Abstract

Simultaneous Machine Translation (SiMT) is a subfield of Natural Language Processing that focuses on translating continuous streams of sentences from one language to another with the lowest latency and highest quality possible. Previous works and empirical research have shown that having supplementary information can improve the performance of a language model. The main hypothesis for this project was that having supplementary information from an image will make up for the missing textual information. Language barriers can have multiple implications in fields like healthcare, education, aviation and multilateral negotiations which is why the development of such a system is required. The project has been using the Multi30K dataset and pysimt library for training and evaluation. The first phase of the project was completed in January and an improvement of roughly 2 BLEU points was been achieved with the help of visual context. The next phase of the project looked into several methods that can be used to improve the performance of the SiMT system. On analyzing the Multi30K dataset, other online available datasets as well as recent publications, I came to the conclusion that creating a new dataset for the task of Simultaneous Machine Translation was the best option. Therefore, a new dataset was created by scraping the video from TED Talks. In the end, the dataset contains 4024 videos with subtitles available in several different languages. The dataset is roughly 118GB in size and using a segmentation algorithm, this model can be run on the current SiMT model. The main hypothesis for this part is that TED Talks that contain slides, animations or other resources for explaining the content will provide valuable information in the form of visual features that can be extracted using a ResNet CNN which will improve the performance of the simultaneous machine translation system.
Acknowledgements

I would like to thank Dr Kong Lingpeng, for his support and advice regarding the techniques which could be applied to improve the results achieved in the paper. His input allowed me to appropriately define the project scope and focus my research in a direction with the most impact. I would also like to thank Zhiyong Wu, for providing me with valuable feedback on my work. Furthermore, I would like to extend my gratitude to Dr Locky Law for giving me a lot of helpful suggestions for the project reports. Lastly, I would like to thank the Department of Computer Science at The University of Hong Kong for providing me with the resources required to work on this project.
Table of Contents

Acknowledgements ii
List of Figures iv
List of Tables v
List of Abbreviations vii

1 Introduction 1
   1.1 Background .................................................. 2
   1.2 Motivation ................................................... 4
   1.3 Objectives ................................................... 4
   1.4 Scope ........................................................ 5
   1.5 Report Outline ................................................ 5

2 Literature review 7
   2.1 Machine Translation ............................................ 7
   2.2 History of Machine Translation ............................... 7
   2.3 Multimodal Machine Translation (MMT) ....................... 8
   2.4 Simultaneous Neural Machine Translation (S-NMT) ........... 9
   2.5 Simultaneous Machine Translation with Visual Context .... 10

3 Methodology 11
   3.1 Dataset ....................................................... 11
      3.1.1 Preprocessing ............................................ 12
      3.1.2 Train/Test Split ......................................... 12
   3.2 Baseline Model ................................................ 13
   3.3 Incorporating Visual Modality ................................. 13
      3.3.1 Object Classification (OC) features ...................... 13
      3.3.2 Object Detection (OD) features ........................ 14
   3.4 Model Architectures .......................................... 14
      3.4.1 Decoder attention (DEC-OC/OD) ......................... 14
      3.4.2 Multimodal encoder (ENC-OD) .......................... 14
3.5 Simultaneous Machine Translation Approaches .................................. 15
3.6 Quality and Latency metrics .............................................................. 15
   3.6.1 BLEU Score ............................................................................. 15
   3.6.2 Average Proportion (AP) ............................................................ 16
   3.6.3 Average Lagging (AL) ............................................................... 16
   3.6.4 Consecutive Wait (CW) .............................................................. 16

4 Project Progress ................................................................................. 17
   4.1 Current Progress ............................................................................ 17
   4.2 Phase I - Reproducing the Original Results ..................................... 18
      4.2.1 English→German .................................................................. 18
      4.2.2 English→French .................................................................... 20
      4.2.3 Key Observations .................................................................. 21
      4.2.4 Reproducibility ..................................................................... 21
   4.3 Phase II - Research ......................................................................... 22
   4.4 CLIP Model - Contrastive Language–Image Pre-training ................. 22
   4.5 Analysis of Available Datasets ......................................................... 24
      4.5.1 Multi30K Dataset .................................................................. 24
      4.5.2 Available Datasets .................................................................. 25
   4.6 TED Talk Video Subtitle Dataset ..................................................... 26
      4.6.1 Procedure for Data Extraction ................................................ 27
      4.6.2 Dataset Samples .................................................................... 28
      4.6.3 Dataset Statistics ................................................................... 30
   4.7 Next Step ....................................................................................... 31
   4.8 Schedule ....................................................................................... 32
   4.9 Challenges .................................................................................... 33
   4.10 Limitations .................................................................................. 34
   4.11 Future Work ................................................................................ 34

5 Conclusion ......................................................................................... 35
List of Figures

1.1 Consecutive Machine Translation ................................................................. 3
1.2 Simultaneous Machine Translation ............................................................... 3

2.1 Example of Machine Translation (English → Swahili) ................................. 7
2.2 Multimodal machine translation model .......................................................... 8

3.1 Sample image and captions from Multi30K dataset ........................................ 12
3.2 Decoder attention model equations ............................................................... 14

4.1 Comparison of wait-k models (en-de) ............................................................ 19
4.2 Comparison of wait-k models (en-fr) ............................................................ 20
4.3 Performance of CLIP is either similar to or better than ResNet [24] ............... 23
4.4 Contrastive pre-training - CLIP pre-trains an image encoder and a text encoder
   to predict which images were paired with which texts in our dataset [24] ......... 24
4.5 Sample images from Multi30k dataset [15] ................................................... 25
4.6 Extracting the links from the webpage ........................................................ 27
4.7 Samples of 3 frames from the TED Talk “The next outbreak? We’re not ready”
   with subtitles in English, German and French [27]. Different frames provide
   different amounts of context. For example, frame 1 does not provide any useful
   context for translation while frame 3 contains the entire text in English on the
   slide. Frame 2 lies in between frames 1 and 3. .............................................. 29
List of Tables

4.1 Simultaneous MMT (en-de) ................................................. 18
4.2 Simultaneous NMT (en-de) ................................................. 18
4.3 Multimodal gains for wait-k models (en-de) ......................... 19
4.4 Simultaneous MMT (en-fr) ................................................. 20
4.5 Simultaneous NMT (en-fr) ................................................. 20
4.6 Multimodal gains for wait-k models (en-fr) ......................... 21
4.7 Number of videos available for different languages in descending order ............................................. 30
4.8 Vocabulary size of different languages .................................. 31
4.9 Subset of the sample vocabulary in different languages ............. 31
4.10 Project Schedule .............................................................. 32
List of Abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>MT</td>
<td>Machine Translation</td>
</tr>
<tr>
<td>MMT</td>
<td>Multimodal Machine Translation</td>
</tr>
<tr>
<td>NMT</td>
<td>Neural Machine Translation</td>
</tr>
<tr>
<td>SiMT</td>
<td>Simultaneous Machine Translation</td>
</tr>
<tr>
<td>S-NMT</td>
<td>Simultaneous Neural Machine Translation</td>
</tr>
<tr>
<td>SOV</td>
<td>Subject Object Verb</td>
</tr>
<tr>
<td>SVO</td>
<td>Subject Verb Object</td>
</tr>
<tr>
<td>BLEU</td>
<td>BiLingual Evaluation Understudy</td>
</tr>
<tr>
<td>DE</td>
<td>German</td>
</tr>
<tr>
<td>EN</td>
<td>English</td>
</tr>
<tr>
<td>FR</td>
<td>French</td>
</tr>
</tbody>
</table>
1. Introduction

