WHAT'S HIDDEN IN A RANDOMLY WEIGHTED NEURAL NETWORK?

Ramanujan & Wortsman et al.

Presented by Andre

I2 Journal Club, 2/28/22

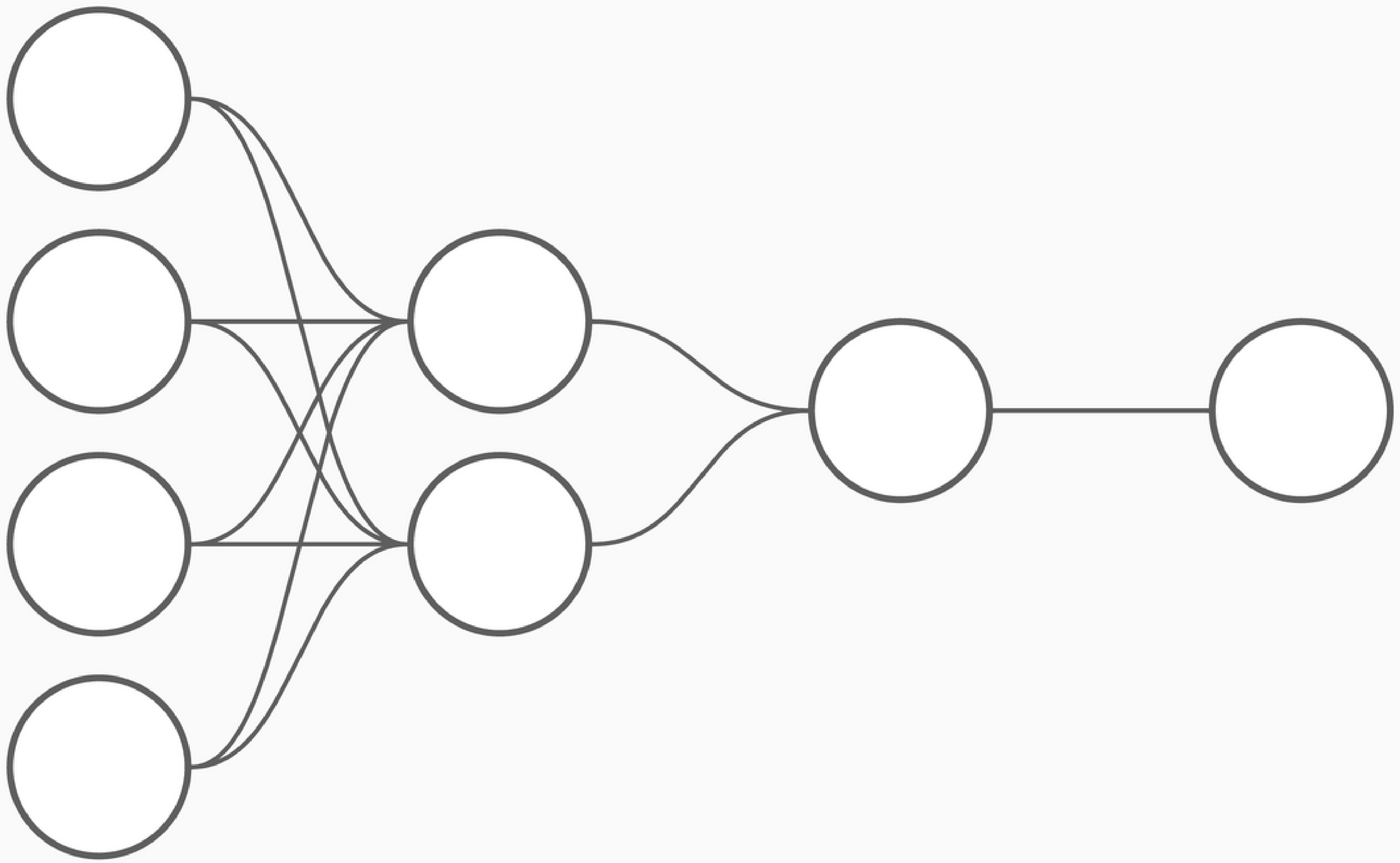
Primary Finding

"If a neural network with random weights is sufficiently overparameterized, it will contain a subnetwork that performs as well as a trained neural network with the same number of parameters."

