DOI: 10.11992/tis.202311011
网络出版地址: https://link.cnki.net/urlid/23.1538.TP.20240828.1041.020

视觉深度学习模型压缩加速综述

丁贵广[1,2], 陈辉[2], 王澳[1,2,3], 杨帆[1,2,3], 熊翊哲[1,2,3], 梁伊雯[1,2,3]
（1. 清华大学 软件学院, 北京 100084; 2. 清华大学 北京信息科学与技术国家研究中心, 北京 100084; 3. 涿溪脑与智能研究所, 浙江 杭州 311121）

摘　要: 近年来, 深度学习模型规模越来越大, 在嵌入式设备等资源受限环境中, 大规模视觉深度学习模型难以实现高效推理部署。模型压缩加速可以有效解决该挑战。尽管已经出现相关工作的综述, 但相关工作集中在卷积神经网络的压缩加速, 缺乏对视觉 Transformer 模型压缩加速方法的整理和对比分析。因此, 本文以视觉深度学习模型压缩技术为核心, 对卷积神经网络和视觉 Transformer 模型 2 个最重要的视觉深度模型进行了相关技术手段的整理, 并对技术热点和挑战进行了总结和分析。本文旨在为研究者提供一个全面了解模型压缩和加速领域的视角, 促进深度学习模型压缩加速技术的发展。

关键词: 视觉深度学习; 模型压缩; 轻量化结构; 模型剪枝; 模型量化; 模型蒸馏; Transformer; 序列剪枝
中图分类号: TP18　　**文献标志码:** A　　**文章编号:** 1673−4785(2024)05−1072−10

中文引用格式: 丁贵广, 陈辉, 王澳, 等. 视觉深度学习模型压缩加速综述 [J]. 智能系统学报, 2024, 19(5): 1072–1081.
英文引用格式: DING Guiguang, CHEN Hui, WANG Ao, et al. Review of model compression and acceleration for visual deep learning[J]. CAAI transactions on intelligent systems, 2024, 19(5): 1072–1081.

Review of model compression and acceleration for visual deep learning

DING Guiguang[1,2], CHEN Hui[2], WANG Ao[1,2,3], YANG Fan[1,2,3],
XIONG Yizhe[1,2,3], LIANG Yiwen[1,2,3]
(1. School of Software, Tsinghua University, Beijing 100084, China; 2. Beijing National Research Center for Information Science and Technology, Tsinghua University, Beijing 100084, China; 3. Zhuoxi Institute of Brain and Intelligence, Hangzhou 311121, China)

Abstract: Deep learning models have increasingly grown in scale in recent years. Large-scale visual deep learning models are difficult to efficiently infer and deploy in resource-constrained environments, such as embedded devices. Model compression and acceleration can effectively solve this challenge. Although reviews of related works are available, they generally focus on the compressing and acceleration of convolutional neural networks and lack the organization and comparative analysis of the compression and acceleration methods for visual Transformer models. This study focuses on visual deep learning model compression technology and summarizes and analyzes the relevant technical means for convolutional neural networks and visual Transformer models. Technical hotspots and challenges are also summarized and explored. This study provides researchers with a comprehensive understanding of model compression and acceleration fields, which promotes the development of compression and acceleration techniques for deep learning models.

Keywords: visual deep learning; model compression; lightweight structure; model pruning; model quantization; model distillation; Transformer; token pruning

基于神经网络的深度学习已成为人工智能领域的主要方法论。其崛起往往被归功于在 2012 年的 ImageNet[1] 比赛上获得冠军的 AlexNet[2] 深度卷积神经网络模型, 之后, 研究者竞相创建更深层次的卷积神经网络 (convolution neural network, CNN)。随着 VGGNet[3]、Inception[4]、ResNet[5] 等在 ImageNet 比赛中连续打破各项计算机视觉

收稿日期: 2023−11−10.　**网络出版日期:** 2024−08−28.
基金项目: 国家自然科学基金项目 (61925107, 62271281); 浙江省自然科学基金项目 (LDT23F01013F01).
通信作者: 陈辉. E-mail: jichenhui2012@gmail.com.

任务的性能记录, 深度学习神经网络参数规模和模型复杂度急剧增加。特别是近年来, 基于 Transformer 架构的视觉深度学习模型 ViT(vision Transformer)[6]、DeiT(data-efficient image Transformers)[7] 和 Swin Transformer[8] 在计算机视觉的多个主要任务上取得了显著的性能, 视觉深度学习模型的规模已经迈进了 10 亿级别大小, 如 ViT-22B[9] 。

尽管日益增长的模型参数给视觉任务带来极大的性能提升, 但也给计算硬件的发展带来了巨大的挑战。在嵌入式设备等资源受限的环境中, 计算硬件算力有限, 内存不足, 往往难以支持大规模视觉模型的高效推理和部署。因此, 如何对视觉深度模型进行压缩加速, 成为学术界和工业界的研究热点, 催生了一系列有效通用的模型压缩加速技术。这些技术极大地优化了深度模型的性能和效率, 从而促进了深度学习在多个实际领域的广泛应用。

视觉深度学习模型压缩加速技术的相关工作不断推陈出新, 在研发方向和技术方案上呈现出新的变化趋势, 特别是对于视觉 Transformer 的压缩加速方法, 相关的综述文章还未深入探讨。因此, 本文以视觉深度学习模型压缩技术作为核心, 对当前领域中两大最重要的视觉深度模型, 即卷积神经网络和视觉 Transformer 模型, 进行相关技术手段的整理, 并对技术热点和技术挑战进行总结和分析。本文旨在为研究者提供一个全面的模型压缩与加速领域的了解, 以便抓住当前的热门研究方向, 促进未来针对深度学习模型的压缩与加速技术的发展, 从而更好地推动这些模型在实际应用中的应用和发展。本文贡献如下:

1) 本文首先梳理了传统的针对卷积神经网络的模型压缩加速方法, 从轻量化结构、模型剪枝、模型量化和模型蒸馏 4 个研究方向进行概述。

2) 进一步, 本文针对热门视觉 Transformer 深度模型, 全面综述了通道剪枝和序列剪枝 2 类核心关键压缩加速技术。

3) 最后, 本文对视觉深度学习模型压缩技术的未来挑战和方向进行探讨, 希望对模型压缩领域产生积极影响。

和相关综述文献[10-12] 不同的是, 本文全面梳理了针对视觉 Transformer 模型的压缩加速技术, 并探讨了视觉模型压缩加速技术的未来发展方向, 包括在下游复杂视觉任务上的迁移以及针对视觉大模型的有效压缩等。

1　卷积神经网络压缩加速

卷积神经网络的突破带动了十年来人工智能浪潮, 图像识别、人脸识别等核心应用性能达到了人类水平, 带动了万亿规模产业应用。当前工业场景中, 卷积神经网络的应用仍然占据主导, 其中, 对其进行压缩加速是加快落地应用的必要环节, 对降低应用成本具有重要意义。本小节将主要介绍目前主流的卷积神经网络压缩加速方法, 分别从轻量化结构、模型剪枝、模型量化和模型蒸馏 4 个方面进行概述。

1.1　轻量化结构

传统的卷积神经网络结构设计追求精度, 忽视了模型结构的复杂度带来的效率问题。近年来, 随着深度卷积神经在智能手机等端侧设备上的大规模应用, 设计精度高且复杂度低的轻量化结构成为了学术界和工业界关注的焦点。

在卷积核层面, Forrest 等[13] 提出了 SqueezeNet, 采用 1×1 点卷积替代 3×3 卷积, 再结合压缩层和扩展层, 在极大降低参数量的前提下, 保持了与 AlexNet 相似的性能。MobileNet[14] 采用深度可分离卷积取代传统的卷积方法, 显著地减少了模型的参数量和计算复杂度。MobileNet V2[15] 采用了倒残差结构以降低高维特征通道的计算负担, 提升了内存使用效率。MobileNet V3[16] 则融合了神经网络架构搜索技术, 同时引入 SENet[17] 中的轻量级注意力模块, 相比 MobileNet V2 在速度上有较大提升。ShuffleNet V1[18] 在分组卷积之上引入了通道重排技术, 以加强各分组的信息交互并提高卷积核捕捉特征的效果。ShuffleNet V2[19] 进一步推出了通道分离策略, 实现了更加高效的特征通道利用和信息传递。在卷积层层面, Li 等[20] 提出合并连续的非张量层和张量层, 以及合并非张量分支和张量单元, 生成新的张量层来模拟原始多个层的功能。Prabhu 等[21] 采用稀疏但高度连通的扩展图模拟 CNN 中神经元之间连接的方法, 解决了轻量化网络设计中复杂性和效率权衡的问题。Jeon 等[22] 将卷积解构为位移操作和逐点卷积, 并提出了一种称为主动位移层的新的位移操作, 可优化位移量并利用共享的位移值来减少参数。在网络结构层面, 嵌套稀疏网络 NestedNet[23] 通过在不同层级的内部网络中学习不同形式的知识, 在多任务场景中具有较高的灵活性和效率。Jin 等[24] 提出权重采样网络, 从一个可学习的密集参数集中采样以学习模型参数。通过结合权值量化技术, 权重采样网络能够显著降低模型的大小

和计算量，从而实现模型的高效性和轻量化。

1.2　模型剪枝

当前模型剪枝方法可以分为非结构化剪枝和结构化剪枝 2 类方法。

非结构化剪枝通常根据参数的重要性进行选择，将较小的权重设置为零或删除，从而减小模型的尺寸和计算量。LeCun 等[25] 计算参数的二阶导数和海森矩阵来确定参数的重要性。Srinivas 等[26] 去除全连接层中的密集连接来降低计算复杂性，而不依赖于训练数据。Han 等[27] 提出了一种基于神经元连接权重范数的不重要连接删除方法，然后通过再训练来恢复性能。Guo 等[28] 开发了一种 DNS (dynamic network surgery) 方法来恢复被错误删除的重要连接。Molchanov 等[29] 使用变分丢弃技术对模型进行压缩，稀疏卷积层和全连接层。

结构化剪枝通过删除整个通道或层来减小模型的大小，可以保持模型的整体结构。相对于非结构化剪枝，结构化剪枝可以达到更高的模型压缩率。结构化剪枝方法涉及到针对特定结构的剪枝操作，常见的有组级剪枝[30] 和滤波器/通道级剪枝[31-34]。组级剪枝方法[30] 将同样的稀疏模式应用于每一层的滤波器，即对每个立方体去除相同的比特，从而生成具有相同结构的稀疏矩阵的小块。滤波器/通道级剪枝方法根据滤波器的 L1 范数、能耗、激活值等评估标准来决定哪些滤波器或通道应该被剪枝[31]。除了上述方法，还有其他的网络剪枝方法。比如，基于特征选择器来确定特征的重要性分数[32]，或者引入额外的识别感知损失来帮助选择有助于识别的通道[33]。此外，研究人员还开发了一些动态迭代剪枝方法，对每个滤波器进行全局评估，并纠正之前错误剪枝的滤波器[34]。

1.3　模型量化

模型量化技术是指用较低位宽的低数值精度的整型数来表示常用的 32 位高精度浮点网络参数，从而实现模型整体大小的压缩。此外，在现代 CPU 架构上，相比于浮点数值，整型数值的计算也具有较大的效率优势。因此，模型量化也能够实现模型的压缩加速。

