**Activation Function**

|  |  |  |
| --- | --- | --- |
| Activation Function | Accuracy | Train Time |
| ReLu | 0.855 | 11.12 |
| Sigmoid | 0.676 | 11.92 |
| Linear | 0.864 | 13.29 |

**Optimizers**

|  |  |  |
| --- | --- | --- |
| Optimizer | Accuracy | Train Time |
| Mini Batch GD (Baseline) | 0.856 | 13.69 |
| Mini Batch GD + Nesterov Momentum | ­0.864 | 13.47 |
| Adagrad | 0.872 | 11.29 |
| Adam | 0.831 | 12.01 |

**Regularization**

|  |  |  |
| --- | --- | --- |
| Regularization | Accuracy | Train Time |
| Baseline | 0.856 | 12.61 |
| Dropout | 0.787 | 12.41 |
| Batch Norm | 0.864 | 15.31 |
| Weight Initializer | 0.822 | 11.96 |

**Personal Observation:**

Before analyzing the observations, I need to state that each dataset has different characteristics when responding to different hyper-parameter in the neural network. In other words, optimizer might affect one training time on dataset A but might affect the accuracy on dataset B. Therefore, the following analyzing focus only on the dataset chosen for the assignment, the IMDB dataset.

For accuracy, sigmoid activation and dropout layer relatively decrease the accuracy of the model compared to other setting. For dropout layer, a possible reason is that I set the dropout rate to 0.75, it may cause the inaccurate model due to the lack of neurons in the network.

For the training time, Adagrad optimizer brings the shortest training time while Batch Norm layer regularization causes the longest training time. Otherwise, in general speaking, the training time of each setting doesn’t vary drastically, they are all in the range around 12 seconds.