Classes Are Not Equal: An Empirical Study on Image Recognition Fairness

Jiequan Cui¹  Beier Zhu¹  Xin Wen²  Xiaojuan Qi²  Bei Yu³  Hanwang Zhang¹
Nanyang Technological University¹  The University of Hong Kong²  The Chinese University of Hong Kong³

Abstract

In this paper, we present an empirical study on image recognition unfairness, i.e., extreme class accuracy disparity on balanced data like ImageNet. We demonstrate that classes are not equal and unfairness is prevalent for image classification models across various datasets, network architectures, and model capacities. Moreover, several intriguing properties of fairness are identified. First, the unfairness lies in problematic representation rather than classifier bias distinguished from long-tailed recognition. Second, with the proposed concept of Model Prediction Bias, we investigate the origins of problematic representation during training optimization. Our findings reveal that models tend to exhibit greater prediction biases for classes that are more challenging to recognize. It means that more other classes will be confused with harder classes. Then the False Positives (FPs) will dominate the learning in optimization, thus leading to their poor accuracy. Further, we conclude that data augmentation and representation learning algorithms improve overall performance by promoting fairness to some degree in image classification.

1. Introduction

In the past decade, significant advancements have been made in image recognition. Researchers have diligently explored a multitude of techniques to continually enhance the recognition capabilities of deep models, such as data augmentations [7, 46, 47], the evolution of model architectures [17, 18, 25, 41], and representation learning [13, 19, 20, 34]. However, it’s worth noting that most of these endeavors primarily focus on achieving state-of-the-art overall accuracy, often overlooking the goal of consistent performance across all interested classes.

Recently, long-tailed recognition has garnered significant attention within the realms of computer vision and machine learning, due to its growing relevance in real-world applications. In practical scenarios, data often exhibits a long-tailed distribution, i.e., a few classes occupy plenty of data while most of the classes only have a few samples. Models trained on long-tailed data show extremely imbalanced performance. Especially, the accuracies of the low-frequency classes are pretty poor. However, as shown in Figure 1, we identify that a significant accuracy disparity can occur even on balanced datasets such as ImageNet [15]. The best class achieves 100% top-1 accuracy while the worst class only achieves 16% top-1 accuracy with a ResNet-50 model on ImageNet. This phenomenon suggests that factors beyond class frequency contribute to imbalanced performance. We hope that our findings will draw further attention from the research community to this pressing issue.

In this paper, we empirically investigate the image classification fairness issue, i.e., extreme accuracy disparity among classes, on 8 balanced datasets including CIFAR and ImageNet. Our analysis also encompasses vision-language models, i.e., CLIP [37] and the stable diffusion model [38]. With a range of network architectures (CNNs and vision transformers), model capacities, and datasets, we identify this as a widespread issue in image classification models.

Representation or classifier bias? Inspired by long-tailed recognition, we investigate the root of unfairness regarding classifier bias including $\ell_2$-norm of classifier weight and class separation angles. Challenging the conventional belief that larger separation angles correlate with higher accuracy, our findings reveal an intriguing contrast: classes with lower performance exhibit larger separation angles. Confirmed with $k$-nearest neighbors ($k$-NN) algorithm and ETF classifier, we conclude that the problematic representation serves as a crucial factor contributing to the unfairness.

Figure 1. The unfairness is prevalent in image classification across various datasets, network architectures, and model capacities. We sort classes by the performance of WideResNet-34-10 on CIFAR-100 and ResNet-50 on ImageNet. For CLIP models, we sort classes by the zero-shot performance on ImageNet of the CLIP ResNet-50 model. Note that data rebalancing is considered in the collection of WIT-400M [37].
We observe that data augmentations and representation learning algorithms achieve better fairness with improved overall accuracy while re-weighting sacrifices the accuracy of easy classes.

2. Related Work

**Fairness.** While unfairness is typically defined in terms of disparities related to sensitive attributes, like gender, race, disabilities, and sexual or political orientation, it is a growing concept. In federated learning [28], fairness can be a measurement of the degree of uniformity in performance across federated client devices. For group distributionally robust optimization (DRO), spurious correlations are studied with group labels on the Waterbirds [40]. In this paper, the unfairness is defined as the class accuracy disparities for image classification models.


