Conditional Prompt Learning for Vision-Language Models

Anonymous CVPR submission

Paper ID 51

Abstract

With the rise of powerful pre-trained vision-language models like CLIP, it becomes essential to investigate ways to adapt these models to downstream datasets. A recently proposed method named Context Optimization (CoOp) introduces the concept of prompt learning—a recent trend in NLP—to the vision domain for adapting pre-trained vision-language models. Specifically, CoOp turns context words in a prompt into a set of learnable vectors and, with only a few labeled images for learning, can achieve huge improvements over intensively-tuned manual prompts. In our study we identify a critical problem of CoOp: the learned context is not generalizable to wider unseen classes within the same dataset, suggesting that CoOp overfits base classes observed during training. To address the problem, we propose Conditional Context Optimization (CoCoOp), which extends CoOp by further learning a lightweight neural network to generate for each image an input-conditional token (vector). Compared to CoOp's static prompts, our dynamic prompts adapt to each instance and are thus less sensitive to class shift. Extensive experiments show that CoCoOp generalizes much better than CoOp to unseen classes, even showing promising transferability beyond a single dataset; and yields stronger domain generalization performance as well.

1. Introduction

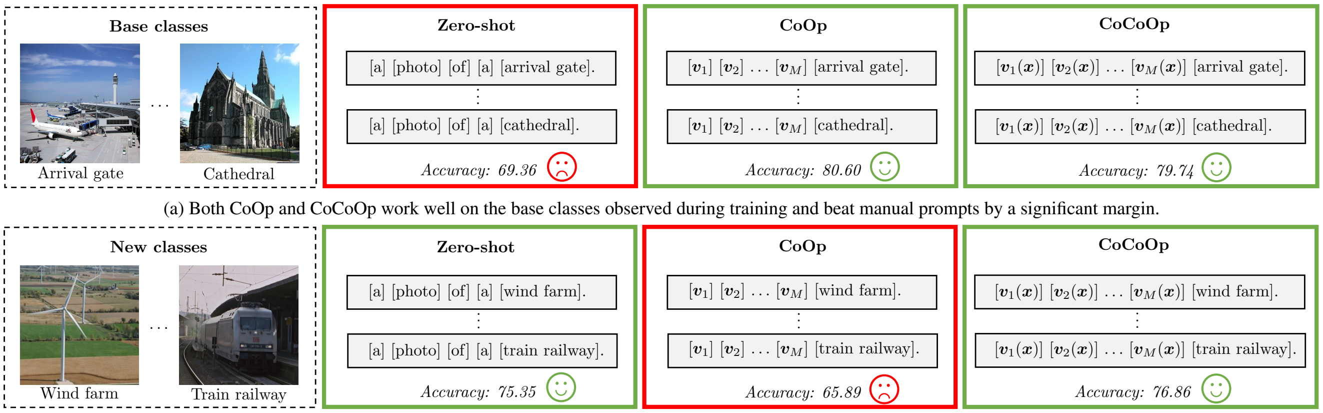
Recent research in large-scale vision-language pre-training has achieved striking performance in zero-shot image recognition [13, 24, 33, 40], demonstrating a potential in learning open-world visual concepts for such a paradigm. The key design lies in how visual concepts are modeled. In traditional supervised learning where labels are discretized, each category is associated with a randomly initialized weight vector that is learned to minimize the distance with images containing the same category. Such a learning method focuses on closed-set visual concepts, limiting the model to a pre-defined list of categories, and is unscalable when it comes to new categories unseen during training.

In contrast, for vision-language models[1] like CLIP [40] and ALIGN [24], the classification weights are diametrically generated by a parameterized text encoder (e.g., a Transformer [48]) through prompting [34]. For instance, to differentiate pet images containing different breeds of dogs and cats, one can adopt a prompt template like "a photo of a {class}, a type of pet" as input to the text encoder, and as a result, class-specific weights for classification can be synthesized by filling in the "{class}" token with real class names. Compared to discrete labels, vision-language models' source of supervision comes from natural language, which allows open-set visual concepts to be broadly explored and has been proven effective in learning transferable representations [24, 40].

With the rise of such powerful vision-language models, the community has recently started to investigate potential solutions to efficiently adapt these models to downstream datasets [14, 53, 56, 61]. To fit web-scale data, such as the 400 million pairs of images and texts used by CLIP, vision-language models are purposefully designed to have high capacity, entailing that the model size would be enormous, typically with hundreds of millions of parameters or even billions. Therefore, fine-tuning the entire model, as often adopted in deep learning research [18], is impractical and might even damage the well-learned representation space.

A safer approach is to tune a prompt by adding some context that is meaningful to a task, like "a type of pet" for the pet dataset mentioned above, which has been found effective in improving performance [40]. However, prompt engineering is extremely time-consuming and inefficient as it has to be based on trial and error, and does not guarantee an optimal prompt either. To automate prompt engineering, Zhou et al. [61] have recently explored the concept of prompt learning—a recent trend in NLP [15, 25, 30, 32, 44, 59]—for adapting pre-trained vision-language models. Their approach, Context Optimization (CoOp), turns context words in a prompt into a set of learnable vectors, taking advantage of the differentiable nature of neural networks. With only a few labeled images for learning, CoOp achieves

[1] We follow existing studies [13, 24, 33, 40] to refer to CLIP-like models as *vision-language models*.


