Learning to Prompt for Vision-Language Models

Kaiyang Zhou · Jingkang Yang · Chen Change Loy · Ziwei Liu

arXiv:2109.01134v6 [cs.CV] 6 Oct 2022

Abstract Large pre-trained vision-language models like CLIP have shown great potential in learning representations that are transferable across a wide range of downstream tasks. Different from the traditional representation learning that is based mostly on discretized labels, vision-language pre-training aligns images and texts in a common feature space, which allows zero-shot transfer to a downstream task via *prompting*, i.e., classification weights are synthesized from natural language describing classes of interest. In this work, we show that a major challenge for deploying such models in practice is prompt engineering, which requires domain expertise and is extremely time-consuming—one needs to spend a significant amount of time on words tuning since a slight change in wording could have a huge impact on performance. Inspired by recent advances in prompt learning research in natural language processing (NLP), we propose *Context Optimization (CoOp)*, a simple approach specifically for adapting CLIP-like vision-language models for downstream image recognition. Concretely, CoOp models a prompt's context words with learnable vectors while the entire pre-trained parameters are kept fixed. To handle different image recognition tasks, we provide two implementations of CoOp: unified context and class-specific context. Through extensive experiments on 11 datasets, we demonstrate that CoOp requires as few as one or two shots to beat hand-crafted prompts with a decent margin and is able to gain significant improvements over prompt engineering with more shots, e.g., with 16 shots the average gain is around 15% (with the highest reaching over 45%). Despite being a learning-based approach, CoOp achieves superb domain generalization performance compared with the zero-shot model using hand-crafted prompts.

Kaiyang Zhou
S-Lab, Nanyang Technological University, Singapore
E-mail: kaiyang.zhou@ntu.edu.sg

Jingkang Yang
S-Lab, Nanyang Technological University, Singapore
E-mail: jingkang001@ntu.edu.sg

Chen Change Loy
S-Lab, Nanyang Technological University, Singapore
E-mail: ccloy@ntu.edu.sg

Ziwei Liu
S-Lab, Nanyang Technological University, Singapore
E-mail: ziwei.liu@ntu.edu.sg

1 Introduction

A common approach for building state-of-the-art visual recognition systems is to train vision models to predict for a fixed set of object categories using discrete labels (He et al., 2016; Dosovitskiy et al., 2021). From a technical point of view, this is achieved by matching image features—produced by a vision model like ResNet (He et al., 2016) or ViT (Dosovitskiy et al., 2021)—with a fixed set of weights that are seen as visual concepts and initialized randomly. Although training categories often have a textual form, such as "goldfish" or "toilet paper," they will be converted into discrete labels just for easing the computation of the cross-entropy loss, leaving the semantics encapsulated in texts largely unexploited. Such a learning paradigm limits visual recognition systems to closed-set visual concepts, making them unable to deal with new categories since additional data are required for learning a new classifier.
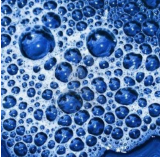
Recently, vision-language pre-training such as CLIP (Radford et al., 2021) and ALIGN (Jia et al., 2021) has emerged as a promising alternative for visual representation learning. The main idea is to align

Caltech101	Prompt	Accuracy
	a [CLASS].	82.68
	a photo of [CLASS].	80.81
	a photo of a [CLASS].	86.29
	$[V]_1 [V]_2 ... [V]_M$ [CLASS].	**91.83**

(a)

Flowers102	Prompt	Accuracy
	a photo of a [CLASS].	60.86
	a flower photo of a [CLASS].	65.81
	a photo of a [CLASS], a type of flower.	66.14
	$[V]_1 [V]_2 ... [V]_M$ [CLASS].	**94.51**

(b)

Describable Textures (DTD)	Prompt	Accuracy
	a photo of a [CLASS].	39.83
	a photo of a [CLASS] texture.	40.25
	[CLASS] texture.	42.32
	$[V]_1 [V]_2 ... [V]_M$ [CLASS].	**63.58**

(c)

EuroSAT	Prompt	Accuracy
	a photo of a [CLASS].	24.17
	a satellite photo of [CLASS].	37.46
	a centered satellite photo of [CLASS].	37.56
	$[V]_1 [V]_2 ... [V]_M$ [CLASS].	**83.53**

(d)

Fig. 1 Prompt engineering vs Context Optimization (CoOp). The former needs to use a held-out validation set for words tuning, which is inefficient; the latter automates the process and requires only a few labeled images for learning.

images and raw texts using two separate encoders—one for each modality. For instance, both CLIP and ALIGN formulate the learning objective as a contrastive loss, which pulls together images and their textual descriptions while pushes away unmatched pairs in the feature space. By pre-training at a large scale, models can learn diverse visual concepts and can readily be transferred to any downstream task through *prompting* (Radford et al., 2021; Jia et al., 2021; Fürst et al., 2021; Li et al., 2021; Singh et al., 2021; Yuan et al., 2021). In particular, for any new classification task one can first synthesize the classification weights by giving sentences describing task-relevant categories to the text encoder, and then compare with image features produced by the image encoder.

We observe that for pre-trained vision-language models, the text input, known as prompt, plays a key role in downstream datasets. However, identifying the right prompt is a non-trivial task, which often takes a significant amount of time for words tuning—a slight change in wording could make a huge difference in performance. For instance, for Caltech101 (Figure 1(a), 2nd vs 3rd prompt), adding "a" before the class token brings more than 5% increase in accuracy. Moreover, prompt engineering also requires prior knowledge about the task and ideally the language model's underlying mechanism. This is exemplified in Figure 1(b-d) where adding task-relevant context can lead to significant improvements, i.e., "flower" for Flowers102, "texture" for DTD and "satellite" for EuroSAT. Tuning the sentence structure could bring further improvements, e.g., putting "a type of flower" after the class token for Flowers102, keeping only "texture" in the context for DTD, and adding "centered" before "satellite photo" for EuroSAT. However, even with extensive tuning, the

resulting prompts are by no means guaranteed to be optimal for these downstream tasks.

Inspired by recent prompt learning research in natural language processing (NLP) (Shin et al., 2020; Jiang et al., 2020; Zhong et al., 2021), we propose a simple approach called *Context Optimization (CoOp)*[1] to automate prompt engineering, specifically for pre-trained vision-language models. Concretely, CoOp models a prompt's context words with learnable vectors, which could be initialized with either random values or pre-trained word embeddings (see Figure 2). Two implementations are provided to handle tasks of different natures: one is based on unified context, which shares the same context with all classes and works well on most categories; while the other is based on class-specific context, which learns a specific set of context tokens for each class and is found to be more suitable for some fine-grained categories. During training, we simply minimize prediction errors using the cross-entropy loss with respect to the learnable context vectors while keeping the entire pre-trained parameters fixed. The gradients can be back-propagated all the way through the text encoder, distilling the rich knowledge encoded in the parameters for learning task-relevant context.

