

Name - Shefali Gupta

KSU ID - 811276732



Objective of the project

The objective is to enhance a language model's performance by analyzing diverse factors, including training data size, vocabulary size, and word embeddings. The aim is to identify optimal combinations of model configurations that improve its effectiveness across various scenarios.

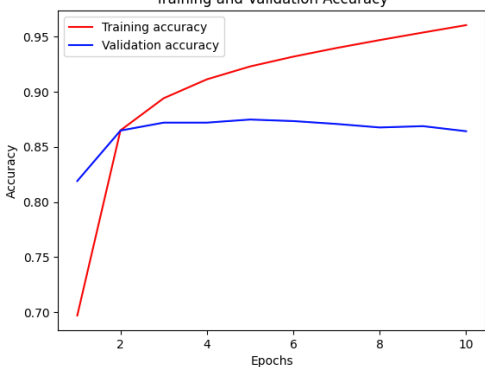
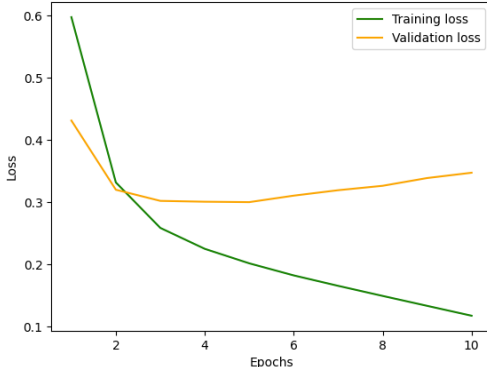
Assignment Overview

For this assignment, we will revisit the IMDB example discussed in Chapter 6, introducing several modifications to investigate their impact on the model's performance. Firstly, reviews will be constrained to a maximum of 150 words. The training dataset will be restricted to just 100 samples, while validation will be conducted on a set of 10,000 samples. Additionally, the analysis will focus on the top 10,000 words. To deepen the exploration, both an embedding layer and a pre-trained word embedding will be considered to determine which approach yields superior results. Furthermore, we will experiment with varying the number of training samples to identify the point at which the embedding layer outperforms the pre-trained word embedding. This multifaceted investigation aims to provide insights into the nuanced interactions of these factors and their influence on the model's overall effectiveness.

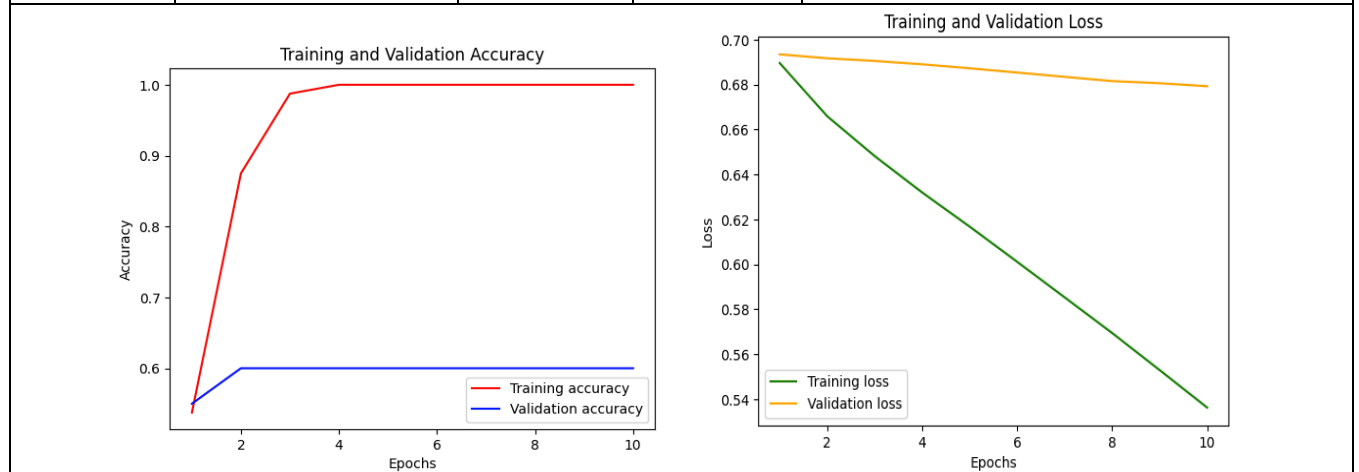
Results

Trained Models with Different Sample Sizes:

