

LAMM: Label Alignment for Multi-Modal Prompt Learning

选自2024年AAAI会议

主讲人：余萍
2024. 3. 13



目录

CONTENTS

1

背景引入

2

论文主体方法

3

实验

01 背景引入



背景引入 CLIP介绍

想象一个图像分类的场景：利用图像数据集训练模型做分类任务，假设现在要训练100个类别，请问在训练完后，此时的模型可以对新出现的第101个类别进行分类吗？

传统模型

不能分类新的类别，因为一般在模型最后都有一层全连接或者softmax,这样一来就把可以分类的类别固定死了（这个操作会把之前100个类别的可能性映射到和为1的概率值上，第101个类别的概率根本不会出现），如果要想模型认识第101个类别，就要重新训练。

VS

CLIP模型

CLIP，全称Contrastive Language-Image Pre-Training，具有较强的迁移能力（泛化能力），和GPT一可以做到zero-shot，在没有学习新类别的情况下也能对其进行分类。OpenAI有提供预训练模型，但没有给源码，网上已经有山寨版的源码了。

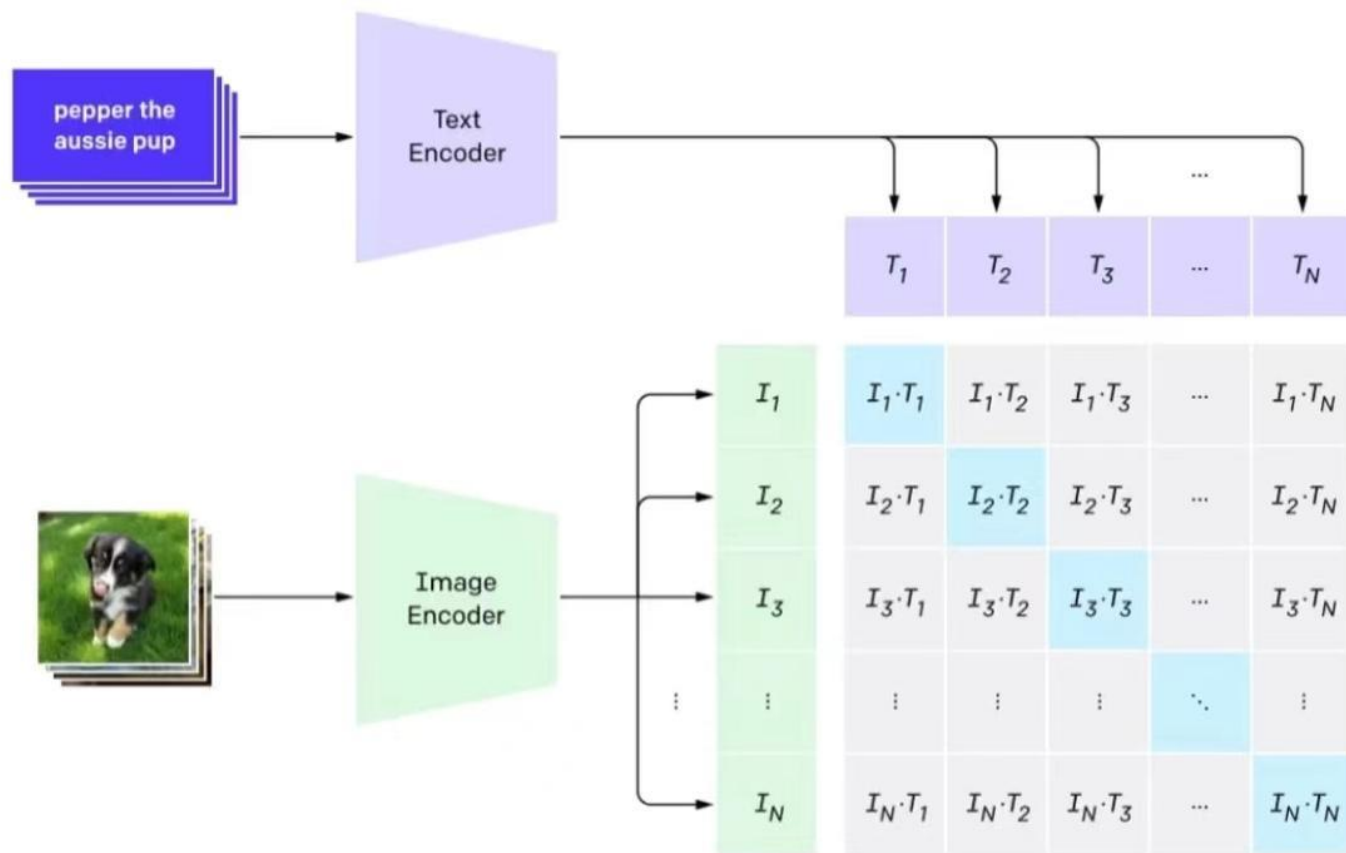
能不能一劳永逸呢？



背景引入 CLIP介绍

如何训练？

1. 使用了4亿对文本图像数据进行训练
2. 两个Encoder：一个文本编码器（提取文本特征），一个图像编码器（提取图像特征）
3. 计算两个特征的余弦相似度
4. 最大化正样本的概率：对角线为正样本，其他为负样本