To demonstrate the effectiveness of CoOp, we benchmark on 11 datasets, which cover a diverse set of visual recognition tasks including classification on generic objects, scenes, actions and fine-grained categories, as well as specialized tasks like recognizing textures and satellite imagery. The results show that CoOp effectively turns pre-trained vision-language models into data-efficient visual learners, requiring as few as one or two shots to beat hand-crafted prompts with a decent margin. The performance can be further boosted

---

[1] CoOp is pronounced as /ku:p/.

by using more shots, e.g., with 16 shots the margin over hand-crafted prompts averages at around 15% and reaches over 45% for the highest. CoOp also outperforms the linear probe model, which is known as a strong few-shot learning baseline (Tian et al., 2020). Furthermore, CoOp demonstrates much stronger robustness than the zero-shot model (which uses manual prompts) to domain shifts, despite being a learning-based approach.

In summary, we make the following contributions:

1. We present a timely study on the adaptation of recently proposed vision-language models in downstream applications and identify a critical problem associated with the deployment efficiency, i.e., prompt engineering.
2. To automate prompt engineering specifically for pre-trained vision-language models, we propose a simple approach based on continuous prompt learning and provide two implementations that can handle different recognition tasks.
3. We for the first time show that the proposed prompt learning-based approach outperforms both hand-crafted prompts and the linear probe model in terms of downstream transfer learning performance and robustness under domain shifts for large vision-language models.
4. We open-source our project at `https://github.com/KaiyangZhou/CoOp`.

We hope the findings together with the open-source code can inspire and facilitate future research on efficient adaptation methods for large vision-language models—an emerging topic related to democratization of foundation models (Bommasani et al., 2021) i.e., making them easier and cheaper to adapt for the wider community.

2 Related Work

2.1 Vision-Language Models

Vision-language models have recently demonstrated great potential in learning generic visual representations and allowing zero-shot transfer to a variety of downstream classification tasks via prompting (Radford et al., 2021; Jia et al., 2021; Zhang et al., 2020; Singh et al., 2021; Yuan et al., 2021).

To our knowledge, the recent developments in vision-language learning, particularly CLIP (Radford et al., 2021) and ALIGN (Jia et al., 2021), are largely driven by advances in the following three areas: i) text representation learning with Transformers (Vaswani et al., 2017), ii) large-minibatch contrastive representation learning (Chen et al., 2020; He et al., 2020; Hénaff et al., 2020), and iii) web-scale training datasets—CLIP benefits from 400 million curated image-text pairs while ALIGN exploits 1.8 billion noisy image-text pairs.

The idea of mapping images and text onto a common embedding space has been studied since nearly a decade ago (Socher et al., 2013; Frome et al., 2013; Elhoseiny et al., 2013), but with drastically different technologies. For text features extraction, early work has mainly utilized pre-trained word vectors (Socher et al., 2013; Frome et al., 2013) or the hand-crafted TF-IDF features (Elhoseiny et al., 2013; Lei Ba et al., 2015). Matching images and text features has been formulated as metric learning (Frome et al., 2013), multi-label classification (Joulin et al., 2016; Gomez et al., 2017), n-gram language learning (Li et al., 2017), and the recently proposed captioning (Desai and Johnson, 2021).

Our work is orthogonal to recent research in vision-language models, aiming to facilitate the adaptation and deployment of such models in downstream datasets.

2.2 Prompt Learning in NLP

Knowledge probing for large pre-trained language models, formally defined by Petroni et al. (2019) as "fill-in-the-blank" cloze tests, has recently sparked interest in prompt learning research in NLP (Shin et al., 2020; Jiang et al., 2020; Li and Liang, 2021; Zhong et al., 2021; Lester et al., 2021; Gao et al., 2020; Liu et al., 2021b).

The basic idea of knowledge probing is to induce pre-trained language models to generate answers given cloze-style prompts, which can benefit a number of downstream tasks, such as sentiment analysis. Jiang et al. (2020) propose to generate candidate prompts through text mining and paraphrasing, and identify the optimal ones that give the highest training accuracy. Shin et al. (2020) introduce a gradient-based approach, which searches for tokens with the largest gradient changes in the label likelihood.

Most related to our work are continuous prompt learning methods (Zhong et al., 2021; Li and Liang, 2021; Lester et al., 2021) which optimize continuous vectors in the word embedding space. A drawback of such methods compared to searching discrete tokens is the lack of a clear way to visualize what "words" are learned for the vectors. We refer readers to Liu et al. (2021a) for a comprehensive survey in the topic of prompt learning in NLP.

It is worth noting that we are the first to apply prompt learning to the adaptation of large vision-
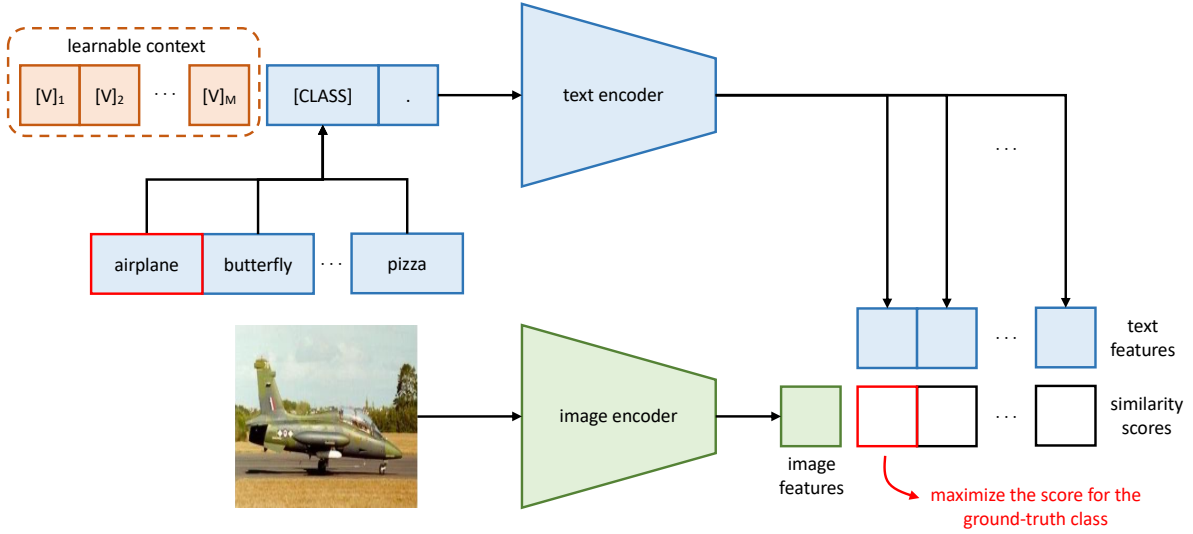
Fig. 2 Overview of Context Optimization (CoOp). The main idea is to model a prompt's context using a set of learnable vectors, which can be optimized through minimizing the classification loss. Two designs are proposed: one is unified context, which shares the same context vectors with all classes; and the other is class-specific context, which learns for each class a specific set of context vectors.

language models in computer vision—which we view as an important topic for democratizing foundation models (Bommasani et al., 2021)—and justify that prompt learning not only brings significant improvements to computer vision tasks in terms of transfer learning performance but also produces robust models that can handle domain shifts.

3 Methodology

3.1 Vision-Language Pre-training

We briefly introduce vision-language pre-training with a particular focus on CLIP (Radford et al., 2021). Our approach is applicable to broader CLIP-like vision-language models.

Models CLIP consists of two encoders, one for images and the other for text. The image encoder aims to map high-dimensional images into a low-dimensional embedding space. The architecture of the image encoder can take the form of a CNN like ResNet-50 (He et al., 2016) or a ViT (Dosovitskiy et al., 2021). On the other hand, the text encoder is built on top of a Transformer (Vaswani et al., 2017) and aims to generate text representations from natural language.

