

## Declaration of Candidate

This is to certify that the work presented in this thesis is the outcome of the analysis and experiments carried out by **Abdoulaye keita**, Student ID: 180041242 , **Mohaman Dairou**, Student ID: 180041251 and **Mazen Asag** Student ID: 180041257, under the supervision of **Md. Hamjajul Ashmafee**, Assitant Professor, Department of Computer Science and Engineering (CSE), Islamic University of Technology (IUT), Dhaka, Bangladesh. It is also declared that neither this thesis nor any part of it has been submitted anywhere else for any degree or diploma. Information derived from the published and unpublished work of others has been acknowledged in the text and a list of references is given.



---

**Md. Hamjajul Ashmafee**  
Assitant Professor  
Department of Computer Science and Engineering  
Islamic University of Technology (IUT)




---

**Mohaman Dairou**  
Student No.: 180041251



---

**Abdoulaye Keita**  
Student No.: 180041242



---

**Mazen Abdulwahab Asag**  
Student No.: 180041257

# **Deep Learning Approach: Image Captioning in French and Arabic Language**

**Abdoulaye Keita**

**Mohaman Dairou Hamadou**

**Mazen Abdulwahab Mahyoub Salem Asag**

Department of Computer Science and Engineering

Islamic University of Technology (IUT)

june 09, 2023.

# **Deep Learning Approach: Image Captioning in French and Arabic Language**

by

Abdoulaye Keita 180041242

Mohaman Dairou Hamadou 180041251

Mazen Abdulwahab Mahyoub Salem Asag 180041257

Supervisor

Md. Hamjajul Ashmafee

Assistant Professor

Department of Computer Science and Engineering

Islamic University of Technology

## **BACHELOR OF SCIENCE IN COMPUTER SCIENCE AND ENGINEERING**



Department of Computer Science and Engineering

Islamic University of Technology (IUT)

Board Bazar, Gazipur-1704, Bangladesh.

june 09, 2023.

## Declaration of Candidate

This is to certify that the work presented in this thesis is the outcome of the analysis and experiments carried out by **Abdoulaye keita**, Student ID: 180041242 , **Mohaman Dairou**, Student ID: 180041251 and **Mazen Asag** Student ID: 180041257, under the supervision of **Md. Hamjajul Ashmafee**, Assitant Professor, Department of Computer Science and Engineering (CSE), Islamic University of Technology (IUT), Dhaka, Bangladesh. It is also declared that neither this thesis nor any part of it has been submitted anywhere else for any degree or diploma. Information derived from the published and unpublished work of others has been acknowledged in the text and a list of references is given.

**Md. Hamjajul Ashmafee**

Assitant Professor

Department of Computer Science and Engineering  
Islamic University of Technology (IUT)

**Mohaman Dairou**

Student No.: 180041251

**Abdoulaye Keita**

Student No.: 180041242

**Mazen Abdulwahab Asag**

Student No.: 180041257

*Dedicated to my parents*

# Table of Contents

	<b>Page</b>
<b>Acknowledgement</b>	<b>viii</b>
<b>Abstract</b>	<b>ix</b>
<b>1 Introduction</b>	<b>1</b>
1.1 Introduction . . . . .	1
1.2 Motivation and Scope . . . . .	2
1.3 Problem Statement . . . . .	3
1.4 Research Challenges . . . . .	3
1.5 Research Contributions . . . . .	4
1.6 Organization . . . . .	5
<b>2 Background Study</b>	<b>6</b>
2.1 Previous work on image captioning . . . . .	6
2.2 Recent work on machine translation for captioning . . . . .	10
2.3 Limitations and gaps in previous work . . . . .	11
<b>3 Proposed Methodology</b>	<b>12</b>
3.1 Data collection and preprocessing . . . . .	12
3.2 Translation using machine learning models . . . . .	12
3.3 Image captioning model . . . . .	17
<b>4 Results and Discussion</b>	<b>19</b>
4.1 FLICKR30K Dataset . . . . .	19
4.2 Evaluation Metric of French Flickr30K dataset . . . . .	19
4.2.1 Human Evaluation using a Webform consisting of images with their corresponding captions . . . . .	20
4.2.2 Model Performance . . . . .	21
4.3 Comparison with existing captioning datasets . . . . .	23
4.3.1 Analysis of translation and captioning errors and limitations . . . . .	25

4.4	Research Solutions . . . . .	26
<b>5</b>	<b>Conclusion</b>	<b>28</b>
5.1	Summary . . . . .	28
5.2	Future Works . . . . .	28
	<b>References</b>	<b>29</b>

## List of Figures

	<b>Page</b>
1.1 Basic Architecture of image captioning [1] . . . . .	1
1.2 Challenges to the creation of French and Arabic Flickr30K dataset . .	4
3.1 Google Translate Architecture [2] . . . . .	13
3.2 Transformer Architecture . . . . .	14
3.3 Language Models . . . . .	15
3.4 Model Implementation Approach . . . . .	17
3.5 Resnet-50-LSTM Architecture [3] . . . . .	18
4.1 Webform consisting of images with their corresponding captions . . .	20
4.2 Performance with T5 Base method . . . . .	21
4.3 Performance with Google Translate . . . . .	22
4.4 Loss Curve with Google Translate . . . . .	23
4.5 EncoderDecoder with T5 Base . . . . .	24
4.6 EncoderDecoder with T5 Small . . . . .	24
4.7 EncoderDecoder with T5 Base . . . . .	25
4.8 EncoderDecoder with Google Translate . . . . .	25



## List of Tables

	<b>Page</b>
3.1 Language Model with Dataset Information . . . . .	12
3.2 Machine Translation models . . . . .	15
3.3 Some examples of translated captions with the different language models(Google Translate, T5 Small, etc. . . . .	16
4.1 Language Model with captions Information . . . . .	19
4.2 Evaluation of Image Captioning Models using different translation methods . . . . .	20
4.3 Evaluation of Image Captioning Models on different Dataset . . . . .	24

## List of Abbreviations

<b>CNN</b>	Convolutional Neural Networks
<b>T5 Base</b>	Text-To-Text Transfer Transformer Base
<b>T5 Small</b>	Text-To-Text Transfer Transformer Small
<b>CV</b>	Computer Vision
<b>NLP</b>	Natural Language Processing
<b>GAN</b>	Generative Adversarial Networks
<b>RL</b>	Reinforcement Learning
<b>LSTM</b>	Long Short-Term Memory
<b>RNN</b>	Recurrent Neural Networks
<b>NMT</b>	Neural Machine Translation
<b>SMT</b>	statistical machine translation
<b>BLEU</b>	BiLingual Evaluation Understudy

## **Acknowledgment**

We express our sincere gratitude to the following individuals for their contributions to this work: our supervisor, Md. Hamjajul Ashmafee, for their guidance and expertise throughout this research; our research committee members, Abdoulaye Keita, Mohaman Dairou, and Mazen Asag, for their valuable inputs and constructive criticism; the faculty members of the CSE department at the Islamic University of Technology (IUT) for their teachings and mentorship; our colleagues and fellow researchers for their support and collaboration; the participants who volunteered their time and efforts for data collection and experiments; and our family and friends for their unwavering support and encouragement. We sincerely thank each and every individual who has contributed to this work, directly or indirectly. Your contributions have made a significant impact, and your support has been invaluable.

## Abstract

This research report introduces a novel dataset of French captions translated from the Flickr30k dataset using different translation models, namely we have Google Translate and the powerful Transformers: T5 Small and T5 base models. A novel dataset of French captions means creating fresh data collection by translating existing captions from the Flickr30k dataset into French. The Flickr30k dataset is valuable for training and evaluating image captioning models in French.

The main objective is to address the problem of generating precise image captions in French. The performance of an image captioning model is evaluated on the translated datasets, employing ResNet-50 for image feature encoding and LSTM network with attention in generating captions. These results demonstrate that the accuracy of image captions varies depending on the translation(or Language) models, with the Transformers models outperforming Google Translate. The proposed approach achieves state-of-the-art performance in generating accurate French captions when combined with ResNet-50 and LSTM network with attention.

The findings contribute to the field of image captioning and machine translation for French speakers, highlighting the importance of using advanced translation models for improved caption accuracy and other NLP tasks in French. Furthermore, this research provides insights into the potential of smaller-scale models in limited data scenarios. Based on our findings, we can explore alternative translation models, and data augmentation techniques, and consider multi-modal approaches that could lead to more accurate and contextually relevant captions and the potential of this approach in other languages.

**Keywords :** Novel dataset, Translation models, Transformers, Image captioning, Natural Language Processing, Multimodal technologies

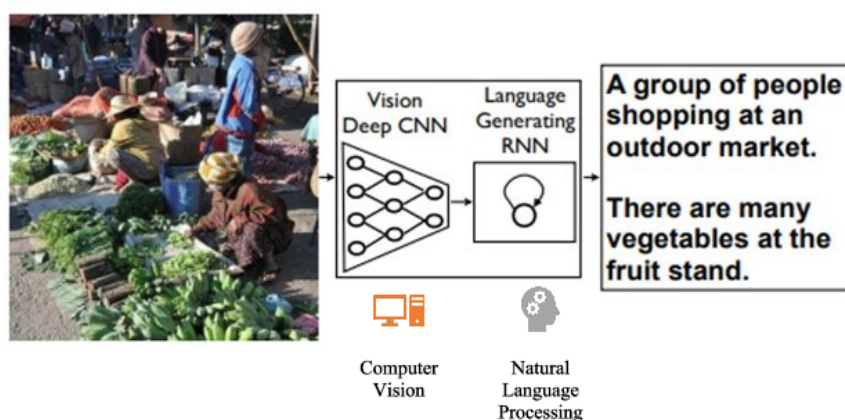
# Chapter 1

## Introduction

### 1.1 Introduction

Image captioning has gained significant attention as an area of research, with the aim of automatically generating descriptive text for images, as illustrated on Figure 1.1 [1] below, which consist of 2 parts: CV and NLP task.

