Consistency-guided Prompt Learning for Vision-Language Models

Shuvendu Roy, Ali Etemad

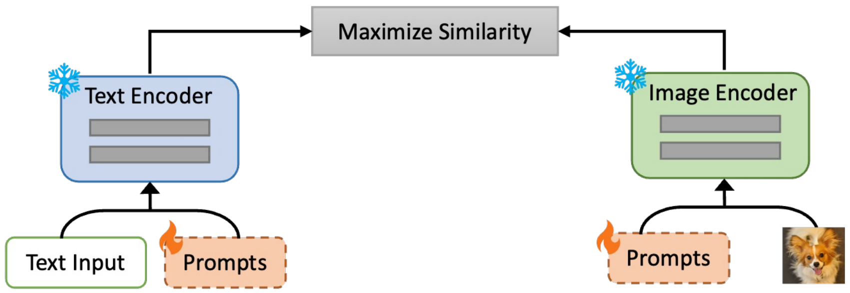
Queen's University

Introduction

- We propose Consistency-guided Prompt learning (CoPrompt)
  - A new fine-tuning method for vision-language models.
  - Improves the generalization of large foundation models when fine-tuned on downstream tasks in a few-shot setting.

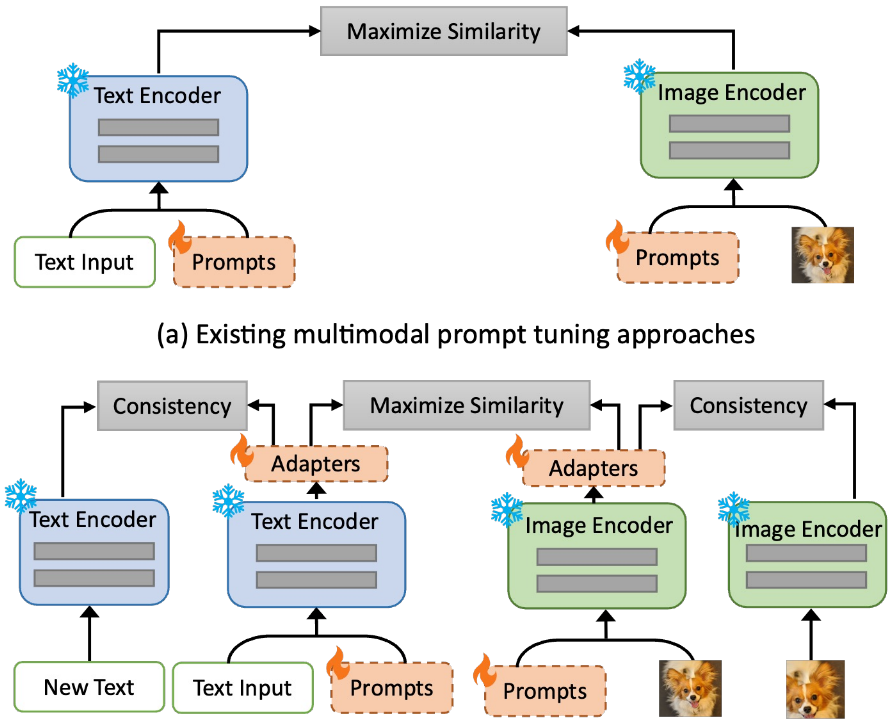
Background

- VL foundation models show excellent generalization on zero-shot task
- Exhibits strong downstream performance when fine-tuned in a few-shot setting
- Often comes at the cost of reduced generalization
- Fine-tuning techniques in existing literature include linear tuning, full fine-tuning, prompt tuning, and adapter tuning.



