GOETHE UNIVERSITÄT FRANKFURT AM MAIN

Seminar Künstliche Intelligenz

Paper:

“Fine tuning Pre trained Models for Robustness Under Noisy Labels”

Marco Cholewik

Fine tuning Pre trained Models for Robustness Under Noisy Labels

- Authors: Sumyeong Ahn[1], Sihyeon Kim[2], Jongwoo Ko[2], Se-Young Yun[2]
- Institutes: [1]Michigan State University, [2]Korea Advanced Institute of Science and Technology

- Publication year: 2024
- Conference: Thirty-Third International Joint Conference on Artificial Intelligence (IJCAI)
- Pages: 3643-3651

Table of Contents

- Phase I
  - Motivation
  - Background
  - Prior work
  - Intuition
- Phase II
  - Mathematic foundation
  - Setup
  - Results
  - Future Work / Relevance

fine-**TU**ning pre-trained models for **R**obustness under **N**oisy labels - **TURN**

- Previous Approaches:
  - Computationally Expensive
  - Datsets Rely on Human Annotations
  - Fine-Tuning Fails under Label Noise
  - Robustness Suffers

- TURN
  - Generally Applicable Methodology
  - Computationally Tolerable
  - Better Robustness

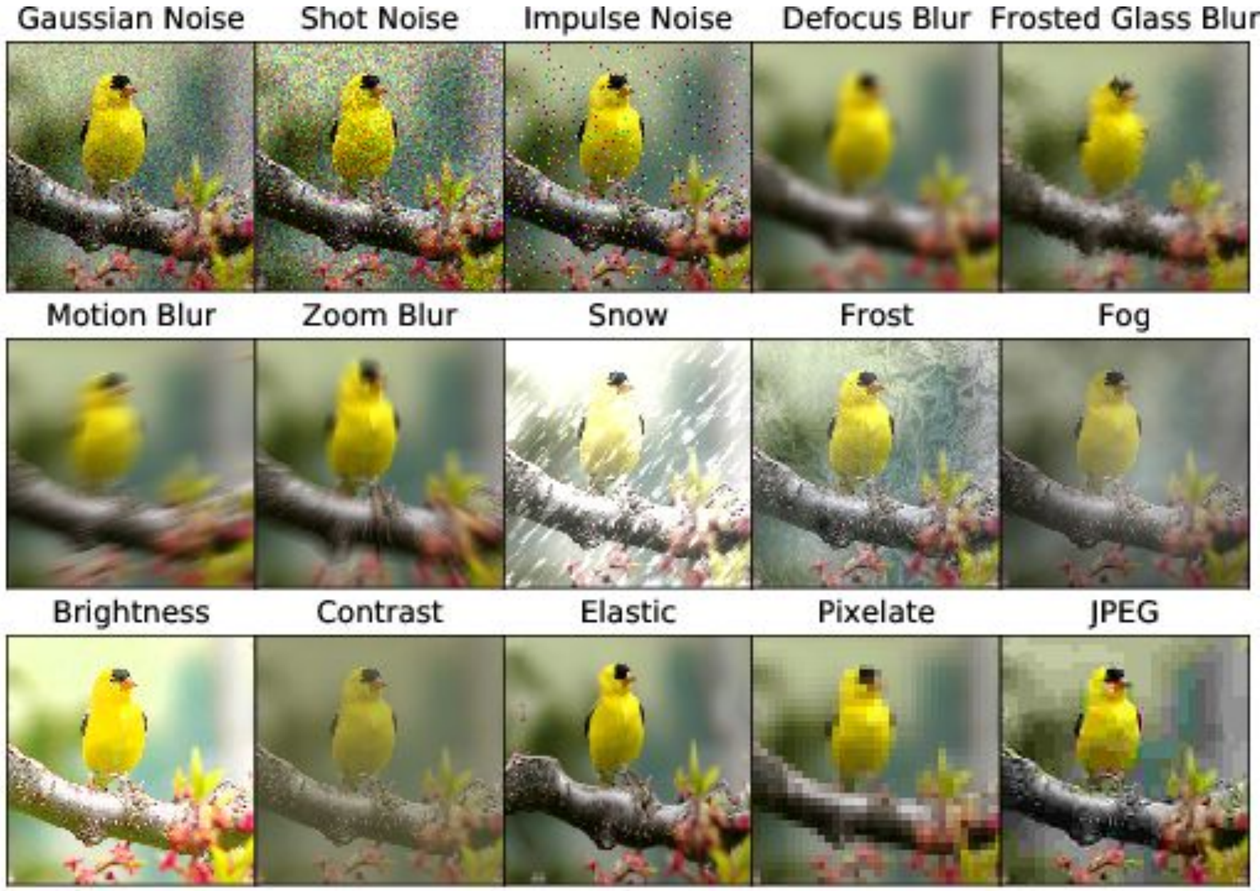
Neural Networks

You all know this:

- Input -> Layers with Weights -> Output (Generative, Classification, …)
- Machine Learning
- Data Hungry
- Different Architectures

