Norwegian University
of Life Sciences

**Master's Thesis 2024 30 ECTS**
Faculty of Science and Technology

# Semantic Enhancements in Image Captioning: Leveraging Neural Networks to Improve BLIP and GPT-2

Sushant Kumar Srivastava
Master of Science in Data Science

This page is intentionally left blank.

# ACKNOWLEDGEMENTS

This thesis concludes my two-year journey in the Data Science Master's program at the Norwegian University of Life Sciences (NMBU), which I began in 2022. It signifies both an academic and personal achievement.

I am deeply grateful to my supervisor, Fadi Al Machot, for his expert guidance in areas like machine learning, deep learning, and neural networks along with valuable feedback on academic writing. His support was crucial to the success of this work.

This thesis has been a passion project for me, strengthening my enthusiasm for data science and artificial intelligence—a field that I believe will continue to impact our lives significantly.

Finally, I owe a huge thank you to my family for their constant support throughout my studies and my life. Their encouragement has been fundamental to my achievements.

Sushant Kumar Srivastava
Ås, May 14th 2024

# ABSTRACT

In the dynamic arena of automated image captioning, significant resources, including energy and manpower, are required to train state-of-the-art models. These models, though effective, necessitate frequent and costly retraining to maintain or enhance their performance. Our Motivation in this thesis has been to explore alternative methods that improve caption accuracy, addressing the unsustainable need for constant retraining.

This study assesses the performance of existing state-of-the-art models like BLIP, and GPT-2 on two key datasets: COCO and FLICKR. It evaluates their effectiveness in generating captions and their potential biases across different image types, using metrics such as **BLEU**, **METEOR**, and **ROUGE**. Our primary goal in this thesis **was to develop innovative approaches that produce captions more akin to human-generated text, aiming to surpass existing models in quality and efficiency without the need for retraining**. We introduced a technique called '**Weighted Summarization**,' combining artificial neural networks with strategic refinements to leverage the strengths of pre-trained models and set a new benchmark in automated image captioning.

Our approach achieved scores on the **COCO** dataset (**BLEU: 0.322, METEOR: 0.328, ROUGE-1 f: 0.452, ROUGE-2 f: 0.187, ROUGE-L f: 0.415**) and on the **FLICKR** dataset (**BLEU: 0.181, METEOR: 0.300, ROUGE-1 f: 0.348, ROUGE-2 f: 0.107, ROUGE-L f: 0.311**), demonstrating enhanced performance over existing models and improved caption quality.

This thesis shares detailed results and discussions about these findings, suggesting a new method that could make automated captioning more accurate and effective, providing a robust foundation for future research and development in this field.

# CONTENTS

# LIST OF FIGURES

# LIST OF TABLES

# LIST OF ABBREVIATIONS

| Abbreviation | Meaning |
|---|---|
| ANN | Artificial Neural Networks |
| NLP | Natural Language Processing |
| BLIP | Bootstrapped Language Image Pretraining |
| GPT-2 | Generative Pre-trained Transformer 2 |
| Pix2Struct | Screenshot Parsing as Pretraining for Visual Language Understanding |
| COCO | Common Objects in Context Dataset |
| BLEU | Bilingual Evaluation Understudy, a Metric for Evaluating Text Translation |
| ROUGE | Recall-Oriented Understudy for Gisting Evaluation, a Metric for Summarization |
| METEOR | Metric for Evaluation of Translation with Explicit Ordering |
| CNN | Convolutional Neural Network |
| RNN | Recurrent Neural Network |
| LSTM | Long Short-Term Memory, a Type of RNN |
| Oscar | Object-Semantic Alignments and Representations |
| VIVO | Visual Vocabulary Pre-training |
| CAMEL | Captioning with Meta-Learning |
| CLIP | Contrastive Language-Image Pre-training |
| ANN | Artificial Neural Network |
| ReLU | Rectified Linear Unit, a type of activation function |

| Abbreviation | Meaning |
|---|---|
| MLP | Multi-Layer Perceptron |
| GRU | Gated Recurrent Unit, a Type of RNN |
| DistilBART | A Distilled Version of the BART Model for Efficient Training |
| API | Application Programming Interface |
| RESNET | Residual Network, a Type of CNN |

# CHAPTER 1

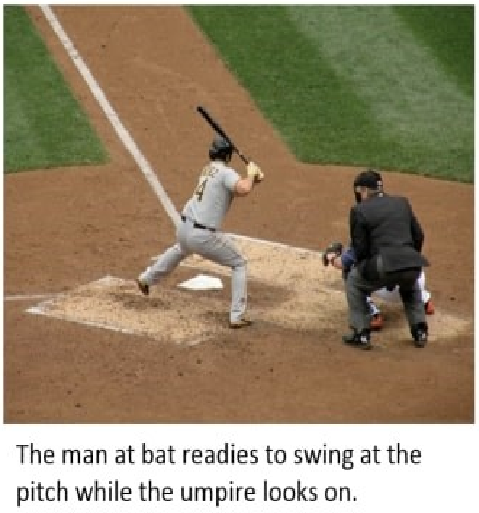
# INTRODUCTION

Artificial intelligence continues to transform various technological domains, notably through the integration of computer vision and natural language processing (NLP). These disciplines converge significantly in the field of image captioning—a technology designed to interpret visual content and translate it into descriptive text. This capability closely mimics human visual and cognitive abilities to perceive, understand, and verbalize visual inputs.

Image captioning has evolved significantly over the years from basic associations of images with tags and labels to the sophisticated task of describing images with complete, contextually relevant sentences. Early methods often relied on simple keyword matching or basic object recognition (Blei and Jordan, 2003; Barnard et al., 2003)[1], [2].

There have been more advancements in this field such as Vinyals et al. (2015) introduced the idea of using deep learning to generate novel captions that not only recognize the objects within an image but also understand their interrelations and the overall scene context [3]. However, these generation-based approaches often grapple with the linguistic nuances and the challenge of producing contextually appropriate and syntactically correct sentences[2].

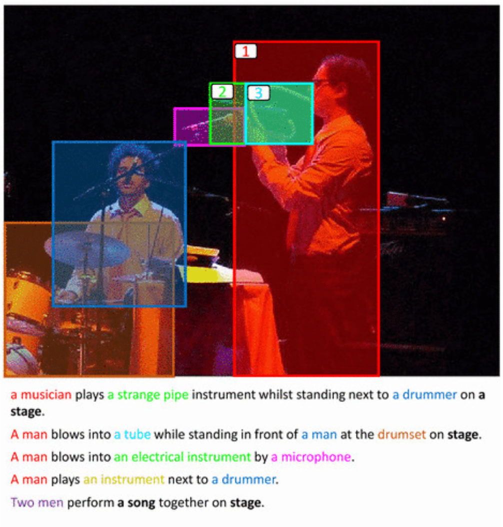
A pivotal shift in the approach to image captioning is highlighted by Hodosh, Young, and Hockenmaier (2013) who propose framing the task as a ranking problem rather than a generation task. They argue that by ranking a predefined pool of accurate and diverse captions according to their relevance to the given images, systems can more effectively and efficiently associate images with the most fitting descriptions. This method not only simplifies the evaluation of captioning systems by comparing their outputs against a benchmark dataset of image-caption pairs but also addresses the semantic accuracy of the captions used [2].



The man at bat readies to swing at the pitch while the umpire looks on.

**Figure 1.1:** A sample image from the COCO dataset showing diverse and complex scenes. Image credit: COCO Dataset [4]

.



a musician plays a strange pipe instrument whilst standing next to a drummer on a stage.
A man blows into a tube while standing in front of a man at the drumset on stage.
A man blows into an electrical instrument by a microphone.
A man plays an instrument next to a drummer.
Two men perform a song together on stage.

**Figure 1.2:** A sample image and captions from the FLICKR dataset,Image credit: FLICKR Dataset [5].

Despite the advancements in algorithms and neural network architectures, creating captions that encapsulate deep semantic meanings of images, capturing their contextual nuances as well as the interplay of elements within remains a challenging task. This thesis explores several advanced models—such as BLIP, GPT-2, and Pix2Struct—that are at the forefront of the image captioning domain that leverages complex neural architectures to enhance the quality and relevance of generated captions and checks the performance of these models on datasets like COCO 1.1 and FLICKR 1.2.

This study further examines how integrating techniques using summarization approaches and assigning weights based on decisions from artificial neural networks can improve performance metrics and caption quality, aiming to synergistically combine outputs from various models to refine the overall effectiveness of image captioning systems. Through the implementation of comprehensive evaluation metrics like BLEU, ROUGE, and METEOR, this research endeavors to set new benchmarks in automated captioning technologies, thus improving both the precision and contextual alignment of generated text [2], [6]–[8].

## 1.1 Background

Building upon the initial discussion in the introduction, this section delves into the historical and technical progression of Image captioning. The journey from simple rule-based methods to sophisticated artificial intelligence models showcases a significant evolution in technology. Early systems were limited by their simplistic algorithms, but the field has dramatically transformed with the introduction of deep learning techniques.

Convolutional neural networks (CNNs), pivotal for detailed image analysis, and recurrent neural networks (RNNs), crucial for processing sequences of data, have significantly boosted the effectiveness of captioning systems [9], [10]. The milestone study by Vinyals et al. demonstrated the potential of neural networks to match images with accurate textual descriptions, establishing a new standard for how machines understand and describe visual content [3].

The availability of extensive datasets such as COCO and FLICKR has fueled advancements by providing diverse sets of images and captions, essential for training more advanced captioning algorithms [2], [5], [11]. These improvements have not only enhanced the precision of object recognition in images but have also deepened the understanding of the contexts in which these objects appear, facilitating the generation of more nuanced and contextually rich captions.

Image captioning involves several essential steps:

1. **Preprocessing:** Images are adjusted in size and quality to facilitate optimal feature extraction.

2. **Feature Extraction:** CNNs are employed to detect and encode vital visual features into a compact form [12].

3. **Sequence Modeling:** RNNs or newer models like transformers translate these features into textual descriptions, often utilizing attention mechanisms to focus adaptively on different parts of the image [13].

4. **Postprocessing:** The resulting captions are polished to ensure they are grammatically sound and stylistically coherent.

The advent of advanced image captioning and object detection technologies has significantly influenced various real-world applications, ranging from enhancing accessibility tools for the visually impaired to improving content management on digital platforms and streamlining indexing in digital archives.

Automatic image captioning has also been advanced by Karpathy and Fei-Fei (2015), who developed a model that effectively combines convolutional neural networks (CNNs) and recurrent neural networks (RNNs) to describe segments of images. Their approach has significantly contributed to the development of technologies that assist the visually impaired by providing real-time, descriptive narratives of visual content, thereby enhancing accessibility [14].

The work by Johnson et al. (2016) on Dense Captioning introduces fully convolutional localization networks that can detect and describe multiple elements within images. This capability is crucial for the efficient indexing of images in digital archives, where precise and comprehensive tagging of numerous image features enhances retrievability and usability. By providing dense, contextual captions that describe various aspects of images, such technologies facilitate more nuanced searches and better organization of visual data in archives[15].

This background aims to seamlessly connect the technical advancements with their practical applications, setting the stage for a detailed exploration of how enhancements in model capabilities could revolutionize automated image captioning.

## 1.2 Problem Statement

While image captioning has made significant advancements, existing models still face substantial challenges in generating semantically rich and contextually accurate captions [16]. Despite their proficiency in detecting individual objects, these systems often struggle to weave these elements into cohesive narratives that accurately reflect human perception [3], [17]. This constraint not only affects their usefulness in areas like assistive devices for visually impaired individuals and automated content creation, but it also underscores a significant technological shortfall. Current models, although sophisticated in object recognition, frequently fall short when it comes to understanding complex interactions and contextual subtleties within images [16], [18]. Furthermore, the integration of emotional and thematic depth in captions remains a significant hurdle, as these aspects require an understanding beyond the visual cues presented [19].

A key issue lies in the resource-intensive nature of training large-scale models, which involves significant computational costs and manpower [15]. Most state-of-the-art models rely on extensive retraining over massive datasets, a process that is not feasible or sustainable, especially when rapid adaptation to new data or contexts is required. This constant updating not only raises the cost of operations but also puts a heavy load on computer systems. Additionally, as the complexity of images increases, there is a growing need for smarter and more sensitive captioning systems. This highlights the importance of finding more sustainable and efficient methods. We are challenged to rethink how we've been doing things and to come up with new, innovative solutions. This research seeks to address these challenges by exploring innovative methods that enhance caption accuracy and depth without the need for extensive retraining.

The focus is on developing techniques that improve upon traditional metrics like BLEU, METEOR, and ROUGE [6]–[8], [15], while also pushing the boundaries of how models understand and interpret complex visual scenes. The ultimate goal is to achieve a level of image understanding better than the state-of-the-art models by performing better in the evaluation metrics considered in this thesis as compared to other models. To achieve this we need to devise an approach, particularly with Artificial neural networks and weighed-based summarization to **reduce dependency on retraining, lower operational costs, and increase the efficiency of generating high-quality image captions**.

## 1.3 Research Questions

Based on the unresearched areas of devising approaches that enhance the quality of captions using modern technologies, the following research question was framed:

**"Can we develop a method that does not rely on retraining, operates at a low cost, and generates captions that are more semantically accurate than those produced by BLIP and GPT-2?"**.

So we explored advanced methodologies in automated image captioning, focusing on the integration of different models(BLIP, GPT-2) and summarization techniques, as well as evaluating their effectiveness across diverse datasets like COCO and FLICKR. To answer this, we formalized four major questions.

1. **Q1: Are advanced machine learning models like BLIP and GPT-2 predisposed to generating more accurate and semantically rich**

**captions for certain categories of images?**

- This question investigates whether state-of-the-art image captioning models exhibit biases towards specific categories of images such as "people and daily activities," "animals and nature," "urban and rural settings," "objects and interiors," "vehicles and transportation," and "food and beverages." A classifier will be utilized to categorize dataset images into these groups, enabling an analysis of the captioning models' performance across varied genres to determine any significant disparities in accuracy and semantic depth.

2. **Q2: How do advanced machine learning models like BLIP and GPT-2 enhance the semantic accuracy and contextual relevance of image captions across varied datasets?**

   - This question seeks to evaluate the effectiveness of cutting-edge machine learning models in generating image captions that are not only accurate but also contextually appropriate across different types of images. It aims to understand the strengths and limitations of these models in handling diverse and complex visual scenes.

3. **Q3: Can integrating outputs from multiple captioning models through advanced summarization techniques and neural networks improve the semantic depth of generated captions?**

   - This question explores the potential of novel summarization strategies that combine outputs from multiple models to produce captions that better reflect the depth and nuances of the visual content. It also examines which integration techniques are most effective and under what conditions.

4. **Q4: To what extent do advanced summarization strategies and ANN findings, aimed at addressing the research questions, surpass state-of-the-art models in terms of semantic accuracy and evaluation metrics for generated captions?**

   - This question assesses the impact of novel summarization strategies and ANN-guided adjustments on the improvement of semantic accuracy and evaluation metrics for image captions. It seeks to quantify the degree of enhancement over state-of-the-art models like BLIP and GPT-2. Furthermore, the question aims to identify which evaluation metrics indicate the most significant increase and what these findings imply about how well the suggested method is working to advance the field of automated picture captioning.

## 1.4 Objectives

Our objectives for this research work include evaluating leading models like BLIP and GPT-2 for their ability to produce semantically rich and contextually detailed captions on diverse datasets and also evaluating their biases towards certain image categories and finally exploring innovative approaches, such as the integration of Artificial Neural Networks (ANNs) to determine the weights for effective summarization and to assess their potential in improving performance metrics. These efforts are intended to refine captioning technologies, ensuring that generated descriptions are both technically sound and practically useful.

## 1.5 Structure of the Thesis

This thesis is systematically divided into several chapters, each addressing a different aspect of automated image captioning. The **Introduction** chapter sets the stage for the research by defining the problem and stating the objectives. The **Literature Review** provides an exhaustive survey of the existing models and methodologies in image captioning. In the **Theoretical Framework**, the principles of artificial neural networks and their application in caption generation, particularly focusing on models like BLIP, GPT-2, and Pix2Struct, are discussed along with selection among various summarizers and evaluation metrics for Image captioning. The **Methodology** chapter details the practical steps involved in dataset preparation, preprocessing, and constructing an ANN model to predict an effective caption generation model along with the weighted summarization approach and post-processing. Also, Bayesian analysis is employed for effective validation of the results. **Results** chapter presents the data and findings of the methodologies applied, while the **Discussion** critically analyses these results to the initial research questions and objectives. Finally, the **Conclusion** summarizes the key findings and discusses the implications and potential avenues for future research. For further detail and practical application, all codes and files used in this research are available in a GitHub repository at this link: GitHub Link with latest commit 7b0a9a9. Read the readme.md file provided for more info.

*In this chapter, we have introduced the concept of image captioning, outlined its historical development, and highlighted significant prior work in this field. We have also detailed our motivation, goals, and the research questions that this thesis aims to address. Moving forward, the next chapter will delve into a comprehensive* *literature review, discussing key methodologies and the evolution of image captioning and summarization techniques, from foundational approaches to the latest advances.*

# CHAPTER 2 LITERATURE REVIEW

In the Literature Review, we review key methodologies and recent progress in image captioning and summarization techniques. We begin by examining foundational image captioning approaches such as CNNs for feature extraction and advances to sophisticated models like RNNs, attention-based systems, and various innovative models. We also cover different summarization strategies, distinguishing between extractive and abstractive methods, and the recent developments made for more relevant summarization.

## 2.1 Image Captioning Techniques

Automated image captioning skillfully combines the fields of computer vision and natural language processing (NLP), turning visual information into written descriptions. This section examines a spectrum of captioning models, evaluating traditional frameworks and innovative systems like I-Tuning, ClipCap, Camel, Oscar, Vivo, and SMALLCAP.

### 2.1.1 CNNs and Feature Extraction

Convolutional Neural Networks (CNNs) have continued to evolve, offering substantial improvements in the feature extraction phase of image captioning. The ResNet architecture introduced by He et al. (2016) marked a significant development by enabling the training of much deeper networks through residual learning [20]. This approach helps mitigate the vanishing gradient problem, allowing the network to learn richer and more sophisticated visual features crucial for detailed image captioning.

In the domain of image captioning, Vinyals et al. (2017) utilized CNNs to encode image data into a dense vector representation before passing it to an RNN for generating descriptive text [21]. Their model demonstrates how deep CNNs can be effectively paired with sequence models to produce coherent captions that not only the objects in an image but also the context in which they exist.

Further advancements by Anderson et al. (2018) introduced the concept of bottom-up and top-down attention mechanisms in CNNs, which allow the model to focus on salient regions of the image dynamically during the caption generation process [18]. This method enhances the relevance of the generated captions by linking specific visual cues to corresponding textual descriptions.

More recently, Cornia et al. (2020) explored how CNNs can be integrated with graph neural networks to capture relationships between objects in an image, thereby providing a more structured semantic understanding that benefits caption generation [22]. Their approach underscores the potential of using CNNs not just for feature extraction but also for constructing a semantic map of the image, which significantly aids in the generation of accurate and context-aware captions.

These studies illustrate the progression of CNNs from basic feature extractors to complex systems capable of providing a deep semantic understanding necessary for generating meaningful image captions.

### 2.1.2 RNNs, LSTMs, and Sequential Processing

Recurrent Neural Networks (RNNs) and Long Short-Term Memory networks (LSTMs) form the backbone of sequential processing in image captioning, enabling models to handle the temporal dynamics of language. The attention mechanism, as pioneered by Xu et al. (2015), revolutionized this field by allowing models to focus adaptively on different parts of an image during the caption generation process [17]. This approach aligns with techniques used in this study, focusing on how objects interact within their scenes to generate contextually rich captions.

Building on these foundations, You et al. (2016) introduced a semantic attention model that combines LSTMs with a dynamic visual attention mechanism to enhance the relevance of generated captions by focusing on salient objects and their attributes [23]. This method emphasizes the importance of not just recognizing objects but understanding their semantic roles in images.

The Transformer model by Vaswani et al. (2017) further advanced sequential processing by eliminating recurrent layers and utilizing multi-headed self-attention to process all words or image regions simultaneously [13]. This architecture has been adapted for image captioning by Herdade et al. (2019), who introduced geometric attention mechanisms to maintain spatial relationships between objects in the captions, thus enhancing the descriptiveness and accuracy of the generated text [24].

These advancements underscore a significant shift towards more sophisticated models that not only process visual and linguistic elements effectively but also contextualize the interactions within the image, a crucial aspect evaluated in this thesis.

### 2.1.3 Transformers Based

Attention mechanisms in transformers have significantly advanced the capabilities of image captioning models by allowing for dynamic allocation of focus on relevant parts of an image. For instance, Li et al. (2019) introduced an area attention mechanism that broadens the focus beyond single points to larger areas, enhancing the model's ability to describe complex scenes [25]. Additionally, a notable advancement is the attention-aligned Transformer, introduced by Fei (2023), which addresses the "deviated focus" problem in existing attention mechanisms. This model enhances the grounding of correct image regions for word generation by employing a perturbation-based self-supervised learning approach, which refines attention distribution without needing manual annotation [26].

Transformers have significantly improved the depth and precision of image captioning by analyzing multiple image parts simultaneously. Carion et al. (2020) presented DETR, a Transformer-based model that directly predicts bounding boxes and object classes in a single stage, which has been adapted to improve spatial awareness in captioning tasks [27]. Moreover, Zhou et al. (2021) introduced a transformer model that incorporates cross-modal features to enhance the correlation between visual elements and textual descriptions, enriching the semantic content of generated captions [28].

### 2.1.4 Innovative Models

Recent developments in image captioning have introduced innovative models like OSCAR and VIVO, which are reshaping traditional approaches by integrating advanced semantic analysis techniques. OSCAR (Object-Semantic Alignments and Representations) utilizes object tagging within images to enhance the semantic richness of the captions. This method leverages pre-trained object detectors to tag visible objects in the images, which are then used as anchor points for generating more contextually accurate captions. The incorporation of these object tags into the language model allows OSCAR to achieve superior performance by grounding the textual elements more firmly in the visual content [29].

Similarly, the VIVO (Visual Vocabulary Pre-training) model extends the capability of image captioning systems by embedding richer semantic information during the training process. VIVO enhances traditional captioning approaches by pre-training on a large-scale dataset to learn visual-semantic embeddings effectively. This enables the model to comprehend and articulate complex image contents with heightened accuracy and nuance, thus approaching a more human-like perception in generated captions [30].

Also, developments in image captioning have introduced several notable models, each bringing unique strengths to the table. **I-Tuning** is an image captioning approach that fine-tunes pre-existing language models using a smaller dataset of image-caption pairs. This method leverages the vast knowledge captured by language models pre-trained on extensive text data, adapting it to the specific task of image captioning with minimal additional training. I-Tuning utilizes a frozen language model combined with image inputs to enhance captioning capabilities [31]. On the other hand, **CaMEL** (Captioning with Meta-Learning) utilizes meta-learning principles to adapt effectively to the image captioning task. By applying meta-learning algorithms, CaMEL can quickly learn from a small number of examples, making it adept at generating captions for new, unseen images after training on a limited dataset. This approach allows for rapid adaptation and robust caption generation in diverse scenarios[32][33]. Also, **ClipCap** combines the capabilities of CLIP (Contrastive Language–Image Pretraining) with

GPT (Generative Pretrained Transformer) to create captions. CLIP's strength in understanding visual concepts combined with GPT's language generation prowess allows ClipCap to produce accurate and relevant image captions [34]. Lastly, **SMALLCAP** represents an efficient method of generating image captions by using retrieval augmentation to streamline the process. It focuses on optimizing cross-attention layers, resulting in a model with fewer parameters yet high captioning performance [33].

In the field of image captioning, these models are examined for their ability to produce descriptive and relevant captions using standard evaluation metrics like BLEU and METEOR. Below is the table 2.1 including the scores for these models

**Table 2.1:** Comparison of Image Captioning Techniques and Models evaluated on COCO Dataset

| **Model** | **BLEU** | **METEOR** |
|---|---|---|
| CNN-LSTM[35] | 0.27 | 0.23 |
| OSCAR[29] | 0.35 | 0.29 |
| VIVO[30] | 0.34 | 0.28 |
| SMALLCAP[33] | 0.37 | 0.27 |
| I-Tuning_Large[33] | 0.34 | 0.29 |
| CaMEL[33] | 0.39 | 0.29 |
| I-Tuning_Medium[33] | 0.35 | 0.28 |
| ClipCap[33] | 0.33 | 0.27 |
| I-Tuning_Base[33] | 0.34 | 0.28 |

The exploration of these technologies provides a comprehensive view of the evolution and current state of automated image captioning.

## 2.2 Summarization Techniques

Text summarization techniques can be broadly categorized into two main approaches: extractive summarization and abstractive summarization. Extractive summarization involves selecting and concatenating the most important sentences or phrases from the source text to form a summary. In contrast, abstractive summarization aims to generate a concise paraphrase of the original content, often producing new sentences that were not in the original text [36], [37].

### 2.2.1 Extractive Summarization

Extractive summarization identifies key sentences or segments within a text and combines them to create a summary. This method often uses criteria such as the location of a sentence within the text, term frequency, and how unique certain words are across documents to pinpoint important sentences[38]. Erkan and Radev (2004) for instance, introduced the concept of implementing a graph-based centrality scoring approach that identifies the most salient sentences to improve summary coherence [39]. Yasunaga et al. (2017) further introduced a novel graph-based model that improves sentence selection accuracy by evaluating how sentences interconnect within the text, thereby enhancing the structural comprehension of the document [40]. Zhong et al. (2020) have advanced this area by incorporating machine learning classifiers that assess sentence importance more effectively, using vast amounts of training data to refine selection accuracy [41]. Narayan et al. (2018) explored a reinforcement learning approach that optimizes the selection process by rewarding the system for selecting sentences that contribute most effectively to a coherent summary [42].

### 2.2.2 Abstractive Summarization

Abstractive summarization methods synthesize the underlying concepts of texts to generate concise new formulations that capture the essence of the original content [38]. This approach has been significantly advanced by deep learning techniques, particularly with the development of sequence-to-sequence models that use attention mechanisms to focus on relevant text segments. Devlin et al. (2019), for instance introduced BERT, a transformer-based model that uses bidirectional training of transformers to generate language understanding that can be fine-tuned for tasks like summarization [43]. Brown et al. (2020) developed GPT-3, an even larger transformer model that excels in generating coherent text based on a given prompt, including summarization tasks [44]. Raffel et al. (2020) introduced the T5 model, which pre-trains on a diverse dataset to handle various NLP tasks, including summarization, by converting all text-based language problems into a unified text-to-text format [45]. Furthermore, Lewis et al. (2020) presented BART, which fine-tunes the transformer approach by pre-training a denoising autoencoder to reconstruct text, aiding the generation of fluent and accurate summaries [46]. Zhang et al. (2020) developed PEGASUS, which introduces a novel pre-training objective focused on gap sentences, optimizing the model's ability to predict and generate relevant content for abstractive summarization [47].

