Understanding and Mitigating the Pre-training Noise on Downstream Tasks

Hao Chen, Jindong Wang, Ankit Shah, Ran Tao
Hongxin Wei, Xing Xie, Masashi Sugiyama, Bhiksha Raj
haoc3@andrew.cmu.edu

https://arxiv.org/pdf/2309.17002.pdf
https://arxiv.org/abs/2403.06869.pdf

Carnegie Mellon University
Microsoft Research 微软亚洲研究院
RIKEN
MOHAMED BIN ZAYED UNIVERSITY OF ARTIFICIAL INTELLIGENCE
SUSTech Southern University of Science and Technology

Content

- Background and Motivation of Noisy Model Learning
- Motivating Experiments on Effects of Pre-training Noise
- Feature Space Analysis of Pre-training Noise
- Mitigation of Pre-training Noise on Downstream Tasks
- More Experiments and Discussions

Background and Motivation

Catastrophic Inheritance of Large Foundation Models

Noisy Model Learning

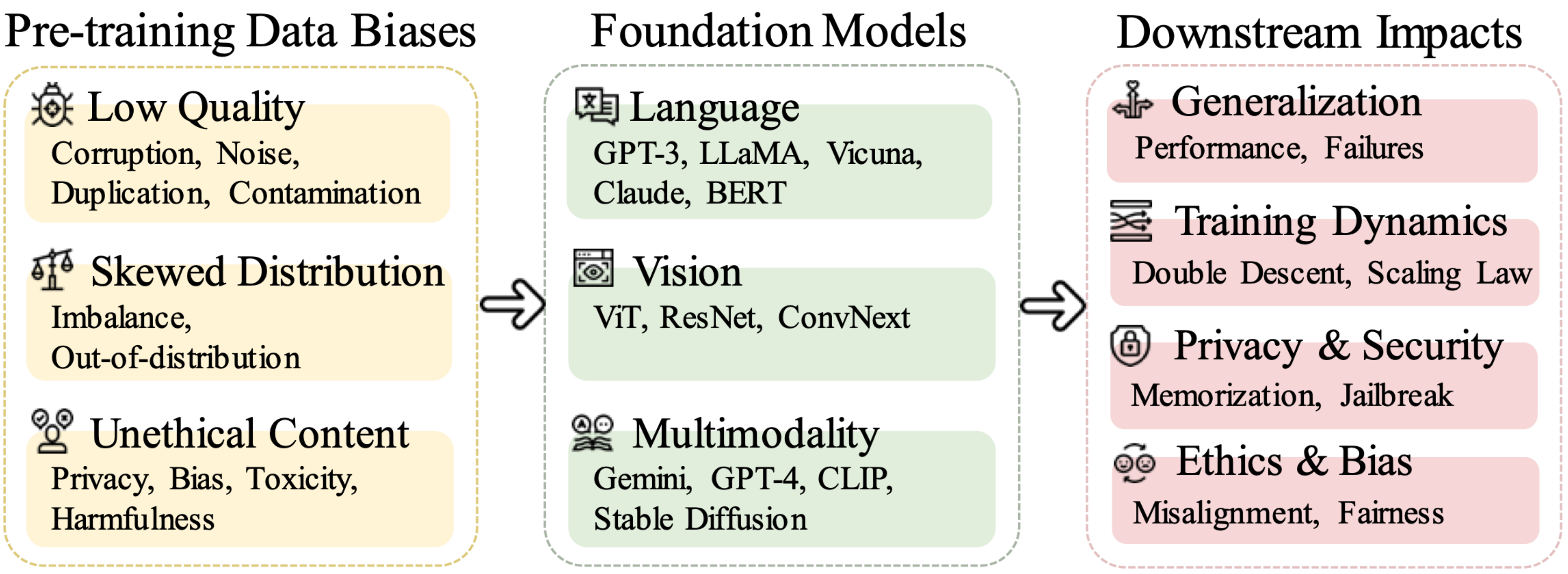
Background – Large Foundation Models

- Large foundation models require massive pre-training data
  - Open CLIP – 2.0 billion image-text pairs
  - Llama – 2.0T tokens

- Adaption of foundation models
  - Pre-training on proxy tasks
  - Tuning on specific downstream tasks (linear probing, parameter efficient tuning, full fine-tuning, etc.)

- Success of foundation models attributed to the pre-training data
  - Large-scale pre-training data are usually collected from web
  - Inevitable noise (and other types of bias) in pre-training data that may lead to unexpected generalization performance and behavior

Pre-training Bias -> Catastrophic Inheritance



- Pre-training biases used to train foundation models may be inherited to downstream tasks with malicious impacts
- Unexplored direction yet very important and interesting

Hao Chen et al. On catastrophic inheritance of large foundation models. 2024.

Examples of Catastrophic Inheritance

Table 1: Realistic examples of catastrophic inheritance from published papers or news.

Example	Domain	Source
Stable Diffusion models was trained on Laion-5B, which contains hundreds of harmful images of child sexual abuse material (CSAM). Then, the model was reported to memorize during training and generate CSAM at production.	Ethics and privacy	[Birhane et al., 2023, Forbes, 2023, Thiel, 2023]
At least 50% of poisoning, adversarial, and backdoor vulnerabilities will be inherited from pre-training data to fine-tuned models, which can be easily triggered at the deployment. Jailbreaks may also relate to pre-training biases.	Security	[Wang et al., 2018, Zhang et al., 2022, Carlini et al., 2023a, Zou et al., 2023]
An MIT student asked AI to make her headshot more 'professional.' It gave her lighter skin and blue eyes. Country bias also found in language models.	Bias	[Boston.com, 2023, Wang et al., 2023a]
Fine-tuning LLMs on only 10 adversarially designed or even benign samples leads to degradation of safety alignment, which costs less than $0.2 using API.	Misalignment	[Qi et al., 2023]
Noisy labels contained in pre-trained data always hurt downstream OOD performance; more than 10% noisy data will hurt in-domain performance.	Generalization	[Chen et al., 2024]
Large language models like GPT-3.5 exhibited an accuracy reduction of 18.12% when answering non-English medical questions. Similar for coding tasks.	Model behaviors	[Jin et al., 2024, Zheng et al.]
Noise in the pre-training data strengthen the double descent phenomena, where the critical point of LFMs overfitting/memorizing data appears earlier.	Training dynamics	[Nakkiran et al., 2019]

This work

Hao Chen et al. On catastrophic inheritance of large foundation models. 2024.

This Work: Inevitable Pre-training Noise

- Evidence in CLIP
  - OpenAI trains CLIP on WIT-400M (not public)
  - OpenCLIP trains CLIP on Laion-2B, with more noisy image-text pairs
  - Yet they achieve similar zero-shot performance

	Data	Arch.	ImageNet	VTAB+
CLIP [55]	WIT-400M	L/14	75.5	55.8
Ours	LAION-2B	L/14	75.2	54.6
Ours	LAION-2B	H/14	<u>78.0</u>	<u>56.4</u>

Mehdi Cherti et al. Reproducible scaling laws for contrastive language-image learning. 2022.

Inevitable Pre-training Noise

- Evidence in LLM
  - Repeated data/corruption [2]
  - Leads to memorization of these noise

OpenWebText		C4		mC4-en		OSCAR		The Pile	
n-gram	Count	n-gram	Count	n-gram	Count	n-gram	Count	n-gram	Count
- - - - - - - - - -	3.4M	? ? ? ? ? ? ? ? ? ?	9M	. . . .	1.76B		773M	- - - - - - - - - -	3.64B
. . . .	1.05M	. . . .	7.27M	- - - - - - - - - -	823M	\ \ \ \ \ \ \ \ \ \	395M	= = = = = = = = = =	602M
= = = = = = = = = =	830K	- - - - - - - - - -	4.41M	� � � � � � � � � �	349M	- - - - - - - - - -	175M	* * * * * * * * * *	188M
* * * * * * * * * *	595K	* * * * * * * * * *	3.87M	* * * * * * * * * *	314M	. . . .	91.6M	) { ref - type = " fig " }	59.1M
# # # # # # # # # #	302K	! ! ! ! ! ! ! ! ! !	1.91M	\ / s \ / files \ / 1 \	183M	* * * * * * * * * *	34.9M	/ / / / / / / / / /	56.2M
amp ; amp ; amp ; amp ; amp ;	278K	. You can follow any responses to this entry through	784K	/ s \ / files \ / 1 \ /	183M	= = = = = = = = = =	22.9M	. . . .	54.9M
; amp ; amp ; amp ; amp ; amp	265K	� � � � � � � � � �	753K	\ / \ / cdn.shopify.com \ / s \ /	182M	(Opens in new window) Click to share on	15.7M	# # # # # # # # # #	38.3M
— — — — — — — — — —	249K	You can follow any responses to this entry through the	752K	/ cdn.shopify.com \ / s \ / files \ /	182M	Log Out / Change) You are commenting using your	13.6M	} - - - - - - - - -	30.1M
… … … … … … … … … …	88.1K	can follow any responses to this entry through the RSS	752K	\ / cdn.shopify.com \ / s \ / files \	182M	(Log Out / Change) You are commenting using	13.6M	{ ref - type = " fig " })	28.9M
~ ~ ~ ~ ~ ~ ~ ~ ~ ~	83.3K	follow any responses to this entry through the RSS 2.0	748K	/ \ / cdn.shopify.com \ / s \ / files	182M	. (Log Out / Change) You are commenting	13.6M	} = = = = = = = = =	21.8M