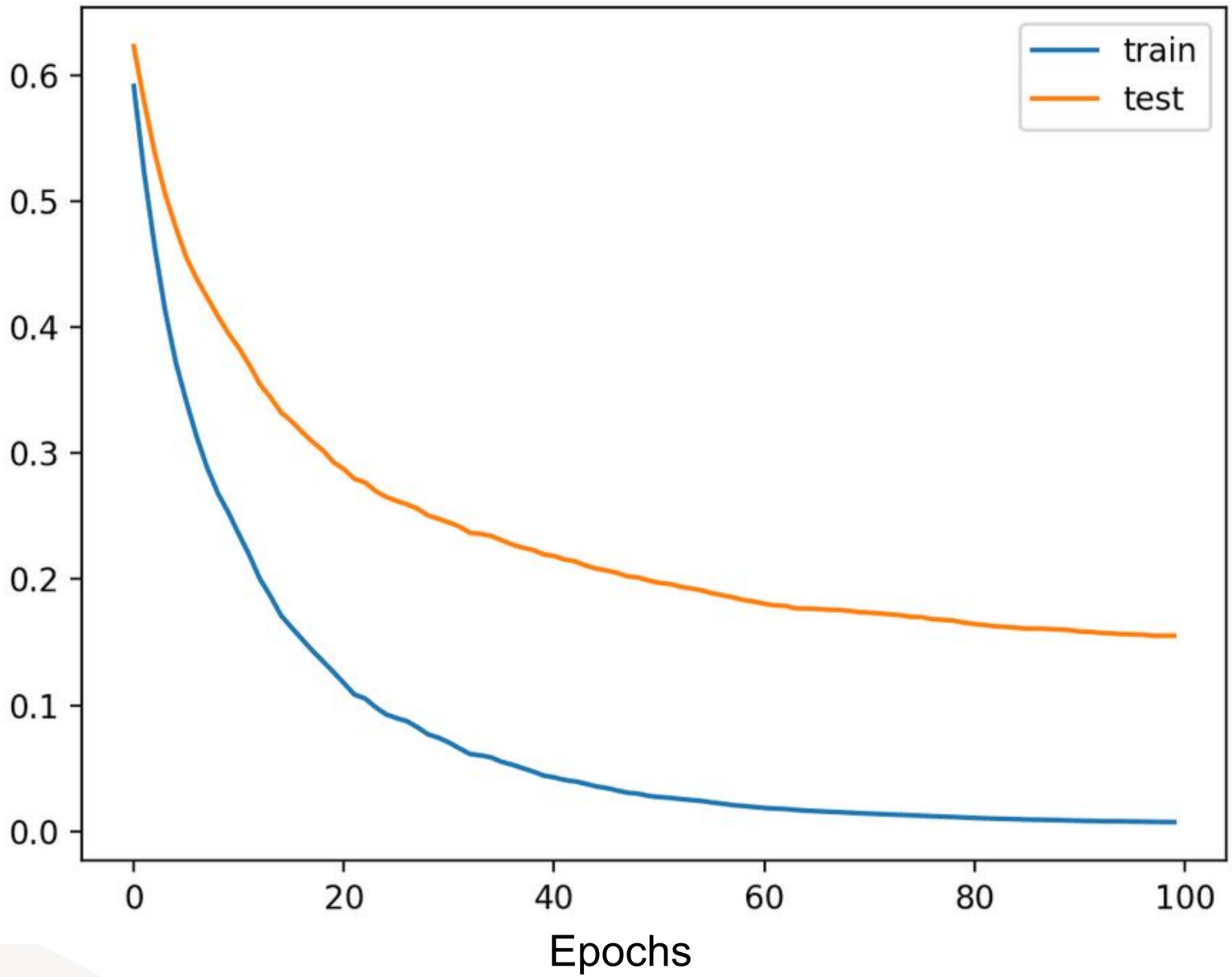
Robustness

- Keep performance under:
  - Noise
  - Out of Distribution (OOD)
  - Domain shift
  - Variations (rotations, color shift, word order,...)
  - Adversarial attacks



[3]

Training



Error

[4]

