSION AND FUTURE SCOPE	23
5.1	Future Work	24
A	LIST OF PUBLICATION	25

List of Tables

2.1	Comparison of different Baseline Models	11
4.1	ViT-based image captioning techniques On MSCOCO Dataset	19
4.2	Conventional image captioning techniques On MSCOCO Dataset	19
4.3	Comparison of different image captioning models on various performance metrics	21

List of Figures

2.1	This diagram shows a CNN architecture,	5
2.2	ViT splits images into patches, processes them with transformers, and outputs class predictions via MLP [1].	6
2.3	A simple feedforward neural network illustrating connections between input, hidden, and output layers.	7
2.4	Diagram illustrating the GRU architecture showing the flow between input, hidden states, update gate, reset gate, and output.	8
2.5	Illustration of an LSTM cell showing the flow through forget, input, and output gates for managing distant relationship in consecutive data.	8
2.6	Illustration of a transformer encoder processing input tokens with a masked token, followed by a classification layer using GELU activation and normalization.	9
3.1	This model have of two main components: a transformer as the decoder and a CNN as the encoder.	12
3.2	This model have of two main components: a ViT as the encoder and a transformer as the decoder.	13
3.3	Proposed image captioning model architecture combines ViT for extracting image features with BERT for generating context-aware captions within an encoder-decoder setup, optimized using the AdamW algorithm.	14
4.1	It shows the training loss for each epoch	21
4.2	ViBert model-generated caption	22
4.3	ViBert model-generated caption	22
A.1	Screenshot	26
A.2	Scopus Index Screenshot	27
A.3	Paper 1 acceptance Screenshot	28
A.4	Paper 2 acceptance Screenshot	28

List of Symbols

ViTs	Vision Transformers
CNNs	Convolutional Neural Networks
RNNs	Recurrent Neural Networks
LSTMs	Long Short Term Memorys
SPICE	Semantic Propositional Image Caption Evaluation
GRUs	Gated Recurrent Units
AdamW	Adam with Weight Decay
BERT	Bidirectional Encoder Representations from Transformers
BLEU	Bilingual Evaluation Understudy
CIDEr	Consensus-based Image Description Evaluation
ROUGE-L	Recall-Oriented Understudy for Gisting Evaluation
SPICE	Semantic Propositional Image Caption Evaluation
DAIT	Dual-Adaptive Interactive Transformer
GCS-M3VLT	Guided Context SelfAttention based Modal Medical Vision Language Transformer
NLP	Natural Language Processing
RBBA	ResNet- BERT- Bahdanau Attention
MSCOCO	Microsoft Common Objects in Context

Chapter 1

INTRODUCTION

1.1 Motivation

In today's digital world, images are everywhere on social media, news articles, educational platforms, and enabling machines to interpret and describe images using natural language continues to be a difficult challenge. Image captioning is in the intersection of computer vision and NLP (Natural Language Processing), can provide additional access to content and enable automatic content creation. Old methods involving CNNs (Convolutional Neural Networks) and RNNs (Recurrent Neural Networks) were good background work, but tend to falter while detecting global context of an image or context relevance in longer sentences. Transformer models changed everything. Vision Transformers (ViTs) can see the whole image by taking the picture as a patch sequence, and BERT (Bidirectional Encoder Representations from Transformers) understands language more effectively by infusing global context of the sentence. The motivation for this project is that uniting the best of ViT and BERT in one architecture has the ability to produce more fluent and accurate captions. Through this fusion, the goal is to create a model that surpasses current boundaries and produces captions that not only can be understood but are human.

1.2 Problem Statement

Traditional models, with the overall approach of combining CNNs for visual feature extraction and RNNs or LSTMs (Long Short Term Memorys) for sequence generation, have performed well enough but are beset with several shortcomings. These include failure to capture long-range dependencies, restricted global image context comprehension, and generation of generic or shallowness in semantics in captions. Moreover, the sequential nature of RNNs makes training less efficient and harder to parallelize. With recent advances in transformer-based models, ViTs have appeared as a powerful choice to CNNs by enabling global attention across image patches, while BERT has proven highly effective in capturing rich, bidirectional context in language modeling. Despite both of them having individual success, the integration of ViT and BERT for captioning images is relatively underdeveloped. The aim of this work is to fill this gap by introducing a cross functional unified transformer called as ViBERT that leverages the strength of ViT in understanding vision and uses BERT for generating fluent, coherent, and contextually relevant captions to overcome the limitations of existing CNN-RNN models.

1.3 Overview

The capacity to create descriptive text based on visual material, more often referred to as image captioning, is a complex and difficult challenge. This type of interdisciplinary problem demands a system to properly comprehend the semantic meaning of an current picture and tell it in native, glib language. Currently image captioning getting more significant over the last few years in real world applications. Conventionally, image captioning models have been developed on top of encoder-decoder architectures, in which CNNs are used to mine visual features from the image and RNNs, including sophisticated variants like LSTM networks and GRUs, are utilized to produce textual descriptions step by step. While these architectures have facilitated tremendous progress in this domain, they suffer from a number of the most important limitations. One of them is their inability to realize long-range relationship and preserve a contextual global context of the image. It especially when the scene includes more than one interacting object or intricate visual relationships. The sequential nature of RNNs is also restrictive in parallelization at the time of training, rendering them computationally expensive for large-scale use. The discovery of transformer architectures has introduced a revolution in computer vision as well as NLP, providing a more efficient and scalable solution. ViTs are introduced as a replacement for CNNs. It have demonstrated a strong ability to model global dependencies by dividing pictures into certain fixed size patches and the method of using attention mechanisms. This enable model to learn intricate spatial relationships across the entire image without relying on hierarchical feature extraction. Instead of analyzing images using convolutional layers like CNNs. ViT treat an image as a arrangement of patches just like a sequential data made up of words. To begin with, the image is fragmented into small, fixed-size patches for instance. These pieces of pictures are then changed into a vector and taken through a linear layer to create patch embeddings. A special classification token is put into the sequence, position embeddings are included so the model can understand the order and position of each patch. Once the patches have been prepared, the entire sequence is put into a encoder, which employs attention in order to enable each patch to talk to every other patch in the image. This is what makes ViTs have a significant edge—ViTs are able to find out relationships between far-apart parts of the image, which CNNs, which are primarily concerned with local features, are not able to. The self-attention process assists the model in determining what components of the picture are applicable for the task in question, whether classifying an object or detecting text. After a number of layers of this feedforward and self-attention processing, the output of the classification token is generally utilized for prediction. On certain tasks, the full output sequence may be employed to take more sophisticated decisions, like producing a caption or dividing out portions of the image.

In parallel, BERT has revolutionized the field of NLP by introducing deep bidirectional learning of textual context. In contrast to earlier language models, BERT handles left and right word context together, enabling richer and more accurate meaning representation. In combination, ViTs and BERT represent an unprecedented chance to develop multimodal systems able to comprehend visual input and produce highly coherent and semantically correct textual output.

In this thesis, we provide an extensive investigation of transformer-based image captioning, specifically the incorporation of Vision Transformers and BERT into a single encoder-decoder model, called ViBERT. The overarching goal of this work is to overcome the in-built restrictions of traditional CNN-RNN models by exploiting ViT’s capacity to encode

global visual context and BERT’s better performance in modeling deep semantic relations in words. The new ViBERT model is stringently tested on standard benchmark data, like Flickr8K, that offers a rich variety of images and human-created captions. In order to measure the model’s performance and efficiency, we utilize commonly accepted metrics such as BLEU (Bilingual Evaluation Understudy), CIDEr (Consensus-based Image Description Evaluation), ROUGE-L (Recall-Oriented Understudy for Gisting Evaluation), and SPICE (Semantic Propositional Image Caption Evaluation). In experiment, the result show that ViBERT perform superior to many baseline models, generating captions that are not just grammatically correct but also contextually pertinent and semantically dense. This research highlights the increasing importance of transformer-based multi-modal architectures and presents a foundation for future research into automated visual comprehension and language translation.

1.4 Objectives

This thesis seeks to make a valuable contribution to the area of image captioning by investigating the effect of different techniques on image captioning tasks. In particular, this thesis seeks to accomplish the following aims:

- To compare Traditional image captioning models and ViT-based image captioning models.
- To investigate the performance of ViTs-based image captioning methods and the traditional image captioning method on various performance metrics.
- To determine the limitation of the traditional and ViTs-based approach.
- To develop an image captioning model by the name of ViBERT which incorporates ViT for image understanding and BERT for caption generation.
- To train and fine-tune ViBERT on Flickr8K dataset.
- To test ViBERT on BLEU, ROUGE-L, CIDEr, and SPICE.

