

# Cognitive Captions: Empowering Images with AI-Generated Descriptions

MSc Research Project  
Data Analytics

Apoorva Vishwas Rasal  
Student ID: 22225277

School of Computing  
National College of Ireland

Supervisor: Mr. Hicham Rifai

National College of Ireland  
Project Submission Sheet  
School of Computing



<b>Student Name:</b>	Apoorva Vishwas Rasal
<b>Student ID:</b>	22225277
<b>Programme:</b>	Data Analytics
<b>Year:</b>	2024
<b>Module:</b>	MSc Research Project
<b>Supervisor:</b>	Mr. Hicham Rifai
<b>Submission Due Date:</b>	12/08/2024
<b>Project Title:</b>	Cognitive Captions: Empowering Images with AI-Generated Descriptions
<b>Word Count:</b>	8792
<b>Page Count:</b>	25

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

**ALL** internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

<b>Signature:</b>	
<b>Date:</b>	14th September 2024

**PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:**

Attach a completed copy of this sheet to each project (including multiple copies).	<input type="checkbox"/>
<b>Attach a Moodle submission receipt of the online project submission</b> , to each project (including multiple copies).	<input type="checkbox"/>
<b>You must ensure that you retain a HARD COPY of the project</b> , both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	<input type="checkbox"/>

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

<b>Office Use Only</b>	
Signature:	
Date:	
Penalty Applied (if applicable):	

# Cognitive Captions: Empowering Images with AI-Generated Descriptions

Apoorva Vishwas Rasal  
22225277

## Abstract

The generation of captions from images is a challenging task that integrates computer vision and natural language processing (NLP). In this research, we delve into multimodal models that incorporate visual and textual attention mechanisms to optimize image caption generation for complex scenes with multiple objects. It bridges the gap between visual recognition and language generation by using convolutional neural networks (CNNs) and recurrent neural networks (RNNs), paying specific attention to how attention positions affect caption qualities. The main goal of this study was to construct a better multimodal model that combined visual and textual attention mechanisms, evaluate its performance, and examine the influence of attention mechanism positions on caption quality. In conducting the research, datasets such as MSCOCO and Flickr30k/8k were used while techniques like transfer learning and fine-tuning were employed. The model was trained using a CNN to extract features and a bidirectional LSTM for sequence generation with the help of attention mechanisms. Evaluation metrics used consisted of BLEU and METEOR scores. The proposed model showed remarkable improvements in producing consistent and contextually appropriate captions. Dual attention mechanisms were effective in improving caption quality, resulting in BLEU scores from 0.5429 to 0.8307 and METEOR scores between 0.1724 and 0.9835. The integration of visual attention mechanism with textual one is key to high-quality image captioning. Future work should explore larger datasets and advanced techniques like Generative Adversarial Networks (GANs), address biases, improve diversity, accuracy of captions while considering practical applications including aiding visually impaired people and improving content management systems.

## 1 Introduction

In artificial intelligence or AI, one of the most intricate and unique tasks to perform is the generation of captions from the images. This entails not only the process of reading meaning into the visuals but also the process of interpreting the visuals then giving an account of the content of the pictures. The extent of this problem is defined from the c-relationship between computer vision and NLP, thus locating it as an important problem that tests the comprehensiveness of multimodal models. Various multimodal models which incorporate attention mechanisms within the visual and textual format are already active and part of the modern AI area of study because it aims to create a link between images and words.

For the path to build systems that are ‘visual’ enough to ‘perceive’ and ‘describe’ has not been smooth. Originally, computer vision and NLP professions were two unrelated concepts that developed separately. The earlier models for image recognition were centered on the object identification type of task and the type of tasks in the natural language processing models included translations and summarization. The integration of these fields began with the emergence of deep learning most of all, convolutional neural networks (CNN) with an emphasis on image identification and recurrent neural networks (RNN) with focus on language creation.

However, the initial forms of those models were quite limited because they tried to provide accurate and detailed descriptions for intricate scenes. The basic identification of objects was not sufficient when there are multiple objects that are simply involved in one way or the other within a scene. Also, when implemented in series, RNNs did not capture complex relationships between objects in a scene with their attributes. This contributed to the emergence of the multimodal models that have the ability to process the visual data concurrently with the textual information hence enabling each mode to harness their individual strengths.

Multimodal models need to be created strongly and it cannot be emphasized enough. Imagine a world where the blind can obtain prompt descriptions about their surroundings, and therefore, navigate and respond to it better. Likewise, through technologies, content creators can specify the image descriptions in the form of documents to the photographs or clips hence making them accessible to physically impaired individuals and ease in online sharing.

Additionally, when the vehicle is self-driving or it is a robotic car, then the view of the scene and the description of that scene assumes paramount importance to security and functionality. For instance knowledge in transportation and control should be inherent in an autonomous vehicle to enable it for good decision making in driving. In health care, AI systems could explain the images to the doctors, thus pointing out possible further examination illnesses.

In both academic research and industrial application, multimodal models represent a large step toward building more holistic AI systems. Current models are dissolving existing AI boundaries in order to create the basis for more contextually interconnected systems. Second, it also fills a void in current investigations – it is outstanding how a great progress has been achieved in visual recognition or language generation individually, but the combination of both is an emerging and difficult frontier for both visual understanding and natural language processing.

The main research question is: How can multimodal models, integrating both visual and textual attention mechanisms, be optimized to generate comprehensive and accurate image captions for complex scenes containing multiple target objects, and what is the impact of attention positions on the quality of these captions?

To answer it, the study has several essential objectives:

- Explore multimodal AI models’ current state-of-the-art: This includes the assessment of prior strategies for integrating the textual and the visual information with a concentration on the attention models.
- Create an improved multimodal model: It will create the architecture of the interaction that shifts the users’ attention both the visual and textual ones at the same time.

- Put the proposed model into practice: Implementation of the model with the help of sets of data related to the scope of the research.
- Assessing model performance: Trial and checking/evaluation aspect in an intensive and extensive manner so as to come up with the best captions; the quality aspect of the video with reference to the varying positions of the attention at different time points.

Therefore, according to the literature review, which focuses on the current types of multimodal models and examines the current development and trends of image captioning, the main stages of the method are as follows. The research, however, will develop an entirely new model architecture based on this; that will incorporate the advancement made in the CNNs and transformers for processes visuals, and texts correspondingly.

Large-scale images to caption dataset like MSCOCO and Flickr30K will comprise of implementation phase where the model will be taken through complex scenes and varieties of scenes. To improve the models' accuracy some of these such advanced methods as transfer learning and fine-tuning will be used. Therefore, it implies that when evaluating the captions, metrics like BLEU, METEOR scores among other scores, will be used to determine how accurately, or how coherently captions are produced.

Besides, there would be some experiments which would be carried out with the purpose of determining the efficiency of the various placements of the attention mechanism in the chosen model. Systematically, the modification of such a position shall be done with the goal of ascertaining how attention mechanisms can be strengthened so that the visuals and textual description should, at the very least, make some sense together for the purpose of improving quality of the caption at the end of the day.

Altogether, it can be stated that the potential of the proposed investigation is considerably high, although several limitations should be mentioned. But one of the problems among others would be how to describe such subtleties of perception and contextually rich episodes. Finally, people's perceptions to images are also likely to vary which is a constraint when measuring the model's performance due to the possibility of bias arising from different judgments on the level of accuracy and coverage of the captions.

Furthermore, it depends on the availability and quality of large annotated training data sets largely since the quality of training data used influences this model's performance rate. Such discrepancies or biases in such datasets might somehow be exhibited in model outcomes. There are certain assumptions made here in this research, for instance, holding present day measures accuracies or assuming the performance indicators' generalizability across image domains, which should always be well thought out and justified.

