Supplementary Material for Rotation Regularization Without Rotation

Takumi Kobayashi

takumi.kobayashi@aist.go.jp

1 Proofs

We denote vectors by bold lowercase letters, e.g., $\boldsymbol{x}$, and normalized vectors by using $\bar{}$, e.g., $\bar{\boldsymbol{x}} = \frac{\boldsymbol{x}}{\|\boldsymbol{x}\|_2}$; thus, $\boldsymbol{x} \in \mathbb{R}^D \Rightarrow \bar{\boldsymbol{x}} \in \mathbb{S}^{D-1}$, the hyper-sphere in $\mathbb{R}^D$. Matrices are denoted by bold uppercase letters, e.g., $\boldsymbol{R}$.

Lemma 1. *A vector $\bar{\boldsymbol{x}} \in \mathbb{S}^{D-1}$ is rotated by an angle α through a rotation matrix $\boldsymbol{R}_\alpha$. So rotated vector is described by using a differential vector $\exists \bar{\boldsymbol{z}} \in \mathbb{S}^{D-2}$ which is in the orthogonal complement space to the input vector $\bar{\boldsymbol{x}}$ as*

$$\boldsymbol{R}_\alpha \bar{\boldsymbol{x}} = \cos\alpha\, \bar{\boldsymbol{x}} + \sin\alpha\, \bar{\boldsymbol{z}}, \ \textit{where } \|\bar{\boldsymbol{z}}\|_2 = 1,\, \bar{\boldsymbol{x}}^\top \bar{\boldsymbol{z}} = 0. \tag{1}$$

Proof. Fig. 1 would be helpful to grasp relationships among the following vectors. Let the rotated vector be denoted by $\bar{\boldsymbol{\xi}} = \boldsymbol{R}_\alpha \bar{\boldsymbol{x}}$ which satisfies

$$\bar{\boldsymbol{x}}^\top \bar{\boldsymbol{\xi}} = \cos\alpha. \tag{2}$$

We define the differential vector $\boldsymbol{z}$ as

$$\boldsymbol{z} = \bar{\boldsymbol{\xi}} - \cos\alpha \bar{\boldsymbol{x}}, \tag{3}$$

and it has the following properties;

$$\bar{\boldsymbol{x}}^\top \boldsymbol{z} = \bar{\boldsymbol{x}}^\top \bar{\boldsymbol{\xi}} - \cos\alpha\, \|\bar{\boldsymbol{x}}\|_2^2 = 0 \tag{4}$$

$$\|\boldsymbol{z}\|_2^2 = \|\bar{\boldsymbol{\xi}}\|_2^2 + \cos^2\alpha\, \|\bar{\boldsymbol{x}}\|_2^2 - 2\cos\alpha\, \bar{\boldsymbol{x}}^\top \bar{\boldsymbol{\xi}} = 1 - \cos^2\alpha = \sin^2\alpha, \tag{5}$$

where we apply (2). Since $0 \le \alpha < \frac{\pi}{2}$, the differential vector $\boldsymbol{z}$ is thus described by using $\bar{\boldsymbol{z}} = \frac{\boldsymbol{z}}{\|\boldsymbol{z}\|_2}$ as

$$\boldsymbol{z} = \sin\alpha\, \bar{\boldsymbol{z}} \quad \text{where } \|\bar{\boldsymbol{z}}\|_2 = 1,\, \bar{\boldsymbol{x}}^\top \bar{\boldsymbol{z}} = 0. \tag{6}$$

It should be noted that $\bar{\boldsymbol{z}}$ is laid on the hyper-sphere $\mathbb{S}^{D-2}$ embedded in the $D-1$ dimensional subspace orthogonal to $\bar{\boldsymbol{x}} \in \mathbb{R}^D$. $\square$

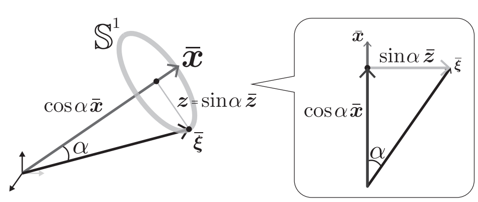


Fig. 1. Reparameterization of a rotated vector.