RedPajama		S2ORC		peS2o		LAION-2B-en		The Stack	
n-gram	Count	n-gram	Count	n-gram	Count	n-gram	Count	n-gram	Count
. . . .	670M	q q q q q q q q q q	30.2M	. . . .	1.42M	- - - - - - - - - -	1.65M	- - - - - - - - - -	4.29B
- - - - - - - - - -	507M	. . . .	5.49M	[1] [2] [3] [	457K	� � � � � � � � � �	1.43M	* * * * * * * * * *	3.87B
\ \ \ \ \ \ \ \ \ \	213M	+ + + + + + + + + +	3.03M	] [2] [3] [4]	453K	. . . .	1.15M	0 0 0 0 0 0 0 0 0 0	2.75B
* * * * * * * * * *	195M	* * * * * * * * * *	1.93M	1] [2] [3] [4	453K	\ \ \ \ \ \ \ \ \ \	809K	= = = = = = = = = =	2.62B
= = = = = = = = = =	145M	° ° ° ° ° ° ° ° ° °	1.73M	[5] [6] [7] [	450K	< br / > < br / > < br	797K	, " resolved " : " https : / /	1.46B
/ / / / / / / / / /	79.3M	. . . .	1.56M	[6] [7] [8] [	448K	/ > < br / > < br / >	796K	" , " resolved " : " https : /	1.46B
. . / . . / . . / .	35.3M	- - - - - - - - - -	1.11M	] [6] [7] [8]	448K	br / > < br / > < br /	796K	" resolved " : " https : / / registry.npmjs.org	1.42B
. / . . / . . / . .	35.3M	[5] [6] [7] [	646K	5] [6] [7] [8	446K	> < br / > < br / > <	576K	resolved " : " https : / / registry.npmjs.org /	1.42B
/ . . / . . / . . /	35.2M	[1] [2] [3] [	645K	] [7] [8] [9]	446K	| Price : 1 Credit (USD $ 1)	437K	, , , , , , , , , ,	1B
# # # # # # # # # #	33M	[6] [7] [8] [	644K	6] [7] [8] [9	444K	vector | Price : 1 Credit (USD $ 1	437K	. tgz " , " integrity " : " sha512	938M

Yanai Elazar et al. What's In My Big Data? 2023.

Noisy Model Learning

- Noisy Label Learning
  - Data of downstream task contain noise
  - Noise hurts downstream performance
  - Improve the model performance when downstream contains noise
  - Many techniques, widely studied

- Noisy Model Learning (of foundation models)
  - Data of pre-training task contain noise
  - Data of downstream tasks are clean (or noisy)
  - **Does the pre-training noise affect the downstream generalization? If so, how?**
  - Unexplored before, perhaps intuitively believe the cleaner, the better

Motivation on Noisy Model Learning

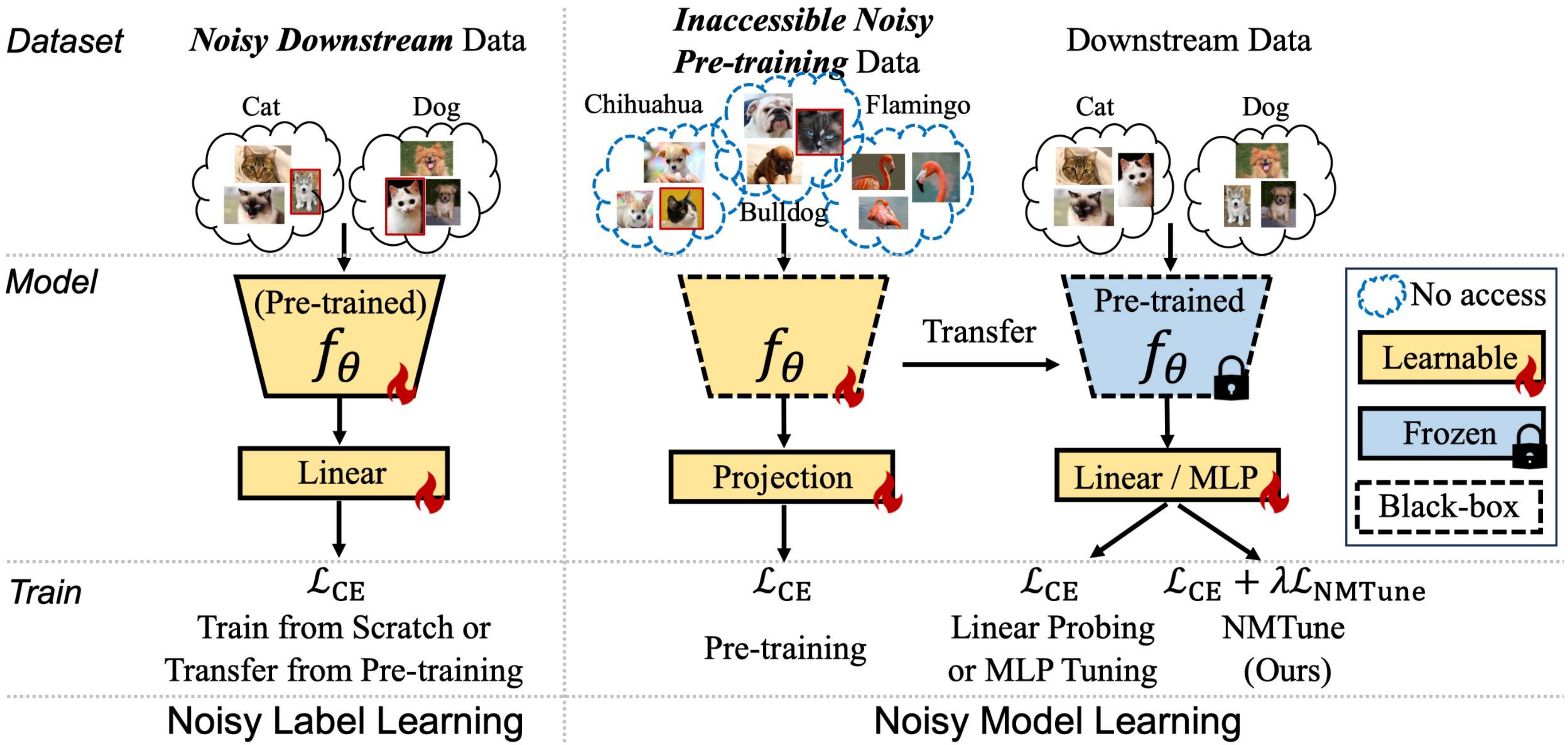
Noisy Model Learning (of foundation models)

- How does the noise in pre-training data affect the performance of pre-trained models on downstream tasks?
- How can we mitigate the detrimental effect of pre-training noise on downstream, if any?

