How Does Fine-Tuning Impact Out-of-Distribution Detection for Vision-Language Models?

Yifei Ming[1] and Yixuan Li[1]

[1]Department of Computer Sciences, University of Wisconsin-Madison.

Contributing authors: ming5@wisc.edu; sharonli@cs.wisc.edu;

Abstract

Recent large vision-language models such as CLIP have shown remarkable out-of-distribution (OOD) detection and generalization performance. However, their zero-shot in-distribution (ID) accuracy is often limited for downstream datasets. Recent CLIP-based fine-tuning methods such as prompt learning have demonstrated significant improvements in ID classification and OOD generalization where OOD labels are available. Nonetheless, it remains unclear whether the model is reliable to semantic shifts without OOD labels. In this paper, we aim to bridge the gap and present a comprehensive study to understand how fine-tuning impact OOD detection for few-shot downstream tasks. By framing OOD detection as multi-modal concept matching, we establish a connection between fine-tuning methods and various OOD scores. Our results suggest that a proper choice of OOD scores is essential for CLIP-based fine-tuning. In particular, the maximum concept matching (MCM) score provides a promising solution consistently. We also show that prompt learning demonstrates the state-of-the-art OOD detection performance over the zero-shot counterpart.

Keywords: CLIP, OOD detection, fine-tuning, multi-modality, vision-language models, prompt learning, few-shot learning, adaptor

1 Introduction

Machine learning (ML) is undergoing a paradigm shift with the rise of models that are trained on massive data and are adaptable to a wide range of downstream tasks. Popular pre-trained large vision-language models (Radford et al., 2021; Jia et al., 2021; Yao et al., 2021; Li et al., 2022) demonstrate remarkable performance, and allow researchers without extensive computation power to benefit from these models. It is now the common practice of the ML community to adopt pre-trained models for transfer learning on downstream tasks rather than learning from scratch. Despite the promise, the safety risks of these large pre-trained models can be potentially inherited by all the fine-tuned models. Without appropriately understanding the safety risks, development on top of pre-trained models can exacerbate and propagate safety concerns writ large, causing profound impacts on society.

In response to these urgent challenges, the overall objective of this paper is to systematically understand the out-of-distribution risks of learning with pre-trained vision-language models. This paper seeks to address the research question that arises in building responsible and ethical AI models: *How does fine-tuning influence out-of-distribution (OOD) detection for large vision-language models?* Detecting OOD samples is crucial for machine learning models deployed in the open world, where samples from unseen classes

naturally emerge, and failure to detect them can have severe consequences. Despite increasing attention (Yang et al., 2021), OOD detection research for large vision-language models has been scant. Among the most recent works, Ming et al. (2022) investigated training-free OOD detection based on the pre-trained CLIP model. However, the impact of fine-tuning on OOD detection has been unexplored in the vision-language literature.

In this paper, we bridge the gap by investigating how fine-tuning large vision-language models affects OOD detection. Parameter-efficient fine-tuning methods have been popularized in recent years. In particular, prompt learning (Zhou et al., 2022a,b) optimizes learnable word embeddings of the prompts, while adaptors directly optimize the internal feature representations (Gao et al., 2021; Zhang et al., 2022). Both methods are parameter-efficient as image and text encoders are frozen during fine-tuning, and have shown significant improvement for few-shot in-distribution (ID) classification. Complementary to existing research, we focus on OOD detection for fine-tuned models using multi-modal concept matching. At the core of the concept matching framework, we use the few-shot ID training set and textual descriptions of the labels to derive a set of visual and textual features that represent the typical features for each ID class. We can measure OOD uncertainty based on the distance between the input feature and the nearest ID prototype.

Based on the concept matching framework, we then present a comprehensive and systematic study to explore how different parameter-efficient fine-tuning methods impact OOD detection performance, and contribute unexplored findings to the community. We disentangle various aspects such as adaptation methods and OOD scoring functions. Interestingly, we observe that parameter-efficient fine-tuning can significantly improve OOD reliability compared to zero-shot CLIP models. In particular, prompt learning methods exhibit very competitive performance when coupled with the maximum concept matching (MCM) score (Ming et al., 2022).

Furthermore, we delve deeper into prompt learning and analyze how the pre-trained features are modified during fine-tuning, and how it impacts OOD detection as a consequence. We study the impact of shots, architectures, and explore the effects of prompt learning on various downstream tasks, including the challenging ImageNet-1k (ID) benchmark. Our results demonstrate that prompt learning perturbs the pre-trained feature space that benefits both ID and OOD performance. More encouragingly, the trend holds consistently across different settings, highlighting its potential for reliable fine-tuning in vision-language modeling.

We summarize the contributions of this work as follows:

- We provide a timely and systematic study on how CLIP-based fine-tuning influences OOD detection in the few-shot setting. Our study disentangles various factors, including adaptation methods and OOD scoring functions.
- We present novel evidence that parameter-efficient fine-tuning does not deteriorate pre-trained features. Instead, they can improve both ID and OOD performance with a proper OOD scoring function, especially the MCM score. We show that prompt learning consistently demonstrates the state-of-the-art OOD detection performance over the zero-shot counterpart.
- We provide an in-depth analysis of prompt learning's impact on the feature space for OOD detection and conduct comprehensive ablations across datasets, architectures, and the number of shots with various OOD detection scores.

2 Preliminaries

Contrastive vision-language models. Recent large vision-language models have shown great potential for various computer vision tasks. In this paper, we focus on CLIP-like models (Radford et al., 2021; Yao et al., 2021), which adopt a dual-stream architecture with one text encoder $f : t \to \mathbb{R}^d$ and one image encoder $g : \boldsymbol{x} \to \mathbb{R}^d$. CLIP is pre-trained on a massive web-scale image-caption dataset with a multi-modal contrastive loss that promotes the alignment of features from different modalities. CLIP learns transferable feature representations and demonstrates promising zero-shot generalization performance (Fort et al., 2021). Despite the promise, existing vision-language models perform zero-shot classification in a *closed-world* setting. That is, it will match an input into a fixed set of categories, even if it is irrelevant. For example, a bird in Figure 1 can

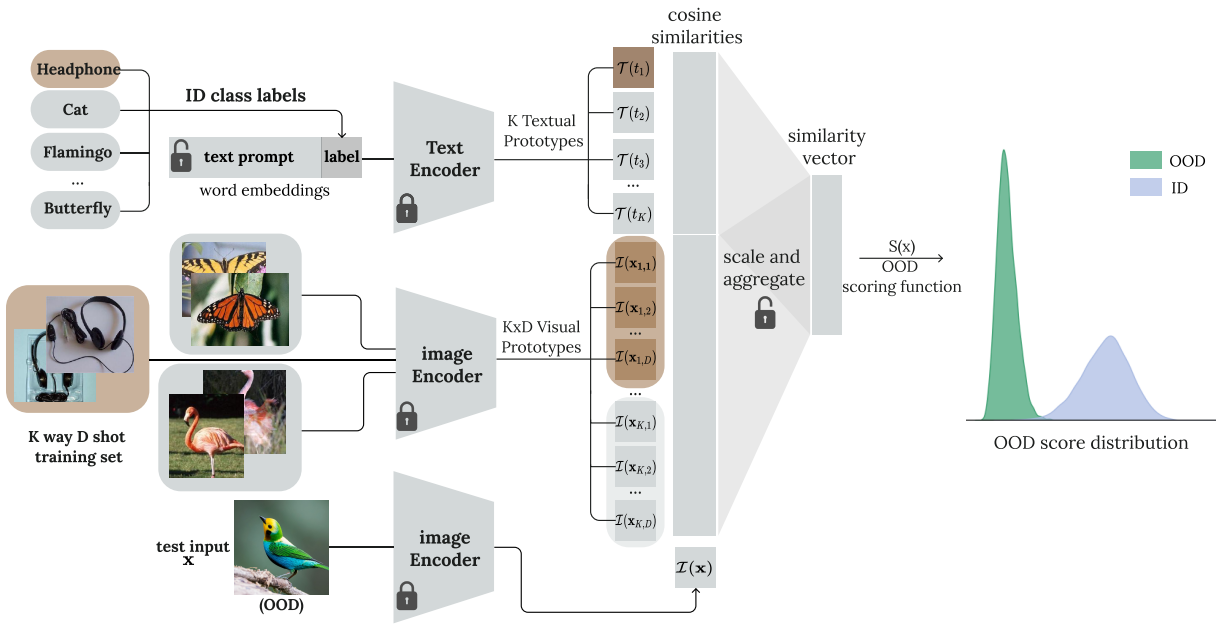

Fig. 1 A unified pipeline for OOD detection with parameter-efficient fine-tuning of CLIP models on few-shot datasets. Given ID text labels $\mathcal{Y}_{\text{in}}$ and a few-shot training set, we view the textual and visual embeddings of ID classes as concept prototypes in the feature space. The OOD uncertainty of an input image can be characterized by the distance from its visual feature to the closest ID prototype from both modalities. See Section 3 for details.

be blindly predicted as one of the in-distribution classes $\mathcal{Y}_{\text{in}}$ ={headphone, cat, flamingo, butterfly}. This motivates the importance of OOD detection for vision-language models.

