ImageNet-RIB Benchmark: Large Pre-Training Datasets Don’t Guarantee Robustness after Fine-Tuning

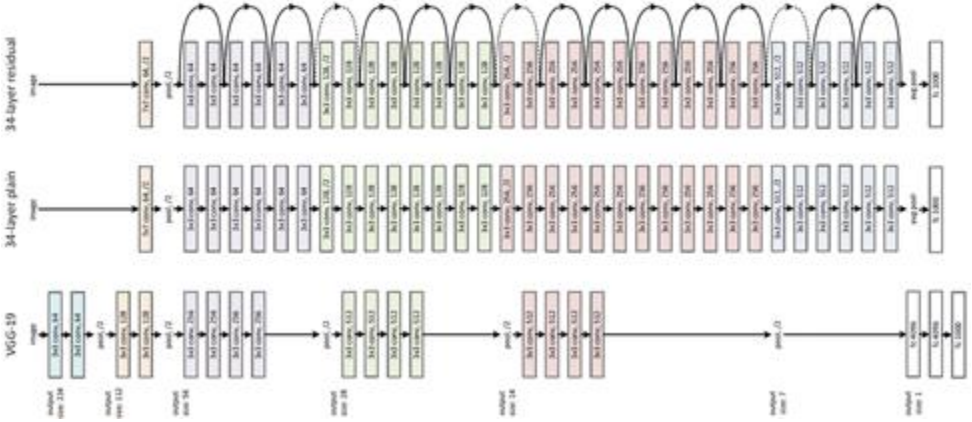
Jaedong Hwang

Massachusetts Institute of Technology

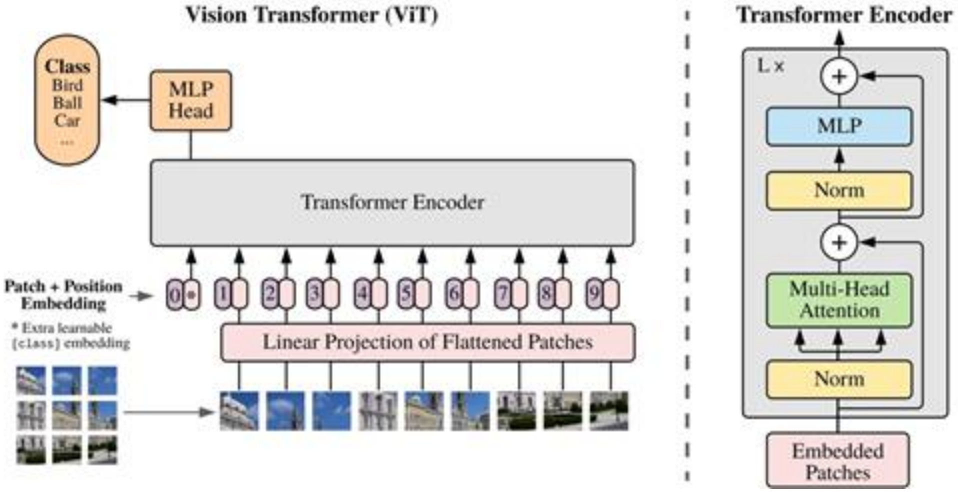
Voxel 51 Boston AI, ML, CV Meetup

Feb 28, 2025

Deep Learning has evolved with Larger Data and Deeper Models



LAION



Training from Scratch vs. Transfer Learning

Training from Scratch

Random Init. Model

Prediction

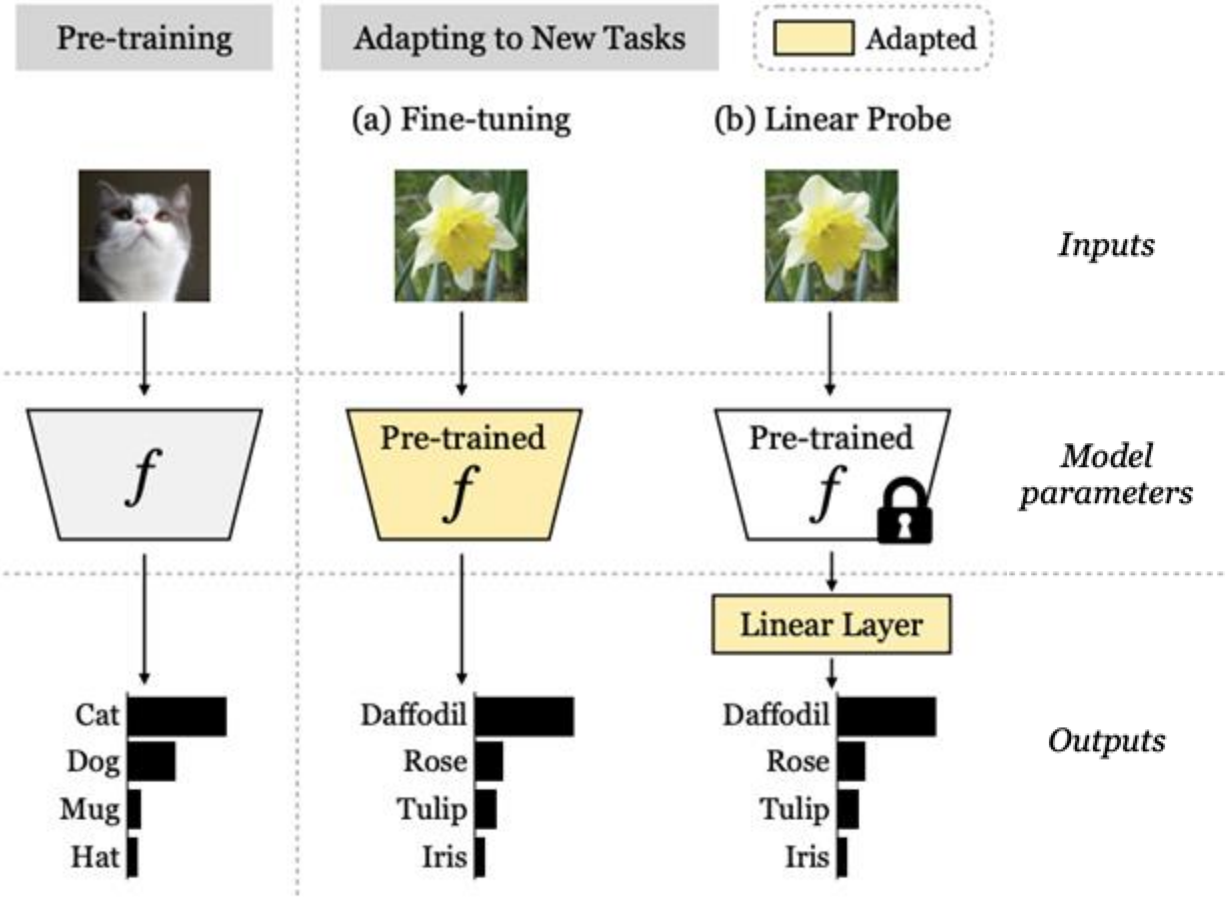
More Common

Transfer Learning

Pre-Trained Model

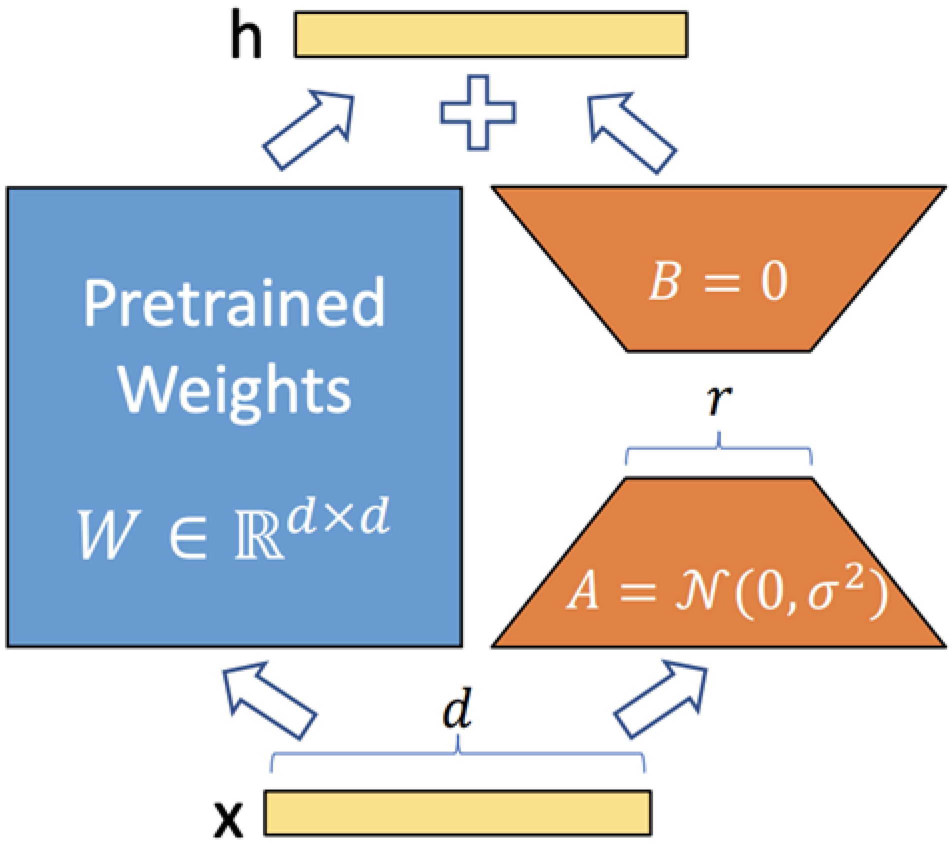
Prediction

Various Adaptation to the New (Downstream) Dataset



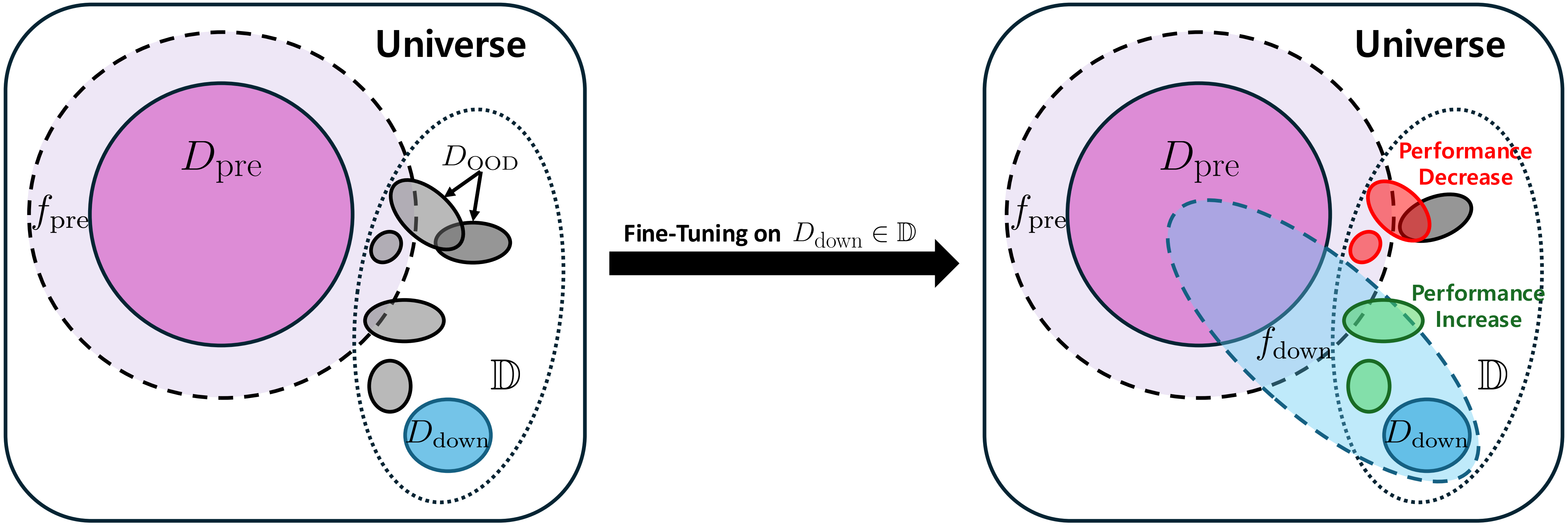
Bahng, Hyojin, et al. "Exploring visual prompts for adapting large-scale models." *arXiv preprint arXiv:2203.17274* (2022).

