uCAP: An Unsupervised Prompting Method for Vision-Language Models Supplementary Material

A. Tuan Nguyen[1⋆], Kai Sheng Tai[1], Bor-Chun Chen[1], Satya Narayan Shukla[1], Hanchao Yu[1], Philip Torr[2], Tai-Peng Tian[1], and Ser-Nam Lim[3]

[1] Meta
[2] University of Oxford
[3] University of Central Florida
a.tuan.nguyen@outlook.com,philip.torr@eng.ox.ac.uk,sernam@ucf.edu

A Gradient of $\log Z_Y(w)$

$$\begin{aligned}
\nabla_w \log Z_Y(w) &= \frac{\nabla_w Z_Y(w)}{Z_Y(w)} = \frac{\int \nabla_w e^{-E(x,Y,w)} dx}{Z_Y(w)} \\
&= \frac{\int e^{-E(x,Y,w)} \nabla_w \log e^{-E(x,Y,w)} dx}{Z_Y(w)} \\
&= \int \frac{e^{-E(x,Y,w)}}{Z_Y(w)} \nabla_w \log e^{-E(x,Y,w)} dx \\
&= -\int p(x|Y,w) \nabla_w E(x,Y,w) dx \\
&= -\mathbb{E}_{p(x|Y,w)} \left[\nabla_w E(x,Y,w)\right] \\
&\approx -\frac{1}{M} \sum_{i=1}^{M} \nabla_w E(\tilde{x}_i, Y, w),
\end{aligned}$$

where $\tilde{x}_i \sim p(x|Y,w), i \in \{1,2,...,M\}$, and the last equation is obtained from Monte Carlo sampling.

B Classification via Descriptions

To show that our method is truly versatile, in this section, we combine uCAP with [1] (which use class descriptions generated by a large language model instead of the default class names). We use the same class descriptions as released in their official github page [4], and additionally learn domain-level prompts with uCAP. The results of this experiment is presented in Table 1, which shows that uCAP can further improve the performance of class descriptions generated by Large Language Models.

⋆ Work done when Tuan was at the University of Oxford and doing an internship at Meta
[4] https://github.com/sachit-menon/classify_by_description_release

Table 1: Comparison with "Classification via Descriptions".

Method	ViT-B/16			ViT-L/14@336px		
	ImageNet	OxfordPets	CUB	ImageNet	OxfordPets	CUB
Classification via Description [1]	67.9	86.9	58.1	76.2	91.7	64.9
uCAP + Descriptions	**69.6**	**90.1**	**58.3**	**77.6**	**94.7**	**65.2**

Table 2: Results for UCF101 action recognition dataset with CLIP.

Method	RN50	RN101	ViT-B/32	ViT-B/16	ViT-L/14	ViT-L/14@336
Classnames Only	61.79	63.68	66.28	68.26	74.23	73.82
Classnames + HTPs	68.39	67.52	69.18	73.83	80.44	79.75
uCAP	**68.64**	**70.50**	**70.71**	**74.46**	**80.77**	**80.74**

C Video Experiment

We also perform an experiment on the UCF101 video dataset [3]. This is an action recognition dataset which consists of 101 actions. For each video, we sample 60 frames randomly, and average the frames' representation to get the video representation. For the hand-tuned prompts baseline, we use the same prompts as proposed in the CLIP paper [2]. As illustrated in Table 2, our method uCAP can be straightforwardly extended to this modality and the learned prompts lead to significant improvement in accuracy.

References

1. Menon, S., Vondrick, C.: Visual Classification via Description from Large Language Models. ICLR (2023)
2. 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. ICML (2021)
3. Soomro, K., Zamir, A.R., Shah, M.: UCF101: A Dataset of 101 Human Actions Classes From Videos in The Wild. arXiv (2012)