TCP: Textual-based Class-aware Prompt tuning for Visual-Language Model

Hantao Yao[1], Rui Zhang[2], Changsheng Xu [1,3]

1State Key Laboratory of Multimodal Artificial Intelligence Systems, Institute of Automation, CAS

2State Key Lab of Processors, Institute of Computing Technology, CAS;

3 University of Chinese Academy of Sciences(CAS),

{hantao.yao,csxu}@nlpr.ia.ac.cn;zhangrui@ict.ac.cn

Summary

- **Prompt Tuning** has been proposed to adapt the pretrained VLM to downstream tasks, achieving a fantastic performance on various few-shot or zero-shot visual recognization task.

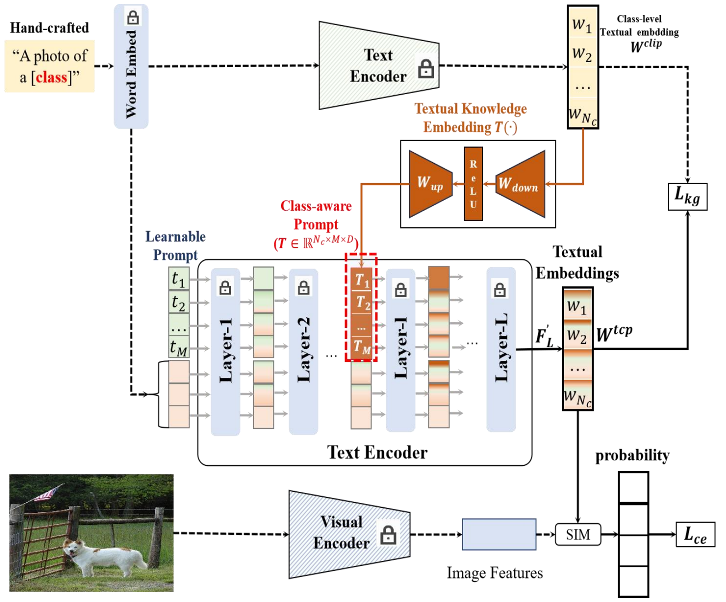
- **Existing methods' shortcoming**:
  - **Domain-shared Prompt Tuning** are derived from labeled training image, their performance is suboptimal for unseen classes and images.
  - **Image-conditional prompt tuning** has a limited ability to reduce the domain shifts at the class-level.

- **Main insight**: The textual embedding generated by the forzen CLIP contains the essential class-level knowledge, which can be injected into the learnable prompt for generating the class-aware prompt.

- **Method**: An Textual Knowledge Emebedding $\mathcal{T}(\cdot)$ is proposed to project the hand-craft textual-based class-level embedding $\boldsymbol{W}^{clip}$ in into the **Class-aware Prompt** $\mathbf{T} \in \mathbb{R}^{N_c \times M \times D}$.

- $\mathbf{T} = [\mathbf{T}_1, \mathbf{T}_2, ..., \mathbf{T}_M]$ is inserted into $\mathbf{F}_l$ for generating the enhanced tokens $\mathbf{F}_l'$,

$$\mathbf{F}_l' = [\mathbf{T}_1, \mathbf{T}_2, ..., \mathbf{T}_M, \mathbf{F}_{l,M+1}, \mathbf{F}_{l,M+1}, ..., \mathbf{F}_{l,N_t}]$$



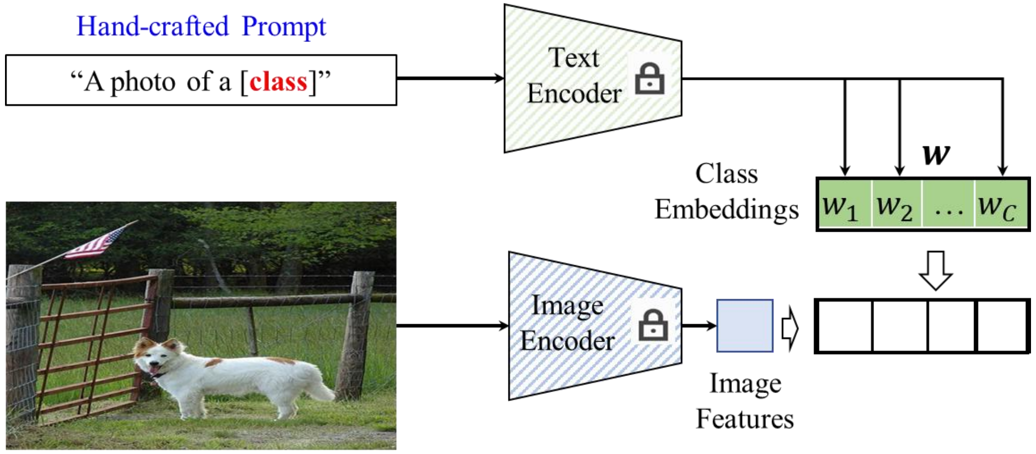
- **Generalization of** TKE**:** lower distance, higher performance.

Methods	*Base*	*New*	*H*
CoOp [50]	82.63	67.99	74.6
CoOp+TKE	83.10(↑ 0.47)	70.17(↑ 2.18)	76.09 (↑ 1.49)
KgCoOp [43]	80.73	73.6	77.00
KgCoOp +TKE	84.13 (↑ 3.40)	75.36(↑ 1.76)	79.51 (↑ 2.51)
ProGrad [51]	82.48	70.75	76.16
ProGrad+TKE	82.61 (↑ 0.13)	72.91(↑ 2.16)	77.46 (↑ 1.30)
PromptSRC [19]	82.07	74.83	78.03
PromptSRC+TKE	83.74 (↑ 1.67)	75.85(↑ 1.02)	79.60 (↑ 1.57)
DAPT [5]	83.18	69.27	75.59
DAPT+TKE	84.15 (↑ 0.97)	74.87(↑ 5.60)	79.24 (↑ 3.65)