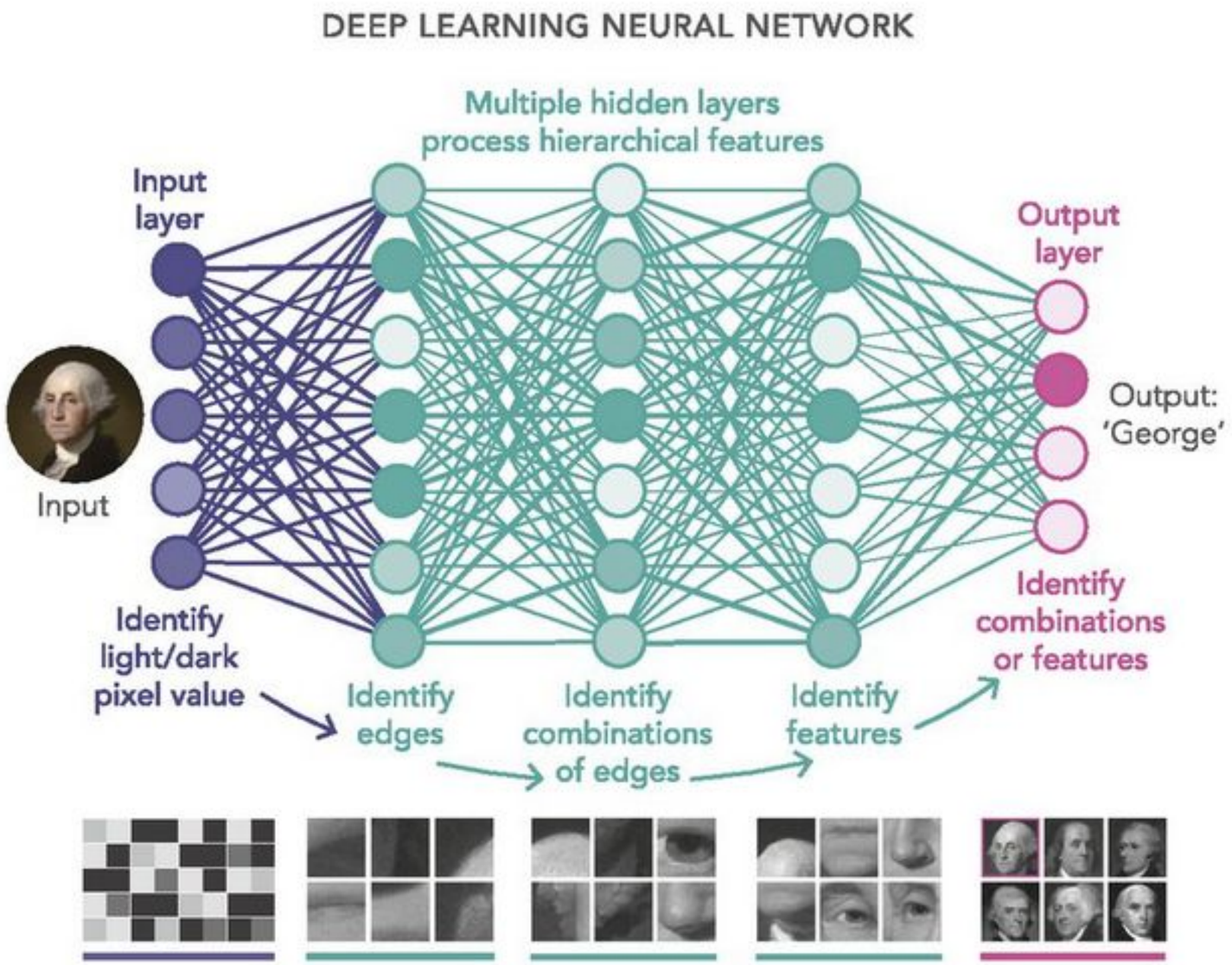
Dataset-Problems

- Size of Datasets
- Human Annotation Impossible
- Severity of Noise Unclear
- Real World is Noisy (Importance of Robustness)

Pre-trained

- Already Trained Model with Saved Weights
- ImageNet [16]
  - 1000 Object Classes
  - 1.281.167 Training Images
- Robust Feature Extractor

Feature Extractor



[2]

Fine-Tuning

- Further Training of Pre-Trained Model
- Higher Accuracy at Specific Task or Better Robustness / Generalization
- Computationally Cheaper than Training from Scratch
- Possibility of Tailor Models
- Different Types:
  - Full Fine-Tuning (FFT)
  - Last Layer Retraining (LLR)

Label Noise

- Change in Training Labels
- Severity
- Symmetric (Uniformly Random Switches of Annotations)
- Asymmetric (Switches within Similar Categories)
- Instance (Switches within Similar Categories dependent on the Instance)

Prior Work

- Detection of Corrupted Instances
  - Co-Teaching (train 2 NNs) [5]
  - DivideMix [6]
  - This is where TURN [1] is Located as Well
- Loss Function and Regularization Terms
  - Generalized Cross Entropy Loss [8]
  - Early Learning Regularization [9]
- Self-Supervised
  - SimCLR [7]
- Limited Research with PTMs under Noisy Labeled Datasets

Intuition

- Observation 1:
  - High Label Noise can Significantly Distort the Feature Extractor under FFT

- Observation 2:
  - FFT can Effectively Enhance the Feature Extractur under low Label Noise

