CoPL: Contextual Prompt Learning for Vision-Language Understanding

Koustava Goswami, Srikrishna Karanam, Prateksha Udhayanan,
K J Joseph and Balaji Vasan Srinivasan

Adobe Research, Bangalore India
koustavag@adobe.com

Abstract

Recent advances in multimodal learning has resulted in powerful vision-language models, whose representations are generalizable across a variety of downstream tasks. Recently, their generalization ability has been further extended by incorporating trainable prompts, borrowed from the natural language processing literature. While such prompt learning techniques have shown impressive results, we identify that these prompts are trained based on global image features which limits itself in two aspects: First, by using global features, these prompts could be focusing less on the discriminative foreground image, resulting in poor generalization to various out-of-distribution test cases. Second, existing work weights all prompts equally whereas intuitively, prompts should be reweighed according to the semantics of the image. We address these as part of our proposed Contextual Prompt Learning (CoPL) framework, capable of aligning the prompts to the localized features of the image. Our key innovations over earlier works include using local image features as part of the prompt learning process, and more crucially, learning to weight these prompts based on local features that are appropriate for the task at hand. This gives us dynamic prompts that are both aligned to local image features as well as aware of local contextual relationships. Our extensive set of experiments on a variety of standard and few-shot datasets show that our method produces substantially improved performance when compared to the current state of the art methods. We also demonstrate both few-shot and out-of-distribution performance to establish the utility of learning dynamic prompts that are aligned to local image features.

Introduction

Fully supervised computer vision models for problems like classification are typically trained on datasets like ImageNet (Deng et al. 2009a), OpenImages (Kuznetsova et al. 2018), JFT300M (Kolesnikov et al. 2020; Sun et al. 2017) etc., and have also proven themselves to be effective for a variety of downstream tasks via transfer learning (Plested and Gedeon 2022; Guo et al. 2019; Wan et al. 2019). Despite this, it is challenging to adapt these models to other domains due to various reasons including limited data and annotation overhead. Additionally, since these models are trained for specific objectives like classification, they tend to capture concepts related to categories seen during training and not to scale to unseen classes during inference.

To enhance the adaptability of such models, there has been recent efforts in tuning the associated prompts (instead of the model weights). Inspired by traditional prompt engineering efforts (Petroni et al. 2019; Brown et al. 2020; Schick and Schütze 2021), there has been work in tuning discrete prompts from predefined prompt templates (Radford et al. 2021) that help in capturing rich semantics from user intents and align them to visual contents. However, since building a rich semantic based prompt templates require domain specific and linguistic knowledge, this apporach is not scalable. The CoOp (Zhou et al. 2022b) algorithm used ideas from soft-prompting in natural language processing (Gao, Fisch, and Chen 2021; Jiang et al. 2020; Lester, Al-Rfou, and Constant 2021; Li and Liang 2021; Liu et al. 2021, 2022) to train dynamic learnable prompt vectors with backpropagation and preserve the semantic relationship between sentences and labels (Liu et al. 2021). However, the context learned with CoOp in this fashion fails to generalize to unseen classes, leading to the need for dynamically updating prompts based on the image context, an idea that was proposed in CoCoOp (Zhou et al. 2022a). This model was trained by explicitly conditioning the prompts on image feature vectors as tokens where a separate lightweight neural network (called meta-net in their work) was used to equally weight all the prompt vectors. Later Yao, Zhang, and Xu (2023) note that CoOp-based methods suffer from catastrophic knowledge forgetting, where these methods gradually miss out on essential general textual knowledge. This leads them to propose Knowledge-guided Context Optimization (KgCoOp) where they try to reduce down the distance between hand-crafted prompts and learnable prompts during training.

In our work, we identify some key issues with the aforementioned architectures. First, the features obtained from meta-net in CoCoOp (Zhou et al. 2022a) are global in nature and hence susceptible to issues like clutter and noise in many few-shot and out-of-distribution test cases (we demonstrate this empirically later on). Next, these features are directly added to all the learned prompt vectors, thus resulting in an equal weighting for each prompt. Consequently, this model is unable to learn which of the prompt vectors are more semantically relevant and contextually meaningful

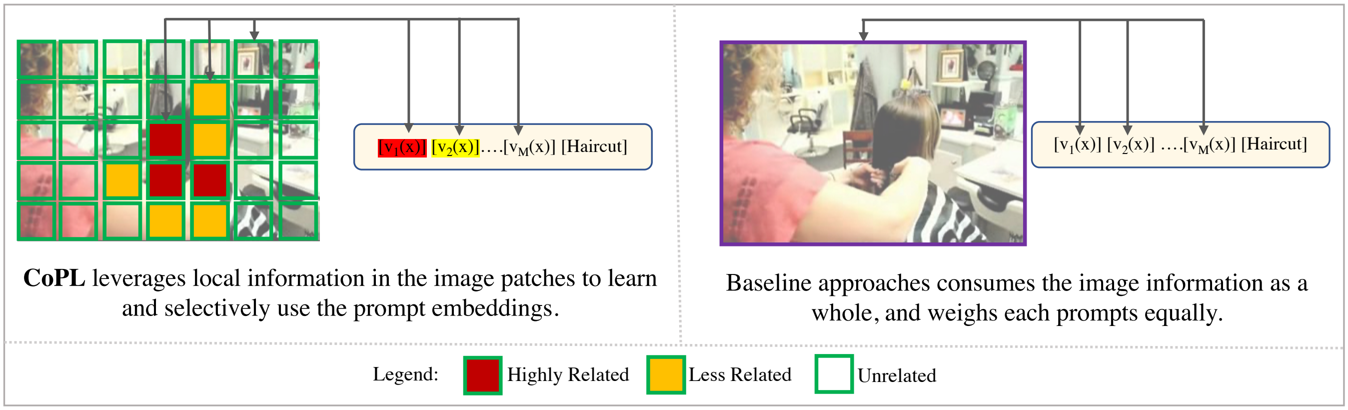


Figure 1: The figure summarises our methodological contribution. CoPL leverages the local information in the image patches to learn contextualized prompts (left). Further, it weights each of the prompts differently, based on its semantic affinity. The coloured patches highlights how the prompts are weighted according to their semantic alignment with the local image features. Baselines approaches (on right) uses global image features for prompt learning, and weights each of the prompts equally.

during inference, which makes model less generalizable on unseen classes in a zero-shot settings. While Yao, Zhang, and Xu (2023) in KgCoOp architecture prevents the catastrophic knowledge forgetting by minimizing the distance between hand-crafted prompts and learnable prompts, there is no clear evidence of what are the efficient discrete prompts. Moreover, it is hard to design generalized hand-crafted discrete prompts and sometimes tend to overfit linguistic features (Liu et al. 2021).

