中国科学 : 信息科学 2025 年 第 55 卷 第 2 期: 326–342
SCIENTIA SINICA Informationis
论文

迈向可持续自监督学习: 基于目标增强的条件掩码重建自监督学习

高尚华[1], 周攀[2], 程明明[1*], 颜水成[3]

1. 南开大学计算机学院, 天津 300350, 中国
2. School of Computing and Information Systems, Singapore Management University, Singapore 178902, Singapore
3. Skywork AI, Singapore 178902, Singapore
* 通信作者. E-mail: cmm@nankai.edu.cn

收稿日期: 2024–06–10; 修回日期: 2024–08–29; 接受日期: 2024–10–12; 网络出版日期: 2025–01–23

国家自然科学基金杰出青年科学基金项目 (批准号: 62225604) 资助

摘要 自监督学习 (self-supervised learning, SSL) 训练成本日益攀升, 而仅有少数最先进的模型会被应用于下游任务. 为了降低自监督学习的训练成本, 本文探索了一个旨在实现可持续 SSL 训练的框架. 该框架通过高效重复利用现有的 SSL 模型 (即基模型), 以低成本的方式训练出性能更优的新 SSL 模型, 从而有效应对了高昂训练成本的挑战. 该框架设计了兼容机制, 确保新 SSL 模型的训练过程能够与具有不同特性的现有 SSL 模型相适配, 实现了对现有 SSL 模型的最大化重复利用. 为实现上述目标, 我们提出了目标增强的条件掩码重建 (target-enhanced conditional, TEC) 方案, 该方案在基于掩码重建的 SSL 算法中引入了两个新组件. 首先, 我们提出了区域间关系增强策略, 以增强基模型提供的重建目标中的区域间关系信息, 并使新模型利用不完整的输入来预测基模型提供的目标, 从而学习基模型中的语义关系知识. 由于该策略使新模型能够处理不完整的输入, 有效地提升了新模型在区域间关系建模方面的能力, 进而助力新模型实现对基模型的性能超越. 其次, 我们引入了一个条件适配器, 通过调整新模型的预测方式以匹配不同基模型提供的重建目标. 大量的实验结果表明, 本文的 TEC 方案不仅可以加快学习速度, 还可以提升如 MAE 和 iBOT 等现在最优的 SSL 基模型的性能. 该方案朝着可持续的自监督学习迈出了探索性的一步. 本文工作代码在 https://github.com/sail-sg/tec 开源.

关键词 可持续, 自监督学习, 预训练, 图像掩码建模

1 介绍

自监督学习 (self-supervised learning, SSL) 在无监督表征学习领域取得了显著的成功, 在分类 [1,2]、目标检测和分割 [3,4] 等许多下游任务中展现出了惊人的性能提升. 在 SSL 训练过程中, 首先需要构建一个代理任务, 如实例判别任务 [5,6] 或掩码图像建模 (masked image modeling, MIM) [3,4], 然后通过该

引用格式: 高尚华, 周攀, 程明明, 等. 迈向可持续自监督学习: 基于目标增强的条件掩码重建自监督学习. 中国科学: 信息科学, 2025, 55: 326–342, doi: 10.1360/SSI-2024-0176
Gao S H, Zhou P, Cheng M-M, et al. Towards sustainable self-supervised learning: target-enhanced conditional mask-reconstruction for self-supervised learning. Sci Sin Inform, 2025, 55: 326–342, doi: 10.1360/SSI-2024-0176

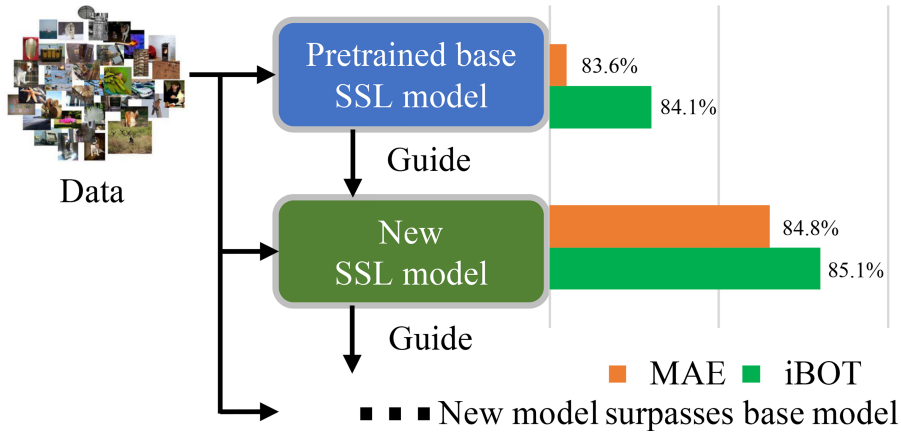

图 1　(网络版彩图) 可持续自监督学习的概念. 一个新的 SSL 模型通过继承预训练 SSL 基础模型的知识, 以实现更优越的表示学习能力, 实现 "可持续" 或可复用的学习, 相比于从头开始训练一个新的 SSL 模型, 这种方式显著提高了学习效率.

Figure 1　(Color online) Concept of sustainable SSL. A new SSL model inherits the knowledge from a pretrained SSL base model to achieve superior representation learning ability for "sustainable" learning. This approach significantly improves learning efficiency compared to training a new SSL model from scratch.

代理任务生成伪标签替代人工标注, 以此训练网络模型. 尽管取得一些成功, 但随着训练数据的增加和模型复杂度的提升, SSL 正朝着需要更高训练成本的方向发展. 例如, MoCo [5] 需要 200 个迭代轮次, 而 MAE (masked autoencoder) [4] 则需要 1600 个迭代轮次才能充分释放其潜力. 不幸的是, 大多数研究人员面临有限的计算预算, 往往难以承担训练大型 SSL 模型所需的巨额成本. 此外, 由于非 SOTA (state-of-the-art) 的预训练 SSL 模型在实践中很少被使用, 且由于 SOTA 性能频繁更新, 之前的 SSL 模型很快便失去价值, 导致大量训练资源的浪费. 因此, 构建一个可持续的 SSL 框架显得尤为重要.

就像人类社会中的知识在代代相传中逐渐扩充一样, 本文试图让新的 SSL 模型在继承先前预训练的 SSL 基模型知识的同时, 增强其相较于基模型的表示学习能力. 以此实现的 "可持续" SSL 相比从头开始训练一个新的 SSL 模型, 在提高学习效率的同时增强了表征能力. 图 1 给出了可持续 SSL 的示意图, 其中本文将待训练的新 SSL 模型简称为新模型, 将预训练的 SSL 模型称为基模型. 为了超越基模型, 在可持续 SSL 中, 新模型不仅要利用基模型隐含的知识, 而且要补充基模型中缺乏的知识. 不同于有监督学习需要标签来实现自我训练 [7,8], 本文的可持续 SSL 学习过程遵循完全自监督的范式. 该过程也可以被视为在自监督学习范式下, 要求新模型比基模型更强大的一种知识蒸馏 [9,10] 的特例.

在这项工作中, 我们通过构建一种能够高效地学习并超越现有预训练 SSL 模型的目标增强的条件掩码重建 (target-enhanced conditional mask reconstruction, TEC) 训练策略, 向可持续 SSL 迈出了探索性的一步. 为了实现这一具有挑战性的目标, 该策略激励新模型不仅学习基模型的知识, 还学习更多与语义相关的新知识. 因此, 本文选择了一种基于掩码重建 [4] 的 SSL 方案来训练新模型, 其中基模型从完整的输入图像中生成重建目标, 而新模型则尝试从随机掩码图像输入中预测该重建目标. 通过这一训练任务, 新模型必须学习输入图像的完整语义及区域间的关系, 以便能够从不完整的输入中推理得到重建目标所需的完整信息. 如图 2 所示, 在 iBOT [1] 预训练策略下的 ViT [11] 模型的注意力图中遗漏了一些语义区域, 例如耳朵. 然而, 以 iBOT 预训练模型为基模型, 经过 TEC 训练后的 ViT 模型捕获了所有区域的语义信息, 并有效地区分了图像的不同组成部分的语义差异. 由于 TEC 具有更强大地捕获综合语义的能力, 因此它有助于实现具有挑战性的可持续 SSL, 并且能够为下游任务提供丰富而灵活的语义表征.

然而, 不同的 SSL 基模型由于其不同的训练目标和策略, 可能展现出各异的属性. 例如, iBOT 预训练模型倾向于具有更多的类别语义特征, 而 MAE [4] 预训练模型则可能保留更多的图像细节特征.

