LEARNED NEAREST-CLASS-MEAN FOR BIASED REPRESENTATIONS IN LONG-TAILED RECOGNITION

Anonymous authors
Paper under double-blind review

ABSTRACT

The problem of long-tailed recognition (LTR) has received attention in recent years due to the fundamental power-law distribution of objects in the real-world. While classifier bias in LTR has been addressed by many works, representation bias has not yet been researched. At the same time, most recent works use softmax classifiers that are unable to cope with representation bias. In this work, we address these shortcomings by firstly making the key observation that intra-class variance in representation space is negatively correlated to class frequency, leading to biased representations; our analysis reveals that high tail variance is due to spurious correlations learned by deep models. Secondly, to counter representation bias, we propose the Learned Nearest-Class-Mean (NCM), which overcomes uncertainty in empirical centroid estimates and jointly learns centroids minimizing average class-distance normalized variance. Further, we adapt the logit adjustment technique in the NCM framework to achieve higher tail class margin. Our Learned NCM with Logit Adjustment achieves 6% gain over state-of-the-art in tail accuracy on the benchmark CIFAR100-LT and ImageNet-LT datasets.

1 INTRODUCTION

Imbalanced datasets are prevalent in the natural world due to the fundamental power-law distribution of objects Van Horn & Perona (2017). Past decades have seen a lot of research in class-imbalanced learning He & Garcia (2009), a challenge that is shared by a diverse set of problems ranging from image classification Van Horn & Perona (2017); Wang et al. (2017), face recognition Yin et al. (2019) and object detection Lin et al. (2017) to sentiment analysisMaas et al. (2011) and anomaly detectionChandola et al. (2009). Prior work in this area primarily focuses on data-resampling Estabrooks et al. (2004); He & Garcia (2009) and the related area of cost-sensitive learning Elkan (2001); Ling & Sheng (2008); Khan et al. (2017). More recently, due to it's far-reaching relevance in the context of deep-learning, the problem of long-tailed recognition (LTR) has received significant attention Cui et al. (2019); Cao et al. (2019); Liu et al. (2019) in the field of computer vision.

A common underlying assumption of recent work in LTR is that softmax classifiers learned by regular sampling are biased towards head classes. They seek to rectify head class bias in the classifier through data-resampling Kang et al. (2020), loss reshaping Ren et al. (2020); Menon et al. (2021); Samuel & Chechik (2021), ensemble-based models Xiang et al. (2020); Wang et al. (2020); Zhou et al. (2020), and knowledge transfer Liu et al. (2019; 2021). Particularly, it was shown in Kang et al. (2020) that deep models have a high correlation of classifier norm to class frequency. To achieve similar classifier norms, they retrain the classifier with balanced class sampling, or alternatively normalize every classifier by a power of its norm. Towards a slightly different end, Menon et al. (2021) propose a logit adjusted cross-entropy loss, whereby they apply label-dependent logit offsets to account for the label distribution shift from imbalanced training set to balanced test set in LTR.

However, the issue of *representation bias* in the LTR literature has so far been ignored, leading to the false belief that head and tail classes share equivalent class-conditional local neighborhoods in the representation space. In this work, we analyze the correlation between class-conditional local neighborhoods and class frequency through a quantitative estimate of the compactness of a class- the *intra-class sample variance*. Our analysis shows that the intra-class sample variance is negatively correlated to the class frequency, indicating that majority classes are more compact while tail classes are diffused in the representation space. In Figure 1, we show a tail class alongside two head classes

from long-tailed CIFAR10, projected down to 2D space using TSNE Van der Maaten & Hinton (2008); the illustration shows that the tail class has a much higher scatter than the head classes. This suggests that deep models learn spurious correlations which show up as high variance noise in tail class representations, as we explore in Section 3. Moreover, softmax classifiers used in prior work are unable to cope with this high tail variance, leading to poorer tail class accuracy.

Motivated by the presence of representation bias in LTR, we avoid the use of an additional softmax classifier layer and instead propose to use *Learned Nearest-Class-Mean* (Learned NCM). Differently from Nearest-Class-Mean (NCM), whereby samples are classified based on distances to *empirical centroids*, we propose to learn the centroids themselves. Since the uncertainty in estimation of class centroids scales as $1/\sqrt{N_y}$, where N_y is the number of samples in class y, learning the centroids leads to a lower classification error while at the same time overcoming centroid estimation error. In Section 4, we show that NCM can be interpreted as logistic regression, and show the form of the gradient update on centroids in Learned-NCM. Further, we impose a higher tail margin during training using the logit adjustment technique Menon et al. (2021), which results in much lower tail classification error. We also extend Learned-NCM using multiple centroids per class, leading to the Multi-NCM model. In Section 6, we discuss how Learned NCM implicitly minimizes the class-distance normalized variance.

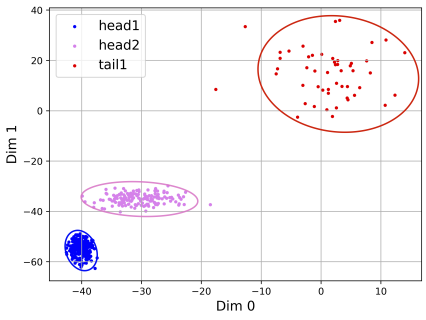


Figure 1: 2-dim TSNE projection of a tail class along with 2 head classes for long-tailed CIFAR10. Head classes have small intra-class variance and are compact, while the tail class has high intra-class variance and is diffused.

Overall, here is a summary of our main contributions in this work: (1) We analyze and show that models learned with regular instance-based sampling exhibit representation bias of the form where the intra-class variance is negatively correlated to the class frequency. This intriguing observation has not been explored in prior work to the best our knowledge, (2) We propose to deal with representation bias using the Learned-NCM, whereby we learn centroids for classification using the Nearest-Class-Mean rule. Further, we impose a higher tail margin by applying the logit adjustment technique to Learned-NCM. We also show how to extend our model to Multi-NCM wherein we employ multiple centroids per class, (3) Our extensive experiments on three benchmark long-tailed datasets, CIFAR10-LT, CIFAR100-LT and ImageNet-LT, with consistently superior results especially on the harder tail classes, demonstrate the soundness of our approach.

2 Related work

The literature on long-tailed recognition can be broadly divided into four main strands: (i) Data resampling, (ii) Loss reshaping, (iii) Ensemble based models, and (iv) Knowledge transfer.

