Optimization and Generalization of Neural Networks at the Edge

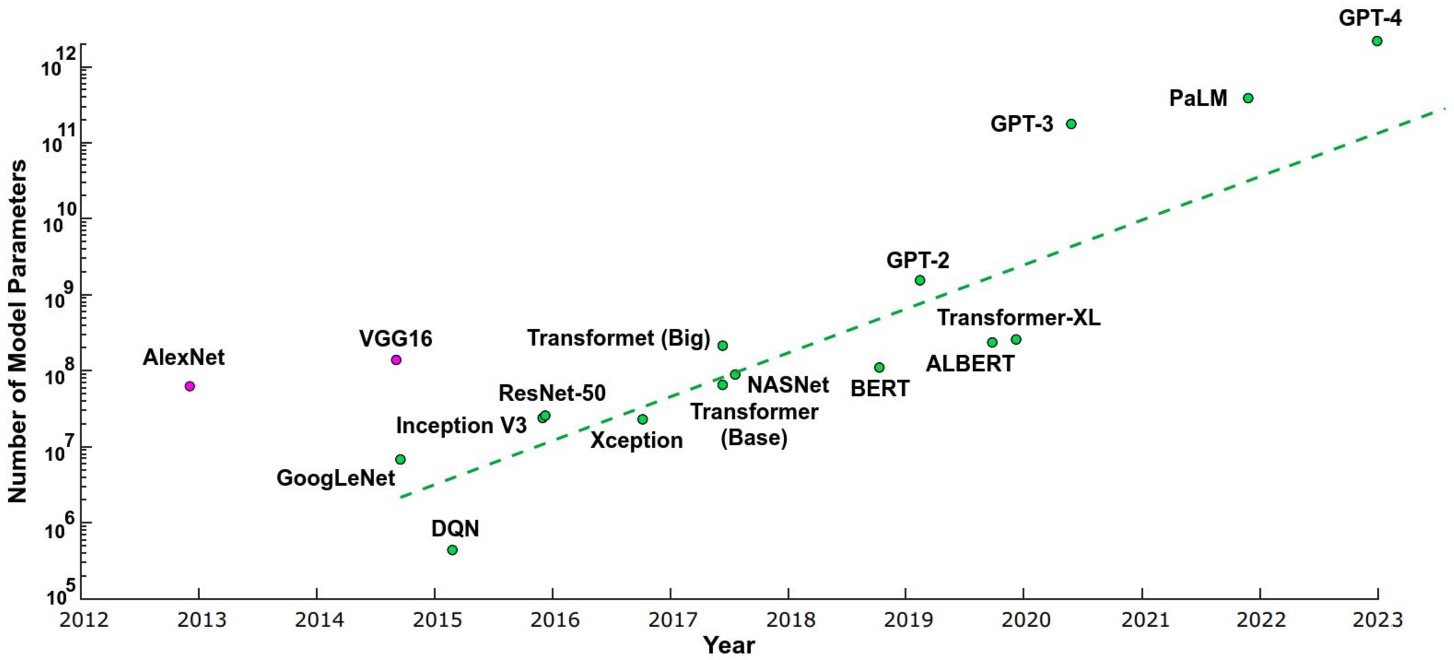
Rahim Entezari

July 17th, 2023

TU Graz

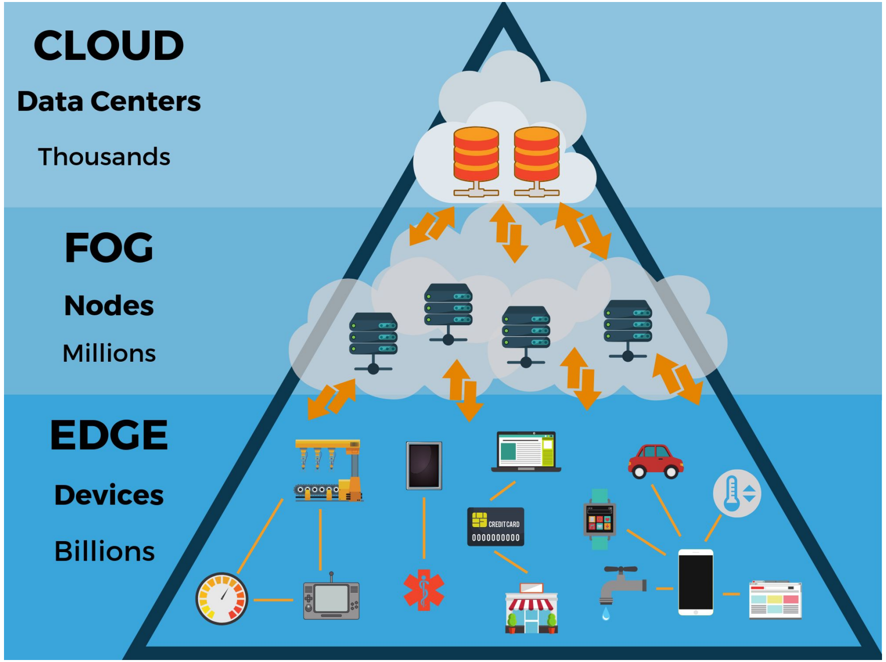
COMPLEXITY SCIENCE HUB VIENNA

Scaling trend



Bernstein, Liane, et al. "Freely scalable and reconfigurable optical hardware for deep learning." *Scientific reports* 11.1 (2021): 3144.
https://the-decoder.com/gpt-4-is-1-76-trillion-parameters-in-size-and-relies-on-30-year-old-technology/

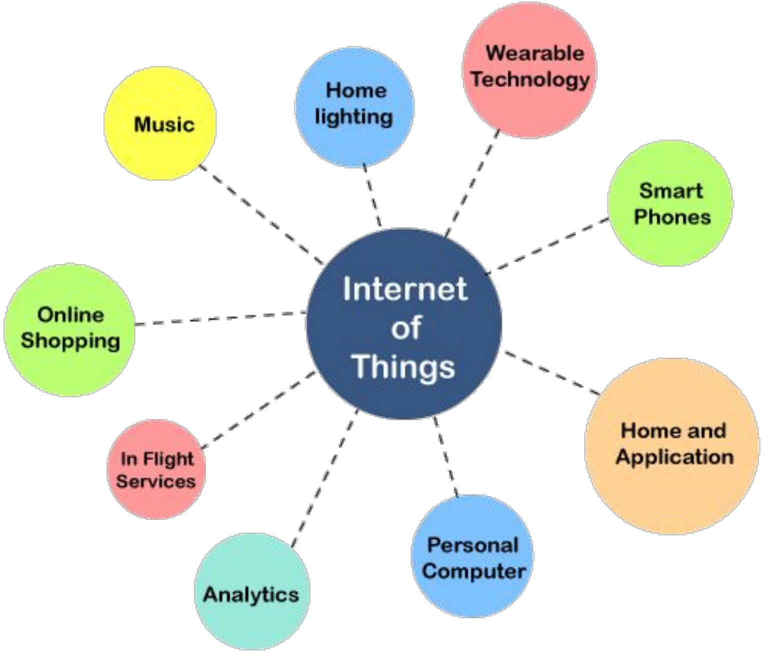
Neural networks at the edge



https://www.thinkebiz.net/what-edge-computing/

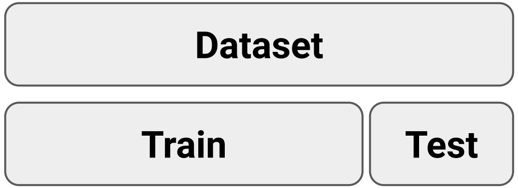
Neural networks at the edge

Pervasive but limited resources → make AI possible on the edge



Generalization

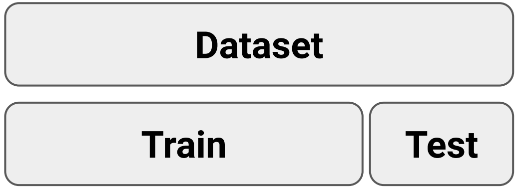
In-Distribution vs. Out-Of-Distribution generalization (ID vs. OOD)



ID generalization

Generalization

In-Distribution vs. Out-Of-Distribution generalization (ID vs. OOD)

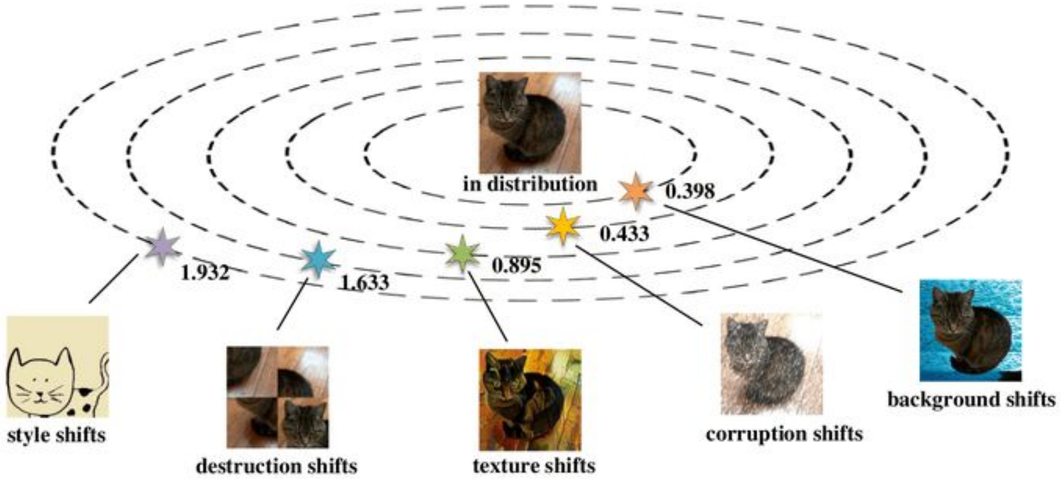


ID generalization

- Regularization
- Dropout
- Early stopping

Generalization

In-Distribution vs. Out-Of-Distribution generalization (ID vs. OOD)



- Transfer learning
- Domain adaptation

Zhang, Chongzhi, et al. "Delving deep into the generalization of vision transformers under distribution shifts."CVPR. 2022.

Generalization of Neural Networks

Parameter

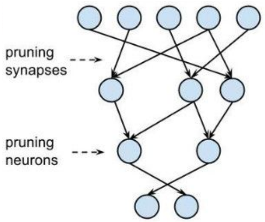
Data

Generalization of Neural Networks

Parameter

Data

Sparsity



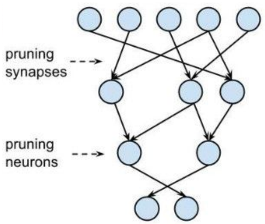
Make neural networks work at the edge
To improve generalization

Generalization of Neural Networks

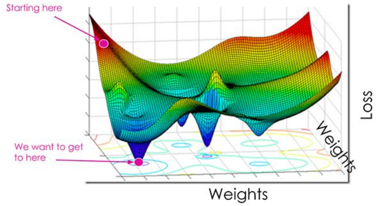
Parameter

Data

Sparsity



Loss Landscape



Make neural networks work at the edge
To improve generalization

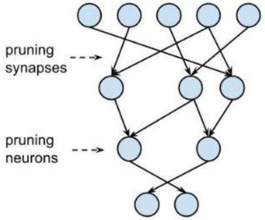
To understand/probe trained networks
ensembles to make neural networks work at the edge

Generalization of Neural Networks

Parameter

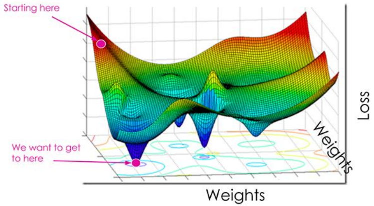
Data

Sparsity



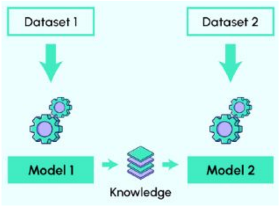
Make neural networks work at the edge
To improve generalization

Loss Landscape



To understand/probe trained networks
ensembles to make neural networks work at the edge

Pre-training Transfer Learning



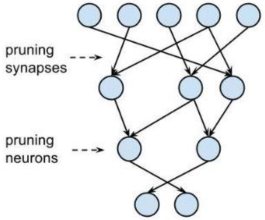
Dynamic environment → dynamic data
Reliable AI applications

Generalization of Neural Networks

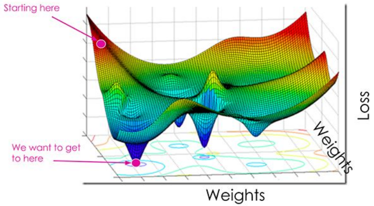
Parameter

Data

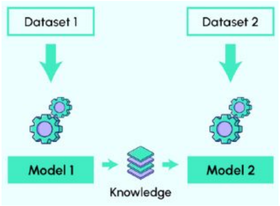
Sparsity



Loss Landscape



Pre-training Transfer Learning



Class-dependent pruning of deep neural networks

Understanding the effect of sparsity on neural networks robustness

Studying the impact of magnitude pruning on contrastive learning methods

Neural Network Pruning for Nuclei Instance Segmentation

The Role of Permutation Invariance in Linear Mode Connectivity of Neural Networks

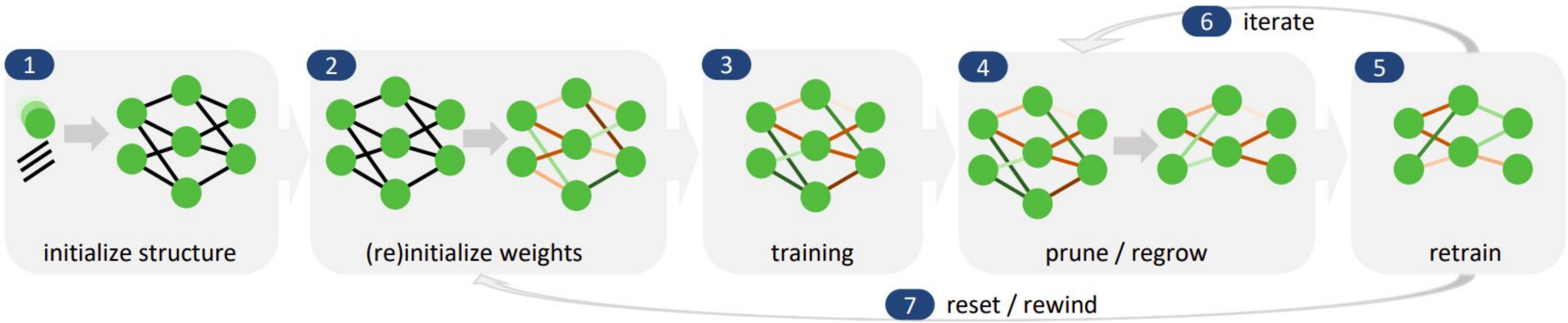
REPAIR: Linear Mode Connectivity of Deep Neural Networks via Permutation Invariance and Renormalization

The Role of Pretraining Data in Transfer Learning

Part 1:
Sparsity

Sparsity

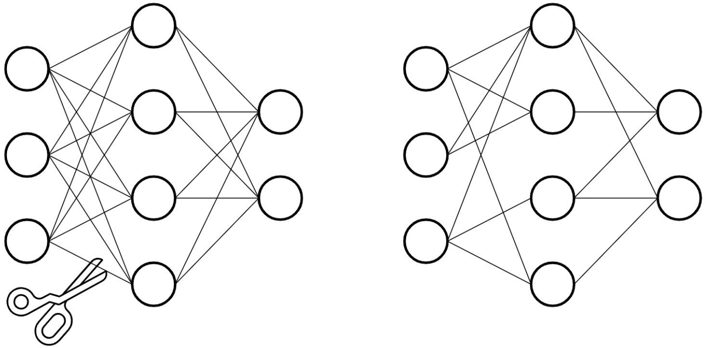
“With all things being equal, the simplest explanation tends to be the right one” (William of Ockham, ~1300)



Hoefler, Torsten, et al. "Sparsity in Deep Learning: Pruning and growth for efficient inference and training in neural networks." *arXiv preprint arXiv:2102.00554* (2021).