**Table 2.2:** Comparison of ROUGE-1 and BLEU-1 Scores for some summarization models. Data credit: [48]

| **Model** | **ROUGE-1** | **BLEU-1** |
|---|---|---|
| Extractive | | |
| PositionRank | 0.3341 | 0.2507 |
| LexRank | 0.3245 | 0.2334 |
| TextRank (pyTextRank) | 0.3126 | 0.2310 |
| Abstractive | | |
| distilBART-CNN-12-6 | 0.4292 | 0.3625 |
| BART-large-CNN | 0.4270 | 0.3156 |
| PEGASUS-CNN_dailymail | 0.4186 | 0.3259 |

In a recent study, *Abstractive vs. Extractive Summarization: An Experimental Review,*[48] Giarelis et al. present a comprehensive comparison of extractive and abstractive summarization techniques using various performance metrics such as ROUGE and BLEU scores. A snippet of their findings is represented in table 2.2, showcasing the ROUGE-1 and BLEU-1 scores for selected models, which illustrate significant differences in performance between extractive and abstractive methods concluding that abstraction summarization produces better-summarized content as compared to the extractive summarization.

### 2.2.3 Hybrid Approaches

Additionally, as natural language processing has advanced, there has been a greater emphasis placed on hybrid summarizing models, which combine extractive and abstractive methods to maximize the coherence and accuracy of generated summaries. These hybrid approaches begin by utilizing extractive techniques to pinpoint specific sentences or phrases that are important for comprehending the primary concepts in a text. The collected content is then paraphrased and reorganized using an abstractive model in an effort to provide a summary that captures the main idea of the original text while being clear and succinct.

Hsu et al. (2018), for instance, presented a unified model that makes use of both extractive and abstractive elements[49]. In order to penalize misalignment between sentence-level attention used for extraction and word-level attention required for abstraction, their model uniquely integrates an inconsistency loss function during training. By ensuring that words in less attended phrases are less likely to be generated, this method helps to keep attention on the most important details. A solid human evaluation revealed that their system achieved state-of-the-art ROUGE scores and was the most readable and informative summarization on the CNN/Daily Mail dataset by training end-to-end with the inconsistency loss along with original losses of extractive and abstractive models[49].

Similarly, a pointer-generator network that effectively combines extractive features with abstractive capabilities was created by See et al. [50]. This model incorporates a pointing mechanism that enables the system to copy words straight from the source text, combined with a coverage technique to prevent repetition, to address frequent problems in abstractive summarization, such as factual errors and repetitions. As a result, on the CNN/Daily Mail summary test, the robust summarization tool performed at least two ROUGE points better than the abstractive state-of-the-art[50].

*In this chapter, we reviewed established and emerging methodologies in image captioning and summarization techniques. We explored various models and strategies, from fundamental architectures to advanced systems, and discussed their effectiveness in creating relevant captions. Moving forward, Chapter 3, 'Theoretical Framework,' will delve into the specifics of Artificial Neural Networks (ANNs), from their basic structures to complex configurations, and their application in image captioning, with a focus on modern generation techniques like BLIP, GPT-2, and Pix2Struct.*

# CHAPTER 3

# THEORETICAL FRAMEWORK

In the Theoretical Framework, we explore the foundational and advanced aspects of Artificial Neural Networks (ANNs), discussing their evolution from simple perceptrons to complex multi-layer networks, along with key concepts like activation functions and optimization techniques and their relevance to tasks such as image captioning. We also examine modern caption generation techniques including BLIP, GPT-2, and PIX2STRUCT, alongside their implications in generating descriptive text from images. Additionally, we address summarization techniques, highlighting their importance in captioning, what model we have selected, and on what basis, concluding with a review of metrics used to assess the effectiveness of image captioning and summarization methods.

## 3.1 Artificial Neural Networks

Artificial Neural Networks (ANNs) are computational models inspired by the human brain, widely used in machine learning and artificial intelligence to emulate complex decision-making processes. These networks are fundamental to modern AI applications, ranging from image recognition to natural language processing [51].

### 3.1.1 Neural Network Fundamentals and Perceptrons

Neural networks are modeled after the biological neural networks found in animal brains. The basic unit of a neural network, the neuron, includes dendrites as input receivers, a cell body, and an axon that transmits signals to other neurons. This biological inspiration is abstracted in artificial neural networks where neurons are simulated by nodes in a network, as illustrated in Figure 3.1.

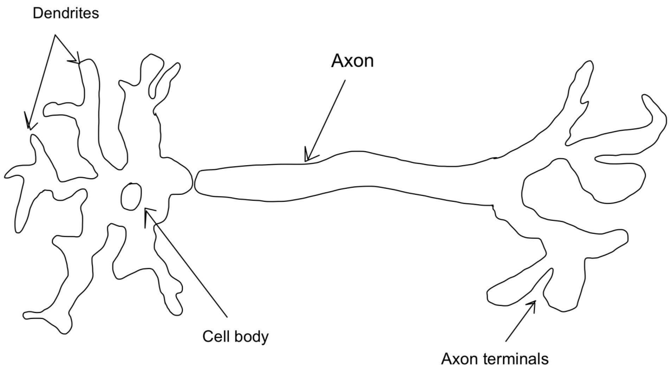


**Figure 3.1:** Diagram of a biological neuron showing the cell body, dendrites, axon, and axon terminals.

Each artificial neuron receives input signals and processes them through a weighted sum which is then passed through an activation function to produce an output. The perceptron, conceptualized in the 1950s by Frank Rosenblatt, represents the simplest form of a feedforward neural network, which is fundamentally a single-layer neural network. Figure 3.2 illustrates a basic perceptron setup.

The perceptron computes its output $y$ using the formula:

$$y = f\left(\sum_{i=1}^{n} w_i x_i + b\right) \tag{3.1}$$

where $x_i$ are inputs, $w_i$ are the associated weights, $b$ is a bias term, and $f$ is an activation function. The activation function, often a step function in basic perceptrons, determines the output based on whether the weighted sum reaches a

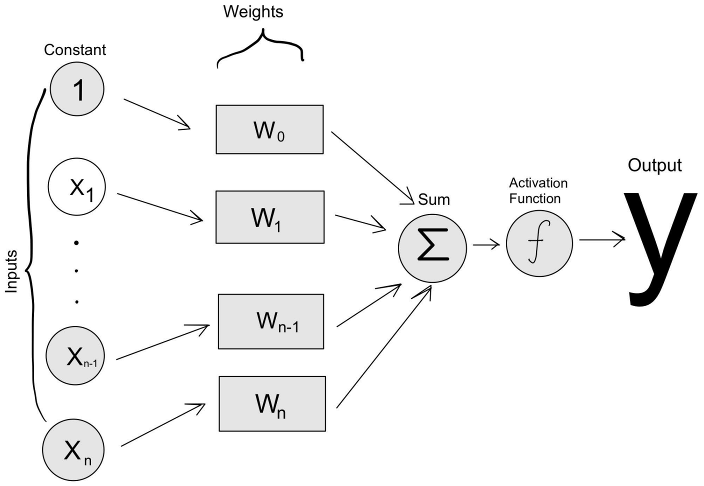


**Figure 3.2:** Illustration of a perceptron showing inputs with weights, the summation function, and output through a step function.

certain threshold. More advanced neural networks utilize different types of activation functions like sigmoid, tanh, or ReLU to introduce non-linearity into the network, enabling them to learn more complex patterns [51].

The development and evolution of perceptrons have paved the way for more complex architectures such as multi-layer perceptrons (MLPs) and modern deep learning networks, which consist of multiple layers of neurons and are capable of capturing high-dimensional patterns in data, far surpassing the capabilities of single-layer perceptrons [52].

### 3.1.2 From Perceptrons to Multi-Layer Networks

The inherent limitations of single-layer perceptrons, particularly their inability to solve non-linear problems such as the XOR dilemma highlighted by Minsky and Papert [53], necessitated the development of more sophisticated architectures. Multi-layer perceptrons (MLPs) address these challenges by incorporating multiple layers of neurons, usually composed of an output layer, an input layer, and one or more hidden layers. This structure allows MLPs to create complex representations and solve non-linear problems by learning hierarchical features.

#### 3.1.2.1 Activation Functions

To enable these networks to capture non-linear relationships, MLPs employ non-linear activation functions. These functions are essential because they give the network non-linear characteristics, which enable it to recognize complicated data patterns:

- Sigmoid: $\sigma(x) = \frac{1}{1+e^{-x}}$, which outputs values between 0 and 1, making it suitable for binary classification tasks [54].
- ReLU: $f(x) = \max(0, x)$, known for allowing models to train faster and more effectively by overcoming problems like the vanishing gradient issue [55].

To complement the basic activation functions, several other types are frequently employed, each with unique characteristics and applications:

- Tanh (Hyperbolic Tangent): The Tanh function, given by $\tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$, outputs values between -1 and 1. This range makes it particularly effective for tasks where negative values have semantic meaning [56].
- Leaky ReLU: This variant of ReLU is designed to address some of its limitations by allowing a small, non-zero gradient when the input is less than zero $f(x) = \max(0.01x, x)$. This helps maintain the gradient flow during training, which can prevent the dying ReLU problem [57].
- Softmax: Often used in the final layer of a classifier, the Softmax function $\sigma(z)_i = \frac{e^{z_i}}{\sum_j e^{z_j}}$ assigns decimal probabilities to each class in a multi-class problem, ensuring the total sum of these probabilities equals 1 [58].

#### 3.1.2.2 Learning and Optimization

MLPs learn by adjusting the synaptic weights to minimize the error between the predicted and actual outputs. This learning process typically uses the backpropagation algorithm, a powerful method for efficiently computing gradients of the loss function concerning the weights [59]. Combined with optimization techniques such as gradient descent, backpropagation updates weights to minimize the network's loss function, effectively training the network to make accurate predictions.

The cross-entropy loss function, frequently used in binary classification tasks, quantifies the cost of predicting probabilities divergent from the actual class labels:

$$L(y, \hat{y}) = -(y \log(\hat{y}) + (1 - y) \log(1 - \hat{y})) \tag{3.2}$$

This function is particularly effective in scenarios where decisions are categorical, and it helps in penalizing incorrect classifications more severely [51].

Recent advances in network architecture and learning algorithms have further evolved MLPs into deeper and more complex networks, capable of tackling a wide range of challenging tasks across various fields. The introduction of techniques such as dropout[60], batch normalization[61], and advanced optimizers like Adam have significantly improved the training dynamics and generalization capabilities of MLPs [62].

### 3.1.3 Relevance of ANNs to Image Captioning

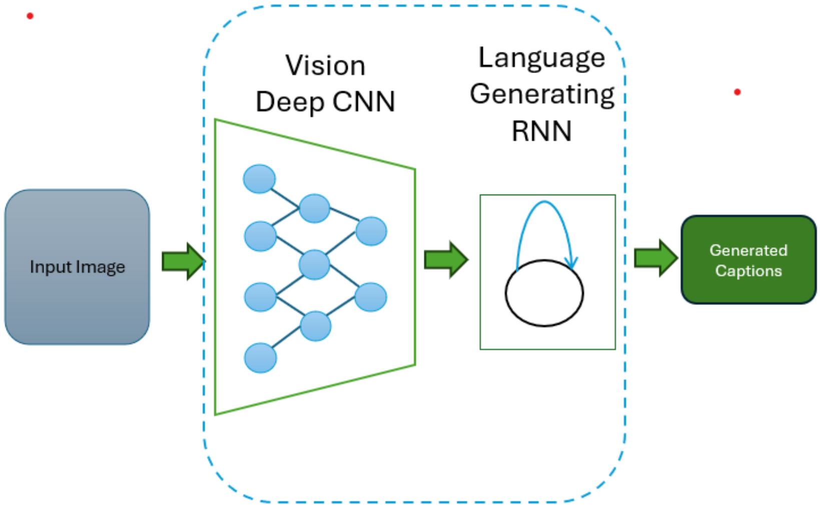


**Figure 3.3:** Redrawn image showing the Neural Image Caption Generator: an end-to-end neural network combining a vision CNN and a language-generating RNN to produce sentences from images, based on the description by Vinyals et al. [3].

Artificial Neural Networks (ANNs), particularly Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), are fundamental to the process of image captioning. CNNs excel in extracting visual features due to their ability to hierarchically process pixel data, making them ideal for understanding the structural and textural details of images. This capability allows them to act as feature extractors in image captioning systems where the visual context is critical [12].

On the other hand, RNNs and their advanced variants like LSTMs and GRUs play a crucial role in forming sequential data processing, making them suitable for generating textual descriptions based on the features provided by CNNs. Their ability to maintain internal states aids in handling the sequence of words in captions, aligning the generated text with the visual content [10], [63].

The integration of RNNs with CNNs in image captioning was popularized by the architecture presented in "Show and Tell: A Neural Image Caption Generator" by Vinyals et al., also seen in figure 3.3, where a CNN encodes an image into a dense vector, followed by an RNN that decodes it to form a coherent caption [3]. This model laid the groundwork for further studies in the field, such as the introduction of attention mechanisms. These mechanisms enable the model to dynamically concentrate on particular areas of an image while generating captions, thereby enhancing the relevance and precision of the captions produced.[17].

Further advancements were made with the introduction of transformer-based models such as BERT and GPT-3, which leverage vast amounts of data and sophisticated self-attention mechanisms to generate even more contextually relevant captions. These models represent a significant shift from traditional RNNs to architectures that can process inputs in parallel, enhancing both the efficiency and effectiveness of caption generation [43], [44].

## 3.2 Caption Generation Techniques

In this study, we explored the performance of state-of-the-art models for generating image captions across the COCO and FLICKR8k datasets. We selected models such as BLIP, GPT-2, and PIX2STRUCT for their innovative approaches to image captioning.

### 3.2.1 BLIP

BLIP (Bootstrapped Language Image Pretraining) utilizes advanced machine learning techniques to integrate visual and textual data, facilitating the creation of robust image captions. Developed by Salesforce, BLIP leverages transformer architectures to efficiently process complex visual scenes and generate descriptive text efficiently [64].

**Architecture and Functionality:** BLIP employs a dual-component setup, incorporating a vision encoder for image analysis and a language decoder for caption generation. It is pre-trained on a vast array of image-text pairs, which enhances its capability to contextualize and articulate visual inputs effectively [64].

### 3.2.2 GPT-2

GPT-2, developed by OpenAI, is a large-scale transformer-based model primarily known for its abilities in natural language understanding and generation [65]. For image captioning, GPT-2 can be effectively combined with Vision Transformers (ViT) to generate descriptive captions from images.

**Architecture and Functionality:** GPT-2's architecture is based on the transformer model, which utilizes self-attention mechanisms to process and generate text based on input sequences. When adapted for image captioning, GPT-2 is paired with a vision model like ViT, which encodes images into a sequence of embeddings that GPT-2 can process [66].

### 3.2.3 Pix2Struct

Pix2Struct introduces a novel pretraining methodology for visual language understanding, focusing on the parsing of screenshots. Developed by Kenton Lee et al., this model leverages an image-encoder-text-decoder architecture to transform screenshots into detailed textual descriptions, enhancing tasks related to automated documentation and accessibility [67].

**Key Features and Architecture:** The architecture employs Vision Transformers (ViT) integrated with modified Transformers for text decoding, adept at interpreting complex visual and textual elements from web content and translating them into coherent structured outputs[67].

**Functionality and Application:** Pretraining involves self-supervised learning from screenshots and HTML pairs, aiming at capturing web page semantics efficiently. This approach prepares Pix2Struct for applications requiring a nuanced understanding of visual layouts, which is critical for technologies such as content accessibility and automated data extraction. Published in 2022, Pix2Struct marks a significant advance in pretraining strategies for multimodal AI systems, with potential applications in various fields requiring robust visual data interpretation [67].

Despite its innovative approach, **Pix2Struct was excluded from further detailed analysis in this study due to its relatively lower performance in preliminary evaluations compared to other models. The decision to focus on models that provided more effective captioning in terms of our evaluation criteria reflects our goal to highlight the most efficient technologies in this thesis**.

## 3.3 Evaluation metrics for Image captioning

Assessing the quality of machine-generated captions involves several metrics, each offering unique insights into different aspects of caption quality:

### 3.3.1 BLEU (Bilingual Evaluation Understudy)

BLEU (Bilingual Evaluation Understudy) is widely used for evaluating the quality of text that has been machine-translated from one natural language to another. However, it has also been used in automated image captioning to assess how much the n-grams in the created captions match with those in a set of reference captions:

$$BLEU = BP \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right) \tag{3.3}$$

where $BP$ represents the brevity penalty to penalize short machine-generated captions, $w_n$ are the weights for each n-gram, and $p_n$ is the precision of n-grams. The brevity penalty and weighting of n-grams aim to balance fluency and adequacy of the generated text [6]. **By default, BLEU@4 is used here, which considers 4-gram precision, and ranges from 0 to 1**.

### 3.3.2 METEOR (Metric for Evaluation of Translation with Explicit Ordering)

METEOR (Metric for Evaluation of Translation with Explicit Ordering) seeks to address some of the BLEU metric's shortcomings by incorporating synonyms and stemming, thus providing a more nuanced evaluation of translation or caption generation:

$$METEOR = \frac{10}{R} \cdot \frac{P \cdot R}{\alpha P + (1 - \alpha)R} \tag{3.4}$$

$P$ denotes precision, $R$ recall, and $\alpha$ is a parameter set to balance precision and recall, typically around 0.9. METEOR correlates better with human judgment by considering synonymy and paraphrasing, aiming for semantic alignment beyond exact word matches [8]. **By default, Meteor score ranges from 0 to 1**.

### 3.3.3 ROUGE (Recall-Oriented Understudy for Gisting Evaluation)

ROUGE (Recall-Oriented Understudy for Gisting Evaluation) metrics are essential for evaluating the quality of text in tasks like summarization or caption generation by comparing machine-generated text to human-generated reference texts[7]. ROUGE metrics involve:

- **Recall (R)**: This measures the fraction of the reference text's n-grams that are also found in the generated text, focusing on content coverage[7].
- **Precision (P)**: This assesses the fraction of the generated text's n-grams that are found in the reference texts, reflecting the pertinence of information provided[7].
- **F-measure (F)**: This is the harmonic mean of precision and recall, balancing both the completeness and the relevance of the generated text against the references[7].

Specifically, **ROUGE-L** uses the longest common subsequence to calculate these metrics, considering the order of words and providing a comprehensive measure of quality:

$$ROUGE - L = \frac{(1+\beta^2) \cdot P_{\text{lcs}} \cdot R_{\text{lcs}}}{R_{\text{lcs}} + \beta^2 \cdot P_{\text{lcs}}} \tag{3.5}$$

where $P_{\text{lcs}}$ (Precision) and $R_{\text{lcs}}$ (Recall) are based on the longest common subsequence. The parameter $\beta$ is typically set to prioritize recall over precision, emphasizing the importance of not missing information over avoiding spurious information[7].

For this research, we particularly utilize the **ROUGE-F score** because it provides a balanced view of both recall and precision, making it ideal for assessing the overall quality of the generated captions in comparison to the reference captions [7]. **The ROUGE-F score effectively combines the aspects of accuracy and completeness, which are crucial for evaluating the semantic depth of automated captions and it ranges from 0 to 1**.

**ROUGE-1F** and **ROUGE-2F** refer specifically to the F-measure applied to unigram and bigram overlap, respectively. These metrics are calculated as follows:

$$ROUGE - nF = \frac{(1+\beta^2) \cdot P_n \cdot R_n}{R_n + \beta^2 \cdot P_n} \tag{3.6}$$

where $P_n$ and $R_n$ denote the precision and recall of the n-gram (either unigram for ROUGE-1 or bigram for ROUGE-2). These variations provide insights into the textual similarity at different levels of granularity[7]. **ROUGE-LF** further incorporates these measures to the longest common subsequence.

### 3.3.4 Cosine Similarity

Cosine Similarity is utilized to gauge the semantic similarity between vectors of two sentences, providing a measure of how conceptually close the machine-generated caption is to the reference captions. This measurement assesses the angle between two vectorized sentences, represented as $A$ and $B$, with a smaller angle indicating a higher similarity. This metric is particularly useful in semantic analyses where the exact choice of words is less critical than the overall conveyed meaning:

$$cosine\_similarity(A, B) = \frac{A \cdot B}{\|A\| \|B\|} \tag{3.7}$$

Cosine similarity evaluates how closely aligned the semantic contexts of two document vectors are, making it invaluable for applications like document retrieval and text similarity assessment[68]. **It ranges from -1 to 1**.

Together, these metrics provide a comprehensive toolkit for evaluating the performance of automated captioning systems, highlighting the field's ongoing evolution towards more nuanced and semantically aware evaluation methodologies.

## 3.4 Summarization Techniques

### 3.4.1 Importance in Captioning Context

In the realm of automated image captioning, the generation of accurate and contextually relevant captions presents a significant challenge, compounded by the variability and complexity of visual content. Summarization techniques, particularly those based on advanced neural network models, offer a promising avenue for refining and enhancing generated captions. By summarizing the captions produced by different models, it is possible to distill the essence of multiple descriptive texts into a cohesive, singular narrative as we see further in our study. This process not only streamlines the content but also potentially increases the relevance and accuracy of the final caption, making summarization an invaluable tool in the captioning context.

### 3.4.2 Selection and Evaluation of Summarizer Models

The choice of the summarizer is crucial in generating concise and semantically coherent captions. To identify the most effective summarizer for our study, we evaluated several state-of-the-art models. Among the evaluated models, **DistilBART-CNN-12-6**[1] emerged as the preferred choice for several reasons:

- **High BLEU Score:** DistilBART demonstrated a competitive BLEU score, by indicating a strong alignment with human-written summaries.
- **Parameter Efficiency:** Despite having a considerable number of parameters, DistilBART balances model complexity and summarization quality

[1] https://huggingface.co/sshleifer/distilbart-cnn-12-6

effectively.

- **Versatility:** Originally fine-tuned on a diverse dataset (CNN/Daily Mail), DistilBART exhibits the versatility necessary for processing captions across different domains.

This decision underscores our commitment to leveraging advanced NLP techniques to close the gap between machine-generated text and human-like expressiveness.

DistilBART—a streamlined BART model optimized for the CNN/Daily Mail dataset—proved to be very successful at summarizing image captions[69].
*We also considered other summarization models but they were not as effective as DistilBART, for more info see comparison table. E.1 in the Appendix. E.*

The evaluation of these summarization techniques involved comparing their generated summaries against original captions from the COCO dataset, utilizing BLEU scores as a quantitative measure of performance. This comparative analysis highlighted the nuanced capabilities of each model, with DistilBART's summaries achieving the highest BLEU scores.

### 3.4.3 Evaluation of Summarization Techniques

Evaluating text summarization extends beyond automated metrics to include qualitative assessments that reflect the summary's utility and readability. While metrics such as ROUGE [7], BLEU [6], and METEOR [8] are standard for preliminary assessments, they do not wholly capture the effectiveness of a summary in conveying the intended message or its linguistic quality. This section discusses comprehensive methods to evaluate summarization outputs effectively.

**Human Judgment:** When evaluating the quality of summaries, human reviewers are essential. They provide insights into:

- **Informativeness:** Human evaluation provides informativeness and reliability of summaries.[70].
- **Coherence and Fluency:** The logical flow and readability of the text, ensure that the summary is not only concise but also well-structured and clear [71].

**Consistency and Fidelity:**

- **Factual Consistency:** Evaluators check for factual accuracy, ensuring the summary does not alter or misrepresent the information presented in the source document [72].
- **Coverage:** This measures how well the summary covers key points and topics from the original text, ensuring that no critical information is omitted [73].

**Hybrid Evaluation Approaches:** Combining automated metrics with human evaluations provides a balanced approach, leveraging the scalability of automated methods and the nuanced understanding of human reviewers [74]. This method ensures that the summaries are not only statistically valid but also practically useful and engaging.

*In this chapter, we explored the theoretical aspects of artificial neural networks, their application in cutting-edge image captioning techniques, along with caption generation models and various evaluation metrics. Building on this foundation, the next chapter, 'Methodology', outlines the practical steps we have taken to enhance our captioning models and refine our evaluation metrics.*

# CHAPTER 4

# METHODOLOGY

In the Methodology chapter, we explain in detail the approach employed in the analysis of our generated caption and evaluation metric improvement in image captioning. We begin with an overview of the datasets used, namely COCO and FLICKR8K, explaining their significance and the specifics of data preparation for each. Then we move further with explaining the overall approach pipeline and explaining each step procedure in detail that was employed. Lastly, we discuss the use of Bayesian analysis to compare models namely BLIP and GPT-2, detailing the calculation of likelihood and posterior probabilities to validate the research findings.

## 4.1 Dataset Overview

### 4.1.1 COCO

The Common Objects in Context (COCO) dataset is a large-scale object detection, segmentation, and captioning dataset. COCO has several features that distinguish it from other image datasets. It contains over 330K images, 1.5 million object instances, and 80 object categories, making it one of the most comprehensive datasets for computer vision research [11].

COCO is designed to spur the development of image recognition, segmentation, and captioning algorithms that are capable of understanding images in the context of their natural environments. Unlike datasets that focus on object classification, COCO provides multiple object annotations per image, which include object segmentation, object categorization, and captioning, offering a richer set of data for training and evaluating AI models [11].

The inclusion of natural language captions for each image also makes COCO uniquely suited for tasks that bridge computer vision and natural language processing, such as image captioning. These captions provide a human-generated description of the scenes depicted in the images, covering a wide range of objects and actions [75].

For this thesis, the COCO dataset is used as the primary corpus for training the ANN and evaluating the image captioning models. The diversity of the dataset, coupled with its rich annotations, provides an ideal setting for exploring advanced machine-learning methodologies for generating descriptive text from visual inputs.

**COCO Dataset in This Study:** Due to constraints on computational resources and the limited timeframe available for this study, our experiment utilizes a subset of the COCO 2014 dataset. Specifically, we have selected a set of 1,100 images, each accompanied by 5 human-generated captions, to evaluate the image captioning models and train the ANN developed for prediction discussed in the coming sections.

