Week 03

Work done

Spent most of the week doing preprocessing. Zhang et al. mention 5 disparate datasets of biomedical abstracts, but provide no actual links. These datasets are disparate, unstandardized and subjected to link decay. Thankfully, all the data has been centralized at http://mars.cs.utu.fi in an XML format.

Naturally, even this XML dataset required parsing and compiling. Thus, a python XML parser was developed along with a class to capture and rebuilt each abstract to allow a dynamic and flexible dataset suited to the researcher's needs.

After analysis, it became clear that most of the dataset consists of less than 100 word long abstracts, with one exception which was removed. Since protein interactions have been grouped by two, the entire dataset was split into interactions. The tested pair of proteins are concatenated at the beginning of the abstract. In total, about 11000 training examples and 2500 test examples were produced.

After the dataset has been such curated, an attempt at learning the keras embedding layer has been made. The advantage of such a layer means it could be integrated directly into the neural network without additional bells and whistles. It can even be customized with pretrained weights, in our example with the glove word bags of size 100. With a padding, we therefore have 100 word vectors of 100 integers each, or a 100 x 100 matrix, easy to process.

To check the overall functioning of the dataset, it was pipelined through two sequential models:

- an MLP model consisting of multiple dense layers with sigmoid activations
- a convolutional model with 2D convolutions reducing in size

Both models were tail ended with a flatten layer and a dense layer to produce a single output(0 or 1). Even with such primitive architectures trained for 50 epochs, the training scores were 70%.

Problems

1. The glove embedding does not know many of the biomedical specialty words in the corpus of study. In fact, it cannot encode about a third of the dataset. Alternatives exist, such as fast text which encodes by syllables, or scispacy, which has been trained on specialty training. Of course, this requires more preprocessing.
2. The two sequential models are simplistic in nature, and likely won't obtain high results, which is why the researcher hasn't bothered with deep training. A more complex network means more development time, and deviates from the sequential model, resulting in a more complex pipeline.

To do for next week:

Solving some embedding layer issues and attempting to implement Zhang et al model.