


Rotation Regularization Without Rotation

Takumi Kobayashi^{1,2} 

¹ National Institute of Advanced Industrial Science and Technology,
Tsukuba 305-8560, Japan

² University of Tsukuba, Tsukuba 305-8577, Japan
`takumi.kobayashi@aist.go.jp`

Abstract. In various visual classification tasks, we enjoy significant performance improvement by deep convolutional neural networks (CNNs). To further boost performance, it is effective to regularize feature representation learning of CNNs such as by considering margin to improve feature distribution across classes. In this paper, we propose a regularization method based on random rotation of feature vectors. Random rotation is derived from cone representation to describe angular margin of a sample. While it induces geometric regularization to randomly rotate vectors by means of rotation matrices, we theoretically formulate the regularization in a statistical form which excludes costly geometric rotation as well as effectively imposes rotation-based regularization on classification in training CNNs. In the experiments on classification tasks, the method is thoroughly evaluated from various aspects, while producing favorable performance compared to the other regularization methods. Codes are available at <https://github.com/tk1980/StatRot>.

Keywords: Regularization, random rotation, CNN, classification

1 Introduction

The last decade has witnessed great success of deep convolutional neural networks (CNNs) in computer vision fields [14,31]. For training deep models equipped with huge amount of parameters, regularization methods effectively work to remedy such as over fitting on scarce training data. Those models are also regularized so as to improve feature representation even on biased learning scenarios [4].

Regularization in training CNNs is roughly categorized into two groups. One is for input signals, i.e., images in computer vision. Injecting perturbation into images increases robustness against such as image noises, object deformation and occlusion [9,40]. While regularization on input signals is designed on the basis of prior knowledge about the input patterns, the other type of regularization is rather generally applicable to feature representation (neuron activations) and weights in deep models. The most common approach is to impose L_2 -norm regularization on weights in the form of weight decay in optimizers [23]. Features composed of neuron activations are subject to normalization such as BatchNorm [20] and its variants [1,33,35] for stabilizing the training process. It is also effective to inject perturbation into features as in input regularization

such as by randomly adding noises [13] and masking some feature components via DropOut [34] in a *stochastic* manner. From the geometrical viewpoint in a feature space, there are regularization approaches to improve feature representation by addressing intra-class compactness and between-class separability. While center loss [38] is simply formulated to reduce within-class variance, the large-margin approaches [8,36] provide effective feature representation through introducing margin into a classification loss. The margin-based losses focus on *angular* margin between an input feature vector and its target classifier vector to enhance margin from the classification boundary.

In this work, we integrate the stochastic and margin-based approaches by means of *rotation*. Rotating a feature vector in any random directions by an angle α is naturally derived from a cone centered on the feature vector (Fig. 1ab). From the geometrical viewpoint, the cone exhibits an angular margin of angle α around the feature vector and thus classification of the cones contributes to improving classification margin. Though the random rotation applies stochastic perturbation across feature components [17], geometric operation of rotating vector in higher dimensional feature space demands considerable amount of computation cost. Thus, we theoretically propose an efficient formulation through reparameterization of geometric rotation. The proposed method imposes rotation regularization on classification as in geometric random rotation while excluding costly geometric operation of rotation to significantly reduce the computation cost. Our contributions are summarized as follows.

- Through analyzing geometric random rotation of feature vectors, we theoretically formulate an efficient rotation-based regularization *without geometric rotation* for improving feature representation from the viewpoint of margin.
- The proposed method works with a low computation cost to regularize classifier logits by using a parameter of rotation angle which is so interpretable as to be set in advance based on general geometric characteristics of classifiers.
- We thoroughly analyze the proposed method through empirical evaluation from various aspects as well as performance comparison to the other regularization methods on various visual classification tasks.

2 Related Works

This paper addresses regularization on feature vectors in linear classification; those features are produced at the penultimate layer of CNN models which is followed by the (fully-connected) linear classifier. This section briefly reviews related regularization approaches which cope with the feature vectors.

Stochastic regularization. DropOut [34] is a representative stochastic regularization to randomly mask (drop) feature components for increasing generalization performance via preventing co-adaptation; it is applicable not only to neuron activations at intermediate layers but also the final feature representation at the penultimate layer [25]. The DropOut has some variants to target such as feature maps [11] and residual paths [18]. In contrast to the component-wise

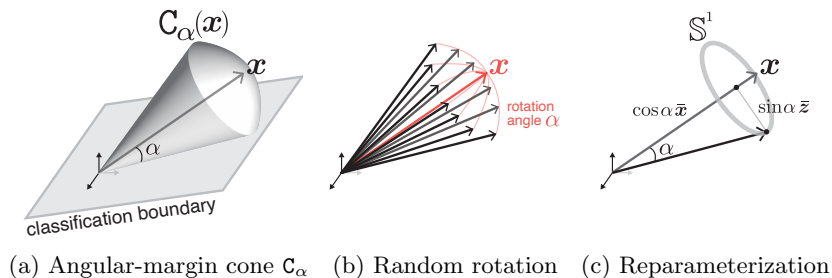


Fig. 1. Angular-margin cone of a vector $\mathbf{x} \in \mathbb{R}^3$ (a) induces random rotation by an angle α (b) and is also reparameterized by using $\bar{\mathbf{z}} \in \mathbb{S}^1$ (circle) (c).

perturbation, our angular perturbation is applied to a whole feature vector via rotation and the stochasticity is derived from random directions of the rotation. Thus, the method is orthogonal to DropOut and their combination could further regularize networks from two distinctive aspects.

Margin-based regularization. In the process of classifying feature vectors, margin-based losses are constructed from a geometrical viewpoint by underestimating (degrading) the angle between the feature vector and the target classifier vector of the assigned class [27,36,8,26]. The regularization further encourages feature vectors to be close to the target classifier, which is also contributive to alleviate saturation of softmax loss [5]. While the margin-based methods pay special attention to the target class by leaving the others untouched, the proposed regularization works symmetrically across classes without requiring class labels assigned to samples. Meanwhile, our statistical formulation of the regularization includes slight connection to the margin-based loss, especially noisy softmax [5] which stochastically degrades the target angle in a large-margin framework.

Rotation. In the literature of computer vision, some works have paid attention to rotation of input *images* from theoretical viewpoints [6,24]. In contrast to 2D image rotation, we focus on rotating feature *vectors* of higher dimension. Rotating features has been addressed in a learning framework [3,17], though rotation is also utilized to analyze interpretability of CNN features [2]. In ensemble learning, rotation is effectively applied to construct diverse base learners of decision trees [3]. RotationOut [17] randomly injects rotation variation into feature representation via sparse rotation matrices in a manner similar to DropOut [34]. Though it is based on the regularization of random rotation similarly to ours, we analyze dense random rotation to theoretically formulate the proposed regularization in a clearly different form than [17].