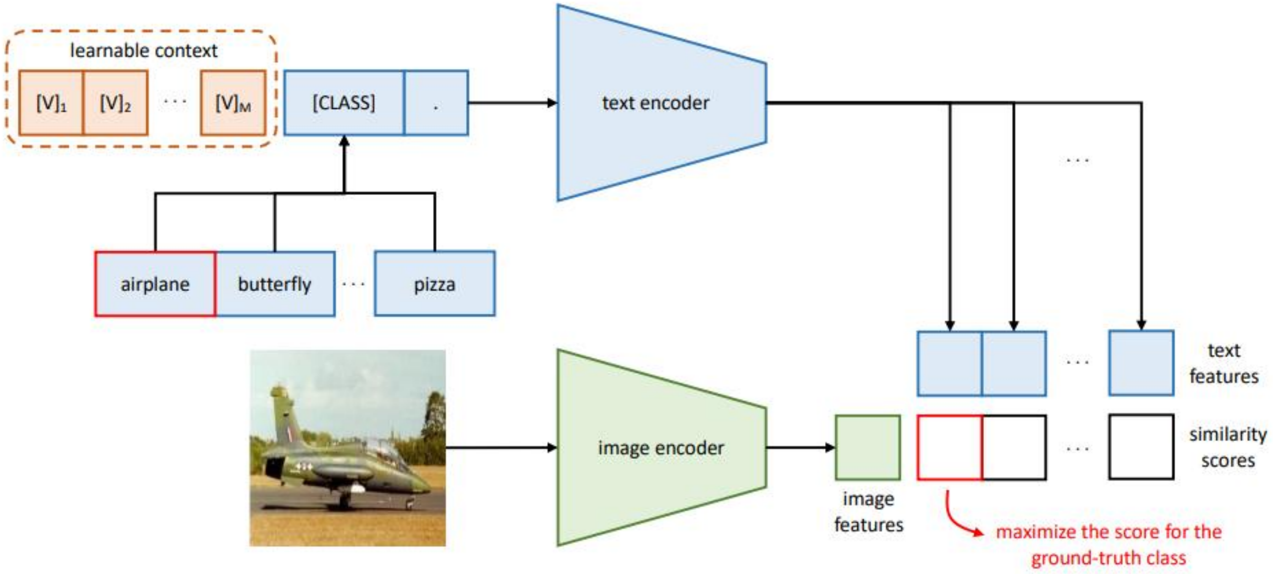
Prompt Tuning

- **Prompt Tuning** has been proposed to adapt the pretrained VLM to downstream tasks, achieving a fantastic performance on various few-shot or zero-shot visual recognization task.

- CLIP uses a **hand-crafted prompts** to model the textual-based class embedding for zero-shot prediction.



- Context Optimization(CoOp) aims to model a prompt's context using **a set of learnable vectors**.

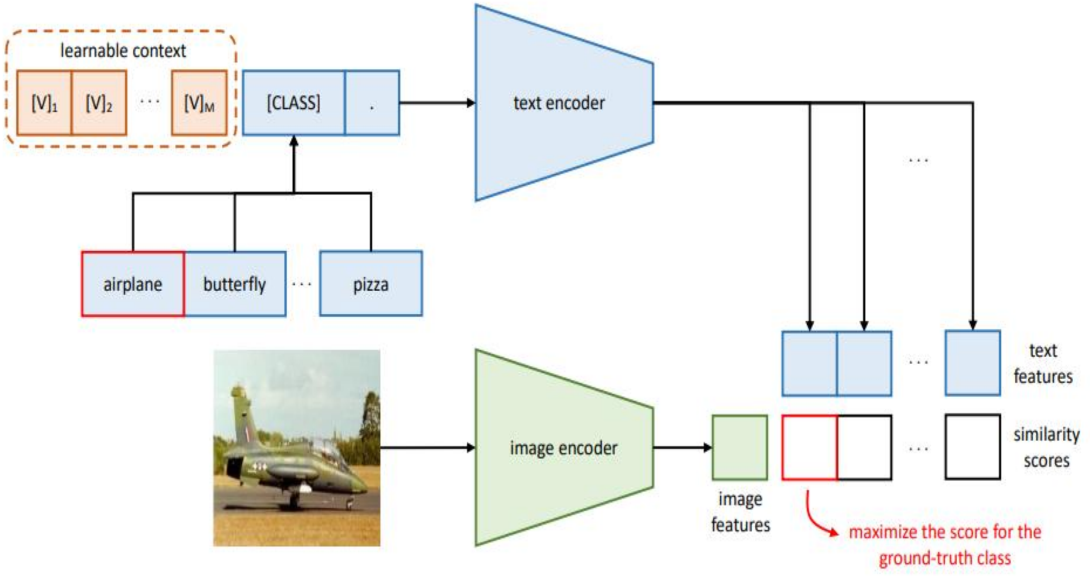


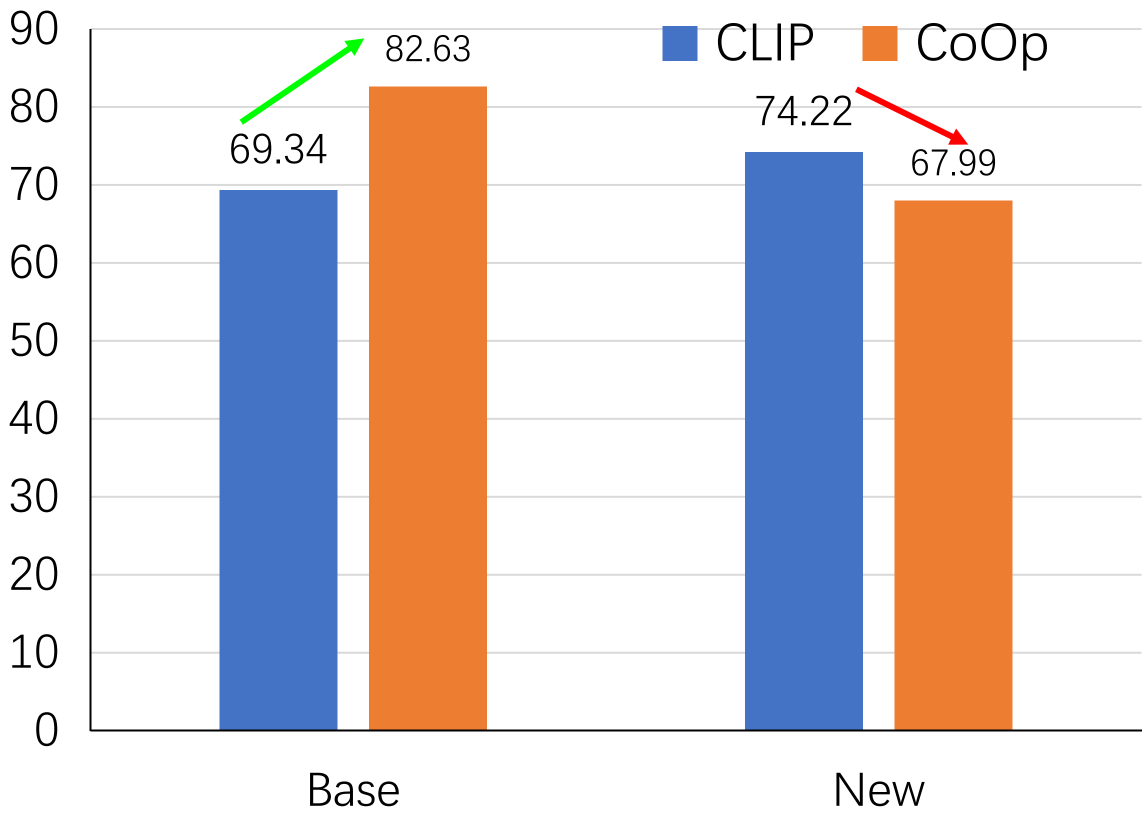
Overview of Context Optimization(CoOp)[1]