Chapter 2

LITERATURE REVIEW

In recent years, image captioning is emerging as a significant research area, integrating computer vision and natural language. Early approaches predominantly utilized CNNs in combination with RNNs, particularly LSTM networks, to extract visual features and generate captions in sequence. Early methods mostly made use of CNNs together with RNNs, especially LSTM networks, to learn visual features and produce captions sequentially. Vinyals et al. [2] were the first to innovate the "Show and Tell" model that made use of a standard model for extracting image features and an LSTM for producing textual descriptions. Xu et al. present the "Show, Attend, and Tell" model in which it includes an attention mechanism to guide it towards the relevant areas of the image while generating the captions [3]. Although CNN-based methods have achieved success, they usually struggle to capture the overall meaning of an picture. It lead them to the creation of incorrect captions for sophisticated images. The current attention mechanism is present by Vaswani et al. for machine text generation that transformed the area of image captioning [4]. Transformers support the modeling of long range dependencies without sequential processing, which makes them suitable for task requiring a deep understanding of global context. Following this premise, Dosovitskiy et al. [1] present the ViT that show that a transformer model without using convolutions will perform well on image classification task. Zhu et al. [5] developed this further for image captioning by embedding ViTs into a transformer-based language model, proving that ViTs were able to outperform Traditional image captioning methods in caption quality, especially for images with complex compositions.

2.1 Image Feature Extraction

In image captioning, feature extraction plays very important role in helping the model understand what's in the image. It starts by using deep learning techniques, most often CNNs such as ResNet or VGG, to pick out key visual elements such as objects, textures, and shapes. These models convert unprocessed image data into semantic vectors that capture the essence of the image. Some state-of-the-art systems take it one step ahead by employing object detectors such as Faster R-CNN to locate specific regions in the image, providing the model with more nuanced input to work with. More recently, ViTs have also been utilized to divide the image into tiny patches and study the interrelationship between them so that subtle detail and wider context can be understood. The better the features are extracted, the more accurate and detailed the resulting captions are likely to be.

2.1.1 Convolutional Neural Network

Visual feature extraction within image captioning has always been dependent largely on CNN, which has been found to be highly effective at learning image patterns. To achieve rich picture information, previous architectures such as VGG16 and VGG19 utilized deep convolutional layers. These models use small convolutional filters to analyze hierarchical characteristics, but struggle with complex spatial relationships. To overcome these difficulties, more sophisticated CNN architectures were developed, such as ResNet, which incorporates residual connections to enable deeper networks without running into issues with vanishing gradients [6].

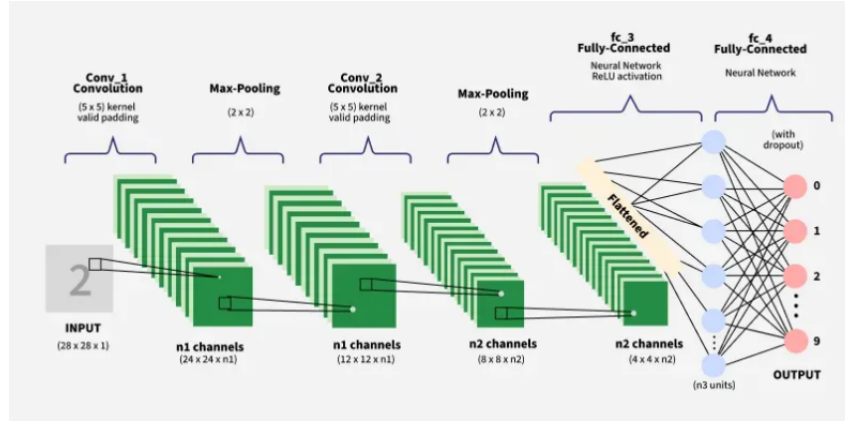


Figure 2.1: This diagram shows a CNN architecture,

ResNet models such as ResNet-50 and ResNet-152 have been extensively employed for image captioning because they can detect deep features of images. Inception also enhances the performance of the model by employing multi-scale convolutional filters within a single layer in order to detect different patterns in images [7]. EfficientNet has emerged as a powerful architecture that combines accuracy with efficiency, making it a strong fit for applications that require real-time image captioning. CNNs are effective at visual processing tasks, but their ability to understand broader spatial relationships in an image is restricted by their attention to localized features. As a result, there has been a growing shift toward using transformer-based models in computer vision.

2.1.2 Vision Transformer

ViTs are a recent image understanding paradigm that have exhibited robust performance in applications like image captioning. In contrast to the localized filters used by conventional CNNs, ViTs illustrated in Fig. 2.2 split an image into fixed-sized pieces and flatten them. Then we feed each piece as a token in an order like sequential data [1].

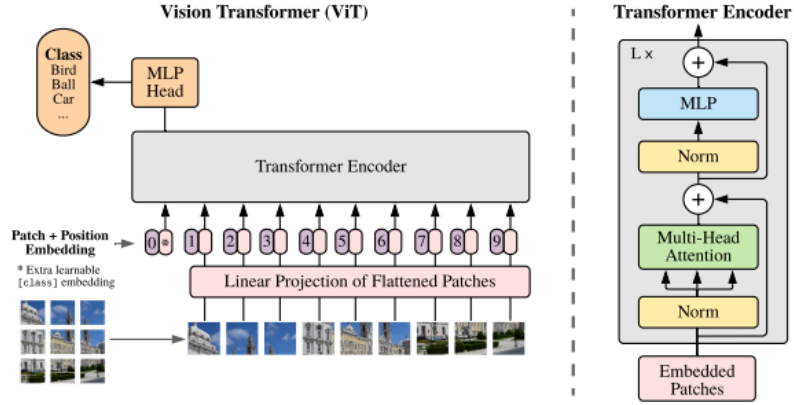


Figure 2.2: ViT splits images into patches, processes them with transformers, and outputs class predictions via MLP [1].

Feeding tokens to the transformer model with self-attention employed for learning patch relations so that the network can learn global context efficiently. Such global reasoning makes ViTs suitable for challenging visual tasks that can be enhanced by understanding the scene as a whole, for example, generating well-aware and contextual captions [5]. Yet they tend to need big sets of data and computation to be trained from scratch. Dosovitskiy et al. [1] introduced ViTs in a significant departure from how deep learning models are learning and representing visual features.

2.2 NLP in Image Captioning

NLP is crucial for transforming visual depictions to useful text descriptions. Previous captioning techniques of images usually employed RNNs, as well as more sophisticated forms such as LSTMs and GRUs, to produce sequences of texts from visual inputs [8]. LSTM is presented to solve the vanishing gradient problem which is common in standard RNNs. It mostly uses memory cells and gated units that retain important information for longer inputs. GRUs minimize this mechanism by integrate the forget gate and input gate into a single update gate. It minimize the computational expense without loss of ability to learn temporal patterns.

2.2.1 Recurrent Neural Networks

RNNs are neural networks suited for handling sequential data. While feedforward neural networks process inputs in isolation. RNNs possess an internal state which is refreshed step by step [8]. It enable them to store context about past inputs as illustrated in Fig. 2.3. RNNs are particularly suitable for most tasks like language modeling, speech recognition and image captioning, where the ordering and dependency of input are significant. As we know they can handle distant relationship, typical RNNs are plagued by vanishing and exploding gradients when trained on long sequences, which greatly hinders their usefulness in practice[9].

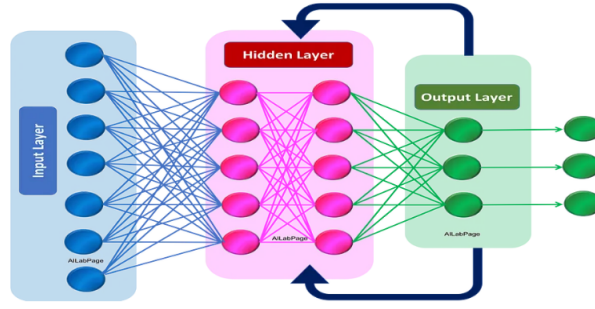


Figure 2.3: A simple feedforward neural network illustrating connections between input, hidden, and output layers.

As back-propagated gradients go deeper in time, they either shrink to the point of being close to zero or grow exponentially, rendering learning troublesome. This constrain has given rise to the construction of better architectures such as LSTM and GRU, which implement gating mechanisms to control information flow with time[10]. In image captioning, RNNs find their application in the form of decoders of encoder-decoder models. A CNN first derives the visual characteristics from the image, and in decoder we use RNN as it generates a caption by iterating through these features as the initial state and outputting one word at a time in order. This was initiated by models such as the "Show and Tell" model, which was able to combine CNNs with RNNs to produce natural-sounding image descriptions [2]. Although RNNs initiated deep learning sequence modeling, inefficiency in dealing with long-range dependencies and sequential training inefficiency have caused deep learning to adopt transformer-based models, which are able to achieve greater parallelism and performance on the majority of natural language processing tasks [8].

2.2.2 Gated Recurrent Units

GRUs are another form of RNN architecture is a simpler, computationally less expensive version than LSTM networks[9]. Whereas LSTMs use three gates which are input, forget, and output gate. GRUs use only two gates.

- The update gate specifies how much of the history must be propagated forward to the future[8].
- The reset gate, determining how much of the history to forget[8].

In Fig.2.4, we can see that GRUs also integrate the cell state and cell state into a single vector. It keeps it architecture simple but still allowing the network to keep useful information over long sequences[9]. This simplified design leads to faster training time with fewer parameters. GRUs make appealing in those applications where computational resources are not plentiful.

