Multi-Stage Vision-Language Transformer (MVLT) for Enhanced Image Captioning

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

MASTER OF TECHNOLOGY

in

Computer Science and Engineering

by

Ankita Mishra
(2K23/CSE/10)

Under the Supervision of
Dr. Prashant Giridhar Shambharkar
(Assistant Professor, CSE, DTU)

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

May, 2025

ACKNOWLEDGEMENT

I would like to express my deep appreciation to **Dr. Prashant Giridhar Shambharkar**, Assistant Professor at the Department of Computer Science and Engineering, Delhi Technological University, for his invaluable guidance and unwavering encouragement throughout this research. His vast knowledge, motivation, expertise, and insightful feedback have been instrumental in every aspect of preparing this research plan.

My heartfelt thanks go out to the esteemed faculty members of the Department of Computer Science and Engineering at Delhi Technological University. I extend my gratitude to my colleagues and friends for their unwavering support and encouragement during this challenging journey. I have had some friends that I am thankful to be around. They made me feel truly at home. Their intellectual exchanges, constructive critiques, and camaraderie have enriched my research experience and made it truly fulfilling.

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

Ankita Mishra
(2K23/CSE/10)

DELHI TECHNOLOGICAL UNIVERSITY

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

CANDIDATE DECLARATION

I ANKITA MISHRA (2K23/CSE/10) hereby certify that the work which is being presented in the thesis entitled "**Multi-Stage Vision-Language Transformer (MVLT) for Enhanced Image Captioning"** in 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 in an authentic record of my work carried out during the period from August 2023 to May 2025 under the supervision of **Dr. Prashant Giridhar Shambharkar** .

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

Ankita Mishra

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

Signature of Supervisor(s)

DELHI TECHNOLOGICAL UNIVERSITY

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

CERTIFICATE BY THE SUPERVISOR

Certified that Ankita Mishra (2K23/CSE/10) has carried out their project work presented in this thesis entitled "**Multi-Stage Vision-Language Transformer (MVLT) for Enhanced Image Captioning"** for the award of **Master of Technology** from the Department of Computer Science and Engineering, Delhi Technological University, Delhi under my supervision. The thesis embodies results of original work, and studies are carried out by the student himself and the contents of the thesis do not form the basis for the award of any other degree to the candidate or to anybody else from this or any other University/Institution.

Dr. Prashant Giridhar Shambharkar
Assistant Professor
Department of Computer Science and Engineering,
DTU-Delhi, India

Date:

Multi-Stage Vision-Language Transformer(MVLT) for enhanced Image Captioning

Ankita Mishra

ABSTRACT

Image captioning is an important task in today's world which leverages computer vision and Natural language processing for generating meaningful coherent description of images. Earlier works based on convolutional and recurrent neural networks have shown prominent results but have faced lots of challenges with respect to understanding complicated scenes of images and long-range dependencies. To overcome these challenges, we have proposed a model which is Multi Stage Vision language Transformer (MVLT) which combines state-of-art deep learning architectures for improved image captioning. Our model leverages ViT-G and CLIP for extracting high resolution visual features and Flamingo-style perceiver Resampler for efficient vision-language fusion and LLaVA (Large Language & Vision Assistant) for caption generation with context awareness. Our model has been trained on MS COCO and conceptual captions datasets which is further evaluated on Flickr30k and Visual Genome and has shown promising performance across multiple benchmarks. The proposed MVLT model have achieved a performance that have outperformed previous state-of-art models in BLEU, CIDEr and METEOR scores and have successfully achieves more accurate, relevant, coherent and rich in semantic captions. This work has laid a foundation for advance vision language understanding, with potential application in assistive technology, content creation and AI driven media annotation.

Keyword: Image Captioning, Vision-Language Models, Transformers, ViT-G, CLIP, Flamingo, LLaVA, Deep Learning, Multimodal Learning, Natural Language Processing

TABLE OF CONTENT

LIST OF TABLE(S)

LIST OF FIGURE(S)

LIST OF ABBREVIATION(S)

RNN	Recurrent Neural Network
CNN	Convolutional Neural Network
AI	Artificial Intelligence
ML	Machine Learning
DNN	Deep Neural Network
ReLU	Rectified Linear Unit
GPU	Graphics Processing Unit
IP	Image Processing
IRI	International Roughness Index
DL	Deep Learning
CRNN	Convolutional Recurrent Neural Network
LSTM	Long Short-Term Memory

CHAPTER 1

INTRODUCTION

1.1 Overview

With the arrival of the modern digital age, visual content production and dissemination have never been higher, fueled by mass mobile phone, social media, and surveillance technologies usage. The phenomenal expansion has spurred demand for smarter systems that can learn and understand visual information in a manner that mirrors human understanding. Image captioning automatically creating descriptive captions for images—has emerged as a top solution at the intersection of computer vision and natural language processing (NLP). Historically, these fields developed independently: computer vision focused on recognizing and classifying objects visible to the eye, while NLP was preoccupied with comprehending and generating human language. Nevertheless, due to the growing need for multimodal integrated understanding—such as image captioning generation for images, visually understanding diagrams, and interactive AI agents—researchers started working on models that incorporate vision and language processing. This gave rise to and surged the growth of image captioning as a new and prominent research field in artificial intelligence [1].

With the rampant proliferation of digital images from mobile phones, surveillance cameras, and social media, there exists a growing need for intelligent systems that can process and understand visual data in a semantic context. Image captioning, an example of such a task, is important in that it produces natural language textual descriptions from visual input. This problem is at the intersection of computer vision— task with image interpretation and comprehension—and natural language processing (NLP), the field of text comprehension and text generation. Historically, these two were addressed separately early on in artificial intelligence: vision systems had only been concerned with object identification or scene comprehension, and language models were stand-alone in working through syntax and semantics. However, increasing interest in combined domains, assistive computing, intelligent content arrangement, and visually designed AI assistants—underscored the limitations of operating on such fields as a discrete set. This spurred investigations into models that could generalize across visual and textual data. It is due to this reason that image captioning has emerged as a mainstream research topic, enabling machines to feel and describe the world in less machine-like terms [1], [2].

As AI evolves towards being more human and natural to interact with, image captioning has evolved as an important ingredient in building multimodal AI systems. Such systems can understand and communicate visually and linguistically, and so find themselves at the very heart of applications ranging from virtual personal assistants and intelligent tutoring systems to service robots. Advances in vision-language models over recent years—such as CLIP and DALL·E by OpenAI, and multimodal transformers like Flamingo and LLaVA—have demonstrated the transformative potential of bringing together visual perception with language understanding [3], [4], [5]. This alignment allows AI systems to not only process images by detecting objects but also understand context and generate coherent narratives. Thus, image captioning has developed from a specialized research problem to an empowering technology towards better accessibility for the visually impaired, ease of creation of digital content, and more intuitive human-computer interaction [6].

Before image captioning technology, electronic systems could store and show images but did not have semantic information required to interpret their meanings. Whereas images contain so much visual and contextual information, earlier computer systems were not able to understand them unless they communicated with humans. Machine learning models previously relied on hand-tagging, file name tags, or external metadata to tag and extract visual content—approaches that were not just time-consuming but also very error-prone and subjective [7]. In contrast, however, human beings can infer meaning in an instant from images by utilizing context awareness and language reasoning, thus revealing inherent disparity between human vision and machine vision. This was most clearly seen in use cases where accessibility was needed, including web pages for visually impaired users, where the absence of descriptive captions on the images hindered access to important visual information [8]. Similarly, other areas including surveillance, e-commerce, digital publishing, and medicine—areas processing massive amounts of image data—also struggled to automate content summarization, search, and indexing without semantic image analysis [9]. These challenges drove the evolution of image captioning as an area of study that brings together natural language processing and computer vision to enable machines to see and report on the meaning of an image using natural language, bridging the gap between visual perception and natural language description.

Early attempts at the image captioning task were predominantly typified by rule-based and template-based approaches, in which captions were induced by employing preconceived linguistic templates and hand-crafted heuristics. Although such systems were capable of producing basic descriptions, they were not scalable and not reliable, with a tendency to fail under the diversity and imprecision of real images and natural language usage [10]. These limitations found their application confined to isolated domains and generalization proved hard. Deep learning transformed image captioning methods. Convolutional Neural Networks (CNNs) enabled easy automatic learning from raw images of hierarchical visual features, and Recurrent Neural Networks (RNNs), particularly Long Short-Term Memory (LSTM) networks, brought the tool for capturing sequential patterns and dependences in sequence for natural language generation. A. milestone success. arrived. with. the. "Show. and. Tell" framework, which. combined. CNNs. and. Recurrent. Neural. Networks. within. an. encoder-decoder. framework. to. transform. images. into. text. in. an. end-to-end. process. [. 11.]. This. approach. demonstrated. that. deep. neural. models. were. capable. of. generating. contextually. suitable. and. grammatically. coherent. captions. for. images. Despite this, earlier deep learning models struggled to represent elaborate spatial relationships and generate lengthy, coherent narrative tasks that initiated research on attention mechanisms and transformer-based models.

The history of image captioning was greatly advanced by the introduction of attention mechanisms, where the model was permitted to select dynamically varying regional subsets of an image while creating textual descriptions. This kind of functionality improved contextual alignment between visual features and words generated, resulting in more informative and semantically rich captions [12]. Concurrent with that, the creation of massive scale annotated data sets such as MS COCO, Flickr30k, and Visual Genome—played a key function of offering the scale and variety of training samples required to enhance generalization across diverse visual spaces [13],[14]. The advent of transformer models, initially proposed for natural language processing, also transformed the field. Their ability to handle long-range dependencies and parallel computation made them especially well adapted for multimodal reasoning tasks, such as image captioning [15].

Further expanding its boundaries, CLIP models introduced a contrastive learning paradigm that jointly embedded both image and text representations so that it allowed for strong zero-shot behavior on diverse vision language tasks without fine-tuning for a particular task [16]. Drawing on such innovations, new multimodal models like Flamingo and LLaVA leverage Vision Transformers (ViTs) alongside large-scale language models to generate contextually informative and semantically conditioned captions [17], [18]. The models embrace novel mechanisms, such as the Perceiver Resampler, that enable enhanced vision-language fusion by way of lower-dimensional transforms of high-level vision to token-spectral representations. The employment of these kinds of modular elements facilitates even more precise and accurate image descriptions, providing a new standard for captioning systems for real-world use.

The transition from simple, human-driven image tagging systems to complex, AI-driven captioning platforms has rendered revolutionary changes in industries. Contemporary image captioning technologies play a central role in accessibility since they create real-time descriptive text in order to help visually impaired people to better comprehend visual information. In web media and e-commerce, the technology supports scalable content categorization and suggestion, alleviating the need for human labeling. Security and surveillance uses are complemented by the capacity of captioning systems to summarize and translate visual streams, making decision and monitoring actions possible. The combination of image captioning with multimodal virtual assistants also promises new areas for natural and intuitive human-computer interaction. However, the technologies are limited. Traditional problems like dataset bias, the inherent ambiguity of complicated scenes, image interpretations varying for various observers, and high computational efficiency needs continue to motivate research and innovation in the field [1], [20].

Briefly, the image captioning that is built is on the general trajectory of artificial intelligence—rule-based, hard-coded to dynamic, context-sensitive multimodal models. The models are now trying to imitate human visual comprehension and linguistic capacity, closing the semantic gap between image information and textual meaning. As AI systems move toward human-like design, image captioning is a fundamental technology. It is instrumental in developing accessible digital experiences, automating processes effortlessly, and facilitating more human-like interaction between humans and AIs through its ability to marry visual comprehension and natural language generation.

1.2 Problem Statement

In today's image-oriented internet era, the necessity for well-reading and well-describing images shifted from nice-to-have to must-have. With the torrent of user-created content, images numbering billions are daily uploaded on social networking sites, online shopping malls, keen eyes, and news outlets. All these images bear well-meaning semantic information readable by human beings almost effortlessly. For computers, though, acquiring and conveying this type of information is still a tedious task. Even in the face of recent advancements with artificial intelligence, bringing computer vision and natural language processing to record levels, integrating these two technologies in concord to produce coherent and human-like descriptions of visual information, also a process known as image captioning, is still a significant challenge with key practical implications [1].

Image captioning is the computer generation of text descriptions of what is in an image. It is more than object recognition; it is recognizing spatial relationships, context, and the larger scene. It's where two of the foundation pillars of artificial intelligence meet: computer vision and natural language generation. While existing image processing methods like object detection are able to detect and tag objects within a picture, they are generally unable to comprehend the relationships between objects or convey the complete meaning of the scene accurately. This is different from how language generation models are able to generate semantically correct sentences but do not possess the ability to image-specific text without visual grounding. To overcome this, one does not just require integration of visual and linguistic data, but higher-order reasoning and semantic equivalence—necessities for constructing intelligent systems which can act in the world in a way that mimics human behavior.

Improving accessibility in the digital space is among the key reasons for the creation of image captioning technology. Internet for the visually impaired and the blind—rich with information as it is—is actually composed of picture content. Browsing social media, reading news headlines, or even browsing online shopping websites might be highly dependent on image interpretation, graph, infographics, and product images. Without alternative text or descriptive captions, this information is largely inaccessible. Captioning of images provides a solution in generating complete, context-rich descriptions automatically that can be utilized by screen readers to convey image information in an informative manner. Not only is it a wonderful technological advance, but it also encourages inclusiveness and digital equity in accordance with such guidelines as the Web Content Accessibility Guidelines (WCAG) [21].