To address the aforementioned problems, we propose a new technique called *Contextual Prompt Learning (CoPL)*. Our key ideas include aligning prompts to local image context, realized with local features, and determining which prompts are more sematically relevant conditioned on such local context. Our insight is that by doing so, we are able to learn a more appropriate weighting of the prompts, unlike equal weighting of prior work above, that is semantically reflective of the actual content of the image under consideration. Inspired by the concept of *global attention* from the NLP literature (Luong, Pham, and Manning 2015), during training, we propose to align each local feature vector (e.g., computed from a local image patch) to a set of dynamic *soft-prompts* using a learned context vector that attends to these prompt vectors. An overview of the methdology is shown in Figure 1. This produces a set of attention weights for the prompt vectors that are semantically aligned to local image regions. This results in CoPL learning being more generalizable features as we demonstrate with an extensive set of zero-shot and few-shot classification evaluation settings.

We conduct a comprehensive set of experiments on visual classification on 11 different datasets and scenarios (zero-shot, one-shot, seen/unseen, and within-dataset and cross-dataset). Across all these experiments, we demonstrate substantial performance improvement when compared to all the baselines, indicating the ability of our method to be adapted across various classification settings with little or no training and much reduced prompt engineering.

Our key contributions are: **(i)** We identify two key shortcomings with existing prompt-based image classification methods: equal weighting of the prompt vectors and no flow of contextual local information of input images to the prompt vectors while learning during back-propagation; **(ii)** We propose *CoPL: Contextualized Prompt Learning*, a new method that addresses the issues above by learning prompt weights dynamically and aligning the resulting prompt vectors with local image features; **(iii)** We conduct extensive experiments with CoPL under a variety of classification scenarios and demonstrate substantial performance improvements, particularly in various unseen and few-shot data scenarios. Importantly, on 11 different image recognition dataset, on average CoPL achieves state-of-the-art performance on unseen class classifications beating state-of-the-art **zero-shot large-scale model CLIP by** 1.4%, **conditional prompting model CoCoOp by** 3.9%, **knowledge pruning model KgCoOp by** 2.0% and **gradient learning method ProGrad by** 4.9% on accuracy. We evaluated CoPL on cross-dataset zero-shot image recognition tasks and on an average on 8 datasets, CoPL outperformed CoCoOp by 2.3% on accuracy. Our extensive experiments on domain generalization in Table 3, established CoPL as state-of-the-art most generalized model by outperforming MaPLe (Khattak et al. 2023) and CoPrompt (Roy and Etemad 2023).

Related Work

Multimodal Models: Vision-language models have shown great potential in learning generic visual representations. The core idea has been to use natural language supervision for image representation learning and align them jointly in the same embedding space (Jia et al. 2021; Radford et al. 2021). Early explorations in this line of research include related problem formulations - metric learning (Frome et al. 2013a), multilabel classification (Gomez et al. 2017; Joulin et al. 2016), n-gram language learning (Li et al. 2017), captioning (Desai and Johnson 2021). Traditionally, hand-crafted descriptors (Elhoseiny, Saleh, and Elgammal 2013; Socher et al. 2013) were the mode of capturing image representations. Later, convolutional neural net based architectures (Frome et al. 2013b; Ba et al. 2015) were introduced. Recent works have focused on learning joint representations of both the modalities using deep learning architectures (Fürst et al. 2021; Jia et al. 2021; Li et al. 2022b; Kamath et al. 2021). With the introduction of transformers (Vaswani et al. 2017), Li et al., (Li et al. 2019) proposed Visual-BERT, where texts and images are jointly encoded in a single transformers architecture. One of the biggest milestone in the multimodal research is the CLIP model (Radford et al. 2021). It is a dual encoder based model and during training it matches pairs of images and texts. One of the main component of CLIP is the carefully designed prompts which most of the times very hard to formulate. To overcome this, Zhou et al., (Zhou et al. 2022b) designed CoOp, which trains dynamic soft-prompts during back-propagation.

Prompting: The intuition behind prompt learning is to capture user intention and instructions to perform certain downstream tasks (Li et al. 2022c; Parmar et al. 2022; Zhu et al. 2022). With the introduction of GPT (Brown et al. 2020), prompt engineering is shown to be performing efficiently in few-shot knowledge adaptation. But, building prompt templates is hard and requires immense skills. Recently, researchers have proposed "soft-prompting". The main intuition is to learn dynamic continuous prompt tokens during back-propagation (Gao, Fisch, and Chen 2021; Jiang et al. 2020; Lester, Al-Rfou, and Constant 2021; Li and Liang 2021; Liu et al. 2021, 2022). Recently, Goswami et al. (Goswami et al. 2023) highlights that the soft prompts can be further tuned with the semantic knowledge of language models without explicitly verbalizers setup. At the same Sarkar et al. (2023) showed tuning these soft prompts can help downstream tasks. CoOp (Zhou et al. 2022b) is designed to train these soft prompts during training. Later Zhou et al. (2022a) introduced CoCoOp on top of CoOp to improve the performance by conditioning on image input. On the other hand, KgCoOp (Yao, Zhang, and Xu 2023) and ProGrad (Zhu et al. 2023) were proposed to improve the performance of CoOp by aligning prompts towards general knowledge. While ProGrad tries to optimize the prompts with the aligned direction, KgCoOp minimises the distance between hand-crafted prompts and learnable prompts during training. As our methodology is directly improving the CoCoOp by infusing the local context of the image during prompt training, we take CoCoOp as our direct baseline. Along with CoOp and CoCoOp, KgCoOp and ProGrad methodologies are directly in line with our way of learning representations, we considered these methods as our standard baselines. Recently, MaPLe (Khattak et al. 2023) and CoPrompt (Roy and Etemad 2023) introduces a new way of learning by infusing prompts to both text and image encoder. Such infusion in a multi-modal setup may affect generalization,and hence in our generalization experiment in Table 3, we compare against these approaches too.

