Simplifying My Model! Sparsity and Beyond

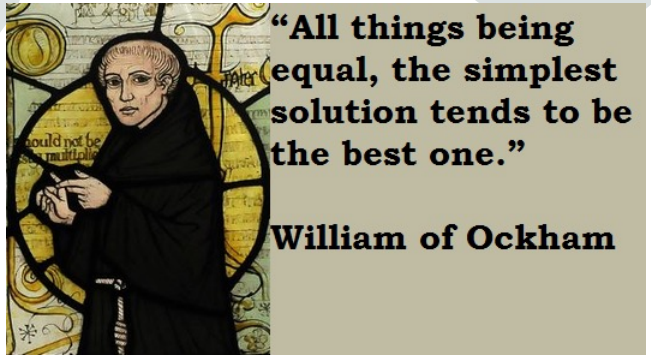
Zhangyang “Atlas” Wang, ECE@UT Austin

VITA group: https://vita-group.github.io/

GitHub: https://github.com/VITA-Group

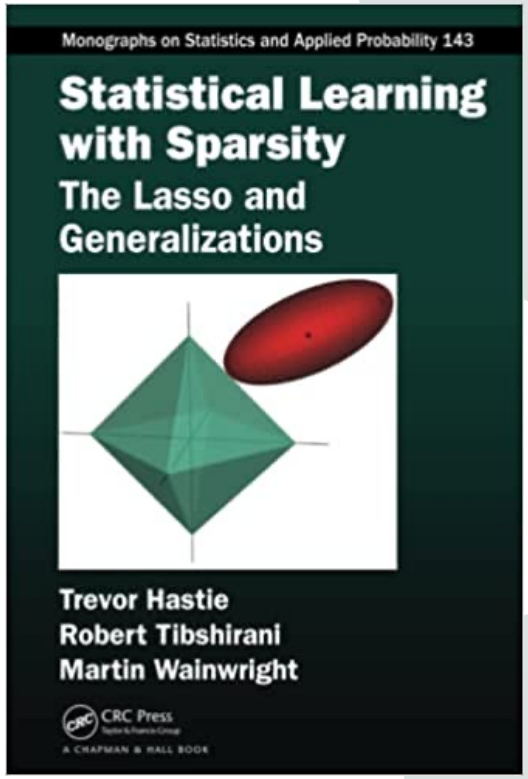
VITA

Everybody knows sparsity, more or less

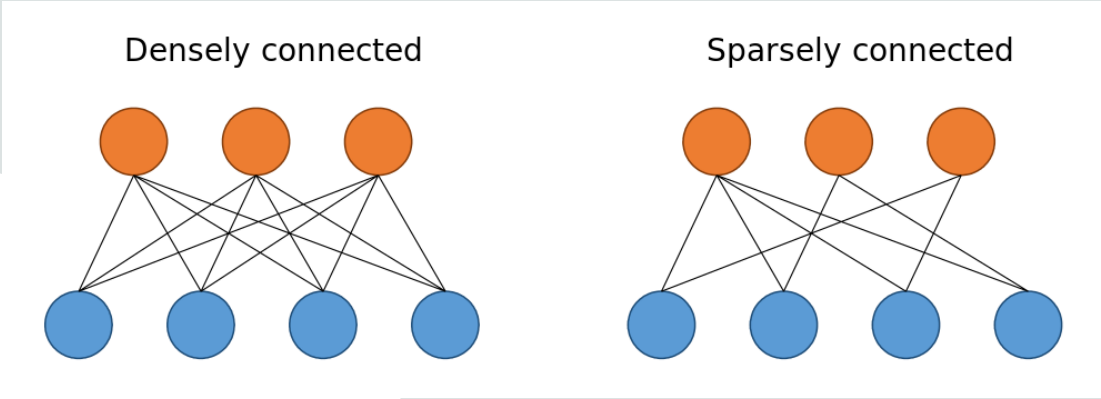


Start from philosophers and artists…

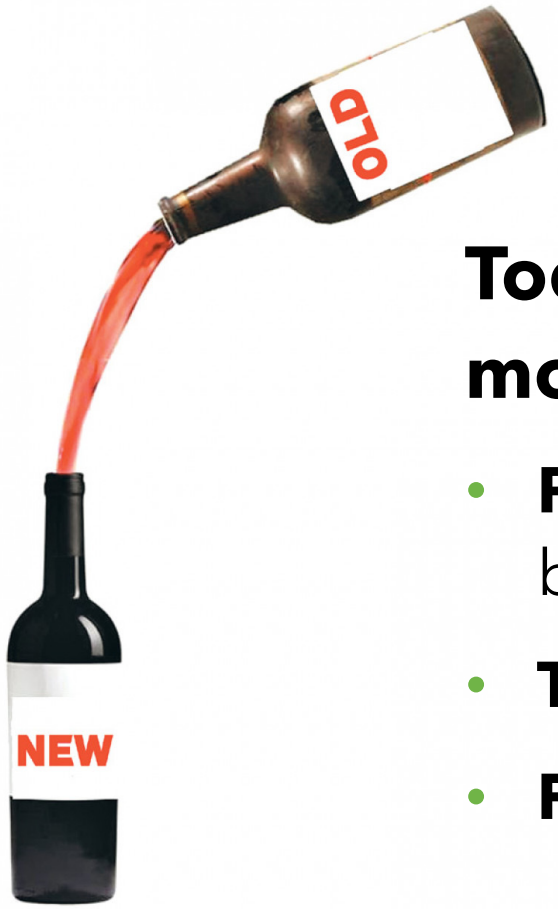
Well-known in the ML world…



Even in the deep learning world…



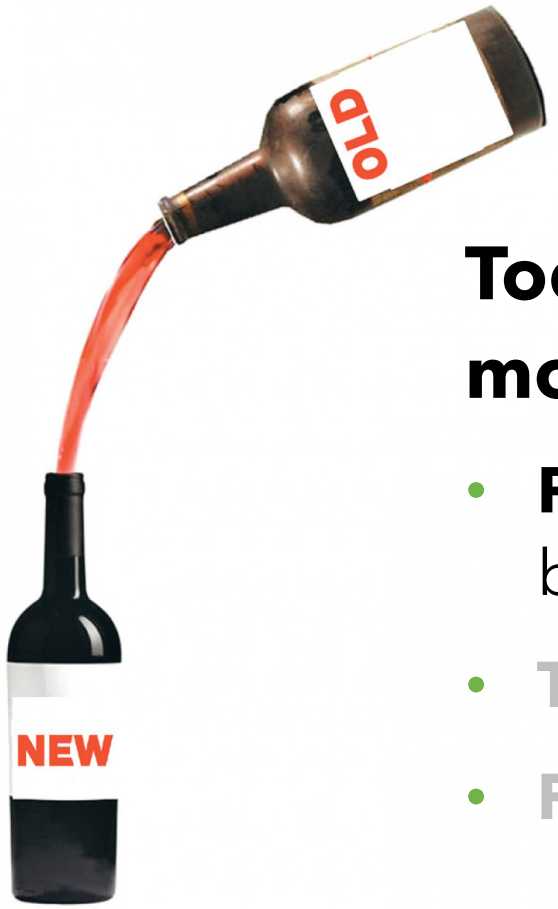
So, anything new that sparsity can offer?



Today we talk about the opportunities of sparsity in modern deep learning

- **Practically**, why we should love a ***sparse neural network*** (NN), beyond just a way of "model compression"
- **Theoretically**, what guarantees we can expect from sparse NNs
- **Future** – what is the new prospect of exploiting sparsity

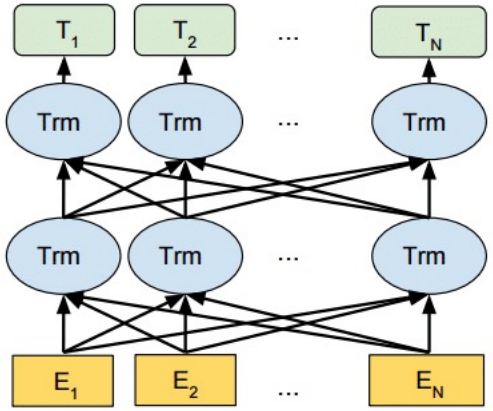
So, anything new that sparsity can offer?

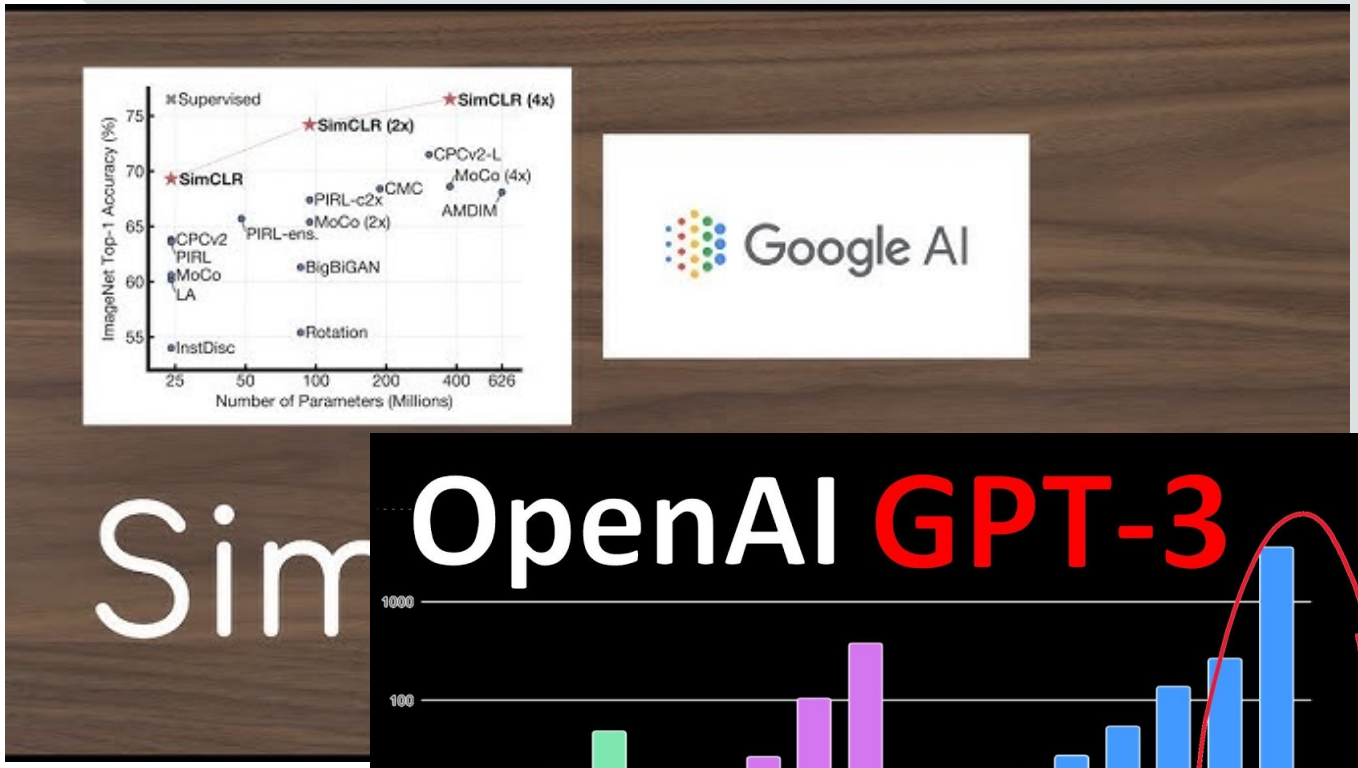


Today we talk about the opportunities of sparsity in modern deep learning

- **Practically**, why we should love a ***sparse neural network*** (NN), beyond just a way of "model compression"
- **Theoretically**, what guarantees we can expect from sparse NNs
- **Future** – what is the new prospect of exploiting sparsity

