软件学报 ISSN 1000-9825, CODEN RUXUEW
[doi: 10.13328/j.cnki.jos.007696] [CSTR: 32375.14.jos.007696]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

安全视角下的模型压缩研究综述*

商 婧[1,2], 王 健[1,2], 王凯崙[1,2], 杨 磊[1,2], 姜 楠[3], 姜 强[1,2], 刘吉强[1,2]

[1](智能交通数据安全与隐私保护北京市重点实验室 (北京交通大学), 北京 100044)

[2](北京交通大学 网络空间安全学院, 北京 100044)

[3](北京工业大学 计算机学院, 北京 100124)

通信作者: 王健, E-mail: wangjian@bjtu.edu.cn; 刘吉强, E-mail: jqliu@bjtu.edu.cn

摘 要: 随着深度学习模型日益广泛地部署于资源受限的场景中, 模型剪枝、量化、知识蒸馏等模型压缩技术已成为提升模型效率与可部署性的关键手段. 然而, 模型压缩在带来性能与效率优势的同时, 也与模型安全和隐私问题产生了复杂关联. 一方面, 压缩过程可能改变模型的表示空间或决策边界, 从而引入新的安全风险或放大已有威胁; 另一方面, 安全与隐私需求也反过来推动模型压缩技术的发展, 使得安全感知的模型压缩逐渐成为一个重要研究方向. 系统梳理近年来安全视角下的模型压缩研究进展, 重点从两个方面进行综述: 一是压缩模型面临的安全与隐私风险, 包括后门攻击、对抗样本、隐私泄露以及前沿应用场景下的新型风险; 二是面向对抗鲁棒性与隐私保护的安全感知模型压缩方法, 总结在压缩过程中引入安全约束与隐私保护机制的代表性研究. 此外, 进一步提炼压缩模型安全与隐私研究面临的关键问题, 给出面向典型部署场景的风险感知选型参考和面向压缩模型安全性的基准化评估维度建议, 并从内在机理分析、真实部署环境下的安全防护等方面对未来研究进行展望.

关键词: 模型压缩; 人工智能安全; 隐私保护; 对抗鲁棒性; 安全感知模型压缩

中图法分类号: TP18

中文引用格式: 商婧, 王健, 王凯崙, 杨磊, 姜楠, 姜强, 刘吉强. 安全视角下的模型压缩研究综述. 软件学报. http://www.jos.org.cn/1000-9825/7696.htm

英文引用格式: Shang J, Wang J, Wang KL, Yang L, Jiang N, Jiang Q, Liu JQ. Survey on Model Compression Research from Security Perspective. Ruan Jian Xue Bao/Journal of Software (in Chinese). http://www.jos.org.cn/1000-9825/7696.htm

Survey on Model Compression Research from Security Perspective

SHANG Jing[1,2], WANG Jian[1,2], WANG Kai-Lun[1,2], YANG Lei[1,2], JIANG Nan[3], JIANG Qiang[1,2], LIU Ji-Qiang[1,2]

[1](Beijing Key Laboratory of Security and Privacy in Intelligent Transportation (Beijing Jiaotong University), Beijing 100044, China)

[2](School of Cyberspace Science and Technology, Beijing Jiaotong University, Beijing 100044, China)

[3](College of Computer Science, Beijing University of Technology, Beijing 100124, China)

Abstract: As deep learning models are increasingly deployed in resource-constrained scenarios, model compression techniques such as pruning, quantization, and knowledge distillation have become key approaches for improving model efficiency and deployability. However, while model compression offers advantages in performance and efficiency, it is also intricately linked to security and privacy issues. On the one hand, compression may alter a model's representation space or decision boundary, thus introducing new security risks or amplifying existing threats. On the other hand, security and privacy requirements have also shaped the development of model compression techniques, making security-aware compression an increasingly important research direction. This study systematically reviews recent progress in model compression research from a security perspective, focusing on two main aspects. First, it examines the security and privacy risks faced by compressed models, including backdoor attacks, adversarial examples, privacy leakage, and emerging risks in frontier application scenarios. Second, it surveys security-aware model compression methods for adversarial robustness and privacy

* 基金项目: 国家重点研发计划 (2023YFB2703700); 中央高校基本科研业务费专项资金 (2025JBZY025); 北京市自然科学基金 (L251062)
收稿时间: 2026-03-22; 修改时间: 2026-04-27; 采用时间: 2026-05-12; jos 在线出版时间: 2026-08-12

protection, summarizing representative studies that incorporate security constraints and privacy-preserving mechanisms into the compression process. In addition, this study identifies key issues in security and privacy research on compressed models, provides risk-aware guidance for typical deployment scenarios, recommends benchmark-based evaluation dimensions for assessing the security of compressed models, and discusses future research directions in terms of underlying mechanism analysis and security protection in real-world deployments.
Key words: model compression; artificial intelligence security; privacy preservation; adversarial robustness; security-aware model compression

1 引 言

随着深度学习在移动终端、嵌入式设备及大规模在线服务中的广泛应用, 模型的推理效率、存储开销与能耗成为影响其实际部署的关键因素. 在此背景下, 模型压缩技术迅速发展, 修剪、量化、知识蒸馏、低秩分解等方法被广泛用于构建轻量、高效、易部署的深度学习模型[1]. 尤其是在大模型快速发展的背景下, 模型压缩技术在边缘部署、隐私计算与资源受限场景中显得愈发重要, 相关研究受到持续关注. 与此同时, 随着模型在安全敏感场景中的应用日益增多, 人工智能 (artificial intelligence, AI) 系统面临的安全与隐私威胁也愈加凸显, 包括后门攻击、对抗样本以及隐私泄露等多种形式. 这些威胁不仅危及系统自身的可靠性, 也可能在实际部署中引发严重的安全后果.

需要指出的是, 模型压缩与安全问题并非彼此孤立, 模型压缩在提升模型效率的同时, 也会改变模型的参数分布、表示空间以及决策边界, 从而可能影响模型在面对对抗攻击、后门攻击或隐私推断时的安全特性. 已有研究表明, 某些攻击在压缩模型上可能表现出更强的攻击效果, 甚至能够绕过针对原始模型设计的防御机制[2,3]. 另一方面, 随着安全与隐私需求的不断提升, 模型压缩技术也逐渐从单纯追求效率优化, 发展为同时考虑模型精度、压缩率以及安全鲁棒性[4]或隐私保护[5]的新型压缩方法. 因此, 模型压缩与安全、隐私之间逐渐形成了相互影响、协同发展的研究关系.

尽管近年来已有大量研究从不同角度探讨模型压缩与安全问题之间的联系, 但目前针对该交叉领域的系统性综述仍然较为有限. 现有综述工作大多集中于分析模型压缩对对抗鲁棒性的影响[6,7], 虽在理解压缩模型的对抗脆弱性方面提供了重要参考, 但其研究范围相对局限, 未能更全面地覆盖后门攻击、隐私泄露等其他安全与隐私威胁; 同时, 近年来出现的安全感知模型压缩方法也尚缺乏系统梳理. 基于上述背景, 本文从安全视角对模型压缩相关研究进行系统综述, 重点从两个方面总结现有工作: (1) 压缩模型面临的安全与隐私风险; (2) 对抗鲁棒与隐私保护的安全模型压缩方法. 在此基础上, 本文进一步归纳压缩模型在典型部署场景下的风险特征与选型参考, 并围绕基准化评估、压缩触发型攻击防护、安全与隐私影响机理以及硬件-算法协同部署防护等问题, 对未来研究方向进行展望.

本文第 2 节介绍模型压缩技术基础. 第 3 节简要介绍深度学习中的主要安全与隐私威胁. 第 4 节总结压缩模型面临的安全与隐私风险. 第 5 节综述对抗鲁棒与隐私保护的安全模型压缩方法. 第 6 节讨论当前挑战并展望未来研究方向. 最后在第 7 节给出全文总结.

2 模型压缩技术

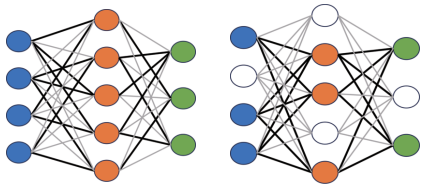
随着深度学习模型规模不断扩大, 其计算与存储成本成为实际部署中的主要瓶颈. 为提升模型在资源受限场景中的效率, 研究者提出了多种模型压缩技术, 从结构层面、参数层面到训练策略层面均有大量研究. 尽管现有方法类型较为丰富, 但在实践中, 模型修剪、模型量化、知识蒸馏和低秩分解是应用最广、最具代表性的 4 类方法, 也构成了多数压缩策略的基础. 因此, 本节将主要围绕这 4 类技术进行介绍, 并在后续讨论其在安全与隐私背景下的特性与研究进展.

2.1 模型修剪

模型修剪 (或剪枝) 旨在移除神经网络中的冗余参数或结构, 在尽量保持模型性能的同时, 显著降低其计算与存储开销, 并加速推理过程. 根据修剪粒度的不同, 模型修剪主要可分为非结构化修剪与结构化修剪两类.

2.1.1 非结构化修剪

非结构化修剪是一类细粒度的模型压缩技术, 其目标是通过移除神经网络中贡献较小或重要性较低的独立权重连接 (即参数) 来降低模型规模. 这使得修剪后的模型在数学上变为稀疏权重矩阵, 但不会改变原始网络架构的拓扑结构, 如图 1(a) 所示. 这种细粒度稀疏化方式具有较高灵活性和通用性, 通常能够获得较好的压缩效果. 非结构化修剪的一类代表性做法是基于权值幅值的修剪, 该方法假设绝对值较小的权重对模型输出影响较小, 可优先被剪除, 从而降低冗余度. 这类策略通常结合训练-剪枝-微调的三阶段流程, 即先训练得到原始稠密模型, 再根据修剪准则置零低重要性权重, 最后对稀疏模型进行微调以恢复性能[8].



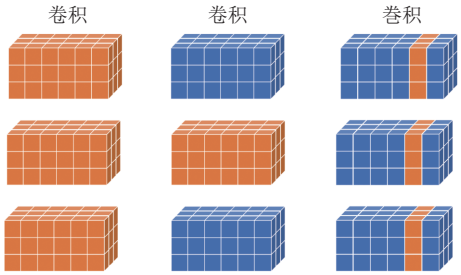
(a) 非结构化修剪 (b) 结构化神经元修剪

图 1 非结构化与结构化模型修剪示意图

除了简单的幅值修剪外, 一些研究采用敏感度分析[9]或梯度信息[10]作为重要性度量, 通过评估权重对损失函数的贡献来指导修剪决策, 这类方法可以更准确地识别对性能影响较小的参数. 此外, 随着高稀疏压缩的需求不断增长, 非结构化修剪也出现了多种改进方向, 例如基于稀疏掩码学习修剪模式[11], 以及从初始化阶段就设计稀疏结构[12]等. 这些方法相较于传统修剪策略, 更加注重与训练过程的协同优化, 从而在获得高稀疏度的同时, 尽量减少性能损失.

2.1.2 结构化修剪

结构化修剪旨在以更粗粒度的方式移除神经网络中的特定结构单元, 如神经元、卷积通道、滤波器甚至整个层等, 如图 1(b) 以及图 2 所示. 与非结构化修剪仅删除单个权重连接不同, 结构化修剪会改变模型的张量形状和结构. 其中, 神经元修剪方法通过分析神经元的功能冗余性或激活特征来决定移除对象.