Machine Translation (MT) is the use of computers for translating text from one language to another without human involvement. Simultaneous Machine Translation (SiMT) aims to reproduce human interpretation, where an interpreter translates spoken utterances as they occur [1]. The interpreter needs to dynamically find the balance between how much context is required to produce the translation reliably and how long the listener has to wait for the translation [1].

Consecutive machine translation systems where source sentences are available in their entirety before translation begins have existed for years. While they seem to provide a seamless experience with accurate translation, one major drawback with this approach is the latency incurred during the translation. In contrast, SiMT aims to translate a continuous input text stream into another language. Different from consecutive machine translation in which only the translation quality matters, the challenge with SiMT is to design a strategy that can find a good trade-off between the quality of the translation and the latency incurred while producing it [2].

Simultaneous translation is more complex due to the differences in syntactic structure between the source and target languages [2]. For example, when simultaneously translating from a verb-final (SOV) language (e.g., German or Japanese) to a verb-media (SVO) language (e.g., English or Chinese), the verb appears much later in the source sequence than in the target language. Due to this, premature translations of source sentences can lead to a significant loss in quality [3].

Previous research has considered several unimodal approaches to overcome the limitations of SiMT mentioned above. However, to tackle these challenges, this project aims to perform SiMT in a multimodal setting and improve the performance of the current state of the
art approach. In addition to the source sentence, the model will be provided with visual information in the form of an image. Previous research has shown that having access to supplementary information should help the models anticipate the missing context by grounding their decisions about ‘when’ and ‘what’ to translate [1].

The remaining part of the introduction section includes detailed background, motivation, the objective of this project, the scope followed by an outline of this report.

1.1 Background

Language is substantially more than the communication of words. It represents people’s culture, society, and beliefs. The need for translation has a growing demand as today’s multicultural and multilingual society demands effective communication between languages and cultures. Due to globalization and the development of technology, the number of international tourist arrivals has approximately doubled from the year 2000 to 2019 [4]. Even more people from different countries, using different languages are interacting with each other on the internet.

Language barriers can have multiple implications in fields like healthcare, education, aviation, and multilateral negotiations and can also affect collaboration between multinational teams at large corporations [5]. Even in the current era of globalization and communication, language continues to remain a barrier in conveying our messages to other people. Thus, an application that can automate simultaneous interpretation can have numerous applications.

While there are a number of tools available on the market such as Google Translate, Baidu Translate etc. to help people eliminate the language barrier, these tools perform consecutive machine translation. (see Figure 1.1). Consecutive machine translation is not ideal for real-world applications where the live translation is needed as there is a significant delay in the translation process. This happens because the system needs to wait for the entire sentence to finish before translating. Apart from this, the translation itself may take some time.
In contrast, Simultaneous Machine Translation (SiMT) is still under development. There is a noticeable gap between the performance of human interpreters and simultaneous machine translation models. The idea behind Simultaneous Machine Translation is for the system to progressively read a stream of input data and provide real-time translation (see Figure 1.2). Since the input is being fed continuously into the model, there is a need to anticipate the next word \[1\]. Human translators can predict this context through several factors such as their linguistic ability and experience. Besides, the human brain, unlike machine learning models, can recognize expressions, tone of voice, visual context etc. efficiently. Therefore in this project, we aim to build a similar context for the machine translation model using visual features extracted from images.
1.2 Motivation

The impact of translation on our everyday lives is more multidimensional than we realize. It enables a global economy and it is necessary for the spread of information, knowledge, and ideas [6]. When there’s a demand for translation, there are opportunities for translators. However, dynamic translation is quite challenging and exhausting for humans and there are a limited number of skilled simultaneous interpreters in the world. On top of that, each person only lasts for about 15-30 minutes after which the rate at which they make errors in their translations increases exponentially [7].

Hence, an assessment of the current situation leads to the conclusion that there is a critical need for a system that can translate text from one language to another, in real-time. It can help reduce the burden on human interpreters and make machine translation more accessible and affordable.

1.3 Objectives

The core objective of this project is to conduct research on the already existing frameworks and state of the art models to perform simultaneous machine translation and then suggest ways to either improve the performance of the current models or develop new ways to solve this challenge. The project also focuses on analysing the effect of multimodality on the quality of translations by using images as the auxiliary data source. Multimodality is the integration of information from many supplementary sources such as sound, images, video, music, and so on. Multimodality is central to understanding how computers learn languages and may learn them in a better way [8].

The project plans to combine the methods and techniques used for Multimodal Machine Translation (MMT) and simultaneous machine translation. Therefore, it ultimately aims to understand and implement the existing SiMT models that use visual aid and then improve upon the theoretical aspects of the model. The purpose of this project is to gain an in-depth understanding of the current SiMT model so as to be able to suggest some modifications.
that can improve the performance or discover new methods which can be used to tackle this challenge. Several previously experimented approaches like rule-based strategies, reinforcement learning and supervised learning will also be inspected. As mentioned earlier, the two main criteria to evaluate the performance of a SiMT model are translation quality and latency. To evaluate the translation quality, this project uses the bilingual evaluation understudy (BLEU) score as opposed to other metrics such as METEOR, NIST, GLEU etc. For measuring the latency, this project like the original paper uses Average Proportion (AP) and Average Lagging (AL) [1].

1.4 Scope

The objectives of this project have been explained in detail above. Multimodal Machine translation can be performed using additional information from a variety of sources like sound, music, video etc. but in this project, only visual context extracted from images will be considered. While the original paper refers to several types of models for performing SiMT, the scope of this project is limited to only the models actually implemented in the paper whose results are reported. The project progress can be checked on the website: [https://wp.cs.hku.hk/2021/fyp21071/](https://wp.cs.hku.hk/2021/fyp21071/)

1.5 Report Outline

The report begins by introducing the concept of Simultaneous Machine Translation as well as demonstrating the need to build such a system. It introduces the project objective and the scope of the project in detail. The rest of the report is structured in this way: Section 2 provides an insight into the history of machine translation and the related work that has been done in this field. Different techniques have been explained in this section. Section 3 presents the methodology followed while developing this project, which covers the data preprocessing, model architectures, implementation and evaluation metrics.