Specifically, given a sequence of words (tokens), such as "a photo of a dog," CLIP first converts each one of the token (including punctuation) into a lower-cased byte pair encoding (BPE) representation (Sennrich et al., 2016), which is essentially a unique numeric ID. The vocabulary size in CLIP is 49,152. To

facilitate minibatch processing, each text sequence is encompassed with the [SOS] and [EOS] tokens and capped at a fixed length of 77. After that, the IDs are mapped to 512-D word embedding vectors, which are then passed on to the Transformer. Finally, the features at the [EOS] token position are layer normalized and further processed by a linear projection layer.

Training CLIP is trained to align the two embedding spaces learned for images and text respectively. Specifically, the learning objective is formulated as a contrastive loss. Given a batch of image-text pairs, CLIP maximizes the cosine similarity for matched pairs while minimizes the cosine similarity for all other unmatched pairs. To learn diverse visual concepts that are more transferable to downstream tasks, CLIP's team collects a large training dataset consisting of 400 million image-text pairs.

Zero-Shot Inference Since CLIP is pre-trained to predict whether an image matches a textual description, it naturally fits zero-shot recognition. This is achieved by comparing image features with the classification weights synthesized by the text encoder, which takes as input textual descriptions specifying classes of interest. Formally, let f be image features extracted by the image encoder for an image x and $\{w_i\}_{i=1}^{K}$ a set of weight vectors generated by the text encoder. K denotes the number of classes and each w_i is derived from a prompt that could have the form of "a photo of a [CLASS]." where the class token is replaced by the specific class name, such as "cat," "dog" or "car." The

prediction probability is then computed as

$$p(y = i | \boldsymbol{x}) = \frac{\exp(\cos(\boldsymbol{w_i}, \boldsymbol{f})/\tau)}{\sum_{j=1}^{K} \exp(\cos(\boldsymbol{w_j}, \boldsymbol{f})/\tau)}, \quad (1)$$

where τ is a temperature parameter learned by CLIP and $\cos(\cdot, \cdot)$ denotes cosine similarity.

Compared with the traditional classifier learning approach where closed-set visual concepts are learned from random vectors, vision-language pre-training allows open-set visual concepts to be explored through a high-capacity text encoder, leading to a broader semantic space and in turn making the learned representations more transferable to downstream tasks.

3.2 Context Optimization

We propose Context Optimization (CoOp), which avoids manual prompt tuning by modeling context words with continuous vectors that are end-to-end learned from data while the massive pre-trained parameters are frozen. An overview is shown in Figure 2. Below we provide several different implementations.

Unified Context We first introduce the unified context version, which shares the same context with all classes. Specifically, the prompt given to the text encoder $g(\cdot)$ is designed with the following form,

$$\boldsymbol{t} = [\text{V}]_1[\text{V}]_2 \dots [\text{V}]_M[\text{CLASS}], \quad (2)$$

where each $[\text{V}]_m$ $(m \in \{1, \dots, M\})$ is a vector with the same dimension as word embeddings (i.e., 512 for CLIP), and M is a hyperparameter specifying the number of context tokens.

By forwarding a prompt $\boldsymbol{t}$ to the text encoder $g(\cdot)$, we can obtain a classification weight vector representing a visual concept (still from the [EOS] token position). The prediction probability is computed as

$$p(y = i | \boldsymbol{x}) = \frac{\exp(\cos(g(\boldsymbol{t_i}), \boldsymbol{f})/\tau)}{\sum_{j=1}^{K} \exp(\cos(g(\boldsymbol{t_j}), \boldsymbol{f})/\tau)}, \quad (3)$$

where the class token within each prompt $\boldsymbol{t_i}$ is replaced by the corresponding word embedding vector(s) of the i-th class name.

Other than placing the class token at the end of a sequence as in Equation (2), we can also put it in the middle like

$$\boldsymbol{t} = [\text{V}]_1 \dots [\text{V}]_{\frac{M}{2}}[\text{CLASS}][\text{V}]_{\frac{M}{2}+1} \dots [\text{V}]_M, \quad (4)$$

which increases flexibility for learning—the prompt is allowed to either fill the latter cells with supplementary descriptions or cut off the sentence earlier by using a termination signal such as full stop.

Class-Specific Context Another option is to design class-specific context (CSC) where context vectors are independent to each class, i.e., $[\text{V}]_1^i[\text{V}]_2^i \dots [\text{V}]_M^i \neq [\text{V}]_1^j[\text{V}]_2^j \dots [\text{V}]_M^j$ for $i \neq j$ and $i, j \in \{1, \dots, K\}$. As an alternative to unified context, we find that CSC is particularly useful for some fine-grained classification tasks.

Training is performed to minimize the standard classification loss based on the cross-entropy, and the gradients can be back-propagated all the way through the text encoder $g(\cdot)$, making use of the rich knowledge encoded in the parameters to optimize the context. The design of continuous representations also allows full exploration in the word embedding space, which facilitates the learning of task-relevant context.

3.3 Discussion

Our approach specifically addresses the emerging problem of the adaptation of recently proposed large vision-language models such as CLIP (Radford et al., 2021). There are some differences that distinguish our approach from the prompt learning methods developed in NLP for language models (e.g., GPT-3 (Brown et al., 2020)). First, the backbone architectures are clearly different for CLIP-like models and language models—the former take both visual and textual data as input and produce alignment scores used for image recognition, while the latter are tailored to handle textual data only. Second, the pre-training objectives are different: contrastive learning vs autoregressive learning. This would lead to different model behaviors and thus require different module designs.