Primary Finding - Rephrased

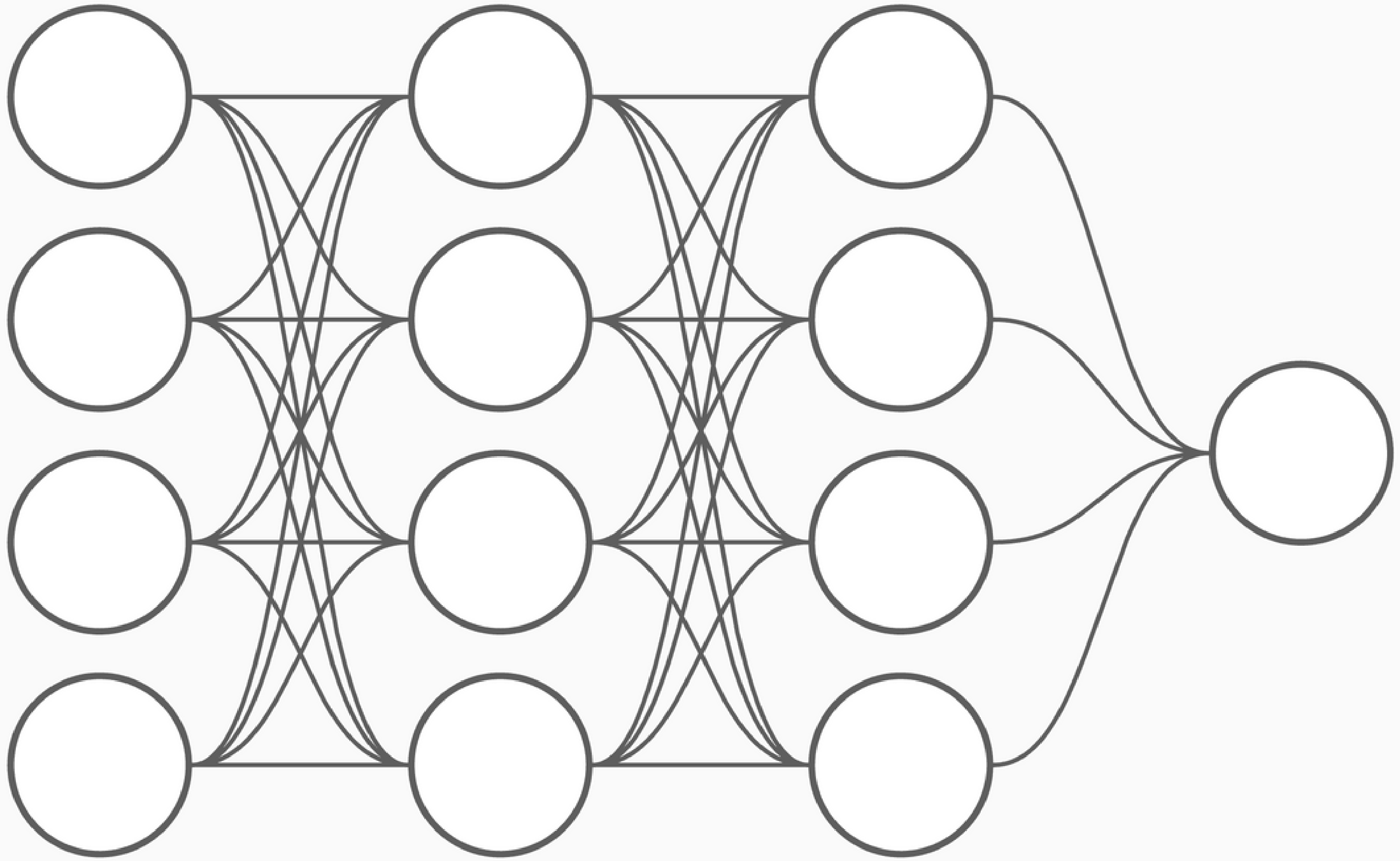
We can find a subnetwork in a randomly initialized network that does pretty well, with <u>no training</u> (<u>parameter updating</u>).

