Test-Time Prompt Tuning for Zero-Shot Generalization in Vision-Language Models

Manli Shu, Weili Nie, De-An Huang, Zhiding Yu, Tom Goldstein, Anima Anandkumar, Chaowei Xiao

Presenter: Tianjiao Yu

Overview

- Background: CLIP
- Prompt Learning for VL models: CoOp
- Conditional Prompt learning for VL models: CoCoOp
- Test-time prompt learning for VL models: TPT

Background: Contrastive Language-Image Pretraining

- A bridge between computer vision and natural language processing
- A multimodal model built on hundreds of millions of images and captions
- Can return the best caption given an image
- Has impressive "zero-shot" capabilities, making it able to accurately predict entire classes it's never seen before

Background: CLIP

Previous datasets might be large but lack of corresponding textual description

- YFCC100M shrunk by a factor of 6 to only 15m photos.
- Constructed a new dataset of 400 million image text pairs
  - Get queries from wikepedia
  - Use queries to search for image-text pairs
- Collect around 20,000 pairs for 500,000 queries so that the data is balanced

Background: CLIP

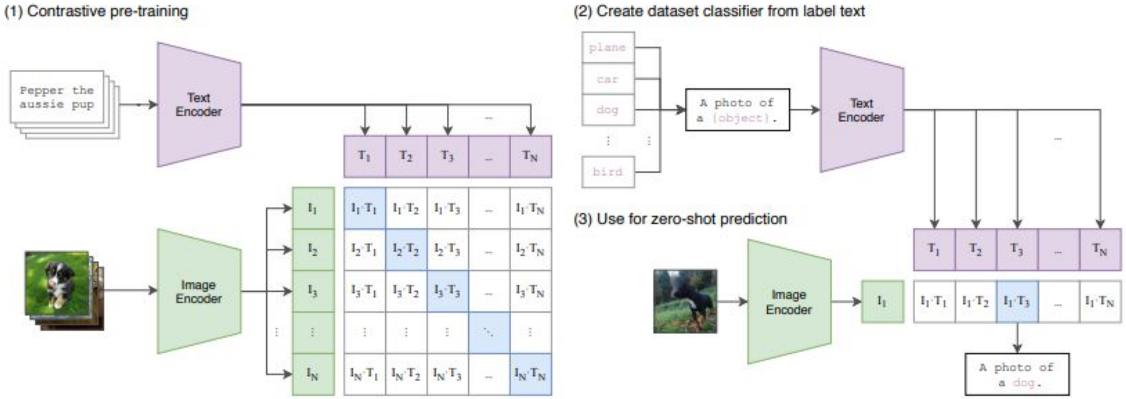
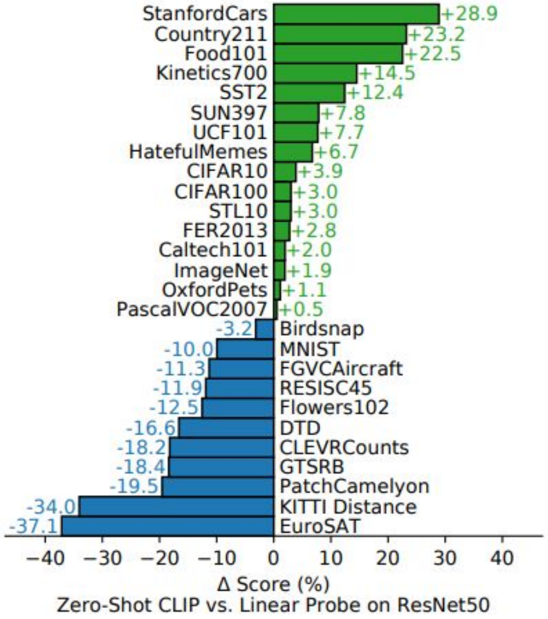
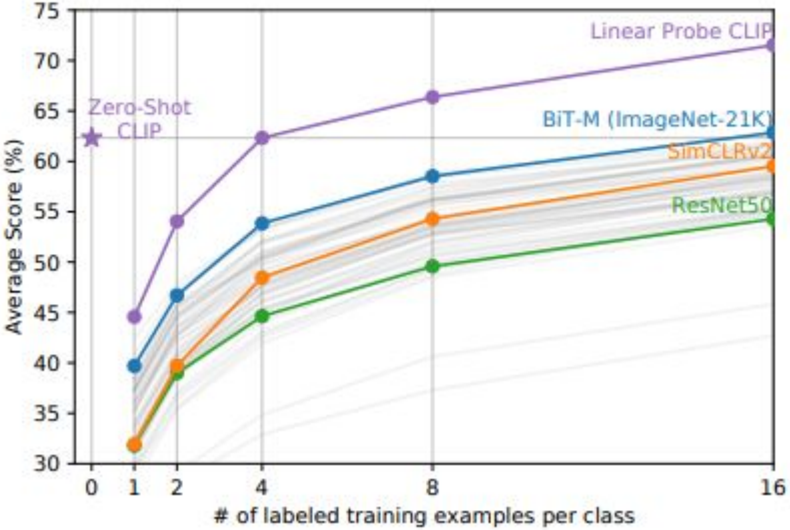


Figure 1. Summary of our approach. While standard image models jointly train an image feature extractor and a linear classifier to predict some label, CLIP jointly trains an image encoder and a text encoder to predict the correct pairings of a batch of (image, text) training examples. At test time the learned text encoder synthesizes a zero-shot linear classifier by embedding the names or descriptions of the target dataset's classes.