### 4.1.2 FLICKR8K

In addition to leveraging the comprehensive COCO dataset, our methodology equally encompasses the FLICKR8k dataset to conduct a parallel analysis. The FLICKR8k dataset, comprising 8,000 images sourced from FLICKR, each with five unique English language captions, serves as a critical component of our study [2]. This dataset is especially valued for its focus on everyday scenes and objects, captured in a variety of settings, offering a complementary perspective to the diverse imagery found in COCO.

The application of our approach to both the COCO and FLICKR8k datasets allows for a multifaceted evaluation of our models. This dual-dataset strategy is designed to uncover common patterns and insights in the evaluation metrics for transformers, ensuring that our findings are not individual to a single dataset's characteristics.

The methodology applied to the FLICKR dataset mirrors that of the COCO dataset in terms of feature extraction, and the subsequent training phases and prediction. This parallel processing ensures that any observed performance trends can be attributed to the models' capabilities rather than differences in methodology or dataset handling.

**FLICKR8k Dataset in This Study:** Similarly, for the FLICKR dataset, we have applied the same criteria to select an equivalent subset of 1,100 images, each with its corresponding five captions. This selection ensures that our analysis across datasets is consistent and comparable, facilitating a direct examination of model performance and generalization across different data sources.

*Note that we also have explored other datasets, but due to their low performance of generating captions and low scores on evaluation metrics, they were not considered in further analysis. Refer Appendix. B for more info.*

## 4.2 Overall Approach

The methodology adopted in this study unfolds through a sequence of distinct but interrelated stages, as visualized in Figure 4.1. Beginning with a comprehensive dataset preparation, we advance through data preprocessing and feature extraction before employing an Artificial Neural Network (ANN) to guide our caption generation strategy.

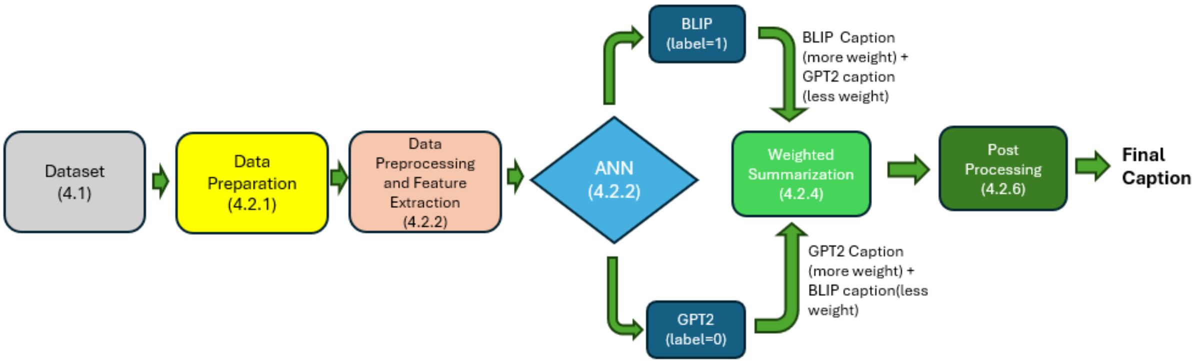

Dataset (4.1)
Data Preparation (4.2.1)
Data Preprocessing and Feature Extraction (4.2.2)
ANN (4.2.2)
BLIP (label=1)
BLIP Caption (more weight) + GPT2 caption (less weight)
GPT2 (label=0)
GPT2 Caption (more weight) + BLIP caption(less weight)
Weighted Summarization (4.2.4)
Post Processing (4.2.6)
Final Caption

**Figure 4.1:** Schematic overview of the methodological framework from dataset handling to final caption generation.

Subsection 4.2.1 delineates the **Data Preparation** stage, where the COCO and FLICKR datasets are curated for processing. In Subsection 4.2.2, we describe the

**Data Preprocessing and Feature Extraction** process, leveraging a pre-trained ResNet-50 model to obtain image features for ANN input.

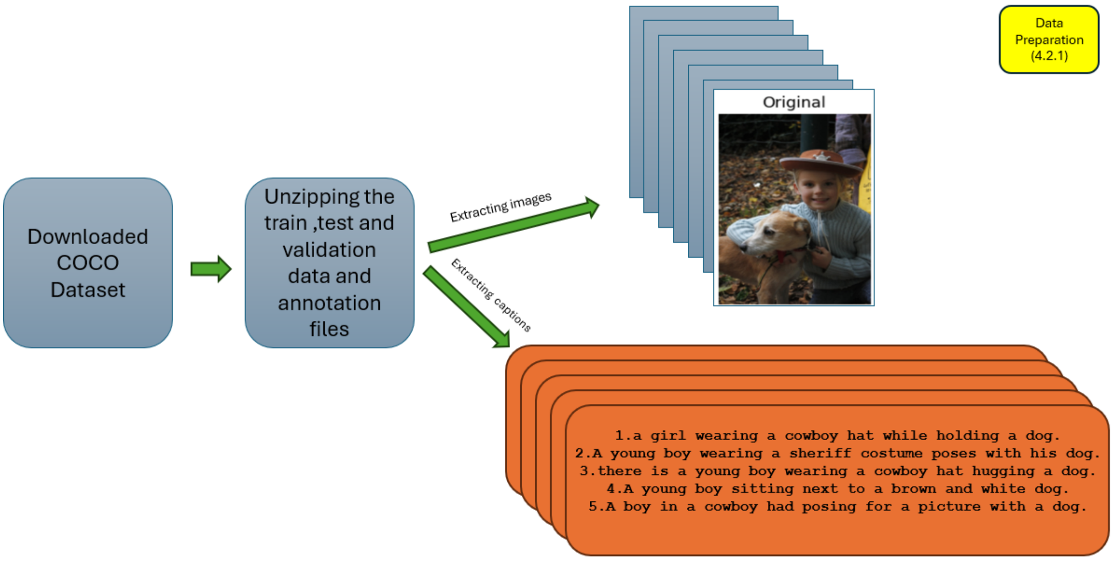
The **ANN Model** (Subsection 4.2.2), instrumental in predicting the most effective captioning model for an image, acts as a decision-making juncture. Depending on the ANN's output, either BLIP or GPT2 model's caption is assigned more weight in the **Weighted Summarization** step (Subsection 4.2.4). This innovative approach intelligently biases the summary towards the model deemed more accurate, ensuring that the final caption reflects the most pertinent description of the image content.

Finally, in Subsection 4.2.6, **Post-Processing** measures are employed to refine the synthesized captions, enhancing their grammatical correctness and semantic coherence. The result is a **Final Caption** that encapsulates the essence of the visual data in a linguistically polished form.

Each subsection is accompanied by a detailed figure to enhance the reader's understanding and provide a visual representation of the complex processes involved.

## 4.2.1 Data Preparation

### 4.2.1.1 COCO Dataset



1.a girl wearing a cowboy hat while holding a dog.
2.A young boy wearing a sheriff costume poses with his dog.
3.there is a young boy wearing a cowboy hat hugging a dog.
4.A young boy sitting next to a brown and white dog.
5.A boy in a cowboy had posing for a picture with a dog.

**Figure 4.2:** Detailed steps for COCO data preparation. The sample picture and captions are taken from the COCO dataset[1].

The data preparation for the COCO dataset involved a structured approach that began with the establishment of a dedicated directory, termed 'coco_data'. The COCO API is streamlined to access the dataset's extensive annotations (Figure 4.2). The process encompassed downloading the dataset and unzipping the train, test, and validation data and annotation files.

Post-download, the COCO API was leveraged to process the annotations and retrieve image IDs, and corresponding metadata—essential elements for the subsequent training of our models. This comprehensive preparation equipped the images ready to be processed for augmentation and further processing.

[1] https://cocodataset.org/

#### 4.2.1.2 FLICKR Dataset

The FLICKR dataset preparation required a custom-tailored solution due to the absence of a dedicated API. A custom dataset class was developed to manage the images and their associated captions effectively (Figure 4.3). This involved the extraction of images and captions, establishing a well-organized dataset ready for further processing.

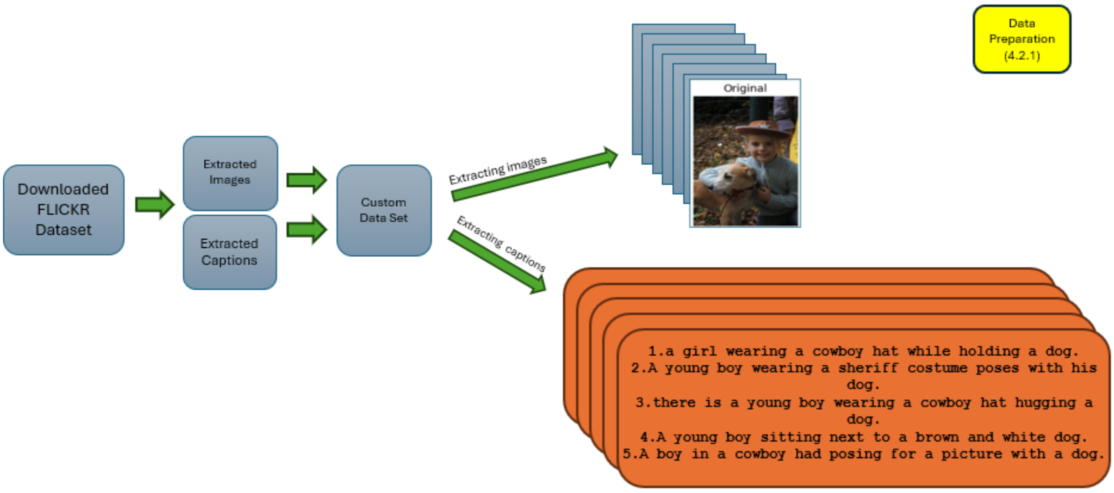

Data Preparation (4.2.1)
Downloaded FLICKR Dataset
Extracted Images
Extracted Captions
Custom Data Set
Extracting images
Extracting captions
Original
1.a girl wearing a cowboy hat while holding a dog.
2.A young boy wearing a sheriff costume poses with his dog.
3.there is a young boy wearing a cowboy hat hugging a dog.
4.A young boy sitting next to a brown and white dog.
5.A boy in a cowboy had posing for a picture with a dog.

**Figure 4.3:** Detailed steps for FLICKR data preparation.The sample picture and captions are taken from the COCO dataset[2].

The subsequent stages entailed applying necessary image transformations to standardize the data according to the requirements of our image captioning models, thus ensuring that the dataset was optimally formatted for efficient model processing.

### 4.2.2 ANN for Transformer Prediction

This section details the construction of an Artificial Neural Network (ANN) model which predicts the most effective caption generation model for an image, based on its extracted features. This prediction helps in determining whether BLIP or GPT-2 would generate a superior caption.
*Note that we also explored other machine learning models apart from ANN, more details are in Appendix. D.*

[2] https://cocodataset.org/

#### 4.2.2.1 Preprocessing and Feature Extraction with ResNet-50

We begin with the preprocessing and feature extraction phase, where a pre-trained ResNet-50 model plays a pivotal role. The model is prepared for translating complex image data into a rich feature vector, setting the stage for the subsequent prediction task. This preprocessing step takes **an image as an input** and produces the output as a **transformed tensor consisting of image features ready to be fed as input to the ann** explained next.

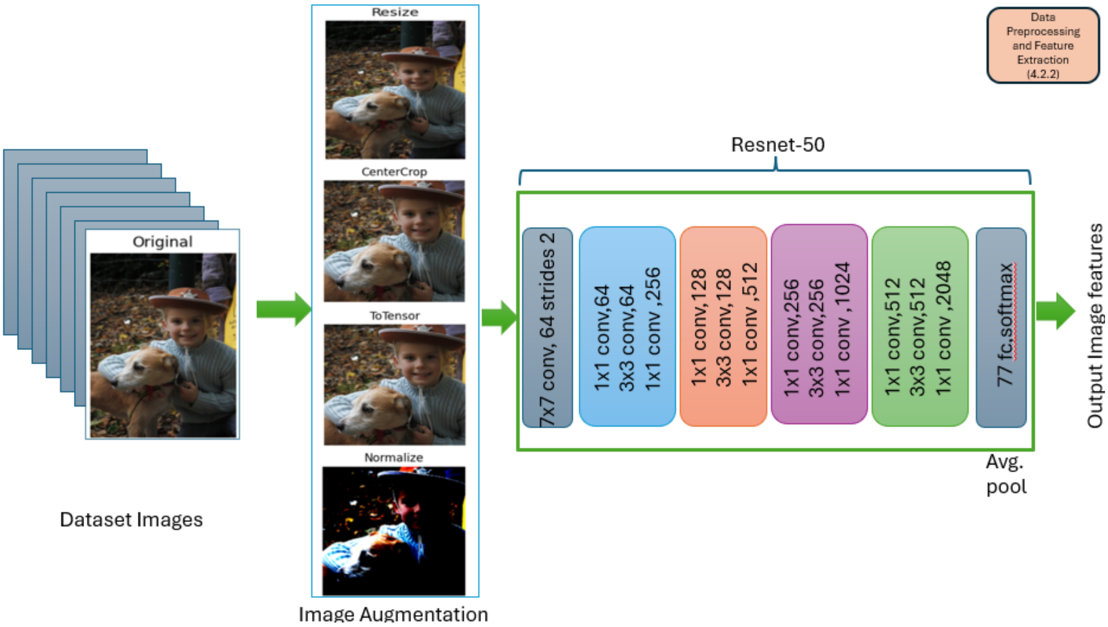


**Figure 4.4:** Preprocessing for ANN (Image Augmentation and Resnet-50). The sample picture is taken from the COCO dataset[3].

Figure 4.4 illustrates the initial processing steps performed on the images, such as resizing and center cropping, followed by normalization which adapts the images to the requirements of ResNet-50, ensuring optimal feature extraction [20].

#### 4.2.2.2 ANN Architecture

The ANN model, depicted in Figure 4.5, is composed of a sequential stack of layers, including a dense layer with 256 neurons and ReLU activation with a subsequent dense layer with 128 neurons and ReLU activation followed by another dense layer

[3] https://cocodataset.org/

with 64 neurons, also with ReLU activation and a final output layer with a single neuron employing sigmoid activation for binary classification [51]

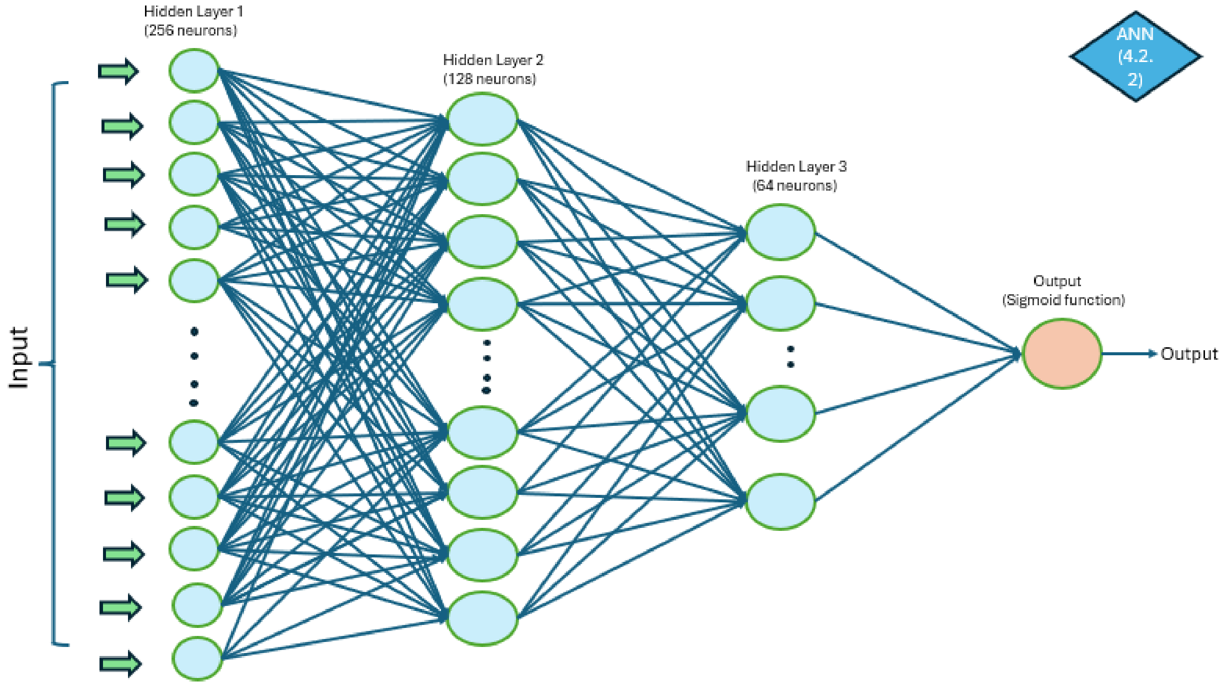


**Figure 4.5:** Architecture of the ANN model showing three sequential dense layers and a final sigmoid output layer for binary classification

This multi-layered architecture allows the model to learn complex patterns in the feature data, leading to a binary prediction that signifies the suitability of either the BLIP or GPT-2 model for generating captions. **This step takes the image features as input**(produced from the previous subsection) and produces **a label based on which system's captions have a higher average cosine similarity as an output.**

#### 4.2.2.3 Label Assignment Based on Cosine Similarity:

Binary labels for training the ANN are determined by comparing the cosine similarity between captions generated by BLIP and GPT-2 and the original dataset captions. A label (1 or 0) is assigned to indicate which model's caption is semantically closer to the human-authored text, serving as the output for this step. The **`TfidfVectorizer`** from Scikit-learn is used to transform the captions into vectorized form, allowing for the computation of cosine similarity. **(input: captions**

**text, output: cosine similarity scores)**.

#### 4.2.2.4 Training of ANN

The ANN undergoes a rigorous training process, iterating over several epochs using a dataset of 1100 image+captions. The data is divided into training (input) and testing (output) sets to validate the model's generalizability. Throughout the training, performance metrics such as loss and accuracy are monitored **(input: feature vectors and labels, output: model performance metrics)** to prevent overfitting and measure model accuracy. This trained ANN model underpins the weighted summarization approach that follows.

### 4.2.3 Caption Generation

For generating the caption from BLIP, GPT-2, or PIX2STRUCT, a common pipeline was developed and used. Here is a pseudocode describing it:

**Algorithm 4.1** Generate Image Captions Using Pre-trained Models

```
Require: Model checkpoint path, Image path
procedure LoadModel(modelCheckpoint)
    Model ← Load model from modelCheckpoint
    Tokenizer ← Load tokenizer from modelCheckpoint
end procedure
procedure GenerateCaption(imagePath)
    Image ← Load image from imagePath
    Inputs ← Preprocess image (Image, returnTensors=”pt”, padding=True)
    Outputs ← Model.generate(Inputs)
    Caption ← Decode tokens from Outputs
    return Caption
end procedure
imagePath ← ”path_to_your_image.jpg”
caption ← GenerateCaption(imagePath)
print ”Generated Caption:”, caption
```

This pseudocode serves as a template for utilizing different models to generate captions. It encapsulates the essential steps from image loading and processing to caption generation and decoding. The ”model-checkpoint” should be replaced with the specific model's checkpoint names such as 'Salesforce/blip-image-captioning-large', 'nlpconnect/vit-gpt2-image-captioning', or any other relevant model identifier depending on the specific architecture being used.

### 4.2.4 Weighted Summarization Approach

This section outlines the utilization of an advanced summarization technique that exploits the ANN model's predictions to enhance the caption quality based on the performance of the BLIP and GPT-2 models. This innovative approach dynamically assigns weights to the outputs of these models, reflecting their reliability as determined by the ANN.

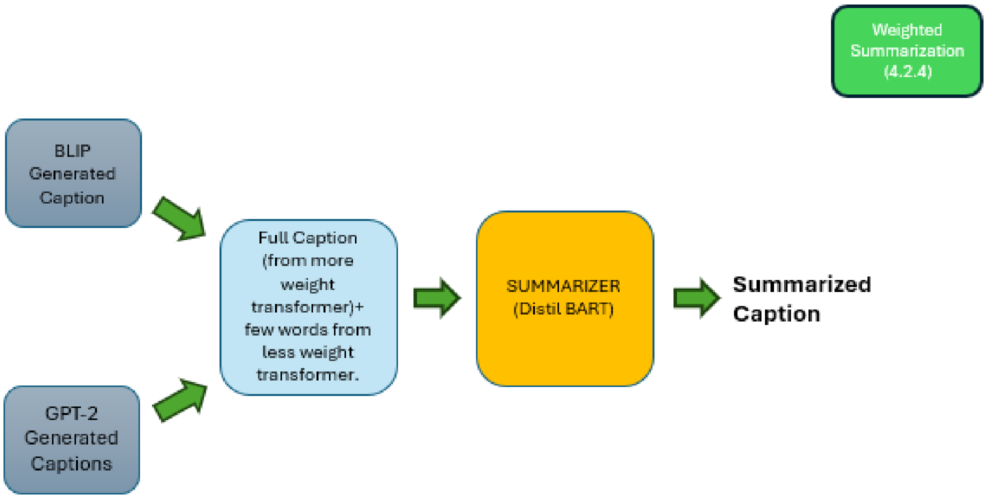

Weighted Summarization (4.2.4)
BLIP Generated Caption
GPT-2 Generated Captions
Full Caption (from more weight transformer)+ few words from less weight transformer.
SUMMARIZER (Distil BART)
Summarized Caption

**Figure 4.6:** Weighted Summarization process combining captions from BLIP and GPT-2, based on their dynamic weights from the ANN predictions.

As depicted in Figure 4.6, the summarization process involves combining the generated captions from both BLIP and GPT-2 models. The weight assigned to each caption is determined by the ANN prediction, where the caption from the more accurate model is given more significance. The selected caption is then concatenated with a portion of the caption from the other model, resulting in an input that is rich in information and diversity.

#### 4.2.4.1 Comprehensive Summarization

The comprehensive summarization strategy is thoroughly crafted to enhance the summary by giving prominence to the more reliable caption as indicated by the ANN's prediction. In practice, this involves an empirical approach to weighting, where the captions generated by GPT, when predicted to be superior, are given double representation in the input to the summarization process compared to those from BLIP. The weighted input then undergoes summarization using the `DistilBART` model. This process is not arbitrarily determined but rather the result of an experimental approach that explores various weight configurations to establish the most effective balance for summarization. Through this, the summarization model is able to amalgamate the captions, preserving the critical details and generating a unified caption that resonates more closely with the intrinsic content of the image.

#### 4.2.4.2 Selective Word Integration and Summarization

Conversely, the selective word integration method incorporates a **fixed number of words(taken from the start of the caption)** from the less weighted model's caption into the entire caption from the more heavily weighted model. The integration level is varied to determine the most effective combination that still preserves valuable context. This concatenated text is then summarized by DistilBART to produce a caption that synergizes the strengths of both BLIP and GPT-2 models.

This weighted summarization approach, underpinned by empirical analysis and ANN predictions, provides a nuanced method to synthesize captions that are not only descriptive but also contextually rich and coherent.

### 4.2.5 Evaluation of Summarization Techniques

For both approaches, the summarized captions were compared with the original dataset captions using cosine similarity to evaluate semantic alignment. The summarized caption exhibiting the highest cosine similarity with any of the original captions was documented, alongside the corresponding original caption and the computed similarity score, into a JSON file. This file served as the foundation for calculating average BLEU, METEOR, and ROUGE scores, facilitating a comprehensive evaluation of the summarized captions' quality and their alignment with human-generated annotations.

#### 4.2.5.1 Adapted Summarization Approach based on Datasets

Building upon the foundational work of developing an ANN model to predict the more effective caption generator, different summarization techniques were explored to further enhance caption quality. The ANN model's primary function is to predict which caption generator, BLIP or GPT-2, is more likely to produce a superior caption for a given image. This prediction informs the dynamic selection of summarization techniques, which were empirically tested to identify the method yielding the highest quality outcomes. **For the COCO dataset, the optimal strategy involved merging the complete caption from the model preferred by the ANN with a carefully chosen trio of keywords from the other model's caption. This choice was not arbitrary; it emerged from methodical testing with different keyword counts to determine which combination enriched the captions most effectively(see Section. 5.4 for more info). In contrast, for the FLICKR dataset, weaving five keywords from the less preferred model into the complete caption from the more capable model proved to be the best course of action. This too was an empirical decision, honed by observing the influence various keyword numbers had on the summarization quality(see Section. 5.4 for more info). Integrating five keywords struck the right chord between a detailed and targeted analysis, consistently leading to the most advantageous results in terms of precision and relevance.** These insights underscore the importance of dataset-specific strategies in achieving semantic richness and cohesiveness in the generated captions.

### 4.2.6 Post-Processing for Semantic and Grammatical Refinement

Once captions have been generated and weighted through the summarization process informed by the ANN model's predictions, we engage in a step-by-step post-processing stage. This stage refines the captions to enhance their semantic coherence and grammatical integrity, as shown in Figure 4.7.

This essential phase ensures that the final, polished caption resonates with the semantic intent of the original annotations while adhering to high linguistic standards.

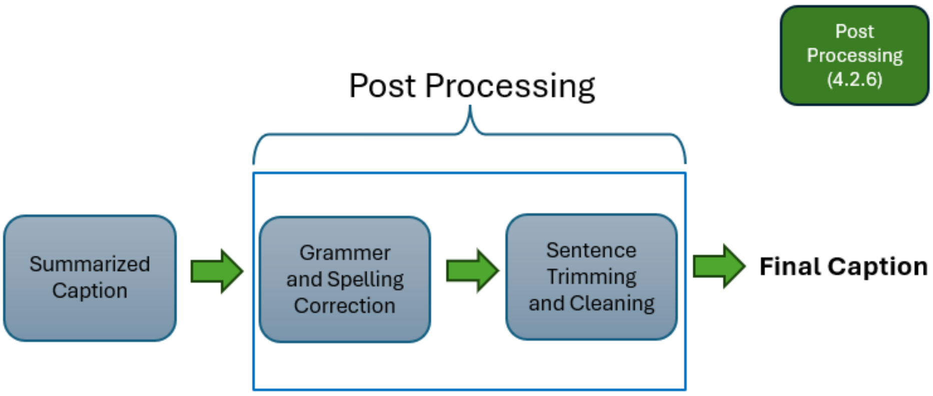


**Figure 4.7:** Post-processing steps consisting of Grammer check and Sentence Trimming

#### 4.2.6.1 Procedure for Sentence Refinement

As depicted in Figure 4.7, the post-processing workflow encompasses a series of pivotal actions to refine each summarized caption:

1. **Grammar and Spelling Corrections:** To ensure that the generated captions are free of grammatical errors and typos, we employ the `language_tool_python` library. This powerful proofreading tool checks each caption for spelling and grammatical inaccuracies, offering suggestions for corrections to create polished and error-free sentences.