SMALL NEURAL NETWORK



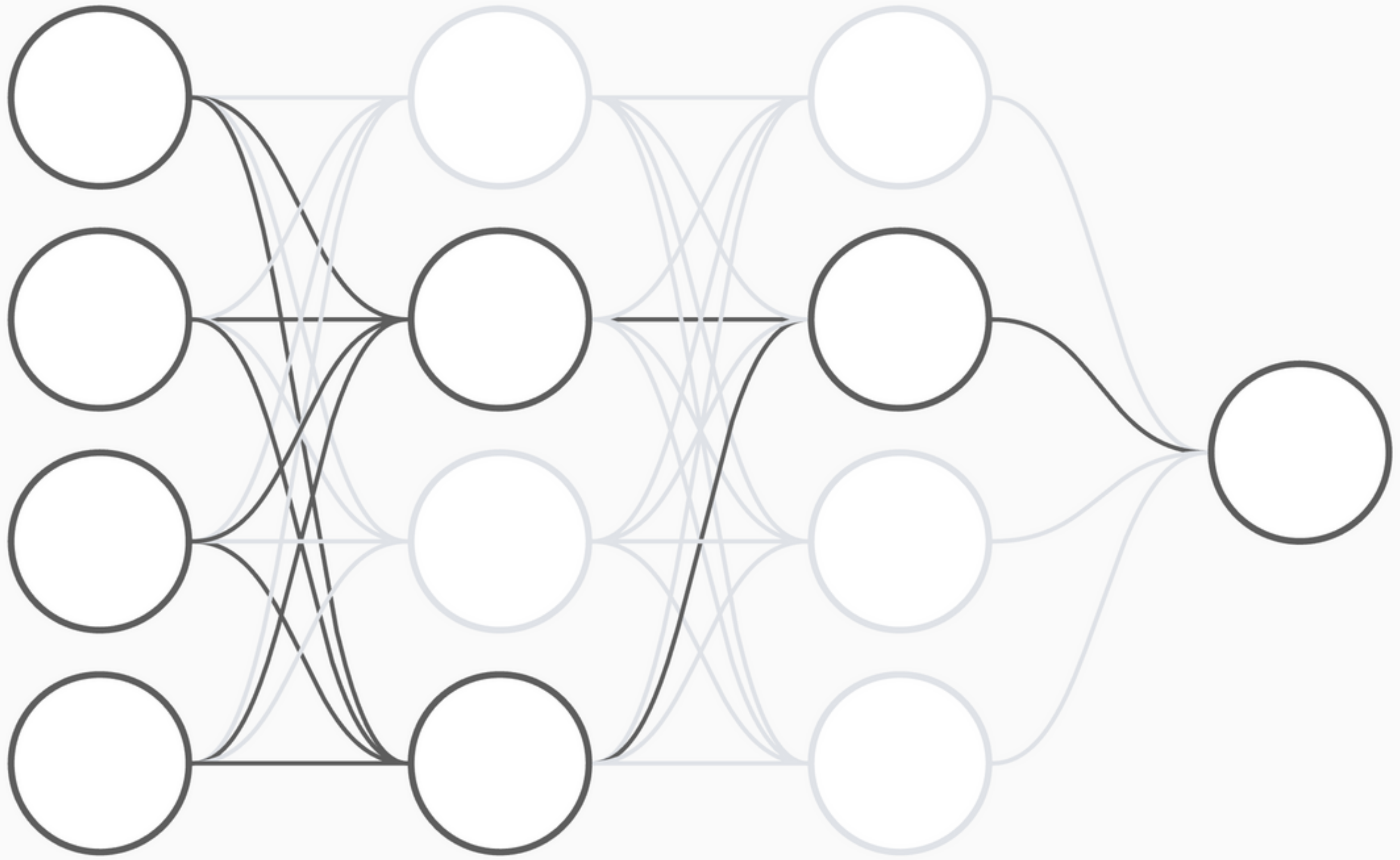
Trained 'small network' *p*