Data resampling. Data resampling or reweighting is the most commonplace strategy against imbalanced datasets. This generally takes three forms: (i) Oversampling minority class samples by adding small perturbations to the data Chawla et al. (2002; 2003), (ii) Undersampling majority class samples by throwing away some data Drummond et al. (2003), and (iii) Uniform or class-balanced sampling based on the number of samples in a class Sun et al. (2019); Xian et al. (2019). However, oversampling minority class samples has been shown to lead to overfitting and undersampling majority class samples can cause poor generalization He & Garcia (2009). Recently, Kang et al. (2020) learn the representations using instance-based sampling and retrain the softmax classifier in a second step using uniform sampling while keeping underlying representations fixed. We also use the two-stage training strategy in our work; however, we avoid learning an additional softmax classifer and instead use NCM as it operates in the representation space to mitigate representation bias.

Loss reshaping. This line of prior work focuses on engineering loss functions suited for imbalanced datasets. Focal loss Lin et al. (2017) reshapes the standard cross-entropy loss to push more weight on misclassified and/or tail class samples. Class-balanced loss Cui et al. (2019) uses a theoretical estimate of the volume occupied by a class to reweigh the loss function. Cao et al. (2019) offset the label's logit score with a power of class frequency to achieve a higher decision boundary margin for tail classes. Ren et al. (2020) apply the offset to all the logits, and change the class frequency

exponent from $1/4$ in prior work to 1 and use a meta-sampler to retrain the classifier. Menon et al. (2021) apply a label-dependent adjustment to the logit score before softmax which is theoretically consistent. In our work, we build upon the logit adjustment approach and show how it can be adapted to the Nearest-Class-Mean framework to achieve lower tail class error.

Ensemble based models. Many papers employ a specialized ensemble of experts to reduce tail bias and model variance. Sharma et al. (2020) train experts on class-balanced subsets of the training data and aggregate them using a joint confidence calibration layer. Xiang et al. (2020) also train experts on class-balanced subsets and distill them into a unifed student model. Wang et al. (2020) explicitly enforce diversity in experts and aggregate them using a routing layer. Zhou et al. (2020) train two experts using regular and reversed sampling together with an adaptive learning strategy. Our Multi-NCM, which uses multiple centroids per class, is also an atypical ensemble model.

Knowledge transfer. Transfer of knowledge from head to tail classes aims to boost data starved tail representations using head data. Yin et al. (2019) generate synthetic tail representations by sampling displacements to class centroids from manyshot samples and transposing onto tail centroids. Liu et al. (2019) enhance representations by attending over a visual memory of class-centroids. However, due to lack of control over the knowledge transfer process these methods have limited applicability.

Orthogonal to all these works, we discard the softmax classifier layer and use NCM. The NCM classifier appears prominently in metric learning and few-shot learning. Mensink et al. (2013) proposed learning a Mahalanobis distance metric on top of fixed representations, which leads to a NCM classifier based on the Euclidean distance in the learned metric space. Snell et al. (2017) learn deep representations for few shot learning by minimizing the NCM based loss on task-specific epsiodes. Guerriero et al. (2018) learn deep representations by optimizing the NCM classification objective on the entire training data. Rebuffi et al. (2017) learn NCM classifiers by maintaining a fixed number of samples to compute prototypes or centroids. In this work, we neither learn a distance metric or finetune representations using the NCM objective; instead our Learned NCM directly updates the centroid locations, which overcomes the uncertainty in the empirical centroid estimates.

3 STRUCTURE OF BIASED REPRESENTATIONS

In long-tailed recognition it is common practice to initially train deep models with regular sampling to learn representations. However, the structure of deep representations learned using regular sampling from imbalanced datasets is not equivalent across all the classes. Most prior work assumes that the local class-specific neighborhoods in the learned representation space are equivalent. They aim to correct the bias in the learned softmax classifier in a second stage by retraining it with uniform sampling and ignore the bias in the representation space.

Suppose the representation space is parameterized by a deep neural network $f_\theta : \mathbb{R}^{3\times H\times W} \longrightarrow \mathbb{R}^d$ with parameters θ and output dimensionality d. We expect that the *number of samples in a class affect the local class-specific neighborhoods* in the learned representation space. More specifically, we are interested in the *intra-class variance* σ_y^θ, which is a quantitative measure of compactness around the class centroid μ_y^θ, as a function of the class frequency N_y. The superscript θ is dropped when it's clear from the context. We now formally define these variables of interest:

$$\mu_y^\theta = \frac{1}{N_y}\sum_{y_i=y} f_\theta(x_i), \qquad \Sigma_y = \frac{(f_\theta(X_y)-\mu_y^\theta)^T(f_\theta(X_y)-\mu_y^\theta)}{N_y}, \qquad \sigma_y^\theta = \max_{i=1}^{d} |\Delta_y^i| \tag{1}$$

where X_y denotes the training images for class y, Σ_y is the maximum likelihood estimator of the sample covariane matrix, and $\Sigma_y = Q_y \Delta_y {Q_y}^T$ is its eigenvalue decomposition.

In Figure 2, we plot the intra-class variance σ_y^θ versus the class frequency N_y on the CIFAR10-LT dataset, with the imbalance ratios 200, 100 and 10 respectively denoting ratio of $\frac{\max N_y}{\min N_y}$. The results indicate that there is a negative correlation between σ_y^θ and N_y, with head classes representations being more compact hence having small variance and tail classes more diffused hence large variance.

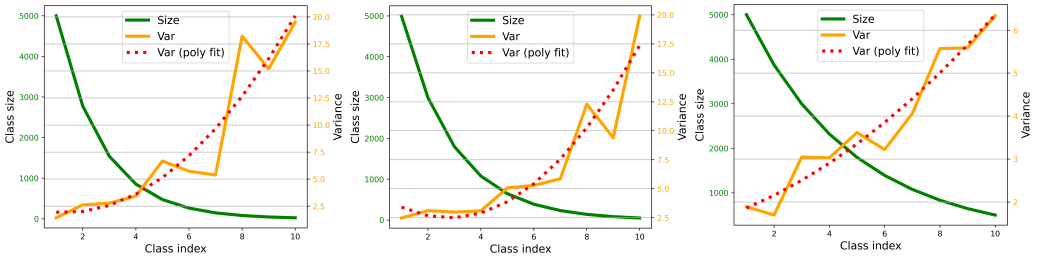


Figure 2: Intra-class variance of representations vs class frequency for long-tailed CIFAR10-LT dataset. Left, middle and right correspond to imbalance ratios of 200, 100 and 10 respectively. Variance is negatively correlated to class frequency.