OOD detection for vision-language models. In the open-world setting, the goal of OOD detection is to detect samples that do not belong to ID classes $\mathcal{Y}_{\text{in}}$. Here ID classes are defined w.r.t. the classification task of interest, instead of the classes used in pre-training. Accordingly, OOD is defined w.r.t. the ID classes, not the data distribution during pre-training. Ming et al. (2022) explore the zero-shot OOD detection for the pre-trained CLIP model, without adapting to the ID dataset. Instead, we focus on the setting where CLIP models are fine-tuned on a few-shot dataset $\mathcal{D}_{\text{in}}$, and hence are better adapted to the downstream ID task. We evaluate the fine-tuned CLIP model on a combination of ID and OOD datasets $\mathcal{D}_{\text{in}} \cup \mathcal{D}_{\text{out}}$, where $\mathcal{D}_{\text{out}} = \{\boldsymbol{x}_i, y_i^{\text{out}}\}_{i=1}^{m}$ contains inputs with semantically different categories $y^{\text{out}} \notin \mathcal{Y}_{\text{in}}$. Formally, given an input $\boldsymbol{x}$, OOD detection can be formulated as:

$$G(\boldsymbol{x};\ f, g) = \begin{cases} 1 & S(\boldsymbol{x};\ f, g) \geq \lambda \\ -1 & S(\boldsymbol{x};\ f, g) < \lambda \end{cases},$$

where $S(\cdot)$ is a scoring function that measures OOD uncertainty. In practice, λ is chosen so that a high fraction of ID data (*e.g.*, 95%) is above the threshold.

Parameter-efficient fine-tuning. To improve the performance on downstream tasks, parameter-efficient approaches are proposed to fine-tune CLIP on datasets of interest. Prompt learning and adaptor tuning have recently gained popularity and demonstrated improved results over zero-shot settings. In particular, prompt learning optimizes the word embeddings of the prompts, while adaptors directly optimize the internal feature representations. Both methods are parameter-efficient as image and text encoders are frozen during fine-tuning. In what follows, we introduce prompt-based and adaptor-based methods respectively.

For a downstream dataset with K in-distribution classes $\mathcal{Y}_{\text{in}} = \{y_1, y_2, ..., y_K\}$, prompt learning method such as `CoOp` (Zhou et al., 2022b) introduces M learnable context vectors $v_i \in \mathbb{R}^e$ to replace hand-engineered text prompts such as "`this is a photo of`", where e is the dimension of word embeddings. For each class y_k, we obtain its contextualized representation $t_k = [v_1, v_2, \cdots, v_M, w_k]$ by concatenating the context vectors and the word embedding $w_k \in \mathbb{R}^e$ of

the label (upper left, Figure 1). To avoid overfitting and improve generalization performance, `CoCoOp` (Zhou et al., 2022a) further introduces instance-conditional prompts via a meta-network which produces a meta token $m(\boldsymbol{x})$ given the visual feature of the input $\boldsymbol{x}$. The meta token is added to each context token $v_i(\boldsymbol{x}) = v_i + m(\boldsymbol{x})$ for $i \in \{1, 2, \cdots, M\}$. Therefore, the prompt for class k is conditioned on each input: $t_k(\boldsymbol{x}) = [v_1(\boldsymbol{x}), v_2(\boldsymbol{x}), \cdots, v_M(\boldsymbol{x}), w_k]$. To learn the context vectors, the cross-entropy loss is used in fine-tuning:

$$p(y_k \mid \boldsymbol{x}) = \frac{\exp\left(s_k(\boldsymbol{x})/\tau\right)}{\sum_{i=1}^{K} \exp\left(s_i(\boldsymbol{x})/\tau\right)}, \tag{1}$$

where $s_k(\boldsymbol{x}) = \frac{g(\boldsymbol{x}) \cdot f(t_k)}{\|g(\boldsymbol{x})\| \cdot \|f(t_k)\|}$ is the cosine similarity of input $\boldsymbol{x}$ with the k-th label, and τ is the temperature.

Alternatively, adaptor-based methods directly optimize the feature representations $g(\boldsymbol{x})$ instead of learning context vectors. Specifically, given a K-way-D-shot ID training set (consisting of K classes with D examples per class), Zhang et al. (2022) propose a training-free adaptation method `TipAdaptor` which extracts all the visual features $W_g = [g(\boldsymbol{x}_{1,1}), g(\boldsymbol{x}_{1,2}), \cdots, g(\boldsymbol{x}_{K,D})] \in \mathbb{R}^{KD \times d}$ from the few-shot training dataset. For each input $\boldsymbol{x}$, we can obtain $K \times D$ cosine similarities $s_{k,d}(\boldsymbol{x}) = \frac{g(\boldsymbol{x}) \cdot g(\boldsymbol{x}_{k,d})}{\|g(\boldsymbol{x})\| \cdot \|g(\boldsymbol{x}_{k,d})\|}$. The cosine similarities are scaled by an exponential function $\tilde{s} : s \mapsto \exp(-\beta + \beta s)$ with a hyperparameter β that modulates the sharpness. Therefore, we can obtain an average similarity vector for each class based on visual features, $\tilde{s}_k(\boldsymbol{x}) = \frac{1}{D} \sum_{d=1}^{D} \tilde{s}_{k,d}(\boldsymbol{x})$. The final similarity for class k is a weighted sum of similarities from the two modalities $\alpha \tilde{s}_k(\boldsymbol{x}) + s_k(\boldsymbol{x})$. To achieve better few-shot ID performance, Zhang et al. (2022) set visual features W_g as learnable parameters and denote the method as `TipAdaptorF`, where F stands for fine-tuning. Despite the stronger downstream classification performance, it remains unknown if fine-tuning leads to more reliable OOD detection at test time. We aim to provide a comprehensive understanding in this paper.

3 Method

3.1 OOD detection with fine-tuning

We investigate OOD detection with parameter-efficient fine-tuning on downstream tasks. We present a unified framework in Figure 1, where the learnable part of the CLIP model is marked with an "unlock" icon while the frozen part is marked with a "lock" icon. For prompt learning methods such as `CoOp` and `CoCoOp`, the cosine similarity of the input feature with the k-th class $s_k(\boldsymbol{x}) = \frac{g(\boldsymbol{x}) \cdot f(t_k)}{\|g(\boldsymbol{x})\| \cdot \|f(t_k)\|}$ is derived based on the adapted textual feature vector t_k. Alternatively, adaptor-based methods such as `TipAdaptor` and `TipAdaptorF` first scale the cosine similarities of visual prototypes and perform a weighted sum with the similarities of textual prototypes. Therefore, we can view `TipAdaptor` as an ensemble method that utilizes multi-modal prototypes.

To summarize, for each adaptation algorithm $\mathcal{A}$, OOD detection can be performed by:

$$G_{\mathcal{A}}(\boldsymbol{x};\ f, g) = \begin{cases} \text{ID} & S(\boldsymbol{x};\ f, g) \geq \lambda \\ \text{OOD} & S(\boldsymbol{x};\ f, g) < \lambda \end{cases},$$

where $\mathcal{A}$ can be instantiated by an adaptation method such as `CoOp`, `CoCoOp`, `TipAdaptor`, or `TipAdaptorF`. Therefore, the OOD detector $G_{\mathcal{A}}(\cdot)$ can be viewed as a "safeguard" for the classification model. Next, we introduce various OOD score functions $S(\boldsymbol{x};\ f, g)$ assuming $G_{\mathcal{A}}(\boldsymbol{x};\ f, g)$ is defined implicitly as each score function corresponds to an OOD detector G.

