Few-Shot Learning with Selective Attack and Cross-Modal Distribution Alignment

Supplementary Materials

1 Derivation of the EMD Upper Bound

In this work, the EMD can be written as [2]:

$$\mathrm{EMD}(\mathbf{VLP}, \mathbf{LP}) = \sum_k \mathrm{EMD}(\mathbf{v}_k, \mathbf{w}_k) = \sum_k \inf \mathbb{E}\|\mathbf{v}_k - \mathbf{w}_k\|, \tag{1}$$

which is the same as Eq. 10 in the manuscript, with $\mathbf{v}_k \sim \mathcal{N}(\boldsymbol{\mu}_{\mathrm{v}}^k, \boldsymbol{\Sigma}_{\mathrm{v}}^k)$ and $\mathbf{w}_k \sim \mathcal{N}(\boldsymbol{\mu}_{\mathrm{w}}^k, \boldsymbol{\Sigma}_{\mathrm{w}}^k)$. According to [3], if we define:

$$\mathbf{w}_k = \boldsymbol{\mu}_{\mathrm{w}}^k + \underbrace{\boldsymbol{\Sigma}_{\mathrm{v}}^{k^{-\frac{1}{2}}} \left(\boldsymbol{\Sigma}_{\mathrm{v}}^{k^{\frac{1}{2}}} \boldsymbol{\Sigma}_{\mathrm{w}}^{k} \boldsymbol{\Sigma}_{\mathrm{v}}^{k^{\frac{1}{2}}}\right)^{\frac{1}{2}} \boldsymbol{\Sigma}_{\mathrm{v}}^{k^{-\frac{1}{2}}}}_{\mathbf{A}^k} \left(\mathbf{v}_k - \boldsymbol{\mu}_{\mathrm{v}}^k\right), \tag{2}$$

since $\mathbf{A}^k = (\mathbf{A}^k)^T$,

$$\mathbb{E}(\mathbf{w}_k) = \boldsymbol{\mu}_{\mathrm{w}}^k + \mathbf{A}^k \left(\mathbb{E}(\mathbf{v}_k) - \boldsymbol{\mu}_{\mathrm{v}}^k\right) = \boldsymbol{\mu}_{\mathrm{w}}^k + \mathbf{A}^k \left(\boldsymbol{\mu}_{\mathrm{v}}^k - \boldsymbol{\mu}_{\mathrm{v}}^k\right) = \boldsymbol{\mu}_{\mathrm{w}}^k, \tag{3}$$

and

$$\begin{aligned}
\mathrm{Var}(\mathbf{w}_k) &= \mathbf{A}^k \boldsymbol{\Sigma}_{\mathrm{v}}^k (\mathbf{A}^k)^T \\
&= \boldsymbol{\Sigma}_{\mathrm{v}}^{k^{-\frac{1}{2}}} \left(\boldsymbol{\Sigma}_{\mathrm{v}}^{k^{\frac{1}{2}}} \boldsymbol{\Sigma}_{\mathrm{w}}^{k} \boldsymbol{\Sigma}_{\mathrm{v}}^{k^{\frac{1}{2}}}\right)^{\frac{1}{2}} \boldsymbol{\Sigma}_{\mathrm{v}}^{k^{-\frac{1}{2}}} \boldsymbol{\Sigma}_{\mathrm{v}}^{k} \boldsymbol{\Sigma}_{\mathrm{v}}^{k^{-\frac{1}{2}}} \left(\boldsymbol{\Sigma}_{\mathrm{v}}^{k^{\frac{1}{2}}} \boldsymbol{\Sigma}_{\mathrm{w}}^{k} \boldsymbol{\Sigma}_{\mathrm{v}}^{k^{\frac{1}{2}}}\right)^{\frac{1}{2}} \boldsymbol{\Sigma}_{\mathrm{v}}^{k^{-\frac{1}{2}}} \\
&= \boldsymbol{\Sigma}_{\mathrm{v}}^{k^{-\frac{1}{2}}} \left(\boldsymbol{\Sigma}_{\mathrm{v}}^{k^{\frac{1}{2}}} \boldsymbol{\Sigma}_{\mathrm{w}}^{k} \boldsymbol{\Sigma}_{\mathrm{v}}^{k^{\frac{1}{2}}}\right) \boldsymbol{\Sigma}_{\mathrm{v}}^{k^{-\frac{1}{2}}} \\
&= \boldsymbol{\Sigma}_{\mathrm{w}}^k,
\end{aligned} \tag{4}$$

then $\mathbf{w}_k \sim \mathcal{N}(\boldsymbol{\mu}_{\mathrm{w}}^k, \boldsymbol{\Sigma}_{\mathrm{w}}^k)$. With Eq. 2, we have