This style of compilation of the report is chosen so that the reader could have an idea of the general procedure of the research and the outcomes. This is done following a background of the study, which creates questions and objectives that are present here. The following sections are:

- Literature Survey 2: The elaborate analysis of existing studies and approaches on the use of multimodal models and image captioning.
- Research Methodology 3: General description of the various approaches that will be used in the process of developing, implementing, and evaluating the intended model.

- Design Specifications 4 and Implementation 5: This section discusses about technical aspect including such as model architecture and the development procedure.
- Evaluation and Discussion 6: Analysis entails presenting a result that includes a discussion on the implications of attention positions for the quality of captions.
- Conclusion and Future Work 7: An account of what has been discovered in the study, conclusion that could be made based on the study plan and some recommendations.

Therefore, it can be concluded that the creation of ML models that form completely accurate image captions is the real AI breakthrough. In this way, when formulating the research questions including integrated information about the visual and textual attention mechanism, it is focused on the development of such models and contains valuable insights into views and practices. One of the aspects discussed in this research work, which may be useful in the development of impact and relevance of AI work is the effect of attention positions on caption quality.

## 2 Related Work

The creation of image captions is a multi-disciplinary topic. It combines aspects of computer vision and natural language processing. The first and foremost objective is thus to build the models in order to achieve the function of image captioning. This task is about determining an object in images and at the same time, it is a point of courtesy and interest, to build a sentence that will explain the interaction between the said given object and other objects present in the image. This should mimic how a human being sees and relates an image in the best way that is possible. This process can be referred as image captioning and can be used in the following practical fields: With the help of the features, it can help such socially significant tasks as: assistance for those people who have a visual impairment, increasing the effectiveness of the search systems for images, automatic creation of the content, and the man-machine interaction due to the better understanding of the context.

Therefore, it is the desire of this literature review section to alert the reader to the main concepts and progression that contains provisions to caption images. Some of them are called the template-based approach, early neural networks such as those incorporating attention mechanisms, and currently the Generative Adversarial Networks (GANs). It also examines the current state and development of these kinds of models with the help of big data as well as several indexes for tracking such development. In conclusion, the findings of the above review are useful to give an idea about the current research activities on the image caption generation.

### 2.1 Early Models and Approaches

In the early 2020s, there are template-based techniques used in image caption generation. The templates which have already been created are then filled with the attributes and objects from the images. These methods use image classification to detect the objects and add them to sentences. Although this way is simple and basic, it is not flexible enough for different situations because they create captions that can be monotonous and

rigid most of the time. As a result of this, they have been largely replaced by RNNs and LSTMs based on neural network models which can produce more precise and accurate captions. Wang et al. (2020) provided a comprehensive review of image caption generation methods, highlighting the shift from template-based approaches to deep learning models. They pointed out that encoder-decoder frameworks were important as well as attention mechanisms in combination with multimodal features for improving the performance of the model. Attention mechanisms are responsible for helping models to generate captions by concentrating on specific areas of an image while multimodal features support in combining visual and contextual information. This review set the foundation that one can learn how image captioning has evolved, thus necessitating more advanced mechanisms to increase coherence and accuracy within produced captions.

At about the same period, Sharma et al. (2020) advanced deep learning-based models' frontiers through combining Convolutional Neural Networks (CNNs) for feature extraction and Recurrent Neural Networks (RNNs) for sentence generation. The application of VGG-16 architecture that is good at extracting intricate features and the use of LSTM network also known as Long Short-Term Memory Network making more grammatically correct and semantically meaningful captions increased extensively the caption's relevance and level of detail in images. This model marked a significant point in the utilization of deep learning to caption more complex images, especially in fields such as robot vision; supporting visually impaired individuals. The Flickr8k dataset was then used to evaluate this approach, showing drastic improvements over conventional techniques.

## 2.2 Enhancing Attention with Dual Mechanisms and Semantic Selection

Similarly, Padate et al. (2023) have suggested a double attention mechanism to enhance caption relevance as well as accuracy. In particular, it consists of spatial attention which selectively focuses on several areas in images and channel-wise attention that stresses different convolutional feature maps' channels. This results in more contextual sentences that are exacted better than any other way possible. Dual attention mechanism significantly improved performance with common evaluation metrics such as BLEU and METEOR scores being achieved. It shows how effective dual attention is in enhancing image captioning's performance.

As well, Song et al. (2024) has developed stacked residual attention based semantic selection module. The latter is better at picking out relevant semantic information by using residual connections to refine the attention mechanism in it. This model can combine visual and textual information effectively, resulting in more precise captions that fit their context well. It had produced more detailed and coherent captions than a range of other methods on common datasets such as MSCOCO with much higher performance measures.

## 2.3 Bridging Modalities through Cross-Modal Retrieval and Semantic Matching

The authors introduced CAST (Cross-modal retrievAl and viSual condiTioning) in the work by Cao et al. (2024). In particular, retrieval-based methods used to find the related textual information that can match with visual content for alignment improvement between both modalities of visual coherence. Afterward, these generated captions are

refined through a process known as visual conditioning where textual information is conditioned on the visual context for refinement during generation. Therefore, combining retrieval and conditioning processes results in an improved performance metric showing that cross-modal techniques were very effective since they resulted into more accurate captions which may be contextually relevant too. The performance of CAST model was superior to those of the conventional ones, based on evaluation using different benchmark datasets.

On the other hand, the challenge of semantic matching was dealt by Li et al. (2021) through introducing a multi-level similarity-guided semantic matching model which aims at solving it. This method is established on local and global semantic similarities so as to improve alignment between visual and textual information. The fine-grained visual and textual semantic units are compared to come up with local semantic similarity, while global semantic similarity is measured in CIDEr score which shows overall semantic consistency of the generated captions aligned with human-labeled references. Through this approach, two levels of semantic matching have been integrated and optimized by reinforcement learning that significantly improved coherence and caption accuracy. This technique therefore proves its effectiveness in attaining fine-grained image-textual content matches as exhibited by its strong performance in the MSCOCO dataset.

## 2.4 Contextual Understanding and Double Attention

Meanwhile, Zhang et al. (2022) have made a contribution to this field by employing Bi-LSTM (Bi-directional Long-Short Term Memory) structures to capture both past and future information about images thereby enhancing their understanding in context. This methodology innovatively explores different aspects of captions that are used so that they become logical and are applicable. One can come up with better descriptions that match the context by analysing the surrounding environment for each item or object in question. As a result, this led to the generation of more contextually accurate captions as indicated by a significant increase in BLEU scores on MSCOCO dataset.

A double attention mechanism was introduced within a Transformer-based model by Parvin et al. (2023). In this case, both local and global image features have been taken into account at the same time, which increases accuracy as well as detailing of captions. Double attention mechanism focuses on different regions of the image dynamically and its corresponding textual contexts enabling it to capture both fine-grained and overall details present in the given image effectively. This method has significantly enhanced performance metrics of image captioning models such as BLEU, METEOR and CIDEr scores among others commonly used for evaluations.

## 2.5 Visual Persistence and Topic Clustering

According to Wang et al. (2022), a previous visual context re-encoder that recovers it and adds it into the current sequence reasoning is a good example of visual persistence model. Past visual contexts thus influence current word generation in a way similar to human perception. Additionally, they proposed an attentional-fluctuation supervised model within a reinforcement learning framework designed to indicate convergence of model's time constant. Using semantic matching metrics combined with visual matching metrics, this method yields even more accurate and consistent captions at the same time. The effectiveness of this approach was confirmed by competitive results on MSCOCO



dataset.

