# Deep Learning-Based Image Captioning and Audio Generation with Flickr8K and Flickr30K Datasets

**Israth Jahan**

**ID: 2020-1-60-044**

**Litaz Saif**

**ID: 2020-1-60-002**

**Sheikh Reshma Sultana**

**ID: 2020-1-60-140**

**Farhana Moni**

**ID: 2020-2-60-036**

**A Capstone project report submitted in partial fulfillment of the requirements for the degree of Bachelor of Science in Computer Science and Engineering**

**Department of Computer Science and Engineering**
**East West University**
**Dhaka-1212, Bangladesh,**
**December 2024**

# Declaration

We, **Israth Jahan, Litaz Anwar Saif, Sheikh Reshma Sultana, and Farhana Moni,** hereby declare that the work presented in this capstone project report is the outcome of our investigation under the supervision of Supervisor **Mahmuda Rawnak Jahan**, Lecturer, Department of Computer Science and Engineering, East West University. We also declare that no part of this project has been or is being submitted elsewhere for any degree or diploma award, except for publication.

| | |
|---|---|
| Countersigned | Signature |
| ..................... | ................................ |
| **Mahmuda Rawnak Jahan** | **Israth Jahan (2020-1-60-044)** |
| Supervisor | |
| | Signature |
| | ................................. |
| | **Litaz Anwar Saif (2020-1-60-002)** |
| | Signature |
| | ................................ |
| | **Sheikh Reshma Sultana (2020-1-60-140)** |
| | Signature |
| | ................................ |
| | **Farhana Moni (2020-2-60-036)** |

# Letter of Acceptance

The capstone project report entitled "**Deep Learning-Based Image Captioning and Audio Generation with Flickr8K and Flickr30K Datasets"**, submitted by, **Israth Jahan, Litaz Anwar Saif**, **Sheikh Reshma Sultana**, and **Farhana Moni** to the Department of Computer Science and Engineering, East West University, Dhaka, Bangladesh is accepted for partial fulfillment of the requirement for the degree of Bachelor of Science in Computer Science and Engineering on December 31, 2024.

Board of Examiners

**1.** ______________________________

**Mahmuda Rawnak**
**Jahan**
**Lecturer**

Department of Computer Science and Engineering
East West University

**2.** ______________________________

**Dr. Maheen Islam**

Chairperson and Associate Professor

Department of Computer Science and Engineering
East West University

# Abstract

This thesis explores a comparative analysis of deep learning models for the tasks of image captioning and audio generation, utilizing models including MobileNetV3 + LSTM, VGG16 + LSTM, and ResNet50 + LSTM. Image captioning, a subset of machine learning, enables the automatic generation of textual descriptions for visual inputs. Additionally, translating these captions into audio speech enhances accessibility, enabling a seamless experience for visually impaired users. This research's primary objective is to evaluate these models' performance in generating accurate, coherent, and contextually aligned captions and corresponding audio outputs for a range of image datasets. We assess each model's performance using objective metrics—BLEU 1, BLEU 2, BLEU 3 and BLEU 4 scores—to measure caption quality and precision. Furthermore, qualitative assessments provide insights into fluency, alignment, and overall descriptive quality. An accompanying web application was developed to facilitate image upload, caption generation, and audio synthesis, thereby enhancing the practical applicability of the research. The findings indicate that each model exhibits distinct strengths in handling complex scenes. This analysis contributes to advancing the field of image captioning by identifying optimal model configurations and integrating audio generation, with implications for accessibility and human-computer interaction design.

# Acknowledgment

I would like to express my deepest gratitude to my advisor, **Mahmuda Rawnak Jahan**, for her invaluable guidance, support, and encouragement throughout the development of this thesis. Her expertise in machine learning and deep learning has been a cornerstone of my thesis journey, providing me with critical insights and direction. Ms. Jahan's constructive feedback, thoughtful suggestions, and unwavering belief in my abilities have been instrumental in shaping the direction and depth of this work. I am truly grateful for her mentorship, patience, and dedication, which have not only enhanced my academic skills but also instilled a passion for continuous learning.

I extend my heartfelt thanks to the faculty members and staff at the Department of Computer Science and Engineering at East West University, Bangladesh. Their commitment to creating an enriching and supportive environment for research and learning has been invaluable. The engaging discussions, collaborative atmosphere, and access to various resources have greatly facilitated my academic growth. I would also like to acknowledge the contributions of my fellow students, whose diverse perspectives and collective enthusiasm for knowledge have made this experience more rewarding.

A special note of appreciation goes to my friends and colleagues who offered their insights, motivation, and assistance during this journey. Their camaraderie and support were vital in overcoming challenges, brainstorming ideas, and maintaining a positive outlook. The memories created during our collaborative efforts and discussions will always hold a special place in my heart.

To my family, I cannot express enough gratitude for your unwavering support, patience, and understanding throughout my academic pursuits. Your belief in my abilities has been a constant source of strength and encouragement. Thank you for being my greatest cheerleaders and for standing by me through every challenge and triumph.

Additionally, I am grateful to the developers and contributors of the open-source libraries and frameworks used in this project, including TensorFlow, PyTorch, and others. Their innovative work and dedication to making advanced tools accessible have played a significant role in enabling me to bring my ideas to life. The collaborative nature of the open-source community is a testament to the spirit of sharing knowledge and fostering innovation.

Finally, I would like to acknowledge all those individuals who have, in one way or another, supported me throughout this journey. Whether through direct assistance or simple acts of kindness, your contributions have not gone unnoticed. I hope to find a more appropriate way to acknowledge each of you in the future.

Thank you all for your contributions to this accomplishment and for being part of my academic journey.

Israth Jahan

December 2024

Sheikh Reshama Sultana

December 2024

Litaz Saif

December 2024

Farhana Moni

December 2024

# Table of Contents

# List of Figures

# List of Tables

# List of Algorithms

1. Feature Extraction: VGG16

2. Feature Extraction: ResNet50

3. Feature Extraction: MobileNetV3

4. Sequence Generation: LSTM (Long Short-Term Memory)

5. Evaluation Metric: BLEU (Bilingual Evaluation Understudy)

# List of Acronyms

| | | |
|---|---|---|
| 1. | gTTS | Google Text-to-Speech |
| 2. | ML | Machine Learning |
| 3. | AI | Artificial Intelligence |
| 4. | LSTM | Long Short-Term Memory. |
| 5. | CNN | Convolutional Neural Network |
| 6. | RNN | Recurrent Neural Network |
| 7. | API | Application Programming Interface |
| 8. | HTML | HyperText Markup Language |
| 9. | CSS | Cascading Style Sheets |
| 10. | NLP | Natural Language Processing |
| 11. | BLEU | Bilingual Evaluation Understudy |

# Chapter 1

# Introduction

This research examines the rapid advancements in artificial intelligence and machine learning, specifically in the area of image captioning, where machines create descriptive captions for images. It compares several multimodal deep learning models, including VGG16, ResNet50 and MobileNetV3, each using different strategies for feature extraction and caption generation. The study aims to evaluate their performance against two standardized datasets Flickr8K and Flickr30K and analyze the audio generation capabilities based on the generated captions, thereby enhancing user engagement and accessibility. Ultimately, this work contributes to the academic discourse in AI and machine learning by offering insights into practical applications, such as assistive technologies and content generation tools.

## 1.1 Background

Artificial Intelligence (AI) and Machine Learning (ML) have significantly transformed various domains, enhancing automation, data analysis, and user engagement. Among the many applications of these technologies, image captioning has emerged as a pivotal area of research, capturing the interest of both academia and industry. Image captioning involves the automatic generation of textual descriptions for images, thereby bridging the gap between visual content and natural language. This capability is essential for various applications, such as improving accessibility for visually impaired individuals, enriching social media platforms, and enabling more intuitive human-computer interactions.

Historically, image captioning methods relied on hand-crafted features and rule-based approaches, which limited their effectiveness and scalability. However, the advent of deep learning has revolutionized this field. Modern image captioning techniques leverage Convolutional Neural Networks (CNNs) to extract high-level visual features from images and employ Recurrent Neural Networks (RNNs) or transformers for natural language generation. This combination allows for a more nuanced understanding of visual content, enabling the generation of coherent and contextually relevant descriptions.

Recent developments, such as the integration of attention mechanisms, have further improved the performance of image captioning models by allowing them to focus on specific areas of an image when generating captions. This has led to enhanced accuracy and fluency in the generated text. Furthermore, the emergence of deep learning approaches, which simultaneously process multiple data types—such as images, text, and audio—has opened new avenues for research. These approaches aim to create more comprehensive models that can understand and describe complex visual scenes in greater detail.

In summary, image captioning represents a convergence of AI, ML, and natural language processing, with the potential to significantly impact various industries, including media, e-commerce, and education. As research continues to evolve, developing more sophisticated models promises to enhance the quality and applicability of automated image descriptions, driving innovation in how we interact with visual content.

## 1.2 Problem Statement and Motivation

With the rapid growth of multimedia content, particularly images, on digital platforms, there is an increasing need for systems that can automatically generate descriptive captions for images. This need extends to various fields, including accessibility for visually impaired users, digital content management, and e-commerce. While substantial progress has been made, current image captioning methods still face challenges in producing accurate, contextually relevant, and human-like descriptions, especially when handling complex scenes with multiple objects or abstract interactions.

Traditional approaches often struggle with understanding finer details within an image, such as identifying relationships between objects, capturing context-specific nuances, and generating fluent, coherent sentences. These limitations affect the practicality of image captioning models in real-world applications, where nuanced and accurate descriptions are essential. Therefore, the motivation behind this research is to develop a robust image captioning model that can enhance the quality of automated captions by leveraging multimodal learning techniques. By integrating models like VGG16, ResNet50 and MobileNetV3, this research aims to create a system capable of delivering improved caption accuracy, contextual understanding, and fluency.

The outcome of this project has the potential to advance applications in accessibility, user engagement, and human-computer interaction, making image content more comprehensible and actionable in diverse environments.

## 1.3 Project Objective

The primary objective of this project is to develop a deep-learning framework for automatic image captioning that enhances caption accuracy, coherence, and contextual relevance. To achieve this, the project will integrate and evaluate different deep learning models—specifically VGG16, ResNet50 and MobileNetV3Large—by analyzing their performance in generating descriptive and contextually accurate captions.

Through comparative analysis, the project aims to identify the strengths and limitations of each model and determine which model or combination of models provides the most effective image-captioning results. Additionally, this project will explore how audio generation can be integrated to produce synthesized speech of generated captions, extending accessibility to visually impaired users. Ultimately, the project seeks to contribute to advancements in image captioning technology, with applications in digital accessibility, content management, and enhanced human-computer interaction.

## 1.4 Project Contributions

This project presents a comprehensive exploration and comparative analysis of multimodal deep learning models—including VGG16, ResNet50 and MobileNetV3Large—specifically designed for image captioning. By leveraging these models, the project evaluates their capabilities to generate contextually relevant and human-like captions, addressing the need for effective automated visual descriptions. An added contribution is an audio generation feature, transforming textual captions into spoken output, making the project especially beneficial for visually impaired users. A web application has been developed to integrate these functionalities, enabling users to upload images and receive both caption and audio responses in real-time.

The project's evaluation includes BLEU 1, BLEU 2, BLEU 3, and BLEU 4 metrics, which provide insights into caption quality and model performance. The findings offer a structured framework for further advancements in image captioning and accessibility applications.

## 1.5 Research Questions

1. How do different deep learning models, such as MobileNetV3, VGG16 and ResNet50 perform in generating accurate and contextually relevant captions for a variety of images?

2. What is the comparative effectiveness of these models in terms of quality metrics such as BLEU1, BLEU2, BLEU3, BLEU4?

3. Can the integration of an audio generation component enhance the accessibility and usability of image captioning systems, particularly for visually impaired users?

4. What are the specific strengths and weaknesses of each model in handling diverse image content, including complex scenes, multiple objects, and varying contexts?

5. How can a web application interface be effectively designed to facilitate user interaction with image captioning and audio feedback in real time.

## 1.6 Project Outline:

The project begins with an explanation of image captioning and audio generation using deep learning models, examining key challenges and current solutions in the field. The project involves a literature review of different image captioning systems implemented by different techniques, tools, and algorithms and previous research in the field, which we can see in **Chapter 2**. Then in **Chapter 3**, the materials and methods include the design and development process, data collection, analysis methods, and the proposed model. The implementation phase focuses on the image captioning features, functionalities, algorithms, and user interface design and demonstrates its capabilities and benefits. In **Chapter 4**, the results and discussion section evaluate the image captioning performance, user feedback, and retention. Lastly, in **Chapter 5**, the conclusion summarizes the project's objectives, methodology, outcomes, contributions, and limitations, and provides recommendations for future research and development.

# Chapter 2

# Related Works

This chapter serves as a comprehensive exploration of the latest developments and research in Image Captioning. Emphasizing significant datasets and cutting-edge techniques that enhance system capabilities. It examines the developments and innovations shaping the landscape of image-captioning models and methods, providing context and foundational insights for this study's comparative analysis.

## 2.1 Survey of the State-of-the-art

In this study, Dessy Santi et al. [1] focus on using Convolutional Neural Networks (CNNs) with residual network architectures—namely ResNet-50, ResNet-101, and ResNet-152—on the Flickr8k dataset. The primary objective is to examine how varying network depths impact the model's ability to comprehend visual structures and contextual details, thereby enhancing the quality of generated image descriptions. The methodology includes stages such as image feature extraction, text preprocessing, and model optimization. Evaluation metrics, including BLEU scores, are employed to assess performance. Experimental results indicate that the ResNet-101 model achieves the highest BLEU score among the models tested, highlighting the efficacy of this approach. This research contributes to ongoing advancements in bridging visual understanding and natural language generation, paving the way for more accurate and contextually relevant image captioning systems.