$$\begin{aligned}
\mathbf{D}^k &= \mathbf{v}_k - \mathbf{w}_k \\
&= \mathbf{v}_k - \boldsymbol{\mu}_{\mathrm{w}}^k - \mathbf{A}^k \left(\mathbf{v}_k - \boldsymbol{\mu}_{\mathrm{v}}^k\right) \\
&= (\mathbf{I} - \mathbf{A}^k)\mathbf{v}_k - \boldsymbol{\mu}_{\mathrm{w}}^k + \mathbf{A}^k \boldsymbol{\mu}_{\mathrm{v}}^k,
\end{aligned} \tag{5}$$

and the expectation of $\mathbf{D}^k$ is

$$\mathbb{E}(\mathbf{D}^k) = \boldsymbol{\mu}_{\mathrm{v}}^k - \boldsymbol{\mu}_{\mathrm{w}}^k. \tag{6}$$

For simplicity, suppose $\boldsymbol{\Sigma}_{\mathrm{v}}^k = \sigma_{\mathrm{v}^k}^2 \mathbf{I}$ and $\boldsymbol{\Sigma}_{\mathrm{w}}^k = \sigma_{\mathrm{w}^k}^2 \mathbf{I}$. Based on Jensen's inequality [1], we have:

$$
\begin{aligned}
(\mathbb{E}\|\mathbf{D}^k\|)^2 &\leq \mathbb{E}(\|\mathbf{D}^k\|^2) \\
&= \left\|\boldsymbol{\mu}_{\mathrm{v}}^k - \boldsymbol{\mu}_{\mathrm{w}}^k\right\|^2 + \operatorname{tr}\left(\boldsymbol{\Sigma}_{\mathrm{v}}^k + \boldsymbol{\Sigma}_{\mathrm{w}}^k - \mathbf{A}^k \boldsymbol{\Sigma}_{\mathrm{v}}^k - \boldsymbol{\Sigma}_{\mathrm{v}}^k \mathbf{A}^k\right) \\
&= \left\|\boldsymbol{\mu}_{\mathrm{v}}^k - \boldsymbol{\mu}_{\mathrm{w}}^k\right\|^2 + \operatorname{tr}\left(\boldsymbol{\Sigma}_{\mathrm{v}}^k\right) + \operatorname{tr}\left(\boldsymbol{\Sigma}_{\mathrm{w}}^k\right) - 2\operatorname{tr}\left(\boldsymbol{\Sigma}_{\mathrm{v}}^{k\,-\frac{1}{2}}\left(\boldsymbol{\Sigma}_{\mathrm{v}}^{k\,\frac{1}{2}} \boldsymbol{\Sigma}_{\mathrm{w}}^k \boldsymbol{\Sigma}_{\mathrm{v}}^{k\,\frac{1}{2}}\right)^{\frac{1}{2}} \boldsymbol{\Sigma}_{\mathrm{v}}^{k\,\frac{1}{2}}\right) \\
&= \left\|\boldsymbol{\mu}_{\mathrm{v}}^k - \boldsymbol{\mu}_{\mathrm{w}}^k\right\|^2 + \operatorname{tr}\left(\boldsymbol{\Sigma}_{\mathrm{v}}^k\right) + \operatorname{tr}\left(\boldsymbol{\Sigma}_{\mathrm{w}}^k\right) - 2\operatorname{tr}\left(\left(\boldsymbol{\Sigma}_{\mathrm{v}}^k \boldsymbol{\Sigma}_{\mathrm{w}}^k\right)^{\frac{1}{2}}\right) \\
&= \left\|\boldsymbol{\mu}_{\mathrm{v}}^k - \boldsymbol{\mu}_{\mathrm{w}}^k\right\|^2 + \|\boldsymbol{\Sigma}_{\mathrm{v}}^{k\,\frac{1}{2}} - \boldsymbol{\Sigma}_{\mathrm{w}}^{k\,\frac{1}{2}}\|^2.
\end{aligned} \tag{7}
$$

Then $\mathrm{EMD}(\mathbf{VLP}, \mathbf{LP})$ can be derived as:

$$
\begin{aligned}
\mathrm{EMD}(\mathbf{VLP}, \mathbf{LP}) &= \sum_k \inf \mathbb{E}\|\mathbf{v}_k - \mathbf{w}_k\| \\
&= \sum_k \inf \mathbb{E}\|\mathbf{D}^k\| \\
&\leq \sum_k \inf(\mathbb{E}\|\mathbf{D}^k\|^2)^{\frac{1}{2}} \\
&= \sum_k \inf(\left\|\boldsymbol{\mu}_{\mathrm{v}}^k - \boldsymbol{\mu}_{\mathrm{w}}^k\right\|^2 + \|\boldsymbol{\Sigma}_{\mathrm{v}}^{k\,\frac{1}{2}} - \boldsymbol{\Sigma}_{\mathrm{w}}^{k\,\frac{1}{2}}\|^2)^{\frac{1}{2}}.
\end{aligned} \tag{8}
$$