3 Method

We first define a rotation regularization from a geometrical viewpoint by showing its connection to angular margin of feature representation. Then, the geometric

formulation of the rotation regularization is theoretically relaxed via statistical analysis into the efficient one *without* geometrically rotating vectors.

3.1 Geometrical rotation regularization

Suppose an input feature vector $\mathbf{x} \in \mathbb{R}^d$ is classified into C classes by linear classifiers, each of which is equipped with a weight vector $\mathbf{w}_c \in \mathbb{R}^d$, $c \in \{1, \dots, C\}$. The feature vector is produced by a backbone CNN ϕ from an input image \mathcal{I} as $\mathbf{x} = \phi(\mathcal{I}; \Theta)$ and thereby the CNN parameters Θ are optimized to provide favorable feature representation. Effective feature \mathbf{x} should be compactly distributed within a class while being separable across classes, which is encouraged by introducing classification margin [8,36]. We introduce a *cone* $\mathbf{C}_\alpha(\mathbf{x})$ pointing toward \mathbf{x} which geometrically describes angular margin around \mathbf{x} :

$$\mathbf{C}_\alpha(\mathbf{x}) = \left\{ \mathbf{z} \mid \frac{\mathbf{z}^\top \mathbf{x}}{\|\mathbf{z}\|_2 \|\mathbf{x}\|_2} = \cos \alpha, \|\mathbf{z}\|_2 = \|\mathbf{x}\|_2, \mathbf{z} \in \mathbb{R}^d \right\}, \quad (1)$$

where $0 \leq \alpha < \frac{\pi}{2}$ is a half cone angle, equivalent to angular margin, as shown in Fig. 1a. Instead of a feature vector \mathbf{x} , we consider to classify a cone $\mathbf{C}_\alpha(\mathbf{x})$ as

$$\mathbf{w}_y^\top \mathbf{x} > \mathbf{w}_c^\top \mathbf{x}, \forall c \neq y \quad \Rightarrow \quad \mathbf{w}_y^\top \mathbf{z} > \mathbf{w}_c^\top \mathbf{z}, \forall c \neq y, \forall \mathbf{z} \in \mathbf{C}_\alpha(\mathbf{x}), \quad (2)$$

where y is the class label of \mathbf{x} . To correctly classify the cone $\mathbf{C}_\alpha(\mathbf{x})$, the vector \mathbf{x} is forced to exhibit angular margin larger than α from classification boundaries (Fig. 1a). The cone representation (1) is only dependent on a vector \mathbf{x} in disregard of a class label y , being contrastive with margin-based losses [8,36,5] which focuses on the angle to the target class y .

We relax the cone classification (1) by means of sampling to facilitate training. A cone $\mathbf{C}_\alpha(\mathbf{x})$ is approximated by randomly *rotating* a vector \mathbf{x} with an angle α during training since random rotation is equivalent to random sampling from the cone. Therefore, by using a random rotation matrix $\mathbf{R}_\alpha \in \mathbb{R}^{d \times d}$ of a rotation angle α , a backbone CNN Θ and a classifier $\mathbf{W} = [\mathbf{w}_1, \dots, \mathbf{w}_C]$ are optimized through minimizing the following loss:

$$\mathbb{E}_{(\mathbf{x}=\phi(\mathcal{I}), y)} \mathbb{E}_{\mathbf{R}_\alpha \in \mathcal{R}_\alpha} [\ell(\mathbf{W}^\top \mathbf{R}_\alpha \mathbf{x}, y)], \quad (3)$$

where ℓ is a classification loss of softmax cross-entropy based on *rotated* logits $\mathbf{W}^\top \mathbf{R}_\alpha \mathbf{x}$ and \mathcal{R}_α is a set of rotation matrices with rotation angle α .

Random rotation matrix \mathbf{R}_α can be computed based on Givens rotation [12]. Let random rotation orientation be described by orthonormal matrix $\mathbf{V} \in \mathbb{R}^{d \times d}$. A rotation matrix toward the orientation \mathbf{V} by an angle α is represented as

$$\mathbf{R}_\alpha = \mathbf{V} \text{blkdiag} \left[\left\{ \begin{pmatrix} \cos \alpha & -\sin \alpha \\ \sin \alpha & \cos \alpha \end{pmatrix} \right\}_{i=1}^{d/2} \right] \mathbf{V}^\top, \quad (4)$$

where **blkdiag** concatenates $d/2$ small rotation matrices of 2×2 in a block-diagonal manner³. To embed stochasticity into the rotation matrix \mathbf{R}_α , the orthonormal matrix \mathbf{V} is randomly drawn in a rather dense manner. In contrast, to

³ For odd d , we apply $(d-1)/2$ with $\mathbf{V} \in \mathbb{R}^{d \times d-1}$.

reduce the computation cost of projection via \mathbf{V} , in [17] sparse orthonormal matrix is sampled so that $\text{card}(\mathbf{V}) = d$, identical to swapping feature components, though lacking the following analysis about subspace of the classifier \mathbf{W} .

The matrix (4) rotates vectors in a (full) d -dimensional feature space, effectively working in the case that the classifier \mathbf{W} spans the full space, i.e., $\text{rank}(\mathbf{W}) = d$ requiring $d \leq C$. It, however, degrades efficacy if \mathbf{W} occupies only a subspace of d -dimensional feature space by $\text{rank}(\mathbf{W}) < d$ such as due to $d > C$. In that case, a rotation by \mathbf{R}_α would project vectors onto orthogonal space of \mathbf{W} and the orthogonal space gives no interferences in the classification (3) by \mathbf{W} . Thus, we consider an essential rotation in the space spanned by \mathbf{W} , reformulating the rotation matrix into

$$\mathbf{R}_\alpha = \mathbf{U}\tilde{\mathbf{R}}_\alpha\mathbf{U}^\top + \mathbf{U}_\perp\mathbf{U}_\perp^\top \text{ where } \tilde{\mathbf{R}}_\alpha = \tilde{\mathbf{V}} \text{blkdiag} \left[\left\{ \begin{pmatrix} \cos \alpha & -\sin \alpha \\ \sin \alpha & \cos \alpha \end{pmatrix} \right\}_{i=1}^{D/2} \right] \tilde{\mathbf{V}}^\top, \quad (5)$$

where $\mathbf{U} \in \mathbb{R}^{d \times D}$ is an orthonormal basis matrix of the subspace spanned by \mathbf{W} and $D = \text{rank}(\mathbf{W})$ is the essential dimension for rotation; $\mathbf{W} = \mathbf{U}\mathbf{U}^\top\mathbf{W}$ and $\mathbf{U}_\perp^\top\mathbf{U} = \mathbf{0}$ for $\mathbf{U}_\perp \in \mathbb{R}^{d \times d-D}$. The essential rotation matrix $\tilde{\mathbf{R}} \in \mathbb{R}^{D \times D}$ is produced by the random orthonormal matrix $\tilde{\mathbf{V}}$ in the D -dimensional space. The rotation matrix (5) is reduced into (4) in case of $D = d$; considering $D = \min(d, C)$ in most cases, it is the case of $d \leq C$.

