

Noname 手稿编号  
(将由编辑插入)

# 学习提示视觉语言模型

Kaiyang Zhou - Jingkang Yang - Chen Change Loy - Ziwei Liu

收到： 日期 / 接受： 日期

**摘要** 像 CLIP 这样的大型预训练视觉语言模型在学习可在各种下游任务中转移的代表符号方面显示出巨大的潜力。与主要基于离散化标签的传统代表学习不同，视觉语言预训练将图像和文本统一在一个共同的特征空间中，这样就可以通过提示零距离转移到下游任务中，即从描述兴趣类别的自然语言中合成分类权重。在这项工作中，我们发现在实践中部署这种模式的一个主要挑战是提示工程，这需要领域专业知识，而且非常耗时--人们需要花大量时间进行词语调整，因为措辞上的细微变化都可能对性能产生巨大影响。受自然语言处理（NLP）中提示学习研究的最新进展的启发，我们提出了 *语境操作化* (CoOp)，这是一种专门用于调整类似 CLIP 的视觉语言模型以进行下游图像识别的简单方法。具体来说，CoOp 利用可学习向量对提示的上下文词语进行建模，而整个预训练参数则保持不变。为了应对不同的图像识别任务，我们提供了两种

CoOp 的实现：统一上下文和特定类上下文。通过在 11 个数据集上进行的大量实验，我们证明了 CoOp 只需要一到两个镜头，就能以相当大的优势击败手工制作的提示语，并且能够在更多镜头的情况下显著改善提示语工程，例如，在 16 个镜头的情况下，平均增益约为 15%（最高超过 45%）。尽管 CoOp 是一种基于学习的方法，但与使用手工制作提示语的零镜头模型相比，CoOp 实现了极佳的领域泛化性能。

## 1 引言

建立最先进的视觉识别系统的一种常见方法是训练视觉模型，以便使用离散拉贝尔对一组固定的物体类别进行预测（He 等人，2016 年；Dosovitskiy 等人，2021 年）。从技术角度看，这是通过将 ResNet（He 等人，2016 年）或 ViT（Dosovitskiy 等人，2021 年）等视觉模型生成的图像特征与一组固定的权重相匹配来实现的，这些权重被视为视觉概念并随机初始化。虽然训练类别通常有文本形式，如“金鱼”

周开阳

新加坡南洋理工大学 S-Lab 电子邮箱：

kaiyang.zhou@ntu.edu.sg

杨景康

新加坡南洋理工大学 S-Lab 电子邮箱：

jingkang001@ntu.edu.sg

arXiv:2109.01134v6 [cs.CV] 2022年10月

Chen Change Loy

新加坡南洋理工大学 S-Lab 电子邮箱:

ccloy@ntu.edu.sg

Ziwei Liu

新加坡南洋理工大学 S-Lab 电子邮箱:

ziwei.liu@ntu.edu.sg

或 "卫生纸", 它们会被转换成离散的词, 只是为了简化交叉熵损失的计算, 而文本中包含的语义在很大程度上未被利用。这种学习范式将视觉识别系统局限于封闭的视觉概念, 使其无法处理新的类别, 因为学习新的分类器需要额外的数据。最近, 视觉语言预训练 (如 CLIP ([Radford 等人, 2021 年](#)) 和 ALIGN ([Jia 等人, 2021 年](#))) 已成为视觉表征学习的一种有前途的替代方法。其主要理念是将





图 1 提示工程与上下文优化 (CoOp) 的对比。前者需要使用保留的验证集来进行单词调整，效率较低；而后者可自动完成这一过程，只需要几张标注过的图像即可进行学习。

CLIP 和 ALIGN 的学习目标都是对比损失 (contrastive loss)。例如，CLIP 和 ALIGN 都将学习目标表述为对比损失，它将图像及其文字描述结合在一起，同时推掉特征空间中不匹配的对。通过大规模的预训练，模型可以学习各种视觉概念，并通过提示随时转移到任何下游任务中 (Radford 等人，2021 年；Jia 等人，2021 年；Fu 等人，2021 年；Li 等人，2021 年；Singh 等人，2021 年；Yuan 等人，2021 年)。特别是，对于任何新的分类任务，我们可以首先通过向文本编码器提供描述任务相关类别的句子来合成分类权重，然后与图像编码器生成的图像特征进行比较。

我们注意到，对于预训练的视觉语言模型来说，文本输入（即提示）在下游数据集中发挥着关键作用。然而，确定正确的提示语并非易事，往往需要花费大量时间进行文字调整--措辞上的细微变化可能会对性能产生巨大影响。例如，对于 Caltech101 (图 1(a)，第 2 次提示与第 3 次提示)，在班级名称前添加 "a" 会使准确率提高 5% 以上。更重要的是，提示工程还需要有关任务的先验知识，最好是语言模型的下卧机制。图 1(b-d) 就是一个很好的例子，

添加与任务相关的上下文可以显著提高准确率，如 Flowers102 中的 "flower"、DTD 中的 "tex-ture" 和 EuroSAT 中的 "satellite"。调整句子结构可带来进一步的改进，例如，在 Flowers102 的类标记后加上 "一种花"，在 DTD 的上下文中只保留 "纹理"。

由此产生的提示绝不保证是这些下游任务的最佳选择。

受最近自然语言处理（NLP）中的提示学习研究（Shin 等人，2020；Jiang 等人，2020；Zhong 等人，2021）的启发，我们提出了一种称为上下文优化（CoOp）的简单方法。<sup>1</sup>的简单方法，特别是针对预训练的视觉语言模型。具体来说，CoOp 用可学习的向量对提示语的上下文词进行建模，这些向量可以用随机值或预先训练好的词嵌入进行初始化（见图 2）。我们提供了两种实现方法来处理不同性质的任务：一种是基于统一语境，它与所有类别共享相同的语境，在大多数类别中都能很好地发挥作用；另一种是基于特定类别的语境，它为每个类别学习一组特定的 DTD，并在“卫星照片”前添加“居中”

欧洲通信卫星组织。不过，即使进行了大量调整，CoOp 的发音仍为 /ku:p/。<sup>1</sup>CoOp 读作 /ku:p/。

语境标记，在某些细粒度类别中更为适用。在训练<sup>3</sup>过程中，我们只需使用与可学习上下文向量相关的交叉熵损失来最小化预测误差，同时保持整个预训练参数固定不变。梯度可以一直反向传播到文本编码器中，提炼出编码在参数中的丰富知识，用于学习与任务相关的上下文。

为了证明 CoOp 的有效性，我们在 11 个数据集上进行了基准测试，这些数据集涵盖了多种视觉识别任务，包括对一般物体、场景、动作和细粒度类别的分类，以及识别文本和卫星图像等专门任务。研究表明，CoOp 能有效地将预先训练好的视觉语言模型转化为数据效率高的视觉学习器，只需一两个镜头就能以相当大的优势击败手工制作的提示。性能还可以进一步提高

例如，在使用 16 次提示时，与手工制作的提示相比，平均差距约为 15%，最高差距超过 45%。CoOp 还优于线性探针模型，后者是众所周知的强少镜头学习基线（Tian 等人，2020 年）。此外，尽管 CoOp 是一种基于学习的方法，但它比零点模型（使用人工提示）具有更强的抗领域变化能力。

总之，我们做出了以下贡献：

1. 我们及时提交了一份研究报告，内容涉及在下行应用中调整最近提出的视觉语言模型，并确定了与部署效率相关的一个关键问题，即及时工程。
2. 为了使提示工程自动化，特别是针对预训练的视觉语言模型，我们提出了一种基于连续提示学习的简单方法，并提供了两种可处理不同识别任务的实现方法。
3. 我们首次证明，对于大型视觉语言模型而言，基于提示学习的方法在下游转移学习性能和领域转移下的鲁棒性方面优于手工制作的提示和线性探测模型。
4. 我们在 <https://github.com/KaiyangZhou/CoOp> 上开源了我们的项目。

我们希望这些研究成果和开源代码能够启发和促进未来对大型视觉语言模型的高效适配方法的研究--这是一个与基础模型民主化相关的新兴课题（Bommasani et al.

