Generative AI Assignment 1

Rohan Mathew Sabu, Advik Sinha

1 Introduction

Neural Machine Translation (NMT) leverages sequence-to-sequence neural architectures to translate text from one language to another. In these systems, encoder-decoder models are trained to capture intricate linguistic patterns and produce fluent, contextually appropriate translations. This report presents a comparative analysis between fine-tuned transformer models and fully trained transformer models using a large-scale parallel corpus.

2 Training Data

2.1 Dataset

The data set used in this study is the IITB English-Hindi Parallel Corpus, which comprises more than 1.66 million parallel sentence pairs. It is a high-quality parallel corpus developed by the Centre for Indian Language Technology (CFILT) at IIT Bombay. It is widely recognized for training and evaluating neural machine translation (NMT) models.

2.2 Preprocessing and Tokenisation

Prior to training, the data set is subjected to a preprocessing pipeline to ensure consistency and optimal performance. Tokenisation is performed using the `MarianTokenizer` from HuggingFace associated with the pretrained model *"Helsinki-NLP/opus-mt-en-hi"*. This tokenizer uses the Byte-Pair Encoding (BPE) [1] technique and efficiently handles rare and out-of-vocabulary words by breaking them into manageable subword units. Furthermore, the tokenised sequences are padded to a fixed maximum length, and any tokens exceeding this limit are truncated. The resulting token sequences are then converted into tensors, which serve as input to the translation models. For the pretrained mBART model, the tokenization is done using the `MBart50TokenizerFast` from HuggingFace. It uses a similar Byte-Pair Encoding (BPE) technique to the `MarianTokenizer`.

3 Method

3.1 Architecture

3.1.1 Fine-Tuned Model

The fine-tuned model is based on the MarianMT [2] transformer architecture using the pretrained model *"Helsinki-NLP/opus-mt-en-hi"*. This architecture utilizes an encoder-decoder framework where the encoder processes the input from the source language, and the decoder generates the translated output. In our approach, all layers of the encoder are kept frozen, with only decoder layers remaining trainable. This selective fine-tuning allows the model to adapt to the specifics of the English-Hindi translation task while retaining the robust general language understanding learned during pretraining. The fine-tuning was performed on the dataset mentioned earlier.

3.1.2 Fully Trained Model

The fully trained model is based on mBART-50 [3], a many-to-many multilingual machine translation model available on Hugging Face [4]. mBART-50 is pretrained on 50 languages using a denoising auto-encoding objective and has been designed to facilitate robust multilingual translation tasks. Its architecture follows a standard encoder-decoder Transformer model, where both the encoder and decoder comprise multiple self-attention and cross-attention layers. The model's large capacity and extensive pretraining on diverse language pairs enable it to learn rich cross-lingual representations. This makes mBART-50 particularly well-suited for low-resource translation scenarios, as it can leverage its multilingual knowledge to perform effective translation even when parallel corpora are limited. The decision to employ mBART-50 as the fully trained model is motivated by its proven performance in many-to-many translation settings and its strong generalization capabilities across a wide range of languages.

3.2 Training

To train the model, we used a subset of 1,000,000 sentences from the dataset. We limited the sequence length to 75 to make the training time more manageable, truncating longer sentences and padding shorter ones. We trained the model for 6 epochs, using a learning rate of 3e-5. The training process utilizes token-wise cross-entropy loss computed over batched inputs.

Each epoch of the training process took around 3 hours on the Kaggle T4 GPUs. As we had limited access to GPUs, we had to keep checkpointing the training process and resuming it on different platforms. As a result, we only have the training loss plot for the first 5 epochs of training, which we have provided in Figure 1.

3.3 Testing

For evaluation, the performance of the translation models was assessed using common machine translation metrics: BLEU, ROUGE, and METEOR. These metrics provide quantitative measures of translation quality by comparing the generated translations

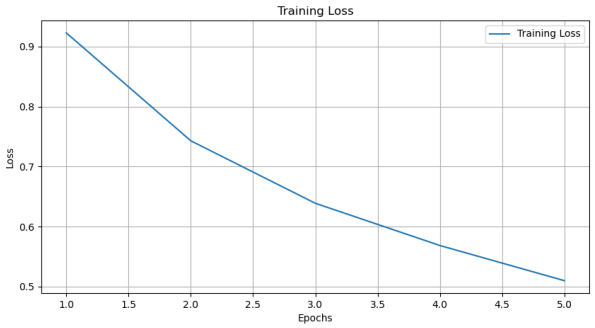

Figure 1: Training Loss Plot for the First 5 epochs

to reference sentences, capturing aspects such as fluency, adequacy, and semantic relevance.

