Contrastive Learning based Hybrid Networks for Long-Tailed Image Classification

Peng Wang[1] Kai Han[2] Xiu-Shen Wei[3] Lei Zhang[4] Lei Wang[1]
[1]University of Wollongong [2]University of Bristol
[3]Nanjing University of Science and Technology [4]Northwestern Polytechnical University

Abstract

Learning discriminative image representations plays a vital role in long-tailed image classification because it can ease the classifier learning in imbalanced cases. Given the promising performance contrastive learning has shown recently in representation learning, in this work, we explore effective supervised contrastive learning strategies and tailor them to learn better image representations from imbalanced data in order to boost the classification accuracy thereon. Specifically, we propose a novel hybrid network structure being composed of a supervised contrastive loss to learn image representations and a cross-entropy loss to learn classifiers, where the learning is progressively transited from feature learning to the classifier learning to embody the idea that better features make better classifiers. We explore two variants of contrastive loss for feature learning, which vary in the forms but share a common idea of pulling the samples from the same class together in the normalized embedding space and pushing the samples from different classes apart. One of them is the recently proposed supervised contrastive (SC) loss, which is designed on top of the state-of-the-art unsupervised contrastive loss by incorporating positive samples from the same class. The other is a prototypical supervised contrastive (PSC) learning strategy which addresses the intensive memory consumption in standard SC loss and thus shows more promise under limited memory budget. Extensive experiments on three long-tailed classification datasets demonstrate the advantage of the proposed contrastive learning based hybrid networks in long-tailed classification.

1. Introduction

In the real world, the image classes are normally presented in a long-tailed distribution [25]. While some common classes (head classes) can have sufficient image samples, some uncommon or rare categories (tail classes) can be underrepresented by limited samples. The data imbalance poses great challenge to learning unbiased classifiers.

Most existing work addresses the data imbalance issue by mitigating the data shortage in tail classes in order to prevent the model from being dominated by the head classes. Typical methods include data re-sampling [1, 8, 35, 4, 28], loss re-weighting [26, 7, 30, 33], margin modification [3], and data augmentation [19, 6, 29]. Recently a new line of work is proposed which approaches long-tailed image classification by decoupling the representation learning and classifier learning into two stages [15, 37, 36]. The shared motivation of such work [15, 37, 36] is that image feature learning and classifier learning may favor different data sampling strategies and thus the focus thereon is to identify suitable sampling strategies for these two tasks. Specifically, they find under cross-entropy loss, random data sampling can benefit feature learning more while class-balanced sampling

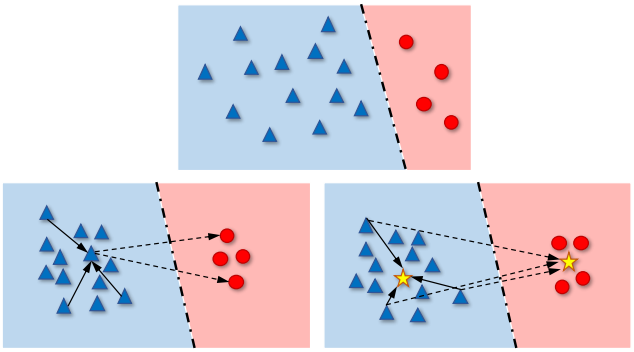

Figure 1. Illustration of cross-entropy (upper), standard supervised contrastive (SC) (bottom left), and prototypical supervised contrastive (PSC) (bottom right) loss based feature learning for long-tailed image classification. Cross-entropy loss learns skewed features, which can result in biased classifiers. Supervised contrastive learning (bottom two) learns more intra-class compact and inter-class separable features, which ease classifier learning. In standard SC learning, an anchor sample together with positive samples from the same class are pulled together and the anchor is pushed away from negatives from other classes. In PSC learning, each sample is pulled towards the prototype (marked by star) of its class and pushed away from prototypes of other classes.

is a better option for classifier learning. Despite promising accuracy achieved, these methods leave the question of *whether typical cross-entropy is an ideal loss for learning features from imbalanced data* untouched. Intuitively, as shown in Fig. 1, the feature distribution learned from typical cross-entropy can be highly skewed, which can lead to biased classifiers [24, 12] that harm long-tailed classification.

In this work, we explore effective contrastive learning strategies and tailor them to learn better image representations from imbalanced data in order to boost long-tailed image classification. Specifically, we propose a novel hybrid network structure composed of a contrastive loss for learning image representations and a cross-entropy loss to learn classifiers. To embody the idea that better features make better classifiers, we follow a curriculum to progressively transit the learning from feature learning to classifier learning. We realize two variants of supervised contrastive learning strategies, as shown in Fig. 1, which vary in the forms but share a common idea of pulling the samples from the same class together in the normalized embedding space and pushing the samples from different classes apart. By doing this, less skewed features and consequently less biased classifiers are expected to be obtained.

The first contrastive learning we explore to learn features in imbalanced scenario is the recently proposed supervised contrastive (SC) learning [18], which is extended from the state-of-the-art unsupervised contrastive learning [5] by incorporating different within-class samples as positives for each anchor. Following unsupervised contrastive learning [5, 9] that have two independent stages for feature learning and classifier learning, the original SC learning [18] learns features using SC loss first and then freezes the features to learn classifiers. We argue in this paper such two-stage learning may not be an optimal choice in fully supervised scenario, which can harm the compatibility of the features and classifiers. We propose a hybrid framework to jointly learn features and classifiers, and empirically demonstrate the advantage of our joint learning mode.