**Algorithm 4.2** Correct spelling and grammar in a caption

```
procedure CorrectSpellingAndGrammar($sentence$)
    Input: $sentence$
    Output: Corrected sentence
    $tool \leftarrow$ language_tool_python.LanguageToolPublicAPI($'en - US'$)
    $matches \leftarrow tool.check(sentence)$    ▷ Identify spelling and grammar issues
    $corrected\_sentence \leftarrow$ language_tool_python.utils.correct($sentence, matches$)
▷ Apply corrections based on identified issues
    return $corrected\_sentence$
end procedure
```

This function is a critical part of the post-processing pipeline, as it helps to enhance the readability and credibility of the captions by correcting any language errors. [76]

2. **Sentence Trimming and Cleaning:** Further refinement involves pruning unnecessary conjunctions, excising repetitive elements, and clarifying ambiguous phrases, thereby bolstering readability and conciseness [77].This process involves using the Python `re` library for regular expressions to identify and modify specific patterns that enhance semantic clarity.

**Algorithm 4.3** Trim and clean sentences in a caption for readability

```
procedure RemovePattern(sentence)
    pattern ← r”\s+is\s+a\s+\w+”
    modified_sentence ← Substitute(pattern,″ , sentence)
    return modified_sentence
end procedure
procedure RemoveTrailingPhrases(sentence)
    Define a set of patterns for common trailing phrases
    for each pattern in the set do
        sentence ← Substitute(pattern,″ , sentence)
    end for
    return sentence
end procedure
procedure RemoveRepeatingParts(sentence)
    Extract and process words in the sentence to eliminate repetitions
    return The processed sentence without repetitions
end procedure
```

## 4.3 Summarization Scenarios

In addition to the above methods and steps discussed, we also tried an alternative method of summarization technique that was **later excluded because our weighted summarization combined with ann gave better results**. However, the result and insights of these scenarios were utilized for the Bayesian analysis(discussed next) which played a pivotal role in refining methodology and validating the strengths and weaknesses of the state-of-the-art models utilized. Hence it is important to discuss this.

**Scenario 1: Comprehensive Summarization** (see Fig. 4.8) begins with captions generated by BLIP (Sentence A) and GPT-2 (Sentence B), which are input into an abstraction-based summarizer like DistilBART[69]. This process yields a new summarized caption known as ”Summarized Combined Caption,” which may

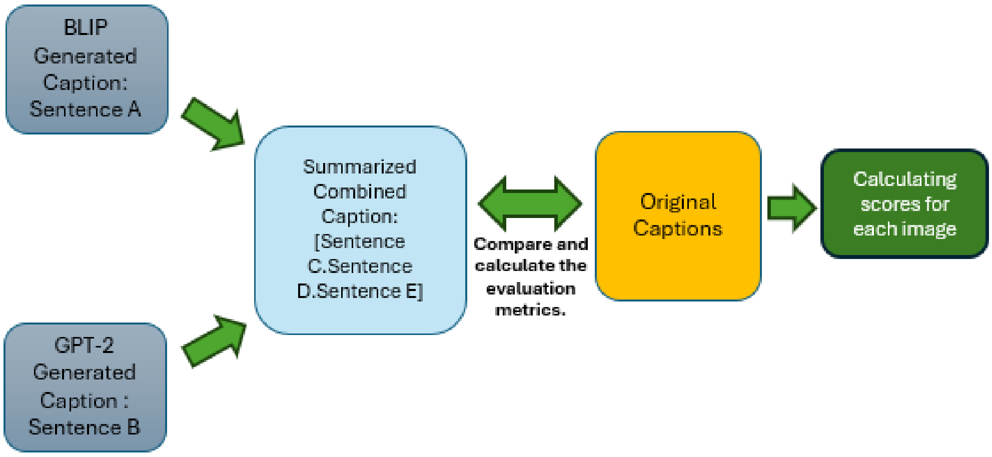

BLIP Generated Caption: Sentence A
GPT-2 Generated Caption : Sentence B
Summarized Combined Caption: [Sentence C.Sentence D.Sentence E]
Compare and calculate the evaluation metrics.
Original Captions
Calculating scores for each image

**Figure 4.8:** Flowchart for Scenario 1: Comprehensive Summarization.

include Sentences C, D, and E. These sentences are not simply combinations of Sentences A and B but may contain new elements or phrasings introduced by the summarizer. In this scenario, the whole summarized combined caption is compared to the original captions. This comprehensive summarization aims to align the machine-generated captions with human-generated annotations and is evaluated for semantic accuracy and linguistic quality using BLEU, METEOR, and ROUGE scores.

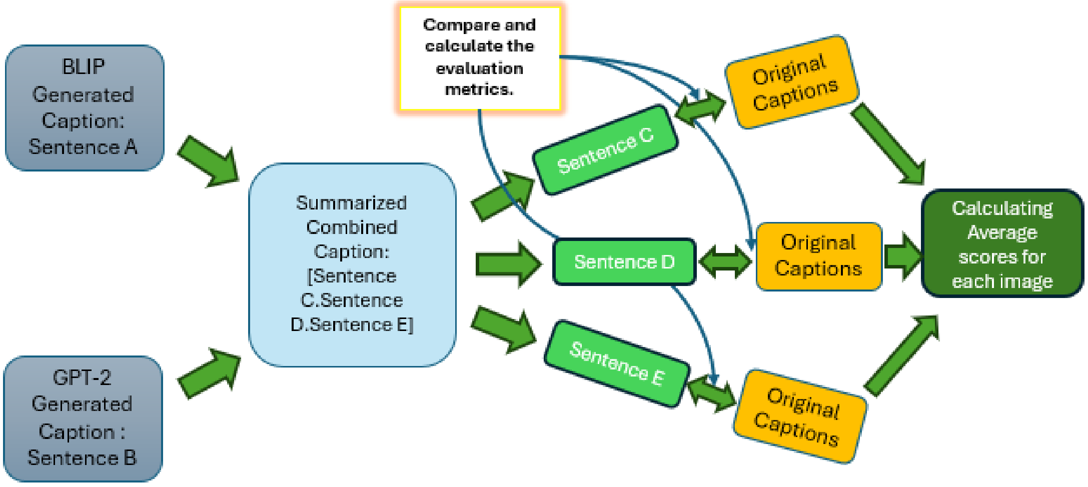

Compare and calculate the evaluation metrics.
BLIP Generated Caption: Sentence A
GPT-2 Generated Caption : Sentence B
Summarized Combined Caption: [Sentence C.Sentence D.Sentence E]
Sentence C
Sentence D
Sentence E
Original Captions
Original Captions
Original Captions
Calculating Average scores for each image

**Figure 4.9:** Flowchart for Scenario 2: Segment-Based Analysis.

**Scenario 2: Segment-Based Analysis** (see Fig. 4.9) builds upon the initial summarization by introducing a more granular, segment-based analysis. This approach further dissects the summarized captions into individual segments (Sentences C, D, and E) to enhance alignment with specific parts of the original human-generated captions. Each segment is then separately compared with the corresponding original captions, and the alignment is quantitatively evaluated using BLEU, METEOR, and ROUGE scores for each segment. This detailed comparison allows for a more nuanced understanding of how well each part of the machine-generated summary captures the nuances of human language. These individual scores are then further averaged to get a single score for an image.

In the coming sections of the next chapters 5.3 and 6.2, we will see in detail how these scenarios help in our analysis and address the research question in the thesis.

## 4.4 Bayesian Analysis for Model Comparison

Bayesian analysis provides a powerful framework for evaluating different models, such as BLIP and GPT-2, by examining how well their generated image captions align with captions created by humans and also further validating the superior performance of one transformer over another showcased on the dataset tested. This approach uses a combination of priors, likelihoods, and posterior probabilities to quantitatively determine which model produces superior results.

The performance of the BLIP and GPT-2 models was assessed using the Bayesian framework, which starts with priors based on the training data. These priors represent our starting assumption about the effectiveness of each model before any test data are observed.

**Key Terms:**

- **Prior Probability (Prior)**: Represents our initial belief about the probability of a hypothesis before observing any evidence. It is denoted as $P(H)$, where $H$ represents a hypothesis about a model's performance.

  $$P(H)$$

  The concept of prior probability was formally introduced by Reverend Thomas Bayes and is foundational in Bayesian statistics [78].

- **Likelihood**: Measures how probable the observed data is, given a particular hypothesis or model. This is represented as $P(D|H)$, where $D$ is the observed data.

$$P(D|H)$$

The likelihood function plays a crucial role in statistical inference and was extensively developed by Sir Ronald A. Fisher [79].

- **Marginal Likelihood (Evidence)**: Represents the probability of observing the data under all possible hypotheses, serving as a normalization constant. It is calculated using the following equation:

  $$P(D) = \sum_{H} P(D|H)P(H)$$

  This concept was further explored by Harold Jeffreys, who introduced the idea of using it as a tool for model comparison [80].

- **Posterior Probability (Posterior)**: The probability of a hypothesis given the observed data. It updates our knowledge after taking the evidence into account and is calculated using Bayes' Theorem:

  $$P(H|D) = \frac{P(D|H)P(H)}{P(D)}$$

  An essential step in Bayesian inference is updating our prior beliefs in light of new information, which is made possible by the Bayes Theorem. [78].

### 4.4.1 Calculation of Likelihood and Posterior Probabilities

Likelihood calculations were performed using the BLEU score to measure how closely the machine-generated captions resemble those in the test dataset. The BLEU score is a well-established metric in natural language processing that quantifies the linguistic quality and relevance of the captions [6].

**Calculation of Priors:** Priors for both GPT-2 and BLIP were calculated based on the average cosine similarity across the training dataset. This statistical foundation supports our initial assumptions about each model's caption generation capabilities. We employed the `TfidfVectorizer` for vectorization and `cosine_similarity` from the `scikit-learn` library to compute these similarities.

**Algorithm 4.4** Calculate Priors

```
Require: $data$
$data\_train, data\_test \leftarrow$ SplitData($data, 0.8$)
$better\_count \leftarrow \{'blip' : 0, 'gpt2' : 0\}$
for $item \in data\_train$ do
    $original\_captions \leftarrow$ CombineCaptions($item['original\_coco\_captions']$)
    $generated\_captions \leftarrow item['generated\_captions']$
    $cos\_similarities \leftarrow$ ComputeCosineSimilarities($generated\_captions,$ )
            $original\_captions$
    $better\_model \leftarrow$ MaxKey($cos\_similarities$)
    $better\_count[better\_model]$ += 1
end for
$prior\_gpt \leftarrow \frac{better\_count['gpt2']}{\textsc{Sum}(better\_count)}$
$prior\_blip \leftarrow \frac{better\_count['blip']}{\textsc{Sum}(better\_count)}$

return $prior\_gpt, prior\_blip$
```

**Calculation of Posterior Probabilities:** The posterior probabilities are calculated by integrating the computed priors with the likelihoods derived from the BLEU scores on the test data. This process offers a detailed quantitative analysis of the effectiveness of each model in generating image captions.

**Algorithm 4.5** Calculate Posterior Probabilities

```
Require: $data\_test, prior\_gpt, prior\_blip$
$bleu\_scores \leftarrow \{'blip' : [], 'gpt2' : []\}$
for $item \in data\_test$ do
    $original\_caps \leftarrow item['original\_coco\_captions']$
    for $model \in ['gpt2', 'blip']$ do
        $bleu\_score \leftarrow \textsc{CalcBLEU}(original\_caps, item['generated\_captions'][model])$
        $bleu\_scores[model].\textsc{Append}(bleu\_score)$
    end for
end for
$likelihood\_gpt \leftarrow \textsc{Average}(bleu\_scores['gpt2'])$
$likelihood\_blip \leftarrow \textsc{Average}(bleu\_scores['blip'])$
$marg\_lik \leftarrow likelihood\_gpt \times prior\_gpt + likelihood\_blip \times prior\_blip$
$posterior\_gpt \leftarrow \frac{likelihood\_gpt \times prior\_gpt}{marg\_lik}$
$posterior\_blip \leftarrow \frac{likelihood\_blip \times prior\_blip}{marg\_lik}$

return $posterior\_gpt, posterior\_blip$
```

**Implementation Details:** We implemented the Bayesian analysis using Python, leveraging the NLTK library for text processing and Scikit-learn for computing cosine similarities and BLEU scores. The dataset was split into 80% training and 20% testing. **Based on our assumption the BLEU score, chosen as the likelihood measure, indicates model alignment with original captions, serving as a probabilistic measure of performance. Similarly, Cosine similarity establishes priors, providing a baseline for each model's caption generation capabilities before testing, thereby anchoring the Bayesian analysis with quantifiable pre-test expectations**.

*In this chapter, we carefully outlined the approach and procedures utilized in our study, starting with the selection and preparation of datasets. We detailed each step of our analytical pipeline, from the initial data handling to the sophisticated methodologies applied. Building on this groundwork, the next chapter 'Results,' will present the findings from our methodology. We will analyze the performance of our image captioning models across different datasets and evaluate how enhancements in summarization techniques with ANN have improved the caption's semantic accuracy and contextual relevance.*

# CHAPTER 5 RESULTS

In the Result chapter, we delve into the findings from the methodology section steps that we employed previously along with addressing the research questions and framing our thesis objectives posed in this study. We start with analyzing the performance of BLIP and GPT-2 along image categories, then evaluating their performance on the datasets used, and exploring how their integration through sophisticated summarization techniques with the use of ANN can enhance the semantic accuracy and contextual relevance of generated image captions across these diverse datasets.

## 5.1 Performance Across Image Categories in Dataset

Before we dive deep into the results of the evaluation metrics of the state-of-the-art models on datasets used, and discuss various techniques involving ANN and weighted summarization to increase semantic accuracy, it is necessary to evaluate the biases of these models towards different categories. Understanding these biases is crucial in ensuring that the comparative analysis of the models is fair and takes into account the inherent strengths and weaknesses of each model's performance across various contexts. This consideration is foundational to addressing the **first research question:Q1** and establishes a baseline for a fair comparison.

We utilized the BART classifier, specifically the "facebook/bart-large-mnli" model[81], to categorize each image into predefined categories. This classification enabled a targeted analysis of the **average cosine similarity** for the captions generated by the BLIP and GPT-2 models across distinct categories. The cosine similarity, in this context, functions as a reliable measure of how closely the machine-generated captions align semantically with the human-created counterparts. Again, `TfidfVectorizer` from Scikit-learn is employed for this similarity calculation.

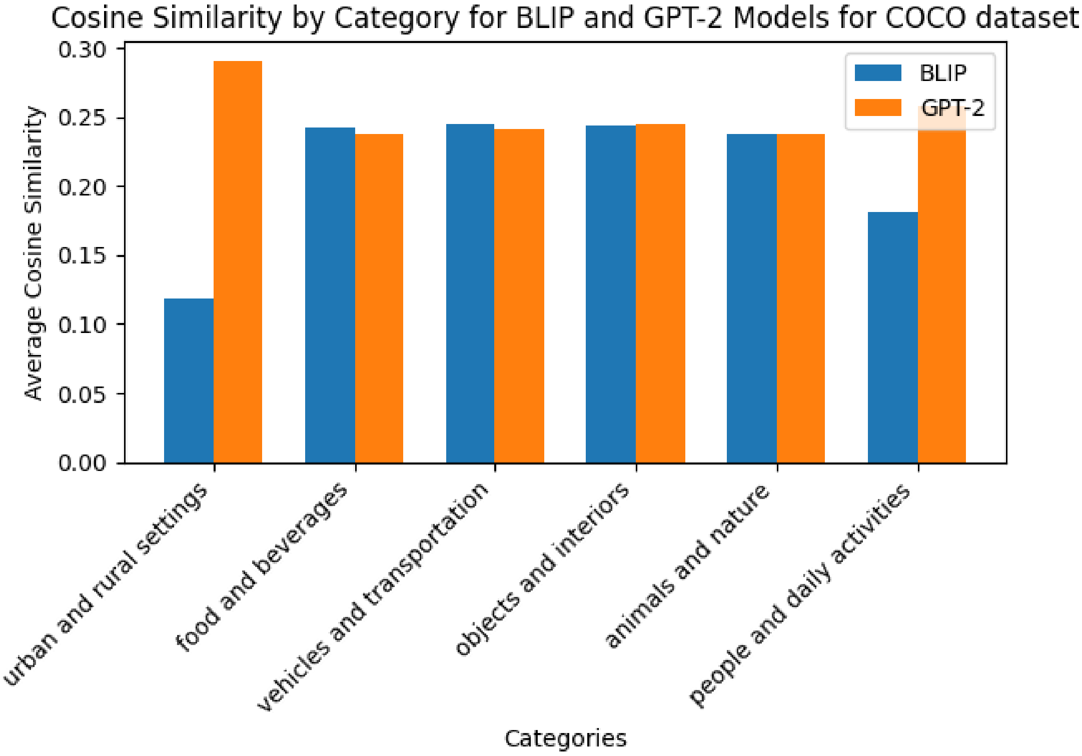


**Figure 5.1:** Category Distribution: COCO (1100 images)

The chart above 5.1 presents a comparative analysis of the **average cosine similarity** scores achieved by the BLIP and GPT-2 models across various image categories in the COCO dataset. While both models show competitive performance across categories such as *"food and beverages"*, *"vehicles and transportation"*, *"objects and interiors"*, and *"animals and nature"*, distinct variances emerge in categories *"urban and rural settings"* and *"people and daily activities"*. Here, GPT-2 outperforms BLIP with notably higher scores, recording a cosine similarity of **0.290** in *"urban and rural settings"* and **0.258** in *"people and daily activities"*, compared to BLIP's **0.119** and **0.181** respectively.

These results are further confirmed and validated in the weighted summarization approach 5.4. Specifically, in scenarios where GPT-2 was assigned more weight, our approach to evaluation led to the highest evaluation metrics. This reinforces the strategy of adapting model weights based on categorical performance to enhance the overall semantic alignment and accuracy of the generated captions, thus confirming the effectiveness of GPT-2 in processing more complex or dynamic scenes within the COCO dataset which is further validated and discussed at 6.3.1.

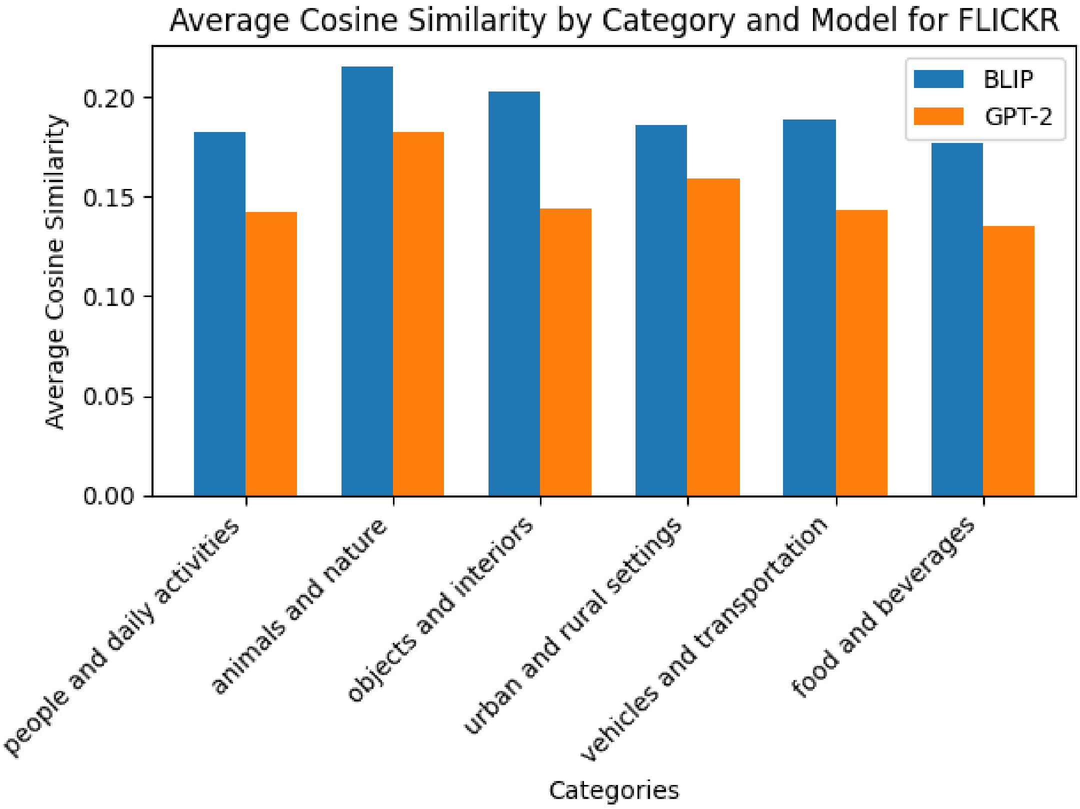


**Figure 5.2:** Category Distribution: FLICKR (1100 images)

Similarly, the chart depicted in 5.2 displays the **average cosine similarity** scores for the BLIP and GPT-2 models across various image categories within the Flickr dataset. This comparison reveals that BLIP consistently outperforms GPT-2 across all categories. Specifically, BLIP achieves its highest scores in *"animals and nature"*, *"objects and interiors"*, and *"people and daily activities"*, with cosine similarities of **0.215**, **0.203**, and **0.182** respectively. Conversely, GPT-2's best performance is observed in the *"animals and nature"* category, albeit with a lower cosine similarity of **0.182**. Even in GPT-2's strongest category, BLIP maintains a higher score of **0.215**.

These insights answer to our **first research question Q1.** as ***"Yes,models like BLIP and GPT-2 are certainly biased towards certain image categories of datasets"*** These insights are again further confirmed and validated for our weighted summarization strategy, discussed in 5.4. When applying more weight to BLIP's in flickr dataset, our methodology yielded better alignment and higher evaluation metrics. This approach has allowed us to fine-tune the semantic accuracy of the generated captions, aligning with our objective to enhance the overall quality of automated captioning, as elaborated further in 6.3.2.

**Exclusion of PIX2STRUCT for this Analysis:** Since PIX2STRUCT showed a very low similarity score, it was not feasible to include this in our categorical analysis. Hence, we excluded the results for PIX2STRUCT here.

## 5.2 Performance on Datasets

This section investigates the performance of advanced machine learning models, specifically BLIP, GPT-2, and PIX2STRUCT, on the COCO and FLICKR datasets to establish a benchmark for semantic accuracy and contextual relevance in image captions. The COCO dataset, known for its diversity and complexity, provides a comprehensive platform for testing the capability of these models to generate contextually relevant captions across a wide array of images. Similarly, the FLICKR dataset, with its unique everyday scenes, offers a distinct challenge, allowing us to evaluate model adaptability and performance in more casual, real-world settings.

By establishing these benchmarks, the study aims to identify the strengths and limitations of each model, thereby setting a foundation upon which our approaches with ANN and weighted summarization techniques can be applied. This initial analysis is crucial as it provides a baseline against which the improvements introduced by integrating multiple model outputs can be measured. The results from these evaluations not only address the **second research question Q2** regarding the enhancement of semantic accuracy and contextual relevance by current models but also set the stage for exploring innovative integration and summarization techniques that will potentially elevate the quality of generated captions.

### 5.2.1 COCO

For our analysis, each model(BLIP, GPT2, and PIX2STRUCT) generated one caption for each of the 1,100 images in the dataset. These generated captions were then compared against the five original human-annotated captions accompanying each image to compute the evaluation metrics.

#### 5.2.1.1 BLIP Generated Captions

**Table 5.1:** Average evaluation scores for BLIP generated captions on the COCO dataset(1100 images).

| Metric | BLEU | METEOR | ROUGE-1 f | ROUGE-2 f | ROUGE-l f |
|---|---|---|---|---|---|
| Average Scores | 0.305 | 0.302 | 0.401 | 0.163 | 0.369 |

Table 5.1 shows BLIP's good performance with a bleu, meteor, rouge-1 f, rouge-2 f, and rouge-l f score of **0.305,0.302,0.401,0.163 and 0.369** respectively, with a single generated caption for each image being compared against five original captions and slightly better performance in **rouge-2f** as compared to GPT-2.

#### 5.2.1.2 GPT-2 Generated Captions

**Table 5.2:** Average evaluation scores for GPT-2 generated captions on the COCO dataset(1100 images).

| Metric | BLEU | METEOR | ROUGE-1 f | ROUGE-2 f | ROUGE-l f |
|---|---|---|---|---|---|
| Average Scores | 0.317 | 0.327 | 0.404 | 0.158 | 0.369 |

As Table 5.2 indicates, GPT-2 marginally outperforms BLIP in bleu and meteor scores of **0.317 and 0.327** respectively, suggesting a slightly better grasp of language intricacies. This model also paves the way for integrating advanced techniques to refine the generated captions.

#### 5.2.1.3 PIX2STRUCT Generated Captions

**Table 5.3:** Average evaluation scores for PIX2STRUCT generated captions on the COCO dataset(1100 images).

| Metric | BLEU | METEOR | ROUGE-1 f | ROUGE-2 f | ROUGE-l f |
|---|---|---|---|---|---|
| Average Scores | 0.054 | 0.165 | 0.227 | 0.029 | 0.206 |

Table 5.3 demonstrates that PIX2STRUCT lags behind its counterparts with lower scores in bleu, meteor scores of **0.054,0.165** and rouge-1 f, rouge-2 f, rouge-l f scores of **0.227,0.029 and 0.206** respectively highlighting its current limitations in addressing the intricacies of the COCO dataset and underscoring the need for enhanced approaches.

### 5.2.2 FLICKR

Similar to COCO, for the FLICKR dataset each model(BLIP, GPT2, and PIX2STRUCT) generated one caption for each of the 1,100 images in the dataset. These generated captions were then compared against the five original human-annotated captions accompanying each image to compute the evaluation metrics.

#### 5.2.2.1 BLIP Generated Captions

**Table 5.4:** Average evaluation scores for BLIP generated captions on the Flickr dataset(1100 images).

| Metric | BLEU | METEOR | ROUGE-1 f | ROUGE-2 f | ROUGE-l f |
|---|---|---|---|---|---|
| Average Scores | 0.177 | 0.292 | 0.321 | 0.097 | 0.282 |

The BLIP model achieved a BLEU score of **0.177**, a METEOR score of **0.292**, and a ROUGE-1 f, ROUGE-2 f, and ROUGE-l f score of **0.321, 0.097 and 0.282** respectively on the Flickr dataset 5.4. While the BLEU score suggests room for improvement, the high METEOR and ROUGE-1 f, ROUGE-2 f, and ROUGE-L f scores indicate BLIP's strong ability to create captions with high semantic relevance.