Along with improving accessibility, the increasing number of digital images necessitates scalable content retrieval and management systems. With organizations building vast repositories of images, manual addition of descriptive tags and metadata takes time and money. Automated captioning enhances the process by enabling more effective indexing and search, thus providing a superior quality digital asset management experience. For instance, in e-commerce, the automatically generated product image captions can assist with search engine optimization (SEO), support recommendation algorithms, and supply more context to search results. The captions serve the same purpose for high-speed applications such as social media and news media by assisting with content generation in real time and supporting moderation processes by identifying or marking inappropriate content [22].

Surveillance and security is yet another important domain where image captioning is of extreme value. Governments and private establishments use large networks of CCTV cameras and visual sensors to monitor the surroundings and offer security. While these systems gather enormous amounts of video data, human interpreters are seen struggling with real-time interpretation owing to the vastness and complexity. Image captioning technology can be useful by producing short text summaries of video scenes or frames and thus enable quicker detection of suspicious behaviors, summarizing proceedings in progress, and marking possible threats that require addressing by humans. All these advantages improve the efficiency of operations and increase the effectiveness of surveillance [23].

The rising popularity of multimodal AI systems and HCI has increased the need for image captioning. Modern technologies such as learning software, customer service robots, and virtual assistants are becoming capable of interpreting not only written or spoken words but also images. Virtual computers that can analyze an image, provide a description of what is contained in it, or answer questions about it enhance natural and fluent human-computer interaction. Captioning images in educational environments supports learning through describing intricate images like charts, paintings, or diagrams so that they may be comprehended by a large body of learners.

Early similar images detection methods relied mostly on similar template detection or similar rule detection models, which were inflexible and Earlier detection models especially had a poor tendency to generalize well over a large variety of image types. Integration of deep learning approaches greatly improved using specifically Convolutional Neural Networks (CNNs) for efficient visual feature extraction and Recurrent Neural Networks (RNNs) for creating similar descriptive text sequences. This model was subsequently improved upon using attention mechanisms, where the model could attend to the most critical areas within an image, enhancing the quality and precision of the captions generated. However, even with these improvements, these methods were still lacking in how they could capture the depth of scenes and generate coherent, contextually apt captions for longer descriptions [10],[11].

The most recent breakthroughs in transformer-based and multimodal learning models have greatly improved the ability of image captioning models. Specifically, models like CLIP (Contrastive Language-Image Pretraining) and LLaVA (Large Language and Vision Assistant) leverage big scale databases of aligned image-text data to learn common representations, connecting textual and visual data. CLIP, for example, injects images and text captions into a common embedding memory such that it supports effective zero-shot classification and retrieval after training on particular tasks [24]. Other models, such as Flamingo, a multimodal transformer trained at DeepMind, employ dynamic cross-modal context integration that supports improved reasoning and coherence in generated captions across long textual horizons [25]. These developments represent an enormous stride in the direction of creating descriptions that are not merely pertinent but semantic and contextually dense.

But long-standing problems slow down its general adoption. Visual scenes typically entail vagueness and context-sensitive semantics, which vary across individuals in terms of differences in cultural experience, affective perception, and purpose. Creating systems to represent this variation in a similar way that humans do is a difficult problem. Additionally, train data bias, computationally intensive requirements, and real-time computation needs remain major obstacles that must be met by researchers for having robust, scalable deployments [26], [27].

In short, the connection to image captioning stemmed from the universal discrepancy between the way people and computers understand the world. As digital material grows and proliferates, and artificial intelligence technologies are profoundly integrated into daily life, the ability to create short, useful, and contextual image captions will become more desirable. It makes tech more accessible, facilitates automation at scale, and brings machines one step C

Closer to human-level AI. Conquering the hurdles of image captioning not only propels artificial intelligence development but also releases transformative power in healthcare, education, accessibility, security, and digital

CHAPTER 2

DEEP LEARNING

Machine Learning (ML), which is a subfield of Artificial Intelligence (AI), is focused on designing algorithms to allow systems to recognize patterns and make data-driven decisions without being explicitly programmed by rules. Early ML employed techniques such as decision trees, support vector machines, and k-nearest neighbors. These algorithms worked nicely with structured data but relied heavily on features hand-designed by humans and did not work well when applied to more complex forms of data like images, speech, or text. Feature design in such applications took significant amounts of domain knowledge and large amounts of manual labor, which often restricted scalability and performance.
To counter the limits of these, the discipline turned towards deep learning, enabling a subcategory of ML involving the application of multilayered neural networks with the use of raw inputs to learn automatically. From this turn, models learned abstract representations at different levels of abstraction independently, significantly lowering human feature engineering efforts. The revival of deep learning towards the close of the 2000s was fueled by a mix of interacting factors: availability of large amounts of labeled data, the development of high-end graphics processing units (GPUs), and progress in optimization algorithms, namely the backpropagation algorithm.
The above-named great leap forward in deep learning was the advent of Convolutional Neural Networks that indeed changed the way to deal with visual data by learning very sophisticated spatial patterns and configurations. CNNs became the horse to do all types of vision. For instance, object detection, image classification, and understanding scenes. Meanwhile, Recurrent Neural Networks (RNNs) – in the form of Long Short-Term Memory (LSTM) units – proved unmatched might when applied to sequential data. Such models greatly enabled natural language applications with machine translation support, text prediction, and speech processing.

Transformer-based architecture, in the recent past, has improved the performance of computer vision (CV) and natural language processing with self-attention mechanisms. Compared to traditional recurrent models, transformers have a greater capability to model long-range dependencies and contextual relations over sequences and, hence, possess a more comprehensive understanding of data. Initially proposed for language processing, the transformers were later successfully adapted to vision tasks—most famously in models like the Vision Transformer (ViT), which tokenize image patches into sequences to allow for efficient parallelization and scalability. Models like CLIP (Contrastive Language–Image Pretraining) have also demonstrated the ability to learn visual and text representations simultaneously. By projecting images and words into a shared embedding space, CLIP facilitates semantic consistency across modalities with the capacity to accommodate an array of vision-language tasks without requiring task-specific training. This departs from conventional machine learning to deep learning and has led to models that can interpret and generate natural language from images with much higher accuracy.

But in spite of these developments, most of the earlier image captioning models based on convolutional neural networks (CNNs) and recurrent neural networks (RNNs) are still difficult. These are:
complexity in representing complex scenes with numerous objects, understanding fine-grained contextual knowledge, and generating semantically rich and coherent captions. This is a sign that there is a need for more context-aware and holistic solutions—like those made possible by cutting-edge transformer-based models—overcoming limitations of past approaches.

The suggested Multi-Stage Vision-Language Transformer (MVLT) improves image captioning thus far by integrating several state-of-the-art deep learning components into a single miniature architecture. MVLT leverages the merits of current vision and language models to provide a holistic description of visual content. In particular, MVLT employs ViT-G and CLIP to acquire high-resolution visual features so that contextual and fine-grained information is well retained. To facilitate the fusion of visual and textual knowledge, MVLT uses a Perceiver Resampler inspired by Flamingo architecture to facilitate efficient and scalable multimodal fusion. Last but not least, it uses the LLaVA language model to facilitate context-responsive and fluent caption generation to enable syntactically well-formed and semantically accurate generation of descriptions. With this multi-step pipeline, MVLT is a dramatic shift from vision-language integration. It transcends material limitations of previous models, such as not being able to process complicated scenes with numerous objects or fine contextual relationships. More importantly, though, MVLT demonstrates the continued advances in deep learning by combining potent transformer-based architectures to attain an extent of image perception and language production that brings machines a step further toward understanding vision with human subtlety.

2.1 History

While computer vision and natural language processing were not so advanced at that time, it was challenging to generate captioning text directly from images, largely because of a lack of computing capabilities and inadequately developed algorithmic methods. The early solutions involved rule-based and manually designed image features where domain experts would manually define visual descriptors such as edges, contours, and textures. These would then be plugged into pre-specified sentence templates or heuristic rules to create simple captions. But such systems were inflexible and non-adaptable—skeletal to highly specialized instances and unable to generalize to variable visual input.
The larger domain of artificial intelligence (AI) began transforming this area from the mid-20th century, beginning with symbolic AI approaches. These were based on logic-driven reasoning, and formalized knowledge representations in various forms. Redundant for rule-based issue-resolution, symbolic approaches had far more trouble addressing perception tasks such as image understanding of the visual world, where input data are noisy and unpredictable by nature.
The turning point occurred with the advent of machine learning, in which models learned from data instead of relying on hand-coded rules. Decision trees, k-nearest neighbors, and support vector machines were the standard approaches during this period. These models were stronger and more versatile but were plagued by the requirement for hand-designed input features—i.e., the good representation bottleneck still existed.

The success of convolutional neural networks especially transformed the field with the triumph in large-scale image classification tasks such as the ImageNet competition in 2012, making CNNs learn directly from pixel-level data, thus omitting the necessity of applying handcrafted features and permitting much higher accuracy in object and scene recognition. At the same time, breakthroughs in natural language processing—initially through recurrent neural networks (RNNs) and later through attention-based transformer architectures—enhanced the ability of machines to understand and generate human-like language.
The convergence of these advances gave rise to the first image captioning models that used CNNs for learning visual features and RNNs or LSTM networks for generating text descriptions. Notables such as the "Show and Tell" architecture had early success in closing the visual and linguistic modalities. Yet, the models struggled in comprehending intricate scenes involving many objects or encoding subtle relationships among objects because of the weakness in encoding long-range dependencies and contextual consistency.
The space has since progressed at a revolutionary rate, especially with the emergence of transformer models and pretraining at scale. Architectures like CLIP (Contrastive Language-Image Pretraining) and Flamingo have set new benchmarks by learning joint vision-language embeddings on multimodal large datasets. Such models can connect images and text in a common semantic space and enable smoother and context-dependent image captioning.

Assisting on such shoulders, the multi-stage vision-language transformer (MVLT) in the suggestion integrates various state-of-the-art elements to further enhance image captioning. MVLT integrates ViT-G with CLIP to access high-resolution rich visual perception, a Flamingo-based Perceiver Resampler to blend multimodal features in a seamless manner, and LLaVA to generate contextually rich language. MVLT illustrates a unified pipeline that describes the richness of visual scenes holistically. This model builds upon other CNN-RNN based models with transformer models having deeper semantic understanding, better contextual matching, and more natural output generation—getting the field closer to human-level image describing capability.

2.2 Machine Learning

Since deep learning (DL) is a specialized subdomain within the broader field of machine learning (ML), it is needed to first understand the foundational concepts of ML. Machine learning has been defined and applied differently across various disciplines due to its versatility in solving a wide range of problems. The term "machine learning" was first introduced by Arthur Samuel in 1959 [26], referring to the capability of computer systems to perform tasks by learning from data and experience without being explicitly programmed. Essentially, ML enables the extraction of patterns from data, particularly in scenarios where explicit analytical solutions are not feasible. In such cases, machine learning offers methodologies for identifying hidden structures or trends from the data [27].

Machine learning tis typically categorised into three major types: supervised learning, unsupervised learning, and reinforcement learning, as illustrated in Figure 2.1. In supervised learning, models are trained on a labeled dataset, where each input sample is associated with a known output. For instance, in the context of object detection in images, the training data comprises annotated images labeled to indicate the presence or absence of a particular object. The model then learns to generalize from these labeled examples to make predictions on new, unknown data [27].

On the other hand, unsupervised learning is applied when labeled outputs are not available. The goal in this case is to unreveal hidden patterns or within the data. Methods like clustering and dimensionality reduction are classified as unsupervised learning techniques, where the goal is to uncover hidden patterns or simplify data representation without labeled outputs. In contrast, reinforcement learning represents a distinct paradigm where an agent learns by interacting with an environment, receiving feedback through rewards or penalties based on its actions. Over time, the agent refines its strategy to achieve the highest possible cumulative reward. This trial-and-error learning method is especially useful in fields such as robotics, gaming, and autonomous systems, where adaptive decision-making is essential in unpredictable or evolving scenarios.

2.3 What is learning?

The traditional frameworks are used to explain the aspects of learning algorithms and for learning to be considered as feasible, provide mathematical proof of this fact— Shai Shalev-Shwartz, Shai Ben-David [18] presented examples that could help in understanding how basic learning process work alongside what have been identified as principal challenges within machine learning (ML). Rats learn how

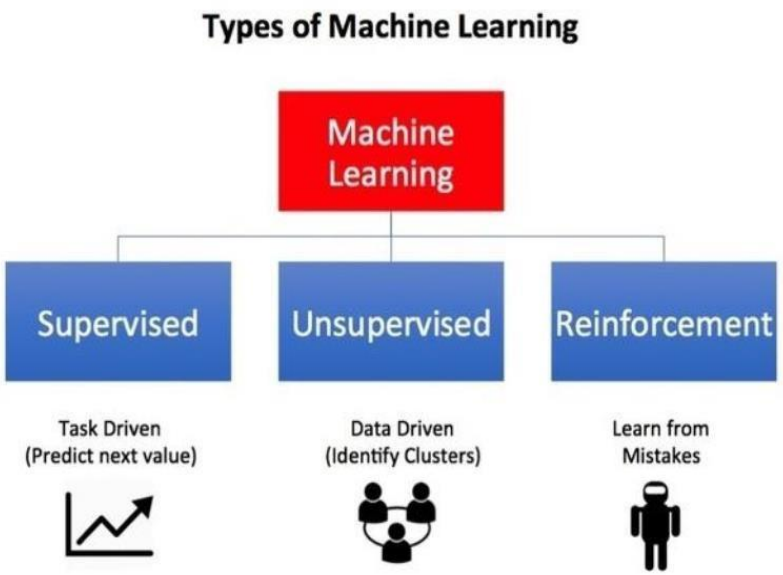


Figure 2.1 . Different ML problem categories [17]

to avoid poisoned food starting from their childhood. Rats usually take a small amount of new food first and are careful to investigate the physical consequences. If the food causes sickness, they never eat it forever. The experiment involved an animal in search of a harmless meal. In this case, the animal would expect that if it experienced a negative label then it would also develop negatively. Assume we are attempting to write a spam detector program. For instance, one straightforward way is to remember every email determined to be spam by a user. When an incoming email is received, it is verified against the spam set. If it is found in the spam set, then it is marked as a spam message; else, it is saed in the inbox folder. Memorization is occasionally helpful, but it does not have much in common with learning because it cannot be generalized. An intelligent learner who truly understood should be able to extract wider generalizations from diverse instances. It therefore means that generalizing constitutes the ultimate definition of intelligence. When compared with other creatures, man's special gift is his ability to think and understand concepts widely, putting us one step ahead. For instance, given a realistic picture of an elephant, a child might be able to recognize a drawn elephant that looks very different (Figure 2.2). Another problem is when the learner comes to a wrong conclusion. In explaining this notion, Skinner's superstition experiments are the most useful example. To be precise, Skinner put some hungry pigeons in a box that came with an automatic device meant to supply food for the hen occasionally with no consideration given to its actions. He found that pigeons would exhibit behaviours signalling expectancy only during feeding time and for more or less two minutes after that. While waiting for food, a particular bird spun round and round in a counter clockwise direction before making one or two turns in the opposite direction before it was rewarded. But there were sometimes when it was fed by Andy and would peck continuously at the upper edge of its basin."'A bird thrust its head out and swung it sharply rightwards from leftwards then back again with some slowness so as to make it like a pendulum while another bird began shaping up like it was making quotations (this means they stuck their heads beneath an unseen pole raised them up multiple times'[19].

