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]:

\[
\text{EMD}(\text{VLP, LP}) = \sum_k \text{EMD}(v_k, w_k) = \sum_k \inf \mathbb{E}||v_k - w_k||, \tag{1}
\]

which is the same as Eq. 10 in the manuscript, with \(v_k \sim \mathcal{N}(\mu_v^k, \Sigma_v^k)\) and \(w_k \sim \mathcal{N}(\mu_w^k, \Sigma_w^k)\). According to [3], if we define:

\[
w_k = \mu_w^k + \Sigma_v^k \frac{1}{2} \left( \Sigma_v^k \Sigma_w^k \Sigma_v^k \right)^{-\frac{1}{2}} \left( \Sigma_v^k \Sigma_w^k \Sigma_v^k \right)^{\frac{1}{2}} (v_k - \mu_v^k), \tag{2}
\]

since \(A^k = (A^k)^T\),

\[
\mathbb{E}(w_k) = \mu_w^k + A^k (\mathbb{E}(v_k) - \mu_v^k) = \mu_w^k + A^k (\mu_v^k - \mu_v^k) = \mu_w^k, \tag{3}
\]

and

\[
\text{Var}(w_k) = A^k \Sigma_v^k (A^k)^T
\]

\[
= \Sigma_v^k \frac{1}{2} \left( \Sigma_v^k \Sigma_w^k \Sigma_v^k \right)^{-\frac{1}{2}} \Sigma_v^k \Sigma_w^k \Sigma_v^k \left( \Sigma_v^k \Sigma_w^k \Sigma_v^k \right)^{\frac{1}{2}} \Sigma_v^k
\]

\[
= \Sigma_v^k \frac{1}{2} \left( \Sigma_v^k \Sigma_w^k \Sigma_v^k \right) \Sigma_v^k \frac{1}{2}
\]

\[
= \Sigma_v^k, \tag{4}
\]

then \(w_k \sim \mathcal{N}(\mu_w^k, \Sigma_w^k)\). With Eq. 2, we have

\[
D^k = v_k - w_k
\]

\[
= v_k - \mu_w^k - A^k (v_k - \mu_v^k)
\]

\[
= (I - A^k)v_k - \mu_w^k + A^k \mu_v^k, \tag{5}
\]

and the expectation of \(D^k\) is

\[
\mathbb{E}(D^k) = \mu_v^k - \mu_w^k. \tag{6}
\]
For simplicity, suppose \( \Sigma^k_v = \sigma^2_v I \) and \( \Sigma^k_w = \sigma^2_w I \). Based on Jensen’s inequality [1], we have:

\[
(\mathbb{E}\|D^k\|)^2 \leq \mathbb{E}(\|D^k\|^2)
\]

\[
= \|\mu^k_v - \mu^k_w\|^2 + \text{tr} \left( \Sigma^k_v + \Sigma^k_w - A^k \Sigma^k_v - \Sigma^k_w A^k \right)
\]

\[
= \|\mu^k_v - \mu^k_w\|^2 + \text{tr} \left( \Sigma^k_v \right) + \text{tr} \left( \Sigma^k_w \right) - 2 \text{tr} \left( \frac{1}{2} \left( \Sigma^k_v \frac{1}{2} \Sigma^k_v \Sigma^k_w \frac{1}{2} \right) \right)
\]

\[
= \|\mu^k_v - \mu^k_w\|^2 + \|\Sigma^k_v^{\frac{1}{2}} - \Sigma^k_w^{\frac{1}{2}}\|^2.
\]  

(7)

Then EMD(VLP, LP) can be derived as:

\[
\text{EMD} = \sum_k \inf \mathbb{E}\|v_k - w_k\|
\]

\[
= \sum_k \inf \mathbb{E}\|D^k\|
\]

\[
\leq \sum_k \inf(\mathbb{E}\|D^k\|^2)^{\frac{1}{2}}
\]

\[
= \sum_k \inf \left( \|\mu^k_v - \mu^k_w\|^2 + \|\Sigma^k_v^{\frac{1}{2}} - \Sigma^k_w^{\frac{1}{2}}\|^2 \right)^{\frac{1}{2}}.
\]  