Sabih Zahra et al. [2] introduce a novel model that combines Convolutional Neural Networks (CNNs) for image feature extraction with a Transformer model for text generation. The aim is to effectively merge visual and textual modalities to produce relevant and semantically coherent image captions. Detailed visual features are extracted from cropped image sections using pre-trained CNNs, such as VGG16 and ResNet50, and are subsequently concatenated. These visual features are then fused with text embeddings from a Transformer model (bert-base-uncased) to achieve a robust multimodal representation. To generate captions sequentially, a Long Short-Term

Memory (LSTM) layer and a dense layer with SoftMax activation are employed to predict each word in succession. The model is trained with categorical cross-entropy loss using the Adam optimizer. Evaluation on the Flickr8k and Flickr30k datasets yielded BLEU1 and BLEU2 scores of 37.533 and 13.8097, respectively, demonstrating the model's ability to produce linguistically accurate and contextually meaningful captions. This study presents an innovative approach to enhancing image captioning systems by integrating CNN-based image feature extraction with Transformer-driven text generation.

Rabin Budhathoki et al. [3] address the limited research and resources available for Nepali-language image captioning by translating and refining the English-language Flickr8k dataset into Nepali, creating a valuable foundation for further study. Recognizing the limitations of traditional RNN-CNN approaches, they employ a streamlined Transformer model that removes the encoder module to reduce complexity, using only the decoder alongside a CNN (MobileNetV3 Large) to efficiently extract and process image features. The model's performance is evaluated with both BLEU and METEOR scores, providing a more comprehensive assessment of caption quality. This research offers an efficient approach to image captioning in under-resourced languages, contributing to multilingual advancements in the field.

Akash Bhadange et al. [4] present an innovative approach to image caption generation using a hybrid neural architecture that combines Convolutional Neural Networks (CNNs) with Long Short-Term Memory (LSTM) networks. This method enhances image captioning by incorporating Word2Vec embeddings, capturing semantic relationships between words in generated captions. The study utilizes the Xception model for robust image feature extraction, advancing state-of-the-art image captioning techniques. Evaluations are conducted on datasets like MS-COCO, rigorously assessing both visual context accuracy and linguistic fluency. Performance metrics, including BLEU, METEOR, CIDEr, and ROUGE, quantitatively measure caption quality. Additionally, the integration of a Text-to-Speech (TTS) feature further enriches the model by enabling audio output of the generated captions.

Ruth-Ann Armstrong et al. [5] explore image captioning on the Flickr 8k dataset, a compact dataset for image captioning, by applying both traditional and state-of-the-art techniques. Various model architectures are tested, including multi-modal paradigms, pretrained encoders, and Transformer decoders with multi-headed attention mechanisms. The top-performing model—a Transformer encoder-decoder network utilizing a pretrained ResNet-18 CNN as the encoder—achieved a BLEU-1 score of 0.879 and a BLEU-4 score of 0.543 on the dataset. This study highlights the effectiveness of Transformer-based architectures in generating accurate and contextually rich captions for image captioning tasks on smaller datasets.

Chaitanya Kulkarni et al. [6] propose a novel framework for combined image captioning and audio generation, leveraging deep neural networks such as Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks, alongside transfer learning techniques. The model operates in two stages: first, it generates captions for a given image, and then it uses Google Text-to-Speech (gTTS) to produce audio for the captions. This framework is particularly beneficial for visually impaired individuals, as it facilitates an auditory understanding of visual content. The model is trained and tested on the Flickr8k dataset, utilizing 6,000 images for training and 1,000 for validation and testing.

K. Kushal et al. [7] examine image captioning techniques using deep learning, focusing on the encoder-decoder framework. This study employs models such as CNN-LSTM, CNN-GRU, Xception-YOLO v4, and a GIT-based model, emphasizing the importance of large, annotated datasets to help algorithms understand the relationship between visual elements and text. In addition to traditional evaluation metrics like BLEU, the study uses METEOR, ROUGE-L, and SPICE to comprehensively cmpare model performance. The findings underscore deep learning's effectiveness in enabling automated caption generation for diverse visual content.

In their research, Rashid Khan et al. [8] introduced a single joint model for automatic image captioning based on CNN and GRU with an attention network. Their study aimed to develop a system utilizing a pre-trained convolutional neural network (CNN) to extract image features, which were then integrated with an attention mechanism. Captions were generated using a recurrent neural network (RNN). To enhance performance, they integrated the Bahdanau attention model with GRU, enabling focused learning on specific portions of the image.

Peter Anderson et al. [9] proposed a combined bottom-up and top-down attention mechanism enabling attention calculation at the level of objects and other salient image regions, which forms a natural basis for attention consideration. In this approach, the bottom-up mechanism, based on Faster R-CNN, suggests image regions, each associated with a feature vector, while the top-down mechanism determines feature weightings. Applying this approach to image captioning, they achieved state-of-the-art results with CIDEr/SPICE/BLEU-4 scores of 117.9, 21.5, and 36.9, respectively. Demonstrating its broad applicability, the same approach applied to the VQA task secured first place in the 2017 VQA Challenge.

In this paper, G. Geetha et al. [10] presented a method for Image Captioning using Deep Convolutional Neural Networks (CNNs). The authors fine-tuned an architecture comprising the encoder of the pre-trained VGG-19 parameters, which had been trained on ImageNet data, along with a GRU (Gated Recurrent Unit) decoder. This approach leveraged the pre-trained features learned by VGG-19 for image classification tasks and adapted them to the image captioning task by fine-tuning the network parameters during training.The proposed method aimed to generate accurate and contextually relevant captions for images.

In the paper by Jyoti et al. [11], Convolutional Image Captioning was presented. In this paper, convolutional image captioning techniques were developed, and a detailed analysis was performed, providing compelling reasons in favor of convolutional language generation approaches. The convolutional (or CNN) approach performed comparably to LSTM (or RNN) based approaches on image captioning metrics. The performance improved with beam search. Adding attention to CNN gave improvements on metrics and outperformed the LSTM+Attn baseline.

Aishwarya et al. [12] focused on VQA: Visual Question Answering which is amenable to automatic evaluation, since many open-ended answers contain only a few words or a closed set of answers that can be provided in a multiple-choice format.The accuracy of the baselines and methods for both the open-ended and multiple-choice tasks on the VQA test-dev for real images.The accuracy of different ablated versions of their best model (deeper LSTM Q + norm I) for both the open ended and multiple-choice tasks on the VQA test-dev for real images.The model trained on filtered version performs worse by 1.13% for open-ended task and by 1.88% for multiple choice task.

In contrast, Md. Zakir et al. [13] proposed an image captioning method utilizing both real and synthetic data for training and testing. They employed a Generative Adversarial Network (GAN)-based text-to-image generator to produce synthetic images and an attention-based image captioning method trained on both real and synthetic images. This approach aimed to describe general traffic targets with more abundant semantic information simultaneously.

In their paper, Aishwarya Maroju et al. [14] proposed an image captioning deep learning model utilizing the ResNet-LSTM architecture. ResNet, known for its convolution layers, was employed to generate captions for given images. The authors concluded that this ResNet-LSTM model exhibited higher accuracy compared to CNN-RNN and VGG models. Furthermore, they noted its efficient performance when run on a Graphic Processing Unit (GPU). This image captioning deep learning model holds significant utility for analyzing large amounts of unstructured and unlabeled data, facilitating tasks such as guiding self-driving cars and aiding visually impaired individuals.

## 2.2 Summary

The body of research on image captioning has evolved significantly, showcasing diverse methodologies and innovative approaches integrating various deep-learning architectures. At the forefront are studies that employ Convolutional Neural Networks (CNNs) alongside recurrent models and Transformers to enhance the quality and contextual relevance of generated captions. For example, Dessy Santi et al. [1] analyzed the impact of different ResNet architectures on the Flickr8k dataset, concluding that deeper networks like ResNet-101 yield a better understanding of visual content and improved BLEU scores. Similarly, Sabih Zahra et al.[2] combined CNNs with Transformers to produce semantically coherent captions, achieving strong performance metrics across multiple datasets.

Several studies also address the challenge of multilingual captioning and the limitations of traditional models. Rabin Budhathoki et al. [3] translated the Flickr8k dataset into Nepali and simplified the model architecture by utilizing a Transformer decoder with MobileNetV3 for efficient image feature extraction. Akash Bhadange et al.[4] introduced a hybrid architecture that incorporates Word2Vec embeddings with CNNs and LSTMs, along with Text-to-Speech capabilities, thereby enhancing accessibility for visually impaired individuals. These efforts reflect a growing emphasis on inclusivity and practical application in image captioning research.

Additionally, attention mechanisms have proven vital in improving model performance, as seen in studies by Peter Anderson et al. [11] and K. Kushal et al.[8], where attention mechanisms enhanced both accuracy and contextual understanding in generating image captions. Overall, these studies collectively contribute to the ongoing evolution of image captioning technologies, paving the way for more sophisticated and user-centric applications across various languages and accessibility needs.

# Chapter 3

# Materials and Methods

In this chapter, we discussed what type of resources we gathered, and the methodology followed to achieve our goal. Here we mentioned what datasets we used, how to preprocess the datasets, which models we trained, and most importantly how we connected the frontend with our backend along with the model.

## 3.1 Materials

Here we discussed the datasets we used to train our model. How we preprocessed captions and extracted features from the image. We also discussed our model and the notebook documents, the necessary libraries to develop the models, and other activities to build our web app.

### 3.1.1 Dataset Acquisition

We use the Flickr 8K [15] and Flickr30K [16] dataset, sourced from Kaggle, was used in this project as a benchmark for image captioning. Comprising 8,091 and 31,783 diverse images with five human-generated captions each, it provides varied scenes, objects, and actions. Each caption offers unique linguistic descriptions for the same image, helping the model learn contextual understanding and accurate language generation. The datasets were organized for training by standardizing captions and preparing images, making it ideal for evaluating our captioning model's performance.

### 3.1.2 Dataset Preprocessing

In dataset preprocessing, images are resized to a specific shape for example, 224x224 for each models and scaled for compatibility with the feature extraction model. Captions are loaded from a text file, split by image ID, and stored in a dictionary, where each ID maps to one or more captions. Captions are tokenized, converted to sequences, and padded to a fixed length to ensure uniformity. This structured data is then ready for input into the model.

### 3.1.3 Feature Extraction

Feature extraction uses a pre-trained convolutional neural network (CNN), like VGG16, MobileNetV3 and ResNet50, to obtain high-level features from images. Each image passes through the CNN, which transforms it into a compact, feature-rich representation. These features act as a summary of the image content, providing essential input data for the captioning model

### 3.1.4 Model Development

In model development, an LSTM-based model is commonly used to generate captions. The model accepts two inputs: the extracted image features and the processed caption sequences. The LSTM learns to map the features and sequences to output the next word in the caption. Training occurs by iterating over the captions and adjusting the model based on prediction errors. After training, the model can generate captions by sequentially predicting words for each image, creating a coherent description.

### 3.1.5 Research Environment and Devices

We used Kaggle as our primary research environment, Kaggle [17] is a cloud-based platform offering free access to powerful computational resources like GPUs and TPUs, pre-installed libraries such as TensorFlow and PyTorch, and an intuitive notebook interface, making it ideal for machine learning and deep learning research. It provides seamless integration with datasets, collaborative features for teamwork, and a supportive community for learning and sharing insights. With tools for version control, participation in competitions, and access to vast datasets, Kaggle eliminates hardware costs and setup complexities, enabling efficient and scalable research in AI and data science.

## 3.2 Methods

Here we discussed our proposed model for this project. We mentioned the framework and algorithm. We showed the architecture how we restructured it and the experimental setup for this project.

### 3.2.1 Proposed Model

This project implements three models for image captioning, each leveraging unique architectural strengths to generate descriptive captions from images. The first model combines **VGG16** with LSTM, where VGG16 serves as the feature extractor, capturing detailed spatial patterns from images, while the LSTM generates coherent text sequences. This combination provides a balanced approach, focusing on spatial detail and efficient sequence generation. The second model integrates **MobileNetV3Large** with LSTM, leveraging MobileNetV3Large's lightweight and efficient design for feature extraction, making it well-suited for applications with resource constraints. The third model employs **ResNet50** with LSTM, using ResNet50's residual connections to capture high-level visual features, ensuring deeper and more detailed feature extraction.

To evaluate these models, **BLEU scores** are utilized to measure the accuracy, fluency, and relevance of the generated captions against the ground truth. By exploring these diverse architectures, the project aims to analyze the trade-offs between accuracy, descriptive richness, and computational efficiency in image captioning tasks.

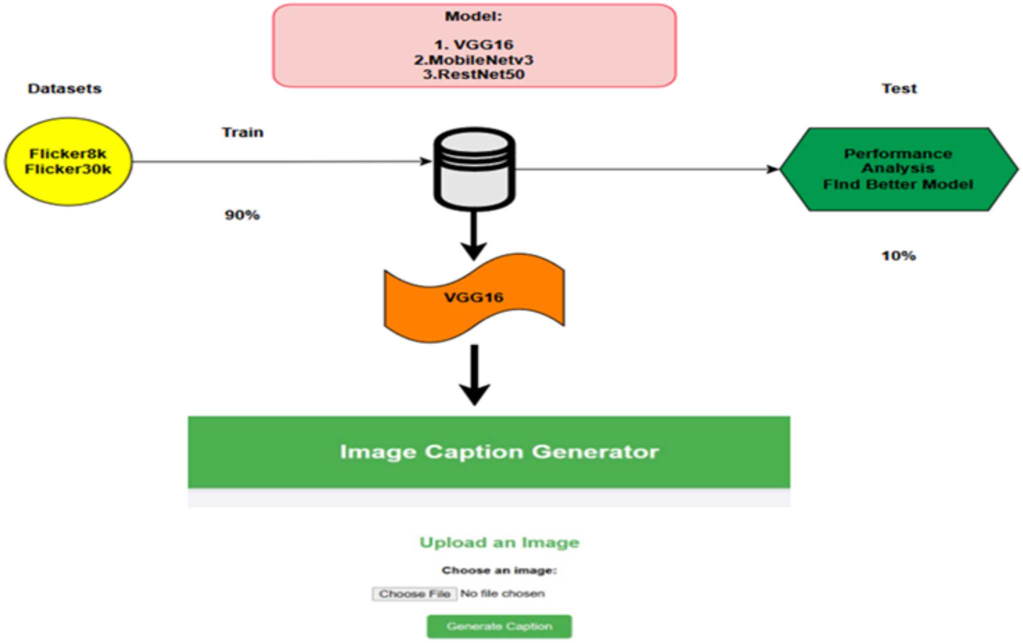


**Fig1: Models Training, Testing, and Web Application Overview**

### 3.2.2 Design/Framework

The design/framework for an image captioning project involves a structured pipeline beginning with data collection and preprocessing, such as cleaning textual captions and resizing images from datasets. The architecture incorporates feature extraction using pre-trained model for visual data and sequence generation through LSTMs or transformer-based networks for textual data. These components are integrated to generate accurate captions, optimized with loss functions like categorical cross-entropy and evaluated using metrics such as BLEU scores. The system features a robust pipeline for seamless input-to-output transformations and can be deployed with an API as the backend and HTML/CSS for the frontend. Continuous iteration through feedback and updates ensures the model remains effective and improves over time.

