Abhinav Reddy Yelipeddi
811300807

Assignment 1
Summary

According to 50,000 reviews in the IMDB database. The word count of these reviews is a total of 10,000highlighted. Below are the methods I used to determine the best combination of parameters to improve test accuracy and validation.

In this report, we describe the steps required to build and run a deep learning model using Keras, integrated with TensorFlow. We start by importing the necessary libraries: OS for file and directory operations, numpy for arithmetic and arrays, and pandas for data manipulation and analysis. To simplify data visualization, we use matplotlib.pyplot. Deep learning features are available through TensorFlow, especially the Keras API, which allows us to build and compile models. The main components imported from Keras are models, layers, corrections, modifiers, loss, and statistics, which provide the necessary tools for model architecture and evaluation. Additionally, we include tools for indexing and loading the IMDb dataset for sentiment analysis. This architectural approach ensures a clear and efficient implementation of our deep learning framework.

After loading the IMDB Dataset:
The dataset is divided into two parts: the training part and the test part. There are 25,000 comments per episode. Then the data is filtered, because we can't put integers into neural networks. The additional training set is divided into sub-training sets and validation sets into two parts.

Model Architectures with Varying Hidden Layers and Units

Combination	Training Accuracy	Validation Accuracy	Test Accuracy
1 Hidden Layer	0.9450	0.8894	0.8824
2 Hidden Layers	0.9983	0.8696	0.8582
3 Hidden Layers	0.9984	0.8440	0.8326

- A model with a hidden layer is better than others for validation and test accuracy.
- The more number of hidden layers leads to overfitting.

The following models are constructed using a single hidden layer, the relu activation function, the rmsprop optimizer, and the binary loss function. The number of hidden units is now 32 and 64. The results are shown in the table below.

Model with the Different Units

Combination	Training Accuracy	Validation Accuracy	Test Accuracy
16 Hidden Units	0.9450	0.8894	0.8824
32 Hidden Units	0.9925	0.8731	0.8615
64 Hidden Units	0.9906	0.8698	0.8591

- 16 Hidden Units is the best as it avoids the Overfitting.

Increasing the hidden units has the same problem of filling and increasing the hidden layers. The best combination to use 16 units is hidden because it is so precise. Convert the loss function to mse and compare it to the binary cross entropy to evaluate the accuracy now. These models have a hidden layer, 16 hidden units, ReLU activation function and rmsprop optimizer.

Loss Function Comparison

Combination	Training Accuracy	Validation Accuracy	Test Accuracy
Binary Cross Entropy	0.9450	0.8894	0.8824
MSE	0.9830	0.8782	0.8701

- Binary Cross Entropy suits best for this Dataset.

Another model is developed using one hidden layer, 16 hidden units, rmsprop optimizer and binary loss function to compare the efficiency of tanh and relu optimization functions.

Activation Function Comparison

Combination	Training Accuracy	Validation Accuracy	Test Accuracy
ReLu	0.9450	0.8894	0.8824
Tanh	0.9925	0.8737	0.8637

- Relu activation function performs better than the Tanh

Dropout Regularizer

For the next model, I opted the Drop out (0.5). In this model, I used one hidden layer, 16 hidden units, rmsprop optimizer, Binary cross entropy loss function and relu activation function.

Combination	Training Accuracy	Validation Accuracy	Test Accuracy
Without Dropout	0.9450	0.8894	0.8824
Dropout(0.5)	0.9772	0.8844	0.8753

- Using an dropout makes it less accurate, suggesting that may not be appropriate in this case.

Regularizer Comparison

Combination	Training Accuracy	Validation Accuracy	Test Accuracy
No Regularizer	0.9450	0.8894	0.8824
L1 Regularizer	0.9993	0.8759	0.8667
L2 Regularizer	0.9925	0.8798	0.8706

- L1 & L2 don't improve the accuracy much affective in this case.

Optimizer Comparison

Combination	Training Accuracy	Validation Accuracy	Test Accuracy
RMSProp	0.9450	0.8894	0.8824

Conclusion

- The best model configuration: for the IMDB data set, a model with a hidden layer containing 16 hidden units, using the ReLU activation function , the binary cross entropy loss, and the RMSprop optimizer perform well.

- Modelling: Although correctors, moderators and regression methods were investigated, the accuracy of the model was not significantly improved. These techniques are more effective when working with larger datasets.
- Improvements: Getting more data will increase the accuracy of the model.

THANK YOU !!