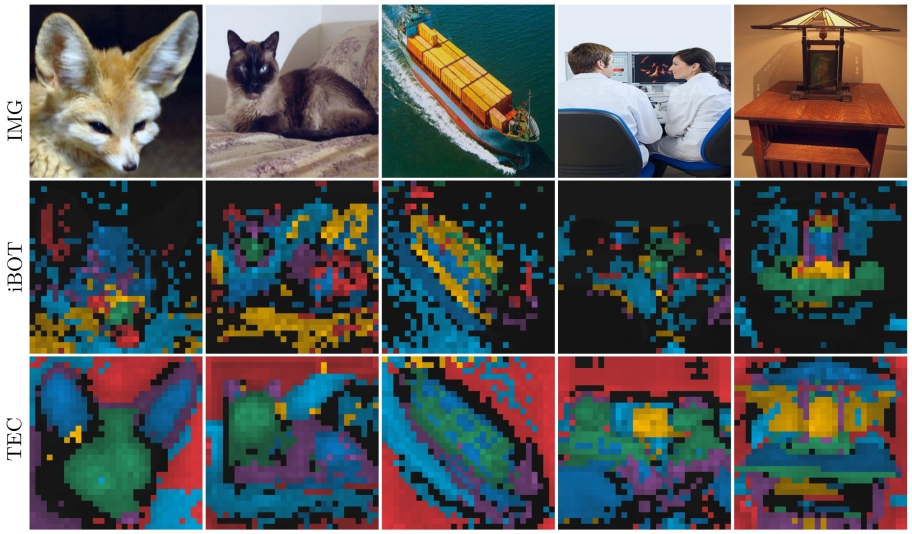

图 2　自注意力模块的可视化图. 该图中每种颜色代表一个自注意力头的注意力. 黑色区域则表明在该区域内, 没有任何一个自注意力头给予显著的关注.

Figure 2　Self-attention visualization. Different colors denote the attention of different heads. The black regions indicate areas where no attention is given by any of the attention heads.

因此, 从基模型中构建高质量且兼容的重建目标显得尤为重要, 这样新模型才能高效地学习到更全面的信息. 一个理想的重建目标应当能够揭示特征的空间语义关系, 比如清晰地展现车轮与车身之间的关联, 从而有利于新模型学习出能够广泛适用于各种下游任务的通用关系特征. 为此, 本文提出采用两个互补的重建目标来提升基模型生成目标的质量: (a) 空间维度归一化的重建目标, 该目标通过对基模型输出特征沿空间维度进行归一化处理, 从而强化特征区域之间的关系属性; (b) 利用基模型中具有丰富语义信息的 token 注意力图作为重建目标, 以此过滤掉噪声并建立整个图像与语义丰富区域之间的紧密联系. 为了兼容来自不同基模型的重建目标, 本文在新模型中引入了条件适配器. 这些适配器使得新模型的预测结构能够灵活地适应具有不同属性的各种基模型. 在给定基模型重建目标的情况下, 适配器能够有条件地激活并调整新模型的中间层特征, 从而更有效地预测重建目标. 这些适配器在预训练阶段完成后通常会被丢弃, 但如果保留下来, 则可用于实现轻量级的微调 [12, 13].

本文将上述实现可持续 SSL 的方法称为目标增强的条件掩码重建 (TEC). 如图 3 所示, 在 ImageNet 数据集上, 无需任何额外训练数据的 TEC 相比 MAE [4] 和 iBOT [1] 等 SSL 基模型的性能有显著改善. 例如, 以 1600 迭代轮次的 iBOT 为基模型, TEC 在仅 800 个迭代轮次下就提高了 1.0% 的分类准确率. 此外, 本文还发现 TEC 可以显著加快 SSL 模型的学习过程, 从而节省训练成本. 例如, 随机初始化的 TEC 在仅训练 100 个批次和 300 个迭代轮次的情况下, 其性能就优于训练了 1600 个批次的 MAE 基模型. 该方法迈出了探索可持续 SSL 的第一步, 希望其能在未来激发更多的工作, 以绿色的方式可持续地改进自监督学习.

2　相关工作

自监督学习. 自监督学习通过利用代理任务进行训练, 而无需人工标注, 来实现表征的预训练. 例如, 实例识别任务 (instance discrimination, ID) 和掩码图像建模任务 (MIM). ID 通过学习拉近一个图像的多个视图中提取的表示, 来学习强类别相关的表征 [14~18]. 从多视图中提取表征相比有监督训练, 通常需要更大的训练成本. MIM 则通过从未掩码部分重建掩码区域 token 的信息来学习语义, 这种方式比 ID 能学习到更多的空间语义细节. 由于需要处理不完整的输入, MIM 通常需要比 ID 更长的

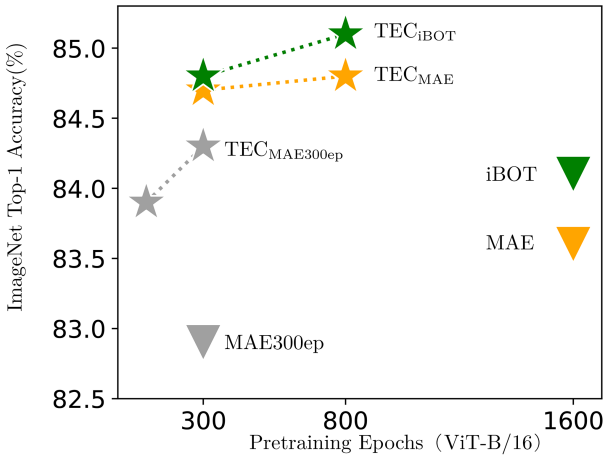

图 3　(网络版彩图) ImageNet-1k 上的 Top1 精度. TEC 预训练的新模型与其对应的基模型采用相同的颜色表示. 三角形代表基模型, 星形代表 TEC 模型. TEC 模型的颜色与其基模型的颜色保持一致. 预训练 1600 个迭代轮次的 iBOT 模型以及预训练 300/1600 个迭代轮次的 MAE 模型被用作此次实验的基模型.

Figure 3　(Color online) Top1 accuracy on ImageNet-1k. The TEC models maintain the same color as their corresponding base models. The triangles represent base models and the stars represent TEC models. TEC models have the same color as their base models. In this experiment, iBOT pretrained with 1600 epochs and MAE pretrained with 300/1600 epochs are used as base models.

训练轮次才能达到收敛. 文献 [19, 20] 探索了结合 MIM 和 ID 的优势, 以进一步提高性能. 最近, 文献 [21] 揭示了 MIM 和 ID 都在学习遮挡不变性特征. 我们观察到一个趋势, 即这些 SSL 方法需要越来越大的计算成本来实现 SOTA 性能, 这阻碍了新的 SSL 方法的发展. 为了解决这个问题, 我们探索了通过从预先训练好的 SSL 模型中学习知识来实现可持续的 SSL.

各种目标上的掩码图像建模. 重建目标指导 MIM 在不同语义空间上进行学习. MIM 已经探索了各种重建目标, 例如, RGB 图像像素和分词器 (tokenizers). 为了使图像的处理方式更接近于自然语言处理 (natural language processing, NLP) 中的离散化语言处理方式 [22], BEiT [3] 采用了 DALLE 预训练的分词器作为预测目标 [23]. CAE [24] 进一步将这种代理任务预测与编码器解耦. MAE 和 SimMIM [25] 的研究表明, 使用 RGB 图像作为重建目标可以实现具有竞争力的完全微调性能. MaskFeat [26] 揭示了手工设计的 HOG 特征 [27] 是一种有效的目标形式. Ge2-AE [28] 和 MFM [29] 发现图像频域信息可以与 RGB 图像目标形成互补. 在 PeCo [30] 中, 感知编码本帮助模型学习语义信息. iBOT 和 data2vec [31] 则利用在线动量网络 [5] 来提供在线更新的预测目标. BootMAE [32] 同时利用了 RGB 图像和在线更新的目标进行训练. 文献 [33] 通过掩码方案增强了从大型教师模型到紧凑学生模型的蒸馏过程. MVP [34] 以 CLIP 预训练模型 [23] 为目标, 引入了从视觉语言预训练中学习到的丰富语义. 与这些强调特定重建目标独特属性的工作不同, 我们的研究表明, 在 TEC 的帮助下, 所有 SSL 预训练模型都可以作为良好的基模型. TEC 中的适配器和目标增强方案使其能够灵活地适应各种基模型产生的目标.

自监督知识蒸馏. 可持续的 SSL 可以被视为自监督知识蒸馏的特殊情况, 因为它们都从 SSL 预训练的模型中学习. Reversed KD [35] 表明在有监督设置下, 一个弱教师模型可以使学生模型受益. ClusterFit [36] 对聚类伪标签进行训练, 以减少对代理任务的过拟合. SEED [37] 使用对比损失将大型 SSL 模型中的知识蒸馏到小型模型中. 文献 [38] 使用 MLP (multilayer perceptron) 头进行特征回归, 将大型 SSL 教师模型蒸馏为紧凑的学生模型. 文献 [39] 使用教师模型对实例进行分组, 并将实例关系知识传递给学生模型. 作为例外, 文献 [40] 表明特征蒸馏改善了基于对比的 SSL 模型, 但对 SOTA 的 MAE [4] 模型带来的增益有限. 本文可持续的 SSL 以一种自监督的方式使新模型优于基模型. 我们将在 4.2 小节中表明, 本文的 TEC 方法比几种 SOTA 自监督蒸馏方法更有优势.

