Deep Long-Tailed Data Learning towards Visual Recognition

Yiu-ming Cheung

Department of Computer Science

Hong Kong Baptist University, Hong Kong SAR, China

香港浸會大學
HONG KONG BAPTIST UNIVERSITY

DEPARTMENT OF
COMPUTER SCIENCE
計算機科學系

September 29, 2024

Outline

I. Introduction

II. Related Works

III. Recent Research Progress

IV. Future Directions and Conclusion

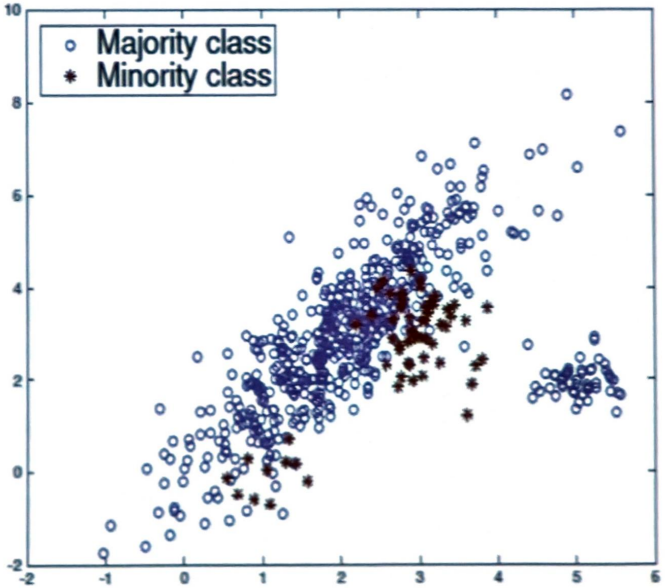
Introduction: Imbalanced Data Learning

1. Class Imbalance Learning
2. Long-tailed Data Classification

香港浸會大學
HONG KONG BAPTIST UNIVERSITY

DEPARTMENT OF
COMPUTER SCIENCE
計算機科學系

What is class imbalance problem?



In classification, class imbalance refers to that the number of data in the majority class is **MUCH MORE** than the number of data in the minority class.

Example 1: Cancer Diagnosis

- The number of cancerous patients is only a small proportion, compared with healthy people.
- During diagnosis, our target is to identify the cancerous patients from healthy people.

Image source: http://www.nccrm.com/wp-content/uploads/2013/10/cancer-diagnosis.jpg

Example 2: Detecting Malicious Behaviors in **Internet**

- Find "**intruders**" in traffic data and identify new network intrusion patterns:

Majority Class: Normal Behavior

Minority Class: Malicious Behavior

- More examples...

Credit Card Fraud Detection

Image source: https://kikinote.net/156014

Anomaly Detection of Biological Data

Image Source:
https://news.pharmacodia.com/web/informationMobileController/getMobileNewInformationById?id=8a2d983760074bc001600a54986a0283
https://zhuanlan.zhihu.com/p/387679170

Industrial Fault Detection

In most of class imbalance situation, the minority class is usually more important, and receives more attention.

Misclassification Cost

In the above examples, the misclassification cost is different.

		Ground Truth	
Prediction		Positive (sick)	Negative (healthy)
	Positive (sick)	True Positive (TP)	False Positive (FP)
	Negative (healthy)	False Negative (FN)	True Negative (TN)

False Positive, misdiagnose a healthy person to be sick

- mental stress
- more payments for further diagnosis

False Negative, misdiagnose a sick person to be health

- treatment delay
- death

The cost of False Negative is usually higher than False Positive.

Measurement

- Classifier is usually evaluated by classification accuracy.
- 10 cancerous patients in 10,000 people:
  - A diagnosis machine just simply predicts all people as health.
  - Is it fair to claim that the machine has 99.9% accuracy?
- Accuracy will produce misleading results on imbalanced data.
- Special measurement should be adopted to assess the performance of classifiers.

Measurement

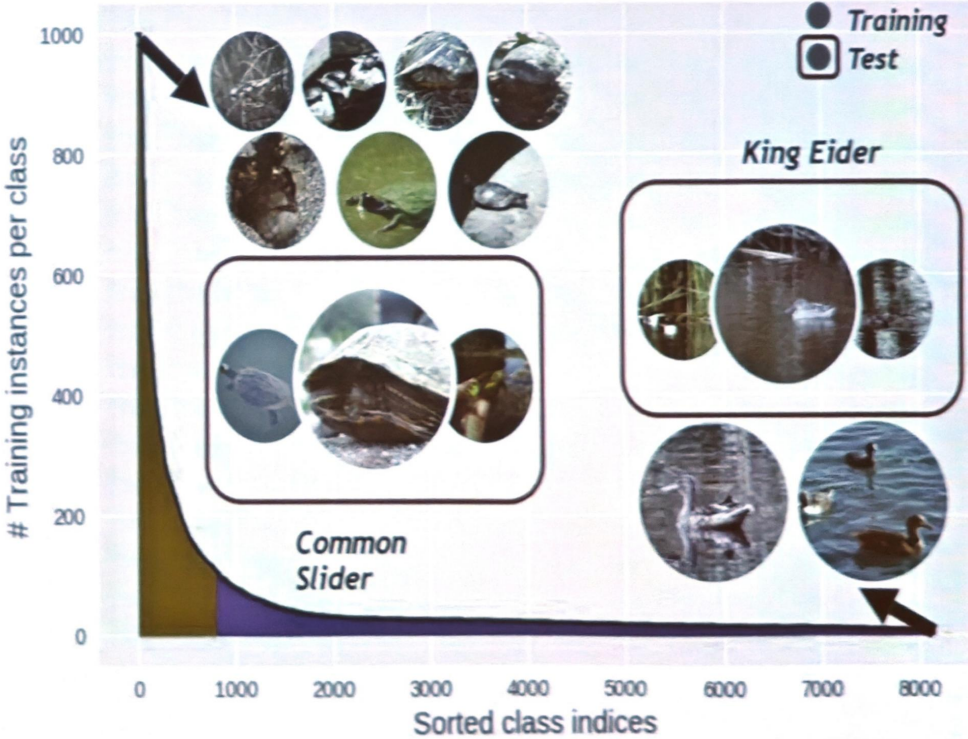
- Composite measurements:
  - F1 score: $F1 = \frac{2 \cdot Recall \cdot Precision}{Recall + Precision}$
    - Harmonic average of the precision and recall.
    - Usually used information retrieval, document classification and query classification.
  - Geometric mean (G-Mean): $Gmean = \sqrt{\frac{TP}{TP+FN} + \frac{TN}{TN+FP}}$
    - Harmonic average of the TPR (TP Rate) and TNR (TN Rate).
    - Supplement of F1 measure, consider true negatives.
  - Receiver Operating Characteristic curve (ROC curve)
    - Visual representation of the trade-off between FPR and TPR.
    - Numerically calculated by Area Under Curve (AUC).

Why Study Long-tailed Data?

- Real data often has **long-tail distribution.**

  Wang et al. first mentioned Long-tailed data for deep learning in 2017 [1].

- The differences from traditional imbalanced data are:
  - Huge number of categories;
  - Sample numbers in each class follows power law;
  - Mainly use deep models to address.



The training set of iNaturalist 2018.

(Image source: Jamal et al in *CVPR* 2020)

[1] Wang et al. Learning to Model the Tail, in NIPS, 2017.

DEPARTMENT OF COMPUTER SCIENCE 計算機科學系
香港浸會大學 HONG KONG BAPTIST UNIVERSITY

Why Study Long-tailed Data?

Tail classes are important.

Example: disease diagnosis

- Rare diseases have wide variety but with few samples compared with common diseases and healthy people;
- Identifying rare diseases from common disease and healthy people is more important.

(Image source: https://m.sohu.com/a/160373651_654956?_f=m-article_24_feeds_14#read)

DEPARTMENT OF COMPUTER SCIENCE 計算機科學系
香港浸會大學 HONG KONG BAPTIST UNIVERSITY

Why Study Long-tailed Data?