**Data Augmentation and Representation Learning.** Data augmentations [7, 8, 16, 46, 47] can significantly improve the generalization ability of models taking nearly no additional cost. Single-image augmentations [7, 8, 16] do image-processing operations on each image independently. Cross-image augmentations [46, 47] mix images and their labels simultaneously. Representation learning is a fundamental task. Contrastive loss [6, 19] measures the similarities of sample pairs from feature space, and positive sample pairs are encouraged to have similar representations. Recently, masked modeling transferred from the neural language processing (NLP) community has made great progress in learning good representation for images, like [1, 20], benefiting a wide spectrum downstream tasks.

Data augmentations and representation learning techniques both improve feature quality for recognition, thus promoting overall accuracy. However, per-class accuracy is usually ignored and we bridge the gap in this paper.

3. Fairness Issue Widely Exists

To showcase the widespread prevalence of fairness issues in image classification models, we conduct experiments on 8 datasets including CIFAR-100 [24] and ImageNet [39] with various network architectures including convolutional neu-
(e) “Hard” of ImageNet
(c) “Easy” of ImageNet
(d) “Medium” of ImageNet
(e) “Few” of ImageNet-LT

Figure 3. Analysis on ImageNet-LT. (a) Data distribution and per-class accuracy with ResNet-50 on ImageNet-LT; (b) The $\ell_2$-norm of classifier weights before/after classifier rebalancing; (c), (d), and (e) evaluate the prediction bias on “Many”, “Medium”, and “Few” classes data separately. We sort classes by the number of samples in the classes on ImageNet-LT.

Figure 4. Model prediction bias. (a) and (b) show the relationship between class accuracy and prediction bias. ImageNet-LT and ImageNet appear contrary conclusions. The details are discussed in Section 4.2. (c), (d), and (e) evaluate the prediction bias on “Easy”, “Medium”, and “Hard” classes data separately. ResNet-50 is used for ImageNet-LT while ViT-B is adopted on ImageNet. Classes are sorted by their frequency on ImageNet-LT. On ImageNet, classes are sorted by the performance of ResNet-50.

Table 1. Top-1 accuracy on ImageNet and CIFAR-100. CLIP models trained on large-scale image-text pairs are also included.

<table>
<thead>
<tr>
<th>Method</th>
<th>Easy</th>
<th>Medium</th>
<th>Hard</th>
<th>All</th>
<th>Best</th>
<th>Worst</th>
</tr>
</thead>
<tbody>
<tr>
<td>ResNet-50</td>
<td>93.1</td>
<td>81.1</td>
<td>59.4</td>
<td>77.8</td>
<td>100.0</td>
<td>16.0</td>
</tr>
<tr>
<td>ResNet-101</td>
<td>94.1</td>
<td>82.6</td>
<td>61.5</td>
<td>79.4</td>
<td>100.0</td>
<td>16.0</td>
</tr>
<tr>
<td>ViT-B</td>
<td>95.4</td>
<td>86.6</td>
<td>68.9</td>
<td>83.6</td>
<td>100.0</td>
<td>20.0</td>
</tr>
<tr>
<td>CIFAR-100</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ResNet-50</td>
<td>89.8</td>
<td>81.4</td>
<td>68.5</td>
<td>79.8</td>
<td>98.0</td>
<td>51.0</td>
</tr>
<tr>
<td>WideResNet-34-10</td>
<td>92.2</td>
<td>83.2</td>
<td>70.1</td>
<td>81.7</td>
<td>98.0</td>
<td>57.0</td>
</tr>
<tr>
<td>CLIP model on ImageNet</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ResNet-50</td>
<td>84.3</td>
<td>61.8</td>
<td>33.4</td>
<td>59.8</td>
<td>100.0</td>
<td>0.0</td>
</tr>
<tr>
<td>ResNet-101</td>
<td>83.6</td>
<td>63.1</td>
<td>40.3</td>
<td>62.3</td>
<td>100.0</td>
<td>0.0</td>
</tr>
</tbody>
</table>

We use CNNs (ResNet-50 and ResNet-101) and vision transformers (ViT-B) as our backbones. For CNNs, we follow the training schedule from [46]. Models are trained in 300 epochs with a cosine learning rate strategy and a batch size of 256 on 8 GPUs. The initial learning rate and the weight decay are set to 0.1 and 1e-4 separately. The SGD optimizer with a momentum of 0.9 is used.