尤其是 CLIP（Radford 等人，2021 年）和 ALIGN（贾等人，2021 年），主要是由以下三个领域的进展推动的：i) 使用 Transformers 的文本表示学习（Vaswani

## 2 相关工作

### 2.1 视觉语言模型

最近，视觉语言模型在学习通用视觉表征和通过提示零距离转移到各种下游分类任务方面展现出巨大潜力（Radford 等人，2021 年；Jia 等人，2021 年；Zhang 等人，2020 年；Singh 等人，2021 年；Yuan 等人，2021 年）。

据我们所知，视觉语言学习领域的最新进展，

等人, 2017 年), ii) 大容量对比性表征学习 (

Chen 等人, 2020 年; He 等人, 2020 年; H'énaff 等人, 2020 年), iii) 网络规模的训练数据集--CLIP 受益于 4 亿个经过策划的图像-文本对, 而 ALIGN 则利用了 18 亿个噪声图像-文本对。

将图像和文本映射到共同嵌入空间的想法早在近十年前就已开始研究 (Socher 等人, 2013 年; Frome 等人, 2013 年; Elhoseiny 等人, 2013 年), 但采用的技术却大相径庭。在文本特征提取方面, 早期的工作主要利用预训练的单词向量 (Socher 等人, 2013 年; Frome 等人, 2013 年) 或手工创建的 TF-IDF 特征 (Elhoseiny 等人, 2013 年; Lei Ba 等人, 2015 年)。图像和文本特征的匹配被归纳为度量学习 (Frome 等人, 2013 年)、多标签分类 (Joulin 等人, 2016 年; Gomez 等人, 2017 年)、n-gram 语言学习 (Li 等人, 2017 年) 以及最近提出的标题学习 (Desai 和 Johnson, 2021 年)。

我们的工作与最近在视觉语言模型方面的研究正相反, 旨在促进这些模型在下游数据集中的适应和部署。

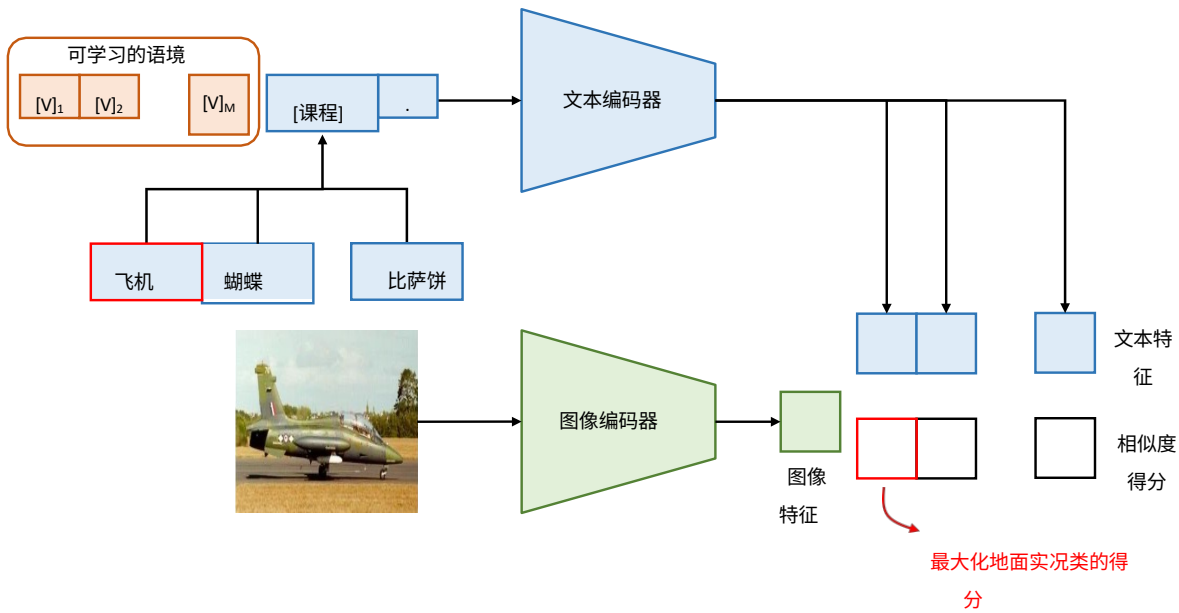
## 2.2 NLP 中的提示学习

Petroni 等人 (2019) 将大型预训练语言模态的知识探究正式定义为 "填空 "招词测试, 该测试最近引发了 NLP 领域及时学习研究的兴趣 (Shin 等人, 2020; Jiang 等人, 2020; Li 和 Liang, 2021; Zhong 等人, 2021; Lester 等人, 2021; Gao 等人, 2020; Liu 等人, 2021b)。

知识探查的基本思想是诱导预先训练好的语言模型来生成给定抬头去尾式提示的答案, 这对情感分析等下游任务大有裨益。Jiang 等人 (2020 年) 建议通过文本挖掘和解析生成候选提示, 并找出训练准确率最高的最优提示。Shin 等人 (2020 年) 引入了一种基于梯度的方法, 该方法可搜索标签似然中具有最大梯度变化的标记

与我们的工作最相关的是连续提示学习方法 (Zhong 等人, 2021 年; Li 和 Liang, 2021 年; Lester 等人, 2021 年), 这种方法可以优化词嵌入空间中的连续向量。与搜索离散词组相比, 这类方法的一个缺点是缺乏一种清晰的方法来直观地显示向量学习到了哪些 "词"。我们建议读者参阅 Liu 等人 (2021a), 以了解有关 NLP 中提示学习主题的全面调查。

值得注意的是, 我们是第一个将快速学习应用于大型视觉系统的适应性研究。



**图 2 上下文优化 (CoOp) 概述。**其主要思想是使用一组可学习的向量对提示语境进行建模，并通过最小化分类损失对其进行优化。我们提出了两种设计方案：一种是统一语境，即所有类别共享相同的语境向量；另一种是特定类别语境，即针对每个类别学习一组特定的语境向量。

我们认为计算机视觉中的语言模型是实现基础模型民主化的重要课题 (Bommasani 等人, 2021 年)，并证明及时学习不仅能显著改善计算机视觉任务的迁移学习性能，还能生成稳健的模型，以应对领域的变化。

3 方法

3.1 视觉语言预培训

我们简要介绍了视觉语言预训练，重点是 CLIP (Radford 等人, 2021 年)。我们的方法适用于更广泛的类似 CLIP 的视觉语言模型。

CLIP 模型由两个编码器组成，一个用于图像，另一个用于文本。图像编码器旨在将高维图像映射到低维嵌入空间。图像编码器的架构可以是类似 ResNet-50 的 CNN (He 等人, 2016 年)，也可以是 ViT (Dosovitskiy 等人, 2021 年)。另一方面，文本编码器建立在 Transformer (Vaswani 等人, 2017 年) 之上，旨在从自然语言生成文本表示。

具体来说，给定一串词 (标记)，如 "狗的照片"，CLIP 首先将每个标记 (包括标点符号) 转换为小写字节对编码 (BPE) 表示法 (Sennrich et al. CLIP 的词汇量为 49,152 个。对



为便于进行小批量处理，每个文本序列都包含 [SOS] 和 [EOS] 标记，长度固定为 77。然后，将 ID 映射到 512-D 字嵌入向量，再将其传递给变换器。最后，对 [EOS] 标记位置的特征进行层归一化处理，并通过线性投影层进一步处理。

**训练** CLIP 的目的是对齐分别为图像和文本学习的两个嵌入空间。具体来说，学习目标被表述为一个限制性损失。给定一批图像-文本对，CLIP 将匹配对的余弦相似度最大化，而将所有其他未匹配对的余弦相似度最小化。为了学习更适用于下游任务的各种视觉概念，CLIP 团队收集了由 4 亿对图像和文本组成的大型训练数据集。

