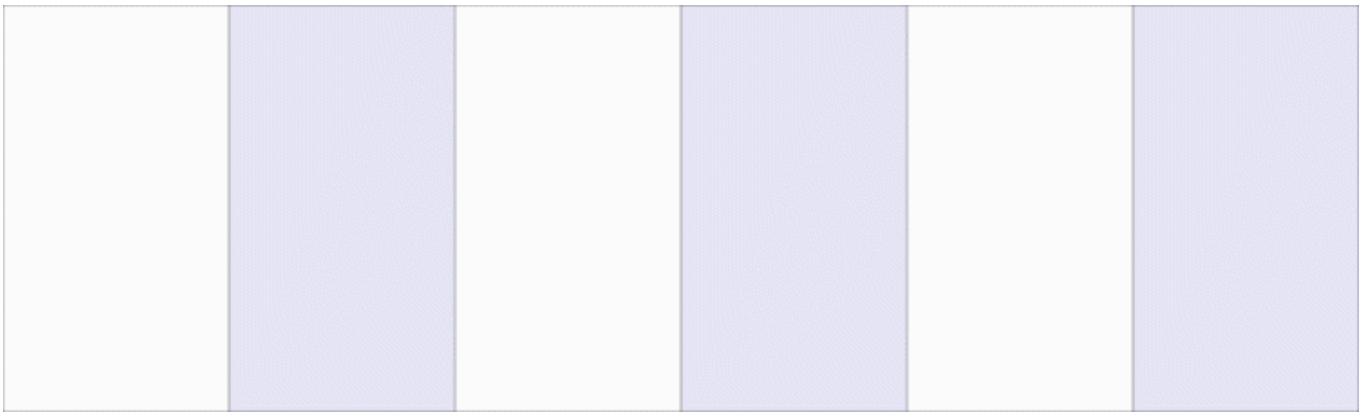
Prompt 学习和微调综述 (Prompt Learning and Tuning)

CV开发者都爱看的 极市平台 2023-04-01 22:01:06 发表于广东 手机阅读



作者丨Jarvis73@知乎（已授权）

来源丨https://zhuanlan.zhihu.com/p/601905339

编辑丨极市平台

极市导读

一文总结Prompt Learning/Tuning。 >>加入极市CV技术交流群，走在计算机视觉的最前沿

Self-Attention 和 Transformer 自从问世就成为了自然语言处理领域的新星. 得益于全局的注意力机制和并行化的训练, 基于 Transformer 的自然语言模型能够方便的编码长距离依赖关系, 同时在大规模自然语言数据集上并行训练成为可能. 但由于自然语言任务种类繁多, 且任务之间的差别不太大, 所以为每个任务单独微调一份大模型很不划算. 在 CV 中, 不同的图像识别任务往往也需要微调整个大模型, 也显得不够经济. Prompt Learning 的提出给这个问题提供了一个很好的方向.

本文关于 NLP 的部分主要参考综述[1].

1. NLP 模型的发展

过去许多机器学习方法是基于**全监督学习 (fully supervised learning)** 的.

由于监督学习需要大量的数据学习性能优异的模型, 而在 NLP 中大规模训练数据(指为特定任务而标注好的数据)是不足的, 因此在深度学习出现之前研究者通常聚焦于**特征工程 (feature engineering)**, 即利用领域知识从数据中提取好的特征;

在深度学习出现之后, 由于特征可以从数据中习得, 因此研究者转向了**结构工程 (architecture engineering)**, 即通过通过设计一个合适的网络结构来把归纳偏置 (inductive bias) 引入模型中, 从而有利于学习好的特征.

在 2017-2019 年, NLP 模型开始转向一个新的模式 (BERT), 即**预训练 + 微调 (pre-train and fine-tune)**. 在这个模式中, 先用一个固定的结构预训练一个**语言模型 (language model, LM)**, 预训练的方式就是让模型补全上下文 (比如完形填空).

由于预训练不需要专家知识, 因此可以在网络上搜集的大规模文本上直接进行训练. 然后这个 LM 通过引入额外的参数或微调来适应到下游任务上. 此时研究者转向了 **目标工程 (objective engineering)**, 即为预训练任务和微调任务设计更好的目标函数.

2. Prompt Learning

2.1 什么是 Prompt ?

在做 objective engineering 的过程中, 研究者发现让下游任务的目标与预训练的目标对齐是有好的. 因此下游任务通过引入**文本提示符 (textual prompt)**, 把原来的任务目标重构为与预训练模型一致的填空题.

比如一个输入 "I missed the bus today." 的重构:

- **情感预测任务.** 输入: "I missed the bus today. **I felt so___**." 其中 "I felt so" 就是**提示词 (prompt)**, 然后使用 LM 用一个表示情感的词填空.
- **翻译任务.** 输入: "**English**: I missed the bus today. **French: ___**" 其中 "English:" 和 "French:" 就是提示词, 然后使用 LM 应该再空位填入相应的法语句子.