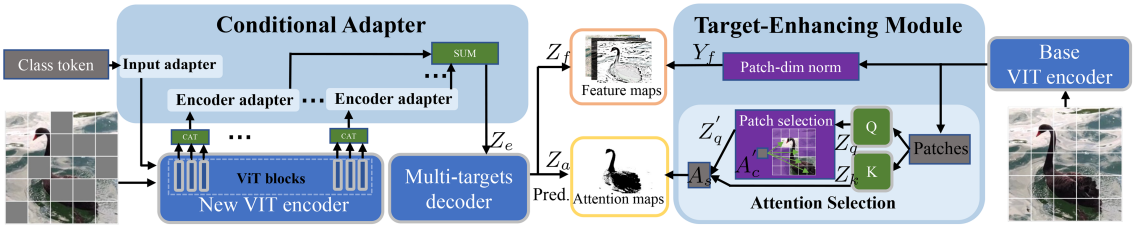

图 4　(网络版彩图) 本文提出的 TEC 的总体框架. TEC 中预训练的 SSL 基模型生成区域间关系增强的重建目标, 即空间维度归一化特征和语义相关的注意力图. 新的 ViT 编码器接收掩码图像和通过输入适配器增强的类别 token, 随后, 将生成的特征依次传递给编码器适配器和多目标解码器, 以预测基模型所给定的目标.

Figure 4　(Color online) Overall framework of the proposed TEC. The pretrained SSL base model in TEC generates patch-relation enhanced reconstruction targets, i.e., patch-dim normalized features and semantic attention maps. The new ViT encoder takes in a masked image and the class token enhanced by the input adapter, and then sequentially passes the generated features into encoder adapters and the multi-target decoder to predict the targets given by the base model.

3　方法

3.1　总体框架

本文所提出的目标增强的条件掩码重建方法的总体框架如图 4 所示. TEC 依照文献 [3, 4] 采用 vision transformer (ViT) [11] 来实现算法. 在掩模重建框架 [4] 下, TEC 由待预训练的新 ViT 编码器、用于条件预训练的条件适配器、用于重建目标预测的多目标解码器、作为基模型的 SSL 预训练 ViT 编码器, 以及一个用于从基模型构建特征关系以增强重建目标的目标增强模块组成. 具体来说, 基模型是 SSL 预训练的 ViT 编码器 (例如 MAE [4]), 并用于生成完整图像的隐式语义特征. 然后, 目标增强模块对隐式语义进行增强, 构建两个互补的重建目标作为新模型的监督信号. 配备适配器的新 ViT 编码器接收掩码图像输入并生成适配后的隐式语义特征, 随后将这些特征输入多目标解码器以预测由基模型提供的重建目标. 在预训练之后, 新 ViT 编码器被保留用于下游任务, 而其他部分则被移除. 我们将在 3.2 小节中介绍通过适配器辅助的条件预训练, 以帮助新模型有效地预测基模型目标, 并在 3.3 小节中详细介绍用于生成高质量基模型重建目标的目标增强模块.

3.2　条件预训练

如前所述, 基模型通常具有不同的属性, 例如, iBOT 中含有更多全局类别语义, 而 MAE 中更多的是局部细节. 因此, 新模型的重建预测应该与任意给定的基模型兼容. 为了解决图像像素重建的类似问题, 文献 [20, 32, 41] 通过试错的方式从编码器的中间层手动选择某些特征, 以更好地与图像像素目标对齐. 然而, 从某些固定层中手动选择与不同基模型兼容的特征几乎是不可能的, 因为它们可能具有不同的属性. 因此, 为了更好地预测基模型目标, 对于给定的 SSL 基模型, 新模型必须具有条件适应能力.

给定一个固定的预训练模型, 轻量化微调方案将具有少量参数的可训练额外模块引入预训练模型, 使其适应视觉 [12, 13] 和自然语言处理 [42~44] 域的下游任务. 例如, 提示 (prompt) 方案 [12, 43, 44] 将可学习的输入 token (例如, 类别 token) 与特征 token 连接起来, 以激活固定的 ViT 模型的某些适用于特定的下游任务的语义特征. 此外, 将轻量级适配器模块 (例如, MLP [13, 42] 和残差 token [45]) 整合到固定模型中, 可以对模型的中间层特征进行调整, 从而预测下游任务所需的特征. 受这些轻量化微调方案的启发, 我们将调整方案引入到预训练阶段, 通过为新模型配备条件适配器来处理基模型的多样性. 由于我们的适配器仅用于预训练, 并将在微调期间删除, 因此它们不会增加额外的推理成本. 此外, 将这些适配器保留在推理阶段可以进一步增强模型的轻量化微调能力. 下面将介绍如何将适配器, 即输入适配器和编码器适配器, 应用到新模型的 ViT 编码器中.

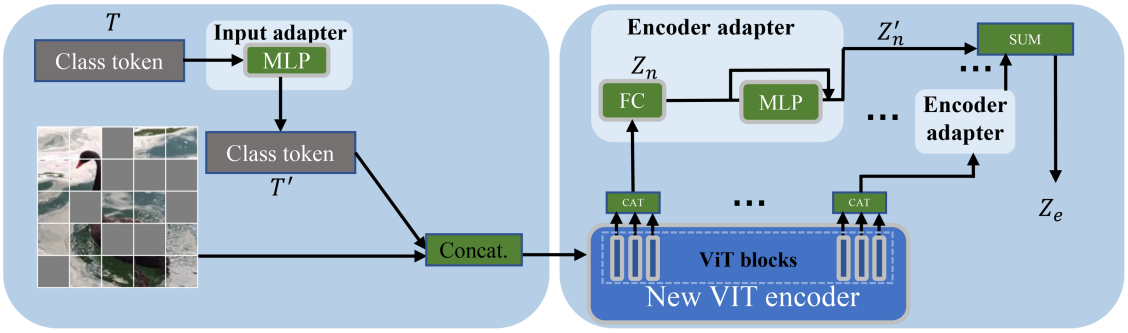

图 5　(网络版彩图) 条件预训练的输入适配器和编码器适配器的细节图.

Figure 5　(Color online) Details of input adapter and encoder adapters for conditional pretraining.

输入适配器. 对于 ViT 网络, 类别 token 通常被与输入特征 token 连接起来, 以学习整个输入的全局语义. 由于提示方案证明了类别 token 对网络的调整能力, 我们提出通过增加输入适配器来进一步增强类别 token 的特征调整能力. 如图 5 所示, 由一个小的两层 MLP 层构成的输入适配器增强了类别 token 的表示能力, 使类别 token 能够更好地根据基模型目标激活新模型中的特征. 具体来说, 我们使用 MLP 层对 ViT 的类别 token $T \in \mathbb{R}^C$ 进行处理, 得到增强的类别 token $T' \in \mathbb{R}^C$: $T' = \mathbf{MLP}(T)$, 其中 C 是特征向量维度. 在预训练期间, T' 被附加到特征 token 中. **MLP** 增强了 T 的表示能力, 使新模型能够更好地预测基模型目标. 对于推理过程, 由于 $\mathbf{MLP}(T)$ 是所有输入样本共享的, 所以可以提前计算得到新的类别 token T', 这意味着 $\mathbf{MLP}(T)$ 的计算不会给推理过程带来额外的开销.

编码适配器. 为了调节新模型中的中间层特征, 使其能够适应基模型的目标, 我们在预训练阶段应用了一个带有残差连接的简单 MLP [13] 作为 ViT 的编码器适配器. 由于我们希望在预训练后去除适配器以获得更高的推理效率, 因此需要在去除适配器后保持编码器网络拓扑结构不变. 于是我们将适配器的输入放置在编码器的中间位置, 并在编码器的输出位置合并所有适配器的输出. 如图 5 所示, 从编码器每个中间层得到的特征集 $X = \{X_i, i = 1, \ldots, D\}$, 其中 D 为编码器中间层的总数, 首先将它们统一分成 N 组, 其中每组默认包含 3 个中间层. 在第 n 组中, 合并来自所有该组中间层的特征

$$Z_n = \mathbf{FC}(\mathbf{Concat}(X_i, \ldots, X_j)).$$

然后将特征 Z_n 输入到适配器中, 得到一个整体特征 Z_e:

$$Z'_n = Z_n + \mathbf{MLP}(Z_n), \qquad Z_e = \sum\nolimits_{n=1}^{N} Z'_n, \tag{1}$$

其中, **MLP** 是一个包含两层全连接层的 MLP. 随后, 将调整后的特征输入多目标解码器来预测基模型的目标, 这将在 3.3 小节中详细介绍.