在具体的量化方法方面，可以将其分为几个主要类别。首先是二值化，在二值化中，可以将权重二值化或同时二值化权重和激活值。例如，Binaryconnect[35]、Bi-real Net[36] 等都是常见的二值化方法。二值化的优点是可以大幅减小存储空间和内存占用，并且加快运算速度，但也带来了训练难度和推理精度下降的问题。其次是三值化，三值化在二值化的基础上引入一个额外的阈值 0，以减少量化误差。TTQ (trained ternary quantization)[37] 和 VNQ (variational network quantization)[38] 等是常见的三值化方法。此外，混合位宽也是一种常见的量化方法。在混合位宽中，可以根据经验手动选择最优的网络参数位宽组合，例如 DoReFa-Net[39] 和 WRPN(wide reduced-precision networks)[40] 等。最后，为了训练量化网络，需要采用特殊的训练技巧，因为量化网络的参数值是离散的，无法直接使用传统的梯度下降方法。常见的训练技巧包括 INQ(incremental network quantization)[41] 和 ADMM(alternating direction method of multipliers)[42] 等。

1.4　模型蒸馏

模型蒸馏的核心思想是将较大模型的知识迁移到较小的深度模型上，以此实现模型复杂度的降低，提升模型推理速度。在深度学习时代，Hinton 等[43] 首次正式定义了模型蒸馏的框架，并将其应用至深度学习模型驱动的图像分类任务中，显示出良好的效果。Romero 等[44] 认为，加深学生模型的层数可以让其学习的效果更好，并据此提出 FitNets 模型，利用教师模型的中间输出作为额外的监督信息训练学生模型的浅层部分。Chen 等[45] 认为，在训练过程中让学生模型逐渐作深度与宽度上的扩展，可以更好地训练更复杂的学生模型。这些工作从模型蒸馏框架出发，研究了在这一框架下的细粒度深度模型设计方法。

在模型结构设计层面，Lan 等[46] 提出，多分支结构深度学习模型天然地构建了一个多学生框架，这个多学生框架不但训练上极为简便，而且可以将不同学生分支组合，形成性能更好的教师网络。与此同时，图卷积网络也被广泛地用于模型蒸馏框架当中[47]，证明了图卷积网络中的领域专业知识经过模型蒸馏后可以使得学生模型保持相当的专业能力。最新的模型蒸馏方法聚焦学习框架上的创新。You 等[48] 通过委员会投票策略从不同教师网络中遴选出不同的中间层特征监督学生模型，从而使学生模型接触教师模型对相同任务的不同视角。Wang 等[49] 在教师模型与学生模型之间插入了一个生成对抗网络作为助理模型，通过判断其输入为教师模型输出或学生模型输出，促使学生模型与教师模型具有同样的特征分布。Mirzadeh 等[50] 在教师模型与学生模型间插入中等规模的深度网络，采用分布蒸馏的方式蒸馏学生模型。

1.5　卷积神经网络压缩加速算法讨论

表 1 列举了针对卷积神经网络压缩加速各类

方法的优缺点比较。此外, 为了展示各个方法的有效性, 本文选取了各类代表性工作, 分别是轻量化模型结构 MobileNetV2[15]、模型剪枝算法 SFP (soft filter pruning)[51]、模型量化算法 Min-Max[52] 和模型蒸馏方法 DKD(decoupled knowledge distillation)[53], 从计算量减少、参数量减少和加速比 3 个层面进行对比分析。如表 2 所示, 轻量化结构、模型剪枝及模型蒸馏均可有效降低计算量和参数量, 而模型量化只对参数存储大小进行优化, 不会降低模型的计算量和参数量。在加速效果上, 模型量化的加速效果最为明显, 这得益于低数值精度在运算上的巨大效率优势, 但其依赖硬件支持。此外, 轻量化结构通过设计更高效的模型结构, 也能体现出明显的加速比, 但其需要针对特定任务重新设计模型。模型剪枝和模型蒸馏虽然加速比相对有限, 但其应用范围广泛, 适用于已有的任意模型。

表 1　卷积神经网络压缩加速方法特点分析

Table 1　Analysis of different compression schemes for CNNs

压缩加速技术	优缺点
轻量化结构	效果优异, 但需针对特定任务和硬件进行设计
模型剪枝	鲁棒性好, 但依赖人工先验和手动设计
模型量化	效率提升大, 但加速效果依赖硬件支持
模型蒸馏	性能提升大, 但对任务和模型结构敏感

表 2　卷积神经网络压缩加速代表性工作

Table 2　Representative compression acceleration method for CNNs

代表性工作	计算量/%	参数量/%	加速比
MobileNetV2	47.8	80.1	1.57×
SFP	59.7	74.6	1.29×
Min-max	100.0	100.0	1.71×
DKD	61.5	58.7	1.36×

2　视觉 Transformer 模型压缩加速

近年来, 基于 Transformer 架构的视觉深度学习模型 ViT[6] 及其变体 DeiT[7], Swin Transformer[8] 在计算机视觉的多个主要任务上取得了显著的性能与传统的卷积神经网络结构不同, ViT 模型将输入图像划分为多个大小相等的独立的块, 在块中嵌入位置编码信息后转换成块序列, 通过多头注意力机制捕捉图像块之间的视觉特征关系。多头注意力机制的计算量与序列长度即图像大小成二次关系, 过长的图像序列会导致 ViT 模型有很高的计算负载。因此, 计算量及内存占用极大地限制了 ViT 模型在实际场景下的视觉任务上的应用, 特别在资源更加受限的边缘设备上, ViT 模型的部署难度更大。因此, 对 ViT 模型进行模型压缩和加速对 ViT 模型的广泛应用具有重大的实际应用价值。

