深度学习模型压缩的挑战与展望

□文 / 杨戈[1]，郭晋阳[2,3*]，柴振华[4]

（1. 北京航空航天大学 沈元学院，北京 100191；2. 北京航空航天大学 复杂关键软件环境全国重点实验室 北京 100191；3. 北京航空航天大学 人工智能研究院，北京 100191；4. 美团公司，北京 100102）

摘要：近年来，深度学习模型在计算机视觉、自然语言处理等领域的表现取得了长足的发展和巨大的进步，但与之俱来的是复杂的网络结构和庞大的参数数量，其高昂的运算和存储成本限制了深度学习在实际应用中的部署。虽然模型轻量化方法有效地减少了模型的体积与运算，但仍面临新的挑战。本文总结了目前主流的五种深度学习模型轻量化方法，深度分析模型轻量化方法在学术界与工业界目前面临的挑战。最后，本文在前沿研究成果的基础上对未来的工作加以展望。

关键词：人工智能；深度学习；模型压缩

中图分类号：TP18　**文献标志码：**A　**文章编号：**2096-5036(2023)03-0098-09

DOI：10.16453/j.2096-5036.2023.03.010

0 引言

近年来，深度学习[1]在人工智能领域取得了许多成就，尤其是深度神经网络（Deep Neural Network，DNN）的优越性能逐渐被重视。例如，GPT 模型（Generative Pre-training Transformer）是自然语言处理领域的里程碑式成果，可以流畅地与用户沟通交流并完成代码编写任务；而 Meta 发布的 SAM[2] 模型（Segment Anything Model）在计算机视觉与模式识别领域取得了革命性的成就，能够在任意给定的分割提示下返回一个有效的分割掩码，在图像分割任务上展现出极强的能力。

然而，高性能的深度神经网络模型在具体落地时还存在许多挑战，例如复杂的隐藏层结构、高达千亿级的参数数量[3]和高昂的计算成本与存储成本，限制了在嵌入式或移动设备上的部署。例如，训练 OPT-175B 模型需要 992 个 80GB A100 GPU，每个 GPU 的利用率达到 147TFLOP/s，而目前最先进的手机芯片 GPU 性能也要小于 3TFLOP/s。因此，对深度学习模型进行压缩和加速变得很有必要。

具体而言，深度学习模型的轻量化是指在不影响模型精度的前提下，针对网络参数和网络结构的冗余性进行删减和优化，从而加快推理速度，降低浮点运算量和内存占用。本文首先对现有的模型轻量化方法进行了分析，随后提出了该领域目前主要面临的一些挑战。最后，本文将对未来的工作进行展望。

基金项目：“双一流”引导专项（ZG216S2320）

1 模型轻量化进展

本章将简要介绍目前主流的深度学习模型轻量化方法，主要分为模型剪枝、模型量化、低秩因子分解、知识蒸馏、紧凑网络设计五类。

1.1 模型剪枝

深度神经网络结构庞大，存在大量冗余参数，这些参数的信息量趋近于零，将它们去除后对模型精度的影响微乎其微，这种情况被称为过参数化[4]。针对这一特性，研究人员提出了模型剪枝方法。本节将剪枝技术分为非结构化剪枝、结构化剪枝、半结构化剪枝三类进行介绍。

非结构化剪枝是站在参数的层级考虑是否需要被修剪的剪枝方法，它通过删除单个参数减少模型的复杂性。例如，最优脑损伤（Optimal Brain Damage，OBD）和最优脑手术（Optimal Brain Surgeon，OBS）算法是比较经典的非结构化剪枝算法，通过计算目标函数的二阶泰勒展开中海森矩阵选择需要修剪的连接。然而，对单个参数进行剪枝会造成参数矩阵非结构化稀疏，产生额外的内存成本，无法有效地加速稀疏操作，使非结构化剪枝无法实现实际加速效果。