(a) 层级 (b) 过滤器级 (c) 通道级

图 2 3 种类型的结构化修剪

对于通道及过滤器级的修剪, 早期 Li 等人[13]基于卷积核的 L1 范数评估通道重要性, 移除贡献较小的滤波器, 实现了卷积神经网络 (convolutional neural network, CNN) 的有效加速. 该方法为后续的大量通道级剪枝工作奠定了基础. 随后, Liu 等人[14]提出了使用批量归一化 (batch normalization, BN) 层缩放因子 γ 作为通道重要性指标, 通过在训练过程中加入稀疏正则化, 使部分通道自然趋于失效, 在修剪时直接将其移除. 对于层级修剪, Fan 等人[15]针对 Transformer 提出了基于重要性评分的层级裁剪策略, 通过移除冗余自注意力层显著减少模型计算量, 同时保持语言建模性能. Jordao 等人[16]通过在特征子空间中利用偏最小二乘评估各层的判别性贡献来决定其重要性, 从而实现对整个卷积层的有效裁剪. 层级修剪从网络结构层面直接约简模型深度, 在大模型压缩和实际部署中具有

重要价值.

近年来, 结构化修剪也与网络架构搜索 (neural architecture search, NAS)、知识蒸馏以及稀疏正则化[17]等技术深度结合. 例如部分工作利用 NAS 直接搜索压缩后的高效结构[18], 而另一些方法通过蒸馏指导剪枝后模型恢复精度[19,20]. 相比非结构化修剪方法, 结构化修剪在计算友好性和硬件加速方面具有显著优势, 因此在实际部署场景中使用更为广泛.

2.2 模型量化

模型量化旨在以低比特整数表示替代浮点表示, 从而降低模型的存储开销与计算成本, 是当前应用最广泛的模型压缩技术之一[21]. 其基本思想是将连续空间中的参数或激活映射到有限个离散取值上, 以较小的精度损失换取更高的压缩率与推理效率. 根据量化表示本身的特征以及量化介入训练的方式, 现有方法可从量化表示形式和训练方式两个角度进行概括.

2.2.1 量化表示形式

量化后, 全精度值被转换为位宽较低的新值, 这些新值被称为量化级别, 从表示形式来看, 量化方法可按照比特宽度、量化映射机制以及量化级别分布进行区分[21].

(1) 按比特宽度. 比特宽度决定了参数、激活等模型组件被映射到多少个离散取值, 是影响表示能力与压缩率的核心因素[22]. 二值量化将权重或激活限制在$\{-1,+1\}$[23]; 三值量化在此基础上引入 0 值[24], 将权重约束为$\{-1,0,+1\}$; 而 8 比特整数量化[25]因在压缩率、精度保持和硬件支持之间取得较好平衡, 已成为实际部署中的主流方案.

(2) 按量化映射机制. 量化映射机制指的是将原始浮点数 x 映射到其量化值 q 的具体过程. 典型的映射方式包括确定性量化与随机量化[21]. 确定性量化通过固定规则 (如就近取整) 将连续值映射到最近的量化级别. 图 3 展示了 $[-\infty,+\infty]$ 范围内的连续值被映射到离散值的确定性量化过程, 其量化公式为: $Q(X)=q_i,\ X\in[r_i,r_{i+1}],i=1,2,\ldots,m$. 量化后的值为: $Q=q_1,q_2,\ldots,q_m$. 随机量化机制[26]在量化区间内按概率随机向上或向下取整, 以缓解梯度估计偏差或避免训练陷入局部极小值.

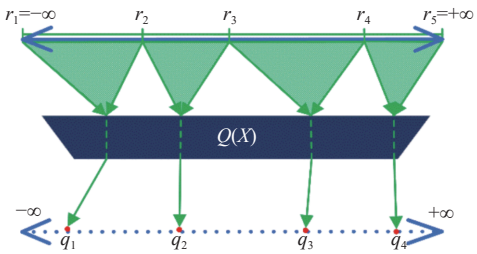


图 3 确定性量化映射示意图[21]

(3) 按量化级别分布. 量化级别的数量由比特宽度决定, 而其在数轴上的分布形式 (均匀或非均匀) 则决定量化值与原始值的逼近方式. 均匀量化[27]假定量化级别等间隔分布, 这是当前整数量化的默认选择. 与之相对, 非均匀量化[28]允许量化级别在数值空间中不等间距分布, 更贴近参数的真实统计特性.

2.2.2 量化训练方式

从训练流程的角度来看, 量化方法可大致分为两类: 训练后量化 (post-training quantization, PTQ) 与量化感知训练 (quantization-aware training, QAT). PTQ 在模型完整训练结束后再进行量化转换, 不需要重新训练或只需极少量的校准数据, 是部署成本最低的一类量化方法. 例如, Jacob 等人[29]系统化地提出了仅使用整数运算进行推理的 INT8 量化方案, 并给出了在工业框架中实现 PTQ 的具体流程, 使得 8 位整数量化成为移动端与服务器端推理的事实标准. 与 PTQ 相比, QAT 在训练阶段显式引入量化算子, 使模型在优化过程中逐步适应离散表示. 这类方法虽然训练开销更高, 但在低比特量化和高精度要求场景下通常具有更好的性能保持能力. 例如, Courbariaux 等人[23]将权重及激活约束为二值形式, 并通过随机或确定性二值化加上直通估计器, 实现了可训练的二值网络.

2.3 知识蒸馏

知识蒸馏是一类通过模型间知识迁移实现模型压缩与性能保持的技术, 其核心思想是利用性能较强、结构较复杂的教师模型来指导轻量的学生模型的训练过程[30]. 知识蒸馏通过引入教师模型输出的“软信息”, 使学生模型能够学习到更丰富的类别关系和决策边界, 从而在显著降低模型规模的同时尽量保持甚至提升预测性能. 知识蒸馏的关键思想在于利用教师模型的 *Softmax* 输出作为额外监督信号[31]. 给定教师模型输出的 logits 向量 $z^{(t)}$ 与学生模型输出的 logits 向量 $z^{(s)}$, 通过引入温度参数 $T>1$ 来平滑类别分布, 得到教师与学生的软概率分布:

$$p_i^{(t)}=\frac{\exp(z_i^{(t)}/T)}{\sum_j \exp(z_j^{(t)}/T)},\ p_i^{(s)}=\frac{\exp(z_i^{(s)}/T)}{\sum_j \exp(z_j^{(s)}/T)} \tag{1}$$

在此基础上, 蒸馏损失 $\mathcal{L}_{\mathrm{KD}}$ 通常定义为教师与学生软输出之间的 Kullback-Leibler (KL) 散度[32]:

$$\mathcal{L}_{\mathrm{KD}}=T^2\cdot KL(p^{(t)}\,\|\,p^{(s)}) \tag{2}$$

其中, 系数 T^2 用于平衡梯度尺度. 实际训练中, 蒸馏损失通常与标准的监督损失, 如交叉熵 (cross-entropy, CE) 损失 $\mathcal{L}_{\mathrm{CE}}$ 联合优化, 完成完整的知识蒸馏过程:

$$\mathcal{L}=(1-\lambda)\mathcal{L}_{\mathrm{CE}}+\lambda\mathcal{L}_{\mathrm{KD}} \tag{3}$$

其中, λ 控制蒸馏监督与真实标签监督之间的权重. 图 4 展示了知识蒸馏的完整过程.

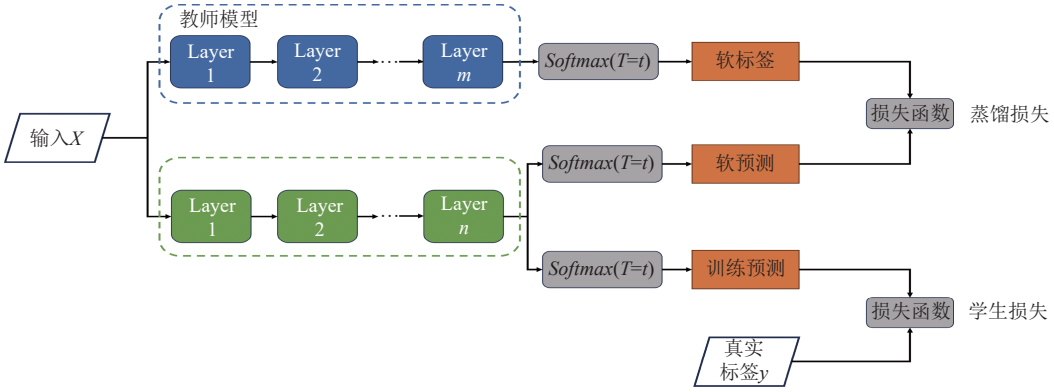


图 4 知识蒸馏过程示意图

2.4 低秩分解

低秩分解是一类通过近似原始参数矩阵的低秩结构来实现模型压缩与计算加速的方法, 其基本假设是深度神经网络 (deep neural network, DNN) 中的权重矩阵或张量在高维空间中往往具有显著的冗余性[33], 可以用秩更低的表示进行有效近似. 与修剪和量化不同, 低秩分解通常不依赖于稀疏化或离散化操作, 而是通过矩阵或张量分解显式重构网络参数, 从而减少模型的存储需求和计算复杂度.

在最基本的形式下, 低秩分解可作用于全连接层或卷积层中的权重矩阵. 给定一个权重矩阵 $W\in\mathbb{R}^{m\times n}$, 其低秩近似可以表示为两个较小矩阵的乘积: $W\approx UV$, 其中 $U\in\mathbb{R}^{m\times r}$, $V\in\mathbb{R}^{r\times n}$, 且 $r\ll\min(m,n)$. 经典的奇异值分解 (singular value decomposition, SVD) 提供了一种最优的低秩近似方式, 即通过保留最大的 r 个奇异值来最小化重构误差. 在深度网络压缩中, 这一思想被用于将原本一次矩阵乘法拆分为两次更小规模的乘法, 从而减少计算量和参数规模. 针对卷积神经网络中的高维卷积核, 研究者进一步将低秩分解推广至张量层面. 例如, Jaderberg 等人[34]提出利用低秩近似加速卷积网络, 通过对卷积核进行分解显著降低了推理成本. 这类方法表明, 卷积层中存在可被系统性利用的低秩结构, 为模型压缩提供了另一条重要路径. 近年来, 随着大规模模型和高维参数空间的广泛应用, 低秩分解思想在参数高效建模与模型压缩场景中得到进一步拓展. 例如, Hu 等人[35]将低秩分解的研究视角拓展到大语言模型 (large language model, LLM), 提出了低秩适应 (low-rank adaptation, LoRA) 方法, 通过将模型参数更新限制在

低秩子空间内, 在显著减少可训练参数数量的同时保持了下游任务性能.

3 深度学习安全与隐私威胁

随着深度学习模型在关键应用场景中的广泛部署, 其安全性与隐私问题逐渐成为研究与实践中不可忽视的重要议题. 深度神经网络通常依赖大规模数据与复杂参数结构, 其训练与推理过程均可能暴露新的攻击面, 使攻击者能够通过数据操纵、模型交互或输出分析等方式破坏模型的可靠性或窃取敏感信息[36]. 因此, 从模型生命周期的角度系统梳理深度学习面临的安全与隐私威胁, 是理解后续压缩模型安全与隐私风险的必要前提. 从攻击发生的阶段来看, 现有的安全与隐私威胁可大致分为训练阶段威胁与推理阶段威胁, 基于这一划分, 本节将对深度学习模型在不同阶段面临的主要威胁类型进行概述性介绍.

3.1 训练阶段威胁

训练阶段是构建深度学习模型的核心环节, 其高度依赖大规模训练数据与复杂的优化过程. 一旦攻击者能够干预训练数据、训练流程或模型参数更新过程, 便可能在模型中植入长期且隐蔽的安全或隐私风险.

3.1.1 训练阶段安全威胁

训练阶段的安全威胁主要指攻击者通过直接或间接操纵模型训练过程, 使模型在部署后表现出非预期甚至恶意的行为. 这类威胁通常发生在攻击者能够影响训练数据、训练管道或模型更新机制的场景中, 其影响往往在模型部署后长期存在, 且难以通过常规性能测试或简单验证手段发现.