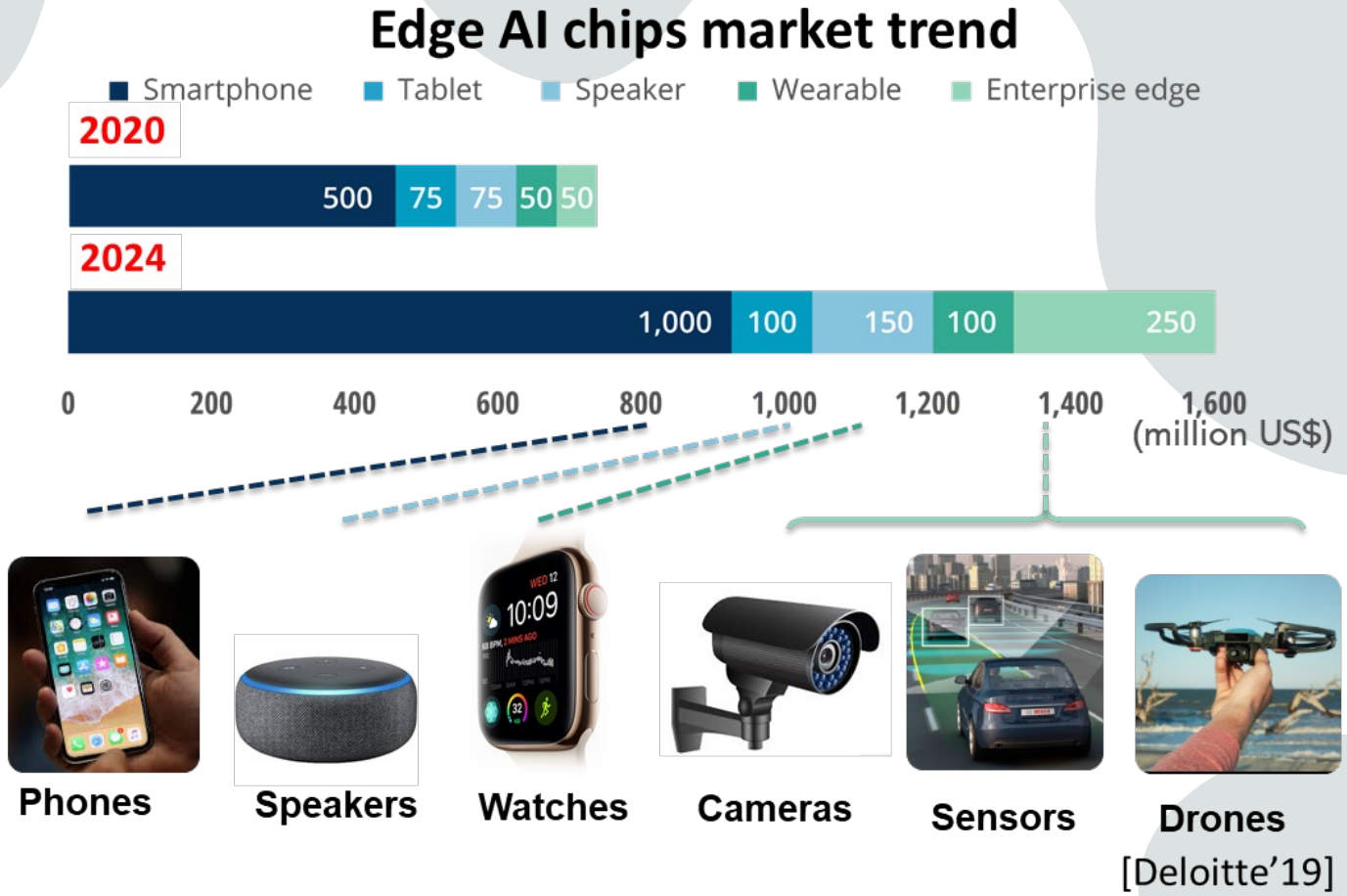
ML researchers like to go BIG



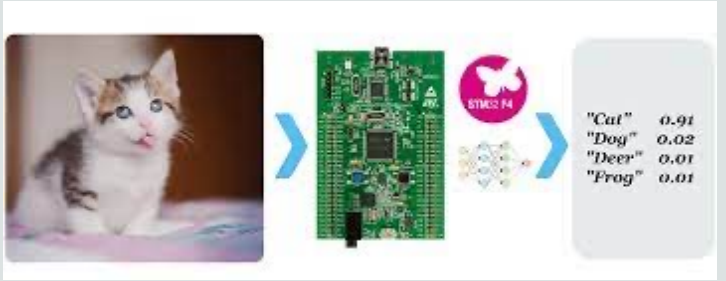


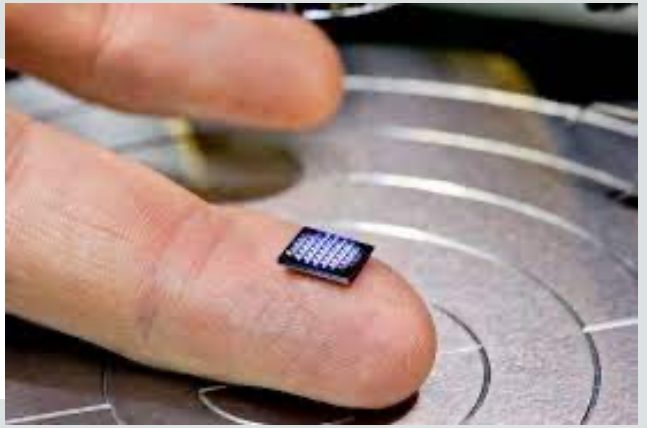
Big NNs seem to be more capable at everything…

….While the world prefers going TINY

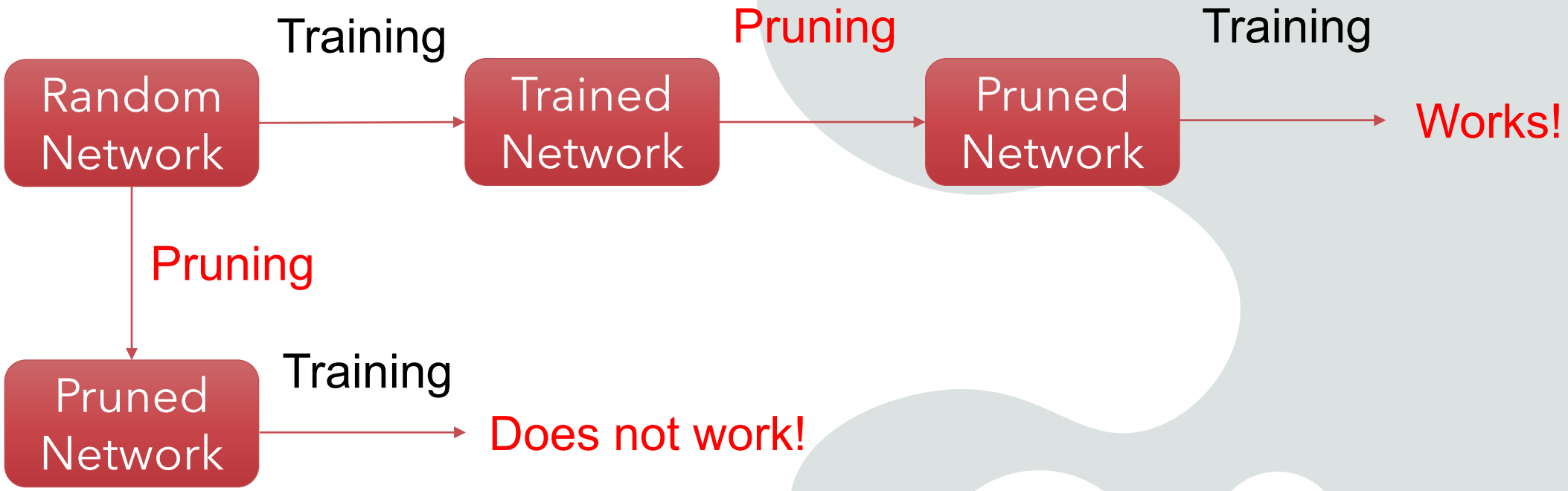


[Deloitte'19]

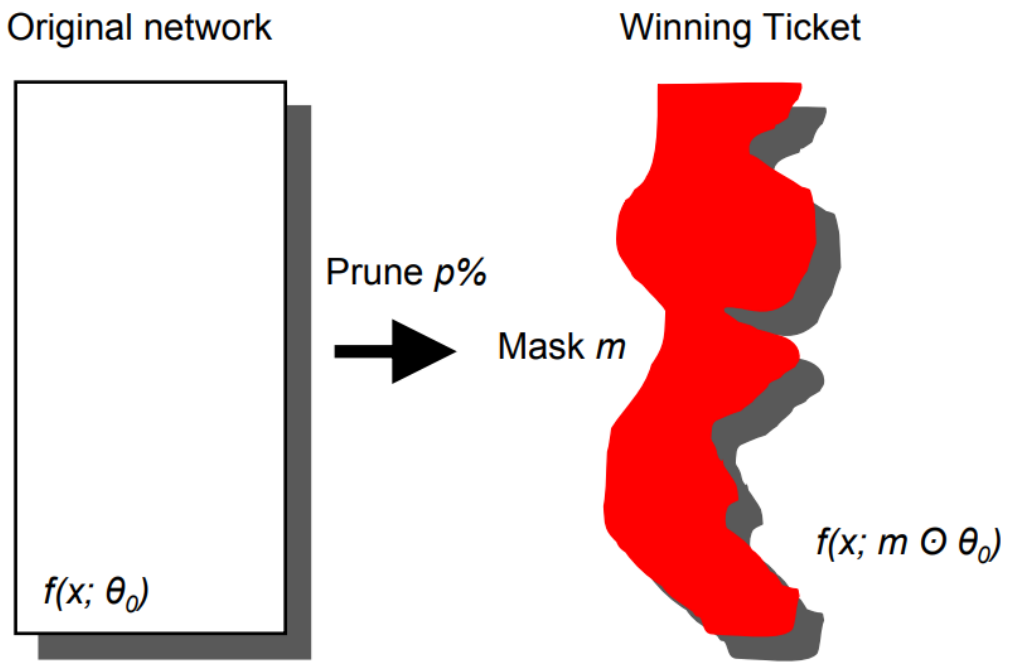




“Old-Fashioned” Sparsity for NNs



The Lottery Ticket Hypothesis. *A randomly-initialized, dense neural network 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.*

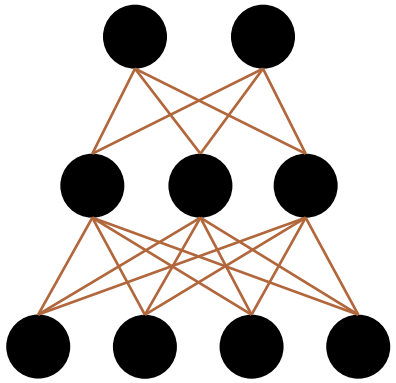


- Winning Ticket gives
  - Better or same results
  - Shorter or same training time
  - Notably fewer parameters
  - Is trainable from the beginning

…As long as we know which sub-network is winning!

