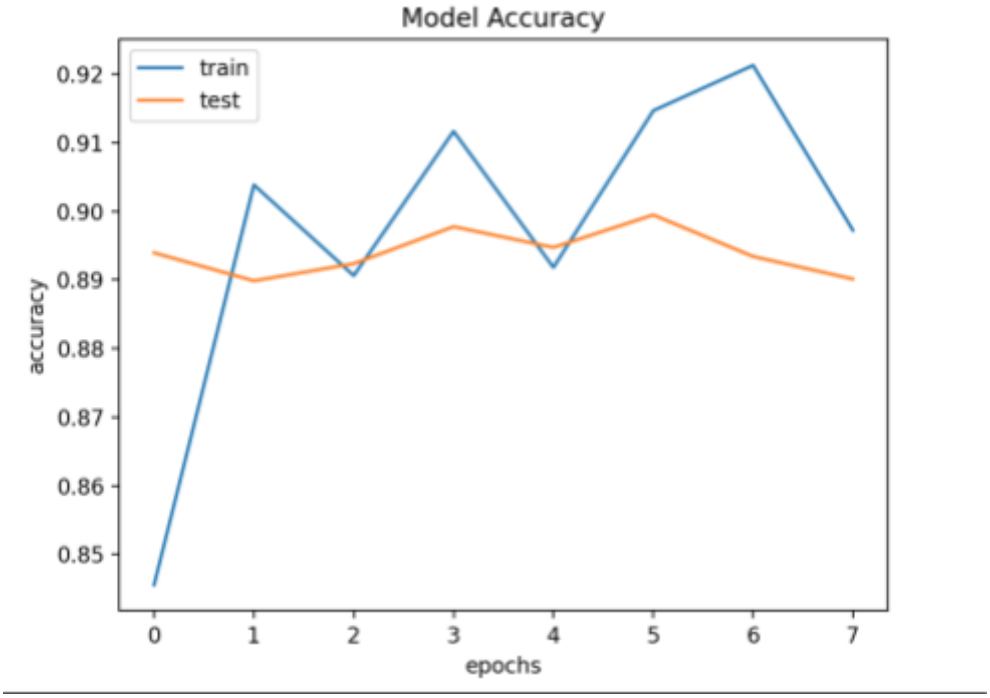
In this project, I used a pretrained model to create a state of the art classifier on the AG News dataset. The AG News dataset is a collection of more than 1 million news articles. Each of these articles are labeled based on what kind of news they are. And in this dataset, there are four kinds: “World”, “Business”, “Sports”, and “Sci/Tech”. The AG News dataset is often used as a benchmark for text classification models. Currently, the state of the art performance on this dataset ranges from roughly 86% to 93%.

In order to achieve similar performance, I decided to use the BERT model which is a transformer-based language model that generates a 768 length embedding vector from each token. These embedded vectors are highly condensed and readily have the relationships between different tokens built into them. As a result, we can apply a very simple classification model on top of BERT to get a model that achieves state of the art performance.

For my classification model, I simply stacked two dense layers on top of the BERT model with the last layer being 4 nodes wide. I originally also used a dropout layer between the two layers but I found that it didn’t really improve performance on the validation set and only made the model take longer to converge.



I used the validation set to find the optimal hyperparameters and after training my model for eight epochs, I was able to achieve an 89% accuracy on the test set. The figure to the right shows the accuracy of my model on both the training and test set as a function of the number of epochs. What’s interesting is that the model is able to achieve that peak 89% accuracy within a single epoch. This shows that by using the rich embeddings developed by the BERT model, a simple shallow MLP is very easily able to pick up on the relationships between tokens and use them for complex NLP tasks like text classification.