3.2 OOD score for vision-language models

Recently, Ming et al. (2022) propose a conceptual framework of CLIP-based OOD detection via concept matching, where the textual feature $f(t_k)$ is viewed as the concept prototype for ID class $k \in \{1, 2, ..., K\}$. OOD uncertainty is then characterized by the distance from the visual feature of the input to the closest ID textual prototype. That is, images closer to one of the ID prototypes are more likely to be ID and vice versa. Ming et al. (2022) suggest that softmax scaling with a proper temperature τ provably leads to state-of-the-art performance under the zero-shot (training-free) setting. Specifically, the maximum

concept matching (MCM) score is defined as:

$$S_{\text{MCM}}(\boldsymbol{x}) = \max_{k\in[K]} \frac{e^{s_k(\boldsymbol{x})/\tau}}{\sum_{j=1}^{K} e^{s_j(\boldsymbol{x})/\tau}}, \tag{2}$$

where the temperature τ needs to be tuned on the downstream dataset. As a special case of MCM, we use MSP to denote the MCM score when the temperature τ_d is set as default for CLIP models at inference time (*e.g.,* 100 for CLIP-B/16).

Additionally, we consider a simpler scoring function based on the maximum similarity (MS) among ID prototypes before applying softmax scaling:

$$S_{\text{MS}}(\boldsymbol{x}) = \max_{k\in[K]} s_k(\boldsymbol{x}), \tag{3}$$

which does not require any hyperparameter tuning. We show in Section 4 that the MS score demonstrates strong OOD detection performance with fine-tuning, especially for fine-grained ID datasets. We now proceed to experiments where we investigate the impact of fine-tuning on real-world tasks.

4 Experiments

4.1 Setup

Datasets. Following Ming et al. (2022), we consider a wide range of real-world ID datasets with various semantics and number of classes: Caltech-101 (Bossard et al., 2014), Stanford-Cars (Krause et al., 2013), Food-101 (Bossard et al., 2014), Oxford-Pets (Parkhi et al., 2012) and ImageNet-1k (Deng et al., 2009). For each ID dataset, we follow Zhou et al. (2022a) and construct the training set with D random samples per class, while the original test set is used for testing. We use $D = 16$ by default and study the impact of shots as ablations in Section 4.3. For OOD test datasets, we use the same ones in Huang and Li (2021), including subsets of iNaturalist (Van Horn et al., 2018), SUN (Xiao et al., 2010), PLACES (Zhou et al., 2017), and TEXTURE (Cimpoi et al., 2014). For each OOD dataset, the categories do not overlap with the ID dataset. For ImageNet-1k as ID, we also consider two additional OOD datasets ImageNet-O (Hendrycks et al., 2021) and OpenImage-O (Wang et al., 2022).

Models and training details. For pre-trained models, we use CLIP-B/16 as the default backbone for main experiments, which uses ViT-B/16 (Dosovitskiy et al., 2021) as the image encoder. The impact of backbones is included in the ablation studies. We use `ZOCLIP` to denote pre-trained CLIP without fine-tuning. For each method, we closely follow the original implementations. Specifically, for `CoOp` and `CoCoOp`, the context length is set to 4, and the context vectors are initialized using the pre-trained word embeddings of "a photo of a". `CoCoOp` is trained with a batch size of 1 for 10 epochs using SGD, while `CoOp` is trained for 100 epochs with a batch size of 32. `TipAdapterF` is trained with a batch size 256 using AdamW (Loshchilov and Hutter, 2019) for 20 epochs. Cosine scheduling is used for all methods and the data preprocessing protocol consists of random re-sizing, cropping, and random horizontal flip.

Evaluation metrics. We consider the following evaluation metrics: (1) the false positive rate (FPR95) of OOD samples when the true positive rate of in-distribution samples is at 95%, (2) the area under the receiver operating characteristic curve (AUROC), and (3) ID classification accuracy (ID ACC).

4.2 Main results and discussions

In this section, we first present novel evidence that parameter-efficient fine-tuning generally improves OOD performance over the zero-shot counterpart with a simple OOD scoring function. Next, we investigate the effects of various OOD scoring functions in the parameter-efficient fine-tuning setting. In particular, we will show that the MCM score consistently demonstrates the most promising performance compared to alternative OOD scores when coupled with prompt learning.

How does parameter-efficient fine-tuning impact OOD detection? We evaluate the OOD detection performance on various ID datasets. The results are summarized in Table 1. We show that adapted CLIP models demonstrate nearly perfect OOD detection performance for ID datasets with fine-grained categories such as Stanford-Cars and Oxford-Pets. Moreover, when the ID dataset contains a diverse collection of categories such as

Table 1 OOD detection performance based on S_{MS} score (w.o. softmax scaling). When ID datasets contain finer-grained categories semantically different from OOD categories, the pre-trained CLIP model demonstrates nearly perfect OOD detection performance. More encouragingly, after adapting the model to downstream datasets, OOD detection performance remains competitive.

ID Dataset	Method	SUN		Places		Textures		iNaturalist		Average	
		FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑
						Training not required					
	ZOCLIP	0.04	99.92	0.12	99.93	4.63	98.29	0.15	99.87	1.24	99.50
	TipAdaptor	0.00	99.94	0.04	99.95	2.87	98.85	0.06	99.90	0.74	99.66
						Requires training					
Food-101	TipAdaptorF	0.00	99.94	0.03	99.95	3.16	98.77	0.05	99.91	0.81	99.64
	CoOp	0.01	99.97	0.00	99.98	1.45	99.68	0.00	99.97	0.36	99.90
	CoCoOp	0.00	99.98	0.00	99.98	1.97	99.51	0.01	99.97	0.49	99.86
						Training not required					
	ZOCLIP	0.03	99.99	0.14	99.96	0.12	99.95	0.00	100.00	0.07	99.97
	TipAdaptor	0.01	100.00	0.07	99.98	0.07	99.99	0.00	100.00	0.04	99.99
						Requires training					
Oxford-Pets	TipAdaptorF	0.02	100.00	0.07	99.98	0.09	99.98	0.00	100.00	0.04	99.99
	CoOp	0.02	100.00	0.18	99.97	0.25	99.92	0.00	100.00	0.11	99.97
	CoCoOp	0.03	99.99	0.19	99.96	0.11	99.96	0.00	100.00	0.08	99.98
						Training not required					
	ZOCLIP	0.02	99.99	0.24	99.94	0.00	100.00	0.00	100.00	0.07	99.98
	TipAdaptor	0.01	100.00	0.08	99.98	0.00	100.00	0.00	100.00	0.02	100.00
						Requires training					
Stanford-Cars	TipAdaptorF	0.01	100.00	0.06	99.98	0.00	100.00	0.00	100.00	0.02	100.00
	CoOp	0.01	100.00	0.07	99.97	0.00	100.00	0.00	100.00	0.02	99.99
	CoCoOp	0.01	100.00	0.07	99.97	0.00	100.00	0.00	100.00	0.02	99.99
						Training not required					
	ZOCLIP	32.03	94.06	33.01	93.39	54.66	89.29	32.14	94.30	37.96	92.76
	TipAdaptor	9.69	98.07	11.25	97.84	20.90	96.68	13.62	97.72	13.86	97.58
						Requires training					
Caltech-101	TipAdaptorF	10.20	97.76	11.60	97.42	23.32	95.54	14.01	97.36	14.78	97.02
	CoOp	5.53	98.56	9.88	97.50	13.10	97.10	4.89	98.76	8.35	97.98
	CoCoOp	2.86	99.19	6.42	98.37	8.81	98.09	5.68	98.68	**5.94**	**98.58**

Caltech-101[1], parameter-efficient fine-tuning still significantly improves the OOD detection performance on average compared to `ZOCLIP`. In particular, `CoCoOp` yields the best performance among other adaptation methods on Caltech-101 (ID). It achieves an average FPR95 of 5.94% using S_{MS}, improving by 32.02% over `ZOCLIP`. While prior works suggest that parameter-efficient fine-tuning methods improve ID accuracy on few-shot datasets, our results complement their findings and show that fine-tuning also improves the OOD detection performance with proper OOD scoring functions.

Effects of OOD scoring functions. We investigate the effect of OOD scoring functions under fine-tuned vision-language models. In Table 2, we contrast the OOD detection performance using MCM (Ming et al., 2022) vs. MS on Caltech-101 (ID). Our findings suggest that: (1) S_{MCM} performs on par with S_{MS} for fine-grained ID tasks across a wide range of adaptation methods (Table 3). (2) However, when ID contains diverse categories, utilizing S_{MCM} generally leads to better performance compared to using S_{MS} for most adaptation methods (Table 2). (3) In particular, prompt learning methods such as `CoCoOp` demonstrate very competitive results with both OOD scores (an average FPR95 of 5.02% with S_{MCM} and 5.94% with S_{MS} in Table 2).

