मौलाना आजाद राष्ट्रीय प्रौद्योगिकी संस्थान - भोपाल

Maulana Azad National Institute of Technology– Bhopal


**Department of Computer Science and Engineering**
**कंप्यूटर विज्ञान और अभियांत्रिकी विभाग**

**Session- 2023-24**

**PROJECT-1 REPORT**
**On**
**"Enhancing accessibility through image captioning and text-to-speech technology "**

**Submitted In partial Fulfilment for the degree of Bachelor of Technology in Computer Science and Engineering**

**SUBMITTED BY:**

**Suhaani Batra (211112079)**

**Ankit Kumar (211112004)**

**Zanjabila Bano (211112048)**

**Aman Anand (211112229)**

**GUIDED BY:**

**Dr. Vikram Garg**

मौलाना आजाद राष्ट्रीय प्रौद्योगिकी संस्थान - भोपाल

Maulana Azad National Institute of Technology– Bhopal


**Department of Computer Science and Engineering**
**कंप्यूटर विज्ञान और अभियांत्रिकी विभाग**

**Session- 2023-24**

**PROJECT-1 REPORT**
**On**
**"Enhancing accessibility through image captioning and text-to-speech technology "**

**Submitted In partial Fulfilment for the degree of Bachelor of Technology in Computer Science and Engineering**

**SUBMITTED BY:**

**Suhaani Batra (211112079)**

**Ankit Kumar (211112004)**

**Zanjabila Bano (211112048)**

**Aman Anand (211112229)**

**GUIDED BY:**

**Dr. Vikram Garg**

मौलाना आजाद राष्ट्रीय प्रौद्योगिकी संस्थान - भोपाल

Maulana Azad National Institute of Technology– Bhopal


**Department of Computer Science and Engineering**

**कंप्यूटर विज्ञान और अभियांत्रिकी विभाग**

# CERTIFICATE

---

This is to certify that "**Suhaani Batra", "Ankit Kumar", "Zanjabila Bano" and "Aman Anand",** student of B.Tech 3rd Year (Computer Science & Engineering), have successfully completed their project **"Enhancing accessibility through Image Captioning and text-to-speech Technology"** in partial fulfilment of their Bachelor of Technology in Computer Science & Engineering.

**Dr. Vikram Garg**

मौलाना आजाद राष्ट्रीय प्रौद्योगिकी संस्थान - भोपाल

Maulana Azad National Institute of Technology– Bhopal


**Department of Computer Science and Engineering**

**कंप्यूटर विज्ञान और अभियांत्रिकी विभाग**

# DECLARATION

---

We hereby declare that the work, which is presented in this Project Report, entitled "**Enhancing accessibility through image captioning and text-to-speech technology**", in partial fulfilment of the requirements for the award of the degree, submitted in the **Department of Computer Science and Engineering, Maulana Azad National Institute of Technology, Bhopal.** It is an authentic record of our work carried out under the noble guidance of our guide "**Dr. Vikram Garg".** The following project and its report, in part or whole, have not been presented or submitted by us for any purpose in any other institute or organization. We hereby declare that the facts mentioned above are true to the best of our knowledge. In case of any unlikely discrepancy that may possibly occur, we will be the ones to take responsibility.

| **SCHOLAR NAME** | **Suhaani Batra** | **Ankit Kumar** | **Zanjabila Bano** | **Aman Anand** |
|---|---|---|---|---|
| **SCHOLAR NO** | **211112079** | **211112004** | **211112048** | **211112229** |

# ACKNOWLEDGEMENT

With due respect, we express our deep sense of gratitude to our respected Guide and project coordinator, **Dr. Vikram Garg** for his valuable help and guidance.

We are thankful for his encouragement in completing this project successfully. His rigorous evaluation and constructive criticism were very helpful.

We are also grateful to our respected director, **Dr. Karunesh Kumar Shukla** for permitting us to utilize all the necessary college facilities.

Needless to mention the additional help and support extended by our respected HOD, **Dr. Deepak Singh Tomar** in allowing us to use the departmental laboratories and other services.

We are also thankful to all the other faculty, staff members, and laboratory attendants of our department for their cooperation and help.

Last but certainly not least, we would like to express our deep appreciation towards our family members and batch mates for providing much-needed support and encouragement.

Place: Bhopal

Date: 25 April 2024

| Scholar Name | Scholar No |
|---|---|
| Suhaani Batra | (211112079) |
| Ankit Kumar | (211112004) |
| Zanjabila Bano | (211112048) |
| Aman Anand | (211112229) |

# ABSTRACT

आज की डिजिटल रूप से संचालित दुनिया में, प्रौद्योगिकी में अंतराल को पाटने और दिव्यांग व्यक्तियों के लिए पहुंच बढ़ाने की शक्ति है। इनमें से, दृष्टिबाधित समुदाय को छवियों में मौजूद दृश्य जानकारी तक पहुँचने में अद्वितीय चुनौतियों का सामना करना पड़ता है। जवाब में, इस परियोजना का लक्ष्य छवियों की सामग्री को समझने में नेत्रहीनों की सहायता के लिए टेक्स्ट-टू-स्पीच (टीटीएस) तकनीक के साथ छवि कैप्शनिंग के लिए एक समाधान विकसित करना है।

प्रस्तावित मॉडल छवि कैप्शनिंग के लिए गहन शिक्षण तकनीकों का उपयोग करता है, जहां एक तंत्रिका नेटवर्क को छवियों का पाठ्य विवरण उत्पन्न करने के लिए प्रशिक्षित किया जाता है। इसके अतिरिक्त, टीटीएस तकनीक को इन पाठ्य विवरणों को बोले गए शब्दों में परिवर्तित करने के लिए एकीकृत किया गया है, जिससे उपयोगकर्ता छवियों (दृश्य) की सामग्री को श्रव्य रूप से समझने में सक्षम हो जाते हैं।

इस परियोजना के माध्यम से, हम दृष्टिबाधित व्यक्तियों के लिए अधिक समावेशिता और स्वतंत्रता को बढ़ावा देने, सहायक प्रौद्योगिकियों की उन्नति में योगदान करने की आकांक्षा रखते हैं। यह अनुमान लगाया गया है कि यह परियोजना न केवल नेत्रहीनों के लिए एक मूल्यवान उपकरण के रूप में काम करेगी बल्कि प्रौद्योगिकी विकास में पहुंच के महत्व के बारे में जागरूकता भी बढ़ाएगी।

In today's digitally driven world, technology has the power to bridge gaps and enhance accessibility for individuals with disabilities. Among these, the visually impaired community faces unique challenges in accessing visual information present in images. In response, this project aims to develop a solution for image captioning coupled with text-to-speech (TTS) technology to aid the blind in comprehending the content of images.

The proposed model utilizes deep learning techniques for image captioning, where a neural network is trained to generate textual descriptions of images. Additionally, TTS technology is integrated to convert these textual descriptions into spoken words, thus enabling users to audibly perceive the content of images(scene).

Through this project, we aspire to contribute to the advancement of assistive technologies, fostering greater inclusivity and independence for individuals with visual impairments. It is anticipated that this project will not only serve as a valuable tool for the blind but also raise awareness about the importance of accessibility in technology development.