One issue of incorporating within-class positive samples in SC learning is that it leads to extra memory consumption. In SC learning [18], the distances to positives from the same class are contrasted with the distances to negatives from other classes, which results in memory consumption linear to the product of the positive size and negative size. Due to this, when under limited memory budget, the negative size needs to be shrunk. This can compromise the quality of the features learned from contrastive loss [5], especially when dealing with dataset that has a large number of classes, *e.g.*, iNaturalist [11].

To address the aforementioned memory bottleneck from SC loss, we further propose a prototypical supervised contrastive (PSC) learning strategy, which shares the similar goal with standard SC learning but avoids explicitly sampling positives and negatives. In PSC learning, we learn a prototype for each class and force each sample to be pulled towards the prototype of its class and pushed away from prototypes of all the other classes. In this sense, the PSC strategy enables more flexible and efficient data sampling akin to softmax-based cross-entropy. It observes advantages when dealing with large-scale dataset under limited memory budget. In addition, the PSC loss has some other appealing properties that can benefit imbalanced classification, such as less sensitive to data sampling and the potential to capture finer within-class data distribution by using multiple prototypes per class.

Experiments on three long-tailed image classification datasets demonstrate the proposed contrastive learning based hybrid networks can obviously outperform the cross-entropy based counterparts and establish new state-of-the-art long-tailed image classification performance. The contributions of this work can be summarized as follows:

- We propose a novel hybrid network structure for long-tailed image classification. The network is designed to be composed of a contrastive loss for feature learning and a cross-entropy loss for classifier learning. These two learning tasks are performed following a curriculum to embody the idea that better features can ease classifier learning.
- We explore effective supervised contrastive learning strategies to learn better features to boost long-tailed classification performance. A prototypical supervised contrastive (PSC) learning is proposed to resolve the memory bottleneck resulted from standard supervised contrastive (SC) learning.
- We unveil supervised contrastive learning can be a better substitute for typical cross-entropy loss for feature learning in long-tailed classification. Benefited from the better features learned, our hybrid network substantially outperforms the cross-entropy based counterparts.

Our code is publicly available at https://k-han.github.io/HybridLT.

2. Related Work

Our work is closely related to both long-tailed classification and contrastive learning.

2.1. Long-tailed image classification

Long-tailed classification is a long-standing research problem in machine learning, where the key is to overcome the data imbalance issue [21, 16]. Given the great success deep neural networks have achieved in balanced classification tasks, increasing attention is being shifted to proposing neural networks based solutions for long-tailed classification. In this work, we mainly focus on the neural networks based

approaches, which can be roughly divided into the following categories.

Data re-sampling Data re-sampling is a commonly used strategy to artificially balance the imbalanced data. Two types of re-sampling techniques are under-sampling [1, 28, 8] and over-sampling [1, 32, 31]. Under-sampling discards part of the data in head classes and over-sampling repetitively samples data from the tail classes. It is revealed that over-sampling can lead to overfitting to the tail classes [4, 28]. Under-sampling can potentially lose information about the head classes but it may yield good results if each sample of a head class is close to other samples of the same class [28].

Data augmentation As analyzed above, although over-sampling enhances the chance to see more data from the tail classes, it does not generate new information and thus leads to overfitting. One remedy is to use strong data augmentation to enrich the tail classes. Existing work approaches this goal from different angles. The work in [29] uses generative model to generate new samples for tail classes as convex combination of existing instances. Another line of studies attempt to transfer the information from head classes to tail classes. In [19], the authors generate data for tail classes by adding learnable noise to head samples. In another work [6], the authors decompose the feature maps of images as class-generic features and class-specific features and compose new tailed data by combining class-generic features from the head image and class-specific features from a tail image. In [23], the intra-class angular variance is transferred from head classes to enlarge the diversity of tail classes.

Loss re-weighting Apart from the aforementioned data-based re-balance strategies, another line of studies propose to mitigate the negative effects of data imbalance by modifying the loss functions. Loss re-weighting is one of the simple but effective ways to tailor the loss function for imbalanced classification, where the basic idea is to upweight the tailed samples and downweight the head samples in the loss function [17]. The existing solutions differ mainly in how to define the weights for different classes. In class-sensitive cross-entropy loss [14], the weight assigned to each class is inversely proportional to the number of samples. In class-balanced loss [7], the authors decide the re-weighting coefficients based on the real volumes of different classes, named effective numbers. In the work [30], the weights to the training examples are optimized to minimize the loss of a held-out evaluation set.

Margin modification It is revealed that the effect of loss re-weighting can diminish when the datasets are separable [2]. An intuitive alternative is to shift the separator closer to a dominant class [27]. In the work [3], the authors propose to integrate per-class margin into the cross-entropy loss. The margin is inversely proportional to the prior probability of a class and thus can enforce larger margins between a tail class and other classes. The work [33] realizes the margin under an alternative motivation, which is to suppress the negative gradients resulted from head samples for each tailed sample.

Decoupled learning Decoupled learning is a recent line of methods towards imbalanced classification. To identify the specific contributions of different factors to the long-tailed recognition capability, the work [15] decouples long-tailed classification into two separate stages: representation learning and classifier learning. They use cross-entropy as loss function for both of these two stages and conclude that feature learning favors random data sampling and class-balanced sampling is a better option for classifier learning. Parallel to this, the work [37] obtains similar conclusions empirically. In addition, a bilateral-branch network is proposed in [37], where one branch uses random sampling to learn head data and the other branch uses revered sampling to emphasize tailed data. One common focus of these two works lies in choosing proper data sampling strategies for different learning tasks underpinning long-tailed classification. But both studies are limited to cross-entropy loss.

