Efficient Test-Time Adaptation of Vision-Language Models

摘要

使用预先训练的视觉语言模型进行测试时的适应对于解决测试期间的分布变化引起了越来越多的关注。尽管先前的研究已经取得了非常有希望的性能，但它们涉及密集的计算，这与测试时间的适应严重不一致。本文设计了 TDA，这是一种无需培训的动态适配器，可以通过视觉语言模型实现有效且高效的测试时间适应。TDA 使用轻量级键值缓存，该缓存维护一个动态队列，其中少量伪标签作为值，相应的测试样本特征作为键。利用键值缓存，TDA 允许通过渐进式伪标签细化逐渐适应测试数据，这是超级高效的，不会产生任何反向传播。此外，本文引入了负伪标签，当模型不确定其伪标签预测时，通过将伪标签分配给某些负类来减轻伪标签噪声的不利影响。两个基准的广泛实验证明了 TDA 与最先进的技术相比具有卓越的有效性和效率。

关键词：测试时间适应；图像识别；视觉语音模型；领域偏移

1 引言

视觉语言模型的最新进展为将人类语言集成到各种计算机视觉任务中打开一扇新的大门。以 CLIP 为例。它可以通过利用从网络规模图像文本对中学习的共享嵌入空间来实现零样本图像分类。在这个共享空间中，可以通过将图像的特征与 CLIP 类的文本嵌入相匹配来直接识别图像。另一方面，CLIP 在处理各种特定下游图像时经常面临挑战，特别是当下游图像与 CLIP 训练图像相比具有明显的域和分布变化时。

测试时适应（Test-Time Adaptation，TTA）旨在让模型在测试阶段快速适应新环境（即测试数据与训练数据的分布差异），这符合现实应用场景的需求。如 TPT 通过学习特定领域的提示来适应视觉 - 语言模型。对于每个测试样本，它利用一组增强函数生成多个随机增强视图，将这些视图输入到 CLIP 模型中生成预测，通过最小化置信预测的边际熵来训练一个可学习的提示。然而，这种方法计算密集，需要大量的计算资源，限制了其在实际场景中的应用。

本文提出了一种免训练动态适配器 Training-free Dynamic Adapter（TDA），TDA 通过维护一个动态队列，以测试样本特征为键，对应的测试样本伪标签为值，能够通过渐进式伪标签细化逐渐适应测试数据。同时，本文受负向学习思想启发，引入负伪标签确定某些类的缺失，当模型对伪标签预测不确定时，给某些负类别分配伪标签。具体是构建额外的 TDA 缓存存储负伪标签，与正缓存结合。

本文的方法有以下优点：

1、高效性：键值缓存是非参数的，测试时不需要反向传播，避免了复杂的计算过程。

2、鲁棒性：引入负伪标签减轻伪标签噪声的不利影响，使 TDA 对伪标签噪声更具鲁棒性，并且能够更好地推广到测试数据。

2 相关工作

2.1 视觉语音模型

视觉语言模型 [1,4] 通过对图像文本数据进行广泛的训练，在有效学习语义表示方面表现出了巨大的潜力。CLIP [7] 在这些模型中脱颖而出，因为它能够在视觉和文本表示之间建立联系，这使其能够在各种下游任务上实现令人印象深刻的零样本结果。为了增强 CLIP 模型在下游分类任务中的迁移学习能力，研究人员提出集成 CoOp [16] 和 CoCoOp [15] 等语言提示学习器，以及 CLIPAdapter [3]、Tip-Adapter [14] 等视觉适配器。尽管这些方法已经显示出相当大的性能改进，但它们通常在下游任务中需要大量的训练数据，这使得它们在现实场景中不太实用。另一方面，这项工作侧重于一种名为测试时间适应的新范式，无需访问原始训练数据。

