# Supplementary Materials for the Paper: Prompt Distribution Learning

## 1. Proof of Proposition 1

**Proposition 1** *Suppose that* $\mathbf{w}_{1:C} = [\mathbf{w}_1^T, .., \mathbf{w}_C^T]^T \in \mathbb{R}^{dC}$ *follows* $\mathcal{N}(\mu_{1:C}(\mathcal{P}^K), \mathbf{\Sigma}_{1:C}(\mathcal{P}^K))$. *Let* $\mathbf{\Sigma}_{ij}(\mathcal{P}^K)$ *be the covariance matrix of* $\mathbf{w}_i$ *and* $\mathbf{w}_j$, $\mu_i(\mathcal{P}^K)$ *be the mean of* $\mathbf{w}_i$, *and* $\mathbf{A}_{i,j} = \mathbf{\Sigma}_{ii} + \mathbf{\Sigma}_{jj} - \mathbf{\Sigma}_{ij} - \mathbf{\Sigma}_{ji}$. *Then it holds that*

$$\mathcal{L}(\mathcal{P}^K) \leq \mathop{\mathbb{E}}_{\mathbf{x}_i, y_i} \left[ -\log \frac{e^{\mathbf{z}_i^T \mu_{y_i}(\mathcal{P}^K)/\tau}}{\sum_c e^{\mathbf{z}_i^T \mu_c(\mathcal{P}^K)/\tau + \mathbf{z}_i^T \mathbf{A}_{c,y_i} \mathbf{z}_i / 2\tau^2}} \right] \tag{1}$$

$$\triangleq \mathcal{L}_{upper}(\mathcal{P}^K). \tag{2}$$

**Proof:** *Given the Gaussian distribution assumption, we have:*

$$\mathcal{L}(\mathcal{P}^K) = \mathop{\mathbb{E}}_{\mathbf{x}_i, y_i} \left[ -\log \mathop{\mathbb{E}}_{\mathbf{w}_{1:C}} \frac{e^{\mathbf{z}_i^T \mathbf{w}_{y_i}/\tau}}{\sum_c e^{\mathbf{z}_i^T \mathbf{w}_c/\tau}} \right] \tag{3}$$

$$\leq \mathop{\mathbb{E}}_{\mathbf{x}_i, y_i} \left[ \mathop{\mathbb{E}}_{\mathbf{w}_{1:C}} \left[ -\log \frac{e^{\mathbf{z}_i^T \mathbf{w}_{y_i}/\tau}}{\sum_c e^{\mathbf{z}_i^T \mathbf{w}_c/\tau}} \right] \right] \tag{4}$$

$$= \mathop{\mathbb{E}}_{\mathbf{x}_i, y_i} \left[ \mathop{\mathbb{E}}_{\mathbf{w}_{1:C}} \left[ \log \sum_c e^{\mathbf{z}_i^T (\mathbf{w}_c - \mathbf{w}_{y_i})/\tau} \right] \right] \tag{5}$$

$$\leq \mathop{\mathbb{E}}_{\mathbf{x}_i, y_i} \left[ \log \sum_c \mathop{\mathbb{E}}_{\mathbf{w}_{1:C}} \left[ e^{\mathbf{z}_i^T (\mathbf{w}_c - \mathbf{w}_{y_i})/\tau} \right] \right] \tag{6}$$

$$= \mathop{\mathbb{E}}_{\mathbf{x}_i, y_i} \left[ \log \sum_c e^{\mathbf{z}_i^T (\mu_c - \mu_{y_i})/\tau + \mathbf{z}_i^T \mathbf{A}_{c,y_i} \mathbf{z}_i / 2\tau^2} \right] \tag{7}$$

$$= \mathop{\mathbb{E}}_{\mathbf{x}_i, y_i} \left[ -\log \frac{e^{\mathbf{z}_i^T \mu_{y_i}(\mathcal{P}^K)/\tau}}{\sum_c e^{\mathbf{z}_i^T \mu_c(\mathcal{P}^K)/\tau + \mathbf{z}_i^T \mathbf{A}_{c,y_i} \mathbf{z}_i / 2\tau^2}} \right], \tag{8}$$