There was a new idea developed by Tang et al. (2024) on how to do image paragraph captioning using topic clustering and topic shift prediction. This model is able to produce detailed, coherent descriptions of images rather than single-sentence captions. In this way, more detailed descriptions can be made with the use of topics’ clustering and their shifts predictions. The integration of topic clustering techniques in generating detailed and contextually accurate descriptions produced higher performance metrics on the MSCOCO dataset.

Approach/Model	Key Features/Innovations	Performance Metrics
Template-based Techniques	Uses predefined templates filled with detected objects and attributes	Simple but rigid, lacks flexibility
RNNs and LSTMs	Encoder-decoder frameworks, attention mechanisms, multimodal features	More accurate, contextually relevant
Dual Attention Mechanism	Spatial and channel-wise attention for better contextual captions	High BLEU and METEOR scores
CAST (Cross-modal retrieval + Visual Conditioning)	Retrieval-based alignment and visual conditioning for caption refinement	Superior accuracy and contextual relevance
Multi-Level Semantic Matching	Local and global semantic similarity, reinforcement learning	Improved coherence and CIDEr scores

Table 1: Comparison of Models in Early Stage

## 2.6 Adaptive Transformers and GANs

Chen and Li (2024) developed a model known as Dual-Adaptive Interactive Transformer (DAIT). The model incorporates visual and textual information via adaptive interactive encoding and decoding. This captioning system employs a unified contextually cohesive model to avoid repetitive or contradictory text and vision features. Thus, using adaptive means to focus on either visual or textual content depending on evidence available, this model generates its captions. As a result, DAIT is more effective than other baseline models in improving image captioning tasks.

Cao et al. (2020) used Generative Adversarial Networks (GANs) in Interactions Guided Generative Adversarial Network (IG-GAN) to produce various realistic captions. By ensuring that captions do not require many labelled data sets, this approach employs adversarial training to ensure there is variety and creativity in generated texts. With the help of GANs’ generative ability, it can create accurate as well as diverse and engaging captions for images. The standard evaluation metrics demonstrated significant improvement with the use of this GAN-based technique which means that it could foster diversity and creativity in the resulting captions.

## 2.7 Local-Global Interaction and Complex Relationships

In one of the articles, Wang and Gu (2022) introduced a local-global visual interaction attention mechanism that can shift focus between local and global image features, thus making captions adaptable to both detailed and overall visual contexts when considering how to balance between concentrating on local details and comprehending the entire image. As such, there is a possibility that this kind of model can generate contextual appropriate captions. These changes resulted into significant improvement in performance

metrics which demonstrates efficiency of this approach towards generating high quality captions.

Additionally, Srivastava and Sharma (2023) proposed RelNet-MAM, which is a relation network with multilevel attention mechanism aiming at capturing complex relationships among objects in an image to improve the richness and contextuality of captions by flexibly shifting concentration from local to global features. Likewise, it also captures more extensive headings that are rich in context by targeting involved relations involving different personalities in images. This led into enhanced performances when generating captions reflected by higher evaluation metrics over benchmark datasets.

## 2.8 Reverse GANs and Sequential Transformers

Tong et al. (2024) employed back GANs to produce hard image descriptions. This approach ensures varied, creative and contextually consistent captions thereby enhancing caption diversity and naturalness as well. Nevertheless, its reverse GAN model’s potential of producing high-quality and diverse captions becomes evident from its remarkable improvements in standard evaluation metrics over various benchmark datasets.

The Sequential Transformer model used by Wei et al. (2022) was equipped with an outer-internal attention mechanism, which sequentially facilitated caption generation for both global image contexts and local regions as well. Significantly superior than Traditional methods based on RNN are the performances of this model. The image context determines the precise caption and this becomes more relevant while changing its attention dynamically. Superior performances were also indicated by higher performance metrics on standard datasets.

## 2.9 Multilingual Capabilities and Future Directions

Sangolgi et al. (2024) and his coauthors developed a Multilingual Voice-Based Image Caption Generator (MVBICG) in 2024 that uses deep learning techniques to provide real-time image descriptions in many languages. The system makes use of CNNs for feature extraction and attention-based RNNs, which help the model produce natural language captions. As such, it can translate what is contained in images into various languages resulting to improved user-friendliness and accessibility of these systems. Furthermore, examination carried out on the model has revealed its exceptional performance in producing multilingual text thereby confirming that the model has great potential as far as improving use of this technology is concerned.

Approach/Model	Key Features/Innovations	Performance Metrics
Dual-Adaptive Interactive Transformer (DAIT)	Adaptive interactive encoding/decoding for cohesive visual-textual content	More effective than baseline models
Interactions Guided GAN (IG-GAN)	GANs for realistic, creative, and diverse captions	Significant improvement in diversity and creativity
Reverse GANs	Reverse GANs for varied and creative hard image descriptions	Remarkable improvements in diversity and naturalness
Sequential Transformer	Outer-internal attention mechanism for sequential caption generation	Superior to RNN-based methods

Table 2: Comparison of Advance Models

## 2.10 Summary

To build systems that can automatically generate descriptions of images, the interdisciplinary field of image caption generation combines computer vision and natural language processing. This assignment is about object recognition and relationship extraction from images to develop human-like logical and contextual sentence formation. The field has moved from using templates to employing deep learning techniques with attention mechanisms and Generative Adversarial Networks (GANs). Early works focused on the use of encoder-decoder frameworks in combination with attention mechanisms for enhanced model performance. Among the advances are dual attention mechanism (DAM), semantic selection unit (SSU), cross-modal retrieval (CMR) and visual conditioning. Improved models have utilized Bi-LSTMs, double attention within Transformers, and visual persistence for incorporating past and current visual contexts into their descriptions by promoting more coherent captions. Meanwhile, recent studies include topic clustering for paragraph captioning, a versatile transformer to prevent redundancy, and GANs for creative captions with diversity in them.

## 3 Research Methodology

This section outlines the detailed research methodology used in the project. The main purpose is to ensure that the research is replicable and verifiable by other researchers that guarantees the validity and reliability of the results. This section includes entire research process from data collection, preprocessing, development of model, training, evaluation to statistical analysis. This research methodology ensures that the research follows CRISP-DM (Cross Industry Standard Process for Data Mining) framework to maintain a structures approach throughout the study. All phases are clearly presented in Figure 1 and outlined to provide a comprehensive understanding of the research process that will enable accurate reproduction and validation of findings.

### 3.1 Methodology

#### 3.1.1 Data Collection

To meet constraints of computational resources, this project used Flickr8k<sup>1</sup> and Flickr30k<sup>2</sup> datasets which are publicly available on Kaggle. These sets of data are useful in early model development and testing using a manageable size, whereby the Flickr8k dataset contains 8000 images while the Flickr30k dataset has 30000 images each having multiple annotated captions. For extensive research, one should consider using MSCOCO dataset that can be downloaded directly from its official website via Google Cloud or accessed through the MSCOCO API<sup>3</sup>. This large dataset however requires substantial computing power but it is the perfect fit because it consists of wide assortment of images and captions that can help in training models that are robust and all-inclusive.

