NLU project exercise lab: 10

Thomas Trevisan (240458)

University of Trento

thomas.trevisan@studenti.unitn.it

1. Introduction

In the first part:

- the RNN was switched with an LSTM
- dropout layers on the embedding, the hidden layers and the output layer were applied
- The optimizer was changed from SGD to AdamW

In the second part, according to [1], the following methods were added:

- Weight tying to bind weights in the embedding layer with the ones in the output layer
- A class that implements variational dropout.
- A non-monotonically trigger to switch from standard SGD to Average SGD

2. Implementation details

As shown in [1], the weight tying aims at reducing the number of parameters to train, and at creating a connection between the initial embedding and the final embedding of the words. This method was implemented in the __init__ function of the model. The embedding and hidden size then must be of the same size, for which 500 was used. Secondly, the variational dropout was used in order to create a mask that:

- is different for every sample in the batch
- is the same for every embedding given as input in a single forward pass of the LSTM

This was implemented throughout a class that gets called at every forward of the LSTM model, generating the dropout mask. The variational dropout was employed both on the embeddings and on the LSTM output. As in [1], a standard dropout was applied to the hidden layers of the LSTM. Lastly the non-monotonical trigger for the ASGD was implemented by switching to ASGD when the following conditions applied:

- SGD is the current optimizer
- The parameter of the optimizer was *NT-ASGD*
- N steps have already been done, with N=5 according to [1]
- The current loss is greater than the minimum of the last N1 losses (The loss is starting to increase)

The check for these conditions was implemented after every validation step, because the logging interval used was 1. The model was trained on 100 epochs and the best one was taken to perform the final testing.

3. Results

The model was trained incrementally, to see the contributions of every method for the increase or decrease in performances. We can see a drop in PPL when using variational dropout. The probabilities are kept the same as in the experiment with weight tying alone, and are:

- emb_dropout = 0.1
- out_dropout = 0.4
- hidden_dropout = 0.3

It is probable that probabilities have to be tuned differently based on the dropout technique. The Non monotonically triggered SGD further worsen the model, probably due to ASGD parameter settings.

Table 1: *Results for part 1*

	Perplexity
LSTM	225.34
LSTM+Dropout	204.39
LSTM+Dropout+AdamW	166.90

Table 2: *Results for part 2*

	Perplexity
LSTM+Weigth Tying	193.51
LSTM+WeightTying+VariationalDropout	195.86
LSTM+WeightTying+VariationalDropout+NT-ASGD	201.24

4. References

[1] S. Merity, N. S. Keskar, and R. Socher, "Regularizing and optimizing lstm language models," 2017.