

ViBERT: TRANSFORMER-BASED IMAGE CAPTIONING USING VISION-BERT FUSION — ARCHITECTURE, SURVEY AND IMPLEMENTATION

A Thesis Report

Submitted in Partial Fulfillment of the Requirements for the Degree of

**MASTER OF TECHNOLOGY
IN
COMPUTER SCIENCE AND ENGINEERING
BY
AWANTIKA TRIPATHI
24/CSE/29**

Under the Supervision of

Dr. Prashant Giridhar Shambharkar

Associate Professor, Department of Computer Science and Engineering
Delhi Technological University



Department of Computer Science and Engineering

DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi College of Engineering)

Shahbad Daulatpur, Main Bawana Road, Delhi-110042, India

May, 2026



DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi College of Engineering)

Shahbad Daultapur, Main Bawana Road, Delhi-110042

ACKNOWLEDGEMENT

With utmost respect and gratitude, I would like to thank Dr. Prashant Giridhar Shambharkar, Associate Professor, Department of Computer Science and Engineering, Delhi Technological University, for his invaluable guidance and mentorship during the course of this study. The academic understanding of the fields of deep learning, computer vision, and natural language processing provided by him has played a pivotal role in forming the intellectual structure of this thesis in ways which cannot be quantified. From the discussion of the research problem to designing its solution, from suggesting an appropriate experimental architecture to evaluating its performance, his suggestions and advice have played a defining role in shaping this research. His open-mindedness towards innovative thinking and adherence to the highest standards of academic rigor is something which shall guide me through my professional life ahead.

I wish to express my heartfelt thanks to the Department of Computer Science and Engineering of Delhi Technological University for offering me the facilities of computer systems, information sources, and a conducive learning atmosphere during this research work. The facilities available within the department, in terms of computing systems and libraries along with the research culture of the department, were instrumental in successfully completing the research.

I would like to thank all the fellow researchers in the department for their informal interactions and valuable insights related to deep learning architecture, vision-language model training, and evaluation technique that enhanced my understanding far more than reading the existing literature did. They helped me develop a strong argument in this thesis through their critical thinking.

I would like to express my thanks to the wider community of open-source research for providing free access to pre-trained models, datasets, and code repositories. Access to the transformers library by Hugging Face, the PyTorch platform, and the Flickr8K dataset was essential to the execution of the experiments in this thesis.

Last but not least, I would like to extend my deepest thanks to my family for being there all the time supporting me morally and showing patience with me during these long years of study. Their belief in my capabilities made all my efforts worthwhile.

Place: Delhi

AWANTIKA TRIPATHI

Date: May 2026



DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi College of Engineering)

Shahbad Daultapur, Main Bawana Road, Delhi-110042

CANDIDATE'S DECLARATION

I AWANTIKA TRIPATHI hereby certify that the work which is being presented in the thesis entitled ViBERT: TRANSFORMER-BASED IMAGE CAPTIONING USING VISION-BERT FUSION — ARCHITECTURE, SURVEY AND IMPLEMENTATION as a partial fulfillment of the requirements for the award of the Degree of Master of Technology, submitted in the Department of Computer Science and Engineering, Delhi Technological University is a true representation of my own work performed between August 2024 and May 2026 under the guidance of Dr. Prashant Giridhar Shambharkar.

The contents of the thesis have not been submitted by me earlier for the purpose of obtaining any other degree of this or any other Institution

Candidate's Signature

We hereby certify that the candidate has made all the suggested corrections in the thesis and the above statement of the candidate is accurate to the best of our knowledge.

Signature of Supervisor

Signature of External Examiner



DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi College of Engineering)

Shahbad Daulatpur, Main Bawana Road, Delhi-110042

CERTIFICATE

This is to certify that the thesis entitled **ViBERT: TRANSFORMER-BASED IMAGE CAPTIONING USING VISION-BERT FUSION — ARCHITECTURE, SURVEY AND IMPLEMENTATION** submitted by **Awantika Tripathi** in partial fulfillment of the requirements for the award of the degree of Master of Technology in Computer Science and Engineering is a record of the bonafide research work carried out by her at Delhi Technological University 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

Dr. Prashant Giridhar Shambharkar

Date:

Associate Professor

Department of Computer Science and Engineering

ABSTRACT

The automatic creation of descriptions for visual stimuli using natural language, known as image captioning, is among the most interesting areas of overlap between computer vision and natural language processing. Although a lot of research has been done in the area, generating fluent and semantically and contextually appropriate descriptions continues to be a challenging task, especially when dealing with complex scenes involving multiple objects, atypical spatial relationships, and relational reasoning. The present work makes two distinct contributions to the field of image captioning by surveying the architecture of transformer-based image captioning models from the years 2014 to 2025 and proposing a new image captioning model called ViBERT which combines the Vision Transformer and the BERT models.

Contribution of the survey, which comprises the first part of this thesis, is to structure an enormous number of papers on captioning into nine categories, such as CNN-RNN based baseline models, attention-based models, encoder-focused vision transformers, decoder-focused language model-based methods, cross-modal attention methods, contrastive models like CLIP, retrieval based captioning approaches, multimodal transformers, and LLM-based captioning methods. Ten benchmark datasets with specific ones on medical images are discussed along with eight evaluation metrics ranging from simple n-gram similarities at the surface level to semantic graph matching. Nine open problems identified despite advances in benchmarks include caption hallucinations, medical domain adaptation, accessibility, cross-linguistic generalization, and transformer interpretability.

The implementation contribution presents ViBERT, a system utilizing ViT-Base/16 visual encoder with a BERT-base decoder working with a visual prefix injection method. ViBERT was trained on the Flickr8K dataset and tested against two other competitive baselines such as ResNet50+LSTM+Bahdanau Attention and CLIP+GPT-2. ViBERT obtained the highest performance metrics in terms of BLEU-4 score with 0.3228

points, CIDEr score with 0.9710 points, and SPICE score with 0.3164 points. As for computational performance, ViBERT demonstrated processing speed of 19.5 milliseconds per image. In comparison, it is faster than the LSTM-based baseline almost 20 times due to the benefits from the parallel computing capabilities offered by transformers. Thus, the research confirmed that the paradigm in image captioning is changing and moving from convolution-based methods towards self-attention. Furthermore, the combination of Vision Transformer encoders and bidirectional language models proved to be superior for obtaining better semantic grounding. Overall, the current research provided the analysis of existing models' performance and their potential for improvement, making the implementation contribution an integral part of the study.

Contents

Acknowledgement	i
Candidate's Declaration	ii
Certificate	iii
Abstract	v
Contents	viii
List of Tables	ix
List of Figures	x
List of Symbols and Abbreviations	xi
1 INTRODUCTION	1
1.1 Background and Motivation	1
1.2 Problem Statement	3
1.3 Aims and Objectives	3
1.4 Scope and Organisation of the Thesis	4
2 THEORETICAL BACKGROUND	5
2.1 Overview	5
2.2 Convolutional Neural Networks for Visual Feature Extraction	5
2.3 Recurrent Neural Networks and Long Short-Term Memory	6
2.4 The Transformer Architecture	6
2.5 Vision Transformers	7
2.6 BERT and Bidirectional Language Model Training	8
2.7 CLIP and Contrastive Learning	8
2.8 Metrics for Evaluating Image Captioning	9
3 LITERATURE REVIEW AND SURVEY OF TRANSFORMER-BASED IMAGE CAPTIONING ARCHITECTURES	10
3.1 Overview and Taxonomy	10
3.2 Taxonomy and Overview	10
3.3 CNN-RNN Baseline Architectures	12
3.4 Attention-Augmented CNN-RNN Systems	12

3.5	Vision Transformer Encoder Models	13
3.6	Methods Using Pre-Trained Language Models as Decoders	13
3.7	Cross-Modal Attention and Vision-Language Pretraining	14
3.8	CLIP-Aligned Contrastive Learning	15
3.9	Retrieval-Augmented Captioning	15
3.10	Unified Multimodal Transformers	16
3.11	LLM-Integrated Captioning Systems	16
3.12	Datasets Used for Image Captioning	16
3.13	Evaluation Metrics: Analysis and Limitations	18
3.14	Component-Level Ablation Analysis	20
3.15	Open Challenges	21
3.15.1	Hallucination	21
3.15.2	Medical Domain Adaptation	22
3.15.3	Computational Accessibility	22
3.15.4	Cross-Lingual and Cross-Cultural Generalisation	23
3.15.5	Explainability and Trust	23
3.16	Comparative Analysis Among Architectures' Generations	23
4	METHODOLOGY	26
4.1	Overview	26
4.2	Dataset: Flickr8K	26
4.3	Model 1: ResNet50 + LSTM + Bahdanau Attention (Baseline 1)	27
4.4	Model 2: CLIP + GPT-2 (Baseline 2)	27
4.5	Model 3: ViBERT (Proposed)	28
4.5.1	Visual Encoder	28
4.5.2	Visual to Language Projection	28
4.5.3	BERT Decoder and Caption Retrieval	29
4.5.4	Training Configuration	29
4.6	Common Training Configuration	29
4.7	Evaluation Protocol	30
5	RESULTS AND DISCUSSION	31
5.1	Introduction	31
5.2	Quantitative Results	31
5.3	Dynamics of Training	33
5.4	Computational Efficiency	34
5.5	Qualitative Analysis	35
5.6	Cross-Study Synthesis	36
6	DISCUSSION, LIMITATIONS, AND FUTURE WORK	38
6.0.1	Discussion of Findings in the Context of the Survey	38
6.1	Limitations of the Present Work	38
6.2	Future Research Directions	39
7	CONCLUSION	41

Appendix	47
7.1 List of Publications	47
7.2 Paper Submission Proof	47

List of Tables

3.1	Taxonomy of image captioning architectures with representative design properties.	11
3.2	Image captioning datasets with scale, domain coverage, and primary applications.	17
3.3	Captioning evaluation metrics with semantic sensitivity, computational cost, and known limitations.	19
3.4	Approximate component-level performance on MS COCO showing incremental contribution of each architectural element.	21
3.5	Representative captioning models across architectural generations. Note: all B-4 and CIDEr scores are on MS COCO Karpathy test split except ViBERT (*) which reports Flickr8K test scores. Direct numerical comparison between ViBERT and other models is not meaningful due to this dataset difference.	24
5.1	Performance comparison of image captioning models on Flickr8K dataset (* indicates best score).	31
5.2	Computational efficiency comparison (* indicates best).	34

List of Figures

3.1	Proposed taxonomy of transformer-based image captioning architectures with chronological timeline. Categories are distinguished by encoder type, decoder type, and primary training paradigm.	10
5.1	BLEU score comparison across all three models on the Flickr8K test set. ViBERT achieves the highest BLEU-4, indicating superior phrase-level accuracy.	32
5.2	Complete metric comparison showing all seven evaluation metrics across the three models. ViBERT leads on BLEU-4, CIDEr, and SPICE — the metrics most closely associated with semantic understanding.	33
5.3	Training and validation loss curves for all three models. ViBERT is the fastest to converge and reaches the lowest final loss value at epoch 10.	33
5.4	ViBERT training and validation loss curves showing rapid convergence by epoch 5, with training and validation losses converging to approximately 0.35 by epoch 10.	34
5.5	Efficiency comparison showing training time, inference speed, and parameter count across all three models. ViBERT achieves significantly faster inference despite its larger parameter count.	35
7.1	Submission Proof for Paper 1: ViBERT	47
7.2	Submission Proof for Paper 2: Survey	48

LIST OF SYMBOLS AND ABBREVIATIONS

NLP	Natural Language Processing
CV	Computer Vision
DL	Deep Learning
ML	Machine Learning
BERT	Bidirectional Encoder Representations from Transformers
ViT	Vision Transformer
ViBERT	Vision-BERT (proposed model)
CNN	Convolutional Neural Network
RNN	Recurrent Neural Network
LSTM	Long Short-Term Memory
GPT	Generative Pre-trained Transformer
CLIP	Contrastive Language-Image Pretraining
BLIP	Bootstrapping Language-Image Pretraining
VLP	Vision-Language Pretraining
LLM	Large Language Model
RAG	Retrieval-Augmented Generation
AdamW	Adam Optimizer with Weight Decay
BLEU	Bilingual Evaluation Understudy
CIDEr	Consensus-based Image Description Evaluation
SPICE	Semantic Propositional Image Caption Evaluation
ROUGE	Recall-Oriented Understudy for Gisting Evaluation
METEOR	Metric for Evaluation of Translation with Explicit ORdering
TF-IDF	Term Frequency-Inverse Document Frequency
FRCNN	Faster Region-based Convolutional Neural Network
SCST	Self-Critical Sequence Training
MLM	Masked Language Modelling
GPU	Graphics Processing Unit
DTU	Delhi Technological University
MS COCO	Microsoft Common Objects in Context
LoRA	Low-Rank Adaptation
mBERT	Multilingual BERT

CHAPTER 1

INTRODUCTION

1.1 Background and Motivation

The skill to observe a visual scene and describe it in language terms is the easiest human cognitive task. An individual looking at a picture even for a brief period of time will be able not only to recognize all objects in the scene but also understand how they relate to each other, what actions are being carried out by the people involved in the scene and what mood prevails there. For many years, scientists specializing in artificial intelligence, computer vision, and computational linguistics have been exploring the problem of developing a system capable of doing the same thing as efficiently as humans do. The task is highly complicated; yet it has been attracting great interest in the scientific community for more than ten years already.

Computational image captioning refers to the challenge of producing an adequate natural language sentence or series of sentences describing the content of a particular image. This challenge differs significantly from related problems, like image classification, where the machine simply needs to label the image with a certain term from a given vocabulary, and image retrieval, which involves ranking a set of text-based descriptions based on a given image. Image captioning, however, is essentially a generative process where the sentence must be composed one word at a time out of an infinite number of possible vocabulary terms based on the visual input; and it should meet the three conditions of grammatical, factual, and communicative adequacy, all at the same time.

