JUNE 18-22, 2023

CVPR VANCOUVER, CANADA

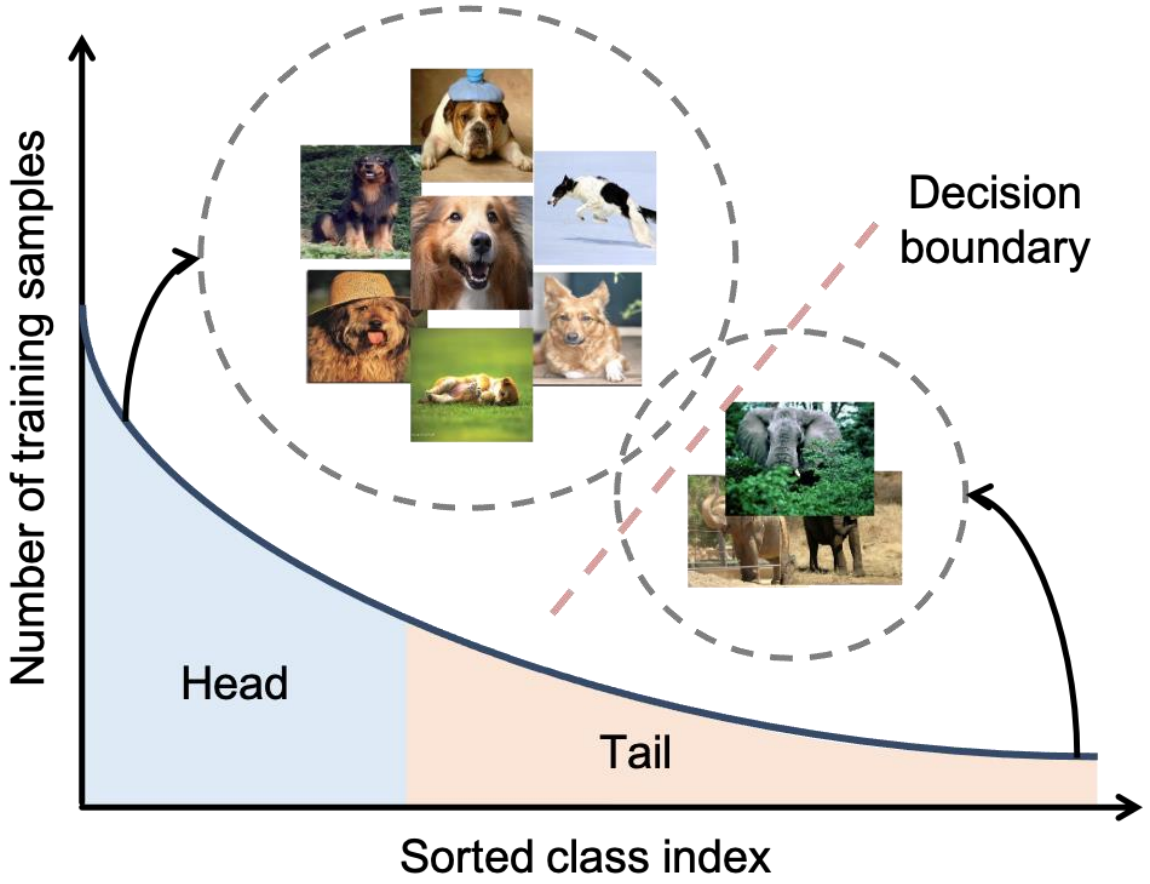
Class-Conditional Sharpness-Aware Minimization for Deep Long-Tailed Recognition

Zhipeng Zhou[1], Lanqing Li[2,4]*, Peilin Zhao[3], Pheng-Ann Heng[2], Wei Gong[1]*

[1]University of Science and Technology of China, [2]The Chinese University of Hong Kong,
[3]Tencent AI Lab, [4]Zhejiang Lab

Poster: # TUE-AM-333

Introduction



Long-tailed Class Distribution

Taxonomy of Long-tailed learning

- Borrow from [1]
- We use **DLTR** short for **D**eep **L**ong-**T**ailed **R**ecognition

[1] Deep Long-Tailed Learning: A Survey. IEEE TPAMI, 2022.

Introduction

- **Cost Sensitive Re-balancing**: LDAM-DRW[1](NeurIPS'19), BalancedSoftmax[2](NeurIPS'20)
  - Loss function
- **Augmentation-based Re-balancing**: M2m[3](CVPR'20), RSG[4](CVPR'21)
  - Data generation
- **Decoupling-based**: MiSLAS[5](CVPR'21), GCL[6](CVPR'22)
  - Representation

But seldom works research on parameter perturbation

[1] Learning Imbalanced Datasets with Label-Distribution-Aware Margin Loss. In NeurIPS 2019.