Success of deep learning

(Image source: https://image-net.org)

Visual Object Classes Challenge 2012 (VOC2012)

PASCAL2
Pattern Analysis, Statistical Modelling and Computational Learning

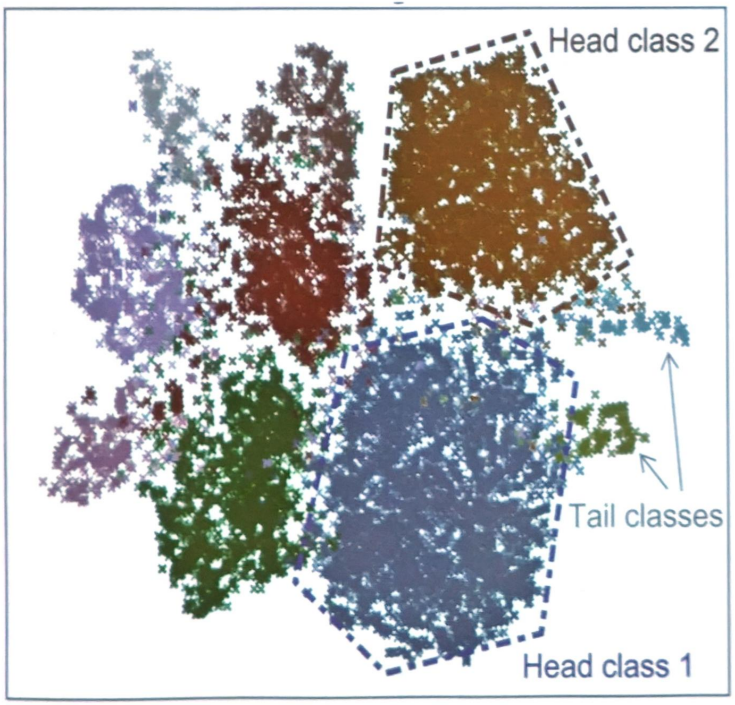
(Image source: http://host.robots.ox.ac.uk/pascal/VOC/voc2012/index.html)

- Large-scale
- Clean label
- Balanced

DEPARTMENT OF COMPUTER SCIENCE 計算機科學系
香港浸會大學 HONG KONG BAPTIST UNIVERSITY

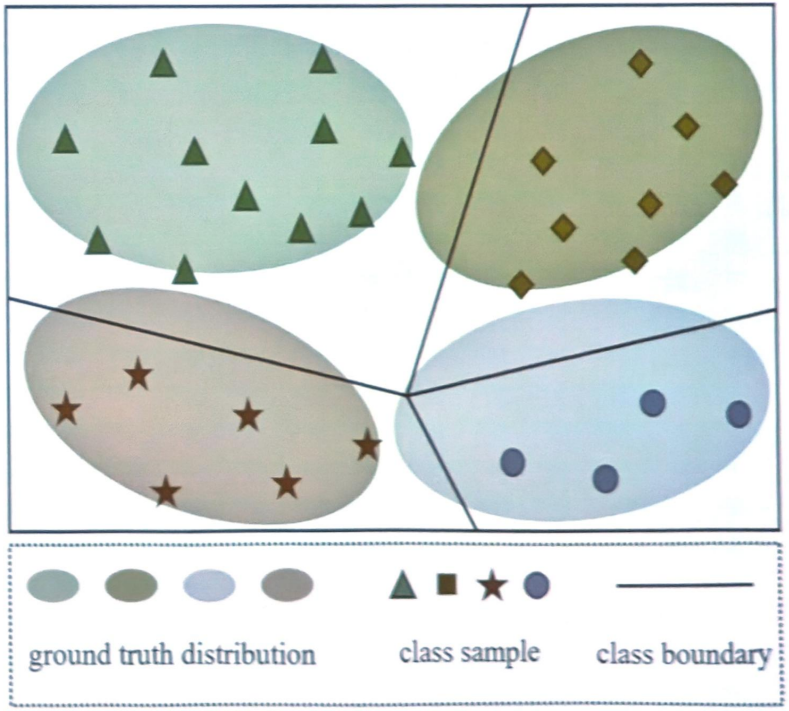
Why Study Long-tailed Data?

Challenges for deep models on long-tailed data:



(Image source: Yang *et al.* in *NeurIPS* 2020)

- Distorted embedding space

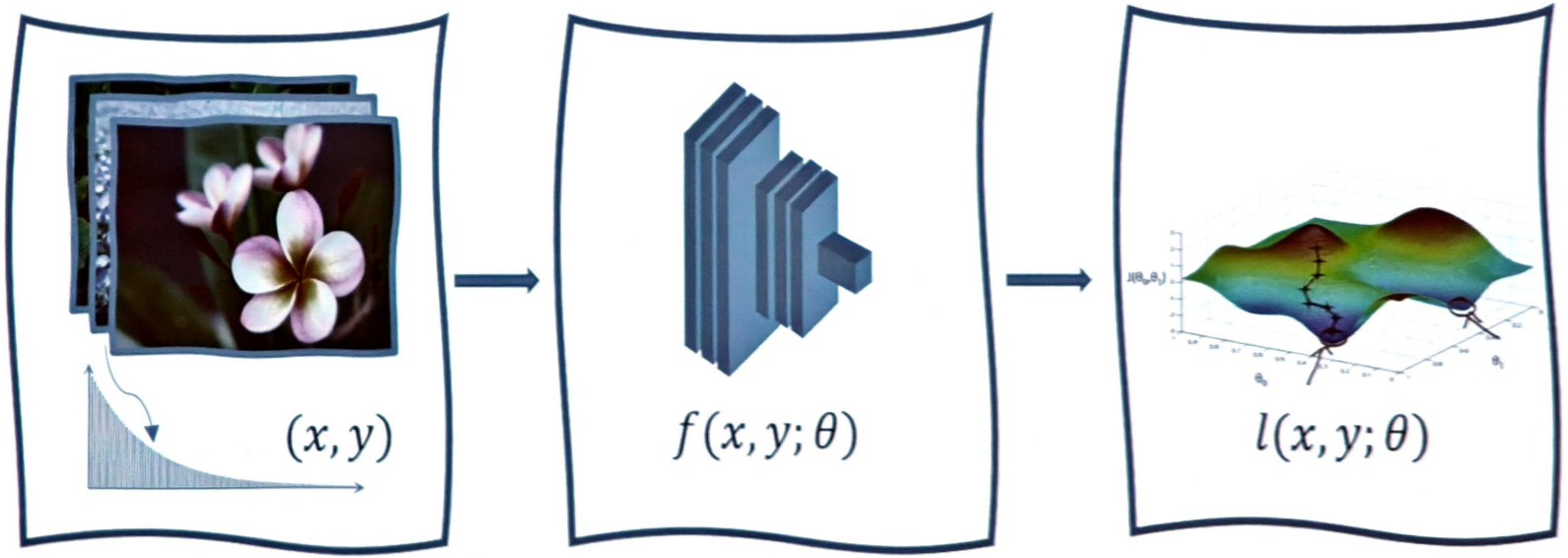


(Image source: Li *et al.* in *CVPR* 2022)

- Biased classifier

II.(i) Related Works

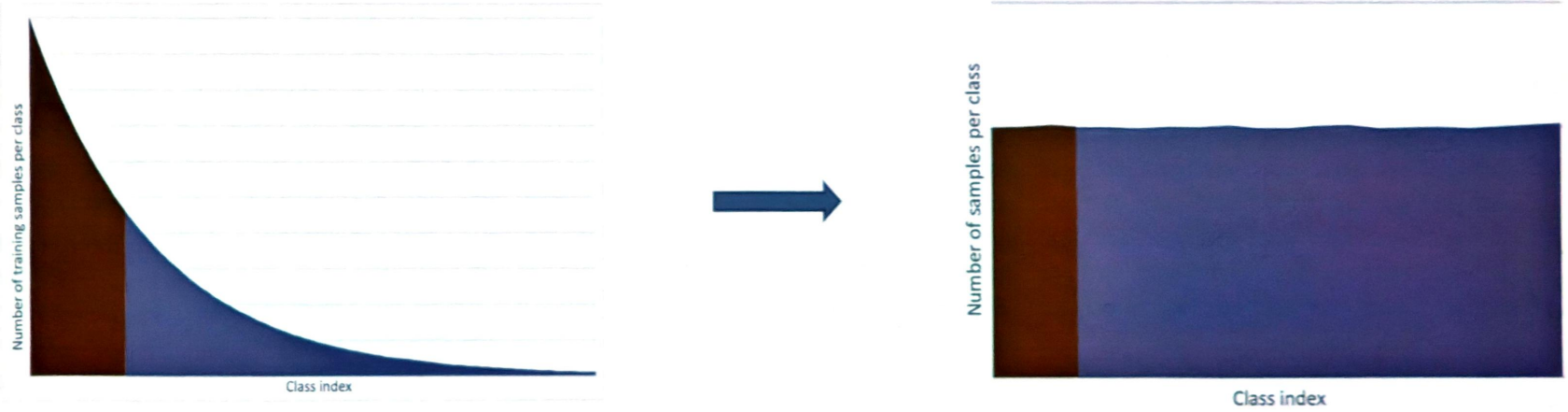
Basic process of deep learning



Data Model Loss function

II.(i) Related Works

Re-sampling



Make the number of samples of each class balanced via sampling.

