Neural Machine Translation for English to German

Using Seq2Seq Model with Attention

Jayesh Nayak
Roll No: 121CS0195

Ashish Padhy
Roll No: 121CS0196

Arman Paikaray
Roll No: 121CS0197

November 13, 2024

Contents

Chapter 1

Introduction

Machine Translation (MT) has become an essential component of natural language processing (NLP), serving as a bridge between languages in an increasingly interconnected world. MT systems are designed to automatically translate text from one language to another, addressing both practical communication needs and enabling cross-lingual access to information. Over the decades, MT has evolved from rule-based approaches, which rely on pre-defined grammatical rules, to statistical methods that leverage probabilities derived from bilingual text corpora. However, with the advent of deep learning, Neural Machine Translation (NMT) has taken precedence as the most effective approach, offering substantial improvements in translation accuracy and fluency.

This project investigates the development of an NMT system specifically tailored for English to German translation. We focus on implementing two distinct deep learning architectures: a Sequence-to-Sequence (Seq2Seq) model with attention and a Transformer model. The Seq2Seq model, a foundational architecture in NMT, allows the system to effectively handle sequences of varying lengths, with the attention mechanism enhancing translation quality by helping the model focus on relevant parts of the input sentence. On the other hand, the Transformer model represents a more recent paradigm, utilizing self-attention mechanisms to capture long-range dependencies, often resulting in faster training times and improved performance.

The primary dataset used in this project is the WMT 14 English-German corpus, a standard benchmark for evaluating translation systems. We further experiment with transfer learning techniques on the Transformer model, aiming to demonstrate its ability to generalize across similar translation tasks. This report provides a comparative analysis of the Seq2Seq and Transformer architectures, evaluates the effectiveness of the attention mechanism, and examines the role of transfer learning in enhancing translation accuracy. Through these efforts, we contribute to the broader field of MT by exploring how modern NMT techniques can improve English to German translation quality.

Chapter 2

Work Done

2.1 Showcase

The project is hosted on GitHub at the following link: https://github.com/shurtugal/neural-machine-translation

Additionally, the project has been deployed and made available through Streamlit, and a Kaggle Notebook is provided for interactive exploration of the model and its performance.

- **Deployed App**: Explore the live version of the app on Streamlit at this link: Deployed Streamlit App
- **Kaggle Notebook**: Access the Kaggle Notebook with model insights and analysis here: Kaggle Notebook

2.2 Project Contribution

The project involved collaborative contributions in various aspects, including code development, training, reporting, and presentation. The specific contributions are as follows:

- **Code Development and Model Training**: Ashish Padhy
- **Graphical User Interface**: Ashish Padhy
- **Report Preparation**: Jayesh Nayak
- **Presentation (PPT) Design**: Arman Paikaray

These collaborative efforts ensured a well-rounded and thoroughly documented project, with live deployment, detailed reporting, and an accessible presentation format.

Chapter 3

Background Theory

3.1 Neural Machine Translation (NMT)

Neural Machine Translation (NMT) leverages deep neural networks to translate text sequences end-to-end without intermediate steps, unlike traditional methods that rely on phrase-based or rule-based systems. NMT uses an encoder-decoder architecture that effectively maps a sentence from a source language to a target language. Here, the encoder converts the input sentence into a fixed-length context vector, which represents the semantic meaning of the sentence, while the decoder translates this vector into the target language. NMT models are capable of capturing complex linguistic patterns and context, resulting in improved translation quality, especially for long and context-dependent sentences.

3.2 Seq2Seq with Attention

The Sequence-to-Sequence (Seq2Seq) model is one of the foundational architectures in NMT. It involves using two neural networks that function sequentially to encode an input sequence and then decode it into the target language.

- **Encoder:** The encoder processes the input sentence token-by-token and encodes it into a fixed-length context vector (latent space), capturing semantic information. This context vector is the final hidden state of the encoder, which contains compressed information about the input sequence.
- **Decoder:** The decoder generates the output sentence step-by-step, using the context vector as its primary reference. It predicts each token in the output sequence based on the context vector and previous token predictions, thus producing a sequentially coherent translation.

