AMA: ASYMPTOTIC MIDPOINT AUGMENTATION FOR MARGIN BALANCING AND MODERATE BROADENING

Anonymous authors
Paper under double-blind review

ABSTRACT

Feature augmentation in neural networks is an effective regularization method to adjust the margin in feature space. However, a similar approach in terms of directly repositioning features, contrastive learning, has reported collapse problems of inter-class and intra-class features. The augmentation approaches are also related to the issues, but have been barely analyzed. In this paper, we show that feature augmentation methods are also affected by the collapse problems and address them by proposing a novel method to generate augmented features gradually approaching the midpoint of inter-class feature pairs, called *asymptotic midpoint augmentation* (AMA). The method induces two effects: 1) balancing the margin for all classes and 2) only moderately broadening the margin until it holds maximal confidence. We empirically analyze *alignment* and *uniformity* to show vulnerability to the problems in a toy task. Then, we validate its impacts in original, long-tailed, and coarse-to-fine transfer tasks on CIFAR-10 and CIFAR-100. To enhance generality, we additionally analyze its relation to a representative input-level augmentation such as Mixup.

1 INTRODUCTION

Augmenting features in neural networks has been effective in regularization by handling margin in feature space(Verma et al. (2019)). The approach generates a *feature*, which indicates a hidden representation of a layer created from an input, and its confidence information from involved original features. A similar approach in the perspective of directly repositioning features, contrastive learning (Chen et al. (2020) He et al. (2020)), learns features distant from a decision boundary by getting centroids of classes further away from each other, and gathering positive pairs closer, which decreases intra-class feature distance and increases inter-class feature distance, measured by *alignment* and *uniformity*, respectively. In the contrastive learning literature, two problems have been recently discussed: collapse of intra-class and inter-class features (Li et al. (2022) Chen et al. (2022)). The first problem is reported in coarse-to-fine transfer learning(Chen et al. (2022)), where all features are closely located on the centroids of each class as the alignment excessively decreases. The second problem is introduced in Supervised Contrastive learning (SupCon) (Khosla et al. (2020)), which uses labels to create positive and negative pairs. The method outperforms other self-supervised learning methods. However, SupCon causes unbalanced margins on long-tailed datasets by overwhelming numerical dominance of the head classes, and it decreases the image classification performance on them. Feature augmentation may also be affected by the collapse problems because of direct feature adjustment. However, the issues have not been deeply analyzed.

In this paper, we show that feature augmentation also suffers from the problems by analyzing alignment and uniformity, and propose a novel feature augmentation method to generate augmented features gradually approaching a decision boundary, called *Asymptotic Midpoint Augmentation* (AMA). AMA has three parts: 1) generating a pool of augmented features by interpolating inter-class feature pairs and pseudo labeling, 2) class-unbiased random sampling, and 3) adaptive interpolation ratio control. The proposed method creates augmented features to make the margin balanced and moderately broad by asymptotically moving them to the midpoint, as shown in Figure 1. As a result, the method shows higher uniformity than before and sufficiently high alignment.

In an experiment on a toy task, we validate the effect of collapses by using alignment and uniformity metrics for AMA and other feature relocation methods such as SupCon(Khosla et al. (2020)) and

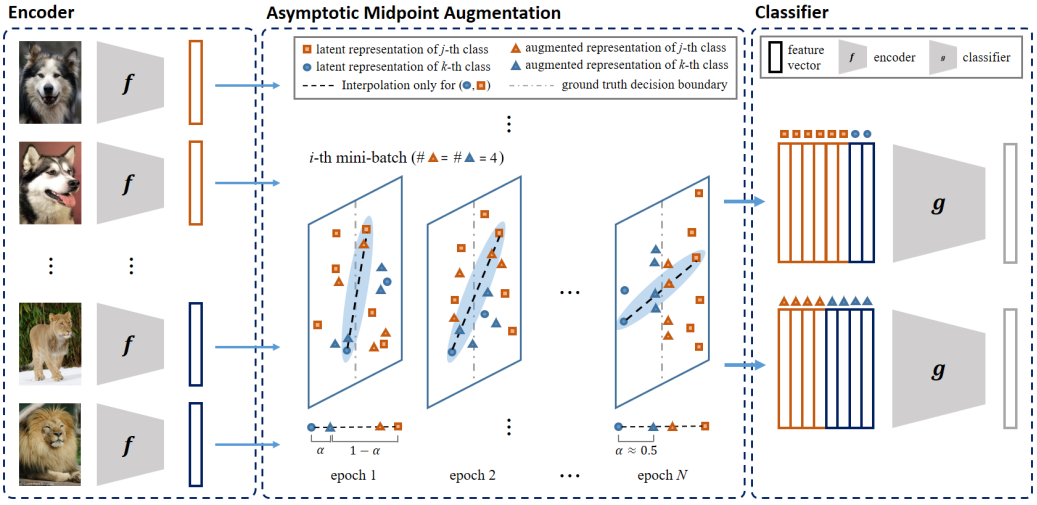


Figure 1: Overview of Asymptotic Midpoint Augmentation. (left) Feature vectors of input samples came from the pre-trained encoder. (middle) Asymptotic Midpoint Augmentation generates augmented features and pseudo labels based on interpolation. Examples for understanding the interpolation are highlighted as cyan. The ratio between two different features is controlled by α and this parameter has asymptotically decreased from 1.0 to 0.5 until the end of training. The augmented features are created as a mini-batch size at the same rate. (right) Finally, original and augmented features are passed to the classifier.