[2] Balanced Meta-Softmax for Long-Tailed Visual Recognition. In NeurIPS 2020.

[3] M2m: Imbalanced Classification via Major-to-minor Translation. In CVPR 2020.

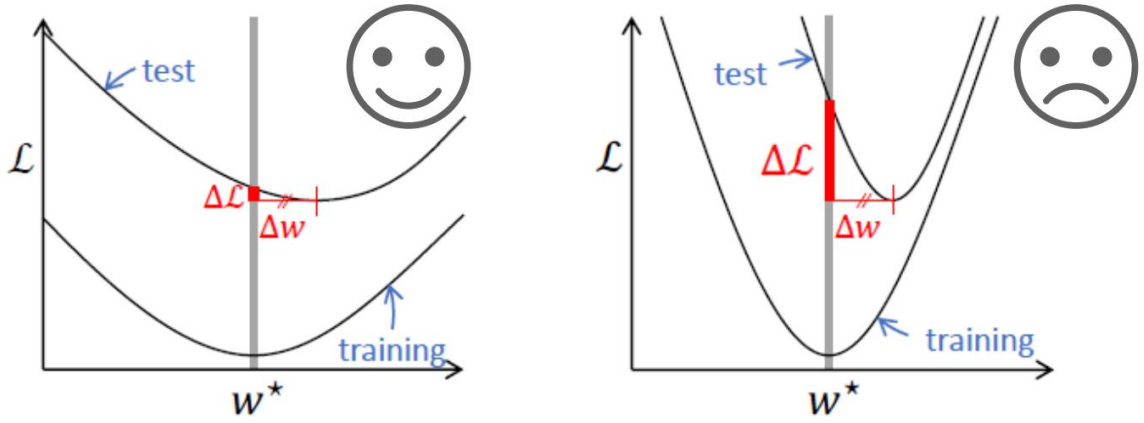
[4] RSG: A Simple but Effective Module for Learning Imbalanced Datasets. In CVPR 2021.

[5] Improving Calibration for Long-Tailed Recognition. In CVPR 2021.

[6] GCL: Long-tailed Visual Recognition via Gaussian Clouded Logit Adjustment. In CVPR 2022.

Introduction

Parameter Perturbation Solutions in General Optimization Problem



(a) Flatter minima (b) Sharper minima

A flatter minima usually indicates better generalization

- General Flattening
  - SWA[1] (NeurIPS'19), SAM[2] (ICLR'21), GPN[3] (ICML'22)
- Post Flattening
  - PoF[4] (ICML'22)
- Applications
  - FS-DGPM[5] (ICLR'21), F2M[6] (NuerIPS'21), rwSAM[7] (ICLR'22)

Continual Learning

Imbalanced Learning

[1] Fantastic generalization measures and where to find them. In NeurIPS 2019.
[2] Sharpness-aware minimization for efficiently improving generalization. In ICLR 2021.
[3] Penalizing Gradient Norm for Efficiently Improving Generalization in Deep Learning. In ICML 2022.
[4] PoF: Post-Training of Feature Extractor for Improving Generalization. In ICML 2022.
[5] Flattening Sharpness for Dynamic Gradient Projection Memory Benefits Continual Learning. In ICLR 2021.
[6] Overcoming Catastrophic Forgetting in Incremental Few-Shot Learning by Finding Flat Minima. In NeurIPS 2021.
[7] Self-supervised Learning is More Robust to Dataset Imbalance. In ICLR 2022.

Are current DLTR models have flat minima?