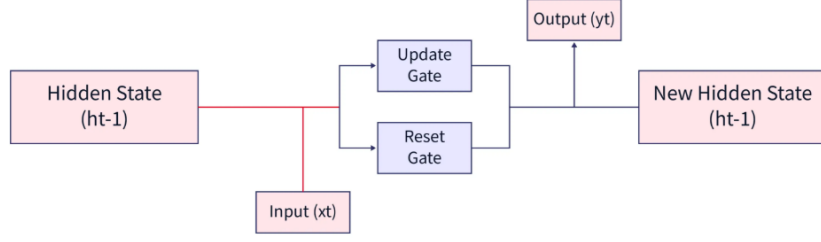


Figure 2.4: Diagram illustrating the GRU architecture showing the flow between input, hidden states, update gate, reset gate, and output.

In image captioning, GRUs tend to serve as decoders in encoder-decoder models. The GRU decoder is given visual feature vectors as input along with a CNN-based encoder and produces captions one word at a time. Since they have a less complex structure yet perform similarly, GRUs have proven to be competitive with LSTMs, particularly on the case of medium-length sequences and low-resource settings[9]. GRUs offer varying modeling power to LSTMs but fail to deal with extremely long sequences in as good a manner as desired fine-grained memory control. However, their faster convergence and lower training cost make them used extensively in real-time systems and light-weight applications.

2.2.3 Long Short Term Memory

LSTM mark the shortcomings of RNNs, more so the vanishing and exploding gradients experienced during training on long sequences[9]. As opposed to regular RNNs, which fail to maintain information over large numbers of time steps, LSTMs are designed specifically to remember past dependencies over long periods via an advanced internal structure revealed in Fig. 2.5[8].

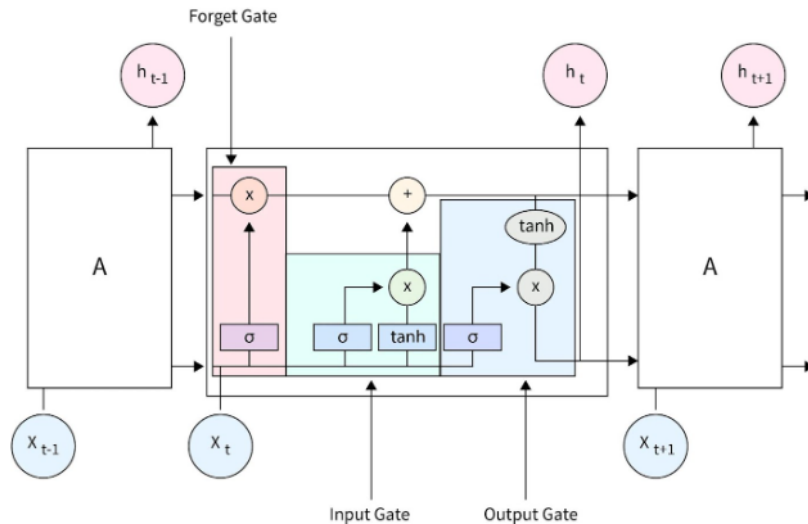


Figure 2.5: Illustration of an LSTM cell showing the flow through forget, input, and output gates for managing distant relationship in consecutive data.

An LSTM unit contains three main gates and they are the input gate, forget gate and output gate, plus a cell state that functions as a memory buffer[9]. These gates guide the information in and out of the memory cell:

- The input gate decide what new information should be saved[9].
- The forget gate decide what information should be forget.
- The output gate decide which cell state component has to be moved to the next concealed state.

Through this gating mechanism, LSTMs can maintain and update context over extended sequences and are therefore particularly effective for applications. In image captioning, in particular, LSTMs are typically employed as decoders within encoder-decoder architectures. While the encoder draws out visual features with an image, the LSTM decoder produces the equivalent textual description, one word at a time[10]. Research has proved that LSTM-based models greatly surpass simple RNNs in generating captions since they can capture sequential dependencies more vigorously. In spite of their effectiveness, LSTMs come with limitations[10]. They handle sequences sequentially, a factor that discourages training in parallel and may lead to more time-consuming training. In addition, although they perform better with moderately long sequences than RNNs, their performance still deteriorates as sequences become extremely long, and hence the emergence of transformer-based architectures that can better manage long-range dependencies with self-attention mechanisms.

2.2.4 Bidirectional Encoder Representations from Transformers

BERT revolutionized NLP as we are building deep contextualized word representations, which enable the models to learn rich word relationships [11]. In contrast to sequential RNNs, transformers process the whole text sequence simultaneously and produce more coherent and context-aware captioning.

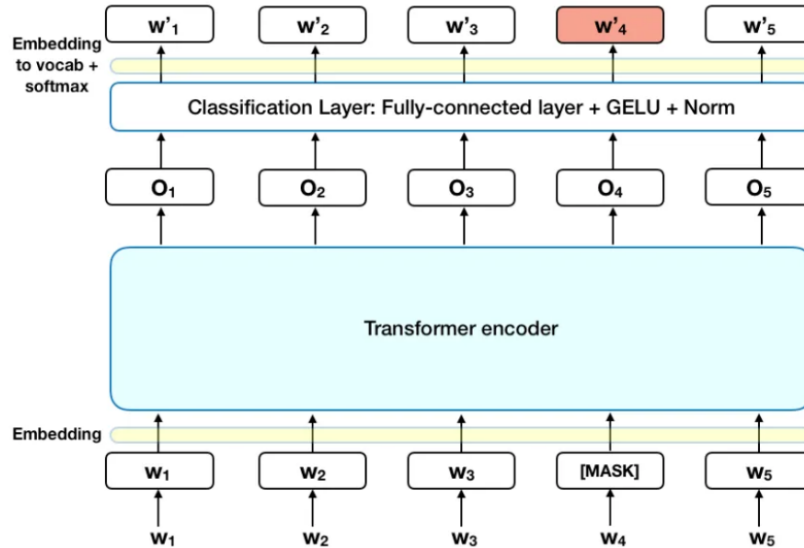


Figure 2.6: Illustration of a transformer encoder processing input tokens with a masked token, followed by a classification layer using GELU activation and normalization.

Another popular transformer model, Generative Pre-trained Transformer, demonstrated robust performance in language generation tasks [11]. Even so, BERT is still the choice of captioning because it's bidirectional and will possess more context sense.

The combination of CNNs for extracting the visual information and transformers that have been used to generate the text has contributed significant impact in image captioning methods[11].

2.3 Image Caption Generators

Image captioning creates contextually appropriate text for images with the fusion of language and vision models. The methods where we employ normal methods to obtain the image features and RNNs to produce the caption. When we employ the attention mechanisms to direct the model to the significant features of the image [9]. This sort of model avoids issues with analyzing parts of an image at once. Thus, it's more suitable to process heavy and complex visual scenes more accurately and more efficiently. Current transformer models have used ViT for vision encoding [10]. ViT splits the image into patches of the same size and processes the patches in parallel fashion [7]. Experiments have shown that transformer models are particularly worth using while working with images and text. They have been dominating all the past CNN-RNN models on most of the benchmarks and still providing reasonable performance. Below Table 2.1 provides a comparative summary of leading image captioning models, describing their key strengths and weaknesses as far as accuracy, complexity, and usability are concerned. It gives a feature-by-feature comparison of six image captioning models and their strengths and weaknesses.

Table 2.1: Comparison of different Baseline Models

S.No.	Model Name	Advantages	Disadvantages
1	ResNet50-BERT-Bahdanau attention (RBBA) [12]	Highest BLEU-1 (0.5321) and BLEU-4 (0.1263) scores, blending a snippet of ResNet50’s feature extraction ability with BERT’s language understanding ability and bahdanau attention enhances semantic alignment	Highest parameter utilization can go up to 119M and long training time of 7488s, thereby drawing in real-time and It is heavy in resource utilization.
2	Guided Context SelfAttention based Multi-modal Medical Vision Language Transformer (GCS-M3VLT) [13]	Combines the visual feature and the diagnostic keywords. Increases focus with Guided Context Attention. Lightweight and less training data intensive	It is not a overall generalizer
3	Entrocap [14]	Supports zero-shot captioning without parallel image-text data Identifies frequent and rare regions with entropy-based retrieval	Highly advanced architecture; strongly dependent on CLIP embeddings Lacks a dedicated visual decoder
4	Dual-Adaptive Interactive Transformer (DAIT) [15]	Employ dual-adaptive mechanism and interactive modules for ultra-interactive multi-modes Adaptive Interactive Decoder guarantees contextual captions	Complex and resource-hungry because of numerous transformer modules Based on textual external descriptions
5	Encoder-decoder based on Fourier Transform [16]	It makes use of Fourier Transform for improved feature representation and to capture periodic and spatial patterns It is compliant with basic transformer architecture	Lower Meteor score compared to CPTR baseline and fourier complexity influences real-time usage
6	ViBERT (Ours)	Combines ViT and BERT for robust visual and textual comprehension and it highly well-balanced BLEU scores, decent ROUGE-L (0.58), and SPICE (0.49) and it have strong structural and semantic performance	Resource-hungry, long training time, overfits on small data and it needs high-end hardware.