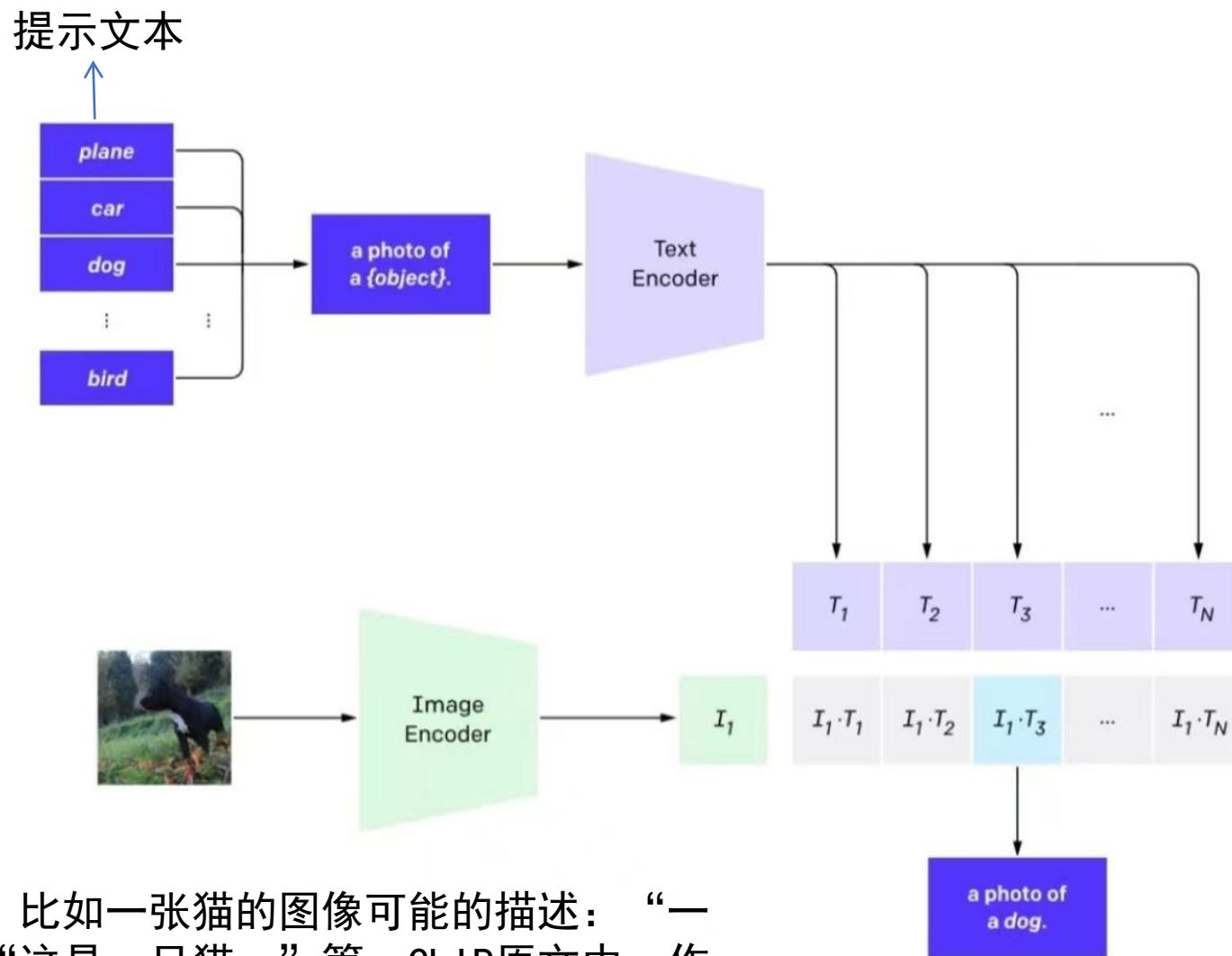


背景引入 CLIP介绍

如何推理？

1. 计算现有图像向量与所有文本向量的相似度，相似度最高的就是分类结果

2. 文本内容：需要给提示文本进行一些修饰和完善，如果提示文本写得不好，那么效果就会差强人意



例如：一张图其实可能有多种描述，比如一张猫的图像可能的描述：“一个猫。”，“一张猫的照片。”，“这是一只猫。”等。CLIP原文中，作者发现了这个问题，并且发现prompt对预测性能的影响较大。



背景引入

论文中涉及的其他模型简介

CoOp 是在CLIP上的改进，提出了一种可学习的prompt，让模型直接学习一个最优的prompt

<https://zhuanlan.zhihu.com/p/492546332>

CoCoOp 是对CoOp的提升，主要提高了学习后的prompt的泛化能力

<https://zhuanlan.zhihu.com/p/493354342>

MaPLE 在CoCoOp上的改进，实现了多模态prompt，不仅文本有prompt，图像也有

<https://zhuanlan.zhihu.com/p/622959116>



背景引入