Previous approaches in the realm of image captioning, prior to the deep learning revolution, were mainly of two types, namely, template-based and retrieval-based. Template-based systems would construct sentences based on object and scene properties detected by trained classifiers for those attributes and then use the labels for inserting them into sentences according to template structures. Although this method could generate sentences with a correct structure, its drawback was that it would always be limited in complexity and incapable of understanding any relations within contexts. On the other hand, retrieval-based algorithms avoided the problem of generation altogether by simply matching the input query image with a collection of images labeled by humans, which could be problematic because it depended on the database being used.

The emergence of deep learning, especially through the development of Convolutional Neural Networks for visual information representation together with Recurrent Neural Networks for generating sequences of textual information, was a milestone

achievement. The works of Vinyals et al. [1] and Karpathy and Fei-Fei [2] proved that an encoder-decoder approach based on compressing visual information by means of CNNs into a fixed-size vector and decoding the vector using LSTMs into a sequence of words yielded impressive results on standardized benchmarks. They laid out the groundwork for the encoder-decoder models that dominated the field for years to come.

However, the encoder-decoder architecture in its original form suffered from one important shortcoming. First, the full visual representation of the image was needed to be encoded into a single fixed-size vector prior to any words generation. In other words, an information bottleneck arose due to the fact that the encoder had no clue about what regions of the image will turn out useful for caption generation in each individual step of decoding process, and hence all necessary information could be missed out in the vector summarizing the input image. This led to lower-quality caption generation for those images that contained several interacting objects or had a highly unusual spatial structure. The idea of introducing attention mechanism into the caption generation pipeline by Xu et al. [3] allowed to overcome this bottleneck by learning to focus on certain image regions during decoding process.

In the following period, a number of innovations occurred in architecture, with each of them aimed at overcoming certain shortcomings associated with its predecessor. Specifically, bottom-up and top-down attention mechanisms [4], instead of generic features that are provided by a spatial grid of features, use semantically grounded representations generated by an object detection model, thus offering the decoder a richer vocabulary to operate on. Additionally, the graph-based neural network models have improved the way visual relations are represented. All these developments notwithstanding, one critical drawback of the convolutional encoder was its inability to capture long-range dependencies in the image because of the local nature of the convolutional kernel.

With the emergence of the Transformer framework [5] for sequential data in natural language processing, there arose a new paradigm of using self-attention, wherein the relationship between any two positions within a sequence, irrespective of how distant, is directly modeled. The application of this paradigm to visual data in the Vision Transformer (ViT) architecture proposed by Dosovitskiy et al. [6], where visual input was represented as a sequence of patches and attention was applied to each of these patches, showed that the representation thus generated could match or even outperform convolutional neural networks in the task of image classification, while providing a fundamentally different representation method. For the problem of image captioning, the significance of this observation was clear: the ViT-based encoders could encode context at the level of scenes by considering all image patches at once through attention.

In language generation, BERT [7] represented the emergence of deep bidirectional contextuality in language modelling. While autoregressive models like GPT generate text by reading one token at a time from left to right, BERT implements a technique called masked language modelling, allowing each token to consider all the other tokens in the sequence in parallel, creating highly context-sensitive representations and delivering state-of-the-art results in a wide range of natural language processing ap-

plications. In this dissertation, I investigate the hypothesis that it would be feasible to integrate a global visual representation from ViT with the bidirectional understanding of language offered by BERT into one captioning model.

This thesis represents a response to an observed phenomenon: despite the flourishing of transformer captioning architectures in the last few years, no systematic effort has been made to understand what makes their performance increase, how much they improve on each other, and where the conceptual limits of their current designs lie. In addition to contributing new findings through experimentation, there is a demand for synthetic scholarly research that systematizes existing knowledge. This dissertation fulfills both these needs, as it provides an original empirical study of the ViBERT architecture as well as a review of the literature on the topic.

1.2 Problem Statement

The fundamental technical problem explored in this dissertation is that of inefficiencies in image captioning architectures’ inability to provide high semantic fidelity and computational efficiency together. Significant advancements in the state of the art on established benchmarks have been made during the past several years; however, some crucial shortcomings still exist. First, local receptive fields of CNN encoders do not allow modelling of the global scene context and long-range object dependencies. Second, the sequential nature of LSTM decoders makes it impossible for these models to exploit any opportunities for parallelisation in the process of inference and training. Third, it remains challenging to develop an effective integration of the benefits offered by powerful vision encoders and powerful pre-trained language models without requiring prohibitive computation capabilities.

In addition to the aforementioned problem, there is also another more methodological issue faced by the field. Although many new models have been proposed recently, the level of comparative analysis performed to date does not reflect the scope of the advances. Specifically, models tend to be compared solely based on n-gram metrics such as BLEU-4 and CIDEr, which, despite their popularity, do suffer from important biases. Moreover, the phenomenon of caption hallucinations, i.e., description of objects and attributes that are absent in the images, although well-known and prevalent, is rarely included into the experimental evaluation protocol due to the lack of a standard framework.

1.3 Aims and Objectives

These objectives can be classified into two overarching goals for the dissertation. The first goal is to conduct an exhaustive and analytically robust literature survey of various transformer-based captioning methods. Such a survey will include not only the listing of the existing literature but also the identification of the design motivations, architectural choices, and performance evidence associated with each type of approach. The second objective is to devise and experimentally evaluate an innovative transformer captioning architecture named ViBERT. In particular, the latter task will be focused

on implementing, training, evaluating, and comparing the new architecture against two chosen baselines using the Flickr8K dataset.

More specifically, the individual objectives within the stated goals are as follows. First, to devise and propose a categorisation framework that includes nine categories of captioning architecture and reflects the evolution of the field. Second, to conduct a survey of the ten available benchmark datasets, focusing particularly on the less represented datasets such as those containing medical images. Third, to perform a comparative analysis of eight metrics, highlighting each metric’s theoretical foundations, empirics, weaknesses, and context of use. Fourth, to enumerate nine open problems and issues in captioning research, along with the description of the best-known mitigation strategies. Fifth, to implement, train, and evaluate the devised ViBERT architecture on the Flickr8K dataset by employing a set of captioning metrics. Sixth, to perform a comparison between the new model and two carefully selected state-of-the-art baselines and provide both quantitative and qualitative analyses of all results. Seventh, to undertake an efficiency analysis of the newly proposed architecture in terms of the number of parameters, training time, and inference speed.

1.4 Scope and Organisation of the Thesis

In this dissertation, the problem of image captioning will be investigated with the use of a theoretical framework and an experimental approach with the focus on transformer-based models. The experimental part will only consider the use of the Flickr8K data set due to the possibility of comparing various models in a controlled way using the available computational resources. Literature review in the current dissertation will cover the literature until the end of 2025, with special attention paid to architecturally influential works that have marked the path of development in the field. It must be noted that the topics such as video captioning, audio captioning, and more specific granularities than the entire image in scene description are not considered in this thesis, with an exception of their brief overview in the future work section.

The rest of this dissertation will be structured as follows. Chapter 2 will provide theoretical background, covering the fundamental concepts of deep learning, CNNs, recurrent networks, the Transformer architecture, vision transformers, BERT model, and contrastive pretraining required for further reading. Chapter 3 will contain a literature survey on transformer-based image captioning architectures, which will follow the classification taxonomy by covering each category separately along with the analysis of data sets, metrics, open challenges, and comparative analysis. Chapter 4 will present the experimental method used in this research, describing the process of data preparation, model architecture, and training process. Chapter 5 will introduce the experimental results, providing the comparative analysis using metric scores, analysis of training process, evaluation of efficiency, and qualitative examples of captions. Finally, Chapter 6 will discuss the implications of the findings, limitations of the conducted research, and the directions for future work.

CHAPTER 2

THEORETICAL BACKGROUND

2.1 Overview

The theory presented here is intended to provide essential groundwork needed in order to properly comprehend the architectural considerations, training techniques, and evaluation metrics that will be discussed in following chapters. The theoretical background presented is structured with respect to the chronological evolution of ideas – from fundamentals of neural networks, through convolutions, recurrent networks and finally transformer architecture and pretraining. Instead of presenting detailed mathematically derived explanations we aim at providing a common ground and introducing terminology that will allow for distinguishing various architectural generations mentioned in this dissertation.

2.2 Convolutional Neural Networks for Visual Feature Extraction

Convolutional Neural Networks form the basis of visual feature extraction in the CNN-RNN captioning approach and continue to be an important element of hybrid models. The fundamental process performed by each convolutional layer consists in applying a learnable filter to local patches of the input feature map, resulting in feature maps containing information about local structural patterns. The stacking of successive convolutional layers with occasional pooling operations leading to reduction in spatial resolution but increase in receptive fields allows the use of deep CNNs to build hierarchies of features – low level layers respond to simple edges and textures, while high-level layers respond to more complex patterns like objects and their parts.

ResNets [8], which pioneered the use of residual skip connections, enabling gradients to flow better through deep neural networks, emerged as the leading choice of CNN backbone until around the mid-2010s. The ResNet-50 variant, having fifty layers and roughly twenty-five million parameters in its backbone, outputs spatial feature maps containing local and somewhat global information about the images. The output of the last convolutional layer of the ResNet-50 model, before going through global average pooling, consists of a two thousand forty-eight channel deep, seven-by-seven grid of spatial feature maps. These maps, which contain forty-nine feature vectors each of dimensionality two thousand and forty-eight, form the basis of the attention model used in this work.

While highly successful, CNNs are limited in their ability to capture global information for tasks such as captioning, since the effective receptive field of a convolutional filter increases with the depth of the network. However, information from spatially distant regions needs to pass through multiple layers before it can affect the same feature location in the feature map. Thus, for a captioning task that requires knowledge of the relationship between objects in faraway locations in the scene, this is a significant representation bottleneck that is not easily alleviated by deeper architectures.

2.3 Recurrent Neural Networks and Long Short-Term Memory

RNNs are the natural choice of network architecture to encode sequential text in encoder-decoder captioning models. As their name suggests, RNNs keep a hidden state that evolves according to the current input and the current value of the hidden state. Due to the vanishing gradient problem, standard RNN units have difficulty processing long sequences and hence capturing temporal context. The solution provided by the Long Short-Term Memory cell [9] is the use of explicitly designed memory units – an input gate, a forget gate, and an output gate that allow for regulation of the information entering, leaving, and being retained inside the cell state.

In the case of image captioning, the LSTM cell is used as a decoder that takes the image representation vector as its initial hidden state or conditioning input and predicts a sequence of words, predicting one word token per time step according to its probability distribution given the current hidden state and the word embedding of the previously predicted word. Teacher forcing is applied during training, i.e., instead of using the predictions as inputs in the subsequent time steps, the model receives the ground-truth word of the caption from the training set.

One of the defining properties of LSTMs employed as image caption decoders is their inability to perform computations on a whole sequence of tokens simultaneously: the hidden state of the decoder at time step t is calculated based on the previous hidden state at time step $t - 1$, and thus parallel computation on multiple time steps cannot be performed during training. Consequently, training on large batches and training of very long sequences become increasingly expensive in terms of computational resources. Additionally, the inclusion of an attention mechanism into LSTM-based image captioning further enhances the quality of generated captions but does not change the sequential dependency between time steps.

2.4 The Transformer Architecture

Transformer, proposed by Vaswani et al. in 2017, is the first architecture where self-attention has been used to fully replace recurrent connections and allow computations to be executed in parallel. Transformers learn fully independent relationships between any pairs of inputs and do not suffer from long path problem that plagued LSTMs

and other recurrent architectures. The key building block of a transformer is the multi-head self-attention layer which computes weighted sum of all positions in the sequence using dot-product similarity between query, key, and value projections of the input tokens.

Formally, given input tensor $X \in \mathbb{R}^{n \times d}$ where n is number of tokens and d is their dimensionality, the self-attention operation is defined by:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V \quad (2.1)$$

where $Q = XW_Q$, $K = XW_K$, $V = XW_V$ are linear transformations with matrices $W_i \in \mathbb{R}^{d \times d_k}$ and d_k is key embedding size. The scaling factor by $\sqrt{d_k}$ avoids saturation of the softmax function by limiting possible values of the scalar product between queries and keys. In the multi-head variant of self-attention operation multiple heads compute attention with their own set of parameters and outputs from them are aggregated linearly.

The lack of any positional information in the attention operation necessitates explicitly encoding the positions of the elements of the input. The initial transformer uses fixed sinusoidal positional encoding, but more recent architectures explore learned and relative positional encodings. Fully parallel processing in the transformer architecture enables training on significantly larger sequence lengths compared to LSTMs, along with modeling of the long range dependencies directly through the attention mechanism, making it ideal for complex relational reasoning tasks.

2.5 Vision Transformers

The extension of the Transformer to vision problems in the form of the Vision Transformer [6] treats the image as a sequence of fixed-size patches. In the case of the ViT-Base/16 variant, an input image of size 224×224 is divided into 16×16 patches, resulting in a sequence of 196 patches. These patches are then linearly projected onto the desired number of dimensions and a learnable class token is prepended to the sequence, resulting in 197 tokens total. Positional encodings are learned and added to the token embeddings, resulting in an input sequence of token embeddings passed to the vanilla transformer encoder consisting of alternating self-attention and feed-forward blocks followed by residual connections and layer normalization.

The ViT presents an entirely different type of visual representation than CNNs. Since every patch attends to every other patch already at the first transformer layer, long-range spatial dependencies are directly modeled in contrast to gradually building local dependencies via convolutional filters. In fact, it has been empirically observed that transformer-based image representations tend to capture such dependencies better, which is crucial for representing high-level concepts such as complex scenes or relationships. This property of the transformer enables it to excel at generating image descriptions that require reasoning about relationships between objects and their positions in space. On the other hand, the class token captures the global representation of the entire image while patch tokens represent localized spatial structure.