Background: CLIP





Background: CLIP

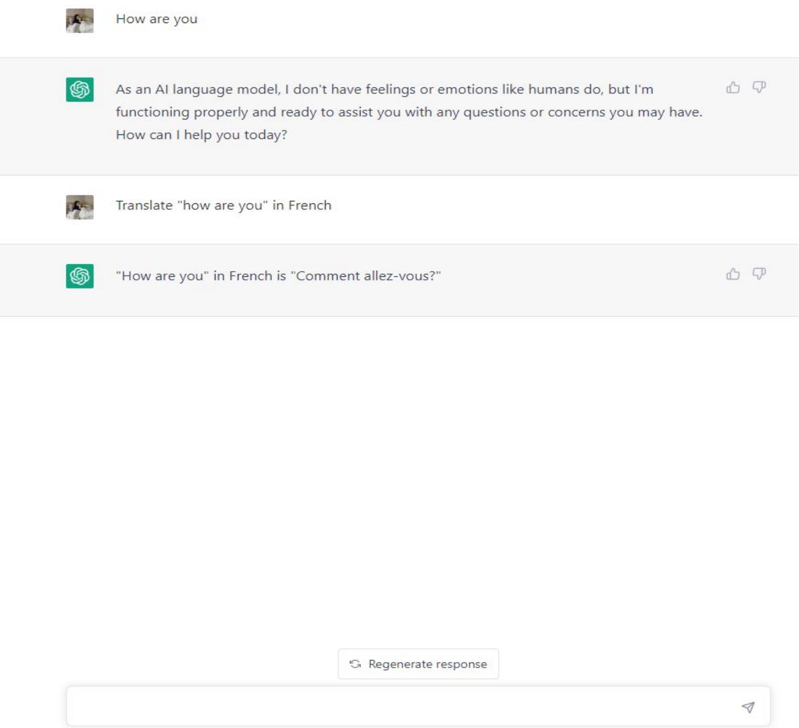
Weaknesses?

- Zero-shot performance well worse than fine-tuned SotA
- Does not work well with image regions
- Sensitive to prompt wording
    - Polysemy, some images are tagged with just a class label and not a full-text prompt
    - “Boxer” as a type of dog, but perceived as an athlete

Learning to Prompt

From NLP:

Large pre-trained language models



Learning to Prompt

Previous visual recognition system:

- ResNet or ViT: Limited in closed-set concepts; New categories requires more data for learning new classifiers
- CLIP and ALIGN: align images and raw texts using two separate encoders; By pre-training at a large scale, models can learn diverse concepts and readily be transferred to different downstream tasks.
- **Natural language** is used to reference learned visual concepts

Learning to Prompt

Text prompt plays a key role in downstream datasets.

Different prompts lead to different performance

But how do we identify the right prompt?

Caltech101

Prompt	Accuracy
a [CLASS].	82.68
a photo of [CLASS].	80.81
a photo of a [CLASS].	86.29
$[V]_1 [V]_2 \dots [V]_M$ [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
$[V]_1 [V]_2 \dots [V]_M$ [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
$[V]_1 [V]_2 \dots [V]_M$ [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
$[V]_1 [V]_2 \dots [V]_M$ [CLASS].	**83.53**

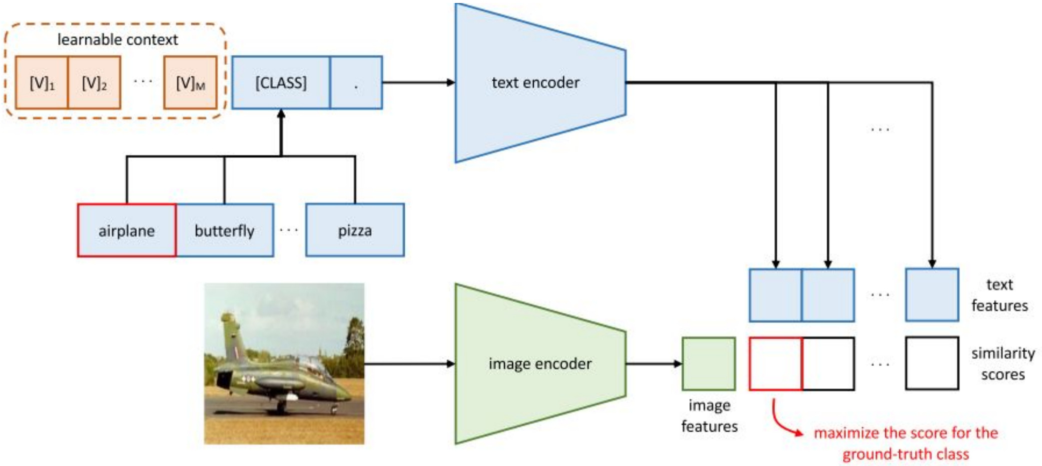
(d)

Learning to Prompt for Vision-Language Models

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

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

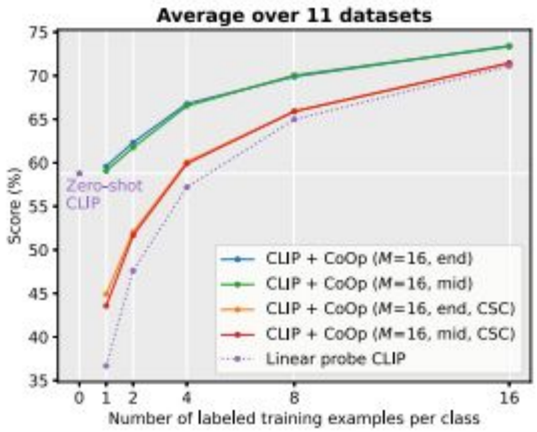
$$\boldsymbol{t} = [\mathrm{V}]_1[\mathrm{V}]_2 \dots [\mathrm{V}]_M[\mathrm{CLASS}]$$



