Disentangling Label Distribution for Long-tailed Visual Recognition CVPR 2021

Youngkyu Hong* Seungju Han* Kwanghee Choi* Seokjun Seo Beomsu Kim Buru Chang *Equal contributions.

{youngkyu.hong,seungju.han,kwanghee.choi,seokjun.seo,beomsu.kim,buru.chang}@hpcnt.com Hyperconnect, Republic of Korea

HYPERCONNECT

Summary

- We borrow the concept of label shift problem to suggest a more practical setting for the long-tailed visual recognition problem.
- To solve the problem, we design a novel loss that directly disentangles the label distribution from the trained model.
- Our method outperforms state-of-the-art long-tailed methods in various settings.

Motivation

Bayes' rule for Image Classification

Given input image x and output class label y, model estimates:

$$p(y|x) = \frac{p(y)p(x|y)}{p(x)},$$

where $p(y)$ is the class label distribution, which is the ratio of the labels in the training dataset.

Long-tailed Classification as Label Distribution Shift

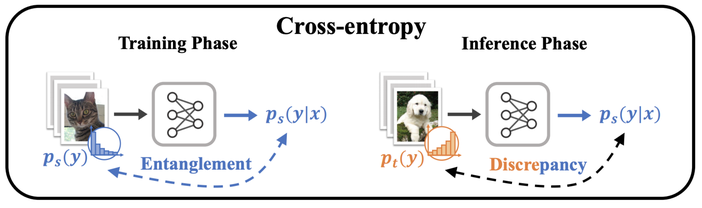

Figure 1:After training with the cross-entropy loss, the model prediction gets entangled with the source label distribution $p_s(y)$, which causes a discrepancy with the target label distribution $p_t(y)$ during the inference phase.

In this paper, we cope with long-tailed visual recognition as one of the label distribution shift problems. We assume $p_s(x|y) = p_t(x|y)$ but $p_s(y) \neq p_t(y)$.

Long-tailed Classification Evaluation Protocol

The current protocol trains the classification model on the long-tailed source label distribution and evaluates its performance on the uniform target label distribution. However, we argue that this is often impractical as it is natural to assume that $p_t(y)$ could be the arbitrary distribution.

Post-Compensated Softmax

The straightforward way to handle the label distribution shift is by replacing $p_s(y)$ with $p_t(y)$. We introduce a post-compensation (PC) strategy that modifies the logit in the inference phase.

The estimation of $p_t(y|x)$ is formulated as:

$$p_t(y|x;\theta) = \frac{\frac{p_t(y)}{p_s(y)} \cdot e^{f_\theta(x)[y]}}{\sum_c \frac{p_t(c)}{p_s(c)} \cdot e^{f_\theta(x)[c]}} = \frac{e^{(f_\theta(x)[y] - \log p_s(y) + \log p_t(y))}}{\sum_c e^{(f_\theta(x)[c] - \log p_s(c) + \log p_t(c))}},$$

where $f_\theta(x)[y]$ is the logit of class y from the Softmax regression model estimating $p_s(y|x)$.

Despite the simplicity of this method, PC Softmax becomes a strong baseline that surpasses previous state-of-the-art long-tailed visual recognition methods. However, no recent literature consider this as a baseline.

LAbel distribution DisEntangling (LADE) loss

Performance gain from the PC strategy shows the efficacy of disentangling the source label distribution. However, for the further disentanglement in the training phase, we design a new modeling objective.

1. Detach $p_s(y)$ from $p_s(y|x)$, which results in $p_s(x|y)/p_s(x)$.
2. Replace $p_s(y)$ in $p_s(x)$ with the uniform prior $p_u(y)$, i.e. $p_u(y=c) = 1/C$, where C is the total number of classes.
3. This yields the modeling objective for the logits: $f_\theta(x)[y] = \log \frac{p_u(x|y)}{p_u(x)}$.

