



视觉提示学习研究进展

Vision Prompt Tuning Learning

分享人：高逸凡

2023.06.12

目录

2

- 作者介绍
- 研究动机
- 本文方法
- 实验效果
- 总结反思

- 作者介绍
- 研究背景
- 研究动机
- 本文方法
- 实验效果
- 总结反思

研究背景



4

□ Prompt Learning

⑤ NLP中的四大范式

Paradigm	Engineering	Task Relation
a. Fully Supervised Learning (Non-Neural Network)	Features (e.g. word identity, part-of-speech, sentence length)	<p>CLS TAG LM</p> <p>GEN</p>
b. Fully Supervised Learning (Neural Network)	Architecture (e.g. convolutional, recurrent, self-attentional)	<p>CLS TAG LM</p> <p>GEN</p>
c. Pre-train, Fine-tune	Objective (e.g. masked language modeling, next sentence prediction)	<p>CLS TAG LM</p> <p>GEN</p>
d. Pre-train, Prompt, Predict	Prompt (e.g. cloze, prefix)	<p>CLS TAG LM</p> <p>GEN</p>



研究背景

5

□ Prompt Learning

◎ 什么是Prompt Learning?

Name	Notation	Example	Description
<i>Input</i>	x	I love this movie.	One or multiple texts
<i>Output</i>	y	++ (very positive)	Output label or text
<i>Prompting Function</i>	$f_{\text{prompt}}(x)$	[X] Overall, it was a [Z] movie.	A function that converts the input into a specific form by inserting the input x and adding a slot [Z] where answer z may be filled later.
<i>Prompt</i>	x'	I love this movie. Overall, it was a [Z] movie.	A text where [X] is instantiated by input x but answer slot [Z] is not.
<i>Filled Prompt</i>	$f_{\text{fill}}(x', z)$	I love this movie. Overall, it was a bad movie.	A prompt where slot [Z] is filled with any answer.
<i>Answered Prompt</i>	$f_{\text{fill}}(x', z^*)$	I love this movie. Overall, it was a good movie.	A prompt where slot [Z] is filled with a true answer.
<i>Answer</i>	z	“good”, “fantastic”, “boring”	A token, phrase, or sentence that fills [Z]



研究背景

6

□ Prompt Learning

◎ NLP中Prompt用于不同的下游任务

Type	Task	Input ($[X]$)	Template	Answer ($[Z]$)
Text CLS	Sentiment	I love this movie.	[X] The movie is [Z].	great fantastic ...
	Topics	He prompted the LM.	[X] The text is about [Z].	sports science ...
	Intention	What is taxi fare to Denver?	[X] The question is about [Z].	quantity city ...
Text-span CLS	Aspect Sentiment	Poor service but good food.	[X] What about service? [Z].	Bad Terrible ...
	NLI	[X1]: An old man with ... [X2]: A man walks ...	[X1]? [Z], [X2]	Yes No ...
Tagging	NER	[X1]: Mike went to Paris. [X2]: Paris	[X1] [X2] is a [Z] entity.	organization location ...
Text Generation	Summarization	Las Vegas police ...	[X] TL;DR: [Z]	The victim ... A woman
	Translation	Je vous aime.	French: [X] English: [Z]	I love you. I fancy you. ...