#### 5.2.2.2 GPT-2 Generated Captions

**Table 5.5:** Average evaluation scores for GPT-2 generated captions on the Flickr dataset(1100 images).

| Metric | BLEU | METEOR | ROUGE-1 f | ROUGE-2 f | ROUGE-l f |
|---|---|---|---|---|---|
| Average Scores | 0.149 | 0.231 | 0.286 | 0.076 | 0.263 |

GPT-2's scores on Flickr, detailed in Table 5.5, reveal a BLEU score of **0.149** and a METEOR score of **0.231**, and a ROUGE-1 f, ROUGE-2 f, and ROUGE-l f score of **0.286, 0.076 and 0.263** respectively, which are somewhat lower than BLIP's, indicating a potential difficulty of GPT-2 in adapting its language model to the diverse imagery and casual context of the Flickr dataset. Also, GPT-2's lower performance on Flickr compared to COCO suggests possible limitations in handling less structured visual data, prompting further investigation into how model training and summarization techniques can be optimized for such scenarios.

#### 5.2.2.3 PIX2STRUCT Generated Captions

**Table 5.6:** Average evaluation scores for PIX2STRUCT generated captions on the Flickr dataset. (1100 images)

| Metric | BLEU | METEOR | ROUGE-1 f | ROUGE-2 f | ROUGE-l f |
|---|---|---|---|---|---|
| Average Scores | 0.042 | 0.154 | 0.215 | 0.027 | 0.198 |

PIX2STRUCT's scores, as seen in Table 5.6, were considerably lower across all metrics, highlighting the model's limitations in effectively capturing the essence of Flickr's image content.

**Exclusion of PIX2STRUCT from Further Analysis:** Despite PIX2STRUCT's innovative approach to image captioning, its performance on both the COCO and Flickr datasets was significantly lower than that of BLIP and GPT-2. Given its lower BLEU, METEOR, and ROUGE scores, PIX2STRUCT was excluded from further analysis to focus on enhancing the performance of the more promising models. This decision was made to ensure that subsequent steps in the research, including summarization and model comparison analyses, were based on the highest quality input data, thereby avoiding the potential negative impact of including lower-performing models on overall results.

## 5.3 Bayesian Analysis Results

This section extends the evaluation of machine learning models BLIP and GPT-2 through Bayesian analysis, offering a probabilistic understanding of their performance on the COCO and FLICKR datasets. This statistical method complements the findings from the previous section, confirming the models' effectiveness with a focus on semantic accuracy and contextual relevance, directly addressing **the research questions Q1 and Q2 posed**.

Both the datasets are analyzed in two scenarios, as described already in Sec. 4.3 Scenario 1 which refers to the comprehensive summarization in which the `DistilBART` model was employed to summarize captions generated by BLIP and GPT-2 models. On the other hand Scenario 2 involves segment-based summarization, which is splitting summarized captions into segments for individual comparison against original dataset annotations. In both scenarios, summaries were compared against original COCO and FLICKR dataset annotations using BLEU, METEOR, and ROUGE scores, to assess semantic alignment, and the best caption with the highest similarity was stored for further use in Bayesian analysis.

### 5.3.1 COCO Dataset Analysis

The results highlight GPT-2's slightly better performance with higher posterior probabilities of **0.5672 and 0.5458** in the two scenarios, compared to BLIP's **0.4328 and 0.4542**. This confirms GPT-2's capability to generate more semantically rich captions, which aligns with its superior BLEU and METEOR scores of **0.317 and 0.327** respectively as compared to BLIP's BLEU and METEOR scores of **0.305 and 0.302** respectively which can be seen from the previous section. The

**Table 5.7:** Bayesian Analysis Results on COCO Dataset(1100 images)

| **Parameter** | **Scenario 1** | **Scenario 2** |
|---|---|---|
| Prior for GPT-2 | 0.5482 | 0.5608 |
| Prior for BLIP | 0.4518 | 0.4392 |
| Likelihood for GPT-2 | 0.3202 | 0.3054 |
| Likelihood for BLIP | 0.2964 | 0.3245 |
| Marginal Likelihood | 0.3095 | 0.3138 |
| Posterior for GPT-2 | 0.5672 | 0.5458 |
| Posterior for BLIP | 0.4328 | 0.4542 |

marginal and likelihood values further substantiate their alignment with the complex visual and textual requirements of the COCO dataset.

### 5.3.2 Flickr Dataset Analysis

**Table 5.8:** Bayesian Analysis Results on FLICKR Dataset(1100 images)

| **Parameter** | **Scenario 1** | **Scenario 2** |
|---|---|---|
| Prior for GPT-2 | 0.3136 | 0.3068 |
| Prior for BLIP | 0.6864 | 0.6932 |
| Likelihood for GPT-2 | 0.1555 | 0.1538 |
| Likelihood for BLIP | 0.1772 | 0.1826 |
| Marginal Likelihood | 0.1704 | 0.1737 |
| Posterior for GPT-2 | 0.2862 | 0.2715 |
| Posterior for BLIP | 0.7138 | 0.7285 |

BLIP's performance on the Flickr dataset, with posterior probabilities of **0.7138 and 0.7285**, significantly surpasses GPT-2's **0.2862 and 0.2715**. This validates the earlier evaluation results where BLIP achieved higher BLEU, METEOR, ROUGE-1 f, ROUGE-2 f, and ROUGE-l f scores of **0.177, 0.292, 0.321, 0.097 and 0.282** respectively which can be seen from the previous section. These metrics underscore BLIP's better adaptability to the casual and diverse imagery of the Flickr dataset compared to GPT-2's scores of **0.149** in BLEU, **0.231** in METEOR, **0.286** in ROUGE-1 f, **0.076** in ROUGE-2 f and **0.263** in ROUGE-l f.

The likelihood and marginal likelihood numbers support BLIP's stronger semantic alignment with the human-like understanding required for the Flickr dataset's captions.

The Bayesian analysis, combined with the detailed summarization scenarios, conclusively demonstrates how advanced machine learning models like BLIP and GPT-2 enhance semantic accuracy and contextual relevance across varied datasets, directly addressing the **second research question Q2**. Additionally, this analysis confirms the effectiveness of using advanced summarization strategies that combine outputs from various models(**discussed next**), thereby enhancing the richness and depth of the generated captions, which addresses the **third research question Q3**. Overall, the Bayesian approach not only validates the effectiveness of the implemented summarization techniques but also reinforces their role in advancing the field of automated image captioning, pointing to significant avenues for future enhancements.

## 5.4 Results for Weighted Summarization Approaches

Following the analysis of BLIP, GPT-2, and PIX2STRUCT on the COCO and FLICKR datasets, and validating these findings with Bayesian analysis, we now explore the impact of advanced summarization techniques aimed at enhancing these initial results. This section examines how weighted summarization approaches can further improve the performance of image captioning models, thereby increasing the semantic alignment with the original human-generated captions. The weighted summarization process involves adjusting the input based on weights assigned to the outputs from different models, informed by an Artificial Neural Network (ANN) that predicts the most effective model for each image. This predictive approach establishes a solid foundation for determining which model's output should be given higher/lower weight. This section describes the different results of the experimental approach that explores various weight configurations to establish the most effective balance for summarization.

## 5.4.1 COCO Dataset

### 5.4.1.1 Weighted Caption Integration

This approach involves adjusting the input to the summarization process based on the assigned weights to BLIP and GPT-generated captions on the COCO Dataset. Heavier captions are fed twice as much as less heavy ones.

**Table 5.9:** Average evaluation metrics for weighted caption integration on COCO dataset (1100 images)

| **Model Weighing** | **BLEU** | **METEOR** | **ROUGE-1 F** | **ROUGE-2 F** | **ROUGE-L F** |
|---|---|---|---|---|---|
| BLIP heavier | 0.266 | 0.314 | 0.422 | 0.160 | 0.383 |
| GPT-2 heavier | 0.270 | 0.320 | 0.426 | 0.164 | 0.386 |

The table 5.9 reveals a slight advantage in all metrics for GPT-2 heavier integration, with Average **BLEU 0.270**, **METEOR 0.320**, **ROUGE-1 F 0.426**, **ROUGE-2 F 0.164**, and **ROUGE-L F 0.386** being marginally higher than BLIP heavier, suggesting GPT-2's refined capability in generating syntactically and semantically richer captions.

### 5.4.1.2 Selective Word Integration

This method includes a fixed number of words(taken from starting of the caption) from the less weighted model's caption combined with the entire caption from the more heavily weighted model.

As seen from Table. 5.10 , incorporating 3 words from BLIP's captions yields the highest Average **BLEU 0.278**, **METEOR 0.313** and **ROUGE-1 F 0.425**, indicating an optimal blend of additional context without overshadowing the main caption's semantic integrity.

In contrast, GPT-2's captions demonstrate a consistent increase in semantic alignment as more words are included, peaking with 5 words, with Average **BLEU 0.271**, which can be seen from Table. 5.11 demonstrating better narrative flow and coherence in the extended context.

**Table 5.10:** Average evaluation metrics for selective word integration from BLIP on COCO dataset (1100 images)

| Words from BLIP | BLEU | METEOR | ROUGE-1 F | ROUGE-2 F | ROUGE-L F |
|---|---|---|---|---|---|
| 2 words | 0.264 | 0.298 | 0.422 | 0.159 | 0.390 |
| 3 words | 0.278 | 0.313 | 0.425 | 0.161 | 0.388 |
| 4 words | 0.275 | 0.311 | 0.421 | 0.161 | 0.384 |
| 5 words | 0.276 | 0.317 | 0.425 | 0.161 | 0.388 |

**Table 5.11:** Average evaluation metrics for selective word integration from GPT-2 on COCO dataset (1100 images)

| Words from GPT-2 | BLEU | METEOR | ROUGE-1 F | ROUGE-2 F | ROUGE-L F |
|---|---|---|---|---|---|
| 2 words | 0.222 | 0.316 | 0.428 | 0.156 | 0.396 |
| 3 words | 0.270 | 0.295 | 0.427 | 0.162 | 0.392 |
| 4 words | 0.268 | 0.297 | 0.425 | 0.161 | 0.389 |
| 5 words | 0.271 | 0.306 | 0.428 | 0.163 | 0.392 |

### 5.4.2 Flickr Dataset

#### 5.4.2.1 Weighted Caption Integration

Similar to COCO, This approach involves adjusting the input to the summarization process based on the assigned weights to BLIP and GPT-generated captions on the Flickr Dataset. Heavier caption are fed twice as much as less heavy one.

The GPT-2 heavier integration on Flickr shows a modest improvement across all metrics, Table. 5.12 with Average **BLEU 0.149**, **METEOR 0.305**, **ROUGE-1 F 0.365**, **ROUGE-2 F 0.105**, and **ROUGE-L F 0.330**, indicating a better fit for the casual imagery in the Flickr dataset compared to BLIP.

**Table 5.12:** Average evaluation metrics for weighted caption integration on Flickr dataset (1100 images)

| **Model Weighing** | **BLEU** | **METEOR** | **ROUGE-1 F** | **ROUGE-2 F** | **ROUGE-L F** |
|---|---|---|---|---|---|
| BLIP heavier | 0.146 | 0.281 | 0.348 | 0.098 | 0.316 |
| GPT-2 heavier | 0.149 | 0.305 | 0.365 | 0.105 | 0.330 |

#### 5.4.2.2 Selective Word Integration

This method includes a fixed number of words from the less weighted model's caption combined with the entire caption from the more heavily weighted model.

**Table 5.13:** Average evaluation metrics for selective word integration from BLIP on FLICKR dataset (1100 images)

| **Words from BLIP** | **BLEU** | **METEOR** | **ROUGE-1 F** | **ROUGE-2 F** | **ROUGE-L F** |
|---|---|---|---|---|---|
| 2 words | 0.117 | 0.232 | 0.310 | 0.076 | 0.284 |
| 3 words | 0.123 | 0.233 | 0.322 | 0.080 | 0.296 |
| 4 words | 0.130 | 0.244 | 0.343 | 0.087 | 0.314 |
| 5 words | 0.129 | 0.244 | 0.344 | 0.088 | 0.314 |

From Table. 5.13, including 4 words from BLIP optimizes the contextual richness, achieving Average **BLEU 0.130**, **METEOR 0.244**, **ROUGE-1 F 0.343**, **ROUGE-2 F 0.087**, and **ROUGE-L F 0.314**.

From Table. 5.14, including 5 words from GPT-2 offers the best balance, yielding Average **BLEU 0.175 which is highest across all cases**, **METEOR 0.298**, **ROUGE-1 F 0.328**, **ROUGE-2 F 0.099**, and **ROUGE-L F 0.288**, demonstrating a slightly better understanding of the image content compared to BLIP in the selective word integration scenario.

The weighted summarization techniques applied to the COCO and FLICKR datasets enhance semantic accuracy and contextual relevance, directly addressing and answering our **third research question Q3**. By comparing Table. 5.2 and

**Table 5.14:** Average evaluation metrics for selective word integration from GPT-2 on FLICKR dataset (1100 images)

| Words from GPT-2 | BLEU | METEOR | ROUGE-1 F | ROUGE-2 F | ROUGE-L F |
|---|---|---|---|---|---|
| 2 words | 0.172 | 0.296 | 0.324 | 0.096 | 0.285 |
| 3 words | 0.173 | 0.296 | 0.326 | 0.098 | 0.287 |
| 4 words | 0.174 | 0.297 | 0.329 | 0.100 | 0.290 |
| 5 words | 0.175 | 0.298 | 0.328 | 0.099 | 0.288 |

Table. 5.9 for COCO, Table. 5.4 and Table. 5.14 for FLICKR, we see that these techniques not only improved BLEU, METEOR, and ROUGE scores but also demonstrated the effectiveness of integrating multiple model outputs to deepen the semantic richness of captions. This approach highlights the potential for advanced summarization to produce captions that better reflect human-like understanding, suggesting that current evaluation metrics may need refinement to fully capture these improvements. Overall, these findings confirm that strategic model integration can significantly advance the field of automated image captioning.

## 5.5 Results using the ANN Model

The ANN model was trained to predict which caption generator would produce a caption more closely aligned with the human-annotated captions in the dataset. This prediction informed a dynamic summarization strategy that adjusted the emphasis on captions generated by the two models. The performance of the ANN model was evaluated separately for the COCO and Flickr datasets.

### 5.5.1 Results on COCO Dataset

To better illustrate the process, let's visit a specific example from our experiment results on the COCO dataset. As the figure 5.3 shows once the image is preprocessed and fed to ann, the ann predicts and gives weights to the generated captions from blip which is **'a bike leaning on a pole'** and from GPT-2 **'a dog standing**

[1] https://cocodataset.org/

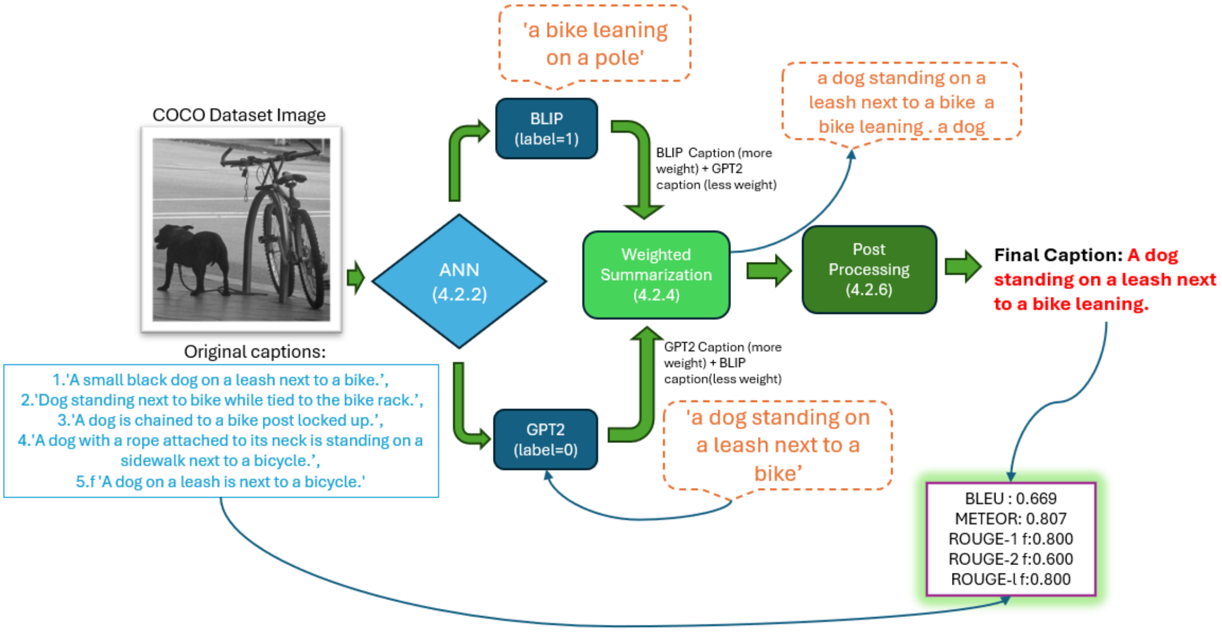

COCO Dataset Image
Original captions:
1.'A small black dog on a leash next to a bike.',
2.'Dog standing next to bike while tied to the bike rack.',
3.'A dog is chained to a bike post locked up.',
4.'A dog with a rope attached to its neck is standing on a sidewalk next to a bicycle.',
5.f 'A dog on a leash is next to a bicycle.'
ANN
(4.2.2)
BLIP
(label=1)
'a bike leaning on a pole'
GPT2
(label=0)
'a dog standing on a leash next to a bike'
BLIP Caption (more weight) + GPT2 caption (less weight)
GPT2 Caption (more weight) + BLIP caption(less weight)
Weighted Summarization
(4.2.4)
a dog standing on a leash next to a bike a bike leaning . a dog
Post Processing
(4.2.6)
Final Caption: A dog standing on a leash next to a bike leaning.
BLEU : 0.669
METEOR: 0.807
ROUGE-1 f:0.800
ROUGE-2 f:0.600
ROUGE-l f:0.800

**Figure 5.3:** Overall Process flow example: COCO(sample image taken from the COCO dataset[1]used.)

**on a leash next to a bike'**. These captions now undergo weighted summarization in which fully generated caption, in this case from GPT-2, and few-words(in this case first 3 words) from BLIP generated caption are combined resulting in the summarized caption **'a dog standing on a leash next to a bike leaning. a dog'**, since the caption is not very meaningful, we post-process it and get the final caption as **'A dog standing on a leash next to a bike leaning'** which when compared to the original caption produces a **BLEU, METEOR, ROUGE-1 f, ROUGE-2 f and ROUGE-l f scores of 0.669,0.807,0.800,0.600 and 0.800** respectively. This process goes on iterating over all the 1100 images considered in the COCO dataset and the resulting average scores of evaluation metrics are presented.

For the COCO dataset, the ANN model achieved a test accuracy of **73.39%**, demonstrating its capability to effectively predict the superior caption generator based on image features. **For more details on the learning curve over epochs, refer to appendix:C.1**.

The table 5.15 illustrates a notable enhancement in the quality of the captions, as evidenced by the improved scores across all evaluation metrics. Without ANN the highest average scores achieved on COCO were from GPT-2:**BLEU 0.317**,

**Table 5.15:** Improvement in Evaluation Metrics with ANN-Based Weighted Summarization for COCO Dataset(1100 images)

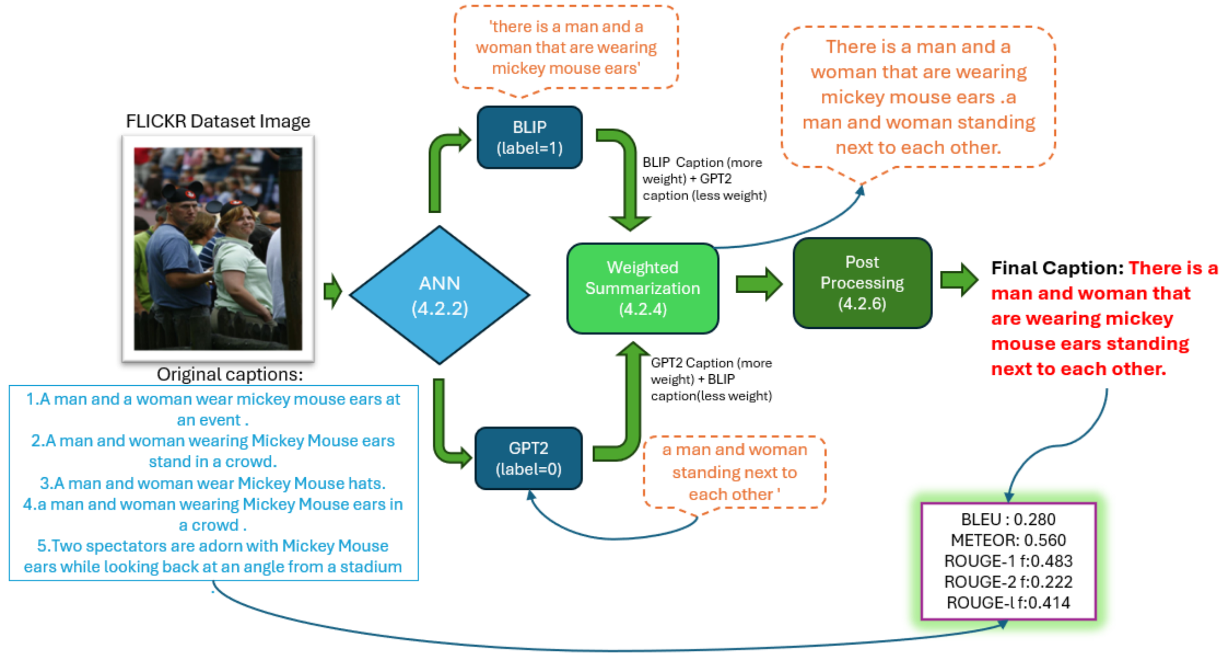
| Metric | Average Score |
|---|---|
| BLEU | 0.322 |
| METEOR | 0.328 |
| ROUGE-1 F | 0.452 |
| ROUGE-2 F | 0.187 |
| ROUGE-L F | 0.415 |

**METEOR 0.327** and **ROUGE-1 f, ROUGE-2 f, ROUGE-l f scores** of **0.404,0.158 and 0.369** respectively, with ANN the performance is indeed improved with scores of Average **BLEU 0.322**, **METEOR 0.328** and **ROUGE-1 f, ROUGE-2 f and ROUGE-l f scores** of **0.452,0.187 and 0.415** respectively. This proves the effectiveness of the ANN model in guiding a more aligned summarization strategy which increases the values of all the metrics considered in this study.

### 5.5.2 Results on FLICKR Dataset

Similarly for the FLICKR dataset, let's have a look at an example from our experiment results as shown in Figure 5.4. Once a preprocessed image is ready, the ANN assigns weights to the generated captions: BLIP's **'there is a man and a woman that are wearing mickey mouse ears'** and GPT-2's **'a man and woman standing next to each other'**. The weighted summarization merges the complete BLIP caption repeated twice with GPT-2 caption repeated only once(in this case), forming the summarized caption **'There is a man and a woman that are wearing mickey mouse ears. a man and woman standing next to each other**. The caption, albeit somewhat redundant, is post-processed to enhance coherence, finally rendering the caption as **'There is a man and woman that are wearing mickey mouse ears standing next to each other.'**. Compared to the dataset's original annotations, this final caption achieves a **BLEU score of 0.280**, a **METEOR score of 0.560**, **ROUGE scores of 0.483 for ROUGE-1 f, 0.222 for ROUGE-2 f,** and **0.414 for ROUGE-L f**. The same

[2] https://www.kaggle.com/datasets/adityajn105/flickr8k



FLICKR Dataset Image
Original captions:
1.A man and a woman wear mickey mouse ears at an event .
2.A man and woman wearing Mickey Mouse ears stand in a crowd.
3.A man and woman wear Mickey Mouse hats.
4.a man and woman wearing Mickey Mouse ears in a crowd .
5.Two spectators are adorn with Mickey Mouse ears while looking back at an angle from a stadium
ANN (4.2.2)
BLIP (label=1)
'there is a man and a woman that are wearing mickey mouse ears'
GPT2 (label=0)
a man and woman standing next to each other '
BLIP Caption (more weight) + GPT2 caption (less weight)
GPT2 Caption (more weight) + BLIP caption(less weight)
Weighted Summarization (4.2.4)
There is a man and a woman that are wearing mickey mouse ears .a man and woman standing next to each other.
Post Processing (4.2.6)
Final Caption: There is a man and woman that are wearing mickey mouse ears standing next to each other.
BLEU : 0.280
METEOR: 0.560
ROUGE-1 f:0.483
ROUGE-2 f:0.222
ROUGE-l f:0.414

**Figure 5.4:** Overall Process flow example: FLICKR(sample image taken from Flickr dataset$^2$used.)

process is iterated similarly to over 1100 images in the Flickr dataset and the average evaluation metrics are presented.

For the Flickr dataset, the ANN model showcased a test accuracy of **60.91%**, indicating its efficiency in predicting the better caption generator for images unique to this dataset. **For more details on the learning curve over epochs, refer to appendix:C.2**.

Similarly, the Flickr dataset experienced an improvement in caption quality and all of the evaluation metrics considered, after applying the ANN-based weighted-summarization technique 5.16. Without ANN the average highest scores achieved on FLICKR were from BLIP:**BLEU 0.177**, **METEOR 0.292** and **ROUGE-1 f, ROUGE-2 f and ROUGE-l f scores** of **0.321,0.097 and 0.282** respectively, with ANN the performance is indeed improved with Average scores of **BLEU 0.181**, **METEOR 0.300** and **ROUGE-1 f, ROUGE-2 f and ROUGE-l f scores** of **0.348,0.107 and 0.311** respectively. This further validates the adaptability and potential of the proposed method across different datasets.

As already seen how the integration of outputs from multiple models with neural networks, as evidenced by the increased scores, supports the enhancement of se-

**Table 5.16:** Improvement in Evaluation Metrics with ANN-Based Weighted Summarization for FLICKR Dataset(1100 images)

| Metric | Average Score |
|---|---|
| BLEU | 0.181 |
| METEOR | 0.300 |
| ROUGE-1 F | 0.348 |
| ROUGE-2 F | 0.107 |
| ROUGE-L F | 0.311 |

mantic depth in captions, which answers our **third research question Q3** as ***"Yes, it does improve the semantic depth of generated captions."*** Additionally, the results from the ANN model significantly bolster semantic accuracy and contextual relevance across various datasets, affirming its effectiveness in enhancing key NLP metrics and **performing better than the state-of-the-art models like BLIP, GPT-2 and PIX2STRUCT**, addressing the first part of the **fourth research question Q4**. It leverages the strengths of different models to produce more nuanced captions that are contextually appropriate. Overall, this analysis underscores the effectiveness of ANN models in guiding summarization strategies, demonstrating how such techniques can bridge the gap between automated systems and human-like captioning capabilities.

