Learning and Memorization: Supplementary Material

Satrajit Chatterjee [1]

1. Extensions

Although our initial goal was to study if memorization alone can generalize it is tempting given our experimental results and the computational efficiency of the algorithm to see if memorization can be extended to a practical learning algorithm. We briefly discuss some possible extensions below.

Weights on training points. This is naturally accommodated by adding up the weights of the corresponding patterns instead of counting them. (Counting may be seen as a special case when all weights are 1.)

Multi-class classification. A simple way to handle m classes is to train m 1-v/s-rest binary classifiers where each classifier is trained with appropriate weights to compensate for the class imbalance. The output lut of each classifier is modified to emit the conditional probability that the output is 1 given the input pattern is p. This probability can be computed from the counts stored in the lut as $\frac{c_{p1}}{c_{p0}+c_{p1}}$. Finally the class with the highest probability is picked as the output. This approach can be used to get a 10-class classifier for binarized MNIST with a test accuracy of 0.86 which is significant compared to random guessing at 0.10. However, a more natural approach would be to have different luts in the network target different classes which also increases diversity inside the network.

Real-valued inputs and outputs. One way to handle real-valued outputs might be by quantizing the outputs and treating it as a multi-class classification problem which has proven useful even with neural networks (van den Oord et al., 2016). For inputs we could explore different coding schemes after quantizing such as one-hot, thermometer, and regular binary coding.

Structural Priors. Currently a lut in a layer could be connected to any lut from the previous layer. By restricting the possible connections geographically and by sharing counts across luts we can build convolutional structures. Similarly, skip connections (He et al., 2016; Srivastava et al., 2015) could increase the diversity of luts in a layer.

Integrated Architectural Exploration. Although in our current setting we fix the structure of the network before training this is not necessary. We can explore the topology while memorizing.

Unsupervised Learning. Instead of learning the target function some luts in a layer could be used to compress their input patterns to 1 bit.

Uncertainty Propagation. A lookup table could emit "unknown" in addition to 0 and 1 if at test time it encounters a pattern p such that c_{p0} and c_{p1} are both zero or close. This extra information could be utilized by downstream lookup tables in deciding their outputs perhaps by averaging over compatible patterns. This observation could also allow memorization to be used as a fast path to handle common cases with a fallback to more accurate (but also more computationally expensive) classifiers when memorization is uncertain. Finally, this scheme could be generalized to propagating quantized probabilities through the network.

2. Additional Experiment

We perform an experiment along the lines of Experiment 9 in the main paper where instead of a classifying points inside and outside the circle, we classify points as being below or above the line $x = y$. The results (shown in Figure 1) are very similar to those of Experiment 9 (reproduced in Figure 2) for convenience.

References

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pp. 770–778, 2016. doi: 10.1109/CVPR.2016.90. URL https://doi.org/10.1109/CVPR.2016.90.

Srivastava, R. K., Greff, K., and Schmidhuber, J. Highway networks. *CoRR*, abs/1505.00387, 2015. URL http://arxiv.org/abs/1505.00387.

van den Oord, A., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A. W., and Kavukcuoglu, K. Wavenet: A generative model for raw audio. *CoRR*, abs/1609.03499, 2016. URL http://arxiv.org/abs/1609.03499.

[1]Two Sigma, New York, NY, USA. Correspondence to: Satrajit Chatterjee <satrajit.chatterjee@twosigma.com>.

Proceedings of the 35^{th} International Conference on Machine Learning, Stockholm, Sweden, PMLR 80, 2018.