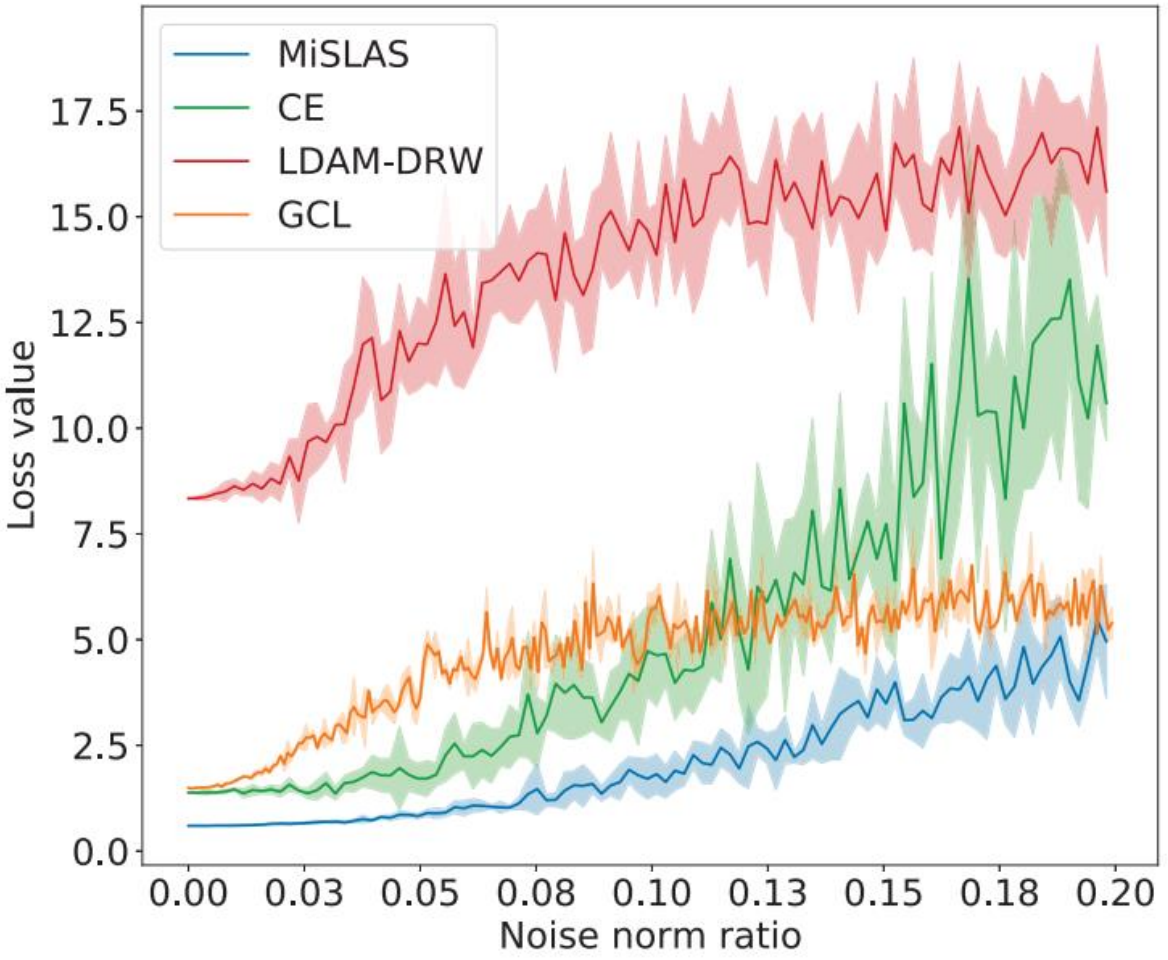
Motivation

- We investigate the local minima status of mainstream DLTR models:
  - ERM (CE), LDAM-DRW, MiSLAS, GCL

- We randomly perturbate the trained model **5** times and observe their vary loss values.

The loss function values of LDAM-DRW and GCL fluctuate more than CE.

LDAM-DRW and GCL have sharper minima than CE, which indicates a castle of DLTR models lack of flat minima.

