Understanding and Mitigating the Label Noise in Pre-training on Downstream Tasks

Hao Chen[1,2]*, **Jindong Wang**[2]†, **Ankit Shah**[1], **Ran Tao**[1],
Hongxin Wei[3], **Xing Xie**[2], **Masashi Sugiyama**[4,5], **Bhiksha Raj**[1,6]

[1]Carnegie Mellon University, [2]Microsoft Research Asia, [3]SusTech,
[4]RIKEN AIP, [5]The University of Tokyo, [6]Mohamed bin Zayed University of AI

ICLR 2024（Spotlight）

Background

- Applying Foundation models on **downstream** tasks requires **pre-training** on very **large amounts of data**, and then fine-tuning on downstream tasks.
- These pre-training data are generally **collected from networks** such as Laion-2B and Common Crawl, so it is inevitable that some **Bias/Noise** data is introduced in the pre-training.

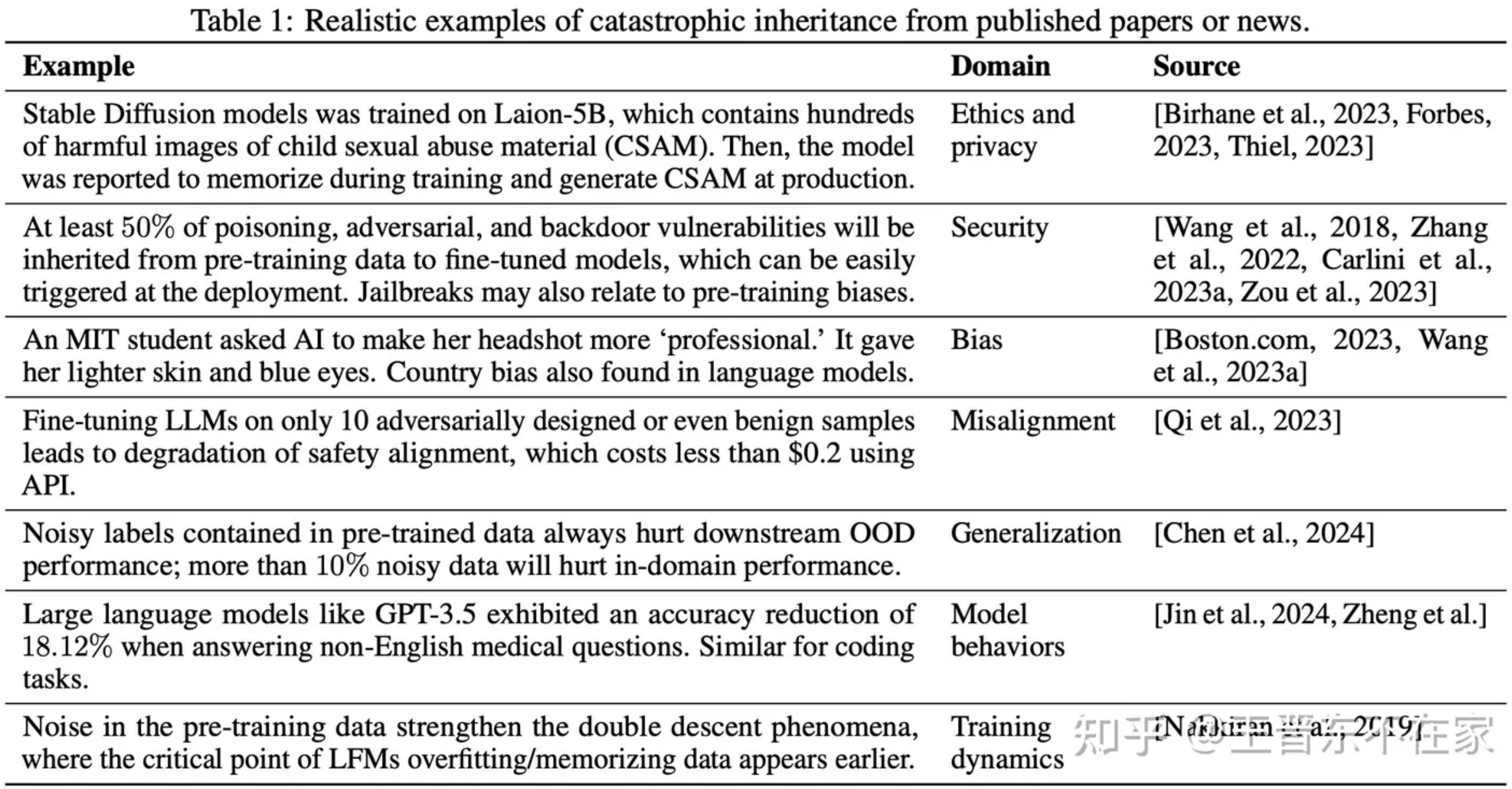
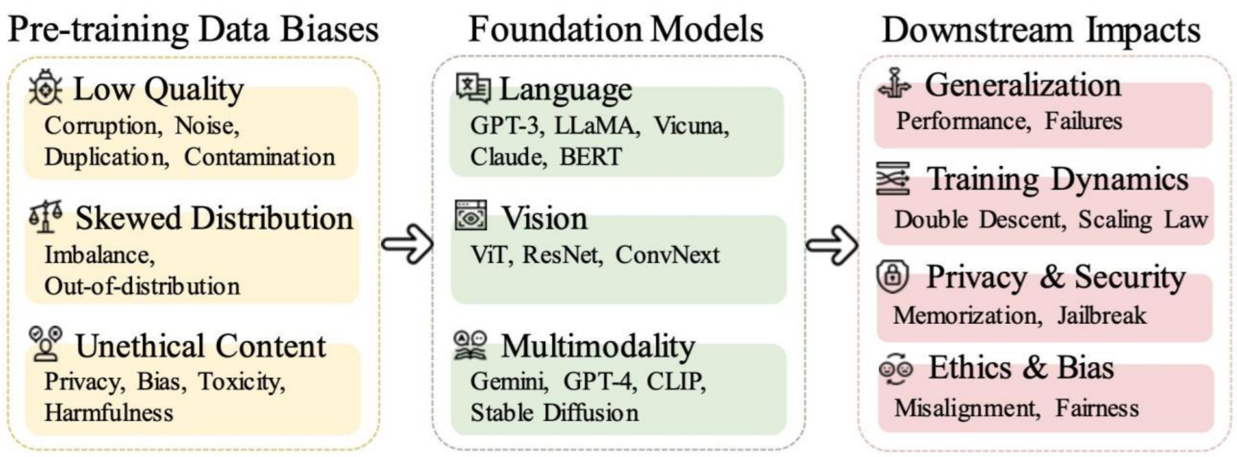


Table 1: Realistic examples of catastrophic inheritance from published papers or news.

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]

Background

The bias/noise in these pre-trained datasets may cause some unpredictable and unavoidable impacts on the Foundation model in downstream tasks.