4 Experiments

4.1 Few-Shot Learning

Datasets We select 11 publicly available image classification datasets used in CLIP: ImageNet (Deng et al., 2009), Caltech101 (Fei-Fei et al., 2004), Oxford-Pets (Parkhi et al., 2012), StanfordCars (Krause et al., 2013), Flowers102 (Nilsback and Zisserman, 2008), Food101 (Bossard et al., 2014), FGVCAircraft (Maji et al., 2013), SUN397 (Xiao et al., 2010), DTD (Cimpoi et al., 2014), EuroSAT (Helber et al., 2019) and UCF101 (Soomro et al., 2012) (see Appendix A for their statistics). These datasets constitute a comprehensive benchmark, which covers a diverse set of vision tasks including classification on generic objects, scenes, actions and fine-grained categories, as well as specialized tasks like recognizing textures and satellite imagery.
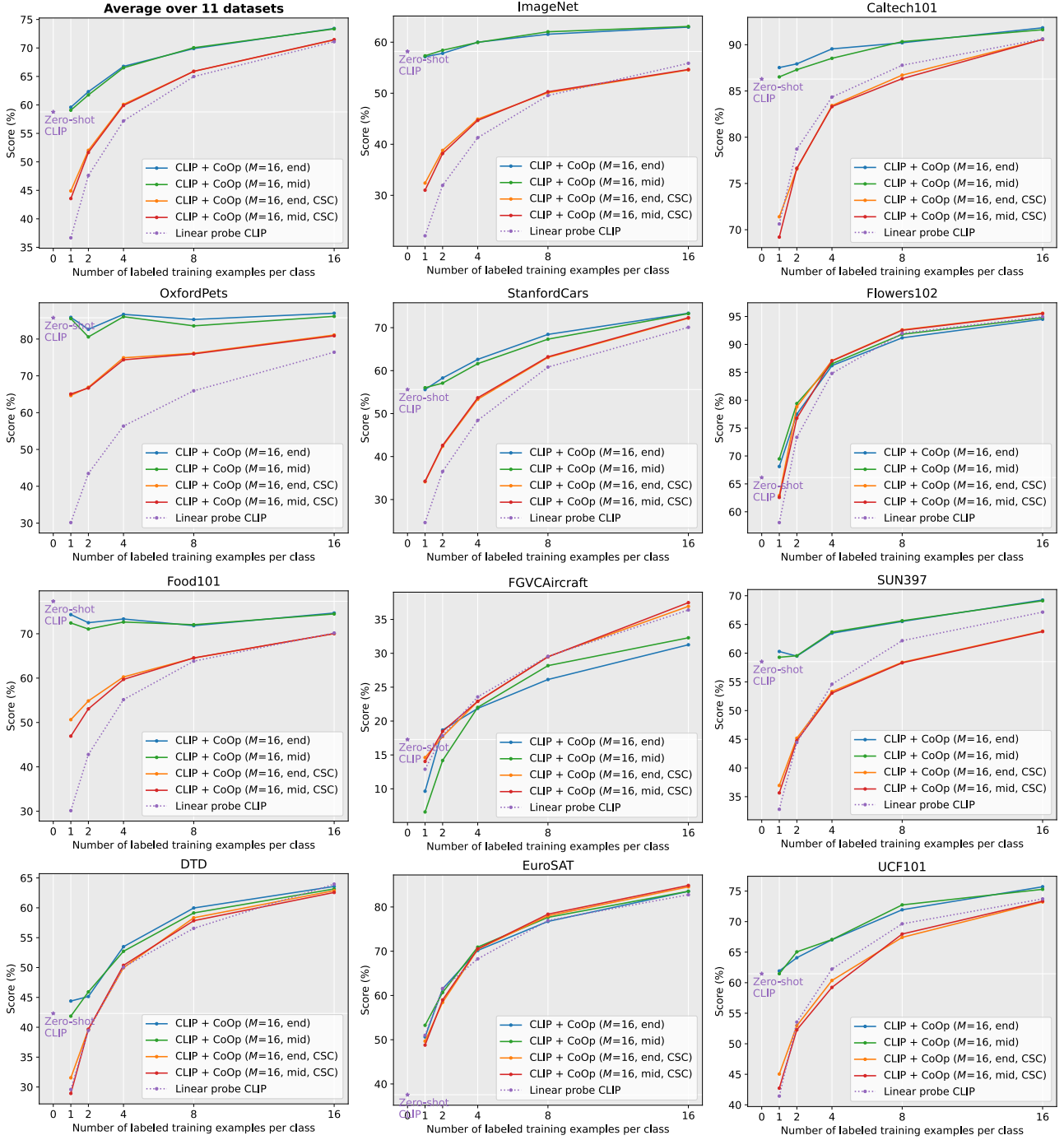
Fig. 3 Main results of few-shot learning on the 11 datasets. Overall, CoOp effectively turns CLIP into a strong few-shot learner (solid lines), achieving significant improvements over zero-shot CLIP (stars) and performing favorably against the linear probe alternative (dashed lines). M denotes the context length. "end" or "mid" means putting the class token in the end or middle. CSC means class-specific context.

We follow the few-shot evaluation protocol adopted in CLIP (Radford et al., 2021), using 1, 2, 4, 8 and 16 shots for training respectively and deploying models in the full test sets. The average results over three runs are reported for comparison.

Training Details CoOp has four versions: positioning the class token in the end or middle; unified context vs CSC. Unless otherwise stated, ResNet-50 (He et al., 2016) is used as the image encoder's backbone and the number of context tokens M is set to 16. Investigations on other design choices are discussed in

Section 4.3. All models are built on top of CLIP's open-source code.[2] CoOp's context vectors are randomly initialized by drawing from a zero-mean Gaussian distribution with standard deviation equal to 0.02. Training is done with SGD and an initial learning rate of 0.002, which is decayed by the cosine annealing rule. The maximum epoch is set to 200 for 16/8 shots, 100 for 4/2 shots, and 50 for 1 shot (except for ImageNet where the maximum epoch is fixed to 50). To mitigate explosive gradients observed in the early training iterations, we use the warmup trick by fixing the learning rate to $1e-5$, only during the first epoch.

Baseline Methods We compare CoOp with two baseline methods. The first is zero-shot CLIP, which is based on hand-crafted prompts. We follow the guideline of prompt engineering introduced by Radford et al. (2021). For generic objects and scenes, "a photo of a [CLASS]." is adopted. For fine-grained categories, task-relevant context is added like "a type of pet" for OxfordPets and "a type of food" for Food101. When it comes to specialized tasks such as recognizing textures in DTD, the prompt is customized as "[CLASS] texture." where the class names are adjectives like "bubbly" and "dotted." See Appendix A for the details. The second baseline is the linear probe model. As suggested by Radford et al. (2021) and a recent study on few-shot learning (Tian et al., 2020), training a linear classifier on top of high-quality pre-trained models' features (like CLIP) can easily achieve performance that is on a par with that of state-of-the-art few-shot learning methods, which are often much more sophisticated. We follow the same training method used by Radford et al. (2021) to train the linear probe model.

Comparison with Hand-Crafted Prompts Figure 3 summarizes the results. Our default model is CLIP+CoOp with the class token positioned in the end. The two different ways of positioning the class token achieve similar performance as their curves highly overlap. From the average performance displayed in the top-left corner, we observe that CLIP+CoOp is a strong few-shot learner, requiring only two shots on average to obtain a decent margin over zero-shot CLIP. Given 16 shots for training, the average gap brought by CoOp can be further increased to around 15%.

Figure 4 ranks the absolute improvements obtained by CoOp at 16 shots over hand-crafted prompts. Huge improvements are observed on specialized tasks namely EuroSAT and DTD where the increase in performance reaches over 45% and 20% respectively. The jumps in performance are also significant (those more than
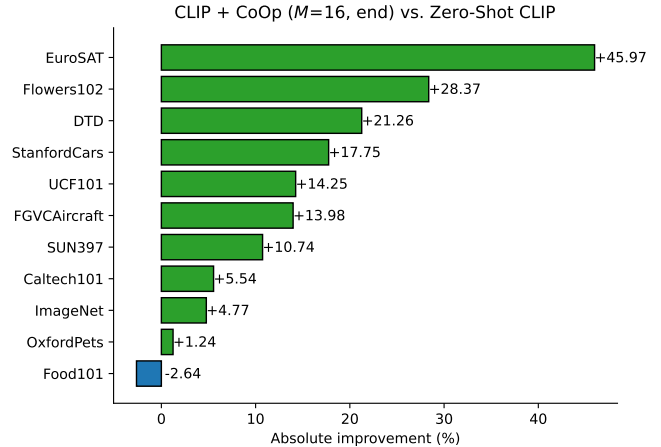
Fig. 4 Comparison with hand-crafted prompts.