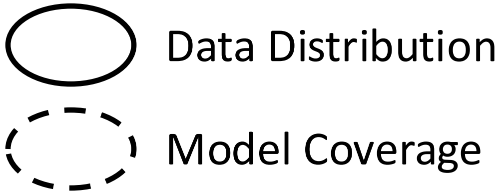
Low-Rank Adaptation (LoRA)



Hu, Edward J., et al. "LoRA: Low-rank adaptation of large language models." arXiv preprint arXiv:2106.09685 (2021)

Model Coverage is Changing in Fine-Tuning





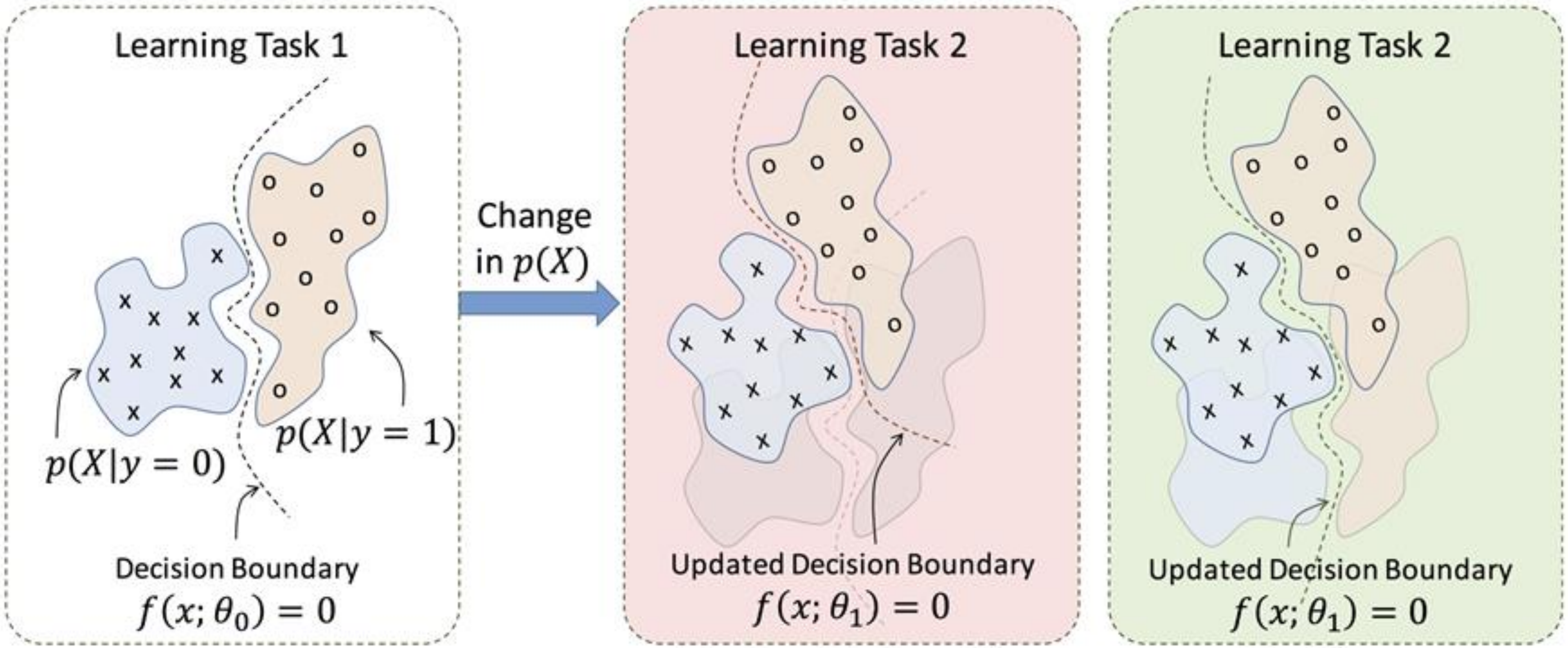
D_{pre} : Pre-Train Dataset

D_{down} : Downstream Dataset

D_{OOD} : Out-of-Distribution Dataset

Catastrophic Forgetting in Transfer Learning / Continual Learning

- Catastrophic Forgetting
  - When we learn a new task, we severely forget the previous task.
- Machine Learning is transudative learning.
  - Meaning that it depends on the data distribution.



Catastrophic Forgetting **Ideal Case**

Kolouri, Soheil, et al. "Attention-based selective plasticity." arXiv preprint arXiv:1903.06070 (2019).

