Transformers From Scratch

Objective

1) To build a Transformer model from scratch with all the components of the original transformer architecture.
2) To implement two types of positional encoding strategies.
   (a) Rotary positional embeddings (RoPE).
   (b) Relative position bias (additive bias to attention scores).
3) To implement and evaluate the following decoding algorithms for inference:
   (i) Greedy decoding – always select the token with the highest probability at each step.
   (ii) Beam search – maintain the top-B sequences (beams) at each step and choose the sequence (or "beam") with the highest overall score.
   (iii) Top-k sampling – sample the next token from the top-k most probable tokens.
4) To evaluate all configurations on the test set and report their final BLEU scores in a table. Analyze the impact of different decoding strategies on translation accuracy and support your observations.
5) To plot the training loss against epochs for both positional encoding methods on the same graph. Indicate which configuration converges faster and support your observations.

Experimental Setup

For each experiment, we use the following configuration:

Hyperparameter	Value
d_model	128
num_encoder	3
num_decoder	3
num_heads	4
seq_len	500

feedforward network hidden_size	256
dropout	0.1
batch_size	32
epochs	9
alpha	0.05
pe_method	Sinusoidal / Relative Bias / RoPE
decode_method	Greedy / Beam / Top_k
Beam Size	3
K (Top K Sampling)	5
Tokenizer Type	BPE / Word Level

Results

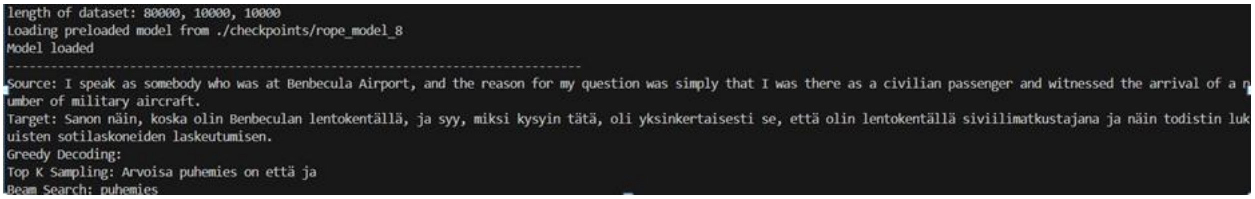
1) *Evaluating all configurations on the test set and reporting their final BLEU scores in a table. Additionally, analyzing the impact of different decoding strategies on translation accuracies.*

The below table summarizes the BLUE score obtained for all possible configuration pairs on the test data.

Configuration	Relative Position Bias	RoPE
Greedy Decoding	0	0
Beam Search	0	0
Top-k sampling	8.36e-236	3.17e-232

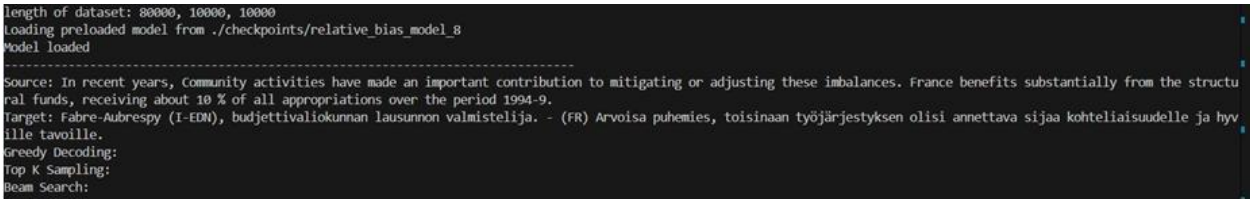
The results obtained for translation are very less for all decoding strategies and positional embeddings. One reason for this is that the model is too small to capture the translation patterns, and we run it only for 9 epochs. In addition to this, methods such as Greedy decode prefer continuously predicting tokens such as [UNK] or punctuations. This leads to a 0 BLEU score.