Relation between sparsity and generalization

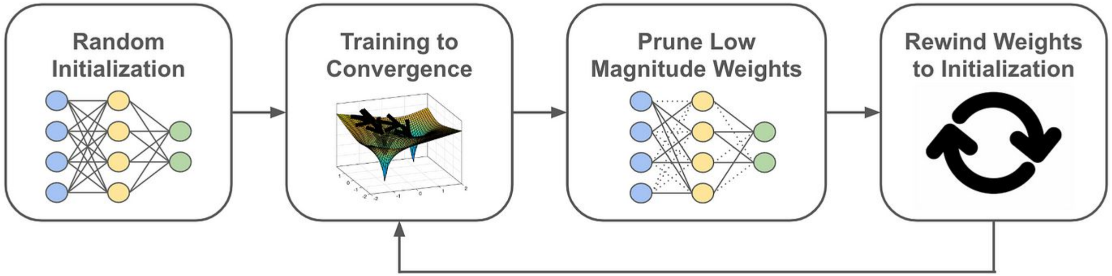
Does sparsity help/hurt generalization?



Before pruning

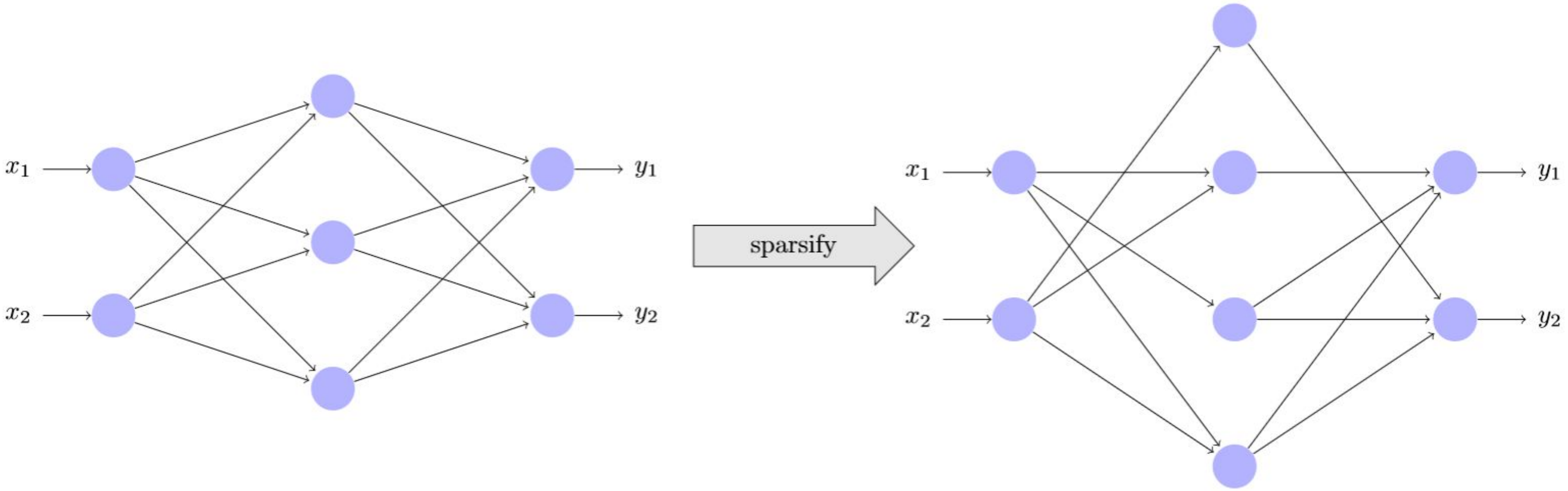
After pruning

Magnitude pruning



Lottery Ticket

Effective model capacity



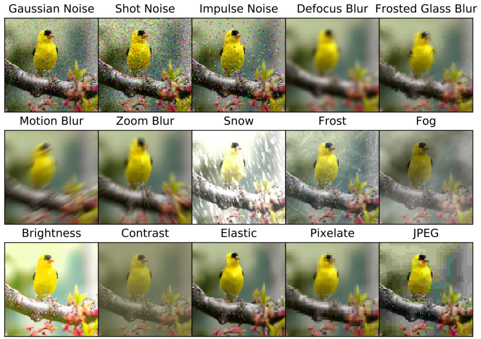
Sparsity and Generalization

1. **Data corruption**

2. **Weight perturbation**

Sparsity and Generalization

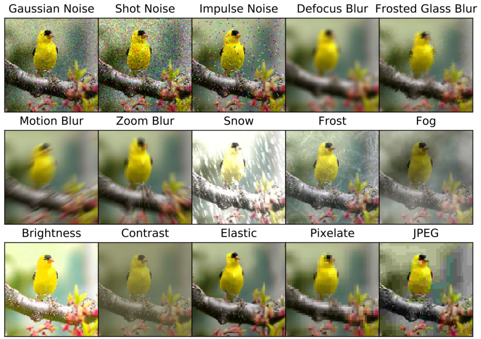
1. Data corruption



- Performance on corrupted datasets
  - MNIST-C
  - CIFAR10-C
  - CIFAR100-C

Sparsity and Generalization

1. Data corruption



- Performance on corrupted datasets
  - MNIST-C
  - CIFAR10-C
  - CIFAR100-C

2. Weight perturbation

- Add Gaussian noise to each weight
  - $z_i \sim N\,(\mu, \omega_i^2\sigma_i^2)$
- Flatness of achieved minima

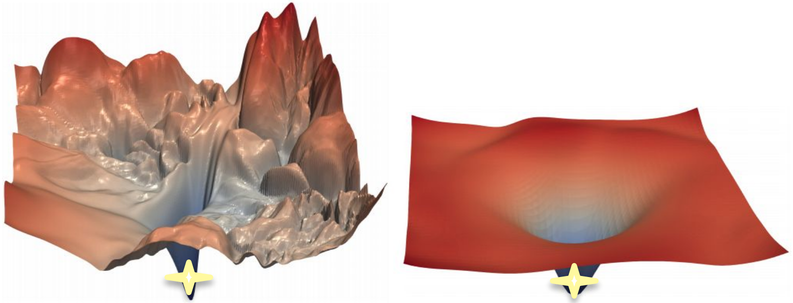
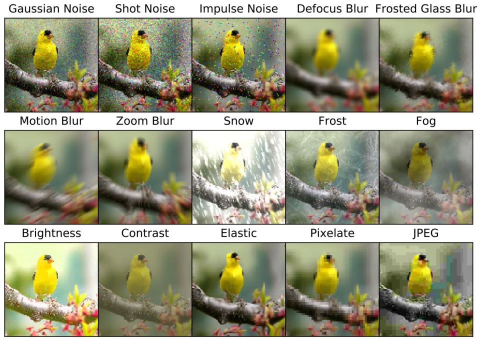


Image source [Li et al. 2018]

Sparsity and Generalization

1. Data corruption



- Performance on corrupted datasets
  - MNIST-C
  - CIFAR10-C
  - CIFAR100-C

2. Weight perturbation

- Add Gaussian noise to each weight
  - $z_i \sim N(\mu, \omega_i^2 \sigma_i^2)$
- Flatness of achieved minima

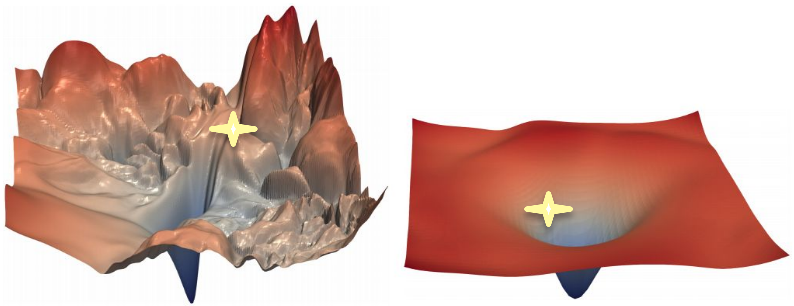
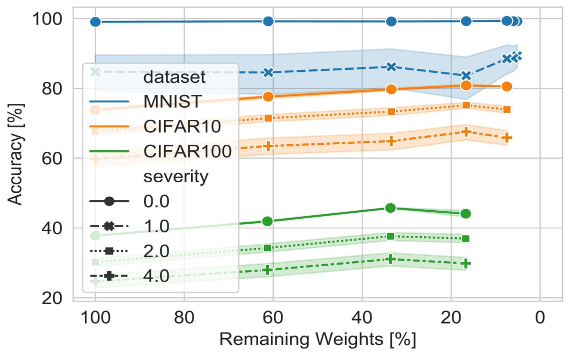


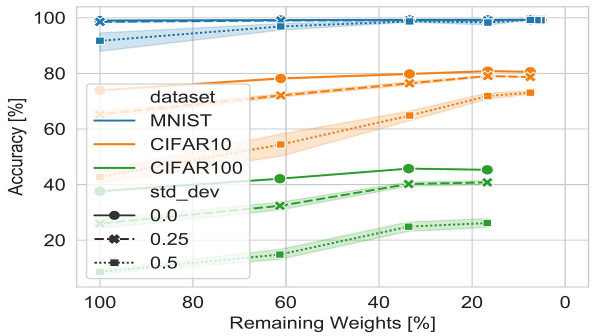
Image source [Li et al. 2018]

Sparsity and Generalization

1. Data corruption



2. Weight perturbation



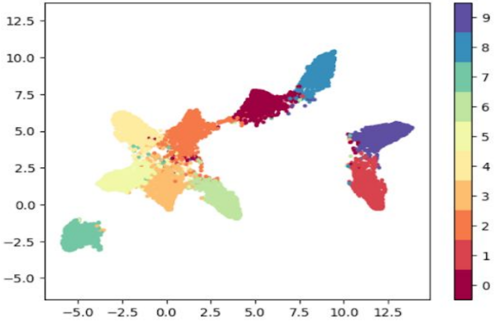
contrary to common belief, sparsity indeed does not hurt network generalization

Timpl, Lukas, et al. "Understanding the effect of sparsity on neural networks robustness." *arXiv preprint arXiv:2206.10915* (2022).

What is the effect of sparsity on learned representations?

Learned representations: UMAP

Supervised

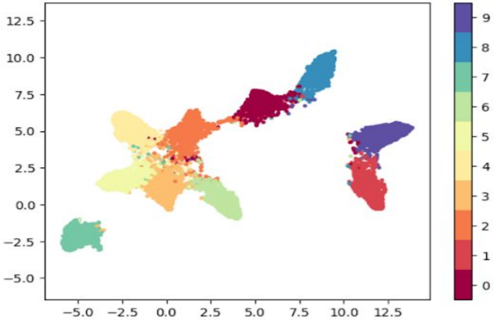


Dense

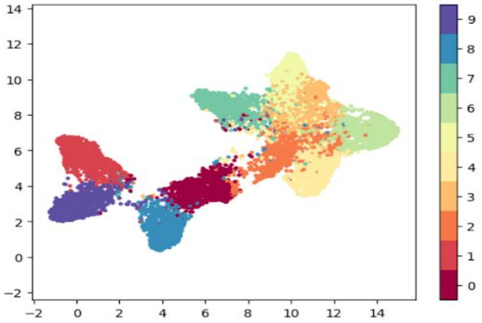
Corti, Francesco, et al. "Studying the impact of magnitude pruning on contrastive learning methods." *arXiv preprint arXiv:2207.00200* (2022).

Learned representations: UMAP

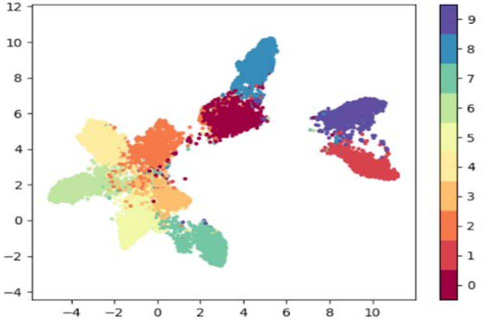
Supervised



Dense



GMP 90%

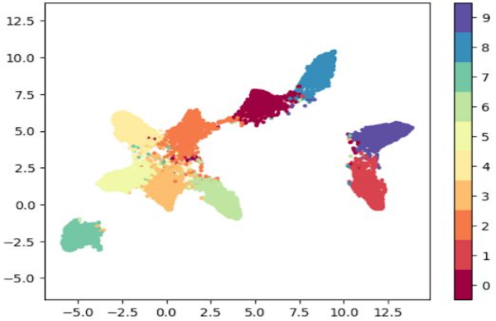


One-shot 90%

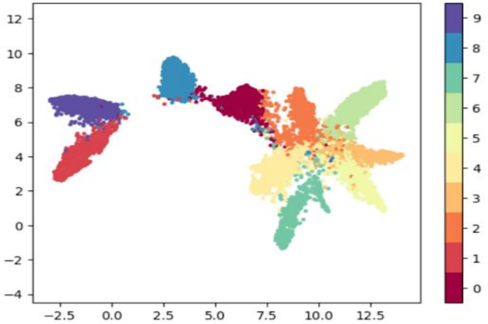
Corti, Francesco, et al. "Studying the impact of magnitude pruning on contrastive learning methods." *arXiv preprint arXiv:2207.00200* (2022).

UMAP: what if we change the training algorithm?