3.2 Statistical rotation regularization

The geometric rotation (5) requires considerable amount of computation for matrix-vector multiplication $\mathbf{R}_\alpha\mathbf{x}$. Apart from such a geometrical point of view, we shed light on statistical aspect of the random rotation, leading to a novel rotation regularization formulation which excludes the geometrical rotation of vectors and thus is computationally efficient.

We consider normalized representation $\bar{\mathbf{x}} = \frac{\mathbf{x}}{\|\mathbf{x}\|_2} \in \mathbb{S}^{D-1}$ and $\bar{\mathbf{w}} = \frac{\mathbf{w}}{\|\mathbf{w}\|_2} \in \mathbb{S}^{D-1}$ of a feature vector $\mathbf{x} \in \mathbb{R}^D$ and a classifier weight vector $\mathbf{w} \in \mathbb{R}^D$ in the essential D -dimensional space since the rotation essentially affects them;

$$\mathbf{w}^\top \mathbf{R}_\alpha \mathbf{x} = \|\mathbf{w}\|_2 \|\mathbf{x}\|_2 \bar{\mathbf{w}}^\top \mathbf{R}_\alpha \bar{\mathbf{x}}. \quad (6)$$

We begin with *reparameterization* of rotating vectors as follows.

Lemma 1. *A vector $\bar{\mathbf{x}}$ is rotated by an angle α through a rotation matrix \mathbf{R}_α . As shown in Fig. 1c, so rotated vector is described by using a differential vector $\exists \bar{\mathbf{z}} \in \mathbb{S}^{D-2}$ which is in the orthogonal complement space to the input vector $\bar{\mathbf{x}}$ as*

$$\mathbf{R}_\alpha \bar{\mathbf{x}} = \cos \alpha \bar{\mathbf{x}} + \sin \alpha \bar{\mathbf{z}}, \text{ where } \|\bar{\mathbf{z}}\|_2 = 1, \bar{\mathbf{x}}^\top \bar{\mathbf{z}} = 0. \quad (7)$$

We then focus on the rotated logits $\bar{\mathbf{w}}^\top \mathbf{R}_\alpha \bar{\mathbf{x}}$ through projection by a classifier vector \mathbf{w} . The following statistical representation is useful for characterizing the projection.

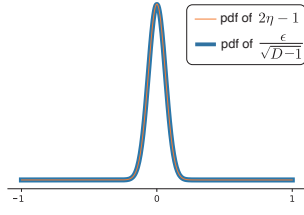


Fig. 2. PDF of Beta η and Gaussian ϵ by $D = 256$.

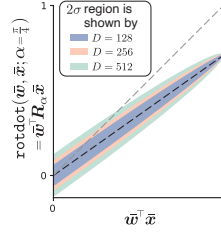


Fig. 3. Rotated logits rot_dot (13) or equivalently $\bar{\mathbf{w}}^\top \mathbf{R}_\alpha \bar{\mathbf{x}}$.

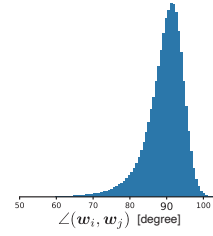


Fig. 4. Angles between classifiers in Sec. 5.1.

Lemma 2. *Projection of random vectors $\bar{\mathbf{a}}$ uniformly distributed on a unit hypersphere \mathbb{S}^{m-1} into a unit-length vector $\bar{\mathbf{b}} \in \mathbb{S}^{m-1}$ follows Beta distribution.*

$$\bar{\mathbf{a}} \in \mathbb{S}^{m-1}, \bar{\mathbf{b}} \sim \text{Unif}(\mathbb{S}^{m-1}), u = \frac{1 + \bar{\mathbf{a}}^\top \bar{\mathbf{b}}}{2} \Rightarrow u \sim \text{Beta}\left(\frac{m-1}{2}, \frac{m-1}{2}\right). \quad (8)$$

For the higher dimensional case $m \gg 1$, it approaches Gaussian distribution as

$$\bar{\mathbf{a}}^\top \bar{\mathbf{b}} \sim \mathcal{N}\left(0, \frac{1}{\sqrt{m}}\right). \quad (9)$$

We apply Lemma 1&2 to the rotated logit $\bar{\mathbf{w}} \mathbf{R}_\alpha \bar{\mathbf{x}}$ to construct the following statistical representation.

Theorem 1. *Random rotation matrix \mathbf{R}_α of an angle α is applied to an inner product between two unit-length vectors $\bar{\mathbf{w}}$ and $\bar{\mathbf{x}}$ where $\bar{\mathbf{w}}^\top \bar{\mathbf{x}} = \cos \theta$. Then, the inner product is endowed with stochasticity by the random \mathbf{R}_α and is statistically described by*

$$\bar{\mathbf{w}}^\top \mathbf{R}_\alpha \bar{\mathbf{x}} = \cos \alpha \cos \theta + (2\eta - 1) \sin \alpha \sin \theta \text{ where } \eta \sim \text{Beta}\left(\frac{D-2}{2}, \frac{D-2}{2}\right). \quad (10)$$

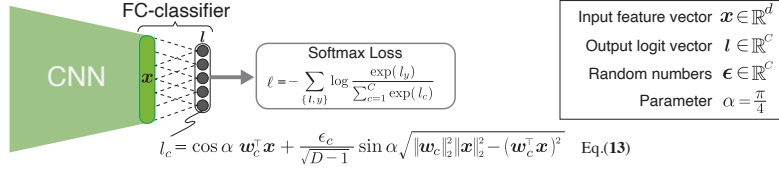
For the higher dimensional case $D \gg 1$, it approaches Gaussian distribution as

$$\bar{\mathbf{w}}^\top \mathbf{R}_\alpha \bar{\mathbf{x}} = \cos \alpha \cos \theta + \frac{\epsilon}{\sqrt{D-1}} \sin \alpha \sin \theta \text{ where } \epsilon \sim \mathcal{N}(0, 1). \quad (11)$$

It is noteworthy that the statistical representation (11) is simply computed by using a Gaussian random number ϵ without geometrically rotating vectors; the correspondence between Gaussian and Beta distributions are depicted in Fig. 2. The inner product (logit) degraded by random rotation is shown in Fig. 3. By using this efficient formulation of random rotation, an objective loss (3) can be rewritten into

$$\mathbb{E}_{(\mathbf{x}=\phi(\mathcal{I}), y)} \mathbb{E}_{\epsilon \sim \mathcal{N}(0,1)} [\ell(\{\text{rot_dot}(\mathbf{w}_c, \mathbf{x}; \epsilon_c, \alpha)\}_{c=1}^C, y)], \quad (12)$$

$$\text{rot_dot}(\mathbf{w}, \mathbf{x}; \epsilon, \alpha) = \cos \alpha \mathbf{w}^\top \mathbf{x} + \frac{\epsilon}{\sqrt{D-1}} \sin \alpha \sqrt{\|\mathbf{w}\|_2^2 \|\mathbf{x}\|_2^2 - (\mathbf{w}^\top \mathbf{x})^2}, \quad (13)$$



1. Sample C normal random numbers $\{\epsilon_c\}_{c=1}^C$.
2. Compute logits $\{l_c\}_{c=1}^C$ by (13) based on $\mathbf{w}_c^\top \mathbf{x}$ and ϵ_c with $\alpha = \frac{\pi}{4}$.
3. Feed the logits into a softmax cross-entropy loss.

