Decoupling representation and classifier for long-tailed recognition

Bingyi Kang, Saining Xie, Marcus Rohrbach, Zhicheng Yan, Albert Gordo, Jiashi Feng, Yannis Kalantidis
Facebook AI, National University of Singapore

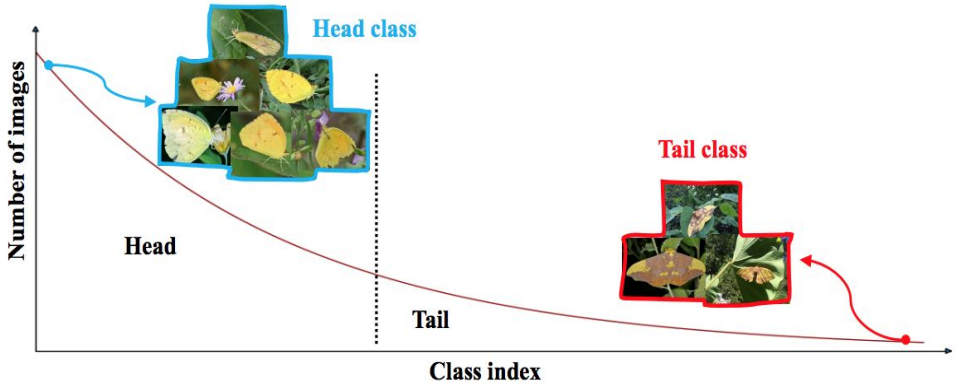
ICLR 2020

Outlines

- **Background**
  - What is long-tailed recognition?
  - What are the existing class-balancing strategies?
- **Decoupling representation and classifier**
  - Sampling strategies for *representation learning*
  - Balanced *classifier learning*
    - Classifier re-training (cRT)
    - Nearest class mean classifier (NCM)
    - τ-normalized classifier
    - Learnable weight scaling (LWS)
- **Experiments and discussions**
- **Conclusions**

Background: long-tailed recognition

Long-tailed recognition - a few classes have many samples (head classes) and many classes have few (tail classes)



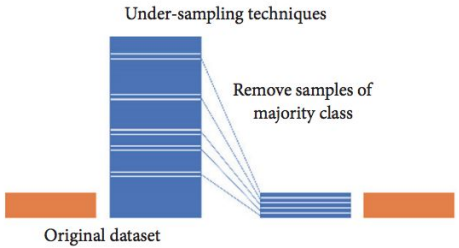
- Problem: head classes dominate the training procedure, which leads to biased predictions and misleading accuracy

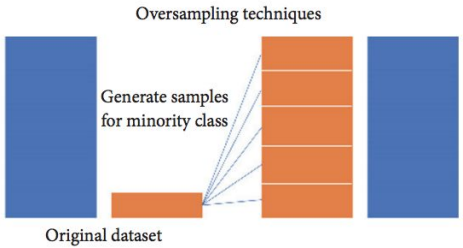
Zhou, Boyan, et al. "BBN: Bilateral-Branch Network with Cumulative Learning for Long-Tailed Visual Recognition." *arXiv preprint arXiv:1912.02413* (2019).

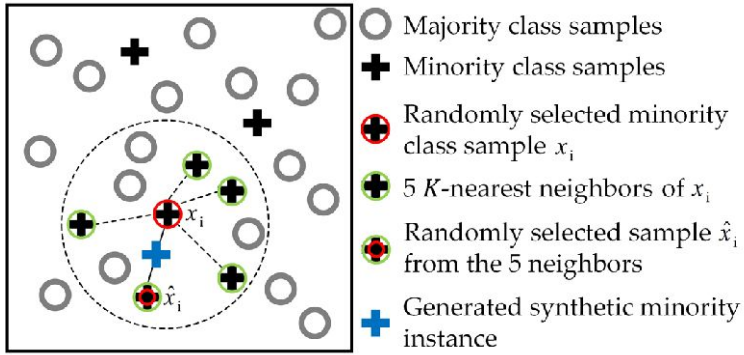
Background: class-balancing strategies

Strategy 1: re-sampling to balance the data distribution

- Under-sampling for the majority classes, hurts generalization ability
- Over-sampling for the minority classes
  - Duplicating or synthesizing examples, eg. SMOTE