尽管之前已有各种各样的方法对卷积神经网络结构进行压缩和加速, 但这些方法由于结构上的差异不能很好地适用于 ViT 模型。在自然语言处理任务上, 也有不少研究设计 Transformer 的压缩和加速方法。然而, 由于视觉和文本的模态差异, 这些方法应用于 ViT 模型时效果受限。因此, 针对 ViT 模型的压缩加速具有独特的技术特点, 研究者在这一前沿领域开发了不同的压缩算法, 包括通道剪枝和序列剪枝。

2.1　通道剪枝

通道剪枝与针对传统的卷积神经网络模型的结构化剪枝类似, 是在视觉 Transformer 的特征维度上进行剪枝, 通过人工设计的评价指标或自动化搜索得到不重要的特征维度并加以移除, 从而减小视觉 Transformer 的模型容量和加快其推理速度。

受卷积神经网络结构中滤波器剪枝的启发, VTP(vision Transformer pruning)[54] 提出在注意块中和前馈块中加入可学的缩放参数, 通过对缩放参数增加稀疏性约束进行端到端训练, 并在训练后对缩放参数大小进行排序, 移除掉小的缩放参数对应的维度, 接着再进行微调以恢复被剪枝模型的性能。CP-ViT(cascade vision Transformer pruning)[55] 进一步提出逐步动态预测视觉 Transformer 模型中的稀疏性来减小冗余结构, 通过借助注意力权重定义累积分数, 并作为依据保留信息量大的通道及头注意力。除对单独的特征维度进行剪枝以外, 视觉 Transformer 的深度通常也存在冗余。WDPruning (width & depth pruning)[56] 基于此提出在视觉 Transformer 的中间层引入额外的分类头, 并通过端到端训练使模型能依赖浅层分类头进行分类。在推理阶段, WDPruning 丢弃浅层分类头之后的所有块, 从而大大减少视觉 Transformer 的深度, 降低参数量及加快推理速度。在对特征维度的剪枝上, WDPruning 通过对参数增加可学习的掩码参数, 并训练自适应的阈值参数来减小特征维度大小。NViT(novel ViT)[57] 从硬件的延迟感知出发, 使用全局的结构化通道剪枝方法去除掉对硬件延迟友好且不重要的通道。在对通道的重要性判别上, NViT 使用损失函数对参数变动的敏感程度来评价参数对视觉特征的影响程

度，通过损失函数对参数的一阶泰勒展开近似来作为敏感程度的度量标准。UVC(unified visual Transformer compression)[58] 进一步在通道剪枝的基础上加上跳跃视觉 Transformer 块配置，除对特征维度进行剪枝以外，UVC 对每个块加上是否可跳过的学习参数，并在端到端训练中进行优化，在推理时则通过是否学习参数决定是否可跳过当前块。考虑到视觉 Transformer 中不同组件之间存在丰富的交互，包括多头注意力模块、前馈模块和嵌入层等，对特征维度进行单独度量并剪枝会忽略其与其他特征维度之间的交互作用。从这个角度出发，SAViT (structure-aware vision Transformer)[59] 提出通过探索不同特征维度之间的相互依赖进行更加合理的剪枝。SAViT 使用损失函数对参数的二阶泰勒近似作为组件之间交互作用强弱的度量信息，并基于此选择交互作用弱的组件加以移除，从而可以保证对模型整体的相互依赖没有影响。GOHSP (graph and optimization-based heterogeneous structured pruning)[60] 从马尔可夫链图的角度出发，通过构建头注意力之间的相互关系图对头的重要性进行度量，并直接移除相互关系弱的头注意力。在此基础上，GOHSP 使用基于软剪枝的掩码参数对特征通道重要性进行度量，在训练中动态修改特征维度。MDMC (multi-dimensional model compression)[61] 从输出特征对前馈层、注意力头及图像序列的依赖关系出发，统一度量不同组件的重要性，实现前馈层、注意力模块及图像序列块的统一修剪。

卷积神经网络结构相比视觉 Transformer 结构计算效率更高，内存占用更小，且硬件设备通常对卷积算子进行特殊优化使得卷积神经结构的运行效率进一步提升。而视觉 Transformer 中的多头注意力机制在一定条件下则可以等价于卷积算子的推理过程，SPViT(single-path vision Transformer)[62] 基于此在注意力模块中同时加入等价的卷积算子，在训练过程中，通过注意力模块调整卷积算子的运算结果，并在训练结束后将注意力模块近似转换为卷积算子，从而加速视觉 Transformer 中注意力机制的运算过程。SViTE (sparse vision Transformer exploration)[63] 从稀疏性训练的角度出发，在视觉 Transformer 训练的过程中动态提取稀疏子网络进行小参数量的运算，从而可以大大降低剪枝过程中训练的代价，并在推理时转换为参数量更小的稀疏子网络。考虑到视觉 Transformer 对视觉图像中的低频信号更加敏感，VTC-LFC (vision Transformer compression with low-frequency components)[64] 基于频域，通过对视觉图像进行低通滤波保留信息量更大的低频信号，并通过一阶泰勒展开近似参数对低频信号的敏感程度，选择对视觉低频信号交互作用不明显的特征维度并加以移除。

2.2　序列剪枝

与卷积神经网络结构中对图像进行结构化的降采样不同，视觉 Transformer 直接对划分得到的图像块序列进行处理。而视觉图像中通常存在很多冗余，如除图像主体外的背景。视觉 Transformer 对视觉特征的捕捉通常只依赖少量有信息量的图像位置，大量含有信息量少的图像位置则对模型的最终性能影响不大，因此这些图像块则可以在视觉 Transformer 推理的过程中去除掉[65]。