Feature Extractor

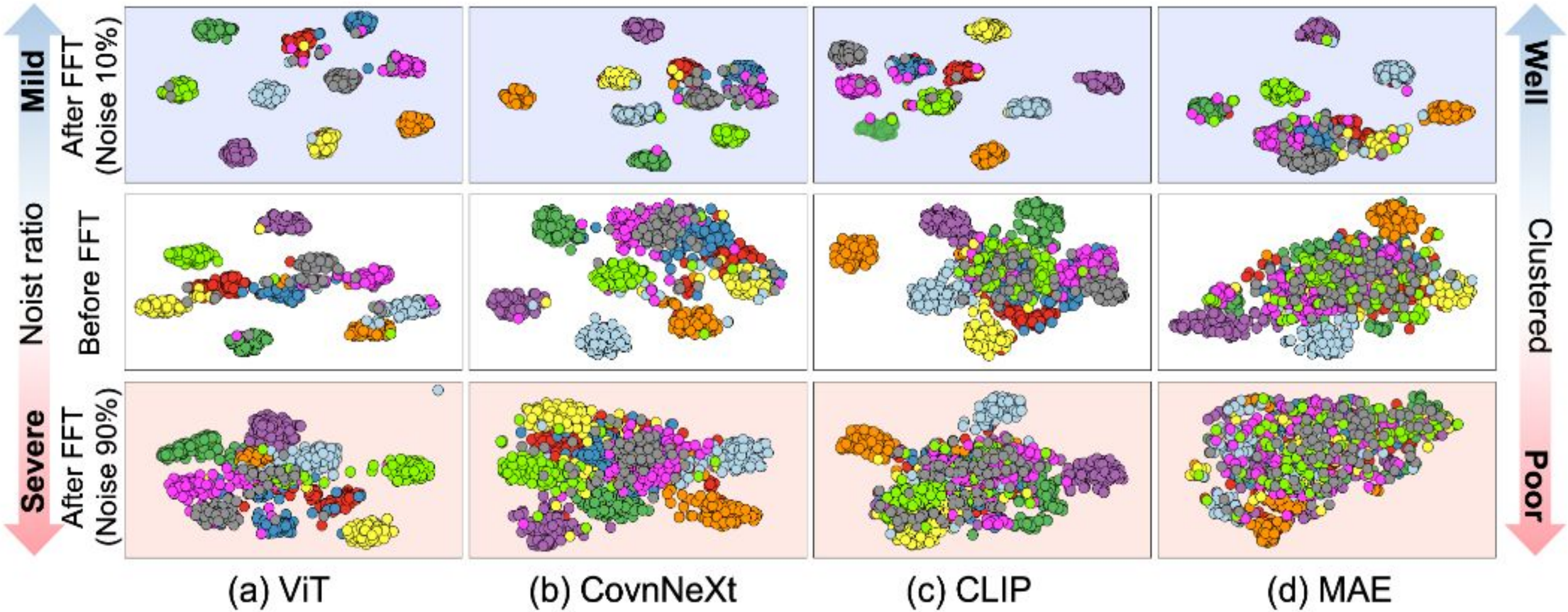


Figure 2: Illustration about tuning characteristics under noisy labeled dataset. We plot t-SNE results before and after FFT on the noise ratio of 90% and 10% datasets. Simply speaking, 60% shows well-clustered features while 90% shows poorly-clustered result.

[1]

Intuition

- Loop:
  - Step 1:
    - Use Pre-Trained Model to Extract Training Data with Correct Labels
  - Step2:
    - Use FFT with the Clean Training Set

Phase II

Mathematical Foundation

Notation:

- **Dataset:** $D_{train} = \{x_i, \hat{y}_i\}_{i=1}^{N}$, with x_i being the image and $\hat{y}_i$ being the given label (can be wrong/noisy)
- **Linear Classifier:** $g(z; \phi)$, taking an input z and with ϕ as the updatable weights
- **Pre-trained Feature Extractor:** $f(x; \theta)$, taking an input x and with θ as the updatable weights
- **Gaussian Mixed Model (GMM) Threshold:** τ for deciding which images to keep in the cleaned dataset
- **Number of Epochs:** E_{LP} and E_{FFT}

TURN-Algorithm

Phase I:

- Linea Probing (LP)
- Training of Classifier $g(z_i; \phi)$ for E_{LP}

- Weights are Updated Using Generalized Cross Entropy Loss:

$$\mathcal{L}_{GCE} = \frac{1 - g(z_i; \phi)^q}{q}$$

With q ∈ (0, 1] as a Hyperparameter

TURN-Algorithm

Phase II:

- For E_{FFT}
  - Create D_{Clean}
  - FFT with Cross-Entropy as Loss-Function

Creation of D_{Clean}

$$n = min_{x \in \{1,...,C\}} \mid D_{clean} \mid$$

$$c = Class$$

$$D_{clean} = \cup_{c=1}^{C} U(D_{clean}^{c}, n)$$

The new dataset is created with n randomly chosen samples from each class with loss $\mathcal{L}_{CE}(g(f(x_i;\theta);\phi), \hat{y}_i) < \tau$

Experimental Setup / Specifics

- Datasets

Dataset	# class	# train	# valid	# test
CIFAR-100	100	47.5K	2.5K	10K
Clothing 1M	14	1M	14K	10K
WebVision	1,000	2.4M	-	50K

