Decoupling representation and classifier for long-tailed recognition

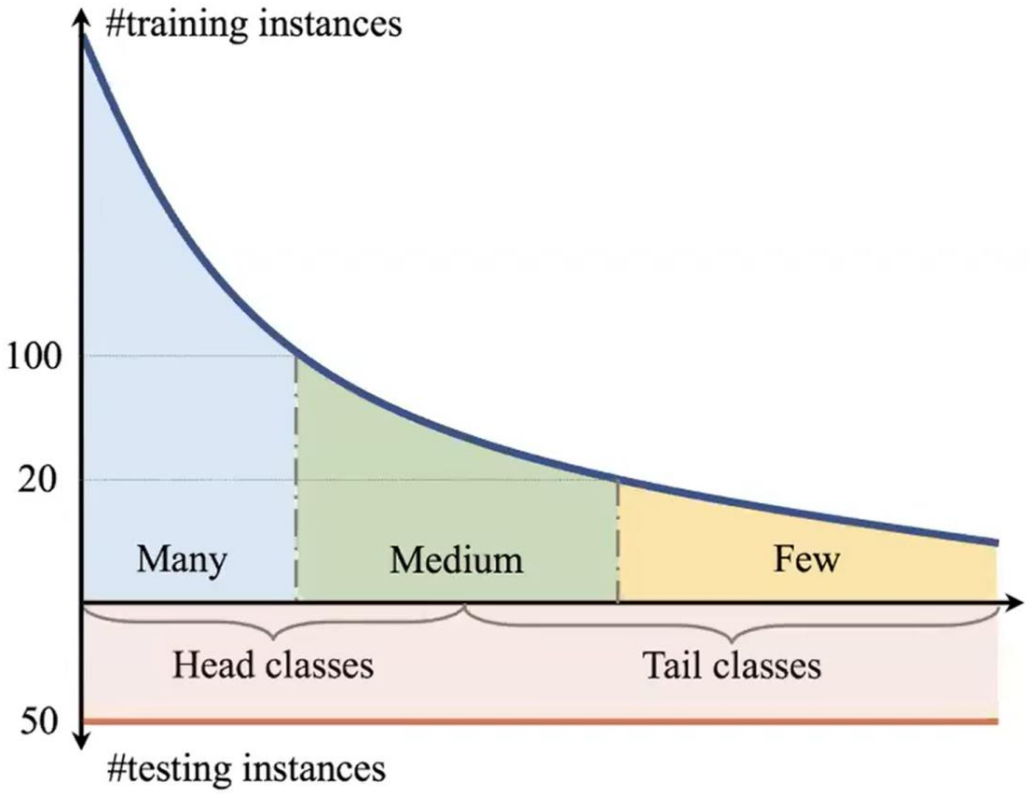
ICLR 2020

Catalogue

- Background
- Related work
- Introduction
- Method
- Conclusion
- My opinions & questions

