Learning to Prompt for Vision-Language Models

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

Motivation

Caltech101

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

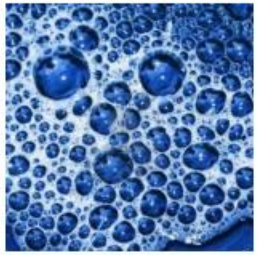
(a)

Flowers102

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

(b)

Describable Textures (DTD)



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

(c)

EuroSAT

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

(d)

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

CotextOptimization

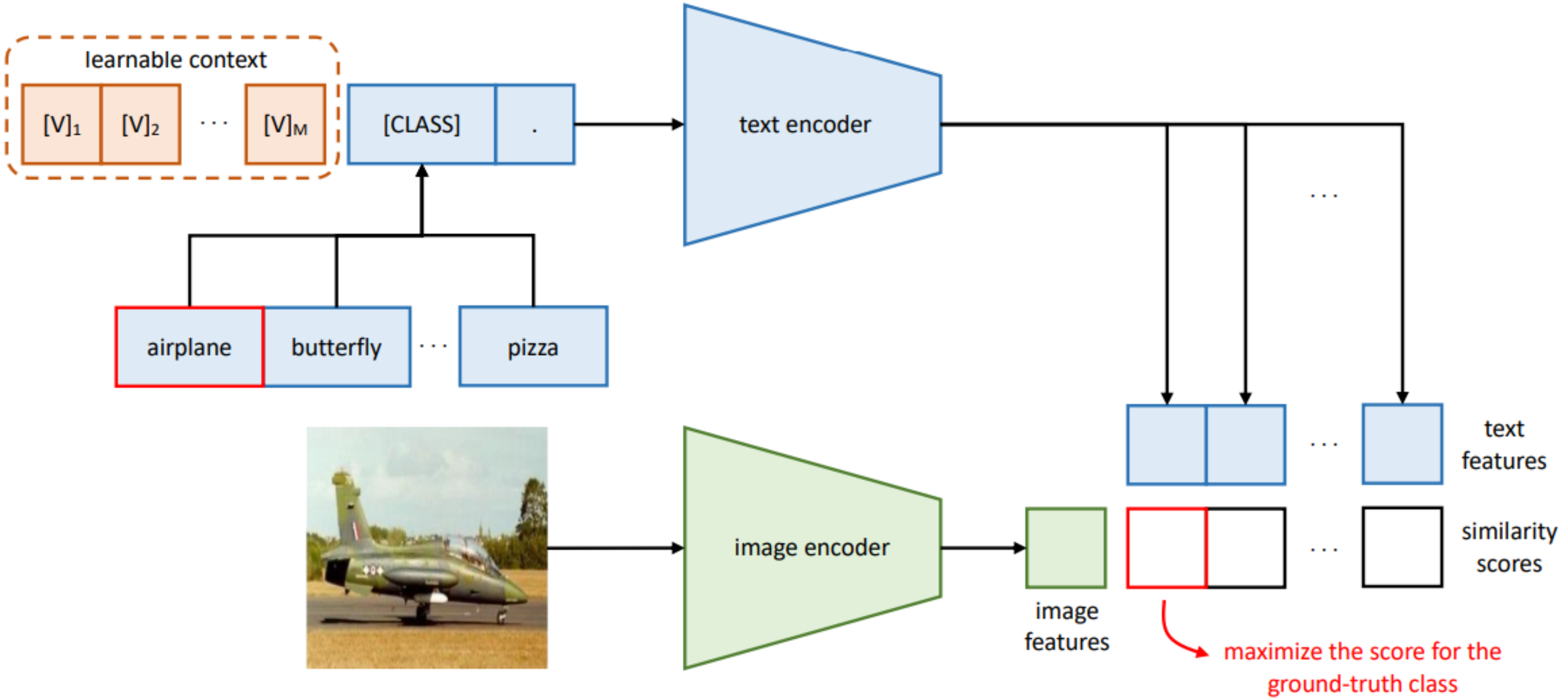


Fig. 2 Overview of Context Optimization (CoOp). The main idea is to model a prompt's context using a set of learnable vectors, which can be optimized through minimizing the classification loss. Two designs are proposed: one is unified context, which shares the same context vectors with all classes; and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

