1 Data Preparation

1.1 Pattern Observation

Most words in the dataset are full words, not parts of words. This suggests that the dataset is well-organized. Also, things like punctuation marks and numbers are separate, which makes it easier to read the data.

1.2 Challenges of Using Whole Words

If we use whole words, the size of our vocabulary would be very large, around 130,000 unique words. This is too much when we have only 3 million words for training. That's why we use a method called Byte Pair Encoding (BPE) to make the vocabulary smaller, down to 10,000 words.

2 Training

2.1 Sentence Pairs in Training and Dev Sets

We have 174,443 pairs of sentences for training and 2,052 for development. The large number for training helps the model learn better. However, the small number for development might not be enough for testing how well the model works.

2.2 Encoder and Decoder Layers

Both the encoder and decoder have 6 layers. This is a common setup, and it's a good balance between being fast and effective.

2.3 Vocabulary Size for Source and Target Languages

The source language has 6,104 words in its vocabulary. The target language has 7,600. These numbers are less than the 10,000 we expected, maybe due to specific things about the dataset or how we prepared it.

2.4 Embedding Table Dimensions

We have tables that turn words into numbers. These tables have sizes of $6,104 \times 512$ and $7,600 \times 512$ for the source and target languages, respectively.

2.5 Model Parameters

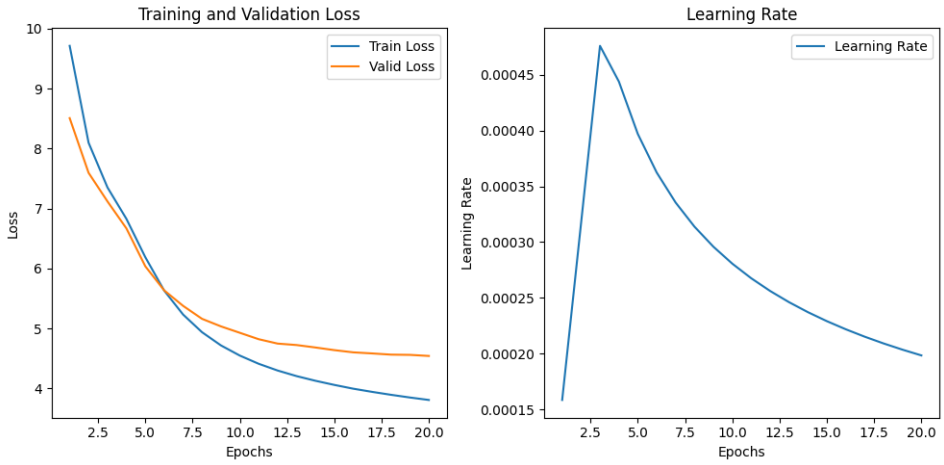
The model has around 51 million adjustable settings, known as parameters. This allows the model to learn a lot of things but also risks learning the training data too well, and not being useful for new data.

2.5.1 Training Objective

The model uses a combined technique called LabelSmoothedCrossEntropy-Criterion to learn better from the training data.

2.5.2 Learning Rate Adaptation

The learning rate goes up and then down during training.

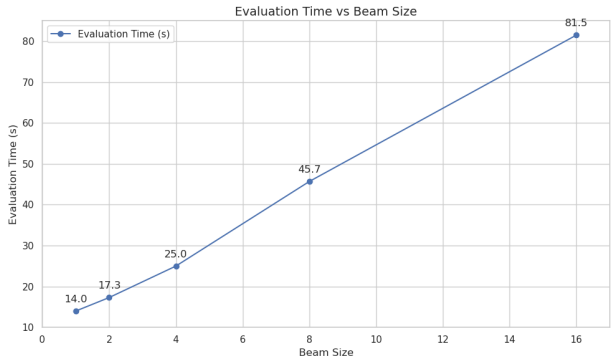


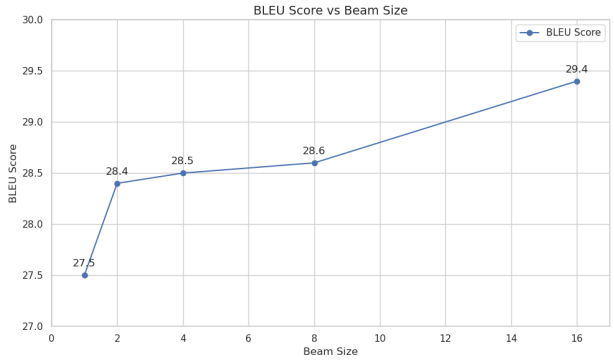
2.5.3 Intuition Behind Learning Rate Adaptation

The learning rate starts high to help the model learn fast. Then, it slows down to help the model fine-tune its learning.

3 Evaluation

Beam Size	BLEU Score	Evaluation Time
1	27.5	14.0s
2	28.4	17.3s
4	28.5	25.0s
8	28.6	45.7s
16	29.4	81.5s





4 Further Considerations

4.1 Impact of ASR Errors on MT

Errors from Automatic Speech Recognition (ASR) can make the text less accurate and affect translation quality.

4.2 Adaptation for Audio Inputs

If we use audio as input, we'd have to change the encoder to understand audio data. The decoder can stay the same if we still want text as output.

5 Notes

5.1 Unusual Vocabulary Sizes

The observed vocabulary sizes don't align with the expected 10,000 tokens based on Byte Pair Encoding (BPE).

5.2 Tokens in Test and Dev Sets

Some tokens appear in the Test and Dev sets but are not present in the Training set, raising questions about data consistency.

5.3 Mismatch in Transformer Input Size

The size of the input for the Transformer model doesn't match the vocabulary sizes observed in the datasets.

5.4 Discrepancy in Token Counts

For English, there are 6,098 unique tokens in the training set, but the Transformer takes an input size of 6,104. This could be due to the addition of the Start-of-Sentence token, among other possible factors.

For German, there are 7,596 unique tokens in the training set, but the Transformer has an input size of 7,600. This might include the End-of-Sentence token and other unknown elements.

5.5 Underutilized BPE Tokens

Only about half of the expected 10,000 BPE tokens are actually used in the training set, requiring further investigation.

5.6 Extreme Train-Test Split

The dataset is heavily skewed towards training, with 96% of the data used for training. The reason for this extreme split needs to be understood.