相比于非结构化剪枝算法，结构化剪枝算法主要针对的是结构化信息，是一种粗粒度的修剪方式。其优点在于直接删除卷积核或通道[5]可以有效地对模型进行压缩与加速。常见的结构化剪枝算法包括通道剪枝[6]、列剪枝、行剪枝等。例如，Ma 等[34]根据梯度信息有选择地删减非必需的耦合结构，最大限度地保留了大语言模型的各种功能。但是，结构化剪枝算法的缺点是灵活性较差，粗粒度的操作可能对模型精度造成较大的影响，而微调难度较大。

半结构化剪枝是一种同时利用结构化和非结构化剪枝思想的剪枝方法，其目标是实现 N:M 稀疏，即每 M 个参数中保留 N 个参数[7]。该方法通常将权重按大小进行排序，然后选择值最大的 N 个权重保留，剩下的权重被剪掉。相比于非结构化剪枝，半结构化剪枝具有更好的硬件加速性能，同时对于卷积层的剪枝效果较为显著。

1.2 模型量化

模型量化是指将高位精度的浮点数网络参数用低位精度的数字表示，从而减少存储和计算所需的内存和处理器时间。最早的模型量化方法由 Fiesler 等[8]和 Balzer 等[9]提出，随后在深度学习领域得到广泛应用。本文将模型量化分为训练后量化（Post-training Quantization，PTQ）和量化感知训练（Quantization Aware Training，QAT）两类进行介绍。

训练后量化旨在模型训练完成之后对参数进行精度转换，因此不需要预训练和标注的数据，使得量化过程中计算消耗较小。在大多数情况下，训练后量化可以达到 8 比特量化，同时精度几乎不下降。例如，Nagel 等[10]提出了仿射均匀量化、对称均匀量化、2 的幂次方量化等具体量化方法。Krishnamoorthi 等[11]的工作表明非对称的逐通道

量化与原始精度最为接近。

由于训练后量化在低比特时可能存在较大误差，难以通过微调进行弥补，因此需要选择更加精细的量化感知训练方法。量化感知训练会在模型训练期间对量化误差进行建模和优化，在推断时插入模拟量化操作，从而得到带有各种量化参数的模型。量化感知训练通常需要更多的超参数调整和计算成本，但与训练后量化相比，它通常可以进一步缩小与原始精度之间的差距。

1.3 低秩因子分解

低秩因子分解是指通过施加低秩约束的方式稀疏化卷积核矩阵，其好处是可以去除冗余参数，并降低存储成本与计算量，主要分为满秩分解以及奇异值分解。

对于任意给定的权重矩阵$\boldsymbol{A}\in\mathbf{R}^{m\times n}$，其秩$r\leq\min\{m,n\}$，那么$\boldsymbol{A}$的满秩分解可以表示为$\boldsymbol{A}=\boldsymbol{WH}$，其中$\boldsymbol{W}\in\mathbf{R}^{m\times r}$，$\boldsymbol{H}\in\mathbf{R}^{r\times n}$，如果$r$远小于$\min\{m,n\}$则可以显著降低空间复杂度，否则难以取得很好的优化效果。

奇异值分解旨在使用低秩矩阵近似权重。对给定的矩阵$\boldsymbol{A}\in\mathbf{R}^{m\times n}$，奇异值分解即找到正交矩阵$\boldsymbol{U}\in\mathbf{R}^{m\times m}$、$\boldsymbol{V}\in\mathbf{R}^{n\times n}$和非负实对角矩阵$\boldsymbol{\Sigma}\in\mathbf{R}^{m\times n}$，使得$\boldsymbol{A}=\boldsymbol{U\Sigma V}^{\boldsymbol{T}}$成立。$\boldsymbol{U}$和$\boldsymbol{V}$的列称作$\boldsymbol{A}$的左奇异向量和右奇异向量，$\boldsymbol{\Sigma}$的对角线值称为$\boldsymbol{A}$的奇异值[12]。