Pre-training Data Biases

- Low Quality: Corruption, Noise, Duplication, Contamination
- Skewed Distribution: Imbalance, Out-of-distribution
- Unethical Content: Privacy, Bias, Toxicity, Harmfulness

Foundation Models

- Language: GPT-3, LLaMA, Vicuna, Claude, BERT
- Vision: ViT, ResNet, ConvNext
- Multimodality: Gemini, GPT-4, CLIP, Stable Diffusion

Downstream Impacts

- Generalization: Performance, Failures
- Training Dynamics: Double Descent, Scaling Law
- Privacy & Security: Memorization, Jailbreak
- Ethics & Bias: Misalignment, Fairness

Why is this issue so difficult?

- The underlying models are either **not open-source**
- Too **big** to be trained from scratch.
- Most of the **pre-training data** is also **not open-source**
- Results in **black-boxed** models of the data

However, the specific manifestations and reasons for these different impacts in downstream tasks are not clear, yet they are very important for the safe application of these models. We refer to this novel research direction as **Catastrophic Inheritance**.

Introduction

Noisy Model Learning

Noisy Label Learning vs Noisy Model Learning

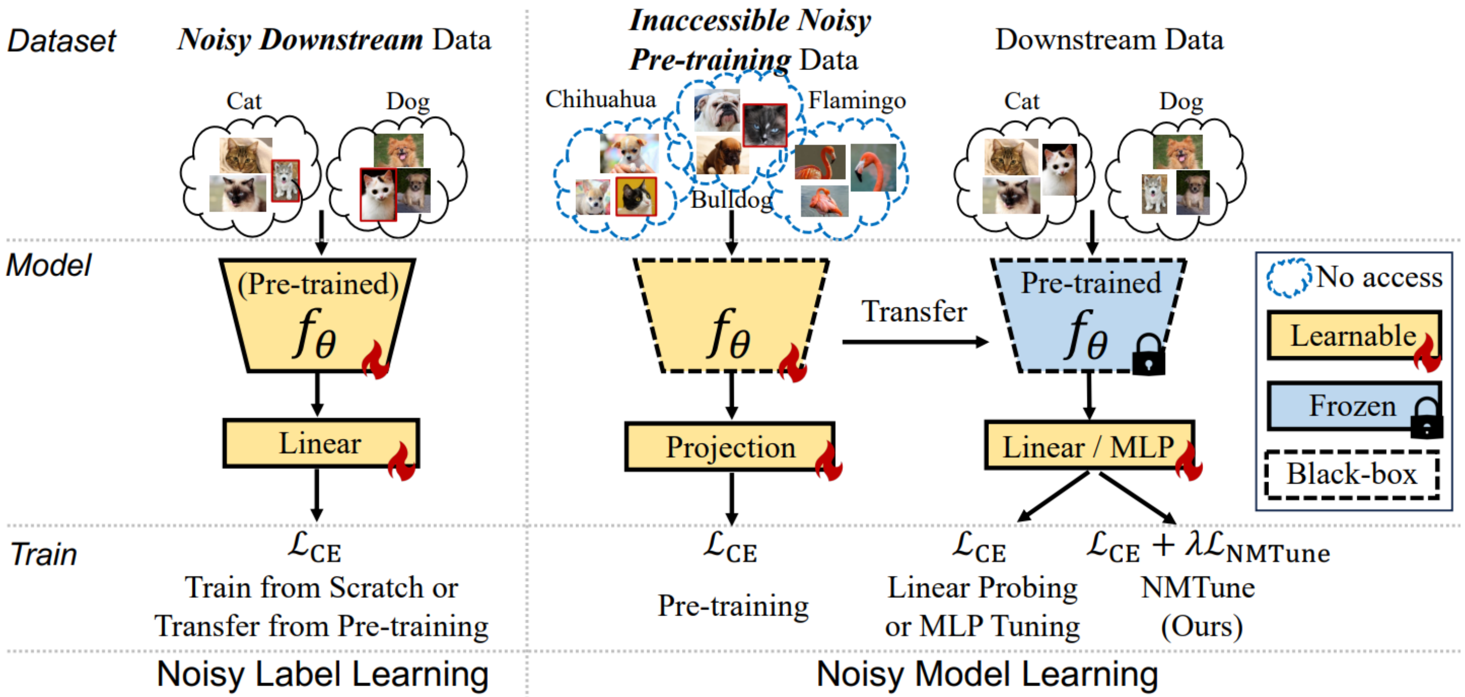


Figure 4: Illustration of noisy label learning (left) and the proposed *Noisy Model Learning* (right). Noisy label learning mainly focuses on robustly training a model from scratch or fine-tuning a model from pre-training on a noisy dataset. Noisy model learning focuses on robustly adapting the black-box noisy pre-trained models to downstream datasets with no assumption on the downstream dataset.

Motivation

Effects of Pre-training Noise

- Pre-training: ResNet-50 and ViT-B-16 for Fully-Supervised (FS) and Image-Text Contrastive (CLIP) Learning

- Pre-training data: For Fully-Supervised (FS) Learning, we use ImageNet-1K as pre-training data and randomly contaminated labels as noise. For CLIP, we use YFCC15M+CC12M as the pre-training data and randomly swap the text in two image-text pairs as noise. For these two data and two pre-training methods, we introduce different degrees of noise 0%, 5%, 10%, 15%, 20%, and 30%.

- Downstream task evaluation: we use two evaluation modalities In-Domain (ID) and Out-of-Domain (OOD). For both evaluations, we use Linear Probing for the pre-trained models. ID uses the 14-sum dataset, and OOD uses DomainNet and ImageNet variants. In addition to Linear Probing, in our recent extensions, we also use ViT-B-16+LoRA and full fine-tuning to study the impact of different tuning approaches.

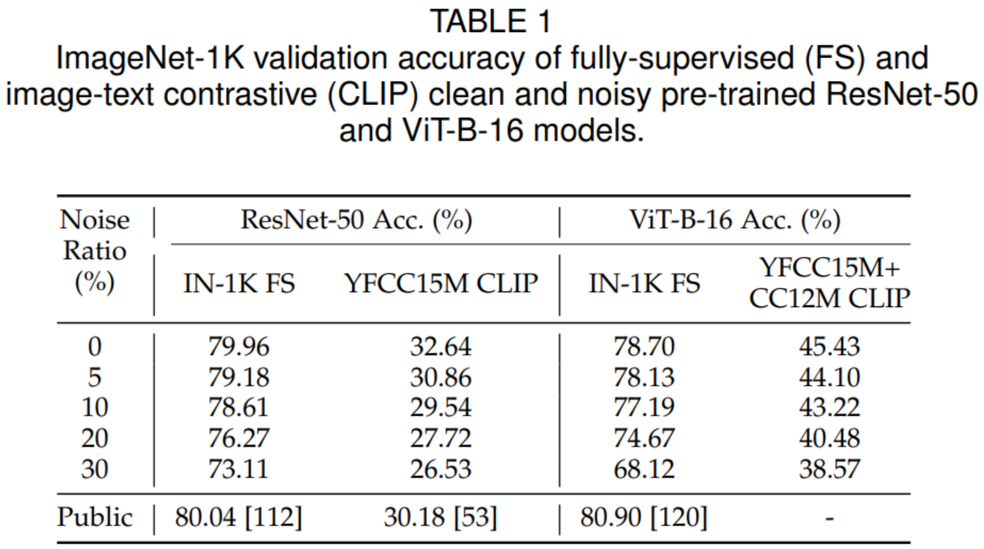


TABLE 1
ImageNet-1K validation accuracy of fully-supervised (FS) and image-text contrastive (CLIP) clean and noisy pre-trained ResNet-50 and ViT-B-16 models.