10%) on most fine-grained datasets including Flowers102, StanfordCars and FGVCAircraft, as well as on scene and action recognition datasets (i.e., SUN397 & UCF101). Since ImageNet is a challenging dataset that contains 1,000 classes, the 4.77% improvement is also noteworthy. In contrast, the increases on the two fine-grained datasets, OxfordPets and Food101, are less appealing.[3] By digging into CLIP+CoOp's curves on these two datasets in Figure 3, we find there is a loss of momentum in performance improvements even with more shots used, seemingly an overfitting problem. A potential solution is to impose higher regularization like increasing the weight decay. Nonetheless, the overall results are strong enough to serve as evidence of CoOp's capability of learning task-relevant prompts in a data-efficient manner.

Comparison with Linear Probe CLIP In terms of the overall performance (Figure 3, top-left), CLIP+CoOp demonstrates clear advantages over the linear probe model. The latter requires more than 4 shots on average to match the zero-shot's performance while CoOp's average gain at 4 shots is already impressive. It is also clear that the gaps in the extreme low-data regime such as one or two shots are much larger, suggesting that CoOp is much more effective than learning a linear classifier from scratch for few-shot learning. We also observe that the linear probe model is comparable to CLIP+CoOp on the two specialized tasks (DTD & EuroSAT) as well as on a couple of fine-grained datasets (Flowers102 & FGVCAircraft)—this is not too surprising as the pre-trained CLIP space has been proved powerful, making the linear probe model a strong competitor. Nevertheless, CoOp's CSC version

Table 1 Comparison with zero-shot CLIP on robustness to distribution shift using different vision backbones. M: CoOp's context length.

	Source	Target			
Method	ImageNet	-V2	-Sketch	-A	-R
ResNet-50					
Zero-Shot CLIP	58.18	51.34	33.32	21.65	56.00
Linear Probe CLIP	55.87	45.97	19.07	12.74	34.86
CLIP + CoOp ($M{=}16$)	62.95	55.11	32.74	22.12	54.96
CLIP + CoOp ($M{=}4$)	**63.33**	**55.40**	**34.67**	**23.06**	**56.60**
ResNet-101					
Zero-Shot CLIP	61.62	54.81	38.71	28.05	64.38
Linear Probe CLIP	59.75	50.05	26.80	19.44	47.19
CLIP + CoOp ($M{=}16$)	**66.60**	**58.66**	39.08	28.89	63.00
CLIP + CoOp ($M{=}4$)	65.98	58.60	**40.40**	**29.60**	**64.98**
ViT-B/32					
Zero-Shot CLIP	62.05	54.79	40.82	29.57	**65.99**
Linear Probe CLIP	59.58	49.73	28.06	19.67	47.20
CLIP + CoOp ($M{=}16$)	**66.85**	58.08	40.44	30.62	64.45
CLIP + CoOp ($M{=}4$)	66.34	**58.24**	**41.48**	**31.34**	65.78
ViT-B/16					
Zero-Shot CLIP	66.73	60.83	46.15	47.77	73.96
Linear Probe CLIP	65.85	56.26	34.77	35.68	58.43
CLIP + CoOp ($M{=}16$)	**71.92**	64.18	46.71	48.41	74.32
CLIP + CoOp ($M{=}4$)	71.73	**64.56**	**47.89**	**49.93**	**75.14**

can beat the linear probe CLIP on the aforementioned datasets, and moreover, shows much better potential when more shots become available. We later show that CoOp obtains much stronger performance than the linear probe model in domain generalization.

Unified vs Class-Specific Context On average, using unified context leads to better performance. In terms of when to apply CSC and when not to, we have the following suggestions. For generic objects (ImageNet & Caltech101), scenes (SUN397) and actions (UCF101), using unified context is clearly better. Unified context also works better on some fine-grained datasets including OxfordPets and Food101, but on others like StanfordCars, Flowers102 and FGVCAircraft the CSC version is preferred. CSC also yields better performance on the two specialized tasks, DTD and EuroSAT, at 16 shots in particular. However, CSC mostly underperforms unified context in challenging low-data scenarios (fewer than 8 shots), which makes sense because CSC has more parameters than unified context and needs more data for training.

4.2 Domain Generalization

Since CoOp requires training on a specific data distribution, it risks learning spurious correlations that are

detrimental to generalization in unseen distributions (domains), as suggested in recent studies (Taori et al., 2020; Zhou et al., 2021). On the contrary, zero-shot CLIP is not tied to a specific data distribution and has exhibited strong robustness to distribution shifts (Radford et al., 2021). In this section, we aim to unveil how robust CoOp is to distribution shifts, in comparison to zero-shot CLIP and the linear probe model.

Datasets The source dataset is ImageNet. The target datasets are ImageNetV2 (Recht et al., 2019), ImageNet-Sketch (Wang et al., 2019), ImageNet-A (Hendrycks et al., 2021b) and ImageNet-R (Hendrycks et al., 2021a), all of which have compatible class names with ImageNet allowing seamless transfer for the prompts learned by CoOp. ImageNetV2 is a reproduced test set using different sources while following ImageNet's data collection process. ImageNet-Sketch contains sketch images belonging to the same 1,000 ImageNet classes. Both ImageNet-A and -R contain 200 classes derived from a subset of ImageNet's 1,000 classes. The former consists of real-world adversarially filtered images that cause current ImageNet classifiers to produce low results, whereas the latter features a rendition of the ImageNet classes in diverse image styles such as paintings, cartoons and sculptures.
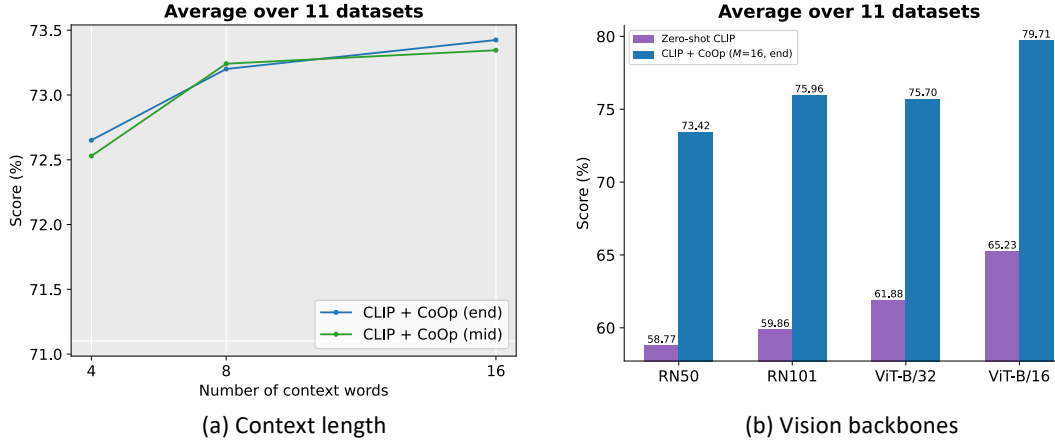
Fig. 5 Investigations on CoOp's context length and various vision backbones.

Table 2 Comparison with prompt engineering and prompt ensembling on ImageNet using different vision backbones.

Method	ResNet-50	ResNet-101	ViT-B/32	ViT-B/16
Prompt engineering	58.18	61.26	62.05	66.73
Prompt ensembling	60.41	62.54	63.71	68.74
CoOp	**62.95**	**66.60**	**66.85**	**71.92**

Table 3 Random vs manual initialization.