Base classes
Arrival gate
Cathedral
Zero-shot
[a] [photo] [of] [a] [arrival gate].
[a] [photo] [of] [a] [cathedral].
Accuracy: 69.36 ☹
CoOp
$[\boldsymbol{v}_1]$ $[\boldsymbol{v}_2]$ … $[\boldsymbol{v}_M]$ [arrival gate].
$[\boldsymbol{v}_1]$ $[\boldsymbol{v}_2]$ … $[\boldsymbol{v}_M]$ [cathedral].
Accuracy: 80.60 ☺
CoCoOp
$[\boldsymbol{v}_1(\boldsymbol{x})]$ $[\boldsymbol{v}_2(\boldsymbol{x})]$ … $[\boldsymbol{v}_M(\boldsymbol{x})]$ [arrival gate].
$[\boldsymbol{v}_1(\boldsymbol{x})]$ $[\boldsymbol{v}_2(\boldsymbol{x})]$ … $[\boldsymbol{v}_M(\boldsymbol{x})]$ [cathedral].
Accuracy: 79.74 ☺
New classes
Wind farm
Train railway
Zero-shot
[a] [photo] [of] [a] [wind farm].
[a] [photo] [of] [a] [train railway].
Accuracy: 75.35 ☺
CoOp
$[\boldsymbol{v}_1]$ $[\boldsymbol{v}_2]$ … $[\boldsymbol{v}_M]$ [wind farm].
$[\boldsymbol{v}_1]$ $[\boldsymbol{v}_2]$ … $[\boldsymbol{v}_M]$ [train railway].
Accuracy: 65.89 ☹
CoCoOp
$[\boldsymbol{v}_1(\boldsymbol{x})]$ $[\boldsymbol{v}_2(\boldsymbol{x})]$ … $[\boldsymbol{v}_M(\boldsymbol{x})]$ [wind farm].
$[\boldsymbol{v}_1(\boldsymbol{x})]$ $[\boldsymbol{v}_2(\boldsymbol{x})]$ … $[\boldsymbol{v}_M(\boldsymbol{x})]$ [train railway].
Accuracy: 76.86 ☺

(a) Both CoOp and CoCoOp work well on the base classes observed during training and beat manual prompts by a significant margin.

(b) The instance-conditional prompts learned by CoCoOp are much more generalizable than CoOp to the unseen classes.

Figure 1. **Motivation of our research: to learn generalizable prompts**. The images are randomly selected from SUN397 [55], which is a widely-used scene recognition dataset.

huge improvements over intensively-tuned manual prompts across a wide range of image recognition datasets.

In our study, we identify a critical problem of CoOp: the learned context is not generalizable to wider unseen classes within the same task. Figure 1 illustrates the problem: the context learned by CoOp works well in distinguishing the base classes like “arrival gate” and “cathedral” but suffers a significant drop in accuracy when it is transferred to the new (unseen) classes, such as “wind farm” and “train railway”—even though the task’s nature remains the same, i.e., recognizing scenes. The results suggest that the learned context overfits the base classes, thus failing to capture more generalizable elements that are vital for broader scene recognition. We argue that such a problem is caused by CoOp’s static design: the context, which is fixed once learned, is optimized only for a specific set of (training) classes. On the contrary, the manually-designed prompts adopted by the zero-shot method are relatively generalizable.

To address the weak generalizability problem, we introduce a novel concept: *conditional prompt learning*. The key idea is to make a prompt conditioned on each input instance (image) rather than fixed once learned. To make the model parameter-efficient, we introduce a simple yet effective implementation of conditional prompt learning. Specifically, we extend CoOp by further learning a lightweight neural network to generate for each image an input-conditional token (vector), which is combined with the learnable context vectors. We call our approach Conditional Context Optimization (CoCoOp).[2] An overview is shown in Figure 2. Interestingly, the paradigm of CoCoOp is analogous to image captioning [49], which explains why instance-conditional prompts are more generalizable: *they are optimized to characterize each instance (more robust to class shift) rather than to serve only for some specific classes.*

We present comprehensive experiments on 11 datasets covering a diverse set of visual recognition tasks. Specifically, we design a base-to-new generalization setting where a model is first learned using base classes and then tested on completely new classes. Compared with the zero-shot method [40] and CoOp [61], our approach achieves the best overall performance (Table 1). Importantly, CoCoOp gains significant improvements over CoOp in unseen classes (Figure 3(a)), allowing the gap between manual and learning-based prompts to be substantially reduced.

In a more challenging scenario where the context learned for one task is directly transferred to other tasks with drastically different classes, CoCoOp still beats CoOp with a clear margin (Table 2), suggesting that instance-conditional prompts are more transferable and have the potential to succeed at larger scale. CoCoOp also obtains stronger domain generalization performance than CoOp (Table 3), further justifying the strengths of dynamic prompts.

In summary, our research provides timely insights into the generalizability problem in prompt learning, and crucially, demonstrates the effectiveness of a simple idea in various problem scenarios. We hope our approach and the findings presented in this work can pave the way for future research in generalizable—and transferable—prompt learning.

2. Related Work

Vision-Language Models We mainly review studies focused on aligning images and texts to learn a joint embedding space [24, 40, 58]. The idea of cross-modality alignment is certainly not new and has been investigated since nearly a decade ago—though with dramatically different technologies than today.

A typical vision-language model consists of three key elements: two for image and text encoding while the third is

[2] Pronounced as /kəʊˌku:p/.

related to the design of loss functions. In early days, models for processing images and texts are often designed and also learned independently, with their outputs connected by extra modules (losses) for alignment. Images are often encoded using hand-crafted descriptors [10, 45] or neural networks [12, 29], while texts are encoded using, for instance, pre-trained word vectors [12, 45] or the frequency-based TF-IDF features [10, 29]. In terms of cross-modality alignment, common approaches include metric learning [12], multi-label classification [16, 26], and n-gram language learning [31]. Recently, a study suggests that training the vision part with an image captioning loss can make the visual representation more transferable [7].

Recent vision-language models [13, 24, 33, 40] bridge the two modalities by learning two encoders jointly. Also, the models are now built with much larger neural networks. As discussed in Zhou et al. [61], recent successes in vision-language models are mainly attributed to the developments in i) Transformers [48], ii) contrastive representation learning [4, 17, 20], and iii) web-scale training datasets [24, 40]. A representative approach is CLIP [40], which trains two neural network-based encoders using a contrastive loss to match pairs of images and texts. After consuming 400 million data pairs, the CLIP model demonstrates a remarkable zero-shot image recognition capability. Similar to CoOp [61], our approach is orthogonal to the research of CLIP-like models [13, 24, 33, 40], aiming to offer an efficient solution for adapting pre-trained vision-language models to downstream applications.

