Figure 0a (GRU)

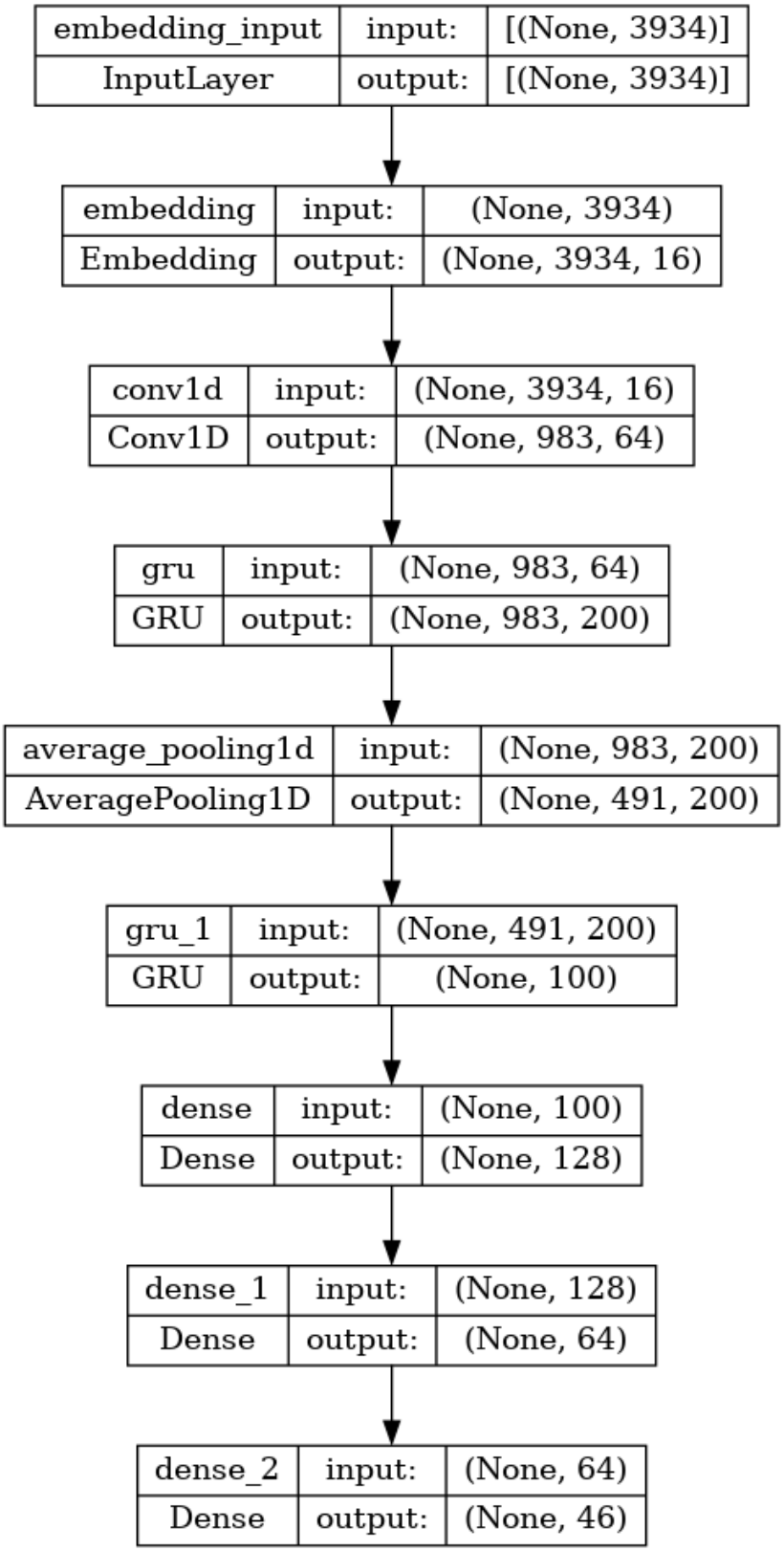


Figure 0b (MHA)