- Models
  - ViT-B/16 [10]
  - ResNet [12]
  - ConvNeXt-t [13]
  - CLIP-ViT-B [11]
  - MAE-ViT-B [14]
  - MSN-ViT-B [15]

- Label Noise
  - Variety for CIFAR-100
  - Real-World Noise in Clothing1M and WebVision

- Training Time
  - 20 Epochs LP
  - 5 Epochs FFT

Results

- Promising Results
- Robust Algorithm for Transfer of Knowledge from Pre-Trained Model to Target Dataset
- Better Results than Previous Algorithms (Compared Ones)
- Usable for Datasets with Unknown Label Noise
- Computationally Affordable (Need to Train Only 1 DNN)

Results

Tuning Type	Alg.	CIFAR-100							
		Symm. 0.6	*Symm.* 0.9	*Asym.* 0.4	*Inst.* 0.4	*Symm.* 0.6	*Symm.* 0.9	*Asym.* 0.4	*Inst.* 0.4
		ViT-B/16				ConvNeXt-T			
FFT	CE	88.45±0.59	62.31±1.51	61.25±1.52	64.42±0.14	79.12±0.32	54.72±1.01	68.31±0.52	57.61±0.40
	GCE	89.82±1.32	46.51±0.62	83.73±0.31	1.31±0.60	81.53±0.52	62.31±0.82	79.52±0.73	1.17±0.21
	ELR	88.52±0.18	63.52±0.52	77.83±0.52	83.34±0.27	78.93±0.68	51.52±0.74	74.62±0.42	67.14±0.32
LP	CE	81.20±0.49	64.17±0.62	61.15±1.24	61.62±0.04	70.67±0.69	53.14±0.26	54.83±0.14	62.15±0.31
	GCE	83.19±0.91	81.21±0.15	76.32±0.63	43.11±0.20	73.76±1.32	65.21±0.83	70.26±0.25	5.00±0.05
	ELR	81.23±0.24	65.58±0.62	64.37±0.83	69.43±0.00	70.95±0.16	52.38±0.83	57.15±0.52	61.31±0.21
	DMix	84.31±0.28	80.72±0.52	82.62±0.73	84.26±0.32	74.92±0.92	68.25±1.14	72.41±0.25	65.73±0.52
	UNC	83.15±0.46	80.23±1.25	83.51±1.18	84.32±0.31	71.12±0.71	60.35±0.76	63.92±0.29	69.25±0.3
LP-FFT	Ours	**90.62±0.42**	**84.35±1.13**	**88.13±1.00**	**87.57±0.15**	**83.83±0.52**	**70.01±1.32**	**81.28±1.12**	**73.40±0.13**
		MAE-ViT-B				MSN-ViT-B			
FFT	CE	60.21±0.52	7.58±0.23	55.48±0.52	50.70±0.32	67.42±0.28	5.52±0.13	57.35±0.74	62.24±0.41
