Online Zero-Shot Classification with CLIP Supplementary

Qi Qian[1]* and Juhua Hu[2]

[1] Alibaba Group, Bellevue, WA 98004, USA
[2] School of Engineering and Technology,
University of Washington, Tacoma, WA 98402, USA
qi.qian@alibaba-inc.com, juhuah@uw.edu

1 Theoretical Analysis

1.1 Proof of Theorem 1

Proof. According to the optimality of the closed-form solution, we have

$$\forall p_i, \quad \mathcal{L}(\rho^{i-1}, p_i^*) \leq \mathcal{L}(\rho^{i-1}, p_i) \tag{1}$$

With the gradient ascent of dual variables, we have

$$\begin{aligned}\mathcal{L}(\rho, p_i^*) - \mathcal{L}(\rho^{i-1}, p_i^*) &= \sum_j (\rho_j^{i-1} - \rho_j)(p_{i,j}^* - \frac{\alpha}{C}) \\ &\leq \frac{\|\rho^{i-1} - \rho\|_2^2 - \|\rho^i - \rho\|_2^2}{2\eta_\rho^i} + \frac{\eta_\rho^i}{2}\end{aligned} \tag{2}$$

Combining Eqns. 1 and 2, we have

$$\mathcal{L}(\rho, p_i^*) - \mathcal{L}(\rho^{i-1}, p_i) \leq \frac{\|\rho^{i-1} - \rho\|_2^2 - \|\rho^i - \rho\|_2^2}{2\eta_\rho^i} + \frac{\eta_\rho^i}{2}$$

Summing i from 1 to n, we have

$$\sum_i \mathcal{L}(\rho, p_i^*) - \mathcal{L}(\rho^{i-1}, p_i) \leq \frac{1}{2}(\sum_i \|\rho^{i-1} - \rho\|_2^2 (\frac{1}{\eta_\rho^i} - \frac{1}{\eta_\rho^{i-1}}) + \sum_i \eta_\rho^i)$$

By assuming $\|\rho\|_2 \leq \gamma$ and $\eta_\rho^i = c_\rho / \sqrt{i}$, we have

$$\sum_i \mathcal{L}(\rho, p_i^*) - \mathcal{L}(\rho^{i-1}, p_i) \leq \gamma^2 \sqrt{n} / c_\rho + c_\rho \sqrt{n}$$

* Corresponding author

1.2 Proof of Theorem 2

Proof. Due to the convexity of the loss function, we have

$$\begin{aligned}&\mathcal{L}(\mathbf{w}^{i-1}, \mathbf{x}_i) - \mathcal{L}(\mathbf{w}, \mathbf{x}_i) \leq \langle \nabla \mathcal{L}(\mathbf{w}^{i-1}, \mathbf{x}_i)), \mathbf{w}^{i-1} - \mathbf{w} \rangle \\ &\leq \frac{\|\mathbf{w}^{i-1} - \mathbf{w}\|_F^2 - \|\mathbf{w}^{i} - \mathbf{w}\|_F^2}{2\eta_w^i} + \frac{\eta_w^i}{2} \|\nabla \mathcal{L}(\mathbf{w}^{i-1}, \mathbf{x}_i))\|_F^2\end{aligned}$$

Assuming the norm of the gradient is bounded by σ as $\forall i, \|\nabla \mathcal{L}(\mathbf{w}^{i-1}, \mathbf{x}_i))\|_F \leq \sigma$ and summing i from 1 to n, we have

$$\begin{aligned}&\mathcal{L}(\mathbf{w}^{i-1}, \mathbf{x}_i) - \mathcal{L}(\mathbf{w}, \mathbf{x}_i) \leq \sum_i 2C(\frac{1}{\eta_w^i} - \frac{1}{\eta_w^{i-1}}) + c_w \sigma^2 \sqrt{n} \\ &\leq 2C\sqrt{n}/c_w + c_w \sigma^2 \sqrt{n}\end{aligned}$$

2 Experiments

2.1 Ablation Study

Effect of text prompts Besides the 7 text prompts for the ensemble, we also benchmark the performance with a single prompt as "a photo of a {class name}" for the comparison in Tab. 1. The performance of ViT-B as vision encoder is also evaluated. We denote the vanilla zero-shot transfer in CLIP as "Baseline".