For training with ViT-B, we follow the training schedule in [20], i.e., firstly pre-train it with masked modeling technique and then fine-tune it 100 epochs. Our implementation is built on their open-sourced code.

Evaluation. In the long-tailed recognition field, “Many”, “Medium”, and “Few” classes are defined for evaluation according to the number of samples in the classes. Similarly, to quantify the unfairness, we define “Hard”, “Medium”, and “Easy” classes with a well-trained reference model. ResNet50 and WideResNet34-10 are adopted as the reference models for ImageNet and CIFAR-100 separately. The number of classes in “Hard”, “Medium”, and “Easy” are evenly divided according to the performance of the reference model. For CLIP [37] models, we directly evaluate their pre-trained model on the ImageNet validation set and sort the classes with the performance of the ResNet-50 CLIP model.

SGD optimizer with a momentum of 0.9 in training.

Training on ImageNet. ImageNet [39] is one of the most challenging datasets for classification, which consists of 1.2 million images for training and 50K images for validation over 1,000 classes. It is worth noting that ImageNet is a carefully curated dataset. Data is balanced over all classes and each class has nearly 1000 training samples.

We use CNNs (ResNet-50 and ResNet-101) and vision transformers (ViT-B) as our backbones. For CNNs, we follow the training schedule from [46]. Models are trained in 300 epochs with a cosine learning rate strategy and a batch size of 256 on 8 GPUs. The initial learning rate and the weight decay are set to 0.1 and 1e-4 separately. The SGD optimizer with a momentum of 0.9 is used.

For training with ViT-B, we follow the training schedule in [20], i.e., firstly pre-train it with masked modeling technique and then fine-tune it 100 epochs. Our implementation is built on their open-sourced code.

Evaluation. In the long-tailed recognition field, “Many”, “Medium”, and “Few” classes are defined for evaluation according to the number of samples in the classes. Similarly, to quantify the unfairness, we define “Hard”, “Medium”, and “Easy” classes with a well-trained reference model. ResNet50 and WideResNet34-10 are adopted as the reference models for ImageNet and CIFAR-100 separately. The number of classes in “Hard”, “Medium”, and “Easy” are evenly divided according to the performance of the reference model. For CLIP [37] models, we directly evaluate their pre-trained model on the ImageNet validation set and sort the classes with the performance of the ResNet-50 CLIP model.
Results Summary. We summarize the experimental results in Figure 1 and Table 1. As plotted in Figure 1, from easy to hard classes, the accuracy significantly decreases for models trained on CIFAR-100, ImageNet, and WIT-400M. The phenomenon is consistent across various backbones including CNNs and vision transformers. Especially, the best class “ostrich” achieves 100% top-1 accuracy while the worst class “screen” only obtains 16% top-1 accuracy on ImageNet with ResNet-50, which demonstrates a large performance disparity on balanced data.

As shown in Table 1, “Easy” classes can get much better performance than “Hard” classes although training samples are evenly distributed among all the classes. For the CLIP ResNet-50 model, the average accuracy of “Easy” classes is 50% higher than that of “Hard” classes. The huge accuracy gap between the best (100%) and the worst (0%) classes indicates a surprising performance imbalance on balanced datasets. It is worth noting that data rebalancing strategies have already been applied in the collection of WIT-400M. Analysis of the other five datasets is included in the supplementary file.

With these experimental analyses, we conclude that the fairness issue in image classification is a general problem across various datasets, network architectures, and model capacities. Further, the similar trends on the same datasets with various models demonstrate that the unfairness highly depends on training data distribution. We expect that our study could raise more attention from the community to this issue and facilitate progress.

