LaFTer: Label-Free Tuning of Zero-shot Classifier using Language and Unlabeled Image Collections

Neurips 2023

Paper Reading by Yixuan Zhang

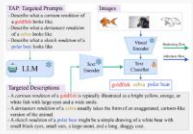
智能多媒体内容计算实验室
Intelligent Multimedia Content Computing Lab

- 作者介绍
- 研究背景
- 本文方法
- 实验效果
- 总结反思

作者介绍

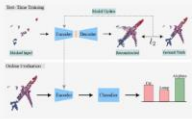
M. Jehanzeb Mirza[†1] **Leonid Karlinsky**[2] **Wei Lin**[1] **Mateusz Kozinski**[1]
Horst Possegger[1] **Rogerio Feris**[2] **Horst Bischof**[1]

[1]Institute for Computer Graphics and Vision, TU Graz, Austria.
[2]MIT-IBM Watson AI Lab, USA.

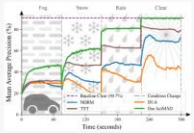


TAP: Targeted Prompting for Task Adaptive Generation of Textual Training Instances for Visual Classification
M. Jehanzeb Mirza, Leonid Karlinsky, Wei Lin, Horst Possegger, Rogerio Feris, Horst Bischof
ArXiv pre-print (Under Review)
[Paper]

LaFTer: Label-Free Tuning of Zero-shot Classifier using Language and Unlabeled Image Collections
M. Jehanzeb Mirza, Leonid Karlinsky, Wei Lin, Mateusz Kozinski, Horst Possegger, Rogerio Feris, Horst Bischof
NeurIPS 2023
[Paper]



MATE: Masked Autoencoders are Online 3D Test-Time Learners
[*]**M. Jehanzeb Mirza**,[*]Inkyu Shin, [*]Wei Lin, Andreas Schriebl, Kunyang Sun, Jaesung Choe, Mateusz Kozinski, Horst Possegger, In So Kweon, Kun-Jin Yoon, Horst Bischof *(*Equal Contribution)*
ICCV 2023
[Paper | Code]



ActMAD: Activation Matching to Align Distributions for Test-Time-Training
M. Jehanzeb Mirza, Pol Jane Soneira, Wei Lin, Mateusz Kozinski, Horst Possegger, Horst Bischof
CVPR 2023
[Paper | Project Page | Code]

Video Test-Time Adaptation for Action Recognition
[*]Wei Lin, [*]**M. Jehanzeb Mirza**, Mateusz Kozinski, Horst Possegger, Hilde Kuehne, Horst Bischof *(*Equal Contribution)*
CVPR 2023
[Paper | Code]

- 作者介绍
- 研究背景
- 本文方法
- 实验效果
- 总结反思

研究背景

- Large-scale Vision&Language (VL) Models
  - Zero-shot downstream tasks
  - CLIP, ALIGN,…
  - Improve image-text representation alignment
    - off-the-shelf object detectors
    - cross-attention and additional objective functions ITM / MIM
    - finer-grained interactions, modern Hopfield networks, optimal transport distillation, cycle consistency , hierarchical feature alignment

研究背景

- 无监督微调VLM
  - Unsupervised prompt learning (UPL)
    - 构造伪标签，learnable prompt
    - https://arxiv.org/abs/2204.03649

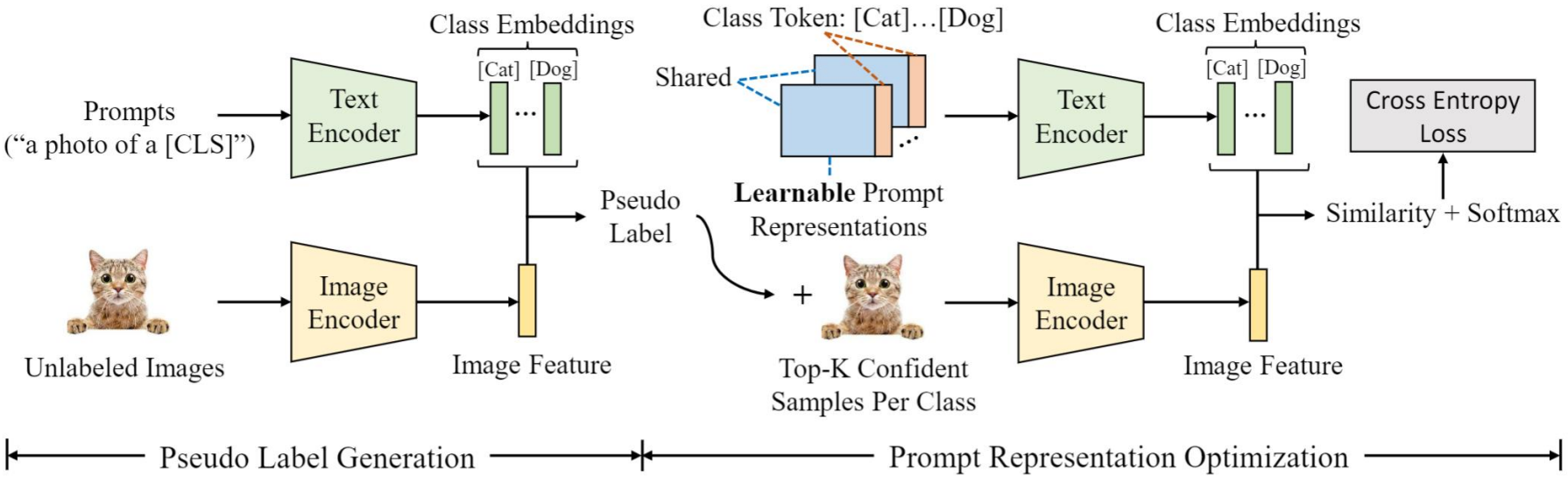


Figure 2: Overview of the proposed unsupervised prompt learning (UPL) framework. Our UPL mainly contains two parts, namely pseudo label generation and prompt representation optimization. We first use CLIP with a simple prompt (e.g., "a photo of a [CLS]") to generate pseudo labels for target datasets and select top-K confident samples per class for subsequent training. Then we define a learnable prompt representation which is optimized on selected pseudo-labeled samples. For inference, we simply swap out the hand-crafted prompts with the well-optimized prompt representations

研究背景

- 无监督微调VLM
  - Improving Zero-Shot Models with Label Distribution Priors (CLIP-PR)
    - 伪标签 + 分布约束

