Report for Homework 1b: Language Model

1 Extension Methods

The training process shows that the model overfitted after dozens of epochs, the average training loss still declining but the validation loss becomes stable or even worse, i.e., the model works much worse on the validation set than it on the training set.

Several reasons cause this problem. This experiment dataset is not so large that the model is complicated enough to remember all samples that existed during the training process. The model overfitted these samples causing the lower robustness and this problem.

In a small dataset, regularization and dropout can always work for the overfitting problem. To mitigate the overfitting problem, I applied several methods including:

- **Batch Size Adjustment.** Training with a larger batch size usually calculates a much stabler gradient. The training curve will be more smooth. Larger batch sizes can improve the robustness and generalization of the model.
- **L2 Regularization.** Overfitting is usually caused by too many parameters or a too complicated network. When we add the L2 penalty to the loss function, too complicated network or parameters will enlarge the loss. Thus, with this modified loss function, the optimizer tends to update parameters with the aim of being more streamlined and more accurate.
- **Embedding Dropout.** When the network has too many parameters and the training dataset is small, it may overfitted to the data. The core idea of embedding dropout is to streamline the network to avoid overfitting.

  Also, I have tried to modify the network structure, including raising the number of layers and the output size of LSTM. It does not work, it is because the base network is already complicated enough to fit the dataset, more parameters will not help much. The key problem here is overfitting.

2 Results

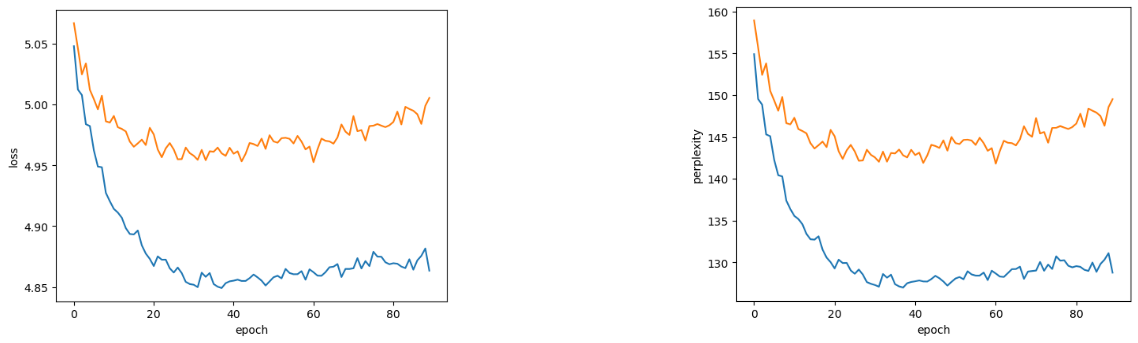

Figure 1: Validation losses and perplexities, the blue line indicates the model training with a better algorithm, data comes from the latter 90 epochs.

The loss and perplexity changing trend is similar during the training process in the base model and the modified model. With the regularization methods, the optimizer works better than before, we reach at least ten values reducing the perplexity.