Here, we introduce and discuss details of our proposed method *CoPL: Contextualized Prompt Learning*. Since CoPL uses the same architectural backbone as CLIP (Radford et al. 2021), we first begin with a brief review followed by a discussion on how our closest baseline, CoCoOp (Zhou et al. 2022a), is trained. While we use the CLIP backbone for simplicity, we are in no way limited by this design choice. In fact, our method is very much applicable to be used in conjunction with a variety of other architectures, e.g., VisualBERT (Li et al. 2019), MDETR (Kamath et al. 2021), GLIP (Li et al. 2022a) etc.

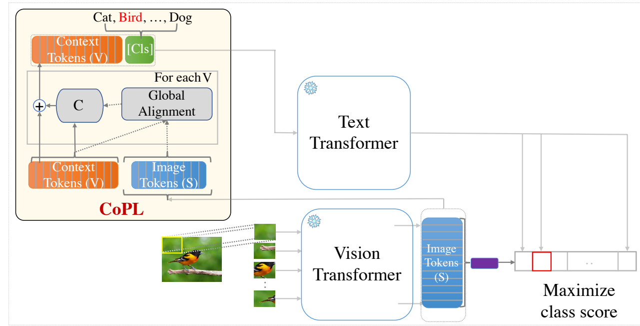


Figure 2: CoPL architecture summary.

Review of CoOp/CoCoOp

CoCoOp: Conditional Context Optimization is built on top of a previously introduced *Context Optimization (CoOp)* (Zhou et al. 2022b) algorithm. It is observed that the choice of prompts plays a major role in vision-language understanding (Radford et al. 2021), though it is quite hard to find the best match between prompts and image descriptions. To overcome such a need for prompt engineering, the CoOp model (Zhou et al. 2022b) trains a set of continuous vectors $\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_M\}$ as context tokens during back-propagation. The CoCoOp (Zhou et al. 2022a) algorithm improves CoOp's (Zhou et al. 2022b) performance by learning to generate prompts conditioned on each image instance, i.e., these now change for each image unlike in CoOp, where they are fixed. CoCoOp does this by training an additional two-layer neural network called meta-net that takes an image feature vector as input and produces a conditional vector that is combined with the prompt vectors to generate the final image-dependent prompts. This is done as follows:

$$\mathbf{v}_m(\mathbf{x}) = \mathbf{v}_m + h_\theta(\mathbf{x}) \tag{1}$$

where h_θ refers to the meta-net and $\mathbf{x}$ the image feature vector.

Methodology

CoPL: Contextual Prompt Learning

While CoCoOp generalizes better than CoOp for unseen-class classification, there is significant scope for improvement. First, since CoCoOp uses global feature vectors for learning the updated prompts, it focuses less on the discriminative regions in images (which tend to be more local). Next, the addition operation performed in Equation 1 does not capture the individual importance of each prompt token. This is particularly important since certain discriminative regions in images may weight certain prompts more when compared to others and this is not captured in the CoCoOp model. To address these issues, we propose CoPL, a simple and intuitive algorithm that operates at the local feature granularity while aligning them with prompts to learn prompt importance weights.

Given an image $\mathbf{I}$, we first compute a set of local feature vectors. For instance, this can be the output of a vision transformer model that generates patch embeddings. Let $\mathbf{s} \in \mathbb{R}^{P \times B \times d}$ (where P is number of patches from image, B is training batch size, and d is feature dimesionality) be this. Conditioned on these local features, we determine the semantically most meaningful prompts. To do this, we take inspiration from the attention work of Luong et al. (Luong, Pham, and Manning 2015) and generate context representations that explicitly consider both the learnable prompt tokens as well as the patch representations. We first learn a lightweight neural network to generate a conditional token for the representation of each patch in $\mathbf{s}$ as:

$$\mathbf{s}_p = h_\theta(\mathbf{s}_p) \tag{2}$$

where $p \in \{1, 2, \ldots, P\}$. This makes the architecture parameter-efficient and easily differentiable during back-propagation. To generate the context representations that can be used to update the prompt tokens, we first learn a variable length alignment vector $\mathbf{a}_p$ ($\mathbf{a} \in \mathbb{R}^{B \times M \times d}$), one for every patch p that attends to each prompt tokens $\mathbf{v}_i$ and compares them to the corresponding image patch representation $\mathbf{s}_p$ as:

$$\begin{aligned}\mathbf{a}_p &= \text{align}(\mathbf{s}_p, \mathbf{v}_i) \\ &= \frac{\exp(\text{score}(\mathbf{s}_p, \mathbf{v}_i))}{\sum_{i=1}^{M}(\exp(\text{score}(\mathbf{s}_p, \mathbf{v}_i)))}\end{aligned} \tag{3}$$

where $i \in 1, 2, \ldots, M$ and M is the number of prompt tokens. In our setup, `score` refers to the *content function* and is implemented as:

$$\text{score}(\mathbf{s}_p, \mathbf{v}_i) = \tanh(\mathbf{W}_a[\mathbf{s}_p; \mathbf{v}_i]) \tag{4}$$

where W_a is the weight vector.

Finally, the per-patch context representation $\mathbf{c}_p$ is calculated as the weighted sum over all prompt tokens as:

$$\mathbf{c}_p = \sum_{i=1}^{M} \mathbf{a}_{pi}\mathbf{v}_i \tag{5}$$

The final prompt tokens are now obtained by conditioning on the context vectors above as:

$$\mathbf{v}_m(\mathbf{x}) = \mathbf{v}_m + \sum_{i=1}^{P} \mathbf{c}_i \tag{6}$$

In a nutshell, CoPL calculates the prompts for the i-th class as $t_i = [v_1(x), v_2(x), ..., v_M(x), cl_i]$, where cl_i is the embedding of the i^{th} class label. The prediction probability is calculated as:

$$p(y|x) = \frac{exp(sim(x, g(t_y(x))/\gamma}{\sum_{i=1}^{k} exp(sim(x, g(t_i(x))/\gamma} \tag{7}$$

where $g()$ is the feature vector produced by the text encoder and γ is a temperature parameter. During our entire pipeline, the pre-trained CLIP model is fixed.

