NLU course projects: LM

Rossana Cervesato (mat. 231499)

University of Trento

rossana.cervesato@studenti.unitn.it

1. Introduction

This is the final report for the Natural Language Understanding course, lab 4 (A.Y.2023/2024). The goal of the project is to implement language model based on LSTMs and to improve its performance on the PennTreebank dataset [1] through some regularization techniques to achieve a perplexity score ≤ 250.
This project includes:

- the analysis of the dataset (Table 1, Table 2);
- the implementation and the training of the baselineLSTM model and the proposed model;
- the proposed model's fine-tuning and the application of different optimization and regularization techniques [2];
- the evaluation and the analysis of the models' behaviour.

The best model achieved a test perplexity of 99.58. The code is available at https://github.com/rossana24/NLU2024

2. Implementation details

2.1. Architecture

LSTMs [3] were used as the backbone architecture for both implemented models. The baseline and the proposed networks are 2-layer LSTM with an embedding size of 300, a hidden size of 200 and two additional dropout layers (one after the embedding layer and one before the last linear layer). These architectures are also designed to handle sequence data and can process variable-length inputs. To prevent overfitting, early stopping was implemented with a patience of 10. Additionally, weight decay with a coefficient of $1.2e^{-6}$ was applied to penalize large weights, aiding in better generalization on unseen data.

2.2. Regularization

Sequence models are prone to overfitting, as they are trained on long sequences and they have to capture long-term dependencies. In this project, several regularization methods were implemented and tested in the proposed model training process:

- *Variational Dropout* [2]: it ensures that the same dropout mask generated using a Bernoulli distribution is applied across all time steps of a sequence;
- *Tied weights* [2]: it reduces the the total number of parameters by sharing parameters between the input and output embeddings;
- *Gradient Clipping*: a clipping threshold of 5 was used to prevent the exploding gradient problem.

2.3. Optimization

To efficiently minimize the loss function and update model parameters, attention was given to both the choice of optimizer and the learning rate. For the baseline model, AdamW was used as the primary optimizer with a learning rate set to 0.001.
In contrast, the proposed model implemented the Non-monotonically Triggered Averaged Stochastic Gradient Descent (AvSGD) optimization strategy [2]. This strategy involves switching from SGD to AvSGD if the recent validation perplexity worsens compared to past values. This approach aims to enhance generalization and stability during the latter stages of training. Some experiments were performed with AdamW without any significant improvement in the model performance. Additionally, a learning rate scheduler was employed to further refine the training process. The scheduler was configured to apply an exponential decay to the learning rate, starting from 1.0 with a gamma value of 0.5.
Finally, the proposed model was trained using both normal and truncated back-propagation through time (TBPTT) (1) [2] to manage long sequences effectively without running into memory constraints.

2.4. Metrics and Evaluation

The language model problem can be framed as a multi-class classification problem, where the classes to predict are the words in the vocabulary. The model is thus trained to minimize the Cross-Entropy Loss (2).
The main metric used to compare and evaluate the models is the perplexity (PP) (3).
An additional metric was added to subjectively evaluate the models' prediction ability through generated sequences of words based on a seed sentence (4).

3. Results and Discussion

Several experiments were conducted on the baseline to achieve the best configuration with a perplexity score ≤ 250 (Fig. 1). The use of ADAMW and the inclusion of the two additional dropout layers set a relatively high standard for comparison (Table 3).
The results of the proposed model (Table 3, Fig.2) highlight the importance of aligning regularization techniques with model depth. When employing Variational Dropout, it is crucial to consider the architecture's capacity to ensure that the regularization enhances rather than hinders performance.
Adding Non-monotonically Triggered AvSGD significantly improves perplexity, indicating that this optimization strategy helps in better convergence and fine-tuning, effectively complementing the benefits of weight tying and Variational Dropout.
However, introducing TBPTT worsens the perplexity, suggesting it might disrupt the learning process in this context. This could be due to improper handling of temporal dependencies or unnecessary segmentation of sequences, which might not align well with the model's training dynamics and the dataset statistics.
Analysis of the generated sentences (5) reveals that models can be compared and evaluated based on syntactic and semantic correctness. This evaluation often diverges from perplexity scores, indicating that low PPL does not always correspond to high-quality and meaningful sentences.