Discussion. When trained with long-tailed data, models tend to perform well on high-frequency classes but show significantly reduced accuracy on low-frequency ones. This pattern suggests to us that class frequency can be the primary driver of the extremely imbalanced performance. Although several works start to consider hard example learning, most works [4, 11, 14, 22] in long-tailed recognition still focus on this aspect. However, our research reveals that significant performance disparities (100% accuracy on the best class while only 16% on the worst class on ImageNet) also occur in balanced datasets. This indicates the presence of other significant factors that influence per-class accuracy, beyond just class frequency.

4. Analysis on Image Recognition Fairness

As shown in Figure 1 and Figure 3 (a), extreme performance disparity happens both in long-tailed recognition and unfairness. We delve into the properties of fairness in image classification by comparing it with long-tailed recognition. The discussion on whether bias in representation or classifier leads to unfairness is presented in Section 4.1. Following that, we conduct an in-depth examination of the underlying optimization challenges for fairness in Section 4.2.  

4.1. Representation or Classifier Bias?

Inspired by long-tailed recognition, we diagnose the classifier bias for unfairness in terms of their weights $\ell_2$-norm and class separation angles. Then we conclude that the problematic representation results in unfairness.

$\ell_2$-norm of Classifier Weights. Previous works [4, 22, 49] study the classifier bias in long-tailed learning and observe $\ell_2$-norm of classifier weights matters. Kang et al. [22] propose a two-stage pipeline: uniform data sampling for representation learning, and then class-balanced re-sampling for classifier learning with previously learned representation fixed. With the method, we train a ResNet-50 model on ImageNet-LT [30], which obeys the long-tailed distribution shown in Figure 3(a).

With the uniform sampling strategy for representation learning, the $\ell_2$-norm of the learned classifier weights positively correlates to class accuracy and class frequency as plotted in Figure 3(a) and Figure 3(b). After classifier re-balancing using the re-sampling strategy, the norm of classifier weights becomes more uniform, which is visualized in Figure 3(b), and the accuracies of low-frequency classes are promoted.

Interestingly, although the models suffering from unfairness display extremely imbalanced accuracy varying from 100% to 16%, we notice that the $\ell_2$-norm of classifier weights is already balanced for models trained on ImageNet shown in Figure 5(a), implying that the $\ell_2$-norm of classifier weights is not the core factor for unfairness.

Definition 1 (Simplex Equiangular Tight Frame). A collection of vectors $m_i \in \mathbb{R}^d$, $i = 1, 2, \ldots, K, d \leq K - 1$, is said to be a simplex equiangular tight frame if:

$$M = \sqrt{\frac{K}{K-1}} U \left( I_K - \frac{1}{K} I_K I_K^T \right),$$

where $M = [m_1, \ldots, m_K] \in \mathbb{R}^{d \times K}$ allows a rotation and satisfies $U^T U = I_K$, $I_K$ is the identity matrix, and $1_K$ is an all-ones vector.

All vectors in a simplex ETF have an equal $\ell_2$-norm and the same pair-wise angle, i.e., $m_i^T m_j \propto -\frac{1}{K-1}$ if $i \neq j$.

Maximal Equiangular Separation. The simplex equiangular tight frame (ETF) in Definition 1 is introduced by Papan et al. [35]. Then several works [45, 50] explore the ETF classifier for long-tailed recognition. ETF classifier enjoys the same pair-wise angle for any two different classes and allows low-frequency classes to have the maximal equiangular separation.

We examine if bias in unfairness comes from the smaller separation angles between “Hard” and other classes by conducting experiments on ImageNet. The mean angle for class
i, denoted by CMA(i), is defined as follows:

$$\text{CMA}(i) = \frac{1}{C} \sum_{j} \frac{w_{i} \cdot w_{j}}{||w_{i}||_2 \cdot ||w_{j}||_2},$$

(2)

where $w_i$ is the classifier weight for class $i$, and $C$ is the number of classes.

As shown in Figure 5(b), most of the “Hard” classes get smaller CMA values than “Easy” ones and thus they even possess larger separation angles with other classes, which is quite contrary to our common belief: larger separation angles usually indicate a higher accuracy.

We hypothesize that the problematic representations lead to the contrary. When samples in two classes look similar or have overlapping scenarios, the model can be confused. During optimization, the cross-entropy loss will enlarge the separation angles between the two classes and meanwhile encourage their representations more distinguishable from each other. However, if there is little effect on the representations, the two classes still suffer from poor accuracy despite large separation angles between them.