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

$$\bar{\boldsymbol{a}} \sim \mathit{Unif}(\mathbb{S}^{m-1}),\ \bar{\boldsymbol{b}} \in \mathbb{S}^{m-1},\ u = \frac{1+\bar{\boldsymbol{a}}^\top \bar{\boldsymbol{b}}}{2} \Rightarrow u \sim \mathit{Beta}\Big(\frac{m-1}{2}, \frac{m-1}{2}\Big). \tag{7}$$

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

$$\bar{\boldsymbol{a}}^\top \bar{\boldsymbol{b}} \sim \mathcal{N}\left(0, \frac{1}{\sqrt{m}}\right) \tag{8}$$

Proof. As shown in Fig. 2, we consider the probability density of dt region at $t = \bar{\boldsymbol{a}}^\top \bar{\boldsymbol{b}}$ along $\bar{\boldsymbol{b}}$. It is proportional to a surface volume dV on $\mathbb{S}^{m-1}$, a gray-colored belt in Fig. 2 which is composed of the length $\frac{dt}{\sqrt{1-t^2}}$ and the hyper volume of $\mathbb{S}^{m-2}$ with radius $\sqrt{1-t^2}$ and thereby computed as

$$\mathtt{p}(t)dt \propto dV \propto (\sqrt{1-t^2})^{m-2} \frac{dt}{\sqrt{1-t^2}} = (1-t^2)^{\frac{m-3}{2}} dt. \tag{9}$$

Due to $\bar{\boldsymbol{a}}^\top \bar{\boldsymbol{b}} = t = 2u - 1$, the probability density function $\mathtt{q}(u)$ is described by

$$\mathtt{q}(u)du = \mathtt{p}(t)dt \propto \left\{1 - (2u-1)^2\right\}^{\frac{m-3}{2}} \cdot 2du = 2(4u - 4u^2)^{\frac{m-3}{2}} du \tag{10}$$

$$= 2^{m-2} u^{\frac{m-3}{2}} (1-u)^{\frac{m-3}{2}} du \propto u^{\frac{m-1}{2}-1} (1-u)^{\frac{m-1}{2}-1} du, \tag{11}$$

which corresponds to the Beta distribution, $\mathit{Beta}(u; \beta, \beta) \propto u^{\beta-1}(1-u)^{\beta-1}$ with $\beta = \frac{m-1}{2}$. Thus, the mean and variance of u are

$$\mathrm{E}_{u\sim \mathit{Beta}}[u] = \frac{1}{2},\ \mathrm{Var}_{u\sim \mathit{Beta}}[u] = \frac{\beta^2}{4\beta^2(2\beta+1)} = \frac{1}{4(2\beta+1)} = \frac{1}{4m}. \tag{12}$$

Asymptotic property of the Beta distribution as $\beta \to \infty$ is

$$\mathit{Beta}(\beta, \beta) \to \mathcal{N}\left(\frac{1}{2}, \frac{1}{2\sqrt{2\beta+1}}\right), \tag{13}$$

the proof of which is found such as in [7]. It is applied to $u \sim \mathit{Beta}(\beta, \beta)$ with $\beta = \frac{m-1}{2}$ to produce

$$\bar{\boldsymbol{a}}^\top \bar{\boldsymbol{b}} = 2u - 1 \sim \mathcal{N}\left(0, \frac{1}{\sqrt{m}}\right). \qquad \square \tag{14}$$

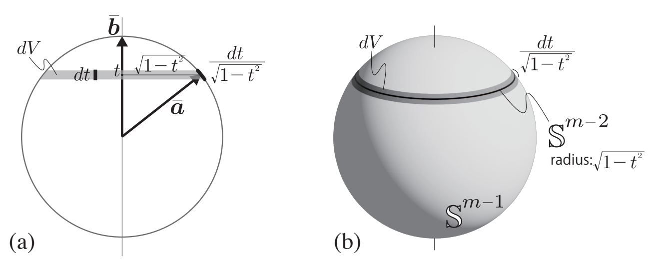


Fig. 2. Projection from a hyper sphere $\mathbb{S}^{m-1}$ into a unit-length vector $\bar{\boldsymbol{b}}$.

