Decoupled Training for Long-tailed Classification With Stochastic Representations

Giung Nam*[1] Sunguk Jang*†[2] Juho Lee[1,2]

[1]KAIST, South Korea, [2]AITRICS, South Korea.

*Equal contribution, †The work was done while the author was a graduate student at KAIST.

Decoupled Training for *Long-tailed Classification*

- The real-world classification data are often *long-tailed*.
- The iNaturalist dataset is a prominent example of this phenomenon.

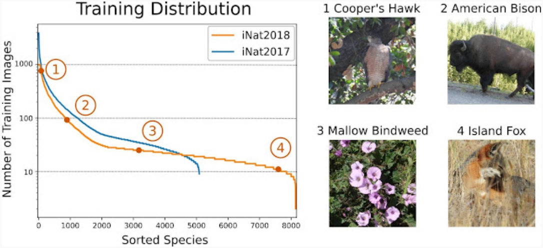

Figure 1: Distribution of the number of train examples per species for iNaturalist datasets, plotted on a log-linear scale[1].

[1] image credit: Grant Van Horn and Oisin Mac Aodha.

Preliminaries

Decoupled Training for Long-tailed Classification

- Decoupling representation learning and classifier learning has been shown to be effective in long-tailed classification [Kang et al., 2020].

> It is also possible to achieve strong long-tailed recognition ability by adjusting only the classifier, with representations learned with the simplest instance-balanced sampling.

- In a nutshell, we can implement *decoupled training* as follows;
  1. Representation learning stage,

$$(\boldsymbol{\theta}^*, \boldsymbol{\phi}^*) = \operatorname*{arg\,min}_{(\boldsymbol{\theta}, \boldsymbol{\phi})} \mathbb{E}_{(\boldsymbol{x}, y) \sim \mathcal{D}} \left[\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\phi}; \boldsymbol{x}, y) \right]. \tag{1}$$

  2. Classifier re-training stage [cRT; Kang et al., 2020],

$$\boldsymbol{\phi}^{**} = \operatorname*{arg\,min}_{\boldsymbol{\phi}} \mathbb{E}_{(\boldsymbol{x}, y) \sim \mathcal{D}_{\text{CB}}} \left[\mathcal{L}(\boldsymbol{\theta}^*, \boldsymbol{\phi}; \boldsymbol{x}, y) \right]. \tag{2}$$

Decoupled Training w/ Stochastic Representations

Constructing an effective decoupled learning scheme

- [Q1] How to train the feature extractor for representation learning so that it provides generalizable representations?
- [Q2] How to re-train the classifier that constructs proper decision boundaries by handling class imbalances in long-tailed data?

Does the success of SWA continue in the long-tailed classification?

- *Stochastic Weight Averaging (SWA)* improves the generalization performance by seeking flat minima in loss surfaces [Izmailov et al., 2018].
- Without classifier re-training, SWA itself *does not* bring significant performance gain for long-tailed classification tasks.
- We diagnose that SWA actually *enhances* the quality of the feature extractor, but the classification layer is acting as a bottleneck.

[A1] Confirming that SWA can benefit long-tailed classification, we apply SWA to obtain more generalizing feature extractor.

Decoupled Training w/ Stochastic Representations

Stochastic representations reflect the difficulty of each input.

- *SWA-Gaussian (SWAG)* further provides a Gaussian approximation that captures the geometry of the posterior over parameters [Maddox et al., 2019].

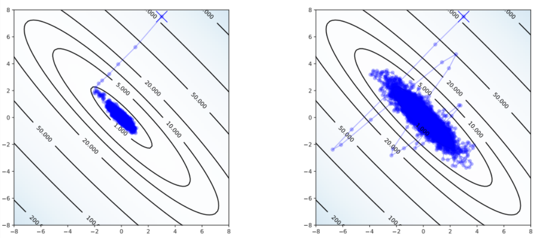

Figure 2: Quadratic loss contour plot and iterates of SGD [Maddox et al., 2019].

- We consider the *stochastic representations*,

$$\{\mathcal{F}(\boldsymbol{x};\boldsymbol{\theta}_m)\}_{m=1}^{M}, \text{ where } \boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_M \sim q(\boldsymbol{\theta}|\mathcal{D}) = \mathcal{N}(\boldsymbol{\theta};\boldsymbol{\theta}_{\mathrm{SWA}},\boldsymbol{\Sigma}_{\mathrm{SWAG}}). \quad (3)$$

Decoupled Training w/ Stochastic Representations

Stochastic representations reflect the difficulty of each input.

- Empirically, the stochastic representations well reflect the uncertainty of inputs, e.g., the head-class instance tends to have smaller dispersion.
- The *dispersion* quantifies how stochastic representations are scattered.

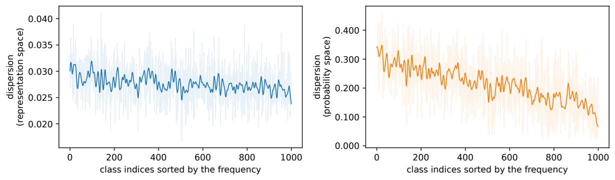

Figure 3: The per-class dispersion along with class indices on ImageNet-LT. It measured in (left) the representation space and (right) the probability space.

[A2] Confirming that the stochastic representations obtained from SWAG well reflect the uncertainty of inputs, we utilize them to build more robust decision boundary.