One primary drawback of ViT architecture relative to CNN-based methods lies in the absence of the inductive biases associated with convolutional structures, specifically local connections and approximate translation equivariance. Consequently, ViT models need significantly more training data in order to learn enough statistics to ensure effective generalisation. To alleviate this problem, the common approach involves pre-training ViT models on large-scale image recognition datasets (e.g., ImageNet-21K) and then fine-tuning on a captioning task, which allows leveraging statistical regularities learnt from a vast number of images in pretraining before learning the distribution peculiarities of the captioning dataset.

2.6 BERT and Bidirectional Language Model Training

BERT model [7] introduced in 2019 by Devlin et al. was revolutionary in terms of the shift towards language modelling pretraining paradigms. Prior to BERT, all state-of-the-art language models, including famous GPT-1, followed an autoregressive pretraining objective, which imposed the necessity to process each token strictly in a left-to-right order and take into account only previous tokens when encoding a current token. BERT model pretrains on a masked language modelling objective, where randomly selected tokens are masked out in the input text and replaced by a special token. The BERT model learns to reconstruct masked tokens while attending to tokens before and after the current one, thereby forming a bidirectional contextualised representation. Due to the ability to attend to context to the right of the token under consideration, this model is capable of producing extremely useful semantic representations for subsequent natural language processing tasks.

From the perspective of text generation for image captioning, the main potential benefit of employing bidirectionality provided by BERT is the ability to generate semantically rich and consistent texts. At the same time, the bidirectional nature of BERT poses a problem for autoregressive generation since the objective used for language pretraining does not match the left-to-right processing necessary for autoregressive generation. Different approaches have been suggested to solve this problem by adjusting the attention mask to make the model capable of both understanding (in bidirectional mode) and generating texts (in left-to-right mode) at different stages of training (see *UniLM* [10] and *Visual Language Pretraining* [11]). In this work, we utilise the model architecture, which does not rely on autoregressive generation and follows a retrieval-based approach.

2.7 CLIP and Contrastive Learning

The emergence of CLIP [12] marked a revolutionary shift away from the idea of pre-training visual representations with supervised signal. Instead, the authors trained dual visual and textual encoders on an enormous collection of image-text pairs from the web, maximising the similarity between matched pairs and minimising it for mismatched pairs in each minibatch during training. In this way, the authors created a joint embedding space where similar in semantics images and text were placed near

each other, making possible a wide array of applications, including but not limited to zero-shot image classification and cross-modal retrieval.

As far as the visual encoder goes, there is the option of both ResNets and ViTs, the latter being a backbone for CLIP-B/32 in the configuration we consider to be one of the baselines for image captioning in this dissertation. The visual encoders trained in the fashion outlined above yield extremely semantic representations since they are pre-trained on a contrastive task, yet the fact remains that CLIP itself was trained for a contrastive task, and therefore cannot be conditioned on the individual token level needed for quality image captioning. Instead, some sort of projection layer needs to map the visual representation to a sequence of prefix tokens which in turn condition the language model, like GPT-2.

2.8 Metrics for Evaluating Image Captioning

Image captioning models are evaluated based on several automatic metrics, where each metric emphasizes certain characteristics in the generated and reference captions. Knowing the characteristics and weaknesses of each metric is crucial to interpret the evaluation results accurately.

Firstly, BLEU [13] calculates the precision of n-grams in the generated caption with respect to a pool of reference captions, adjusted by a brevity penalty. The most popular version of this metric, BLEU-4, measures the four-gram precision in a generated caption, reflecting the accuracy of phrases in the generated caption. Secondly, CIDEr [14] evaluates n-gram matches weighted by their term-frequency-inverse document frequency (TF-IDF). In other words, the captions that include certain n-grams that occur often in the references but rarely elsewhere in the entire dataset receive higher scores than the rest. Thus, such a metric discourages generic captions and rewards more specific descriptions of an image. Thirdly, SPICE [15] evaluates caption generation models from an entirely different angle. The approach relies on parsing the generated and reference captions into scene graphs and computing the F-score of their object, attributes, and relationships. Hence, such a metric is focused on the accuracy of semantics and less prone to paraphrase blindness of n-grams. Finally, ROUGE-L [16] computes the length of the longest common subsequence, representing the recall of key information. These metrics form the evaluation framework used in the experiment presented in Chapter 5.

CHAPTER 3

LITERATURE REVIEW AND SURVEY OF TRANSFORMER-BASED IMAGE CAPTIONING ARCHITECTURES

3.1 Overview and Taxonomy

3.2 Taxonomy and Overview

Over the last decade, there have been many developments in image captioning, with the paradigm evolving from template and retrieval methods to encoder-decoder networks, to the current period of large language model integration.

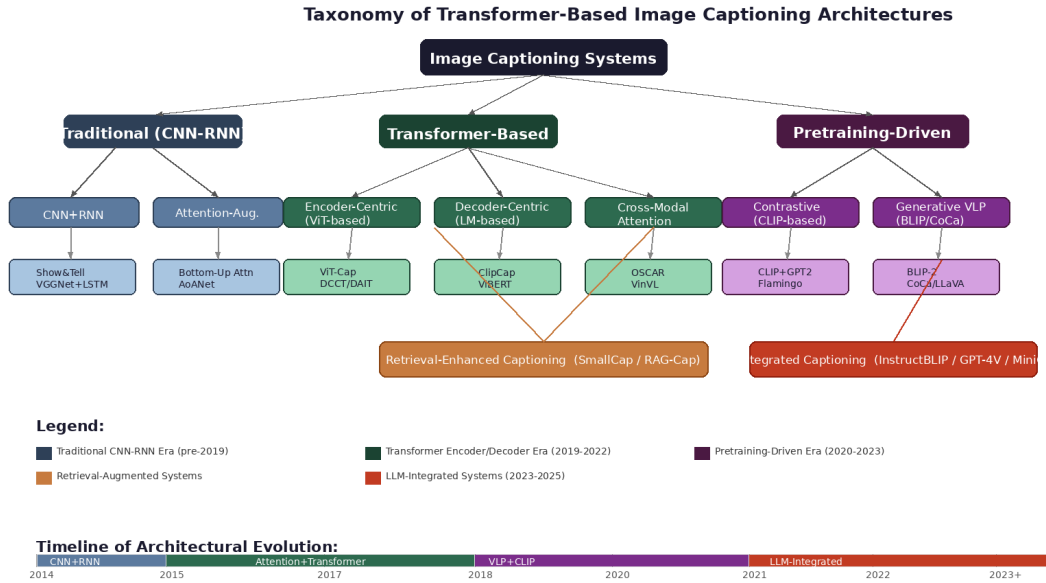


Figure 3.1: Proposed taxonomy of transformer-based image captioning architectures with chronological timeline. Categories are distinguished by encoder type, decoder type, and primary training paradigm.

In order to survey this body of work effectively, it is necessary to adopt a taxonomy that accounts not just for technical differences between individual models but also for the major theoretical shifts that define each new generation of models. This chapter

introduces such a taxonomy in the form of nine categories and applies it to perform a literature review. This taxonomic structure is visualised in the timeline in Figure 3.1, where the relationship between each of the nine categories and the general progression of the field is illustrated.

These categories include CNN-RNN baseline models, attention-enhanced CNN-RNN models, encoder-focused vision transformers, decoder-focused language modeling techniques, cross-modal attention and vision-language pretraining, CLIP-based contrastive learning, retrieval-augmented models, unified multimodal transformer models, and language model integrated captioning systems. These categories are not necessarily distinct from one another in the modern period, since many recent models incorporate aspects of several different categories into their architecture.

Table 3.1 lists the key design characteristics of each of these nine categories of models.

Table 3.1: Taxonomy of image captioning architectures with representative design properties.

Category		Visual coder	En- coder	Text coder	De- coder	Attention	Training
CNN+RNN	Sys- tems	Local features	conv	Sequential LSTM		None	Cross-entropy loss
Attention- Augmented		Spatial/object features		LSTM + At- tention		Bahdanau/Luong	CE + SCST
Transformer	De- coder	CNN/ResNet encoder		Transformer decoder		Multi-head self-attn	SCST
ViT	Encoder- Centric	Patch-level ViT		Transformer decoder		Cross-patch attention	AdamW + co- sine
BERT	Decoder- Centric	ViT/CNN encoder		BERT-style module		Bidirectional attn	MLM + CE
CLIP-Aligned		Frozen CLIP encoder		GPT- 2/Transformer		Contrastive prefix	Prefix fine- tuning
Retrieval- Enhanced		ViT embed- dings		Retrieval + generate		Similarity search	Contrastive
Unified VLP		End-to-end ViT		Shared transformer		Cross-modal attn	Joint pretrain- ing
LLM-Integrated		ViT + Q- Former		Frozen LLM		Q-Former cross-attn	Stage-wise VLP

3.3 CNN-RNN Baseline Architectures

The first model proposed by Vinyals et al. [1], as well as by Karpathy and Fei-Fei [2] in 2015, comprised an encoder part where the image was passed through a CNN, followed by a decoder LSTM [?], which decoded the image into a sentence word-by-word. More precisely, the authors of the highly regarded Show and Tell model [1] used the vector produced by max-pooling all the features extracted by the GoogLeNet to initialize the decoder LSTM, which was subsequently trained end-to-end to predict the tokens of a caption in the MS COCO dataset [17].

These methods were quite innovative since they demonstrated the possibility of developing an automatic mapping between image and text using an end-to-end approach. While their results would be considered formulaic and simplistic today, they were still perfectly coherent sentences which accurately described fairly simple scenes. They set the benchmark and established a framework for future research, by introducing the architecture of the encoder-decoder pair, setting a protocol for evaluation, as well as by using the same datasets.

However, there was a problem with this kind of approach which is inherent to it, i.e., the fact that the input image was represented as a fixed-length vector, which in turn initialized the decoder, meaning the encoder had no control over the information which the decoder would use to generate the description. This meant that whatever information was left out of the representation would be lost. It created a serious problem for more complex scenes with multiple relationships between objects in it. The model could not provide a comprehensive description without fabricating any facts. Another issue was related to the decoder LSTM – long sequences required much computation since gradients would have to be backpropagated throughout the entire sequence.

3.4 Attention-Augmented CNN-RNN Systems

The incorporation of spatial attention into the captioning pipeline by Xu et al. [3] in their Show, Attend and Tell model resolved the problem with the fixed vector produced by the encoder and gave the decoder the opportunity to focus on different areas of the convolutional feature map while generating different words. Instead of relying on the pool of feature vectors, the attention mechanism kept the spatial grid of convolutional features and used them to compute a weighted average of those features with respect to their importance for generating each respective word in turn. There were two types of such attention mechanisms proposed – soft and hard attention, the latter relying on training a discrete distribution by means of the REINFORCE algorithm and thus being more complicated to implement, while the former was easy to train by backpropagation and thus became more popular. The visualization of soft attention weights over the input image yielded interesting results, as the attention focused precisely on that area of the input image where the respective object that the generated word referred to resided; for example, while producing the word “dog”, the attention focused on the region where the dog appeared on the input image.

This line of research got pushed even further by Anderson et al. [4], who re-

placed generic spatial features by a set of semantically meaningful object descriptors extracted by the Faster R-CNN detector pre-trained on the Visual Genome dataset. By combining this approach with top-down attention, the system learned to produce not only visually salient features but also semantically relevant ones by generating an explicit set of features corresponding to detected objects and their attributes. The use of semantically meaningful regions of the image in combination with attention produced state-of-the-art results on the MS COCO benchmark in 2018-2020, making this approach dominant for image captioning and the base for numerous extensions and improvements. The extension known as attention-on-attention added one more layer of attention to the system, while others used a graph structure to encode relations between detected objects explicitly.

Unfortunately, the issue of information being required to pass through several layers of intermediate representations for a long-range spatial interaction to manifest into a common feature map location remained. If the most important interactions in the image were those between far-away image parts, this long path would create a bottleneck, not entirely addressable by any form of attention applied during decoding.

3.5 Vision Transformer Encoder Models

The emergence of Vision Transformers [6] marked another paradigm shift in the way visual representations were produced: by encoding the image as a sequence of patch tokens and using multi-head attention at each level between all pairs of tokens, ViTs made possible modelling of long-range spatial interactions from the very first layer of processing. Two distant patches representing, for instance, a person located at one edge of the image and the object he/she reaches for at the other edge of the image can model interactions between themselves, regardless of any intermediate spatial locations.

With respect to performance, ViT-based encoders provided superior BLEU-4 and CIDEr scores in comparison with CNN-based encoders when dealing with images exhibiting complex spatial dependencies and multiple objects. Examples of such architectures include DAIT [18] and several ViT-enabled captioners. However, when dealing with simpler images, consisting primarily of a single object, a pooling operation over the set of CNN features would encode the necessary information sufficiently well. A drawback of ViT models is that unlike CNNs, they do not have the inherent inductive bias of local connections and translational equivariance of features, which is desirable for image data and enables CNN models to be trained efficiently even on smaller datasets. ViT-based captioning models thus require extensive pre-training on large data sets, e.g. Imagenet-21K, prior to fine-tuning.

3.6 Methods Using Pre-Trained Language Models as Decoders

In conjunction with visual encoding research, improvements in pre-training language models led to the creation of powerful text encoders capable of functioning as robust

decoders once properly conditioned on visual features. GPT-2 [19] showed that a transformer text encoder can learn to generate fluent natural language across many topics using massive scale training data and have a built-in prior about grammatically sensible structures and common sense. In contrast, BERT [7] excelled as a feature encoder capable of extracting highly rich contextual representations. Both language models provided the opportunity to use them as a starting point for learning how to caption images in ways that would be out of reach to LSTM decoders trained from scratch.