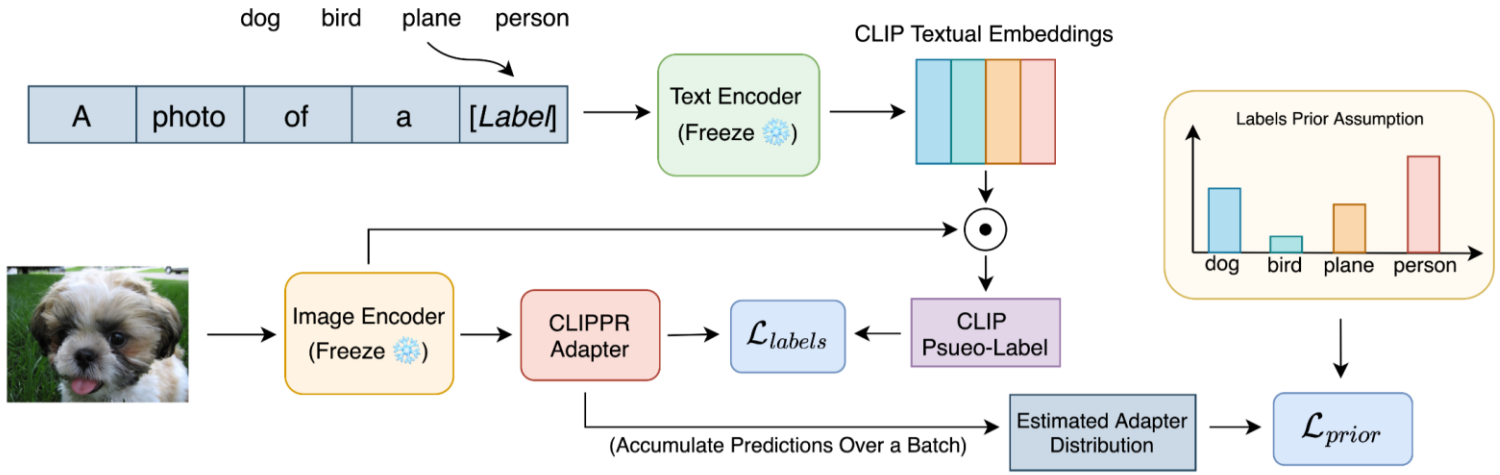
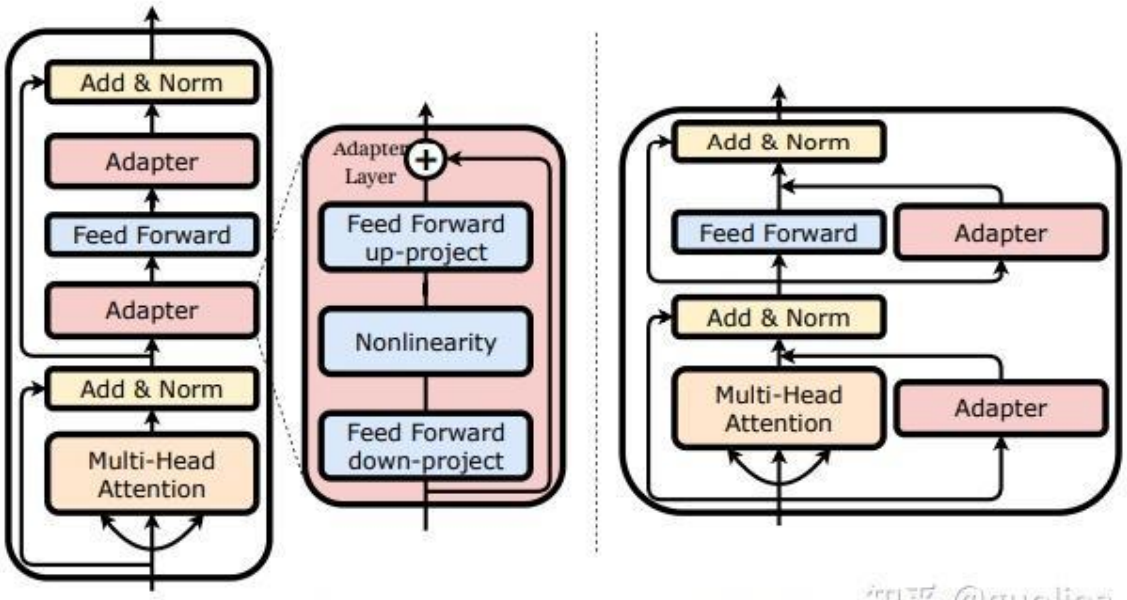


Figure 1. An illustration of our method. Our method trains an adapter module on top of a frozen V&L model image encoder. The adapter is trained with two competing objectives: (i) Predicting labels close to the original V&L model zero-shot predictions. (ii) Predicting a labels distribution similar to the given prior distribution. Together, these two objectives adapt the original zero-shot predictions to the distributional prior, resulting in better performance.

研究背景

- 预训练大模型的微调方法
  - Fine-tuning
  - Parameter efficient fine-tuning（PEFT）
    - Adapter



(a) Series Adapter　　(b) Parallel Adapter

研究背景

- 预训练大模型的微调方法
  - Fine-tuning
  - Parameter efficient fine-tuning（PEFT）
    - Adapter
    - Low-Rank Adaption(LoRA)

$$h = W_0 x + \Delta W x = W_0 x + BAx$$

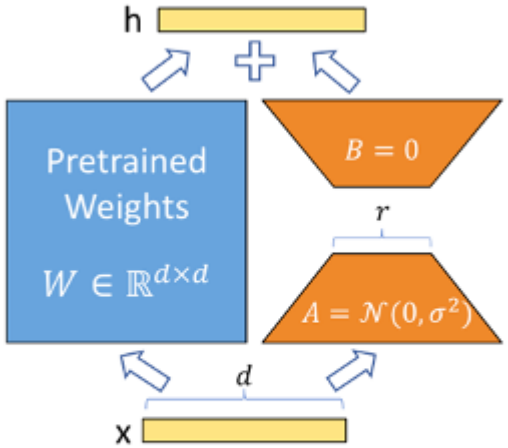


Figure 1: Our reparametrization. We only train A and B.