(8)

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

\[
\mathcal{L}_{\text{EMD}} = \sum_k \left( \|\mu^k_v - \mu^k_w\|^2 + \|\Sigma^k_v^{\frac{1}{2}} - \Sigma^k_w^{\frac{1}{2}}\|^2 \right).
\]  

(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, \( z_{i,j} \) is the feature of the current input image, \( w_{y_i} \) is the text feature of class \( y_i \), \( v_{y_i} \) is the current vision-language prototype of class \( y_i \), and \( z'_{i,j} = (1 - \alpha)z_{i,j} + \alpha v_{y_i} \) is the image feature calibrated (aligned) by \( v_{y_i} \). The circle with its center \( w_{y_i} \) in Fig. 1 has the radius \( \|w_{y_i} - z_{i,j}\| \). Since the target of the loss function \( \mathcal{L}_{\text{EMD}} \) is to align \( v_{y_i} \) and \( w_{y_i} \), the prototype \( v_{y_i} \) is usually closer to \( w_{y_i} \) than \( z_{i,j} \) after some iterations. It is easy to prove that as long as \( v_{y_i} \) is inside the circle, \( z'_{i,j} \) must be closer to \( w_{y_i} \) than \( z_{i,j} \), meaning that after the alignment, the image feature \( z_{i,j} \) becomes \( z'_{i,j} \) that is closer to the text feature \( w_{y_i} \).
Fig. 1. Geometric explanation of why $(1 - \alpha)z_{i,j} + \alpha v_{y_i}$ 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].

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Classes Train Size</th>
<th>Test Size</th>
<th>Accuracy Metric</th>
</tr>
</thead>
<tbody>
<tr>
<td>ImageNet</td>
<td>1000</td>
<td>1281167</td>
<td>50000</td>
</tr>
<tr>
<td>CIFAR-10</td>
<td>10</td>
<td>50000</td>
<td>10000</td>
</tr>
<tr>
<td>CIFAR-100</td>
<td>100</td>
<td>50000</td>
<td>10000</td>
</tr>
<tr>
<td>STL-10</td>
<td>10</td>
<td>1000</td>
<td>8000</td>
</tr>
<tr>
<td>Food-101</td>
<td>101</td>
<td>75750</td>
<td>25250</td>
</tr>
<tr>
<td>Stanford Cars</td>
<td>196</td>
<td>8144</td>
<td>8041</td>
</tr>
<tr>
<td>FGVC Aircraft</td>
<td>100</td>
<td>6667</td>
<td>3333</td>
</tr>
<tr>
<td>Oxford-IIIT Pets</td>
<td>37</td>
<td>3680</td>
<td>3669</td>
</tr>
<tr>
<td>Caltech-101</td>
<td>102</td>
<td>3060</td>
<td>6086</td>
</tr>
<tr>
<td>DTD</td>
<td>47</td>
<td>3760</td>
<td>1880</td>
</tr>
<tr>
<td>UCF-101</td>
<td>101</td>
<td>9537</td>
<td>1794</td>
</tr>
</tbody>
</table>