Experimental Evaluation

Datasets We follow Zhou et al. (Zhou et al. 2022b) to evaluate our model on 11 image classification dataset of varying complexity. The datasets include: generic classification datasets like ImageNet (Deng et al. 2009b) and Caltech-101 (Fei-Fei, Fergus, and Perona 2004); curated fine-grained datasets like OxfordPets (Parkhi et al. 2012), StanfordCars (Krause et al. 2013), Flowers102 (Nilsback and Zisserman 2008), Food101 (Bossard, Guillaumin, and Gool 2014) and FGVCAircraft (Maji et al. 2013); scene, action, texture and satellite image recognition datasets from SUN397 (Xiao et al. 2010), UCF101 (Soomro, Zamir, and Shah 2012), DTD (Cimpoi et al. 2014) and EuroSat (Helber et al. 2018) respectively. For the few-shot experiments, we follow Zhou et al. (Zhou et al. 2022a) to randomly sample datapoints for training and evaluate on the entire test set.

Training Details In CLIP architecture, we use ViT-B/16 as image encoder. Our prompt token length is 4. All our models are trained with a batch size 1 for 10 epochs on a single 16 GB Tesla T4 GPU system. Our starting learning rate is 0.002 and used cosine learning rate scheduler. Our warm-up with a constant learning rate is 0.00001.

Baseline Models We compare our approach with CoCoOp (Zhou et al. 2022a), CoOp (Zhou et al. 2022b), KgCoOp (Yao, Zhang, and Xu 2023), ProGrad (Zhu et al. 2023) and also large-scale zero-shot methodology CLIP (Radford et al. 2021). While comparing with CLIP, we indeed compare our learned prompt embeddings with manually designed prompts. We closely follow most of our experimental setting with that of Zhou et al., (Zhou et al. 2022a).

Knowledge Transfer to Unseen Classes

One of the main downsides of CoOp is that it is unable to generalize to the unseen classes while being trained on the base classes. The other methods improve over CoOp, but they miss out on capturing the local features of the image and do not give weightage to the semantically aligned learnable prompts, which is the key focus area of our work. Following the implementation of Zhou et al., (Zhou et al. 2022a), we conducted our experiments on the above mentioned 11 datasets both for seen and unseen classes. While training is only conducted on the base classes, during testing we transfer the learnt knowledge to classify unseen classes as well as seen classes.

From Table 1, we observe that there is a decrease in performance for CoOp methodology over the zero-shot large-scale method CLIP for unseen classes. Though introduction of conditional prompting improve the performance of CoCoOp, in most of the cases it fails to generalize to unseen classes with the dynamically learnt prompts (in 7 out of 11 datasets CLIP outperform CoCoOp methodology). In fact KgCoOp, though it is not conditioned on the image features during prompt learning, outperforms CoCoOp in many cases.

Interestingly, our proposed approach CoPL, improves the accuracy of unseen classes over CoCoOp and CoOp for 10 different tasks. It also outperforms KgCoOp and Pro-

Methodology	Protocols	C101	OP	SC	F102	F101	FGVCA	SUN397	DTD	IN	ESAT	UFC101	Average
CLIP	Seen	96.8	91.1	62.3	72.0	90.1	27.1	69.3	53.2	72.4	56.4	70.5	69.3
	Unseen	94.0	97.2	74.8	**77.8**	91.2	**36.2**	75.3	**59.9**	68.1	64.0	**77.5**	74.2
	HM	95.4	94.1	68.6	74.8	90.6	31.0	72.2	56.3	70.2	60.0	73.8	71.7
CoOp	Seen	98.0	93.6	78.1	**97.6**	88.3	40.4	80.6	**79.4**	76.4	**92.1**	**84.6**	**82.7**
	Unseen	89.8	95.2	60.4	59.6	82.2	22.3	65.8	41.1	67.8	54.7	56.0	63.2
	HM	93.7	94.4	68.1	74.0	85.1	28.7	72.5	54.2	71.9	68.9	67.4	71.6
CoCoOp	Seen	97.9	95.2	70.4	94.8	90.7	33.4	79.7	77.0	75.9	87.4	82.3	80.5
	Unseen	93.8	97.6	73.4	71.7	91.2	23.7	76.8	56.0	70.4	60.0	73.4	71.7
	HM	95.8	96.4	72.0	81.7	90.9	27.7	78.2	**64.8**	73.1	71.2	77.6	75.8
KgCoOp	Seen	97.7	94.6	71.7	95.0	90.5	36.2	80.2	77.5	75.8	85.6	82.8	80.7
	Unseen	94.3	97.7	**75.0**	74.7	91.7	33.5	76.5	54.9	69.9	**64.3**	76.6	73.6
	HM	96.0	96.1	**73.3**	83.6	91.0	**34.8**	78.3	64.3	72.7	**73.4**	79.6	77.0
ProGrad	Seen	98.0	95.0	**77.6**	95.5	90.3	**40.5**	**81.2**	77.3	77.0	90.1	84.3	82.4
	Unseen	93.8	97.6	68.6	71.8	89.5	27.5	74.1	52.3	66.6	60.8	74.9	70.7
	HM	95.9	96.3	72.8	82.0	89.9	32.8	77.5	62.4	71.4	72.6	79.3	76.1
CoPL	Seen	**98.1**	**95.6**	70.7	96.1	**90.9**	36.1	80.2	78.0	**77.8**	89.2	83.1	81.4
	Unseen	**94.9**	**97.8**	74.4	72.1	**91.4**	31.3	**77.3**	50.0	**71.3**	34.2	76.6	**75.6**
	HM	**96.5**	**96.7**	72.6	**84.1**	**91.1**	33.7	**78.7**	64.0	**74.5**	61.7	**79.8**	**78.5**

Table 1: Performance of CoPL and the baselines on 11 classification datasets; each training dataset consists 16-shots per class. These results highlights that CoPL has better generalization than other state-of-the-art methods. The results for the base methods are borrowed from KgCoOp (Yao, Zhang, and Xu 2023). HM refers to Harmonic Mean between the seen and unseen classes. Here “C101” is Caltech101; “OP” is OxfordPets; “SC” is StandfordCars; “F102” is Flowers102; “F101” is Food101; “FGVCA” is FGVCAircraft; “IN” is ImageNet; “ESAT” is EuroSAT.

