

Appendix A. Glimpse Trajectories

Here we show visualizations of the learned policy for our best models and distribution of attention weights with examples from each class in the MNIST (Figure 6) and cluttered MNIST (Figure 7) test datasets. Columns 1-6 illustrate the trajectory of individual glimpses, which are cropped from the original image, in red from the first timestep on the left up to the point of prediction. The class prediction y is labeled next to the target image t . Note that some samples in Figure 7 are incorrectly classified, and evidently have a poor trajectory. Columns 7-11 visualize the distribution of self-attention weights from the 4 attention heads and their associated mean.

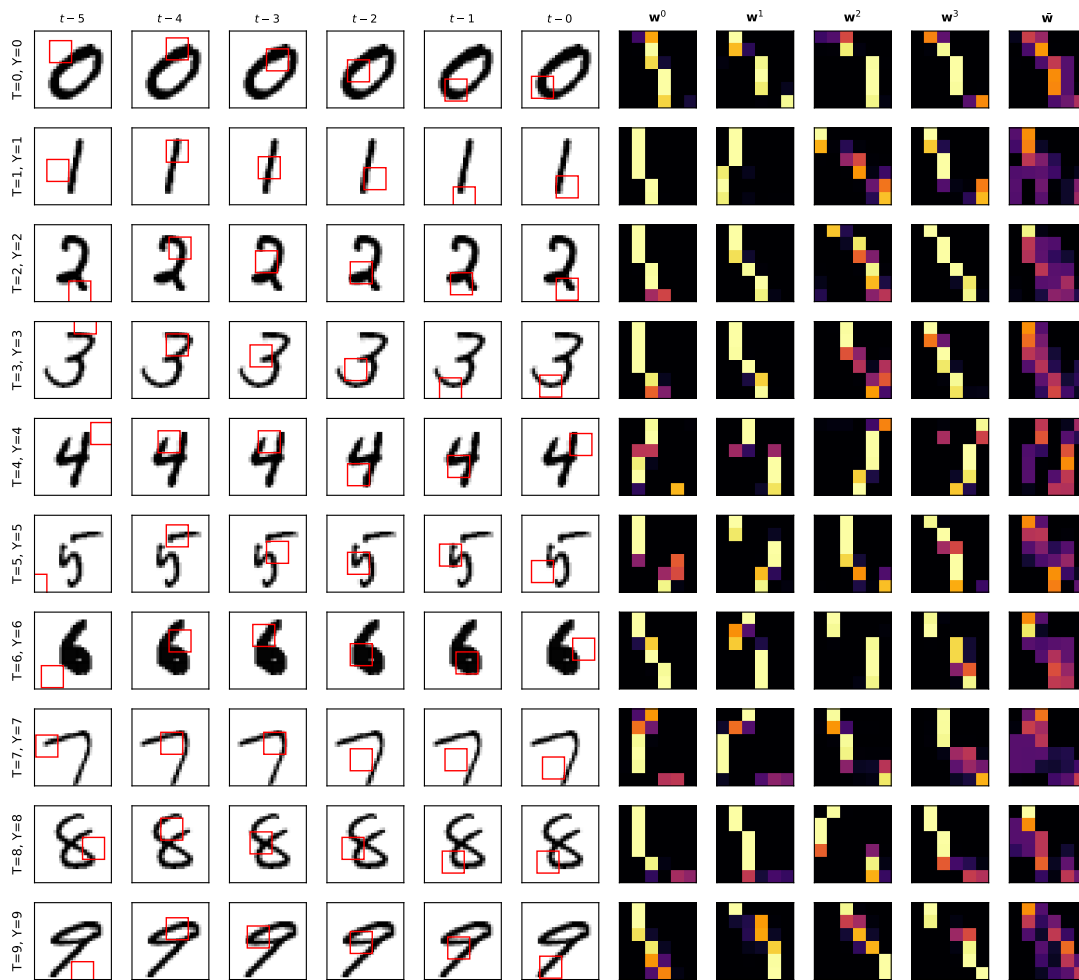


Figure 6: Example trajectories and distribution of attention weights for our best MNIST model.

