CS 474 Final Write Up

For my final project, I decided to do some work on the project for the big data capstone class that I am also enrolled in. The project is from Proctor & Gamble, and they would like us to look at possible ways to detect anomalous behavior in the electric servo motors on their production lines, and then see if we can build models that will predict their failure from these anomalies, or possible directly from the raw sensor data. This would be very valuable to P&G, because unanticipated stoppage of a production line results in far more downtime than planned stoppages. If the failed motor does not have a replacement in stock, this downtime extends even further since the replacement part mush be shipped. For this project, I did some work towards this goal on my own (work on the project doesn't begin until the second semester) and focusing only on deep learning. The data consists of sensor readings at fixed time frequencies, although it can vary between datasets. To help better understand the data and its context, my group and I had a web conference with production engineers at P&G's headquarters, and then later traveled to their production facility in Utah to observe the processes behind the data and ask employees and operators questions about both the data and the manufacturing processes behind them. It's a lot of data, and it isn't labeled in the data where servo motors fail. They do have a dataset of those failures, but it's separate from the sensor data.

Since the challenge of integrating this data will exceed the scope of this project, after understanding the data's nature I looked for a drop-in replacement dataset for the purposes of this project. I was able to find one in a NASA repository regarding the readings of sensors on wheel bearings, which were being turned at a very close to constant rate until at least one bearing failed. This dataset is similar enough that I should be able to swap out the data files and not have to change much else. It's also from a similar context. In addition to the miracle of finding this very similar dataset while reading papers about anomaly detection with machine learning, I also found a paper where someone had performed analyses on one subset of this dataset and had determined approximately when the anomalous behavior began. This allowed me to drop the data right in, requiring no cleaning or mangling.

I then spent the rest of the time thinking of possible models to identify anomalies or predict failures, thinking of different variations of those models, and creating/debugging/testing them. In the end, the number of models I thought of far exceeded the number I was able to actually get working before my time ran out. I started with anomaly detection, since I figured it may be advantageous to have those working to feed data into the failure predictors (since anomalies may not necessarily mean that something is about to fail). In the end, I didn't even get through all of the ideas for anomaly detectors. Most of them were some kind of encoder-decoder network, although one of them was also a plain linear network. The idea of the encoder-decoder network is that if the normal data behaves fairly consistently, then the network should be able to compress the information down into fewer nodes and then reconstruct a decent approximation of it out the other end. This means that normal data would incur a low loss going through the model, while anomalous data would have measurably higher loss as the model attempted to reconstruct it. The linear model required labeled data and would simply classify data as normal or anomalous. The failure predictors I would have gotten to would have been (at this point at least, since I came up these ideas over time) to have some sort of LSTM network, a GRU network, and some other simpler models looking only at the anomaly data fed in from the detectors.

In the end, the anomaly detectors which I finished performed fairly well at their task, with anomalous data being clearly identified by the network. I am happy with the ideas I generated and the groundwork I laid for possible solutions which my team can explore together next semester. I was also careful to build everything to be as flexible as possible with things like data shapes so that different segments of P&G's data can be run through the models with little or no modification.