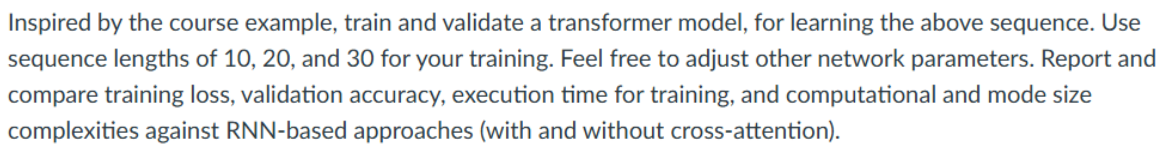
Source Code

Remote Repository: https://github.com/Mikestriken/Deep-Learning-Class
Name: Michael Marais
Student ID: 801177649
Homework Number: 5

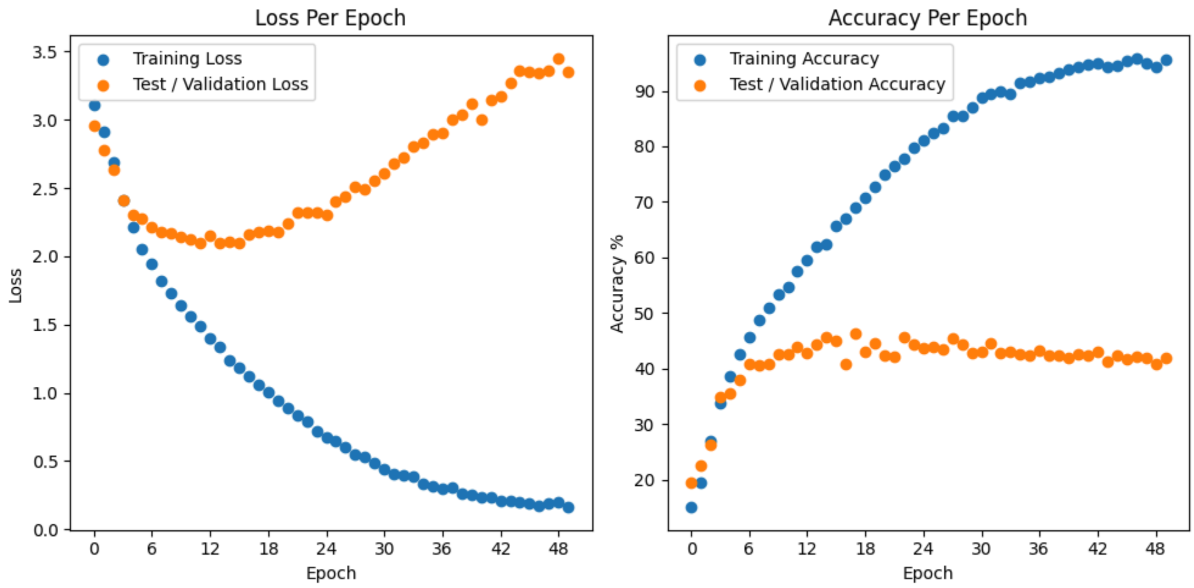
Problem 1



Inspired by the course example, train and validate a transformer model, for learning the above sequence. Use sequence lengths of 10, 20, and 30 for your training. Feel free to adjust other network parameters. Report and compare training loss, validation accuracy, execution time for training, and computational and mode size complexities against RNN-based approaches (with and without cross-attention).

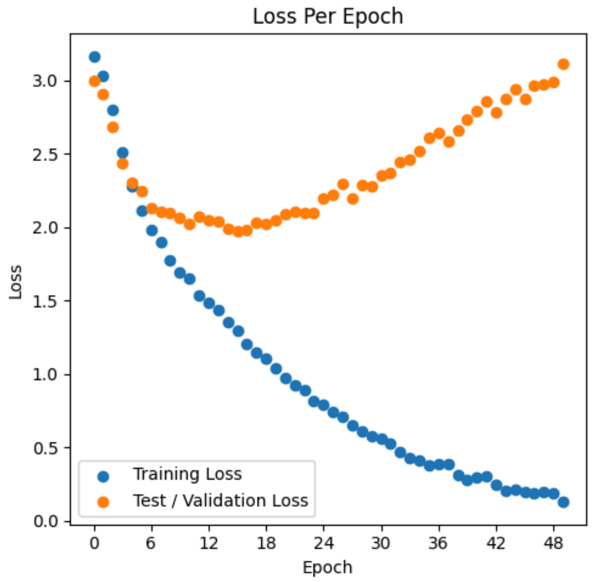
Transformer

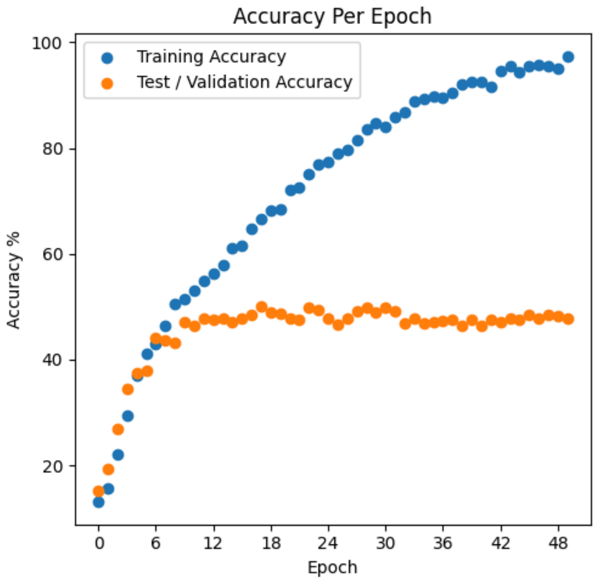
Sequence Length 10



Last Best Epoch: 17, Train Time ≈ 0 hr, 0 min, 46 sec
Train Loss: 1.05581, Train Accuracy: 68.95%
Test Loss: 2.18086, Test Accuracy: 46.43%
Model Parameters: 3,188,781, MACS: 369,844,252

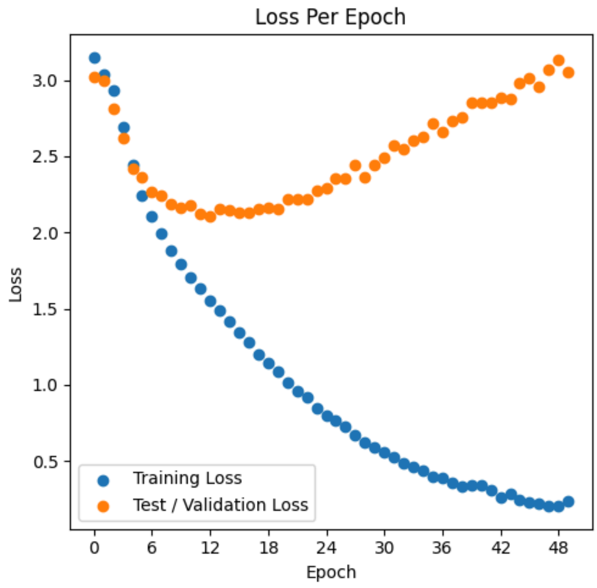
Sequence Length 20

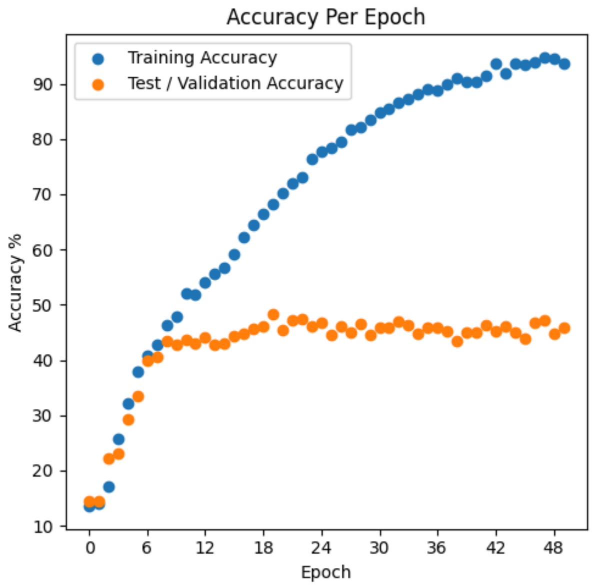




Last Best Epoch: 17, Train Time ≈ 0 hr, 0 min, 46 sec
Train Loss: 1.14845, Train Accuracy: 66.54%
Test Loss: 2.03023, Test Accuracy: 50.00%
Model Parameters: 3,188,781, MACS: 705,716,252

Sequence Length 30

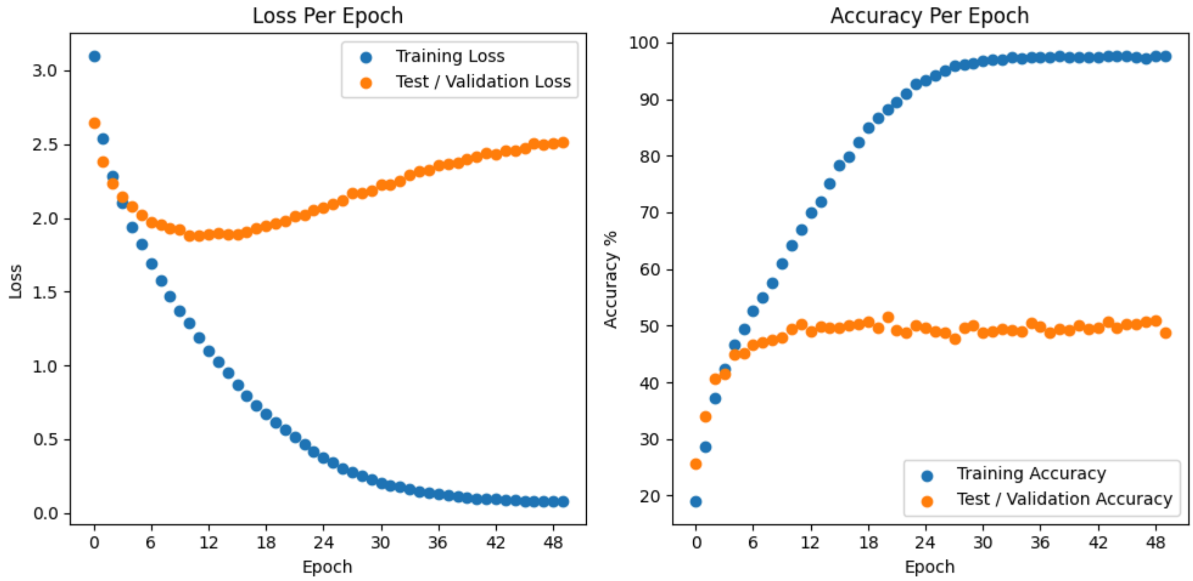




Last Best Epoch: 19, Train Time ≈ 0 hr, 0 min, 46 sec
Train Loss: 1.08561, Train Accuracy: 68.26%
Test Loss: 2.15687, Test Accuracy: 48.31%
Model Parameters: 3,188,781, MACS: 1,041,588,252