Effects of softmax scaling. Previously, Ming et al. (2022) observed that the commonly used maximum softmax score (S_{MSP}) is suboptimal for zero-shot OOD detection with vision-language models. We investigate whether MSP is competitive for OOD detection with fine-tuned models. To better illustrate the effects, we plot the score distributions for Stanford-Cars (ID) vs. SUN (OOD) in Figure 2 when the model is fine-tuned with `CoOp`, `CoCoOp`, and `TipAdaptorF` respectively. For each fine-tuning method, we can clearly see that the S_{MS} leads to superior ID-OOD separability, while S_{MSP} displays significant overlapping. Quantitatively, compared to S_{MSP}, the average FPR95 is significantly decreased with S_{MS} (Table B4). Our findings highlight that directly

[1]Similar trends also hold for ImageNet-1k as ID.

Table 2 OOD detection performance with S_{MS} and S_{MCM} score when the ID dataset contains diverse categories. Prompt learning methods display clear advantages over zero-shot models. The results are based on Caltech-101 (ID).

OOD Score	Method	SUN		Places		Textures		iNaturalist		Average	
		FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑
S_{MS}	ZOCLIP	32.03	94.06	33.01	93.39	54.66	89.29	32.14	94.30	37.96	92.76
	TipAdaptor	9.69	98.07	11.25	97.84	20.90	96.68	13.62	97.72	13.86	97.58
	TipAdaptorF	10.20	97.76	11.60	97.42	23.32	95.54	14.01	97.36	14.78	97.02
	CoOp	5.53	98.56	9.88	97.50	13.10	97.10	4.89	98.76	8.35	97.98
	CoCoOp	2.86	99.19	6.42	98.37	8.81	98.09	5.68	98.68	**5.94**	**98.58**
S_{MCM}	ZOCLIP	14.83	97.20	20.45	96.00	14.98	97.35	10.84	97.76	15.28	97.08
	TipAdaptor	5.12	98.83	8.05	98.34	4.65	99.05	6.94	98.77	6.19	**98.75**
	TipAdaptorF	4.83	98.79	8.09	98.07	6.41	98.11	4.94	98.98	6.07	98.49
	CoOp	3.62	99.01	8.15	97.89	6.29	98.62	7.57	98.35	6.41	98.47
	CoCoOp	4.26	98.94	6.76	98.00	4.33	98.88	4.71	98.68	**5.02**	98.62

Table 3 OOD detection performance based on S_{MCM} score.

ID Dataset	Method	SUN		Places		Textures		iNaturalist		Average	
		FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑
Food-101	ZOCLIP	1.75	99.46	2.04	99.35	5.54	98.05	2.80	99.17	3.03	99.01
	TipAdaptor	0.63	99.75	0.64	99.71	3.76	98.59	1.32	99.55	**1.59**	**99.40**
	TipAdaptorF	1.77	99.57	1.57	99.53	4.43	98.34	1.85	99.40	2.40	99.21
	CoOp	2.00	99.46	1.60	99.47	5.85	98.39	1.37	99.54	2.71	99.22
	CoCoOp	1.06	99.69	1.01	99.63	4.17	98.42	1.40	99.53	1.91	99.32
Oxford-Pets	ZOCLIP	1.18	99.73	3.37	99.28	1.37	99.73	6.17	98.84	3.02	99.40
	TipAdaptor	0.05	99.97	0.62	99.87	0.17	99.96	0.11	99.87	**0.24**	**99.92**
	TipAdaptorF	0.48	99.89	1.74	99.66	0.43	99.88	0.93	99.53	0.90	99.74
	CoOp	0.06	99.96	0.55	99.85	0.39	99.90	2.07	99.37	0.77	99.77
	CoCoOp	0.08	99.95	0.53	99.85	0.25	99.91	1.12	99.55	0.49	99.82
Stanford-Cars	ZOCLIP	0.02	99.96	0.31	99.89	0.02	99.96	0.10	99.74	0.11	99.89
	TipAdaptor	0.01	99.98	0.11	99.94	0.00	99.97	0.00	99.84	**0.03**	99.93
	TipAdaptorF	0.03	99.98	0.19	99.94	0.00	99.99	0.00	99.93	0.06	**99.96**
	CoOp	0.01	99.98	0.17	99.93	0.00	99.98	0.02	99.84	0.05	99.93
	CoCoOp	0.02	99.98	0.15	99.93	0.00	99.97	0.00	99.87	0.04	99.94
Caltech-101	ZOCLIP	14.83	97.20	20.45	96.00	14.98	97.35	10.84	97.76	15.28	97.08
	TipAdaptor	5.12	98.83	8.05	98.34	4.65	99.05	6.94	98.77	6.19	**98.75**
	TipAdaptorF	4.83	98.79	8.09	98.07	6.41	98.11	4.94	98.98	6.07	98.49
	CoOp	3.62	99.01	8.15	97.89	6.29	98.62	7.57	98.35	6.41	98.47
	CoCoOp	4.26	98.94	6.76	98.00	4.33	98.88	4.71	98.68	**5.02**	98.62

applying MSP is not competitive for fine-tuned vision-language models.

4.3 Delving into parameter-efficient fine-tuning for OOD detection

The impact of fine-tuning on feature geometry. To better understand how fine-tuning leads to improved OOD detection performance, we examine the geometry of the feature representations. For illustration, we use the simple S_{MS} score as it provides an intuitive geometric interpretation. For each test input, S_{MS} captures the angular distance between its visual features and the closest ID prototype. Figure 3 shows S_{MS} for ID and each OOD test dataset, where radians are converted to degrees for better readability. Intuitively, one desires to learn compact ID clusters such that ID inputs are closer to the nearest ID prototypes than OOD inputs. We illustrate the effects of prompt learning in Figure 4. Compared to zero-shot CLIP, `CoOp` and `CoCoOp` decrease the angular distance for ID inputs to the nearest concept prototype while simultaneously increasing the angular distance for OOD inputs. In particular, `CoCoOp` decreases the angular distance for ID inputs more significantly, resulting in better ID-OOD separability. Although prompt learning methods introduce perturbations to the feature space, the overall effect is modest, with only a slight deviation of a few degrees from the pre-trained model[2]. Nonetheless, these perturbations play a crucial role in enhancing both ID classification and OOD detection performance.

Exploring prompt learning for OOD detection on challenging large-scale benchmarks

[2]Similar observations can also be verified for adaptor-based methods.

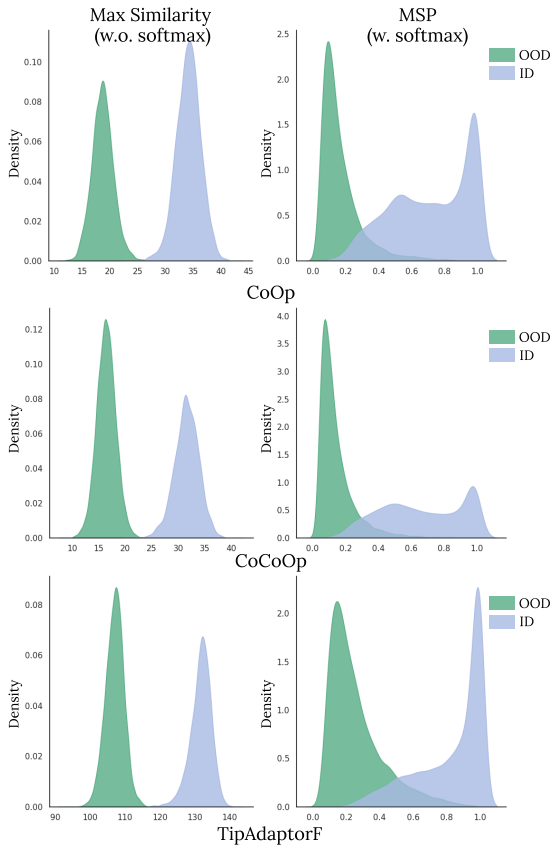

Fig. 2 The impact of softmax scaling. We use StanfordCars (ID) vs. SUN (OOD) for illustration. Applying softmax scaling significantly decreases ID-OOD separability for `CoOp` (top row), `CoCoOp` (second row), and `TipAdaptorF` (last row), resulting in worse OOD detection performance.

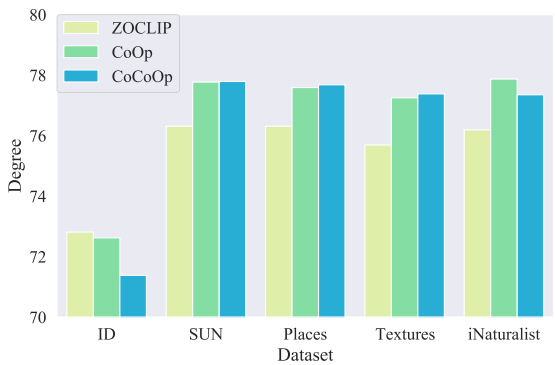

Fig. 3 Average S_{MS} for ID (Caltech-101) and OOD test sets. Prompt learning methods decrease the angular distance for ID inputs while increasing the angular distance for OOD inputs to the nearest concept prototype, leading to better ID-OOD separability (Figure 4).