4. References

[1] M. P. Marcus, M. A. Marcinkiewicz, and B. Santorini, “Building a large annotated corpus of english: The penn treebank,” *Comput. Linguist.*, vol. 19, no. 2, p. 313–330, jun 1993.

[2] S. Merity, N. S. Keskar, and R. Socher, “Regularizing and optimizing LSTM language models,” *CoRR*, vol. abs/1708.02182, 2017. [Online]. Available: http://arxiv.org/abs/1708.02182

[3] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[4] R. J. Williams and J. Peng, “An efficient gradient-based algorithm for on-line training of recurrent network trajectories,” *Neural Computation*, vol. 2, no. 4, pp. 490–501, 1990.

Table 1: *PTB dataset statistics.*

	Train set	Validation set	Test set
Total number of words	887521	70390	78669
Total number of sentences	42068	3370	3761
Average sentence length	21	21	21
Standard deviation	10.14	9.98	10.19
Min sentence length	1	1	1
Maximum sentence length	82	74	77
Sentences split percentage	85.51	6.85	7.64
# OOV words	-	0	0

Table 2: *Top 5 word frequencies in the train, validation and test sets.*

Word	Train set	Validation set	Test set
the	5.720	5.856	6.094
<unk>	5.073	4.951	5.757
N	3.660	3.698	3.207
of	2.749	2.603	2.790
to	2.663	2.486	2.596

1: *Details of the TBPTT implementation [2, 4].*

Sentences are divided into smaller chunks, and error backpropagation is performed only through these chunks. The splitting length is sampled dynamically from a Gaussian N(μ, σ^2) with probability p=0.95 and N($\frac{\mu}{2}$, σ^2) with probability $1 - p$. The splitting length is set to 30, $\mu = 30$, $\sigma^2 = 5$

2: *Cross-Entropy Loss equation: where q is the approximating distribution on data W*

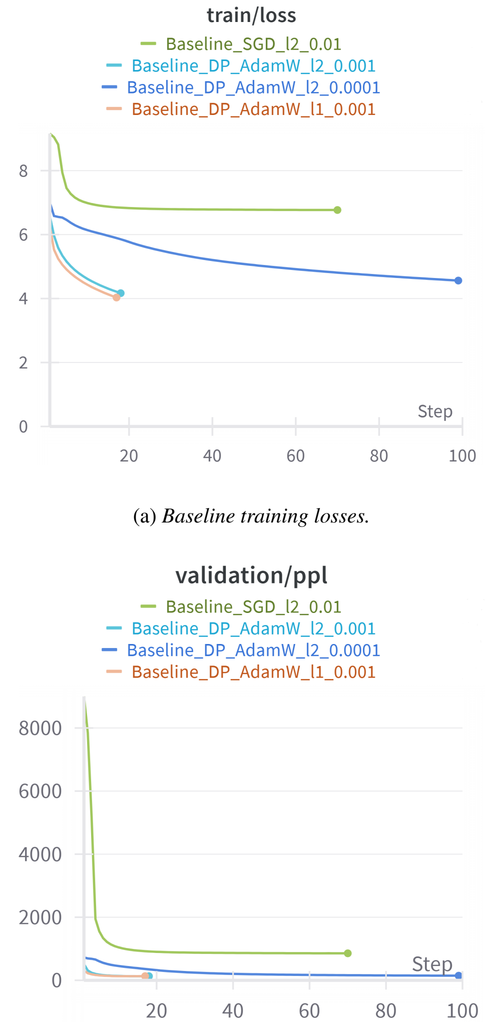
$$H(W) = -\frac{1}{N} log(p(W))$$

3: *Perplexity equation: where H(W) is the cross-entropy defined in* (2)

$$PP(W) = 2^{H(W)}$$

4: *Details of the additional subjective metric.*

The metric implementation uses the trained language models to generate a sequence of words based on a seed sentence. The ability of the models of generating meaningful sentences is tested on the outputs with maximum length of 20 and avoiding the ’unk’ tokens.