Manifold Mixup(Verma et al. (2019)). We empirically verify the impact of AMA in comparison with the feature augmentation methods in image classification tasks on long-tailed, coarse-to-fine transfer, and original data sets. Additionally, we also analyze the relation of AMA to a representative input-level augmentation method that enhances the different types of information, Mixup(Zhang et al. (2017))

In summary, our main contributions are four-fold:

- we raise the inter-class and intra-class collapse issues in feature augmentation approaches and show their impacts by analyzing alignment and uniformity.
- we propose a novel feature augmentation method, *asymptotic midpoint augmentation*, to address the problems by balancing and moderately broadening the margin in feature space.
- we empirically analyze the effects and performance of AMA and other feature augmentation methods in image classification tasks on long-tailed datasets and coarse-to-fine transfer learning, which are sensitive to collapses.
- we additionally confirm that it maintains performance in the original dataset to inhere uncertain portion of the problems, compare AMA with a representative input-level augmentation method, and analyze their relation.

2 Background

Intra-class collapse Contrastive loss leads the features of positive pairs to be closed to invariant on the noise factor. In contrastive learning, the encoder is forced to ensure that similar samples must be placed at a similar location in the feature space. However, the attraction between positive pairs makes features gather at one point. This phenomenon limits the expressiveness of the model, and it is especially critical for some tasks such as coarse-to-fin transfer learning. More specifically, if a model is pre-trained by coarse-grained labels and then fine-tuned by fine-grained labels, the model would likely not classify fine-grained samples due to the collapsed features. Especially, features in the same class are prone to collapse on the centroids of the class in supervised contrastive learning. We called this problem as *intra-class collapse*. To measure the intra-class collapse, *intra-class alignment* has been proposed, which represents the closeness of positive pairs (Wang & Isola

(2020) Li et al. (2022)). The intra-class alignment can be measured by following:

$$\mathbf{A} = \frac{1}{C}\sum_{i=1}^{C}\frac{1}{|\mathrm{F}_i|^2}\sum_{\boldsymbol{v}_j,\boldsymbol{v}_k\in\mathrm{F}_i}\|\boldsymbol{v}_j-\boldsymbol{v}_k\|_2 \tag{1}$$

, where C is the number of classes, $\boldsymbol{v}$ is a feature vector, and F_i is the set of features from class i. $\|\cdot\|_2$ means L2-norm.

Inter-class collapse Common contrastive learning methods achieve high performance thanks to the property that centroids of the class get further away through repulsion between negative samples. However, supervised contrastive learning tends to make collapse between features in different classes when the dataset is imbalanced, such as long-tailed datasets. More specifically, the model naturally concentrates on getting a large distance between head classes to minimize the loss. For this reason, the contrastive loss is not evenly weighted on all classes. In this situation, features in tail classes would be collapsed each other. We called this collapse as *inter-class collapse*, and it prevents the model from learning regular simplex of features, which is a crucial factor when training on imbalanced datasets in contrastive learning. The inter-class collapse can be measured by *inter-class and neighborhood uniformity*, which are metrics that favor the uniform distribution of representations on the unit hypersphere (Wang & Isola (2020) Li et al. (2022)). The inter-class uniformity measures the pair-wise distance between different classes, and the neighborhood uniformity inspects the convergence of tail classes. These two kinds of metrics can be measured by following $\mathbf{U}$ and $\mathbf{U}_k$, respectively:

$$\mathbf{U} = \frac{1}{C(C-1)}\sum_{i=1}^{C}\sum_{j=1,j\neq i}^{C}\|\bar{\boldsymbol{v}}_i-\bar{\boldsymbol{v}}_j\|_2 \tag{2}$$

$$\mathbf{U}_k = \frac{1}{Ck}\sum_{i=1}^{C}\min_{j_1,\cdots,j_k}\left(\sum_{l=1}^{k}\|\bar{\boldsymbol{v}}_i-\bar{\boldsymbol{v}}_{j_l}\|_2\right) \tag{3}$$

, where $\bar{\boldsymbol{v}}_i$ is the center of samples from class i on the hypersphere: $\bar{\boldsymbol{v}}_i = \frac{\sum_{\boldsymbol{v}_j\in\mathrm{F}_i}\boldsymbol{v}_j}{\|\sum_{\boldsymbol{v}_j\in\mathrm{F}_i}\boldsymbol{v}_j\|_2}$.

In this paper, we do not normalize the center of samples by their norm for a fair comparison with the original method and feature augmentation methods, which do not purpose to learning representations on the hypersphere.

3 Asymptotic Midpoint Augmentation

In this section, we first present our motivation based on preliminary experiments about alignment and uniformity for augmentation and contrastive learning methods. Then, we introduce *asymptotic midpoint augmentations* (AMA) and analyze its effects to feature distribution and decision boundaries.

3.1 Motivation

Experimental Setting To quantitatively measure the intra-class and inter-class collapses, we inspect intra-class alignment, inter-class uniformity, and top-3 neighborhood uniformity in an image classification task on long-tailed CIFAR-100 where the imbalance factor was set to 100. We analyzed those metrics by Eq. 1, 2, and 3. The thing to note here is that we did not normalize the uniformity by class centers for a fair comparison. The experimental settings here are the same as Section 4.3.