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.
<table>
<thead>
<tr>
<th>Method</th>
<th>#Shots</th>
<th>Average</th>
<th>ImageNet-1k</th>
<th>CIFAR10</th>
<th>Caltech101</th>
<th>Oxford Pets</th>
<th>Food101</th>
<th>STL10</th>
<th>UCFC101</th>
<th>DTD</th>
<th>CIFAR100</th>
<th>FGVC Aircraft</th>
<th>PETS100</th>
</tr>
</thead>
<tbody>
<tr>
<td>Linear Probe CLIP</td>
<td>1</td>
<td>35.1</td>
<td>22.1</td>
<td>44.3</td>
<td>30.2</td>
<td>31.4</td>
<td>80.6</td>
<td>41.2</td>
<td>29.8</td>
<td>24.3</td>
<td>18.2</td>
<td>13.0</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>44.9</td>
<td>31.9</td>
<td>53.5</td>
<td>40.5</td>
<td>45.1</td>
<td>86.9</td>
<td>53.8</td>
<td>41.4</td>
<td>36.8</td>
<td>28.6</td>
<td>17.9</td>
<td></td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>54.9</td>
<td>41.4</td>
<td>62.2</td>
<td>79.0</td>
<td>56.4</td>
<td>56.8</td>
<td>92.2</td>
<td>61.7</td>
<td>51.9</td>
<td>49.5</td>
<td>35.6</td>
<td>23.9</td>
</tr>
<tr>
<td></td>
<td>8</td>
<td>62.6</td>
<td>49.4</td>
<td>70.1</td>
<td>84.3</td>
<td>67.5</td>
<td>67.1</td>
<td>94.3</td>
<td>67.2</td>
<td>59.0</td>
<td>61.2</td>
<td>43.6</td>
<td>29.4</td>
</tr>
<tr>
<td></td>
<td>16</td>
<td>68.2</td>
<td>55.9</td>
<td>73.8</td>
<td>87.3</td>
<td>75.1</td>
<td>73.7</td>
<td>95.0</td>
<td>72.8</td>
<td>64.3</td>
<td>70.0</td>
<td>59.5</td>
<td>36.0</td>
</tr>
<tr>
<td>CoOp</td>
<td>4</td>
<td>64.0</td>
<td>58.3</td>
<td>73.6</td>
<td>86.1</td>
<td>86.8</td>
<td>77.9</td>
<td>94.3</td>
<td>64.1</td>
<td>45.3</td>
<td>54.3</td>
<td>43.2</td>
<td>19.8</td>
</tr>
<tr>
<td></td>
<td>8</td>
<td>65.5</td>
<td>59.0</td>
<td>75.1</td>
<td>86.7</td>
<td>87.0</td>
<td>78.3</td>
<td>94.8</td>
<td>67.1</td>
<td>50.3</td>
<td>57.5</td>
<td>44.1</td>
<td>22.5</td>
</tr>
<tr>
<td></td>
<td>16</td>
<td>67.5</td>
<td>59.7</td>
<td>76.5</td>
<td>88.5</td>
<td>87.3</td>
<td>78.5</td>
<td>95.2</td>
<td>70.0</td>
<td>56.0</td>
<td>61.8</td>
<td>46.3</td>
<td>25.3</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>60.7</td>
<td>60.7</td>
<td>77.5</td>
<td>88.7</td>
<td>87.9</td>
<td>78.6</td>
<td>95.4</td>
<td>73.7</td>
<td>61.0</td>
<td>65.9</td>
<td>49.8</td>
<td>30.2</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>69.0</td>
<td>62.5</td>
<td>77.3</td>
<td>91.5</td>
<td>88.3</td>
<td>78</td>
<td>94.6</td>
<td>71.2</td>
<td>58.0</td>
<td>64.4</td>
<td>46.9</td>
<td>26.2</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>71.2</td>
<td>64.7</td>
<td>77.7</td>
<td>92.1</td>
<td>89.1</td>
<td>78.4</td>
<td>94.8</td>
<td>74.8</td>
<td>61.7</td>
<td>68.8</td>
<td>48.4</td>