- 模型是过参数化的，依赖于这个低的内在维度（low intrinsic dimension）去做任务适配。低秩自适应（LoRA）方法对大模型进行微调。

- Adapter由于增加了模型的深度从而额外增加了模型推理的延时
- Prompt较难训练，同时减少了模型的可用序列长度

研究背景

- 预训练大模型的微调方法
  - Fine-tuning
  - Parameter efficient fine-tuning（PEFT）
    - Adapter
    - LoRA
    - Prompt tuning : text prompting

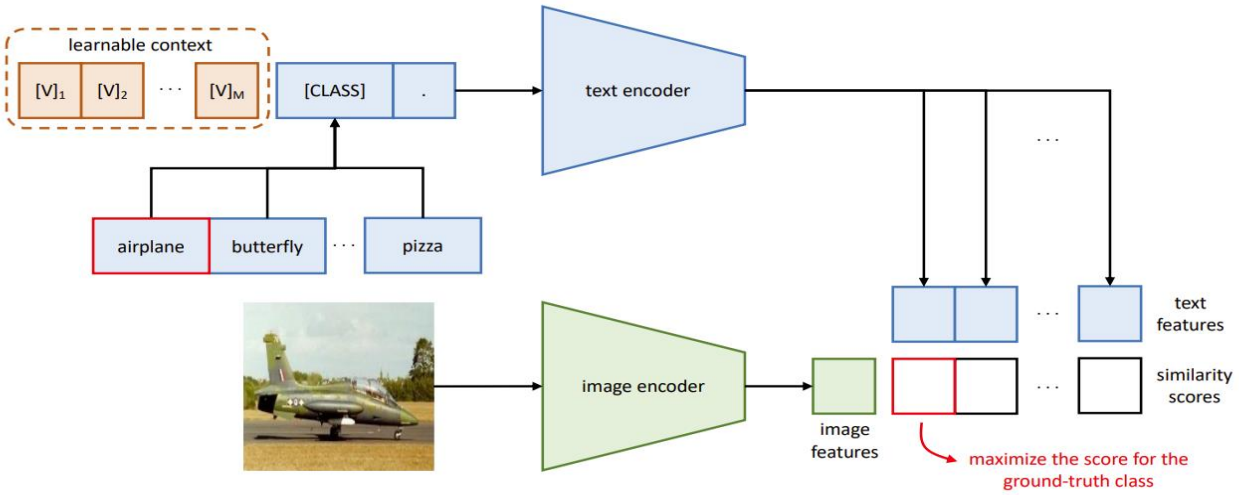


Fig. 2 Overview of Context Optimization (CoOp). The main idea is to model a prompt's context using a set of learnable vectors, which can be optimized through minimizing the classification loss. Two designs are proposed: one is unified context, which shares the same context vectors with all classes; and the other is class-specific context, which learns for each class a specific set of context vectors.

研究背景

- 预训练大模型的微调方法
  - Fine-tuning
  - Parameter efficient fine-tuning（PEFT）
    - Adapter
    - LoRA
    - Prompt tuning : text prompting

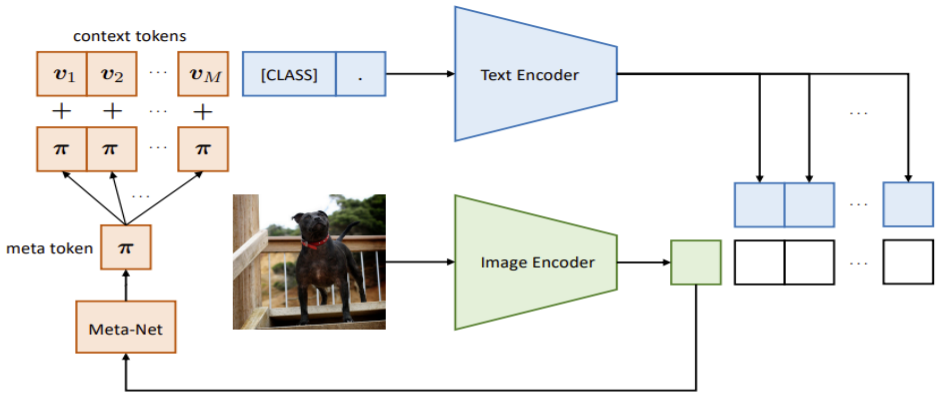
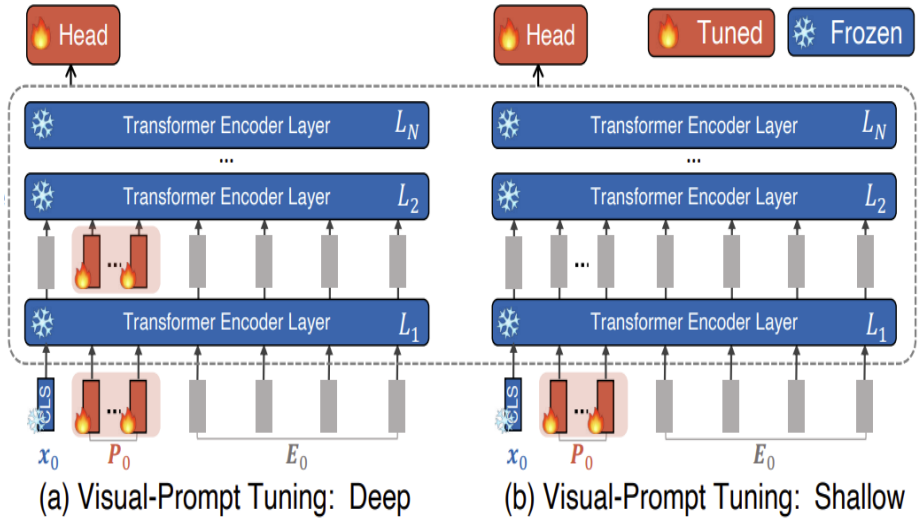
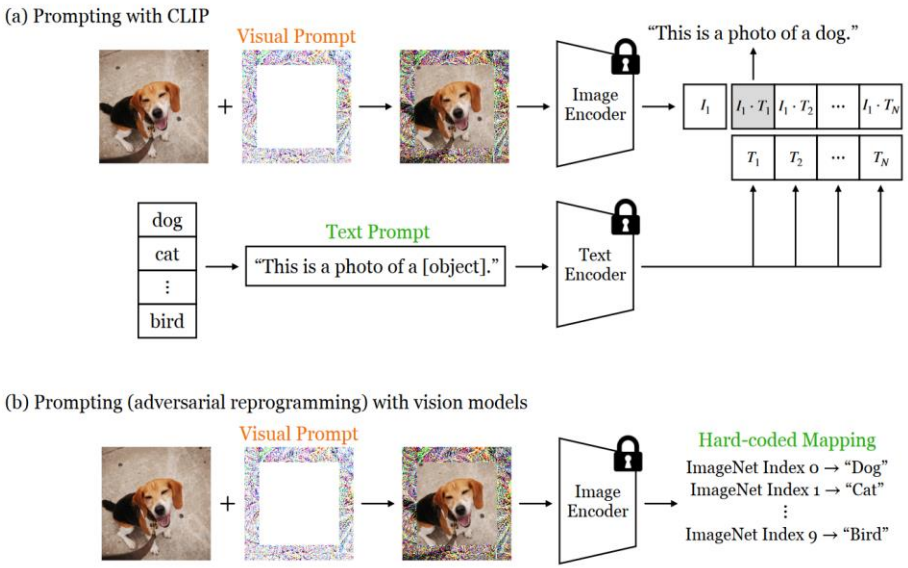


Figure 2. Our approach, Conditional Context Optimization (CoCoOp), consists of two learnable components: a set of context vectors and a lightweight neural network (Meta-Net) that generates for each image an input-conditional token.