Table 1: Alignment and uniformity in an image classification task on long-tailed CIFAR-100 with imbalance factor 100. Superscript $\uparrow$ and $\downarrow$ means that higher is better and lower is better, respectively. Subscript in $\mathbf{U}_k$ means the number of neighbors. Acc.(%) means test accuracy in each model. Best in bold. (Acc.(%): *mean* ± *std*)

	Orig.	SupCon	Mixup*	Manifold Mixup	AMA
$\mathbf{A}^{\downarrow}$	9.36 ± 0.12	**3.34** ± 0.09	5.89 ± 0.07	6.80 ± 0.07	8.85 ± 0.08
$\mathbf{U}^{\uparrow}$	**8.69** ± 0.14	3.83 ± 0.11	5.19 ± 0.06	6.84 ± 0.08	8.14 ± 0.14
$\mathbf{U}_3^{\uparrow}$	**4.25** ± 0.08	1.87 ± 0.06	2.61 ± 0.04	3.49 ± 0.04	4.10 ± 0.04
Acc.$^{\uparrow}$ (%)	43.23 ± 0.39	36.43 ± 0.76	37.00 ± 0.17	40.46 ± 0.42	**45.98** ± 0.31

Collapse Problems Are Important in Feature Augmentation In Table 1, the evidence of collapses and their unignorable impact are observed. First of all, augmentation methods show higher intra-class alignment than SupCon. Optimal intra-class alignment is uncertain and varies by many factors, but SupCon is known as having excessively low intra-class alignment when intra-class collapse occurs. Therefore, it is reasonable that the augmentation methods are alleviating the collapse effect. According to the background, inter-class collapse reduces inter-class uniformity and neighborhood uniformity, and the augmentation methods gradually get higher values in more recent methods. The two observations show the possibility of resolving collapses via feature augmentation, and the corresponding significant increase in accuracy implies that the impact of the collapses can not be ignored. Additionally, Mixup is a data augmentation method on input space, but it also improves the measures, which shows the difference between the augmentation approach to contrastive learning. We introduce this extended experiment in Section 4.6.

3.2 Proposed Method

Notations Let $\mathbf{D} = \{(\boldsymbol{x}_i, c_i) | 1 \leq i \leq n, i \in \mathbb{N}\}$ be the set of pairs of an input vector and its label where $\boldsymbol{x}_i \in \mathbb{R}^d$ and $c_i \in C$ for the class index set C and the pair index i. We define $\boldsymbol{y}_i = [y_1, y_2, ..., y_{|C|}] \in \mathbb{R}^{|C|}$ as one-hot encoding vector for c_i, where $y_{c_i} = 1$. The feature vector of i-th input sample $\boldsymbol{x}_i$, is notated as $\boldsymbol{z}_i \in \mathbb{R}^{|C|}$. The confidence $\boldsymbol{p}$ comes from $\sigma(\boldsymbol{z})$, where $\sigma(\cdot)$ is a function that normalizes an input vector into a range that leads to probabilistic interpretations, similarly to softmax. In this paper, we used softmax function for $\sigma(\cdot)$. Θ and Φ represent the parameters of the networks.

Interpolation-Based Feature Generation and Pseudo Labeling In AMA, augmented features and labels are created as

$$\begin{aligned} \boldsymbol{z}^{(i,j)} &= \alpha \cdot \boldsymbol{z}_i + (1-\alpha) \cdot \boldsymbol{z}_j \\ c^{(i,j)} &= \begin{cases} c_i, & \text{if } \alpha \geq 0.5 \\ c_j, & \text{if } \alpha < 0.5 \end{cases} \end{aligned} \tag{4}$$

, where $\boldsymbol{z}^{(i,j)}$ is an augmented feature generated via interpolation of $\boldsymbol{z}_i$ and $\boldsymbol{z}_j$ selected from different classes, and the pseudo label is $c^{(i,j)}$. This process occurs in the feature space, and the pseudo labels are determined by controlling a parameter α for asymptotically moving them close to the decision boundary. In different with other interpolation-based methods, the labels are definitely determined as one side.

Class-Unbiased Random Sampling We consider how to sample original features for interpolation from two different classes to balance pair-wise margins between them. For this purpose, original features are randomly selected from probabilistic distribution in every mini-batches. Let $\mathbf{D}_B = \{(\boldsymbol{x}_{B,i}, c_{B,i}) | 1 \leq i \leq m, i \in \mathbb{N}\}$ be the pairs of input samples and labels in the mini-batch, where the mini-batch size is m. Then, the probability of selecting $(\boldsymbol{x}_i, c_i)$ from $\mathbf{D}_B$ for interpolation is illustrated in Eq. 5:

$$\mathrm{P}(\boldsymbol{x}_{B,i}) = \frac{1}{C_B} \cdot \frac{1}{N_{c_i}} \tag{5}$$

, where C_B is the number of classes in the mini-batch and N_{c_i} is the number of samples of c_i-th class in the mini-batch. This sampling method allows the decision boundary to be placed in the middle of two engaged classes while maximizing the margin.

Asymptotic move of Augmented Features Confidence is an important factor in estimating the decision boundary. However, it is unreliable to use the pseudo labels as ground truth in early training because neural networks are prone to predict wrong. To reduce this risk, we propose a scheduler that relies on the training accuracy to update α more sensitively, as illustrated in Eq. 6.

$$\alpha = f(v_{acc}) = e^{-\beta \cdot v_{acc}} \tag{6}$$

, where N is the number of epochs and $v_{acc} \in [0, 1]$ means the real value of training accuracy at each epoch. β is a hyperparameter to decide how α decreases as the training accuracy. We set β as 0.67 where α exponentially decreased from 1.0 to about 0.5, and empirically figured out the performance consistently shows best when $\beta = 0.67$ except coarse-to-fine transfer learning environment.