### 3.2.3 Algorithm/Model Formulation

Our image captioning model leverages pre-trained deep learning architectures for a two-stage process: feature extraction and caption generation. First, VGG16, MobileNetV3Large and ResNet50 extract visual features from each image, condensing complex details into feature vectors. These features are then input into an LSTM-based Recurrent Neural Network (RNN) to generate coherent captions, word by word, in a contextually relevant manner.This combined architecture effectively produces natural language captions that align well with the visual content.

**VGG16:**

VGG16, a widely used CNN, features 16 layers: 13 convolutional layers followed by max-pooling, and 3 fully connected layers. Convolutional layers use 3x3 filters with a stride of 1 and 2x2 max-pooling with a stride of 2. Fully connected layers consist of 4096 neurons each, culminating in an output layer with 1000 neurons for classification. Key operations include convolution, pooling, and matrix multiplication for hierarchical feature extraction.

The convolutional operation:

$$y_{i,j}^{k} = \sigma \left( \sum_{l=1}^{F} \quad \sum_{m=1}^{H} \quad \sum_{n=1}^{W} \quad \omega_{m,n,l}^{k} x_{i+m-1,j+n-1,l} + b_k \right)$$

where $y_{i,j}^{k}$ is the output feature map at position $(i, j)$ and channel k, σ is the activation function, $\omega_{m,n,l}^{k}$ is the weight, $x_{i+m-1,j+n-1,l}$ is the input feature map, and $b_k$ is the bias term. The max pooling operation: $y_{i,j}^{k} = \max\left(x_{2i-1,2j-1}^{k}, x_{2i,2j-1}^{k}, x_{2i-1,2j}^{k}, x_{2i,2j}^{k}\right)$

where $y_{i,j}^{k}$ is the output feature map at position $(i, j)$ and channel k, and $x_{i,j}^{k}$ are the input feature map values at positions $(2i-1,2j-1)$,$(2i, 2j-1)$, $(2i-1,2j), and\ (2i, 2j)$.

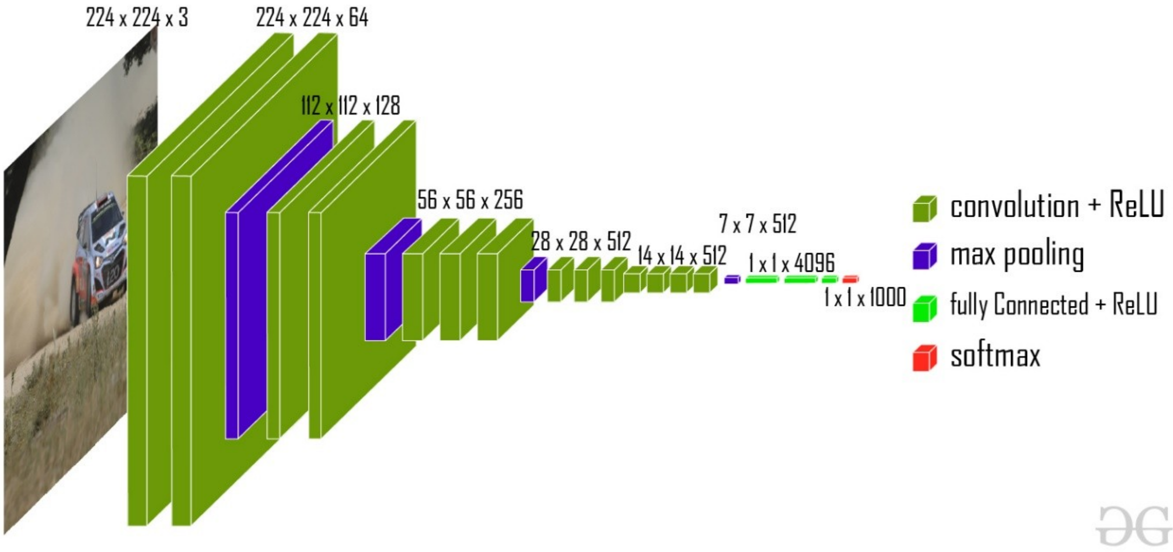


**Fig2: VGG16 Architecture [18]**

**Resnet50:**

ResNet50 stands out as a convolutional neural network design within the ResNet series, distinguished by its 50 layers. Frequently applied in computer vision tasks like image classification and feature extraction, it owes its efficacy to deep architecture and the integration of residual connections. These connections facilitate better optimization and counteract the issue of vanishing gradients. Renowned for its exceptional performance on standard datasets, ResNet50 serves as a robust foundation for cutting-edge deep learning models in image recognition.

The ResNet50 architecture utilizes residual blocks, which can be represented mathematically as:

$$\boldsymbol{y_l} = \boldsymbol{x_l} + \mathrm{F}\ (x_l, W_l)$$

Where:

- $x_l$ represents the input to the lth layer.
- F denotes the residual mapping to be learned by the layer, parameterized by $W_l$ .
- $\boldsymbol{y_l}$ is the output of the lth layer.

The core idea is that the layer learns the residual mapping F, which is added to the input $\boldsymbol{x_l}$ to obtain the output $\boldsymbol{y_l}$ . This allows for the preservation of information from previous layers and helps alleviate the vanishing gradient problem during training.

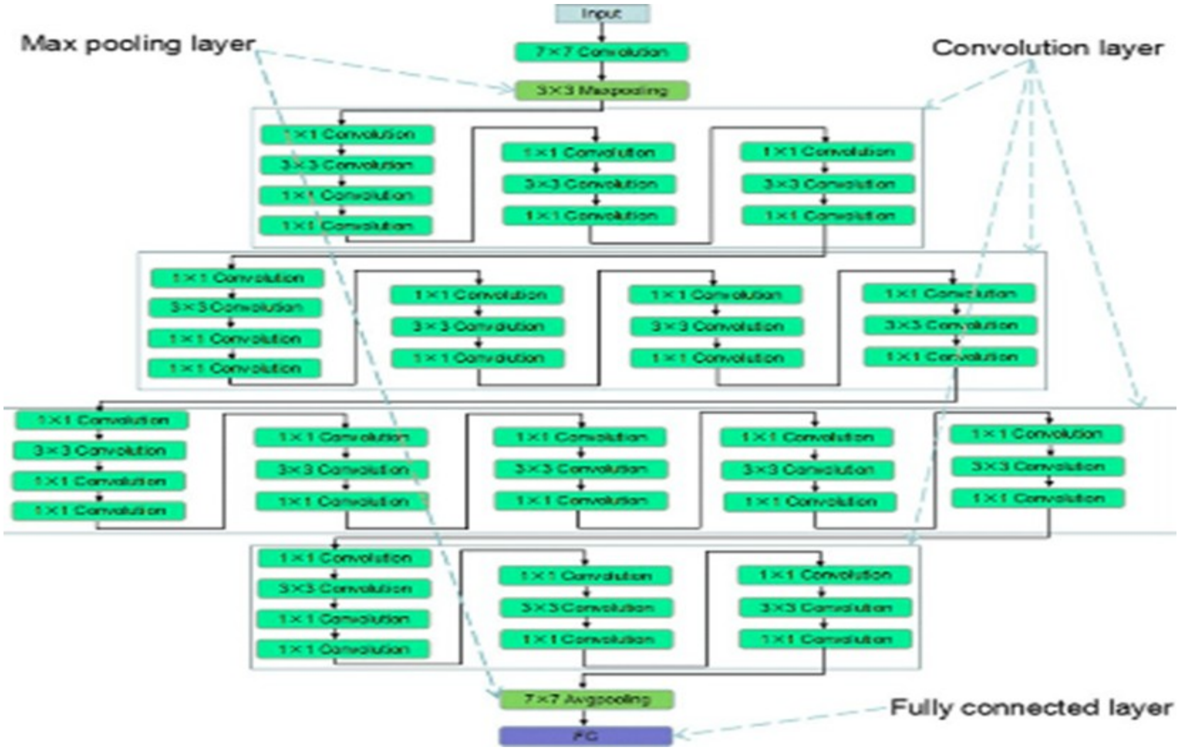


**Fig3: Resnet50 Architecture [19]**

**MobileNetV3Large:**

MobileNetV3 is a state-of-the-art, lightweight deep learning model for mobile devices, optimized for speed and accuracy using advanced techniques like platform-aware neural architecture search (NAS) and NetAdapt. Unlike previous hand-designed versions, MobileNetV3 leverages AutoML to tailor architecture for mobile vision tasks. This design integrates MnasNet's reinforcement learning for initial architecture selection, followed by NetAdapt's refinement of activation channels to enhance efficiency. Available in both Large and Small variants, MobileNetV3's Large model maximizes accuracy with minimal computational costs, making it ideal for real-time mobile applications.

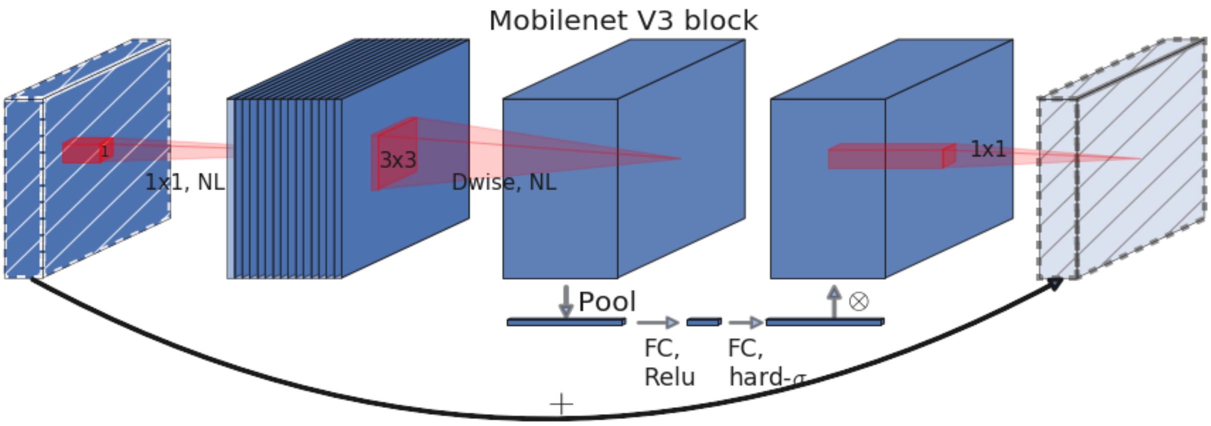


**Fig:4 Architecture of MobileNetV3 [20]**

The MobileNetV3 Large model is optimized for high accuracy in mobile applications, improving upon earlier MobileNet versions through neural architecture search (NAS) and novel design techniques that minimize computational costs while retaining performance.

Key components include depth-wise separable convolutions within bottleneck blocks, squeeze-and-excite modules for highlighting critical features, and alternating ReLU and hard-swish activations optimized for mobile hardware. The model also uses 1x1 convolutions for expansion layers, enabling efficient learning with low computational demand. In segmentation tasks, MobileNetV3 Large employs Lite Reduced Atrous Spatial Pyramid Pooling (LR-SPP) to enhance semantic feature processing with minimal latency. This architecture, tailored by NAS, balances speed and accuracy, making it ideal for mobile vision applications, including image classification and segmentation.

Secondly, we split the dataset into training and test sets, with 90% images for training and 10% for testing.

Finally, our model integrates image and text features to generate descriptive sequences. It processes image features and text sequences separately, then aligns them through an attention mechanism to create a context vector. This vector, combined with the image features, is fed through dense layers to output text tokens, enabling image-based sequence generation. The model uses Adam optimization and categorical cross-entropy with learning rate 0.001 for effective deep learning.

This is the summary of the Custom VGG16 with LSTM model. Here we showed the layer name, Output shape, and number of parameters. We have two input layers because while training we will use the images and captions/text. That is the reason we have two input layers.