Prompt Learning This topic originates from the NLP domain. The motivation was to view pre-trained language models, such as BERT [8] or GPT [41], as knowledge bases from which information useful to downstream tasks is elicited [39]. Concretely, given a pre-trained language model, the task is often formulated as a "fill-in-the-blank" cloze test, such as asking the model to predict the masked token in "No reason to watch. <u>It was</u> [MASK]" as either "positive" or "negative" for sentiment classification. The key lies in how to design the underlined part, known as prompt (template), in such a format familiar to the model.

Instead of manually designing a prompt, research in prompt learning aims to automate the process with the help of affordable-sized labeled data. Jiang et al. [25] use text mining and paraphrasing to generate a group of candidate prompts, within which the optimal ones are chosen to have the highest training accuracy. Shin et al. [44] propose AutoPrompt, a gradient-based approach that selects from a vocabulary the best tokens that cause the greatest changes in gradients based on the label likelihood. Our research is most related to continuous prompt learning methods [30, 32, 59], where the main idea is to turn a prompt into a set of continuous vectors that can be end-to-end optimized with respect to an objective function. See Liu et al. [34] for a more comprehensive survey.

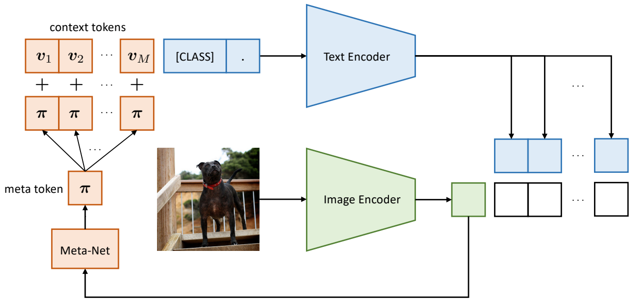


Figure 2. Our approach, Conditional Context Optimization (CoCoOp), consists of two learnable components: a set of context vectors and a lightweight neural network (Meta-Net) that generates for each image an input-conditional token.

In computer vision, prompt learning is a nascent research direction that has only been explored very recently [27, 42, 56, 61]. Our research is built on top of CoOp [61], which is the earliest work to bring continuous prompt learning to the vision domain for adaptation of pre-trained vision-language models. Crucially, our approach solves the weak generalizability problem of CoOp [61], based on a simple idea of conditional prompt learning—*which to our knowledge is also novel in the context of NLP and thus could be of interest to the NLP community as well.*

Zero-Shot Learning (ZSL) is another relevant research area where the goal is similar to ours, i.e., to recognize novel classes by training only on base classes [3, 51, 54, 57]. Moreover, the generalization problem where a model trained on base classes often fails on novel classes is also linked to the "seen-class bias" issue raised in the ZSL literature [54]. The most common approach to ZSL is to learn a semantic space based on auxiliary information such as attributes [23] or word embeddings [12, 52]. Different from existing ZSL methods, our work addresses the emerging problem of adapting large vision-language models and uses drastically different techniques based on prompting.

3. Methodology

An overview of our approach is shown in Figure 2. Below we first provide brief reviews on CLIP [40], which is the base model used in this paper, and CoOp [61]. Then, we present the technical details of our approach as well as the rationale behind the design. Same as CoOp, our approach is applicable to broader CLIP-like vision-language models.

3.1. Reviews of CLIP and CoOp

Contrastive Language-Image Pre-training known as CLIP [41], has well demonstrated the potential of learning open-set visual concepts. CLIP is built using two encoders, one for image and the other for text, as shown in Figure 2.

The image encoder can be either a ResNet [18] or a ViT [9], which is used to transform an image into a feature vector. The text encoder is a Transformer [48], which takes as input a sequence of word tokens and again produces a vectorized representation.

During training, CLIP adopts a contrastive loss to learn a joint embedding space for the two modalities. Specifically, for a mini-batch of image-text pairs, CLIP maximizes for each image the cosine similarity with the matched text while minimizes the cosine similarities with all other unmatched texts, and the loss is computed in a similar fashion for each text too. After training, CLIP can be used for zero-shot image recognition. Let $\boldsymbol{x}$ be image features generated by the image encoder and $\{\boldsymbol{w}_i\}_{i=1}^{K}$ a set of weight vectors produced by the text encoder, each representing a category (suppose there are K categories in total). In particular, each $\boldsymbol{w}_i$ is derived from a prompt, such as "a photo of a {class}" where the "{class}" token is filled with the i-th class name. The prediction probability is then

$$p(y|\boldsymbol{x}) = \frac{\exp(\text{sim}(\boldsymbol{x}, \boldsymbol{w}_y)/\tau)}{\sum_{i=1}^{K}\exp(\text{sim}(\boldsymbol{x}, \boldsymbol{w}_i)/\tau)}, \tag{1}$$

where $\text{sim}(\cdot,\cdot)$ denotes cosine similarity and τ is a learned temperature parameter.

Context Optimization (CoOp) aims to overcome the inefficiency problem in prompt engineering for better adapting pre-trained vision-language models to downstream applications [61]. The key idea in CoOp is to model each context token using a continuous vector that can be end-to-end learned from data. Concretely, instead of using "a photo of a" as the context, CoOp introduces M learnable context vectors, $\{\boldsymbol{v}_1, \boldsymbol{v}_2, \dots, \boldsymbol{v}_M\}$, each having the same dimension with the word embeddings. The prompt for the i-th class, denoted by $\boldsymbol{t}_i$, now becomes $\boldsymbol{t}_i = \{\boldsymbol{v}_1, \boldsymbol{v}_2, \dots, \boldsymbol{v}_M, \boldsymbol{c}_i\}$ where $\boldsymbol{c}_i$ is the word embedding(s) for the class name. The context vectors are shared among all classes.[3] Let $g(\cdot)$ denote the text encoder, the prediction probability is then

$$p(y|\boldsymbol{x}) = \frac{\exp(\text{sim}(\boldsymbol{x}, g(\boldsymbol{t}_y))/\tau)}{\sum_{i=1}^{K}\exp(\text{sim}(\boldsymbol{x}, g(\boldsymbol{t}_i))/\tau)}. \tag{2}$$

To adapt CLIP to a downstream image recognition dataset, a cross-entropy loss can be used as the learning objective. Since the text encoder $g(\cdot)$ is differentiable, gradients can be propagated all the way back to update the context vectors. Note that the base model of CLIP is frozen in the entire training process (ours too).