Learning to Prompt for Vision-Language Models

CoOp is a strong few-shot learner, requiring only two shots on average to get decent margin over CLIP

Given 16 shots for training, the average gap brought by CoOp can be further increased to around 15%

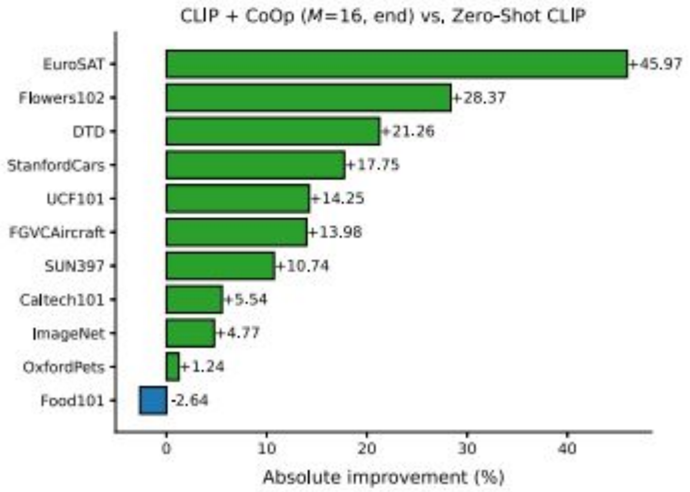


Learning to Prompt for Vision-Language Models

Specialized tasks (e.g. EuroSAT, DTD) increase over 45% and 20% respectively

Better performance on most fine-grained datasets (e.g. Flowers102, StanfordCars)

Improvement on OxfordPets and Food101 are less appealing

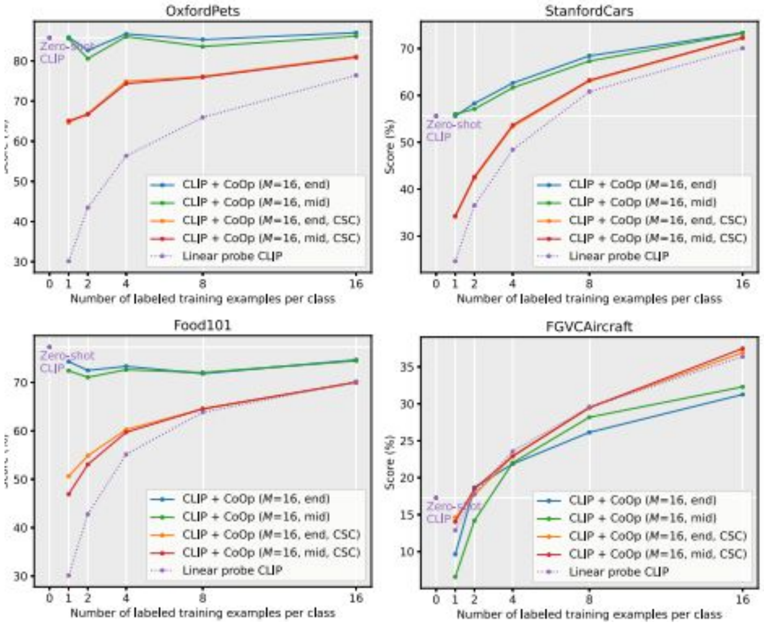


Learning to Prompt for Vision-Language Models

Loss of momentum in performance improvement for OxfordPets and Food101. Could be overfitting

CoOp demonstrates clear advantages over the linear probe model

On average, using unified context leads to better performance



Learning to Prompt for Vision-Language Models

Domain Generalization:

- Comparison with zero-shot CLIP on robustness to distribution shift using different vision backbones
- CoOp enhances CLIP's robustness to distribution shifts, despite the exposure to the source dataset
- Linear probe model obtains much worse results, exposing its weakness in domain generalization.

	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**

Learning to Prompt for Vision-Language Models

Further Analysis:

- Shorter context length benefits domain generalization, longer for better performance
- CoOp outperforms prompt ensembling
- Random initialization is sufficient

Conditional Prompt Learning for VL Models

- To fit web-scale data, such as the 400 million pairs of images and texts (CLIP)
- VL models are intentionally designed to have high capacity. Sometimes, even fine-tuning is impractical.
- A safer approach is to tune a prompt by adding some context that is meaningful to a task
- However, prompt engineering is extremely time-consuming as it has to be based on trial and error, hence the CoOp model.
- But in CoOp, the learned context is not generalizable to wider unseen classes.

Conditional Prompt Learning for VL Models

This suggests that the learned context overfits the base classes, thus failing to capture more generalizable elements.

The context is fixed once learned in CoOp.