The below results are sample examples generated by the model when trained with RoPE for all decoding strategies. Top K sampling seems to work better than others because it just doesn't pick the word with the highest probability. However, even the BLEU score of Top K is trivial.



```
length of dataset: 80000, 10000, 10000
Loading preloaded model from ./checkpoints/rope_model_8
Model loaded
--------------------------------------------------------------------------------
Source: I speak as somebody who was at Benbecula Airport, and the reason for my question was simply that I was there as a civilian passenger and witnessed the arrival of a n
umber of military aircraft.
Target: Sanon näin, koska olin Benbeculan lentokentällä, ja syy, miksi kysyin tätä, oli yksinkertaisesti se, että olin lentokentällä siviilimatkustajana ja näin todistin luk
uisten sotilaskoneiden laskeutumisen.
Greedy Decoding:
Top K Sampling: Arvoisa puhemies on että ja
Beam Search: puhemies
```

Below are the samples generated by Relative Positional Bias for the same configuration.



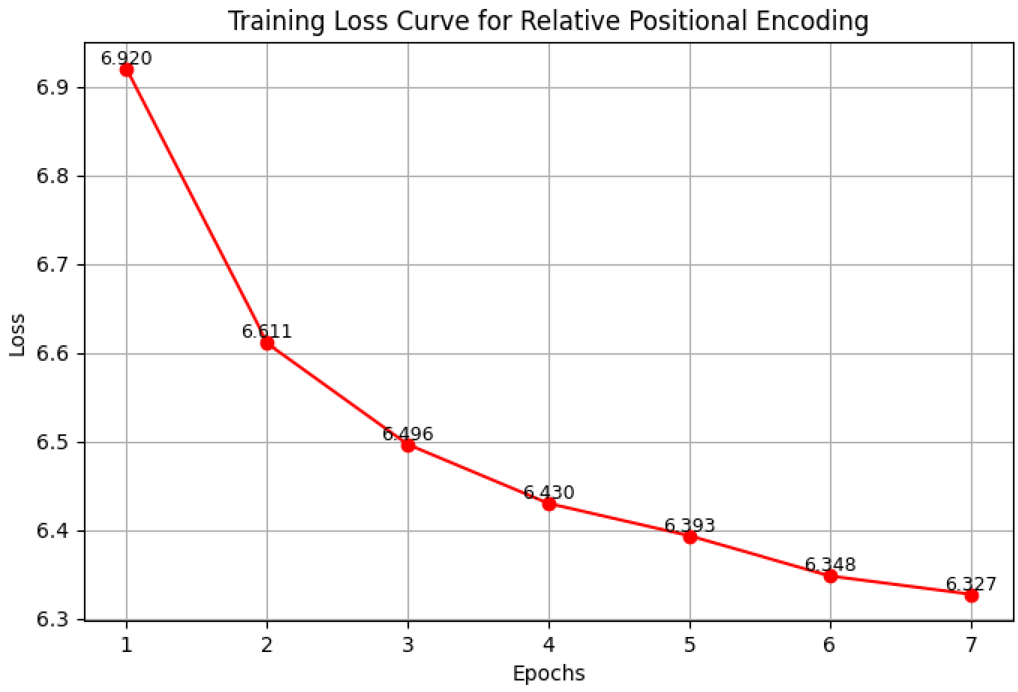
```
length of dataset: 80000, 10000, 10000
Loading preloaded model from ./checkpoints/relative_bias_model_8
Model loaded
--------------------------------------------------------------------------------
Source: In recent years, Community activities have made an important contribution to mitigating or adjusting these imbalances. France benefits substantially from the structu
ral funds, receiving about 10 % of all appropriations over the period 1994-9.
Target: Fabre-Aubrespy (I-EDN), budjettivaliokunnan lausunnon valmistelija. - (FR) Arvoisa puhemies, toisinaan työjärjestyksen olisi annettava sijaa kohteliaisuudelle ja hyv
ille tavoille.
Greedy Decoding:
Top K Sampling:
Beam Search:
```

Greedy Decoding produces the lowest BLEU scores due to its nature to choose the highest tokens. Top-K sampling performs better than all methods. Beam search performs slightly better than greedy as per the results, but still it performs worse compared to Top-K. In order to improve it, we tried using word level tokenizer and BPE. We also tried filtering the punctuations to force greedy methods to provide some output. However, neither of this resulted in improved scores.