Chapter 3

METHODOLOGY

This research proposes a novel image captioning framework named ViBERT, in which we uses ViT as an encoder and in decoder part we uses BERT. The core aim is to address the limitations of conventional CNN-RNN-based models, particularly in their incapacity to proficiently seize global visual context and maintain coherent, semantically rich language during caption generation. This methodology is shaped by a comprehensive review and experimental validation presented across the referenced literature.

3.1 An Overview of Image Captioning Model

In Fig. 3.1 we show the architecture of an image captioning model with a CNN and a transformer-based decoder for tasks such as image captioning. The CNN processes the input to derive picture features, which are input into linear layers and integrated with positional and output embeddings. The decoder use a masked attention and feedforward layers with Add and Norm blocks to produce output probabilities via a softmax layer. CNNs are constructed using convolutional layers designed to extract local spatial features hierarchically, helping in capturing the patterns such as edges, shapes, and texture of the images.

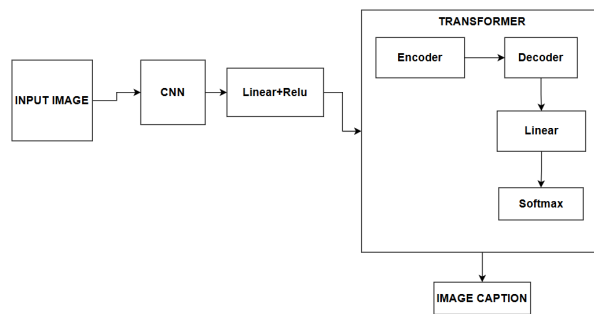


Figure 3.1: This model have of two main components: a transformer as the decoder and a CNN as the encoder.

CNNs make use of kernels to examine particular areas within an image and allow them to run with immense computing power. Their approach makes them very efficient where there is a need for smaller receptive fields. Faster R-CNN enhances the regular CNN model with a region proposal network that efficiently picks out region-of-interest from an image [9]. It allow Faster R-CNN to excel in object detection and segmentation by enabling shared features between the region proposal network and downstream classification layers.

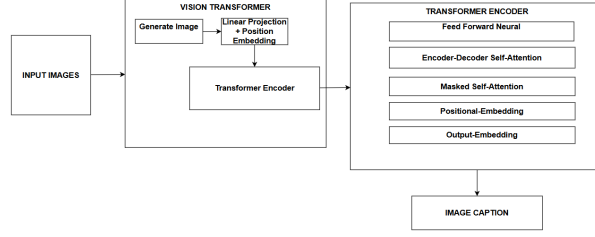


Figure 3.2: This model have of two main components: a ViT as the encoder and a transformer as the decoder.

It helps improve both computational performance and accuracy. In figure 3.2, inspired by ViTs, the picture is segmented into minor patches. These patches are further flattened and transformed into input tokens with positional embeddings. The tokens are refined by an encoder using self-attention and feedforward layers, while the decoder generates text outputs through masked self-attention, cross-attention with image embeddings, and feedforward networks. Contrarily to this, ViTs utilize self-attention mechanisms whereby images are decomposed into sequentially embedded pieces and processed in parallel. This architecture captures global dependencies by allowing every image patch to see all other patches and therefore resulting in a broader contextual understanding. Computational complexity of ViTs results from the quadratic scaling of their attention mechanism with the image resolution since as we know it's a major issue. These developments have attempted to counter this by using hybrid methods that combine convolutional feature learning and transformer blocks in order to get the advantages of both methods.

3.2 Architecture of Proposed Model

The ViT encoder processes input images by segmenting them into uniform patches, such as 16×16 pixels, and then flattening them, and feeding the resulting sequences into a stack of self-attention layers. Unlike CNN architectures such as ResNet and InceptionNet, which operate hierarchically on localized features, ViT attends to all image patches globally from the very beginning. This allows it to develop a better spatial feeling for long-range relations and world context—a critical necessity for good image captioning. ResNet, due to its residual connections, and InceptionNet, due to its multi-scale filter banks, are good at feature extraction at different scales but not good at doing something where one needs to pick up on the semantic overall structure of an image. ViT surpasses these constraints by enabling us to have cross-patch attention, thus complicated images, especially those with multiple interactive parts, can be understood.

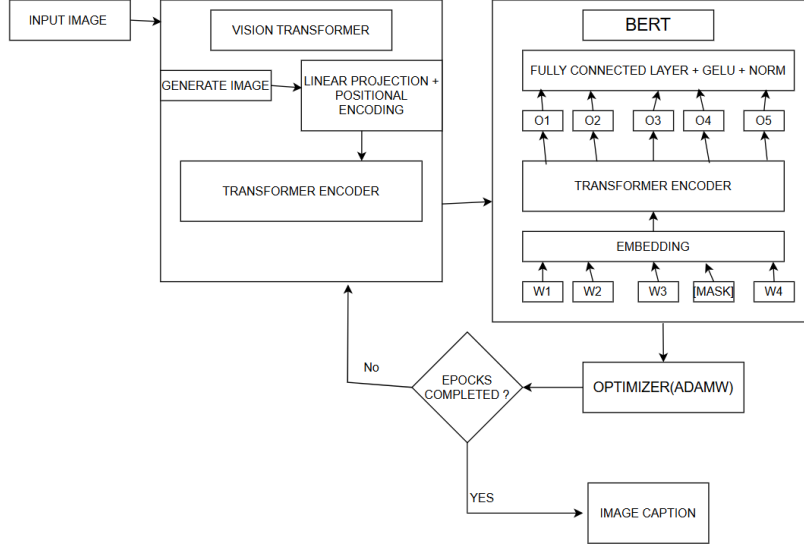


Figure 3.3: Proposed image captioning model architecture combines ViT for extracting image features with BERT for generating context-aware captions within an encoder-decoder setup, optimized using the AdamW algorithm.

In the Fig. 3.3 depicts an image captioning operation utilizing the blend of ViT and BERT in an encoder decoder model. First, ViT transforms the input image, which segment the image into patches. The patches undergo positional encoding and linear transformation before the transformer encoder is processed to obtain information about the image. These features are then forwarded to the BERT module, which functions as the decoder. BERT receives a sequence of word tokens (W1, W2, W3, [MASK], W4), applies embedding, and integrates the visual context to generate outputs (O1 to O5). These outputs are passed through fully connected layers that use GELU activation and normalization to refine predictions. An AdamW optimizer is used to optimize loss and scale the model parameters. A decision block verifies if training has reached the optimal number of epochs. If not, the loop runs until completion after which the model generates an image caption.

In the decoder, BERT is used for language modeling. It was designed to be used in question answering and text analysis. BERT strength is its bidirectional contextual encoding of the text using self-attention. In this model, BERT is used as a decoder, where it takes both visual context embeddings of ViT and partially generated text tokens as input. BERT is repurposed as a decoder by modifying its architecture to accept both visual embeddings from ViT and partially generated text tokens. BERT’s self-attention layers evaluate the relationship between all words in a sequence at once, allowing it to generate fluent, grammatically accurate, and contextually aligned captions. In addition, Tasks are run on pretrained BERT over large corpora such as Masked Language Modeling, resulting in it having a rich syntax, semantics, and contextual word usage. This allows the model to generate semantically and syntactically accurate captions during pretraining. For other intents of achieving additional integration between textual and visual modalities, visual features extracted from ViT are merged with word token embeddings before their input to BERT. The integration enables BERT to realize how visual context affects word choice and sentence composition. Second, the BERT’s attention weights assist in giving greater weight to crucial aspects of the image as well as its partially captioned image, and this results in better sentence construction. The union of BERT and ViT under the ViBERT

framework is fueled to a great extent by the results in Paper 1, which emphasized how the ViTs perform better than CNN-based feature extractors traditionally employed in learning global visual semantics.

The review also highlighted the advantage of transformer-based decoders to produce fluent and detailed captions with minimal use of sequential memory. It also presented the evidence that ViTs are able to excel in common benchmarks like MS COCO with good scores on measures like BLEU, METEOR, CIDEr, and SPICE since they can utilize a self attention mechanism instead of convolutional filter to process image input. This piece also recognized a couple of inherent limitations of traditional models, including the way they typically overlook cross-object relations and context dependencies, which are both directly resolved in the ViBERT design process.

3.3 Baseline Model

The DAIT model efficiently integrates textual and visual contextual information to improve image captioning techniques [15]. DAIT’s two new components: the Adaptive Interactive Encoder and the Adaptive Interactive Decoder. Encoder uses an Interactive Fusion Module to discover correlations between image features (obtained through Swin-T) and text descriptions (from BERT), while a Correlation Aware Module minimizes noise across modalities. The decoder learns to produce based on multimodal features with an Adaptive Guidance Module that considers global context from CLIP embeddings. Together, these break throughs allow the model to more closely match visual content with descriptive text. Large-scale experiments on Flickr30K and MS COCO demonstrate that DAIT surpasses different models in important metrics [15].