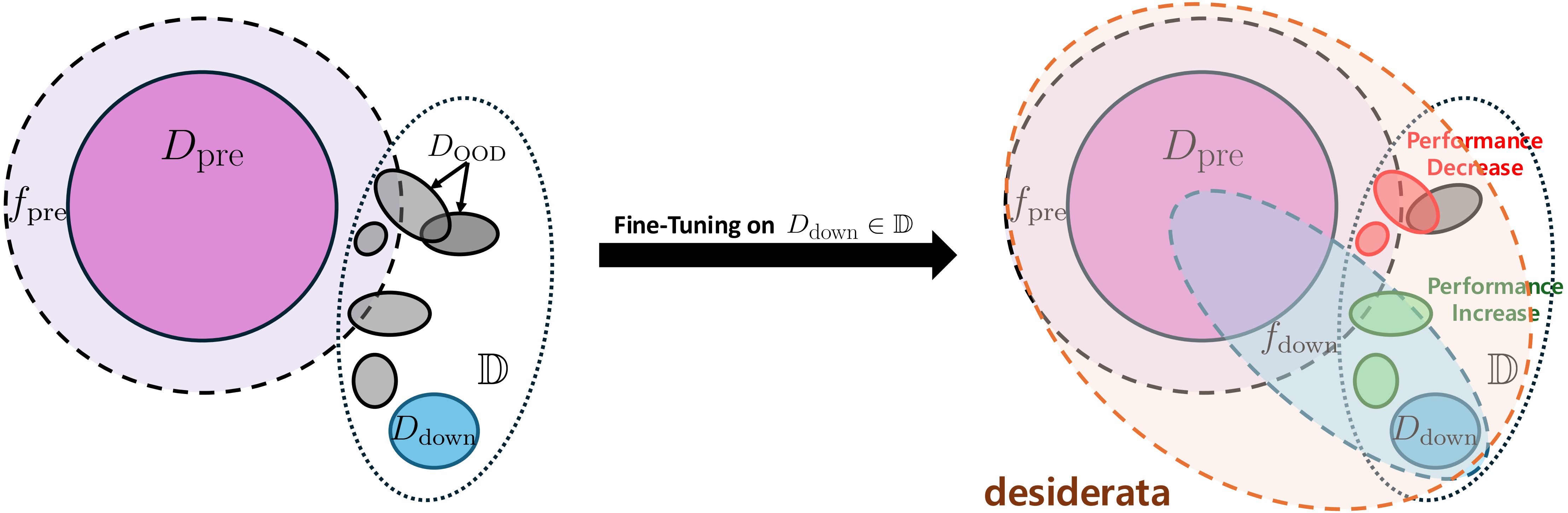
Robustness

- How much we can maintain the performance on out-of-distribution samples.

Out-of-Distribution

Robust Fine-Tuning

- Aims to maintain / improve robustness to OOD data while in fine-tuning



Existing Robust Fine-Tuning Benchmark (Taori et al., 2020)

- Fine-Tuning Pre-Trained model on ImageNet and evaluating on 5 Realistic ImageNet variants

Fine-Tuning

Pre-Trained Model
(e.g., CLIP)

Fine-Tuned Model

Evaluation

ImageNet-V2

ImageNet-A

ImageNet-R

ImageNet-Sketc

ObjectNet

Taori, Rohan, et al. "Measuring robustness to natural distribution shifts in image classification." NeurIPS. 2020

Problem of Robust Fine-Tuning Benchmark (Taori et al. 2020)

Fine-Tuning

Pre-Trained Model

Fine-Tuned Model

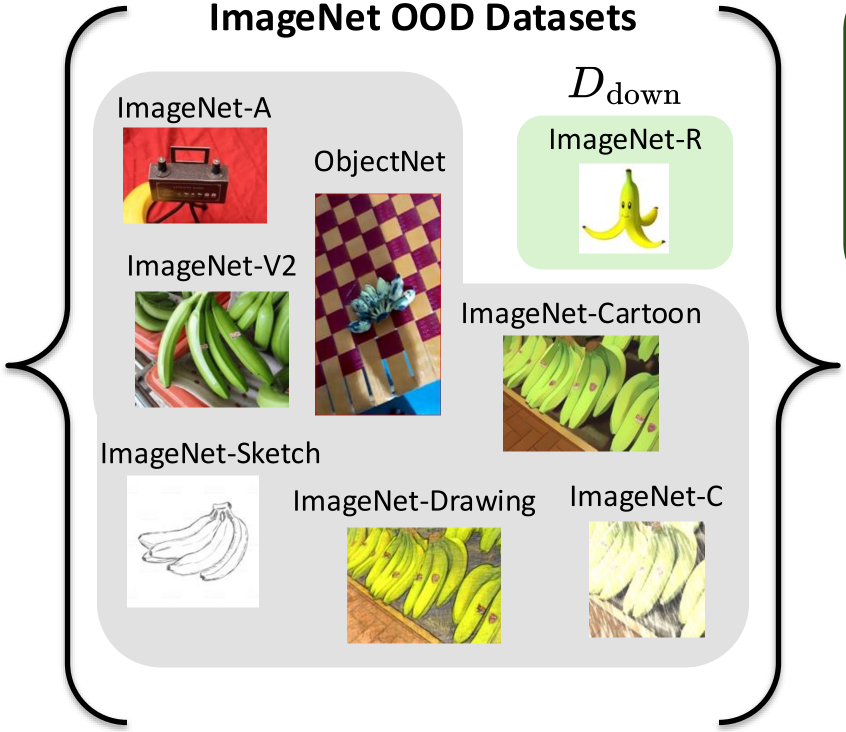
ImageNet-V2	ImageNet-A	ImageNet-R	ImageNet-Sketch	ObjectNet
				

- Some pre-training datasets may contain downstream dataset, ImageNet.
- Fine-tuning on only one dataset.
- No study regarding relationship between downstream dataset and OOD dataset.

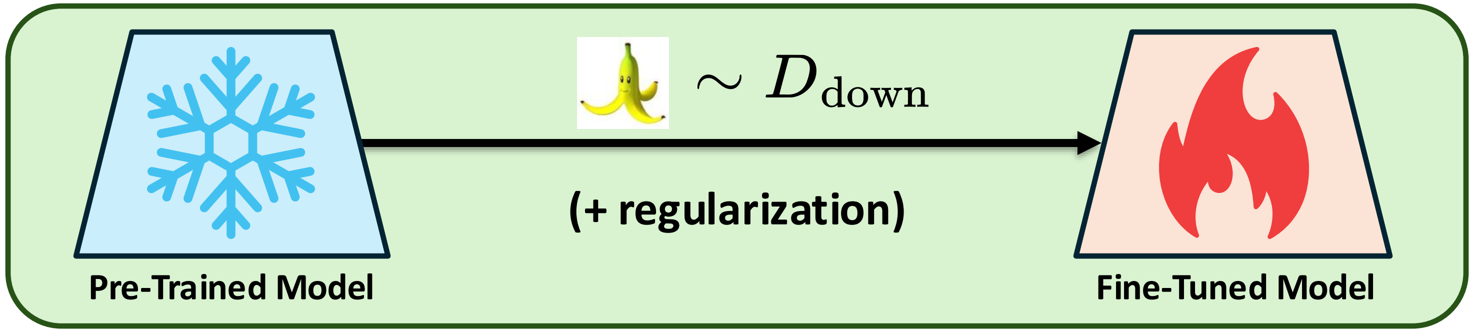
Taori, Rohan, et al. "Measuring robustness to natural distribution shifts in image classification." NeurIPS. 2020

ImageNet-RIB (Robustness Improvement Benchmark)