- Possible black-box and noisy pre-training data
  - Massive size, expired urls…

- Possible (partially) black-box pre-trained models
  - Private models
  - Expensive computational requirement of full fine-tuning

Noisy Model vs. Noisy Label



Understanding the Effects of Pre-training Noise

Empirical Study

Effect of Pre-training Noise on Downstream

- Two pre-training paradigms/dataset
  - YFCC15M (and CC12M) – Image-Text Pair Contrastive Learning (CLIP)
  - ImageNet1K – Fully-Supervised Learning (FS)

- Introduce noise into the datasets
  - YFCC15M (and CC12M) – randomly swap the image-text pairs
  - ImageNet1K – randomly swap the label

- Two models pre-trained of different scales: ResNet-50 and ViT-B-16
  - for CLIP, ViT-B-16 is trained on YFCC15M+CC12M, and ResNet-50 on YFCC15M
  - for FS, both are trained on ImageNet-1K

- Train models with noise ratios {0, 5, 10, 20, 30}%
  - Heavy regularizations are adopted during pre-training

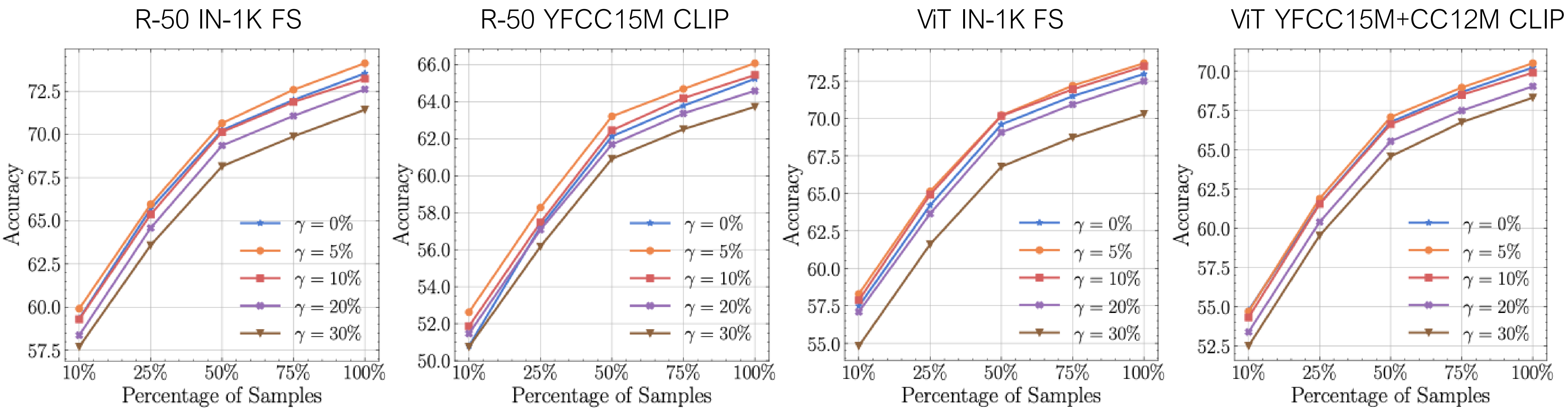
Downstream Classification Generalization

- In-Domain (ID) Evaluation
  - 14 vision datasets, including CIFAR-100, Flowers102, Food101, RESISC45, DTD, etc.
  - The training set and the testing set are of the same distribution

- Out-of-Domain (OOD) Evaluation
  - DomainNet: Clipart, Real, Sketch, Inpainting
  - ImageNet-Variants: IN-v2, IN-R, IN-Sketch, IN-A, IN-Vid, ObjectNet
  - The training set and the testing sets are of different distribution

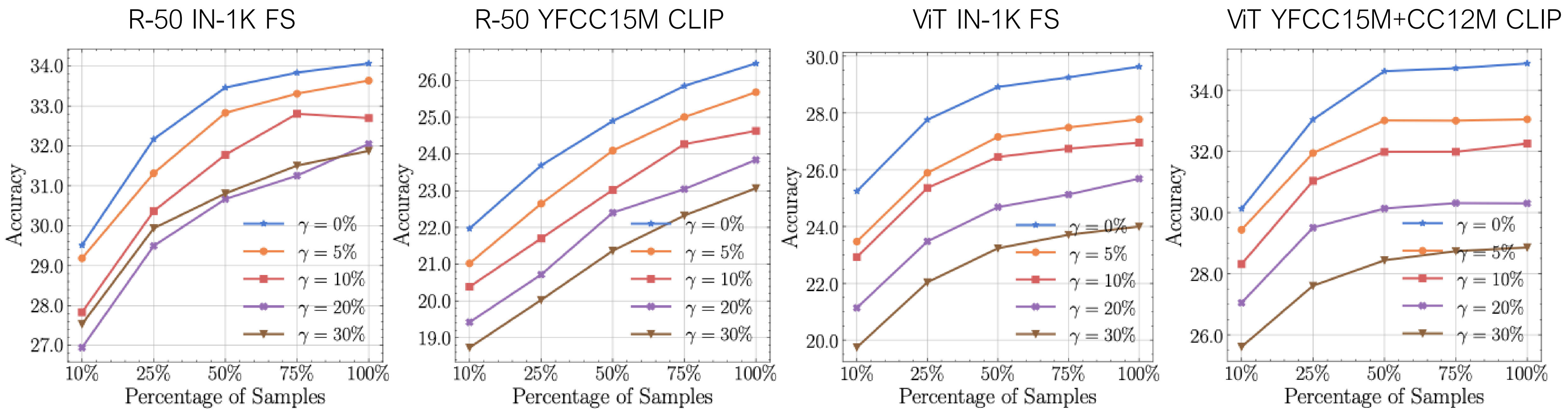
- Report average performance over all datasets with various tuning
  - Linear probing, LoRA (of ViT-B-16), full fine-tuning

ID Linear Probing Evaluation



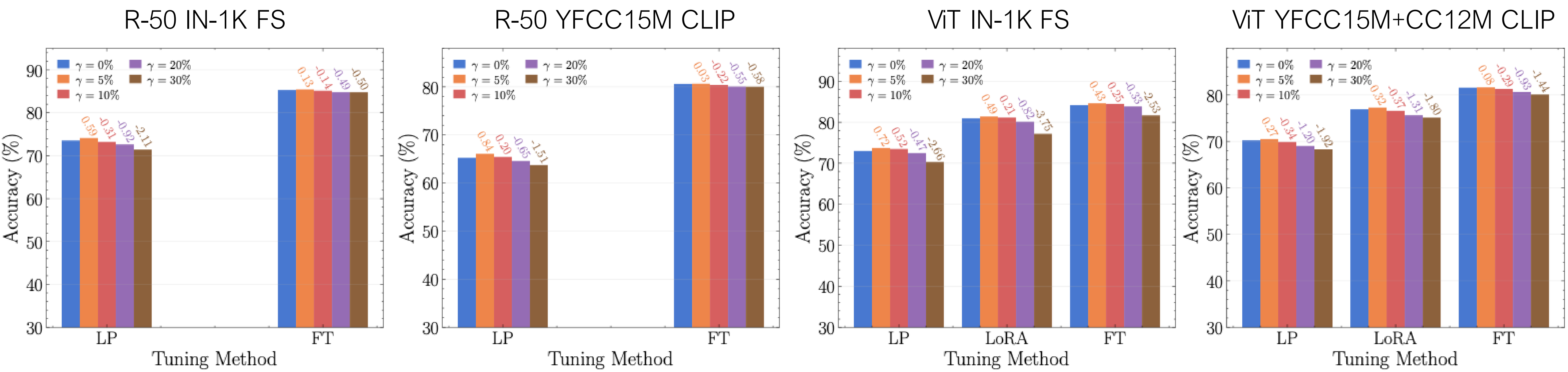
- Slight pre-training noise (5% or 10%) benefits ID classification tasks
- Further increase noise in pre-training hurts downstream performance

