Neural Machine Translation with Transformers

1. Introduction

Neural Machine Translation (NMT) is a Natural Language Processing (NLP) technique that uses deep neural networks to automatically translate text from one language to another. In this project, we implemented a Transformer-based model to translate sentences from English to French.

2. Dataset Used

We used a subset of the Tatoeba dataset consisting of 50,000 English-French sentence pairs. This limited dataset size is due to hardware and computational resource constraints.

3. Preprocessing

- Removal of punctuation and digits

- Conversion to lowercase

- Unicode normalization

- Special tokens [start] and [end] added to French sentences

4. Tokenization and Sequence Padding

Two separate tokenizers were trained: one for English and one for French. Padding was applied to all sequences to ensure a uniform maximum sequence length of 15 tokens.

5. Model Hyperparameters

Parameter	Value
Architecture	Transformer Encoder-Decoder
Number of attention heads	8
Embedding dimension	256
English vocabulary size	6,048 words
French vocabulary size	12,197 words

Max sequence length	15 tokens
Trainable parameters	9,784,229

6. Model Architecture Details

The model is composed of the following layers:
- Encoder Input Layer: Accepts integer-encoded English sentences
- Token and Position Embedding for the encoder
- TransformerEncoder block with 8 attention heads
- Decoder Input Layer: Accepts integer-encoded French sentences
- Token and Position Embedding for the decoder (with mask_zero=True)
- TransformerDecoder block
- Dropout Layer: Rate = 0.4
- Final Dense Layer: Softmax activation over the French vocabulary

7. Training

The model was trained for a maximum of 50 epochs with EarlyStopping based on validation accuracy.

Epoch	Training Accuracy	Validation Accuracy	Validation Loss
1	0.8612	0.6176	0.5650
11	0.8917	0.7563	0.4212
13	0.8894	0.7552	0.4356

EarlyStopping was triggered after epoch 13 due to validation accuracy stagnation.

8. Conclusion

The Transformer model trained on a limited dataset showed promising results with a final validation accuracy of 75.63%. Better results can be expected with:
- A larger dataset
- Extended training time
- More powerful computational resources

- Next Steps:
- - Use larger corpora such as OpenSubtitles or WMT datasets
- - Experiment with learning rate scheduling and label smoothing
- - Add beam search decoding for improved translation quality