In previous sections, we show that prompt learning with both S_{MS} and S_{MCM} scores display competitive performance. Next, we consider a more challenging large-scale benchmark ImageNet-1k (ID). The results in FPR95 and AUROC are shown in Figure 5 and Figure 6. While S_{MS} outperforms S_{MSP} score, we can clearly see that S_{MCM} is particularly advantageous compared to the simpler S_{MS} baseline. In particular, S_{MCM} outperforms S_{MS} by 7.44% in FPR95 averaged across the four OOD test sets. Moreover, `CoOp` with S_{MCM} achieves an average FPR95 of 37.74%

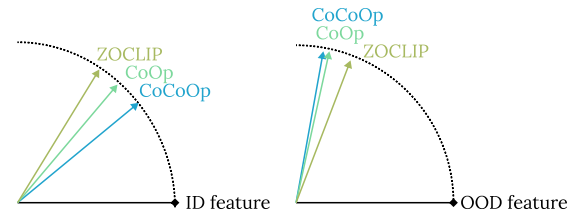

Fig. 4 Illustration of how prompt learning methods impact the hyperspherical features. Left: feature of an ID sample and its nearest ID prototype; Right: feature of an OOD sample and its nearest ID prototype.

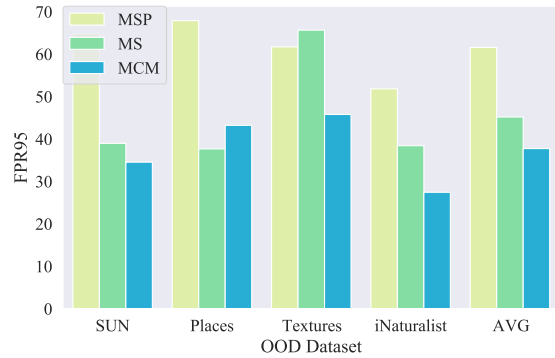

Fig. 5 OOD detection performance (FPR95) on ImageNet-1k (ID). Using S_{MCM} score leads to significant improvement over S_{MSP}.

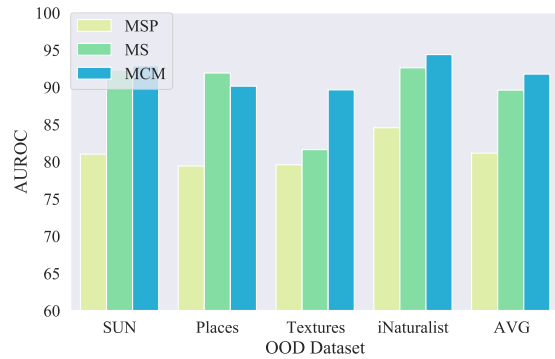

Fig. 6 OOD detection performance (AUROC) on ImageNet-1k (ID). The trend is consistent with Fig 5.

on the benchmark, surpassing the zero-shot performance of the large backbone CLIP-L/14 model which has an FPR95 of 38.17% (Ming et al., 2022). These results further demonstrate the effectiveness of S_{MCM} in CLIP-based prompt learning for challenging scenarios.

The impact of shots. We investigate the impact of shots for `CoOp` and `CoCoOp` with various OOD detection scores. The results are shown in Figure 7 and Figure 8, where each point represents the average FPR95 over the four OOD test sets. We highlight two key findings. First, the OOD detection performance with both S_{MS} and S_{MCM} score improves as the number of shots increases. This trend is consistent with the ID classification accuracy reported in Zhou et al. (2022b), suggesting that using a suitable OOD uncertainty score can enhance the representation quality as more data is incorporated during prompt learning. Second, the performance of S_{MCM} is promising even with a low number of shots, demonstrating its effectiveness in resource-constrained settings.

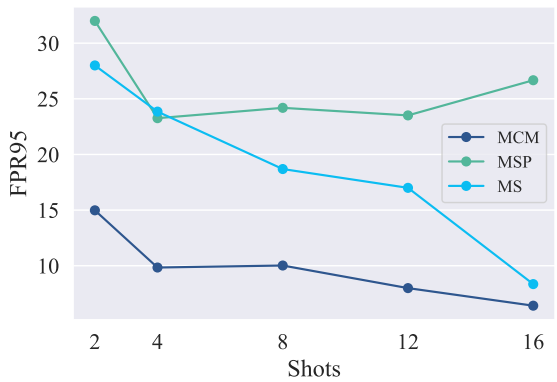

Fig. 7 The effects of shots for `CoOp` with various OOD detection scores on Caltech-101 (ID). The performance is averaged over the four OOD test sets.

The impact of backbone architecture. We conduct another ablation study on the impact of model architectures. We consider CLIP with ResNet backbones (N50, RN101) and ViT backbones (CLIP-B/32, CLIP-L/14), where the vision encoder is based on ViT-B/32 and ViT-L/14, respectively. We train with `CoOp` with hyperparameters following the original implementation for each architecture (Zhou et al., 2022b). We evaluate the models using S_{MSP}, S_{MS}, and S_{MCM} score and summarize the results in Table 4 and Table 5. Interestingly, compared to S_{MSP}, S_{MS} brings more significant improvements under ViT backbones than ResNet backbones. In contrast,

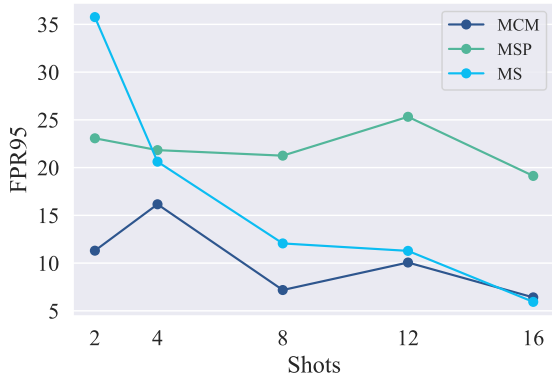

Fig. 8 The effects of shots for `CoCoOp` with various OOD detection scores on Caltech-101 (ID). The performance is averaged over the four OOD test sets.

Table 4 The impact of model architecture on ResNet backbones with CoOp on Caltech-101 (ID).

Arch	Score	OOD Dataset	FPR95↓	AUROC↑
RN50	S_{MSP}	SUN	29.93	93.95
		Places	37.64	91.96
		Textures	35.69	93.58
		iNaturalist	43.42	91.27
		AVG	36.67	92.69
	S_{MS}	SUN	6.02	98.45
		Places	9.02	97.79
		Textures	23.17	95.25
		iNaturalist	12.39	97.37
		AVG	12.65	97.22
	S_{MCM}	SUN	8.56	98.03
		Places	17.02	95.88
		Textures	12.09	97.56
		iNaturalist	21.00	95.93
		AVG	14.67	96.85
RN101	S_{MSP}	SUN	23.60	95.20
		Places	29.37	93.94
		Textures	21.29	96.24
		iNaturalist	34.18	94.05
		AVG	27.11	94.86
	S_{MS}	SUN	19.08	96.56
		Places	20.79	96.25
		Textures	36.97	94.39
		iNaturalist	30.89	95.41
		AVG	26.93	95.65
	S_{MCM}	SUN	6.19	98.42
		Places	11.57	97.16
		Textures	5.83	98.49
		iNaturalist	10.56	97.69
		AVG	8.54	97.94

Table 5 The impact of model architecture on ViT backbones with CoOp on Caltech-101 (ID).