*where Inequalities 4 and 6 are from Jensen's inequality* ($\mathbb{E}(\log X) \leq \log \mathbb{E}(X)$ *[6]). Besides, due to* $\mathbf{z}_i^T(\mathbf{w}_c - \mathbf{w}_{y_i})$ *being a Gaussian variable following* $\mathcal{N}(\mathbf{z}_i^T(\mu_c - \mu_{y_i}), \mathbf{z}_i^T \mathbf{A}_{c,y_i} \mathbf{z}_i)$, *the expectation in Equation 7 is obtained by leveraging the moment-generating function:*

$$\mathbb{E}(e^{tX}) = \mathbb{E}(e^{t\mu + \frac{1}{2}\sigma^2 t^2}), X \sim \mathcal{N}(\mu, \sigma^2). \tag{9}$$

## 2. Datasets

The details of the 12 downstream datasets are shown in Tabel 1. The accuracy metric of each dataset follows CLIP [13].

## 3. Baselines

The regularization of Linear Probe CLIP is selected by the validation set on each dataset, following the hyperparameter sweep strategy in CLIP [13]. Note that the validation set, which is used to select task-specific hyperparameters, is only used in Linear Probe CLIP.

For CoOp [14], we use the SGD optimizer with learning rate of 0.001 and the batch size of 20. The learning rate has a cosine decay schedule. The number of training epochs is 100. The prompt length is 16. Our implementation of Prompt Tuning has slightly better results than those reported in [14].

## 4. Results

Table 2 shows the detailed results of various methods with the same pre-trained CLIP model (RN50) on the 12 datasets.

## References


[1] Lukas Bossard, Matthieu Guillaumin, and Luc J. Van Gool. Food-101 – mining discriminative components with random forests. In *ECCV*, 2014. 2

[2] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *CVPR*, 2014. 2

[3] Adam Coates, Andrew Y. Ng, and Honglak Lee. An analysis of single-layer networks in unsupervised feature learning. In *AISTATS*, 2011. 2

[4] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009. 2

[5] Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *J-STARS*, 2019. 2

[6] Johan Ludwig William Valdemar Jensen. Sur les fonctions convexes et les inégalités entre les valeurs moyennes. *Acta mathematica*, 1906. 1

| Dataset | Classes | Train Size | Test Size | Accuracy metric |
|---|---|---|---|---|
| ImageNet [4] | 1000 | 1,281,167 | 50,000 | accuracy |
| CIFAR-10 [8] | 10 | 50,000 | 10,000 | accuracy |
| CIFAR-100 [8] | 100 | 50,000 | 10,000 | accuracy |
| STL-10 [3] | 10 | 1,000 | 8,000 | accuracy |
| Food-101 [1] | 101 | 75,750 | 25,250 | accuracy |
| Stanford Cars [7] | 196 | 8,144 | 8,041 | accuracy |
| FGVC Aircraft [10] | 100 | 6,667 | 3,333 | mean per-class |
| Oxford-IIIT Pets [12] | 37 | 3,680 | 3,669 | mean per-class |
| Caltech-101 [9] | 102 | 3,060 | 6,086 | mean per-class |
| Oxford 102 Flowers [11] | 102 | 2,040 | 6,149 | mean per-class |
| EuroSAT [5] | 10 | 10,000 | 5,000 | accuracy |
| Describable Textures (DTD) [2] | 47 | 3,760 | 1,880 | accuracy |

Table 1. Datasets in our experiments.