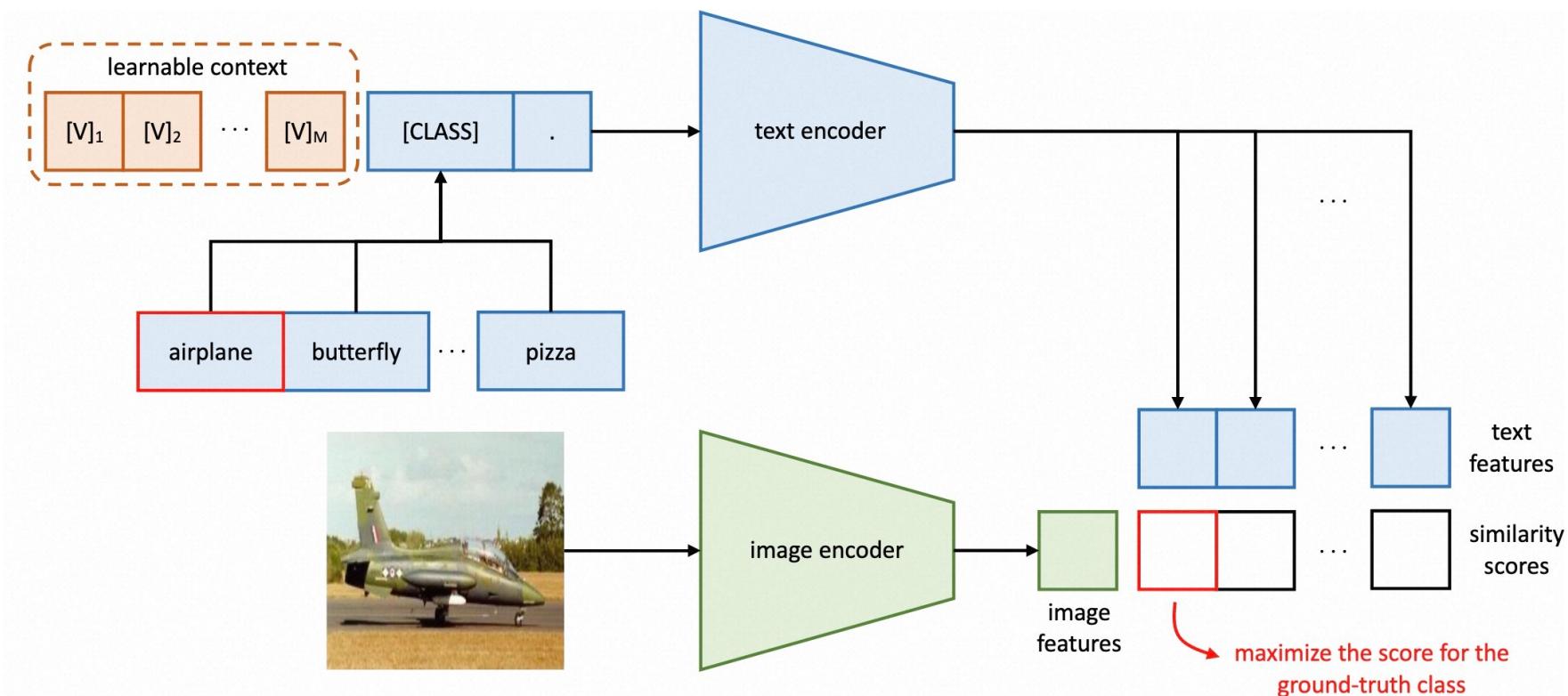
研究背景

7

□ Prompt Learning in CV

○ VLM with PL

- CoOp (IJCV 2022)、CoCoOp(CVPR 2022)





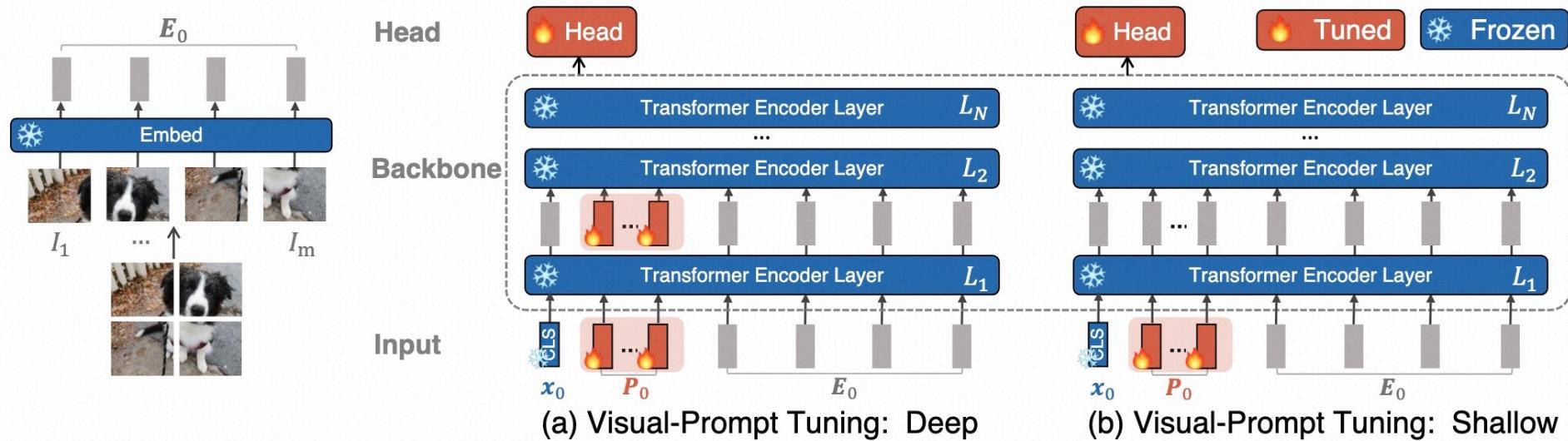
研究背景

8

□ Prompt Learning in CV

○ VM with PL

■ VPT (ECCV 2022)



- 作者介绍
- 研究动机
- CLIP-ReID
- LPT
- 实验效果
- 总结反思

CLIP-ReID

10

CLIP-ReID: Exploiting Vision-Language Model for Image Re-Identification without Concrete Text Labels