3.1 Local projection to reduce tail variance

We hypothesize that higher tail class variance in representation space is due to overfitting on high frequency noise in data starved tail classes. To confirm our intuition, we learn a local projection in representation space to throw out undesirable dimensions of variation around the class centroid. Below we provide some preliminaries and then detail this model.

The Nearest-Class-Mean (NCM) classifier with Euclidean distance metric uses the following rule:

$$d_{xy} = \|x - \mu_y^\theta\|_2 = \sqrt{(x - \mu_y^\theta)^T (x - \mu_y^\theta)} \qquad y^* = \underset{y \in \{1,\ldots,\bar{y}\}}{\text{argmin}}\ d_{xy} \tag{2}$$

Mensink et al. (2013) propose to learn a low-rank Mahalanobis distance metric such that:

$$L = -\frac{1}{N}\sum_i^N \log p(y_i|x_i) \qquad p(y|x) \propto -\frac{1}{2} d_{xy}^W \qquad d_{xy}^W = (x - \mu_y^\theta)^T W^T W (x - \mu_y^\theta) \tag{3}$$

We note that W is independent of the class y and induces a global distance metric. This model assumes that the local neighborhoods around class centroids in the representation space are equivalent.

We now describe our local projection model:

$$p(y|x) \propto -\frac{1}{2}\hat{d}_{xy} \qquad \hat{d}_{xy} = (x - \mu_y^\theta)^T Q_y^T \hat{\Delta}_y Q_y (x - \mu_y^\theta) \qquad \hat{\Delta}_y = \text{diag}\sigma(\{w_y^1, \ldots, w_y^d\}) \tag{4}$$

where Q_y are the eigenvectors of the covariance matrix Σ_y as before, σ denotes the sigmoid operation, and $w_y^1, \ldots, w_y^d$ denote class-specific projection parameters that are learned by minimizing the negative log-likelihood loss function in Eq.3. The local projection parameter w_y^i controls how much dimension i in local eigenbasis Q_y contributes to distance metric $\hat{d}_{xy}$.

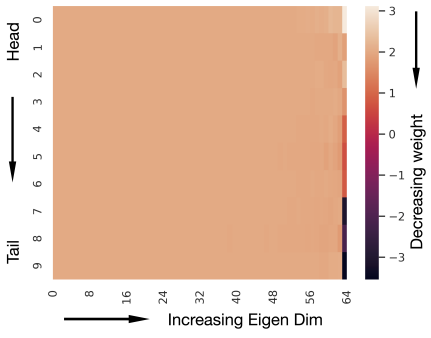


Figure 3: Heatmap of learned local projections on long-tailed CIFAR10-LT with imbalance ratio 200. Higher eigendimensions have little or no contribution for tail classes 4-9.

In Figure 3, we show the learned local projections w_y^i as a heatmap. We note that the higher eigendimensions corresponds to higher eigenvalues. The local projection model learns to downweigh or throw away the contribution of higher eigendimensions only for tail classes while preserving them in head classes. This confirms our intuition that all the dimensions in the learned representations for head classes are meaningful while higher variance dimensions for tail classes correspond to high-frequency noise or spurious correlations. *This high tail variance in the representation space leads to poorer performance on tail classes even after a classifier rebalancing step* Kang et al. (2020).

4 OUR APPROACH

Motivated by the presence of representation bias in LTR, we propose to use a classifier that directly operates in the representation space, the Nearest-Class-Mean classifier. Our approach folows a two-stage framework: (i) Learn representations and softmax classifier jointly with regular sampling, (ii) Fix representations and train the Learned NCM model using uniform sampling.

4.1 LEARNING REPRESENTATIONS

We follow prior work Kang et al. (2020); Ren et al. (2020); Samuel & Chechik (2021) and train a standard model with a softmax classification layer and regular instance-based sampling, i.e, each instance is given equal weight in the training proceduce. Instance-based sampling works better than class-balanced sampling for learning representations. In Figure4, we show the accuracy over manyshot, mediumshot, and fewshot class splits vs the number of training epochs of instance-based sampling. The model first improves accuracy on manyshot data and only after manyshot accuracy saturates it starts to improve accuracy on tail classes, showing an implicit kind of curriculum learning. On the other hand, with class-balanced sampling the optimization gets harder as it lacks this implicit curriculum learning.

4.2 LEARNED NEAREST-CLASS-MEAN FOR LTR

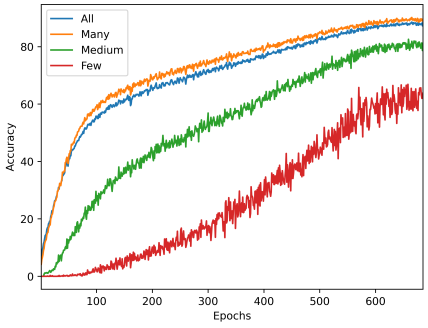


Figure 4: Learning curve for CIFAR100-LT (imbalance ratio 200). Many accuracy goes up immediately, followed by medium and then few.

The Nearest-Class-Mean classifier does not suffer from the bias learned by softmax classifiers, making it an ideal choice for long-tailed recognition. While higher class frequency leads to higher classifier norm Kang et al. (2020) in the softmax classifier, NCM naturally avoids this problem by computing Euclidean or cosine distances to class-centroids. In its most plain form, the class centroids are directly computed from the learned representations in the first stage as given in Eq. 1. However, since the learned representations are biased as discussed in Section 3, this is quite sub-optimal. Alternatively, the class centroids may be learned while keeping the representations fixed and minimizing the negative log-likelihood loss function in Eq.3. We use Euclidean distance instead of squared Euclidean distance due to it's stable gradient, as discussed below. We call this the *Learned NCM* model. Similar to other two-stage models, we use uniform sampling in this step.