The Flickr30k dataset developed by Hodosh and Associates at Berkeley University was created in the context of their research into image descriptions and captioning. This dataset was intended for the benefit of facilitating CV and NLP research. The primary task associated with the Flickr30k [4] dataset was to generate image descriptions and do captioning tasks. Also, the dataset only consists of a large collection of images and human-annotated captions, enabling researchers to develop and evaluate algorithms for automated captioning systems.



**Figure 1.1:** Basic Architecture of image captioning [1]

Often, they are used as a sample dataset containing a subset of images and their corresponding captions from the larger pool of images available on the Flickr platform

providing a wide variety of representative images and captions and sometimes be selected based on certain criteria or methods, such as popularity, relevance, and quality, but may not represent the entire population of images and captions on Flickr.

Although the Flickr30k dataset [4] has been a widely used benchmark for image captioning research, its availability is currently limited to English, which posed a challenge to researchers who need to evaluate their models in other languages as well. In this paper, we present our efforts to extend the Flickr30k dataset [4] to multiple languages like French, and Arabic and discuss its motivation.

Some applications of image captioning areas are:

- Self-driving cars.
- Visually impaired.
- CCTV cameras and relevant captions.
- Improve Image Search in Search Engines.

## **1.2 Motivation and Scope**

The motivation behind creating a French and Arabic version of the Flickr30K [4] image captioning dataset using machine translation is to enable and facilitate French-speaking and Arab-speaking users to benefit from the dataset in their NLP tasks. This would improve the accessibility of the dataset but also enhance the quality of French and Arabic image captions, ultimately researchers and developers can better understand and analyze the linguistic and cultural nuances present in French and Arabic captions.

When utilizing machine translation, it is important to acknowledge that achieving 100% accuracy in translation is challenging. This becomes particularly evident when translating artistic works such as poems from one language to another. Machine translation may struggle to capture the full essence, subtleties, and nuances of the original text, resulting in potential loss or alteration of meaning in the translated document.

The limited accuracy or research of implemented models for the French and Arabic languages in image captioning tasks is a significant challenge that needs to be addressed. While image captioning models have shown impressive performance in English, they often struggle to achieve similar levels of accuracy and fluency in other languages such as French and Arabic. One of the main reasons for the limited accuracy in French and Arabic image captioning is the lack of sufficient training data. Altogether, addressing these challenges will contribute to the development of contextual and accurate image caption systems for French and Arabic, expanding the application of this technology to a wider range of languages and cultural contexts

### **1.3 Problem Statement**

Our main intention is to construct high-quality French and Arabic versions of the Flickr30K image captioning dataset by using the power of machine translation, addressing the challenges of accuracy and linguistic nuances, to improve the quality of image captions in these languages.

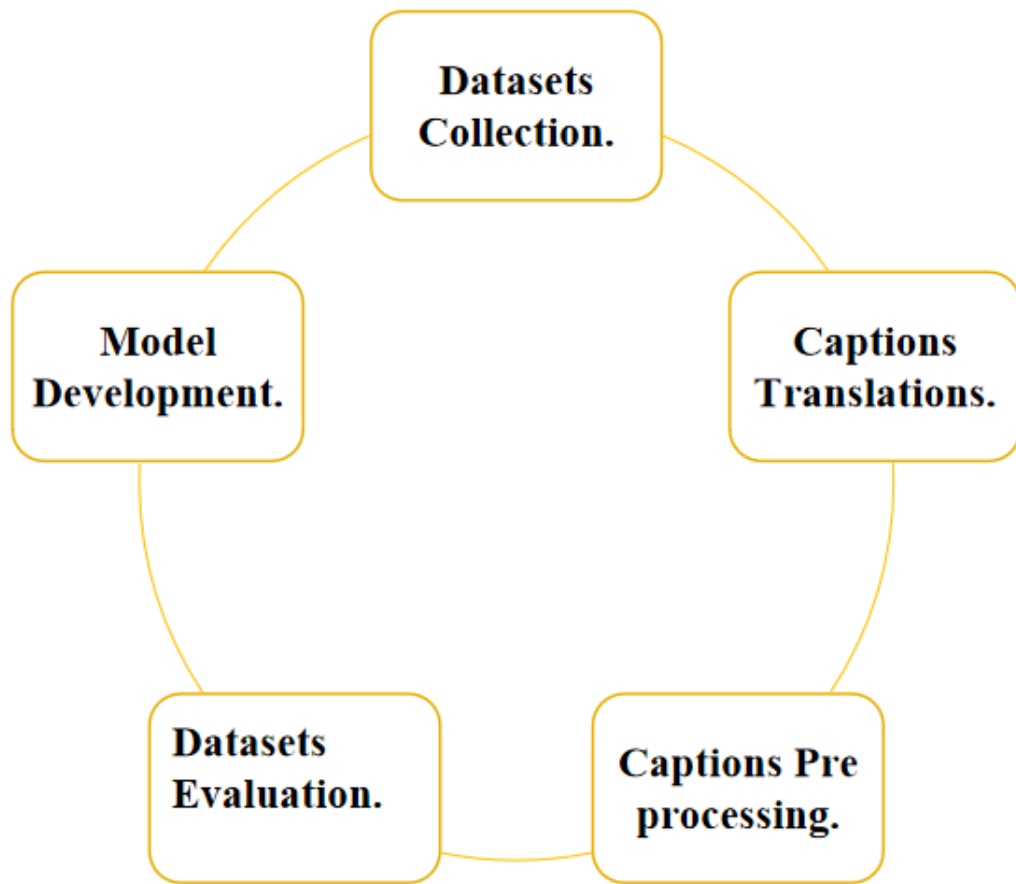
This research paper will explore the potential of using machine translation to construct a French and Arabic Flickr30K Image Caption Dataset. We will focus on three main research questions:

1. What are the best methods for improving French and Arabic image captions using machine translation?
2. How can we effectively construct a French and Arabic Flickr30K [4] Image Caption Dataset?
3. How can we evaluate the fluency and accuracy of machine-translated French and Arabic image captions, and what are the most appropriate metrics for this task?

By exploring these research questions, we hope to gain insights into how machine translation can be used and create a comprehensive dataset for further usage.

### **1.4 Research Challenges**

Creating a French dataset for research comes with many challenges, Figure 1.2 illustrates the main challenges. First, gathering a complete and diverse French corpus can be difficult due to availability limitations and restrictions. Second, the correct translation of captions from different languages into French requires expertise in the source and target languages, ensuring semantic consistency. Third, captions preprocessing involves managing language variations, dialects, cultural nuances and idioms specific to the French language, ensuring accurate and meaningful translations. Fourth, it is important to assess the quality and consistency of the dataset, considering factors such as translation accuracy, and relevance to the intended research task. Fifth, developing effective models for using French datasets requires addressing domain-specific grammatical, syntactic, and lexical differences.



**Figure 1.2:** Challenges to the creation of French and Arabic Flickr30K dataset

## 1.5 Research Contributions

The key contributions in this research are summarized as follows:

1. We identified effective methods for improving French and Arabic image captions using machine translation,
2. We developed a process to construct a comprehensive French and Arabic Flickr30K Image Caption Dataset
3. We evaluated the advantages and disadvantages of machine translation for image captions, and proposed appropriate evaluation metrics to check the fluency and accuracy of machine-translated captions in French and Arabic.



## 1.6 Organization

The remaining dissertation is organized as follows. **Chapter 2** discusses the background and motivation for image captioning research in French and Arabic languages and identifies the problems persistent in the existing literature. **Chapter 3** presents a new methodology that addresses these issues and discusses its implementation. **Chapter 4** analyzes the performance of our proposed methodology and presents the results and analysis of the evaluation metrics and comparisons with existing captioning datasets. **Chapter 5** concludes the dissertation by summarizing the key findings, highlighting contributions to the field, and offering recommendations for future research.

# Chapter 2

## Background Study

### 2.1 Previous work on image captioning

In the field of image captioning, some datasets in English including PASCAL, Flickr30k [4], and MS-Coco [5], were developed by different research groups to provide diverse images with corresponding captions for training and evaluating models.

A paper by Katiyar et al. [6] examines the progress made in image caption generation using deep Learning techniques. They focus on evaluating different CNN architectures for feature extraction in caption generation. Surprisingly, they find that the complexity of CNN models, which is measured by the number of parameters and object recognition accuracy, does not necessarily correlate with their effectiveness in feature extraction for caption generation. This study highlights the importance of systematically comparing different CNN architectures [6] for the task of feature extraction [6].

Another paper titled "Image Caption Generation for News Articles" by Yang et al. [7] addresses the challenging task of news image captioning, where the description of an image is generated based on both the image and its corresponding article body. The authors propose a Transformer model that combines text and image modalities to generate captions considering both text and visual features. Experiments, assessed by automated metrics and human evaluation, found that article text played an important role in reproducing news captions written by journalists. The proposed model [7] goes beyond the state-of-the-art model and incorporates visual features to further improve the quality of the news captions, it produces. This study provides valuable insight into the integration of text and image information in news captions.