3.2. CoCoOp: Conditional Context Optimization

CoOp is a data-efficient approach allowing the context vectors to be trained with only a few labeled images in a downstream dataset. However, as discussed CoOp is not generalizable to wider unseen classes within the same task. We argue that instance-conditional context can generalize better because it shifts the focus away from a specific set of classes—for reducing overfitting—to each input instance, and hence to the entire task.

A straightforward way to implement CoCoOp is to build M neural networks to get M context tokens. However, such a design would require $M\times$ the size of a neural network, which is much larger than having M context vectors as in CoOp. Here we propose a parameter-efficient design that works very well in practice. Specifically, on top of the M context vectors, we further learn a lightweight neural network, called Meta-Net, to generate for each input a conditional token (vector), which is then combined with the context vectors. See Figure 2 for a sketch of the architecture.

Let $h_{\boldsymbol{\theta}}(\cdot)$ denote the Meta-Net parameterized by $\boldsymbol{\theta}$, each context token is now obtained by $\boldsymbol{v}_m(\boldsymbol{x}) = \boldsymbol{v}_m + \boldsymbol{\pi}$ where $\boldsymbol{\pi} = h_{\boldsymbol{\theta}}(\boldsymbol{x})$ and $m \in \{1, 2, ..., M\}$. The prompt for the i-th class is thus conditioned on the input, i.e., $\boldsymbol{t}_i(\boldsymbol{x}) = \{\boldsymbol{v}_1(\boldsymbol{x}), \boldsymbol{v}_2(\boldsymbol{x}), \dots, \boldsymbol{v}_M(\boldsymbol{x}), \boldsymbol{c}_i\}$. The prediction probability is computed as

$$p(y|\boldsymbol{x}) = \frac{\exp(\text{sim}(\boldsymbol{x}, g(\boldsymbol{t}_y(\boldsymbol{x})))/\tau)}{\sum_{i=1}^{K}\exp(\text{sim}(\boldsymbol{x}, g(\boldsymbol{t}_i(\boldsymbol{x}))/\tau)}. \tag{3}$$

During training, we update the context vectors $\{\boldsymbol{v}_m\}_{m=1}^{M}$ together with the Meta-Net's parameters $\boldsymbol{\theta}$. In this work, the Meta-Net is built with a two-layer bottleneck structure (Linear-ReLU-Linear), with the hidden layer reducing the input dimension by 16×. The input to the Meta-Net is simply the output features produced by the image encoder. We leave exploration of more advanced designs for future work.

4. Experiments

Our approach is mainly evaluated in the following three problem settings: 1) generalization from base to new classes within a dataset (Section 4.1); 2) cross-dataset transfer (Section 4.2); 3) domain generalization (Section 4.3). All models used in our experiments are based on the open-source CLIP [40].[4] Before discussing the results, we provide the details of the experimental setup below.

Datasets For the first two settings, i.e., base-to-new generalization and cross-dataset transfer, we use the 11 image recognition datasets as in Zhou et al. [61], which cover a diverse set of recognition tasks. Specifically, the benchmark includes ImageNet [6] and Caltech101 [11] for classification on generic objects; OxfordPets [38], StanfordCars [28], Flowers102 [36], Food101 [2] and FGVCAircraft [35] for fine-grained classification; SUN397 [55] for scene recognition; UCF101 [46] for action recognition; DTD [5] for tex-

[3] CoOp has an alternative version that learns class-specific context, which is not considered here because it is not straightforward to transfer class-specific context to unseen classes.

[4] https://github.com/openai/CLIP.

Table 1. **Comparison of CLIP, CoOp and CoCoOp in the base-to-new generalization setting**. For learning-based methods (CoOp and CoCoOp), their prompts are learned from the base classes (16 shots). The results strongly justify the strong generalizability of conditional prompt learning. H: Harmonic mean (to highlight the generalization trade-off [54]).

(a) **Average over 11 datasets**.

	Base	New	H
CLIP	69.34	**74.22**	71.70
CoOp	**82.69**	63.22	71.66
CoCoOp	80.47	71.69	**75.83**

(b) ImageNet.

	Base	New	H
CLIP	72.43	68.14	70.22
CoOp	**76.47**	67.88	71.92
CoCoOp	75.98	**70.43**	**73.10**

(c) Caltech101.

	Base	New	H
CLIP	96.84	**94.00**	95.40
CoOp	**98.00**	89.81	93.73
CoCoOp	97.96	93.81	**95.84**

(d) OxfordPets.

	Base	New	H
CLIP	91.17	97.26	94.12
CoOp	93.67	95.29	94.47
CoCoOp	**95.20**	**97.69**	**96.43**

(e) StanfordCars.

	Base	New	H
CLIP	63.37	**74.89**	68.65
CoOp	**78.12**	60.40	68.13
CoCoOp	70.49	73.59	**72.01**

(f) Flowers102.

	Base	New	H
CLIP	72.08	**77.80**	74.83
CoOp	**97.60**	59.67	74.06
CoCoOp	94.87	71.75	**81.71**

(g) Food101.

	Base	New	H
CLIP	90.10	91.22	90.66
CoOp	88.33	82.26	85.19
CoCoOp	**90.70**	**91.29**	**90.99**

(h) FGVCAircraft.

	Base	New	H
CLIP	27.19	**36.29**	**31.09**
CoOp	**40.44**	22.30	28.75
CoCoOp	33.41	23.71	27.74

(i) SUN397.

	Base	New	H
CLIP	69.36	75.35	72.23
CoOp	**80.60**	65.89	72.51
CoCoOp	79.74	**76.86**	**78.27**

(j) DTD.

	Base	New	H
CLIP	53.24	**59.90**	56.37
CoOp	**79.44**	41.18	54.24
CoCoOp	77.01	56.00	**64.85**

(k) EuroSAT.