我们发现用不同的 prompt 加到相同的输入上, 就能实现不同的任务, 从而使得下游任务可以很好的对齐到预训练任务上, 实现更好的预测效果.

后来研究者发现, 在同一个任务上使用不同的 prompt, 预测效果也会有显著差异, 因此现在有许多研究开始聚焦于 prompt engineering.

2.2 有哪些预训练模型?

- Left-to-Right LM: GPT, GPT-2, GPT-3
- Masked LM: BERT, RoBERTa
- Prefix LM: UniLM1, UniLM2

- Encoder-Decoder: T5, MASS, BART

2.3 有哪些 Prompt Learning 的方法?

- 按照 prompt 的形状划分: 完形填空式, 前缀式.
- 按照人的参与与否: 人工设计的, 自动的(离散的, 连续的)

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 ...
Text-pair CLS	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. ...

Table 3: Examples of *input*, *template*, and *answer* for different tasks. In the **Type** column, "CLS" is an abbreviation for "classification". In the **Task** column, "NLI" and "NER" are abbreviations for "natural language inference" (Bowman et al., 2015) and "named entity recognition" (Tjong Kim Sang and De Meulder, 2003) respectively.

人工设计的 Prompt

3. Prompt Tuning

3.1 Fine-tune 的策略

在下游任务上微调大规模预训练模型已经成为大量 NLP 和 CV 任务常用的训练模式. 然而, 随着模型尺寸和任务数量越来越多, 微调整个模型的方法会储存每个微调任务的模型副本, 消耗大量的储存空间. 尤其是在边缘设备上存储空间和网络速度有限的情况下, 共享参数就变得尤为重要.

一个比较直接的共享参数的方法是只微调部分参数, 或者向预训练模型中加入少量额外的参数. 比如, 对于分类任务:

- Linear: 只微调分类器 (一个线性层), 冻结整个骨干网络.

- Partial-k: 只微调骨干网络最后的 k 层, 冻结其他层[2][3].
- MLP-k: 增加一个 k 层的 MLP 作为分类器.
- Side-tuning[4]: 训练一个 "side" 网络, 然后融合预训练特征和 "side" 网络的特征后输入分类器.
- Bias: 只微调预训练网络的 bias 参数[5][6].
- Adapter[7]: 通过残差结构, 把额外的 MLP 模块插入 Transformer.

近年来, Transformer 模型在 NLP 和 CV 上大放异彩. 基于 Transformer 的模型在大量 CV 任务上已经比肩甚至超过基于卷积的模型.

Transformer 与 ConvNet 比较: Transformer 相比于 ConvNet 的一个显著的特点是: 它们在对于空间(时间)维度的操作是不同的.

- ConvNet: 卷积核在空间维度上执行卷积操作, 因此空间内不同位置的特征通过卷积(可学习的)操作融合信息, 且只在局部区域融合.
- Transformer: 空间(时间)维度内不同位置的特征通过 Attention (非学习的) 操作融合信息, 且在全局上融合.

Transformer 在特征融合时非学习的策略使得其很容易的通过增加额外的 feature 来扩展模型.

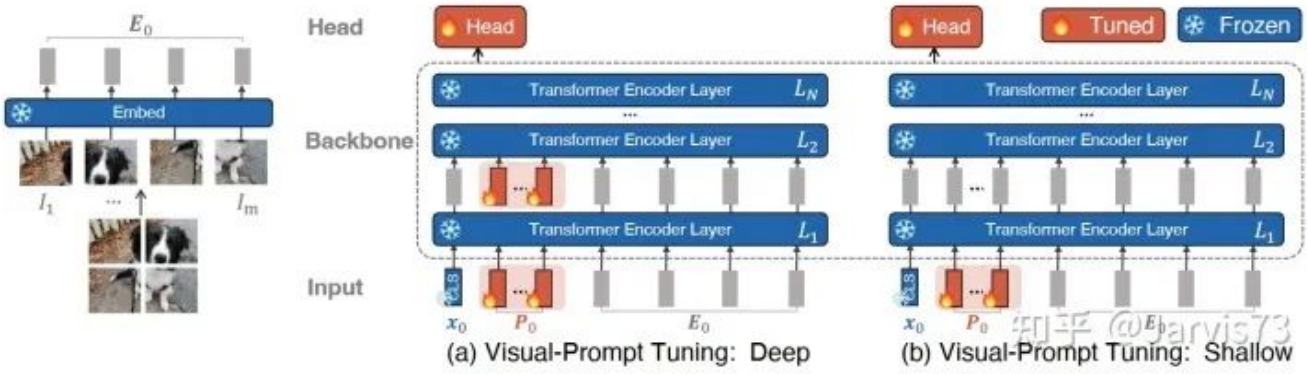
3.2 NLP 中 基于 Prompt 的 fine-tune

- Prefix-Tuning
- Prompt-Tuning
- P-Tuning
- P-Tuning-v2

3.3 CV 中 基于 Prompt 的 fine-tuning

3.3.1 分类

Visual Prompt Tuning[8]