**零镜头推理** 由于 CLIP 经过预先训练，可以预测图像是否与文本描述相匹配，因此它自然适合零镜头识别。这是通过比较图像特征和文本编码器合成的分类权重来实现的，文本编码器将指定兴趣类别的文本描述作为输入。形式上，假设  $\mathbf{f}$  是提取的图像特征

和  $\{\mathbf{w}_i\}_{i=1}^K$  一组由文本编码器生成的权重向量。 $K$  表示类别的数量，而每个  $\mathbf{w}_i$  都来自一个提示，其形式可以是 "一张 [CLASS] 的照片。" 其中类别标记由具体的类别名称代替，如 "猫"、"狗" 或 "汽车"。该

$$i=1$$

预测概率计算公式为

$$p(y = i | \mathbf{x}) = \frac{\exp(\cos(\mathbf{w}_i, \mathbf{f})/\tau)}{\sum_{j=1}^K \exp(\cos(\mathbf{w}_j, \mathbf{f})/\tau)} \quad (1)$$

其中,  $\tau$  是 CLIP 学习到的温度参数,  $\cos(-, -)$  表示余弦相似度。

与传统的分类器学习相比

与从随机向量中学习封闭式视觉概念的方法不同, 视觉语言预训练可以通过大容量文本编码器探索开放式视觉概念, 从而获得更广阔的语义空间, 并反过来使学习到的表述更容易迁移到下游任务中。

### 3.2 背景优化

我们提出了上下文优化 (CoOp) 方案, 通过使用从数据中端到端学习的连续向量对上下文单词进行建模, 同时冻结大量的预训练参数, 从而避免了人工提示调整。图 2 显示了其概览。下面我们将提供几种不同的实现方法。

**统一语境** 我们首先介绍统一语境版本, 它与所有类别共享相同的语境。具体来说, 文本  $\mathbf{e}_i$  的提示。编码器  $g(-)$  的设计形式如下

$$\mathbf{t} = [\mathbf{V}]_1 [\mathbf{V}]_2 \dots [\mathbf{V}]_M [\text{CLASS}], \quad (2)$$

其中, 每个  $[\mathbf{V}]_m$  ( $m \in \{1, \dots, M\}$ ) 都是一个与词嵌入维度相同的向量 (即 512 为词嵌入维度)。

CLIP), 而  $M$  是一个超参数, 指定了上下文标记的数量。

将提示  $\mathbf{t}$  发送给文本编码器  $g(-)$ , 我们可以得到一个分类权重向量, 代表视觉概念 (仍来自  $[\mathbf{E}OS]$  标记位置)。预测概率的计算公式为

$$p(y = i | \mathbf{x}) = \frac{\exp(\cos(g(\mathbf{t}_i), \mathbf{f})/\tau)}{\sum_{j=1}^K \exp(\cos(g(\mathbf{t}_j), \mathbf{f})/\tau)} \quad (3)$$

其中, 每个提示  $\mathbf{t}_i$  中的类标记被替换为的相应词嵌入向量的

第  $i$  个类别名称。

除了等式 (2) 中将类标记放在序列末尾外,

**特定类别上下文** 另一种方法是去标识特定类别上下文

(CSC), 其中上下文向量对每一类都是独立的, 即  $[\mathbf{V}]^i_1 [\mathbf{V}]^i_2 \dots [\mathbf{V}]^i_M$  for  $i \neq j$  and  $i, j \in \{1, \dots, K\}$ . 由于作为统一语境的替代, 我们发现 CSC 是对于某些细粒度分类任务尤其有用。

在交叉熵的基础上进行训练, 以最小化标准分类损失, 并将粒度反向传播到整个系统。

文本编码器  $g(-)$ , 利用丰富的知识, 可以对文本进行编码。

在参数中编码, 以优化上下文。参数

连续表征的设计还允许在单词嵌入空间中进行充分展示, 这有助于学习与任务相关的语境。

### 3.3 讨论

我们的方法专门解决了最近提出的大型视觉语言模型 (如 CLIP, Radford 等人, 2021 年) 的适应性问题。我们的方法与 NLP 领域为语言模型开发的快速学习方法 (如 GPT-3 (Brown 等人, 2020 年)) 有一些不同之处。首先, CLIP 类模型和语言模型的骨干架构明显不同--前者将视觉和文本数据作为输入, 并产生用于图像识别的配准分数, 而后者则仅为处理文本数据而定制。其次, 预训练目标不同: 强制学习与自回归学习。这将导致不同的模型行为, 因此需要不同的模块设计。

## 4 实验

### 4.1 少量学习

我们选择了 CLIP 中使用的 11 个公开图像分类数据集

: ImageNet (Deng

我们还可以将其放在中间, 如

$$\mathbf{t} = [\mathbf{V}]_1 \dots [\mathbf{v}]_m [\text{class}] [\mathbf{v}]_{m+1} \dots [\mathbf{V}]_M, \quad (4)$$

[et al., 2009](#)), Caltech101 ([Fei-Fei et al., 2004](#)), Oxford- Pets ([Parkhi et al., 2012](#)), StanfordCars ([Krause et al., 2013](#)), Flowers102 ([Nilsback and Zisserman, 2008](#)), Food101 ([Bossard et al., 2014](#) 年) 、FGVCAircraft ([Maji 等人, 2013](#) 年) 、

2

2

这就增加了学习的灵活性--提示语既可以用补充说明来填充后面的单元格,也可以使用句号等终止信号提前截断句子。

SUN397 ([Xiao 等人, 2010](#) 年)、DTD ([Cim-<sup>7</sup>poi 等人, 2014](#) 年)、EuroSAT ([Helber 等人, 2019](#) 年) 和 UCF101 ([Soomro 等人, 2012](#) 年) 。