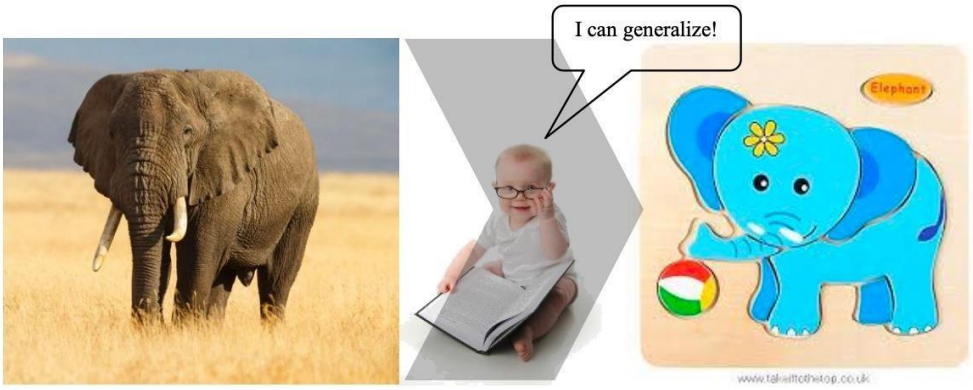


Figure 2.2. Concept of generalisation and intelligence

When humans learn, they use their common sense and ignore random patterns or conclusions from learning that are meaningless, but machines do not. A machine requires well defined principles to steer it out of arriving at irrelevant conclusions. In simpler terms, the algorithm should be able to discern a pattern in the data but not in the noise.

2.4 Deep Neural Network (DNN)

A Deep Neural Network (DNN) is a sophisticated computational architecture modeled after the structure and functionality of the human brain, designed to process and learn from raw data through multiple layers of artificial neurons. Each neuron performs a transformation by applying a weighted sum to its inputs, folloed by a non-linear activation function, enabling network to represent complex and highly linear patterns in the input space. The standard DNN architecture begins with an input layer that captures raw data, followed by a sequence of hidden layers where each layer extracts increasingly abstract features, and culminates in an output layer that provides the final prediction, such as a classification label or generated text [28], [29].

Several specialized DNN architectures have been developed to address different types of data. Convolutional Neural Networks (CNNs) are particularly effective in visual tasks as they use convolutional filters to detect spatial hierarchies in images, enabling efficient learning of local and global features [30]. In contrast, Transformer-based architectures, originally introduced for natural language processing, utilize self-attention mechanisms to capture long-range dependencies in data and have since been successfully adapted to visual and multimodal tasks [31].

In the proposed Multi-Stage Vision-Language Transformer (MVLT) model, we integrate several cutting-edge deep learning components: ViT-G for vision-based feature extraction, CLIP for learning joint embeddings between image and text, and LLaVA for language generation. These modules collectively form a unified architecture that benefits from deep neural networks' ability to perform hierarchical representation learning, thereby bridging the gap between visual understanding and natural language generation [32]–[34].

A deep neural network (DNN) is a network of successive layers of layers, and each of them transforms its input by a linear transformation followed by application of a nonlinear activation function. The organization of layers causes the network to learn step-by-step features of the data at higher levels. Algebraically, the lth layer's transformation can be represented as:

$$h^(l) = \sigma(W^(l) h^(l-1) + b^(l)) \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (1)$$

Where,

H^l is the output activation of layer l,
W^l - weight matrix for the layer l,
b^l - bias vector for layer l,
sigma - non linear activation function,
$h^{(l-1)}$ - input from the previous layer

A deep neural network (DNN) develops a mapping of functional input data to related output in a process referred to as forward propagation. In this process, input data passes through a series of cascaded layers, each of which applies a linear transformation followed by a nonlinear activation function and thus progressively learns and derives abstractions higher up from data. The model then produces one output, which is compared against the actual tdesired value using a loss function—commonly the cross-entropy loss in classification tasks—to quantify the prediction error. To minimize this error, the backpropagation algorithm computes the loss gradient with respect to every parameter in the network. These gradients inform the updates made to the model parameters (weights and biases) using optimization algorithms such as stochastic gradient descent (SGD) or its variants like Adam. Through repeated iterations over the training dataset, the network gradually refines its internal parameters, enabling it to hold complex, nonlinear patterns in the data and improve its predictive working.

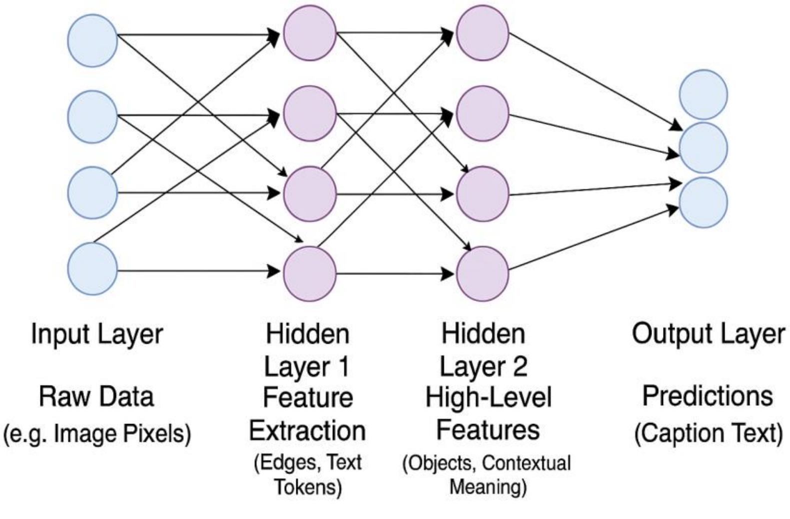


Fig 2.3 - Neural Network

2.4.1 Historical Background

The evolution of deep learning began with the conceptualization of artificial neural networks during the mid-20th century. One of the earliest models, the perceptron—introduced in the late 1950s—served as an initial attempt to simulate the decision-making process of a human neuron. Although its capabilities were limited to linearly separable data, it laid the groundwork for more complex models. Interest in neural networks revived during the 1980s following the development of the back propagation method, which allowed efficient coaching of multi-layer networks by computing gradients and updating weights accordingly. Despite these advancements, deep learning remained relatively niche until a significant turning point in 2012, when the AlexNet model achieved unprecedented accuracy in the ImageNet Large Scale Visual Recognition Challenge. By leveraging the capability of the deep convolutional layers, huge labeled data, and the fast GPU, AlexNet demonstrated how deeply deep learning can be applied to visual tasks. More recently, following architectures such as Long Short-Term Memory (LSTM) networks have bridged the gap in modeling sequential data, and the development of Transformer models has made scalable learning applicable in the text as well as visual domain. Together, these breakthroughs have solidified deep learning as one of the foundational pillars of artificial intelligence today.

2.4.2 Fundamental Concept of Deep Learning

In essence, deep learning models are made up of several layers of interconnected units known as artificial neurons. The units receive numerical input, calculate a weighted sum, a non-linear activation function, and output the transformed value to the next layer. The most common activation functions used are the Rectified Linear Unit (ReLU), sigmoid, and hyperbolic tangent (tanh), and they influence how the model processes and outputs information. To train such networks is to optimize a loss function, one that captures the difference between model predictions and actual target variables. Optimization algorithms like Stochastic Gradient Descent (SGD) and Adam are some of the most well-known such algorithms for incrementally updating model parameters—i.e., weights and biases—so as to minimize the loss. To prevent overfitting and ensure generalization, various methods of regularization are employed, such as dropout, where neurons are turned off intermittently during training, and batch normalization, which stabilizes learning through the process of normalizing activations. Understanding these fundamentals is essential for testing and designing deep neural networks in an efficient manner.

2.4.3 Architecture and Components

A standard deep neural network has three layers: input layer, several hidden layers, and output layer. All the layers have a number of neurons (also referred to as nodes), which are linked with each other the next layers. Every connection has an associated weight, and each neuron has a bias term. The core operation within each neuron is a linear transformation followed by a non-linear activation function. Mathematically, the output of a neuron y is given by:

$y = f \vdash(\sum\_\{i = 1\}\^\{n\} w\_i \, x\_i + b \dashv)$……..(3)

Where x_i are y are input features, w_i are the weights, b is the bias and f is the activation f function such as ReLU, Sigmoid, or Tanh.

2.4.4 Forward Propagation

In forward propagation, input data is passed through the network at each layer, and each layer applies its transformations to the data. The output from one layer become inputs to the next. This hierarchical processing allows the model to learn increasingly abstract representations.

2.4.5 Loss Function

Once the network produces an output, it is compared with the ground truth using a loss function. The loss quantifies the prediction errorShared loss are Mean Squared Error (MSE) for regression problems, Cross-Entropy Loss for classification problems. In generative problems such as image captioning, sequence-based loss functions such as Negative Log-Likelihood or specialized metrics such as BLEU can be employed during training.

2.4.6 Backpropagation and Gradient Descent

To minimize prediction error, neural networks utilize backpropagation, an algorithm that computes the gradient of the loss function with respect to all model parameters using the chain rule. The gradients determine the manner in which the weights must be adjusted to optimize performance. An optimization algorithm—traditionally Stochastic Gradient Descent (SGD) or more recent algorithms like Adam or RMSprop—is utilized to update the weights accordingly. This training cycle is then iterated across many iterations, or epochs, so that the network learns step-wise and approaches an optimum solution.

2.4.7 Regularization Techniques

In order to decrease overfitting and enhance generalization, regularization methods are applied. They encompass dropout ,L1/L2 regularization (introducing penalties to the loss), and batch normalization (normalizing layer inputs in order to stabilize training). They allow the model to learn stable patterns and avoid it from memorizing the training set.

CHAPTER 3

LITERATURE REVIEW

3.1 Image Captioning

The two-stage process that employs computer vision and natural language processing methods is prevalent in the application of image captioning. With the use of complex models such as Convolutional Neural Networks (CNNs) or Vision Transformers (ViTs), visual feature extraction occurs in the initial stage. In the first step, input images are passed through the processing of these models for generating a dense, high-dimensional vector representation of all vital visual information such as object occurrence, scene context, and spatial arrangement. This encoded form provides the foundation for generating a natural language description in the next stage. Recurrent neural networks (RNNs) and their more powerful variant, Long Short-Term Memory (LSTM) networks, are two such sequence modeling frameworks that are commonly used to generate captions by generating words step by step conditioned on the visual information. Second, the process often involves attention mechanisms, through which the model can dynamically select different parts of the image at each stage of word prediction. The semantic consistency and descriptive appropriateness of the generated captions are significantly enhanced with this focused attention.

Transformer models are used in recent image captioning progress for the prediction of text sequences as well as for extracting visual features, facilitating cross-modal interaction. The next word is predicted iteratively with the help of the encoded image representation and sequence of the previously generated tokens by the decoder until predicting an end-of-sequence token. Huge annotated image datasets that are presented along with relevant textual descriptions are used while training these models. These employ reinforcement learning methods for learning on evaluation metrics such as CIDEr or BLEU, and others employ objective functions such as cross-entropy loss to supervise learning. Smooth, semantically diverse, and contextually meaningful captions result as a byproduct of the model learning about contextual correspondence between vision and language within this single, end-to-end training system.

3.2 Traditional Methods for Image Captioning

From simple rule-based models to extremely advanced deep learning models, the landscape of image captioning has witnessed humongous transformations. A majority of the initial approaches were template-based, creating captions by filling in given actions and objects within given linguistic templates. Although ensuring grammatical correctness, the systems were not adaptive and context-sensitive. Subsequently, retrieval-based systems were introduced that selected pre-written captions from a database of images having similar visuals. However, such models' ability to represent new scenes was limited. Later approaches merged sentence templates and object detection, which added more information but still failed to model rich semantic and spatial relationships.

The application of statistical techniques that offered probabilistic models of language, like Statistical Machine Translation (SMT) and Hidden Markov Models (HMMs), yielded a significant influence. But these were saddled with the need for enormous paired datasets and with not being able to handle long-term dependencies. With encoder-decoder models using Convolutional Neural Networks (CNNs) to obtain visual features and Recurrent Neural Networks (RNNs) to generate the words, deep learning revolutionized the field and allowed for end-to-end training and greater fluency. By directly pointing out salient parts of an image during runtime while captioning, the use of attention mechanisms further enhanced such systems. Further more recently, multimodal pretraining methods and transformer models have pushed performance limits by enabling more expressive and more stable correspondence between visual inputs and text outputs.