1.4 知识蒸馏

知识蒸馏是一种基于迁移学习的模型轻量化方法，依据教师模型的输出训练一个结构简单的学生模型，将知识从前者转移至后者。知识蒸馏主要分为离线蒸馏和在线蒸馏，下文将分别进行介绍。

早期的大多数知识蒸馏方法都属于离线蒸馏，离线蒸馏是基于已经完全训练好的教师模型进行的知识迁移，这种模式使得学生模型的训练更加简单可控，也降低了训练成本。例如，Hinton 等[29]提出了利用网络的输出分布作为知识进行蒸馏；Romero 等[30]提出了利用网络中间层的特征作为知识进行蒸馏，从而训练学生模型；Zhang 等[31]利用蒸馏技术对大语言模型进行轻量化。然而，很多离线蒸馏的实验表明，教师模型与学生模型的能力差距始终存在，并且学生模型的能力很大程度上依赖于教师模型，同时也无法根据学习情况对教师模型进行调整。

在线蒸馏与离线蒸馏最显著的区别是教师模型会与学生模型同步更新参数与知识。在线蒸馏有多种学习形式，例如相互学习、共享学习等，它们共同的特点是整个知识蒸馏框架都是端到端可训练的。例如，Jin 等[32]提出在蒸馏过程中将教师模型与学生模型同时更新，以消除教师与学生间的性能鸿沟；Mirzadeh 等[33]提出利用一个助教模型作为中介，传递蒸馏信息。在线蒸馏更适合多任务知识迁移，教师模型与学生模型在训练中实现“教学相长”、优势互补。

1.5 紧凑网络设计

模型剪枝、模型量化、低秩因子分解从宏观角度而言，都是在参数的层级对网络进

行压缩与加速；紧凑网络设计则不同，其与知识蒸馏类似，是从改变网络结构的层面进行模型压缩。近年来，紧凑网络设计研究有 Iandola 等 [13] 提出的 SqueezeNet 网络，使用 1×1 卷积替换 3×3 卷积；谷歌提出了 MobileNet 系列轻量级网络 [14] 将普通卷积操作拆解成一个深度卷积和一个逐点卷积；旷视提出 ShuffleNet[15]，使用逐点组卷积降低了 1×1 卷积的复杂度，并使用通道混洗操作改善跨特征通道的信息流动，极大地降低了计算成本。紧凑网络设计的优点是可以在相近的精度下实现网络的低存储和低计算量，但其缺点是由于网络结构的特殊性难以配合其他模型压缩方法一起使用，且适用范围较窄、泛化性差。

2 目前面临的挑战

虽然模型轻量化方法可以减小模型规模、便于边缘部署，但随着深度学习的发展以及如 ChatGPT 等大模型的涌现，模型轻量化方法依旧面临着以下挑战。

2.1 大规模深度模型难压缩

大规模深度模型指的是参数量巨大、网络结构复杂、深度较深的深度学习模型。随着深度学习技术的快速发展，大规模深度模型在计算机视觉、自然语言处理等领域获得了广泛应用。例如，GPT、SAM、ViT 等预训练模型在自然语言处理和计算机视觉领域都取得了重要的研究进展。近年来，研究人员也在尝试构建更大的模型以实现更高的精度和泛化能力，同时也希望探索实现通用人工智能的路径。但是，目前对于这些拥有数百亿，甚至千亿个参数的大语言模型，尚未出现更为有效的压缩方法。

首先，大规模深度模型轻量化的困难主要在于网络结构的复杂性。当模型参数规模达到十亿级别时，会出现高量级的离群特征，这些离群特征可能存在于多个网络层中，导致在模型剪枝、模型量化时更加难以精准地找出冗余参数和特征。此外，大规模深度模型中特征复杂，使得知识蒸馏方法无法准确地将知识迁移至小模型，导致难以应用。因此，研究人员需要探索更加有效的网络轻量化方法，以实现对大规模深度模型的有效压缩和优化。

