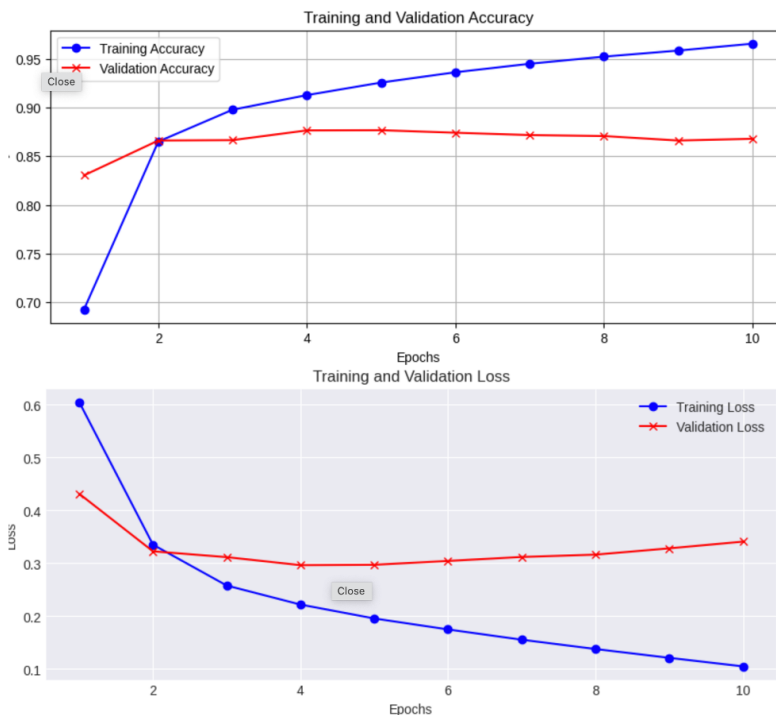


ASSIGNMENT 3

WORD EMBEDDING USING EMBEDDING LAYERS AND PRE-TRAINED MODEL

Our base model uses Keras embedding layers on the IMDB dataset, keeping 10,000 words and cutting reviews at 150 words. It trains and validates with 20,000 samples each for thorough assessment.

```
625/625 [=====] - 2s 2ms/step - loss: 0.6790 - acc: 0.5868 - val_loss: 0.6418 - val_acc: 0.6770
Epoch 2/10
625/625 [=====] - 1s 2ms/step - loss: 0.5618 - acc: 0.7425 - val_loss: 0.5361 - val_acc: 0.7284
Epoch 3/10
625/625 [=====] - 1s 2ms/step - loss: 0.4708 - acc: 0.7850 - val_loss: 0.5053 - val_acc: 0.7438
Epoch 4/10
625/625 [=====] - 1s 2ms/step - loss: 0.4281 - acc: 0.8044 - val_loss: 0.4981 - val_acc: 0.7472
Epoch 5/10
625/625 [=====] - 1s 2ms/step - loss: 0.4009 - acc: 0.8202 - val_loss: 0.4988 - val_acc: 0.7548
Epoch 6/10
625/625 [=====] - 1s 2ms/step - loss: 0.3791 - acc: 0.8332 - val_loss: 0.5012 - val_acc: 0.7532
Epoch 7/10
625/625 [=====] - 1s 2ms/step - loss: 0.3591 - acc: 0.8466 - val_loss: 0.5069 - val_acc: 0.7516
Epoch 8/10
625/625 [=====] - 1s 2ms/step - loss: 0.3405 - acc: 0.8569 - val_loss: 0.5134 - val_acc: 0.7504
Epoch 9/10
625/625 [=====] - 1s 2ms/step - loss: 0.3225 - acc: 0.8670 - val_loss: 0.5215 - val_acc: 0.7504
Epoch 10/10
625/625 [=====] - 1s 2ms/step - loss: 0.3049 - acc: 0.8766 - val_loss: 0.5291 - val_acc: 0.7444
```



Validation Accuracy is 75.48 and the optimal number of Epochs is 10

Considering the embedding layer in model from scratch:

Number of epochs is set to 10.