CotextOptimization

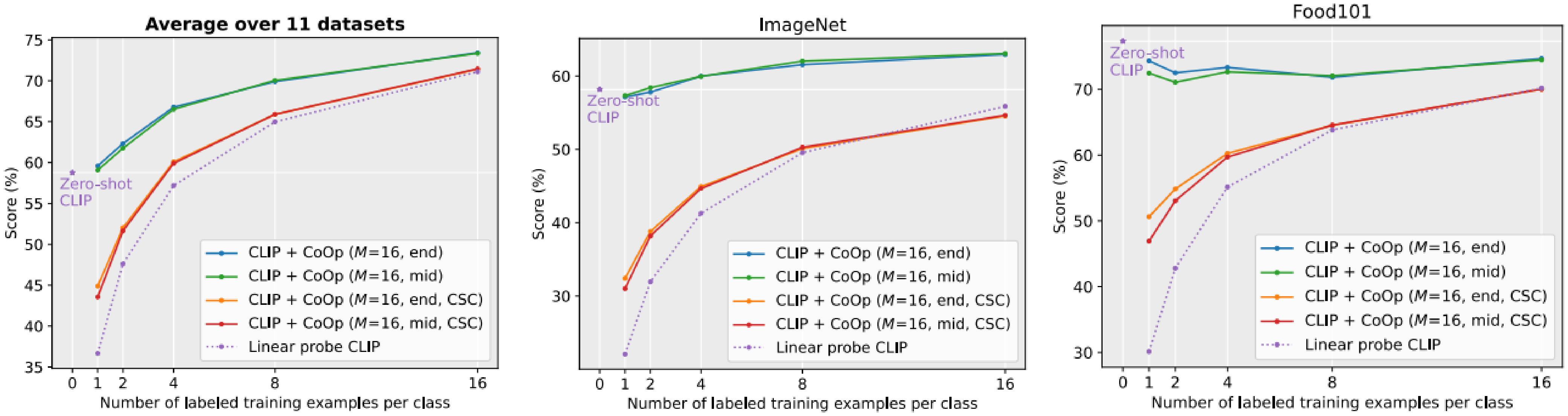
$$p(y=i|\boldsymbol{x}) = \frac{\exp(\cos(\boldsymbol{w_i}, \boldsymbol{f})/\tau)}{\sum_{j=1}^{K} \exp(\cos(\boldsymbol{w_j}, \boldsymbol{f})/\tau)}$$

$$\boldsymbol{t} = [\mathrm{V}]_1[\mathrm{V}]_2 \ldots [\mathrm{V}]_M[\mathrm{CLASS}]$$

$$\boldsymbol{t} = [\mathrm{V}]_1 \ldots [\mathrm{V}]_{\frac{M}{2}}[\mathrm{CLASS}][\mathrm{V}]_{\frac{M}{2}+1} \ldots [\mathrm{V}]_M$$

$$p(y=i|\boldsymbol{x}) = \frac{\exp(\cos(g(\boldsymbol{t}_i), \boldsymbol{f})/\tau)}{\sum_{j=1}^{K} \exp(\cos(g(\boldsymbol{t}_j), \boldsymbol{f})/\tau)}$$

Experiment


Average over 11 datasets
ImageNet
Food101
Score (%)
Number of labeled training examples per class
Zero-shot CLIP
CLIP + CoOp (M=16, end)
CLIP + CoOp (M=16, mid)
CLIP + CoOp (M=16, end, CSC)
CLIP + CoOp (M=16, mid, CSC)
Linear probe CLIP