其次，大规模深度模型轻量化的调优成本高，主要是因为大模型在压缩后精度下降严重。传统的剪枝方法需要在剪枝期间或剪枝后进行额外训练，计算成本可达到原始训练的十倍。此外，剪枝步骤高度复杂，修剪后需要针对特定的阶段重新训练，并引入额外的超参数进行调整 [16]。如果使用模型剪枝、知识蒸馏等压缩手段，压缩后的网络结构可能会发生多维度的变化，使得研究人员在根据不同测试情况调整参数时调优成本过高。

最后，大规模深度模型中的一些特殊结构使得常用的模型轻量化方法难以奏效。例如，对于 Transformer 模型而言，其结构中的非线性多头注意力层和前馈神经网络层使得许多可以用在卷积神经网络的结构化剪枝策略无用武之地。在设计紧凑网络时，也难以对这些特殊结构实现高效的删减并维持原有精度。

因此，大规模深度模型的模型轻量化依旧面临挑战，需要进一步设计针对该类模型的压缩策略。

2.2 轻量模型生产成本高

深度压缩模型的生产成本主要由训练成本和硬件成本构成。在训练成本方面，深度模型层数多，层间连接结构复杂，且在训练前需要获取大量的高质量可用数据。无论是收集现有数据还是自行制作数据，都需要耗费大量时间和人力资源。另外，由于收集到的数据中可能存在虚假、恶意和无效数据，在投入训练前还需要进行数据清洗。在训练时，由于数据集和模型的规模十分庞大，在轻量化的过程中需要重新训练以达到性能要求。但是，使用不同的压缩手段进行压缩可能会改变原有的结构，导致更难收敛。因此，轻量化模型需要消耗高昂的训练成本，即使具有高性能的计算资源也难以在短时间内完成再训练。

硬件成本方面，高性能计算资源造价昂贵。在轻量化过程中，长时间的再训练需要消耗大量电力等能源成本，如 OpenAI 发布的 GPT-3 在训练时需要 3640pfs-day[3]，若使用 100 张 A100 GPU，则需要 220 天才能训练完成。此外，在轻量化过程中需要进行超参数配置，这将导致训练时间增加最多 10 倍。超长的训练周期和大量运算使得大部分研究人员难以负担。除计算成本外，模型再训练过程中产生的中间信息也导致存储成本大幅上涨，中小型企业难以承担。另外，大规模的参数在训练时可能会超过单个 GPU 的可用内存，增加了大模型训练的难度。此外，深度模型在不同硬件芯片上的兼容性差，移植时需要重新设计硬件架构，费时费力。

2.3 轻量模型可解释性差

模型的可解释性是深度模型一个十分重要的属性。例如，在医疗领域，模型对其诊断的症状作出解释才能增强结果的说服力，盲目相信模型推断可能导致严重的医疗事故；在法治领域，如果任由模型决策会导致案件误审、误判等过失；此外，财经、运输等领域的业务场景都需要对模型最终的输出作出解释，以免造成大规模经济损失。我们需要从人类的视角理解模型的决策，了解模型的预测依据，提高模型的透明度，这样才可以建立人与深度模型间的信任，降低其在实际应用中的风险。

目前，有许多人对模型的可解释性给出了定义，例如 Lipton 定义了两种模型的可解释性，分别是事前可解释性与事后可解释性 [17]。压缩后的深度神经网络损失了一些展示参数间联系的层或卷积核，这会损伤我们对模型内部网络结构的认知以及对模型输出的理解，某种程度上降低了其可解释性；同时，压缩后得到新模型的具体层结构原理我们也不得而知，难以解释其泛化性能。总体而言，深度神经网络结构复杂，在压缩时又产生了一系列黑箱操作，使其更加难以被人类解释。