Finally, based on Eq. 8, we define the loss function as:

$$
\mathcal{L}_{\mathrm{EMD}} \triangleq \sum_k (\|\boldsymbol{\mu}_{\mathrm{v}}^k - \boldsymbol{\mu}_{\mathrm{w}}^k\|^2 + \|\boldsymbol{\Sigma}_{\mathrm{v}}^{k\,\frac{1}{2}} - \boldsymbol{\Sigma}_{\mathrm{w}}^{k\,\frac{1}{2}}\|^2). \tag{9}
$$

2 Geometric Explanation of Cross-Modal Distribution Alignment

To give an intuitive analysis, we regard the feature space as 2-dimensional. In Fig. 1, $\mathbf{z}_{i,j}$ is the feature of the current input image, $\mathbf{w}_{y_i}$ is the text feature of class y_i, $\mathbf{v}_{y_i}$ is the current vision-language prototype of class y_i, and $\mathbf{z}'_{i,j} = (1-\alpha)\mathbf{z}_{i,j} + \alpha \mathbf{v}_{y_i}$ is the image feature calibrated (aligned) by $\mathbf{v}_{y_i}$. The circle with its center $\mathbf{w}_{y_i}$ in Fig. 1 has the radius $|\mathbf{w}_{y_i} - \mathbf{z}_{i,j}|$. Since the target of the loss function $\mathcal{L}_{\mathrm{EMD}}$ is to align $\mathbf{v}_{y_i}$ and $\mathbf{w}_{y_i}$, the prototype $\mathbf{v}_{y_i}$ is usually closer to $\mathbf{w}_{y_i}$ than $\mathbf{z}_{i,j}$ after some iterations. It is easy to prove that as long as $\mathbf{v}_{y_i}$ is inside the circle, $\mathbf{z}'_{i,j}$ must be closer to $\mathbf{w}_{y_i}$ than $\mathbf{z}_{i,j}$, meaning that after the alignment, the image feature $\mathbf{z}_{i,j}$ becomes $\mathbf{z}'_{i,j}$ that is closer to the text feature $\mathbf{w}_{y_i}$.

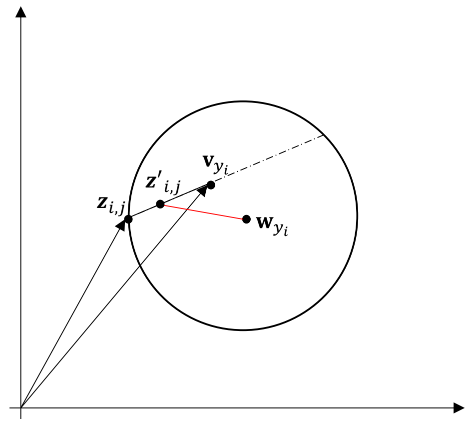

Fig. 1. Geometric explanation of why $(1-\alpha)\mathbf{z}_{i,j} + \alpha\mathbf{v}_{y_i}$ in Eq. 12 in the manuscript helps the alignment.

3 Datasets

The details of the 11 downstream datasets are shown in Table 1. The accuracy metric of each dataset follows CLIP [4].

Dataset	Classes	Train Size	Test Size	Accuracy Metric
ImageNet	1000	1281167	50000	accuracy
CIFAR-10	10	50000	10000	accuracy
CIFAR-100	100	50000	10000	accuracy
STL-10	10	1000	8000	accuracy
Food-101	101	75750	25250	accuracy
Stanford Cars	196	8144	8041	accuracy
FGVC Aircraft	100	6667	3333	mean per-class
Oxford-IIIT Pets	37	3680	3669	mean per-class
Caltech-101	102	3060	6086	mean per-class
DTD	47	3760	1880	accuracy
UCF-101	101	9537	1794	accuracy

Table 1. Datasets in our experiments.

4 Results

Table 2 shows the detailed results of 5 methods with the same pre-trained CLIP model (vision encoder=ResNet50) on the 11 datasets. In addition to the methods in the manuscript, Linear Probe CLIP is also used for the evaluation. Our SADA outperforms the others.