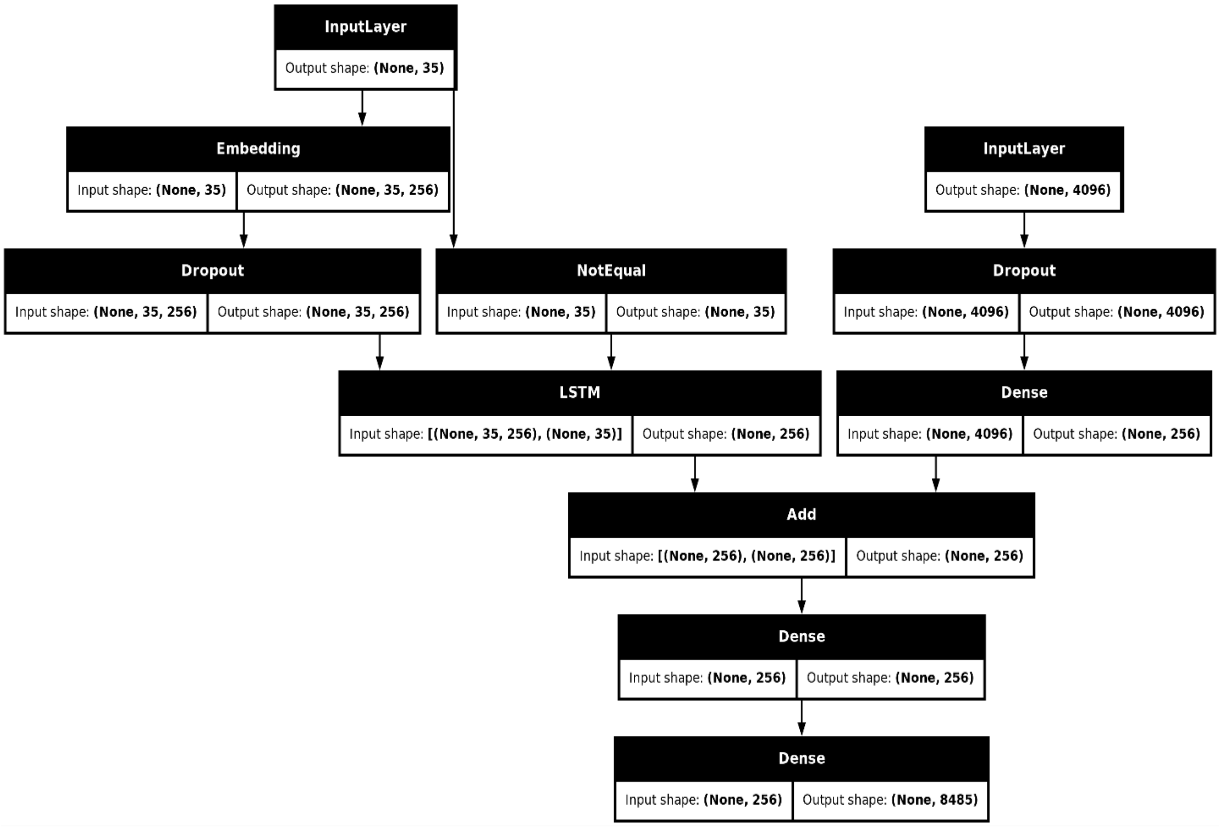

InputLayer
Output shape: (None, 35)
Embedding
Input shape: (None, 35)
Output shape: (None, 35, 256)
InputLayer
Output shape: (None, 4096)
Dropout
Input shape: (None, 35, 256)
Output shape: (None, 35, 256)
NotEqual
Input shape: (None, 35)
Output shape: (None, 35)
Dropout
Input shape: (None, 4096)
Output shape: (None, 4096)
LSTM
Input shape: [(None, 35, 256), (None, 35)]
Output shape: (None, 256)
Dense
Input shape: (None, 4096)
Output shape: (None, 256)
Add
Input shape: [(None, 256), (None, 256)]
Output shape: (None, 256)
Dense
Input shape: (None, 256)
Output shape: (None, 256)
Dense
Input shape: (None, 256)
Output shape: (None, 8485)

**Fig:5 Model Summary Overview (Plot)**

### 3.2.4 Proposed Model Architecture

#### CNN + LSTM Model:

In our thesis, image captioning models like VGG16, MobileNetV3Large, and ResNet50 combine a pre-trained Convolutional Neural Network (CNN) and a Long Short-Term Memory (LSTM) network. The input image, sized 224×224×3 is passed through a CNN pre-trained on the ImageNet dataset to extract a feature vector at the fully connected (fc) layer, representing the visual content of the image. A Linear layer then processes this feature vector to adapt its dimensions for the LSTM decoder.

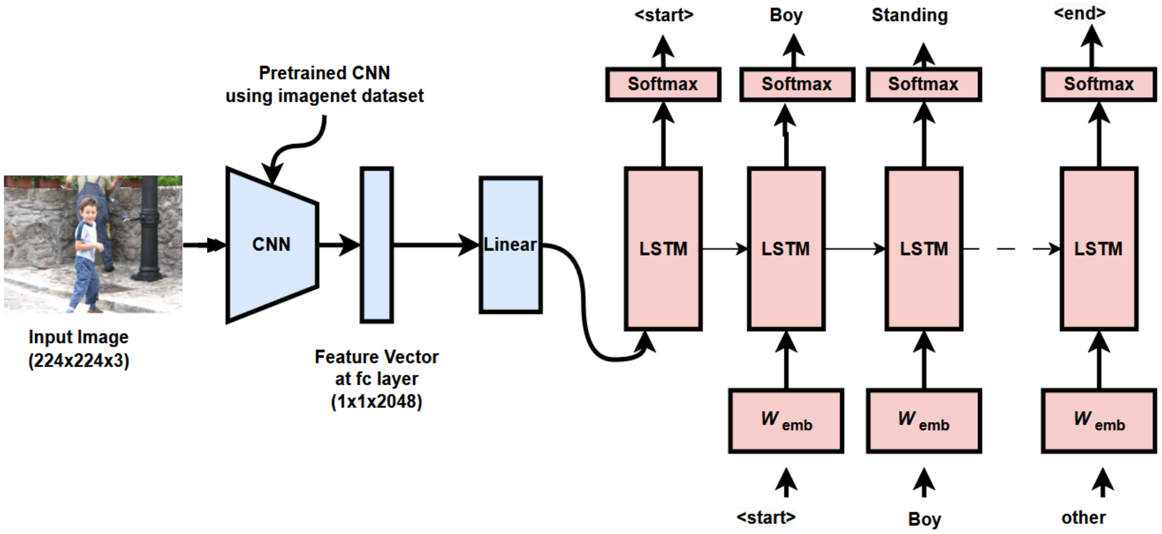


**Fig 6: Process of Caption Generation [21]**

The LSTM generates a caption for the image, starting with a <start> token and predicting words sequentially. At each time step, the LSTM uses the image feature vector, the current input word (embedded into a dense vector using an embedding layer), and its previous hidden state to predict the next word via a Softmax layer. This process continues until the <end> token is generated, completing the caption. For example, the model might generate the caption: <start> → "Boy" → "Standing" → <end>. This architecture effectively combines CNNs for visual understanding and LSTMs for sequence generation.

### gTTS (Google Text-to-Speech):

In our thesis, the gTTS (Google Text-to-Speech) library converts text captions generated by our model into spoken audio. Once the captions are generated (by the LSTM or other captioning model), gTTS takes these text captions as input, synthesizes them into speech, and creates an audio file that can be played back. This can be useful for adding an auditory output to an image captioning system, providing an accessible way to "hear" descriptions of images.

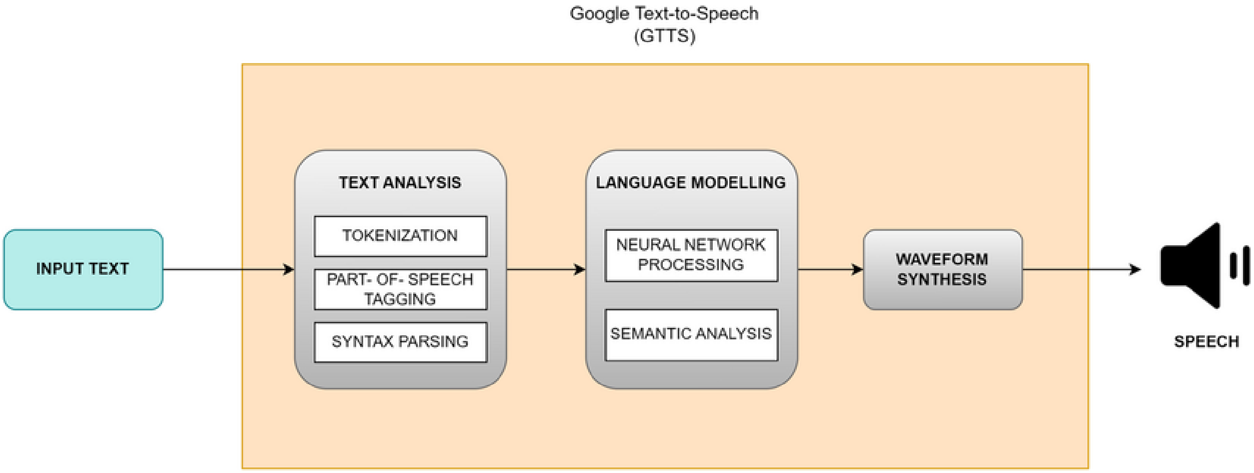
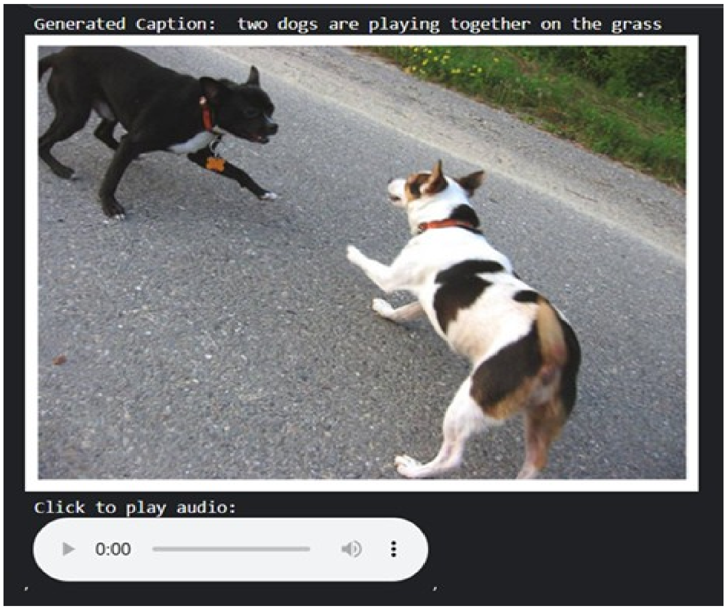


**Fig 7: Generated text caption converted to audio speech [22]**



Generated Caption:  two dogs are playing together on the grass

Click to play audio:

**Fig 8: a) The LSTM model generates captions from images using CNN model ; b) The gTTS (Google Text-to-Speech) library converts text captions generated by CNN model into spoken audio**

### 3.2.5 Experimental Setup

We set the learning rate at 0.001. Then we set 30 epochs with Early Stopping. Suppose there is no progress while training, it will stop the training automatically. Then we started our training model. To evaluate the model, we used metrics like BLEU1 to BLEU4 scores, measuring caption accuracy against ground-truth references, with qualitative analysis for fluency and alignment, ensuring a thorough comparison of each model's ability to produce relevant and coherent captions.

We setup the environment and models for this experiment by fine tuning some hyper parameters. The selected values of the hyperparameters shown in Table 1.

**Table 1:** The selected values for the hyperparameters tuning

| **Hyperparameters** | **Value(s)** | |
|---|---|---|
| Image Size | 224 x 224 x 3 | |
| Class Mode (CNN) | Categorical | |
| Transfer Learning Weights | ImageNet | |
| Training Split | 90% | |
| Testing Split | 10% | |
| Pooling | Max-Pooling | |
| Activation | Hidden Layers | ReLu |
| | Output Layer | Softmax |
| Optimizer | Adam | |
| Learning Rate (CNN) | 0.001 with learning rate warmup for first 10 epochs | |
| Loss | CNN | Categorical Cross Entropy |
| Metrics | CNN | BLEU1 to BLEU4 |
| Epoch | CNN | 30 |
| Batch size | CNN | 32 |

## 3.3 Summary

To develop "Image Captioning", we need to focus on three main keys. The deep learning model, Backend API, and Frontend UI. To build a deep learning model we took the Kaggle dataset. Preprocessed the images and feature extraction using the MobileNetV3Large, VGG16 and ResNet50 models. We used the TensorFlow framework and Neural Networks to train the model. Python serves as our backend server, while HTML and CSS handles the frontend, providing a user-friendly interface for our web application.

# Chapter 4

# Data Handling

To achieve our goal, we used the open-source Flickr8k and Flickr30k datasets from Kaggle. In this chapter, we will give an overview of our datasets.

## 4.1 Dataset Description

The Flickr8K and Flickr30K datasets consist of an **images folder** (containing .jpg, .png, and .jpeg images) and a **captions.txt file**. The Flickr8K dataset includes 8,091 images, each paired with five descriptive captions, totaling 40,456 captions, while the Flickr30K dataset contains 31,783 images with five captions per image, amounting to 158,915 captions. These datasets cover a wide range of everyday scenarios, including people, animals, objects, and various activities, providing rich visual and contextual diversity for training image captioning models. This variety makes them suitable for enhancing models' ability to associate visual content with accurate and meaningful textual descriptions.A sample overview of the dataset is given below:

| Dataset | Image | Caption.txt |
|---|---|---|
| Flickr8K [15] | | A child in a pink dress is climbing up a set of stairs in an entry way .<br>A girl going into a wooden building .<br>A little girl climbing into a wooden playhouse .<br>A little girl climbing the stairs to her playhouse .<br>A little girl in a pink dress going into a wooden cabin. |
| Flickr30K [16] | | Two young guys with shaggy hair look at their hands while hanging out in the yard .<br>Two young , White males are outside near many bushes .<br>Two men in green shirts are standing in a yard .<br>A man in a blue shirt standing in a garden .<br>Two friends enjoy time spent together . |

**Table 2: Dataset Overview**

## 4.2 Data Analysis Plan

In this thesis, we utilize **MobileNetV3Large**, **VGG16**, and **ResNet50** architectures to extract essential image features and generate descriptive captions. **MobileNetV3Large**, designed for efficiency and performance, employs depth-wise separable convolutions to preprocess datasets and extract critical visual features, making it ideal for lightweight and high-accuracy applications. Each image is paired with five captions from the captions.txt file, and the vocabulary size is computed from all captions to enhance the language generation process. The datasets is split into 90% for training and 10% for testing to ensure robust model evaluation.

The **VGG16** model, known for its deeper architecture with 16 layers, is used to generate feature vectors by capturing spatial patterns and texture information from images. These features serve as input for an LSTM decoder to generate coherent and contextually accurate captions. On the other hand, **ResNet50**, with its residual connections, excels in handling deeper networks by avoiding the vanishing gradient problem. It extracts high-level features effectively, enabling the generation of detailed captions through its integration with an LSTM decoder.

