Large Language Models are Good Prompt Learners for Low-Shot Image Classification

Supplementary Material

Introduction

In the supplementary material, we provide extra discussions that did not fit in the main paper due to the space limitation, including i) Ablation study on the textual augmentation described in Sec 3.3; ii) Few-shot classification results on 8/4/2/1 shots with comparisons with previous methods.

Ablation on Textual Augmentations

As discussed in Sec 3.3, we perform textual augmentations in two steps: i) When computing $\hat{g}$, we replace the original template "`A photo of [STH]`", with "`A photo of [STH] with [NP]`", thereby enriching the descriptive content with noun phrases extracted from LLM responses; ii) We create new LLM prompt templates similar to "`In one sentence, describe the distinctive appearance of [STH]`" through GPT-4 [11], and average the scores for final prediction.

We show ablation study results in Tab. 1, with TA1 and TA2 referring to the step i) and ii) mentioned above. Results show that, even without textual augmentations, LLaMP still outperforms PSRC, the previous state-of-the-art, by 0.68% on base accuracy, 0.94% on novel accuracy and 0.83% on the harmonic mean. Moreover, we observe that both augmentation steps further improve the performance of LLaMP. More specifically, TA1 improves the HM by 0.35% while TA2 brings in another boost of 0.12%.

Method	TA1	TA2	Base	Novel	HM
PSRC [7]			84.26	76.10	79.97
LLaMP			84.94	77.04	80.80
		✓	84.78	77.31	80.86
	✓		85.16	77.50	81.15
	✓	✓	**85.16**	**77.71**	**81.27**

Table 1. Ablation study on textual augmentations.

Few-shot Classification

In addition to the 16-shot classification result reported in the main paper, we present few-shot classification results with with 8/4/2/1 numbers of shots in Tab. 2 and compare LLaMP against previous baseline models.

Results in Tab. 2 show that LLaMP outperforms previous SOTAs under all settings, on average of all 11 benchmarks, with 0.88% improvement with 8 shots. In particular, we observe that LLaMP surpasses PSRC [7] consistently on FGVCAircraft (Aircraft) [9] and Food [1] with all numbers of shots. Such observation aligns with our argument in the main paper that the knowledge from LLMs provides richer semantic information for fine-grained classification.

References

[1] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101–mining discriminative components with random forests. In *ECCV*, pages 446–461. Springer, 2014. 1, 2

[2] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *CVPR*, pages 3606–3613, 2014. 2

[3] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255. Ieee, 2009. 2

[4] Li Fei-Fei, Rob Fergus, and Pietro Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *CVPR workshop*, pages 178–178. IEEE, 2004. 2

[5] Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(7):2217–2226, 2019. 2

[6] Muhammad Uzair Khattak, Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. Maple: Multi-modal prompt learning. In *CVPR*, pages 19113–19122, 2023. 2

[7] Muhammad Uzair Khattak, Syed Talal Wasim, Muzammal Naseer, Salman Khan, Ming-Hsuan Yang, and Fahad Shahbaz Khan. Self-regulating prompts: Foundational model adaptation without forgetting. In *ICCV*, pages 15190–15200, 2023. 1, 2

[8] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *ICCV workshops*, pages 554–561, 2013. 2

[9] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013. 1, 2

[10] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing*, pages 722–729. IEEE, 2008. 2

[11] OpenAI. Gpt-4 technical report, 2023. 1

[12] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. Cats and dogs. In *CVPR*, pages 3498–3505. IEEE, 2012. 2

	8-Shot Classification											
	Average	ImageNet [3]	Caltech [4]	Pets [12]	Cars [8]	Flowers [10]	Food [1]	Aircraft [9]	SUN397 [15]	DTD [2]	EuroSAT [5]	UCF101 [14]
CLIP [13]	74.47	62.23	93.41	78.36	73.67	96.10	79.79	39.35	69.08	63.46	84.43	79.34
CoOp [17]	76.98	70.63	94.37	91.27	79.30	94.97	82.67	39.00	71.53	64.77	78.07	80.20
CoCoOp [16]	72.96	70.63	95.04	93.45	70.44	84.30	86.97	26.61	70.84	58.89	68.21	77.14
MaPLe [6]	78.89	70.30	95.20	92.57	79.47	95.80	83.60	42.00	73.23	66.50	87.73	81.37
PSRC [7]	80.69	**72.33**	95.67	93.50	80.97	**96.27**	86.90	43.27	**75.73**	69.87	88.80	**84.30**
LLaMP	**81.57**	72.30	**96.57**	**93.69**	**82.15**	96.20	**87.39**	**47.48**	75.18	**71.14**	**91.15**	84.06

