LoCoOp: Few-Shot Out-of-Distribution Detection via Prompt Learning

Atsuyuki Miyai[1] **Qing Yu**[1,2] **Go Irie**[3] **Kiyoharu Aizawa**[1]
[1]The University of Tokyo [2]LY Corporation [3]Tokyo University of Science
`{miyai,yu,aizawa}@hal.t.u-tokyo.ac.jp` `goirie@ieee.org`

Supplementary material

This supplementary material provides additional experimental results (Appendix A) and dataset details (Appendix B).

A Additional experimental results

A.1 Sensitivity to λ

In Fig. A, we show the sensitivity to a hyperparameter λ in Eq.(6) of the main paper, which controls the weight of OOD regularization loss. We use MCM and GL-MCM as test-time detection methods for LoCoOP. We report average FPR95 and AUROC scores on four OOD datasets in a 16-shot setting. We found that, when λ is smaller than 1, LoCoOp outperforms CoOp for both MCM and GL-MCM. In this paper, we set λ to 0.25. However, we observe that LoCoOp is not sensitive to λ so much except for $\lambda = 1$.

A.2 Detailed results on few-shot OOD detection

In this section, we show the detailed results of few-shot OOD detection with different numbers of ID samples. Fig. 2 in the main paper shows the average FPR and AUROC scores, and we omit the detailed results due to space limitations. In Table A, we show the results of 2, 4, and 8 shots detection results on all OOD datasets in detail. These results demonstrate that LoCoOp is the most effective method among comparison methods.

A.3 The effectiveness of LoCoOp on small-scale datasets

In this section, we show the effectiveness of LoCoOp on small-scale datasets. As for the datasets, we use ImageNet-100 (a 100-class subset of ImageNet) as the ID dataset. As for the OOD datasets, we adopt the same ones as the ImageNet-1K OOD datasets. In Table B, we show the OOD detection result. From this result, we find that LoCoOp outperforms CoOp on ImageNet-100.

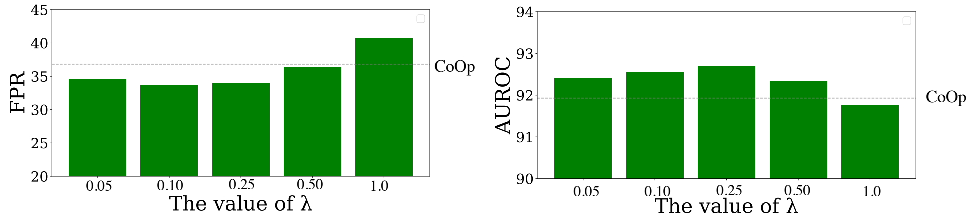
A.4 Relationship between the OOD detection performance and ID accuracy

In Table C, we show the ID classification accuracies for the OOD detection methods. We discuss the relationships between ID accuracy and OOD detection performance in the following three points.

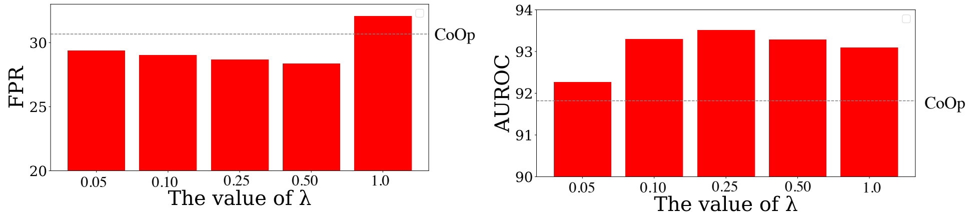
1. Why do zero-shot and prompt learning methods outperform fully supervised methods in OOD detection performance while their ID accuracies are considerably lower?

The key point in OOD detection is to avoid incorrectly assigning a high confidence score to OOD samples. In this respect, zero-shot and prompt learning methods calculate OOD scores based on the similarity between the text and the image, so models are less likely to produce unnaturally high

37th Conference on Neural Information Processing Systems (NeurIPS 2023).



(a) MCM



(b) GL-MCM

Figure A: Analysis of the sensitivity to a hyper-parameter λ

Table A: Few-shot OOD detection with different numbers of ID samples.

	iNaturalist		SUN		Places		Texture		**Average**	
Method	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑
Prompt learning					*two-shot (two label per class)*					
CoOp$_{\text{MCM}}$	38.89	92.12	39.38	91.58	44.18	88.98	44.92	89.16	41.85	90.46
CoOp$_{\text{GL}}$	21.17	95.36	35.00	91.08	42.25	88.32	49.23	85.79	36.91	90.14
LoCoOp$_{\text{MCM}}$ (ours)	35.38	92.76	33.95	93.31	41.15	90.38	45.07	89.76	38.89	91.55
LoCoOp$_{\text{GL}}$ (ours)	23.39	95.14	24.32	94.89	34.15	91.53	47.36	88.27	32.30	92.46
					four-shot (four labels per class)					
CoOp$_{\text{MCM}}$	35.36	92.60	37.06	92.27	45.38	89.15	43.74	89.68	40.39	90.92
CoOp$_{\text{GL}}$	18.95	95.52	29.58	92.90	38.72	89.64	48.03	85.87	33.82	90.98
LoCoOp$_{\text{MCM}}$ (ours)	29.45	93.93	33.06	93.24	41.13	90.32	44.15	90.54	36.95	92.01
LoCoOp$_{\text{GL}}$ (ours)	18.49	96.07	22.85	95.00	32.38	91.86	44.72	89.10	29.61	93.01
					eight-shot (eight labels per class)					
CoOp$_{\text{MCM}}$	35.17	92.96	34.45	92.50	41.17	89.76	43.29	89.92	38.52	91.29
CoOp$_{\text{GL}}$	15.23	96.69	27.78	93.08	35.93	90.22	48.26	85.91	31.80	91.47
LoCoOp$_{\text{MCM}}$ (ours)	27.12	94.60	33.87	93.23	40.53	90.53	42.49	90.98	36.00	92.34
LoCoOp$_{\text{GL}}$ (ours)	16.34	96.47	22.40	94.96	31.86	91.83	42.20	89.81	28.20	93.27

confidence scores for OOD samples. On the other hand, most fully-supervised methods do not use the language, and use the probability distribution through the last fc layer to calculate OOD scores. Therefore, even if the ID accuracy is high, there is a higher possibility that the model will produce an incorrect high confidence score for an OOD sample due to some reasons (*e.g.*, noisy activation signal [6]).

