T-distributed Spherical Feature Representation for Imbalanced Classification

Xiaoyu Yang[1], **Yufei Chen**[1*], **Xiaodong Yue**[2,3,4], **Shaoxun Xu**[1], **Chao Ma**[5]

[1] College of Electronics and Information Engineering, Tongji University, Shanghai, China
[2] School of Computer Engineering and Science, Shanghai University, Shanghai, China
[3] Artificial Intelligence Institute of Shanghai University, Shanghai, China
[4] VLN Lab, NAVI MedTech Co., Ltd. Shanghai, China
[5] Department of Radiology, Changhai Hospital of Shanghai, Shanghai, China
yufeichen@tongji.edu.cn

Abstract

Real-world classification tasks often show an extremely imbalanced problem. The extreme imbalance will cause a strong bias that the decision boundary of the classifier is completely dominated by the categories with abundant samples, which are also called the head categories. Current methods have alleviated the imbalanced impact from mainly three aspects: class re-balance, decoupling and domain adaptation. However, the existing criterion with the winner-take-all strategy still leads to the crowding problem in the eigenspace. The head categories with many samples can extract features more accurately, but occupy most of the eigenspace. The tail categories sharing the rest of the narrow eigenspace are too crowded together to accurately extract features. Above these issues, we propose a novel T-distributed spherical metric for equalized eigenspace in the imbalanced classification, which has the following innovations: 1) We design the T-distributed spherical metric, which has the characteristics of high kurtosis. Instead of the winner-take-all strategy, the T-distributed spherical metric produces a high logit only when the extracted feature is close enough to the category center, without a strong bias against other categories. 2) The T-distributed spherical metric is integrated into the classifier, which is able to equalize the eigenspace for alleviating the crowding issue in the imbalanced problem. The equalized eigenspace by the T-distributed spherical classifier is capable of improving the accuracy of the tail categories while maintaining the accuracy of the head, which significantly promotes the intraclass compactness and interclass separability of features. Extensive experiments on large-scale imbalanced datasets verify our method, which shows superior results in the long-tailed CIFAR-100/-10 with the imbalanced ratio $\mathbf{IR} = 100/50$. Our method also achieves excellent results on the large-scale ImageNet-LT dataset and the iNaturalist dataset with various backbones. In addition, we provide a case study of the real clinical classification of pancreatic tumor subtypes with 6 categories. Among them, the largest number of PDAC accounts for 315 cases, and the least CP has only 8 cases. After 4-fold cross-validation, we achieved a top-1 accuracy of 69.04%.

Introduction

In visual classification, the extremely imbalanced problem is increasingly a critical challenge. The head categories represent categories with a lot of samples, and the tail categories denote categories with only a few samples. Due to the limitation of sampling, the head categories with abundant samples are in minority, and the most of categories only have a few samples, which are tail categories. Under the extremely imbalanced condition, the classifier will fall into the local optimal solution, that only improves the accuracy of head categories and ignores the tail categories. In the most extreme case, the classifier will fail to train that all samples are predicted as the class with the largest number. Especially as large-scale classification problems, such as Image-Net (Russakovsky et al. 2015) and iNaturalist (Cui et al. 2018), become more granular and sophisticated, the head categories will dominate the whole classifier, resulting in a strong bias in the eigenspace that totally ignores the tail categories. To address this critical issue, a number of methods are proposed for extremely imbalanced recognition. The current mainstream methods are mainly divided into three categories: class re-balance, decoupling and domain adaptation.

Class re-balance. There are two main aspects of class re-balance: re-sampling and re-weighting. Re-sampling (Chou et al. 2020; Yang and Xu 2020) alleviates the extremely imbalanced problem by simply oversampling the tail categories and undersampling the head categories. The class re-weighting (Byrd and Lipton 2019; Cao et al. 2019; Cui et al. 2019a; Jamal et al. 2020) believes that as the training samples number of one certain category increases, the improvement of accuracy will gradually decrease. Therefore, the loss function is used to assign different weights to different numbers of categories to improve the accuracy of the tail categories. However, the re-balance method improves the performance of the tail categories at the cost of the head categories, that abundant samples in head categories are not fully utilized.

Decoupling. BBN (Zhou et al. 2020) and LWS (Kang et al. 2019) are representative of the method of decoupling. They think that the distribution of image features and the distribution of category labels are essentially uncoupled. So the essence of the decoupling method should separate the distribution of image features from the distribution of categories. (Kang et al. 2019; Zhong et al. 2021) use a two-stage strategy. In the first stage, a traditional model is normally trained upon the imbalanced data to construct the feature space. In

*Corresponding author

the second stage, the parameters of the feature extractor are fixed, and a new classifier is fine-tuned to elaborate the decision boundaries of the eigenspace. (Zhou et al. 2020) merges these two steps into a neural network to achieve decoupling of the distributions between image features and categories. In addition, (Wang et al. 2020; Cai, Wang, and Hwang 2021) utilize multiple expert networks to predict categories of different number-level, and different expert networks are connected through a routing module. So the classification of the tail category is not affected by the head category. Despite their efficiency, these methods either require complex training strategies or drastically increase the parameters.

Domain Adaption. LDA (Peng et al. 2021), as the representative, formulates the imbalanced recognition as domain adaption, by modeling the imbalanced distribution as an unbalanced domain and the general distribution as a balanced domain to alleviate the extremely imbalanced problem. In addition, (Kobayashi 2021) makes improvement from the view of regularization for cosine similarity, which is known for normalizing features but causing larger within-class variance of features. Besides,(Tang, Huang, and Zhang 2021) uses the normalization of multi-head to remove the bad momentum and keep the good momentum in the optimizer, which makes the optimization direction of the imbalanced dataset consistent with the optimization direction of the balanced dataset. But some of these methods need pre-trained models, and the similarities and disparities between the head domain and the tail domain still need further research.