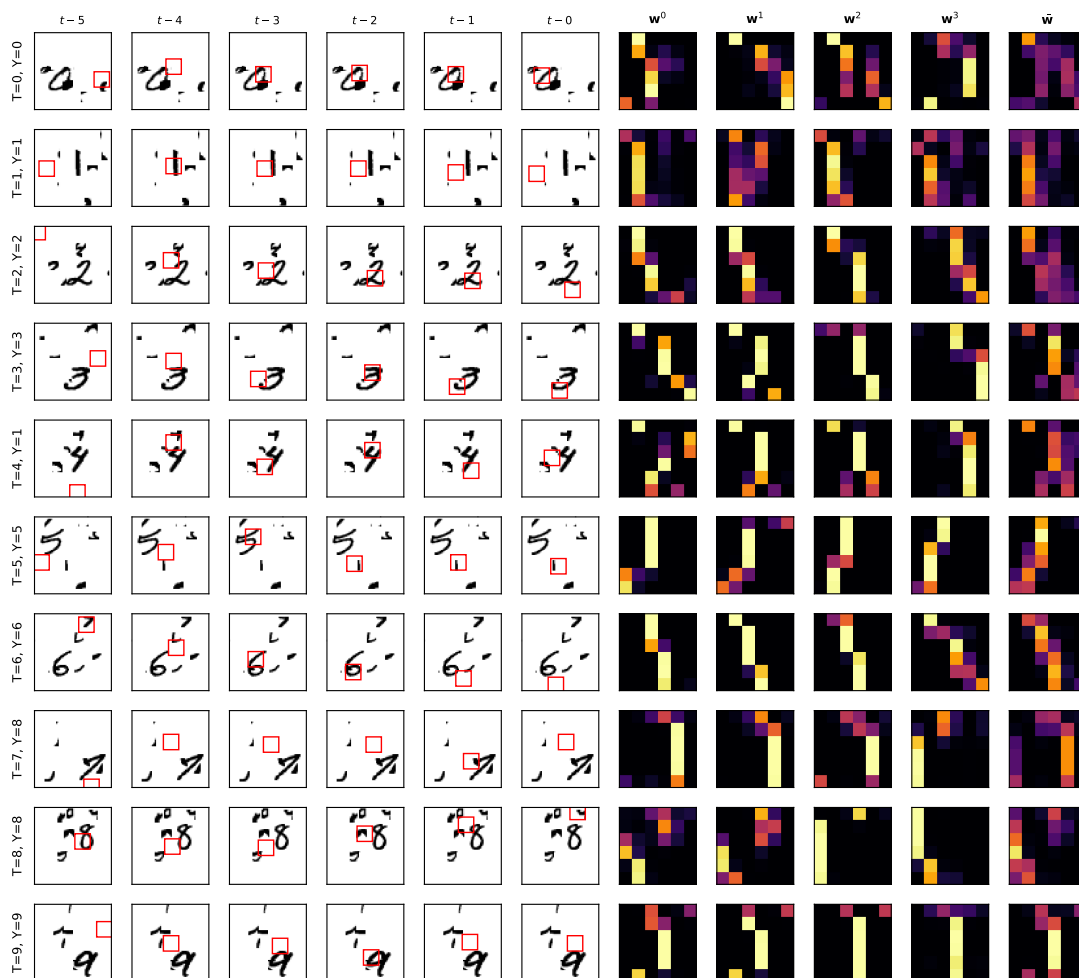


Figure 7: Example trajectories and distribution of attention weights for our best cluttered MNIST model.

Appendix B. Visualizing Input Units

The glimpse network learns the “what” features from a glimpse of the input image. A fully-connected layer takes as input a glimpse before combining it with the “where” features. Our best model on the MNIST data uses 256 input units to represent this fully-connected layer. After we have trained the network, and made updates to the weights in this layer, we can visualize each unit separately. These visualizations are made to better understand, to some degree, how the network represents a glimpse and arrives at its prediction.

In no particular order, we show these units in Figure 8. These units operate on the unstandardized glimpse of MNIST digits with intensity values between $[0, 1]$. The majority of weights are positive (as shown in red), but reveal some interesting patterns. That is, there are strong positive gradients that outline the shape of certain lines and curves at

different rotations and angles. It is unclear why strong negative weights are in some of the corners. We speculate these could be to better inform the location network and it would be interesting to view the correlation of neuron activations and change in location values.