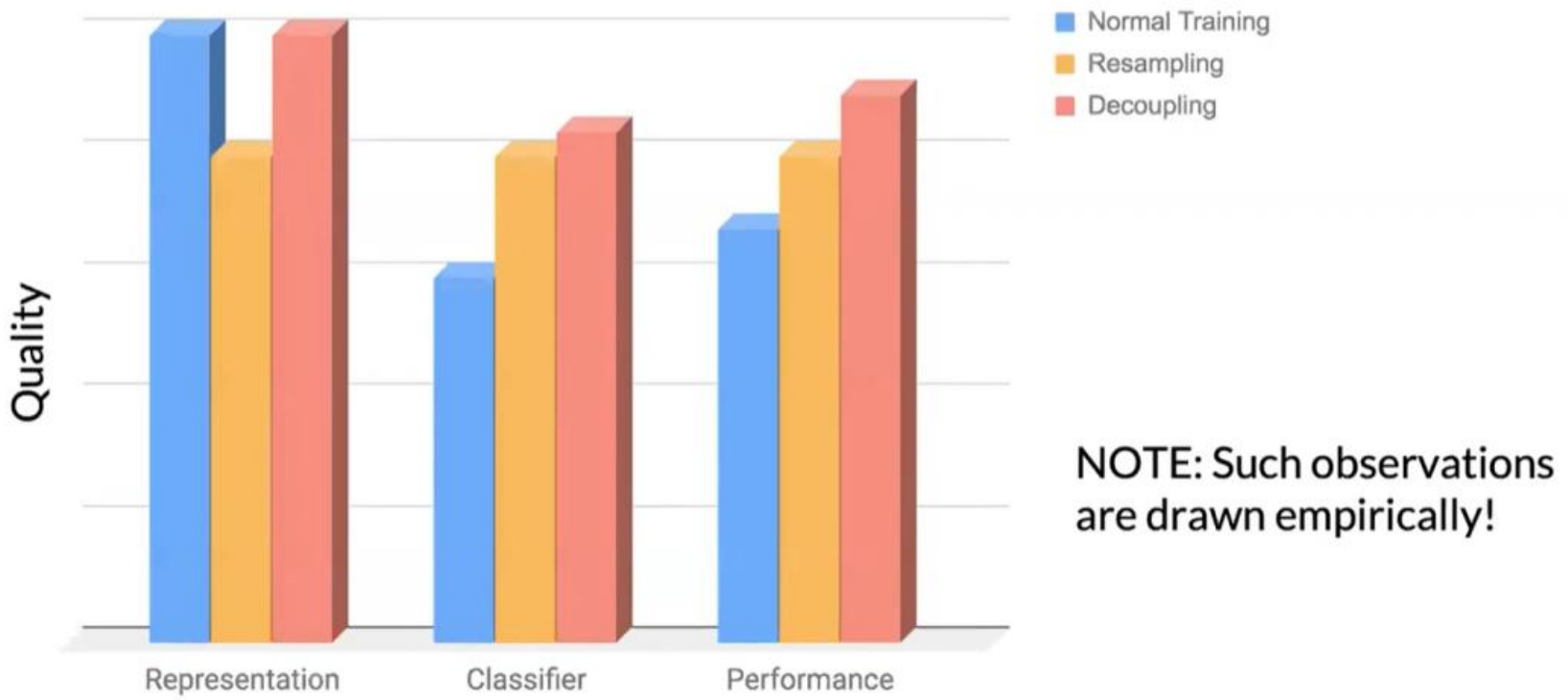
Background

- The common belief is that the existing approaches are useful for learning high-quality representations for long-tailed recognition.
- Thus, most aforementioned approaches learn the classifiers jointly with the data representations.
- However, such a scheme makes it unclear how the long-tailed recognition ability is achieved—is it from learning a better representation or by handling the data imbalance better via shifting classifier decision boundaries?