Noise Ratio (%)	ResNet-50 Acc. (%)		ViT-B-16 Acc. (%)	
	IN-1K FS	YFCC15M CLIP	IN-1K FS	YFCC15M+ CC12M CLIP
0	79.96	32.64	78.70	45.43
5	79.18	30.86	78.13	44.10
10	78.61	29.54	77.19	43.22
20	76.27	27.72	74.67	40.48
30	73.11	26.53	68.12	38.57
Public	80.04 [112]	30.18 [53]	80.90 [120]	-

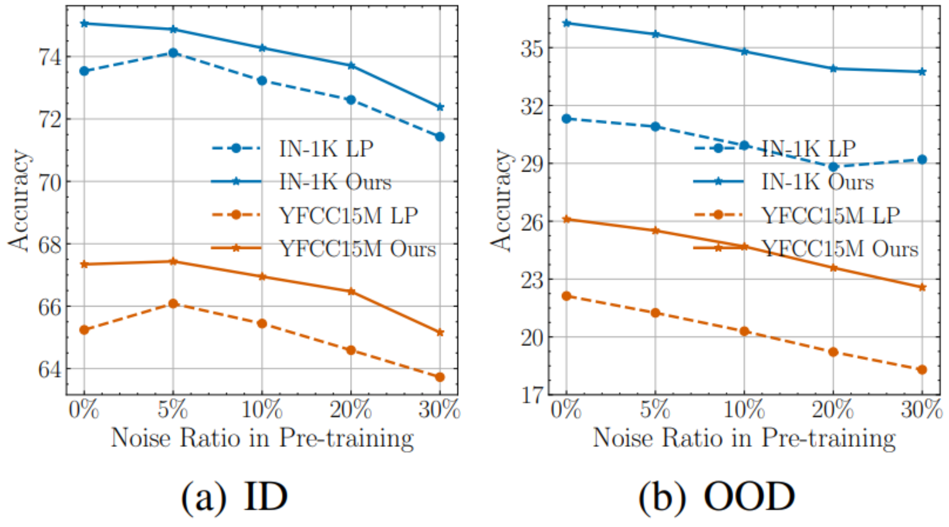
Motivation

Effects of Pre-training Noise in the Downstream Classification Task Linear Probing

- There are some interesting findings from the results of noisy pre-trained models doing linear probing on downstream classification tasks:

- For the ID task, slight noise (up to 5% or 10%) in the pre-training leads to better resultsFor the OOD task, noise in pre-training leads to monotonically decreasing results.

- The results on the ID task are contrary to traditional Noisy Label Learning perceptions and our intuition. Therefore we did more experiments to verify it.



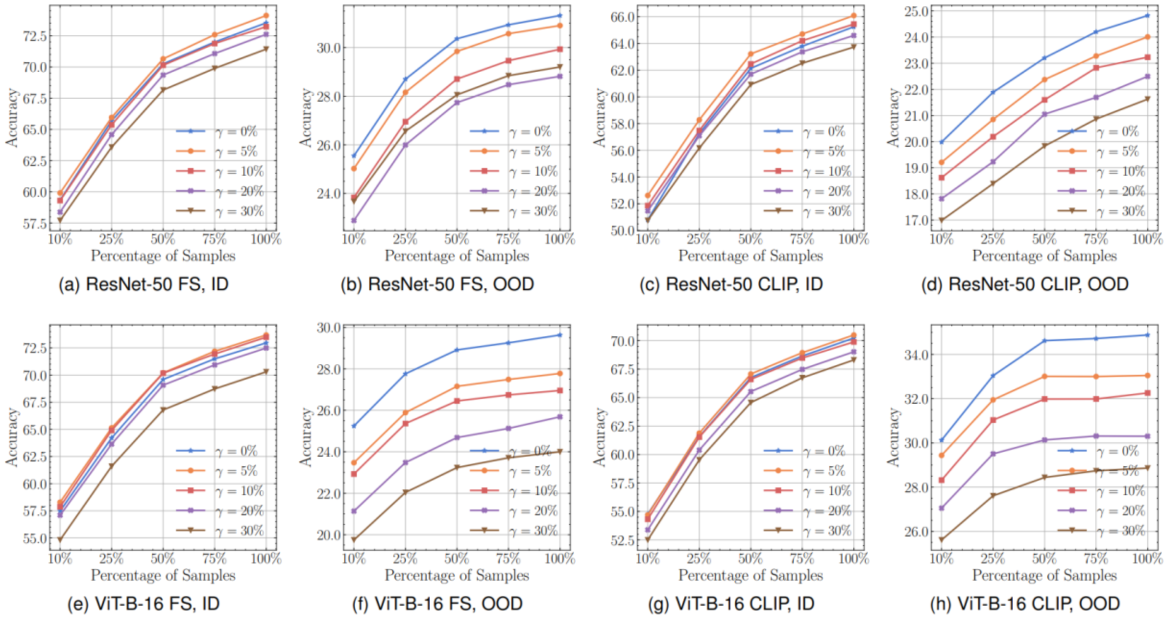
(a) ID

(b) OOD

Figure 1: In-domain (ID) and out-of-domain (OOD) downstream performance when supervised pre-training the model on synthetic noisy ImageNet-1K (IN-1K) and YFCC15M of various noise ratios. We compare linear probing (LP) and the proposed method on 14 ID and 4 OOD tasks. On ID, 5% noise in pre-training benefits the LP performance. Our method not only boosts the general performance but also rectifies the model pre-trained on clean data to be comparable to 5% noise. On OOD, noise in pre-training is detrimental to robustness performance when conducting LP. Our method improves the transferability on OOD tasks significantly compared to LP.