2.2 测试时适应

测试时适应是指使模型适应可能与训练数据存在分布差异的测试数据的过程。这对于需要模型在不同环境中部署的实际应用特别有益，例如各种天气条件下的自动驾驶、不同医院的医疗诊断等。最近的一些工作 [10,11] 利用每批测试样本来更新部分权重，归一化统计 [8]，或两者的组合 [12]。为了避免使用多个测试样本更新模型，MEMO [13] 建议根据测试数据流中每个样本的不同增强来强制执行不变预测。TPT [9] 通过对每个测试样本微调可学习的提示来解决视觉语言模型的相同挑战。DiffTPT [2] 通过利用预先训练的扩散模型来增强 TPT 中使用的测试数据样本的多样性，从而创新了测试时提示调整。尽管 TPT 和 DiffTPT 在解决视觉语言模型的测试时适应方面是有效的，但即时学习在计算上是昂贵且耗时的。

3 本文方法

3.1 本文方法概述

本文的目标是通过从测试数据流中收集足够的知识来进行测试时调整，并通过调整图像特征来改进预测。受 TipAdapter 概念的启发，本文提出了一种免训练动态适配器（TDA），以通过 CLIP 实现高效且有效的测试时间适应。如图 1所示，TDA 包括两个轻量级键值缓存，其中每个缓存存储一个由少量测试特征组成的动态队列作为键，并将相应的伪标签作为值存储。第一个缓存用于正向学习，它通过高置信度预测动态更新键值对以提高准确性。第二个缓存是为负向学习而设计的，旨在通过引入负伪标签来识别类的缺失而不是存在来解决噪声伪标签的不利影响。通过结合正缓存和负缓存，所提出的 TDA 可以在速度和准确性方面实现卓越的性能。

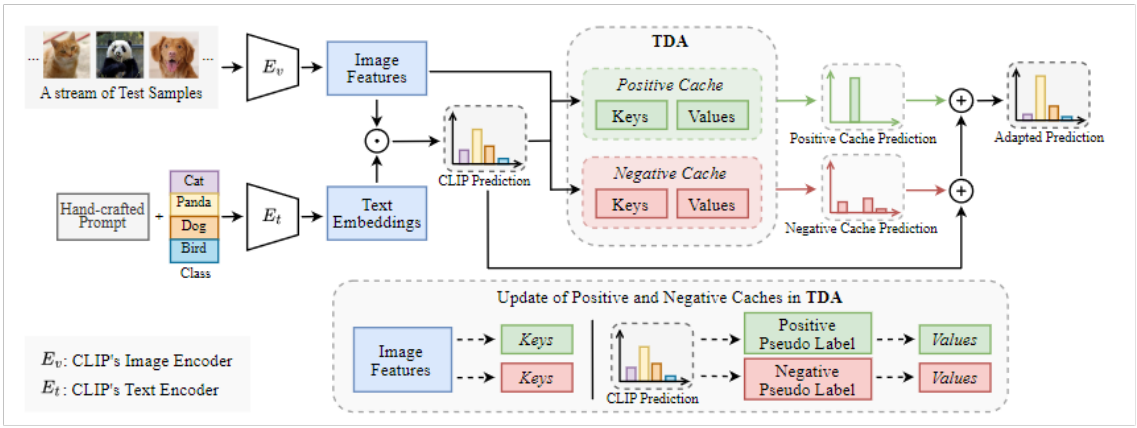

图 1. TDA 方法示意图

3.2 正向缓存

TDA 中的正缓存是键值缓存，其中键和值表示为动态队列，如图 2。它的目的是收集高质量的小样本伪标签 L_p 作为正值，并收集相应的特征 Q_p 作为键。键值缓存最初是空的，然后在测试时适配期间累积足够数量的键值对。为了保持高质量的伪标签，TDA 逐步合并具有较低熵的测试预测，同时限制正缓存中的容量。需要注意的是，每个类都有自己的队列来维护缓存中每个类的顺序和正确的数据结构。