2.2. Contrastive learning

Recently, contrastive learning has shown great promise in unsupervised representation learning [5, 9]. The basic idea is to learn a hidden space in which the agreement between differently augmented views of the same image is maximized by contrasting to the agreement between different images. Some key components enable the success of contrastive loss in learning useful representations include proper data augmentations, a learnable nonlinear transformation between the representation and contrastive loss, and large batch size for negative data [5]. Supervised contrastive (SC) learning [18] is an extension to contrastive learning by incorporating the label information to compose positive and negative images. Following unsupervised feature learning, SC learning also adopts a two-stage learning fashion, where the first stage learns features by using contrastive loss and the second stage learns classifiers using cross-entropy loss.

3. Main Approach

In this section, we firstly present the framework for the contrastive learning based hybrid networks proposed for long-tailed classification. Then, we elaborate on the two supervised contrastive learning schemes used as part of the hybrid networks for image representation learning.

3.1. A Hybrid Framework for Long-tailed Classification

Fig. 2 shows the overview of the proposed hybrid framework for long-tailed image classification. The network consists of two branches: one contrastive learning branch for image representation learning and one cross-entropy driven branch for classifier learning. The feature learning branch

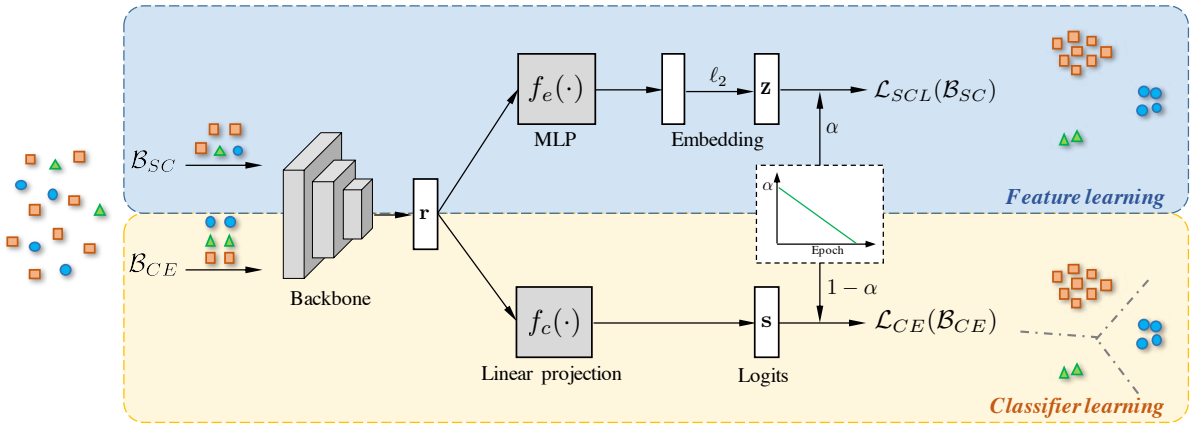

Figure 2. Overview of the proposed contrastive learning based hybrid network structure. The network is composed of a supervised contrastive learning (SCL) based feature learning branch and a cross-entropy (CE) loss based classifier learning branch. A backbone is shared between these two branches to extract image representations, after which a non-linear MLP $f_e(\cdot)$ combined with ℓ_2-normalization is adopted to translate the image representation for contrastive loss, and a single linear layer $f_c(\cdot)$ is applied on top of the image representation to predict classification logits. A curriculum is designed to control the weightings of these two branches, *i.e.*, α and $1-\alpha$, during network training.

aims to learn a feature space which has the property of intra-class compactness and inter-class separability. The classifier learning branch is expected to learn less biased classifier based on the discriminative features obtained from the sibling branch. To realize the idea that better features ease classifier learning and consequently result in more generalizable classifiers, we follow a curriculum [37] to adjust the weightings of these two branches during the training phase. Concretely, the feature learning plays a leading role at the beginning of the training, and then classifier learning gradually dominates the training.

A backbone network, *e.g.*, ResNet [10], is shared between these two learning branches to learn the image representation $\mathbf{r} \in \mathcal{R}^{D_E}$ for each image $\mathbf{x}$. A projection head $f_e(\cdot)$ maps the image representation $\mathbf{r}$ into a vector representation $\mathbf{z} \in \mathcal{R}^{D_S}$ which is more suitable for contrastive loss. We implement this projection head $f_e(\cdot)$ as a nonlinear multiple-layer perceptron (MLP) with one hidden layer. Such projection module is proven important in improving the representation quality of the layer before it [5]. Then, the ℓ_2 normalization is applied to $\mathbf{z}$ in order that inner product can be used as distance measurements. To avoid abuse of notations, unless otherwise stated we use $\mathbf{z}$ as the normalized representation of $\mathbf{x}$ for contrastive loss computation. After that, a supervised contrastive loss $\mathcal{L}_{SCL}$ is applied on top of the normalized representations for feature learning. The classifier learning branch is simpler which applies a single linear layer $f_c(\cdot)$ to the image representation $\mathbf{r}$ to predict the class-wise logits $\mathbf{s} \in \mathcal{R}^{D_C}$, which are used to compute the cross-entropy loss $\mathcal{L}_{CE}$. Due to different natures of the two loss functions, the feature learning and classifier learning branches have different data sampling strategies. The feature learning branch takes as input anchor point $\mathbf{x}_i$ together with positive samples $\{\mathbf{x}_i^+\} = \{\mathbf{x}_j | y_i = y_j, i \neq j\}$ from the same class and negative samples $\{\mathbf{x}_i^-\} = \{\mathbf{x}_j | y_j \neq y_i\}$ from other classes. The input batch of the feature learning branch is denoted as $\mathcal{B}_{SC} = \{\mathbf{x}_i, \{\mathbf{x}_i^+\}, \{\mathbf{x}_i^-\}\}$. The classifier learning branch directly takes image and label pairs as input $\mathcal{B}_{CE} = \{\{\mathbf{x_i}, y_i\}\}$. The final loss function for the hybrid network is:

$$\mathcal{L}_{hybrid} = \alpha \cdot \mathcal{L}_{SCL}(\mathcal{B}_{SC}) + (1-\alpha) \cdot \mathcal{L}_{CE}(\mathcal{B}_{CE}), \quad (1)$$

where α is a weighting coefficient inversely proportional to the epoch number, as shown in Fig. 2.