Section 4 presents the current status of the project including what has been implemented. This section also reports the results that have been obtained using the models mentioned in the paper. After reporting the results, we move on and report the ways which can be
used to develop upon the current research. Two main approaches have been identified and the second approach, which is the creation of a new dataset was prioritized. Section 4 also talks about the limitations of the current datasets, analyzes the need for the creation of a new dataset and then informs the procedures followed for the development of this dataset. The challenges faced, limitations of the new approach and the future work which will be done before submitting the final paper to the EMNLP (Empirical Methods in Natural Language Processing) conference are also mentioned along with the timeline. Finally, section 5 concludes this final year project report.
2. Literature review

This section presents the literature related to the field of machine translation field. Different approaches relevant to this project like Multimodal Machine Translation (MMT), Simultaneous Neural Machine Translation (S-NMT) and Simultaneous Machine Translation with Visual Context (SiMT) are studied and summarized in this section.

2.1 Machine Translation

Machine Translation is the task of automatically converting the source text in one language into another language without human intervention while preserving the meaning of the input sentence and producing fluent and grammatically correct text in the required output language. An example of machine translation is shown in Figure 2.1.

![Figure 2.1: Example of Machine Translation (English → Swahili)](image)

2.2 History of Machine Translation

Machine Translation has been in existence for a long time. Some of the first translations turned Sumerian poems into Asian languages about 4500 years ago [9][10]. In the 1954s, IBM introduced machine translation based on grammar rules [10]. However, machine translation based on grammar rules was quite limited as it needed broad dictionaries and the grammar rules had to be set manually. Then came Statistical machine translation based on models whose parameters were derived from the analysis of bilingual text corpus. As time went
by, the development of machine learning progressed and neural machine translation based on deep learning was born. The SiMT model in this project is based on neural machine translation.

2.3 Multimodal Machine Translation (MMT)

Multimodal machine translation or simply MMT aims to enhance the performance of a machine translation model by drawing information from additional modalities under the assumption that these different modalities will provide useful information to the model [11] (see Figure 2.2). As mentioned earlier, humans rely on a combination of visual, auditory, tactile and other stimuli being processed simultaneously to draw complex relationships which improve the quality of our perception [11]. Therefore, MMT aims to mimic this relationship by allowing the language model to take information from different sources.

![Figure 2.2: Multimodal machine translation model](image)

Since these models make use of information from supplementary sources, the various available models differ primarily in the techniques used by them for feature extraction and also, in deciding when and how these features should be used during translation. Generally, for images, these models rely on features extracted from state-of-the-art CNN models pre-trained on large-scale visual tasks. Similarly, the SiMT model makes use of the information extracted from images by a ResNet CNN.
Since one of the main objectives of a SiMT system is to reduce latency, minimizing the time required to extract the visual features from an image is very important. Therefore, the choice of the techniques for feature extraction can also make an impact on the performance of this system. The methods used for feature extraction from images can be broadly classified into two categories. First being, the models based on multimodal attention which uses an intra-sequential mapping between source and target representations [11]. These models incorporate visual features into different parts of both the encoder and the decoder [11]. The second approach makes use of a pooled attention-based neural network that incorporates features from both, visual and linguistic representations [12]. During translation, auxiliary information like images and text are not equally important. Noise can be created by encoding auxiliary information directly which also needs to be managed.

2.4 Simultaneous Neural Machine Translation (S-NMT)

Simultaneous Neural Machine Translation was first explored in 2016 by Cho and Esipova in a greedy decoding framework [13]. It is different from consecutive translation where only the translation quality matters. In simultaneous translation both, the quality and the latency matter. For minimizing delay while maximizing the quality, a simultaneous NMT must start generating symbols in the target language before the full source sentence is received [13].

This greedy approach is called “wait-k” and can be seen as a policy-driven translation algorithm. At each “state” or an “input”, the interpreter has 3 choices: whether to keep reading the next word, write the translation of this current word to the output or predict the next word based on the context gained from the “k” previous words. “k” refers to the number of words the interpreter reads before choosing whether to write or predict.

Thus, these approaches form a basis for our SiMT model. Essentially, we need to design a way in which we can decide at each time step, which action would yield the least latency without compromising the quality of the translation. We can also consider other strategies like heuristics or alignment-based approaches that could help in this decision-making process [1].
2.5 Simultaneous Machine Translation with Visual Context

This is the main idea on which this project is based [1]. Simultaneous translation with an additional modality like a visual feature can help in improving the accuracy of the model as it aids the model in filling the missing textual context.

Simultaneous Machine Translation (SiMT) aims to reproduce human interpretation. Humans interpreters can translate spoken utterances as they occur in real-time [1]. The interpreter can dynamically find the balance between how much context is required to produce the translation reliably and how long the listener needs to wait for the translation [1]. Thus, the translation begins with an incomplete text source which is being fed into the model progressively along with the visual context extracted from the images using a ResNet CNN. Simultaneous machine translation (SiMT) aims to translate a continuous input text stream into another language with the lowest latency and highest quality possible.
3. Methodology

This section explains the general procedures and technologies that have been followed during the development of phase one of the project. The project was divided into two phases for convenience to make sure it was completed successfully. The first phase of the project focused on reproducing the results achieved in the original paper [1]. For this phase, PyTorch’s library *pysimt* was used which is available on Github [14]. *Pysimt* is a PyTorch-based sequence-to-sequence framework geared towards the recent approaches in Simultaneous MT and is aimed at facilitating research in unimodal and multi-modal machine translation. The results achieved have been reported in Section 4.2.

The second phase of the project focused on research work. This phase was started after the submission of the interim report in February. The objective of the second phase of the project was to find ways to improve the current Simultaneous Machine Translation framework and build upon the existing work. The second phase of the project has been explained in Section 4.3. This section mainly discusses six major components. The dataset that was used to reproduce the original results, the baseline model, the method followed to incorporate visual modality in the baseline model, model architectures, SiMT approaches and the evaluation metrics used to measure the performance of the model.

3.1 Dataset

The Multi30K dataset was used for the development and evaluation of this project [15]. It is an extension of the Flickr30K dataset [16]. It contains 31,014 images and translations present in several different languages [15]. For this project, I used German (de) and French (fr) languages as the original paper also focused on these two languages. The qualitative results achieved for English→German and English→French are mentioned in Section 4.2 below. A sample image from the Multi30K dataset is shown in Figure 3.1.
3.1.1 Preprocessing

The project used Moses scripts [17] to lowercase, punctuation-normalise and tokenise the sentences with hyphen splitting. Word vocabularies were then created on the training subset of the dataset. The resulting English, German and French vocabularies contained 9.8K, 18K and 11K tokens, respectively.

3.1.2 Train/Test Split

The project followed a standard split of 29,000 instances for training and 1,014 for validation. Flickr and MS COCO datasets have been used for testing the model. The statistics for the English→German dataset are mentioned below for reference.