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

$$\bar{\boldsymbol{w}}^\top \boldsymbol{R}_\alpha \bar{\boldsymbol{x}} = \cos\alpha\cos\theta + (2\eta - 1)\sin\alpha\sin\theta \text{ where } \eta \sim Beta\Big(\frac{D-2}{2}, \frac{D-2}{2}\Big). \quad (15)$$

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

$$\bar{\boldsymbol{w}}^\top \boldsymbol{R}_\alpha \bar{\boldsymbol{x}} = \cos\alpha\cos\theta + \frac{\epsilon}{\sqrt{D-1}}\sin\alpha\sin\theta \text{ where } \epsilon \sim \mathcal{N}(0,1). \quad (16)$$

Proof. By applying Lemma 1, we can obtain

$$\bar{\boldsymbol{w}}^\top \boldsymbol{R}_\alpha \bar{\boldsymbol{x}} = \bar{\boldsymbol{w}}^\top (\cos\alpha\, \bar{\boldsymbol{x}} + \sin\alpha\, \bar{\boldsymbol{z}}) = \cos\alpha\cos\theta + \sin\alpha\, \bar{\boldsymbol{w}}^\top \bar{\boldsymbol{z}}. \quad (17)$$

Due to the randomness of the rotation matrix $\boldsymbol{R}_\alpha$, $\bar{\boldsymbol{z}}$ is uniformly drawn from $\mathbb{S}^{D-2}$ embedded in the subspace $\mathbb{R}^{D-1}$ perpendicular to $\bar{\boldsymbol{x}}$; so, $\bar{\boldsymbol{z}} = (\boldsymbol{I} - \bar{\boldsymbol{x}}\bar{\boldsymbol{x}}^\top)\bar{\boldsymbol{z}}$. In that subspace, the inner product $\bar{\boldsymbol{w}}^\top \bar{\boldsymbol{z}}$ is described by

$$\bar{\boldsymbol{w}}^\top \bar{\boldsymbol{z}} = \bar{\boldsymbol{w}}^\top (\boldsymbol{I} - \bar{\boldsymbol{x}}\bar{\boldsymbol{x}}^\top)\bar{\boldsymbol{z}} = \sin\theta\, \hat{\boldsymbol{w}}^\top \bar{\boldsymbol{z}}, \quad (18)$$

where $\hat{\boldsymbol{w}} = \frac{(\boldsymbol{I} - \bar{\boldsymbol{x}}\bar{\boldsymbol{x}}^\top)\bar{\boldsymbol{w}}}{\|(\boldsymbol{I} - \bar{\boldsymbol{x}}\bar{\boldsymbol{x}}^\top)\bar{\boldsymbol{w}}\|_2}$ is in the subspace $\mathbb{R}^{D-1}$ and $\|(\boldsymbol{I} - \bar{\boldsymbol{x}}\bar{\boldsymbol{x}}^\top)\bar{\boldsymbol{w}}\|_2 = \sin\theta$ as shown in Fig. 3. Then, based on the stochasticity $\bar{\boldsymbol{z}} \sim Unif(\mathbb{S}^{D-2})$, Lemma 2 with $m = D-1$ statistically rewrites the inner product into

$$\bar{\boldsymbol{w}}^\top \bar{\boldsymbol{z}} = \sin\theta(2\eta - 1), \quad (19)$$

where $\eta \sim Beta(\frac{D-2}{2}, \frac{D-2}{2})$. (17) and (19) lead to (15) and the asymptotic property of Beta distribution shown in Lemma 2 produces (16) with $m = D-1$. □

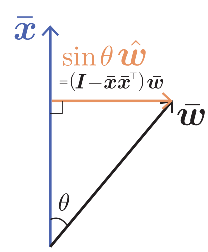


Fig. 3. Projection from $\bar{\boldsymbol{w}}$ into the subspace perpendicular to $\bar{\boldsymbol{x}}$.

2 Hyper parameters of comparison methods

In DropOut [13], we employ small probability $p = 0.2$ to mask feature elements according to the analysis [8] and preliminary experiments. For the margin-based losses, based on the preliminary experiments, we apply angular margin parameter of $m = 0.1$ in ArcFace [2], $m = 0.1$ in CosFace [14] and $\alpha = \sqrt{0.1}$ in NoisySoftmax [1].

3 Datasets

The details of the datasets that we use in Sec. 5 of the main manuscript are shown below. Except for CALTECH101 [3], we use the train/test splits provided in the respective datasets; for SUN397 [16], we use the first split out of 10 splits given in the dataset. In CALTECH101, following the standard protocol, we randomly draw 30 training samples per category and use the remaining samples as test.