Supervised



Supervised Contrastive

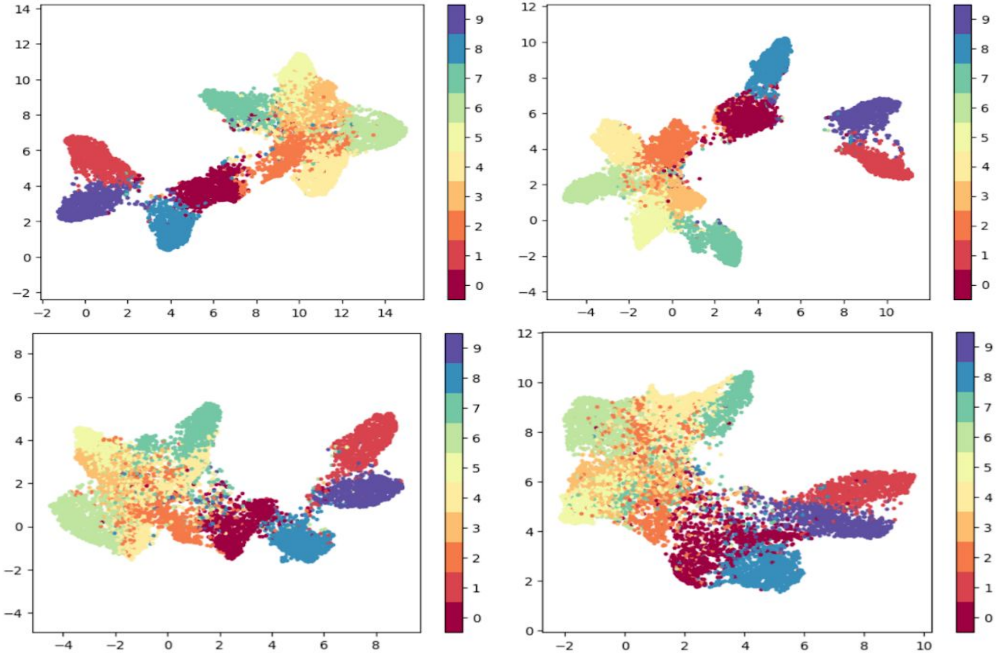


Dense

UMAP: supervised vs. semi-supervised

Supervised

Supervised Contrastive

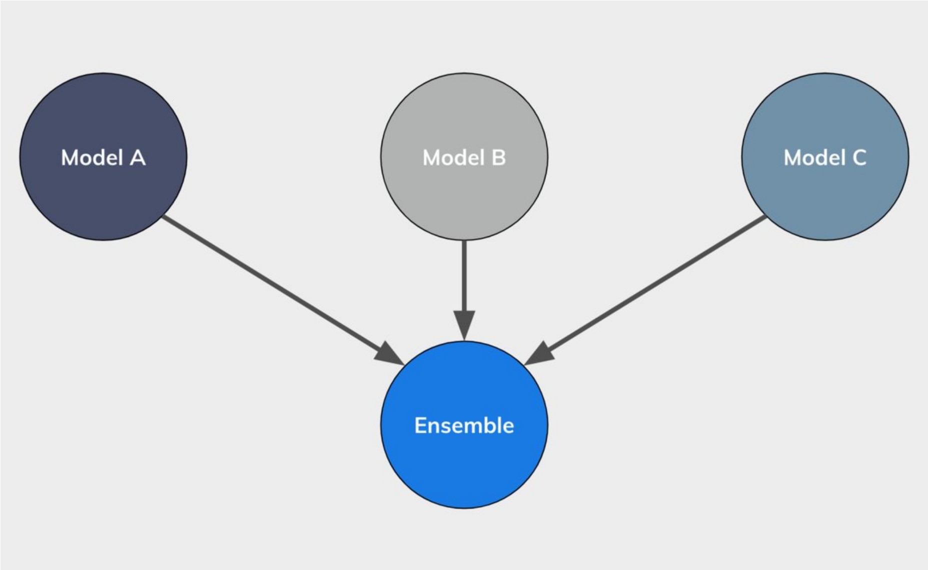


GMP 90% **One-shot 90%**

Part 2:
Loss Landscape

Motivation

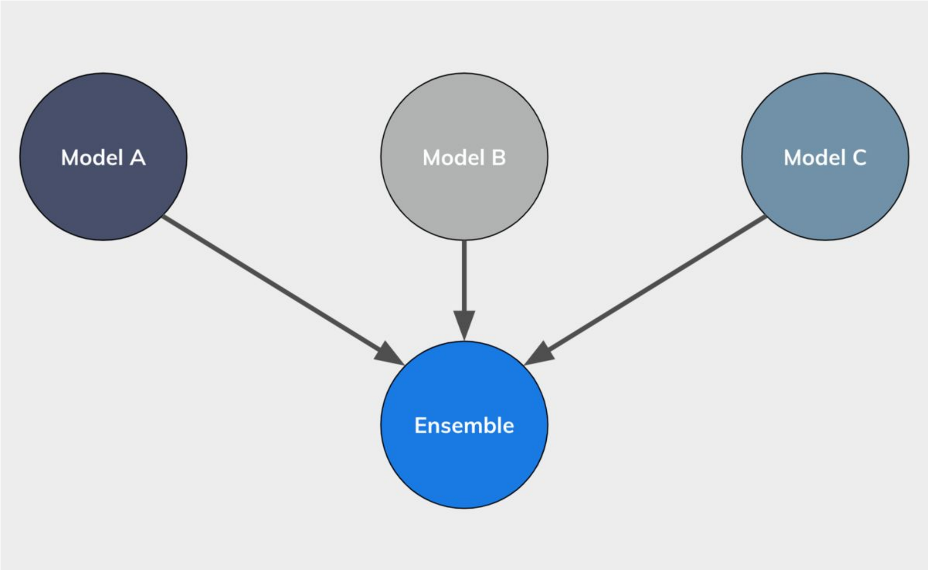
Ensembling helps generalization



Motivation

Form an ensemble model

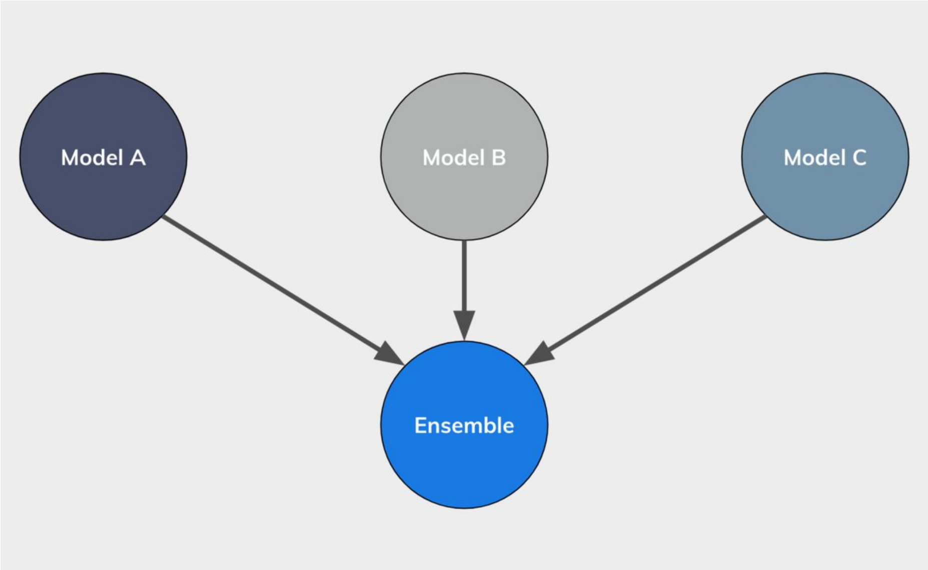
1. In output space



Motivation

Form an ensemble model

1. In output space
2. **In weight space (Embedded ML)**



Weight Averaging

Ensemble by weight averaging

Requirements:

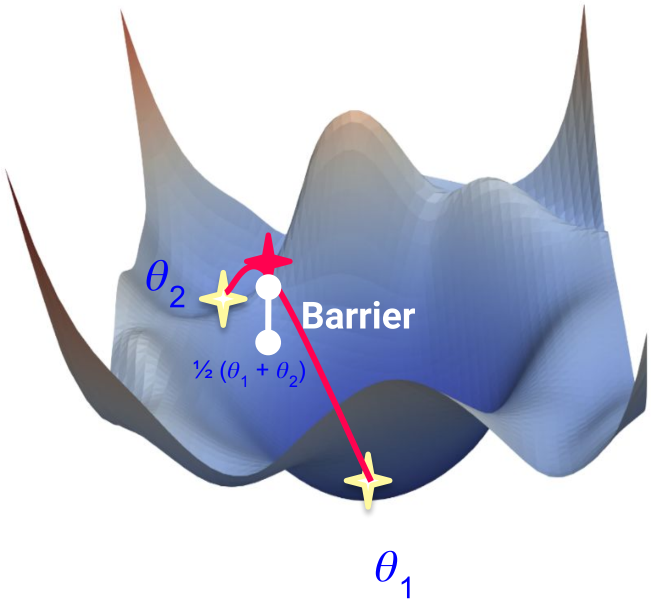
1. Solutions should be functionally diverse

Weight Averaging

Ensemble by weight averaging

Requirements:

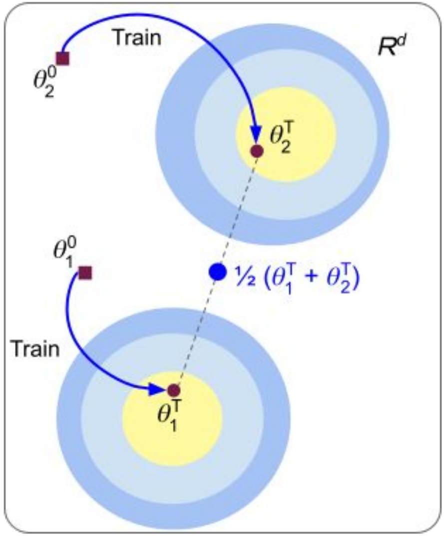
1. Solutions should be functionally diverse
2. Solutions should reside in one basin



Related works

Functionally different solutions

Weight space averaging

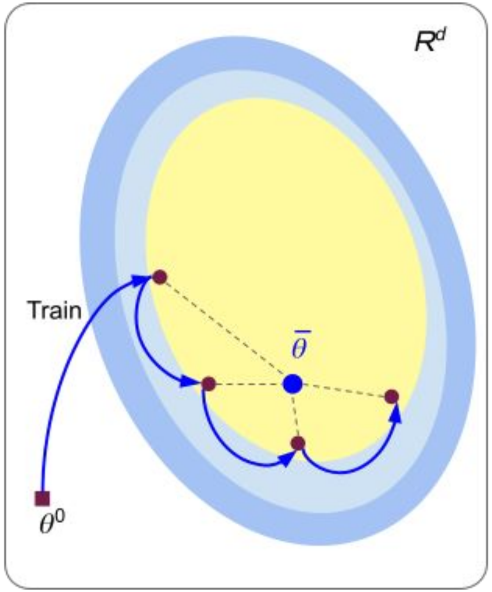


Frankle, Jonathan, et al. "Linear mode connectivity and the lottery ticket hypothesis." *International Conference on Machine Learning*. PMLR, 2020.

Related works

Functionally different solutions

Weight space averaging



Izmailov, Pavel, et al. "Averaging weights leads to wider optima and better generalization." *arXiv preprint arXiv:1803.05407* (2018).

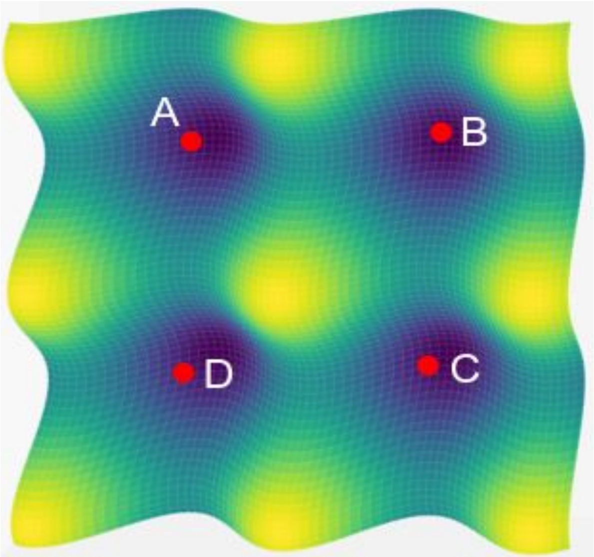
Is there any way to make different solutions in one basin?

Functionally different solutions

Weight space averaging

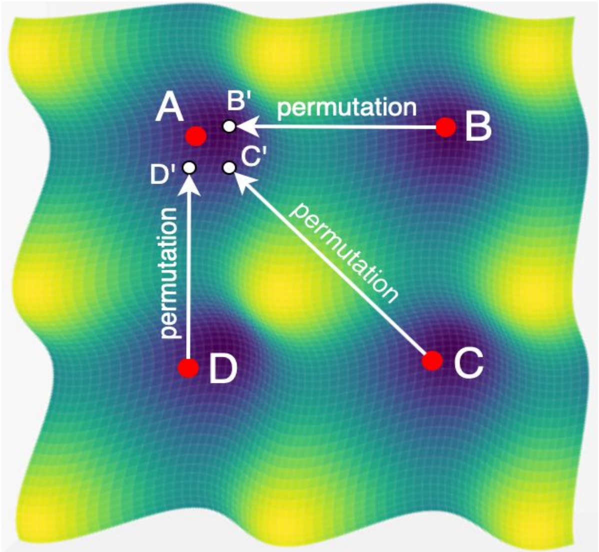
Conjecture

A, B, C, and D are minimas in different basins with barriers between pairs.



Conjecture

Taking permutations into account, there is likely no barrier in the linear interpolation between SGD solutions.



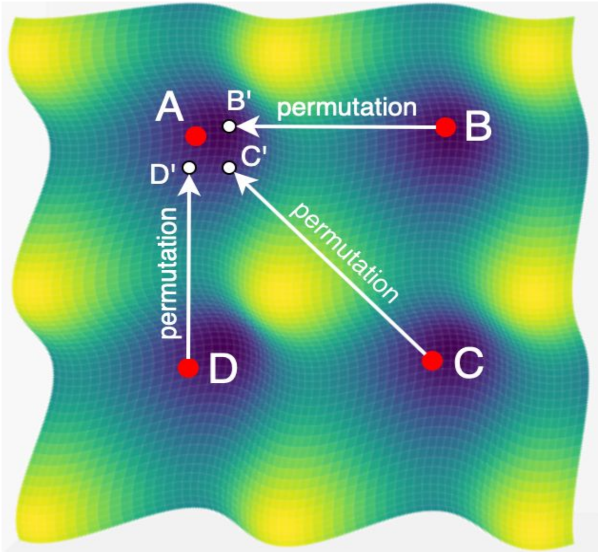
Entezari, Rahim, et al. "The role of permutation invariance in linear mode connectivity of neural networks." *arXiv preprint arXiv:2110.06296* (2021).

Conjecture

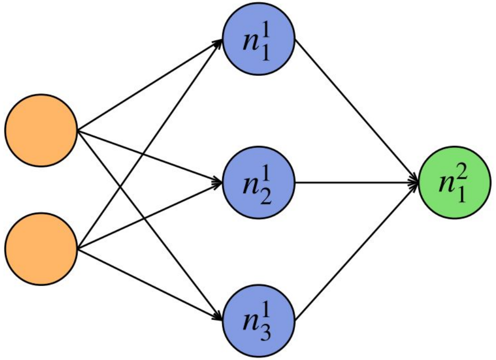
Taking <u>permutations</u> into account, there is likely no <u>barrier</u> in the linear interpolation between SGD solutions.

Functionally different solutions ✓

Weight space averaging ✓

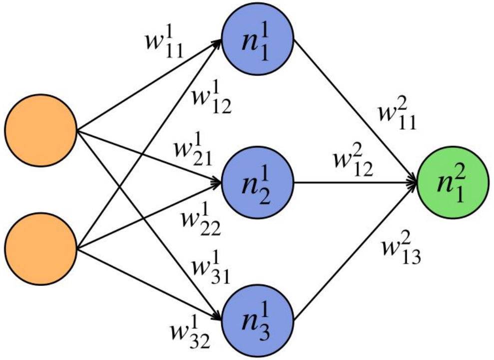


Permutations in Neural Networks

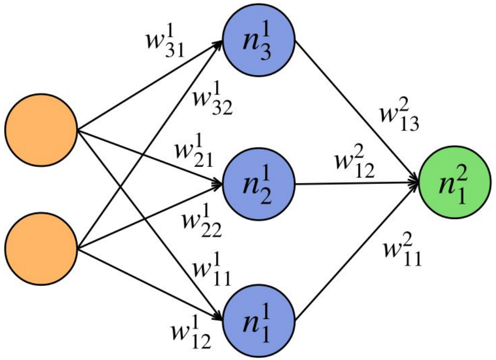


3! = 6 permutations

Permutations does not change the function!