**Problematic Representation in Unfairness.** We verify the hypothesis with the following empirical studies:

• Evaluate the model with $k$-NN algorithm and examine the existence of unfairness.

• Train models with a fixed ETF classifier and examine the existence of unfairness.

For the $k$-NN algorithm, with $k$ nearest neighbors in training data, the prediction is determined on the basis of a majority vote, i.e., the label most frequently represented around its neighbors. Following DINO [5], $k$ is set to 20 for evaluation. ETF classifier satisfies the maximal equiangular separation principle. We follow the implementation in previous works [45, 50]. As shown in Figure 5(c) and Figure 5(d), the class performance still exhibits extreme imbalance, confirming that problematic representation is the core issue for unfairness.

**Discussion.** This study highlights the fundamental distinctions between unfairness and long-tailed recognition. Biased classifiers can exacerbate performance disparities in long-tailed recognition. Conversely, unfairness in balanced data stems from problematic representation. It provides valuable insights indicating that effective representation learning algorithms are needed to explicitly improve feature quality for the recognition of “Hard” classes. Building on these findings, we delve into strategies to enhance fairness in Section 5.

### 4.2. Model Prediction Bias

Models trained on balanced datasets such as ImageNet exhibit problematic representations, leading to unfairness. However, the underlying cause of this issue in terms of optimization remains a mystery. In this section, we are prompted to explore this question with the introduced concept of **Model Prediction Bias** in Definition 2.

**Definition 2 (Model Prediction Bias).** For a well-calibrated [26] classification model $\mathcal{M}$, $x_i \in X$ is an input sample. Then the prediction bias for $\mathcal{M}$ is calculated on $X$ with the following formulation:

$$G_n = \sum_{x_i \in X} \sigma(\mathcal{M}(x_i)) \odot (1 - y_i),$$

(3)

$$G_p = \sum_{x_i \in X} \sigma(\mathcal{M}(x_i)) \odot y_i,$$

(4)

where $\sigma$ is the softmax activation. $\mathcal{M}(x_i) \in \mathbb{R}^C$ is the logits output. $y_i$ is the one-hot label. $C$ is the number of classes. Normalization can be applied to $G_n$ or $G_p$.

$\mathcal{M}$ is unbiased when $X$ is balanced and $G_n$ is uniform. $G_p$, with the normalization indicates the class accuracy for the well-calibrated model $\mathcal{M}$.

**High Accuracy Partially from Prediction Bias in LT.** With the concept of **Model Prediction Bias**, we calculate the prediction bias vector $G_n$ of the model trained on ImageNet-LT. As shown in Figure 3(b), the $\ell_2$-norm of classifier weights for high-frequency classes is much larger than that for low-frequency ones. This leads to the model being biased toward high-frequency classes on all the data (“Many”, “Medium”, and “Few”) when giving predictions, as demonstrated in Figs 3(c), 3(d), 3(e) and 4(a). Thus, the model prediction bias results in a much better performance of high-frequency classes than low-frequency classes. In a word, in long-tailed recognition, the high performance of high-frequency classes benefits from the prediction bias of the trained model to some degree.
Low Accuracy from Prediction Bias for Unfairness. On the contrary, as shown in Figure 4(b), the model suffering from unfairness tends to be biased towards “Hard” classes which are with poor accuracy. In particular, the harder the class, the higher the prediction bias the model exhibits. We further calculate the prediction bias $G_n$ on “Easy”, “Medium”, and “Hard” subsets data individually. Illustrated by Figure 4(c), we observe that the model has little prediction bias with “Easy” subset data. On “Medium” subset data, the slight tendency of prediction bias towards “Medium” and “Hard” classes appears. Figure 4(e) indicates an obvious prediction bias towards “Hard” classes of the model when evaluating on the “Hard” subset data, which implies that the unfairness mainly comes from the confusion among “Hard” classes and thus results in their pretty low performance.

High Accuracy or Low Accuracy from Prediction Bias?