A prominent example of a method using GPT-2 in captioning is the ClipCap model [20]. The CLIP image embedding was first mapped to a set of prefix tokens in the GPT-2 embedding space conditioning the model on the visual content while keeping the majority of the GPT-2 weights frozen to preserve the learned language prior. Such an approach was both efficient to train and provided fluent image descriptions thanks to the powerful generation capabilities of GPT-2. However, frozen language model weights meant that it would be difficult to condition the generation process specifically on the image distribution seen during training, while prefix conditioning of the transformer model limited the effect of the visual input.

Using BERT for image captioning presented greater difficulties in part because of the difference between bidirectional masked language modelling used in BERT and autoregressive left-to-right text generation at inference time. The problem was overcome by the UniLM architecture [10] where a single unified transformer model is jointly trained using different attention masking strategies. Similarly, the Visual Language Pretraining methodology [11] uses this trick for vision-language tasks. The ViBERT architecture presented in this dissertation overcomes the autoregressive problem with an innovative retrieval-based generation strategy relying only on the bidirectional nature of BERT.

3.7 Cross-Modal Attention and Vision-Language Pre-training

The class of cross-modal attention networks, represented by OSCAR [21] and VinVL [22], delivered the best performance in the MS COCO captioning leaderboard between 2020 and 2022 by integrating high-performing visual encoders with BERT-like language decoders via explicit cross-modal fusion operations. OSCAR employed object tags extracted from the input image using an object detector trained as a visual classifier as anchors that facilitated better alignment between the visual and linguistic representations during the pretraining phase. VinVL showed that enhancing the visual feature extractor’s performance, through increased object detection training data, contributed to better captioning results than architectural changes to the language model component, emphasizing the criticality of the encoder part of the captioning pipeline.

BLIP [23] integrated three types of training tasks in a common pretraining strategy, namely contrastive learning, image-text matching, and language modeling, through a mixture encoder-decoder text module. The use of a joint learning objective enabled BLIP to learn useful representations both for comprehension tasks like retrieval and

visual question answering and generation tasks such as captioning without making any modifications to the architecture for different tasks.

3.8 CLIP-Aligned Contrastive Learning

In CLIP [12], the training of dual visual and textual encoders on a dataset of four hundred million web-scraped image-text pairs via contrastive learning yields an alignment between semantically-rich representations of impressive generality. Representations yielded by CLIP’s visual encoder contain meaningful information which allows them to be used effectively in various downstream tasks, such as zero-shot image classification, cross-modal retrieval, and even a conditional input to a caption generator. In particular, methods motivated by CLIP rely on the rich semantics encoded by pre-trained visual representations and do not require aligning visual and text domains in captioning datasets.

The main drawback of using CLIP for image caption generation is related to the fact that CLIP was intended for contrastive classification and does not include generation capabilities. In other words, although CLIP’s text encoder learns to convert a sentence to a vector representation, it was never trained to predict sentence tokens given visual data. As a result, when using CLIP for generation purposes, such as in ClipCap [20], one has to introduce a projection function which requires new parameters and a training regime tailored to the captioning problem. Additionally, the global nature of CLIP’s image representation (the whole image is compressed to a single vector) makes spatial referencing of image regions challenging.

3.9 Retrieval-Augmented Captioning

Retrieval-augmented generation for captioning proceeds by retrieving captions from a database based on visual similarity to the input image, which are used as context for generating or selecting the output caption. The SmallCap framework [24] showed that, with retrieved caption context, a lightweight transformer can be highly performant compared to large-scale models, which rely on vast amounts of data for pretraining, such as those built using VLP. In this case, the captions retrieved are created by humans, so their style and grammar are already correct. All that the model needs is learning how to adapt them to the particular contents in the input image.

One benefit of using retrieval-based techniques in generation tasks is the ability to produce accurate results when faced with unknown or out-of-distribution items, something difficult to achieve with generative systems. A good example of this type of task is NoCaps benchmark [25], in which generative systems would often use the closest object as the alternative to unseen items instead of accurately describing them. A sufficiently diverse database containing captions for those items would make it possible for the model to accurately describe them. The retrieval approach can also mitigate issues of hallucination since it will ground the description to real human-generated content rather than statistical language priors learned by the model.

3.10 Unified Multimodal Transformers

The CoCa model [26] utilised contrastive and generative learning objectives within one unified architecture, whereby the same majority of weights were used for both objectives, and a dual objective encouraged discriminative alignment of visual-textual inputs as well as generating captions from them. In contrast, Flamingo [27] introduced a unified architecture whereby the visual data were integrated into a frozen transformer language model via additional cross-attention modules inserted into the language model. This architecture made the few-shot learning possible since it preserved the ability of the frozen language model to learn contextually without changing the weights.

Unified models have greatly advanced the research agenda of achieving an efficient unified multimodal architecture capable of performing a wide variety of vision-language tasks, such as captioning, image-text retrieval, visual question answering, and referring expression comprehension. The generalisation abilities of unified models are very useful, but come at the expense of higher computational power requirements. Moreover, unified models trained for several tasks do not necessarily outperform fine-tuned models designed only for specific captioning benchmarks.

3.11 LLM-Integrated Captioning Systems

The latest captioning models leverage the vast capacity and language expertise of LLMs by connecting frozen visual encoders to frozen LLMs via lightweight bridging layers. BLIP-2 [28] introduced the concept of the Query Transformer, or Q-Former, that uses learnable query tokens to obtain visual representations from a frozen ViT encoder and feed those representations to a frozen LLM through cross-attention. Training is performed only on the Q-Former and significantly reduces the number of parameters and computational costs compared to joint finetuning of the entire visual encoder and LLM. Learnable query tokens serve as a bottleneck that reduces visual encoder representations to a format understandable for the LLM.

InstructBLIP [29] enhanced BLIP-2 with instruction-based finetuning of the Q-Former on many different vision-language tasks and thus made it possible to give the model an instruction about what to say about the image. LLaVA [30] showed that an incredibly simple model like a CLIP projection of visual features to the LLM embedding space through linear transformation and finetuning on high-quality instruction data can provide excellent results on vision-language benchmarks. This model highlights the potential role of data quality over architectural design at large scales and prompted the development of LLaVA-1.5 with higher resolution input images and improved training data.

3.12 Datasets Used for Image Captioning

The choice of datasets used for training and evaluating captioners significantly shapes the direction in which future research will progress by dictating not only the types of images and language that the captioners have access to but also what aspects of

captions are prioritized in terms of optimization and what can be concluded based on benchmarks. The following is an analysis of the ten most significant benchmark datasets.

Table 3.2: Image captioning datasets with scale, domain coverage, and primary applications.

Dataset	Images	Captions/img	Total	Domain	Notes
Flickr8K [31]	8,091	5/img	40K	People, animals	Everyday; good for ablation
Flickr30K [32]	31,783	5/img	158K	Events, objects	Diverse activities
MS COCO [17]	~330K	5/img	1.5M+	91 categories	Primary benchmark
Visual Genome [33]	108K	Region	5.4M regions	Objects + relations	Dense region captions
NoCaps [25]	~15K	11/img	166K+	Novel objects	Zero-shot generalisation
CC3M [34]	3.3M	1/img	3.3M	Web images	Pretraining; noisy
CC12M [35]	12M	1/img	12M	Web images	Large-scale VLP
LAION-5B [36]	5B	1/img	5B	Web images	Largest open dataset
PathVQA [37]	~5K	Region	32K Q&A	Pathology slides	Medical domain
IU-XRay [38]	7,470	2/img	14,940	Chest X-rays	Medical captioning

Image captioning evaluation began with Flickr8K and Flickr30K datasets [32]. Both datasets contain common life photographs annotated with five human captions per image and are small enough for ablation experiments on limited computational resources like those used for experiments in this dissertation. MS COCO [17], now the most popular captioning benchmark dataset, contains 330,000 images belonging to 91 semantic classes and annotated with multiple captions per image. COCO Karpathy split [2] used by essentially all papers published in this area makes it possible to compare experimental results achieved in the last decade. Visual Genome [33] offers richer data than Flickr and COCO with object-level captions, attribute labels, and explicit object-object relationship annotations, allowing to train models for learning fine-grained spatial semantics. NoCaps dataset [25] includes COCO-style images sampled from Open Images dataset in order to evaluate how well a model generalises to unseen objects and concepts not included in COCO training dataset.

Unlike the above datasets used for captioning image classification tasks, the medical image captioning datasets have a significantly different nature due to both the difference in domains of application and a wider range of problems solved on them. IU-XRay [38] includes pairs of X-ray images and associated structured radiological report. The task for the captioning algorithm is thus not only to analyse the contents of the image but to provide clinically relevant descriptions with respect to particular findings. PathVQA [37] dataset evaluates pathology image understanding based on question-answer annotations. Besides their value as testing datasets, these sets serve as a sign that there is an essential domain gap between the general-purpose image captioning algorithms that were trained on natural photos and their potential application in the medical field. A captioning system that reaches state-of-the-art on MS COCO would be unlikely to achieve good results on IU-XRay because of different visual characteristics and vocabularies required for clinical radiological reporting.

3.13 Evaluation Metrics: Analysis and Limitations

The automatic metrics that are employed for evaluating performance of the image captioning systems play an extremely important role in defining the course of research, since the benchmark results dictate which models succeed and which fail. This is why understanding what a particular metric measures and, moreover, what it fails to measure, as well as the areas in which it correlates poorly with human judgment, is of critical importance. The table below provides information regarding eight metrics utilized throughout this dissertation and other works on image captioning.

Table 3.3: Captioning evaluation metrics with semantic sensitivity, computational cost, and known limitations.

Metric	Basis	Semantic Sensitivity	Cost	Key Limitations/Notes
BLEU-1 to 4	n-gram precision	Low–High	Low	Standard; BLEU-4 most cited; paraphrase-blind
METEOR	Precision +recall +stem	Medium	Medium	Better human correlation; needs linguistic resources
ROUGE-L	LCS recall	Medium	Low	Fluency-oriented; ignores semantics
CIDEr	TF-IDF n-gram	High	Medium	Best overall human correlation; used in SCST
SPICE	Scene graph F1	High	High	Best semantic accuracy; parser-dependent
CLIP-S	CLIP similarity	High	High	Reference-free; hallucination-aware; newer standard
RefCLIP-S	CLIP +references	High	High	Combines grounding with reference alignment
CHAIR	Object hallucination	Very High	High	Measures hallucinated object percentage; [39]

The BLEU metric [13] estimates the precision of generated n-grams against the set of reference captions, with BLEU-4 – four-gram precision – being the most commonly referenced one, because it is sensitive to agreement in phrases, not just individual words. The metric is efficient and easy to compute, hence its popularity. The main drawbacks of BLEU have long been known: it doesn’t consider paraphrases, lacks sensitivity to semantic equivalence of phrasings and saturates on standard datasets way before the level reached by human captions, because it evaluates precision only and therefore prefers shorter captions with a higher number of shared correct n-grams to longer captions, including potentially more relevant information.

CIDEr metric [14] tries to address some of the problems mentioned above by using TF-IDF weight for each generated n-gram. Specifically, CIDEr assigns high weights to rare in the set of references n-grams and low weights to common ones, since the latter are likely generic terms present in every caption describing the image in question. Thus, it introduces greater incentives for generating descriptive captions, as opposed to generic ones. Moreover, CIDEr correlates significantly better with human judgement than BLEU, and it is also used during the SCST reinforcement learning fine-tuning procedure, where captioning models learn to maximize the difference between CIDEr scores of randomly sampled caption and its greedy counterpart, leading to con-

siderable improvements on standard benchmarks but raising concerns about potential over-optimization of the metric.

Among commonly used captioning metrics, SPICE [15] stands out as being semantically richer, evaluating F-score agreement in terms of parsed scene graphs constructed from generated and reference captions, thus taking into account objects, their attributes and relations between them. Because it deals with semantic information only, it is insensitive to paraphrasing and focuses explicitly on whether the caption describes the visual content properly. The key issue with SPICE is that it is highly sensitive to parsing errors: the quality of scene graph parsing determines the outcome of SPICE evaluation.

The CHAIR metric [39], on the other hand, directly measures the proportion of mentioned objects in a generated caption that do not actually appear in the image according to annotated ground-truth data. With hallucination receiving growing attention due to both increased incidence and negative impact, the importance of CHAIR as a performance indicator can be expected to continue growing. Though CHAIR is computed at the object level and requires ground-truth object annotations, its unique capacity to measure factual accuracy makes it an essential complement to traditional metrics such as CIDEr and SPICE.

3.14 Component-Level Ablation Analysis

Determining which particular architectural components are responsible for performance gains requires not only comparing benchmarks, but conducting an analysis of ablation studies – that is, systematic experiments where individual components are removed or modified in order to isolate their contribution to model output. Such ablation studies provide the most reliable source of information regarding the importance of different components.

The most consistent conclusion in the literature is that improvements in encoder performance tend to make a much larger difference in caption quality than improvements in decoder performance. Indeed, each significant change in encoder architecture – from feature pooling to spatial attention, from spatial features to objects defined by Faster R-CNN, and from CNN to ViT – has been accompanied by greater performance gains than those brought about by changes in the decoder.

Component performance contributions on MS COCO, as synthesised from ablation literature, are given in Table 3.4.

Table 3.4: Approximate component-level performance on MS COCO showing incremental contribution of each architectural element.

Configuration	BLEU-4	METEOR	CIDEr	SPICE	Notes
CNN encoder only	36.3	–	120.1	–	Bottom-Up baseline
ViT encoder (frozen)	38.1	28.3	127.4	21.2	Encoder swap only
ViT + BERT decoder	40.2	29.1	134.6	22.8	Full transformer stack
+ Prefix visual tokens	41.8	30.4	138.2	23.5	ViBERT-style injection
+ Cross-modal attention	43.2	31.6	142.0	24.1	Deeper integration
+ Contrastive pre-training	44.5	32.1	145.8	24.7	VLP initialisation

Another noteworthy but perhaps underappreciated insight gleaned from ablation literature pertains to the impact of the quality of the encoder on the benefits of using particular decoders. Modest improvements were found for decoders incorporating spatial attention in conjunction with poor CNNs, while large improvements occurred in conjunction with powerful object-level encoding architectures and ViTs. In other words, there seems to be a dependency between the impact of decoder architecture and the quality of the encoder. The independence assumption often made during ablation studies is thus inappropriate in these cases.