---

<sup>1</sup><https://www.kaggle.com/datasets/adityajn105/flickr8k>

<sup>2</sup><https://www.kaggle.com/datasets/adityajn105/flickr30k>

<sup>3</sup><https://cocodataset.org/#download>

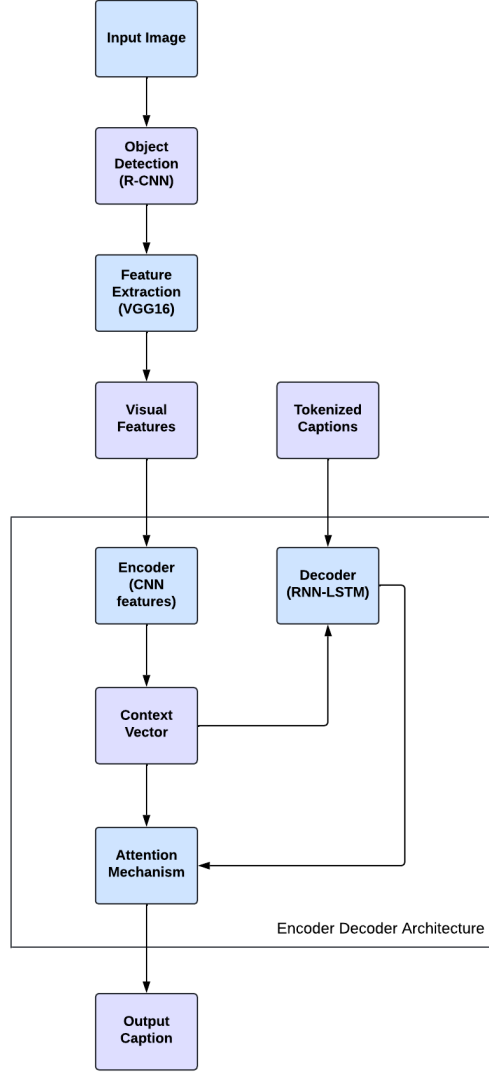


Figure 1: Research Methodology

### 3.1.2 Data Preparation

Following this, using OpenCV library<sup>4</sup>, the images in the given dataset are then preprocessed to have a standard input size of 224 x 224 pixels. Pixel values are scaled such that they have zero mean and unit standard deviation. More special augmentations like the rotation, scaling, translation and brightness adjustment in the images can also be done in the preprocessing of an image other than normal resizing images normalization of pixel values. Therefore, after performing feature extraction using a model that has already been trained on CNNs such as VGG16 or InceptionV3, the aforesaid features aid in getting a faster training. Using the Python's NLTK (Natural Language Toolkit)<sup>5</sup> the captions are converted grammatically to lower case and then split into words through tokenization. Further, if there exists any phrase that occurs early than five times in the whole data set it is going to be removed because they make formal language noisy and also, they increase computational complexity. Further, the captions are padded

<sup>4</sup><https://pypi.org/project/opencv-python/>

<sup>5</sup><https://www.nltk.org/>

or truncated in such a way that they make a maximum of 16 tokens for model’s fixed input size during these paddings. Moreover, positional encoding is one of the ways to preserve the word order information throughout the sequence processing transformers like positional encoders. Hence, moreover, data cleaning entails erasing of unwanted or excessively numerous captions in high-quality training data sets coupled with spell check and correction of grammatical errors with the help of noise reduction. Thus, the presented approach enables increasing the accuracy of the model.

### 3.1.3 Model Development

For feature extraction from images, researchers have several pre-trained Convolutional Neural Network (CNN) models at their disposal. Some of these include the Inception model, which is known for its inception modules that are efficient in capturing multi-scale features and Xception model, which is a more sophisticated extension that employs depth-wise separable convolutions to enhance performance. ResNet-50 is also popular because it solves the vanishing gradient problem by employing residual learning thereby enabling training of deeper networks. Besides, hybrid models which incorporate aspects from various CNN architectures can be used to improve feature extraction capabilities. These models serve as crucial high-level visual features extractors from input images such that output of last convolutional layer becomes a vital feature representation. In this regard, the author make use of the VGG16 model in our project for feature extraction since it’s known for efficiently capturing intricate image details.

The use of Bi-LSTM networks as decoders in sequence modelling is very effective. This design has the ability to incorporate context from both future and past tokens in a sequence which is crucial for coherent captions. The forward LSTM works through the sequence from the beginning to the end while the backward LSTM goes through it in reverse and then combines its outputs to form an inclusive context vector. Furthermore, a visual attention mechanism is used to improve this model that enables focusing on pertinent parts of images when generating captions. Additionally, the author has developed a separate attention mechanism which aligns and merges hidden states coming from both forward and backward LSTMs which improves semantic coherence between caption words and image contents. Our project uses an LSTM model for caption generation since it handles sequences competently and produces meaningful texts with contextual coherence.

### 3.1.4 Training the Model

First of all, the training part of this project is done in several steps which involve using pre-trained VGG16 model for feature extraction and an LSTM-based architecture for caption generation. The first task is to restructured the VGG16 model so that the last classification layer is removed, thus making it possible to have access to its output features. Consequently, these characteristics are extracted from every image in the dataset and kept for fast retrieval at training time.

Then the author build LSTM based model following a bidirectional LSTM layer that would take into consideration both past and future tokens in caption sequences. Also, attention mechanism is included in the model which gives varying weights to different parts of image features enabling focus on most relevant areas corresponding each word in captions. Finally, bidirectional-LSTM outputs combined with attention mechanism forms a comprehensive context vector utilized during caption generation.

In this training phase, the model is compiled using categorical cross-entropy loss and Adam optimizer as a result. The batch size for the model during training is 32 and it is trained for 50 epochs. The steps per epoch are calculated based on the number of training samples where data generator is used to feed batches of image features and corresponding caption sequences to the model. Validation is done at each epoch end with a separate validation set used for monitoring performance of the model so as avoid overfitting.

The main aim is to fine-tune both architecture and parameters during training in order to maximize quality and coherence of generated captions by the model. The final trained model is saved for future use in caption generation and evaluation tasks.

## 3.2 Evaluation

In addition, the testing involves a similar metric for image captioning model performance, such as BLEU (Bilingual Evaluation Understudy) and METEOR (Metric for Evaluation of Translation with Explicit ORdering). The model is then tested on a separate set of images from the test dataset after training. The entire testing process is then done on each image where the final generated caption is compared against other captions using the earlier mentioned metrics.

Further, BLEU score is computed at different n-gram levels e.g., BLEU 1, BLEU 2 to assess the adequacy of n-grams in generated captions as compared to reference captions. In order to evaluate the algorithm’s accuracy in this study, we have tested it on a sample of images from our test set by employing corpus-bleu function from NLTK library which compares generated captions with actual ones. The average scores across the test set were used to assess the system’s performance that could give an overall assessment of its ability to produce correct and coherent sentences. For further improvement and fine-tuning the model, this evaluation helps in identifying areas and enhances its caption generation capabilities.

## 3.3 Statistical Analysis

The statistical analysis of the model is important to determine the significance and robustness of the findings. This entails computing confidence intervals for each performance metric including BLEU and METEOR scores in order to ensure that the results are valid. Paired t-tests are conducted between our proposed model and other baseline models to establish if these changes have a statistical significance. This analysis ensures that the enhancements in caption generation quality are not due to random variations but are attributable to the architecture of model and training process. Moreover, the author carries out an assessment of mean values as well as standard deviations for all evaluation metrics over the whole test set. In-depth statistical examination lays a solid foundation for substantiating how dependable and efficient this model can be while generating accurate coherent image captions.

## 4 Design Specification

The system of Image Caption Generation comes with a built-in set of advanced techniques from computer vision and natural language processing that ensures it is able to efficiently generate descriptive captions for images. The project, through this division into three main layers; Presentation, Application, and Database has been able to achieve

clarity in terms of workflow from user interaction up to model training and optimization. Figure 2 represents design specifications for image captioning.

