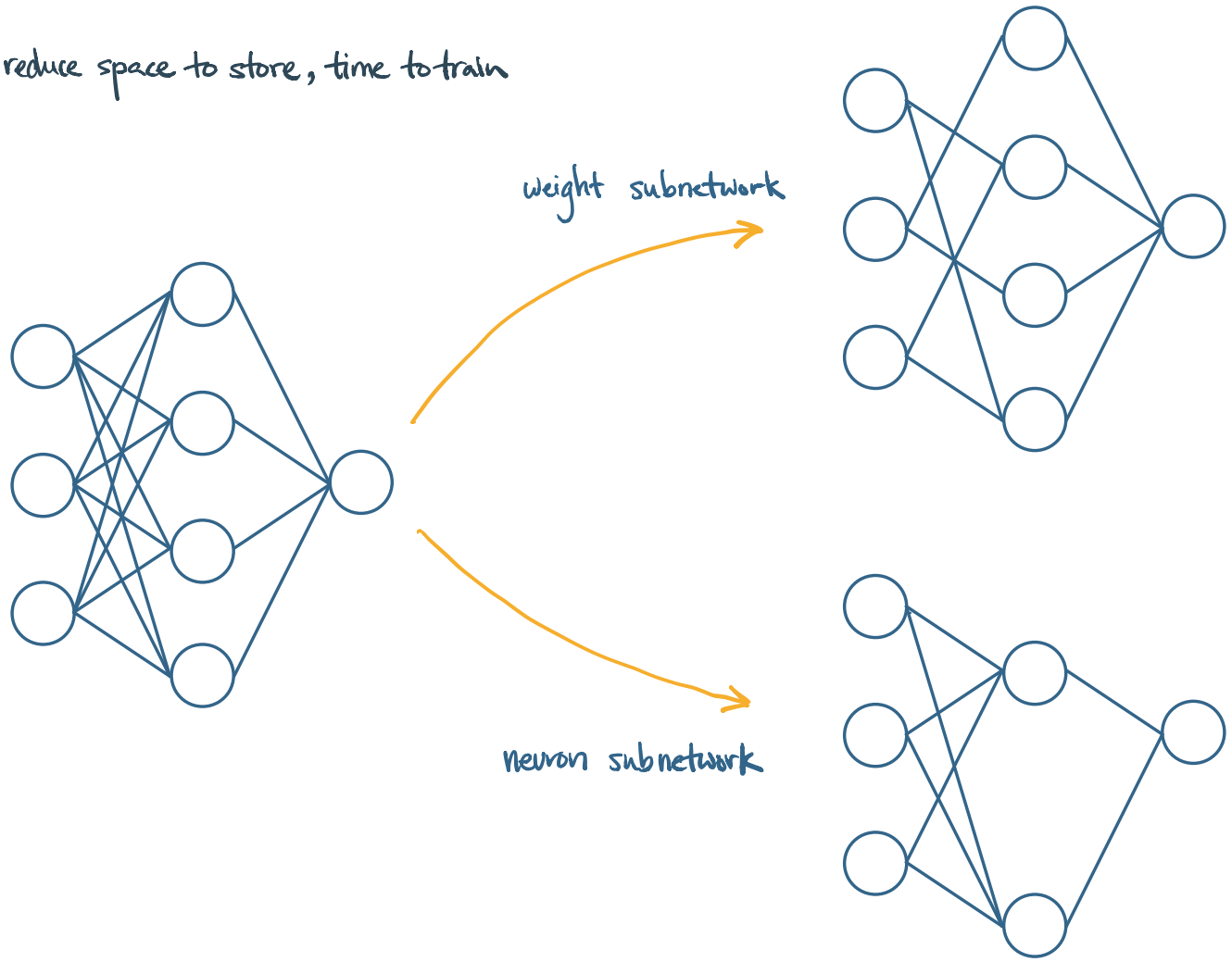
Proving the Lottery Ticket Hypothesis

ICML 2020, Eran Malach, Gilad Yehudai, Shai Shalev-Shwartz, Ohad Shamir

KMLR Reading Group, Geelon So, Aug. 12, 2020

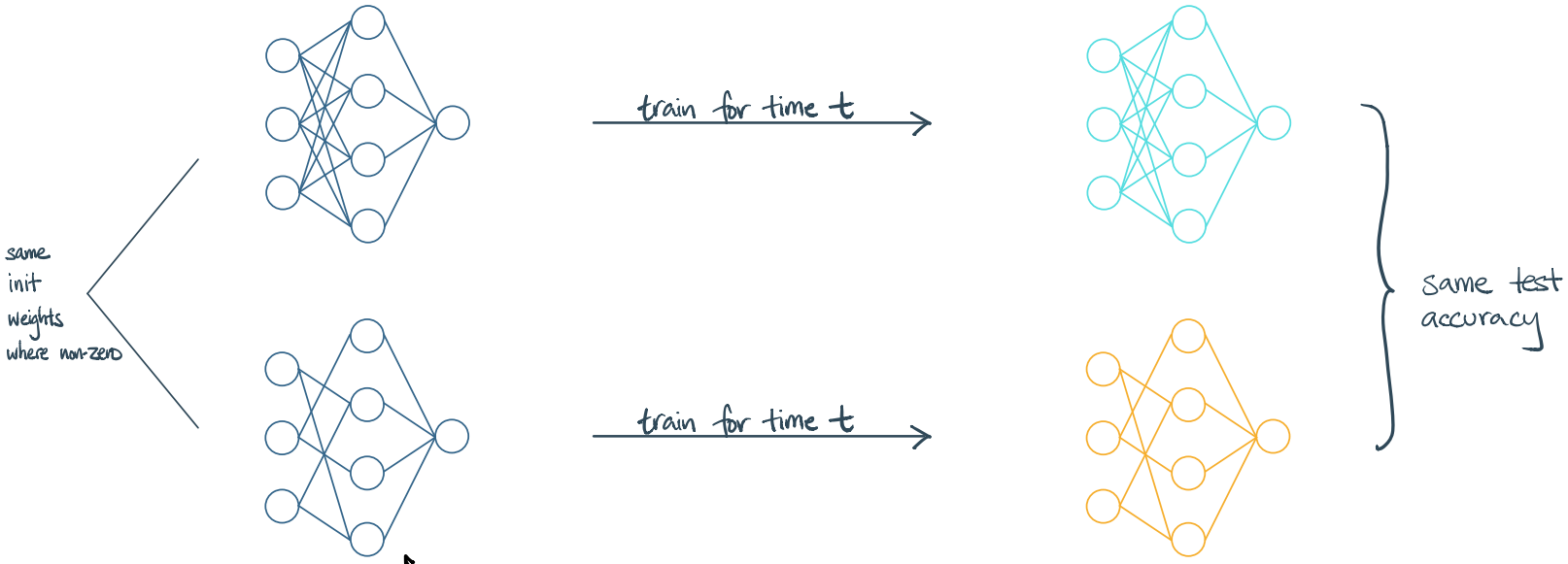
PRUNING NEURAL NETS

- goal: reduce space to store, time to train



THE LOTTERY TICKET HYPOTHESIS

