CSE 587: Deep Learning for NLP

Midterm Project

Spam Detection Using CNN and GloVe Embeddings

Jeremy Huang

Date: 2/14/2025

A. Problem Definition and Dataset Curation

Spam messages are a common issue nowadays, especially in emails and SMS. Spam detection helps in filtering out unwanted messages. The goal of this project is to classify messages as either spam or non-spam using a Convolutional Neural Network (CNN) with word embeddings.

The dataset used for this project is the SMS Spam Collection Dataset, which contains 5,574 messages labeled as either spam or ham (non-spam). The dataset was downloaded from Kaggle and processed to remove unnecessary characters, convert text to lowercase, and tokenize the words.

B. Word Embeddings, Algorithm, and Training Process

Word Embeddings

For text representation, we used pre-trained GloVe word embeddings (100-dimensional). GloVe captures the semantic meaning of words based on co-occurrence in a large corpus. The embeddings were loaded and mapped to words in the dataset.

Model Architecture

A CNN was used to classify the messages. The model consists of the following layers:

- Embedding Layer: Maps words to dense vector representations using the pre-trained GloVe embeddings.

- 1D Convolutional Layer: Detects important word patterns and n-grams in messages.
- Global Max Pooling Layer: Reduces dimensions while preserving important features.
- Dense Layer: Fully connected layer with ReLU activation.
- Dropout Layer: Prevents overfitting by randomly deactivating neurons.
- Output Layer: Uses sigmoid activation to classify messages as spam or ham.

Training Details

- Loss Function: Binary Cross-Entropy
- Optimizer: Adam
- Epochs: 5
- Batch Size: 32
- Validation Split: 20% of the dataset used for validation

C. Results and Presentation

After training for 5 epochs, the model achieved the following results:

	Training Set	Validation Set
Accuracy	0.9984	0.9874
Loss	0.0103	0.0448

Figure 1. Model Training and Validation Performance

The accuracy is very high, indicating that the model is learning effectively. The low validation loss shows that the model can work well with new data.

D. In-Depth Analysis and Experiments

Checking for Overfitting

The training accuracy (99.84%) is slightly higher than validation accuracy (98.74%), but the gap is small. This suggests that the model is not overfitting heavily.

Testing on New Messages

We tested new spam and ham messages to check real-world performance. The model correctly classified promotional and phishing messages as spam while keeping normal conversations as ham.

Input	Predicted Probability	Interpretation
Congratulations! You won a free iPhone. Click here!	9.6652335e-01	Spam
Hey, are we meeting tomorrow?	8.1521016e-08	Ham

Figure 2. Predictions on New Messages

E. Lessons&Experience

- GloVe embeddings improved how the model understood words.
- CNN worked well for short text, making it a good fit for SMS spam detection.
- The model performed well even with fewer training epochs.
- Testing on new messages showed that it classified spam accurately.
- Future work could explore using LSTM models to see if they offer any advantage in capturing sequential patterns.