3.3　区域间关系增强的重建目标

为了更好地利用基模型的知识以实现可持续的 SSL, 我们的目标增强模块设计了两个具有增强区域间关系的互补目标: 其中一个是在空间维度上进行归一化的特征级目标, 旨在增强特征 token 之间的关系; 另一个是语义相关的注意力图, 用于学习语义特征 token 与其他特征 token 之间的关联. 特征级目标旨在揭示特定 token 的语义特性, 而注意力图则更加聚焦于特征 token 之间的相互关系.

空间维度归一化的特征级目标. 在给定一个基模型的输出特征作为目标时, 我们提出将该特征沿空间维度进行归一化, 以增强空间区域间的关系. 具体来说, 对于一个输入, 假设其基模型目标为 $Y \in \mathbb{R}^{L \times C}$, 其中 L 和 C 分别表示 token 数量和通道维度. 然后, 沿着 token 维度对 Y 进行归一化

$$Y_f = (Y - \mu_L)/\sigma_L, \tag{2}$$

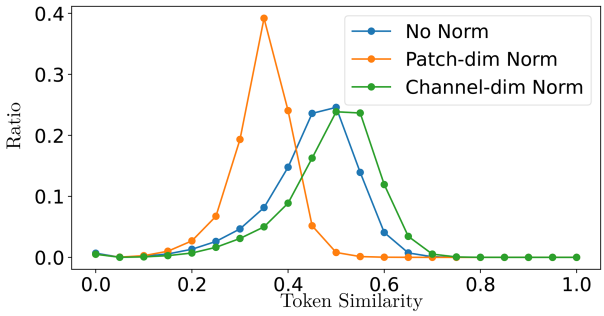

图 6　(网络版彩图) MAE 模型中 token 之间相似度的具体分布情况.

Figure 6　(Color online) Distribution of token similarity in the MAE model.

其中, μ_L 和 σ_L 分别为沿 token 维度的均值和方差. 对于 MIM, 这种空间维度归一化比广泛使用的在通道维度上的特征归一化 [26, 31, 40] 能更好地增强 token 之间的空间关系. 这是因为, 从图 6 中可以看出, 在使用 MAE 预训练得到的模型中, 由于所有 token 可能更多地反映了图像的全局语义, 因此不同 token 的基模型特征值较为相似, 即相似度值很高. 这并不能很好地揭示这些 token 之间的空间关系. 因此, 由于掩码 token 的特征与可见 token 的特征具有较高的相似性, 模型可以很容易地重建掩码 token 的特征目标. 通道维度归一化仅考虑了 token 内的均值和方差, 难以增强 token 之间的关系. 实际上, 如图 6 所示, 通道维度归一化甚至扩大了 token 之间的相似性. 而空间维度归一化则确保了每个通道内的值有明显的差异, 通过显著降低 token 之间的相似性, 增强了 token 之间可能存在的空间关系. 此外, 从后文的实验可以看出, 本文提出的归一化方法可以显著提高新模型的性能. 在归一化之后, 依照文献 [4], 新模型在解码器生成后使用一个全连接层生成 Z_f, 用于预测掩码区域的基模型目标 Y_f

$$L_{\text{fea}} = \|M \circ (Y_f - Z_f)\|_2^2, \tag{3}$$

其中, M 是掩码矩阵, $\circ$ 表示元素乘积.

语义相关注意力图作为目标. 预训练 ViT 模型中的自注意力具有很强的捕获特征 token 之间语义关系的能力 [46~48]. 我们随之提出利用自注意力图作为 MIM 的一种重建目标, 进一步增强新模型的语义关系建模能力. 根据之前关于自注意力图在 KD 中的作用的研究 [49, 50], 我们发现并不是所有的注意力图都包含有用的语义关系, 有严重噪声的注意力甚至会阻碍学生模型的学习. 因此, 有必要选择部分注意力图, 以减少可能出现的严重噪声, 同时也有助于降低训练成本. 本文利用包含足够全局语义的基模型的类别 token 来选择最相似的特征 token, 从而过滤掉可能的噪声. 如图 7 所示, 给定基模型中最后一个 ViT 层的类别 token 和特征 token 之间的注意力图 $A_c \in \mathbb{R}^{H\times L}$, 其中 L 和 H 分别表示 token 数量和注意力头数量, 我们对注意力图 A_c 沿注意力头的维度求平均, 得到 $A'_c \in \mathbb{R}^{1\times L}$. 然后, 如图 7 所示, 选取 A'_c 中值最大的 top-k 个 token, 然后计算 top-k token 与所有特征 token 之间的注意力图 $A_p \in \mathbb{R}^{H\times k\times L}$. 考虑到类别 token 的重要性, 进一步将其自身与所选 A_p 之间的注意力图合并, 得到最终的重建目标, 即 $A_s \in \mathbb{R}^{H\times(k+1)\times L}$. 在计算 A_s 时, 在 Softmax 操作之前添加了一个温度 τ 来调整注意力的清晰度. 对于新模型, 分别使用两个全连接层, 及图 4 中的 Q 和 K, 将其解码器输出映射到两个预测 $Z_q \in \mathbb{R}^{L\times C}$ 和 $Z_k \in \mathbb{R}^{L\times C}$ 中. 我们从 Z_q 中选取与 A_s 中相同的 token, 形成 $Z'_q \in \mathbb{R}^{k\times C}$. 然后将新模型中的类别 token cls 与 Z'_q 拼接, 并计算出 KQ 注意力图 $Z_a = \text{Softmax}([Z'_q, \mathsf{cls}]^{\mathrm{T}} Z_k) \in \mathbb{R}^{H\times(k+1)\times L}$. 最后, 计算预测 Z_a 与目标 A_s 之间的预测熵损失

$$L_{\text{att}} = -A_s \log Z_a. \tag{4}$$

多目标解码器. 由于特征目标和注意目标这两个重建目标的性质不同, 在新模型中, 一个解码器难以同时处理这两个重建目标, 且往往会导致预测冲突. 然而, 对每个目标使用单独的解码器会增加

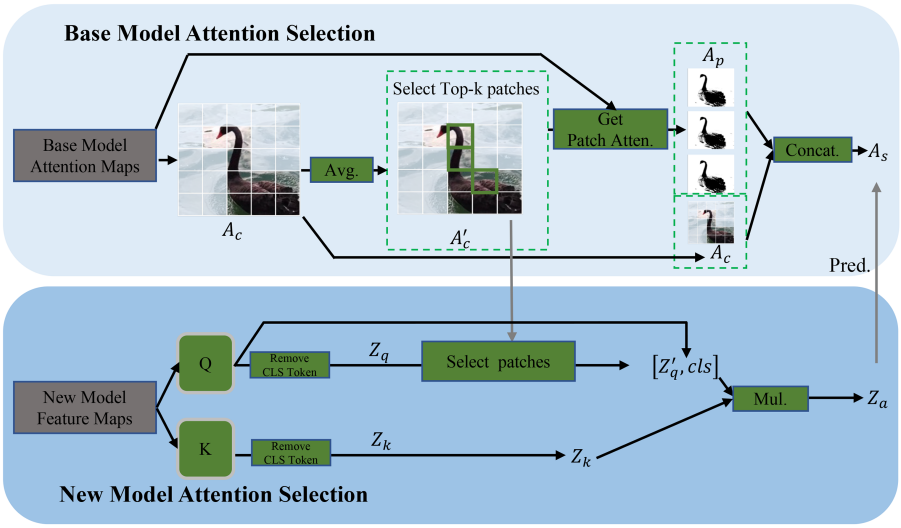

图 7　(网络版彩图) 语义注意力图选择模块的细节图.

Figure 7　(Color online) Details of the semantic attention map selection.

可训练参数的数量, 从而减慢训练速度. 为了解决这个问题, 我们采用了一个简单的解码器适应方案, 即构造特定于目标的输入特征, 然后将它们馈送到共享解码器中. 具体来说, 将新模型编码器的输出特征 Z_e (见式 (1)) 输入到一个全连接层中, 然后用一个可学习的掩码 token 填充掩码 token 以获得 Z'_f. 类似地, 给定 Z_e, 也使用一个全连接层和一个可学习的掩码 token 来获得 Z'_m. 接下来, 将 Z'_f 和 Z'_m 分别输入到一个共享的基于 transformer 的解码器中, 用于预测基模型的特征和注意力映射目标. 与 MAE 中编码器输出和普通图像之间的巨大语义差距不同, 基模型目标与新模型预测具有相似的语义. 因此, 一个浅的两层解码器就足够了, 并且其效果比 MAE 中使用的 8 层解码效果器更好. 这种设计也大大降低了训练成本.