考虑到多头注意力机制计算复杂度与图像块序列长度成二次关系，移除图像序列则可以大大降低视觉 Transformer 的计算量。从这个角度出发，DynamicViT[65] 通过基于全连接的预测模块评估图像序列块的重要性，在训练时通过掩码注意力机制实现预测模块的端到端训练。在推理时，通过预测模块可以选出不重要的图像序列块，如背景位置等，移除这些信息量不大的图像块则可以加速视觉 Transformer 的处理速度。考虑到直接丢掉不重要的图像序列块会导致信息损失，SP-ViT(soft pruning vision Transformer)[66] 提出通过聚合模块将这些序列块合并到一个图像块中参与后面的推理过程，从而可以有效利用图像块中的信息。SP-ViT 进一步使用多头注意力机制的预测模块，根据多头的注意力权重实现更加细粒度的图像序列块评估能力。

考虑到动态移除冗余的图像块会破坏视觉图像的空间结构的完整性，这会导致剪枝方法不能很好地适用于具有层次化结构的视觉 Transformer 中，Evo-ViT (self-motivated slow-fast token evolution approach for vision Transformers)[67] 提出保持图像序列的完整性，通过选用不同的计算路径来分别更新重要图像块和不重要图像块，即对不重要图像块使用计算量更小的快速更新，对重要图像块使用计算量更大的正常更新。在选择重要图像块的过程中，与 DynamicViT 和 SP-ViT 会引入额外的评估模块不同，Evo-ViT 使用视觉 Transformer 固有的注意力权重来选择与分类聚合块相关性弱的图像块作为信息量小的图像块，避免了引入额外的计算量。通过选择不重要的图像块并移除可以有效降低视觉 Transformer 处理的图像序列长度，并且保留的图像块通常跟分类聚合图像块相

关性更高, 但是这些图像块之间信息较相似, 图像块本身包含的其他视觉特征信息在某种程度上是被忽略的。从这个角度出发, 考虑到视觉 Transformer 中多头注意力模块对图像中的低频信号更加敏感, 拥有更多低频信息成分的图像块可以由模型捕获更多信息, 因此 VTC-LFC[64] 提出使用图像块中的低频信号能量来量化其低频信息占比, 并使其作为图像块重要性程度的评价指标来移除掉部分图像块。SiT (self-slimmed vision Transformer)[68] 同时提出通过聚合模块将图像序列块合并为更少的图像块, 并通过图像块特征级别的蒸馏校准剪枝后的模型与原模型的视觉特征输出。SiT 使用类似自注意力机制的轻量网络对图像块进行聚合, 进一步降低由剪枝带来的额外延迟。

考虑到直接丢弃冗余的图像块可能会损失图像中的有效信息, ToMe (token merging)[69] 提出在视觉 Transformer 的每个块中合并相似的图像块而不是丢弃图像块。ToMe 使用注意力机制中的键来计算图像块之间的相似性, 并采用二分匹配的算法来确定哪些图像块应该合并。这样通过合并图像块的方式相比直接丢弃图像块则可以保留更多的图像信息。与此类似, TPS (token pruning and squeezing)[70] 提出图像块先剪枝后合并的模块, 以提高压缩视觉 Transformer 的效率。TPS 首先通过图像块剪枝得到保留和剪枝的图像块集合, 然后通过单向最近邻匹配和基于相似性的融合步骤, 将剪枝图像块的信息压缩到部分保留图像块中。保留过程通过基于相似性的融合进行实现。尽管 ToMe 与 TPS 通过图像块合并取得了很好的效果, 但这 2 类方法在使用中均依赖人工先验确定每层 Transformer 块的剪枝图像块数和保留图像块数, 因此自适应性差。为解决这个问题, DiffRate (differentiable compression rate)[71] 进一步提出了可微压缩率方法, 其结合了图像块剪枝和图像块合并 2 种策略来实现更有效的压缩。具体来说, DiffRate 通过引入一个可微分的离散代理模块来实现压缩率的可微分化。该模块首先对图像块进行排序, 然后使用重参数化技巧来确定最优的压缩率。DiffRate 则可以实现在不同计算成本约束下, 通过梯度优化来学习压缩率, 以获得自适应的压缩参数。

2.3　视觉 Transformer 压缩加速算法讨论

表 3 列举了针对视觉 Transformer 压缩加速各类方法的优缺点比较。表 4 列举了各类方法的代表性工作, 包括通道剪枝算法 SPViT[62] 和序列剪枝算法 ToMe[69]。从这 2 个表中可以看出, 序列剪枝技术可明显降低模型的计算量, 但不会影响模型的参数量大小。而通道剪枝技术则可以同时降低模型的计算量和参数量。在加速效果上, 序列剪枝的加速比更为明显。这是由于视觉 Transformer 推理复杂度与图像块数量成二次比例关系, 序列剪枝通过有效降低图像块数量, 则可明显提升视觉 Transformer 的推理效率。但是序列剪枝也面临着剪枝后模型图像空间性不完整的问题, 这会很大程度影响模型的迁移性。与之对比, 通道剪枝则只改变模型计算量和参数量, 适用性广。

表 3　视觉 Transformer 压缩加速方案特点分析

Table 3　Analysis of different compression schemes for ViTs

压缩加速技术	优缺点
通道剪枝	适用性广, 鲁棒性好, 但需要大量时间微调
序列剪枝	加速效果明显, 但可能会破坏图像完整性, 导致迁移性差

表 4　视觉 Transformer 压缩加速代表性工作

Table 4　Representative compression acceleration method for ViTs

代表性工作	计算量/%	参数量/%	加速比
ToMe	58.7	100.0	1.5×
SPViT	76.9	85.9	1.1×

3　未来方向

尽管深度学习模型的压缩加速已经取得了显著的技术成果, 但在实际部署应用仍然面临不少挑战。

首先, 当前压缩加速技术的验证方式仍然采用预训练任务进行, 然后, 作为主干模型在下游任务上进行使用。这种方式无法保证压缩后的模型在下游任务上保持较好的性能, 特别是在一些少样本、有噪学习等弱监督场景中, 现有压缩算法可能难以很好泛化。