This comparative analysis between MobileNetV3Large, VGG16, ResNet50, evaluates their performance in terms of caption accuracy and quality, offering valuable insights into the strengths and limitations of each architecture for image captioning tasks.

## 4.3 Summary

In this project, we aim to build an advanced image captioning model using the deep learning models combined with an LSTM network. First, we preprocess the images and captions of the datasets, which involves resizing and normalizing each image to extract high-quality features and make captions machine readable. Using the models, each image is passed through a pre-trained convolutional neural network to obtain rich, compact feature representations that capture essential visual details. For caption generation, these extracted features are then fed into an LSTM-based language model. The LSTM network is responsible for learning the temporal dependencies in the sequence of words and generating coherent captions by predicting each word in the sentence one by one. The model is trained on pairs of images and corresponding captions, allowing it to learn an association between visual content and textual descriptions. By the end of the training, our MobileNetV3Large +LSTM , VGG16 + LSTM and ResNet50 + LSTM models will be capable of generating meaningful and contextually accurate captions for unseen images, providing a robust approach to automatic image description.

# Chapter 5

# Result and Discussion

We present a comprehensive analysis of our findings and the results observed during training. From developing a deep learning model to successfully launching the application, we achieved our objectives through a systematic and iterative approach. This section explains the model evaluation process and provides a detailed rationale behind the methods and metrics used.

## 5.1 Obtained Results

In the Flickr8K dataset, three models—VGG16, MobileNetV3Large, and ResNet50—were trained for 30 epochs, achieving respective training accuracies and losses of 48.62% (1.9498), 83.78% (0.4943), and 48.29% (1.9735). During testing, BLEU scores were evaluated, with VGG16 achieving BLEU-1 to BLEU-4 scores of 0.5483, 0.3214, 0.2019, and 0.1202, respectively; MobileNetV3Large scoring 0.4999, 0.2734, 0.1616, and 0.0915; and ResNet50 scoring 0.5422, 0.3177, 0.2013, and 0.1218. While MobileNetV3Large excelled in training with the highest accuracy and lowest loss, VGG16 and ResNet50 slightly outperformed it in higher-order BLEU scores, indicating better performance for generating more complex captions during testing.

In our other dataset Flickr30K, three models—MobileNetV3Large, VGG16, and ResNet50—were trained for 30 epochs. During training, MobileNetV3Large achieved the highest accuracy (67.18%) with a loss of 1.1720, compared to VGG16 (39.25%, loss: 2.6345) and ResNet50 (39.97%, loss: 2.5859). In the testing phase, MobileNetV3Large achieved BLEU scores of 0.4781 (BLEU-1), 0.2376 (BLEU-2), 0.1310 (BLEU-3), and 0.0690 (BLEU-4), while VGG16 performed better with BLEU scores of 0.5364, 0.2868, 0.1663, and 0.0899, respectively. ResNet50 closely followed VGG16 with BLEU scores of 0.5347, 0.2926, 0.1730, and 0.0963. These results indicate that while MobileNetV3Large excelled during training with higher accuracy and lower loss, VGG16 and ResNet50 demonstrated stronger performance in generating complex captions in the testing phase.

The Image Caption Generator application uses HTML and CSS for the frontend, integrated with an API as the backend to utilize the best-performing model for generating captions. Users can upload an image by choosing a file, and clicking "Generate Caption" sends the image data to the backend model, which returns an appropriate caption. The simple and clean design ensures a user-friendly experience, while leveraging the model that showed the best performance, likely chosen based on BLEU scores or accuracy during testing, for optimal caption generation results.

The purpose of creating this application is to demonstrate the practical implementation of deep learning, combining computer vision and natural language processing to solve real-world problems. It serves as an accessible tool for generating meaningful image descriptions, which can be used in applications like aiding visually impaired users, improving image search engines, or automating content creation. Additionally, it showcases the power of AI in understanding and describing visual content, making it an engaging educational and functional tool.

## 5.2 In-depth Result Analysis

After splitting the datasets for the custom models, we obtained **28,604 images for training** and **3,179 images for testing** from the Flickr30K dataset. Similarly, for the Flickr8K dataset, we utilized **7,281 images for training** and **810 images for testing**. To visualize the model's performance during training, we ensured that the training history was saved, allowing us to generate loss and accuracy plots once the training process was complete. The plots for loss and accuracy of the trained model are presented below:

### Flickr8K Dataset:

**Table 3: Loss and Accuracy scores for each model**

| Algorithm | Training Accuracy | Trainig Loss |
|---|---|---|
| MobileNetV3Large | 0.8378 | 0.4943 |
| VGG16 | 0.4862 | 1.9498 |
| ResNet50 | 0.4829 | 1.9735 |

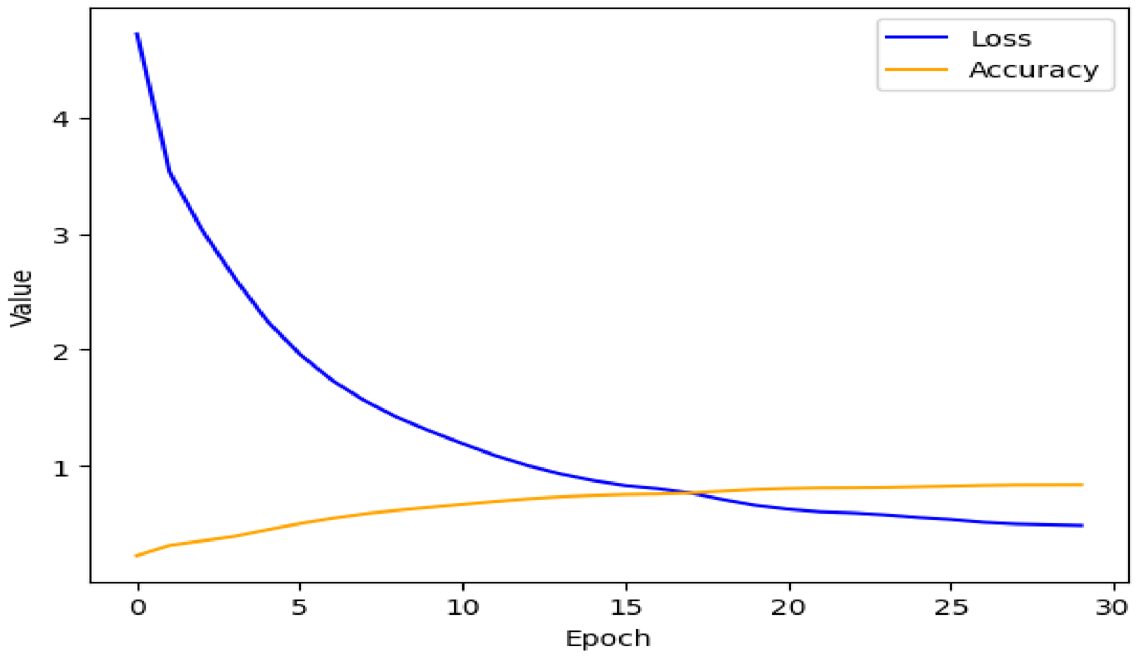


**Fig 9: Loss Accuracy plot of MobileNetV3Large**

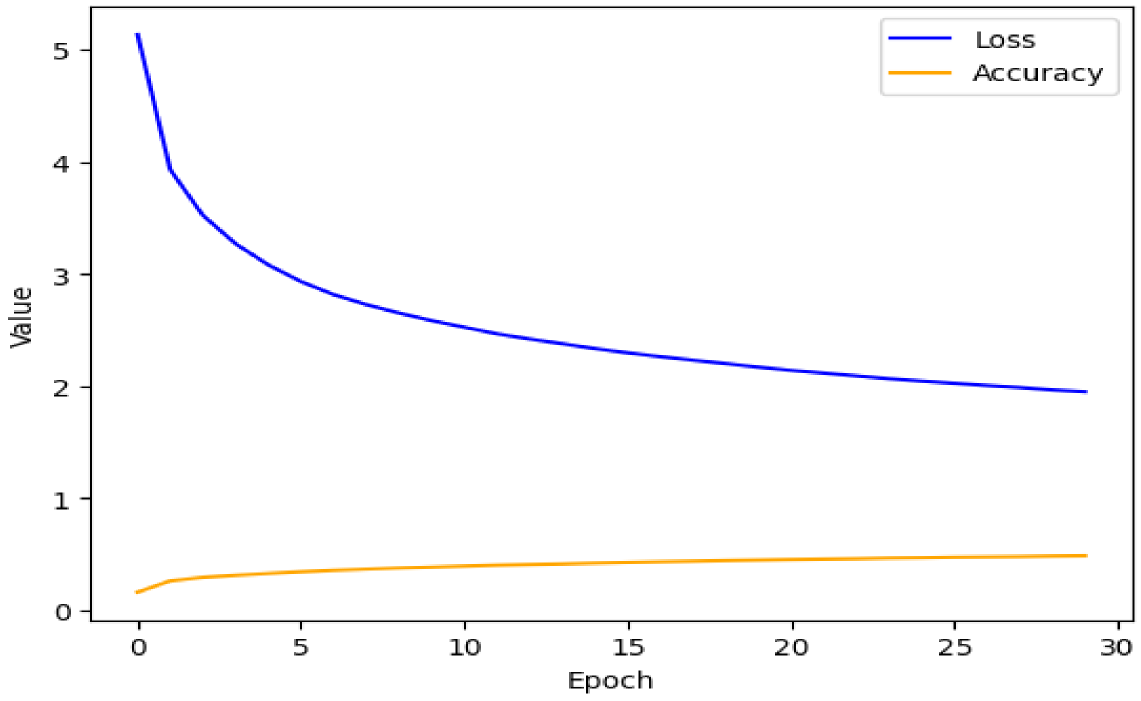


**Fig 10: Loss Accuracy plot of VGG16**

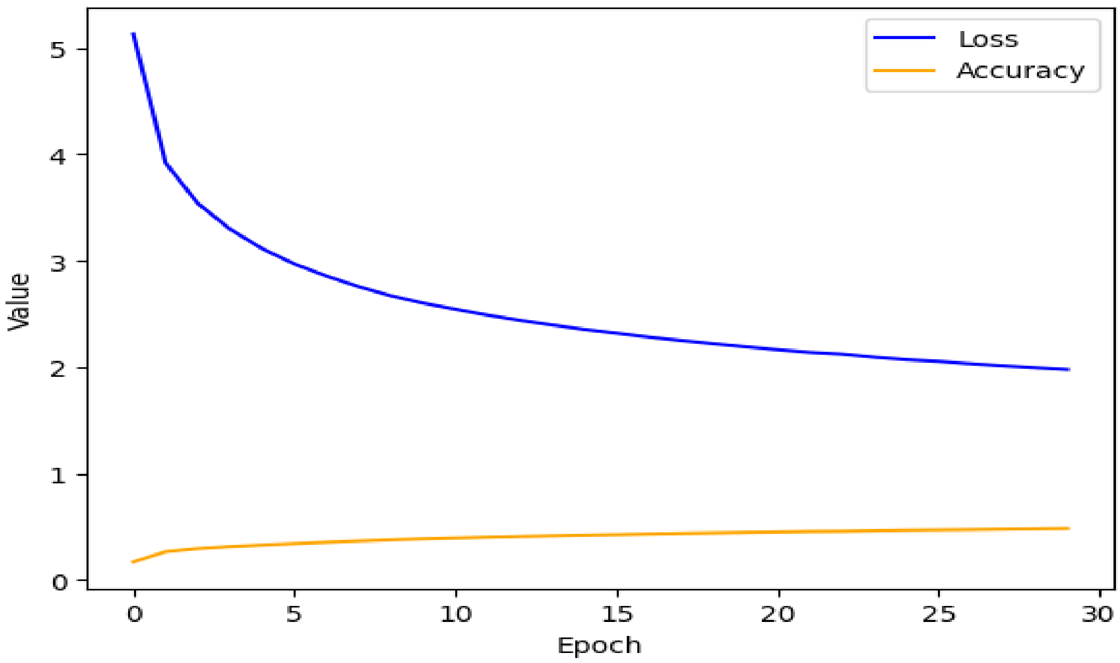


**Fig 11: Loss Accuracy plot of ResNet50**

## Flickr30k Dataset:

**Table 4: Loss and Accuracy scores for each model**

| Algorithm | Training Accuracy | Training Loss |
|---|---|---|
| MobileNetV3Large | 0.6718 | 1.172 |
| VGG16 | 0.3925 | 0.6345 |
| ResNet50 | 0.3997 | 2.5859 |