Fig. 5. Computational procedure of the proposed method.

where $D = \min(d, C)$ indicates an essential dimensionality in the linear classification as described in Sec. 3.1. In (13), we assume less correlation among classifier vectors $\{\mathbf{w}_c\}_{c=1}^C$ to simply draw a random number ϵ_c in an *i.i.d.* manner. It practically holds since the learned classifier vectors are close to orthogonal as shown in Fig. 4. This orthogonality of classifiers also inspires us to set the rotation angle as half of the orthogonality, $\alpha = \frac{\pi}{4}$, so as to maximize the angular margin α within a gap ($\frac{\pi}{2}$) between classifiers. Fig. 5 shows a computational procedure of the proposed method, which first computes the logit by (13) and then feeds it into a loss function such as softmax cross-entropy loss.

4 Discussion

We analyze the proposed regularization (13) through comparing it with its variants and related margin losses [8,5]. The methods mentioned in this section are also empirically evaluated in Sec. 5.

4.1 Comparison to geometric regularization

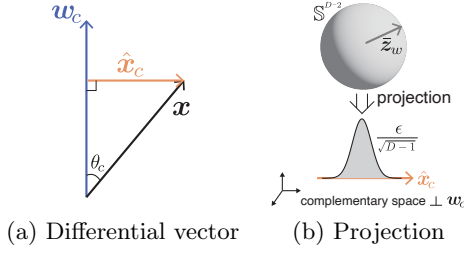
The geometrical formulation (3) is different from the statistical one (13) due to reparameterization, though both of them are derived from the regularization of randomly rotating input vector \mathbf{x} . We delve deeper into the difference by contrasting gradients of the c -th logit l_c with respect to input vector \mathbf{x} as

$$(3) \Rightarrow \frac{\partial l_c}{\partial \mathbf{x}} = \mathbf{R}_\alpha^\top \mathbf{w}_c = \|\mathbf{w}_c\|_2 [\cos \alpha \bar{\mathbf{w}}_c + \sin \alpha \bar{\mathbf{z}}_w], \quad (14)$$

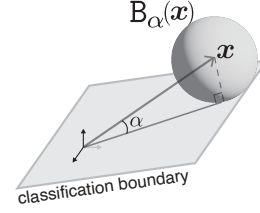
$$(13) \Rightarrow \frac{\partial l_c}{\partial \mathbf{x}} = \cos \alpha \mathbf{w}_c + \sin \alpha \frac{\epsilon}{\sqrt{D-1}} \frac{\|\mathbf{w}_c\|_2^2 \mathbf{x} - \mathbf{w}_c \mathbf{w}_c^\top \mathbf{x}}{\sqrt{\|\mathbf{w}_c\|_2^2 \|\mathbf{x}\|_2^2 - (\mathbf{w}_c^\top \mathbf{x})^2}} \quad (15)$$

$$= \|\mathbf{w}_c\|_2 \left[\cos \alpha \bar{\mathbf{w}}_c + \sin \alpha \frac{\epsilon}{\sqrt{D-1}} \frac{\hat{\mathbf{x}}_c}{\|\hat{\mathbf{x}}_c\|_2} \right], \quad (16)$$

where we apply Lemma 1 to (14) with a rotation matrix $\mathbf{R}_\alpha^\top = \mathbf{R}_{-\alpha}$ and $\hat{\mathbf{x}}_c = (\mathbf{I} - \bar{\mathbf{w}}_c \bar{\mathbf{w}}_c^\top) \mathbf{x}$ as shown in Fig. 6a. A critical difference between (14) and (16) is found in $\bar{\mathbf{z}}_w$ and $\frac{\epsilon}{\sqrt{D-1}} \frac{\hat{\mathbf{x}}_c}{\|\hat{\mathbf{x}}_c\|_2}$ which involve randomness of $\bar{\mathbf{z}}_w$ on \mathbb{S}^{D-2} and ϵ



(a) Differential vector (b) Projection

**Fig. 7.** Margin ball B_α in accordance with cone C_α .**Fig. 6.** Gradient of rotated logits w.r.t. \mathbf{x} is essentially on the plane of \mathbf{w}_c and a differential vector $\hat{\mathbf{x}}_c$ (a) through projecting random unit vector $\bar{\mathbf{z}}_w$ onto it (b).

from \mathcal{N} , respectively. Updating feature representation \mathbf{x} based on the logit l_c connected to the classifier \mathbf{w}_c is supposed to be essentially performed on the plane spanned by \mathbf{w}_c and \mathbf{x} , or equivalently $\hat{\mathbf{x}}_c$. It inspires us to consider the projection of $\bar{\mathbf{z}}_w$ onto $\hat{\mathbf{x}}_c$ (Fig. 6b) as

$$\frac{\hat{\mathbf{x}}_c \hat{\mathbf{x}}_c^\top \bar{\mathbf{z}}_w}{\|\hat{\mathbf{x}}_c\|_2^2} \approx \frac{\epsilon}{\sqrt{D-1}} \frac{\hat{\mathbf{x}}_c}{\|\hat{\mathbf{x}}_c\|_2}, \quad (17)$$

where we apply Lemma 2 to $\bar{\mathbf{z}}_w$ uniformly drawn from \mathbb{S}^{D-2} . From this viewpoint, the statistical formulation (13) provides an effective updating on the plane of \mathbf{w}_c and $\hat{\mathbf{x}}_c$ and it corresponds to the geometric one (14) when random differential vector $\bar{\mathbf{z}}_w$ is projected onto the direction of $\hat{\mathbf{x}}_c$.

It is noteworthy that the statistical form (13) does not require explicit projection onto the classifier subspace via \mathbf{U} in (5) but implicitly controls it by the dimensionality D . In other words, it might be possible to regard D as a tunable *parameter* for virtually exploiting the more essential feature dimensionality in the classification; such an approach is empirically evaluated in Sec. 5.1.