Motivation

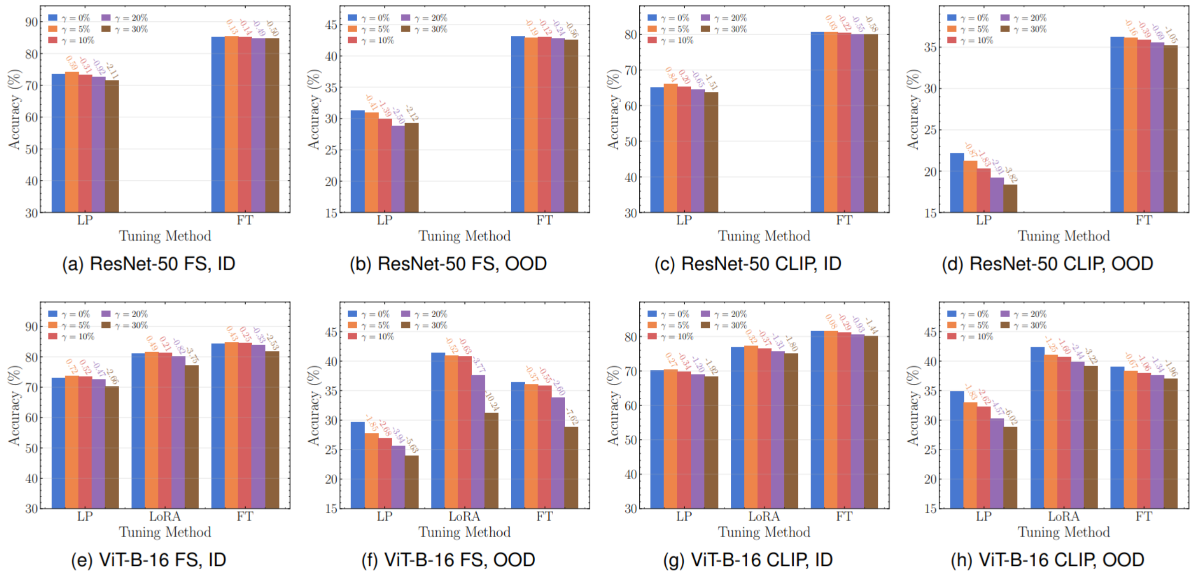
Effects of Pre-training Noise in the Downstream Classification Task LoRA/Full Fine-Tuning



(a) ResNet-50 FS, ID (b) ResNet-50 FS, OOD (c) ResNet-50 CLIP, ID (d) ResNet-50 CLIP, OOD

(e) ViT-B-16 FS, ID (f) ViT-B-16 FS, OOD (g) ViT-B-16 CLIP, ID (h) ViT-B-16 CLIP, OOD

Fig. 3. Average ID and OOD evaluation results of ResNet-50 (top row) and ViT-B-16 (bottom row), using ImageNet-1K (IN-1K) fully-supervised pre-training ((a), (b), (e), (f)) and YFCC15M (and CC12M) CLIP pre-training ((c), (d), (g), (h)) on downstream tasks with various percentages of data. For both ResNet-50 and ViT-B-16 pre-trained on datasets of different scales, on ID evaluation, the transferring performance first increases as noise increases (to 5% or 10%) and then decreases with more noise. On OOD evaluation, the robustness performance constantly decreases.



(a) ResNet-50 FS, ID (b) ResNet-50 FS, OOD (c) ResNet-50 CLIP, ID (d) ResNet-50 CLIP, OOD

(e) ViT-B-16 FS, ID (f) ViT-B-16 FS, OOD (g) ViT-B-16 CLIP, ID (h) ViT-B-16 CLIP, OOD

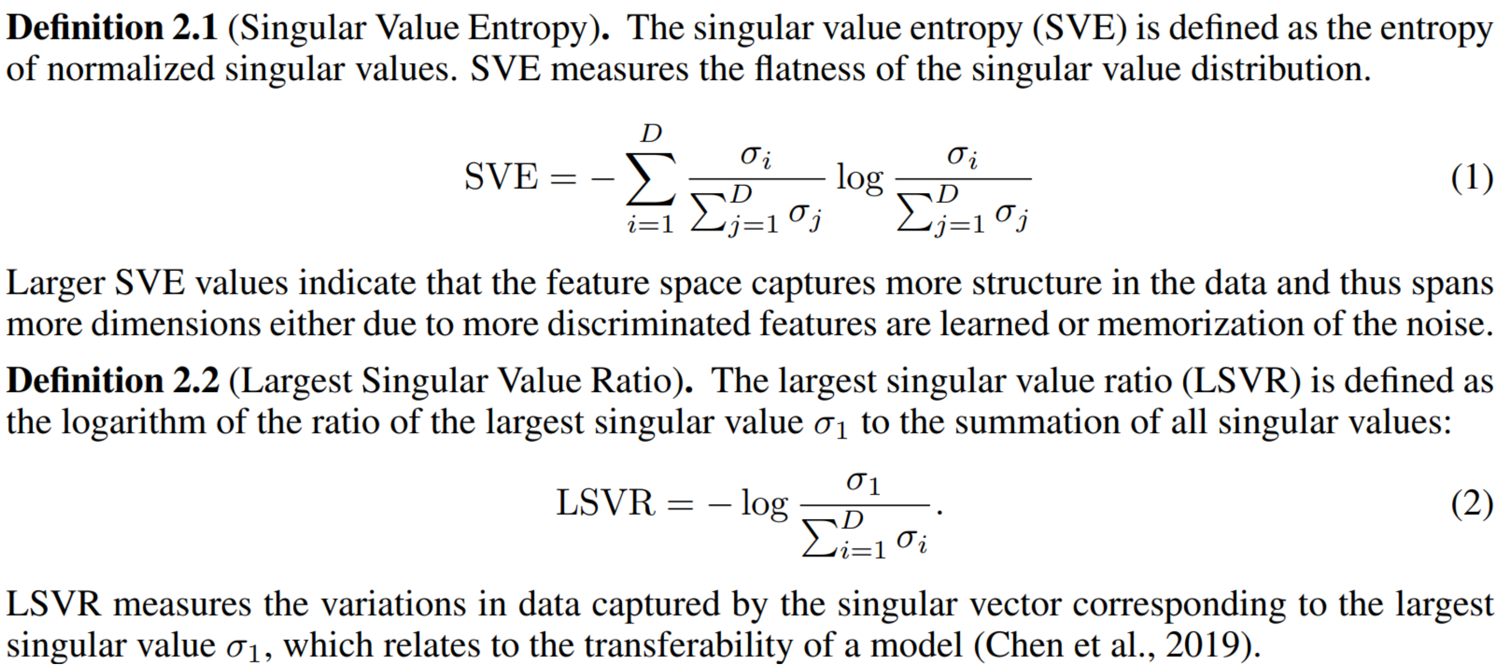
Fig. 4. Average ID and OOD tuning results of ResNet-50 (top row) and ViT-B-16 (bottom row), using ImageNet-1K (IN-1K) fully-supervised pre-training ((a), (b), (e), (f)) and YFCC15M (and CC12M) CLIP pre-training ((c), (d), (g), (h)) on downstream tasks with full data. For ResNet-50, we adopt linear probing (LP) and full fine-tuning (FT). For ViT-B-16, we additionally adopt LoRA. On different tuning methods, we find similar observations for downstream tasks, where slight noise in pre-training benefits the model's ID performance but always hurts the OOD performance. As more pre-trained parameters are modified on downstream tasks, i.e., from LP (to LoRA) to FT, the difference (shown on the top of each bar) between noisy pre-trained models becomes smaller in terms of both ID benefits (with slight noise) and OOD deterioration.

- The slight noise from pre-training still leads to a boost in results when doing LoRA and full fine-tuning in downstream ID tasks. But as more weights/features can be changed, the boost becomes smaller.