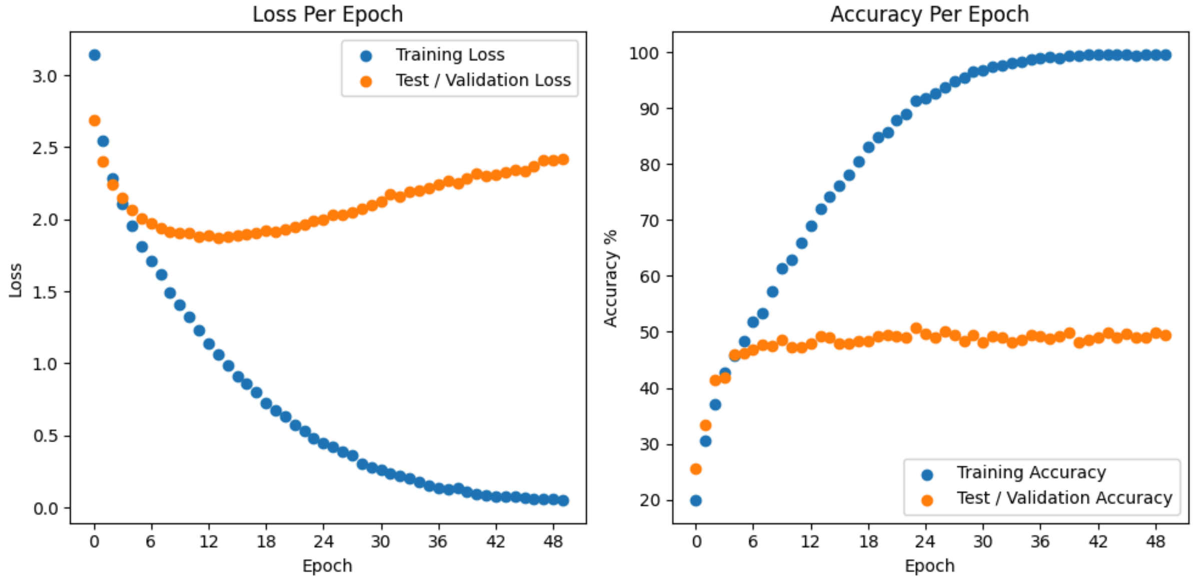
nn.GRU

Sequence Length 10



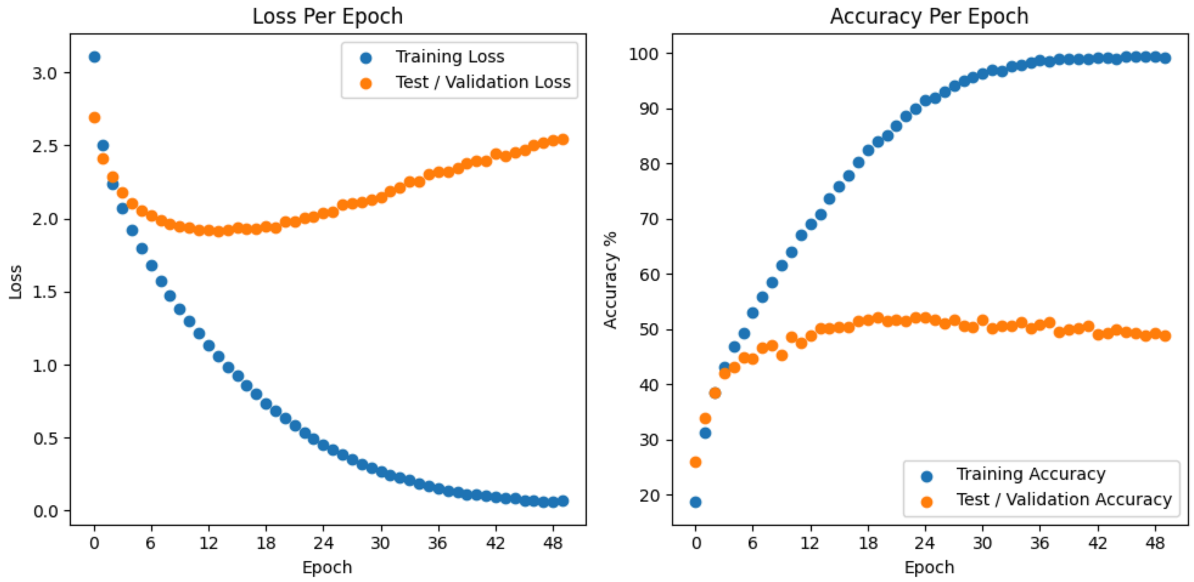
Last Best Epoch: 20, Train Time ≈ 0 hr, 0 min, 37 sec
Train Loss: 0.56217, Train Accuracy: 88.26%
Test Loss: 1.97599, Test Accuracy: 51.68%
Model Parameters: 143,404, MACS: 180,224

Sequence Length 20



Last Best Epoch: 23, Train Time ≈ 0 hr, 0 min, 10 sec
Train Loss: 0.48214, Train Accuracy: 91.38%
Test Loss: 1.99033, Test Accuracy: 50.84%
Model Parameters: 143,404, MACS: 180,224

Sequence Length 30

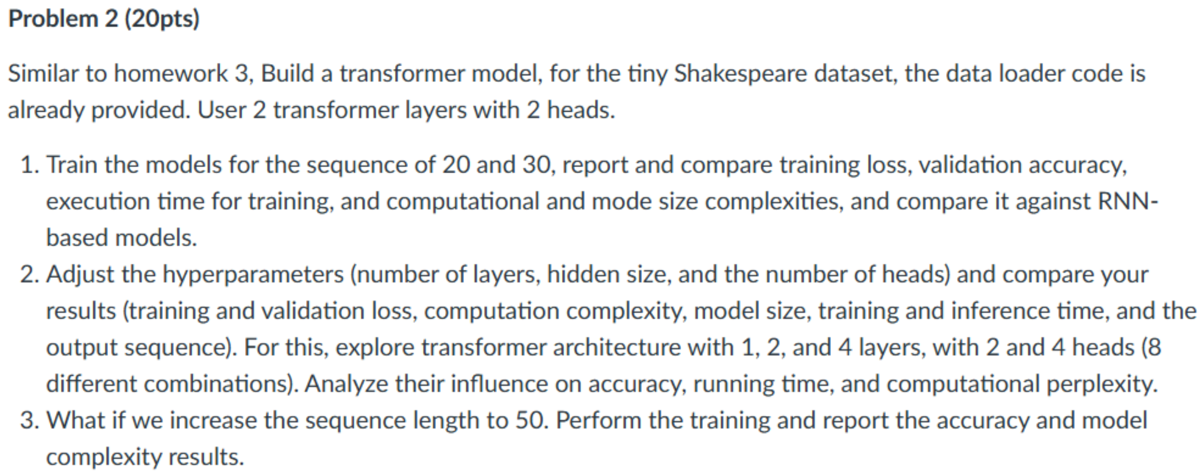


Last Best Epoch: 19, Train Time ≈ 0 hr, 0 min, 9 sec
Train Loss: 0.68354, Train Accuracy: 84.02%
Test Loss: 1.93846, Test Accuracy: 52.12%
Model Parameters: 143,404, MACS: 180,224

Observations

RNN architecture performed better for this problem, most likely due to the dataset being small.

Problem 2



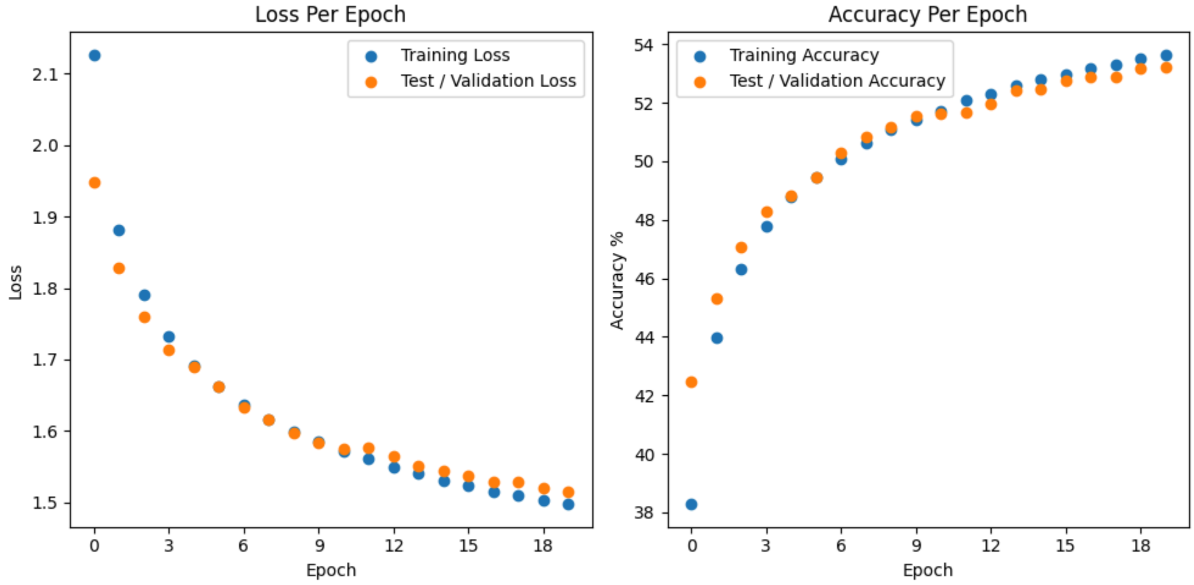
Problem 2 (20pts)

Similar to homework 3, Build a transformer model, for the tiny Shakespeare dataset, the data loader code is already provided. User 2 transformer layers with 2 heads.

1. Train the models for the sequence of 20 and 30, report and compare training loss, validation accuracy, execution time for training, and computational and mode size complexities, and compare it against RNN-based models.
2. Adjust the hyperparameters (number of layers, hidden size, and the number of heads) and compare your results (training and validation loss, computation complexity, model size, training and inference time, and the output sequence). For this, explore transformer architecture with 1, 2, and 4 layers, with 2 and 4 heads (8 different combinations). Analyze their influence on accuracy, running time, and computational perplexity.
3. What if we increase the sequence length to 50. Perform the training and report the accuracy and model complexity results.