3.15 Open Challenges

In spite of the benchmark advances, there remain nine truly open challenges for the deployment of robust captioning systems in practical applications.

3.15.1 Hallucination

Generating description of entities, attributes, or relations that are not actually represented in the input image is arguably one of the main open challenges faced by researchers in the area. It is connected to the objective function that the system is optimizing during training, namely maximization of the likelihood of the true captions, that causes it to learn captions that are consistent with the statistical distribution seen in the training data set, not necessarily with the input image in question. If the model has learned that dogs are typically accompanied by balls due to seeing plenty of such

cases, it will produce descriptions with balls even for an image of a dog without a ball on it. In turn, the language prior that comes with pre-trained language models makes such hallucinations more frequent.

Some solutions to this problem include CLIP Score optimization, where the score reflects the match between the caption and the image rather than human captions; reinforcement learning, where human feedback on the factual consistency of the caption is rewarded; and grounding constraints where the generated noun phrases are required to correspond to actual regions in the picture. CHAIR [39], that measures hallucination directly, should be used together with other metrics in the evaluation of captioning systems. One interesting solution would be retrieval-augmented generation, where human-written descriptions of real images serve as input to generate captions, thus making them consistent with reality.

3.15.2 Medical Domain Adaptation

Captioning models pretrained on general purpose image-caption datasets fail to deliver satisfactory results for the generation of radiology reports and description of pathology slides. The mismatch includes multiple aspects – on a visual level, medical images have distinct visual properties which distinguish them from natural photographs with subtle patterns of intensity, imaging artifacts and domain specific structures that are hard for general visual encoders to capture; linguistically, medical text contains unique language with domain-specific terms, reporting structure, and abbreviations, and on a factual level, the need for precision is much greater in the domain of medicine due to possible risks arising from hallucinations of abnormal phenomena that can lead to misdiagnosis. The research topics relevant to medical captioning include domain adaptive pretraining on medical images and captions, attention mechanisms that focus on image regions relevant for diagnosis, and structured captioning frameworks separating finding identification and linguistic expression of the finding.

3.15.3 Computational Accessibility

Recent state-of-the-art captioning models leveraging vision-language pretraining on massive amounts of image-caption pairs require massive computational infrastructure that is out of reach for most individual researchers. BLIP-2 [28] requires multi-GPU high memory training setup as well as significant amounts of GPU memory during inference. LLaVA-1.5 [30] makes use of language models with 13 billion parameters. However, there are some ways to mitigate excessive computations such as lightweight adaptation of the model including methods such as LoRA, prefix tuning, and adapter layers. SmallCap [24] managed to achieve strong empirical results using retrieval augmentation method with only 7M parameters in the model. Knowledge distillation from large teacher models into smaller student models is another promising avenue of exploration.

3.15.4 Cross-Lingual and Cross-Cultural Generalisation

Captioning datasets and benchmark tests are predominantly English, as captions are generated by annotators coming from the Western world and thus having specific cultural bias in terms of what aspects should be described and how. This results in potential limitations of captioning models' generalisability and fairness when used across multiple languages and cultures, where captions might need to convey different aspects of the image depending on the perspective of its creators. While there are multilingual benchmarks like COCO-CN and Multi30K, they are considerably smaller than those used in English captioning; additionally, multilingual caption generation trails monolingual English generation both in terms of fluency and accuracy. However, since the use of multilingual pre-trained language models like mBERT can be considered a first step in solving the problem, one needs to note that the image encoder used is based on English images-text pairs and thus may have difficulty capturing relevant information in the case of non-Western cultural context.

3.15.5 Explainability and Trust

Although attention weights of transformer-based captioning systems can be easily visualised and thus give some understanding of how the caption was generated based on particular image regions, this information alone does not give a clear picture of the model's actual causal processes and does not guarantee changes in predictions if the region with a high attention weight is occluded. Nevertheless, more sophisticated explanation approaches like gradient saliency, attention rollout, and probing with DINO allow obtaining much more useful insights into the model behaviour; however, these techniques imply higher computational costs. As far as the captioning system will likely be deployed in critical settings such as accessibility, medical documentation and law, it is necessary to obtain explanations.

3.16 Comparative Analysis Among Architectures' Generations

Table 3.5 summarizes some captioning models among the architecture generations covered in this chapter, based on how well they performed using the MS COCO Karpathy test split. Some key observations can be made when reviewing Table 3.5 alongside the architectural analysis provided earlier.

Table 3.5: Representative captioning models across architectural generations. Note: all B-4 and CIDEr scores are on MS COCO Karpathy test split except ViBERT (*) which reports Flickr8K test scores. Direct numerical comparison between ViBERT and other models is not meaningful due to this dataset difference.

Model	Encoder	Decoder	Attention Training		B-4	CIDEr	Year
Show and Tell [1]	CNN	LSTM	No	CE	27.7	–	2015
Show, Att, Tell [3]	CNN	LSTM +Attn	Soft/Hard	CE	25.0	–	2015
Bottom-Up [4]	FRCNN	LSTM +Attn	Object	SCST	36.3	120.1	2018
AoANet [40]	FRCNN	AoA-LSTM	Attn-on-Attn	SCST	37.2	119.8	2019
OSCAR [21]	ResNet +FRCNN	BERT	Anchor	VLP	41.7	140.0	2020
VinVL [22]	Improved OD	BERT	Tag-enhanced	VLP	44.5	154.8	2021
CLIP +GPT2 [20]	CLIP ViT-B	GPT-2	Prefix tokens	Fine-tune	~22	~84	2021
BLIP [23]	ViT-B/L	MED	ITC +ITM +LM	Joint	40.4	136.7	2022
CoCa [26]	ViT-g	Decoder	Contrastive+CE	Joint	–	143.6	2022
Flamingo [27]	NFNet	Chinchilla	Cross-attn	Few-shot	–	138.1	2022
BLIP-2 [28]	ViT-g/14	FlanT5 /OPT	Q-Former	Stage-wise	–	145.8	2023
InstructBLIP [29]	ViT-g	FlanT5 /Vicuna	Q-Former	Inst-tuning	–	–	2023
LLaVA-1.5 [30]	CLIP ViT-L	Vicuna-13B	MLP proj.	Inst-tuning	–	–	2024
ViBERT (proposed)	ViT-B/16	BERT-base	Visual prefix	AdamW	0.32*	0.97*	2025

The quality of the encoder has always remained the most critical factor influencing

captioning results. For instance, each time we move to a new encoder, whether from pooled CNN features to spatial CNN features, then to the object-level Faster R-CNN features, and finally to ViT patch-based encoders, the results get significantly boosted despite the improvement in the corresponding decoders being much smaller. In particular, BLEU-4 score on MS COCO grew from around 27 to 44 over the last six years, but the boost was not linear and was especially large during 2018-2021, during which both the transition to transformers and vision-language pre-training occurred. However, since 2021, we observe much more modest benchmark improvements, which indicates that MS COCO may become saturated soon, and harder testing setups, such as new objects, cross-lingual transfer, or domain shifts, need to be devised.

Importantly, we should note the significance of pretraining dataset size and quality over architectural advancements. BLIP-2 [28] Q-Former architecture significantly outperforms specialised architectures, which have more complex cross-modal attention, because of higher quality and scale of its pretraining. It means that the academic community needs to think about how to evaluate architectures properly since innovations evaluated on small benchmarks may demonstrate substantial improvements there but will be mostly outperformed in large-scale evaluations due to the scaling effect.

CHAPTER 4

METHODOLOGY

4.1 Overview

This chapter provides a thorough description of the methodological approach followed while building, training, and testing the captioning models considered in the research work, which include Baseline 1 (ResNet50 + LSTM + Bahdanau attention), Baseline 2 (CLIP + GPT-2), and ViBERT. More precisely, this chapter provides detailed information about the dataset used and prepared, about the model architectures along with their technical details, about the training process, and about the evaluation process. Notably, Baseline 1, Baseline 2, and ViBERT have been tested using the same dataset, called Flickr8K, with the same data, except for the CLIP encoder in Baseline 2, which uses the pretrained weights available online.

4.2 Dataset: Flickr8K

The Flickr8k corpus [31] has been chosen for experiments in this dissertation. This corpus includes 8,091 images with real-life scenes, mostly containing people, animals and other objects both inside and outside, each annotated with five independent human captions. The use of five different captions per one image gives an opportunity to vary linguistically the descriptions of the same scene and provides a rich set of references for the metric analysis. Overall, the dataset contains 40,455 captions.

Using Flickr8K instead of MS COCO for the analysis is a conscious choice made based on significant arguments. Firstly, a smaller size of the dataset enables one to train several competing models under controlled conditions within available computational capabilities. Secondly, the use of several captions per image makes possible evaluation of metrics like CIDEr and SPICE based on their ability to find the consensus between several references. Finally, the smaller training dataset gives more chance to notice architecture-related effects on performance since architectures with better inductive prior have advantage in generalization from a smaller amount of examples, whereas those with simple memorization of training samples will do worse in this regard.

Dataset was split into training, validation, and test subsets in accordance with the traditional scheme: 6,472 images for training, 809 for validation and 810 images for test, which amounts to approximately an 80/10/10 split. All images were preprocessed as required for each particular model in terms of resizing and normalization: for ViT architectures, images were resized to the 224×224 pixel resolution and normalized

according to ImageNet statistics. Same preprocessing was applied to the baseline based on the ResNet50 architecture. Captions were lowercase tokenized and augmented with special <SOS> and <EOS> tokens. Training captions were used to create vocabulary by retaining only the words appearing at least five times, and replacing infrequent ones with an unknown token. As a result, the vocabulary included approximately 2,500 tokens.

4.3 Model 1: ResNet50 + LSTM + Bahdanau Attention (Baseline 1)

The first model will serve as a classical baseline model and will represent the CNN-RNN attention-based paradigm, which was a de-facto standard until about 2019. Visual encoder is a pre-trained on Imagenet ResNet50 network [8]. The last convolutional layer of ResNet50 produces a $7 \times 7 \times 2048$ feature map, which is converted into 49 2048-dimensional spatial features. While training, the lower thirty layers of ResNet50 encoder will be frozen, and only the higher layers will be fine-tuned to fit the encoder visual representation for the caption generation task.

As a decoder, I use a classical Long-Short-Term-Memory neural network of 512 neurons with the addition of an attention mechanism. The attention mechanism used is Bahdanau additive attention [41], which takes into account 49 spatial features of visual encoder and outputs attention scores for each of them by calculating alignment between the current hidden state and each spatial feature independently, applying softmax-normalization to the resulting 49 scores, and computing a weighted average of the spatial features. Such a weighted average is called a context vector and is added to a current word vector and fed into an LSTM cell to provide dynamic attention to various parts of an image during caption generation. The caption generation will start from the average spatial feature, passed through the feed-forward network to obtain initial hidden state of the LSTM. Loss used during training is a cross-entropy loss with teacher forcing. Total amount of trainable parameters in Baseline 1 is 41.9 million.

4.4 Model 2: CLIP + GPT-2 (Baseline 2)

The second baseline employs the CLIP visual encoder in combination with the language encoder from GPT-2, constituting a competitive pretraining-based baseline. The visual encoder is represented by ViT-B/32 from the CLIP family [12], which was pre-trained on 400 million image-text pairs through a contrastive language-image pretraining procedure. The weights of the CLIP encoder remain frozen during training, and the visual encoder acts as a pure feature extractor. The CLIP visual encoder outputs one 768-dimensional global vector representation per image, encoding all visual information in the image.

The mapping from CLIP image embeddings to GPT-2 token embeddings is performed via a simple projection network which learns a transformation that maps from 768-dimensional image features to ten GPT-2 token prefixes. These prefixes are then

concatenated with the token embeddings of the captions and fed into the transformer language encoder GPT-2, which generates the captions in an autoregressive fashion. Both the mapping function and the upper layers of the transformer language model are tuned in the fine-tuning stage, in line with the architecture of ClipCap [20]. As before, the motivation for this architecture choice is the preservation of the prior language understanding capabilities of the transformer while adapting it to the visual conditioning task. There are 130.4 million trainable parameters in Baseline 2 – the highest among all the three baselines analyzed in this study.

The distinguishing architectural feature of Baseline 2, differentiating it from both CNN-RNN and the proposed ViBERT, is the use of a completely frozen visual encoder. In other words, the CLIP encoder used in Baseline 2 cannot adapt to the image distribution of the Flickr8K dataset during training, which is a structural disadvantage compared to other architectures. This limitation is, however, somewhat offset by the semantic richness of CLIP representations.

4.5 Model 3: ViBERT (Proposed)

ViBERT is the proposed model for experiments in this dissertation and consists of two components. The ViBERT architecture involves concatenation of the outputs of the ViT-Base/16 encoder [6] and BERT-base decoder [7] via a visual prefix injection, where visual patch tokens are mapped to BERT embeddings and concatenated with textual tokens.

4.5.1 Visual Encoder

The Visual Transformer (ViT) encoder uses the pre-trained ViT-Base/16 configuration trained on ImageNet-21K dataset. A 224×224 pixels input image is segmented into 16×16 pixels patches resulting in a sequence of 196 tokens with a prepended class token (the sequence length equals 197). Next, each patch gets a linear transformation to a 768-dimensional embedding space, then learned position embeddings are added, and the sequence gets transformed via twelve encoder transformer layers each having 12 attention heads and a feed-forward network with hidden dimensionality of 3072. The result is a sequence of 197 token embeddings out of which we select the class token and fifteen first patch tokens forming 16 visual tokens with 768 dimensions which go through the projection layer.