One limitation of basic Seq2Seq models is that the fixed-size context vector can struggle to retain information from long input sequences. This bottleneck can cause degraded performance in translations involving lengthy sentences. To mitigate this, the attention mechanism is introduced, allowing the model to focus dynamically on relevant portions of the input sequence during decoding.

3.2.1 Attention Mechanism

The attention mechanism enhances the Seq2Seq model by allowing the decoder to focus on different parts of the input sequence at each time step, improving the handling of long sequences and intricate grammatical structures. Specifically, at each decoding step, an attention score α_{ij} is calculated to weigh the relevance of each input token. This score between the i-th encoder hidden state h_i and the j-th decoder hidden state s_j is computed as:

$$\alpha_{ij} = \frac{\exp(e_{ij})}{\sum_{k=1}^{T_x} \exp(e_{ik})} \tag{3.1}$$

where $e_{ij} = s_j^T W h_i$ represents a compatibility function, which quantifies how well each encoder hidden state aligns with the current decoder state. This mechanism allows the model to dynamically allocate focus to different input tokens, resulting in translations that better capture context and nuances.

3.3 Transformer Architecture

The Transformer architecture introduces a shift from RNNs by relying entirely on self-attention mechanisms, which enable parallelization and make training significantly faster. The central component of the Transformer model is the self-attention mechanism, which computes the dependencies between all pairs of input tokens in a single pass, thereby capturing context efficiently. Self-attention operates as follows:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right) V \tag{3.2}$$

where Q, K, and V are the query, key, and value matrices derived from the input embeddings, and d_k is the dimension of the keys. The result is a weighted representation of the input sequence, where each token attends to relevant tokens in the sequence, capturing relationships irrespective of their positions. This mechanism underpins the Transformer's ability to model complex dependencies in text data efficiently, leading to state-of-the-art translation results.

Chapter 4

Dataset: WMT 14 EN-DE

The WMT 2014 English-German dataset is an essential benchmark in machine translation (MT) research, widely recognized for its role in evaluating and advancing MT models. Created for the annual WMT shared task, it enables consistent comparison across different MT approaches and allows researchers to track progress over time in translation quality and accuracy. With approximately 4.5 million parallel English-German sentence pairs, the dataset provides a robust and expansive foundation for training and testing, supporting a range of MT applications.

Key features of the WMT 2014 dataset contribute to its utility and reliability. Sourced from diverse domains—including Europarl proceedings, News Commentary, and TED Talks—the dataset offers a rich variety of sentence structures, vocabulary, and idiomatic expressions, reflecting realistic language use. This diversity enables MT models trained on WMT 2014 to generalize effectively, handling the nuances of different contexts and text types. Furthermore, the dataset is meticulously preprocessed, involving cleaning and normalization steps that ensure data consistency. This preprocessing streamlines model training, helping MT systems achieve higher efficiency and compatibility by reducing noise and inconsistencies in the data.