统计)。这些数据集构成了一个全面的该基准涵盖了多种视觉任务,包括一般物体、场景、行动和细粒度类别的分类,以及识别纹理和卫星图像等专门任务。

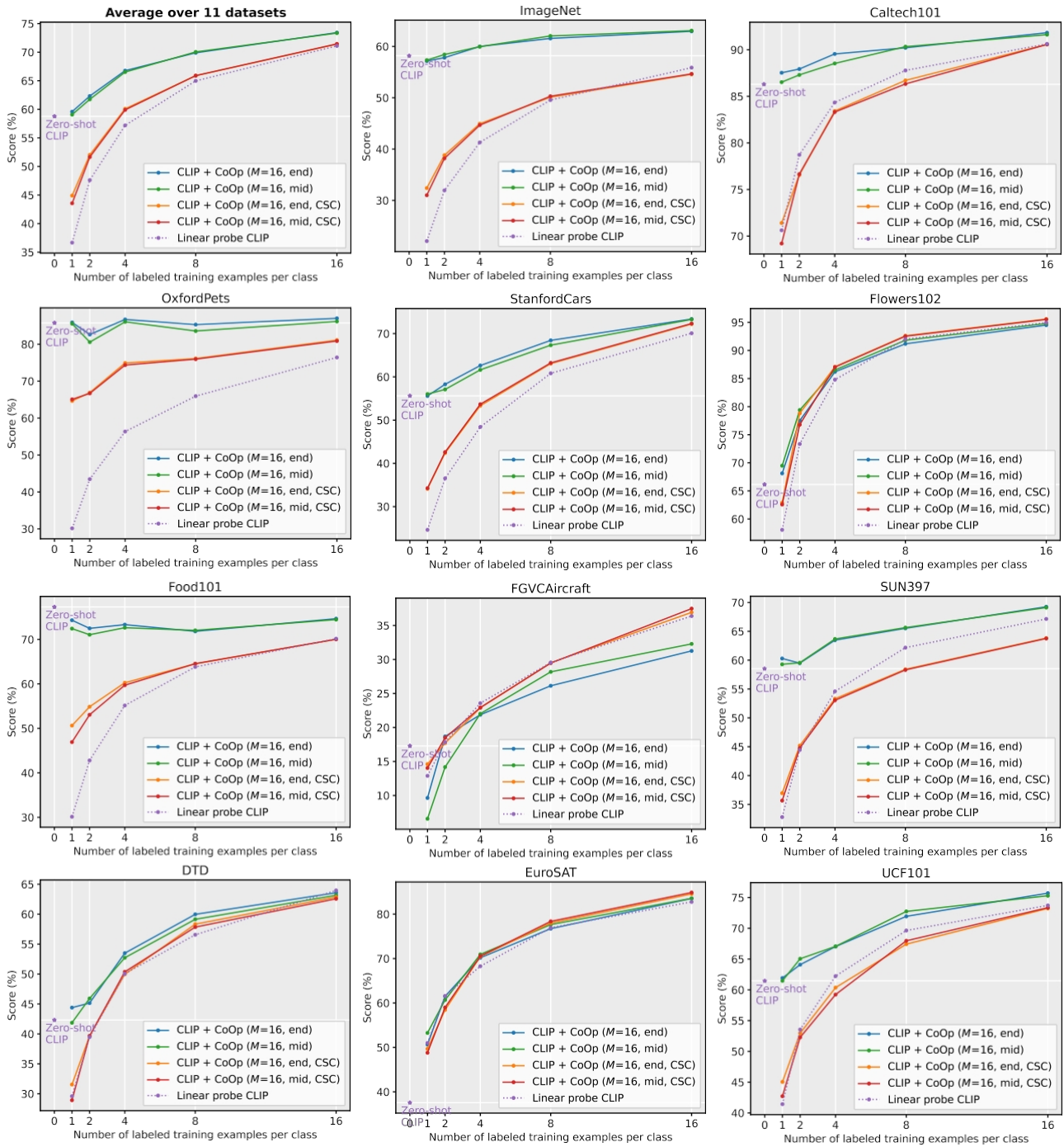


图 3 在 11 个数据集上进行少点学习的主要结果。总体而言，CoOp 有效地将 CLIP 变成了一个强大的少次学习器（实线），与零次学习的 CLIP 相比有显著提高（星线），与线性探针替代方案（虚线）相比也表现出色。 $M$  表示上下文长度。“末端”或“中间”表示将类别标记放在末端或中间。CSC 表示特定类别上下文。

我们沿用了 CLIP (Radford 等人, 2021 年) 中采用的少数几次评估协议，分别使用 1、2、4、8 和 16 次进行训练，并在完整测试集中部署模型。为便于比较，我们报告了三次运行的平均结果。

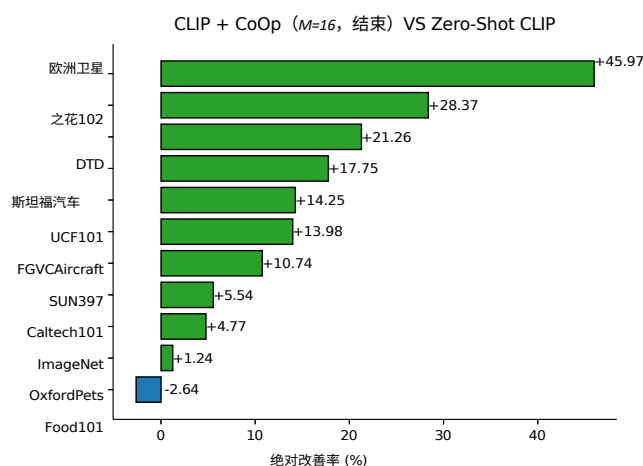
**训练细节** CoOp 有四个版本：将类标记置于末尾或中间；统一上下文与 CSC。除非另有说明，ResNet-50 (He et al. 关于其他设计选择的讨论见

第 4.3 节。所有模型都建立在 CLIP 的开源代码之上。<sup>2</sup>CoOp 的上下文向量是从标准差等于 0.02 的零均值高斯分布中随机初始化的。训练采用 SGD 算法，初始学习率为 0.002，并通过余弦退火规则进行衰减。16/8 个镜头的最大历时设置为 200，4/2 个镜头的最大历时设置为 100，1 个镜头的最大历时设置为 50（ImageNet 除外，其最大历时固定为 50）。为了减少在早期训练迭代中观察到的利用梯度，我们使用热身技巧，将学习率固定为  $1e-5$ ，仅在第一个历元期间使用。

**基准方法** 我们将 CoOp 与以下两种方法进行了比较

基线方法。第一种是零镜头 CLIP，它基于手工制作的提示。我们遵循 Radford 等人（2021 年）提出的提示工程指导路线。对于一般物体和场景，我们采用“一张[CLASS]的照片”。对于细粒度类别，我们会添加与任务相关的上下文，如 OxfordPets 的“一种宠物”和 Food101 的“一种食物”。当涉及到专业任务时，如识别 DTD 中的纹理，提示会被定制为“[CLASS] texture.”，其中类名是形容词，如“bubbly”和“dotted”。详见附录 A。第二个基线是线性探测模型。正如 Radford 等人（2021 年）和最近一项关于少点学习的研究（Tian 等人，2020 年）所建议的，在高质量预训练模型特征（如 CLIP）的基础上训练线性分类器，可以轻松达到与最先进的少点学习方法相当的性能，而后者通常要复杂得多。我们采用与 Radford 等人（2021 年）相同的训练方法来训练线性探测模型。

**与手工制作的提示比较** 图 3 总结了结果。我们的默认模型是 CLIP+CoOp，类标记位于末尾。两种不同的类标记定位方式取得了相似的性能，因为它们的曲线高度重合。从左上角显示的平均性能中，我们可以看出 CLIP+CoOp 是一个很强的少镜头学习器，与零镜头的 CLIP 相比，平均只需要两次镜头就能获得相当大的优势。如果有 16 次训练机会，CoOp



带来的平均差距可以进一步扩大到 15% 左右。

图 4 列出了 CoOp 在 16 个镜头上比手工制作的提示所取得的绝对改进。在专门任务（即 EuroSAT 和 DTD）上可以看到巨大的改进，性能分别提高了 45% 和 20%。性能的跃升也很明显（超过

图 4 与手工制作的提示比较。

在大多数细粒度数据集（包括 Flow-ers102<sup>2</sup>、StanfordCars 和 FGVCAircraft）以及场景和动作识别数据集（即 SUN397 和 UCF101）上的改进幅度为 10%。由于 ImageNet 是一个包含 1,000 个类别的高难度数据集，因此 4.77% 的改进也是值得注意的。相比之下，OxfordPets 和 Food101 这两个细粒度数据集的改进就不那么吸引人了。<sup>3</sup> 通过深入研究图 3 中 CLIP+CoOp 在这两个数据集上的曲线，我们发现即使使用了更多的镜头，性能提升的势头也会减弱，这似乎是一个过度拟合的问题。潜在的解决方案是采用更高的正则化，如增加权重衰减。尽管如此，总体结果足以证明 CoOp 能够以数据高效的方式学习与任务相关的提示。

<sup>2</sup><https://github.com/openai/CLIP>。

**与线性探针 CLIP 的比较** 周开阳等人  
在整体性能方面（图 3，左上角），CLIP+CoOp 与线性探针模型相比具有明显优势。后者平均需要 4 次以上的拍摄才能达到零拍摄的性能，而 CoOp 在 4 次拍摄时的平均增益已经非常可观。同样明显的是，在数据量极少的情况下，比如只有一两次机会时，差距要大得多，这表明 CoOp 比从头开始学习线性分类器更有效。我们还观察到，在两个专业任务（DTD 和 EuroSAT）以及几个细粒度数据集（Flowers102 和 FGVCAircraft）上，线性探测模型与 CLIP+CoOp 不相上下-- 这并不太令人惊讶，因为预训练的 CLIP 空间已被证明是强大的，这使得线性探测模型成为强有力的竞争对手。尽管如此，CoOp 的 CSC 版本

