Neural Machine Translation

Prerna Tripathi

June 17, 2025

1. EDA on the dataset

The dataset used the Multi30K dataset.The dataset was already split into train, validation and test datasets having features "en" for english sentence and "de" with the corresponding german sentence. The train dataset had about 29000 sentence pairs, the validation dataset 1014 sentence pairs and the test dataset had about 1000 sentence pairs.

2. Data Preprocessing

It begins by tokenizing English and German sentences using spaCy, with English tokens reversed to improve learning in the encoder-decoder model. It then builds vocabularies for both source and target languages, including special tokens like PAD, UNK, SOS, and EOS, and filters out infrequent tokens based on a minimum frequency threshold. Next, it converts each tokenized sentence into a sequence of indices using the respective vocabularies (numericalization), appending ¡SOS¿ and ¡EOS¿ tokens to mark sentence boundaries. Finally, it prepares PyTorch DataLoaders for training, validation, and testing by padding sequences in each batch to the same length and converting them to tensors, enabling efficient mini-batch training.

3. Seq2Seq without Attention

The seq2seq model without attention is built on the encoder decoder architechture with the encoder being made of bidirectional lstms with 2 layers in the encoder. The hyperparameters like the hidden state size is set at 512, the word embeddings are also learned while training,no pretrained embeddings are used with the embedding dimension set at 256. The teacher forcing ratio is set at 0.5 such that during training 50 percent of the time the actual output tokens are used as input and the other times the predictions of the decoder are only used as the input. The model is trained using the Adam optimiser with the default learning rate of

0.001 and the number of epochs set at 17 since after that the validation loss was continuously increasing which means the model was overfitting the training dataset. Then by tracking the validation loss during train the best model was saved and finally after training the model was tested on the test dataset and the bleu score was calculated. BLEU score:0.19

4. Seq2Seq with Attention

The seq2seq model with attention was similar to the prev model with the only difference being the addition of Luong Attention which calculates the alignment scores by taking the dot product of the current hidden state of the decoder with all the hidden states of the encoder and then concatenating the weighted sum with the current output of the decoder. Masking is also applied for tokens like ¡PAD¿ which are of no use in determining the output and hence need to be masked from the attention calculation. The attention mechanism led to quite an improvement in the BLEU score from the previous part as the model now learns on which parts to focus more on to predict a given token. Rest all the hyperparameters are same but the no. of epochs is set at 15 as the validation loss was increasing continuously and the model was overfitting on the training datatset. Then the BLEU was calculated. To clearly visualize the effect of attention the attention heat map was also printed in order to understand for a given word where the model focus on the most which helps us understand the working behind attention. BLEU score:0.24

5. Analysis of the effect of attention mechanism

Although by adding the Luong attention my BLEU score increased by a decent 0.05 units but the attention heat maps showcase that the attention is following a diagonal pattern mostly. We can see most of the predicted translations are very close to the reference translations. But the attention heat maps are following a diagonal pattern and is not focussing exactly on the correct token but still the BLEU improved and the translations are close to correct. This could be due to the translations are more positional than being contextual or semantic but this is a failure of this attention model and must be improved.

6. Transformer

This code fine-tunes the pre-trained facebook/mbart-large-50-one-to-many-mmt model for English-to-German translation using the Multi30k dataset from Hugging Face. It sets up the tokenizer with source and target languages (en for english and de for german), tokenizes both

inputs and targets with truncation to a maximum length of 128, and maps this preprocessing function across the dataset splits. The model is trained using Seq2SeqTrainer with a custom compute metrics function that evaluates BLEU score using sacrebleu. Special care is taken to replace -100 (ignored tokens) with padding tokens before decoding predictions and references. Training is performed for 3 epochs with model checkpoints saved and the best model loaded based on validation loss. After training, the final BLEU score is evaluated on the test set. Due to time constraints and time taking fine tuning of the pre trained model the training was done for only 3 epochs. Performance can be improved furthur by increasing the number of epochs. BLEU score:0.45