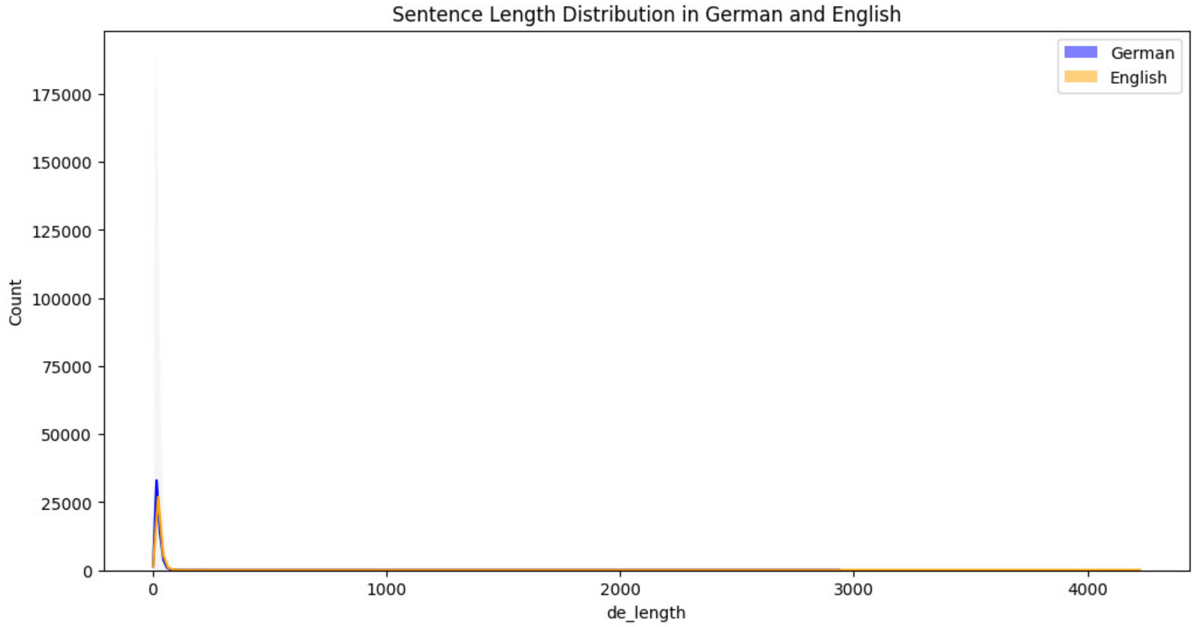


Figure 4.1: Dataset Details

Originally developed for the WMT 2014 News Translation Task, this dataset is not only valuable for that specific task but is applicable across a broad spectrum of MT research areas. Its size and quality make it particularly suitable for both statistical and neural machine translation approaches, including neural architectures like Seq2Seq with attention and Transformers. For our work, the WMT 2014 English-German dataset serves as a comprehensive resource that supports rigorous evaluation, enabling us to measure translation performance through standard metrics like BLEU scores and refine our model's handling of language complexities.

Chapter 5

Model Architectures

This section describes the architectures used in our NMT models, highlighting the core components of each design.

5.1 Seq2Seq with LSTM

Our first architecture is a Sequence-to-Sequence (Seq2Seq) model using Long Short-Term Memory (LSTM) layers. The Seq2Seq model with LSTM is designed to capture sequential dependencies in the input language and translate it to the target language effectively. The architecture consists of two main parts:

- **Encoder:** The encoder comprises an LSTM layer with 512 units, designed to process the input sequence one token at a time. As each token passes through the LSTM, the encoder updates its hidden state to capture information about both recent and past tokens, preserving long-range dependencies crucial for understanding the sentence structure. The final hidden state of the encoder serves as a fixed-length context vector that summarizes the meaning of the input sentence.
- **Decoder:** The decoder is another LSTM layer that uses the context vector from the encoder to generate the output sequence. It predicts each token in the target language based on the context vector and previously generated tokens. By feeding the predicted token back into the model at each step, the decoder learns to create coherent translations.

The probability of each token y_t in the target language, conditioned on previous tokens and the input sentence, is calculated as:

$$P(y_t|y_{<t}, x) = \text{softmax}(W \cdot h_t) \tag{5.1}$$

where h_t is the hidden state of the decoder at time t, and W is a weight matrix. This design forms the foundation of sequence-based NMT but can struggle with long input sequences due to the fixed-size context vector.

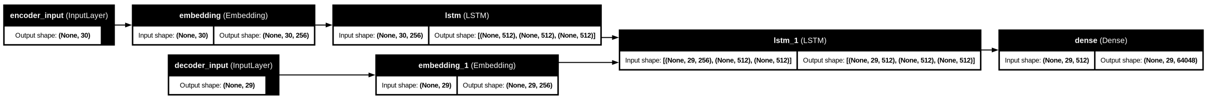

encoder_input (InputLayer)
Output shape: (None, 30)
embedding (Embedding)
Input shape: (None, 30)
Output shape: (None, 30, 256)
lstm (LSTM)
Input shape: (None, 30, 256)
Output shape: [(None, 512), (None, 512), (None, 512)]
decoder_input (InputLayer)
Output shape: (None, 29)
embedding_1 (Embedding)
Input shape: (None, 29)
Output shape: (None, 29, 256)
lstm_1 (LSTM)
Input shape: [(None, 29, 256), (None, 512), (None, 512)]
Output shape: [(None, 29, 512), (None, 512), (None, 512)]
dense (Dense)
Input shape: (None, 29, 512)
Output shape: (None, 29, 64048)

