CONTRASTIVE PROMPT TUNING IMPROVES GENERALIZATION IN VISION-LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Prompt tuning, which focuses on learning continuous text prompts for adapting large vision-language models, has attracted much attention in recent years. While prior works show promising performance over the hand-crafted prompts, they typically use cross-entropy loss for learning prompts, which limits their generalization capability in many real-world scenarios. Motivated by the effectiveness of contrastive learning for improved generalization, we introduce Contrastive Prompt Tuning (CPT), an incredibly simple yet highly efficient framework that explicitly optimizes for the learned prompts to be consistent with the image space. In particular, combined with cross-entropy loss, our contrastive losses help learning prompts so that the model has consistent predictions across different views of an image while also maintaining the consistency of pairwise similarities among different images. Extensive experiments on a battery of datasets demonstrate that our proposed method significantly outperforms the existing methods in improving model's generalization, while also achieving consistent improvements in few-shot in-domain performance for a wide variety of vision-language models.

1 INTRODUCTION

Large vision-language models (VLMs) (Radford et al., 2021; Jia et al., 2021; Li et al., 2021; 2022; Wu et al., 2021), with appropriately designed text prompts have achieved promising progress on several downstream recognition tasks. For instance, one can prepend a category name with a prompt "a photo of a" (e.g., "a photo of a cat") and then use as input to the CLIP (Radford et al., 2021) text encoder to classify images. However, identifying the right hand-crafted prompt is a non-trivial task, which often requires significant amount of time and domain-specific heuristics.

This has motivated much work on prompt tuning (Zhou et al., 2022b;a; Lu et al., 2022), which aims to learn soft prompts using few labeled data from the downstream tasks, while keeping the pretrained model parameters fixed. Although ubiquitous in finding better prompts compared to hand-crafted ones, the prompts learned using such methods often have poor generalization to different natural distribution shifts (i.e, when transferred to recognizing new classes within the same or different datasets or even recognizing same classes across different domains of data). We argue that such a problem is caused by the use of only cross-entropy loss for learning prompts, which has shown sub-optimal generalization and instability in many works (Liu et al., 2016; Cao et al., 2019).

Meanwhile, recent research shows that representations learned with contrastive learning have larger intra-class variations and hence are more generalizable over their supervised counterparts (Sun et al., 2019; Islam et al., 2021; Zhao et al., 2020). However, despite recent progress, self-supervised contrastive objectives have never been used for prompt learning in VLMs. Motivated by this, in this paper, we explore the following natural, yet important question: *whether and how contrastive learning could be exploited for improved generalization of prompts in vision-language models?*

To this end, we introduce Contrastive Prompt Tuning (`CPT`), a simple yet effective framework that explicitly optimizes for the learned prompts to be consistent in the image space. Specifically, given a few labeled examples, we augment the standard cross-entropy loss with two additional contrastive loss terms driven by an hypothesis that contrastive losses can improve generalization by making the model outputs invariant to small input perturbations. The first term helps learning prompts by encouraging the model to have consistent predictions across different views of an image while the second term maintains the consistency of pairwise similarities among different images. To the best

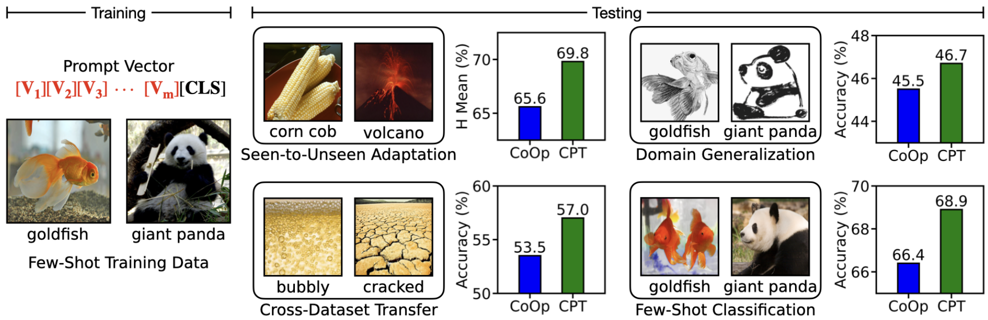


Figure 1: **Prompt Generalization in Vision-Language Models.** Figure shows four testing scenarios and the corresponding bar charts comparing the average performance of CoOp (Zhou et al., 2022b) and **CPT** on CLIP with ResNet50 (Radford et al., 2021). Our **CPT** approach, which learns the prompts to be consistent in the image space, outperforms CoOp across all the settings. Best viewed in color.

of our knowledge, our work is the first to successfully integrate a self-supervised contrastive learning objective for prompt tuning of vision-language models.

We evaluate the generalization of **CPT** in four different image classification settings that can occur naturally in real-world scenarios: seen-to-unseen classes adaptation within a dataset, cross-dataset transfer, domain generalization and the standard few-shot classification setting without any distribution shift, as shown in Figure 1. Extensive experiments on a battery of datasets (in total 15) with a diverse set of vision-language models (in total 10 models) demonstrate the superiority of **CPT** over existing methods. For the setting of seen-to-unseen classes generalization with CLIP-RN50 (Radford et al., 2021), **CPT** yields an average 4.2% improvement over CoOp (Zhou et al., 2022b), while also very competitive with the most recent SOTA method (Zhou et al., 2022a) that requires an additional specialized meta-network for learning prompts (17 times more trainable parameters compared to **CPT**). The gains over CoOp are as large as 3.5% and 1.2% for the cross-dataset transfer and domain generalization settings without the need for additional unlabeled data. Further, **CPT** consistently outperforms CoOp while it is on par or better than SOTA adaptation methods (e.g., Tip-Adapter (Zhang et al., 2021)) with significantly less number of trainable parameters in the few-shot in-domain setting. In summary, our findings conclusively show that **CPT** improves performance of prompt tuning across most evaluations by a significant margin, an encouraging signal for the general utility of contrastive learning in the context of generalizable prompt tuning for VLMs.

Our approach also brings another advantage in terms of easy implementation compared to many recent adaptation methods (e.g., CoCoOp (Zhou et al., 2022a), UPL (Huang et al., 2022), ProDA (Lu et al., 2022), Tip-Adapter (Zhang et al., 2021)): with only few lines of code change in PyTorch, **CPT** can be applied to a wide variety of vision-language models, like, CLIP (Radford et al., 2021), DeCLIP (Li et al., 2021), FILIP (Yao et al., 2021), CLOOB (Fürst et al., 2021), and CyCLIP (Goel et al., 2022). We hope our simple approach and efforts in benchmarking the results of different methods will open up avenues for future research in prompt learning for VLMs. We will make all our codes, data and models publicly available upon acceptance.