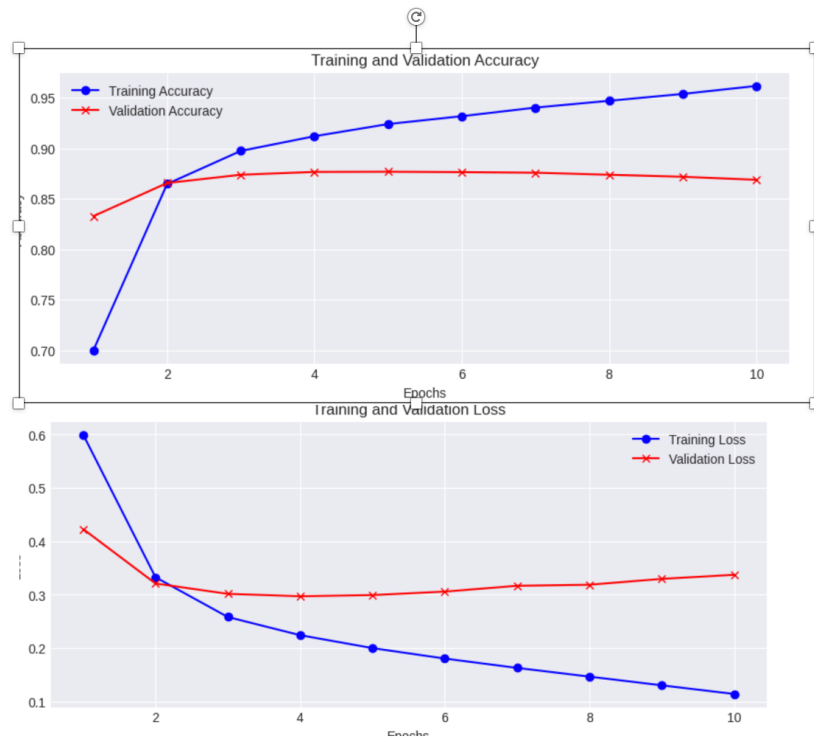
Re-running the example modifying the following:

Cut-off Words	Training Sample	Validation Sample	Word Count	Validation Accuracy
20	10000	10000	10000	74.70
150	100	10000	10000	50.38
100	200	10000	10000	50.40
150	20000	20000	10000	87.11
150	15000	10000	10000	86.84

Pre-trained model with glove embedding and freezing the embedding layer.

Cut-off Words	Training Sample	Validation Sample	Word Count	Validation Accuracy
100	10000	10000	10000	50.1

```
Epoch 1/10
313/313 [=====] - 4s 11ms/step - loss: 0.7132 - acc: 0.4994 - val_loss: 0.6934 - val_acc: 0.4949
Epoch 2/10
313/313 [=====] - 3s 9ms/step - loss: 0.6969 - acc: 0.5048 - val_loss: 0.6936 - val_acc: 0.4960
Epoch 3/10
313/313 [=====] - 3s 10ms/step - loss: 0.6947 - acc: 0.5161 - val_loss: 0.6999 - val_acc: 0.4978
Epoch 4/10
313/313 [=====] - 4s 14ms/step - loss: 0.6782 - acc: 0.5683 - val_loss: 0.7013 - val_acc: 0.5024
Epoch 5/10
313/313 [=====] - 4s 12ms/step - loss: 0.6338 - acc: 0.6362 - val_loss: 0.7549 - val_acc: 0.5021
Epoch 6/10
313/313 [=====] - 3s 11ms/step - loss: 0.5681 - acc: 0.6947 - val_loss: 0.7732 - val_acc: 0.4990
Epoch 7/10
313/313 [=====] - 4s 11ms/step - loss: 0.5059 - acc: 0.7424 - val_loss: 0.8840 - val_acc: 0.4994
Epoch 8/10
313/313 [=====] - 4s 14ms/step - loss: 0.4325 - acc: 0.7973 - val_loss: 0.9206 - val_acc: 0.4945
Epoch 9/10
313/313 [=====] - 4s 12ms/step - loss: 0.3658 - acc: 0.8337 - val_loss: 1.1086 - val_acc: 0.5004
Epoch 10/10
313/313 [=====] - 3s 10ms/step - loss: 0.3023 - acc: 0.8719 - val_loss: 1.1311 - val_acc: 0.5014
```