To compute the BLEU, ROUGE, and METEOR scores, we used the HuggingFace implementations of these metrics. The two models were tested on all the examples present in the testing data of the IITB dataset.

4 Results

The average BLEU, ROUGE, and METEOR scores over the entire test dataset are provided in Table 1. The average BLEU, ROUGE, and METEOR for different sequence lengths is displayed in Figure 2 for the fine-tuned Marian-MT model and Figure 3 for the pretrained mBART model.

Some sample translations from the fine-tuned Marian-MT model are given below.

- *Input*: A black box in your car?
  Prediction: अपनी कार में एक काला बॉक्स?
  Ground Truth: आपकी कार में ब्लैक बॉक्स?

- *Input*: As America's road planners struggle to find the cash to mend a crumbling highway system, many are beginning to see a solution in a little black box that fits neatly by the dashboard of your car.
  Prediction: जैसे अमेरिका के मार्गदर्शक योजना के लिए अमेरिका की प्रस्ताव में करने के लिए करीब
  Ground Truth: जबकि अमेरिका के सड़क योजनाकार, ध्वस्त होते हुए हाईवे सिस्टम को सुधारने के लिए धन की कमी से जूझ रहे हैं, वहीं बहुत–से लोग इसका समाधान छोटे से ब्लैक बॉक्स में देख रहे हैं, जो आपकी कार के डैशबोर्ड पर सफ़ाई से फिट हो जाता है।

- *Input*: The devices, which track every mile a motorist drives and transmit that information to bureaucrats, are at the center of a controversial attempt in Washington and state planning offices to overhaul the outdated system for funding America's major roads.
  Prediction: जो उन उपकरणों को लगाते हैं, जो हर मिल जाते हैं और जो प्रकार की जानकारी के लिए
  Ground Truth: यह डिवाइस, जो मोटर–चालक द्वारा वाहन चलाए गए प्रत्येक मील को ट्रैक करती है तथा उस सूचना को अधिकारियों को संचारित करती है, आजकल अमेरिका की प्रमुख सड़कों का वित्त–पोषण करने के लिए पुराने हो चुके सिस्टम का जीर्णोद्धार करने के लिए वाशिंगटन और राज्य नियोजन कार्यालय के लिए एक विवादास्पद प्रयास का मुद्दा बन चुका है।

Some sample translations from the pretrained mBART model are given below.

- *Input*: A black box in your car?
  Prediction: कार में एक काला बक्सा?
  Ground Truth: आपकी कार में ब्लैक बॉक्स?

- *Input*: As America's road planners struggle to find the cash to mend a crumbling highway system, many are beginning to see a solution in a little black box that fits neatly by the dashboard of your car.
  Prediction: अमेरिका के सड़क योजनाकारों को एक गिरते हुए राजमार्ग प्रणाली को सुधारने के लिए नकद ढूंढने में कठिनाई पड़ रही है, इसलिए बहुत से लोग एक छोटी–सी काली बक्से में एक समाधान देखना शुरू कर रहे हैं जो अपने कार के डैशबोर्ड के पास ठीक से फिट हो।
  Ground Truth: जबकि अमेरिका के सड़क योजनाकार, ध्वस्त होते हुए हाईवे सिस्टम को सुधारने के लिए धन की कमी से जूझ रहे हैं, वहीं बहुत–से लोग इसका समाधान छोटे से ब्लैक बॉक्स में देख रहे हैं, जो आपकी कार के डैशबोर्ड पर सफ़ाई से फिट हो जाता है।