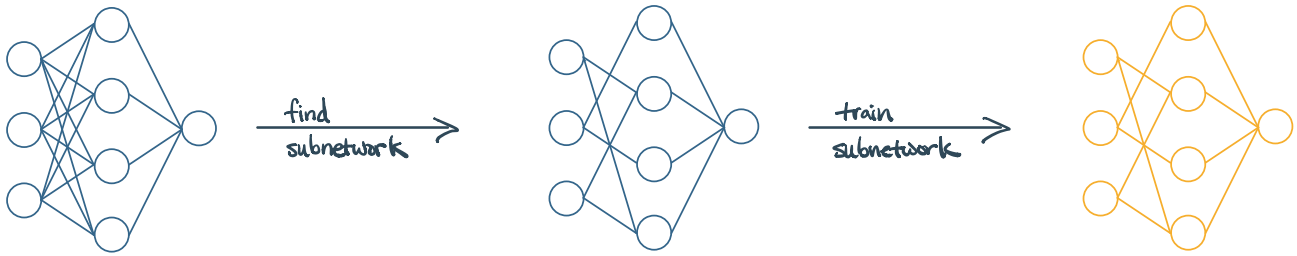
CONJECTURE (Frankle & Carbin 2018). A randomly-initialized, dense neural net contains a subnetwork that is initialized such that when trained in isolation, it can match the test accuracy of the original network after training for at most the same number of iterations.