4.2 Correlated stochasticity

In (13), we apply *i.i.d.* random number $\epsilon \sim \mathcal{N}(0, 1)$ on the assumption of less correlation among classifiers which is practically plausible (Fig. 4). Meanwhile, to take into account the correlation among \mathbf{W} , we can explicitly draw random unit vector $\bar{\mathbf{z}}$ in (7) to modify the regularization into

$$\text{rot}_{\text{corr}}(\mathbf{w}, \mathbf{x}; \bar{\mathbf{z}}, \alpha) = \cos \alpha \mathbf{w}^\top \mathbf{x} + \frac{\hat{\mathbf{w}}^\top \bar{\mathbf{z}}}{\|\hat{\mathbf{w}}\|_2} \sin \alpha \sqrt{\|\mathbf{w}\|_2^2 \|\mathbf{x}\|_2^2 - (\mathbf{w}^\top \mathbf{x})^2}, \quad (18)$$

where $\hat{\mathbf{w}} = (\mathbf{I} - \bar{\mathbf{x}} \bar{\mathbf{x}}^\top) \mathbf{w}$ and a random vector $\bar{\mathbf{z}}$ satisfies $\|\bar{\mathbf{z}}\|_2 = 1$ and $\bar{\mathbf{x}}^\top \bar{\mathbf{z}} = 0$.

4.3 Margin ball

Sec. 3.1 introduces a *cone* $C_\alpha(\mathbf{x})$ to assign an angular margin with each sample \mathbf{x} . Sample-wise margin can be embedded by a *ball* $B_\alpha(\mathbf{x}) = \{\mathbf{z} \mid \|\mathbf{z} - \mathbf{x}\|_2 = \sin \alpha \|\mathbf{x}\|_2\}$

as shown in Fig. 7. Similarly to Sec. 3, classifying the ball leads to the following regularization on logits through reparameterization of the ball.

$$\mathbf{z} = \mathbf{x} + \sin \alpha \|\mathbf{x}\|_2 \bar{\mathbf{z}} \in \mathbb{B}_\alpha \text{ where } \bar{\mathbf{z}} \in \mathbb{S}^{D-1}, \quad (19)$$

$$\mathbf{w}^\top \mathbf{z} = \mathbf{w}^\top \mathbf{x} + \sin \alpha \|\mathbf{x}\|_2 (\mathbf{w}^\top \bar{\mathbf{z}}) = \mathbf{w}^\top \mathbf{x} + \frac{\epsilon}{\sqrt{D}} \sin \alpha \|\mathbf{x}\|_2 \|\mathbf{w}\|_2, \quad (20)$$

where we apply Lemma 2 to $\mathbf{w}^\top \bar{\mathbf{z}}$ and $\epsilon \sim \mathcal{N}(0, 1)$. From geometrical viewpoint, a ball \mathbb{B}_α contains perturbation that exhibits $\angle(\mathbf{z}, \mathbf{x}) < \alpha$ while a cone \mathbb{C}_α strictly imposes $\angle(\mathbf{z}, \mathbf{x}) = \alpha$, which implies that balls provide modest regularization than cones in terms of angular margin. From the arithmetic viewpoint, the stochastic term in (20) is simply composed of norms $\|\mathbf{w}\|_2$ and $\|\mathbf{x}\|_2$ in contrast to (13) containing correlation $\mathbf{w}^\top \mathbf{x}$ similarly to margin-based losses as discussed next.

4.4 Comparison to margin-based losses

The statistical form (13) is rewritten in

$$\text{rot\dot{dot}}(\mathbf{w}_c, \mathbf{x}; \epsilon_c, \alpha) = \|\mathbf{w}_c\|_2 \|\mathbf{x}\|_2 \left(\cos \alpha \cos \theta_c + \frac{\epsilon_c}{\sqrt{D-1}} \sin \alpha \sin \theta_c \right), \quad (21)$$

where $\mathbf{w}_c^\top \mathbf{x} = \|\mathbf{w}_c\|_2 \|\mathbf{x}\|_2 \cos \theta_c$ and $c \in \{1, \dots, C\}$. Thus, $\text{rot\dot{dot}}(\mathbf{w}_y, \mathbf{x}; \epsilon_y, \alpha)$ is reduced to $\cos(\theta_y + \alpha)$ when $\epsilon_y = -\sqrt{D-1}$ which is the degraded logit in the margin-based loss [8] underestimating the angle to the target class y . The margin-based loss computes the degraded logits in a deterministic way with a margin parameter, while our regularization works in a stochastic fashion.

Noisy softmax loss [5] introduces stochasticity into the margin-based loss by

$$\text{noisy\dot{dot}}(\mathbf{w}_y, \mathbf{x}; \epsilon, \gamma) = \|\mathbf{w}_y\|_2 \|\mathbf{x}\|_2 \{\cos \theta_y - \gamma |\epsilon| (1 - \cos \theta_y)\}, \quad (22)$$

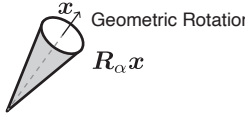
where γ is a scale parameter and $\epsilon \sim \mathcal{N}(0, 1)$. It is similar to our logit (21) as

$$\frac{1}{\cos \alpha} \text{rot\dot{dot}}(\mathbf{w}_y, \mathbf{x}; \epsilon_y, \alpha) = \|\mathbf{w}_y\|_2 \|\mathbf{x}\|_2 \left\{ \cos \theta_y + \frac{\tan \alpha}{\sqrt{D-1}} \epsilon_y (1 - \cos^2 \theta_y)^{\frac{1}{2}} \right\}. \quad (23)$$

Their differences are as follows. (i) Our formulation gives theoretically clear interpretation to the scaling factor regarding $\alpha = \frac{\pi}{4}$ and $D = \min(d, C)$ while the hyper parameter γ in (22) is heuristically determined. (ii) Noisy softmax assigns positive random number $|\epsilon|$ in an ad-hoc way while a random number ϵ in our regularization is derived from random rotation in a theoretical manner.

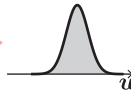
It should be noted that our regularization works on any classes symmetrically while the margin-based losses [8,36,5] touch only the target logit of class y in an asymmetric way. Such a difference might also motivate us to modify the target random number ϵ_y into $-|\epsilon_y|$ as in (22) to follow the asymmetric approach toward larger margin.

Table 1. Ablation study of the proposed statistical rotation regularization (13) on IMAGENET-LT dataset of long-tailed recognition using ResNet10.



