**Assignment -4**

**Advanced Machine Learning**

**Title: Comparative Analysis of Sentiment Analysis Models Using IMDb Movie Reviews Datase**

**Introduction:**

This project uses the IMDB movie reviews dataset to analyze sentiment using TensorFlow and Keras. The process includes preprocessing the data, building a deep learning model, training it on a portion of the dataset, and assessing its performance using validation data. To possibly enhance model performance by utilizing word semantic relationships, the project also investigates the integration of pre-trained GloVe embeddings. The goal of this investigation is to demonstrate how important model architecture design and the use of pre-trained embeddings are to improving sentiment analysis accuracy.

**Experimental Configuration:**

The experiment utilizes the IMDB dataset, which comprises movie reviews annotated with binary sentiment labels. To ensure consistency, reviews are truncated to a maximum length of 150 words. Moreover, only the top 10,000 most frequent words are considered for analysis, simplifying the vocabulary while retaining critical information.

For training efficiency, the experiment is conducted on a limited subset of 100 samples, expediting the iterative process of model development and evaluation. To ensure the reliability of results, model performance is rigorously assessed on a separate set of 10,000 samples for validation. This approach allows for robust evaluation, providing insights into the model's ability to generalize to unseen data and accurately predict sentiment across a diverse range of movie reviews.

**Models:**

In this experiment, two distinct models are utilized for sentiment analysis on the IMDB dataset. Model 1 is constructed with a conventional architecture, comprising an Embedding layer followed by a Bidirectional LSTM layer and a Dense layer. This model learns embeddings from scratch during training, capturing the contextual relationships between words directly from the dataset. In contrast, Model 2 leverages pre-trained GloVe word embeddings to initialize the Embedding layer. By incorporating these pre-existing embeddings, Model 2 enhances its ability to capture nuanced semantic meanings from the text data. Following the Embedding layer, both models include a Bidirectional LSTM layer to capture sequential dependencies effectively and a Dense layer for final sentiment classification. This experimental setup allows for a comparative analysis of the performance between a model trained from scratch (Model 1) and a model leveraging external linguistic knowledge through pre-trained embeddings (Model 2).

The training parameters for both models are standardized to ensure consistency and facilitate a fair comparison. Each model is trained for 10 epochs, with a batch size of 32 samples per iteration. The RMSprop optimizer is employed to optimize the model's parameters during training, adjusting the learning rate adaptively to accelerate convergence. For evaluating the model's performance and guiding the optimization process, binary crossentropy is utilized as the loss function. By adhering to these predefined training parameters, the experiment aims to systematically assess the effectiveness of different model architectures and embedding strategies in sentiment analysis on the IMDB dataset.

**Training configurations:**

The training parameters for both models are standardized to ensure consistency and facilitate a fair comparison. Each model is trained for 10 epochs, with a batch size of 32 samples per iteration. The RMSprop optimizer optimizes the model's parameters during training, adjusting the learning rate adaptively to accelerate convergence. Binary cross entropy is utilized as the loss function to evaluate the model's performance and guide the optimization process. By adhering to these predefined training parameters, the experiment aims to systematically assess the effectiveness of different model architectures and embedding strategies in sentiment analysis on the IMDB dataset.

**Results:**

**A model that includes an Embedding Layer**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Epoch** | **Training Loss** | **Training Accuracy** | **Validation Loss** | **Validation Accuracy** |
| 1 | 0.6906 | 0.61 | 0.6935 | 0.5027 |
| 2 | 0.6863 | 0.58 | 0.6944 | 0.5027 |
| 3 | 0.6818 | 0.58 | 0.6954 | 0.5027 |
| 4 | 0.6784 | 0.58 | 0.6965 | 0.5027 |
| 5 | 0.6743 | 0.58 | 0.696 | 0.5027 |
| 6 | 0.6711 | 0.58 | 0.6969 | 0.5027 |
| 7 | 0.6648 | 0.59 | 0.7001 | 0.5027 |
| 8 | 0.6555 | 0.58 | 0.7012 | 0.5027 |
| 9 | 0.6462 | 0.59 | 0.7016 | 0.5027 |
| 10 | 0.6325 | 0.59 | 0.697 | 0.5032 |

**A model utilizing pre-trained word embedding**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Epoch** | **Training Loss** | **Training Accuracy** | **Validation Loss** | **Validation Accuracy** |
| 1 | 0.6968 | 0.47 | 0.704 | 0.5043 |
| 2 | 0.6803 | 0.58 | 0.7231 | 0.5036 |
| 3 | 0.6704 | 0.58 | 0.7044 | 0.5056 |
| 4 | 0.6524 | 0.6 | 0.7216 | 0.5028 |
| 5 | 0.6473 | 0.59 | 0.7097 | 0.5035 |
| 6 | 0.6404 | 0.6 | 0.7247 | 0.5045 |
| 7 | 0.6368 | 0.6 | 0.7085 | 0.5057 |
| 8 | 0.6247 | 0.7 | 0.739 | 0.5027 |
| 9 | 0.6207 | 0.59 | 0.702 | 0.5091 |
| 10 | 0.6142 | 0.75 | 0.7052 | 0.5051 |

**Conclusion:**

The provided code exemplifies two contrasting approaches to sentiment analysis using the IMDb dataset: one employing randomly initialized word embeddings and the other leveraging pre-trained GloVe embeddings. Through the comparison of these approaches, several noteworthy conclusions can be drawn. Firstly, while the randomly initialized embeddings serve as a foundational baseline, they inherently lack the semantic depth and context captured by pre-trained embeddings. Secondly, the utilization of pre-trained GloVe embeddings presents distinct advantages, including the encapsulation of semantic relationships from a vast corpus, thereby offering richer representations for sentiment analysis tasks. Furthermore, the pre-trained embeddings facilitate faster convergence and enhanced generalization, particularly beneficial when working with limited training data, as observed in the IMDb dataset. By freezing the pre-trained embedding layer, the model preserves the acquired semantic knowledge while allowing subsequent layers to adapt to the task at hand. Ultimately, the comparison underscores the critical role of pre-trained embeddings in augmenting model performance and efficiency, underscoring their relevance in various natural language processing applications, where harnessing contextual understanding from extensive text corpora can significantly enhance predictive capabilities.