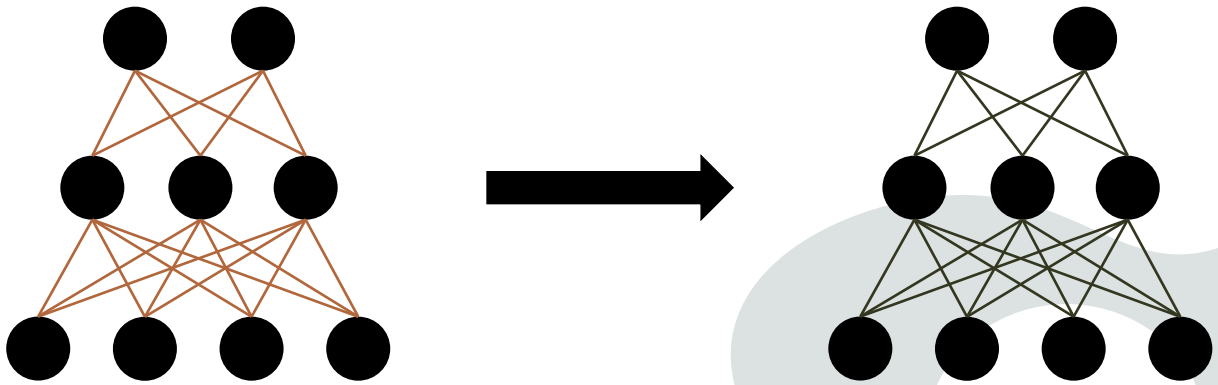
Iterative Magnitude Pruning

a) Randomly initialize a dense network
b) Train it like normal and prune unimportant weights
c) Reset remaining weights to their values from a) exactly
d) Repeat steps b-c) iteratively



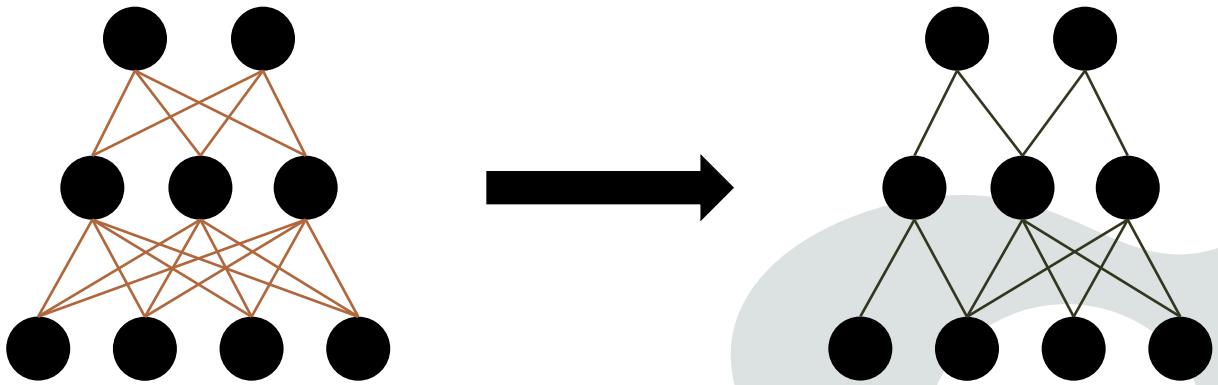
Iterative Magnitude Pruning

a) Randomly initialize a dense network
b) Train it like normal and prune unimportant weights
c) Reset remaining weights to their values from a) exactly
d) Repeat steps b-c) iteratively



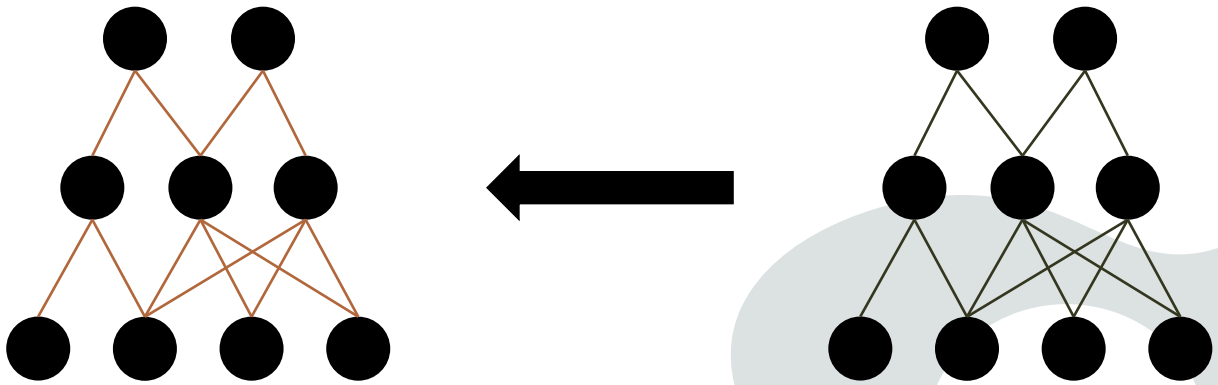
Iterative Magnitude Pruning

a) Randomly initialize a dense network
b) Train it like normal and prune unimportant weights
c) Reset remaining weights to their values from a) exactly
d) Repeat steps b-c) iteratively



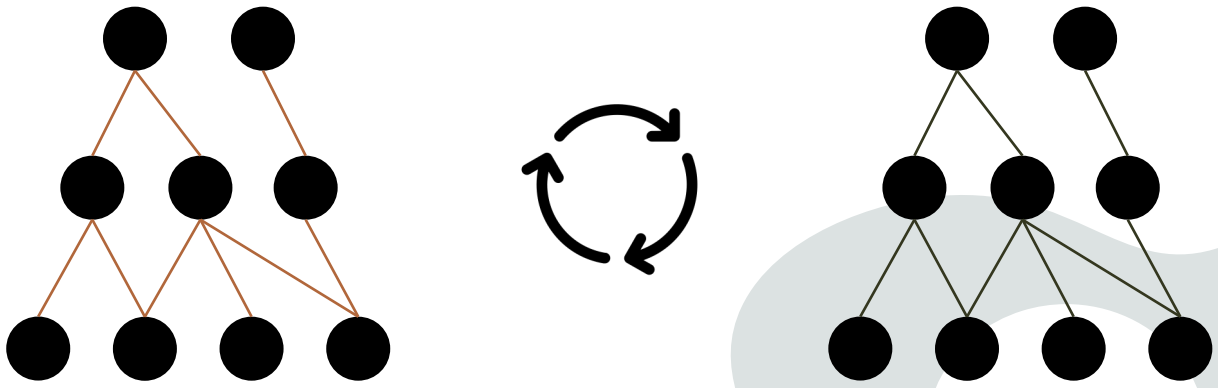
Iterative Magnitude Pruning

a) Randomly initialize a dense network
b) Train it like normal and prune unimportant weights
c) Reset remaining weights to their values from a) exactly
d) Repeat steps b-c) iteratively



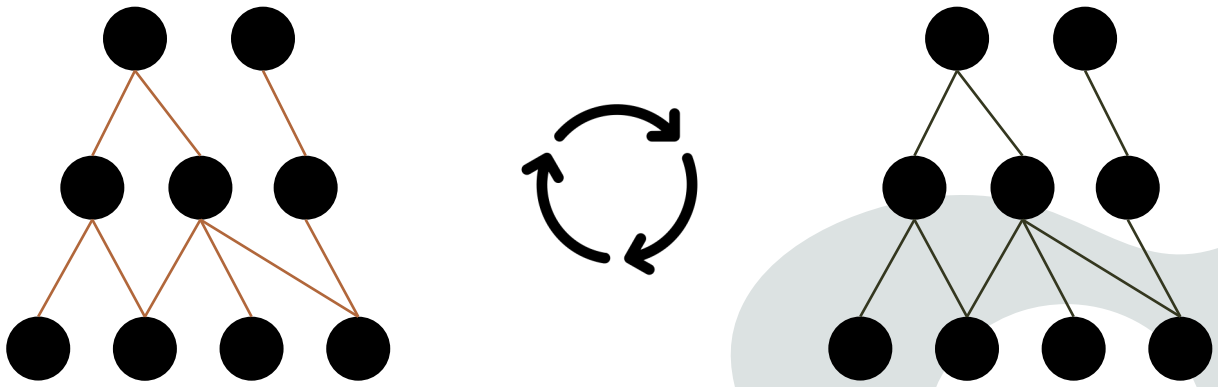
Iterative Magnitude Pruning

a) Randomly initialize a dense network
b) Train it like normal and prune unimportant weights
c) Reset remaining weights to their values from a) exactly
d) Repeat steps b-c) iteratively



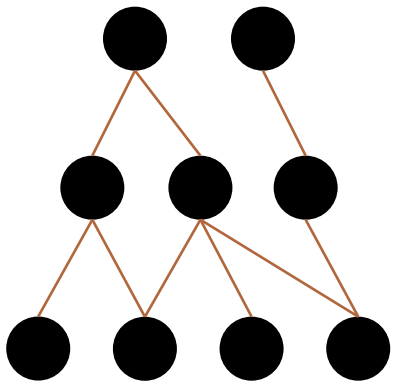
Iterative Magnitude Pruning

a) Randomly initialize a dense network
b) Train it like normal and prune unimportant weights
c) Reset remaining weights to their values from a) exactly
d) Repeat steps b-c) iteratively



Iterative Magnitude Pruning

a) Randomly initialize a dense network
b) Train it like normal and prune unimportant weights
c) Reset remaining weights to their values from a)
d) Repeat steps b-c) iteratively

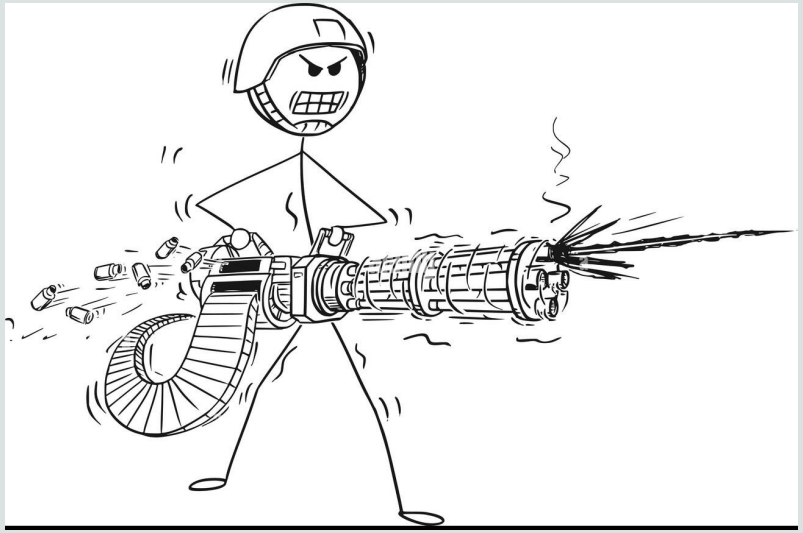


Winning Tickets!

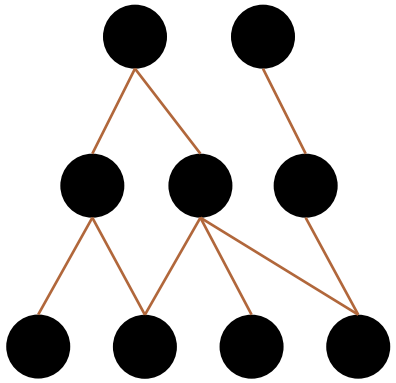
Sparse mask

Original initialization

The Problem?