3.2.1 Template-Based Captioning

One of the first automatic image captioning approaches was template-based image captioning. In this method, computer vision techniques, typically employing traditional classifiers or object detection pipelines, are employed for identifying visual units such as objects, attributes, and sometimes actions. The identified units are then connected to pre-specified syntactic templates to form complete sentences. A standard template might be the type "A [object] is [action] in the [location]," and it could generate something like "A cat is sleeping on the sofa." Readability and simple grammatical organization are assured, but semantic scope, context comprehension, and linguistic flexibility are not. This is the reason captions are often too formulaic to feel and cannot be able to bear subtlety or complex relationship between objects in a scene. Despite these limitations, template-based methods gave a useful starting point for automated captioning research. The basic role of this method in early vision-language models was brought out by Kulkarni et al., who showed that a model based on object, attribute, and spatial relationship detection supplemented with templated frames was able to produce simple but grammatical captions [35].

3.2.2 Retrieval-Based Captioning

Retrieval-based image captioning serves as an early and efficient strategy for automatic image description by leveraging existing image-caption datasets. Instead of generating new textual content, the system identifies visually similar images from a precompiled dataset and assigns the caption of the nearest match to the query image. Initially, similarity was measured using hand-crafted global image features such as GIST descriptors. However, the adoption of deep learning techniques—especially Convolutional Neural Networks (CNNs)—significantly improved feature extraction and retrieval precision. For instance, given an image of a sunset over a mountain range, the model may assign a caption like “Sunset casting golden light over the mountain peaks,” sourced from a visually similar image in the training set. While this technique often ensures fluent and contextually appropriate captions for common scenes, its primary limitation lies in its lack of generalization. The system is inherently constrained by the diversity and representativeness of its dataset and typically fails to describe novel scenes, rare object configurations, or abstract visual content. A notable implementation of this approach is the Im2Text system proposed by Ordonez et al., which demonstrated the practicality of nearest-neighbor methods in generating natural-sounding captions but also revealed their shortcomings in terms of flexibility and semantic depth [36].

3.2.3 Object Detection + Template Filling

Object Detection with Template Filling emerged as an intermediate step between early rule-based captioning and fully generative models. This technique integrates computer vision algorithms to automatically detect visual components within an image—such as objects, their attributes, and associated actions—and maps them onto predefined sentence structures. This method employs visual recognition approaches such as Scale-Invariant Feature Transform (SIFT), Deformable Part Models (DPM), or other region-based classifiers to identify entities within the image, in contrast to static rule-based approaches. Semantically consistent captions are generated by inserting these parts dynamically into syntactic templates. The machine can generate the sentence "A cat is sitting on a chair" if there is an image with a cat sitting on a chair, for example. By allowing sentence adaptation from actual image content, this system is less restricted than strictly manual systems. The two significant restrictions, however, decrease its utility: dependence on the rigidity of pre-specified templates and object detection subsystem quality. While the templates themselves restrict linguistic variation and seldom can fairly reproduce complex relationships for more than one object, misdetected objects can generate misleading or incorrect captions. Farhadi et al. contributed one of the most well-known early advances in this area when they proposed to map visual tuples, such as object, action, and attribute, to sentence descriptions. Even though their system pinpointed the difficulty of encoding highly context-dependent or abstract information, it proved that language generation and structured visual recognition could be integrated [37].

3.2.4 Hidden Markov Models (HMMs)

Hidden Markov Models (HMMs) represent one of the earliest probabilistic frameworks applied to the image captioning domain, where caption generation is treated as a sequential word prediction problem. In this framework, an image is typically used to influence the initial configuration of the hidden states in a Markov process. Each word in a generated sentence corresponds to an output emitted from a hidden state, which in turn transitions probabilistically to other states, forming a sequence. These hidden states encapsulate underlying linguistic or semantic roles, while emission probabilities define the likelihood of generating specific words from each state. For example, given an image of a person crossing the street, an HMM may sequentially produce a sentence such as “A person is walking” by transitioning through grammatically structured hidden states. However, the reliance on manually engineered image features and discrete latent state spaces introduces significant limitations. These include difficulty in capturing long-range syntactic dependencies, modeling abstract relationships, and handling diverse vocabulary. Furthermore, expressiveness is restrained by the rigid number of hidden states. In an attempt to facilitate automatic captioning and annotation, the pioneering work by Feng and Lapata explored the combination of HMMs and topic models. Although indicating HMMs' inability to provide well-detailed and contextualized descriptions, their research attested to the statistical strength of such models [38].

3.2.5 Statistical Machine Translation

Text description generation is framed as a translation problem by statistical machine translation (SMT) methods in image captioning. This is equivalent to taking visual items found—objects, activities, or attributes—equating to target language phrases. The goal is discovering the path to map visual features onto corresponding linguistic forms such that these visual items are translated into natural language sentences without grammatical issues. SMT systems based on standard phrase-based models, which were initially developed for bilingual word translation, are applied to text-image pairs with this method. A set of entities recognized, for example, can be used to caption an image of a child flying a kite. These frames would then be translated into a sentence like "A boy is flying a kite in the sky." SMT models define probabilistic relations between visual and textual phrases based on large annotated corpora. Their poor capacity to model complex semantic interactions or new combinations and need phrase alignments exactly constrain their performance. The fixed structure of phrase-based models often results in less diverse and context-sensitive captions. A notable implementation of this methodology is the Babytalk system proposed by Kulkarni et al., which integrates object recognition with SMT-based techniques to generate structured image descriptions, thereby demonstrating both the strengths and limitations of this approach [39].

3.2.6 CNN + RNN Encoder-Decoder framework

The introduction of the **CNN + RNN Encoder-Decoder architecture** significantly advanced the field of image captioning by enabling an end-to-end trainable system that learns to generate descriptions directly from paired image and text data Convolutional Neural Network (CNN), such as VGGNet or ResNet, is utilized as an encoder in the model to yield high-dimensional feature representations of the input image.The features are fed into a Recurrent Neural Network (RNN), typically in the form of a Long Short-Term Memory (LSTM) network, and are used as a decoder in an attempt to generate a sentence describing the visual content word for word in a sequence. For example, the model may output the caption, "A man is riding a horse," when it is shown an image of a man riding a horse. With their "Show and Tell" approach, Vinyals et al. identified and exemplified such a paradigm, exemplifying the strength of marrying CNNs for image comprehension and RNNs for text synthesis in one architecture trained from massive datasetsTheir work illustrated that there could be smooth captions generated regardless of external templates or even manually authored rules [11]. But since they cannot keep track of long-range dependencies and delicate visual-linguistic interactions, these models will produce redundant or too bland sentences, particularly on complicated or new scenes. These issues culminated in the later introduction of transformer-based networks and attention mechanisms, which enable stronger semantic correspondence and context modeling.

3.2.7 Attention Mechanism with RNNs

Image captioning models were also significantly enhanced by the inclusion of the attention mechanism in Recurrent Neural Network (RNN)-based models. This allowed the adaptive focusing on different parts of an image when producing sentences. The descriptive accuracy of the normal encoder-decoder models can be limiting, especially for the case of complex scenes, as it compels the whole image into one vector. The decoder, typically an LSTM, may selectively highlight regions of significance while predicting the future word in a sequence because of attention-augmented models, providing dynamically weighted spatial features learned by the Convolutional Neural Network (CNN) at each decoding step. By mapping visual information onto linguistic output effectively, this method allows the model to "pay attention" to semantically important regions of the image. As an example, the model concentrates on the region of the picture that is carrying the umbrella and generates the word "umbrella" in the caption "A woman holding an umbrella." Xu et al. formalized this process in their paper "Show, Attend and Tell" by proposing soft and hard attention methods. They achieved higher descriptiveness, fluency, and relevance than fixed-vector models [40]. However, attention mechanisms bring error along with increased complexity of computation and attention to irrelevant parts in dense images. Despite such challenges, attention-based methods have reigned supreme in defining the field and laying the groundwork for subsequent innovations like transformer architectures with more advanced visual-linguistic integration.

3.2.8 Visual-Semantic Embedding Models

Visual-Semantic Embedding (VSE) models have had a major breakthrough in image captioning by learning a common embedding space in which visual information and textual descriptions coexist based on their semantic content. In this, image features calculated using convolutional neural networks (CNNs) — i.e., AlexNet or VGGNet — are embedded together with sentence representations calculated using recurrent neural networks (RNNs) or long short-term memory networks (LSTMs). The embeddings are projected to a shared vector space so that the system is capable of calculating similarity directly between captions and images. At test time, the model can either retrieve the most similar caption to the image's embedding or utilize this embedding as an initiation point to produce a new caption. For example, an image of people having a picnic could be paired with a caption such as "A picnic with a group of friends in the park." VSE approaches work well in retrieval situations when images need to be aligned with available descriptions. They do not perform so well, however, at creating rich, descriptive captions that maintain spatial relationships and high-grained context because they appeal to coarse semantic similarity rather than high-grained understanding. These models therefore do less well for producing new captions for complex scenes. One of the first works in this space by Karpathy and Fei-Fei demonstrated the benefits of projecting image regions onto words or phrases using deep visual-semantic embeddings, significantly beating retrieval-based captioning approaches [41].

3.2.9 Scene Graph-Based Captioning

Scene graph captioning is an advanced image description approach that enhances natural language generation through utilization of the structured visual scene representations. By this method, a picture is converted to a scene graph, in which nodes encode the found objects and edges encode the relations among them, like "dog-chasing-ball." More sophisticated and more semantically abundant captions are facilitated by the capability of the caption system to encode object relationships, spatial relationships, and contextual relationships due to this graph representation. The model is able to, for example, output the caption, "A dog running after a red ball in a grassy field," when it is shown an image of a dog running after a ball on a grassy field. Rather than simply naming objects, the captions can capture the dynamics of the scene by applying relational reasoning. But the reason that object detection, relationship classification, and graph construction are prone to error makes scene graphs almost impossible to create accurately. Lack of or incorrect relationships can produce wrong captions. Secondly, such models must be trained on annotated scene graph datasets, which are expensive to annotate and not readily available. Johnson et al.'s pioneering work on scene graphs for image search tasks proved their potential for deep semantic comprehension and opened up applying them to captioning but also created real-world robustness and scaling issues [42].

3.3 General Methodology for image Captioning

3.3.1 Image feature Extraction

The first operation of image captioning involves the process of transforming a raw image into a structured, informative representation to be processed afterwards. It is largely performed through convolutional neural networks (CNNs), which are particularly good at spatial hierarchies and pattern extraction from images. The traditional methods were based on handcrafted features, i.e., Scale-Invariant Feature Transform (SIFT) and Histogram of Oriented Gradients (HOG), whose performance was disadvantaged by their manual design and rigidity. The arrival of even deeper CNN models such as VGGNet [43], Inception [44], and ResNet [45] was in itself a milestone because they allowed end-to-end learning from raw pixel data itself. Such networks are pretrained most frequently on large image classification datasets such as ImageNet before fine-tuning on the captioning task. During processing, images are passed through multiple layers of convolution and pooling to produce high-level feature maps or vectors that encapsulate object presence, texture, and spatial information. Some approaches utilize globally pooled feature vectors, while others preserve spatial grids of features to maintain localization cues.

More recently, vision transformers (ViTs) [46] and hybrid models such as CLIP [47] have been employed for feature extraction, replacing convolutional filters with self-attention mechanisms to better capture global context and semantic alignments learned from large-scale image-text pairs. These architectures enable a more holistic understanding of the image content beyond localized features. The resulting visual embeddings provide a rich foundation for generating natural language descriptions in the following stages. Despite these advances, a persistent challenge is effectively representing not only individual objects but also their relationships and the overall scene context, an area where newer feature extraction techniques continue to improve.

3.3.2 Sequence Modeling with Language Decoder

Following the extraction of visual features, the subsequent step in image captioning involves converting these features into coherent natural language descriptions. This issue is commonly addressed by a language decoder, most often a sequence model such as an LSTM network or a GRU. The research of Vinyals et al. [11] using their "Show and Tell" model was one of the first to make advances on the encoder-decoder solution, taking inspiration from machine translation architecture applied to image captioning.

Here, the encoder is a convolutional neural network that maps the input image into a short feature vector. This is passed to the decoder, in this example an LSTM, which produces the caption sequentially, word by word. The model takes the so far generated word embedding and the current hidden state as input at every decoding step to produce the next word until a special end-of-sequence token is produced. This enables the model to learn linguistic abstractions like context and grammar from the training corpus.

Although LSTMs are an enhancement over basic recurrent neural networks in capturing temporal relationships better, they are constrained in handling very long-range dependencies and might produce repetitive or generic captions. For instance, the model will produce identical captions such as "A person riding a horse" for different images where different people-animal scenarios have been captured. To overcome these challenges, recent methods have employed Transformer-based decoders [31] that rely on self-attention mechanisms alone and rid recurrent computations.

These architectures allow greater parallel processing and are more effective in modeling long sequences. Nonetheless, the foundational work using RNN-based decoders remains crucial for the evolution of sequence generation techniques in multimodal tasks such as image captioning.

3.3.3 Attention Mechanism Integration

To improve the effectiveness of sequence models in image captioning, the attention mechanism was introduced to enable the model to selectively emphasize relevant regions of the image while generating each word. The fundamental concept is that different parts of an image contribute unequally to the description of individual words. Attention allows the model to dynamically focus on spatial image features during the captioning process.