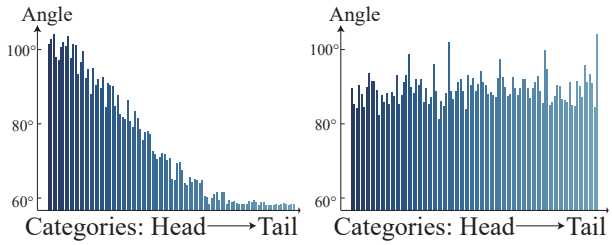


Figure 1: The cosine distance of each center in the hypersphere classifier. Left: Trained on the long-tailed CIFAR-100. Right: Trained on the original CIFAR-100 (balanced)

We reconsider the imbalanced problem from the perspective of metric learning. Ideally, the more the number of training samples, the more accurate the feature extraction, and the less the eigenspace occupied by the head categories. However, current metrics improve the accuracy of the head category at the expense of squeezing the tail category eigenspace and allocating them to the head, which causes the crowding problem.

The crowding problem in the eigenspace under the imbalanced condition is shown in Fig.1. The softmax-based cosine classifier is trained on the long-tailed CIFAR-100 dataset (imbalanced ratio $\mathbf{IR} = 100$) and the balanced CIFAR-100 dataset, respectively. Fig.1 shows the average angle distance of each category center in the $\mathbb{R}^{128}$ hyper-sphere. A smaller average angle means more crowded, and vice versa. Compared with the balanced circumstance, the centers of tail categories are more crowded with the eigenspace of only about 60° under the extremely imbalanced condition, that head categories dominate most of the hyper-spherical eigenspace with an angle of more than 100°. The head categories can extract accurate features due to their large number of samples, but occupy most of the eigenspace, causing a huge waste. The tail categories need a lot of eigenspace because of their rare training samples, but they are crowded together, sharing the rest of the small eigenspace. The crowding problem poses a huge challenge for imbalanced tasks.

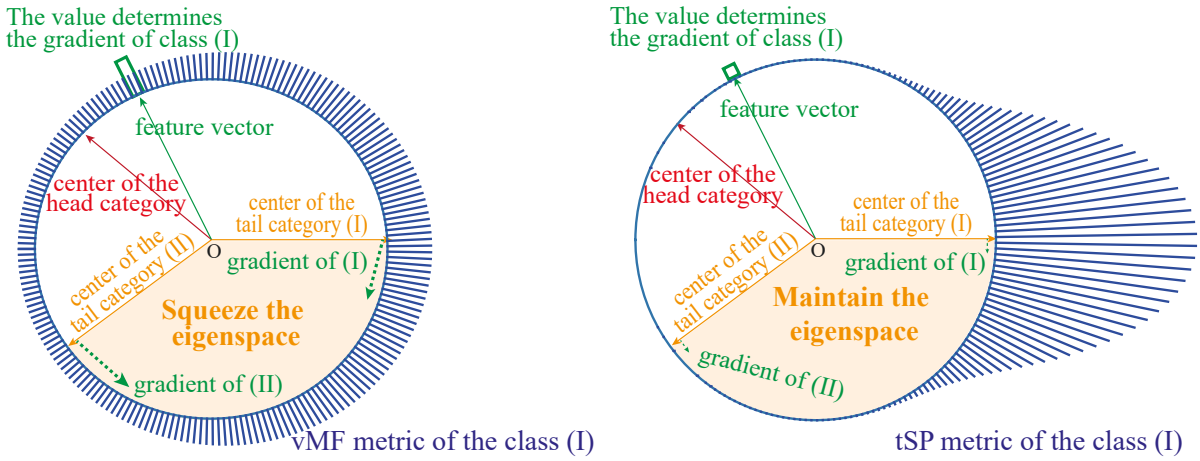
The cause of the crowding is the winner-take-all strategy of the current metrics, that one category with only marginally higher logit than other categories will receive a disproportionately large share of the eigenspace. Due to the abundant samples of the head categories, they are probabilistically more likely to appear than other categories, resulting in a marginally higher logit against other categories. With the winner-take-all strategy, the head categories will dominate the whole classifier, occupy most of the eigenspace, and keep squeezing the tail categories for a larger eigenspace, even though their accuracy is already high. The strong bias brought by the winner-take-all strategy intensely affects the classifier, causing serious crowding problems.

Based on the above issues and expectations, from the perspective of metric learning, we design a T-distributed spherical metric (tSP) on the hyper-sphere, and embed it into the classifier as the distance metric. In general, our paper mainly makes the following contributions:

1) We propose a T-distributed spherical metric, which has the characteristic of high kurtosis. Instead of current methods with the winner-take-all strategy, the T-distributed spherical metric produces a high logit only when the extracted feature is close enough to the category center without a strong bias against other categories.
2) Integrating the T-distributed spherical metric, we design a T-distributed spherical classifier to equalize the eigenspace for alleviating the crowding issue in the imbalanced problem, which significantly promotes the intraclass compactness and interclass separability of features without any pre-trained model.
3) We have verified our method on various public imbalanced datasets. Our model achieved superior results in the long-tailed CIFAR-100/-10 dataset with imbalanced ratio $\mathbf{IR} = 100/50$, and also got excellent results on the ImageNet-LT and iNaturalist dataset with different backbones. In addition, we also provided a case study that applied our model to the real clinical classification of pancreatic tumor subtype with Top-1 accuracy of 69.04%.