Visual Prompt Tuning

- VPT-Shallow

$$
\begin{aligned}
[\mathbf{x}_1, \mathbf{Z}_1, \mathbf{E}_1] &= L_1([\mathbf{x}_0, \mathbf{P}, \mathbf{E}_0]) \\
[\mathbf{x}_i, \mathbf{Z}_i, \mathbf{E}_i] &= L_i([\mathbf{x}_{i-1}, \mathbf{Z}_{i-1}, \mathbf{E}_{i-1}]) \qquad i = 2, 3, \ldots, N \\
y &= \text{Head}(x_N)
\end{aligned}
$$

- VPT-Deep

$$
\begin{aligned}
[\mathbf{x}_i, \_, \mathbf{E}_i] &= L_i([\mathbf{x}_{i-1}, \mathbf{P}_{i-1}, \mathbf{E}_{i-1}]) \qquad i = 1, 2, \ldots, N \\
y &= \text{Head}(x_N)
\end{aligned}
$$

Table 1. ViT-B/16 pre-trained on supervised ImageNet-21k. For each method and each downstream task group, we report the average test accuracy score and **number of wins in** (·) compared to FULL. "Total params" denotes total parameters needed for all 24 downstream tasks. "Scope" denotes the tuning scope of each method. "Extra params" denotes the presence of additional parameters besides the pre-trained backbone and linear head. Best results among all methods except FULL are **bolded**. VPT outshines the full fine-tuning 20 out of 24 cases with significantly less trainable parameters

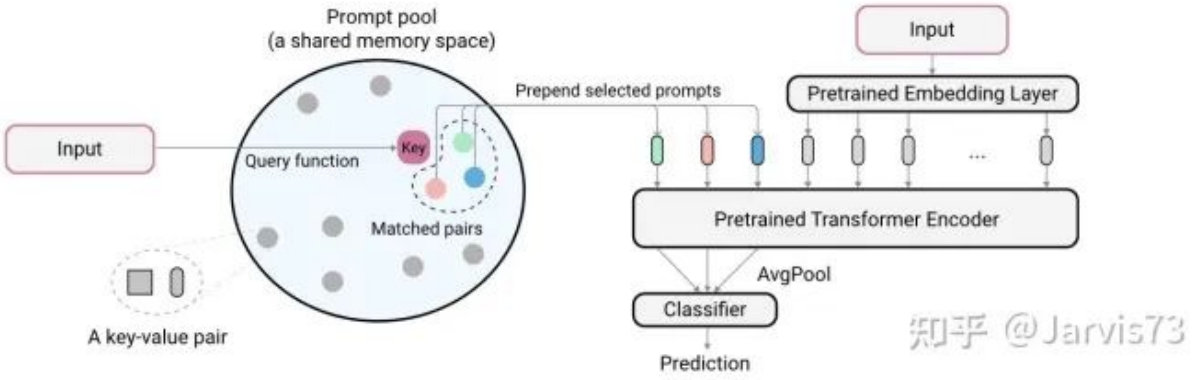
	ViT-B/16 (85.8M)	Total params	Scope Input	Scope Backbone	Extra params	FGVC	VTAB-1k Natural	VTAB-1k Specialized	VTAB-1k Structured
	Total # of tasks					5	7	4	8
(a)	FULL	24.02×		✓		88.54	75.88	83.36	47.64
(b)	LINEAR	1.02×				79.32 (0)	68.93 (1)	77.16 (1)	26.84 (0)
	PARTIAL-1	3.00×				82.63 (0)	69.44 (2)	78.53 (0)	34.17 (0)
	MLP-3	1.35×			✓	79.80 (0)	67.80 (2)	72.83 (0)	30.62 (0)
(c)	SIDETUNE	3.69×		✓	✓	78.35 (0)	58.21 (0)	68.12 (0)	23.41 (0)
	BIAS	1.05×		✓		88.41 (3)	73.30 (3)	78.25 (0)	44.09 (2)
	ADAPTER	1.23×		✓	✓	85.66 (2)	70.39 (4)	77.11 (0)	33.43 (0)
(ours)	VPT-SHALLOW	1.04×	✓		✓	84.62 (1)	76.81 (4)	72.69 (0)	46.98 (4)
	VPT-DEEP	1.18×				**89.11 (4)**	**78.48 (6)**	**82.43 (2)**	**54.98 (8)**

VPT Results

3.3.2 持续学习

Learning to Prompt for Continue Learning[9]

引入一个 prompt pool, 对每个 input, 从 pool 中取出与其最近的 N 个 prompts 加入 image tokens. input 和 prompts 距离的度量通过计算 input feature 和每个 prompt 的 key 的距离来得到, 这些 key 通过梯度随分类目标一起优化.