*This chapter presents the results of our creative approaches, including how the performance of GPT-2 and BLIP across datasets was evaluated using a variety of metrics, as well as the effects of advanced summarization strategies and the incorporation of Artificial Neural Networks (ANNs) on the contextual relevance and semantic accuracy of the generated captions. We go into more detail about how to evaluate these data in the upcoming chapter and go over how each finding relates to the initial research questions and thesis objectives..*

# CHAPTER 6

# DISCUSSION

In the Discussion Chapter, we discuss in detail the results and findings from the previous chapter with a focus on evaluation metrics, what the higher/lower metrics with comparison signify and indicate for a model or an approach, and how they answer our research question and thesis objectives. We start by discussing the performance of BLIP and GPT-2 on the COCO and FLICKR dataset, then we move on to discussing the summarization scenarios that assisted in the confirmation of our Bayesian analysis, further discussing the results from weighted summarization and finalizing the discussion with the overall results from the ANN integration to our approach and the state-of-the-art models performance on datasets.

## 6.1 Performance on Image Categories and Datasets

Let's start the discussion with answering and discussing our **first research question Q1: Are advanced machine learning models like BLIP and GPT-2 predisposed to generating more accurate and semantically rich captions for certain categories of images?** The evidence from our evaluations across two distinct datasets, COCO and FLICKR, as seen from 5.1 affirms this, highlighting significant biases in how each model performs depending on the image category. In the COCO dataset, GPT-2 exhibited a strong preference for "urban and rural settings" with a cosine similarity score of **0.290**, significantly outperforming BLIP's **0.119** in the same category. Additionally, GPT-2 also led in "people and daily activities" with a score of **0.258**, compared to BLIP's **0.181**. Conversely, BLIP performed more consistently across categories in the FLICKR dataset, **notably outperforming GPT-2 in all categories**, with its highest scores in "animals and nature" **(0.215)** and "objects and interiors" **(0.203)**. This distinct variances in model performance suggest that **GPT-2 may be more adept at processing complex, dynamic scenes, as evidenced by its higher scores in urban settings and active human scenes in the COCO dataset. In contrast, BLIP shows a robust capability across a broader range of image types, particularly excelling in more static and natural scenes, as seen in the FLICKR dataset**. These insights confirm that while both models are capable of producing high-quality captions, their effectiveness can vary dramatically with the image content. This understanding is crucial for deploying these models in real-world applications, where choosing the right model for a specific type of image can greatly enhance the accuracy and relevance of the generated captions.

Moving forward, the evaluation of **BLIP, GPT-2, and PIX2STRUCT** on the **COCO and FLICKR** datasets provides a comprehensive look at their capabilities to address the **second research question Q2: How do advanced machine learning models like BLIP and GPT-2 enhance the semantic accuracy and contextual relevance of the image captions across varied datasets?**, which probes the enhancement of semantic accuracy and contextual relevance across varied image datasets. Table 5.1 shows BLIP's solid performance with a BLEU score of **0.305**, a METEOR score of **0.302**, and a ROUGE-1 f, ROUGE-2 f, and ROUGE-l f score of **0.401, 0.163 and 0.369** respectively, indicating its capability to produce relevant and detailed captions that resonate with human annotations. This performance sets the stage for applying advanced weighted summarization techniques aimed at further enhancing caption quality. Conversely, as detailed in Table 5.2, GPT-2 exhibits slightly superior BLEU and METEOR scores of **0.317** and **0.327**, underscoring its ability to capture the semantic essence of the images more effectively, which hints at its robust capability to translate visual content into semantically rich text. However, PIX2STRUCT lagged significantly behind in both datasets, especially evident from its lower scores, which highlight its challenges in translating complex visual scenes into coherent textual descriptions. On the FLICKR dataset, known for its casual and less structured visual content, both BLIP and GPT-2 showed a dip in performance, necessitating a higher level of abstraction and contextual interpretation. Despite this, BLIP maintained relatively stable METEOR scores, as shown in Table 5.4, with scores of **0.292**, emphasizing its robustness in varied contexts. GPT-2, however, experienced a decrease in performance as compared to BLIP which is evidenced in Table

5.5, with a BLEU score of **0.149**, a METEOR score of **0.231** and a ROUGE-1 f, ROUGE-2 f, and ROUGE-l f score of **0.286, 0.076 and 0.263** respectively, indicating potential difficulties in adapting its model to the diverse imagery of the FLICKR dataset. These observations across datasets underline the strengths and limitations of each model, offering insights into how they meet the challenges posed by different visual contents and setting a foundation for further advancements in image captioning technology.

## 6.2 Evaluation and Discussion of Summarization Scenarios

Now let's assesses the impact of summarization techniques on the alignment of generated captions with human-generated annotations across two detailed scenarios. As already discussed in Sec. 4.3 about Scenario 1:Comprehensive Summarization and Scenario 2: Segment-based analysis, further following along these individual assessments, an intercomparison identifies the captions that best align with original annotations, which are then evaluated using average BLEU, METEOR, and ROUGE scores to assess their linguistic quality and semantic accuracy. This methodical approach not only advances caption precision but also integrates findings from the previous chapter on the efficacy of weighted summarization techniques.

**Table 6.1:** Evaluation of Summarization Scenarios on COCO Dataset.(1100 images)

| **Metric(Average Scores)** | **Scenario 1** | **Scenario 2** | **Intercomparison** |
|---|---|---|---|
| BLEU | 0.243 | 0.231 | 0.281 |
| METEOR | 0.334 | 0.336 | 0.309 |
| ROUGE-1 f | 0.427 | 0.404 | 0.439 |
| ROUGE-2 f | 0.169 | 0.157 | 0.171 |
| ROUGE-L f | 0.386 | 0.361 | 0.401 |

In the summarization scenarios evaluated on the COCO dataset (Table. 6.1), **Scenario 1** demonstrated a slight edge over **Scenario 2** in terms of average **BLEU, ROUGE-1 f, ROUGE-2 f and ROUGE-l f** scores, of **0.243, 0.427, 0.169 and 0.386** respectively. This suggests that **Scenario 1**'s comprehensive approach, utilizing the `DistilBART` model, is more proficient in capturing detailed narrative structures. Conversely, **Scenario 2**, which employs segment-based summarization, shows very slight edge with **METEOR** score of **0.336** against **Scenario 1**'s **0.334**.

The **Intercomparison** phase, aiming to select the best captions based on the highest cosine similarity with original annotations, consistently showed superior results across BLEU and ROUGE scores. Improvements in **BLEU** score to **0.281** along with **ROUGE-1 f, ROUGE-2 f and ROUGE-l f** score to **0.439,0.171 and 0.401** respectively showcasing a superior grasp of both the syntactic and narrative elements of the captions compared to the initial scenarios.

**Table 6.2:** Evaluation of Summarization Scenarios on FLICKR Dataset.(1100 images)

| **Metric(Average Scores)** | **Scenario 1** | **Scenario 2** | **Intercomparison** |
|---|---|---|---|
| BLEU | 0.152 | 0.156 | 0.158 |
| METEOR | 0.298 | 0.287 | 0.289 |
| ROUGE-1 f | 0.326 | 0.323 | 0.324 |
| ROUGE-2 f | 0.094 | 0.094 | 0.094 |
| ROUGE-L f | 0.285 | 0.282 | 0.282 |

In the summarization evaluations on the FLICKR dataset, the results showcased some nuanced differences between the two scenarios. **Scenario 1** scored more as compared to **Scenario 2** in **METEOR, ROUGE-1 f and ROUGE-l f** scores with a recorded **0.298, 0.326 and 0.285** compared to **0.287, 0.323 and 0.282** respectively in **Scenario 2**. This indicates that **Scenario 1**, which involves comprehensive summarization using the `DistilBART` model, might be slightly better at maintaining semantic integrity with the original annotations.

However, **Scenario 2**, focusing on segment-based summarization, recorded a slightly higher **BLEU** score of **0.156** against **0.152** in **Scenario 1**, suggesting it could be slightly more effective at capturing syntactic structures relevant to the evaluated captions.

The **Intercomparison** results, which aggregate the best outcomes from both scenarios based on their alignment with original human annotations, indicate a general improvement across all metrics. Notably, **BLEU** increased to **0.158**, showing that while both scenarios contribute uniquely to caption quality, the selection of optimally aligned captions enhances overall performance, aligning closely with human perceptions of relevance and coherence.

In this study, the detailed exploration of scenarios 1 and 2 becomes a solid based for answering and exploring the answers to **third research question Q3** regarding the integration of outputs from multiple captioning models. These scenarios demonstrate that advanced summarization techniques, specifically the comprehensive and segment-based approaches, can indeed enhance the semantic depth of generated captions.

Despite the success of these approaches in enhancing caption alignment with human-generated annotations and increasing performance metrics, **this strategy was not directly incorporated into the final approach. Instead, the insights gained from comparing Bayesian metrics such as likelihood and posterior probabilities were instrumental in refining the overall approach by validating the performance of BLIP and GPT-2 on datasets as seen from table 5.7 and 5.8**. The final methodology incorporated an ANN-driven model combined with weighted summarization, optimized based on the findings from these analytical scenarios. This highlights the adaptive nature of research in automated image captioning, where findings from one phase of experimentation inform the strategic decisions in subsequent developments, ensuring the adoption of the most effective captioning techniques.

## 6.3 Comparative Insights from Weighted Summarization and Bayesian Analysis

In the last section, we discussed how the two scenarios helped to contribute to the final approach adopted. Now we discuss how weighted summarization techniques have yielded distinct outcomes for BLIP and GPT-2 on the datasets used and how they addressed our second and third research questions.

### 6.3.1 COCO Dataset: The Advantage of GPT-2

For the COCO dataset, the weighted summarization approach highlighted GPT-2's enhanced performance when assigned more weight, reflecting its capability to generate contextually relevant and semantically rich captions for the dataset's diverse and complex images. According to the table, when GPT-2 was weighted more heavily, it showed a slight advantage across all metrics, achieving a **BLEU score of 0.270**, **METEOR score of 0.320**, **ROUGE-1 F score of 0.426**, **ROUGE-2 F score of 0.164**, and **ROUGE-L F score of 0.386** as shown in

Table 5.9. These results surpass those observed when BLIP was given more weight, underscoring GPT-2's refined ability to handle syntactic complexity and semantic depth effectively.

The selective word integration method, where a few words from BLIP were combined with captions from the more heavily weighted GPT-2, found that incorporating 3 words from BLIP resulted in the highest scores, with a **BLEU of 0.278** referring from Table 5.10 as compared to all the tables 5.9,5.10 and 5.11, once again proving GPT-2 effectiveness over BLIP. Additionally, the Bayesian analysis, focusing on prior, likelihood, and posterior probabilities, reinforced GPT-2's superior performance, confirming its higher likelihood of producing better quality captions as evidenced by its performance metrics, as shown in Table 5.7.

### 6.3.2 Flickr Dataset: The Superiority of BLIP

Conversely, the Flickr dataset revealed BLIP's strengths when it was assigned more weight in the summarization process. This outcome highlights BLIP's capability to produce more descriptive and accurate captions for Flickr's everyday scenes and activities. The difference in performance between the two datasets illustrates the variability in model effectiveness based on the content and context of the dataset being analyzed. Bayesian analysis for the Flickr dataset similarly supported BLIP's superior performance, aligning with the empirical data from the weighted summarization approach.

In the context of the Flickr dataset, interestingly with comprehensive weighting, GPT-2's performance when given more weight had a modest improvement in performance with **BLEU, METEOR, ROUGE-1 F, ROUGE-2 F, and ROUGE-L F scores of 0.149**, **0.305**, **0.365**, **0.105**, and **0.330** respectively, as shown in Table 5.12.

In the selective word integration, conversely, integrating 5 words from GPT-2 resulted in the highest **BLEU score of 0.175** and **ROUGE-1 F score of 0.328** as seen from Table 5.14 as compared across all the tables 5.12, 5.13 and 5.14 demonstrating a refined grasp of image content nuances. The Bayesian analysis supported BLIP's superior performance in generating captions that align closely with human annotations, reinforcing the empirical data from the weighted summarization approach as shown in Table 5.8.

## 6.4 Performance of the ANN Model

The preceding section demonstrated how experimenting with varying weights guided the selection of optimal weights for our datasets. As we progress to the concluding discussion, we focus on analyzing the metric scores obtained from implementing our final approach, which integrates ANN predictions with the weighted summarization of the generated captions.

**Employing the ANN model marked a substantial enhancement in performance compared to the state-of-the-art use of BLIP, GPT-2, and PIX2STRUCT models on both the COCO and FLICKR datasets.** The tables below compare evaluation metrics to underscore the improved efficacy of the ANN models, highlighting the ANN model's superior performance.

**Table 6.3:** Average Scores of Evaluation Metrics for Different Models on COCO Dataset(1100 images)

| Model | BLEU | METEOR | ROUGE-1 f | ROUGE-2 f | ROUGE-l f |
|---|---|---|---|---|---|
| BLIP | 0.305 | 0.302 | 0.401 | 0.163 | 0.369 |
| GPT-2 | 0.317 | 0.327 | 0.404 | 0.158 | 0.369 |
| PIX2STRUCT | 0.054 | 0.165 | 0.227 | 0.029 | 0.206 |
| ANN Model+Weighted Summarization | 0.322 | 0.328 | 0.452 | 0.187 | 0.415 |

For the COCO dataset, our integrated approach using the ANN model and weighted summarization demonstrates significant performance enhancements over traditional models. In the provided table, the ANN model with weighted summarization shows superior results: a **BLEU** score of **0.322**, improving by **1.6%** over GPT-2's **0.317**; a **METEOR** score of **0.328**, marginally higher by **0.3%** than GPT-2's **0.327**; a **ROUGE-1 f** score of **0.452**, marking an **11.9%** increase from GPT-2's **0.404**; a **ROUGE-2 f** score of **0.187**, up by **14.7%** from BLIP's **0.163**; and a **ROUGE-l f** score of **0.415**, a substantial **12.5%** improvement from **0.369**. These results also illustrated in the above column chart 6.1, validate the ANN model combined with weighted summarization as a potent enhancement to automated image captioning.

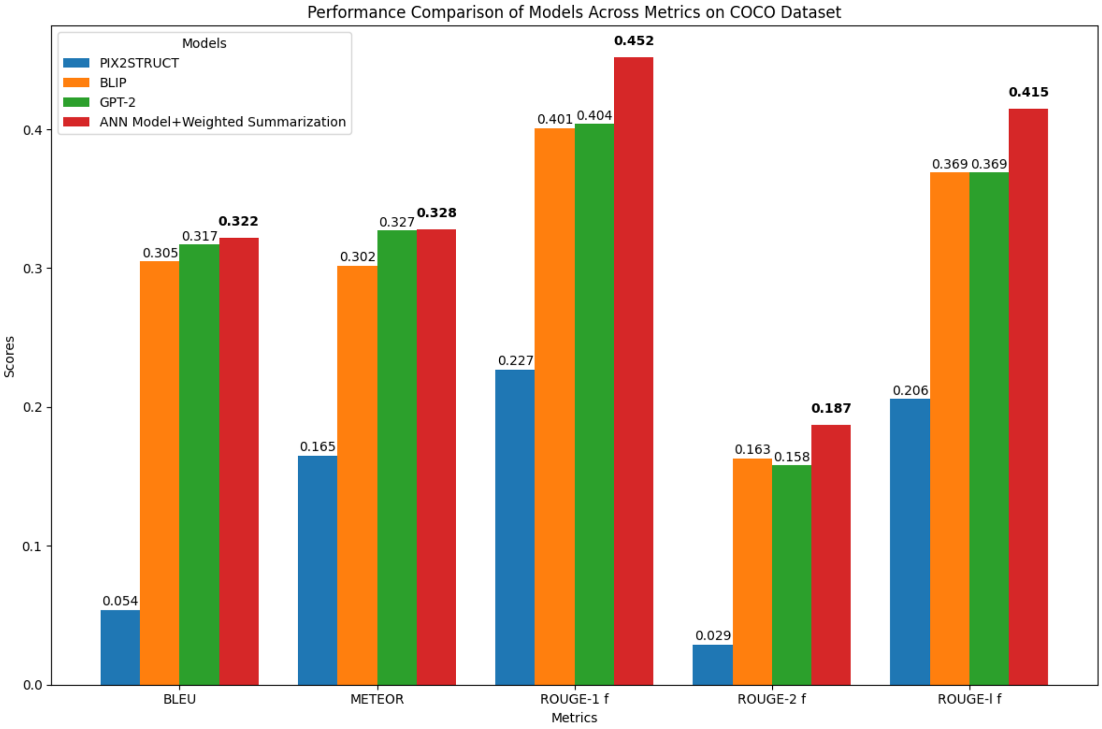


**Figure 6.1:** COCO: Column chart showing performance comparison of models across metrics (1100 images)

**Table 6.4:** Average Scores of Evaluation Metrics for Different Models on FLICKR Dataset(1100 images)

| **Model** | **BLEU** | **METEOR** | **ROUGE-1 f** | **ROUGE-2 f** | **ROUGE-l f** |
|---|---|---|---|---|---|
| BLIP | 0.177 | 0.292 | 0.321 | 0.097 | 0.282 |
| GPT-2 | 0.149 | 0.231 | 0.286 | 0.076 | 0.263 |
| PIX2STRUCT | 0.042 | 0.154 | 0.215 | 0.027 | 0.198 |
| ANN Model+Weighted Summarization | 0.181 | 0.300 | 0.348 | 0.107 | 0.311 |

For the Flickr dataset, the combined approach of the ANN model and weighted summarization outperforms established models, highlighting its efficacy in handling diverse image contexts. The ANN model with weighted summarization scores highest across all metrics: **BLEU** score at **0.181**, showing a **2.3%** increase over

BLIP's **0.177**; **METEOR** score of **0.300**, significantly higher by **2.7%** than BLIP's **0.292**; **ROUGE-1 f** score of **0.348**, marking an **8.4%** improvement from BLIP's **0.321**; **ROUGE-2 f** score of **0.107**, which is **10.3%** higher than BLIP's **0.097**; and **ROUGE-l f** score of **0.311**, up by **10.3%** from BLIP's **0.282**. These substantial performance enhancements are visually represented in the column chart 6.2, effectively showcasing the advanced capabilities of our ANN-driven summarization method.

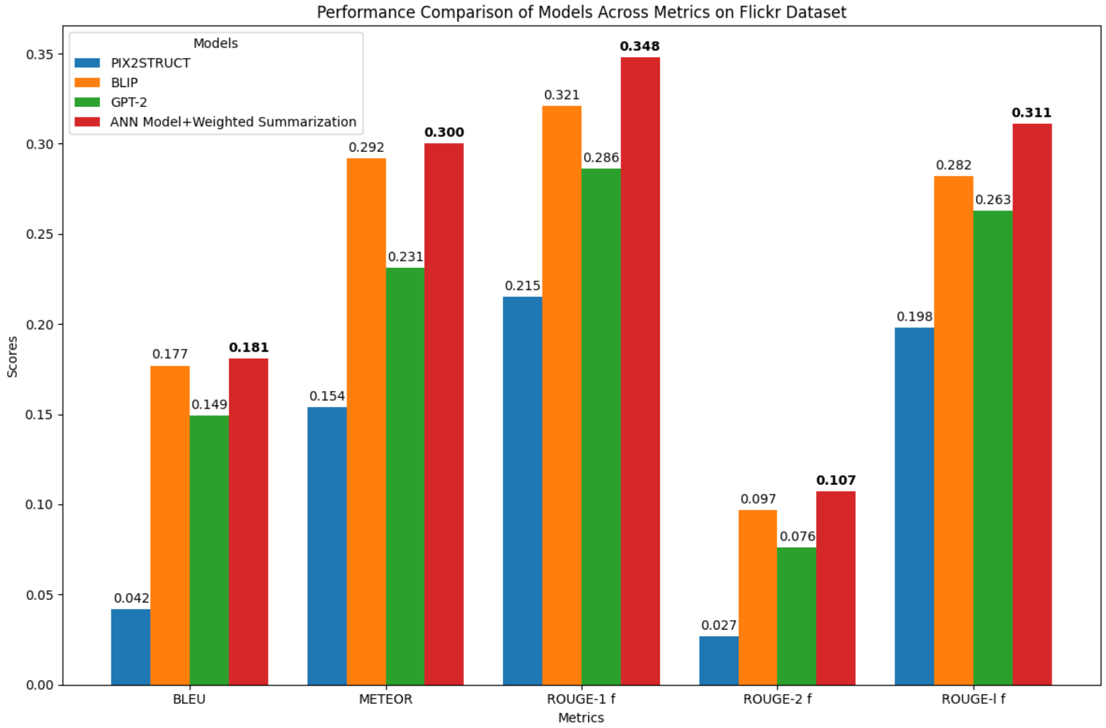


**Figure 6.2:** FLICKR: Column chart showing performance comparison of models across metrics (1100 images)

***These tables elucidate that the ANN model not only enhances the BLEU, METEOR, and ROUGE scores significantly but also presents a new benchmark in the domain of automated image captioning.*** This achievement underscores the ANN model's capacity to predict the most suitable caption generator, resulting in captions that are more aligned with human annotations and thus, more natural and contextually accurate.

*Having thoroughly examined the results and their implications in this chapter, we now move to the next chapter, 'Conclusion.' This final chapter synthesizes the key* *findings of the thesis, outlines the limitations encountered during our research, and suggests directions for future work.*

# CHAPTER 7 CONCLUSION

## 7.1 Conclusion

This research embarked on an ambitious journey to explore and enhance automated image captioning through methodical experiments across the COCO and FLICKR datasets. By harnessing the capabilities of advanced machine learning models such as BLIP, GPT-2, and PIX2STRUCT, along with innovative summarization techniques, this study aimed to narrow the gap between machine-generated captions and human-level annotations.

Our research successfully addressed **all posed research questions Q1, Q2, Q3 and Q4** through thorough experimentation, analysis, and discussion, affirmatively responding to the initial problem statement. We began by exploring the evolution of image captioning techniques and assessed how state-of-the-art models like BLIP and GPT-2 could significantly contribute to our research. Also, the evaluation of state-of-the-art models on diverse image categories validates the approach of using a weighted summarization strategy based on the model's performance across different categories. The integration of an ANN model with weighted summarization proved crucial, optimizing caption generation by leveraging the strengths of the most effective model for each specific context.

Through the analysis of different datasets, it was observed that GPT-2 generally performed better on the COCO dataset while BLIP showed stronger results on FLICKR. This led to tailored weight assignments in our summarization process—favoring GPT-2 for COCO and BLIP for FLICKR. When trained on image features, the ANN was able to accurately predict the most suitable caption generator, which, when combined with our weighted summarization approach, resulted in superior evaluation metrics compared to those achieved by the state-of-the-art models.

A significant takeaway from our research is the demonstration that high-quality image captioning can be achieved without the need of retraining. Despite using only 1100 images, our approach yielded better evaluation metrics, such as **METEOR score of 0.328** on the COCO dataset, than those obtained by models that requires extensive training on vast datasets which can be seen from the table 2.1 achieving a maximum **METEOR score of 0.29 by OSCAR, I-Tuning_Large and CaMEL**. This finding underscores the efficiency of our method which requires only a pre-trained ANN model and its potential to reduce the computational costs associated with advanced image captioning technologies.

Interestingly, when calculated, the **total number of parameters for our approach**—which includes transformers (BLIP, GPT-2), a summarizer (DistilBART), and an ANN model—amounts to approximately **791 million**. In comparison, a recent study by Ramos et al. reported that the highest performing model, **LEMON**, utilizes **675 million** parameters but achieves a **lower METEOR score of 0.308[33] compared to ours 0.328**. It is important to note that within our model, the majority of parameters are frozen: 247 million for BLIP, 239 million for GPT-2, and 305 million for DistilBART. The ANN model, crucial for optimizing our captioning process, requires only **0.3 million trainable parameters**, significantly fewer than SMALLCAP's 7 million trainable parameters and a total of 218 million parameters, which includes the frozen CLIP encoder and GPT-2 decoder as described by Ramos et al.[33] (more details for comparison are provided in **Appendix. E.2**.) This highlights that while our approach utilizes a higher total number of parameters, the portion that requires active training and thus computational resources during operation is substantially lower. Additionally, our approach is computationally efficient because it requires very low trainable parameters. The time required to generate captions from an image is **6.87 seconds for BLIP** and **3.84 seconds for GPT-2**, while the summarization process takes **13.36 seconds**. Thus, the total time for generating and summarizing captions is approximately **24.07 seconds**. While this may seem significant, it is important to consider that in many applications of image captioning, **real-time processing is not a critical requirement**. For example, data warehouses analyze archived images to extract detailed insights for decision-making and historical data analysis. Social media platforms periodically process images to ensure content compliance, focusing on accuracy over speed. In e-commerce, detailed captions enhance product image searchability and user experience. Educational content creators and museums use images to enhance learning materials and document artifacts, respectively, where detailed and accurate descriptions are essential. Given these contexts, **the extended processing time is a reasonable trade-off for the enhanced quality and depth of information provided, underscoring the suitability of our method where precision and factual accuracy are prioritized over immediacy**. Therefore, despite the higher number of total parameters, our method's ability to achieve better quality captions with minimal active learning justifies the additional computational investment, especially for applications where caption quality is critical. This configuration not only underscores the efficiency of our approach—which does not require retraining and primarily utilizes a pre-trained ANN model—but also demonstrates significant potential to reduce the computational costs associated with advanced image captioning technologies.

Finally, addressing our second part of the **last research question Q4.**, for COCO dataset we notice that the **ROUGE-1 f** score improved significantly to **0.452**, an **11.9%** increase from GPT-2's **0.404**, while the **ROUGE-2 f** score rose to **0.187**, up by **14.7%** from BLIP's **0.163**, and the **ROUGE-L f** score improved by **12.5%** to **0.415** from **0.369**. These advancements, visually underscored in the column chart 6.1, particularly, point to our approach's enhanced ability to capture not just the keywords but the overall structure and fluency of the human reference captions. This indicates that our model excels in maintaining narrative continuity and detail, which are crucial for producing contextually rich and coherent captions. The pronounced improvements in ROUGE metrics suggest that our methodology is particularly effective in enhancing the comprehensiveness and detail-oriented aspects of caption generation, which are vital for applications requiring high fidelity to the original content's context and subtleties.