因此，模型的可解释性研究是轻量化模型领域的一个重要挑战，需要加强对模型网络结构和输出推断的理解。此外，研究人员也可以基于对模型的解释发现重要的影响因素，进而反馈到网络结构设计，对模型进行完善。

2.4 轻量模型隐私与安全问题

深度学习模型在未来应用于日常生活场景中的另一个重要难题是隐私和安全问题。

企业需要考虑如何确保用户输入深度模型的信息不被恶意获取，并保护深度模型不被恶意攻击。在面对恶意输入时，如何正确处理输入数据，以及如何在训练过程中防止数据被恶意篡改，从而避免模型性能降低。例如，在自动驾驶领域，Cao 等 [18] 发现，目前自动驾驶使用的多传感器融合感知技术存在着严重的技术漏洞，恶意攻击者可以通过放置一个刻意设计的障碍物使模型检测不到并最终导致车祸。而在金融领域，攻击者可能会通过模型输出的置信度信息逆向获取企业机密数据 [19]。

轻量模型由于更少的参数量以及更简单的网络结构，更容易受到攻击。因此，如何在模型轻量化过程中考虑模型的隐私与安全问题，以确保深度学习模型在实际应用中的安全性和隐私性，是目前模型轻量化领域急需考虑的问题。

3 未来的工作展望

目前，虽然深度学习模型压缩领域已成为研究热点之一，各项成果也频繁出现，但是压缩技术仍有很大的发展空间，对各类模型的压缩算法仍不完善。为了应对上述挑战，模型轻量化领域需要突破性进展，主要集中在特定算力和延迟约束下的按需轻量化模型。同时，未来的模型压缩研究不应仅限于算法领域，更多新的模型压缩研究成果应该在硬件领域涌现，以实现从实验室到具体任务场景的转变。通过综合分析目前模型压缩领域的研究成果，本文对未来的发展方向进行展望。

3.1 软硬协同设计

随着硬件部署平台的种类不断增长，主流深度模型与硬件平台的不适配性越来越普遍。为了能够充分挖掘移动硬件平台的潜力，软硬件协同设计近些年受到广泛关注。一方面，硬件部署时，可以根据资源的具体特点设计深度神经网络，有针对性地利用有限的计算以及存储资源设计专项化的网络结构与数据结构，甚至对模型中每一层参数的数量、位数等信息进行具体地限制，从而最大化地提升模型性能。从结合硬件的压缩方法入手，结合硬件反馈信息对模型进行轻量化，最终达到根据硬件资源信息设计合理的深度神经网络架构，实现最优性能的目的。另一方面，我们可以根据模型压缩方法设计适配的硬件架构，定制人工智能芯片以实现模型加速。这样使软硬件各司其职、协同配合，发挥最佳的效果。例如，Mudigere 等 [35] 将高性能可扩展软件栈与 ZionEX 平台配对，实现了 40 倍的训练速度增幅；Zhou 等 [36] 提出了一个基于内存的 Transformer 加速器，在软件层面采用基于令牌的数据流，并通过微调硬件架构支持高效通信，成功提高了计算效率。

3.2 多方法混合压缩

随着对深度学习模型大小和推理速度的要求不断提高，目前已有多种模型轻量化方法可供选择。尽管这些方法已经表现出很好的效果，但仍有许多改进的空间。不同的压缩方法针对深度神经网络的不同特点而设计，但其均有各自的局限性。例如，模型量化通过修改模型权重的表达粒度，实现计算加速；模型剪枝通过缩小模型结构，去除模型

冗余，实现模型压缩。因此，合理地进行多方法混合压缩可以打破不同方法间的障碍，实现优势互补，达到更高的压缩率和更好的模型性能，同时将精度损失降低到最小，甚至可以提升原模型的精度。Li 等 [37] 联合剪枝与张量分解方法，同时通过学习模型的稀疏度和低秩性压缩卷积神经网络。Aghli 等 [38] 则是将剪枝和知识蒸馏结合，对网络进行部分修剪、部分知识提取，大幅压缩了模型的大小。