- Training Dataset: Multi30K (English-German)
  - (en) 29000 sentences, 377534 words, 13.0 words/sent
  - (de) 29000 sentences, 360706 words, 12.4 words/sent

- Validation Dataset: Multi30K (English-German)
  - (en) 1014 sentences, 13308 words, 13.1 words/sent
  - (de) 1014 sentences, 12828 words, 12.7 words/sent

- Test Datasets:
  - 2016_flickr, 2017_flickr, 2017_mscoco, 2018_flickr
3.2 Baseline Model

The baseline model for reproducing the original results used a Neural Machine Translation (NMT) model. Neural Machine Translation is different from SiMT as there is no need for anticipation since all the information regarding the source text will be available before the translation begins. Another difference between them is that NMT is missing the multimodal information extracted from images as it is made unavailable.

The consecutive baseline model for this project consists of a 2-layer Gated Recurrent Unit (GRU) encoder and a 2-layer conditional GRU decoder with attention mechanism [1]. The encoder can be thought of as a black box that converts input text into some abstract hidden representation. The decoder then uses this hidden representation to generate the output text. For SiMT the encoder is made unidirectional so that no information from the words present in the sentence later that the word currently entered into the model can be encoded in the hidden representation which was computed by the encoder. The source sentences are then read progressively from left to right.

3.3 Incorporating Visual Modality

Two kinds of visual features called Object Classification (OC) features and Object Detection (OD) features have been extracted from the images. These features are then added to the baseline encoder and decoder architectures which were explained above.

3.3.1 Object Classification (OC) features

These features represent the global image information extracted from the final layer of a ResNet-50 CNN [18]. The feature extracted image is represented by a feature tensor $V \in \mathbb{R}^{8 \times 8 \times 20}$. 

Methodology  13
3.3.2 Object Detection (OD) features

These features represent the explicit object information generated by the bottom up-top-down (BUTD) extractor which is a faster R-CNN/ResNet-101 object detector [19]. The feature extracted image is represented by a feature tensor $V \in R^{36 \times 2048}$.

3.4 Model Architectures

The two multimodal model architectures used for Simultaneous Machine Translation that encode the information from the visual features (Object Classification and Object Detection) extracted from a ResNet CNN and the text have been explained in this section.

3.4.1 Decoder attention (DEC-OC/OD)

Decoder attention has been a standard way of integrating visual modality into NMT. It works by applying secondary attention at each decoding timestep [12]. This approach was used to construct the baseline model for MMT. This method has been used with both, object classification and object detection features. The decoder attends to both the source hidden states and the visual features. They are added together to form the multimodal context vector.

$$c_t^T = \text{Attention}^T(H, \text{query}=d_t)$$
$$c_t^V = \text{Attention}^V(V, \text{query}=d_t)$$
$$c_t = c_t^T + c_t^V$$

Figure 3.2: Decoder attention model equations

3.4.2 Multimodal encoder (ENC-OD)

A multimodal encoder can be thought of as a black-box model that incorporates the source sentence representations and object detection features extracted from the image in the encoder itself. It is expected that these hidden representations containing the visual features from the beginning could be more appropriate for SiMT to fill in the missing context. Visual
features are first linearly projected to the dimension of textual representations H and then normalized.

### 3.5 Simultaneous Machine Translation Approaches

This section summarises the Simultaneous Machine Translation approaches used during phase one of the project. As mentioned in the original paper [1], the two approaches that have been implemented so far are:

- **Consecutive Baseline** - The model simply takes in the words progressively and translates them into the target language. The results achieved after running this model with and without taking in the visual features are mentioned in section 4.2.

- **Wait-k Policy** - This model reads exactly k words before it begins the translation [20]. The results achieved after running this model with and without taking in the visual features, with values of k = 1,2,3,5,7 are mentioned in section 4.2.

Both these approaches have been tested using English→German and English→French datasets and DEC-OC/OD and ENC-OD model architectures. The improvement in the performance between the model having visual features compared to the one without visual features is also reported in Section 4.2.

### 3.6 Quality and Latency metrics

As mentioned in section 1.3 there are two components of SiMT that need to be optimized. The first is quality (how accurate the output text is) and the second is latency (time delay between translation output and input text stream). The different quantitative metrics used to compare and optimize the different models have been explained below.

#### 3.6.1 BLEU Score

Bilingual Evaluation Understudy or BLEU score is one of the widely used metrics for measuring machine translation quality [21]. BLEU is a numerical measurement assigned to the translated
result by comparing it with several reference sentences previously set as ground truth. It can have a value ranging from 0 to 1. The translated sentence which is the same as the reference sentence achieves a score of 1 [21].

### 3.6.2 Average Proportion (AP)

Average Proportion computes a normalized score between 0 and 1, which represents the average number of source tokens required to commit a translation. A major drawback of this evaluation criterion is that it produces different scores for two samples with the same underlying latency but with different source and target sentence lengths.

### 3.6.3 Average Lagging (AL)

To fix the problem with Average Proportion, Average Lagging was introduced. It measures the number of tokens the output is lagging behind the input [20].

### 3.6.4 Consecutive Wait (CW)

Consecutive Wait is another latency metric that measures the number of source tokens read consecutively between generating two translated tokens [22]. Therefore, we want to develop a system having a low consecutive wait score which implies that less context will be needed for translations.
4. Project Progress

This section reports all the work that has been done so far as well as the work which will be done before submitting the final paper in the EMNLP conference in June 2022. Section 4.1 and 4.2 discuss the updates made during phase one: reproducing the results of the original paper. Section 4.3 provides updates on the work done during the second phase of the project. During the second phase, two ideas were explored which looked the most promising. In the end, it was decided that the development of a new dataset was the way going forward. Analysis of the current dataset and its limitations, the need for creating a new dataset and the methodology followed while creating the TED Talk dataset have been reported in this section. This section also reports the key milestones achieved, challenges faced so far, limitations of the new dataset, project schedule that was followed throughout the development of this project as well as the future work which can be done.

4.1 Current Progress

The project since its inception in August 2021 has been consistently making progress. Based on the methodology, there are two main phases in the project. The first phase of the project was to reproduce the results of the original paper. This part was completed in January 2022.

In October 2021, a comprehensive literature review was carried out for getting familiar with the existing technologies in the field of Simultaneous Machine Translation. The documentation available for the `pysimt` was also understood to become familiar with the codebase.

In November 2021, the data pipeline and the baseline model was implemented. In December 2021, the final SiMT model was completed and trained on the English→German dataset. In January 2021, the training on the English→French dataset was also completed. I used an Nvidia 2080Ti GPU available on the GPU farm provided by the Department of Computer
Science at the University for training and testing the models. An average model takes roughly around 2 hours to train. Therefore, over the past few months, the data pipeline and the baseline model was implemented and the initial findings have been reported in section 4.2.