4　实验

在 ImageNet-1k [51] 上, 通过预训练随机初始化的 ViT [11] 模型来评估本文的 TEC 方案. 训练过程中, 使用了 16×16 的 token 尺寸和 224×224 的图像分辨率, 并通过 AdamW [52] 优化器, 以 4096 的批次大小训练 300/800 个迭代轮次. 为了确保性能的改进完全来自于 TEC, 我们没有使用任何显式或隐式的额外训练数据, 也没有使用比新模型更强的基模型. 实际上, 我们使用了在 ImageNet-1k 上由 iBOT [1] 和 MAE [4] 预训练的 ViT 模型作为基模型, 这些基模型都是从它们的官方公开发布版本中获取的. 此外, 我们采用了与 MAE 相同的掩码策略, 例如使用了 75% 的随机掩码比.

4.1　性能对比

4.1.1　在 ImageNet 数据集分类任务的性能对比

ImageNet-1k 数据集分类任务微调结果. 表 1 [1~4, 6, 19, 24~26, 28~32, 34, 40, 46, 53~59] 总结了在 ImageNet-1k 上的分类任务微调性能. 可以观察到, 以 iBOT 为基模型, 从随机初始化开始训练 300 个轮次后, TEC 超出了基模型 0.7%, 经过 800 个训练轮次 TEC 提高了 1.0%. 同样, 在 300/800 训练轮次下, TEC 相对 MAE 基模型分别带来 1.1% 和 1.2% 的提升. 这些结果表明, 我们提出的目标增强的条件掩码重建方案 (TEC) 实际上可以进一步优化基于 MIM 的强大方法, 例如本文使用的 MAE 和 iBOT. 此外, 表 1 还显示, 在训练成本相似甚至更低的情况下, TEC 优于其他先进的 SSL 方法, 包括使用隐式额外数据进行训练的方法, 如 MVP [34] 和 FD-CLIP [40]. 更令人惊讶的是, 仅使用 ImageNet-1k

表 1　在使用 ViT 进行 ImageNet-1k 微调下与现有 SSL 方法的比较. † 和下划线分别表示使用了隐式/显式额外数据. TEC 的预训练 epoch 数是指在基模型的指导下, 随机初始化权重模型进行训练的迭代轮次, 该数字并不包括基模型的迭代轮次. 我们用于比较的结果均来源于各方法官方报告的数据.

Table 1 Comparison with existing SSL methods under ImageNet-1k finetuning using ViT. † and underline mean the usage of implicit/explicit extra data, respectively. The pretraining epoch number of TEC denotes the one from randomly initialized weights under the guidance of base models, and does not include that of the base model. The compared results are obtained from their reported results.

Model	Method	Epoch	Guidance	Top1 acc. (%)
ViT-Base	Deit III [53]	800	Supervised	83.8
	DINO [46]	300	NA	82.8
	MoCov3 [6]	300	NA	83.2
	MixMIM [54]	300	RGB	83.2
	MFM [29]	300	Frequency	83.1
	BEiT [3]	800	DALLE†	83.2
	SplitMask [55]	300	NA	83.6
	ConMIM [56]	800	Momentum	83.7
	SimMIM [25]	800	RGB	83.8
	SIM [57]	1600	Momentum	83.8
	CAE [24]	1600	DALLE†	83.9
	MaskFeat [26]	1600	HOG	84.0
	LoMaR [58]	1600	RGB	84.1
	BootMAE [32]	800	RGB+Momentum	84.2
	data2vec [31]	800	Momentum	84.2
	Mugs [2]	1600	NA	84.3
	MVP [34]	300	CLIP†	84.4
	PeCo [30]	800	Perceptual codebook	84.5
	CMAE [19]	1600	RGB	84.7
	Ge2-AE [28]	800	RGB+Frequency	84.8
	FD-CLIP [40]	300	CLIP†	84.9
	MAE [4]	300	RGB	82.9
	TEC	100	MAE300ep	$83.9_{+1.0}$
	TEC	300	MAE300ep	$84.3_{+1.4}$
	MAE [4]	1600	RGB	83.6
	FD-MAE [40]	300	MAE	$83.8_{+0.2}$
	TEC	300	MAE	$84.7_{+1.1}$
	TEC	800	MAE	$84.8_{+1.2}$
	iBOT-ImageNet-22K	–	Momentum	84.4
	iBOT [1]	1600	Momentum	84.1
	SemMAE [59]	800	iBOT	$84.5_{+0.4}$
	TEC	300	iBOT	$84.8_{+0.7}$
	TEC	800	iBOT	$85.1_{+1.0}$
ViT-Large	MAE [4]	1600	RGB	85.9
	TEC	300	MAE	$86.5_{+0.6}$

数据的 TEC 比使用 ImageNet-22k 训练的 iBOT 提高了 0.7%, 这表明 TEC 的预训练效果比单纯依赖更多训练数据还要有效. 据我们所知, 仅使用 ImageNet-1k 时, TEC 配合 ViT-B 模型达到的 85.1% 性

表 2　在轻量化微调下, ImageNet-1k 数据集的分类 Top1 精度.

Table 2　Top1 accuracy on the ImageNet-1k dataset under parameter-efficient finetuning.

Method	Epoch	Setting	Top 1 acc. (%)
MAE	1600	Linear probing	68.0
TEC_{MAE}	800	Linear probing	69.8
		+Input adapter FT	72.6
		+Encoder adapter FT	79.9

表 3　在 ImageNet-S 数据集上的半监督语义分割任务结果.

Table 3　Semi-supervised semantic segmentation on the ImageNet-S dataset.

Pretrain	Method	Epoch	mIoU_{val} (%)
SSL	MAE	1600	38.3
	TEC_{MAE}	800	42.9
SSL+FT	MAE	1600+100	61.0
	TEC_{MAE}	800+100	62.0

能创下了新的 SOTA 记录, 显示了可持续 SSL 学习的巨大潜力. 我们还使用 ViT-Large 研究了 TEC 的扩展能力, 并观察到, 从随机初始化开始训练 300 个迭代轮次后, TEC 比基于 MAE 预训练的基模型性能高出 0.6%.

ImageNet-1k 数据集分类任务轻量化微调结果. 例如线性预测 (linear probing) 等轻量化微调方法旨在通过微调少量参数来适应下游任务. 我们在线性预测设置下测试了 TEC, 该设置冻结预训练模型中的参数, 仅对输出线性分类器进行微调. 表 2 展示了在线性预测下, VIT-B 在 ImageNet-1k 上的分类精度. 与 MAE 基模型相比, TEC 的分类精度提高了 1.8%, 这表明新学习到的模型中包含更多与类别相关的语义信息. 事实上, 我们用于预训练的输入适配器和编码器适配器同样适用于轻量化微调. 通过对输入适配器进行微调, 可以显著提高 4.6% 的精度. 同时微调输入适配器和编码器适配器, 与 MAE 基模型相比, 可以提高 11.9% 的精度. 这也进一步凸显了本文所提出的适配器的优势.

ImageNet-S 数据集的语义分割任务微调结果. 为了测试 TEC 预训练模型的像素级表示能力, 我们在拥有像素级训练标签的 ImageNet 子数据集 ImageNet-S [60] 上进行了语义分割微调. 由于预训练和微调数据间不存在域偏移, 我们采用了不带额外分割头的 ViT-B 作为分割模型. 从表 3 可以看出, TEC_{MAE} 在 mIoU 上相较于 MAE 基模型提高了 4.6%. 当使用经过有监督 ImageNet 完全微调的预训练模型时, TEC_{MAE} 相比 MAE 有 1.0% 的性能提升.

4.1.2　下游任务的迁移学习性能

本小节研究了 TEC 模型在下游任务中的迁移学习能力.

语义分割. 在 ADE20k 数据集 [61] 上进行语义分割时, 我们采用了配备 ViT-B 的 Upernet [62] 作为分割模型. 从表 4 可以看出, TEC_{iBOT} 相较于 iBOT 基模型, mIoU 提高了 1.0% 而 TEC_{MAE} 则比其 MAE 基模型提高了 1.8%. 因此, 与它们的基模型相比, TEC 预训练模型在语义分割任务中展示出了更强的迁移学习能力. 此外, 在预训练轮次较少的情况下, TEC 相较于竞争对手, 如 MAE, CAE [24] 和 CMAE [19], 分别高出了 2.9%, 0.8% 和 0.9%, 实现了新的 SOTA 性能.

实例分割. 对于 COCO 数据集上的实例分割任务 [63], 为了确保公平性, 将 iBOT [1] 和 ViTDet [45] 所实现的 Cascade MaskRCNN [64] 应用于基于 iBOT/MAE 基模型的 TEC 之上. 根据表 5 的数据显示, 当使用 iBOT 的实现时, TEC 在 box AP 上相较于 iBOT 基模型提升了 1.5%, 在 mask AP 上则提升了 1.2%; 而当采用 ViTDet 的实现时, TEC 在 box AP 上也获得了 0.6% 的提升, 在 mask AP 上则

表 4　使用 Upernet 和 ViT-B 在 ADE20k 数据集上语义分割任务的性能对比.

