NLU project exercise lab: 9

Massimo Stefan (240998)

University of Trento

massimo.stefan@studenti.unitn.it

1. Introduction

In this report, I detail the iterative improvements made to a baseline language model. Starting with the replacement of an RNN with an LSTM, I introduced dropout layers, experimented with the AdamW optimizer, and further incorporated advanced regularization techniques. The goal of these enhancements was to reduce the perplexity (PPL) metric, indicating better performance of the language model.

2. Implementation details

2.1. Step 1: LSTM Introduction and Dropout

- **LSTM Replacement**: I substituted the RNN in the baseline model with an LSTM to better capture long-term dependencies. This modification led to an immediate reduction in PPL.
- **Dropout Addition**: To mitigate overfitting, I integrated two dropout layers - post-embedding and pre-output. These layers further aided in decreasing PPL.
- **Optimizer Experimentation**: Transitioning from SGD, I adopted the AdamW optimizer. This change was motivated by AdamW's potential for improved convergence, which manifested in our results with a further decline in PPL.

2.2. Step 2: Advanced Regularization

- **Variational Dropout** : I adopted Variational Dropout, ensuring consistent dropout masks across time steps. While the advantages of this method can be context-dependent, in our scenario, it contributed to PPL reduction.
- **Weight Tying**: By tying weights between the embedding and output layers, I reduced the model's trainable parameters. This not only ensures computational efficiency but also aids in better generalization.

2.3. Step 3: Stochastic Weight Averaging (SWA)

- **SWA Integration**: To potentially find a broader minimum in the loss landscape, I incorporated SWA, which averages model weights over multiple epochs.
- **Non-monotonic Trigger**: An adaptive approach was employed to start SWA. The trigger activates when validation loss plateaus, ensuring we leverage model averaging at optimal epochs. This strategy further optimized the PPL metric.

3. Results

In the first part of the exercise, I evaluated the model using two main functions: train_loop and eval_loop.

The train_loop function is responsible for training the model on the given data. Inside this function, I initialized the model to training mode and iterated through the training data. For each batch, I zeroed the gradients, computed the model's output, and calculated the loss using the given criterion. I then performed backpropagation and applied gradient clipping to prevent exploding gradients. Finally, I updated the model's weights using the specified optimizer.

The eval_loop function is used to evaluate the model on a validation or test set. Inside this function, I set the model to evaluation mode and iterated through the evaluation data without computing gradients. I computed the model's output and loss for each batch, and then calculated the Perplexity (PPL) by taking the exponential of the mean loss.

For early stopping, I implemented a patience mechanism within the main training loop. If the validation PPL did not improve for a specified number of epochs (controlled by the patience variable), the training was terminated early.

In the second part of the exercise, I made some changes to the evaluation process to incorporate Stochastic Weight Averaging (SWA). I introduced a new function train_and_evaluate_avg, which includes the initialization of the SWA model and scheduler. Inside the main training loop, I added logic to update the SWA model's parameters and adjust the learning rate using the SWA scheduler if the validation loss had not improved for a specified number of epochs (controlled by the non_monotonic_trigger variable). This allowed the model to benefit from the averaging of weights over multiple epochs, potentially leading to better generalization and performance on the evaluation set.

In summary, I achieved the expected PPL below 250 for each model, except for LM_LSTM_Dropout_SWA. This model obtained an evaluation PPL ranging from 230 to 187 during the training and evaluation phase, but yielded 10007 at test time for an unknown reason.

Model	PPL Results
LM_LSTM_SGD	202.72
LM_LSTM_Dropout_SGD	189.39
LM_LSTM_Dropout_AdamW	178.59
LM_LSTM_Dropout_SWA	10007.33
LM_LSTM_Weight_Tying	206.39.0
LM_LSTM_Var_Dropout	178.04

Table 1: *Model Performance in terms of Perplexity (PPL) on the validation set*