- For OOD, LoRA and full fine-tuning can still be observed to deteriorate results once noise is introduced for pre-training. The magnitude still gets smaller as more weights/features can be changed.

Method

Feature Space Analysis



Definition 2.1 (Singular Value Entropy)**.** The singular value entropy (SVE) is defined as the entropy of normalized singular values. SVE measures the flatness of the singular value distribution.

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

Larger SVE values indicate that the feature space captures more structure in the data and thus spans more dimensions either due to more discriminated features are learned or memorization of the noise.

Definition 2.2 (Largest Singular Value Ratio)**.** The largest singular value ratio (LSVR) is defined as the logarithm of the ratio of the largest singular value σ_1 to the summation of all singular values:

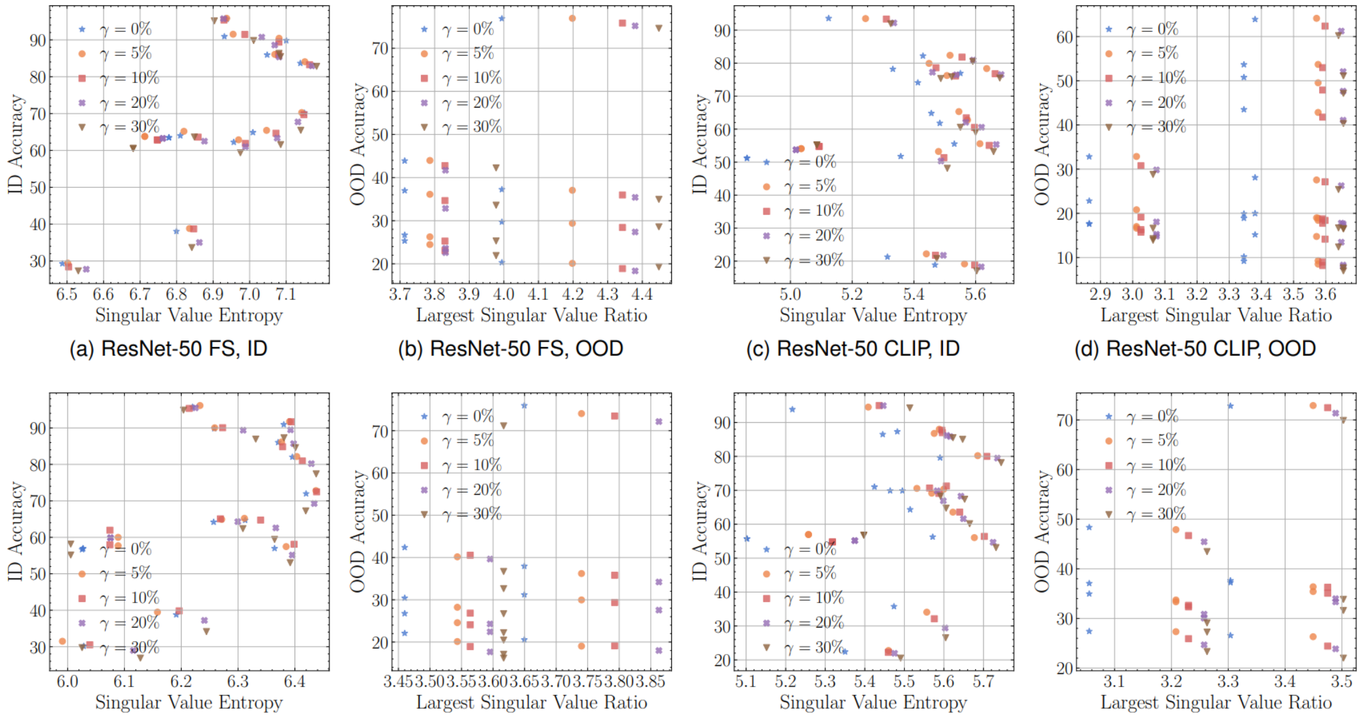
$$\mathrm{LSVR} = -\log \frac{\sigma_1}{\sum_{i=1}^{D} \sigma_i}. \tag{2}$$

LSVR measures the variations in data captured by the singular vector corresponding to the largest singular value σ_1, which relates to the transferability of a model (Chen et al., 2019).

- SVE measures the distribution of singular value. If the singular value is more evenly distributed, the SVE is larger, indicating that the model has more effective dimension/capacity on the feature space.
- LSVR measures the ratio of maximum singular value to the sum of singular values. Here we take the negative log, the smaller the percentage, the larger the LSVR is.

Method

Feature Space Analysis



(a) ResNet-50 FS, ID

(b) ResNet-50 FS, OOD

(c) ResNet-50 CLIP, ID

(d) ResNet-50 CLIP, OOD

(e) ViT-B-16 FS, ID

(f) ViT-B-16 FS, OOD

(g) ViT-B-16 CLIP, ID

(h) ViT-B-16 CLIP, OOD

Fig. 5. Feature SVD analysis of ResNet-50 (top row) and ViT-B-16 (bottom row). We compute the singular value entropy (SVE) for in-domain (ID) tasks and the largest singular value ratio (LSVR) for out-of-domain (OOD) tasks. Both metrics are computed for ImageNet-1K fully-supervised pre-trained ((a), (b), (e), (f)) and YFCC15M (and CC12M) CLIP pre-trained ((c), (d), (g), (h)) models. For both model architectures, the SVE first slightly improves as the noise ratio increases to 5% or 10%, indicating better generalization. As the noise ratio increases, the SVE further improves, and the LSVR drops significantly, corresponding to worse generalization on OOD tasks. The dominant singular components become less transferable.