其次, 现有压缩技术主要集中在图像分类模型的压缩加速上, 对于对象检测、语义识别等更复杂视觉模型的压缩加速的研究相对比较少。如何在下游更复杂视觉任务上开展压缩加速的研究, 对于视觉模型的大规模落地应用至关重要。

最后, 大模型的出现振奋了所有人, 但将大模型适用在各种不同场景和领域, 仍然需要模型压缩加速技术的进步。如何在压缩大模型计算复杂度的同时尽可能保持大模型与众不同的智能能

力，是大模型压缩加速领域需要思考和解决的关键问题。

4　结束语

本文首先介绍了视觉深度学习模型压缩加速技术的研究背景和研究意义，接着从卷积神经网络压缩加速和视觉 Transformer 压缩加速 2 个方面进行分类总结，最后，探讨了视觉模型压缩加速技术的未来发展方向。希望本文能够为研究者提供一个全面了解模型压缩和加速领域的视角，为深度学习模型压缩加速技术的发展作出贡献。

参考文献：

[1] DENG Jia, DONG Wei, SOCHER R, et al. ImageNet: a large-scale hierarchical image database[C]//2009 IEEE Conference on Computer Vision and Pattern Recognition. Miami: IEEE, 2009: 248−255.

[2] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. ImageNet classification with deep convolutional neural networks[J]. Communications of the ACM, 2017, 60(6): 84−90.

[3] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[EB/OL]. (2014−09−04) [2023−11−10]. https://arxiv.org/abs/1409.1556.

[4] SZEGEDY C, LIU Wei, JIA Yangqing, et al. Going deeper with convolutions[C]//2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston: IEEE, 2015: 1−9.

[5] HE Kaiming, ZHANG Xiangyu, REN Shaoqing, et al. Deep residual learning for image recognition[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016: 770−778.

[6] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An image is worth 16x16 words: Transformers for image recognition at scale[EB/OL]. (2020−10−22) [2023−11−10]. https://arxiv.org/abs/2010.11929.

[7] TOUVRON H, CORD M, DOUZE M, et al. Training data-efficient image Transformers & distillation through attention[C]//International Conference on Machine Learning. Virtual Event: PMLR, 2021: 10347−10357.

[8] LIU Ze, LIN Yutong, CAO Yue, et al. Swin Transformer: hierarchical vision Transformer using shifted windows[C]//2021 IEEE/CVF International Conference on Computer Vision. Montreal: IEEE, 2021: 9992−10002.

[9] DEHGHANI M, DJOLONGA J, MUSTAFA B, et al. Scaling vision Transformers to 22 billion parameters[C]//International Conference on Machine Learning. Honolulu: PMLR, 2023: 7480−7512.

[10] GOU Jianping, YU Baosheng, MAYBANK S J, et al. Knowledge distillation: a survey[J]. International journal of computer vision, 2021, 129(6): 1789−1819.

[11] CHENG Yu, WANG Duo, ZHOU Pan, et al. A survey of model compression and acceleration for deep neural networks[J]. IEEE signal processing magazine, 2018, 35(1): 126−136.

[12] LIANG Tailin, GLOSSNER J, WANG Lei, et al. Pruning and quantization for deep neural network acceleration: a survey[J]. Neurocomputing, 2021, 461: 370−403.

[13] FORREST N, SONG Han, MATTHEW W, et al. SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and < 0.5 MB model size[EB/OL]. (2016−11−04) [2023−11−10]. https://arxiv.org/abs/1602.07360.

[14] HOWARD A G, ZHU Menglong, CHEN Bo, et al. MobileNets: efficient convolutional neural networks for mobile vision applications[EB/OL]. (2017−04−17) [2023−11−10]. https://arxiv.org/abs/1704.04861.

[15] SANDLER M, HOWARD A, ZHU Menglong, et al. MobileNetV2: inverted residuals and linear bottlenecks[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018: 4510−4520.

[16] HOWARD A, SANDLER M, CHEN Bo, et al. Searching for MobileNetV3[C]//2019 IEEE/CVF International Conference on Computer Vision. Seoul: IEEE, 2019: 1314−1324.

[17] HU Jie, SHEN Li, SUN Gang. Squeeze-and-excitation networks[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018: 7132−7141.

[18] ZHANG Xiangyu, ZHOU Xinyu, LIN Mengxiao, et al. ShuffleNet: an extremely efficient convolutional neural network for mobile devices[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018: 6848−6856.

[19] MA Ningning, ZHANG Xiangyu, ZHENG Haitao, et al. Shufflenet v2: practical guidelines for efficient CNN architecture design[C]//European Conference on Computer Vision. Cham: Springer, 2018: 122−138.

[20] LI Dawei, WANG Xiaolong, KONG Deguang. DeepRebirth: accelerating deep neural network execution on mobile devices[C]//Proceedings of the AAAI Conference on Artificial Intelligence. New Orleans: AAAI, 2018: 2322−2330.

[21] PRABHU A, VARMA G, NAMBOODIRI A. Deep ex-

pander networks: efficient deep networks from graph theory[C]// European Conference on Computer Vision. Cham: Springer International Publishing, 2018: 20−36.

[22] JEON Y, KIM J. Constructing fast network through deconstruction of convolution[C]//Proceedings of the 32nd International Conference on Neural Information Processing Systems. Montreal: ACM, 2018: 5955−5965.

[23] KIM E, AHN C, OH S. NestedNet: learning nested sparse structures in deep neural networks[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018: 8669−8678.

[24] JIN Xiaojie, YANG Yingzhen, XU Ning, et al. WSNet: compact and efficient networks through weight sampling [C]//International Conference on Machine Learning. Stockholm: PMLR, 2018: 2352−2361.