We utilize the optimal form of the regularized Donsker-Varadhan (DV) representation (Choi et al., 2020) to model the log-likelihood ratio above.

$$\log \frac{d\mathbb{P}}{d\mathbb{Q}} = \arg\max_{T:\Omega\to\mathbb{R}} (\mathbb{E}_\mathbb{P}[T] - \log(\mathbb{E}_\mathbb{Q}[e^T]) - \lambda(\log(\mathbb{E}_\mathbb{Q}[e^T]))^2)$$

$$\mathcal{L}_{LADERc} = -\frac{1}{N_c}\sum_{i=1}^{N} 1_{y_i=c} \cdot f_\theta(x_i)[c] + \log(\frac{1}{N}\sum_{i=1}^{N} \frac{p_u(y_i)}{p_s(y_i)} \cdot e^{f_\theta(x_i)[c]}) + \lambda(\log(\frac{1}{N}\sum_{i=1}^{N} \frac{p_u(y_i)}{p_s(y_i)} \cdot e^{f_\theta(x_i)[c]}))^2$$

The final loss is derived as follows:

$$\mathcal{L}_{LADE-CE}(f_\theta(x), y) = -\log(p_s(y|x;\theta)) = -\log(\frac{p_s(y) \cdot e^{f_\theta(x)[y]}}{\sum_c p_s(c) \cdot e^{f_\theta(x)[c]}})$$

$$\mathcal{L}_{LADE}(f_\theta(x), y) = \mathcal{L}_{LADE-CE}(f_\theta(x), y) + \alpha \cdot \sum_{c\in\mathbb{S}} \alpha_c \cdot \mathcal{L}_{LADERc}$$

Comparison with Other Methods

Dataset	Previous SotA	PC-Softmax	LADE
CIFAR-100 (IR=100)	45.1	45.3	**45.4**
CIFAR-100 (IR=50)	50.3	49.5	**50.5**
CIFAR-100 (IR=10)	61.6	61.6	**61.7**
Places-LT	38.6	38.7	**38.8**
ImageNet-LT	52.0	52.8	**53.0**
iNaturalist 2018	69.8	69.3	**70.0**

Table 1: Previous SotA is the max accuracy among these methods: Focal Loss (Lin et al., 2017), LDAM and LDAM-DRW (Cao et al., 2019), BBN (Zhou et al., 2020), Causal Norm (Tang et al., 2020), Balanced Softmax (Ren et al., 2020), OLTR (Liu et al., 2019), Decouple-* (Kang et al., 2020). IR denotes the imbalance ratio.

Dataset	Forward					Uniform	Backward				
IR	50	25	10	5	2	1	2	5	10	25	50
SotA	66.3	63.9	60.4	57.8	54.6	52.1	49.6	46.5	44.1	41.4	39.7
SotA+PC	66.7	64.3	60.9	58.1	54.6	52.1	50.2	48.8	48.3	48.5	49.0
LADE	**67.4**	**64.8**	**61.3**	**58.6**	**55.2**	**53.0**	**51.2**	**49.8**	**49.2**	**49.3**	**50.0**

Table 2: ImageNet-LT accuracy results on variously test-time shifted label distributions. Dataset sorted from most to least similar to the source label distribution. SotA is the max accuracy among these methods: Vanilla Softmax, Causal Norm (Tang et al., 2020), Balanced Softmax (Ren et al., 2020). SotA+PC denotes SotA methods with PC strategy applied. PC strategy shows consistent performance gain, and LADE outperforms all the other methods in every imbalance settings.

Further Analysis

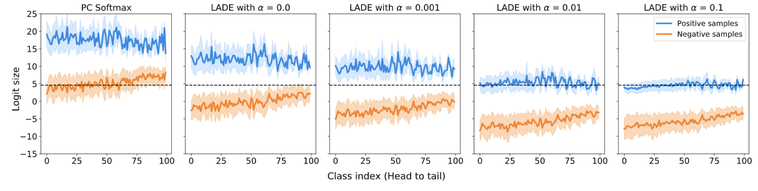

Figure 2: Logits of each class where the model is trained on CIFAR-100-LT. *positive samples* denote the sample corresponds to the class c, and *negative samples* denote the other.

By disentangling the source label distribution, the logit value $f_\theta(x)[y]$ should converge to $\log C$ for the positive samples. As α increases, the logit values gradually converge to the theoretical value, indicating that LADER successfully regularizes the logit values as we intended.