# LIST OF FIGURES

# LIST OF TABLES

# TABLE OF CONTENTS

# CHAPTER 1

# INTRODUCTION

## 1.1 Introduction

The problem of producing a description based on an image is called image caption generation. Automatically creating captions for an image is a challenge at the core of scene perception, which is one of computer vision's main objectives. Not only must caption generation models be capable of overcoming the computer challenges of deciding what objects are present in an image, but they must also be capable of capturing and expressing their relationships in natural language.

Early approaches to this task involved a combination of object detection, attribute discovery, and caption templating. However, with the advent of attention mechanisms and deep learning, the field has advanced considerably, particularly with the rise of encoder-decoder models.

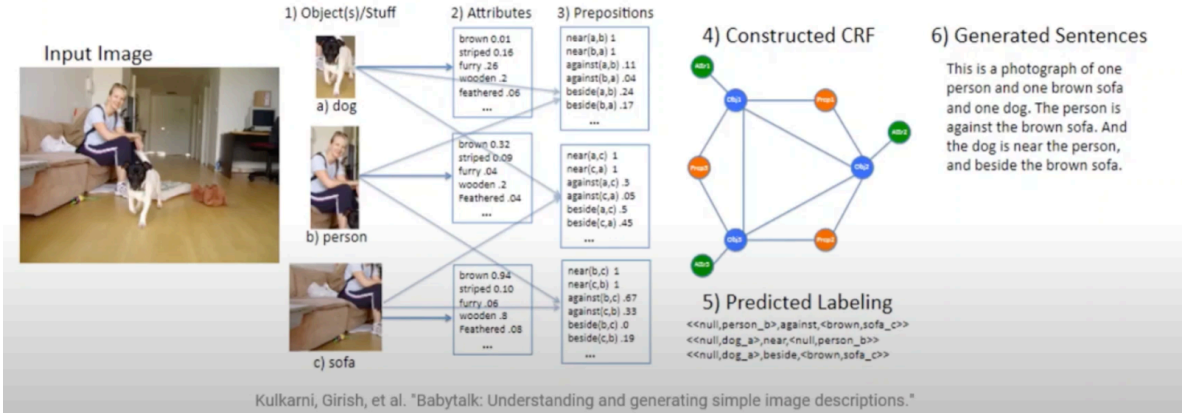


Figure 1.1 Image captioning model based on object detection and caption templating

Despite these advancements, existing methods for captioning images, particularly for the visually impaired, often overlook the textual data present within images. Given that many visual scenes contain text crucial for understanding, this oversight poses a significant barrier to accessibility.

The inability to access visual information hinders the independence and autonomy of the visually impaired in daily life. Activities such as navigating unfamiliar environments, shopping for groceries, or reading instructions become daunting challenges without access to descriptive information about the surroundings. By bridging the gap between visual content and accessible descriptions, our project aims to empower individuals with visual impairments to engage more fully with the world around them.

This project addresses this gap by developing a multi-modal caption generator capable of leveraging both visual and textual information present in images. By harnessing convolutional neural networks to extract image features and decoders to generate descriptive captions, our system aims to provide succinct and informative descriptions of visual content. The ultimate goal is to empower the visually impaired community by enabling them to access and comprehend visual information through audio descriptions generated from images.

## 1.2 Understanding the Dataset

The proposed model is trained using the Flickr30k dataset. The Flickr30k dataset serves as the backbone of our image caption generation project, providing a diverse collection of images paired with corresponding descriptive captions. This dataset comprises 31,000 images, each associated with five unique captions, totalling approximately 155,000 captions.

### 1.2.1 Data Collection and Characteristics:

Data Source - University of Illinois

The images in the Flickr30k dataset were sourced from the photo-sharing platform Flickr, ensuring a broad spectrum of visual content encompassing various scenes, objects, and contexts. Each image is accompanied by human-generated captions, capturing the essence of the depicted scene with linguistic diversity and richness.

### 1.2.2 Annotation and Preprocessing:

Prior to utilizing the dataset for training our image captioning model, several preprocessing steps were undertaken to ensure data quality and consistency. These steps included:

- Removal of duplicate images and captions to mitigate redundancy.
- Cleaning and normalization of text, including punctuation removal and lowercasing.

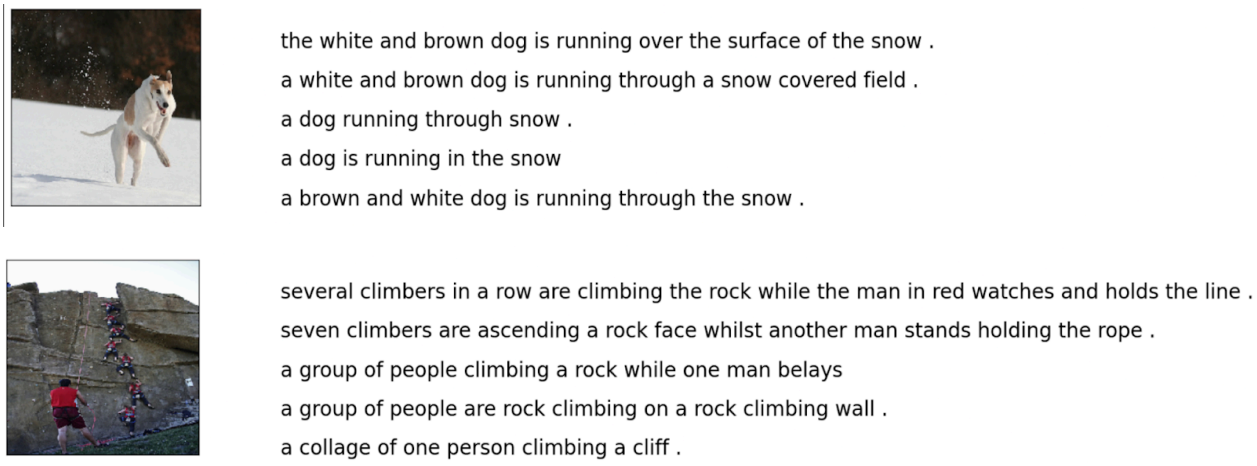


Figure 1.2 Sample Images with captions from the Flickr30k dataset each image has 5 captions.

# 1.3 Exploratory Data Analysis

## 1.3.1 Caption Statistics:

- Analysis of the most frequently and least frequently occurring words

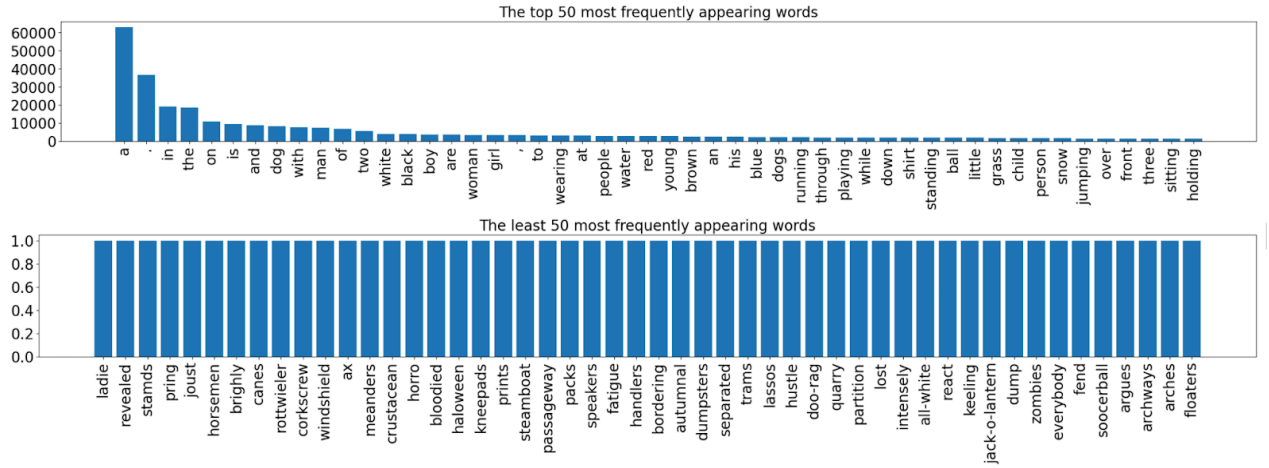


Figure 1.3 Frequencies of the most frequently and least frequently occurring words in the dataset

- Examination of caption length distribution (number of words per caption).