- *Input*: The devices, which track every mile a motorist drives and transmit that information to bureaucrats, are at the center of a controversial attempt in Washington and state planning offices to overhaul the outdated system for funding America's major roads.
  Prediction: ये उपकरण, जो प्रत्येक मील को ट्रैक करते हैं जो एक मोटरिस्ट ड्राइव करता है और उस जानकारी को नौकरशाहों को प्रेषित करते हैं, अमेरिका के प्रमुख सड़कों के लिए निधिकरण के लिए पुरानी प्रणाली का पुनर्गठन करने के लिए वाशिंगटन और राज्य योजना कार्यालयों में एक विवादास्पद प्रयास के केंद्र में हैं
  Ground Truth: यह डिवाइस, जो मोटर–चालक द्वारा वाहन चलाए गए प्रत्येक मील को ट्रैक करती है तथा उस सूचना को अधिकारियों को संचारित करती है, आजकल अमेरिका की प्रमुख सड़कों का वित्त–पोषण करने के लिए पुराने हो चुके सिस्टम का जीर्णोद्धार करने के लिए वाशिंगटन और राज्य नियोजन कार्यालय के लिए एक विवादास्पद प्रयास का मुद्दा बन चुका है।

These results suggest that fine-tuning the Marian-MT on a limited dataset likely lead to poor generalization of the model. On the other hand, the pretrained mBART model, even without additional fine-tuning, might already have strong cross-lingual representations that allow it to perform better.

Model	BLEU Score	ROUGE Score	METEOR Score
Fine-Tuned MarianMT	0.03	0.10	0.19
PreTrained mBART	0.18	0.17	0.42

Table 1: Translation performance of fine-tuned and pretrained models.