3.2. Supervised contrastive loss and its memory issue

Supervised contrastive (SC) loss [18] is an extension to unsupervised contrastive (UC) loss [5]. The key difference between SC loss and UC loss lies in the composition of the positive and negative samples of an anchor image. In UC loss, the positive image is an alternatively augmented view of the anchor image. In SC loss, apart from the alternatively augmented counterpart, the positives also include some other images from the same class. In this paper, we unify all the positive images of an anchor $\mathbf{x}_i$ as $\{\mathbf{x}_i^+\} = \{\mathbf{x}_j | y_j = y_i, i \neq j\}$ (we assume different views of the same image have different indexes). The definitions for positives and negatives of $\mathbf{x}_i$ also apply to $\mathbf{z}_i$ as $\{\mathbf{z}_i^+\}$ and $\{\mathbf{z}_i^-\}$. Assuming the minibatch size is N, the SC loss function is written as:

$$\mathcal{L}_{SCL} = \sum_{i=1}^{N} \mathcal{L}_{SCL}(\mathbf{z}_i), \quad (2)$$

$$\mathcal{L}_{SCL}(\mathbf{z}_i) = \frac{-1}{|\{\mathbf{z}_i^+\}|} \sum_{\mathbf{z}_j \in \{\mathbf{z}_i^+\}} \log \frac{\exp(\mathbf{z}_i \cdot \mathbf{z}_j/\tau)}{\sum_{\mathbf{z}_k, k \neq i} \exp(\mathbf{z}_i \cdot \mathbf{z}_k/\tau)}, \tag{3}$$

where $|\{\mathbf{z}_i^+\}|$ denotes the number of positive samples of anchor $\mathbf{z}_i$, and $\tau > 0$ is a scalar temperature parameter.

Comparing to the UC loss [5], the SC loss can flexibly incorporate arbitrary number of positives. It optimizes the agreements between such positives by contrasting against negative samples. However, the consequence of using within-class positives in SC loss is that it results in memory consumption linear to the product of positive size and negative size. For example, when one different within-class image along with an alternative view are used as positives in SC loss, the memory consumption will be doubled comparing to the UC loss with the same size of negatives. This limits the application of SC loss when under limited GPU memory budget. One solution is to shrink the size of negatives. But this can be problematic when dealing with dataset that has large number of classes because small negative size samples small fraction of negative classes, which can compromise the quality of the learned representation.

3.3. Prototypical supervised contrastive loss

To simultaneously resolve the memory bottleneck issue and mostly retain the feature learning property of SC loss, we propose a prototypical supervised contrastive (PSC) loss. In PSC loss, we aim to attain similar goal of SC loss by learning a prototype for each class and force differently augmented views of each sample to be close to the prototype of their class and far away from the prototypes of the remaining classes. The benefits of using prototypes are two-fold. Firstly, it enables more flexible data sampling by avoiding explicitly sampling positives and negatives. Thus, we can flexibly adopt the data sampling strategies readily available in long-tailed classification, such as random sampling and class-balanced sampling. Secondly, data sampling efficiency is enhanced in PSC loss. In PSC loss, we contrast each sample against the prototypes of all other classes. If a dataset has $\mathcal{C}$ classes, this is essentially equivalent to a negative size of $\mathcal{C} - 1$. This is practically important when dealing with dataset with large number of classes, *e.g.*, iNaturalist [11]. The PSC loss function is:

$$\mathcal{L}_{PSC}(\mathbf{z}_i) = -\log \frac{\exp(\mathbf{z}_i \cdot \mathbf{p}_{y_i}/\tau)}{\sum_{j=1, j \neq y_i}^{\mathcal{C}} \exp(\mathbf{z}_i \cdot \mathbf{p}_j/\tau)}, \tag{4}$$

where $\mathbf{p}_{y_i}$ is the prototype representation for class y_i, which is normalized to the unit hypersphere in $\mathcal{R}^{D_S}$ and $\mathbf{z}_i$ is the normalized representation of $\mathbf{x}_i$.

Extension to multiple prototypes per class In the above section, we learn one prototype per class. But PSC loss can be simply extended to multiple prototypes for each class. The rationale behind is that the samples within a class may follow a multimodal distribution, which can be better modeled by using multiple prototypes. The multiple prototype supervised contrastive (MPSC) loss function can be designed as:

$$\mathcal{L}_{MPSC}(\mathbf{z}_i) = \frac{-1}{M} \sum_{k=1}^{M} \log \frac{w_{i,k} \exp(\mathbf{z}_i \cdot \mathbf{p}_{y_i}^k/\tau)}{\sum_{j=1, j \neq y_i}^{C} \sum_{m=1}^{M} \exp(\mathbf{z}_i \cdot \mathbf{p}_j^m/\tau)}, \tag{5}$$

where M is the number of prototypes per class, $\mathbf{p}_j^i$ denotes the representation for the i-th prototype of class j, and $w_{i,k} (w_{i,k} \geq 0, \sum_{k=1}^{M} w_{i,k} = 1)$ denotes the affinity value between $\mathbf{z}_i$ and the k-th prototype of its class, which is used to control the affinity of each sample in finer level. We leave detailed evaluation of MPSC loss as future work.