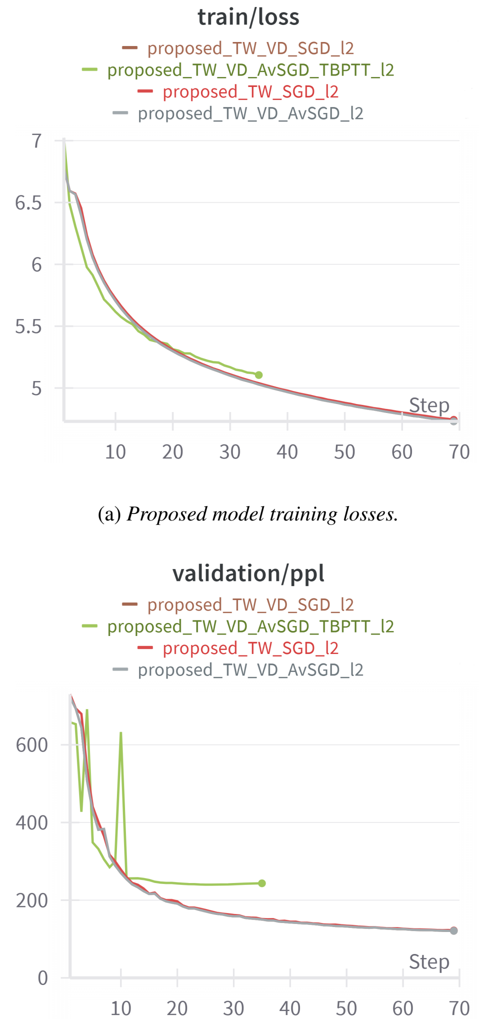
(a) *Baseline training losses.*

(b) *Baseline validation perplexities.*

Figure 1: *Baseline experiments.*
The required modifications (LSTM, two dropout layers (DP), AdamW) were implemented incrementally. Additionally the number of layers was modified to match the one of the proposed model.
The names of the models are defined as Baseline_{Dropout layers if added}_{optimizer}_{# layers}_{learning rate}

Table 3: *Results (PP) of the experiments.*
Where TW = tie_weights, VD = Variational Dropout, AvSGD = Non- monotonically Triggered Averaged Stochastic Gradient Descent, TBPTT = truncated back-propagation through time.

Model	Validation	Test
Baseline_LSTM_AdamW	131.16	120.04
proposed_TW_SGD	103.68	100.19
proposed_TW_VD_SGD	104.20	100.98
proposed_TW_VD_AvSGD	102.75	99.58
proposed_TW_VD_AvSGD_TBPTT	234.33	214.37



(a) *Proposed model training losses.*

(b) *Proposed model validation perplexities.*

Figure 2: *Proposed model experiments.*
The required modifications (Weight Tying, Variational Dropout (VD), Non-monotonically Triggered AvSGD) were implemented incrementally. Additionally the number of layers and the use of TBPTT is reported.
The names of the models are defined as proposed_TW_{VD if added}_{optimizer}_{TBPTT if added}_{# layers}

5: *Result of the generated sequences.*
The initial tokens of the generated sequences are [”the”, ”los angeles”, ”in the first half”]:

Baseline

- *the offering plo casualty everyday phony motorola misstated expense input*
- *los angeles slowed ruth bond-equivalent democratic lowest drifted highs kane*
- *in the first half restricting slated source grabbed bolster boost*

proposed_TW

- *the N magazine divisions per-share districts illuminating earning hugo fanfare*
- *los angeles slowed abuse bennett describe gamble opposing colony closings*
- *in the first half thailand civil condemn campbell england chan*

proposed_TW_VD_l2

- *the government extraordinary progress farrell gold plan rudolph detailing thieves*
- *los angeles slowed army distinct describe publicity carefully motor helm*
- *in the first half agreements civil styles warnings eggs bottling*

proposed_TW_VD_AvSGD

- *the government only moment united ideal jamie split industries banking*
- *los angeles slowed army distinct describe publicity carefully motor removed*
- *in the first half agreements civil condemn warnings additional affordable*

proposed_TW_VD_AvSGD_TBPTT

- *the government wait easier wall died attribute detergent interpreted reasonable*
- *los angeles motors midmorning prize exception chapter conduct protect wild*
- *in the first half developing answers financial-services shearson touted i.*