| Method | # Shot | ImageNet | CIFAR-10 | CIFAR-100 | STL-10 | Food-101 | Stanford Cars | FGVC Aircraft | Oxford Pets | Caltech-101 | Oxford Flowers | EuroSAT | DTD |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Zero-Shot CLIP | 0 | 59.8 | 71.6 | 40.6 | 94.4 | 80.6 | 54.3 | 17.0 | 85.5 | 84.5 | 65.5 | 41.8 | 41.2 |
| Linear Probe CLIP | 1 | 22.1 | 44.3 | 18.2 | 80.6 | 31.4 | 24.3 | 13.0 | 30.2 | 56.7 | 56.1 | 46.8 | 29.8 |
| | 2 | 31.9 | 53.5 | 26.6 | 86.9 | 45.1 | 36.8 | 17.9 | 40.5 | 68.7 | 74.9 | 55.5 | 41.4 |
| | 4 | 41.4 | 62.2 | 35.6 | 92.2 | 56.8 | 49.5 | 23.9 | 56.4 | 79.0 | 86.3 | 65.7 | 51.9 |
| | 8 | 49.4 | 70.1 | 43.6 | 94.3 | 67.1 | 61.2 | 29.4 | 67.5 | 84.3 | 91.5 | 75.1 | 59.0 |
| | 16 | 55.9 | 73.8 | 50.5 | 95.0 | 73.7 | 70.0 | 36.0 | 75.1 | 87.3 | 95.6 | 80.7 | 64.3 |
| CoOp | 1 | 53.4 | 71.8 | 41.3 | 94.1 | 77.5 | 54.0 | 17.7 | 86.5 | 83.6 | 70.0 | 47.1 | 44.1 |
| | 2 | 55.7 | 73.2 | 42.4 | 94.3 | 76.4 | 57.0 | 19.9 | 86.7 | 84.2 | 77.8 | 60.4 | 48.4 |
| | 4 | 57.9 | 75.4 | 45.7 | 94.9 | 77.0 | 61.4 | 22.7 | 87.2 | 85.5 | 84.8 | 69.4 | 53.4 |
| | 8 | 60.5 | 76.7 | 49.7 | 95.3 | 77.8 | 65.5 | 26.3 | 87.8 | 87.4 | 89.9 | 76.5 | 58.7 |
| | 16 | 62.3 | 78.4 | 53.3 | 95.6 | 79.3 | 70.5 | 30.1 | 88.4 | 89.6 | 93.4 | 81.4 | 65.0 |
| ProDA (ours) | 1 | 61.8 | 74.6 | 47.8 | 95.1 | 80.8 | 60.1 | 22.2 | 88.2 | 86.7 | 77.5 | 58.5 | 50.9 |
| | 2 | 62.3 | 76.4 | 49.4 | 95.3 | 80.6 | 63.7 | 24.8 | 88.4 | 87.1 | 85.2 | 69.0 | 56.2 |
| | 4 | 63.6 | 78.3 | 51.7 | 95.7 | 80.8 | 67.9 | 27.5 | 89.0 | 88.7 | 90.6 | 75.3 | 60.0 |
| | 8 | 64.7 | 79.6 | 54.3 | 96.1 | 81.7 | 72.1 | 31.5 | 89.4 | 89.8 | 93.6 | 80.1 | 65.0 |
| | 16 | 65.3 | 80.9 | 57.0 | 96.3 | 82.4 | 75.5 | 36.6 | 90.0 | 91.3 | 95.5 | 84.3 | 70.1 |

Table 2. Detailed performance (%) of various methods on the 12 downstream datasets. "# Shot" denotes the number of training samples per class.


[7] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *ICCVW*, 2013. 2

[8] Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, 2009. 2

[9] Ang Li, Allan Jabri, Armand Joulin, and Laurens van der Maaten. Learning visual n-grams from web data. In *ICCV*, 2017. 2

[10] S. Maji, J. Kannala, E. Rahtu, M. Blaschko, and A. Vedaldi. Fine-grained visual classification of aircraft. *arXiv:1306.5151*, 2013. 2

[11] M.-E. Nilsback and A. Zisserman. Automated flower classification over a large number of classes. In *ICVGIP*, 2008. 2

[12] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and C. V. Jawahar. Cats and dogs. In *CVPR*, 2012. 2

[13] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 1

[14] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *arXiv:2109.01134*, 2021. 1