然而，目前的模型压缩领域并没有一个评估各种压缩方法效果的统一准则，因此在具体的任务场景中，需要选择哪些压缩方法进行组合，需要考虑多个因素。未来的研究应该注重探索不同类型任务的属性，针对这些属性选用不同的压缩方法进行组合。同时，未来的研究应该打破不同压缩方法间的壁垒，汲取各种压缩思想的长处，在各种压缩方法中融会贯通，以实现更好的模型压缩和性能提升。

3.3 自动化压缩

随着模型压缩领域的不断发展，深度神经网络在压缩后的表现越来越接近原模型的精度，甚至有时表现更加优秀。然而，设计压缩后的网络结构和确定超参数取值仍然严重依赖专业知识和经验，成本和门槛高。因此，研究人员开始探索使模型自行探索超参数设置和适当的网络结构的方法。目前，自动化机器学习方法和神经网络搜索算法（Neural Architecture Search，NAS）已经出现在模型设计领域中，可以自动评估神经网络结构的性能，并找到最优的网络结构。已有部分研究者将神经网络搜索算法应用于模型压缩中，例如 Yu 等提出的基于软障碍惩罚的混合精度量化模型（BP-NAS）[20]，以及 Gu 等 [21] 将神经网络搜索与知识蒸馏结合以自动搜索最佳的学生模型架构，这些研究已经取得了一定的成果。然而，对于大型网络而言，搜索空间过于庞大，对算力要求高，耗时较长。因此，未来的研究可以将这些技术更好地移植到模型压缩领域，最终实现将算法高度封装至一键自动化压缩，算法自行调整压缩参数，选择压缩方法以达到最大压缩率和最优表现，以此降低模型压缩技术的应用门槛，提高模型压缩的效率和精度。

3.4 针对不同任务的模型轻量化方法

尽管深度学习模型压缩领域已经取得了许多进展，但是目前大部分工作都集中在用于图像分类和识别的卷积神经网络 [22] 上，而常用于机器翻译和语音识别领域的循环神经网络（Recurrent Neural Networks，RNN）[23] 以及用于图像合成和超分辨率的生成对抗网络（Generative Adversarial Networks，GAN）[24] 等其他网络模型的研究相对较少。这些模型与卷积神经网络在结构、推理思想和训练策略等方面存在差异 [25]，因此卷积神经网络的压缩算法难以直接移植到其他模型上。

近年来，一些学者开始针对非卷积神经网络进行压缩研究。例如，Xu 等 [26] 提出了基于交替方向法的循环神经网络多比特量化方法，该方法可以将推理速度提高 3 倍，并将内存占用降低 10 倍。Li 等 [27] 将原始教师生成器的中间表示知识迁移到对应的压缩学生生成器中，并利用神经架构搜索技术寻找更高效的结构，最终完成了一个通用的压缩框架，以降低条件生成对抗网络的推理时间和模型大小。

基于这些研究，未来的发展方向应该是将深度模型压缩技术向其他类型和特点的模

型进行推广，以覆盖更广泛的应用场景[28]。因此，需要针对不同类型的模型开发对应的压缩技术，以实现模型的高效压缩和加速。

4 结束语

深度学习在计算机视觉、自然语言处理等领域成功应用，然而受限于部署设备的算力，无法大规模应用到移动式设备上。针对此问题，本文首先介绍了主流的深度学习模型压缩技术。随后，根据目前深度模型轻量化中学术界与工业界的障壁，总结了当前深度学习模型压缩领域面临的一些挑战。最后，基于这些挑战和对前沿学术成果的归纳分析，对未来的工作加以展望。