Method	#Shots	Average	ImageNet-1k	CIFAR10	Caltech101	Oxford Pets	Food101	STL10	UCF101	DTD	Stanford Cars	CIFAR100	FGVC Aircraft
Linear Probe CLIP	1	35.1	22.1	44.3	44.3	30.2	31.4	80.6	41.2	29.8	24.3	18.2	13.0
	2	44.9	31.9	53.5	68.7	40.5	45.1	86.9	53.8	41.4	36.8	26.6	17.9
	4	54.9	41.4	62.2	79.0	56.4	56.8	92.2	61.7	51.9	49.5	35.6	23.9
	8	62.6	49.4	70.1	84.3	67.5	67.1	94.3	67.2	59.0	61.2	43.6	29.4
	16	68.2	55.9	73.8	87.3	75.1	73.7	95.0	72.8	64.3	70.0	50.5	36.0
CoOp	1	62.4	53.4	71.8	83.6	86.5	77.5	94.1	60.0	44.1	54.0	41.3	17.7
	2	63.8	55.7	73.2	84.2	86.7	76.4	76.4	64.2	48.4	57.0	42.4	19.9
	4	66.1	57.9	75.4	85.5	87.2	77.0	94.9	65.8	53.4	61.4	45.7	22.7
	8	68.6	60.5	76.7	87.4	87.8	77.8	95.3	70.0	58.7	65.5	49.7	26.3
	16	71.3	62.3	78.4	89.6	88.4	79.3	95.6	74.6	65.0	70.5	53.3	30.1
CLIP-Adapter	1	64.0	58.3	73.6	86.1	86.8	77.9	94.3	64.1	45.3	54.3	43.2	19.8
	2	65.5	59.0	75.1	86.7	87.0	78.3	94.8	67.1	50.3	57.5	44.1	22.5
	4	67.5	59.7	76.5	88.5	87.3	78.5	95.2	70.0	56.0	61.8	46.3	25.3
	8	69.6	60.7	77.5	88.7	87.9	78.6	95.4	73.7	61.0	65.9	49.8	30.2
	16	72.2	61.3	79.1	90.1	88.5	79.3	95.6	76.7	65.8	71.0	52.9	38.0
Tip-Adapter	1	65.9	61.3	75.2	89.7	87.3	77.7	94.4	64.8	50.3	58.4	43.9	21.7
	2	67.0	61.7	75.9	89.9	87.9	77.8	94.5	67.5	52.9	61.5	44.3	23.1
	4	69.0	62.5	77.3	91.5	88.3	78	94.6	71.2	58.0	64.4	46.9	26.2
	8	71.2	64	77.7	92	89.1	78.4	94.8	74.8	61.7	68.8	48.4	33.8
	16	73.4	65.5	78.5	93.1	89.3	78.9	94.9	77.5	66.2	75.6	49.5	38.7
ProDA	1	66.8	61.8	74.6	86.7	88.2	80.8	95.1	66.4	50.9	60.1	47.8	22.2
	2	68.4	62.3	76.4	87.1	88.4	80.6	95.3	68.7	56.2	63.7	49.4	24.8
	4	70.3	63.6	78.3	88.7	89.0	80.8	95.7	71.5	60.0	67.9	51.7	27.5
	8	72.3	64.7	79.6	89.8	89.4	81.7	96.1	74.7	64.0	72.1	54.3	31.5
	16	74.3	65.3	80.9	90.0	90.0	82.4	96.3	77.3	68.7	75.5	57.0	36.6
SADA	1	**68.7**	**63.9**	**78.0**	**90.6**	**89.2**	**81.9**	**95.8**	**69.2**	**51.8**	**61.7**	**49.3**	**23.9**
	2	**69.9**	**64.3**	**78.7**	**91.4**	**89.4**	**82.1**	**96.0**	**69.9**	**57.7**	**64.5**	**50.0**	**25.3**
	4	**71.9**	**65.5**	**80.0**	**93.7**	**89.5**	**82.5**	**96.2**	**72.7**	**61.5**	**68.4**	**52.6**	**28.3**
	8	**74.1**	**66.9**	**80.9**	**93.2**	**89.9**	**82.7**	**96.3**	**75.9**	**65.0**	**73.6**	**55.0**	**31.9**
	16	**76.2**	**67.7**	**81.8**	**94.5**	**90.7**	**83.1**	**96.5**	**78.5**	**69.9**	**76.2**	**57.9**	**37.2**

Table 2. Accuracies (%) by 5 methods. #Shots: the number of training samples per class.

References

1. Chandler, D.: Introduction to modern statistical. In: Mechanics. Oxford University Press, Oxford, UK. vol. 5, p. 449 (1987)
2. Clement, P., Desch, W.: An elementary proof of the triangle inequality for the wasserstein metric. In: Proceedings of the American Mathematical Society. vol. 136, pp. 333–339 (2008)
3. Knott, M., Smith, C.S.: On the optimal mapping of distributions **43**, 39–49 (1984)
4. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: ICML (2021)