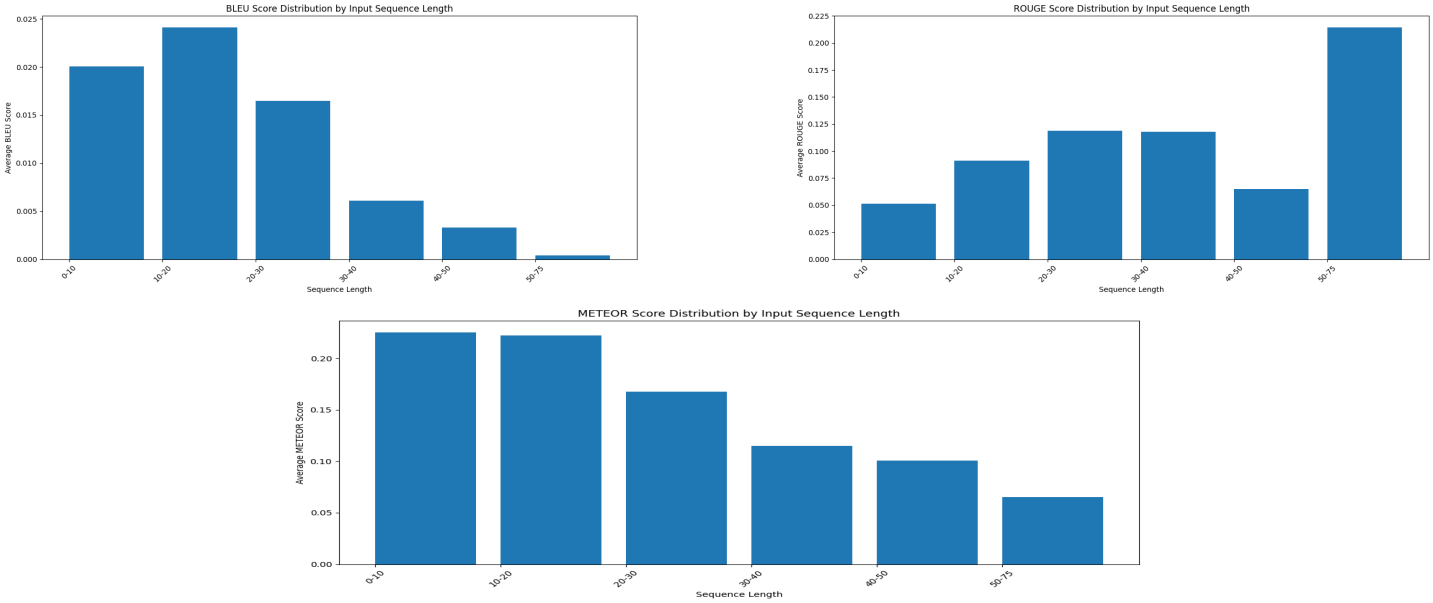

Figure 2: Average BLEU, ROUGE, and METEOR Scores for Fine-Tuned Marian-MT

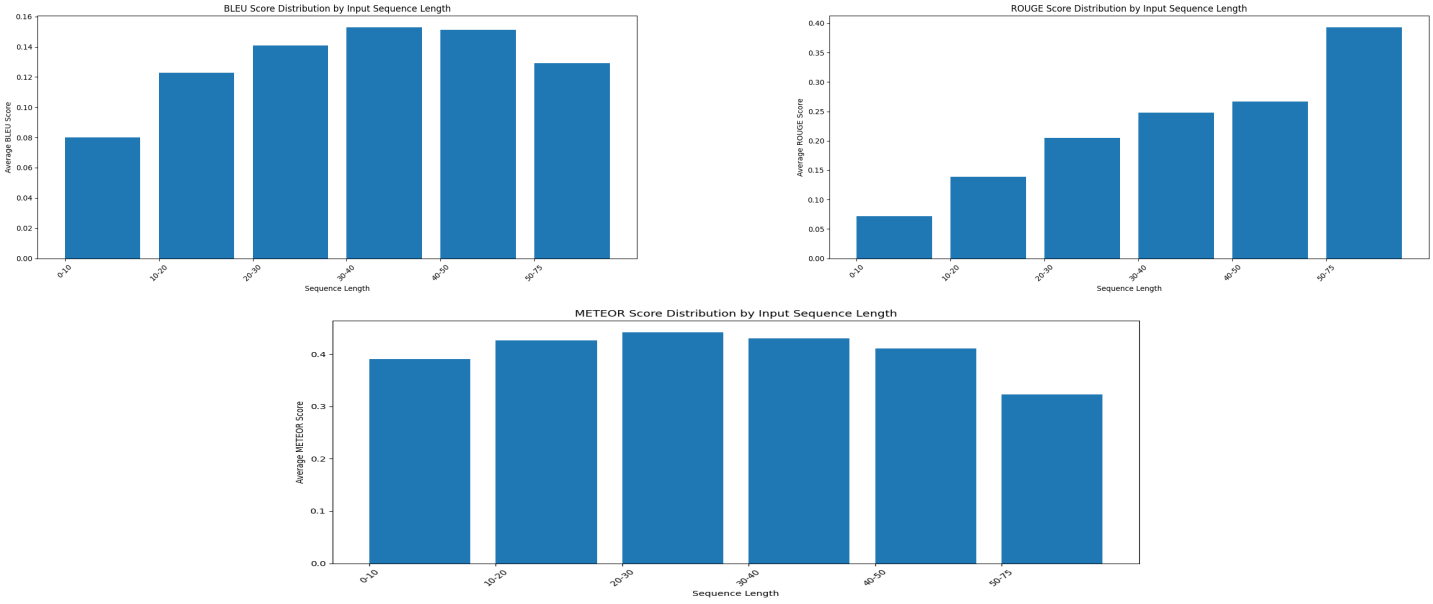

Figure 3: Average BLEU, ROUGE, and METEOR Scores for PreTrained mBART

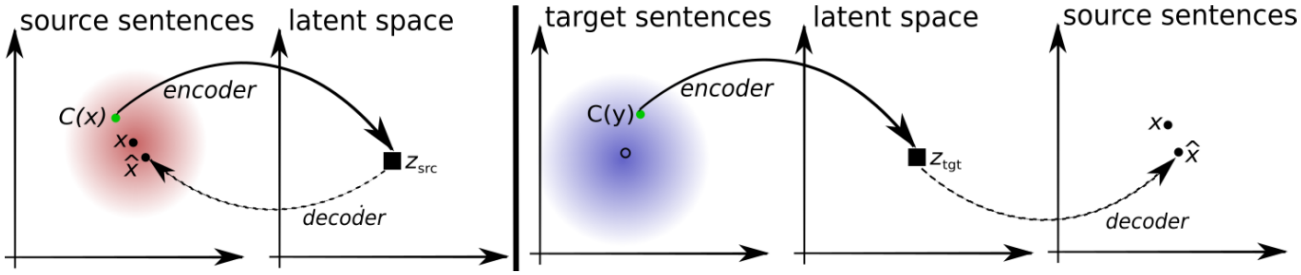

Figure 4: Principles Guiding the Architecture [5]

The diagram to the left (auto-encoding) depicts where the model reconstructs a sentence from its noisy input. The diagram to the right (back-translation) depicts how the input language is first mapped to a poor translation (as the model is still training) and is then translated back to the input language. This helps the model learn the relationships between languages without explicitly needing a parallel dataset.