Siyuan Li¹, Li Sun^{1,2 *}, Qingli Li¹

¹ Shanghai Key Laboratory of Multidimensional Information Processing

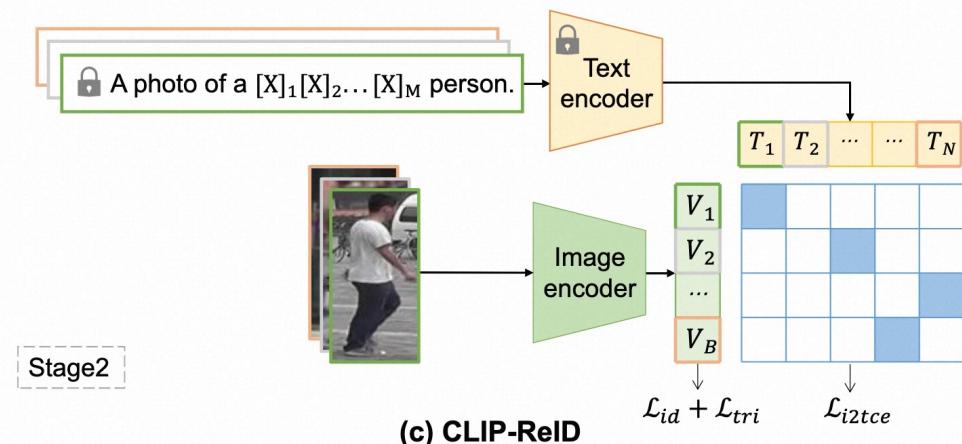
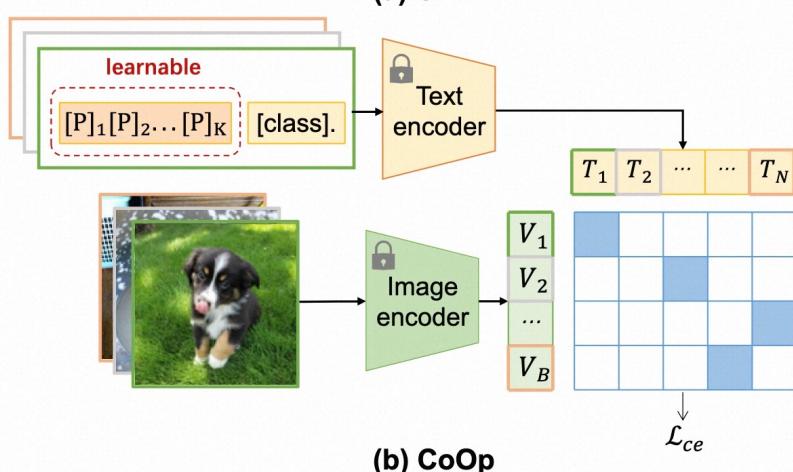
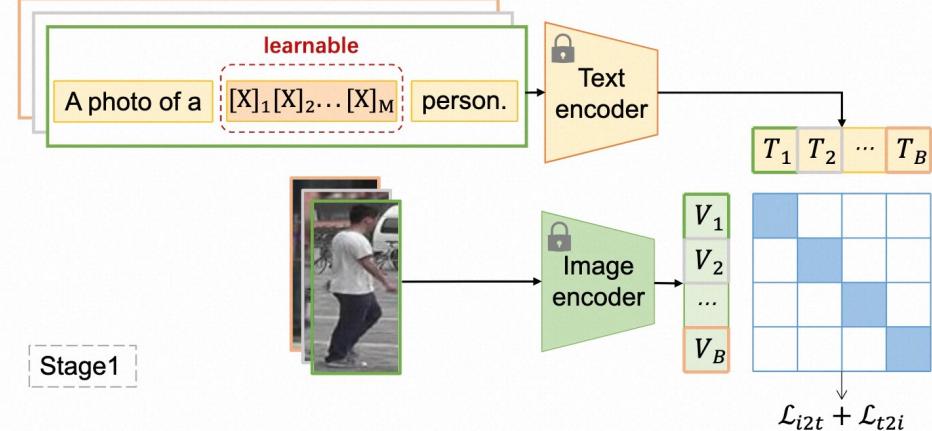
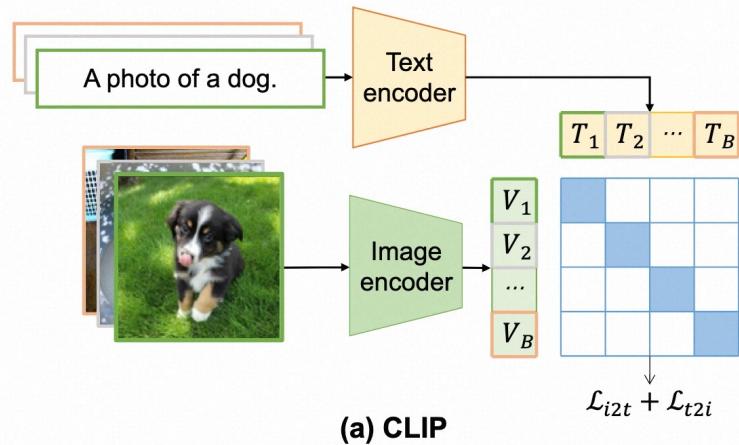
² Key Laboratory of Advanced Theory and Application in Statistics and Data Science
East China Normal University, Shanghai, China