Algorithm 1 Example of Applying AMA to Training a Neural Network for Classification

Input: model parameter Θ and Φ, cross-entropy loss $\mathcal{L}_{\text{CE}}$, AMA loss $\mathcal{L}_{\text{AMA}}$, mini-batch size M, # mini-batches N, balancing parameter α, learning rate η
Output: balanced and moderately broad margin

$\mathbf{D} \leftarrow$ a set of pairs of input samples and labels
$f_\Theta \leftarrow$ encoder, which parameters are Θ
$g_\Phi \leftarrow$ classifier, which parameters are Φ
$\alpha \leftarrow 1.0$
for epoch = $1, 2, \ldots, T$ **do**
  for $i = 1, 2, \ldots, N$ **do**
    $\mathbf{D}_B \leftarrow$ a set of pairs of input samples and labels in the i-th mini-batch
    $\mathbf{X} \leftarrow \{\boldsymbol{x}_{B,1}, \boldsymbol{x}_{B,2}, \ldots, \boldsymbol{x}_{B,M}\}$
    $\mathbf{Z} \leftarrow f_\Theta(\mathbf{X})$
    $\mathbf{Z}_s \leftarrow$ a set of original features selected via class-unbiased random sampling by Eq. 5
    Generate augmented features $\mathbf{Z}^{(\cdot,\cdot)}$ and labels $\boldsymbol{c}^{(\cdot,\cdot)}$ from $\mathbf{Z}_s$ by Eq. 4
    $\mathcal{L}_{\text{CE}} \leftarrow$ cross-entropy loss from $\mathbf{Z}$ by Eq. 7
    $\mathcal{L}_{\text{AMA}} \leftarrow$ AMA loss from $\mathbf{Z}^{(\cdot,\cdot)}$ by Eq. 8
    $\mathcal{L} \leftarrow \mathcal{L}_{\text{CE}} + \mathcal{L}_{\text{AMA}}$
    $\Theta \leftarrow \Theta - \eta \nabla_\Theta \mathcal{L}$
    $\Phi \leftarrow \Phi - \eta \nabla_\Phi \mathcal{L}$
    Update α by Eq. 6
  end for
end for

Training Loss for Augmented Features AMA uses cross-entropy for the augmented features as original features and integrated with original cross-entropy loss as follows.

$$\mathcal{L}_{\text{CE}} = \sum_{\boldsymbol{z} \in \mathbf{Z}} \sum_{k=1}^{C} -y_k \log p_k, \quad \text{where } \boldsymbol{p} = \sigma(\boldsymbol{z}) \tag{7}$$

$$\mathcal{L}_{\text{AMA}} = \sum_{\boldsymbol{z}^{(i,j)} \in \mathbf{Z}^{(\cdot,\cdot)}} \sum_{k=1}^{C} -y_k^{(i,j)} \log p_k^{(i,j)}, \quad \text{where } \boldsymbol{p}^{(i,j)} = \sigma(\boldsymbol{z}^{(i,j)}) \tag{8}$$

where $\mathbf{Z}$ and $\mathbf{Z}^{(i,j)}$ are the set of features and selected augmented features, respectively, and p_k is the probability for the k-th class. An example of integration with a usual classification is shown in Algorithm 1.

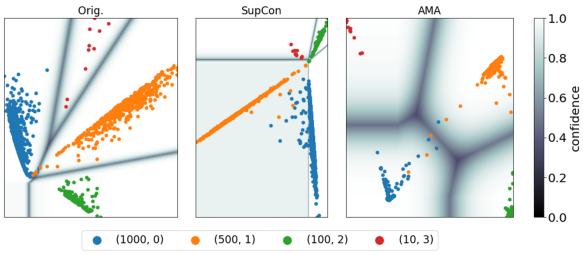
3.3 Effect Analysis

We explain the margin-balancing and moderate margin-broadening effects of AMA and empirically figure out the effects of a simple classification task on a long-tailed toy dataset via qualitative and quantitative analysis.

Margin Balancing AMA forces a decision boundary to locate near the midpoint of inter-class features, because the optimum of AMA loss is obtained when the boundary passes the midpoint for the following reasons: 1) class-unbiased random sampling selects the same number of augmented features for every class, 2) the expected distance of two augmented features to their midpoint is equal, and 3) the sum of their confidences determined by the distance d is $2\sigma(0.5 + d)$ that has the maximum at the midpoint ($d = 0$). Using the guidance to the midpoint repeatedly over many updates, the asymptotic move of the augmented features toward the midpoint reduces the possibility of locating the boundary at the intermediate points between the original and augmented features. Because of this convergence to midpoint by AMA loss, its mixture with other losses is still adjusted to balance margin.

Moderate Margin Broadening AMA broadens margin than original networks. Generally, loss to maximize confidence increases margin in a simple relation of a feature and a decision boundary. AMA adds the gradient of augmented features to the guidance in the same direction because the features are interpolations of original features and have the same label. On the other side, the original features stop being further away from the boundary after obtaining maximal confidence. Because of nearly zero gradients at the state, the distance of intra-class features to their centroids is moderately preserved without excessive converging pressure.