Table 4　Performance comparison of semantic segmentation on ADE20k using Upernet and ViT-B.

Method	Epoch	mIoU (%)
BEiT	800	47.1
PeCo	800	48.5
GE2-AE	800	48.9
CAE	1600	50.2
CMAE	1600	50.1
MAE	1600	48.1
TEC_{MAE}	800	49.9
iBOT	1600	50.0
TEC_{iBOT}	800	51.0

表 5　使用 Cascade MaskRCNN 和 ViT-B 在 COCO 数据集上实例分割任务的性能对比.

Table 5　Performance comparison of instance segmentation on COCO using Cascade MaskRCNN and ViT-B.

	Method	AP_{box} (%)	AP_{mask} (%)
Implementation from [1]	iBOT	51.2	44.2
	TEC_{iBOT}	52.7	45.4
Implementation from [45]	MAE	54.0	46.7
	TEC_{MAE}	54.6	47.2

提升了 0.5%. 这些结果均表明, TEC 能够稳定地提升模型在实例分割任务的性能.

4.2　消融实验和分析

本小节对 TEC 进行了消融实验和分析. 在默认情况下, 模型会进行 300 个轮次的预训练, 随后在 ImageNet-1k 数据集上进行微调和评估.

条件预训练. 条件适配器有助于基于不同基模型进行 SSL 预训练. 表 6(a) 显示, 当使用 MAE 和 iBOT 作为基模型时, 适配器分别稳定地提高了 0.4% 和 0.2% 的性能. 为了观察对不同基模型的适应差异, 我们在图 8 中展示了编码器适配器在编码器输出中所占的平均比例, 即式 (1) 中的 Z'_n/Z_e. iBOT 基模型需要适配器从更深的层提取更多特征, 而 MAE 基模型则使适配器更关注浅层特征. 该结果与基模型的特性相吻合, 即 iBOT 基模型含有更多的高级分类语义, 而 MAE 模型则保留了更多的低级图像细节.

不同维度上的特征归一化对比. 我们在空间维度上对目标特征进行归一化, 以凸显 token 之间的相对关系, 这与现有的在通道维度上进行特征归一化的方法不同. 在表 6(c) 中, 对空间维度进行归一化相较于通道维度归一化, 性能提升了 0.3%. 相反, 通道维度归一化与无归一化相比, 并未带来任何性能提升. 通道维度归一化强调通道间的特征差异. 而我们的空间维度归一化则强调 token 之间的关系, 这与 MIM 方案中的 token 预测相匹配. 表 6(a) 显示, 相较于 MAE/iBOT 基模型, 使用空间维度归一化特征进行训练分别带来了 0.6%/0.4% 的性能提升, 显示了其相对于基模型的优越性.

语义相关注意力图. KQ 注意力图通常揭示了 token 之间的语义关系, 因此被用作强化基模型中区域间关系的目标. 表 6(a) 显示, 引入注意力图进一步提升了使用空间维度归一化训练的模型性能. 表 6(f) 比较了不同类型注意力图的效果. 仅使用类别 token 的注意力图并没有改善, 而使用语义相关 token 的注意力比基线提高了 0.2%. 因此, token 之间的关系有助于 MIM 训练. 与使用所有的注意力图相比, 选择语义相关的注意力图可以降低噪声, 从而获得更大的性能提升.

TEC 加快基模型的训练进程. 默认情况下, 我们使用完全预训练的 SSL 模型作为基模型. 为了

表 6　使用 ViT-B 在 ImageNet-1K 数据集分类任务上进行微调的消融分析.

Table 6　Ablation study on ImageNet-1K fully finetuning setting using ViT-B.

(a) Ablation study of proposed modules.

Patch-norm.	Attention	Adapters	MAE base (%)	iBOT base (%)
Base model performance			83.6	84.1
✓			84.2	84.5
✓	✓		84.3	84.7
✓		✓	84.6	84.7
✓	✓	✓	84.7	84.8

(b) Effect of adapters.

Setting	Top1 acc. (%)
MAE base	83.6
No adapter	84.2
+ Input adapter	84.3
+ Encoder adapter	84.6

(c) Patch-norm features.

Setting	Top1 acc. (%)
MAE base	83.6
NA	83.9
Feature dim.	83.9
Patch dim.	84.2

(d) Initialization with base model pretraining.

Setting	Top1 acc. (%)
iBOT base	84.1
Load	84.4
Not load	84.8

(e) TEC accelerates MAE training.

Setting	Epoch	Top1 acc. (%)
MAE	1600	83.6
$\text{TEC}_{\text{MAE1600ep}}$	300	$84.7_{+1.1}$
MAE	300	82.9
$\text{TEC}_{\text{MAE300ep}}$	100	$83.9_{+1.0}$
$\text{TEC}_{\text{MAE300ep}}$	300	$84.3_{+1.4}$

(f) Effect of semantic-related patch attention.

Setting	Top1 acc. (%)
iBOT base	84.1
No attention	84.5
Cls token only	84.5
All attention	84.6
Attention select	84.7

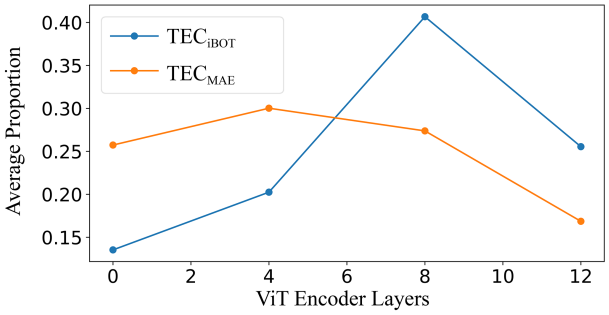

图 8　(网络版彩图) 编码器适配器占编码器输出 Z_e 的平均比例.

Figure 8　(Color online) Average proportion of encoder adapters contributing to the encoder output Z_e.

验证 TEC 是否可以改善未收敛的 SSL 模型, 采用经过 300 个迭代轮次 MAE 预训练的 ViT-B 作为基模型, 并从随机初始化开始, 对 TEC 进行 100/300 个迭代轮次的训练. 如表 6(e) 所示, 300 轮次预训练的 MAE 的 Top.1 准确率为 82.9%. 相比之下, $\text{TEC}_{\text{MAE300ep}}$ 在 300/100 轮次时达到 84.3%/83.9%, 超过了经过 300 个轮次预训练的 MAE 基模型 1.4%/1.0%. 值得注意的是, $\text{TEC}_{\text{MAE300ep}}$ 在仅进行 100 个轮次训练的情况下, 甚至比经过 1600 个训练轮次预训练的 MAE 还要高出 0.3%, 这表明 TEC 可以显著加速基模型的训练过程. 并且, 使用训练了 1600 个轮次的 MAE 模型作为基模型, $\text{TEC}_{\text{MAE1600ep}}$ 相比 $\text{TEC}_{\text{MAE300ep}}$ 进一步提高了 0.4% 的性能, 表明本文可持续学习策略可以依靠好的基模型来取得更好的表现.

是否使用基模型的权重初始化新模型. 在 TEC 框架中, 新模型是通过随机初始化来开始训练的. 表 6(d) 对比了加载与不加载预训练的基模型权重时, 新模型的性能差异. 结果显示, 随机初始化的新模型性能比使用基模型权重进行初始化的模型要好 0.4%. 我们推测, 随机初始化的新模型能够避免陷入基模型的局部极小值, 同时, 新模型能够从基模型中学习到不同的权重分布.

迈向通用的可持续 SSL. 这项工作旨在以现有的预训练 SSL 模型为基础, 向可持续 SSL 迈出第一步. 为了在向可持续的 SSL 方向更进一步, 我们采用 TEC 预训练模型作为新一轮 TEC 预训练的

表 7　使用 TEC 作为新的基模型迈向一般的可持续 SSL.

Table 7　Towards general sustainable SSL using the TEC as the new base model.

Model	Base	Epoch	Top1 acc. (%)
iBOT	–	1600	84.1
TEC_{iBOT}	iBOT	800	85.1
TEC	TEC_{iBOT}	800	85.2

表 8　TEC 与现有 SSL 方法的预训练成本的对比. 我们使用了配置有 8xA100 的服务器来配置评估训练所需的时间. FLOPs (带梯度) 表示需要反向传播的训练 FLOPs, 而 FLOPs (无梯度) 指的是不需要反向传播的 FLOPs.

Table 8　Comparison of pretraining costs between TEC and current SSL methods. Training time is assessed using an 8xA100 server configuration. FLOPs (with grad) represent the training FLOPs requiring backpropagation, while FLOPs (no grad) indicate FLOPs that do not require backpropagation.