研究背景

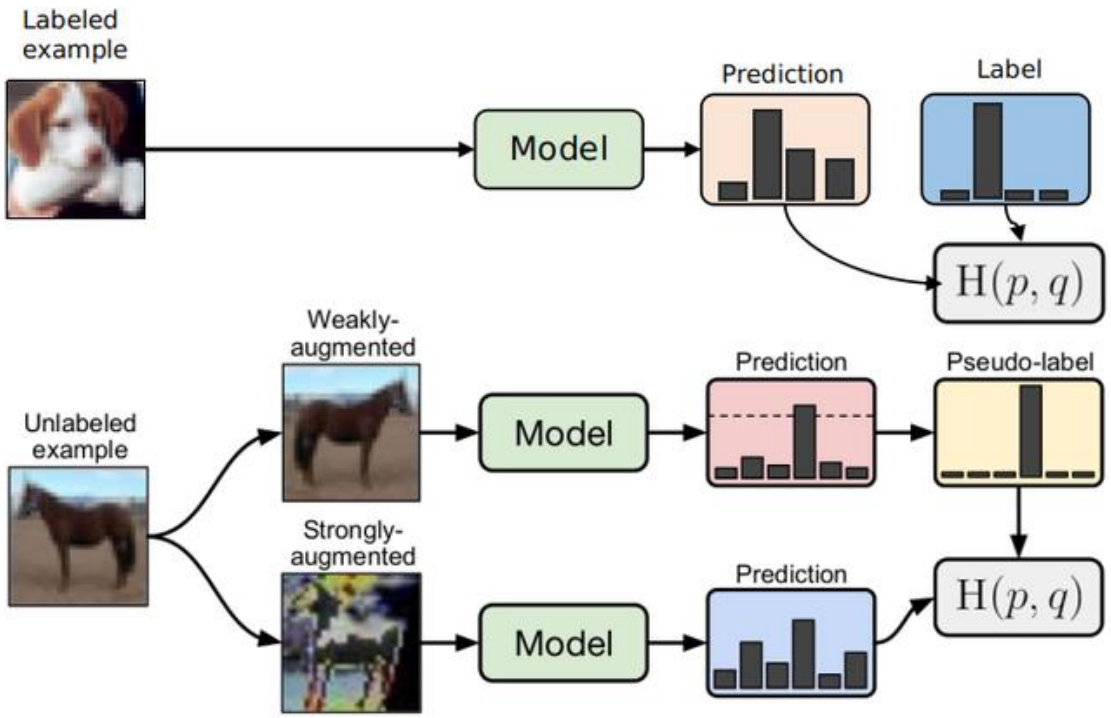
- 预训练大模型的微调方法
  - Fine-tuning
  - Parameter efficient fine-tuning（PEFT）
    - Adapter
    - LoRA
    - Prompt tuning : visual prompting