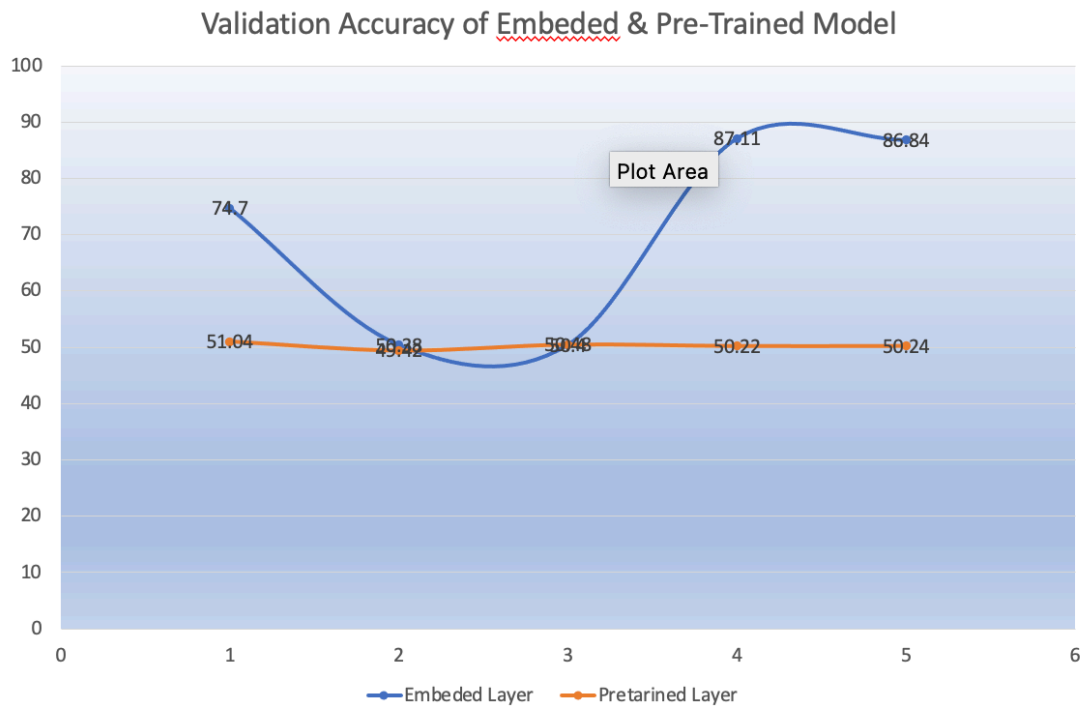


The validation accuracy drops to 50.14% the model overfits.

Validation Accuracy for the pretrained model:

Cut-off Words	Training Sample	Validation Sample	Word Count	Validation Accuracy
20	10000	10000	10000	51.04
150	100	10000	10000	49.42
100	200	10000	10000	50.48
150	20000	20000	10000	50.22
150	15000	10000	10000	50.24

Results & Findings:



Summary:

- The model from the scratch with embedding layer was able to give better accuracy of 50.38 % than the pretrained model with when we have less training samples of 100. The model built from scratch with an embedding layer outperformed the pretrained model with fewer training samples (100) due to potential overfitting in the more complex model.
- Through fine-tuning hyperparameters, specifically adjusting the dimensions of the embedding vectors and implementing regularization techniques such as dropout, the accuracy of the model utilizing an embedding layer peaked to 87.11%.
- when we increase the cut-off words from 20 to 150 there is a significant improvement in the accuracy of the model.