4. Experiments

In this section, we firstly introduce the three long-tailed image classification datasets used for our experiments. Then we present some key implementation details of our methods. After that, we compare our proposed hybrid networks with state-of-the-art long-tailed image classification methods. Finally, some ablation studies are given to highlight some important properties of our hybrid networks.

4.1. Datasets

We conduct experiments on three long-tailed image classification datasets. Two of them, long-tailed CIFAR-10 and long-tailed CIFAR-100, are derived artificially from balanced CIFAR [20] datasets by re-sampling. The third dataset, iNaturalist 2018 [11], is a large-scale image dataset, in which the image categories exhibit long-tailed distribution.

Long-tailed CIFAR-10 and CIFAR-100 The original CIFAR-10 and CIFAR-100 datasets are balanced datasets. They consist of 50,000 training images and 10,000 validations images of size 32×32 in 10 and 100 classes respectively. Following [7, 3], the long-tailed versions are created by reducing the number of training examples per class but with the validation set unchanged. An imbalance ratio β is used to denote the ratio between sample sizes of the most frequent and least frequent class, *i.e.*, $\beta = N_{max}/N_{min}$. The sample size follows an exponential decay across different classes. Similar to most existing work [7, 3, 37], we use imbalance ratios of 10, 50 and 100 in our experiments.

iNaturalist 2018 The iNaturalist 2018 [11] is a large-scale real-world species classification dataset. It consists of 8,142 species, with 437,513 training and 24,424 validation images. The dataset observes severe imbalance in the sample sizes across different specie categories. We use the official training and validation splits for our experiments.

4.2. Implementation details

In this section, we present some key implementation details for experiments on long-tailed CIFAR and iNaturalist respectively.

Implementation details for long-tailed CIFAR For both long-tailed CIFAR-10 and CIFAR-100, we use ResNet-32 [10] as backbone network to extract image representation. Our hybrid network has two branches, which have independent input data as shown in Fig. 2. The basic set of data augmentation shared by both branches include random cropping with size 32×32, horizontal flip and random grayscale with probability of 0.2. Following SC loss, we also derive different views of an image by using different data augmentations in PSC loss. In our experiments, we simply use with and without color jitter as two different augmentation views. We use batch size of 512 for both SC and PSC based hybrid networks. The classifier learning branch uses class-wise balanced data sampling. We use SGD with a momentum of 0.9 and weight decay of 1×10^{-4} as optimizer to train the hybrid networks. The networks are trained for 200 epochs with the learning rate being decayed by a factor of 10 at the 120^{th} epoch and 160^{th} epoch. The initial learning rate is 0.5. For the curriculum coefficient α, we use a parabolic decay w.r.t the epoch number [37], *i.e.*, $\alpha = 1 - (T/T_{max})^2$, where T denotes the current epoch number and T_{max} indicates the maximum epoch number. For SC based hybrid network, the temperature τ in Eq. (3) is fixed to be 0.1. For PSC base hybrid network, τ is set to be 1 for CIFAR-10 and 0.1 for CIFAR-100.

Implementation details for iNaturalist 2018 For iNaturalist 2018, following most of the existing work, we use ResNet-50 [10] as backbone network. The data augmentation is similar to that used in long-tailed CIFAR datasets except that random cropping with size 224×224 is used. To fit two NVIDIA 2080Ti GPUs, we use a batch size of 100 for both SC and PSC based hybrid networks. The networks are trained for 100 epochs using SGD with momentum 0.9 and weight decay 1×10^{-4}. The initial learning rate is 0.05, which is decayed by a factor of 10 at epoch 60 and epoch 80. Motivated by the fact that iNaturalist has a large number of classes which can make classifier learning more difficult, we assign higher weighting to the classifier learning branch by using a linearly decayed weighting factor α, *i.e.*, $\alpha = 1 - T/T_{max}$. The temperature τ is set to be 0.1 for both SC and PSC loss functions. For SC loss function, the number of positive samples for each anchor is fixed to 2.

4.3. Comparison to state-of-the-art methods

In this section, we compare the proposed hybrid networks, including both SC and PSC loss based networks, to existing long-tailed classification methods on long-tailed CIFAR and iNaturalist datasets, respectively.

Experimental results on long-tailed CIFAR The comparison between the proposed hybrid networks and existing methods on long-tailed CIFAR datasets is presented in Table 1. The compared methods cover various categories of ideas for imbalanced classification, including loss re-weighting [7], margin modification [3], data augmentation [19], decoupling [37] and some other newly proposed ideas [34, 13]. As can be seen from the table, our hybrid networks outperform the compared methods on almost all the settings.

Among these methods, CE denotes the simplest baseline which directly uses cross-entropy to train the network on the long-tailed datasets. As expected, this baseline method achieves the worst performance, which reveals the limitation of cross-entropy in dealing with imbalanced data. Although the performance can be improved by using advanced loss functions tailored for long-tailed data [3, 7, 22], these methods ignore the different properties of feature learning and classifier learning. BBN [37] takes a step further by decoupling the head data and tailed data modeling. But several factors of BBN compromise the full potential of decoupling learning: 1) It unifies the representation for two data streams with different properties in the penultimate layer; 2) Cross-entropy loss is not an ideal loss for imbalanced data in both streams; 3) The final predication in testing phase is calculated as the sum of two prediction functions from two branches with equal weights, which is inconsistent with the training phase. Our methods address such limitations in that: 1) The projection module in our feature learning branch adapts the image representation to a space more suitable for contrastive loss; 2) We use different loss functions to learn the features and classifiers and conclude supervised contrastive loss can be a better substitute for cross-entropy in learning features from imbalanced data; 3) We use a single classifier learning function to predict the class labels for each sample, which avoids the gap between training and testing. Within our methods, SC based hybrid network, a.k.a Hybrid-SC, performs better than the PSC counterpart, a.k.a Hybrid-PSC, but the latter still performs on par with or better than the compared methods.