1. Choose one dataset



2. Fine-tune on the downstream dataset



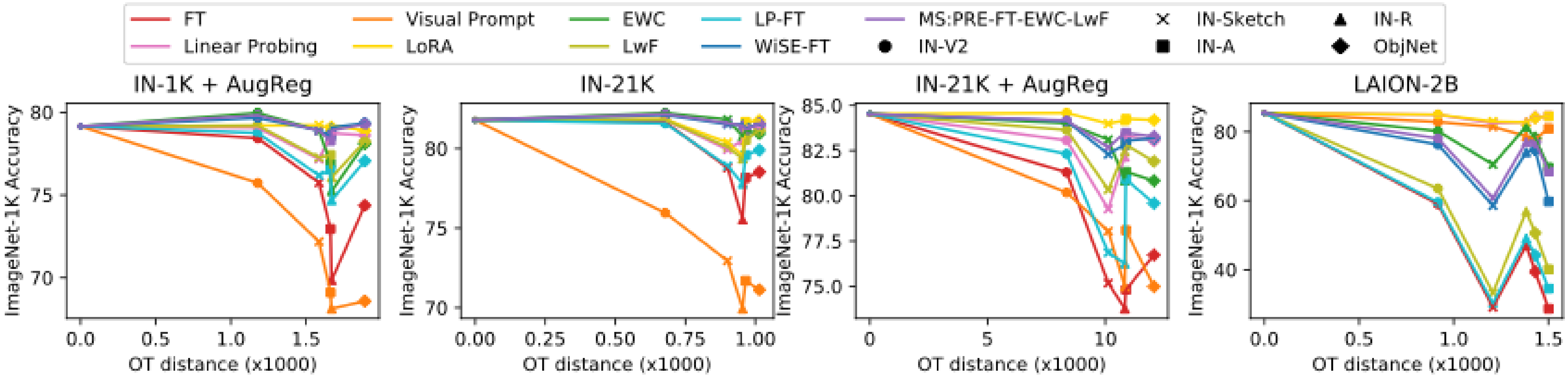
3. Evaluate on other OOD datasets

	IN-V2	IN-A	IN-Sketch	ObjNet	IN-Cartoon	IN-Drawing	IN-C
Acc(❄)	66	15	28	26	66	39	56
Acc(🔥)	59	21	47	32	61	51	52
Robustness Change	↓	↑	↑	↑	↓	↑	↓

4. Repeat 1-3

Dataset Distance can Estimate Performance Drop in Pre-Train Dataset

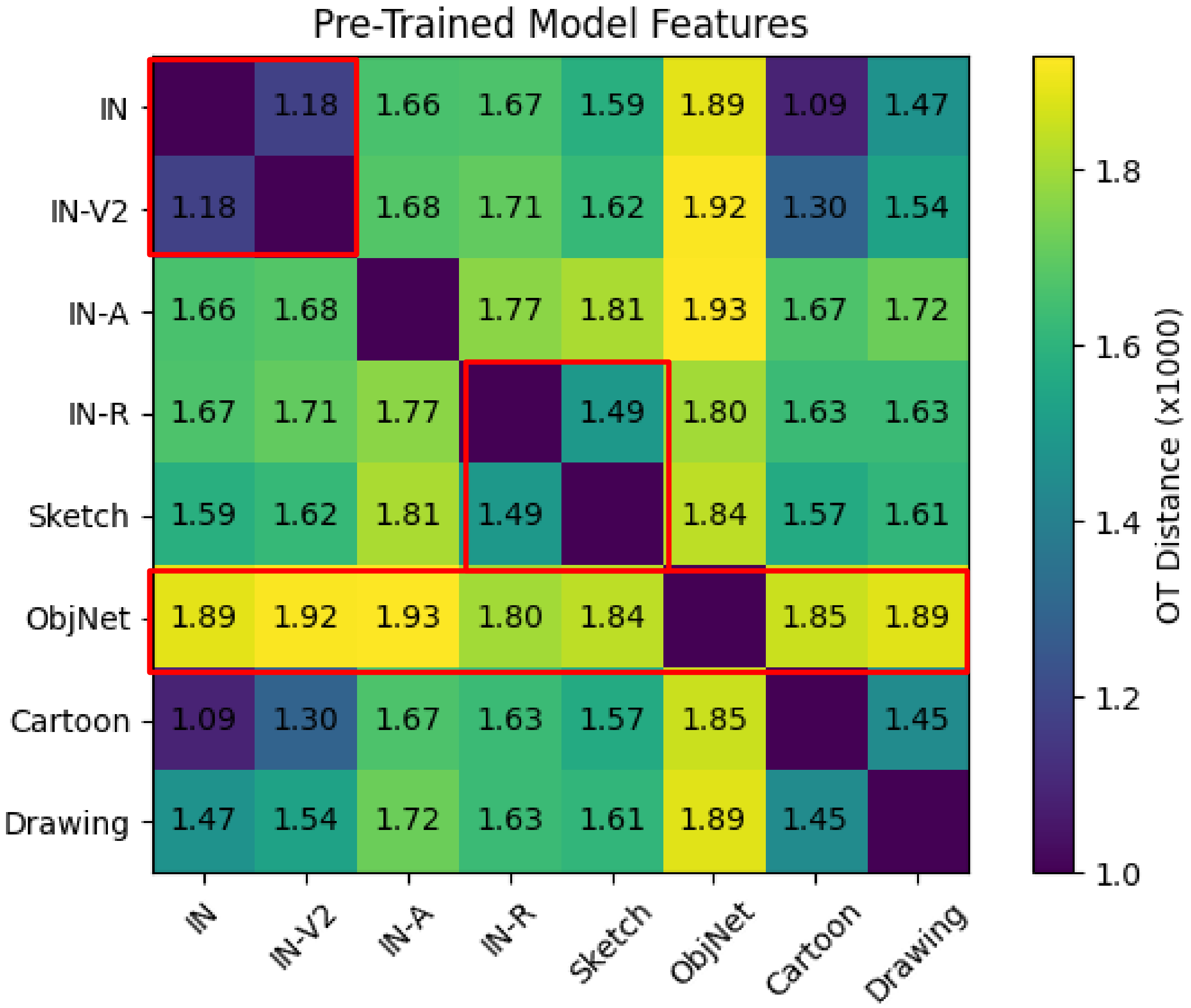
- Measure Performance in ImageNet-1K after fine-tuning on each downstream dataset.
- The performance algins with Optimal Transport Dataset Distance (Alvarez-Melis and Fusi, 2020)
  - Measured in feature space of pre-trained models.

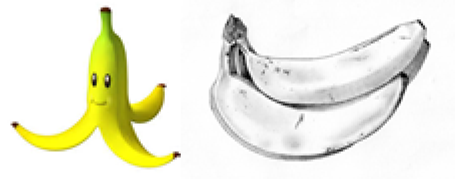


Alvarez-Melis, David, and Fusi, Nicolo. "Geometric dataset distances via optimal transport." *NeurIPS*. 2020

OT Distance Matches with Semantic Different

- Distance is measured using extracted feature from ImageNet pre-trained ViT-B/16





ImageNet-R

ImageNet-Sketch

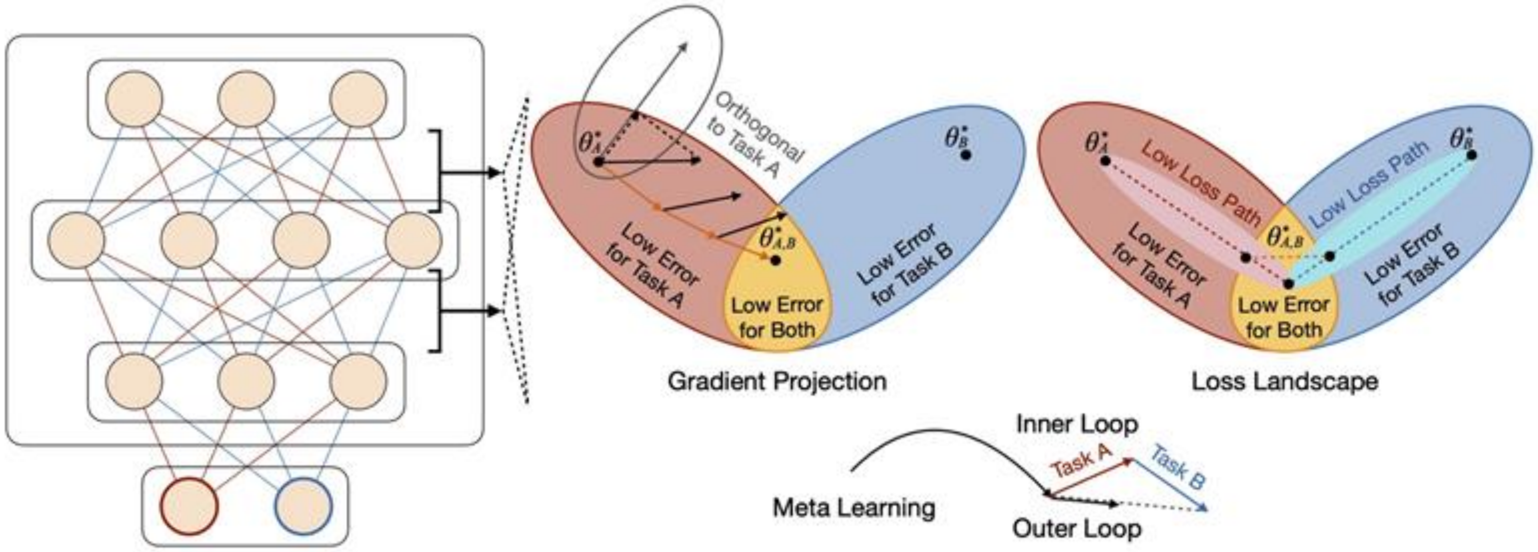
ObjectNet

Alvarez-Melis, David, and Fusi, Nicolo. "Geometric dataset distances via optimal transport." *NeurIPS*. 2020

Methods

- Fine-Tuning
  - (Vanilla) Fine-Tuning, Linear Probing, Visual Prompting, LoRA
- Regularization-Based Continual Learning
  - EWC, LwF
- Robust Fine-Tuning
  - WiSE-FT, LPFT
- Model Soup
  - Average multiple weights.