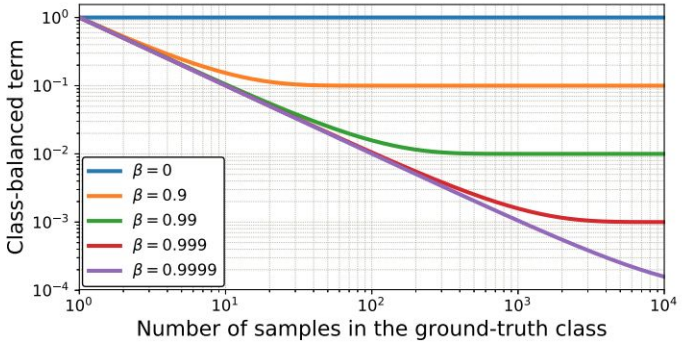
Le, Tuong, et al. "A hybrid approach using oversampling technique and cost-sensitive learning for bankruptcy prediction." *Complexity* 2019 (2019).
Chawla, Nitesh V., et al. "SMOTE: synthetic minority over-sampling technique." *Journal of Artificial Intelligence Research* 16 (2002): 321-357.
Ma, Huiqin, et al. "Integrating growth and environmental parameters to discriminate powdery mildew and aphid of winter wheat using bi-temporal Landsat-8 imagery." *Remote Sensing* 11.7 (2019): 846.

Background: class-balancing strategies

Strategy 2: re-weighting the loss

- Vanilla scheme: weighting classes proportionally to the inverse of their class frequency

- Class balanced (CB) loss: effective number of samples
  - As the number of samples increase, the additional benefit of a newly added data point will diminish
  - The effective number of samples is defined as $(1-\beta^n)/(1-\beta)$ where n is the number of samples and $\beta \in [0, 1)$ is a hyperparameter
  - Class balanced term - the inverse of the effective number
  - Weighting classes by the inverse of the ***effective*** number of samples



Cui, Yin, et al. "Class-balanced loss based on effective number of samples." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2019.

Background: class-balancing strategies

Strategy 3: two-stage fine-tuning

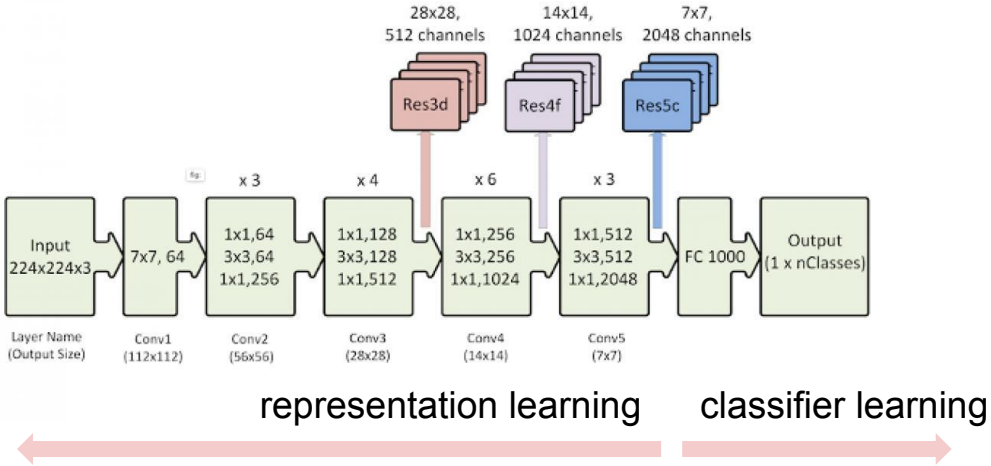
- (Stage 1) train the network without data balancing
- (Stage 2) utilize re-balancing to fine-tune the *whole* network

- Related literatures:
  - Cao, Kaidi, et al. "Learning imbalanced datasets with label-distribution-aware margin loss." *Advances in Neural Information Processing Systems*. 2019.
  - Cui, Yin, et al. "Large scale fine-grained categorization and domain-specific transfer learning." Proceedings of the IEEE conference on computer vision and pattern recognition. 2018.
  - ...

Decoupling representation and classifier

- With the aforementioned strategies, it remains unclear whether the ability to learn with class balancing is from a better representation or a better classifier decision boundary

- Example: ResNet-50



He, Kaiming, et al. "Deep residual learning for image recognition." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016.