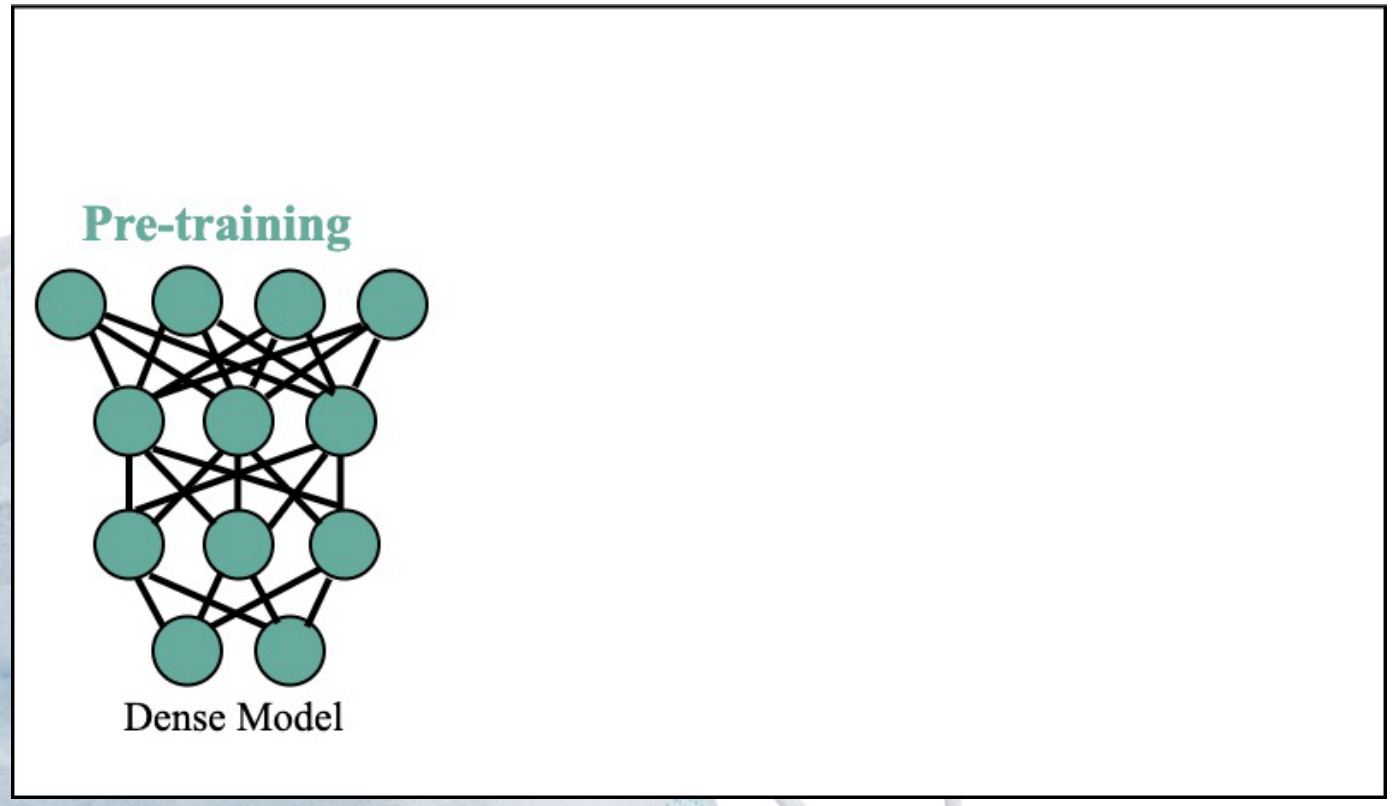
Obtaining this good sparse mask is expensive...

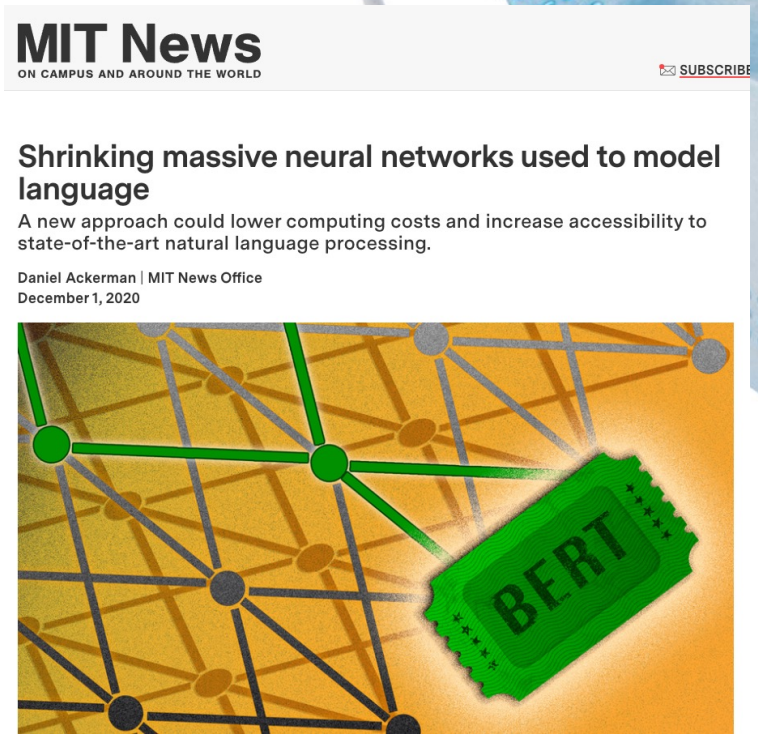


Winning Tickets!

- **Sparse mask**
- Original initialization

Finding Mask is Expensive? **Re-using a pre-made mask**

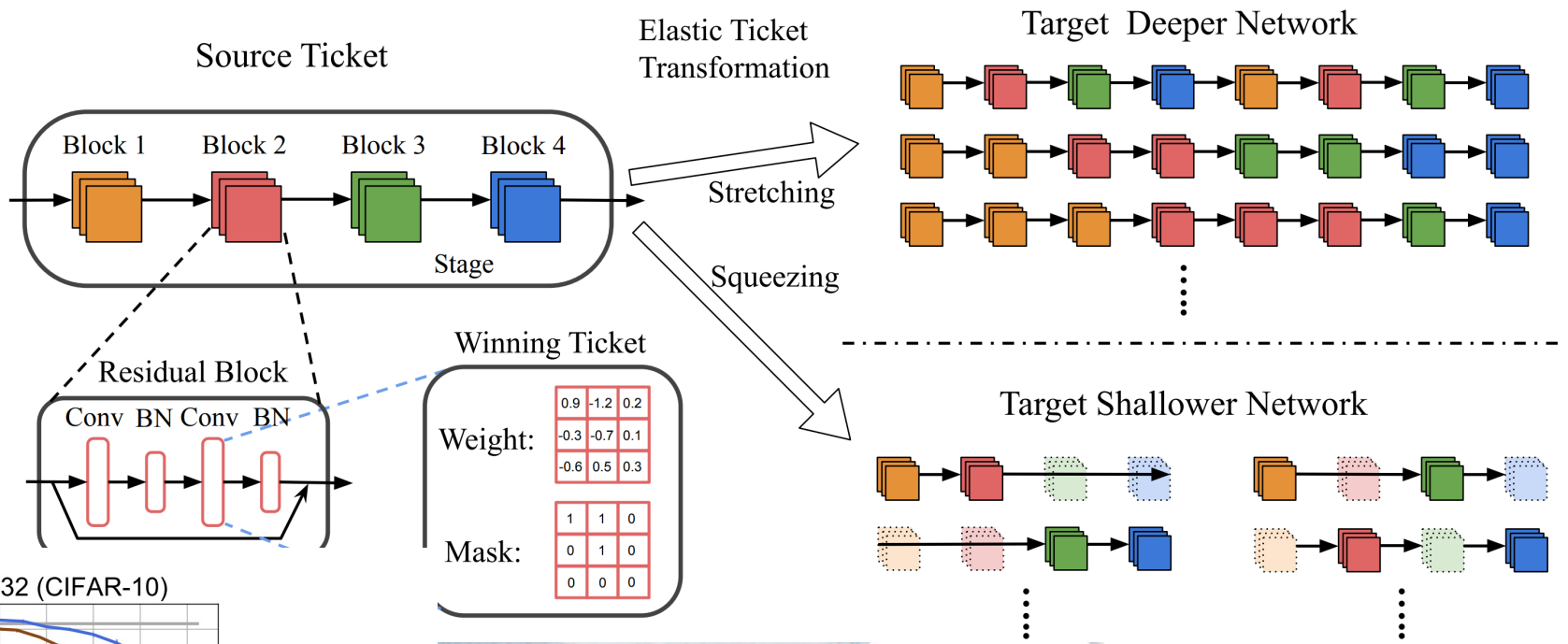


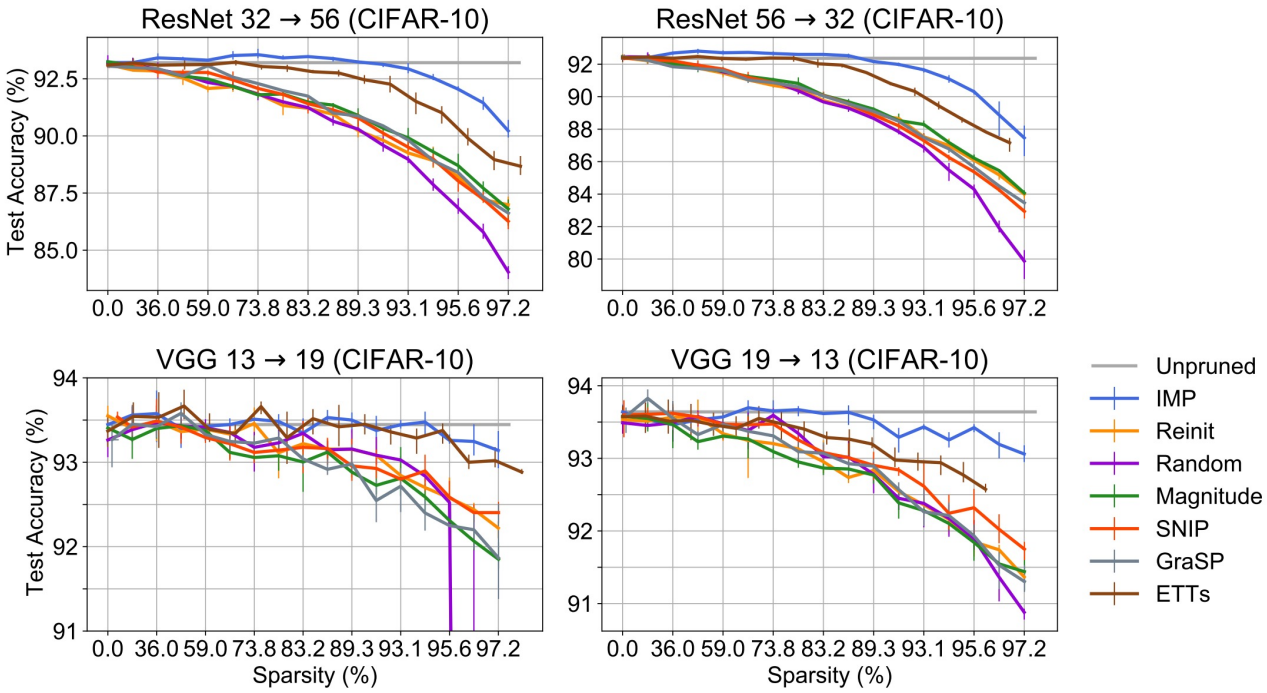


Take Home Message: IMP can find you a good mask on pre-trained models (supervised or self-supervised), in NLP, CV and even multi-modality, so the sparse subnetwork is the same transferrable!