OOD Linear Probing Evaluation



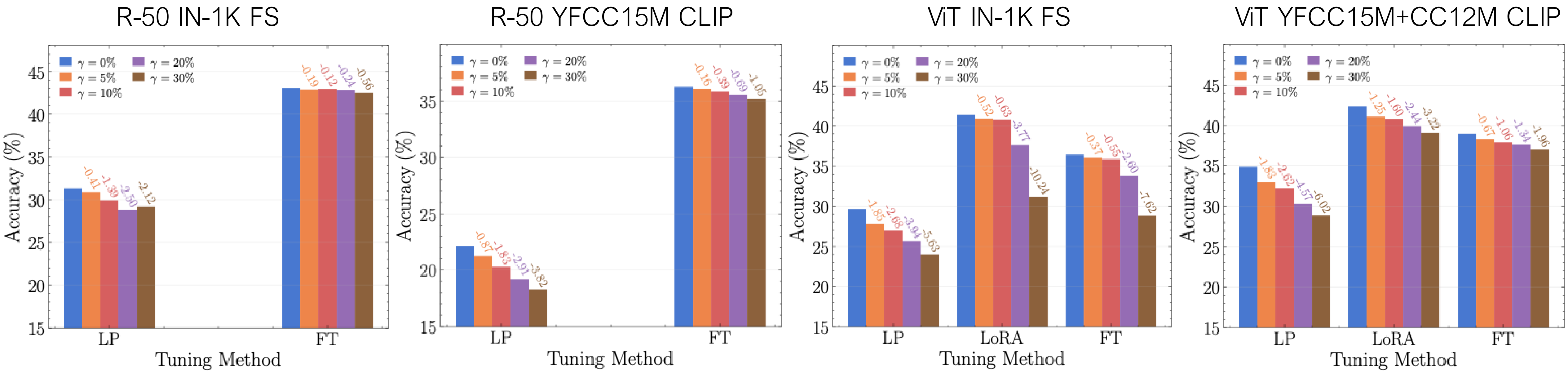
- Pre-training noise always deteriorates OOD tasks
- As noise ratio increases, the performance consistently decreases

ID Eval. with Different Tuning Methods



- Different tuning methods on ID tasks present similar trends
  - up to 5% or 10% can benefit ID performance
- Differences between clean and noisy models become smaller
  - with more pre-trained parameters modified at downstream tasks

OOD Eval. with Different Tuning Methods



- Different tuning methods on OOD tasks present similar trends
  - pre-train noise consistently hurts the performance
- Differences between clean and noisy models become smaller
  - with more pre-trained parameters modified at downstream tasks

Detection and Segmentation Tasks

TABLE 4: Object detection results on COCO 2017 of IN-1K ResNet-50 noisy FS pre-trained models.

Detection	Noise (%)	AP^{box}	AP^{box}_{50}	AP^{box}_{75}
Faster R-CNN [19]	0	38.5	59.8	41.7
	5	**38.6**	**60.1**	**41.9**
	10	**38.6**	60.0	**41.9**
	20	38.4	59.7	41.6
	30	37.9	59.1	40.9
RetinaNet [20]	0	38.3	58.2	40.9
	5	**38.4**	**58.4**	40.9
	10	**38.4**	58.1	**41.1**
	20	37.9	57.7	40.4
	30	37.0	56.8	39.1

TABLE 5: Instance segmentation results on COCO 2017 of IN-1K ResNet-50 noisy FS pre-trained models.

Detection	Noise (%)	AP^{mask}	AP^{mask}_{50}	AP^{mask}_{75}
Mask R-CNN [21]	0	31.3	**51.3**	33.0
	5	**31.4**	**51.3**	**33.2**
	10	**31.3**	**51.3**	32.9
	20	31.2	51.1	32.8
	30	30.30	49.9	32.1
SOLOv2 [140]	0	32.2	52.7	33.6
	5	**32.7**	**53.2**	**34.2**
	10	32.4	52.8	33.9
	20	32.0	52.2	33.6
	30	31.4	51.3	32.5

- Evaluate IN-1K noisy pre-trained on COCO Detection and Segmentation
- Slight pre-training noise can also benefit other downstream tasks than classification

Feature Space Analysis

Empirical Study

Singular Values Analysis

- Where do the superior ID performance (with slight noise) and the inferior OOD performance stem from?

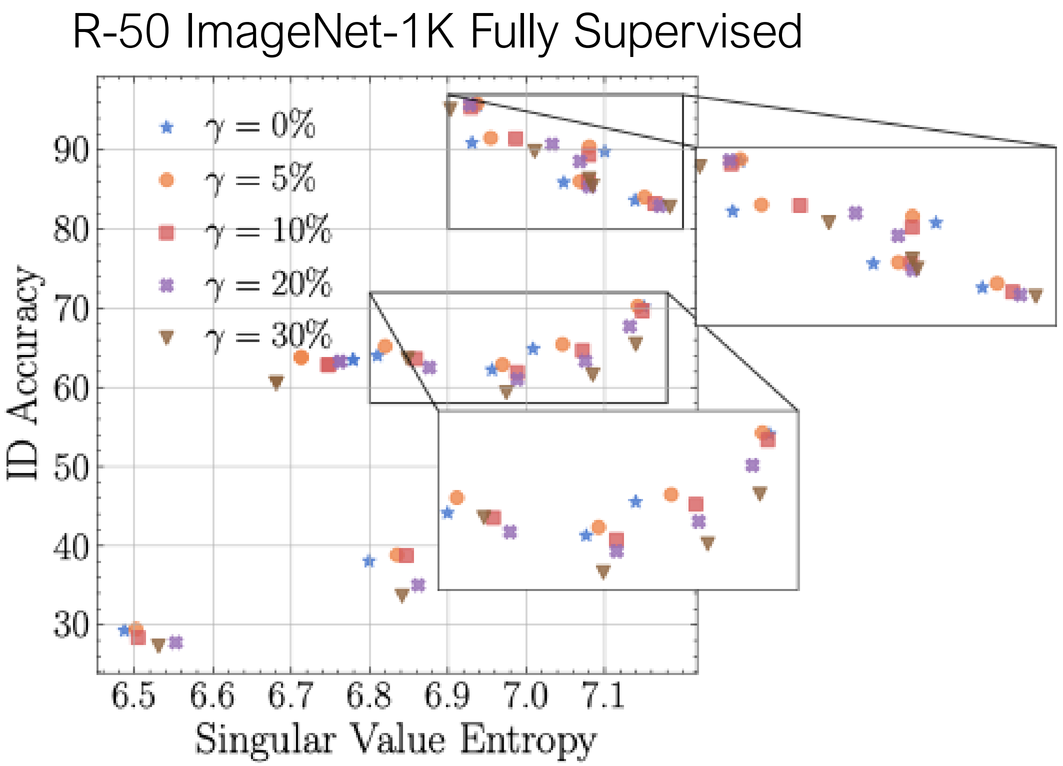
- We conduct SVD on features of pre-trained models on downstream tasks
  - Singular Value Entropy (SVE): measures the flatness of singular value distribution

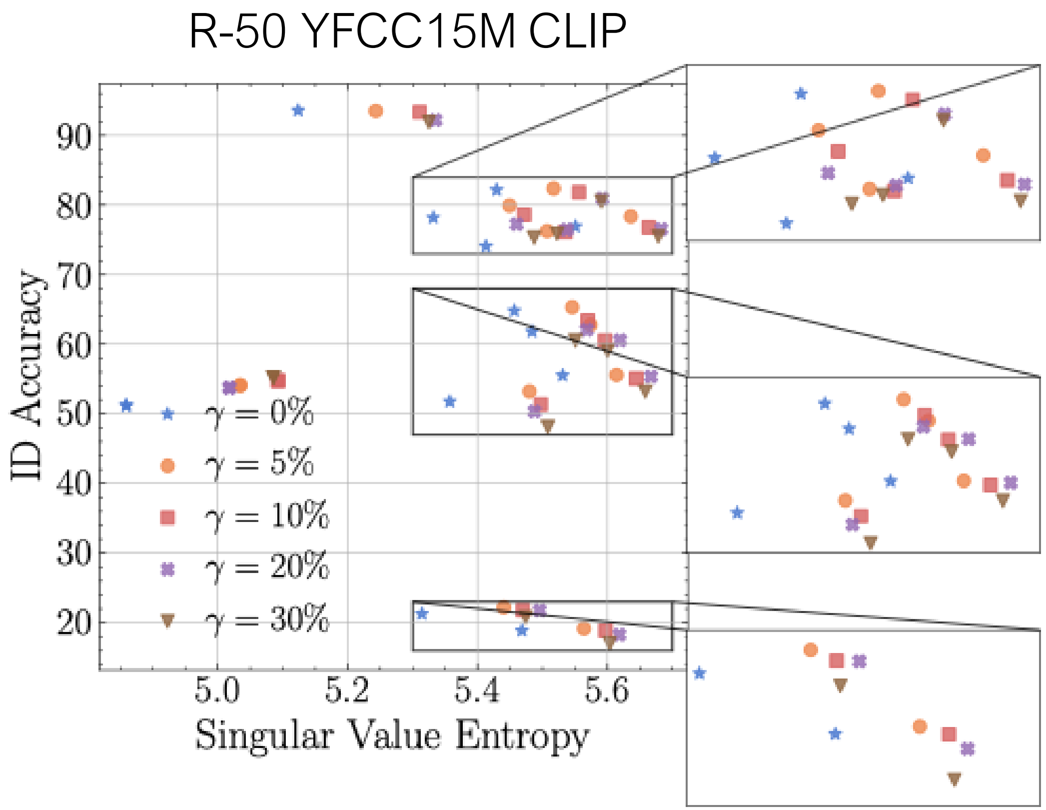
$$\mathrm{SVE} = -\sum_{i=1}^{D} \frac{\sigma_i}{\sum_{j=1}^{D} \sigma_j} \log \frac{\sigma_i}{\sum_{j=1}^{D} \sigma_j}$$