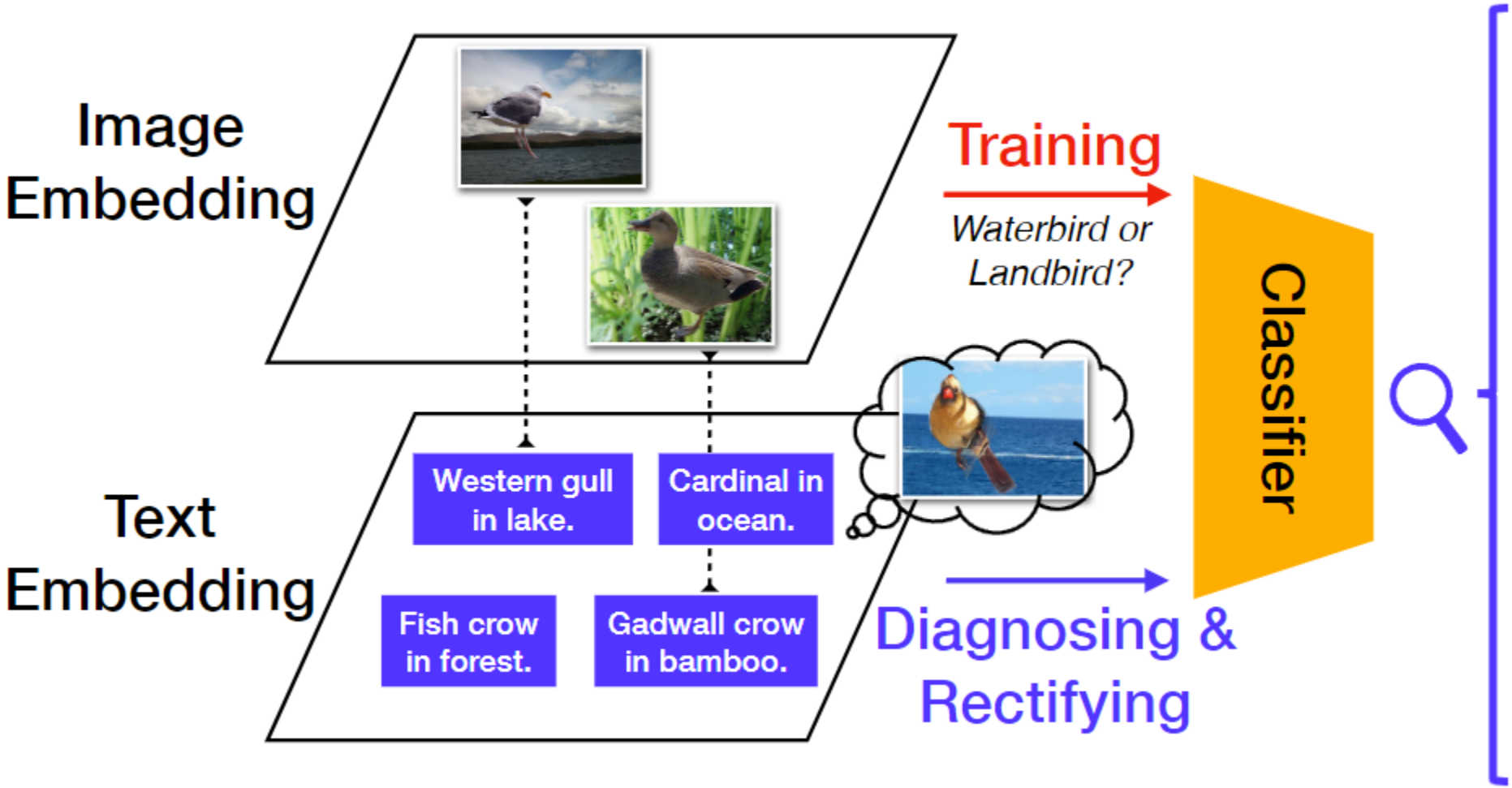




研究背景

- 半监督学习
  - Fixmatch

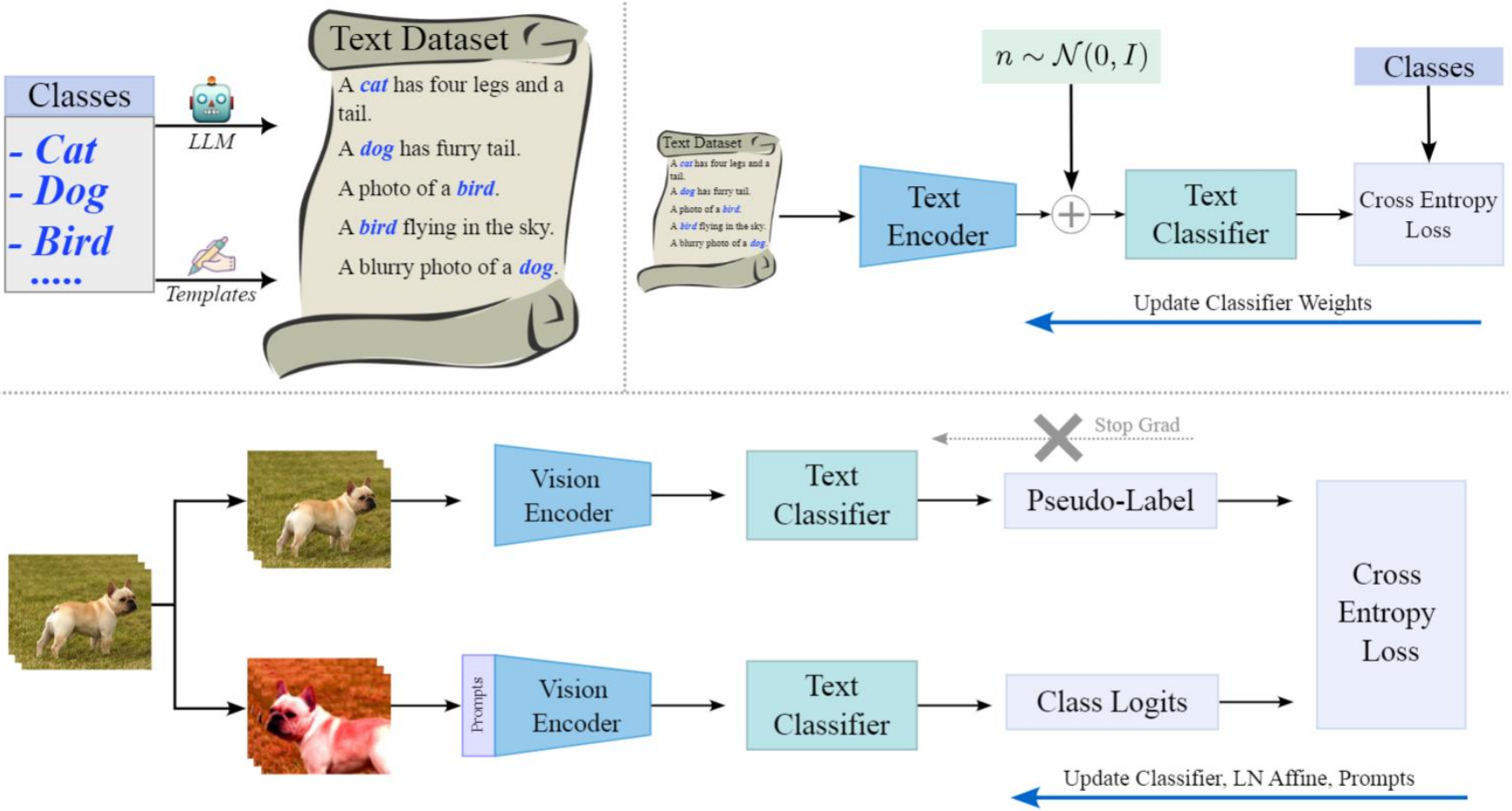




可以通过文本数据上微调零样本图像分类器

- 作者介绍
- 研究背景
- 本文方法
- 实验效果
- 总结反思

本文方法


Classes
- Cat
- Dog
- Bird
.....
LLM
Templates
Text Dataset
A cat has four legs and a tail.
A dog has furry tail.
A photo of a bird.
A bird flying in the sky.
A blurry photo of a dog.
$n \sim \mathcal{N}(0, I)$
Classes
Text Encoder
Text Classifier
Cross Entropy Loss
Update Classifier Weights
Stop Grad
Vision Encoder
Pseudo-Label
Prompts
Class Logits
Update Classifier, LN Affine, Prompts

本文方法

- step1: LLM构造数据集
  - 采用的LLM模型：GPT3
  - 50 descriptions for each category: GPT3生成+clip模板

- Describe what a category looks like.
- How can you identify a category?
- What does a category look like?
- Describe an image from the internet of a category.
- A caption of an image of a category?

- A quail is a small game bird with a rounded body and a small head.
- A quail is a small, plump bird with a round body and a short tail.
- A quail can be identified by its plump body, short legs, and small head with a pointed beak.
- A quail can be identified by its small, rounded body and short tail.
- A quail looks like a small chicken.
- A quail is a small, crested game bird.
- This image is of a quail in a natural setting.
- In the image, there is a brown and white quail perched on a branch.
- A quail hiding in some foliage.
- A young quail pecks at the ground in search of food.