Introduction of a Fourier transform layer subsequent to the encoder’s attention mechanism gives us the vanilla transformer model [16]. This modification enables the model to look at the image in the frequency domain, and therefore it’s enhanced to identify borders, textures, and recurring structures. Transferring pixel information to frequency information, the model becomes more aware of spatial relationships in images. The encoder-decoder setup remains intact, with residual connections and layer normalization ensuring stable training. Evaluation on the Flickr8k dataset showed improvement through out the the CPTR model, especially in ROUGE and BLEU-1 scores. Despite having the extra Fourier layer, the model maintained comparable training time. This design demonstrates help us to understand how deep learning will mix with signal processing in a new way to enhance picture captioning outcomes [16].

RBBA technique is a strong architecture engineered for image captioning tasks. It combines ResNet50 for visual feature extraction and BERT for text data encoding, forming a strong multimodal representation. For better alignment between visual features and produced text, the model uses Bahdanau attention, which enables it to attend to appropriate regions of the image when producing each word [12]. The decoder consists of LSTM layers, which well capture the sequential nature of the caption. The RBBA model, which was trained with the Flickr8K dataset. It recorded a BLEU-1 measure of 0.532 alongside a BLEU-4 measure of 0.126, compared to several baselines. These results demonstrate the approach’s ability to generate detailed and precise image descriptions [12].

The GCS-M3VLT model is a novel approach for retina captioning of images, which efficiently merges text as well as visual information to produce accurate medical descriptions [13]. It uses a Guided Context Self-Attention mechanism to highlight important spatial and channel features in retinal images. The model combines diagnostic words via a lan-

guage encoder and combines them with visual data via a Vision-Language TransFusion Encoder. By combining them in this way, the model is capable of understanding more complex clinical environments better. A Transformer-based decoder then generates coherent and accurate captions [13]. GCS-M3VLT, being trained on the DeepEyeNet dataset, surpasses current models in BLEU, CIDEr, and ROUGE scores. Accuracy and efficiency are prioritized in its design even with the use of limited annotated data.

The EntroCap model presents a new way of doing zero-shot image captioning through the integration of GPT-2 and CLIP with custom-designed modules. The hierarchical projector is employed to extract broad background information using CLIP embedded data and allow it to comprehend basic and a high-level aspects. Regarding inference, an Entropy-based Retrieval Strategy assists in retrieving not only salient objects but also small, informative targets that are easily overlooked by other models [14]. These local and global signals are equilibrated by a Balancing Gate. This encourages the linguistic model to produce captions which are more precise and comprehensive. In contrast to supervised methods, no training on image-text pairs is required for EntroCap. The model is shown to work really well on several datasets such as Microsoft Common Objects in Context (MSCOCO), Flickr30k, NoCaps. Therefore, EntroCap is a better and more robust real-world captioning model [14].

3.4 Dataset Description

3.4.1 Flickr 8k Dataset

Flickr8K is a smaller but well-curated dataset that is often employed for initial exploration and model inspection in image captioning research. The dataset contains 8,000 images, each of which contains five alternative captions created by human annotators. The images were crawled from Flickr and typically depict easier scenes, often involving people, animals, and common activities in open or event-driven scenes. Compared to MS COCO with complex, dense scenes, Flickr8K has clear, clean scenes and is thus a better model test bed for models with clean, less complex visual data. Although smaller in scale, the dataset is useful for fine-tuning larger models or training light models and enables standard evaluation using widely available captioning metrics. For researchers, the simplicity of the dataset enables easy baseline performance testing before being scaled up to more complicated datasets such as MS COCO.

3.4.2 MSCOCO dataset

MSCOCO is one of the most popular datasets used in image captioning because it is so diverse and complex. It contains over 328,000 images with five captions for which a human has written. It have total of over 1.5 million captions. The images normally have several objects interacting in a natural setting with an average of 7.7 object instances per image across 91 groups of items, such as people, animals, cars, and common items. This makes MS COCO well adapted to training models that must learn not just individual objects but how they relate to each other and are arranged in space. Because of its richness in visual and linguistic diversity, MS COCO is a robust benchmark for assessing the contextual and descriptive strength of image captioning models.

3.5 Parameter Settings

Training is performed over Flickr8K dataset by resizing the images of size 224X224 and caption tokens with BERT tokenizer. Training is performed using PyTorch with batch size 32, learning rate of $6e-5$, and AdamW (Adam with Weight Decay) optimizer with a number of epochs as 40. To resolve the issue of incorrect weight decay in Adam, a modification of the base Adam optimizer has been created, known as the AdamW optimizer. Unlike its original version that induces bad generalization due to its imposition of L2 regularization, AdamW decouples weight decay from gradient update. This allows easy regulation of model complexity and avoids overfitting, especially in large transformer models such as ViT and BERT. AdamW has the benefit of adaptive learning rates and convergence along with good regularization and hence is best suited for fine-tuning pre-trained transformer models. Pre-trained weights for both BERT and ViT are fine-tuned during training to fit them into image captioning. Training is conducted on an NVIDIA Tesla P100 GPU, and regularization methods including dropout and early stopping are applied to facilitate generalization and prevent overfitting.

3.6 Performance Metrics

Performance metrics are crucial for testing how well human-provided captions match those created by machines. Linguistic and semantic properties are included in these metrics to give an overall insight into model performance [17]. BLEU score is the most widely used, estimating the precision of n grams by measuring the coincident between the captions generated by the model and the references. BLEU-1 to BLEU-4 score unigrams up to 4-grams, thus capturing both single word accuracy and short phrase consistency. However, BLEU’s sensitivity to surface matching makes it weaker when detecting semantic equivalence, especially when there are paraphrasing or synonyms involved [18]. In opposition to this lack, the METEOR score is used, which uses stemming, synonym mapping, and word order mapping to produce improved correlation with human preference. ROUGE-L also assists in complementing the scoring by highlighting recall using longest common subsequence analysis, an indication of the proportion of key content preserved. The CIDEr measure provides a collective-oriented view by taking TF-IDF weighting of n -gram overlap, encouraging captions to closely match the unified human understanding of an image and penalizing over-generic outputs [19]. Lastly, SPICE measures captions against scene graph structures of objects, attributes, and relationships and thus estimates a human-level semantic content understanding [20]. Together, these measures constitute a strong evaluation framework that supports a multidimensional assessment of caption quality. Their use for both papers guarantees uniform benchmarking and allows for an unbiased comparison of the proposed ViBERT model and standard CNN-RNN-based systems for captioning.

Chapter 4

RESULTS and DISCUSSION

Flickr8K dataset was utilized for testing the recently presented ViBERT model. In which merged BERT into a decoder for natural language synthesis with ViT for visual feature extraction. The model’s performance was benchmarked using widely accepted metrics in image captioning. These measures provided a comprehensive evaluation across syntactic precision, fluency, semantic alignment, and contextual richness.

4.1 Result of Review of ViT Based Image captioning Technique

4.1.1 Performance Comparison of ViT-based and Conventional Image Captioning Technique

In this thesis, we are comparing how well image captioning models work with and without ViTs on the MSCOCO datasets. We use different metrics like SPICE, METEOR, ROUGE-L, CIDEr, and BLEU to judge their performance. BLEU evaluates machine-generated text in image captioning by comparing n-grams with human-written text. BLEU-1 places extremely high emphasis on unigram accuracy, whereas BLEU-4, better recognized by its common usage, keeps word and phrase matching in equilibrium for better evaluation[17]. METEOR is a text score measure widely employed in image captioning, with greater accuracy kept by preserving word order and synonyms. It generates a longer evaluation, therefore appropriate for operations with more language comprehension [21]. Some of these include the vintage critical performance measure we’re used to, i.e., ROUGE, used for image captioning text quality estimation. It is recall-based as it is a metric for how much the generated text mirrors common facts in human sources, employing n-grams and longest common subsequences[18]. SPICE measures caption meaning by segmenting them into propositions and matching to human-written content. Paying attention to attributes, objects, and relations, it provides a human-oriented measurement that values meaning over word-to-word similarity[20]. CIDEr measures captions with n-gram overlap with human captions, applying term frequency-inverse document frequency weighting in order to highlight significant words and suppress frequent words. It effectively captures alignment with image context, supporting tasks with diverse valid captions[19]. ViTs are’ multi-head attention methods and capacity to accumulate complex relationships between picture features have shown an important effect on model performance. On the MS COCO dataset, ViTs-based models demonstrate a clear advantage. DAIT achieves 83.3 on BLEU-1, 41.6 on BLEU-4, 31.2 on METEOR, and 143.6 on CIDEr,

which is the best in this category. This improvement with transformers is because they can listen to multiple components of the image simultaneously and process challenging image features. In Table 4.1.1 self-supervised modal optimization transformer (SMOT) also achieves good performance with 81.4 in BLEU-1, 39.9 in BLEU-4, and 136.2 in CIDEr. The performance shows how the ViT model achieves better in generating descriptive, relevant captions that are appropriate to the content of the images, revealing hidden descriptions that are usually hard for traditional methods to compete with.