Method	Epoch	Time (8xA100) (h)	FLOPs (with grad) (G)	FLOPs (no grad) (G)	Parameters (M)	Top1 acc. (%)
VIT-B	–	–	17.6	–	86.6	–
iBOT	1600	361	19.2	19.2	96.3	84.1
TEC_{iBOT}	300	25	8.3	17.6	118.6	84.8
MAE	1600	125	9.8G	0	111.9	83.6
TEC_{MAE}	300	25	8.3	17.6	118.6	84.7

基模型. 表 7 显示, 以第 1 轮 TEC 基模型为基础训练的第 2 轮 TEC 达到了 85.2% 的准确率. 第 2 轮提升幅度较小的原因可能是网络容量有限, 或者是两轮 TEC 预训练所学习的知识具有较高的相似性.

训练成本比较. 本文添加了 TEC 与其他 SSL 方法的 FLOPs、训练时间和参数比较, 如表 8 所示. TEC 需要的训练时间更短, 且性能优于基模型. 例如, 在训练时间仅为 iBOT/MAE 的 7%/20% 的情况下, TEC 的 Top 1 准确率高出 0.7%/1.1%. TEC 的参数数量与 MAE 相似, 这是因为虽然适配器增加了参数, 但是更少层数的解码器减少了参数. 由于配备了额外的解码器, TEC 和 MAE 与 iBOT 相比, 参数更多. 然而, 得益于解码器仅处理编码器中可见的 token, 它们的训练成本比 iBOT 更低. 在某些 SSL 方法中, 模型只有部分需要梯度, 例如, TEC 中的基模型和 iBOT 中的在线模型, 它们无需反向损失来计算梯度. 因此, 本文比较了有/无梯度的网络部分的 FLOPs. 由于编码器仅处理未掩码的 token, 并且配备了一个浅的两层解码器, TEC 需要的带梯度训练 FLOPs 比 iBOT 和 MAE 更少. TEC 中基模型的额外 FLOPs (无梯度 FLOPs) 小于 iBOT 中的在线网络所需的 FLOPs, 因为 TEC 的基模型不需要额外的预测头. 与 MAE 相比, TEC 中带梯度的较小 FLOPs 可以部分抵消基模型的额外 FLOPs. 因此, TEC 在每次训练迭代中与 MAE 的训练时间相似.

轻量化微调的性能对比. 表 9 报告了 (1) 线性预测的准确性和 (2) 仅通过微调适配器和线性分类器实现的轻量化适配器微调的准确性. 可以观察到: (1) TEC 的线性预测性能受基模型影响; (2) 适配器微调能显著提升性能. 事实上, 大多数基于 MIM 的模型, 例如 BEiT 和 MAE, 其线性预测性能要比全微调低得多. 这是因为它们没有采用有助于分类任务的全局语义学习损失, 例如聚类损失或实例对比判别损失. 这也解释了 TEC 与采用全局语义学习方法 (如 iBOT) 相比性能略低的原因. 然而, 通过微调适配器和线性分类器, TEC 相较于 iBOT 展现出了 3.9% 的显著优势. 原因在于, 如图 2 所示, iBOT 更侧重于区分与全局语义相关的 token, 却忽视了其他 token 的语义, 而 TEC 能够将这些 token 划分为几个语义组, 并进一步识别每个组的语义. 因此, 微调适配器有助于激活下游任务所需的与全局语义相关的语义组, 从而提升模型对全局语义的分辨能力, 并展现出出色的轻量化微调性能.

与自监督蒸馏方法的对比. 本文比较了最近在 ImageNet 上完全微调的几种自监督蒸馏方法的性能. 表 10 显示, 与其他自监督蒸馏方法相比, TEC 方法有着明显的性能提升. 当使用 MAE ViT-B 作为基模型时, TEC 相比 FD 方法提高了 0.9%. 而与同样采用 MIM 方案的 MaskFeat 相比, 当 TEC 使

表 9　线性预测 (LP)、适配器微调 (adapter FT) 和完全微调 (fully FT) 设定下, ImageNet-1k 数据集上的分类 Top1 精度.

Table 9　Top1 accuracy on the ImageNet-1k dataset under linear probing (LP), adapter finetuning (adapter FT), and fully finetuning (fully FT).

Method	Epoch	Settings	Top1 acc. (%)	Fully FT Top1 acc. (%)
BEiT	800	LP	56.7	83.2
SimMIM	800	LP	56.7	83.8
BootMAE	800	LP	66.1	84.2
CAE	800	LP	68.6	83.8
SemMAE	800	LP	68.7	84.5
CMAE	800	LP	73.9	84.7
Ge2-AE	800	LP	75.3	84.8
MAE	1600	LP	68.0	83.6
TEC_{MAE}	800	LP	69.8	84.7
TEC_{MAE}	800	Adapter FT	79.9	84.7
iBOT	1600	LP	79.8	84.1
TEC_{iBOT}	800	LP	78.0	84.8
TEC_{iBOT}	800	Adapter FT	81.9	85.1

表 10　与自监督蒸馏方法的对比.

Table 10　Comparison with self-supervised distillation methods.

Method	Base	Arch	Epoch	Top 1 acc. (%)
MAE	–	ViT-B	1600	83.6
FD_{MAE}	MAE-ViT-B	ViT-B	300	83.8
TEC_{MAE}	MAE-ViT-B	ViT-B	300	84.7
MoCov3	–	ViT-B	300	83.2
$\text{MaskFeat}_{\text{MoCov3}}$	MoCov3-ViT-B	ViT-B	300	83.9
$\text{TEC}_{\text{MoCov3}}$	MoCov3-ViT-B	ViT-B	300	84.5

用 MoCov3 ViT-B 作为基模型时, 性能提升了 0.6%.

5　总结

本文通过向预训练的 SSL 模型学习, 探索了可持续的自监督学习路径. 我们提出了一种目标增强的条件掩码重建学习方案, 旨在学习和超越现有的自监督学习模型. 适配器不仅可以在预训练期间帮助新模型适应各种基模型, 还能作为轻量化的微调模块使用. 我们将掩模重建方案作为基础, 并通过构建具有增强空间关系的预测目标, 来辅助 MIM 预训练, 并以此超越基模型. 本文方法对已有的强大 MIM 预训练方法, 如 MAE 和 iBOT, 进行了进一步的改进, 从而证明了可持续学习的可行性. 本文工作标志着向可持续自监督学习迈出的第一步, 未来我们将继续探索更为通用的多回合可持续自监督学习框架.

补充材料　本文的补充材料见网络版 infocn.scichina.com. 补充材料为作者提供的原始数据, 作者对其学术质量和内容负责.

参考文献

1 Zhou J, Wei C, Wang H, et al. iBOT: image BERT pre-training with online tokenizer. In: Proceedings of International Conference on Learning Representations, 2022
2 Zhou P, Zhou Y, Si C, et al. Mugs: a multi-granular self-supervised learning framework. 2022. ArXiv:2203.14415
3 Bao H, Dong L, Wei F. BEiT: BERT pre-training of image transformers. 2021. ArXiv:2106.08254
4 He K, Chen X, Xie S, et al. Masked autoencoders are scalable vision learners. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2022. 16000–16009
5 He K, Fan H, Wu Y, et al. Momentum contrast for unsupervised visual representation learning. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2020
6 Chen X, Xie S, He K. An empirical study of training self-supervised vision transformers. 2021. ArXiv:2104.02057
7 Xie Q, Luong M T, Hovy E, et al. Self-training with noisy student improves imagenet classification. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2020. 10687–10698
8 Yalniz I Z, Jégou H, Chen K, et al. Billion-scale semi-supervised learning for image classification. 2019. ArXiv:1905.00546
9 Hinton G, Vinyals O, Dean J, et al. Distilling the knowledge in a neural network. 2015. ArXiv:1503.02531
10 Gou J, Yu B, Maybank S J, et al. Knowledge distillation: a survey. Int J Comput Vis, 2021, 129: 1789–1819
11 Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16×16 words: transformers for image recognition at scale. 2020. ArXiv:2010.11929
12 Jia M, Tang L, Chen B C, et al. Visual prompt tuning. 2022. ArXiv:2203.12119
13 Chen S, Ge C, Tong Z, et al. Adaptformer: adapting vision transformers for scalable visual recognition. 2022. ArXiv:2205.13535
14 Chen T, Kornblith S, Norouzi M, et al. A simple framework for contrastive learning of visual representations. In: Proceedings of International Conference on Machine Learning (ICML), 2020. 1597–1607
15 Grill J B, Strub F, Altché F, et al. Bootstrap your own latent — a new approach to self-supervised learning. In: Proceedings of Advances in Neural Information Processing Systems, 2020
16 Chen X, He K. Exploring simple siamese representation learning. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2021
17 Zbontar J, Jing L, Misra I, et al. Barlow twins: self-supervised learning via redundancy reduction. 2021. ArXiv:2103.03230
18 Caron M, Misra I, Mairal J, et al. Unsupervised learning of visual features by contrasting cluster assignments. In: Proceedings of Advances in Neural Information Processing Systems, 2020
19 Huang Z, Jin X, Lu C, et al. Contrastive masked autoencoders are stronger vision learners. 2022. ArXiv:2207.13532
20 Wang L, Liang F, Li Y, et al. RePre: improving self-supervised vision transformer with reconstructive pre-training. 2022. ArXiv:2201.06857
21 Kong X, Zhang X. Understanding masked image modeling via learning occlusion invariant feature. 2022. ArXiv:2208.04164
22 Devlin J, Chang M W, Lee K, et al. BERT: pre-training of deep bidirectional transformers for language understanding. 2018. ArXiv:1810.04805
23 Ramesh A, Pavlov M, Goh G, et al. Zero-shot text-to-image generation. In: Proceedings of International Conference on Machine Learning (ICML), 2021. 8821–8831
24 Chen X, Ding M, Wang X, et al. Context autoencoder for self-supervised representation learning. 2022. ArXiv:2202.03026
25 Xie Z, Zhang Z, Cao Y, et al. SimMIM: a simple framework for masked image modeling. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2022. 9653–9663
26 Wei C, Fan H, Xie S, et al. Masked feature prediction for self-supervised visual pre-training. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2022. 14668–14678
27 Dalal N, Triggs B. Histograms of oriented gradients for human detection. In: Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), 2005. 886–893
28 Liu H, Jiang X, Li X, et al. The devil is in the frequency: geminated gestalt autoencoder for self-supervised visual pre-training. 2022. ArXiv:2204.08227
29 Xie J, Li W, Zhan X, et al. Masked frequency modeling for self-supervised visual pre-training. 2022. ArXiv:2206.07706
30 Dong X, Bao J, Zhang T, et al. PeCo: perceptual codebook for bert pre-training of vision transformers. 2021. ArXiv:2111.12710
31 Baevski A, Hsu W N, Xu Q, et al. Data2vec: a general framework for self-supervised learning in speech, vision and