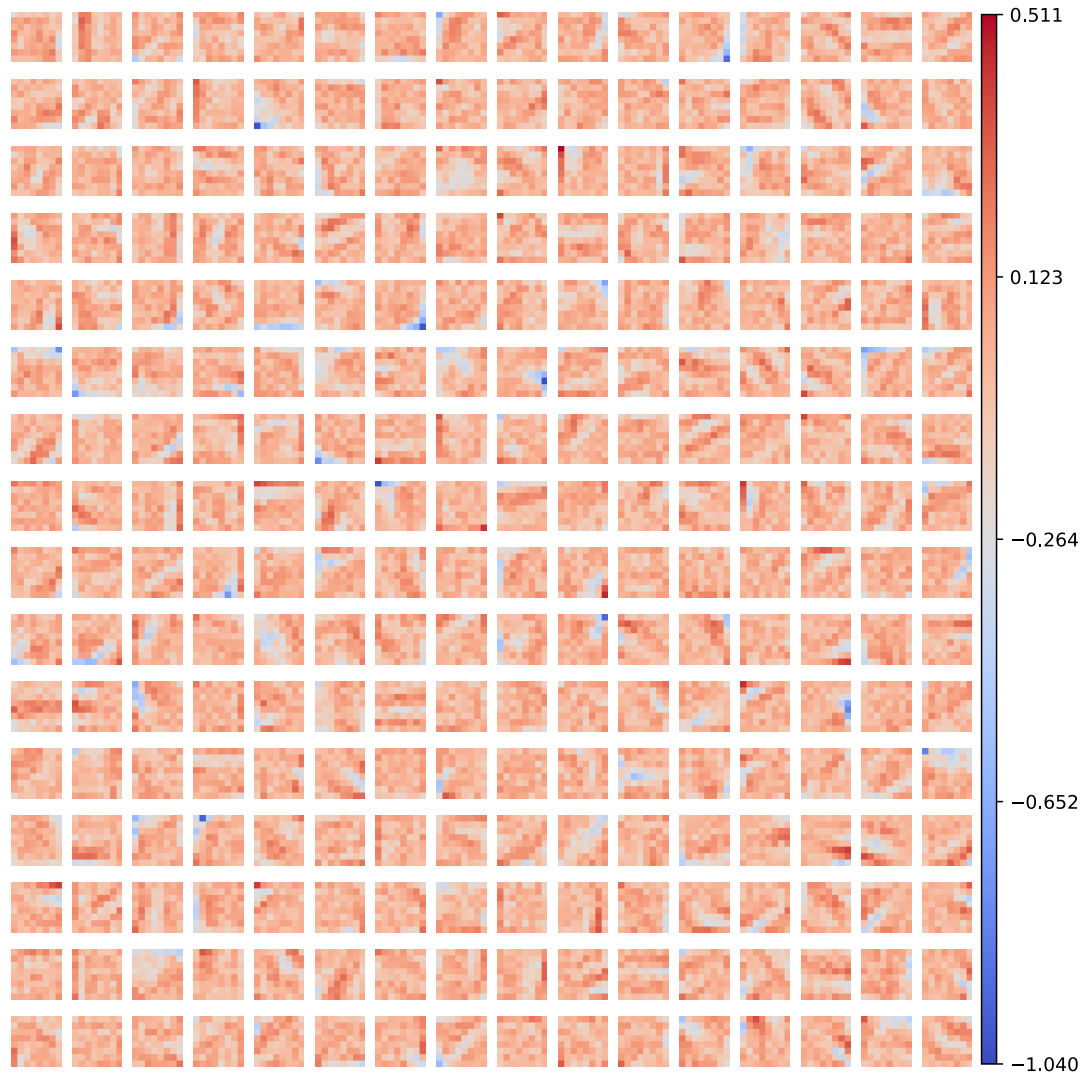


Figure 8: Input unit weights for the fully-connected layer in the glimpse network that takes as input an 8×8 glimpse of MNIST digits. In red are higher, positive weight values and blue are smaller and negative.