Experimental Setting We randomly generated [1000, 500, 100, 10] training samples and [200, 200, 200, 200] test samples around (-3, 3), (3, 3), (3, -3), and (-3, -3) for four different classes in $\mathbb{R}^2$, respectively. All points were randomly sampled from the Gaussian distribution, where mean



(a) Features and confidence map on 2-dimensional space

	Orig.	SupCon	AMA
Train Acc. ↑ (%)	**99.69**	98.01	**99.69**
Test Acc. ↑ (%)	98.00	96.50	**99.75**
D_{max}	65.0550	110.5938	15.9769
$D_{centroid}$	33.6430	43.9860	10.1800
$D_{relative}$ ↑	0.5171	0.3977	**0.6372**

(b) Accuracy and Distance

Figure 2: Effects of AMA to margin and feature distribution in an image classification task on long-tailed toy dataset. The legend of (a) means *(the number of points, label)*. We removed the axis ticks for the simplicity, but it does not mean they have the same range to each other. (D_{max}: the Euclidean distance of the farthest pair of features, $D_{centroid}$: average distance between all pairs of centroids for classes, $D_{relative}$: $\frac{D_{centroid}}{D_{max}}$.)

and variance are set to 0 and 1, respectively. We used a 4-layer neural network, which has 128-64-2 hidden units in each layer for baselines and AMA. We set the optimizer as SGD at the momentum of 0.9 and weight decay of 5e-4, and the initial learning rate as 0.1. We used 16 mini-batches, and the total number of epochs was 100. In SupCon, we used the first three layers as an encoder and trained the encoder while maintaining the same settings except for epochs set to 600. Then, the last hidden layer was used as a classifier to predict labels with the same settings. To compare the margin, we visualized feature vectors of input samples as points and their confidences as a heat map on 2-dimensional space. Moreover, we analyzed various distances to quantitatively compare how they affect the margin.

Result and Analysis In Figure 2a, AMA learns more balanced margin than the original and SupCon methods. It is shown by the critically narrow area for tail classes (label 2 and 3) compared to the area for head classes (label 0 and 1). Especially, SupCon assigns an extremely large area to the head classes while AMA maintains a relatively similar distance from all boundaries. To investigate the effect of moderate margin-broadening, we quantitatively analyze original, SupCon, and AMA, as shown in Figure 2b. $D_{relative}$ indicates the relative margin of inter-class features compared to the total size of feature distribution. AMA shows the best $D_{relative}$, which is helpful in increasing inter-class uniformity and neighborhood uniformity while maintaining low D_{max}. SupCon improves inter-class uniformity by increasing $D_{centroid}$, but D_{max} increases more than about 7× of AMA. The observation implies that AMA only moderately broadens the margin without an excessive expansion of feature distribution as SupCon.

4 Experiments

We selected two methods as baselines to compare with AMA. SupCon shows our target problem well, and Manifold Mixup is a representative method of feature augmentation. In the followings, all experiments have been run on three different random seeds, and their performances are represented as the mean *mean* and standard deviation *std*. In AMA, β was set to 0.67 as default, and we only annotate when it has a different value.

4.1 Common Settings

We conducted experiments on CIFAR-10, CIFAR-100, and Tiny-ImageNet, which are generally used in image classification benchmarks. Also, we used VGG11(Simonyan & Zisserman (2014)), ResNet32, ResNet50(He et al. (2016)) and DenseNet-BC with 12 growth rate(Huang et al. (2017)). SupCon and Manifold Mixup used the same environmental settings with the following explanation for each task. In Manifold Mixup, we interpolated features only right before the classifier for a fair comparison.

4.2 COARSE-TO-FINE TRANSFER LEARNING TASK

Experimental Setting We conducted coarse-to-fine transfer learning on CIFAR-10 and CIFAR-100. We first trained the ResNet50 with a coarse-grained dataset and fine-tuned the linear classifier with a fine-grained dataset. We used 128 mini-batches and the SGD at the momentum of 0.9 and weight decay of 5e-4. For CIFAR-100, we set the initial learning rate as 0.1 and divided it by five at the 60th, 120th, and 160th epochs, where the total number of epochs is 200. We composed the coarse-grained dataset by splitting the original dataset into a super-class of them. The fine-grained dataset is the same as the original dataset. For CIFAR-10, we followed the hyperparameter and coarse-to-fine dataset settings in Chen et al. (2022).

Table 2: Retention of training ability in coarse-to-fine transfer learning. Best in bold and Second best in underline. (Accuracy (%): *mean* ± *std*)

Method	CIFAR-10	CIFAR-100
Orig.	**66.66** ± 1.51	**62.57** ± 1.53
SupCon	62.46 ± 0.25	57.16 ± 0.22
Manifold Mixup	52.10 ± 1.19	55.65 ± 0.67
AMA ($\beta = 0.3$)	<u>64.65</u> ± 0.12	<u>61.90</u> ± 0.66

Result and Analysis As shown in Table 2, AMA achieved the second-best test accuracy, while SupCon suffers intra-class collapse noticed by low accuracy. In a similar context, Manifold Mixup and AMA also have intra-class collapse by showing lower accuracy than the original method. However, AMA achieves better than SupCon and Manifold Mixup, and it means that AMA alleviates intra-class collapse in coarse-to-fine transfer learning.

