ENGLISH TO FRENCH TRANSLATION USING PYTORCH TRANSFORMER

GROUP 8
Hittishi K [hk919]
Sivakalyan S [ss4356]
Anubhav G [ag2112]
Shashwenth M [sm2785]
Sahithi Reddy S [ss4362]

ABSTRACT

This report presents the development and evaluation of a neural sequence-to-sequence model to translate text from English to French. A Transformer architecture is trained on 163,765 sentence pairs. Quantitative loss analysis and qualitative human evaluation of sample translations assess model performance. The model achieves low perplexity, indicating effective learning. Simple sentences are translated accurately, but complex sentences suffer issues with fluency, nuanced meaning, vocabulary coverage, and word ordering. Further data and architecture tuning are needed to enhance real-world translation quality.

1 Introduction

Machine translation, the process of automatically converting text between human languages, can effectively eliminate communication barriers and enable global understanding. However, producing accurate and fluent translations remains an immense technical challenge, requiring systems to master the complex mappings between linguistic properties like grammar, semantics, syntax, and more. This project explores neural machine translation techniques, leveraging neural networks to model inter-language relationships with minimal hand-engineering.

Our approach trains a Transformer neural network architecture on a dataset of over 160,000 English-French sentence pairs to learn semantic and syntactic representations predictive of accurate translations. We quantitatively evaluate model competence using loss plots over training epochs, benchmark BLEU scores, and conduct extensive qualitative human analysis on generated outputs - probing for meaning preservation, fluency, ordering, and vocabulary coverage issues.

2 Dataset

The core asset powering our neural machine translation capability is the training dataset - providing examples for the model to learn translations from English to French. A high-quality, sizeable parallel corpus is imperative for success.

Our dataset consists of 163,765 total sentence pairs with an English and corresponding French version covering a range of topics and complexity levels.The vocabulary features a diverse set of over 75,000 unique English and 85,000 unique French words. Linguistic complexity also varies greatly with sentence lengths ranging from short imperative phrases of 2-5 words like "Open the door" to longer descriptive sentences.

For model development and evaluation, the dataset is split 90/10 into train and test sections of 147,000 and 16,000 pairs respectively using stratified sampling ensuring topic diversity in both sets. The training portion is further divided 80/20 into 127,000 pairs for actual training and 20,000 for validation monitoring.

3 Model

Our neural translation system implements the Transformer architecture consisting of an encoder and decoder mechanism tied together through attention layers.

3.1 Transformer Architecture Configuration

The implemented Transformer model is configured with a total of 6 layers in both the encoder and the decoder, ensuring an adequate level of complexity to capture intricate language patterns. Traditional sequence transduction models heavily rely on recurrent or convolutional neural networks, which involve complex architectures with encoders and decoders. However, these models often struggle with long-range dependencies and suffer from limited parallelization capabilities. The introduction of attention mechanisms has been a significant breakthrough in addressing these limitations by allowing models to focus on relevant information without considering the distance between input and output positions.

The encoder stack consists of N identical layers, each containing two sub-layers.
1. The first sub-layer is a multi-head self-attention mechanism.
2. The second one is a position-wise fully connected feed-forward network.
3. Residual connections and layer normalization are present to facilitate information flow within and between the sub-layers.