(a) Neural network f_1



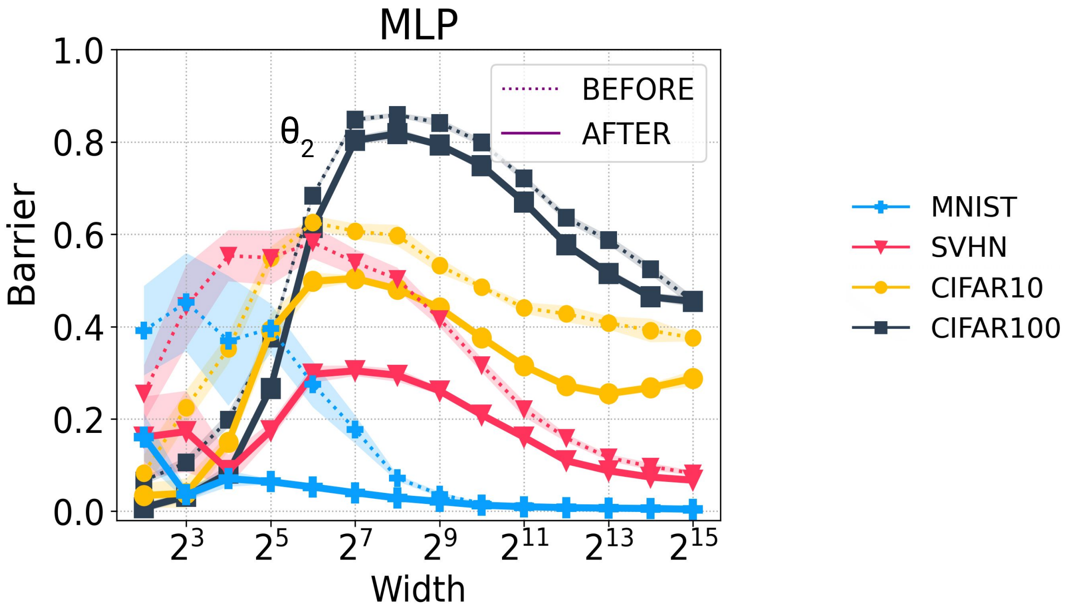
(b) Neural network f_2

How to find the appropriate permutation?

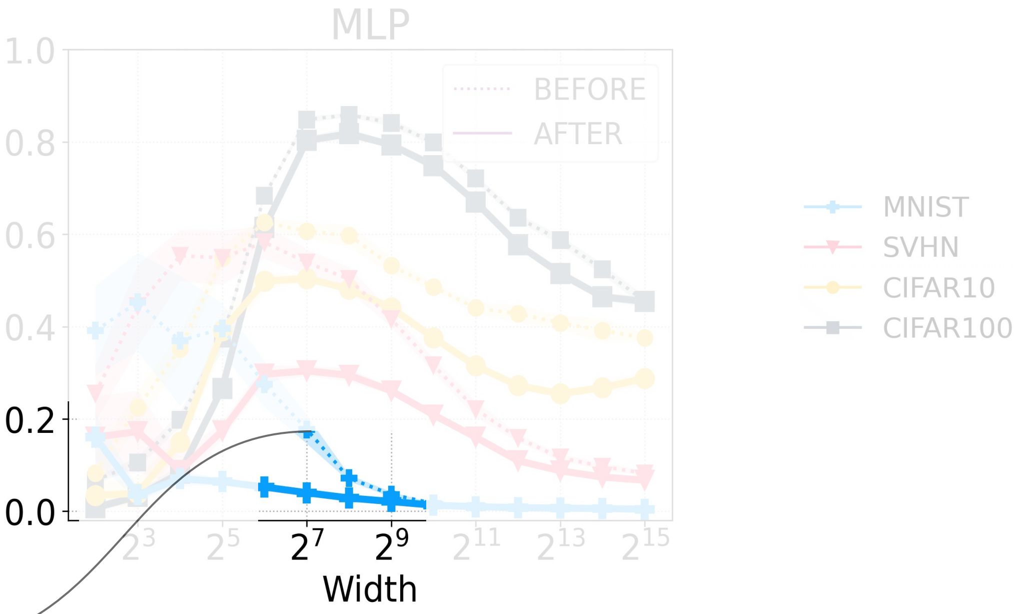
Permutation by brute force

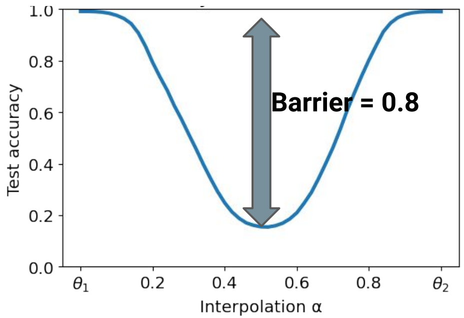
- ResNet-50 → 10^{55109}
- For comparison, the number of atoms in universe is about 10^{82}

Permutation by Simulated Annealing

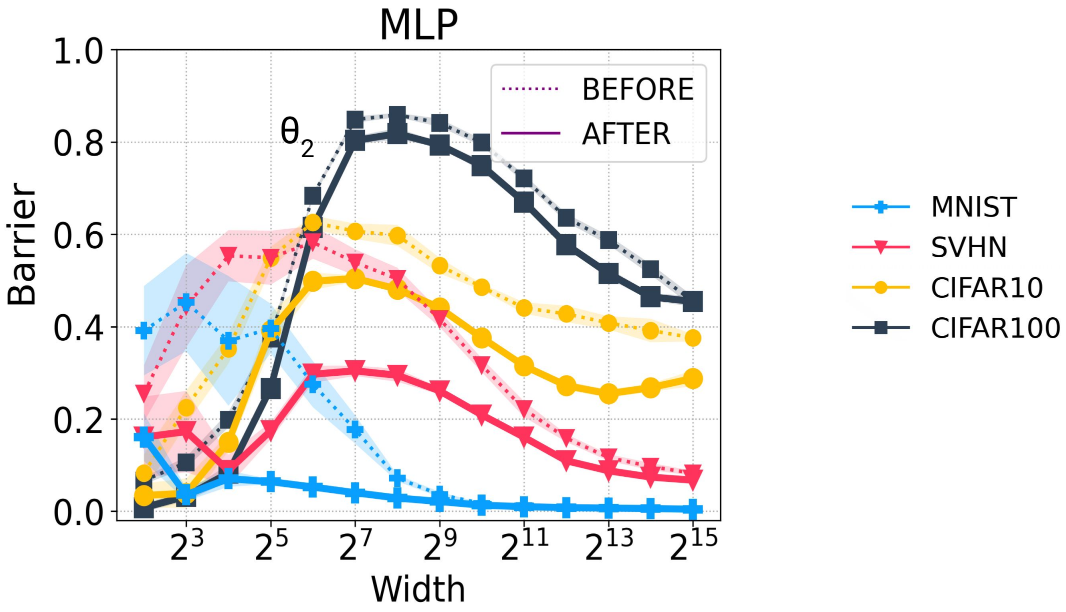


Permutation by Simulated Annealing





Permutation by Simulated Annealing

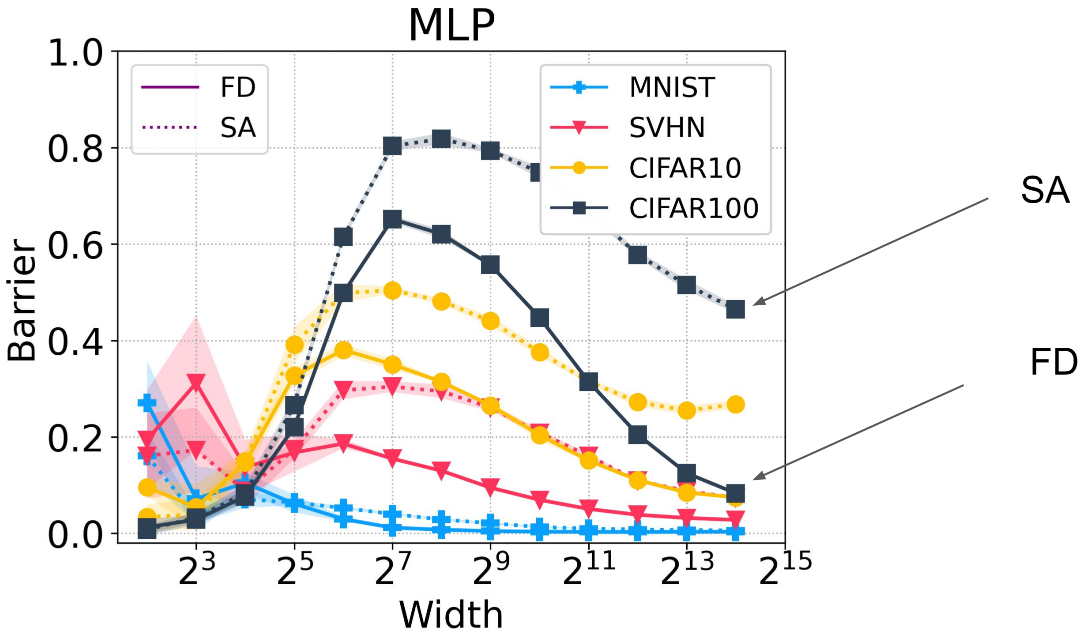


Neuron Alignment: **F**unctional **D**ifference

$$\delta E_l^{opt} = \frac{1}{2}(\tilde{\mathbf{w}}_{l,i}^A - \tilde{\mathbf{w}}_{l,j}^B)^\top \cdot \left((\tilde{\mathbf{H}}_{l,i}^A)^{-1} + (\tilde{\mathbf{H}}_{l,j}^B)^{-1}\right)^{-1} \cdot (\tilde{\mathbf{w}}_{l,i}^A - \tilde{\mathbf{w}}_{l,j}^B)$$

He, Xiaoxi, Zimu Zhou, and Lothar Thiele. "Multi-task zipping via layer-wise neuron sharing." *arXiv preprint arXiv:1805.09791* (2018).

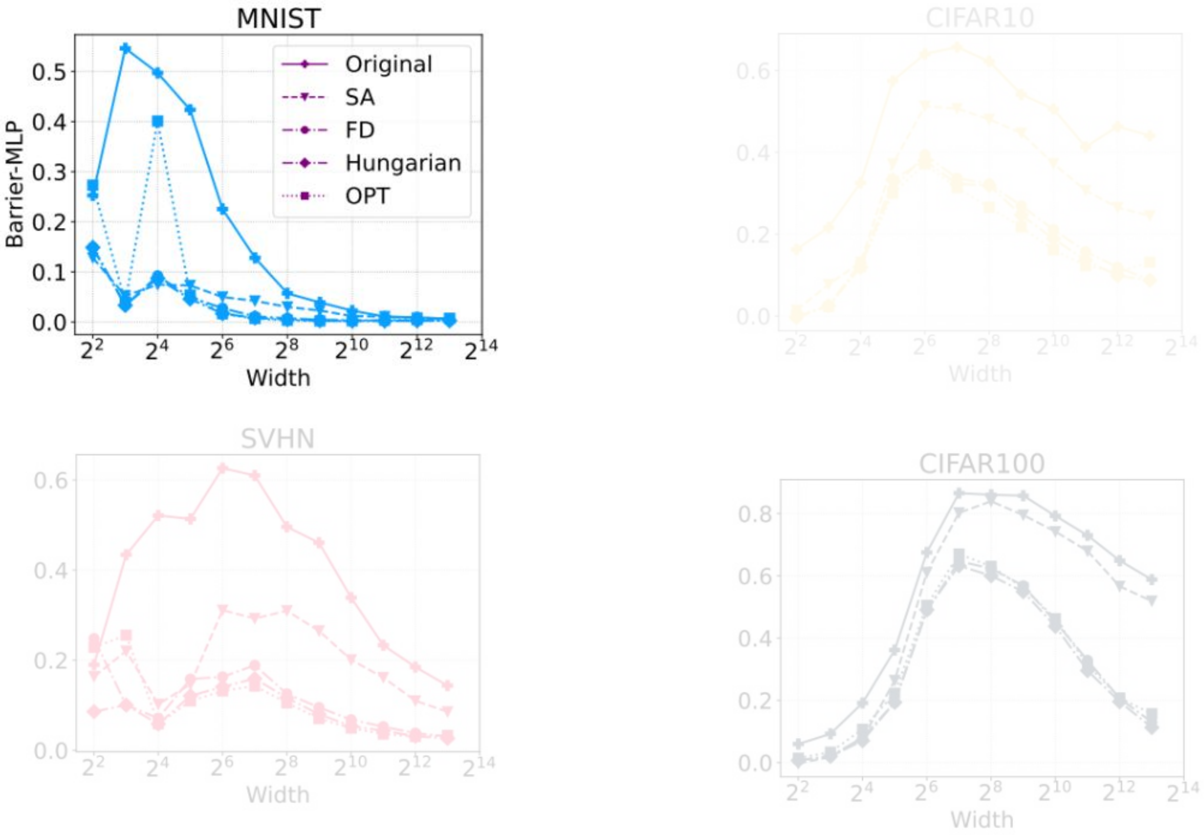
Neuron Alignment: **F**unctional **D**ifference



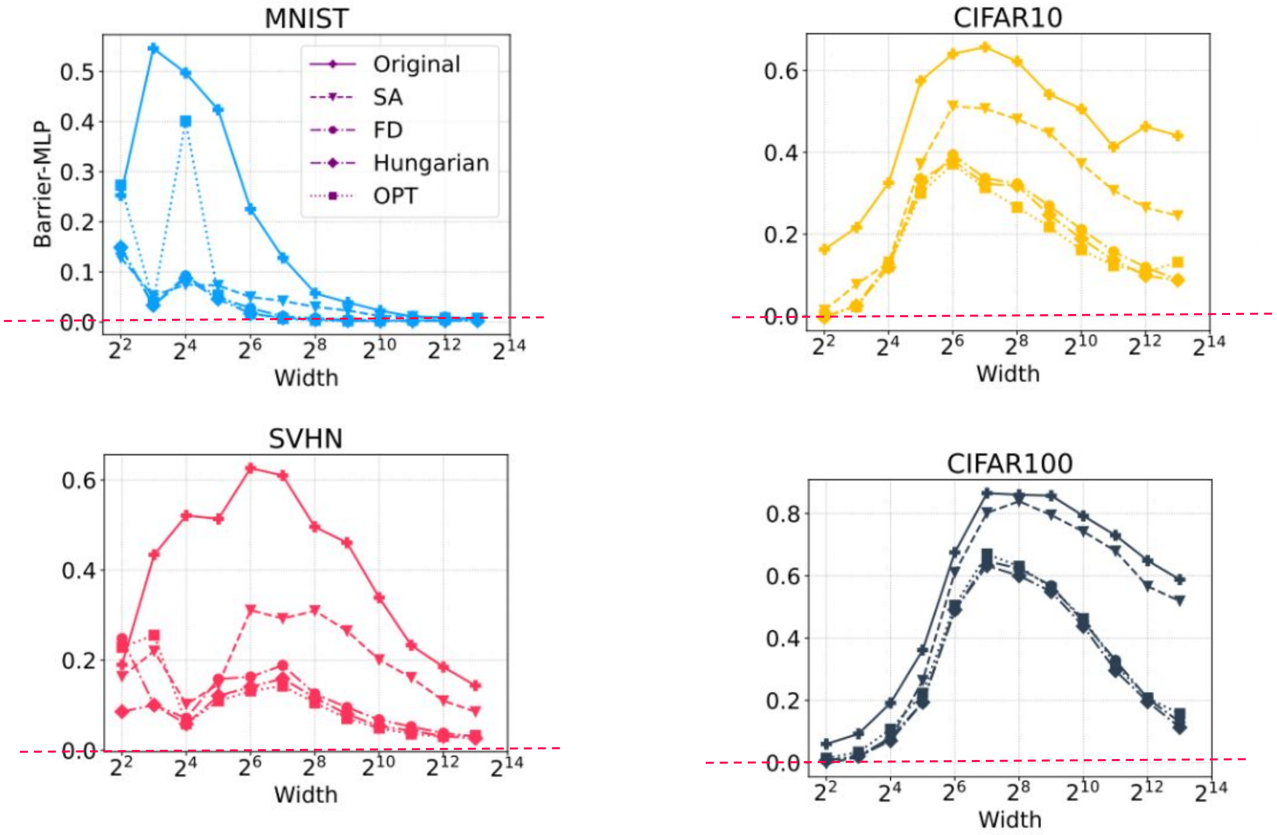
$$\delta E_l^{opt} = \frac{1}{2}(\tilde{\mathbf{w}}_{l,i}^A - \tilde{\mathbf{w}}_{l,j}^B)^\top \cdot \left((\tilde{\mathbf{H}}_{l,i}^A)^{-1} + (\tilde{\mathbf{H}}_{l,j}^B)^{-1}\right)^{-1} \cdot (\tilde{\mathbf{w}}_{l,i}^A - \tilde{\mathbf{w}}_{l,j}^B)$$

He, Xiaoxi, Zimu Zhou, and Lothar Thiele. "Multi-task zipping via layer-wise neuron sharing." *arXiv preprint arXiv:1805.09791* (2018).

