NLU course project

Jie Chen (mat. 256177)

University of Trento

jie.chen-2@studenti.unitn.it

1. Introduction (approx. 100 words)

The objective of this project is to improve the model performance by modifying the baseline RNN with a set of regularization techniques. First, the standard RNN language model is replaced with the Long Short-Term Memory (LSTM) network, this allows to better capture long-range dependencies. Then, dropout layers are added trying to prevent overfitting and SGD optimizer is replaced with AdamW to enhance convergence. From the LSTM model, additional regularization techniques including weight tying, variational dropout, and Non-monotonically Triggered Averaged SGD (AvSGD) are applied. The goal is to achieve perplexity below 250 on the validation set. The dataset used is the Penn Treebank dataset.

2. Implementation details (max approx. 200-300 words)

In the first part of the assignment, I replaced the RNN with an LSTM network and applied two dropout layers, one after the embedding layer and one before the last linear layer. The dropout layer randomly sets a fraction of neurons' outputs to zero, in the layers where it is applied, to prevent overfitting. To enhance the model performance, I experimented with different hyperparameters settings hoping to achieve the best PPL. The hyperparameters I tried to fine-tune include: hidden size, embedding size, learning rate, number of epochs, output dropout, and embedding dropout. Although many other parameters could be changed, these are the main ones I played with to see if modifying them would improve performance. The best setting was then kept for subsequent modifications. The SGD optimizer was replaced with AdamW to improve training speed and stability, and the learning rate was adjusted.

In the second part of the project, few additional techniques were introduced and applied as suggested by the paper [1]. Weight tying shares the weights between the embedding layer and the output (softmax) layer. It reduces the memory usage for large vocabularies, improving generalization capability and allowing for better performance. This technique requires the hidden layer dimensions to match the embedding layer dimension. Standard dropout randomly zeros out neurons independently at every timestep, while variational dropout maintains the same dropout mask across all timesteps for a given sequence. The same neurons are dropped at every timestep, so the network sees a consistent "subnetwork" throughout the sequence. Finally, the Non-monotonically Triggered Average SGD (NT-ASGD) optimizer is applied to provide more stable convergence in the critical phase of training, as it averages the parameters over several steps to improve model generalization. This regularization technique is triggered by monitoring the validation loss: it tracks the best validation loss so far and if the validation loss does not improve for n non-monotonic steps, averaging is triggered. This ensures averaging occurs only when normal SGD stops improving, adapting to the actual training dynamics.

3. Results

First part. The table 1 below shows all the parameter settings I have experimented with and the results I obtained. I started experimenting with different RNN model sizes and concluded that larger models achieve better PPL. This can be motivated by the fact that bigger embeddings capture richer lexical information and a larger hidden state models longer/complex dependencies better, provided that enough data is available. For the different size models, I also tried different learning rates. After replacing the model with LSTM, I observed that a smaller model with SGD learning rate of 2 achieve lower PPL. Specifically, replacing RNN with LSTM reduced the PPL from 157.09 to 142.01. By applying dropout layers with different emb_dropout and out_dropout values, I further reduced the PPL to 141.98. Finally, using AdamW optimizer with different weight decay values led to the best PPL of 114.33.

Second part. The table 2 below shows the result I got with different learning rate settings, using the same hid_size and emb_size. Starting from LSTM with weight tying, the best result is achieved with a learning rate of 5 and PPL of 103.28. By applying additional variational dropout regularization technique I managed to reduce the PPL to 100.03, but this time with a even higher leaning rate. After several trials incorporating non-monotonically triggered averaged SGD, I obtained the best PPL of 97.64,with optimized dropout and learning rate values.

The table 3 below reports the best result obtained and the figure 1 and 2 shows the evolution of the loss and perplexity during training for the best configuration.

4. References

[1] S. Merity, N. S. Keskar, and R. Socher, "Regularizing and optimizing lstm language models," *arXiv preprint arXiv:1708.02182*, 2017. [Online]. Available: https://arxiv.org/abs/1708.02182

Model	hid_size	emb_size	lr	PPL	emb_dropout	out_dropout	weight_decay
RNN	200	300	2	166.30	-	-	-
RNN	200	300	1	160.12	-	-	-
RNN	300	500	2	165.89	-	-	-
RNN	300	500	1	157.09	-	-	-
LSTM	300	500	1	145.63	-	-	-
LSTM	200	300	1	145.98	-	-	-
LSTM	200	300	2	142.01	-	-	-
LSTM_DROPOUT	200	300	1	143.03	0.1	0.1	-
LSTM_DROPOUT	200	300	2	141.98	0.2	0.1	-
LSTM_AdamW	200	300	0.001	123.28	0.1	0.1	0.05
LSTM_AdamW	200	300	0.001	122.99	0.2	0.1	0.05
LSTM_AdamW	200	300	0.001	114.33	0.1	0.1	0.1

Table 1: *Results of different model configurations and hyperparameter settings.*

Method	emb_dropout	out_dropout	lr	PPL
LSTM wt	0.5	0.1	3	108.28
LSTM wt	0.5	0.1	5	103.28
LSTM wt + vd	0.5	0.1	5	103.61
LSTM wt + vd	0.5	0.1	7	100.03
LSTM wt + vd + avsgd	0.5	0.1	2	105.79
LSTM wt + vd + avsgd	0.5	0.1	2.8	103.95
LSTM wt + vd + avsgd	0.6	0.2	4	97.64

Table 2: *Results of modified models with weight tying (wt), variational dropout (vd), and ASGD optimizer, showing embedding/output dropout, learning rate (lr), and perplexity (PPL).*

Model	PPL
RNN	157.09
LSTM	142.01
+ Dropout	141.98
+ AdamW	**114.33**
wt	103.28
wt + vd	100.03
wt + vd + AvSGD	**97.64**

Table 3: *Perplexity of best models.*

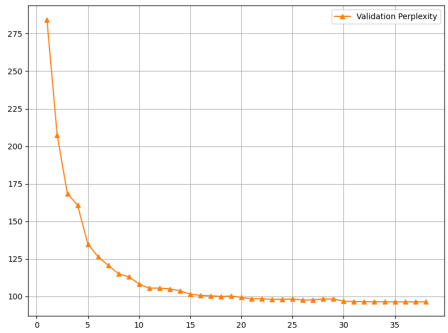

Figure 1: *Perplexity loss of LSTM + weight tying + variational dropout + non monotonically triggered AvSGD*

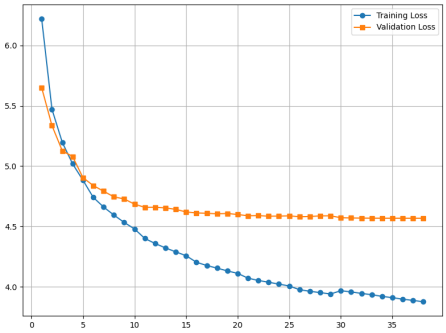

Figure 2: *Training and validation loss of LSTM + weight tying + variational dropout + non monotonically triggered AvSGD*