TECHNISCHE UNIVERSITÄT WIEN

Revisiting the Transformer Model from Attention Is All You Need: Is Attention Almost All You Need?

Deep Learning for Natural Language Processing

Prepared by

The Capybaras Team

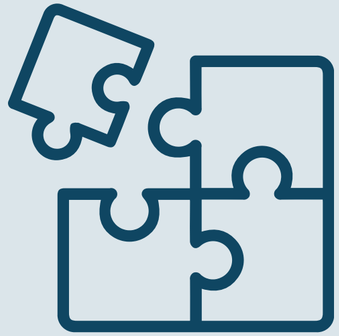
Sofiia Zholubak, Meryem Ben yahia, Cveta Capova

30 January, 2025

Project Objectives

Attention Is All You Need

- Understand the transformers architecture



Transformer Implementation

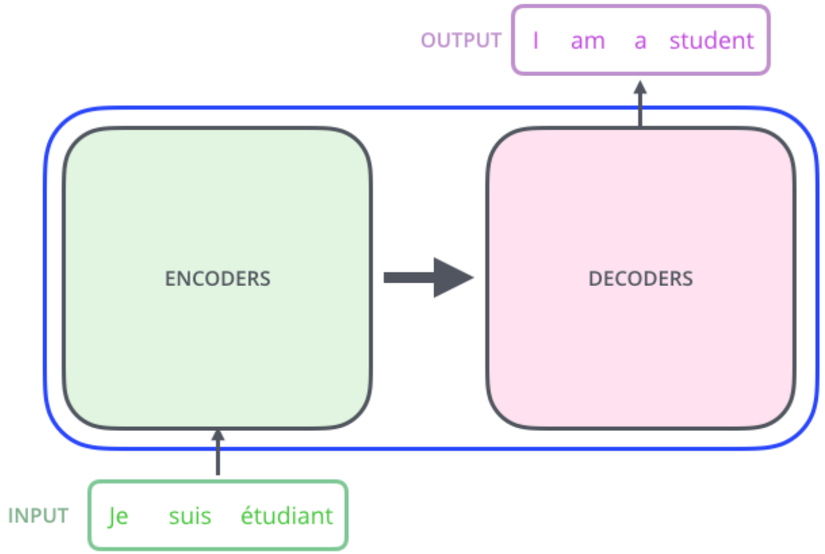
- From Scratch Transformer Models
- Hugging Face Transformer

Evaluate

- Evaluate the performance of different model configs
- Compare them

Transformer Architecture

High-level view



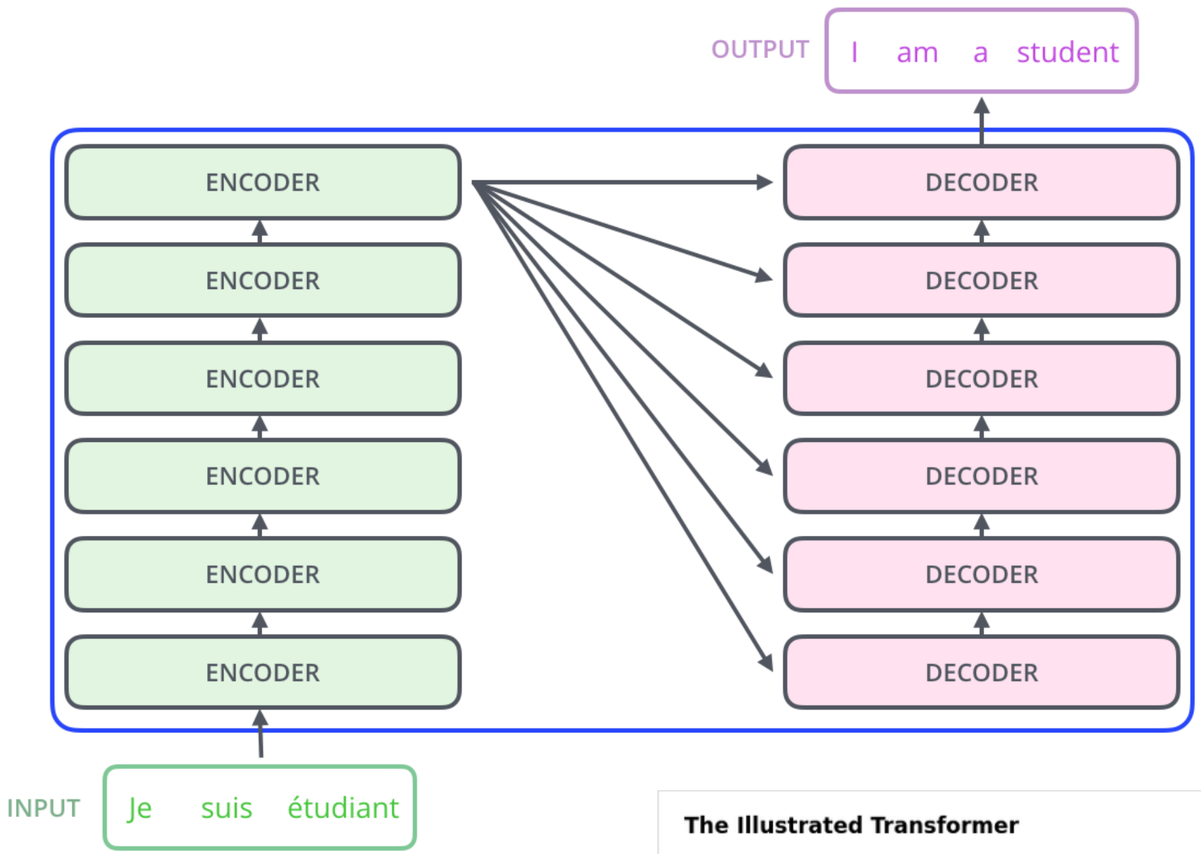
The Illustrated Transformer

Discussions: Hacker News (65 points, 4 comments), Reddit r/MachineLearning (29 points, 3 comments) Translations: Arabic, Chinese (Simplified) 1, Chinese (Simplified) 2, French 1, French 2, Italian,...

github.io /

Transformer Architecture

The encoder module consists of multiple encoders stacked on top of each other: in the paper's example, six are used. The decoder module likewise consists of the same number of stacked decoders.



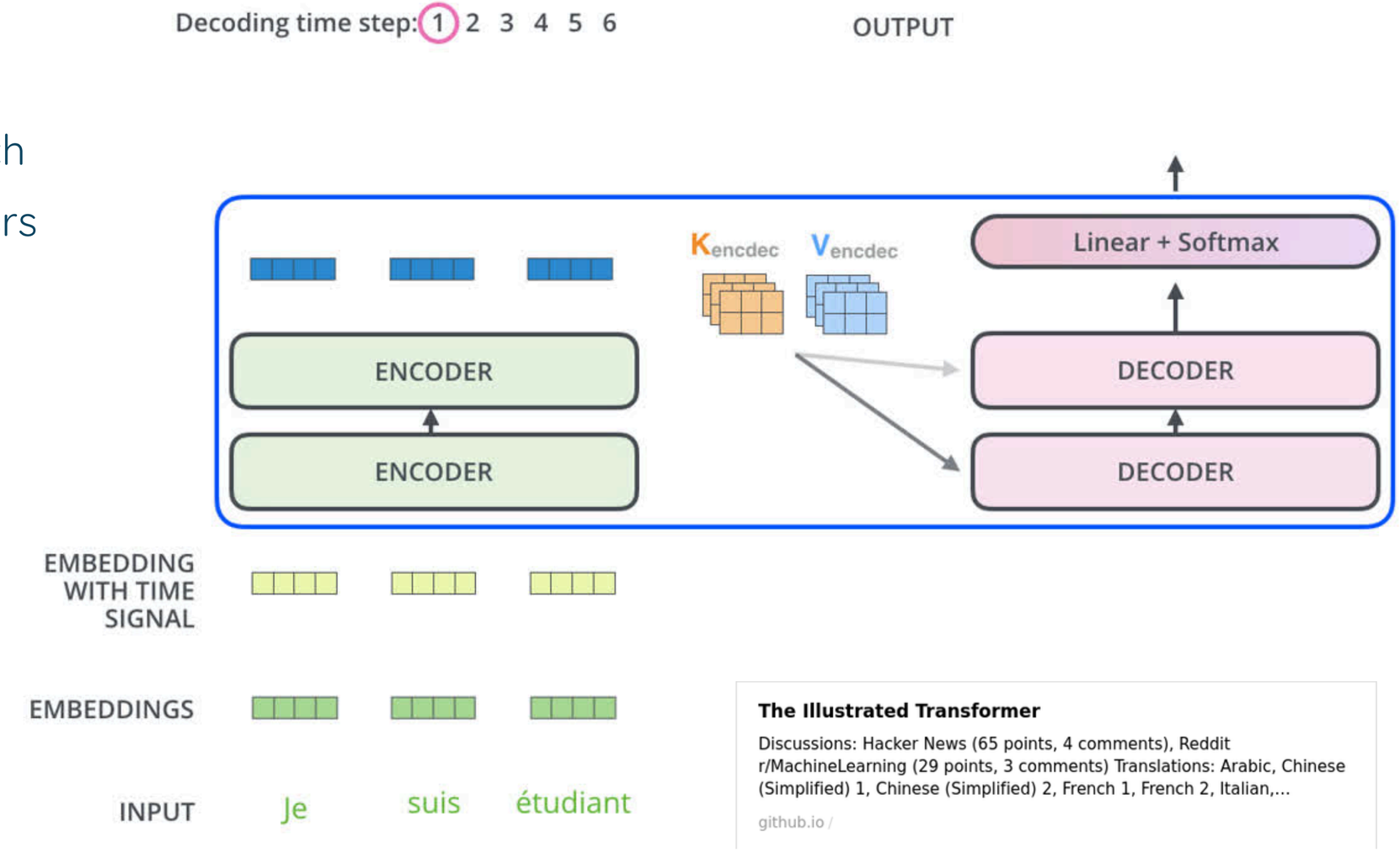
The Illustrated Transformer

Discussions: Hacker News (65 points, 4 comments), Reddit r/MachineLearning (29 points, 3 comments) Translations: Arabic, Chinese (Simplified) 1, Chinese (Simplified) 2, French 1, French 2, Italian,...

github.io /

Transformer Architecture

The final encoder's output is converted into two sets of attention vectors, K and V. Each decoder then uses these vectors within its "encoder-decoder attention" layer, enabling it to focus on the relevant parts of the input sequence.



Transformer Architecture

Some bad news!