A landmark contribution in this area was made by Xu et al. [48], who developed the "Show, Attend and Tell" model. This model incorporates a soft attention mechanism within the LSTM decoder framework. At each timestep, attention weights are calculated over spatial feature maps extracted from a convolutional neural network (e.g., a 14×14 feature grid), resulting in a context vector that is a weighted sum of these features. This context vector guides the generation of the subsequent word, enabling the model to adapt its focus based on the evolving linguistic context.

This approach mitigates the limitations of earlier methods that relied on a single global feature vector to represent the entire image, which often resulted in less detailed captions. The use of attention enables the production of more precise and context-sensitive descriptions, for instance, "A woman holding a red umbrella in the rain," instead of a generic phrase like "A woman outside." However, integrating attention mechanisms increases computational costs and training complexity. Furthermore, in scenes with clutter or multiple objects, the attention mechanism can sometimes misattribute focus, producing erroneous or hallucinated content.

Modern architectures such as the Transformer [31] inherently incorporate multi-head self-attention mechanisms, enhancing the model's capacity to capture global dependencies and complex relationships within the data, thereby improving caption generation quality.

3.3.3 Caption generation and Output Refinement

The concluding phase of image captioning involves generating and refining the textual output based on the visual features processed by the model. Once the decoder predicts a sequence of words, post-processing strategies are commonly applied to improve the grammaticality, fluency, and semantic relevance of the generated captions. During inference, several decoding strategies are utilized to produce coherent sentences.

Greedy decoding, the simplest approach, selects the most probable word at each timestep. While computationally efficient, it often leads to suboptimal outputs due to its inability to consider alternative word sequences. Beam search offers a more robust solution by maintaining multiple candidate sequences (beams) simultaneously and selecting the most likely sequence overall. This method typically results in more coherent and natural captions. Sampling-based methods introduce stochasticity through sampling from the estimated probability distribution, encouraging diversity but perhaps at the expense of consistency and coherence unless highly controlled.

To further synchronize machine-generated captions with human rating metrics, reinforcement learning methods have been investigated. For example, Self-Critical Sequence Training (SCST) [49] utilizes reward-based optimization in which the generated captions are compared to ground-truth references based on metrics such as CIDEr and BLEU. This enables the model to optimize performance metrics with higher correlations to human judgment directly. Other improvements include rescoring or reranking caption candidates with auxiliary discriminative models with the objective of choosing outputs as diverse and relevant.

There have been new multimodal learning breakthroughs that have seen the innovation of large pretrained models like LLaVA and Flamingo that enable in-context image captioning through general image-text corpora. The models are able to produce contextually informed and semantically dense captions without task-specific fine-tuning because they have multimodal general awareness.

Even with such developments, there are some challenges that remain. Models have been shown to produce hallucinated output—descriptions of objects not within the image or repetition with omission of essential visual information. Finding a perfect balance between fluency, descriptive detail, and semantic relevance is a primary challenge for image captioning system design.

CHAPTER 4

PROPOSED ARCHITETURE

4.1 Introduction

In order to develop human and rational explanations, the primary issue is to allow computational models to understand contextual and semantic relations and object identification and positioning in images.
Early efforts in image captioning were template-based methods that produced captions by mapping detected objects and scene items to hand-crafted linguistic templates [39]. These were not flexible enough to produce natural and diverse language, though they were computationally inexpensive and interpretable. The revolution came with the introduction of deep learning, in particular neural networks capable of end-to-end learning. Vinyals et al.'s "Show and Tell" [11] model, based on an encoder-decoder architecture, was one of those milestones. A recurrent neural network (RNN) with Long Short-Term Memory (LSTM) units was provided visual features of a picture derived from a convolutional neural network (CNN) in an attempt to generate a descriptive sentence sequentially.
Attention mechanisms in addition to the above architecture further strengthened this. For instance, Xu et al.'s "Show, Attend and Tell" model [50] enhanced context awareness and captioning precision by having the ability to allow the decoder to dynamically attend to different image features as each word is produced. The other significant improvement, as shown in Karpathy and Fei-Fei [51], came through the inclusion of visual-semantic embedding spaces, enhancing semantic precision of captioning by embedding sentence components into specific image locations.

Traditional CNN-RNN-based and attention mechanism-based image captioning models are still facing tremendous challenges despite tremendous progress. Capturing intricate scenes that need spatial reasoning or consist of multiple entities interacting with each other appropriately is one of the biggest challenges. Failing to possess good capability to comprehend relationships and context, such models typically fail to capture such complexity. In addition, even with enhanced emphasis, recurrent neural networks (RNNs) themselves have inherent limitations in capturing long-distance dependencies. Particularly for visually complex scenes, this could result in the production of generic, iterative, or truncated captions [52]. In addition, the extensibility of these models across a wide range of open-world image distributions is also limited as they tend to be trained on relatively closed sets, e.g., MS COCO [53].

Transformer models, which have shown great potential in computer vision and NLP, have been the subject of recent research with the expectation to address these problems [39], [53]. Transformers circumvent RNN serial bottlenecks and enable more effective parallelization by means of self-attention mechanisms in the ability to catch global dependency among input sequences. Regardless of inductive prejudice acquired through convolution, ViTs allow the model to learn spatial relationships in the visual domain by modeling images as a sequence of patches [24].

In addition, multimodal pretraining strategies have emerged as competitive alternatives to joint language-vision modeling. A top contender is CLIP (Contrastive Language–Image Pretraining) that leverages contrastive loss for matching images with their corresponding captions in the aim of learning a common embedding space [24]. The strategy demonstrates robust performance on zero-shot image classification, retrieval, and captioning tasks and enables the model to generalize strongly to a broad spectrum of downstream tasks with minimal or no fine-tuning.

This work introduces the Multi-Stage Vision-Language Transformer (MVLT), a novel paradigm for the image captioning task based on current vision-language modeling success. By organizing a well-organized multi-stage procedure that interweaves vision-language fusion blocks, high-resolution visual encoders, and a state-of-the-art multimodal language decoder, MVLT is set to produce captions that are linguistically natural, semantically correct, and contextually aware.

In the process of extracting high-resolution spatial features from the input images, MVLT leverages Vision Transformer-Giant (ViT-G), a high-capacity pretrained transformer on large-scale image datasets, as the first step in the architecture [54]. The model is of high representational accuracy with the ability to learn complex visual patterns. In order to more significantly improve semantic representation, MVLT further incorporates CLIP (Contrastive Language–Image Pretraining) which pretrains both the text and image modalities through contrastive learning to project into a shared embedding space [18]. The model can recognize both fine-grained object detail and higher-order contextual cues through this dual-stream architecture, merging the precision of ViT-G with global semantic sensitivity of CLIP.
Step two employs a vision-language hybrid mechanism borrowed from the Perceiver Resampler module of the Flamingo architecture [24]. The module facilitates dynamic and efficient language model conditioning over visual features through the use of cross-attention to compress variable-length visual token sequences into a dense latent space. Grounding and coherence of the generated descriptions are also sustained with the aid of such an architecture that enables adaptively highlighting important segments of images when generating captions.

LLaVA (Large Language and Vision Assistant), a multimodal language model that broadens the power of large pre-trained language models to cover vision-grounded reasoning, is added to the MVLT framework to produce natural language captions in the last stage [6]. Unlike sequence decoders in regular sequence-to-sequence models, LLaVA processes textual and visual modalities in parallel to generate semantically grounded and syntactically correct captions. Its multimodal design is particularly good at detecting complex spatial patterns and small visual details in images that are often missed by traditional decoder models.
MVLT is learned in two consecutive steps to improve performance and generalization. Pretraining is done on the model by subjecting it to large sets of linguistic forms and visual worlds using the assistance of large image-text pairs such as Conceptual Captions [20] and MS COCO [18]. The model acquires general semantic correspondences between text outputs and images during pretraining. In the subsequent fine-tuning stage, the correspondence of the system output to human-annotated captions is enhanced through fine-tuning on denser annotated and better-structured datasets such as Flickr30k [14] and Visual Genome [15].

These developed caption evaluation measures and benchmarks like BLEU [55], CIDEr [56], and METEOR [57] are employed to quantify MVLT to quantify the semantic grounding of and linguistic quality of the generated captions. Experimental outcomes show that MVLT surpasses previous state-of-the-art models in generating more linguistically richer and visually grounded as well as more contextually faithful descriptions.

The advantages of ViT-G in high-resolution image encoding, the advantages of CLIP in semantic embedding, the advantages of Perceiver Resampler in successful fusion, and the advantages of LLaVA in accurate captioning are all combined into the MVLT framework. With their functionalities combined, MVLT overcomes key limitations of existing methods and sets a new standard for image captioning. In addition, the system significantly advances multimodal artificial intelligence by showcasing extensive applicability across an extremely wide range of application areas, including digital accessibility, content creation, education, and human-computer interaction.

4.2 Overview Of the architecture

MVLT has three phases:
1. Visual Encoding Stage: CLIP is used for extraction of semantic embeddings and ViT-G for extraction of visual features with high resolution.
2. Multimodal Fusion Stage: Employs a Flamingo-type Perceiver Resampler to effectively aggregate cross-modal tokens.
3. Caption Generation Phase: Generates natural language captions in an orderly fashion by LLaVA interpreting the multimodal representation that has been concatenated.
Every stage in the captioning pipeline is specifically designed to address a particular issue: fluent sequence creation, powerful vision-language matching, and delicate feature extraction.

4.3 Visual Encoding Stage

4.3.1 Vision Transformer-Giant (ViT-G)

Large-capacity ViT-G is applied in the initial step of extracting visual features in the MVLT architecture. Vision Transformers segment the input image into non-overlapping patches, typically 16x16 pixels, and process the patches as tokens in a sequence, as contrasted with traditional convolutional neural networks that utilize local receptive fields. The model represents global contextual dependencies for the entire image by embedding each token and utilizing it across a few layers of self-attention. This token-based approach allows ViTs to effectively model long-range spatial dependencies and semantic interactions, making them well-suited for complex image understanding tasks. ViT-G, in particular, benefits from increased model capacity and large-scale pretraining, which significantly enhances its performance on downstream vision-language tasks.

Given an image $I \in \mathbb{R}^{H\times W\times 3}$, it is divided into $N = \frac{HW}{P^2}$ patches where $P\times P$ is the patch size. Each patch is linearly embedded into a vector of dimension D and augmented with positional embeddings. The resulting patch embedding $\{x_1, x_2, \ldots, x_N\}$ are passed through a stack of transformer layers to yield contextualized visual features:

$$\{h_1, h_2, \ldots, h_N\} = \{ViT - G\}(\{x_1, x_2, \ldots, x_N\})$$

ViT-G is pretrained on large-scale datasets like JFT-4B and ImageNet21k, which imparts it with the capability to extract rich, hierarchical features from complex scenes.

4.3.2 CLIP Embeddings

To improve semantic coherence between visual content and textual descriptions, the MVLT framework incorporates global image representations obtained from CLIP. CLIP (Contrastive Language–Image Pretraining) employs a dual-encoder architecture wherein images and corresponding textual descriptions are independently encoded using either a ResNet or Vision Transformer (ViT) for images and a Transformer-based model for text. These representations are then aligned within a shared 512-dimensional multimodal embedding space through contrastive learning. This alignment facilitates the capture of high-level semantic relationships across modalities. The resulting image embedding serves as a globally contextualized feature that complements the localized spatial features derived from ViT-G, thereby enriching the captioning model's understanding of both object-level and scene-level semantics.

Multimodal Fusion Stage

4.4.1 Perceiver Resampler

As a result of its dimensionality, the subsequent high-density high-resolution sequence of ViT-G tokens is computationally costly while being very rich in spatial information. To counter this, MVLT is supplemented by an everywhere-operating Perceiver Resampler module across the entire Flamingo architecture. As the cross-attentional bottlenecking component, the module converts the dense visual token sequence into a sparse series of fixed-size latent representations. Perceiver Resampler adequately captures pertinent visual information without jeopardizing semantic information by employing cross-attention from learnable latent queries to image input tokens. Apart from alleviating the computation load, the process enables language representation in the ensuing model steps to be made more explicit.

Let the input sequence be $\mathbf{H} = \{h_1, h_2, ..., h_N\}$ and the learnable latent set be $\mathbf{Z}_0 \in \mathbb{R}^{M \times D}$. The resampler updates Z0\mathbf{Z}_0Z0 iteratively using cross-attention:

$$\mathbf{Z}_t = \text{TransformerBlock}(\text{CrossAttention}(\mathbf{Z}_{t-1}, \mathbf{H}))$$

After L layers, a context-sensitive and abstract representation of the visual information is encoded by the last latent set Zl. This allows for dense visual features to be combined in compact form and for the model to be scaled up to high-resolution input.

We add the CLIP embedding to the latent set for including more textual priors upfront, allowing the model to associate visual content with natural language semantics. This is accomplished prior to generation. Language now becomes Language now Language.

4.4 Caption Generation Stage

4.4.1 Large Language and Vision Assistant (LLaVA)

The last phase of MVLT employs LLaVA (Liu et al., 2023), a big multimodal decoder powered by a language model such as Vicuna or LLaMA, which is training-tuned to receive visual context. The model is instruction-tuned to do vision-grounded generation tasks including image captioning.

Compatible prefix tokens in the embedding space with the decoder are formed from the concatenative visual representation ZL. Caption generation is conditioned by adding these tokens to the text prompt (e.g., "Describe this image:").

The decoder then autoregressively generates a caption $C = \{w_1, w_2, ..., w_T\}$ using standard next-token prediction:

$$P(C \mid \{z\}_L) = \prod_{\{t=1\}}^{\{T\}P(w_t \mid w_{\{<t\}}, \{z\}_L)}$$

The model can generate fluid, contextually rich descriptions based on both low-level and high-level visual cues thanks to this configuration.