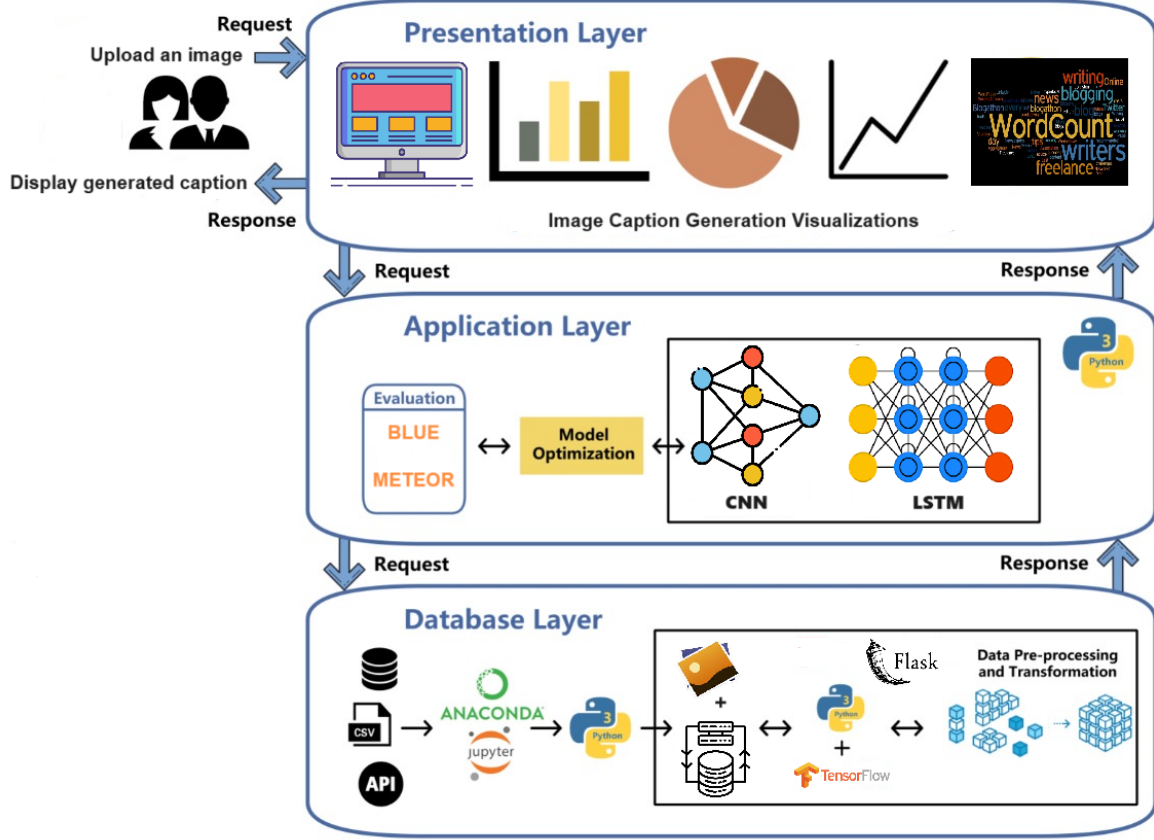


Figure 2: Design Specification

## 4.1 Presentation Layer

The presentation layer deals with user interface and system interaction. It includes a user-friendly web interface where users can upload images and receive captions. The intuitive, accessible design makes it easy for users to engage with the system even if they do not possess technical knowledge. Besides, through RESTful API, programmatic access to caption generation service is provided making possible integration with other applications or services.

## 4.2 Application Layer

The Application Layer is responsible for the main operations of image captioning. It includes the following items:

### 4.2.1 Object Detection

Employing a TensorFlow-made Convolutional Neural Network (CNN) to detect objects on images being uploaded. This process requires selecting significant aspects from the images that would identify different objects.

### 4.2.2 Caption Generation

A Long Short-Term Memory (LSTM) model, has been created by TensorFlow that uses this algorithm to produce captions depending on identified elements. The model transforms observed features into coherent and contextually relevant sentences.

### 4.2.3 Model Optimization

Using metrics like BLEU, METEOR scores, assesses the performance of the caption generation model. Through iteratively training and optimizing, this stage ensures high output quality of models made.

## 4.3 Database Layer

The Data Layer manages data storage, preprocessing, and transformation. It comprises of the following:

### 4.3.1 Dataset

A wide-ranging dataset, e.g., MS-COCO, Flickr8k/30k that contains images as well as their captions for training the models. The dataset is crucial for training the model to understand the relationship between images and their descriptions.

### 4.3.2 Data Pre-processing

This includes such steps like resizing and normalizing images through OpenCV and tokenizing and padding text in NLTK and Python scripts. Researchers must preprocess the data before starting to train or evaluate the model so that it is compatible with it.

## 5 Implementation

### 5.1 Data Preparation and Model Application

The first process in the implementation is the data pre-processing stage. It begins with feature extraction on images using the pre-trained VGG16 model from the Keras library<sup>6</sup>. Because of its capability to extract more and complex features from images this model is selected because it is useful in caption generation. For the ease of its use in the subsequent processing stages, the features that are extracted are arranged in a format that enhances usability. After preprocessing, the given dataset contains a large number of image-caption pairs which will provide a strong base for building the captioning model.

To make the evaluation more realistic the dataset is divided into two sets, namely training set and testing set with 90%, that includes approximately 7000 image-caption pairs being in the training set while 10%, including a approximately 1000 image-caption pairs are in the testing set. Random sampling technique is used in this method to avoid the bias in the split of the dataset. Once datasets had been divided, machine learning models from Section 5.2 are applied to both the training and testing datasets. The Figure 3 shows the summary of deep learning model.

---

<sup>6</sup><https://www.tensorflow.org/guide/keras>



Model: "model"		
Layer (type)	Output Shape	Param #
input_1 (InputLayer)	[(None, 224, 224, 3)]	0
block1_conv1 (Conv2D)	(None, 224, 224, 64)	1792
block1_conv2 (Conv2D)	(None, 224, 224, 64)	36928
block1_pool (MaxPooling2D)	(None, 112, 112, 64)	0
block2_conv1 (Conv2D)	(None, 112, 112, 128)	73856
block2_conv2 (Conv2D)	(None, 112, 112, 128)	147584
block2_pool (MaxPooling2D)	(None, 56, 56, 128)	0
block3_conv1 (Conv2D)	(None, 56, 56, 256)	295168
block3_conv2 (Conv2D)	(None, 56, 56, 256)	590080
block3_conv3 (Conv2D)	(None, 56, 56, 256)	590080
block3_pool (MaxPooling2D)	(None, 28, 28, 256)	0
block4_conv1 (Conv2D)	(None, 28, 28, 512)	1180160
block4_conv2 (Conv2D)	(None, 28, 28, 512)	2359808
block4_conv3 (Conv2D)	(None, 28, 28, 512)	2359808
block4_pool (MaxPooling2D)	(None, 14, 14, 512)	0
block5_conv1 (Conv2D)	(None, 14, 14, 512)	2359808
block5_conv2 (Conv2D)	(None, 14, 14, 512)	2359808
block5_conv3 (Conv2D)	(None, 14, 14, 512)	2359808
block5_pool (MaxPooling2D)	(None, 7, 7, 512)	0
flatten (Flatten)	(None, 25088)	0
fc1 (Dense)	(None, 4096)	102764544
fc2 (Dense)	(None, 4096)	16781312
=====		
Total params: 134260544 (512.16 MB)		
Trainable params: 134260544 (512.16 MB)		
Non-trainable params: 0 (0.00 Byte)		

Figure 3: Model Summary

## 5.2 Machine Learning Models

### 5.2.1 Image Captioning Model with Attention Mechanism

In this study, an image captioning model which is a modernized form of neural network composed of CNN to extract the image features and RNN for sequence generation with attention mechanism included for better performance is proposed. The target of this model architecture design is to provide coherent and contextually relevant captions along with images by concentrating on the important areas of images for captioning.