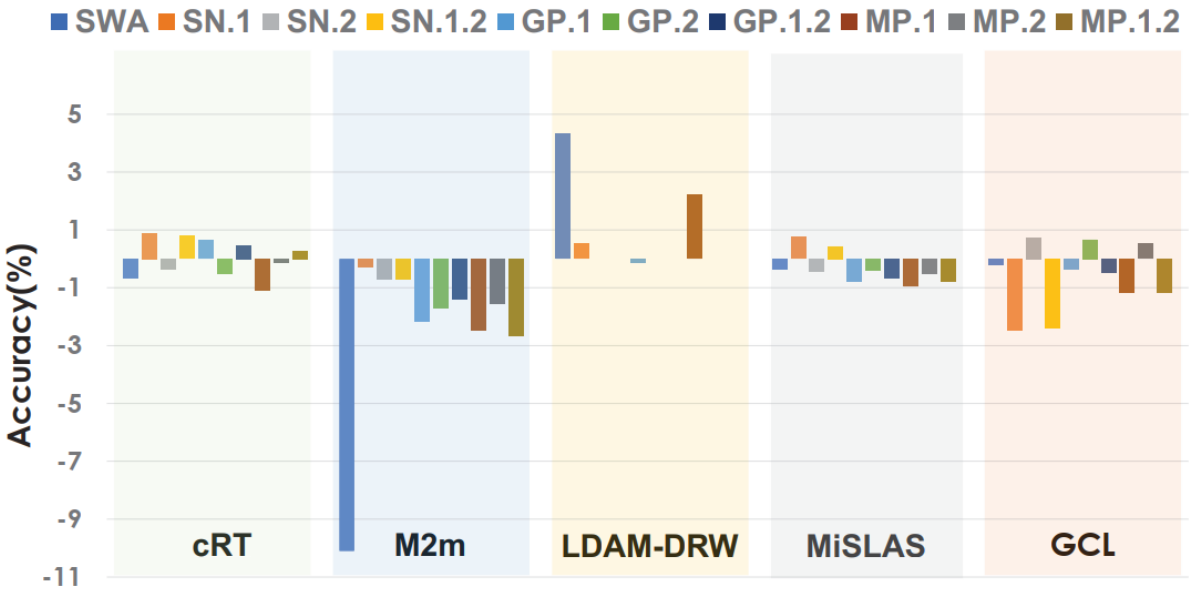


(b) 1D local loss landscape.

Motivation

Can current flattening operations help?

- We investigate **4** popular flattening operations:
  - SWA[1], Spectral Normalization (SN)[2], Gradient Penalization (GP)[3], Model Perturbation (MP)[4]



The results demonstrate that a naïve integration of general flattening operations with DLTR models can hardly bring improvements.

[1] Asymmetric valleys: Beyond sharp and flat local minima. In NeurIPS 2019.
[2] Large-scale gan training for high fidelity natural image synthesis . In ICLR 2019.
[3] Penalizing Gradient Norm for Efficiently Improving Generalization in Deep Learning. In ICML 2022.
[4] Overcoming Catastrophic Forgetting in Incremental Few-Shot Learning by Finding Flat Minima. In NeurIPS 2021.

Framework

A Close Look at the *Characteristic Radius* of Flat Minima:

Theorem 1: ***(Perturbative PAC-Bayesian Generalization Bound)***

$$L_T(\omega) \leq \max_{\|\epsilon\|_2 \leq \sqrt{k}\rho} L_S(\omega+\epsilon) + \sqrt{\frac{\frac{\|\omega\|_2^2}{4\rho^2} + \log\left(\frac{n}{\delta}\right) + \mathcal{O}(1)}{n-1}}$$

$$\hat{\epsilon}^*(\omega) \approx \sqrt{k}\rho^* \frac{\nabla_\omega L_S(\omega)}{\|\nabla_\omega L_S(\omega)\|_2}$$

$$\rho^* = \left(\frac{\|\omega\|_2}{2\|\nabla_\omega L_S(\omega)\|_2}\right)^{\frac{1}{2}} k^{-\frac{1}{4}}(n-1)^{-\frac{1}{4}}$$

1st-order Taylor Expansion

$$\max_{\|\epsilon\|_2 \leq \sqrt{k}\rho} L_S(\omega+\epsilon) \approx \max_{\|\epsilon\|_2 \leq \sqrt{k}\rho} [L_S(\omega) + \epsilon^{\mathrm{T}}\nabla_\omega L_S(\omega)]$$

$$= L_S(\omega) + \sqrt{k}\rho\|\nabla_\omega L_S(\omega)\|_2$$

Approximated optimal generalization bound:

$$\hat{L}_T(\omega) \approx \max_{\|\epsilon\|_2 \leq \sqrt{k}\rho^*} L_S(\omega+\epsilon) + \frac{1}{2\sqrt{n-1}}\frac{\|\omega\|_2}{\rho^*}$$

More details please refer to our paper!

Framework

A Close Look at the *Characteristic Radius* of Flat Minima:

Theorem 1: ***(Perturbative PAC-Bayesian Generalization Bound)***

$$L_T(\omega) \leq \max_{\|\epsilon\|_2 \leq \sqrt{k}\rho} L_S(\omega+\epsilon) + \sqrt{\frac{\frac{\|\omega\|_2^2}{4\rho^2} + \log\left(\frac{n}{\delta}\right) + \mathcal{O}(1)}{n-1}}$$

$$\hat{\epsilon}^*(\omega) \approx \sqrt{k}\rho^* \frac{\nabla_\omega L_S(\omega)}{\|\nabla_\omega L_S(\omega)\|_2}$$

$$\rho^* = \left(\frac{\|\omega\|_2}{2\|\nabla_\omega L_S(\omega)\|_2}\right)^{\frac{1}{2}} k^{-\frac{1}{4}}(n-1)^{-\frac{1}{4}}$$

In long-tailed problem:

$\boldsymbol{p_s(y) \neq p_t(y)}$ ✗

$\boldsymbol{p_s(x|y) = p_t(x|y)}$ ✓

Class-Conditional Perturbation:

$$\hat{\epsilon}_c{}^*(\omega) \approx \sqrt{k}\rho_c^* \frac{\nabla_\omega L_S^c(\omega)}{\|\nabla_\omega L_S^c(\omega)\|_2}$$

$$\rho_c^* = \left(\frac{\|\omega\|_2}{2\|\nabla_\omega L_S^c(\omega)\|_2}\right)^{\frac{1}{2}} k^{-\frac{1}{4}}(n_c-1)^{-\frac{1}{4}}$$

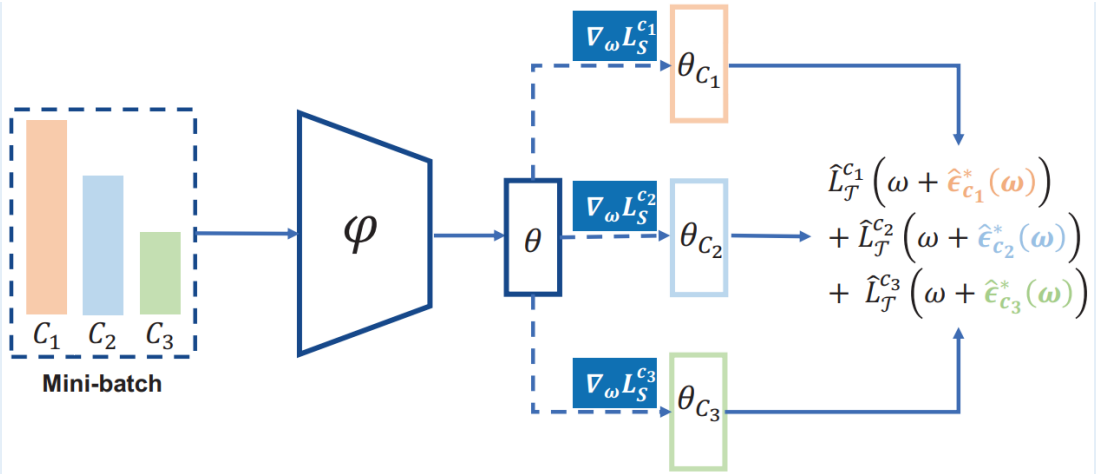
Framework

Stage 1

$$\hat{\epsilon}_c{}^*(\omega) \approx \sqrt{k}\rho_c^* \frac{\nabla_\omega L_S^c(\omega)}{\|\nabla_\omega L_S^c(\omega)\|_2}$$

$$\omega \leftarrow \omega - \eta \nabla_\omega L_S^{\text{CC-SAM}}(\omega)$$

$$\approx \omega - \eta \sum_{c=1}^{k} \nabla_\omega \hat{L}_T^c(\omega)\Big|_{\omega+\hat{\epsilon}_c^*(\omega)}$$

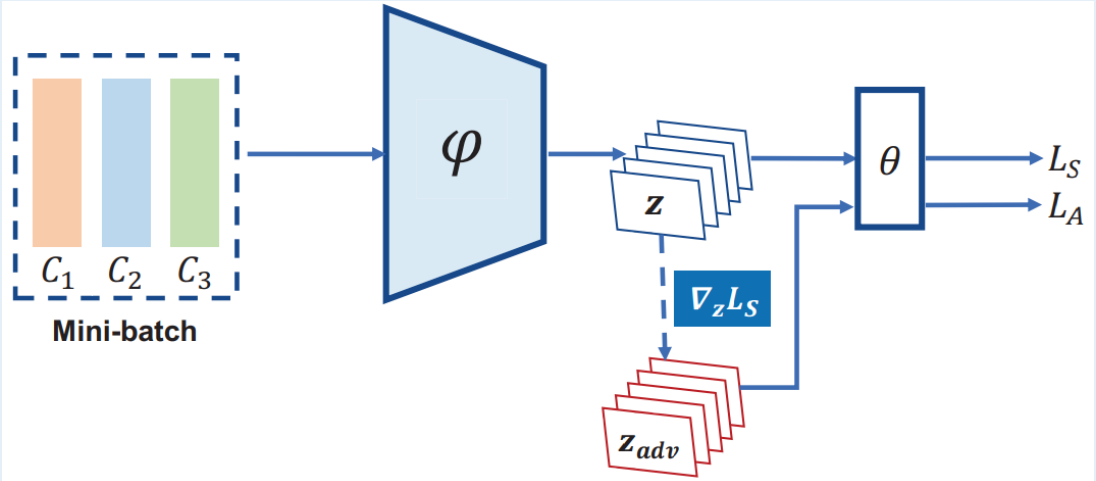


(a) **Stage 1**: Class-conditional sharpness-aware minimization. We take three classes and classifier perturbation as examples for illustration.