Decoupling representation and classifier

- Two papers during the same period disentangle representation learning from classifier learning

  Decoupling representation and classifier for long-tailed recognition, ICLR 2020

  BBN: Bilateral-Branch Network with Cumulative Learning for Long-Tailed Visual Recognition, CVPR 2020

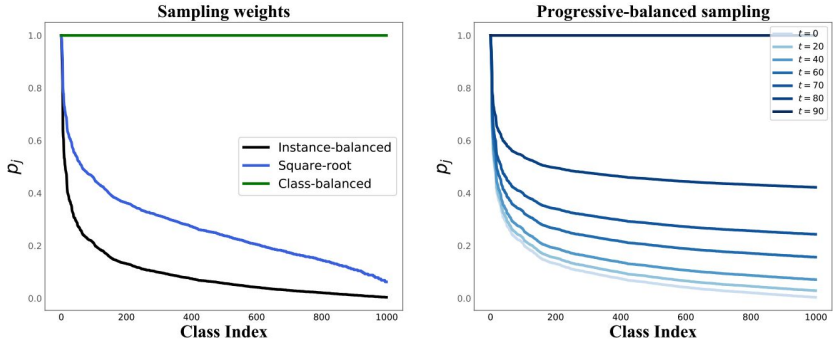
Sampling strategies for representation learning

The probability p_j of sampling a data point from class j is given by

$$p_j = \frac{n_j^q}{\sum_{i=1}^{C} n_i^q}$$

where n_j is the number of training sample for class j, C is the number of training classes.

(1) q = 1, Instance-balanced sampling: each instance has equal probability of being selected
(2) q = 0, Class-balanced sampling: a class is selected uniformly
(3) q = ½, Square-root sampling
(4) Progressively-balanced sampling: $p_j^{\mathrm{PB}}(t) = (1 - \frac{t}{T})p_j^{\mathrm{IB}} + \frac{t}{T}p_j^{\mathrm{CB}}$ where T is the total number of epochs, it progressively "interpolates" between instance-balanced sampling and class-balanced sampling as learning progresses



Balanced classifier learning

- Classifier re-training (cRT)
  - Similar to two-stage fine-tuning
  - In the second stage, instead of training the *whole* network using a smaller learning rate, cRT keeps the *representations fixed*, re-initializes and optimizes the *classifier* only

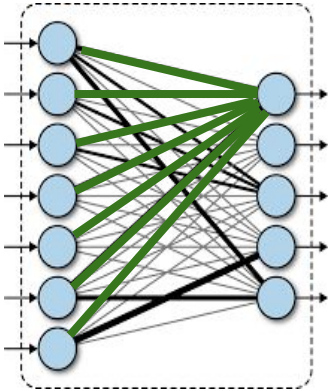
- Nearest class mean classifier (NCM)
  - Non-parametric
  - First compute the mean feature representation for each class, and then perform nearest neighbor search using cosine similarity or Euclidean distance

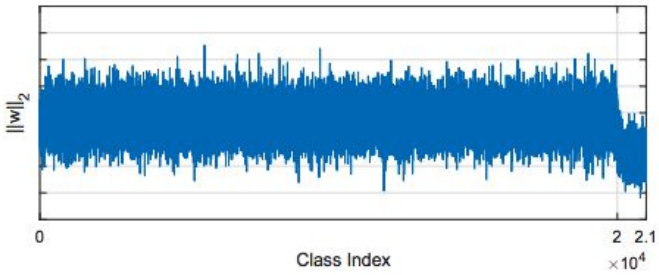
Decoupling representation and classifier

- τ-normalized classifier (τ-normalized)

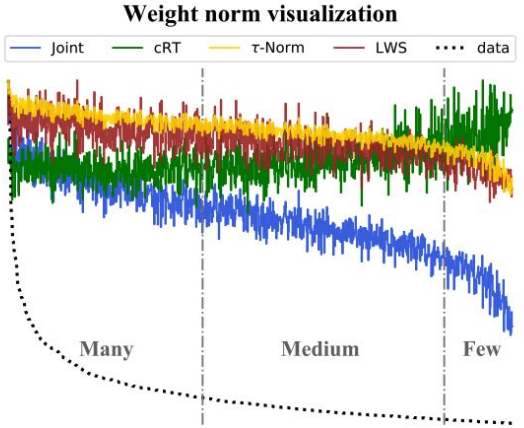
$$\widetilde{w_i} = \frac{w_i}{||w_i||^\tau}$$

where τ is a hyper-parameter controlling the "temperature" of the normalization. In the paper, they empirically choose $\tau \in (0, 1)$ via cross validation





Norm of the weights for low-shot face recognition



Norm of the weights for ImageNet-LT (artificially truncated)

Guo, Yandong, and Lei Zhang. "One-shot face recognition by promoting underrepresented classes." arXiv preprint arXiv:1707.05574 (2017).

