Intro Deep Learning Homework 3

Jaskin Kabir
Student Id: 801186717
GitHub:
https://github.com/jaskinkabir/Intro_Deep_Learning/tree/master/HM3

February 2025

1 Problem 1: Character Prediction Small Dataset

1a. *Training Curves*

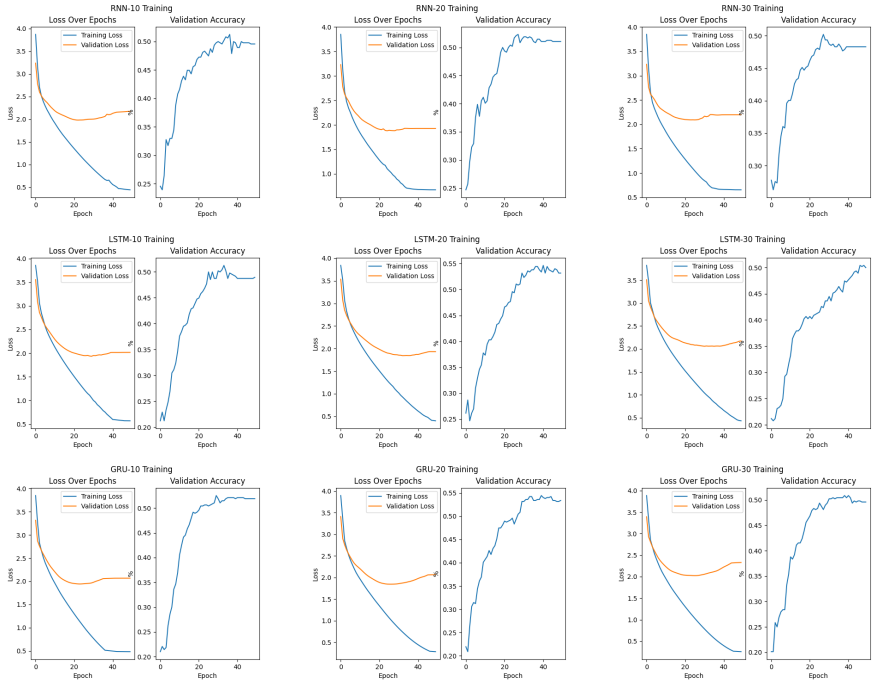

Figure 1: Problem 1 Training Curves

1b. *Results Comparison*

Model	Parameter Count	Training Time (s)	Overfit (%)	Accuracy (%)
RNN-10	44846	1.93	389.68	49.58
RNN-20	44846	3.65	185.97	51.05
RNN-30	44846	5.24	235.41	48.31
LSTM-10	143918	5.79	253.80	48.95
LSTM-20	143918	11.21	381.00	53.16
LSTM-30	143918	18.23	399.18	50.00
GRU-10	110894	5.30	329.01	51.89
GRU-20	110894	9.88	347.95	54.01
GRU-30	110894	15.41	460.00	50.85

Table 1: Problem 1 Data Comparison

1c. *Discussion*

- Parameter Count

  The parameter count of the LSTM network was more than triple that of the basic RNN network, and the GRU was almost halfway between the two. This is to be expected, as the LSTM adds significant complexity to the base RNN architecture, and the GRU reduces that complexity. This complexity comes from the number of gates in each architecture.

- Training Time

  The RNN was the quickest to train, followed by the GRU and then the LSTM. Training time seemed to increase linearly with sequence length for all architectures.

- Effect of Sequence Length

  The sequence length of 10 was too short for the models to have enough relevant information, while the length of 30 was too difficult for the models to learn. The sequence length of 20 was the best compromise between the two.

- Overfitting

  For the best sequence length of 20, the RNN had the lowest overfit and the LSTM the highest. This is likely a result of the difference in complexity between the architectures. The most complex model, the LSTM, had the highest overfit.

Interestingly, for the sequence length of 10, the RNN the highest overfit while the LSTM had the lowest. This could be due to the RNN not having enough information to memorize the training data, while the LSTM was able to memorize more from the shorter sequences.

For the sequence length of 30, the GRU had the highest overfit and the RNN the lowest. As the GRU was the most effective at generalizing from the training data, it makes sense that it would also overfit the most

- Accuracy

  Across all sequence lengths, the GRU was the most accurate and the RNN the least. The LSTM was in the middle. This is likely due to the complexity of the LSTM and the simplicity of the RNN. The GRU is a good compromise between the two, which is why it performed the best.

