Report for PA2

Part 1

Model Architecture

The model architecture used for this task is a Transformer Encoder-based classifier. The encoder consists of multiple layers with self-attention mechanisms and feedforward neural networks. Positional encoding was added to the input embeddings to incorporate sequential information, allowing the model to handle order within the input sequences effectively.

- **Total Parameters**: 496,611

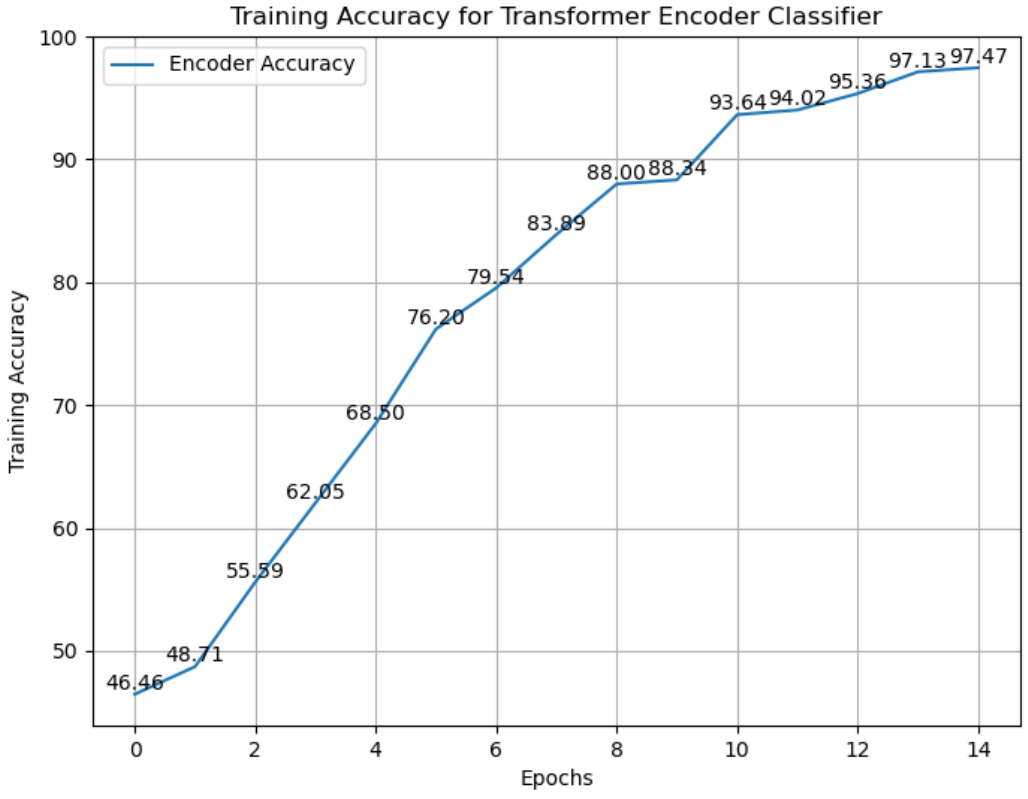
Training Process and Results

The model was trained for multiple epochs, and the training accuracy was recorded at each epoch. The training accuracy curve (shown in `part1.png`) illustrates a steady increase in accuracy over the epochs, indicating that the model effectively learned from the training data.

- **Final Training Accuracy**: 97.47% (on training data)
- **Final Evaluation Accuracy**: 84.27% (on test data)