AAAI2023

CLIP-ReID

11

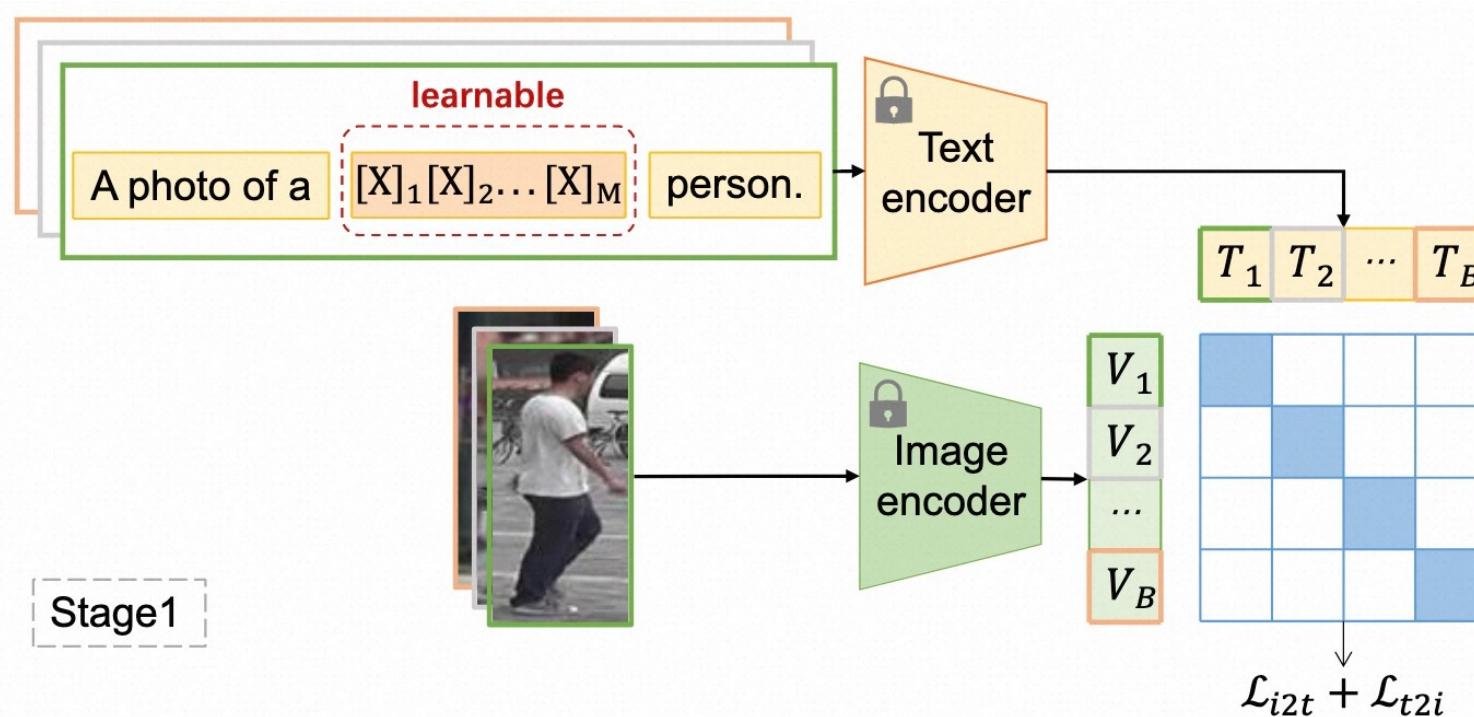
方法



CLIP-ReID

12

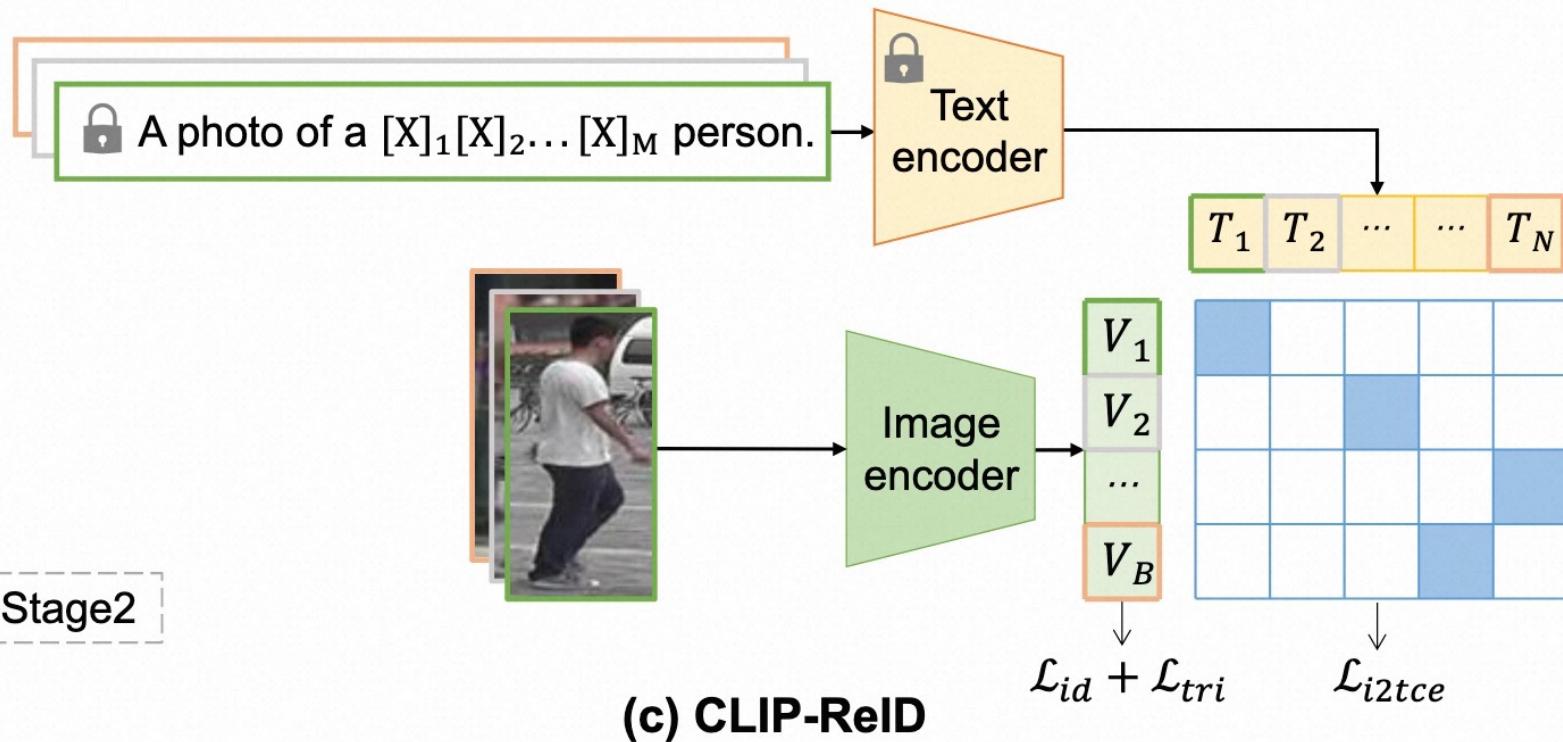
- 第一阶段
 - ◎ Learn prompt



CLIP-ReID

13

- 第二阶段
 - Fine-tune Image Encoder





CLIP-ReID

14

□ Person

Backbone	Methods	References	MSMT17		Market-1501		DukeMTMC		Occluded-Duke	
			mAP	R1	mAP	R1	mAP	R1	mAP	R1
CNN	PCB*	ECCV (2018)	-	-	81.6	93.8	69.2	83.3	-	-
	MGN*	MM (2018)	-	-	86.9	95.7	78.4	88.7	-	-
	OSNet	ICCV (2019)	52.9	78.7	84.9	94.8	73.5	88.6	-	-
	ABD-Net*	ICCV (2019)	60.8	82.3	88.3	95.6	78.6	89.0	-	-
	Auto-ReID*	ICCV (2019)	52.5	78.2	85.1	94.5	-	-	-	-
	HOReID	CVPR (2020)	-	-	84.9	94.2	75.6	86.9	43.8	55.1
	ISP	ECCV (2020)	-	-	88.6	95.3	80.0	89.6	52.3	62.8
	SAN	AAAI (2020b)	55.7	79.2	88.0	96.1	75.5	87.9	-	-
	OfM	AAAI (2021a)	54.7	78.4	87.9	94.9	78.6	89.0	-	-
	CDNet	CVPR (2021)	54.7	78.9	86.0	95.1	76.8	88.6	-	-
	PAT	CVPR (2021b)	-	-	88.0	95.4	78.2	88.8	53.6	64.5
	CAL*	ICCV (2021)	56.2	79.5	87.0	94.5	76.4	87.2	-	-
	CBDB-Net*	TCSVT (2021)	-	-	85.0	94.4	74.3	87.7	38.9	50.9
	ALDER*	TIP (2021b)	59.1	82.5	88.9	95.6	78.9	89.9	-	-
	LTReID*	TMM (2022)	58.6	81.0	89.0	95.9	80.4	90.5	-	-
	DRL-Net	TMM (2022)	55.3	78.4	86.9	94.7	76.6	88.1	50.8	65.0