Experiment

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

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

Experiment

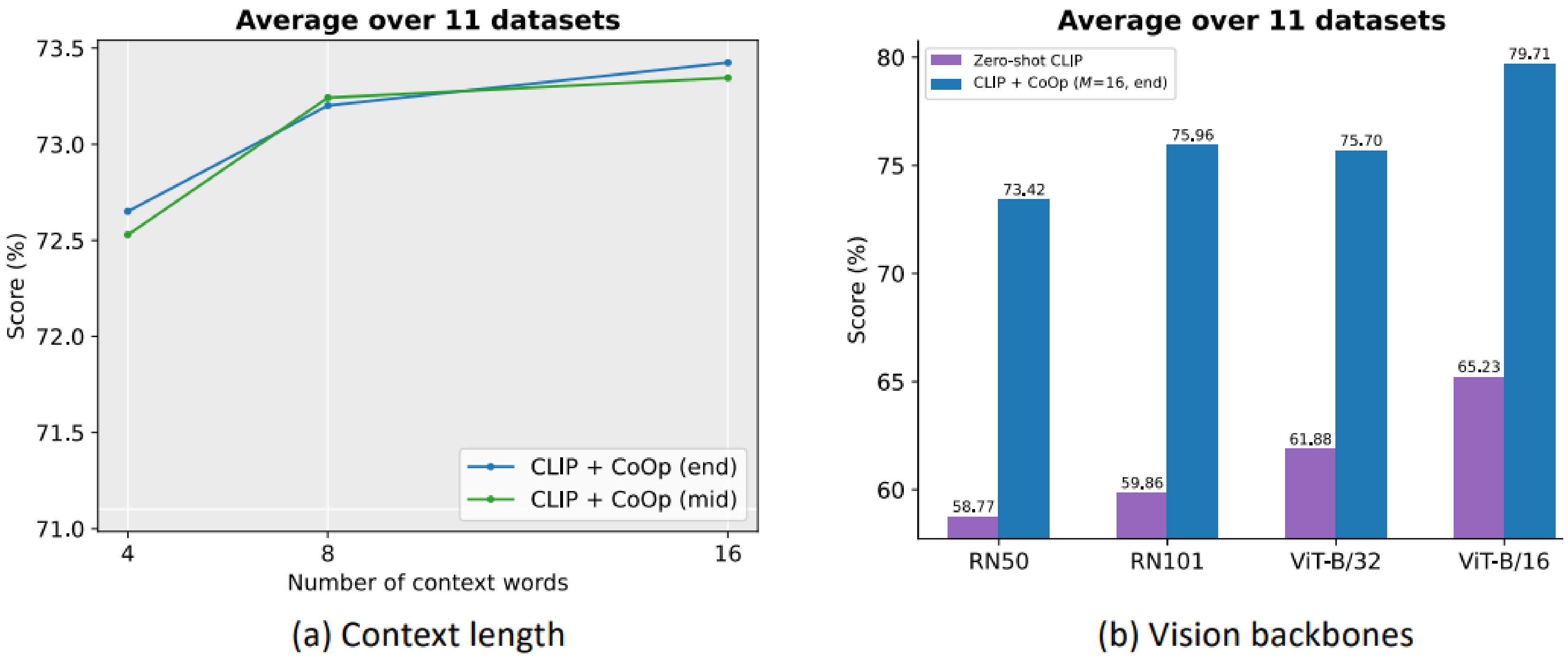


Fig. 5 Investigations on CoOp's context length and various vision backbones.

Experiment

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

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

Table 3 Random vs manual initialization.