4.2.1 GRADIENT UPDATES ON CENTROIDS

In Learned NCM, we use the Euclidean distance in Eq2 due to it's stable gradient optimization. We now describe the gradient updates on the centroids μ_y:

$$\begin{aligned}\partial L_{xy}/\partial \mu_z &= \partial(-\log \frac{e^{\sqrt{2x^T\mu_y - \mu_y^T\mu_y + \text{constants}}}}{Z(x)})/\partial \mu_z \\ &= \frac{\delta_{yz}(1-p(z|x))(\mu_z - x) + (1-\delta_{yz})p(z|x)(x-\mu_z)}{d_{xz}}\end{aligned} \tag{5}$$

$$\mu_z' = \mu_z + \alpha \frac{\delta_{yz}(1-p(z|x)(x-\mu_z)) + (1-\delta_{yz})p(z|x)(\mu_z - x)}{d_{xz}} \tag{6}$$

where $Z(x)$ is the normalization term over all the classes, δ_{yz} is the Kronecker delta, α the learning rate and μ_z' is the updated centroid location. From the gradient update, we can see that if sample x belongs to class z, then the updated centroid is shifted in the direction $(x - \mu_z)$, weighted by misclassification probability $(1 - p(z|x)$, otherwise it gets shifted in the opposite direction $\mu_z - x$ weighted by the misclassification probability $p(z|x)$. In both cases, the magnitude of the shift has a $\|\cdot\|_2$ norm = 1, leading to stable gradient updates.

4.2.2 INTERPRETATION AS LOGISTIC REGRESSION

In actual fact, the NCM classifier is closely related to the logistic regression classifier. We consider below a generalized NCM classifier parameterized by a Mahalanobis distance metric, W, and centroids μ_y,

$$d_{xy}^W = (x - \mu_y)^T W^T W (x - \mu_y) = -2x^T W^T W \mu_y + \mu_y^T W^T W \mu_y + \text{constants} \tag{7}$$

Thus, the NCM classifier is a logistic regression classifier with weight term $W^T W \mu_y$ and bias term $-\frac{1}{2}\mu_y^T W^T W \mu_y$. In Learned NCM, we only learn μ_y and keep the usual Euclidean distance metric by keeping $W = I$. Note that the classification rule remains the same even though we use Euclidean distance instead of squared Euclidean during optimization.

4.3 LOGIT ADJUSTMENT

Recent work has shown the benefits of associating a higher decision boundary margin for tail classes. More precisely, denote by γ_j the margin for class j, and $err(t)$ the probability that the loss exceeds threshold t, then the error on the balanced test set is bounded as follows Ren et al. (2020):

Theorem 1. *Let $t \geq 0$ be any threshold, for all $\gamma_j > 0$, with probability at least $1 - \delta$, we have*

$$err_{bal}(t) \underset{\sim}{\leq} \frac{1}{k}\sum_{j=1}^{k}\left(\frac{1}{\gamma_j}\sqrt{\frac{C}{N_j}} + \frac{\log N}{\sqrt{N_j}}\right); \qquad \gamma_j^* = \frac{\beta N_j^{-1/4}}{\sum_{i=1}^k n_i^{-1/4}} \tag{8}$$

where $err_{bal}(t)$ is the error on the balanced test set, $\underset{\sim}{\leq}$ is used to hide constant terms and C is some measure on complexity. With a constraint on $\sum_{j=1}^k = \beta$, Cauchy-Schwarz inequality gives us the optimal γ_j*.

Therefore, we should set the class margin to be inversely proportional to $1/N_j^{-1/4}$, i.e, higher margin for tail classes. This is achieved by modifying the loss function through a label-dependent offset to the logit score Ren et al. (2020); Menon et al. (2021). We present below the *logit-adjusted cross-entropy loss for Learned NCM*:

$$L_{xy} = -\log \frac{e^{-\frac{1}{2}d_{xy} + \tau \cdot \log N_y}}{\sum_{y'} e^{-\frac{1}{2}d_{xy'} + \tau \cdot \log N_{y'}}} \tag{9}$$

After optimizing this loss, we use the Nearest-Class-Mean argmin d_{xy} as usual for prediction.

The logit-adjusted loss forces a higher weight on tail class samples and enforces the margin constraints in Theorem 1. In practice, we set $\tau = 1/8$ instead of $1/4$ as this gives us better results empirically. We discuss the effect of changing τ further in the Section 5.

4.4 EXTENDING NCM TO MULTIPLE CENTROIDS

The NCM model can be extended to multiple centroids:

$$d_{xy} = \min_i d(x, \mu_y^i) \qquad \mu_y^i = \mu_y + \delta_i, \forall i \in \{1, \ldots, C\} \tag{10}$$

where C is number of centroids per class and δ_i are shared displacements from class-centroid μ_y that are learned. We use the centroids from the learned NCM model for μ_y. The shared displacements reduce model complexity and force shared geometric structure. We call this model Multi-NCM.

5 EXPERIMENTS

5.1 DATASETS AND EVALUATION PROTOCOL

We evaluate our proposed method on the following three benchmark long-tailed datasets:

1. **CIFAR10-LT** Cao et al. (2019): This is a long-tailed split of CIFAR10. CIFAR10 consists of 60K images from 10 classes. Following prior work, we control the degree of data imbalance with an imbalance factor β. $\beta = N_{max}/N_{min}$, where N_{max} and N_{min} are the maximum and minimum number of training images per class respectively. N_{max} is kept fixed at 5000, and $\beta \in [200, 100, 10]$.

2. **CIFAR100-LT** Cao et al. (2019): This is a long-tailed split of CIFAR100. CIFAR100 consists of 60K images from 100 classes. Similar to CIFAR10-LT, we experiment with varying degrees of imbalance $\beta \in [200, 100, 10]$. N_{max} is kept fixed at 500.

3. **ImageNet-LT** Liu et al. (2019): This is a long-tailed split of ImageNet. ImageNet-LT has an imbalanced training set with 115,846 images for 1,000 classes from ImageNet-1K Deng et al. (2009). The class frequencies follow a natural power-law distribution Van Horn & Perona (2017) with a maximum number of 1,280 images per class and a minimum number of 5 images per class. The validation and testing sets are balanced and contain 20 and 50 images per class respectively.

Following prior work, we report average top-1 accuracy on balanced test sets across four splits, *Many*: classes with $\geq$ 100 samples, *Med*: classes with 20 $\sim$ 100 samples, *Few*: classes $<$ 20 samples, and *All* classes.

5.2 Implementation details

Representation learning: We follow prior work and train all models with SGD optimizer and cosine learning rate schedule with instance-balanced sampling. For CIFAR10-LT and CIFAR100-LT we use a ResNet-32 backbone and for ImageNet-LT we use a ResNet-10 backbone.