可以看到，自于自然语言处理（NLP）的提示式调优范式在VL（visual-language）领域取得了重大进展。但是，之前的方法主要集中于构建文本和视觉输入的提示模板，而忽略了VL模型与下游任务之间的类标签表示上的差距。



举例：在训练模型时使用了一些文本标签，但在将其应用于图像分类时，可能需要不同的标签表示方式，比如更具体的类别或不同的标准。如果在这个过程中存在标签表示的不匹配，就可能导致性能下降或不准确的结果。

为了解决这个问题，本文提出了一种创新的标签对齐方法LAMM，以及分层损失函数。作者在11个下游视觉数据集上进行了实验，证明了该方法显著提高了现有的多模态提示学习模型在少样本训练场景中的性能

02 论文主体方法



论文主体方法

文本模板中的<类>标记在将图像分类为适当的类别中至关重要。例如，如图1所示，美洲驼和羊驼是两种非常相似的动物。在CLIP中，由于训练前数据集中羊驼数据的过度代表，存在将美洲驼误分类为羊驼的倾向。

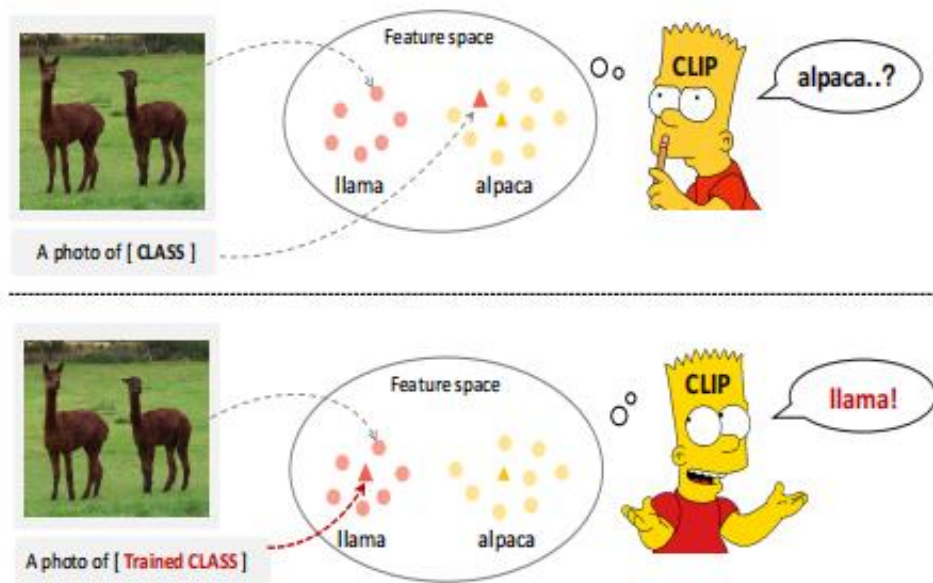


Figure 1: CLIP is more inclined to classify an image as belonging to a similar category. Altering the category text feature's position can enhance CLIP's recognition capabilities.