Introduction

- Enforce a consistency constraint in the prediction of the trainable and pre-trained models to prevent overfitting on the downstream task.
- This facilitates more effective adaptation to downstream tasks in a few-shot learning setting
- Also, improves the zero-shot performance



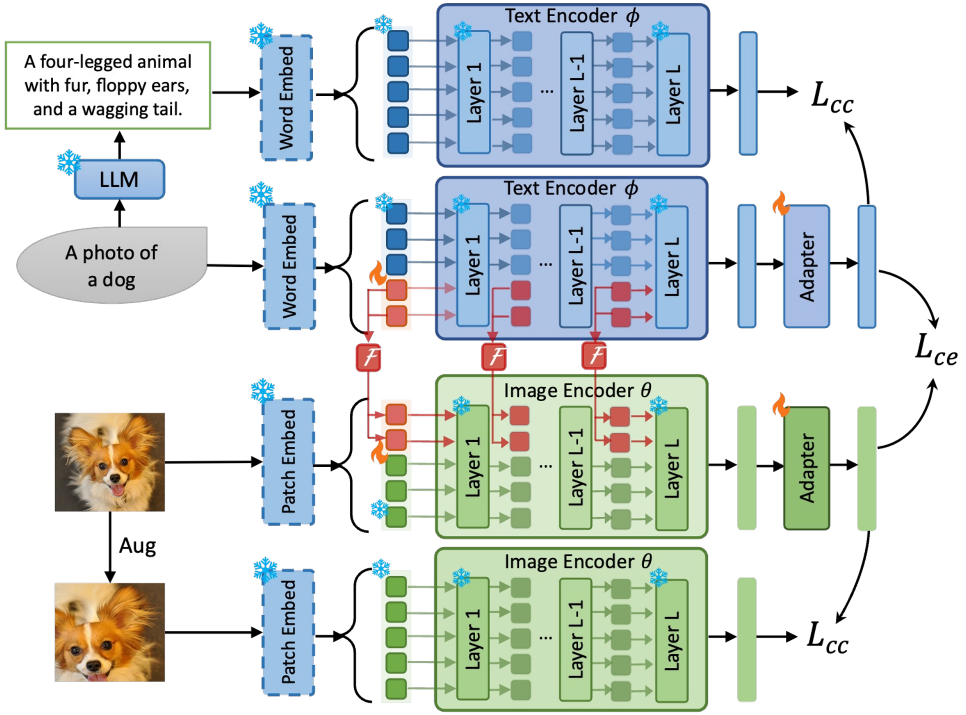
(a) Existing multimodal prompt tuning approaches

(b) Our consistency-guided multimodal prompt tuning (CoPrompt)

Method

- Consistency constraint
  - We use cosine distance as the consistency constraint between the embeddings of the pre-trained encoder and the learnable encoder. This constraint is applied on image and text branches. We can denote the consistency constraint as:

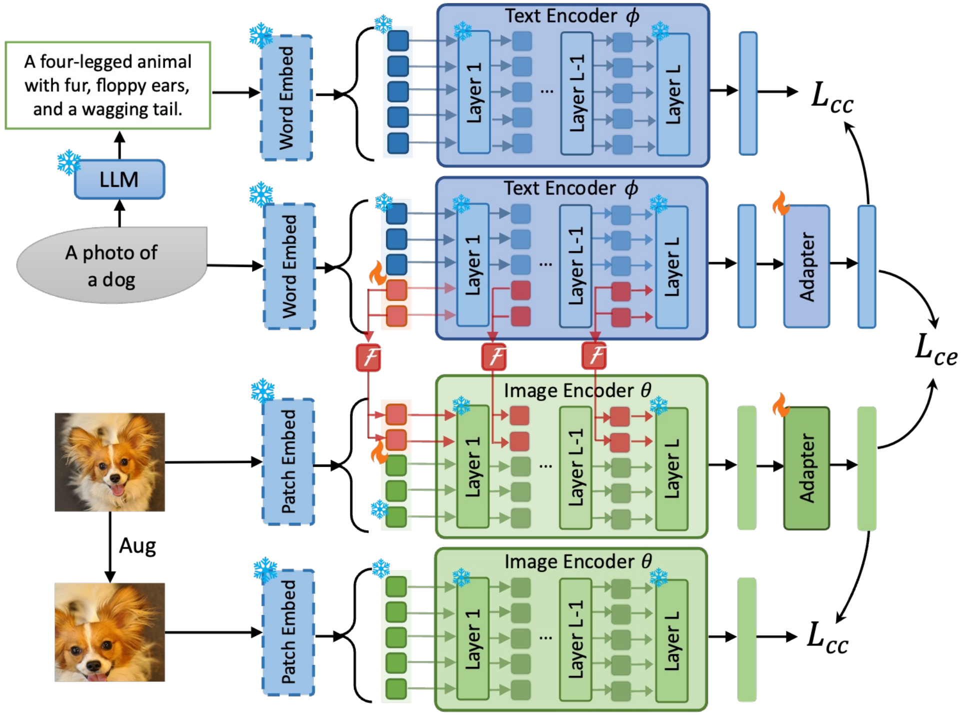
$$\mathcal{L}_{cc} = 2 - \frac{w_y \cdot \phi(t_y)}{||w_y||\ ||\phi(t_y)||} - \frac{z \cdot \theta(i)}{||z||\ ||\theta(i)||}$$