	Avg %
$[V]_1[V]_2[V]_3[V]_4$	72.65
"a photo of a"	72.65

Results Table 1 summarizes the results (with a variety of vision backbones). It is surprising that CoOp enhances CLIP's robustness to distribution shifts, *despite the exposure to the source dataset.* This suggests that the learned prompts are also generalizable. Moreover, it is interesting to see that using fewer context tokens leads to better robustness. In contrast, the linear probe model obtains much worse results on these target datasets, exposing its weakness in domain generalization. In Appendix B, we provide the domain generalization results on DOSCO-2k (Zhou et al., 2022b), a recently proposed benchmark focusing on contextual domain shift.

4.3 Further Analysis

Context Length How many context tokens should be used? And is it better to have more context tokens? The results in Section 4.2 suggest having a shorter context length benefits domain generalization (probably due to less overfitting as fewer parameters are learned). Here we study this hyperparameter for source datasets. Specifically, we repeat experiments on the 11 datasets by varying the context length from 4 to 8 to 16. The average results are shown in Figure 5(a), which indicate that having more context tokens leads to better performance and that positioning the class token in the middle gains more momentum with longer context length. To sum up, there is no golden rule for selecting perfect context length since one needs to balance between performance and robustness to distribution shift.

Vision Backbones Figure 5(b) summarizes the results on the 11 datasets using a variety of vision backbones covering both CNNs and ViTs. The results are expected: the more advanced the backbone, the better the performance. The gap between CoOp and hand-crafted prompts is significant across all architectures.

Comparison with Prompt Ensembling The authors of CLIP (Radford et al., 2021) have suggested that additional improvements can be obtained by ensembling over multiple zero-shot classifiers generated using different hand-crafted prompts, such as "a photo of the large [CLASS].", "a bad photo of the [CLASS]." and "a origami [CLASS].", which reflect a different scale, view and abstraction respectively for an image. We are interested to know whether the prompts learned by CoOp can still maintain advantages when compared with prompt ensembling. For fair comparison, we use the select prompts from Radford et al. (2021), which have been extensively tuned on ImageNet, to construct the ensemble classifier. Table 2 shows the comparison and justifies the superiority of

CoOp. Given the potential of prompt ensembling, future work could investigate how to improve CoOp from the ensembling perspective.

Comparison with Other Fine-tuning Methods
We further compare CoOp with other fine-tuning methods: i) fine-tuning CLIP's image encoder; ii) optimizing a transformation layer added to the text encoder's output; iii) optimizing a bias term added to the text encoder's output. The results are shown in Table 5. Obviously, fine-tuning the image encoder does not work well. Adding a transformation layer slightly improves upon the zero-shot model. Adding a bias term shows promising results, but still largely underperforms CoOp, which suggests that the gradients that went through the text encoder provide more useful information.

Initialization We compare random initialization with manual initialization. The latter uses the embeddings of "a photo of a" to initialize the context vectors for the 11 datasets. For fair comparison, we also set the context length to 4 when using random initialization. Table 3 suggests a "good" initialization does not make much difference. Though further tuning of the initialization words might help, in practice we suggest using the simple random initialization method.

Interpreting the Learned Prompts is difficult because the context vectors are optimized in a continuous space. We resort to an indirect way by searching within the vocabulary for words that are closest to the learned vectors based on the Euclidean distance. Note that CLIP (Radford et al., 2021) uses the BPE representation (Sennrich et al., 2016) for tokenization, so the vocabulary includes subwords that frequently appear in text, such as "hu" (subsumed by many words like "hug" and "human"). Table 4 shows the searched results on some datasets. We observe that a few words are somewhat relevant to the tasks, such as "enjoyed" for Food101, "fluffy" and "paw" for OxfordPets, and "pretty" for DTD. But when connecting all the nearest words together, the prompts do not make much sense. We also observe that when using manual initialization (like "a photo of a"), the nearest words for the converged vectors are mostly the ones used for initialization. We conjecture that the learned vectors might encode meanings that are beyond the existing vocabulary. Overall, we are unable to draw any firm conclusion based on the observations because using nearest words to interpret the learned prompts could be inaccurate—the semantics of the vectors is not necessarily correlated with the nearest words.

5 Conclusion, Limitations and Future Work

Large pre-trained vision-language models have shown surprisingly powerful capabilities in diverse downstream applications. However, these models, also called vision foundation models given their "critically central yet incomplete" nature (Bommasani et al., 2021), need to be adapted using automated techniques for better downstream performance and efficiency.

Our research provides timely insights on how CLIP-like models can be turned into a data-efficient learner by using prompt learning, and reveals that despite being a learning-based approach, CoOp performs much better in domain generalization than manual prompts. The results serve as strong evidence that prompt learning has potential for large vision models. It is worth noting that our paper presents the first comprehensive study about adapting large vision models with prompt learning.

Though the performance is excellent, the results of CoOp are relatively difficult to interpret, like other continuous prompt learning methods in NLP. The experiments also reveal that CoOp is sensitive to noisy labels given the weak performance on Food101.

Nevertheless, the simplicity of CoOp allows easy extension for future work and there remain many interesting questions to explore, such as cross-dataset transfer (Zhou et al., 2022a) and test-time adaptation (Wang et al., 2020). It would also be interesting to investigate more generic adaptation methods for mega-size vision models (Jia et al., 2022; Bahng et al., 2022; Gao et al., 2021). In summary, we hope the empirical findings and insights presented in this work could pave the way for future research on efficient adaptation methods for emerging foundation models, which is still a nascent research topic.

Appendix

A Datasets Details

The detailed statistics of the 11 datasets, as well as the four variants of ImageNet, are shown in Table 6. The hand-crafted prompts used for zero-shot CLIP are also detailed in the table. For Caltech101, the "BACKGROUND_Google" and "Faces_easy" classes are discarded. For the video dataset, UCF101, the middle frame of each video is used as input to the image encoder.

Table 4 The nearest words for each of the 16 context vectors learned by CoOp, with their distances shown in parentheses. N/A means non-Latin characters.

#	ImageNet	Food101	OxfordPets	DTD	UCF101
1	potd (1.7136)	lc (0.6752)	tosc (2.5952)	boxed (0.9433)	meteorologist (1.5377)
2	that (1.4015)	enjoyed (0.5305)	judge (1.2635)	seed (1.0498)	exe (0.9807)
3	filmed (1.2275)	beh (0.5390)	fluffy (1.6099)	anna (0.8127)	parents (1.0654)
4	fruit (1.4864)	matches (0.5646)	cart (1.3958)	mountain (0.9509)	masterful (0.9528)
5	,... (1.5863)	nytimes (0.6993)	harlan (2.2948)	eldest (0.7111)	fe (1.3574)
6	° (1.7502)	prou (0.5905)	paw (1.3055)	pretty (0.8762)	thof (1.2841)
7	excluded (1.2355)	lower (0.5390)	incase (1.2215)	faces (0.7872)	where (0.9705)
8	cold (1.4654)	N/A	bie (1.5454)	honey (1.8414)	kristen (1.1921)
9	stery (1.6085)	minute (0.5672)	snuggle (1.1578)	series (1.6680)	imam (1.1297)
10	warri (1.3055)	∼ (0.5529)	along (1.8298)	coca (1.5571)	near (0.8942)
11	marvelcomics (1.5638)	well (0.5659)	enjoyment (2.3495)	moon (1.2775)	tummy (1.4303)
12	.: (1.7387)	ends (0.6113)	jt (1.3726)	lh (1.0382)	hel (0.7644)
13	N/A	mis (0.5826)	improving (1.3198)	won (0.9314)	boop (1.0491)
14	lation (1.5015)	somethin (0.6041)	srsly (1.6759)	replied (1.1429)	N/A
15	muh (1.4985)	seminar (0.5274)	asteroid (1.3395)	sent (1.3173)	facial (1.4452)
16	.# (1.9340)	N/A	N/A	piedmont (1.5198)	during (1.1755)

Table 5 CoOp vs other fine-tuning methods on ImageNet (w/ 16 shots). Δ: difference with the zero-shot model.