- **Resource-intensive Training**
- **Memory Footprint and Quadratic Complexity:** The standard self-attention operation grows $O(n^2)$ in time and memory with sequence length n.
- **Very Data Hungry**
- **Attention Doesn't Equal Explanation:** Internal Representations are complex.
- **Fine-Tuning is expensive given all the above factors!**

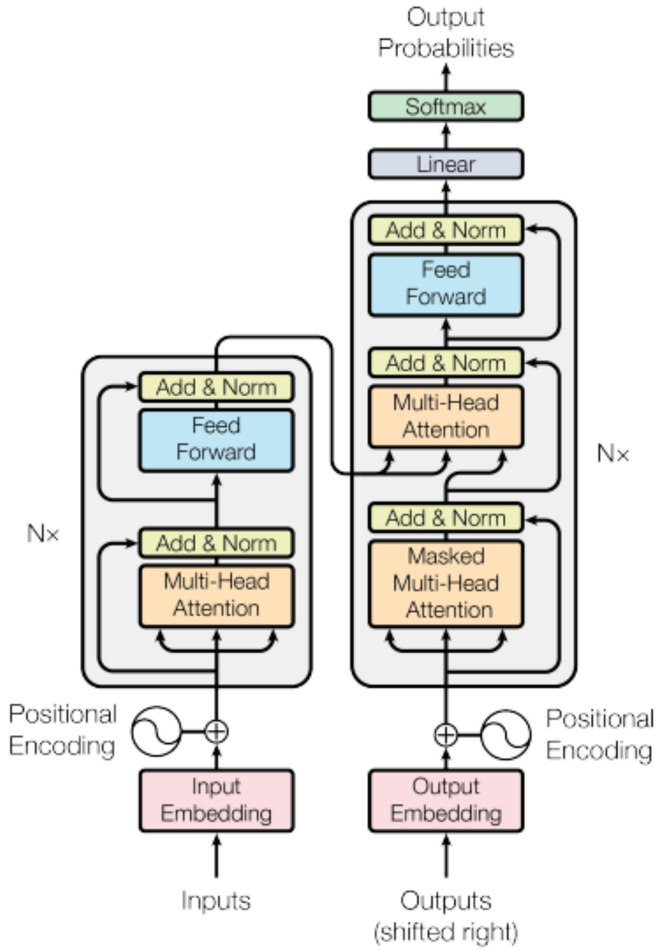
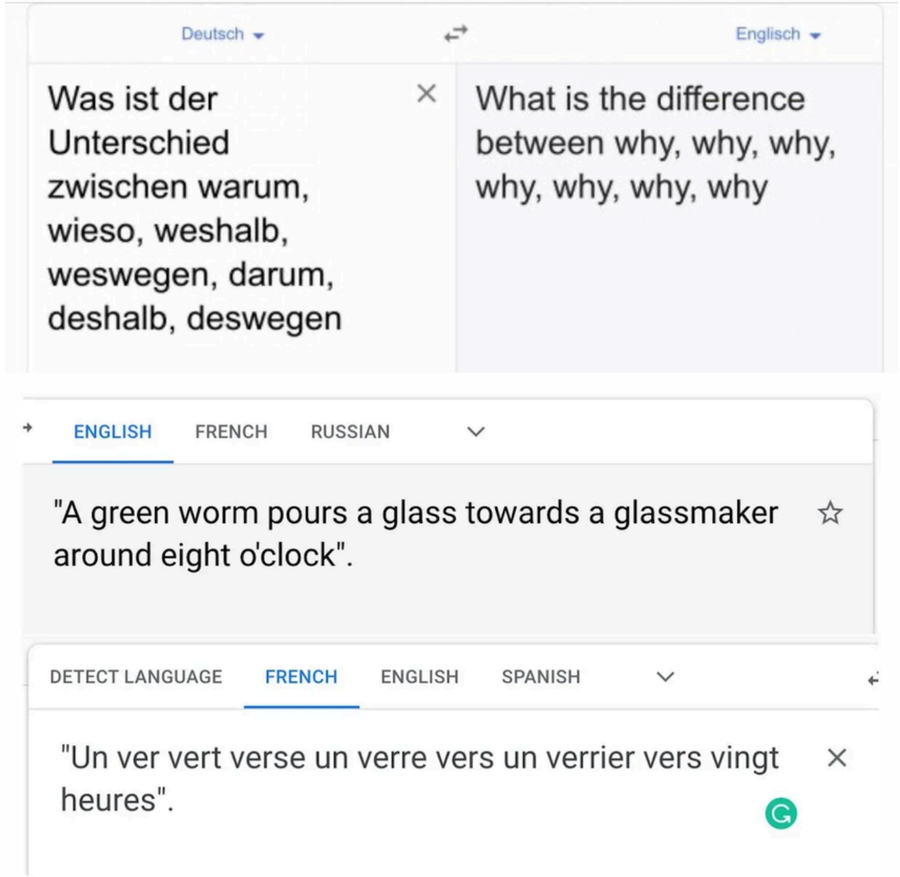


Figure 1: The Transformer - model architecture.

Data Overview

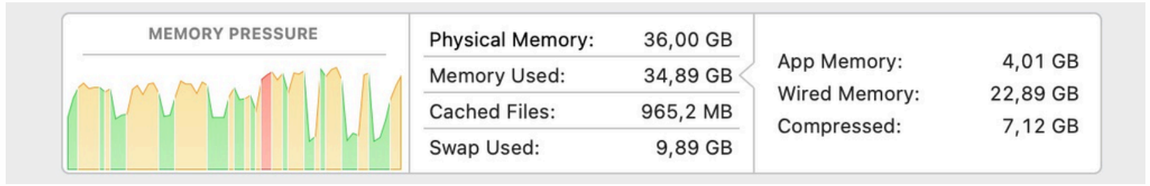
- **WMT 2014 English-German dataset**
  - consisting of about 4.5 million sentence pairs
- **WMT 2014 English-French dataset**
  - consisting of 36M sentences

- **Train, Validation, Test splits**
  - Training subset for small models: 40000 pairs
  - Training subset for bigger models: 300000 pairs
  - Provided validation and test splits: 3000 pairs
  - In some cases, 1% subset.

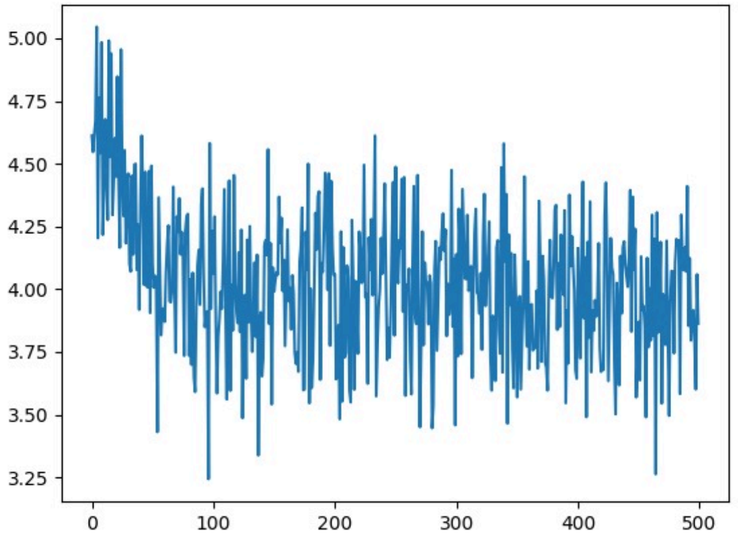


Challenges

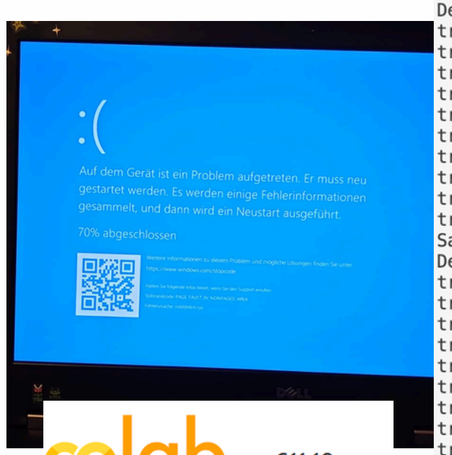
Challenges




```
(.venv) szholubak@Sofiias-MacBook-Pro training % python evaluate.py
[1, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15]
Input: Hi, how are you?
Output: , , , , , , , , , , ,
```



-



colab -€11.10

-€11.10

```
training loss: 9.297120094299316, validation Loss
training loss: 9.259479522705078, validation Loss
training loss: 9.354104995727539, validation Loss
training loss: 9.382797241210938, validation Loss
Saving checkpoint to accelerator_checl
Deleting 1 checkpoints to make room f
training loss: 9.413219451904297, val
training loss: 9.36528205871582, vali
training loss: 9.216065406799316, val
training loss: 9.23721981048584, vali
training loss: 9.325044631958008, val
training loss: 9.321989059448242, val
training loss: 9.253366470336914, val
training loss: 9.340381622314453, val
training loss: 9.518147468566895, val
training loss: 9.231877326965332, val
Saving checkpoint to accelerator_checl
Deleting 1 checkpoints to make room f
training loss: 9.218393325805664, val
training loss: 9.316357612609863, validation Loss
training loss: 9.345953941345215, validation Loss
training loss: 9.382152557373047, validation Loss
training loss: 9.381168365478516, validation Loss
training loss: 9.27957820892334, validation Loss
training loss: 9.369402885437012, validation Loss
training loss: 9.268611907958984, validation Loss
training loss: 9.29889965057373, validation Loss
training loss: 9.325664520263672, validation Loss
Saving checkpoint to accelerator_checkpoints_7...
Deleting 1 checkpoints to make room for new checkpoint.
training loss: 9.27908992767334, validation Loss
training loss: 9.27347183227539, validation Loss
training loss: 9.26522445678711, validation Loss
training loss: 9.353264808654785, validation Loss
training loss: 9.215087890625, validation Loss
training loss: 9.318567276000977, validation Loss
training loss: 9.330349922180176, validation Loss
training loss: 9.39588737487793, validation Loss
training loss: 9.419000625610352, validation Loss
training loss: 9.259446144104004, validation Loss
Saving checkpoint to accelerator_checkpoints_7...
Deleting 1 checkpoints to make room for new checkpoint.
training loss: 9.205376625061035, validation Loss
training loss: 9.428646087646484, validation Loss
training loss: 9.434494018554688, validation Loss
training loss: 9.372218132019043, validation Loss
training loss: 9.335967063903809, validation Loss
training loss: 9.325461387634277, validation Loss
training loss: 9.290207862854004, validation Loss
training loss: 9.37196159362793, validation Loss
training loss: 9.231799125671387, validation Loss
training loss: 9.188459396362305, validation Loss
Saving checkpoint to accelerator_checkpoints_7...
Deleting 1 checkpoints to make room for new checkp
```

BLEU=0.00

Python 3 Google Compute Engine backend (GPU)
Showing resources from 4:02 AM to 5:07 AM