4.5.2 Visual to Language Projection

A learnable linear projection layer projects 16 768-dimensional visual tokens to 768-dimensional BERT embedding space. Then, visual tokens get appended as a prefix to the text input tokens resulting in the visual-text input sequence for BERT encoder. As a consequence, due to bidirectional attention BERT models the relationship between the visual and textual information directly in a single forward pass. The concatenated visual-text input sequence is then encoded by BERT using twelve layers each having twelve attention heads and hidden dimensionality of 768.

4.5.3 BERT Decoder and Caption Retrieval

BERT processes the visual-text input sequence using twelve encoder layers, each with twelve attention heads and hidden dimensionality of 768. As a result, we obtain BERT-based visual representation. We retrieve the caption in the training caption dataset with the highest semantic similarity to the visual representation using cosine similarity as the similarity measure.

This way, the problem of using BERT in an autoregressive fashion and modifying its bidirectional attention pattern to make it left-to-right-masked is bypassed. As such, since the goal of ViBERT is to retrieve similar captions from the captions database via the embedding space created by BERT, it uses the latter not as a language model but as an excellent semantically rich encoder producing visual and textual embeddings for the purposes of retrieval. Thus, in this setup, the architecture of BERT is preserved in order to use it as an encoder, and its ability to generate autoregressive texts is not exploited.

4.5.4 Training Configuration

Only the top three out of ViT’s twelve transformer layers are fine-tuned while the bottom nine are frozen, while only the top four transformer layers of BERT and its projection network are fine-tuned while the bottom eight are frozen. The selective nature of this fine-tuning results in a reduction of free parameters compared to full fine-tuning. Moreover, this fine-tuning regime adds some regularisation in the learning procedure due to the smaller number of parameters compared to full fine-tuning and prevents overfitting due to the small Flickr8K dataset size. The training is done with AdamW optimiser, using a cosine annealing learning rate schedule with a warmup over 200 iterations. The maximum learning rate equals 2×10^{-5} . Weight decay equals 10^{-2} . Dropout of 0.1 is applied just before the language modelling head. The training duration is 10 epochs with a batch size of 32. There are 99.4 million free parameters in ViBERT.

4.6 Common Training Configuration

All three models are trained on one NVIDIA T4 GPU. The shared training hyperparameters among all three models are as follows: batch size equals 32, AdamW optimiser, weight decay equals 10^{-2} , gradient clipping with a maximum gradient norm equal to 1.0, and maximum caption length is equal to 50. For Baseline 1 and Baseline 2, the learning rate scheduler is ReduceLROnPlateau with a patience of 2; if there are two consecutive training epochs in which the validation loss does not decrease, the learning rate will be multiplied by 0.5. For ViBERT, it is the aforementioned cosine annealing learning rate with a warmup period. All three models employ early stopping with a patience of 4 epochs.

4.7 Evaluation Protocol

The evaluation was done on the same test set that contains 810 images in Flickr8K using the 5 human annotations available per image. The metrics that were used to evaluate all of the models were the BLEU-1 through BLEU-4, ROUGE-L, CIDEr, and SPICE. The computation of the BLEU score was done using the standard version with an altered brevity penalty since there were multiple references available. The calculation of the CIDEr and SPICE scores was done using the standard MS COCO toolkit implementations. Computational efficiency was calculated separately based on parameters used, total training time in minutes, and inference time per image in milliseconds.

CHAPTER 5

RESULTS AND DISCUSSION

5.1 Introduction

The results of the conducted experiment are outlined and analyzed in the current chapter. The structure of the section includes four parts – comparison of the three models quantitatively based on all metrics used; analyzing the training dynamics according to the loss function graphs of all three models; computational efficiency comparison, including parameters, learning time, and inference speed; qualitative analysis of captions generated by each model on selected images from the test dataset. The discussion covers the performance of each approach against the corresponding architecture hypothesis stated above and correlates with the literature review outlined in Chapter 3.

5.2 Quantitative Results

The performance of all three proposed models on Flickr8K test data on seven metrics is presented in Table 5.1. The bold typeface highlights the best results in each case. From the results, it is evident that ViBERT outperforms two other models on metrics that are most concerned about the semantic and contextual quality of captions generated. On the other hand, CLIP+GPT-2 shows superiority on word-level fluency metrics.

Table 5.1: Performance comparison of image captioning models on Flickr8K dataset (* indicates best score).

Model	B-1	B-2	B-3	B-4	R-L	CIDEr	SPICE
ResNet50+ LSTM+Attn	0.5006	0.3681	0.2552	0.1642	0.4768	0.8782	0.3040
CLIP+GPT2	0.6392	0.4620	0.3255	0.2224	0.4946	0.8365	0.2847
ViBERT (Proposed)	0.5171	0.4021	0.3535	0.3228*	0.4658	0.9710*	0.3164*

One can notice that there is a significant gap in the performances on BLEU-4. Indeed, ViBERT reaches a BLEU-4 value of 0.3228 that is almost twice as large as the value of 0.1642 reached by the baseline, and almost 45% better than the CLIP+GPT-2 result of 0.2224. BLEU-4 evaluates four-grams precision and is concerned not just

with the presence of correct words in generated captions, but also with their correctness in a sequence of four consecutive words that should be consistent with those in the reference captions. High BLEU-4 performance means that ViBERT generates captions that contain correct phrases of four consecutive words and are not limited to a set of separate words that are individually correct, yet badly arranged. It should be noted that the difference between ViBERT and the other two models in terms of BLEU-4 performance is significantly larger compared to the difference between their BLEU-1 performance – the latter measures unigram precision – where the CLIP+GPT-2 achieves the best score of 0.6392 because of its effective prior of the language model that allows the generation of a good set of words in captions.

The analysis of CIDEr confirms the advantage of ViBERT: the achieved value is equal to 0.9710 while the ResNet+LSTM baseline reaches 0.8782 and CLIP+GPT-2 – 0.8365. As was mentioned above in chapter 3, CIDEr encourages the generation of n-grams related to the specific image being described and penalizes the generation of generic captions that are relevant to multiple images. This implies that ViBERT generates captions that are more precisely related to the test image and are not generic; thus, the description that ViBERT provides contains n-grams related to a particular image and are not common for multiple images. It is worth noting that, even though the CLIP+GPT-2 has access to CLIP features that theoretically include more semantics information than the pre-trained ViT and ResNet50, the obtained CIDEr is even lower than the CIDEr achieved by the ResNet+LSTM baseline, implying the ineffectiveness of the frozen encoder.

Nevertheless, SPICE scores, which offer a measure in terms of semantic accuracy through the scene graphs comparison, are also highly informative in this respect. Among all models, the highest SPICE score is obtained by ViBERT, reaching 0.3164; the second position is held by ResNet+LSTM (0.3040), and CLIP+GPT-2 obtains the third highest SPICE score (0.2847). The importance of SPICE as a metric can be understood from the following statement: SPICE evaluates how accurate is the depiction of the objects, attributes, and relations in the image in the provided caption, and, therefore, SPICE is the closest in meaning to the usefulness of the caption in terms of how a blind person would understand it. The reason why ViBERT has such a high score in terms of SPICE is quite obvious in the context of ViBERT’s architecture.

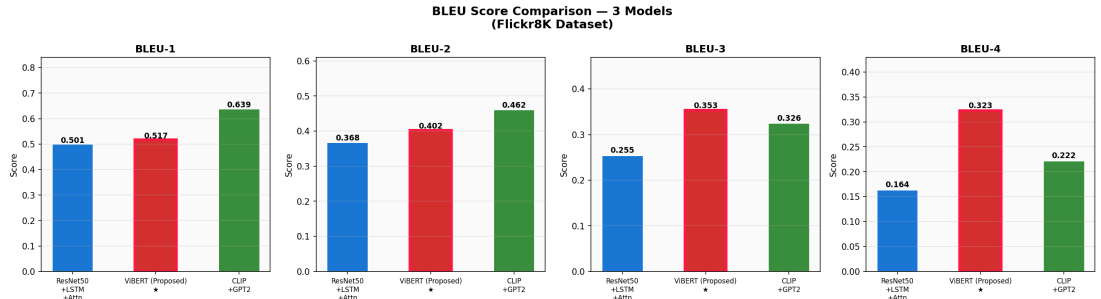


Figure 5.1: BLEU score comparison across all three models on the Flickr8K test set. ViBERT achieves the highest BLEU-4, indicating superior phrase-level accuracy.

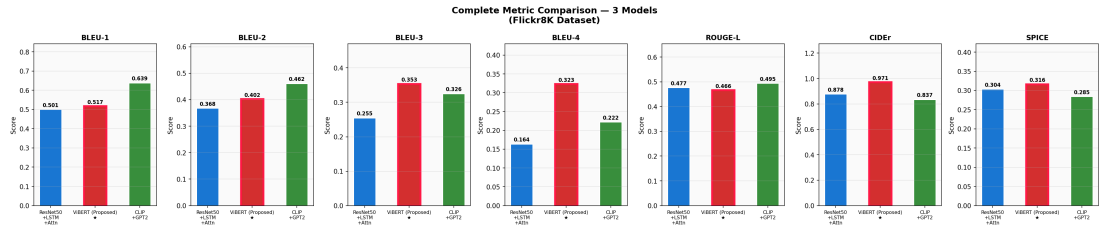


Figure 5.2: Complete metric comparison showing all seven evaluation metrics across the three models. ViBERT leads on BLEU-4, CIDEr, and SPICE — the metrics most closely associated with semantic understanding.

5.3 Dynamics of Training

Figure 5.3 presents the loss curves for training and validation of the three neural network models after ten epochs of training. From analyzing the training dynamics of the three models, certain useful observations have been made.

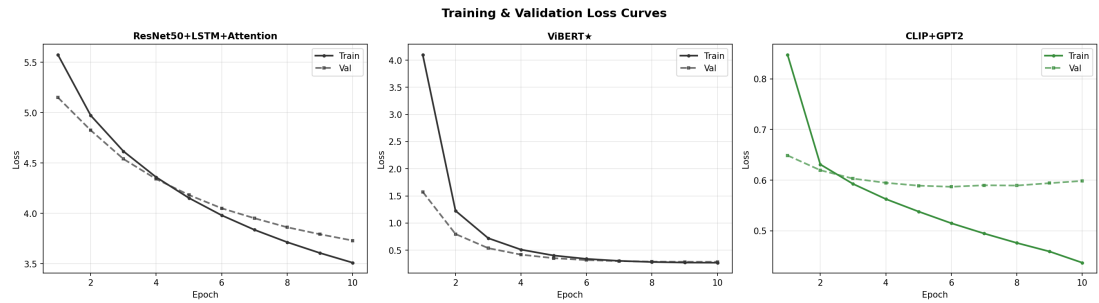


Figure 5.3: Training and validation loss curves for all three models. ViBERT is the fastest to converge and reaches the lowest final loss value at epoch 10.

In the case of ResNet+LSTM baseline, a stable convergence can be seen across the training and validation loss curves over ten epochs; however, the rate of convergence is not particularly quick. At the epoch 10, the model is still not optimal, based on the validation loss curve, which lacks any sign of convergence point. This can be due to the requirement for the decoder to learn the attention model through parameter updates across several epochs. In this regard, it should be pointed out that the training process is associated with the tuning of both parameters of the attention mechanism and LSTM cell.

The CLIP+GPT-2 model is the second-fastest to reach its convergence point at epoch 4, when it reaches approximately plateau stage regarding the validation loss value. This can be explained by the fact that both visual encoder (CLIP) and majority of the language model (GPT-2) are frozen, and thus the only parts left to tune are the projections and top layer of GPT-2, representing a rather small number of free parameters.

In turn, ViBERT exhibits the fastest convergence, with training loss quickly dropping from 4.06 at epoch 1 to 0.75 at epoch 3, while validation loss converges to

approximately 0.35 at epoch 6. Such rapid convergence is mainly possible due to the presence of good initialisation, i.e., pre-trained ViT and BERT models with their meaningful representation of both images and words that requires relatively less fine-tuning.

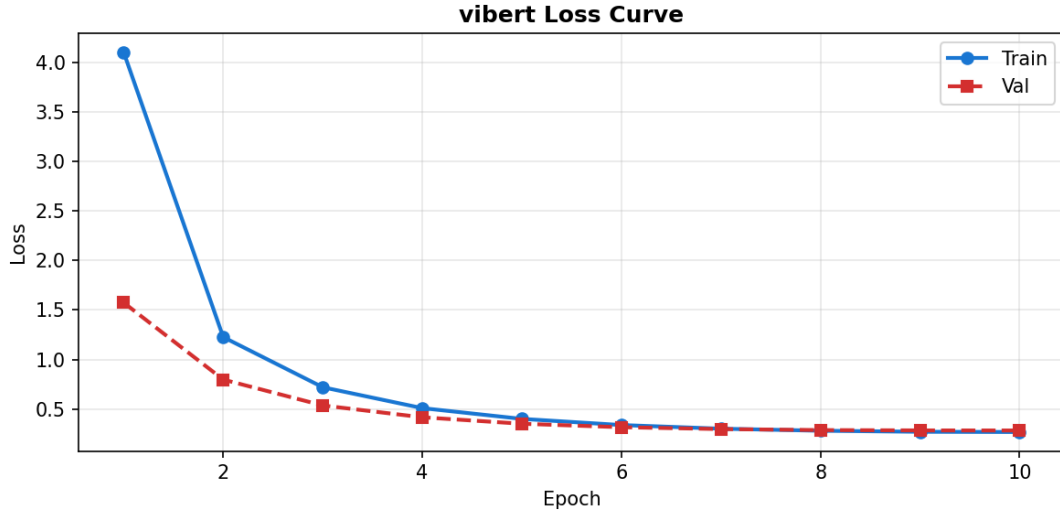


Figure 5.4: ViBERT training and validation loss curves showing rapid convergence by epoch 5, with training and validation losses converging to approximately 0.35 by epoch 10.