1 Image comes from "Learning to Prompt for Vision-Language Models"

Context Optimization(CoOp)

- Context Optimization(CoOp) aims to model a prompt's context using a set of learnable vectors.
- CoOp is overfitted on the trained seen domain(Base), **leading a worse generalization on the unseen domain(New)**.





CoOp-based Methods

- **Existing methods can be divided into:**
  - Domain-shared prompt tuning: a unique learnable prompt for both seen and unseen domains
  - Image-Conditional prompt tuning: injecting the image's knowledge into each prompt.

Domain-shared Prompt Tuning

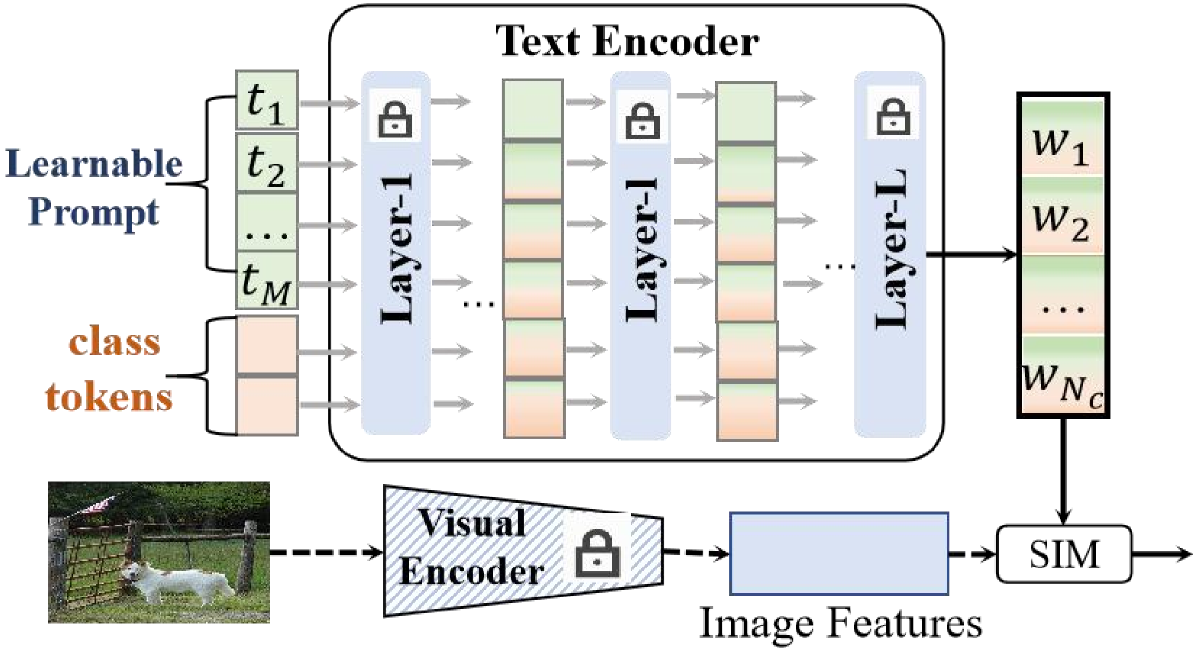
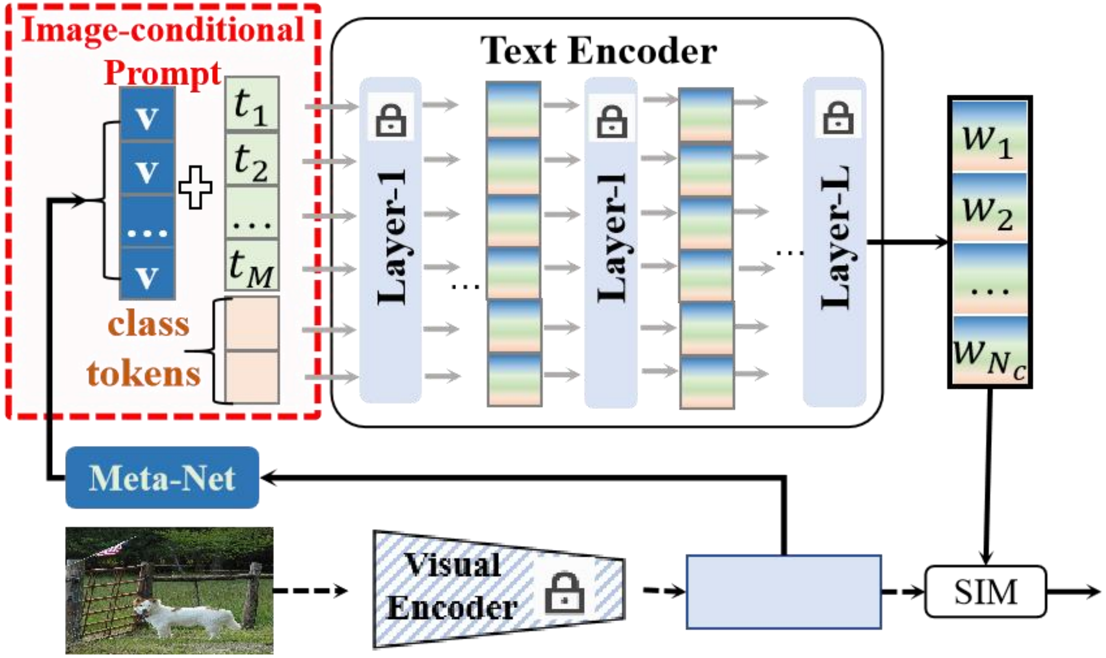