Figure 5.1: LSTM Model Details

5.2 Seq2Seq with Attention

To improve upon the limitations of the fixed-size context vector, the Seq2Seq model with Attention introduces an attention mechanism in the decoder. This enhancement allows the decoder to focus on different parts of the input sequence dynamically, especially beneficial when translating longer sentences. The key components of this model include:

- **Attention Layer:** This layer calculates attention weights α_{ti} for each encoder hidden state, which measure the relevance of each input token to the current decoding step. These weights help the model focus on the most pertinent parts of the input sequence during each output prediction.

- **Context Vector:** At each decoding step, the context vector c_t is computed as a weighted sum of the encoder's hidden states, allowing the decoder to adaptively retrieve information from different parts of the input sequence.

The attention-based decoding step is formulated as:

$$c_t = \sum_{i=1}^{T_x} \alpha_{ti} h_i \tag{5.2}$$

where c_t is the context vector at time t, and h_i represents the encoder hidden states. This attention mechanism significantly improves translation quality by allowing the model to retain and focus on relevant information from the source sentence dynamically.

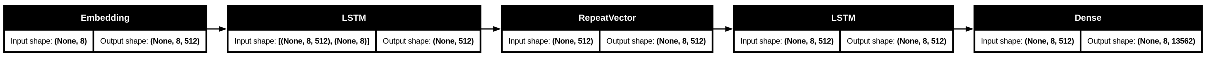

Embedding
Input shape: (None, 8)
Output shape: (None, 8, 512)
LSTM
Input shape: [(None, 8, 512), (None, 8)]
Output shape: (None, 512)
RepeatVector
Input shape: (None, 512)
Output shape: (None, 8, 512)
LSTM
Input shape: (None, 8, 512)
Output shape: (None, 8, 512)
Dense
Input shape: (None, 8, 512)
Output shape: (None, 8, 13562)

Figure 5.2: Seq2Seq Attention Model Details

5.3 Transformer with Transfer Learning

The Transformer model, an advanced architecture in neural machine translation, consists of stacked layers of self-attention and feed-forward networks. Unlike Seq2Seq models, Transformers enable parallel computation across all input tokens through self-attention, significantly improving training efficiency and translation quality. Key components include:

- **Input Embeddings:** The input tokens are mapped to embeddings of size 512. Positional encodings are added to retain sequence information, as Transformers do not have inherent sequential order.

- **Multi-Head Attention:** The model utilizes 8 attention heads per layer, each capturing unique relationships between tokens. Multi-head attention allows the model to focus on different parts of the sequence simultaneously, enhancing contextual understanding.

- **Feed-Forward Layers:** Each layer consists of a fully connected network with ReLU activation, enabling the model to learn non-linear transformations effectively.

- **Output Layer:** The final layer generates probabilities over the target vocabulary, producing the translation output.

To accelerate training and improve generalization on smaller datasets, we applied transfer learning by initializing our Transformer model with weights from a pre-trained model on a similar language task and fine-tuning it on our dataset. This approach enhances translation performance, especially when training data is limited.

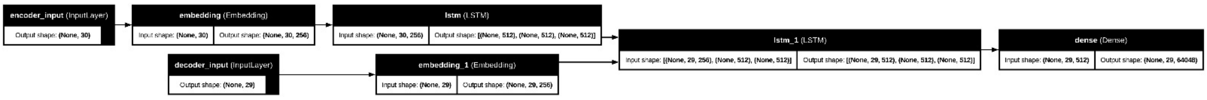

encoder_input (InputLayer)
Output shape: (None, 30)
embedding (Embedding)
Input shape: (None, 30)
Output shape: (None, 30, 256)
lstm (LSTM)
Input shape: (None, 30, 256)
Output shape: [(None, 512), (None, 512), (None, 512)]
decoder_input (InputLayer)
Output shape: (None, 29)
embedding_1 (Embedding)
Input shape: (None, 29)
Output shape: (None, 29, 256)
lstm_1 (LSTM)
Input shape: [(None, 29, 256), (None, 512), (None, 512)]
Output shape: [(None, 29, 512), (None, 512), (None, 512)]
dense (Dense)
Input shape: (None, 29, 512)
Output shape: (None, 29, 64048)