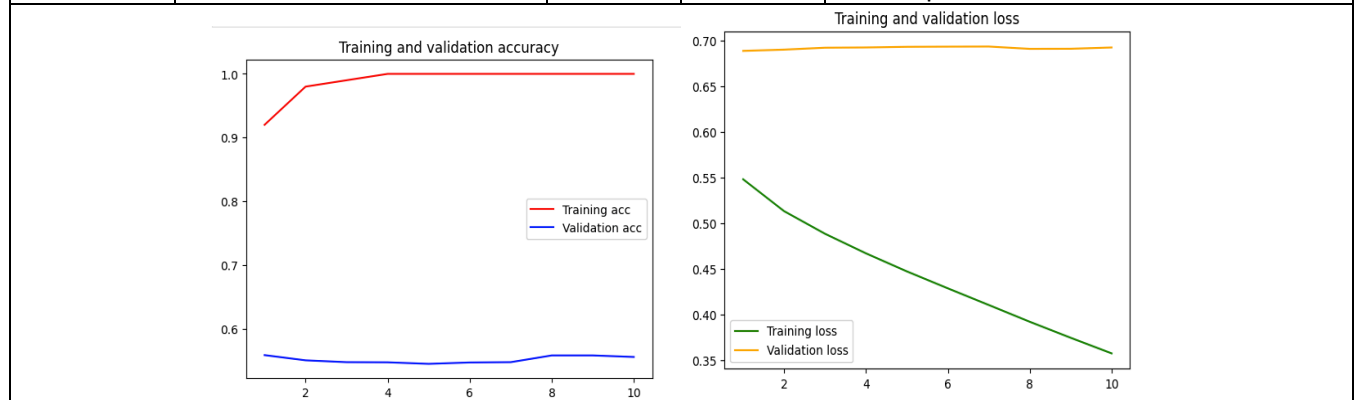
The model underwent training with varying sample sizes. Employing diverse techniques including one-hot encoded sequence, embedded, embedded masked, and pre-trained methods, the accuracy and loss were computed using test data. The recorded observations are summarized in the table below:

Model	Step	Loss	Accuracy	Observation																																				
Model 1	Training a model with a constraint on the review length, limiting it to a maximum of 150 words.	34.8%	86.5%	The step exhibits a high accuracy rate, yet there remains room for improvement, evident from the observed loss percentage. Further refinement through additional training data and fine-tuning could enhance the model's overall performance.																																				
	<div><div><p>Training and Validation Accuracy</p><table><caption>Training and Validation Accuracy Data</caption><tr><th>Epochs</th><th>Training accuracy</th><th>Validation accuracy</th></tr><tr><td>2</td><td>0.70</td><td>0.82</td></tr><tr><td>4</td><td>0.90</td><td>0.87</td></tr><tr><td>6</td><td>0.93</td><td>0.87</td></tr><tr><td>8</td><td>0.95</td><td>0.86</td></tr><tr><td>10</td><td>0.96</td><td>0.86</td></tr></table></div><div><p>Training and Validation Loss</p><table><caption>Training and Validation Loss Data</caption><tr><th>Epochs</th><th>Training loss</th><th>Validation loss</th></tr><tr><td>2</td><td>0.60</td><td>0.45</td></tr><tr><td>4</td><td>0.25</td><td>0.30</td></tr><tr><td>6</td><td>0.18</td><td>0.31</td></tr><tr><td>8</td><td>0.15</td><td>0.33</td></tr><tr><td>10</td><td>0.12</td><td>0.35</td></tr></table></div></div>				Epochs	Training accuracy	Validation accuracy	2	0.70	0.82	4	0.90	0.87	6	0.93	0.87	8	0.95	0.86	10	0.96	0.86	Epochs	Training loss	Validation loss	2	0.60	0.45	4	0.25	0.30	6	0.18	0.31	8	0.15	0.33	10	0.12	0.35
Epochs	Training accuracy	Validation accuracy																																						
2	0.70	0.82																																						
4	0.90	0.87																																						
6	0.93	0.87																																						
8	0.95	0.86																																						
10	0.96	0.86																																						
Epochs	Training loss	Validation loss																																						
2	0.60	0.45																																						
4	0.25	0.30																																						
6	0.18	0.31																																						
8	0.15	0.33																																						
10	0.12	0.35																																						

Model 2	Training the embedded sample with a dataset size of 10,000 and a training sample of 100.	69.5%	50.1%	The adjustment in the number of training samples was made to determine the point at which the embedding layer surpassed the capabilities of the pre-trained word embedding. Notably, the embedding layer exhibited superior performance with 10,000 training samples, achieving a test accuracy of 59% in contrast to a test loss of 66%.
---------	--	-------	-------	---