	baseline		60.7	82.1	88.1	94.7	79.3	88.6	47.4	54.2
	CLIP-ReID		63.0	84.4	89.8	95.7	80.7	90.0	53.5	61.0
ViT	AAformer*	arxiv (2021)	63.2	83.6	87.7	95.4	80.0	90.1	58.2	67.0
	TransReID+SIE+OLP	ICCV (2021)	67.4	85.3	88.9	95.2	82.0	90.7	59.2	66.4
	TransReID+SIE+OLP*		69.4	86.2	89.5	95.2	82.6	90.7	-	-
	DCAL	CVPR (2022)	64.0	83.1	87.5	94.7	80.1	89.0	-	-
	baseline		66.1	84.4	86.4	93.3	80.0	88.8	53.5	60.8
	CLIP-ReID		73.4	88.7	89.6	95.5	82.5	90.0	59.5	67.1
	CLIP-ReID+SIE+OLP		75.8	89.7	90.5	95.4	83.1	90.8	60.3	67.2



CLIP-ReID

15

□ Car

Back -bone	Methods	VeRi-776		VehicleID	
		mAP	R1	R1	R5
CNN	PRN (2019)	74.3	94.3	78.4	92.3
	PGAN (2019)	79.3	96.5	77.8	92.1
	SAN (2020)	72.5	93.3	79.7	94.3
	UMTS (2020a)	75.9	95.8	80.9	-
	SPAN (2020)	68.9	94.0	-	-
	PVEN (2020)	79.5	95.6	84.7	97.0
	SAVER (2020)	79.6	96.4	79.9	95.2
	CFVMNet (2020)	77.1	95.3	81.4	94.1
	CAL (2021)	74.3	95.4	82.5	94.7
	EIA-Net (2018)	79.3	95.7	84.1	96.5
	FIDI (2021)	77.6	95.7	78.5	91.9
	baseline	79.3	95.7	84.4	96.6
ViT	CLIP-ReID	80.3	96.8	85.2	97.1
	TransReID (2021)	80.6	96.9	83.6	97.1
	TransReID!	82.0	97.1	85.2	97.5
	DCAL (2022)	80.2	96.9	-	-
	baseline	79.3	95.7	84.2	96.6
	CLIP-ReID	83.3	97.4	85.3	97.6
	CLIP-ReID!	84.5	97.3	85.5	97.2