Chen et al. [8] presents a novel framework for captioning that combines conditional GANs and traditional RL techniques. Their approach addresses the challenge of inconsistent scoring of audio metrics by introducing a discrimination network. These distinctions gradually determine whether the labels generated are human-generated or machine-generated. He examined two types of classifier architectures, CNN and RNN,

each with its own advantages. Our algorithm is flexible and can be applied to enhance existing RL-based captioning frameworks. This approach continuously improves the language score metrics in various state-of-art captioning models [8]. Moreover, well-trained identifiers act as objective evaluators of captions.

Tavakoli et al. [9] explore the fascinating world of image description. Their goal was to understand how humans describe what they see, and how machines can learn the same. They found that when people talk about images, they tend to mention the most important things first. They also looked at how well a machine could represent an image and found that the better a machine performed, the better its description matched the human description. To help machines better represent images, They experimented with special techniques that improve our understanding of important visual details. Surprisingly, we found that this technique does not significantly improve performance for images we are already familiar with. However, we observed that machines were better at describing new and unseen images. Taken together, our results shed light on the fascinating field of image description and provide insight into how machines can come to understand and describe the world around us.

Huang and al. [10] developed an AI system that generates diverse captions and rich images [10]. Here, users were given the ability to imagine an image, associate it with multiple captions, and their system will draw a faithful and detailed representation of the image. Similarly, when a user uploads an image, the system will generate several different captions. Their multimodal framework combines image and text representations, promoting diversity in training and offering various labeling suggestions using transformation networks. The real-time inference is enabled by a non-autoregressive decoding strategy. Their system creates visually appealing images but also offers various captions suggestions.

Furthermore, in neural image captioning systems, Tanti et al. [11] investigated two approaches for incorporating image features into the model: injecting them directly into the RNNs or merging them with the final representation of the RNNs. RNNs is an encoders for language features, and the image features are merged later. Our key findings in this study compare these two architectures and concluded the merge approach to generally be superior to injection. This suggests that RNNs are more effective as encoders rather than generators in captioning tasks.

Katiyar et al. [12] consider the use of CNN-based decoders to generate image captions, which is a sequence modeling task. The authors analyze various aspects of a CNN-based decoder, including network complexity, data augmentation, attention mechanisms, and sentence length during training. Tests conducted on the Flickr8k and Flickr30k datasets show that increasing the network depth with stacked convolutional

layers and using data enhancement techniques generally do not improve the solver performance of CNN code. Furthermore, the use of attention mechanisms has limited effect. The study also found that the CNN decoder performed well when trained with shorter sentences (up to 15 words), but struggled with longer sentences, indicating a limitation in the long efficient dependency model term. In addition, compared with repeater decoders, CNN decoders tend to perform worse on the CIDER rating. These results contribute to an understanding of the strengths and limitations of CNN-based decoders for image annotation generation tasks.

Automatic description generation [13] from natural images has attracted considerable attention in the fields of CV and NLP. In this study, the problem is classified according to different concepts: generation-based approach and retrieval-based approach using visualization or multimodal representation. Their survey provides a comprehensive review of existing models, highlighting their strengths and limitations. In addition, it processes reference image datasets and develops evaluation metrics to evaluate the quality of machine-generated image descriptors. Their survey concluded by discussing future directions in the field of automatic image description generation [13].

Most research has been focused only on English captions, creating a lack of resources for languages like Japanese, Chinese, and French. To address this, the authors Miyazaki et al. [14] developed a Japanese version of the MS Coco [5] image caption dataset and a generative model using a deep recurrent architecture. This model transfers knowledge from the English portion to generate Japanese captions. Experimental results show that leveraging a bilingual corpus improves performance compared to a monolingual one, demonstrating the benefits of using a resource-rich language. This work contributes to expanding image captioning resources for non-English languages.

Furthermore, research was focused on Text-based image captions [9], also known as TextCap [9], which played an important role in machine understanding of complex scene environments by combining visual and textual information. However, current methods struggle to comprehensively describe the complex text and visual details of images. To solve this problem, a proposed Anchor-Captioner [9] was used. This approach uses anchor tokens to guide attention, builds anchor-focused graphics (ACGs) to represent relationships, and generates multiple captions with diverse content. This approach achieved the highest performance while generating diverse annotations.

Existing image annotation models are often evaluated based on their performance on a set of saved images, ignoring their ability to generalize to unseen concepts. Nikolaus et al. [15] focuses on composition generalization, which measures the ability of a model to describe novel conceptual combinations in image captions. Modern image annotation models struggle with this task. To solve this problem, Nikolaus et al. [15] propose

a multi-task model that combines caption generation and image-sentence ranking. The model uses a decoding mechanism to re-rank the generated captions based on their similarity to the image. The experimental results demonstrate that this model significantly improves the generalization ability to novel combinations of concepts compared with existing annotated models.

Some researchers showed that reinforcement learning methods can be used to effectively train an image annotation system. Rennie et al. [16] introduce the critical sequence training (SCST) [16] method. It uses model-tested inference output to improve performance. Their intention was to directly optimize the CIDER index using SCST and a simple greedy decoding strategy during testing, we achieved significant improvements in captioning performance. The test results on the review server MS-Coco have established a new state-of-art in image captioning, with a significant increase in CIDER score from 104.9 to 114.7.

Also, the "Show and Tell" paper by Vinyals et al. [1], proposes a generative model that combines both computer vision and machine translation techniques to automatically generate natural language descriptions for images. The model is trained to generate sentences that accurately describe the content of the image by maximizing the likelihood of the target description sentence. The authors conduct experiments on multiple datasets and demonstrate the accuracy and fluency of the generated descriptions. The results showed significant improvements in the BLEU-1 score compared to the state-of-the-art methods, indicating the high performance of their approach. The model achieves impressive results on various datasets, including Pascal, Flickr30k, SBU, and COCO, surpassing previous benchmarks and even approaching human-level performance.

However, while there has been significant progress in English image captioning, there is a lack of similar resources for non-English languages. To address this gap, researchers have constructed datasets like STAIR Captions [17]; a large-scale Japanese image caption dataset that emphasizes the importance of high-quality translated image captioning for non-English languages.

Arabic, being a semantic language heavily dependent on root words, is an important element of our approach. They use RNNs and Deep Neural Networks based on root words to directly generate Arabic captions for images. Through experiments on datasets from Middle Eastern newspaper websites, They achieved the first reported BLEU score for direct Arabic caption generation [18]. A comparison with English-Arabic translated captions highlights the superior performance of their approach.

Another paper focuses on developing and evaluating Arabic-language image annotation models using metrics established on public benchmarks. We initialized the models

with pre-trained transformers on the Arabic corpus and refined them using OSCAR, a learning method that uses object tags for semantic alignment. Their best model achieved improved scores on BLEU-1,2,3,4 measures compared to previous results. However, using a pure Arabic dataset with Arabic object tags is better based on their experience. These datasets offer valuable resources for CV and NLP researchers working on non-English languages.

## **2.2 Recent work on machine translation for captioning**

It is worth noting that there have been several previous works on using machine translation in multilingual image captioning. For instance, the study by Barz and Sonntag [19] focused on improving German image captions using machine translation and transfer learning using a two-step approach: first, translating the original English captions of the Flickr30k [4] dataset into German using a neural machine translation model, and second, fine-tuning a pretrained image captioning model [20] on the translated German captions using transfer learning. Their work highlights the potential of machine translation and transfer learning for improving the quality of image captions in languages other than English. Similarly, the study made by Lee et al. [21] proposed a multilingual image captioning model that utilizes machine translation to generate captions in multiple languages.

Translation-based approaches have been explored to generate multilingual captions for various datasets, such as COCO [5] and Flickr30k [4], or to create an image captioning dataset in the target language for training language-specific models. One pioneering study was conducted by Elliott et al. [22] where they utilized features from both source and target language models and generated captions using a decoder based on LSTM networks.

Regarding pre-trained language models, Text to Text Transformer [23] has attracted attention due to its versatility in various text-based NLP tasks. It features a consistent "text-to-text" format. This format is particularly useful for generative tasks such as machine translation and summarization, as the model can generate text based on given inputs. T5 takes a different approach to the classification task, where it is trained to output the actual text labels instead of the class index, allowing consistent training with a single set of hyperparameters across all tasks. T5 uses a transformer-based encoder/decoder architecture, similar to the original proposal by Vaswani et al. [24] follows. It is pre-trained on a masked language modeling target to reconstruct the masked region of tokens. T5 is famous for its large-scale models ranging from 60-11 billion parameters, pre-trained on a huge dataset of about 1 trillion tokens obtained from public

common crawl web scraping. Neural Machine Translation has received much attention in recent years and has been the subject of extensive research. Moving from traditional statistical machine translation (SMT) to NMT has significantly improved translation quality. The Google Neural Machine Translation (GNMT) [25] system tackles challenges in Neural Machine Translation (NMT) by using a deep LSTM network with 8 encoder and 8 decoder layers, along with residual and attention connections. It employs low-precision arithmetic during inference and divides words into sub-word units for handling rare words. GNMT incorporates beam search with length normalization and a coverage penalty for improved translation quality. While reinforcement learning was explored, it did not yield significant improvements in human evaluation. On benchmark tests, GNMT achieves competitive results and reduces translation errors by 60% compared to Google’s previous system. Overall, GNMT [25] addresses NMT challenges, improving accuracy and efficiency.