The conclusion appears to be contradictory:

- In long-tail learning, higher prediction biases are found in “high-frequency” classes which enjoy high accuracies. In particular, all classes are biased towards high-frequency classes as shown in Figs. 3(c), 3(d), and 3(e).
- On balanced data, higher prediction biases are observed in “Hard” classes which suffer from low accuracies. Specifically, only “Hard” and “Medium” classes are biased towards “Hard” classes in Figs. 4(c), 4(d), and 4(e).

We detail the reasons behind the phenomenon by introducing the following Lemma 1.

Lemma 1 (Gradients Convergence Condition). When the training of model $\mathcal{M}$ converges, the gradients with respect to parameters will be $0$. Then, the following equation is established.

$$G_n + G_p = \mathbf{CF},$$

(5)

where $\mathbf{CF}[i]$ represents the number of samples in class $i$.

Proof. With the rule of backpropagation, the gradients with sample $x_i$ on $\mathcal{M}(x_i)$ is what follows,

$$\frac{\partial \mathcal{L}}{\partial \mathcal{M}(x_i)} = \sigma(\mathcal{M}(x_i)) - y_i.$$  \hspace{1cm} (6)

The gradients integrated on the whole training data should be $0$ for the gradients convergence of training. Then we derive the equation:

$$\sum_{x_i \in \mathcal{X}} \sigma(\mathcal{M}(x_i)) - y_i = 0.$$  \hspace{1cm} (7)

Expand Eq. (7) with Eq. (3), and Eq. (4),

$$\sum_{x_i \in \mathcal{X}} \sigma(\mathcal{M}(x_i)) \odot (1 - y_i) + \sum_{x_i \in \mathcal{X}} \sigma(\mathcal{M}(x_i)) \odot y_i = \sum_{x_i \in \mathcal{X}} y_i.$$  \hspace{1cm} (8)

Eq. (8) is equal to Eq. (5), thus completing the proof.

In Lemma 1, we theoretically prove the relations among the prediction bias $G_n$, the class performance $G_p$, and the class frequency $\mathbf{CF}$. As demonstrated by Eq. (5), for long-tailed recognition, the large number of samples (CF) still allows their good performance ($G_p$) for high-frequency classes, although the model shows higher prediction bias ($G_n$) towards them. Meanwhile, in terms of the unfairness on balanced data, the higher prediction bias ($G_n$) toward “Hard” classes necessarily results in their low accuracy ($G_p$) because the class frequency is uniform (CF).

Intuitively, with Model Prediction Bias in Definition 2, $G_n$ can be regarded as the false positive frequency (FPF) and the class frequency $\mathbf{CF}$ is our expected true positive frequency (TPF). During training, we improve the score of the ground-truth label and decrease the scores for other classes with cross-entropy loss, which means we minimize the FPF and expect the TPF to match the class frequency (CF). To achieve better accuracy for one specific class, the training should be dominated by true positives (TPs) instead of false positives (FPs) learning. Otherwise, FPs will overwhelm the learning of TPs, leading to hard optimization and resulting in low performance.

As shown in Figure 6(a), in long-tailed recognition, the model displays higher $G_n$ towards high-frequency classes. However, TPs still dominate the training due to their high CF. On the contrary, illustrated by Figure 6(b), ImageNet enjoys a uniform CF and the high $G_n$ has overwhelmed their TPs learning for “Hard” classes, resulting in their hard optimization and poor accuracy.

The Unfairness Can Stem from Data Bias. We have analyzed the reason for the unfairness phenomenon in terms of training optimization. From the perspective of input data, bias in the datasets can lead to unfairness.

Class co-occurrence happens in the ImageNet [2]. For example, “keyboard” and “monitor” frequently appear in conjunction. Thus, single-label image annotations in ImageNet can lead to confusion and hard optimization. The ImageNet validation set is relabeled as ImageNet-ReaL where multiple labels can be assigned per image. Evaluating on ImageNet-ReaL, the accuracy of the worst class improves from 16% to 45% but is still far behind the best class 100%.
Figure 7. Improvements of per class with data augmentations on ImageNet and CIFAR-100. (a) Trained with Autoaug, Mixup, and CutMix on ImageNet. (b) Trained with CutMix on CIFAR-100. ViT-B backbone is used on ImageNet while WideResNet-34-10 is used on CIFAR-100.

demonstrating the existence of other potential factors leading to the imbalance.