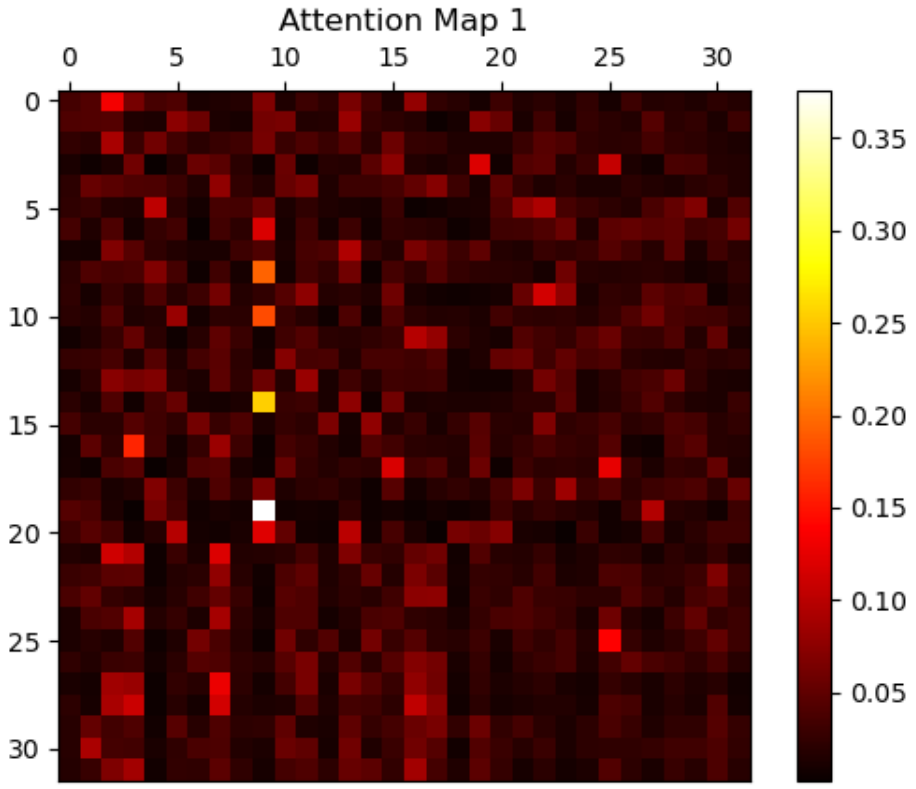
The model achieved a high training accuracy, suggesting that it could capture the underlying patterns in the training data. However, the final accuracy on the test set is 84.27%, indicating a possible overfitting issue where the model performs slightly less effectively on unseen data.

Visualization of Training Accuracy



The training accuracy plot shows the progression of accuracy across epochs. The model demonstrates a consistent improvement, achieving approximately 97.47% accuracy by the end of training. This curve reflects that the model converged well, with rapid gains in the early epochs, which plateaued as training progressed.

Attention Matrix Visualization



The attention matrix visualizes the self-attention weights from the first layer in the encoder. The sentence I choose is "That is in Israel's interest, Palestine's interest, America's interest, and the world's interest." This matrix highlights how different tokens in a sequence attend to each other, with lighter colors indicating higher attention weights.

In this particular attention map:

- Tokens with higher weights (lighter cells) suggest a stronger focus or dependency between certain positions within the input.
- The distribution of attention weights suggests that the model is learning contextual relationships between tokens, which is essential for capturing semantic information in text classification.

Conclusion

The Transformer Encoder model showed promising results, achieving a high training accuracy and a satisfactory evaluation accuracy. The training and attention visualizations indicate that the model is effectively learning meaningful patterns, although there may be a degree of overfitting. Future work could focus on regularization techniques or hyperparameter tuning to improve generalization on the test data.

model parameter size: 496611

84.27%

Part 2

The model uses a Transformer Decoder architecture designed for text generation. The decoder consists of multiple layers, each containing masked self-attention mechanisms to allow for auto-regressive generation, where each token is generated based on previously generated tokens. Positional encoding was included to enable the model to capture sequential information effectively.

