Performance Comparison of Nano GPT Models
Character-Level ReLU Model vs Word-Level GeLU Model

Pankaja Balasooriya (SKF2400104)
August 3, 2024

1 Introduction

This report compares two versions of Nano GPT model trained with Tiny Shakespeare dataset using the character-level tokenization and ReLU activation, and the word-level tokenization model using GeLU activation. The changes aimed to improve the model's understanding of word-level context and potentially enhance its performance through the use of GeLU, which is common in modern language models.

2 Methodology

Both models are evaluated using **Training and validation loss, Perplexity, Training time, Number of Parameters and generated text**.Both models were trained for 15000 iterations on the same Tiny Shakespeare dataset, using identical hyperparameters except for the mentioned changes. Training of both models were performed on a Intel Core i7-13650HX, NVIDIA RTX 4060 6GB with 16GB of RAM and pytorch v2.3.1, Cuda v12.1.

3 Results

3.1 Quantitative Metrics

Metric	Character-Level (ReLU)	Word-Level (GeLU)
Final Training Loss	1.5327	2.5475
Final Validation Loss	1.7193	9.2367
Final Validation Perplexity	5.6016	11693.125
Training Time	7.69 minutes	8.60 minutes
Model Parameters	0.209729 M	3.512903 M
Vocabulary size	65	25671

3.2 Qualitative Assessment

Character-Level (ReLU) generated text:

```
KING RICHARD II:
I am spelw you love.  I cannot to the hope.
I the goldesson,- A what erichle, and doubt,
Then, in Francous
```

Word-Level (GeLU) generated text:

```
ISABELLA:
No, sir:  why, lady!  'tis as to tarry in I
could all affected as all and as free as if
But is it in a thing with the scene,
```

4 Discussion

1. **Loss and Perplexity**: The word-level model shows higher loss and perplexity values. This is expected due to the increased vocabulary size and complexity of predicting entire words rather than characters. However, it doesn't necessarily indicate poorer performance, as possible word combinations are much wider than character combinations.

2. **Training Time**: The word-level tokenized model took a little longer to train, likely due to the increased vocabulary size.

3. **Model Size**: The word-level model has more parameters, primarily due to the larger embedding layer needed for the expanded vocabulary.

4. **Generated Text Quality**: Both models generated text in a similar form to Shakespeare's plays, but the word-level model generated more natural-sounding language, even though both generated sentences with no meaning. Charater level model also generated words with incorrect spellings.

5 Conclusion

The transition to a word-level tokenization with GeLU activation has resulted in a more complex model with higher computational requirements. While metrics like loss and perplexity appear worse, this is largely due to the increased difficulty of the predicting at word level compaired to the character level. The qualitative assessment suggests that the word-level model produces higher quality, more naturally sounding text. This improvement in output quality may justify the increased computational cost and complexity.

Further tuning of hyperparameters for the word-level model may help improve its quantitative metrics while maintaining its qualitative advantages. Additionally, experimenting with different context lengths and model architectures could potentially enhance performance further.