Moreover, Some class objects occur in much more complex scenes than others. Example classes in “Easy” and “Hard” subsets on ImageNet are shown in Figures 8 and 9 of the supplementary file. We calculate “the $t_2$-norm of the class mean feature variance” with a well-trained ResNet-50 model on ImageNet. As shown in Figure 2, the feature distribution of “Hard” classes has a higher variance than that of “Easy” classes. It demonstrates that “Hard” classes cover much more complex scenarios which can have some overlaps with other classes. In this case, without enough training data for the “Hard” classes, the model can overfit some scenario background and lead to prediction confusion with other classes that share similar scenarios.

4.3. Discussion

As illustrated in Section 3, unfairness is a general problem in image classification models. In this section, we both empirically and theoretically uncover the mystery behind the unfairness from the perspective of gradient optimization during training. Moreover, regarding dataset bias, we observe that features of “Hard” classes usually have much higher variances than “Easy” classes as shown by Figure 2, implying that data for “Hard” classes enjoys larger diversity and possesses lower sample density. From this point of view, the unfairness can be considered as a general imbalance problem, i.e., data diversity imbalance. Besides, although we usually assume classes are independent, the relations among classes (e.g., class co-occurrence) will make the data diversity imbalance heavier. To our knowledge, there are few efforts to explicitly solve the data diversity imbalance. We expect our study to raise more attention from the community to the problem.

5. Improving Fairness

In this section, we investigate how regular training techniques, like data augmentation and representation learning, affect fairness in image classification. Previous work including Mixup [47], CutMix [46], AutoAug [7], RandAug [8], and self-supervised pertaining [20] boosts overall performance. However, fairness is rarely discussed and explored. We bridge the gap in this work.

5.1. Data Augmentation

Mixup and CutMix. Mixup [47] combines image-label pairs $(x_i, y_i)$ and $(x_j, y_j)$ with a uniform sampled $\lambda$,

$$x_{mix} = \lambda x_i + (1 - \lambda) x_j, \quad (9)$$

$$y_{mix} = \lambda y_i + (1 - \lambda) y_j, \quad (10)$$

where $\lambda \in [0, 1]$.

CutMix [46] extends Mixup and randomly exchanges an image region in $x_i$ with another one in image $x_j$. Then the mixed label is determined by mixing $y_i$ and $y_j$ with $\lambda$ which depends on the area of $x_i$ region in the mixed image.

AutoAug and RandAug. AutoAug [7] is one data augmentation strategy that is automatically learned from the data. It is learned from a pre-defined search space including various image process operations, like translation, rotation, or shearing. RandAug [8] selects image transformations in a well-designed search space with a uniform probability.

Training and Evaluation. We use the same training pipeline as described in Section 3. For evaluation, we calculate the improvements for each class between models with and without using data augmentation strategies. We also calculate the improvements of the “Head”, “Medium”, and “Easy” subsets separately. WideResNet-34-10 is used on CIFAR-100 while ViT-B is adopted for ImageNet.

Results and Analysis. Table 3 shows the improvements on the “Hard”, “Medium”, and “Easy” subsets on ImageNet and CIFAR-100 datasets with data augmentations. With the experimental results, the consistent observation is that improvements of “Hard” classes are much larger than that of “Easy” classes. We also plot per-class improvement on CIFAR-100 and ImageNet. As shown in Figure 7, most of the “Hard” classes usually enjoy higher performance improvement than “Easy” classes after adopting data augmentation. It indicates that data augmentation strategies can improve the performance fairness of image classification models. Moreover, data augmentation strategies can be complementary to each other. Combining all of these techniques, the “Hard” classes obtain more performance gains than applying only a single one.

5.2. Representation Learning

Contrastive Learning. We examine the effects of contrastive learning on fairness in two cases: 1) with the pre-trained weights by MoCo [19] as initialization; 2) using the trained model by GPaCo [13]. For fair comparisons, we use the same training setting in GPaCo [13] for baseline models. ResNet-50 backbone is used in experiments.

