Context Optimization–自动化视觉-语言模型的提示工程

摘要

本文综述了 CLIP（Contrastive Language-Image Pre-training）和 CoOp（Context Optimization）两种多模态模型。CLIP 通过对比学习将图像和文本信息融合，能够在多种任务中表现出色，如图像分类、物体检测和文本生成等。然而，CLIP 需要大量数据进行训练，且对预训练数据质量要求较高。CoOp 则通过自监督学习生成图像描述，不依赖大量标记数据，具有更广泛的通用性。本文还探讨了 CoOp 的改进版本 CoCoOp（Conditional Context Optimization），其通过引入轻量级网络为每个输入样本生成特定的 prompt，提高了模型在不可见类上的泛化性能。

关键词： CLIP；CoOp；多模态；提示工程；

1 引言

图像识别技术的发展经历了从传统算法到深度学习的转变。最初，图像识别主要基于人工设计的特征提取和匹配方法，如 Hough 变换、边缘检测等。这些方法需要人工设计特征，并且对于复杂的图像识别任务，效果有限。随着模式识别和人工智能的发展，图像识别技术得到了一定的进展，主要基于神经网络、决策树等机器学习方法，如多层感知器、回归树等。这些方法可以自动学习特征，但对于大规模、高维的图像数据，效果仍然有限。

进入 21 世纪，深度学习技术的发展，尤其是卷积神经网络（CNN）的出现，极大地推动了图像识别技术的发展。CNN 通过多层卷积和池化操作，能够自动提取图像特征，显著提高了识别精度。AlexNet、VGG 和 ResNet 等深度学习模型在 ImageNet 等大型数据集上取得了突破性成果，标志着深度学习在计算机视觉领域的广泛应用 [4][5][6]。

在自然语言处理（NLP）领域，预训练语言模型的发展同样取得了显著进展。这些模型通过大规模文本数据的训练，能够捕捉语言的深层次语义和上下文关系，为多种 NLP 任务提供了强大的支持 [7]。预训练模型的成功引发了对其在视觉-语言结合任务中应用的探索，特别是在视觉-语言预训练领域 [8]。

2 相关工作

CLIP 模型的提出背景在于，如何将自然语言处理技术的进步应用于计算机视觉领域，特别是在图像识别任务中。CLIP 通过对比学习的方式，将图像和文本嵌入到同一个特征空间

中，使得模型能够理解图像内容和对应的文本描述之间的关系。CLIP 的核心思想是让模型学会匹配图像和其对应的文本描述，而不是生成或预测具体的文本，从而高效地学习图像和文本的联合嵌入空间 [1]。

课题的研究目标是探索如何利用自然语言监督来提高图像识别模型的泛化能力和减少对标注数据的依赖。CLIP 模型通过从自然语言描述中学习视觉表示，并实现零样本迁移学习，为这一目标提供了新的可能性。在此基础上，CoOp 模型被提出，其主要思想是自动设计提示文本，先保持预训练参数不变，然后利用少量数据去学习合适的 prompt，这样的 prompt 比人工设计的提示文本在测试时更有效。

2.1 CLIP 模型

CLIP（Contrastive Language-Image Pre-training）模型是一种强大的多模态预训练模型，由 OpenAI 提出。它通过对比学习将图像和文本信息融合在一起，能够在多种任务中表现出色。CLIP 的核心作用在于实现零样本学习，即在没有直接标注样本的情况下，对全新类别的数据进行准确预测。此外，CLIP 还广泛应用于文本到图像检索、图像到文本检索、视觉问题回答和图像描述生成等任务，展示了其在多模态领域的广泛适用性和高效性。

2.2 CoOp

CoOp（Context Optimization）是一种用于优化预训练视觉语言模型（如 CLIP）的提示工程的方法。它通过自动学习提示的上下文词，避免了手动调整提示的复杂性。CoOp 使用可学习的向量来建模提示的上下文，这些向量在训练过程中从数据中端到端学习，同时保持预训练参数不变。这种方法不仅提高了模型在下游任务中的性能，还减少了人工干预，使模型能够更有效地适应不同的图像识别任务。

3 本文方法

3.1 本文方法概述

CLIP 模型由两个编码器组成：一个用于图像，另一个用于文本。图像编码器的目标是将高维图像映射到低维嵌入空间，可以采用 CNN（如 ResNet-50）或 ViT 架构。文本编码器基于 Transformer，旨在从自然语言生成文本表示。给定一个单词序列，CLIP 首先将每个单词（包括标点符号）转换为小写字节对编码（BPE）表示，这是一种独特的数值 ID。CLIP 的词汇量为 49,152。为了便于小批量处理，每个文本序列都以 [SOS] 和 [EOS] 标记开始和结束，并限制在固定长度 77。之后，这些 ID 被映射到 512-D 的词嵌入向量，然后传递给 Transformer。最后，[EOS] 标记位置的特征进行层归一化，并由线性投影层进一步处理。