	Base	New	H
CLIP	56.48	**64.05**	60.03
CoOp	**92.19**	54.74	68.69
CoCoOp	87.49	60.04	**71.21**

(l) UCF101.

	Base	New	H
CLIP	70.53	**77.50**	73.85
CoOp	**84.69**	56.05	67.46
CoCoOp	82.33	73.45	**77.64**

ture classification; and finally EuroSAT [19] for satellite imagery recognition. For domain generalization experiments, we use ImageNet as the source dataset and four other variants of ImageNet that contain different types of domain shift as the target datasets, namely ImageNetV2 [43], ImageNet-Sketch [50], ImageNet-A [22] and ImageNet-R [21].

Following Zhou et al. [61], we randomly sample for each dataset a few-shot training set while using the original test set for testing. We only evaluate the highest shot number studied in Zhou et al. [61], i.e., 16 shots, which is sufficient to justify our approach. For learning-based models, the results are averaged over three runs.

Baselines The direct rival to our approach is CoOp [61], which essentially learns *static* prompts (in comparison to our *dynamic* prompts). The zero-shot method, i.e., CLIP [40] is also compared, which is based on manual prompts. It is worth mentioning that the manual prompt for each dataset was intensively tuned using *all classes in the test data* [40].

Training Details Our implementation is based on CoOp's code.[5] Throughout the experiments, we use the best available vision backbone in CLIP, i.e., ViT-B/16. Zhou et al. [61] have suggested that a shorter context length and a good initialization can lead to better performance and stronger robustness to domain shift. Therefore, we fix the context length to 4 and initialize the context vectors using the pre-trained word embeddings of "a photo of a" for both CoOp and CoCoOp. Due to the instance-conditional design, our approach is slow to train and consumes much more GPU memory than CoOp. Therefore, to ensure the model can fit into a GPU and meanwhile reduce the training time, we train CoCoOp with batch size of 1 for 10 epochs. Such a limitation is discussed in more detail in Section 5.

4.1. Generalization From Base to New Classes

Solving the weak generalizability problem of CoOp is the main focus in this research. On each of the 11 datasets, we split the classes equally into two groups, one as base classes and the other as new classes. Learning-based models, i.e., CoOp and CoCoOp, are trained using only the base classes while evaluation is conducted on the base and new classes *separately* to test generalizability. The detailed results are shown in Table 1.

[5] https://github.com/KaiyangZhou/CoOp.

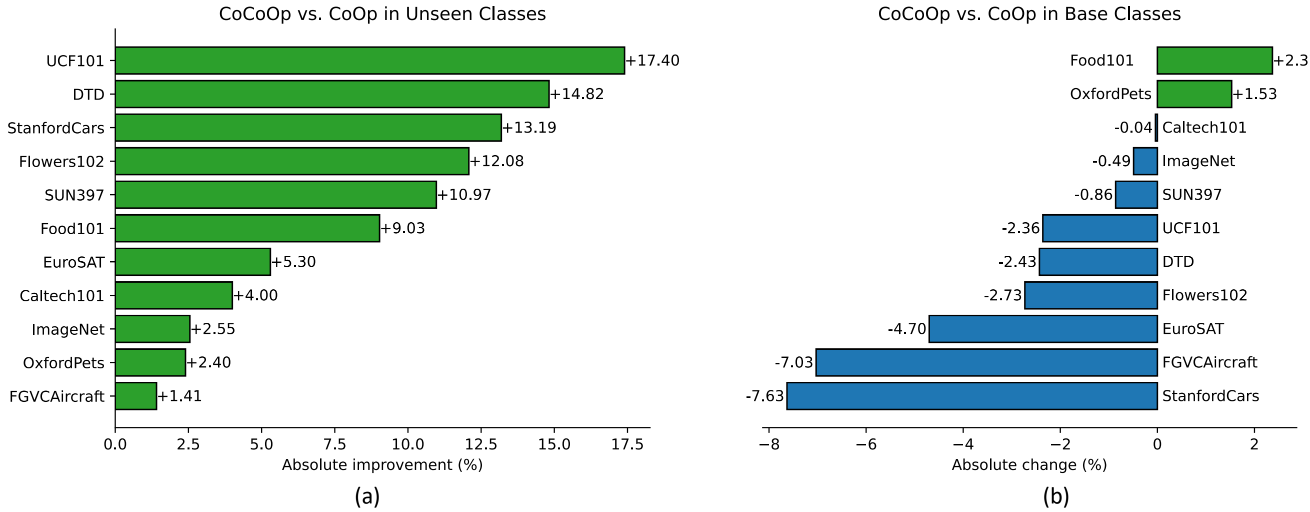


Figure 3. **Comprehensive comparisons of CoCoOp and CoOp in the base-to-new generalization setting**. (a) CoCoOp is able to gain consistent improvements over CoOp in unseen classes on all datasets. (b) CoCoOp's declines in base accuracy are mostly under 3%, which are far outweighed by the gains in generalization.

Failures of CoOp in Unseen Classes The split does not guarantee that the two class groups are equally difficult, as evidenced in CLIP's bumpy results: the base and new accuracy numbers are dramatically different.[6] Nonetheless, CoOp's new accuracy is consistently much weaker than the base accuracy on nearly all datasets, leaving a huge gap of almost 20% on average (82.69% vs 63.22%). Despite maintaining an advantage over CLIP in terms of average performance, CoOp's gains in the base classes are nearly zeroed out by the catastrophic failures in the new classes, highlighting the need to improve generalizability for learning-based prompts.

CoCoOp Significantly Narrows Generalization Gap As shown in Table 1(a), CoCoOp improves the accuracy in unseen classes from 63.22% to 71.69%, which largely reduces the gap with manual prompts. The results confirm that *instance-conditional prompts are more generalizable*. A more detailed breakdown of per-dataset improvement is visualized in Figure 3(a) where we observe more than 10% increases in accuracy on 5 out of 11 datasets. Notably, on the challenging ImageNet dataset, CoCoOp's surge from 67.88% to 70.43% represents a non-trivial progress (the 70.43% accuracy even surpasses CLIP's 68.14%).