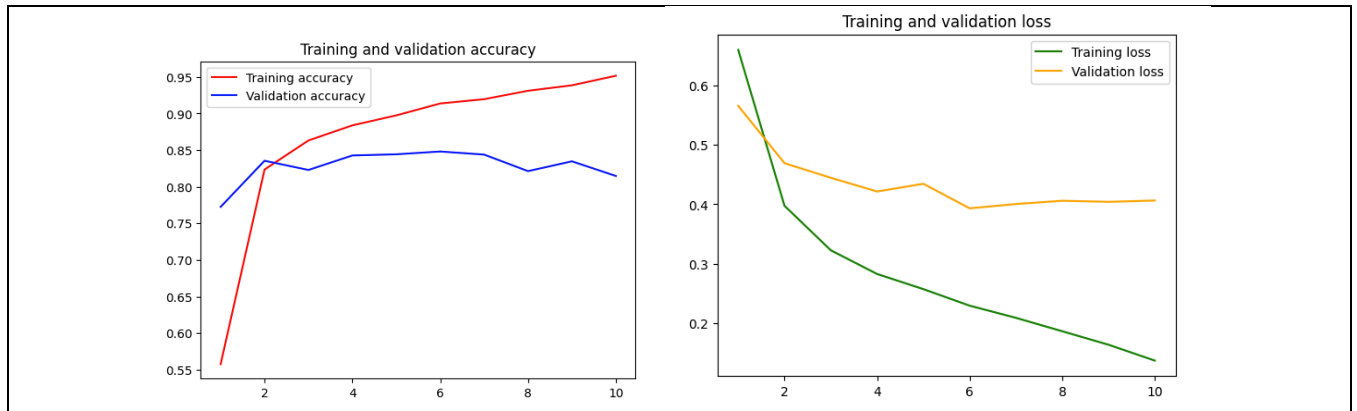


Model 3	Implemented Glove word embeddings while considering only the top 10,000 samples.	70.9	35.7	By implementing Glove word embeddings and limiting the analysis to the top 10,000 samples, the model achieved a test loss of 70.9% with a corresponding test accuracy of 35.7%. These results suggest a need for further investigation or potential adjustments to improve the overall performance of the model.
---------	--	------	------	--



Model 4	Enhancing Glove word embeddings Model Performance via Evaluation with an Increased Training Sample Size.	68.2%	55.89	Increasing the training sample size resulted in a notable improvement in model performance. The test accuracy increased to 68.2%, while the test loss decreased to 55.89%. This emphasizes the
----------------	--	-------	-------	--

				significance of a larger training dataset in refining the model's predictive capabilities.
<div> <div> <p>Training and validation accuracy</p> </div> <div> <p>Training and validation loss</p> </div> </div>				
Evaluating the performance of embedding layers by varying the sample size, while leveraging pretrained word embeddings				
Model 5	Training the embedded sample with a dataset size of 20,000.	42.1%	81.3%	Leveraging both conv1D and embedding layers while increasing the size of the training sample can enhance the model's efficacy by providing a more diverse set of data for learning.
<div> <div> <p>Training and validation accuracy</p> </div> <div> <p>Training and validation loss</p> </div> </div>				
Model 8	Training the embedded sample with a dataset size of 30,000.	42.7%	80%	This aids the algorithm in better comprehending the subtleties of different reviews, leading to more accurate results.



Interpretation

Based on the analysis, employing RNNs on the IMDB data exhibited notable performance, particularly in terms of both test loss and test accuracy, when utilizing embedded layers compared to alternative word embedding techniques. As the sample size increased from 1000 to 30,000, the RNN model demonstrated improved performance, with a corresponding increase in test accuracy. This underscores the positive correlation between larger sample sizes and enhanced model performance, as the model benefits from a more extensive dataset for learning.

A specific comparison between standard embedded and masked embedded layers revealed that the standard embedded layer outperformed in terms of test accuracy. The masking technique, designed to focus solely on actual word embeddings, did not exhibit a discernible impact on the IMDB dataset.

Additionally, when incorporating pretrained word embeddings, GloVe embeddings proved to yield a superior and more effective model compared to training the embedded layer from scratch. This suggests that leveraging pre-existing knowledge in the form of GloVe embeddings contributes to the overall effectiveness of the model, surpassing the performance achieved through training the embedded layer independently.

Conclusion

The model's performance is intricately linked to the volume of data encountered during training. An increased training dataset invariably results in improved performance, as the model gains more information to learn from, enhancing its ability to generalize well on unseen data.

Adapting the model architecture and sample size based on the specific task and requirements is crucial. Experimentation with different configurations is necessary to identify the optimal size for a given scenario. In this study, both masked and standard embedded layers, as well as GloVe embeddings, demonstrated enhanced performance across various embedding techniques.

Notably, the GloVe pre-trained embeddings consistently outperformed other models in terms of accuracy and loss, regardless of sample size. This suggests that GloVe embeddings are particularly efficient for sentiment analysis tasks, capturing comprehensive semantic and syntactic information. Their advantages

include reducing the need for extensive training data, providing a standardized representation, and being easy to implement, making them a preferred choice for sentiment analysis applications.

Moreover, the assignment highlights crucial considerations in the decision-making process between embedding layers and pre-trained word embeddings:

- The use of embedding layers or pre-trained word embeddings can enhance deep learning models for text and sequence tasks, such as sentiment analysis or language translation.
- The decision between the two methodologies is influenced by several criteria, including the amount and quality of the training data, domain-specific vocabulary, and task constraints.
- While pre-trained embeddings based on big corpora can offer a broad representation of words, embedding layers can adapt to the specific job and learn domain-specific characteristics from training data.
- Furthermore, utilizing embedding layers can assist in reducing overfitting and increasing the generalization capacity of the model.
- The neural network learns to discover patterns and correlations in data by modifying its parameters to minimize the gap between expected and goal outputs. Additionally, by limiting overfitting, the model can avoid becoming overly specialized to the pre-training data when exposed to higher samples.