<sup>3</sup>我们发现，对于基于学习的模型（包括 CoOp 和线性探针）来说，Food101 的负面结果是由以下原因造成的：嘈杂的训练数据，"色彩浓烈，有时还错误的标签"（Bossard et al.）

表 1 使用不同视觉骨干与零镜头 CLIP 对分布偏移的鲁棒性比较。 $M$ : CoOp 的上下文长度。

方法	来源	目标			
	图像网	-V2	-草图	-A	-R
<b>ResNet-50</b>					
零点射击 CLIP	58.18	51.34	33.32	21.65	56.00
线性探头夹	55.87	45.97	19.07	12.74	34.86
CLIP + CoOp ( $M = 16$ )	62.95	55.11	32.74	22.12	54.96
CLIP + CoOp ( $M = 4$ )	<b>63.33</b>	<b>55.40</b>	<b>34.67</b>	<b>23.06</b>	<b>56.60</b>
<b>ResNet-101</b>					
零点射击 CLIP	61.62	54.81	38.71	28.05	64.38
线性探头夹	59.75	50.05	26.80	19.44	47.19
CLIP + CoOp ( $M = 16$ )	<b>66.60</b>	<b>58.66</b>	39.08	28.89	63.00
CLIP + CoOp ( $M = 4$ )	65.98	58.60	<b>40.40</b>	<b>29.60</b>	<b>64.98</b>
<b>ViT-B/32</b>					
零点射击 CLIP	62.05	54.79	40.82	29.57	<b>65.99</b>
线性探头夹	59.58	49.73	28.06	19.67	47.20
CLIP + CoOp ( $M = 16$ )	<b>66.85</b>	58.08	40.44	30.62	64.45
CLIP + CoOp ( $M = 4$ )	66.34	<b>58.24</b>	<b>41.48</b>	<b>31.34</b>	65.78
<b>ViT-B/16</b>					
零点射击 CLIP	66.73	60.83	46.15	47.77	73.96
线性探头夹	65.85	56.26	34.77	35.68	58.43
CLIP + CoOp ( $M = 16$ )	<b>71.92</b>	64.18	46.71	48.41	74.32
CLIP + CoOp ( $M = 4$ )	71.73	<b>64.56</b>	<b>47.89</b>	<b>49.93</b>	<b>75.14</b>

在上述数据集上，CoOp 可以击败线性探针 CLIP，而且，当有更多镜头可用时，CoOp 显示出更好的潜力。随后，我们将展示 CoOp 在领域泛化方面比线性探针模型更强的性能。

**统一上下文与特定于类的上下文** 平均而言，使用统一上下文能带来更好的性能。关于何时应用 CSC，何时不应用 CSC，我们有以下建议。对于通用对象（ImageNet 和 Caltech101）、场景（SUN397）和动作（UCF101），使用统一上下文显然更好。在一些细粒度数据集（包括 OxfordPets 和 Food101）上，统一上下文的效果也更好，但在其他数据集（如 StanfordCars、Flowers102 和 FGVC Aircraft）上，CSC 版本则更受欢迎。在 DTD 和 EuroSAT 这两个特殊任务上，CSC 也能获得更好的性能，尤其是在 16 个镜头上。不过，在具有挑战性的低数据场景（少于 8 个镜头）中，CSC 的表现大多不如统一上下文，这是因为 CSC 比统一上下文有更多的参数，需要更多的数据进行训练。

4.2 领域通用化

由于 CoOp 需要在特定的数据分布上进行训练，因此它有可能学习到虚假的相关性，而这些相关性是

正如最近的研究 (Taori 等人, 2020 年; Zhou 等人, 2021 年) 所指出的那样, CLIP 在未见过的分布 (域) 中不利于泛化。相反, 零镜头 CLIP 不依赖于特定的数据分布, 对分布变化表现出很强的鲁棒性 (Radford 等人, 2021 年)。在本节中, 我们将揭示 CoOp 对分布变化的稳健性, 并与零点 CLIP 和线性探测模型进行比较。

**数据集** 源数据集是 ImageNet。目标数据集是 ImageNetV2 (Recht 等人, 2019 年)、ImageNet-Sketch (Wang 等人, 2019 年)、ImageNet-A (Hendrycks 等人, 2021b) 和 ImageNet-R (Hendrycks 等人, 2021a), 所有这些数据集的类名都与 ImageNet 兼容, 可以无缝传输 CoOp 学习到的提示。ImageNetV2 是一个重现的测试集, 使用了不同的数据源, 同时遵循了 ImageNet 的数据收集流程。ImageNet-Sketch 包含属于相同的 1000 个 ImageNet 类别的素描图像。ImageNet-A 和 -R 都包含从 ImageNet 的 1000 个类别中提取的 200 个类别。前者由现实世界中经过逆向过滤的图像组成, 这些图像会导致当前的 ImageNet 分类器产生较低的结果, 而后者则以不同的图像风格 (如绘画、卡通和雕塑) 呈现 ImageNet 类别。



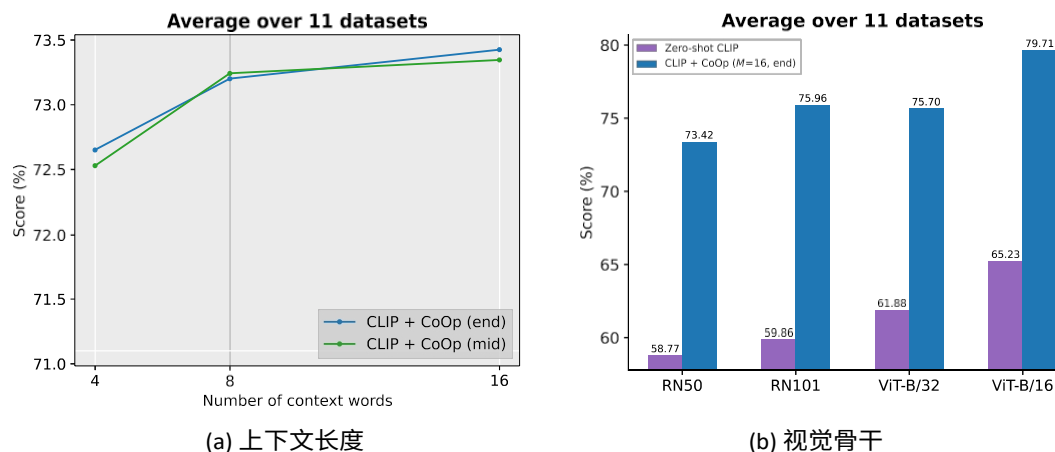


图 5 对 CoOp 的上下文长度和各种视觉骨干的研究。

表 2 使用不同视觉骨干在 ImageNet 上与即时工程和即时集合进行的比较。

方法	ResNet-50	ResNet-101	ViT-B/32	ViT-B/16
及时的工程设计	58.18	61.26	62.05	66.73
迅速组装	60.41	62.54	63.71	68.74
CoOp	<b>62.95</b>	<b>66.60</b>	<b>66.85</b>	<b>71.92</b>

表 3 随机初始化与手动初始化。

平均
百分比 [V] <sub>1</sub> [V] [V] [V] <sub>234</sub>
72.65
"一张照片"
72.65

的上下文长度有利于领域泛化（可能是由于学习的参数较少，过拟合程度较低）。在此，我们针对源数据集研究了这一超参数。具体来说，我们在以下 11 个数据集上重复实验

结果 表 1 总结了结果（包括多种视觉骨干）。令人惊讶的是，尽管接触了源数据集，CoOp 仍增强了 CLIP 对分布变化的稳健性。这表明学习到的提示也具有通用性。此外，有趣的是，使用较少的上下文标记会带来更好的鲁棒性。相比之下，线耳探针模型在这些目标数据集上的结果要差得多，这暴露了它在领域泛化方面的弱点。在附录 B 中，我们提供了在 DOSCO-2k (Zhou 等人, 2022b) 上的领域泛化结果，这是最近提出的一个侧重于上下文领域转换的基准。