2. Why does LoCoOp have higher ID accuracy than CoOp in a 1-shot setting?

This is because CoOp does not have enough training samples in a 1-shot setting. As shown in Fig. 2 (main), CoOp and LoCoOp require about 16-shot image-label pairs to reach the upper score. On the other hand, even in a 1-shot setting, LoCoOp can learn from many OOD features, so LoCoOp outperforms CoOp in ID accuracy in a 1-shot setting.

3. Why does LoCoOp have lower ID accuracy than CoOp in a 16-shot setting?

In a 16-shot setting (sufficient training data for prompting methods), excluding OOD nuisances that are correlated with ID objects will degrade the ID accuracy. For example, in some images of dogs, the presence of green grass in the background may help identify the image as a dog. Therefore, learning to remove the background information could make it difficult to rely on such background information to determine that the image is a dog. However, this study reveals that excluding such backgrounds improves OOD detection performance.

Table B: Few-shot OOD detection on ImageNet-100.

	iNaturalist		SUN		Places		Texture		**Average**	
Method	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑
					16-shot (16 label per class)					
CoOp$_{\text{MCM}}$	16.66	97.05	7.58	98.48	13.72	97.18	20.33	95.78	14.57	97.12
CoOp$_{\text{GL}}$	8.11	98.22	8.38	98.24	14.41	96.95	24.37	94.32	13.82	96.93
LoCoOp$_{\text{MCM}}$ (ours)	14.98	97.34	6.20	98.72	11.11	97.71	18.42	96.18	12.68	97.49
LoCoOp$_{\text{GL}}$ (ours)	8.44	98.21	4.77	99.01	9.47	98.03	20.41	95.41	10.77	97.67

Table C: The relationship between ID accuracy and OOD detection performance.

	Average		**ID acc.**
Method	FPR95↓	AUROC↑	
Zero-shot			
MCM	42.82	90.76	67.01
GL-MCM	35.47	90.83	67.01
Fine-tuned			
ODIN	47.75	88.80	79.64
ViM	50.20	87.82	79.64
KNN	42.19	90.97	79.64
NPOS	37.93	91.22	79.42
Prompt learning		*one-shot (one label per class)*	
CoOp$_{\text{MCM}}$	44.81	90.03	66.23
CoOp$_{\text{GL}}$	44.81	90.03	66.23
LoCoOp$_{\text{MCM}}$	40.17	91.53	66.88
LoCoOp$_{\text{GL}}$	33.52	92.14	66.88
		16-shot (16 labels per class)	
CoOp$_{\text{MCM}}$	36.83	91.93	72.10
CoOp$_{\text{GL}}$	30.67	91.82	72.10
LoCoOp$_{\text{MCM}}$	33.98	92.69	71.70
LoCoOp$_{\text{GL}}$	28.66	93.52	71.70

B Dataset details

B.1 ID dataset

We use ImageNet-1K [2] as the ID data. We download the dataset via the official URL link `https://www.image-net.org/`. For the few-shot training, we follow the few-shot evaluation protocol adopted in CLIP [5] and CoOp [10], using 1, 2, 4, 8, and 16 shots for training, respectively. The average results over three runs are reported for comparison. For evaluation, we use the ImageNet validation dataset, which consists of 50,000 images with 1,000 classes following existing studies [4, 3].

B.2 OOD dataset

We use the following four datasets as OOD datasets following existing studies [4, 3]. We download all OOD datasets via `https://github.com/deeplearning-wisc/large_scale_ood`.

iNaturalist. iNaturalist [7] contains 859,000 plant and animal images across over 5,000 different species. We evaluate on 10,000 images randomly sampled from 110 classes that are disjoint from ImageNet-1K following [3].

SUN. SUN [8] contains over 130,000 images of scenes spanning 397 categories.We evaluate on 10,000 images randomly sampled from 50 classes that are disjoint from ImageNet-1K following [3].

Places. Places [9] is another scene dataset with similar concept coverage as SUN. We evaluate on 10,000 images randomly sampled from 50 classes that are disjoint from ImageNet-1K following [3].

TEXTURE. TEXTURE [1] contains 5,640 real-world texture images under 47 categories. We use the entire dataset for evaluation following [3].

References

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

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

[3] Rui Huang and Yixuan Li. Mos: Towards scaling out-of-distribution detection for large semantic space. In *CVPR*, 2021.

[4] Yifei Ming, Ziyang Cai, Jiuxiang Gu, Yiyou Sun, Wei Li, and Yixuan Li. Delving into out-of-distribution detection with vision-language representations. In *NeurIPS*, 2022.

[5] 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 supervision. In *ICML*, 2021.

[6] Yiyou Sun and Yixuan Li. Dice: Leveraging sparsification for out-of-distribution detection. In *ECCV*, 2022.

[7] Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The inaturalist species classification and detection dataset. In *CVPR*, 2018.

[8] Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *CVPR*, 2010.

[9] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *TPAMI*, 40(6):1452–1464, 2017.

[10] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *IJCV*, 2022.