(a) Long-tailed datasets

	IMAGENET-LT [9]	*i*NAT2018 [5]	PLACES-LT [9]
category	1000 objects	8142 species	365 scenes
training sample	115,846	437,513	62,500
test sample	50,000	24,426	36,500
majority : minority	1,280 : 5	1,000 : 2	4,980 : 5

(b) Downstream datasets

	CUB200 [15]	AIRCRAFT100 [11]	CAR196 [6]	SUN397 [16]	CALTECH101 [3]
category	200 birds	100 planes	196 cars	397 scenes	101 objects
trn. sample	5,994	6,667	8,144	19,850	3,030
tst. sample	5,794	3,333	8,041	19,850	5,647

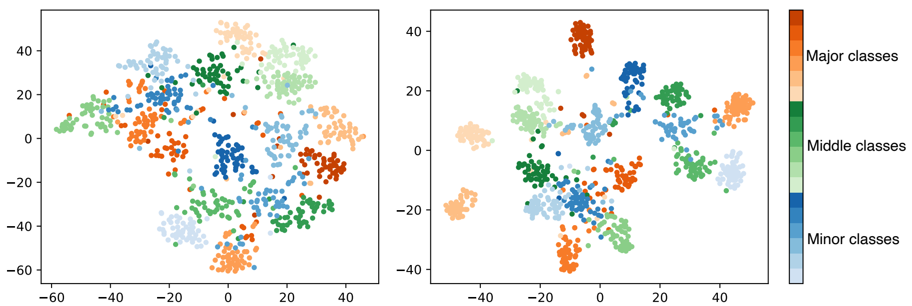
(c) Person re-identification datasets

	MARKET1501 [17]	DUKEMTMC [12]
trn. category	750 identities	702 identities
trn. sample	12,936	16,522
tst. category	751 identities	702 identities
query sample	3,368	2,228
gallery sample	19,732	17,661

4 Discriminativity of feature representation

We apply t-SNE [10] to show the feature distributions of ResNet10 backbone trained on ImageNet-LT dataset by using the baseline softmax loss and ours with statistical rotation regularization. Fig. 4 demonstrates that our method improves feature distribution in comparison to the softmax loss.

The discriminativity of features is quantitatively measured by means of discriminant score [4], the ratio of between-class feature variance to within-class one; $\mathsf{tr}(\boldsymbol{\Sigma}_B)/\mathsf{tr}(\boldsymbol{\Sigma}_W)$. Table 1 demonstrates that our method improves the score, contributing to intra-class compactness as well as inter-class separability.



(a) Softmax Loss (b) Ours

Fig. 4. Visualization of ResNet10 feature distributions on ImageNet-LT via t-SNE [10]. Each point indicates a sample drawn from major, middle and minor classes on the validation set.

Table 1. Discriminant score $\mathsf{tr}(\boldsymbol{\Sigma}_B)/\mathsf{tr}(\boldsymbol{\Sigma}_W)$. Higher is better.

	ImageNet-LT	*i*Nat2018	Places-LT	
	ResNet10	ResNet50	ResNet10	ResNet152
Softmax Loss	0.385	1.475	0.270	0.372
Ours	**0.692**	**1.747**	**0.495**	**0.672**

References

1. 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)
2. Deng, J., Guo, J., Niannan, X., Zafeiriou, S.: Arcface: Additive angular margin loss for deep face recognition. In: CVPR (2019)

3. 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)
4. Fukunaga, K.: Introduction to Statistical Pattern Recognition. Academic Press, Boston (1990)
5. iNatrualist: The inaturalist 2018 competition dataset. `https://github.com/visipedia/inat_comp/tree/master/2018` (2018)
6. Krause, J., Stark, M., Deng, J., Fei-Fei, L.: 3d object representations for fine-grained categorization. In: Workshop on 3D Representation and Recognition (2013)
7. Leemis, L.: `http://www.math.wm.edu/~leemis/chart/UDR/PDFs/BetaNormal.pdf`
8. 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)
9. 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)
10. van der Maaten, L., Hinton, G.: Visualizing data using t-SNE. Journal of Machine Learning Research **9**, 2579–2605 (2008)
11. Maji, S., Rahtu, E., Kannala, J., Blaschko, M.B., Vedaldi, A.: Fine-grained visual classification of aircraft. arXiv:1306.5151 (2013)
12. 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)
13. 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)
14. 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)
15. 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)
16. Xiao, J., Hays, J., Ehinger, K.A., Oliva, A., Torralba, A.: Sun database: Large-scale scene recognition from abbey to zoo. In: CVPR (2010)
17. Zheng, L., Shen, L., Tian, L., Wang, S., Wang, J., Tian, Q.: Scalable person re-identification: A benchmark. In: ICCV (2015)