Image-Conditional Prompt Tuning



CoOp-based Methods

- **Existing methods can be divided into:**
  - Domain-shared prompt tuning: a unique learnable prompt for both seen and unseen domains
  - Image-Conditional prompt tuning: injecting the image's knowledge into each prompt.

Domain-shared Prompt Tuning

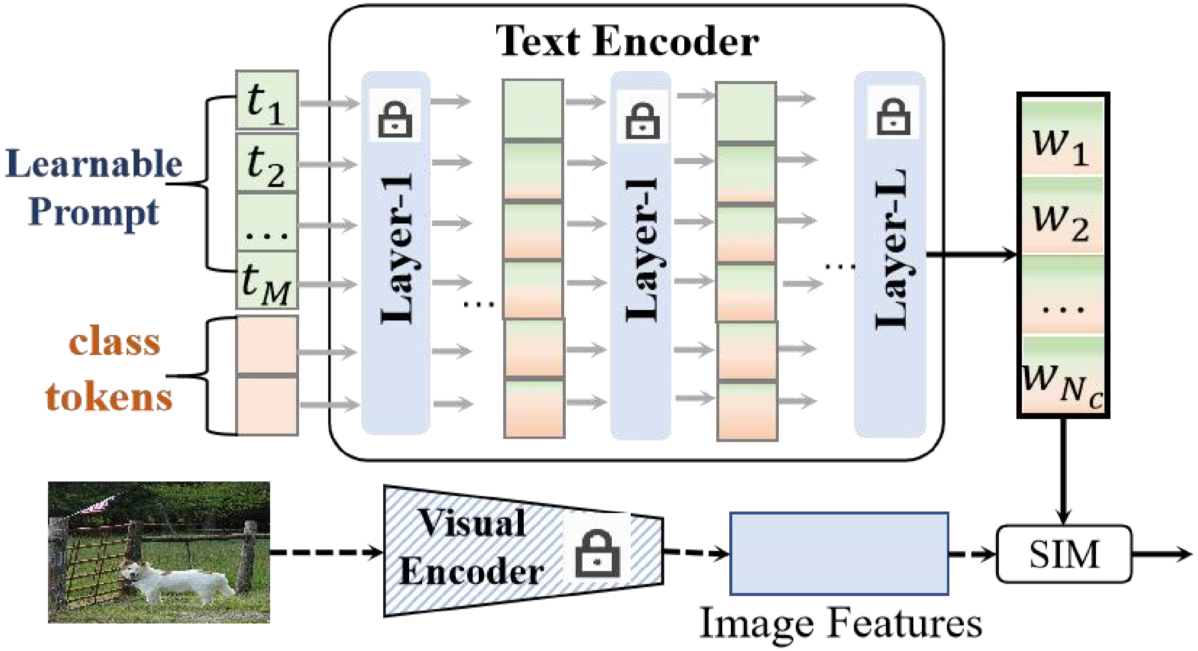
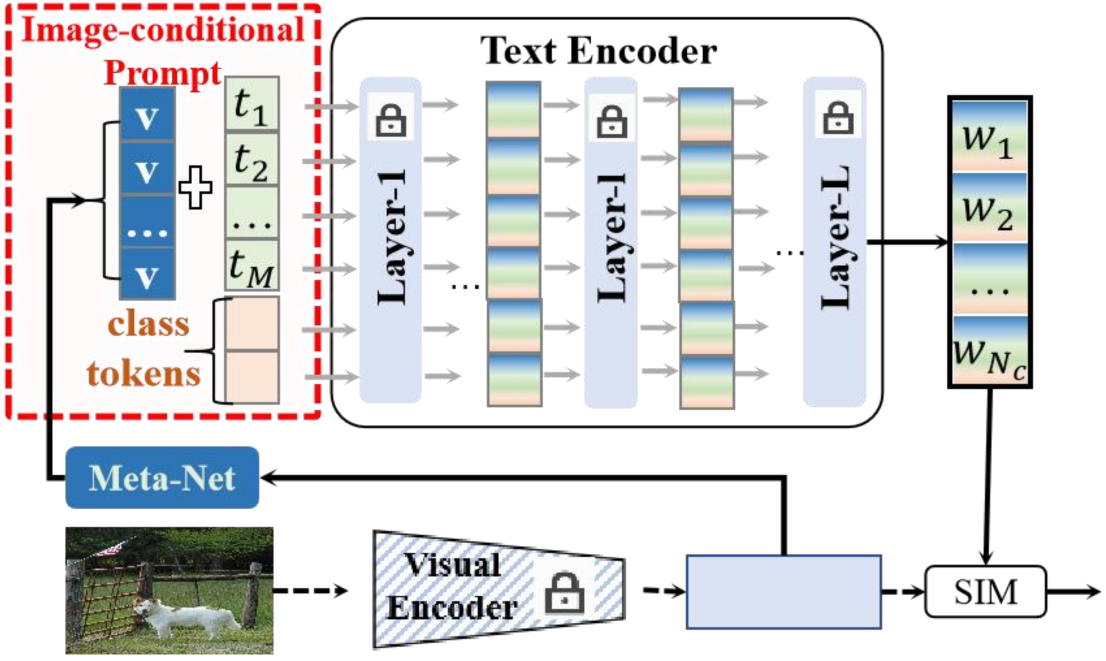