- **Total Parameters**: 863,883

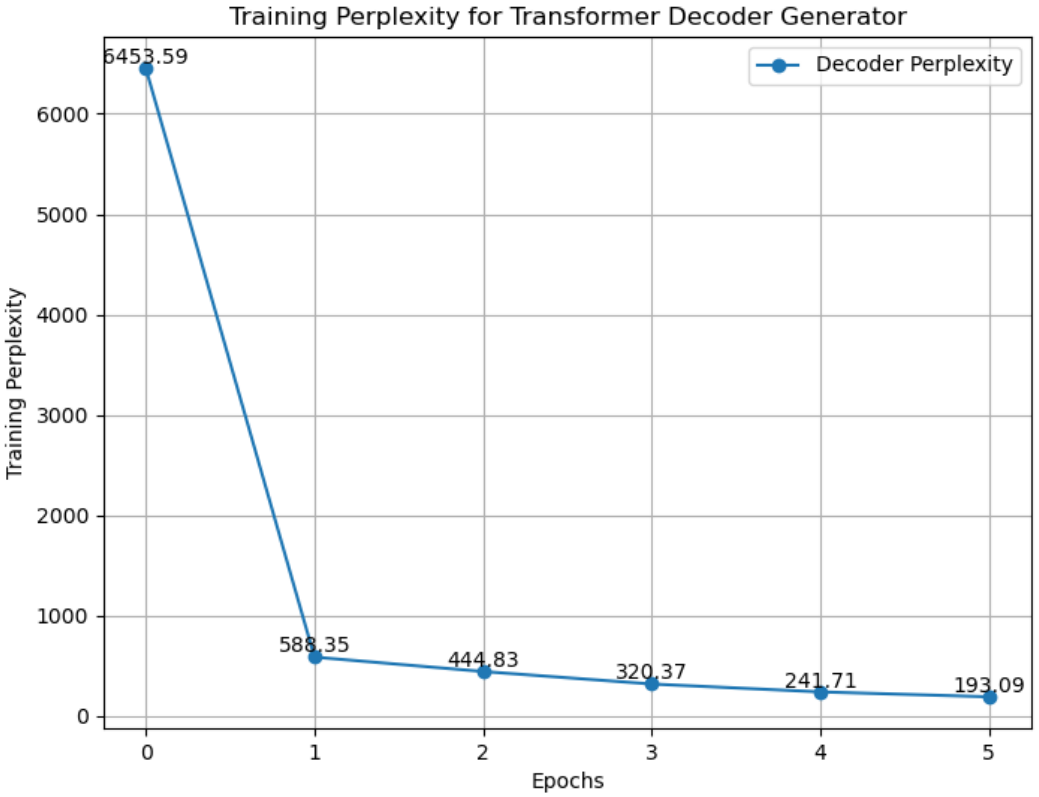
Training Process and Results

The model was trained on a corpus of speeches and evaluated using perplexity, a common metric for language models. Lower perplexity indicates a better fit to the training data, meaning the model can more accurately predict the next word in a sequence.

- **Final Training Perplexity**: 191.91
- **Perplexity on Obama Test Set**: 375.77
- **Perplexity on H. Bush Test Set**: 424.80
- **Perplexity on W. Bush Test Set**: 499.03

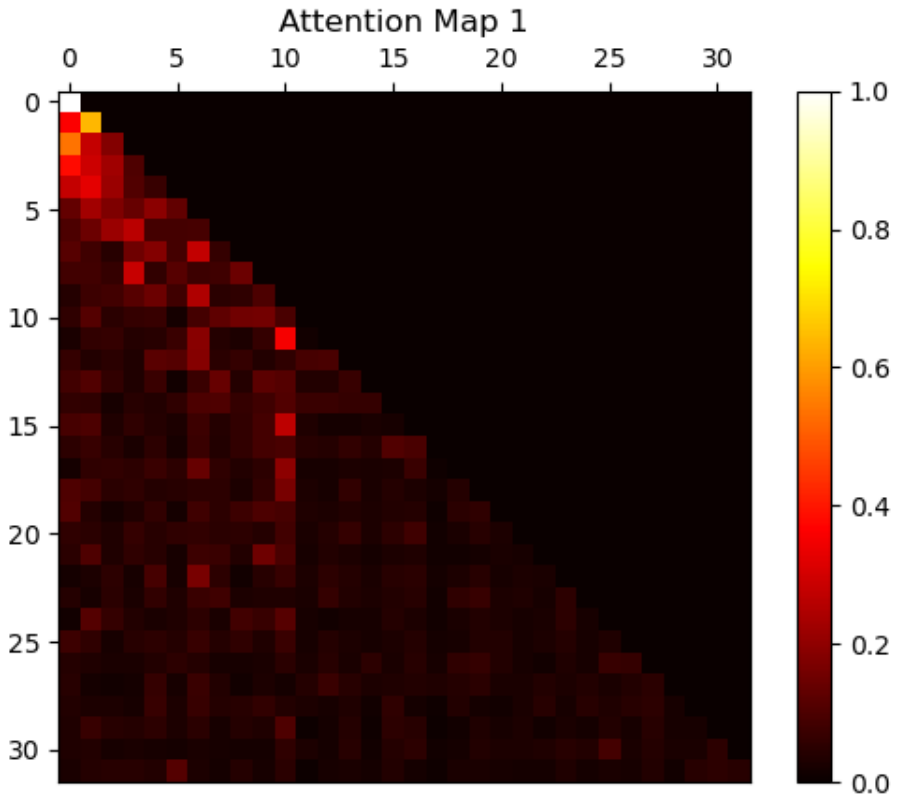
The model achieved a final training perplexity of 191.91, which suggests a reasonable fit to the training data. However, the perplexity on test sets for each politician is higher, indicating a degree of overfitting and suggesting that the model may not generalize perfectly to unseen data. The differences in test perplexity across politicians might be due to variations in their linguistic styles and vocabulary usage.

Visualization of Training Perplexity



The training perplexity plot illustrates the decline in perplexity over the training epochs. The model's perplexity decreased rapidly during the initial epochs, reaching a more stable value in later epochs. This trend indicates that the model was able to learn a meaningful representation of the data, allowing it to better predict the next token as training progressed.