- T. Chen, J. Frankle, S. Chang, S. Liu, Y. Zhang, M. Carbin, and Z. Wang, *"The Lottery Tickets Hypothesis for Supervised and Self-supervised Pre-training in Computer Vision Models"*, **CVPR 2021**
- T. Chen, J. Frankle, S. Chang, S. Liu, Y. Zhang, Z. Wang, and M. Carbin, *"The Lottery Ticket Hypothesis for Pre-trained BERT Networks"*, **NeurIPS 2020.**

Finding Mask is Expensive? **Your Mask can be Elastic**





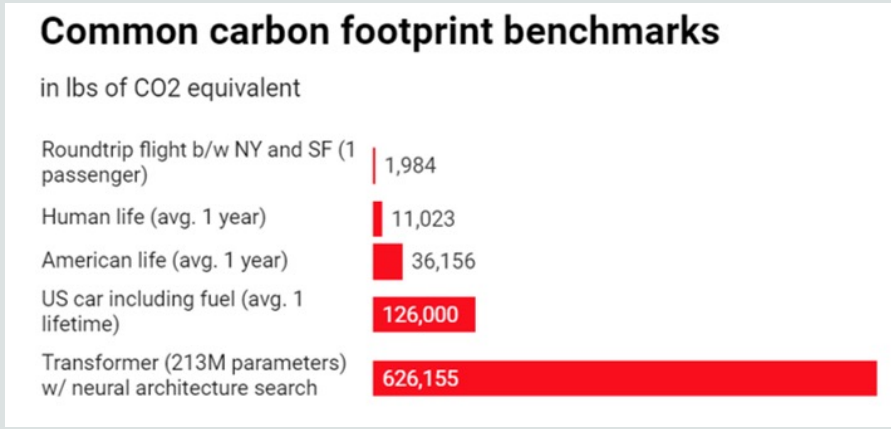
Take Home Message: we can mindfully "transform" the winning ticket found in one network, to another with a different but related architecture, directly yielding a winning ticket for the latter without re-doing the expensive IMP

- X. Chen, Y. Cheng, S. Wang, Z. Gan, J. Liu, and Z. Wang, "The Elastic Lottery Ticket Hypothesis", **NeurIPS 2021.**

Training Big NNs is Expensive.
Can Sparsity Help?

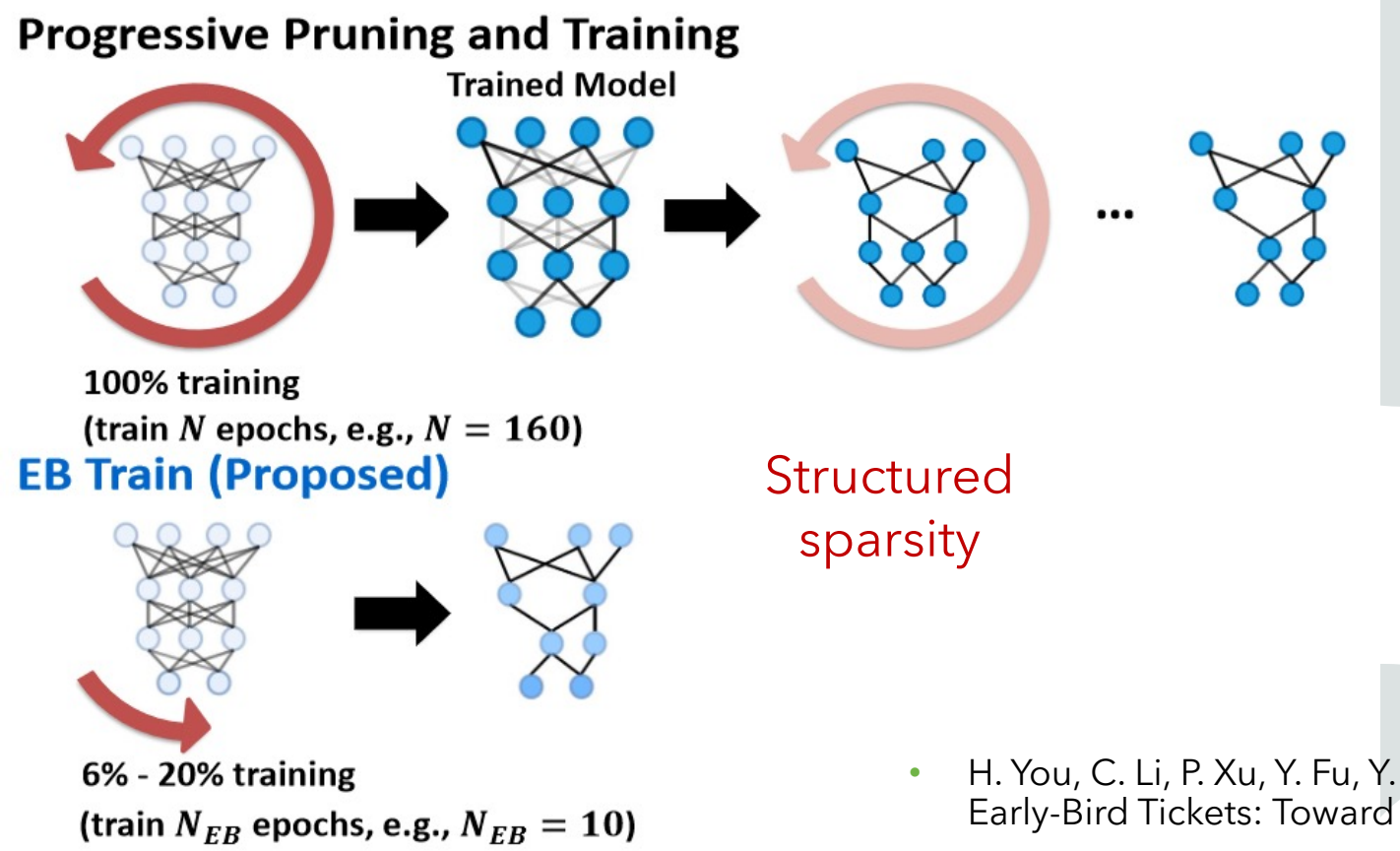
Energy and Policy Considerations for Deep Learning in NLP

Emma Strubell **Ananya Ganesh** **Andrew McCallum**
College of Information and Computer Sciences
University of Massachusetts Amherst
{strubell, aganesh, mccallum}@cs.umass.edu



Model	Hardware	Power (W)	Hours	kWh·PUE	CO_2e	Cloud compute cost
$\text{Transformer}_{base}$	P100x8	1415.78	12	27	26	$41–$140
Transformer_{big}	P100x8	1515.43	84	201	192	$289–$981
ELMo	P100x3	517.66	336	275	262	$433–$1472
BERT_{base}	V100x64	12,041.51	79	1507	1438	$3751–$12,571
BERT_{base}	TPUv2x16	—	96	—	—	$2074–$6912
NAS	P100x8	1515.43	274,120	656,347	626,155	$942,973–$3,201,722
NAS	TPUv2x1	—	32,623	—	—	$44,055–$146,848
GPT-2	TPUv3x32	—	168	—	—	$12,902–$43,008

Surprise - Good Sparsity Can be Found Early!



- In the first 15-20% of total epochs, the "important" connection subset already emerges and keeps stable!
- **So, pruning can happen early!**
- Training an "early emerging" sparse NN is more efficient not in not only resource, but also data
- Validated on CNNs, BERT & ViTs

- H. You, C. Li, P. Xu, Y. Fu, Y. Wang, X. Chen, R. Baraniuk, Z. Wang, and Y. Lin, "Drawing Early-Bird Tickets: Toward More Efficient Training of Deep Networks", **ICLR 2020**
- X. Chen, Y. Cheng, S. Wang, Z. Gan, Z. Wang, and J. Liu, "EarlyBERT: Efficient BERT Training via Early-Bird Lottery Tickets", **ACL 2021.**

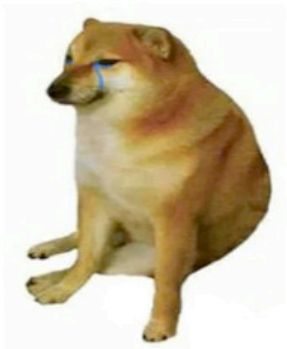
Wait … Do we really need pay the tax of finding expensive “good mask”?

What is the **naivest and cheapest** baseline for producing “sparse NNs”?

(Perhaps) True for post-training pruning….

All the new, fancy pruning methods that can publish in top conferences …

Random Pruning



- S. Liu, T. Chen, X. Chen, L. Shen, D. Mocanu, Z. Wang, and M. Pechenizkiy, “The Unreasonable Effectiveness of Random Pruning: Return of the Most Naive Baseline for Sparse Training”, **ICLR 2022**

Surprise Again- No more paying for the tax!

What is the **naivest and cheapest** baseline for producing "sparse NNs"?

Not necessary, if training a sparse NN from the scratch! **Random pruning can be as good**

All the new, fancy pruning methods that can publish in top conferences

Random Pruning

- S. Liu, T. Chen, X. Chen, L. Shen, D. Mocanu, Z. Wang, and M. Pechenizkiy, "The Unreasonable Effectiveness of Random Pruning: Return of the Most Naive Baseline for Sparse Training", **ICLR 2022**

What really matters for good sparse training?

The network size matters the most

- As the original dense networks grow **wider** and deeper, the performance of training a randomly pruned sparse network will quickly match that of **its dense equivalent**, *besides just matching other pruning competitors*, even at high sparsity ratios

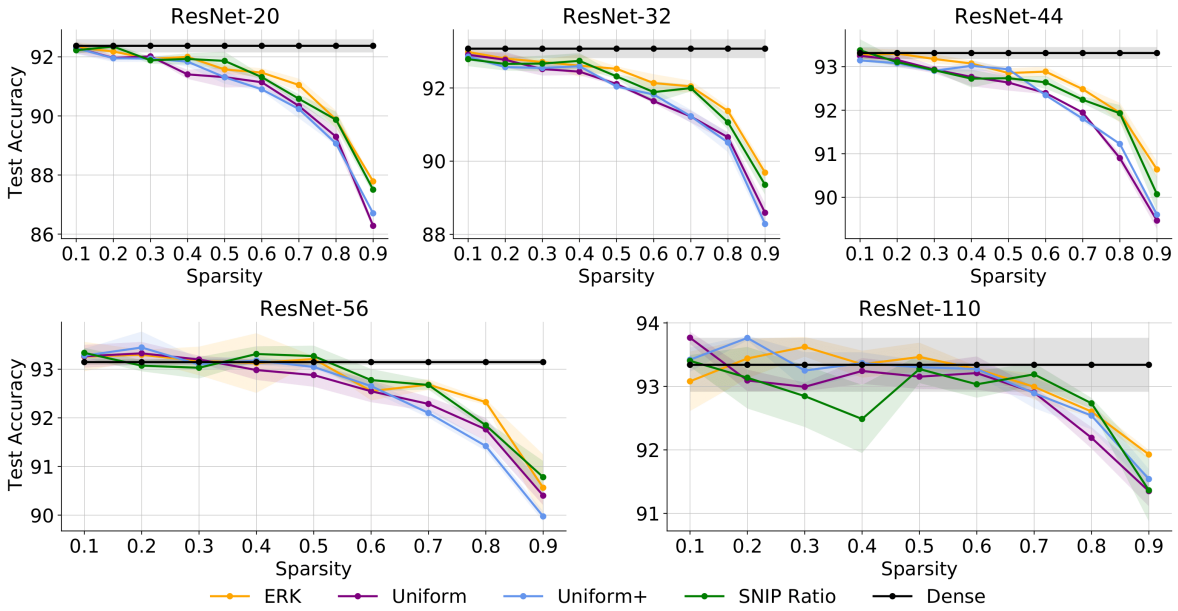
Appropriate layer-wise sparsity ratios can be pre-configured