Methodology

In this part, we first introduce the preliminaries about directional statistics in metric learning, and revisit the softmax-based cosine distance classifier from the perspective of directional statistics, as the representative of the winner-take-all strategy that causes the crowding problem. Then we propose the T-distributed spherical metric with the characteristic of high kurtosis. Integrating with the T-distributed spherical metric of good properties, a novel classifier is designed



(a) vMF (b) tSP

Figure 2: The crowding performance of different metrics in $\mathbb{S}^1$ hyper-sphere.

to alleviate the crowding problem for the imbalanced classification.

Directional Statistics

Given a training dataset $D = \{(z_i, y_i)|i = 0, 1, ..., n\}$, where z_i and y_i represent the i-th pair of image and label respectively, and n denotes the number of pairs in the training dataset. The feature extractor establishes the mapping $f_\theta : z_i \to x_i$ from image z_i to feature $x_i \in \mathbb{S}^{d-1}$, where $\mathbb{S}^{d-1} = \{x \in \mathbb{R}^d : ||x||_2 = 1\}$ denotes the hyper-spherical set. A key idea in directional distribution is the tangent-normal decomposition. Any unit vector x can be decomposed as:

$$x = t\mu + (1 - t^2)^{\frac{1}{2}} v \tag{1}$$

with $t \in [-1, 1]$ and $v \in \mathbb{S}^{d-2}$ a tangent to $\mathbb{S}^{d-1}$ at μ (Mardia and Jupp 2000; Cao and Aziz 2020), where v and t are independent and v is uniform on $\mathbb{S}^{d-2}$. Thus, the intersection of $\mathbb{S}^{d-1}$ with the hyperplane through $t\mu$ and normal to μ is a $(d-2)$-dimensional sphere of radius $\sqrt{1-t^2}$, that t has density as following:

$$p_T(t; d) \propto (1 - t^2)^{\frac{d-3}{2}} \tag{2}$$

with $t \in [-1, 1]$. Therefore, through the marginal density p_T and p_v, we can estimate the density of the entire spherical distribution.

Revisiting Softmax-based Cosine Classifier

Taking the softmax-based cosine distance classifier as a classical example with the winner-take-all strategy. The cosine distance classifier with softmax can be considered as a spherical distribution that follows the density:

$$p_X(x; \mu) \propto \exp(\mu^T x) \tag{3}$$

where the feature $x \in \mathbb{S}^{(d-1)}$ and the center $\mu \in \mathbb{S}^{(d-1)}$. The density conforms to the form of the von Mises-Fisher distribution (vMF) (Banerjee et al. 2005) under the condition of concentration $\kappa = 1$, that follows the density:

$$p_X(x; \mu, \kappa) \propto \exp(\kappa \mu^T x) \tag{4}$$

where exp represents the exponential function. So, combined with Eq.2, we can get the normalizer of the marginal distribution $p_T(t; \kappa, d)$:

$$\begin{aligned} C_T(\kappa, d) &= \int_{\mathbb{S}^{d-1}} \exp(\kappa \mu^T x) \mathrm{d}x \\ &= \left(\frac{\kappa}{2}\right)^{\frac{d}{2}-1} \left\{ \Gamma\left(\frac{d-1}{2}\right) \Gamma\left(\frac{1}{2}\right) I_{\frac{d-1}{2}}(\kappa) \right\}^{-1} \end{aligned} \tag{5}$$

where I_m denotes the modified Bessel function of the first kind at order m, and $\Gamma(\cdot)$ represents the gamma function. Further, according to the Eq.1, the normalizer $C_X(\kappa, d)$ of density $p_X(x; \mu, \kappa)$ is the product of the normalizer $C_T(\kappa, d)$ and the uniform distribution on $\mathbb{S}^{d-2}$, which is:

$$C_X(\kappa, d) = C_T(\kappa, d) \cdot A_{d-2} = \frac{(2\pi)^{d/2} I_{d/2-1}(\kappa)}{\kappa^{d/2-1}} \tag{6}$$

where $A_{d-1} = (2\pi^{\frac{d}{2}})/\Gamma\left(\frac{d}{2}\right)$ denotes the surface area of the hyper-sphere $\mathbb{S}^{d-1}$. Thus, the density of von Mises-Fisher is:

$$p(x) = C_X(\kappa, d)^{-1} \exp(\kappa \mu^T x), \quad x \sim \text{vMF}(\mu, \kappa) \tag{7}$$

Based on the above analysis, we discuss the shortcomings of the softmax-based cosine distance classifier in the imbalanced problem. Due to the Eq.4 characteristic of vMF, it is essentially a winner-take-all classifier, which will receive an exponential share of the eigenspace with only a marginally higher logit. Fig.2a also illustrates that even though the head samples are close enough to the head center, it will produce a high gradient to the tail categories and squeeze them. Under the imbalanced circumstance, the head categories with many samples will dominate the whole softmax-based cosine classifier and cause the decision boundary to expand

continuously. Most of the eigenspace will be occupied by the head categories, leading to the squeezing of the eigenspace of other categories, especially when the concentrate $\kappa = 1$. The rest of the small eigenspace is shared by other categories, and the classifier will be stuck into a local optimum.