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

Interpreting the Learned Prompts

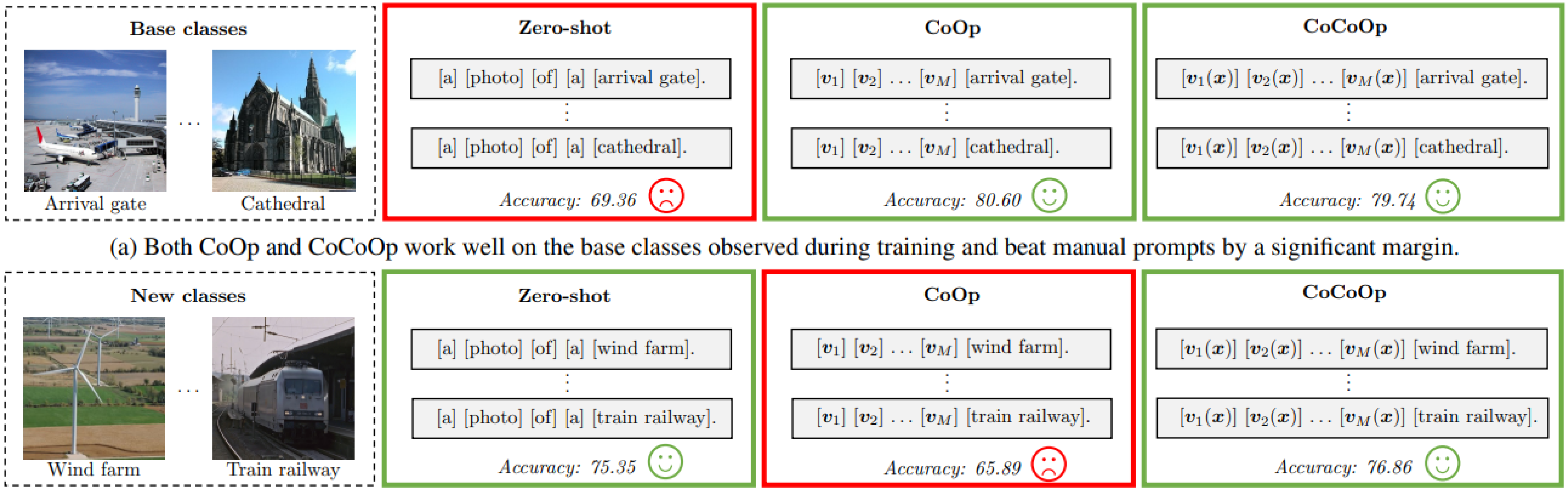
Table 4 The nearest words for each of the 16 context vectors learned by CoOp, with their distances shown in parentheses. N/A means non-Latin characters.

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

Conditional Prompt Learning for Vision-Language Models

Kaiyang Zhou Jingkang Yang Chen Change Loy Ziwei Liu ✉
S-Lab, Nanyang Technological University, Singapore
{kaiyang.zhou, jingkang001, ccloy, ziwei.liu}@ntu.edu.sg

Overfitting problem



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

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

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

Assumption

- The context, which is fixed once learned, is optimized only for a specific set of (training) classes.
- Make a prompt conditioned on each input instance (image) rather than fixed once learned could be more generalizable