THE LOTTERY TICKET HYPOTHESIS

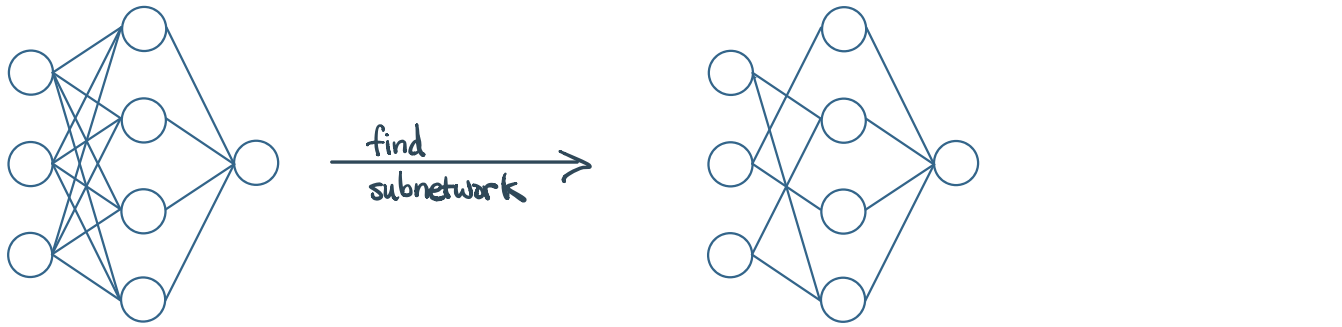
IMPLICATION. Instead of training a dense neural net, find suitable subnetwork to train.

i.e. it might be possible to speed up training large networks.



STRONG LOTTERY TICKET HYPOTHESIS

CONJECTURE (Ramanujan, et. al.). Within a sufficiently overparametrized neural net with random weights, there exists a subnetwork that achieves competitive accuracy.



↑ performs well (no training).

- Proved in this paper.

RESULTS of this paper

THEOREM (weight-subnetwork, informal).

GIVENS:

- $f : \mathbb{R}^d \to \mathbb{R}$ an arbitrary ReLU neural net, depth ℓ, width n
- $g : \mathbb{R}^d \to \mathbb{R}$ a random ReLU neural net, depth 2ℓ, width $\text{poly}(d, n, \ell, \frac{1}{\varepsilon}, \log\frac{1}{\delta})$

If f satisfies mild conditions, then $\exists \tilde{g}$ a weight-subnetwork of g s.t. w.p. $1-\delta$,

$$\sup_{x \in B(0,1)} |f(x) - \tilde{g}(x)| < \varepsilon.$$

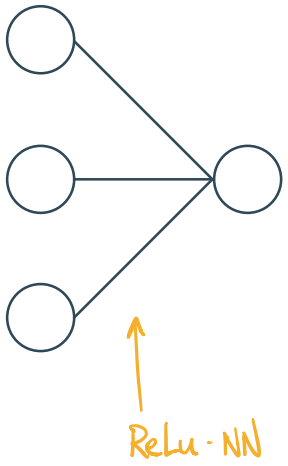
Furthermore, the number of active weights in $\tilde{g}$ is $O(dn + n^2\ell)$.

↖ same order of weights in f.

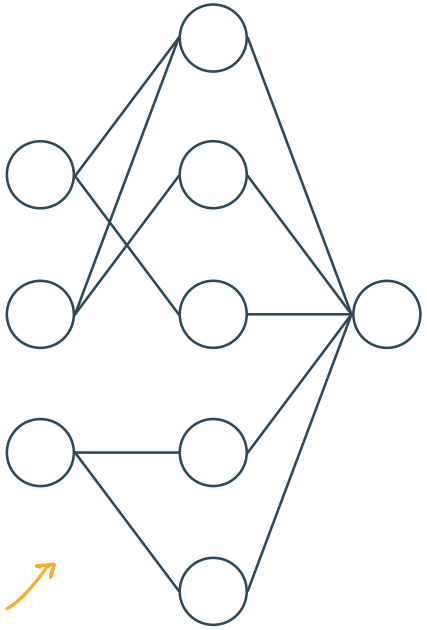
PROOF IDEA

- Show that a single ReLu neuron $f:\mathbb{R}^d \to \mathbb{R}$ is closely approximated by a weight-subnetwork of a random 2-layer net with high probability

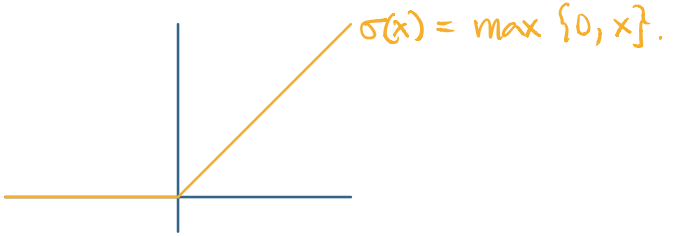
$$f : x \mapsto \sigma\left(\sum_{i=1}^{d} w_i x_i\right)$$