T-distributed Spherical Metric

To solve the above problems, inspired by the T-SNE(Van der Maaten and Hinton 2008), we design a T-distribution spherical metric (tSP) with one degree of freedom, which follows the density:

$$p_X(x) \propto (1 - \frac{1+\mu^T x}{2})^{-1} \tag{8}$$

where the feature $x \in \mathbb{S}^{d-1}$ and the center $\mu \in \mathbb{S}^{d-1}$. Through the Eq.2, we can get the marginal normalizer:

$$N_T(d) = \int_{\mathbb{S}^{d-1}} \frac{2}{1-\mu^T x} \mathrm{d}x = 2^{\alpha+\beta} \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)} \tag{9}$$

where $\alpha = \frac{d-1}{2}$ and $\beta = \frac{d-3}{2}$. Combined with Eq.1, the normalizer $N_X(d)$ of density $p_X(x;\mu)$ is:

$$N_X(d) = N_T(d) \cdot A_{d-2} = 2^{d-1}\pi^{\frac{d}{2}} \frac{\Gamma(\beta)}{\Gamma(\alpha+\beta)} \tag{10}$$

Thus, the probability density function is as followed:

$$p(x) = N_X(d)^{-1}(1 - \frac{1+\mu^T x}{2})^{-1}, \quad x \sim \text{tSP}(\mu) \tag{11}$$

According to Eq.11, the basic properties of the tSP distribution are given in Table.1, and the $\psi(\cdot)$ denotes the digamma function. And the vMF distribution with concentrate $\kappa = 1$ and tSP distribution on the hyper-sphere are shown in Fig.3.

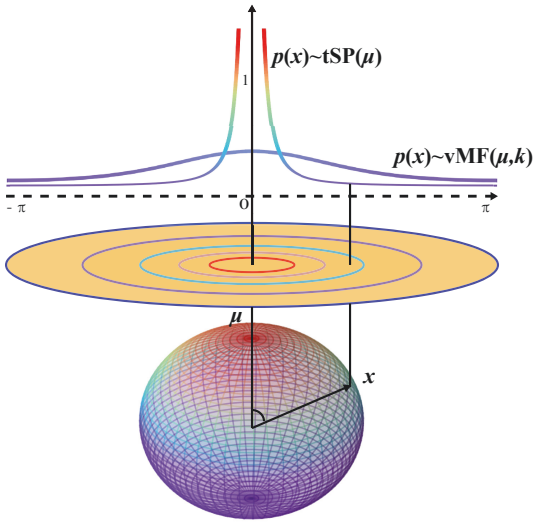


Figure 3: tSP distribution and vMF distribution in $\mathbb{S}^2$ hyper-sphere.

As we can see from Fig.3, tSP is a distribution with the characteristic of high kurtosis, that has a higher tip and slim body than the vMF distribution. Only when the extracted features are sufficiently obvious, that the feature is close enough to the category center, the classifier will generate high confidence in this category, which is in line with the human thinking mechanism. Otherwise, the classifier will only generate particularly low confidence without a bias against other categories. Furthermore, the movement of the category center parameters is shown in Fig.2. Compared with the vMF, the feature vector belonging to the head category in the tSP will only achieve a small logit in the tail category. So, the gradients of the tail will also be small, and the tail categories will not be crowded. It means that the eigenspace of those head categories in the long-tailed dataset occupies less, and the limited feature space can be more allocated to the features of tail categories. The tSP metric can equalize the eigenspace as much as possible, alleviating the crowding issue in the extremely imbalanced problem.

Properties	**Value**
$\mathbb{E}(X)$	$(\alpha-\beta)/(\alpha+\beta)\mu$
Var[X]	$\frac{2\alpha}{(\alpha+\beta)^2(\alpha+\beta+1)}\left((\beta-\alpha)\mu\mu^\top + (\alpha+\beta)I_d\right)$
Mode	μ
H(x)	$\log N_X(d) - (\log 2 + \psi(\alpha) - \psi(\alpha+\beta))$

Table 1: Properties of tSP distribution.

T-distributed Spherical Classifier

Based on the good properties of the T-distributed spherical metric, we designed the T-distributed hyper-spherical classifier to alleviate the crowding problem in the imbalanced classification. There are K centers $\boldsymbol{\mu} = \{\mu_k | k = 0, 1, ..., K-1\}$ in the T-distributed hyper-spherical classifier, where K denotes the number of categories. For the i-th sample z_i, the extracted feature x_i is obtained through the backbone neural network and L2-based normalization. And the posterior probability p_{ik} of the k-th class are obtained by the feature x_i through the T-distributed hyper-spherical classifier is given by:

$$p_{ik} = \frac{(1 - \frac{1+\mu_k^T x_i}{2})^{-1}}{\sum_{j=0}^{K}(1 - \frac{1+\mu_j^T x_i}{2})^{-1}} \tag{12}$$

Since the dimension of the feature x of each sample is the same, the normalization term of the T-distributed spherical distribution $N_X(d)$ is ignored in the calculation of logits. Besides, high kurtosis will bring a narrow margin, making training hard. Therefore, we use a trainable parameter to relax the margin of the T-distributed spherical classifier, reducing the training difficulty.

The workflow of our method is shown in Fig.4. First, the input image is extracted by the backbone network to obtain the feature. The feature will be mapped on the hyper-sphere through the L2-based normalization method. Finally, the result of logits is obtained through the T-distributed hyper-spherical classifier.