  - Largest Singular Value Ratio (LSVR): measures the ratio of the largest singular value

$$\mathrm{LSVR} = -\log \frac{\sigma_1}{\sum_{i=1}^{D} \sigma_i}$$

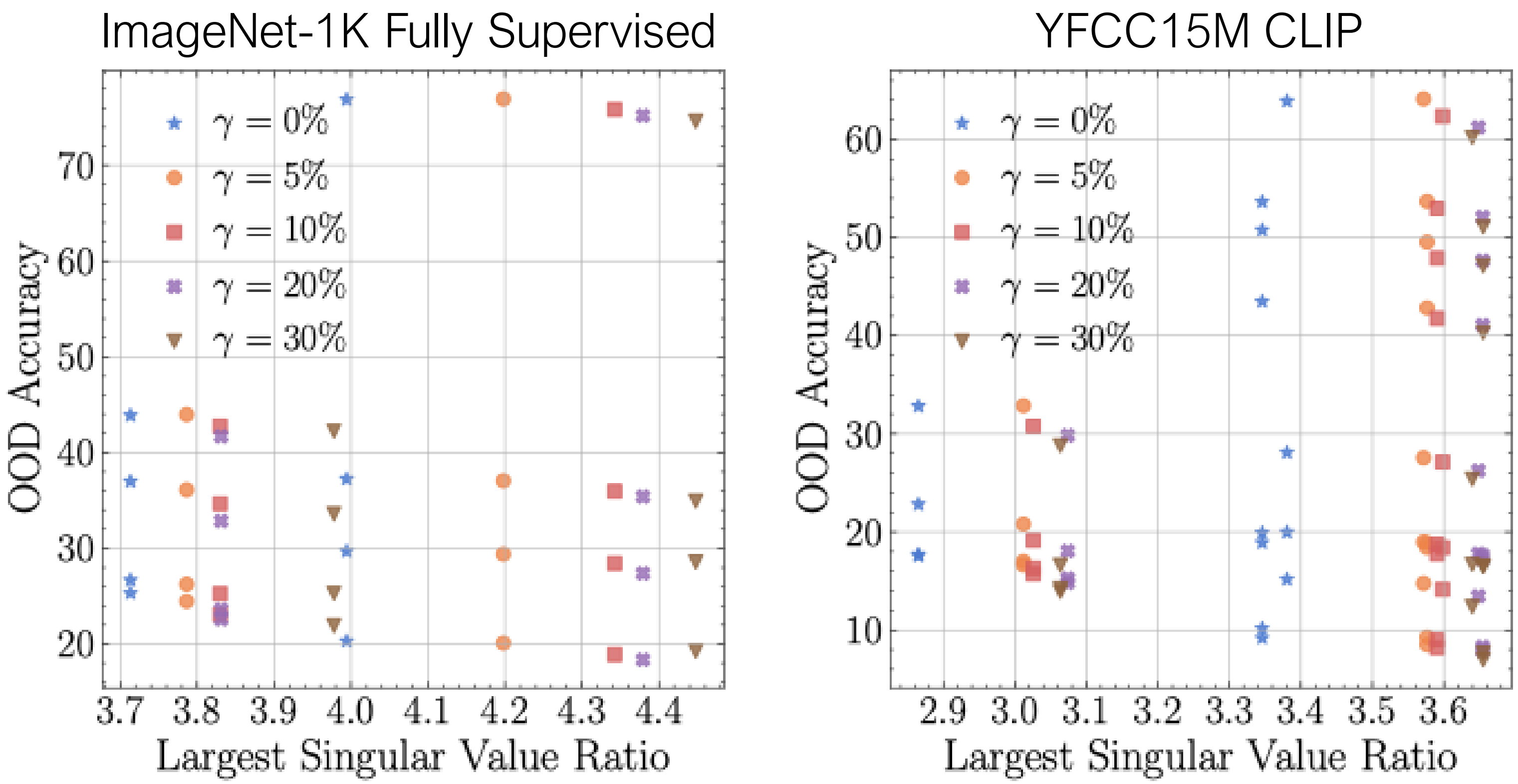
ID – Singular Value Entropy





- SVE and ID accuracy first increases then decreases, as the noise ratio increases
- Slight pre-training noise encourages the model to use more capacity to fit the noise
- A higher dimension of feature space, better-initialized features at the downstream
- Noise further increases, more dimensions fitting the noise, less useful features at downstream

OOD – Largest Singular Value Ratio



- LSVR consistently increases and OOD consistently decreases, as the noise ratio increases
- More capacity in feature space is used for fitting noise, and less transferable/dominant singular vectors are learned during pre-training

Mitigating the Noise on Downstream

- We propose a black-box fine-tuning method
  - with an MLP projection head and a linear classification layer
  - MLP is used for affine transformation of pre-trained features F to get Z

- NMTune defines 3 regularization terms during black-box fine-tuning
  - encouraging consistency between pre-trained features and MLP-transformed features

$$\mathcal{L}_{\text{MSE}} = \left\| \frac{\mathbf{F}}{\|\mathbf{F}\|_2} - \frac{\mathbf{Z}}{\|\mathbf{Z}\|_2} \right\|_2^2$$

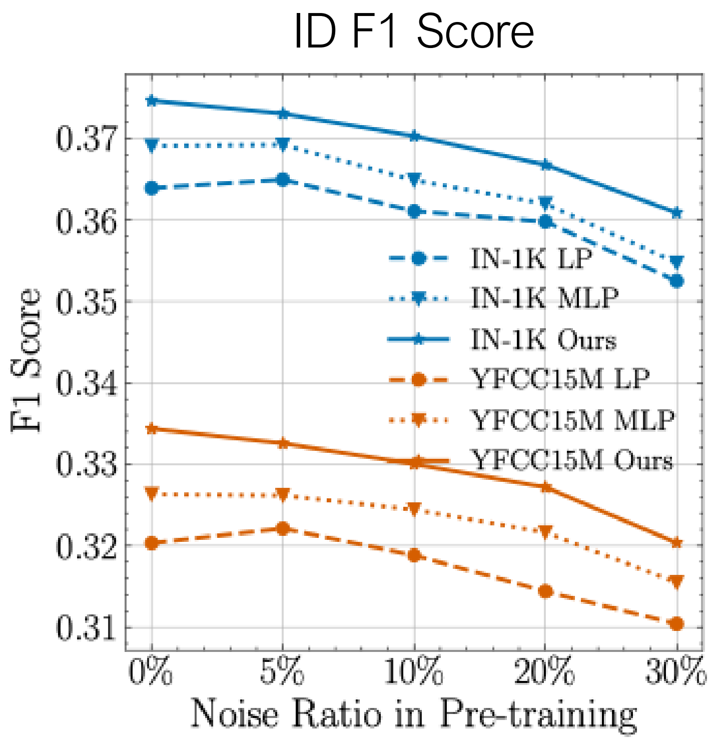
  - minimizing the covariance matrix of MLP-transformed features