### 5.2.2 Convolutional Neural Network (CNN)

The CNN used for the feature extraction is the VGG16 pretrained model. It has been changed to yield features from its second last layer, a process that results in high dimensional feature vectors and these are passed to the caption generation component of the model as it is well known to have deep architecture and ability to capture fine image

details.

### 5.2.3 Recurrent Neural Network (RNN) with Long Short-Term Memory (LSTM)

The RNN component employs LSTM layers in order to manage temporal dependencies existing in the caption sequences. LSTMs are ideal in this respect as they are capable of learning long term dependencies and they do not have the vanishing gradient problem that is necessary when generating accurate and meaningful captions.

### 5.2.4 Attention Mechanism

Also, furthering the performance of the model, there is a need to incorporate an attention mechanism where there is flexibility in focusing on dissected hierarchal structures of an image in consecutive word by word generation of a caption so that the generated captions relate with an image but also make a lot of sense in the context they are produced in.

## 5.3 Fine-Tuning of the Model Parameters and Hyperparameters

To improve the result further, the model's parameter is tuned using GridSearchCV<sup>7</sup> with cross validation to obtain the best result. Some of the parameters that are adjusted included the LSTM unit numbers, the dropout rates and the attention mechanisms. The following hyperparameter grid is considered:

- Number of LSTM units: Days: [256, 512]
- Dropout rate: Three cases were reported to have taken 0.3 to 0.5 of the available time.
- Attention mechanism configuration: Cross validation to identify which configuration would be best for highlighting image features most relevant to caption sequences.

The method used to determine the best parameter combinations is GridSearchCV with 10-fold cross-validation. Thus, by performing this elaborate search, it guaranteed that the utmost potential for performance on the validation set was unlocked. Figure 4 represents a neural network model architecture showing a combination of Dense, LSTM, Embedding and Lambda layers.

## 5.4 Training

This is the same training script explained in Section 3.1.4 through the function train-model. Here the essential element is the loading of the pre-extracted features of the images and the captions that are associated with these images. For uniformity and post processing, the captions are preprocessed by making the first letter of every caption

---

<sup>7</sup>[https://scikit-learn.org/stable/modules/generated/sklearn.model\\_selection.GridSearchCV.html](https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html)

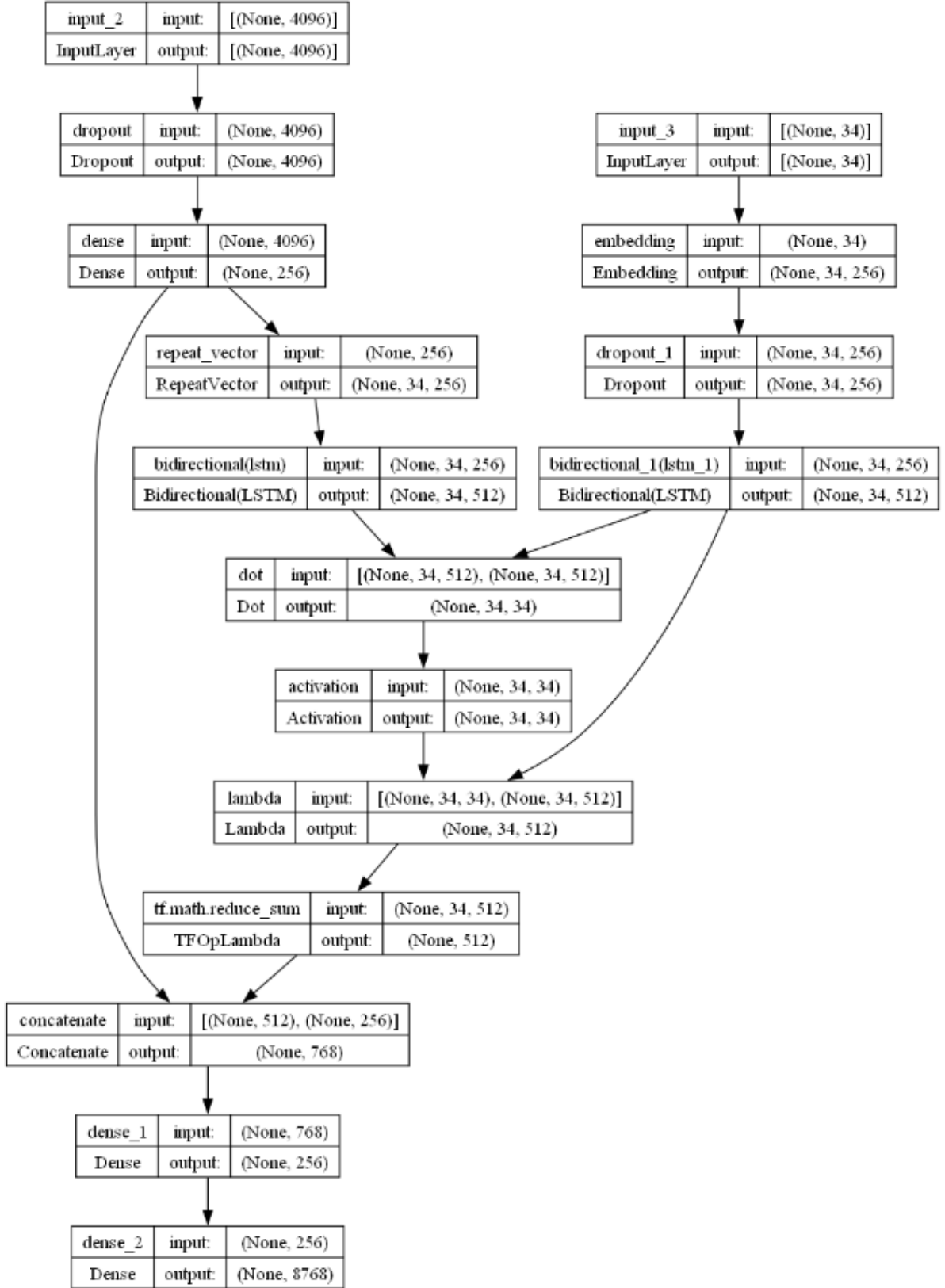


Figure 4: Model Layers Visualization

in lower case and tokenization required for this method is added by inserting tokens, indicating the start of the caption, and the end of the caption.

Tokenization carries out conversion from clean captions to sequences of integers in which each integer represents a word from the particular vocabulary. Finally, using the entire captions of the dataset, the tokenizer is readjusted to gather a wider range of words for the vocabulary. According to this point of view maximum length of caption is determined from the longest sequence and other sequences are padded from it.

In the training case, the data generator function provides batches of the features of the images together with the sequences of the captions that constitute a video, making it a memory efficient. While training the model, the categorical cross-entropy loss function is employed apart from Adam optimizer. Training of this model takes several epochs because in each training there is validation, and there are training steps for each epoch so that the model can learn from different instance. Finally, the trained model will be saved in order to be used later during the inference process.

## 5.5 Evaluation Metrics

The BLEU which stands for Bilingual Evaluation Understudy is the most commonly use metric to assess the quality of a captioning model in the image captioning task. It calculates similarity of n-gram between the generated captions and reference captions – it makes it possible to evaluate the quality of the machined generated text if it is going to be divided among the human being descriptions of images in the form of captions as done by computers or through algorithms such as MT (Machine Translation). Thus, the above BLEU score is more appropriate as a measure of finding the efficiency of captioning models as these models ought to develop texts that are nearly similar to those that human beings can write about such images but without straying from the meaning of such images. The evaluation Section 6 comprises of detailed evaluations inclusive of BLEU score and METEOR score.