System RAM	GPU RAM	Disk
4.0 / 83.5 GB	6.3 / 40.0 GB	235.7 / 235.7 GB

rpuiggari/**bert2**

20 Contributors 0 Issues 0 Stars 0 Forks

GPT-
SMALL

DECODER

• • •

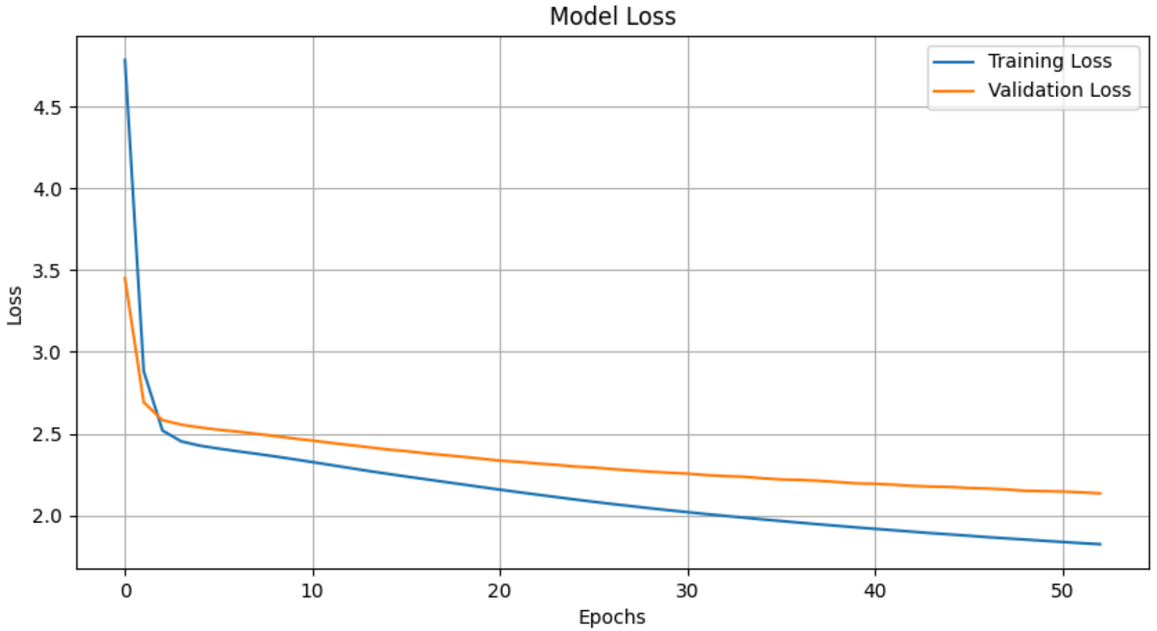
1 DECODER

English-to-German

Training (Config 1)

Total training time: 4.5 hours

- Batch Size: 32
- Allowed Sequence Length: 512
- Vocabulary Size: 2,000
- Model Dimension (D_MODEL): 128
- Feedforward Dimension (D_FF): 512
- Number of Layers: 4
- Number of Attention Heads: 2
- Epochs: 53
- Learning Rate: 1e-4
- Beta 1: 0.9
- Beta 2: 0.98
- Epsilon (EPS): 1e-9
- Label Smoothing (EPS_LS): 0.1
- Number of Samples: 45 000

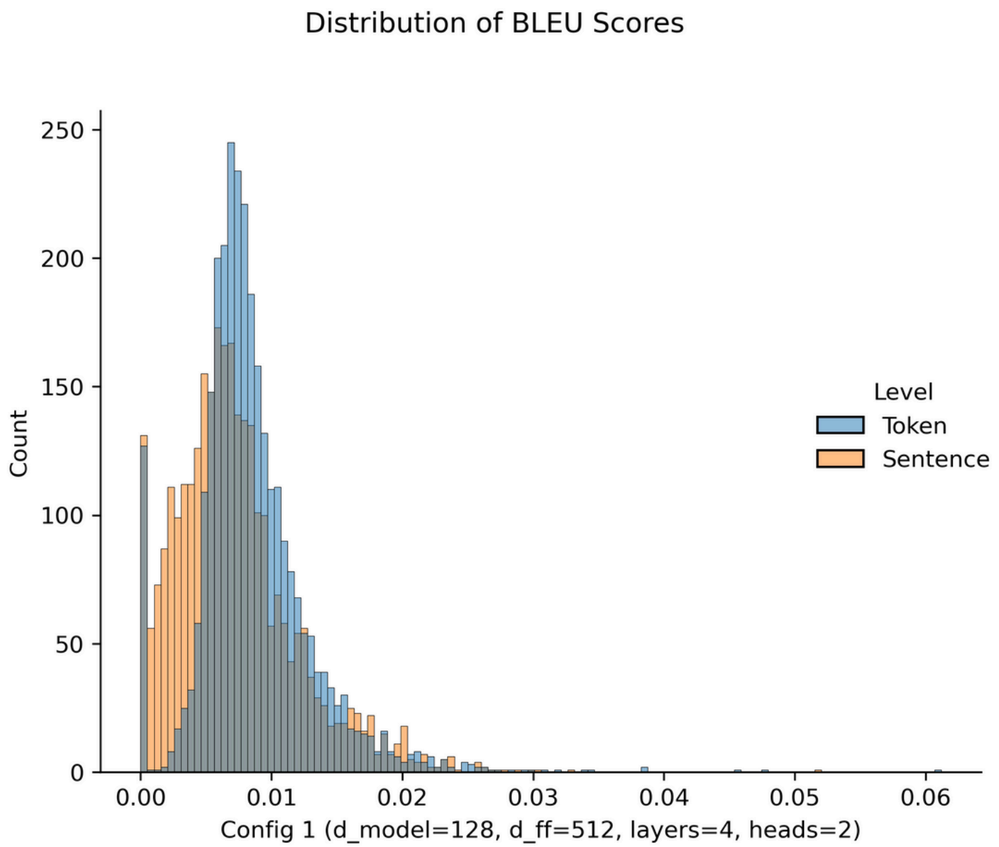


Results (Config 1)

- Final training loss: 1.8228
- Final validation loss: 2.1341

- Test loss: 1.6067

- Avg. token level BLEU score: 0.0086
- Avg. sentence level BLEU score: 0.0072

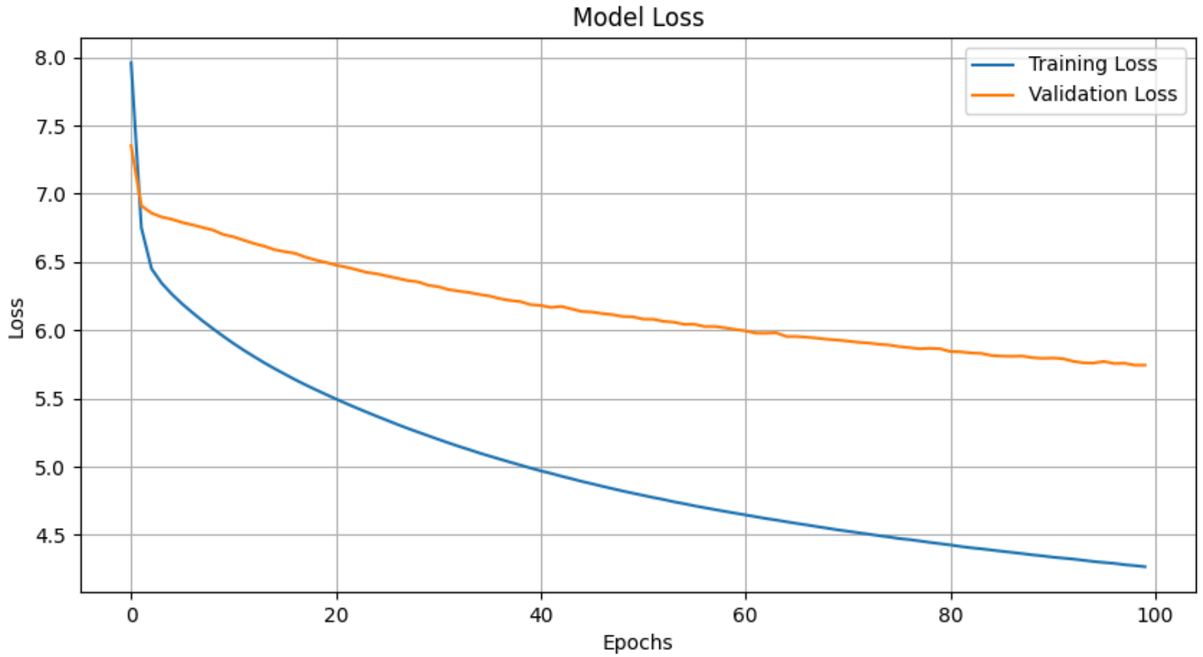


Problem! Example output: w i r w e i t e r e n s i c h t d e r e u r o p ä i s c h e n t e n , d a ß d i e s e r e i n e n t s c h e i d e n e n t s c h l i e ß u n g s a n t r a g e n , d a ß d i e s e r e n t s c h l i e ß u n g s a f t z e n , d a ß d i e s e n t s c h l i e ß u n g d e r e n t e r s t e l l t e n , d a ß d i e s e s e n t s c h l i e ß u n g s a f t z u s t e l l e n , d a ß d i e s e s e n t w i r d .

Training (Config 2)

Total training time: 4.5 hours

- Batch Size: 32
- Allowed Sequence Length: 1000
- Vocabulary Size: 10,000
- Model Dimension (D_MODEL): 128
- Feedforward Dimension (D_FF): 512
- Number of Layers: 4
- Number of Attention Heads: 2
- Epochs: 100
- Learning Rate: 1e-4
- Beta 1: 0.9
- Beta 2: 0.98
- Epsilon (EPS): 1e-9
- Label Smoothing (EPS_LS): 0.1
- Number of Samples: 45 000

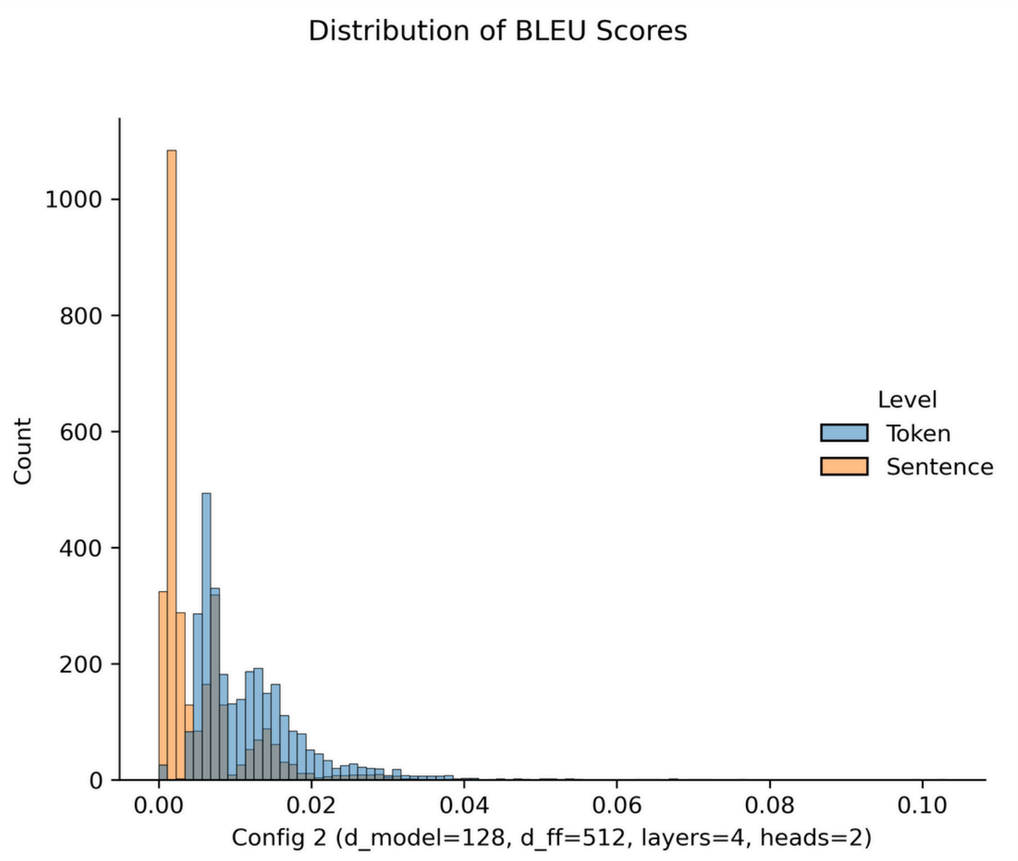


Results (Config 2)