Due to the high kurtosis characteristics of the T-distributed spherical distribution, the T-distributed spherical

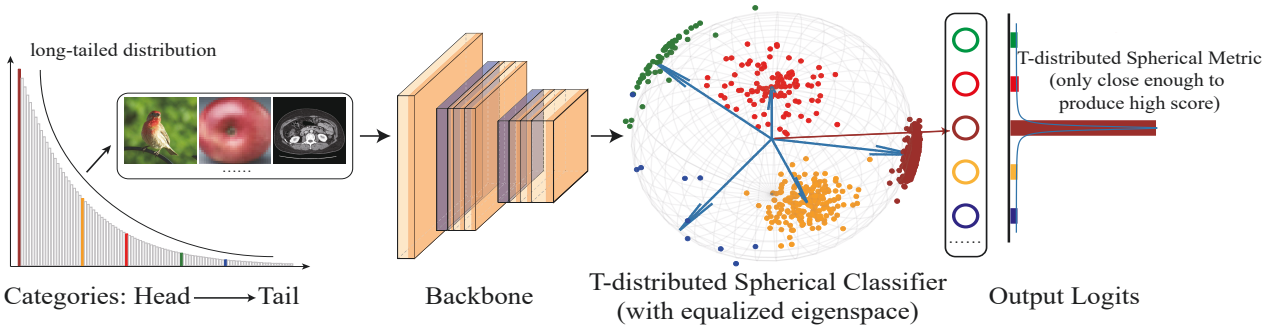


Figure 4: The workflow of our proposed model.

classifier is able to equalize the eigenspace for alleviating the crowding issue in the imbalanced problem. The equalized eigenspace is capable of improving the accuracy of the tail categories while maintaining the accuracy of the head categories, which significantly promotes the intraclass compactness and interclass separability of features. Instead of sacrificing the head categories to improve the tail categories, our model allocates the eigenspace more equally and more sufficiently, so that each category can get as much eigenspace as possible without affecting other categories.

Experiments

Dataset and Evaluation

Our proposed method was evaluated on major public imbalanced datasets and various backbones. The improvements across different tasks show the generalization and feasibility of our method.

1. Long-tailed CIFAR-10/-100: we constructed the long-tailed CIFAR-10/-100 dataset following the (Zhou et al. 2020; Yang and Xu 2020), which consists of approximately 10K∼13K training samples and 10K test images. The controllable degrees of imbalanced ratio $\mathbf{IR} = \frac{N_{max}}{N_{min}}$ controls the distribution of training sets, where N is the number of samples in each category. In addition, the validation dataset is also set according to the same ratio as the training dataset, meaning the distribution of the training dataset represents the distribution of the real world.

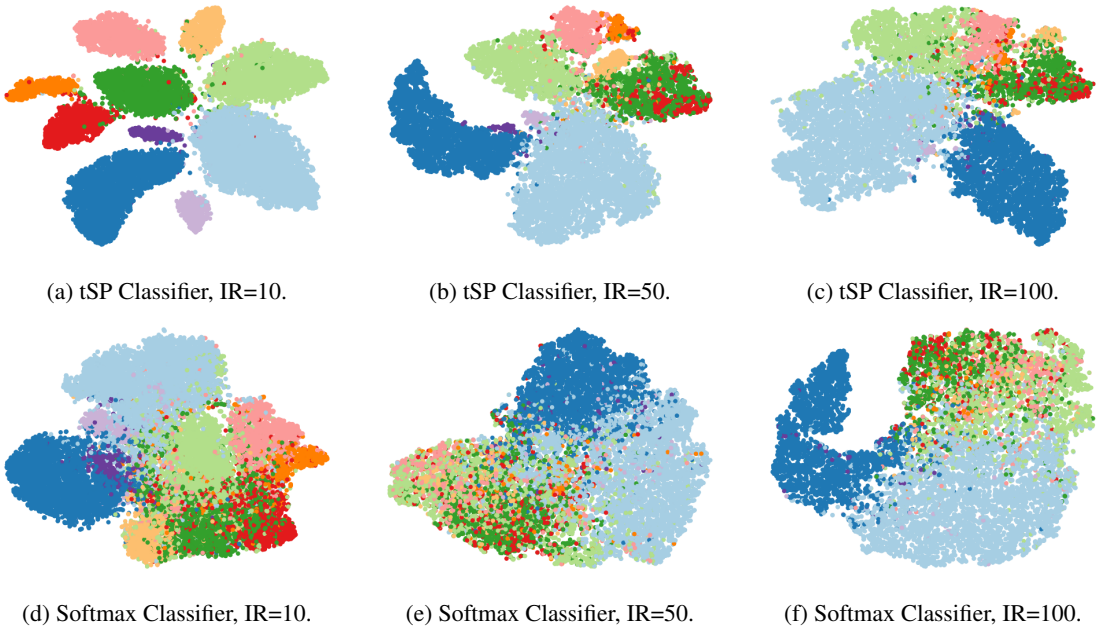
2. ImageNet-LT (Liu et al. 2019): ImageNet-LT is a long-tailed subset from the ImageNet-2012 dataset (Russakovsky et al. 2015), which consists of 1000 classes and 115.8K samples. The imbalanced ratio $\mathbf{IR} = 256$, that the category with the largest number has 1280 images and the smallest category has only 5 images in the training dataset. Besides, different from the test and validation dataset in Long-tailed CIFAR-10/-100, each category in the test and validation of ImageNet-LT contains the same number of images, which is 50 and 20 respectively. It means that the long-tailed distribution of the training dataset is caused by sampling rather than its own characteristic. The distribution of the real world is not long-tailed.

