Exercise 1 – Report

Antigen Discovery for SARS-CoV-2 (“Corona”) Virus Vaccine

# Data Handling

Each data item is a 9-lengthed sequence of characters, and all of the characters are in a vocabulary of 20 characters. We decided to represent each sequence to 9\*20 vector – concatenation of nine (one for character) one-hot encoding vectors, each one is of size 20. This seems like a natural choice because we do not have prior assumptions about the data, we want to keep the order and we do not know if there is any connection between some of the characters.

# Models Architectures

We will try multi-layered perceptron networks with different parameters. The parameters are numbers of layers, number of neurons in each hidden layer and activation function. In all cases the first layer will get 9\*20-sized input vector and the last layer will output 1-sized output vector after going through sigmoid activation.

# First Training

For the first training, we tried feed forward network with one hidden layer with 128 neurons with relu activation. We trained the network for 5 epochs. We used batch size=1 for simplicity.

Results on the test set:

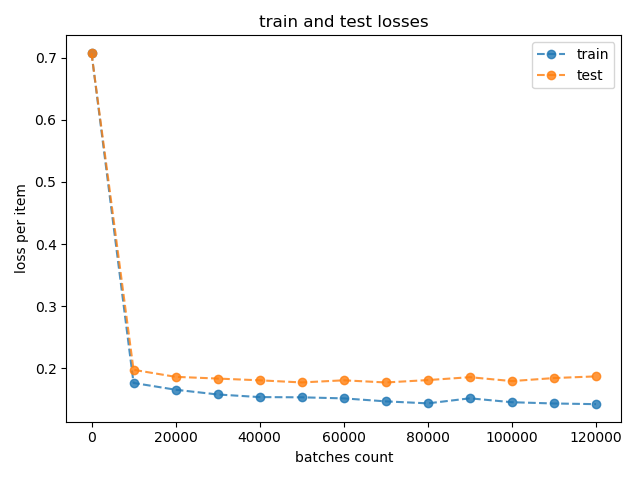
Accuracy: 0.917

Recall: 0.968

Precision: 0.941

f1: 0.954

Those are the learning curves:



According to those results, I decided that comparing the different architectures with 2 epochs for each (about 50,000 batches) would be enough.

## Architectures comparison