Methodology	Oxford Pets	Standford Cars	Food101	DTD	EuroSat	Flowers102	FGVC Aircraft	UCF101	SUN397	Avg
CoCoOp	89.6	61.4	88.2	47.0	**63.5**	66.6	20.9	65.5	70.3	63.7
CoPL	**89.8**	**61.9**	**89.5**	**49.5**	58.1	**71.3**	**23.8**	**70.3**	**71.9**	**65.1**

Table 2: Table shows the zero-shot image recognition results of CoPL across multiple dataset. We trained models on Caltech-101 dataset and tested on the other datasets. These results suggests very strong generalization ability of CoPL.

Models	Src.	Tgt.				
	IN	INV2	IN-Sk	IN-A	IN-R	Avg.
CLIP	66.7	60.8	46.1	47.7	73.9	57.1
CoCoOp	71.0	64.0	48.7	50.6	76.1	59.9
CoOp	71.5	64.2	47.9	49.7	75.2	59.2
ProGrad	72.2	64.7	47.6	49.3	74.5	59.0
KgCoOp	71.2	64.1	48.9	50.6	76.7	60.1
MaPLe	70.7	64.0	49.1	**50.9**	76.9	60.2
CoPrompt	70.8	64.2	49.4	50.5	77.5	60.4
CoPL	**71.7**	**65.2**	**49.4**	50.8	**78.6**	**61.0**

Table 3: CoPL v. baselines in a domain generalization setting (“IN” refers to ImageNet).

Grad during unseen class classification in 6 tasks. Moreover, the harmonic mean for all 7 tasks are greater than the current state-of-the-art model, which specifies that the learned prompts are more generalizable across domains and tasks. It is important to note that CoPL comfortably outperforms manual prompt based methodology CLIP on 5 different tasks. With the relatively complex dataset like EuroSAT, where localization is hard to be aligned due to the nature of satellite imagery, for the seen classes CoPL outperformed CoCoOp by 1.8%, KgCoOp by 3.6% and CLIP by 32.8% on accuracy.

In Figure 3, we further analyse the absolute performance gain on the unseen classes over CoOp and CoCoOp methodologies. When compared with CoOp on UFC101 dataset, CoPL improve the accuracy by 20.6%, showing the capability of the dynamic prompt vectors to learn semantic relatedness conditioned on the local image features. On FGVC-Aircraft dataset, the absolute performance gain of 9.0% and 7.6% for CoOp and CoCoOp respectively. Performance boost of 8.9% on the DTD dataset, confirms the capability of COPL in identifying different textures, opening up the possibility to deploy the model for design understanding. CoPL outperforms CLIP by 2.7% on the FGVC-Aircraft dataset, bringing out its ability to transfer the learnt knowledge to the relatively complex unseen classes.

In general, for both seen and unseen dataset, CoPL outperforms **CoCoOp by** 2.7%, **KgCoOp by** 1.5% and **ProGrad by** 2.4% in accuracy. Also, it achieves a gain of 6.8% in

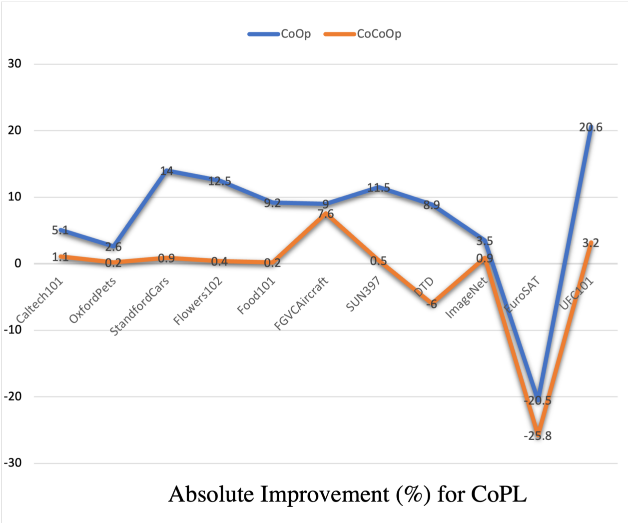


Figure 3: Performance comparison of CoPL with CoOp and CoCoOp for the unseen class detection while trained on base classes within the dataset.

accuracy over manual prompt base methodology CLIP, suggesting that conditional alignment of prompts to local features of the image can generalize the model better to diverse recognition tasks. Ideally, the manually designed prompts should have been well aligned with the images, thus CLIP can be considered as the best performing model. It is interesting to note that CoPL, on an average across all 11 datasets, achieves state-of-the-art performance by outperforming CLIP by 1.4% in accuracy. This justifies our assertion of making dynamic prompt vectors more contextualized like human annotated prompts by observing the different local features of the image.

Inter Dataset Zero-shot Performance

Here, we test the mettle of our approach on an even more challenging setting. We learn the model on a dataset, and evaluate the model on a different dataset to see how transferable the learned representations are. Concretely, we train the model on Caltech101 dataset and evaluate its performance on the rest of the datasets. Observe, in Table 2, CoPL outperforms CoCoOp across most of the datasets. For complex dataset like FGVCAircraft, CoPL gains 2.9% Caltech101 being a general purpose object classification dataset, we observe impressive performance on OxfordPets and Food101 dataset. Interestingly, though trained on object classification dataset, CoPL is able to transfer the knowledge to do texture recognition task on DTD dataset, outperforming CoCoOp by 2.5% on accuracy. In the most challenging EuroSAT dataset, consists of images taken from satellite, CoPL achieves 58.1% accuracy, comparable to the performance of the CoCoOp. These highlights that CoPL able to learn the semantics of the images by identifying the localized features and align it with the learnable textual prompts.