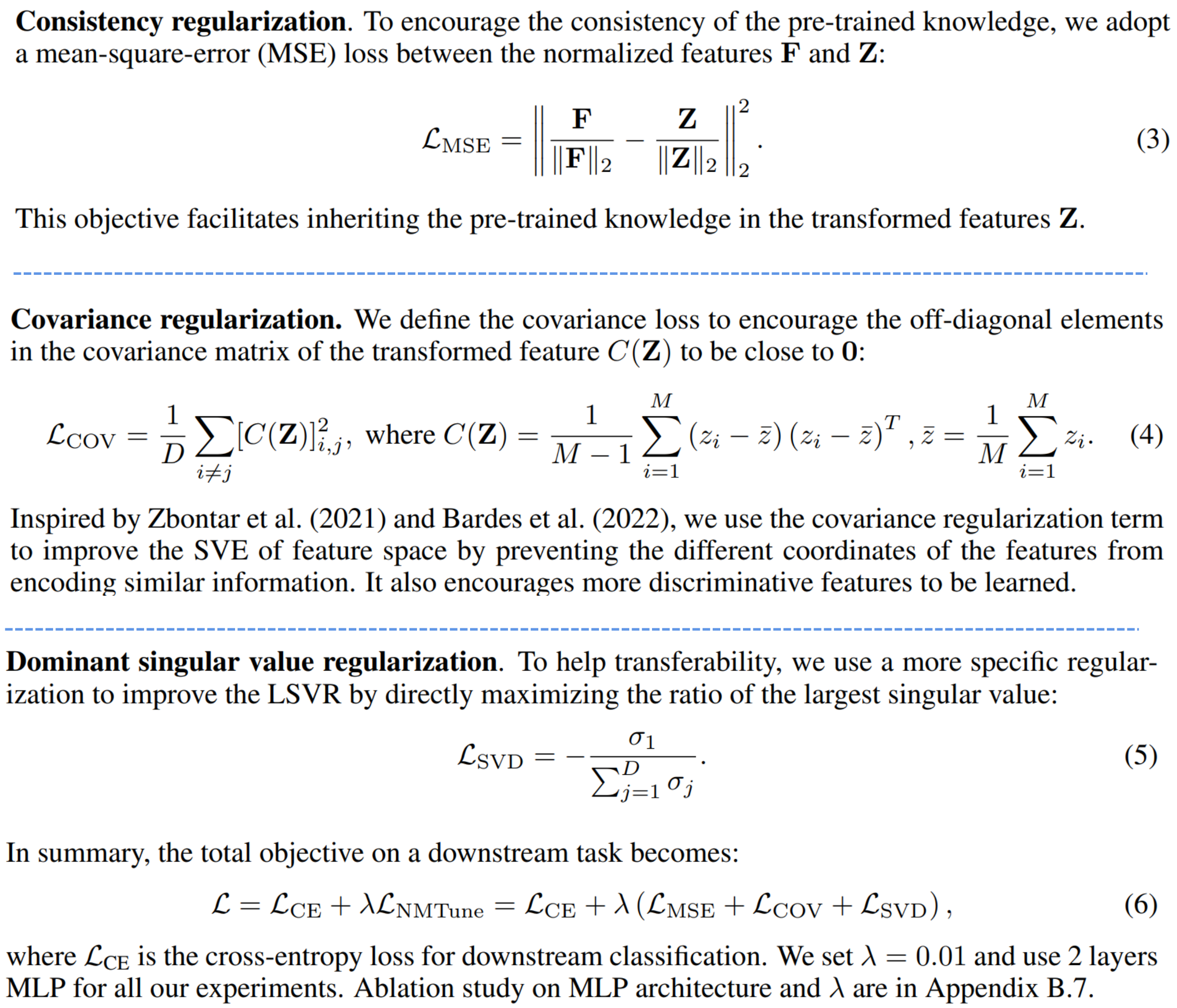
- On the ID task, SVE has a very similar pattern on different datasets. As the pre-training noise increases, both SVE and ID Accuracy first increase and then decrease. This suggests that the model uses more dimension/capacity to fit the noises in the presence of pre-trained noises. a small amount of noises (5%, 10%) leads to a more uniform distribution of the model's feature space, better initialized features and better performance on the ID downstream task.

- On the OOD task, LSVR monotonically increases as pre-training noise increases. It indicates that the model allocates more singular value to the tail singular vector when there is noise in the pre-training, resulting in the feature space going to fewer dominant/most transferable singular vectors.

Method

Elimination of impacts: the NMTune approach

When ADAPT pre-training models on downstream tasks, we propose three regularity terms:



Consistency regularization. To encourage the consistency of the pre-trained knowledge, we adopt a mean-square-error (MSE) loss between the normalized features $\mathbf{F}$ and $\mathbf{Z}$:

$$\mathcal{L}_{\mathrm{MSE}} = \left\| \frac{\mathbf{F}}{\|\mathbf{F}\|_2} - \frac{\mathbf{Z}}{\|\mathbf{Z}\|_2} \right\|_2^2 . \tag{3}$$

This objective facilitates inheriting the pre-trained knowledge in the transformed features $\mathbf{Z}$.

Covariance regularization. We define the covariance loss to encourage the off-diagonal elements in the covariance matrix of the transformed feature $C(\mathbf{Z})$ to be close to $\mathbf{0}$:

$$\mathcal{L}_{\mathrm{COV}} = \frac{1}{D}\sum_{i\neq j}[C(\mathbf{Z})]_{i,j}^2, \text{ where } C(\mathbf{Z}) = \frac{1}{M-1}\sum_{i=1}^{M}\left(z_i - \bar{z}\right)\left(z_i - \bar{z}\right)^T, \bar{z} = \frac{1}{M}\sum_{i=1}^{M} z_i. \tag{4}$$

Inspired by Zbontar et al. (2021) and Bardes et al. (2022), we use the covariance regularization term to improve the SVE of feature space by preventing the different coordinates of the features from encoding similar information. It also encourages more discriminative features to be learned.

Dominant singular value regularization. To help transferability, we use a more specific regularization to improve the LSVR by directly maximizing the ratio of the largest singular value:

$$\mathcal{L}_{\mathrm{SVD}} = -\frac{\sigma_1}{\sum_{j=1}^{D}\sigma_j}. \tag{5}$$

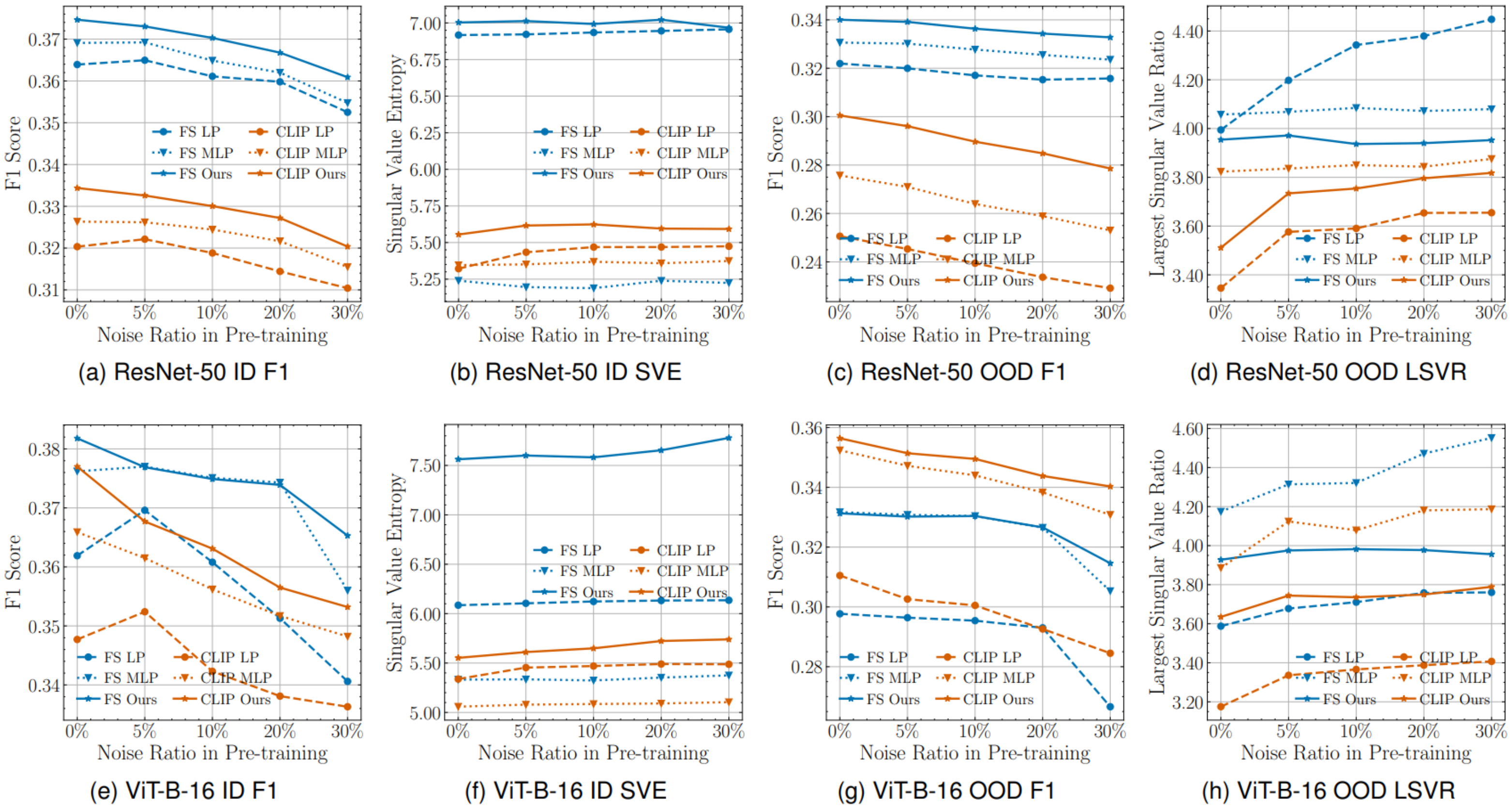
In summary, the total objective on a downstream task becomes:

$$\mathcal{L} = \mathcal{L}_{\mathrm{CE}} + \lambda\mathcal{L}_{\mathrm{NMTune}} = \mathcal{L}_{\mathrm{CE}} + \lambda\left(\mathcal{L}_{\mathrm{MSE}} + \mathcal{L}_{\mathrm{COV}} + \mathcal{L}_{\mathrm{SVD}}\right), \tag{6}$$

where $\mathcal{L}_{\mathrm{CE}}$ is the cross-entropy loss for downstream classification. We set $\lambda = 0.01$ and use 2 layers MLP for all our experiments. Ablation study on MLP architecture and λ are in Appendix B.7.

- The MSE term helps the model to keep the pre-trained knowledge.
- The Cov term helps the model learn a more uniform feature space.
- Svd directly encourages the model to have a larger maximum singular value.

Experiment



(a) ResNet-50 ID F1

(b) ResNet-50 ID SVE

(c) ResNet-50 OOD F1

(d) ResNet-50 OOD LSVR

(e) ViT-B-16 ID F1

(f) ViT-B-16 ID SVE

(g) ViT-B-16 OOD F1

(h) ViT-B-16 OOD LSVR

Fig. 7. Evaluation of our method (NMTune) in a black-box manner on ID and OOD downstream tasks, compared to MLP tuning and LP of ResNet-50 (top row) and ViT-B-16 (bottom row) FS pre-trained on ImageNet-1K (IN-1K) and CLIP pre-trained YFCC15M (and CC12M). (a) Average F1 score of ResNet-50 on ID tasks; (b) SVE of ResNet-50 on ID tasks; (c) Average F1 of ResNet-50 score on OOD tasks; (d) LSVR of ResNet-50 on OOD tasks; (e) F1 of ViT-B-16 on ID tasks; (f) SVE of ViT-B-16 on ID tasks; (g) F1 of ViT-B-16 on OOD tasks; (h) LSVR of ViT-B-16 on OOD tasks. Our method presents better SVE and LSVR on both ID and OOD tasks with better generalization performance. Our method also rectifies the malignant noise effect: the feature extractor pre-trained on clean data now exhibits better performance than others on noisy data on ID tasks; and the performance gap between the clean one and the noisy ones becomes smaller on OOD tasks.

For ID, NMTune not only improves the overall model performance, but also makes clean pre-trained models perform better than noisy pre-trained models.

For OOD, NMTune improves the performance of the model and also mitigates the performance degradation caused by noisy to some extent.