数据投毒攻击是训练阶段最早被系统研究的一类安全威胁. 该类攻击通过向训练集中注入恶意构造的样本, 或篡改部分训练数据分布, 诱导模型学习到攻击者期望的错误模式. 早期代表性工作如 Biggio 等人[37]从优化视角系统分析了投毒样本对分类器决策边界的影响, 揭示了即便少量恶意样本也可能显著破坏模型性能. 随后, Steinhardt 等人[38]进一步表明, 在现代深度学习设置下, 投毒攻击可以在不显著降低训练精度的前提下, 使模型在测试阶段出现系统性错误. 这类攻击通常以降低整体性能或引入全局偏置为目标, 其隐蔽性相对有限, 但在安全关键应用中仍具有严重风险.

在数据投毒攻击的基础上, 后门攻击作为一种更具隐蔽性的训练阶段威胁受到广泛关注. 后门攻击同样通过污染训练数据或训练过程实现[39], 但其目标并非破坏模型整体性能, 而是诱导模型在遇到特定触发条件, 即触发器时输出攻击者指定的结果, 而在正常输入下仍保持接近原模型的准确率. Gu 等人[40]提出的 BadNets 是最早展示后门攻击可行性的工作之一, 该研究表明, 即使只注入极少量带有触发器的样本, 也足以在模型中植入稳定且隐蔽的后门行为. 后续研究进一步拓展了后门攻击的形式, 包括更隐蔽的触发器设计[41]、与语义相关的后门[42]等.

3.1.2 训练阶段隐私威胁

训练阶段的隐私威胁主要指模型在训练过程中对敏感训练数据产生过度记忆[43,44], 从而在模型参数或行为中隐式编码了与个体样本相关的信息. 这类威胁并不一定依赖攻击者在训练阶段的主动干预, 而是源于模型容量、训练目标以及数据分布等因素, 使得模型在无意中学习并保留了敏感内容.

早期研究表明, 深度神经网络在训练过程中可能显著记忆训练样本中的细节信息, 尤其是在模型容量较大、正则化不足或训练数据存在稀有样本的情况下. Song 等人[43]系统分析了这一现象, 指出现代深度学习模型可能在参数中存储可被恢复的敏感训练数据. 该工作强调, 即便模型在标准测试集上表现良好, 其内部仍可能包含严重的隐私泄露风险. 进一步的研究发现, 模型对训练数据的记忆并非均匀分布, 而是往往集中于异常样本、低频样本或噪声样本[44]. 这意味着在实际应用中, 包含敏感信息的少量样本更容易被模型“记住”, 从而成为潜在的隐私泄露源. 尤其在自然语言处理 (natural language processing, NLP) 和生成模型等场景[45]中, 模型可能在生成结果中直接复现训练数据, 使得训练阶段的隐私风险具有直接可观测的后果.

3.2 推理阶段威胁

推理阶段威胁发生在模型部署和对外提供服务的过程中, 攻击者无需接触训练数据或训练过程, 仅通过与模

型的交互即可实施攻击. 因攻击门槛相对较低, 这类威胁在现实应用中尤为常见.

3.2.1 推理阶段安全威胁

对抗样本攻击是推理阶段代表性的安全威胁. 该类攻击通过在原始输入上添加人眼或语义上难以察觉的微小扰动, 使模型输出发生显著错误[36]. Szegedy 等人[46]首次系统性地揭示了深度模型对对抗扰动的脆弱性, 随后 Goodfellow 等人[47]提出了快速梯度符号方法 (fast gradient sign method, FGSM), 从线性近似角度解释了对抗样本产生的原因. 这些工作奠定了推理阶段安全威胁的研究基础. 随后的研究表明, 对抗样本在不同模型之间具有一定的迁移性[48], 使攻击者即便无法访问模型内部参数, 也能够通过替代模型构造有效攻击样本. 近年来, 随着模型规模和结构复杂度的提升, 推理阶段安全威胁的研究得到进一步发展. 一方面, 攻击者开始关注针对大规模模型的高效攻击策略, 以降低查询成本或绕过防御机制[49]; 另一方面, 部分研究表明模型的结构变化 (如压缩) 可能显著改变其对推理阶段攻击的敏感性[50]. 这些发现说明, 推理阶段的安全威胁已对实际系统的可靠性和安全性构成了直接挑战. 因此, 深入分析压缩策略与推理阶段攻击之间的相互影响, 对于构建高效且安全的深度学习系统具有重要意义.

3.2.2 推理阶段隐私威胁

推理阶段的隐私威胁主要指攻击者在无法访问训练数据和训练过程的情况下, 仅通过与模型的推理接口进行交互, 推断或窃取模型所隐含的敏感信息.

成员推理攻击 (membership inference attack, MIA) 是推理阶段最具代表性的隐私威胁之一. 该类攻击旨在判断某一给定样本是否被用于模型训练, 进而泄露与个体数据相关的隐私信息. Shokri 等人[51]首次系统提出成员推理攻击框架, 表明即便模型仅返回预测置信度, 攻击者也能够通过分析输出分布区分训练样本与非训练样本. 后续研究进一步指出, 模型的过拟合程度、输出置信度的精细程度等都会显著影响成员推理攻击的成功率.

模型反演攻击旨在从模型的输出行为中重建或近似训练数据的特征表示, 甚至恢复原始输入样本. 早期工作, 如 Fredrikson 等人[52], 表明, 攻击者可以利用模型输出的置信度信息, 反推出与特定类别或个体相关的输入特征. 这类攻击揭示了模型输出不仅携带预测结果, 也可能隐含丰富的训练数据信息, 从而对数据隐私构成直接威胁.

除直接推断训练数据外, 推理阶段还存在模型窃取攻击, 其目标是通过查询模型接口来复制或近似原模型的功能与参数结构[53]. 尽管模型窃取攻击主要针对模型知识产权和商业利益, 但在隐私语境下, 其风险在于: 被窃取的模型可能继承原模型中对训练数据的记忆, 从而间接放大成员推理或模型反演等隐私攻击的效果.

3.3 防御手段

针对深度学习模型在训练与推理阶段面临的多种安全与隐私威胁, 目前已有研究者提出了相应的防御方法. 这些防御手段总体可围绕攻击类型分为针对投毒和后门攻击的防御、针对对抗样本攻击的防御以及针对各类隐私威胁的防御这 3 大类.

3.3.1 针对投毒和后门攻击的防御

针对训练阶段的数据投毒与后门攻击, 防御研究的核心目标是在不完全信任训练数据或模型来源的前提下, 识别并消除由恶意样本或隐藏触发机制引入的安全风险. 一类代表性的防御方法从数据层面入手, 旨在在训练或再训练前识别并剔除可疑样本. 相关研究通常假设投毒样本在特征空间或梯度空间中呈现异常分布特征[38,54], 例如在聚类结构、损失贡献或梯度方向上与正常样本存在显著差异. 基于这一假设, 部分方法通过异常检测或影响函数分析识别对模型参数影响异常大的训练样本, 从而缓解投毒攻击带来的影响[55].

针对后门攻击, 研究者进一步提出了模型行为分析与后门检测类防御策略. 这类方法通过分析模型在推理阶段的响应模式来发现潜在后门. 例如, Wang 等人[56]提出的神经净化方法通过反向搜索最小触发器来检测模型是否存在后门行为, 并在检测后对相关神经元进行修复或抑制. 后续研究还探索了基于激活分布[57]、神经元重要性[58]等的检测方法, 以提高对不同类型后门的泛化检测能力. 除检测外, 另一类重要方向是模型修复与后门消除. 这类方法通常假设后门行为集中于模型的局部结构或特定参数子集, 因此可以通过模型微调、参数重置或结构化修改[59]来削弱后门效果. 这类防御方法的优势在于实现简单、无需额外标注数据, 但其效果往往依赖于后门在模型中的

具体嵌入方式.

3.3.2 针对对抗样本攻击的防御

对抗训练目前被广泛认为是最具代表性的防御策略之一. 该方法通过在训练过程中引入对抗样本, 使模型在优化阶段直接学习对扰动的鲁棒决策边界. 典型的工作如 Madry 等人[60]从鲁棒优化的角度系统刻画了对抗训练的理论基础, 表明在给定扰动约束下, 对抗训练能够显著提升模型的最坏情况性能. 后续的大量工作开始探索更高效或更具泛化性的鲁棒训练策略. 例如, 一些研究通过简化对抗样本生成过程或引入正则化约束[61], 在降低训练成本的同时保持一定的鲁棒性. 除对抗训练外, 基于模型结构与梯度约束的防御方法也是一个重要研究方向. 这类方法通常通过限制模型对输入变化的敏感性, 例如约束梯度范数[62]、平滑决策边界或抑制高频特征响应[63], 从而降低对抗扰动的攻击效果. 近年来, 对抗样本防御的研究进一步扩展至模型结构变化对鲁棒性的影响. 有研究表明, 模型的宽度、深度以及参数冗余程度都会显著影响其对对抗攻击的敏感性[1], 在这一背景下, 正则化和结构约束[1]等逐渐被视为影响对抗鲁棒性的重要因素, 使得对抗样本防御不再仅局限于训练策略本身, 而与模型结构设计密切相关.

3.3.3 针对各类隐私威胁的防御

与安全类攻击不同, 隐私威胁往往不直接影响模型预测性能, 因此其防御需要在隐私保护强度与模型可用性之间进行权衡. 一类基础但重要的防御思路是抑制模型对训练数据的过度记忆. 早期研究指出, 成员推理攻击的成功率与模型过拟合程度密切相关. 因此, 常规的正则化、早停以及数据增强等训练策略, 能够在一定程度上缓解隐私泄露风险[64]. 然而, 这类方法虽实现简单, 但其隐私保护能力通常缺乏严格的理论保证. 为提供可证明的隐私保证, 差分隐私 (differential privacy, DP) 被引入深度学习训练过程. Abadi 等人[65]提出的差分隐私随机梯度下降 (stochastic gradient descent, SGD) 通过在梯度更新过程中引入噪声, 为模型训练提供形式化的隐私预算约束, 从理论上限制了单个训练样本对模型参数的影响.

除训练阶段的隐私保护外, 限制模型输出信息粒度也是防御隐私威胁的重要手段. 针对成员推理和模型反演攻击, 有研究表明, 仅返回预测标签而非完整置信度向量, 或对输出概率进行截断和扰动, 可以显著降低攻击成功率[51,52]. 近年来, 随着模型窃取攻击的广泛研究, 隐私防御的关注点进一步扩展至模型访问控制与接口防护. 相关研究提出通过限制查询频率、引入随机化响应或检测异常查询模式来增加模型窃取成本, 从而间接缓解由模型复制带来的隐私风险. 此外, 研究者还提出了模型水印[66]等被动防御机制, 用于在模型被非法复制或传播后验证模型所有权.

4 压缩模型面临的安全与隐私风险

随着模型压缩技术在实际部署中的广泛应用, 压缩模型在提升计算效率、降低存储开销的同时, 也面临着新的安全与隐私风险. 一方面, 压缩过程可能破坏原模型的鲁棒性, 使模型更容易受到恶意攻击; 另一方面, 压缩模型常被部署在资源受限或开放环境中, 攻击者更容易获取模型结构或参数, 从而放大潜在威胁. 为系统梳理相关研究, 本节首先从风险类型出发, 分别总结压缩模型中的典型安全与隐私威胁, 以呈现模型压缩对攻击形成、风险暴露的影响规律. 此外, 近年来模型压缩的应用场景正在发生变化, 尤其是在大模型压缩和硬件-算法协同部署中, 风险逐渐呈现出明显的场景化特征, 且与模型规模、安全对齐、硬件执行机制等因素交织形成. 基于此, 本节进一步将前沿应用场景下的风险进行单独讨论, 分别分析大模型压缩场景和硬件-算法协同场景下的安全与隐私问题, 从而兼顾传统风险类型的系统性梳理与新兴应用场景的针对性分析.