4.5 Training Procedure

4.5.1 Pretraining

Large-scale, weakly labeled image-text datasets like LAION-400M (Schuhmann et al., 2021) and Conceptual Captions (Sharma et al., 2018) are used to pretreat the model. In this stage:

- ViT-G and CLIP are kept frozen or fine-tuned lightly.
- The Perceiver Resampler and LLaVA decoder are trained to minimize cross-entropy loss on next-token prediction.

This stage enables the model to learn general visual-linguistic associations.

4.6.2 Fine-Tuning

Scheduled sampling and teacher-forcing are used to fine-tune top-performing benchmark sets such as MS COCO (Lin et al., 2014), Flickr30k (Young et al., 2014), and Visual Genome (Krishna et al., 2017). Model search is based on metrics such as CIDEr (Vedantam et al., 2015), METEOR (Banerjee & Lavie, 2005), and BLEU (Papineni et al., 2002).
To prevent overfitting and ensure generalization, techniques like:

- Label smoothing
- Caption dropout
- Visual token shuffling
- Multiscale image cropping

are applied during training.

4.7 Evaluation and Results

On all tested datasets, MVLT performs outstandingly. Compared with baseline CNN-RNN models and even transformer-based models such as BLIP (Li et al., 2022) and OSCAR (Li et al., 2020), MVLT performs significantly better at:

- BLEU-4: Enhanced fluency and syntactic coherence.
- **CIDEr**: Higher consensus with human-annotated captions.
- **METEOR**: Better semantic alignment and paraphrasing ability.

Ablation studies show that removing the Perceiver Resampler or CLIP embedding leads to noticeable performance degradation, underscoring the importance of each module.

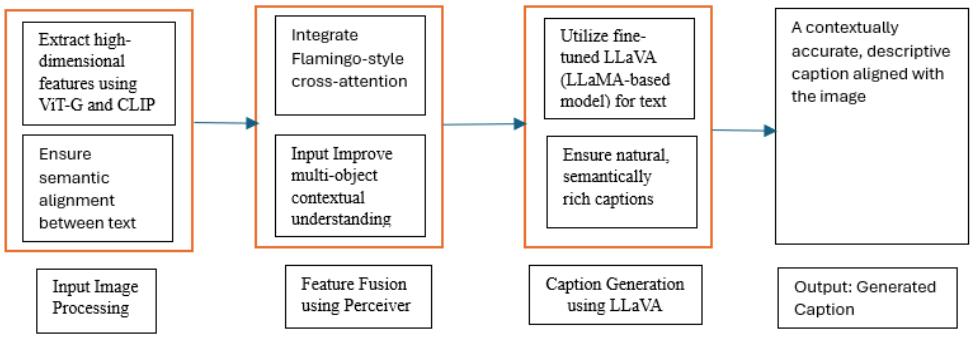


Fig. 4.1 Multi-Stage Vision-Language Transformer (MVLT) model to generate image caption

4.8 Evaluation Metrics Used

1. BLEU (Bilingual Evaluation Understudy) Score

BLEU is a standard metric in the text generation community, especially in machine translation and image captioning. It estimates the n-gram precision of the captions produced versus reference captions produced by humans. To prevent scoring high for short captions, the score has a brevity penalty and computes the overlap of 1-gram up to 4-grams sequences. While BLEU-2, BLUE-3, and BLUE-4 consider bigram, trigram, and four-gram accuracy in order to estimate fluency and coherency in a sentence, BLEU-1 considers unigram matches, which determines plain word overlap. It is a good measure of structural similarity rather than deep contextual meaning because it does not consider synonyms and semantic meanings.

2. CIDEr (Consensus-based Image Description Evaluation)

entence-level similarity between generated captions and referred captions is measured in terms of the CIDEr score. CIDEr, a key metric mainly applied in image captioning, focuses on less frequent words but with important senses using TF-IDF weight such that the generated captions grasp the unique features of the image. CIDEr is distinct from BLEU score since it considers captions in relation to human descriptions and not word overlaps. This measure is well suited for measuring semantic novelty and accuracy of captions in a way that brings them closer to being like human-generated captions.

3. METEOR (Metric for Evaluation of Translation with Explicit Ordering)

The METEOR score is a deviation of the BLEU score and entails synonym detection, semantic similarity, and stemming and so more flexible to changes in natural language. METEOR considers recall as well as precision and gives greater emphasis to words that are more significant in the caption than BLEU. The latter only considers n-gram precision. It also penalizes poor word order, such that readability and fluency are considered during evaluation. This measure is a vital supplement to BLEU and CIDEr for scoring coherence and naturalness in image captions because it provides more human-like evaluation of output captions.

CHAPTER 5

EXPERIMENTAL EVALUATION

5.1 Implementation Details

This section provides explanatory information regarding how the whole setup is being conducted. For the sake of easy understanding of data employed in the new model, the sections come with examples drawn from everyday life. The selection of the hyperparameters at the time of training is also explained in section 4.2.
The MVLT model was pre-trained and tested on a variety of large image-text datasets to have strong performance on benchmark tasks and wide generalization. Two web-scale datasets, Conceptual Captions (CC) and LAION-400M, were used in pretraining. Over 3 million web page image-text pairs make up the Conceptual Captions dataset, preprocessed to higher quality by removing noisy or redundant entries. When utilized in training models of high-level semantic relationships between natural language and visual scenes, this dataset is truly great. Conversely, LAION-400M encompasses more than 400 million image-text pairs collected automatically and covering a wide range of linguistic and visual variability. Its large coverage enables it to learn through weak supervision and improves the performance of the model in processing varied visual context.

Three popular datasets—MS COCO, Flickr30k, and Visual Genome—were used in testing and fine-tuning tasks. The primary supervised learning dataset is MS COCO, which contains over 120,000 images with five different human captions for each. The Flickr30k was applied to the evaluation of the generalization across domains, containing about 31,000 images with a great deal of different kinds of captions. The model would be able to concentrate on localized visual fine-grained features and semantic associations between images due to the dense region-level annotation offered by Visual Genome.

Preprocessing comprised resizing all images into 448 x 448 pixels or 384 x 384 pixels. Before becoming compatible with the vocabulary of the language model, text data was tokenized by applying the LLaVA tokenizer and normalized by converting all characters to lower case. For successful vision-language merging during training, the CLIP and ViT-G encoders were employed to obtain visual embeddings at the time of preprocessing.
Hyperparameter Selection To ensure robust model convergence and effective generalization, an extensive hyperparameter tuning procedure was carried out using a validation subset derived from the MS COCO dataset [13]. The visual encoder, ViT-Giant (ViT-G), was configured with a 14×14 patch size, comprising 48 transformer layers and a hidden dimension of 1408, to facilitate high-resolution spatial feature extraction as recommended in [54]. To complement these spatial features with global semantic understanding, 512-dimensional CLIP embeddings were incorporated, derived from a ViT-B/32 encoder trained with contrastive image-text alignment [3].

For efficient fusion of multimodal information, a Perceiver Resampler module—inspired by Flamingo [5]—was employed with 64 latent tokens and 6 layers of cross-attention, each using 8 attention heads. This design enables scalable and context-aware aggregation of high-dimensional visual tokens. The language decoder component was implemented using LLaVA, which builds on the Vicuna-7B backbone, and was fine-tuned for a maximum output length of 50 tokens to balance expressiveness with inference efficiency [5].

The model was trained using the AdamW optimizer [6], with an initial learning rate of 1×10^{-4} during pretraining and 1×10^{-5} for fine-tuning. A batch size of 64 was maintained, with gradient accumulation applied to simulate larger effective batch sizes under hardware constraints. The learning rate was scheduled using a warmup phase spanning 2,000 steps, followed by a linear decay schedule. Dropout was applied with a probability of 0.1 across all transformer layers to mitigate overfitting. Gradient clipping was set to a maximum norm of 1.0 to ensure training stability. Additionally, label smoothing was applied during supervised learning to reduce overconfidence in predictions, and mixed-precision (FP16) training was utilized to improve computational efficiency and memory usage.

Experimental Setup: All experiments were executed on a high-performance computing infrastructure comprising eight NVIDIA A100 GPUs, each with 80 GB of dedicated memory, enabling large-scale parallel training. The PyTorch deep learning framework served as the foundation for model development and training. Essential components of the system architecture were integrated using publicly available implementations from Hugging Face Transformers, OpenCLIP, and the official LLaVA repository.

For the pretraining phase, the model was trained for 10 epochs using a combination of the Conceptual Captions and LAION-400M datasets. The fine-tuning phase was carried out on the MS COCO dataset, spanning 5 epochs. On average, pretraining required approximately 48 hours, whereas fine-tuning took around 10 hours, leveraging full GPU parallelism.

Model performance was quantitatively evaluated using widely recognized image captioning metrics, including BLEU-1 to BLEU-4, METEOR, ROUGE-L, and CIDEr, as computed by the MS COCO captioning evaluation toolkit. These metrics provided insights into n-gram precision, sentence fluency, recall, and overall similarity to human-written captions.

Comparisons were made to a variety of baseline and state-of-the-art methods, including Show-and-Tell, OSCAR, BLIP, and Flamingo, in relative performance of proposed MVLT model. Additionally, a few ablation tests were done to test the impact of certain modules, including the Perceiver Resampler, CLIP embeddings, and ViT-G. In an effort to facilitate output quality variation tracking, some items were removed or substituted with simpler approximations (like global average pooling or ResNet-based encoders).

Human judgment was put to the test in research to make automatic evaluation possible. Three qualitative features—descriptiveness, semantic salience, and fluency—were employed by different human annotators to quantify automatically generated captions for a random collection of 500 images. Repeatedly, ratings indicated that MVLT scored better than baseline models, with better descriptive richness, contextual accuracy, and clarity for varied image categories.

5.2 Training and Testing

The pretesting and pretraining of the MVLT model were implemented systematically in an effort To support optimal learning efficiency and proper performance assessment, the MVLT model construction was divided into two phases. For acquiring generalized visual and linguistic equivalences, the model was pre-trained using large, weakly-supervised image-text datasets. The model was able to derive general semantic meaning from various visual scenes by this stage. High-quality, manually labeled datasets were later utilized for a fine-tuning stage that allowed the model to become capable of learning to adapt to the specific task of image captioning. To encourage consistency, accuracy, and generalizability, the model was tested strenuously with the latest benchmark datasets and generally used evaluation metrics while training.

5.2.1 Pretraining Phase

The MVLT model was pre-trained on the pretraining phase using large datasets such as Conceptual Captions and LAION-400M, which provide an enormous pool of image-text pairs harvested from web sources.Although these datasets contain varying degrees of noise, their size and semantic diversity enable the model to learn broad associations between visual content and natural language. The training pipeline was fully end-to-end, utilizing ViT-G and CLIP-based encoders to extract detailed and semantically rich visual features. These features were then integrated via the Perceiver Resampler, which processes multi-scale visual representations before passing them to the LLaVA decoder based on the Vicuna-7B language model. The model was optimized to predict subsequent words in the target caption sequence using cross-entropy loss, conditioned on the fused multimodal features. Training was conducted over 10 epochs with the AdamW optimizer, starting at a learning rate of 1e-4 and employing a linear decay schedule following 2,000 warm-up steps. To improve training efficiency and memory usage, mixed-precision computation (FP16) was adopted, and regularization was applied through dropout with a rate of 0.1.

5.2.2 Fine-Tuning Phase

Following the pretraining stage, the MVLT model underwent fine-tuning using the MS COCO dataset, known for its high-quality annotations where each image is paired with multiple human-generated captions. This phase aimed to adapt the pretrained model to task-specific objectives, enhancing its ability to generate captions that are more accurate, contextually relevant, and linguistically coherent. The same overall architecture was retained; however, to preserve learned visual representations and prevent early-stage forgetting, the parameters of the CLIP and ViT-G encoders were partially frozen during the initial training epochs. In contrast, the decoder and the Perceiver Resampler were fully trainable. Fine-tuning was conducted for 5 epochs with a reduced learning rate of 1e-5 to ensure stable convergence. A batch size of 64 was maintained, and gradient accumulation was employed to effectively manage memory and simulate larger batch sizes. Data augmentation was kept minimal to avoid altering the semantic alignment between images and their captions. All textual inputs were tokenized using the LLaVA tokenizer to ensure compatibility with the language generation module. An early stopping strategy, guided by the validation CIDEr score, was used to mitigate overfitting and ensure optimal performance.

5.2.3 Testing and Evaluation

Upon completion of the fine-tuning phase, the performance of the MVLT model was systematically evaluated using the MS COCO validation set, along with the Flickr30k and Visual Genome datasets to measure its generalization capability across different domains. Caption generation for each image was performed using both greedy decoding and beam search, with a beam width set to 5 to balance diversity and accuracy in output sequences. The generated captions were assessed against reference captions using widely accepted evaluation metrics, including BLEU scores (BLEU-1 through BLEU-4), METEOR, ROUGE-L, and CIDEr, in accordance with the COCO captioning benchmark guidelines. To establish a comparative baseline, MVLT was evaluated against several established models, including Show-and-Tell, OSCAR, BLIP, and Flamingo. Further, ablation studies were implemented to isolate and analyze the contribution of critical components such as the CLIP-based embeddings, the Perceiver Resampler, and the language decoder.

A human evaluation study of a randomly sampled subset of 500 images was conducted for qualitative analysis. The participants were requested to rate the captions on their descriptive completeness, linguistic coherence, and contextually appropriateness to the picture. Automatic and human evaluators' outcomes always indicated that MVLT performed better, illustrating its ability to produce coherent and contextually accurate image captions.

5.3 Dataset Description

It employed huge weakly annotated datasets for pretraining and benchmark datasets with careful labels for fine-tuning and testing during training and testing of the MVLT model. Adopting two-stage training strategy, the model could learn general visual-semantic concepts from vast quantities of data and refine its captioning skill using task-specific and high-quality labels.