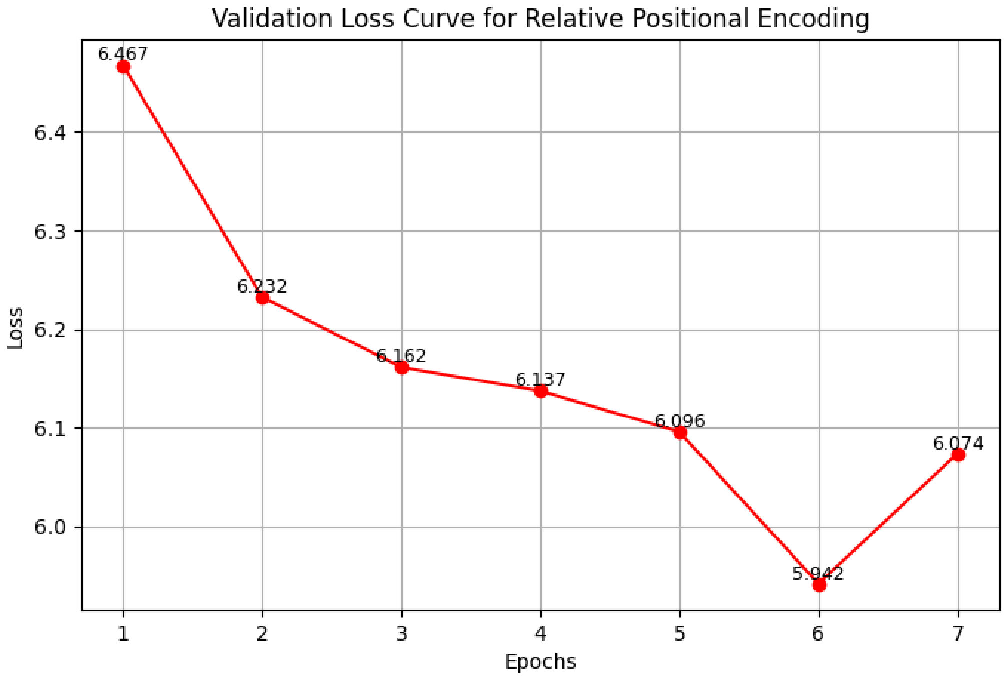
Overall, the best performance is observed with the RoPE + Top-K combination, achieving a BLEU score of 3.17e-232

2) Plotting the training loss against epochs for both positional encoding methods on the same graph and indicating which configuration converges faster.

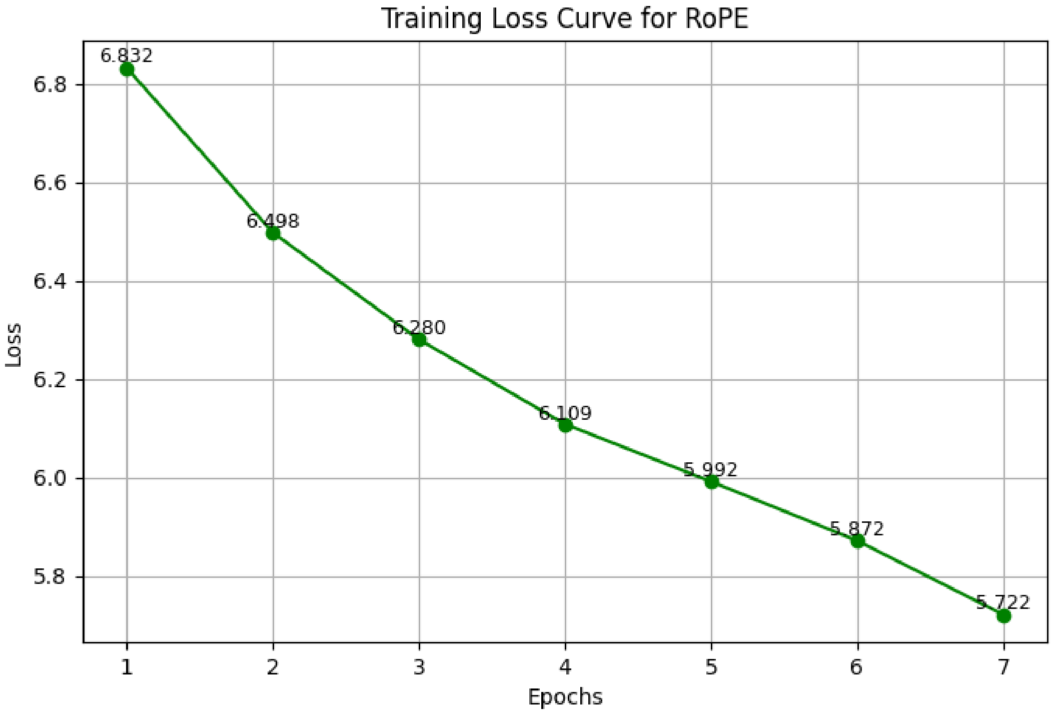
2.1) Train loss curve for Relative Positional Encoding