[25] LECUN Y, DENKER J, SOLLA S. Optimal brain damage[C]//Advances in Neural Information Processing Systems. Denver: ACM, 1990: 598−605.

[26] SRINIVAS S, SUBRAMANYA A, BABU R V. Training sparse neural networks[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops. Honolulu: IEEE, 2017: 455−462.

[27] HAN Song, POOL J, TRAN J, et al. Learning both weights and connections for efficient neural networks[C]// Proceedings of the 28th International Conference on Neural Information Processing Systems. Montreal: ACM, 2015: 1135−1143.

[28] GUO Yiwen, YAO Anbang, CHEN Yurong. Dynamic network surgery for efficient DNNs[C]//Proceedings of the 30th International Conference on Neural Information Processing Systems. Barcelona: ACM, 2016: 1387−1395.

[29] MOLCHANOV D, ASHUKHA A, VETROV D. Variational dropout sparsifies deep neural networksC]//Proceedings of the 34th International Conference on Machine Learning. Sydney: JMLR, 2017: 2498−2507.

[30] WEN Wei, WU Chunxia, WANG Yongan, et al. Learning structured sparsity in deep neural networks[C]//Proceedings of the 30th International Conference on Neural Information Processing Systems. Barcelona: ACM, 2016: 2082–2090.

[31] LI Hao, KADAV A, DURDANOVIC I, et al. Pruning filters for efficient ConvNets[EB/OL]. (2016−08−31) [2023−11−10]. https://arxiv.org/abs/1608.08710.

[32] YU Ruichi, LI Ang, CHEN Chunfu, et al. NISP: pruning networks using neuron importance score propagation[C]// 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018: 9194−9203.

[33] ZHUANG Zhuangwei, TAN Mingkui, ZHUANG Bohan, et al. Discrimination-aware channel pruning for deep neural networks[C]//Proceedings of the 32nd International Conference on Neural Information Processing Systems. Montreal: ACM 2018: 875−886.

[34] LIN Shaohui, JI R, LI Yuchao, et al. Accelerating convolutional networks via global & dynamic filter pruning [C]//Proceedings of the 27th International Joint Conference on Artificial Intelligence. Stockholm: AAAI, 2018: 2425−2432.

[35] COURBARIAUX M, BENGIO Y, DAVID J P. Binaryconnect: Training deep neural networks with binary weights during propagations[C]//Proceedings of the 28th International Conference on Neural Information Processing Systems. Montreal: ACM, 2015: 3123−3131.

[36] LIU Zechun, WU Baoyuan, LUO Wenhan, et al. Bi-real net: enhancing the performance of 1-bit CNNs with improved representational capability and advanced training algorithm[C]//European Conference on Computer Vision. Cham: Springer, 2018: 747−763.

[37] ZHU Chi, HAN Song, MAO Huimin, et al. Trained ternary quantization[EB/OL]. (2016−12−04) [2023−11−10]. https://arxiv.org/abs/1612.01064.

[38] ACHTERHOLD J, KOEHLER J, SCHMEINK A, et al. Variational network quantization[C]//International Conference on Learning Representations. Vancouver: ICLR, 2018: 1−18.

[39] ZHOU Shuchang, WU Yuxin, NI Zekun, et al. DoReFa-net: training low bitwidth convolutional neural networks with low bitwidth gradients[EB/OL]. (2016−06−20) [2023−11−10]. https://arxiv.org/abs/1606.06160.

[40] MISHRA A, NURVITADHI E, COOK J et al. WRPN: wide reduced-precision networks[EB/OL]. (2017−09−04) [2023−11−10]. https://arxiv.org/abs/1709.01134.

[41] ZHOU Aojun, YAO Anbang, GUO Yiwen, et al. Incremental network quantization: towards lossless CNNs with low-precision weights[EB/OL]. (2017−02−10) [2023−11−10]. https://arxiv.org/abs/1702.03044.

[42] BOYD S. Distributed optimization and statistical learning via the alternating direction method of multipliers[J]. Foundations and trends in machine learning, 2010, 3(1): 1−122.

[43] HINTON G E, VINYALS O, DEAN J. Distilling the knowledge in a neural network[EB/OL]. (2015−03−09) [2023−11−10]. https://arxiv.org/abs/1503.02531.

[44] ROMERO A, BALLAS N, KAHOU S E, et al. FitNets:

hints for thin deep nets[EB/OL]. (2014−12−19) [2023−11−10]. https://arxiv.org/abs/1412.6550.

[45] CHEN Tianqi, GOODFELLOW I, SHLENS J. Net2Net: accelerating learning via knowledge transfer[EB/OL]. (2015−11−18) [2023−11−10]. https://arxiv.org/abs/1511.05641.

[46] LAN Xu, ZHU Xiatian, GONG Shaogang. Knowledge distillation by on-the-fly native ensemble[C]// Proceedings of the 32nd International Conference on Neural Information Processing Systems. Montreal: ACM, 2018: 7528−7538.

[47] LASSANCE C, BONTONOU M, HACENE G B, et al. Deep geometric knowledge distillation with graphs[C]// 2020 IEEE International Conference on Acoustics, Speech and Signal Processing. Barcelona: IEEE, 2020: 8484−8488.

[48] YOU Shan, XU Chang, XU Chao, et al. Learning from multiple teacher networks[C]//Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Halifax: ACM, 2017: 1285−1294.

[49] WANG Yunhe, XU Chang, XU Chao, et al. Adversarial learning of portable student networks[C]//Proceedings of the AAAI Conference on Artificial Intelligence. New Orleans: AAAI, 2018: 4260−4267.

[50] MIRZADEH S I, FARAJTABAR M, LI Ang, et al. Improved knowledge distillation via teacher assistant[C]// Proceedings of the AAAI Conference on Artificial Intelligence. New Orleans: AAAI, 2020: 5191−5198.

