Assignment 4

Your name: Willie Zeng

Your GTID: wzeng40@gatech.edu

Seq2Seq Results

- Put your results from training before and after hyperparameter tuning here.
- Explain what you did here as well, you can use another slide if needed

Using Default Parameters (BEFORE)
EPOCH 10:
Training Loss: 4.6025. Validation Loss: 4.5244.
Training Perplexity: 99.7379. Validation Perplexity: 92.2412.

Using Tweaked Parameters (AFTER)
EPOCH 12:
Training Loss: 3.4325. Validation Loss: 3.5224.
Training Perplexity: 79.7379. Validation Perplexity: 78.9712

- I increased the epochs to 12, increased the learning rate (0.002) because I feel like the model converged too slow. I changed the model to LSTM because I was curious of the affects of keeping track of the “Cell” tensor (in addition to the hidden tensor). Furthermore, I lowered the batch size (64) because it lacks the ability to generalize; large batch models converges to sharp minimizers (https://arxiv.org/abs/1609.04836)

Transformer Results

- Put your results from training before and after hyperparameter tuning here.
- Explain what you did here as well, you can use another slide if needed

Using Default Parameters (BEFORE)
EPOCH 10:
Training Loss: 2.2988. Validation Loss: 2.9797.
Training Perplexity: 9.9625. Validation Perplexity: 19.6814

Using Tweaked Parameters (AFTER)
Epoch 8
Training Loss: 2.5345. Validation Loss: 2.9388.
Training Perplexity: 12.6097. Validation Perplexity: 18.8925.

- The Transformer results are better than seq2seq. One thing that the transformer does better is that the encoder and decoder sees the entire inputs at once by using multi-head attention.
- The results show that the model was overfitting. This is because after a certain epoch, the training loss & perplexity kept decreasing, however, the validity loss & perplexity stayed the same and occasionally increased. In response to this, I lowered the **learning rate** (0.0009) and **decreased the epoch to 8**. Furthermore, I lowered batch size (64) because it lacks the ability to generalize; large batch models converges to sharp minimizers (https://arxiv.org/abs/1609.04836)