Learned NCM: We freeze representations and compute empirical class centroids to initialize the learned NCM model. We optimize for the centroids using SGD with gradient clipping above 0.1 and use a learning rate of 5 with momentum 0.9.

Multi-NCM: We freeze the learned centroids from Learned NCM and learn the displacements for Multi-NCM using SGD with gradient clipping above 0.1 and use a learning rate of 5 with momentum 0.9. The number of displacements hyperparameter is fixed at 20 after cross-validation. We also include a non-learnable null displacement to keep the Learned NCM centroids fixed in Multi-NCM.

5.3 Ablation Study

The benefits gained by our various models above the baseline are investigated in an ablation study over CIFAR100-LT, shown in Table1. We make several observations: (i) NCM using the empirical centroids alone makes considerable gains above the softmax, particularly for tail classes. We attribute this to the neural collapse phenomenon Galanti et al. (2021), whereby feature representations collapse towards the class centroid in the limit of training with large number of samples. This is further discussed in Section 6. (ii) Learned-NCM lifts us above NCM significantly on all the metrics, indicating that NCM performance can be boosted quite a bit by learning the centroids and overcoming uncertainty in the empirical centroid estimates. (iii) Multi-NCM improves results over Learned-NCM in most cases, however the improvements are very slight. This indicates that even though Multi-NCM is more expressive, a single centroid usually suffices. (iv) Logit adjustment yields significant improvement for both Learned-NCM and Multi-NCM, especially on the harder tail classes. This validates the design choices of our approach.

Table 1: Ablation study for CIFAR100-LT. Our models with Logit Adjustment are denoted as +LA.

	Imba 200				Imba 100				Imba 10		
Method	All	Many	Med	Few	All	Many	Med	Few	All	Many	Med
Softmax	41.2	76.1	46.6	10.2	46.0	73.8	46.0	13.6	62.3	68.7	48.0
NCM	43.4	64.6	50.4	21.6	48.7	60.7	52.1	30.8	59.0	61.7	52.9
Learned NCM	45.7	66.2	**52.4**	24.6	50.7	64.6	53.1	31.7	62.1	64.5	56.8
+ LA	45.9	62.9	50.8	**28.9**	50.9	60.5	52.9	**37.2**	62.4	63.7	**59.7**
Multi-NCM	45.7	**67.3**	51.8	24.3	50.8	64.5	53.0	32.3	62.4	**64.8**	57.3
+ LA	**46.0**	66.1	51.9	25.7	**51.0**	**64.7**	53.0	32.6	**62.5**	64.6	57.9

5.4 COMPARISON TO STATE-OF-THE-ART

We compare our proposed Learned NCM and Multi-NCM to the following state-of-the-art methods in the solution space: ***(A) Data resampling:*** Decoupling classifier and representation learning via classifier retraining using uniform sampling (**Decoupling**) Kang et al. (2020), ***(B) Loss reshaping:*** **Focal** loss Lin et al. (2017) and Label-distribution aware margin (**LDAM**) loss Cao et al. (2019), ***(C) Ensemble-based models:*** Class-balanced ensemble of experts (**CB Experts**) Sharma et al. (2020) and ***(D) Knowledge transfer:*** Attend over visual memory of class centroids **OLTR** Liu et al. (2019).

We report *All* accuracy results on CIFAR10-LT and CIFAR100-LT for varying imbalance ratios in Table 2. In Table 3 we report *All*, *Many*, *Med* and *Few* accuracy results on CIFAR100-LT (imba 200) and ImageNet-LT. Our Learned-NCM lifts us above the compared state-of-the-art in most cases. Further, Learned-NCM + Logit Adjustment pushes performance on tail classes significantly ahead, gaining 6% over state-of-the-art in tail accuracy on benchmark CIFAR100-LT and ImageNet-LT.

Table 2: Comparison to state-of-the-art on CIFAR10-LT and CIFAR100-LT. We report *All* accuracy on three imbalance ratios $\beta \in [200, 100, 10]$, where $\beta = N_{max}/N_{min}$. † denotes reproduced results. Learned-NCM consistently achieves superior results across varying imbalance ratios.

Dataset	CIFAR10-LT			CIFAR100-LT		
Imbalance ratio	200	100	10	200	100	10
Focal Lin et al. (2017)	65.3	70.4	86.7	35.6	38.4	55.8
LDAM Cao et al. (2019)	-	77.0	88.2	-	42.0	58.7
Decoupling Kang et al. (2020) †	80.2	81.4	91.3	45.3	50.7	**62.5**
NCM	79.7	79.8	91.2	43.4	48.7	59.0
Learned-NCM	80.8	**81.6**	91.4	45.7	50.7	62.1
Learned-NCM + Logit Adjustment	**81.3**	81.4	**91.7**	**45.9**	**50.9**	62.4

Table 3: Comparison to state-of-the-art on CIFAR100-LT and ImageNet-LT. † denotes reproduced results. Learned-NCM consistently achieves superior results, especially on tail classes.

Dataset	CIFAR100-LT				ImageNet-LT			
Method	All	Many	Med	Few	All	Many	Med	Few
Focal Lin et al. (2017)	35.6	60.4	41.7	15.7	30.5	36.4	29.9	16.0
Decoupling Kang et al. (2020) †	45.3	**69.5**	50.6	22.4	**41.2**	**51.7**	37.9	23.1
CB Experts Sharma et al. (2020)	-	-	-	-	39.2	48.2	37.0	21.5
LFME Xiang et al. (2020)	-	-	-	-	38.8	47.0	37.9	19.2
OLTR Liu et al. (2019)	-	-	-	-	35.6	43.2	35.1	18.5
NCM	43.4	64.6	50.4	21.6	34.4	42.5	32.1	19.7
Learned NCM	45.7	66.2	**52.4**	24.6	40.0	48.4	37.7	24.5
Learned NCM + Logit Adjustment	**45.9**	62.9	50.8	**28.9**	39.6	45.1	**38.1**	**29.2**

5.5 EFFECT OF LOGIT ADJUSTMENT WEIGHT

In Figure5, we study the effect of the logit adjustment weight on the different performance metrics. We observe a clear trade-off between *Few* accuracy on one hand and *All*, *Many* and *Med* on the other, with higher weight favoring *Few*. Consider that the logit adjustment $\log N_y$ leads to a higher loss on tail classes during the optimization process. Therefore, increasing the weight parameter leads to tail loss being over-emphasized during training and causing lower tail class error at the price of higher head class error. We choose $\tau = 1/8$ as the optimal weight achieving good accuracy across the spectrum.