CLIP 的训练目标是使图像和文本各自的嵌入空间对齐。具体来说，学习目标是制定为对比损失。给定一批图像-文本对，CLIP 最大化匹配对的余弦相似度，同时最小化所有其他不匹配对的余弦相似度。为了学习多样化的视觉概念，CLIP 团队收集了由 4 亿图像-文本对组成的大型训练数据集。

由于 CLIP 预训练用于预测图像是否与文本描述匹配，它自然适合零样本识别。这是通过比较图像特征与文本编码器合成的分类权重来实现的，文本编码器的输入是指定类别的文

本描述。形式上，设 f 为图像编码器提取的图像特征，文本编码器生成的一组权重向量。K 表示类别数量，每个 w 都是从提示中派生的，提示可能是 “a photo of a [CLASS]” 的形式，其中类别标记被特定类别名称替换，如 “cat”、“dog” 或 “car”。预测概率计算如下，如图 1所示：

$$p(y=i|\boldsymbol{x}) = \frac{\exp(\cos(\boldsymbol{w_i}, \boldsymbol{f})/\tau)}{\sum_{j=1}^{K} \exp(\cos(\boldsymbol{w_j}, \boldsymbol{f})/\tau)},$$

图 1. 概率公式

3.2 提示设计

接下来，我们提出了上下文优化（CoOp），它避免了手动提示调整，通过从数据中端到端学习连续向量来建模上下文词，同时冻结大量的预训练参数。我们提供了几种不同的实现方式。统一上下文版本与所有类别共享相同的上下文。具体来说，给文本编码器 g(·) 的提示设计如图 2，另一种选择是设计特定于类别的上下文（CSC），其中上下文向量独立于每个类别，作为统一上下文的替代方案，我们发现 CSC 特别适用于一些细粒度分类任务。训练是为了最小化基于交叉熵的标准分类损失，并且梯度可以一直通过文本编码器 g(·) 反向传播，利用编码在参数中的丰富知识来优化上下文。连续表示的设计也允许在词嵌入空间中进行全面探索，这有助于学习与任务相关的上下文。总的来说，CoOp 方法特别针对最近提出的大型视觉语言模型（如 CLIP）的适应问题。与 NLP 中为语言模型开发的语言提示学习方法（例如 GPT-3）不同，CLIP 类模型和语言模型的背景架构明显不同——前者同时处理视觉和文本数据，并产生用于图像识别的对齐分数，而后者仅处理文本数据。

$$t = [V]_1^i [V]_2^i ... [V]_M^i [CLASS] \neq t = [V]_1^j [V]_2^j ... [V]_M^j [CLASS]$$
$$或 t = [V]_1^i ... [V]_{\frac{M}{2}}^i [CLASS] [V]_{\frac{M}{2}+1}^i ... [V]_M^i \neq t = [V]_1^j ... [V]_{\frac{M}{2}}^j [CLASS] [V]_{\frac{M}{2}+1}^j ... [V]_M^j$$
$$i \neq j \ and \ i, j \ \epsilon 1, 2...K$$

图 2. 模板设计

3.3 损失函数定义

$$t = [V]_1^i [V]_2^i ... [V]_M^i [CLASS] \neq t = [V]_1^j [V]_2^j ... [V]_M^j [CLASS]$$
$$或 t = [V]_1^i ... [V]_{\frac{M}{2}}^i [CLASS] [V]_{\frac{M}{2}+1}^i ... [V]_M^i \neq t = [V]_1^j ... [V]_{\frac{M}{2}}^j [CLASS] [V]_{\frac{M}{2}+1}^j ... [V]_M^j$$
$$i \neq j \ and \ i, j \ \epsilon 1, 2...K$$

图 3. 模板设计

4 复现细节

训练的目标是最小化基于交叉熵的标准分类损失，梯度可以一直通过文本编码器 g( ) 反向传播，利用编码在参数中的丰富知识来优化上下文。连续表示的设计还允许在词嵌入空间中进行全面探索，这有助于学习与任务相关的上下文。

4.1 与已有开源代码对比

引用代码地址:https://github.com/KaiyangZhou/CoOp

4.2 实验环境搭建

操作系统：Ubuntu 22.04 GPU: RTX4070s 内存：32G

4.3 创新点

CLIP 的 encode_text $token_embedding$ $positional_embedding$ $token_embedding$ $nn.Embedding$ $clip.$

```python
class CustomCLIP(nn.Module):
    def __init__(self, cfg, classnames, clip_model):
        super().__init__()
        self.prompt_learner = PromptLearner(cfg, classnames, clip_model)
        self.tokenized_prompts = self.prompt_learner.tokenized_prompts
        self.image_encoder = clip_model.visual
        self.text_encoder = TextEncoder(clip_model)
        self.logit_scale = clip_model.logit_scale
        self.dtype = clip_model.dtype
    # 训练CoOp时的前向阶段
    def forward(self, image):
        image_features = self.image_encoder(image.type(self.dtype))

        prompts = self.prompt_learner()
        tokenized_prompts = self.tokenized_prompts
        text_features = self.text_encoder(prompts, tokenized_prompts)

        image_features = image_features / image_features.norm(dim=-1, keepdim=True)
        text_features = text_features / text_features.norm(dim=-1, keepdim=True)

        logit_scale = self.logit_scale.exp()
        logits = logit_scale * image_features @ text_features.t()

        return logits
```