However, unlike these previous works, our approach focuses specifically on enhancing French and Arabic image captions and constructing a French Flickr30K image caption dataset. By narrowing down the scope to these languages, hence addressing the limitations and challenges specific to these languages, our work aims to contribute to the advancement of multilingual image captioning in French and Arabic.

### **2.3 Limitations and gaps in previous work**

The previous work on multilingual image captioning using machine translation has made significant advancements in generating captions in multiple languages. Firstly, the dataset used in the study may not have been diverse enough to capture the complexities and variations of different languages. A more comprehensive and diverse dataset [26], encompassing a wide range of subjects, contexts, and linguistic styles, would provide a more realistic representation of multilingual image captioning challenges. Secondly, the choice of the source language for machine translation can impact the quality and accuracy of the generated captions. The study did not explicitly explore the impact of different source languages on translation performance.

# Chapter 3

## Proposed Methodology

### 3.1 Data collection and preprocessing

The Flickr30k dataset [4] is a popular benchmark dataset for image captioning and is widely used in the research community. It consists of 31,783 images from the Flickr website, each of which is paired with five human-written captions. The dataset was created to address the limitations of previous datasets, which had relatively small sizes and lacked diversity in terms of image content and caption styles. The Flickr30k dataset [4] has become a standard benchmark for evaluating image captioning models, and many state-of-the-art models have been trained on this dataset. However, the original dataset is in English, and so there is a need to create translations of the captions in other languages to enable researchers and practitioners from non-English speaking countries to use the dataset. In this experiment, the dataset is separated into three parts, with 75% of the images utilized for training, 10% for validating, and the remaining 15% for testing, illustrated on Table 3.1

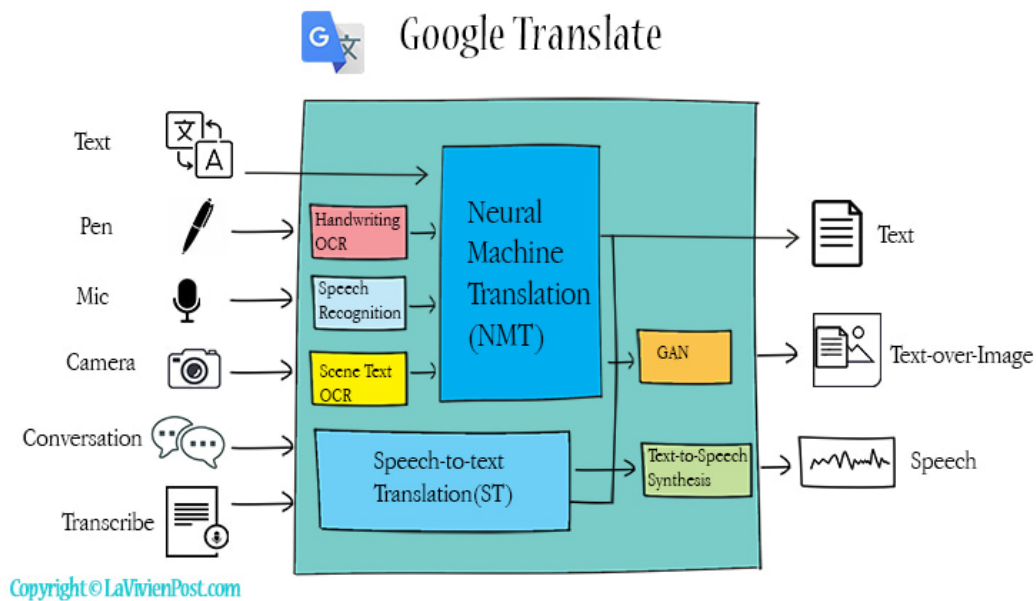
<b>Model</b>	<b>Dataset</b>	<b>Training</b>	<b>Validation</b>	<b>Test</b>
Google Translation [2]	13000	8002	1000	3000
T-5 Small [27]	158915	111250	20000	27665
T-5 Base [28]	158915	111250	20000	27665

**Table 3.1:** Language Model with Dataset Information

### 3.2 Translation using machine learning models

Google translate [2] is widely used for translating text from one language to another. Figure 3.1 displays the model which follows a specific process to generate the target language, French.

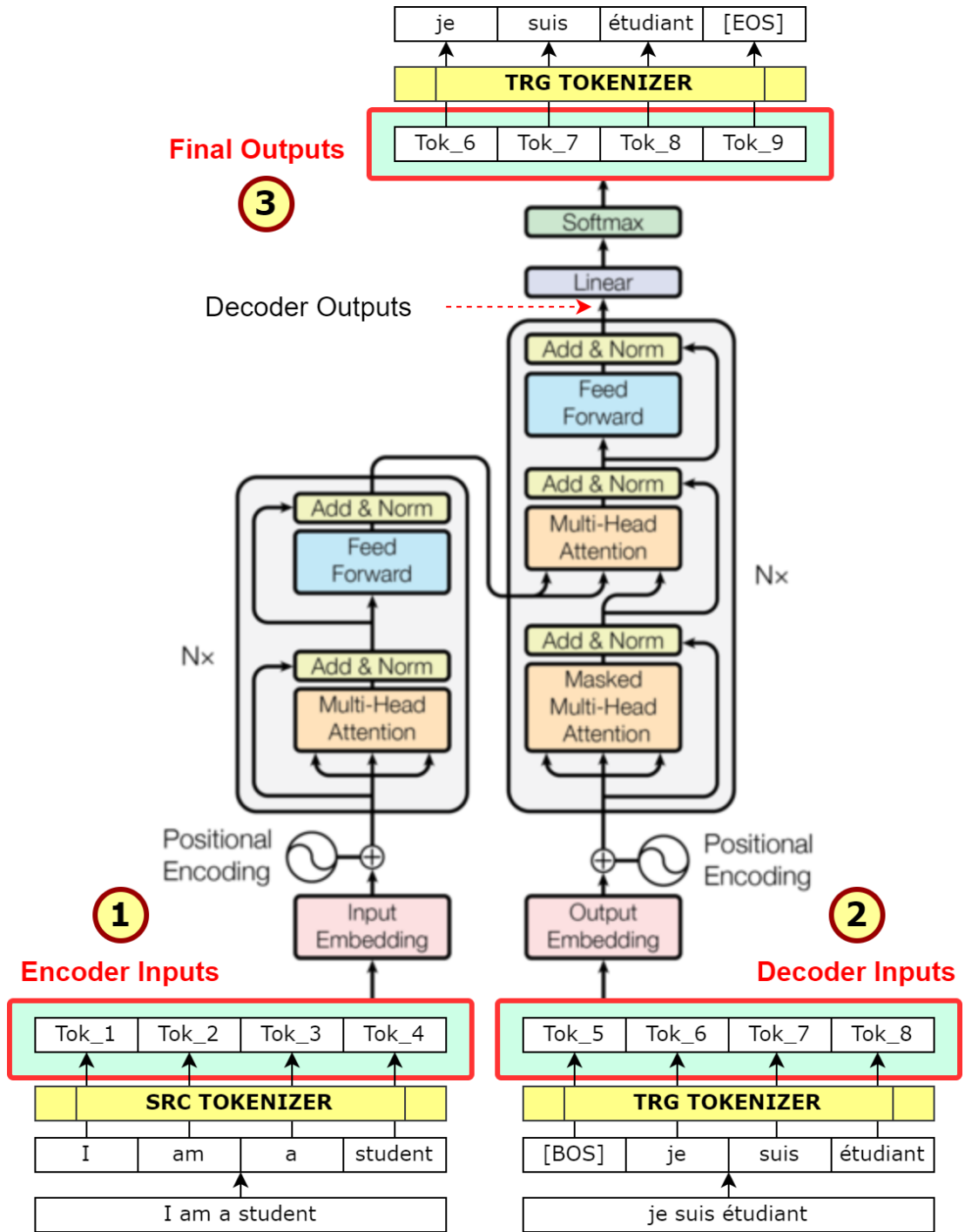




**Figure 3.1:** Google Translate Architecture [2]

The translation process begins by analyzing the input English text to identify grammatical structures, vocabulary choices, and syntactic patterns. This model applies a set of predefined linguistic rules and patterns carefully designed to handle a wide variety of translation scenarios. These rules dictate how different parts of the English text should be transformed into the corresponding French text. It takes into account the context of the source text and considers the surrounding words and phrases to ensure the consistency and naturalness of the resulting French translation. One of the Google Translate model's greatest strengths is its extensive training data. Trained on a large corpus of bilingual and multilingual texts, it can capture common translation patterns and improve overall translation quality. In addition, the model considers user feedback to continuously refine and improve its translation functionality.

With the advancement of machine learning models, the task of translating between languages has become much easier. We will use Google Translate [2], T5 Small [27], T5 Base [28], a state-of-the-art translation model to translate English captions into French and Arabic and then construct a French and Arabic Flickr30k [4] Image Caption Dataset. By using machine translation, we can ensure that the captions are accurate and meaningful in both languages.