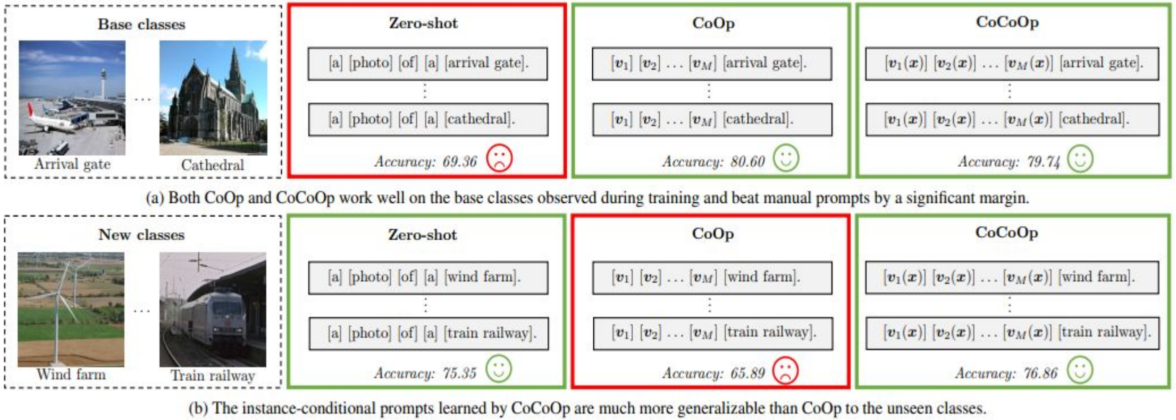


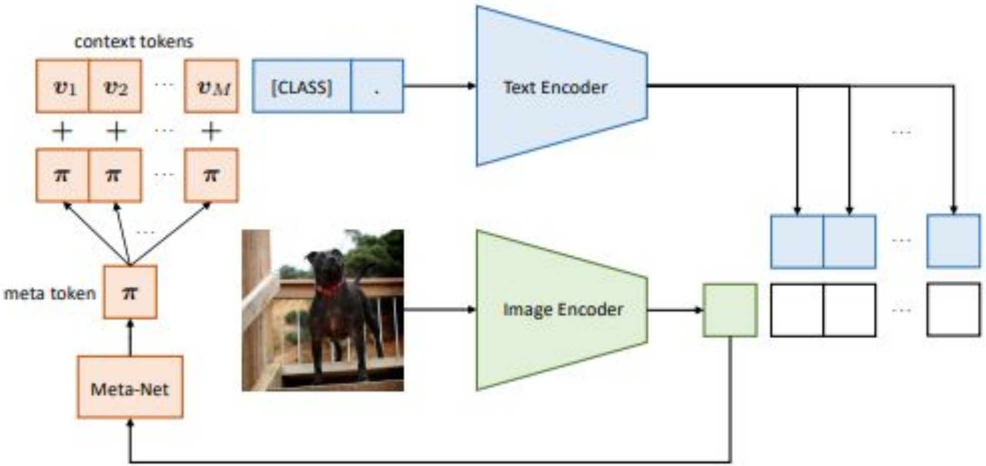
Figure 1. **Motivation of our research: to learn generalizable prompts**. The images are randomly selected from SUN397 [55], which is a widely-used scene recognition dataset.

Conditional Prompt Learning for VL Models

The key idea is to make a prompt conditioned on each input instance (image) rather than fixed once learned

Extend CoOp by further learning a lightweight neural network to generate for each image an input-conditional token (vector)

Similar to *Show and Tell (Vinyals et. al 2015)*, which validates that it is more robust to class shift

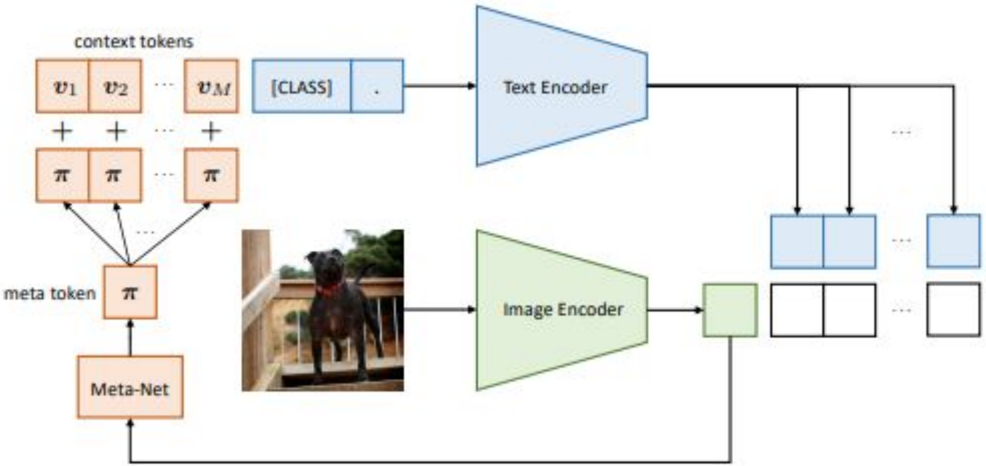


Conditional Prompt Learning for VL Models

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

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

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



Conditional Prompt Learning for VL Models

CoOp's new accuracy is consistently much weaker than the base accuracy on nearly all datasets

CoCoOp Significantly Narrows Generalization Gap

CoCoOp Is More Compelling Than CLIP

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

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

(b) ImageNet.

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

(c) Caltech101.

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

(d) OxfordPets.

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

(e) StanfordCars.

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

(f) Flowers102.

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

(g) Food101.

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

(h) FGVCAircraft.

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

(i) SUN397.

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

(j) DTD.

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

(k) EuroSAT.

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

(l) UCF101.

