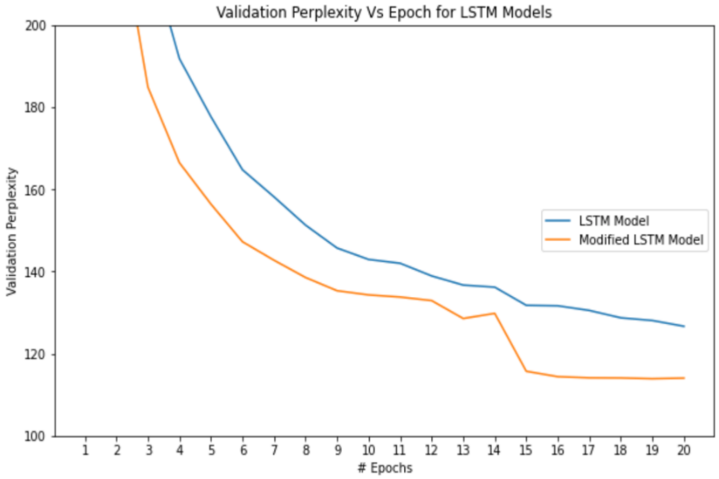
Kristy Lee
CS288 Project 1 Report

The improvement I made to my LSTM model included adding a learning rate scheduler inside the train function of the model that would decrease the learning rate after 14 epochs, which is when the validation perplexity of the LSTM model approaches the 130s. Originally, the Adam optimizer's parameter groups had a default learning rate of 0.001, but after 14 epochs I decrease the learning rate to 0.0001. I also replaced the last linear layers of the unmodified LSTM model to be the following series of linear layers in the modified LSTM model: linear layer projecting input of size 512 to output of size 216, a linear layer projecting size 216 to size 128, a linear layer projecting size 128 to vocab_size.

The motivation behind using a smaller learning rate during the last training epochs is to have the model converge to a more optimal set of weights. I use a larger, default learning rate initially for training and observing patterns in text, and then decrease the learning rate once the decrease in perplexity has become stable so that the model isn't focused on learning new patterns of data but instead on converging to an optimal set of weights, which works to decrease perplexity appropriately due to optimality weights and of the outputted logits. The use of 3 linear layers instead of 2 was for stylistic purposes since more hidden layers can be used to learn more features, which can improve accuracy.

The modifications were effective. See the graph below:



After # Epochs [1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20]
Corresponding Validation Perplexities for Unmodified LSTM = [453.4776967238335, 275.52673874811387, 218.86994837933827, 191.77715281120956, 177.64086208456428, 164.72814133185773, 158.13904962522918, 151.25879691149336, 145.6856994327846, 142.90372706181878, 141.9750297491395, 138.89187343110947, 136.66953385355254, 136.16349725154072, 131.74431426993357, 131.6262423794258, 130.49292044940927, 128.6966587884459, 128.041226166693, 126.64037174756169]
Corresponding Validation Perplexities for modified LSTM = [343.375079907425, 229.92692720027924, 184.85568691095145, 166.44385083791815, 156.3707099153785, 147.22088254707143, 142.71207151311998, 138.5135838977063, 135.29676811341025, 134.2746138862812, 133.75218319926648, 132.8895621922218, 128.5346552429466, 129.78749278635334, 115.6959497201583, 114.37626766488985, 114.09606182530938, 114.06036198831498, 113.88000188350168, 114.02266658767375]

Observe that for both the modified LSTM and unmodified LSTM, the perplexity mostly decreased across all epochs, and that modified LSTM consistently had lower perplexity values than the unmodified LSTM across all epochs. This shows that modified LSTM predicts validation dataset text better than the unmodified LSTM throughout training, demonstrating the effectiveness of the model architecture.
My modified LSTM model reached a final validation perplexity of 113.88000188350168, while my non-modified LSTM model reached a final validation perplexity of 126.64037174756169 after 20 epochs. We can see that the perplexity of the modified LSTM model was <120, and this reflects the improvement in using the modified LSTM model to predict text compared to the original LSTM: the text has a higher probability of being predicted as such with modified LSTM model, which results in a lower perplexity.