Report: Binary Classification with CBOW and CNN

Cheikh Ahmed Tidiane Mane

1 Introduction

This report presents the results of a binary classification task of movie reviews (positive or negative) using two main approaches: CBOW (with and without a hidden layer) and CNN. The objective is to evaluate the models' performance in terms of accuracy and loss, as well as their generalization capacity on a validation set.

2 Methods

2.1 Data Preprocessing

The textual data is cleaned to remove special characters and convert words to lowercase. A dictionary is created to map each unique word to an integer, allowing sentences to be transformed into sequences of indices. The data is then split into *train*, *dev*, and *test* sets.

2.2 Tested Models

- **CBOW Without Hidden Layer**: Averages the word embeddings in each sentence, followed by a single linear layer.
- **CBOW With Hidden Layer**: Averages the word embeddings, followed by a hidden linear layer with ReLU activation and a second linear layer to produce the binary output.
- **CNN**: Embedding of words, 2D convolutions with filter sizes $[2, 3, 4]$, followed by max pooling, vector concatenation, and a fully connected layer.

2.3 Hyperparameter Selection

- **Vocabulary Size**: Number of unique words in the training set plus 1.
- **Embedding Dimension**: Set to 50.
- **Loss Function**: Binary Cross-Entropy Loss.
- **Optimizer**: Adam with a learning rate of 0.001.
- **Batch Size**: 32.
- **Number of Epochs**: 25.

For the CBOW model with a hidden layer, the hidden layer dimension is set to 25. For CNN, the number of filters is set to 5 for each filter size.

3 Results

3.1 CBOW Without Hidden Layer

- **Train Accuracy**: 94.11%
- **Dev Accuracy**: 78.66%
- This model demonstrates good generalization despite its simplicity.

3.2 CBOW With Hidden Layer

- **Train Accuracy**: 97.04%
- **Dev Accuracy**: 76.02%
- Adding a hidden layer improves learning capacity, but generalization is slightly reduced.

3.3 CNN

- **Train Accuracy**: 99.60%
- **Dev Accuracy**: 76.69%
- CNN effectively captures local patterns (n-grams), but the gains in generalization remain limited.

4 Discussion

4.1 Model Comparison

Model	Train Accuracy	Dev Accuracy	Strengths
CBOW Without Hidden Layer	94.11%	78.66%	Simple, fast to train
CBOW With Hidden Layer	97.04%	76.02%	Improved learning capacity
CNN	99.60%	76.69%	Captures local patterns

For weaknesses: CNN is more computationally expensive and prone to overfitting.

4.2 Result Analysis

The CBOW model without a hidden layer quickly achieves acceptable accuracy but lacks flexibility to capture complex relationships. The CBOW model with a hidden layer improves expressiveness at the cost of generalization. The CNN offers interesting capabilities for capturing local patterns, but the gains in generalization are limited, suggesting possible overfitting.

5 Conclusion

- The CBOW without a hidden layer is a simple, efficient, and fast model, achieving a **Dev Accuracy** of 78.66%.
- The CBOW with a hidden layer adds non-linear learning capacity but appears to overfit slightly.
- The CNN offers interesting capabilities for capturing local patterns, but the gains in generalization remain limited.