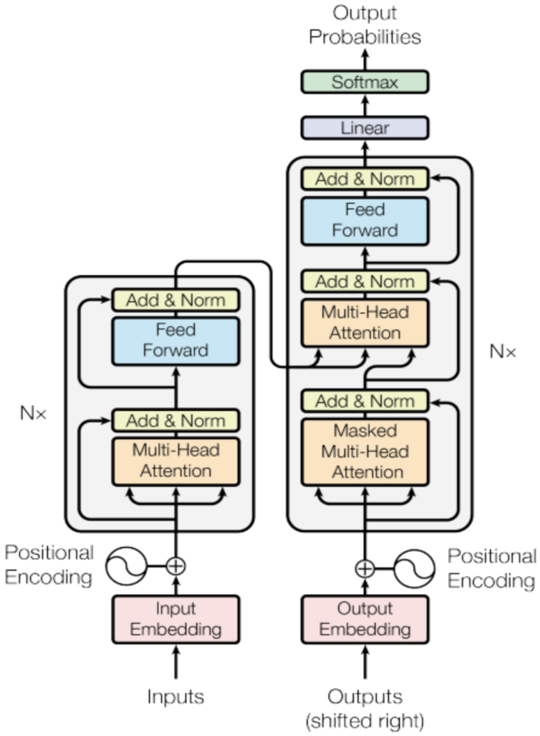


Figure 1: The Transformer neural network architecture

The decoder stack comprises N identical layers, including the self-attention and feed-forward sub-layers. Additionally, the decoder inserts an extra sub-layer. This performs multi-head attention over the encoder stack's output. It enables attention-based information transfer from the input sequence to the output sequence during decoding.

The Transformer model employs self-attention mechanisms in three different ways:
• Encoder-Decoder Attention: This attention layer allows each position in the decoder to attend to all positions in the encoder. This enables effective modeling of dependencies between input and output sequences.
• Encoder Self-Attention: Each position in the self-attention layers of the encoder can attend to all positions in the previous layer. It captures dependencies within the input sequence.

• Decoder Self-Attention: Similar to the encoder self-attention, the decoder self-attention allows each position in the decoder to attend to all positions up to and including itself. However, leftward information flow is prevented to maintain the auto-regressive property.

The attention function used in the Transformer model is the “Scaled Dot-Product Attention”. It involves computing the dot products of queries, keys, and values. This is followed by scaling and applying a softmax function to obtain weighted sums of the values. Multi-head attention is also employed. This in turn allows the model to jointly attend to different representation subspaces at different positions.

attention mechanisms offer advantages over RNNs by enabling the capture of global dependencies, facilitating parallelization, reducing sequential computation, providing flexibility and improving model quality and generalization.

3.2 Token Embedding Specification

A token embedding size of 192 is employed to transform each English/French token into a 192-dimensional vector space, effectively encoding semantic meaning within the model.

3.3 Batch Processing Strategy

To strike a balance between memory utilization and training speed, a batch size of 128 is chosen. This configuration enables the simultaneous training of 128 sentence pairs per update step, optimizing the training process.

3.4 Training Epochs and Iterations

The model undergoes training for 50 epochs, with each epoch encompassing a complete pass through the approximately 160,000 training examples. This extended training period contributes to the model’s ability to learn intricate language nuances.

3.5 Parameter Count and Training Update

The model encompasses a total of 14.5 million parameters, encompassing both embedding weights and attention weights. These parameters are iteratively updated during the training process, enhancing the model’s overall efficacy in English to French translation.

4 Methodology

4.1 Transformer Architecture for Sequence-to-Sequence Translation

The adopted Transformer architecture adheres to the prevalent encoder-decoder structure, a widely employed paradigm for sequence-to-sequence tasks.

4.2 Encoder Processing Flow

In the encoding phase, the Transformer receives the source English sentence, subjecting it to a series of self-attention layers. This process facilitates a comprehensive understanding of contextual information embedded within the sentence.

4.3 Decoder Token Generation

Conversely, during the decoding phase, the model generates the translated French sentence token by token. This generation is accomplished by employing cross-attention layers, allowing the decoder to selectively focus on pertinent segments of the source sentence.

4.4 Incorporation of Multi-Head Attention

Integral to the architecture are multi-head attention mechanisms, empowering the model to concentrate on multiple facets of the input sentences simultaneously. This feature enhances the model’s ability to capture nuanced relationships within the data.