论文主体方法

主要贡献点:

- 1、引入了一种名为LAMM的标签对齐技术，它通过梯度优化自动搜索最优的<类>嵌入。实现优化下游数据集中不同类别的类别嵌入，以增加每个图像与其对应的类别描述之间的相似性。
- 2、引入了一个分层损失，以防止整个提示模板的语义特征偏离太远

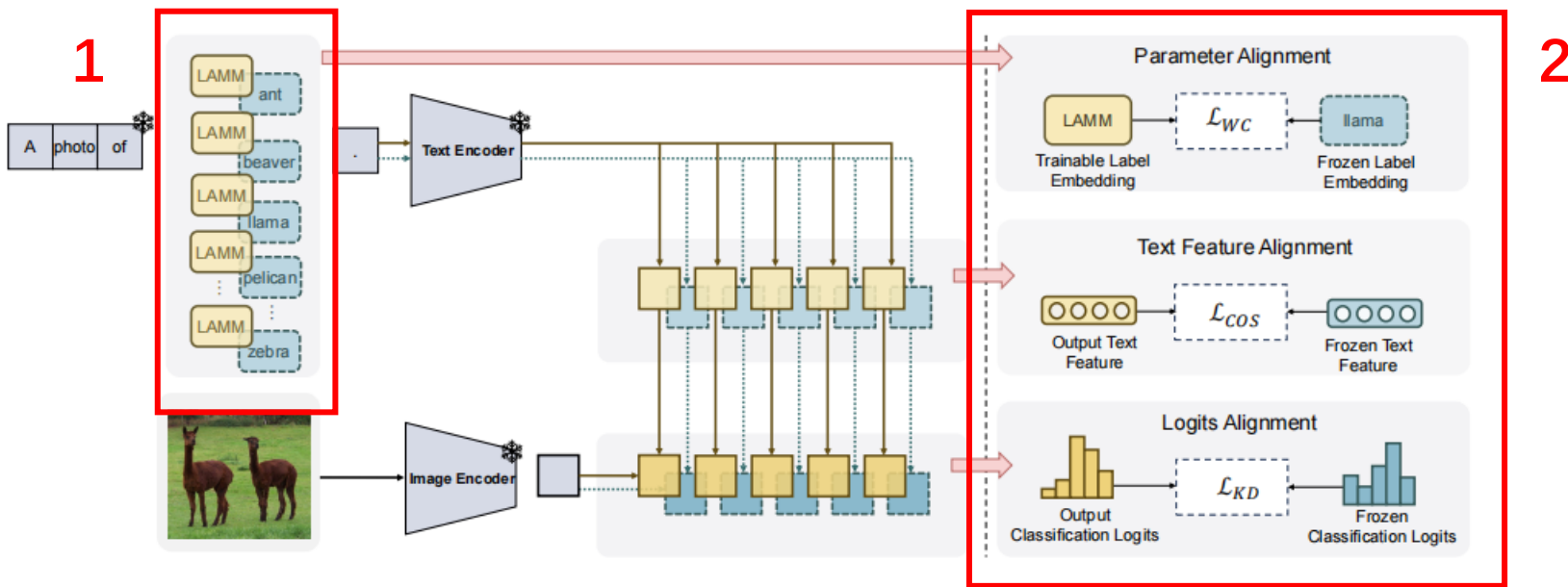


Figure 2: The whole architecture of LAMM. We replace the category tokens in the downstream dataset with trainable vectors and incorporate a hierarchical loss to preserve the CLIP's generalization ability of each category. The gray boxes represent the prompt template and frozen model, the blue boxes indicate the original label embeddings/features/logits, and the yellow ones denote the label embeddings/features/logits during training.




实现标签对齐

最基础的损失函数：

$$p(y = i | I) = \frac{\exp(\cos(I_x, \psi(y_i)) / \tau)}{\sum_{j=1}^k \exp(\cos(I_x, \psi(y_j)) / \tau)} \quad (1)$$

where τ is a temperature parameter acquired by CLIP, while function \cos represents cosine similarity.

