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

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

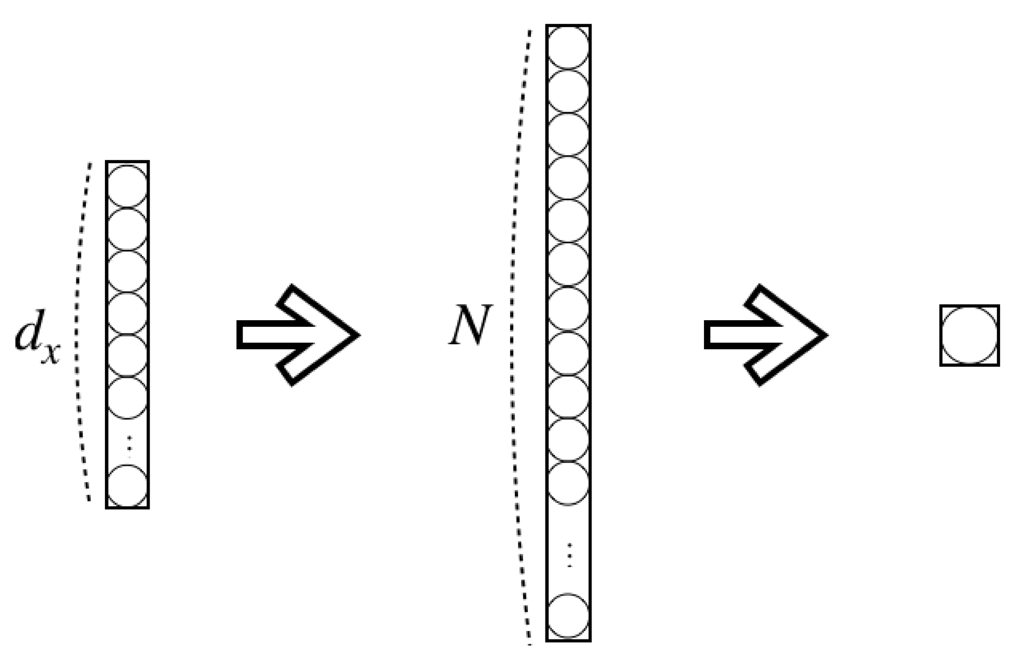
Laboratory for Information and Decision Systems, Massachusetts Institute of Technology

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

- Overparametried 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 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 $f_\theta(\cdot)$ as its ability to memorize arbitrary dataset $\{(x_i, y_i)\}_{i=1}^N \in \mathbb{R}^{(d_x+d_y)\times N}$ 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 $f_\theta(\cdot)$ can shatter $\{y_i\}_{i=1}^N \in \{\pm 1\}^N \implies$ Memorization capacity $\leq$ VC dim.

Problem Settings

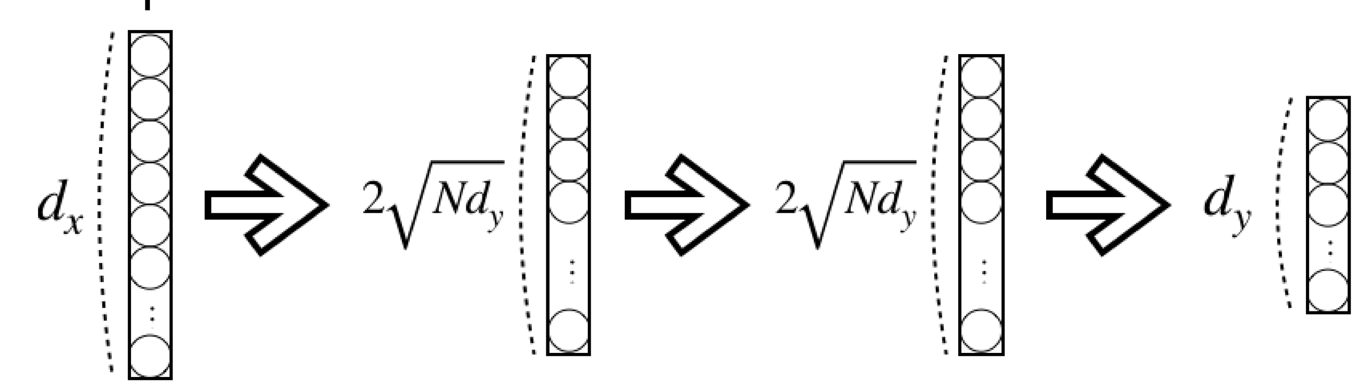
- ReLU(-like) fully-connected neural network:

$$a^0(x) = x, \quad a^l(x) = \sigma(\boldsymbol{W}^l a^{l-1}(x) + \boldsymbol{b}^l), \; l = 1, \dots, L,$$
$$f_\theta(x) = \boldsymbol{W}^{L+1} a^L(x) + \boldsymbol{b}^{L+1},$$
$$\sigma(t) = \max\{s_+ t, s_- t\}, \; s_+ > s_- \geq 0.$$

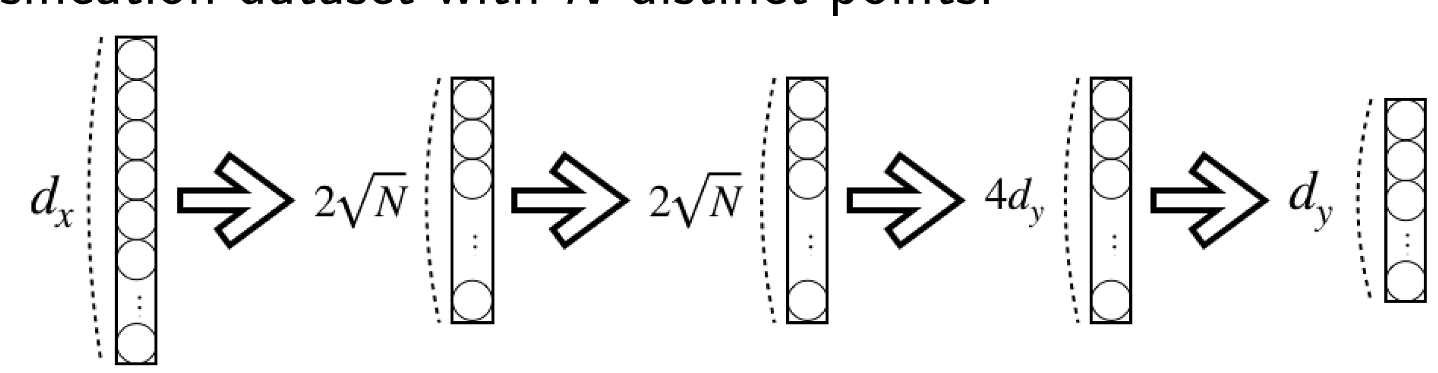
- d_l is the width of the l-th hidden layer. $d_0 = d_x, d_{L+1} = d_y$.
- $\boldsymbol{W}^l \in \mathbb{R}^{d_l \times d_{l-1}}, \boldsymbol{b}^l \in \mathbb{R}^{d_l}, \boldsymbol{\theta} = (\boldsymbol{W}^l, \boldsymbol{b}^l)_{l=1}^{L+1}$

Main Results

Theorem 3.1. A 2-hidden-layer ReLU network with hidden layer dimensions $d_1 d_2 \geq 4Nd_y$ can memorize any arbitrary dataset with N distinct points.



Proposition 3.2. A 3-hidden-layer ReLU network with hidden layer dimensions $d_1 d_2 \geq 4N$ and $d_3 \geq 4d_y$ can memorize any arbitrary classification dataset with N distinct points.