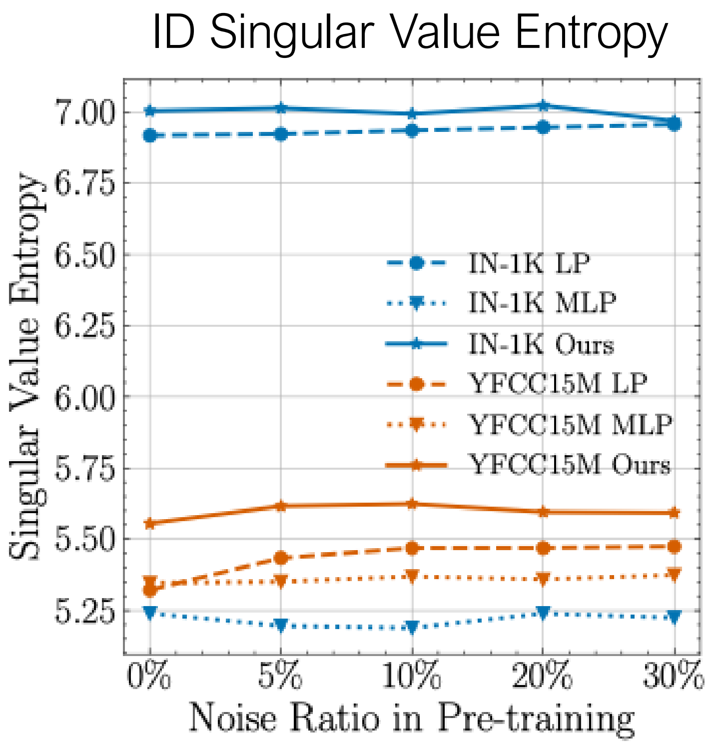
$$\mathcal{L}_{\text{COV}} = \frac{1}{D} \sum_{i \neq j} [C(\mathbf{Z})]_{i,j}^2$$

  - maximizing the largest singular value ratio of MLP-transformed features

$$\mathcal{L}_{\text{SVD}} = -\frac{\sigma_1}{\sum_j^D \sigma_{j=1}}$$

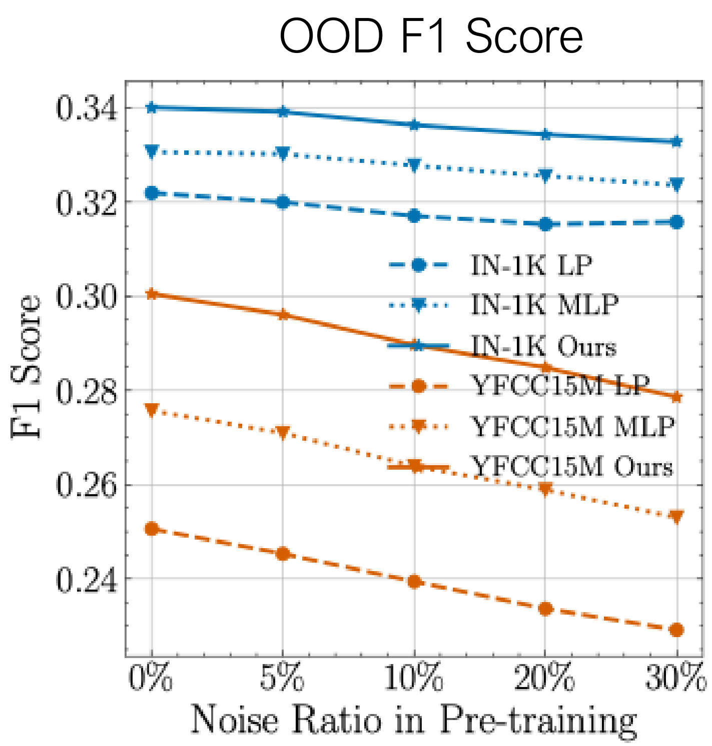
NMTune for ID tasks

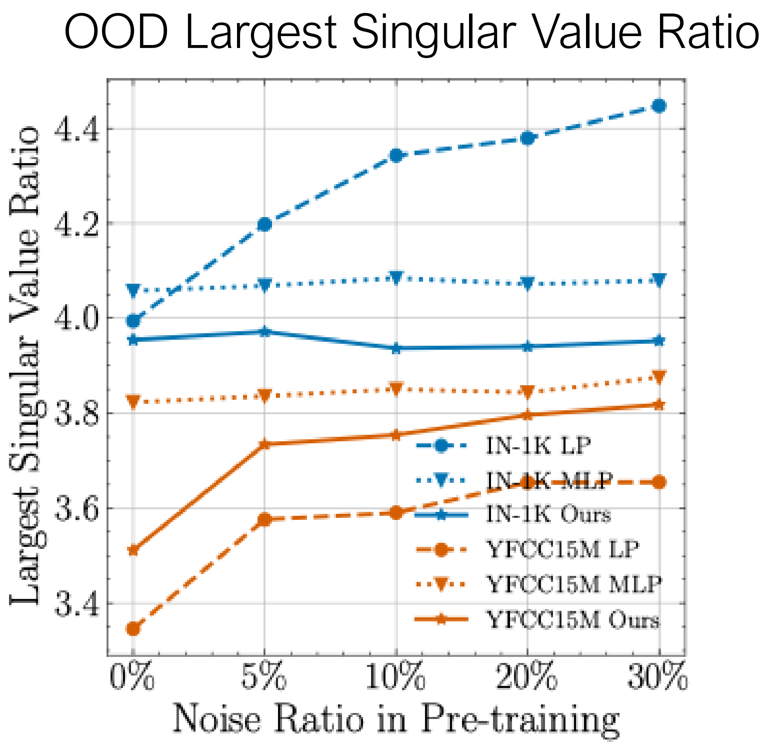




- Our method helps improve F1 score and SVE for ID tasks for both noisy ImageNet-1K and YFCC15M pre-trained models
- Adding MLP only helps with F1 but produces lower SVE

NMTune for OOD tasks

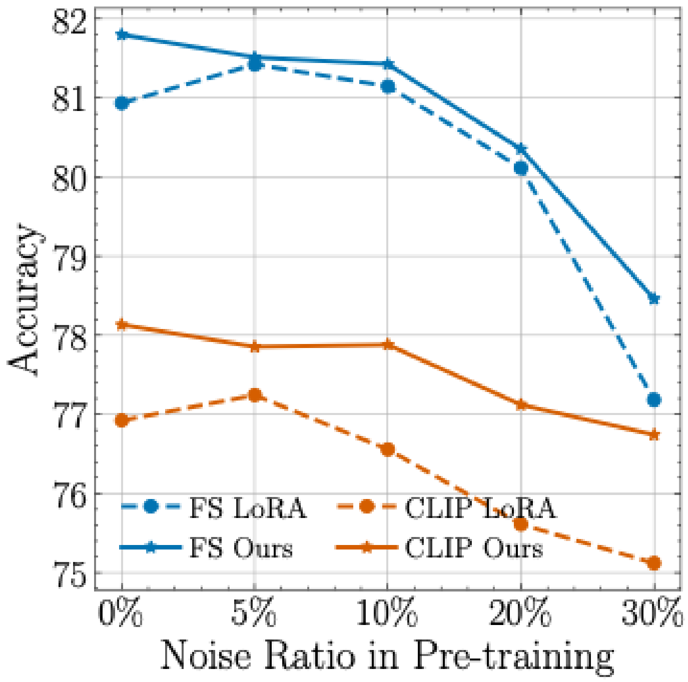




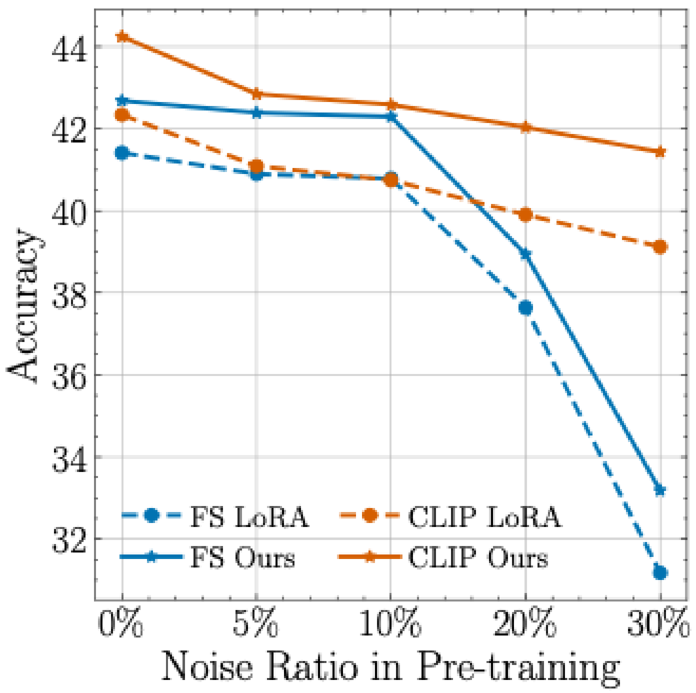
- Our method helps improve F1 score for OOD tasks
- Our method produces more consistent LSVR across noise ratios (MLP also does)

NMTune for LoRA

ViT ID Acc



ViT OOD Acc



- NMTune also can be applied with LoRA to mitigate the pre-training noise

Practical Large Models

- Vision Models
  - JFT300M Semi-Supervised Pre-trained EfficientNet-B3
  - ImageNet-21K Fully-Supervised Pre-trained ResNetv2-152x2
  - ImageNet-21K Fully-Supervised Pre-trained Swin-L
  - Laion-2B CLIP Pre-trained ConvNext-L
  - Laion-2B CLIP Pre-trained ViT-L
  - ID: 14 datasets, OOD: DomainNet

- Language Models
  - BERT-L, RoBERTa-L, GPT-2, text-ada-002 embedding API
  - ID: GLUE, OOD: GLUE-X

Practical Large Models

Table 1: Results on popular vision models that are pre-trained on noisy datasets. We use 14 in-domain (ID) and 4 out-of-domain (OOD) tasks.