Geometric Rotation

$R_\alpha \mathbf{x}$



Statistical Rotation

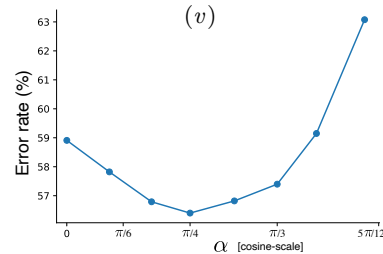
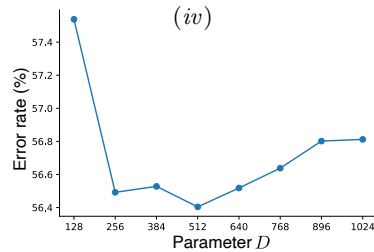
$\cos \alpha \mathbf{w}^\top \mathbf{x} + \sin \alpha \frac{\epsilon}{\sqrt{D-1}} \sqrt{\|\mathbf{w}\| \|\mathbf{x}\| - (\mathbf{w}^\top \mathbf{x})^2}$

(ii) ϵ

(iii) \mathbf{w}

(iv) $\sqrt{D-1}$

Method	Err. (%)	Method	Err. (%)
Baseline (softmax loss)	58.91	<i>vi) Variants</i>	
Statistical Rot (13)	56.40	Margin ball	56.94
<i>i)</i> Geometric Rot (3)	56.68	Asymmetric $\epsilon_y = - \epsilon $	56.45
<i>ii)</i> Correlated ϵ (18)	56.41	Combination with Dropout	56.12
<i>iii)</i> Scaled logit $\cos \alpha \mathbf{w}^\top \mathbf{x}$	58.25		



5 Experimental Results

The proposed method is applied to regularize training CNN models on classification tasks of long-tailed recognition, transfer learning and person reidentification, which follows ablation study to analyze the method in detail.

5.1 Ablation study

The proposed rotation regularization is analyzed on long-tailed recognition [21] by applying ResNet-10 [14] to IMAGENET-LT [28]; the detail of training protocol is shown in Sec. 5.2. We can analyze the method (13) with $\alpha = \frac{\pi}{4}$ from various aspects outlined in Table 1 following the discussion in Sec. 4.

i) Rotation formulation. Sec. 3 presents two formulations from geometrical (3) and statistical (13) viewpoints, performances of which are compared in Table 1*i*. Both approaches improve performance of baseline using softmax loss and particularly, the statistical formulation outperforms the geometrical one, implying that the effective updating discussed in Sec. 4.1 works in training. It is noteworthy that the statistical approach is computationally efficient without applying rotation matrix; rotating logits (13) requires $O(Cd)$ while the geometrical rotation (3) performs in $O(Cd + d^3)$. Actually, our statistical rotation requires 0.72 ms while the geometrical one takes 160 ms for batch size of 256 on TitanV GPU.

ii) Correlation among classifiers. Stochasticity ϵ in the statistical formulation (13) is built upon a simple assumption that classifiers \mathbf{W} are less correlated. It is contrasted with the correlated form (18) which directly considers random unit vector. Those two approaches are different only in terms of stochasticity and are compared in Table 1ii without showing performance difference. This result validates our assumption about the correlation from the performance viewpoint, which is also plausible according to the observation of $\angle(\mathbf{w}_i, \mathbf{w}_j)$ in Fig. 4.

iii) Scaled logits. The formulation (13) is composed of two parts. One is a deterministic scaling of logit via $\cos \alpha \mathbf{w}_c^\top \mathbf{x}$ and the other gives stochasticity derived from random rotation via a random number ϵ . The former scaling term can be regarded as changing temperature of softmax loss by $\cos \alpha$. To disentangle its effect, Table 1iii shows performance of the scaled logits $\cos \alpha \mathbf{w}_c^\top \mathbf{x}$ which is significantly inferior to that of the proposed method, being close to the baseline performance. Therefore, the performance improvement of our method is actually brought by the random rotation regularization beyond a trivial logit scaling.

iv) Stochasticity scale $\frac{1}{\sqrt{D-1}}$. Given an angle $\alpha = \frac{\pi}{4}$, scale of the stochastic term is theoretically determined by $\frac{1}{\sqrt{D-1}}$ based on random rotation; $D = \min(d, C) = 512$ in this experiment⁴. To validate this theoretical scale, we regard D as a tunable *hyper-parameter* and evaluate performances across various D as shown in Table 1iv. Increasing D means that random rotation is performed in a higher dimensional feature space by padding redundant features. Decreasing D indicates that the classifiers are shrunk into a lower dimensional subspace, reducing $\text{rank}(\mathbf{W})$. In Table 1iv, one can see effectiveness of the theoretical $D = \min(d, C) = 512$ improving performance. On the other hand, smaller $D = 128$ imposes an impractical assumption that the classifiers can be packed into a fewer dimensional subspace, thereby degrading performance.

v) Rotation angle α . While the angle $\alpha = \frac{\pi}{4}$ is determined based on the geometrical analysis of classifier orthogonality, we evaluate performances over various $0 < \alpha < \frac{\pi}{2}$. The smaller α works as weak regularization while the larger α highly regularize training. The favorable performance is found at $\alpha = \frac{\pi}{4}$ which is half of orthogonality of classifiers, while the larger α significantly degrades performance. The setting of $\alpha = \frac{\pi}{4}$ is applied to the other experiments as well.

vi) Variants. Some variants of the method are conceivable as discussed in Sec. 4. In contrast to the angular margin *cone* \mathcal{C}_α (Fig. 1a), the margin *ball* (Fig. 7) could endow regularization regarding Euclidean margin as described in Sec. 4.3. Table 1vi shows that the ball approach is inferior to our cone-based method. A margin ball contains not only perturbation affecting angles but also variations of norm $\|\mathbf{x}\|_2$ irrelevant to angular margin.

Our symmetric formulation can also be transformed into an asymmetric one in a manner similar to margin-based losses [8,36,5] by touching only the target logit via $\epsilon_y \rightarrow -|\epsilon_y|$ as in (22). The performance result in Table 1vi shows that the asymmetric approach is competitive with the simple symmetric one (13).

⁴ ResNet10 produces $d = 512$ -dimensional features for $C = 1000$ IMAGENET classes.

Thus, the symmetric regularization form derived from random rotation even works well without requiring label information.

Our simple regularization is compatible with the others such as DropOut [34] to impose regularization from various aspects, further improving performance as shown in Table 1*vi*. For fair comparison in the following experiments, however, we apply only the proposed method (13) without such a combination technique.

5.2 Performance comparison

We evaluate the proposed method in comparison with the other regularization methods. On the basis of the baseline softmax cross-entropy loss, we apply DropOut [34] as stochastic regularization and CosFace [36], ArcFace [8] and NoisySoftmax [5] for large-margin losses; their parameter settings are shown in the supplementary material. The proposed statistical rotation (StatRot) (13) is also compared to the geometrical rotation regularization which is formulated by means of geometrical rotation (GeoRot) using a random orthonormal matrix (3) and a sparse matrix [17]; they are equipped with $\alpha = \frac{\pi}{4}$ for fair comparison.

Long-tailed recognition. In a real-world scenario, the number of available samples per category is occasionally biased across class categories to form a long-tailed distribution, in contrast to the standard benchmark datasets composed of well balanced number of training samples. The imbalanced training dataset biases CNNs toward majority classes through disregarding minority ones. To cope with the imbalance issue, we follow the two-stage training procedure [21] which first learns feature representation in a standard training protocol and then finetunes only the linear classifier by balanced batch sampling while freezing the backbone feature extractor. Regularization methods are compatible with the first-stage training to improve feature representation.