- We choose Erdos-Renyi-Kernel (ERK) ratio: *larger layers tend to be sparser*

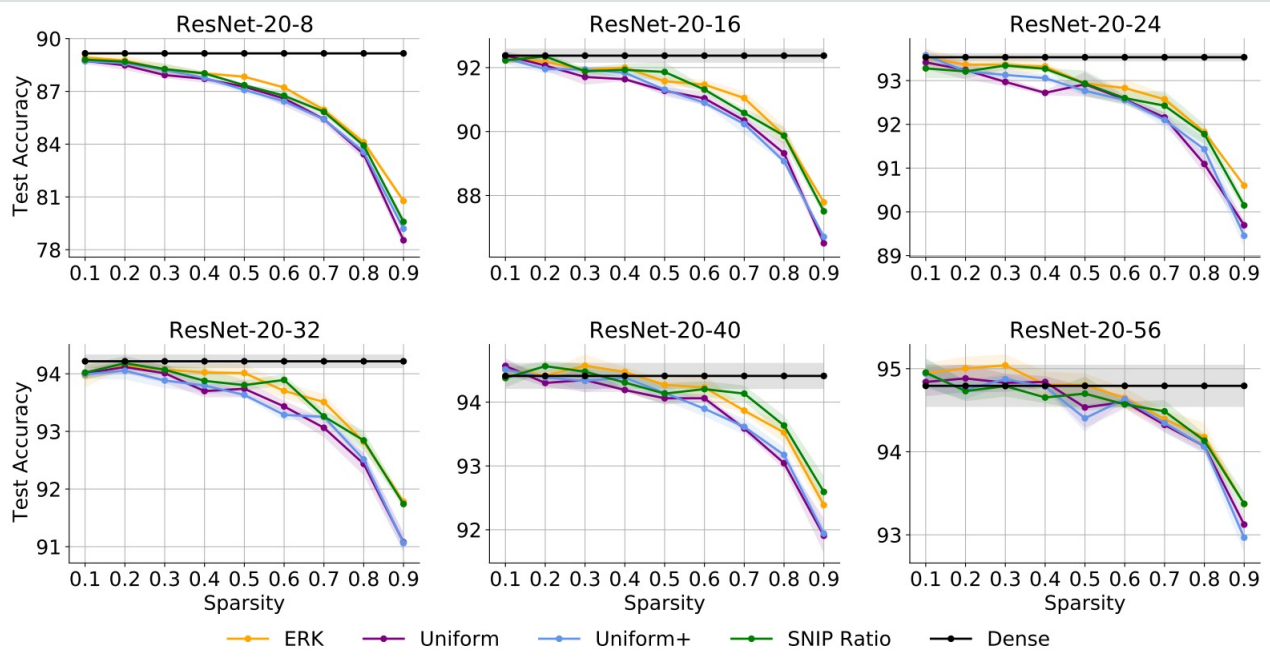
- S. Liu, T. Chen, X. Chen, L. Shen, D. Mocanu, Z. Wang, and M. Pechenizkiy, "The Unreasonable Effectiveness of Random Pruning: Return of the Most Naive Baseline for Sparse Training", **ICLR 2022**

Take-Home: a randomly pruned subnetwork of Wide ResNet-50 can be sparsely trained to match the accuracy of a dense Wide ResNet-50, on ImageNet

- … Plus superior out-of-distribution detection, uncertainty estimation, and adversarial robustness
- Performance gap between different sparse masks becomes negligible as models get bigger



From shallow to deep



From narrow to wide

- S. Liu, T. Chen, X. Chen, L. Shen, D. Mocanu, Z. Wang, and M. Pechenizkiy, "The Unreasonable Effectiveness of Random Pruning: Return of the Most Naive Baseline for Sparse Training", **ICLR 2022**

More good things are yet to uncover…

Sparse NN appears to be **more data-efficient** (*only verified under SGD, yet to find a proof…*)

- Empirically, validated on both image generation (NeurIPS'21) and recognition (under review) tasks. Also good for communication efficiency (AAAI'22)
- Theoretically, related to the neuroscience foundation of modularity and neural-constraint theory – we are collaborating with BCI experts now…

Sparse NN also appears to be **more robust**

- Empirically, validated for adversarial attacks (ICLR'22), backdoor attacks (CVPR'22), and outlier detection (under review). Even better if ensembled (ICLR'22)
- Theoretically, certified robustness also improves! (ICML'22)

So, anything new that sparsity can offer?

Today we talk about the opportunities of sparsity in modern deep learning

- **Practically**, why we should love a ***sparse neural network*** (NN), beyond just a way of "model compression"
- **Theoretically**, what guarantees we can expect from sparse NNs
- **Future** – what is the new prospect of exploiting sparsity

Optimization Analysis of Sparse NNs

- We study the behavior of ultra-wide neural networks when their weights are randomly pruned at the initialization, through the lens of *neural tangent kernels* (**NTKs**)
  - Once the width of a neural network approaches infinity, training will incur increasingly smaller changes and the neural network *f* starts to behave like a simple kernel defined as:

$$\ker(\mathbf{x}, \mathbf{x}') = \left\langle \frac{\partial f(\boldsymbol{\theta}, \mathbf{x})}{\partial \boldsymbol{\theta}}, \frac{\partial f(\boldsymbol{\theta}, \mathbf{x}')}{\partial \boldsymbol{\theta}} \right\rangle$$

  where θ is the parameters of the neural network (Jacot et. al. 2018; Arora et. al. 2019; …)

- Since random pruning is the cheapest avenue towards sparsity, **how good a random pruned subnetwork could actually be, compared to the original unpruned network**, even in a simplified analysis framework of ultra-wide NNs?
  - *Plus, we already observed the strong experiment results in wide NNs!*

Asymptotic Convergence

- The asymptotic analysis is straightforward
  - The weights being pruned are staying at zero always during the training process, so the gradient of the pruned NN is simply the masked gradient of the unpruned NN

Theorem 1.1 (The NTK of randomly pruned networks). *Consider an L-hidden-layer fully connected neural network with ReLU activation. Suppose we prune the weights in this neural network except the input layer i.i.d. with probability $1-\alpha$ at the initialization. Then, as the width of each layer goes to infinity sequentially,*

$$\lim_{d_1,d_2,\ldots,d_L\to\infty} \tilde{\Theta}(\mathbf{x},\mathbf{x}') = \alpha^L \Theta_\infty(\mathbf{x},\mathbf{x}')$$

If we perform random pruning at the initialization, as the NN width approaches infinity, the empirical NTK of the pruned NN indeed approaches the original one!

- ...up to some multiplicative scaling factor that depends on the pruning probability
- Implying **unimpaired optimization + generalization** of sparse NN w.r.t. original

- H. Yang and Z. Wang, "On the Neural Tangent Kernel Analysis of Randomly Pruned Wide Neural Networks", arXiv 2022

Non-Asymptotic Convergence

- Once we fix pruning probabilities, *how wide the pruned NN needs to be*, in order to approach the NTK regime and show its effective optimization/generalization?
- **It is misleading to think this non-asymptotic proof can be trivially adapted!**

Theorem 1.2 (Non-asymptotic Bound of Pruned Network's NTK). *Consider an L-hidden-layer fully connected neural network with all the weights initialized with i.i.d. standard Gaussian distribution. Suppose all the weights except the input layer are pruned with probability $1-\alpha$ at the initialization and after pruning we rescale the weights by $1/\sqrt{\alpha}$. Fix $\epsilon > 0$ and $\delta \in (0,1)$. Suppose we use ReLU activation $\sigma(z) = \max(0,z)$ and $d_h \geq \Omega(\frac{1}{\alpha^2}\frac{L^6}{\epsilon^4}\log\frac{Ld_{h+1}\sum_{i=1}^{h}d_i+\sum_{h'=h}^{L}d_{h'}}{\delta})$. Then for any inputs $\mathbf{x}, \mathbf{x}' \in \mathbb{R}^{d_0}$ such that $\|\mathbf{x}\|_2 \leq 1$, $\|\mathbf{x}'\|_2 \leq 1$, with probability at least $1-\delta$ we have*

$$\left|\left\langle \frac{\partial f(\boldsymbol{\theta},\mathbf{x})}{\partial \boldsymbol{\theta}}, \frac{\partial f(\boldsymbol{\theta},\mathbf{x}')}{\partial \boldsymbol{\theta}} \right\rangle - \boldsymbol{\Theta}_{\infty}(\mathbf{x},\mathbf{x}')\right| \leq (L+1)\epsilon$$

at initializatoin). *Fix $\epsilon > 0$ and $\delta \in (0,1)$. Suppose $(\frac{L^6}{\epsilon^4}\log(L/\delta))$. Then for any inputs $\boldsymbol{x}, \boldsymbol{x}' \in \mathbb{R}^{d_0}$ such that ast $1-\delta$ we have:*

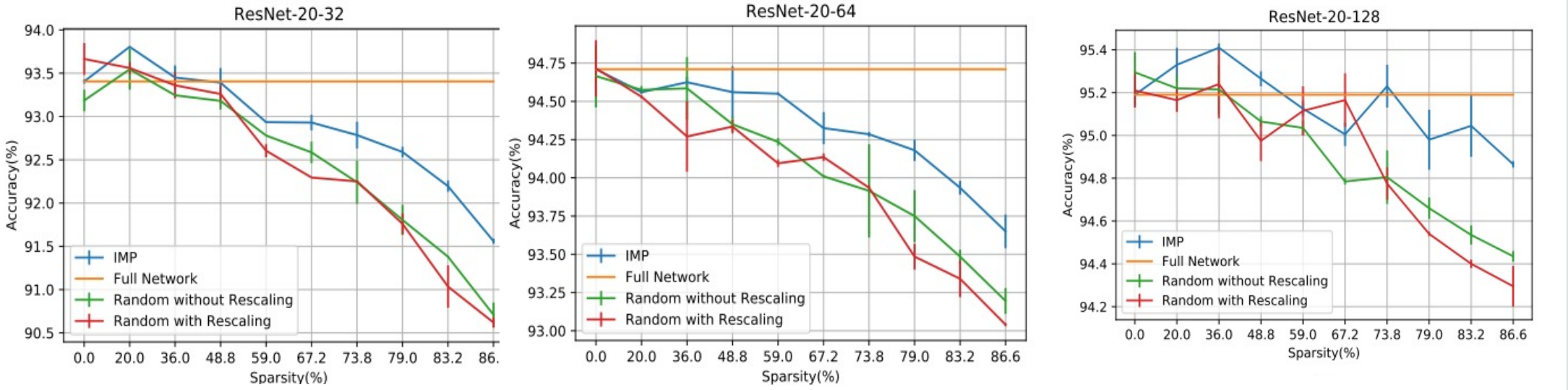
$$\left|\left\langle \frac{\partial f(\boldsymbol{\theta},\boldsymbol{x})}{\partial \boldsymbol{\theta}}, \frac{\partial f(\boldsymbol{\theta},\boldsymbol{x}')}{\partial \boldsymbol{\theta}} \right\rangle - \Theta^{(L)}(\boldsymbol{x},\boldsymbol{x}')\right| \leq (L+1)\epsilon.$$

hask!
er
t
en

- Will need tackle new concentration with Bernoulli

- S. Arora, et. al., "On exact computation with an infinitely wide neural net", **NeurIPS 2019**
- H. Yang and Z. Wang, "On the Neural Tangent Kernel Analysis of Randomly Pruned Wide Neural Networks", arXiv 2022

Results and Implications



Gap up to 2.7%.

Gap up to 1.5%.