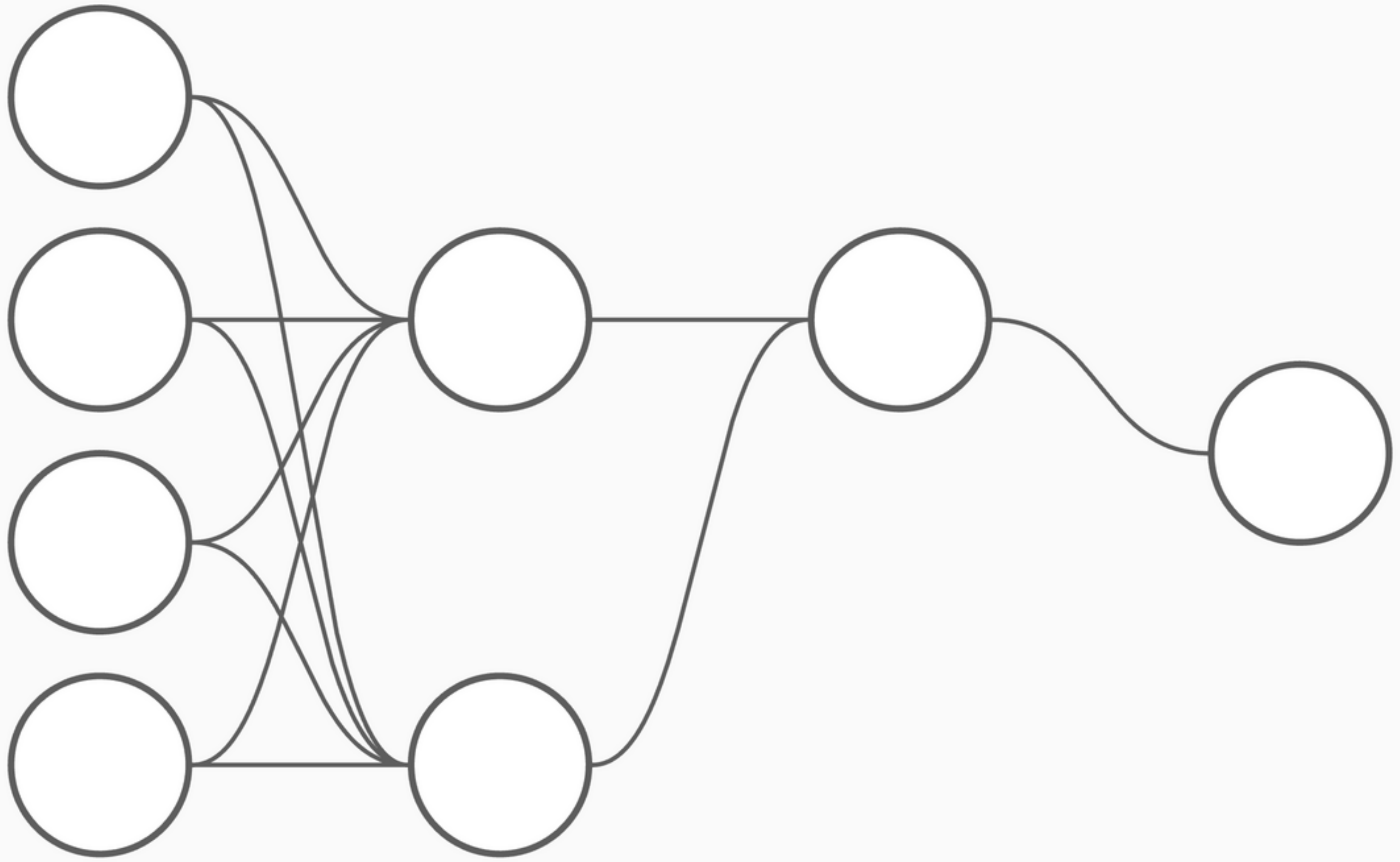
STANDARD NEURAL NETWORK



Untrained *large network q*

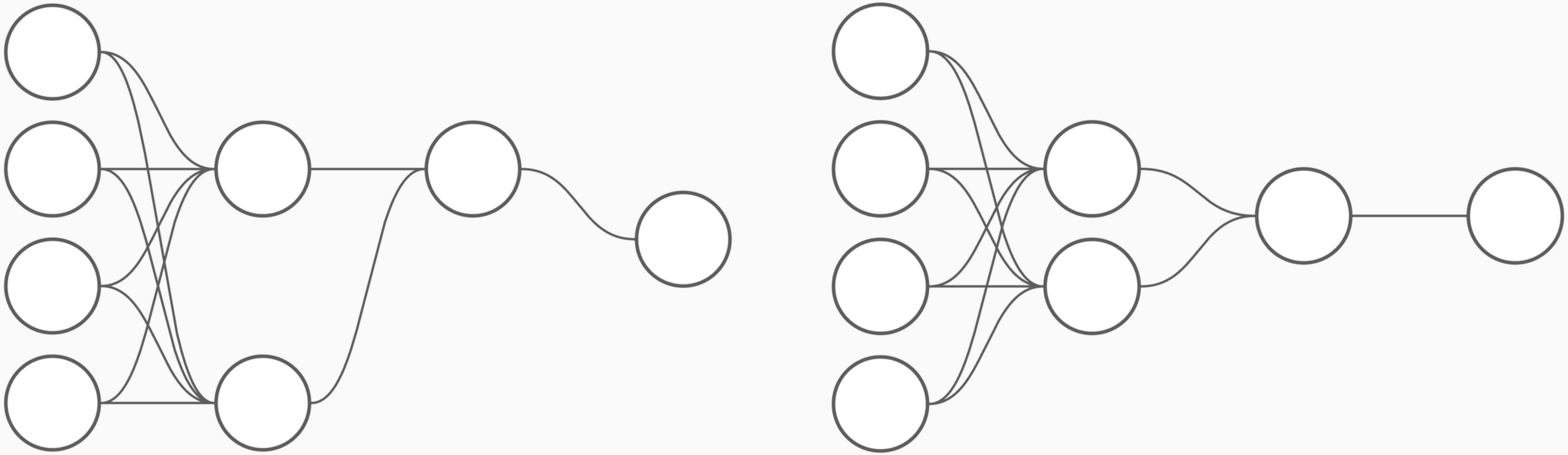
