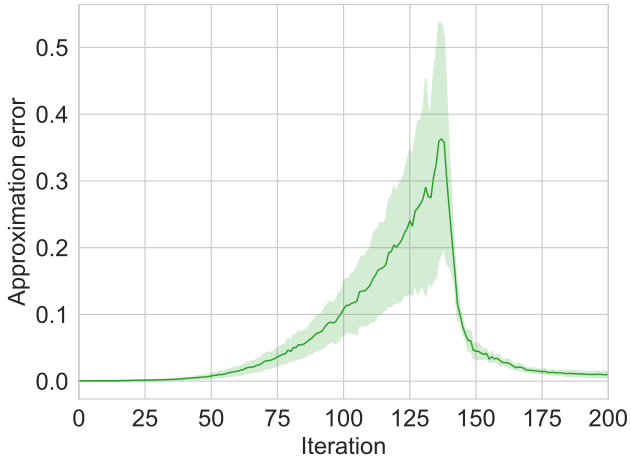

Figure 1: Approximation error across iterations between the attention output $\boldsymbol{y}_n$ and the skip connection $\boldsymbol{x}_n$ for $p = 0.2$ and $q = 0.3$, averaged across 5 runs. The figure shows that the attention layer is active during training but becomes negligible compared to the skip connection at convergence.

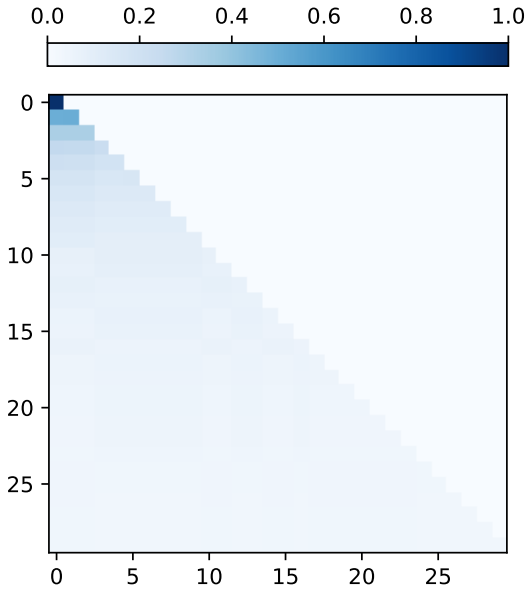

Figure 2: First 30 rows and columns of the attention matrix at convergence, averaged across 5 random test sequences. Note that attention attains every past symbol uniformly. This is due to the fact that in our setting the final contribution of the attention layer is negligible compared to the skip connection, as also shown in Figure 1 above.

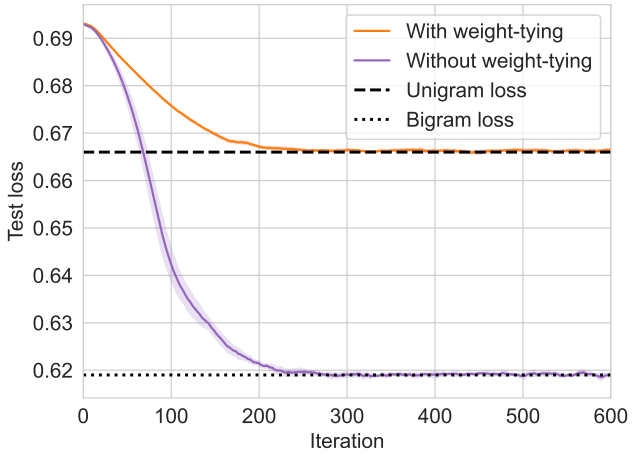

Figure 3: Test loss across iterations for the network considered in the paper with the attention layer removed, for $p = 0.5$ and $q = 0.8$. Similarly to the case examined in the paper, the network gets stuck in the local minimum when weight-tying is used. This does not happen whitout weight-tying.