CS572 Assignment 1 Report

Austin Brown

February 2024

1 Introduction

In this experiment, we aim to enhance the following Baseline GRU with batch size 64, 15 epochs, 32 BPTT length, 2 layers, hidden size of 768, embedding size of 128, and a gradient clipping threshold of 5. We will explore techniques to improve its performance. Currently it achieves a perplexity ≈ 124

2 Extensions and Modifications

2.1 Extension 1: Deeper Network

After experimenting with different modifications to the network architecture, including increasing the number of layers within the GRU and adjusting the size of the hidden layer, I found that a GRU with three layers yielded the best performance. Therefore, I decided to maintain this configuration and explored further enhancements with (new technique). The results of these experiments are summarized in the table. Additionally, the learning curve for the best-performing GRU is visualized in Figure 5.1.

2.2 Extension 2: Weight Decay

Based on the success of Weight Decay in improving the performance of the Neural N-gram model, I decided to apply it to this model as well. Setting the weight decay to 1e-5, I also reduced the dropout percentages from 0.5 to 0.3 to accommodate this regularization technique. This adjustment aimed to address slight overfitting observed in the model. With the introduction of weight decay reducing the risk of overfitting, I increased the number of epochs from its original value of 15 to 20 for further model refinement. With the introduction of weight decay, reducing the risk of overfitting, and an increase in the number of epochs from 15 to 20, the model achieved a perplexity of approximately 118, as shown in figure 5.2.

3 Conclusion

In conclusion, the experiments conducted showcased the effectiveness of leveraging a deeper GRU architecture, coupled with weight decay regularization, to mitigate overfitting and enhance model generalization. By fine-tuning hyperparameters and applying advanced regularization techniques, such as weight decay and dropout, I achieved a notable reduction in perplexity from 124 to 118. These findings highlight the importance of meticulous parameter tuning and regularization strategies in improving the performance of language models. Moving forward, deeper exploration into more sophisticated techniques holds promise for achieving even greater performance gains in language modeling tasks.

4 Experimental Results

Table 1: Summary of Experimental Results

Model	**Validation Perplexity**
Baseline GRU	124.70
Enhanced Deep GRU (3 Layers)	120.38
Advanced Deeper GRU (3 layers, Hidden = 1024)	157.46
Modified Less-Deep Deeper GRU (3 layers, Hidden = 512)	143.61
Optimized GRU (3 layers, Dropout = 0.3, Weight Decay = 1e-5, Epochs = 20)	118.15

5 Figures

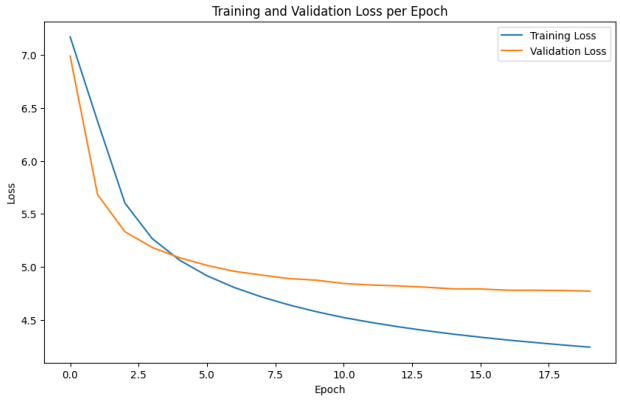

Figure 1: 3 Layer GRU Learning Curve

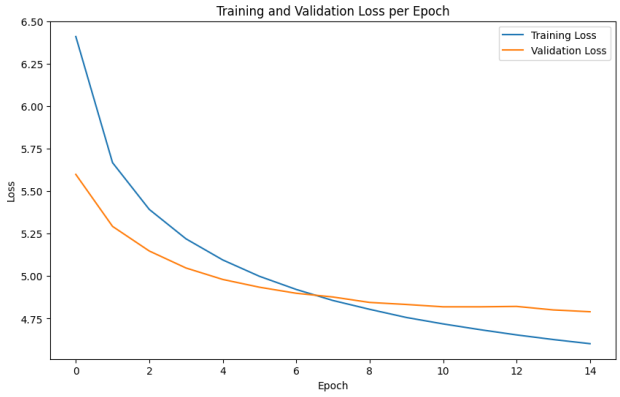

Figure 2: Improved GRU