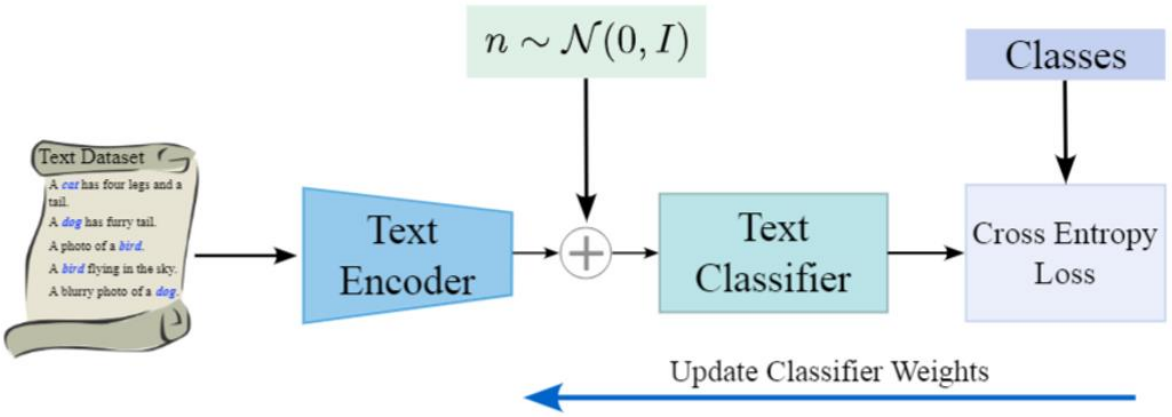
- a bad photo of the category.
- a category in a video game.
- a origami category.
- a photo of the small category.
- art of the category.
- a photo of the large category.
- itap of a category.

本文方法

- step2: 根据文本数据训练分类器

$$\min_{\theta} \sum_{\substack{(t,c)\in T \\ n\sim\mathcal{N}(0,I)}} \mathcal{L}_{\text{SCE}}\left(f_\theta\left(\frac{u(t)}{\|u(t)\|}+n\right),c\right), \tag{2}$$

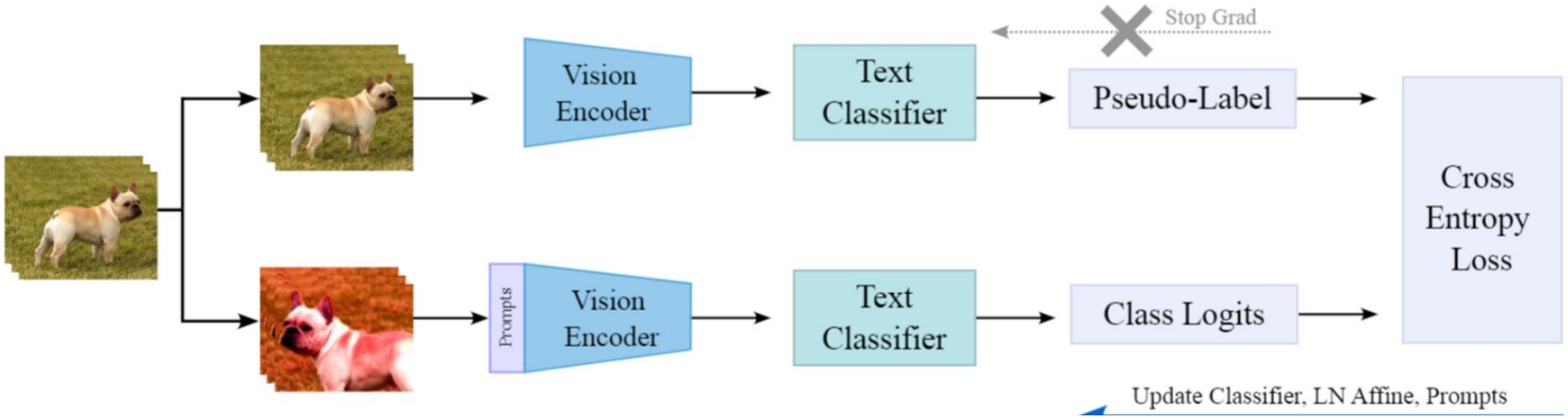
where θ is the parameter of the classifier. Training of the text-only classifier is very efficient. For example, 3000 epochs of training the classifier on the data set of 130000 text sentences, representing the 1000 classes of the ImageNet [53] dataset is completed in $\sim$ 120 seconds on an NVIDIA 3090 graphics card.



本文方法

- Step3: Unsupervised Finetuning on Target Domain Images
  - 伪标签生成

$$\hat{c}(x) = \arg\max_{c \in C} p_c, \quad \text{where} \quad p = f\Big(v\big(\alpha_w(x)\big)\Big).$$



The network is trained with the smoothed cross entropy loss $\mathcal{L}_{\text{SCE}}$ [52]. We denote the set of unlabelled target domain images by D, and formalize the training objective as

$$\min_{\theta,\eta} \sum_{x \in D} \mathcal{L}_{\text{SCE}}\Big(f_\theta\Big(v_\eta^p\big(\alpha_h(x)\big)\Big), \hat{c}(x)\Big), \tag{4}$$

where θ and η denote the finetuned parameters of the classifier and of the vision encoder, respectively. The parameters η are the visual prompts and the scale and shift (affine) parameters of the normalization layers of the vision encoder. The selection of η is motivated by keeping the adaptation *parameter-efficient*. For LaFTer, the number of trainable parameters are less than 0.4% of the entire model parameters, making adaptation extremely lightweight.

- 作者介绍
- 研究背景
- 本文方法
- 实验效果
- 总结反思

定量结果

	ImageNet	CIFAR-10	CIFAR-100	EuroSat	DTD	CALTECH-101
CLIP	61.9	88.8	64.2	45.1	42.9	90.5
CLIP-PR	60.4	89.3	63.2	44.2	40.1	84.8
UPL	61.2	89.2	65.8	62.2	**48.0**	90.6
LaFTer*	63.4	95.1	73.1	69.7	44.2	92.2
LaFTer	**64.2**	**95.8**	**74.6**	**73.9**	46.1	**93.3**
	UCF-101	**Flowers-102**	**SUN-397**	**ImageNet-A**	**ImageNet-S**	**ImageNet-R**
CLIP	61.0	66.6	60.8	29.6	40.6	65.8
CLIP-PR	57.9	57.7	54.7	11.6	38.6	54.1
UPL	63.9	**71.5**	**66.0**	26.9	42.4	65.6
LaFTer*	67.7	68.4	62.9	28.9	41.3	70.2
LaFTer	**68.2**	71.0	64.5	**31.5**	**42.7**	**72.6**

Table 1: Top-1 Classification Accuracy (%) while using the CLIP pre-trained ViT-B/32 backbone for 12 image classification benchmarks. We provide two versions of our method. LaFTer* represents all components of our methodology but without the *text-only* pre-training, whereas, LaFTer represents results obtained by first pre-training the visual classifier on text-only data and then performing unsupervised finetuning on the unlabeled image data. Highest accuracy is shown in bold, while second best is underlined.