3. iNaturalist (Cui et al. 2018): iNaturalist is a real-world large-scale dataset for species recognition of animals and plants. We verify our methods on the iNaturalist 2018 to demonstrate the feasibility of our method. iNaturalist 2018 contains 438K images for over 8K categories with extreme imbalance $\mathbf{IR} = 512$ and challenging fine-grained problems.

Top-1 accuracy is used to evaluate the performance of our model. Especially, in long-tailed recognition, the categories are divided into many (with more than 100 training samples), medium (with 20∼100 samples) and few (with less than 20 samples) splits (Cai, Wang, and Hwang 2021), which calculate Top-1 accuracy respectively. Representative methods of the current extremely imbalanced approaches are chosen for our comparison, which are Focal loss (Lin et al. 2017), CB Focal loss (Cui et al. 2019b), BBN (Zhou et al. 2020), ACE (Cai, Wang, and Hwang 2021), OLTR (Liu et al. 2019), LDAM-DRW (Cao et al. 2019), cRT (Yang and Xu 2020), LWS (Kang et al. 2019), NCM (Kang et al. 2019), τ-norm (Kang et al. 2019), Mixup (Huang, Zhang, and Zhang 2020), De-confound-TDE (Tang, Huang, and Zhang 2021), LADE (Hong et al. 2021), LDA (Peng et al. 2021), CAM (Zhang et al. 2021) and RIDE (Wang et al. 2020), respectively. And † denotes that the results are copied from the corresponding paper.

Method	Cifar100		Cifar10	
	100	50	100	50
Baseline †	38.3	42.1	70.4	74.8
Focal loss †	38.4	44.3	70.4	76.7
CB Focal loss †	39.6	45.3	74.6	79.3
BBN †	42.6	47.0	79.8	82.2
ACE †	49.6	51.9	81.4	84.9
OLTR †	41.2	-	-	-
LDAM-DRW †	43.4	47.1	77.8	82.1
cRT †	45.1	50.9	79.1	84.2
LWS †	44.2	50.7	76.3	82.6
Mixup †	39.5	45.0	73.1	77.8
De-confound-TDE †	44.1	50.3	80.6	83.6
LADE †	45.4	50.5	-	-
LDA †	50.6	54.6	-	-
CAM †	47.8	51.7	80.0	83.6
Ours	**51.7**	**55.1**	**84.5**	**86.9**

Table 2: Top-1 Accuracy on Long-tailed CIFAR-10/-100.



(a) tSP Classifier, IR=10. (b) tSP Classifier, IR=50. (c) tSP Classifier, IR=100.

(d) Softmax Classifier, IR=10. (e) Softmax Classifier, IR=50. (f) Softmax Classifier, IR=100.

Figure 5: The visualization of the eigenspace of the T-distributed classifier and softmax classifier respectively, which is resulted by Long-tailed CIFAR-10 with different imbalanced ratios IR.

Ablation Experiments

We perform ablation experiments on long-tailed CIFAR-10/-100 datasets with various imbalanced ratios (100 and 50). The ResNet-32 is used as our backbone, as the same as the methods of comparison. And the network is trained by the stochastic gradient descent (SGD) with a momentum of 0.9 for 90 epochs, following the experiment setting of Deconfound-TDE(Tang, Huang, and Zhang 2021). The result is shown in Table.2. The red denotes the best result. Especially, to intuitively demonstrate the improvement by our method, there is no data augmentation and pre-trained model in our training phase, that the improvement is totally brought by our T-distributed spherical classifier. It can be seen that we have achieved superior results in the long-tailed CIFAR-10 and long-tailed CIFAR-100 datasets, showing our model can significantly improve the Top-1 accuracy in the long-tailed problems, alleviating the extreme imbalanced issue brought by the real-world sampling.

To show the eigenspace of our T-distributed spherical classifier, we use T-SNE to visualize the extracted features in Fig.5, with the comparison of the softmax classifier as the representation of the winner-take-all strategy. The results are achieved in long-tailed CIFAR-10 with difference imbalanced ratios **IR**. It can be seen that, our model significantly maintains the separation of each category with clearer decision boundaries, and promotes the intraclass compactness and interclass separability of features. Instead of the most of eigenspace occupied by the head categories in the softmax classifier, the head categories in our T-distributed spherical classifier still occupy a small eigenspace, so that the redundant eigenspace is provided to other categories, alleviating the crowding problem.

Furthermore, we measure the cosine distance of each center in our T-distributed spherical classifier in left of Fig.6, which shows the eigenspace occupied by each category. The larger the angle, the more eigenspace occupied by the category, which means less crowded. Compared with the performance of cosine metric trained on the long-tailed CIFAR-100 and the balanced CIFAR-100 in Fig.1, our model not only equalizes the eigenspace for each category, but also raises the angle to more than 100°, which means more available eigenspace for each category. Our model can make full use of the eigenspace. Besides, we visualize the categories centers of our tSP classifier and cosine classifier in right of Fig.6, which shows that our category center distribution is more sparse, alleviating crowding problems.

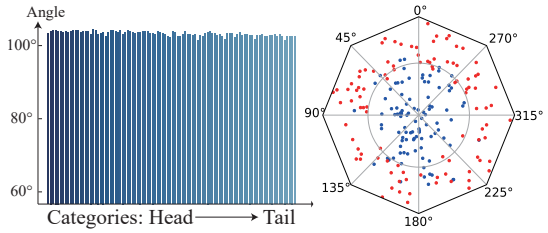


Figure 6: Left: The cosine distance of each center in the tSP classifier trained on the long-tailed CIFAR-100. Right: The visualization of the classifier centers by t-SNE on the long-tailed CIFAR-100. tSP classifier is in red, and the cosine classifier is in blue.