- Static prompt -> Dynamic prompt

Method

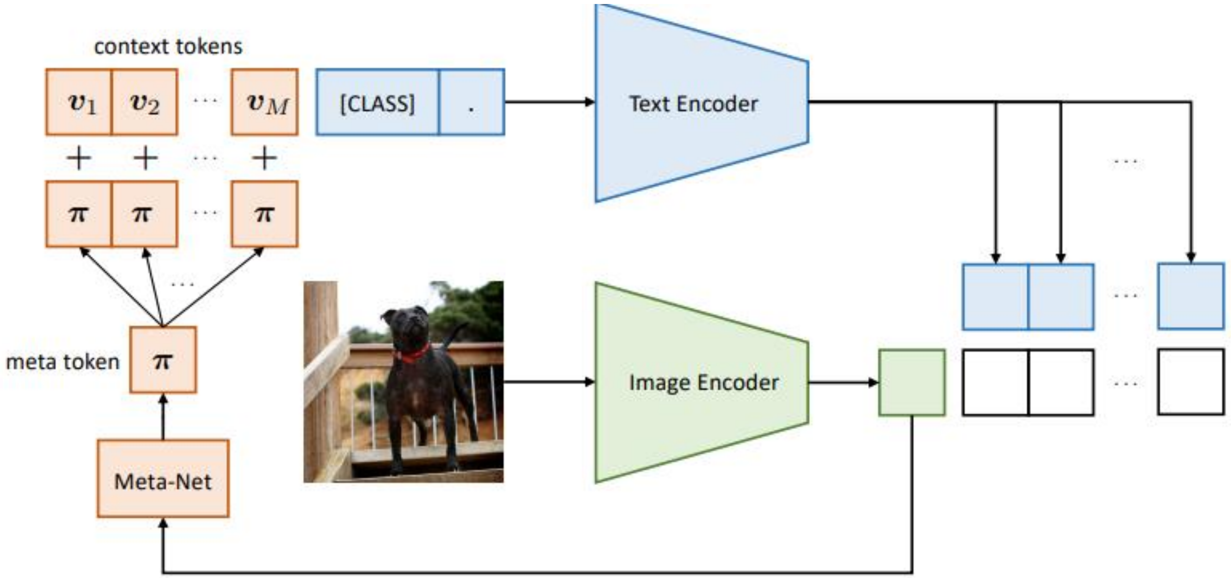


Figure 2. Our approach, Conditional Context Optimization (CoCoOp), consists of two learnable components: a set of context vectors and a lightweight neural network (Meta-Net) that generates for each image an input-conditional token.

Method

vm(x) = vm + π where π = hθ(x) and m ∈ {1, 2, ..., M}.

ti(x) = {v1(x), v2(x), . . . , vM(x), ci}.