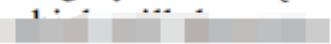
改变提示文本后最基础的损失函数：

 In this way, the prompt template of LAMM converts to:

$$\text{第}i\text{个类别的可学习标签} \quad z_i = [a][photo][of][< M_i >][.] \quad (2)$$

where $< M_i >$ ($i=1, 2, \dots, k$) represents a learnable token of the i -th category. Similar to Equation 1, the prediction probability of LAMM is computed as:

$$p(y = i | I) = \frac{\exp(\cos(I_x, \psi(z_i)) / \tau)}{\sum_{j=1}^k \exp(\cos(I_x, \psi(z_j)) / \tau)} \quad (3)$$

During training, we only update the category vectors $\{< M_i >\}_{i=1}^k$ in each downstream dataset, 

图像特征向量

文本特征向量

温度系数的作用是调节对困难样本的关注程度：越小的温度系数越关注于将本样本和最相似的其他样本分开)

https://blog.csdn.net/qq_27590277/article/details/120620493

以往对文本模板的提示式调优工作主要集中于“一张照片”的训练，而忽略了“<类>”的优化



分层损失函数

为了防止在少样本训练情况下的过拟合，作者提出了分层损失（Hierarchical Loss），由三部分组成 Parameter Space、Feature Space 和 Logits Space

Parameter Space

(WC) (Kirkpatrick et al. 2017) loss is employed as follows:

$$\mathcal{L}_{WC} = \sum_i (\theta_i - \bar{\theta}_i)^2 \quad (5)$$

where θ is the trainable parameters of the current model, and $\bar{\theta}$ is the reference ones. In LAMM, θ represents the trainable label embeddings, whereas $\bar{\theta}$ represents the original label embeddings.

为了将下游任务中真实的类标签
与可学习标签对齐

Feature Space

In this approach, for a category template z_i in Equation 2, the original prompt template y_i from Equation 1 serves as the center of its optimization region. The cosine loss is formulated as follows:

$$\mathcal{L}_{COS} = \sum_i 1 - \cos(\psi(z_i), \psi(y_i)) \quad (6)$$

为了缓解每个类别的文本特征的
过拟合，采用文本特征对齐损失来限制
文本特征的优化区域（减少对非必要
图片区域特征的关注）



论文主体方法


Logits Space

we introduce a knowledge distillation loss in the classification logits space, which allows for the transfer of generalization knowledge from CLIP to LAMM. The distillation loss can be formulated as follows:

$$\mathcal{L}_{\text{KD}} = - \sum \cos(I_x, \psi(z_i)) \log(\cos(I_x, \psi(y_i))) \quad (7)$$

损失的目标是LAMM模型可以学到如何调整其对数分布，使其更接近CLIP模型的对数分布，相当于在预训练的CLIP模型中蕴含的泛化知识被传递给了LAMM模型

cross-entropy Loss

 To train the LAMM for downstream tasks, cross-entropy (CE) loss is applied to the similarity score as same to finetuning the CLIP model:

$$\mathcal{L}_{\text{CE}} = \frac{1}{N} \sum_{i=1}^N \text{CE}(\tau \cdot \cos(I_x, \psi(z_i)), y_i) \quad (8)$$

预训练学习到的

交叉熵（CE）损失应用于相似度评分，和对CLIP模型进行微调



Total Loss

where $\lambda_1, \lambda_2, \lambda_3$ are hyper-parameters. In this way, the total loss is:

$$\mathcal{L} = \mathcal{L}_{CE} + \lambda_1 * \mathcal{L}_{WC} + \lambda_2 * \mathcal{L}_{COS} + \lambda_3 * \mathcal{L}_{KD} \quad (9)$$

where $\lambda_1, \lambda_2, \lambda_3$ are hyper-parameters. To prevent the redundancy of adjusting parameter, we set $\lambda_1 = 1/n$, $\lambda_2 = 1$, $\lambda_3 = 0.05$ for all our experiments empirically, where n represents the number of training shots.