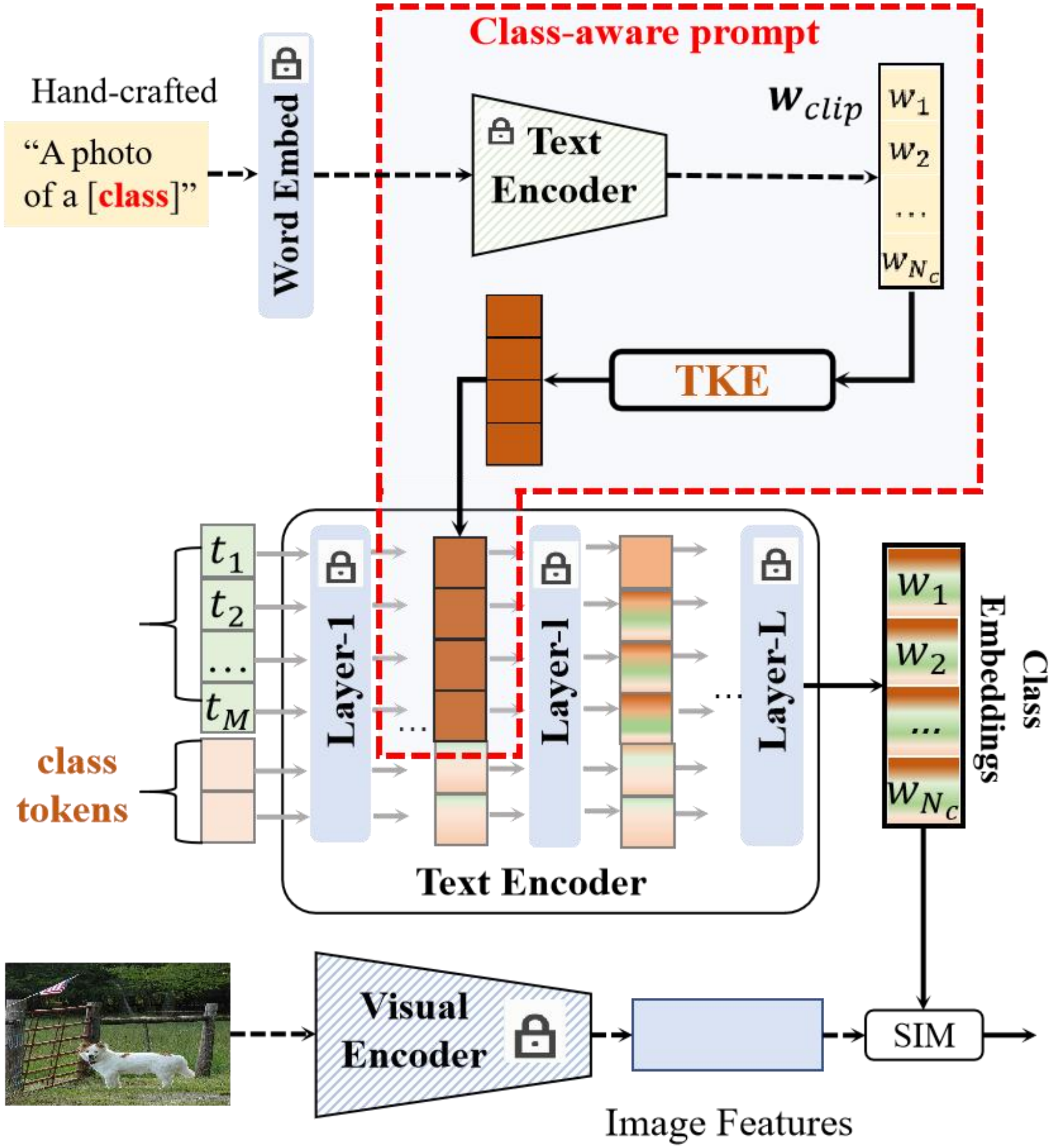
Image-Conditional Prompt Tuning



They all have a limted ability to boost class-level discriminative ability, especially for the unseen domain.

Main Insight

- The textual embedding generated by the forzen CLIP contains the essential class-level knowledge, which can be injected into the learnable prompt for generating the class-aware prompt.
  - Improving discriminative of the class-level textual classifier on the seen domain
  - Improving generalization on the unseen domain with considering the prior knowledge of unseen domain.