Figure 5.3: Transformer Model Details

5.4 Summary of Models

Model	**Embedding Dim**	**LSTM/ Transformer Layers**	**Attention Heads**	**Dropout**
Seq2Seq (LSTM)	256	1	-	0.5
Seq2Seq (Attention)	256	1	1	0.5
Transformer (Transfer Learning)	512	6	8	0.1

Table 5.1: Model Configurations

Chapter 6

Evaluation Metrics

To comprehensively assess our neural machine translation models, we employed three primary evaluation metrics: BLEU Score, Perplexity, and Accuracy. These metrics collectively provide insights into translation quality, model confidence, and precision, capturing different aspects of model performance and aiding in balanced evaluation.

- **BLEU Score:** The Bilingual Evaluation Understudy (BLEU) score is one of the most widely accepted metrics in machine translation, designed to measure the similarity between the machine-generated translation and a reference translation. It evaluates the overlap of n-grams (phrases of consecutive words) up to a specified length, commonly up to four words, across the model's output and reference translations. The BLEU score is essential in this context as it quantifies the overall quality of the translation, where a higher score indicates greater similarity to human-translated references. BLEU is particularly useful for capturing fluency and contextual accuracy in translation models.

- **Perplexity:** Perplexity serves as a measure of uncertainty in language models, indicating how confidently the model predicts the sequence of tokens in the target language. This metric is particularly relevant for assessing the model's internal structure and capacity to generate coherent sequences. Lower perplexity values reflect greater confidence and reliability in the model's predictions, translating to more consistent and contextually accurate translations. Perplexity thus helps in understanding model performance at the linguistic and contextual levels, which is critical in machine translation tasks.

- **Accuracy:** Accuracy measures the token-level match between the generated translation and the reference translation, providing a straightforward metric of precision. By calculating the proportion of correct token predictions, accuracy directly reflects how often the model's predictions align with the target language's syntax and vocabulary. This metric is important for gauging the precision of individual word choices within the translation, especially in high-stakes applications where word-level accuracy is essential. High accuracy in token prediction is indicative of effective language model training and alignment with reference translations.

Chapter 7

Results and Analysis

7.1 Model Performance

The following table summarizes the performance of our models on the WMT 14 EN-DE test set.