2 Related Work

Vision-Language Models. Much progress has been made in developing VLMs using single-stream (Chen et al., 2020b; Li et al., 2019; Su et al., 2019; Li et al., 2020) or dual-stream paradigms (Radford et al., 2021; Jia et al., 2021; Li et al., 2021; Goel et al., 2022; Tan & Bansal, 2019; Li et al., 2022). Our approach is most related to the dominant dual-stream paradigm that decouples the image encoder and text encoder and extracts features for images and texts respectively. Representative works like CLIP (Radford et al., 2021) and ALIGN (Jia et al., 2021) have greatly revolutionized computer vision by allowing zero-shot transfer to a variety of downstream classification tasks. A very few methods have recently attempted learning transferable features more efficiently, using additional supervision (Li et al., 2021; Mu et al., 2021), finer-grained interactions (Yao et al., 2021), modern Hopfield networks (Fürst et al., 2021), optimal transport distillation (Wu et al., 2021), cycle consistency (Goel et al., 2022), and hierarchical feature alignment (Gao et al., 2022). Orthogonal to developing new learning strategies or VLM architectures, our work addresses the emerging problem of efficiently adapting large pretrained vision-language models to downstream applications.

Prompt Tuning. Prompt tuning for efficient adaptation of vision-language models has been studied from multiple perspectives (Zhou et al., 2022b; Huang et al., 2022). Inspired by prompt tuning from NLP (Zhong et al., 2021; Lester et al., 2021), CoOp (Zhou et al., 2022b) minimizes the prediction error using the cross-entropy loss with respect to the learnable prompt vectors. While ProDA (Lu et al., 2022) learns diverse prompts from data to handle the variance of visual representations, UPL (Huang et al., 2022) proposes an unsupervised prompt learning framework without requiring any annotations of the target dataset. A test-time prompt tuning framework that does not need any training data or annotations to optimize the prompt is also proposed in (Shu et al., 2022). Similar in spirit, CLIP-Adapter (Gao et al., 2021) and Tip-Adapter (Zhang et al., 2021) propose to adapt vision-language models by training an additional adapter network on top of the pretrained models using a small set of labeled data. While these approaches show reasonable improvements over hand-crafted prompts, they often suffer from poor generalization under different data distribution shifts. Recently, CoCoOp (Zhou et al., 2022a) utilizes a Meta-Net to generate image-dependent prompt vectors for improved generalization. Alternately, we propose a much simpler yet effective method which leverages contrastive losses to learn more generalizable prompts without any additional network, making it significantly more parameter efficient than CoCoOp. In addition, **`CPT`** makes prompt learning extremely fast and more computationally efficient than CoCoOp, which is unwieldy to train and requires very small batch sizes during training for memory constraints.

Contrastive Learning. Contrastive learning is becoming increasingly attractive for learning robust representations of both unimodal (Chen et al., 2020a; Grill et al., 2020; He et al., 2020; Oord et al., 2018) and multimodal data (Yuan et al., 2021; Akbari et al., 2021; Radford et al., 2021). Many variants have been recently proposed that learn representations by modeling the relationship between different instances (Dwibedi et al., 2021; Zheng et al., 2022; Abbasi Koohpayegani et al., 2020; Wei et al., 2020). Contrastive learning has also been used in supervised settings, where labels are used to guide the choice of positive and negative pairs (Khosla et al., 2020). While our approach is inspired by these methods, we propose contrastive prompt tuning for improving generalization in vision-language models, which to our best knowledge has not been explored in the literature.

3 METHODOLOGY

Given a pretrained vision-language model (e.g., CLIP Radford et al. (2021)), the goal of our proposed **`CPT`** is to learn a single prompt using only a few labeled training images for efficient yet generalizable adaptation of the model to several downstream tasks. Below we first describe basic functioning of VLMs with prompt tuning, then we elaborate on the technical details and the working principle of **`CPT`** in Section 3.2. An overview of our approach is illustrated in Figure 2.

3.1 PRELIMINARIES

Vision-Language Models. Dual stream VLMs jointly train an image encoder $f(.)$ and a text encoder $g(.)$ on data composed of image-text pairs. Given an image $\mathbf{x}$, the image encoder maps it to the feature space and outputs the l_2-normalized image embedding $\mathbf{z} = f(\mathbf{x})/||f(\mathbf{x})||_2 \in \mathbb{R}^d$ of dimension d. Similarly, the corresponding natural language description of $\mathbf{x}$ is preprocessed using an embedding layer to get $\mathbf{t}$ and is then fed to the text encoder to obtain the normalized text embedding $\mathbf{w} = g(\mathbf{t})/||g(\mathbf{t})||_2 \in \mathbb{R}^d$. Recent VLMs (e.g. CLIP Radford et al. (2021), DeCLIP Li et al. (2021), etc.) use variants of the InfoNCE loss (Oord et al., 2018) to train on large image-text data with the idea of learning perception from supervision contained in natural language.

Prompt Engineering. Once the encoders $f(.)$ and $g(.)$ are pretrained, using them for zero-shot prediction requires designing specific text descriptions (a.k.a prompts) to pair the test images. Given the C class names of a downstream task, generally a default prompt of "`a photo of a {class}`" is used to generate the natural language class descriptions $\{\mathbf{t}_c\}_{c=1}^C$ resulting in text embeddings $\{\mathbf{w}_c\}_{c=1}^C$. For a test image $\mathbf{x}$ with embedding $\mathbf{z}$, the prediction probability is calculated as:

$$p(y|\mathbf{x}) = \frac{e^{\mathbf{z}^\mathrm{T}\mathbf{w}_y/\tau}}{\sum_{c=1}^{C} e^{\mathbf{z}^\mathrm{T}\mathbf{w}_c/\tau}} \tag{1}$$

Prompt engineering focusses on designing prompts customised to the downstream dataset to significantly improve zero-shot performance. E.g. "`a photo of a {label}, a type of pet`" is a more appropriate prompt for a pet classification dataset. However, prompt engineering is a manual, intuition-guided trial and error process, which can take a long time for appropriate design.

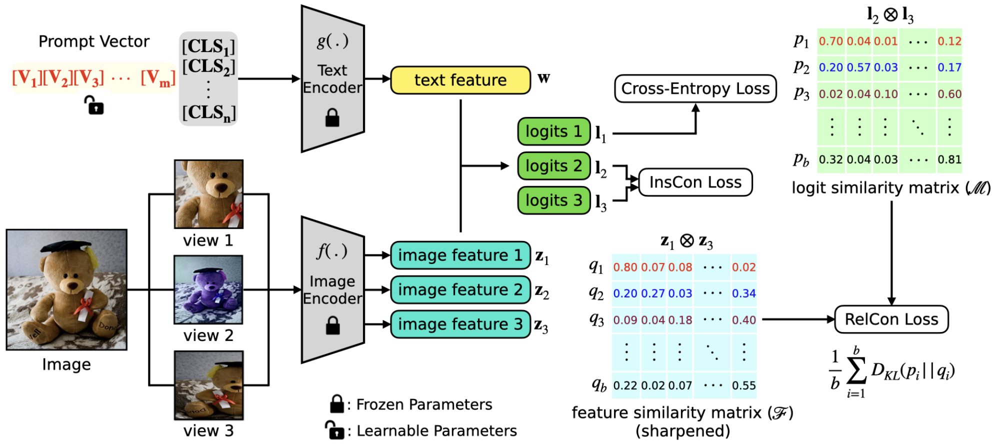


Figure 2: **An overview of our Contrastive Prompt Tuning (CPT) approach**. `CPT` learns prompt by augmenting the cross entropy loss with two self-supervised contrastive losses. The instance contrastive (InsCon) loss encourages learning instance discriminative features invariant to different views. The relational consistency (RelCon) loss makes the logit space consistent with the image space with respect to various inter-image semantic relationships. Despite being frustratingly simple, `CPT` is effective in learning generalizable prompts without any additional use of parameters. See Section 3 for more details. Best viewed in color.