L2P

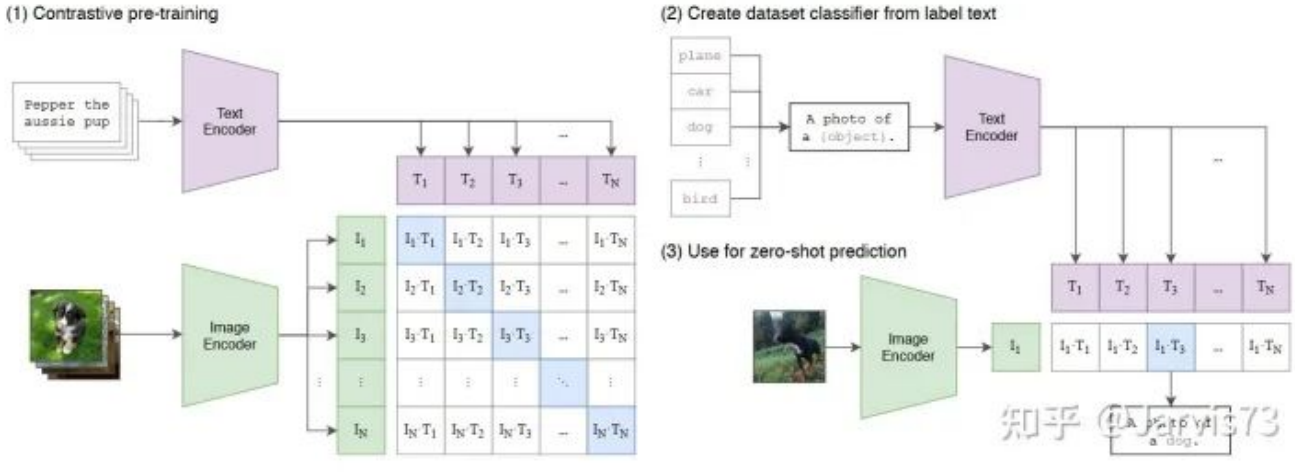
$$\min_{\mathbf{P},\mathbf{K},\phi} \mathcal{L}(g_\phi(f_r^{avg}(\mathbf{x}_p)), y) + \lambda \sum_{\mathbf{K_x}} \gamma(q(\mathbf{x}), \mathbf{k}_{s_i})$$

注意, 最后使用 prompt 来分类.

3.3.3 多模态模型

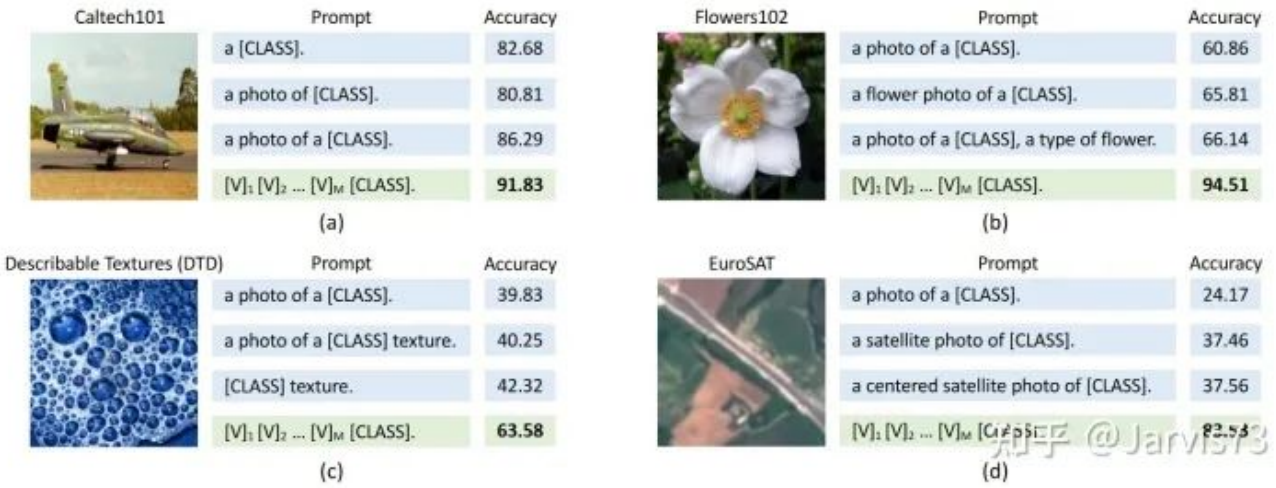
Vision-Language Model: Context Optimization (CoOp)[10]

多模态学习的预训练模型. 比如 CLIP, 通过对比学习对齐文本和图像的特征空间.



CLIP

选择不同的文本 prompt 对于精度影响较大.