训练样本越少，偏离的惩罚就越大

03 实验





实验

Datasets (11个)

Caltech101、ImageNet、OxfordPets、StanfordCars、Flowers102、Food101、FGVCAircraft、SUN397、UCF101、DTD、EuroSAT

实验设置

- 使用1、2、4、8和16张图像对模型进行训练，所有实验结果均为在种子1、2、3上进行实验所得结果的平均值
- 基线模型：zero-shot CLIP、CoOp、MaPLe、LAMM+CoOp和LAMM+MaPLe
- 实验中的所有提示模板都是从“<类>的照片”初始化的
- 在每个模型的原始设置中保持相同的训练参数（如学习率、epoch和其他提示参数）
- 相应的超参数在所有数据集中都是固定的



实验

与最先进的方法比较

这表明，标签嵌入比预先训练的模型到下游任务的提示模板更重要。此外，随着训练样本数的增加，用LAMM观察到的改善变得更加明显。同时与LAMM结合后，CoOp和MaPLE的表现提升，表明LAMM在少数显示场景中可以显著提高现有模型的性能

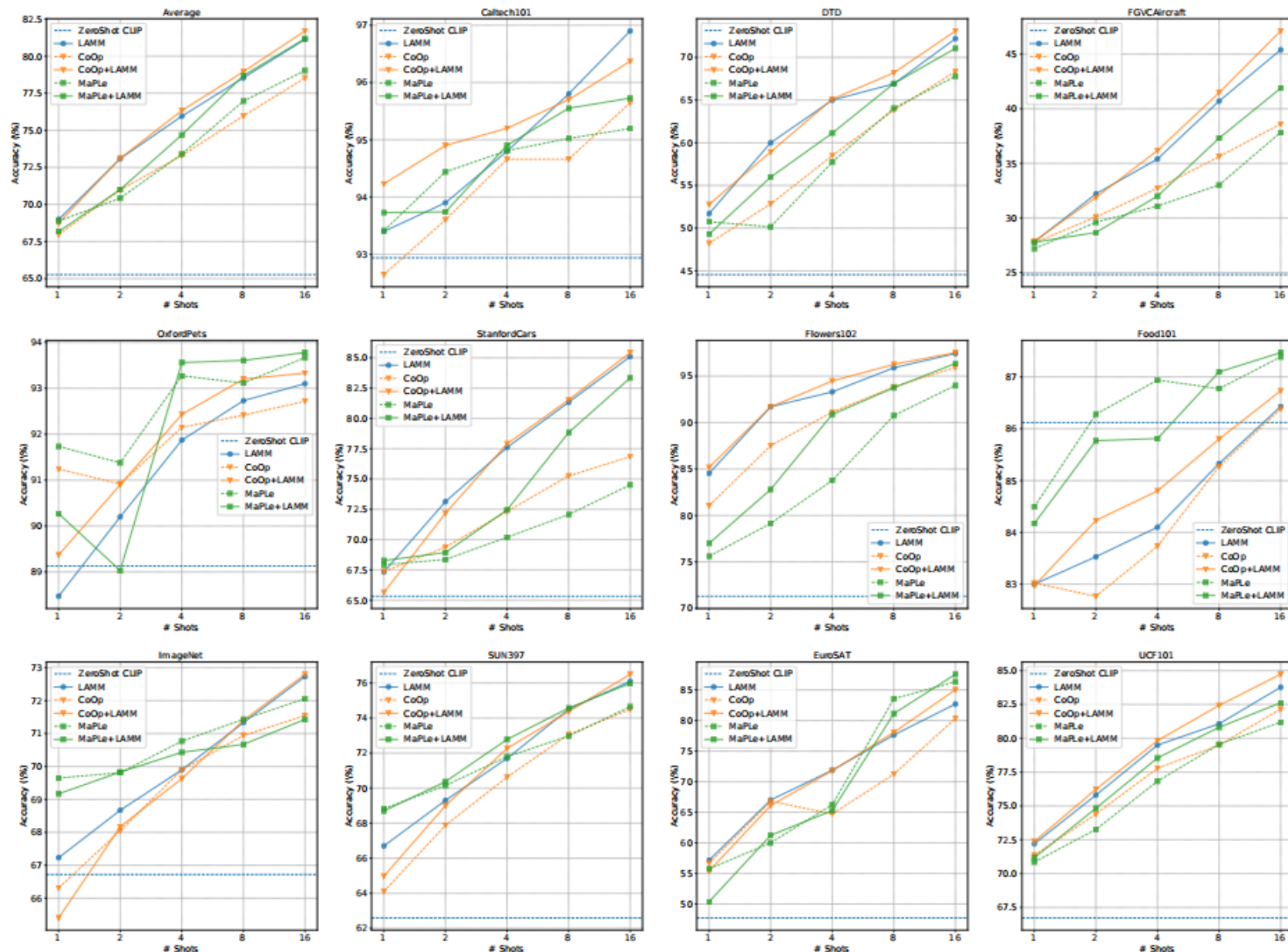


Figure 3: Main results over 11 datasets under the few-shot learning setting. We report the average accuracy (%) of 1/2/4/8/16 shots over three runs. Overall, the proposed LAMM enhances the performance of CLIP, CoOp, and MaPLE.



跨数据集泛化能力比较

Method	Source ImageNet	Target		Average
		-V2	-Sketch	
CLIP	66.73	60.83	46.15	57.90
CoOp	71.51	64.20	47.99	61.23
CoCoOp	71.02	64.07	48.75	61.28
MaPLe	70.02	64.07	49.15	61.08
LAMM	72.73	65.13	48.11	61.99

Table 1: Comparison of LAMM with existing methods in domain generalization setting. LAMM shows highest performance on average.

在ImageNet上训练，在ImageNetV2和Imagenet-Sketch上评估。

评估数据集与训练集具有相同的类别，但这三个数据集在域分布上有所不同。

假设训练一个图像分类模型，用于识别猫和狗。训练数据集包含来自家庭、宠物店和动物园的图像，其中猫和狗的照片均匀分布。然而，在评估数据集中，由于特殊原因，大部分照片都来自宠物店，而几乎没有来自动物园的照片。在这种情况下，虽然训练集和评估集中都包含相同的猫和狗类别，但它们的领域分布存在差异。模型可能在宠物店环境下表现得更好，而在动物园环境下可能表现不佳，因为它在训练中更多地接触到了宠物店领域的样本