Neuron Alignment methods: a comparison

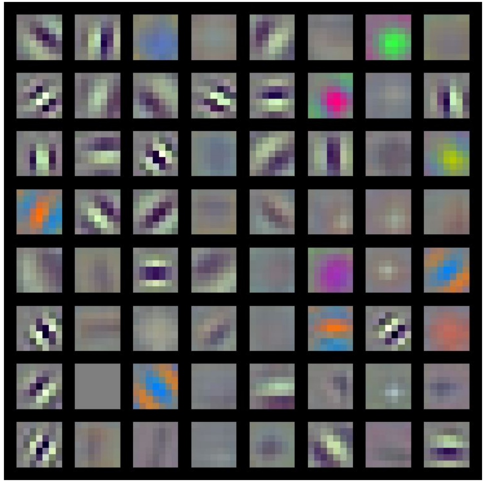


Neuron Alignment methods: a comparison

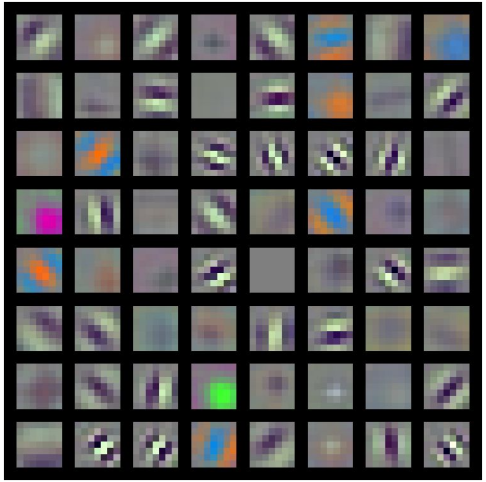


Neuron Alignment: **Correlation Matching**

Network A

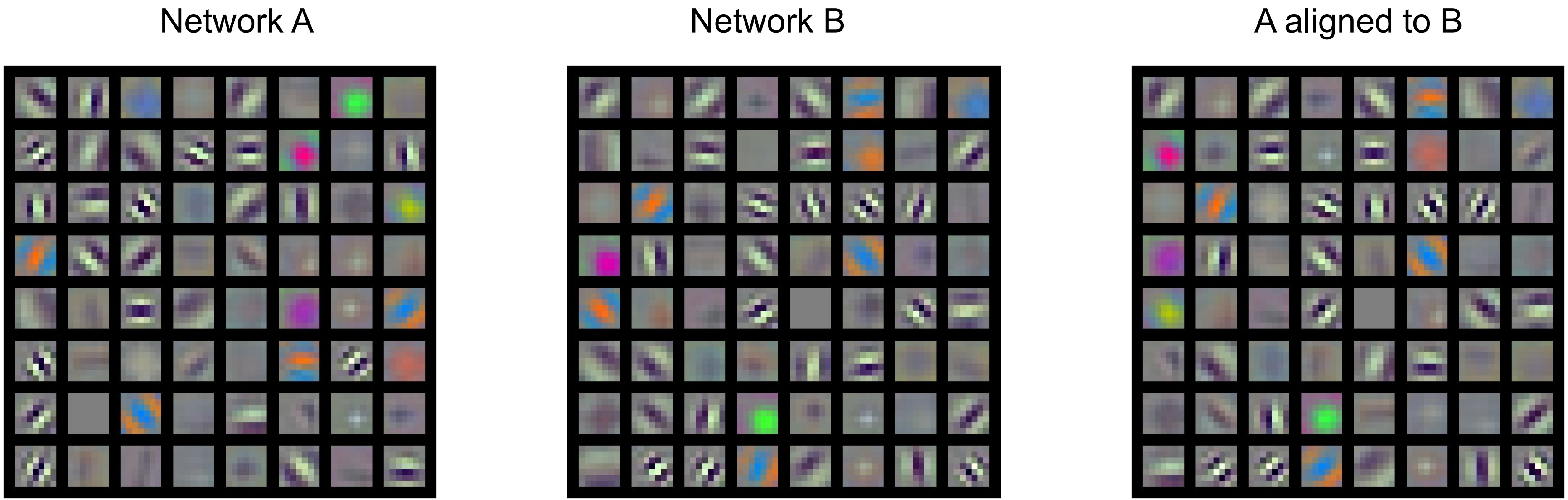


Network B



- Resnet-50
- ImageNet
- First layer: 64 filters

Neuron Alignment: **Correlation Matching**

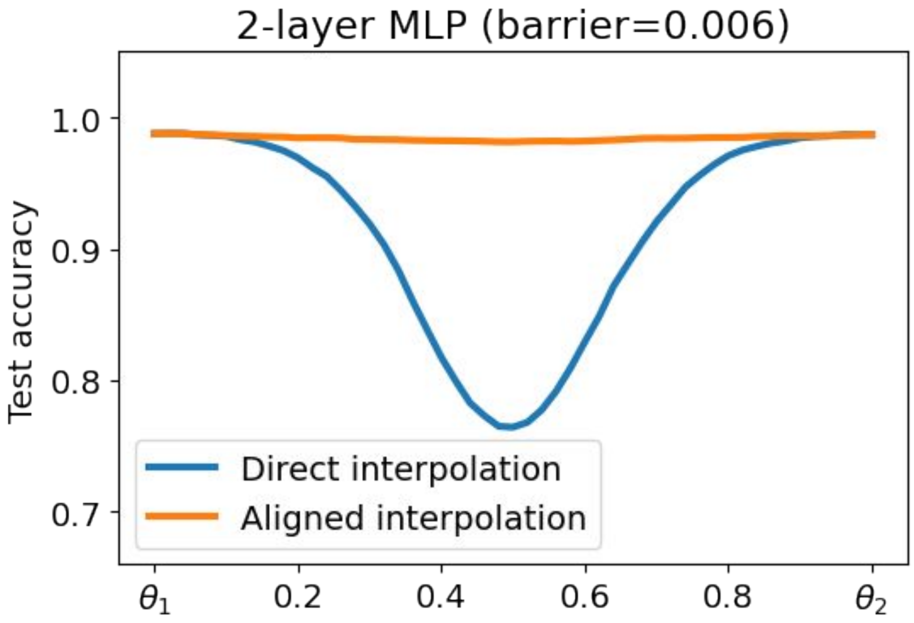


$$\sum_i \mathrm{corr}\left(X_{l,i}^{(1)}, X_{l,P_l(i)}^{(2)}\right)$$

Li, Yixuan, et al. "Convergent learning: Do different neural networks learn the same representations?." *arXiv preprint arXiv:1511.07543* (2015).

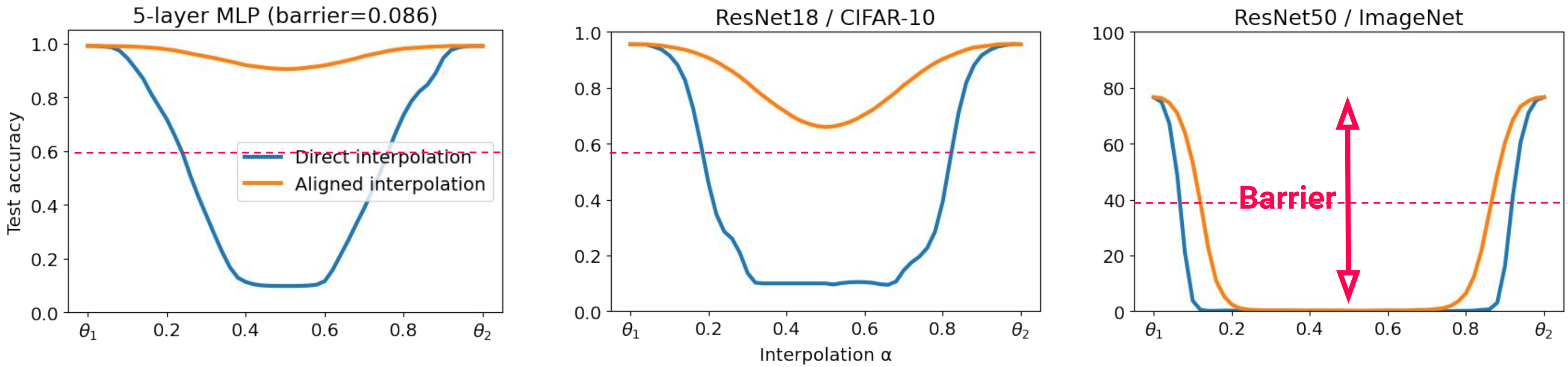
Neuron Alignment: **Correlation Matching**

Works for shallow+wide MLPs



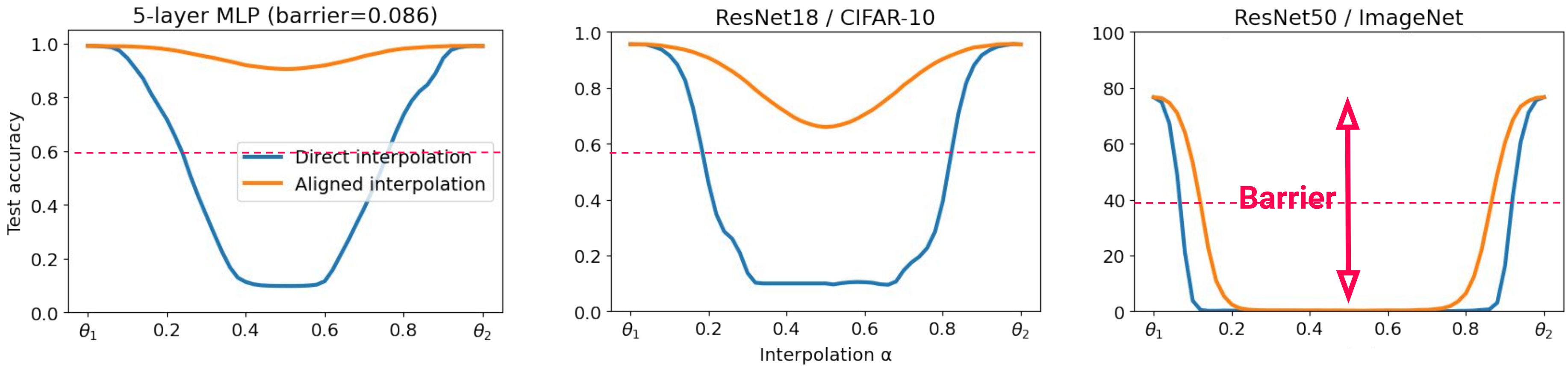
Jordan, Keller, et al. "Repair: Renormalizing permuted activations for interpolation repair." *arXiv preprint arXiv:2211.08403* (2022).

Correlation Matching breaks for deeper networks



Jordan, Keller, et al. "Repair: Renormalizing permuted activations for interpolation repair." *arXiv preprint arXiv:2211.08403* (2022).

Correlation Matching breaks for deeper networks

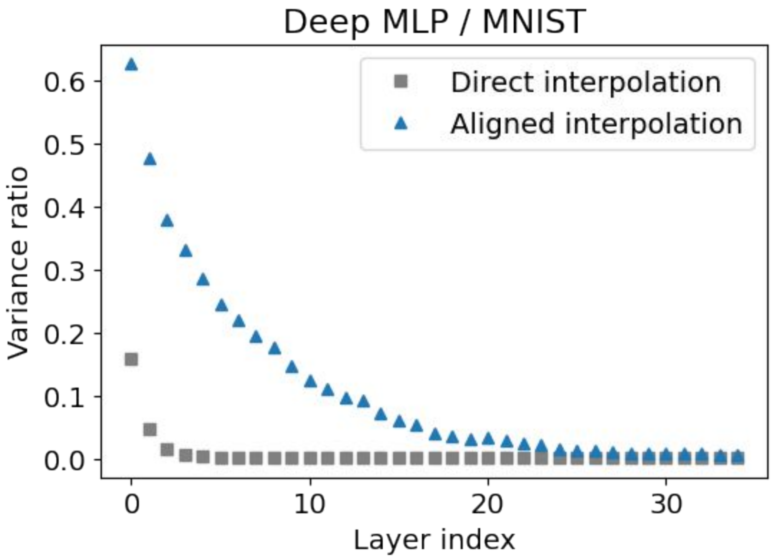


But why?

Jordan, Keller, et al. "Repair: Renormalizing permuted activations for interpolation repair." *arXiv preprint arXiv:2211.08403* (2022).