language. 2022. ArXiv:2202.03555
32 Dong X, Bao J, Zhang T, et al. Bootstrapped masked autoencoders for vision bert pretraining. In: Proceedings of European Conference on Computer Vision, 2022
33 Yang Z, Li Z, Shao M, et al. Masked generative distillation. 2022. ArXiv:2205.01529
34 Wei L, Xie L, Zhou W, et al. MVP: multimodality-guided visual pre-training. 2022. ArXiv:2203.05175
35 Yuan L, Tay F E, Li G, et al. Revisiting knowledge distillation via label smoothing regularization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020. 3903–3911
36 Yan X, Misra I, Gupta A, et al. ClusterFit: improving generalization of visual representations. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2020. 6509–6518
37 Fang Z, Wang J, Wang L, et al. SEED: self-supervised distillation for visual representation. 2021. ArXiv:2101.04731
38 Navaneet K, Koohpayegani S A, Tejankar A, et al. SimReg: regression as a simple yet effective tool for self-supervised knowledge distillation. 2022. ArXiv:2201.05131
39 Xu H, Fang J, Zhang X, et al. Bag of instances aggregation boosts self-supervised distillation. In: Proceedings of International Conference on Learning Representations, 2021
40 Wei Y, Hu H, Xie Z, et al. Contrastive learning rivals masked image modeling in fine-tuning via feature distillation. 2022. ArXiv:2205.14141
41 Gao P, Ma T, Li H, et al. ConvMAE: masked convolution meets masked autoencoders. 2022. ArXiv:2205.03892
42 Houlsby N, Giurgiu A, Jastrzebski S, et al. Parameter-efficient transfer learning for NLP. In: Proceedings of International Conference on Machine Learning (ICML), 2019. 2790–2799
43 Li X L, Liang P. Prefix-tuning: optimizing continuous prompts for generation. 2021. ArXiv:2101.00190
44 Liu X, Zheng Y, Du Z, et al. GPT understands, too. 2021. ArXiv:2103.10385
45 Li Y, Mao H, Girshick R, et al. Exploring plain vision transformer backbones for object detection. 2022. ArXiv:2203.16527
46 Caron M, Touvron H, Misra I, et al. Emerging properties in self-supervised vision transformers. In: Proceedings of IEEE International Conference on Computer Vision, 2021
47 Li Z Y, Gao S, Cheng M M. Exploring feature self-relation for self-supervised transformer. 2022. ArXiv:2206.05184
48 Ziegler A, Asano Y M. Self-supervised learning of object parts for semantic segmentation. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2022. 14502–14511
49 Wu H, Gao Y, Zhang Y, et al. Self-supervised models are good teaching assistants for vision transformers. In: Proceedings of International Conference on Machine Learning (ICML), 2022. 24031–24042
50 Wang S, Gao J, Li Z, et al. A closer look at self-supervised lightweight vision transformers. 2022. ArXiv:2205.14443
51 Deng J, Dong W, Socher R, et al. ImageNet: a large-scale hierarchical image database. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2009. 248–255
52 Loshchilov I, Hutter F. Decoupled weight decay regularization. 2017. ArXiv:1711.05101
53 Touvron H, Cord M, Jegou H. DeiT III: revenge of the ViT. In: Proceedings of European Conference on Computer Vision, 2022
54 Liu J, Huang X, Liu Y, et al. MixMIM: mixed and masked image modeling for efficient visual representation learning. 2022. ArXiv:2205.13137
55 El-Nouby A, Izacard G, Touvron H, et al. Are large-scale datasets necessary for self-supervised pre-training? 2021. ArXiv:2112.10740
56 Yi K, Ge Y, Li X, et al. Masked image modeling with denoising contrast. 2022. ArXiv:2205.09616
57 Tao C, Zhu X, Huang G, et al. Siamese image modeling for self-supervised vision representation learning. 2022. ArXiv:2206.01204
58 Chen J, Hu M, Li B, et al. Efficient self-supervised vision pretraining with local masked reconstruction. 2022. ArXiv:2206.00790
59 Li G, Zheng H, Liu D, et al. SemMAE: semantic-guided masking for learning masked autoencoders. 2022. ArXiv:2206.10207
60 Gao S, Li Z Y, Yang M H, et al. Large-scale unsupervised semantic segmentation. IEEE Trans Pattern Anal Mach Intell, 2023, 45: 7457–7476
61 Zhou B, Zhao H, Puig X, et al. Semantic understanding of scenes through the ADE20K dataset. Int J Comput Vis, 2019, 127: 302–321
62 Xiao T, Liu Y, Zhou B, et al. Unified perceptual parsing for scene understanding. In: Proceedings of European Conference on Computer Vision, 2018. 418–434
63 Lin T Y, Maire M, Belongie S, et al. Microsoft COCO: common objects in context. In: Proceedings of European

Conference on Computer Vision, 2014. 740–755
64 Cai Z, Vasconcelos N. Cascade R-CNN: high quality object detection and instance segmentation. IEEE Trans Pattern Anal Mach Intell, 2021, 43: 1483–1498

Towards sustainable self-supervised learning: target-enhanced conditional mask-reconstruction for self-supervised learning

Shanghua GAO[1], Pan ZHOU[2], Ming-Ming CHENG[1*] & Shuicheng YAN[3]

1. *College of Computer Science, Nankai University, Tianjin 300350, China*
2. *School of Computing and Information Systems, Singapore Management University, Singapore 178902, Singapore*
3. *Skywork AI, Singapore 178902, Singapore*
* Corresponding author. E-mail: cmm@nankai.edu.cn

Abstract Although increasingly training expensive, most self-supervised learning (SSL) models have repeatedly been trained from scratch but not fully utilized since only a few SOTAs are employed for downstream tasks. To mitigate the high training costs of SSL, this work explores a framework aimed at sustainable SSL training. This framework efficiently reuses existing SSL models (referred to as "base" models) to train new SSL models with improved performance at a reduced cost. Additionally, it incorporates an adaptation mechanism that ensures the training of new SSL models is compatible with various base models, maximizing their reuse. To achieve this, we propose a target-enhanced conditional (TEC) scheme, which introduces two components to the existing mask-reconstruction based SSL. Firstly, we propose patch-relation enhanced targets that enhance the target given by the base model and encourage the new model to learn semantic relation knowledge from the base model by using incomplete inputs. This hardening and target-enhancing help the new model surpass the base model, since they enforce additional patch relation modeling to handle incomplete input. Secondly, we introduce a conditional adapter that adaptively adjusts new model prediction to align with the target of different base models. Extensive experimental results show that our TEC scheme can accelerate the learning speed and improve SOTA SSL base models, e.g., MAE and iBOT, taking an explorative step towards sustainable SSL. The source code is publicly available at https://github.com/sail-sg/tec.

Keywords sustainable, self-supervised learning, pretraining, mask image modeling