Improved Visual Fine-tuning with Natural Language Supervision
Supplementary

Junyang Wang1+ Yuanhong Xu2 Juhua Hu3 Ming Yan2 Jitao Sang1,4 Qi Qian5†
1 School of Computer and Information Technology & Beijing Key Lab of Traffic Data Analysis and Mining, Beijing Jiaotong University, Beijing, China
2 DAMO Academy, Alibaba Group, Hangzhou, China
3 School of Engineering and Technology, University of Washington, Tacoma, WA 98402, USA
4 Peng Cheng Lab, Shenzhen, China
5 DAMO Academy, Alibaba Group, Bellevue, WA 98004, USA
{junyangwang, jtsang}@bjtu.edu.cn, {yuanhong.xuyh, yml19608, qi.qian}@alibaba-inc.com, juhuah@uw.edu

1. Theoretical Analysis

1.1. Proof of Theorem 1

Proof. Note that with the fixed features, the function $\mathcal{L}(\theta^0, W)$ is convex in $W$. Assuming the function is $m$-strongly convex such that for the arbitrary $(W_1, W_2)$, we have

$$\mathcal{L}(W_1) \geq \mathcal{L}(W_2) + \langle \nabla_w \mathcal{L}(W_2), W_1 - W_2 \rangle + \frac{m}{2} \|W_1 - W_2\|^2_F$$

Since $W^0$ is the optimal solution for $\mathcal{L}(\theta^0, W)$, we have

$$\|W^T - W^0\|^2_F \leq \frac{2}{m} (\mathcal{L}(\theta^0, W^T) - \mathcal{L}(\theta^0, W^0))$$

$$= \frac{2}{m} (\mathcal{L}(\theta^0, W^T) - \mathcal{L}(\theta^T, W^T) + \mathcal{L}(\theta^T, W^T) - \mathcal{L}(\theta^0, W^0))$$

$$\leq \frac{2}{m} (\mathcal{L}(\theta^0, W^T) - \mathcal{L}(\theta^T, W^T))$$

(1)

The last inequality is due to that fine-tuning can obtain a better performance than linear probing, i.e., $\mathcal{L}(\theta^T, W^T) \leq \mathcal{L}(\theta^0, W^0)$.

For fine-tuning, the loss function $\mathcal{L}$ is non-convex but can be Lipschitz continuous. With $L/2$ as the parameter of Lipschitz continuous, we have

$$\mathcal{L}(\theta^0, W^T) - \mathcal{L}(\theta^T, W^T) \leq \frac{L}{2} \|\theta^0 - \theta^T\|_F \leq \frac{L}{2} \epsilon$$

where the last inequality is from the constraint of fine-tuning. Taking it back to the Eqn. [1], the result is obtained.

1.2. Proof of Proposition 1

Proof. Note that the backbone is updated by SGD

$$\theta^t = \theta^{t-1} - \eta_t \nabla \mathcal{L}_{\theta^{t-1}}$$

Adding $t$ from 0 to $T$, we have $\theta^T = \theta^0 - \sum_{t} \eta_t \nabla \mathcal{L}_{\theta^{t-1}}$.

By applying the triangle inequality, the difference between $\theta^T$ and $\theta^0$ can be bounded as

$$\|\theta^0 - \theta^T\|_F = \|\sum_{t} \eta_t \nabla \mathcal{L}_{\theta^{t-1}}\|_F$$

$$\leq \sum_{t} \eta_t \|\nabla \mathcal{L}_{\theta^{t-1}}\|_F \leq \sum_{t} \eta_t \epsilon$$

With a cosine decay strategy and the initial learning rate as $\eta_0$, we have

$$\|\theta^0 - \theta^*\|_F \leq 0.5 \eta_0 \int_0^T 1 + \cos(x) dx = 0.5 \eta_0 \pi \epsilon$$

1.3. Proof of Theorem 2

Proof. According to the definition, we have

$$P_{i,k} = \frac{\exp((x_i - w_{yi})^T w_k + w_{yi}^T w_k)}{\sum_{j} \exp((x_i - w_{yi})^T w_j + w_{yi}^T w_j)}$$

With Cauchy-Schwarz inequality, we have

$$-\gamma \|x_i - w_{yi}\|_2 \leq (x_i - w_{yi})^T w_k \leq \gamma \|x_i - w_{yi}\|_2$$

Due to the fact that exponential function is monotone, we have

$$P_{i,k} \leq \frac{c \exp(w_{yi}^T w_k)}{\sum_{j} \exp(w_{yi}^T w_j)/c} = c^2 P_{yi,k}$$
<table>
<thead>
<tr>
<th>Method</th>
<th>Aircraft</th>
<th>Caltech</th>
<th>Cars</th>
<th>C10</th>
<th>C100</th>
<th>CUB</th>
<th>DTD</th>
<th>Flower</th>
<th>Food</th>
<th>Pet</th>
<th>SUN</th>
<th>Avg.</th>
</tr>
</thead>
<tbody>
<tr>
<td>CE + LS (mean)</td>
<td>76.80</td>
<td>94.76</td>
<td>89.21</td>
<td>98.02</td>
<td>88.59</td>
<td>78.79</td>
<td>75.95</td>
<td>96.12</td>
<td>88.29</td>
<td>91.57</td>
<td>69.92</td>
<td>86.17</td>
</tr>
<tr>
<td>CE + LS (std)</td>
<td>0.46</td>
<td>0.17</td>
<td>0.12</td>
<td>0.08</td>
<td>0.16</td>
<td>0.08</td>
<td>0.09</td>
<td>0.50</td>
<td>0.08</td>
<td>0.04</td>
<td>0.34</td>
<td>0.21</td>
</tr>
<tr>
<td>TeS (mean)</td>
<td>77.80</td>
<td>94.78</td>
<td>90.01</td>
<td>97.97</td>
<td>88.48</td>
<td>80.01</td>
<td>77.01</td>
<td>96.74</td>
<td>88.49</td>
<td>92.17</td>
<td>70.98</td>
<td>86.77</td>
</tr>
<tr>
<td>TeS (std)</td>
<td>0.16</td>
<td>0.10</td>
<td>0.10</td>
<td>0.11</td>
<td>0.10</td>
<td>0.32</td>
<td>0.12</td>
<td>0.10</td>
<td>0.08</td>
<td>0.13</td>
<td>0.11</td>
<td>0.13</td>
</tr>
</tbody>
</table>

Table 1. Comparison with ViT pre-trained by CLIP. The significantly better method examined by Student’s t-test is bolded.

\[ P_{i,k} \geq \frac{\exp(w_{y_i}^\top w_k)/c}{\sum_j c \exp(w_{y_i}^\top w_j)} = \frac{1}{c^2} P_{y_i,k} \]

where \( c = \exp(\gamma \|x_i - w_{y_i}\|_2) \).

2. Repeated Experiments on CLIP

We repeat experiments for the vision encoder of CLIP by 3 times and conduct Student’s t-test at the 95% confidence level in Table 1. It confirms that our method is significantly better than the best baseline on average.