We evaluate the methods on IMAGENET-LT dataset [28] using ResNet10 [14], iNaturalist2018 (*i*NAT2018) [19] using ResNet50 and PLACES-LT [28] using ResNet10 and ResNet152. While IMAGENET-LT and PLACES-LT are artificially constructed from large-scale IMAGENET [7] and PLACES365 [42], respectively, *i*NAT2018 is a real-world long-tailed dataset. At the first-stage learning, ResNet10 and ResNet50 are trained from random initial weights by SGD optimizer with momentum 0.9, weight decay 10^{-4} and cosine-scheduled learning rates starting from 0.2 over 180 epochs; the second-stage training is similarly performed over 30 epochs. In PLACES-LT, we apply ResNet152 pre-trained on IMAGENET and then trained it at the first stage over 30 epochs by SGD with cosine-scheduled learning rate starting from 0.1 on a linear classifier and 0.001 on the backbone ResNet152; the second-stage training takes 10 epochs. The performance results are shown in Table 2 demonstrating the effectiveness of the proposed statistical rotation regularization compared to the other approaches. It also works for training the pre-trained ResNet152 on PLACES-LT.

Transfer learning. The methods are then evaluated on transfer learning. Deeper CNN models pretrained on a large-scale dataset are transferable to downstream tasks which are equipped with limited amount of training samples. The

Table 2. Performance results (error rates %) on long-tailed recognition in the two-stage learning framework [21].

	IMAGENET-LT [28]	<i>i</i> NAT2018 [19]	PLACES-LT [28]	
	ResNet10 [14]	ResNet50	ResNet10	ResNet152
SoftmaxLoss [21]	58.91	32.82	72.99	61.14
ArcFace [8]	59.68	32.79	75.40	60.21
CosFace [36]	59.40	32.59	75.59	60.23
NoisySoftmax [5]	57.42	34.41	72.92	70.57
DropOut [34]	56.82	31.05	72.30	60.98
GeoRot <i>sparse</i> [17]	56.53	30.98	71.96	60.41
GeoRot <i>dense</i> (3)	56.68	30.92	71.89	60.09
StatRot (13)	56.40	30.39	71.85	59.93

Table 3. Performance results (error rates %) on transfer learning by applying RegNetY-32gf [31] pre-trained on IMAGENET. The last column shows performance gain compared to the baseline.

	CUB [37]	AIRCRAFT [30]	CAR [22]	SUN [39]	C101 [10]	Avg.Gain
SoftmaxLoss	16.05	17.61	11.07	33.31	5.91	–
ArcFace [8]	14.67	19.78	12.00	33.86	4.64	-0.20
CosFace [36]	14.46	19.89	12.32	33.87	4.75	-0.27
NoisySoftmax [5]	14.77	22.89	10.43	34.36	5.65	-0.83
DropOut [34]	15.52	17.16	10.72	33.09	5.76	0.34
GeoRot <i>sparse</i> [17]	15.22	16.20	10.56	32.51	5.56	0.78
GeoRot <i>dense</i> (3)	14.00	17.79	9.87	32.28	5.15	0.97
StatRot (13)	14.22	15.96	9.76	32.23	4.66	1.42

regularization methods contribute to exploit the discriminative power of the deeper models even on those scarce training data.

We finetune RegNetY-32gf [31] pretrained on IMAGENET dataset by means of SGD with 0.9 momentum and 10^{-4} weight decay over 60 epochs with 128 batch size by cosine-scheduled learning rate; the initial learning rates are 0.1 for linear classifiers and 0.001 for backbone CNN models. Table 3 shows performance results on various downstream classification tasks, CUB200 [37], AIRCRAFT100 [30], CAR196 [22], SUN397 [39] and CALTECH101 [10]. The deeper models are stably finetuned by the proposed method to improve performance.

Person reidentification. We finally apply the methods to regularize feature representation learning on person re-identification. The task demands CNN backbones to capture effective features from diverse camera images so that identical person images are matched across multiple cameras. We follow the baseline procedure [29] integrating three types of losses, triple loss [15], softmax loss and

Table 4. Performance results (accuracy %) on person re-identification.

Method	MARKET1501 [41]				DUKEMTMC [32]			
	ResNet50 [14]		SE-ResNeXt50 [16]		ResNet50		SE-ResNeXt50	
	Rank1	mAP	Rank1	mAP	Rank1	mAP	Rank1	mAP
SoftmaxLoss [29]	94.1	85.7	94.9	87.8	86.2	75.9	88.7	78.7
CenterLoss [29]	94.5	85.9	94.7	87.7	86.4	76.4	88.8	78.9
ArcFace [8]	89.2	72.4	82.2	67.6	79.6	60.6	72.9	53.5
CosFace [8]	85.2	72.4	85.0	71.0	79.5	62.0	76.2	56.9
NoisySoftmax [5]	94.7	87.1	94.9	87.7	86.7	75.3	88.0	76.9
DropOut [34]	93.8	85.3	94.8	87.7	86.0	75.9	88.5	78.8
GeoRot sparse [17]	94.1	85.5	95.2	87.6	86.7	76.1	89.1	78.5
GeoRot dense (3)	93.8	85.9	95.1	88.5	87.0	76.7	88.7	79.5
StatRot (13)	94.8	87.2	95.6	89.1	87.9	77.8	89.9	80.2

center loss [38], in which the regularization method is applicable to replace the center loss while keeping the other modules and training protocols the same.

The CNN backbones of ResNet50 [14] and SE-ResNeXt50 [16] pretrained on IMAGENET dataset [7] are applied to extract features from 128×256 bounding-box images. We evaluate performance by rank-1 accuracy (Rank1) and mean average precision (mAP) [41] on MARKET1501 [41] and DUKEMTMC [32] datasets as shown in Table 4. The proposed method effectively improves performance on both metrics of Rank1 and mAP which comprehensively evaluate matching performance, i.e., feature representation.

As shown in these experimental results, the proposed regularization theoretically derived from random rotation of feature vectors is stably contributive to performance improvement on various tasks outperforming the other types of regularization. Besides, it is also demonstrated that the statistical formulation effectively connects the geometric formulation with regularizing CNNs at classification in a superior manner to the naive geometric formulations.

6 Conclusion

We have proposed a regularization method based on random rotation of feature vectors. The random rotation is derived from sample-wise cone representation to geometrically embed angular margin into classification. Beyond straightforward geometric formulation to rotate vectors by random rotation matrices, we established a novel regularization formulation through theoretically analyzing the random rotation from a statistical viewpoint. It excludes laborious operation of rotating vectors as well as improves backward updating for effective training with only one hyper-parameter of a rotation angle α which can be geometrically set as $\alpha = \frac{\pi}{4}$. The experimental results on various visual classification tasks demonstrate that the method effectively contributes to performance improvement.