图 5. 代码 2

5 实验结果分析

CoOp 在领域泛化方面表现出色，比 zero-shot CLIP 和线性探针模型更具鲁棒性。在 ImageNetV2、ImageNet-Sketch、ImageNet-A 和 ImageNetR 等目标数据集上，CoOp 的性能优于这些基线模型。这表明 CoOp 学习到的提示不仅在训练数据上有效，还能在未见过的数据分布上保持良好的性能。

实验表明，使用更多的上下文 token 通常会带来更好的性能，但同时也可能导致过拟合。实验结果显示，将类标记定位在中间时，随着上下文长度的增加，性能提升更为显著。然而，更短的上下文长度在领域泛化方面表现更好。CoOp 支持随机初始化和手动初始化两种方法。实验结果表明，简单的随机初始化方法已经足够有效，尽管进一步调整初始化词可能会带来一些改进。

	Source	Target			
	ImageNet	V2	Sketch	A	R
Resnet-101					
ZS-CLIP	58.14	51.34	33.32	21.65	56.06
Linear Probe CLIP	54.66	45.97	19.07	12.74	34.86
CLIP+CoOp（M=16）	64.25	55.11	32.74	22.14	55.01
CLIP+CoOp（M=4）	66.42	55.54	34.89	23.60	58.04
Vit-B/16					
ZS-CLIP	66.45	61.05	47.14	47.77	73.68
Linear Probe CLIP	64.58	56.78	33.76	35.68	58.34
CLIP+CoOp（M=16）	72.05	65.03	46.17	49.1	74.23
CLIP+CoOp（M=4）	71.45	65.64	47.89	50.01	75.41

图 6. 代码 2

6 总结与展望

CoCOp 模型的主要优势在于其数据效率和性能提升。通过自动学习提示模板，CoCOp 能够将预训练的视觉语言模型转化为数据高效的视觉学习者，仅需少量样本就能显著提高性能，击败手工制作的提示。此外，CoCOp 在领域泛化方面表现出色，即使在领域转移的情况下也能保持较好的性能，这比传统的手动提示方法更为鲁棒。自动化的提示工程减少了人工干预的需求，提高了模型部署的效率，而且 CoCOp 项目的开源代码为社区提供了进一步研究和应用的便利。

然而，CoCOp 模型也存在一些局限性。一个主要问题是学到的提示向量难以形成正常的语言表达，这使得模型的解释性降低。此外，CoCOp 对噪声标签较为敏感，这可能影响模型在含有噪声数据的数据集上的表现。训练难度也是一个问题，CoCOp 需要更多的 GPU 资源，且收敛速度较慢，这增加了模型训练的成本和时间。

总的来说，CoCOp 模型在提高视觉语言模型的适应性和性能方面取得了显著进展，但也面临着解释性、对噪声数据的敏感性以及训练资源需求等方面的挑战。未来的工作需要在这

些领域进行改进，以进一步提升 CoCOp 模型的实用性和泛化能力。

参考文献

[1]Radford, A., et al. ”Learning Transferable Visual Models From Natural Language Supervision.” OpenAI, 2021.

[2]Zhou, K., et al. ”Learning to Prompt for Vision-Language Models.” International Journal of Computer Vision (IJCV), 2022.

[3]Zhou, K., et al. ”Conditional Prompt Learning for Vision-Language Models.” IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022.

[4]K. He, X. Zhang, S. Ren, and J. Sun, ”Deep Residual Learning for Image Recognition,” in Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), 2016.

[5]K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, ”Momentum Contrast for Unsupervised Visual Representation Learning,” in Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), 2020.

[6]P. Helber, B. Bischke, A. Dengel, and D. Borth, ”EuroSAT: A Novel Dataset and Deep Learning Benchmark for Land Use and Land Cover Classification,” IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2019.

[7]O. J. Hénaff, A. Srinivas, J. D. Fauw, A. Razavi, C. Doersch, S. M. A. Eslami, and A. van den Oord, ”Data-Efficient Image Recognition with Contrastive Predictive Coding,” in Proc. Intl. Conf. on Machine Learning (ICML), 2020.

[8]D. Hendrycks, S. Basart, N. Mu, S. Kadavath, F. Wang, E. Dorundo, R. Desai, T. Zhu, S. Parajuli, M. Guo, D. Song, J. Steinhardt, and J. Gilmer, ”The Many Faces of Robustness: A Critical Analysis of Out-of-Distribution Generalization,” in Proc. IEEE Intl. Conf. on Computer Vision (ICCV), 2021.