The next phase of the project was to research on finding ways to improve the work being done in the field of Simultaneous Machine Translation. The second phase began in February 2022. Over the course of the past few months, several options were analyzed to improve the performance of Simultaneous Machine Translation systems. In the end, we decided to go ahead with creating a new dataset. More details about the second phase are mentioned in section 4.3.

4.2 Phase I - Reproducing the Original Results

The results achieved after training all the models have been reported in this section. The first section covers the English→German (en-de) dataset findings and the second section covers the English→French (en-fr) dataset findings. The third section goes over some key observations made using this data.

4.2.1 English→German

- Simultaneous MMT vs Simultaneous NMT

Tables 4.1 and 4.2 show that all 3 multimodal models performed better than the Neural MT model on the en-de dataset.

<table>
<thead>
<tr>
<th>Multimodal Model</th>
<th>BLEU Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>DEC-OC</td>
<td>36.76</td>
</tr>
<tr>
<td>DEC-OD</td>
<td>37.11</td>
</tr>
<tr>
<td>ENC-OD</td>
<td>37.01</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Baseline Model</th>
<th>BLEU Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>NMT</td>
<td>36.07</td>
</tr>
</tbody>
</table>

Table 4.1: Simultaneous MMT (en-de)

Table 4.2: Simultaneous NMT (en-de)
• Wait-k Policy for \( k = \{1, 2, 3, 5, 7\} \)

Figure 4.1 shows that the BLEU score increases as the value of \( k \) increases. This is expected because as the value of \( k \) gets larger, the model reads more words before it starts translating which in turn provides more context and increases the accuracy of the model.

![Comparison of wait-k models](image)

Figure 4.1: Comparison of wait-k models (en-de)

• Multimodal gains in BLEU for MMT wait-k models compared to NMT

Table 4.3 shows the gains achieved between MMT and NMT for different values of \( k \). As we can observe, the gains decrease as the value of \( k \) increases. This might be happening because, for large values of \( k \), the models lose their property of simultaneous MT as they have to wait for quite a few words before beginning the translation.

<table>
<thead>
<tr>
<th>( k )</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>5</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>NMT</td>
<td>20.49</td>
<td>28.3</td>
<td>32.51</td>
<td>36.10</td>
<td>36.42</td>
</tr>
<tr>
<td>DEC-OC</td>
<td>0.97↑</td>
<td>0.50↑</td>
<td>0.09↑</td>
<td>0.54↓</td>
<td>0.18↑</td>
</tr>
<tr>
<td>DEC-OD</td>
<td>1.82↑</td>
<td>0.70↑</td>
<td>0.47↑</td>
<td>0.41↑</td>
<td>0.79↑</td>
</tr>
<tr>
<td>ENC-OD</td>
<td>1.05↑</td>
<td>0.52↑</td>
<td>0.07↑</td>
<td>0.39↓</td>
<td>0.04↓</td>
</tr>
</tbody>
</table>

Table 4.3: Multimodal gains for wait-k models (en-de)
4.2.2 English→French

- Simultaneous MMT vs Simultaneous NMT

Tables 4.4 and 4.5 show that both the decoder based models performed better than the NMT model while the encoder based model performed worse on the en-fr dataset.

<table>
<thead>
<tr>
<th>Multimodal Model</th>
<th>BLEU Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>DEC-OC</td>
<td>56.79</td>
</tr>
<tr>
<td>DEC-OD</td>
<td>56.78</td>
</tr>
<tr>
<td>ENC-OD</td>
<td>55.41</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Baseline Model</th>
<th>BLEU Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>NMT</td>
<td>56.76</td>
</tr>
</tbody>
</table>

Table 4.4: Simultaneous MMT (en-fr)

Table 4.5: Simultaneous NMT (en-fr)

- Wait-k Policy for \(k = \{1,2,3,5,7\}\)

Figure 4.2 shows that the BLEU score increases as the value of \(k\) increases.

![Comparison of wait-k models](image)

Figure 4.2: Comparison of wait-k models (en-fr)
- Multimodal gains in BLEU for MMT wait-k models compared to NMT

Table 4.6 shows the gains achieved between MMT and NMT for different values of k.

<table>
<thead>
<tr>
<th>k</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>5</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>NMT</td>
<td>36.64</td>
<td>47.65</td>
<td>52.99</td>
<td>56.48</td>
<td>56.67</td>
</tr>
<tr>
<td>DEC-OC</td>
<td>1.45 ↑</td>
<td>0.76 ↑</td>
<td>0.39 ↑</td>
<td>0.11 ↓</td>
<td>0.31 ↑</td>
</tr>
<tr>
<td>DEC-OD</td>
<td>1.56 ↑</td>
<td>1.20 ↑</td>
<td>0.40 ↑</td>
<td>0.11 ↑</td>
<td>0.34 ↓</td>
</tr>
<tr>
<td>ENC-OD</td>
<td>1.30 ↑</td>
<td>1.20 ↑</td>
<td>0.63 ↑</td>
<td>0.13 ↑</td>
<td>0.79 ↓</td>
</tr>
</tbody>
</table>

Table 4.6: Multimodal gains for wait-k models (en-fr)

4.2.3 Key Observations

- The BLEU scores for French tend to be higher than that for German. This could be because of the fact that multimodal systems are able to translate correctly from English into gender-marked languages like French.

- The maximum BLEU score improved by almost 2 between MMT and NMT models. This result is slightly different from the one observed in the paper.

- DEC-OD system exhibits the best overall performance which is the same as the result observed in the paper.

4.2.4 Reproducibility

The values of hyperparameters that can reproduce these results are mentioned in this section.

- l2_reg: 1e-05
- lr_decay_factor: 0.5
- optimizer: adam
4.3 Phase II - Research

As mentioned above, the project had been divided into two parts. This section will focus on the second phase of the project, which was conducting research and identifying ways to improve the current state-of-the-art model for Simultaneous Machine Translation [1].

This phase of the project began in February, after the completion of the interim report and presentation. In our first meeting in February, Dr. Kong suggested that there were two feasible options that could be followed to build upon the current work, which were:

1. Extracting features from the images using the CLIP (Connecting Text and Images) model [23] to incorporate the visual modality (instead of using the ResNet CNN)

2. Building a new dataset, specifically for the task of Simultaneous Machine Translation using TED Talks and their subtitles

Both of these options had great potential and their own advantages and disadvantages which have been explained in detail in sections 4.4 and 4.5. After conducting initial research on both options, we decided that option two was the more appropriate path going forward. Therefore, for the purpose of this project, we have created a new dataset tailored to the task of SiMT.

4.4 CLIP Model - Contrastive Language–Image Pre-training

The current implementation of the SiMT model uses ResNet-50 CNN and ResNet-101 object detector to extract Object Classification (OC) and Object Detection (OD) features from the images present in the Multi30K dataset.