Experiment

Real noise pre-training model

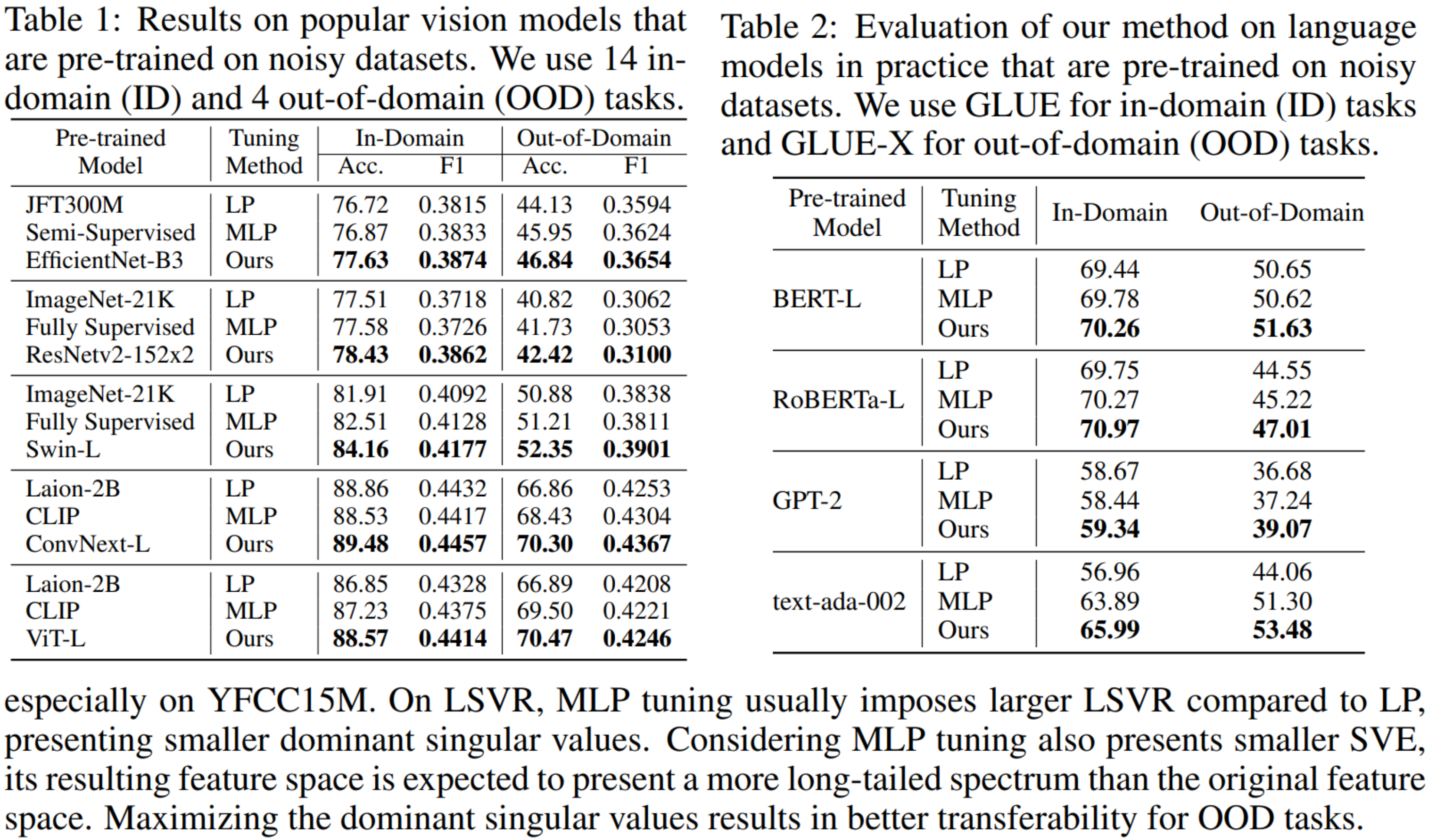


Table 1: Results on popular vision models that are pre-trained on noisy datasets. We use 14 in-domain (ID) and 4 out-of-domain (OOD) tasks.

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**

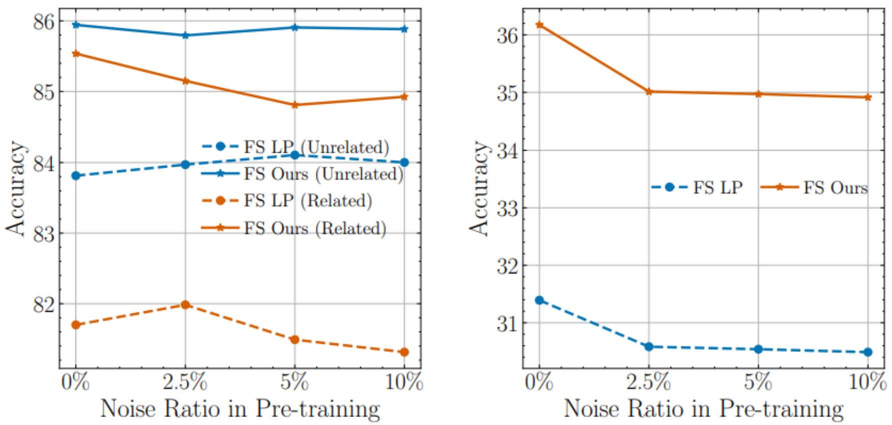
Table 2: Evaluation of our method on language models in practice that are pre-trained on noisy datasets. We use GLUE for in-domain (ID) tasks and GLUE-X for out-of-domain (OOD) tasks.

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**

especially on YFCC15M. On LSVR, MLP tuning usually imposes larger LSVR compared to LP, presenting smaller dominant singular values. Considering MLP tuning also presents smaller SVE, its resulting feature space is expected to present a more long-tailed spectrum than the original feature space. Maximizing the dominant singular values results in better transferability for OOD tasks.

Experiment

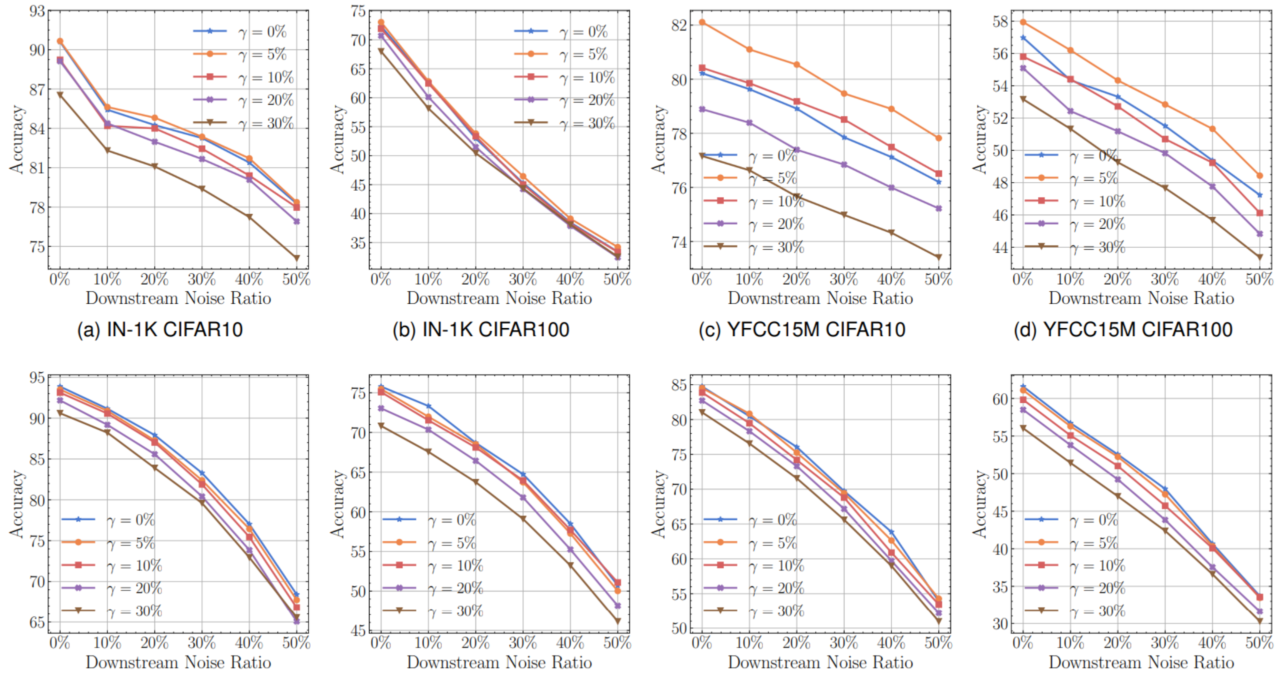
Pre-training Asymmetric noise



(a) ResNet-50 ID (b) ResNet-50 OOD

Fig. 10. Evaluation of IN-1K FS pre-trained ResNet-50 with asymmetric noise using our method (NMTune) in a black-box manner on ID and OOD downstream tasks, compared to LP. (a) Average accuracy on ID tasks, divided to tasks related to the asymmetric noise and unrelated to the noise; (b) Average accuracy on OOD tasks. With asymmetric noise, slight noise still benefits ID performance, even on noise related tasks. Our method presents better performance with smaller difference between the clean and noisy pre-trained models.

Noisy Model Learning + Noisy Label Learning



(a) IN-1K CIFAR10 (b) IN-1K CIFAR100 (c) YFCC15M CIFAR10 (d) YFCC15M CIFAR100

(e) IN-1K CIFAR10 (f) IN-1K CIFAR100 (g) YFCC15M CIFAR10 (h) YFCC15M CIFAR100

Fig. 9. Linear Probing and NMTune of noisy ResNet-50 models on noisy CIFAR-10 and CIFAR-100.

Thanks