Advanced Machine Learning

Assignment 3

Text And Sequence Data

Rohith Desamseety

Summary:

In order to forecast whether a movie review would be favorable or unfavorable, our study's main objective was to conduct binary classification on the IMDB review dataset. We trained our models using a range of sample sizes, including 100, 500, 1,000, and 10,000 reviews, and the dataset has a total of 50,000 reviews. We only focused on the top 10,000 most frequently used terms while validating our findings on 10,000 samples to ensure consistency. The dataset was pre-processed before being fed into the embedding layer and a pretrained embedding model. After putting the models through various tests, we adjusted them as necessary to optimize their performance.

Technique:

Dataset and Preprocessing: The IMDB dataset consists of a set of film reviews with labels for their overall emotion (positive or negative). Every review in the dataset is transformed into a series of word embeddings, where each word is represented by a vector of fixed size, as part of the preprocessing stage. 10,000 words is the specified maximum for the vocabulary. The reviews were also converted from their original format of words into an integer format in which each number stood for a particular word. We now have a list of numbers; however, they are inappropriate for the input of our neural network. It is necessary to create tensors from the integers. The list of integers might become a tensor with an integer data type and structure (samples, word indices). In order to do that, we must ensure that each sample has the same length; as a result, we must bolster each review with dummy words (integers) in order to ensure that they are of the same length.

Sizes of the training samples and their outcomes

Model	Training size	Scratch or pre-trained	Test loss	Accuracy %
1	-	Scratch	0.35	0.864
2	100	scratch	0.69	0.501
3	100	Pre-trained	3.3	0.500
4	1000	Embedding- layer	0.68	0.562
5	15000	Embedding- layer and conv1D	0.41	0.813
6	30000	Embedding layer and conv1D	0.43	0.800

7	15000	Pre-trained	4.8	0.500
8	30000	Pre-trained	0.95	0.517

- Comparing models 1 and 2, the first model, which had no alterations done, had a test accuracy of 86%, demonstrating its accuracy in classifying reviews as positive or negative just on their textual content. The model's performance, however, significantly declined when trained on just 100 data, with a test accuracy of just 50%.

- With a maximum review length of 150 words, a training sample limit of 100, validation on 10,000 samples, and consideration of just the top 10,000 words, the accuracy of the models using an embedding layer and a pre-trained word embedding was nearly comparable at 50%.

- We tried with various training sample sizes to find the threshold at which the embedding layer outperforms the pre-trained word embedding. According to our research, using 1,000 training samples makes the embedding layer perform better than the pre-trained word embedding. In particular, the embedding layer scored a test accuracy of 0.56 whereas the pre-trained word embedding, trained on just 100 samples, achieved a test accuracy of 0.50.

- As even a minor increase in the training sample size had minimal effect on accuracy, we next investigated the usage of conv1D in combination with the embedding layers and raised the training sample size to 15000 and 30000. This led to increases in test accuracy of 81% and 80%, respectively.

- When Conv1D was added to the embedding layer, the models performed better than the pre-trained word embedding model and had higher test accuracies at training sample sizes of 15000 and 30000.

Findings:

The effectiveness of pretrained embeddings might be impacted by the quantity of the training sample. Pretrained embeddings may be less accurate as the size of the training sample grows because they may be unable to accurately represent the subtleties of the job. Additionally, utilizing pretrained embeddings with greater training sample sizes might result in overfitting right away, lowering accuracy. Since the "best" solution depends on the requirements and constraints of the activity, it is challenging to make that determination. Model Performance is Dependent on Parameters The parameters utilized, such as the quantity of training data, the word embedding, the maximum review duration, among others, have a significant impact on the model's performance. To obtain the optimum performance, it is necessary to carefully choose and adjust these factors. Efficiency of the Embedding Layer with Fewer Datasets When working with less datasets, using an embedding layer could be more effective. By doing this, the model may concentrate on the variations of the data, thereby enhancing accuracy.