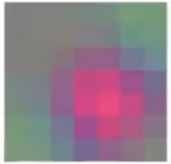
Variance collapse

$$\frac{\sigma_\alpha^2}{\frac{\sigma_0^2+\sigma_1^2}{2}}$$



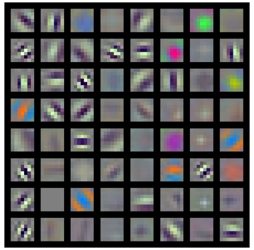
Variance collapse

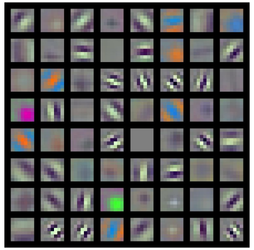
x_1



x_2

Filter 9

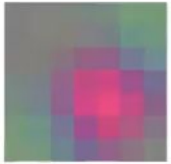




Variance collapse

$\boldsymbol{x}_1$ $\boldsymbol{x}_2$ $\boldsymbol{x}_\alpha$

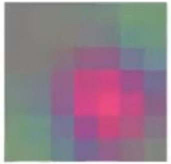
Filter 9



Variance collapse

Filter 9

$\boldsymbol{x}_1$



var=1.58

$\boldsymbol{x}_2$

var=2.94

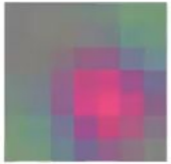
$\boldsymbol{x}_\alpha$

$$
\begin{aligned}
\mathrm{Var}(X_\alpha) &= \mathrm{Var}\left(\frac{X_1 + X_2}{2}\right) \\
&= \frac{\mathrm{Var}(X_1) + \mathrm{Var}(X_2) + 2\mathrm{Cov}(X_1, X_2)}{4} \\
&= \frac{\mathrm{std}^2(X_1) + \mathrm{std}^2(X_2) + 2 \cdot \mathrm{corr}(X_1, X_2) \cdot \mathrm{std}(X_1)\mathrm{std}(X_2)}{4} \\
&= \left(\frac{\mathrm{std}(X_1) + \mathrm{std}(X_2)}{2}\right)^2 - \frac{(1 - \mathrm{corr}(X_1, X_2))}{2}\mathrm{std}(X_1)\mathrm{std}(X_2)
\end{aligned}
$$

Variance collapse

Filter 9

$\boldsymbol{x}_1$



var=1.58

$\boldsymbol{x}_2$

var=2.94

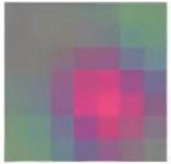
$\boldsymbol{x}_\alpha$

var=1.18 < 2.26

Variance collapse

Filter 9

$\boldsymbol{x}_1$



var=1.58

$\boldsymbol{x}_2$

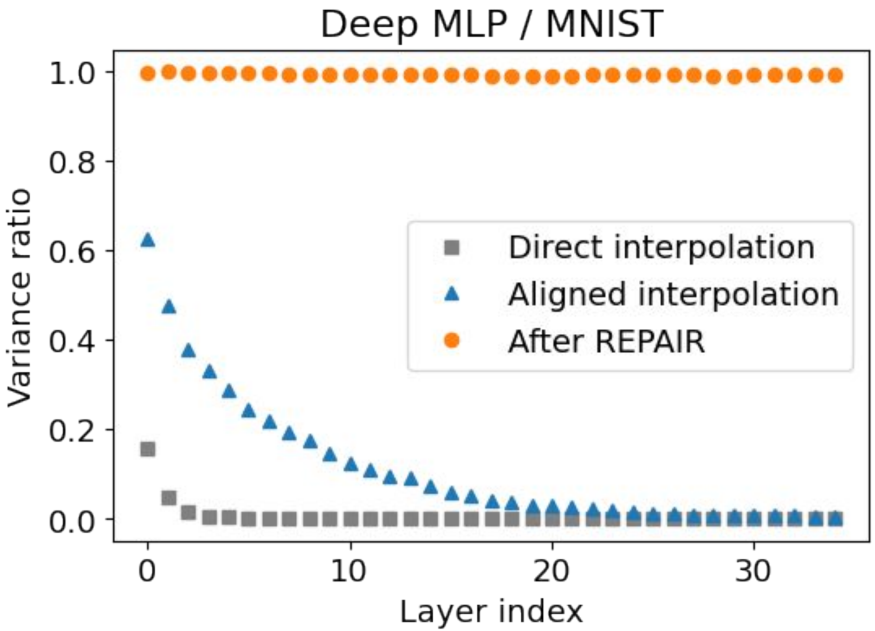
var=2.94

$\boldsymbol{x}_\alpha$

var=1.18 $<$ 2.26

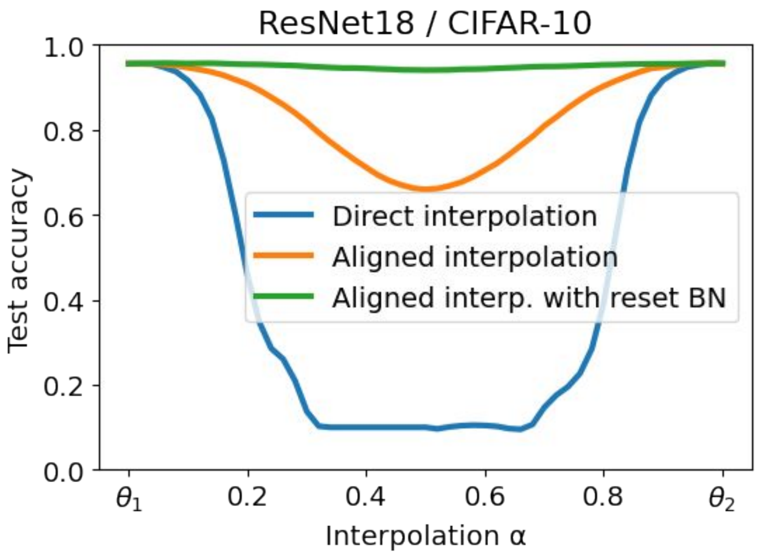
$$\mathrm{Var}\left(x_\alpha\right) = \mathrm{Var}\left(\frac{x_1 + x_2}{2}\right) = \frac{1}{4}(\mathrm{std}(x_1)^2 + \mathrm{std}(x_2)^2 + \mathrm{Corr}(x_1, x_2) \cdot \mathrm{std}(x_1) \cdot \mathrm{std}(x_2))$$

REPAIR: Re-estimate Batchnorm statistics



Jordan, Keller, et al. "Repair: Renormalizing permuted activations for interpolation repair." *arXiv preprint arXiv:2211.08403* (2022).

REPAIR: Re-estimate Batchnorm statistics



Jordan, Keller, et al. "Repair: Renormalizing permuted activations for interpolation repair." *arXiv preprint arXiv:2211.08403* (2022).

Part 3:
Pre-training Data

Research questions

➔ **R1: role of pre-training data**

- Given a target task, which dataset to pre-train?

➔ **R2: role of pre-training method**

- Given a target task, which pre-train method to choose?
- supervised ImageNet or contrastive LAION?

Experimental setup

1

Pre-training

CLIP

LAION, YFCC, WIT, Conceptual captions, Redcaps, Shutterstock

2

Finetuning

Few-shot: 1/5/10/20/all samples per class

CIFAR100, DTD, CALTECH101, PETS, REAL (domain net), CLIPART (domain net), CameraTraps, Cassava Leaf Disease, EuroSAT

Pre-training datasets

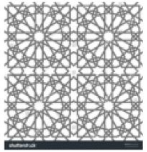
LAION

Yellow sandals for women pointy and low heeled Beatnik Françoise Mustard

3 Bedrooms Terraced House for sale in Eastbourne Road, Walton, Liverpool, Merseyside, L9

Minimum Wage Barbie

Conceptual captions



Islamic vector geometric ornaments based on traditional arabic art. Oriental seamless pattern. Muslim mosaic. Turkish, Arabian tile on a white background. Mosque ...

Illustration of hand holding the id card. Vector illustration flat design.

<PERSON>: U. <PERSON> in United States Army. First <PERSON> appointed to that position. First, &, so far, only <PERSON> to serve on Joint Chiefs of Staff. Black H...

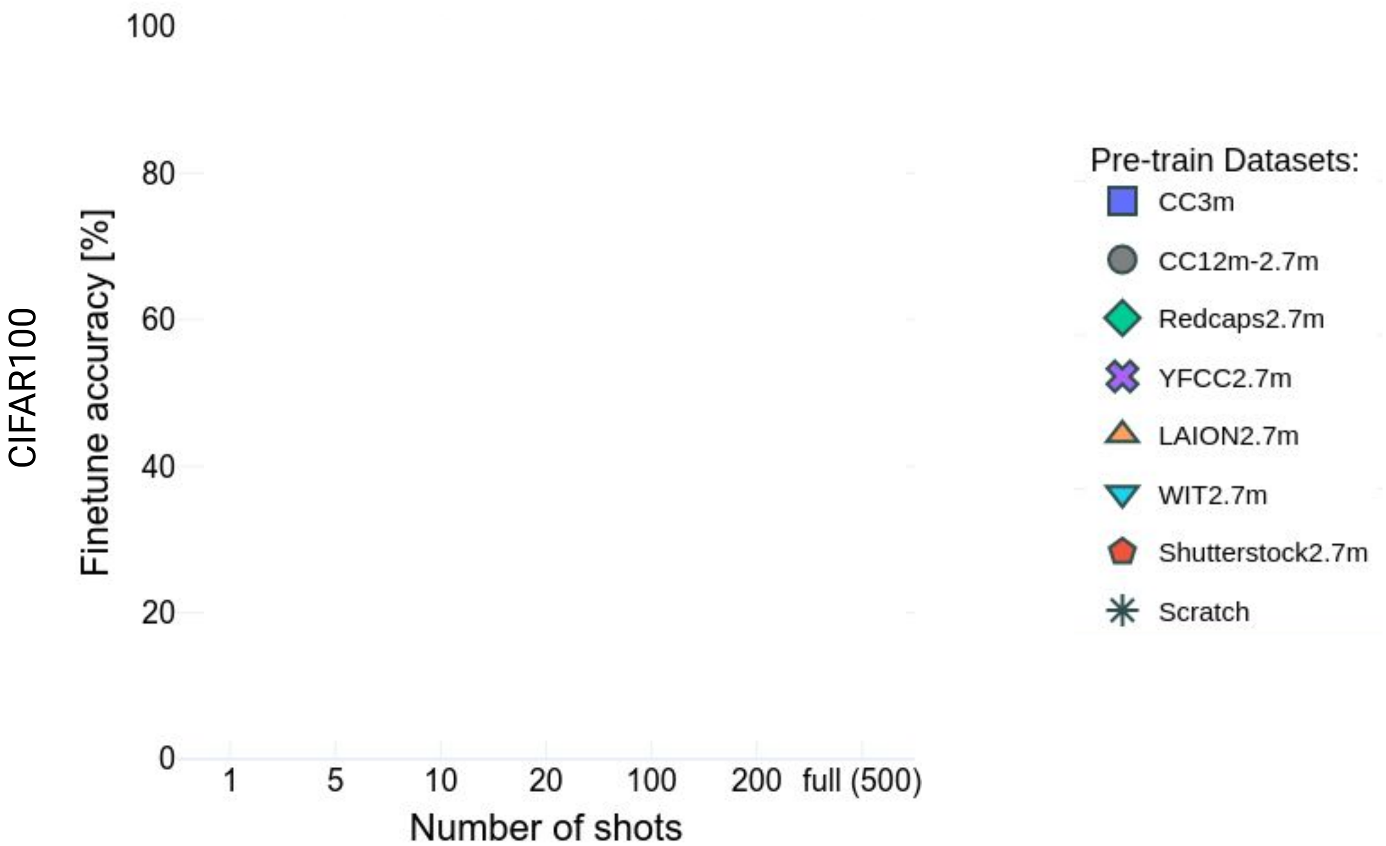
Finetuning datasets

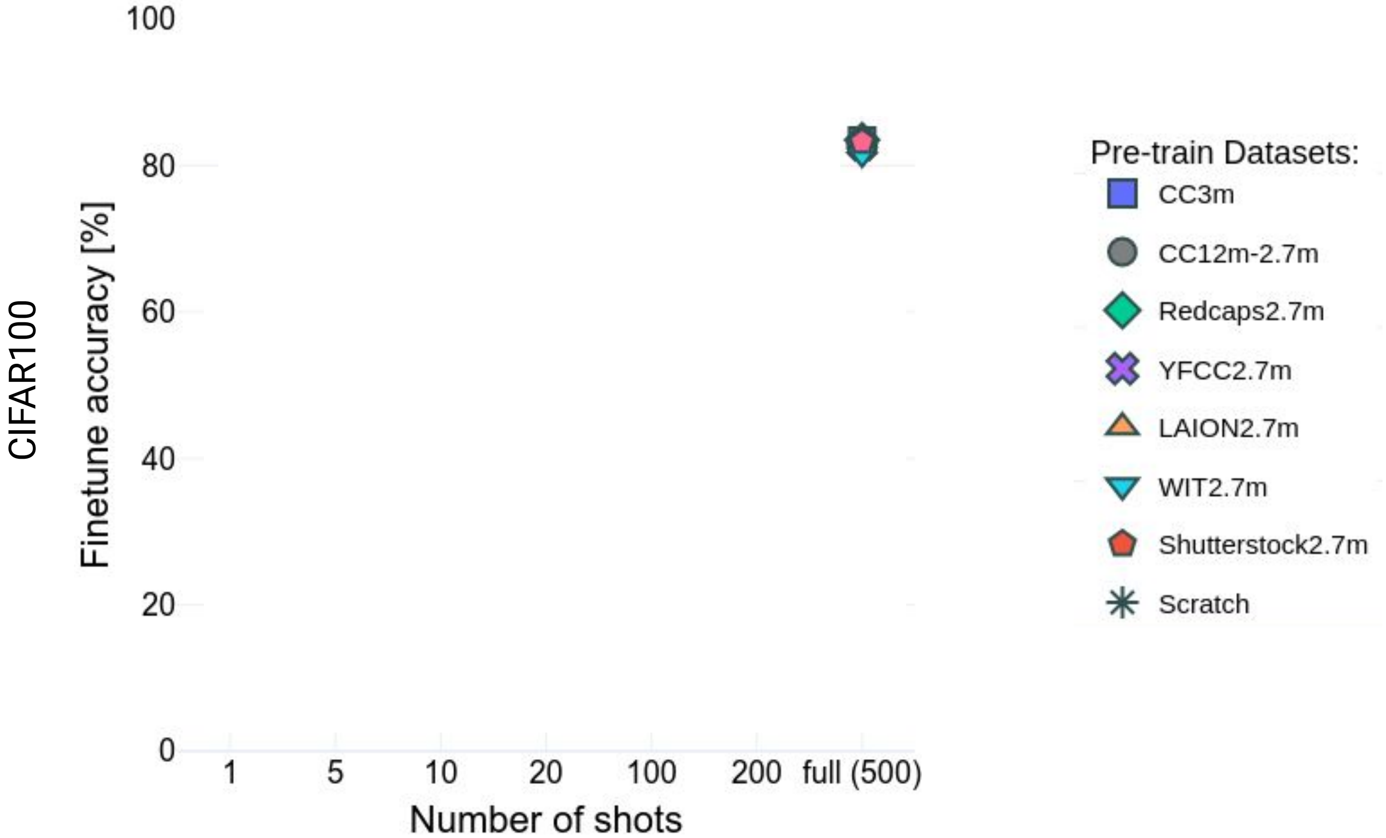
name	CIFAR100	DTD	REAL (domain net)	CLIPART (domain net)	Camera traps	Cassava leaf disease	EuroSAT
samples	50K	5.6K	172K	172K	58K	21K	27K
classes	100	47	345	345	15	5	10