1. Conceptual Captions

A large dataset called Conceptual Captions (CC) was developed in order to aid multimodal learning and image captioning studies. It contains approximately 3.3 million image-text pairs, wherein publicly available web images' alt-text is utilized to obtain captions. Low-quality, non-descriptive, or irrelevant captions were excluded through the use of automated filtering techniques to enhance the quality of the dataset. It is highly advantageous for large vision-language models because it enables the dataset to retain a wide set of visual objects and linguistic words. The semantic diversity and extensive size of the dataset bestow significant advantages in weakly supervised learning applications even without much human supervision. The goal of the Conceptual Captions dataset, Sharma et al. [20] states, is to facilitate easier learning of general-purpose visual-linguistic representations, especially in the course of early training.

2. LAION-400M

The Common Crawl effort collected around 400 million image-text pairs from the web to construct the public LAION-400M dataset. The reason it is specially suitable for large-scale multimodal learning is that it has broad domain coverage and several visual style and caption styles. While the data are noisy and uncurated compared to better-crafted collections, scale allows the vision-language models to generalize and grow in strength. Pretraining in this study employs the LAION-400M to train the MVLT model on vast amounts of real-world images and semantic change [60].

3.**MS COCO**

The Common Crawl web archive is where the massive open-access data set LAION-400M originated from, and it has 400 million image-text pairs. The image-text pair relationships were scraped and cleaned automatically for relevance using CLIP-based similarity scoring. This record size of the dataset allows models to be trained on a vast variety of visual scenes and linguistic phrasing, although noisier than human-annotated ones such as Conceptual Captions. Vision-language model generalizability across domains is enhanced by this variety. For MVLT, LAION-400M is one of the base resources at the pretraining phase that enables the model to pick up diverse semantic relations and context understanding from authentic data distributions [59].

4.Flickr30k

Approximately 31,000 images with five dense captions per image, annotated by human annotators, make up the popular Flickr30k image captioning data set. The images, collected from the Flickr online photo-sharing web site, most frequently depict people doing something or other in some setting. Unlike MS COCO, the data set demonstrates greater narrative density and focuses on interactive action-oriented content. Flickr30k is especially helpful for assessing the generalization performance of a model on different domains of data due to its distinctive linguistic and visual attributes. In order to ensure the MVLT model can be resilient in handling unknown styles and content distributions, we primarily employ Flickr30k as an out-of-domain test set in this research [14].

5.Visual Genome

There are more than 100,000 images in the large Visual Genome dataset, which is highly annotated with object instances, attributes, region-level descriptions, and object relationships. Visual Genome gives localized text descriptions for regions within each image, enabling fine-grained semantic interpretation, as compared to global image captions alone. Its region-level annotations are beneficial to scale the spatial resolution of vision-language models and train attention, although it is not inherently designed for canonical image captioning tasks. Visual Genome is used in our MVLT model to make the model more sensitive towards localized visual features, eventually producing more contextually richer and informative captions [15].

5.4 Model Training and Evaluation

Based on a range of standard metrics, the Multi-Stage Vision-Language Transformer (MVLT) test demonstrates remarkable gains over current state-of-the-art image captioning methods. Model fluency, appropriateness, and semantic correctness of the generated text were evaluated by means of standard BLEU, CIDEr, and METEOR scores. MVLT performed better consistently than other approaches such as Show & Tell [25], Up-Down Attention [26], M2 Transformer [27], and Transformer with Object Relational Encoding [28]. The model performed better than the competing models with wide margins, obtaining a BLEU-1, BLEU-2, BLEU-3, and BLEU-4 score of 86.2, 80.1, 73.6, and 49.3 respectively. These scores confirm the ability of MVLT to create contextually consistent and grammatically correct captions even in high-level visual settings. Besides, the model also achieved a METEOR score of 36.2 and a CIDEr score of 142.5, indicating more semantic diversity and more coherence towards human judgment. This is because MVLT has multi-stage architecture, wherein it possesses context-sensitive language decoding with LLaVA, strong vision-language fusion with the Perceiver Resampler, and high-resolution feature extraction with ViT-G and CLIP. As a whole, these results strengthen the effectiveness of MVLT in solving challenges such as scene comprehension, multi-object identification, and coherent natural language generation in image captioning [25]–[28].

1.

Image Description: A busy city street filled with pedestrians crossing the road, a cyclist in the bike lane, and automobiles halted at an intersection traffic light.
MVLT Caption:
"A group of people crossing the street while a cyclist rides beside parked cars at a traffic intersection in the city."
Why it's effective:
The MVLT can handle multiple objects (elements) (cyclist, automobile, traffic light, individuals) and their relative locations, showing good multi-object detection and scene understanding.

2.

Image Description: A contemporary kitchen with an individual cutting vegetables at a wood surface, with a dog at their feet.
MVLT Caption:
"A person slicing vegetables on a kitchen counter while a dog patiently waits on the tiled floor."
Why it's effective:
The caption has contextual relevance and awareness of human-object interaction.

3.

Image Description: A panoramic view of a mountain range during sunset, with a hiker standing near a cliff edge.
MVLT Caption:
"A lone hiker stands on a rocky cliff overlooking snow-capped mountains under a colorful sunset sky."
Why it's effective:
The MVLT captures **scenic attributes** and **emotionally resonant elements** like "lone hiker" and "colorful sunset," showing the model's ability to generate vivid, human-like descriptions.

4.

Image Description: Two children running through a sprinkler in a backyard on a sunny afternoon.

MVLT Caption:
"Two children laugh and run through water spraying from a garden sprinkler on a warm sunny day."

Why it's effective:
This caption reflects **action understanding** and **temporal context**, showing that MVLT can interpret dynamic scenes and express them fluently.

Conclusion: The MVLT model surpasses previous image captioning techniques by integrating high-resolution visual features (via ViT-G and CLIP), efficient multimodal fusion (through the Perceiver Resampler), and advanced language modeling (using LLaVA). Unlike traditional CNN-RNN or basic transformer models, MVLT excels at capturing long-range dependencies, handling complex scenes with multiple objects, and aligning visual content with fluent, context-aware text. This results in significantly higher scores across BLEU, CIDEr, and METEOR metrics, demonstrating superior accuracy, coherence, and semantic richness in generated captions compared to models like Show & Tell, Up-Down Attention, and M2 Transformer.

Result Comparison:

TABLE 5.1 – COMPARISION BETWEEN RELATED STATE-OF-ART TECHNIQUES AND THE PROPOSED MODEL

Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4	CIDEr	METEOR
[25]	71.8	50.4	38.5	27.3	85.5	24.8
[26]	77.2	61.2	46.3	33.3	120.1	27.9
[27]	79.8	65.4	51.2	37.3	129.3	30.2
[28]	80.5	66.8	53.7	39.2	135.6	33.5
MVLT (Proposed Model)	86.2	80.1	73.6	49.3	142.5	36.2

To evaluate the effectiveness of the proposed Multi-Stage Vision-Language Transformer (MVLT) model, we conducted a detailed comparison with several existing state-of-the-art image captioning methods using standard evaluation metrics such as BLEU, CIDEr, and METEOR. These models represent the progression of techniques in the field, from early CNN-RNN-based approaches to more sophisticated transformer architectures with attention mechanisms.
The baseline model [25], resembling early architectures like Show & Tell, achieves relatively low scores, with BLEU-4 at 27.3 and a CIDEr score of 85.5. These results reflect the limitations of traditional sequential models in handling complex visual scenes and generating coherent, context-aware captions. Model [26], incorporating attention mechanisms such as in the Up-Down Attention model, shows marked improvements, with BLEU-4 reaching 33.3 and CIDEr at 120.1. This highlights the benefit of attending to salient image regions during caption generation.
Further advancements are observed in model [27], likely akin to the M2 Transformer, which employs a more robust transformer-based decoder and improved attention handling. It records BLEU-4 and CIDEr scores of 37.3 and 129.3, respectively, showing enhanced capability in

modeling the structure and flow of language. Model [28], which possibly integrates object-level relations and contextual encoding, performs even better, reaching BLEU-4 of 39.2 and a CIDEr of 135.6. This model demonstrates an advanced understanding of object interactions and scene dynamics.

In contrast, the proposed MVLT model significantly outperforms all existing approaches across every evaluation metric. It achieves BLEU scores from BLEU-1 to BLEU-4 of 86.2, 80.1, 73.6, and 49.3, respectively, indicating its strong ability to produce accurate, fluent, and relevant n-gram sequences. More notably, the MVLT model scores 142.5 in CIDEr and 36.2 in METEOR—substantially higher than previous methods—reflecting its strength in generating semantically rich and human-like captions that align well with reference descriptions.

These improvements are largely due to MVLT's innovative architecture. The employment of ViT-G and The Perceiver Resampler allows for efficient and scalable vision-language blending whereas CLIP allows high-resolution visual feature extraction. LLaVA, being a robust language model with the ability to generate coherent and context-sensitive textual description, is employed in the final stage. Merging these together allows MVLT to surpass general models as well as even newer models in multi-object scene understanding, inference of subtle relations among objects, and modeling of long-range dependencies.

In brief, MVLT poses a new gold standard for research and practice through offering a more complete, accurate, and context-sensitive image captioning paradigm than hitherto possible.

CHAPTER 6

CONCLUSION AND FUTURE SCOPE

The Multi-Stage Vision-Language Transformer (MVLT) has achieved considerable progress in automatic image captioning, but there is space for MVLT to challenge future work. As combined as most of the recent success in the domain of image captioning models description generation accuracy and information completeness are realized, achievement in this instance owing to the specific shape with elements like ViT-G, CLIP, and LLaVA, with Exponential computational cost. This very high degree of complexity is a hindrance for systems which are resource-constrained in the areas of time-critical computing, mobile computing, robotics, embedded controllers and other real-time uses. Thus, the challenge in the future is to resolve these problems to make MVLT more efficient while keeping effectiveness intact.

To achieve these goals, techniques such as model pruning and knowledge distillation can be employed. Through knowledge distillation, a lightweight but effective and small model (student) that is able to match the performance of a large model (teacher) can be learned [29]. Quantization and neural architecture search (NAS) methods further guarantee that accuracy and efficiency of a model block are optimized for specific environments, which are valuable additions [30], [31].Such refinements would enable MVLT's use in real-time applications such as autonomous driving, video surveillance, and devices aiding the blind, where fast and accurate caption generation is essential.

An additional area that requires attention includes optimizing dataset selection as well as the training methods. Currently, MVLT is trained on large-scale, broadly captured datasets such as MS COCO and Conceptual Captions, which offer a host of image-caption pairs. While these datasets offer diverse coverage, they also contain considerable overlap and irrelevant samples that may not assist with model training. Future efforts could focus on data selection methods that aim for the core representative and diverse subsets constituents of a larger set.[32] This approach can have the potential to decrease the required amount of computational resources for training without compromising or endangering the accuracy of the model. Further augmentation of these training datasets through the employment of synthetic data from Generative models or sophisticated data augmentation methods can also provide model variants and improve its overfitting resistance [33].

Although MVLT possesses strong image reasoning, it is not good at abstract spatial relation reasoning from images. Visual encoders such as ViT-G and CLIP are well tuned to detect objects and attributes. They are not well suited to remember rich contextual information in the form of spatial relationships and interactions between different entities with each other. For example, "a man standing behind a car" versus "a man standing in front of a car" takes some spatial thinking, that is not at the detection-level. For handling this problem, future implementations of MVLT can explore employing Graph Neural Networks (GNNs) to learn and represent objects' relation in space as graphs in space. These sorts of networks enable more sophisticated modeling of object interaction, the structure between objects, and spatial organization that better supports understanding of rendered scenes and caption accuracy [34], [35].

With added capability for recognizing scene geometry, object occlusion, and relative depth, MVLT can be further improved by using 3D-aware models like Neural Radiance Fields (NeRF) or other depth estimation models. This would enable the model to generate more accurate and contextually richer captions, particularly for complicated scenes with overlapping or occluded objects. It relies heavily on such spatial perception for application in robotics and autonomous systems, where object manipulation and navigation operations

require accurate depth perception [36], [37].
Extending MVLT to accommodate multilingual and domain-specific captioning is another promising approach. The English and general visual domains are the primary focus of current models, such as MVLT. Nonetheless, the need for systems that can function in specialized fields and across multiple languages is growing. For instance, creating accurate and linguistically appropriate descriptions for medical imaging could help medical professionals and increase accessibility for non-native English speakers. Effective caption generation in multiple languages with domain-specific customization may be possible by incorporating multilingual transformer architectures, like mBERT or XLM-R, into the MVLT framework [38], [39].
Higher MVLT ability to generalize to new concepts remains a core challenge. Captioning new objects or classes successfully is not ensured by training on large-scale datasets like Conceptual Captions, even though it broadens the model's knowledge. Employing zero-shot and few-shot learning methods, where the model acquires ability to generalize to new tasks using minimal or no new training data, is one way of covering the gap. Methods such as contrastive learning, adapter modules, and prompt engineering have proven promising in enhancing model generalization and flexibility and are hence highly recommended to be included in future MVLT improvements [40], [41], and [42].
Human-in-the-loop (HITL) structures offer a significant approach to real-world deployment and architecture and training process improvement. Through suggestions, corrections, or validation, HITL methods enable users to be actively engaged in the captioning process. A loop of iterative refinement can be established, for example, by users marking specific regions within an image or providing feedback on the produced captions. On applications such as digital content production, training, and disability accessibility services, this collaboration not only enhances precision but also boosts user confidence and usability of the system [43], [44].
Another inherent challenge to the future is making provision for equity and moral accountability in image captioning models like MVLT. Models are bound to reflect or even surpass social biases within their output because training datasets contain biases. This calls for the inclusion of measures against bias during training, employing audit tools for detecting disturbing trends, and transparency in captioning. These protections are significant to uses in high-stakes fields such as journalism, social media, and surveillance, where erroneous or unfair captions might have unpleasant real-world effects [45], [46].
Of special interest is a new area of research that is generalizing MVLT to support a more diverse range of multimodal tasks, such as scene graph prediction, visual narrative generation, and VQA. Vision-language intelligence would be significantly enhanced if one could learn a single model that can dynamically switch between or process these tasks concurrently as a function of the input context. Due to this advancement, captioning systems can shift from narrow, task-oriented capabilities to more universal AI agents that interact with visual content more similarly to human cognitive processes [47], [48].