Prompt Tuning. In order to overcome the inefficiency of handcrafted prompts, prompt tuning attempts to learn continuous vectors of each token position utilizing a few labeled data. Specifically, M learnable vectors $\{\mathbf{v}_i\}_{i=1}^{M}$ along with the C class name word embeddings $\{\mathbf{c}_i\}_{i=1}^{C}$ are used to form the prompts as $\{\mathbf{t}_c\}_{c=1}^{C} = \{\mathbf{v}_1, \mathbf{v}_2, ..., \mathbf{v}_M, \mathbf{c}_c\}_{c=1}^{C}$. The vectors $\mathbf{v}_i$'s can be optimized to adapt to a downstream task by propagating gradients of any loss function through the text encoder $g(.)$. Till now, the use of only cross-entropy loss for prompt tuning has limited the generalization ability of the prompt to various real-world downstream tasks.

3.2 CPT: Contrastive Prompt Tuning

We propose Contrastive Prompt Tuning (`CPT`) which leverages self-supervised contrastive learning to learn prompts that are more generalizable to unseen classes and domains. Specifically, we achieve this by encouraging the prompt to be instance-wise discriminative while retaining the interrelationships between various images. As shown in Figure 2, given a few-shot dataset with C classes and the VLM encoders $\{f(.), g(.)\}$, our goal is to learn the M prompt vectors $\mathbf{t} = \{\mathbf{v}_1, \mathbf{v}_2, ..., \mathbf{v}_M\}$. Given a batch of labeled images $\mathbf{x}^b = \{x_i, y_i\}_{i=1}^{b}$ of batchsize b, we first obtain three different views of the images $\mathbf{x}^b_{\text{view1}}$, $\mathbf{x}^b_{\text{view2}}$, and $\mathbf{x}^b_{\text{view3}}$ using a weak, strong, and weak augmentation, respectively. The images are then forwarded through the image encoder $f(.)$ to obtain the corresponding image embeddings $\mathbf{z}_1, \mathbf{z}_2$, and $\mathbf{z}_3$ each of dimension $b \times d$. On the other hand, the learnable prompt vectors $\mathbf{t}$ along with the C class name word embeddings $\{\mathbf{c}_i\}_{i=1}^{M}$ are used to form the prompts as $\{\mathbf{t}_c\}_{c=1}^{C} = \{\mathbf{v}_1, \mathbf{v}_2, ..., \mathbf{v}_M, \mathbf{c}_c\}_{c=1}^{C}$ and then are forwarded through the text encoder $g(.)$ to obtain the text embedding $\mathbf{w}$ of dimension $C \times d$. Prediction logits are then computed as $\mathbf{l}_1 = \mathbf{z}_1\mathbf{w}^{\text{T}}$, $\mathbf{l}_2 = \mathbf{z}_2\mathbf{w}^{\text{T}}$, $\mathbf{l}_3 = \mathbf{z}_3\mathbf{w}^{\text{T}}$ each of dimension $b \times C$. In order to capture the categorical information from the given ground truth labels, we apply a cross-entropy loss on the logit $\mathbf{l}_1$ as:

$$\mathcal{L}_{\text{CE}}(x_i, y_i) = -\sum_{k=1}^{C} (y_i)_k \log(\text{softmax}(\mathbf{l}_{1i}))_k \tag{2}$$

Use of only cross-entropy loss in few-shot setting is prone to overfitting to small training data and restricts generalization of learned prompts to unseen images. In order to tackle this, we incorporate two additional contrastive losses as follows. First, we apply an instance contrastive loss (Chen et al., 2020a) on the logits $\mathbf{l}_1$ and $\mathbf{l}_2$ such that the prompt learns to predict different views of same image (positives) similarly, while views of different images (negatives) differently as follows:

$$\mathcal{L}_{\text{InsCon}}(\mathbf{l}_1, \mathbf{l}_2) = -\log \frac{\exp(\text{sim}(\mathbf{l}_{1i}, \mathbf{l}_{2i})/\tau)}{\sum_{j=1}^{b} \mathbb{1}_{[j \neq i]} \exp(\text{sim}(\mathbf{l}_{1i}, \mathbf{l}_{2j})/\tau)} \tag{3}$$

where, sim(u,v) = $u^T v/||u||||v||$ denotes cosine similarity between l_2-normalized vectors u and v, and τ is temperature parameter. $\mathcal{L}_{\text{InsCon}}$ encourages the prompt to learn instance discriminative features from the images. Moreover, it is essential for a good prompt to capture various semantic relationships between different images to be generalizable across distribution shifts. Thus, inspired by (Fang et al., 2021), we propose to use a relational consistency loss for prompt learning as follows:

$$\mathcal{L}_{\text{RelCon}}(\mathcal{F}, \mathcal{M}) = \frac{1}{b}\sum_{i=1}^{b} D_{KL}(p_i||q_i) = \sum_{i=1}^{b}\sum_{j=1}^{b}(p_i)_j \log\frac{(p_i)_j}{(q_i)_j} \tag{4}$$

where, $D_{KL}(.)$ represents Kullback-leibler (KL) divergence, $\mathcal{F}$, $\mathcal{M}$ are the feature similarity matrix and the logit similarity matrix obtained as the outer products $\mathbf{z}_1 \otimes \mathbf{z}_3$ and $\mathbf{l}_2 \otimes \mathbf{l}_3$. p_i's and q_i's are the softmax normalized rows of $\mathcal{M}$, $\mathcal{F}$, with sharpening temperatures $\tau_{\mathbf{l}}$ and $\tau_{\mathbf{z}}$ respectively, as shown in Figure 2. With this cross-modal design, we want to learn prompts to make the logit space consistent with various inter-image semantic relationships in the image feature space. Note that we only use the features of weakly augmented views to compute the feature similarity matrix $\mathcal{F}$, since using aggressive augmentations can distort and hamper the capture of semantic information between different images. Finally, we iteratively optimize the total loss function $\mathcal{L}_{total} = \mathcal{L}_{\text{CE}} + \lambda(\mathcal{L}_{\text{InsCon}} + \mathcal{L}_{\text{RelCon}})$ and update the learnable prompt through standard backpropagation. λ is a weight to balance the impact of contrastive loss terms. To reduce the number of hyper-parameters, we use the same weight λ for both $\mathcal{L}_{\text{InsCon}}$ and $\mathcal{L}_{\text{RelCon}}$ in all our experiments. Note that the pretrained vision-language model is frozen and the prompt is the only learnable parameter in our approach.

Additionally, we incorporate a memory buffer of size $100\times$ of the batch size, to instill information from a diverse set of images and their views, which we find is essential for learning a generalizable prompt. Furthermore, as a regularizer and making learning prompts invariant to the position of the class token, we randomly choose the position of the class token between "start", "mid", and "end" in every iteration following an uniform distribution. Algorithm 1 summarizes the proposed approach in PyTorch-style pseudocode. Once the training is completed, the learned prompt can be used with the class embeddings appended at the end for any desired downstream task.