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

Conditional Prompt Learning for VL Models

Comparison of prompt learning methods in the cross-dataset transfer setting

CoCoOp exhibits much stronger transferability than CoOp

	Source	Target										
	ImageNet	Caltech101	OxfordPets	StanfordCars	Flowers102	Food101	FGVCAircraft	SUN397	DTD	EuroSAT	UCF101	*Average*
CoOp [62]	**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

Prompt Learning Limitations

CoOp:

- Interpreting the learned prompts is hard

CoCoOp:

- It is slow to train and would consume a significant amount of GPU memory if the batch size is set larger than one, as each image needs an independent forward pass.
- Unseen classes still lags behind CLIP (7 out of 11 datasets)

Test-Time Prompt Tuning: Intro

- Vision-language pre-training, such as CLIP[1] and ALIGN[11], present a promising direction for developing foundation models for vision tasks
  - encode a wide range of visual concepts after training on millions of noisy image-text pairs
  - can be applied to downstream tasks in a zero-shot manner
  - This is made possible by designed appropriate instruction prompts
- Recent works address this by proposing prompt tuning to directly learn prompts using training data
  - We can fine-tune prompts with training data in the same way we finetune model parameters
  - But the learned prompts are limited to the distribution and tasks corresponding to training data
  - It also requires training data which can be expensive or not available for zero-shot tasks

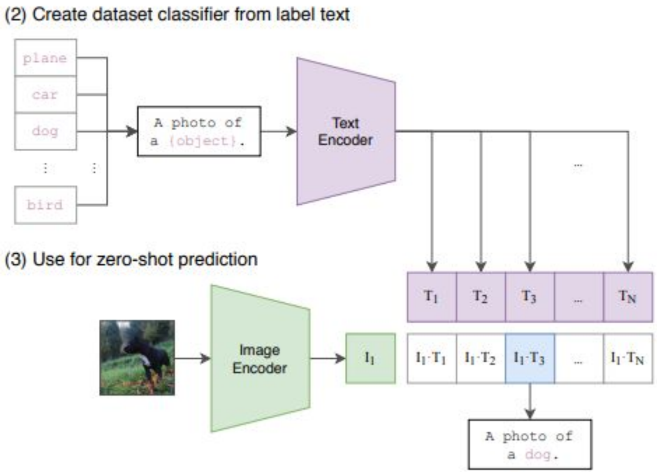
Test-Time Prompt Tuning: Related Work

- Prompting for foundation models
  - Large-scale heterogeneous foundation models
  - Prompt for different downstream tasks
  - NLP -> VL; Require annotations -> single test sample
- Generalization under data distribution shifts
  - Need to handle the discrepancy between the underlying distributions of the test and the training data
  - CLIP can generalize to downstream tasks with various distribution shifts in a zero-shot manner
  - Better the CLIP by using consistency regularization as an additional objective with the confidence selection module.
- Test-time optimization
  - Adapting machine learning models to test samples on the fly
  - TENT [9] proposes a test-time objective by minimizing the entropy of the batch-wise prediction probability distributions
  - Zhang et al. [10] bypass the multi-sample requirements using data augmentations
  - Refine the entropy minimization by proposing confidence selection

Test-Time Prompt Tuning: Method

CLIP[1] with a hand-crafted prompt:

1. We prepend a hand-crafted prompt prefix to every class
2. Feed them to the text encoder
3. Each text feature is paired with the image feature.
4. Find the best pair base on similarity score



Test-Time Prompt Tuning: Method

- text inputs {p; Y} = {{p; yi} for yi ∈ Y} provide the model with the most helpful context information about the task

$$\boldsymbol{p}^* = \arg\min_{\boldsymbol{p}} \mathbb{E}_{(X,y)\sim\mathcal{D}_{\text{train}}} \mathcal{L}(\mathcal{F}_{\boldsymbol{p}}(X), y),$$
$$\text{where } \mathcal{F}_{\boldsymbol{p}}(X) = \texttt{sim}(\mathbf{E}_{\texttt{text}}(\{\boldsymbol{p}; \mathcal{Y}\}), \mathbf{E}_{\texttt{visual}}(X)).$$

- TPT optimizes the prompt p at test time based on the single test sample

$$\boldsymbol{p}^* = \arg\min_{\boldsymbol{p}} \mathcal{L}(\mathcal{F}, \boldsymbol{p}, X_{\text{test}})$$

Test-Time Prompt Tuning: Method

TPT for image classification:

- Must select an unsupervised loss
- The objective promotes the consistency across different augmented views of a given test image

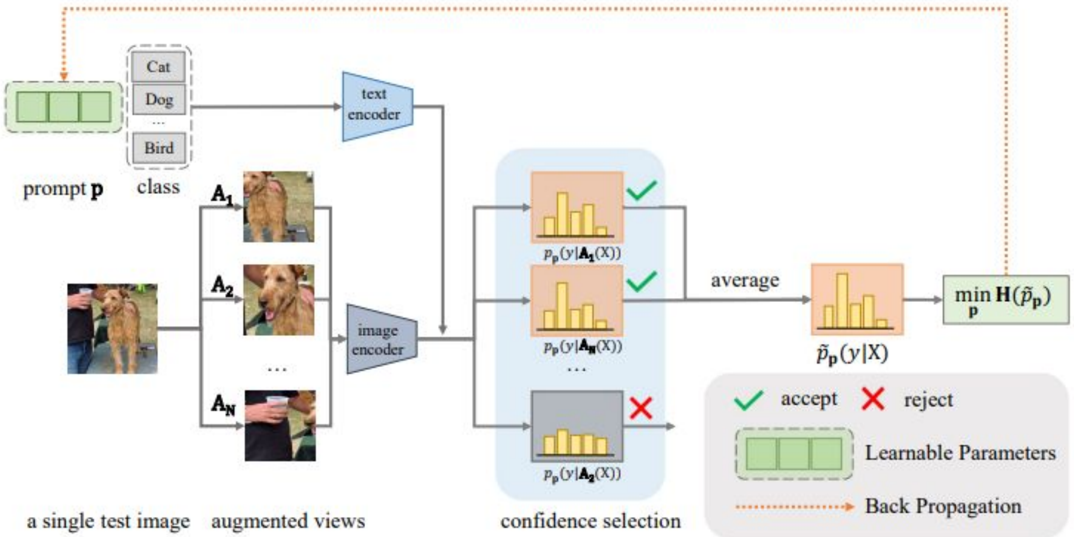
$$\boldsymbol{p}^* = \arg\min_{\boldsymbol{p}} -\sum_{i=1}^{K} \tilde{p}_{\boldsymbol{p}}(y_i|X_{\text{test}}) \log \tilde{p}_{\boldsymbol{p}}(y_i|X_{\text{test}}),$$

$$\text{where } \tilde{p}_{\boldsymbol{p}}(y_i|X_{\text{test}}) = \frac{1}{N}\sum_{i=1}^{N} p_{\boldsymbol{p}}(y_i|\mathcal{A}_i(X_{\text{test}})).$$