Stage 2

$$z_{adv} \approx z + \lambda \frac{\nabla_z L_S(z;\theta)}{\|\nabla_z L_S(z;\theta)\|_2}$$

$$L = \left(1 - \frac{t}{T}\right) L_S + \frac{t}{T} L_A, L_A = h(z_{adv};\theta)$$



(b) **Stage 2**: Robust training of the classifier by progressively generating adversarial features.

Evaluation

	CIFAR-10-LT			CIFAR-100-LT		
Imbalance Ratio	200	100	50	200	100	50
CE	65.68	70.70	74.81	34.84	38.43	43.90
CE + Mixup [57]	65.84	72.96	79.48	35.84	40.01	45.16
LDAM-DRW [7]	73.52	77.03	81.03	38.91	42.04	47.62
De-confound-TDE [45]	-	80.60	83.60	-	44.15	50.31
CE + Mixup + cRT [20]	73.06	79.15	84.21	41.73	45.12	50.86
BBN [63]	73.47	79.82	81.18	37.21	42.56	47.02
Contrastive Learning [51]	-	81.40	85.36	-	46.72	51.87
BGP [49]	-	-	-	41.20	45.20	50.50
MiSLAS [62]	77.31	82.06	85.16	42.33	47.50	52.62
VS + SAM [38]	-	82.40	-	-	46.60	-
GCL [26]	79.03	82.68	85.46	44.88	48.71	53.55
CC-SAM	**80.94**	**83.92**	**86.22**	**45.66**	**50.83**	**53.91**

CC-SAM achieves the state-of-the-art performance

Evaluation

Dataset	Method	Backbone	Many	Medium	Few	Overall
ImageNet-LT	CE	ResNeXt-50	**65.9**	37.5	7.7	44.4
	Decouple-τ-norm [20]	ResNet-50	56.6	44.2	27.4	46.7
	Balanced Softmax [40]	ResNeXt-50	64.1	48.2	33.4	52.3
	LADE [17]	ResNeXt-50	64.4	47.7	34.3	52.3
	RSG [50]	ResNeXt-50	63.2	48.2	32.2	51.8
	DisAlign [58]	ResNet-50	61.3	52.2	31.4	52.9
		ResNeXt-50	62.7	52.1	31.4	53.4
	ResLT [9]	ResNeXt-50	63.0	53.3	35.5	52.9
	BGP [49]	ResNet-50	-	-	-	51.5
	MiSLAS [62]	ResNet-50	-	-	-	52.7
	LDAM-DRW + SAM [38]	ResNet-50	62.0	52.1	34.8	53.1
	GCL [26]	ResNet-50	-	-	-	54.9
	CC-SAM	ResNet-50	61.4	49.5	37.1	52.4
		ResNeXt-50	63.1	**53.4**	**41.1**	**55.4**
Places-LT	CE	ResNet-152	**45.7**	27.3	8.2	30.2
	Decouple-τ-norm [20]	ResNet-152	37.8	40.7	31.8	37.9
	Balanced Softmax [40]	ResNet-152	42.0	39.3	30.5	38.6
	LADE [17]	ResNet-152	42.8	39.0	31.2	38.8
	RSG [50]	ResNet-152	41.9	41.4	32.0	39.3
	DisAlign [58]	ResNet-152	40.4	42.4	30.1	39.3
	ResLT [9]	ResNet-152	39.8	**43.6**	31.4	39.8
	MiSLAS [62]	ResNet-152	-	-	-	40.2
	GCL [26]	ResNet-152	-	-	-	40.6
	CC-SAM	ResNet-152	41.2	42.1	**36.4**	**40.6**
iNaturalist 2018	CE	ResNet-50	**72.2**	63.0	57.2	61.7
	Decouple-τ-norm [20]	ResNet-50	65.6	65.3	65.9	65.6
	Balanced Softmax [40]	ResNet-50	-	-	-	70.6
	LADE [17]	ResNet-50	-	-	-	70.0
	RSG [50]	ResNet-50	-	-	-	70.3
	DisAlign [58]	ResNet-50	-	-	-	70.6
	ResLT [9]	ResNet-50	-	-	-	70.2
	BGP [49]	ResNet-50	70.0	69.9	69.6	70.5
	MiSLAS [62]	ResNet-50	-	-	-	71.6
	LDAM-DRW + SAM [38]	ResNet-50	64.1	70.5	71.2	70.1
	GCL [26]	ResNet-50	-	-	-	**72.0**
	CC-SAM	ResNet-50	65.4	**70.9**	**72.2**	70.9

- CC-SAM shows competitive on large scale datasets.
- CC-SAM generally improves medium and tail classes.
- As a theoretical motivated variant, CC-SAM outperforms the naïve application of SAM, which demonstrates the superiority of class-conditional perturbation.