4.3 LONG-TAILED TASK

Table 3: Performance in an image classification on long-tailed datasets. CIFAR-10-LT and CIFAR-100-LT mean the long-tailed CIFAR-10 and CIFAR-100, respectively. Best in bold. (Accuracy (%): *mean* ± *std*)

Method	CIFAR-100-LT			CIFAR-10-LT		
Imbalance Factor	100	50	10	100	50	10
Orig.	43.23 ± 0.39	47.71 ± 0.24	59.37 ± 0.17	79.08 ± 0.08	83.06 ± 0.26	89.77 ± 0.08
SupCon	36.43 ± 0.76	39.97 ± 0.23	49.99 ± 0.43	71.96 ± 0.13	82.44 ± 0.15	**90.68** ± 0.12
Manifold Mixup	40.46 ± 0.42	44.39 ± 0.58	54.92 ± 0.38	79.26± 0.16	**83.32** ± 0.41	89.66 ± 0.11
TSC† (Li et al. (2022)	43.8	47.4	59.0	79.7	82.9	88.7
AMA	**45.98** ± 0.31	**50.04** ± 0.27	**59.93** ± 0.46	**80.01** ± 0.45	83.27 ± 0.14	89.44 ± 0.15

Experimental Setting We used ResNet32, 256 mini-batches, the SGD at the momentum of 0.9 and weight decay of 5e-4, and the number of epochs is 400. We set the initial learning rate as 0.0 and warmed up for ten epochs by 0.015. After that, we divided the learning rate by ten at 360th and 380th epochs. The more specific settings are illustrated in Cui et al. (2021).

Result and Analysis As shown in Table 3, AMA attains the best performance except for the imbalance factor set as 50 and 10 in CIFAR-10-LT. Whereas, SupCon shows the worst performance in a high imbalance factor, which means SupCon has inter-class collapse in the long-tailed datasets while AMA learns balanced margin. For this reason, AMA achieved the highest performance by alleviating inter-class collapse between tail classes.

4.4 ORIGINAL IMAGE CLASSIFICATION BENCHMARKS

Experimental Setting We conducted image classification experiments on CIFAR-10, CIFAR-100, and Tiny-ImageNet. For CIFAR-10, we set the initial learning rate as 0.05 and divided the learning rate by two at every 30 epochs among the total of 300 epochs for all networks. For CIFAR-100, we used the hyperparameter same as Section 4.2 for all networks. For Tiny-ImageNet on VGG11 and ResNet50, we used 256 mini-batches, the SGD at a momentum of 0.9 without weight decay, and the number of epochs is 200. We set the initial learning rate as 0.1 and multiplied it by 0.9 at every 20 epochs. For DenseNet-BC ($k = 12$) on Tiny-ImageNet, we used 64 mini-batches, the SGD at a momentum of 0.9 without weight decay, and the number of epochs is 300. We set the initial learning rate as 0.1 and divided it by ten at 150 and 225 epochs.

Table 4: Performance in Image Classification Benchmarks. Best in bold (Accuracy (%): *mean* ± *std*)

Network	Method	CIFAR-10	CIFAR-100	Tiny-ImageNet
VGG11	Orig.	8.23±0.02	31.73±0.19	46.90±0.09
	Manifold Mixup	7.86±0.19	30.62±0.18	47.58±0.29
	AMA	**7.32**±0.20	**29.51**±0.15	**45.52**±0.33
ResNet50	Orig.	4.98±0.12	23.43±0.18	43.15±0.74
	SupCon	**4.43**±0.17	24.23±0.11	41.85±0.17
	Manifold Mixup	5.45±0.39	23.58±0.69	42.17±1.63
	AMA	4.62±0.07	**22.95**±0.65	**41.64**±0.16
DenseNet-BC	Orig.	**5.08**±0.20	**23.04**±0.27	39.77±0.34
	Manifold Mixup	5.43±0.18	23.45±0.11	**37.57**±0.09
	AMA	5.36±0.18	23.07±0.12	38.36±0.42

Result and Analysis As shown in Table 4, AMA achieved competitive or even high performance with other representation augmentation based-models. Specifically for VGG11, AMA retained the highest performance overall. It implies AMA sustains proper alignment and high uniformity without interruption for representation learning.

4.5 Ablation Study

We conducted the ablation study to clarify the effects of all parts in AMA: interpolation, class-unbiased random sampling and asymptotic move of augmented features. Table 5 shows the effect of components in AMA. In this experiment, we did experiments in coarse-to-fine transfer on CIFAR-100 and in the image classification on CIFAR-100-LT (imbalance factor: 100) with the same settings each. In coarse-to-fine transfer learning, AMA without CR shows the second-best performance. It implies the asymptotic move of augmented features is more stable than simply locating augmented features at the midpoint since the beginning. Class-unbiased random sampling exhibits its impact in the long-tailed dataset. By mitigating unbiased augmented features, the model could learn more balanced margins. Overall, using these two components together shows the best performance proving their synergy in AMA.

Table 5: Ablation study on AMA. When AM not applied, $\alpha = 0.51$ (I: Interpolation, CR: Class-unbiased Random sampling, AM: Asymptotic move of augmented features)

I	CR	AM	*Coarse-to-Fine Transfer*	*Long-tailed*
✓	✓	✓	**61.90** ± 0.66	**45.98** ± 0.31
✓	✓		56.95 ± 1.05	44.23 ± 0.31
✓		✓	59.70 ± 1.32	42.10 ± 0.02
✓			57.13 ± 1.34	41.33 ± 0.13

4.6 Analysis with Mixup

In our motivation experiments, we found that two collapse problems also occur in the data augmentation method as Mixup(Zhang et al. (2017)). For the exploration of AMA to data augmentation approach, we first apply AMA to Mixup and figured out that AMA is helpful to alleviate the collapses in long-tailed and coarse-to-fine transfer learning tasks. In this analysis, experimental settings are the same as Sections 4.1, 4.2, and 4.3.