References

1. Ba, J.L., Kiros, J.R., Hinton, G.E.: Layer normalization. arXiv, 1607.06450 (2016)
2. Bau, D., Zhou, B., Khosla, A., Oliva, A., Torralba, A.: Network dissection: Quantifying interpretability of deep visual representations. In: CVPR (2017)
3. Blaser, R., Fryzlewicz, P.: Random rotation ensembles. *Journal of Machine Learning Research* **17**(4), 1–26 (2016)
4. Cao, K., Wei, C., Gaidon, A., Arechiga, N., Ma, T.: Learning imbalanced datasets with label-distribution-aware margin loss. In: NeurIPS (2019)
5. Chen, B., Deng, W., Du, J.: Noisy softmax: Improving the generalization ability of dcnn via postponing the early softmax saturation. In: CVPR. pp. 4021–4030 (2017)
6. Cohen, T.S., Welling, M.: Group equivariant convolutional networks. In: ICML (2016)
7. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: CVPR. pp. 248–255 (2009)
8. Deng, J., Guo, J., Niannan, X., Zafeiriou, S.: Arcface: Additive angular margin loss for deep face recognition. In: CVPR (2019)
9. DeVries, T., Taylor, G.W.: Improved regularization of convolutional neural networks with cutout. arXiv, 1708.04552 (2017)
10. Fei-Fei, L., Fergus, R., Perona, P.: Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In: Computer Vision and Pattern Recognition Workshop (2004)
11. Ghiasi, G., Lin, T.Y., Le, Q.V.: Dropblock: A regularization method for convolutional networks. In: NeurIPS. pp. 3917–3924 (2018)
12. Golub, G.H., Loan, C.F.V.: *Matrix Computations*. Johns Hopkins Univ. Press, London, 3 edn. (1996)
13. Goodfellow, I., Bengio, Y., Courville, A.: *Deep Learning*. MIT Press (2016), <http://www.deeplearningbook.org>
14. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR. pp. 770–778 (2016)
15. Hermans, A., Beyer, L., Leibe, B.: In defense of the triplet loss for person re-identification. arXiv:1703.07737 (2017)
16. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: CVPR. pp. 7132–7141 (2018)
17. Hu, K., Póczos, B.: Rotationout as a regularization method for neural network. arXiv:1911.07427 (2019)
18. Huang, G., Sun, Y., Liu, Z., Sedra, D., Weinberger, K.Q.: Deep networks with stochastic depth. In: ECCV. pp. 646–661 (2016)
19. iNaturalist: The inaturalist 2018 competition dataset. https://github.com/visipedia/inat_comp/tree/master/2018 (2018)
20. Ioffe, S., Szegedy, C.: Batch normalization: Accelerating deep network training by reducing internal covariate shift. *Journal of Machine Learning Research* **37**, 448–456 (2015)
21. Kang, B., Xie, S., Rohrbach, M., Yan, Z., Gordo, A., Feng, J., Kalantidis, Y.: Decoupling representation and classifier for long-tailed recognition. In: ICLR (2020)
22. Krause, J., Stark, M., Deng, J., Fei-Fei, L.: 3d object representations for fine-grained categorization. In: Workshop on 3D Representation and Recognition (2013)
23. Krogh, A., Hertz, J.A.: A simple weight decay can improve generalization. In: NeurIPS. pp. 950–957 (1991)

24. Lenc, K., Vedaldi, A.: Understanding image representations by measuring their equivariance and equivalence. In: CVPR. pp. 991–999 (2015)
25. Li, X., Chen, S., Hu, X., Yang, J.: Understanding the disharmony between dropout and batch normalization by variance shift. In: CVPR. pp. 2682–2690 (2019)
26. Liu, W., Wen, Y., Yu, Z., Li, M., Raj, B., Song, L.: SpheroFace: Deep hypersphere embedding for face recognition. In: CVPR. pp. 212–220 (2017)
27. Liu, W., Wen, Y., Yu, Z., Yang, M.: Large-margin softmax loss for convolutional neural networks. In: ICML. pp. 507–516 (2016)
28. Liu, Z., Miao, Z., Zhan, X., Wang, J., Gong, B., Yu, S.X.: Large-scale long-tailed recognition in an open world. In: CVPR. pp. 2537–2546 (2019)
29. Luo, H., Gu, Y., Liao, X., Lai, S., Jiang, W.: Bag of tricks and a strong baseline for deep person re-identification. In: CVPR workshop (2019)
30. Maji, S., Rahtu, E., Kannala, J., Blaschko, M.B., Vedaldi, A.: Fine-grained visual classification of aircraft. arXiv:1306.5151 (2013)
31. Radosavovic, I., Kosaraju, R.P., Girshick, R., He, K., Dollár, P.: Designing network design spaces. In: CVPR. pp. 10428–10436 (2020)
32. Ristani, E., Solera, F., Zou, R., Cucchiara, R., Tomasi, C.: Performance measures and a data set for multi-target, multi-camera tracking. In: ECCV Workshop (2016)
33. Salimans, T., Kingma, D.P.: Weight normalization: A simple reparameterization to accelerate training of deep neural networks. In: NeurIPS (2016)
34. Srivastava, N., Hinton, G.E., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.: Dropout : A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research* **15**, 1929–1958 (2014)
35. Ulyanov, D., Vedaldi, A., Lempitsky, V.: Instance normalization: The missing ingredient for fast stylization. arXiv:1607.08022 (2016)
36. Wang, H., Wang, Y., Zhou, Z., Ji, X., Gong, D., Zhou, J., Li, Z., Liu, W.: CosFace: Large margin cosine loss for deep face recognition. In: CVPR. pp. 5265–5274 (2018)
37. Welinder, P., Branson, S., Mita, T., Wah, C., Schroff, F., Belongie, S., Perona, P.: Caltech-UCSD Birds 200. Tech. Rep. CNS-TR-2010-001, California Institute of Technology (2010)
38. Wen, Y., Zhang, K., Li, Z., Qiao, Y.: A discriminative feature learning approach for deep face recognition. In: ECCV. pp. 499–515 (2016)
39. Xiao, J., Hays, J., Ehinger, K.A., Oliva, A., Torralba, A.: Sun database: Large-scale scene recognition from abbey to zoo. In: CVPR (2010)
40. Yun, S., Han, D., Oh, S.J., Chun, S., Choe, J., Yoo, Y.: Cutmix: Regularization strategy to train strong classifiers with localizable features. In: ICCV. pp. 6023–6032 (2019)
41. Zheng, L., Shen, L., Tian, L., Wang, S., Wang, J., Tian, Q.: Scalable person re-identification: A benchmark. In: ICCV (2015)
42. Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., Torralba, A.: Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **40**(6), 1452–1464 (2018)