- Final training loss: 4.2679
- Final validation loss: 5.7450

- Test loss: 7.3119

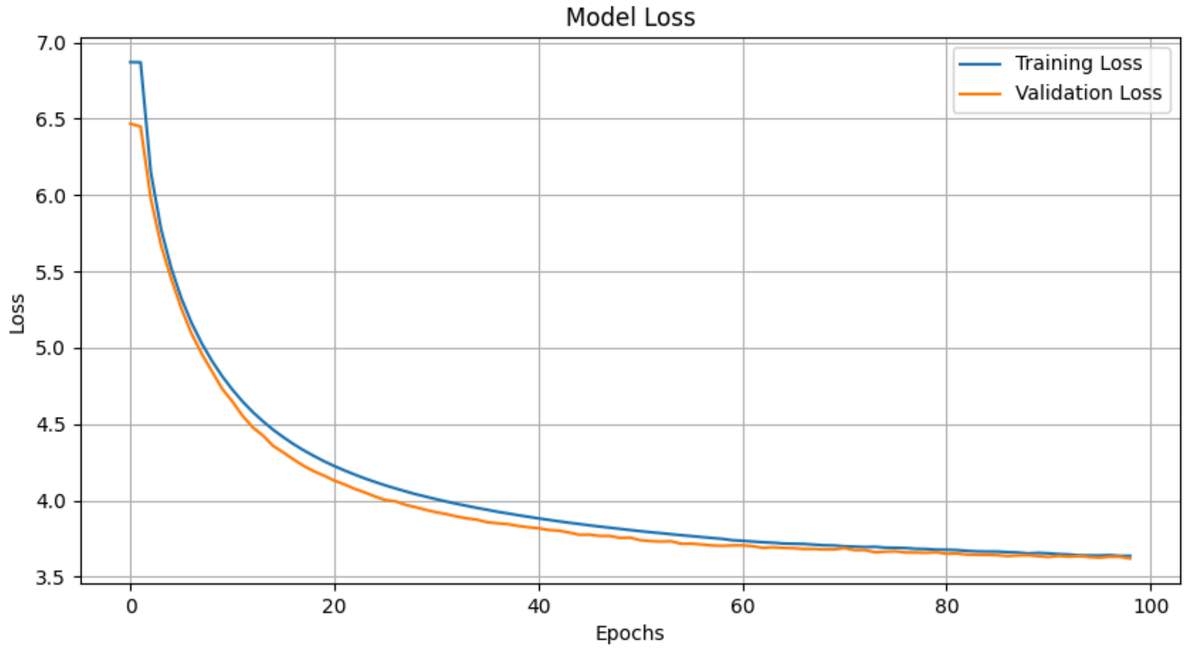
- Avg. token level BLEU score: 0.0119
- Avg. sentence level BLEU score: 0.0055



Training (Config 3)

Total training time: 8 hours

- Batch Size: 128
- Allowed Sequence Length: 150
- Vocabulary Size: 15,000
- Model Dimension (D_MODEL): 256
- Feedforward Dimension (D_FF): 1024
- Number of Layers: 4
- Number of Attention Heads: 4
- Epochs: 100
- Learning Rate: 1e-4
- Beta 1: 0.9
- Beta 2: 0.98
- Epsilon (EPS): 1e-9
- Label Smoothing (EPS_LS): 0.1
- Number of Samples: 300 000

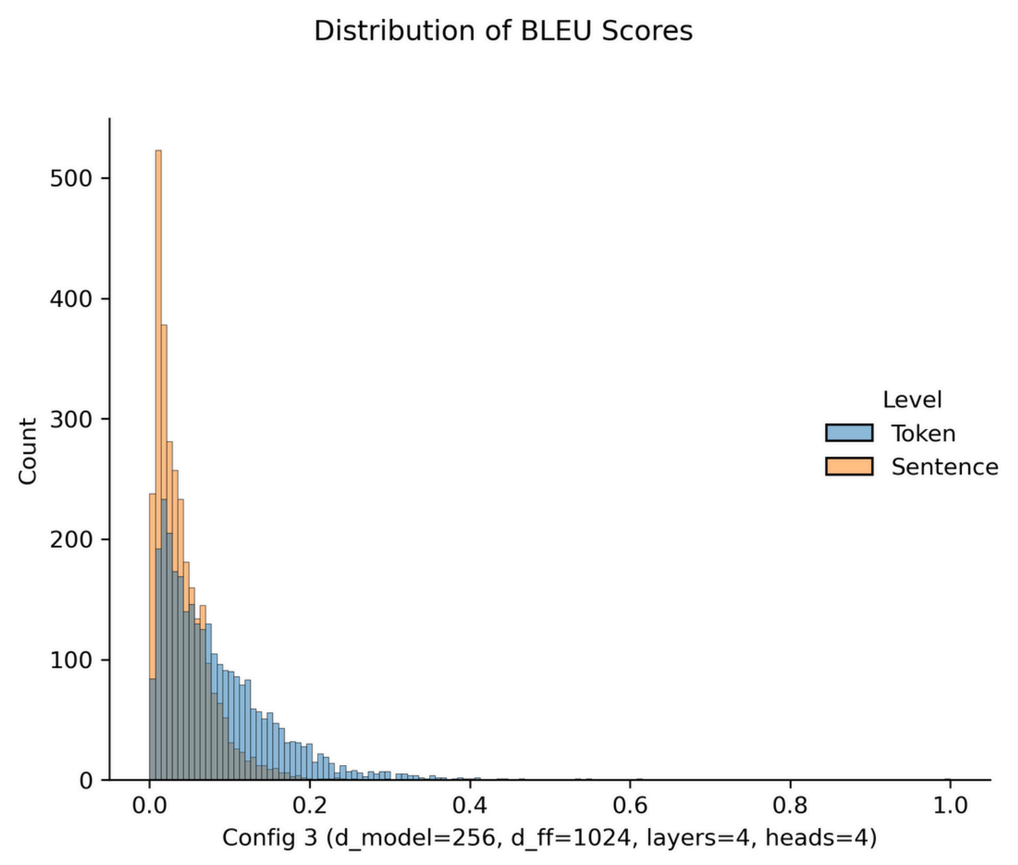


Results (Config 3)

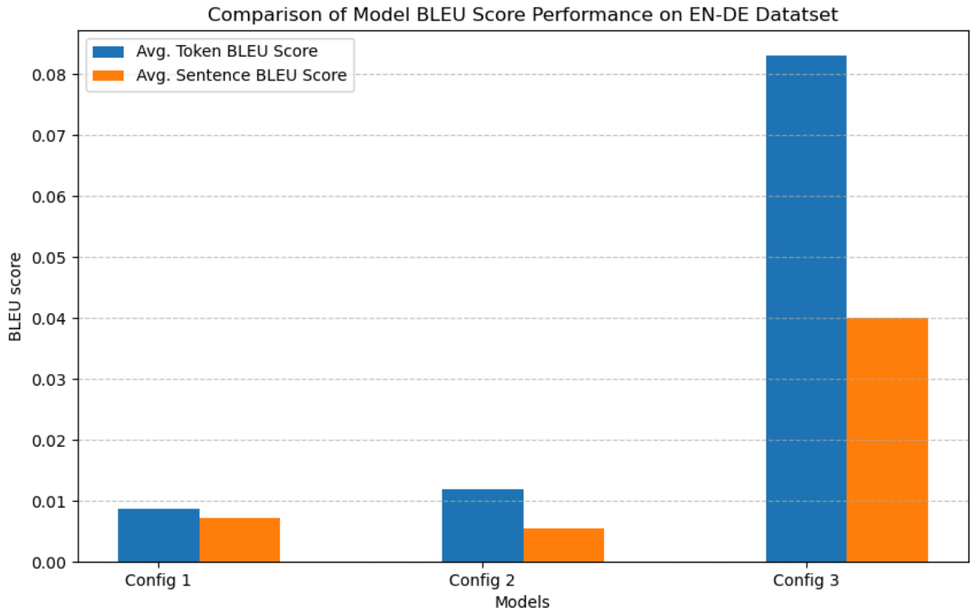
- Final training loss average: 3.6366
- Final validation loss average: 3.6197

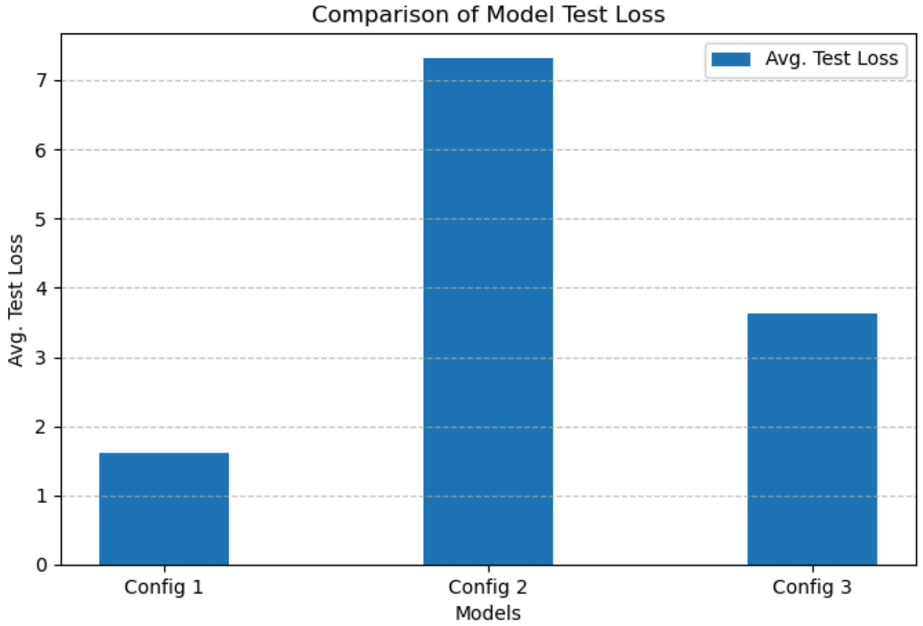
- Avg. test loss: 3.6226

- Avg. token level BLEU score: 0.0831
- Avg. sentence level BLEU score: 0.0400



Model comparison



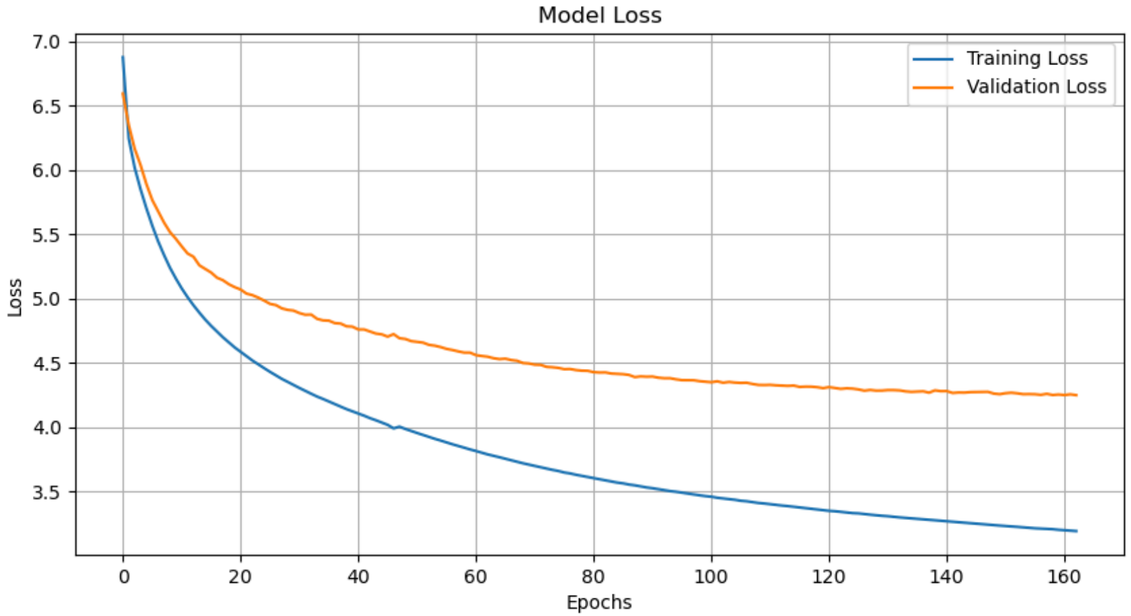


English-to-French

Training (Config 1)