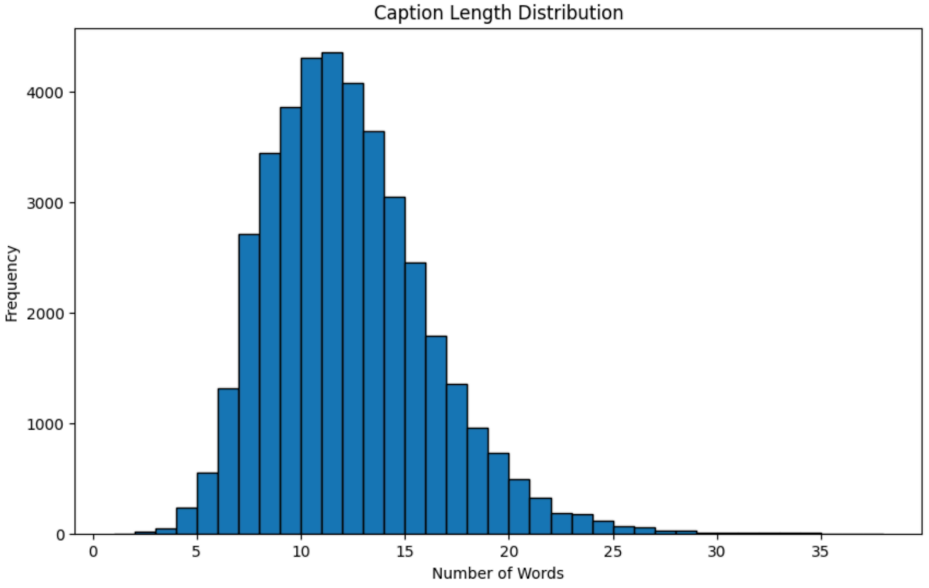


Figure 1.4 Analysis of caption lengths, number of words per caption

# CHAPTER 2

# LITERATURE REVIEW

The review encompasses a diverse range of methodologies, including attention mechanisms, transformer-based models, multi-modal approaches, and deep visual-semantic alignments.

Bahdanau et al. proposed an attention mechanism to enhance traditional encoder-decoder models' performance by selectively focusing on different parts of the input sequence. Unlike basic models that compress the entire input into a fixed-length vector, this approach translates input sequences part by part, dynamically adjusting focus based on relevance. This iterative process mitigates information loss, particularly with longer inputs, resulting in more accurate translations. This attention-based approach has become a standard technique in various natural language processing tasks, improving sequence-to-sequence modelling significantly.

"Show, Attend and Tell: Neural Image Caption Generation with Visual Attention", introduced the concept of attention mechanisms in image captioning. It proposed a model that dynamically focuses on different parts of the image while generating captions. The attention mechanism enhances the model's ability to capture relevant visual information and generate contextually relevant captions.

Another End-to-End Transformer Based Model presented a novel approach to image captioning using Transformer-based architectures. Adapt transformers, originally designed for natural language processing tasks, to the task of image captioning. Achieves state-of-the-art results by eliminating the need for preprocessing steps such as feature extraction.

These advancements have significantly improved the accuracy and contextual relevance of generated captions, addressing diverse challenges such as visual attention, multi-modal understanding, and inclusivity for visually impaired users.

# CHAPTER 3

# RESEARCH GAPS

Many existing image captioning models excel on datasets they were trained on but struggle to generalize to unseen or out-of-domain data. Enhancing the generalization capabilities of these models is essential for real-world deployment across diverse domains and scenarios.

While multi-modal approaches have shown promise in generating descriptive captions, there is room for improvement in understanding and integrating information from different modalities, such as text and vision, more effectively.

Efforts have been made to cater to the needs of visually impaired individuals through multi-modal captioning approaches. However, further research is needed to enhance the accessibility and inclusivity of image captioning systems for individuals with diverse abilities and needs.

Some image captioning models, particularly those based on complex architectures like transformers, may face challenges related to computational efficiency and scalability. Developing efficient and scalable models is essential for deploying image captioning systems in resource-constrained environments.

# CHAPTER 4
# METHODOLOGY

## 4.1 Concept of Attention and the Attention Mechanism

Attention is a widely investigated concept that has often been studied in conjunction with arousal, alertness, and engagement with one's surroundings.

*Visual* attention is one of the areas most often studied from both the neuroscientific and psychological perspectives. Since the human brain has a limited memory capacity, then selecting which information to store becomes crucial in making the best use of the limited resources. The human brain does so by relying on attention, such that it dynamically stores in memory the information that the human subject most pays attention to.

### 4.1.1 Need for attention

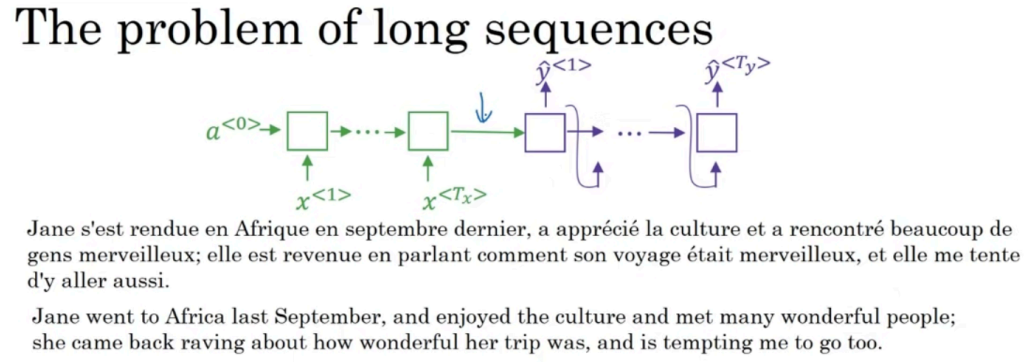


Figure 4.1 The problem of long sequences and the need for attention mechanism for learning context

#### 4.1.1.1 Attention in Machine Learning

1. A process that "reads" raw data (such as source words in a source sentence), and converts them into distributed representations, with one feature vector associated with each word position.
2. A list of feature vectors storing the output of the reader. This can be understood as a "memory" containing a sequence of facts, which can be retrieved later, not necessarily in the same order, without having to visit all of them.
3. A process that "exploits" the content of the memory to sequentially perform a task, at each time step having the ability to put attention on the content of one memory element (or a few, with a different weight).