Masked Modeling. Besides contrastive learning, we empirically study how the masked modeling pre-training meth-


Table 2. Fairness improvements on ImageNet with contrastive learning and masked modeling. “w/o” represents the model is trained from scratch in 300 epochs.

<table>
<thead>
<tr>
<th>Method</th>
<th>Easy</th>
<th>Medium</th>
<th>Hard</th>
<th>All</th>
</tr>
</thead>
<tbody>
<tr>
<td>Contrastive Learning (ResNet-50)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>w/o pre-training</td>
<td>93.4</td>
<td>82.0</td>
<td>62.0</td>
<td>79.1</td>
</tr>
<tr>
<td>w/ pre-training</td>
<td>93.4(+0.0)</td>
<td>82.1(+0.1)</td>
<td>62.2(+0.2)</td>
<td>79.2</td>
</tr>
<tr>
<td>GPaCo [13]</td>
<td>93.8(+0.4)</td>
<td>82.7(+0.7)</td>
<td>62.6(+0.6)</td>
<td>79.7</td>
</tr>
<tr>
<td>Masked Modeling (ViT-B)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>w/o pre-training</td>
<td>92.2</td>
<td>79.6</td>
<td>59.6</td>
<td>77.1</td>
</tr>
<tr>
<td>w/o pre-training*</td>
<td>94.3</td>
<td>84.3</td>
<td>64.8</td>
<td>81.1</td>
</tr>
<tr>
<td>w/ pertaining</td>
<td>95.4(+1.1)</td>
<td>86.6(+2.3)</td>
<td>68.9(+4.1)</td>
<td>83.6</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Contrastive Learning and Masked Modeling (ViT-B)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GPaCo [13]</td>
<td>95.5(+1.2)</td>
<td>87.2(+2.9)</td>
<td>69.1(+4.3)</td>
<td>84.0</td>
</tr>
</tbody>
</table>

Results. We list the results in Table 2. Illustrated by the empirical study, both contrastive learning and masked pretraining can achieve better performance gains on “Hard” classes than “Easy” classes. Further, we also observe that methods specifically for fairness can be boosted by combining contrastive learning and masked modeling pretraining.

5.3. Re-weighting and Other Methods

Re-weighting plays an important role in long-tailed recognition [14, 31]. It assigns larger weights for low-frequency classes and smaller weights for high-frequency ones and expects the low-frequency classes to achieve better performance. We transfer the re-weighting method to address the fairness issue. With a validation set, we calculate the effective frequency (EF) for each class as follows. For class \( c \),

\[
EF(c) = \frac{1}{|X_c|} \sum_{x_i \in X_c} \sigma(M(x_i))[c],
\]

(11)

where \( X_c \) is a subset of the validation set \( X \). Its samples belong to class \( c \). \( \sigma \) is the softmax activation.

With the derived effective frequency (EF), we conduct experiments on ImageNet using [14, 31]. Besides, we investigate the debiased method LfF [32] and worst-group performance optimization method JTT [29] on CIFAR-100.

Results. As shown in Table 4, re-weighting methods can improve the accuracy of “Hard” classes at the cost of some performance degradation on “Easy” and “Medium” classes. However, it fails to promote overall accuracy and even achieves inferior performance than the baseline, which is different from data augmentations and representation learning techniques. Moreover, we observe that methods for worst-group performance or classifier debiasing cannot handle complex data distributions, like CIFAR-100.

6. Conclusion and Future Work

In this paper, we investigate the image recognition fairness issue, i.e., extreme class performance disparity on balanced data. This unfairness phenomenon widely exists in image classification models. We identify several properties of the fairness issue by comparing it with long-tailed recognition: (1) The unfairness comes from problematic representation instead of classifier bias; (2) The harder the class, the more other classes can be confused with it, resulting in their low performance. Finally, we verify that data augmentation and representation learning tricks can improve fairness with better overall performance.

With our empirical study, representation learning is promising for improving fairness. However, it hasn’t been explored to explicitly promote fairness learning. Designing proper data augmentation strategies or self-supervised pretraining techniques specifically for fairness is an interesting direction. We hope the study will attract more researchers to solve this problem.

Acknowledgements. This research was supported by the funding MOE AcRF Tier 2 (MOE2019-T2-2-062).