<td>33.8</td>
</tr>
<tr>
<td></td>
<td>8</td>
<td>73.4</td>
<td>65.5</td>
<td>79.1</td>
<td>91.1</td>
<td>89.3</td>
<td>78.9</td>
<td>94.9</td>
<td>77.6</td>
<td>66.2</td>
<td>75.6</td>
<td>49.5</td>
<td>38.7</td>
</tr>
<tr>
<td></td>
<td>16</td>
<td>68.7</td>
<td>61.8</td>
<td>74.6</td>
<td>86.7</td>
<td>88.2</td>
<td>80.8</td>
<td>95.1</td>
<td>66.4</td>
<td>59.9</td>
<td>60.1</td>
<td>47.8</td>
<td>22.2</td>
</tr>
<tr>
<td>CLIP-Adapter</td>
<td>1</td>
<td>66.8</td>
<td>61.8</td>
<td>74.6</td>
<td>86.7</td>
<td>88.2</td>
<td>80.8</td>
<td>95.1</td>
<td>66.4</td>
<td>59.9</td>
<td>60.1</td>
<td>47.8</td>
<td>22.2</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>68.4</td>
<td>62.3</td>
<td>76.4</td>
<td>87.1</td>
<td>88.4</td>
<td>80.6</td>
<td>95.3</td>
<td>68.7</td>
<td>56.2</td>
<td>63.7</td>
<td>49.4</td>
<td>24.8</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>70.3</td>
<td>63.6</td>
<td>78.3</td>
<td>88.7</td>
<td>89.0</td>
<td>80.8</td>
<td>95.7</td>
<td>71.5</td>
<td>60.0</td>
<td>67.9</td>
<td>51.7</td>
<td>27.5</td>
</tr>
<tr>
<td></td>
<td>8</td>
<td>72.3</td>
<td>64.7</td>
<td>79.6</td>
<td>89.8</td>
<td>89.4</td>
<td>81.7</td>
<td>96.1</td>
<td>74.7</td>
<td>64.0</td>
<td>72.1</td>
<td>54.3</td>
<td>31.5</td>
</tr>
<tr>
<td></td>
<td>16</td>
<td>74.3</td>
<td>65.3</td>
<td>80.9</td>
<td>90.0</td>
<td>90.0</td>
<td>82.4</td>
<td>96.3</td>
<td>77.3</td>
<td>68.7</td>
<td>75.5</td>
<td>57.6</td>
<td>30.6</td>
</tr>
<tr>
<td>Tip-Adapter</td>
<td>1</td>
<td>68.7</td>
<td>63.9</td>
<td>78.0</td>
<td>90.6</td>
<td>89.2</td>
<td>81.9</td>
<td>95.8</td>
<td>69.2</td>
<td>51.8</td>
<td>61.7</td>
<td>49.3</td>
<td>23.9</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>69.9</td>
<td>64.3</td>
<td>78.7</td>
<td>91.4</td>
<td>89.4</td>
<td>82.1</td>
<td>96.0</td>
<td>69.9</td>
<td>57.7</td>
<td>64.5</td>
<td>50.0</td>
<td>25.3</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>71.9</td>
<td>65.5</td>
<td>80.0</td>
<td>93.7</td>
<td>89.5</td>
<td>82.5</td>
<td>96.2</td>
<td>72.7</td>
<td>61.5</td>
<td>68.4</td>
<td>52.6</td>
<td>28.3</td>
</tr>
<tr>
<td></td>
<td>8</td>
<td>74.1</td>
<td>66.9</td>
<td>80.9</td>
<td>93.2</td>
<td>89.9</td>
<td>82.7</td>
<td>96.3</td>
<td>75.9</td>
<td>65.0</td>
<td>73.6</td>
<td>55.0</td>
<td>31.9</td>
</tr>
<tr>
<td></td>
<td>16</td>
<td>76.2</td>
<td>67.7</td>
<td>81.8</td>
<td>94.5</td>
<td>90.7</td>
<td>83.1</td>
<td>96.5</td>
<td>78.5</td>
<td>69.9</td>
<td>76.2</td>
<td>57.9</td>
<td>37.2</td>
</tr>
</tbody>
</table>

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

**References**