Reflection: HW 6

1.
For the Multi-headed attention, The final layer needed to be converted to a probability distribution so I added a flatten layer and a dense layer with n output neurons (46 in our case).

2. Both the models have a very high accuracy with GRU models at 96 – 98% accuracy and MHA models at over 99% accuracy on the test set.

3. The GRU network required a lot of computation where each epoch with 100 steps was taking about 5 – 6 mins. Where as the MHA network required a lot less computation where each epoch was training in 1- 2 mins.

*Please note that on the number of training epochs while training the GRU network, repeat was set to off due to an error with number of steps in the results creation part. It was later fixed in the MHA part. And did not have enough time to retrain all the GRU models. GRU might have taken longer than MHA as sometimes MHA stoppped training at 29 epochs.