定量结果

	ImageNet	CIFAR-10	CIFAR-100	EuroSat	DTD	CALTECH-101
LaFTer (no-shot)	64.2	95.8	74.6	73.9	46.1	93.3
CoOp (1-shot)	60.6	83.0	55.6	58.4	40.1	91.7
CoOp (5-shot)	61.3	86.6	63.2	71.8	41.1	93.2
CoOp (10-shot)	62.3	88.5	66.6	81.6	65.8	94.6
PEFT (1-shot)	50.7	62.7	50.2	37.5	42.6	90.6
PEFT (5-shot)	59.3	80.0	67.3	55.3	59.9	94.5
PEFT (10-shot)	62.8	87.9	74.1	67.9	67.3	96.1
	UCF-101	**Flowers-102**	**SUN-397**	**ImageNet-A**	**ImageNet-S**	**ImageNet-R**
LaFTer (no-shot)	68.2	71.0	64.5	31.5	42.7	72.6
CoOp (1-shot)	63.8	71.2	64.1	24.5	39.9	60.0
CoOp (5-shot)	74.3	85.8	67.3	30.0	46.5	61.6
CoOp (10-shot)	77.2	92.1	69.0	35.0	49.1	63.6
PEFT (1-shot)	60.5	66.9	58.3	20.9	38.5	57.2
PEFT (5-shot)	72.6	91.1	68.7	33.3	55.3	66.4
PEFT (10-shot)	79.8	95.2	72.3	40.2	61.1	71.0

Table 2: Top-1 Accuracy (%) for our LaFTer (no-shot) compared to few-shot methods. We compare to CoOp [30] in 1-, 5- and 10-shot supervised finetuning regimes. Parameter Efficient Finetuning (*PEFT*) represents tuning the same parameters as in LaFTer (prompts, classifier, affine) but in a few-shot manner. For each dataset/compared method, blue highlights the highest number of shots outperformed by *no-shot* LaFTer. Notably, LaFTer improves over 10-shot and all compared methods in 4 datasets, including ImageNet, where 10-shot = 10K labeled samples.

定量结果

	Extra Head	DTD [10]	CUB200 [48]	NABirds [22]	Dogs [28]	Flowers [40]	Food101 [3]	CIFAR100 [29]	CIFAR10 [29]	GTSRB [46]	SVHN [37]	Average
Data diversity	-	78.7	76.0	74.8	73.4	72.7	72.7	70.9	70.2	67.5	61.8	-
Fully-Tuning	✓	64.3	87.3	82.7	89.4	98.8	84.9	68.9	97.4	97.1	87.4	85.8
Linear	✓	63.2	85.3	75.9	86.2	97.9	84.4	63.4	96.3	68.0	36.6	75.7
Adapter [24, 42]	✓	62.7	87.1	84.3	89.8	98.5	86.0	74.2	97.7	91.1	36.3	80.8
VP [1]	✓	59.5	84.6	77.7	84.5	97.7	80.5	78.7	94.2	89.4	87.6	83.4
VPT [26]	✓	65.8	88.5	84.2	90.2	99.0	83.3	78.8	96.8	90.7	78.1	85.5
DAM-VP (10 epochs)	✓	72.4	86.3	81.5	92.2	98.6	86.2	80.1	90.5	87.8	81.1	85.7
DAM-VP (50 epochs)	✓	73.1	87.5	82.1	92.3	99.2	86.9	88.1	97.3	90.6	87.9	**88.5**

Table 4. Head-tuning adaption performance of different methods on ViT-B-22K, where we report image classification accuracy and all of baseline methods are trained for **100 epochs**.

	Extra Head	DTD [10]	CUB200 [48]	NABirds [22]	Dogs [28]	Flowers [40]	Food101 [3]	CIFAR100 [29]	CIFAR10 [29]	GTSRB [46]	SVHN [37]	Average
Data diversity	-	78.7	76.0	74.8	73.4	72.7	72.7	70.9	70.2	67.5	61.8	-
Fully-Tuning	✓	72.4	89.7	86.8	86.2	98.3	91.7	73.3	98.3	97.1	91.2	88.5
Linear	✓	73.6	88.6	85.2	85.9	99.4	88.2	61.6	96.3	83.8	43.5	80.6
Adapter [24, 42]	✓	73.9	88.5	84.6	86.8	98.9	88.7	85.7	96.5	83.6	71.3	85.9
VP [1]	✓	75.1	86.5	82.9	81.3	98.6	83.4	80.6	94.8	82.4	80.3	84.6
VPT [26]	✓	78.5	90.0	85.4	84.8	99.3	90.1	80.5	96.9	86.2	87.8	87.9
DAM-VP (10 epochs)	✓	77.0	89.4	86.8	88.3	99.6	90.2	85.5	96.4	84.7	79.0	87.7
DAM-VP (50 epochs)	✓	80.0	90.4	86.9	88.5	99.6	90.5	88.1	97.3	86.8	81.7	**89.0**

Table 5. Head-tuning adaption performance of different methods on Swin-B-22K, where we report image classification accuracy and all of baseline methods are trained for **100 epochs**.

消融实验

	IN	CIFAR-10	CIFAR-100	IN-A	IN-S	IN-R	Mean
Class Name	58.4	87.1	59.0	30.7	37.7	65.5	56.4
Simple-Template	61.1	88.1	63.1	30.4	40.3	65.9	58.1
Dataset-Templates	60.1	89.3	64.7	31.1	41.2	67.5	59.0
Llama Descriptions	54.8	88.4	59.4	25.7	36.3	58.7	53.9
GPT Descriptions	60.5	87.8	63.0	30.2	39.6	63.6	57.4
GPT + Templates	61.9	89.3	65.4	31.5	40.9	67.8	59.5

Table 3: Top-1 Classification Accuracy (%) for a ViT-B/32 model while ablating different text generation strategies for training the text-only classifier in the first stage of LaFTer . To obtain these results, we evaluate the test set of the respective datasets by using the text-only pre-trained classifier on top of the frozen vision encoder from CLIP. IN = ImageNet.