- Propose confidence selection to filter out views that generate high-entropy

$$\tilde{p}_{\boldsymbol{p}(y|X_{\text{test}})} = \frac{1}{\rho N}\sum_{i=1}^{N} \mathbb{1}[\mathbf{H}(p_i) \leq \tau] p_{\boldsymbol{p}}(y|\mathcal{A}_i(X_{\text{test}}))$$

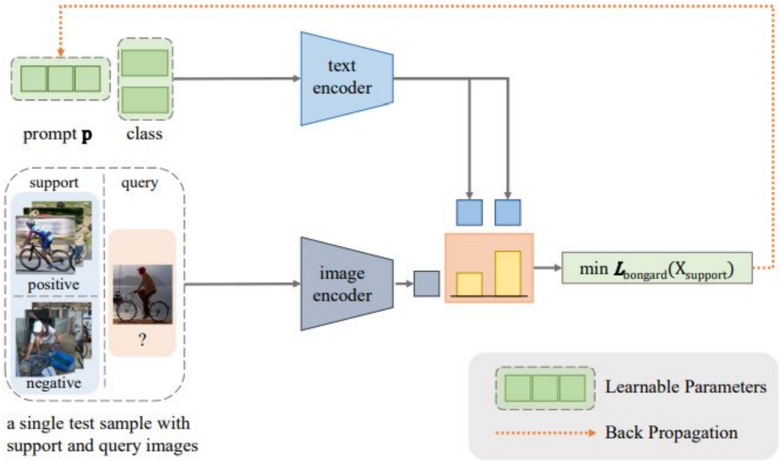
Test-Time Prompt Tuning: TPT



Test-Time Prompt Tuning: Method

Context-dependent visual reasoning:

- correctness of the prediction depends on the context
- Learn an optimal label token cls on the example images



Experiments: Robustness to Distribution Shifts

Datasets:

- Follow the setting in CLIP[1]
- Evaluation robustness on 4 ImageNet Variants:
  - ImageNet-V2 - test sets were re-sampled; independent of existing models so less overfitting. [5]
  - ImageNet-A - test set of natural adversarial examples [6]
  - ImageNet-R - collects images of ImageNet categories but with artistic renditions [8]
  - ImageNet-Sketch - black and white sketches [7]

Experiments: Robustness to Distribution Shifts

Baselines:

- CoOp [2]
- CoCoOp [3]
- CLIP-default-prompt: “a photo of a"
- CLIP-ensemble-prompt: ensemble of 80 hand-crafted prompts

Experiments: Robustness to Distribution Shifts

Method	ImageNet Top1 acc. ↑	ImageNet-A Top1 acc. ↑	ImageNet-V2. Top1 acc. ↑	ImageNet-R. Top1 acc. ↑	ImageNet-Sketch Top1 acc. ↑	Average	OOD Average
CLIP-RN50	58.16	21.83	51.41	56.15	33.37	44.18	40.69
Ensemble	59.81	23.24	52.91	**60.72**	35.48	46.43	43.09
CoOp	63.33	23.06	55.40	56.60	34.67	46.61	42.43
CoCoOp	62.81	23.32	55.72	57.74	34.48	46.81	42.82
TPT	60.74	26.67	54.70	59.11	35.09	47.26	43.89
TPT + CoOp	**64.73**	**30.32**	**57.83**	58.99	**35.86**	**49.55**	**45.75**
TPT + CoCoOp	62.93	27.40	56.60	59.88	35.43	48.45	44.83
CLIP-ViT-B/16	66.73	47.87	60.86	73.98	46.09	59.11	57.2
Ensemble	68.34	49.89	61.88	**77.65**	48.24	61.20	59.42
CoOp	71.51	49.71	64.20	75.21	47.99	61.72	59.28
CoCoOp	71.02	50.63	64.07	76.18	48.75	62.13	59.91
TPT	68.98	54.77	63.45	77.06	47.94	62.44	60.81
TPT + CoOp	**73.61**	**57.95**	**66.83**	77.27	**49.29**	**64.99**	**62.83**
TPT + CoCoOp	71.07	58.47	64.85	78.65	48.47	64.30	62.61