Table 6: Coarse-To-Fine Grained Transfer Learning. Best in bold (Accuracy (%): *mean* ± *std*)

Method	CIFAR-10	CIFAR-100
Orig.	**66.66** ± 1.51	**62.57** ± 1.53
Mixup	62.22 ± 0.30	60.13 ± 1.01
AMA	64.65 ± 0.12	61.90 ± 0.66
AMA + Mixup	66.50 ± 1.30	61.92 ± 1.57

Table 7: Image Classification on Long-Tailed Dataset. Best in bold (Accuracy (%): *mean* ± *std*)

Method	CIFAR-100-LT			CIFAR-10-LT		
Imbalance Factor	100	50	10	100	50	10
Orig.	43.23 ± 0.39	47.71 ± 0.24	59.37 ± 0.17	79.08 ± 0.08	83.06 ± 0.26	**89.77** ± 0.08
Mixup	37.00 ± 0.17	40.41 ± 0.20	51.14 ± 0.26	74.50 ± 0.71	79.11 ± 0.56	87.13 ± 0.09
AMA	**45.98** ± 0.31	**50.04** ± 0.27	**59.93** ± 0.46	**80.01** ± 0.45	**83.27** ± 0.14	89.44 ± 0.15
AMA + Mixup	43.83 ± 0.26	47.38 ± 0.56	58.11 ± 0.14	75.93 ± 0.18	80.01 ± 0.80	88.56 ± 0.34

Results and Analysis In both experiments, Mixup causes performance degradation overall. However, the mixture of AMA and Mixup shows better performance than using only Mixup and almost recovers the original performance. As a result, feature augmentation helps Mixup alleviate intra-class and inter-class collapses.

5 Related Work

5.1 Augmentation

Data augmentation has been one of the effective regularization techniques(Zhang et al. (2017) Shorten & Khoshgoftaar (2019) DeVries & Taylor (2017) Cubuk et al. (2018) Zhong et al. (2020) Moreno-Barea et al. (2018)). Mixup(Zhang et al. (2017)), a generally used approach among data augmentations, interpolates each pair of input samples and labels in the input space. Using this interpolation, it is possible for models to improve their inductive bias. In other streams, data augmentation has been applied to features in feature space, called feature augmentation Verma et al. (2019) Li et al. (2021) Kuo et al. (2020) Lee et al. (2021) Wang et al. (2021)). In Manifold Mixup(Verma et al. (2019)), models get a smoother decision boundary than before, and it results in the improvement of robustness. However, they have not focused on margin, which is an important component to make decision boundary robust, while our proposed method creates augmented features in the feature space and adjusts the augmentation to make the margin balanced and moderately wide.

5.2 Contrastive Learning

Contrastive learning achieved state-of-the-art performance in image classification tasks, which is an example of focusing on the margin(Chen et al. (2020) He et al. (2020) Caron et al. (2020) Li et al. (2020) Gutmann & Hyvärinen (2010) Koch et al. (2015) Khosla et al. (2020)). Contrastive learning attracts positive samples and repulses negative samples from the anchor. In supervised approaches, SupCon(Khosla et al. (2020)) uses label information to choose positive pairs and negative pairs. SupCon can effectively get considerable uniformity between inter-class and minor alignment between intra-class. This property leads to ideal representations, which have a large margin between other classes. In spite of these advantages, Supcon has an unavoidable problem of *collapse*(Jing et al. (2021)) because each sample converged toward the class centroid. This collapse makes features indistinguishable from each other and can lead to poor performance in coarse-to-fine transfer learning(Chen et al. (2022)). In addition, prior works have focused on relatively low performance in long-tailed tasks when using SupCon(Zhu et al. (2022) Li et al. (2022)). In the long-tailed tasks, SupCon leads to overwhelming concentration on head classes, and it encourages the collapse between tail classes. To solve this problem, BCL(Zhu et al. (2022)) used class-average and class-complement with SupCon loss and TSC(Li et al. (2022)) forced class centroids to form a regular simplex on the hypersphere. In contrast, we learn balanced and moderately broad margin while avoiding collapse by creating augmented features as asymptotically moving to the midpoint.

6 Conclusion

In this paper, we raised the two collapse problems of feature augmentation, which are recently discussed in contrastive learning literature. We found that the problems were still important in state-of-the-art feature augmentation method as Manifold Mixup by analyzing alignment and uniformity used as indicators of the collapse problems. To address the collapse problems, we proposed *Asymptotic Midpoint Augmentation* to generate effective features via 1) interpolation of features with pseudo labeling, 2) class-unbiased random sampling of augmented features, and 3) their asymptotic move. The method showed the two effects of margin balancing and moderate-broadening, and their impact on the collapse problems in quantitative and qualitative analysis of a toy long-tailed classification task. In more practical long-tailed and coarse-to-fine transfer learning experiments on CIFAR-10 and CIFAR-100 datasets, which suffered from inter-class and intra-class collapse respectively, AMA significantly alleviated the performance compared to SupCon and Manifold Mixup. Ablation study and relation to data augmentation method as Mixup are also analyzed for validating their deep and broader impact. A limit is that AMA may require additional tuning of hyperparameter β to obtain the best performance because of different intensities of the collapse problems by tasks.