Similarly,for the FLICKR dataset,the **ROUGE-1 f** score sees a notable **8.4%** improvement to **0.348** from BLIP's **0.321**, the **ROUGE-2 f** score rises by **10.3%** to **0.107** from **0.097**, and the **ROUGE-L f** score enhances by **10.3%** to **0.311** from **0.282**. These marked improvements in performance metrics are visually delineated in the column chart 6.2, distinctly highlighting the proficiency of our ANN-driven weighted summarization strategy in producing captions that more accurately reflect the nuances and context of the original imagery.

In conclusion, this study not only enhances the understanding and application of image captioning techniques but also sets the stage for future research to further refine and innovate in this rapidly evolving field.

## 7.2 Contributions of the Research

This thesis has made several significant contributions to the field of automated image captioning:

- **Evaluation Across Categories and Datasets:** Provided a comprehensive comparison of the performance of BLIP, GPT-2 models across the COCO and FLICKR datasets and image categories, offering insights into their strengths and limitations in various contexts along with their performance on different image categories.
- **Summarization Strategies:** Explored two novel summarization scenarios, contributing to the understanding of how different approaches to summarization can enhance the quality and relevance of generated captions.
- **Weighted Summarization:** Introduced and evaluated the effectiveness of weighted summarization techniques, demonstrating their potential to leverage the comparative advantages of different models based on dataset-specific characteristics.
- **Innovative Use of ANN Models:** Pioneered the use of ANN models to predict the most effective caption generator for a given image, setting a new benchmark in the quality of automated image captioning.
- **Foundation for Future Research:** Established a methodological framework and baseline that can inform and inspire future research in the domain, particularly in advancing captioning techniques and model optimization.

## 7.3 Limitations and Future Work

We have framed the Limitations and Future work based on the results and groundwork done in this thesis discussed more in the Appendix. F.

### 7.3.1 Limitations

This research encountered several limitations that should be considered when interpreting the findings:

- **Dataset Scope:** While the COCO and FLICKR datasets provide a wide range of visual content, they do not encompass all possible domains or types of images, potentially limiting the generalizability of the results.
- **Computational Resources:** Limited computational power restricted the research to a subset of 1100 images from the datasets, which may affect the robustness and scalability of the proposed methods.

### 7.3.2 Future Work

Future research can be expanded in several directions:

- **Extending Dataset Coverage:** Testing the proposed methods on a broader array of datasets, including those that cover more specialized or niche content, to enhance the versatility and applicability of captioning models.
- **Scaling Computational Resources:** Leveraging more powerful computational resources to apply the developed techniques to the entirety of the COCO and FLICKR datasets, thereby validating and potentially enhancing the robustness of the findings.
- **Advanced Model Development:** Further refining and optimizing ANN models and summarization techniques to improve accuracy, context-awareness, and semantic richness of the generated captions.
- **Exploring New Summarization Techniques:** Investigating additional summarization strategies like including POS(Part-of-speech tagging) instead of just taking some words from the start of the caption that could offer further improvements in caption quality, particularly in challenging or ambiguous visual contexts.
- **Contributing to Academic Discourse:** Publishing a research paper based on the methods and findings of this thesis, contributing valuable insights and methodologies to the field of automated image captioning.

# BIBLIOGRAPHY

[1] D. M. Blei and M. I. Jordan, 'Modeling annotated data,' ser. SIGIR '03, Toronto, Canada: Association for Computing Machinery, 2003, pp. 127–134, ISBN: 1581136463. DOI: `10.1145/860435.860460`. [Online]. Available: `https://doi.org/10.1145/860435.860460`.

[2] M. Hodosh, P. Young and J. Hockenmaier, 'Framing image description as a ranking task: Data, models and evaluation metrics,' *Journal of Artificial Intelligence Research*, vol. 47, pp. 853–899, 2013.

[3] O. Vinyals, A. Toshev, S. Bengio and D. Erhan, 'Show and tell: A neural image caption generator,' *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3156–3164, 2014. [Online]. Available: `https://api.semanticscholar.org/CorpusID:1169492`.

[4] X. Chen, H. Fang, T.-Y. Lin *et al.*, 'Microsoft coco captions: Data collection and evaluation server,' *ArXiv*, vol. abs/1504.00325, 2015. [Online]. Available: `https://api.semanticscholar.org/CorpusID:2210455`.

[5] B. Plummer, L. Wang, C. Cervantes, J. Caicedo, J. Hockenmaier and S. Lazebnik, 'Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models,' *International Journal of Computer Vision*, vol. 123, 2017. DOI: `10.1007/s11263-016-0965-7`.

[6] K. Papineni, S. Roukos, T. Ward and W.-J. Zhu, 'Bleu: A method for automatic evaluation of machine translation,' in *Annual Meeting of the Association for Computational Linguistics*, 2002, pp. 311–318. [Online]. Available: `https://api.semanticscholar.org/CorpusID:11080756`.

[7] C.-Y. Lin, 'ROUGE: A package for automatic evaluation of summaries,' in *Text Summarization Branches Out*, Barcelona, Spain: Association for Computational Linguistics, Jul. 2004, pp. 74–81. [Online]. Available: https://aclanthology.org/W04-1013.

[8] S. Banerjee and A. Lavie, 'Meteor: An automatic metric for mt evaluation with improved correlation with human judgments,' in *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, Available: https://www.aclweb.org/anthology/W05-0909/, Association for Computational Linguistics, 2005, pp. 65–72.

[9] Y. LeCun, Y. Bengio and G. Hinton, 'Deep learning,' *Nature*, vol. 521, no. 7553, pp. 436–444, 1st May 2015, ISSN: 1476-4687. DOI: 10.1038/nature14539. [Online]. Available: https://doi.org/10.1038/nature14539.

[10] S. Hochreiter and J. Schmidhuber, 'Long short-term memory,' *Neural Computation*, vol. 9, pp. 1735–1780, 1997. [Online]. Available: https://api.semanticscholar.org/CorpusID:1915014.

[11] T.-Y. Lin, M. Maire, S. J. Belongie *et al.*, 'Microsoft coco: Common objects in context,' in *European Conference on Computer Vision*, 2014. [Online]. Available: https://api.semanticscholar.org/CorpusID:14113767.

[12] A. Krizhevsky, I. Sutskever and G. E. Hinton, 'Imagenet classification with deep convolutional neural networks,' in *Advances in neural information processing systems*, 2012, pp. 1097–1105. [Online]. Available: http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf.

[13] A. Vaswani, N. Shazeer, N. Parmar *et al.*, 'Attention is all you need,' in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio *et al.*, Eds., vol. 30, Curran Associates, Inc., 2017. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

[14] A. Karpathy and L. Fei-Fei, 'Deep visual-semantic alignments for generating image descriptions,' in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 3128–3137. [Online]. Available: https://cs.stanford.edu/people/karpathy/cvpr2015.pdf.

[15] J. Johnson, A. Karpathy and L. Fei-Fei, 'Densecap: Fully convolutional localization networks for dense captioning,' *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4565–4574, 2015. [Online]. Available: https://api.semanticscholar.org/CorpusID:14521054.

[16] A. Jamil, Saif-Ur-Rehman, K. Mahmood *et al.*, 'Deep learning approaches for image captioning: Opportunities, challenges and future potential,' *IEEE Access*, vol. PP, pp. 1–1, Jan. 2024. DOI: 10.1109/ACCESS.2024.3365528.

[17] K. Xu, J. Ba, R. Kiros *et al.*, 'Show, attend and tell: Neural image caption generation with visual attention,' in *International Conference on Machine Learning*, 2015. [Online]. Available: https://api.semanticscholar.org/CorpusID:1055111.

[18] P. Anderson, X. He, C. Buehler *et al.*, 'Bottom-up and top-down attention for image captioning and visual question answering,' *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6077–6086, 2017. [Online]. Available: https://api.semanticscholar.org/CorpusID:3753452.

[19] A. Mathews, L. Xie and X. He, 'Senticap: Generating image descriptions with sentiments,' *ArXiv*, vol. abs/1510.01431, 2015. [Online]. Available: https://api.semanticscholar.org/CorpusID:2875390.

[20] K. He, X. Zhang, S. Ren and J. Sun, 'Deep residual learning for image recognition,' *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2015. [Online]. Available: https://api.semanticscholar.org/CorpusID:206594692.

[21] O. Vinyals, A. Toshev, S. Bengio and D. Erhan, 'Show and tell: Lessons learned from the 2015 mscoco image captioning challenge,' *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 4, pp. 652–663, 2017, ISSN: 0162-8828. DOI: 10.1109/TPAMI.2016.2587640. [Online]. Available: https://doi.org/10.1109/TPAMI.2016.2587640.

[22] M. Cornia, M. Stefanini, L. Baraldi and R. Cucchiara, 'Meshed-memory transformer for image captioning,' *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10 575–10 584, 2019. [Online]. Available: https://api.semanticscholar.org/CorpusID:219635470.

[23] Q. You, H. Jin, Z. Wang, C. Fang and J. Luo, 'Image captioning with semantic attention,' *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4651–4659, 2016. [Online]. Available: https://api.semanticscholar.org/CorpusID:3120635.

[24] S. Herdade, A. Kappeler, K. Boakye and J. Soares, 'Image captioning: Transforming objects into words,' in *Neural Information Processing Systems*, 2019. [Online]. Available: https://api.semanticscholar.org/CorpusID:189898359.

[25] Y. Li, L. Kaiser, S. Bengio and S. Si, 'Area attention,' in *International Conference on Machine Learning*, 2018. [Online]. Available: https://api.semanticscholar.org/CorpusID:53085166.

[26] Z. Fei, 'Attention-aligned transformer for image captioning,' *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 1, pp. 607–615, 2022. DOI: 10.1609/aaai.v36i1.19940. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/19940.

[27] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov and S. Zagoruyko, 'End-to-end object detection with transformers,' *ArXiv*, vol. abs/2005.12872, 2020. [Online]. Available: https://api.semanticscholar.org/CorpusID:218889832.

[28] J. Lei, L. Li, L. Zhou *et al.*, 'Less is more: Clipbert for video-and-language learning via sparse sampling,' *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7327–7337, 2021. [Online]. Available: https://api.semanticscholar.org/CorpusID:231880022.

[29] X. Li, X. Yin, C. Li *et al.*, 'Oscar: Object-semantics aligned pre-training for vision-language tasks,' *ECCV*, 2020. [Online]. Available: https://www.microsoft.com/en-us/research/publication/oscar-object-semantics-aligned-pre-training-for-vision-language-tasks/.

[30] X. Hu, X. Yin, K. Lin *et al.*, 'Vivo: Visual vocabulary pre-training for novel object captioning,' in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, 2021, pp. 1575–1583. [Online]. Available: https://doi.org/10.1609/aaai.v35i2.16249.

[31] Z. Luo *et al.*, 'I-tuning: Tuning language models with image for caption generation,' *DeepAI*, 2022. [Online]. Available: https://deepai.org/publication/i-tuning-tuning-language-models-with-image-for-caption-generation.

[32] Y. Wang, Q. Yao, J. T.-Y. Kwok and L. M.-s. Ni, 'Generalizing from a few examples,' *ACM Computing Surveys (CSUR)*, vol. 53, pp. 1–34, 2019. [Online]. Available: https://api.semanticscholar.org/CorpusID:226931458.

[33] R. P. Ramos, B. Martins, D. Elliott and Y. Kementchedjhieva, 'Smallcap: Lightweight image captioning prompted with retrieval augmentation,' *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2840–2849, 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:252668790.

[34] R. Mokady, S. Benaim and L. Wolf, 'Clipcap: Clip prefix for image captioning,' *arXiv preprint arXiv:2111.09734*, 2021. [Online]. Available: https://arxiv.org/abs/2111.09734.

[35] Gaurav and P. Mathur, 'A survey on various deep learning models for automatic image captioning,' *Journal of Physics: Conference Series*, vol. 1950, no. 1, p. 012 045, 2021, International Conference on Mechatronics and Artificial Intelligence (ICMAI) 2021, 27 February 2021, Gurgaon, India. DOI: 10.1088/1742-6596/1950/1/012045.

[36] A. Nenkova and K. McKeown, *Automatic Summarization* (Foundations and trends in information retrieval). Now Publishers, 2011, ISBN: 9781601984708. [Online]. Available: https://books.google.no/books?id=IAExe8b_HMoC.

[37] A. See, P. J. Liu and C. D. Manning, 'Get to the point: Summarization with pointer-generator networks,' in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, R. Barzilay and M.-Y. Kan, Eds., Vancouver, Canada: Association for Computational Linguistics, Jul. 2017, pp. 1073–1083. DOI: 10.18653/v1/P17-1099. [Online]. Available: https://aclanthology.org/P17-1099.

[38] N. Munot and S. Govilkar, 'Comparative study of text summarization methods,' *International Journal of Computer Applications*, vol. 102, pp. 33–37, Sep. 2014. DOI: 10.5120/17870-8810.

[39] G. Erkan and D. R. Radev, 'Lexrank: Graph-based lexical centrality as salience in text summarization,' *ArXiv*, vol. abs/1109.2128, 2004. [Online]. Available: https://api.semanticscholar.org/CorpusID:506350.

[40] M. Yasunaga, R. Zhang, K. Meelu, A. Pareek, K. P. Srinivasan and D. R. Radev, 'Graph-based neural multi-document summarization,' *ArXiv*, vol. abs/1706.06681, 2017. [Online]. Available: https://api.semanticscholar.org/CorpusID:6532096.

[41] M. Zhong, P. Liu, Y. Chen, D. Wang, X. Qiu and X.-J. Huang, 'Extractive summarization as text matching,' pp. 6197–6208, 2020. DOI: 10.18653/v1/2020.acl-main.552. [Online]. Available: https://aclanthology.org/2020.acl-main.552.

[42] S. Narayan, S. B. Cohen and M. Lapata, 'Ranking sentences for extractive summarization with reinforcement learning,' in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, Association for Computational Linguistics, 2018, pp. 1747–1759. DOI: 10.18653/v1/N18-1158. [Online]. Available: https://aclanthology.org/N18-1158.

[43] J. Devlin, M.-W. Chang, K. Lee and K. Toutanova, 'Bert: Pre-training of deep bidirectional transformers for language understanding,' in *Proceedings of NAACL-HLT 2019*, Association for Computational Linguistics, 2019,

pp. 4171–4186. DOI: 10.18653/v1/N19-1423. [Online]. Available: https://aclanthology.org/N19-1423.

[44] T. B. Brown *et al.*, ‘Language models are few-shot learners,’ vol. 33, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan and H. Lin, Eds., pp. 1877–1901, 2020. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

[45] C. Raffel, N. M. Shazeer, A. Roberts *et al.*, ‘Exploring the limits of transfer learning with a unified text-to-text transformer,’ *J. Mach. Learn. Res.*, vol. 21, 140:1–140:67, 2019. [Online]. Available: https://api.semanticscholar.org/CorpusID:204838007.

[46] M. Lewis, Y. Liu, N. Goyal *et al.*, ‘Bart: Denoising sequence-to-sequence pretraining for natural language generation, translation, and comprehension,’ in *Annual Meeting of the Association for Computational Linguistics*, 2019. [Online]. Available: https://api.semanticscholar.org/CorpusID:204960716.

[47] J. Zhang, Y. Zhao, M. Saleh and P. J. Liu, ‘Pegasus: Pre-training with extracted gap-sentences for abstractive summarization,’ *ArXiv*, vol. abs/1912.08777, 2019. [Online]. Available: https://api.semanticscholar.org/CorpusID:209405420.

[48] N. Giarelis, C. Mastrokostas and N. Karacapilidis, ‘Abstractive vs. extractive summarization: An experimental review,’ *Applied Sciences*, vol. 13, no. 13, 2023, ISSN: 2076-3417. DOI: 10.3390/app13137620. [Online]. Available: https://www.mdpi.com/2076-3417/13/13/7620.

[49] W.-T. Hsu, C.-K. Lin, M.-Y. Lee, K. Min, J. Tang and M. Sun, ‘A unified model for extractive and abstractive summarization using inconsistency loss,’ in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Melbourne, Australia: Association for Computational Linguistics, Jul. 2018, pp. 132–141. DOI: 10.18653/v1/P18-1013. [Online]. Available: https://aclanthology.org/P18-1013.

[50] A. See, P. J. Liu and C. D. Manning, ‘Get to the point: Summarization with pointer-generator networks,’ in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Vancouver, Canada: Association for Computational Linguistics, Jul. 2017, pp. 1073–1083. DOI: 10.18653/v1/P17-1099. [Online]. Available: https://aclanthology.org/P17-1099.

[51] I. Goodfellow, Y. Bengio and A. Courville, *Deep Learning.* MIT Press, 2016, http://www.deeplearningbook.org.

[52] Y. LeCun, Y. Bengio and G. Hinton, ‘Deep learning,’ *Nature*, vol. 521, no. 7553, pp. 436–444, 1st May 2015, ISSN: 1476-4687. DOI: 10.1038/nature14539. [Online]. Available: https://doi.org/10.1038/nature14539.

[53] M. Minsky and S. Papert, *Perceptrons: An Introduction to Computational Geometry.* MIT Press, 1969.

[54] J. Han and C. Moraga, ‘The influence of the sigmoid function parameters on the speed of backpropagation learning,’ in *International Work-Conference on Artificial and Natural Neural Networks*, 1995. [Online]. Available: https://api.semanticscholar.org/CorpusID:2828079.

[55] V. Nair and G. E. Hinton, ‘Rectified linear units improve restricted boltzmann machines,’ in *International Conference on Machine Learning*, 2010. [Online]. Available: https://api.semanticscholar.org/CorpusID:15539264.

[56] Y. Lecun, L. Bottou, G. Orr and K.-R. Müller, ‘Efficient backprop,’ Aug. 2000.

[57] A. L. Maas, ‘Rectifier nonlinearities improve neural network acoustic models,’ 2013. [Online]. Available: https://api.semanticscholar.org/CorpusID:16489696.

[58] J. S. Bridle, ‘Probabilistic interpretation of feedforward classification network outputs, with relationships to statistical pattern recognition,’ *Neurocomputing: Algorithms, Architectures and Applications*, 1990.

[59] D. E. Rumelhart, G. E. Hinton and R. J. Williams, ‘Learning representations by back-propagating errors,’ *Nature*, vol. 323, pp. 533–536, 1986. [Online]. Available: https://api.semanticscholar.org/CorpusID:205001834.

[60] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever and R. Salakhutdinov, ‘Dropout: A simple way to prevent neural networks from overfitting,’ *Journal of Machine Learning Research*, vol. 15, no. 56, pp. 1929–1958, 2014. [Online]. Available: http://jmlr.org/papers/v15/srivastava14a.html.

[61] S. Ioffe and C. Szegedy, ‘Batch normalization: Accelerating deep network training by reducing internal covariate shift,’ *ArXiv*, vol. abs/1502.03167, 2015. [Online]. Available: https://api.semanticscholar.org/CorpusID:5808102.

[62] D. P. Kingma and J. Ba, ‘Adam: A method for stochastic optimization,’ *CoRR*, vol. abs/1412.6980, 2014. [Online]. Available: https://api.semanticscholar.org/CorpusID:6628106.

[63] K. Cho, B. van Merriënboer, C. Gulcehre *et al.*, 'Learning phrase representations using RNN encoder–decoder for statistical machine translation,' in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, A. Moschitti, B. Pang and W. Daelemans, Eds., Doha, Qatar: Association for Computational Linguistics, 2014, pp. 1724–1734. DOI: `10.3115/v1/D14-1179`. [Online]. Available: `https://aclanthology.org/D14-1179`.

[64] J. Li, D. Li, C. Xiong and S. C. H. Hoi, 'Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation,' in *International Conference on Machine Learning*, 2022. [Online]. Available: `https://api.semanticscholar.org/CorpusID:246411402`.

[65] A. Radford *et al.*, 'Language models are unsupervised multitask learners,' *OpenAI Blog*, 2019. [Online]. Available: `https://openai.com/blog/better-language-models/`.

[66] A. Dosovitskiy, L. Beyer, A. Kolesnikov *et al.*, 'An image is worth 16x16 words: Transformers for image recognition at scale,' *ArXiv*, vol. abs/2010.11929, 2020. [Online]. Available: `https://api.semanticscholar.org/CorpusID:225039882`.

[67] K. Lee, M. Joshi, I. Turc *et al.*, 'Pix2struct: Screenshot parsing as pretraining for visual language understanding,' *arXiv*, vol. abs/2210.03347, 2022. [Online]. Available: `https://api.semanticscholar.org/CorpusID:252762394`.

[68] C. Manning, P. Raghavan and H. Schütze, *Introduction to Information Retrieval.* Cambridge University Press, 2008, ISBN: 9781139472104. [Online]. Available: `https://books.google.no/books?id=t1PoSh4uwVcC`.

[69] M. Lewis, Y. Liu, N. Goyal *et al.*, *Leveraging pre-trained checkpoints for sequence generation tasks*, `https://huggingface.co/sshleifer/distilbart-cnn-12-6`, 2020. arXiv: `1907.12461 [cs.CL]`.

[70] N. Iskender, T. Polzehl and S. Möller, 'Reliability of human evaluation for text summarization: Lessons learned and challenges ahead,' in *Proceedings of the Workshop on Human Evaluation of NLP Systems (HumEval)*, A. Belz, S. Agarwal, Y. Graham, E. Reiter and A. Shimorina, Eds., Online: Association for Computational Linguistics, Apr. 2021, pp. 86–96. [Online]. Available: `https://aclanthology.org/2021.humeval-1.10`.

[71] E. Lloret and M. Palomar, 'Text summarisation in progress: A literature review,' *Artificial Intelligence Review*, vol. 37, pp. 1–41, 2011. [Online]. Available: `https://api.semanticscholar.org/CorpusID:254232944`.

[72] J. Maynez, S. Narayan, B. Bohnet and R. McDonald, 'On faithfulness and factuality in abstractive summarization,' in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, D. Jurafsky, J. Chai, N. Schluter and J. Tetreault, Eds., Online: Association for Computational Linguistics, 2020, pp. 1906–1919. DOI: 10.18653/v1/2020.acl-main.173. [Online]. Available: https://aclanthology.org/2020.acl-main.173.

[73] M. Grusky, M. Naaman and Y. Artzi, 'Newsroom: A dataset of 1.3 million summaries with diverse extractive strategies,' in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, M. Walker, H. Ji and A. Stent, Eds., New Orleans, Louisiana: Association for Computational Linguistics, 2018, pp. 708–719. DOI: 10.18653/v1/N18-1065. [Online]. Available: https://aclanthology.org/N18-1065.

[74] M. Bhandari, P. N. Gour, A. Ashfaq, P. Liu and G. Neubig, 'Re-evaluating evaluation in text summarization,' in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Online: Association for Computational Linguistics, 2020, pp. 9347–9359. DOI: 10.18653/v1/2020.emnlp-main.751. [Online]. Available: https://aclanthology.org/2020.emnlp-main.751.

[75] X. Chen, H. Fang, T.-Y. Lin *et al.*, 'Microsoft coco captions: Data collection and evaluation server,' in *arXiv:1504.00325*, 2015.

[76] Y. Gondaliya, P. Kalariya, B. Y. Panchal and A. Nayak, 'A rule-based grammar and spell checking,' *SAMRIDDHI*, vol. 14, no. 1, 2022, Available at SSRN: https://ssrn.com/abstract=4139315.

[77] W. Strunk, *The Elements of Style*. Arcturus Publishing, 2023, ISBN: 9781398833913. [Online]. Available: https://books.google.no/books?id=oq60EAAAQBAJ.

[78] T. Bayes and n. Price, 'An essay towards solving a problem in the doctrine of chances,' *Philosophical Transactions of the Royal Society of London*, vol. 53, pp. 370–418, 1763. DOI: 10.1098/rstl.1763.0053. eprint: https://royalsocietypublishing.org/doi/pdf/10.1098/rstl.1763.0053. [Online]. Available: https://royalsocietypublishing.org/doi/abs/10.1098/rstl.1763.0053.

[79] R. A. Fisher, 'On the mathematical foundations of theoretical statistics,' *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 222, pp. 309–368, 1922. DOI: 10.1098/rsta.1922.0009.

[80] H. Jeffreys, *Theory of Probability* (International series of monographs on physics). Clarendon Press, 1983, ISBN: 9780198531937. [Online]. Available: https://books.google.no/books?id=EbodAQAAMAAJ.

[81] M. Lewis, Y. Liu, N. Goyal *et al.*, 'Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension,' *arXiv preprint arXiv:1910.13461*, 2019. [Online]. Available: https://huggingface.co/facebook/bart-large-mnli.

[82] Z. J. Wang, E. Montoya, D. Munechika, H. Yang, B. Hoover and D. H. Chau, 'Large-scale prompt gallery dataset for text-to-image generative models,' *arXiv:2210.14896 [cs]*, 2022. [Online]. Available: https://arxiv.org/abs/2210.14896.

[83] P. Sharma, N. Ding, S. Goodman and R. Soricut, 'Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning,' in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Melbourne, Australia: Association for Computational Linguistics, 2018, pp. 2556–2565. DOI: 10.18653/v1/P18-1238. [Online]. Available: https://aclanthology.org/P18-1238.

[84] H. Fu, Q. Zhang and G. Qiu, 'Random forest for image annotation,' vol. 7577, Oct. 2012, ISBN: 978-3-642-33782-6. DOI: 10.1007/978-3-642-33783-3_7.

[85] G. Hoxha and F. Melgani, 'A novel svm-based decoder for remote sensing image captioning,' *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–14, 2022. DOI: 10.1109/TGRS.2021.3105004.

[86] Callidior, 'Bert2bert model for text summarization,' 2020. [Online]. Available: https://huggingface.co/Callidior/bert2bert-base-arxiv-titlegen.

[87] Einmalumdiewelt, 'T5-base model for german news article dataset (gnad),' 2020. [Online]. Available: https://huggingface.co/Einmalumdiewelt/T5-Base_GNAD.

[88] G. AI, 'Bigbird-pegasus: Large-scale models for summarization,' 2020. [Online]. Available: https://huggingface.co/google/bigbird-pegasus-large-pubmed.

[89] VietAI, 'Vit5: A pre-trained model for vietnamese text summarization,' 2020. [Online]. Available: https://huggingface.co/VietAI/vit5-large-vietnews-summarization.