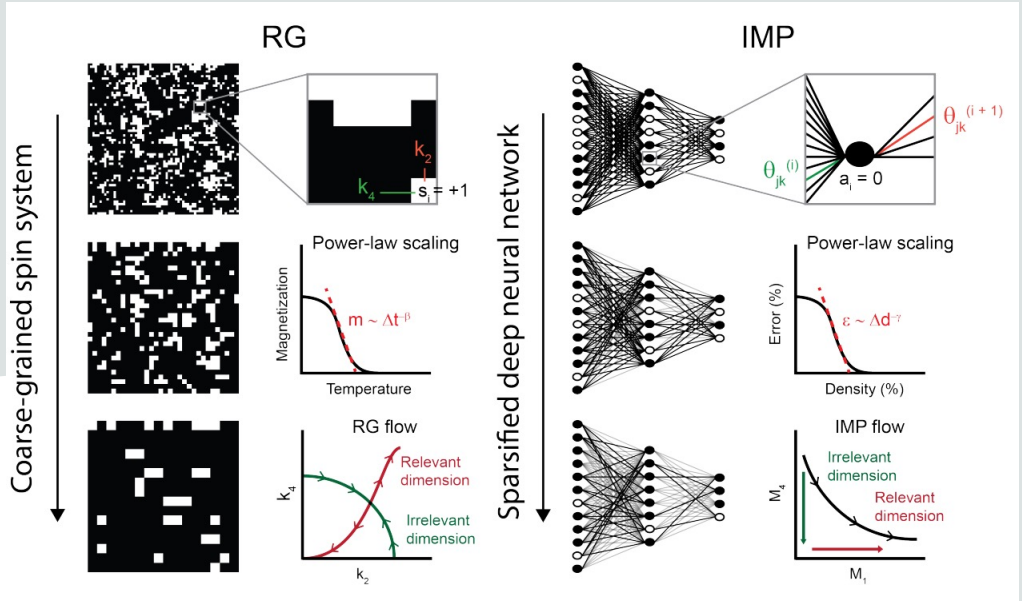
Gap up to 0.8%.

- Both pruned and unpruned NNs, **in the NTK regime**, can achieve similar fast convergence rate to zero training error, and yield similar generalization after training
- For practitioners, this result is likely to **bring random pruning back to the spotlight** of sparse training, as NNs are constantly scaled wider and deeper, e.g., gigantic "foundational" models

What Else "Theory" We Seek for Sparse NNs?

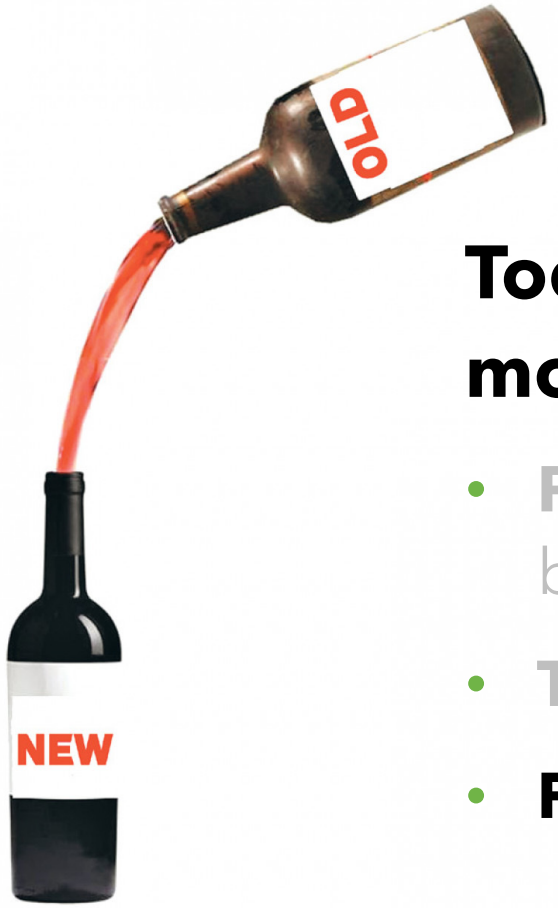
Transferability: why there exist "universal" sparse masks?

- We model the emergence of sparse NNs as physical system phase transition, and use a classical statistical physics tool called *Renormalization Group (RG) / Ising model*
- We prove that IMP is an instantiation of RG operator; and then that lottery ticket represents a fixed point of applying RG operators recursively, which is **dependent** on the architecture topology but (nearly) **independent** of actual weights!
- W. Redman, T. Chen, Z Wang and A. Dogra, *"Universality of Winning Tickets: A Renormalization Group Perspective"*, **ICML 2022**



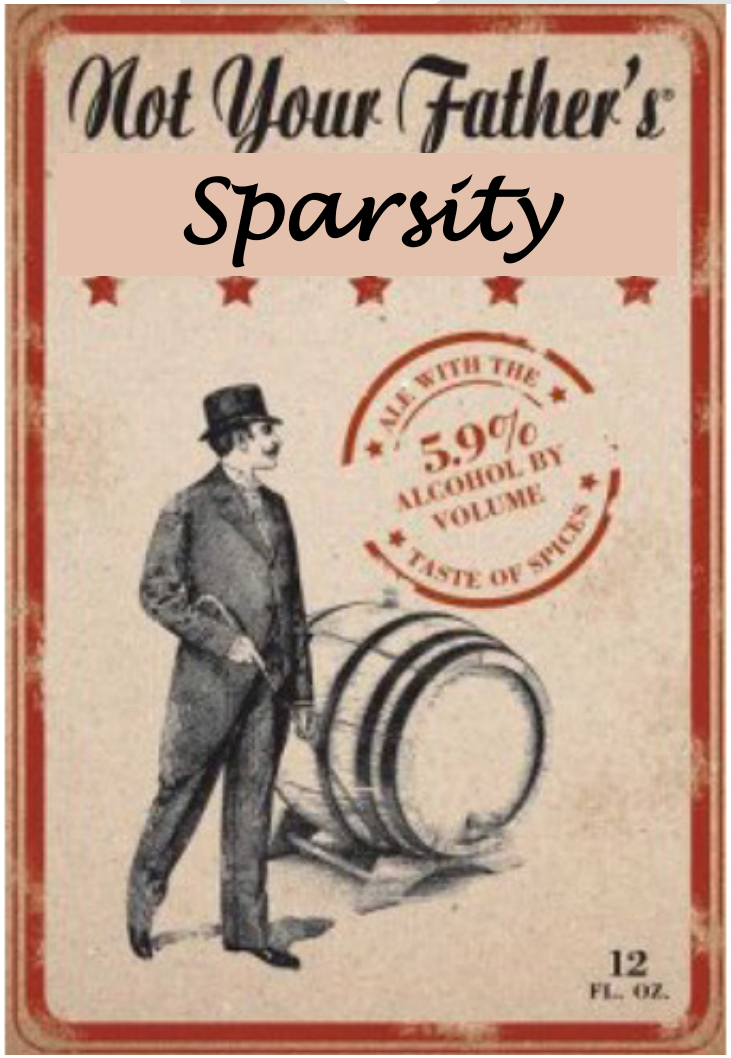
Generalization: sparsity versus overfitting, under overparameterization …

So, anything new that sparsity can offer?



Today we talk about the opportunities of sparsity in modern deep learning

- **Practically**, why we should love a ***sparse neural network*** (NN), beyond just a way of "model compression"
- **Theoretically**, what guarantees we can expect from sparse NNs
- **Future** – what is the new prospect of exploiting sparsity