Textual-based Class-aware Prompt tuning

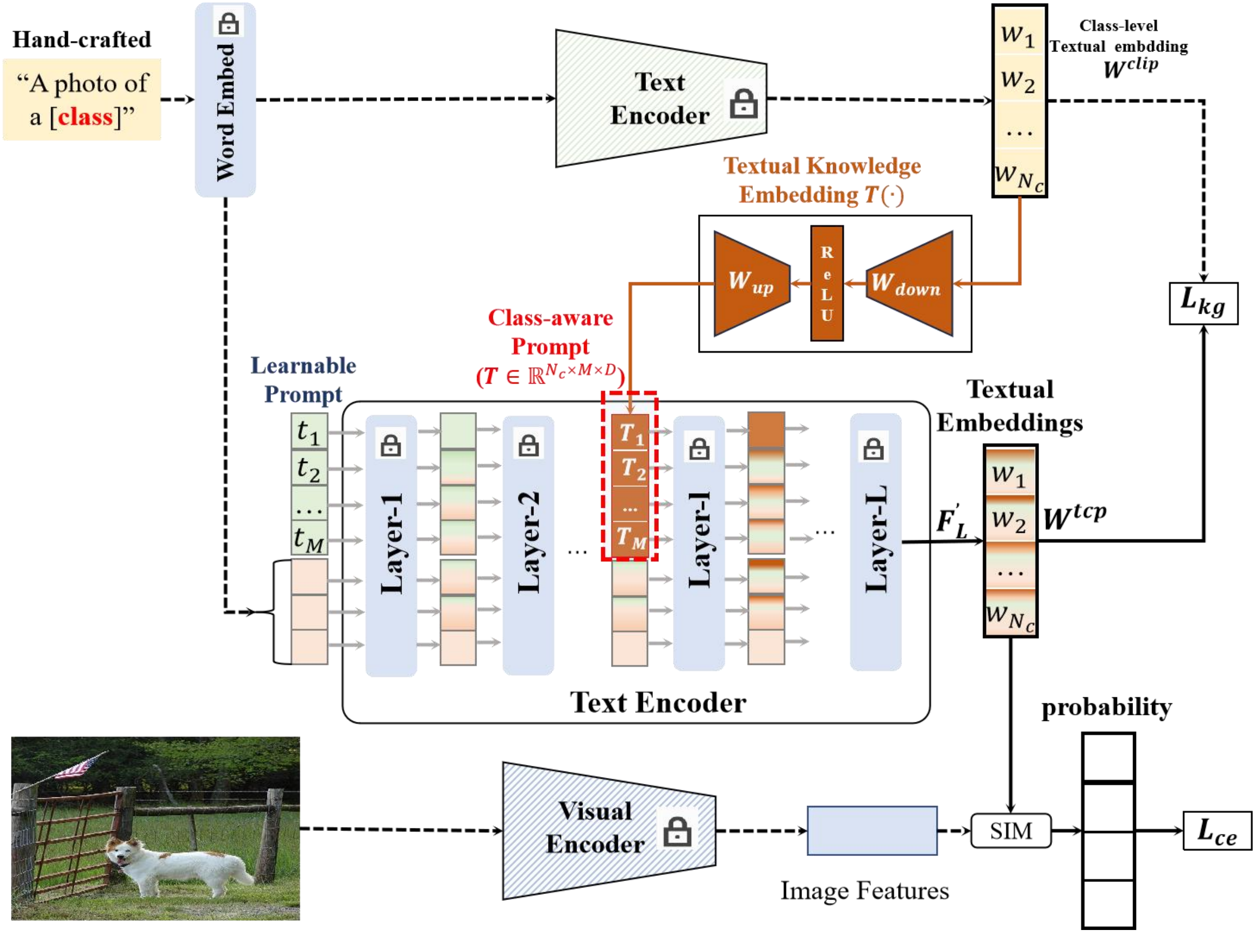
- Based on the standard KgCoOp method, an Textual Knowledge Emebedding $\mathcal{T}(\cdot)$ is proposed to project the hand-craft textual-based class-level embedding $\boldsymbol{W}^{clip}$ in into the **Class-aware Prompt T ∈ $\mathbb{R}^{N_c\times M\times D}$**

- Aftering obtaining the middle tokens $\mathbf{F}_l$ of l-th encoder layer, the class-aware prompt $\mathbf{T}=[\mathbf{T}_1,\mathbf{T}_2,...,\mathbf{T}_M]$ is inserted into $\mathbf{F}_l$ for generating the enhanced tokens $\mathbf{F}_l'$,

$$\mathbf{F}_l' = [\mathbf{T}_1,\mathbf{T}_2,...,\mathbf{T}_M,\mathbf{F}_{l,M+1},\mathbf{F}_{l,M+1},...,\mathbf{F}_{l,N_t}]$$

- $\mathbf{F}_l'$ is fed into the following layers for generating the textual embedding,

$$\mathbf{F}_i' = \theta_i(\mathbf{F}_{i-1}'), i \in [l+1, L]$$



Experiment

- **Generalization of** TKE**:**
  - Textual Knowledge Embedding(TKE) is a **plug-and-play module** that can quickly insert existing prompt tuning methods to improve their performance further.

Methods	*Base*	*New*	*H*
CoOp [50]	82.63	67.99	74.6
CoOp+TKE	83.10(↑ 0.47)	70.17(↑ 2.18)	76.09 (↑ 1.49)
KgCoOp [43]	80.73	73.6	77.00
KgCoOp +TKE	84.13 (↑ 3.40)	75.36(↑ 1.76)	79.51 (↑ 2.51)
ProGrad [51]	82.48	70.75	76.16
ProGrad+TKE	82.61 (↑ 0.13)	72.91(↑ 2.16)	77.46 (↑ 1.30)
PromptSRC [19]	82.07	74.83	78.03
PromptSRC+TKE	83.74 (↑ 1.67)	75.85(↑ 1.02)	79.60 (↑ 1.57)
DAPT [5]	83.18	69.27	75.59
DAPT+TKE	84.15 (↑ 0.97)	74.87(↑ 5.60)	79.24 (↑ 3.65)

Experiment

- **Effectiveness of templates:**

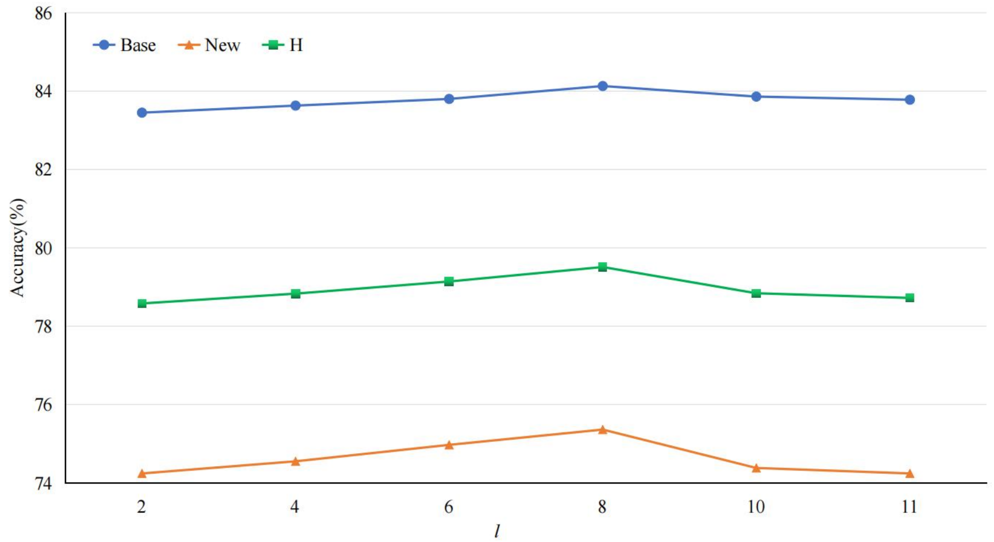
Templates	Base	New	H
'X X X X '	**84.13**	**75.36**	**79.51**
'a photo of a '	83.94	74.94	79.18
'this is a picture '	84.00	74.66	79.06

- Domain-shared prompt vs Class-aware prompt:

Baseline	Domain-Shared	Class-Aware	Base	New	H
√	√		80.73	73.6	77
√		√	84.05	75.18	79.36
√	√	√	**84.13**	**75.36**	**79.51**

Experiment

- Effect of insert layer:



- Comparison with single-layer vs multiple layers prompt tuning:

Modes	Layers	Base	New	H
TCP-Shallow	{8}	84.13	**75.36**	**79.51**
TCP-Deeper	{4;8}	84.02	75.13	79.33
TCP-Deeper	{8;10}	**84.24**	75.29	79.51
TCP-Deeper	{4;8;10}	84.11	74.51	79.02
TCP-Deeper	{4;6;8;10}	84.17	74.66	79.13

Experiment

■ Effectiveness of TCP: *Base-to-new setting*

Table 1. Comparison on the base-to-new generalization setting with 16-shot samples.'tp','dtp','vp',and 'dvp' denote the 'textual prompt', 'deep textual prompt','visual prompt', and 'deep visual prompt', respectively. PromptSRC are based on deep visual-textual prompt tuning('dvp+dtp'). '*' denote the performance obtained by our re-implementation.

Datasets	Sets	CoOp* (IJCV22)	CoCoOp (CVPR22)	DAPT* (ICCV23)	ProGrad* (ICCV23)	ProDA (CVPR22)	KgCoOp (CVPR23)	RPO (ICCV23)	PLOT* (ICLR23)	LFA (ICCV23)	MaPLe (CVPR23)	DePT (Arxiv23)	PromptSRC* (ICCV23)	TCP
		tp	tp	tp+vp	tp	tp	tp	dtp+dvp	tp+vp	–	dtp+dvp	tp	dtp+dvp	tp
Average	Base	82.38	80.47	83.18	82.48	81.56	80.73	81.13	83.98	83.62	82.28	83.62	84.12	**84.13**
	New	67.96	71.69	69.27	70.75	72.30	73.6	75.00	71.72	74.56	75.14	75.04	75.02	**75.36**
	H	74.48	75.83	75.59	76.16	76.65	77.0	77.78	77.37	78.83	78.55	79.10	79.31	**79.51**
ImageNet	Base	76.46	75.98	76.83	77.02	75.40	75.83	76.60	77.30	76.89	76.66	77.03	**77.75**	77.27
	New	66.31	70.43	69.27	66.66	70.23	69.96	**71.57**	69.87	69.36	70.54	70.13	70.70	69.87
	H	71.02	73.10	72.85	71.46	72.72	72.78	74.00	73.40	72.93	73.47	73.42	**74.06**	73.38
Caltech101	Base	97.80	97.96	97.83	98.02	98.27	97.72	97.97	**98.53**	98.41	97.74	98.30	98.13	98.23
	New	93.27	93.81	93.07	93.89	93.23	94.39	94.37	92.80	93.93	94.36	94.60	93.90	**94.67**
	H	95.48	95.84	95.39	95.91	95.68	96.03	96.03	95.58	96.13	96.02	96.41	95.97	**96.42**
OxfordPets	Base	94.47	95.20	95.00	95.07	95.43	94.65	94.63	94.50	95.13	95.43	94.33	**95.50**	94.67
	New	96.00	97.69	95.83	97.63	**97.83**	97.76	97.50	96.83	96.23	97.76	97.23	97.40	97.20
	H	95.23	96.43	95.41	96.33	**96.62**	96.18	96.05	95.65	95.68	96.58	95.76	96.44	95.92
Cars	Base	75.67	70.49	75.80	77.68	74.70	71.76	73.87	79.07	76.32	72.94	79.13	78.40	**80.80**
	New	67.53	73.59	63.93	68.63	71.20	75.04	**75.53**	74.80	74.88	74.00	75.47	74.73	74.13
	H	71.37	72.01	69.36	72.88	72.91	73.36	74.69	76.88	75.59	73.47	77.26	75.52	**77.32**
Flowers	Base	97.27	94.87	96.97	95.54	97.70	95.00	94.13	97.93	97.34	95.92	**98.00**	97.90	97.73
	New	67.13	71.75	60.90	71.87	68.68	74.73	76.67	73.53	75.44	72.46	76.37	**76.77**	75.57
	H	79.44	81.71	74.81	82.03	80.66	83.65	84.50	83.99	85.00	82.56	85.84	**86.06**	85.23
Food101	Base	89.37	**90.70**	90.37	90.37	90.30	90.5	90.33	89.80	90.52	90.71	90.50	90.63	90.57
	New	88.77	91.29	91.30	89.59	88.57	91.7	90.83	91.37	91.48	**92.05**	91.60	91.50	91.37
	H	89.07	90.99	90.83	89.98	89.43	91.09	90.58	90.58	91.00	**91.38**	91.05	91.06	90.97
Aircraft	Base	39.67	33.41	39.97	40.54	36.90	36.21	37.33	42.13	41.48	37.44	**43.20**	42.30	41.97
	New	31.23	23.71	29.80	27.57	34.13	33.55	34.20	33.73	32.29	35.61	34.83	**36.97**	34.43
	H	34.95	27.74	34.14	32.82	35.46	34.83	35.70	37.46	36.31	36.50	38.57	**39.46**	37.83
SUN397	Base	80.85	79.74	80.97	81.26	78.67	80.29	80.60	82.20	82.13	80.82	82.33	**82.83**	82.63
	New	68.34	76.86	76.97	74.17	76.93	76.53	77.80	73.63	77.20	78.70	77.80	**79.00**	78.20
	H	74.07	78.27	78.92	77.55	77.79	78.36	79.18	77.68	79.59	79.75	80.00	**80.87**	80.35
DTD	Base	79.97	77.01	82.23	77.35	80.67	77.55	76.70	81.97	81.29	80.36	82.20	82.60	**82.77**
	New	48.60	56.00	54.23	52.35	56.48	54.99	62.13	43.80	60.63	**59.18**	59.13	57.50	58.07
	H	60.46	64.85	65.36	62.45	66.44	64.35	68.61	57.09	**69.46**	68.16	68.78	67.80	68.25
EuroSAT	Base	90.10	87.49	**94.73**	90.11	83.90	85.64	86.63	93.70	93.40	94.07	89.03	92.40	91.63
	New	53.00	60.04	50.33	60.89	66.00	64.34	68.97	62.67	71.24	73.23	71.07	68.43	**74.73**
	H	66.74	71.21	65.74	72.67	73.88	73.48	76.79	75.11	80.83	82.3	79.04	78.63	**82.32**
UCF101	Base	84.53	82.33	84.30	84.33	85.23	82.89	83.67	86.60	86.97	83.00	85.80	86.93	**87.13**
	New	67.37	73.45	76.33	74.94	71.97	76.67	75.43	75.90	77.48	78.66	77.23	78.33	**80.77**
	H	74.98	77.67	80.12	79.35	78.04	79.65	79.34	80.90	81.95	80.77	81.29	82.41	**83.83**