#### 4.1.1.2 Attention in Machine Translation

If we are processing an input sequence of words, then this will first be fed into an encoder, which will output a vector for every element in the sequence. This corresponds to the first component of our attention-based system, as explained above.

A list of these vectors (the second component of the attention-based system above), together with the decoder's previous hidden states, will be exploited by the attention mechanism to dynamically highlight which of the input information will be used to generate the output.

At each time step, the attention mechanism then takes the previous hidden state of the decoder and the list of encoded vectors, using them to generate unnormalized *score* values that indicate how well the elements of the input sequence align with the current output. Since the generated score values need to make relative sense in terms of their importance, they are normalized by passing them through a softmax function to generate the *weights*. Following the softmax normalization, all the weight values will lie in the interval [0, 1] and add up to 1, meaning they can be interpreted as probabilities. Finally, the encoded vectors are scaled by the computed weights to generate a *context vector*. This attention process forms the third component of the attention-based system above. It is this context vector that is then fed into the decoder to generate a translated output.

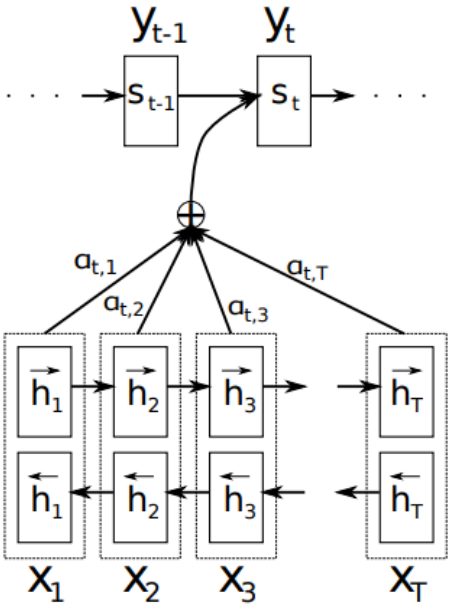


Figure 4.2 Graphical illustration of attention model trying to generate the t-th target word $y_t$ given a source sentence ($x_1$, $x_2$.... $x_T$).

## 4.2 Model Architecture

Attention mechanism - Visual Attention

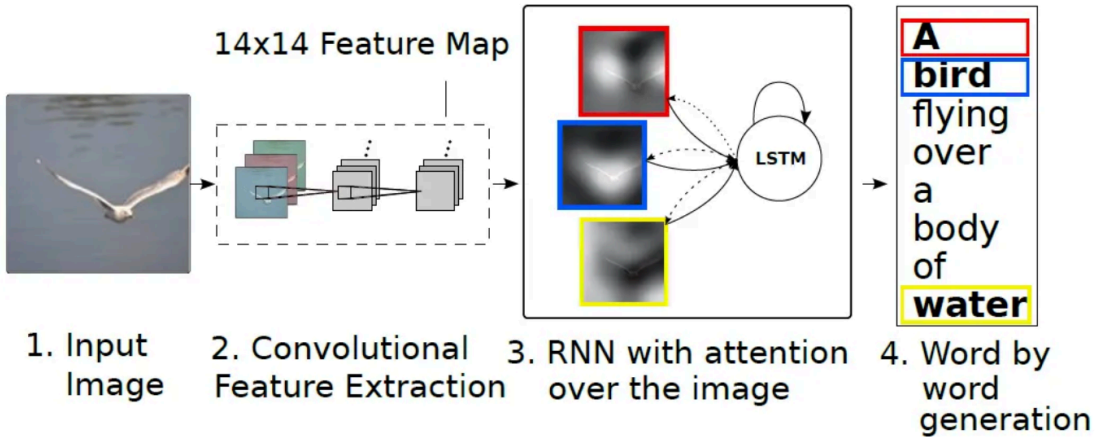


Figure 4.3 Show, Attend and Tell network architecture

### 4.2.1 CNN Encoder

The model takes a single raw image and generates a caption y encoded as a sequence of 1-of-K encoded words.

$$y = \{\mathbf{y}_1, \ldots, \mathbf{y}_C\}, \ \mathbf{y}_i \in \mathbb{R}^K$$

where K is the size of the vocabulary and C is the length of the caption.

A convolutional neural network (CNN) is used to extract a set of feature vectors which we refer to as annotation vectors. The extractor produces L vectors, each of which is a D-dimensional representation corresponding to a part of the image.

$$a = \{\mathbf{a}_1, \ldots, \mathbf{a}_L\}, \ \mathbf{a}_i \in \mathbb{R}^D$$

Features are extracted from a lower convolutional layer. This allows the decoder to selectively focus on certain parts of an image by weighting a subset of all the feature vectors.

Here, we have used the pre-trained model EfficientNetB0.

#### 4.2.1.1 Efficient Net Architecture

The EfficientNet B0 architecture is employed for feature extraction in the image captioning model. Developed by Google researchers, EfficientNet represents a family of convolutional neural network (CNN) models that achieve state-of-the-art performance while maintaining computational efficiency. The "B0" variant refers to the baseline model in the EfficientNet family.

*How EfficientNet Works:*

1. Compound Scaling: EfficientNet introduces a novel compound scaling method that uniformly scales the network's depth, width, and resolution simultaneously. This approach allows for optimal resource allocation and achieves better performance with fewer parameters.
2. Mobile-First Design: The architecture is designed with a mobile-first approach, prioritizing efficiency and computational cost. By incorporating efficient building blocks such as depthwise separable convolutions and squeeze-and-excitation blocks, EfficientNet achieves impressive accuracy while minimizing computational resources.

*Role in the Model Architecture:*

In our model, the EfficientNet B0 architecture serves as the backbone for feature extraction from input images. Pre-trained on a large-scale image classification dataset (such as ImageNet), the EfficientNet B0 model learns to extract high-level visual features that are relevant for generating descriptive captions.

By leveraging the EfficientNet B0 architecture, the image captioning model benefits from both the accuracy of the extracted features and the computational efficiency of the network. This allows for faster inference times and enables the model to be deployed in resource-constrained environments without sacrificing performance.

Overall, the integration of the EfficientNet B0 architecture enhances the effectiveness and efficiency of the image captioning model, contributing to its ability to generate accurate and contextually relevant captions for a wide range of images.

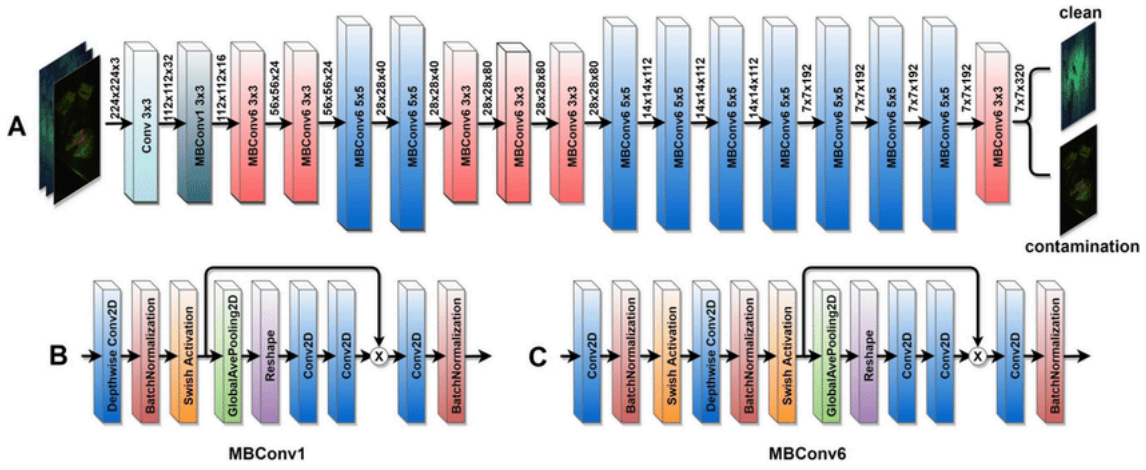


Figure 4.4 EfficientNet Architecture

### 4.2.2 Attention Decoder

A long short-term memory (LSTM) network is used that produces a caption by generating one word at every time step conditioned on a context vector, the previous hidden state and the previously generated words.
The context vector ^zt is a dynamic representation of the relevant part of the image input at time t.