5 Low Resource Translation

Monolingual datasets play a crucial role in bridging the gap for low-resource language translation by enabling unsupervised methods to learn shared linguistic representations without relying on large parallel corpora. As illustrated in Figure 4, the model can first reconstruct a sentence from its own noisy input (de-noising auto-encoding), thereby learning the syntactic and semantic structures of each language. The de-noising autoencoding step involves the model reconstructing sentences from noisy inputs generated by random token dropping and shuffling.

Subsequently, a back-translation step forces the model to map input sentences into a shared latent space and translate them into another language before converting them back, thus reinforcing bidirectional translation capabilities even in the absence of extensive parallel data. An anchor language strategy further reduces computational complexity by limiting back-translation to pairs involving a single, fixed language. The use of pretrained tokenizers and embedding models such as mBART [3] ensures that words across multiple languages are aligned in a unified embedding space, allowing the encoder to generate consistent latent representations. Training methods such as teacher forcing can also be applied during training to ensure rapid convergence. This involves feeding the ground truth back to the RNN at each timestep during the initial stages of training to help the model learn the rough semantics better and quicker. Methods such as beam search with length normalization can also be implemented during inference, which helps in generating more fluent translations.

6 Conclusion

Our exploration into fine-tuning versus fully training transformer architectures for English-Hindi translation highlights the effectiveness of leveraging pretrained models. This approach not only accelerated convergence but also demonstrated improved robustness by capitalizing on the rich linguistic representations learned during pretraining.

Fine-tuning a pretrained model allows us to use prior knowledge, requiring less data and compute as compared to training a model from scratch. However, fine-tuning a model may introduce biases from the pretrained model, or may not perform well for some specific tasks if the training data is not sufficient or if the domain is very different.

However, it is important to note that while we fine-tuned MarianMT, our pretrained baseline was mBART—a model that is generally considered more powerful due to its broader pretraining. As a result, despite the typical benefits of fine-tuning, our fine-tuned MarianMT model still performs worse than the pretrained mBART when compared directly. This suggests that improvements from fine-tuning are most evident when comparing within the same model architecture; differences in the underlying models can have a significant impact on translation performance.

The insights drawn from our analysis imply that fine-tuning pretrained transformer models is a practical and efficient strategy for real-world machine translation, particularly in scenarios where bilingual corpora are limited. At the same time, the choice of base model is crucial—models with stronger pretrained representations, such as mBART, may offer superior performance even without fine-tuning, underscoring the need for careful model selection in low-resource environments.

Such strategies are highly applicable in industries ranging from localization services and content translation, enabling more accessible and accurate translation systems in multilingual settings. One such example would be the live translation services and translation of languages in comments of social media sites such as youtube and instagram.

References

[1] V. Zouhar, C. Meister, J. L. Gastaldi, L. Du, T. Vieira, M. Sachan, and R. Cotterell, "A formal perspective on byte-pair encoding," 2024.

[2] M. Junczys-Dowmunt, R. Grundkiewicz, T. Dwojak, H. Hoang, K. Heafield, T. Neckermann, F. Seide, U. Germann, A. F. Aji, N. Bogoychev, A. F. T. Martins, and A. Birch, "Marian: Fast neural machine translation in c++," 2018.

[3] Y. Tang, C. Tran, X. Li, P.-J. Chen, N. Goyal, V. Chaudhary, J. Gu, and A. Fan, "Multilingual translation with extensible multilingual pretraining and finetuning," 2020.

[4] Y. Tang, C. Tran, X. Li, P.-J. Chen, N. Goyal, V. Chaudhary, J. Gu, and A. Fan, "Multilingual translation with extensible multilingual pretraining and finetuning," 2020.

[5] G. Lample, L. Denoyer, and M. Ranzato, "Unsupervised machine translation using monolingual corpora only," *CoRR*, vol. abs/1711.00043, 2017.