Attention Matrix Visualization



The attention matrix for the decoder visualizes the self-attention weights. This matrix highlights how each token attends to previous tokens in the sequence, with lighter colors representing higher attention values. The sentence I choose is "t is costly and politically difficult to continue this conflict."

Key observations from the attention map:

- The diagonal structure shows that tokens mainly attend to nearby tokens, reflecting the model's reliance on local context.
- Stronger attention values for certain positions may indicate learned dependencies, allowing the model to maintain consistency in generated sequences.

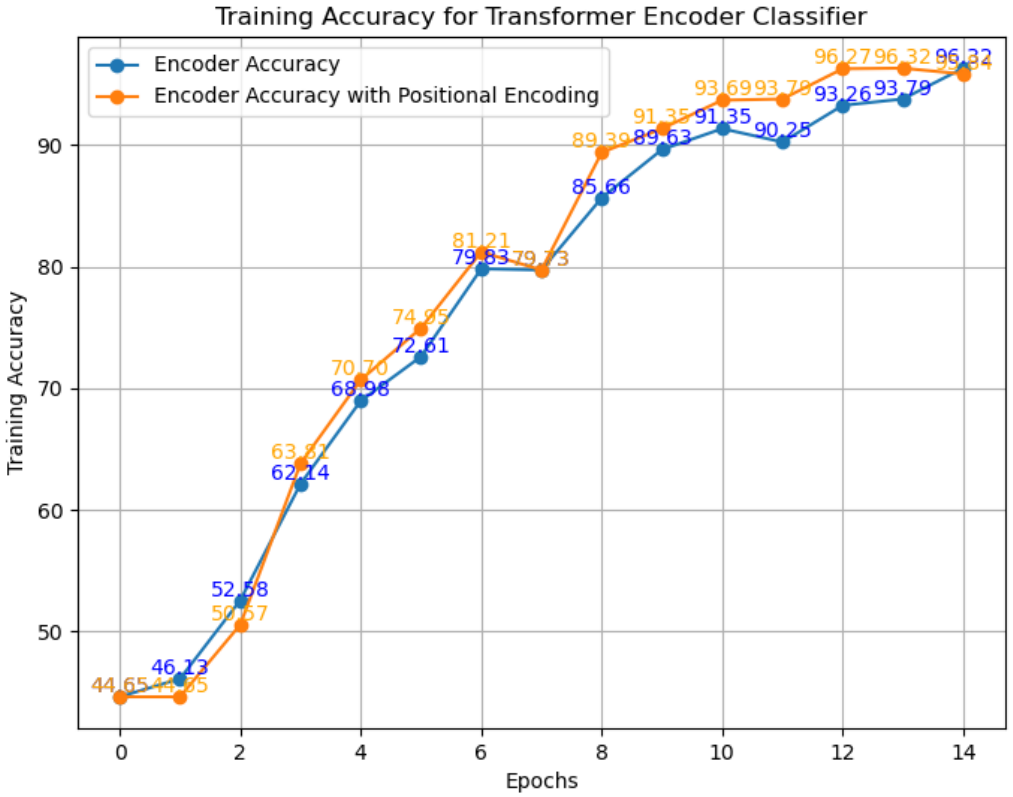
Conclusion

The Transformer Decoder model was able to learn patterns within the training data, as evidenced by the significant reduction in training perplexity. However, the higher perplexity on the test sets suggests that while the model performs well on the training data, it may require further tuning or regularization to generalize better on new, unseen data.

Part 3

I use Positional Encoding to explore alternatives to traditional positional encodings, i.e. AliBi.

Encoder Results



The encoder was trained as a classifier, and training accuracy was monitored over epochs. The models' final test accuracies were compared to evaluate the effectiveness of AliBi positional encoding.

- Encoder without AliBi: **Final Test Accuracy**: 84.20%
- Encoder with AliBi **Final Test Accuracy**: 85.00%

Training Accuracy Progression

The training accuracy plot shows the accuracy progression over 15 epochs for both encoder configurations. The accuracy with AliBi consistently matches or slightly outperforms the model without AliBi. The AliBi-enhanced encoder reaches higher accuracy sooner, indicating that it may facilitate faster convergence during training.

AliBi introduces a slight improvement in final test accuracy, suggesting that it helps the model generalize better on unseen data. This improvement, while modest, demonstrates the potential of AliBi to enhance performance in classification tasks.