SUBNETWORKS





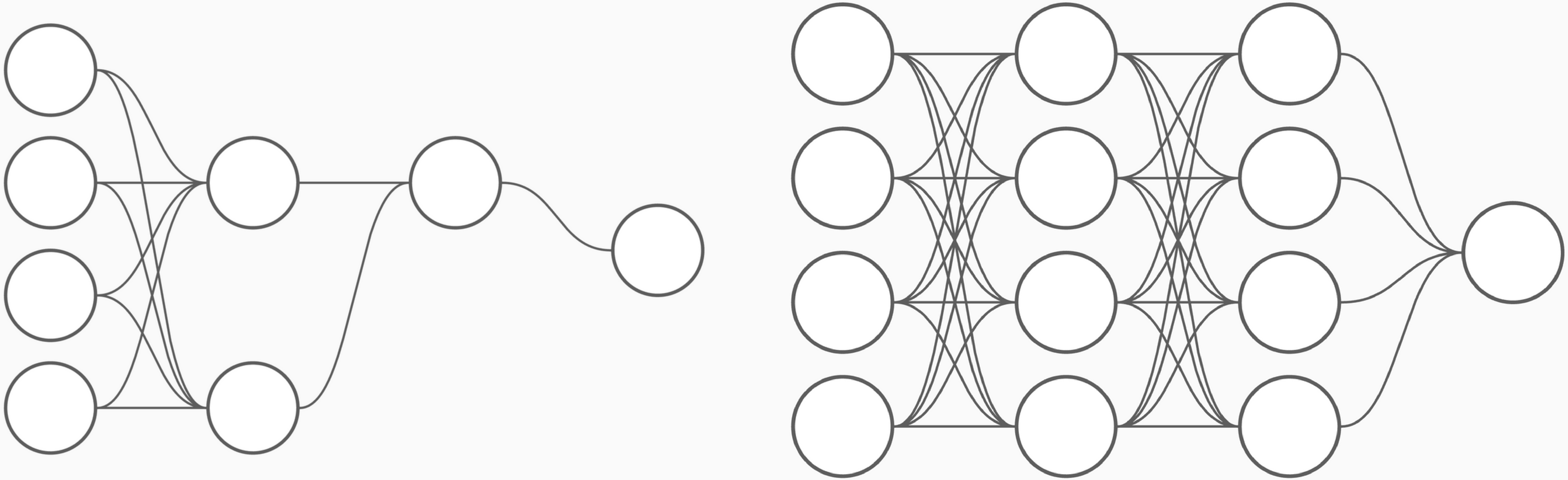
Finding an subnetwork q^* in q (untrained)