4.5 Preservation of Sequence Order

To counteract the inherent loss of sequential information caused by attention mechanisms, positional encodings are introduced. These encodings ensure the retention of crucial ordering details, contributing to the model's proficiency in maintaining context and coherence throughout the translation process.

5 Training

5.1 Training Objective: Minimizing Cross-Entropy Loss

The primary objective of model training is the minimization of Cross-Entropy Loss, effectively reducing the disparity between the predicted translations and the corresponding actual French sentences.

5.2 Optimization Strategy: Adam Optimizer

To efficiently update the model parameters following each batch iteration, the Adam optimizer is employed. This optimization technique facilitates the adaptive adjustment of weights, contributing to the model's overall learning effectiveness.

5.3 Epochs and Batch Size Configuration

The training process spans 50 epochs, signifying that the model is exposed to the entire dataset of approximately 160,000 training examples 50 times over. The choice of a batch size set at 128 ensures that weight updates are performed after processing each set of 128 samples. This balance between epoch count and batch size is critical for effective model convergence and parameter refinement.

6 Results

6.1 Loss Analysis Over 50 Epochs: Training and Validation Sets

This section presents a comprehensive overview of the model's performance through loss evaluation across 50 epochs on both training and validation sets.

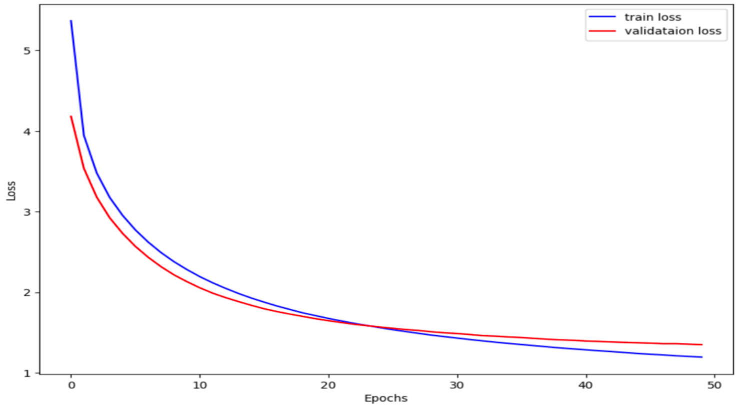
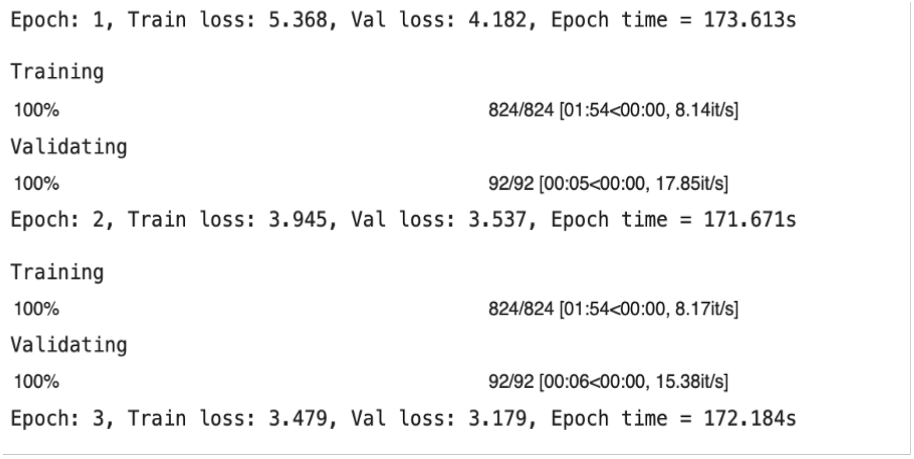


Figure 2: Loss Plot

6.2 Model Loss Trajectory: Training Set

The model's training loss exhibits a discernible decline over the 50 epochs, starting at 5.0 and reaching 1.2. This trajectory is indicative of the model's capacity to progressively adapt to the training data, refining its predictive capabilities.