- Random over-sampling [2]; • Random under-sampling [3]; • Hybrid-sampling [4], etc.

Can cooperate with data augmentation.

[2] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: synthetic minority over-sampling technique," *JAIR*, vol. 16, pp. 321–357, 2002.
[3] Tomek, "Two modifications of CNN," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 6, pp. 769–772, 1976.
[4] G. E. A. P. A. Batista, A. L. C. Bazzan, and M. C. Monard, "Balancing training data for automated annotation of keywords: A case study," in *WOB*, Macaé, RJ, Brazil, Nov. 2003, pp. 10–18.

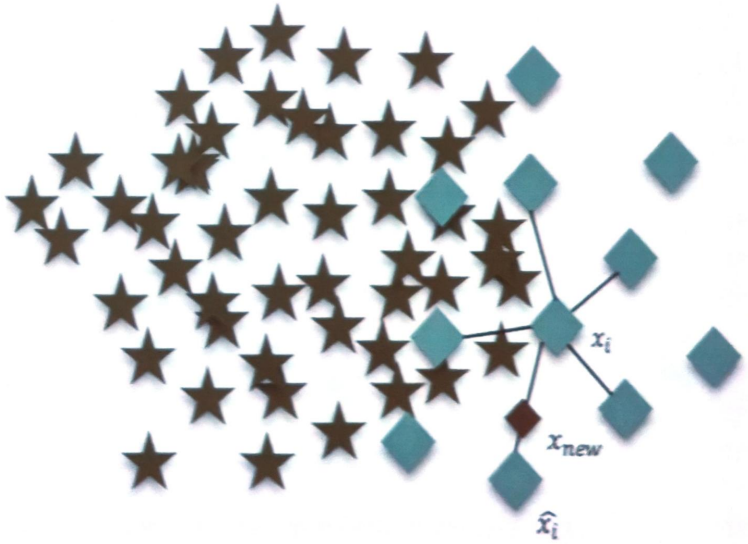
An Example: Synthetic Minority Oversampling Technique (SMOTE)

- SMOTE synthesizes new data points into the minority class by interpolating in the neighborhood of minority class data.

$$x_{new} = x_i + (\hat{x_i} - x_i)\delta$$

x_i is the minority class data and $\hat{x_i}$ is one of its k-nearest neighbors. δ is a random number in [0,1].

  - More noises will be synthesized if the selected minority class data is a noise in the area of majority class.

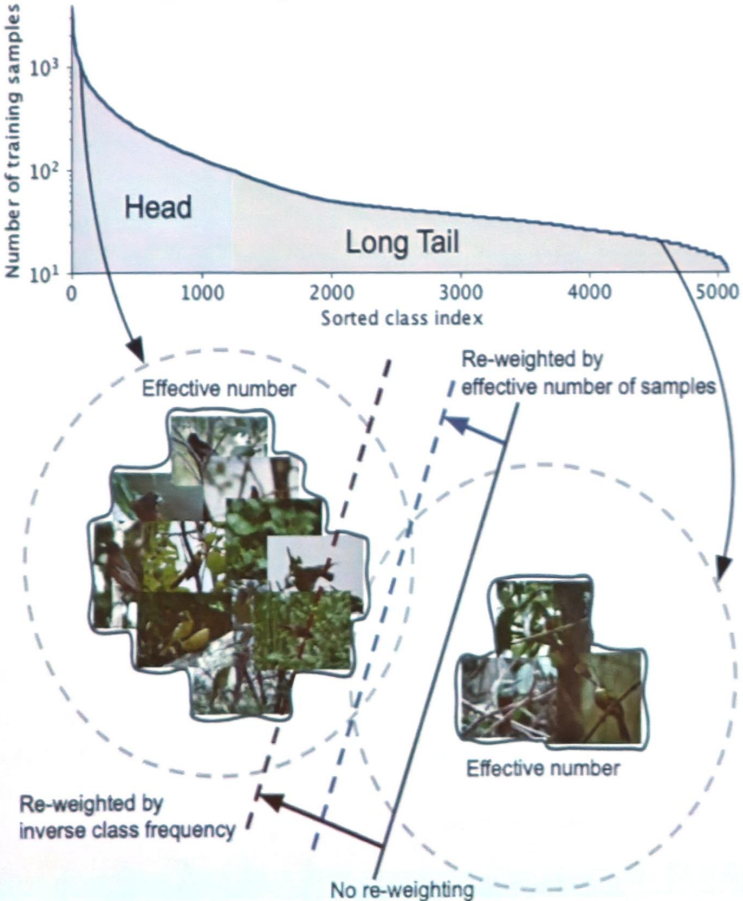


Chawla, Nitesh V., et al. "SMOTE: synthetic minority over-sampling technique." *Journal of artificial intelligence research* 16 (2002): 321-357.

II.(i) Related Works

Re-weighting: *Class-balance loss*

Class-Balanced Loss Based on Effective Number of Samples, Y. Cui, M Jia, TY Lin, Y Song, in *CVPR* 2019.



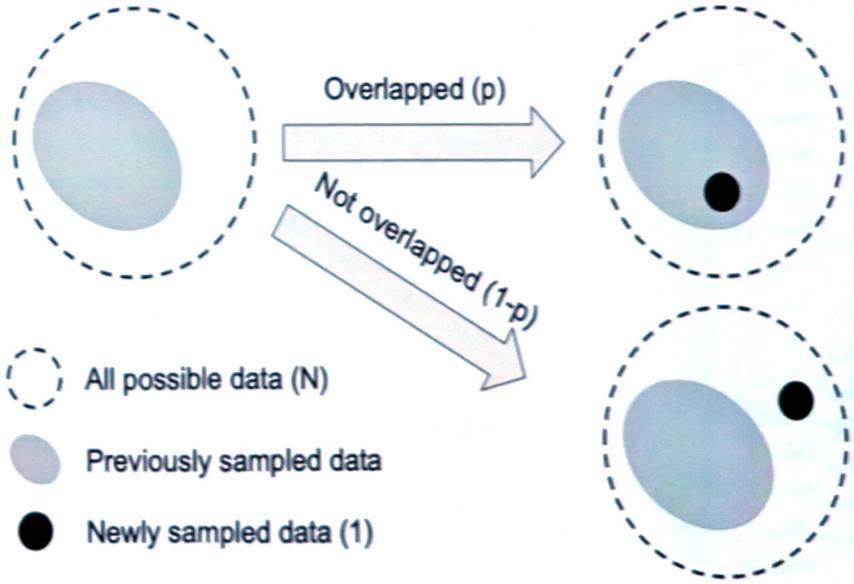
- Observation: Set w_y as inverse class frequency yields poor performance on the dataset with high class imbalance.

II.(i) Related Works

Re-weighting: *Class-balance loss*

Class-Balanced Loss Based on Effective Number of Samples, Y. Cui, M Jia, TY Lin, Y Song, in *CVPR* 2019.

- Motivation: as the number of samples increases, the additional benefit of a newly added data point will diminish.
- Method: treat data sampling process as a random covering. Then, data overlap can be measured by associating each sample with a small neighboring region instead of a single point.



II.(i) Related Works

Re-weighting: *Class-balance loss*

Class-Balanced Loss Based on Effective Number of Samples, Y. Cui, M Jia, TY Lin, Y Song, in *CVPR* 2019.