Algorithm 1 : **`CPT`** in a PyTorch-like style.

```
# image_encoder f, text_encoder g
# I[b, h, w, c] - minibatch of images
# T[b, l] - minibatch of texts
# feature memory buffer - FMB [d, 100b]
# logits memory buffer - LMB [n_cls, 100b]
# generate views
I_view1 = weak_aug(I)
I_view2 = strong_aug(I)
I_view3 = weak_aug(I)
# extract feature representations
z1 = f(I_view1) #[b, d]
z2 = f(I_view2) #[b, d]
z3 = f(I_view3) #[b, d]
dequeue_and_enqueue(FMB, z3)
w = g(Prompt Vector) #[n_cls, d]
# obtain logits [b, n_cls]
l1 = (z1 @ w.T)
l2 = (z2 @ w.T)
l3 = (z3 @ w.T)
dequeue_and_enqueue(LMB, l3)
# compute cross-entropy loss
loss_CE = CrossEntropy(l1, labels)
# compute instance-wise contrastive loss
loss_InsCon = SimCLR(l2, l3)
# compute relational consistency loss
feat_sim_mat = softmax(z1 @ FMB.T / tau_z,
        dim=1) # [b, 100b]
logit_sim_mat = softmax(l2 @ LMB.T / tau_l
        , dim=1) # [b, 100b]
loss_RelCon = torch.sum(-feat_sim_mat *
        log(logit_sim_mat), dim=-1).mean()
# total loss
loss = loss_CE + lossInsCon + loss_RelCon
# compute gradients and optimize
loss.backward()
optimizer.step()
```

4 EXPERIMENTS

In this section, we examine our contrastive prompt tuning approach to answer three key research questions. Q1: To what extent contrastive learning benefits generalization of prompt tuning when learned prompts are transferred across different classes and datasets? Q2: Can **`CPT`** be universally effective across a wide range of pretrained VLMs of different sizes? Q3: Beyond generalization, can **`CPT`** still improve prompt tuning in few-shot classification setting without distribution shift?

4.1 EXPERIMENTAL SETUP

Datasets. Following (Zhou et al., 2022b;a), we evaluate the performance of **`CPT`** using 15 downstream classification datasets, including general object recognition (ImageNet (Deng et al., 2009) and Caltech101 (Fei-Fei et al., 2004)), fine-grained object recognition (OxfordPets (Parkhi et al., 2012), StanfordCars (Krause et al., 2013), Flowers102 (Nilsback & Zisserman, 2008), Food101 (Bossard et al., 2014) and FGVCAircraft (Maji et al., 2013)), scene recognition (SUN397 (Xiao et al., 2010)), texture recognition (DTD (Cimpoi et al., 2014)), satellite image recognition (EuroSAT (Helber et al., 2019), action recognition (UCF101 (Soomro et al., 2012)), and four

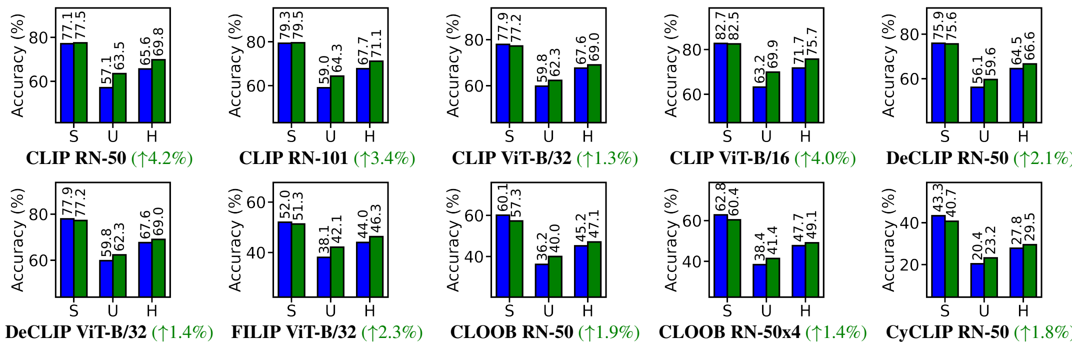


Figure 3: **Seen to Unseen Classes Adaptation.** Figure shows bar charts comparing the average performance on 11 datasets of **CPT** with CoOp on seen classes (S), unseen classes (U), and their harmonic mean (H) on 10 varieties of VLM backbones. Our **CPT** approach outperforms CoOp consistently on all the models. The blue bars represent CoOp, green bars represent **CPT**. Best viewed in color.

variants of ImageNet with domain shifts (ImageNetV2 (Recht et al., 2019), ImageNet-Sketch (Wang et al., 2019), ImageNet-A (Hendrycks et al., 2021b) and ImageNet-R (Hendrycks et al., 2021a)). We use the first 11 datasets for seen-to-new classes adaptation, cross-dataset transfer and few-shot classification experiments. For domain generalization experiments, we use ImageNet as source dataset and four of its variants as target datasets. We use the standard splits provided in (Zhou et al., 2022b) for training and the original test/validation set for testing on all datasets.

Models. We experiment with 10 publicly available pretrained VLMs of varying architectures and sizes from CLIP (Radford et al., 2021), DeCLIP (Li et al., 2021), FILIP (Yao et al., 2021), CLOOB (Fürst et al., 2021), and CyCLIP (Goel et al., 2022): CLIP ResNet-50, CLIP ResNet-101, CLIP ViT-B/32, CLIP ViT-B/16, DeCLIP ResNet-50, DeCLIP ViT-B/32, FILIP ViT-B/32, CLOOB ResNet-50, CLOOB ResNet-50x4, CyCLIP ResNet-50.

Baselines. We compare our approach with the following baselines. (1) Zero-shot CLIP that uses hand-crafted prompts for downstream classification, (2) CoOp (Zhou et al., 2022b) that learns prompt by only minimizing the cross-entropy loss, (3) a state-of-the-art prompt tuning method for CLIP, CoCoOp Zhou et al. (2022a) that uses an additional meta-network for predicting prompts, (4) linear-probe CLIP that uses a logistic regression classifier on the features of training images. We also compare with recent CLIP adaptation methods including CLIP-Adapter (Gao et al., 2021), and Tip-Adapter (Zhang et al., 2021) in few-shot classification settings. We directly quote numbers reported in published papers when possible or use the source code released by CoOp (Zhou et al., 2022b) authors under same experimental settings for a fair comparison.

Implementation Details. Following (Zhou et al., 2022b), we set the number of tokens in each prompt to 16 with random intialization for all the experiments except for the seen-to-unseen classes adaptation and experiments in Table 2 and Table 3, where we set it to 4 and initialize with the word embeddings of "a photo of a" as in (Zhou et al., 2022a). For few-shot classification, we follow CLIP (Radford et al., 2021), which learns with 1, 2, 4, 8, and 16 labeled samples per class on each downstream task. The loss weight coefficient is set to $\lambda = 0.1$. The temperature values τ, $\tau_{\mathbf{z}}$ and $\tau_{\mathbf{l}}$ are set to 0.5, 0.04 and 0.07, respectively. We use a batch size of either 8 or 32, except in ImageNet for which we used a batch size of 128, and train for either 50 or 200 epochs based on the dataset with a learning rate of 0.002. For generating multiple views, we compose strong augmentations using RandAug, color jittering, random grayscaling and blurring, while for weak augmentation we simply use random resized cropping and random horizontal flipping. We run all the experiments for three times with different random seeds and report the mean numbers in all our testing scenarios. We use one NVIDIA Tesla A100 GPU for training all our models.