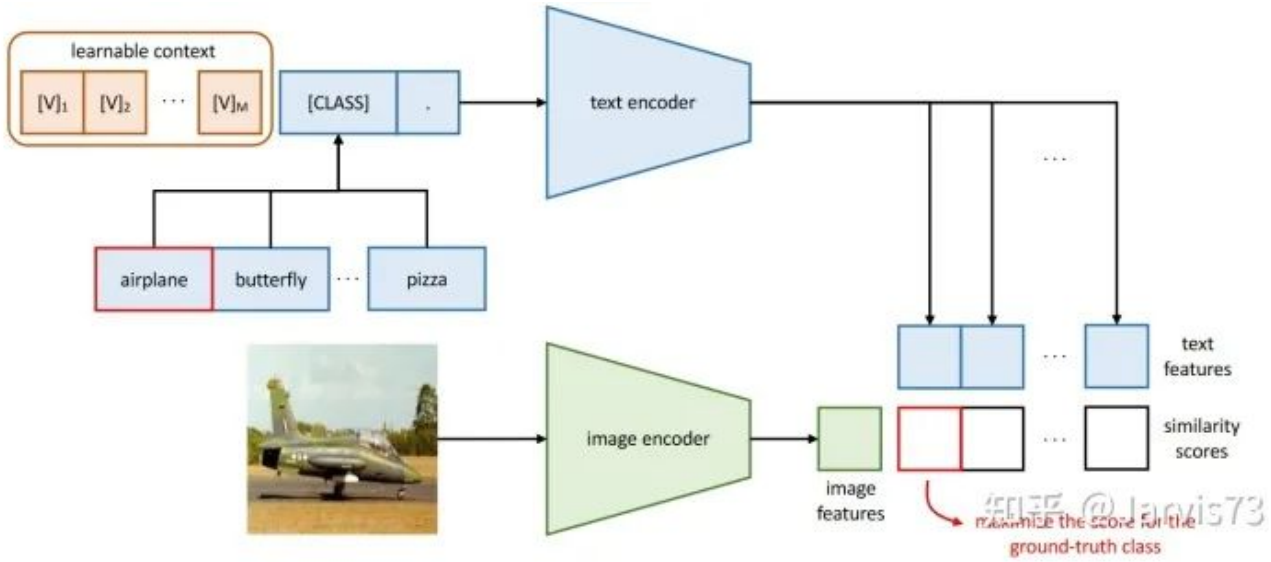


Prompt engineering vs Context Optimization (CoOp)

把人工设定的 prompt 替换为 learnable 的 prompt:

- [CLASS] 放在后面:$t = [\mathrm{V}]_1[\ \mathrm{V}]_2 \dots [\mathrm{V}]_M[\mathrm{CLASS}]$
- [CLASS] 放在中间:$t = [\mathrm{V}]_1 \dots [\mathrm{V}]_{\frac{M}{2}}[\mathrm{CLASS}][\mathrm{V}]_{\frac{M}{2}+1} \dots [\mathrm{V}]_M$

Prompt 可以在不同类之间公用, 也可以为每个类使用不同的 prompts (对于细粒度分类任务更有效).



Learning to Prompt for Vision-Language Model

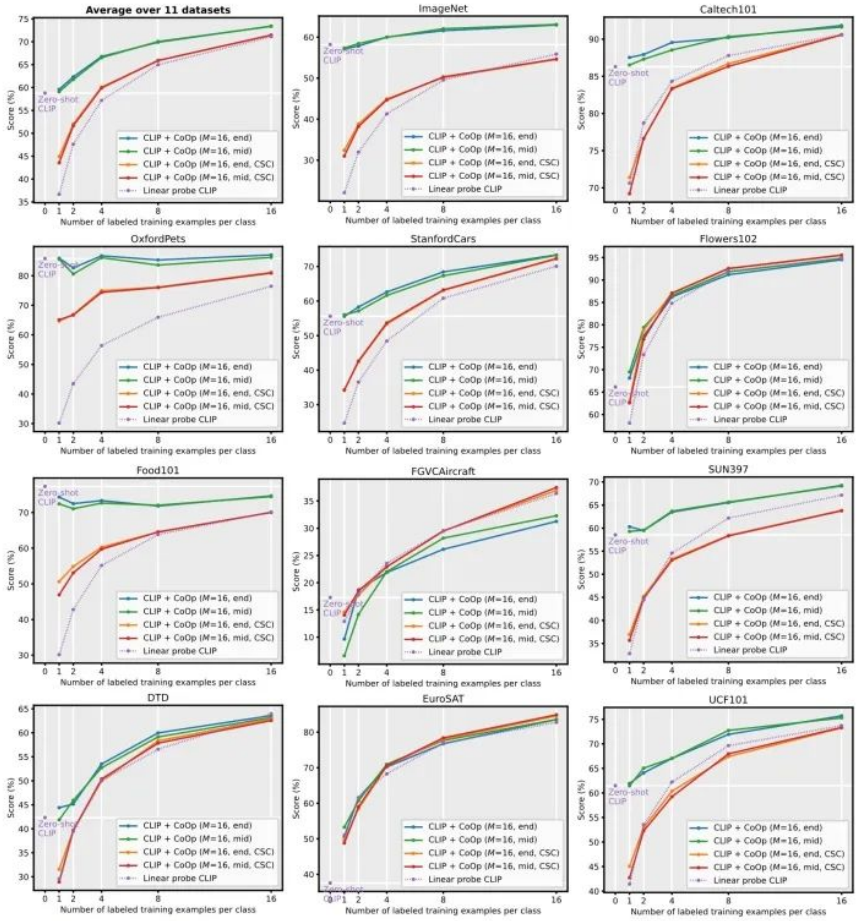
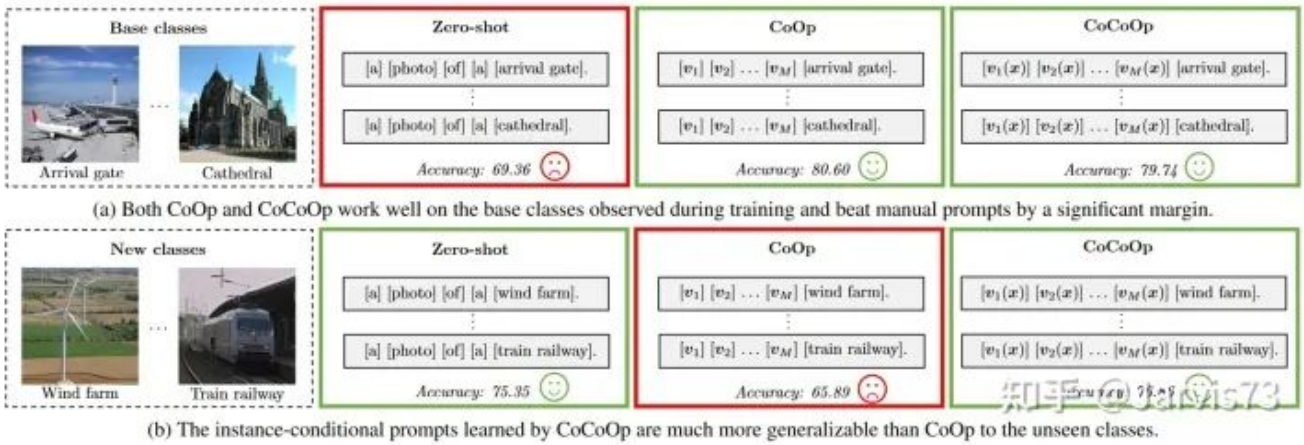


