NLU course project - Lab4

Mauro Antonio de Palma (256175)

University of Trento

mauroantonio.depalma@studenti.unitn.it

1. Introduction

This report presents the implementation and evaluation of neural language models based on Long Short-Term Memory (LSTM) networks, trained and evaluated on the Penn TreeBank corpus. The work is divided into two parts:

- **Part A**: Incremental improvements over a vanilla LSTM baseline through dropout regularization and optimizer selection (SGD vs AdamW).
- **Part B**: Advanced regularization techniques including Weight Tying, Variational Dropout, and Non-monotonically Triggered Averaged SGD (NT-AvSGD).

The primary objective was to minimize perplexity (PPL) through progressive architectural enhancements and hyperparameter optimization. All model configurations were required to achieve a test perplexity below 250 ($PPL < 250$), with Part B models expected to outperform the Part A baseline.

2. Implementation details

All models share a common architecture: an embedding layer maps discrete tokens to dense vector representations, an LSTM network processes sequential dependencies and captures long-range context, and a linear projection layer outputs probability distributions over the vocabulary. Training employs CrossEntropyLoss with padding tokens excluded from computation, gradient clipping (maximum norm of 5) to prevent exploding gradients, and early stopping based on validation perplexity. Models are evaluated using perplexity (PPL), computed as the exponentiation of the average cross-entropy loss.

2.1. Part A: Baseline and Incremental Improvements

The baseline architecture comprises a single-layer LSTM with 200 hidden units and 300-dimensional word embeddings, trained with Stochastic Gradient Descent (SGD) with momentum.

The second configuration introduces dropout regularization ($p = 0.3$) at two critical points: immediately following the embedding layer and preceding the output projection. This dual-dropout strategy mitigates overfitting by preventing co-adaptation of hidden units during training.

The third configuration extends the architecture depth to two LSTM layers and replaces SGD with the AdamW optimizer, which incorporates decoupled weight decay regularization for improved generalization.

2.2. Part B: Advanced Regularization

Building upon the LSTM baseline, three regularization techniques were implemented following Merity et al. [1]:

Weight Tying enforces the constraint that the input embedding matrix and the output projection matrix share identical weights. This reduces the number of trainable parameters and improves generalization by ensuring consistent word representations across input and output spaces. When weight tying is enabled, the hidden dimension is set equal to the embedding dimension.

Variational Dropout applies a consistent dropout mask across all time steps within a sequence, rather than sampling a new mask at each step as in standard dropout. This approach preserves temporal correlations and has been shown to be more effective for recurrent architectures. The same mask is applied after the embedding layer and before the output projection.

Non-monotonically Triggered Averaged SGD (NT-AvSGD) is an optimization strategy that begins with standard SGD training and switches to weight averaging when the validation metric fails to improve. Once triggered, the algorithm maintains a running average of model parameters across subsequent epochs, which typically yields smoother convergence and better generalization than the final iterate alone.

3. Results

This study explored various regularization and optimization techniques to improve language modeling performance on the Penn TreeBank corpus. While the baseline LSTM achieved a test perplexity of 157.47 using SGD optimizer, the introduction of advanced regularization techniques in Part B yielded significant improvements.

The combination of Weight Tying and Variational Dropout proved to be the most effective configuration, achieving the lowest test perplexity of **105.31**, well below the target threshold of 250. Although NT-AvSGD theoretically offers better convergence properties, in our experiments it did not surpass the performance of the Variational Dropout model, likely due to the hyperparameter search space or the limited number of epochs after triggering the averaging phase. Overall, the results highlight the critical importance of weight tying and consistent dropout masks in recurrent neural network training.

Table 1: *Result of an LSTM model with dropout and AdamW/SGD optimizers*

Architecture	**LR**	PPL/Val	PPL/Test	Epochs
LSTM SGD	0.01	212.613	205.938	100
	0.02	177.734	171.026	100
	0.03	163.534	157.467	100
LSTM SGD drop ($p = 0.3$)	0.01	240.233	232.138	100
	0.02	197.125	189.677	100
	0.03	178.740	171.91	100
LSTM AdamW	0.0005	166.56	154.504	37
	0.001	156.989	142.832	22
	0.002	154.07	138.106	15

Table 2: *Results of an LSTM model with advanced regularization techniques. The start averaging period for NTAvgSGD is shown in parentheses.*

Architecture	**LR**	PPL/Val	PPL/Test	Epochs
LSTM (WT)	0.01	114.771	118.547	43
	0.02	111.747	115.527	21
	0.03	111.503	113.211	16
LSTM (WT, VarDrop)	0.01	109.345	113.628	91
	0.02	110.772	105.307	51
	0.03	113.093	116.95	34
LSTM NTAvSGD	0.0005	109.609	113.547	100(94)
	0.001	111.588	112.444	100(56)
	0.002	113.766	113.481	100(53)

4. References

[1] S. Merity, N. S. Keskar, and R. Socher, "Regularizing and optimizing lstm language models," in *ICLR*, 2018.