Regularization-Based Continual Learning

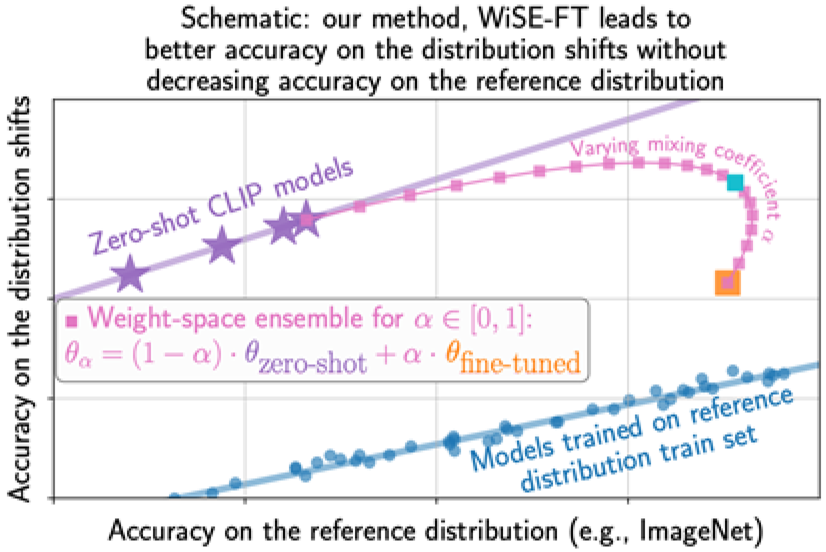


- EWC (Elastic Weight Consolidation)
  - Weight Regularization with pre-trained model's weight
- LwF (Learning without Forgetting)
  - Logit Distillation with pre-trained model's logit

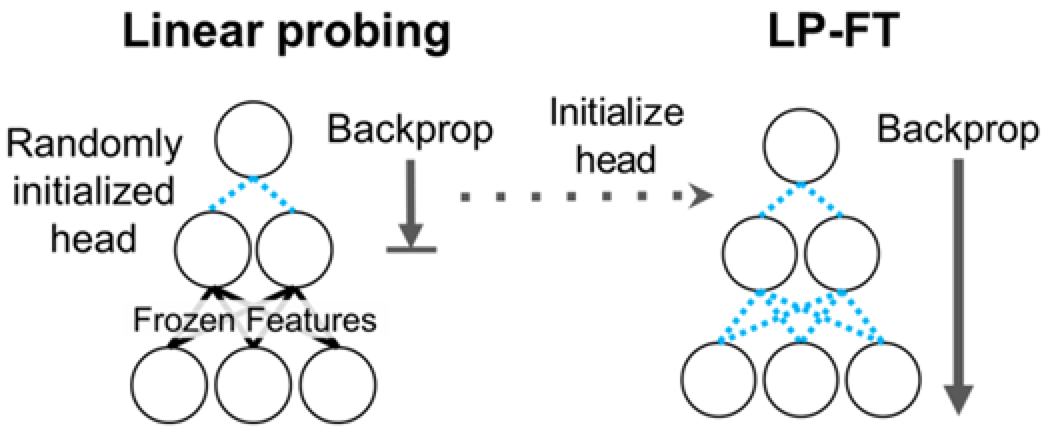
Kirkpatrick, James, et al. "Overcoming catastrophic forgetting in neural networks." *PNAS.* 2017.
Li, Zhizhong, and Derek Hoiem. "Learning without forgetting." *TPAMI.* 2017,
Wang, Liyuan, et al. "A comprehensive survey of continual learning: Theory, method and application." *arXiv preprint arXiv:2302.00487* (2023).

Robust Fine-Tuning Methods

- WiSE-FT
  - Linearly interpolate pre-trained and fine-tuned models.



- LP-FT
  - Linear Probing first
  - Then, Fine-Tuning



Wortsman, Mitchell, et al. "Robust fine-tuning of zero-shot models." *CVPR.* 2022

Kumar, Ananya, et al. "Fine-tuning can distort pretrained features and underperform out-of-distribution." *ICLR.* 2022

Combination of Continual Learning and Weight Interpolation Helps

- Metric: Robust Improvement

$$RI_i = \frac{1}{n-1} \sum_{j=1, j\neq i}^{n} A_i^{(j)} - A_{\text{pre}}^{(j)} \quad mRI = \frac{1}{n} \sum_i^n RI_i$$

average accuracy difference on OOD datasets

- Mean robustness Improvement across various pre-trained dataset

Architecture	ViT-B/16					
Method	IN-1K	IN-1K + AugReg	IN-21K	IN-21K + AugReg	OpenAI	LAION-2B
FT	-6.9	1.3	-0.1	-5.5	-38.0	-38.1
Linear Probing	0.4	0.7	0.4	-0.3	**-2.0**	**-2.0**
Visual Prompt	-7.5	-4.5	-9.4	-8.8	-8.4	-8.2
LoRA	0.5	0.9	-0.3	-2.1	-3.6	-3.6
EWC	0.1	2.8	1.4	0.6	-12.7	-12.5
LwF	-3.6	3.1	1.6	-1.0	-33.1	-33.9
LP-FT	-5.8	2.3	0.5	-2.6	-36.9	-37.1
WiSE-FT	**1.5**	3.6	2.5	1.7	-18.1	-21.6
MS	1.4	**3.9**	**2.7**	**2.2**	-16.0	-17.9

- **WiSE-FT:** Weight Interpolation of Pre-trained model and FT
- **MS**: Model Soup. Weight Interpolation of Pre-trained model, FT, EWC, LwF

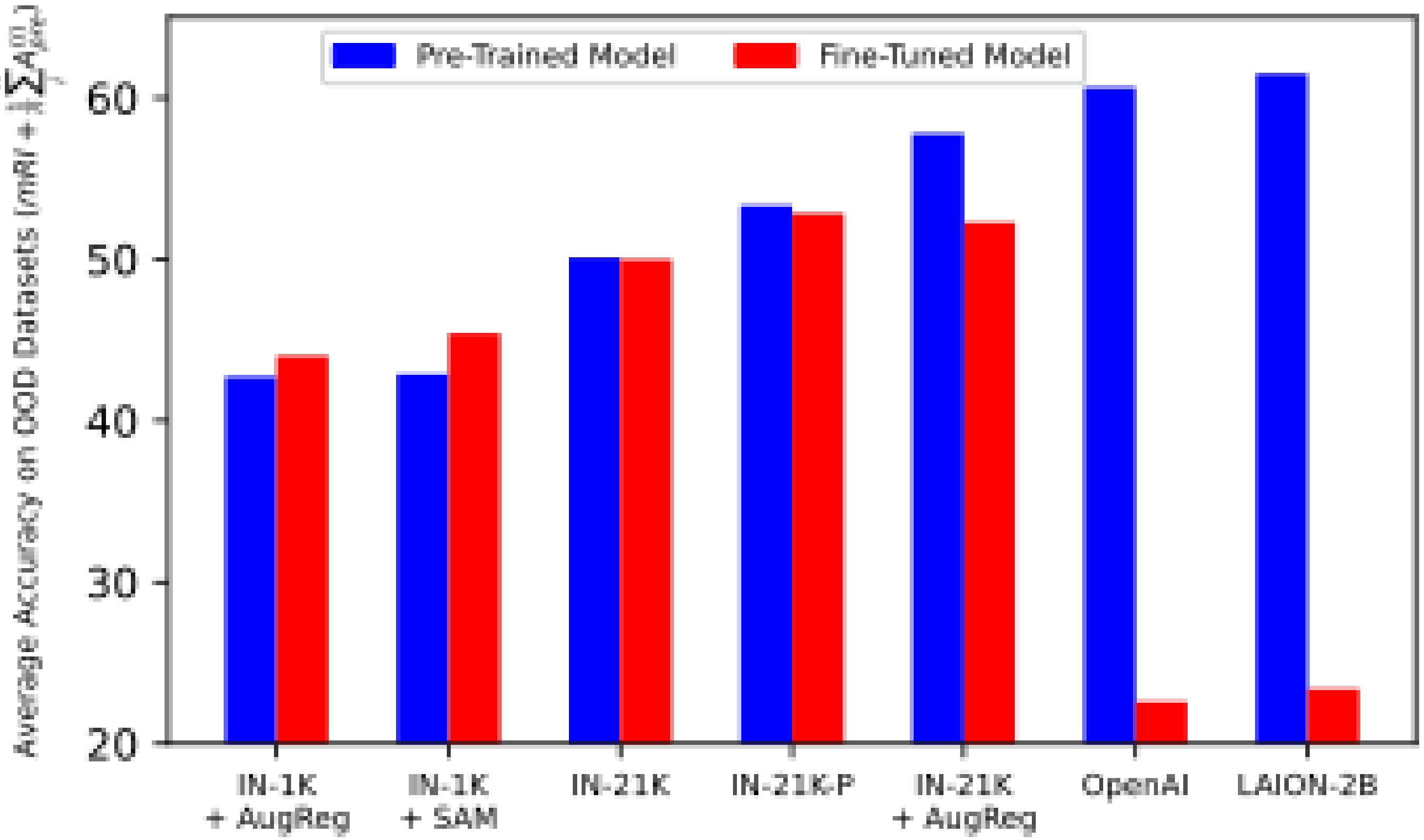
Every pre-trained model was fine-tuned on ImageNet-1K before conducting experiments.