Table 4.1: ViT-based image captioning techniques On MSCOCO Dataset

MODEL	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	CIDEr	ROUGE-L	SPICE
DAIT [15]	83.3	-	-	41.6	31.2	143.6	60.9	24.9
EHAT [22]	81.9	-	-	40.1	29.6	133.5	59.4	-
ETFT [16]	0.33	-	-	-	0.33	-	0.15	-
DCCT [23]	83.2	-	-	42.7	30.6	141.7	60.8	24.6
SMOT [24]	81.4	-	-	39.9	29.9	136.2	59.5	23.8

Table 4.2: Conventional image captioning techniques On MSCOCO Dataset

MODEL	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	CIDEr	ROUGE-L	SPICE
AST [25]	51.77	-	-	14.05	23.98	34.93	30.03	17.64
CATNet [26]	82	66.9	52.3	40.2	29.5	130.4	59.5	-
TDAN [27]	84.6	64.5	52.4	36.2	39.3	133	62.3	-
LLAFN-Generator [28]	73	56.7	43.5	33.7	26.5	105.9	54.5	19.3
RCT [29]	81.3	66.7	52	40	29.5	129.7	59.3	-
SAMT-generator [30]	73.8	57.2	44.1	34.3	27.3	108.7	55.2	20
SG-DLCT [31]	81.8	-	-	40	29.5	134.5	59.2	-
TLGG [32]	86.1	66.5	49.1	37.8	39.2	132.9	65.1	-
TCCTN [33]	81.3	-	-	39.4	29.2	132.8	58.9	-
DVAT [34]	81.2	-	-	39.4	29.3	133.1	59.1	23.9
ETransCap [35]	79.5	67	54.6	38.6	28.3	68.2	54.2	-
PSNet [36]	82	67.1	52.6	40.4	29.8	132.9	59.7	-
XGL-T [37]	81.5	67.1	51.6	39.9	29.8	134	59.9	23.8

4.1.2 Limitation of Conventional and ViTs-based Technique

Firstly, the techniques that we analyzed are primarily evaluated on the MS COCO dataset. Although widely utilized, reliance on a single dataset restricts the generalization of the results. By applying these techniques to more diverse datasets could provide deeper insight into model performance measures across various context. Second problem that arose is insufficient semantics in both text and image present a major challenge, impacting the model ability to accurately represent complex relationships, which often results in incorrect captions. Lastly, the absence of multilingual datasets poses a language barrier. Variations across language families and insufficient resources for many languages hinder the development of effective multilingual image captioning systems.

4.2 Result of Proposed ViBERT Model

4.2.1 Performance Evaluation of ViBERT Compared to Other Leading Captioning Models

In Table 4.3, the results indicate that ViBERT performs better than traditional models, such as ResNet50-BERT with Bahdanau attention and GCS-M3VLT. The ViBERT model shows a good results, with a BLEU-4 score of 0.45, showing high phrase-level correctness in the produced captions. Comparison of performance shows the strength and weakness of certain picture captioning models according to ROUGE-L, BLEU, CIDEr, and SPICE scores. Among them, the one proposed in this paper, ViBERT, shows consistent high performance in every level of BLEU with 0.49 for BLEU-1, 0.48 for BLEU-2, 0.47 for BLEU-3, and 0.45 for BLEU-4. This steady progression indicates ViBERT’s capability to not only match single words but also generate coherent multi-word phrases, which is crucial for natural-sounding captions. In addition, ViBERT achieves the highest ROUGE-L score (0.58) among all models, reflecting superior fluency and structural alignment with reference captions. It can generate subtitles that is meaningful as well as important and closely connected with human-generated descriptions, demonstrated by its CIDEr score of 1.71 and SPICE score of 0.49. The ResNet50-BERT model with Bahdanau attention, on the other hand, only obtains BLEU-1 scores of 0.52 in BLEU-2, 0.17 in BLEU-3, and 0.12 in BLEU-4, which are marginally higher than ViBERT in comparison. The reduction indicates the model has difficulty with phrase-level coherence yet it can identify individual words. Its ROUGE-L score of 0.49 reflects reasonably good fluency but short of that of ViBERT. The GCS-M3VLT model scores somewhat good on BLEU-2 and BLEU-3 at 0.34 and 0.31 respectively but is limited by a comparatively low BLEU-4 (0.23) and a low CIDEr score of 0.55, reflecting lower relevance and specificity in its responses. The Encoder-decoder with Fourier Transform model provides low BLEU-1 score (0.33333) and no other metric scores, which means a quite simple or trial-and-error type architecture that may not have enough depth for good captioning performance.

Entrocap, with semantic attention, has low CIDEr improvement (41.5) and SPICE improvement (11.7), but low ROUGE-L (0.15) and very low BLEU-4 (18.3) indicate that it is producing short or truncated captions of poor structural quality. The DAIT (Dual-Adaptive Interactive Transformer) model, however, performs well in most of the metrics, including BLEU-1 of 75.7, CIDEr of 80.1, SPICE of 19.5, and ROUGE-L of 54.3. However, the absence of higher-order BLEU scores (BLEU-2 to BLEU-4) makes it difficult to fully assess its phrase-level precision and fluency. While DAIT performs well in terms of relevance and semantic richness, ViBERT’s balanced performance across all levels of evaluation—particularly in BLEU-4 and ROUGE-L—demonstrates its ability to produce more complete, contextually accurate, and linguistically fluent captions. In conclusion, ViBERT not only competes with but in some areas surpasses existing models by delivering consistent, high-quality captions that closely resemble human descriptions.

Overall, these findings point to the fact that ViBERT model is on par with baseline and even behemoth models in terms of reasoning through complex visual relations and producing naturalistic, coherent captions.

In Fig. 4.1 the reduction in loss is graphed versus 40 training epochs. The loss begins very high at 6234.02, yet decreases very quickly with further training. The loss drops below 2500 by epoch 10, representing effective learning. It keeps falling steadily to 283.08 at epoch 39.

Table 4.3: Comparison of different image captioning models on various performance metrics

Model Name	BLEU-1	BLEU-2	BLEU-3	BLEU-4	CIDEr	SPICE	ROUGE-L
ResNet50-BERT-Bahdanau [12]	0.53	0.22	0.17	0.12	-	-	-
GCS-M3VLT [13]	0.43	0.34	0.31	0.23	0.55	-	0.49
Encoder-decoder framework [16]	0.33333	-	-	-	-	-	0.15
Entrocap [14]	-	-	-	18.3	41.5	11.7	-
DAIT [15]	75.7	-	-	36.4	80.1	19.5	54.3
ViBERT (Ours)	0.49	0.48	0.47	0.45	1.71	0.49	0.58

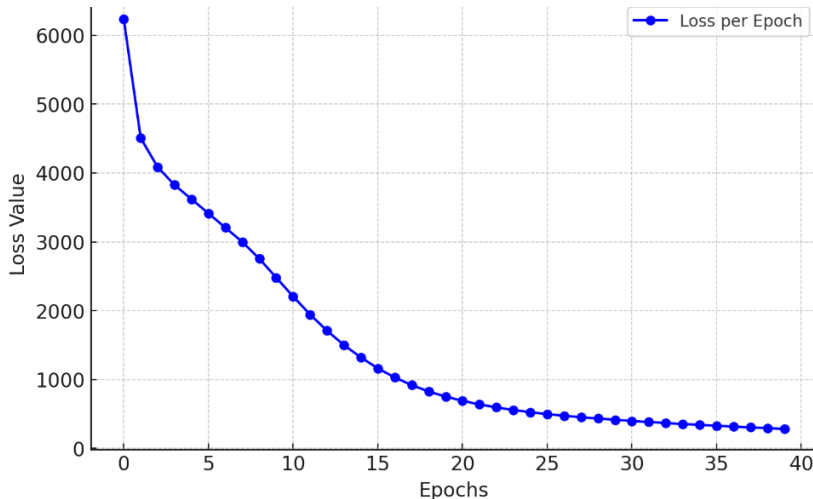


Figure 4.1: It shows the training loss for each epoch

4.3 Overall Findings

In comparison with base CNN-RNN models such as ResNet50-LSTM or InceptionNet-GRU, ViBERT outperforms in all the metrics utilized. While CNN-based models are adept at mimicking local visual features, it does not capture the spatial relationships between multiple objects in an image and global context as well.

In comparison with base CNN-RNN models such as ResNet50-LSTM or InceptionNet-GRU, ViBERT outperforms in all the metrics utilized. While CNN-based models are adept at mimicking local visual features, it does not capture the spatial relationships between multiple objects in an image and global context as well. Figures 4.2 and 4.3 display the captions that the ViBERT model produced for the different images. The model graph indicates that it is convergent and improving with every iteration, as is established by the declining trend. ViT’s attention mechanism, nevertheless, allows each token in an image to see all the rest, improve the model’s comprehension of the whole image. Using BERT it is pretrained on large corpora and capable of bidirectional context modeling, enables more fluent and semantically consistent caption generation. The collaboration between ViT and BERT contributes for ViBERT’s outstanding efficiency. ViT provides a powerful mechanism to extract dense, globally-aware image representations, while BERT excels in transforming these representations into fluent, contextually rich language. Unlike traditional models that may generate generic or repetitive captions, ViBERT can produce descriptions that closely match human references while still being diverse.