For each location i, the mechanism generates a positive weight i which can be interpreted either as the probability that location i is the right place to focus for producing the next word (stochastic attention mechanism), or as the relative importance to give to location i in blending the ai's together (deterministic attention mechanism).

The weight of each annotation vector ai is computed by an attention model fatt for which a multilayer perceptron is used which is conditioned on the previous hidden state ht-1.
To emphasize, the hidden state varies as the output RNN advances in its output sequence: "where" the network looks next depends on the sequence of words that have already been generated.

$$e_{ti} = f_{\text{att}}(\mathbf{a}_i, \mathbf{h}_{t-1}) \quad \text{(eq. 4.1)}$$

$$\alpha_{ti} = \frac{\exp(e_{ti})}{\sum_{k=1}^{L} \exp(e_{tk})} \quad \text{(eq. 4.2)}$$

Once the weights (which sum to one) are computed, i.e. softmax, the context vector ^zt is computed by:

$$\hat{\mathbf{z}}_t = \phi\left(\{\mathbf{a}_i\}, \{\alpha_i\}\right) \quad \text{(eq. 4.3)}$$

where Φ is a function that returns a single vector given the set of annotation vectors and their corresponding weights.

### 4.2.3 Stochastic "Hard" Attention & Deterministic "Soft" Attention

Deterministic models are based on precise inputs and produce the same output for a given set of inputs. These models assume that the future can be predicted with certainty based on the current state. On the other hand, stochastic models incorporate randomness and uncertainty into the modelling process. They consider the probability of different outcomes and provide various possible results.

The hard attention focuses only on the part it wants and ignores other parts. Only the most important parts of the image are chosen, the rest are ignored (skipped). It is evaluated using probabilistic methods, instead of weights, wherein, each feature of the image has a certain 'chance' of getting picked up.
Soft attention is smoother. It involves weights, wherein, each feature is assigned some weight and an overall context vector is created.

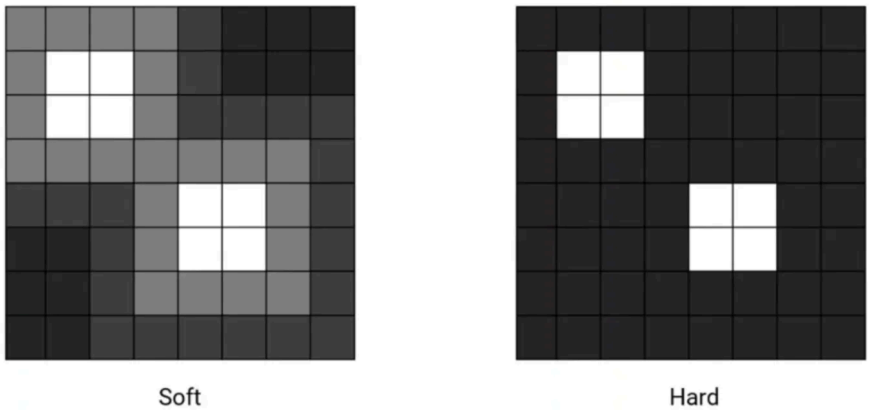


Figure 4.5 Soft and Hard Attention Mechanism

*Limitations of CNN-LSTM Method:*

The conventional approach of using a combination of Convolutional Neural Networks (CNNs) for image encoding and Long Short-Term Memory (LSTM) networks for caption generation, known as the CNN-LSTM method, has been widely used in image captioning tasks. However, this approach suffers from several limitations:

1. Sequential Processing: In the CNN-LSTM method, the image features extracted by the CNN are fed into the LSTM sequentially, word by word, to generate captions. This sequential processing can be computationally expensive and time-consuming, especially for long sequences or large datasets. As a result, training CNN-LSTM models may require significant computational resources and time.
2. Limited Context: LSTMs have a fixed-size memory and struggle to capture long-range dependencies within the input sequences effectively. This limitation can lead to difficulties in capturing complex relationships between different parts of the image features and generating accurate captions. In tasks like image captioning, where understanding the context of the entire image is crucial for generating meaningful captions, the limited context provided by LSTMs may be inadequate.
3. CNN Processing Overhead: Furthermore, the reliance on CNNs for image encoding in the CNN-LSTM method can introduce additional processing overhead. CNNs are primarily designed for tasks like object detection and feature extraction, which may require extensive computation and time. In contrast, the Transformer-based approach eliminates the need for CNN-based encoding, resulting in faster processing and improved efficiency.

### 4.2.4 Transformer-based model

In recent years, Transformer-based architectures have gained popularity in various natural language processing tasks, including image captioning. The Transformer model replaces the CNN encoder and RNN decoder with self-attention mechanisms, offering better performance and scalability.

- Transformer Encoder Block: The Transformer encoder block consists of multiple layers, each containing self-attention mechanisms followed by feed-forward neural networks (FFNs). Self-attention allows the model to capture global dependencies within the input sequence, enabling it to understand relationships between different parts of the image features. The FFNs process the outputs of the self-attention layers to further refine the representations.

- Transformer Decoder Block: The Transformer decoder block also comprises multiple layers, but it incorporates additional attention mechanisms to attend to both the encoder outputs and previous decoder outputs. This enables the model to generate captions by attending to relevant parts of the image features and previously generated words. The decoder also includes positional embeddings to encode the order of words in the caption.

#### 4.2.4.1 Transformer-based Model Advantages

1. Parallelization: Transformers can process input sequences in parallel, making them more efficient for training and inference compared to RNNs, which process sequences sequentially.
2. Long-range Dependencies: Self-attention mechanisms in Transformers allow the model to capture long-range dependencies within the input sequences more effectively, leading to better performance in tasks like image captioning.
3. Scalability: Transformers can handle input sequences of variable lengths without the need for recurrent connections, making them more scalable to larger datasets and longer sequences.

#### 4.2.4.2 Multi-head Attention Scheme

The Multi-head attention scheme enables the model to focus on different parts of the input sequence simultaneously, facilitating the capture of complex relationships and dependencies. Each attention head attends to different parts of the input, and the outputs are combined to produce a comprehensive representation. This scheme enhances the model's ability to extract meaningful information from the input features and generate accurate captions.