实验

增量学习的效果

仿照MaPLe，作者将数据集划分为基类和新类，将基类指定为set1，将新类指定为set2。

由于LAMM在保留现有类嵌入的同时只操作新类的嵌入，因此在对新类的持续训练中，LAMM在以前类上的性能保持稳定。相反，在CoOp和MaPLe的情况下，当需要学习的类别数量增加时，如果不对所有类别从头开始进行再训练，模型的表现将显著下降。

Method	Subset	ImageNet	Catech101	DTD	FGVCAircraft	StanfordCars	Flowers102	OxfordPets	Food101	EuroSAT	UCF101	SUN397	Average	Degradation
Zero-shot CLIP	Set1	72.43	96.84	53.24	27.19	63.37	72.08	91.17	90.10	56.48	70.53	69.36	69.34	69.34 (0)
	Set2	68.14	94.00	59.90	36.29	74.89	77.80	97.26	91.22	64.05	77.50	75.35	74.22	
CoOp	Set1	71.77	97.07	54.63	23.40	61.63	61.97	93.63	87.10	72.37	71.03	72.53	69.74	82.69 (-12.95)
	Set2	73.67	96.37	76.57	54.53	87.07	97.50	97.70	91.47	90.27	88.13	83.77	85.18	
MaPLe	Set1	74.37	97.10	70.83	30.27	65.77	77.53	95.03	90.6	82.73	78.17	77.60	76.36	82.28 (-5.92)
	Set2	74.13	96.50	77.40	53.73	83.27	97.00	98.17	92.27	92.60	88.33	83.95	85.21	
LAMM	Set1	77.23	98.40	83.30	43.27	81.63	97.80	95.17	89.83	90.60	86.17	82.13	84.14	84.14 (0)
	Set2	74.57	96.03	79.47	61.47	91.97	98.33	97.93	91.73	91.80	89.23	84.80	87.03	

Table 2: Comparison of LAMM and other prompting methods on 16-shot classes incremental learning. Initially, models are trained on Set 1, followed by training on Set 2. The contents within the parentheses of term "Degradation" refers to the decline in evaluation results on Set 1 subsequent to further training on Set 2.