Generated Caption: a man and a young girl float on an innertube in the water an inner



Generated Caption: a woman stands on an urban sidewalk holding a beige handbag a waits for a train

Figure 4.2: ViBert model-generated caption



Generated Caption: three children do jumping jacks in a public place while a man sits and looks on



Generated Caption: a man is standing by rocks and water and is holding a stick

Figure 4.3: ViBert model-generated caption

Chapter 5

CONCLUSION AND FUTURE SCOPE

In this thesis provides an in-depth summary of image captioning to allow the reader to get a general idea regarding the research work being done on this subject. We discuss various image captioning methods on various evaluation metrics on widely used benchmark datasets. ViTs become the strong rival to standard CNNs because of their ability to capture global relationship between image regions. This enables the ability to know what the overall visual content is about in a deeper way, which is necessary for making more descriptive and meaningful captions. ViTs utilizes attention mechanism that aid them in identifying the global relationship within the various patches of the image and improving the quality of generated texts. They handle images with complex structures or more than one object more effectively because they process the whole image efficiently, and this gives rise to more comprehensive and coherent captions. Lastly, this review will enable the practitioner to select the appropriate technique for the image captioning technique. The research gap or image captioning based on ViTs is highlighted in this review. We introduce ViBERT, a captioning transformer model with the strength of BERT as a natural language generation process and ViT as an observable feature extractor. The shortcomings of conventional CNN-RNN architectures, which frequently fail to capture the distant dependencies and global implications necessary for producing precise and significant image captions, served as the catalyst for our work.

ViBERT overcomes these issues by using ViT’s self-attention mechanism to capture intricate spatial relationships and BERT’s bidirectional language model to generate fluent, contextually accurate captions. Flickr8K dataset was used for model training and testing according to performance measures. These findings demonstrate the effectiveness of ViBERT can capture both the grammatical and semantically content of image description, outperforming baseline models with robust results across all measures. Utilizing the AdamW optimizer also played a major role in stable training and also improved it generalization. Finally, the increasing significance of transformer-based multimodal systems in picture captioning is confirmed by this study. ViBERT is an important breakthrough, providing a scalable and flexible solution for practical uses. It sets the stage for further advances in terms of multilingual support, computational complexity, and more profound semantic correspondence between vision and language.

5.1 Future Work

To show generalization across a wide range of visual environments, the ViBERT model will be run on large databases such as MS COCO and Flickr30K in future studies. Multilingual translation can be integrated using the likes of mBERT to support cross-lingual captioning. Lighter or hybrid models can be studied for use in low-resource scenarios for better efficiency. Region-based attention or object detection can be integrated into the model to make the captions more specific. Moreover, cross-modal pretraining will have better performance on less supervised data. Last but not least, human judgment with automatic measurement will play a crucial role in measuring caption quality and usability in real use.

Appendix A

LIST OF PUBLICATION

1. Priya Singh & Aman Rawat (2025). Exploring Vision Transformers for Enhanced Image Captioning: A Review (ICCCNT 2025).[**Scopus Indexed**][**Accepted**]
2. Priya Singh & Aman Rawat (2025). Enhancing Image Captioning with Vision Transformers and BERT: A Performance-Driven Study (ICCCNT 2025).[**Scopus Indexed**][**Accepted**]



THE 16th INTERNATIONAL IEEE CONFERENCE ON COMPUTING, COMMUNICATION AND NETWORKING TECHNOLOGIES (ICCCNT)

July 6 - 11, 2025, IIT - Indore, Madhya Pradesh, India.

A 6 Day Hybrid Technical Fest - participate online / offline

[Home](#) [General Info](#) [About](#) [Authors](#) [Program](#)

Welcome to
THE 16th INTERNATIONAL IEEE CONFERENCE
ON COMPUTING, COMMUNICATION AND
NETWORKING TECHNOLOGIES (ICCCNT)

July 6 - 11, 2025, IIT - Indore, Madhya Pradesh, India.



Figure A.1: Screenshot

CFP

16th ICCCNT 2025: 16th International IEEE Conference on Computing Communication and Networking Technologies

IIT-Indore

Indore, India, July 6-11, 2025

Conference website	https://16iccnt.com/
Submission link	https://easychair.org/conferences/?conf=16thiccnt2025
Submission deadline	April 15, 2025

Topics: [computing computer vision iot/biometrics](#) [communication 5g wireless](#) [networks robotics cyber security](#) [data science photonics electronics packaging](#)

The 16th International Conference on Computing, Communication, and Networking Technologies (ICCCNT) is a premier conference that is being organized from July 6-11, 2025 at IIT - Indore, Madhya Pradesh, India. Computing, communication, and networking are the most compelling areas of research because of its rich applications. So far ICCCNT has completed 15 versions of conferences in different locations across the globe. It continuously receives research papers from different countries and at different levels of colleges/universities. All the conference papers were successfully published in IEEE Digital Library Xplore® and indexed in Scopus. It is a prestigious event organized every year with the motivation to provide an excellent platform for the leading academicians, researchers, industrial participants, and budding students to share their research findings with renowned experts.

Submission Guidelines

All papers must be original and not simultaneously submitted to another journal or conference. The following paper categories are welcome:

paper must be 6 page limit, if it is beyond 6 page (excluding references) you have to pay 500 INR more

visit www.16iccnt.com for more details. Submission opens on Feb 24, 2025

List of Topics, but not limited to (see link for more details and additional topics)

[THE 16th INTERNATIONAL IEEE CONFERENCE ON COMPUTING, COMMUNICATION AND NETWORKING TECHNOLOGIES \(ICCCNT\) - IIT-Indore, Madhya Pradesh \(India\) \(16iccnt.com\)](https://www.16iccnt.com/)