**Figure 3.2:** Transformer Architecture

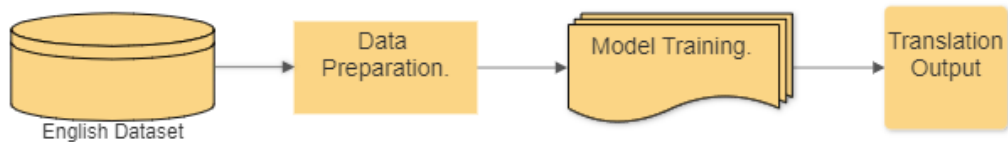
Translation using machine learning models involves the following key steps: data preparation, model training, model deployment, translation inference, and translation output. In this particular case, pretrained models, specifically the T5-small model [27] and Google Translate [28] are employed to translate the Flickr30k dataset. The first step is to prepare the data. The Flickr30k dataset [4], which consists of images and their corresponding captions in a source language (e.g., English), is acquired. The

dataset is preprocessed to ensure proper formatting and alignment between the images and captions. The pretrained T5-small model, T5-Base model, and Google Translate are utilized for translation tasks. Model training involves leveraging the pretrained weights of the T5-small [27] model and fine-tuning it on the Flickr30k [4] dataset. The pretrained T5-small model, T5-base model, and Google Translate are used to generate translations in the target language as shown in Figure 3.3. The input data is passed through the models, which analyze the content and context to produce translated outputs. The translated captions are obtained from the models, providing transformed versions of the original captions in the target language.

When comparing the machine translation models mentioned in Table 3.2 (Google Translate, T5 Small, and T5 Base), there are several general points of comparison to consider:

Features	Google translate [2]	T5 Small [27]	T5 Base [28]
Model Type	Rule-based	Transformer-based	Transformer-based
Model Size	1 billion parameters	60 million parameters	220 million parameters
Translation Accuracy	High	Moderate	Very High
Contextual Understanding	Moderate	Limited	High
Multilingual Support	Yes	Yes	Yes
Availability	Publicly Available	OpenAI Subscription	OpenAI Subscription

**Table 3.2:** Machine Translation models



**Figure 3.3:** Language Models

Image	English caption	Google translation	T5 small caption	T5 base caption	Human Caption
	Wedding photo of the bride and groom jumping for joy on the Great Wall of China.	Photo de mariage des mariés sautant de joie sur la Grande Muraille de Chine.	Photographie de mariage de la mariée et du mari qui saute pour joie sur le Grand Mur de Chine.	Photo du mariage de la mariée et du marié sautant pour la joie sur la Grande Muraille de Chine.	Les mariés sautant en l'air sont sur la Grande Mur de Chine
	A young baby wearing green is playing with the vacuum hose on the floor.	Un jeune bébé vêtu de vert joue avec le tuyau d'aspiration sur le sol.	Un jeune bébé portant le vert joue avec le tuyau de vide sur le plancher.	Un jeune bébé vert joue avec un tuyau d'aspirateur sur le plancher.	Un bébé habillé en vert joue avec un tuyau d'aspirateur sur le plancher.
	The yellow dog walks on the beach with a tennis ball in its mouth.	Le chien jaune se promène sur la plage avec une balle de tennis dans la gueule.	Car pour en venir au moindre détail, nul ne doit pratiquer un travail quelconque s'il n'en tire aucun bénéfice	Le chien jaune marche sur la plage avec une balle de tennis dans sa bouche.	Un chien jaune avec une balle jaune dans la gueule marchant sur une plage.
	Beautiful red car among many at a car show full of excited car enthusiasts held on a tree-lined street.	Belle voiture rouge parmi tant d'autres lors d'un salon de l'automobile rempli d'amateurs de voitures enthousiastes qui se tiennent dans une rue bordée d'arbres.	Une belle voiture rouge parmi de nombreuses personnes lors d'un spectacle de voitures pleines d'enthousiastes automobiles tenue sur une rue couverte d'arbres	Une belle voiture rouge parmi beaucoup à un salon automobile plein d'enthousiastes automobiles tenu sur une rue bordée d'arbres.	Plusieurs personnes sont en train de regarder des voitures dont une de couleur rouge qui s'approche.

**Table 3.3:** Some examples of translated captions with the different language models(Google Translate, T5 Small, etc).

### 3.3 Image captioning model

The aim of this section is to present a summary of the implemented image captioning model, which consists of an Encoder and a Decoder. The model utilizes an encoder-decoder framework with an attention [1] mechanism to generate captions for images. The following report provides a detailed overview of the model's architecture and functionality, as illustrated in Figure 3.4

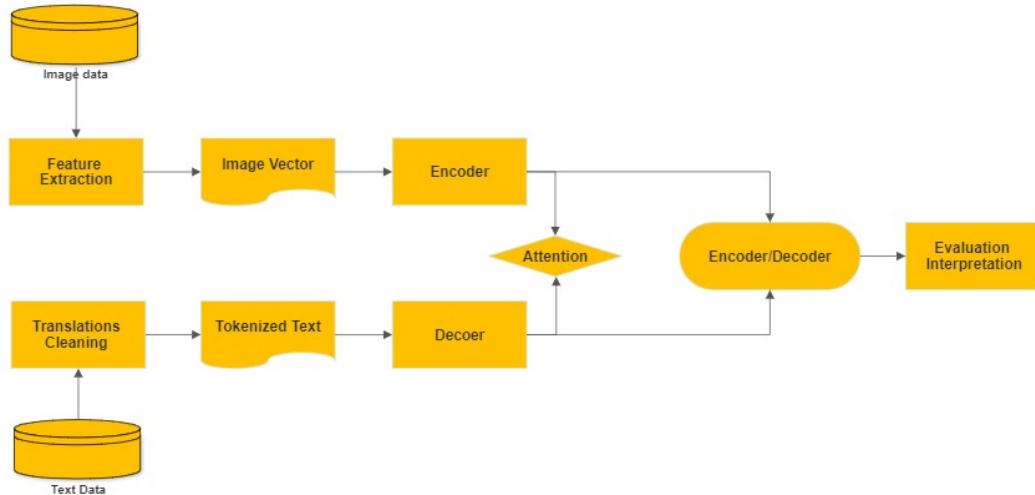
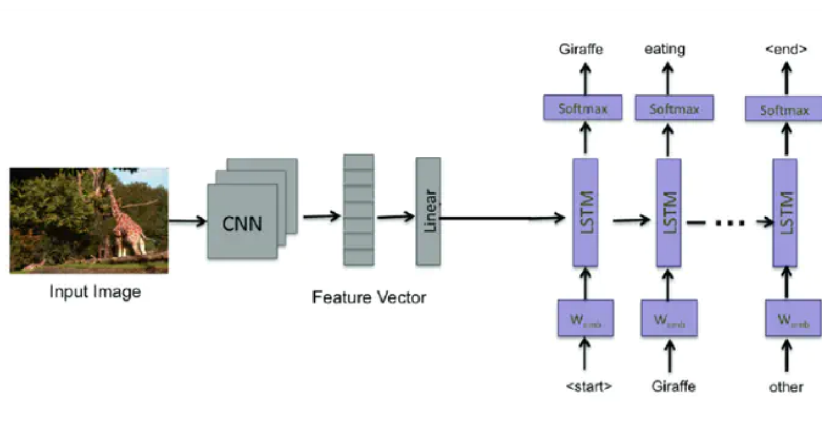


Figure 3.4: Model Implementation Approach

- **Encoder** component employs a pretrained ResNet-50 model to extract visual features from input images. By leveraging the ResNet-50's [3] capabilities, the model is able to capture rich visual representations. The ResNet-50 parameters are frozen to prevent further training and ensure the integrity of the pre-trained weights. The output features are obtained by reshaping the output of the ResNet-50 [3] model and passing them through a linear layer.
- **Attention [1]** module plays a crucial role in aligning the visual features with the hidden state of the decoder. It consists of linear transformations and a softmax function to calculate attention scores. These scores are then applied to the visual features to compute context vectors. By utilizing attention, the model can focus on different regions of the image during the caption generation process.
- **Decoder** generates captions based on the image features and previous word embeddings. It utilizes LSTM cells [3], which enable sequential word generation. The decoder takes the image features, embeds the previous word, and applies attention to obtain context vectors [29]. The LSTM cell takes the concatenated input of the word embedding and context vector, updating its hidden state and

cell state accordingly. Predicted word probabilities are obtained using a linear layer, and dropout is applied to enhance generalization. During training, the model iterates over the sequence length, generating word predictions and storing attention scores. During inference, the model generates captions word by word until it reaches a maximum length or an end token.

- **EncoderDecoder** class combines the Encoder and Decoder components to form the complete translation model. It takes image features and captions as input and passes them through the respective encoder and decoder. The final outputs of the decoder are returned as the predicted translations. The model is trained using cross-entropy loss, which measures the dissimilarity between the predicted word probabilities and the ground truth captions. The Adam optimizer is utilized to optimize the model parameters and update them during training. The provided hyperparameters, including the embedding size, vocabulary size, attention dimension, encoder dimension, decoder dimension, and learning rate, play a crucial role in the training process, as illustrated in Figure 3.5 below



**Figure 3.5:** Resnet-50-LSTM Architecture [3]

In conclusion, the implemented image captioning model leverages an encoder-decoder architecture with an attention mechanism to generate accurate and meaningful image captions. The combination of visual and textual information enables the model to capture the essence of the images and produce coherent translations.

# Chapter 4

## Results and Discussion

### 4.1 FLICKR30K Dataset

- **Dataset Size:** 31,783 images with 158,915 captions.
- **Multimodal:** Combines images and captions for analyzing their relationship.
- **Annotation Format:** Five human-generated captions per image, capturing multiple perspectives.
- **Language:** Captions written in English.

