CSE 5525: HW-2 Report

By Krish Patel

Abstract

This document contains the discoveries and findings from Homework 2 of CSE 5525, which involved implementing a transformer encoder and training a model to predict future tokens.

1 Part-1

For Part 1 of the assignment, a transformer model was implemented from scratch, according to the instructions provided in the assignment PDF. The transformer was used to predict how many times a token has appeared in a text, classified into 0, 1, or 2 for 2 or more occurrences. So, it is a classification task. The details of implementations can be found in the code. Based on experimentation the best learning rate was 0.001. The standard attention mechanism ended up with an accuracy of 0.9935 on the test data set, achieving decent results.

For the explorations, I did both the exploration from this section. For the first exploration, I explored how the attention graph was generated by the standard attention mechanism. An interesting pattern emerges along the diagonal band, showing that the model strongly attends to the current character and its close neighbors. This is to be expected as neighboring characters are most informative for predicting labels. The model also attends to repeated characters earlier in the sequence, suggesting the capability of this model to observe long-range dependencies.

For the second part of the exploration, I implemented a different kind of attention than the standard self-attention in the transformer. Here, I have compared standard self-attention with ALiBi(Attention with Linear Biases). Based on performance, the standard self-attention achieved an almost perfect accuracy of 0.9935, with the attention map showing clear diagonals. On the other hand, ALiBi achieved a low accuracy of 0.698. The attention maps for ALiBi has a smoother diagonal patter with attention fades as the distance increases. Hence, it loses its ability to create a strong link between distant characters and hence achieves lower accuracy during testing. Overall, this suggests that ALiBi is beneficial when extrapolation for longer unseen sequences is required, but fails for absolute positional precision heavy tasks such as this one.

2 Part-2

For Part 2, I implemented a character-level Transformer language model that uses the transformer mechanism at its core, including embeddings, positional encodings, multiple encoder layers, and a linear output with log-softmax, trained the model with NLLoss and Adam on a fixed-length character window. The default base uniform model has a perplexity of about 27, basically random guessing each character. The trained transformer language model achieved a perplexity of about 5.5. It shows that the trained model can successfully capture the character-level patterns and relations in text. It suggests that a small Transformer can model text structures effectively, which is why it is such a widely used model.

3 Conclusion

To conclude, the standard attention with positional encoding was highly effective for the Part-1 classification task, achieving almost perfect accuracy and interpretable results. Whereas ALiBi biased the model towards recent tokens, which reduced the accuracy but showcased why it is good at extrapolating long sequences. For Part-2, the Trafomer language model performed much better than the uniform base model, changing perplexity from 27 to 5.5. Hence, both tasks provided decent insights into the inner workings of transformers.

