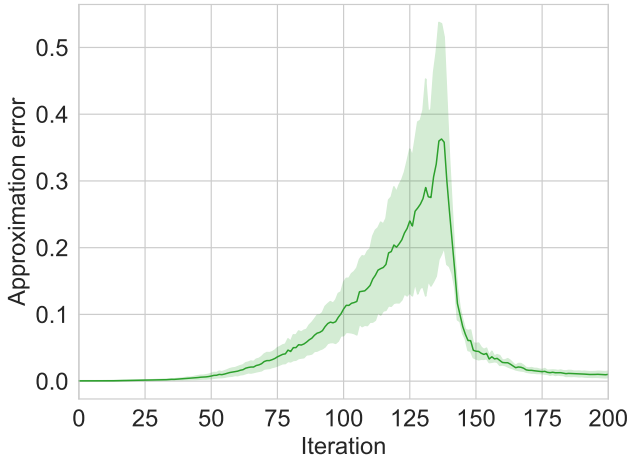

Figure 1: Approximation error across iterations between the attention output $\boldsymbol{y}_n$ and the skip connection $\boldsymbol{x}_n$ for $p = 0.2$ and $q = 0.3$, averaged across 5 runs. The figure shows that the attention layer is active during training but becomes negligible compared to the skip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

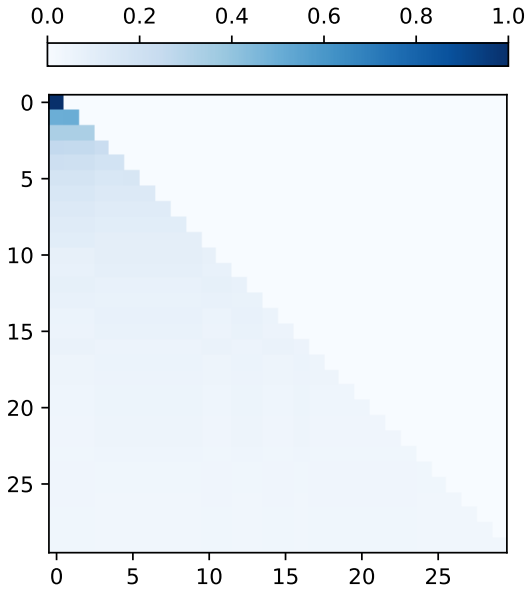

Figure 2: First 30 rows and columns of the attention matrix at convergence, averaged across 5 random test sequences. Note that attention attains every past symbol uniformly. This is due to the fact that in our setting the final contribution of the attention layer is negligible compared to the skip connection,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

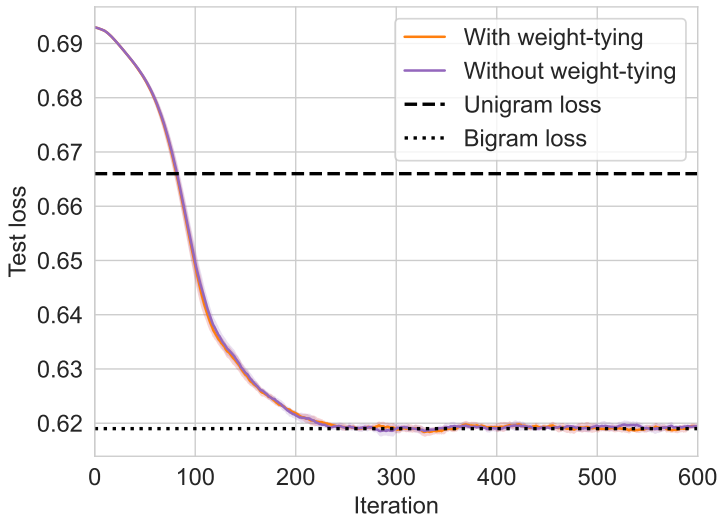

Figure 3: Test loss for a network as the one considered in the paper with the attention layer removed (i.e., the network is composed of word and positional embedding, MLP layer and linear layer with softmax). The network is able to correctly solve the prediction problem as expected. However, weight-tying does not negatively impact the convergence of the test loss: the network converges to the global minimum in both cases.