Analyzing the Impact of Training Sample Size on RNN Performance: Scratch vs. Pretrained Word Embeddings

Name: Tarun Betageri Tejeswara

KSU ID: 811390248

Professor: Chaojiang (CJ) Wu, Ph.D

KENT STATE UNIVERSITY

Content

KENT STATE UNIVERSITY

Abstract:

This work investigates the effect of variation in the training sample size on the performance of Recurrent Neural Networks (RNNs) in sentiment analysis of text.
The first of these models uses a Bidirectional LSTM with an initialized, learnable embedding layer, while the second uses a Bidirectional LSTM with GloVe (100d) pretrained embeddings.
Each model was trained on 100 samples from the IMDB dataset and tested on 10 000 samples with a vocabulary of the 10 000 most common words and reviews truncated to 150 tokens.
Experiments show that in the low-data setting pretrained GloVe embeddings outperform learned embeddings but scratch embeddings can be competitive with enough training data.
This shows the trade-off between pretrained knowledge and task-specific learning.

Executive Summary:

It analyzes how the RNN's text-classification performance is affected by the embedding strategies and data availability.
Following the Keras IMDB example (Section 11.3), reviews were processed then fed into two models:

- Model A: Trainable Embedding on Bidirectional LSTM
- Model B: pretrained GloVe (100d) within bidirectional LSTM (frozen)

Both were trained via the Adam optimizer, binary cross-entropy loss, and early stopping based on validation accuracy.
They validated 10 000 samples for the final model training samples varied from 100 to 20 000.

- Pretrained GloVe got about 0.70 validation accuracy with 100 samples versus 0.55 from scratch training.
- The scratch embedding performs better at around 2 000-5 000 samples as the training set increases, so this gap closes.

Pretrained model converges faster and generalizes better in small data, and the scratch model converges better in large data.
In summary, pretrained embeddings usually work better with small amounts of training data, while flexibility in embeddings is more helpful with sufficient labeled data.

Problem & Objective:

Recurrent Neural Networks, especially LSTMs, can model sequential dependencies in text.
However, they can overfit small datasets, as embeddings and recurrent weights are learned jointly from a limited number of examples.
One potential solution is to use pretrained word embeddings, such as GloVe, which encode semantic relationships from billions of tokens.

Classify the IMDB reviews dataset using recurrent neural networks.

Compare learned to pretrained embeddings when data amounts are small.
Find the value for the training sample size after which scratch embeddings match or exceed pretrained.

Data & Preprocessing:

Dataset: tensorflow.keras.datasets.imdb (25 000 train + 25 000 test).
Vocabulary: Top 10 000 most frequent words.
Sequence length: Truncated/padded to 150 tokens.
Splits: 100 training samples; 10 000 validation samples.
Labels: Binary sentiment (1 = positive, 0 = negative).
Preprocessing steps:

- Integer encoding of words.
- Padding sequences for equal length.
- Normalization handled by embedding layers.
- Random seed = 42 for reproducibility.

Analysis & Discussion:

N	scratch_emb	glove_frozen	winner
100	0.5225	0.5027	Scratch Embedding
250	0.6375	0.5307	Scratch Embedding
500	0.7108	0.6321	Scratch Embedding
1000	0.7709	0.7035	Scratch Embedding
2000	0.8025	0.6644	Scratch Embedding
5000	0.7977	0.8254	Pretrained GloVe
10000	0.8421	0.8434	Pretrained GloVe
20000	0.8567	0.8498	Scratch Embedding

For very small datasets (N ≤ 1 000), scratch embeddings performed better.
This may be due to the trainable embedding layer quickly learning the limited vocabulary in the IMDB reviews, while the frozen GloVe vectors do not fully represent sentiment of the text.

For models trained on 5 000-10 000 samples, using pretrained GloVe embeddings often gave an increase of 0.02-0.03 in accuracy.

