

Natural Language Processing - IMDB Movie Review							
	Description	Hyperparameters	Number of Epochs	Training Loss	Training Accuracy	Test Accuracy	Comments
Part 1a	Given model - Word Embedding Layer + Mean Pooling + Fully Connected Layer + Batch Norm + Relu + Dropout + Output Layer	ADAM optimizer with LR=0.001, BatchSize=200, VocabularySize=8000, HiddenUnits=500	6	0.1444	94.51	86.98	Describe more about the model/results such as why certain hyperparamters were chosen or the effect it had on the accuracy/training time/overfitting/etc.
	Underfit model - Word Embedding Layer + Mean Pooling + Fully Connected Layer + Batch Norm + Relu + Dropout + Output Layer	ADAM optimizer with LR=0.001, BatchSize=200, VocabularySize=8000, HiddenUnits=5	6	0.5226	74.84	78.97	Significantly lower hidden_units would make the model underfit since the model is not able to capture enough features and both training and test accuracy are lower. Training time is only half of the given model as there are fewer parameters to learn.
	Overfit model - Word Embedding Layer + Mean Pooling + 2Fully Connected Layer + 2Batch_norm + 2Relu + Output Layer	ADAM optimizer with LR=0.001, BatchSize=200, VocabularySize=8000, HiddenUnits=1000	6	0.0907	97.19	84.81	Increasing the no_of_hidden_units, removing dropout, and adding a second hidden layer cause the model to overfit(high training accuracy, low test accuracy). More parameters are learned to fit the training data and make it harder to generalize on the test data. Without dropout, the chance to learn independent representation is lower. Also, training time is two times of the given model as there are more parameters to learn.
	Custom 3						
Part 1b	Given Model - (Fully Connected Layer + Batch Norm + Dropout + Relu + Output Layer)	ADAM optimizer with LR=0.001, BatchSize=200, VocabularySize=100000, HiddenUnits=300	6	0.3039	87.45	85.87	Training time is significantly shorter (~0.6s/epoch) since we do not have embedding layer to train.Using GloVe features, the tranning accuracy stops much earlier and the model is not likely to overfit.
	Underfit Model - (Fully Connected Layer + Batch Norm + Dropout + Relu + Output Layer)	ADAM optimizer with LR=0.001, BatchSize=200, VocabularySize=100000, HiddenUnits=3	6	0.4739	76.06	84.99	Significantly lower hidden_units would make the model underfit since the model is not able to capture enough features and training accuracy is much lower.
	Overfit Model - (2Fully Connected Layer + 2Batch Norm + 2Relu + Output Layer)	ADAM optimizer with LR=0.001, BatchSize=200, VocabularySize=100000, HiddenUnits=2000	6	0.2151	90.95	78.04	Increasing the no_of_hidden_units, removing dropout, and adding a second hidden layer cause the model to overfit(high training accuracy, low test accuracy). More parameters are learned to fit the training data and make it harder to generalize on the test data. Without dropout, the chance to learn independent representation is lower. Also, training time is two times of the given model as there are more parameters to learn.
	Custom 3						
Part 2a	Given Model - (Word Embedding Layer + Reset_state + Lstm + Batch Norm + Dropout + MaxPool + Output Layer)	ADAM optimizer with LR=0.001, BatchSize=200, VocabularySize=8000, HiddenUnits=500, Sequence_len=100	20	0.0825	96.98	87.06	LSTM is a overkill for this problem and the model overfits with seq_len=100. The training time is 51.5427s per epoch.
	Underfit Model - (Word Embedding Layer + Reset_state + Lstm + Batch Norm + Dropout + MaxPool + Output Layer)	ADAM optimizer with LR=0.001, BatchSize=200, VocabularySize=8000, HiddenUnits=5, Sequence_len=100	20	0.5657	71.65	70.02	Significantly lower hidden_units would make the model underfit since the model is not able to capture enough features and training accuracy is much lower (~17s/epoch).
	Overfit Model - (Word Embedding Layer + Reset_state + 2Lstm + 2Batch Norm + 2Dropout + MaxPool + Output Layer)	ADAM optimizer with LR=0.001, BatchSize=200, VocabularySize=8000, HiddenUnits=500, Sequence_len=150	20	0.0349	98.9	86.78	Adding a second hidden LSTM layer and larger seq_len(150) cause the model to overfit(high training accuracy, low test accuracy). More parameters and longer dependencies are learned to fit the training data and make it harder to generalize on the test data. Also, training time is much longer(183s/epoch) as there are more paremters to learn
	Custom 3						
Part 2b	Given Model - (Reset_state + Lstm + Batch Norm + Dropout + MaxPool + Output Layer)	ADAM optimizer with LR=0.001, BatchSize=200, VocabularySize=100000, HiddenUnits=500, Sequence_len=100	20	0.2165	91.1	89.65	This is the best result and the difference between training the test accuracy is very small. GloVe features help improve the result. Since we do not have to train the embedding layer, it is unlikely to overfit the tranning time is shorter(41s/epoch)

Underfit Model - (Reset_state + Lstm + Batch Norm + Dropout + MaxPool + Output Layer)	ADAM optimizer with LR=0.001, BatchSize=200, VocabularySize=100000, HiddenUnits=5, Sequence_len=100	20	0.4992	75.73	75.74	Significantly lower hidden_units would make the model underfit since the model is not able to capture enough features and training accuracy is much lower (~24s/epoch).
Overfit Model - (Reset_state + 2Lstm + 2Batch Norm + MaxPool + Output Layer)	ADAM optimizer with LR=0.001, BatchSize=200, VocabularySize=100000, HiddenUnits=500, Sequence_len=150	20	0.0499	98.34	89.98	Adding a second hidden LSTM layer, removing dropout layers and larger seq_len(150) cause the model to overfit(high training accuracy, low test accuracy). More parameters and longer dependencies are learned to fit the training data and make it harder to generalize on the test data. Also, training time is much longer(186s/epoch) as there are more parameters to learn
Custom 3						