### 4.2.5 Data Augmentation

The technique of artificially increasing the training set by creating modified copies of a dataset using existing data. Includes making minor changes to the dataset or using deep learning to generate new data points

Augmented Data is driven from original data with some minor changes. Like making geometric or colour space transformations (flipping, resizing, cropping, brightness, contrast) to increase the size and diversity of the training set

Synthetic data is generated artificially using the original dataset using DNNs and GANs.

#### 4.2.5.1 When to use Data Augmentation?

1. To prevent models from overfitting
2. The initial training set is too small
3. To improve the model's accuracy
4. To reduce the operational cost of labelling and cleaning the raw dataset

#### 4.2.5.2 Limitations of Data Augmentation

- Biases in the original dataset persist
- Quality assurance of data augmentation is expensive
- Research and development are required to build. A system with advanced applications

#### 4.2.5.3 Image Augmentation

Image augmentation may involve the following:

1. Geometric transformations random flip, crop, rotate, stretch, zoom
2. Color space transformations - randomly change RGB colour channels, contrast and brightness
3. Kernel filters - randomly change sharpness or blurring
4. Random erasing
5. Mixing images

# CHAPTER 5
# TRAINING AND TESTING

The training phase of an image captioning model is a crucial aspect of the model development process, determining its ability to accurately generate captions for images. Typically, the model is trained over multiple epochs, with each epoch representing one complete pass through the dataset. The number of epochs required for training varies depending on factors such as the complexity of the model architecture, the size of the dataset, and the desired level of performance.

During training, it is essential to monitor the model's performance using key metrics such as accuracy and loss. These metrics provide insights into how well the model is learning from the data and making predictions.

Accuracy vs. Epoch:
The accuracy vs. epoch curve illustrates the model's accuracy as it progresses through training epochs. In the early epochs, the model's accuracy tends to increase rapidly as it learns to recognize patterns in the training data. However, as training continues, the rate of improvement may diminish, and the accuracy may plateau. This plateau indicates that the model has captured most of the information present in the training data and may not benefit significantly from further training.

Loss vs. Epoch:
The loss vs. epoch curve depicts the model's loss, typically measured using metrics such as cross-entropy loss, over the course of training epochs. Initially, the loss decreases as the model learns to make more accurate predictions. However, as training progresses, the loss may reach a minimum value and then begin to increase. This increase in loss suggests that the model is overfitting to the training data, capturing noise rather than genuine patterns.

Analyzing the accuracy and loss curves provides valuable insights into the training process and helps determine the optimal number of epochs needed to train the model effectively. It is essential to strike a balance between achieving high accuracy on the training data and ensuring the model generalizes well to unseen data. Techniques such as early stopping can be employed to halt training when the model's performance begins to deteriorate, preventing overfitting and improving generalization.

In our study, we conducted extensive experiments to analyze the accuracy vs. epoch and loss vs. epoch curves for our image captioning model. Through careful examination of these

curves, we determined the optimal number of epochs needed to train our model effectively, balancing performance and generalization. Additionally, we employed techniques such as early stopping to prevent overfitting and improve the model's ability to generate accurate captions for a wide range of images.

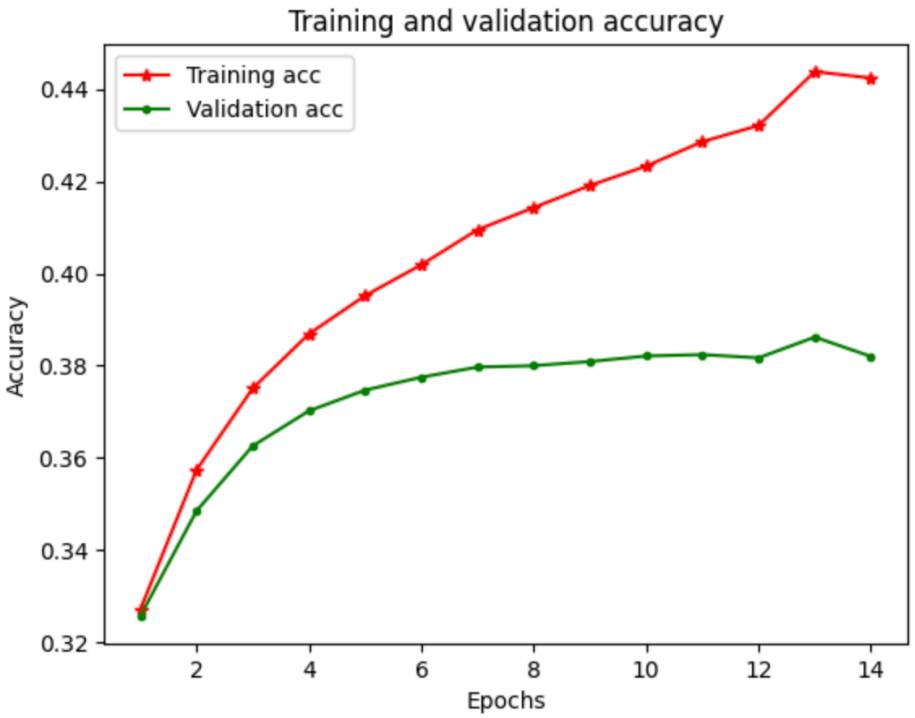


Figure 5.1 Training and validation accuracy vs Number of epochs

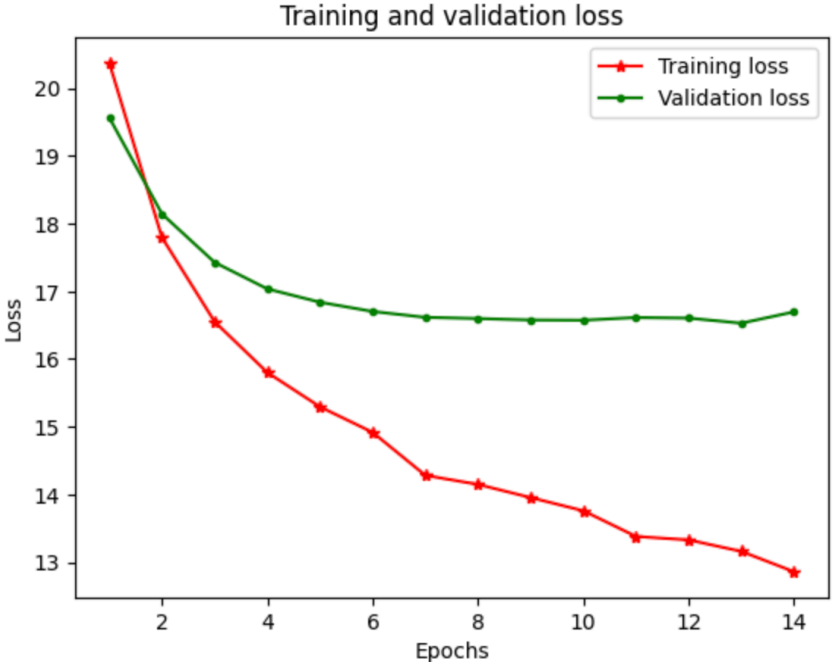


Figure 5.2 Training and validation loss vs Number of epochs

# CHAPTER 6
# RESULTS

## 6.1 Evaluation Metrics - BLEU Score

The BLEU (Bilingual Evaluation Understudy) score is a widely used metric in natural language processing tasks, such as machine translation and image captioning, to measure the quality of generated text compared to reference text. It addresses the challenge of multiple possible answers to the same question by providing a quantitative measure of the similarity between the generated text and reference text.