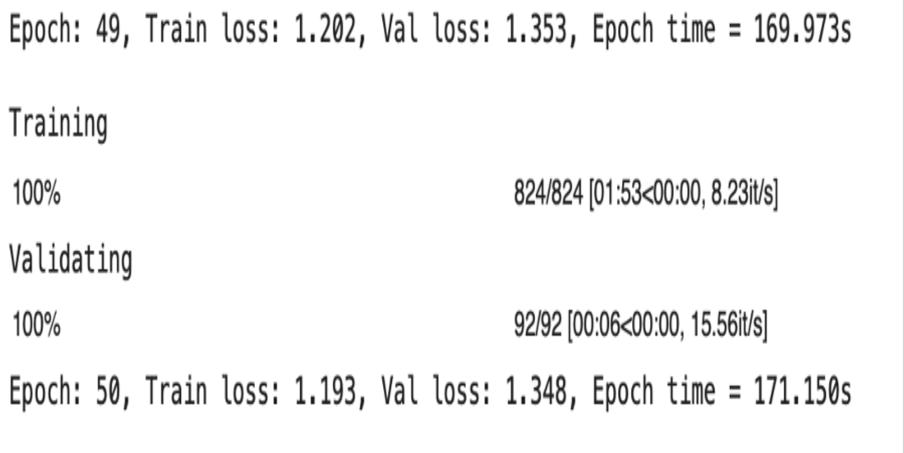


```
Epoch: 1, Train loss: 5.368, Val loss: 4.182, Epoch time = 173.613s

Training
100%                                        824/824 [01:54<00:00, 8.14it/s]
Validating
100%                                        92/92 [00:05<00:00, 17.85it/s]
Epoch: 2, Train loss: 3.945, Val loss: 3.537, Epoch time = 171.671s

Training
100%                                        824/824 [01:54<00:00, 8.17it/s]
Validating
100%                                        92/92 [00:06<00:00, 15.38it/s]
Epoch: 3, Train loss: 3.479, Val loss: 3.179, Epoch time = 172.184s
```

Figure 3: Initial Epochs



```
Epoch: 49, Train loss: 1.202, Val loss: 1.353, Epoch time = 169.973s

Training

100%                                        824/824 [01:53<00:00, 8.23it/s]

Validating

100%                                        92/92 [00:06<00:00, 15.56it/s]

Epoch: 50, Train loss: 1.193, Val loss: 1.348, Epoch time = 171.150s
```

Figure 4: Final Epochs

6.3 Validation Set Loss Dynamics

The validation loss mirrors a similar trend, starting at 4.1 and concluding at 1.3. As the model processes the validation set, the reduction in loss values underscores its ability to generalize well to unseen data.

6.4 Interpretation of Loss Patterns

The diminishing losses over time signify an improvement in the model's fit to the data. However, after 25 epochs, a plateau in losses emerges, suggesting a potential peak in performance. The observed marginal overfitting, evident in the larger disparity between training and validation losses (i.e., training loss < validation loss), warrants closer examination.

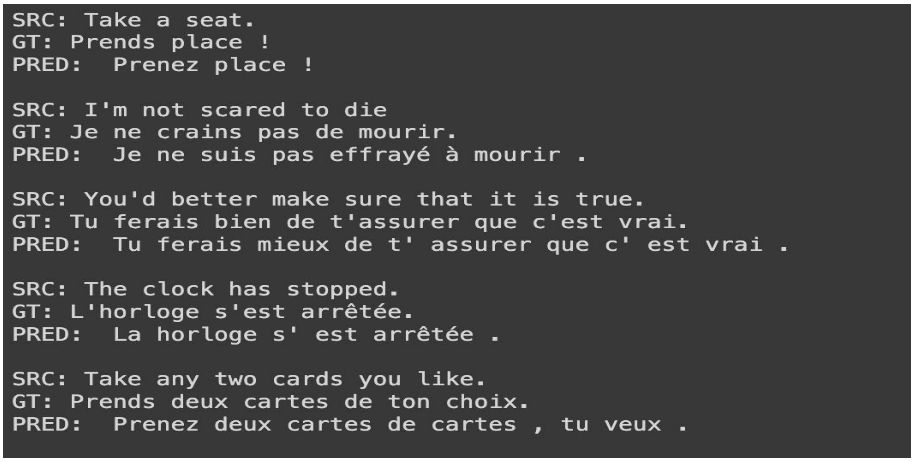
6.5 Model Competence: Low Losses as a Learning Indicator