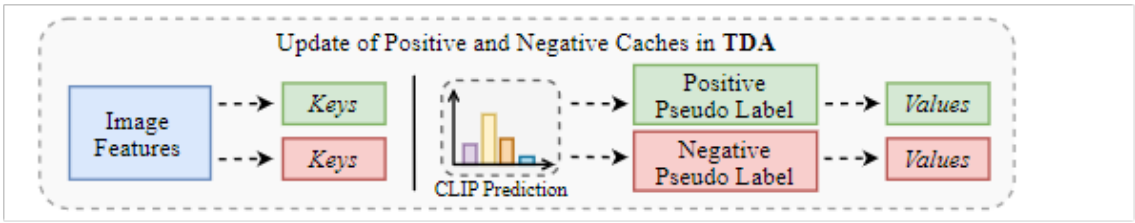

图 2. TDA 缓存示意图

对于每个被添加到 Q_p 的测试样本 f_{test}，根据其预测结果生成伪标签 l。这通过将模型的预测输出转换为独热编码形式实现：

$$\hat{\mathbf{L}}_p = \text{OneHot}(P(Q_p)) \tag{1}$$

本文采用的正向缓存条件如下：

1. 若正向缓存 $\mathbf{L}_p$ 中伪标签的样本数量（每个类收集对数）小于最大样本容量 k，则将伪标签 l 和对应的图像特征 f_{test} 分别作为新值和新键添加到 $\mathbf{L}_p$ 和 $\mathbf{Q}_p$：

$$\text{如果}|\mathbf{L}_p| < k, \quad \mathbf{L}_p \leftarrow \mathbf{L}_p \cup \{l\}, \quad \mathbf{Q}_p \leftarrow \mathbf{Q}_p \cup \{f_{\text{test}}\} \tag{2}$$

2. 如果 $\mathbf{L}_p$ 的样本数量达到 k，当测试样本的熵值低于对应队列中样本的熵值，将替换队列中熵值最高的样本：

$$\text{如果}|\mathbf{L}_p| = k \text{ 且} H(f_{\text{test}}W_c^\top) < H(q^{\text{ent}}W_c^\top), \quad \text{替换} \arg\max(H(\mathbf{Q}_p)) \text{ 中的样本} \tag{3}$$

在测试时适应过程中，将测试图像的特征 f_{test} 作为查询，搜索正向存储的键值对获取缓存信息来调整模型的预测结果。

$$P_{\text{pos}}(f_{\text{test}}) = A(f_{\text{test}}\mathbf{Q}_p^\top)\hat{\mathbf{L}}_p \tag{4}$$

3.3 负向缓存

TDA 的负向缓存与正向缓存类似，通过识别模型预测不确定的类别来减少噪声伪标签的影响。键值对形式为Q_n: 测试样本特征, L_n: 负伪标签，通常为每个类别分别维护一个队列。对于模型预测不确定的样本 f_{test}，负向缓存会生成负伪标签，表示这些样本不属于某些类别。通过对不确定预测（由预测的熵衡量）的类概率应用负掩码获得:

$$\hat{L}_n = -1\left[p_l < P(Q_n)\right] \tag{5}$$

其中，p_l 为负伪标签阈值，高于 p_l 的概率被选作负伪标签。L_n 是向量，元素大于 p_l 为 -1，否则为 0。与现有的负学习方法 [5,6] 从所有噪声标签中选择负标签不同，TDA 从不确定的预测中选择负伪标签，以避免对某些预测的数据产生偏差。

本文采用的负向缓存条件如下：当测试样本特征 f_{test} 满足条件 (f_{test}) 即预测的熵在指定区间内时将其包含在负向缓存中

$$\gamma(f_{\text{test}}) : \tau_l < H(f_{\text{test}} W_c^T) < \tau_h. \tag{6}$$