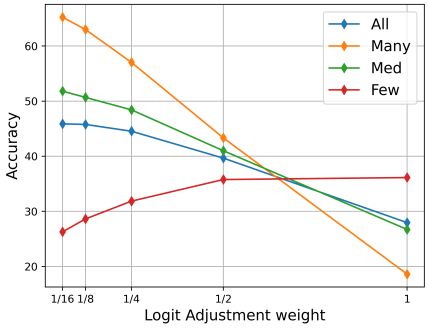
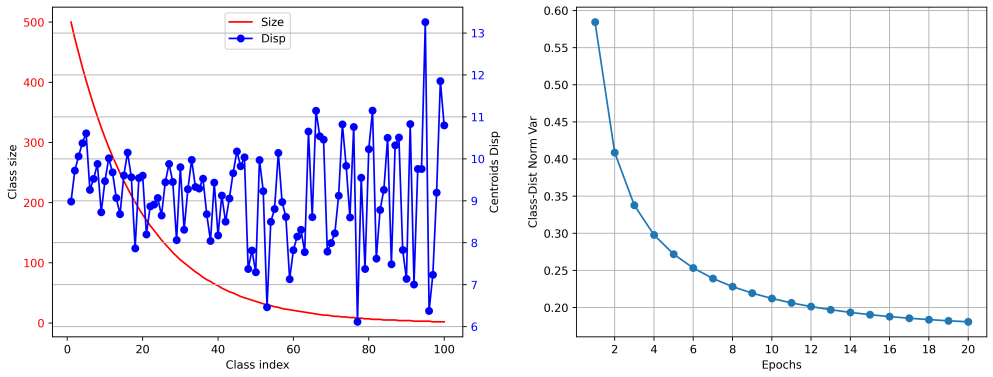


Figure 5: Effect of the logit-adjustment weight τ on CIFAR100-LT.



(a) (b)

Figure 6: L: The displacement of learned centroids from their empirical estimates as measured by the $\|\cdot\|_2$ norm. The class size doesn't affect the magnitude of the displacement, suggesting that empirical estimates of centroids are worse regardless of class size. R: The class-distance normalized variance (CDNV) as a function of training epochs. CDNV decreases as training proceeds, corresponding to greater separation between learned class centroids and thus lower classification error.

5.6 Centroid displacement in Learned NCM

In Figure6a, we investigate how far centroids move away from their empirical estimates in Learned NCM. Even though learned centroids are optimized solely for classification, they are expected to stay close to the empirical centroids which are good candidate centroids. It is interesting to see that there is actually considerable displacement from their original position. This validates our hypothesis that learning the centroids helps in overcoming the uncertainty in the empirical centroid estimate. However, even though the estimation error grows as $1/N_y$ and is expected to be worse for tail classes, there is no correlation between class size and centroid displacement.

6 Discussion

In this work, we showed that representations learned from long-tailed datasets exhibit bias in the form of higher intra-class variance for tail classes. High tail variance hurts tail class accuracy and is due to spurious correlations learned by deep models. We proposed to mitigate representation bias in LTR by the Learned Nearest-Class-Mean classifier, which overcomes uncertainty in empirical centroid estimates and jointly learns centroids minimizing the average class-distance normalized variance (CDNV). Our work aligns with recent theoretical resultsGalanti et al. (2021) that bound the generalization error on the NCM model using the average pairwise CDNV. Briefly, the error of NCM on the test set is bounded above by the average CDNV, multiplied with terms dependent on the number of classes and average class size. Informally, the CDNV between two classes y_1 and y_2 is defined as $\frac{Var(y_i)+Var(y_j)}{2\|\mu_{y_i}-\mu_{y_j}\|^2}$. Roughly, if the *inter-centroid separation* of two classes is higher than the sum of their *intra-class variances*, the CDNV is expected to be lower. In Learned-NCM, we only learn the centroids and assume that the intra-class variance does not change. In Figure6b, we show that the Learned NCM indeed leads to lower CDNV as the number of training epochs increase, and consequently lower classification error.

We also note that training with larger amounts of data leads to the neural collapse phenomenon Papyan et al. (2020); Galanti et al. (2021), whereby feature representations cluster close to the class centroid and the average CDNV approaches 0. Learned NCM can be an effective tool for long-tailed recognition where high tail variance can lead to the opposite of neural collapse. We hope our work deepens the understanding into representation bias and the efficacy of Learned NCM in long-tailed recognition, and offers inspiration for further work in this regard.

REFERENCES

Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Arechiga, and Tengyu Ma. Learning imbalanced datasets with label-distribution-aware margin loss. In *NeurIPS*, 2019.

Varun Chandola, Arindam Banerjee, and Vipin Kumar. Anomaly detection: A survey. *ACM computing surveys (CSUR)*, 41(3):1–58, 2009.

Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. *JAIR*, 2002.

Nitesh V Chawla, Aleksandar Lazarevic, Lawrence O Hall, and Kevin W Bowyer. Smoteboost: Improving prediction of the minority class in boosting. In *European conference on principles of data mining and knowledge discovery*, 2003.

Yilun Chen, Ami Wiesel, Yonina C Eldar, and Alfred O Hero. Shrinkage algorithms for mmse covariance estimation. *IEEE Transactions on Signal Processing*, 58(10):5016–5029, 2010.

Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. Class-balanced loss based on effective number of samples. In *CVPR*, 2019.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009.

Chris Drummond, Robert C Holte, et al. C4. 5, class imbalance, and cost sensitivity: why under-sampling beats over-sampling. In *Workshop on learning from imbalanced datasets II*, 2003.

Charles Elkan. The foundations of cost-sensitive learning. In *International joint conference on artificial intelligence*, volume 17, pp. 973–978. Lawrence Erlbaum Associates Ltd, 2001.

Andrew Estabrooks, Taeho Jo, and Nathalie Japkowicz. A multiple resampling method for learning from imbalanced data sets. *Computational intelligence*, 2004.

Tomer Galanti, András György, and Marcus Hutter. On the role of neural collapse in transfer learning. In *International Conference on Learning Representations*, 2021.

Samantha Guerriero, Barbara Caputo, and Thomas Mensink. Deepncm: Deep nearest class mean classifiers. In *International Conference on Learning Representations Workshop*, 2018.

Haibo He and Edwardo A Garcia. Learning from imbalanced data. *TKDE*, 2009.