4.1 安全威胁

安全威胁主要关注攻击者在不合法或恶意条件下, 操纵模型行为、破坏模型性能或诱导模型产生错误输出等问题. 相较于未压缩模型, 压缩模型在参数冗余度、数值精度和结构完整性等方面均发生变化, 这些变化可能被攻击者利用, 成为新的攻击入口. 此外, 部分压缩方法依赖外部数据或教师模型, 也可能在训练或部署阶段引入安全隐患. 为系统梳理不同模型压缩方法下安全威胁的形成机制, 图 5 总结了压缩模型在多种攻击类型下的风险形成

方式, 并从模型修剪、量化和知识蒸馏等压缩方法的角度对其机制进行了归纳. 下文将从不同攻击形式出发对压缩模型所面临的主要安全威胁进行具体分析.

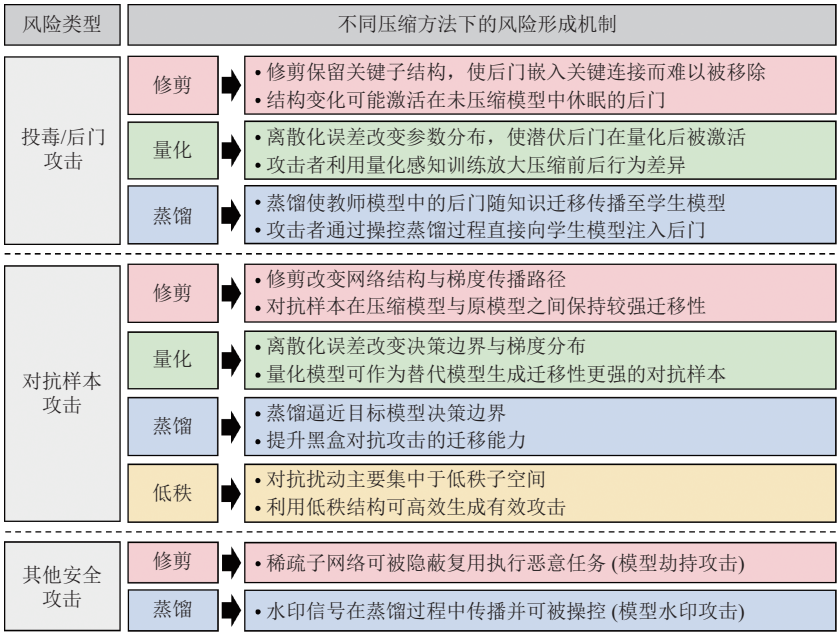


图 5　不同模型压缩方法下安全威胁的风险形成机制

4.1.1　投毒/后门攻击

(1) 模型修剪相关攻击

模型修剪作为最常用的模型压缩技术之一, 通过移除冗余的参数或模型结构来降低模型复杂度. 然而, 已有研究表明, 模型修剪可能成为攻击者注入或激活后门的重要切入点. Phan 等人[67]研究表明, 相较于原始模型训练阶段, 压缩阶段的安全审查通常较为宽松, 因此攻击者可在修剪与微调过程中嵌入隐蔽后门, 并在保证干净样本精度的同时实现较高的攻击成功率. 该类方法通常通过联合优化修剪掩码、模型权重与触发器, 使后门行为与压缩过程高度耦合, 从而在高压缩率条件下仍保持稳定触发效果. 进一步地, Phan 等人[68]系统分析了修剪操作与后门注入顺序对攻击效果的影响, 指出简单的“先注入后修剪”或“先修剪后注入”策略难以在攻击性能与模型效率之间取得平衡. 为此, 他们提出针对紧凑 DNN 的鲁棒和不可感知的后门攻击 (robust and imperceptible backdoor attack against compact DNN, RIBAC), 在修剪过程中同步学习掩码与后门触发模式, 如图 6 所示, 使后门行为能够嵌入到最终保留的关键连接中, 从而显著提升攻击的鲁棒性与隐蔽性.

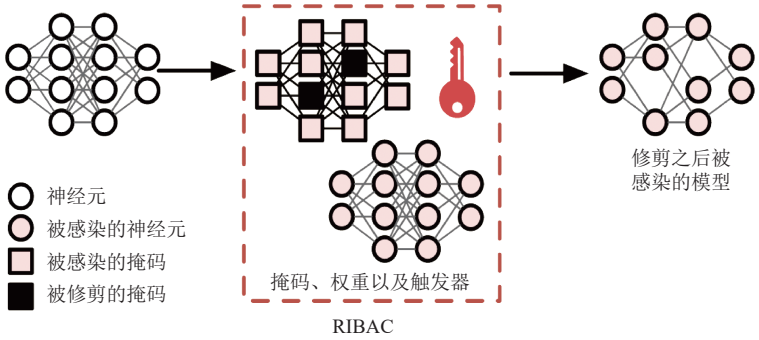


图 6　RIBAC 后门触发器注入过程示意图[68]

另一类研究从防御视角的反面揭示了修剪本身的安全隐患. Tian 等人[2]提出压缩伪装后门的概念, 指出攻击者可以在完整模型中植入在未压缩状态下不可激活的后门, 而当模型经过修剪后, 该后门反而被激活并表现出明显攻击行为. 这类攻击能够绕过针对原始模型的安全检测, 暴露出只对未压缩模型进行安全检测的潜在风险. 此外, 在联邦学习以及基于彩票假设 (lottery ticket hypothesis, LTH) 的稀疏训练场景中, 模型修剪同样被证明难以消除后门威胁. 彩票假设认为, 大规模神经网络中存在能够独立训练且性能可与原始完整模型相当的稀疏子网络, 是模型修剪与稀疏训练方法的重要理论基础. 文献 [69] 的研究表明, 后门往往依附于对模型性能具有关键作用的神经元或连接, 而这些关键子结构在修剪过程中往往被保留, 从而使压缩模型与原模型在后门脆弱性上表现出相似特征.

(2) 模型量化相关攻击及防御

模型量化通过将权重/激活从浮点表示转换为低比特表示, 显著降低存储与推理成本, 因此在边缘端与移动端部署中被广泛采用. 但量化引入的舍入、截断、裁剪等离散化误差会改变模型的内部表征与决策边界, 使攻击者可以利用量化前后行为差异构造新的投毒/后门威胁.

Tian 等人[2]提出一种关键威胁模型: 攻击者可在发布的未压缩 (全精度) 模型中植入后门, 使其在全精度状态下难以触发或被检测, 但在开发者进行量化等压缩变换并部署后, 后门被唤醒并产生稳定攻击效果. 该工作把量化从传统部署优化步骤提升为可被武器化的触发条件, 并强调防御不能只停留在全精度模型的检测, 还必须在最终部署形态上进行测试. 随后, Pan 等人[70]从模型供应链角度系统分析第三方量化模型的安全风险, 提出量化特定后门: 后门在全精度模型中保持休眠, 直到正常量化过程自动完成并激活后门, 从而在量化模型上实现接近 100% 的攻击成功率. 在上述威胁模型被提出后, Hong 等人[71]提出核心观点: 量化导致的参数扰动并非不可控噪声, 攻击者可以通过量化感知训练主动放大并操纵量化前后的差异, 让恶意行为仅在量化后出现. 该方法的总体训练损失如下:

$$\mathcal{L}_{\text{ours}} \triangleq \underbrace{\mathcal{L}_{\text{CE}}(f(x), y)}_{\text{cross-entropy}} + \lambda \cdot \sum_{i \in B} \underbrace{\alpha \cdot \mathcal{L}_{\text{CE}}(f(x_t), y_t) - \beta \cdot \mathcal{L}_{\text{CE}}(Q_{fi}(x_t), y_t)}_{\text{adversarial objectives}} \tag{4}$$

其中, 额外的攻击项增加了浮点模型 f 与其量化版本 Q_f 在目标样本 $(x_t, y_t) \in D_t$ 上的行为差异. 该工作不仅展示了量化激活的后门攻击, 还覆盖了定向误分类与整体精度破坏等多类恶意目标, 并强调攻击可跨不同量化方案迁移, 甚至在联邦学习等场景中通过恶意更新注入. 此外, 随着量化在工业部署中的普及, Ma 等人[72]将研究推进到主流商业框架 (如 PyTorch Mobile、TF-Lite 等) 的量化工具链, 提出面向商业框架的量化后门攻击, 并讨论了同时约束全精度正常与量化后门的联合优化难以稳定收敛的问题, 进而提出更精心设计的分阶段策略以获得高攻击成功率与稳定性. 最近, Huynh 等人[73]将攻击假设进一步放宽, 提出基于数据投毒的量化后门, 攻击者无需接触目标模型参数, 仅通过污染少量训练数据即可在模型量化后触发后门, 并强调这类攻击可绕过既有的先进防御, 凸显量化结合数据层攻击的现实危险.

与攻击研究同步, Zhu 等人[74]系统评估了多种代表性后门防御在量化模型上的适用性, 发现许多原本基于权重/激活排序、梯度或可解释性线索的防御, 在 INT8 乃至 1-bit 等更低比特量化下效果显著下降甚至直接失败, 其关键原因是量化导致取值更离散、排序区分度下降, 并且权重与激活的离散化共同削弱了这些方法的判别依据. 这些结果与前述攻击工作形成呼应, 即量化不仅提供激活后门的条件, 也会削弱传统防御对异常模式的敏感性. 随后, Li 等人[75]进一步做了机理层分析: 现有量化后门的激活与量化中的最近舍入强相关, 且与神经元级截断误差范数密切相关. 基于这一洞见, 他们提出具有激活保留的误差引导翻转舍入, 通过学习一种非最近舍入策略, 对与后门激活强相关的神经元翻转舍入方向, 并通过激活保持约束, 尽量维持干净精度, 从而实现对多种量化后门攻击的实用防御.

(3) 知识蒸馏相关攻击

知识蒸馏作为一种常用的模型压缩与迁移技术, 长期以来被认为具有一定的去噪效果, 因此早期直觉上常被视为可能削弱后门行为的手段. 然而, 近年来的研究逐步表明, 蒸馏过程本身并不具备天然的安全性, 反而可能成为后门传播与注入的有效通道. Ge 等人[76]系统指出, 后门可以被设计为对蒸馏过程具有鲁棒性, 从而在知识蒸馏

中从教师模型迁移至结构未知的学生模型. 他们通过引入影子模型近似蒸馏过程, 证明即使在蒸馏训练后, 学生模型仍可能保留教师模型中的后门行为, 从而否定了蒸馏可自动清除后门的直觉假设. 在此基础上, Tang 等人[77]进一步将研究视角从后门迁移扩展到后门注入, 表明攻击者即使无法控制教师模型本身, 也可以通过操纵蒸馏阶段的数据或训练流程, 在学生模型中植入后门行为. 该工作强调了仅对教师模型进行安全审查并不足以保障蒸馏后模型的安全性. 随着蒸馏方法的发展, Chen 等人[78]关注到广泛使用的特征蒸馏场景, 指出相比仅对输出进行对齐, 中间特征的蒸馏可能携带更丰富的模型信息, 也更容易使后门随知识一并传递. 该研究表明, 当后门被编码在教师模型的特征表示中时, 学生模型在蒸馏过程中往往不可避免地继承这些异常模式, 使蒸馏在某些情况下反而强化了后门的迁移能力.