此条件旨在通过合并表现出中等程度的预测不确定性的测试样本来减轻由于高熵或对某些预测（以非常低的熵为特征）造成的预测错误的风险。一旦完成 $_(f_{\text{test}})$ 检查，在负缓存中收集不确定样本的剩余步骤遵循在正缓存中设计的相同的两个条件。与正缓存类似，TDA 也限制了负向缓存容量 k。

负向缓存预测：在测试时适应期间，通过从负缓存中的Q_n，L_n 检索缓存信息，测试样本特征 f_{test} 可以快速适应目标域，并且适应的预测可以如下获得：

$$P_{\text{neg}}(f_{\text{test}}) = -A(f_{\text{test}} \mathbf{Q}_n^T)\hat{L}_n \tag{7}$$

最终本文 TDA 方法的预测是将正向缓存、负向缓存和预训练 CLIP 模型的预测相结合相结合：

$$P_{\text{TDA}}(f_{\text{test}}) = f_{\text{test}} \mathbf{W}_c^T + P_{\text{pos}}(f_{\text{test}}) + P_{\text{neg}}(f_{\text{test}}) \tag{8}$$

通过这种方式，TDA 利用缓存中的信息逐步适应测试数据，提高预测准确性，同时避免反向传播带来的计算负担，实现高效的测试时适应。

4 复现细节

4.1 与已有开源代码对比

本文的开源代码在作者的 github 中有展示，但下载下来并不能跑通，需要在后续对环境进行进一步配置和修改文件中的训练部分和模型部分的代码后才能运行成功。同时本论文的代码在测试时只是使用正向缓存和负向缓存来调整模型的预测结果，并没有更改模型预测使用的 Text Embeddings，与源代码相比，我在测试时新增了一个模块，使用正向缓存的数据来调整微调模型，从而使模型在测试时使用的 Text Embeddings 能够根据正向缓存的数据进行即时调整，从而使模型适应测试数据的分布，同时我也微调了原论文的部分超参数。

4.2 实验环境搭建

本文方法复现在 linux 系统下进行, 使用的是 python=3.7 的环境，同时使用 A100 的 GPU, 使用 CUDA 版本为 11.6, 并导入本论文对应所需要的库。

4.3 创新点

在源代码的基础上新增了一个模块，使用正向缓存的数据来调整微调模型，从而使模型在测试时使用的 Text Embeddings 能够根据正向缓存的数据进行即时调整，从而使模型适应测试数据的分布。

5 实验结果分析

5.1 实验说明

本论文采用的数据集包括 OOD 基准测试和跨域基准测试

OOD 基准测试：用于评估模型对未见数据的泛化能力，包括 4 个源自 ImageNet 的分布外数据集：ImageNet-A、ImageNet-V2、ImageNet-R 和 ImageNet-S。

跨域基准测试：用于评估模型在不同图像分类数据集上的性能，涵盖 10 个不同领域的数据集，如 Aircraft、Caltech101、Cars 等。

同时所有实验模型都基于预训练的 CLIP 模型，包括图像编码器和文本编码器，评价指标为 top-1accuracy（准确率），整个一次实验需要大约一天至两天的时间。

5.2 实验结果

首先评估 TDA 中两种缓存设计的功效。如图 3所示。正缓存和负缓存都显着超过了基线模型 CLIP，这表明可以通过引入具有正伪标记或负伪标记的动态适配器来改进测试时间自适应。TDA 中的两种缓存设计可以相互补充，因为在具有挑战性的 ImageNet 数据集上，TDA（即两种设计的组合）明显优于正缓存或负缓存。它们的组合产生了百分之 61.03 的准确率，从而凸显了每种类型缓存的重要性提高 TDA 的性能。