Evaluation

	ImageNet-LT				Places-LT			
Method	Many	Medium	Few	F-measure	Many	Medium	Few	F-measure
CE	40.1	10.4	0.4	0.295	45.9	22.4	0.4	0.366
Lifted Loss [36]	34.8	29.3	17.4	0.374	41.0	35.2	23.8	0.459
Focal Loss [28]	35.7	29.3	15.6	0.371	41.0	34.8	22.3	0.453
Range Loss [59]	34.7	29.4	17.2	0.373	41.0	35.3	23.1	0.457
OpenMax [3]	35.8	30.0	17.6	0.368	41.1	35.4	23.2	0.458
OLTR [32]	41.9	33.9	17.4	0.474	44.6	36.8	25.2	0.464
IEM [65]	46.1	42.3	20.1	0.525	48.8	42.4	28.9	0.486
LUNA [5]	48.2	44.7	23.6	**0.579**	48.1	41.6	29.0	0.491
CC-SAM	61.4	49.5	37.1	0.552	41.2	41.8	36.4	**0.510**

OLTR evaluation[1] demonstrates CC-SAM learns a robust representation

[1] Open Long-Tailed Recognition In A Dynamic World. IEEE TPAMI, 2022.

Micro Benchmark

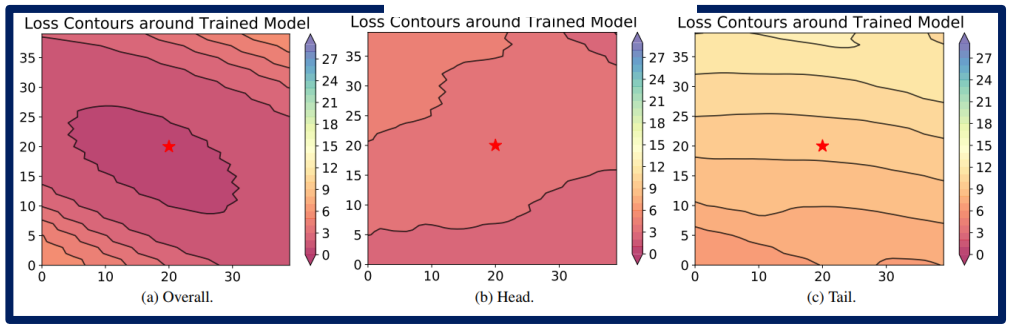
cRT	Stage 1 + dir	Stage 1 + mag	Stage 2	Acc
✔				37.8
✔	✔			38.9
✔		✔		37.5
✔	✔	✔		40.1
✔	✔	✔	✔	**40.6**

Both magnitude and direction contribute to the performance, which demonstrates the superiority of our derived class-conditional perturbation.

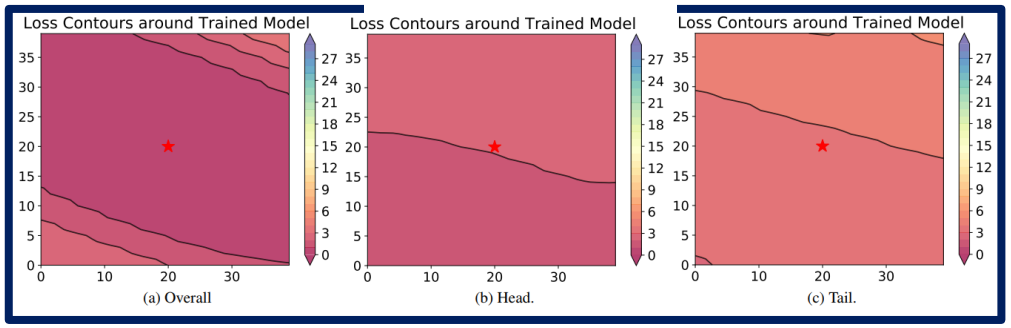
- **Stage 1 + dir**: enforce parameter perturbation along the recommended direction with magnitude of 1.
- **Stage 1 + mag**: enforce perturbation with the recommended magnitude in a random direction.