	GCE	58.47±0.92	3.06±0.41	60.54±0.85	1.00±0.00	65.51±0.77	7.16±0.32	61.58±0.52	1.00±0.00
	ELR	63.24±0.62	7.84±0.13	61.47±0.52	48.24±0.52	67.19±0.63	5.00±0.24	70.58±0.75	58.14±0.42
LP	CE	48.31±0.86	20.29±0.15	38.98±0.53	44.62±0.75	60.01±0.65	22.82±0.62	47.72±0.86	63.85±0.53
	GCE	49.82±0.73	14.13±0.72	48.27±0.65	1.79±0.36	47.75±0.86	14.15±0.83	42.49±0.82	1.45±0.74
	ELR	47.88±0.72	17.26±0.62	39.32±0.83	46.52±0.53	60.21±0.46	20.72±0.65	51.04±0.25	61.13±0.54
	DMix	59.46±0.93	24.89±0.86	55.64±0.72	51.28±0.43	70.28±0.52	42.58±0.67	65.51±0.85	61.45±0.26
	UNC	37.13±0.52	21.32±0.57	34.21±0.86	39.15±1.24	67.15±0.98	51.82±0.96	61.02±0.74	66.32±1.23
LP-FFT	Ours	**64.33±0.26**	**28.83±0.75**	**65.97±1.00**	**56.53±1.32**	**79.52±0.73**	**54.35±0.64**	**75.33±0.24**	**69.13±1.42**
		CLIP-ViT-B				ResNet-50			
FFT	CE	80.17±0.50	26.84±0.94	64.31±0.85	72.66±0.30	66.12±1.32	0.75±0.61	51.98±1.07	56.12±2.58
	GCE	81.56±1.01	3.18±0.68	78.35±0.87	1.13±0.12	55.78±0.42	5.14±1.52	57.04±0.87	1.21±0.25
	ELR	76.24±0.51	32.27±1.18	75.38±1.17	71.66±0.56	65.38±0.69	8.51±1.59	61.21±1.10	56.60±1.55
LP	CE	74.24±0.91	52.17±1.18	53.99±1.79	63.09±1.33	67.19±0.52	49.17±1.70	53.52±2.00	54.95±2.22
	GCE	79.66±1.13	65.49±1.35	72.91±0.36	19.87±0.43	65.21±1.52	49.32±0.76	58.24±2.19	57.58±1.80
	ELR	73.92±1.21	51.94±0.60	56.57±2.67	65.11±1.72	65.14±0.93	49.53±1.09	55.08±1.49	54.51±1.21
	DMix	77.97±0.99	69.55±0.90	75.17±1.70	71.12±0.38	71.03±0.92	56.54±0.54	62.85±1.45	60.40±1.18
	UNC	73.54±0.52	59.55±1.07	67.37±1.38	72.47±2.68	70.03±1.53	58.08±0.92	66.41±0.89	67.79±0.61
LP-FFT	Ours	**84.12±0.82**	**72.55±1.45**	**78.41±0.89**	**80.96±1.97**	**73.32±0.93**	**59.64±0.60**	**69.38±1.00**	**69.78±0.76**