Fig. 3 Main results of few-shot learning on the 11 datasets. Overall, CoOp effectively turns CLIP into a strong few-shot learner (solid lines), achieving significant improvements over zero-shot CLIP (stars) and performing favorably against the linear probe alternative (dashed lines). M denotes the context length. "end" or "mid" means putting the class token in the end or middle. CSC means class-specific context.

Learning to Prompt for Vision-Language Model

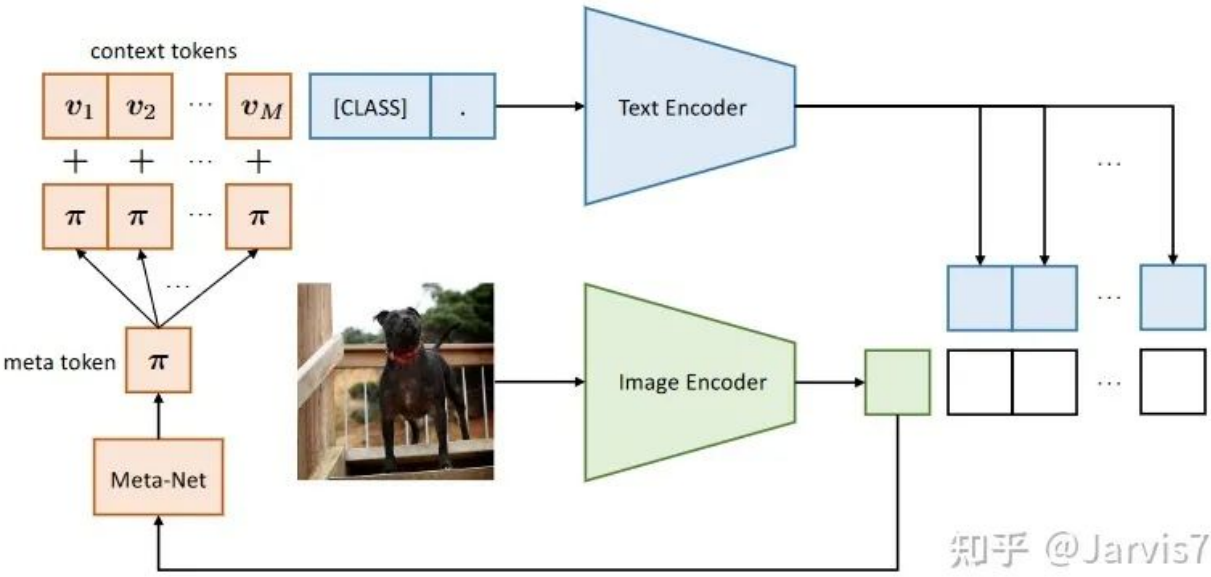
Conditional Prompt Learning for Vision-Language Models[11]

CoOp 在泛化到新的类别上时性能不好.



To learn generalizable prompts

所以把 prompt 设计为 instance-conditional 的.