4.2 Generalization from Seen to Unseen Classes

Following (Zhou et al., 2022a), we show the generalization performance of different prompt tuning methods, namely CoOp (Zhou et al., 2022b) and CoCoOp (Zhou et al., 2022a) including **CPT** by training on seen (base) classes while evaluating on both seen and unseen (new) classes.

Comparison with Vanilla Prompt Tuning (CoOp). We first compare our approach **CPT** with the vanilla prompt tuning, CoOP to show how much performance improvement **CPT** can achieve across different VLMs. As shown in Figure 3, **CPT** consistently outperforms CoOp in improving generalization performance across a wide variety of models. The gains over CoOp are as large as 4.2% on CLIP-RN-50, conforming the hypothesis that contrastive learning can significantly improve generalization of learned prompts to recognize unseen classes. Figure 4 shows absolute improvement over CoOp on both seen and unseen classes for each of the 11 downstream datasets. As expected, **CPT** improves the performance of CoOp in unseen classes on all datasets (see Figure 4(top)), while performance drops marginally in the base classes of few datasets (see Figure 4(bottom)).

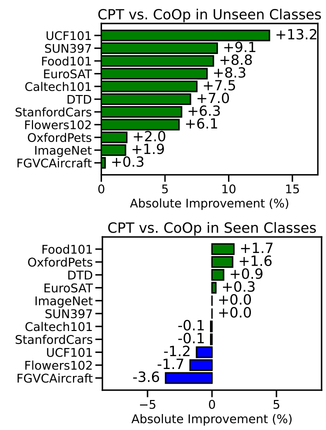


Figure 4: **Absolute improvement over CoOp w/ CLIP RN-50.** Bar charts show improvement over CoOp on seen and unseen classes for each datasets.

Comparison with CoCoOp. Table 1 shows the comparison of our **CPT** approach with CoCoOp including Zero-shot CLIP (ZS CLIP) and CoOp under the same experimental settings (w/ CLIP ViT-B/16). As expected, **CPT** significantly outperforms ZS CLIP on all datasets as handcrafted prompts are naturally worse in generalization, while learnable prompts has the ability to learn the intricate differences between the finely differing categories from data. When compared to CoCoOp, **CPT** achieves very competitive average performance without requiring any additional meta-network as in CoCoOp (Zhou et al., 2022a). However, by simply ensembling two learned prompts, **CPT** $_{2E}$ can outperform CoCoOp on majority of the datasets, to obtain the best average performance of 77.1% across all datasets. This is especially significant as our approach achieves greater performance at the cost of significantly less number of trainable parameters compared to CoCoOp ($8.6\times$ lower).

Table 1: **Generalization from seen to unseen classes.** We report accuracy with CLIP ViT-B/16 model on the seen classes (S), unseen classes (U), and the harmonic mean of both of them (H). **CPT** outperforms CoOp by $+4.0\%$ while performing at par with parameter heavy CoCoOp. To compete with CoCoOp, we adopt a $2\times$ ensemble **CPT** $_{2E}$ which easily outperforms CoCoOp while still having significantly less parameters.

Method	#Params	Average			ImageNet			Caltech101			OxfordPets			StanfordCars			Flowers102		
		S	U	H	S	U	H	S	U	H	S	U	H	S	U	H	S	U	H
ZS CLIP	–	69.3	**74.2**	71.7	72.4	68.1	70.2	96.8	94.0	95.4	91.2	97.3	94.1	63.4	**74.9**	68.7	72.1	**77.8**	74.8
CoOp	2.05K	82.7	63.2	71.7	76.5	67.9	71.9	98.0	89.8	93.7	93.7	95.3	94.5	**78.1**	60.4	68.1	97.6	59.7	74.1
CoCoOp	35.38K	80.5	71.7	75.8	76.0	70.4	73.1	98.0	93.8	95.8	95.2	97.7	96.4	70.5	73.6	72.0	94.9	71.84	81.7
CPT	2.05K	82.5	69.9	75.7	76.4	69.1	72.6	98.0	**94.3**	96.1	95.4	97.8	96.6	74.5	71.6	73.0	97.5	66.6	79.2
CPT $_{2E}$	4.10K	**83.2**	71.8	**77.1**	**76.6**	**70.6**	**73.5**	**98.2**	**94.3**	**96.2**	**95.9**	**98.2**	**97.1**	75.6	72.3	**73.9**	**97.7**	72.2	**83.0**

Method	#Params	Food101			FGVCAircraft			SUN397			DTD			EuroSAT			UCF101		
		S	U	H	S	U	H	S	U	H	S	U	H	S	U	H	S	U	H
ZS CLIP	–	90.1	91.2	90.7	27.2	**36.3**	31.1	69.4	75.4	72.2	53.2	**59.9**	56.4	56.5	**64.1**	60.0	70.5	**77.5**	73.9
CoOp	2.05K	88.3	82.3	85.2	40.4	22.3	28.8	80.6	65.9	72.5	79.4	41.2	54.2	92.2	54.7	68.7	84.7	56.1	67.5
CoCoOp	35.38K	**90.7**	91.3	**91.0**	33.4	23.7	27.7	79.7	**76.9**	78.3	77.0	56.0	**64.9**	87.5	60.0	71.2	82.3	73.5	77.6
CPT	2.05K	89.9	90.9	90.4	38.7	28.4	32.8	80.9	74.4	77.5	80.3	47.7	59.8	92.2	58.1	**71.3**	83.5	70.3	76.3
CPT $_{2E}$	4.10K	90.2	**91.4**	90.8	**40.5**	31.6	**35.5**	**81.5**	76.3	**78.8**	**81.8**	51.5	63.2	**93.1**	57.7	71.2	**84.3**	73.8	**78.7**

4.3 Cross-Dataset Transfer

In this section, we show CPT's ability to transfer learned prompt beyond a single dataset. This is fundamentally more challenging compared to generalizing well while remaining within the same data distribution. In this setting, we train using the generic and natural image dataset ImageNet and test the efficacy of the learned prompt in 10 different datasets comprising of images coming from finegrained categories like Cars, Flowers, Food, Aircraft etc or texture classification like DTD. As seen from Figure 5, while using CLIP RN-50 as the backbone, **CPT** demonstrates better transferability than CoOp on all the datasets, leading to an average accuracy of 57.0%, which is $+3.5\%$ better than CoOp. Likewise for CLIP ViT-B/16, Table 2 shows that performance of **CPT** is comparatively better

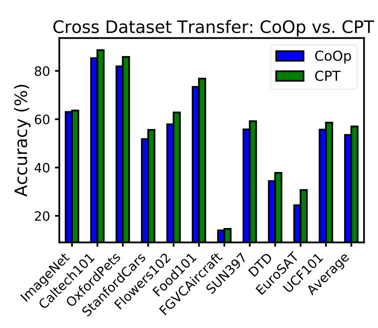


Figure 5: **Cross Dataset Transfer.**

than CoOp which uses same number of learnable parameters. We also achieve similar performance to CoCoOp while being $8.6\times$ parameter efficient with $2\times$ ensemble **`CPT`** and $17.2\times$ parameter efficient with **`CPT`**. In summary, these results show that a contrastively learned prompt not only transfers the knowledge to very different settings but also does it in much more parameter efficient manner.