Pre-trained Model	Tuning Method	In-Domain		Out-of-Domain	
		Acc.	F1	Acc.	F1
JFT300M	LP	76.72	0.3815	44.13	0.3594
Semi-Supervised	MLP	76.87	0.3833	45.95	0.3624
EfficientNet-B3	Ours	**77.63**	**0.3874**	**46.84**	**0.3654**
ImageNet-21K	LP	77.51	0.3718	40.82	0.3062
Fully Supervised	MLP	77.58	0.3726	41.73	0.3053
ResNetv2-152x2	Ours	**78.43**	**0.3862**	**42.42**	**0.3100**
ImageNet-21K	LP	81.91	0.4092	50.88	0.3838
Fully Supervised	MLP	82.51	0.4128	51.21	0.3811
Swin-L	Ours	**84.16**	**0.4177**	**52.35**	**0.3901**
Laion-2B	LP	88.86	0.4432	66.86	0.4253
CLIP	MLP	88.53	0.4417	68.43	0.4304
ConvNext-L	Ours	**89.48**	**0.4457**	**70.30**	**0.4367**
Laion-2B	LP	86.85	0.4328	66.89	0.4208
CLIP	MLP	87.23	0.4375	69.50	0.4221
ViT-L	Ours	**88.57**	**0.4414**	**70.47**	**0.4246**

Table 2: Evaluation of our method on language models in practice that are pre-trained on noisy datasets. We use GLUE for in-domain (ID) tasks and GLUE-X for out-of-domain (OOD) tasks.

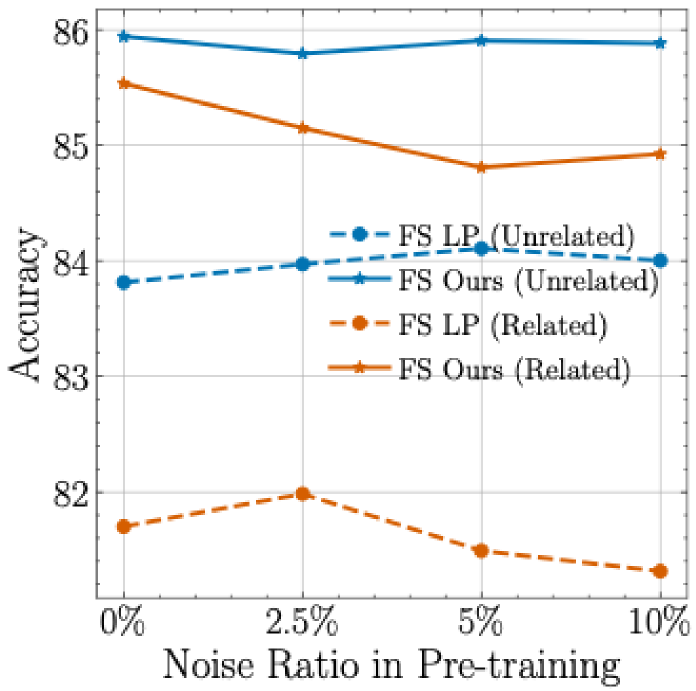
Pre-trained Model	Tuning Method	In-Domain	Out-of-Domain
BERT-L	LP	69.44	50.65
	MLP	69.78	50.62
	Ours	**70.26**	**51.63**
RoBERTa-L	LP	69.75	44.55
	MLP	70.27	45.22
	Ours	**70.97**	**47.01**
GPT-2	LP	58.67	36.68
	MLP	58.44	37.24
	Ours	**59.34**	**39.07**
text-ada-002	LP	56.96	44.06
	MLP	63.89	51.30
	Ours	**65.99**	**53.48**

Asymmetric Pre-training Noise

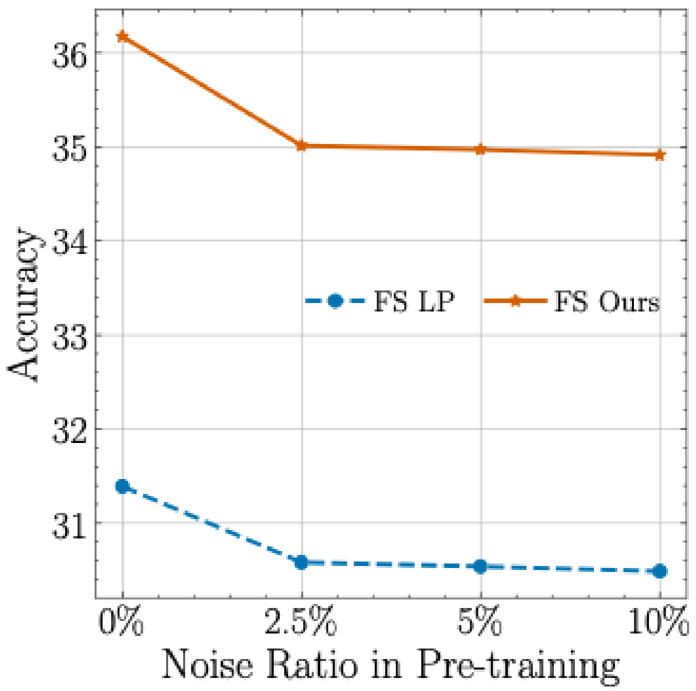
- Previous experiments mainly involve random pre-training noise
  - noise can exist in all classes/concepts uniformly
- We also study asymmetric noise in ImageNet-1K
  - find overlapped classes in IN-1K with CIFAR-100 using wordnet
  - introduce noise only within these overlapped classes
- Downstream linear probing evaluation:
  - noise-related ID: CIFAR-10, CIFAR-100
  - noise-unrelated ID: Food-101, Caltech101, EuroSAT
  - OOD: DomainNet