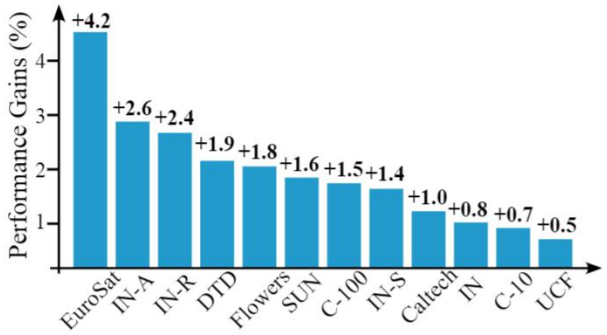


Figure 3: Performance gains of using the text-only pre-trained visual classifier vs not using it in LaFTer pseudo-labeling pipeline scheme.

	C-10	C-100	UCF	EuroSat	IN-R
CLIP (B/16)	89.2	68.1	64.7	48.4	73.8
LaFTer	96.5	76.3	67.2	72.1	81.4
CLIP (L/14)	95.3	75.8	72.0	60.3	84.9
LaFTer	99.0	87.2	77.2	77.2	91.5

Table 5: Top-1 Accuracy (%) for CLIP pre-trained ViT-B/16 and ViT-L/14 backbones. Results are provided for Base VL model (CLIP) and our LaFTer.

消融实验

	Clip	w/o Aug	w/o Prompts	w/o Affine	w/o Cls	w/o Stop Grad	LaFTer
CIFAR-10	88.8	93.5	92.5	94.6	94.1	94.7	95.8
CIFAR-100	64.2	72.6	71.7	72.9	67.4	73.2	74.6
UCF-101	61.0	65.3	66.1	63.6	63.4	67.5	68.2
EuroSat	45.1	63.2	61.2	64.2	60.9	69.2	73.9
ImageNet-A	29.6	29.9	29.8	30.8	30.1	31.1	31.5
ImageNet-S	40.6	40.3	40.7	41.1	40.9	42.1	42.7
ImageNet-R	65.8	68.0	67.7	66.8	66.1	71.8	72.6
Average	56.4	61.8	61.4	62.0	60.4	64.2	65.6

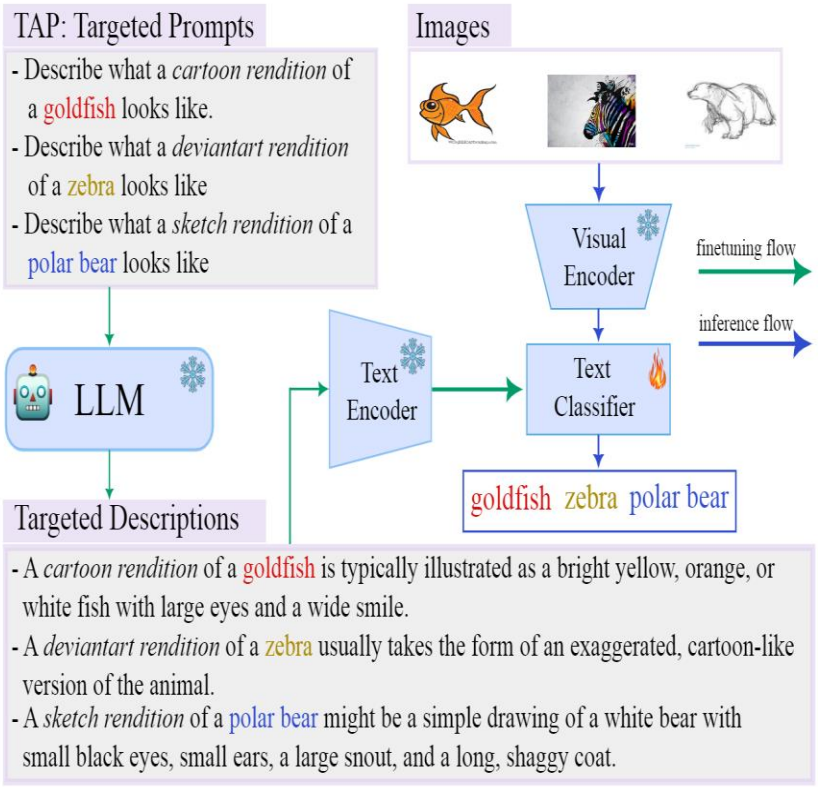
Table 4: Top-1 Accuracy (%) for our LaFTer while ablating the various critical design choices in our methodology. For each of these experiments, we disable one component from our framework and test the resulting method. Aug: Augmentations, Cls: Classifier.

- 作者介绍
- 研究背景
- 本文方法
- 实验效果
- 总结反思

总结反思

- 总结
  - 更好利用图文匹配的方法 --- 文本端优化、视觉端受益
- 思考
  - 任务设定

总结反思



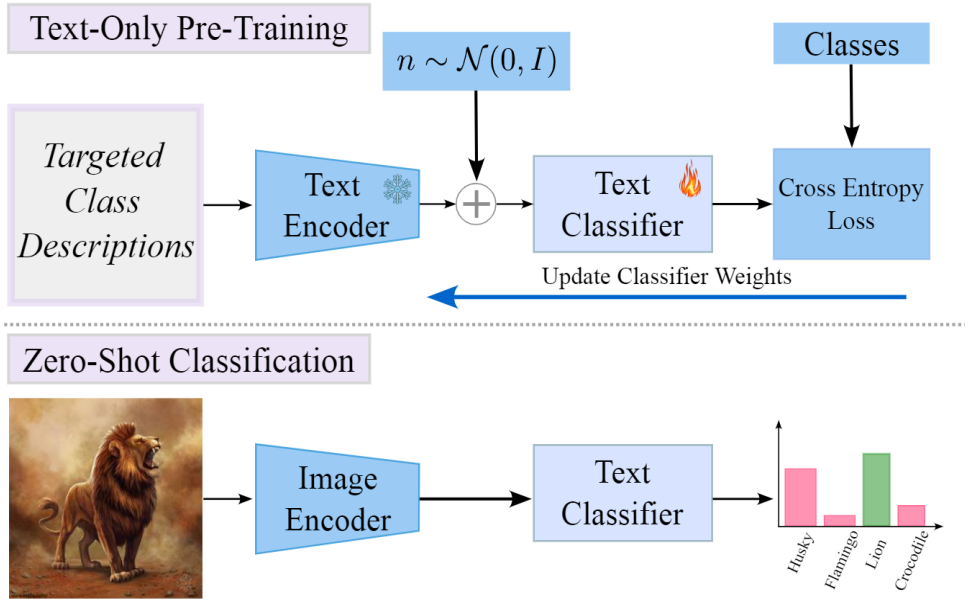


Fig. 2: The targeted descriptions generated by the LLM are automatically matched with the class names to generate the text dataset. This dataset is used to train a text classifier in a *supervised* setting (top). The trained text classifier is used to classify the visual data (bottom).

Thanks for Attention!