Visualization

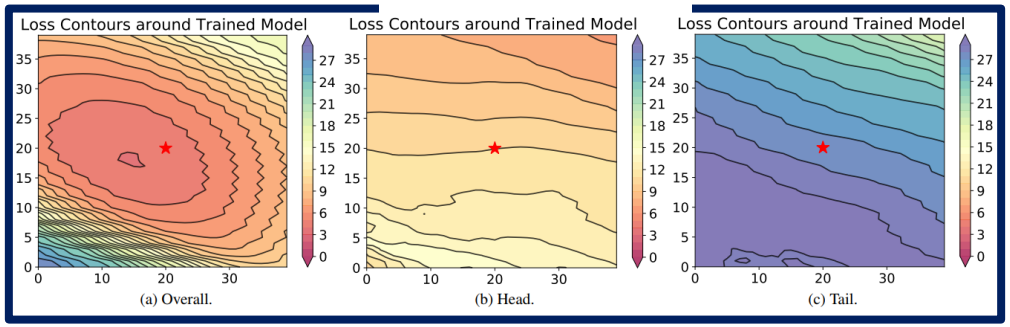
CE



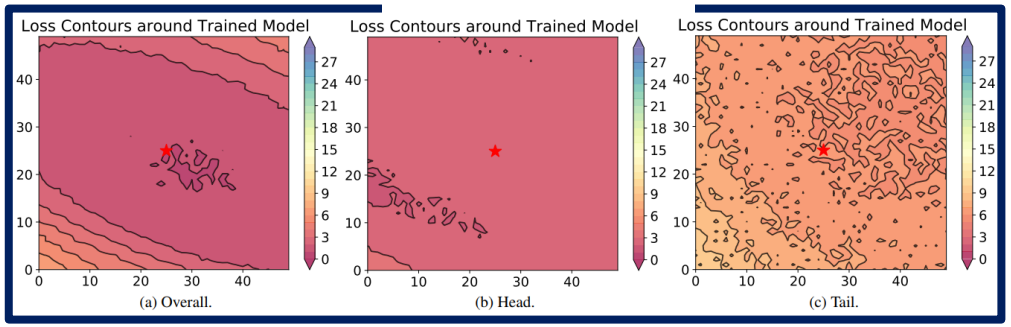
MiSLAS



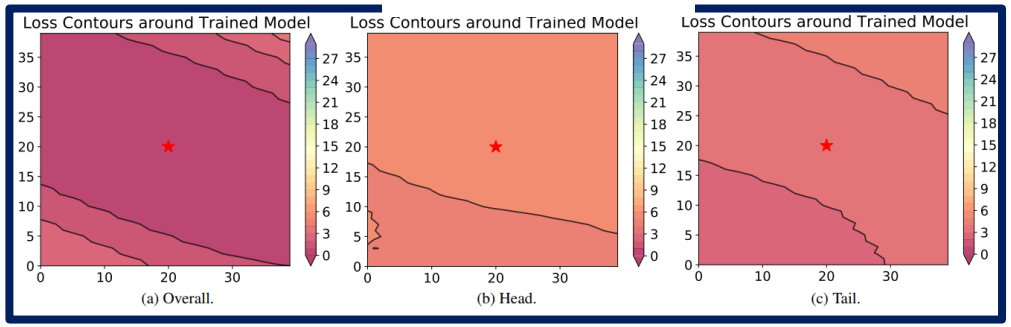
LDAM-DRW



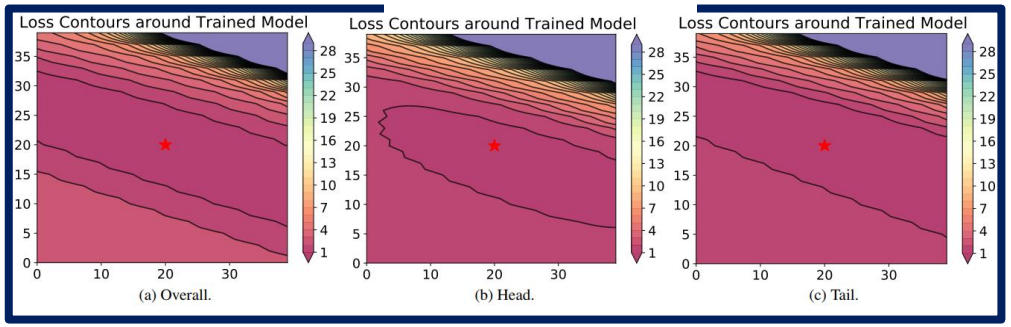
GCL



M2m



CCSAM



Thanks for your listening!

Contact Information

Feel free to contact us if you are interested in our work!

zzp1994@mail.ustc.edu.cn, lanqingli1993@gmail.com

github.com/zzpustc/CC-SAM

Poster: # TUE-AM-333