Domain Generalization

Domain generalization matrix establishes the claim of a model being more adaptable to target dataset from same class but having different data distribution compared to sourced domain. Here with the existing baseline we have also compared our method with MaPLe (Khattak et al. 2023) and CoPrompt (Roy and Etemad 2023). We trained the model on Imagenet and tested the model on ImageNetV2, ImageNet-Sketch, ImageNet-A, and ImageNet-R. Observe, in Table 3, CoPL on an average outperforms all the baseline methodologies. This indicates that, by learning to focus on local image features, CoPL able to figure out the subjective image-prompt mapping, which makes CoPL more robust while working on out-of-distribution dataset.

One-shot Training

We evaluate our approach in an extremely low data regime. Here, we train CoPL and CoCoOp with 1 training instance per class for each of the image recognition task and test the accuracy on both seen and unseen classes within the dataset. In our experiments, we valuate across 7 datasets, and discuss their performance on the seen and unseen classes next.

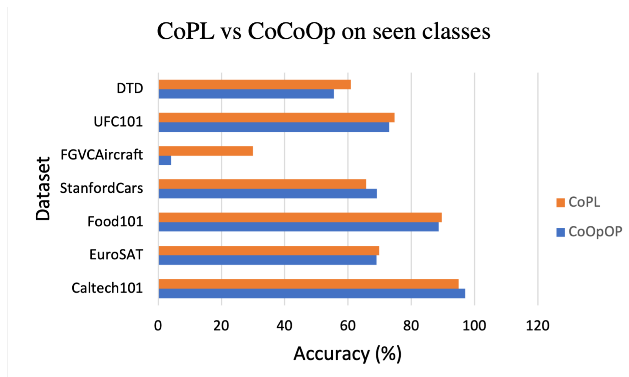


Figure 4: CoPL v. CoCoOp on 1-shot setting.

Seen Class In Figure 4, we observe that for most of the datasets, CoPL outperforms CoCoOp indicating the capacity of adapting to the in-training datasets. It is interesting to see that, for FGVCAircraft dataset, CoPL outperforms CoCoOp by large margin (25.9% on accuracy).

Unseen Class The performance of the CoPL on unseen classes while only trained on 1-shot seen classes is showcased in Figure 5. CoPL outperforms CoCoOp on 6 datasets. It is interesting to observe that, on FGVCAircraft dataset CoPL improves the accuracy from 4.9% to 28.5%.

These results indicates that, the alignment of local image features to prompts helps to capture the semantic meaning of the images while only trained with one training instance per class. This makes the model easily adaptable to diverse set of image recognition task, even for low-resource scenarios.

Discussions and Analysis

Local vs Global Image Features A key contribution of CoPL is to utilize the localized image features for prompt

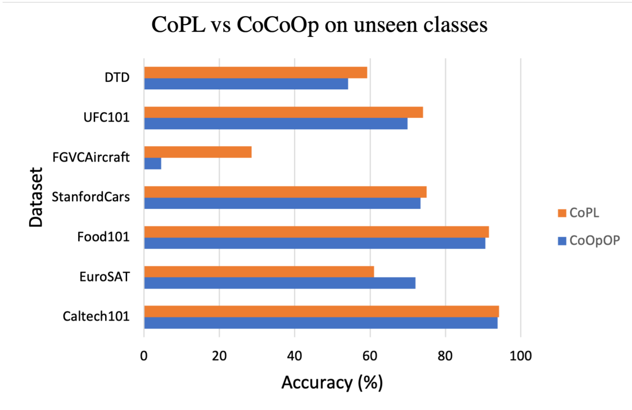


Figure 5: CoPL v. CoCoOp on unseen classes.

Methodology	Class Type	Caltech101	DTD
GA + GF	Seen	97.2	77.4
	Unseen	94.1	39.9
GA + LF	Seen	**98.1**	**78.0**
	Unseen	**94.9**	**50.0**

Table 4: Ablation Study: importance of aligning the local contextual image features with the prompt tokens. Here GA is "Global Attention"; GF is "Global Features"; LF is "Local Features".

learning. Further, we give weightage to the prompts based on the semantic relatedness. In this section, we critically analyse the contribution of using local features an opposed to global features. To evaluate this, we have conduct experiments on Caltech101 and DTD datasets. We selected these two dataset as they are primarily targeted for two different recognition tasks: object recognition and texture recognition. Similar to our earlier evaluation protocol, we evaluate the performance on the seen classes and unseen classes separately. From Table 4, we can can clearly understand that for both seen and unseen classes, aligning local contextual features with prompts, helps CoPL to generalize better.

Incremental Test Following the experimental analysis presented in CoCoOp (Zhou et al. 2022a), we evaluate the model efficacy where test dataset consists of both seen and unseen classes. In this case, during training the model weights are updated based on the seen classes but has not considered the unseen classes. Thus, during testing model will be performing zero-shot classification on unseen classes. We observe in Table 5, that CoPL comfortably outperforms all the baseline models, suggesting that local image feature alignment with prompts help to generalize better.

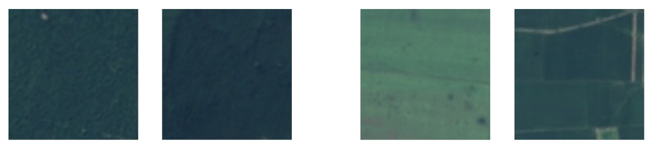
Run-time Analysis We experimentally analyse and quantify the extra wall-clock time that is required for CoPL when compared to the closest baseline CoCoOp. We use Caltech101 for this experiment. For training 800 data-points

Method	CLIP	CoOp	CoCoOp	CoPL
Accuracy	65.2	65.6	69.1	**74.7**

Table 5: Average accuracy on 11 datasets, when the test set contains both the seen and unseen classes.

(corresponding to 50 classes) , CoPL takes 28 minutes whereas evaluation on 1549 data-points took 2 minutes 05 seconds. On the other hand CoCoOp takes 21 minutes 35 seconds for training and 1 minute 34 second for inference. From this, we understand that CoPL improves generalization without sacrificing too much on the computational overhead.

Limitation Our exhaustive experimental analysis across 11 datasets greatly helped us to test the mettle of our method. Our performance on the EuroSat dataset in the unseen class category, is lower than CoCoOp. While analysing the reasons for the drop in performance, we could uncover that images from EuroSat does not contain images with salient objects. We visualise some such examples in Fig. 6; it can be seen that there are no "local" regions in these images, which can contribute towards modelling the prompt better. This results in lower performance of CoPL on such datasets.