Method

- Input perturbation
  - Given the template text ‘a photo of a [category]’, we use a pre-trained LLM to generate a more descriptive sentence as $s_k = \phi_{GPT}$(‘a photo of a[category]$_k$’).
  - On the image branch, we use an augmentation module to generate perturbed image x' = δ(x).

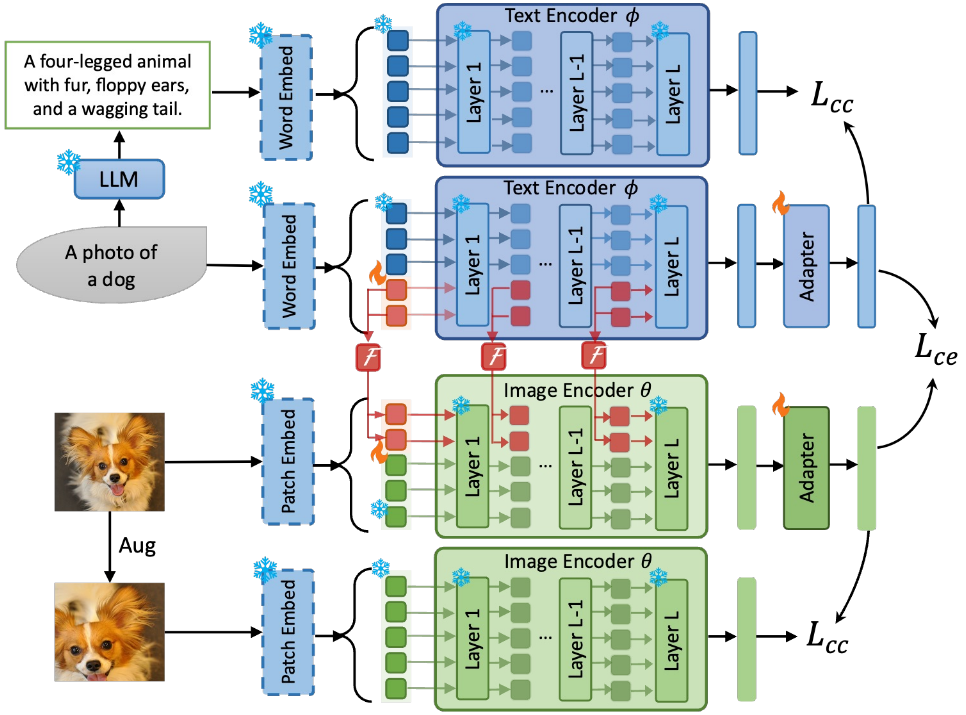
$$\mathcal{L}_{cc} = 2 - \frac{\phi(s_y) \cdot \phi(t_y)}{||\phi(s_y)||\ ||\phi(t_y)||} - \frac{\theta(x') \cdot \theta(i)}{||\theta(x')||\ ||\theta(i)||}$$



Method

- Adapters
  - Adapters are trainable parameters that are added on top of the encoder to transform the embedding vector.
  - Let $\emptyset^a$ be the text adapter that takes a text embedding w_k as input and transforms it as $\emptyset^a(w_k)$.

$$\mathcal{L}_{cc} = 2 - \frac{\phi(s_y)\cdot\phi^a(\phi(t_y))}{||\phi(s_y)||\;||\phi^a(\phi(t_y))||} - \frac{\theta(x')\cdot\theta^a(\theta(i))}{||\theta(x')||\;||\theta^a(\theta(i))||}$$



Queen's University

Experiments

Table 1: Comparison with state-of-the-art methods on base-to-novel generalization.