Transformer Model: 1 Layer; 2 Heads

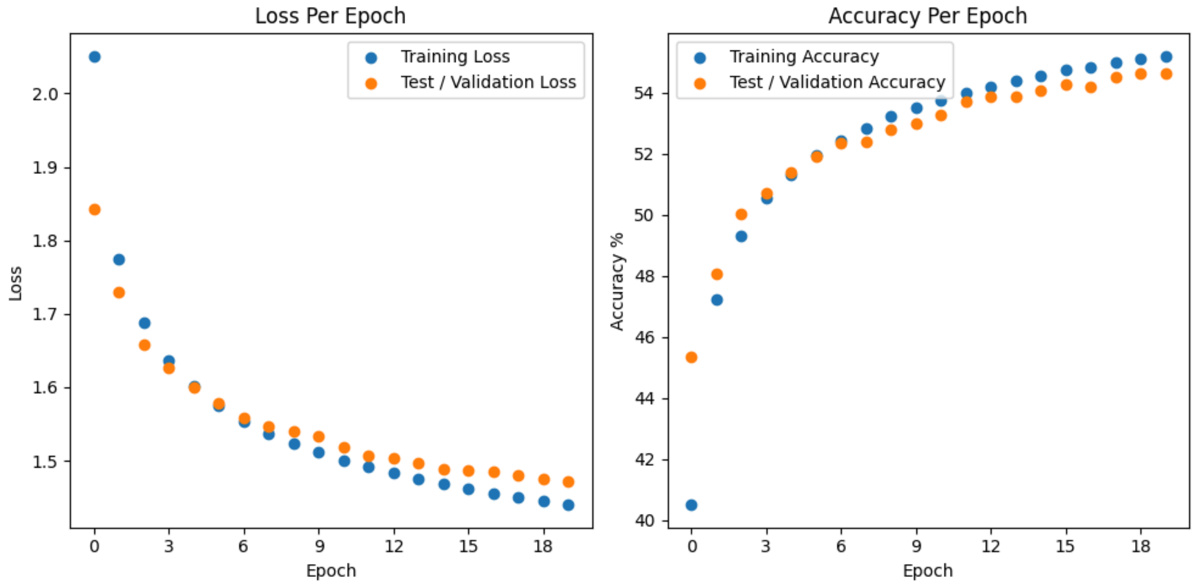
Sequence Length 20



Last Best Epoch: 19, Train Time ≈ 0 hr, 22 min, 35 sec
Train Loss: 1.49754, Train Accuracy: 53.64%
Test Loss: 1.51511, Test Accuracy: 53.23%
Model Parameters: 2,144,322, MACS: 1,412,857,870

Transformer Model: 1 Layer; 4 Heads

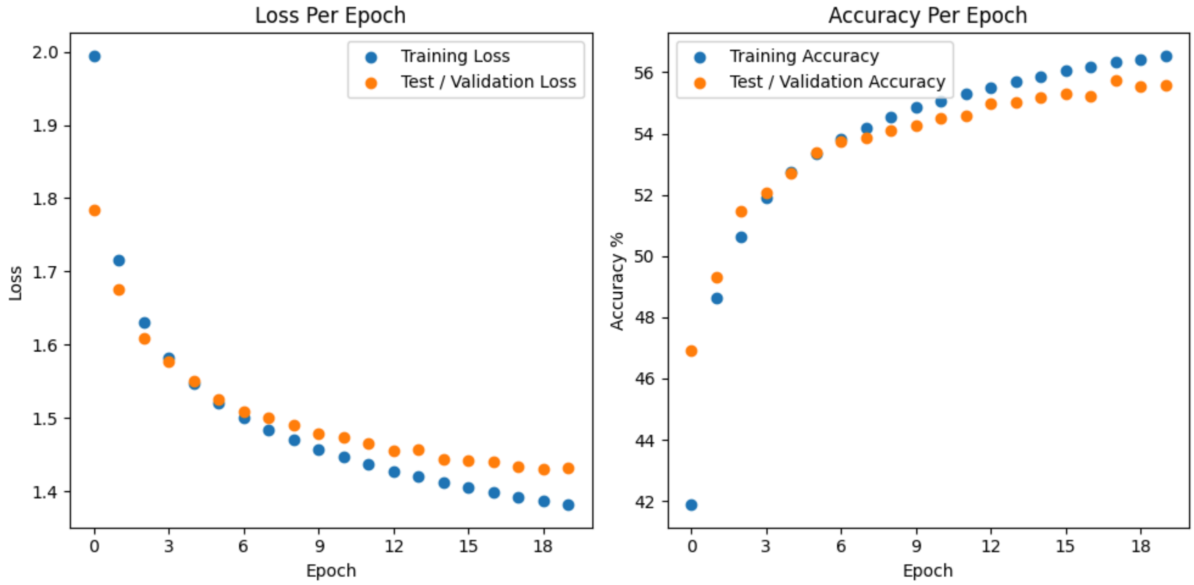
Sequence Length 20



Last Best Epoch: 19, Train Time ≈ 0 hr, 22 min, 58 sec
Train Loss: 1.43971, Train Accuracy: 55.20%
Test Loss: 1.47135, Test Accuracy: 54.63%
Model Parameters: 2,144,322, MACS: 1,412,857,870

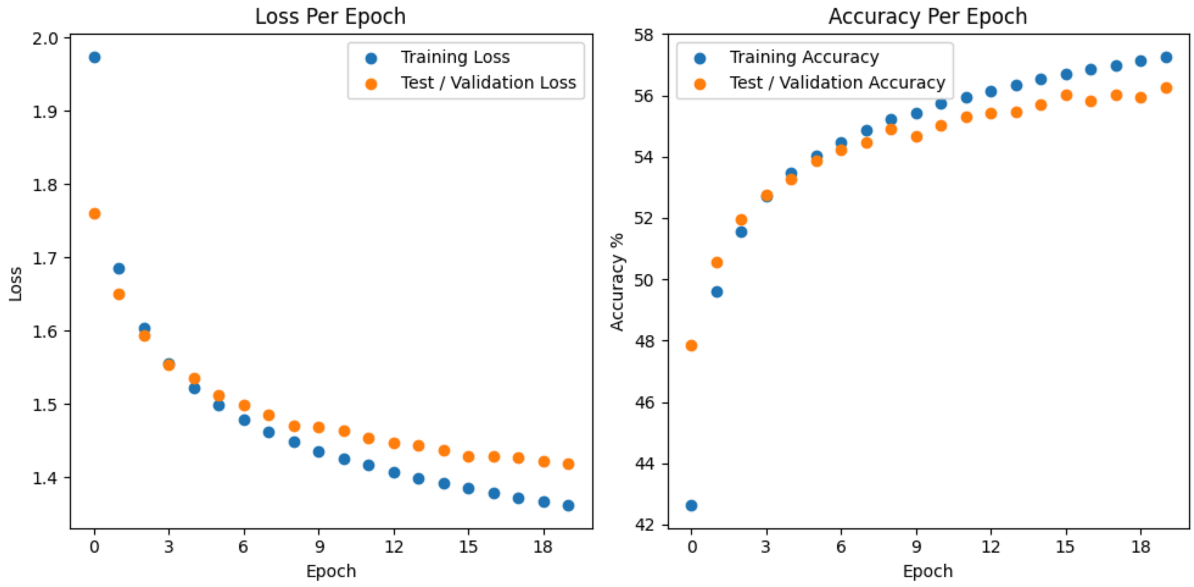
Transformer Model: 2 Layers; 2 Heads

Sequence Length 20



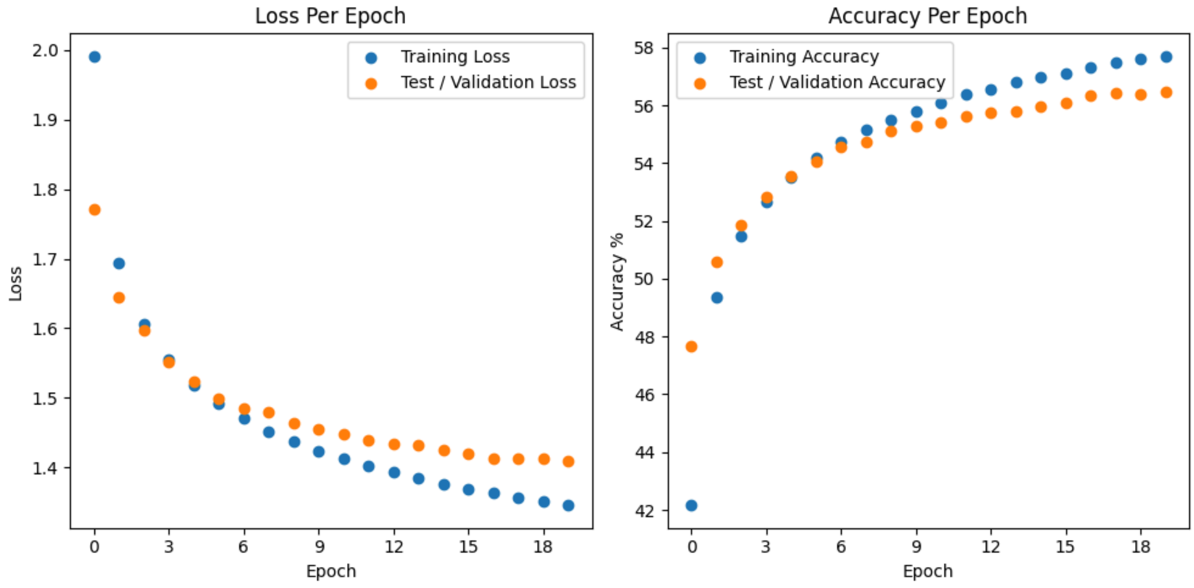
Last Best Epoch: 17, Train Time ≈ 0 hr, 39 min, 17 sec
Train Loss: 1.39237, Train Accuracy: 56.33%
Test Loss: 1.43322, Test Accuracy: 55.75%
Model Parameters: 3,199,554, MACS: 2,823,553,052

Sequence Length 30



Last Best Epoch: 19, Train Time ≈ 0 hr, 37 min, 58 sec
Train Loss: 1.36142, Train Accuracy: 57.27%
Test Loss: 1.41831, Test Accuracy: 56.25%
Model Parameters: 3,199,554, MACS: 4,167,041,052

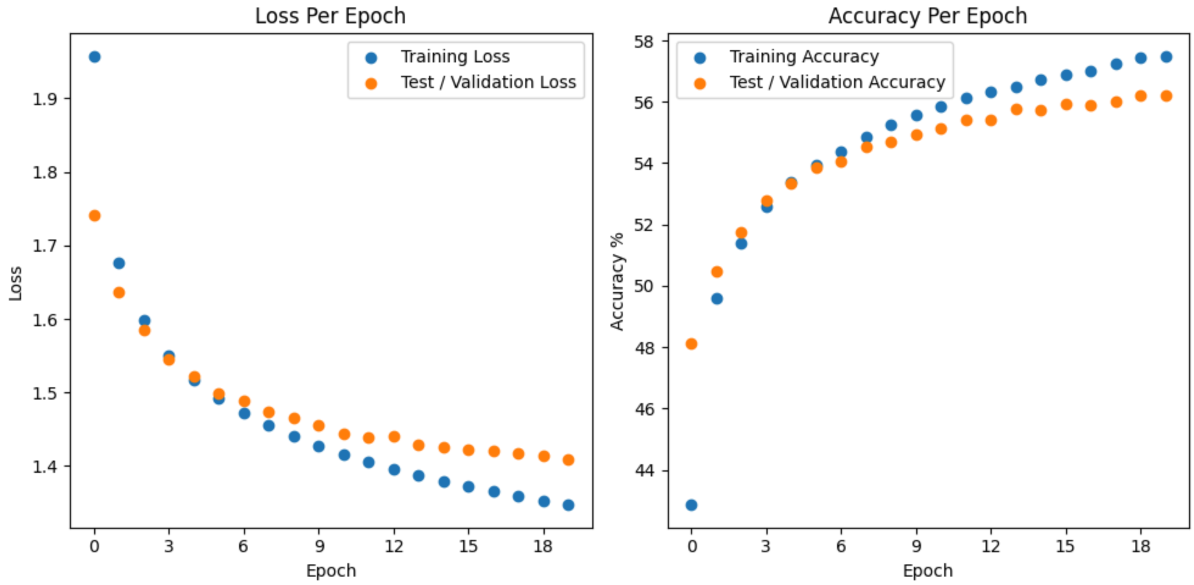
Sequence Length 50



Last Best Epoch: 19, Train Time ≈ 0 hr, 50 min, 2 sec
Train Loss: 1.34516, Train Accuracy: 57.71%
Test Loss: 1.40958, Test Accuracy: 56.47%
Model Parameters: 3,199,554, MACS: 6,854,017,052