Arch	Score	OOD Dataset	FPR95↓	AUROC↑
CLIP-B/32	S_{MSP}	SUN	24.20	96.02
		Places	27.94	94.99
		Textures	24.54	96.09
		iNaturalist	28.90	95.37
		AVG	26.40	95.62
	S_{MS}	SUN	13.81	97.41
		Places	16.49	96.48
		Textures	25.23	95.24
		iNaturalist	13.00	97.60
		AVG	17.13	96.68
	S_{MCM}	SUN	4.06	98.92
		Places	7.31	98.01
		Textures	4.61	98.81
		iNaturalist	8.70	98.17
		AVG	6.17	98.48
CLIP-L/14	S_{MSP}	SUN	7.73	98.36
		Places	10.96	97.71
		Textures	19.18	96.60
		iNaturalist	11.33	97.71
		AVG	15.85	97.41
	S_{MS}	SUN	13.81	97.41
		Places	16.49	96.48
		Textures	25.23	95.24
		iNaturalist	13.00	97.60
		AVG	12.30	97.59
	S_{MCM}	SUN	2.15	99.33
		Places	5.60	98.30
		Textures	2.32	99.31
		iNaturalist	3.94	99.06
		AVG	3.50	99.00

S_{MCM} score consistently demonstrates competitive performance for all the architectures considered. For instance, with CLIP-B/32, S_{MCM} achieves an average FPR95 of 6.17%, a 20.23% improvement over the S_{MSP} baseline. We observe similar improvements for RN101 (18.57%) and RN50 (22%). Moreover, larger backbones lead to superior performance when fixing the OOD detection score as MCM. For example, with CLIP-L/14, the average FPR95 is improved by 11.17% compared to RN50 and 2.67% compared to CLIP-B/32. A similar trend has been shown for ID classification (Radford et al., 2021), where larger models yield better feature representation.

5 Related works

Parameter-efficient fine-tuning of vision-language models. Large-scale vision-language models have shown impressive performance on various downstream tasks (Radford et al., 2021; Jia et al., 2021; Yao et al., 2021; Li et al., 2022). These models learn transferable feature representations via pre-training on web-scale heterogeneous datasets. However, as downstream datasets can have a limited number of samples, adapting these large models in a parameter and data-efficient manner is crucial for effective knowledge transfer. Recent works propose various ways to tackle this challenge. Zhou et al. (2022b) propose to tune a set of soft prompts (Li and Liang, 2021; Lester et al., 2021) while freezing the encoders of CLIP. Zhou et al. (2022a) aims to improve the generalization ability of CoOp by introducing a meta-network that learns input-dependent tokens. Huang et al. (2022) propose to learn prompts in an unsupervised manner while TPT (Manli et al., 2022) uses test-time prompt tuning to learn adaptive prompts on the fly. Beyond textual prompt learning, Bahng et al. (2022) propose to tune visual prompts for CLIP-based fine-tuning. Another line of work focuses on adaptor-style fine-tuning, where instead of tuning prompts, the feature embedding is directly optimized using an adaptor module (Gao et al., 2021; Zhang et al., 2022; Udandarao et al., 2023). Prior works demonstrate significant improvement over zero-shot CLIP for few-shot ID classification and OOD generalization where OOD labels are given. However, it is unclear how reliable these parameter-efficient fine-tuning methods are for OOD detection tasks. Our work bridges this gap and explores how fine-tuning impacts OOD detection for few-shot downstream datasets.

OOD detection with vision-language representations. A plethora of OOD detection methods have been proposed on visual inputs (Lee et al., 2018; Liang et al., 2018; Hendrycks et al., 2019; Tack et al., 2020; Sun et al., 2022; Ming et al., 2022; Du et al., 2022; Wang et al., 2022; Ming et al., 2023). With the rise of large-scale pre-trained models on vision language inputs, an increasing number of works utilize textual information for visual OOD detection and demonstrate promising performance. Fort et al. (2021) propose a scheme where pre-trained CLIP models are provided with candidate OOD labels for each target dataset, and show that the output probabilities summed over the OOD labels effectively capture OOD uncertainty. Without the assumption of OOD labels, Esmaeilpour et al. (2022) propose to train a decoder based on the visual encoder of

CLIP to generate candidate labels for OOD detection. However, training a high-quality decoder incurs significant computational costs and requires extra data. While both Esmaeilpour et al. (2022) and Radford et al. (2021) focus on small-scale inputs, Ming et al. (2022) propose an OOD label-free method MCM which demonstrates promising results on a wide range of large-scale and challenging tasks (Ming et al., 2022). However, Ming et al. (2022) only investigate pre-trained CLIP models. For multi-modal OOD detection benchmarks, Bitterwolf et al. (2023) curate a new OOD test set for ImageNet-1k while Gu et al. (2023) provide new OOD datasets for document understanding. In contrast, our work focuses on the impact of parameter-efficient fine-tuning methods for OOD detection in few-shot downstream tasks, which has not been explored.

6 Conclusion

In this paper, we provide a timely study on the impact of parameter-efficient fine-tuning methods for OOD detection with large vision-language models. We focus on the few-shot setting without access to OOD labels, which has been largely unexplored in the literature. We show that parameter-efficient fine-tuning methods can improve both ID and OOD performance when coupled with a proper OOD score, with prompt learning-based methods showing the strongest performance under the MCM score. We analyze the feature space and provide insights into the effectiveness of such methods through the lens of multi-modal concept matching. We hope our findings will inspire and motivate future research on designing reliable fine-tuning methods for large vision-language models.

Acknowledgement

Support for this research was provided by American Family Insurance through a research partnership with the University of Wisconsin–Madison's Data Science Institute, Office of Naval Research under award number N00014-23-1-2643, AFOSR Young Investigator Program under award number FA9550-23-1-0184, and National Science Foundation (NSF) CAREER Award No. IIS-2237037.

Appendix A Dataset Details

Details on ID and OOD dataset construction For ID datasets, we follow the same construction as in previous works (Zhang et al., 2022; Zhou et al., 2022a,b). Detailed instructions on dataset installation can be found in https://github.com/KaiyangZhou/CoOp/blob/main/DATASETS.md. For OOD datasets, Huang and Li (2021) curate a collection of subsets from iNaturalist Van Horn et al. (2018), SUN Xiao et al. (2010), Places Zhou et al. (2017), and Texture Cimpoi et al. (2014) as large-scale OOD datasets for ImageNet-1k, where the classes of the test sets do not overlap with ImageNet-1k. Detailed instructions can be found in https://github.com/deeplearning-wisc/large_scale_ood.

Appendix B Additional Results

B.1 ID accuracy

While we primarily focus on the OOD detection performance of CLIP-based fine-tuning methods, we present the results of the ID accuracy for each dataset based on CLIP-B/16 in Table B1 for completeness. Further results on the ID accuracy with various datasets and architectures can be seen in Zhou et al. (2022a), Zhou et al. (2022b), and Zhang et al. (2022).

B.2 OOD detection performance based on visual features alone

In this section, we explore several commonly used OOD detection scores solely based on the visual branch of CLIP models. Specifically, we consider the Mahalanobis score (Lee et al., 2018) on the penultimate layer of the visual encoder and MSP (Hendrycks and Gimpel, 2017), Energy (Liu et al., 2020), and KL Matching (Hendrycks et al., 2022) scores on the logit layer after linear probing the visual encoder. The results are summarized in Table B2, based on 16-shot Caltech-101 (ID). We can see that the Mahalanobis score does not yield promising performance because 1) the feature embeddings from the visual encoder of CLIP may not follow class-conditional Gaussian distributions, 2) it is challenging to estimate the mean

Table B1 ID accuracy on the downstream datasets for CLIP-based fine-tuning methods with CLIP-B/16.

ID Dataset	Method	ID Acc
Caltech-101	ZOCLIP	92.90
	TipAdaptor	95.01
	TipAdaptorF	95.66
	CoOp	95.30
	CoCoOp	95.00
Food-101	ZOCLIP	86.10
	TipAdaptor	86.49
	TipAdaptorF	87.43
	CoOp	85.50
	CoCoOp	87.30
Stanford-Cars	ZOCLIP	65.27
	TipAdaptor	75.29
	TipAdaptorF	83.40
	CoOp	78.50
	CoCoOp	72.30
Oxford-Pets	ZOCLIP	89.10
	TipAdaptor	91.85
	TipAdaptorF	92.91
	CoOp	93.40
	CoCoOp	93.30
ImageNet-1k	ZOCLIP	68.77
	TipAdaptor	70.26
	TipAdaptorF	73.70
	CoOp	71.63
	CoCoOp	71.20

and especially covariance matrix when the number of samples is much smaller than the feature dimension in the few-shot setting. On the other hand, the OOD scores based on fine-tuned logit layer result in worse performance compared to the MCM score. One major reason is that fine-tuning CLIP in the few-shot setting is prone to overfitting the downstream ID dataset, making the model less reliable. This further highlights the importance of choosing OOD detection scores fitted to parameter-efficient fine-tuning methods.