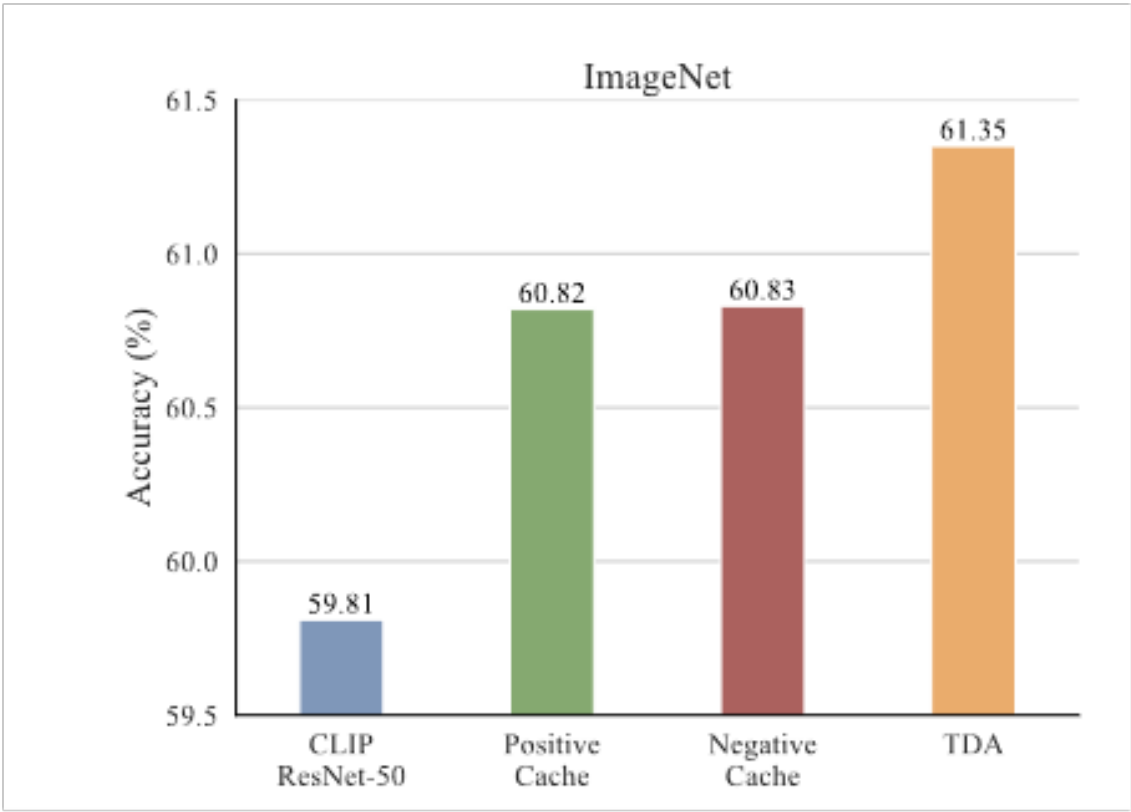

图 3. 图 3：ImageNet 数据集结果

在 OOD 基准上，实验结果如下所示：

Methond	ImageNet	ImageNet-A	ImageNet-V2	ImageNet-R	ImageNet-S	Average
CLIP-ViT-B/16	68.34	49.89	61.88	77.65	48.24	61.2
TDA	69.51	60.11	64.67	80.24	50.54	65.01
TDA优化	69.65	60.21	65.1	80.65	51.02	65.33

图 4. OOD 基准实验结果

在来自 ImageNet 的各种分布外数据集上，TDA 表现出相当优越的性能，验证了 TDA 在分布外测试数据集上提高测试时适应性能的有效性。

在跨域基准上，实验结果如下所示：

Methond	Aircraft	Caltech101	Cars	DTD	EuroSAT	Flower102	Food101	Pets	UCF101	SUN397	Average
CLIP-ViT-B/16	23.22	93.55	66.11	45.04	50.42	66.99	82.86	86.92	65.63	65.16	64.59
TDA	23.91	94.24	67.28	47.4	58	71.42	86.14	88.63	70.66	67.62	67.53
TDA优化	26.1	94.59	67.44	46.92	59	70.65	86.24	89.49	70.83	67.81	67.91