Transformer Model: 2 Layers; 4 Heads

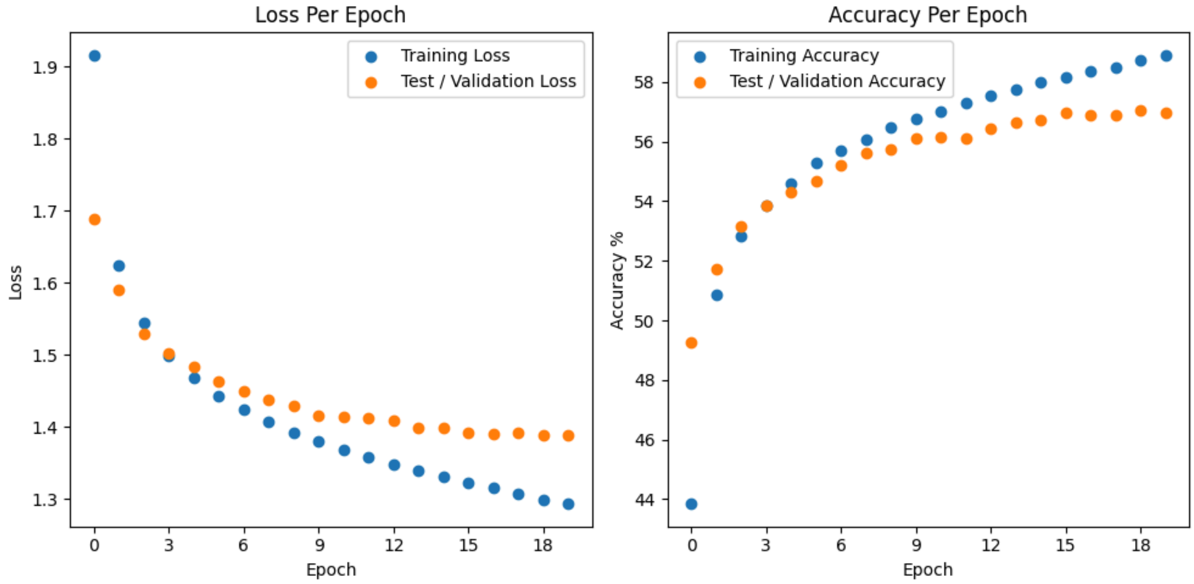
Sequence Length 20



Last Best Epoch: 18, Train Time ≈ 0 hr, 38 min, 56 sec
Train Loss: 1.35226, Train Accuracy: 57.43%
Test Loss: 1.41375, Test Accuracy: 56.23%
Model Parameters: 3,199,554, MACS: 2,823,553,052

Transformer Model: 4 Layers; 2 Heads

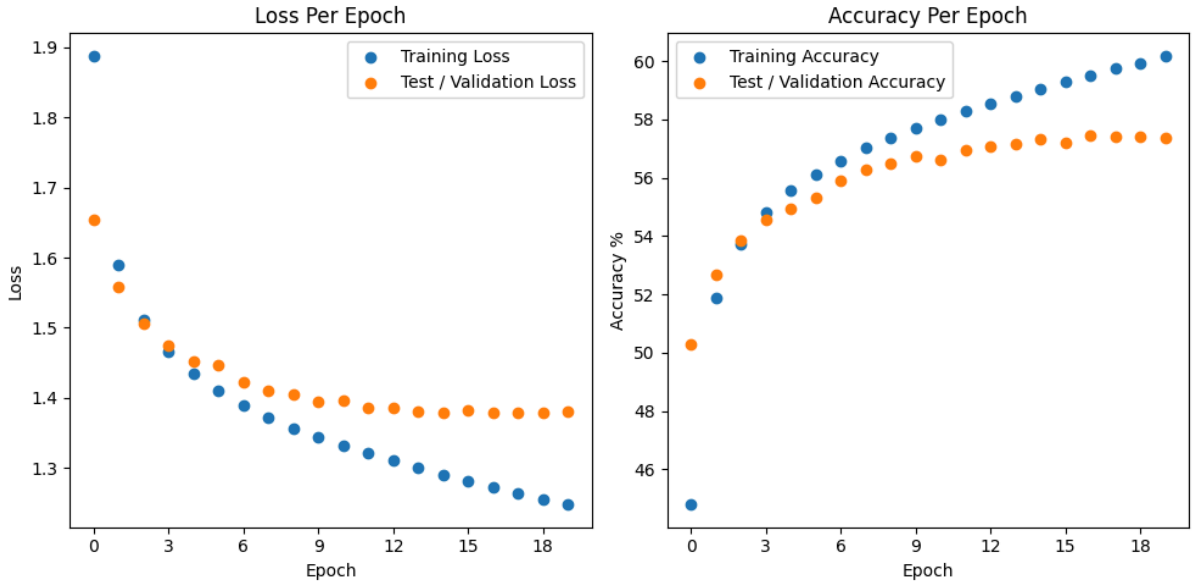
Sequence Length 20



Last Best Epoch: 18, Train Time ≈ 1 hr, 8 min, 1 sec
Train Loss: 1.29986, Train Accuracy: 58.73%
Test Loss: 1.38917, Test Accuracy: 57.04%
Model Parameters: 5,310,018, MACS: 5,644,943,416

Transformer Model: 4 Layers; 4 Heads

Sequence Length 20



Last Best Epoch: 16, Train Time ≈ 1 hr, 10 min, 3 sec

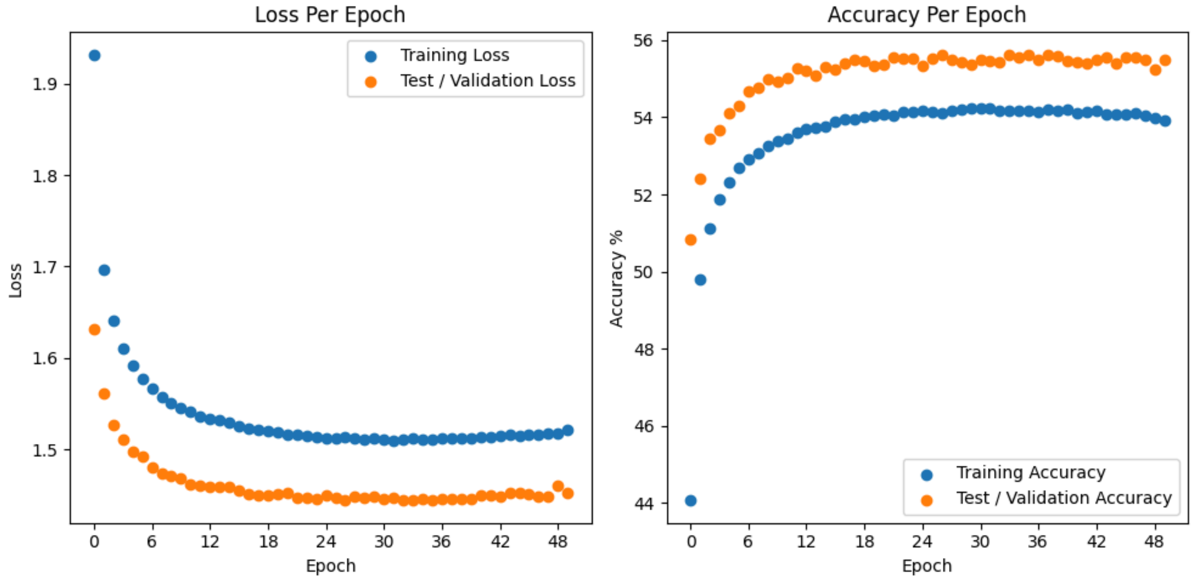
Train Loss: 1.27310, Train Accuracy: 59.51%

Test Loss: 1.37890, Test Accuracy: 57.45%

Model Parameters: 5,310,018, MACS: 5,644,943,416

nn.GRU

Sequence Length 20



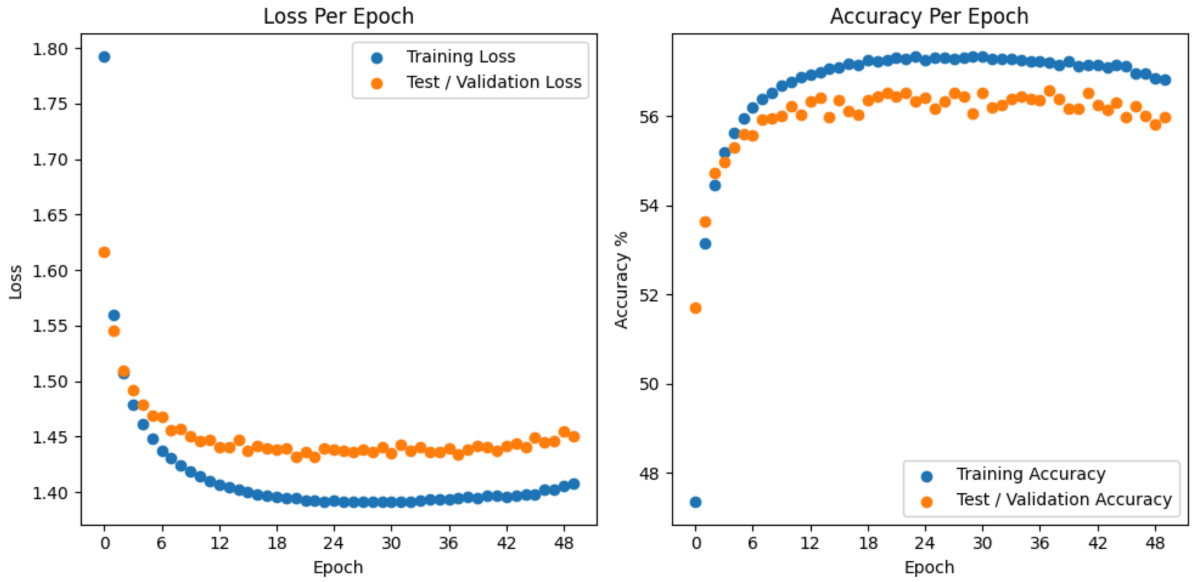
Last Best Epoch: 35, Train Time ≈ 0 hr, 22 min, 18 sec