Table 2: **Cross-dataset transfer.** Promtps trained on ImageNet using **`CPT`** are more generalizable to other datasets than CoOp, while competent with CoCoOp. A simple $2\times$ ensemble **`CPT`** $_{2E}$ fills the gap while being $8.6\times$ parameter efficient. All the baseline use CLIP ViT-B/16 backbone under the same experimental settings.

		Source	Target										
	# Params	IN1K	Caltech	Pets	Cars	Flowers	Food	Aircraft	SUN	DTD	EuroSAT	UCF	*Avg*
CoOp	2.05K	71.5	93.7	89.1	64.5	68.7	85.3	18.5	64.2	41.9	46.4	66.6	63.9
CoCoOp	35.38K	71.0	**94.4**	90.1	65.3	**71.9**	86.1	**22.9**	**67.4**	**45.7**	45.4	68.2	**65.7**
`CPT`	2.05K	71.3	94.1	90.2	**64.8**	70.6	85.9	21.8	66.3	43.6	46.0	68.0	65.1
`CPT` $_{2E}$	4.10K	**71.6**	94.2	**90.3**	64.5	70.9	**86.3**	21.7	66.8	45.3	**47.7**	**69.3**	**65.7**

4.4 Domain Generalization

Domain or distribution shifts are very common in the real-world. In order to study the robustness of the learned prompts to out-of-distribution data, following (Zhou et al., 2022a), we learned a prompt on the ImageNet dataset and tested its performance on 4 of its specially designed benchmarks possessing distribution shift, like ImageNetV2, ImageNet-Sketch, etc. Table 3 clearly shows the dominating performance of **`CPT`** over CoOp and CoCoOp even with a single prompt. Using an additional prompt to ensemble even pushes the performance further by 0.38% on average over the target datasets, while still using 8.6 times less parameters than CoCoOp. This highlights the effectiveness of **`CPT`** in learning domain invariant prompts while being highly parameter efficient.

Table 3: **Domain generalization**. Prompts learned on ImageNet are transferred to four of its domain-shifted variants. **`CPT`** outperforms both CoOp and CoCoOp, while using same numer of tunable parameters as CoOp.

		Source	Target			
	# Params	ImageNet	ImageNetV2	ImageNet-Sketch	ImageNet-A	ImageNet-R
CoOp	2.05K	71.5	64.2	48.0	49.7	75.2
CoCoOp	35.38K	71.0	64.1	48.8	50.6	76.2
`CPT`	2.05K	71.3	64.1	49.0	50.7	76.4
`CPT` $_{2E}$	4.10K	**71.6**	**64.6**	**49.2**	**51.1**	**76.8**

4.5 Few-Shot Classification

Beyond generalization, we also consider the standard few-shot classification setting in which we study the performance on test data belonging to the same classes and the same domain as the training data.

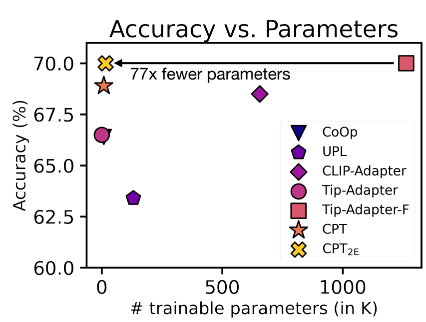


Figure 6: **Comparison with SOTA methods using CLIP RN-50.** **`CPT`** achieves very competitive performance while being highly parameter efficient. Best viewed in color.

Comparison with CoOp. Figure 7 shows the results on 7 different VLMs. While using CLIP RN-50 with single labeled images per class, CoOp outperforms handcrafted prompts (Zero shot) by 0.8% on average, while **`CPT`** provides 4.1% improvement. When compared to CoOp, the gain is particularly significant in the low-shot scenarios, which are practically important cases. E.g. for CLIP RN-50 backbone, the improvement over CoOp in 1-shot is 5.9 times that in 16-shot, which concretely affirms the advantage of self-supervised contrastive learning in overcoming overfitting and better generalization. Similarly, in the one-shot setting of DeCLIP ViT-B/32, **`CPT`** outperforms CoOp by $+1.1\%$, showing its effectiveness in few-shot classification across different models.

Comparison with SOTA Methods. Figure 6 compares different methods in terms of number of parameters vs. average accuracy over all the 11 datasets. We observe that by just using two prompts

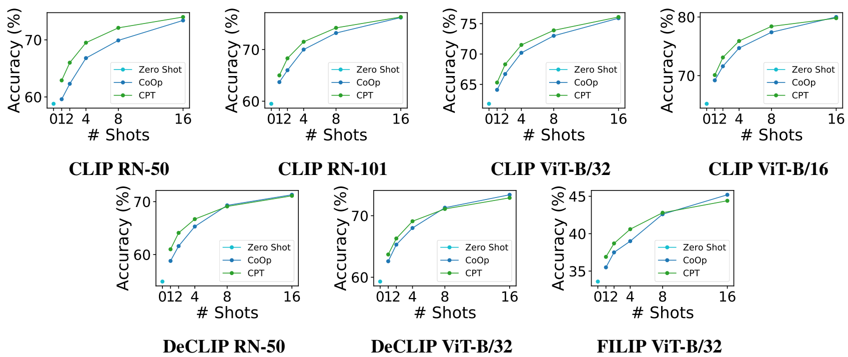


Figure 7: **Few-shot Classification.** **CPT** consistently outperforms CoOp across all seven VLMs, showing the effectiveness of contrastive prompt tuning for efficient adaptation of pretrained models. We report the average accuracy across the 11 datasets for each few-shot setting. Best viewed in color.

for ensembling, **CPT** $_{2E}$ achieves the same performance as the current SOTA method Tip-Adapter-F (Zhang et al., 2021), while using a significant **77 times fewer** trainable parameters. Despite being very simple, **CPT** establishes new SOTA performance for parameter efficient adaptation of CLIP.

4.6 Ablation Studies

Effect of Losses. To study the effectiveness of both the self-supervised contrastive losses, we obtain the few-shot classification performance by using either of the losses $\mathcal{L}_{\text{InsCon}}$ and $\mathcal{L}_{\text{RelCon}}$, independently with $\mathcal{L}_{\text{CE}}$. For ImageNet 1-shot CLIP RN-50 setting, using only $\mathcal{L}_{\text{InsCon}}$ yields an accuracy of 62.4% while using only $\mathcal{L}_{\text{RelCon}}$ produces 62.3%. The best performance of 62.9% was obtained when both losses are employed, showing the effectiveness of both instance discrimination and relational consistency in learning effective generalizable prompts.

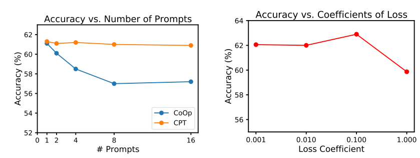


Figure 8: **Ablation Studies.** Left: variation of accuracy with the number of learnable tokens in ImageNet 1-shot using CLIP RN-50. Right: studies the effect of different values of the hyperparameter λ.