Asymmetric Pre-training Noise

R-50 ID Acc

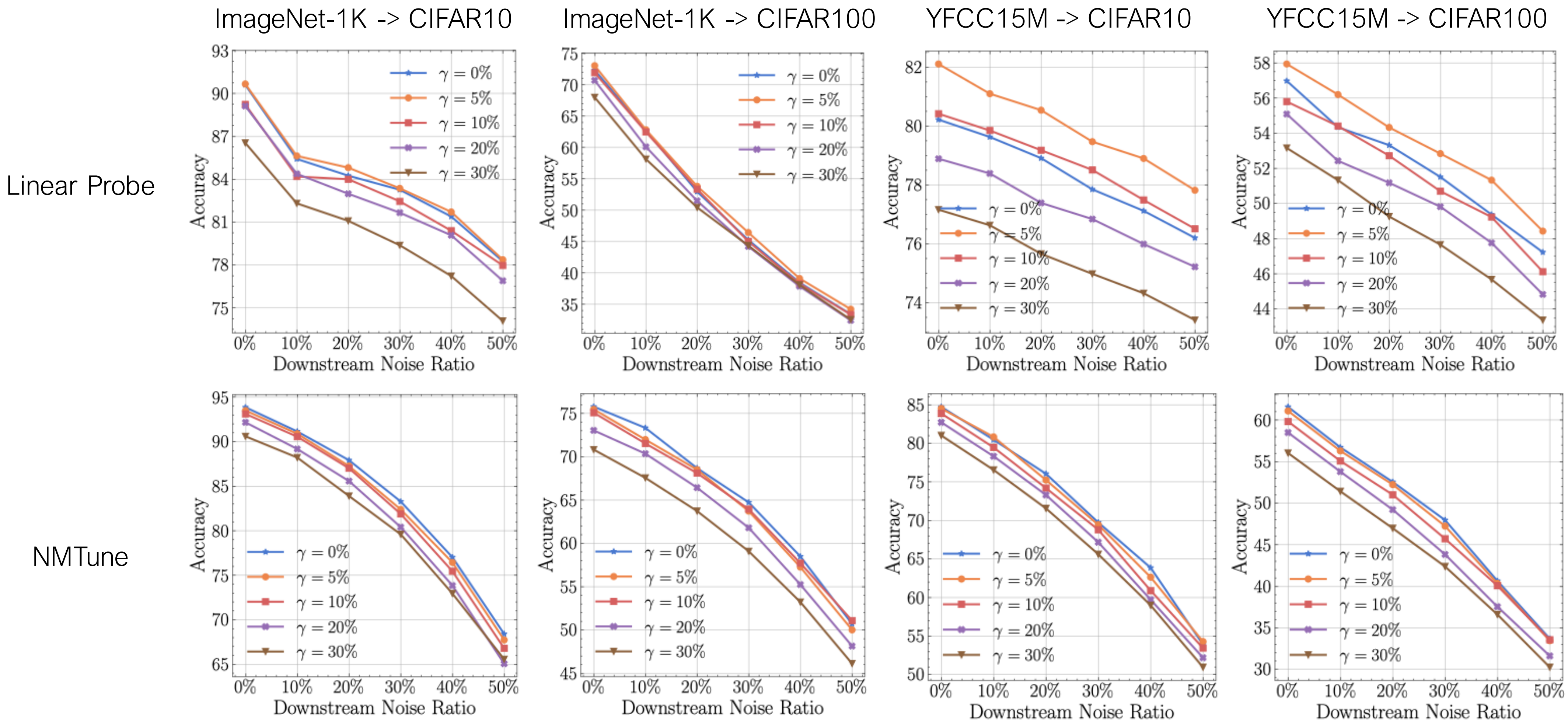


R-50 OOD Acc



- Previous observations still manifest on asymmetric pre-training noise

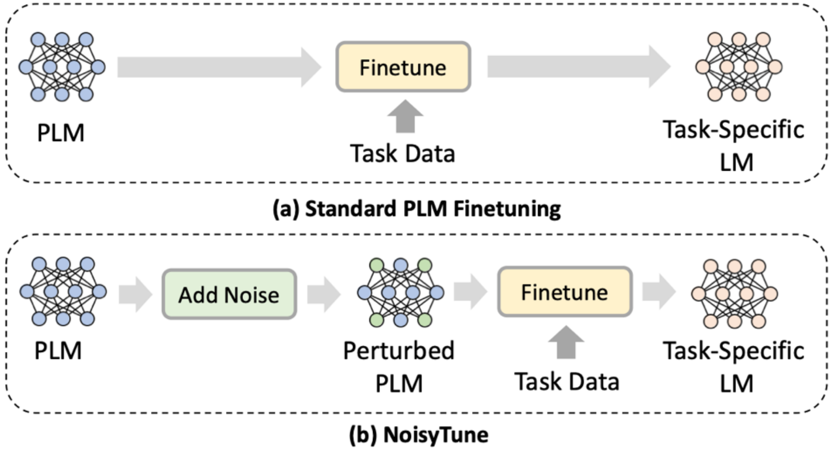
Combining with Noisy Label Learning



- Similar observation holds on NLL and NMTune also helps

Related Works on Pre-training Noise/Data

- NoisyTune



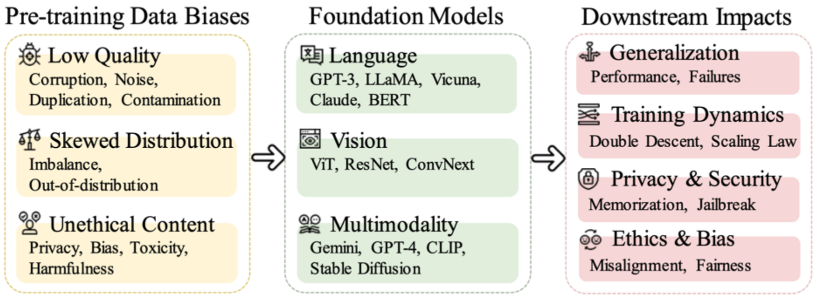
- NEFTune

Algorithm 1 NEFTune: **N**oisy **E**mbedding Instruction **F**ine**tun**ing

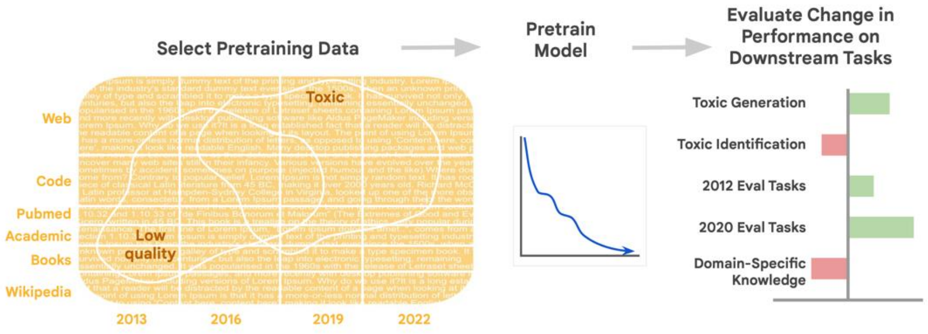
Input: $\mathcal{D} = \{x_i, y_i\}_1^N$ tokenized dataset, embedding layer emb($\cdot$), rest of model $f_{/\text{emb}}(\cdot)$, model parameters θ, loss($\cdot$), optimizer opt($\cdot$)
NEFT Hyperparameter: base noise scale $\alpha \in \mathbb{R}^+$
Initialize θ from a pretrained model.
repeat $(X_i, Y_i) \sim \mathcal{D}$ ▷ sample a minibatch of data and labels
  $X_{\text{emb}} \leftarrow \text{emb}(X_i), \mathbb{R}^{B\times L\times d}$ ▷ batch size B, seq. length L, embedding dimension d
  $\epsilon \sim \text{Uniform}(-1, 1), \mathbb{R}^{B\times L\times d}$ ▷ sample a noise vector
  $X'_{\text{emb}} \leftarrow X_{\text{emb}} + (\frac{\alpha}{\sqrt{Ld}})\epsilon$ ▷ add scaled noise to embeds [a]
  $\hat{Y}_i \leftarrow f_{/\text{emb}}(X'_{\text{emb}})$ ▷ make prediction at noised embeddings
  $\theta \leftarrow \text{opt}(\theta, \text{loss}(\hat{Y}_i, Y_i))$ ▷ train step, e.g., grad descent
until Stopping criteria met/max iterations.

[a] If sequence lengths in a batch are not equivalent, then L is a vector $\in \mathbb{Z}^B_{>0}$ and the scaling factor $(\alpha/\sqrt{Ld})$ is computed independently for each sequence in batch.

- Catastrophic Inheritance



- Pre-trainer's Guide to LLM training data



Chuhan Wu, et al. NoisyTune: A Little Noise Can Help You Finetune Pretrained Language Models Better.
Neel Jain, et al. NEFTUNE: Noisy Embedding Improve Instruction Fine-Tuning.
Hao Chen et al. On catastrophic inheritance of large foundation models.
Shayne Longpre et al. A pre-trainer's guide to training data.

Conclusion

- We propose Noisy Model Learning
  - A novel research topic for studying and mitigating the pre-training noise
- We found:
  - Slight noise in pre-training benefits ID tasks, agnostic to model architectures, pre-training proxy objectives, pre-training noise types, downstream tuning methods, and downstream applications
  - However, pre-training noise always hurts OOD tasks
  - Malicious effects of pre-training noise can be mitigated at downstream tasks through NMTune
- Future work includes other pre-training paradigms and other types of pre-training biases

Thanks

Hao Chen, haoc3@andrew.cmu.edu

https://arxiv.org/pdf/2309.17002.pdf

Carnegie Mellon University

Microsoft Research 微软亚洲研究院

RIKEN

MOHAMED BIN ZAYED UNIVERSITY OF ARTIFICIAL INTELLIGENCE

SUSTech Southern University of Science and Technology