CoCoOp's Gains in Generalization Far Outweigh Losses in Base Accuracy In comparison to CoOp, performance drops in the base classes occur for CoCoOp on most datasets (see Figure 3(b)). This is reasonable because CoOp optimizes specifically for base classes whereas *CoCoOp optimizes for each instance in order to gain more generalization over an entire task*. But it is worth noting that on the 9 datasets where CoCoOp's base accuracy drops below CoOp's, most losses are under 3% (precisely on 6 out of 9 datasets), which are far outweighed by the gains in unseen classes shown in Figure 3(a); even for those where CoCoOp suffers the biggest losses, the boosts in generalization are mostly significant enough to turn the averages into positives, e.g., StanfordCars sees the worst base accuracy drop of -7.63% but has the third-highest accuracy gain of +13.19% in the new classes, which together bring a 5.56% positive improvement for CoCoOp.

CoCoOp Is More Compelling Than CLIP When taking into account both the base and new classes, CoCoOp shows a gain of more than 4% over CLIP (75.83% vs 71.70), suggesting that *instance-conditional prompts have a better potential in capturing more generalizable elements that are relevant for a recognition task*. Theoretically, learning-based prompts have a much higher risk of overfitting base classes than manual prompts. Therefore, CLIP is a strong competitor to beat in unseen classes. Different from CoOp, we obtain promising results for CoCoOp: the new accuracy is even better than CLIP's on 4 out of 11 datasets (i.e., ImageNet, OxfordPets, Food101 and SUN397) and not too far away from CLIP's on the rest except FGVCAircraft where the gap between manual and learning-based prompts is generally large. In the ablation study on context length, we find that FGVCAircraft benefits from longer context, which is aligned with the findings in Zhou et al. [61]. To close or even overturn the gaps between manual and learning-based prompts in unseen classes, more efforts are required and we hope the insights presented in this research can help the community tackle the generalizability issue in

[6] For convenience, we refer to base accuracy as the performance in base classes; and similarly for new accuracy.

Table 2. **Comparison of prompt learning methods in the cross-dataset transfer setting**. Prompts applied to the 10 target datasets are learned from ImageNet (16 images per class). Clearly, CoCoOp demonstrates better transferability than CoOp. Δ denotes CoCoOp's gain over CoOp.

	Source	Target										
	ImageNet	Caltech101	OxfordPets	StanfordCars	Flowers102	Food101	FGVCAircraft	SUN397	DTD	EuroSAT	UCF101	*Average*
CoOp [61]	**71.51**	93.70	89.14	64.51	68.71	85.30	18.47	64.15	41.92	**46.39**	66.55	63.88
CoCoOp	71.02	**94.43**	**90.14**	**65.32**	**71.88**	**86.06**	**22.94**	**67.36**	**45.73**	45.37	**68.21**	**65.74**
Δ	**-0.49**	**+0.73**	**+1.00**	**+0.81**	**+3.17**	**+0.76**	**+4.47**	**+3.21**	**+3.81**	**-1.02**	**+1.66**	**+1.86**

Table 3. **Comparison of manual and learning-based prompts in domain generalization**. CoOp and CoCoOp use as training data 16 images from each of the 1,000 classes on ImageNet. In general, CoCoOp is more domain-generalizable than CoOp.

		Source	Target			
	Learnable?	ImageNet	ImageNetV2	ImageNet-Sketch	ImageNet-A	ImageNet-R
CLIP [40]		66.73	60.83	46.15	47.77	73.96
CoOp [61]	✓	**71.51**	**64.20**	47.99	49.71	75.21
CoCoOp	✓	71.02	64.07	**48.75**	**50.63**	**76.18**

prompt learning.

4.2. Cross-Dataset Transfer

Having demonstrated CoCoOp's generalizability within a dataset, we further show that CoCoOp has the potential to transfer beyond a single dataset, which is a much more challenging problem because the fundamentals can be totally changed across different datasets (e.g., from object recognition to texture classification). We only consider prompt learning methods in this setting.

We compare CoCoOp with CoOp by transferring context learned from ImageNet, with all 1,000 classes used, to each of the other 10 datasets. The results are detailed in Table 2. On the source dataset, the two models perform similarly. Whereas on the target datasets, CoCoOp mostly outperforms CoOp by a clear margin. Since the ImageNet classes mainly contain objects, as well as a fair amount of dog breeds, it is reasonable to see high accuracy for both models on the relevant target datasets including Caltech101 and OxfordPets.

By comparison, the performance on other datasets with distant—and more fine-grained or specialized—categories is much lower, such as FGVCAircraft and DTD (containing various textures) where the accuracy numbers are well below 50%. Nonetheless, CoCoOp exhibits much stronger transferability than CoOp on the two mentioned datasets as well as on most other fine-grained or specialized datasets.

4.3. Domain Generalization

Generalization to out-of-distribution data is a capability essential for machine learning models to succeed in practical applications [47, 60]. Zhou et al. [61] have revealed that their learnable prompts are more robust than manual prompts to domain shift. We are also interested to know if instance-conditional prompts still maintain the advantages as in previous experiments.

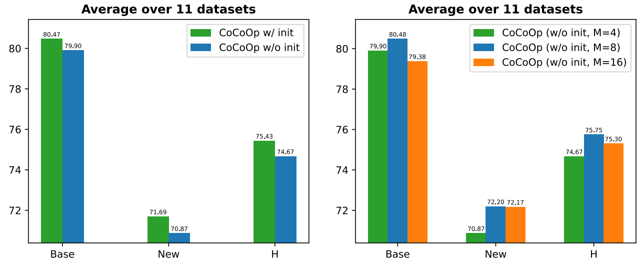
Following Zhou et al. [61], we evaluate CoCoOp's domain generalization performance by transferring the context learned from ImageNet to the four specially designed benchmarks. We also include the comparison with CLIP. Table 3 shows the results. Both prompt learning methods clearly beat CLIP on all target datasets. Compared to CoOp, CoCoOp performs slightly worse on ImageNetV2 but better on the other three. The results confirm that *instance-conditional prompts are more domain-generalizable*.