Results

Architecture	Clothing1M								
	LP					FFT			LP+FFT
	CE	GCE	ELR	DivideMix	UNICON	CE	GCE	ELR	Ours
ViT-B/16	67.83 / 67.54	67.46 / 67.46	66.91 / 66.91	68.13 / 68.13	68.42 / 68.42	68.98 / 68.98	69.74 / 69.74	68.73 / 68.73	**70.28 / 70.28**
ConvNeXt-T	64.82 / 64.81	64.59 / 64.59	64.17 / 64.17	66.12 / 65.42	67.33 / 66.92	68.80 / 68.80	68.92 / 68.92	69.19 / 68.52	**69.63 / 69.63**
MAE-ViT-B	5.06 / 5.06	5.92 / 5.92	8.28 / 8.28	8.04 / 8.04	8.52 / 8.52	61.31 / 61.31	60.80 / 60.80	61.51 / 61.51	**61.96 / 61.96**
MSN-ViT-B	6.77 / 6.77	6.20 / 6.20	7.64 / 7.64	6.42 / 6.42	6.31 / 6.31	66.88 / 63.38	67.06 / 65.41	66.32 / 66.32	**69.13 / 69.13**
ResNet-50	7.08 / 7.08	7.18 / 7.18	6.68 / 6.68	8.13 / 8.13	8.24 / 8.24	66.10 / 66.02	66.19 / 66.19	66.19 / 66.19	**66.31 / 66.31**