	Base	Novel	HM
CLIP	69.34	74.22	71.70
CoOp	82.69	63.22	71.66
Co-CoOp	80.47	71.69	75.83
ProGrad	82.48	70.75	76.16
KgCoOp	80.73	73.60	77.00
MaPLe	82.28	75.14	78.55
PromptSRC	**84.26**	76.10	79.97
CoPrompt	84.00	**77.23**	**80.48**

Queen's UNIVERSITY

Experiments

Table 2: Performance of CoPrompt on cross-dataset evaluation and its comparison to existing methods. Here, the model is trained on the ImageNet dataset and evaluated on ten other datasets in a zero-shot setting.

	Source	**Target**										
	ImNet	Caltech	Pets	Cars	Flowers	Food	Aircraft	SUN397	DTD	EuroSAT	UCF	*Ave.*
CoOp	**71.51**	93.70	89.14	64.51	68.71	85.30	18.47	64.15	41.92	46.39	66.55	63.88
Co-CoOp	71.02	94.43	90.14	65.32	71.88	86.06	22.94	67.36	45.73	45.37	68.21	65.74
MaPLe	70.72	93.53	90.49	65.57	72.23	86.20	24.74	67.01	46.49	48.06	68.69	66.30
Bayesian Prompt	70.93	93.67	90.63	65.00	70.90	86.30	**24.93**	67.47	46.10	45.87	68.67	65.95
PromptSRC	71.27	93.60	90.25	**65.70**	70.25	86.15	23.90	67.10	46.87	45.50	68.75	65.81
CoPrompt	70.80	**94.50**	**90.73**	65.67	**72.30**	**86.43**	24.00	**67.57**	**47.07**	**51.90**	**69.73**	**67.00**

Queen's UNIVERSITY

Experiments

Table 3: Performance on domain generalization.

	Source	Target				
	ImNet	ImNetV2	ImNetS	ImNetA	ImNetR	Ave.
CLIP	66.73	60.83	46.15	47.77	73.96	57.17
UPT	**72.63**	64.35	48.66	50.66	76.24	59.98
CoOp	71.51	64.20	47.99	49.71	75.21	59.28
Co-CoOp	71.02	64.07	48.75	50.63	76.18	59.90
ProGrad	72.24	64.73	47.61	49.39	74.58	59.07
KgCoOp	71.20	64.10	48.97	50.69	76.70	60.11
MaPLe	70.72	64.07	49.15	50.90	76.98	60.26
Bayesian Prompt	70.93	64.23	49.20	**51.33**	77.00	60.44
PromptSRC	71.27	**64.35**	**49.55**	50.90	**77.80**	**60.65**
CoPrompt	70.80	64.25	49.43	50.50	77.51	60.42

Queen's UNIVERSITY

Experiments

Table 4: Analysis of different components of CoPrompt.

(a) Cons. modalities.

Consistency	Accuracy
Image only	79.59
Text only	80.02
Both	80.48

(b) Consistency criterion.

Criterion	Accuracy
Cosine	80.48
L1	80.40
MSE	79.33

(c) Text input.

Input	Accuracy
Same Text	80.09
LLM (GPT-2)	80.46
LLM (GPT-3)	80.48

(d) Image input.

Input	Accuracy
Same Image	80.16
Simple Aug.	80.48
Hard Aug.	79.90

(e) Adapter choices.

Adapter	Accuracy
Text only	80.35
Image only	80.10
Both	80.48

(f) No. of Adapter layers.

Layers	Accuracy
Single layer	80.40
2 layers	80.48
3 layers	79.75

Experiments

Table 5: Ablation Study

Cons.	In. Pert.	Adp.	Base	Novel	HM
✓	✓	✓	84.00	77.23	80.48
✓	✓	✗	83.40	76.90	80.02
✓	✗	✓	83.01	76.39	79.56
✓	✗	✗	82.90	76.36	79.50
✗	✗	✓	83.10	74.31	78.45
✗	✗	✗	82.28	75.14	78.55

Queen's UNIVERSITY

Thank you