	4-Shot Classification											
	Average	ImageNet [3]	Caltech [4]	Pets [12]	Cars [8]	Flowers [10]	Food [1]	Aircraft [9]	SUN397 [15]	DTD [2]	EuroSAT [5]	UCF101 [14]
CLIP [13]	68.01	54.85	92.05	71.17	63.38	92.02	73.19	32.33	63.00	55.71	77.09	73.28
CoOp [17]	74.02	68.73	94.40	92.57	74.47	92.17	84.47	30.83	69.97	58.70	70.80	77.10
CoCoOp [16]	71.21	70.39	94.98	92.81	69.39	78.40	86.88	24.79	70.21	55.04	65.56	74.82
MaPLe [6]	75.37	67.70	94.43	91.90	75.30	92.67	81.77	34.87	70.67	61.00	84.50	78.47
PSRC [7]	78.35	71.07	95.27	93.43	**77.13**	93.87	86.17	37.47	74.00	65.53	**86.30**	81.57
LLaMP	**78.83**	**71.37**	**95.84**	**93.61**	76.79	**93.96**	**87.17**	**40.02**	**74.05**	**66.37**	86.16	**81.80**

	2-Shot Classification											
	Average	ImageNet [3]	Caltech [4]	Pets [12]	Cars [8]	Flowers [10]	Food [1]	Aircraft [9]	SUN397 [15]	DTD [2]	EuroSAT [5]	UCF101 [14]
CLIP [13]	57.98	44.88	89.01	58.37	50.28	85.07	61.51	26.41	53.70	40.76	61.98	65.78
CoOp [17]	70.65	67.07	93.07	89.80	70.50	87.33	84.40	26.20	66.53	53.60	65.17	73.43
CoCoOp [16]	67.65	69.78	94.82	92.64	68.37	75.79	86.22	15.06	69.03	52.17	46.74	73.51
MaPLe [6]	72.58	65.10	93.97	90.87	71.60	88.93	81.47	30.90	67.10	55.50	78.30	74.60
PSRC [7]	75.29	69.77	94.53	92.50	**73.40**	**91.17**	85.70	31.70	71.60	59.97	79.37	78.50
LLaMP	**75.89**	**70.12**	**95.66**	**92.75**	72.20	89.16	**86.33**	**33.41**	**72.64**	**61.29**	**81.71**	**79.56**

	1-Shot Classification											
	Average	ImageNet [3]	Caltech [4]	Pets [12]	Cars [8]	Flowers [10]	Food [1]	Aircraft [9]	SUN397 [15]	DTD [2]	EuroSAT [5]	UCF101 [14]
CLIP [13]	45.83	32.13	79.88	44.06	35.66	69.74	43.96	19.61	41.58	34.59	49.23	53.66
CoOp [17]	67.56	66.33	92.60	90.37	67.43	77.53	84.33	21.37	66.77	50.23	54.93	71.23
CoCoOp [16]	66.79	69.43	93.83	91.27	67.22	72.08	85.65	12.68	68.33	48.54	55.33	70.30
MaPLe [6]	69.27	62.67	92.57	89.10	66.60	83.30	80.50	26.73	64.77	52.13	71.80	71.83
PSRC [7]	72.32	68.13	93.67	**92.00**	69.40	**85.93**	84.87	27.67	69.67	**56.23**	**73.13**	74.80
LLaMP	**72.42**	**69.12**	**94.59**	91.91	**70.02**	84.03	**85.83**	**30.39**	**69.69**	54.98	70.36	**75.72**

Table 2. Few shot classification results with 8/4/2/1 shots. All numbers, excepts ours, are obtained from [7].

[13] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language super-

vision. In *ICML*, pages 8748–8763. PMLR, 2021. 2

[14] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012. 2

[15] Jianxiong Xiao, Krista A Ehinger, James Hays, Antonio Torralba, and Aude Oliva. Sun database: Exploring a large collection of scene categories. *IJCV*, 119(1):3–22, 2016. 2

[16] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *CVPR*, pages 16816–16825, 2022. 2

[17] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *IJCV*, 130(9):2337–2348, 2022. 2