Experiments: Robustness to Distribution Shifts

Method	ImageNet Top1 acc. ↑	ImageNet-A Top1 acc. ↑	ImageNet-V2. Top1 acc. ↑	ImageNet-R. Top1 acc. ↑	ImageNet-Sketch Top1 acc. ↑	Average	OOD Average
CLIP-RN50	58.16	21.83	51.41	56.15	33.37	44.18	40.69
Hand-crafted ensemble	59.81	23.24	52.91	**60.72**	35.48	46.43	43.09
CoOp	63.33	23.06	55.40	56.60	34.67	46.61	42.43
CoOp (ensemble 3 seeds)	61.66	22.96	54.14	57.89	34.94	46.32	42.48
CoOp + hand-crafted ensemble	63.60	23.23	55.63	57.07	34.84	46.87	42.69
CoCoOp	62.81	23.32	55.72	57.74	34.48	46.81	42.82
CoCoOp (ensemble 3 seeds)	63.34	24.27	56.12	58.24	35.46	47.49	43.52
CoCoOp + hand-crafted ensemble	63.03	24.16	55.73	57.88	35.22	47.20	43.25
CoCoOp + CoOp	63.86	23.69	56.45	57.7	35.5	47.44	43.34
TPT (ours)	60.74	26.67	54.7	59.11	35.09	47.26	43.89
TPT + CoOp	**64.73**	**30.32**	**57.83**	58.99	**35.86**	**49.55**	**45.75**
TPT + CoCoOp	62.93	27.4	56.6	59.88	35.43	48.45	44.83

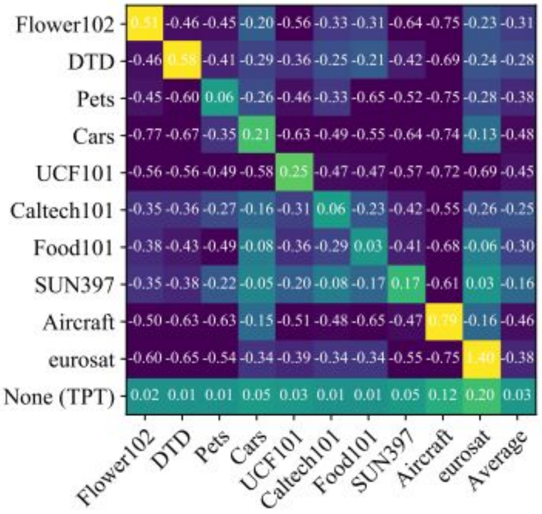
Experiments: Cross-Datasets Generalization

Cross-dataset generalization:

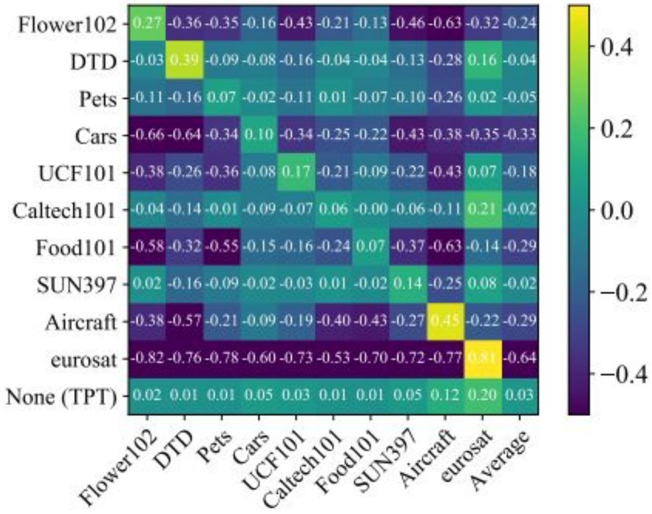
- 10 datasets including plants, animals, scenes, textures etc.
- Two settings:
  - ImageNet as a comprehensive source dataset, fine-tuned datasets for evaluation
  - Fine-tuned datasets are both source and target with no overlaps

Method	Flower102	DTD	Pets	Cars	UCF101	Caltech101	Food101	SUN397	Aircraft	EuroSAT	Average
CLIP-RN50	61.75	40.37	83.57	55.70	58.84	85.88	73.97	58.80	15.66	23.69	55.82
Ensemble	62.77	40.37	82.97	55.89	59.48	**87.26**	74.82	60.85	16.11	25.79	56.63
CoOp	61.55	37.29	87.00	55.32	59.05	86.53	75.59	58.15	15.12	26.20	56.18
CoCoOp	**65.57**	38.53	**88.39**	56.22	57.10	87.38	**76.2**	59.61	14.61	**28.73**	57.23
TPT	62.69	**40.84**	84.49	**58.46**	**60.82**	87.02	74.88	**61.46**	**17.58**	28.33	**57.66**
CLIP-ViT-B/16	67.44	44.27	88.25	65.48	65.13	93.35	83.65	62.59	23.67	42.01	63.58
Ensemble	66.99	45.04	86.92	66.11	65.16	93.55	82.86	65.63	23.22	**50.42**	64.59
CoOp	68.71	41.92	89.14	64.51	66.55	93.70	**85.30**	64.15	18.47	**46.39**	63.88
CoCoOp	**70.85**	45.45	**90.46**	64.90	**68.44**	**93.79**	83.97	**66.89**	22.29	39.23	64.63
TPT	68.98	**47.75**	87.79	**66.87**	68.04	94.16	84.67	65.5	**24.78**	42.44	**65.10**

Experiments: Cross-Datasets Generalization



(a) CoOp with CLIP-RN50.



(b) CoCoOp with CLIP-RN50.