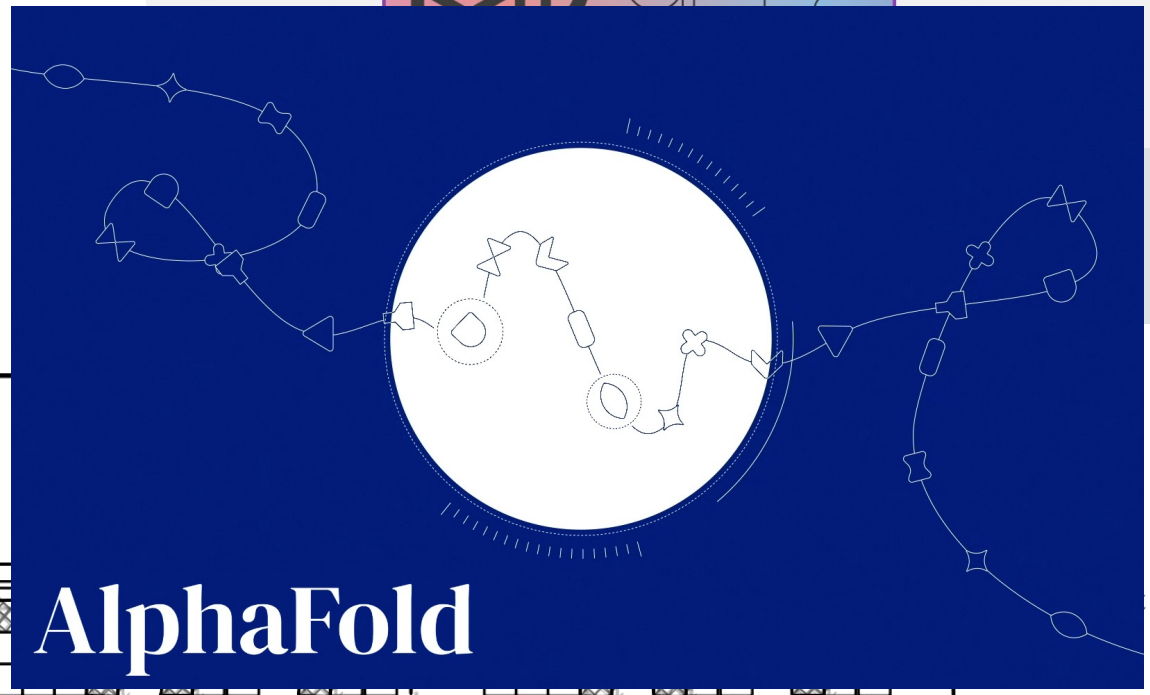
Beyond Sparsity in NNs: New Prospects

Scaling up *Training with Sparsity*

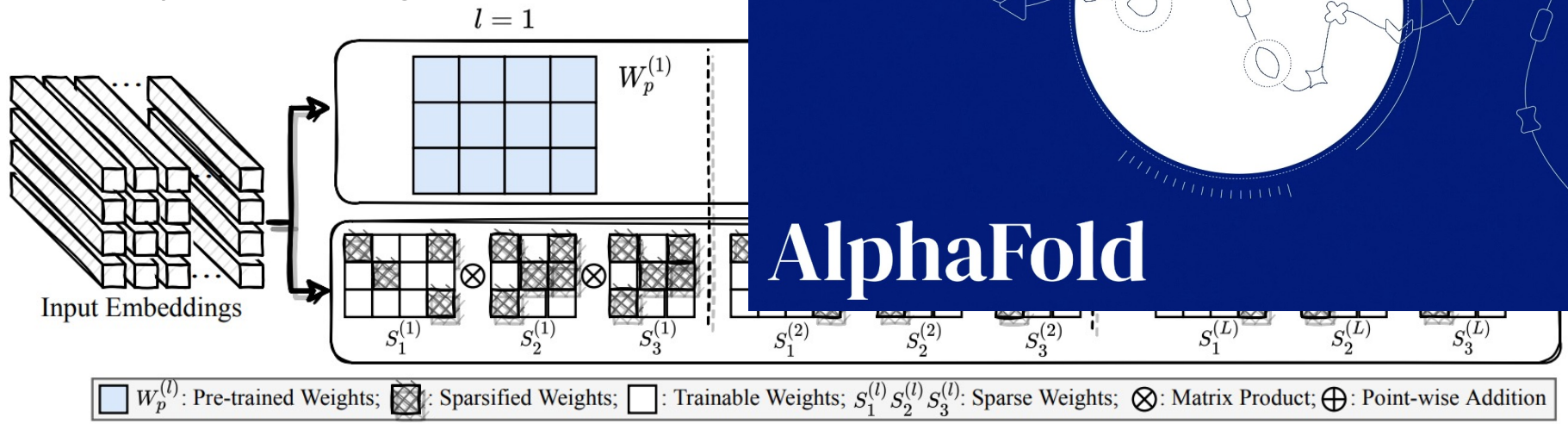
From Static to ***Dynamic*** *Sparsity*

Baking Sparsity into ***New Backbone****!*

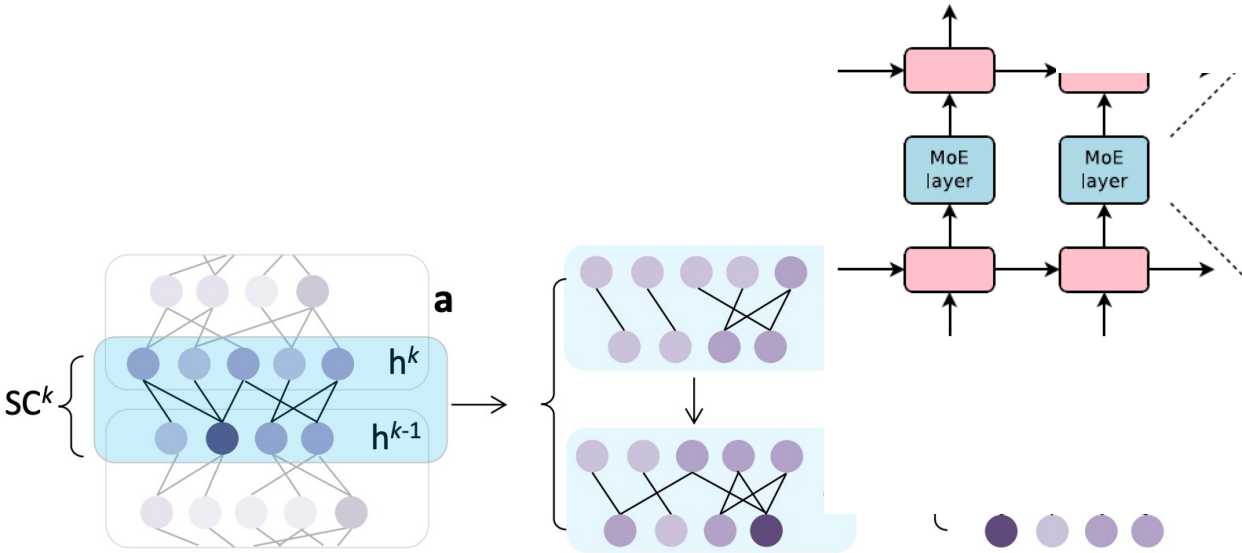
Sparsity is a Bridge towards Scaling-Up Training & Tuning



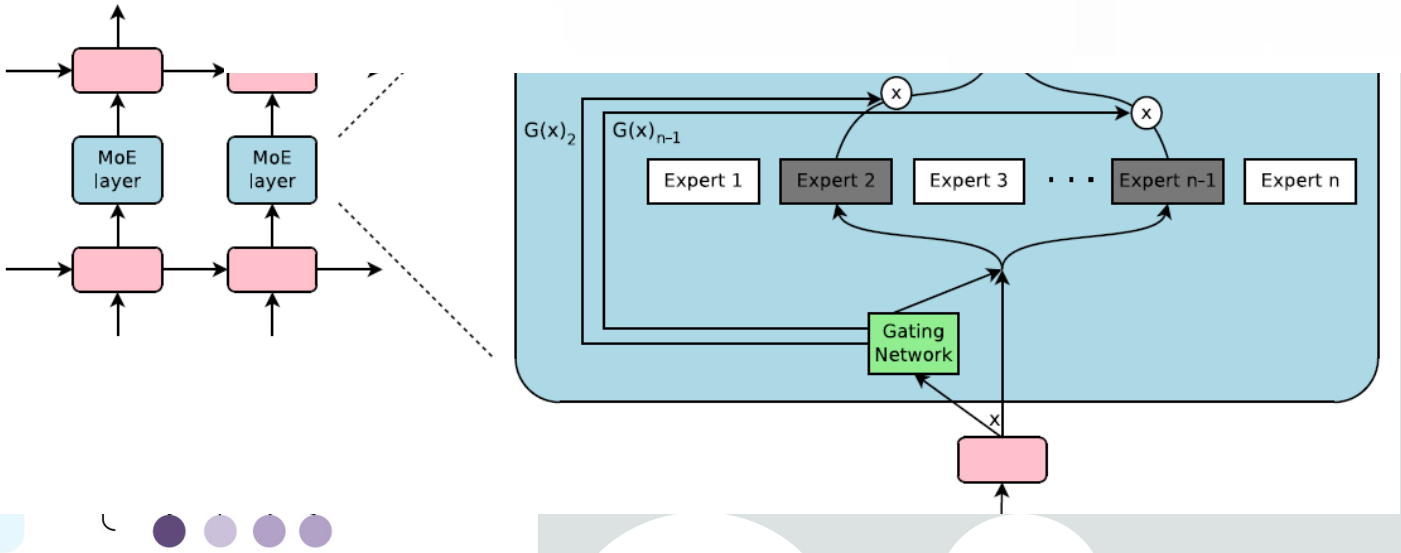
factorized sparse training



Dynamic Sparsity: *You do not wear just one mask!*

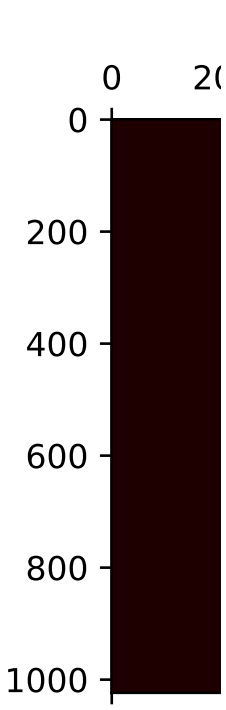


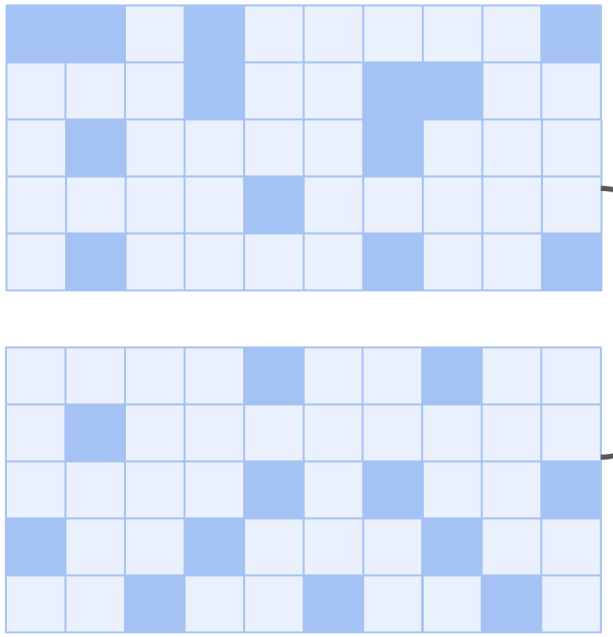
Mocanu, D.C., et al. "Scalable training of artificial neural networks with adaptive sparse connectivity inspired by network science." *Nature communications* 9.1 (2018): 1-12.

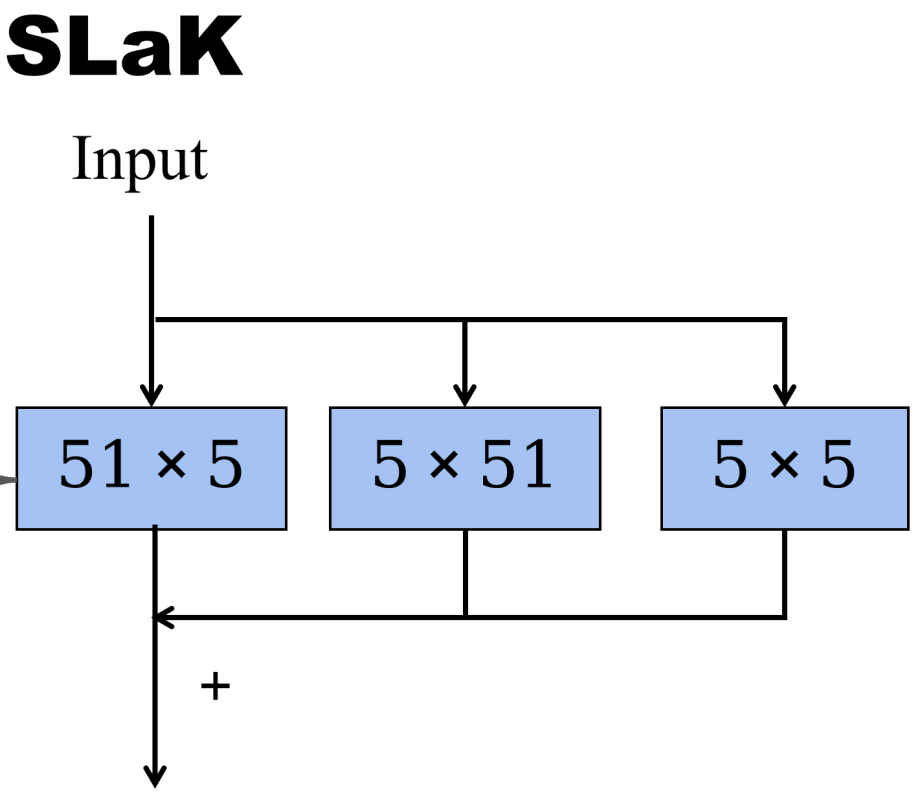


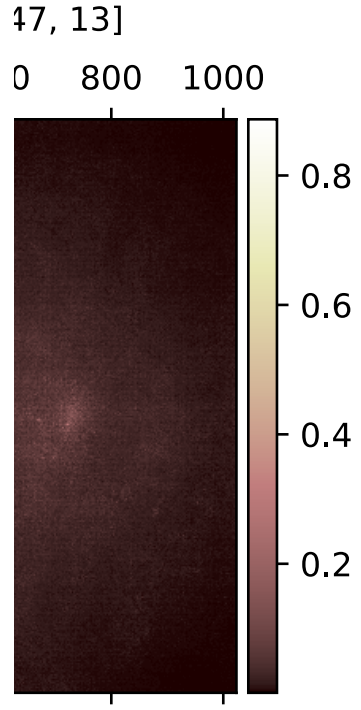
Shazeer M. et. al. "Outrageously Large Neural Networks: The Sparsely-Gated Mixture-of-Experts Layer", ICLR 2017

More ConvNets in the 2020s: Scaling up Kernels Beyond **51 x 51** using Sparsity

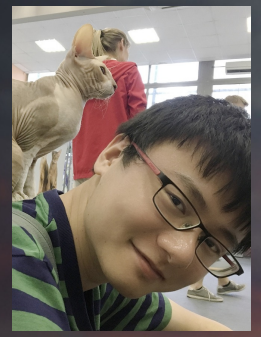






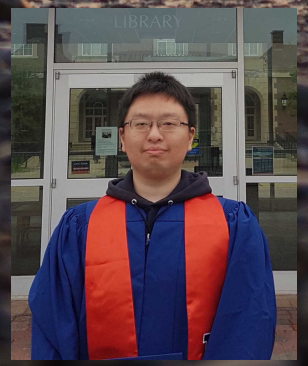


RepLKNe				
Swin-B [5				
ConvNeXt-B [56]	224×224	89M	15.4G	83.8
SLaK-B	224×224	95M	17.1G	**84.0**



I do not know what I may appear to the world, but to myself I seem to have been only like a boy playing on the seashore, and diverting myself in now and then finding a smoother pebble or a prettier shell than ordinary, whilst the great ocean of tr[illegible] lay all undiscovered before me.

Sparsity



THANK YOU!
Q&A, Please