Train Loss: 1.51130, Train Accuracy: 54.18%

Test Loss: 1.44501, Test Accuracy: 55.63%

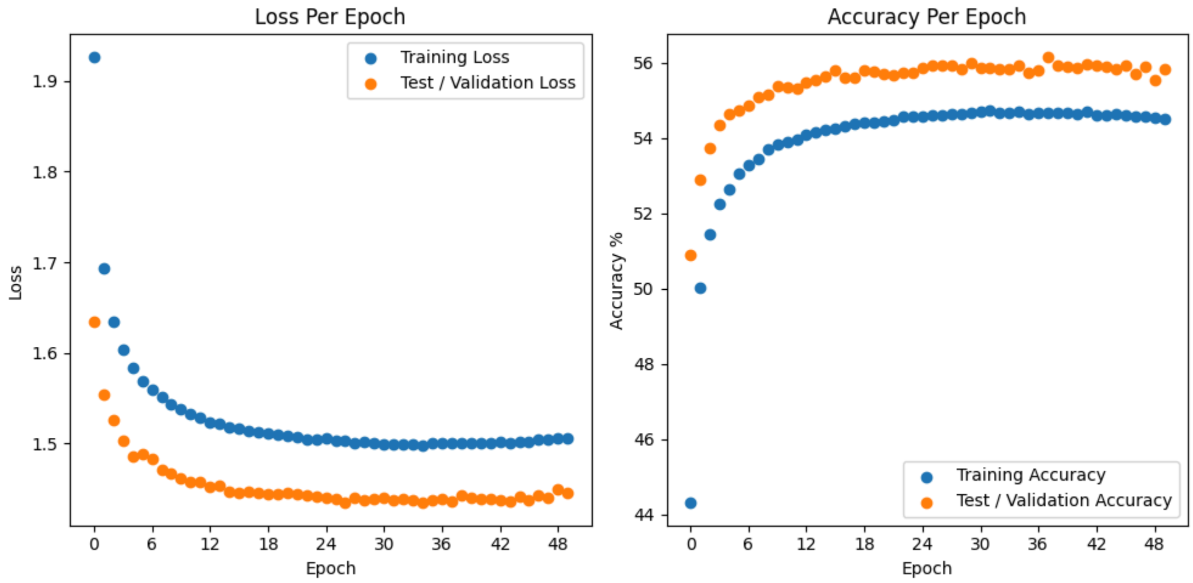
Model Parameters: 132,545, MACS: 3,178,496

Sequence Length 30



Last Best Epoch: 37, Train Time ≈ 0 hr, 21 min, 19 sec
Train Loss: 1.39453, Train Accuracy: 57.20%
Test Loss: 1.43341, Test Accuracy: 56.57%
Model Parameters: 132,545, MACS: 3,178,496

Sequence Length 50



Last Best Epoch: 43, Train Time ≈ 0 hr, 22 min, 19 sec
Train Loss: 1.38558, Train Accuracy: 57.42%
Test Loss: 1.42354, Test Accuracy: 56.77%
Model Parameters: 132,545, MACS: 3,178,496

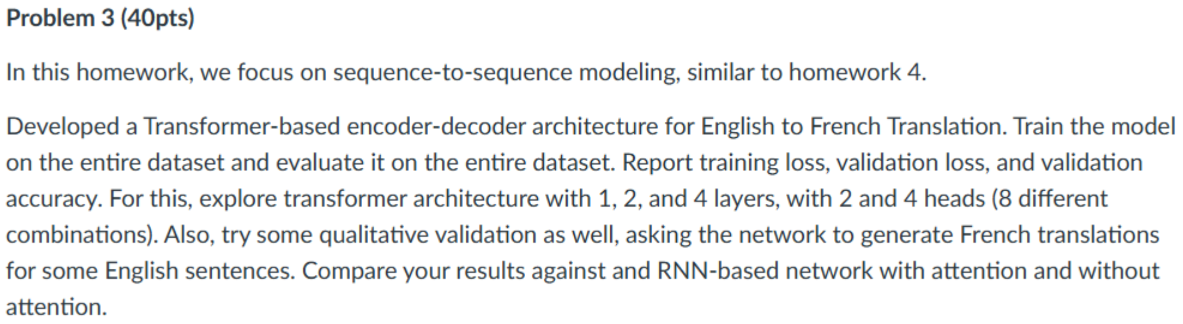
Observations

The performance results of the Transformer model in comparison with the RNN model are similar within percent error. Despite the Transformer model being trained for half the number of epochs, they both plateau at around the 20th epoch.

Computation Perplexity for a transformer scales linearly with sequence length due to the benefits of the transformer model being able to parallelize a lot more than an RNN, which doesn’t scale at all. Ironically, for some reason, however, the Transformer model took longer to train.

Performance increased with higher head and layer counts. However, it should be noted that it isn’t by much, and there is a high likelihood that if given sufficient training time their final training results will be the same.

Problem 3

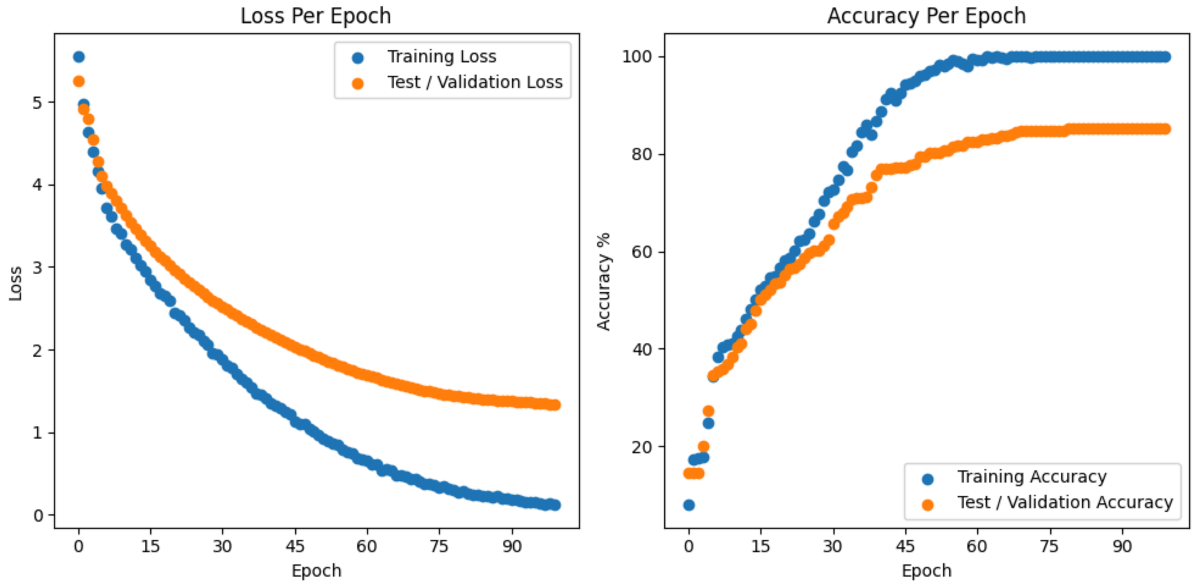


Problem 3 (40pts)

In this homework, we focus on sequence-to-sequence modeling, similar to homework 4.

Developed a Transformer-based encoder-decoder architecture for English to French Translation. Train the model on the entire dataset and evaluate it on the entire dataset. Report training loss, validation loss, and validation accuracy. For this, explore transformer architecture with 1, 2, and 4 layers, with 2 and 4 heads (8 different combinations). Also, try some qualitative validation as well, asking the network to generate French translations for some English sentences. Compare your results against and RNN-based network with attention and without attention.

Transformer Model: 1 Layer; 2 Heads

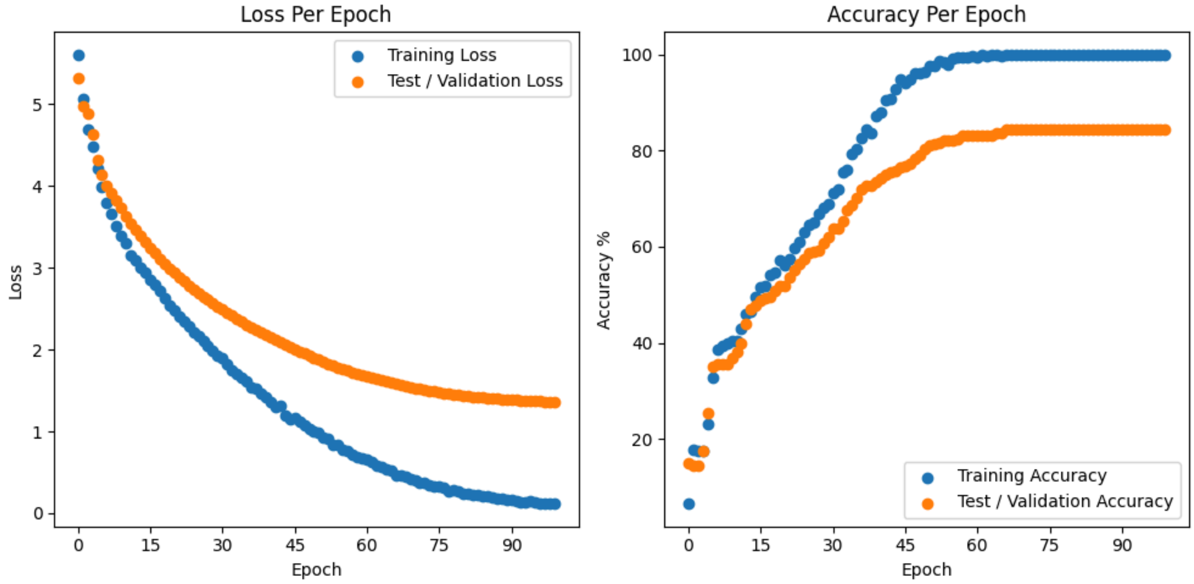


Last Best Epoch: 79, Train Time ≈ 0 hr, 0 min, 8 sec
Train Loss: 0.27481, Train Accuracy: 100.00%
Test Loss: 1.43561, Test Accuracy: 85.07%

Example Output:

- Input: We cook dinner on Sundays EOS
- Output: Nous cuisinons le dîner le le EOS
- Expected: Nous cuisinons le dîner le dimanche EOS

