Supplementary Material of Model-Agnostic Metric for Zero-Shot Learning

Jiayi Shen[1], Haochen Wang[1], Anran Zhang[1], Qiang Qiu[2], Xiantong Zhen[3], Xianbin Cao[1,4,5*]
[1]School of Electronic and Information Engineering, Beihang University, Beijing, China
[2]Duke University, Durham, NC, USA
[3]Inception Institute of Artificial Intelligence, Abu Dhabi, UAE
[4]Key Laboratory of Advanced Technology of Near Space Information System (Beihang University), Ministry of Industry and Information Technology of China, Beijing, China
[5]Beijing Advanced Innovation Center for Big Data-Based Precision Medicine, Beijing, China
shenjiayi@buaa.edu.cn, haochen.hobot@gmail.com, zhanganran@buaa.edu.cn, qiang.qiu@duke.edu, zhenxt@gmail.com, xbcao@buaa.edu.cn

1. Proof of Proposition 1

Proof. According to the marginal probability density of the component b_i in Theorem 1, we get the variation is

$$\mathrm{Var}\left[b_i\right] = \kappa_D \cdot \int_{-1}^{+1} {b_i}^2 \left(1 - {b_i}^2\right)^{\frac{(D-3)}{2}} \mathrm{d}b_i. \tag{1}$$

Due to the symmetry of b_i value range, the integral item can be equivalent to $2\int_0^1 {b_i}^2 \left(1 - {b_i}^2\right)^{\frac{(D-3)}{2}} \mathrm{d}b_i$. Further let the $x = {b_i}^2$, according to the recurrence property of Gamma function $\Gamma\left(x+1\right) = x\Gamma\left(x\right)$, Equation (1) can be reformulated as:

$$\begin{aligned}\mathrm{Var}\left[b_i\right] =& \kappa_D \cdot \int_0^1 x^{\frac{1}{2}} (1-x)^{\frac{(D-3)}{2}} \mathrm{d}x \\ =& \kappa_D \cdot \frac{\Gamma\left(\frac{3}{2}\right)\Gamma\left(\frac{D-1}{2}\right)}{\Gamma\left(\frac{D}{2}+1\right)} \\ =& \frac{\Gamma\left(\frac{D}{2}\right)}{\Gamma\left(\frac{1}{2}\right)\Gamma\left(\frac{D-1}{2}\right)} \cdot \frac{\Gamma\left(\frac{3}{2}\right)\Gamma\left(\frac{D-1}{2}\right)}{\Gamma\left(\frac{D}{2}+1\right)} = \frac{1}{D}.\end{aligned} \tag{2}$$

Equation (2) shows that the variance of any component b_i of the normalized embedded semantic vector decreases as the dimensionality increases on the unit sphere. □

2. Proof of Proposition 2

Proof. Euclidean distance between the normalized projected visual feature $\mathbf{a}$ and the normalized embedded semantic vector $\mathbf{b}_i (i = 1, 2)$ can be viewed as the chord length of them on the unit sphere, which is given by

$$\left\|\mathbf{a} - \mathbf{b}_i\right\|^2 = \left\|\mathbf{a}\right\|^2 + \left\|\mathbf{b}_i\right\|^2 - 2\mathbf{a}^T\mathbf{b}_i.$$

$\mathbf{a}$ and $\mathbf{b}_i$ are points on the unit sphere and the norms of them are equal to 1. The distance can be rewritten by

$$\left\|\mathbf{a} - \mathbf{b}_i\right\|^2 = 2(1 - \mathbf{a}^T\mathbf{b}_i),$$

and its expected value is

$$\mathrm{E}_{\mathcal{X}}\left[\left\|\mathbf{a} - \mathbf{b}_i\right\|^2\right] = 2(1 - \mathrm{E}_{\mathcal{X}}[\mathbf{a}]^T\mathbf{b}_i) = 2(1 - \varepsilon{\mathbf{a}^*}^T\mathbf{b}_i).$$

Substituting the expectations of the distance in the Proposition 2, we get

$$\Delta = 2\varepsilon(\cos\left(\mathbf{a}^*, \mathbf{b}_1\right) - \cos\left(\mathbf{a}^*, \mathbf{b}_2\right)) = 2\varepsilon\gamma\sigma. \tag{3}$$

Due to the normalized embedded semantic vector $\mathbf{b}$ follows a uniform distribution, there is no difference between $\mathbf{a}^*$ and $\hat{\mathbf{a}} = (..., 0, 1, 0, ...)$ to the whole normalized embedded semantic vectors while calculating the variance of $\cos(\mathbf{a}^*, \mathbf{b})$. Meanwhile, Proposition 1 proves that the variance of the component b_i is $\frac{1}{D}$. From that, we can get the variance σ^2 of $cos\left(\mathbf{a}^*, \mathbf{b}\right)$ is

$$\begin{aligned}\sigma^2 &= \mathrm{Var}_{\mathcal{S}}\left[\cos(\mathbf{a}^*, \mathbf{b})\right] = \mathrm{Var}_{\mathcal{S}}\left[\cos(\hat{\mathbf{a}}, \mathbf{b})\right] \\ &= \mathrm{Var}_{\mathcal{S}}\left[\hat{\mathbf{a}}^T\mathbf{b}\right] = \mathrm{Var}_{\mathcal{S}}\left[b_i\right] = \frac{1}{D}.\end{aligned} \tag{4}$$

From (3) and (4), we obtain $\Delta = \frac{2\varepsilon\gamma}{\sqrt{D}}$. □

3. Extend Experiments

Table 1: Accuracy(%) of our proposed method with original visual features (No PCA) and PCA-projected features.

Dataset	AWA1	AWA2	SUN	CUB	aPY
No PCA (D=2048)	70.7	65.5	60.7	52.1	37.7
PCA (D=2048)	**72.7**	**72.0**	**62.6**	**59.6**	**47.3**

*Corresponding author

Table 2: Accuracy(%) of our proposed method on AWA2 with different dimensional metric space by MLP and PCA.

Dim.	64	128	256	512	1024	2048	2560	3072	4096
MLP	46.6	56.1	60.3	61.7	65.1	66.0	65.1	64.6	64.3
PCA	**65.5**	**67.1**	**68.8**	**69.4**	**70.3**	**72.0**	**70.4**	**68.1**	**67.7**

Performance of our method using original features and learning low dimensional features using MLP are shown in Table 1&2, respectively.

Table 1 shows that PCA(D=2048)-based visual features have better performance consistently on five benchmarks than original(D=2048) visual feature. This is due to the PCA's statistical benefits. PCA decorrelates the dimensions of visual features such that embedded semantic vectors can predict these dimensions independently rather than jointly for the better discriminative ability.

Table 2 shows that the PCA-based method outperforms the MLP-based method on different-dimensional embedding space by a large margin. Compared with the non-parametric strategy (PCA), the MLP with parameters needs more training times and is more prone to over-fitting [5]. Thus, in our paper, we choose PCA as the dimensional reduction strategy.