#### 4.3 进一步分析

上下文长度 应该使用多少上下文标记？上下文标记是否越多越好？第 4.2 节中的结果表明，较短

将上下文长度从 4 到 8 再到 16 变化。平均结果如图 5 (a) 所示, 结果表明, 拥有更多的上下文标记会带来更好的性能, 而将类标记定位在中段则会随着上下文长度的增加而获得更多的动力。总之, 选择合适的上下文长度并没有金科玉律, 因为我们需要在性能和对分布变化的稳健性之间取得平衡。

**视觉骨干图** 5(b) 总结了在 11 个数据集上使用各种视觉骨干 (包括 CNN 和 ViT) 的结果。结果在意料之中: 骨干越先进, 性能越好。在所有架构中, CoOp 与手工制作的提示之间的差距都很大。

**与提示集合的比较** CLIP (Radford 等人, 2021 年) 的支持者提出, 通过集合使用不同手工制作的提示生成的多个零镜头分类器, 例如 "一张大 [CLASS]. 的照片"、"一张 [CLASS]. 的坏照片" 和 "一张折纸 [CLASS].", 可以获得额外的改进, 这些提示分别反映了图像的不同比例、视图和抽象程度。我们很想知道, 与提示组合相比, CoOp 学习到的提示是否仍能保持优势。为了进行公平比较, 我们使用 Radford 等人 (2021 年) 选择的提示语构建集合分类器, 这些提示语已在 ImageNet 上进行了广泛调整。表 2 显示了比较结果, 并证明了

考虑到快速组装的潜力，今后的工作可以研究如何从组装的角度改进 CoOp。

**与其他微调方法的比较** 我们进一步将 CoOp 与其他微调方法进行了比较：i) 微调 CLIP 的图像编码器；ii) 优化添加到文本编码器输出中的转换层；iii) 优化添加到文本编码器输出中的偏置项。结果如表 5 所示。显然，微调图像编码器的效果并不理想。添加转换层可略微改善零镜头模型。添加偏置项的结果很有希望，但在很大程度上仍然不如 CoOp，这表明通过文本编码器的梯度提供了更有用的信息。

**初始化** 我们比较了随机初始化和人工初始化。后者使用 "a 的照片" 的嵌入来初始化 11 个数据集的上下文向量。为了进行公平比较，我们还将使用随机初始化时的上下文长度设置为 4。表 3 显示，"好的" 初始化并不会产生太大的影响。虽然进一步调整初始化词可能会有所帮助，但在实践中，我们建议使用简单的随机初始化方法。

由于上下文向量是在连续空间中优化的，因此解释学习到的提示非常困难。我们采用了一种间接的方法，即根据欧氏距离在词汇表中搜索与所学向量最接近的单词。请注意，CLIP (Radford 等人, 2021 年) 使用 BPE representation (Sennrich 等人, 2016 年) 进行标记化，因此词汇中包含了文本中经常出现的子词，如 "hu" (由 "hug" 和 "human" 等多个词组成)。表 4 显示了在一些数据集上的搜索结果。我们发现有几个词与任务有一定的相关性，如 Food101 中的 "enjoyed"，OxfordPets 中的 "fluffy" 和 "paw"，以及 DTD 中的 "pretty"。但是，当把所有近义词连在一起时，提示就没有什么意义了。我们还观察到，在使用手动初始化 (如 "a photo of a") 时，收敛向量的最近词大多是初始化时使用的词。我们推测，学习到的向量可能编码了现有词汇之外的含义。总

之，我们无法根据观察结果得出任何肯定的结论，因为使用最近词来解释所学提示可能并不准确--向量的语义并不一定与最近词相关。

大型预训练视觉语言模型在各种下游应用中显示出了惊人的强大能力。然而，这些模型也被称为视觉基础模型，因为它们具有 "至关重要的核心但却不完整" 的性质 (Bommasani 等人, 2021 年)，需要使用自动化技术对其进行调整，以获得更好的下游性能和效率。

我们的研究为如何通过提示学习将类似 CLIP 的模型转化为数据效率高的学习器提供了及时的见解，并揭示了尽管 CoOp 是一种基于学习的方法，但它在领域泛化方面的表现要比人工提示好得多。这些结果有力地证明了提示学习在大型视觉模型中的应用潜力。值得注意的是，我们的论文首次提出了利用提示学习调整大型视觉模型的综合研究。

虽然表现出色，但 CoOp 的结果与 NLP 中的其他连续提示学习方法一样，相对难以解释。实验还表明，鉴于 CoOp 在 Food101 上的微弱表现，它对噪声标签很敏感。

尽管如此，CoOp 的简易性为未来工作提供了便利，仍有许多相互关联的问题有待探索，例如跨数据集转换 (Zhou 等, 2022a) 和测试时间适应 (Wang 等, 2020)。此外，对超大规模虚拟空间模型的更通用适应方法进行研究也很有意义 (Jia 等, 2022; Bahng 等, 2022; Gao 等, 2021)。总之，我们希望本文提出的经验发现和见解能为未来研究新兴基础模型的高效适应方法铺平道路，这仍是一个新兴的研究课题。

**致谢** 本研究得到了南洋理工大学国家重点实验室、教育部AcRF Tier 2 (T2EP20221-0033)和RIE2020产业联盟基金-产业合作项目 (IAF-ICP) 资助计划的支持，以及来自产业合作伙伴的现金和实物捐助。通讯作者：刘紫薇 ()  
：刘紫薇 (ziwei.liu@ntu.edu.sg).

附录

A 数据集详情

周开阳等人  
表 6 列出了 11 个数据集以及 ImageNet 四个变体的详细统计数据。表中还详细列出了用于零镜头 CLIP 的手工制作的提示。对于 Caltech101，"BACKGROUND Google"和 "Faces easy"类被舍弃。对于视频数据集 UCF101，每个视频的中间帧都被用作图像编码器的输入。

表 4 CoOp 学习到的 16 个上下文向量中每个向量的最近单词，括号中显示的是它们之间的距离。N/A 表示非拉丁字符。

#	图像网	食品101	牛津宠物	DTD	UCF101
1	potd (1.7136)	LC (0.6752)	TOSC (2.5952)	盒装 (0.9433)	气象学家 (1.5377)
2	即 (1.4015)	享有 (0.5305)	法官 (1.2635)	种子 (1.0498)	exe (0.9807)
3	锉刀 (1.2275)	beh (0.5390)	绒毛 (1.6099)	安娜 (0.8127)	父母 (1.0654)
4	水果 (1.4864)	匹配 (0.5646)	购物车 (1.3958)	山 (0.9509)	精湛 (0.9528)
5	,...(1.5863)	纽约时报 (0.6993)	哈兰 (2.2948)	长子 (0.7111)	fe (1.3574)
6	° (1.7502)	普罗 (0.5905)	爪 (1.3055)	漂亮 (0.8762)	thof (1.2841)
7	排除 (1.2355)	较低 (0.5390)	incase (1.2215)	面孔 (0.7872)	其中 (0.9705)
8	冷 (1.4654)	不适用	bie (1.5454)	蜂蜜 (1.8414)	克里斯汀 (1.1921)
9	stery (1.6085)	分钟 (0.5672)	依偎 (1.1578)	系列 (1.6680)	伊玛目 (1.1297)
10	勇士 (1.3055)	~ (0.5529)	沿 (1.8298)	古柯 (1.5571)	近 (0.8942)
11	marvelcomics (1.5638)	井 (0.5659)	享受 (2.3495)	月亮 (1.2775)	腹部 (1.4303)
12	.: (1.7387)	两端 (0.6113)	jt (1.3726)	升 (1.0382)	hel (0.7644)
13	不适用	误差 (0.5826)	改善 (1.3198)	韩元 (0.9314)	噉 (1.0491)
14	(1.5015)	某物 (0.6041)	Srsly (1.6759)	作出答复 (1.1429)	不适用
15	muh (1.4985)	研讨会 (0.5274)	小行星 (1.3395)	发送 (1.3173)	面部 (1.4452)
16	.# (1.9340)	不适用	不适用	皮埃蒙特 (1.5198)	期间 (1.1755)