Bingyi Kang, Saining Xie, Marcus Rohrbach, Zhicheng Yan, Albert Gordo, Jiashi Feng, and Yannis Kalantidis. Decoupling representation and classifier for long-tailed recognition. In *Eighth International Conference on Learning Representations (ICLR)*, 2020.

Salman H Khan, Munawar Hayat, Mohammed Bennamoun, Ferdous A Sohel, and Roberto Togneri. Cost-sensitive learning of deep feature representations from imbalanced data. *TNNLS*, 2017.

Olivier Ledoit and Michael Wolf. A well-conditioned estimator for large-dimensional covariance matrices. *Journal of multivariate analysis*, 88(2):365–411, 2004.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *ICCV*, 2017.

Charles X Ling and Victor S Sheng. Cost-sensitive learning and the class imbalance problem. *Encyclopedia of machine learning*, 2011:231–235, 2008.

Bo Liu, Haoxiang Li, Hao Kang, Gang Hua, and Nuno Vasconcelos. Gistnet: a geometric structure transfer network for long-tailed recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8209–8218, 2021.

Ziwei Liu, Zhongqi Miao, Xiaohang Zhan, Jiayun Wang, Boqing Gong, and Stella X Yu. Large-scale long-tailed recognition in an open world. In *CVPR*, 2019.

Andrew Maas, Raymond E Daly, Peter T Pham, Dan Huang, Andrew Y Ng, and Christopher Potts. Learning word vectors for sentiment analysis. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies*, pp. 142–150, 2011.

Aditya Krishna Menon, Sadeep Jayasumana, Ankit Singh Rawat, Himanshu Jain, Andreas Veit, and Sanjiv Kumar. Long-tail learning via logit adjustment. In *International Conference on Learning Representations*, 2021.

Thomas Mensink, Jakob Verbeek, Florent Perronnin, and Gabriela Csurka. Distance-based image classification: Generalizing to new classes at near-zero cost. *TPAMI*, 2013.

Vardan Papyan, XY Han, and David L Donoho. Prevalence of neural collapse during the terminal phase of deep learning training. *Proceedings of the National Academy of Sciences*, 117(40): 24652–24663, 2020.

Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H Lampert. icarl: Incremental classifier and representation learning. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2017.

Jiawei Ren, Cunjun Yu, Xiao Ma, Haiyu Zhao, Shuai Yi, et al. Balanced meta-softmax for long-tailed visual recognition. *Advances in Neural Information Processing Systems*, 33:4175–4186, 2020.

Dvir Samuel and Gal Chechik. Distributional robustness loss for long-tail learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021.

Saurabh Sharma, Ning Yu, Mario Fritz, and Bernt Schiele. Long-tailed recognition using class-balanced experts. In *DAGM German Conference on Pattern Recognition*. Springer, 2020.

Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. In *NeurIPS*, 2017.

Qianru Sun, Yaoyao Liu, Tat-Seng Chua, and Bernt Schiele. Meta-transfer learning for few-shot learning. In *CVPR*, 2019.

Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

Grant Van Horn and Pietro Perona. The devil is in the tails: Fine-grained classification in the wild. *arXiv preprint arXiv:1709.01450*, 2017.

Xudong Wang, Long Lian, Zhongqi Miao, Ziwei Liu, and Stella Yu. Long-tailed recognition by routing diverse distribution-aware experts. In *International Conference on Learning Representations*, 2020.

Yu-Xiong Wang, Deva Ramanan, and Martial Hebert. Learning to model the tail. In *NeurIPS*, 2017.

Yongqin Xian, Saurabh Sharma, Bernt Schiele, and Zeynep Akata. f-vaegan-d2: A feature generating framework for any-shot learning. In *CVPR*, 2019.

Liuyu Xiang, Guiguang Ding, and Jungong Han. Learning from multiple experts: Self-paced knowledge distillation for long-tailed classification. In *European Conference on Computer Vision*. Springer, 2020.

Xi Yin, Xiang Yu, Kihyuk Sohn, Xiaoming Liu, and Manmohan Chandraker. Feature transfer learning for face recognition with under-represented data. In *CVPR*, 2019.

Boyan Zhou, Quan Cui, Xiu-Shen Wei, and Zhao-Min Chen. Bbn: Bilateral-branch network with cumulative learning for long-tailed visual recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9719–9728, 2020.

A Appendix

A.1 Dataset statistics

We detail the dataset statistics for the three benchmark long-tailed recognition datasets in Table 4.

Table 4: Statistics for training data in CIFAR10-LT, CIFAR100-LT and ImageNet-LT.

Dataset	Attribute	Many	Medium	Few	All
CIFAR10-LT (Imba 200)	Classes	7	3	0	10
	Samples	11052	151	0	11203
CIFAR10-LT (Imba 100)	Classes	8	2	0	10
	Samples	12273	133	0	12406
CIFAR10-LT (Imba 10)	Classes	10	0	0	10
	Samples	20431	0	0	20431
CIFAR100-LT (Imba 200)	Classes	31	30	39	100
	Samples	7753	1445	304	9502
CIFAR100-LT (Imba 100)	Classes	35	35	30	100
	Samples	8824	1718	305	10847
CIFAR100-LT (Imba 10)	Classes	70	30	0	100
	Samples	17743	2130	0	19573
ImageNet-LT	Classes	391	473	136	1,000
	Samples	89,293	24,910	1,643	115,846

A.2 Estimation of intra-class variance

Estimating the covariance matrix from sample data is a non-trivial problem. In this work, we choose the empirical estimator of sample covariance as described in Eq1, which is the maximum likelihood estimator. However, alternate estimators such as the Ledoit-Wolf Ledoit & Wolf (2004) and Oracle Approximating Shrinkage Chen et al. (2010) estimators are also commonly used to estimate covariance. We did not find any difference in our results due to the choice of estimator; the intra-class variance in Eq2 is always negatively correlated to the class frequency.

In Figure7, we plot the intra-class variance for long-tailed CIFAR100-LT with imbalance ratios $\beta \in [200, 100, 10]$. Since this is a more fine-grained dataset with semantic overlap between classes of the sort orchids and poppies or bicycle and motorcycle, the representations of various classes can overlap and affect the estimation of intra-class variance. Therefore, we also consider the 20 superclasses of CIFAR100 to construct CIFAR20-LT, which is coarse-grained and lacks semantic overlap. In Figure8 we plot the the intra-class variance for long-tailed CIFAR20-LT for the various imbalance ratios. Combined, Figure7 and Figure8 indicate that the negative correlation of intra-class variance to class frequency is not dataset specific and is a more general phenomenon. The high degree of variation in the intra-class variance estimate is attributed to (i) the inverse scaling of the MSE in variance estimation to the class frequency, and (ii) the semantic overlap between various classes due to which intra-class variance is not purely class-conditional.