$\approx$



PROOF IDEA



Let $\sigma: \mathbb{R} \to \mathbb{R}$ be the ReLu activation

- $x = \sigma(x) - \sigma(-x)$

$$\Rightarrow \quad \sigma\left(\sum_{i=1}^{d} w_i x_i\right) = \sigma\left(\sum_{i=1}^{d} \sigma(w_i x_i) - \sum_{i=1}^{d} \sigma(-w_i x_i)\right)$$

Therefore, a single ReLu neuron can be written as a 2-layer ReLu NN.

PROOF IDEA

Let g be a random 2-layer neural net,

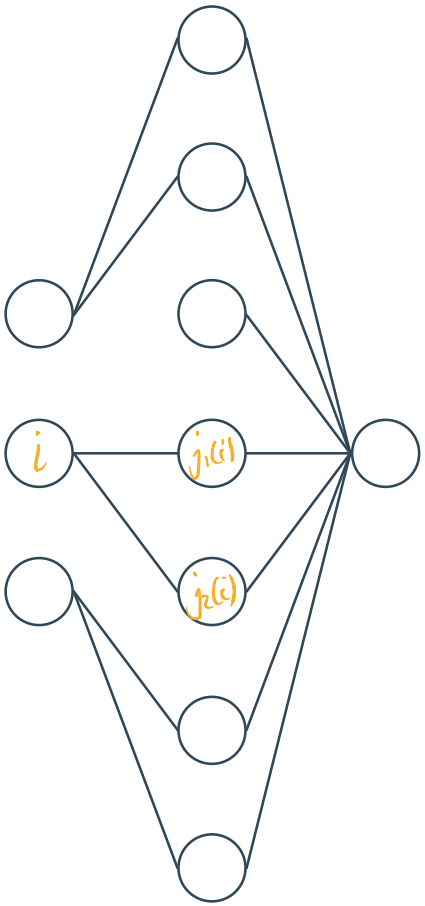
$$g(x) = \sigma\left(\sum_{j=1}^{k} A_j \cdot \sigma\left(\sum_{i=1}^{d} W_{j,i}\, x_i\right)\right)$$

where $A_j, W_{j,i}$ are random variables, $A \in \mathbb{R}^k$, $W \in \mathbb{R}^{k \times d}$

- Let $B \in \{0,1\}^{k \times d}$ be a bit-mask
  - $B \odot W$ is a pruning of W
    - ↑ component-wise multiplication (Hadamard product)
    - $(B \odot W)_{ji} = \begin{cases} W_{ji} & \text{if } B_{ji} = 1 \\ 0 & \text{if } B_{ji} = 0 \end{cases}$

PROOF IDEA

Let B be a bit-mask so that $\forall i \in [d]$,



$$j = j_1(i) \text{ or } j = j_2(i) \iff B_{ji} = 1$$

$$\text{and } B_{ji} = 1 \implies B_{j'i} = 0 \text{ if } j' \neq j.$$

i.e. each neuron in the first layer is connected to two unique neurons in the hidden layer.

PROOF IDEA

If B as above,

$$\tilde{g}(x) = \sigma\left(\sum_{i=1}^{d} A_{j_1(i)} \cdot \sigma(W_{j_1(i),i}\, x_i) + \sum_{i=1}^{d} A_{j_2(i)} \cdot \sigma(W_{j_2(i),i}\, x_i) \right)$$

$$= \sigma\left(\sum_{i=1}^{d} \operatorname{sign}(A_{j_1(i)}) \cdot \sigma\left(|A_{j_1(i)}| \cdot W_{j_1(i),i}\, x_i \right) + \sum_{i=1}^{d} \operatorname{sign}(A_{j_2(i)}) \cdot \sigma\left(|A_{j_2(i)}| \cdot W_{j_2(i),i}\, x_i \right) \right)$$

Recall: $$f(x) = \sigma\left(\sum_{i=1}^{d} \sigma(w_i x_i) - \sum_{i=1}^{d} \sigma(-w_i x_i) \right)$$

- want $\operatorname{sign}(A_{j_1(i)}) \neq \operatorname{sign}(A_{j_2(i)})$

  and $\left| |A_{j_1(i)}| \cdot W_{j_1(i),i} - \operatorname{sign}(A_{j_1(i)}) \cdot w_i \right| < \frac{\varepsilon}{2d}$ (similarly for $j_2(i)$).