B.3 Additional results on ImageNet-1k

In this section, we consider two additional OOD test sets ImageNet-O (Hendrycks et al., 2021) and OpenImage-O (Wang et al., 2022) for ImageNet-1k (ID). OpenImage-O is a subset curated from the test set of OpenImage-V3 (Krasin et al., 2017) containing a diverse set of categories. ImageNet-O is a challenging OOD dataset that contains naturally adversarial examples for ImageNet-1k. The results are shown in Table B3. The model (CLIP-B/16) is trained with CoOp. We can see that: 1) The performance on ImageNet-O is generally worse than the rest of OOD test sets (iNaturalist, Textures, SUN, Places) in Section 4.3, suggesting that this task remains challenging in the context of few-shot prompt learning. 2) MCM score still performs the best compared to MS and MSP on both OOD test sets, consistent with our previous observations, which further highlights the importance of softmax and temperature scaling for OOD detection with fine-tuning.

Table B2 Additional results for OOD scores based on visual encoder only. ID dataset is Caltech-101 (16 shot).

OOD Score	OOD Dataset	FPR95↓	AUROC↑
Maha	SUN	34.15	95.20
	Places	20.50	96.21
	Textures	64.10	92.43
	iNaturalist	66.62	92.97
	AVG	46.34	94.20
Energy	SUN	15.02	97.05
	Places	21.10	95.75
	Textures	15.60	97.00
	iNaturalist	33.77	95.49
	AVG	21.37	96.32
KL Matching	SUN	4.56	98.21
	Places	8.92	97.52
	Textures	42.64	94.47
	iNaturalist	9.70	97.35
	AVG	16.46	96.89
MSP	SUN	16.23	96.59
	Places	20.98	95.97
	Textures	7.15	98.33
	iNaturalist	11.79	97.31
	AVG	14.04	97.05

Table B3 OOD detection performance on two OOD additional test sets for ImageNet-1k (ID). We train CLIP-B/16 with CoOp.

OOD Dataset	OOD Score	FPR95↓	AUROC↑
ImageNet-O	S_{MSP}	77.20	74.01
	S_{MS}	70.75	82.30
	S_{MCM}	61.50	84.13
OpenImage-O	S_{MSP}	56.89	83.73
	S_{MS}	39.18	91.48
	S_{MCM}	36.68	92.76

B.4 Alternative OOD scores

In this section, we investigate the performance with several alternative OOD scoring functions based on the cosine similarities of input $\boldsymbol{x}$ with the k-th label $s_k(\boldsymbol{x})$, $k \in \{1, 2, ..., K\}$ (defined in Section 3.2). Specifically, we consider the energy and the KL matching score for each adaptation method and summarize the results based on Caltech-101 (ID) in Table B5. We observe that 1) using the energy score, all adaptation methods significantly enhance the performance over the zero-shot baseline (ZOCLIP). 2) the general performance vastly improves when utilizing the KL Matching score. However, even the highest achieved performance (FPR95 at 7.91 with CoCoOp) falls short when compared to the MCM score (FPR95 at 5.02 with CoCoOp).

References

Bahng, H., A. Jahanian, S. Sankaranarayanan, and P. Isola. 2022. Exploring visual prompts for adapting large-scale models. *arXiv preprint arXiv:2203.17274* .

Bitterwolf, J., M. Mueller, and M. Hein 2023. In or out? fixing imagenet out-of-distribution detection evaluation. In *International Conference on Machine Learning (ICML)*.

Bossard, L., M. Guillaumin, and L. Van Gool 2014. Food-101 – mining discriminative components with random forests. In *The European Conference on Computer Vision (ECCV)*.

Cimpoi, M., S. Maji, I. Kokkinos, S. Mohamed, and A. Vedaldi 2014. Describing textures in the wild. In *The IEEE / CVF Computer Vision and Pattern Recognition Conference (CVPR)*.

Deng, J., W. Dong, R. Socher, L.J. Li, K. Li, and L. Fei-Fei 2009. Imagenet: A large-scale hierarchical image database. In *The IEEE / CVF Computer Vision and Pattern Recognition Conference (CVPR)*.

Dosovitskiy, A., L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*.

Du, X., G. Gozum, Y. Ming, and Y. Li 2022. Siren: Shaping representations for detecting out-of-distribution objects. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Esmaeilpour, S., B. Liu, E. Robertson, and L. Shu 2022. Zero-shot open set detection by extending clip. In *The AAAI Conference on Artificial Intelligence (AAAI)*.

Fort, S., J. Ren, and B. Lakshminarayanan 2021. Exploring the limits of out-of-distribution detection. In *Conference on Neural Information Processing Systems (NeurIPS)*.

Gao, P., S. Geng, R. Zhang, T. Ma, R. Fang, Y. Zhang, H. Li, and Y. Qiao. 2021. Clip-adapter: Better vision-language models with feature adapters. *arXiv preprint arXiv:2110.04544* .

Gu, J., Y. Ming, Y. Zhou, J. Kuen, V.I. Morariu, H. Zhao, R. Zhang, N. Barmpalios, A. Liu, Y. Li, T. Sun, and A. Nenkova 2023. A critical analysis of out-of-distribution detection for document understanding. In *Findings of Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Hendrycks, D. and K. Gimpel 2017. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *International Conference on Learning Representations (ICLR)*.

Hendrycks, D., M. Mazeika, S. Kadavath, and D. Song 2019. Using self-supervised learning can improve model robustness and uncertainty. In *Conference on Neural Information Processing Systems (NeurIPS)*, pp. 15663–15674.

Hendrycks, D., M. Mazeika, S. Kadavath, and D. Song 2022. Scaling out-of-distribution detection for real-world settings. In *International Conference on Machine Learning (ICML)*.

Table B4 OOD detection performance based on S_{MSP} score. The average performance for most adaptation methods is much worse than using S_{MS} (Table 1) and S_{MCM} (Table 3).

ID Dataset	Method	SUN		Places		Textures		iNaturalist		Average	
		FPR95↓	**AUROC↑**	**FPR95↓**	**AUROC↑**	**FPR95↓**	**AUROC↑**	**FPR95↓**	**AUROC↑**	**FPR95↓**	**AUROC↑**
		FPR95	AUROC	FPR95	AUROC	FPR95	AUROC	FPR95	AUROC	FPR95	AUROC
Food-101	ZOCLIP	11.48	97.76	13.11	97.48	15.04	96.08	16.65	96.73	14.07	97.01
	TipAdaptor	7.32	98.51	9.03	98.31	11.88	96.94	14.47	97.21	**10.68**	**97.74**
	TipAdaptorF	15.08	97.26	15.38	97.24	17.57	95.99	20.95	96.18	17.25	96.67
	CoOp	19.66	96.20	21.15	95.95	28.33	93.62	23.80	95.51	23.23	95.32
	CoCoOp	8.67	98.28	10.56	98.03	14.77	96.23	14.33	97.26	12.08	97.45
Oxford-Pets	ZOCLIP	24.67	94.72	28.54	93.71	19.01	96.42	39.77	93.01	28.00	94.47
	TipAdaptor	15.66	97.11	18.83	96.45	12.50	97.92	25.19	95.90	18.04	96.84
	TipAdaptorF	16.79	96.77	20.33	96.04	12.22	97.90	26.62	95.80	18.99	96.63
	CoOp	8.46	98.50	10.75	98.13	11.21	98.09	32.13	94.08	15.64	97.20
	CoCoOp	9.06	98.31	10.43	98.13	7.39	98.70	27.97	95.11	**13.71**	**97.56**
Stanford-Cars	ZOCLIP	6.99	98.49	10.33	97.68	8.24	98.39	32.85	92.56	14.60	96.78
	TipAdaptor	1.94	99.58	3.30	99.31	1.97	99.56	12.52	97.80	**4.93**	**99.06**
	TipAdaptorF	15.39	97.19	14.01	97.32	8.39	98.49	21.88	95.90	14.92	97.22
	CoOp	9.88	98.05	14.07	97.12	10.71	97.71	36.73	91.51	17.85	96.10
	CoCoOp	9.99	97.81	11.87	97.15	10.46	97.69	31.58	92.59	15.97	96.31
Caltech-101	ZOCLIP	16.17	96.47	22.45	94.96	17.89	96.33	15.01	96.96	17.88	96.18
	TipAdaptor	12.98	97.40	17.79	96.77	13.74	97.72	20.08	96.65	**16.15**	**97.13**
	TipAdaptorF	17.94	96.68	22.92	95.74	15.16	97.40	24.18	96.01	20.05	96.46
	CoOp	24.07	96.11	29.91	94.59	26.29	95.72	26.35	95.92	26.66	95.58
	CoCoOp	14.92	97.32	20.67	95.91	19.20	96.56	21.74	96.33	19.13	96.53

Table B5 Comparison with additional OOD scores on Caltech-101 (ID). S_{KL} stands for the KL matching score (Hendrycks et al., 2022) and S_{Energy} denotes the energy score (Liu et al., 2020).