- 作者介绍
- 研究动机
- CLIP-ReID
- LPT
- 实验效果
- 总结反思



LPT: LONG-TAILED PROMPT TUNING FOR IMAGE CLASSIFICATION

Bowen Dong¹ Pan Zhou² Shuicheng Yan² Wangmeng Zuo^{1,3✉}

¹Harbin Institute of Technology ²National University of Singapore ³Peng Cheng Laboratory
`{cndongsky, panzhou3, shuicheng.yan}@gmail.com, wmuo@hit.edu.cn`

□ 研究动机

◎ 1) VPT帮助对齐数据分布 2) VPT提高模型判别性

Table 1: Prompt tuning results on Places-LT (Zhou et al., 2017a). Prompt tuning achieves better accuracy on all classes and tail classes (*i.e.* “Few” in the table) with different training settings.

Method	Balanced Sampling	Tuned Params (w/o classifier)	Overall	Many	Medium	Few
Linear VPT	-	0 92K	33.29% 37.52%	46.48% 50.42%	29.45% 32.78%	18.77% 23.29%
Linear VPT	✓ ✓	0 92K	41.33% 44.17%	49.47% 45.79%	41.31% 46.73%	27.51% 36.18%

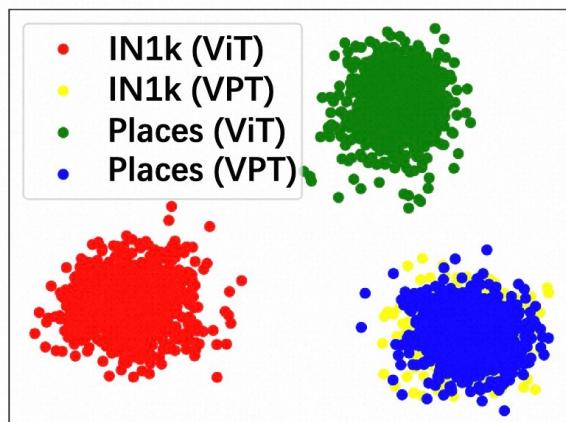
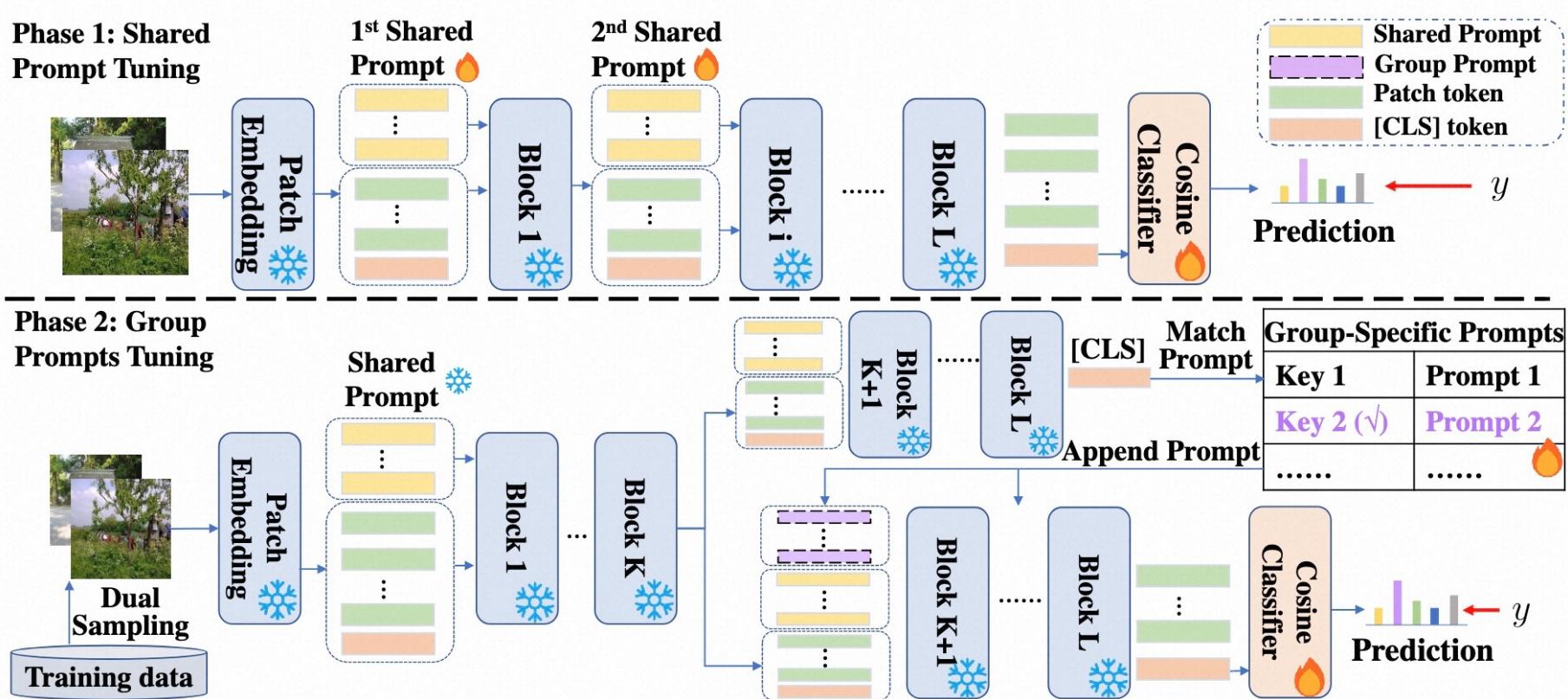