By equalizing the eigenspace and alleviating the crowding problems, our model can improve the accuracy of tail

categories while maintaining the accuracy of the head categories shown in Table.3. The test dataset is split into three parts: many (with more than 100 training samples), medium (with 20∼100 samples) and few (with less than 20 samples) split. It can be seen that our model achieves excellent results on the many and median splits, and significantly improve the accuracy of few split compared with the baseline.

Methods	ALL	Many	Median	Few
BaseLine †	39.1	66.1	37.3	10.6
Focal Loss †	41.2	61.8	41.4	17.6
Remix †	40.9	69.6	40.7	8.8
Mixup †	41.2	70.7	40.4	8.8
BBN †	39.4	47.2	49.4	19.8
LDAM-DRW †	42.0	61.5	41.7	20.2
τ-norm †	43.2	65.7	43.6	17.3
cRT †	43.3	64.0	44.8	18.1
RIDE †	49.1	69.3	49.3	26.0
Ours	**51.7**	**70.9**	**62.6**	**23.9**

Table 3: Performances on Long-tailed CIFAR-100 with the evaluation of Many-shot, Median-shot, Few-shot and all accuracy.

Results and Discussion

In addition, our model still achieves excellent results on the large-scale extremely imbalanced dataset with various backbones, such as ImageNet-LT (Russakovsky et al. 2015) and iNaturalist (Cui et al. 2018). We use ResNet-10, ResNet-50 and ResNeXt-50 as the backbone, and the network is trained by the SGD optimizer for 300 epochs, with the batch size of 2048. We use Auto-Augmentation and class-aware sampler to perform data augmentation. The results are shown in Table. 4. The red denotes the best result.

Method	ImageNet-LT			iNat2018
	Res10	Res50	ResX50	Res50
Baseline †	20.9	41.6	44.4	61.2
Focal loss †	30.5	-	43.7	60.3
CB Focal loss †	-	-	-	61.1
BBN †	-	48.3	49.3	66.3
OLTR †	34.1	-	41.9	63.9
NCM †	35.5	44.3	47.3	-
LDAM-DRW †	36.0	-	-	68.0
cRT †	41.8	47.3	49.6	65.2
τ-norm †	40.6	46.7	49.4	65.6
LWS †	41.4	47.7	49.9	65.9
RIDE †	-	54.4	55.9	71.4
Ours	**44.7**	**52.3**	**54.2**	**70.3**

Table 4: Top-1 Accuracy on ImageNet-LT and iNaturalist2018.

Compared with the baseline, our model improves 23.8% accuracy with the ResNet-10 in ImageNet-LT, which achieves superior top-1 accuracy. Besides, extensive experiments demonstrate the robustness and generalizability of our model, with the improvement of 10.7% under ResNet-50 of ImageNet-LT, 9.8% under ResNeXt-50 of ImageNet-LT and 9.1% under ResNeXt-50 of iNaturalist2018. By equalizing the eigenspace, our model can largely alleviate the crowding problem and boost the performance of the extremely imbalanced classification.

Case Study for Pancreatic Tumor Subtypes

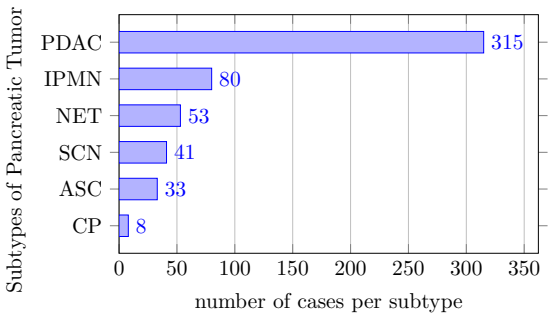


Figure 7: The number of cases in each category of the pancreatic subtype classification dataset.

Besides, we applied our method to the real clinical application of pancreatic tumor subtypes classification, which has the extremely imbalance issue. 530 patients' abdominal CT images are collected as shown in Fig.7, including six categories that doctors are most concerned about, which are pancreatic ductal adenocarcinoma (PDAC), intraductal papillary mucinous neoplasm (IPMN), pancreatic neuroendocrine tumor (NET), serous cystic neoplasm (SCN), adenosquamous carcinoma (ASC) and Chronic pancreatitis (CP). Our model achieves the Top-1 accuracy of 69.04% and Top-5 accuracy of 83.62%. Most of the other methods crashed in training, all predicted as PDAC.

Conclusion

The extremely imbalanced problem often shows in the real-world. We analyze that current metrics with the winner-take-all strategy will lead to the crowding problem in the eigenspace, causing a huge waste of the eigenspace. Therefore, we designed the T-distributed spherical metric which has the good characteristic of high kurtosis. Based on the T-distributed spherical metric, we design the T-distributed spherical classifier embedding the neural network for the extremely imbalanced classification. We achieved superior performances on long-tailed CIFAR-10/-100 with a large imbalanced ratio. And the accuracy of both the tail and the head categories has been improved. We have also verified our model on the large-scale Imagenet-LT and the iNaturalist2018 datasets, which also achieved excellent results. The results show that our model equalizes the eigenspace for each category, alleviating crowding problems. In addition, we successfully apply our model to the classification of pancreatic tumor subtypes without crashing.

Acknowledgments

This work was supported by National Natural Science Foundation of China (Serial Nos. 62173252, 61976134), and Natural Science Foundation of Shanghai (Serial No. 21ZR1423900).