OOD Score	Method	SUN		Places		Textures		iNaturalist		Average	
		FPR95↓	**AUROC↑**	**FPR95↓**	**AUROC↑**	**FPR95↓**	**AUROC↑**	**FPR95↓**	**AUROC↑**	**FPR95↓**	**AUROC↑**
S_{MS}	ZOCLIP	32.03	94.06	33.01	93.39	54.66	89.29	32.14	94.30	37.96	92.76
	TipAdaptor	9.69	98.07	11.25	97.84	20.90	96.68	13.62	97.72	13.86	97.58
	TipAdaptorF	10.20	97.76	11.60	97.42	23.32	95.54	14.01	97.36	14.78	97.02
	CoOp	5.53	98.56	9.88	97.50	13.10	97.10	4.89	98.76	8.35	97.98
	CoCoOp	2.86	99.19	6.42	98.37	8.81	98.09	5.68	98.68	5.94	98.58
S_{MCM}	ZOCLIP	14.83	97.20	20.45	96.00	14.98	97.35	10.84	97.76	15.28	97.08
	TipAdaptor	5.12	98.83	8.05	98.34	4.65	99.05	6.94	98.77	6.19	**98.75**
	TipAdaptorF	4.83	98.79	8.09	98.07	6.41	98.11	4.94	98.98	6.07	98.49
	CoOp	3.62	99.01	8.15	97.89	6.29	98.62	7.57	98.35	6.41	98.47
	CoCoOp	4.26	98.94	6.76	98.00	4.33	98.88	4.71	98.68	**5.02**	98.62
S_{Energy}	ZOCLIP	53.83	90.22	50.51	90.21	74.10	83.20	56.00	90.13	58.61	88.44
	TipAdaptor	11.71	97.72	12.20	97.61	30.48	95.73	16.42	97.30	17.70	97.09
	TipAdaptorF	11.57	97.46	11.89	97.30	29.38	94.70	16.18	96.90	17.26	96.59
	CoOp	6.58	98.29	11.16	97.20	18.19	96.32	5.92	98.53	10.46	97.59
	CoCoOp	5.22	98.87	8.80	98.13	17.30	96.87	11.28	97.95	10.65	97.95
S_{KL}	ZOCLIP	5.51	97.57	9.48	96.61	7.41	97.64	11.43	96.22	14.02	97.31
	TipAdaptor	5.54	97.63	7.69	97.13	5.74	97.96	8.00	97.37	6.74	97.52
	TipAdaptorF	8.52	96.89	13.00	95.92	7.02	98.02	10.71	97.11	9.81	96.98
	CoOp	7.15	98.06	12.37	96.60	8.74	97.62	9.33	98.00	9.40	97.57
	CoCoOp	4.07	98.95	9.61	97.59	5.30	98.77	12.67	97.57	7.91	98.22

Hendrycks, D., K. Zhao, S. Basart, J. Steinhardt, and D. Song. 2021. Natural adversarial examples. *the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* .

Huang, R. and Y. Li 2021. Mos: Towards scaling out-of-distribution detection for large semantic space. In *The IEEE / CVF Computer Vision and Pattern Recognition Conference (CVPR)*.

Huang, T., J. Chu, and F. Wei. 2022. Unsupervised prompt learning for vision-language models. *arXiv preprint arXiv:2204.03649* .

Jia, C., Y. Yang, Y. Xia, Y.T. Chen, Z. Parekh, H. Pham, Q. Le, Y.H. Sung, Z. Li, and T. Duerig 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning (ICML)*.

Krasin, I., T. Duerig, N. Alldrin, V. Ferrari, S. Abu-El-Haija, A. Kuznetsova, H. Rom, J. Uijlings, S. Popov, A. Veit, S. Belongie, V. Gomes, A. Gupta, C. Sun, G. Chechik, D. Cai, Z. Feng, D. Narayanan, and K. Murphy. 2017. Openimages: A public dataset for large-scale multi-label and multi-class image classification. *Dataset available from https://github.com/openimages* .

Krause, J., M. Stark, J. Deng, and L. Fei-Fei 2013. 3d object representations for fine-grained categorization. In *4th International IEEE Workshop on 3D Representation and Recognition (3dRR-13)*, Sydney, Australia.

Lee, K., K. Lee, H. Lee, and J. Shin 2018. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In *Conference on Neural Information Processing Systems (NeurIPS)*.

Lester, B., R. Al-Rfou, and N. Constant 2021. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the Empirical Methods in Natural Language Processing (EMNLP)*, pp. 3045–3059.

Li, X.L. and P. Liang 2021. Prefix-tuning: Optimizing continuous prompts for generation. In *59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, pp. 4582–4597.

Li, Y., F. Liang, L. Zhao, Y. Cui, W. Ouyang, J. Shao, F. Yu, and J. Yan 2022. Supervision exists everywhere: A data efficient contrastive language-image pre-training paradigm. In *International Conference on Learning Representations (ICLR)*.

Liang, S., Y. Li, and R. Srikant 2018. Enhancing the reliability of out-of-distribution image detection in neural networks. In *International Conference on Learning Representations (ICLR)*.

Liu, W., X. Wang, J. Owens, and Y. Li 2020. Energy-based out-of-distribution detection. In *Conference on Neural Information Processing Systems (NeurIPS)*.

Loshchilov, I. and F. Hutter 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*.

Manli, S., N. Weili, H. De-An, Y. Zhiding, G. Tom, A. Anima, and X. Chaowei 2022. Test-time prompt tuning for zero-shot generalization in vision-language models. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Ming, Y., Z. Cai, J. Gu, Y. Sun, W. Li, and Y. Li 2022. Delving into out-of-distribution detection with vision-language representations. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Ming, Y., Y. Fan, and Y. Li 2022. Poem: Out-of-distribution detection with posterior sampling. In *International Conference on Machine Learning (ICML)*.

Ming, Y., Y. Sun, O. Dia, and Y. Li 2023. How to exploit hyperspherical embeddings for out-of-distribution detection? In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Ming, Y., H. Yin, and Y. Li 2022. On the impact of spurious correlation for out-of-distribution detection. In *The AAAI Conference on Artificial Intelligence (AAAI)*.

Parkhi, O.M., A. Vedaldi, A. Zisserman, and C.V. Jawahar 2012. Cats and dogs. In *The IEEE / CVF Computer Vision and Pattern Recognition Conference (CVPR)*.

Radford, A., J.W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*.

Sun, Y., Y. Ming, X. Zhu, and Y. Li 2022. Out-of-distribution detection with deep nearest neighbors. In *International Conference on Machine Learning (ICML)*.

Tack, J., S. Mo, J. Jeong, and J. Shin 2020. Csi: Novelty detection via contrastive learning on

distributionally shifted instances. In *Conference on Neural Information Processing Systems (NeurIPS)*.

Udandarao, V., A. Gupta, and S. Albanie 2023. Sus-x: Training-free name-only transfer of vision-language models. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

Van Horn, G., O. Mac Aodha, Y. Song, Y. Cui, C. Sun, A. Shepard, H. Adam, P. Perona, and S. Belongie 2018. The inaturalist species classification and detection dataset. In *The IEEE / CVF Computer Vision and Pattern Recognition Conference (CVPR)*.

Wang, H., Z. Li, L. Feng, and W. Zhang 2022. Vim: Out-of-distribution with virtual-logit matching. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4921–4930.

Xiao, J., J. Hays, K.A. Ehinger, A. Oliva, and A. Torralba 2010. Sun database: Large-scale scene recognition from abbey to zoo. In *The IEEE / CVF Computer Vision and Pattern Recognition Conference (CVPR)*.

Yang, J., K. Zhou, Y. Li, and Z. Liu. 2021. Generalized out-of-distribution detection: A survey. *arXiv preprint arXiv:2110.11334* .

Yao, L., R. Huang, L. Hou, G. Lu, M. Niu, H. Xu, X. Liang, Z. Li, X. Jiang, and C. Xu. 2021. Filip: Fine-grained interactive language-image pre-training. *International Conference on Learning Representations (ICLR)* .

Zhang, R., W. Zhang, R. Fang, P. Gao, K. Li, J. Dai, Y. Qiao, and H. Li 2022. Tip-adapter: Training-free adaption of clip for few-shot classification. In *17th European Conference on Computer Vision (ECCV)*, pp. 493–510.

Zhou, B., A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba. 2017. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)* .

Zhou, K., J. Yang, C.C. Loy, and Z. Liu 2022a. Conditional prompt learning for vision-language models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Zhou, K., J. Yang, C.C. Loy, and Z. Liu. 2022b. Learning to prompt for vision-language models. *International Journal of Computer Vision (IJCV)* .