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

Experiment

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

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

Experiment

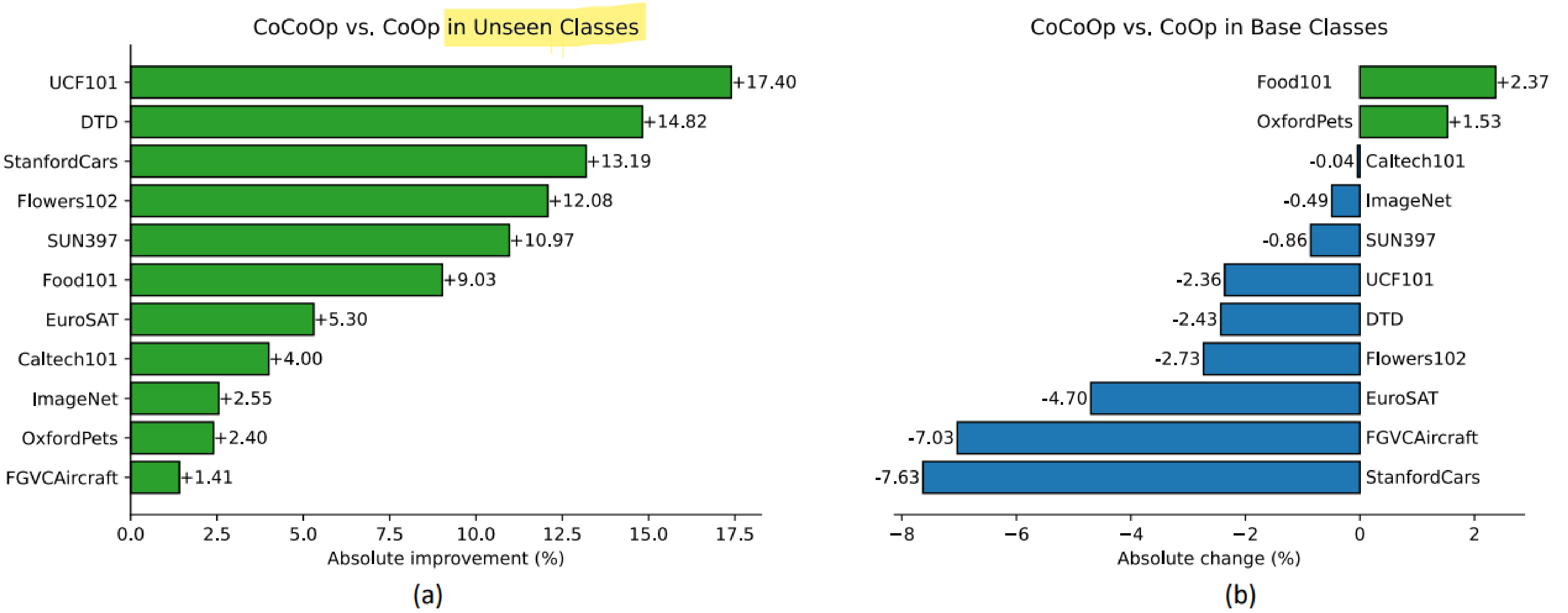


Figure 3. **Comprehensive comparisons of CoCoOp and CoOp in the base-to-new generalization setting**. (a) CoCoOp is able to gain consistent improvements over CoOp in unseen classes on all datasets. (b) CoCoOp's declines in base accuracy are mostly under 3%, which are far outweighed by the gains in generalization.

Experiment

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

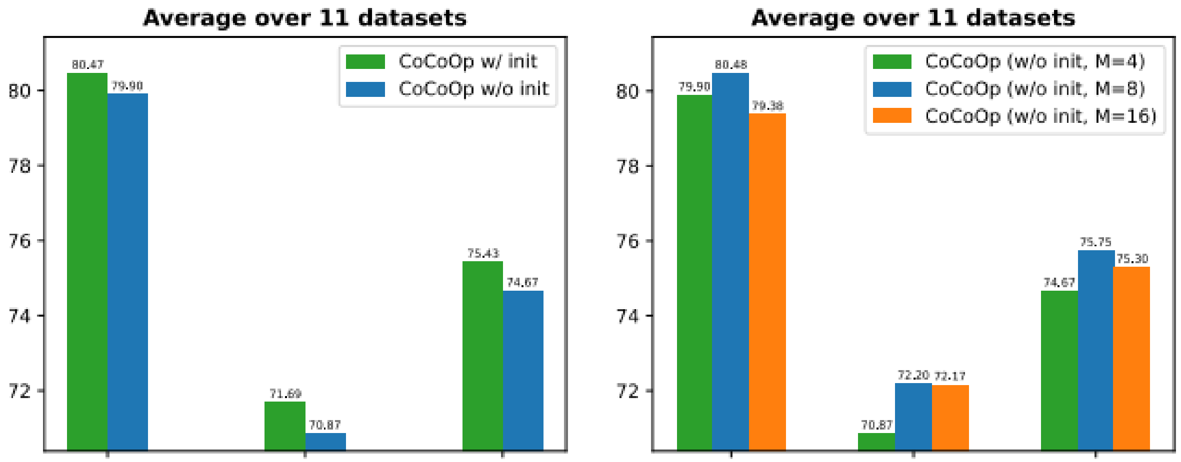
	Source	Target										
	ImageNet	Caltech101	OxfordPets	StanfordCars	Flowers102	Food101	FGVCAircraft	SUN397	DTD	EuroSAT	UCF101	*Average*
CoOp [63]	**71.51**	93.70	89.14	64.51	68.71	85.30	18.47	64.15	41.92	**46.39**	66.55	63.88
CoCoOp	71.02	**94.43**	**90.14**	**65.32**	**71.88**	**86.06**	**22.94**	**67.36**	**45.73**	45.37	**68.21**	**65.74**
Δ	-0.49	+0.73	+1.00	+0.81	+3.17	+0.76	+4.47	+3.21	+3.81	-1.02	+1.66	+1.86

Experiment

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

		Source	Target			
	Learnable?	ImageNet	ImageNetV2	ImageNet-Sketch	ImageNet-A	ImageNet-R
CLIP [40]		66.73	60.83	46.15	47.77	73.96
CoOp [63]	✓	**71.51**	**64.20**	47.99	49.71	75.21
CoCoOp	✓	71.02	64.07	**48.75**	**50.63**	**76.18**

Ablation Study



(a) Ablation on initialization.

(b) Ablation on context length.

Figure 4. Ablation studies.

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

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