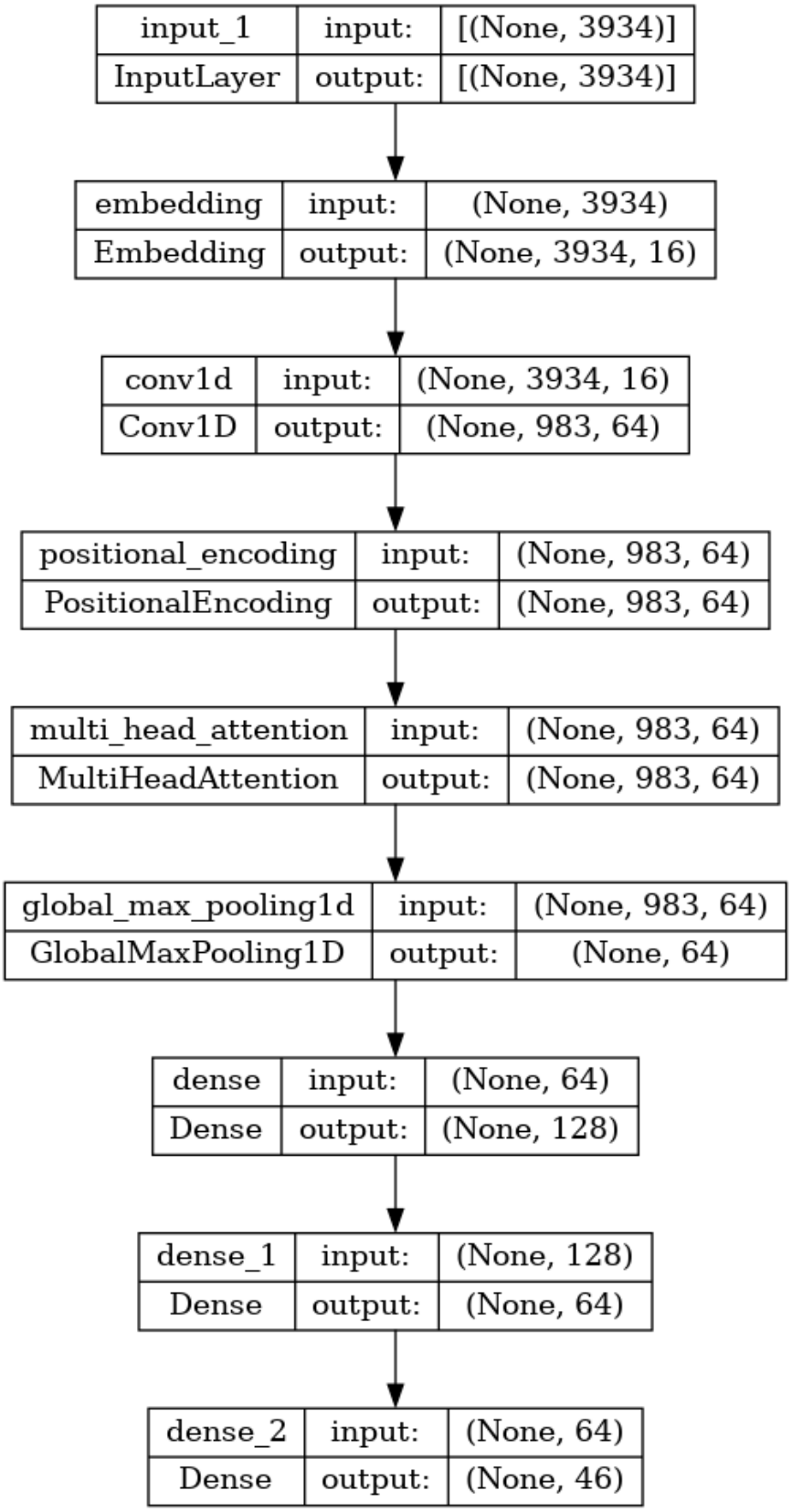


Figure 1

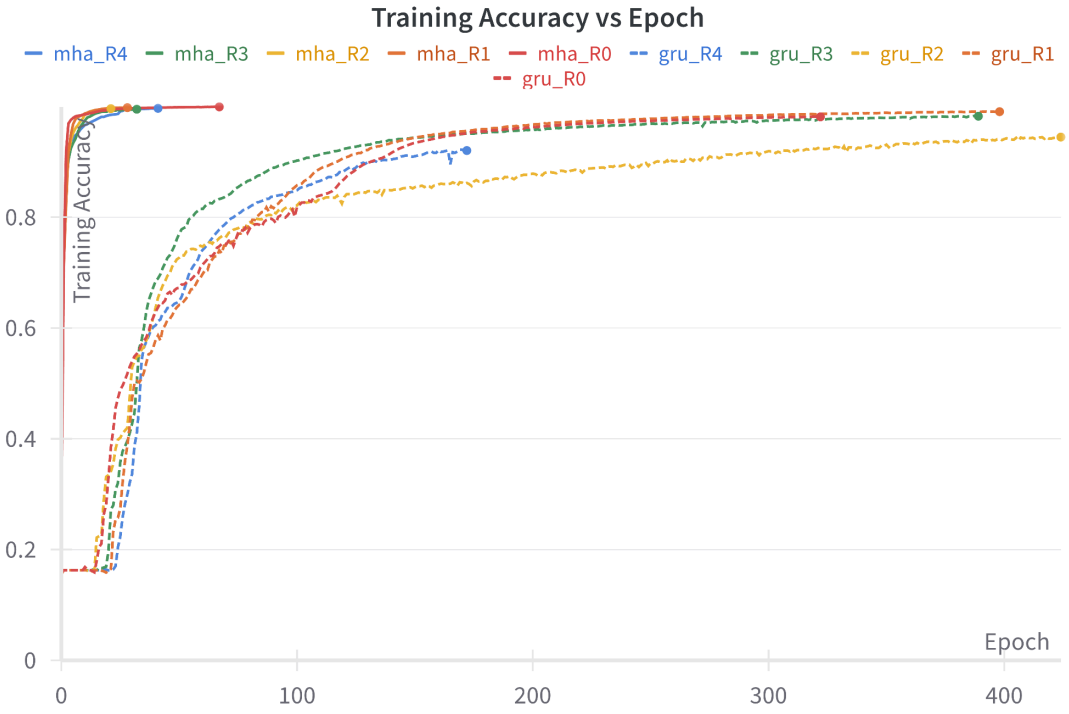


Figure 2

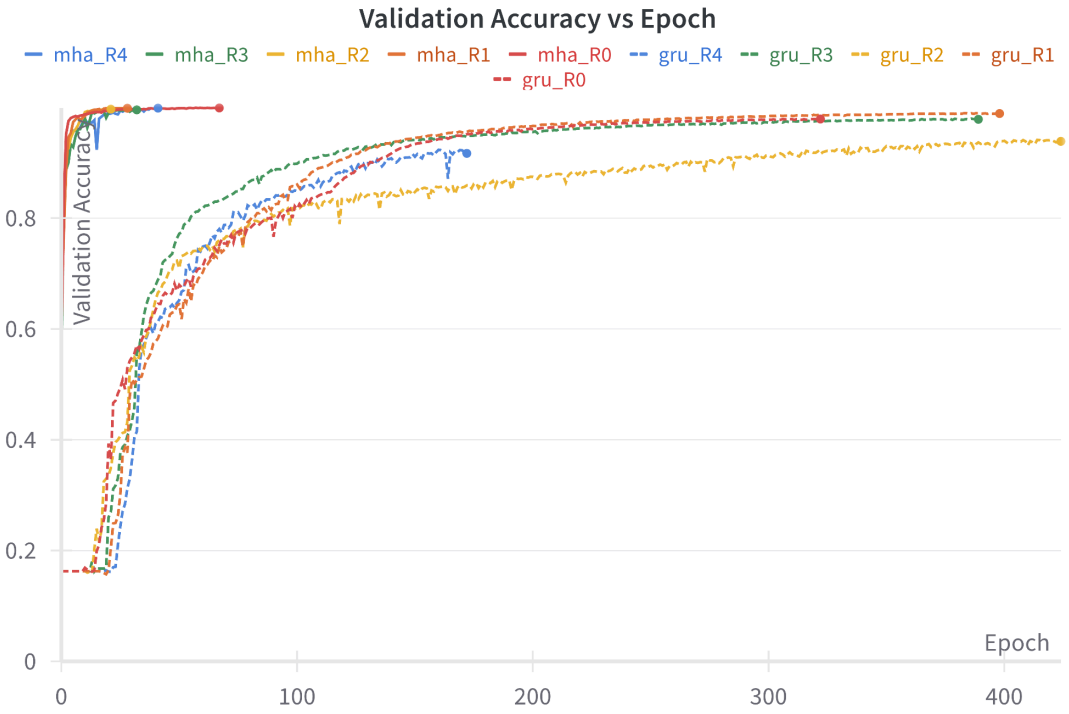


Figure 3

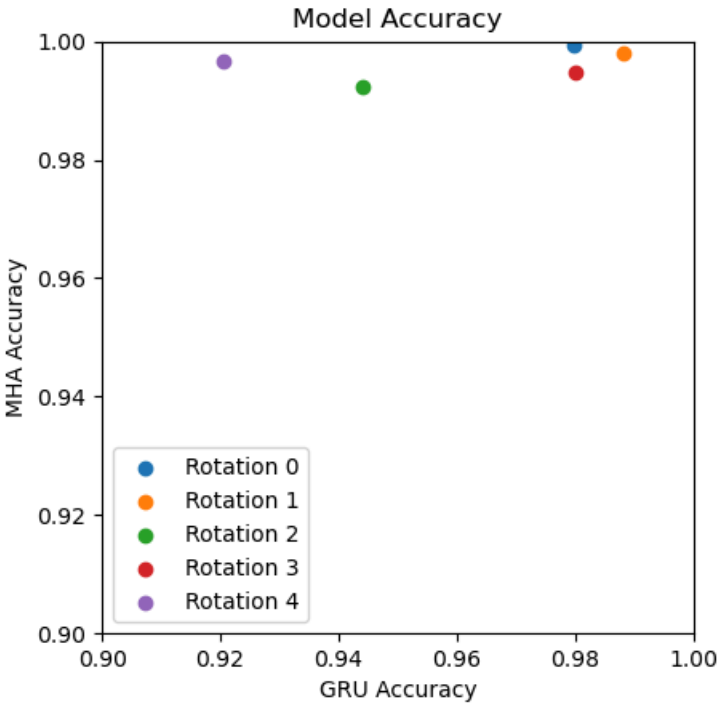
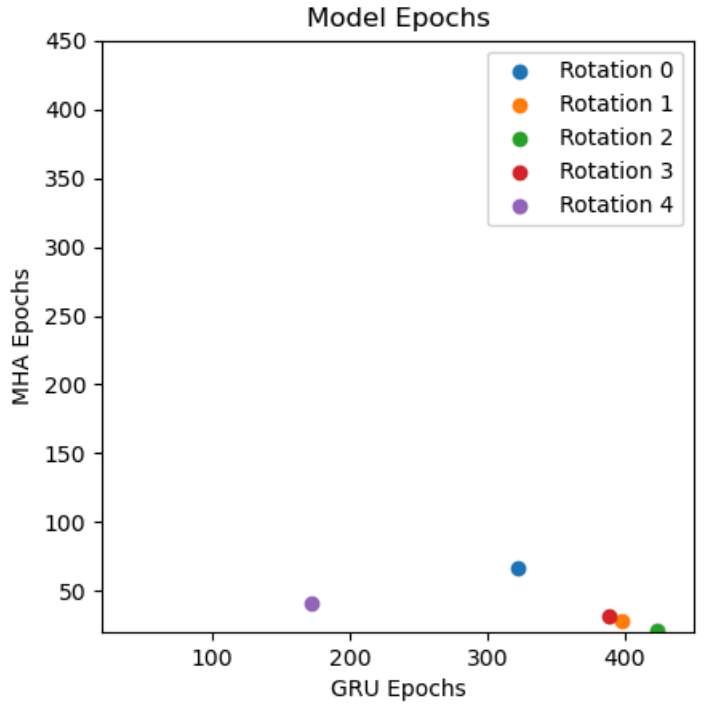


Figure 4



Reflection

1. For your Multi-Headed Attention implementation, explain how you translated your last MHA layer into an output probability distribution.
   I used global max pooling to translate the set of hyper tokens into a single hyper token.

2. Is there a difference in performance between the two model types?
   There is a significant difference in performance. The multihead attention model was able to achieve drastically better performance in a fraction of the number of epochs that the GRU model used.

3. How much computation did you need for the training for each model type in terms of the number of epochs and time?
   The multihead attention models needed around 50 epochs and 50 minutes to train, whereas the GRU models needed around 350 epochs and 1.5 hours to train.