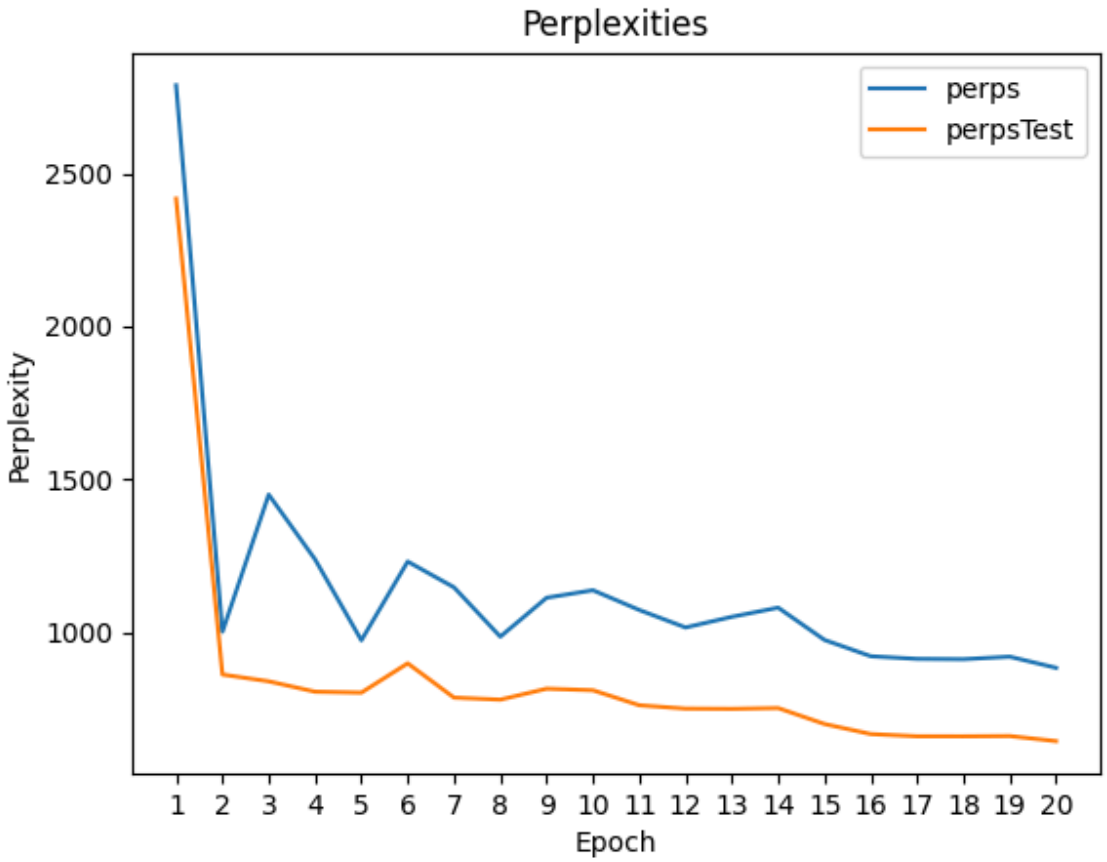
Homework 1 write up
Joseph Blom and Grayson Snyder

1. Modifying the starter code to work as a transformer only model was both easy in some ways and very challenging in others. Removing the cross attention and the encoder layer was relatively straightforward.
   Originally, we used a TokenDataset and DataLoader as our method of feeding batches into the model. We were having major issues with training speed and suspected that our data loading method was part of the problem. So, we changed it to a custom data loader defined in line 323. Beyond this, the setup is standard, with a loop for epochs and a loop to iterate over the data and train the model.
   Big shout out to Harrison Bounds for his assistance and code to reference in correcting our model.

2. Below are our results from training for 20 epochs. We ran our script using “python3 starter.py” and we generally did not change the default hyper parameters. We did change the learning rate to “0.00003” but this should probably be tweaked even more. Our final Perplexities were around 850 training, 600 tests. Included in the zip are the weights.



3. Rate: 105,595 TPS

I think given that one epoch on wikitext103 only took around 10 minutes we should be able to train for the next homework without issue overnight. Our most significant improvements on speed have come from changing our data transformer and loader so if we need more performance that is the first thing we will try to refine.

4. Below are our results from training for 20 epochs. We ran our script using “python3 starter.py”. For this iteration, the batch size had to be decreased to 8 to run successfully due to the euclidean distance calculations taking a lot of space. Additionally, the learning rate was lowered to 0.00001 to experiment with a slower rate. The remaining hyperparameters were the same. Our final Perplexities were around 625 training, 450 tests. This method took much longer to run than using the standard attention model and would be even worse for larger datasets such as wikitext103, so it is likely we will continue the work from part 2 in future assignments instead.