SUBNETWORKS



q^* (**untrained**, left) and p (**trained**, right) have comparable performance

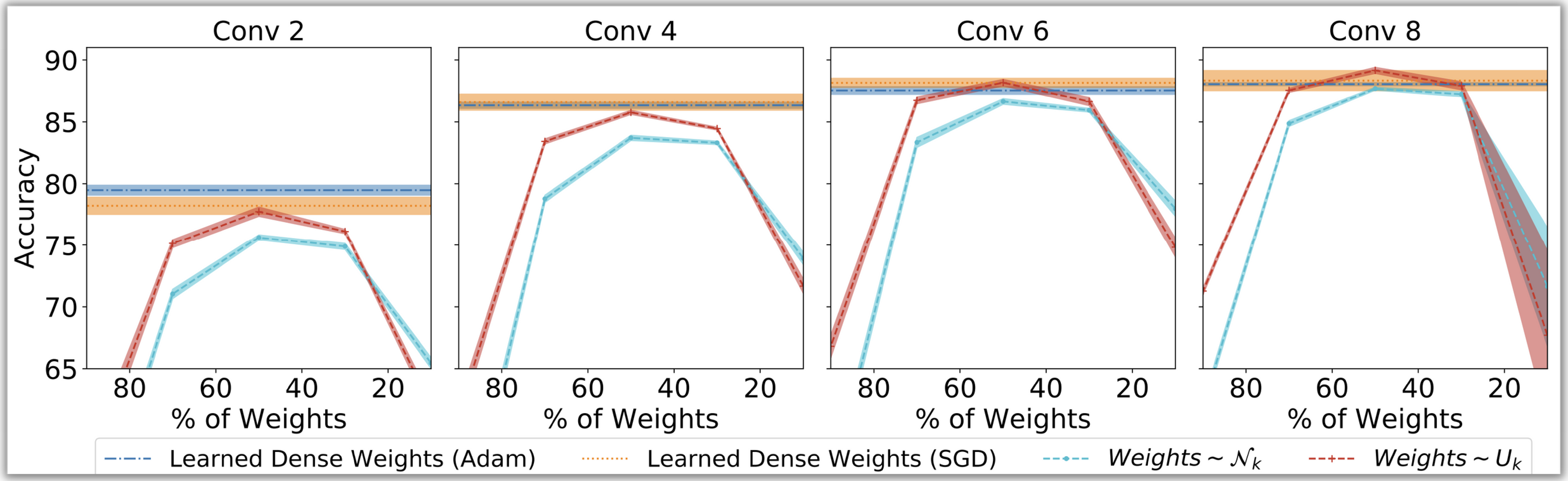
SUBNETWORKS



Sometimes, q^* (**untrained**, left) and a trained version of q (**trained**, right) have comparable performance

Large neural networks contain/encode a 'solution' upon initialization.

REPORTED RESULTS



Straight lines – performance of trained original network q.
Bent lines – performance of selected subnetwork q^* for different sizes.