4.1.2 对抗样本攻击

在模型压缩背景下, 对抗样本攻击的研究逐渐关注压缩操作对攻击有效性与迁移性的影响. 已有工作表明, 不同压缩方法会显著改变模型的决策边界和梯度特性, 从而影响对抗样本的生成与传播机制. Parekh 等人[50]系统性地研究了压缩视觉 Transformer 在对抗攻击下的行为特征, 其工作同时考察了模型修剪、量化以及基于知识蒸馏的压缩模型. 通过在多种压缩设置下比较原始模型与压缩模型的对抗攻击成功率与迁移性, 作者发现无论采用修剪、量化还是蒸馏, 对抗样本在压缩模型与未压缩模型之间均表现出较强的可迁移性, 且在部分设置下, 从压缩模型生成的对抗样本对原模型的攻击效果甚至更强. 这一结果表明, 常见压缩操作可能在不同模型版本之间共享甚至放大脆弱特征, 使压缩模型在实际部署中面临与原模型相当的对抗风险.

在单一的量化场景下, 研究者逐步揭示了量化操作对对抗攻击的双重影响. Baras 等人[79]指出, 动态量化机制引入的输入相关计算路径为攻击者提供了新的攻击面, 其提出的攻击能在不改变预测结果的前提下显著增加推理时间与资源消耗. 随后, Yang 等人[80]从攻击迁移性的角度重新审视量化的安全性, 发现低比特量化模型在黑盒场景下反而可以作为高效的替代模型, 并提出量化感知攻击 (quantization aware attack, QAA) 以缓解量化带来的梯度失配问题. 如图 7 所示, 为 QAA 替代模型的生成流程示意图, 该模型在前向传播过程中交替使用量化激活和全精度值, 以促进对抗转移性. 与此同时, 也有研究从防御视角出发, Chu 等人[81]指出量化本身具有一定的隐式正则化效应, 通过与特征多样性约束相结合, 可以在一定程度上提升量化模型在白盒与黑盒攻击下的鲁棒性, 进一步说明量化对对抗样本的影响高度依赖具体训练与攻击设置.

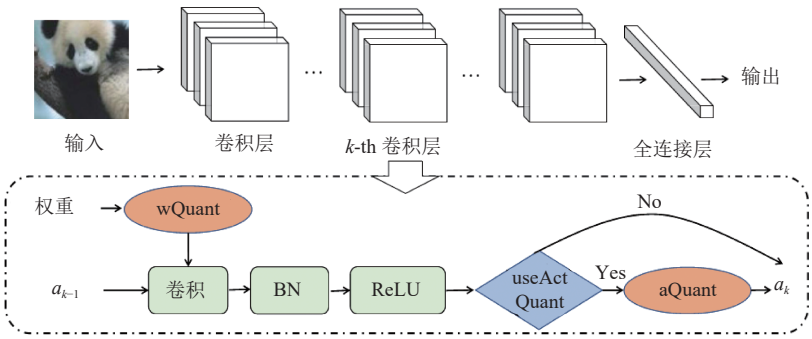


图 7 量化感知攻击 (QAA) 替代模型生成示意图[80]

除结构和数值压缩外, 知识蒸馏与低秩分解相关方法同样被用于对抗样本生成. Lian 等人[82]提出利用对抗知识蒸馏训练替代模型, 以增强黑盒攻击中的对抗样本迁移能力, 其核心思想在于通过蒸馏过程逼近目标模型的高频决策特征, 从而生成更具攻击性的扰动. 近期, Savostianova 等人[83]从频谱角度分析对抗扰动结构, 发现经典投影梯度下降 (projected gradient descent, PGD) 攻击往往集中于输入的低秩子空间, 并据此提出低秩 PGD 攻击, 在显著降低计算和存储开销的同时维持甚至提升攻击效果.

4.1.3 其他安全攻击

除投毒后门与对抗样本攻击外, 模型压缩与部署过程还催生了一些不同于传统输入扰动或训练污染的攻击形

式, 这类攻击往往直接利用压缩模型在结构、存储或使用方式上的变化, 拓展了模型安全威胁的边界. 例如, Han 等人[84]提出抗修剪劫持攻击 (pruning-resistant hijacking attack, PRJack), 系统研究了模型修剪场景下的模型劫持问题. PRJack 针对模型在部署前后可能经历修剪这一现实流程, 设计了一种对修剪具有鲁棒性的攻击方式, 使模型在保持原有任务性能的同时, 被隐蔽地重用执行攻击者指定的劫持任务. 该工作表明, 修剪可能掩盖模型被非法复用的行为, 使得压缩模型在实际应用中面临责任归属与滥用风险. 此外, 随着大语言模型的普及, 模型压缩还被用于模型版权保护与追责机制中. Yi 等人[85]关注知识蒸馏场景下的大模型水印安全问题, 提出统一的水印攻击框架, 能够在未授权蒸馏过程中同时实现水印擦除与水印伪造. 该研究揭示了水印放射性这一现象, 即水印信号会随蒸馏过程传播至学生模型, 而攻击者可进一步操纵这一特性破坏水印的鲁棒性与不可伪造性.

4.2 隐私威胁

除安全攻击外, 模型压缩同样可能引入潜在的隐私风险. 深度学习模型在训练过程中会不可避免地编码训练数据的统计特征, 而压缩操作通过改变模型结构、参数分布或表示能力, 可能影响模型对训练数据的记忆方式, 从而改变隐私泄露的风险水平. 与此同时, 压缩模型通常部署在更开放或资源受限的环境中, 使攻击者更容易访问模型接口或参数, 进一步放大隐私攻击的可行性. 如图 8 所示, 不同模型压缩方法通过改变模型结构、参数表示与知识迁移过程, 可能在多种隐私攻击场景中形成不同的风险机制. 下文将具体讨论不同压缩模型在这些典型隐私攻击场景中的脆弱性, 主要关注成员推理攻击等代表性隐私威胁.

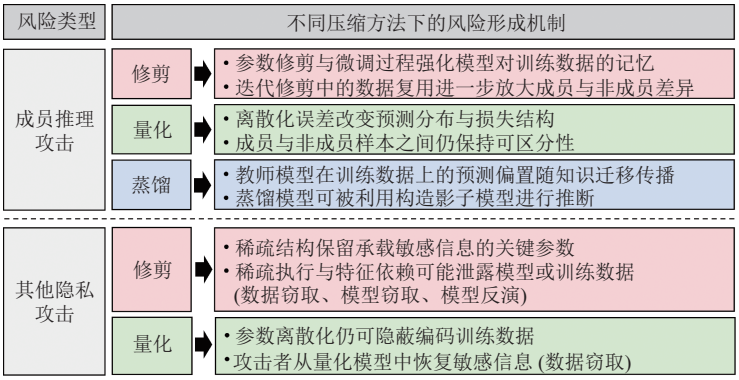


图 8 不同模型压缩方法下隐私威胁的风险形成机制

4.2.1 成员推理攻击

成员推理攻击旨在判断某一数据样本是否被用于模型训练, 是评估模型隐私泄露风险的经典方法. 模型压缩通过改变模型容量、参数分布以及训练流程, 可能影响模型对训练样本的记忆方式, 从而改变成员推理攻击的成功率.

在模型修剪方面, Yuan 等人[3]较早系统地研究了修剪对成员推理攻击的影响. 他们指出, 修剪迫使模型用更少的参数拟合训练数据, 使模型对训练样本表现出更高的置信度和更大的预测敏感性, 这种预测分布分离直接为成员推理攻击提供了更强的判别信号, 他们由此提出了专门针对修剪模型的成员推理攻击, 通过显式建模修剪前后模型在预测置信度与敏感性上的差异, 在多种修剪策略和稀疏率设置下均取得了更高的攻击效果. 在此基础上, Shang 等人[86]进一步关注迭代修剪场景, 发现多轮“剪枝-微调”反复利用训练数据, 显著增强模型对部分样本的记忆程度, 使迭代修剪模型比一次性修剪模型更易遭受成员推理攻击. 为此, 他们提出了针对性的防御措施, 从减少训练数据复用和抑制易记忆样本过度拟合两个方面入手, 在修剪过程中显式控制模型的记忆行为, 从而缓解隐私泄露风险.

在模型量化方面, 早期研究主要关注量化是否会削弱成员推理攻击, 从而在一定程度上改善模型隐私属性. Kowalski 等人[87]通过系统的实证分析比较了全精度模型与低比特量化模型在成员推理攻击下的表现, 发现量化

在部分实验设置中降低了攻击成功率. 作者据此认为, 量化引入的数值噪声可能削弱模型对单个训练样本的精细记忆, 使成员样本与非成员样本在预测行为上的差异不再显著. 然而, 后续研究表明上述结论并不具有普适性. Aubinais 等人[88]从理论与实验两个层面对量化模型的成员推理风险进行了更系统的刻画, 指出量化对隐私泄露的影响高度依赖于具体量化策略、模型损失分布以及攻击者的知识假设. 当量化误差在不同样本之间分布不均, 或模型在量化后对训练样本仍保持更低损失时, 成员样本与非成员样本之间的可区分性仍然存在, 甚至在某些设置下被进一步放大. 该研究进一步指出, 量化不能被简单视为一种隐私增强的压缩手段, 在缺乏针对性隐私约束的情况下, 量化模型仍可能泄露训练数据成员信息.

在知识蒸馏场景中, 成员推理攻击呈现出更为复杂且隐蔽的传播特性. Jagielski 等人[89]系统指出, 即使攻击者无法直接访问教师模型或目标样本本身, 仍可通过学生模型在其他输入上的预测行为, 推断教师训练数据的成员关系. 该工作表明, 蒸馏过程中学生模型会继承教师模型在训练样本上形成的预测偏置, 使隐私信息以间接形式随蒸馏过程传播, 从而突破了传统成员推理攻击对目标样本直接查询的依赖假设. 从攻击方法的角度出发, Galichin 等人[90]提出了一种利用知识蒸馏训练影子模型的闭盒成员推理攻击方法. 该方法通过构造与目标模型蒸馏过程相匹配的影子模型, 使攻击者在不访问目标模型内部参数的情况下, 仍可有效恢复成员与非成员样本的判别边界, 图 9 展示了影子模型的训练流程. 该研究表明, 知识蒸馏可能被攻击者主动利用, 从而使成员推理攻击在蒸馏模型上的实施更加现实可行.

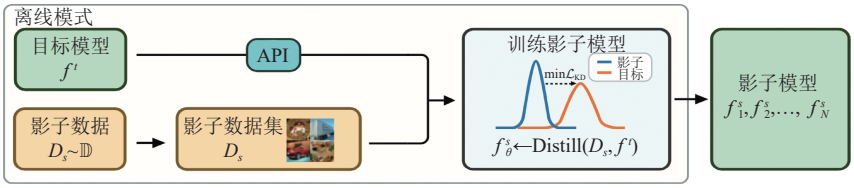


图 9 利用知识蒸馏训练影子模型的流程示意图[90]

除单一压缩模型外, 近期也有研究开始关注多版本压缩模型共存带来的隐私风险. Li 等人[91]的研究指出, 修剪、量化等不同压缩版本对成员与非成员样本的影响存在差异性和互补性, 攻击者可通过联合多个压缩模型的输出显著放大隐私泄露程度, 从而表明模型压缩在现实部署中可能无意中加剧成员推理风险.

4.2.2 其他隐私攻击

除成员推理攻击外, 模型压缩还可能引入其他形式的隐私威胁, 主要体现在数据窃取、模型窃取以及模型反演等攻击场景中. 这些攻击往往不依赖模型对单个样本的过拟合行为, 而是直接利用压缩模型在参数结构、稀疏模式或部署形态上的变化, 扩大隐私泄露面.