Number of Learnable Prompt Tokens. In Figure 8(left), we study the effect of number of learned prompt tokens on few-shot classification performance for the 1-shot setting in ImageNet dataset. An interesting observation is that the performance for CoOp increases with the reduction in the number of tokens (an increase of 3.9% from 16 to 1 tokens). This highlights the problem of overfitting in the low-shot setting and how CoOp is prone to it. On the other hand, **CPT** maintains a very stable performance (a drop of only 0.4% from 1 to 16 tokens) across the number of tokens, demonstrating the importance of contrastive learning in learning effective prompts in low-shot settings.

Initialization of Prompts. We initialized the learnable prompts with word embeddings of a hand-crafted prompt before training, and saw no major changes in the performance. E.g. learning 4 tokens for ImageNet 1-shot using "a photo of a" as initialization gave 61.1% accuracy compared to 61.2% using random initialization, in consistent with the findings in (Zhou et al., 2022b).

Effect of Hyperparameters. Figure 8(right) shows the effect of loss coefficient λ, where we vary λ with values $0.001, 0.01, 0.1, 1.0$ for 1-shot setting in every dataset and find $\lambda = 0.1$ to be the best value and use it for all experiments. Study of performance by varying τ_1 can be found in the Appendix A. Additional experimental results are also included in the Appendix.

5 Conclusion

In this paper, we explore contrastive prompt tuning for improved generalization in pretrained vision-language models. Specifically, we augment the standard cross-entropy loss with two additional contrastive losses that optimizes for the learned prompts to be consistent with the image space. The first loss encourages learning instance discriminative features invariant to different views, while the second one makes the logit space consistent with the image space with respect to various inter-image semantic relationships. We demonstrate the effectiveness of our approach on multiple diverse datasets, outperforming state-of-the-art methods, without any additional use of parameters.

REFERENCES

Soroush Abbasi Koohpayegani, Ajinkya Tejankar, and Hamed Pirsiavash. Compress: Self-supervised learning by compressing representations. *Advances in Neural Information Processing Systems*, 33:12980–12992, 2020.

Hassan Akbari, Liangzhe Yuan, Rui Qian, Wei-Hong Chuang, Shih-Fu Chang, Yin Cui, and Boqing Gong. Vatt: Transformers for multimodal self-supervised learning from raw video, audio and text. *Advances in Neural Information Processing Systems*, 34:24206–24221, 2021.

Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101–mining discriminative components with random forests. In *European conference on computer vision*, pp. 446–461, 2014.

Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Arechiga, and Tengyu Ma. Learning imbalanced datasets with label-distribution-aware margin loss. *Advances in neural information processing systems*, 32, 2019.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pp. 1597–1607, 2020a.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *European conference on computer vision*, pp. 104–120, 2020b.

Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3606–3613, 2014.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255, 2009.

Debidatta Dwibedi, Yusuf Aytar, Jonathan Tompson, Pierre Sermanet, and Andrew Zisserman. With a little help from my friends: Nearest-neighbor contrastive learning of visual representations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9588–9597, 2021.

Zhiyuan Fang, Jianfeng Wang, Lijuan Wang, Lei Zhang, Yezhou Yang, and Zicheng Liu. Seed: Self-supervised distillation for visual representation. *arXiv preprint arXiv:2101.04731*, 2021.

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 *2004 conference on computer vision and pattern recognition workshop*, pp. 178–178, 2004.

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.

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.

Yuting Gao, Jinfeng Liu, Zihan Xu, Jun Zhang, Ke Li, and Chunhua Shen. Pyramidclip: Hierarchical feature alignment for vision-language model pretraining. *arXiv preprint arXiv:2204.14095*, 2022.

Shashank Goel, Hritik Bansal, Sumit Bhatia, Ryan A Rossi, Vishwa Vinay, and Aditya Grover. Cyclip: Cyclic contrastive language-image pretraining. *arXiv preprint arXiv:2205.14459*, 2022.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284, 2020.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9729–9738, 2020.

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

Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8340–8349, 2021a.

Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15262–15271, 2021b.

Tony Huang, Jack Chu, and Fangyun Wei. Unsupervised prompt learning for vision-language models. *arXiv preprint arXiv:2204.03649*, 2022.

Ashraful Islam, Chun-Fu Richard Chen, Rameswar Panda, Leonid Karlinsky, Richard Radke, and Rogerio Feris. A broad study on the transferability of visual representations with contrastive learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8845–8855, 2021.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*, pp. 4904–4916, 2021.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *Advances in Neural Information Processing Systems*, 33:18661–18673, 2020.

Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *Proceedings of the IEEE international conference on computer vision workshops*, pp. 554–561, 2013.

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

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. *arXiv preprint arXiv:2201.12086*, 2022.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*, 2019.

Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *European Conference on Computer Vision*, pp. 121–137, 2020.

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.

Weiyang Liu, Yandong Wen, Zhiding Yu, and Meng Yang. Large-margin softmax loss for convolutional neural networks. *arXiv preprint arXiv:1612.02295*, 2016.

Yuning Lu, Jianzhuang Liu, Yonggang Zhang, Yajing Liu, and Xinmei Tian. Prompt distribution learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5206–5215, 2022.

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

Norman Mu, Alexander Kirillov, David Wagner, and Saining Xie. Slip: Self-supervision meets language-image pre-training. *arXiv preprint arXiv:2112.12750*, 2021.

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

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. Cats and dogs. In *2012 IEEE conference on computer vision and pattern recognition*, pp. 3498–3505, 2012.

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 *International Conference on Machine Learning*, pp. 8748–8763, 2021.

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *International Conference on Machine Learning*, pp. 5389–5400, 2019.

Manli Shu, Weili Nie, De-An Huang, Zhiding Yu, Tom Goldstein, Anima Anandkumar, and Chaowei Xiao. Test-time prompt tuning for zero-shot generalization in vision-language models. *arXiv preprint arXiv:2209.07511*, 2022.

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.

Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. Vl-bert: Pre-training of generic visual-linguistic representations. *arXiv preprint arXiv:1908.08530*, 2019.

Yu Sun, Eric Tzeng, Trevor Darrell, and Alexei A Efros. Unsupervised domain adaptation through self-supervision. *arXiv preprint arXiv:1909.11825*, 2019.

Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*, 2019.

Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P Xing. Learning robust global representations by penalizing local predictive power. *Advances in Neural Information Processing Systems*, 32, 2019.

Chen Wei, Huiyu Wang, Wei Shen, and Alan Yuille. Co2: Consistent contrast for unsupervised visual representation learning. *arXiv preprint arXiv:2010.02217*, 2020.

Bichen Wu, Ruizhe Cheng, Peizhao Zhang, Peter Vajda, and Joseph E Gonzalez. Data efficient language-supervised zero-shot recognition with optimal transport distillation. *arXiv preprint arXiv:2112.09445*, 2021.

Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE computer society conference on computer vision and pattern recognition*, pp. 3485–3492, 2010.

Lewei Yao, Runhui Huang, Lu Hou, Guansong Lu, Minzhe Niu, Hang Xu, Xiaodan Liang, Zhenguo Li, Xin Jiang, and Chunjing Xu. Filip: Fine-grained interactive language-image pre-training. *arXiv preprint arXiv:2111.07783*, 2021.