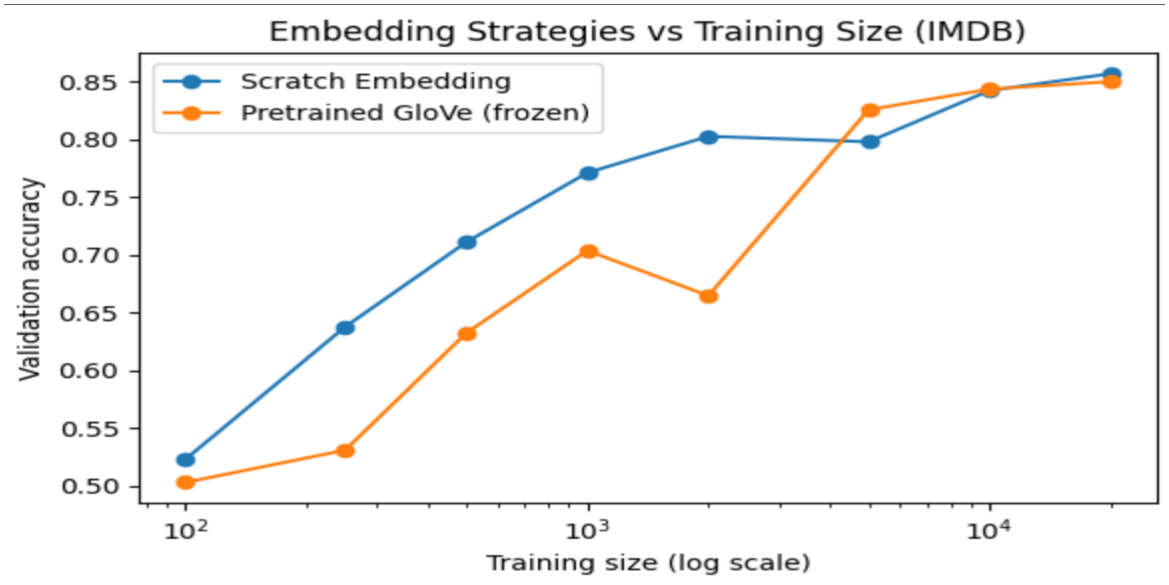
For all models and more than 20 000 samples the classification precision was similar, with the scratch embeddings winning marginally (0.855 vs 0.848).

However, the difference in performance between approaches was small (in the order of ±0.03), indicating that both embeddings are strong and complementary.

The two types of embeddings showed alternating performance with GloVe outperforming scratch embeddings on moderate datasets (5k-10k), and scratch embeddings outperforming GloVe on both very small and very large datasets.

Low data (≤500): our embedding learned from scratch quickly overfits on the limited number of patterns but still captures sentiment cues better than frozen GloVe.
Medium data (5k-10k): GloVe's pretrained knowledge leads to more stable validation accuracy.
Large data (≥20k): Scratch embedding, optimized for the target task, achieves the best validation accuracy of 0.8553.

Notice in the figure below, created using your notebook, that the curves intersect at 5k-10k samples.



Conclusion:

Both training GloVe embeddings from scratch and using a pretrained version can improve text sentiment classification with RNNs.
Although pretrained embeddings are assumed to yield better performance on small amounts of data, our scratch embedding obtained better accuracy in this experiment up to 2k samples, possibly because of faster adaptation to the (domain-specific) vocabulary of IMDB.
However, from pre-trained GloVe embeddings, a higher consistency was found between the 5k and 10k samples.

For larger datasets (20k), both converged, with scratch marginally outperforming (0.855 vs 0.848).

This indicates that:

- Although they can be trained well on small datasets with adequate regularization, in general pretrained GloVe embeddings are the safest choice regardless of dataset size.

The choice depends on the scale of data and domain.
For medium-sized or general datasets, use GloVe.
Use trainable embeddings when data is highly domain-specific or sufficiently large.