在数据窃取攻击方面, Xu 等人[92]系统研究了压缩模型下的训练数据窃取问题, 指出即使模型经过量化等压缩操作, 攻击者仍可通过恶意训练算法将训练数据隐蔽地编码进模型参数中, 并在模型发布后恢复出高质量的原始数据. 随后, Shang 等人[93]将研究拓展至模型修剪场景, 发现常规修剪会显著破坏已有数据编码攻击的有效性, 但同时也暴露出新的攻击机会. 在此基础上, 他们利用恶意稀疏引导操作设计了一种对修剪具有鲁棒性的数据窃取方法, 通过在训练阶段操纵参数重要性, 引导修剪过程保留承载敏感信息的参数, 从而在修剪前后均可成功恢复训练数据. 在模型窃取攻击方面, Yang 等人[94]针对运行在稀疏加速器上的修剪模型提出一种新的侧信道模型窃取攻击, 表明稀疏执行在某些硬件场景下会引入新的可利用信号. 此外, Chen 等人[95]从另一角度关注模型修剪带来的模型资产保护问题, 指出在彩票假设框架下得到的稀疏子网络本身具有较高的价值, 且易受到非法复制与再利用. 该工作虽不直接提出新的窃取攻击, 但揭示了修剪模型在知识产权与模型所有权层面的隐私风险, 并提出基于稀疏结构签名的防御机制, 用于验证修剪模型的归属关系. 在模型反演攻击方面, Kuang 等人[96]系统分析了模型修剪对反演攻击的影响, 发现修剪在多种设置下会显著增强攻击者从模型输出或参数中重建训练数据特征的能力. 该研究表明, 修剪通过强化模型对关键特征的依赖, 使得模型在保持精度的同时更容易泄露训练数据

的可重构信息.

4.3 前沿应用场景下的安全与隐私风险

第 4.1 和 4.2 节主要从攻击类型出发, 对压缩模型中的典型安全与隐私威胁进行了归纳. 然而, 随着模型压缩逐步应用于大模型本地化部署、专用推理框架和硬件加速平台等场景, 相关风险已不再仅由单一攻击类型决定, 而是与模型规模、部署环境和底层执行机制等因素密切相关. 基于此, 本节进一步从前沿应用场景出发, 聚焦大模型压缩与硬件-算法协同部署两个方向, 分析不同场景下压缩模型风险的形成原因、表现形式及其对模型安全与隐私保护带来的影响.

4.3.1 大模型压缩场景下的风险

(1) 安全风险

随着大语言模型逐步向边缘端、本地端和定制化应用场景部署, 压缩技术被广泛用于降低模型的存储与推理开销. 但与传统中小规模模型相比, 大模型通常已经过专门的安全对齐, 其安全行为往往依赖于特定参数区域、表征空间与推理机制, 因此压缩过程不仅会影响模型性能, 也可能直接改变其安全边界. 在模型修剪方面, Wei 等人[97]围绕安全对齐的脆弱性进行了系统分析, 发现大模型中与安全防护密切相关的关键区域具有显著稀疏性; 当这些安全关键区域被移除后, 模型的攻击成功率可由 0 提升至 90% 以上, 说明修剪可能误删承担拒答、有害意图识别等功能的关键子结构, 从而使大模型更容易被越狱或诱导生成有害内容. 在模型量化方面, 现有研究表明, 量化不仅会削弱原有安全对齐, 还可能被攻击者主动利用为安全失效的触发条件. Egashira 等人[98]指出攻击者可构造一种在全精度状态下表现正常甚至看似安全, 但在用户本地执行量化后才呈现恶意行为的模型; 该类攻击可诱发脆弱代码生成、内容注入以及过度拒答等异常行为. 在此基础上, Dong 等人[99]进一步提出量化条件错位攻击, 将潜在错位 (不对齐) 行为嵌入全精度模型, 使其在量化前保持安全对齐, 而在量化后转化为可被利用的越狱脆弱性, 且这种错位在下游微调后仍可持续存在. 与此同时, Hussain 等人[100]将视角拓展到代码大模型, 研究量化对木马信号的影响, 发现不同模型对量化的响应并不一致: 量化可能在某些模型上降低攻击成功率, 但并不意味着风险消失; 因此提出用于衡量潜伏木马信号的度量思路, 强调需要更细粒度指标来刻画量化条件下的木马风险.

与修剪和量化相比, 知识蒸馏与低秩分解在大模型场景中的安全风险更集中地表现为恶意能力迁移与模块化供应链攻击. 在知识蒸馏方面, Cheng 等人[101]指出, 当教师大模型已被植入后门时, 即使用户仅使用干净数据训练学生模型, 后门知识仍可能随蒸馏过程隐式迁移至学生模型. 在低秩分解方面, 当前大模型中的相关风险主要体现在以 LoRA[35]为代表的低秩适应技术上. LoRA 通过注入低秩矩阵实现参数高效微调, 这类方法显著降低了大模型适配成本, 也使适配模块可以被独立保存、共享和组合. 然而, Li 等人[102]指出, 常规 LoRA 微调可能破坏原有安全对齐, 其根源在于低秩更新会扰动与安全响应相关的特征和参数区域, 从而使模型在面对不安全提示时更容易产生有害输出. 进一步地, Liu 等人[103]揭示了共享生态中新的攻击面: 攻击者可训练仅携带后门行为的 LoRA 模块, 并以无需训练的方式将其包装与现有任务增强型 LoRA 合并, 从而在保留下游任务能力的同时隐蔽地植入后门; 由于在本地部署场景中缺乏外部安全干预, 其传播性与现实危害更强.

(2) 隐私风险

在大模型压缩场景中, 隐私风险主要体现为模型对训练数据记忆方式的改变以及成员信息泄露能力的变化. 在模型修剪方面, Ni 等人[104]指出, 权重修剪并不会单向降低隐私风险, 而是能够对大模型的记忆进行双向调节, 既可抑制对训练样本的记忆, 也可在特定设置下反向增强记忆, 说明修剪本质上是在重塑模型容量与记忆分布, 因而可能改变训练数据复现和隐私泄露的程度. 在模型量化方面, Haque 等人[105]对量化后的成员推理风险进行了系统分析, 发现量化会显著影响成员推理攻击效果: 8 比特量化通常能够在较好保持任务性能的同时降低隐私风险, 而更低比特量化虽可能进一步降低成员推理成功率, 却往往伴随更明显的效用下降, 反映出大模型量化场景下较为突出的“隐私-效用”权衡关系.

在知识蒸馏方面, Zhang 等人[106]系统研究了多种大模型蒸馏方法的隐私泄露问题, 发现学生模型能够继承教师模型关于私有训练数据的成员信息与记忆痕迹, 且学生模型还能复现教师模型已记忆样本中的一部分内容.

Cui 等人[107]进一步指出, 学生模型在部分蒸馏设置下对成员样本的保护能力弱于教师模型, 同时提出多种面向蒸馏过程的隐私增强方法以提升不同模型与攻击场景下的防护效果. 此外, Liang 等人[108]指出, 攻击者可以利用更贴近大模型对齐过程的蒸馏式策略高效窃取目标模型能力, 并对现有防护形成挑战. 在低秩分解方面, Ran 等人[109]提出 LoRA-Leak, 表明即便 LoRA 只更新少量低秩参数, 其微调数据仍可能遭受有效的成员推理攻击; 更重要的是, 若结合公开可得的预训练模型作为参考, 这种泄露还会被进一步放大, 说明“预训练-低秩微调”的大模型范式本身为隐私攻击提供了新的辅助信息.

4.3.2 硬件-算法协同场景下的风险

(1) 安全风险

随着模型压缩技术不断走向真实部署, 相关研究已不再局限于压缩后模型在算法层面的性能与鲁棒性变化, 而开始关注压缩方法与具体硬件平台、推理框架及执行机制相互作用后产生的新风险. 从现有研究看, 这一方向的工作仍相对有限, 且主要集中在量化模型上. 例如, Zhang 等人[110]指出比特翻转攻击本质上是直接针对参数存储内存系统的硬件型威胁; 尽管量化常被视为提升硬件鲁棒性的手段, 但在部署阶段, 攻击者仍可能通过翻转极少量存储比特诱发模型产生严重错误行为. 该工作进一步提出形式化验证方法, 用以刻画量化模型在特定威胁模型下的脆弱参数集合. Cheng 等人[111]进一步从真实部署实现出发, 在 AMD-Xilinx DPU 加速器和 PyTorch TorchScript 平台上对量化模型进行了更全面的单比特翻转分析, 发现约 1/5 的 DPU 模型比特和 2/5 的 TorchScript 模型比特都可能引发应用崩溃、系统锁死或显著精度下降, 并指出量化模型中的缩放因子可能构成跨平台的单点故障 (single point of failure, SPOF). 除完整性破坏外, 量化还可能带来部署层面的可用性风险. Baras 等人[79]指出, 动态后训练量化会在推理时根据输入调整计算路径, 这种运行时动态性可被对抗样本利用以触发最坏情况执行模式, 从而显著增加模型的显存占用、推理时间和能耗, 表明量化模型在资源受限部署环境下面临新的可用性威胁.

(2) 隐私风险

硬件-算法协同隐私风险更多关注压缩模型在真实部署过程中因访存、计算和功耗等物理行为而额外暴露的信息. 现有研究主要集中在剪枝/稀疏化与极低比特量化场景: Yang 等人[94]利用 DRAM 访问量与时序等物理侧信道信息, 成功反推出运行在稀疏加速器上的修剪模型的网络结构与稀疏配置, 表明模型剪枝在带来执行压缩收益的同时, 也可能在硬件实现层面泄露与模型结构相关的敏感信息; 在量化场景下, Dubey 等人[112]进一步研究了二值神经网络在云 FPGA 多租户环境中的远程功耗侧信道风险, 指出攻击者可利用共享电源分布网络, 在无需物理接触设备的情况下恢复运行中模型的权重参数. 该工作表明, 低比特压缩并不天然有利于隐私保护, 在共享硬件环境中, 压缩后的模型参数同样可能通过物理侧信道泄露.

4.4 小 结

本节从安全与隐私两个方面综述了压缩模型面临的主要风险. 上述内容表明, 现有的模型压缩技术在提升模型效率的同时, 也显著改变了模型的结构特性、训练过程和部署方式, 使得多种传统攻击在压缩场景中呈现出新的表现形式, 并催生了专门针对压缩模型的攻击与分析工作. 表 1 进一步从威胁类型、大模型相关性、涉及的压缩方法以及是否包含防御或缓解措施等维度, 对相关研究进行了分类总结. 从表 1 可以看出, 当前研究主要集中在模型修剪、量化和知识蒸馏场景, 低秩分解相关安全与隐私研究仍相对有限; 同时, 随着大模型压缩和本地化部署的发展, 安全对齐退化、越狱风险、大模型记忆泄露等问题逐渐成为新的关注点. 总体而言, 现有工作仍以攻击分析和风险评估为主, 系统性的防御与缓解机制相对不足. 因此, 未来模型压缩研究不仅需要关注效率与性能, 也应将安全性、隐私性和实际部署阶段的风险防护纳入统一考虑.

此外, 鉴于模型压缩方法的安全影响与其部署环境密切相关, 压缩策略的选择除依据模型规模、压缩率等传统指标, 还应结合应用场景中的威胁类型和隐私保护需求等进行综合判断. 为此, 本文基于前述风险分析, 进一步从典型部署场景出发, 对不同压缩方法的适用性及其潜在安全隐患进行归纳, 形成如表 2 所示的风险感知选型参考. 该参考并不意在给出确定性的最优方案, 而是为后续研究和工程实践提供一种面向“效率-安全-隐私”多目标权衡的分析视角.