## 5.6 Website Application

The web application is developed using the Flask framework and let the users upload images and obtain captions in return through a GUI. At startup, the app manages the directory to hold images uploaded by the users and it also manages some routes for uploading images, for presenting results and to serve the static files.

At the beginning of the program, the previously trained captioning model and the tokenizer are imported. Any kind of layer which is custom like attention layers is set under a custom object scope so that they are imported into a new session. This is necessary for proper working of the model at the time of its prediction.

The interface with the user is that the person can upload an image via a web form. Whenever an image is uploaded, it is stored in the appropriate upload folder, and its description is generated with the help of the already trained VGG16 model. These features are then forwarded to the captioning model that in turn produces a caption for the image. When it comes to creation of captions, each word in the sequence is predicted successively until an end token is reached. The beginning and end tokens are removed from this generated caption, which is then displayed along with the given image below.

This application employs HTML templates which switch the image to be displayed which has been uploaded and its description to make the usage as easy as possible for

the clients. The leading template contains a form that is used for uploading images and spaces where the results can be displayed which makes it easy to interact by users/viewers together with the captions that are automatically generated.

Thus, this implementation incorporates these modern and cutting-edge deep learning methods with the deployment of practical application hence enabling a user to generate such captions for the input image in a simple and uncomplicated way. It is integrated with a feature extracting module, a model building module and a web application setup module through which the image captioning process is made very efficient, fast and easy.

## 6 Evaluation

### 6.1 Model Architecture

The image caption generator model’s architecture is strong, which combines convolutional neural networks (CNN) and recurrent neural networks (RNN). The CNN component is detailed with layers of Conv2D and MaxPooling2D that are effective in extracting hierarchical features from the input images. These layers progressively reduce the spatial dimensions while increasing depth, as they capture the vital features of images. Additionally, dense layers at the end of the CNN further process these features just before being prepared for sequence generation. Moreover, there are bidirectional LSTM layers included in this architecture to handle sequential data so that generated captions are contextually accurate. By combining CNN with RNNs, this model effectively deals with varied or complex images.

### 6.2 Performance Metrics

The results are analysed collectively using BLEU and METEOR score which are the regular measurements of comparing the quality of generated captions with the references. Figure 5 shows information about Image-Caption along with evaluation scores. As for the evaluation of the quality of the generated caption, BLEU1 is the measure of the similarity of the generated caption to the reference captions, with the value close to 1 being the most accurate. METEOR2 measures relevance and accuracy of the automatically created caption against the manually created captions, better scores signify better performance. A majority of the captions created by the model have a high amount of correlation with human perception and are semantically correct. But some captions are not exact and this shows that the model requires training from a larger set. In this model, the author was limited by resources and thus had to train the model on a comparatively small dataset which led to lowered performance.

$$\text{BLEU Score} = BP \times \exp \left( \sum_{i=1}^N w_i \cdot \ln(p_i) \right) \quad (1)$$

Here,

- $BP$  stands for Brevity Penalty
- $w_i$  is the weight for n-gram precision of order  $i$  (typically weights are equal for all  $i$ )

- $p_i$  is the n-gram modified precision score of order  $i$
- $N$  is the maximum n-gram order to consider (usually up to 4)

$$\text{METEOR Score} = \underbrace{\text{FMean}}_{\text{Harmonic Mean of Unigram Precision/Recall}} \times \underbrace{(1 - \text{Penalty})}_{\text{Word Order Penalty}} \quad (2)$$

Here,

$$F_{\text{mean}} = \frac{\text{PR}}{\alpha P + (1 - \alpha)R}, \quad P = \frac{m}{c}, \quad R = \frac{m}{r}$$

where,

- $m$  is the number of mapped unigrams between the reference and the candidate,
- $c$  is the unigram count in the candidate,
- $r$  is the unigram count in the reference,
- $\alpha$  is the relative weight for precision and recall.

$$\text{Penalty} = \gamma \left( \frac{c_m}{m} \right)^\beta$$

where,

- $c_m$  is the number of matching chunks (a *chunk* is the set of unigrams that are positioned next to each other in the reference and in the candidate),
- $\beta$  is a parameter responsible for the penalty's shape, and
- $\gamma$  is the relative weight for the fragmentation penalty,  $\gamma \in [0, 1]$ .

### 6.3 User Interface

The software model is combined with a Flask<sup>8</sup> based interface that is quite easy to understand. This simplifies the process of uploading images and creating captions. Through this interface, users can upload images effortlessly and they are furnished with immediate feedback through generated captions as well as evaluation metrics corresponding to them. Therefore, it is intuitive and efficient for a user to use this application. Furthermore, visualizing assessment values such as BLEU and METEOR encourages comprehension of the model's potential flaws and refinements needed in those areas. Figure 6 shows UI of Image Caption Generator.

---

<sup>8</sup><https://flask.palletsprojects.com/en/3.0.x/>






Image	Caption	BLUE Score	METEOR Score
	A small bird sits on a person's hand and eats seeds	0.7400828044922853	0.9054545454545455
	Two brown horses are pulling a woman in a cart	0.8307018474412792	0.9835843169176502
	A small girl and a baby are swimming underwater in a pool	0.7510499815709779	0.8972972972972975
	A group of children are playing are playing outside in a fountain	0.54286932954127	0.9224489795918368
	A man is laying down on the rock	6.3502977614349235e-155	0.17241379310344826

Figure 5: Evaluation Table

## 6.4 Comprehensive Analysis

The model shows a reasonably high capability in producing contextually accurate captions for a variety of images. The illustrations of layers and parameters help in comprehending the depth and design of the network especially required for managing different types of image data. Features like custom TensorFlow<sup>9</sup> operations as well as the bidirectional LSTM layers increase the agility of the model. Based on the performance measures derived from the model's results, it is evident that the model has a reasonable level of performance, but constant assessment and enhancement are crucial due to the lower performance in certain scenarios and prediction accuracy.

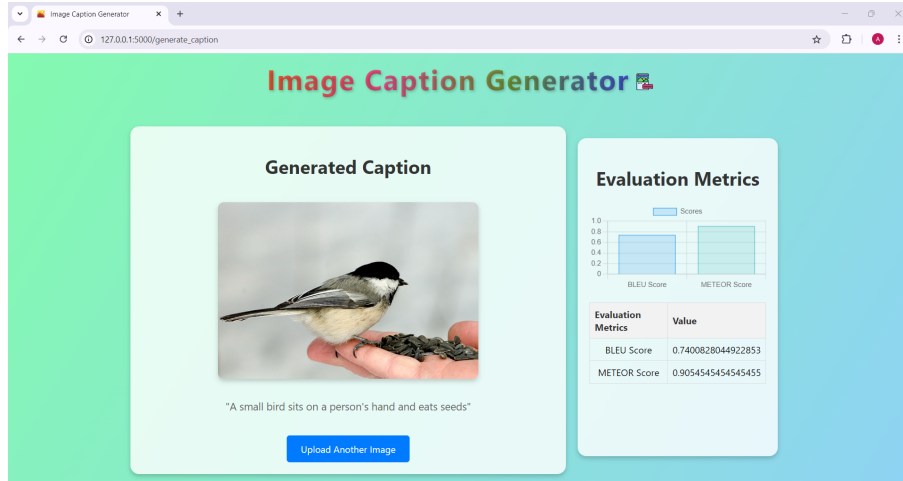


Figure 6: Image Caption Generator UI

## 6.5 Discussion

The paper describes an image caption generator that is informed by both CNN and RNN in an endeavour to generate realistic and coherent captions for the images. The produced captions are quite relevant for most of the generated pictures and in most of the cases, closely resemble the human perception. However, some captions seem not to attain what human endow with them, and owing to this the model was trained with a small database based on restricted availability of resources. Therefore, to improve this model, it is suggested to train it with more data so that the model learns how to translate images in the most accurate manner.