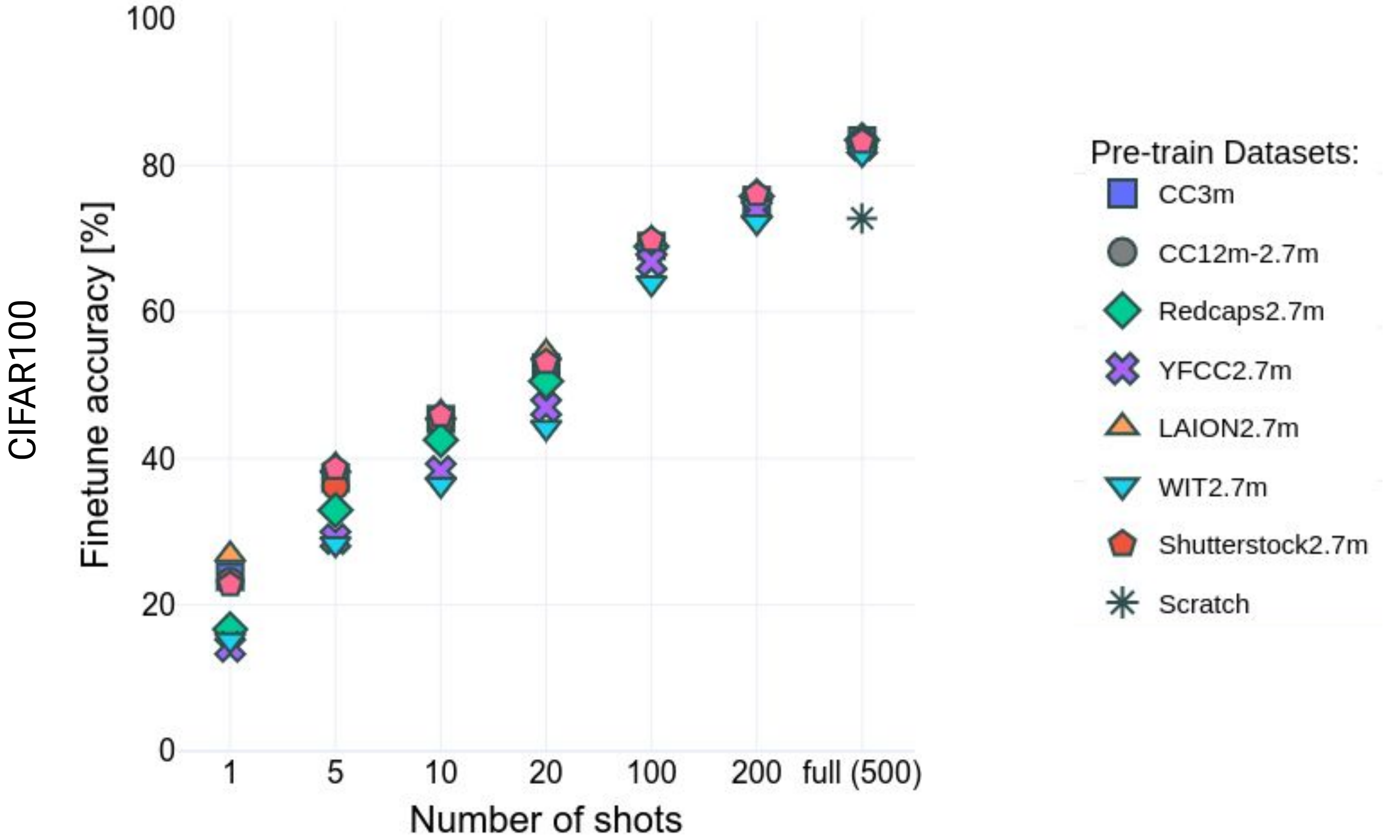


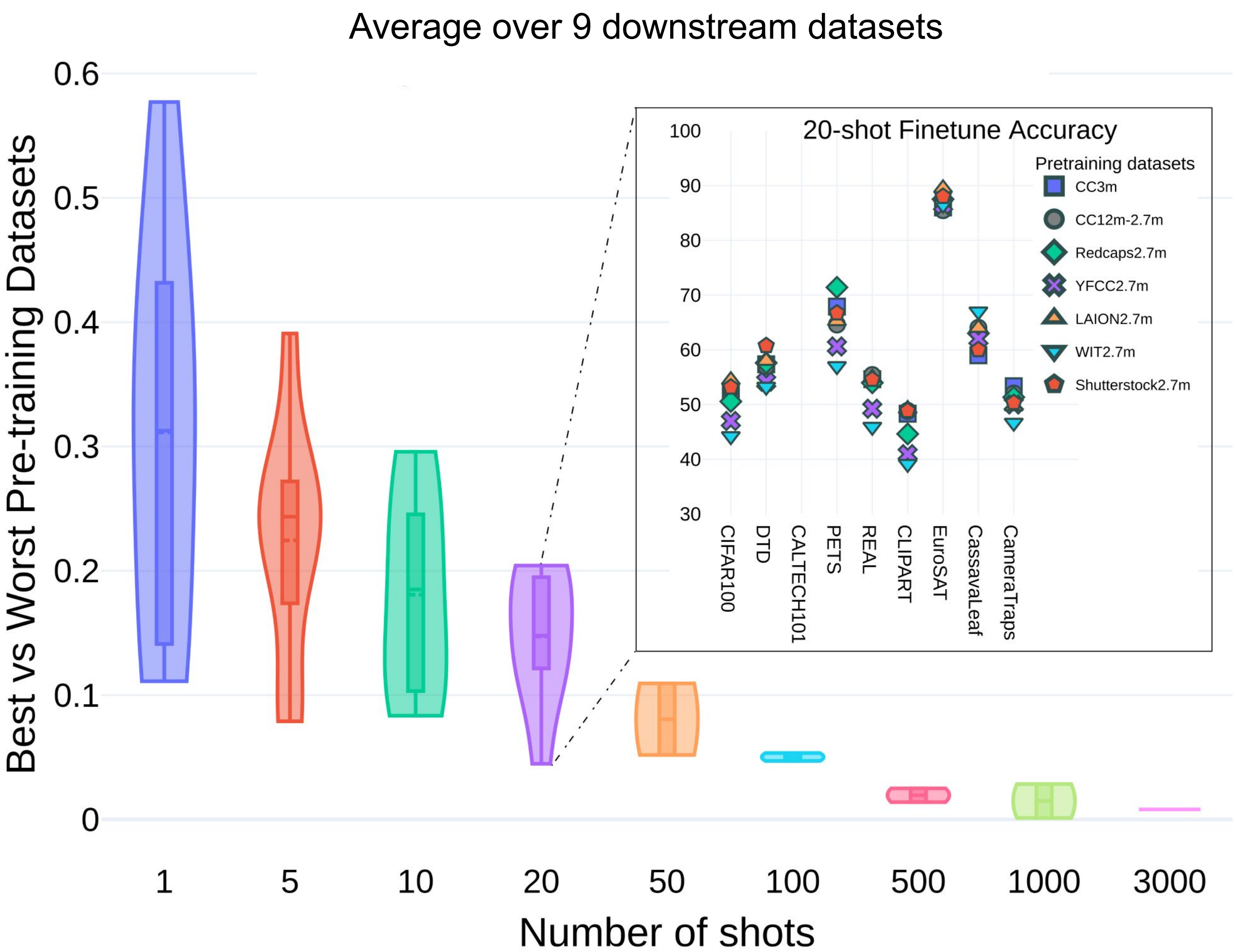


Which dataset to pre-train?

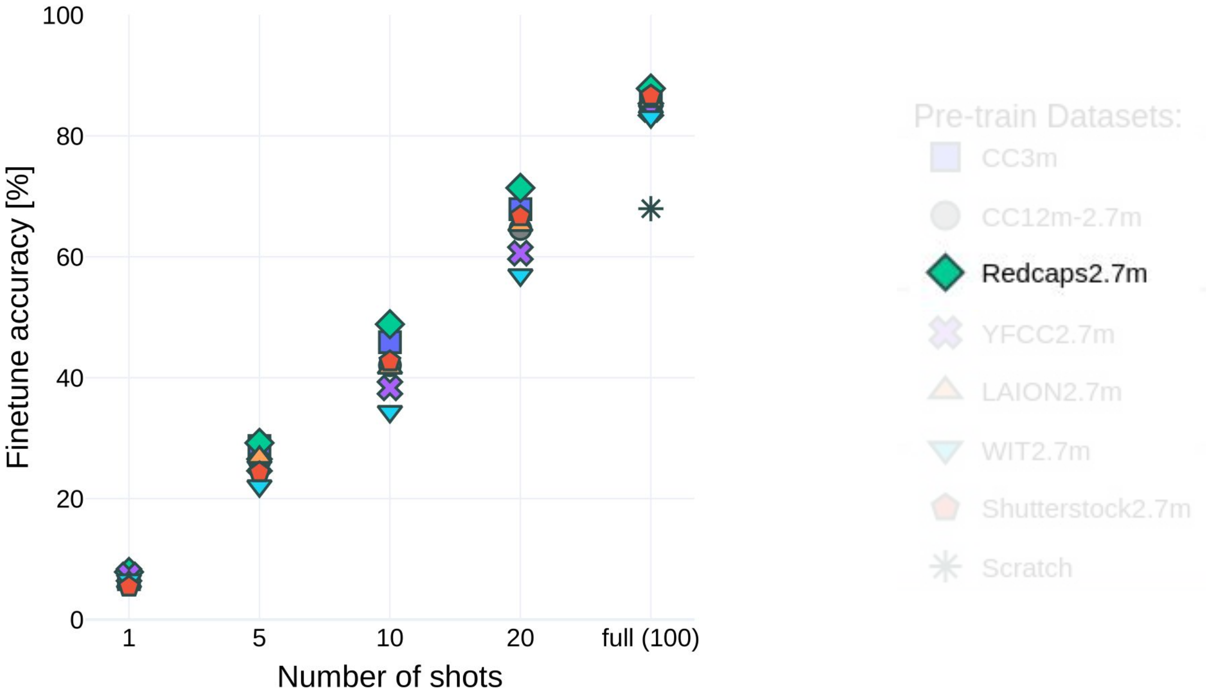



CIFAR100
Finetune accuracy [%]
100
80
60
40
20
0
1
5
10
20
100
200
full (500)
Number of shots
Pre-train Datasets:
CC3m
CC12m-2.7m
Redcaps2.7m
YFCC2.7m
LAION2.7m
WIT2.7m
Shutterstock2.7m
Scratch

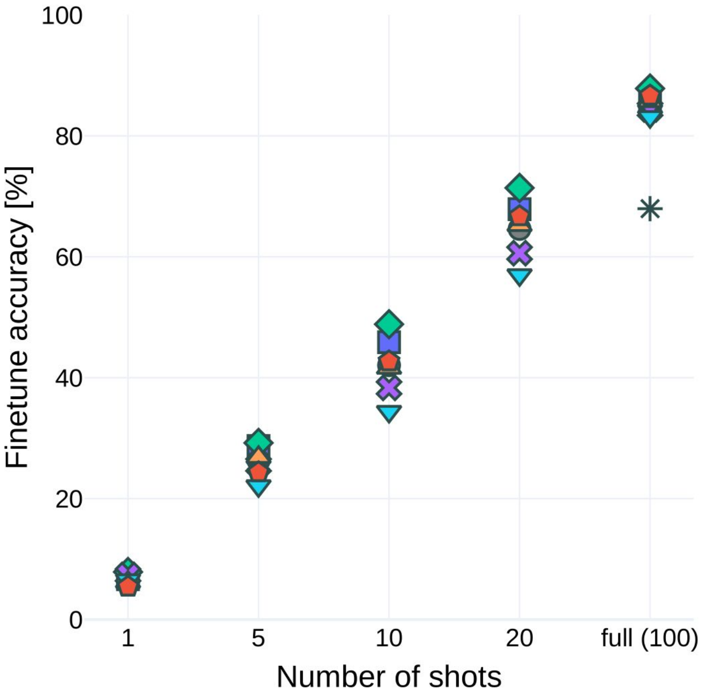

CIFAR100
Finetune accuracy [%]
100
80
60
40
20
0
1
5
10
20
100
200
full (500)
Number of shots
Pre-train Datasets:
CC3m
CC12m-2.7m
Redcaps2.7m
YFCC2.7m
LAION2.7m
WIT2.7m
Shutterstock2.7m
Scratch

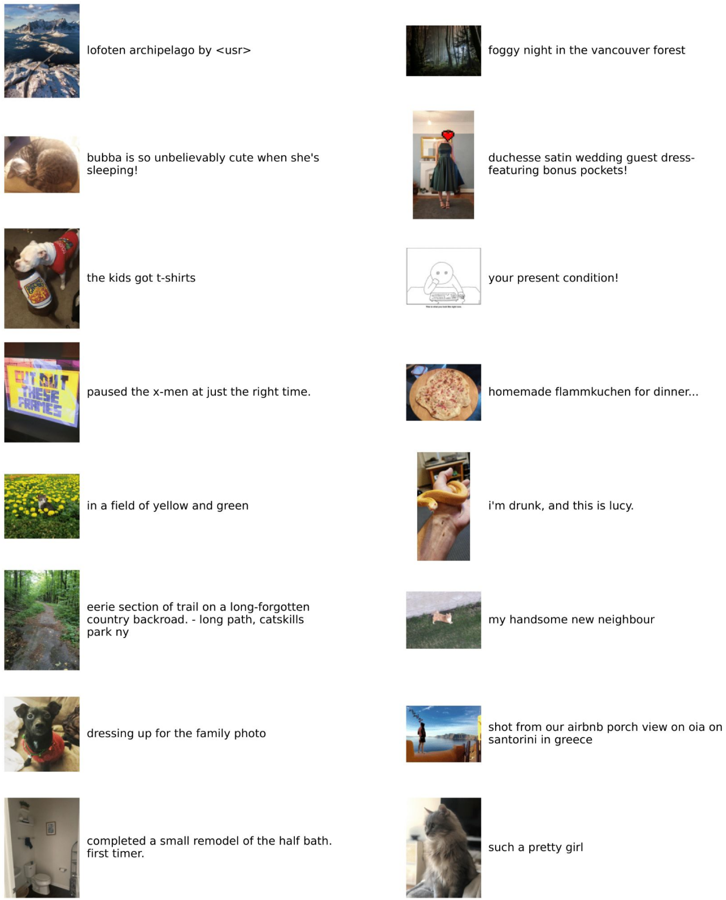

Average over 9 downstream datasets
Best vs Worst Pre-training Datasets
0.6
0.5
0.4
0.3
0.2
0.1
0
1
5
10
20
50
100
500
1000
3000
Number of shots
20-shot Finetune Accuracy
100
90
80
70
60
50
40
30
CIFAR100
DTD
CALTECH101
PETS
REAL
CLIPART
EuroSAT
CassavaLeaf
CameraTraps
Pretraining datasets
CC3m
CC12m-2.7m
Redcaps2.7m
YFCC2.7m
LAION2.7m
WIT2.7m
Shutterstock2.7m