4.4. Further Analysis

Class-Incremental Test We consider a practical problem scenario where the recognition targets originally composed of base classes are expanded to include completely new classes. This problem is relevant to the existing continual learning literature [37] but different in that *the model here does not have access to any training data from new classes and needs to perform zero-shot recognition on them*. We compare CLIP, CoOp and CoCoOp using the 11 datasets. The average results are reported in Table 4.

Table 4. Recognition accuracy (average over 11 datasets) on a combination of base and new classes. The learnable models only have access to training data from base classes.

	Learnable?	Accuracy
CLIP [40]		65.22
CoOp [61]	✓	65.55
CoCoOp	✓	**69.13**



(a) Ablation on initialization. (b) Ablation on context length.

Figure 4. Ablation studies.

Table 5. CoCoOp (last row) vs a bigger CoOp on ImageNet.

Model	# params	Base	New	H
CoOp (ctx=4)	2,048	**76.47**	67.88	71.92
CoOp (ctx=60)	30,720	76.16	65.34	70.34
CoOp (ctx=4) + Meta-Net	34,816	75.98	**70.43**	**73.10**

Clearly, CoOp loses competitiveness against CLIP as their performance is similar but the former needs training data. Again, CoCoOp beats the two competitors with a significant margin.

Initialization We compare word embeddings-based initialization with random initialization, which samples from a zero-mean Gaussian distribution with 0.02 standard deviation. Figure 4(a) suggests that a proper initialization is more beneficial to both the base and new classes.

Context Length Following Zhou et al. [61], we study 4, 8 and 16 context tokens. For fair comparison, we use random initialization for all context tokens. Figure 4(b) summarizes the results on the 11 datasets. The differences in the base classes are fairly small whereas in the new classes the models with a longer context length clearly perform better.

CoCoOp vs a Bigger CoOp Since CoCoOp introduces more parameters than CoOp, namely the Meta-Net, one might question if the improvements simply come from an increased learning capacity. To clear the doubt, we remove the Meta-Net part and increase the number of context tokens in CoOp to the maximum such that CoOp's and CoCoOp's sizes are similar. The results in Table 5 show that increasing the parameter size is not the key.

5. Limitations

The first limitation is about training efficiency: CoCoOp is slow to train and would consume a significant amount of GPU memory if the batch size is set larger than one. The reason is because CoCoOp is based on an instance-conditional design that requires for each image an independent forward pass of instance-specific prompts through the text encoder. This is much less efficient than CoOp that only needs a single forward pass of prompts through the text encoder for an entire mini-batch of any size.

The second limitation is that on 7 out of the 11 datasets (see Table 1), CoCoOp's performance in unseen classes still lags behind CLIP's, indicating that more efforts are needed from the community to fully close or overturn the gaps between manual and learning-based prompts.

6. Discussion and Conclusion

Our research addresses an important issue that arises with the availability of large pre-trained AI models, i.e., how to adapt them to downstream applications. These models, also called foundation models [1], have received increasing attention from academia and industry in both the vision and NLP communities because they are so powerful in terms of their capabilities for diverse downstream tasks. However, foundation models are costly to pre-train in terms of data scale and compute resources; and typically contain an enormous number of parameters in order to develop sufficient capacity. For instance, the CLIP model [40] based on ViT-B/16 used in our experiments has a whopping 150M parameter size. These factors together highlight the need for research of *efficient adaptation methods for democratizing foundation models*.

Our studies, which follow the line of parameter-efficient prompt learning [61], provide timely insights into the generalizability issue of static prompts, and more importantly, demonstrate that a simple design based on conditional prompt learning performs superbly in a variety of problem scenarios, including generalization from base to new classes, cross-dataset prompt transfer, and domain generalization.

References

[1] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021. 8

[2] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101–mining discriminative components with random forests. In *ECCV*, 2014. 4

[3] Wei-Lun Chao, Soravit Changpinyo, Boqing Gong, and Fei Sha. An empirical study and analysis of generalized zero-

shot learning for object recognition in the wild. In *ECCV*, 2016. 3

[4] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, 2020. 3

[5] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *CVPR*, 2014. 4

[6] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009. 4

[7] Karan Desai and Justin Johnson. Virtex: Learning visual representations from textual annotations. In *CVPR*, 2021. 3

[8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*, 2019. 3

[9] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021. 4

[10] Mohamed Elhoseiny, Babak Saleh, and Ahmed Elgammal. Write a classifier: Zero-shot learning using purely textual descriptions. In *ICCV*, 2013. 3

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

[12] Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Marc'Aurelio Ranzato, and Tomas Mikolov. Devise: A deep visual-semantic embedding model. *NeurIPS*, 2013. 3

[13] Andreas Fürst, Elisabeth Rumetshofer, Viet Tran, Hubert Ramsauer, Fei Tang, Johannes Lehner, David Kreil, Michael Kopp, Günter Klambauer, Angela Bitto-Nemling, et al. Cloob: Modern hopfield networks with infoloob outperform clip. *arXiv preprint arXiv:2110.11316*, 2021. 1, 3

[14] Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. Clip-adapter: Better vision-language models with feature adapters. *arXiv preprint arXiv:2110.04544*, 2021. 1

[15] Tianyu Gao, Adam Fisch, and Danqi Chen. Making pre-trained language models better few-shot learners. *arXiv preprint arXiv:2012.15723*, 2020. 1

[16] Lluis Gomez, Yash Patel, Marçal Rusiñol, Dimosthenis Karatzas, and CV Jawahar. Self-supervised learning of visual features through embedding images into text topic spaces. In *CVPR*, 2017. 3

[17] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, 2020. 3

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 1, 4

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

[20] Olivier J. Hénaff, Aravind Srinivas, Jeffrey De Fauw, Ali Razavi, Carl Doersch, S. M. Ali Eslami, and Aäron van den Oord. Data-efficient image recognition with contrastive predictive coding. In *ICML*, 2020. 3

[21] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, Dawn Song, Jacob Steinhardt, and Justin Gilmer. The many faces of robustness: A critical analysis of out-of-distribution generalization. *ICCV*, 2021. 5