- **Effective number of samples** (E_{n_i}) is the expected volume of samples:

$$E_{n_i} = \frac{1-\beta^{n_i}}{1-\beta},$$

where $\beta \in [0,1]$ is a hyperparameters that controls how fast E_{n_i} grows.

- The class weight should be the inverse of effective number. Loss function:

$$L_{cb} = -\frac{1}{E_{n_i}}\log(p_y)$$

II.(i) Related Works

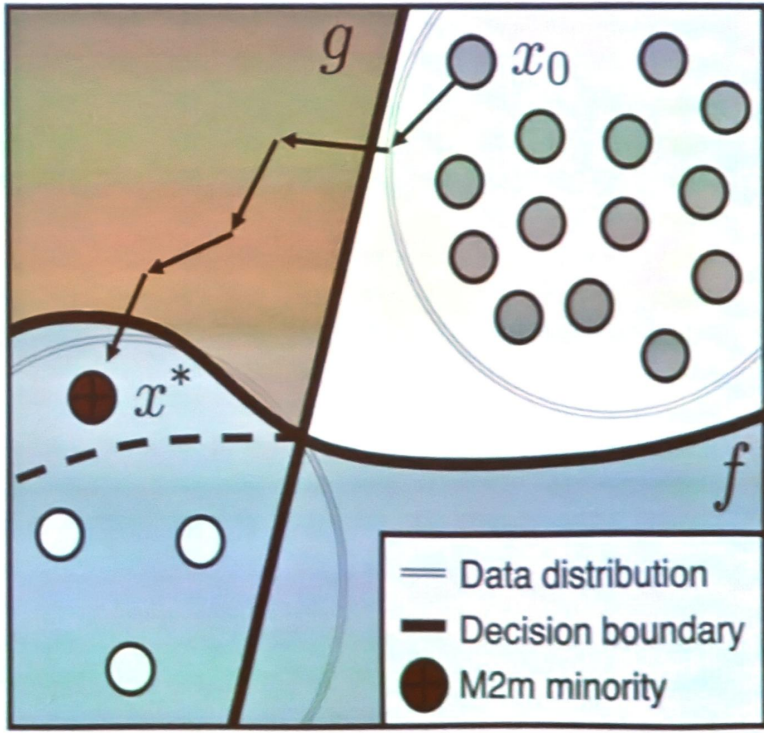
Augmentation: *M2m*

M2m: Imbalanced Classification via Major-to-minor Translation, J. Kim, H. Jeong, J. Shin, in *CVPR* 2020.

- Objective:

$$x^* = arg \min_{x:=x_0+\delta} -g(x; y_t) + \lambda f(x; y_s)$$

- Generated x^* satisifies:
  - classified into target tail class y_t by g
  - not classified into the source class y_s by f

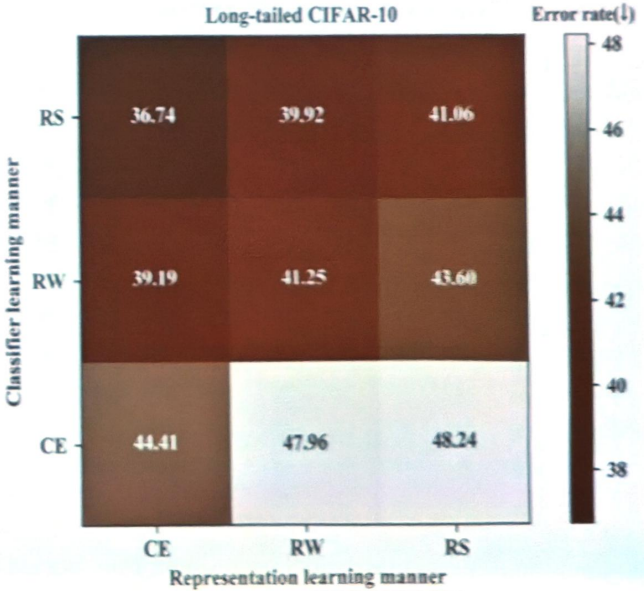


II.(i) Related Works

Decoupling: *BBN*

BBN: Bilateral-Branch Network with Cumulative Learning for Long-Tailed Visual Recognition, B. Zhou, Q. Cui, X.-S. Wei, Z.-M. Chen, in *CVPR* 2020.

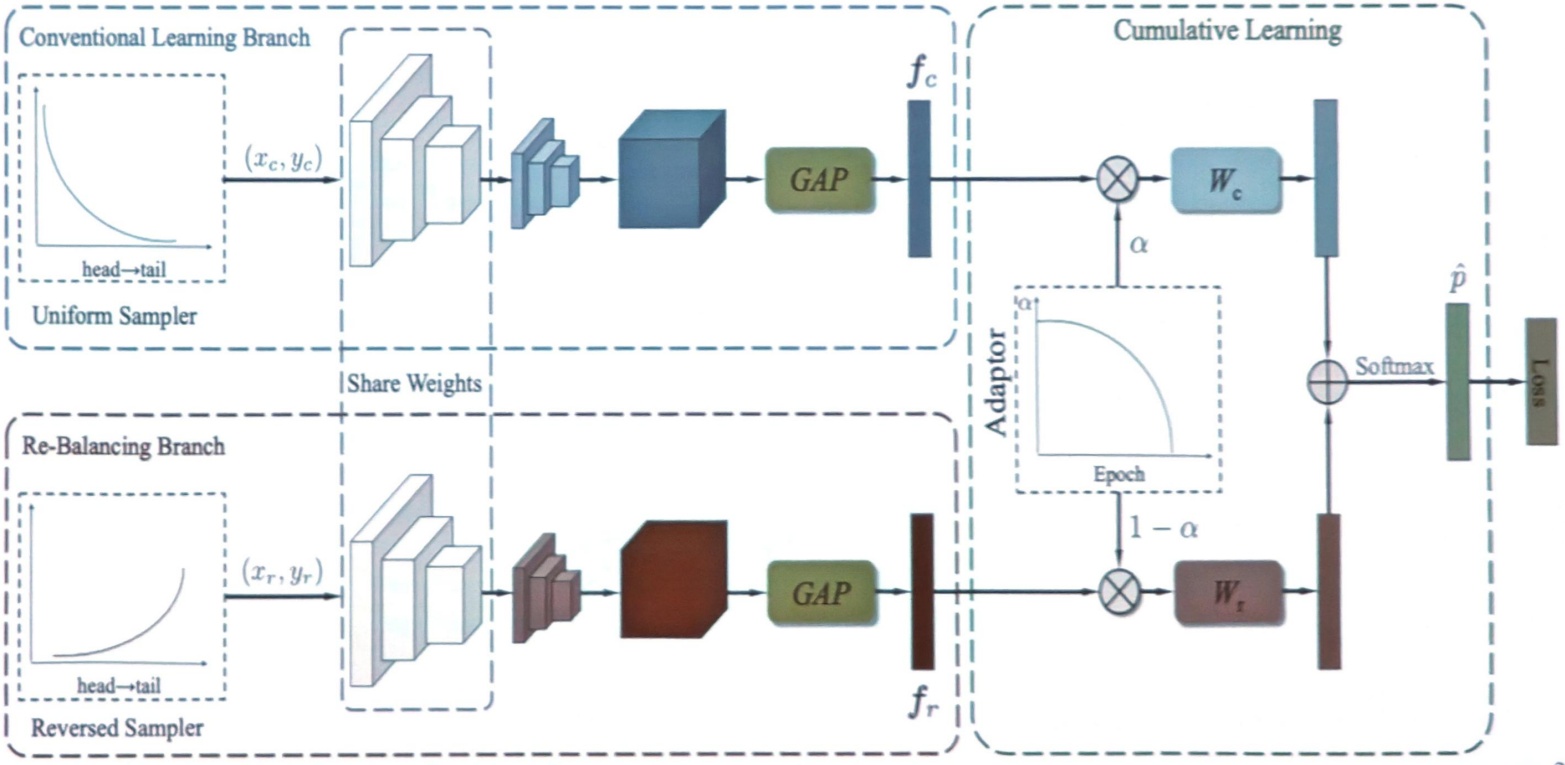
- Deep learning contains 1. Representation and 2. Classifier.
- A question: How does re-sampling/re-weighting affect representation learning and classifier learning?
- Observation:
  - Representation learning on the original long-tailed data obtains the best performance.
  - Classifier learning with RS obtains the best performance.
  - Only classifier learning can be beneficial by RS/RW .