参考文献

[1] LECUN Y, BENGIO Y, HINTON G. Deep learning[J]. Nature, 2015, 521(7553): 436-444.
[2] KIRILLOV A, MINTUN E, RAVI N, et al. Segment anything[J]. arXiv preprint arXiv:2304.02643, 2023.
[3] BROWN T, MANN B, RYDER N, et al. Language models are few-shot learners[J]. Advances in Neural Information Processing Systems, 2020, 33: 1877-1901.
[4] CHENG Y, WANG D, ZHOU P, et al. A survey of model compression and acceleration for deep neural networks[J]. arXiv preprint arXiv:1710.09282, 2017.
[5] GUO J, ZHANG W, OUYANG W, et al. Model compression using progressive channel pruning[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2020, 31(3): 1114-1124.
[6] GUO J, OUYANG W, XU D. Channel pruning guided by classification loss and feature importance[C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2020, 34(7): 10885-10892.
[7] NIU W, MA X, LIN S, et al. PatDNN: achieving real-time DNN execution on mobile devices with pattern-based weight pruning[C]//Proceedings of the Twenty-Fifth International Conference on Architectural Support for Programming Languages and Operating Systems, 2020: 907-922.
[8] FIESLER E, CHOUDRY A, CAULFIELD H J. Weight discretization paradigm for optical neural networks[C]// Optical Interconnections and Networks. SPIE, 1990, 1281: 164-173.
[9] BALZER W, TAKAHASHI M, OHTA J, et al. Weight quantization in Boltzmann machines[J]. Neural Networks, 1991, 4(3): 405-409.
[10] NAGEL M, FOURNARAKIS M, AMJAD R A, et al. A white paper on neural network quantization[J]. arXiv preprint arXiv:2106.08295, 2021.
[11] KRISHNAMOORTHI R. Quantizing deep convolutional networks for efficient inference: a whitepaper[J]. arXiv preprint arXiv:1806.08342, 2018.
[12] KALMAN D. A singularly valuable decomposition: the SVD of a matrix[J]. The College Mathematics Journal, 1996, 27(1): 2-23.
[13] IANDOLA F N, HAN S, MOSKEWICZ M W, et al. SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and < 0.5 MB model size[J]. arXiv preprint arXiv:1602.07360, 2016.
[14] HOWARD A G, ZHU M, CHEN B, et al. MobileNets: efficient convolutional neural networks for mobile vision applications[J]. arXiv preprint arXiv:1704.04861, 2017.
[15] ZHANG X, ZHOU X, LIN M, et al. Shufflenet: an extremely efficient convolutional neural network for mobile devices[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. IEEE, 2018: 6848-6856.
[16] YAO Z, MA L, SHEN S, et al. MlPruning: A multilevel structured pruning framework for transformer-based models[J]. arXiv preprint arXiv:2105.14636, 2021.
[17] LIPTON Z C. The mythos of model interpretability: in machine learning, the concept of interpretability is both important and slippery[J]. Queue, 2018, 16(3): 31-57.
[18] CAO Y, WANG N, XIAO C, et al. Invisible for both camera and lidar: security of multi-sensor fusion based perception in autonomous driving under physical-world attacks[C]//2021 IEEE Symposium on Security and Privacy (SP). IEEE, 2021: 176-194.
[19] FREDRIKSON M, JHA S, RISTENPART T. Model inversion attacks that exploit confidence information and basic countermeasures[C]//Proceedings of the 22nd ACM SIGSAC conference on computer and communications security. ACM, 2015: 1322-1333.
[20] YU H, HAN Q, LI J, et al. Search what you want: barrier panelty nas for mixed precision quantization[C]// Computer Vision–ECCV 2020: 16th European Conference, August 23–28, 2020, Glasgow, UK. Springer International Publishing, 2020: 1-16.
[21] GU J, TRESP V. Search for better students to learn distilled knowledge[J]. arXiv preprint arXiv:2001.11612, 2020.
[22] GUO J, OUYANG W, XU D. Multi-dimensional pruning: a unified framework for model compression[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. IEEE, 2020: 1508-1517.