Surprisingly, Pre-Trained on Large-Scale Data Suffers more Forgetting

- ViT-B/16 with different backbone
- All Models are pre-trained on IN-1K before fine-tuning on downstream datasets

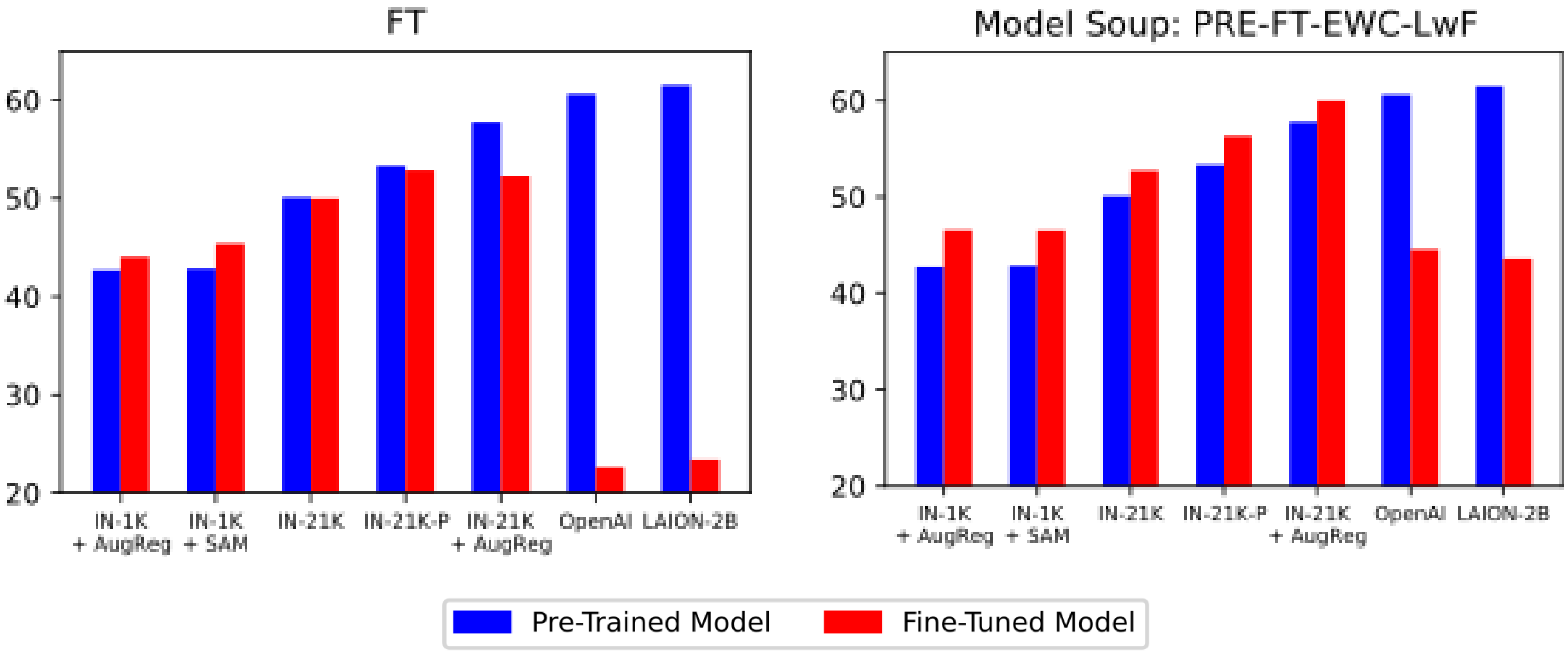
$$\frac{1}{n}\sum_j \frac{1}{n-1}\sum_{i,i\neq j}^{n} A_{\text{down}}^{(i)}$$

$$= mRI + \frac{1}{n}\sum_i^n A_{\text{pre}}^{(i)}$$



Combination of Continual Learning and Weight Interpolation Relax this Issue

- Continual learning methods with post-hoc robust fine-tuning methods can relax the problem
- However, it is not a fundamental solution.