表 5 CoOp 与其他微调方法在 ImageNet 上的对比（16 次拍摄）。Δ: 与零镜头模型的差异。CoCoOp 在解决迁移学习问题方面潜力巨大。

	图像网	Δ
零镜头剪辑	58.18	-
线性探头	55.87	-2.31
微调 CLIP 的图像编码器	18.28	-39.90
优化转换层（文本）	58.86	0.68
优化偏差（文本）	60.93	+2.75
CoOp	62.95	+4.77

B DOSCO-2k 的结果

**DOSCO-2k** DOSCO（DObain Shift in COntext）基准（Zhou 等人，2022b）包含 7 个图像识别数据集，涵盖了广泛的分类问题，如通用对象识别、飞机模型的细粒度识别和动作识别。与前述领域泛化数据集不同的是，DOSCO-2k侧重于更广泛的上下文领域偏移，这种偏移由在 Places数据集上预先训练的神经网络自动检测（Zhou等人，2017年）。效仿 Zhou 等人（2022b）的做法，我们使用 2k 版本，每个数据集的训练和验证分区共有 2,000 幅图像（1,600 幅用于训练，400 幅用于验证）。

**结果** 我们研究了三种方法在 DOSCO-2k 上的领域泛化性能：CLIP、CoOp 和 CoCoOp（Zhou 等人，2022a）。所有模型都是在训练集上训练的，具有最佳验证性能的检查点被用于未见领域的最终测试。表 7 显示了四种不

**参考资料**

Bahng H、Jahanian A、Sankaranarayanan S、Isola P (2022) 视觉提示：修改像素空间以适应预  
arXiv preprint arXiv:220317274

Bommasani R、Hudson DA、Adeli E、Altman R、Arora S、von  
Arx S、Bernstein MS、Bohg J、Bosselut A、Brunskill E、et al. (2021) On the opportunities and risks of foundation  
同架构的结果。很明显，尽管只需调整少量参数，但两种  
学习方法的性能远远优于 "零射击 "方法。在 7 个数据集  
中，CoCoOp 有 4 个数据集优于 CoOp，但 CoOp 的平均  
性能更高。总之，研究结果表明，像 CoOp 和 CoOp 这样  
的高效适配方法，在 7 个数据集中有 4 个胜过 CoOp，但  
CoOp 的平均性能更高。

- 
- Bossard L, Guillaumin M, Van Gool L (2014) Food-101-mining discriminative components with random forests. In : ECCV
- Brown TB, Mann B, Ryder N, Subbiah M, Kaplan J, Dhariwal P, Neelakantan A, Shyam P, Sastry G, Askell A, et al. (2020) Language models are few-shot learners. arXiv preprint arXiv:200514165
- Chen T, Kornblith S, Norouzi M, Hinton G (2020) A simple framework for contrastive learning of visual representations. In: ICML
- Cimpoi M, Maji S, Kokkinos I, Mohamed S, Vedaldi A (2014) Describing textures in the wild. In: CVPR
- Deng J, Dong W, Socher R, Li LJ, Li K, Fei-Fei L (2009) Imagenet: A large-scale hierarchical image database. In : CVPR
- Desai K, Johnson J (2021) Virtex: Learning visual representations from textual annotations. In: CVPR
- Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S, et al. (2021) An image is worth 16x16 words: 规模图像识别变换器。 In: ICLR
- Elhoseiny M, Saleh B, Elgammal A (2013) Write a classifier: 使用纯文本描述的零点学习。 In: ICCV
- Fei-Fei L, Fergus R, Perona P (2004) Learning generative visual models from few training examples: 在 101 个物体类别上测试的渐增贝叶斯方法。 In: CVPR-W

表 6 数据集统计

数据集	班级	火车	瓦尔	测试	手工制作的提示
图像网	1,000	1.28M	不适用	50,000	"一张 [CLASS] 的照片"。
Caltech101	100	4,128	1,649	2,465	"一张 [CLASS] 的照片"。
牛津宠物	37	2,944	736	3,669	"[CLASS]的照片，一种宠物"。
斯坦福汽车	196	6,509	1,635	8,041	"一张 [CLASS] 的照片"。
鲜花102	102	4,093	1,633	2,463	"[CLASS]的照片，一种花"。
食品101	101	50,500	20,200	30,300	"[CLASS]的照片，一种食物"。
FGVCAircraft	100	3,334	3,333	3,333	"[CLASS]的照片，一种飞机"。
SUN397	397	15,880	3,970	19,850	"一张 [CLASS] 的照片"。
DTD	47	2,820	1,128	1,692	"[CLASS]纹理"。
欧洲卫星	10	13,500	5,400	8,100	"[CLASS]的居中卫星照片"。
UCF101	101	7,639	1,898	3,783	"一个人做[CLASS]的照片"。
ImageNetV2	1,000	不适用	不适	10,000	"一张 [CLASS] 的照片"。
			用		
ImageNet-Sketch	1,000	不适用	不适用	50,889	"一张 [CLASS] 的照片"。
ImageNet-A	200	不适用	不适用	7,500	"一张 [CLASS] 的照片"。
ImageNet-R	200	不适用	不适用	30,000	"一张 [CLASS] 的照片"。

表 7 在 DOSCO-2k 上的领域泛化结果，DOSCO-2k 是最近提出的一个基准，侧重于更广泛的上下文领域转换。在这三种方法中，CoOp 及其后续方法 CoCoOp 包含可学习的成分，而 CLIP 在此表示零点模型。CoOp 和 CoCoOp 都使用四个可学习的上下文标记，并以 "a photo of a "的词嵌入作为初始化。粗体表示特定架构在每个数据集上的最佳性能。

	P-Air	P-Cars	P-Ctech	P-Ins	P-Mam	P 宠 物	P-UCF	平均值
<b>ResNet-50</b>								
剪辑	16.1	56.1	86.7	62.7	59.7	84.0	60.6	60.9
CoOp	<b>22.1</b>	<b>60.7</b>	89.4	66.3	61.6	83.8	<b>69.2</b>	64.7
CoCoOp	20.1	59.8	<b>90.4</b>	<b>67.9</b>	<b>63.8</b>	<b>87.6</b>	69.1	<b>65.5</b>
<b>ResNet-101</b>								
剪辑	17.5	63.2	89.5	62.4	62.2	84.2	61.3	62.9
CoOp	<b>24.6</b>	<b>68.2</b>	92.0	68.3	65.4	88.2	<b>72.7</b>	<b>68.5</b>
CoCoOp	22.5	65.2	<b>93.3</b>	<b>69.9</b>	<b>67.5</b>	<b>88.6</b>	71.5	68.4
<b>ViT-B/32</b>								
剪辑	18.2	60.1	91.6	61.3	61.8	85.5	61.3	62.8
CoOp	<b>24.0</b>	<b>63.0</b>	93.6	67.3	65.7	<b>88.5</b>	<b>74.5</b>	<b>68.1</b>
CoCoOp	19.5	60.4	<b>93.8</b>	<b>69.8</b>	<b>67.3</b>	<b>88.5</b>	72.7	67.4
<b>ViT-B/16</b>								
剪辑	24.4	64.9	92.6	67.5	67.9	87.4	66.1	67.2
CoOp	<b>32.4</b>	<b>72.4</b>	94.7	73.2	72.1	90.1	<b>78.2</b>	<b>73.3</b>
CoCoOp	30.4	68.7	<b>94.8</b>	<b>73.5</b>	<b>73.6</b>	<b>91.6</b>	76.3	72.7