CLIP model was released in early 2021 by OpenAI, and it can efficiently learn visual concepts from natural language supervision. It is designed in such a way that it can perform a great
variety of classification benchmarks and can be instructed using natural language. CLIP model follows a “zero-shot” learning approach where the images used for testing the model were not used during training [24]. As reported in the paper, CLIP model performs better than ResNet on images not seen by the model during training (see Figure 4.3).  

The pre-training method used by CLIP is also similar to what we are trying to achieve in Simultaneous Machine Translation. The model takes in text and an image and pre-trains an image encoder and a text encoder before the classification task (see Figure 4.4). Thus, we planned to use this pretrained encoding as an input into the current model which may have provided better results than using a ResNet CNN.

Thus, it seemed that using the CLIP model was a promising option that may have provided better results than the original representation. However, we came to the conclusion that the major limitation of the current model was not the pre-training architecture used, but the dataset itself. Multi30K dataset used in the original paper was designed for the task of Multimodal Machine Translation which is somewhat different from Simultaneous Machine Translation.
4.5 Analysis of Available Datasets

This section will analyze the Multi30K dataset which is being used currently, explain the drawbacks of the current dataset, look into other open source datasets available online and will ultimately justify, why creating a new dataset was necessary for this problem.

4.5.1 Multi30K Dataset

To understand the need for creating a new dataset, first we need to analyze the dataset we are using currently and understand its drawbacks. Multi30K dataset was an extension of Flickr30K dataset [15] and was originally designed for machine translation, multimodal machine translation and caption generation tasks. A couple of sample images from the Multi30k dataset can be seen in Figure 4.5.
1. Brick layers constructing a wall.

2. Maurer bauen eine Wand.

1. Trendy girl talking on her cellphone while gliding slowly down the street

2. Ein schickes Mädchen spricht mit dem Handy während sie langsam die Straße entlangschweibt.

Translations

As we can observe, these are still images with captions and translations describing the tasks being performed in the image. The purpose of Simultaneous Machine Translation is to reproduce human interpretation where the translation has to start with an incomplete text source which is read progressively. Some possible applications are generating captions for live news, TED Talks, Youtube videos etc. in different languages automatically using machine learning. This dataset is not suitable for any of the applications which we want to target through such technology.

Therefore, we would like the supplementary data (images) to change as the context of the sentence changes which can be achieved through a video based dataset. Because of this, we concluded that there was a need for a video based dataset specifically designed for the task of SiMT where the context changes with time and the model can work on a stream of input frames like in a video as opposed to still images.

4.5.2 Available Datasets

The next part of the project was to figure out if there are already any existing datasets which can help us solve this problem. Since I was focusing on TED Talks, I analyzed several
datasets related to TED Talks available online. None of the datasets available online provided the video and the subtitle files related to a TED talk. The most comprehensive dataset related to TED talks I found was the Multitarget TED Talks Task (MTTT) which provided the transcripts of TED talks in different languages [25]. These translations were not time separated like in a subtitle file in “.srt” format which would not solve the main problem of changing the supplementary context with time.

A model developed using the Multi30K dataset or any of the other available datasets present online will not be suitable for practical implementations for the applications like generating captions for live news, TED Talks, Youtube videos etc. Wu et al. [26] has also reported that visual context helps Multimodal Machine Translation, in a similar way as regularization. The paper concluded that MMT systems needed to address the current shortcoming and emphasized the need for the creation of a new dataset, moving forward. This is why there is a need for a new dataset in the field of MMT and SiMT.

4.6 TED Talk Video Subtitle Dataset

We have established that there is a need for a new dataset specifically designed for the task of Multimodal and Simultaneous machine translation. There were several ways to create a new video based dataset but TED Talks seemed like the most feasible solution. Another reason why TED Talks were used is that the subtitles created for videos on TED are made by a team of expert translators and then, their work is reviewed by a reviewer. Thus, the quality of translations in different languages for TED talks would superior.

This section of the report focuses on the steps used to extract the video and subtitle files from the TED website. Samples from the dataset and some statistics are also mentioned later in this section. The basic steps used to extract the dataset from the website were:

- Sort the TED talks by popularity and extract the URLs from the webpage
- For each of the URLs, click the share button to extract the video download link
- Use ffmpeg to extract the subtitle file from the video and then compress it
These steps have been explained in more detail in section 4.6.1.

4.6.1 Procedure for Data Extraction

The first step in the process of data extraction was to sort the TED talks by popular using a link in this format: ["https://www.ted.com/talks?sort=popularpage="] The page number was increased in a while loop until there were no more videos left. Using Python libraries like requests and beautifulsoup, each page was converted into HTML and then links were extracted from it. This provided an initial URLs file consisting of the links to 5497 TED Talks.

The next step in the process was to figure out a way to download the video file with embedded subtitles in all available languages. While the transcript of each video could be extracted from the HTML directly, to download the video with subtitles, each TED Talk had to be opened on a browser window and then required the clicking of “Share” button. A pop-up window appeared after clicking the “Share” button which contained the video and audio download URLs (see Figure 4.6). To automate this process, I used the Python library Selenium with Chromium to open a tab for each TED talk. Then the script clicked on the button, extracted the link and saved them in a CSV file. Out of the 5497 videos, 4163 had download links.

Finally the last step was to download each video and extract the embedded subtitles. The
videos also needed to be compressed as the original files were in 1080p quality which required a huge amount of storage space. To achieve this part, I wrote a shell script along with a couple of helper functions. The script essentially downloaded the video, extracted the subtitle files in “.srt” format and then compressed it to 288p quality which reduced the file size by approximately 90%. After all the steps were completed, the dataset created was of approximately 118GB in size and contained 4163 videos along with the subtitles. Some statistics related to the dataset are mentioned in the next section.

### 4.6.2 Dataset Samples

This section gives a brief introduction of the dataset as well as shows some statistics related to it. On further analyzing the folders, it was discovered that out of the 4163 videos, 139 videos had no embedded subtitle files. Thus, the final dataset contains 4024 videos. Each language contains a different number of videos associated with it. A few sample images from a random video are shown in Figure 4.7.

Figure 4.7 shows three sample frames captured from the TED Talk “The next outbreak? We’re not ready” [27]. Subtitles in English, German and French are also added to the frame using the “.srt” files for visualization. In reality, the model will have two inputs, the video frame and the text extracted from the subtitle file separately.

The key point to observe from Figure 4.7 would be that different frames can provide different amounts of information. For example, frame 1 only contains the image of a person which may not be very helpful in capturing the missing context. In this case, the information extracted from the image may only act as a regularization method (e.g., weight decay) [27].