Table 1: Comparison of accuracy (%) for different text prompt strategies on ImageNet. A single prompt is applied as "a photo of a {class name}" while 7 prompts are suggested in [4, 5].

Method	Single prompt	7 prompts	Single prompt	7 prompts
ResNet-50:			*ViT-B/16:*	
Baseline	58.15	60.32	66.72	68.75
OnZeta	61.49	62.50	70.01	70.99

First, we can observe that the ensemble of prompts outperforms the single prompt with a clear margin of about 2% on the baseline. It illustrates that the vanilla zero-shot transfer is sensitive to the selection of text prompts. With the proposed online learning strategy, OnZeta outperforms the baseline with both the single prompt and 7 prompts. Concretely, it surpasses the baseline by more than 3% on the single prompt with different vision encoders and is more than 2% better than the baseline with 7 prompts. The superior performance of OnZeta demonstrates the effectiveness of side information from the arrived examples in the proposed online scenario. Finally, the gap between different prompts shrinks by 1% when applying OnZeta and it shows that our method is less sensitive to the text prompts for obtaining the text proxy.

Comparision of similarity with different proxies Following [4], we have the averaged cosine similarity between each image and its nearest class proxy to depict the modality gap. First, given text proxy from class names, the similarity between images and the text proxy is only 0.26. After online learning by OnZeta, that between examples and the vision proxy increases to 0.39, which shows the efficacy of mitigating the modality gap.

Comparison on different pre-trained models Besides CLIP [5], there are many other pre-trained vision-language models. Tab. 2 summarizes the performance of our method with backbones from ALBEF [3] and BLIP [2], respectively. Compared with CLIP, OnZeta achieves even larger improvement than the standard zero-shot classification baseline. It shows that our method is applicable to different pre-trained models.

Table 2: Comparison of accuracy (%) on ImageNet with different pre-trained models.

Models	ALBEF [3]	BLIP [2]
Baseline	33.82	47.17
OnZeta	41.86	54.85

Comparison on variants of ImageNet We evaluate the performance of OnZeta on variants of ImageNet with shifted distributions, which include ImageNet-V2 [6], ImageNet-A [1] and ImageNet-Sketch [7], as shown in Tab. 3. With appropriate parameters, our method consistently improves the baseline and it demonstrates the robustness of OnZeta.

Table 3: Comparison of accuracy (%) for different variants of ImageNet with ViT-L as vision encoder.

Method	ImageNet-V2	ImageNet-A	ImageNet-Sketch
Baseline	70.21	70.85	59.70
OnZeta	71.23	72.61	62.71

Running time Finally, we demonstrate the efficiency of online learning. For ImageNet with ResNet-50 as the vision encoder, the total cost including feature extraction for zero-shot CLIP is 1.74 (milliseconds) per image and that of our method is 2.12 (ms). Both can work for real-time applications and the overhead introduced by OnZeta is mild.

Limitations

In this work, we only consider a single vison-language pre-trained model for online zero-shot transfer, while other pre-trained models, *e.g.*, large language models and diffusion models, can be leveraged to mitigate the challenges from the vision-language models.

References

1. Hendrycks, D., Zhao, K., Basart, S., Steinhardt, J., Song, D.: Natural adversarial examples. In: CVPR. pp. 15262–15271 (2021)
2. Li, J., Li, D., Xiong, C., Hoi, S.C.H.: BLIP: bootstrapping language-image pre-training for unified vision-language understanding and generation. In: ICML. vol. 162, pp. 12888–12900 (2022)
3. Li, J., Selvaraju, R.R., Gotmare, A., Joty, S.R., Xiong, C., Hoi, S.C.: Align before fuse: Vision and language representation learning with momentum distillation. In: NeurIPS. pp. 9694–9705 (2021)
4. Qian, Q., Xu, Y., Hu, J.: Intra-modal proxy learning for zero-shot visual categorization with clip. In: NeurIPS (2023)
5. 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. pp. 8748–8763. PMLR (2021)
6. Recht, B., Roelofs, R., Schmidt, L., Shankar, V.: Do imagenet classifiers generalize to imagenet? In: ICML. vol. 97, pp. 5389–5400. PMLR (2019)
7. Wang, H., Ge, S., Lipton, Z.C., Xing, E.P.: Learning robust global representations by penalizing local predictive power. In: NeurIPS. pp. 10506–10518 (2019)