Figure A.2: Scopus Index Screenshot

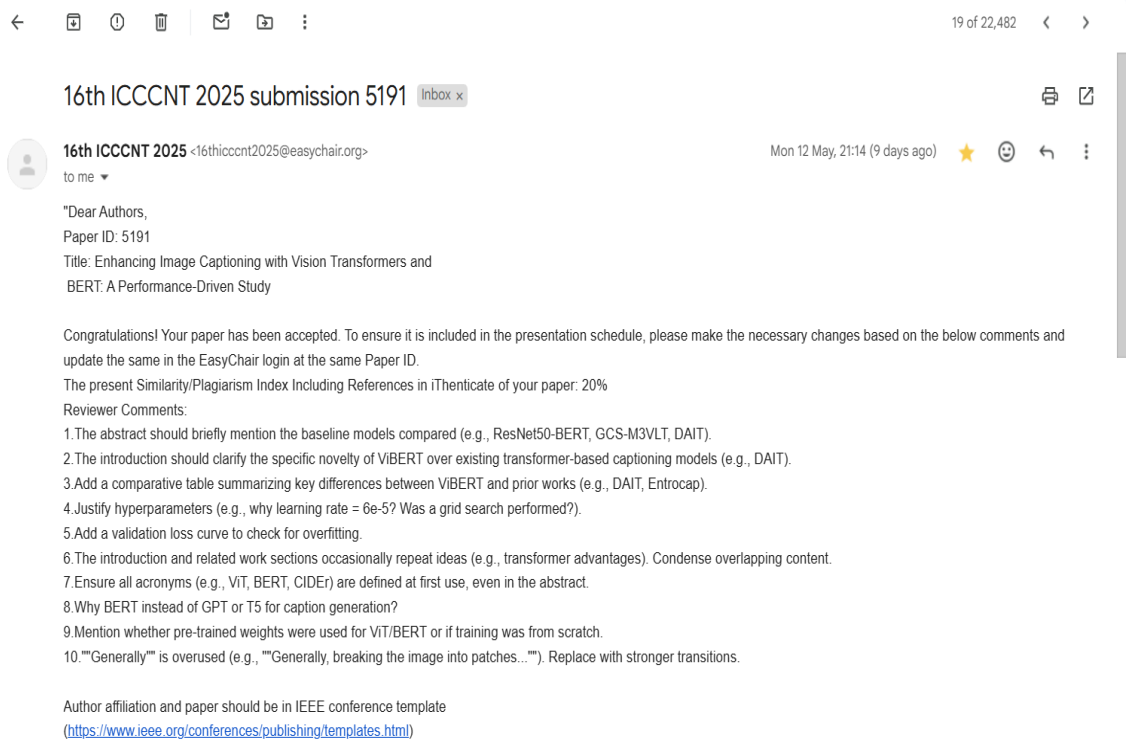


Figure A.3: Paper 1 acceptance Screenshot

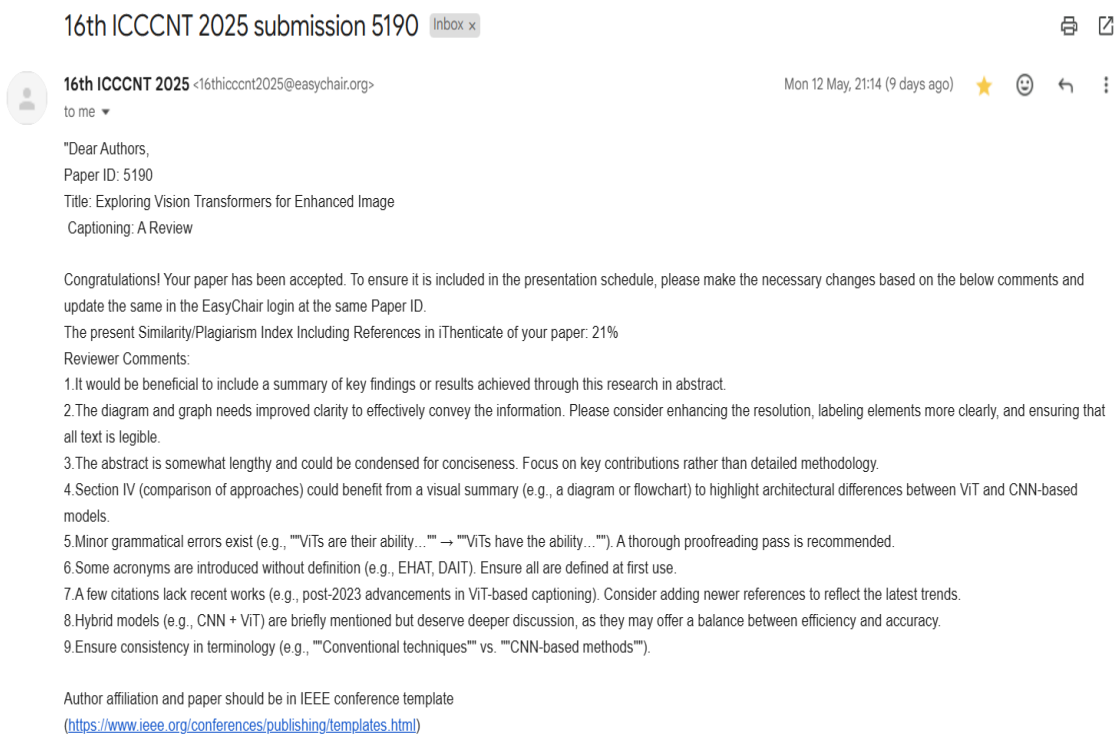


Figure A.4: Paper 2 acceptance Screenshot

Bibliography

- [1] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, “An image is worth 16x16 words: Transformers for image recognition at scale,” *arXiv preprint arXiv:2010.11929*, 2020.
- [2] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, “Show and tell: A neural image caption generator,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3156–3164.
- [3] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio, “Show, attend and tell: Neural image caption generation with visual attention,” in *International conference on machine learning*. PMLR, 2015, pp. 2048–2057.
- [4] V. Ashish, “Attention is all you need,” *Advances in neural information processing systems*, vol. 30, p. I, 2017.
- [5] M. Zhu, Y. Tang, and K. Han, “Vision transformer pruning,” *arXiv preprint arXiv:2104.08500*, 2021.
- [6] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.
- [7] —, “Deep residual learning for image recognition,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.
- [8] O. Ondeng, H. Ouma, and P. Akuon, “A review of transformer-based approaches for image captioning,” *Applied Sciences*, vol. 13, no. 19, p. 11103, 2023.
- [9] H. Tsaniya, C. Fatichah, and N. Suciati, “Transformer approaches in image captioning: A literature review,” in *2022 14th International Conference on Information Technology and Electrical Engineering (ICITEE)*. IEEE, 2022, pp. 1–6.
- [10] P. Singh, F. Raja, and H. Sharma, “Generating image captions in hindi based on encoder-decoder based deep learning techniques,” in *Reliability Engineering for Industrial Processes: An Analytics Perspective*. Springer, 2024, pp. 81–94.
- [11] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding,” in *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, 2019, pp. 4171–4186.

- [12] D.-H. Hoang, A.-K. Tran, D. N. M. Dang, P.-N. Tran, H. Dang-Ngoc, and C. T. Nguyen, “Rbba: Resnet-bert-bahdanau attention for image caption generator,” in *2023 14th International Conference on Information and Communication Technology Convergence (ICTC)*. IEEE, 2023, pp. 430–435.
- [13] T. K. Cherukuri, N. S. Shaik, J. D. Bodapati, and D. H. Ye, “Gcs-m3vlt: Guided context self-attention based multi-modal medical vision language transformer for retinal image captioning,” in *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2025, pp. 1–5.
- [14] J. Yan, Y. Xie, S. Zou, Y. Wei, and X. Luan, “Entrocap: Zero-shot image captioning with entropy-based retrieval,” *Neurocomputing*, vol. 611, p. 128666, 2025.
- [15] L. Chen and K. Li, “Dual-adaptive interactive transformer with textual and visual context for image captioning,” *Expert Systems with Applications*, vol. 243, p. 122955, 2024.
- [16] D. M. Joshy, A. Das, D. T. Sunil, S. Safar *et al.*, “Enriching transformer using fourier transform for image captioning,” in *2023 3rd International Conference on Intelligent Technologies (CONIT)*. IEEE, 2023, pp. 1–6.
- [17] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, “Bleu: a method for automatic evaluation of machine translation,” in *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, 2002, pp. 311–318.
- [18] C.-Y. Lin, “Rouge: A package for automatic evaluation of summaries,” in *Text summarization branches out*, 2004, pp. 74–81.
- [19] R. Vedantam, C. Lawrence Zitnick, and D. Parikh, “Cider: Consensus-based image description evaluation,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 4566–4575.
- [20] P. Anderson, B. Fernando, M. Johnson, and S. Gould, “Spice: Semantic propositional image caption evaluation,” in *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part V 14*. Springer, 2016, pp. 382–398.
- [21] S. Banerjee and A. Lavie, “Meteor: An automatic metric for mt evaluation with improved correlation with human judgments,” in *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, 2005, pp. 65–72.
- [22] Z. Song, Z. Hu, Y. Zhou, Y. Zhao, R. Hong, and M. Wang, “Embedded heterogeneous attention transformer for cross-lingual image captioning,” *IEEE Transactions on Multimedia*, 2024.
- [23] J. Hu, Z. Li, Q. Su, Z. Tang, and H. Ma, “Exploring refined dual visual features cross-combination for image captioning,” *Neural Networks*, vol. 180, p. 106710, 2024.
- [24] Y. Wang, D. Li, Q. Liu, L. Liu, and G. Wang, “Self-supervised modal optimization transformer for image captioning,” *Neural Computing and Applications*, vol. 36, no. 31, pp. 19 863–19 878, 2024.

- [25] K. Zhou, Z. Tang, P. Meng, J. Huang, Y. Xu, and X. Rao, “A novel method for sign judgment of defects based on phase correction in power cable,” *IEEE Transactions on Instrumentation and Measurement*, vol. 73, pp. 1–9, 2023.
- [26] X. Yang, Y. Wang, H. Chen, J. Li, and T. Huang, “Context-aware transformer for image captioning,” *Neurocomputing*, vol. 549, p. 126440, 2023.
- [27] H. Parvin, A. R. Naghsh-Nilchi, and H. M. Mohammadi, “Image captioning using transformer-based double attention network,” *Engineering Applications of Artificial Intelligence*, vol. 125, p. 106545, 2023.
- [28] X. Yang, X. Tian, J. Wu, X. Yang, S. Ma, X. Qi, and Z. Hou, “Llafn-generator: Learnable linear-attention with fast-normalization for large-scale image captioning,” *Computer Vision and Image Understanding*, vol. 248, p. 104088, 2024.
- [29] L. Chen, Y. Yang, J. Hu, L. Pan, and H. Zhai, “Relational-convergent transformer for image captioning,” *Displays*, vol. 77, p. 102377, 2023.
- [30] X. Yang, Y. Yang, S. Ma, Z. Li, W. Dong, and M. Woźniak, “Samt-generator: A second-attention for image captioning based on multi-stage transformer network,” *Neurocomputing*, vol. 593, p. 127823, 2024.
- [31] X. Zhu, J. Deng, K. Zhang, and J. Zhang, “Sg-dlct: Saliency-guided dual-level collaborative transformer for image captioning,” in *2024 IEEE 13th Data Driven Control and Learning Systems Conference (DDCLS)*. IEEE, 2024, pp. 233–238.
- [32] H. Parvin, A. R. Naghsh-Nilchi, and H. M. Mohammadi, “Transformer-based local-global guidance for image captioning,” *Expert Systems with Applications*, vol. 223, p. 119774, 2023.
- [33] J. Li, W. Zhou, K. Wang, and H. Hu, “Triple-stream commonsense circulation transformer network for image captioning,” *Computer Vision and Image Understanding*, vol. 249, p. 104165, 2024.
- [34] Y. Ren, J. Zhang, W. Xu, Y. Lin, B. Fu, and D. N. Thanh, “Dual visual align-cross attention-based image captioning transformer,” *Multimedia Tools and Applications*, pp. 1–20, 2024.
- [35] A. Mundu, S. K. Singh, and S. R. Dubey, “Etranscap: efficient transformer for image captioning,” *Applied Intelligence*, vol. 54, no. 21, pp. 10 748–10 762, 2024.
- [36] L. Xue, A. Zhang, R. Wang, and J. Yang, “Psnet: position-shift alignment network for image caption,” *International Journal of Multimedia Information Retrieval*, vol. 12, no. 2, p. 42, 2023.
- [37] D. Sharma, C. Dhiman, and D. Kumar, “Xgl-t transformer model for intelligent image captioning,” *Multimedia Tools and Applications*, vol. 83, no. 2, pp. 4219–4240, 2024.