[22] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. In *CVPR*, 2021. 5

[23] Dat Huynh and Ehsan Elhamifar. Fine-grained generalized zero-shot learning via dense attribute-based attention. In *CVPR*, 2020. 3

[24] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V Le, Yunhsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *ICML*, 2021. 1, 2, 3

[25] Zhengbao Jiang, Frank F Xu, Jun Araki, and Graham Neubig. How can we know what language models know? *ACL*, 2020. 1, 3

[26] Armand Joulin, Laurens Van Der Maaten, Allan Jabri, and Nicolas Vasilache. Learning visual features from large weakly supervised data. In *ECCV*, 2016. 3

[27] Chen Ju, Tengda Han, Kunhao Zheng, Ya Zhang, and Weidi Xie. Prompting visual-language models for efficient video understanding. *arXiv preprint arXiv:2112.04478*, 2021. 3

[28] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *ICCV-W*, 2013. 4

[29] Jimmy Lei Ba, Kevin Swersky, Sanja Fidler, et al. Predicting deep zero-shot convolutional neural networks using textual descriptions. In *ICCV*, 2015. 3

[30] Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. *arXiv preprint arXiv:2104.08691*, 2021. 1, 3

[31] Ang Li, Allan Jabri, Armand Joulin, and Laurens van der Maaten. Learning visual n-grams from web data. In *ICCV*, 2017. 3

[32] Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. *arXiv preprint arXiv:2101.00190*, 2021. 1, 3

[33] Yangguang Li, Feng Liang, Lichen Zhao, Yufeng Cui, Wanli Ouyang, Jing Shao, Fengwei Yu, and Junjie Yan. Supervision exists everywhere: A data efficient contrastive language-image pre-training paradigm. *arXiv preprint arXiv:2110.05208*, 2021. 1, 3

[34] Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *arXiv preprint arXiv:2107.13586*, 2021. 1, 3

[35] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classi-

fication of aircraft. *arXiv preprint arXiv:1306.5151*, 2013. 4

[36] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *ICVGIP*, 2008. 4

[37] German I Parisi, Ronald Kemker, Jose L Part, Christopher Kanan, and Stefan Wermter. Continual lifelong learning with neural networks: A review. *Neural Networks*, 2019. 7

[38] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. Cats and dogs. In *CVPR*, 2012. 4

[39] Fabio Petroni, Tim Rocktäschel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, Alexander H Miller, and Sebastian Riedel. Language models as knowledge bases? In *EMNLP*, 2019. 3

[40] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 1, 2, 3, 4, 5, 7, 8

[41] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 2019. 3

[42] Yongming Rao, Wenliang Zhao, Guangyi Chen, Yansong Tang, Zheng Zhu, Guan Huang, Jie Zhou, and Jiwen Lu. Denseclip: Language-guided dense prediction with context-aware prompting. In *CVPR*, 2022. 3

[43] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *ICML*, 2019. 5

[44] Taylor Shin, Yasaman Razeghi, Robert L Logan IV, Eric Wallace, and Sameer Singh. Autoprompt: Eliciting knowledge from language models with automatically generated prompts. In *EMNLP*, 2020. 1, 3

[45] Richard Socher, Milind Ganjoo, Hamsa Sridhar, Osbert Bastani, Christopher D Manning, and Andrew Y Ng. Zero-shot learning through cross-modal transfer. In *NeurIPS*, 2013. 3

[46] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012. 4

[47] Rohan Taori, Achal Dave, Vaishaal Shankar, Nicholas Carlini, Benjamin Recht, and Ludwig Schmidt. Measuring robustness to natural distribution shifts in image classification. In *NeurIPS*, 2020. 7

[48] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017. 1, 3, 4

[49] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *CVPR*, 2015. 2

[50] Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P Xing. Learning robust global representations by penalizing local predictive power. In *NeurIPS*, 2019. 5

[51] Wei Wang, Vincent W Zheng, Han Yu, and Chunyan Miao. A survey of zero-shot learning: Settings, methods, and applications. *TIST*, 2019. 3

[52] Xiaolong Wang, Yufei Ye, and Abhinav Gupta. Zero-shot recognition via semantic embeddings and knowledge graphs. In *CVPR*, 2018. 3

[53] Mitchell Wortsman, Gabriel Ilharco, Mike Li, Jong Wook Kim, Hannaneh Hajishirzi, Ali Farhadi, Hongseok Namkoong, and Ludwig Schmidt. Robust fine-tuning of zero-shot models. *arXiv preprint arXiv:2109.01903*, 2021. 1

[54] Yongqin Xian, Bernt Schiele, and Zeynep Akata. Zero-shot learning-the good, the bad and the ugly. In *CVPR*, 2017. 3, 5

[55] Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *CVPR*, 2010. 2, 4

[56] Yuan Yao, Ao Zhang, Zhengyan Zhang, Zhiyuan Liu, Tat-Seng Chua, and Maosong Sun. Cpt: Colorful prompt tuning for pre-trained vision-language models. *arXiv preprint arXiv:2109.11797*, 2021. 1, 3

[57] Kai Yi, Xiaoqian Shen, Yunhao Gou, and Mohamed Elhoseiny. Exploring hierarchical graph representation for large-scale zero-shot image classification. *arXiv preprint arXiv:2203.01386*, 2022. 3

[58] Yuhao Zhang, Hang Jiang, Yasuhide Miura, Christopher D Manning, and Curtis P Langlotz. Contrastive learning of medical visual representations from paired images and text. *arXiv preprint arXiv:2010.00747*, 2020. 2

[59] Zexuan Zhong, Dan Friedman, and Danqi Chen. Factual probing is [mask]: Learning vs. learning to recall. In *NAACL*, 2021. 1, 3

[60] Kaiyang Zhou, Ziwei Liu, Yu Qiao, Tao Xiang, and Chen Change Loy. Domain generalization in vision: A survey. *arXiv preprint arXiv:2103.02503*, 2021. 7

[61] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *arXiv preprint arXiv:2109.01134*, 2021. 1, 2, 3, 4, 5, 6, 7, 8