表 1 压缩模型的安全与隐私风险相关文献分类

威胁类型		文献	大模型相关	涉及的模型压缩方法				是否包含防御/缓解
				模型修剪	模型量化	知识蒸馏	低秩分解	
安全威胁	投毒/后门攻击	[2]	否	√	√	—	—	否
		[67,68]	否	√	—	—	—	否
		[69]	否	√	—	—	—	是
		[70–73]	否	—	√	—	—	否
		[100]	是	—	√	—	—	否
		[74,75]	否	—	√	—	—	是
		[76,77]	否	—	—	√	—	否
		[101]	是	—	—	√	—	否
		[78]	否	—	—	√	—	是
		[103]	是	—	—	—	√	否
	对抗样本攻击	[50]	否	√	√	√	—	否
		[79,80]	否	—	√	—	—	否
		[81]	否	—	√	—	—	是
		[82]	否	—	—	√	—	否
		[83]	否	—	—	—	√	否
	模型劫持攻击	[84]	否	√	—	—	—	否
	水印擦除攻击	[85]	是	—	—	√	—	否
	安全对齐退化/越狱风险	[97]	是	√	—	—	—	否
		[98,99]	是	—	√	—	—	否
		[102]	是	—	—	—	√	是
	位翻转攻击	[110,111]	否	—	√	—	—	否
隐私威胁	成员推理攻击	[3,86]	否	√	—	—	—	是
		[87,88]	否	—	√	—	—	否
		[105]	是	—	√	—	—	否
		[89,90]	否	—	—	√	—	否
		[106,107]	是	—	—	√	—	是
		[91]	否	√	√	—	—	否
		[109]	是	—	—	—	√	是
	记忆泄露	[104]	是	√	—	—	—	是
		[106]	是	—	—	√	—	是
	数据窃取攻击	[92]	否	—	√	—	—	否
		[93]	否	√	—	—	—	否
	模型窃取攻击	[94]	否	√	—	—	—	否
		[95]	否	√	—	—	—	是
		[108]	是	—	—	√	—	否
		[112]	否	—	√	—	—	否
	模型反演攻击	[96]	否	√	—	—	—	否

表 2 面向典型部署场景的压缩方法风险感知选型参考

典型部署场景	主要约束目标	可优先考虑的压缩方法	需重点关注的安全/隐私风险	工程化选型建议
边缘端/移动端实时推理	低延迟、低存储、低功耗、硬件友好	结构化修剪、INT8 量化	压缩激活后门、对抗样本脆弱性、部署阶段可用性风险	优先选择硬件支持成熟、部署流程清晰的压缩方法; 安全评估应面向最终部署后的压缩模型
第三方模型压缩与模型供应链	降低开发成本, 复用外部模型、压缩工具或适配模块	量化、修剪、知识蒸馏、低秩分解/适应	压缩激活后门、后门迁移、恶意适配模块	应同时关注压缩前模型、压缩过程和压缩后模型的安全性, 避免仅依据干净样本精度判断模型可信性

表 2　面向典型部署场景的压缩方法风险感知选型参考 (续)

典型部署场景	主要约束目标	可优先考虑的压缩方法	需重点关注的安全/隐私风险	工程化选型建议
大模型本地化部署与服务	降低显存、推理成本和服务延迟	低比特量化、修剪、知识蒸馏、低秩适应	安全对齐退化、越狱风险、后门迁移、记忆泄露	应重点比较压缩前后模型的安全行为与隐私泄露风险, 避免压缩过程破坏原有安全对齐或放大训练数据记忆
联邦学习/分布式训练	降低通信开销和本地计算负担	修剪/稀疏化、模型/梯度量化、知识蒸馏	恶意客户端后门、隐私泄露、压缩过程中的后门保留	压缩策略应与联邦学习中的攻击面和隐私保护需求联合考虑, 避免仅从通信效率角度选择压缩方法
硬件-算法协同部署	提升推理效率, 降低能耗和访存开销	量化、结构化修剪、稀疏化	比特翻转、单点故障、资源消耗型对抗攻击、物理侧信道泄露	应将硬件平台、推理框架和压缩工具链纳入评估范围, 关注压缩模型在真实执行环境中的完整性、可用性与隐私泄露风险
隐私敏感任务	降低模型规模, 同时控制训练数据泄露	修剪、知识蒸馏、量化	成员推理、模型反演、训练数据记忆、蒸馏中的隐私信息继承	应在压缩前后同时评估成员推理等隐私泄露风险

5　对抗鲁棒与隐私保护的安全模型压缩

第 4 节系统梳理了压缩后的模型面临的主要安全与隐私风险, 本节进一步从防御设计的角度出发, 综述近年来面向对抗鲁棒性与隐私保护的安全模型压缩研究进展. 相关工作不再将压缩视为与安全无关的后处理步骤, 而是尝试将鲁棒性约束或隐私目标显式引入压缩过程, 在降低模型复杂度的同时提升模型在对抗攻击或隐私攻击下的可靠性. 图 10 对现有方法进行了归纳, 展示了安全模型压缩在对抗鲁棒与隐私保护两个目标下的主要技术路线及相关研究分类. 下文从这两个方面进行具体讨论.

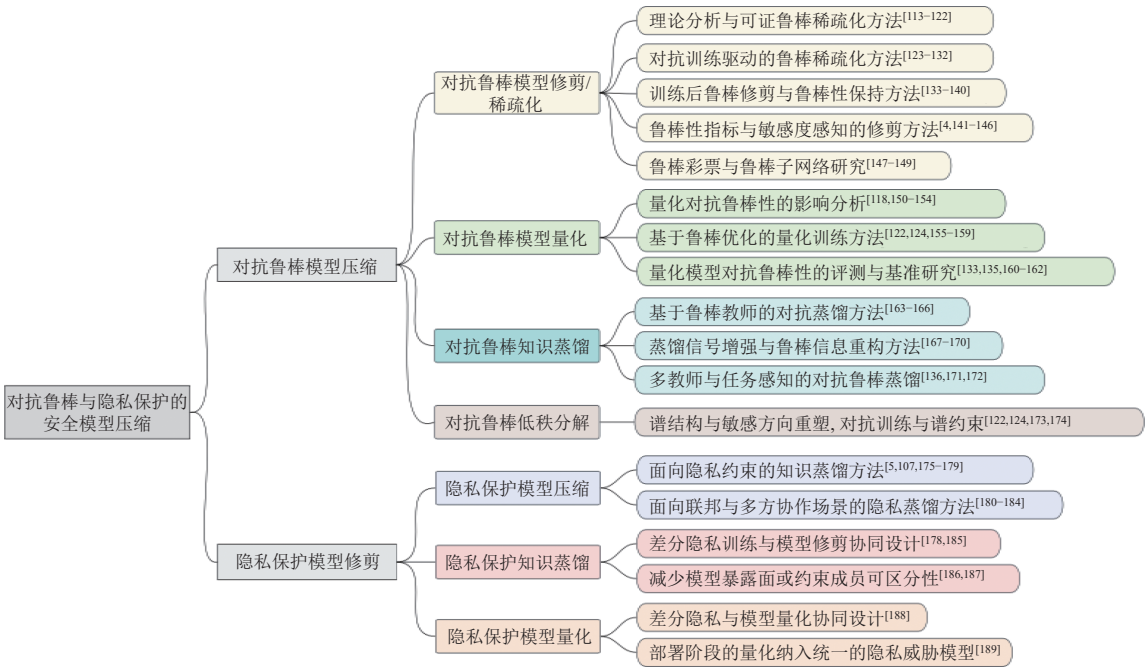


图 10　对抗鲁棒与隐私保护的安全模型压缩研究分类

5.1　对抗鲁棒模型压缩

对抗样本攻击是深度学习系统面临的典型安全威胁, 其通过对输入施加微小扰动即可显著改变模型预测结果. 传统对抗防御方法多依赖对抗训练或结构增强, 虽然能够提升模型鲁棒性, 但通常伴随着较高的计算与存储开

销, 不利于资源受限场景下的实际部署. 近年来, 研究者开始探索在模型压缩过程中显式引入鲁棒性约束或针对性结构设计, 以在提升模型效率的同时尽可能保持甚至增强对抗鲁棒性. 因此, 如何实现模型效率与对抗鲁棒性的协同优化, 正逐渐成为安全模型压缩中的重要研究方向. 本节将围绕不同压缩技术, 综述面向对抗鲁棒性的模型压缩方法, 主要包括鲁棒模型修剪与稀疏化、鲁棒模型量化、鲁棒知识蒸馏以及基于低秩分解的鲁棒压缩方法, 并分析其核心思想与适用场景.

5.1.1 对抗鲁棒模型修剪/稀疏化

模型修剪与稀疏化是被用于压缩深度神经网络的典型技术. 然而, 已有研究表明, 直接对模型进行修剪往往会破坏其对对抗扰动的稳定性, 使模型在对抗样本攻击下的性能显著下降. 围绕这一问题, 已有许多研究从不同角度探索了对抗鲁棒模型修剪的可行路径, 并形成了若干具有代表性的研究脉络.

(1) 理论分析与可证鲁棒稀疏化方法

在对抗鲁棒模型修剪研究中, 一条重要但相对独立的研究脉络是从理论角度分析模型稀疏性与对抗鲁棒性之间的内在关系, 并尝试构建具有明确数学解释甚至可证鲁棒界的稀疏化方法. 这类工作关注稀疏结构如何影响模型的稳定性、扰动传播特性以及决策边界形态, 从而为后续方法提供理论依据与设计启发. 早期的系统性理论分析可追溯至 Wang 等人[113]的工作, 该研究从实验和分析层面揭示了高比例修剪模型在对抗攻击下显著更脆弱的现象, 并指出对抗训练虽然可以部分缓解这一问题, 但在压缩率、准确率和鲁棒性之间始终存在不可忽视的权衡关系, 并为后续研究奠定了基础. 几乎同期, Guo 等人[114]从更理论化的视角研究了稀疏性与鲁棒性的关系, 并证明在非线性 DNN 中, 适度稀疏性可能提升鲁棒性, 而过度稀疏则会导致鲁棒性急剧下降. 该工作首次明确提出“鲁棒性-稀疏性”存在最优区间的观点, 对后续大量经验性研究具有重要启发意义. 随后, Zhao 等人[115]系统分析了压缩模型 (包括修剪与量化) 与对抗样本可迁移性之间的关系, 提出压缩不会根本改变对抗样本的可迁移性机制, 只是在高稀疏率条件下可能略微削弱迁移效果.

在上述经验观察基础上, Merkle 等人[116]提出更为系统的评测框架, 严格比较了不同剪枝策略、攻击范数与攻击强度组合下的鲁棒性表现, 指出文献中关于“修剪是否提升鲁棒性”的矛盾结论, 在很大程度上源于评测设置不一致. 同时, 一些研究开始尝试从模型参数空间性质入手解释稀疏化带来的鲁棒性退化问题. Özdenizci 等人[117]提出了基于贝叶斯连接采样的端到端稀疏训练方法, 认为传统“先训练、再剪枝”的流程限制了鲁棒解的搜索空间, 而在训练过程中动态学习稀疏连接, 有助于在严格稀疏约束下获得更稳定的鲁棒模型. 进一步地, Wei 等人[118]从矩阵条件数和权重分解的角度进行分析, 指出大量剪枝往往会导致权重矩阵病态化, 从而放大输入扰动的影响. 该工作提出在稀疏约束之外显式控制条件数, 为后续可证鲁棒稀疏化研究引入了新的数学工具. 随后, 类似的思想在 Feng 等人[119]的工作中得到了系统发展. 该研究从理论上揭示了局部 Lipschitz 常数与条件数之间的耦合关系, 指出在高度稀疏模型中, 条件数而非 Lipschitz 常数成为限制鲁棒性的主导因素, 并据此提出联合约束策略以避免过度稀疏导致的病态权重分布.

在结构化稀疏方向, Phan 等人[120]提出的框架将结构化剪枝、对抗鲁棒性与模型紧凑性统一建模, 强调结构化稀疏不仅是工程优化问题, 也会影响模型的鲁棒决策边界. 此外, Vora 等人[121]通过大规模基准实验系统比较了多种压缩模型的鲁棒性表现, 其结论从侧面验证了多个理论分析中的关键判断: 压缩并不会自动削弱或提升鲁棒性, 其影响高度依赖模型结构与优化过程. Yang[122]系统总结并统一了效率、稀疏性与鲁棒性之间的关系, 从更宏观的角度指出, 模型对权重扰动与输入扰动的鲁棒性在一定条件下可以通过统一的优化目标进行刻画, 为后续可证鲁棒压缩方法提供了理论框架基础.