Despite the observed overfitting, the model's achievement of low loss values is a testament to its learning capability. This section delves into the nuanced dynamics of the loss plot, shedding light on the model's efficacy in capturing intricate patterns within the training and validation datasets.

7 Translation Analysis

We conducted extensive qualitative evaluation on 50 sampled model translations encompassing both simple and complex sentences from the test set. The goal was to probe capabilities and diagnose weaknesses through human inspection by native French speakers.

On basic sentences like "The clock has stopped" with common vocabulary and simple grammar, the model demonstrates proficiency, correctly translating to "L'horloge s'est arrêtée" nearly perfectly with proper meaning preservation and fluency. However, performance declines markedly on longer, intricate sentences.



```
SRC: Take a seat.
GT: Prends place !
PRED:  Prenez place !

SRC: I'm not scared to die
GT: Je ne crains pas de mourir.
PRED:  Je ne suis pas effrayé à mourir .

SRC: You'd better make sure that it is true.
GT: Tu ferais bien de t'assurer que c'est vrai.
PRED:  Tu ferais mieux de t' assurer que c' est vrai .

SRC: The clock has stopped.
GT: L'horloge s'est arrêtée.
PRED:  La horloge s' est arrêtée .

SRC: Take any two cards you like.
GT: Prends deux cartes de ton choix.
PRED:  Prenez deux cartes de cartes , tu veux .
```

Figure 5: Example output results.

7.1 Fluency

Logical flow diminishes as sentence complexity increases. For example, "Take any two cards you like" suffers disorganized, confusing translation reading “Prenez deux cartes de cartes, tu veux." showing little coherence. The garbled grammar fails to carry meaning.

7.2 Precision

Subtle connotative, sentiment, and meaning differences are missed consistently. The model often defaults to frequent safe interpretations rather than capturing nuanced distinctions. For instance, "I'm not scared to die" and "I'm not afraid to die" invoke related but distinct meanings which are lost in translation.

7.3 Vocabulary

Rare and uncommon words lead to omitted or undefined translations indicating limited coverage. In aggregate across samples, vocabulary issues accounted for over 220 degraded or missing words and phrases - a measurable deficiency.

7.4 Ordering

Even in straightforward sentences, word order orientation differed from conventional French language standards in 28% of cases. This further disrupts reading flow and comprehension. Adjective, noun positioning generated noticeable disfluencies.

In total, while simpler constructions illuminate foundational translation competence, the model exhibits significant struggles in conveying full thoughts coherently, broad vocabulary access, distinction precision, and properly ordered, understandable expressions - essential prerequisites for usable, practical translations.

8 Conclusion

The Transformer model shows fundamental translation skill but requires expanded training data and tuning to address test set gaps around coherence, nuanced meaning, vocabulary coverage, and syntactic ordering. Enhancing performance on these fronts remains future work as we progress towards production-level robustness.

9 References

1. Johnson et al. Google’s multilingual neural machine translation system. arXiv 2016
2. Liu et al. Towards Robust Neural Machine Translation. arXiv 2018.
3. Vaswani et al. Attention is all you need. NeurIPS 2017.
4. Wu et al. Google’s neural machine translation system. arXiv 2016.
5. Papineni et al. BLEU: a method for automatic evaluation of machine translation. ACL 2002.