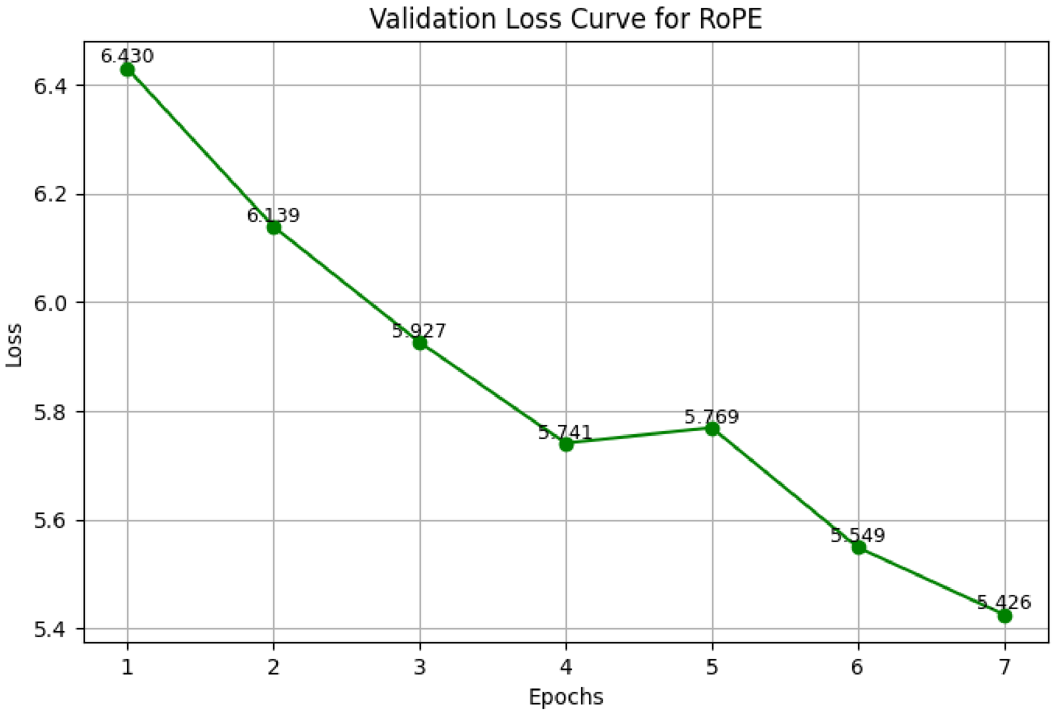
2.2) Validation loss curve for Relative Positional Encoding



2.3) Train loss curve for RoPE



2.4) Validation loss curve for RoPE



The training loss curves show that RoPE converges faster than Relative Position Bias. For instance, if we check the 4th epoch for both cases, RoPE gets a training loss of 6.109, while Relative Position Bias gets a score of 6.430. Thus, RoPE converges faster than Relative Position Bias.

In addition to this, the validation loss curve for RoPE is better than Relative Position Embedding.

Conclusion

Hence, implemented transformer architecture from scratch and explored different types of positional encodings (Relative Position Bias and RoPE) and decoding strategies (Greedy Decoding, Beam Search and Top K sampling). Additionally, experimented with various configurations of positional encodings and decoding strategies.

References

1) Blog explaining RoPE: https://mfaizan.github.io/2023/04/02/sines.html
2) Blog on Decoding Strategies: https://huggingface.co/blog/mlabonne/decoding-strategies
3) Video by Umar Jamil on creating vanilla transformer architecture from scratch: https://www.youtube.com/watch?v=ISNdQcPhsts