Xin Yuan, Zhe Lin, Jason Kuen, Jianming Zhang, Yilin Wang, Michael Maire, Ajinkya Kale, and Baldo Faieta. Multimodal contrastive training for visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6995–7004, 2021.

Renrui Zhang, Rongyao Fang, Peng Gao, Wei Zhang, Kunchang Li, Jifeng Dai, Yu Qiao, and Hongsheng Li. Tip-adapter: Training-free clip-adapter for better vision-language modeling. *arXiv preprint arXiv:2111.03930*, 2021.

Nanxuan Zhao, Zhirong Wu, Rynson WH Lau, and Stephen Lin. What makes instance discrimination good for transfer learning? *arXiv preprint arXiv:2006.06606*, 2020.

Mingkai Zheng, Shan You, Fei Wang, Chen Qian, Changshui Zhang, Xiaogang Wang, and Chang Xu. Relational self-supervised learning. *arXiv preprint arXiv:2203.08717*, 2022.

Zexuan Zhong, Dan Friedman, and Danqi Chen. Factual probing is [mask]: Learning vs. learning to recall. *arXiv preprint arXiv:2104.05240*, 2021.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16816–16825, 2022a.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022b.

A Additional Experiments

In this section, we provide some additional experimental results and figures considering different VLM backbones:

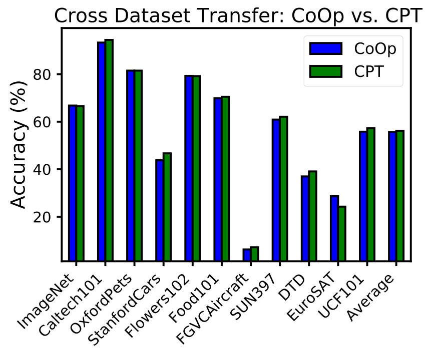
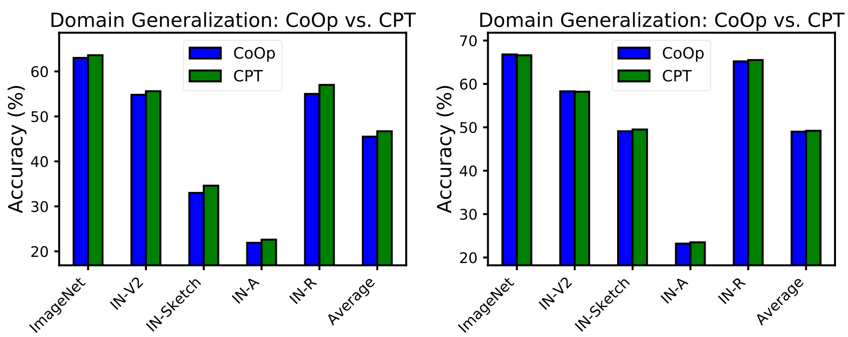


Figure 9: **Cross-Dataset performances** Figure shows bar charts comparing generalization performance of prompts trained on ImageNet towards 10 other datasets using **`CPT`** and CoOp with DeCLIP ViT-B/32 backbone.



CLIP RN-50 **DeCLIP ViT-B/32**

Figure 10: **Domain Generalization.** Figure shows bar charts comparing cross-domain generalization performance of prompts trained using **`CPT`** and CoOp on two VLM backbones. Evidently from the results, prompt learned using **`CPT`** are more robust to distribution shifts.

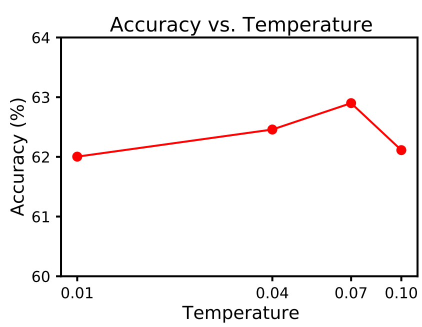


Figure 11: **Ablation on sharpening temperature τ_1.** To study the effect of varying the sharpening temperature for the relational contrastive loss, we fix the value of $\tau_{\mathbf{z}}$ to 0.04 and vary τ_1 to values 0.01, 0.04, 0.07, 0.10 and find the best performance at $\tau_1 = 0.07$.

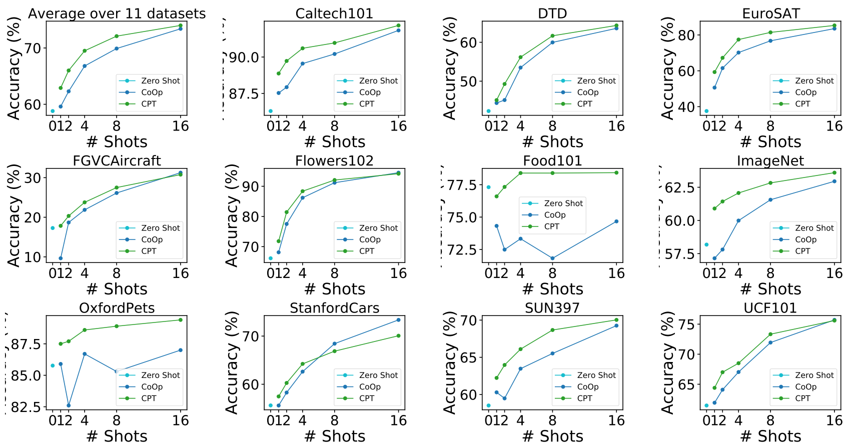


Figure 12: **Few-Shot Classification using CLIP RN-50** Figure shows performance few-shot classification performance for each of the 11 datasets using the CLIP RN-50 VLM backbone.

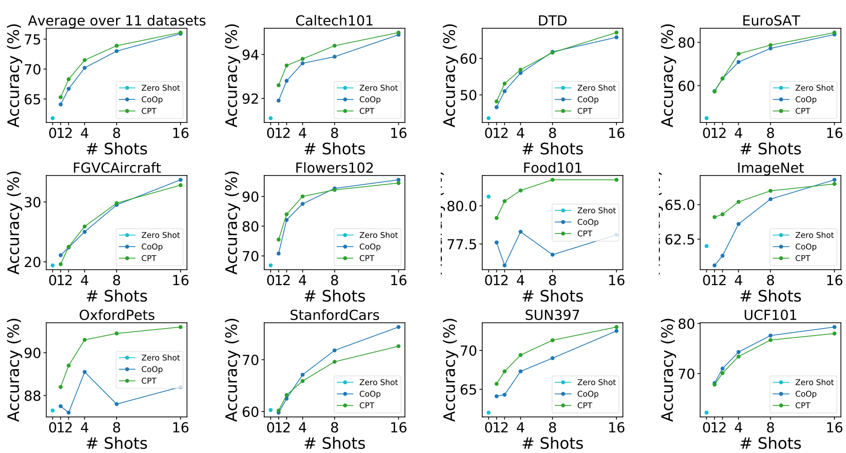


Figure 13: **Few-Shot Classification using CLIP ViT-B/32** Figure shows performance few-shot classification performance for each of the 11 datasets using the CLIP ViT-B/32 VLM backbone.