Decoupling representation and classifier

- Learnable weight scaling (LWS)

$$\widetilde{w_i} = f_i * w_i, \text{where } f_i = \frac{1}{||w_i||^\tau}$$

f_i is a learnable parameter

Fix the representations and classifier and only learns the weights re-scaling factor

Experiments and discussions

Many-shot (more than 100 images), *Medium-shot* (20-100 images), *Few-shot* (less than 20 images)

- *Sampling matters when training jointly.* Consistent gains in performance when using better sampling strategies in medium- and few-shot cases compared with instance-balanced sampling
- *Joint or decoupled learning?* Decoupled methods outperform in medium- and few-shot cases, especially with cRT and τ-norm
- *Instance-balanced sampling gives the most generalizable representations,* expect for many-shot case

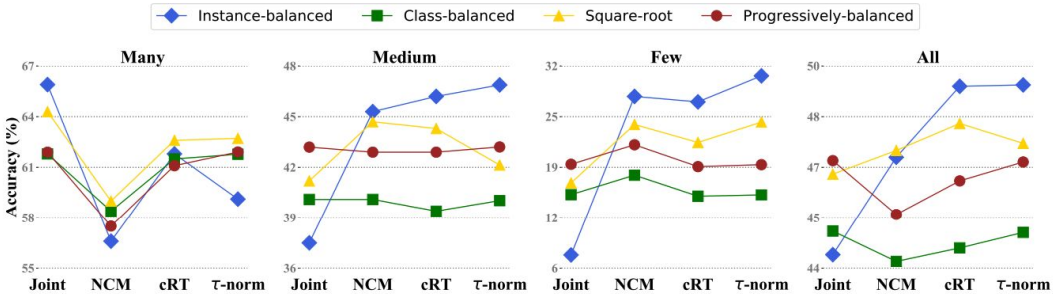


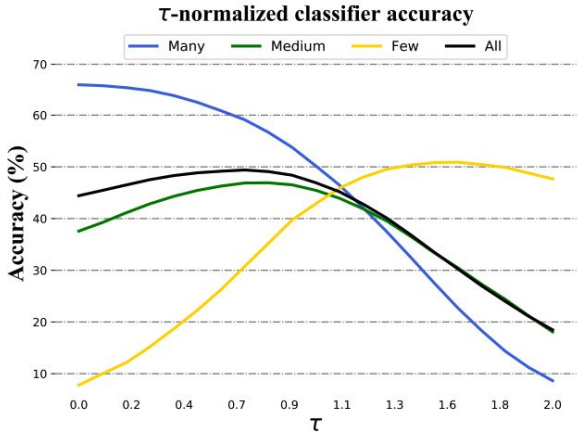
Figure 1: The performance of different classifiers for each split on ImageNet-LT with ResNeXt-50. Colored markers denote the sampling strategies used to learn the representations.

Experiments and discussions

- *Is it really beneficial to decouple representation and classification?*
- *How does the "temperature" influence the performance for the τ-normalized classifier?*

Table 1: Retraining/finetuning different parts of a ResNeXt-50 model on ImageNet-LT. B: backbone; C: classifier; LB: last block.

Re-train	Many	Medium	Few	All
B+C	55.4	45.3	24.5	46.3
B+C(0.1×lr)	**61.9**	45.6	22.8	48.8
LB+C	61.4	45.8	24.5	48.9
C	61.5	**46.2**	**27.0**	**49.5**



Classifier accuracy for ImageNet-LT (artificially truncated)

Experiments and discussions

- *Comparison with the state-of-the-art on long-tailed datasets*

Table 2: Long-tail recognition accuracy on ImageNet-LT for different backbone architectures. † denotes results directly copied from Liu et al. (2019). * denotes results reproduced with the authors' code. ** denotes OLTR with our representation learning stage.