Model	Dataset	Training	Validation	Test
Google Translation [2]	13000	8002	1000	3000
T-5 Small [27]	158915	111250	20000	27665
T-5 Base [28]	158915	111250	20000	27665

**Table 4.1:** Language Model with captions Information

### 4.2 Evaluation Metric of French Flickr30K dataset

In order to assess the accuracy of our experiment, we employ the BLEU evaluation metric. Introduced by Papineni et al. [2], BLEU [2] has become a widely adopted evaluation metric in various NLP and CV applications, including machine translation and image captioning. BLEU [2] compares the number of n-gram sequences in the generated sentence with those in the reference sentence to calculate a BLEU score. Different BLEU scores are computed for different n-gram sizes, such as unigram BLEU-1, bigram BLEU-2, trigram BLEU-3, and so on. By utilizing BLEU, we can quantitatively measure the similarity between the generated captions and the reference captions, providing insights into the accuracy and quality of the translations.

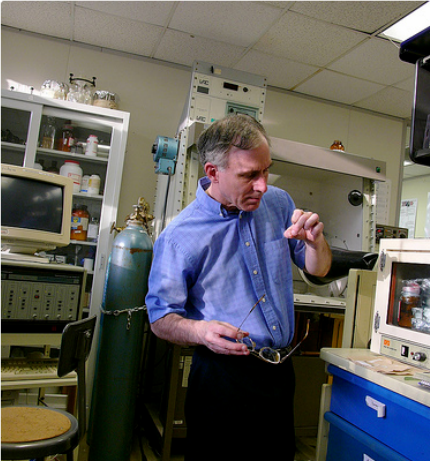
Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4
EncoderDecoder with Google Translate	93.51	69.52	50.34	28.67
EncoderDecoder with T5 Small	81.85	57.43	49.20	26.45
EncoderDecoder with T5 Base	<b>94.87</b>	<b>71.83</b>	<b>52.39</b>	<b>29.81</b>

**Table 4.2:** Evaluation of Image Captioning Models using different translation methods

The BLEU evaluation measure was used to assess the image captions generated by our proposed approach. Table 4.2 depicts a comparison between the proposed model with the different translations. The EncoderDecoder model with Google Translate achieved high scores of 93.51, 69.52, 50.34, and 28.67 for BLEU-1, BLEU-2, BLEU-3, and BLEU-4, respectively. The EncoderDecoder model with T5 Small obtained slightly lower scores, with 81.85, 57.43, 49.20, and 26.45 for BLEU-1, BLEU-2, BLEU-3, and BLEU-4. However, the EncoderDecoder model with T5 Base outperformed the other models with scores of 94.87, 71.83, 52.39, and 29.81 for BLEU-1, BLEU-2, BLEU-3, and BLEU-4, respectively.

#### 4.2.1 Human Evaluation using a Webform consisting of images with their corresponding captions

Image Caption Comparison



**ENGLISH:**

A gray and black-haired male is holding his glasses in one hand while looking at something in the other hand; surrounded by numerous amounts of machines.

**FRENCH:**

Un homme aux cheveux gris et noirs tient ses lunettes dans une main tout en regardant quelque chose dans l'autre main entouré de nombreuses machines.

Yes  No

Please provide the correct caption:

Enter correct caption

Submit

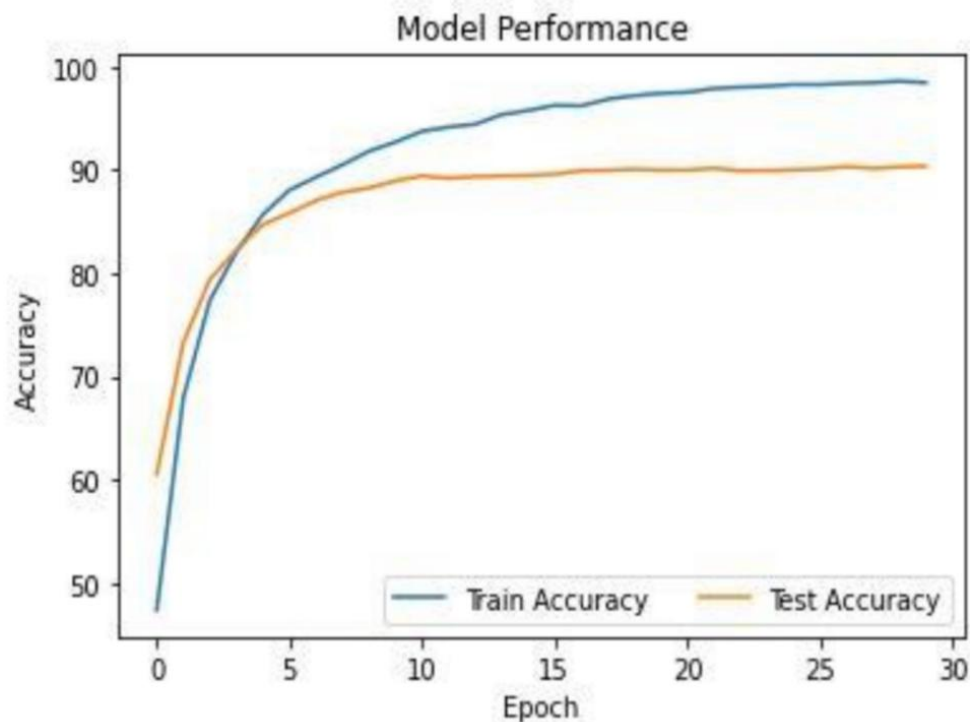
**Figure 4.1:** Webform consisting of images with their corresponding captions



The purpose of the webform, Figure 4.1 is to conduct a human evaluation of image captioning models. The evaluation provides valuable insights into the quality of the generated captions, which can be compared to the model's performance metrics like BLEU, and METEOR. The webform consists of an image displayed along with its corresponding captions. Each image-caption pair is presented together to ensure evaluators have a clear understanding of the context. Clear instructions or guidelines are provided to evaluators before they start the evaluation process. Depending on the evaluation criteria, we included a rating system (e.g., 'yes' or 'no') for evaluators to rate the quality or relevance of the captions either good or bad. Alternatively, if a 'no' is chosen, a text input box was provided for evaluators to provide qualitative correct captions. Responses provided by raters can be collected and stored in a database or saved as a file for further analysis. It is important to ensure the privacy and confidentiality of reviewer data. Once the evaluation is complete, the collected data can be analyzed to assess the performance of the image captioning models.

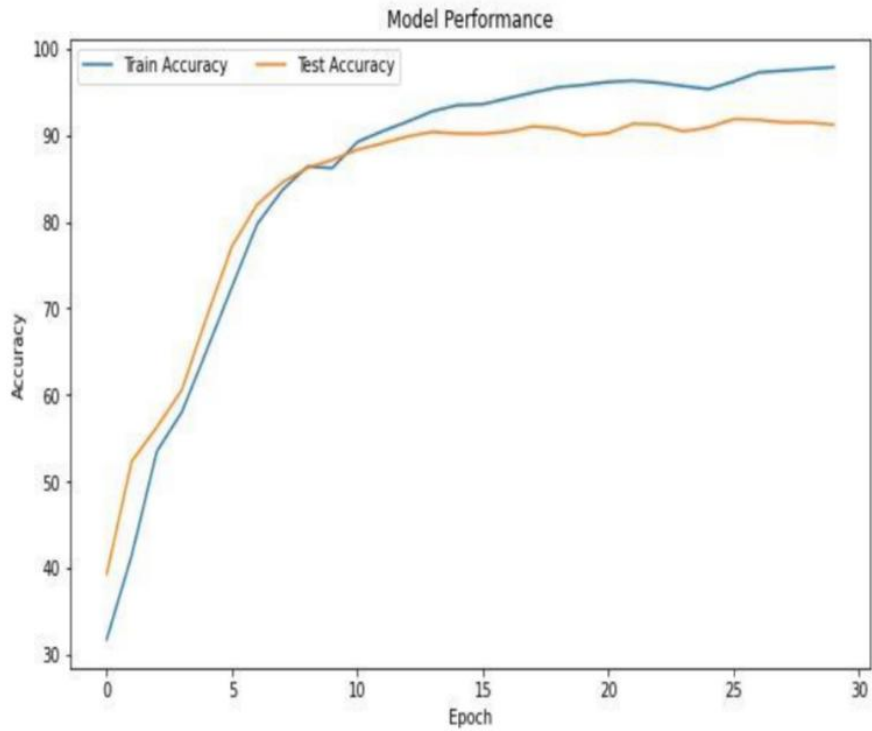
#### 4.2.2 Model Performance

The accuracy curve of the EncoderDecoder model (consisting of ResNet50+LSTM+Attention), is shown below respectively in Figure 4.2.