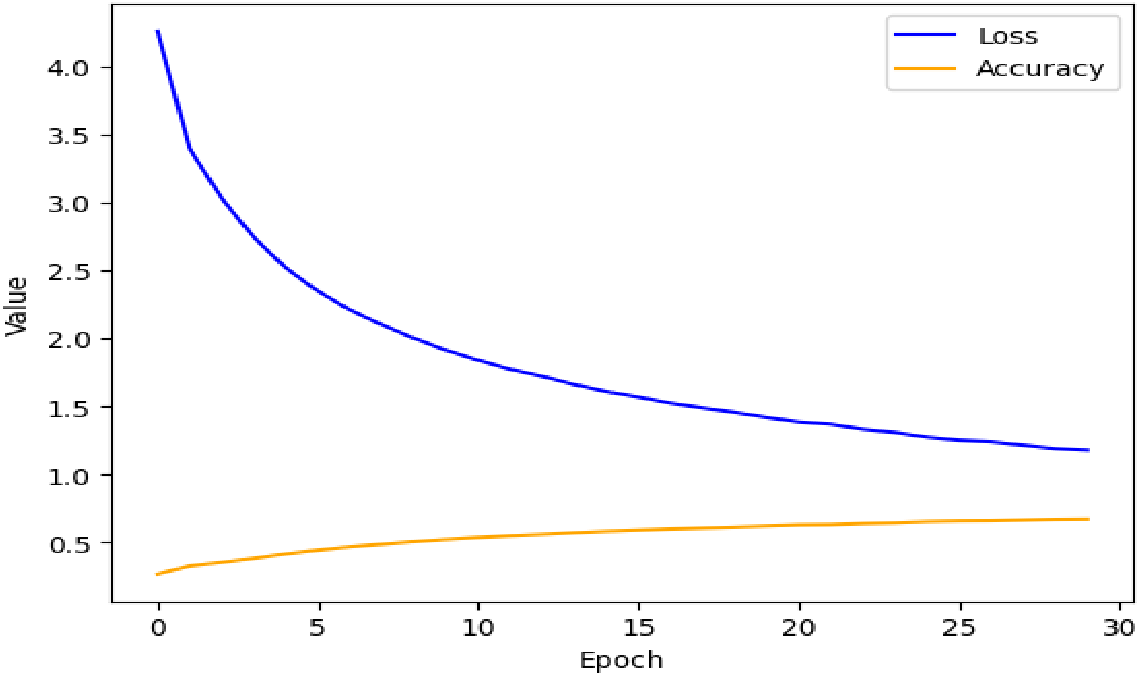


**Fig 12: Loss Accuracy plot of MobileNetV3Large**

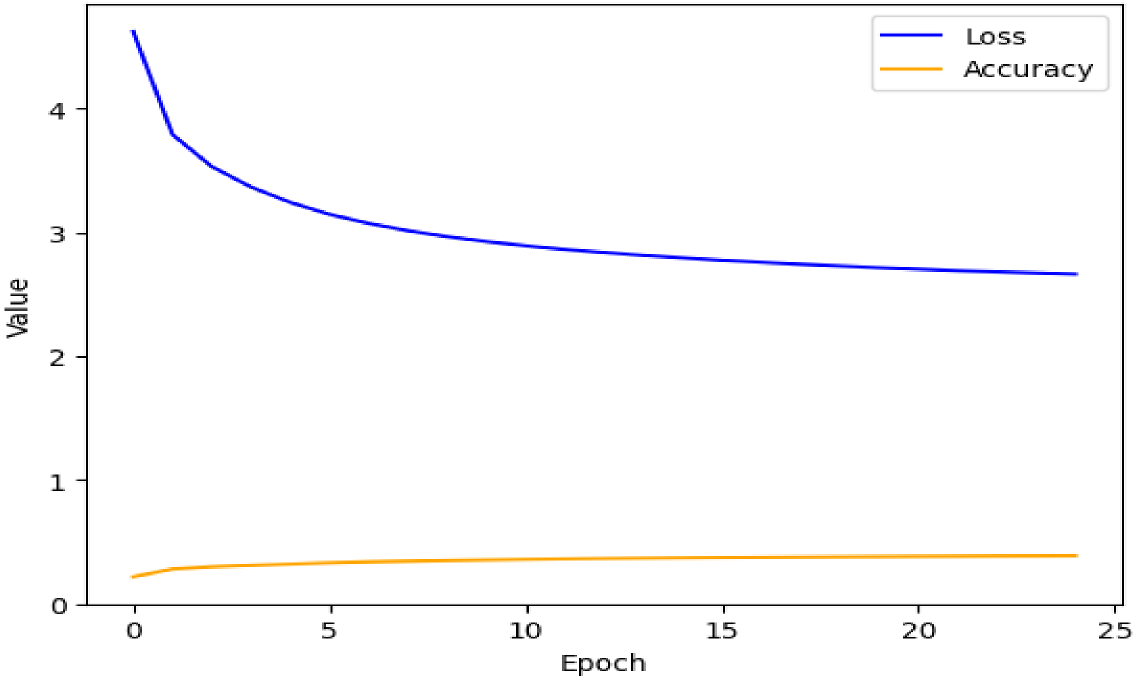


**Fig 13: Loss Accuracy plot of VGG16**

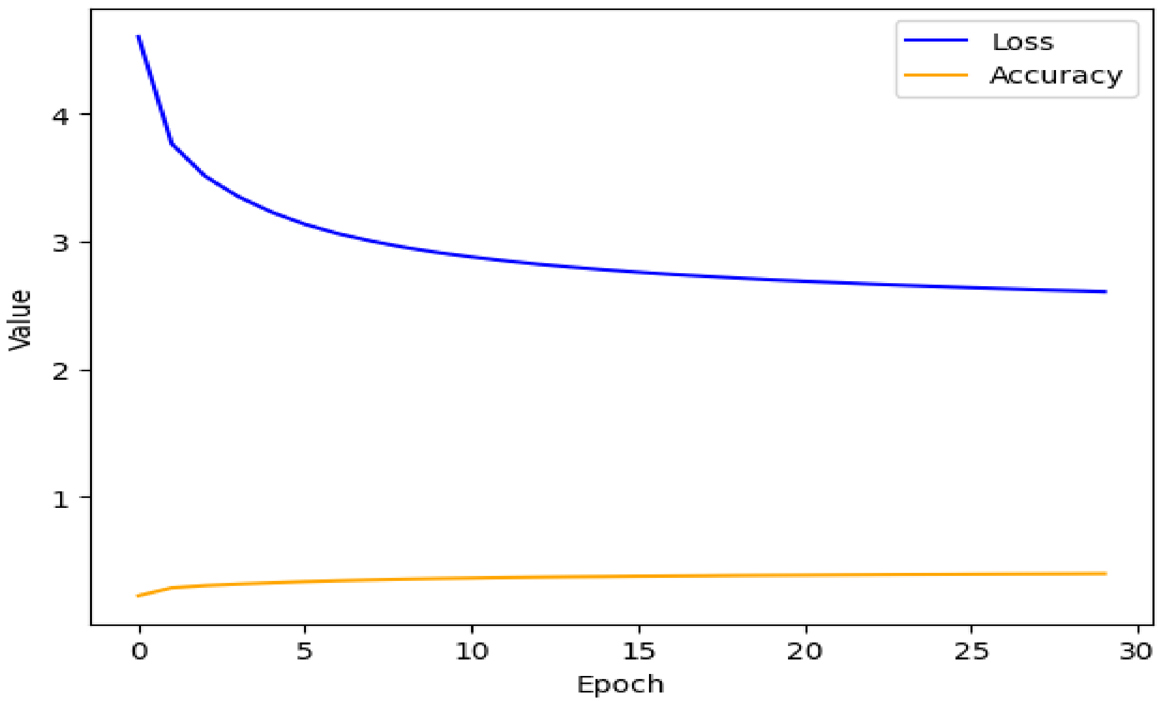


**Fig 14: Loss Accuracy plot of ResNet50**

## 5.3 Performance Evaluation

The BLEU (Bilingual Evaluation Understudy) score is a widely used metric for evaluating the quality of machine-generated text, such as captions or translations, by comparing it to one or more reference texts. BLEU scores range from 0 to 1, with higher scores indicating better performance. BLEU scores are often reported at different levels (BLEU-1 to BLEU-4) to capture varying aspects of quality:

**1. BLEU-1**

- What it measures: The unigram (1-gram) precision, which means it checks how many individual words in the generated text match the reference text, irrespective of their order.
- Focus: Captures the overall vocabulary overlap.
- Limitation: Does not account for the order or context of words.

  BLEU-1: Precision with unigrams (n = 1)

$$\mathrm{P1}= \frac{Number\ of\ matching\ words}{Total\ words\ in\ candidate}$$

$$\mathrm{BLEU1 = BP{\cdot}P1}$$

**2. BLEU-2**

- What it measures: The bigram (2-gram) precision, focusing on matching pairs of consecutive words between the generated and reference texts.
- Focus: Incorporates some degree of word order and short-range context.
- Use case: Helps detect if phrases are formed correctly.

  BLEU-2: Precision with bigrams (n = 2)

$$\mathrm{P2} = \frac{Number\ of\ matching\ bigrams}{Total\ bigrams\ in\ candidateNumber}$$

$$\mathrm{BLEU2{=}BP{\cdot}}\surd P1.P2$$

**3. BLEU-3**

- What it measures: The trigram (3-gram) precision, looking at three consecutive words.
- Focus: Evaluates a deeper level of phrase coherence and order.
- Limitation: Performance might decrease if the generated text diverges more from reference phrases.

BLEU-3: Precision with trigrams (n = 3)

$$\mathrm{P3} = \frac{Number\ of\ matching\ trigrams}{Total\ trigrams\ in\ candidate}$$

$$\mathrm{BLEU2{=}BP}\cdot\sqrt[3]{P1.P2.P3}$$

### 4. BLEU-4

- What it measures: The 4-gram precision, focusing on four-word sequences.
- Focus: Assesses the fluency and grammatical correctness of the generated text.
- Use case: Often used as a summary score for tasks where sentence-level structure is crucial.

BLEU-4: Precision with 4 - grams (n = 4)

$$\mathrm{P4} = \frac{Number\ of\ matching\ 4{-}grams}{Total\ 4{-}\ grams\ in\ candidate}$$

$$\mathrm{BLEU2{=}BP}\cdot\sqrt[4]{P1.P2.P3.P4}$$

The performance evaluation of the models was conducted using **BLEU-1, BLEU-2, BLEU-3, and BLEU-4** scores to measure the accuracy and quality of the generated captions compared to the ground truth captions. The results are summarized below:

**Table 5:** Outcomes of the testing part for three models on two datasets (Flickr8K and Flickr30K)

| Dataset | Algorithm | BLEU-1 | BLEU-2 | BLEU-3 | BLEU-4 |
|---|---|---|---|---|---|
| Flickr8k | VGG16 + LSTM | 0.5483 | 0.3214 | 0.2019 | 0.1202 |
| | MoblieNetV3Large + LSTM | 0.4999 | 0.2734 | 0.1616 | 0.0915 |
| | ResNet50 + LSTM | 0.5422 | 0.3177 | 0.2013 | 0.1218 |
| Flickr30k | VGG16 + LSTM | 0.5364 | 0.2868 | 0.1663 | 0.0899 |
| | MoblieNetV3Large + LSTM | 0.4781 | 0.2376 | 0.131 | 0.069 |
| | ResNet50 + LSTM | 0.5347 | 0.2926 | 0.173 | 0.0963 |