Experimental results on iNaturalist 2018 The experimental comparison to some existing work on iNaturalist 2018 is provided in Table 2. Again, we compare our hybrid networks to various lines of methods. Among these compared methods, Decoupling [15] and BBN [37] are most closely related to our proposal, which are both based on the idea of decoupled learning. The advantage of our methods over BBN has been analyzed above. On iNaturalist, Hybrid-PSC outperforms BBN by 1.8%. Classifier re-training (cRT) is a well-performed method we choose to compare in [15]. It is a two-stage method, where the first stage learns image features and the second stage freezes the features to learn the classifiers. They use cross-entropy as loss function for

Table 1. Top-1 accuracy (%) on long-tailed CIFAR datasets based on ResNet-32. (Best and second best results are marked in bold.)

Dataset	Long-tailed CIFAR-10			Long-tailed CIFAR-100		
Imbalance ratio	100	50	10	100	50	10
CE	70.36	74.81	86.39	38.32	43.85	55.71
Focal loss [22]	70.38	76.72	86.66	38.41	44.32	55.78
CB-Focal [7]	74.57	79.27	87.10	39.60	45.17	57.99
CE-DRW [3]	76.34	79.97	87.56	41.51	45.29	58.12
CE-DRS [3]	75.61	79.81	87.38	41.61	45.48	58.11
LDAM-DRW [3]	77.03	81.03	88.16	42.04	46.62	58.71
CB-DA [13]	80.00	82.23	87.40	44.08	49.16	58.00
M2m [19]	79.10	–	87.50	43.50	–	57.60
Casual model [34]	**80.6**	83.60	88.50	44.10	**50.30**	59.60
BBN [37]	79.82	81.18	88.32	42.56	47.02	59.12
Hybrid-SC (ours)	**81.40**	**85.36**	**91.12**	**46.72**	**51.87**	**63.05**
Hybrid-PSC (ours)	78.82	**83.86**	**90.06**	**44.97**	48.93	**62.37**

both stages but using different data sampling strategies. We argue this method suffers from two limitations: 1) The two-stage learning strategy harms the compatibility between the learned features and classifiers; 2) Cross-entropy loss is not an ideal choice for learning image features from imbalanced data. Our hybrid network addresses the first limitation by using a curriculum based learning strategy to smoothly transit from feature learning to classifier learning. The second limitation is also observed in BBN, which can be addressed by our hybrid network as analyzed above. Our Hybrid-PSC network outperforms Decoupling [15] by nearly 3%.

Another interesting observation is that Hybrid-PSC performs better than Hybrid-SC. This result is consistent with our expectation. Note that for the two hybrid network versions, we use the same batch size of 100 for contrastive loss. This batch size is too small comparing to the number of classes in the iNaturalist dataset, which fails to provide the SC loss with sufficient negative samples to learn high-quality features [5]. PSC loss avoids this issue because, as analyzed in Sec. 3.3, each sample will contrast with all the negative prototypes regardless of the batch size. Due to this reason, Hybrid-PSC obtains superior classification performance. Generally, we can state that the PSC based hybrid network can observe advantage over the SC loss when dealing with imbalanced dataset with large number of classes under limited GPU memory budget.

4.4. Ablation studies and discussions

In this section, we conduct some ablation studies to characterize our hybrid networks. Concretely, we study whether the proposed PSC loss is less sensitive to data sampling, the advantage of using PSC loss in feature learning comparing to cross-entropy loss, and the advantage of our curriculum based joint training comparing to the two-stage learning strategy.

Table 2. Top-1 accuracy (%) on iNaturalist 2018 dataset based on ResNet-50. For Decoupling [15], the well-performed Classifier Re-training (cRT) is reported as it is closely related to our method. By default, the methods are trained for up to 100 epochs. The number in brackets indicates the accuracy obtained by training for 200 epochs. (Best and second best results are marked in bold.)

Dataset	iNaturalist2018
CE	57.16
CB-Focal [7]	61.12
CE-DRW [3]	63.73
CE-DRS [3]	63.56
LDAM-DRW [3]	**68.00**
CB-DA [13]	67.55
FeatAug [6]	65.91
Decoupling [15]	65.20 (67.6)
BBN [37]	66.29 (**69.62**)
Hybrid-SC (ours)	66.74
Hybrid-PSC (ours)	**68.10** (**70.35**)

Sensitivity of PSC loss to data sampling In the decoupled learning work [15, 37], the authors find cross-entropy loss is sensitive to data sampling when it is used to learn features. Concretely, they find random sampling obviously outperforms class-wise balanced sampling for feature learning. For example, in [15], the class-balanced sampling can lead to around 5% accuracy drop comparing to random sampling under cross-entropy loss. As PSC loss in our work has the same data sampling manner as cross-entropy loss, we verify the sensitivity of our PSC loss to data sampling in Table 3. From the table we can see, our Hybrid-PSC network achieves comparable performance by using random sampling and class-balanced sampling, which indicates our PSC can alleviate the overfitting issue resulted from over-sampling (class-balanced sampling belongs to over-sampling). We conjecture that two possible factors

Table 3. Evaluation of the sensitivity of PSC loss to data sampling. Hybrid-PSC with random PSC and Hybrid-PSC with CB-PSC denote in the PSC based hybrid network, we use random data sampling and class-balanced data sampling for the feature learning branch respectively. Classification accuracy (%) on long-tailed CIFAR-100 is reported.