Transformer Model: 1 Layer; 4 Heads

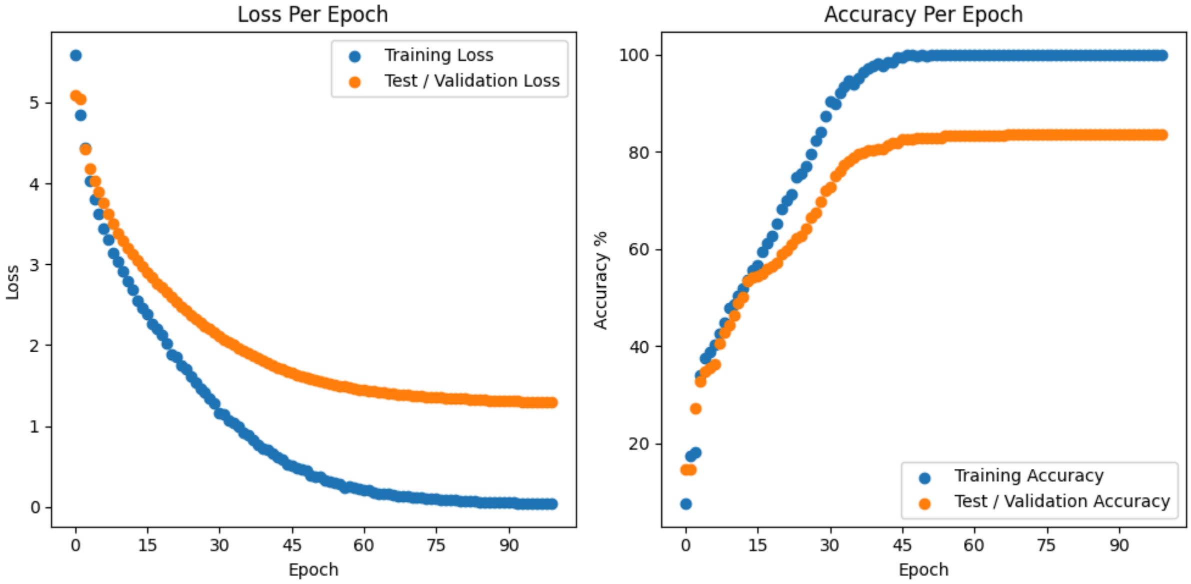


Last Best Epoch: 66, Train Time ≈ 0 hr, 0 min, 9 sec
Train Loss: 0.46940, Train Accuracy: 100.00%
Test Loss: 1.57877, Test Accuracy: 84.33%

Example Output:

- Input: We cook dinner on Sundays EOS
- Output: Nous cuisinons le dîner le dîner EOS
- Expected: Nous cuisinons le dîner le dimanche EOS

Transformer Model: 2 Layers; 2 Heads

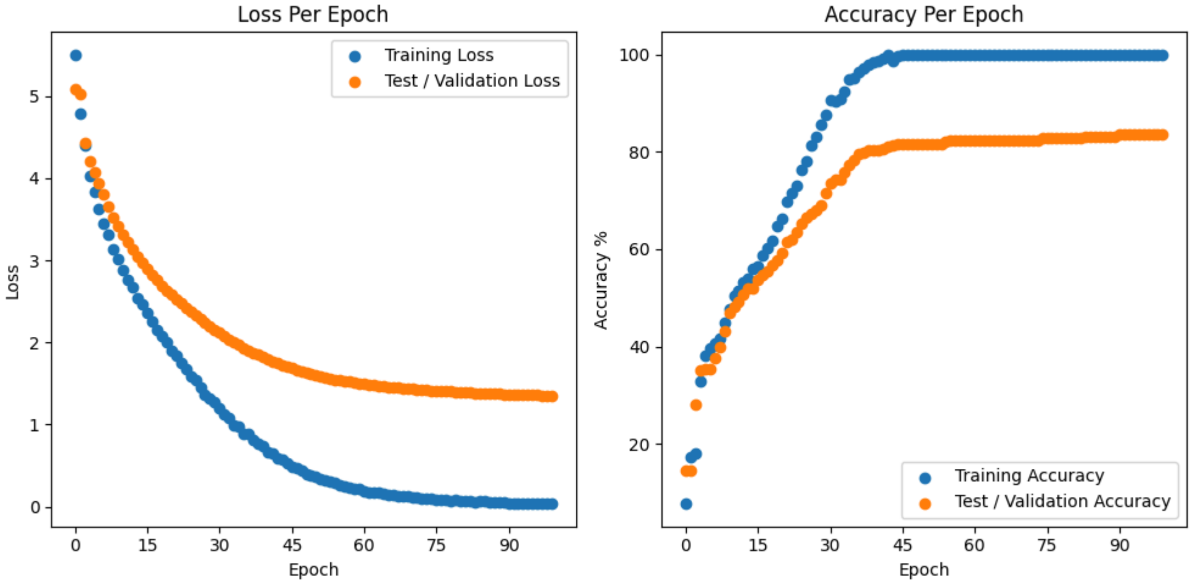


Last Best Epoch: 67, Train Time ≈ 0 hr, 0 min, 16 sec
Train Loss: 0.13697, Train Accuracy: 100.00%
Test Loss: 1.39452, Test Accuracy: 83.58%

Example Output:

- Input: We cook dinner on Sundays EOS
- Output: Nous cuisinons le dîner le cuisinons EOS
- Expected: Nous cuisinons le dîner le dimanche EOS

Transformer Model: 2 Layers; 4 Heads

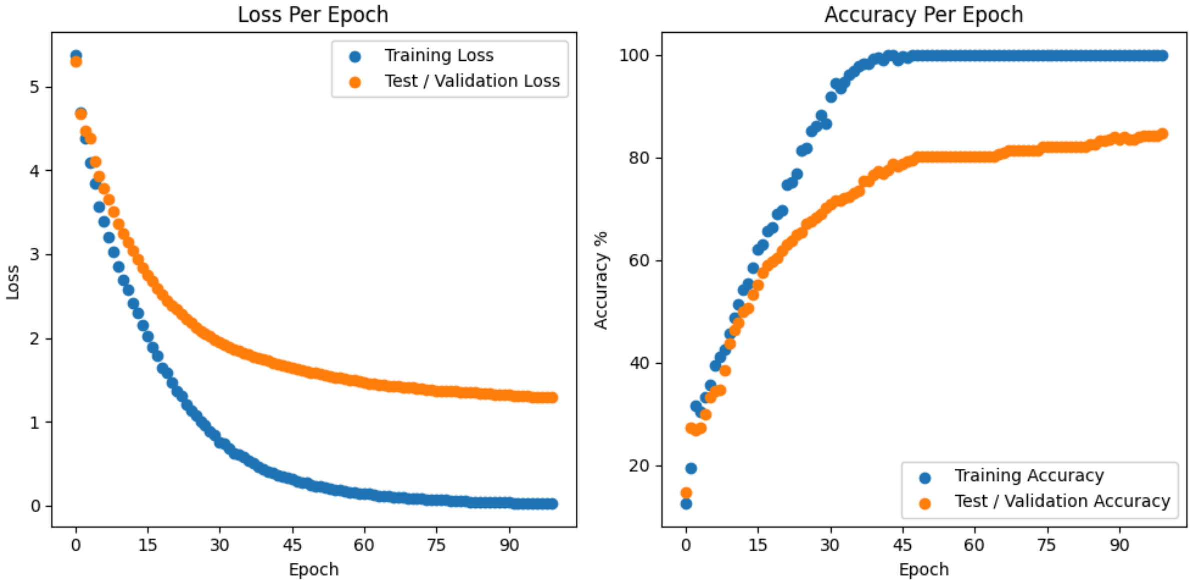


Last Best Epoch: 90, Train Time ≈ 0 hr, 0 min, 17 sec
Train Loss: 0.04759, Train Accuracy: 100.00%
Test Loss: 1.36885, Test Accuracy: 83.58%

Example Output:

- Input: We cook dinner on Sundays EOS
- Output: Nous cuisinons le dîner le dîner EOS
- Expected: Nous cuisinons le dîner le dimanche EOS

Transformer Model: 4 Layers; 2 Heads

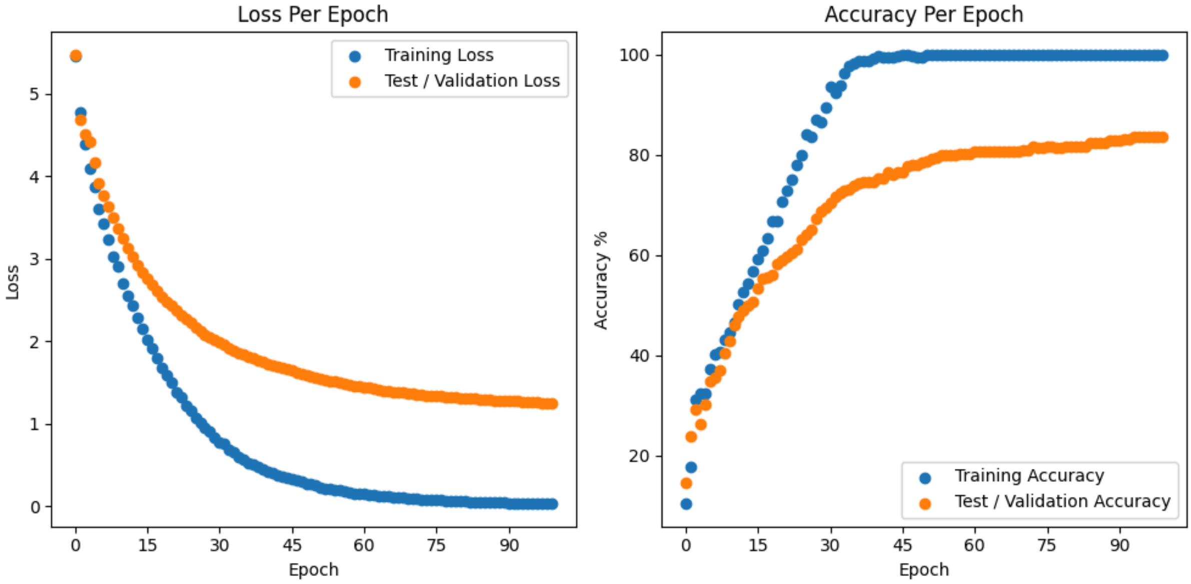


Last Best Epoch: 99, Train Time ≈ 0 hr, 0 min, 33 sec
Train Loss: 0.02387, Train Accuracy: 100.00%
Test Loss: 1.29102, Test Accuracy: 84.70%

Example Output:

- Input: We cook dinner on Sundays EOS
- Output: Nous cuisinons le dîner le dîner EOS
- Expected: Nous cuisinons le dîner le dimanche EOS

Transformer Model: 4 Layers; 4 Heads

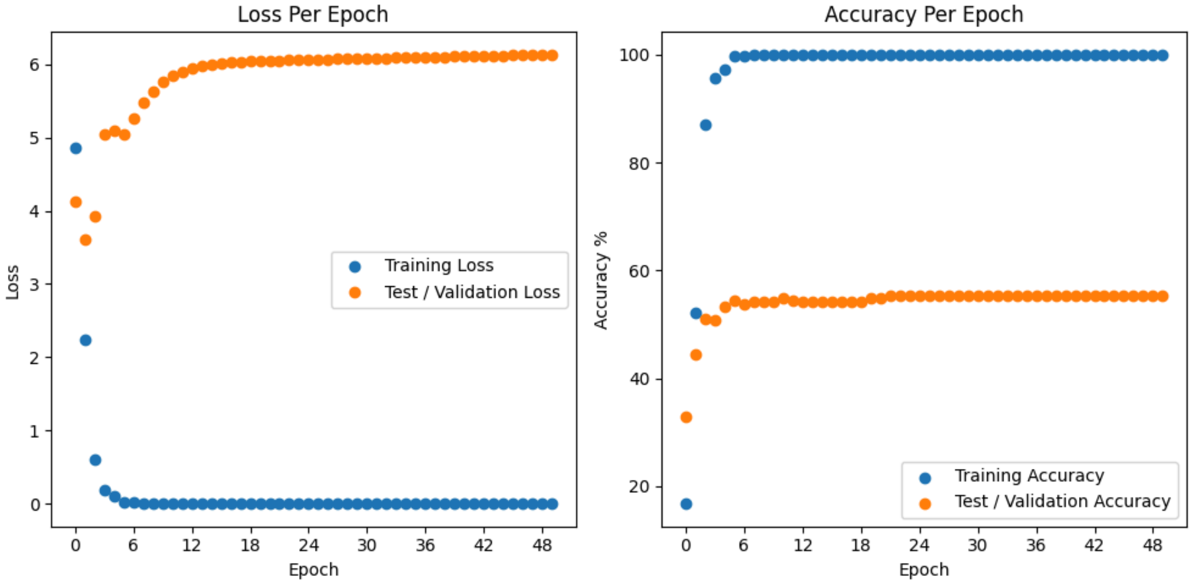


Last Best Epoch: 93, Train Time ≈ 0 hr, 0 min, 33 sec
Train Loss: 0.03137, Train Accuracy: 100.00%
Test Loss: 1.26685, Test Accuracy: 83.58%