[23] HOPFIELD J J. Neural networks and physical systems with emergent collective computational abilities[J]. Proceedings of the national academy of sciences, 1982, 79(8): 2554-2558.
[24] GOODFELLOW I, POUGET-ABADIE J, MIRZA M, et al. Generative adversarial networks[J]. Communications of the ACM, 2020, 63(11): 139-144.
[25] SHEN Q, QIAO L, GUO J, et al. Unsupervised learning of accurate siamese tracking[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. IEEE, 2022: 8101-8110.
[26] XU C, YAO J, LIN Z, et al. Alternating multi-bit quantization for recurrent neural networks[J]. arXiv preprint arXiv:1802.00150, 2018.
[27] LI M, LIN J, DING Y, et al. GAN compression: efficient architectures for interactive conditional gans[C]// Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. IEEE, 2020: 5284-5294.
[28] HU Z, LU G, GUO J, et al. Coarse-to-fine deep video coding with hyperprior-guided mode prediction[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. IEEE, 2022: 5921-5930.
[29] HINTON G, VINYALS O, DEAN J. Distilling the knowledge in a neural network[J]. arXiv preprint arXiv:1503.02531, 2015.
[30] ROMERO A, BALLAS N, KAHOU S E, et al. FitNets: hints for thin deep nets[J]. arXiv preprint arXiv:1412.6550, 2014.
[31] ZHANG C, YANG Y, WANG Q, et al. AutoDisc: automatic distillation schedule for large language model compression[J]. arXiv preprint arXiv:2205.14570, 2022.
[32] JIN X, PENG B, WU Y, et al. Knowledge distillation via route constrained optimization[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. IEEE, 2019: 1345-1354.
[33] MIRZADEH S I, FARAJTABAR M, LI A, et al. Improved knowledge distillation via teacher assistant[C]// Proceedings of the AAAI conference on artificial intelligence. AAAI, 2020, 34(04): 5191-5198.
[34] MA X Y, FANG G F, WANG X C, et al. LLM-Pruner: on the structural pruning of large language models[J]. arXiv preprint arXiv:2305.11627, 2023.
[35] MUDIGERE D, HAO Y, HUANG J, et al. Software-hardware co-design for fast and scalable training of deep learning recommendation models[C]//Proceedings of the 49th Annual International Symposium on Computer Architecture. ACM, 2022: 993-1011.
[36] ZHOU M, XU W, KANG J, et al. TransPIM: a memory-based acceleration via software-hardware co-design for transformer[C]//2022 IEEE International Symposium on High-Performance Computer Architecture (HPCA). IEEE, 2022: 1071-1085.
[37] LI Y, LIN S, LIU J, et al. Towards compact cnns via collaborative compression[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. IEEE, 2021: 6438-6447.
[38] AGHLI N, RIBEIRO E. Combining weight pruning and knowledge distillation for cnn compression[C]// Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. IEEE, 2021: 3191-3198.

杨 戈

北京航空航天大学沈元学院软件工程专业本科在读。主要研究方向为人工智能、机器学习与模式识别。

郭晋阳

北京航空航天大学人工智能研究院助理教授，博士。主要研究方向为深度学习及其模型轻量化。已在国际权威学术期刊和顶级会议等发表论文 10 余篇。担任 TPAMI、IJCV 等国际顶级期刊审稿人与 CVPR、ICCV 等国际顶级会议程序委员会委员。荣获 ICCV Doctoral Consortium、悉尼大学全额奖学金等奖励。
* 通信作者 email：jinyangguo@buaa.edu.cn

柴振华

美团计算机视觉高级算法专家。博士毕业于中国科学院自动化研究所模式识别国家重点实验室。研究方向包括基础模型结构设计、大规模预训练、模型压缩、数据智能算法，以及通用视觉应用。曾在计算机视觉领域国际刊物上发表学术论文 30 多篇，拥有发明专利 40 余项，在多个国际知名挑战赛上获得冠军和亚军。