Figure 2: LDA visualization of VPT.

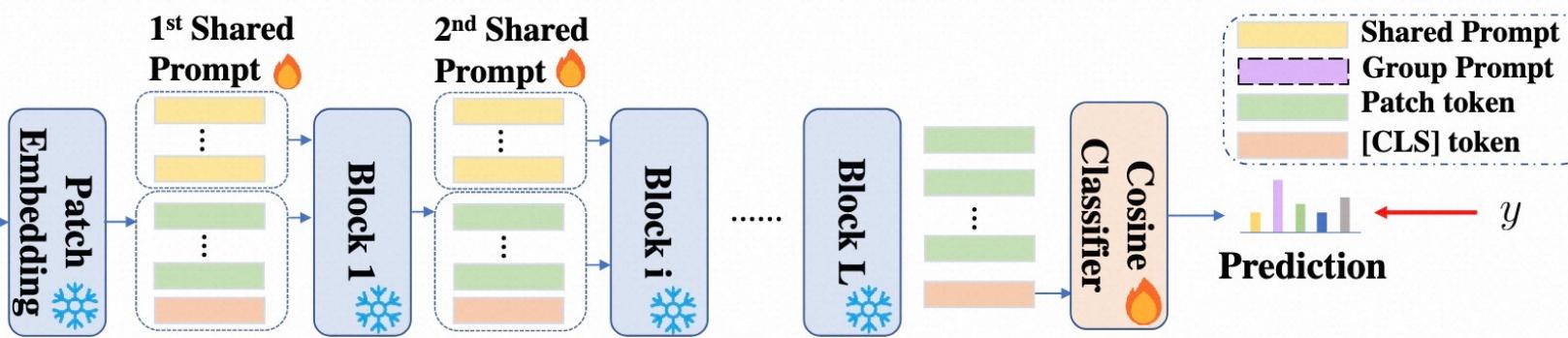
Table 2: Quantitative analysis of features learned by pretrained ViT-B and VPT. Features from VPT obtain better discriminative ability in terms of cluster compactness and also KNN accuracy.

Method	ViT-B	VPT
Pretrain Data	IN21k	IN21k
Fine-tuned	-	✓
Inner-class distance R_i	2.36 ± 0.52	1.82 ± 0.43
Inner-class / inter-class γ	0.171	0.128
K-NN Acc	30.80%	31.90%

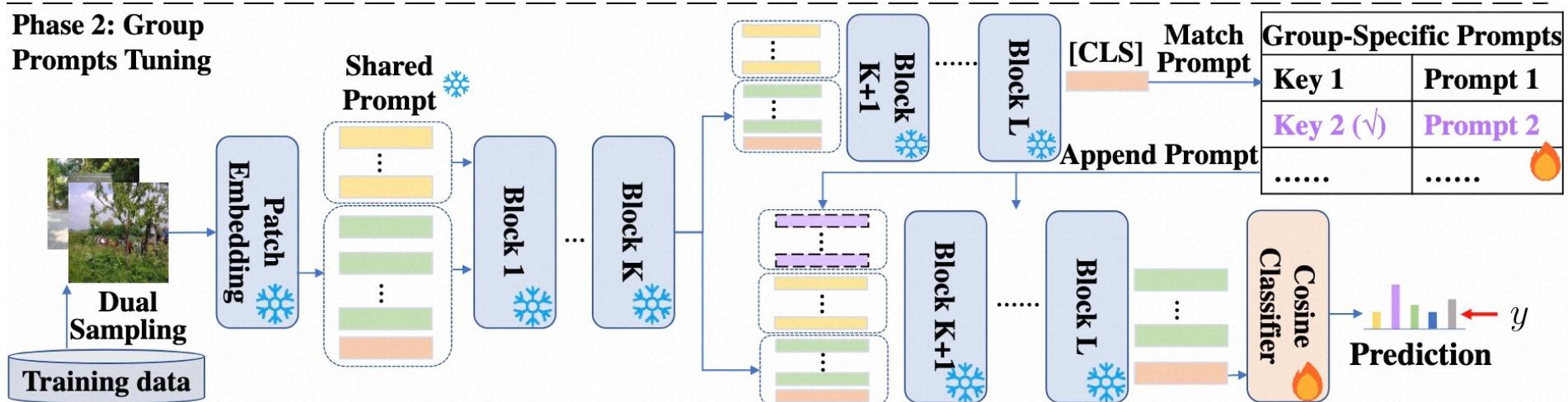
□ 方法


□ 第一阶段

Phase 1: Shared Prompt Tuning



□ 第二阶段



$$\text{prompts } \mathcal{R} = \{(\mathbf{k}_1, \mathbf{r}^1), \dots, (\mathbf{k}_m, \mathbf{r}^m)\}$$