Example Output:

- Input: We cook dinner on Sundays EOS
- Output: Nous cuisinons le dîner le dîner EOS
- Expected: Nous cuisinons le dîner le dimanche EOS

nn.GRU



Last Best Epoch: 21, Train Time ≈ 0 hr, 0 min, 17 sec
Train Loss: 0.00050, Train Accuracy: 100.00%
Test Loss: 6.05060, Test Accuracy: 55.22%

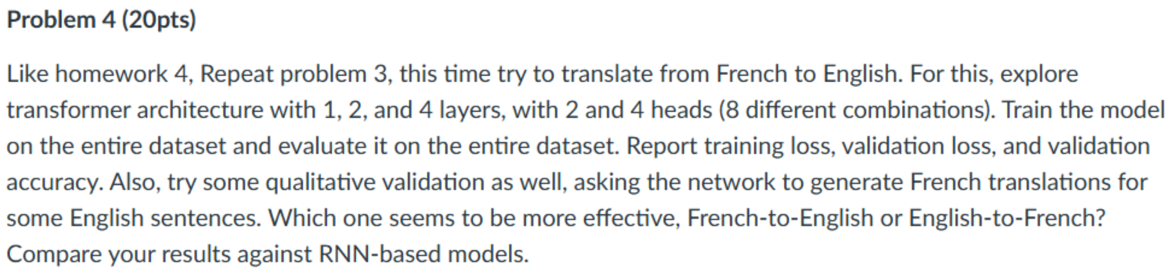
Example Output:

- Input: He plays soccer with friends EOS
- Output: Il joue de la guitare EOS
- Expected: Il joue au football avec des amis EOS

Observations

Transformer models improved performance by around 28% over their RNN counterparts. Varying the transformer complexity had minimal impact on performance.

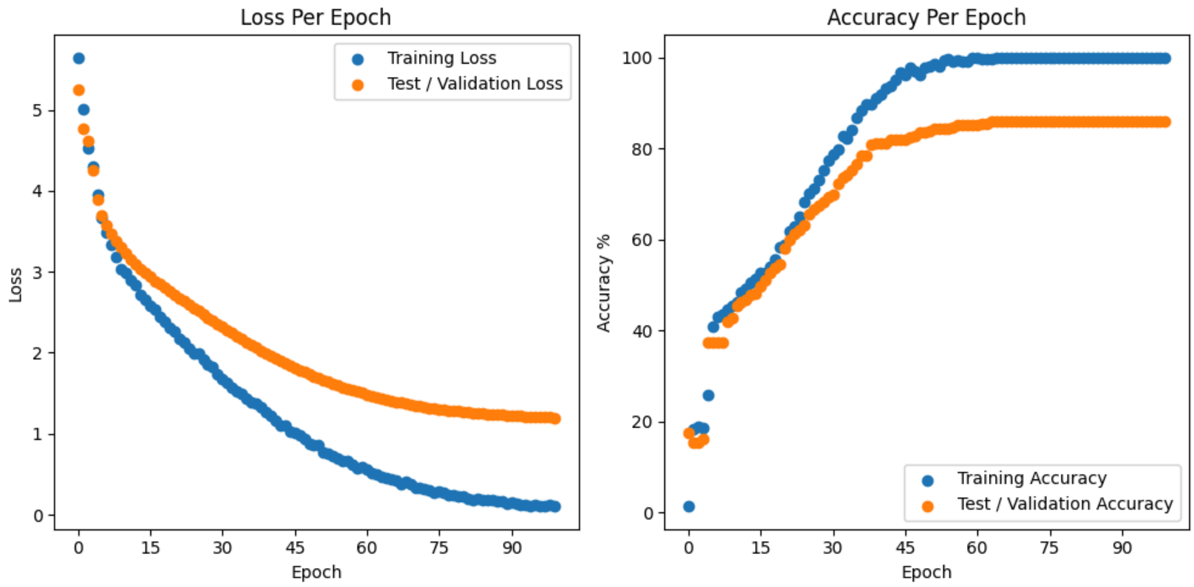
Problem 4



Problem 4 (20pts)

Like homework 4, Repeat problem 3, this time try to translate from French to English. For this, explore transformer architecture with 1, 2, and 4 layers, with 2 and 4 heads (8 different combinations). Train the model on the entire dataset and evaluate it on the entire dataset. Report training loss, validation loss, and validation accuracy. Also, try some qualitative validation as well, asking the network to generate French translations for some English sentences. Which one seems to be more effective, French-to-English or English-to-French? Compare your results against RNN-based models.

Transformer Model: 1 Layer; 2 Heads

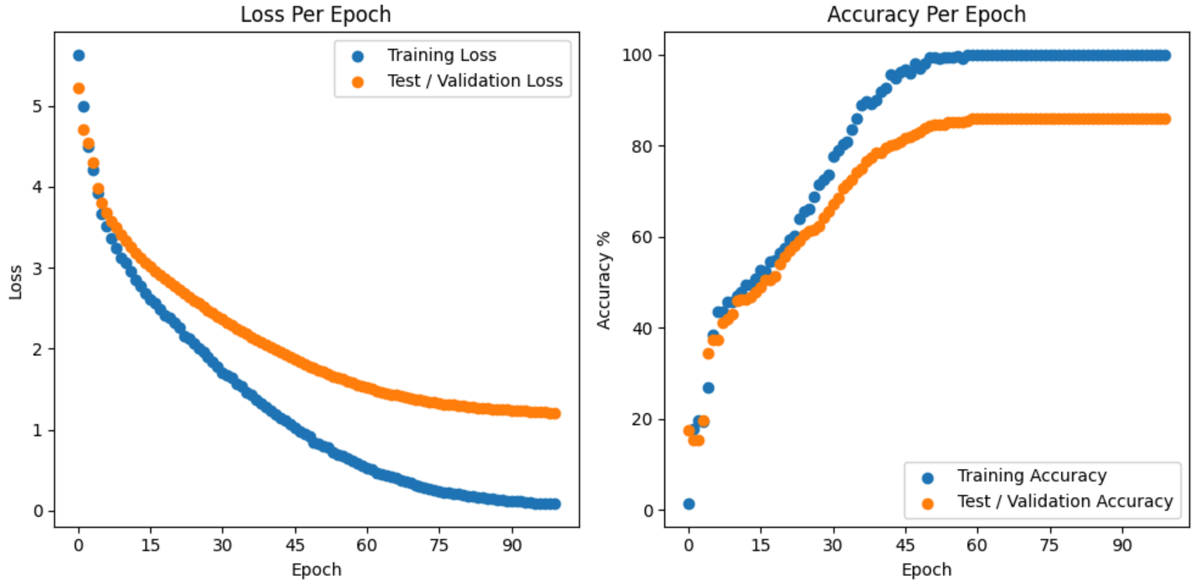


Last Best Epoch: 63, Train Time ≈ 0 hr, 0 min, 52 sec
Train Loss: 0.46505, Train Accuracy: 99.67%
Test Loss: 1.43617, Test Accuracy: 85.88%

Example Output:

- Input: Nous cuisinons le dîner le dimanche EOS
- Output: We cook dinner on on EOS
- Expected: We cook dinner on Sundays EOS

Transformer Model: 1 Layer; 4 Heads

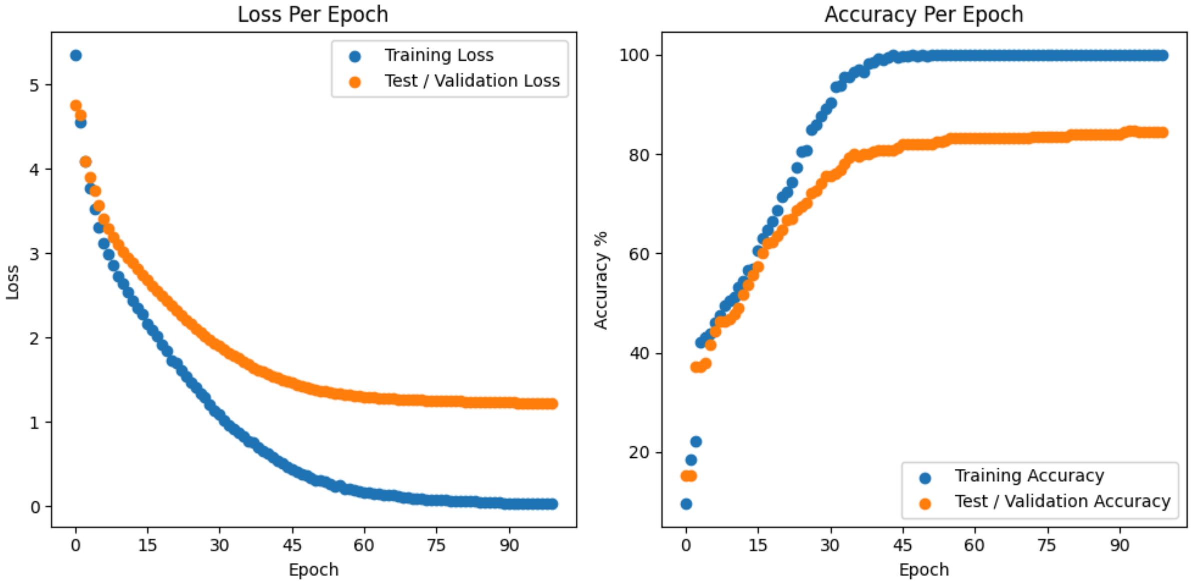


Last Best Epoch: 59, Train Time ≈ 0 hr, 0 min, 15 sec
Train Loss: 0.55776, Train Accuracy: 99.84%
Test Loss: 1.53814, Test Accuracy: 85.88%

Example Output:

- Input: Nous cuisinons le dîner le dimanche EOS
- Output: We cook dinner on on EOS
- Expected: We cook dinner on Sundays EOS