Architecture	WebVision								
	LP					FFT			LP+FFT
	CE	GCE	ELR	DivideMix	UNICON	CE	GCE	ELR	Ours
ViT-B/16	84.62 / 84.48	84.32 / 84.24	84.48 / 84.32	84.72 / 84.72	85.68 / 85.68	84.20 / 83.04	83.40 / 83.40	84.92 / 83.72	**85.96 / 85.92**
ConvNeXt-T	85.24 / 85.24	85.12 / 85.04	86.28 / 86.28	86.40 / 86.40	86.24 / 86.24	84.00 / 82.68	85.40 / 84.92	84.52 / 83.44	**87.16 / 86.44**
MAE-ViT-B	48.00 / 48.00	47.32 / 47.28	49.76 / 49.76	59.40 / 58.44	56.96 / 53.80	67.48 / 65.64	63.16 / 62.84	67.80 / 67.80	**69.45 / 68.45**
MSN-ViT-B	77.40 / 77.40	74.40 / 74.40	74.00 / 74.00	76.56 / 76.40	77.72 / 77.34	77.04 / 77.80	72.28 / 72.28	74.88 / 72.28	**78.36 / 75.40**
ResNet-50	84.88 / 84.72	81.68 / 81.68	84.96 / 84.96	85.16 / 85.16	85.04 / 85.04	78.00 / 76.44	77.04 / 70.92	80.44 / 77.44	**85.36 / 85.36**

Relevance / Future Work / Open Questions

- Robustness and Computational Cost are Essential
- Dataset Generation is Almost Impossible in that Size -> Need for Methodology

- There is a Need for Further Analysis of Fine-Tuning Options Given its Importance
- Influence of Pre-Training on Later Fine-Tuning

- Show of Robustness ?
- Compared Algorithms Seem Insufficient
- Does it Work on Small Datasets?

Questions?

Literature

- [1]: Ahn, S., Kim, S., Ko, J., & Yun, S. (2024). Fine-tuning Pre-trained Models for Robustness under Noisy Labels. *International Joint Conference on Artificial Intelligence*, 3643–3651. https://doi.org/10.24963/ijcai.2024/403
- [2]: M.M. Waldrop (2019). What are the limits of deep learning?, *Proc. Natl. Acad. Sci. U.S.A. 116 (4)* 1074-1077, https://doi.org/10.1073/pnas.1821594116
- [3]: Hendrycks, D., & Dietterich, T. (2019). Benchmarking Neural Network Robustness to Common Corruptions and Perturbations. *Proceedings of the International Conference on Learning Representations*.
- [4]: Brownlee, J. (2021, March 15). *Tune XGBooSt performance with learning curves*. Machinelearningmastery. Retrieved July 8, 2025, from https://machinelearningmastery.com/tune-xgboost-performance-with-learning-curves/

Literature

- [5]: B. Han, Q. Yao, X. Yu, G. Niu, M. Xu, W. Hu, I. Tsang, and M. Sugiyama. Co-teaching: Robust training of deep neural networks with extremely noisy labels. *Advances in neural information processing systems*, 31, 2018.
- [6]: Li, J., Hoi, S. C., & Socher, R. (2020). DivideMix: Learning with Noisy Labels as Semi-supervised Learning. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2002.07394
- [7]: Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. (2020). A simple framework for contrastive learning of visual representations. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2002.05709
- [8]: Zhang, Z., & Sabuncu, M. R. (2018). Generalized Cross Entropy Loss for Training Deep Neural Networks with Noisy Labels. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.1805.07836
- [9]: Liu, S., Niles-Weed, J., Razavian, N., & Fernandez-Granda, C. (2020). Early-Learning regularization prevents memorization of noisy labels. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2007.00151

Literature

- [10]: A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.
- [11]: A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al. Learning transferable visual models from natural language supervision. In I*nternational Conference on Machine Learning, pages 8748–8763*. PMLR, 2021.
- [12]: K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.
- [13]: Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11976–11986, 2022.

Literature

- [14]: K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16000–16009, 2022.
- [15]: M. Assran, M. Caron, I. Misra, P. Bojanowski, F. Bordes, P. Vincent, A. Joulin, M. Rabbat, and N. Ballas. Masked siamese networks for label-efficient learning. In *European Conference on Computer Vision*, pages 456–473. Springer, 2022.
- [16]: *ImageNet*. (n.d.). https://www.image-net.org/update-mar-11-2021.php