Redcaps on PETS

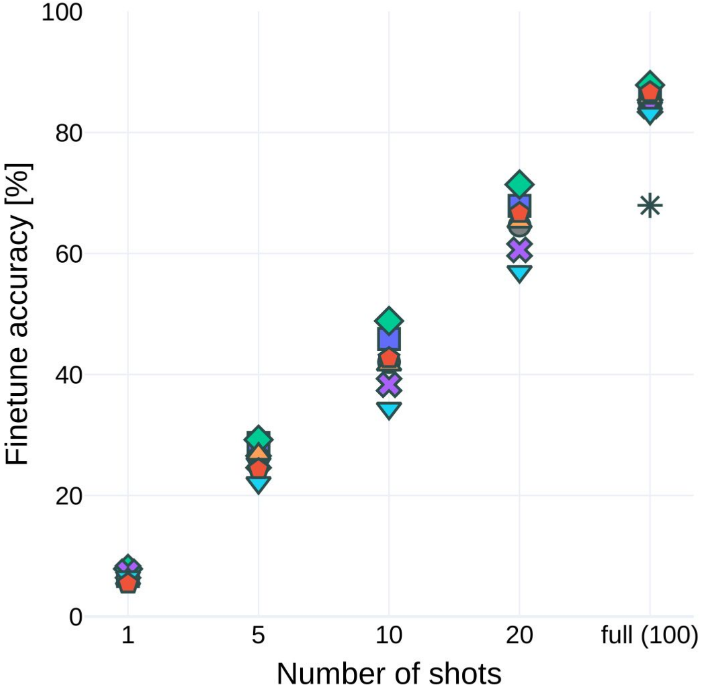


Redcaps on PETS





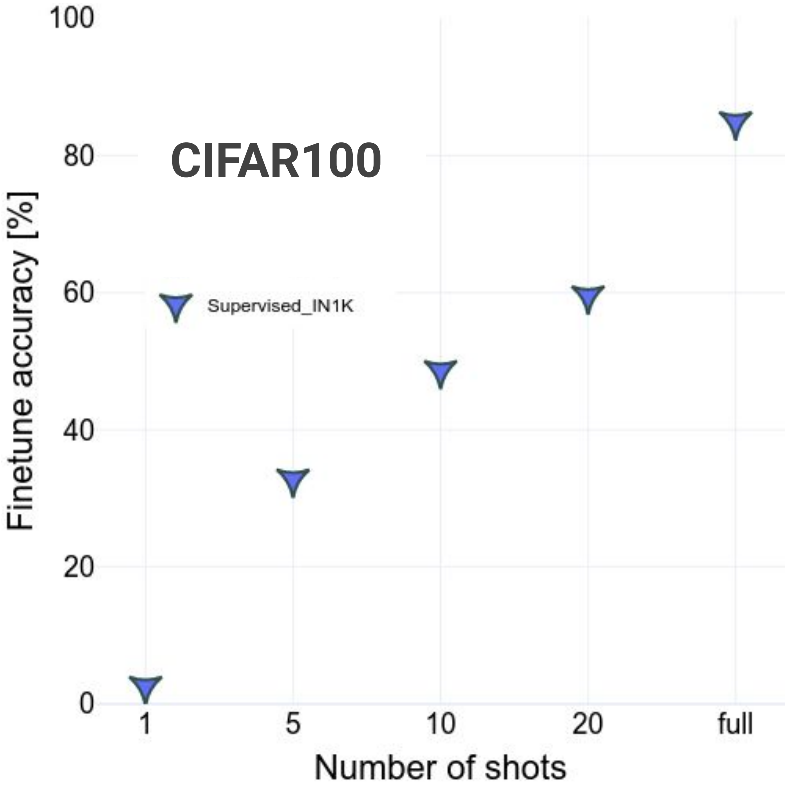
Redcaps on PETS



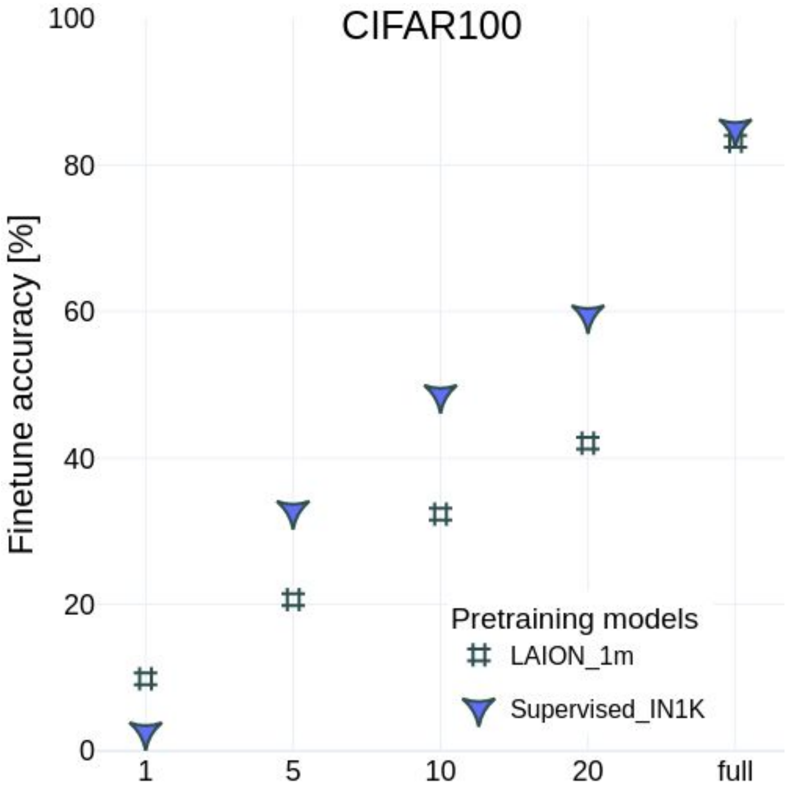
Pre-training dataset	Top 20 words in 1M sample of captions
Shutterstock	background, vector, illustration, design, icon, pattern, texture, style, woman, concept, hand, color, flower, view, template, line, business, logo, card, symbol
Redcaps	day, today, year, time, **cat**, plant, friend, anyone, picture, baby, guy, week, **dog**, home, morning, night, month, way, boy, work
YFCC-15m	photo, day, park, street, city, picture, view, time, world, year, house, state, center, part, garden, shot, image, building, road, museum
LAION-15m	photo, stock, image, black, woman, design, set, vector, white, print, home, men, blue, dress, art, card, sale, gold, bag, cover
CC-12m	illustration, stock, art, design, photo, image, background, room, vector, house, home, woman, wedding, style, photography, royalty, car, fashion, girl, world
CC-3m	background, actor, artist, player, illustration, view, woman, man, football, team, tree, premiere, city, vector, day, girl, beach, game, hand, people
WIT	view, church, station, map, house, building, hall, museum, city, location, street, park, river, state, john, county, town, center, bridge, world

Table 2: Most common words in captions of pre-training distributions

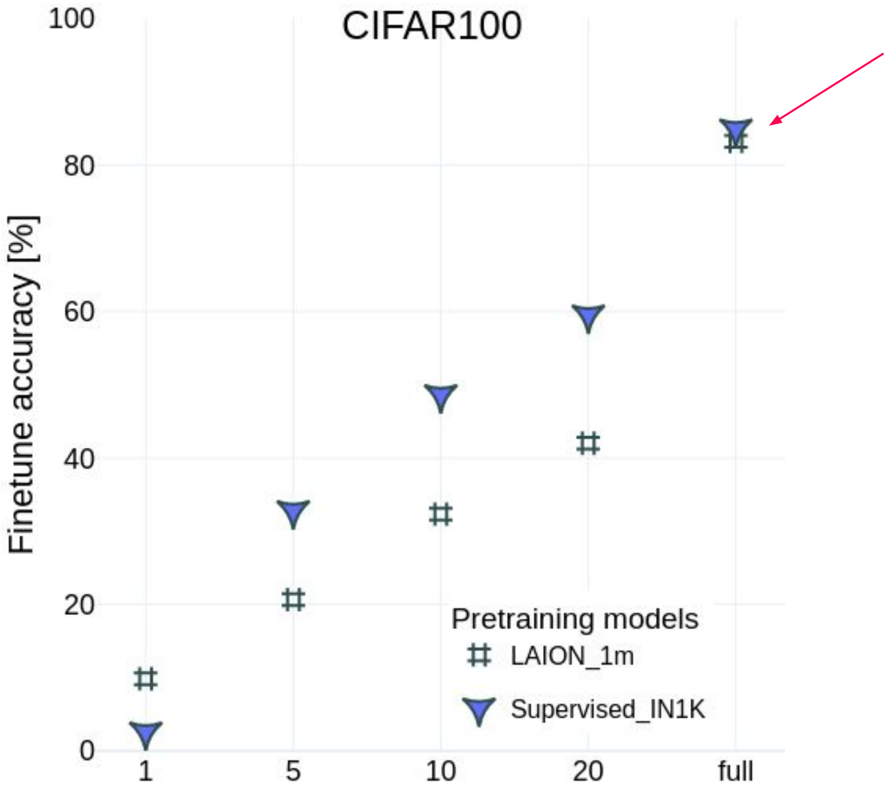
Which pre-train method?



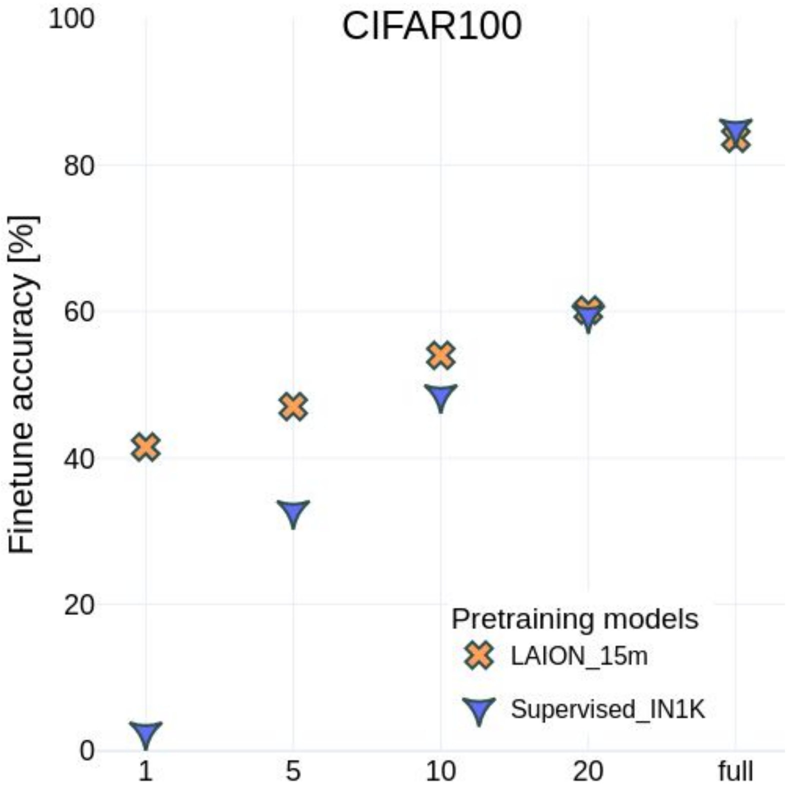
Supervised vs. CLIP



Supervised vs. CLIP

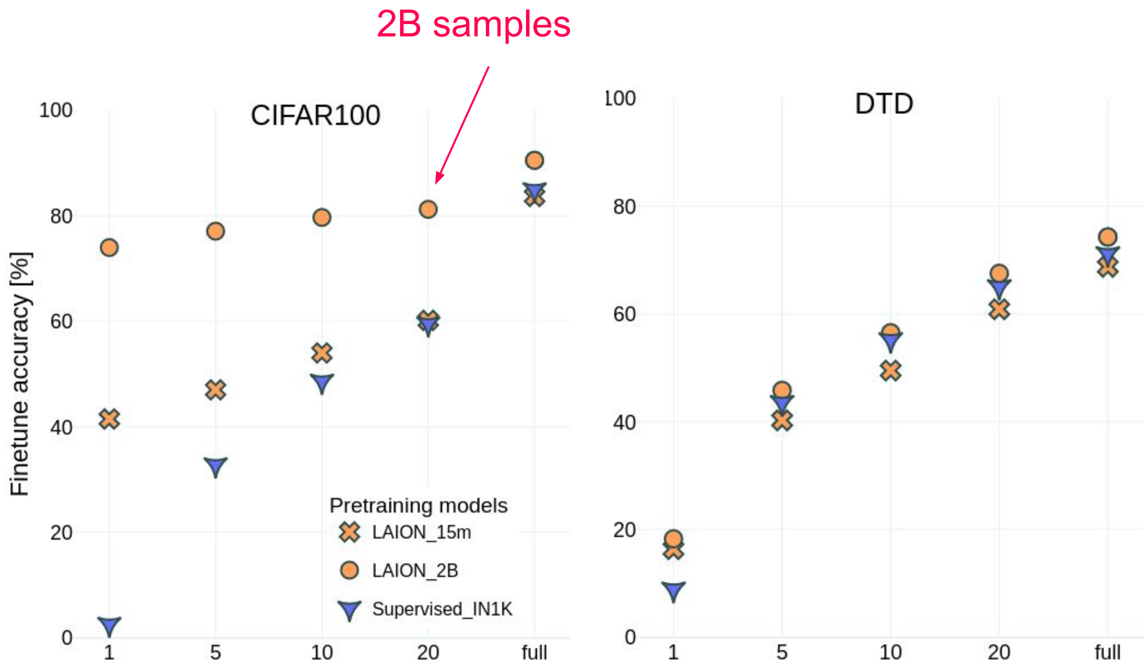


Adding 15x more data?

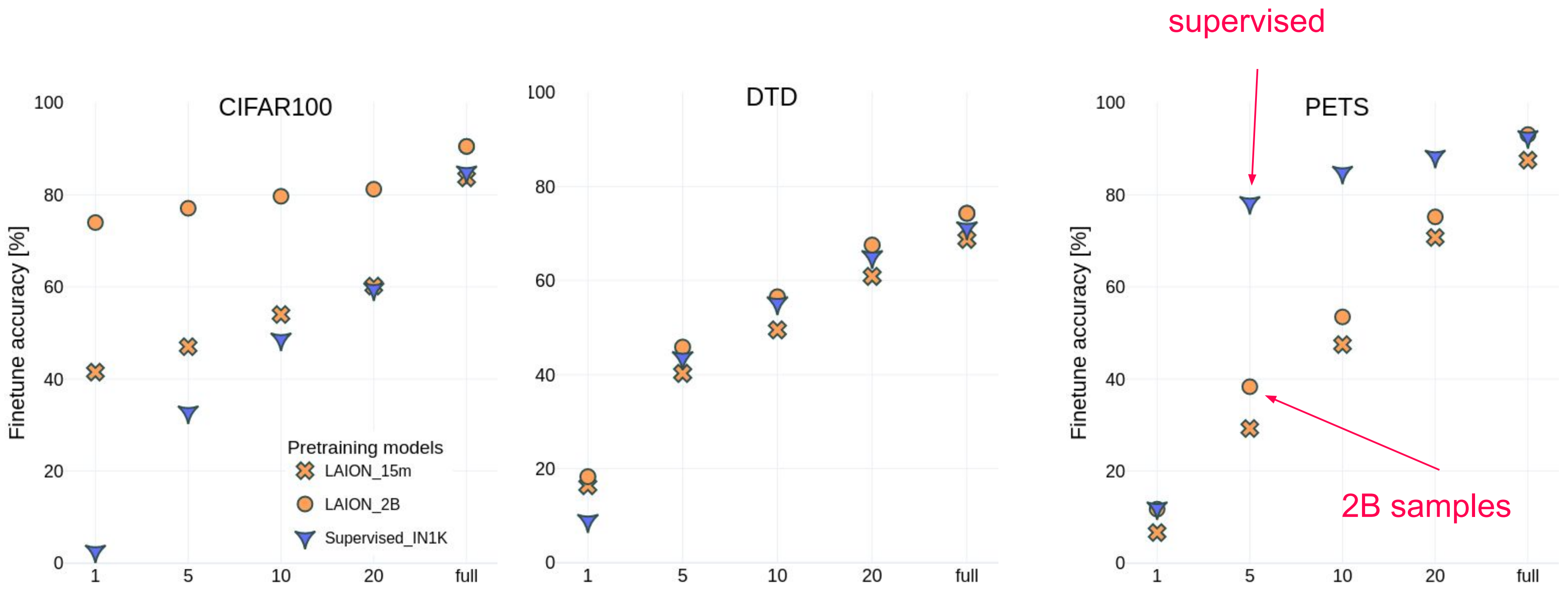


Adding 15x more data?

What if we scale to 2B samples?



What if we scale to 2B samples?



Take away

- Sparsity:
    - There are several motivations for sparity, one is improving generalization.
    - Sparsity has different effects when combined with supervised and semi-supervised training.

Take away

- Loss landscape:
  - Studying the loss landscape of neural networks has implications on model generalization.
  - Accounting for permutation invariance, barriers can be eliminated.
  - New lens to loss landscape: we took the first steps towards understanding ensembles and distributed training.

Take away

- Role of data:
  - Changing the pre-training dataset leads to noticeable differences in few-shot transfer performance.
  - Specific datasets like shutterstock perform well on almost all studied target tasks.
  - Data curation matters. We need 15-2000X more data to compensate for labeling.

Thanks for your attention