Total training time: 4 hours

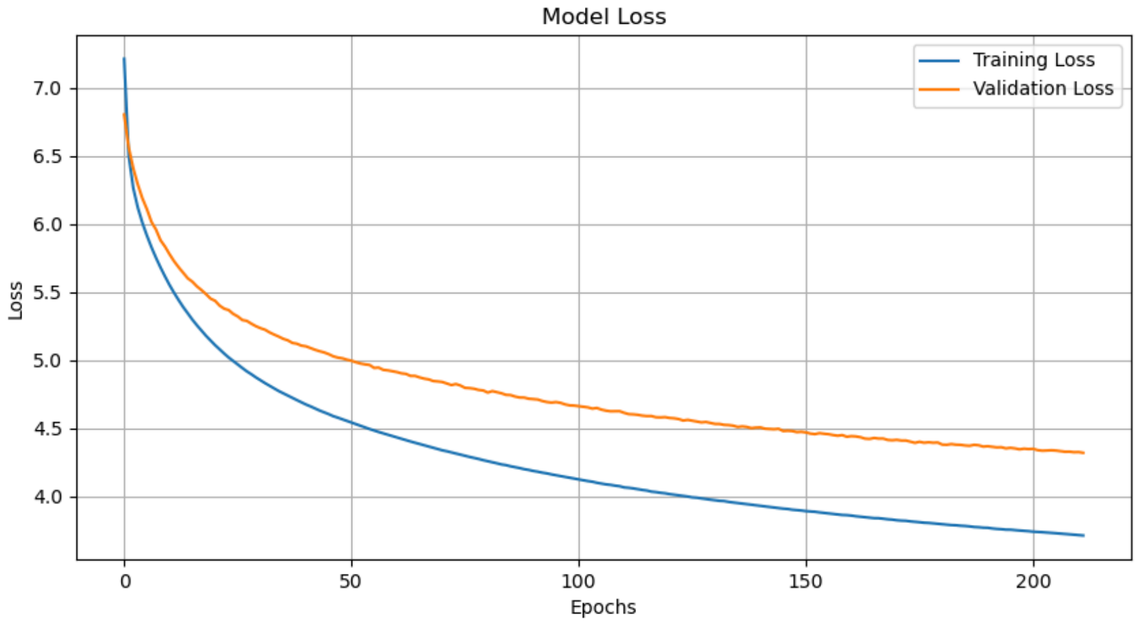
- Batch Size: 64
- Allowed Sequence Length: 500
- Vocabulary Size: 10,000
- Model Dimension (D_MODEL): 256
- Feedforward Dimension (D_FF): 1,048
- Number of Layers: 2
- Number of Attention Heads: 4
- Epochs: 246
- Learning Rate: 1e-4
- Beta 1: 0.9
- Beta 2: 0.98
- Epsilon (EPS): 1e-9
- Label Smoothing (EPS_LS): 0.1
- Number of Samples: 40,000



Training (Config 2)

Total training time: 7.5 hours

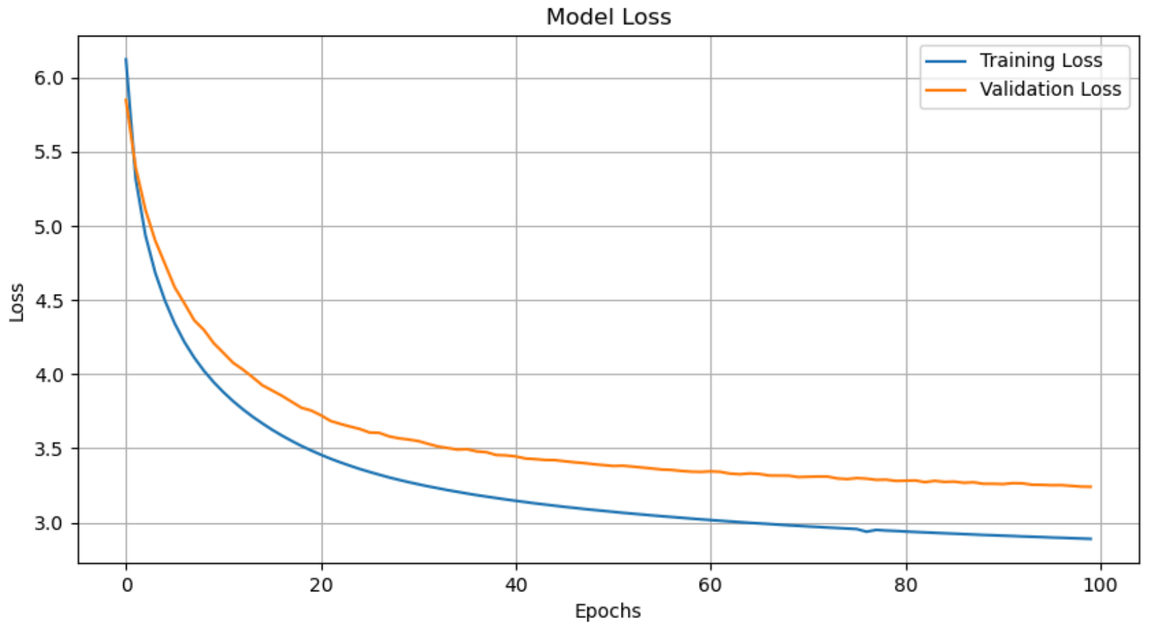
- Batch Size: 64
- Allowed Sequence Length: 500
- Vocabulary Size: 10,000
- Model Dimension (D_MODEL): 128
- Feedforward Dimension (D_FF): 512
- Number of Layers: 4
- Number of Attention Heads: 2
- Epochs: 246
- Learning Rate: 1e-4
- Beta 1: 0.9
- Beta 2: 0.98
- Epsilon (EPS): 1e-9
- Label Smoothing (EPS_LS): 0.1
- Number of Samples: 40,000



Training (Config 3)

Total training time: 14.5 hours

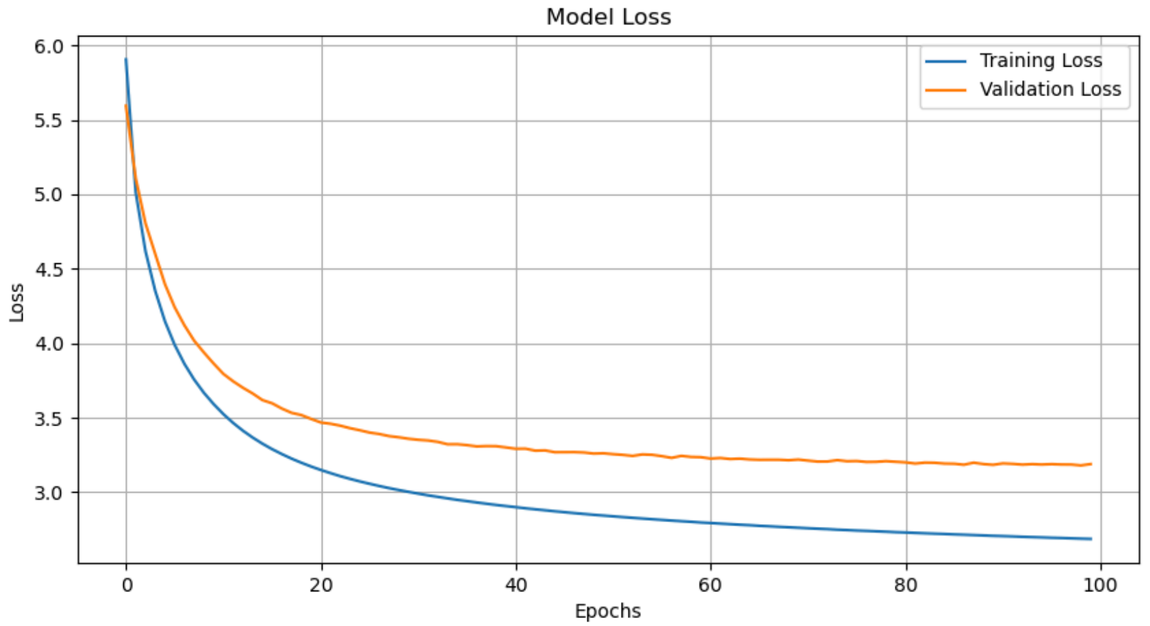
- Batch Size: 128
- Allowed Sequence Length: 150
- Vocabulary Size: 15,000
- Model Dimension (D_MODEL): 256
- Feedforward Dimension (D_FF): 1024
- Number of Layers: 4
- Number of Attention Heads: 4
- Epochs: 100
- Learning Rate: 1e-4
- Beta 1: 0.9
- Beta 2: 0.98
- Epsilon (EPS): 1e-9
- Label Smoothing (EPS_LS): 0.1
- Number of Samples: 300,000



Training (Config 4)

Total training time: 14 hours

- Batch Size: 128
- Allowed Sequence Length: 150
- Vocabulary Size: 15,000
- Model Dimension (D_MODEL): 384
- Feedforward Dimension (D_FF): 1536
- Number of Layers: 3
- Number of Attention Heads: 6
- Epochs: 100
- Learning Rate: 1e-4
- Beta 1: 0.9
- Beta 2: 0.98
- Epsilon (EPS): 1e-9
- Label Smoothing (EPS_LS): 0.1
- Number of Samples: 300,000