Decoupled Training w/ Stochastic Representations

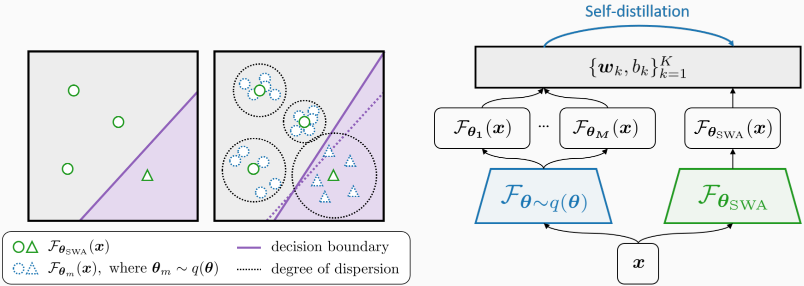

Figure 4: Schematic diagrams depicting the overall concepts of the paper. **Left:** An illustration of two-dimensional representation space. **Right:** Our proposed self-distillation strategy obtaining more robust decision boundaries.

Table 1: Ablation studies of proposed methods on ImageNet-LT: classification accuracy (ACC), negative log-likelihood (NLL), and expected calibration error (ECE).

Method	ACC (↑)	NLL (↓)	ECE (↓)
SGD w/ classifier re-training	50.97	2.231	0.063
+ (a) introducing SWA for the representation learning	51.62	2.206	0.077
+ (b) classifier re-training w/ stochastic representation	51.84	2.208	0.090
+ (c) classifier re-training w/ self-distillation	**52.12**	**2.130**	**0.037**

Table 2: Results on ImageNet-LT: classification accuracy (ACC), negative log-likelihood (NLL), and expected calibration error (ECE).

	ACC (↑)					
ImageNet-LT	Many	Medium	Few	All	NLL (↓)	ECE (↓)
SGD	66.84$\pm$0.26	40.78$\pm$0.24	12.05$\pm$0.23	46.91$\pm$0.22	2.546$\pm$0.009	0.158$\pm$0.003
+ cRT [Kang et al., 2020]	62.83$\pm$0.23	46.92$\pm$0.26	26.33$\pm$0.16	50.25$\pm$0.18	2.364$\pm$0.008	0.110$\pm$0.001
+ LWS [Kang et al., 2020]	63.23$\pm$0.26	47.57$\pm$0.24	27.78$\pm$0.23	50.91$\pm$0.15	**2.197**$\pm$0.007	**0.054**$\pm$0.001
+ LA [Menon et al., 2021]	60.79$\pm$0.20	48.11$\pm$0.14	33.20$\pm$0.34	50.97$\pm$0.13	2.231$\pm$0.004	0.063$\pm$0.001
+ DisAlign [Zhang et al., 2021]	61.63$\pm$0.39	48.68$\pm$0.11	32.71$\pm$0.45	**51.49**$\pm$0.15	2.596$\pm$0.012	0.202$\pm$0.002
SWA	67.71$\pm$0.11	40.74$\pm$0.15	11.01$\pm$0.10	47.08$\pm$0.12	2.631$\pm$0.009	0.187$\pm$0.002
+ cRT [Kang et al., 2020]	63.54$\pm$0.18	47.68$\pm$0.16	26.85$\pm$0.28	50.95$\pm$0.12	2.353$\pm$0.012	0.120$\pm$0.002
+ LWS [Kang et al., 2020]	63.51$\pm$0.30	48.53$\pm$0.07	28.66$\pm$0.45	51.60$\pm$0.10	2.189$\pm$0.007	0.077$\pm$0.002
+ LA [Menon et al., 2021]	61.60$\pm$0.07	48.70$\pm$0.03	33.68$\pm$0.34	51.62$\pm$0.05	2.206$\pm$0.009	0.077$\pm$0.002
+ DisAlign [Zhang et al., 2021]	62.43$\pm$0.20	49.48$\pm$0.15	32.65$\pm$0.43	**52.18**$\pm$0.11	2.673$\pm$0.014	0.215$\pm$0.002
+ SRepr (ours)	62.52$\pm$0.26	49.44$\pm$0.18	32.14$\pm$0.41	**52.12**$\pm$0.06	**2.130**$\pm$0.006	**0.037**$\pm$0.001

Experimental Results

To summarize:

- We first apply SWA to obtain better generalizing feature extractors for long-tailed classification.
- We then propose a new classifier re-training algorithm using stochastic representation obtained from SWA-Gaussian.
- Our approach improves both accuracy and uncertainty estimation.

More experimental results are available in the paper!

- Results on CIFAR-10-LT, CIFAR-100-LT, and iNaturalist-2018.
- Ablations with various balancing strategies.
- Further analysis on proposed methods.

References

Pavel Izmailov, Dmitrii Podoprikhin, Timur Garipov, Dmitry Vetrov, and Andrew Gordon Wilson. Averaging weights leads to wider optima and better generalization. In *34th Conference on Uncertainty in Artificial Intelligence 2018, UAI 2018*, 2018.

Bingyi Kang, Saining Xie, Marcus Rohrbach, Zhicheng Yan, Albert Gordo, Jiashi Feng, and Yannis Kalantidis. Decoupling representation and classifier for long-tailed recognition. In *International Conference on Learning Representations (ICLR)*, 2020.

Wesley J Maddox, Pavel Izmailov, Timur Garipov, Dmitry P Vetrov, and Andrew Gordon Wilson. A simple baseline for bayesian uncertainty in deep learning. In *Advances in Neural Information Processing Systems 32 (NeurIPS 2019)*, 2019.

Aditya Krishna Menon, Sadeep Jayasumana, Ankit Singh Rawat, Himanshu Jain, Andreas Veit, and Sanjiv Kumar. Long-tail learning via logit adjustment. In *International Conference on Learning Representations (ICLR)*, 2021.

Songyang Zhang, Zeming Li, Shipeng Yan, Xuming He, and Jian Sun. Distribution alignment: A unified framework for long-tail visual recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.