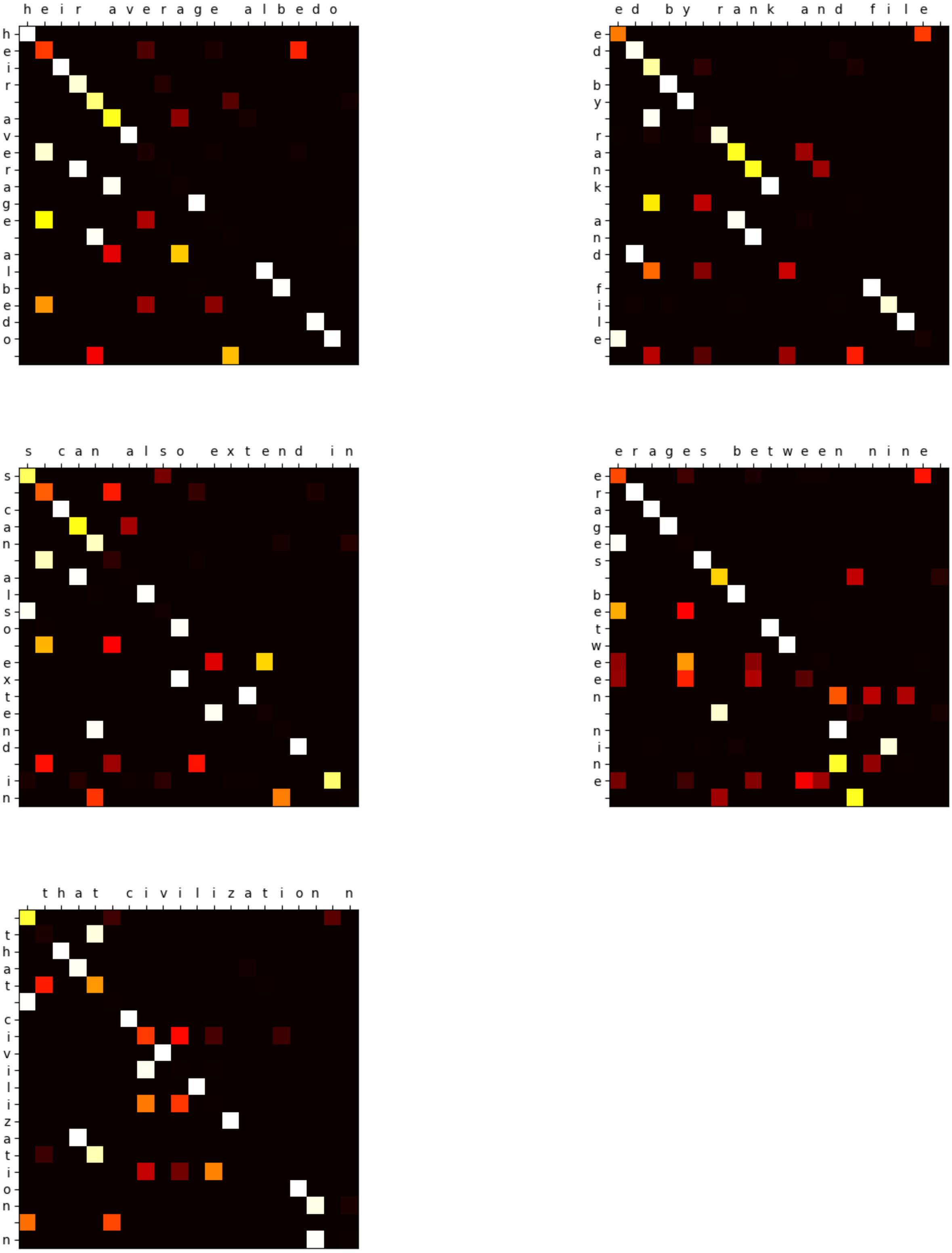

Figure 1: Standard attention maps .

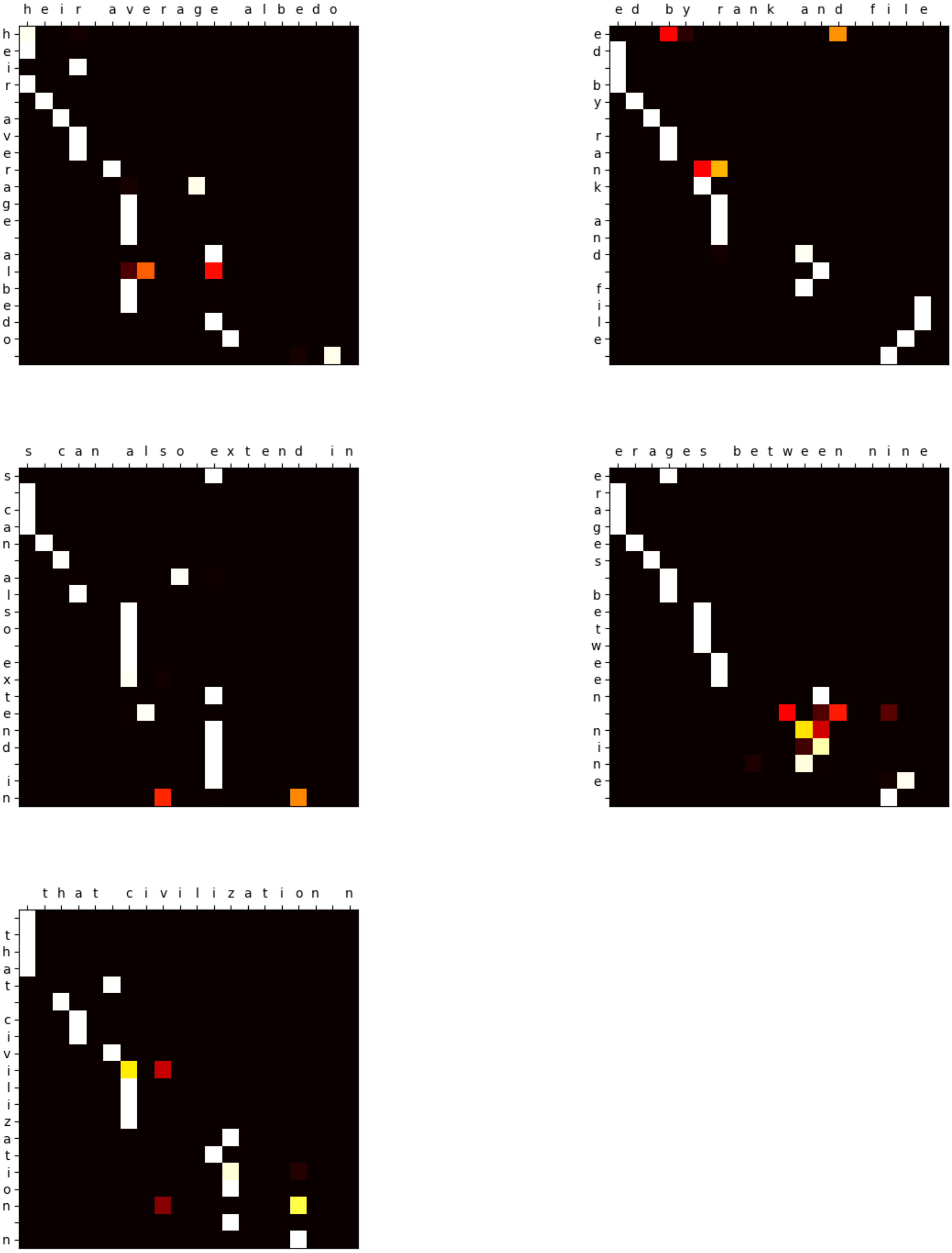

Figure 2: ALiBi attention maps .