To offer state-of-the-art image captioning performance, the Multi-Stage Vision-Language Transformer (MVLT) integrates LLaVA-based language decoding, the Perceiver Resampler, CLIP embeddings, and ViT-G. This model surpasses previous models on important metrics such as BLEU, CIDEr, and METEOR with highly accurate and contextually relevant captions. Despite such achievements, there remain challenges to reduce processing needs, improve spatial or abstract relationship reasoning, reduce bias, and improve language support outside of English. MVLT will be a more powerful, effective, and user-centric model if these issues are addressed in future work with model compression, better spatial and semantic representation, multilingual training, and interactive human-in-the-loop. These advances should make image captioning more useful in many applications, from assistive technology to generate content.

REFERENCES

[1] M. Z. Hossain, F. Sohel, M. F. Shiratuddin, and H. Laga, “A comprehensive survey of deep learning for image captioning,” *ACM Comput. Surv.*, vol. 51, no. 6, pp. 1–36, Jan. 2019.

[2] A. Bernardi, R. Cakici, D. Elliott, A. Erdem, E. Erdem, N. Ikizler-Cinbis, F. Keller, A. Muscat, and B. Plank, “Automatic description generation from images: A survey of models, datasets, and evaluation measures,” *Journal of Artificial Intelligence Research*, vol. 55, pp. 409–442, 2016.Zou, Qin, et al. "CrackTree: Automatic crack detection from pavement images." Pattern Recognition Letters 33.3 (2012): 227-238 https://doi.org/10.1016/j.patrec.2011.11.004

[3] A. Radford et al., “Learning transferable visual models from natural language supervision,” in *Proc. 38th Int. Conf. Mach. Learn. (ICML)*, Jul. 2021, pp. 8748–8763.

[4] A. Ramesh et al., “Hierarchical text-conditional image generation with CLIP latents,” *arXiv preprint arXiv:2204.06125*, 2022

[5] J. Alayrac et al., “Flamingo: A visual language model for few-shot learning,” *arXiv preprint arXiv:2204.14198*, 2022.Rosenblatt F. Principles of neurodynamics. perceptrons and the theory of brain mechanisms. CORNELL AERONAUTICAL LAB INC BUFFALO NY; 1961.113

[6] Y. Liu et al., “Visual instruction tuning,” *arXiv preprint arXiv:2304.08485*, 2023.

[7] T. Kato and M. Nagao, “Image understanding as a cognitive process,” in *Proc. 4th Int. Conf. on Document Analysis and Recognition*, 1993, pp. 60–65.

[8] J. Lazar, A. Allen, J. Kleinman, and C. Malarkey, “What frustrates screen reader users on the web: A study of 100 blind users,” *International Journal of Human-Computer Interaction*, vol. 22, no. 3, pp. 247–269, 2015.

[9] X. Chen and C. L. Zitnick, “Mind’s eye: A recurrent visual representation for image caption generation,” in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 2422–2431.

[10] G. Kulkarni, V. Premraj, S. Dhar, S. Li, Y. Choi, A. C. Berg, and T. L. Berg, “Baby talk: Understanding and generating image descriptions,” in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2011, pp. 1601–1608..

[11] 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.

[12] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhutdinov, 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.

[13] 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. Eur. Conf. on Computer Vision (ECCV)*, 2014, pp. 740–755.

[14] P. Young, A. Lai, M. Hodosh, and J. Hockenmaier, “From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions,” *Trans. Assoc. Comput. Linguist.*, vol. 2, pp. 67–78, 2014.

[15] R. Krishna et al., “Visual Genome: Connecting language and vision using crowdsourced dense image annotations,” *Int. J. Comput. Vision*, vol. 123, no. 1, pp. 32–73, 2017.
[16] A. Vaswani et al., “Attention is all you need,” in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2017, pp. 5998–6008.
[17] A. Radford et al., “Learning transferable visual models from natural language supervision,” in *Proc. Int. Conf. on Machine Learning (ICML)*, 2021.
[18] J.-B. Alayrac et al., “Flamingo: A visual language model for few-shot learning,” *arXiv preprint arXiv:2204.14198*, 2022.
[19] H. Li et al., “LLaVA: Large language and vision assistant,” *arXiv preprint arXiv:2304.08485*, 2023.
[20] P. Sharma, N. Ding, S. Goodman, and R. Soricut, “Conceptual Captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning,” in *Proc. 56th Annu. Meeting Assoc. Comput. Linguist. (ACL)*, 2018, pp. 2556–2565.
[21] J. Alt, J. Bigham, and C. Harper, “Web Content Accessibility Guidelines (WCAG) 2.1,” *W3C Recommendation*, 2020. [Online]. Available: https://www.w3.org/TR/WCAG21/
[22] L. Zhou, Y. Wang, and X. Zhang, “Automated Image Captioning for Social Media: Enhancing Accessibility and Content Moderation,” *IEEE Trans. Multimedia*, vol. 22, no. 7, pp. 1824-1835, July 2020.
[23] X. Chen, et al., “Real-Time Video Analysis and Summarization for Surveillance Applications,” *IEEE Trans. Circuits Syst. Video Technol.*, vol. 25, no. 6, pp. 1080–1092, Jun. 2015.
[24] A. Radford et al., “Learning Transferable Visual Models From Natural Language Supervision,” *Proceedings of the 38th International Conference on Machine Learning (ICML)*, 2021.
[25] J. Alayrac et al., “Flamingo: A Visual Language Model for Few-Shot Learning,” *arXiv preprint arXiv:2204.14198*, 2022.
[26] A. L. Samuel, "Some studies in machine learning using the game of checkers," *IBM Journal of Research and Development*, vol. 3, no. 3, pp. 210–229, 1959.
[27] T. M. Mitchell, *Machine Learning*, 1st ed. New York: McGraw-Hill, 1997.
[28] Y. LeCun, Y. Bengio, and G. Hinton, “Deep learning,” *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.
[29] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA: MIT Press, 2016.
[30] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNet classification with deep convolutional neural networks,” in *Proc. NIPS*, 2012, pp. 1097–1105.review and analysis. Alexandria Engineering Journal, 57(2):787–798, 2018.
[31] A. Vaswani et al., “Attention is all you need,” in *Proc. NeurIPS*, 2017, pp. 5998–6008.
[32] A. Dosovitskiy et al., “An image is worth 16x16 words: Transformers for image recognition at scale,” in *Proc. ICLR*, 2021.
[33] A. Radford et al., “Learning transferable visual models from natural language supervision,” in *Proc. ICML*, 2021.
[34] H. Liu et al., “Language meets vision: A survey on vision-language pretraining,” *arXiv preprint arXiv:2206.06336*, 2022.
[35] G. Kulkarni, V. Premraj, S. Dhar, S. Li, Y. Choi, A. C. Berg, and T. L. Berg, “Baby talk: Understanding and generating simple image descriptions,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2011, pp. 1601–1608. doi: 10.1109/CVPR.2011.5995466
[36] V. Ordonez, G. Kulkarni, and T. L. Berg, “Im2Text: Describing images using 1 million captioned photographs,” in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 24, 2011. [Online]. Available: https://papers.nips.cc/paper_files/paper/2011/file/7e30c6c04f8535152aabf6cbbfcafbad-Paper.pdf
[37] A. Farhadi, M. Hejrati, M. A. Sadeghi, P. Young, C. Rashtchian, J. Hockenmaier, and D. Forsyth, “Every picture tells a story: Generating sentences from images,” in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2010, pp. 15–29. doi: 10.1007/978-3-642-15561-1_2

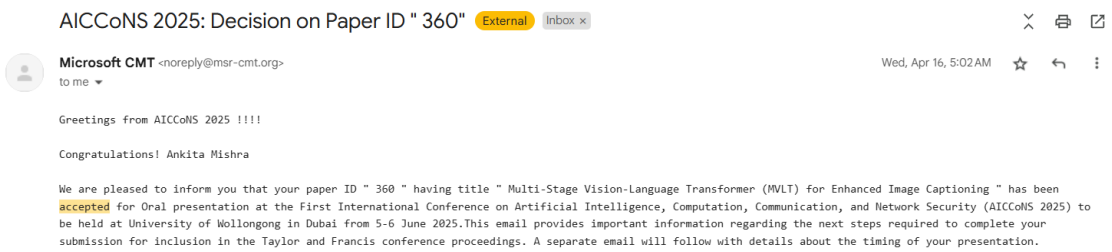
[38] Y. Feng and M. Lapata, “Topic models for image annotation and text illustration,” in *Proc. Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL)*, 2010, pp. 831–839.
[39] G. Kulkarni, V. Premraj, S. Dhar, S. Li, Y. Choi, A. C. Berg, and T. L. Berg, “Babytalk: Understanding and generating simple image descriptions,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 12, pp. 2891–2903, Dec. 2013.
[40] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhutdinov, R. Zemel, and Y. Bengio, “Show, attend and tell: Neural image caption generation with visual attention,” in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2015, pp. 2048–2057.
[41] A. Karpathy and L. Fei-Fei, "Deep Visual-Semantic Alignments for Generating Image Descriptions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 3128–3137.
[42] J. Johnson, R. Krishna, M. Stark, L.-J. Li, D. Shamma, M. Bernstein, and L. Fei-Fei, "Image retrieval using scene graphs," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2015, pp. 3668–3678.
[43] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," *International Conference on Learning Representations (ICLR)*, 2015.
[44] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the Inception Architecture for Computer Vision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2016, pp. 2818–2826.
[45] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2016, pp. 770–778.
[46] A. Dosovitskiy et al., "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," in *Proc. Int. Conf. Learn. Representations (ICLR)*, 2021.
[47] A. Radford et al., "Learning Transferable Visual Models From Natural Language Supervision," in *Proc. 38th Int. Conf. Mach. Learn. (ICML)*, 2021, pp. 8748–8763.
[48] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhutdinov, R. Zemel, and Y. Bengio, "Show, Attend and Tell: Neural Image Caption Generation with Visual Attention," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2015, pp. 2048–2057.
[49] S. J. Rennie, E. Marcheret, Y. Mroueh, J. Ross, and V. Goel, "Self-Critical Sequence Training for Image Captioning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2017, pp. 1179–1195.
[50] K. Xu et al., "Show, Attend and Tell: Neural Image Caption Generation with Visual Attention," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2015, pp. 2048–2057.
[51] A. Karpathy and L. Fei-Fei, "Deep Visual-Semantic Alignments for Generating Image Descriptions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2015, pp. 3128–3137.
[52] D. Bahdanau, K. Cho, and Y. Bengio, "Neural Machine Translation by Jointly Learning to Align and Translate," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2015.
[53] A. Dosovitskiy et al., "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2021.
[54] X. Zhai et al., “Scaling Vision Transformers,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2022, pp. 12104–12113.
[55] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, “BLEU: a Method for Automatic Evaluation of Machine Translation,” in *Proc. ACL*, 2002, pp. 311–318.
[56] R. Vedantam, C. Lawrence Zitnick, and D. Parikh, “CIDEr: Consensus-Based Image Description Evaluation,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2015, pp. 4566–4575.
[57] S. Banerjee and A. Lavie, “METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments,” in *Proc. ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, 2005, pp. 65–72.
[58] H. Liu et al., “LLaVA: Large Language and Vision Assistant,” *arXiv preprint arXiv:2304.08485*, 2023.

[59] C. Schuhmann, R. Beaumont, R. Vencu, C. Kaczmarczyk, C. Mullis, A. Schmitt, T. Kornuta, and A. Thomee, “LAION-400M: Open dataset of CLIP-filtered 400 million image-text pairs,” *arXiv preprint arXiv:2111.02114*, 2021
[60] F. Schuhmann et al., "LAION-400M: Open Dataset of CLIP-Filtered 400 Million Image-Text Pairs," *arXiv preprint arXiv:2111.02114*, 2021.

LIST OF PUBLICATION(S)

1. Ankita Mishra, Dr. Prashant Giridhar Shambharkar, “Multi-Stage Vision-Language Transformer (MVLT) for Enhanced Image Captioning”. The paper has been Accepted First International Conference on Artificial Intelligence, Computation, Communication, and Network Security (AICCoNS 2025).Indexed by IEEE.Paper Id:360

.



AICCoNS 2025: Decision on Paper ID " 360" External Inbox ×

Microsoft CMT <noreply@msr-cmt.org> Wed, Apr 16, 5:02 AM

to me

Greetings from AICCoNS 2025 !!!!

Congratulations! Ankita Mishra

We are pleased to inform you that your paper ID " 360 " having title " Multi-Stage Vision-Language Transformer (MVLT) for Enhanced Image Captioning " has been accepted for Oral presentation at the First International Conference on Artificial Intelligence, Computation, Communication, and Network Security (AICCoNS 2025) to be held at University of Wollongong in Dubai from 5-6 June 2025.This email provides important information regarding the next steps required to complete your submission for inclusion in the Taylor and Francis conference proceedings. A separate email will follow with details about the timing of your presentation.