Experiments: Visual Reasoning

Baselines:

- The CNN classifier, trained to map both support and query images to a binary output
- The Meta-baseline regards each sample as a few shot task.
- The transformer-based HOITrans

Experiments: Visual Reasoning

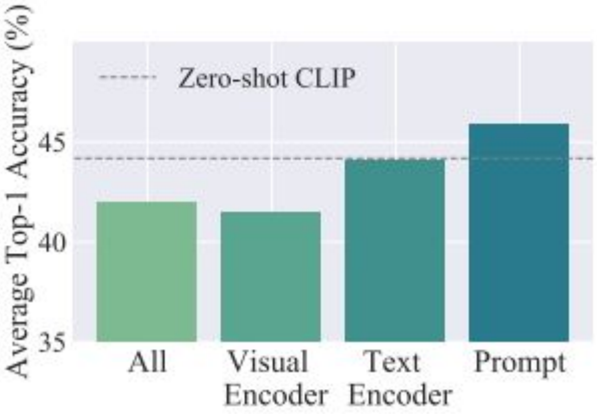
Method	Test Splits				Average
	seen act., seen obj.,	unseen act., seen obj.,	seen act., unseen obj.,	unseen act., unseen obj.,	
CNN-baseline	50.03	49.89	49.77	50.01	49.92
Meta-baseline*	58.82	58.75	58.56	57.04	58.30
HOITrans	59.50	64.38	63.10	62.87	62.46
TPT (w/ CLIP-RN50)	**66.39**	**68.50**	**65.98**	**65.48**	**66.59**

Ablation Study

Test-time optimization

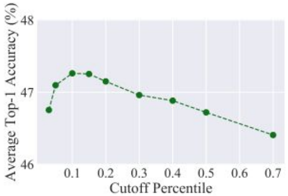
Four different parameter groups (adopt the same setup as MEMO):

- the entire model
- the text encoder
- the visual encoder
- the text prompt



Ablation Study

Method	ImageNet Top1 acc. ↑	ImageNet-A Top1 acc. ↑	ImageNet-V2. Top1 acc. ↑	ImageNet-R. Top1 acc. ↑	ImageNet-Sketch Top1 acc. ↑	Average	OOD Average
CLIP-RN50	58.16	21.83	51.41	56.15	33.37	44.18	40.69
baseline TPT	60.31	23.65	53.66	57.48	34.31	45.88	42.28
+ confidence selection	60.74 (+0.43)	26.67 (+3.02)	54.70 (+1.04)	59.11 (+1.63)	35.09 (+0.78)	47.26 (+1.38)	43.89 (+1.61)

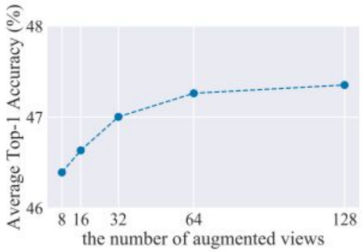


Method	ImageNet Top1 acc. ↑	ImageNet-A Top1 acc. ↑	ImageNet-V2. Top1 acc. ↑	ImageNet-R. Top1 acc. ↑	ImageNet-Sketch Top1 acc. ↑	Average	OOD Average
ResNet-50	76.13	0.00	63.20	36.17	24.09	39.92	30.87
MEMO	77.23	0.75	65.03	41.34	27.72	42.41	33.71
MEMO ($\rho = 0.7$)	77.56	0.92	65.51	41.93	28.20	42.82	34.14
MEMO ($\rho = 0.5$)	**77.72**	1.15	65.77	42.29	**28.55**	43.10	34.44
MEMO ($\rho = 0.3$)	77.57	1.43	**65.85**	42.64	28.33	43.16	34.56
MEMO ($\rho = 0.1$)	77.38	**2.59**	65.37	**42.90**	28.04	**43.26**	**34.72**

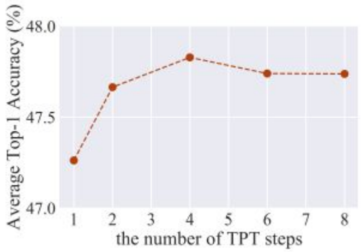
Ablation Study

Analyze two factors that affect TPT's efficiency:

- The number of augmented views
- The number of optimization steps



(a) Different number of augmented views.



(b) Different number of optimization steps.

Strengths

1. The proposed method does not requires additional data or supervision
2. Even without additional pre-training, the model improves the performance
3. The one step of optimization can increase the performance

Weaknesses

1. The most significant gap comes from the ensemble of CoOp/CoCoOp and TPT. However, an ensemble in general brings improvements by itself. How do we validas the TPT?
2. The qualitative study was merely presenting the results. More discussions should be appreciated. (e.g. the confidence selection)
3. The performance of TPT still behind the fine-tuning methods

Future work

One aspect of prompt tuning is, of course, improve the performance and reduce the computational cost.

On the other hand, prompts can mitigate model's bias. This study showed that the proposed method has good generalization ability. Future works can extend on generalization and provide deeper analysis on how prompts eliminates biases

Reference

[1] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In ICML, 2021
[2]Zhou, K., Yang, J., Loy, C.C. *et al*. Learning to Prompt for Vision-Language Models. *Int J Comput Vis* 130, 2337–2348 (2022). https://doi.org/10.1007/s11263-022-01653-1
[3]Zhou, Kaiyang, et al. "Conditional prompt learning for vision-language models." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022.
[4]Zhang, Marvin, Sergey Levine, and Chelsea Finn. "Memo: Test time robustness via adaptation and augmentation." *arXiv preprint arXiv:2110.09506* (2021).
[5]Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, ICML, 2019.
[6]Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. In CVPR, pages 15262–15271, 2021.
[7]Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P Xing. Learning robust global representations by penalizing local predictive power. In NeurIPS, 2019
[8]Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, Dawn Song, Jacob Steinhardt, and Justin Gilmer. The many faces of robustness: A critical analysis of out-of-distribution generalization. In ICCV, 2021.
[9]Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno A. Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. In ICLR, 2021.
[10]Marvin Zhang, Sergey Levine, and Chelsea Finn. MEMO: test time robustness via adaptation and augmentation. CoRR, abs/2110.09506, 2021.
[11]Jia, Chao, et al. "Scaling up visual and vision-language representation learning with noisy text supervision." *International Conference on Machine Learning*. PMLR, 2021.