In essence, the BLEU score quantifies how well the generated text aligns with the reference text. Given a machine-generated caption and one or more reference captions for an image, the BLEU score is computed automatically to determine the quality of the generated caption.

The intuition behind the BLEU score is straightforward: the closer the machine-generated caption is to the reference captions, the higher the BLEU score it receives. This indicates that the generated caption captures the essence and context of the image description effectively.

The BLEU score is calculated based on several components, including precision measures for unigrams, bigrams, and trigrams. These measures assess how well the individual words and phrases in the generated caption match those in the reference captions. Additionally, a brevity penalty is applied to account for cases where the generated caption is significantly shorter than the reference captions, penalizing overly concise outputs.

The term "understudy" in BLEU refers to its role as a substitute for human evaluators in assessing the quality of machine-generated text. By providing an automated and objective evaluation metric, BLEU facilitates the comparison of different models and enables researchers to track improvements in performance over time.

Overall, the BLEU score serves as a valuable tool for evaluating the accuracy and fluency of generated text in natural language processing tasks, offering insights into the performance of machine learning models and guiding further refinement and development efforts.

## 6.2 Evaluation Methods

### 6.2.1 Greedy Search

Greedy search is a simple yet commonly used decoding strategy in sequence generation tasks, including image captioning. In the context of image captioning, greedy search involves generating the caption one word at a time, selecting the most likely word at each step based on the model's predictions.

The process of greedy search begins with an initial input, often a special start-of-sequence token, which serves as the first input to the model. The model then generates a probability distribution over the entire vocabulary for the next word in the sequence. In greedy search, the word with the highest probability is selected as the next word in the sequence.

After selecting the next word, it is appended to the sequence, and the process is repeated iteratively until a special end-of-sequence token is generated or a predefined maximum sequence length is reached.

Greedy search is computationally efficient and straightforward to implement, making it a popular choice for generating captions in real-time applications. However, it has some limitations:

1. Lack of Global Optimization: Greedy search selects each word independently based solely on its local probability, without considering how the entire sequence fits together. As a result, the generated captions may lack coherence and fail to capture long-range dependencies in the data.

2. Suboptimal Solutions: Because greedy search makes locally optimal decisions at each step, it may lead to suboptimal solutions overall. In some cases, the most likely word at each step may not lead to the best overall sequence, resulting in captions that are less fluent or descriptive.

### 6.2.2 Beam Search

Beam search is an advanced decoding algorithm used in sequence generation tasks, such as machine translation and image captioning. Unlike greedy search, which selects the most likely word at each step, beam search explores multiple possible sequences simultaneously, maintaining a set of candidate sequences known as the beam.

The beam search algorithm begins by initializing the beam with a single sequence, often consisting of a special start-of-sequence token. At each step, the model generates a probability distribution over the entire vocabulary for the next word in each candidate

sequence. Instead of selecting only the most likely word, beam search retains the top-k words (where k is the beam width) based on their probabilities.

After generating the probability distributions for each candidate sequence, beam search expands the beam by selecting the top-k sequences with the highest combined probabilities. These selected sequences become the new candidates for the next step, and the process is repeated iteratively until a special end-of-sequence token is generated or a maximum sequence length is reached.

The beam width, or beam size, is a crucial parameter in beam search that determines the number of candidate sequences retained at each step. A larger beam width allows beam search to explore a more extensive search space and consider a broader range of possible sequences. However, increasing the beam width also comes with computational costs, as it requires more memory and computational resources to maintain and evaluate a larger set of candidate sequences.

The beam width has a significant impact on the quality and diversity of the generated captions. A smaller beam width may lead to faster decoding but may also result in less diverse and potentially suboptimal captions, as the algorithm may converge to a locally optimal solution too quickly. On the other hand, a larger beam width can produce more diverse captions by exploring a broader range of possibilities, but it may also increase computational overhead and decoding time.

# 6.3 Result analysis

## 6.3.1 Testing Observations

Upon testing the image captioning model on various images, several notable observations were made:

1. **Detection of Intricate Details:** Adjusting the number of epochs the model is trained on resulted in improved detection of intricate details within images. The model demonstrated the ability to identify subtle elements and background context, indicating enhanced comprehension of image content.
2. **Partial Scene Understanding:** In some instances, even when the model failed to detect the main subject of the image, it still generated captions that partially described the scene or context accurately. This suggests a degree of contextual understanding and the model's capability to infer meaning from visual cues.
3. **Challenges with Rare Elements:** Images containing rare or uncommon elements, such as a woman wearing a saree or a man holding a shell, posed challenges for the model. These instances highlighted potential limitations in the model's training data diversity and its ability to generalize to less common visual concepts.
4. **Colour Detection Accuracy:** Training the model on fewer epochs occasionally resulted in inaccuracies in colour detection. Objects in the images were sometimes

misclassified or labelled with incorrect colours, indicating the importance of sufficient training iterations for colour recognition tasks.

5. **Gender Misclassification:** In some cases, individuals with longer hair were misclassified as women by the model. This observation suggests a potential bias or limitation in the model's gender classification capabilities, emphasizing the need for further refinement and diversity in training data.
6. **Effect of Maximum Sequence Length:** The fixed maximum sequence length of 25 tokens impacted the performance of captions generated by beam search. Although the model may have been capable of detecting more details in the image, the constrained sequence length limited the complexity and expressiveness of the captions, resulting in lower BLEU scores compared to greedy search.

### 6.3.2 Impact of Training Epochs and Search Strategies:

1. Effect of Training Epochs: Altering the number of training epochs influenced the model's ability to detect intricate details and accurately describe image content. Higher epochs resulted in improved caption quality and scene understanding, while lower epochs led to reduced accuracy and occasional misclassifications.
2. Greedy Search vs. Beam Search: Comparative analysis between greedy search and beam search revealed differences in caption diversity and quality. Greedy search tended to produce more conservative captions, while beam search allowed for greater exploration of diverse caption alternatives. However, beam search required longer inference times and increased computational resources compared to greedy search.

| **Epochs** | **Method** | **Average BLEU Score** |
|---|---|---|
| 15 | Greedy | 0.131 |
| 15 | Beam (k = 2) | 0.137 |
| 15 | Beam (k = 3) | 0.139 |
| 50 | Greedy | 0.304 |
| 50 | Beam (k = 2) | 0.311 |
| 50 | Beam (k = 3) | 0.319 |

Table 6.1 Evaluation of model

*Results on a sample image*

Figure 6.1 Sample image from the dataset

Reference captions:

1. A woman is sitting on a beach with her shoes off and on a cellphone talking , scratching her shoulder , while a group of 5 other people are walking in the background.
2. A girl in a hat sit on a beach while talking on her cellphone .
3. Woman in hat and talking on cellphone sitting on the beach .
4. A woman sits on the beach while talking on her cellphone .
5. A lady sits in the sand on the beach .