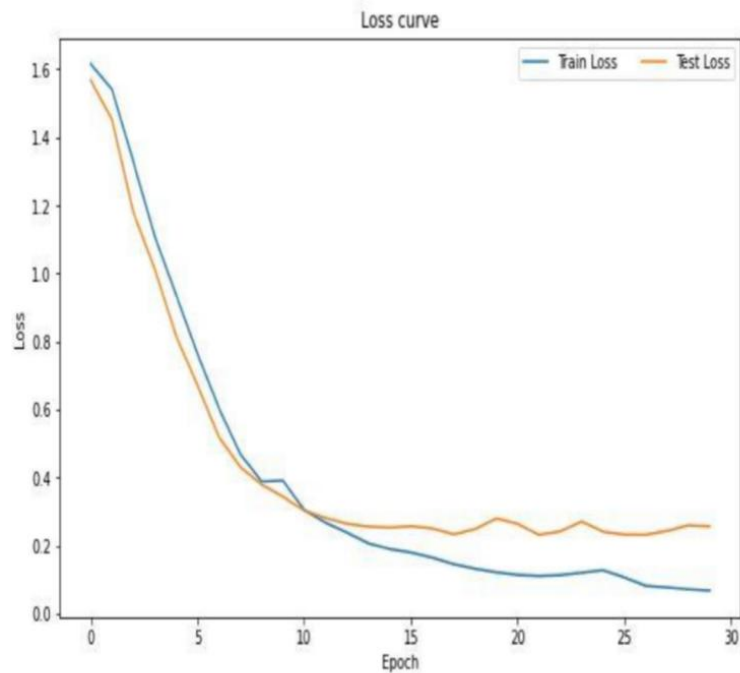


**Figure 4.2:** Performance with T5 Base method

From the graph in Figure 4.2 of the accuracy curve EncoderDecoder with T5 Base, we can clearly observe that starting from epoch 1 up to epoch 5, both the train and test accuracy are increasing exponentially until they reached around 85% accuracy. After epoch 5, the test accuracy started moving slower than the training accuracy until it reached a constant accuracy of around 90% starting from epoch 15 and then onward.



**Figure 4.3:** Performance with Google Translate



**Figure 4.4:** Loss Curve with Google Translate

The accuracy curve, as well as the loss Curve, are shown respectively in Figure 4.3 and Figure 4.4 below: From the graph of the accuracy curve, we can observe that starting from epoch 1 up to epoch 10, both the train and test accuracy are increasing exponentially until they reached around 87% accuracy. After epoch 10, the test accuracy started moving slower than the training accuracy until it reached a constant accuracy of around 90% starting from epoch 15 onward.

### 4.3 Comparison with existing captioning datasets

When comparing the models below it is important to consider the dataset size and diversity. In Table 4.3, the larger and more diverse MS COCO dataset [5] generally provides a richer training data distribution, which may contribute to better performance. The models trained on MS COCO (Ja-generator [17] and monolingual Japanese [14]) achieve moderate scores across the BLEU metrics. On the other hand, the EncoderDecoder with Google Translate method, despite being trained on a relatively minor portion of the FLickr30k dataset, achieves high BLEU scores, indicating strong performance. The EncoderDecoder + T5 Small model, trained on the Flickr30k dataset, performs reasonably well but has lower BLEU scores than the other models.

The comparison highlights the importance of dataset size and diversity in captioning tasks. The models trained on larger and more diverse datasets tend to perform

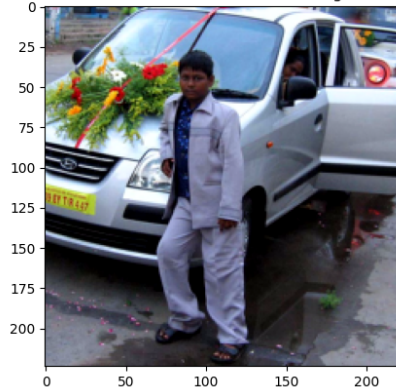
Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4
Ja-generator on MS-COCO [3]	76.3	61.4	49.2	<b>38.5</b>
monolingual Japanese on MS-COCO [14]	71.5	57.3	46.8	37.9
EncoderDecoder with Google Translate	93.51	69.52	50.34	28.67
EncoderDecoder with T5 Small	81.85	57.43	49.20	26.45
EncoderDecoder with T5 Base	<b>94.87</b>	<b>71.83</b>	<b>52.39</b>	29.81

**Table 4.3:** Evaluation of Image Captioning Models on different Dataset

better, although models with limited training data can still achieve competitive results.

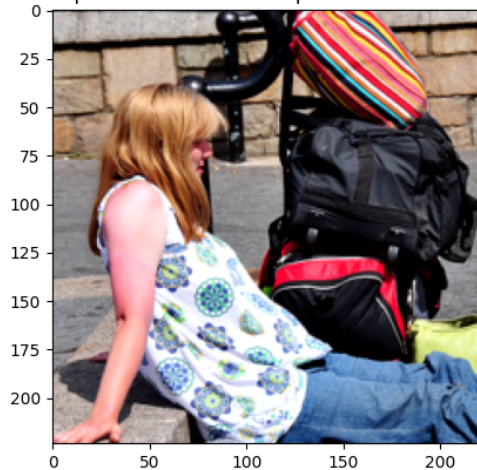
### Examples of generated image captions from our model.

un jeune garçon en costume blanc se tient devant une voiture blanche avec de gros bouquets de fleurs attachés au toit et au capot .

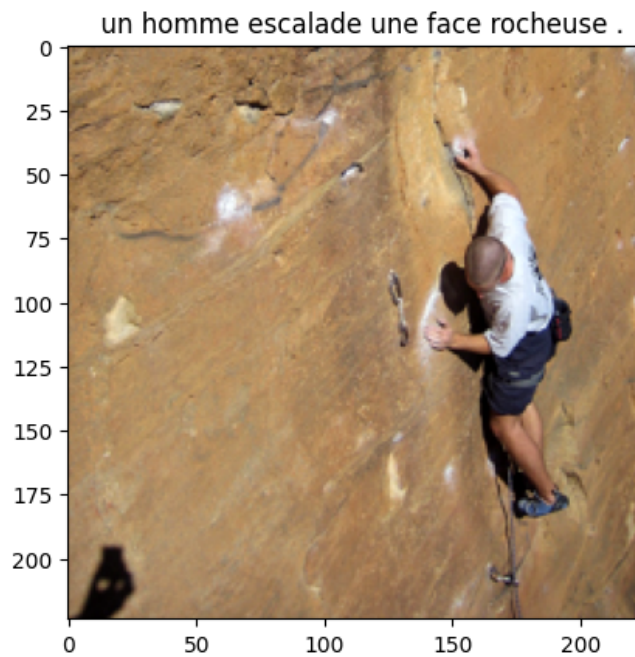


**Figure 4.5:** EncoderDecoder with T5 Base

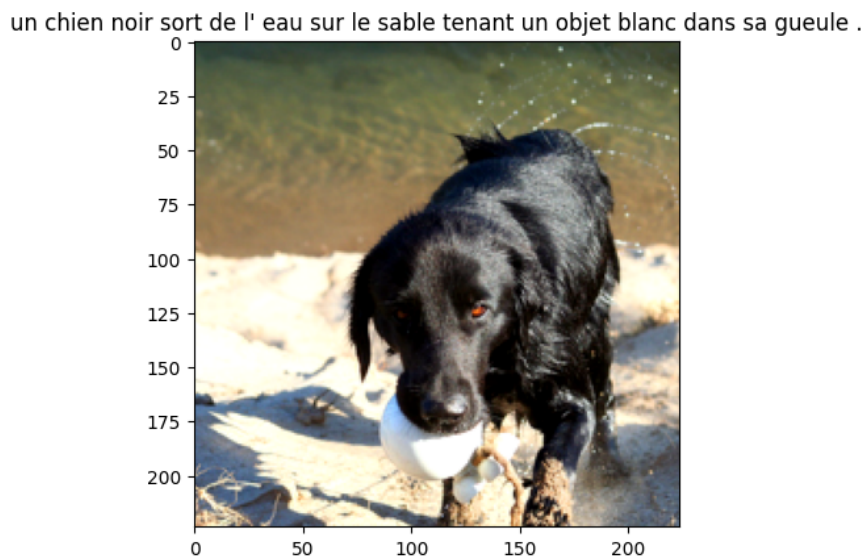
une femme blanche blonde brûlée par le soleil est assise par terre à l' extérieur près de trois sacs polochons .



**Figure 4.6:** EncoderDecoder with T5 Small



**Figure 4.7:** EncoderDecoder with T5 Base



**Figure 4.8:** EncoderDecoder with Google Translate

### 4.3.1 Analysis of translation and captioning errors and limitations

Machine translation is an important tool in the field of natural language processing, and it has been used to improve image captions in French and Arabic. However, there are still some errors and limitations that need to be addressed.

- **Vocabulary and terminology:** Machine translation systems may struggle with accurately translating specialized vocabulary and technical terms.

- **Ambiguity and polysemy:** Translation ambiguity can arise from words or phrases with multiple meanings. Identifying errors related to ambiguity helps in refining translation models and introducing context-aware techniques.
- **Syntax and grammar:** Maintaining the syntactic structure and grammatical rules of the source language in the target language can be challenging. Analyzing errors related to syntax and grammar can guide the improvement of translation models and adherence to target language rules.
- **Cultural and contextual factors:** Translating cultural references, humor, and context-dependent expressions accurately is necessary and crucial. Identifying errors in handling cultural and contextual factors can help develop approaches for cultural adaptation and context-aware translation.

#### 4.4 Research Solutions

- What are the best methods for improving French and Arabic image captions using machine translation?

The best methods for improving French and Arabic image captions using machine translation can vary depending on the specific task and dataset. However, based on the results, EncoderDecoder with T5 base or T5 small shows promising performance in improving image captions, achieving high BLEU [2] scores in both languages. This suggests that utilizing a pretrained translation model like Google Translate [2] can be an effective method for improving the quality of captions in French and Arabic.