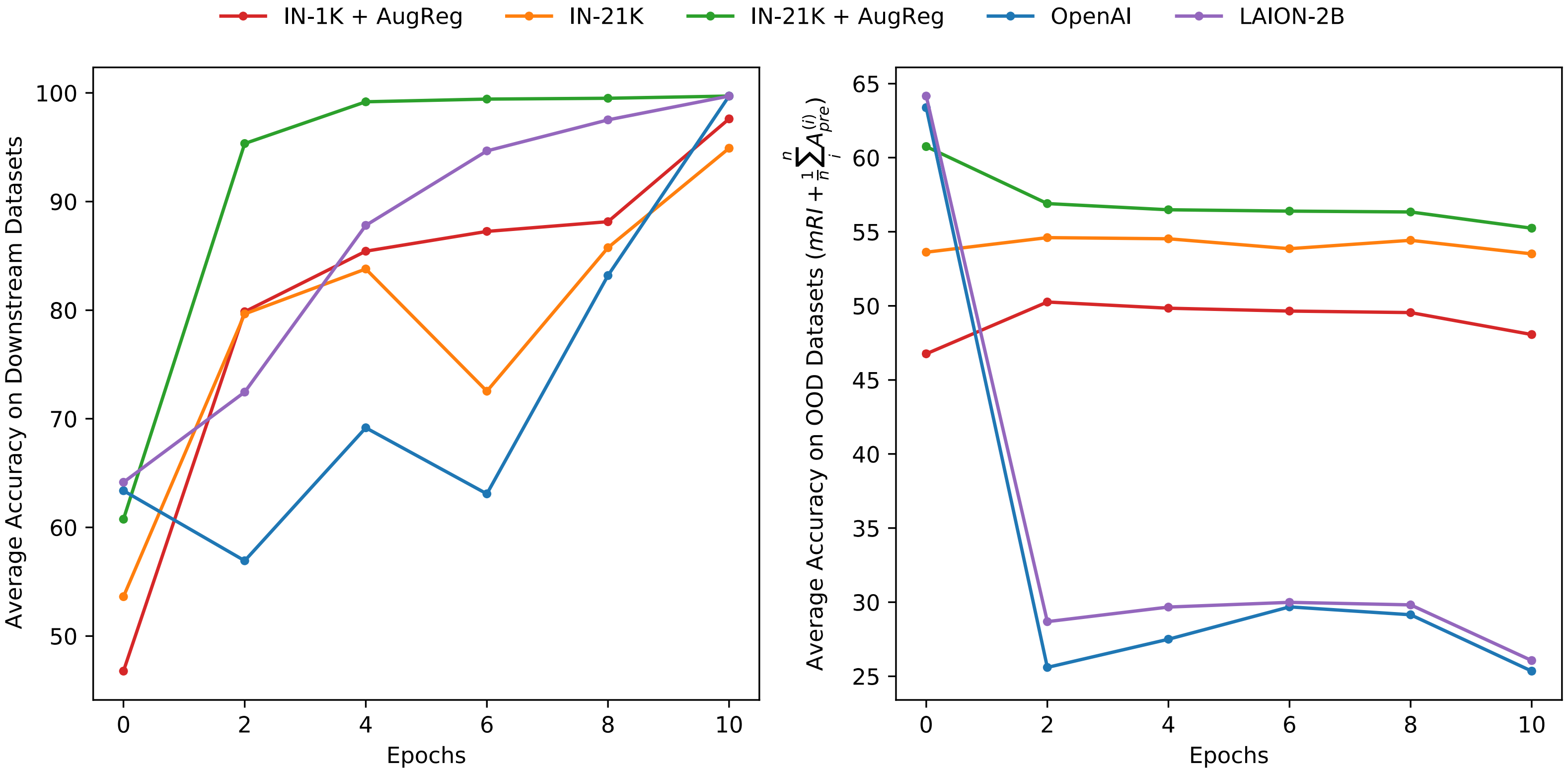


Hypothesis of Performance Degradation

- Overfitting
  - Pre-trained on larger dataset leads better OOD generalization before fine-tuning.
- Texture
  - Models fine-tuned on ImageNet-1K had good generalizability in Taori et al. (2020)
  - Downstream datasets in ImageNet-RIB have various styles, e.g., cartoon, drawing, and sketch.
- Dataset Size
  - ImageNet-1K has 1.2M images while downstream datasets have 50K images in general

Severe Catastrophic Forgetting Happens before Overfitting

- Measure accuracy on the downstream datasets and OOD datasets during fine-tuning.
- IN-21K with AugReg pre-trained model learns the fastest and OpenAI pre-trained model learns slowest.
- But only OpenAI and LAION-2B pre-trained model suffers huge robustness drop.



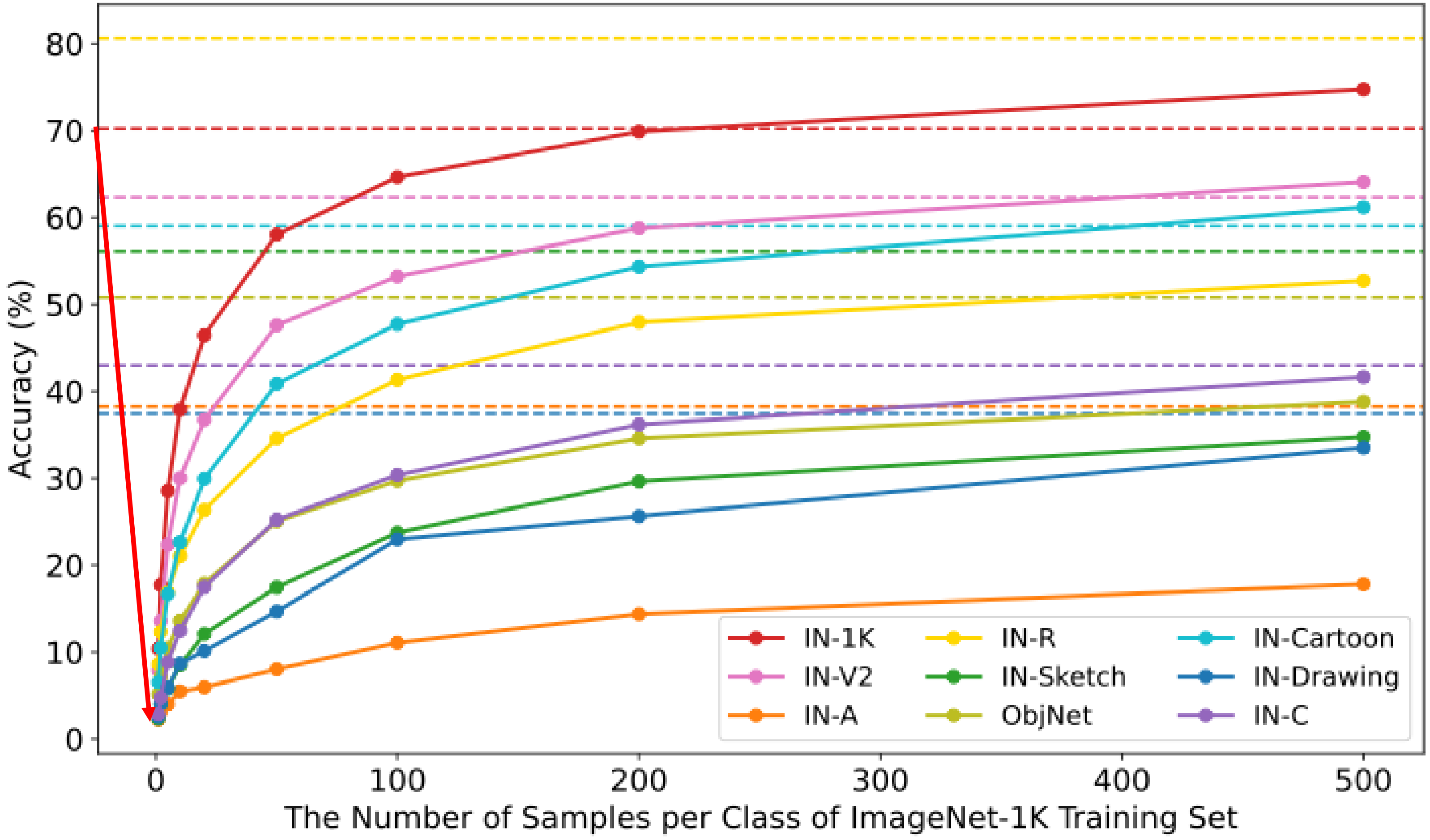
Downstream Dataset Texture Does Not Account for Forgetting

- Considering good robustness of LAION-2B CLIP fine-tuned on ImageNet-1K train set, downstream style may be the cause.
- Fine-tuning pre-trained model on ImageNet-1K validation set also leads to the severe forgetting.

Pre-Training Dataset	IN-1K	IN-V2	IN-A	IN-R	IN-Sketch	ObjNet	IN-Cartoon	IN-Drawing	IN-C
IN-1K + AugReg	97.5 (+18.3)	66.9 (+0.4)	23.3 (+8.3)	40.9 (+2.9)	29.5 (+1.5)	37.2 (+4.2)	71.1 (+4.9)	41.0 (+1.9)	59.5 (+3.5)
IN-1K + SAM	87.3 (+7.1)	69.4 (+1.2)	17.7 (+8.7)	41.8 (+1.7)	30.1 (+2.4)	38 (+3.8)	72.1 (+5.2)	42.9 (+0.6)	56.9 (+2.3)
IN-21K	94.7 (+12.9)	71.6 (+0.2)	38.5 (+6.5)	49.9 (+2.6)	36.7 (+0.9)	45.2 (+2.7)	73.9 (+4.5)	44.1 (0.0)	59.8 (+1.5)
IN-21K-P	96.9 (+12.6)	73.0 (-1.0)	41.4 (+7.3)	51.5 (0.0)	39.8 (-0.4)	45.8 (-0.9)	76.4 (+2.9)	44.3 (-0.8)	61.7 (+0.3)
IN-21K + AugReg	99.9 (+15.4)	70.6 (-3.4)	42.2 (-1.0)	54.1 (-2.7)	39.4 (-3.8)	47.9 (-0.5)	84.5 (+9.4)	55.5 (+0.6)	69.7 (+3.2)
OpenAI	99.9 (+14.6)	59.9 (-15.8)	13.9 (-33.4)	34.9 (-31.0)	19.7 (-31.2)	30.5 (-20.2)	75.0 (-1.3)	33.4 (-22.3)	45.7 (-16.9)
LAION-2B	99.9 (+14.4)	59.4 (-16.2)	12.6 (-28.9)	36.3 (-32.5)	23.4 (-32.0)	30.4 (-20.7)	73.0 (-5.2)	30.6 (-27.8)	41.8 (-21.2)

