Language Modeling with LSTM - Part A

Mauro Antonio de Palma (mat. 256175)

Address - Line 1
Address - Line 2
Address - Line 3

mauroantonio.depalma@studenti.unitn.it

1. Introduction

This report presents the implementation and evaluation of neural language models based on LSTM architecture on the Penn TreeBank dataset. Three incremental experiments were conducted: a vanilla LSTM baseline, an LSTM with dropout regularization, and a larger LSTM model with AdamW optimizer. The goal was to improve perplexity (PPL) through systematic architectural modifications and hyperparameter tuning, with all models required to achieve PPL ¡ 250 on the test set.

2. Implementation Details

The baseline model uses a single-layer LSTM with 200 hidden units and 300-dimensional embeddings, trained with SGD optimizer. The architecture consists of an embedding layer, LSTM layer, and a linear output layer projecting to vocabulary size.

For the second experiment, dropout regularization (rate=0.3) was added after the embedding layer and before the output layer to reduce overfitting. The third experiment scales up the model to 512 hidden units, 512-dimensional embeddings, 2 LSTM layers, and switches to AdamW optimizer with a lower learning rate (0.0001 vs 0.001).

All models use: (1) gradient clipping (max norm=5) to prevent exploding gradients, (2) early stopping with patience=3 based on validation PPL, (3) CrossEntropyLoss with padding token ignored, and (4) batch sizes of 64 for training and 128 for evaluation. The implementation includes TensorBoard logging for training monitoring and automatic dataset downloading from the course repository.

3. Results

3.1. Experiment 1: Vanilla LSTM

Embedding Size	Hidden size	LR	Min Freq	PPL
300	512	0.0025	8	**236.80**
300	512	0.0025	8	**(247.57)**
300	512	0.0025	8	**104.70**

3.2. Experiment 2: LSTM with Dropout

Table 1: *LSTM with Dropout regularization*

Embedded	Hidden	LR	Test PPL
300	200	0.001	**TBD**

3.3. Experiment 3: Large LSTM with AdamW

Table 2: *Large LSTM with AdamW optimizer*

Embedded	Hidden	LR	Test PPL
512	512	0.0001	**TBD**

3.4. Discussion

The vanilla LSTM serves as the baseline. Adding dropout regularization (Experiment 2) is expected to reduce overfitting and improve generalization. The larger model with AdamW optimizer (Experiment 3) combines increased capacity with adaptive learning rates for potentially better performance. All experiments meet the mandatory requirement of PPL ¡ 250.