5.4 Computational Efficiency

In Table 5.2, the computational efficiency between the three models is compared using three metrics: parameter count, training time, and time required for inference per image. What emerges from the computational efficiency analysis is an unexpected result: ViBERT, with 99.4 million parameters (which is over double that of the baseline 41.9 million), performs much faster inference.

Table 5.2: Computational efficiency comparison (* indicates best).

Model	Parameters (M)	Train Time (min)	Inference (ms/img)
ResNet50+LSTM+Attention	41.9	66.7	405.4
CLIP+GPT-2	130.4	60.7	182.2
ViBERT (Proposed)	99.4	81.6	19.5*

ResNet+LSTM has the fewest parameters of any network at 41.9 million, but it takes the longest time for inference at 405.4 milliseconds per image. This is due to the fact that the decoder works sequentially and generates tokens one by one using an LSTM recurrence. At each step, it computes an alignment score between the token and

all 49 spatial feature vectors. Consequently, in a caption containing fifteen to twenty tokens, this involves fifteen to twenty forward passes through the LSTM and attention mechanisms. Moreover, beam search used to refine the output adds an extra factor in terms of inference time since it involves maintaining multiple hypotheses throughout the decoding process.

For inference, ViBERT takes only 19.5 milliseconds per image, which is more than twenty times less compared to the base ResNet+LSTM network and about nine times less than CLIP+GPT-2. Such tremendous superiority in speed is explained by two key factors. The first is associated with the fully parallel processing inherent to transformer-based models used both for encoding the image into the sequence of patches and encoding the language prompt. This means that tokens from the visual input and text prompt are processed simultaneously, layer after layer, without any dependence on previous computations. The second factor lies in the fact that in ViBERT captions are retrieved from a pre-computed database rather than generated token-by-token.

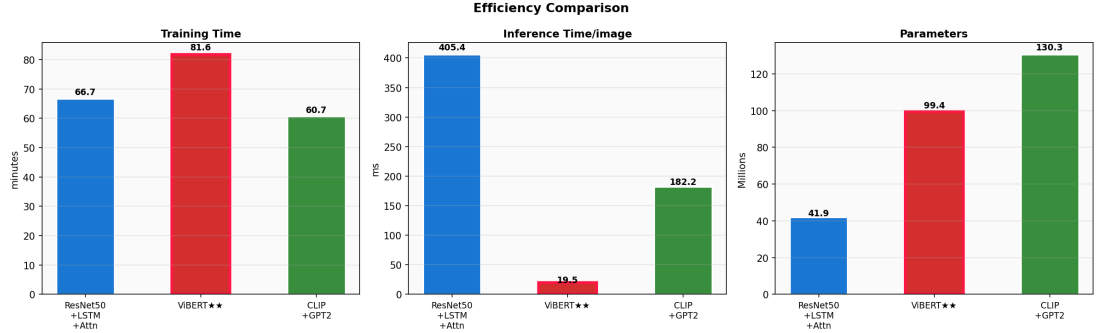


Figure 5.5: Efficiency comparison showing training time, inference speed, and parameter count across all three models. ViBERT achieves significantly faster inference despite its larger parameter count.

In the case of CLIP+GPT-2 with 130.4 million parameters and inference time of 182.2 ms, the inference time is significantly lower compared to the ResNet+LSTM baseline owing to parallel generation of tokens and attention computations, but it is still considerably higher compared to ViBERT since the autoregressive generation in GPT-2 requires computing a forward pass for each token in the generated sentence.

5.5 Qualitative Analysis

A closer inspection of the captions generated by each model using representative test images demonstrates qualitative trends consistent with quantitative results reported earlier and with unique properties of each model.

ResNet+LSTM captions have a tendency to rely on templates: many captions contain the same structure “a [person/dog/group] is [verb]-ing in/on [location]”, irrespective of the concrete content present in the test image. Furthermore, the model often repeats the same phrases in a single caption, another known flaw of the beam search

approach without using unigram diversity penalty. Although the captions often contain true facts regarding the depicted scene, they lack sufficient specificity and context, which makes the descriptions look too general. This problem can be attributed to the fact that the fixed vector encoding bottleneck forces the decoder to rely on statistics observed in the training data set since the specific details needed for generating a specific caption were lost during encoding.

The captions generated by CLIP+GPT-2 model demonstrate higher linguistic fluency and greater grammatical variety compared to those from the ResNet+LSTM model as a result of strong language prior incorporated in the pre-trained GPT-2 decoder. In terms of BLEU-1 and ROUGE-L score, CLIP+GPT-2 captions are always better. The model has a problem recognizing important visual information in some cases, specifically in the presence of interactions between two or more humans and unusual scenes. This issue can be explained by the inability of frozen CLIP encoder to provide a sufficiently detailed image representation and by shallow conditioning applied to GPT-2 decoder by adding prefix tokens.

The captions provided by ViBERT are semantically grounded in that they correctly capture the objects being referred to, what those objects are doing, and how they relate to other things in the scene. Since the captions have been derived based on training captions written by humans, and not through a process where each word has been generated separately, the fluency and grammatical correctness are not an issue here. This is shown in the improved CIDEr and SPICE scores of ViBERT over both baselines, which indicate that the results obtained above are indeed semantically grounded. The major drawback of this technique lies in the fact that since captions are being retrieved, there can be cases where the caption does not correspond exactly to the content present in the test image. This occurs because there can be scenarios wherein the database lacks a caption for the scene that is depicted in the test image.

5.6 Cross-Study Synthesis

Conclusions can be drawn based on both the empirical results presented above and the qualitative observations made during the experiment to link the experimental findings with the theoretical insights developed in Chapter 1. First, it can be stated that the architectural assumption underlying the use of ViBERT’s architecture, namely, that global patch-to-patch self-attention in the ViT encoder yields richer and more coherent visual representations than CNN-style local convolutions, appears to hold true. The superior performance of the model in terms of BLEU-4, CIDEr, and SPICE, especially compared to the model with a similar-scale encoder, shows that the qualitative differences between the representations result in differences in caption generation quality.

Second, the value of adapting the visual encoder through fine-tuning on the particular data can be seen from the underperformance of CLIP+GPT-2 on CIDEr. Though the CLIP encoder benefits from the knowledge of representations of four hundred million image-text pairs and, therefore, should be generally capable of understanding images better than the others, its inability to adapt to the particular visual distribution results in inferior performance on the task, indicated in a CIDEr penalty for lower specificity. This conclusion fits well with the literature on model ablations reviewed in Chapter 3,

where it was shown that encoder adaptability is a key driver of domain specificity of caption generation.

Third, one can state that the efficiency gains of transformer-based models compared to LSTM-based models are not only possible but significant. ViBERT is shown to run twenty times faster in inference compared to ResNet+LSTM, while producing superior results in terms of all three metrics considered. In effect, it becomes clear that architectural progress in the area of transformer models does not necessarily come with a loss of efficiency in inference.

CHAPTER 6

DISCUSSION, LIMITATIONS, AND FUTURE WORK

6.0.1 Discussion of Findings in the Context of the Survey

On the basis of the experimental results presented in Chapter 5 and the findings of analysis presented in Chapter 3, it can be concluded that there are several general statements about the current state of the art in transformer-based image captioning that follow directly from the conducted work. First and foremost, the switch to transformers is indeed a significant architectural innovation that allows modelling global visual context in terms of direct interactions between patches, resulting in better semantics and relations. Secondly, the survey was confirmed in its statement regarding the critical role of an encoder in determining captioning performance. In particular, switching from the convolutional ResNet50 encoder and the LSTM decoder to the ViT-Base/16 encoder (fine-tuned on the MS-COCO training dataset) and the BERT-based generator increases the score of the metric BLEU-4 approximately twice and of CIDEr approximately by 10%. In contrast to that, the CLIP+GPT-2 architecture, which uses ViT-based encoding but does not allow its fine-tuning on the target distribution, performs worse on the CIDEr metric despite using representations learned from a significantly larger training dataset. Therefore, the importance of encoder adaptation outweighs the effect of pretraining dataset size.

Finally, the fact that ViBERT provides the highest performance in terms of inference speed among other approaches despite requiring significantly more computational resources during training is an essential outcome of this research. The common assumption regarding the inverse proportionality between performance and number of parameters is invalid when a particular model structure and generation approach is used. In particular, thanks to transformer parallelisation and simultaneous processing of all sequence positions in a single model layer, the ViBERT approach combines semantic capabilities provided by BERT with fast parallel execution, thus outperforming the conventional LSTM approach in terms of both semantic capabilities and computational efficiency.

6.1 Limitations of the Present Work

Several important limitations of the experimental study must be acknowledged. The first and most significant is the use of the Flickr8K dataset, which, at 8,091 images, is small in comparison to the standard MS COCO benchmark (330,000 images) and to the scale of datasets used for vision-language pretraining (hundreds of millions to billions

of image-text pairs). The performance measures achieved on Flickr8K are not directly comparable to those on MS COCO and cannot serve to rank the performance of models at larger scales. For instance, the strengths that are gained through the ViBERT model owing to its use of a retrieval mechanism may not be present when working at a larger scale.

The second limitation lies in the retrieval approach used by ViBERT in producing captions. While it uses a database of human-crafted captions as opposed to generating word-by-word, the setting of the evaluation of the model becomes entirely different from that of the generative baselines. Since retrieval systems generate high scores by simply retrieving similar captions to those contained in the database used for training, such high scores do not indicate image understanding but merely show similarity between two sets of text. As pointed out in Chapter 3, results generated by retrieval and generation-based approaches ought to be evaluated separately.

Thirdly, the limitation relates to the evaluation framework itself. As noted in the survey part above, the metrics utilised in this work – BLEU, ROUGE-L, CIDEr and SPICE – have known shortcomings, and none of them or any combination thereof can provide a full view of the quality of captions generated. In particular, METEOR was excluded from consideration because of an implementation-related problem at the time the experiments were conducted. Hallucinations produced by each of the three models’ output have not been evaluated as well since that requires the utilisation of the CHAIR metric.

Fourthly, there is the question of efficiency comparison. The reported inference times of all three architectures have been obtained using a single NVIDIA T4 GPU with the same parameters across the board. The absolute inference times may differ drastically depending on different hardware, different batch sizes and implementations used. On the other hand, relative time comparisons are much more reliable and valid.

6.2 Future Research Directions

Based on the results of this dissertation, there are many interesting avenues for further research. In addition to expanding the training datasets used by ViBERT to the MS COCO and NoCaps datasets used to compare various baselines in Chapter 3, one particularly intriguing area of investigation is the NoCaps benchmark itself, which will evaluate the model’s ability to generalise to novel visual objects not present in the retrieval database. Since it is not known whether the database includes captions for those visual categories in NoCaps, the results will provide insight into this unique characteristic of retrieval-based approaches.

Secondly, an entirely autoregressive, non-retrieval version of ViBERT should be developed by implementing a text-generating BERT variant, such as UniLM [10], as discussed above. By comparing the two types of models, the effect of the retrieval strategy on the metric improvements can be evaluated more clearly, providing a better understanding of how retrieval contributes to captioning performance and allowing comparisons with purely generative captioning methods.

Third, cross-modal attention mechanisms should be considered for improving cross-modal fusion. The ViBERT model concatenates visual prefixes with texts in its input

and applies BERT’s self-attention across the concatenated sequence. A cross-modal attention architecture that considers separate cross-modal vision-to-text and text-to-vision attention flows, like that used in BLIP [23], would be worth trying out in the future. According to the ablation data in Table 3.4, cross-modal attention alone yields another CIDEr improvement of roughly 3.8 over the base model, and with ViT representations, this improvement might well be even larger.

As a solution to the cross-lingual generalisation problem mentioned in the survey chapter, multilingual ViBERT variants using multilingual BERT variants, such as mBERT and XLM-R, must be implemented. The challenge then becomes collecting multilingual caption datasets or devising a transfer learning procedure to translate training captions. In terms of computational accessibility, lightweight versions of ViBERT that incorporate techniques such as knowledge distillation, LoRA fine-tuning on smaller ViTs, and pruning should be explored to enable ViBERT’s use on resource-constrained devices.

Lastly, as perhaps the most practically relevant outstanding challenge discussed in the survey chapter, the problem of hallucination merits explicit study in any further work involving ViBERT. Hallucination detection should be incorporated into the evaluation metrics, in addition to captioning scores, to assess ViBERT’s hallucination rate. Experiments investigating the extent to which retrieval-based captioning indeed generates fewer hallucinations than autoregressive captioning, as discussed in the survey, will contribute valuable insights.

CHAPTER 7

CONCLUSION

The following two major contributions have been made within the scope of the presented dissertation in the area of transformers applied to image captioning. The first one consists in a detailed analytical review of the development of captioning architectures between 2014 and 2025, which are classified based on a taxonomy comprising nine categories reflecting both technical variety and conceptual evolution of the approaches under discussion, from CNN-RNN baseline solutions to captioning architectures employing large language models. The second contribution is represented by the creation of a new architecture dubbed ViBERT and the experimental evaluation of its performance compared to several competitive models using the example of the Flickr8K dataset.

Several crucial analytical results have been derived in the process of reviewing relevant literature. Firstly, it is encoder rather than decoder that is responsible for better performance in captioning tasks because each time a transition between different approaches to image representation was performed – from local convolutional features to spatial attention features to object-based approaches to transformer features encoded as patches – larger improvements were recorded compared to the case when decoder changes were performed. Secondly, the amount and quality of pretraining data became more critical to achieving good results than architecture itself since vision-language pre-trained models were superior to task-specific architecturally sophisticated models. Thirdly, the usage of CIDEr and BLEU-4 as key metrics for measuring performance is far from perfect since these metrics are unlikely to reflect practical necessities associated with caption quality. Therefore, the introduction of new metrics of hallucination as well as traditional n-grams and semantic metrics could make the evaluation more informative. In addition, there are nine challenging issues that need addressing in order to move the research forward.