Forest Class Annual Crop Class

Figure 6: Visualization of samples from EuroSat dataset. We note that majority of the sampled does not contain any salient objects, which makes local feature less effective.

Conclusion

In this paper, we present *CoPL: Contextual Prompt Learning*, which can align prompts to the corresponding contextual local image features. During alignment, we also produces a set of attention weights for the prompt vectors, that are semantically related to local image regions. Extensive experimental evaluation on 11 image recognition datasets showcases the efficacy of CoPL in understanding the semantic relationship between the images and the prompts. Moreover, the state-of-the-art zero-shot and few-shot results justify our claim of making CoPL better in generalization by aligning the local features of the images to prompts. In future, we plan to make CoPL capable of understanding user intents to make local edits on images.

References

Ba, L. J.; Swersky, K.; Fidler, S.; and Salakhutdinov, R. 2015. Predicting Deep Zero-Shot Convolutional Neural Networks Using Textual Descriptions. In *ICCV*.

Bossard, L.; Guillaumin, M.; and Gool, L. V. 2014. Food-101 - Mining Discriminative Components with Random Forests. In *ECCV*.

Brown, T. B.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; Agarwal, S.; Herbert-Voss, A.; Krueger, G.; Henighan, T.; Child, R.; Ramesh, A.; Ziegler, D. M.; Wu, J.; Winter, C.; Hesse, C.; Chen, M.; Sigler, E.; Litwin, M.; Gray, S.; Chess, B.; Clark, J.; Berner, C.; McCandlish, S.; Radford, A.; Sutskever, I.; and Amodei, D. 2020. Language Models are Few-Shot Learners. In *NeurIPS*.

Cimpoi, M.; Maji, S.; Kokkinos, I.; Mohamed, S.; and Vedaldi, A. 2014. Describing Textures in the Wild. In *CVPR*.

Deng, J.; Dong, W.; Socher, R.; Li, L.; Li, K.; and Fei-Fei, L. 2009a. ImageNet: A large-scale hierarchical image database. In *CVPR*.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009b. Imagenet: A large-scale hierarchical image database. In *CVPR*.

Desai, K.; and Johnson, J. 2021. Virtex: Learning visual representations from textual annotations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 11162–11173.

Elhoseiny, M.; Saleh, B.; and Elgammal, A. M. 2013. Write a Classifier: Zero-Shot Learning Using Purely Textual Descriptions. In *ICCV*.

Fei-Fei, L.; Fergus, R.; and Perona, P. 2004. 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*.

Frome, A.; Corrado, G. S.; Shlens, J.; Bengio, S.; Dean, J.; Ranzato, M.; and Mikolov, T. 2013a. Devise: A deep visual-semantic embedding model. *NeurIPS*.

Frome, A.; Corrado, G. S.; Shlens, J.; Bengio, S.; Dean, J.; Ranzato, M.; and Mikolov, T. 2013b. DeViSE: A Deep Visual-Semantic Embedding Model. In *NeurIPS*.

Fürst, A.; Rumetshofer, E.; Tran, V.; Ramsauer, H.; Tang, F.; Lehner, J.; Kreil, D. P.; Kopp, M.; Klambauer, G.; Bitto-Nemling, A.; and Hochreiter, S. 2021. CLOOB: Modern Hopfield Networks with InfoLOOB Outperform CLIP. *CoRR*, abs/2110.11316.

Gao, T.; Fisch, A.; and Chen, D. 2021. Making Pre-trained Language Models Better Few-shot Learners. In *ACL/IJCNLP*.

Gomez, L.; Patel, Y.; Rusinol, M.; Karatzas, D.; and Jawahar, C. 2017. Self-supervised learning of visual features through embedding images into text topic spaces. In *CVPR*.

Goswami, K.; Lange, L.; Araki, J.; and Adel, H. 2023. SwitchPrompt: Learning Domain-Specific Gated Soft Prompts for Classification in Low-Resource Domains. *CoRR*, abs/2302.06868.

Guo, Y.; Shi, H.; Kumar, A.; Grauman, K.; Rosing, T.; and Feris, R. S. 2019. SpotTune: Transfer Learning Through Adaptive Fine-Tuning. In *CVPR*.

Helber, P.; Bischke, B.; Dengel, A.; and Borth, D. 2018. Introducing Eurosat: A Novel Dataset and Deep Learning Benchmark for Land Use and Land Cover Classification. In *2018 IEEE International Geoscience and Remote Sensing Symposium, IGARSS 2018, Valencia, Spain, July 22-27, 2018*.

Jia, C.; Yang, Y.; Xia, Y.; Chen, Y.; Parekh, Z.; Pham, H.; Le, Q. V.; Sung, Y.; Li, Z.; and Duerig, T. 2021. Scaling Up Visual and Vision-Language Representation Learning With Noisy Text Supervision. In *ICML*.

Jiang, Z.; Xu, F. F.; Araki, J.; and Neubig, G. 2020. How Can We Know What Language Models Know. *Trans. Assoc. Comput. Linguistics*, 8: 423–438.

Joulin, A.; Van Der Maaten, L.; Jabri, A.; and Vasilache, N. 2016. Learning visual features from large weakly supervised data. In *ECCV*.

Kamath, A.; Singh, M.; LeCun, Y.; Synnaeve, G.; Misra, I.; and Carion, N. 2021. MDETR - Modulated Detection for End-to-End Multi-Modal Understanding. In *ICCV*.

Khattak, M. U.; Rasheed, H.; Maaz, M.; Khan, S.; and Khan, F. S. 2023. Maple: Multi-modal prompt learning. In *CVPR*, 19113–19122.

Kolesnikov, A.; Beyer, L.; Zhai, X.; Puigcerver, J.; Yung, J.; Gelly, S.; and Houlsby, N. 2020. Big Transfer (BiT): General Visual Representation Learning. In *ECCV*.

Krause, J.; Stark, M.; Deng, J.; and Fei-Fei, L. 2013. 3D Object Representations for Fine-Grained Categorization. In *ICCV*.