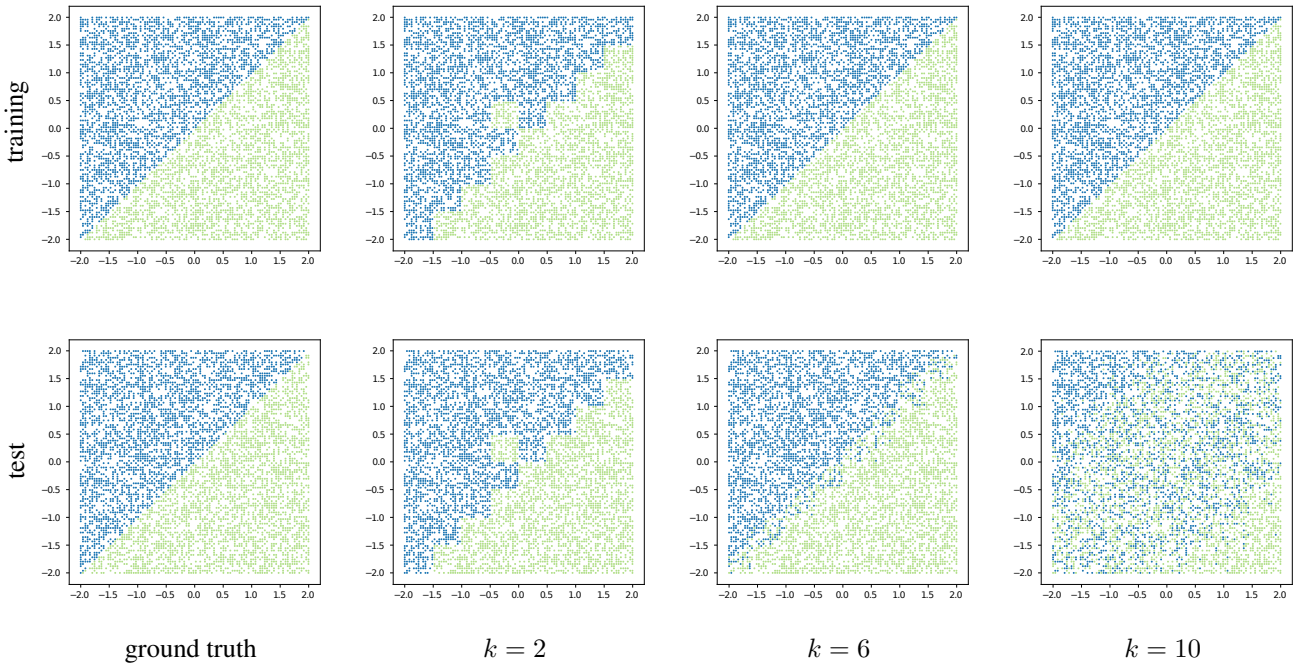

Figure 1. The decision boundaries learned on a synthetic dataset.

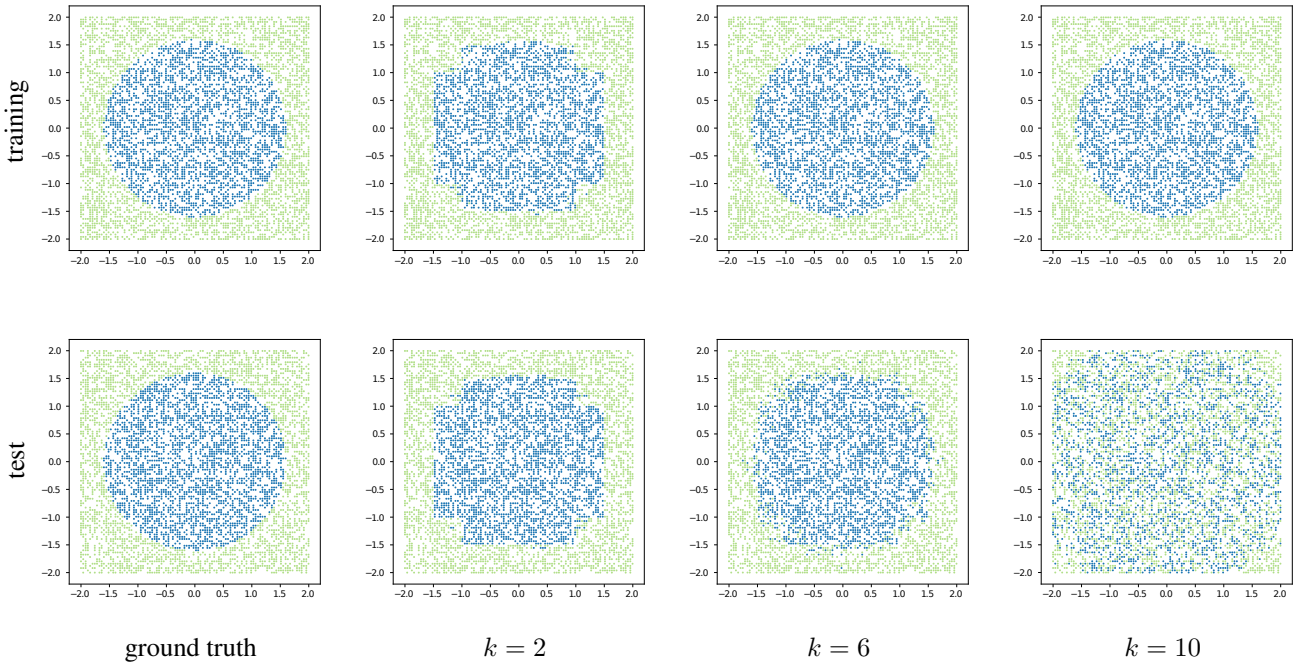

Figure 2. The decision boundaries learned on a synthetic dataset.