ETHICS STATEMENT

In this paragraph, we address potential concerns below:

- studies that involve human subjects: N/A
- practices to data set releases: CIFAR-10, CIFAR-100, Tiny-ImageNet, CIFAR-10-LT, CIFAR-100-LT.(See Sections 4.2, 4.3, and 4.4)
- potentially harmful insights, methodologies and applications: N/A
- potential conflicts of interest and sponsorship, discrimination/bias/fairness concerns, privacy and security issues, legal compliance, and research integrity issues: N/A

REPRODUCIBILITY STATEMENT

In this paragraph, we capsulize contents for reproducing our results.

- Experiment settings
  1. A Simple Classification Task on Long-Tailed Toy Dataset: Section 3.3
  2. Coarse-to-Fine Transfer Learning: Section4.2
  3. Image Classification on Long-Tailed Dataset: Section4.3
  4. Image Classification on Classic Dataset: Section4.4
- Code Description in Supplementary material
  1. Experimental Details
  2. Requirements
  3. Training and Evaluation
     (a) How to run Coarse-to-Fine Transfer Learning
     (b) How to run Image Classification on Long-Tailed Dataset
     (c) How to run Image Classification on Classic Dataset
  4. Reference

REFERENCES

Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *Advances in Neural Information Processing Systems*, 33:9912–9924, 2020.

Mayee Chen, Daniel Y Fu, Avanika Narayan, Michael Zhang, Zhao Song, Kayvon Fatahalian, and Christopher Ré. Perfectly balanced: Improving transfer and robustness of supervised contrastive learning. In *International Conference on Machine Learning*, pp. 3090–3122. PMLR, 2022.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pp. 1597–1607. PMLR, 2020.

Ekin D Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V Le. Autoaugment: Learning augmentation policies from data. *arXiv preprint arXiv:1805.09501*, 2018.

Jiequan Cui, Zhisheng Zhong, Shu Liu, Bei Yu, and Jiaya Jia. Parametric contrastive learning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 715–724, 2021.

Terrance DeVries and Graham W Taylor. Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552*, 2017.

Michael Gutmann and Aapo Hyvärinen. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pp. 297–304. JMLR Workshop and Conference Proceedings, 2010.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9729–9738, 2020.

Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4700–4708, 2017.

Li Jing, Pascal Vincent, Yann LeCun, and Yuandong Tian. Understanding dimensional collapse in contrastive self-supervised learning. *arXiv preprint arXiv:2110.09348*, 2021.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *Advances in Neural Information Processing Systems*, 33:18661–18673, 2020.

Gregory Koch, Richard Zemel, Ruslan Salakhutdinov, et al. Siamese neural networks for one-shot image recognition. In *ICML deep learning workshop*, volume 2, pp. 0. Lille, 2015.

Chia-Wen Kuo, Chih-Yao Ma, Jia-Bin Huang, and Zsolt Kira. Featmatch: Feature-based augmentation for semi-supervised learning. In *European Conference on Computer Vision*, pp. 479–495. Springer, 2020.

Jungsoo Lee, Eungyeup Kim, Juyoung Lee, Jihyeon Lee, and Jaegul Choo. Learning debiased representation via disentangled feature augmentation. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 25123–25133. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper/2021/file/d360a502598a4b64b936683b44a5523a-Paper.pdf.

Junnan Li, Pan Zhou, Caiming Xiong, and Steven CH Hoi. Prototypical contrastive learning of unsupervised representations. *arXiv preprint arXiv:2005.04966*, 2020.

Pan Li, Da Li, Wei Li, Shaogang Gong, Yanwei Fu, and Timothy M Hospedales. A simple feature augmentation for domain generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8886–8895, 2021.

Tianhong Li, Peng Cao, Yuan Yuan, Lijie Fan, Yuzhe Yang, Rogerio S Feris, Piotr Indyk, and Dina Katabi. Targeted supervised contrastive learning for long-tailed recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6918–6928, 2022.

Francisco J Moreno-Barea, Fiammetta Strazzera, José M Jerez, Daniel Urda, and Leonardo Franco. Forward noise adjustment scheme for data augmentation. In *2018 IEEE symposium series on computational intelligence (SSCI)*, pp. 728–734. IEEE, 2018.

Connor Shorten and Taghi M Khoshgoftaar. A survey on image data augmentation for deep learning. *Journal of big data*, 6(1):1–48, 2019.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

Vikas Verma, Alex Lamb, Christopher Beckham, Amir Najafi, Ioannis Mitliagkas, David Lopez-Paz, and Yoshua Bengio. Manifold mixup: Better representations by interpolating hidden states. In *International Conference on Machine Learning*, pp. 6438–6447. PMLR, 2019.

Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International Conference on Machine Learning*, pp. 9929–9939. PMLR, 2020.

Yulin Wang, Gao Huang, Shiji Song, Xuran Pan, Yitong Xia, and Cheng Wu. Regularizing deep networks with semantic data augmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017.

Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. Random erasing data augmentation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pp. 13001–13008, 2020.

Jianggang Zhu, Zheng Wang, Jingjing Chen, Yi-Ping Phoebe Chen, and Yu-Gang Jiang. Balanced contrastive learning for long-tailed visual recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6908–6917, 2022.