Kuznetsova, A.; Rom, H.; Alldrin, N.; Uijlings, J. R. R.; Krasin, I.; Pont-Tuset, J.; Kamali, S.; Popov, S.; Malloci, M.; Duerig, T.; and Ferrari, V. 2018. The Open Images Dataset V4: Unified image classification, object detection, and visual relationship detection at scale. *CoRR*, abs/1811.00982.

Lester, B.; Al-Rfou, R.; and Constant, N. 2021. The Power of Scale for Parameter-Efficient Prompt Tuning. In *EMNLP*.

Li, A.; Jabri, A.; Joulin, A.; and Van Der Maaten, L. 2017. Learning visual n-grams from web data. In *ICCV*.

Li, L. H.; Yatskar, M.; Yin, D.; Hsieh, C.; and Chang, K. 2019. VisualBERT: A Simple and Performant Baseline for Vision and Language. *CoRR*.

Li, L. H.; Zhang, P.; Zhang, H.; Yang, J.; Li, C.; Zhong, Y.; Wang, L.; Yuan, L.; Zhang, L.; Hwang, J.; Chang, K.; and Gao, J. 2022a. Grounded Language-Image Pre-training. In *CVPR*.

Li, X. L.; and Liang, P. 2021. Prefix-Tuning: Optimizing Continuous Prompts for Generation. In *ACL/IJCNLP*.

Li, Y.; Liang, F.; Zhao, L.; Cui, Y.; Ouyang, W.; Shao, J.; Yu, F.; and Yan, J. 2022b. Supervision Exists Everywhere: A Data Efficient Contrastive Language-Image Pre-training Paradigm. In *ICLR*.

Li, Y.; Yin, Y.; Li, J.; and Zhang, Y. 2022c. Prompt-Driven Neural Machine Translation. In *ACL*.

Liu, X.; Ji, K.; Fu, Y.; Tam, W.; Du, Z.; Yang, Z.; and Tang, J. 2022. P-Tuning: Prompt Tuning Can Be Comparable to Fine-tuning Across Scales and Tasks. In *ACL*.

Liu, X.; Zheng, Y.; Du, Z.; Ding, M.; Qian, Y.; Yang, Z.; and Tang, J. 2021. GPT Understands, Too. *CoRR*, abs/2103.10385.

Luong, T.; Pham, H.; and Manning, C. D. 2015. Effective Approaches to Attention-based Neural Machine Translation. In *EMNLP*.

Maji, S.; Rahtu, E.; Kannala, J.; Blaschko, M. B.; and Vedaldi, A. 2013. Fine-Grained Visual Classification of Aircraft. *CoRR*, abs/1306.5151.

Nilsback, M.; and Zisserman, A. 2008. Automated Flower Classification over a Large Number of Classes. In *ICVGIP*.

Parkhi, O. M.; Vedaldi, A.; Zisserman, A.; and Jawahar, C. V. 2012. Cats and dogs. In *CVPR*.

Parmar, M.; Mishra, S.; Purohit, M.; Luo, M.; Mohammad, M.; and Baral, C. 2022. In-BoXBART: Get Instructions into Biomedical Multi-Task Learning. In *NAACL*, 112–128. Association for Computational Linguistics.

Petroni, F.; Rocktäschel, T.; Riedel, S.; Lewis, P. S. H.; Bakhtin, A.; Wu, Y.; and Miller, A. H. 2019. Language Models as Knowledge Bases? In *EMNLP-IJCNLP*.

Plested, J.; and Gedeon, T. 2022. Deep transfer learning for image classification: a survey. *CoRR*, abs/2205.09904.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; Krueger, G.; and Sutskever, I. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *ICML*.

Roy, S.; and Etemad, A. 2023. Consistency-guided Prompt Learning for Vision-Language Models. *CoRR*, abs/2306.01195.

Sarkar, R.; Goswami, K.; Arcan, M.; and McCrae, J. P. 2023. PICKD: In-Situ Prompt Tuning for Knowledge-Grounded Dialogue Generation. In Kashima, H.; Idé, T.; and Peng, W., eds., *PAKDD*.

Schick, T.; and Schütze, H. 2021. Exploiting Cloze-Questions for Few-Shot Text Classification and Natural Language Inference. In *EACL*, 255–269. Association for Computational Linguistics.

Socher, R.; Ganjoo, M.; Manning, C. D.; and Ng, A. Y. 2013. Zero-Shot Learning Through Cross-Modal Transfer. In Burges, C. J. C.; Bottou, L.; Ghahramani, Z.; and Weinberger, K. Q., eds., *NeurIPS*.

Soomro, K.; Zamir, A. R.; and Shah, M. 2012. UCF101: A Dataset of 101 Human Actions Classes From Videos in The Wild. *CoRR*, abs/1212.0402.

Sun, C.; Shrivastava, A.; Singh, S.; and Gupta, A. 2017. Revisiting Unreasonable Effectiveness of Data in Deep Learning Era. In *ICCV*.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; and Polosukhin, I. 2017. Attention is All you Need. In *NeurIPS*.

Wan, R.; Xiong, H.; Li, X.; Zhu, Z.; and Huan, J. 2019. Towards Making Deep Transfer Learning Never Hurt. In *ICDM*.

Xiao, J.; Hays, J.; Ehinger, K. A.; Oliva, A.; and Torralba, A. 2010. SUN database: Large-scale scene recognition from abbey to zoo. In *CVPR*.

Yao, H.; Zhang, R.; and Xu, C. 2023. Visual-language prompt tuning with knowledge-guided context optimization. In *CVPR*, 6757–6767.

Zhou, K.; Yang, J.; Loy, C. C.; and Liu, Z. 2022a. Conditional Prompt Learning for Vision-Language Models. In *CVPR*.

Zhou, K.; Yang, J.; Loy, C. C.; and Liu, Z. 2022b. Learning to Prompt for Vision-Language Models. *Int. J. Comput. Vis.*, 130(9): 2337–2348.

Zhu, B.; Niu, Y.; Han, Y.; Wu, Y.; and Zhang, H. 2023. Prompt-aligned Gradient for Prompt Tuning. ICCV 2023.

Zhu, Y.; Zhou, X.; Qiang, J.; Li, Y.; Yuan, Y.; and Wu, X. 2022. Prompt-Learning for Short Text Classification. *CoRR*, abs/2202.11345.