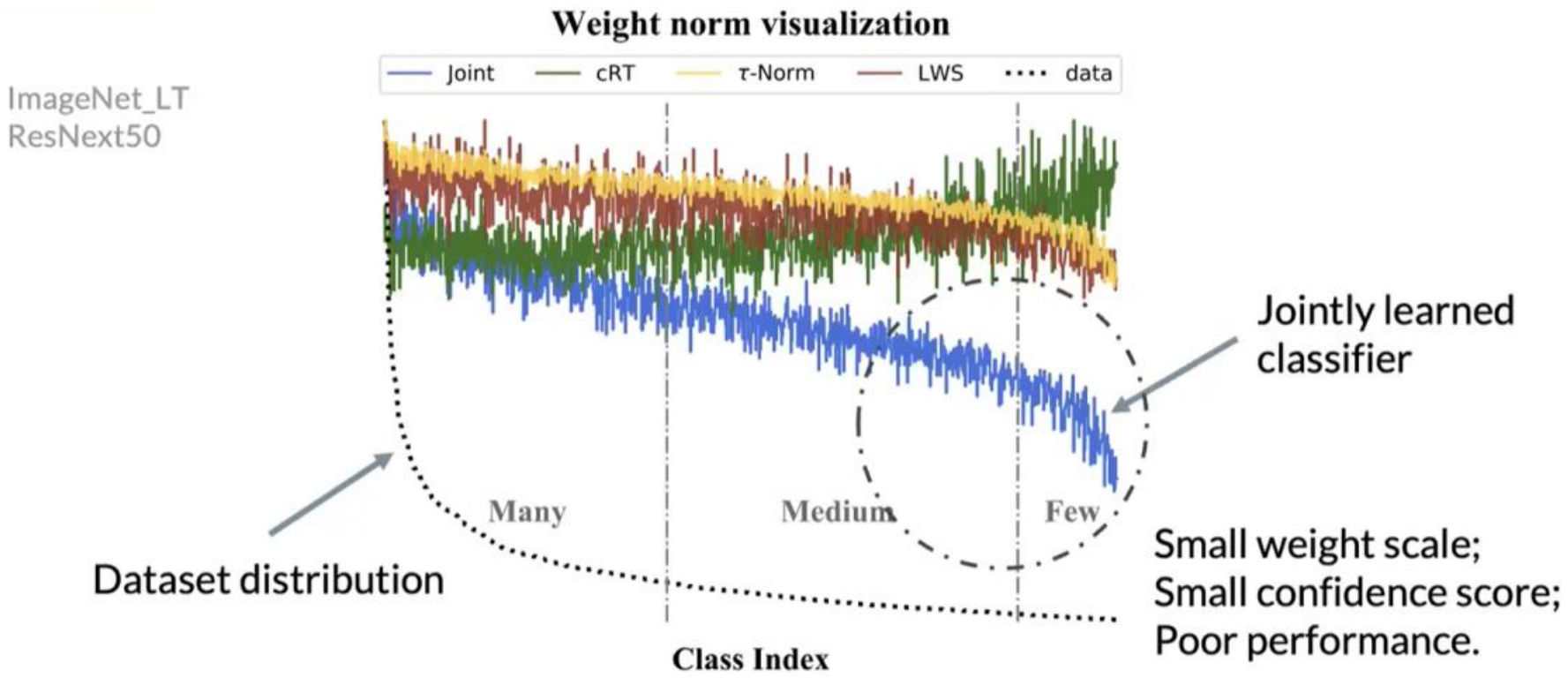
Background- The problem behind long-tail

- Classification performance = Representation Quality + Classifier Quality



Background- The problem behind long-tail

- The problem with the classifier



Related work

Imbalanced learning

- Re-sampling(under-sampling, over-sampling)
- Re-weighting(Focal loss, CB loss…)
- Transfer learning from head to tail class
- Domain adaption
- Semi-supervised learning or Self-supervised learning

Introduction- How to decouple?

- For learning representations, the model is trained through different losses or sampling strategies(the standard instance-based sampling, class-balanced sampling and a mixture of them).

- For classification, upon the learned representations, the model recognizes the long-tailed classes through various classifiers(the re-training linear classifier with class-balanced manner, non-parametric nearest class mean classifier and normalizing the classifier weights).

Introduction-Value of re-adjusting the decision boundaries

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**

Method- first stage

$$p_j = \frac{n_j^q}{\sum_j n_j^q}$$

Different sample strategies

- Instance-balanced sampling: $q = 1$

(each class is sampled proportionally to the number of samples of in this class)

- Class-balanced sampling: $q = 0.5$

(each class has an equal probability to be selected)

- Progressively-balanced sampling:

(a mix of IB and CB)

$$p_j(t) = \left(1 - \frac{t}{T}\right) p_j^{IB} + \frac{t}{T} p_j^{CB}$$

Method- second stage

Different ways of re-adjusting classifiers

- Classifier Re-training (cRT)