$\Rightarrow \sup_{x \in B(0,1)} |f(x) - \tilde{g}(x)| < \varepsilon.$

PROOF IDEA

The mild condition on $f(x) = \sigma\left(\sum_{i=1}^{d} w_i x_i\right)$, $\quad w_i \in [-\frac{1}{\sqrt{d}}, \frac{1}{\sqrt{d}}]$

w_i's are bounded

Random initialization of $A_{j(i)}$, $W_{j(i),i} \sim \text{Unif}([-1,1])$

$\Rightarrow \quad \text{sign}(A_{j_1(i)}) \neq \text{sign}(A_{j_2(i)})$

and $\quad \left| |A_{j_1(i)}| \cdot W_{j_1(i),i} - \text{sign}(A_{j_1(i)}) \cdot w_i \right| < \frac{\varepsilon}{2d}$ (similarly for $j_2(i)$).

with probability $\Omega(\varepsilon/d)$.

$\rightsquigarrow$ setting $k \gg \frac{d}{\varepsilon}$ $\Rightarrow$ for each i, we can find $j_1(i)$ and $j_2(i)$ satisfying above conditions w.h.p.

FORMAL STATEMENT

Theorem 1. *Fix $\varepsilon, \delta \in (0,1)$. Let $f : \mathbb{R}^d \to \mathbb{R}$ be an arbitrary ReLu neural network of depth ℓ and width n,*

$$f(x) = \sigma(w^{(\ell)} \cdot \sigma(w^{(\ell-1)} \cdots \sigma(w^{(1)} \cdot x))),$$

where $w^{(1)} \in \mathbb{R}^{n \times d}$, $w^{(\ell)} \in \mathbb{R}^{1 \times n}$ and $w^{(i)} \in \mathbb{R}^{n \times n}$ for $1 < i < \ell$. Assume bounded weights, $\left\|w^{(i)}\right\|_2 \leq 1$, $\left\|w^{(i)}\right\| \leq \frac{1}{\sqrt{n_i}}$ where $n_1 = d$ and $n_i = n$ otherwise.

Let $g : \mathbb{R}^d \to \mathbb{R}$ be a random ReLu neural network with depth 2ℓ and:

$$\text{width} = \text{poly}\left(d, n, \ell, \frac{1}{\varepsilon}, \log \frac{1}{\delta}\right),$$

with weights $W^{(j)} \sim \text{Unif}([-1,1])$ for $1 \leq j \leq 2\ell$.

With probability at least $1 - \delta$, there is a weight-subnetwork $\widetilde{g}$ of g such that:

$$\sup_{x \in B(0,1)} |\widetilde{g}(x) - f(x)| \leq \varepsilon.$$

Furthermore, the number of active (non-zero) weights in $\widetilde{g}$ is $O(dn + n^2\ell)$.

conditions on f

conditions on g

there is a good pruning

Proof. Apply construction for a single neuron n times for each layer, then compose layers. □

ADDITIONAL NOTE

- the paper also shows a result for neuron-subnetworks, with a relatively simple argument reducing to random features model.

THEOREM (neuron-subnetwork, informal).

Let D be a distribution over $\mathcal{X} \times [-1,1]$, σ be L-Lipschitz.

Let D^* be a distribution over weights $\{w : \|w\| \le 1\}$ s.t. if $w_1, \dots, w_n \sim D^*$ w.h.p. $\exists u_1, \dots, u_n \in [-C, C]$ s.t.

$$f(x) = \sum_{i=1}^{n} u_i \sigma(\langle w_i, x \rangle)$$

achieves small loss $L_D(f) \le \varepsilon$.

(random features model)

Let g be a random 2-layer neural network, width $= \text{poly}\left(C, n, L, \frac{1}{\varepsilon}, \frac{1}{\delta}\right)$ where first layer random weights $\sim D^*$, second layer $\sim \text{Unif}([-1,1])$.

$\Rightarrow \exists \tilde{g}$ neuron-subnetwork and constant $c > 0$ s.t. $L_D(c\tilde{g}) \le \varepsilon$.

This shows the equivalence between random features models and neuron-subnetworks.

- random features models cannot efficiently approximate a single Relu neuron.