II.(i) Related Works

Decoupling: *BBN*

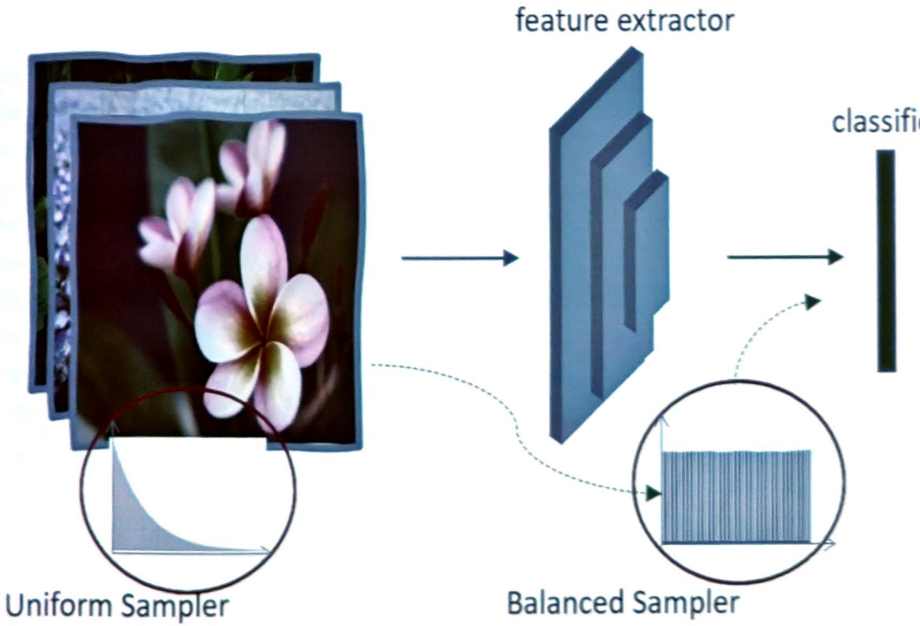
BBN: Bilateral-Branch Network with Cumulative Learning for Long-Tailed Visual Recognition, B. Zhou, Q. Cui, X.-S. Wei, Z.-M. Chen, in *CVPR* 2020.



II.(i) Related Works

Decoupling: *Decoupling representation*

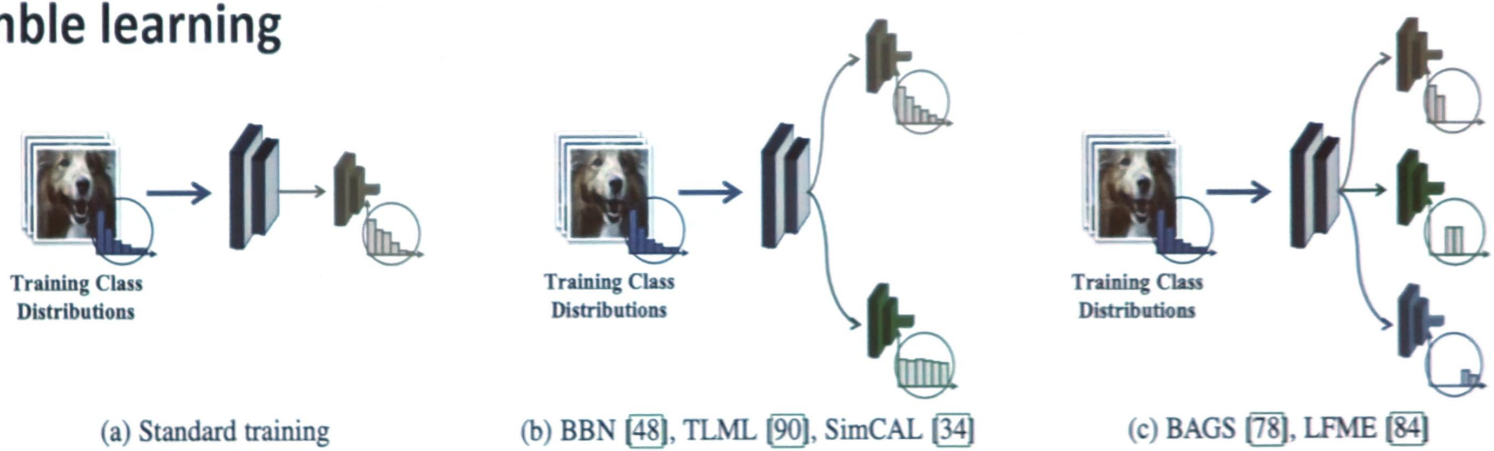
Decoupling representation and classifier for long-tailed recognition, B. Kang, S. Xie, et al, in *ICLR* 2020.



- Stage I: use CE on original long-tailed data to obtain a good feature extractor.
- Stage II: fix the feature extractor, use balanced sampled data to obtain an unbiased classifier.

II.(i) Related Works

Ensemble learning



(a) Standard training

(b) BBN [48], TLML [90], SimCAL [34]

(c) BAGS [78], LFME [84]

(d) ACE [104], ResLT [152]

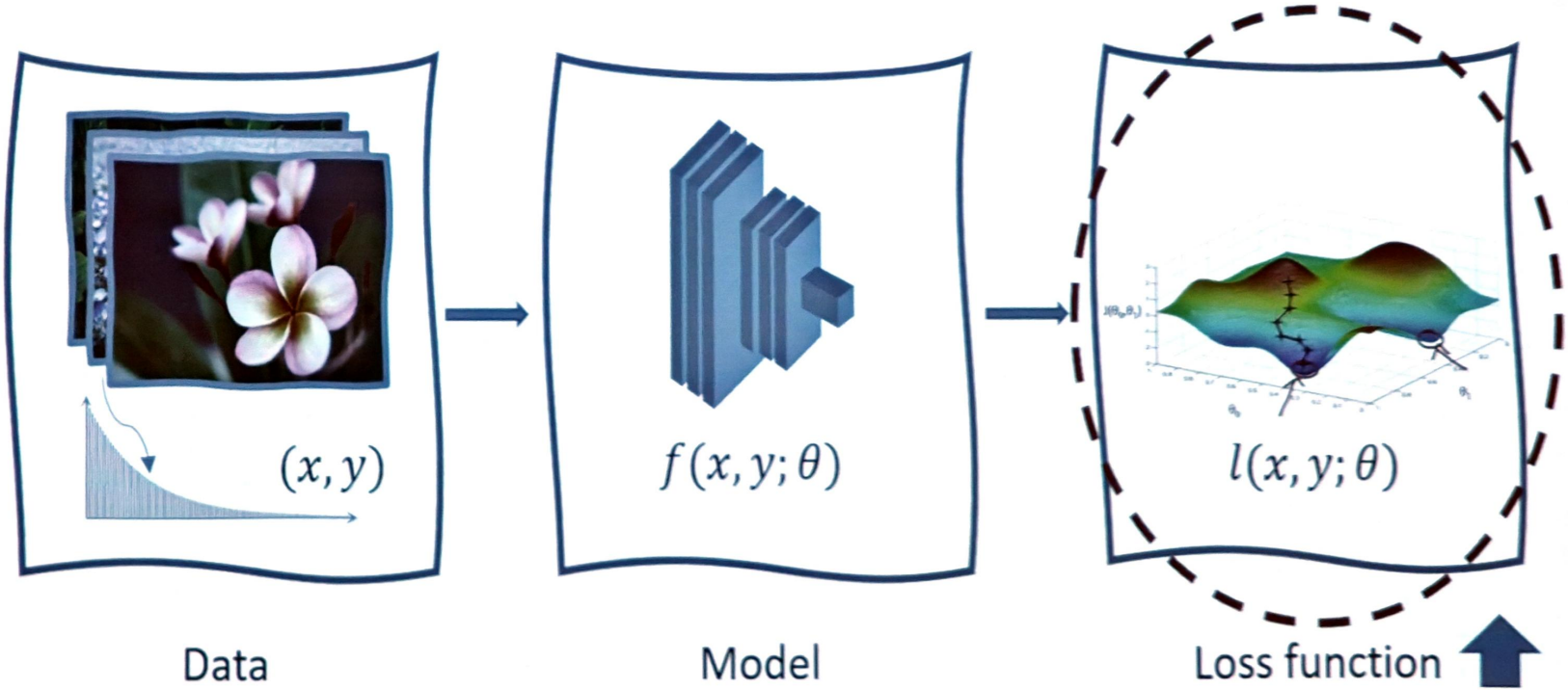
(e) RIDE [17]