$$[\mathbf{w}_1, \dots, \mathbf{w}_k] = \text{top-}k(\langle \mathbf{q}, [\mathbf{k}_1, \dots, \mathbf{k}_m] \rangle, k)$$

$$\mathbf{r} = \text{sum}([\mathbf{r}^{w_1}, \dots, \mathbf{r}^{w_k}])/k,$$

$$\mathcal{L}_{P_2} = \beta \mathcal{L}_{cls}(\hat{\mathbf{s}}, \mathbf{y}) + (1 - \frac{1}{k} \sum_{i \in w} \langle \mathbf{q}, \mathbf{k}_i \rangle),$$



□ 对比实验 on Places-LT.

Table 3: Comparison with state-of-the-art long-tailed classification methods on Places-LT dataset (Zhou et al., 2017a). Our LPT achieves state-of-the-art performance among vision-only pretrained methods and achieves the same performance with state-of-the-art VL-based methods.

Method	Backbone	Tuned Params	Total Params	Extra Data (Inference)	Overall	Many	Medium	Few
Vision-only Pretrained								
OLTR (Liu et al., 2019)	Res152	60.34M	60.34M	-	35.9	44.7	37.0	25.3
TADE (Zhang et al., 2021a)	Res152	60.34M	60.34M	-	38.8	42.8	39.0	31.2
LWS (Kang et al., 2020)	Res152	60.34M	60.34M	-	37.6	40.6	39.1	28.6
MisLAS (Zhong et al., 2021)	Res152	60.34M	60.34M	-	40.4	39.6	43.3	36.1
ALA (Zhao et al., 2022)	Res152	60.34M	60.34M	-	40.1	43.9	40.1	32.9
PaCo (Cui et al., 2021)	Res152	60.34M	60.34M	-	41.2	36.1	47.9	35.3
VPT (Jia et al., 2022)	ViT-B	0.09M	86.66M	-	37.5	50.4	33.8	23.3
LPT (Ours)	ViT-B	1.01M	87.58M	-	50.1	49.3	52.3	46.9
Vision-Languge Pretrained with Extra Data								
RAC (Long et al., 2022)	ViT-B	86.57M	236.19M	IN21k Feat	47.2	48.7	48.3	41.8
BALLAD (Ma et al., 2021)	ViT-B	149.62M	149.62M	-	49.5	49.3	50.2	48.4
VL-LTR (Tian et al., 2022)	ViT-B	149.62M	149.62M	Wiki Text	50.1	54.2	48.5	42.0



□ 对比实验 on CIFAR100-LT and iNaturalist 2018

Table 4: Comparison with state-of-the-art methods on CIFAR100-LT with various imbalanced ratio τ . LPT performs best among all methods.

Imb Ratio τ	200	100	50	10
Training from Scratch				
PaCo	-	52.0	56.0	64.2
Zhu et al. (2022)	-	51.9	56.6	64.9
Li et al. (2022)	44.9	48.7	53.6	-
Vision-only Pretrained				
VPT	72.8	81.0	84.8	89.6
LPT (Ours)	87.9	89.1	90.0	91.0
VL Pretrained with Extra Data				
Ma et al. (2021)	-	77.8	-	-

Table 5: Comparison with state-of-the-art methods on iNaturalist 2018. LPT performs best among vision-only pretrained methods.

Method	Overall	Few-shot
Vision-only Pretrained		
TADE	72.9	-
PaCo	75.2	74.7
ViT-B/16	73.2	-
Iscen et al. (2021)	75.3	73.2
ViT-L/16	75.9	-
LPT (Ours)	76.1	79.3
VL Pretrained with Extra Data		
VL-LTR	76.8	-
RAC	80.2	81.0



□ 消融实验

Table 6: ImageNet-Sketch evaluation results from different fine-tuning methods.

Method	Backbone	Overall
Linear Probe	ViT-B	31.55%
Full fine-tune	ViT-B	32.25%
LPT (Ours)	ViT-B	36.22%

Table 8: Ablation study of each phase in LPT on Places-LT benchmark (Zhou et al., 2017a).

Method	Prompt	Phase 1	$\mathcal{L}_{A\text{-GCL}}$	Phase 2	Overall	Many	Medium	Few
Linear VPT	- ✓	-	-	-	33.29% 37.52%	46.48% 50.42%	29.45% 33.78%	18.77% 23.29%
(a)	-	✓	-	-	41.33%	49.47%	41.31%	27.51%
(b)	✓	✓	-	-	49.10%	49.62%	51.53%	43.25%
(c)	✓	✓	✓	-	49.41%	46.89%	52.54%	47.32%
(d)	✓	✓	✓	✓	50.07%	49.27%	52.31%	46.88%

- 作者介绍
- 研究动机
- CLIP-ReID
- LPT
- 实验效果
- 总结反思



总结反思

26

- PL在不同子任务上的应用有一定研究价值

Thank for your attention !