[51] HE Yang, KANG Guoliang, DONG Xuanyi, et al. Soft filter pruning for accelerating deep convolutional neural networks[EB/OL]. (2018−08−21) [2023−11−10]. https://arxiv.org/abs/1808.06866.

[52] JACOB B, KLIGYS S, CHEN Bo, et al. Quantization and training of neural networks for efficient integer-arithmetic-only inference[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018: 2704−2713.

[53] ZHAO Borui, CUI Quan, SONG Renjie, et al. Decoupled knowledge distillation[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans: IEEE, 2022: 11943−11952.

[54] ZHU Mingjian, TANG Yehui, HAN Kai. Vision Transformer pruning[EB/OL]. (2021−04−17) [2023−11−10]. https://arxiv.org/abs/2104.08500.

[55] SONG Zhuoran, XU Yihong, HE Zhezhi, et al. CP-ViT: cascade vision Transformer pruning via progressive sparsity prediction[EB/OL]. (2022−03−09) [2023−11−10]. https://arxiv.org/abs/2203.04570.

[56] YU Fang, HUANG Kun, WANG Meng, et al. Width & depth pruning for vision Transformers[C]//Proceedings of the AAAI Conference on Artificial Intelligence. Vancouver: AAAI, 2022: 3143−3151.

[57] YANG Huanrui, YIN Hongxu, SHEN Maying, et al. Global vision Transformer pruning with hessian-aware saliency[C]//2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver: IEEE, 2023: 18547−18557.

[58] YU Shixing, CHEN Tianlong, SHEN Jiayi, et al. Unified visual Transformer compression[EB/OL]. (2022−03−15) [2023−11−10]. https://arxiv.org/abs/2203.08243.

[59] ZHENG Chuanyang, ZHANG Kai, YANG Zhi, et al. SAViT: structure-aware vision Transformer pruning via collaborative optimization[C]//Proceedings of the 36th International Conference on Neural Information Processing System. New Orleans: ACM, 2022: 9010−9023.

[60] YIN Miao, UZKENT B, SHEN Yilin, et al. GOHSP: a unified framework of graph and optimization-based heterogeneous structured pruning for vision Transformer[C]// Proceedings of the AAAI Conference on Artificial Intelligence. Washington: AAAI, 2023: 10954−10962.

[61] HOU Zejiang, KUNG S Y. Multi-dimensional model compression of vision Transformer[C]//2022 IEEE International Conference on Multimedia and Expo. Taipei: IEEE, 2022: 1−6.

[62] HE Haoyu, LIU Jing, PAN Zizheng, et al. Pruning self-attentions into convolutional layers in single path[EB/OL]. (2021−11−23) [2023−11−10]. https://arxiv.org/abs/2111.11802.

[63] CHEN Tianlong, CHENG Yu, GAN Zhe, et al. Chasing sparsity in vision Transformers: an end-to-end exploration [C]//Proceedings of the 35th International Conference on Neural Information Processing System. Online: ACM, 2021: 19974−19988.

[64] WANG Zhenyu, LUO Haowen, WANG Pichao, et al. VTC-LFC: vision Transformer compression with low-frequency components[C]//Proceedings of the 36th International Conference on Neural Information Processing Syste. New Orleans: ACM, 2022: 13974−13988.

[65] RAO Yongming, ZHAO Wenliang, LIU Benlin, et al. DynamicViT: efficient vision Transformers with dynamic token sparsification[C]//Proceedings of the 35th International Conference on Neural Information Processing System. Online: ACM, 2021: 13937−13949.

[66] KONG Zhenglun, DONG Peiyan, MA Xiaolong, et al. SPViT: enabling faster vision Transformers via Latency-aware soft token pruning[C]//Lecture Notes in Computer Science. Cham: Springer, 2022: 620−640.

[67] XU Yifan, ZHANG Zhijie, ZHANG Mengdan, et al. Evo-ViT: slow-fast token evolution for dynamic vision Transformer[C]//Proceedings of the AAAI Conference on Artificial Intelligence. Vancouver: AAAI, 2022: 2964−2972.

[68] ZONG Zhuofan, LI Kunchang, SONG Guanglu, et al. Self-slimmed vision Transformer[C]//Lecture Notes in Computer Science. Cham: Springer, 2022: 432−448.

[69] BOLYA D, FU Chengyang, DAI Xiaoliang, et al. Token merging: your ViT but faster[EB/OL]. (2022−10−17) [2023−11−10]. https://arxiv.org/abs/2210.09461.

[70] WEI Siyuan, YE Tianzhu, ZHANG Shen, et al. Joint token pruning and squeezing towards more aggressive compression of vision Transformers[C]//2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver: IEEE, 2023: 2092−2101.

[71] CHEN Mengzhao, SHAO Wenqi, XU Peng, et al. DiffRate: differentiable compression rate for efficient vision Transformers[EB/OL]. (2023−05−29) [2023−11−10]. https://arxiv.org/abs/2305.17997.

作者简介:

丁贵广, 教授, 博士, 主要研究方向为多媒体信息处理、计算机视觉感知。主持和参与国家自然科学基金面上项目等国家级项目数十项。曾获国家科技进步奖二等奖、吴文俊人工智能科技进步奖一等奖、中国电子学会技术发明奖一等奖等。发表学术论文近百篇, 引用量超 17000 次。E-mail: dinggg@tsinghua.edu.cn。

陈辉, 助理研究员, 主要研究方向为计算机视觉、多媒体信息处理。主持国家自然科学基金面上项目 1 项、科技部—“新一代人工智能 2030”子课题 1 项。E-mail: jichenhui2012@gmail.com。

王澳, 博士研究生, 主要研究方向为深度学习模型设计和优化。E-mail: wa22@mails.tsinghua.edu.cn。