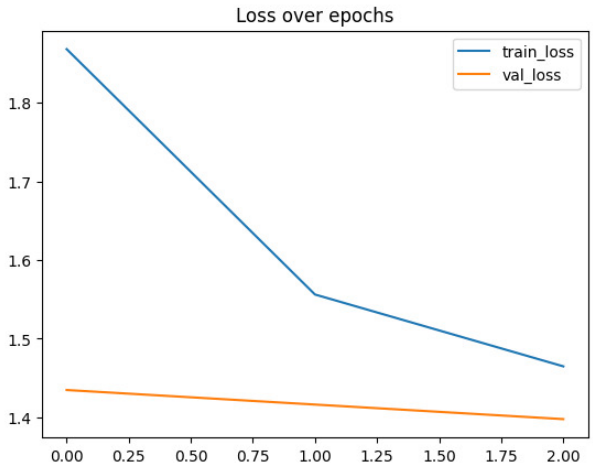


Figure 7.1: Loss during training and validation over epochs

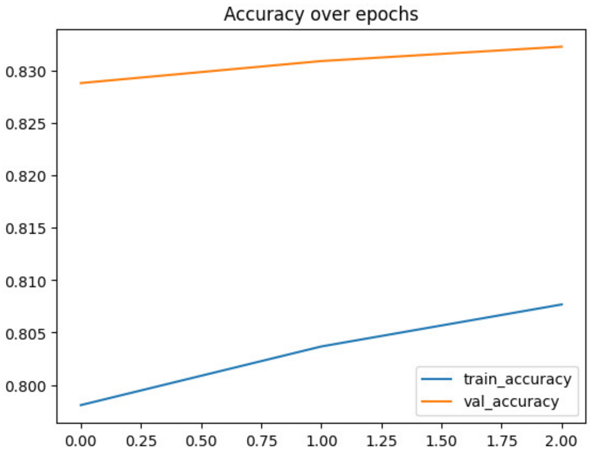
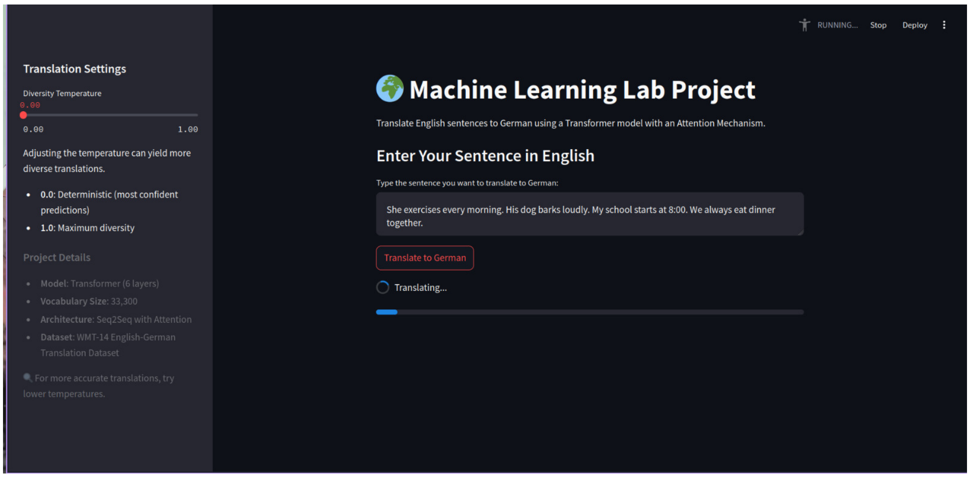


Figure 7.2: Accuracy during training and validation over epochs

Chapter 8

Streamlit GUI for Translation

We developed a Streamlit GUI for interactive translation. The GUI allows users to input English text and receive German translations. A screenshot of the GUI is shown below:



Translation Settings

Diversity Temperature

0.00

0.00 1.00

Adjusting the temperature can yield more diverse translations.

- **0.0:** Deterministic (most confident predictions)
- **1.0:** Maximum diversity

Project Details

- **Model:** Transformer (6 layers)
- **Vocabulary Size:** 33,300
- **Architecture:** Seq2Seq with Attention
- **Dataset:** WMT-14 English-German Translation Dataset

🔍 For more accurate translations, try lower temperatures.

Machine Learning Lab Project

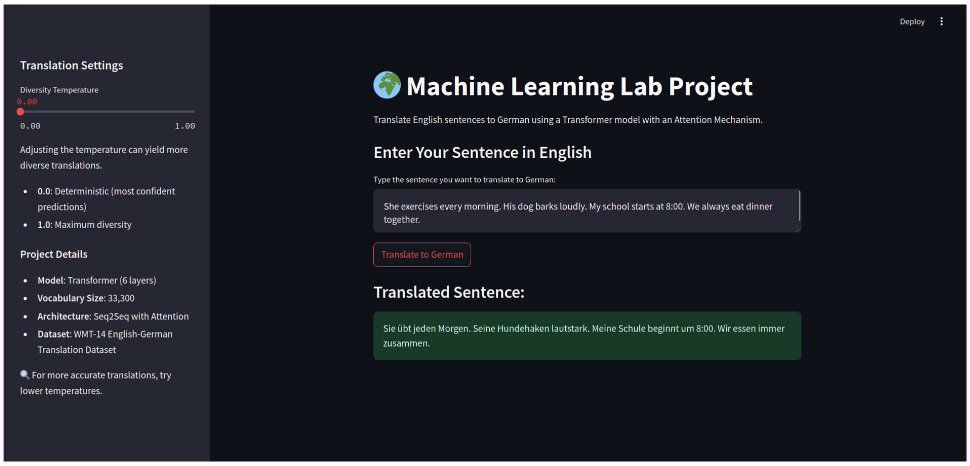
Translate English sentences to German using a Transformer model with an Attention Mechanism.

Enter Your Sentence in English

Type the sentence you want to translate to German:

She exercises every morning. His dog barks loudly. My school starts at 8:00. We always eat dinner together.