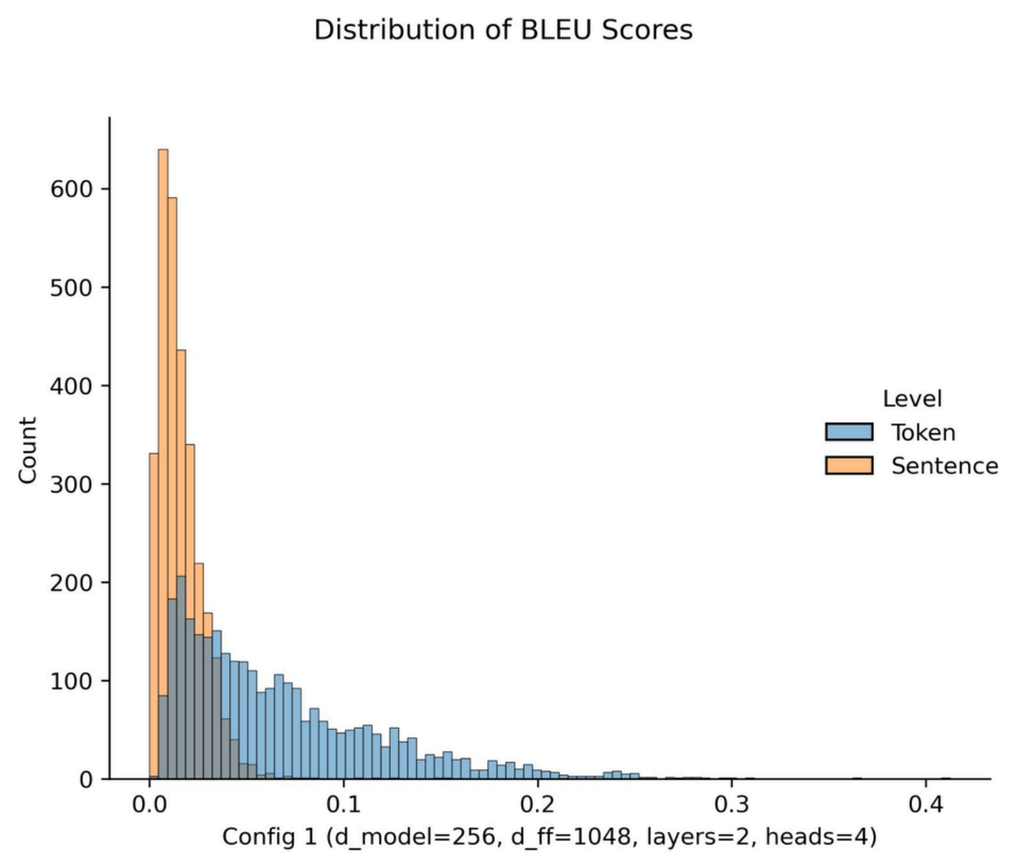


Results (Config 1)

- Final training loss: 3.1923
- Final validation loss: 4.2486

- Test loss: 4.1159

- Avg. token level BLEU score: 0.0673
- Avg. sentence level BLEU score: 0.0161

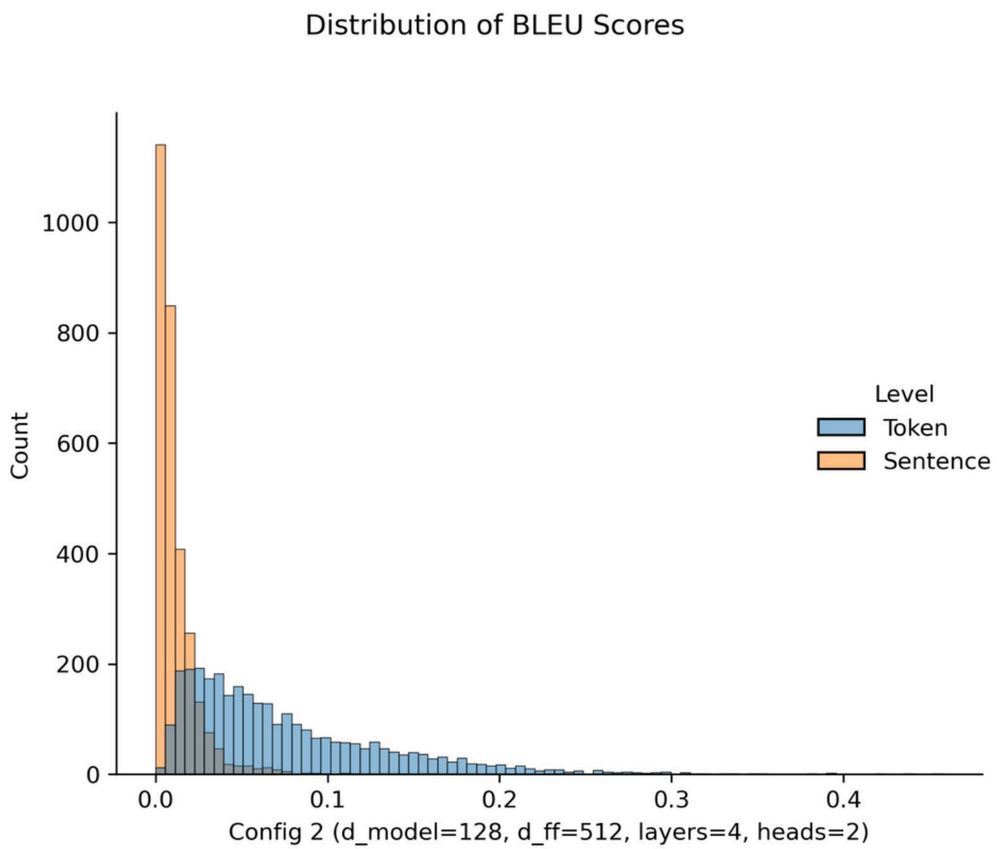


Results (Config 2)

- Final training loss: 3.7120
- Final validation loss: 4.3190

- Test loss: 4.2412

- Avg. token level BLEU score: 0.0748
- Avg. sentence level BLEU score: 0.0116

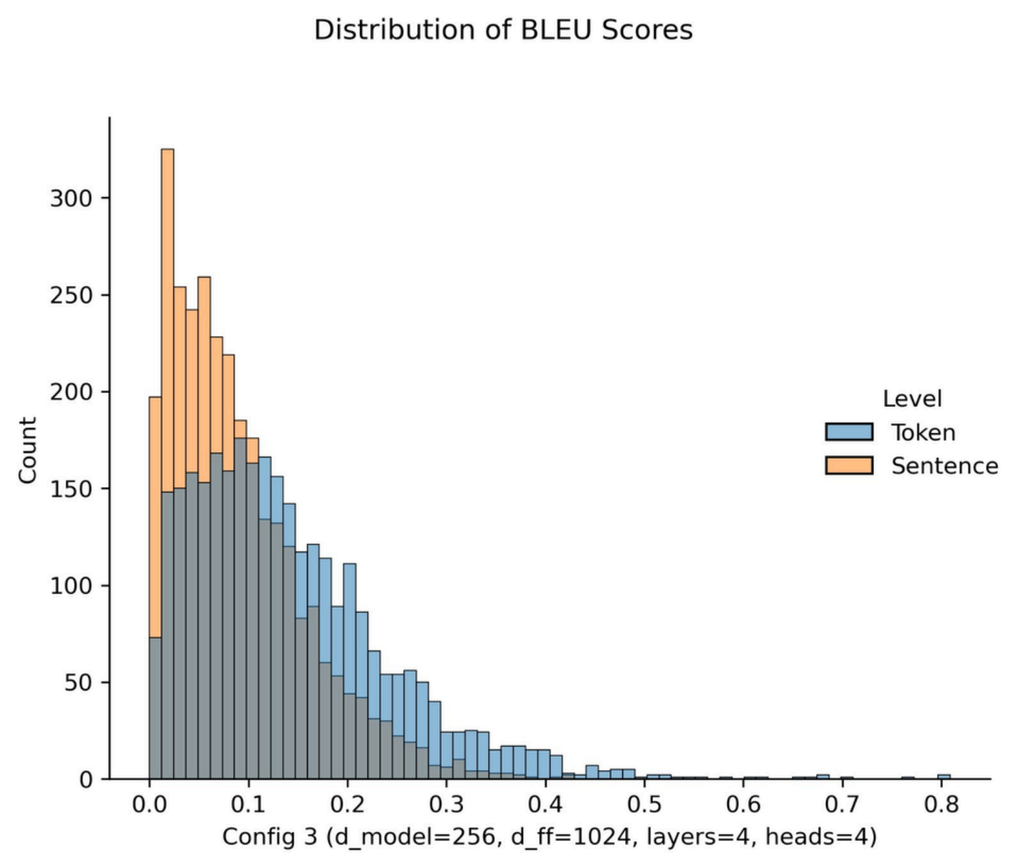


Results (Config 3)

- Final training loss: 2.890
- Final validation loss: 3.2412

- Test loss: 3.0509

- Avg. token level BLEU score: 0.1419
- Avg. sentence level BLEU score: 0.0887

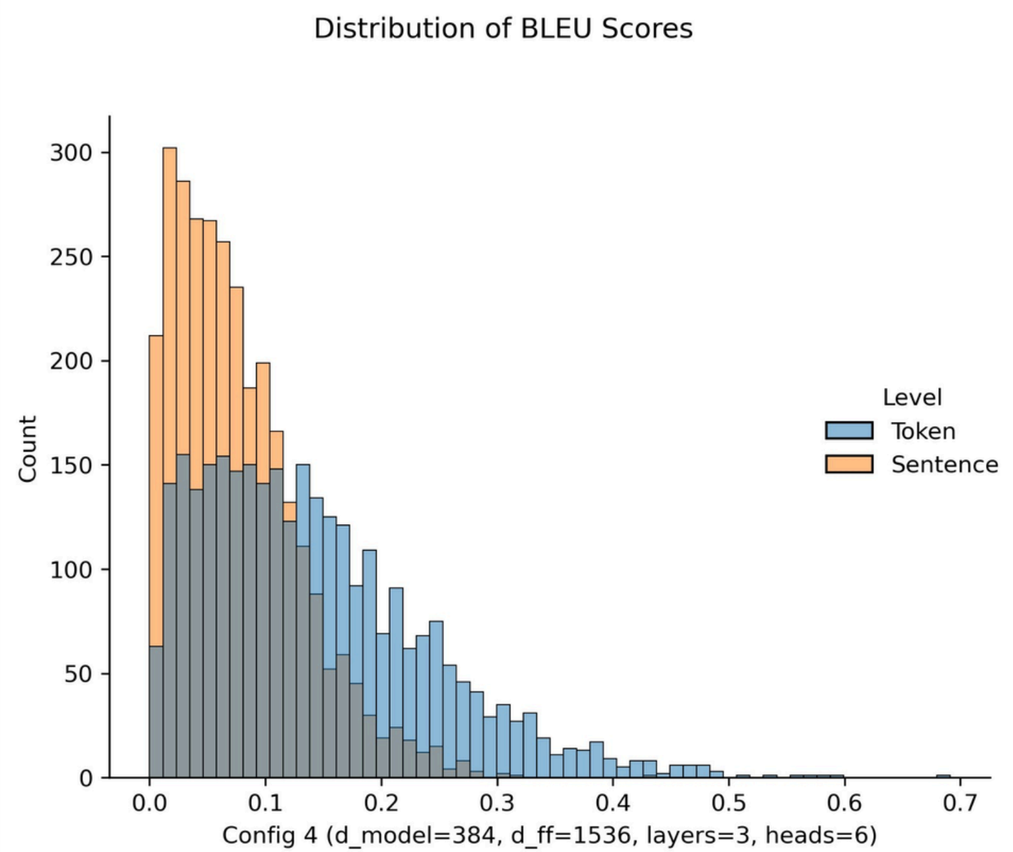


Results (Config 4)