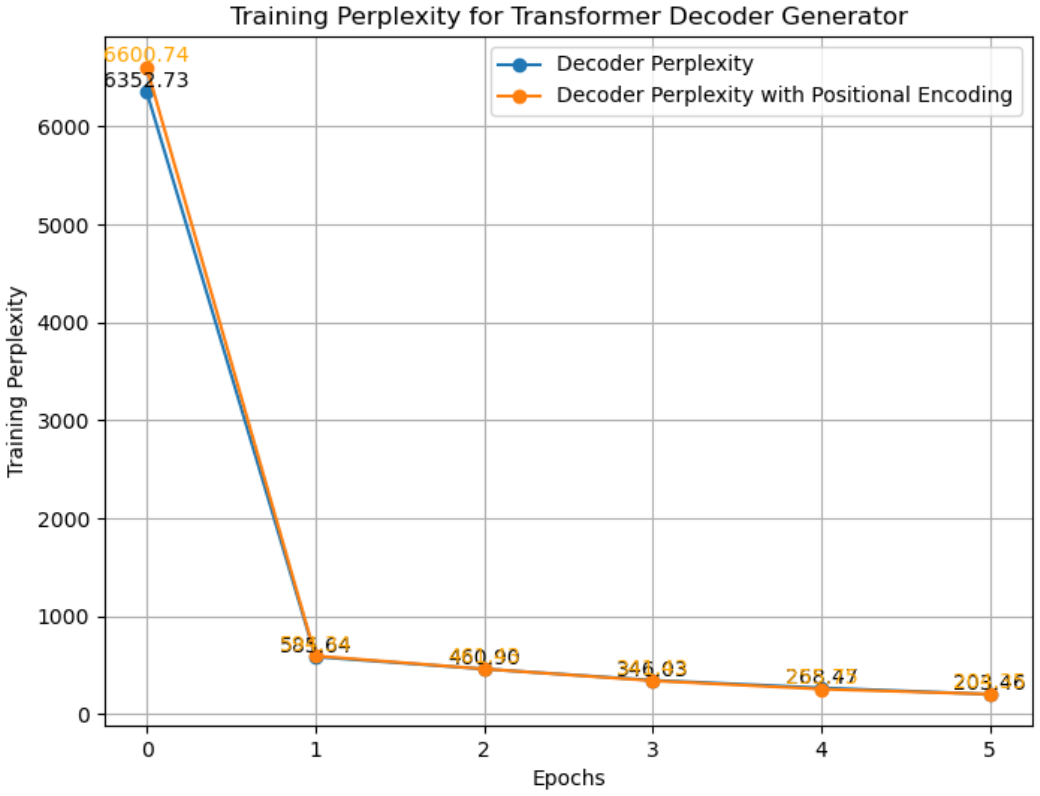
Decoder Results

The decoder was trained for a text generation task, and its performance was evaluated using perplexity. Lower perplexity indicates better model performance, as it reflects the model's confidence in generating accurate sequences.

- **Decoder without AliBi**:
  - **Final Training Perplexity**: 200.63
  - Test Perplexity
    :

        - **Obama**: 396.53
        - **H. Bush**: 421.98
        - **W. Bush**: 498.47
- **Decoder with AliBi**:
    - **Final Training Perplexity**: 190.68
    - Test Perplexity
      :
        - **Obama**: 372.32
        - **H. Bush**: 411.67
        - **W. Bush**: 474.61

Training Perplexity Progression



The training perplexity plot illustrates the reduction in perplexity over 5 epochs for both configurations. The decoder with AliBi consistently achieved lower perplexity, indicating that it learned to generate more accurate predictions and was more confident in its outputs.

The test perplexity also shows improvements across all three datasets (Obama, H. Bush, and W. Bush speeches) with AliBi. The perplexity reduction in the test sets suggests that AliBi helps the model generalize more effectively, particularly by lowering uncertainty when generating sequences in the style of different speakers.

Summary of Results

Model	Configuration	Final Training Perplexity / Accuracy	Test Accuracy / Perplexity (Obama, H. Bush, W. Bush)
Encoder (Classifier)	Without AliBi	97.47%	84.20%
	With AliBi	96.32%	85.00%
Decoder (Generator)	Without AliBi	200.63	396.53, 421.98, 498.47
	With AliBi	190.68	372.32, 411.67, 474.61

Conclusion

Introducing AliBi positional encoding to the Transformer model led to improvements in both the encoder's classification accuracy and the decoder's perplexity. The gains are most noticeable in the decoder's generalization on test sets, where AliBi helped reduce perplexity across different speakers' styles. This suggests that AliBi not only improves the model's ability to understand context within sequences but also enhances its robustness in text generation tasks.

Overall, AliBi proved to be a valuable addition to the Transformer architecture, contributing to both classification and text generation tasks. Further exploration could include fine-tuning AliBi-specific parameters or combining it with other attention optimization techniques.