Results on number of epochs = 10

```
Greedy: a woman sitting on a beach a sunny day on the beach
BLEU score for greedy -> 0.34172334076593075

Beam_2: a woman sitting on the beach the sand on the beach the sun of the ocean
BLEU score for beam search 2 -> 0.35831291876413535

Beam_3: a young girl sitting on a beach sand dune a sunny day on the beach the sun the background
BLEU score for beam search 3 -> 0.21951524426618454

Beam_4: a young girl sitting on a beach sand dune on a sunny day the beach the sun the background
BLEU score for beam search 4 -> 0.19835441454182887

Beam_5: a young girl is sitting on the sand on the beach of the sand on the beach
BLEU score for beam sesarch 5 -> 0.3278603771984186
```

Figure 6.2 Captions generated on the sample image with a model trained on 10 epochs

Results on number of epochs = 12

```
Greedy: a woman in a bikini top and shorts is sitting on a beach a sunny day
BLEU score for greedy -> 0.32069326294908396

Beam_2: a woman in a bikini top is sitting on the beach the sand from the beach
BLEU score for beam search 2 -> 0.27882410979222033

Beam_3: a woman sitting on the beach with her purse on the beach the background the background a man in the background
BLEU score for beam search 3 -> 0.23198210427894822

Beam_4: a woman is sitting on the beach the sun the background of the ocean the background the background
BLEU score for beam search 4 -> 0.3457913759237496

Beam_5: a woman is sitting on the beach the sun the background of the ocean the background the background
BLEU score for beam sesarch 5 -> 0.3457913759237496
```

Figure 6.3 Captions generated on the sample image with a model trained on 12 epochs

Results on number of epochs = 14

```
Greedy: a woman sitting on the beach the sun from
BLEU score for greedy -> 0.36889397323344053

Beam_2: a woman sitting on the beach the ocean from the beach
BLEU score for beam search 2 -> 0.28997844147152074

Beam_3: a young girl sitting on the beach the ocean from the beach
BLEU score for beam search 3 -> 0.24384183193426084

Beam_4: a woman sitting on the beach with her head on the beach from the beach
BLEU score for beam search 4 -> 0.2722589423069702

Beam_5: a young girl laying on the beach the sun of the ocean from the beach
BLEU score for beam sesarch 5 -> 0.08085298080223222
```

Figure 6.4 Captions generated on the sample image with a model trained on 14 epochs

Results on number of epochs = 15

```
Greedy: a woman sitting on the beach the sand the ocean
BLEU score for greedy -> 0.3549481056010053

Beam_2: a woman sitting on the beach with her purse on the beach the background in the foreground background a
BLEU score for beam search 2 -> 0.24065223308491276

Beam_3: a woman is sitting on the beach with her purse on her shoulder a cloudy day the background in the background
BLEU score for beam search 3 -> 0.40931841311707223

Beam_4: a woman is sitting on the beach with her purse on her shoulder a cloudy day the background in the background
BLEU score for beam search 4 -> 0.40931841311707223

Beam_5: a woman is sitting on the beach with her purse on her shoulder a bridge
BLEU score for beam sesarch 5 -> 0.5169731539571706
```

Figure 6.5 Captions generated on the sample image with a model trained on 15 epochs

Results on number of epochs = 50

```
Greedy: a woman sitting on the beach the sand of a beach
BLEU score for greedy -> 0.3508439695638686

Beam_2: a woman sitting on the beach with her phone in the sand of the ocean
BLEU score for beam search 2 -> 0.3290385879986622

Beam_3: a woman is sitting on the beach the sand of a beach from the sand
BLEU score for beam search 3 -> 0.4480304273880272

Beam_4: a woman is sitting on the beach the sand of a beach
BLEU score for beam search 4 -> 0.5773502691896257

Beam_5: a woman is laying on the beach with her purse on her shoulder walking along the beach
BLEU score for beam sesarch 5 -> 0.15287680537655193
```

Figure 6.6 Captions generated on the sample image with a model trained on 50 epochs

# CHAPTER 7
# FUTURE SCOPE

Key areas for future exploration include:

1. **Advanced Model Training:** Further refinement of the model training process can significantly improve its performance and generalization capabilities. This includes exploring novel training techniques, such as curriculum learning and self-supervised learning, to enhance the model's ability to understand diverse visual contexts and nuances.
2. **Training on Larger Datasets:** Scaling up the training dataset to include a more extensive and diverse range of images can improve the model's robustness and adaptability to real-world scenarios. Leveraging large-scale image datasets, such as COCO (Common Objects in Context) and Open Images, can enable the model to learn from a broader spectrum of visual concepts and contexts.
3. **Expansion to Video Captioning:** Extending the image captioning model to support video captioning represents a promising direction for future research. By incorporating temporal information and sequential dependencies present in video data, the model can generate descriptive captions that capture dynamic visual narratives and events, opening up new opportunities for applications in video content analysis and understanding.
4. **Accessibility for the Visually Impaired:** Integrating the image captioning model into assistive technologies, such as wearable devices and mobile applications, can empower individuals with visual impairments to access and interact with visual content more independently. By providing real-time audio descriptions of the surrounding environment, the model can enhance accessibility and inclusivity for visually impaired users, facilitating their participation in various activities and interactions.
5. **Ethical Considerations and Bias Mitigation:** Addressing ethical considerations and mitigating biases in the image captioning model are essential aspects of future research. This includes evaluating the model's performance across diverse demographic groups and cultural contexts, as well as implementing strategies to minimize biases in training data and model predictions, ensuring fairness and equity in its applications.

# CHAPTER 8
# REFERENCES


1. D. Bahdanau, K. Cho, and Y. Bengio, “Neural machine translation by jointly learning to align and translate,” *arXiv.org*, Sep. 01, 2014. https://arxiv.org/abcs/1409.0473
2. K. Xu *et al.*, “Show, Attend and Tell: Neural Image Caption Generation with Visual Attention,” *arXiv.org*, Feb. 10, 2015. https://arxiv.org/abs/1502.03044
3. H. Ahsan, N. Bhalla, D. Bhatt, and K. Shah, “Multi-Modal image captioning for the visually impaired,” *arXiv.org*, May 17, 2021. https://arxiv.org/abs/2105.08106
4. H. Aldabbas, M. Asad, M. A. Hashem, K. Razzaq, and M. Zubair, “Data augmentation to stabilize Image caption Generation models in deep learning,” *International Journal of Advanced Computer Science and Applications/International Journal of Advanced Computer Science & Applications*, vol. 10, no. 10, Jan. 2019, doi: 10.14569/ijacsa.2019.0101074.
5. Y. Wang, J. Xu, and Y. Sun, “End-to-End transformer-based model for image captioning,” *arXiv.org*, Mar. 29, 2022. https://arxiv.org/abs/2203.15350
6. J. Brownlee, “A gentle introduction to calculating the BLEU score for text in Python,” *MachineLearningMastery.com*, Dec. 18, 2019. https://machinelearningmastery.com/calculate-bleu-score-for-text-python/
7. A. A. Awan, “A complete guide to data augmentation,” Nov. 23, 2022. https://www.datacamp.com/tutorial/complete-guide-data-augmentation
8. C. Goyal, “Part 5: Step by step guide to Master NLP – Word embedding and Text Vectorization,” *Analytics Vidhya*, Jun. 22, 2021. https://www.analyticsvidhya.com/blog/2021/06/part-5-step-by-step-guide-to-master-nlp-text-vectorization-approaches/
9. M. Hodosh, P. Young and J. Hockenmaier (2013) "Framing Image Description as a Ranking Task: Data, Models and Evaluation Metrics", Journal of Artificial Intelligence Research, Volume 47, pages 853-899 http://www.jair.org/papers/paper3994.html