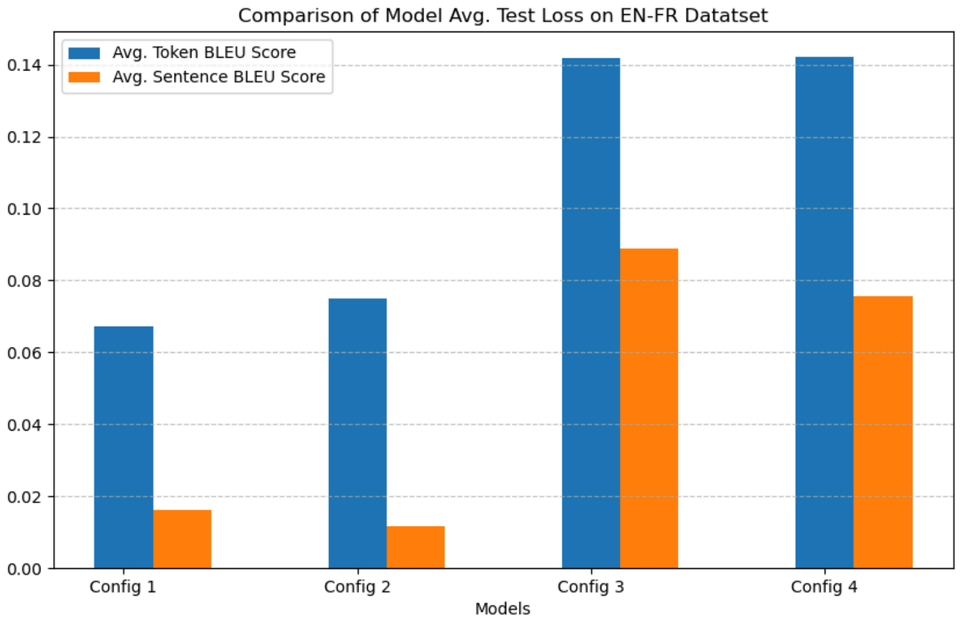
- Final training loss average: 2.6859
- Final validation loss average: 3.1885

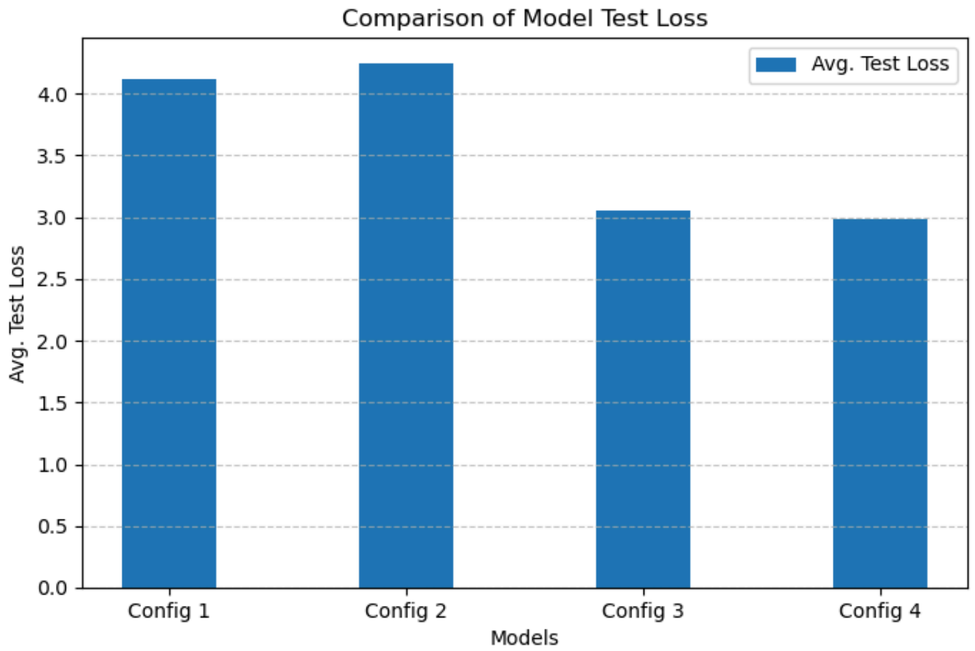
- Avg. test loss: 2.9796

- Avg. token level BLEU score: 0.14216
- Avg. sentence level BLEU score: 0.0755



Model comparison





Output selection (Config 3)

Input: The federal tax itself, 18.4 cents per gallon, hasn't gone up in 20 years.
Output: l ' impôt fédéral elle - même , 18 , 4 % de l ' ensemble des services fiscaux , n ' a pas été re produit en 20 ans .
Actual: La taxe fédérale elle-même, qui est de 18,4 cents par gallon, n'a pas augmenté depuis 20 ans.

Input: A feast for fans
Output: a fe ast for f ans a fe ast for f ans ans f ans ans a fe ast for f ans f ans ans f ans ans a fe ast for f ans f ans f ans a fe ast for f ans f ans f ans a fe
Actual: Un festin pour ses fans

Input: However, Atomica is not the only track to have been released.
Output: toutefois , l ' atom ique n ' est pas la seule voie à avoir été libé rée .
Actual: Mais Atomica n'est pas le seul titre à dévoiler ses charmes.

Input: People are paying more directly into what they are getting.
Output: les personnes pay ent plus directement dans ce qu ' elles ob tiennent .
Actual: Les gens paient plus directement pour les avantages qui leur sont procurés.

Output selection (Config 4)

Input: The federal tax itself, 18.4 cents per gallon, hasn't gone up in 20 years.
Output: l ' impôt fédéral lui - même , 18 , 4 % de l ' impôt fédéral , n ' a pas augmenté de 20 ans , mais n ' a pas augmenté de 20 ans .
Actual: La taxe fédérale elle-même, qui est de 18,4 cents par gallon, n'a pas augmenté depuis 20 ans.

Input: A feast for fans
Output: a fe ast for f ans f ans f ans f ans f ans (f ans) a fe ast for f ans f ans (f ans) s . a fe ast (f ans) s . a fe ast (f ans) s .
Actual: Un festin pour ses fans

Input: People are paying more directly into what they are getting.
Output: les gens reçoivent plus de services directement dans ce qu ' ils ob tiennent .
Actual: Les gens paient plus directement pour les avantages qui leur sont procurés.

Input: However, Atomica is not the only track to have been released.
Output: toutefois , l ' atom ica n ' est pas la seule voie à suivre ; il n ' est pas le seul à suivre ; il a été publié de telles choses .
Actual: Mais Atomica n'est pas le seul titre à dévoiler ses charmes.

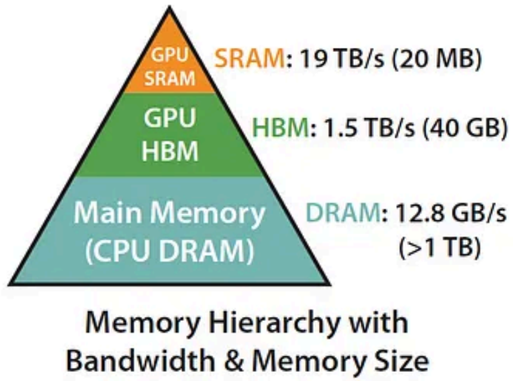
English-to-German

Flash Attention Transformer

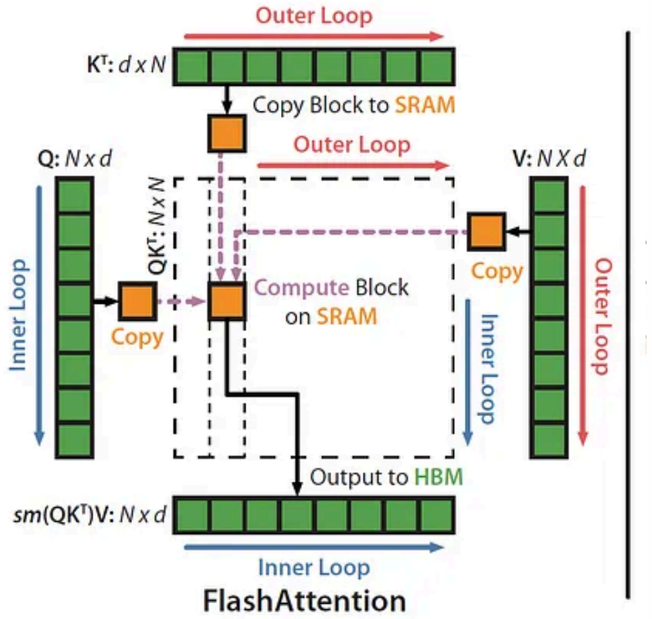
Flash Attention

FlashAttention: Fast and Memory-Efficient Exact Attention with IO-Awareness

- **Memory-efficient:** compared to vanilla attention, which is quadratic in sequence length, $O(n^2)$, this method is sub-quadratic/linear in N (O(n)).
- **Exact :** meaning it's not an approximation of the attention mechanism like e.g. sparse, or low-rank matrix approximation methods, its outputs are the same as in the "vanilla" attention mechanism.

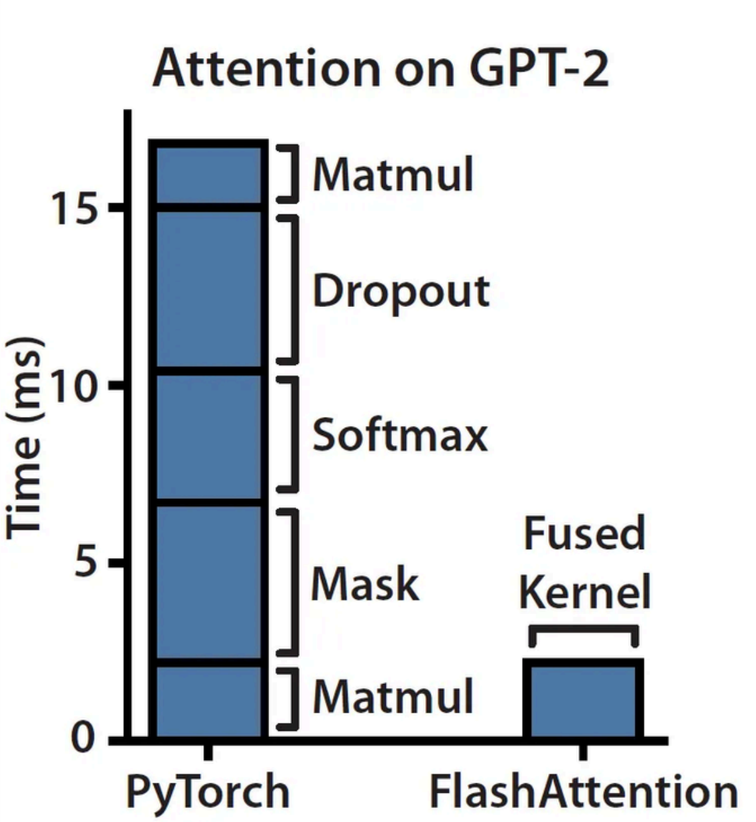


Memory Hierarchy with Bandwidth & Memory Size



FlashAttention

Flash Attention



Simplified: FlashAttention reduces memory usage and speeds up attention computation by splitting queries into blocks and computing attention in a tiled, GPU-efficient manner, avoiding unnecessary memory reads/writes.