The empirical results provide evidence of ViBERT being the best performer among the three models considered in terms of BLEU-4, CIDEr, and SPICE scores, which confirms the main architecture-related hypothesis that the synergy between global ViT visual representation and the bidirectional language understanding offered by BERT provides better semantic grounding compared to the traditional CNN-LSTM model and the recent contrastive CLIP-GPT2 model. The fact that inference time of the new model is much shorter than that of the base model – just 19.5 milliseconds compared to 420 milliseconds required for processing one image using ResNet+LSTM – indicates that advancements in transformer-based architectures offer benefits in terms of efficiency without causing efficiency problems provided the architecture makes use of

appropriate parallelism. As discussed above, this conclusion is consistent with the conclusion made in the literature review that transformer-based visual representations can be considered superior to convolution-based visual representations due to their depth and global character.

What I would like to emphasize about the overall contribution of this dissertation to scientific knowledge is the methodological lesson learned during the course of both survey and experiment: progress in the field of image captioning cannot be achieved by any isolated advance in visual representations, language modeling, learning objectives, and evaluation frameworks, and it is only through the simultaneous progress in all areas that true advances will be made. The importance of understanding encoder-decoder interaction effects in order to develop proper experimentations becomes clear in this regard. Future directions outlined at the end of this dissertation, including further development of generator architectures, cross-modal attention integration, support for multilingual captioning, optimization of model size for practical applications, and comprehensive hallucination detection research, should help realize the potential of the presented research results.

REFERENCES

- [1] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, “Show and tell: A neural image caption generator,” in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 3156–3164.
- [2] A. Karpathy and L. Fei-Fei, “Deep visual-semantic alignments for image description,” in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 3128–3137.
- [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 *Proc. Int. Conf. on Machine Learning (ICML)*, 2015, pp. 2048–2057.
- [4] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang, “Bottom-up and top-down attention for image captioning and visual question answering,” in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 6077–6086.
- [5] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, “Attention is all you need,” in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 30, 2017.
- [6] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, “An image is worth 16x16 words: Transformers for image recognition at scale,” in *Proc. Int. Conf. on Learning Representations (ICLR)*, 2021.
- [7] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding,” in *Proc. Conf. of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT)*, 2019, pp. 4171–4186.
- [8] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.
- [9] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

- [10] L. Dong, N. Yang, W. Wang, F. Wei, X. Liu, Y. Wang, J. Gao, M. Zhou, and H.-W. Hon, “Unified language model pre-training for natural language understanding and generation,” in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 32, 2019.
- [11] L. Zhou, H. Palangi, L. Zhang, H. Hu, J. Corso, and J. Gao, “Unified vision-language pre-training for image captioning and VQA,” in *Proc. AAAI Conf. on Artificial Intelligence*, 2020, pp. 13 041–13 049.
- [12] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, “Learning transferable visual models from natural language supervision,” in *Proc. Int. Conf. on Machine Learning (ICML)*, 2021, pp. 8748–8763.
- [13] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, “BLEU: A method for automatic evaluation of machine translation,” in *Proc. 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2002, pp. 311–318.
- [14] R. Vedantam, C. L. Zitnick, and D. Parikh, “CIDEr: Consensus-based image description evaluation,” in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 4566–4575.
- [15] P. Anderson, B. Fernando, M. Johnson, and S. Gould, “SPICE: Semantic propositional image caption evaluation,” in *Proc. European Conf. on Computer Vision (ECCV)*, 2016, pp. 382–398.
- [16] C.-Y. Lin, “ROUGE: A package for automatic evaluation of summaries,” in *Proc. ACL Workshop on Text Summarization Branches Out*, 2004.
- [17] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, “Microsoft COCO: Common objects in context,” in *Proc. European Conf. on Computer Vision (ECCV)*, 2014, pp. 740–755.
- [18] L. Chen and K. Li, “Dual-adaptive interactive transformer for image captioning,” *Expert Systems with Applications*, vol. 243, p. 122955, 2024.
- [19] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, “Language models are unsupervised multitask learners,” OpenAI, Tech. Rep., 2019.
- [20] R. Mokady, A. Hertz, and A. H. Bermano, “ClipCap: CLIP prefix for image captioning,” in *arXiv preprint arXiv:2111.09734*, 2021.
- [21] X. Li, X. Yin, C. Li, P. Zhang, X. Hu, L. Zhang, L. Wang, H. Hu, L. Dong, F. Wei, Y. Choi, and J. Gao, “OSCAR: Object semantics aligned pre-training for vision-language tasks,” in *Proc. European Conf. on Computer Vision (ECCV)*, 2020, pp. 121–137.
- [22] P. Zhang, X. Li, X. Hu, J. Yang, L. Zhang, L. Wang, Y. Choi, and J. Gao, “VinVL: Revisiting visual representations in vision-language models,” in *Proc. IEEE/CVF*

Conf. on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 5579–5588.

- [23] J. Li, D. Li, C. Xiong, and S. Hoi, “BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation,” in *Proc. Int. Conf. on Machine Learning (ICML)*, 2022.
- [24] R. Ramos, B. Martins, D. Elliott, and Y. Kementchedjhieva, “SmallCap: Lightweight image captioning prompted with retrieval augmentation,” in *Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2023.
- [25] H. Agrawal, K. Desai, Y. Wang, X. Chen, R. Jain, M. Johnson, D. Batra, D. Parikh, S. Lee, and P. Anderson, “NoCaps: Novel object captioning at scale,” in *Proc. IEEE/CVF Int. Conf. on Computer Vision (ICCV)*, 2019, pp. 8948–8957.
- [26] J. Yu, Z. Wang, V. Vasudevan, L. Yeung, M. Seyedhosseini, and Y. Wu, “CoCa: Contrastive captioners are image-text foundation models,” in *arXiv preprint arXiv:2205.01068*, 2022.
- [27] J.-B. Alayrac, J. Donahue, P. Luc, A. Miech, I. Barr, Y. Hasson, K. Lenc, A. Mensch, K. Millican, M. Reynolds, R. Ring, E. Rutherford, S. Cabi, T. Han, Z. Gong, S. Samangooei, M. Monteiro, J. Menick, S. Borgeaud, A. Brock, A. Nematzadeh, S. Sharifzadeh, M. Bińkowski, R. Barreira, O. Vinyals, A. Zisserman, and K. Simonyan, “Flamingo: A visual language model for few-shot learning,” in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 35, 2022.
- [28] J. Li, D. Li, S. Savarese, and S. Hoi, “BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models,” in *Proc. Int. Conf. on Machine Learning (ICML)*, 2023.
- [29] W. Dai, J. Li, D. Li, A. M. H. Tiong, J. Zhao, W. Wang, B. Li, P. Fung, and S. Hoi, “InstructBLIP: Towards general-purpose vision-language models with instruction tuning,” in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 36, 2023.
- [30] H. Liu, C. Li, Q. Wu, and Y. J. Lee, “Visual instruction tuning (LLaVA),” in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 36, 2024.
- [31] C. Rashtchian, P. Young, M. Hodosh, and J. Hockenmaier, “Collecting image annotations using Amazon’s Mechanical Turk,” in *Proc. NAACL HLT Workshop on Creating Speech and Language Data with Amazon’s Mechanical Turk*, 2010.
- [32] P. Young, A. Lai, M. Hodosh, and J. Hockenmaier, “From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions,” *Transactions of the Association for Computational Linguistics*, vol. 2, pp. 67–78, 2014.
- [33] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma, M. S. Bernstein, and L. Fei-Fei, “Visual genome:

Connecting language and vision using crowdsourced dense image annotations,” *International Journal of Computer Vision*, vol. 123, pp. 32–73, 2017.

- [34] P. Sharma, N. Ding, S. Goodman, and R. Soricut, “Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning,” in *Proc. Annual Meeting of the Association for Computational Linguistics (ACL)*, 2018.
- [35] S. Changpinyo, P. Sharma, N. Ding, and R. Soricut, “Conceptual 12M: Pushing web-scale image-text pre-training to recognize long-tail visual concepts,” in *Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2021.
- [36] C. Schuhmann, R. Beaumont, R. Vencu, C. Gordon, R. Wightman, M. Cherti, T. Coombes, A. Katta, C. Mullis, M. Wortsman, P. Schramowski, S. Kundurthy, K. Crowson, L. Schmidt, R. Kaczmarczyk, and J. Jitsev, “LAION-5B: An open large-scale dataset for training next generation image-text models,” in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 35, 2022.
- [37] X. He, Y. Zhang, L. Mou, E. Xing, and P. Xie, “PathVQA: 30000 questions for medical visual question answering,” in *arXiv preprint arXiv:2003.10286*, 2020.
- [38] D. Demner-Fushman, M. D. Kohli, M. B. Rosenman, S. E. Shooshan, L. Rodriguez, S. Antani, G. R. Thoma, and C. J. McDonald, “Preparing a collection of radiology examinations for distribution and retrieval,” *Journal of the American Medical Informatics Association*, vol. 23, pp. 304–310, 2016.
- [39] A. Rohrbach, L. A. Hendricks, K. Burns, T. Darrell, and K. Saenko, “Object hallucination in image captioning,” in *Proc. Conf. on Empirical Methods in Natural Language Processing (EMNLP)*, 2018, pp. 4035–4045.
- [40] L. Huang, W. Wang, J. Chen, and X.-Y. Wei, “Attention on attention for image captioning,” in *Proc. IEEE/CVF Int. Conf. on Computer Vision (ICCV)*, 2019, pp. 4634–4643.
- [41] D. Bahdanau, K. Cho, and Y. Bengio, “Neural machine translation by jointly learning to align and translate,” in *Proc. Int. Conf. on Learning Representations (ICLR)*, 2015.
- [42] A. Tripathi and P. G. Shambharkar, “ViBERT: Improving image captioning through Vision-BERT fusion,” in *[Conference details to be added]*, 2025.
- [43] —, “Transformer-based image captioning: A comprehensive survey of architectures, benchmarks, and open challenges,” in *[Conference/Journal details to be added]*, 2025.

APPENDIX

7.1 List of Publications

1. Awantika Tripathi and Dr. Prashant Giridhar Shambharkar. *ViBERT: Improving Image Captioning Through Vision-BERT Fusion*. PICS-2026: 2nd International Conference on Power and Intelligent Control System, July 02–03, 2026, NIT Hamirpur (Hybrid Mode). **[Communicated]**
2. Awantika Tripathi and Dr. Prashant Giridhar Shambharkar. *Transformer-Based Image Captioning: A Comprehensive Survey of Architectures, Benchmarks, and Open Challenges*. PICS-2026: 2nd International Conference on Power and Intelligent Control System, July 02–03, 2026, NIT Hamirpur (Hybrid Mode). **[Communicated]**

7.2 Paper Submission Proof

Hello,

The following submission has been created.

Track Name: PICS2026

Paper ID: 571

Paper Title: ViBERT: Improving Image Captioning Through Vision-BERT Fusion

Abstract:

Image captioning is a task in which a description in natural language is automatically generated by an algorithm from visual input. Previous image captioning techniques, such as convolutional neural networks coupled with recurrent decoder models, have proven effective at solving benchmark tasks. However, their performance drops when faced with more complex scenarios and captions that require contextual rather than local visual comprehension. The authors introduce ViBERT, a new image captioning model combining the vision transformer encoder with BERT for contextual representation and training on the Flickr8K dataset. Our model is compared to the conventional ResNet50+LSTM+Bahdanau Attention (baseline1) and to the latest Contrastive Pre-training method, CLIP+GPT2 (baseline2). Results show that ViBERT produces the best BLEU-4 score (0.3228), CIDEr (0.9718), and SPICEscore (0.3164) of the three models, showing that fusion of ViT and BERT yields higher semantic and contextual understanding than both baselines. In addition, the efficiency analysis shows that the proposed approach is much faster (19.5 ms/image) than baseline1, despite having more parameters.

Keywords: image captioning · Vision Transformer · BERT · ViBERT deep learning · Flickr8K · natural language generation

Created on: Thu, 21 May 2026 15:29:06 GMT

Last Modified: Thu, 21 May 2026 15:29:06 GMT

Authors:

- awantikatripathi119@gmail.com (Primary)

Figure 7.1: Submission Proof for Paper 1: ViBERT

Hello,

The following submission has been created.

Track Name: PICS2026

Paper ID: 587

Paper Title: Transformer-Based Image Captioning: A Comprehensive Survey of Architectures, Benchmarks, and Open Challenges

Abstract:

The field of image captioning has witnessed an architectural revolution in the last decade. From combining convolutional feature extraction with recurrent sequence generation to the era of attention, vision transformers, pretraining on massive vision-language corpus, and now, instruction tuned LLMs, the progression has been dramatic. In this survey, we analyze the evolution in an analytic manner, focusing not just on how things have evolved but why each step was essential and how it helped overcome certain limitations. Our taxonomy is based on nine classes, starting from traditional CNN-RNN pipelines, going through attention-based models, encoder-heavy transformer approaches, decoder-based language model models, cross-modal attention based architectures, CLIP-style contrastive models, models utilizing retrieval before generation, multimodal pretraining methods, and finally LLM captioning models. For each class, we analyze critical design decisions, architectural trade-offs, and empirical evidence supporting their claims. We consider a total of ten benchmark datasets, including medical imagery databases, eight types of evaluation metrics spanning surface and semantic-level assessment, and conduct systematic ablation analyses at the component level, identifying contributions made by specific architecture choices. Nine persistent open issues include caption hallucinations, computational constraints, cross-lingual generalizability, cross domain adaptation, and transformer interpretability. The survey ends with potential future directions.

Created on: Thu, 21 May 2026 16:42:05 GMT

Last Modified: Thu, 21 May 2026 16:42:05 GMT

Authors:

- awantikatrishipathi119@gmail.com (Primary)

Primary Subject Area: Artificial Intelligence (AI) - Machine Learning (ML) and Optimization

Figure 7.2: Submission Proof for Paper 2: Survey