Method	ResNet-10	ResNeXt-50	ResNeXt-152
FSLwF† (Gidaris & Komodakis, 2018)	28.4	-	-
Focal Loss† (Lin et al., 2017)	30.5	-	-
Range Loss† (Zhang et al., 2017)	30.7	-	-
Lifted Loss† (Oh Song et al., 2016)	30.8	-	-
OLTR† (Liu et al., 2019)	35.6	-	-
OLTR*	34.1	37.7	24.8
OLTR**	37.3	46.3	50.3
Joint	34.8	44.4	47.8
NCM	35.5	47.3	51.3
cRT	**41.8**	**49.5**	52.4
τ-normalized	40.6	**49.4**	**52.8**
LWS	**41.4**	**49.9**	**53.3**

Experiments and discussions

- *Comparison with the state-of-the-art on long-tailed datasets*

Table 3: Overall accuracy on iNaturalist 2018. Rows with † denote results directly copied from Cao et al. (2019). We present results when training for 90/200 epochs.

Method	ResNet-50	ResNet-152
CB-Focal†	61.1	-
LDAM†	64.6	-
LDAM+DRW†	68.0	-
Joint	61.7/65.8	65.0/69.0
NCM	58.2/63.1	61.9/67.3
cRT	65.2/67.6	68.5/71.2
τ-normalized	65.6/**69.3**	68.8/**72.5**
LWS	65.9/**69.5**	69.1/72.1

Table 4: Results on Places-LT, starting from an ImageNet pre-trained ResNet152. † denotes results directly copied from Liu et al. (2019).

Method	Many	Medium	Few	All
Lifted Loss†	41.1	35.4	24.0	35.2
Focal Loss†	41.1	34.8	22.4	34.6
Range Loss†	41.1	35.4	23.2	35.1
FSLwF†	43.9	29.9	29.5	34.9
OLTR†	44.7	37.0	25.3	35.9
Joint	**45.7**	27.3	8.2	30.2
NCM	40.4	37.1	27.3	36.4
cRT	42.0	37.6	24.9	36.7
τ-normalized	37.8	**40.7**	**31.8**	**37.9**
LWS	40.6	39.1	28.6	37.6

Experiments and discussions

CVPR 2020 "BBN: Bilateral-Branch Network with Cumulative Learning for Long-Tailed Visual Recognition"

- Classifier re-training (cRT)

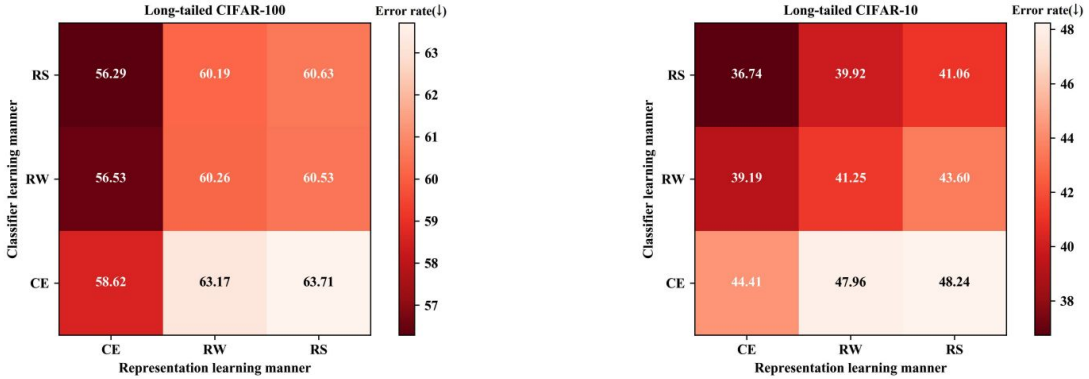


Figure 2. Top-1 error rates of different manners for representation learning and classifier learning on two long-tailed datasets CIFAR-100-IR50 and CIFAR-10-IR50 [3]. "CE" (Cross-Entropy), "RW" (Re-Weighting) and "RS" (Re-Sampling) are the conducted learning manners. As observed, when fixing the representation (comparing error rates of three blocks in the vertical direction), the error rates of classifiers trained with RW/RS are reasonably lower than CE. While, when fixing the classifier (comparing error rates in the horizontal direction), the representations trained with CE surprisingly get lower error rates than those with RW/RS. Experimental details can be found in Section 3.

Conclusions

- Sampling strategies matter when jointly learning representation and classifier

- With a properly re-balanced classifier, instance sampling gives more generalizable representations that can achieve state-of-the-art performance