References

Banerjee, A.; Dhillon, I. S.; Ghosh, J.; Sra, S.; and Ridgeway, G. 2005. Clustering on the Unit Hypersphere Using von Mises-Fisher Distributions. *Journal of Machine Learning Research*, 6(9): 1345–1382.

Byrd, J.; and Lipton, Z. 2019. What Is the Effect of Importance Weighting in Deep Learning? In *International Conference on Machine Learning*, 872–881. PMLR.

Cai, J.; Wang, Y.; and Hwang, J.-N. 2021. Ace: Ally Complementary Experts for Solving Long-Tailed Recognition in One-Shot. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 112–121.

Cao, K.; Wei, C.; Gaidon, A.; Arechiga, N.; and Ma, T. 2019. Learning Imbalanced Datasets with Label-Distribution-Aware Margin Loss. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Cao, N. D.; and Aziz, W. 2020. The Power Spherical distribution. *CoRR*, abs/2006.04437.

Chou, H.-P.; Chang, S.-C.; Pan, J.-Y.; Wei, W.; and Juan, D.-C. 2020. Remix: Rebalanced Mixup. In Bartoli, A.; and Fusiello, A., eds., *Computer Vision – ECCV 2020 Workshops*, 95–110. Cham: Springer International Publishing. ISBN 978-3-030-65414-6.

Cui, Y.; Jia, M.; Lin, T.-Y.; Song, Y.; and Belongie, S. 2019a. Class-Balanced Loss Based on Effective Number of Samples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9268–9277.

Cui, Y.; Jia, M.; Lin, T.-Y.; Song, Y.; and Belongie, S. 2019b. Class-Balanced Loss Based on Effective Number of Samples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9268–9277.

Cui, Y.; Song, Y.; Sun, C.; Howard, A.; and Belongie, S. 2018. Large Scale Fine-Grained Categorization and Domain-Specific Transfer Learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4109–4118.

Hong, Y.; Han, S.; Choi, K.; Seo, S.; Kim, B.; and Chang, B. 2021. Disentangling Label Distribution for Long-Tailed Visual Recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6626–6636.

Huang, L.; Zhang, C.; and Zhang, H. 2020. Self-Adaptive Training: Beyond Empirical Risk Minimization. *Advances in neural information processing systems*, 33: 19365–19376.

Jamal, M. A.; Brown, M.; Yang, M.-H.; Wang, L.; and Gong, B. 2020. Rethinking Class-Balanced Methods for Long-Tailed Visual Recognition from a Domain Adaptation Perspective. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7610–7619.

Kang, B.; Xie, S.; Rohrbach, M.; Yan, Z.; Gordo, A.; Feng, J.; and Kalantidis, Y. 2019. Decoupling Representation and Classifier for Long-Tailed Recognition. In *Eighth International Conference on Learning Representations (ICLR)*.

Kobayashi, T. 2021. T-vMF Similarity For Regularizing Intra-Class Feature Distribution. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 6612–6621. Nashville, TN, USA: IEEE. ISBN 978-1-66544-509-2.

Lin, T.-Y.; Goyal, P.; Girshick, R.; He, K.; and Dollár, P. 2017. Focal Loss for Dense Object Detection. In *Proceedings of the IEEE International Conference on Computer Vision*, 2980–2988.

Liu, Z.; Miao, Z.; Zhan, X.; Wang, J.; Gong, B.; and Yu, S. X. 2019. Large-Scale Long-Tailed Recognition in an Open World. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2532–2541. Long Beach, CA, USA: IEEE. ISBN 978-1-72813-293-8.

Mardia, K. V.; and Jupp, P. E. 2000. *Directional Statistics*. Wiley Series in Probability and Statistics. Chichester ; New York: J. Wiley. ISBN 978-0-471-95333-3.

Peng, Z.; Huang, W.; Guo, Z.; Zhang, X.; Jiao, J.; and Ye, Q. 2021. Long-Tailed Distribution Adaptation. In *Proceedings of the 29th ACM International Conference on Multimedia*, 3275–3282. New York, NY, USA: Association for Computing Machinery. ISBN 978-1-4503-8651-7.

Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; and Bernstein, M. 2015. Imagenet Large Scale Visual Recognition Challenge. *International journal of computer vision*, 115(3): 211–252.

Tang, K.; Huang, J.; and Zhang, H. 2021. Long-Tailed Classification by Keeping the Good and Removing the Bad Momentum Causal Effect. In *Advances in Neural Information Processing Systems*.

Van der Maaten, L.; and Hinton, G. 2008. Visualizing Data Using T-SNE. *Journal of machine learning research*, 9(11).

Wang, X.; Lian, L.; Miao, Z.; Liu, Z.; and Yu, S. 2020. Long-Tailed Recognition by Routing Diverse Distribution-Aware Experts. In *International Conference on Learning Representations*.

Yang, Y.; and Xu, Z. 2020. Rethinking the Value of Labels for Improving Class-Imbalanced Learning. In *Advances in Neural Information Processing Systems*.

Zhang, Y.; Wei, X.-S.; Zhou, B.; and Wu, J. 2021. Bag of Tricks for Long-Tailed Visual Recognition with Deep Convolutional Neural Networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 3447–3455.

Zhong, Z.; Cui, J.; Liu, S.; and Jia, J. 2021. Improving Calibration for Long-Tailed Recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16489–16498.

Zhou, B.; Cui, Q.; Wei, X.-S.; and Chen, Z.-M. 2020. Bbn: Bilateral-Branch Network with Cumulative Learning for Long-Tailed Visual Recognition. In *Proceedings of*

the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 9719–9728.