Transformer Model: 2 Layers; 2 Heads

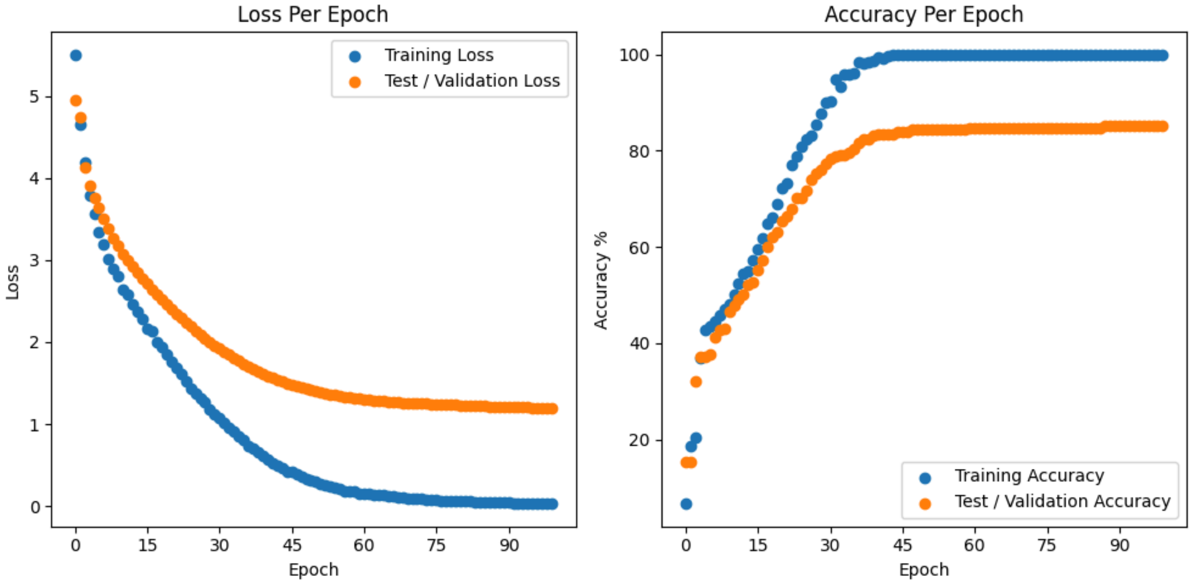


Last Best Epoch: 92, Train Time ≈ 0 hr, 0 min, 34 sec
Train Loss: 0.03998, Train Accuracy: 100.00%
Test Loss: 1.22951, Test Accuracy: 84.71%

Example Output:

- Input: Nous cuisinons le dîner le dimanche EOS
- Output: We cook dinner on dinner EOS
- Expected: We cook dinner on Sundays EOS

Transformer Model: 2 Layers; 4 Heads

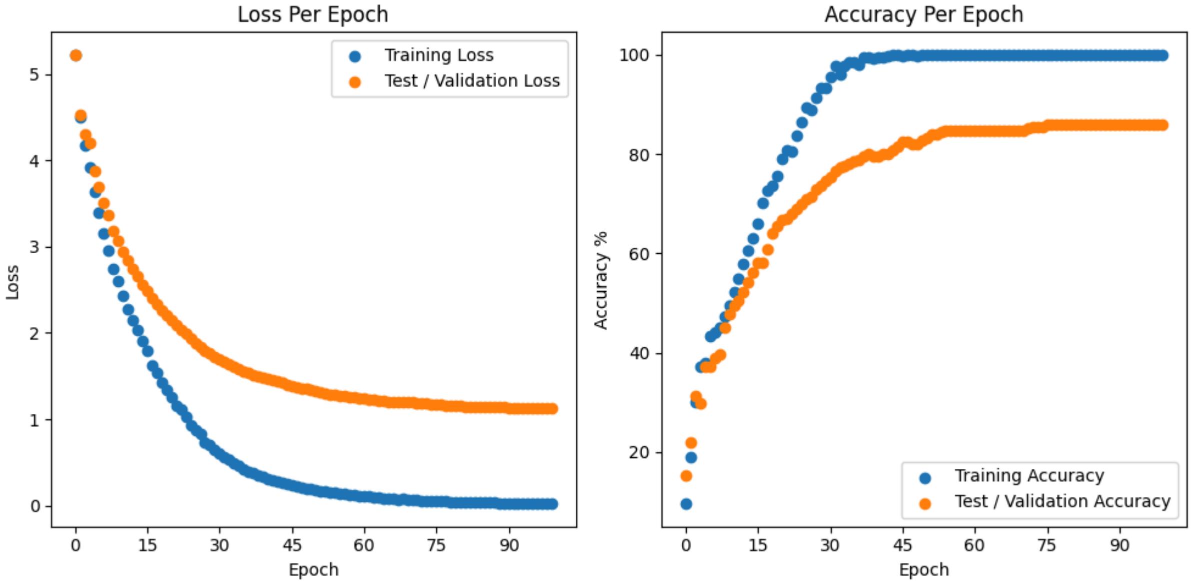


Last Best Epoch: 87, Train Time ≈ 0 hr, 0 min, 16 sec
Train Loss: 0.04443, Train Accuracy: 100.00%
Test Loss: 1.21285, Test Accuracy: 85.10%

Example Output:

- Input: Nous cuisinons le dîner le dimanche EOS
- Output: We cook dinner on dinner EOS
- Expected: We cook dinner on Sundays EOS

Transformer Model: 4 Layers; 2 Heads

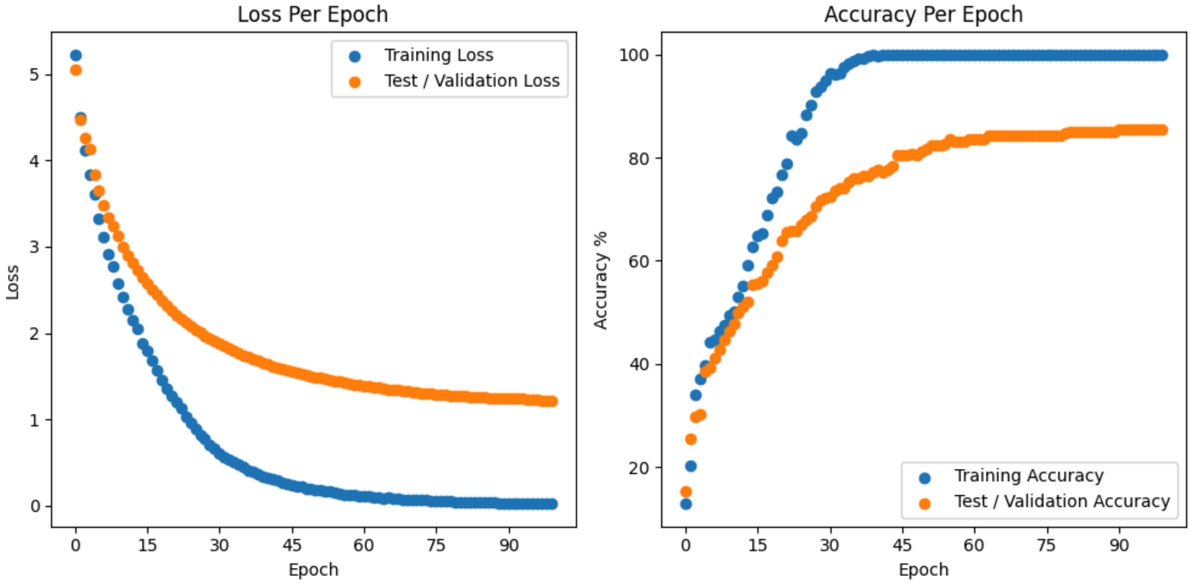


Last Best Epoch: 75, Train Time ≈ 0 hr, 0 min, 53 sec
Train Loss: 0.05009, Train Accuracy: 100.00%
Test Loss: 1.16915, Test Accuracy: 85.88%

Example Output:

- Input: Nous cuisinons le dîner le dimanche EOS
- Output: We cook dinner on dinner EOS
- Expected: We cook dinner on Sundays EOS

Transformer Model: 4 Layers; 4 Heads

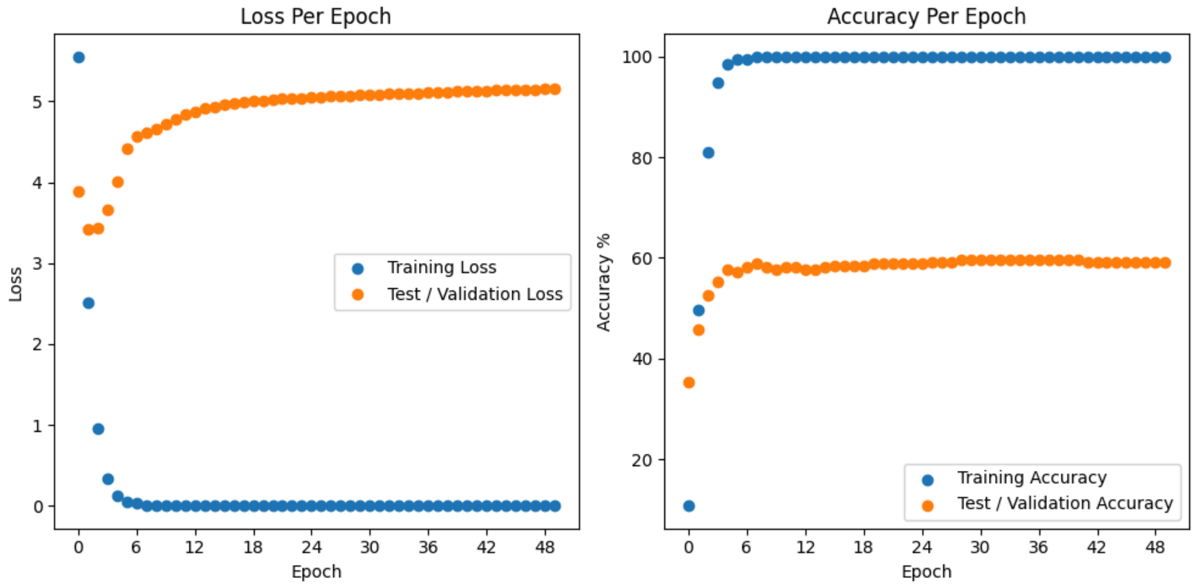


Last Best Epoch: 90, Train Time ≈ 0 hr, 0 min, 31 sec
Train Loss: 0.02739, Train Accuracy: 100.00%
Test Loss: 1.23543, Test Accuracy: 85.49%

Example Output:

- Input: Nous cuisinons le dîner le dimanche EOS
- Output: We cook dinner on dinner EOS
- Expected: We cook dinner on Sundays EOS

nn.GRU



Last Best Epoch: 28, Train Time ≈ 0 hr, 0 min, 25 sec
Train Loss: 0.00044, Train Accuracy: 100.00%
Test Loss: 5.06931, Test Accuracy: 59.61%

Example Output:

- Input: Nous aimons la musique française EOS
- Output: We love music EOS
- Expected: We love French music EOS

Observations

Transformer models improved performance by around 24% over their RNN counterparts. Varying the transformer complexity had minimal impact on performance.

Translating French to English has around 1.5% additional performance over English to French for transformer models.