Small ReLU networks are powerful memorizers: a tight analysis of memorization capacity

Chulhee Yun (chulheey@mit.edu), Suvrit Sra, and Ali Jadbabaie

Laboratory for Information and Decision Systems, Massachusetts Institute of Technology

TL;DR

We prove for 3-layer ReLU fully-connected neural nets that \( \Theta(\sqrt{N}) \) hidden nodes are necessary and sufficient for memorizing arbitrary \( N \) data points. For deeper networks, we prove that \( \Theta(N) \) parameters are sufficient. These results give (almost) tight bounds on memorization capacity.

Introduction

• Overparametrized NNs with SGD memorize even random noise.
• Q: Given a network, can it memorize arbitrary datasets? How large should it be to memorize any \( N \) data points?
• Recent results on fully-connected, residual, convolutional networks require \( N \) hidden nodes to memorize \( N \) data points!

Def. We define (universal) finite sample expressivity of a neural network \( \theta_0(\cdot) \) as its ability to memorize arbitrary dataset \( \{(x_i, y_i)\}_{i=1}^N \subset \mathbb{R}^{d_X} \times \mathbb{R} \) with \( N \) points.

Def. We define memorization capacity as the maximum of \( N \) for which the network has finite sample expressivity, when \( d_y = 1 \).

cf. VC dim: there exists \( \{x_i\}_{i=1}^N \) such that \( \theta_0(\cdot) \) can shatter \( \{y_i\}_{i=1}^N \subset \{-1, 1\}^N \Longrightarrow \) Memorization capacity \( \leq \) VC dim.

Main Results

Problem Settings

• ReLU(-like) fully-connected neural network:
  \[ a^l(x) = x, \quad a^l(x) = \sigma(W^l a^{l-1}(x) + b^l), \quad l = 1, \ldots, L, \]
  \[ f_\theta(x) = W^{L-1} a^1(x) + b^{L-1}, \]
  \[ \sigma(\cdot) = \max\{s, t, s, t\}, \quad s_0 > s_1 \geq 0. \]
• \( d_i \) is the width of the \( l \)-th hidden layer. \( d_0 = d_y, d_{L+1} = d_y \).
• \( W^l \in \mathbb{R}^{d_l \times d_{l-1}}, b^l \in \mathbb{R}^{d_l}, \theta = (W^l, b^l)_{l=1}^{L+1} \)

Theorem 3.1. A 2-hidden-layer ReLU network with hidden layer dimensions \( d_1d_2 \geq 4Nd_y \) can memorize any arbitrary dataset with \( N \) distinct points.

Proposition 3.2. A 3-hidden-layer ReLU network with hidden layer dimensions \( d_1d_2 \geq 4N \) and \( d_3 \geq 4d_y \) can memorize any arbitrary classification dataset with \( N \) distinct points.

Theorem 3.3. For a 1-hidden-layer ReLU network with \( d_1 + 2 < N \) or a 2-hidden-layer ReLU network with \( 2d_1 + d_2 + 2 < N \), \exists a dataset \( \{(x_i, y_i)\}_{i=1}^N \) with \( N \) points that the network cannot memorize.

• Depth-width trade-offs in finite sample memorization.
• Tight bounds \( \Theta(d_1) \) and \( \Theta(d_1d_2) \) on memorization capacity for 1- and 2-hidden-layer ReLU nets, resp.

Extension to Deeper Networks

Proposition 3.4 (informal). A \( L \)-hidden-layer ReLU network with \( W \) parameters between hidden layers can memorize arbitrary \( N \) dataset if \( W = \Omega(Nd_y) \).

• Gives a lower bound on mem. capacity: \( \Omega(W) \). Almost tight!
• cf. memorization capacity \( \leq \) VC dim = \( O(WL\log W) \).

Results on ResNets and SGD

Theorem 4.1 (informal). A deep residual network with \( \frac{d_y}{2} + 6d_y \) can memorize any classification dataset with \( N \) points if \( x_i \)'s are in general position.

• Under a different assumption, reduce \( N + d_y \) nodes to \( \frac{d_y}{2} + 6d_y \).
• CIFAR-10 \((N = 50k, d_y = 3072, d_y = 10)\): 50,016 vs 126 nodes
  • Given a dataset \( \{(x_i, y_i)\}_{i=1}^N \) with empirical risk \( \mathcal{R}(\theta) = \Sigma_i \ell(f_\theta(x_i); y_i) \).
  • \( \ell(z; y) \) is strictly convex and three times differentiable in \( z \).
  • \( \forall y, \exists \) a global minimizer \( z \) of \( \ell(z; y) \).
• Def. A point \( \theta^* \) is a memorizing global minimum of \( \mathcal{R} \) if \( \ell(f_\theta(x_i); y_i) = 0 \) for all \( i \).
• We consider without-replacement SGD, i.e., random shuffling.

Theorem 5.1 (informal). If \( \theta^{(0)} \) satisfies \( \|\theta^{(0)} - \theta^*\| \leq \rho \) for some memorizing global minimum and a small constant \( \rho \), SGD starting from \( \theta^{(0)} \) quickly finds a point \( \theta \) that satisfies

\[ \mathcal{R}(\theta) - \mathcal{R}(\theta^*) = O(||\theta^{(0)} - \theta^*||), \quad \|\theta - \theta^*\| \leq 2||\theta^{(0)} - \theta^*||. \]