NLU Course Project – Language Model Experiments

Stefano Racca (mat. 256173)

University of Trento

stefano.racca@studenti.unitn.it

1. Introduction

I conducted two complementary experiments on Penn Treebank Dataset (PTB) to evaluate methods for reducing test perplexity (PPL). In **Part A**, I began with a single-layer LSTM (embedding size = 650, hidden size = 650) trained with vanilla SGD, then added dropout layers and finally switched from SGD to AdamW for the optimizer. Each modification was applied incrementally, retraining to convergence with early stopping and hyperparameter tuning. In **Part B**, I built on that LSTM backbone by integrating three techniques from Merity [1]: (i) *weight tying* between embedding and output projections; (ii) *variational dropout* at the embedding and output stages; and (iii) *non-monotonically triggered Averaged SGD* (NT-AvSGD). I compared four configurations to measure their impact on final test PPL. The first configuration used only weight tying; the second added variational dropout; the third included NT-AvSGD; and the fourth replaced SGD with AdamW as the initial optimizer. Each model was evaluated under the same PTB splits, employing early stopping on validation PPL with a patience of 5 epochs.

2. Implementation details

2.1. Part A

My baseline model is a single-layer LSTM with an embedding dimension of 650 and hidden state size of 650. I trained it using SGD (learning rate = 5, gradient clipping = 0.25). The vocabulary was built from PTB training text, and I appended an end-of-sentence token (`<eos>`). After verifying a base test PPL of approximately 131.11, I inserted dropout layers: a 0.2 dropout on the embedding outputs and a 0.3 dropout before the final linear projection. Introducing dropout reduced overfitting significantly, yielding a test PPL of about 110.16.

Next, I replaced SGD with AdamW, using a learning rate of 0.0001 and a weight decay = 1e-6. The AdamW optimizer takes more epochs to train but gets better results. Early stopping was triggered when validation PPL did not improve for five consecutive epochs. Under AdamW, the final test PPL decreased slightly to 109.79, demonstrating a modest gain from the adaptive optimizer.

2.2. Part B

All experiments in Part B share a single-layer LSTM backbone (no bidirection), using embedding and hidden dimensions of 800. Two 0.5 variational dropout was implemented (one dropout mask per sequence), following the method proposed in [2], at two points: immediately after the embedding lookup and immediately before the linear classifier. The linear layer's weights are tied to the embedding matrix when embed = hidden, reducing parameter count.

The four training configurations progressively integrate the techniques under evaluation:

- **WT**: Baseline with weight tying only, trained using SGD with a learning rate of 5.
- **WT+VD**: Adds variational dropout to the previous setup, trained with the same SGD schedule.
- **WT+VD+AvSGD**: Further includes non-monotonically triggered Averaged SGD (NT threshold = 5), starting with SGD (lr = 5) and switching to AvSGD (lr = 1) when triggered.
- **WT+VD+AvSGD+AdamW**: Same as the previous configuration, but starts with AdamW (lr = 0.005), followed by AvSGD (lr = 1) upon triggering.

For all four runs, I used a mini-batch size of 64 for training and 128 for dev and test sets, padding/tokenization as per PTB conventions, and applied gradient clipping at a norm of 5. Early stopping monitored validation PPL with a patience of 5 epochs.

3. Results

Table 1 reports Part A test PPLs. The baseline LSTM trained with SGD (no dropout) reached 131.41. Introducing dropout reduced PPL to 110.16, confirming dropout's strong regularization effect. Switching to AdamW gave a further modest gain (final PPL 109.79), as AdamW stabilized weights during later epochs, though its effect was smaller than that of dropout alone.

Table 1: *Part A: Test Perplexities*

Experiment	**Test PPL**
LSTM base	131.11
LSTM + dropout layers	110.16
LSTM + dropout layers + AdamW optimizer	109.79

Table 2 reports the test perplexity (PPL) results for the four configurations evaluated in Part B. The baseline configuration, **WT**, using only weight tying, achieved a PPL of 120.11. Adding variational dropout (**WT+VD**) led to a substantial improvement, reducing test PPL to 96.86. Further gains were observed with the addition of non-monotonically triggered Averaged SGD (**WT+VD+AvSGD**), which achieved the best PPL of 95.83. All these configurations outperformed those explored in Part A, confirming the effectiveness of these regularization techniques. Finally, I experimented with using AdamW instead of SGD as the initial optimizer (**WT+VD+AvSGD+AdamW**) out of personal interest; however, this resulted in a slightly higher PPL of 99.41, suggesting that in this setting SGD remains the more suitable choice. The figure 1 shows a smooth and gradual decline in validation perplexity for WT+VD+AvSGD+AdamW, with convergence occurring slowly across the full training schedule.

Table 2: *Part B: Test Perplexities with Regularization*

Configuration	**Test PPL**
WT	120.11
WT+VD	96.86
WT+VD+AvSGD	95.83
WT+VD+AvSGD+AdamW	99.41

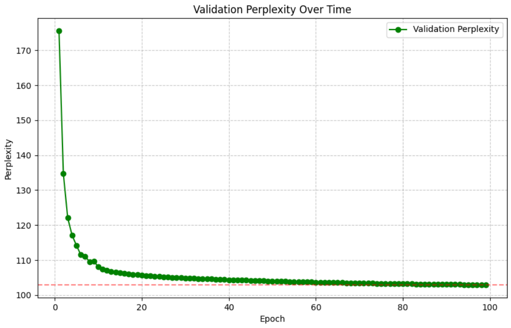

Figure 1: *WT+VD+AvSGD+AdamW: Validation perplexity over training epochs.*

4. References

[1] N. S. K. . R. S. Stephen Merity, “Regularizing and optimizing lstm language models,” 2018.

[2] Z. G. Yarin Gal, “A theoretically grounded application of dropout in recurrent neural networks,” 2016.