图 5. 跨域基准实验结果

可以看到，TDA 在测试时适应不同类别数据集的有效性，对视觉 - 语言模型在图像分类中对任意类别进行分类具有重大价值。同时使用正向缓存的数据来调整微调模型可以是预测的准确率略微提升，其中 Aircraft 的结果能从 23.91 提升到 26.1，大部分数据集的结果有略微的提升，不过也有的数据集的结果会下降。

6 总结与展望

本文对 TDA 方法进行了解读，TDA 采用键值缓存，维护一个动态队列，以测试样本特征作为键，相应的少样本伪标签作为值，允许通过渐进的伪标签改进来逐渐适应测试数据。此外，TDA 引入了负缓存，当模型不确定其预测时，通过将负伪标签分配给某些类来减轻噪声伪标签的不良影响。并通过大量实验结果表明了复现结果的正确性，同时充分证明了 TDA 方法在测试时间适应方法有一定程度上的优势。同时，TDA 方法可以通过正向缓存的数据来进行优化，不过会增加 TDA 的运行时间，降低了原本的高效率。未来可以考虑充分利用 TDA 框架中正向缓存和负向缓存的数据。

参考文献

[1] Karan Desai and Justin Johnson. Virtex: Learning visual representations from textual annotations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11162–11173, 2021.

[2] Chun-Mei Feng, Kai Yu, Yong Liu, Salman Khan, and Wangmeng Zuo. Diverse data augmentation with diffusions for effective test-time prompt tuning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2704–2714, 2023.

[3] Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. Clip-adapter: Better vision-language models with feature adapters. *International Journal of Computer Vision*, 132(2):581–595, 2024.

[4] Xiaowei Hu, Zhe Gan, Jianfeng Wang, Zhengyuan Yang, Zicheng Liu, Yumao Lu, and Lijuan Wang. Scaling up vision-language pre-training for image captioning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 17980–17989, 2022.

[5] Youngdong Kim, Junho Yim, Juseung Yun, and Junmo Kim. Nlnl: Negative learning for noisy labels. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 101–110, 2019.

[6] Youngdong Kim, Juseung Yun, Hyounguk Shon, and Junmo Kim. Joint negative and positive learning for noisy labels. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9442–9451, 2021.

[7] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[8] Steffen Schneider, Evgenia Rusak, Luisa Eck, Oliver Bringmann, Wieland Brendel, and Matthias Bethge. Improving robustness against common corruptions by covariate shift adaptation. *Advances in neural information processing systems*, 33:11539–11551, 2020.

[9] Manli Shu, Weili Nie, De-An Huang, Zhiding Yu, Tom Goldstein, Anima Anandkumar, and Chaowei Xiao. Test-time prompt tuning for zero-shot generalization in vision-language models. *Advances in Neural Information Processing Systems*, 35:14274–14289, 2022.

[10] Yu Sun, Xiaolong Wang, Zhuang Liu, John Miller, Alexei Efros, and Moritz Hardt. Test-time training with self-supervision for generalization under distribution shifts. In *International conference on machine learning*, pages 9229–9248. PMLR, 2020.

[11] Thomas Varsavsky, Mauricio Orbes-Arteaga, Carole H Sudre, Mark S Graham, Parashkev Nachev, and M Jorge Cardoso. Test-time unsupervised domain adaptation. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part I 23*, pages 428–436. Springer, 2020.

[12] Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. *arXiv preprint arXiv:2006.10726*, 2020.

[13] Marvin Zhang, Sergey Levine, and Chelsea Finn. Memo: Test time robustness via adaptation and augmentation. *Advances in neural information processing systems*, 35:38629–38642, 2022.

[14] Renrui Zhang, Wei Zhang, Rongyao Fang, Peng Gao, Kunchang Li, Jifeng Dai, Yu Qiao, and Hongsheng Li. Tip-adapter: Training-free adaption of clip for few-shot classification. In *European conference on computer vision*, pages 493–510. Springer, 2022.

[15] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16816–16825, 2022.

[16] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022.