分了两次训练，（）中表示在更过set2训练后对比只经过set1训练的变动情况



实验

消融实验

去除整个分层损失

Method	1-shot	2-shot	4-shot	8-shot	16-shot
Zero-shot CLIP	65.27	65.27	65.27	65.27	65.27
LAMM(w/o HL)	61.39	66.8	72.13	76.21	79.93
LAMM	68.99	73.09	75.95	78.54	81.13
CoOp	67.97	70.98	73.3	75.97	78.53
CoOp+LAMM(w/o HL)	61.23	66.33	72.06	76.19	79.94
CoOp+LAMM	68.74	73.11	76.33	78.95	81.71
MaPLe	68.88	69.22	73.4	76.97	79.03
MaPLe+LAMM(w/o HL)	61.25	67.70	73.39	77.64	80.74
MaPLe+LAMM	68.17	71.01	74.69	78.69	81.18

Table 3: Comparison of with or without hierarchical loss among vanilla CLIP, CoOp, and MaPLe.

在引入损失后，结果显示出显著的差异，证明分层损失对提高LAMM的性能是非常必要的。此外，当涉及的训练样本较少时，分层损失所带来的改进程度变得更加明显。这是由于微调标签与少量样本对齐，特别是只有单个图像时，会导致标签对图像内的其他噪声信息过拟合，而分层损失函数会抑制过拟合的情况



实验

分别去除分层损失的每个部分

基线

\mathcal{L}_{CE}	\mathcal{L}_{WC}	\mathcal{L}_{COS}	\mathcal{L}_{KD}	Average
✓				79.93
✓	✓	✓	✓	81.13
✓		✓	✓	80.98
✓	✓		✓	80.86
✓	✓	✓		81.08
✓			✓	80.82
✓		✓		80.75
✓	✓			80.63

不同
损失
组合
的
结
果

Table 4: Ablations on hierarchical loss function on 16-shot.

$$\mathcal{L} = \mathcal{L}_{CE} + \lambda_1 * \mathcal{L}_{WC} + \lambda_2 * \mathcal{L}_{COS} + \lambda_3 * \mathcal{L}_{KD} \quad (9)$$

这三种损失对最终结果都有积极的贡献



类别词初始化和随机初始化比较

类别词初始化使用每个类别的原始词嵌入作为初始化的类别嵌入，而随机初始化则使用随机的初始化类别嵌入

Setting	Random Words	
1-shot	62.71	68.99
2-shot	62.71	73.09
4-shot	71.68	75.95
8-shot	76.65	78.54
16-shot	80.71	81.13

Table 5: Comparision of initializations from random and category words.



不同神经网络架构上的效果

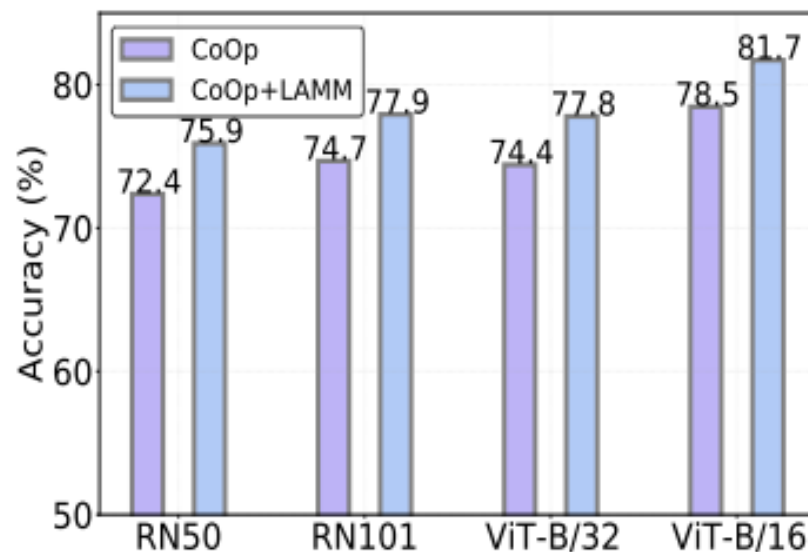


Figure 4: Ablations among different vision backbones

表明在不同网络架构下，模型与LMM都可以得到提升，进一步表明了LMM的实用性和泛化性



西南财经大学
SOUTHWESTERN UNIVERSITY OF FINANCE AND ECONOMICS

Questions and Discussions

主讲人：余萍
2024. 3. 13