Flash Attention

Architecture

- **Encoder: 6 layers of self-attention + feed-forward sublayers.**
  - Multi-Head Self-Attention using FlashAttention instead of standard scaled dot-product attention.
- **Decoder: 6 layers of masked self-attention, encoder–decoder cross-attention, and feed-forward sublayers.**
  - Masked Self-Attention with causal masking, also using FlashAttention.
  - Cross-Attention to encoder outputs.
- **Token Embeddings:** We learn an embedding for each token in the source and target vocabularies using Byte Pair Encoding.

Intended Improvements

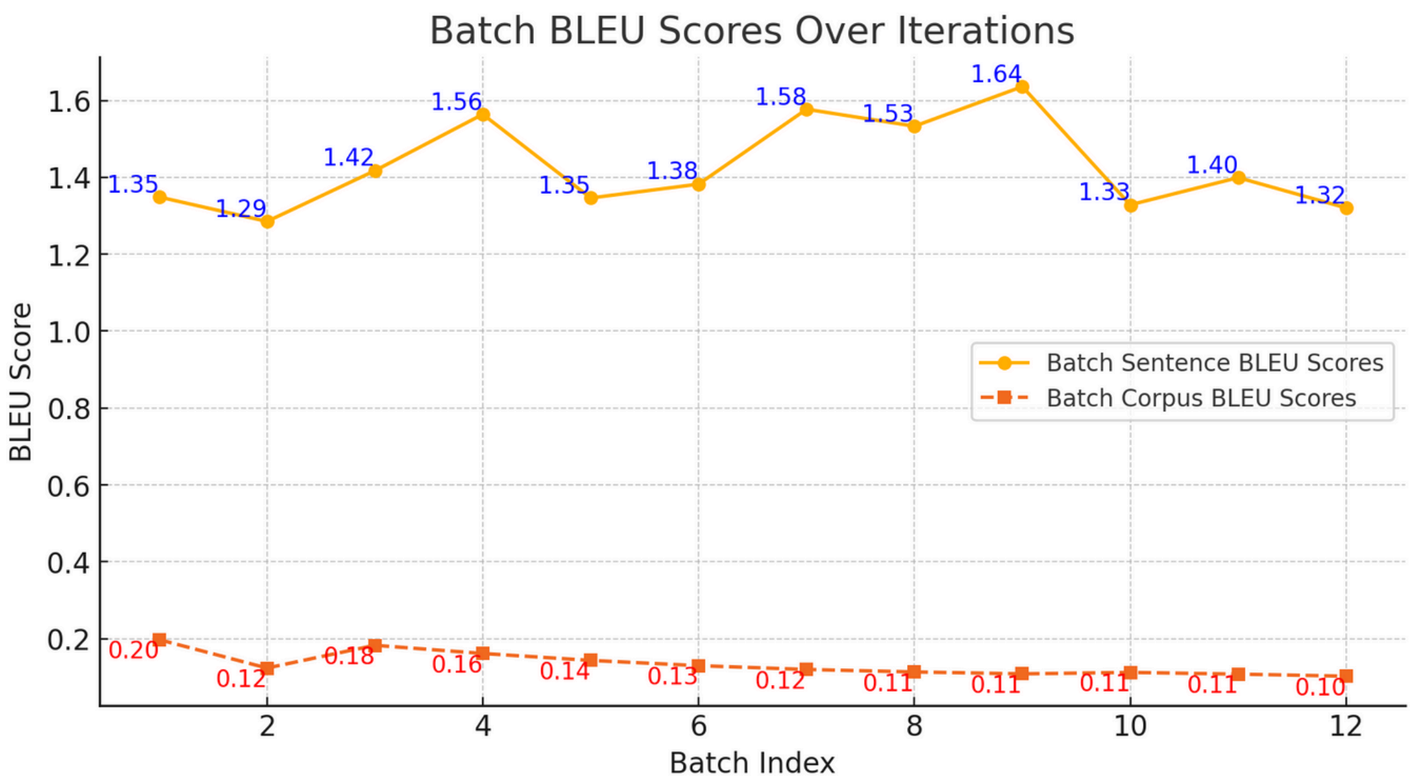
- **FlashAttention**
  - **Speed:** Leveraging specialized CUDA kernels, Flash Attention achieved up to 2× or 3× speedups for long sequences and large batch sizes in the original paper.
- **Mixed-Precision Training (AMP):** mixed precision for both training and inference, thus the matrix multiplications in attention and feed-forward layers run in half precision (FP16).
- **Flexible, Modular Repository: for future continued (Expensive) work.**

Flash Attention

Total training time: 7.12 hours

Final Corpus BLEU on 20% subset: 0.10
Average Sentence-Level BLEU (batch-averaged): 1.43

training:
  batch_size: 50
  learning_rate: 0.0001
  warmup_steps: 2000
  max_epochs: 200
  mixed_precision: true
  gradient_accumulation_steps: 2
  device: "cuda"



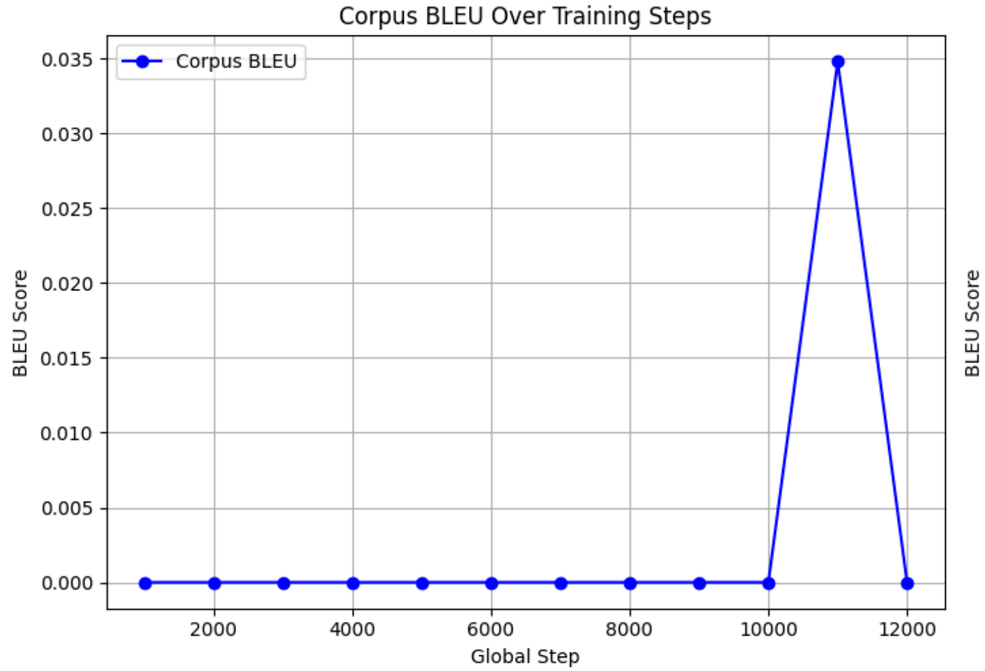
English-to-French

Hugging Face Transformer

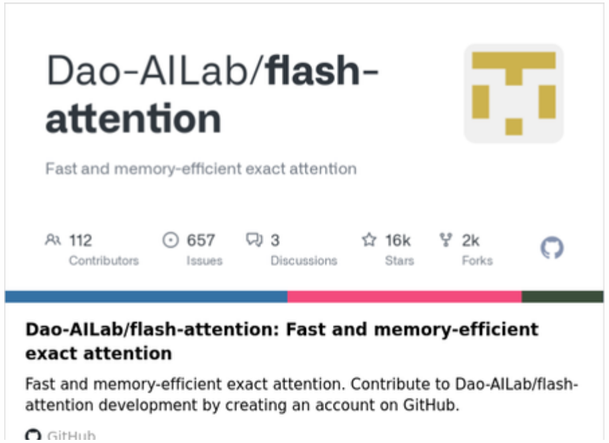
HuggingFace

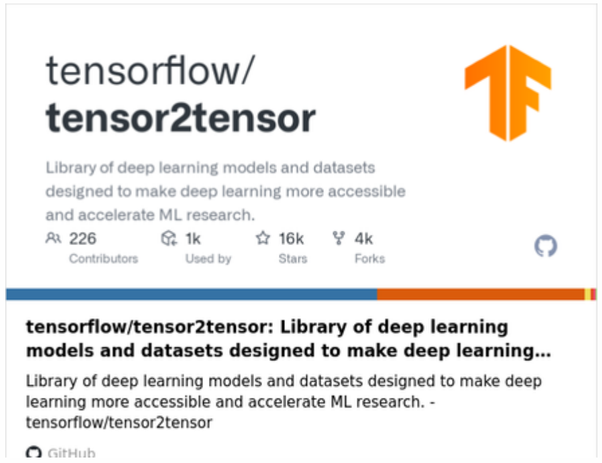
The idea was to use BERT as the encoder and GPT-2 as the decoder with the hopes that they would produce fairly good results on a limited dataset and less training.

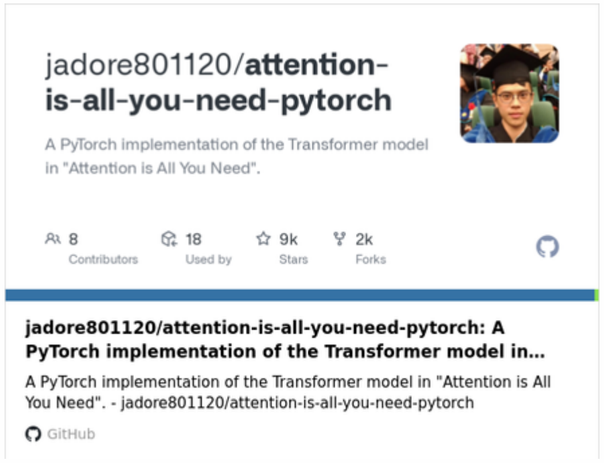
- **Bert**
  - bidirectional attention instead of vanilla self-attention.
  - masked language model.
- **GPT-2**
  - original Transformer used absolute positional encodings while gpt2 introduces **learned positional embeddings**

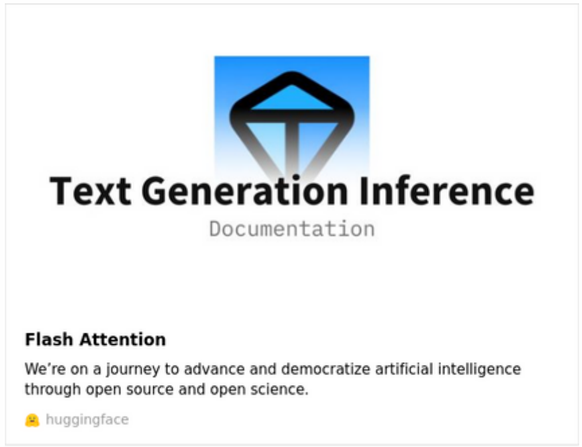


References









Attention is all you need. 2017.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, Illia Polosukhin.