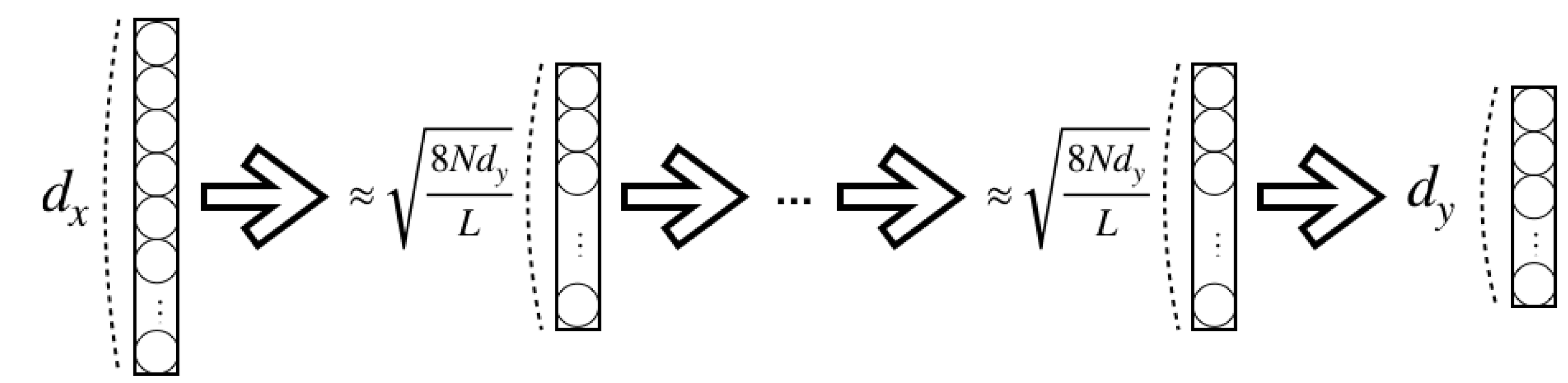


Theorem 3.3. For a 1-hidden-layer ReLU network with $d_1 + 2 < N$ or a 2-hidden-layer ReLU network with $2d_1 d_2 + d_2 + 2 < N$, $\exists$ a dataset ($d_y = 1$) with N points that the network cannot memorize.

- **Depth-width trade-offs** in finite sample memorization.
- **Tight** bounds $\Theta(d_1)$ and $\Theta(d_1 d_2)$ 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 memo. 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 $\frac{4N}{d_x} + 6d_y$ 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 $\frac{4N}{d_x} + 6d_y$.
- CIFAR-10 ($N = 50$k, $d_x = 3072$, $d_y = 10$): 50,010 vs 126 nodes

- Given a dataset $\{(x_i, y_i)\}_{i=1}^N$, empirical risk $\mathfrak{R}(\boldsymbol{\theta}) = \sum_i \ell(f_\theta(x_i); y_i)$.
- $\ell(z; y)$ is strictly convex and three times differentiable in z.
- $\forall y$, $\exists$ a global minimizer z of $\ell(z; y)$.
- **Def.** A point $\boldsymbol{\theta}^*$ is a memorizing global minimum of $\mathfrak{R}$ 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 $\boldsymbol{\theta}^{(0)}$ satisfies $\|\boldsymbol{\theta}^{(0)} - \boldsymbol{\theta}^*\| \leq \rho$ for some memorizing global minimum and a small constant ρ, SGD starting from $\boldsymbol{\theta}^{(0)}$ quickly finds a point $\boldsymbol{\theta}$ that satisfies

$$\mathfrak{R}(\boldsymbol{\theta}) - \mathfrak{R}(\boldsymbol{\theta}^*) = O(\|\boldsymbol{\theta}^{(0)} - \boldsymbol{\theta}^*\|^4), \; \|\boldsymbol{\theta} - \boldsymbol{\theta}^*\| \leq 2\|\boldsymbol{\theta}^{(0)} - \boldsymbol{\theta}^*\|.$$