A.3 Learning distance metric and NCM jointly

Beyond learning just class centroids in Learned NCM, we investigated learning the Mahalanobis distance metric in Eq 3 jointly with the centroids. More precisely, we learn the matrix W which parameterizes the Mahalanobis distance. We experimented with three strategies: (i) Global distance metric shared by all classes, and (ii) Local distance metrics, corresponding to a class-conditional matrix W_y for each class y. Our results for long-tailed CIFAR10-LT and CIFAR100-LT (imba 200 for both) are summarized in Table 5.

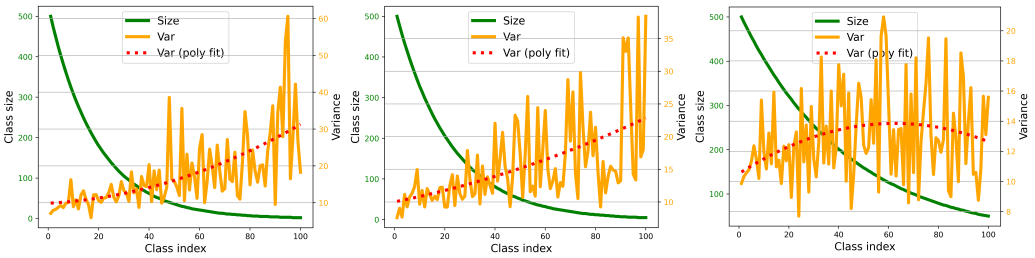


Figure 7: Intra-class variance of representations vs class frequency for long-tailed CIFAR100-LT dataset. Left, middle and right correspond to imbalance ratios of 200, 100 and 10 respectively. Variance is negatively correlated to class frequency.

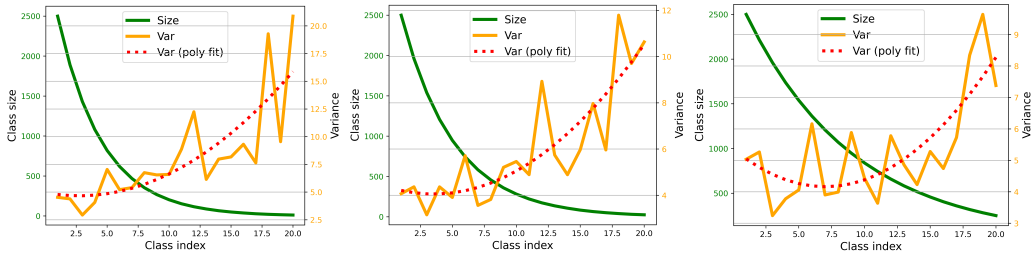


Figure 8: Intra-class variance of representations vs class frequency for long-tailed CIFAR20-LT dataset, consisting of 20 superclasses from the CIFAR100 dataset. Left, middle and right correspond to imbalance ratios of 200, 100 and 10 respectively. Variance is negatively correlated to class frequency.

The results indicate that the learned distance metric only improves *Many* accuracy and in all other cases underperforms Learned NCM. This suggests that Learned NCM is sensitive to choice of distance metric, and keeping the Euclidean distance metric leads to the best results for Learned NCM.

Table 5: Comparision of jointly learned distance metrics and NCM on long-tailed CIFAR10-LT and CIFAR100-LT with imbalance ratio 200.

Dataset	CIFAR10-LT			CIFAR100-LT			
Method	All	Many	Med	All	Many	Med	Few
NCM	79.7	82.2	73.8	43.4	64.6	50.4	21.6
Learned NCM	**80.8**	82.2	**77.6**	**45.7**	66.2	**52.4**	**24.6**
Global Metric + Learned NCM	76.6	**83.1**	61.3	43.4	69.7	48.8	18.9
Local Metric + Learned NCM	80.2	82.1	75.9	43.8	**69.9**	51.0	18.0

A.4 Effect of batch normalization

Batch normalization uses running estimates of the mean and standard deviations statistics to normalize intermediate activations for deep models. For two-stage models used in long-tailed recognition, the batch statistics are used only in stage 1 and after that are kept fixed. However, during training the representations are evolving and so are the batch statistics. Therefore, we experiment with *posthoc* running estimates of mean and standard deviations in the second stage. Since the neural network parameters θ are fixed, the estimates are more precise and can moreover alleviate the biased representations issue discussed in Section 3.

The results are detailed in Table 6. We observe gain in *Few* accuracy due to BN in both Learned NCM and Multi NCM, and gain in *All* accuracy as well. This aligns with our intuition that proper batch normalization can mitigate representation bias in LTR and points to future research directions.

Table 6: Results on the long-tailed CIFAR10-LT and CIFAR100-LT dataset. BN indicates we use posthoc running estimates of mean and standard deviation for the batchnorm layer.

Dataset	CIFAR10-LT			CIFAR100-LT			
Method	All	Many	Medium	All	Many	Medium	Few
Learned NCM	80.8	82.2	77.6	45.7	66.2	52.4	24.6
Learned NCM (BN)	79.7	78.2	**83.2**	45.6	63.4	**52.6**	**26.5**
Multi-NCM	80.8	**82.2**	77.6	45.7	**67.3**	51.8	24.2
Multi-NCM (BN)	**81.4**	81.1	82.2	**45.8**	65.3	52.1	25.8

A.5 Detailed results on CIFAR10-LT

Table 7: Extended results on the long-tailed CIFAR10-LT dataset.

	Imba 200			Imba 100			Imba 10
Method	All	Many	Medium	All	Many	Medium	All
Softmax	74	**83**	52.9	80.3	**81.7**	74.6	90.3
NCM	79.7	82.2	73.8	79.8	79.0	82.8	91.2
Learned NCM	80.8	82.2	77.6	81.6	81.0	83.9	91.4
+LA	**81.3**	81.8	**79.9**	81.4	79.6	**87.4**	**91.7**
Multi-NCM	80.8	82.2	77.6	**81.6**	81.2	83.1	91.3
+LA	81.1	82.4	77.8	81.4	80.1	87.0	91.6