- How can we effectively construct a French and Arabic Flickr30k Image Caption Dataset?

Constructing an effective French and Arabic Flickr30k Image Caption Dataset involves collecting a large number of diverse images along with their corresponding captions in the target languages. The dataset should cover various topics, scenes, and linguistic variations. To ensure high quality, manual annotation and verification of the captions by native speakers are crucial. Additionally, considering the cultural and linguistic nuances specific to French and Arabic languages is essential to create an accurate and comprehensive dataset.

- What are the advantages and disadvantages of using machine translation to improve French and Arabic image captions?

**Advantages:**

- Machine translation can provide a quick and efficient way to generate captions multilingually.
- It can help bridge the language barrier and enable a wider audience to access and understand image content.
- Using machine translation allows for automation and scalability in generating captions for large datasets.

**Disadvantages:**

- Machine translation may produce inaccurate or unnatural translations, especially when dealing with complex or ambiguous language constructs.
- Machine translation may lack context and understanding of visual elements, resulting in captions that do not fully capture the essence of the image.

- How can we evaluate the fluency and accuracy of machine-translated French and Arabic image captions, and what are the most appropriate metrics for this task?

‘ The fluency and accuracy of machine-translated French and Arabic image captions can be evaluated through various metrics, such as BLEU [2](Bilingual Evaluation Understudy) scores as used in the discussed results. BLEU measures the n-gram overlap between the machine-translated captions and reference captions. However, it is essential to note that BLEU scores alone may not capture the full picture of translation quality. Additional metrics, such as METEOR, ROUGE, or human evaluation through annotation studies, can provide more comprehensive assessments by considering semantic meaning, linguistic fluency, and alignment with the image content.

# Chapter 5

## Conclusion

### 5.1 Summary

In conclusion, this study investigated the effectiveness of machine translation in improving French and Arabic image captions. The results revealed several key findings. Firstly, the EncoderDecoder with the T5 base model exhibited the highest BLEU scores in both languages, indicating its ability to generate accurate translations. However, it was noted that the performance of the models varied depending on the dataset size, with the MS COCO dataset, used in conjunction with the Ja-generator [3] and monolingual Japanese models [21], being larger compared to the Flickr30k dataset [4] used with the T5 Small [27] and Google Translate [2] models. With its valuable insights into machine translation for enhancing French and Arabic image captions. The findings emphasize the significance of dataset construction, present comparative results, and suggest future research directions. By addressing these research objectives and questions, the study advances the field of image captioning and machine translation in multilingual settings. This study makes significant contributions to image captioning and machine translation. Evaluating different models on diverse datasets provides insights into their performance and effectiveness in improving French and Arabic image captions. Additionally, the comparison between the MS COCO and Flickr30k datasets sheds light on the impact of dataset size on translation performance. These findings can guide future research and development efforts in the field.

### 5.2 Future Works

It is suggested to conduct experiments and evaluations on larger and more diverse datasets( Dataset expansion like MS coco [5]) in French and Arabic to gain a deeper understanding of the machine translation models' capabilities. Exploring alternative evaluation metrics, such as METEOR or human annotation studies, would provide a



more comprehensive assessment of translation quality beyond the BLEU scores. Exploring larger pre-trained models like T5 large would help gain valuable insights into image captioning and machine translation

## REFERENCES

- [1] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, “Show and tell: A neural image caption generator,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3156–3164.
- [2] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, “Bleu: a method for automatic evaluation of machine translation,” in *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*. Philadelphia, Pennsylvania, USA: Association for Computational Linguistics, Jul. 2002, pp. 311–318. [Online]. Available: <https://aclanthology.org/P02-1040>
- [3] T. Yao, Y. Pan, Y. Li, Z. Qiu, and T. Mei, “Boosting image captioning with attributes,” in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 4894–4902.
- [4] B. A. Plummer, L. Wang, C. M. Cervantes, J. C. Caicedo, J. Hockenmaier, and S. Lazebnik, “Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models,” in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 2641–2649.
- [5] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, “Microsoft coco: Common objects in context,” in *Computer Vision—ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*. Springer, 2014, pp. 740–755.
- [6] S. Katiyar and S. K. Borgohain, “Comparative evaluation of cnn architectures for image caption generation,” *arXiv preprint arXiv:2102.11506*, 2021.
- [7] Z. Yang and N. Okazaki, “Image caption generation for news articles,” in *Proceedings of the 28th International Conference on Computational Linguistics*, 2020, pp. 1941–1951.
- [8] C. Chen, S. Mu, W. Xiao, Z. Ye, L. Wu, and Q. Ju, “Improving image captioning with conditional generative adversarial nets,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, 2019, pp. 8142–8150.

- [9] H. R. Tavakoli, R. Shetty, A. Borji, and J. Laaksonen, “Paying attention to descriptions generated by image captioning models,” in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 2487–2496.
- [10] Y. Huang, B. Liu, J. Fu, and Y. Lu, “A picture is worth a thousand words: A unified system for diverse captions and rich images generation,” in *Proceedings of the 29th ACM International Conference on Multimedia*, 2021, pp. 2792–2794.
- [11] M. Tanti, A. Gatt, and K. P. Camilleri, “What is the role of recurrent neural networks (rnns) in an image caption generator?” *arXiv preprint arXiv:1708.02043*, 2017.
- [12] S. Katiyar and S. K. Borgohain, “Analysis of convolutional decoder for image caption generation,” *arXiv preprint arXiv:2103.04914*, 2021.
- [13] R. Bernardi, R. Cakici, D. Elliott, A. Erdem, E. Erdem, N. Ikizler-Cinbis, F. Keller, A. Muscat, and B. Plank, “Automatic description generation from images: A survey of models, datasets, and evaluation measures,” *Journal of Artificial Intelligence Research*, vol. 55, pp. 409–442, 2016.
- [14] T. Miyazaki and N. Shimizu, “Cross-lingual image caption generation,” in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2016, pp. 1780–1790.
- [15] M. Nikolaus, M. Abdou, M. Lamm, R. Aralikkatte, and D. Elliott, “Compositional generalization in image captioning,” *arXiv preprint arXiv:1909.04402*, 2019.
- [16] S. J. Rennie, E. Marcheret, Y. Mroueh, J. Ross, and V. Goel, “Self-critical sequence training for image captioning,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 7008–7024.
- [17] Y. Yoshikawa, Y. Shigeto, and A. Takeuchi, “Stair captions: Constructing a large-scale japanese image caption dataset,” *arXiv preprint arXiv:1705.00823*, 2017.
- [18] V. Jindal, “Generating image captions in arabic using root-word based recurrent neural networks and deep neural networks,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, no. 1, 2018.
- [19] R. Biswas, M. Barz, M. Hartmann, and D. Sonntag, “Improving german image captions using machine translation and transfer learning,” in *Statistical Language and Speech Processing: 9th International Conference, SLSP 2021, Cardiff, UK, November 23–25, 2021, Proceedings 9*. Springer, 2021, pp. 3–14.
- [20] W. Zhao, B. Wang, J. Ye, M. Yang, Z. Zhao, R. Luo, and Y. Qiao, “A multi-task learning approach for image captioning.” in *IJCAI*, 2018, pp. 1205–1211.

- [21] S. Kwon, B.-H. Go, and J.-H. Lee, “A text-based visual context modulation neural model for multimodal machine translation,” *Pattern Recognition Letters*, vol. 136, pp. 212–218, 2020.
- [22] D. Elliott, S. Frank, and E. Hasler, “Multilingual image description with neural sequence models,” *arXiv preprint arXiv:1510.04709*, 2015.
- [23] A. Mastropaolo, S. Scalabrino, N. Cooper, D. N. Palacio, D. Poshyvanyk, R. Oliveto, and G. Bavota, “Studying the usage of text-to-text transfer transformer to support code-related tasks,” in *2021 IEEE/ACM 43rd International Conference on Software Engineering (ICSE)*. IEEE, 2021, pp. 336–347.
- [24] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, “Attention is all you need,” *Advances in neural information processing systems*, vol. 30, 2017.
- [25] Y. Wu, M. Schuster, Z. Chen, Q. V. Le, M. Norouzi, W. Macherey, M. Krikun, Y. Cao, Q. Gao, K. Macherey *et al.*, “Google’s neural machine translation system: Bridging the gap between human and machine translation,” *arXiv preprint arXiv:1609.08144*, 2016.
- [26] A. Farhadi, M. Hejrati, M. A. Sadeghi, P. Young, C. Rashtchian, J. Hockenmaier, and D. Forsyth, “Every picture tells a story: Generating sentences from images,” in *Computer Vision—ECCV 2010: 11th European Conference on Computer Vision, Heraklion, Crete, Greece, September 5–11, 2010, Proceedings, Part IV 11*. Springer, 2010, pp. 15–29.
- [27] A. Roberts, C. Raffel, K. Lee, M. Matena, N. Shazeer, P. J. Liu, S. Narang, W. Li, and Y. Zhou, “Exploring the limits of transfer learning with a unified text-to-text transformer,” 2019.
- [28] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, “Exploring the limits of transfer learning with a unified text-to-text transformer,” *The Journal of Machine Learning Research*, vol. 21, no. 1, pp. 5485–5551, 2020.
- [29] B. Dzmitry and B. Yoshua, “Neural machine translation by jointly learning to align and translate,” in *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7–9, 2015, Conference Track Proceedings*, 2015.