Figure 8.1: GUI during translation



Translation Settings

Diversity Temperature

0.00

0.00 1.00

Adjusting the temperature can yield more diverse translations.

- **0.0:** Deterministic (most confident predictions)
- **1.0:** Maximum diversity

Project Details

- **Model:** Transformer (6 layers)
- **Vocabulary Size:** 33,300
- **Architecture:** Seq2Seq with Attention
- **Dataset:** WMT-14 English-German Translation Dataset

🔍 For more accurate translations, try lower temperatures.

Machine Learning Lab Project

Translate English sentences to German using a Transformer model with an Attention Mechanism.

Enter Your Sentence in English

Type the sentence you want to translate to German:

She exercises every morning. His dog barks loudly. My school starts at 8:00. We always eat dinner together.

Translated Sentence:

Sie übt jeden Morgen. Seine Hundehaken lautstark. Meine Schule beginnt um 8:00. Wir essen immer zusammen.

Figure 8.2: Streamlit GUI for English-German Translation

Chapter 9

Future Work

There are several promising directions for future work that could improve upon the current model and extend its capabilities.

- **Dataset Expansion:** Increasing the dataset size could enhance model generalization and performance, particularly when dealing with nuanced language patterns or rare linguistic structures. Using larger and more diverse parallel corpora, such as the WMT 20 dataset or open multilingual resources, could help models achieve better accuracy and a richer understanding of language variability. This would be especially beneficial for fine-tuning models to handle informal language, idiomatic expressions, and specialized vocabularies.
- **Enhanced Architectures:** Future research could investigate more advanced architectures, such as BERT (Bidirectional Encoder Representations from Transformers) or GPT (Generative Pretrained Transformer) models, which have demonstrated state-of-the-art performance in natural language understanding and generation tasks. Integrating these models in translation tasks could enhance context comprehension and translation fluency. Specifically, using BERT's bidirectional attention could improve translation accuracy by capturing richer dependencies, while GPT-based models could offer improved language generation capabilities.
- **Multilingual Models:** Extending the translation system to a multilingual setting would allow a single model to translate between multiple languages. Recent advancements, such as multilingual BERT and mBART (Multilingual Bidirectional and Auto-Regressive Transformers), could be employed to develop models capable of translating between various language pairs without retraining. This approach could improve the efficiency of translation systems, particularly for low-resource languages, by enabling shared learning across language pairs.
- **Domain Adaptation and Customization:** A promising area for future exploration is domain-specific translation. Models could be fine-tuned on specialized datasets, such as medical or legal documents, to handle domain-specific terminology and style. This could make the model more adaptable to professional applications, where translation accuracy is critical.
- **Evaluation Metrics Enhancement:** Lastly, exploring enhanced evaluation metrics that go beyond n-gram overlap, such as METEOR, TER, and BERTScore, could provide a more comprehensive assessment of translation quality. These metrics, which consider factors like synonymy and semantic alignment, could give a more accurate reflection of the model's performance, particularly in cases where literal translation is less important than meaning preservation.

Chapter 10

Conclusion

This project highlights the effectiveness of Seq2Seq and Transformer architectures in neural machine translation, underscoring the significance of attention mechanisms and transfer learning in achieving high-quality translations. By comparing models with and without attention, we demonstrated that attention allows the model to focus on specific parts of the input sequence, significantly enhancing translation quality, especially for longer and more complex sentences. The Transformer model, leveraging transfer learning, further improved translation performance by reducing the required training time and allowing the model to generalize better on unseen data.

The results underscore that integrating transfer learning from pre-trained language models is a valuable approach for machine translation, as it not only boosts performance but also optimizes computational resources. The application of such models has promising implications for real-world translation systems, as they can quickly adapt to new languages and specialized domains.

As advancements in neural network architectures and language processing continue, machine translation models will likely see even greater improvements in accuracy and efficiency. Future research can build upon this work by exploring larger datasets, more advanced architectures, and multilingual models, paving the way for systems capable of handling a broader range of languages and contexts. The continued evolution of neural machine translation offers a powerful tool for bridging language barriers, making information accessible on a global scale.