(f) TADE [30]

(Image source: Zhang *et al.* in *arXiv:2110.04596v1*)

II.(i) Related Works

Basic process of deep learning



- Reduce negative gradient suppression (e.g. seesaw loss (Wang et al. 2021 in CVPR'21))
- Re-margining (e.g. LDAM loss (Cao et al. 2019 in NeurIPS'2019))

II.(i) Related Works

Loss modification: *LDAM loss*

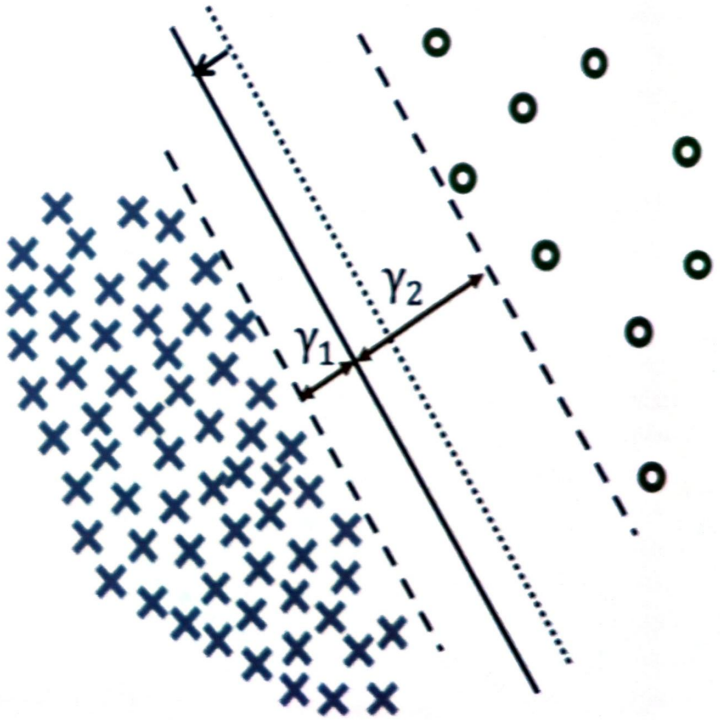
Learning imbalanced datasets with label-distribution-aware margin loss, K. Cao, A. Gaidon, N. Arechiga, T. Ma, in *NeurIPS* 2019.

- Motivation: encourage large margin for tail class.
- Method:

$$L_{ldam} = -\log\left(\frac{e^{z_y-\Delta_y}}{e^{z_y-\Delta_y}+\sum_{j\neq y} e^{z_y}}\right)$$

$$\Delta_j = \frac{C}{n_j^{1/4}},$$

where , z_j is the logit of class j and Δ_j is the class-based margin, C is a hyper-parameter.

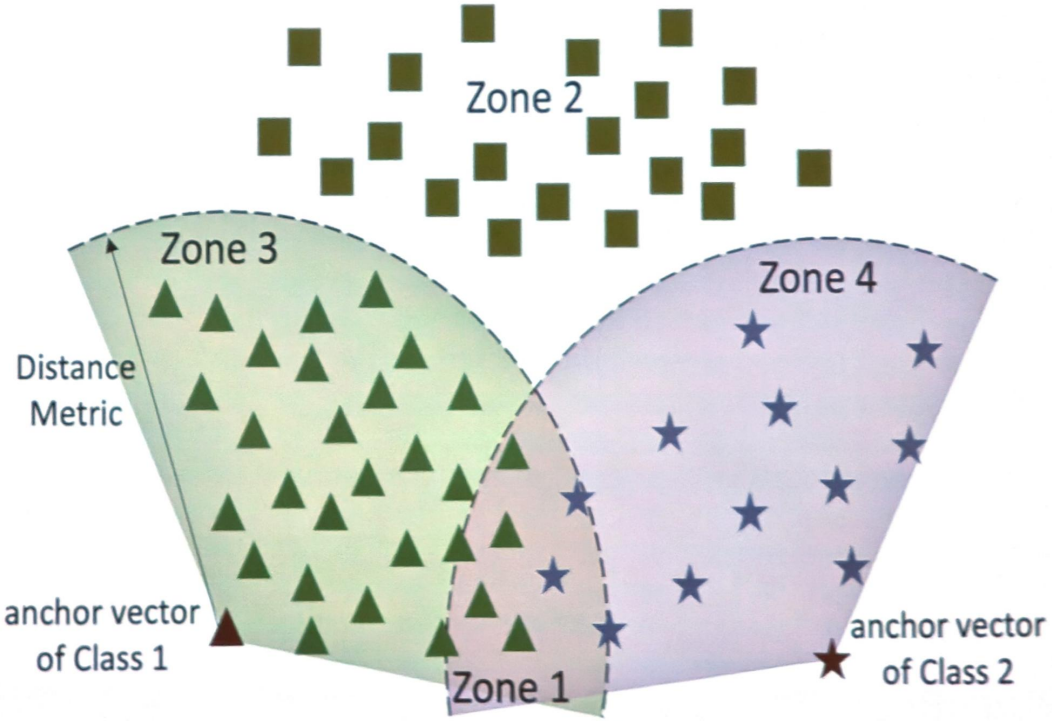


II.(ii) Our work

Key Point Sensitive Loss

Key Point Sensitive Loss for Long-tailed Visual Recognition, M. Li, Y.-M. Cheung, Z.K. Hu, in *TPAMI* 2022, DOI: 10.1109/TPAMI.2022.3196044.