	ImageNet	Δ
Zero-shot CLIP	58.18	-
Linear probe	55.87	**-2.31**
Fine-tuning CLIP's image encoder	18.28	**-39.90**
Optimizing transformation layer (text)	58.86	**0.68**
Optimizing bias (text)	60.93	**+2.75**
CoOp	**62.95**	**+4.77**

B Results on DOSCO-2k

DOSCO-2k The DOSCO (DOmain Shift in COntext) benchmark (Zhou et al., 2022b) contains 7 image recognition datasets, which cover a wide range of classification problems, such as generic object recognition, fine-grained recognition on aircraft models, and action recognition. Unlike existing domain generalization datasets where the domain labels are manually defined and often limited to image style variations, DOSCO-2k focuses on broader contextual domain shift, which is automatically detected by a neural network pre-trained on the Places dataset (Zhou et al., 2017). Following Zhou et al. (2022b), we use the 2k version where the training and validation splits in each dataset have 2,000 images in total (1,600 for training and 400 for validation).

Results We study three methods' domain generalization performance on DOSCO-2k: CLIP, CoOp and CoCoOp (Zhou et al., 2022a). All models are trained on the training set and the checkpoints with the best validation performance are used for final test in unseen domains. Table 7 shows the results of four different architectures. It is clear that the two learnable methods outperform the zero-shot method with a large margin, despite having only a small number of parameters to tune. CoCoOp beats CoOp on 4 out of 7 datasets but CoOp's average performance is higher. In summary, the results suggest that efficient adaptation methods like CoOp and CoCoOp have great potential in tackling transfer learning problems.

References

Bahng H, Jahanian A, Sankaranarayanan S, Isola P (2022) Visual prompting: Modifying pixel space to adapt pre-trained models. arXiv preprint arXiv:220317274

Bommasani R, Hudson DA, Adeli E, Altman R, Arora S, von Arx S, Bernstein MS, Bohg J, Bosselut A, Brunskill E, et al. (2021) On the opportunities and risks of foundation models. arXiv preprint arXiv:210807258

Bossard L, Guillaumin M, Van Gool L (2014) Food-101–mining discriminative components with random forests. In: ECCV

Brown TB, Mann B, Ryder N, Subbiah M, Kaplan J, Dhariwal P, Neelakantan A, Shyam P, Sastry G, Askell A, et al. (2020) Language models are few-shot learners. arXiv preprint arXiv:200514165

Chen T, Kornblith S, Norouzi M, Hinton G (2020) A simple framework for contrastive learning of visual representations. In: ICML

Cimpoi M, Maji S, Kokkinos I, Mohamed S, Vedaldi A (2014) Describing textures in the wild. In: CVPR

Deng J, Dong W, Socher R, Li LJ, Li K, Fei-Fei L (2009) Imagenet: A large-scale hierarchical image database. In: CVPR

Desai K, Johnson J (2021) Virtex: Learning visual representations from textual annotations. In: CVPR

Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S, et al. (2021) An image is worth 16x16 words: Transformers for image recognition at scale. In: ICLR

Elhoseiny M, Saleh B, Elgammal A (2013) Write a classifier: Zero-shot learning using purely textual descriptions. In: ICCV

Fei-Fei L, Fergus R, Perona P (2004) Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In: CVPR-W

Table 6 Datasets statistics.

Dataset	Classes	Train	Val	Test	Hand-crafted prompt
ImageNet	1,000	1.28M	N/A	50,000	"a photo of a [CLASS]."
Caltech101	100	4,128	1,649	2,465	"a photo of a [CLASS]."
OxfordPets	37	2,944	736	3,669	"a photo of a [CLASS], a type of pet."
StanfordCars	196	6,509	1,635	8,041	"a photo of a [CLASS]."
Flowers102	102	4,093	1,633	2,463	"a photo of a [CLASS], a type of flower."
Food101	101	50,500	20,200	30,300	"a photo of [CLASS], a type of food."
FGVCAircraft	100	3,334	3,333	3,333	"a photo of a [CLASS], a type of aircraft."
SUN397	397	15,880	3,970	19,850	"a photo of a [CLASS]."
DTD	47	2,820	1,128	1,692	"[CLASS] texture."
EuroSAT	10	13,500	5,400	8,100	"a centered satellite photo of [CLASS]."
UCF101	101	7,639	1,898	3,783	"a photo of a person doing [CLASS]."
ImageNetV2	1,000	N/A	N/A	10,000	"a photo of a [CLASS]."
ImageNet-Sketch	1,000	N/A	N/A	50,889	"a photo of a [CLASS]."
ImageNet-A	200	N/A	N/A	7,500	"a photo of a [CLASS]."
ImageNet-R	200	N/A	N/A	30,000	"a photo of a [CLASS]."

Table 7 Domain generalization results on DOSCO-2k, a recently proposed benchmark focusing on broader contextual domain shift. Among the three approaches, CoOp and its follow-up, CoCoOp, contain learnable components while CLIP here denotes the zero-shot model. Both CoOp and CoCoOp use four learnable context tokens initialized with the word embeddings of "a photo of a". Bold denotes the best performance on each dataset for a specific architecture.

	P-Air	P-Cars	P-Ctech	P-Ins	P-Mam	P-Pets	P-UCF	*Avg*
ResNet-50								
CLIP	16.1	56.1	86.7	62.7	59.7	84.0	60.6	60.9
CoOp	**22.1**	**60.7**	89.4	66.3	61.6	83.8	**69.2**	64.7
CoCoOp	20.1	59.8	**90.4**	**67.9**	**63.8**	**87.6**	69.1	**65.5**
ResNet-101								
CLIP	17.5	63.2	89.5	62.4	62.2	84.2	61.3	62.9
CoOp	**24.6**	**68.2**	92.0	68.3	65.4	88.2	**72.7**	**68.5**
CoCoOp	22.5	65.2	**93.3**	**69.9**	**67.5**	**88.6**	71.5	68.4
ViT-B/32								
CLIP	18.2	60.1	91.6	61.3	61.8	85.5	61.3	62.8
CoOp	**24.0**	**63.0**	93.6	67.3	65.7	**88.5**	**74.5**	**68.1**
CoCoOp	19.5	60.4	**93.8**	**69.8**	**67.3**	**88.5**	72.7	67.4
ViT-B/16								
CLIP	24.4	64.9	92.6	67.5	67.9	87.4	66.1	67.2
CoOp	**32.4**	**72.4**	94.7	73.2	72.1	90.1	**78.2**	**73.3**
CoCoOp	30.4	68.7	**94.8**	**73.5**	**73.6**	**91.6**	76.3	72.7