To learn generalizable prompts

为 prompt 加上一个跟当前图像相关的特征以提高泛化性能. 具体来说, 先用 Image Encoder 计算当前图像的 feature, 然后通过一个 Meta-Net 把 feature 映射到 prompt 的特征空间, 加到 prompt 上面.

Table 1. **Comparison of CLIP, CoOp and CoCoOp in the base-to-new generalization setting**. For learning-based methods (CoOp and CoCoOp), their prompts are learned from the base classes (16 shots). The results strongly justify the strong generalizability of conditional prompt learning. H: Harmonic mean (to highlight the generalization trade-off [54]).

(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.

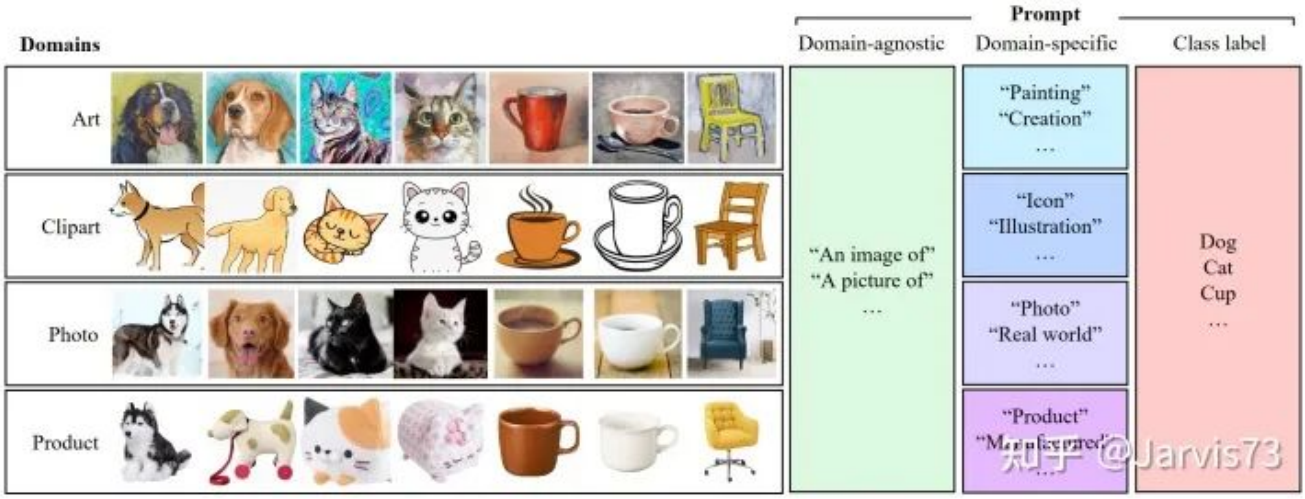
	Base	New	H
CLIP	70.53	**77.50**	73.85
CoOp	**84.69**	56.05	67.46
CoCoOp	82.33	73.45	**77.64**

To learn generalizable prompts

3.3.4 域适应

Domain Adaptation via Prompt Learning[12]

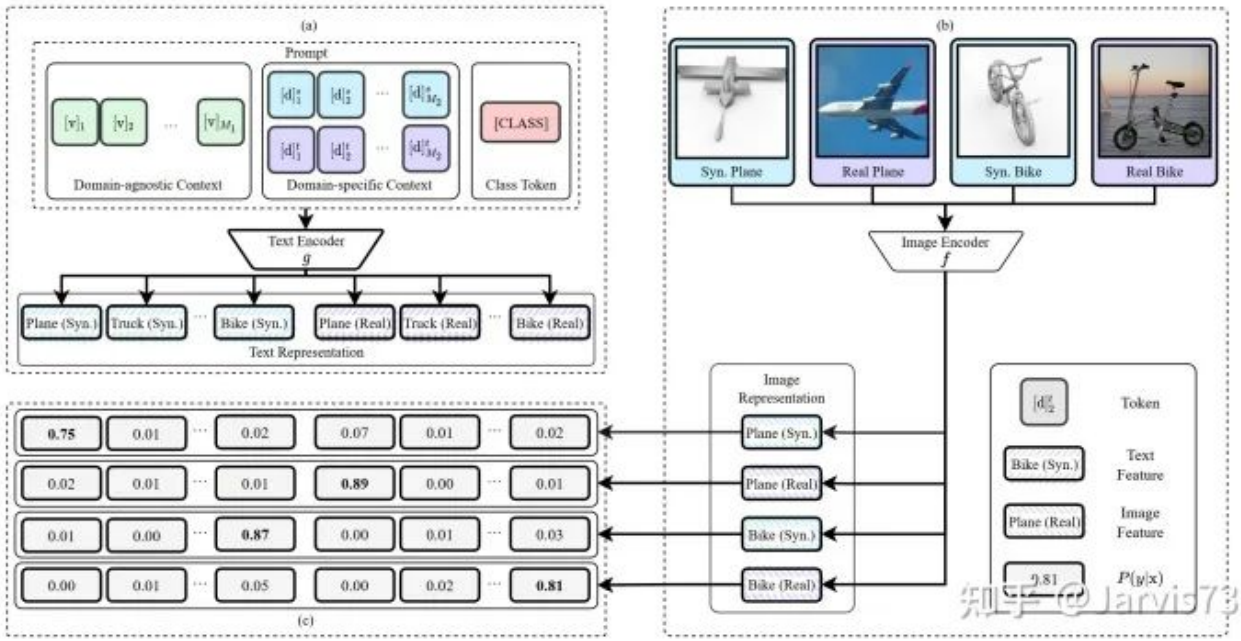
用 prompt 来标识 domain 的信息.