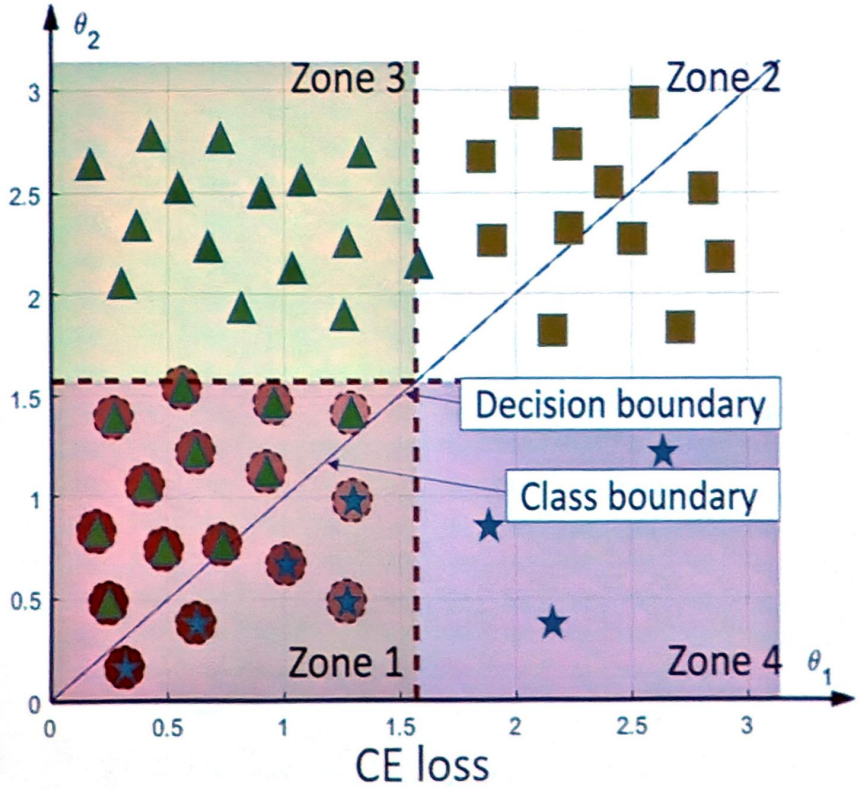
- Motivation

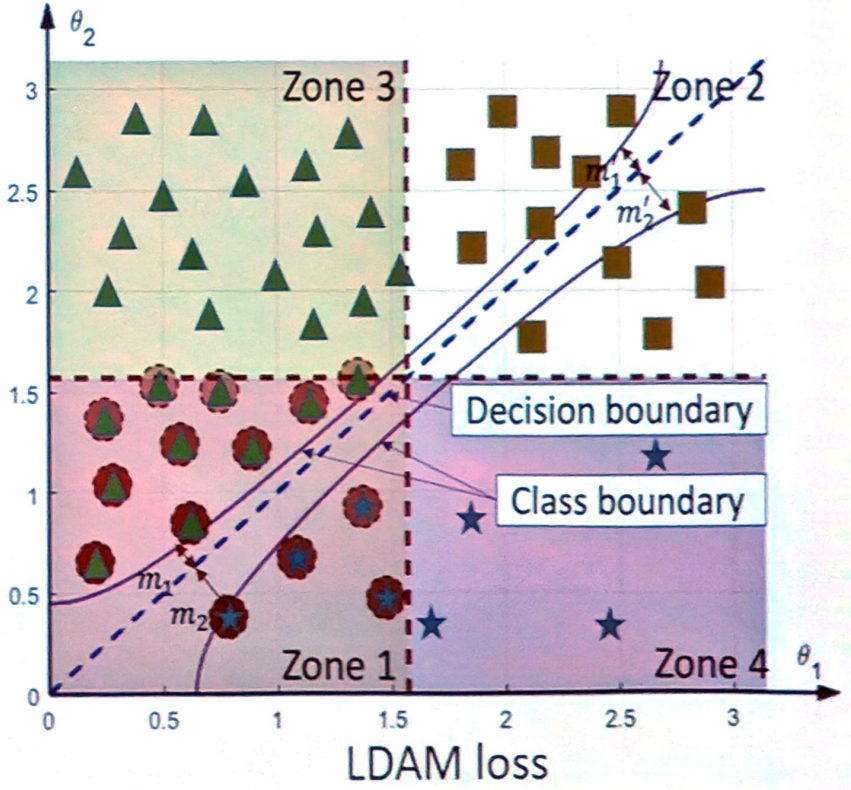


II.(ii) Our work

Key Point Sensitive Loss

- Motivation





II.(ii) Our work

Key Point Sensitive Loss

- Method: *Adjust class margin*

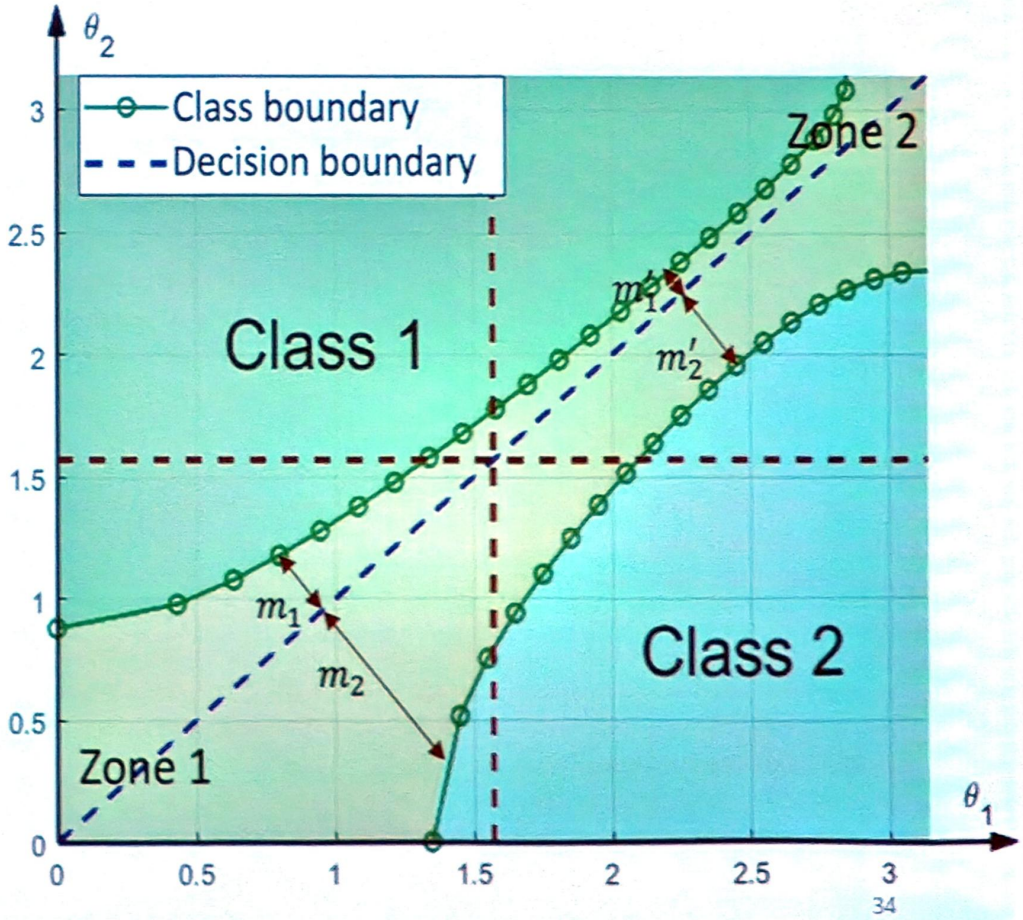
- The tail class has large margin, namely, $m_2 > m_1$ and $m'_2 > m'_1$.
- Key points have large margin, namely, $m_1 > m'_1$ and $m_2 > m'_2$.

For two classes, we set:

$$r_2 \cdot \cos\theta_2 - m_2 > r_1 \cdot \cos\theta_1$$

Margin, and m_i negatively correlated with n_i

Radius, and r_i positive correlated with n_i



II.(ii) Our work

Key Point Sensitive Loss

- Method: *Adjust class margin*

  Extending to Multiple Classes :

  1. extent the binary classification loss to hinge loss;
  2. use LogSumExp to replace $\max$ function;
  3. use softplus to smoothly relax $\max(x, 0)$.

  The key point sensitive (KPS) loss function is expressed as:

$$L_{KPS}(x,y) = -\log \frac{e^{s\cdot(r_y \cos\theta_y - m_y)}}{e^{s\cdot(r_y \cos\theta_y - m_y)} + \sum_{j=1, j\neq y}^{C} e^{s\cdot r_j \cos\theta_j}}.$$

Selection of margin m_i:
(according on lemma 2 in [1])

$$m_i = C_{m_i} - \log n_i$$

Selection of margin r_i:

$$r_i = \log n_i + C_{r_i}$$

Constant

II.(ii) Our work

Key Point Sensitive Loss

- Method: *Adjust gradient*

Balance the gradients:

- Increase tail class gradients;
- Decrease head class gradients.

Adjust the scale parameter s: $p_y = \frac{e^{s\tilde{z}_y}}{\sum_{j=1}^{C} e^{s\tilde{z}_j}}$

II.(ii) Our work

Key Point Sensitive Loss

- Experiment: *Comparison results*

Table 1: Top-1 error rate on CIFAR datasets

Dataset	CIFAR-10-LT		CIFAR-100-LT	
Backbone	ResNet-32			
Imbalance ratio	100	50	100	50
CE loss	29.64	24.78	63.32	56.15
Focal loss [36]	29.62	24.75	62.75	55.68
CosFace [98]	27.92	22.60	60.79	56.89
ArcFace [95]	26.24	21.81	60.94	56.60
LDAM-DRW [67]	22.97	18.97	57.96	53.41
CB-Focal [35]	25.43	20.73	60.40	53.79
Equalised loss*[71]	26.02	-	57.26	-
LA loss [76]	**19.08**	-	56.11	-
BBN [50]	20.18	17.82	57.44	52.98
Meta-learning*[106]	20.00	17.77	55.92	50.84
Meta-learning†[106]	21.10	**17.12**	**55.30**	**49.92**
KPS loss (Ours)	**18.77**	**15.41**	**54.97**	**50.82**