Frome A, Corrado G, Shlens J, Bengio S, Dean J, Ranzato M, Mikolov T (2013) Devise: A deep visual-semantic embedding model. In: NeurIPS

Fürst A, Rumetshofer E, Tran V, Ramsauer H, Tang F, Lehner J, Kreil D, Kopp M, Klambauer G, Bitto-Nemling A, et al. (2021) Cloob: Modern hopfield networks with infoloob outperform clip. arXiv preprint arXiv:211011316

Gao P, Geng S, Zhang R, Ma T, Fang R, Zhang Y, Li H, Qiao Y (2021) Clip-adapter: Better vision-language models with feature adapters. arXiv preprint arXiv:211004544

Gao T, Fisch A, Chen D (2020) Making pre-trained language models better few-shot learners. arXiv preprint arXiv:201215723

Gomez L, Patel Y, Rusiñol M, Karatzas D, Jawahar C (2017) Self-supervised learning of visual features through embedding images into text topic spaces. In: CVPR

He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: CVPR

He K, Fan H, Wu Y, Xie S, Girshick R (2020) Momentum contrast for unsupervised visual representation learning. In: CVPR

Helber P, Bischke B, Dengel A, Borth D (2019) 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

Hénaff OJ, Srinivas A, Fauw JD, Razavi A, Doersch C, Eslami SMA, van den Oord A (2020) Data-efficient image recognition with contrastive predictive coding. In: ICML

Hendrycks D, Basart S, Mu N, Kadavath S, Wang F, Dorundo E, Desai R, Zhu T, Parajuli S, Guo M, Song D, Steinhardt J, Gilmer J (2021a) The many faces of robustness: A critical analysis of out-of-distribution generalization. ICCV

Hendrycks D, Zhao K, Basart S, Steinhardt J, Song D (2021b) Natural adversarial examples. In: CVPR

Jia C, Yang Y, Xia Y, Chen YT, Parekh Z, Pham H, Le QV, Sung Y, Li Z, Duerig T (2021) Scaling up visual and vision-language representation learning with noisy text supervision. In: ICML

Jia M, Tang L, Chen BC, Cardie C, Belongie S, Hariharan B, Lim SN (2022) Visual prompt tuning. arXiv preprint arXiv:220312119

Jiang Z, Xu FF, Araki J, Neubig G (2020) How can we know what language models know? ACL

Joulin A, Van Der Maaten L, Jabri A, Vasilache N (2016) Learning visual features from large weakly supervised data. In: ECCV

Krause J, Stark M, Deng J, Fei-Fei L (2013) 3d object representations for fine-grained categorization. In: ICCV-W

Lei Ba J, Swersky K, Fidler S, et al. (2015) Predicting deep zero-shot convolutional neural networks using textual descriptions. In: ICCV

Lester B, Al-Rfou R, Constant N (2021) The power of scale for parameter-efficient prompt tuning. arXiv preprint arXiv:210408691

Li A, Jabri A, Joulin A, van der Maaten L (2017) Learning visual n-grams from web data. In: ICCV

Li XL, Liang P (2021) Prefix-tuning: Optimizing continuous prompts for generation. arXiv preprint arXiv:210100190

Li Y, Liang F, Zhao L, Cui Y, Ouyang W, Shao J, Yu F, Yan J (2021) Supervision exists everywhere: A data efficient contrastive language-image pre-training paradigm. arXiv preprint arXiv:211005208

Liu P, Yuan W, Fu J, Jiang Z, Hayashi H, Neubig G (2021a) Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. arXiv preprint arXiv:210713586

Liu X, Zheng Y, Du Z, Ding M, Qian Y, Yang Z, Tang J (2021b) Gpt understands, too. arXiv preprint arXiv:210310385

Maji S, Rahtu E, Kannala J, Blaschko M, Vedaldi A (2013) Fine-grained visual classification of aircraft. arXiv preprint arXiv:13065151

Nilsback ME, Zisserman A (2008) Automated flower classification over a large number of classes. In: ICVGIP

Parkhi OM, Vedaldi A, Zisserman A, Jawahar C (2012) Cats and dogs. In: CVPR

Petroni F, Rocktäschel T, Lewis P, Bakhtin A, Wu Y, Miller AH, Riedel S (2019) Language models as knowledge bases? In: EMNLP

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

Recht B, Roelofs R, Schmidt L, Shankar V (2019) Do imagenet classifiers generalize to imagenet? In: ICML

Sennrich R, Haddow B, Birch A (2016) Neural machine translation of rare words with subword units. In: ACL

Shin T, Razeghi Y, Logan IV RL, Wallace E, Singh S (2020) Autoprompt: Eliciting knowledge from language models with automatically generated prompts. In: EMNLP

Singh A, Hu R, Goswami V, Couairon G, Galuba W, Rohrbach M, Kiela D (2021) Flava: A foundational language and vision alignment model. arXiv preprint arXiv:211204482

Socher R, Ganjoo M, Sridhar H, Bastani O, Manning CD, Ng AY (2013) Zero-shot learning through cross-modal transfer. In: NeurIPS

Soomro K, Zamir AR, Shah M (2012) Ucf101: A dataset of 101 human actions classes from videos in the wild. arXiv preprint arXiv:12120402

Taori R, Dave A, Shankar V, Carlini N, Recht B, Schmidt L (2020) Measuring robustness to natural distribution shifts in image classification. In: NeurIPS

Tian Y, Wang Y, Krishnan D, Tenenbaum JB, Isola P (2020) Rethinking few-shot image classification: a good embedding is all you need? In: ECCV

Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I (2017) Attention is all you need. In: NeurIPS

Wang D, Shelhamer E, Liu S, Olshausen B, Darrell T (2020) Tent: Fully test-time adaptation by entropy minimization. arXiv preprint arXiv:200610726

Wang H, Ge S, Lipton Z, Xing EP (2019) Learning robust global representations by penalizing local predictive power. In: NeurIPS

Xiao J, Hays J, Ehinger KA, Oliva A, Torralba A (2010) Sun database: Large-scale scene recognition from abbey to zoo. In: CVPR

Yuan L, Chen D, Chen YL, Codella N, Dai X, Gao J, Hu H, Huang X, Li B, Li C, et al. (2021) Florence: A new foundation model for computer vision. arXiv preprint arXiv:211111432

Zhang Y, Jiang H, Miura Y, Manning CD, Langlotz CP (2020) Contrastive learning of medical visual representations from paired images and text. arXiv preprint arXiv:201000747

Zhong Z, Friedman D, Chen D (2021) Factual probing is [mask]: Learning vs. learning to recall. In: NAACL

Zhou B, Lapedriza A, Khosla A, Oliva A, Torralba A (2017) Places: A 10 million image database for scene recognition. IEEE transactions on pattern analysis and machine intelligence 40(6):1452–1464

Zhou K, Liu Z, Qiao Y, Xiang T, Loy CC (2021) Domain generalization: A survey. arXiv preprint arXiv:210302503

Zhou K, Yang J, Loy CC, Liu Z (2022a) Conditional prompt learning for vision-language models. arXiv preprint arXiv:220305557

Zhou K, Zhang Y, Zang Y, Yang J, Loy CC, Liu Z (2022b) On-device domain generalization. arXiv preprint arXiv:220907521