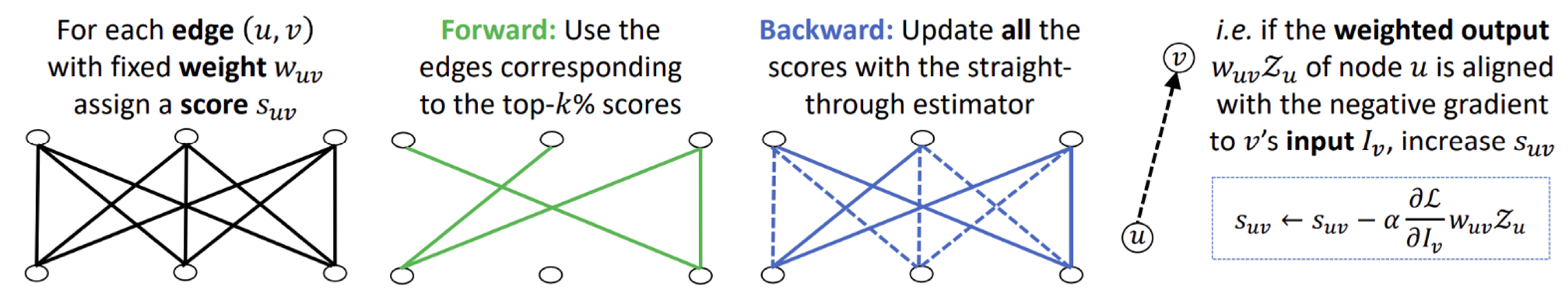
PRUNING

1. Initialize a neural network architecture.
2. Train the network 'a bit' on the dataset.
3. Identify the connections with the least importance.
4. 'Prune' the *k*% least important connections (i.e. fix them at 0).
5. Repeat steps 2-4.

50-90% of parameters in most neural networks with minimal performance impact.

EDGE-POPUP ALGORITHM

'Apply' pruning to an initialized neural network: optimize which connections are kept/pruned to maximize subnetwork performance.



INTUITION

1. Consider an initialized network *N*.
2. Let τ be *N*, but trained.
3. Let *q* be the probability that a subnetwork of *N* obtains the same or better performance as τ.
4. *q* is very small, but nonzero. The probability that a subnetwork of *N* does not obtain same or better performance is $(1-q)$.
5. The probability that no subnetwork of *N* obtains the same or better performance to τ is $(1-q)^s$, where *s* is the # subnetworks.
6. As *s* increases w/ network size, the chance *any* subnetwork of *N* performs just as well as τ becomes fairly high.

NAPKIN CALCULATIONS

$$(99.999\%)^{500000} < 0.1\%$$

$q = 99.999\%;\ s = 500{,}000$

(Combinatorics blow up quickly. $10! \approx 3.6$m.)

Neural network training as a process of discovery rather than one of update.

Discussion Questions/Topics

- Do the results surprise you? Why/why not?
- Implications for how we think about NNs
- How is this (not) related to learning in the brain?
    - Parallels in neuroscience - randomness, learning, understanding, update vs. discovery.
- In what ways does this study demonstrate desirable & undesirable attributes of neural networks?

More Reading

- "What's Hidden in a Randomly Weighted Neural Network?" Ramanujan & Wortsman et al. arXiv, 2020. https://arxiv.org/pdf/1911.13299.pdf.
- "The Lottery Ticket Hypothesis: Finding Sparse, Trainable Neural Networks". Frankle & Carbin. arXiv, 2019. https://arxiv.org/pdf/1803.03635.pdf.
- "Understanding Deep Learning Requires Re-thinking Generalization". Zhang et al. arXiv, 2017. https://arxiv.org/pdf/1611.03530.pdf.
- "Distinct Sources of Deterministic and Stochastic Components of Action Timing Decisions in Rodent Frontal Cortex". Murakami et al. Neuron, 2017. https://doi.org/10.1016/j.neuron.2017.04.040.
- "Individual Differences Among Deep Neural Network Models". Mehrer & Kietzmann et al. Nature, 2020. https://www.nature.com/articles/s41467-020-19632-w.
- "Artificial Neural Nets Finally Yield Clues to How Brains Learn". Ananthaswamy. Quanta, 2021. https://www.quantamagazine.org/artificial-neural-nets-finally-yield-clues-to-how-brains-learn-20210218/.