II.(ii) Our work

Key Point Sensitive Loss

- Experiment: *Comparison results*

Table 2: Top-1 error rate on long-tailed ImageNet and iNaturalist 2018

Dataset	ImageNet-LT		iNaturalist 2018
Backbone	ResNet-50	ResNeXt-50	ResNet-50
CE loss	55.49	53.35	39.89
Focal loss [36]	54.2	-	39.70
CosFace [98]	55.05	53.22	33.28
ArcFace [95]	55.46	51.51	36.14
LDAM-DRW [67]	51.20	-	32.00
CB-Focal [35]	-	-	38.88
Equalised loss [71]	52.70	50.90	38.37
LA loss [76]	**51.11**	-	31.56
BBN [50]	55.30	-	**30.38**
Meta-learning*[106]	-	-	32.45
Decoupling [12]	52.30	**50.10**	30.70
KPS loss (Ours)	**48.72**	**47.17**	**29.65**

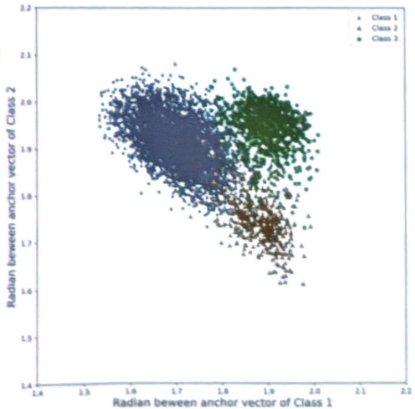
Key Point Sensitive Loss

- Experiment: *Effectiveness of gradient adjustment (GA) strategy*

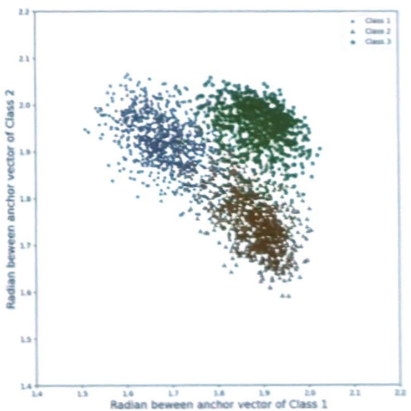
Table 3: Comparison of different optimization strategies (top-1 error rate)

Dataset	CIFAR-10-LT		CIFAR-100-LT	
ρ	100	50	100	50
KPS w/o GA and DRW	19.85	16.40	55.31	50.92
KPS-DRW	19.09	16.11	58.95	54.30
KPS-GA	**18.77**	**15.41**	**54.97**	**50.82**

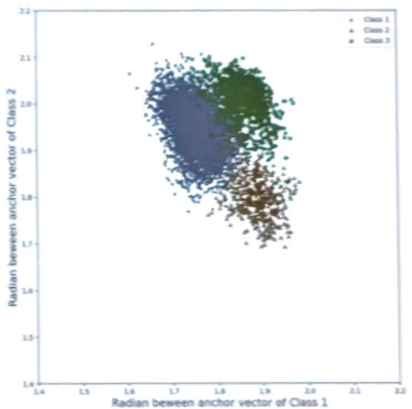
Key Point Sensitive Loss



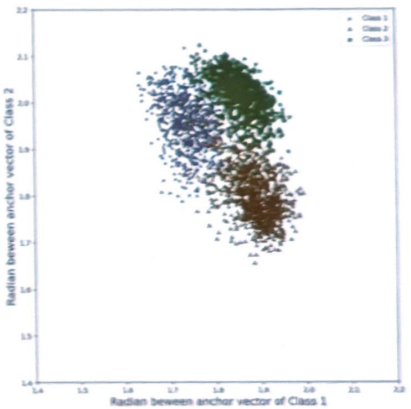
(a1) CE loss on training data



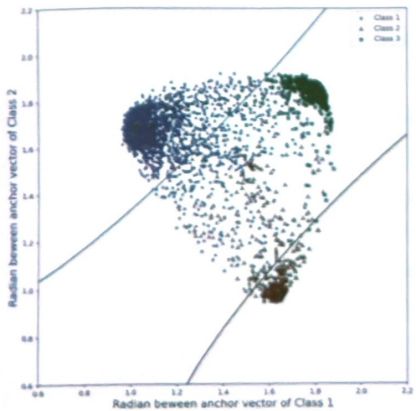
(a2) CE loss on testing data



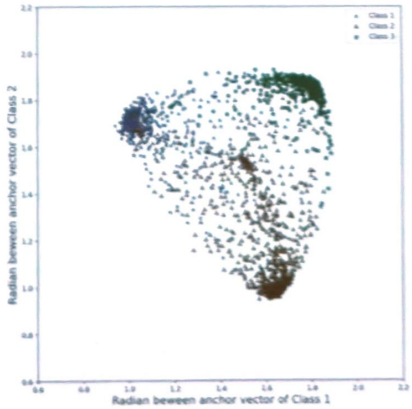
(b1) Focal loss on training data



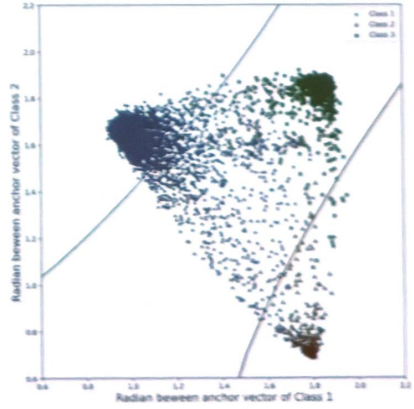
(b2) Focal loss on testing data



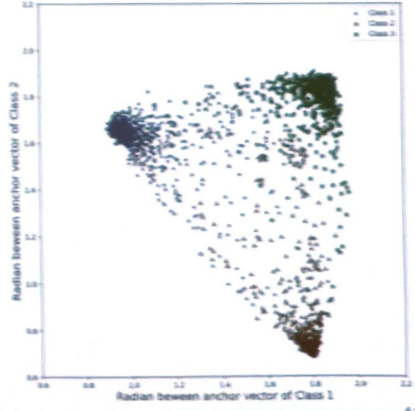
(c1) LDAM on training data



(c2) LDAM on testing data



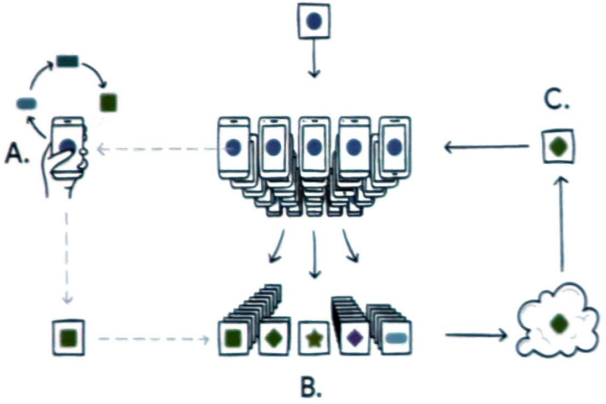
(d1) KPS loss on training data



(d2) KPS loss on testing data

Future Direction

Future direction



(Image source: https://zhuanlan.zhihu.com/p/396739530)

- Federated learning
  - Bank
  - Hospital
  - ...

(Image source: J. Han et al. Deep Self-Learning From Noisy Labels in *ICCV* 2019)

- Noisy label learning

DEPARTMENT OF

IEEE TRANSACTIONS ON EMERGING TOPICS IN COMPUTATIONAL INTELLIGENCE (IF: 5.3)

香港浸會大學
HONG KONG BAPTIST UNIVERSITY

A PUBLICATION OF THE IEEE COMPUTATIONAL INTELLIGENCE SOCIETY

Editor-in-Chief: Yiu-ming Cheung

Scope
The IEEE Transactions on Emerging Topics in Computational Intelligence (TETCI) publishes original articles on emerging aspects of computational intelligence, including theory, applications, and surveys.

TETCI is an electronics only publication. TETCI publishes six issues per year.

Authors are encouraged to submit manuscripts in any emerging topic in computational intelligence, especially nature-inspired computing topics not covered by other IEEE Computational Intelligence Society journals. A few such illustrative examples are glial cell networks, computational neuroscience, Brain Computer Interface, ambient intelligence, non-fuzzy computing with words, artificial life, cultural learning, artificial endocrine networks, social reasoning, artificial hormone networks, computational intelligence for the IoT and Smart-X technologies.