Frome A, Corrado G, Shlens J, Bengio S, Dean J, Ranzato M, Mikolov T (2013) Devise: 深度视觉语义 床垫模型。 In: NeurIPS

Fu`rst A, Rumetshofer E, Tran V, Ramsauer H, Tang F, Lehner J, Kreil D, Kopp M, Klambauer G, Bitto-Nemling A, et al. (2021) Cloob: arXiv preprint arXiv:211011316

Gao P, Geng S, Zhang R, Ma T, Fang R, Zhang Y, Li H, Qiao Y (2021) Clip-adapter: 使用 特征适配器建立更好的视觉语言模型。 arXiv 预印本 arXiv:211004544

Gao T, Fisch A, Chen D (2020) Making pre-trained lan- guage models better few-shot learners. ArXiv preprint arXiv:201215723

Gomez L, Patel Y, Rusin`ol M, Karatzas D, Jawahar C (2017) Self-supervised learning of visual features through embed- ding images into text topic spaces.In: CVPR

- 
- He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: CVPR
- He K, Fan H, Wu Y, Xie S, Girshick R (2020) Momentum contrast for unsupervised visual representation learning. In: CVPR
- Helber P, Bischke B, Dengel A, Borth D (2019) Eurosat: 用于土地利用和土地覆被分类的新型数据集和深度学习基准。 IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing H'enaiff OJ, Srinivas A, Fauw JD, Razavi A, Doersch C, Es- lami SMA, van den Oord A (2020) Data-efficient image. 使用对比预测编码进行识别。 In: ICML Hendrycks D, Basart S, Mu N, Kadavath S, Wang F, Dorundo E, Desai R, Zhu T, Parajuli S, Guo M, Song D, Steinhardt J, Gilmer J (2021a) The many faces of robustness: 分布外泛化的批判性分析。 ICCV



- Hendrycks D, Zhao K, Basart S, Steinhardt J, Song D (2021b) 自然对抗示例。In: CVPR
- Jia C, Yang Y, Xia Y, Chen YT, Parekh Z, Pham H, Le QV, Sung Y, Li Z, Duerig T (2021) Scaling up visual and vision-language representation learning with noisy text supervision. In: ICML
- Jia M, Tang L, Chen BC, Cardie C, Belongie S, Hariharan B, Lim SN (2022) Visual prompt tuning. arXiv preprint arXiv:2203.12119
- Jiang Z, Xu FF, Araki J, Neubig G (2020) How can we know what language models know? ACL
- Joulin A, Van Der Maaten L, Jabri A, Vasilache N (2016) Learning visual features from large weakly supervised data. In: ECCV
- Krause J, Stark M, Deng J, Fei-Fei L (2013) 3d object representations for fine-grained categorization. In: ICCV-W
- Lei Ba J, Swersky K, Fidler S, et al. (2015) Predicting deep zero-shot convolutional neural networks using textual descriptions. In: ICCV
- Lester B, Al-Rfou R, Constant N (2021) The power of scale for parameter-efficient prompt tuning. arXiv preprint arXiv:2104.08691
- Li A, Jabri A, Joulin A, van der Maaten L (2017) Learning visual n-grams from web data. In: ICCV
- Li XL, Liang P (2021) Prefix-tuning: 优化连续提示生成。arXiv 预印本 arXiv:2101.00190
- Li Y, Liang F, Zhao L, Cui Y, Ouyang W, Shao J, Yu F, Yan J (2021) 监督无处不在: 一种数据高效的对比语言图像预训练范式。arXiv preprint arXiv:2110.05208
- Liu P, Yuan W, Fu J, Jiang Z, Hayashi H, Neubig G (2021a) Pre-train, prompt, and predict: 自然语言处理中提示方法的系统调查。arXiv preprint arXiv:2107.13586
- Liu X, Zheng Y, Du Z, Ding M, Qian Y, Yang Z, Tang J (2021b) Gpt understands, too. arXiv preprint arXiv:2103.10385
- Maji S, Rahtu E, Kannala J, Blaschko M, Vedaldi A (2013) Fine-grained visual classification of aircraft. arXiv preprint arXiv:1306.5151
- Nilsback ME, Zisserman A (2008) Automated flower classification over a large number of classes. In: ICVGIP
- Parkhi OM, Vedaldi A, Zisserman A, Jawahar C (2012) 猫和狗。中: CVPR
- Petroni F, Rocktaschel T, Lewis P, Bakhtin A, Wu Y, Miller AH, Riedel S (2019) Language models as knowledge bases? In: EMNLP
- Radford A, Kim JW, Hallacy C, Ramesh A, Goh G, Agarwal S, Sastry G, Askell A, Mishkin P, Clark J, et al. (2021) Learning transferable visual models from natural language supervision. In: ICML
- Recht B, Roelofs R, Schmidt L, Shankar V (2019): 图像 genet 分类器能推广到图像网吗? In: ICML
- Sennrich R, Haddow B, Birch A (2016): 利用子词单元对罕见词进行神经机器转。In: ACL
- Shin T, Razeghi Y, Logan IV RL, Wallace E, Singh S (2020) Autoprompt: 通过自动生成的提示从语言模型中获取知识。In: EMNLP
- Singh A, Hu R, Goswami V, Couairon G, Galuba W, Rohrbach M, Kiela D (2021) Flava: arXiv preprint arXiv:2112.04482
- Socher R, Ganjoo M, Sridhar H, Bastani O, Manning CD, Ng AY (2013) Zero-shot learning through cross-modal transfer. In: NeurIPS

- Soomro K, Zamir AR, Shah M (2012) Ucf101: 从野外视频中提取的 101 个人类动作类别的数据集。ArXiv preprint arXiv:12120402
- Taori R, Dave A, Shankar V, Carlini N, Recht B, Schmidt L (2020) Measuring robustness to natural distribution shifts in image classification. In: NeurIPS
- Tian Y, Wang Y, Krishnan D, Tenenbaum JB, Isola P (2020) Rethinking few-shot image classification: a good embedding is all you need? In: ECCV
- Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser L-, Polosukhin I (2017) Attention is all you need. In: NeurIPS
- Wang D, Shelhamer E, Liu S, Olshausen B, Darrell T (2020) Tent: 通过熵最小化实现完全测试时间适应。arXiv preprint arXiv:200610726
- Wang H, Ge S, Lipton Z, Xing EP (2019) 通过惩罚局部预测能力来学习稳健的全局表征。In: NeurIPS
- Xiao J, Hays J, Ehinger KA, Oliva A, Torralba A (2010) Sun 数据库: 从修道院到动物园的大规模场景识别。In: CVPR
- Yuan L, Chen D, Chen YL, Codella N, Dai X, Gao J, Hu H, Huang X, Li B, Li C, et al. (2021) Florence : ArXiv preprint arXiv:211111432
- Zhang Y, Jiang H, Miura Y, Manning CD, Langlotz CP (2020) Contrastive learning of medical visual representations from paired images and text. arXiv preprint arXiv:201000747
- Zhong Z, Friedman D, Chen D (2021) Factual probing is [掩码]: 学习与学习回忆。In: NAACL
- Zhou B, Lapedriza A, Khosla A, Oliva A, Torralba A (2017) Places: 用于场景识别的千万级图像数据库。IEEE transactions on pattern analysis and machine intelligence 40(6):1452-1464
- Zhou K, Liu Z, Qiao Y, Xiang T, Loy CC (2021) Domain generalization: arXiv preprint arXiv:210302503
- Zhou K, Yang J, Loy CC, Liu Z (2022a) Conditional prompt learning for vision-language models. arXiv preprint arXiv:220305557
- Zhou K, Zhang Y, Zang Y, Yang J, Loy CC, Liu Z (2022b) On-device domain generalization.