# APPENDIX A

# PYTHON MODULES

**Table A.1:** Overview of Python modules and packages utilized in this thesis, along with their functions:

| **Module/Package** | **Purpose** |
|---|---|
| collections | Introduces specialized container data types that serve as alternatives to Python's standard built-in containers. |
| evaluate | Deliver a comprehensive suite of evaluation metrics tailored for appraising the performance of machine learning models. |
| evaluate | A Hugging Face library for model evaluation and comparison. |
| huggingface/evaluate | Supports the assessment of machine learning models through a diverse array of benchmarks and dataset evaluations. |
| huggingface/transformers | Offers a large collection of pre-trained models using the transformer architecture, ideal for various natural language processing (NLP) tasks. |
| io | Provide access to the Python interfaces for managing streams. |
| language_tool_python | A grammar and style checker. |
| matplotlib | A plotting library for NumPy. |

*Continued on next page*

Table A.1 – *Continued from previous page*

| **Module/Package** | **Purpose** |
|---|---|
| nltk | A framework on which Python programs can be developed to handle data in human languages. |
| numpy | supports a variety of mathematical functions to be applied to big, multidimensional arrays and matrices. |
| os | Gives users a portable method to access operating system-dependent features. |
| pandas | Provides structured data operations in Python by acting as a potent toolkit for data analysis and manipulation. |
| PIL (Pillow) | Adds image processing capabilities to your Python interpreter. |
| pickle | Carries out binary protocols to serialize and decode Python object structures. |
| pycocoevalcap | A GitHub repository containing Python tools for caption evaluation, particularly for the COCO dataset. |
| pycocotools | Official COCO dataset utility tools. |
| skimage | A collection of algorithms for image processing in Python. |
| sklearn (scikit-learn) | Machine learning in Python. |
| torch | A fast and highly flexible deep learning research platform. |
| torchvision | A collection of widely used model architectures, datasets, and standard image transformations for computer vision. |
| transformers | Cutting-edge Natural Language Processing with TensorFlow 2.0 and PyTorch. |
| tqdm | A fast, extensible progress bar for loops and or code blocks. |
| urllib | A package for opening and reading URLs. |
| zipfile | Allows the creation, reading, writing, appending, and listing of ZIP files. |

# APPENDIX B

# ADDITIONAL DATASETS CONSIDERED

Although we did explore other datasets for our study, their poor performance in preliminary evaluations led us to omit them from our primary analysis.

## B.1 Diffusion Database

The image-caption pairs in the Diffusion Database come from a variety of internet sources and are frequently less structured and more diversified, which presents more difficulties for caption generation[82].
Hugging Face: `https://huggingface.co/datasets/poloclub/diffusiondb` was the access point for this dataset. Table B.1 summarizes the various models' performance on this dataset.

**Table B.1:** Performance of BLIP, GPT-2, and PIX2STRUCT on the Diffusion Database

| Model | BLEU | METEOR | ROUGE-1 F | ROUGE-2 F | ROUGE-L F |
|---|---|---|---|---|---|
| BLIP | 0.003 | 0.043 | 0.127 | 0.018 | 0.101 |
| Pix2Struct | 0.002 | 0.037 | 0.089 | 0.003 | 0.083 |
| GPT-2 | 0.001 | 0.034 | 0.101 | 0.010 | 0.095 |

## B.2 Google Conceptual Dataset

The Google Conceptual Captions dataset is made up of image-caption pairings that are taken from internet photos linked with alt-text descriptions. Its purpose is to help train algorithms for image caption generation[83].
Similarly, Hugging face `https://huggingface.co/datasets/conceptual_captions` was the access point to this dataset. Table B.2 summarizes the performance of different models on this dataset.

**Table B.2:** Performance of BLIP, GPT-2, and PIX2STRUCT on the Google Conceptual Dataset

| Model | BLEU | METEOR | ROUGE-1 F | ROUGE-2 F | ROUGE-L F |
|---|---|---|---|---|---|
| BLIP | 0.087 | 0.238 | 0.284 | 0.122 | 0.217 |
| GPT-2 | 0.009 | 0.074 | 0.148 | 0.013 | 0.096 |
| Pix2Struct | 0.003 | 0.038 | 0.077 | 0.000 | 0.077 |

As seen from the above two tables, none of the models produce well-generated captions that are semantically aligned with the original captions which is reflected in the very low metrics scores.

# APPENDIX C

# ANN TRAINING GRAPH ON DATASETS

## C.1 COCO Dataset

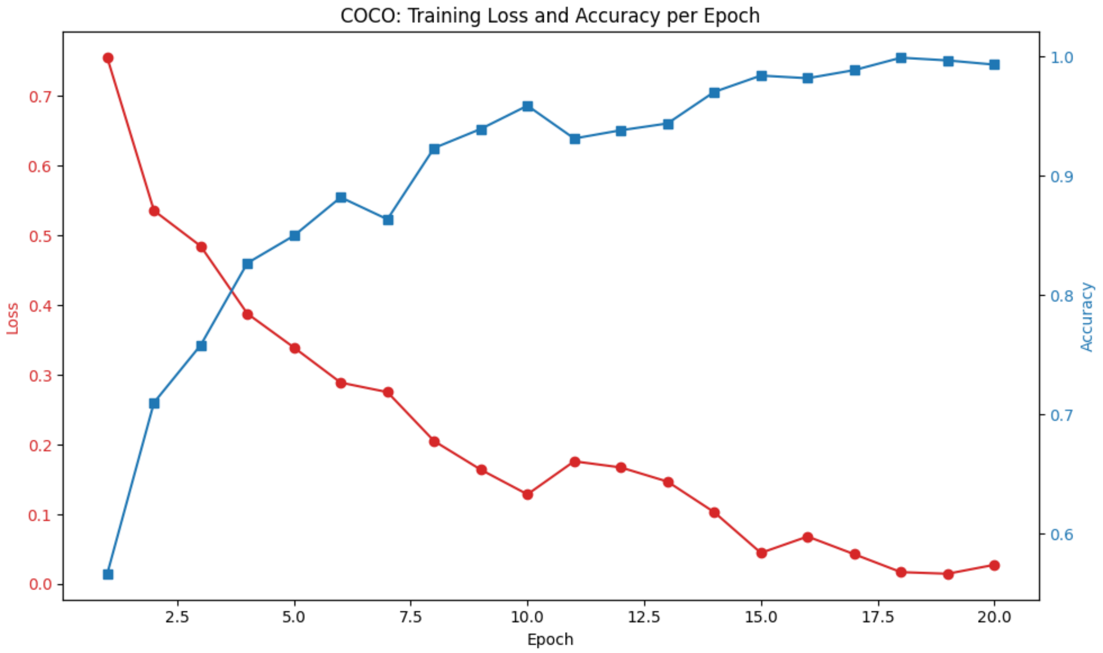


**Figure C.1:** COCO: Training Loss and Accuracy per Epoch (1100 images)

The training curve for the ANN model, trained on the COCO dataset, demonstrates the model's significant progress over 20 epochs. This model employs ResNet-extracted image features to assign binary labels—1 or 0—based on whether BLIP or GPT-2 models' captions exhibit higher cosine similarity with the original captions associated with each image.

The training initiated with a loss of 0.7552 and an accuracy of 56.65%, indicating the beginning phase of learning. Notable progress is evident by the second epoch, with the loss decreasing to 0.5358 and accuracy improving to 70.99%. By the third epoch, the accuracy further increases to 75.80%, with a loss of 0.4849, showcasing the model's rapid adaptation to the features.

A consistent improvement is observed in the training trajectory, with the model reaching a significant accuracy of 95.87% by the tenth epoch. Despite a slight fluctuation in the subsequent epochs, where accuracy briefly dips to 93.12% in the eleventh epoch, it quickly recovers, reaching a peak accuracy of 98.89% by the eighteenth epoch. This fluctuation may reflect the model's response to the more complex or diverse data within the training set.

Throughout the final stages, the model fine-tunes and stabilizes, achieving an impressive accuracy of 99.31% by the twentieth epoch. However, during post-training validation, the model exhibits a loss of 1.1811 with an accuracy of 73.39% on unseen data, suggesting potential overfitting to the training data. While the model shows high performance on the training set, its performance on new, unseen images is significantly lower, indicating a need for further refinement to enhance its generalization capabilities across diverse image contexts.

## C.2 FLICKR Dataset

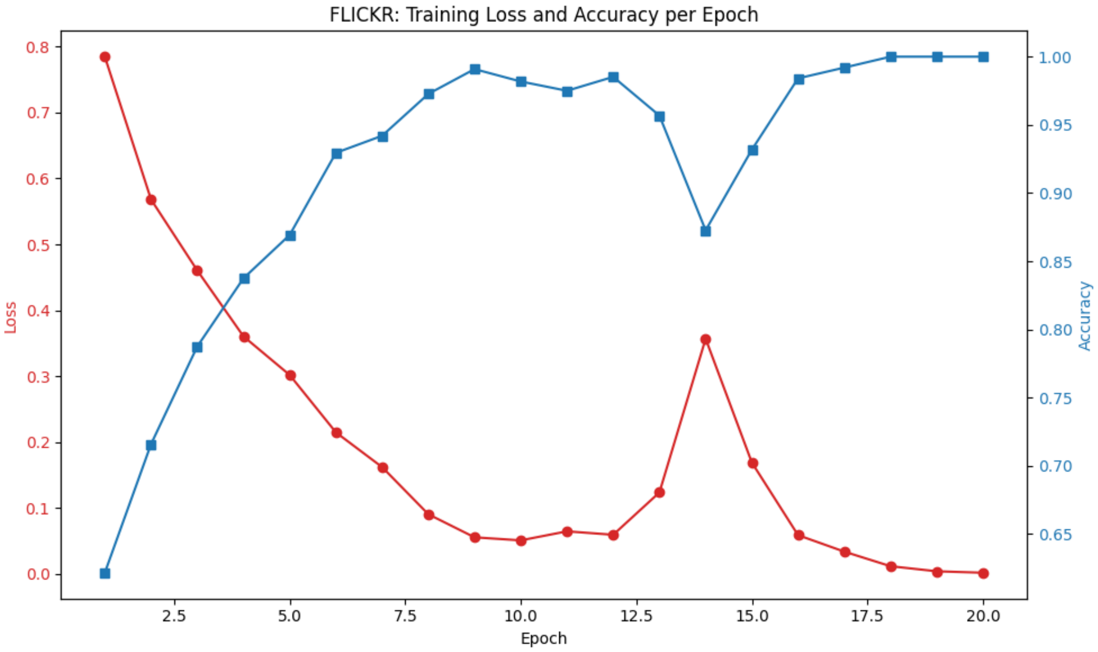


**Figure C.2:** FLICKR: Training Loss and Accuracy per Epoch (1100 images)

For FLICKR, the training began with an initial loss of 0.7847 and an accuracy of 62.16%, showing the model's initial phase of adapting to the feature set. As training progressed, significant improvements were seen by the second epoch with a reduction in loss to 0.5681 and an increase in accuracy to 71.59%. By the fifth epoch, the model achieved an accuracy of 86.93%, with a further reduced loss of 0.3019, indicating strong learning dynamics.

The model demonstrated substantial gains in learning efficiency, with accuracy reaching 97.27% by the eighth epoch and peaking at 100% by the eighteenth epoch. This peak suggests optimal internal model adjustments and learning of the dataset's nuances. However, a notable challenge is observed in the stability of the learning curve; the model experienced a significant drop in accuracy to 87.27% during the fourteenth epoch, possibly due to overfitting or an anomalous batch of data.

Stabilization and fine-tuning continued towards the latter epochs, with the model consistently maintaining a perfect accuracy of 100% from the eighteenth to the twentieth epoch. Despite these high training accuracies, the model's performance on the validation set, as indicated by a final test loss of 1.7454 and an accuracy of 60.91%, suggests that the model might be overfitting the training data and not generalizing well to new, unseen images from the Flickr dataset. This highlights the potential need for additional regularization techniques or training data diversification to enhance the model's ability to generalize more effectively across various real-world scenarios.

# APPENDIX D

# EXPLORATION OF ADDITIONAL MACHINE LEARNING MODELS

In addition to ANN, we also investigated alternate approaches and other ensemble methods to help with the decision-making process for weights assigned to captions created by various models, such as BLIP and GPT-2, and to optimize the automated captioning process.

## D.1 Random Forest Model

We looked at the Random Forest model because it was reliable and could handle complex patterns without a lot of parameters adjusting[84]. The hyperparameters were set to employ 100 trees, and the model was trained using an 80-20 train-test split. On the test set with an 80-20 train-test split, the Random Forest model yielded an accuracy of 68.8% as seen from Figure. D.1.

The training accuracy initially improves sharply with the number of trees, indicating that the model quickly captures the variance in the training data and then attains 100% accuracy indicating overfitting with the test accuracy coming out to be 68%.

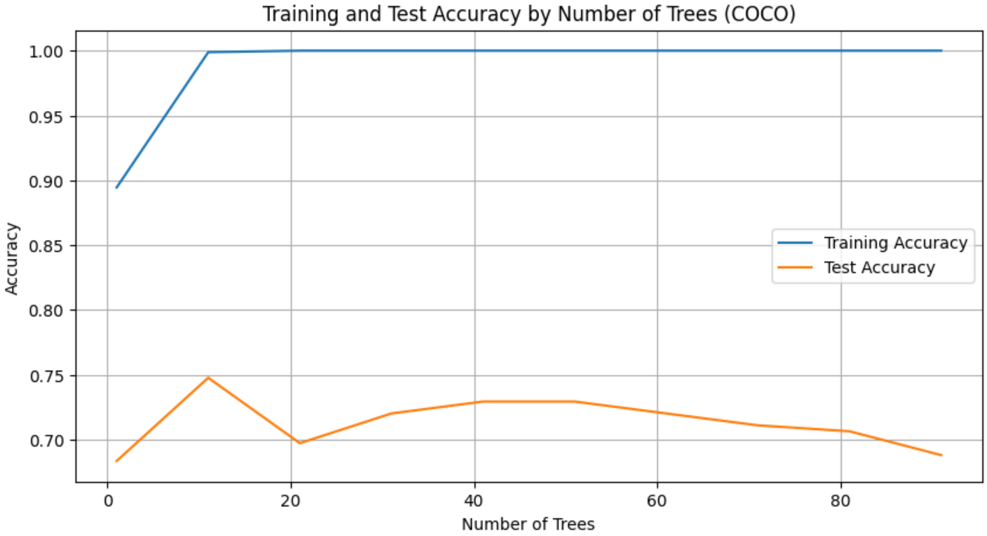


**Figure D.1:** Training and Test accuracy of the Random Forest model (COCO dataset,1100 images).

## D.2 Support Vector Machine (SVM)

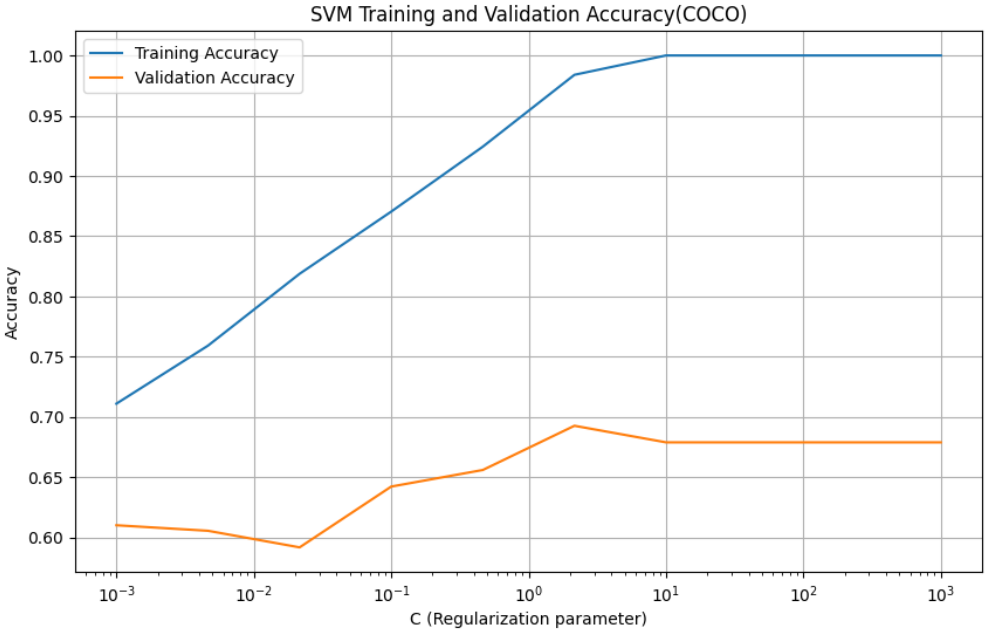


**Figure D.2:** SVM Training and Validation Accuracy (COCO dataset,1100 images).

The SVM's efficiency in high-dimensional spaces—a common scenario for feature sets obtained from text and image processing tasks—was also taken into consideration[85]. Because SVMs are sensitive to the scale of input characteristics, feature scaling was essential. The linear kernel in the original SVM model has $C = 1.0$. On the test set, the first SVM model's accuracy was 67.8%. Figure D.2 shows the SVM's performance with a variable regularization parameter $C$. The training accuracy greatly increases as $C$ rises, indicating that the model gets more adaptable and adept at capturing the training data. However, at increased complexity, the validation accuracy peaks, which may suggest that the model has limited generalization power beyond the training set.

To refine the SVM's performance, a grid search was implemented to explore a range of parameters systematically. The grid search explored combinations of $C$, $\gamma$, and kernels. The parameter grid was defined as:

```
param_grid = {
    'C': [0.1, 1, 10],
    'gamma': [1, 0.1, 0.01],
    'kernel': ['rbf', 'poly', 'sigmoid']
}
```

It was determined that $C = 10$, $\gamma = 0.01$, and kernel='rbf' were the optimal values. With these parameters, 68.35% test accuracy was attained.

**The Artificial Neural Network (ANN) model was selected as the main model for continued development in our approach, despite Random Forest and SVM's respectable results**. Outperforming the other models, the ANN had the maximum accuracy of **73.3% on the COCO dataset**(Appendix. C). Because neural networks can collect complex information in images through layers that may be customized, they are especially well-suited for image captioning. Additionally, the same analytical methodology used for the COCO dataset was applied to the FLICKR dataset to provide a thorough and impartial assessment.

APPENDIX E

# SUMMARIZATION MODELS AND PARAMETERS COMPARISON

## E.1 Comparison of Summarization Models

The following table provides a detailed comparison of summarization models based on their total parameters, trainable parameters, and BLEU scores. This comparison was crucial in selecting the most suitable model for our research.

**Table E.1:** Comparison of Summarization models considered in this thesis

| **Model** | **Total Params** | **Trainable Params** | **BLEU Score** |
|---|---|---|---|
| DistilBART-CNN-12-6 [69] | 305,510,400 | 305,510,400 | 0.277 |
| Callidior/Bert2Bert [86] | 247,363,386 | 247,363,386 | 0.007 |
| Einmalumdiewelt/T5-Base GNAD [87] | 222,882,048 | 222,882,048 | 0.132 |
| Google/BigBird-Pegasus [88] | 576,891,904 | 576,891,904 | 0.121 |
| VietAI/VIT5 [89] | 791,276,544 | 791,276,544 | 0.243 |

As we can see from the above Table. E.1, Distilabart performs the best with **0.277** bleu score.

## E.2 Comparison of Model Parameters

This table provides a detailed comparison of the total number of parameters used in our approach with those used in various models reported by Ramos et al.[33]

**Table E.2:** Detailed comparison of the total number of parameters and METEOR scores in our approach versus those in models studied by Ramos et al., Data taken from:[33]

| **Model** | **Number of Parameters (Millions)** | **METEOR Score** |
|---|---|---|
| **ANN+BLIP+GPT-2+Summarizer** | **791** | **32.8** |
| LEMONHuge[33] | 675 | 30.8 |
| SimVLMHuge[33] | 632 | 33.7 |
| OSCARLarge[33] | 338 | 30.7 |
| I-TuningLarge[33] | 95 | 29.3 |
| CaMEL[33] | 76 | 29.4 |
| I-TuningMedium[33] | 44 | 28.8 |
| ClipCap[33] | 43 | 27.5 |
| I-TuningBase[33] | 14 | 28.3 |
| SMALLCAP[33] | 7 | 27.9 |
| SMALLCAP(d=16, Large)[33] | 47 | 28.3 |
| SMALLCAP(d=16, Med)[33] | 22 | 28.1 |
| SMALLCAP(d=8, Base)[33] | 3.6 | 27.8 |
| SMALLCAP(d=4, Base)[33] | 1.8 | 27.4 |

Our model generates captions of superior quality even with a higher computational demand thanks to its high METEOR score, which is achieved despite the larger number of parameters. This thorough analysis demonstrates the efficacy and efficiency of our approach, especially in situations when caption quality takes precedence over computing cost.

# APPENDIX F

# CAPTIONS WITH LOWEST SCORE

Here we see some of the generated captions in this thesis work with one of the lowest average BLEU, METEOR, and ROUGE scores. These examples illustrate the limitations of the current approach and suggest directions for future work.

**Table F.1:** Analysis of Final captions with their evaluation scores and discussions for the COCO and FLICKR datasets.

| **Section** | **Content** |
|---|---|
| | **COCO Dataset** |
| **Example 1:** | |
| Generated Caption (BLIP) | A woman and a baby in a kitchen |
| Generated Caption (GPT-2) | A woman and a baby are in a kitchen |
| Final Caption | A woman |
| Scores | BLEU: 0.0013, METEOR: 0.0645, ROUGE-1 f: 0.25, ROUGE-2 f: 0.125, ROUGE-L f: 0.25 |

Continued on next page

Table F.1 – continued from previous page

| **Section** | **Content** |
|---|---|
| Discussion | The analysis for the Final caption suggests that the post-processing can sometimes **overly simplify generated captions making it too short to handle details, indicating a need to develop parameters that decide the levels of pruning or post-processing on generated captions**. |
| **Example 2:** | |
| Generated Caption (BLIP) | A dog is standing next to a car |
| Generated Caption (GPT-2) | A car with a surfboard and a dog on the back of it |
| Final Caption | A dog with a surfboard |
| Scores | BLEU: 0.0722, METEOR: 0.0949, ROUGE-1 f: 0.3, ROUGE-2 f: 0.0, ROUGE-L f: 0.2 |
| Discussion | The Final caption failed to capture keywords and does not accurately reflect the context or the idea of the image, **suggesting the need for more experimentation on different weighted summarization methods that better grasp the subject matter**. |
| | **FLICKR Dataset** |
| **Example 3:** | |
| Generated Caption (BLIP) | There is a plane that is flying over a rock formation |
| Generated Caption (GPT-2) | A rock wall with a bird perched on top of it |
| Final Caption | Plane is flying over a rock formation with a rock wall with a plane. |
| Scores | BLEU: 0.0382, METEOR: 0.0575, ROUGE-1 f: 0.0, ROUGE-2 f: 0.0, ROUGE-L f: 0.0 |

Continued on next page

Table F.1 – continued from previous page

| Section | Content |
|---|---|
| Discussion | Here the final caption, in contrast to the second example, has succeeded in capturing essential word details but is not being able to frame it correctly. This suggests an improvement with the way ANN predictions lead to weight assignment for summarization. Hence, **it calls for extensive training of ANN over a fairly large dataset that comprises varied types of scenes and categories**. |
| **Example 4:** | |
| Generated Caption (BLIP) | There are two women standing on the sidewalk talking to each other |
| Generated Caption (GPT-2) | Two women walking down a sidewalk with luggage |
| Final Caption | Two women walking down a sidewalk with luggage there are two women standing |
| Scores | BLEU: 0.0383, METEOR: 0.0711, ROUGE-1 f: 0.1333, ROUGE-2 f: 0.0, ROUGE-L f: 0.0667 |
| Discussion | In this example, we see that the Final caption has fairly captured all the details required but still lacks proper grammar understanding, suggesting the need for **using a more robust grammar checking and language framing library or tuning the already used library/package according to specific needs** implemented in the post-processing. |

# APPENDIX G

# RESOURCES AND MODELS USED

**Table G.1:** List of tools and resources used in this thesis

| Resource Type | Links |
|---|---|
| **Caption Generators** | BLIP Image Captioning Large, ViT-GPT2 Image Captioning, Pix2Struct TextCaps Base |
| **Datasets** | COCO: Dataset Link<br>Flickr8K: Dataset Link |
| **Summarizers** | Primary: DistilBART<br>Others(for comparison): BERT2BERT, T5 Base GNAD, Meta-Llama-3-8B, ViT5 Large, Fine-tuned BART |
| **Zero Shot Classification(classifying image into categories)** | BART Large MNLI |

# Compliance with REALTEK AI Usage Guidelines

In accordance with the guidelines provided by the Faculty of Science and Technology (REALTEK) for the use of artificial intelligence (AI) in academic work, I hereby affirm my awareness and adherence to these principles throughout the preparation of this document.

## Declaration of AI Usage

AI tools have been employed in the following mentioned aspects of this work, strictly following REALTEK's guidelines to ensure academic integrity and reliability:

- **Enhancing LaTeX Documentation:** AI was utilized to generate ideas for improved LaTeX syntaxes, facilitating the creation of better table structures and overall page layouts, ensuring that the document adheres to high standards of academic presentation and readability.
- **Code Debugging Assistance:** AI was instrumental in debugging complex code snippets that were initially challenging. It provided suggestions and solutions that were not readily available on the web.

# APPENDIX H

# COPYRIGHT FOR USED DATASETS

## H.1 COCO Dataset

The COCO (Common Objects in Context) dataset is provided under the ***Creative Commons Attribution 4.0 License***, which requires users to credit the dataset when results derived from it are published. Complete license details can be found at the provided link.

Available at: `https://cocodataset.org/#download`
License details: `https://creativecommons.org/licenses/by/4.0/legalcode`

## H.2 FLICKR Dataset

The FLICKR8k dataset is covered by the ***Creative Commons CC0 1.0 Universal (Public Domain Dedication)*** license, permitting unrestricted use. More details on the licensing can be found on the Creative Commons website.

Available at: `https://www.kaggle.com/datasets/adityajn105/flickr8k/data`
License details: `https://creativecommons.org/publicdomain/zero/1.0/`

## H.3 Diffusion DB Dataset

Similar to the FLICKR dataset, the Diffusion DB dataset is covered by the ***Creative Commons CC0 1.0 Universal (Public Domain Dedication)*** license, which allows for unrestricted use.

Available at: `https://huggingface.co/datasets/poloclub/diffusiondb`
License details: `https://creativecommons.org/publicdomain/zero/1.0/`

## H.4 Google Conceptual Captions Dataset

It is **free and open source** to utilize the Google Conceptual Captions dataset for academic and research purposes.

Available at: `https://ai.google.com/research/ConceptualCaptions/`

Thank you.

**Norges miljø- og biovitenskapelige universitet**
Noregs miljø- og biovitskapelege universitet
Norwegian University of Life Sciences

Postboks 5003
NO-1432 Ås
Norway