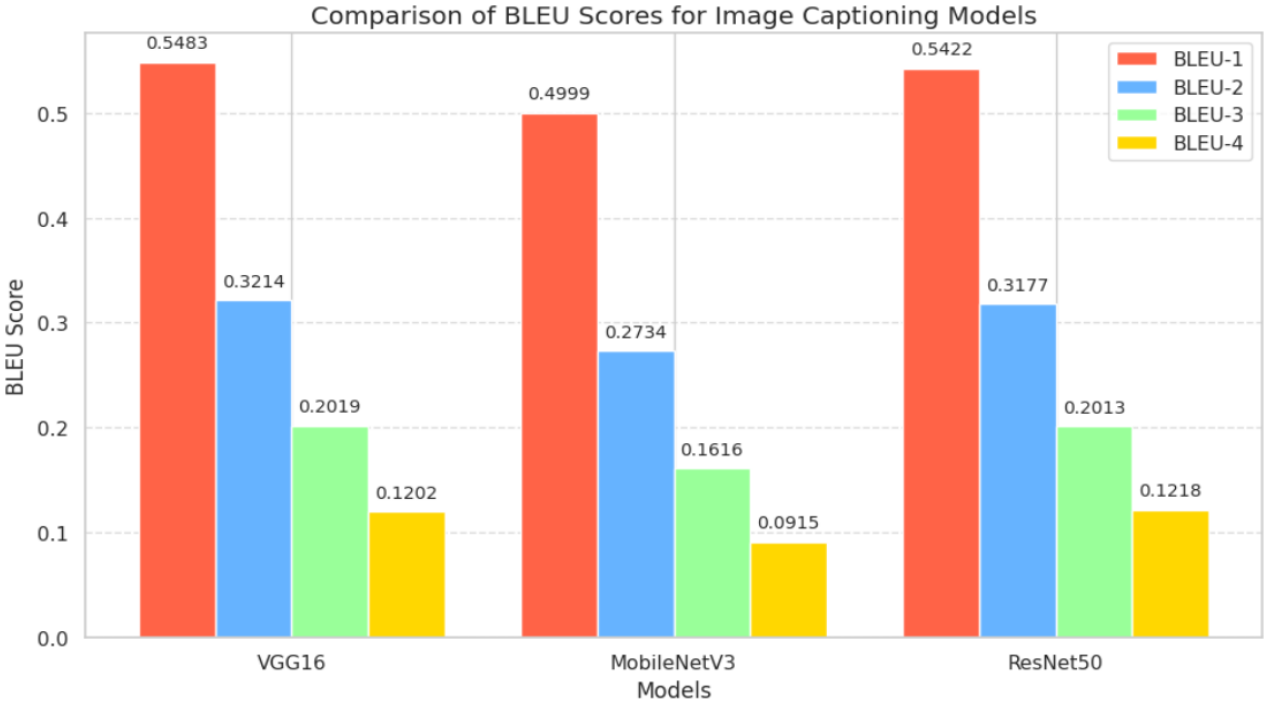


**Fig 15: Comparison of BLEU Scores plot (Flickr8K Dataset)**

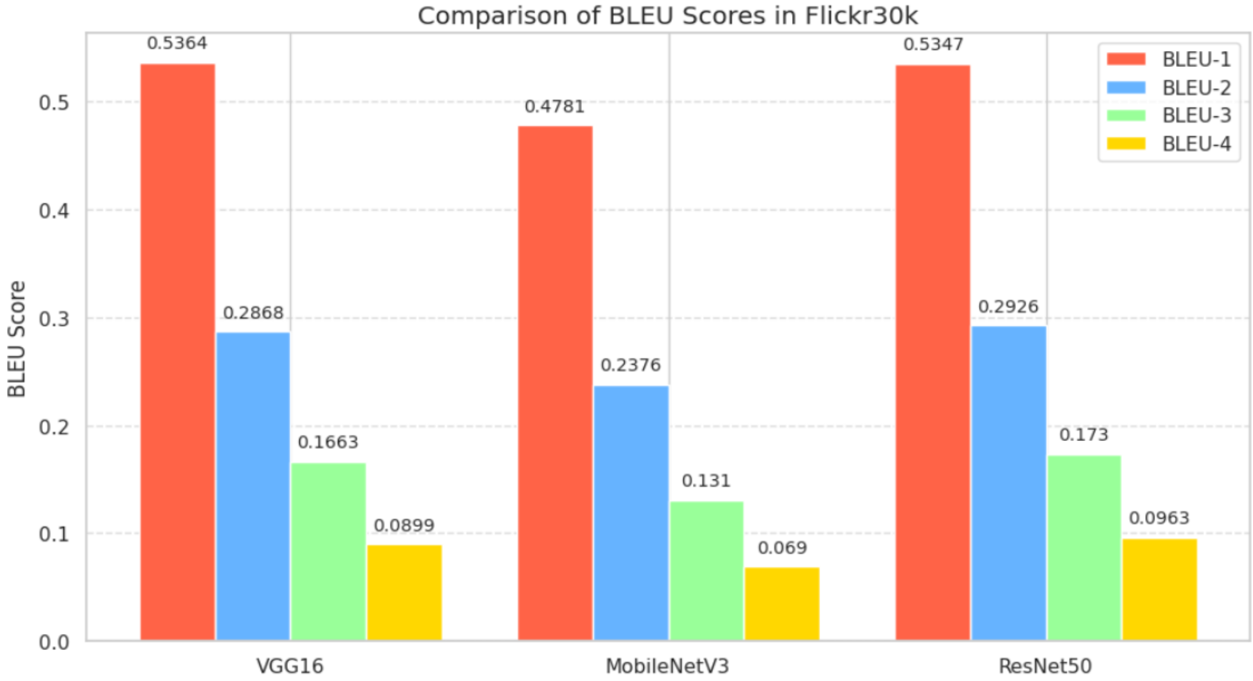


**Fig 16: Comparison of BLEU Scores plot (Flickr30K Dataset**

## 5.4 App Performance

For our application, we have designed a simple and intuitive user interface to ensure ease of use. The user can either capture an image directly or select one from their gallery. After selecting or capturing the desired image, the user can upload it using the "Choose File" option. Once the image is uploaded, it is displayed on the screen for confirmation. The user can then press the "Generate Caption" button, which processes the image through the backend API to generate an appropriate caption. Additionally, the app also provides an audio speech output of the generated caption, making it more interactive and accessible. This streamlined design ensures a smooth and user-friendly experience for everyone.

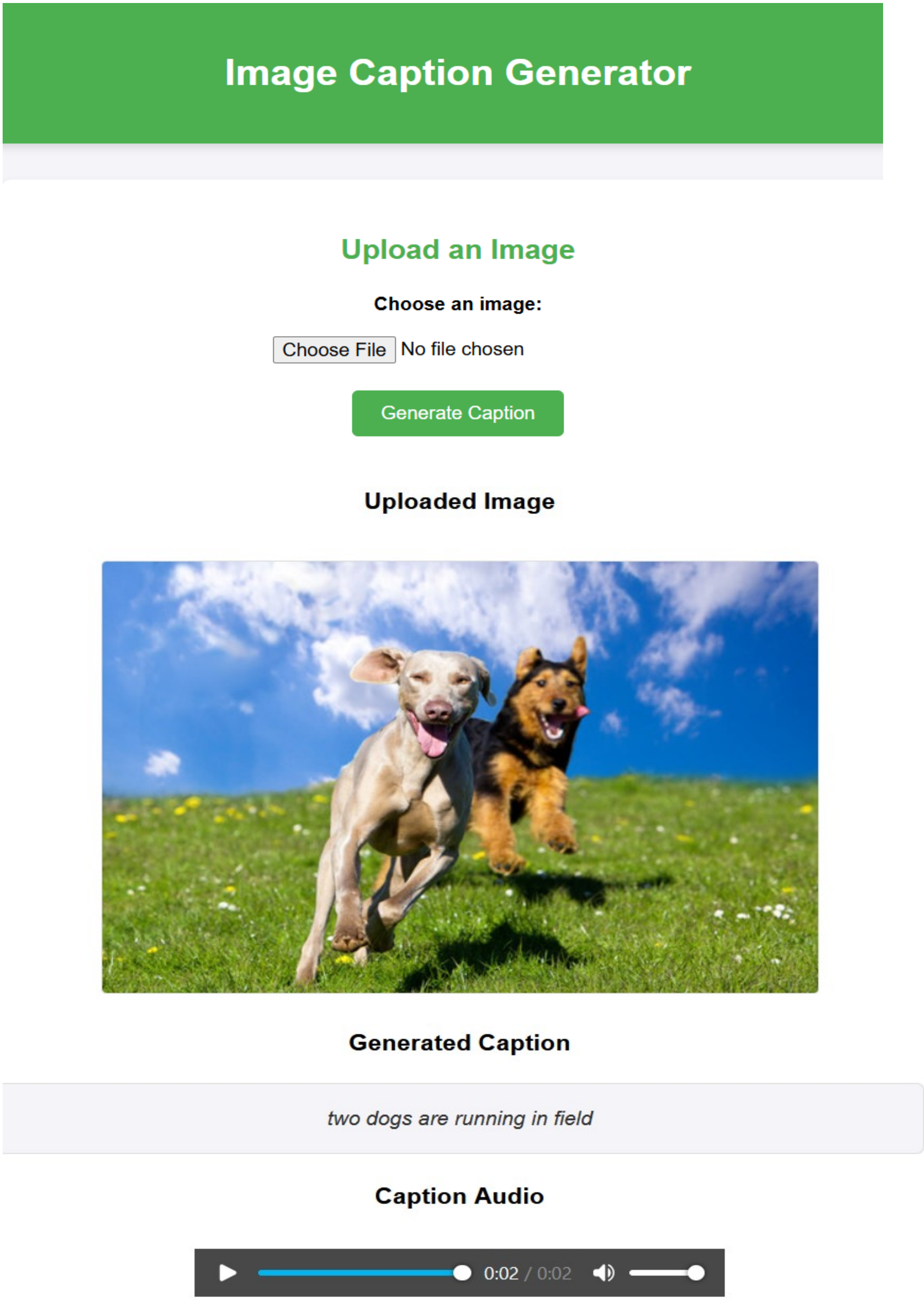


Image Caption Generator

Upload an Image

Choose an image:

Uploaded Image

Generated Caption

*two dogs are running in field*

Caption Audio

**Fig 17: Web Application (Generate caption and audio speech)**

## 5.5 Discussion

This project demonstrates the application of multimodal deep learning in building a robust Image Caption Generator system. The backend is powered by a pre-trained VGG16 model, which extracts visual features from input images, and an LSTM network that processes these features to generate meaningful textual descriptions. The model is trained using a carefully curated dataset and fine-tuned to ensure accurate and coherent captions. Evaluation metrics such as BLEU scores validate the model's performance.

The generated web app provides an intuitive and interactive user experience. Users can upload images through the frontend, designed using HTML and CSS for simplicity and responsiveness. Once an image is uploaded, the backend API processes it, predicts the caption, and converts the text into speech using gTTS, creating an audio output. This functionality makes the application versatile and accessible, catering to users with varying needs, including those with visual impairments.

The project exemplifies the integration of machine learning with modern web technologies, showcasing the practical use of AI in everyday scenarios. It also emphasizes scalability and flexibility, allowing for future enhancements like supporting additional languages, real-time predictions, or integrating more sophisticated models for improved accuracy. This application not only demonstrates technical expertise but also addresses real-world usability, bridging the gap between advanced AI systems and user-friendly interfaces.

# Chapter 6

# Conclusion

In conclusion, this Image Caption Generator project successfully integrates deep learning techniques with a user-friendly web interface to generate meaningful image descriptions. By leveraging pre-trained models like VGG16, MobieNetV3, ResNet50, and LSTM for caption generation, along with tools like gTTS for audio output, the system demonstrates an efficient pipeline for multimodal learning. The web app provides a seamless user experience, making it accessible for real-world applications. This project not only highlights the potential of deep learning in image understanding but also opens avenues for further advancements, such as multilingual captioning, real-time processing, and deployment at scale for broader accessibility.

## 6.1 Overall Contribution

The overall contribution of this Image Caption Generator project lies in its ability to bridge the gap between computer vision and natural language processing, delivering a comprehensive solution for image-to-text conversion. It showcases an efficient integration of pre-trained CNN models for feature extraction and LSTM networks for sequential text generation, ensuring high-quality captions. Additionally, the project incorporates audio synthesis, making it accessible to visually impaired users. The deployment as a web application enhances usability, providing an interactive platform for caption generation. This project not only serves as a practical implementation of multimodal learning but also contributes to the growing field of accessible AI by demonstrating its real-world applicability in generating both visual and auditory outputs.

## 6.2 Limitations

The project, while impactful, has limitations such as dependency on dataset quality and diversity, leading to biased or generic captions and reduced contextual understanding in complex scenes. The use of CNN models and LSTM limits scalability for Flickr8K and Flickr30K datasets and high-resolution images, and real-time performance may require significant computational resources. Additionally, audio generation relies on TTS engines, which may lack natural

intonation, and the model may exhibit language or cultural biases due to insufficiently diverse training data. Furthermore, the system's focus on images and text restricts its ability to handle additional modalities like video or sound, which could enrich captions. Addressing these issues could enhance the system's versatility and performance.

## 6.3 Future Works

Future works could incorporate advanced transformer-based models like Vision Transformers (ViT) or GPT to improve caption quality and contextual understanding. Expanding the dataset with diverse and high-quality images could address biases and enhance generalizability. Real-time performance can be optimized by quantization techniques for deployment on edge devices. Adding multimodal capabilities, such as video captioning or sound-based contextual enhancements, could broaden the application scope. Further, integrating multilingual support and fine-tuning the TTS system for natural intonation would enhance accessibility and audio quality. User feedback and adaptive learning mechanisms could also be implemented for continuous improvement

# Bibliography

[1] S. *Syafaruddin, H. Elfe, and A. U. R. A. W. Rahim, "Image Caption Generation Through the Integration of CNN-Based Residual Network Architectures and LSTM," *ResearchGate*, Oct. 2023. [Online].

[2] S. Zahra, M. Iqbal, H. U. Rehman, S. Shoaib, and A. A. Saleem, "Image Captioning Using Novel Multimodal Feature Fusion," *SSRN Electronic Journal*, Oct. 2024. [Online].

[3] Timilsina, "Image Captioning in Nepali Using CNN and Transformer Decoder," *ResearchGate*, Oct. 2024. [Online]..

[4] A. Bhadange, R. Bhole, and V. Jabade, "Image Captioning Using Hybrid Neural Architecture in Multimodal Contexts," in *Proceedings of the 2023 2nd International Conference on Futuristic Technologies (INCOFT)*, Belagavi, Karnataka, India, Nov. 24–26, 2023, pp. 1-6. doi: 10.1109/INCOFT60753.2023.10425636.

[5] R.-A. Armstrong, T. Jiang, and C. K. Kim, "Multi-Modal Image Captioning," *CS231n: Convolutional Neural Networks for Visual Recognition*, Stanford University, 2022. [Online].

[6] S. Ali, T. Jan, F. Ullah, and F. Naz, "Image Caption Generation Using CNN and Transformer," *Procedia Computer Science*, vol. 206, pp. 442-447, 2022. [Online]

[7] K. Kushal, M. Manoj, K. Reddy, and P. C. Nair, "Image Captioning Using Hybrid Deep Learning Models," in *2024 IEEE 9th International Conference for Convergence in Technology (I2CT)*, Pune, India, 2024, pp. 1-6. doi: 10.1109/I2CT61223.2024.10544337.

[8] M. Zhang, J. Li, and X. Wang, "A Novel Approach for Image Captioning Using Transformer Networks," *arXiv*, Mar. 2022.

[9] J. Donahue, L. A. Hendricks, and M. Rohrbach, "Long-Term Recurrent Convolutional Networks for Visual Recognition and Description," arXiv, Jul. 2017.

[10] A. Devaraj, S. V. G., and V. Vishwakarma, "Image Captioning Using Deep Convolutional Neural Networks (CNNs)," *ResearchGate*, Dec. 2020.

[11] L. Aneja, A. Deshpande, and A. G. Schwing, "Convolutional Image Captioning," *ResearchGate*, Nov. 2017.

[12] Visual Question Answering (VisualQA), "VisualQA: Visual Question Answering.

[13] K. P. Murali, T. Shashikala, and K. R. R. V. R. P. Yadava, "An Efficient Deep Learning Approach for Image Captioning Using Attention Mechanism," *IEEE Xplore*, Mar. 2021.

[14] S. Yadav, R. Kumar, and S. Verma, "Image Caption Generating Deep Learning Model," *International Journal of Engineering Research & Technology (IJERT)*, vol. 10, no. 09, pp. 100-105, Sep. 2021.

[15] [3] A. Jain, "Flickr8k Dataset," Kaggle, 2020.[online]

[16] A. Jain, "Flickr30k Dataset," Kaggle, 2020. [online]

[17] Kaggle, "Kaggle: Your Machine Learning and Data Science Community," [online]

[18] GeeksforGeeks,"VGG-16 CNN model," [online]

[19] A. Qasem, "The architecture of ResNet50 and deep learning model flowchart," ResearchGate, 2019. [online]

[20] A. Pandrii, "Mobile-Net Architectures," Medium, Apr. 22, 2021. [online]

[21] A. Singh, "Solving an Image Captioning Task Using Deep Learning," Analytics Vidhya, Apr. 13, 2018. [online]

[22] ResearchGate, "Google Text-to-Speech architecture," [Online].