Experiment

■ **Effectiveness of TCP:** *Cross-Dataset Generalization*

Table 2. Comparison of cross-dataset evaluation.'tp','dtp','vp',and 'dvp' denote the 'textual prompt', 'deep textual prompt','visual prompt', and 'deep visual prompt', respectively. Note that DAPT and MaPLe are based on visual-textual prompt tuning('vp+tp').

Datasets	CLIP	CoOp	ProGrad	KgCoOp	DePT	VPT	PLOT	PromptSRC	MaPLe	DAPT	**TCP**
		tp	tp	tp	tp	tp+vp	tp+vp	dtp+tvp	dtp+dvp	tp+vp	tp
ImageNet	66.70	71.51	72.24	70.66	**72.77**	69.73	71.60	71.27	70.72	71.60	71.40
Caltech101	93.30	93.70	91.52	93.92	**94.23**	93.67	92.07	93.60	93.53	93.50	93.97
OxfordPets	89.10	89.14	89.64	89.83	90.03	89.27	90.10	90.25	90.49	90.67	**91.25**
StandfordCars	65.70	64.51	62.39	65.41	65.57	65.5	65.70	65.70	65.57	**65.93**	64.69
Flowers	70.70	68.71	67.87	70.01	70.57	70.2	69.23	70.25	72.20	71.70	**71.21**
Food101	85.90	85.30	85.40	86.36	86.37	86.27	86.23	86.15	86.20	86.10	**86.69**
FGVCAircraft	24.90	18.47	20.16	22.51	23.27	22.13	**25.00**	23.90	24.74	23.03	23.45
SUN397	62.60	64.15	62.47	66.16	66.67	66.57	61.67	67.10	67.01	67.00	**67.15**
DTD	44.30	41.92	39.42	46.35	45.97	46.93	38.60	**46.87**	46.49	44/00	44.35
EuroSAT	48.30	46.39	43.46	46.04	43.53	47.43	47.83	45.50	48.06	**52.47**	51.45
UCF101	67.60	66.55	64.29	68.50	**69.30**	67.20	67.00	68.75	68.69	68.73	68.73
Avg.	65.24	63.88	62.71	65.51	65.55	65.52	64.34	65.81	66.30	**66.31**	66.29

Experiment

■ **Effectiveness of TCP:** *Few-shot Learning(K=4)*

Table 4. Comparison of few-shot learning with 4-shot samples.

	CLIP	CoOp	CoCoOp	ProGrad	KgCoOp	MaPLe	TIP-Adapter-F	DAPT	PromptSRC	PLOT	TaskRes	**TCP**
ImageNet	66.70	69.37	70.55	70.21	70.19	70.67	70.78	70.80	**70.80**	70.40	62.87	70.48
Caltech101	93.30	94.44	94.98	94.93	94.65	94.30	94.77	94.23	94.77	**95.13**	94.67	95.00
OxfordPets	89.10	91.30	**93.01**	93.21	93.20	92.05	92.26	92.17	93.23	92.55	92.00	91.90
StandfordCars	65.70	72.73	69.10	71.75	71.98	68.70	74.42	74.40	71.83	74.93	75.90	**76.30**
Flowers	70.70	91.14	82.56	89.98	90.69	80.80	92.98	92.37	91.31	92.93	91.50	**94.40**
Food101	85.90	82.58	86.64	85.77	86.59	**86.90**	86.18	83.60	86.06	86.46	86.03	85.3
FGVCAircraft	24.90	33.18	30.87	32.93	32.47	29.03	35.49	32.47	32.80	35.29	33.80	**36.20**
SUN397	62.60	70.13	70.50	71.17	71.79	71.47	70.65	72.20	**72.80**	70.42	72.70	72.11
DTD	44.30	58.57	54.79	57.72	58.31	54.73	61.70	61.37	60.64	62.43	59.57	**63.97**
EuroSAT	48.30	68.62	63.83	70.84	71.06	54.87	78.27	72.73	75.02	**80.70**	72.87	77.43
UCF101	67.60	77.41	74.99	77.82	78.40	73.70	79.73	79.40	79.35	79.76	76.10	**80.83**
Avg.	65.37	73.59	71.98	74.21	74.48	70.66	76.11	75.07	75.33	76.45	74.36	**76.72**

Conclusion

- An effectively Textual-based Class-aware Prompt tuning is proposed by injecting the textual class-aware prompts generated by Textual Knowledge Embedding(TKE) into the Text Encoder.
- We demonstrate that explicitly incorporating the prior knowledge of each class into the learnable prompt tokens can enhance the discriminative of the class distribution.
- https://github.com/htyao89 /Textual-based_Class-aware_prompt_tuning

Methods	Prompts	Accuracy			Training-time
		Base	New	H	
CLIP	Hand-crafted	69.34	**74.22**	71.70	-
CoOp	Textual	82.63	**67.99**	74.60	6ms/image
ProGrad	Textual	82.48	**70.75**	76.16	22ms/image
CoCoOp	Textual+visual	80.47	**71.69**	75.83	160ms/image

TCP: Textual-based Class-aware Prompt tuning for Visual-Language Model

Hantao Yao[1], Rui Zhang[2], Changsheng Xu [1,3]

1State Key Laboratory of Multimodal Artificial Intelligence Systems, Institute of Automation, CAS

2State Key Lab of Processors, Institute of Computing Technology, CAS;

3 University of Chinese Academy of Sciences(CAS),

{hantao.yao,csxu}@nlpr.ia.ac.cn;zhangrui@ict.ac.cn