Dataset	Long-tailed CIFAR-10			Long-tailed CIFAR-100			iNaturalist 2018
Imbalance ratio	100	50	10	100	50	10	-
Hybrid-PSC with random PSC	78.82	83.86	90.06	44.91	48.93	62.37	68.10
Hybrid-PSC with CB-PSC	78.84	82.85	89.85	44.21	49.66	61.93	67.71

Table 4. Evaluation of the advantage of supervised contrastive losses over cross-entropy loss for feature learning in long-tailed classification. CE-CE denotes both feature learning and classifier learning adopt cross-entropy loss, *i.e.*, our supervised contrastive loss is replaced by cross-entropy loss. Classification accuracy (%) on long-tailed CIFAR-100 is reported.

Dataset	Long-tailed CIFAR-100		
Imbalance ratio	100	50	10
CE-CE	41.40	46.68	59.14
Hybrid-SPC	44.97	48.93	62.37
Hybrid-SC	46.72	51.87	63.05

Table 5. Evaluation of the advantage of the curriculum based joint training over two-stage training. Two-stage SC denotes we train the features and classifiers in separate stages. Hybrid-SC w/o curriculum means we use equal and fixed weighting for the feature and classifier learning during the training process. Classification accuracy (%) on long-tailed CIFAR-100 is reported.

Dataset	Long-tailed CIFAR-100		
Imbalance ratio	100	50	10
Two-stage SC	42.73	46.76	60.62
Hybrid-SC w/o curriculum ($\alpha = 0.5$)	42.58	47.45	60.48
Hybrid-SC	46.72	51.87	63.05
Hybrid-SPC	44.91	48.93	62.37

contribute to the insensitivity of the PSC loss on data sampling. Firstly, in PSC loss, the image features and prototypes are both ℓ_2-normalized, which breaks the strong correlations between class frequency and feature norms. Secondly, assuming the affinity score between a sample and its prototype is $s_i^{y_i} = \mathbf{z}_i \cdot \mathbf{p}_{y_i}/\tau$. For a sample $\mathbf{x}_i$ with label $y_i \in \{1, 2, \ldots, \mathcal{C}\}$, the gradient of the PSC loss $\mathcal{L}_{PSC}(\mathbf{z}_i)$ w.r.t $s_i^{y_i}$ is constant, and the gradient w.r.t the affinity to a prototype from a negative class $c \in \{1, 2, \ldots, \mathcal{C}\} \backslash y_i$, is $\exp(s_i^c) / \sum_{y \in \{1,2,\cdots,\mathcal{C}\}, y \neq y_i} \exp(s_i^y)$. The denominator excludes the dominating term of $s_i^{y_i}$ and thus results in a prominent gradient. The constant gradient for positive class and prominent gradients for negative classes can help to alleviate the overfitting in over-sampling and enhance the inter-class separability of the features.

Is PSC loss a better substitute for cross-entropy loss for feature learning? In this work, we claim the supervised contrastive losses are expected to learn better features from imbalanced features and consequently lead to better long-tailed classification performance. To verify this, we replace the contrastive loss in our hybrid networks with cross-entropy loss. The results are shown in Table 4. As can be seen, when using cross-entropy to learn the image features, the performance drops significantly.

Two-stage learning v.s. curriculum based joint learning In this work, we use a curriculum to smoothly transit the training from feature learning to classifier learning. To justify the advantage of this learning strategy, we firstly choose the original two-stage SC work [18] as our baseline, which trains the features using SC loss in the first stage and then fixes the features to train classifiers in the second stage. From Table 5 we can see, this two-stage training scheme results in obviously inferior performance to our curriculum based training, because it harms the compatibility between the features and classifiers. To further highlight the importance of the curriculum, we set the weighting coefficient α in Eq. (1) to be 0.5. Still, unsatisfactory results are obtained. When the curriculum is used, we allow the supervised contrastive losses to dominate the training first in order to fully exploit their capacity to learn discriminative features, which can benefit the classifier learning in later phase.

5. Conclusion

In this work, we approached long-tailed image classification by proposing a novel hybrid network, which consists of a supervised contrastive loss to learn image features and a cross-entropy loss to learn classifiers. To embody the idea that better features make better classifiers, a curriculum is followed to smoothly transit the training from feature learning to classifier learning. A new prototypical supervised contrastive loss was proposed to learn features from imbalanced data, which observes advantage under limited GPU memory budget. Experiments on three long-tailed classification datasets showed that our proposal not only significantly outperforms existing methods but also has some other appealing properties that can benefit imbalanced classification. To our knowledge, this is the first work that explores how to maximize the value of supervised contrastive learning in long-tailed image classification. We will continue this direction as our future work, with the deeper exploration of MPSC as the first step.