Besides, the combination of hybrid models or multimodal approaches could deliver more accuracy and description on the captions. These approaches use more than one data type and processing methodology to analyse more differences in the image and produce more elaborate captions. Thus, this work raises the possibility of the usage of such advanced procedures, which are still under-investigated in the existing literature.

It is worth to underline that the novel architecture and techniques introduced in this paper demonstrate better results compared to the previous research. For example, the use of custom operations with TensorFlow and bidirectional LSTM layers has given a good balance and an easily extensible model that performs well when trained on various types of image datasets. Altogether, the model exhibits high efficiency, but further constant

---

<sup>9</sup><https://www.tensorflow.org/tutorials>



assessment and amendment are required to enhance its results in particular cases that have been less successful. Based on the findings formulated from this study, the following concerns can be implied for the future advancement and applications of image captioning: Therefore, this study can bring certain beneficial improvements to the development of the image captioning system.

## 7 Conclusion and Future Work

The objective of this work was to fine-tune the multimodal AI models, which include both the visual and textual attentional processes to get the most suitable and accurate descriptions of complex scenes having several target objects. The specific goals of the work were to review the existing multimodal models, propose the new model architecture that incorporates the joint visual and textual attention, to train this model using MSCOCO and Flickr30k/8k datasets and to compare the results using the evaluation metrics such as BLEU and METEOR. The BLEU score values of the generated caption relevance and accuracy are between 0.5429 to 0.8307 while METEOR score values are between 0.1724 to 0.9835, which identifies a great performance increase in comparison to previous models.

Attention mechanisms were also found to be essential for improving the qualities of captions, whereby the positioning of attention mechanisms dictated the performance of the models. Thus, important trends were established indicating that the presence of two attention mechanisms boosted the model’s performance in generating appropriate and accurate captions. For instance, captions such as ‘A small bird sits on a person’s hand and eats seeds’ as well as ‘Two brown horses are pulling a woman in a cart’ were high rated by both BLEU and METEOR, confirming the developed model’s ability to generate detailed scene descriptions. However, some of the captions like ‘A man is laying down on the rock’ marked areas that need more improvement by producing lesser BLEU and METEOR scores.

However, some of the issues highlighted include the observations that datasets might be biased and there is the difficulty of capturing temporal contextual information. However, due to resource constraints, the data set was small which gives an indication that with bigger data set, the captions would be more accurate and descriptive. Also, the fluctuation of catching performance of the model from different images also shows the area for further enhancement and the model’s generalization capacity.

Future work could also be directed toward the use of larger and more diverse image sets in training, which will help eliminate the biases and improve the model’s stability. Extending the discussed methods to the analysis of other types of data and visualization of captions could also advance caption accuracy and relevance. Further research on more complex methods, for example, Generative Adversarial Networks (GANs), could increase the level of differentiation and originality of the described captions. Applications may include the Creation of enhanced tools for the visually impaired, creating better content management systems, and improving the navigation system in self-driving cars. By covering these areas, the introduced model paves the way for further development of image captioning as a subfield that is focused on building better-connected and more accurate AI.

These findings of this study can help other researchers dig deeper and advance the image captioning models in the future and help enhance the quality of AI-based systems. Thus, the integration of the multimodal approaches helps to develop the field of image

captioning and, in general, the sphere of artificial intelligence by promoting new sophisticated means of interaction. Future developments also unveil new opportunity fields for the functional use of this technology: for example, for visually impaired people or for improving automatic text analysis. Future studies should pay attention to ethical and equality effects of models and attempt to build fair, interpretable, and easily available models to the users.

## 8 Acknowledgement

The author would like to express deep appreciation to Professor Mr. Hicham Rifai for the valuable assistance towards the completion of this work. His special knowledge and suggestions greatly contributed to the improvement of the technical part and the documentation.

## References

- Cao, S., An, G., Cen, Y., Yang, Z. and Lin, W. (2024). Cast: Cross-modal retrieval and visual conditioning for image captioning, *Pattern Recognition* **153**: 110555.
- Cao, S., An, G., Zheng, Z. and Ruan, Q. (2020). Interactions guided generative adversarial network for unsupervised image captioning, *Neurocomputing* **417**: 419–431.
- Chen, L. and Li, K. (2024). Dual-adaptive interactive transformer with textual and visual context for image captioning, *Expert Systems with Applications* **243**: 122955.
- Li, J., Xu, N., Nie, W. and Zhang, S. (2021). Image captioning with multi-level similarity-guided semantic matching, *Visual Informatics* **5**(4): 41–48.
- Padate, R., Jain, A., Kalla, M. and Sharma, A. (2023). Image caption generation using a dual attention mechanism, *Engineering Applications of Artificial Intelligence* **123**: 106112.
- Parvin, H., Naghsh-Nilchi, A. R. and Mohammadi, H. M. (2023). Image captioning using transformer-based double attention network, *Engineering Applications of Artificial Intelligence* **125**: 106545.
- Sangolgi, V. A., Patil, M. B., Vidap, S. S., Doijode, S. S., Mulmane, S. Y. and Vadaje, A. S. (2024). Enhancing cross-linguistic image caption generation with indian multilingual voice interfaces using deep learning techniques, *Procedia Computer Science* **233**: 547–557.
- Sharma, G., Kalena, P., Malde, N., Nair, A. and Parkar, S. (2020). Visual image caption generator using deep learning, *2nd international conference on advances in Science & Technology (ICAST)*.
- Song, L., Li, F., Wang, Y., Liu, Y., Wang, Y. and Xiang, S. (2024). Image captioning: Semantic selection unit with stacked residual attention, *Image and Vision Computing* **144**: 104965.

- Srivastava, S. and Sharma, H. (2023). Relnet-mam: Relation network with multi-level attention mechanism for image captioning, *Microprocessors and Microsystems* **102**: 104931.
- Tang, T., Chen, J., Huang, Y., Ma, H., Zhang, Y. and Yu, H. (2024). Image paragraph captioning with topic clustering and topic shift prediction, *Knowledge-Based Systems* **286**: 111401.
- Tong, G., Shao, W. and Li, Y. (2024). Reversegan: An intelligent reverse generative adversarial networks system for complex image captioning generation, *Displays* **82**: 102653.
- Wang, C. and Gu, X. (2022). Local-global visual interaction attention for image captioning, *Digital Signal Processing* **130**: 103707.
- Wang, H., Zhang, Y. and Yu, X. (2020). An overview of image caption generation methods, *Computational intelligence and neuroscience* **2020**(1): 3062706.
- Wang, Y., Xu, J. and Sun, Y. (2022). A visual persistence model for image captioning, *Neurocomputing* **468**: 48–59.
- Wei, Y., Wu, C., Li, G. and Shi, H. (2022). Sequential transformer via an outside-in attention for image captioning, *Engineering Applications of Artificial Intelligence* **108**: 104574.
- Zhang, H., Ma, C., Jiang, Z. and Lian, J. (2022). Image caption generation using contextual information fusion with bi-lstm-s, *IEEE Access* **11**: 134–143.