# 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. Yes, the GRU model seems to be more stable and consistent in performance compared to the MHA model, based on the training and validation accuracy graphs The MHA model displays more fluctuation

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