The third frame contains the picture of a slide on which is it written “We are not ready for the next epidemic.” This is exactly the text that the speaker is going to speak. Therefore in this case, even though the words spoken by the speaker will be fed into the model progressively, the features extracted from the visual context entered into the model can help in providing the missing textual context. This should improve the quality of translation in a SiMT system. Frame 2 lies somewhat in-between frames 1 and 3. It contains the image of the person as...
Figure 4.7: Samples of 3 frames from the TED Talk “The next outbreak? We’re not ready” with subtitles in English, German and French [27]. Different frames provide different amounts of context. For example, frame 1 does not provide any useful context for translation while frame 3 contains the entire text in English on the slide. Frame 2 lies in between frames 1 and 3.

well as the slide which contains a graphical representation of doctors. The speaker is talking about “Médecins Sans Frontières” and “volunteers” during this frame. Thus, in this case, the graphics present on the slide should be able to provide some missing context using the visual context.

This dataset contains the videos and subtitles files from 4024 videos. Many speakers use slides, animations and other visual aids in TED talks for explaining clearly. This is also a major advantage of the dataset as we hypothesize that images that contain some useful context related to the text will ultimately help in improving the translation quality, while the images where the speaker or the audience is shown will help in regularization. We plan to use a segmentation algorithm to identify the frames in the video and segment them by the amount of information present in that frame. This algorithm will be implemented using
OpenCV in Python which can detect whether there is a face present in an image or not. Thus, we can split the frames using OpenCV and put more weights on the image where some useful context is present. More details about this part can be found in section 4.7.

4.6.3 Dataset Statistics

This section of the report mentions some statistics related to this newly scraped dataset.

- **Size:** 118GB
- **Number of videos:** 4024
- **Number of subtitle files:** 93486
- **Total number of languages:** 108

The number of videos available in different languages for the top 22 languages sorted in descending order are mentioned in Table 4.7

<table>
<thead>
<tr>
<th>Language</th>
<th>Code</th>
<th>Number of Videos</th>
</tr>
</thead>
<tbody>
<tr>
<td>English</td>
<td>ENG</td>
<td>4013</td>
</tr>
<tr>
<td>Spanish</td>
<td>SPA</td>
<td>3990</td>
</tr>
<tr>
<td>Arabic</td>
<td>ARA</td>
<td>3971</td>
</tr>
<tr>
<td>Chinese</td>
<td>ZHO</td>
<td>3912</td>
</tr>
<tr>
<td>Portuguese</td>
<td>POR</td>
<td>3896</td>
</tr>
<tr>
<td>French</td>
<td>FRA</td>
<td>3833</td>
</tr>
<tr>
<td>Korean</td>
<td>KOR</td>
<td>3692</td>
</tr>
<tr>
<td>Turkish</td>
<td>TUR</td>
<td>3606</td>
</tr>
<tr>
<td>Russian</td>
<td>RUS</td>
<td>3546</td>
</tr>
<tr>
<td>Italian</td>
<td>ITA</td>
<td>3519</td>
</tr>
<tr>
<td>Japanese</td>
<td>JPN</td>
<td>3397</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Language</th>
<th>Code</th>
<th>Number of Videos</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hébreu</td>
<td>HEB</td>
<td>3074</td>
</tr>
<tr>
<td>Persian</td>
<td>FAS</td>
<td>2990</td>
</tr>
<tr>
<td>Vietnamese</td>
<td>VIE</td>
<td>2959</td>
</tr>
<tr>
<td>Romanian</td>
<td>RON</td>
<td>2915</td>
</tr>
<tr>
<td>Hungarian</td>
<td>HUN</td>
<td>2875</td>
</tr>
<tr>
<td>Polish</td>
<td>POL</td>
<td>2758</td>
</tr>
<tr>
<td>Dutch</td>
<td>NLD</td>
<td>2638</td>
</tr>
<tr>
<td>German</td>
<td>DEU</td>
<td>2525</td>
</tr>
<tr>
<td>Modern Greek</td>
<td>ELL</td>
<td>2479</td>
</tr>
<tr>
<td>Serbian</td>
<td>SRP</td>
<td>2335</td>
</tr>
<tr>
<td>Indonesian</td>
<td>IND</td>
<td>2024</td>
</tr>
</tbody>
</table>

Table 4.7: Number of videos available for different languages in descending order

The vocabularies were extracted using the raw data without pre-processing for upper-case and lower-case characters. The punctuations were removed from each word. Duplicated words have also been removed and the number of unique tokens in each language for the languages on which we are focusing on (English, German and French) and a few others for reference are mentioned in Table 4.8.
### Table 4.8: Vocabulary size of different languages

<table>
<thead>
<tr>
<th>Language</th>
<th>Code</th>
<th>Number of Tokens</th>
</tr>
</thead>
<tbody>
<tr>
<td>English</td>
<td>ENG</td>
<td>94603</td>
</tr>
<tr>
<td>French</td>
<td>FRA</td>
<td>126818</td>
</tr>
<tr>
<td>German</td>
<td>DEU</td>
<td>161763</td>
</tr>
<tr>
<td>Czech</td>
<td>CES</td>
<td>145448</td>
</tr>
<tr>
<td>Italian</td>
<td>ITA</td>
<td>136772</td>
</tr>
<tr>
<td>Spanish</td>
<td>SPA</td>
<td>129609</td>
</tr>
</tbody>
</table>

A few samples from the vocabulary from each of the languages used in table 4.8 are mentioned in table 4.9.

### Table 4.9: Subset of the sample vocabulary in different languages

<table>
<thead>
<tr>
<th>Language</th>
<th>Code</th>
<th>Sample Vocabulary</th>
</tr>
</thead>
<tbody>
<tr>
<td>English</td>
<td>ENG</td>
<td>'hopefully', 'enclosures', 'nervous', 'scrambling', 'sociologists', 'habits', 'optus', 'territories', 'convulsions', 'goes', 'wooded' ...</td>
</tr>
<tr>
<td>French</td>
<td>FRA</td>
<td>'sectarisme', 'lénigmatique', 'gravons', 'difficultés', 'rêverie', 'surveillées', 'partagerions', 'infantile', 'concorde', 'emparera', 'myéline' ...</td>
</tr>
<tr>
<td>German</td>
<td>DEU</td>
<td>'Speerfahren', 'Staatskunst', 'anzweifelt', 'Veranstaltungsplanerin', 'Rechtsdienst', 'Tragweite', 'elWafi', 'Neiman', 'Dienstleistungsmodelle'...</td>
</tr>
<tr>
<td>Czech</td>
<td>CES</td>
<td>'kulisovým', 'Osy', 'cítíš', 'Enclosures', 'pokřivený', 'obdivují', 'četná', 'celė', 'trčící', 'diktovat', 'cenový', 'Jackie', 'dýmování', 'přijaty'...</td>
</tr>
<tr>
<td>Italian</td>
<td>ITA</td>
<td>'metalinguaggio', 'prevedessero', 'abbinati', 'Enclosures', 'bufere', 'fobiche', 'sciarpia', 'sialici', 'aggregano', 'conclusivi', 'rimpiazzarle' ...</td>
</tr>
<tr>
<td>Spanish</td>
<td>SPA</td>
<td>'reptando', 'permitirnos', 'convivan', 'presupuesto', 'respondiésemos', 'contengan', 'estratega', 'temernos', 'extirpó', 'invertirlo' ...</td>
</tr>
</tbody>
</table>

### 4.7 Next Step

Since the dataset has now been created and analyzed, the next step for the project is to generate the English-German and English-French sentence pairs using the subtitle files. After creating the sentence pairs, we plan to use the time mentioned in the “.srt” file to capture the frames from the video using the segmentation algorithm and OpenCV. Once a script that can extract the sentence pairs and frames from the video has been implemented, we plan to use a ResNet-50 CNN to extract the visual context from the frames captured from the video and then run the original Simultaneous Machine Translation model on our new dataset. The results obtained on this new dataset will be published in the paper. We plan to finish this
task by mid June and submit the final paper to the EMNLP (Empirical Methods in Natural Language Processing) Conference.