Downstream Dataset Size is a Major Determinant

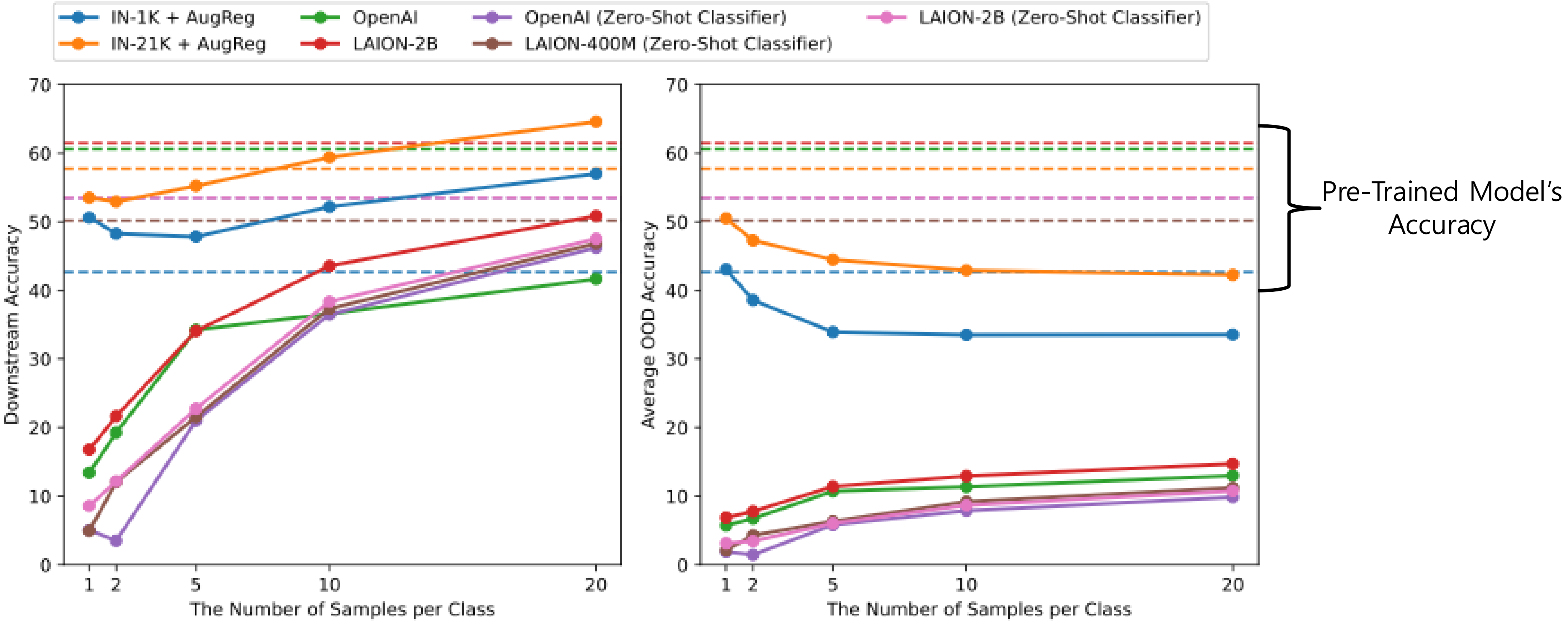
- Fine-tuning LAION-2B pre-trained CLIP (without fine-tuning on ImageNet-1K) with zero-shot classifier on portion of ImageNet-1K train set.



Pre-trained model's performance

Downstream Dataset Size is a Major Determinant

- Fine-tuning on K-images / class in downstream dataset.
- IN-1K / 21K pre-trained models drop performance on downstream dataset shortly and recover.
- LAION-2B / OpenAI pre-trained models suffer huge forgetting even in 1-shot.



Conclusion

- Propose a new robustness fine-tuning benchmark for understanding the impact of down stream datasets.
  - ➓ Model Soup has the best *mRI* with ImageNet pre-trained model.
  - ➓ Linear Probing has the best *mRI* with LAION-2B pre-trained model.
- Pre-train on large dataset and then fine-tune on small dataset leads huge catastrophic fo rgetting.
  - ➓ Full Fine-Tuning is required when the downstream dataset is far from pre-training dataset.
  - ➓ It challenges the common belief that pre-trained on the largest dataset is always better.
- Question remains whether this problem happens in other domains.

Thank You!

paper

Jaedong Hwang

Akhilan Boopathy

Ila Fiete

Zhang-Wei Hong

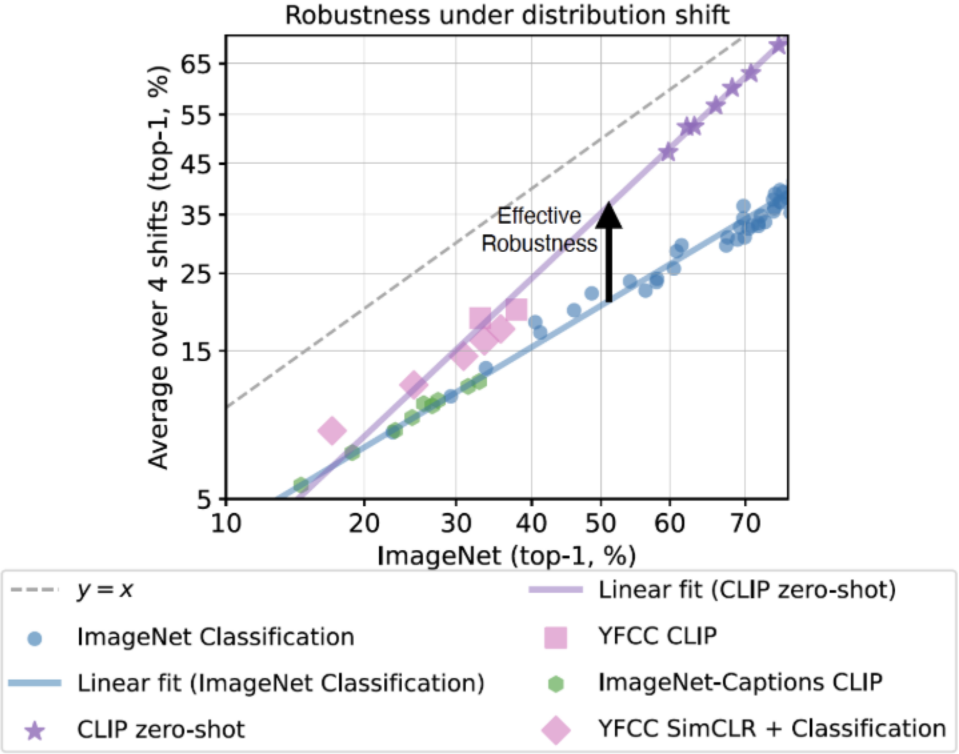
Brian Cheung

Pulkit Agrawal

Contact: jdhwang@mit.edu

Fang et al. (2020)

- **CLIP's robustness is related to pre-training data distribution not contrastive loss.**
- Models pre-trained with various contrastive objectives on ImageNet do not achieve the same effective robustness as CLIP models



Self-sup (SimCLR, SwAV,...) does not affect

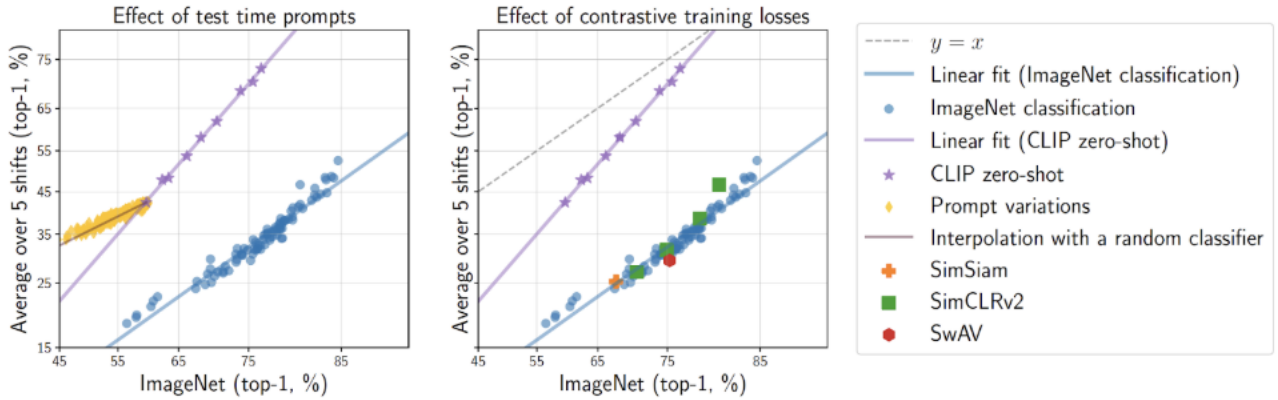


Figure 6. Effect of prompting strategies and contrastive objectives on robustness. (Left) On most natural distribution shifts, effect of prompting on effective robustness is similar to that of random interpolation. (Right) Models pre-trained with various contrastive objectives on ImageNet do not achieve the same effective robustness as CLIP models.

Fang, Alex, et al. "Data determines distributional robustness in contrastive language image pre-training (clip)." ICML 2022