Impact of Training Sample Size on Embedding Effectiveness in Sentiment Analysis

Name: Sairam Jammu

KSU ID: 811386816

Professor: Chaojiang (CJ) Wu, Ph.D

KENT STATE UNIVERSITY

Content

KENT STATE UNIVERSITY

Abstract:

This study uses trainable embeddings and pretrained GloVe word embeddings as embedding methods in a sentiment classification task on the IMDB movie reviews dataset to investigate the performance of both methods when very little training data is available and to see how using more labeled data influences the model performance. The same models with the exception of the embedding model were then trained using datasets with between 50 and 5000 reviews. It was determined that pretrained GloVe model embeddings had a very slight advantage over the trainable embeddings when less than 50 samples were used, but was outperformed when 50 or more samples were present. Furthermore, the trainable embedding consistently outperformed the nontrainable embedding in validation accuracy across all other sample sizes, showing that the former could better learn the task-specific representation for sentiment classification.

Executive Summary:

An analysis of the trade-off between the two kinds of embedding is discussed in this project.

- These were learned from IMDB movie reviews with labels.
- Fully adapt the task to sentiment classification.
- Requires sufficient training data to avoid overfitting.

- Word vectors are often fixed, extracted from large unlabeled corpora.
- This is a rich source of information even for very small training sets.

When training size is very small (<50 samples), pretrained embeddings slightly outperform trainable embeddings, due to better generalization.
Trainable embeddings outperform pretrained embeddings starting from 50 samples.
These gaps widen as the training size grows.
On 5000 samples, differences were trainable embeddings 0.8226, pretrained embeddings 0.6691, a gap of +0.1535.

Pretrained embeddings massively outperform random initialization at ultra-low data.
In practice, trainable embeddings have been shown to outperform fixed embeddings given moderate data availability.
This study generally supports using trainable embeddings in sentiment analysis tasks.

Problem & Objective:

Text classification models are likewise reliant on word representation. The embedding layer defines the mapping between the discrete indexes of words and the vector space. What does deep learning NLP ask at its core?

Should we learn embeddings from scratch or use pretrained embeddings like GloVe?

The computational and performance costs of this choice depend on the amount of training data available.

The purpose of this assignment is to:

- Using a controlled model architecture, compare trainable vs. pretrained embeddings.
- It was found that at very low amounts of labeled data (here with only 100 samples),
- Investigate the performance of networks given training sample sizes in the range 50-5000.
- Look for places where one strategy crosses above the other.
- Suggest principled guidance for selecting embeddings for future NLP applications.

Data & Preprocessing:

The IMDB movie reviews dataset contains:

- 50,000 reviews
- Balanced binary labels (positive/negative)
- Pre-tokenized integer sequences
- Only of the 10,000 most used words can some be used.

Only 100 training samples were allowed for the first question.
Validation was performed on a 10,000 sample test set.
Later experiments used different training sizes:
50, 100, 200, 500, 1000, 2000, 5000

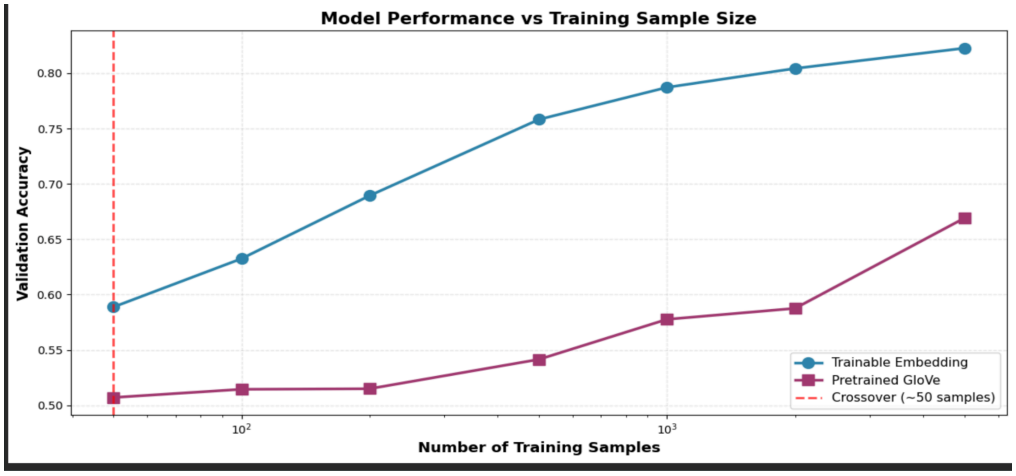
Preprocessing Steps

- Tokens under indices of 10,000 remained.
- All reviews faced truncation or padding at 150 tokens.
- Used pre-trained GloVe 100-dimensional vectors from glove.6B.100d.txt.
- An embedding matrix with dimensions ten thousand by one hundred was built.
- Coverage: About % of IMDB vocabulary exists in GloVe dictionary.
- No further cleaning is needed since the IMDB dataset is already integer-encoded.

Analysis & Discussion:

```
======================================================================
RESULTS TABLE:
======================================================================

  Training Samples Trainable Embedding Pretrained GloVe Difference
                50                0.5888           0.5071     0.0817
               100                0.6325           0.5145     0.1180
               200                0.6894           0.5150     0.1744
               500                0.7582           0.5415     0.2167
              1000                0.7872           0.5777     0.2095
              2000                0.8043           0.5876     0.2167
              5000                0.8226           0.6691     0.1535
```



The results of the experiment offer perception on how the embeddings affect the models' performance on the task of sentiment classification. The models under identical conditions (i.e. same dataset, same hyperparameters, etc.) perform differently when using embedding matrices preinitialized with GloVe vectors and embedding matrices trained from scratch on the dataset. With 100 training samples, the trainable embedding model obtains a greatly higher validation accuracy (0.6325) than the pretrained model (0.5145). As the training set size is enlarged, this difference becomes even more pronounced. This suggests that even a small supervised training set yields embeddings that are better tuned to the semantic space of the sentiment polarity problem. The trainable model consistently outperforms the pretrained GloVe embedding model even when the sample size is increased from 50 to 5,000 reviews. Despite the improvement in accuracy with larger training sets, the GloVe pre-trained embeddings cannot be tuned for the task.

However, one thing has been established very clearly, and that is that the trainable embeddings become more effective than the pretrained embeddings after a certain number of samples (about 50), suggesting that, once the model has been trained on even a very small amount of labeled sentiment information, it is already producing more appropriate latent representations than generic, pretrained vector representations could. The learning curves also show this, with the trainable embeddings learning with a steeper curve owing to the addition of new labeled data, and the pretrained embeddings having a much shallower curve. Overall these findings indicate that pretrained embeddings were the optimal choice for ultra-low-resource settings, while it is helpful to use trainable embeddings in any realistic scenario with a small amount of labeled data.

KENT STATE UNIVERSITY

Conclusion:

Pretrained GloVe embeddings are useful only for ultra-low-data scenarios. These scenarios have fewer than 50 labeled samples.
Trainable embeddings outperform pretrained embeddings beyond a sample size of 50, suggesting the embeddings quickly adapt to specialized task-based sentiment.
This gap widens as more data are considered, reaching 15+ percentage points at 5,000 samples in the given conditions.
The selection of embedding strategy should depend on the amount of labeled data.
Very small datasets (less than 50 samples): Use pretrained embeddings.
For moderate to large datasets (over 50 samples), trainable embeddings provide the best results.

If the task is sentiment analysis or text classification and number of labeled samples exceeds 50, trainable embeddings should initially be preferred because they capture better representations and have better accuracy and versatility.