References

[1] Mateusz Buda, Atsuto Maki, and Maciej Mazurowski. A systematic study of the class imbalance problem in convolutional neural networks. *Neural Networks*, 2017. 1, 3

[2] Jonathon Byrd and Zachary Chase Lipton. What is the effect of importance weighting in deep learning? In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *ICML*, 2019. 3

[3] Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Arechiga, and Tengyu Ma. Learning imbalanced datasets with label-distribution-aware margin loss. In *NeurIPS*, 2019. 1, 3, 5, 6, 7

[4] Nitesh Chawla, Kevin Bowyer, Lawrence Hall, and W. Kegelmeyer. SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 2002. 1, 3

[5] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, 2020. 2, 3, 4, 5, 7

[6] Peng Chu, Xiao Bian, Shaopeng Liu, and Haibin Ling. Feature space augmentation for long-tailed data. In *ECCV*, 2020. 1, 3, 7

[7] Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. Class-balanced loss based on effective number of samples. In *CVPR*, 2019. 1, 3, 5, 6, 7

[8] Chris Drummond and Robert Holte. C4.5, class imbalance, and cost sensitivity: Why under-sampling beats oversampling. *ICML Workshop on Learning from Imbalanced Datasets*, 2003. 1, 3

[9] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, 2020. 2, 3

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 4, 6

[11] Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The iNaturalist species classification and detection dataset. In *CVPR*, 2017. 2, 5

[12] Chen Huang, Yining Li, Chen Change Loy, and Xiaoou Tang. Learning deep representation for imbalanced classification. In *CVPR*, 2016. 2

[13] Muhammad Abdullah Jamal, Matthew Brown, Ming-Hsuan Yang, Liqiang Wang, and Boqing Gong. Rethinking class-balanced methods for long-tailed visual recognition from a domain adaptation perspective. In *CVPR*, 2020. 6, 7

[14] Nathalie Japkowicz and Shaju Stephen. The class imbalance problem: A systematic study. *Intelligent Data Analysis*, 2002. 3

[15] Bingyi Kang, Saining Xie, Marcus Rohrbach, Zhicheng Yan, Albert Gordo, Jiashi Feng, and Yannis Kalantidis. Decoupling representation and classifier for long-tailed recognition. In *ICLR*, 2020. 1, 3, 6, 7

[16] Grigoris Karakoulas and John Shawe-Taylor. Optimizing classifiers for imbalanced training sets. In *NIPS*, 1999. 2

[17] S. H. Khan, M. Hayat, M. Bennamoun, F. A. Sohel, and R. Togneri. Cost-sensitive learning of deep feature representations from imbalanced data. *IEEE Transactions on Neural Networks and Learning Systems*, 2018. 3

[18] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *NeurIPS*, 2020. 2, 3, 4, 8

[19] Jaehyung Kim, Jongheon Jeong, and Jinwoo Shin. M2m: Imbalanced classification via major-to-minor translation. In *CVPR*, 2020. 1, 3, 6, 7

[20] A. Krizhevsky and G. Hinton. Learning multiple layers of features from tiny images. *Technical report*, 2009. 5

[21] Miroslav Kubat and Stan Matwin. Addressing the curse of imbalanced training sets: One-sided selection. In *ICML*, 1997. 2

[22] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollar. Focal loss for dense object detection. In *ICCV*, 2017. 6, 7

[23] Jialun Liu, Yifan Sun, Chuchu Han, Zhaopeng Dou, and Wenhui Li. Deep representation learning on long-tailed data: A learnable embedding augmentation perspective. In *CVPR*, 2020. 3

[24] Weiyang Liu, Yandong Wen, Zhiding Yu, and Meng Yang. Large-margin softmax loss for convolutional neural networks. In *ICML*, 2016. 2

[25] 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. 1

[26] Aditya Menon, Harikrishna Narasimhan, Shivani Agarwal, and Sanjay Chawla. On the statistical consistency of algorithms for binary classification under class imbalance. In *ICML*, 2013. 1

[27] Aditya Krishna Menon, Sadeep Jayasumana, Ankit Singh Rawat, Himanshu Jain, Andreas Veit, and Sanjiv Kumar. Long-tail learning via logit adjustment. *arXiv*, 2020. 3

[28] Ajinkya More. Survey of resampling techniques for improving classification performance in unbalanced datasets. *arXiv*, 2016. 1, 3

[29] Sankha Subhra Mullick, Shounak Datta, and Swagatam Das. Generative adversarial minority oversampling. In *ICCV*, 2019. 1, 3

[30] Mengye Ren, Wenyuan Zeng, Bin Yang, and Raquel Urtasun. Learning to reweight examples for robust deep learning. In *ICML*, 2018. 1, 3

[31] Nikolaos Sarafianos, Xiang Xu, and Ioannis A. Kakadiaris. Deep imbalanced attribute classification using visual attention aggregation. In *ECCV*, 2018. 3

[32] Li Shen, Zhouchen Lin, and Qingming Huang. Relay backpropagation for effective learning of deep convolutional neural networks. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *ECCV*, 2016. 3

[33] Jingru Tan, Changbao Wang, Buyu Li, Quanquan Li, Wanli Ouyang, Changqing Yin, and Junjie Yan. Equalization loss for long-tailed object recognition. In *CVPR*, 2020. 1, 3

[34] Kaihua Tang, Jianqiang Huang, and Hanwang Zhang. Long-tailed classification by keeping the good and removing the bad momentum causal effect. In *NeurIPS*, 2020. 6, 7

[35] Byron C. Wallace, Kevin Small, Carla E. Brodley, and Thomas A. Trikalinos. Class imbalance, redux. In *ICDM*, 2011. 1

[36] Junjie Zhang, Lingqiao Liu, Peng Wang, and Chunhua Shen. To balance or not to balance: A simple-yet-effective approach for learning with long-tailed distributions. *arXiv*, 2020. 1

[37] Boyan Zhou, Quan Cui, Xiu-Shen Wei, and Zhao-Min Chen. BBN: Bilateral-branch network with cumulative learning for long-tailed visual recognition. In *CVPR*, 2020. 1, 3, 4, 5, 6, 7