4.8 Schedule

The schedule which was followed throughout the development of this project has been mentioned in Table 4.10.

<table>
<thead>
<tr>
<th>Time Period</th>
<th>Tasks</th>
</tr>
</thead>
<tbody>
<tr>
<td>September</td>
<td>• Meet with the supervisor</td>
</tr>
<tr>
<td></td>
<td>• Write a detailed project plan</td>
</tr>
<tr>
<td></td>
<td>• Develop the project website</td>
</tr>
<tr>
<td></td>
<td>• Start literature review on SiMT</td>
</tr>
<tr>
<td>October</td>
<td>• Finish the literature review on SiMT</td>
</tr>
<tr>
<td></td>
<td>• Read the documentation of pysiwt</td>
</tr>
<tr>
<td></td>
<td>• Download and explore the dataset</td>
</tr>
<tr>
<td>November - December</td>
<td>• Reproduce the results from the original paper [1]</td>
</tr>
<tr>
<td></td>
<td>• More literature reviews on existing works</td>
</tr>
<tr>
<td>January</td>
<td>• Finish training on the en-fr dataset</td>
</tr>
<tr>
<td></td>
<td>• Finish interim report</td>
</tr>
<tr>
<td></td>
<td>• Finalize research topic</td>
</tr>
<tr>
<td>February</td>
<td>• Start researching on scraping TED Talks</td>
</tr>
<tr>
<td></td>
<td>• Get familiar with selenium, requests and beautiful soup libraries</td>
</tr>
<tr>
<td>March - April</td>
<td>• Complete scraping the dataset</td>
</tr>
<tr>
<td></td>
<td>• Finish final report</td>
</tr>
<tr>
<td></td>
<td>• Prepare for final presentation</td>
</tr>
<tr>
<td></td>
<td>• Design the poster and prepare for the exhibition</td>
</tr>
</tbody>
</table>

Table 4.10: Project Schedule
4.9 Challenges

This section summarizes some of the challenges I encountered over the course of this project.

- Lack of documentation for the pysimt library made this project quite difficult as it is the most important framework required for this project. Besides, there are several bugs on the original Github repository [13] which required manual debugging by understanding 1000s of lines of code.

- The size of the Multi30K dataset was also a problem as it took days to train the models whose results have been reported in section 4.2. Each model took around 2 hours for training.

- While scraping the new TED Talks dataset, one of the biggest problems was the scale of the dataset. When scraping the links from the TED website for 5497 URLs, the Python script took over 4 days to complete while running parallelly on 6 terminals. Manual supervision was also required as after sending too many HTTP requests in a short time span, the HKU Wifi network disconnects that device. Due to this, I had to manually restart the laptop and run the process multiple times.

- Another major challenge I faced during the second phase was that, after I had written the scripts to download the videos, there was some error on the TED server due to which downloads could not succeed. I reported the error to them via email and it got fixed after a few days which wasted precious time.

- Downloading 4163 TED Talks from the website took over a week. Due to the scale of the dataset, the computational resources available on the HKU servers were not enough since each script takes hours to run. Thus, performing experiments on the new dataset was very time consuming and was also computational intensive.

Overall, I believe that I have learned a lot over the course of this project, especially about the latest research in the field of Natural Language Processing and conducting academic research. Though I did encounter some challenges, I actively worked to come up with solutions to mitigate them.
4.10 Limitations

Even though the newly created dataset using TED Talks is quite robust, it still has its own limitations. A major limitation would be the scale of the dataset. Since the size of the dataset is quite large, reproducing the results obtained using the SiMT model on this dataset would be an intensive task even for modern GPUs and could take days. This limits the number of people who will be able to benefit from such a dataset as the resources required to work on such large datasets are mostly available in the industry.

The dataset also contains a lot of noise as it contains videos from 4024 TED talks. A major hypothesis while working on this project was that the videos that contain slides, animations or other resources used for explaining will be better as they would be able to provide better quality visual context for the translation. However, the majority of the TED Talks only show the face of the speaker and the audience in the video which will generate an inefficacious context that will not be useful for multimodal or simultaneous machine translation. A possible method to overcome this limitation would be to use audio as the source of auxiliary information instead of images.

4.11 Future Work

There can be numerous other applications of the TED Talk dataset that has been created for this project. For example, there are a total number of 93486 subtitle files present in the dataset. Using these files, one can train models for machine translation for not just English to an arbitrary language but train models for any set of languages available in the dataset. Another major application could be, that even though we are focusing on the task of Simultaneous machine translation in this project, the video file contains the audio associated with the speech. Therefore the audio files can be used to analyse the effect of audio as supplementary input for the machine translation task. Thus, this would be a significantly important dataset in the field of Multimodal Machine Translation and should help out several researchers in this field.
5. Conclusion

This report discusses the application of NLP in the field of simultaneous machine translation. SiMT is the process of translating one language into another instantaneously by a computer program without human intervention. Therefore this project aims to develop and improve a SiMT system that will take an input sentence and an image and output the sentence in the target language. Speed and accuracy are the two top priorities for this project. Empirical research has shown that having data from a supplementary source can improve the performance of a language model [1].

The project was divided into two parts for convenience: reproducing the results of the original paper [1] and research. The first phase of the project was completed in January. Some key points to note are that under low-latency wait-k policies, the visual cues were highly impactful and improved the translation by 1.82 BLEU points. Besides, the qualitative analysis also shows that SiMT models perform better on gender marked languages which is what was observed as the model performed better for French which is a gender marked language as compared to German.

During the second phase, a few possible options were analyzed and after much discussion, I decided to go ahead with creating a new dataset. Since the current dataset was not designed for this task and recent research suggested that there was a need for a new dataset in the field of simultaneous and multimodal machine translation [26], this path was finalized. The new dataset created during this project consists of 4024 video files along with subtitles in several languages. Statistics related to the dataset are mentioned in section 4.6.3.

Several problems came up during the development of this project but I actively came up with solutions to tackle them. Overall, I believe that this project was a great learning experience as I learned a lot about the latest advancements in the field of Natural Language Processing.
References


