Improved Visual Fine-tuning with Natural Language Supervision Supplementary

Junyang Wang[1*] Yuanhong Xu[2] Juhua Hu[3] Ming Yan[2] Jitao Sang[1,4] Qi Qian[5†]
[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, ym119608, 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

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

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

$$\begin{aligned}
&\|W^T - W^0\|_F^2 \leq \frac{2}{m}(\mathcal{L}(\theta^0, W^T) - \mathcal{L}(\theta^0, W^0)) \\
&= \frac{2}{m}(\mathcal{L}(\theta^0, W^T) - \mathcal{L}(\theta^T, W^T) + \mathcal{L}(\theta^T, W^T) - \mathcal{L}(\theta^0, W^0)) \\
&\leq \frac{2}{m}(\mathcal{L}(\theta^0, W^T) - \mathcal{L}(\theta^T, W^T)) \qquad (1)
\end{aligned}$$

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. □

*Work done during internship at DAMO Academy, Alibaba Group.
†Corresponding author

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 $\theta^T = \theta^0 - \sum_t^T \eta_t \nabla\mathcal{L}_{\theta^{t-1}}$. By applying the triangle inequality, the difference between θ^T and θ^0 can be bounded as

$$\begin{aligned}
\|\theta^0 - \theta^T\|_F &= \|\sum_t^T \eta_t \nabla\mathcal{L}_{\theta^{t-1}}\|_F \\
&\leq \sum_t^T \eta_t \|\nabla\mathcal{L}_{\theta^{t-1}}\|_F \leq \sum_t^T \eta_t \delta
\end{aligned}$$

With a cosine decay strategy and the initial learning rate as η_0, we have

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

□

1.3. Proof of Theorem 2

Proof. According to the definition, we have

$$P_{i,k} = \frac{\exp((\mathbf{x}_i - \mathbf{w}_{y_i})^\top \mathbf{w}_k + \mathbf{w}_{y_i}^\top \mathbf{w}_k)}{\sum_j^C \exp((\mathbf{x}_i - \mathbf{w}_{y_i})^\top \mathbf{w}_j + \mathbf{w}_{y_i}^\top \mathbf{w}_j)}$$

With Cauchy-Schwarz inequality, we have

$$-\gamma\|\mathbf{x}_i - \mathbf{w}_{y_i}\|_2 \leq (\mathbf{x}_i - \mathbf{w}_{y_i})^\top \mathbf{w}_k \leq \gamma\|\mathbf{x}_i - \mathbf{w}_{y_i}\|_2$$

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

$$P_{i,k} \leq \frac{c\exp(\mathbf{w}_{y_i}^\top \mathbf{w}_k)}{\sum_j^C \exp(\mathbf{w}_{y_i}^\top \mathbf{w}_j)/c} = c^2 P_{y_i,k}$$

Method	Aircraft	Caltech	Cars	C10	C100	CUB	DTD	Flower	Food	Pet	SUN	Avg.
CE + LS (mean)	76.80	94.76	89.21	98.02	88.59	78.79	75.95	96.12	88.29	91.57	69.92	86.17
CE + LS (std)	0.46	0.17	0.12	0.08	0.16	0.08	0.09	0.50	0.31	0.04	0.34	0.21
TeS (mean)	**77.80**	94.78	**90.01**	97.97	88.48	**80.01**	**77.01**	**96.74**	88.49	**92.17**	**70.98**	**86.77**
TeS (std)	0.16	0.10	0.10	0.11	0.10	0.32	0.12	0.10	0.08	0.13	0.11	0.13

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

and

$$P_{i,k} \geq \frac{\exp(\mathbf{w}_{y_i}^\top \mathbf{w}_k)/c}{\sum_j^C c \exp(\mathbf{w}_{y_i}^\top \mathbf{w}_j)} = \frac{1}{c^2} P_{y_i,k}$$

where $c = \exp(\gamma \|\mathbf{x}_i - \mathbf{w}_{y_i}\|_2)$. $\square$

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.