(randomly re-initialize and optimize the linear classifier weights W and b with class-balanced sampling strategy)

- Nearest Class Mean classifier (NCM)

(perform nearest neighbor search either using cosine similarity or the Euclidean distance computed on L2 normalized mean features)

- τ -normalized classifier (τ -normalized): $w_i' = \frac{w_j}{||w_i||^\tau}$

- Learnable weight scaling (LWS):

(freezing representation & classifier and only learning τ)

Method- experiments

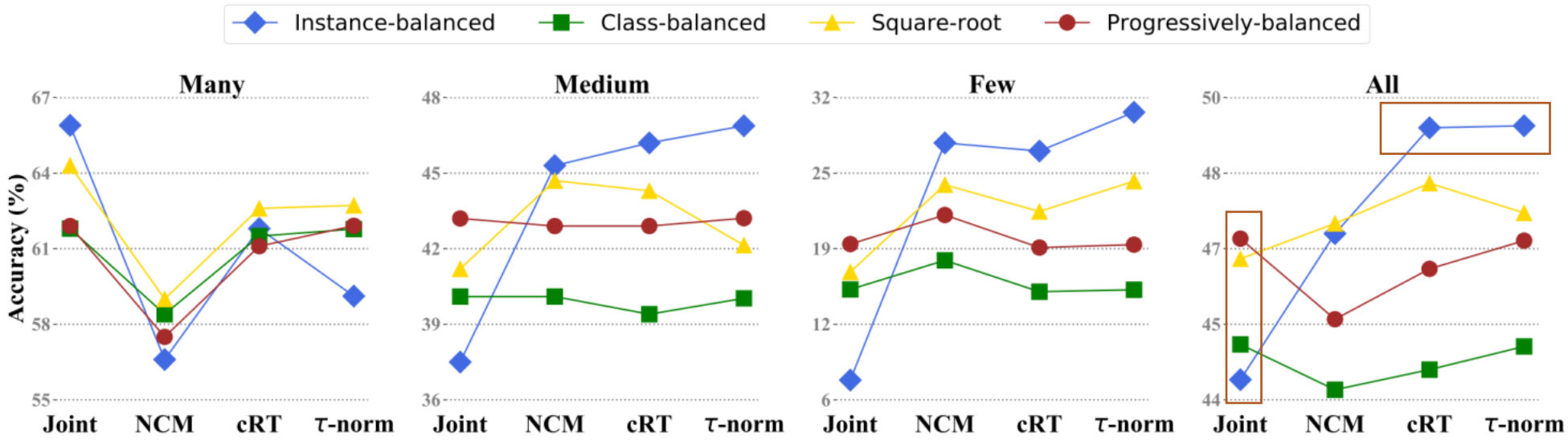


Figure 1: The performance of different classifiers for each split on ImageNet-LT with ResNeXt-50. Colored markers denote the sampling strategies used to learn the representations.

- With representations learned with the simplest sample strategy (instance-balance sampling), it is also possible to achieve strong long-tailed recognition ability by adjusting only the classifier.
- Data imbalance might not be an issue in learning high-quality representations;

Method- experiments

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**

- Fine-tuning the whole network yields the worst performance (46.3% and 48.8%),
- keeping the representation frozen performs best (49.5%)

My opinion & questions about the paper

Opinion:

The strategy of considering the representation and classifier is really easy and useful. When it comes to the reasons of its effectiveness, in my opinion, the decoupling strategy combines the statistics-based method and end-to-end manner. Those re-training classifier methods are based on statistics and thus well alleviates class bias, while end-to-end manner in first stage can obviously obtain the representation based on the total big data.

Question:

- Maybe it will be better to explain the reasons of the changes of decision boundary by using some visualization tools like t-SNE.
- It seems that there are no experiment of using the decoupling strategy and some loss like Focal loss at the same time.

Thanks for listening.