# Budget Cost of Work Schedule (BCWS)

| Task name | Sub Task | Unit Cost (BDT) | Total |
|---|---|---|---|
| **Development Cost** | Front-End Development | 10,000 | 30,000 |
| | Back-End Development | 20,000 | |
| **Hosting and Domain** | Hosting | 12,000 | 12,000 |
| **Machine Learning Model Hosting** | Automated testing tools | 6,000 | 28,500 |
| | Automated testing tools | 8,500 | |
| | Maintenance and Updates | 10,000 | |
| | Ongoing support for bug fixes and feature updates | 5,500 | |
| | Data Engineer | 20,000 | |
| **Total (BDT)** | | | 70,500 |

# Appendix

## Mapping of Course and Program Outcomes

## CSE400-A

**Program Outcomes:**

**PO1 (Engineering Knowledge):** The problem we have chosen can be resolved with the help of our engineering and computer science expertise. We chose this project by taking the state of the world into account and using both our old and new knowledge to tackle it. We have applied programming language, data science ideas, deep learning, artificial intelligence, and machine learning to solve this issue.

**PO4 (Investigation):** Gaining knowledge about the prior work completed on a project is essential. Learning about the constraints that our project can overcome will be aided by reading a lot of articles. We have studied numerous study articles about our job to have a better knowledge of the instruments, strategies, and tactics they have employed to accomplish their own objectives.

| CO | Details | Knowledge Profile(K) | Engineering Problem (EP) |
|---|---|---|---|
| CO1 | To locate a real-life hard challenge for our capstone project, we have acquired expertise from data science, artificial intelligence, deep learning, and machine learning. Solving this problem will solve many other ones. | **(i) Identify a real-life problem [K1, K2, K3, K4]**<br>**KI:** Natural sciences based on theory: We work on data science, machine learning, deep learning and statistical tools. Thus, from earlier courses, we have acquired both theoretical and practical knowledge. We require in-depth understanding.<br>.<br>**K2:** Formal parts of computer and information science, statistics, numerical analysis, and mathematics with a conceptual foundation: Our capstone project | **(i) Identify a real-life problem [EP1, EP2, EP3, EP4, EP5, EP6, EP7]**<br>**EP1:** Required level of expertise: We attempted to utilize our engineering, design, and practice knowledge in this project.<br>**EP2:** Extensive engineering, contradictory technical, and other problems, like several model types and data augmentation techniques, are all part of our system. |

| | | | |
|---|---|---|---|
| | | was chosen with machine learning and deep learning in mind. Thus, data science is frequently employed ideas. All of these rely on analysis of numbers.<br>**K3:** Fundamentals of theory-based engineering: To construct our project, we must be familiar with programming languages. We require a solid grasp of programming language and its foundations because we are utilizing machine learning and deep learning technologies. We also require development knowledge to complete our project flawlessly.<br>**K4:** Practice-oriented expertise in front-end engineering: We possess expertise in machine learning, deep learning, Python, and other areas. | **EP3:** Required level of analysis: The problem cannot be solved. To identify the best solution to our challenge, we have tried and implemented a number of different ways.<br>**EP4:** Problem familiarity: Getting a lot of data proved a little challenging. However, we've easily gathered a respectable amount of information.<br>**EP5:** Range of applicable codes: To address this issue, we have employed deep learning and machine learning architectures. The engineering standard is upheld. This project was constructed using both implementation and coding expertise.<br>**EP6:** Stakeholder engagement level and competing requirements: The project's stakeholders' perspective is taken into account.<br>**EP7:** Interdependence: We have tackled numerous sub-problems within high-level issues. |
| CO2 | Fast and precise picture understanding is essential in a time when AI innovation is revolutionizing industries. We are working on a sophisticated image caption | **(i) Define the problems [K8]**<br>**K8:** Research Literature: We have examined a sizable quantity of research articles that are pertinent to our work. After that, we came up with methods to deal with our issues. We have learned about many machine learning and deep learning models, among other things, from the research articles. | **(i) Define the problems [EP1, EP2, EP3, EP4, EP5, EP6, EP7]**<br>**[Same as (CO1)]** |

| | | | |
|---|---|---|---|
| | generator prototype. Our system creates accurate and evocative captions for photos using state-of-the-art technologies like computer vision and natural language processing. Through this project, students will be able to improve accessibility, drive automation, and understand deep learning frameworks. It seeks to transform the way images are viewed through its practical applications, providing significant solutions in a variety of fields. | | |

## CSE400-B

| CO | Details | Knowledge Profile(K) | Engineering Problem (EP) |
|---|---|---|---|
| CO3 | Evaluate several facets of the goals in order to create a solution for the capstone project. | **(i) Problem Analysis [K1, K2, K3, K4]**<br>**K1:** There are several kinds of photographs in this section.<br>**K2:** We performed statistical and quantitative analysis after reviewing previous research.<br>**K3:** Theory-based engineering expertise includes image processing, which was utilized to extract data from machine learning algorithms to anticipate outcomes.<br>**K4:** To obtain the best accurate outcome, many machine learning models are implemented. | **(i) Problem Analysis [EP1, EP2, EP3, EP6, EP7]**<br>**EP1:** Theory-based engineering knowledge includes using image processing techniques to retrieve data for machine learning algorithms that are intended to predict outcomes. This strategy is further supported by using a variety of machine learning models to achieve the most accurate outcomes.<br>**EP2:** The Flickr8k dataset's small size, Flickr30k's high processing costs, and the requirement for fine-tuning models like VGG16, Mobile-Net Large, and ResNet50 were some of the difficulties we encountered. Choosing appropriate evaluation criteria, controlling training durations, and optimizing hyperparameters were among the implementation challenges. Overcoming these challenges required effective project management.<br>**EP3:** We concentrated on overcoming obstacles by using transfer learning to enhance model performance and using data augmentation. VGG16, Mobile-Net Large, and ResNet50 are examples of pre-trained models that we fine-tuned and adjusted for efficiency. Improved assessment metrics and resource management made it |

| | | | |
|---|---|---|---|
| | | | possible to guarantee equitable comparisons and efficient advancement of our goals.<br>**EP6:** We carried out a thorough analysis of VGG16, Mobile-Net Large, and ResNet50, highlighting important findings and determining which model performed the best. Throughout the project, we considered the difficulties encountered, the solutions put in place, and the lessons discovered. Lastly, we suggested some next approaches, such as investigating novel architectures, growing datasets, and using the models for more extensive tasks.<br>**EP7:** A major challenge in deployment and scalability of the model were the main challenges. This involved making sure the models were stable with unseen data, integrating them into production systems, optimizing them for real-time performance, and modifying the solution to effectively manage bigger datasets or higher loads. |
| CO4 | Develop and put into practice capstone project solutions that consider cultural, social, and environmental factors | **K5:** Design and Implementation [K5] Statistical analysis and visualization, machine learning algorithms implementation. | **(i) Design and Implementation [EP1, EP2, EP4, EP5, EP6, EP7]**<br>**EP1:** The project's goal is to create an image caption generator using the VGG16, MobileNetV3 Large, and ResNet50 models on the |

|  |  |  |  |
|---|---|---|---|
|  |  |  | Flickr8k and Flickr30k datasets.<br>**EP2:** The Flickr8k and Flickr30k datasets should be gathered and preprocessed to ensure data diversity for improved model generalization.<br>**EP4:** Based on their individual strengths, choose and modify VGG16, MobileNetV3 Large, and ResNet50 for the creation of image captions.<br>**EP5:** Assess the models' ability to generate image captions and make adjustments in response to metrics and comments<br>**EP6:**Use a user-friendly application to implement the learned model, making sure it is accessible and scalable on many devices.<br>**EP7:** Complete the model, taking into account various user inputs, and specify potential paths for ongoing development.<br>information from it. |
| CO5 | Determine and use cutting-edge IT and engineering tools for the capstone project's design and development. | **K6:** We utilized the Flickr8k and Flickr30k datasets to conduct our experiments. For model architectures, we employed VGG16, Mobile-Net Large, and ResNet50 to extract features and analyse performance across various image-related tasks. | We leveraged the Flickr8k and Flickr30k datasets to conduct comprehensive experiments on Kaggle notebooks **[EP1, EP2, EP4, EP5]** , focusing on image caption generation. For model architectures, we implemented VGG16, Mobile-Net Large, and ResNet50 **[EP1].** , |

| | | | |
|---|---|---|---|
| | | | utilizing TensorFlow and PyTorch libraries to extract meaningful features and evaluate their performance across a range of image-related tasks[EP2], ensuring robust and accurate caption generation**[EP5].** |
| CO6 | Examine and address the societal, safety, legal, and cultural issues of the capstone project's implementation while taking into account pertinent engineering and professional practices and solutions. | Consequences for Safety and Societal Issues [K7] There have been no negative impacts on the legal, cultural, safety, or societal fronts. Respect for professional ethics is maintained [K7]. | Legal and ethical challenges may arise related to data usage, particularly concerning the ownership and potential biases in the Flickr8k and Flickr30k datasets, which could impact societal fairness **[EP2]**. Trust issues emerge regarding the reliability and contextual relevance of captions generated by models like VGG16, Mobile-Net Large, and ResNet50, especially in representing diverse cultures and environments **[EP5].** Varying stakeholder opinions on the safety and implications of the generated captions reflect differing concerns about potential misuse or misinterpretation of content, leading to debates on model transparency and accountability **[EP6].** |

# CSE400-C

**Program Outcomes:**

**PO7 (Environment and Sustainability):** Both society and the environment may be significantly impacted by our capstone project. By creating captions that emphasize environmentally friendly behaviors and lessen the use on paper-based items, it raises awareness of environmental issues. Through the analysis of medical photos and the provision of accurate captions, it can help healthcare practitioners improve patient care and diagnosis. By explaining images for those with visual impairments, the project improves accessibility and provides an affordable tool for marketing, education, and conservation initiatives. Action toward sustainability and societal advancement is encouraged by increasing awareness of global issues like pollution and climate change.

**PO8 (Ethics):** Our capstone project adheres to professional and engineering ethical standards to guarantee the responsible design and development of our intelligent tutoring system.to prevent bias and produce accurate, truthful captions, train the model on a variety of datasets. Respect user privacy by abstaining from identifiable personal information and following data protection rules. Make sure captions are accessible and inclusive, especially for users who are blind or visually challenged. Reduce your influence on the environment by using resources that use less energy. Put measures in place to stop abuse, like creating subtitles for offensive or fraudulent information. Responsible use also requires openness regarding the model's limits.

**PO9 (Individual Work and Teamwork):** The success of our capstone project depends on both individual and teamwork since they allow us to pool our skills and abilities to accomplish a shared objective. While working alone enables us to hone our own project-related abilities and expertise, cooperation promotes idea exchange, constructive criticism, and the creation of a thorough and workable solution as a whole. Collaboration, mutual support, and the capacity to tackle complicated problems that could call for a variety of viewpoints and levels of experience are further benefits of teamwork. Our capstone project benefits greatly from teamwork as it allows us to reach our full potential and produce significant results.

**P10 (Communication):** For our capstone project, communication with our supervisor, teammates, and other stakeholders is crucial to ensuring that we receive helpful criticism, direction, and support all along the way. Speaking with our supervisor enables us to get prompt feedback, resolve any problems or worries, and get advice on how to make our work better. Maintaining cohesiveness and staying on course with our aims and objectives is made possible by regular contact with our teams.

**P11 (Project Management and Finance):** We demonstrated effective project planning using an agile framework and budget management using functional points, team member engagement, outcome assessment feedback, and rating in order to finish our project for the update summarizing PO11. Those working in the computing industry must be proficient in project management and finance.

**P12 (Life-Long Learning):** We gain lifelong knowledge from our capstone project. We gained useful expertise using machine learning and natural language processing methods to create picture captioning systems, examine huge datasets, and assess their efficacy. Additionally, we acquired critical thinking, communication, problem-solving, and teamwork abilities that are beneficial in both academic and professional contexts. All things considered, the initiative helped us get ready for the opportunities and difficulties of the digital age.

| | | | |
|---|---|---|---|
| CO7 | Evaluate and discuss the capstone project's sustainability and effects on society and the environment. | **(i)Societal and environmental contexts [K7]**<br>**K7:**<br>There are important societal and environmental ramifications to the sustainability and impact of your picture caption generator project. By creating culturally appropriate subtitles, it ensures broader involvement with a variety of social contexts and advances accessibility and inclusivity in society. To maintain justice and avoid harm, ethical issues pertaining to bias and cultural sensitivity must be addressed. Adopting energy-efficient technology or cloud resources powered by renewable energy sources can assist lessen the project's environmental impact, as it requires significant computational resources. Additionally, resource usage can be further reduced by optimizing the model using strategies like transfer learning. The project must be flexible enough to accommodate upcoming dataset expansions in order to be sustainable over the long run and maintain its relevance. Sharing or open-sourcing the technology could promote sustainable innovation and public education. Fair portrayal and transparent data usage are examples of ethical AI activities. | (i) **Societal and Environmental Contexts [EP1, EP2, EP3, EP4, EP5, EP6, EP7]**<br>**Depth of knowledge required**: We reviewed various research papers to gather literature on image caption generation, including methods, algorithms, and machine learning techniques. Based on these insights, we are utilizing machine learning to detect and interpret images, with VGG16, Mobile-Net largeV3, Restnet50 for feature extraction.<br>**Range of conflicting requirements**: Given the range of conflicting engineering requirements, we found that the generation of captions may vary in accuracy and cultural sensitivity depending on factors like the dataset used and model training, highlighting the challenge of addressing societal diversity while maintaining technical efficiency. |

|  |  |  |  |
|---|---|---|---|
|  |  |  | **EP3: Depth of analysis required**: Our depth of analysis involves processing all the images to a consistent size. We use Neural Networks (NN) and Recurrent Neural Networks (RNN) to detect and analyze the images.<br>**EP4: Familiarity of issues**: Collecting datasets of diverse types of images presents challenges in ensuring they cover a wide range of scenarios for accurate captioning.<br>**EP5: Extent of applicable codes**: The extent of applicable codes from various methods are observed here, with most of them being applied to professional engineering standards for image processing and caption generation. |

| | | | |
|---|---|---|---|
| | | | **EP6: Extent of stakeholder involvement and conflicting requirements**: Our main stakeholders are visually impaired individuals, whose needs must be carefully considered when developing an accessible and accurate image captioning system.<br>**EP7: Interdependence**: While preprocessing data, we encountered several challenges, such as the inability to correctly detect all objects in the images, affecting the accuracy of the captions generated. |