Example prompt structure

通过对比学习解耦 representation 中的 class 和 domain 的表示.

$$P(\hat{y}_i^s = k|\mathbf{x}_i^s) = \frac{\exp(\langle g(\mathbf{t}_k^s), f(\mathbf{x}_i^s)\rangle/T)}{\sum_{d\in\{s,u\}}\sum_{j=1}^{K}\exp(\langle g(\mathbf{t}_j^s), f(\mathbf{x}_i^s)\rangle/T)}$$



Domain Adaptation with Prompt Learning

参考

1. ^Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing. Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Ji

ang, Hiroaki Hayashi, Graham Neubig. In arXiv 2021 https://arxiv.org/abs/2107.13586

2. ^How transferable are features in deep neural networks? Jason Yosinski, Jeff Clune, Yoshua Bengio, Hod Lipson. In NeruIPS 2014 https://proceedings.neurips.cc/paper/2014/hash/375c71349b295fbe2dcdca9206f20a06-Abstract.html

3. ^Masked autoencoders are scalable vision learners. Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, Ross Girshick. In arXiv 2021 https://arxiv.org/abs/2111.06377

4. ^Side-tuning: a baseline for network adaptation via additive side networks. Jeffrey O. Zhang, Alexander Sax, Amir Zamir, Leonidas Guibas, Jitendra Malik. In ECCV 2020 https://link.springer.com/chapter/10.1007/978-3-030-58580-8_41

5. ^Bitfit: Simple parameter-efficient fine-tuning for transformer-based masked language-models.Elad Ben Zaken, Shauli Ravfogel, Yoav Goldberg. In ACL 2022 https://arxiv.org/abs/2106.10199

6. ^TinyTL: Reduce memory, not parameters for efficient on-device learning. Han Cai, Chuang Gan, Ligeng Zhu, Song Han. In NeurIPS 2020 https://proceedings.neurips.cc/paper/2020/hash/81f7acabd411274fcf65ce2070ed568a-Abstract.html

7. ^Parameter-efficient transfer learning for nlp. Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, Sylvain Gelly. In ICML 2019 http://proceedings.mlr.press/v97/houlsby19a.html

8. ^Visual Prompt Tuning. Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, Ser-Nam Lim. In arXiv 2022 https://arxiv.org/abs/2203.12119

9. ^Learning to Prompt for Continual Learning. Zifeng Wang, Zizhao Zhang, Chen-Yu Lee, Han Zhang, Ruoxi Sun, Xiaoqi Ren, Guolong Su, Vincent Perot, Jennifer Dy, Tomas Pfister. In CVPR 2022 https://arxiv.org/abs/2112.08654

10. ^Learning to Prompt for Vision-Language Models. Kaiyang Zhou, Jingkang Yang, Chen Change Loy, Ziwei Liu. In arXiv 2021 https://arxiv.org/abs/2109.01134

11. ^Conditional Prompt Learning for Vision-Language Models. Kaiyang Zhou, Jingkang Yang, Chen Change Loy, Ziwei Liu. In CVPR 2022 https://arxiv.org/abs/2203.05557

12. ^Domain Adaptation via Prompt Learning. Chunjiang Ge, Rui Huang, Mixue Xie, Zihang Lai, Shiji Song, Shuang Li, Gao Huang. In arXiv 2022 https://arxiv.org/abs/2202.06687

极市干货

极视角动态：「无人机+AI」光伏智能巡检，硬核实力遇见智慧大脑！ | 「AI 警卫员」上线，极视角守护龙大食品厂区安全！ | 点亮海运指明灯，极视角为海上运输船员安全管理保驾护航！

CVPR2023：CVPR'23 最新 125 篇论文分方向整理 | 检测、分割、人脸、视频处理、医学影像、神经网络结构、小样本学习等方向

数据集：自动驾驶方向开源数据集资源汇总 | 医学影像方向开源数据集资源汇总 | 卫星图像公开数据集资源汇总

点击阅读原文进入CV社区

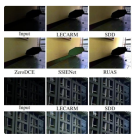
收获更多技术干货

阅读原文

喜欢此内容的人还喜欢

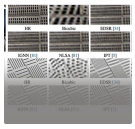
ICCV23｜将隐式神经表征用于低光增强，北大张健团队提出NeRCo

极市平台



ICCV 2023｜南开程明明团队提出适用于SR任务的新颖注意力机制（已开源）

极市平台



实践教程｜使用 OpenCV 进行特征提取（颜色、形状和纹理）

极市平台