2 Problem 2: Shakespeare Character Prediction

1a. *Training Curves*

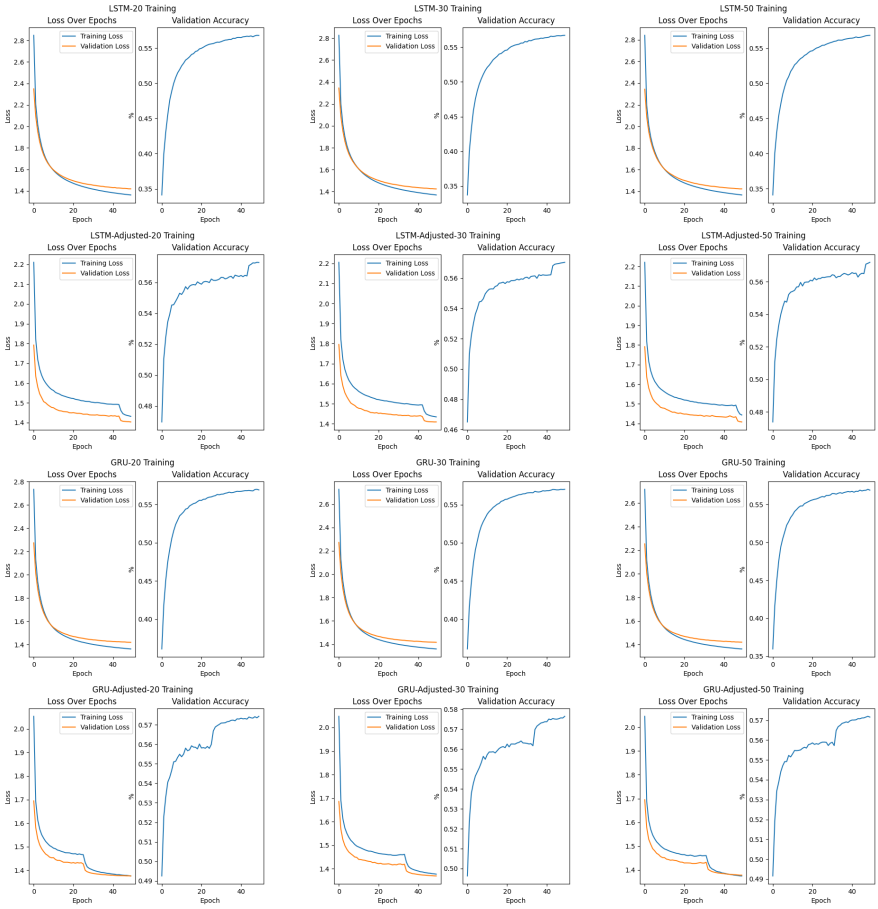

Figure 2: Problem 2 Training Curves

1b. *Results Comparison*

Model	Parameter Count	Training Time (s)	Inference Time (s)	Overfit (%)
LSTM-20	148801	346.11	3.83	56.62
LSTM-30	44846	3.65	185.97	51.05
LSTM-50	44846	5.24	235.41	48.31
LSTM-Adjusted-20	143918	5.79	253.80	48.95
LSTM-Adjusted-30	143918	11.21	381.00	53.16
LSTM-Adjusted-50	143918	18.23	399.18	50.00
GRU-20	115777	5.30	329.01	51.89
GRU-30	110894	9.88	347.95	54.01
GRU-50	110894	15.41	460.00	50.85
GRU-Adjusted-20	110894	5.30	329.01	51.89
GRU-Adjusted-30	110894	9.88	347.95	54.01
GRU-Adjusted-50	110894	15.41	460.00	50.85

Table 2: Problem 2 Data Comparison

1c. *Discussion*

- Hyperparameter and Architecture Modifications To improve the LSTM's performance, the sequence length was reduced to 100 and two dense layers with batch normalization and dropout were added to the end of the model. These layers had 128 and 64 neurons. The GRU was also modified in the same way, but the dense layers both had 256 neurons. Additionally, because of the dropout layers, the learning rate was increased tenfold. These same modifications were made for all sequence lengths.
- Parameter Count
  The LSTM had more parameters than the GRU because it is the more complex architecture. However, when the models were adjusted to

improve their performance, the LSTM had fewer parameters even after the additional layers because the reduction of sequence length was more significant for the LSTM. Additionally, the GRU's dense layers are 4 and 2 times larger than the LSTM's, respectively, which caused its adjusted parameter count to be the highest of the 4 models.

- Training Time

  Sequence length did not have a significant effect on training time. However, the GRU was significantly faster to train than the LSTM. Because the adjustments to the GRU model adds complexity, it took longer to train than the base model, whereas the adjusted LSTM model was faster to train than the base model.

- Inference Time

  Average inference time was on the order of milliseconds for each model. This is likely due to the small size of the models and the simplicity of the task. The recorded data shows fluctuations between the models, but there is no discernible pattern. These are likely due to inaccuracies in recording inference time when all samples in the batch are computed in parallel.

- Overfitting

  Both models were almost equally susceptible to overfit, and the adjustments made to the models eliminated this overfit thanks to the dropout layers.

- Accuracy

  The GRU was more accurate than the LSTM for all sequence lengths, and the adjustments made to the models improved their accuracy. However, these differences are small and likely not significant. The adjusted GRU trained on a sequence length of 30 was the most accurate model, and the LSTM trained on a sequence length of 20 the least. The difference between the two was just 1%.