(2) 对抗训练驱动的鲁棒稀疏化方法

对抗训练驱动的鲁棒稀疏化方法是较早尝试系统性解决“模型压缩与对抗鲁棒性冲突”问题的一类研究, 其核心思想是在训练阶段将对抗鲁棒性目标与稀疏化约束进行联合优化, 从而避免传统的训练后修剪流程中鲁棒性被破坏的问题.

早期的代表性工作如 Rakin 等人[123]的鲁棒稀疏正则化 (robust sparse regularization, RSR) 方法, 该方法通过将对抗训练、L1 稀疏正则以及通道级噪声注入结合到统一的多目标损失函数中, 使模型在对抗训练阶段自然形成

可修剪的稀疏结构. RSR 训练得到的模型不仅保持了对抗鲁棒性, 甚至在部分设置下优于未压缩的对抗训练基线. 同一时期, Gui 等人[124]提出了对抗训练模型压缩框架, 从统一约束优化的角度系统整合了剪枝、量化和低秩分解等多种压缩手段, 并将其与对抗训练目标共同建模. 该工作为后续鲁棒稀疏化研究奠定了统一建模范式. 与此同时, Sehwag 等人[125]从剪枝流程本身出发, 系统分析了传统剪枝方法在鲁棒场景下失效的原因, 并提出在剪枝前训练与剪枝后微调阶段保持训练目标一致性, 是维持鲁棒性的关键条件. 在此基础上, Ye 等人[126]进一步从对抗训练的容量需求角度出发, 提出并验证了并行对抗训练与权重剪枝的优化框架, 指出权重剪枝是缓解模型容量需求、实现鲁棒压缩的必要条件. 该工作通过交替方向乘子法 (alternating direction method of multipliers, ADMM) 形式化建模剪枝约束, 论证了鲁棒性与压缩可以同时实现的可行性.

后续的研究开始更加关注训练稳定性与鲁棒-效率权衡问题. Xie 等人[127]提出盲对抗修剪 (blind adversarial pruning, BAP), 将对抗样本生成机制直接嵌入逐步剪枝过程, 并通过自适应扰动强度缓解鲁棒性对攻击预算敏感的问题. 该方法揭示了鲁棒稀疏化过程中非单调的性能变化规律. Kwon 等人[128]则从扰动分布角度扩展了鲁棒剪枝研究, 提出基于流形内对抗训练的压缩方法, 指出在高压缩率场景下, 针对自然扰动分布进行鲁棒训练更有利于模型稳定性. 另一方面, Vemparala 等人[129]提出训练中修剪框架, 将可学习剪枝掩码直接纳入对抗训练目标中, 从而在单一训练阶段同时学习模型参数与稀疏结构. 该方法在保证对抗鲁棒性的同时显著降低了剪枝搜索和微调成本, 提升了方法的工程实用性. 围绕训练效率问题, Kundu 等人[130,131]先后提出动态网络重连接 (dynamic network rewiring, DNR/DNR++) 框架, 利用混合损失函数和动量驱动的连接重构机制, 在单次训练过程中实现高比例剪枝与鲁棒性保持. 最近的研究开始将鲁棒稀疏化与结构搜索相结合. 例如, Li 等人[132]提出基于可学习剪枝策略的对抗鲁棒神经架构搜索 (adversarial neural architecture search by the pruning policy, ANAS-P), 通过可微分二值掩码自动搜索通道宽度, 在高稀疏率条件下显著提升对抗鲁棒精度.

(3) 训练后鲁棒修剪与鲁棒性保持方法

除在训练阶段联合优化鲁棒性与稀疏结构之外, 另一类研究关注训练后修剪场景, 即在已有标准模型或已训练完成的模型基础上进行压缩, 并通过额外机制尽可能维持或恢复模型的对抗鲁棒性. Wijayanto 等人[133]系统比较了压缩模型与从头设计的小模型在对抗鲁棒性上的差异, 指出在相同参数规模下, 压缩得到的模型往往比紧凑设计模型更鲁棒. 该工作强调, 训练后修剪并非必然削弱鲁棒性, 其影响与模型结构继承性密切相关. 与此同时, Yan 等人[134]针对边缘计算场景提出了一种设备-服务器协同的鲁棒压缩框架, 在完成模型压缩后引入针对梯度型对抗攻击的防御机制, 并结合权重分布特性进行后处理调整, 体现了训练后防御思路在系统层面的可行性. 随后, Matachana 等人[135]从攻击视角系统研究了不同修剪方式下模型对通用对抗扰动 (universal adversarial perturbations, UAP) 的鲁棒性与迁移特性, 发现不同训练后修剪策略会显著改变模型的特征空间结构, 在某些场景下可以削弱黑盒迁移攻击效果, 这为训练后鲁棒性保持提供了新的评估维度. 在此基础上, Lee 等人[136]提出了一个面向安全敏感嵌入式系统的鲁棒 CNN 压缩框架, 将训练后剪枝、知识蒸馏与有限轮次的对抗微调相结合, 并采用近端梯度方法降低优化过程中的内存开销. 该工作表明, 即使不进行完整对抗训练, 通过合理的训练后补偿机制, 仍可以在较高压缩率下获得可接受的鲁棒性能.

Weiss 等人[137]针对迁移攻击提出贪婪对抗剪枝 (greedy adversarial pruning, GAP) 方法, 其核心思想是剪除在对抗梯度下贡献最大的权重, 从而破坏对抗样本在模型之间的可迁移性. 随着模型规模向语言模型和大模型扩展, 训练后鲁棒修剪的研究也开始向 NLP 场景延伸, 例如, Li 等人[138]提出了一种面向语言模型的知识保持型训练后修剪策略, 通过最小化稀疏模型与原始稠密模型在嵌入空间和特征空间中的偏差, 显式保留预训练知识, 从而提升稀疏模型对文本对抗扰动的鲁棒性. 该方法避免了重训练, 在准确率、稀疏率和鲁棒性之间取得了较好的平衡. 近期研究进一步探索了训练后修剪与集成防御思想的结合. Jung 等人[139]提出的方法通过对同一基模型采用不同训练后剪枝准则生成多个子模型, 并在推理阶段进行动态集成, 从而在保持模型整体紧凑性的同时显著提升了对抗鲁棒性. 此外, Kraidia 等人[140]提出一种面向多种攻击场景的训练后压缩防御框架, 结合权重压缩、多专家结构与随机化推理机制, 在不重新训练完整模型的前提下提升了模型对黑盒对抗攻击的抵抗能力.

(4) 鲁棒性指标与敏感度感知的修剪方法

这类工作通常强调修剪时应当依据与对抗鲁棒性直接相关的指标来评估参数/结构的重要性, 而不是沿用自然精度导向的启发式准则, 如权重幅值等. 这一方向在 Sehwag 等人[141]的研究中被系统化推进. 他们认为传统剪枝准则在对抗训练/可证鲁棒训练下会失效, 原因在于其重要性定义与鲁棒目标不一致; 因此应当让鲁棒训练目标本身来指导剪枝搜索. 这一工作也为后续大量基于鲁棒损失/梯度等的指标研究提供了范式. 随后, Wei 等人[142]提出批归一化辅助的对抗修剪, 从对抗扰动的频域特性出发, 分析 BN 的缩放因子 γ 与高频成分学习之间的关联, 提出用 γ 的幅值作为结构化剪枝的判据, 并进一步定义结合 γ 与权重幅度的"有效权重"用于非结构化剪枝. 图 11 为基于权重和有效权重实现的非结构化剪枝的比较. 对于按权重的剪枝, 模型中的所有权重被收集并排序, 如图 11(a)所示, 最小的 6 个权重将被设置为 0. 而基于有效权重的剪枝同时考虑了权重和 γ, 可以更好地度量对抗模型中各个权重的重要性 (图 11(b)). 文献 [142] 通过 BN 参数捕捉模型对高频扰动的敏感性, 从而剪掉更可能携带脆弱特征的通道/权重. 这类方法的价值在于计算开销更小, 且给出了较强的机制解释.

(a) 基于权重的非结构化修剪　　(b) 基于有效权重的非结构化修剪

图 11　基于权重和有效权重实现的非结构化修剪示意图[142]

Lee 等人[143]进一步把敏感度提升到更直接的鲁棒目标度量, 提出掩蔽对抗损伤 (masking adversarial damage, MAD) 方法, 用对抗损失的二阶信息来估计对抗显著性, 并强调参数之间的相关性在对抗场景下尤其关键. 此外, MAD 还指出初始层参数往往对对抗扰动更敏感、剪枝需要更谨慎, 这类经验结论在后续结构化鲁棒剪枝中经常被复用. 在随后的研究中, 这一类工作将鲁棒性指标用于滤波器/通道/核等结构单元, 以便获得真实推理加速. Zhuang 等人[144]直接估计每个滤波器对对抗训练损失的贡献, 进而按鲁棒贡献剪除滤波器, 并强调在极高稀疏率 (如 90%) 下保持鲁棒性是现有方法的局限. 同时, Zhong 等人[145]针对结构化剪枝方法在对抗攻击下普遍很脆弱的现象, 基于分组核剪枝 (grouped kernel pruning, GKP) 的细粒度结构优势, 引入与鲁棒性相关的核级度量 (如 kernel smoothness) 对剪枝过程进行修正, 提出训练后几乎不增加额外成本的鲁棒增强结构化剪枝方案, 并强调鲁棒提升可以在不增加训练代价的情况下通过剪枝粒度与指标选择获得.

在最近的研究中, Luo 等人[4]提出用多类型对抗样本结合输入无关扰动生成器构造混合噪声样本, 并设计特征评分模块 (feature scoring module, FSM) 与贡献差异损失, 从中间特征在干净/噪声输入下贡献一致性的角度动态评估鲁棒特征, 再通过门控机制保留鲁棒通道, 同时引入压缩控制损失以满足期望的压缩比, 整体流程如图 12 所示. Han 等人[146]把这一思路应用到通信系统中的自动调制分类 (automatic modulation classification, AMC) 任务中. 他们使用基于通道激活贡献的剪枝指标, 同时配合梯度掩蔽等策略来扩大不确定性空间、增强对多类攻击的抵抗能力.

(5) 鲁棒彩票与鲁棒子网络研究

这类研究的核心是考虑在一个过参数化的稠密网络中, 是否存在某些稀疏子网络能够在对抗攻击下同样表现出较强鲁棒性. 与前面几类方法不同, 这里更强调鲁棒性是否能作为一种子网络属性被发现、继承或在初始化时就已隐含存在. Cosentino 等人[147]进行了早期的系统性探索, 他们直接把彩票假设的迭代剪枝流程与对抗训练结

合, 提出通过“迭代训练-剪枝-重置”的方式寻找既稀疏又鲁棒的子网络. 实验结果表明, 在对抗训练条件下, 彩票假设框架能够找到在较高稀疏率下仍保持对抗鲁棒性的稀疏鲁棒网络. 同时, 该工作也强调, 保留特定初始化往往更有利于子网络性能与稳定性. 随后, Fu 等人[148]提出了更加直接的鲁棒“刮刮乐”彩票 (robust scratch ticket, RST), 声称在随机初始化的稠密网络中, 存在某些子网络可以在不更新权重, 甚至不进行模型训练的情况下, 表现出与对抗训练模型相当的鲁棒性能. 其方法本质上是固定随机权重, 仅学习一个可稀疏化的二值掩码, 使掩码搜索过程能够感知鲁棒目标, 最终得到具有先天鲁棒性的子网络. 在最近的研究中, Shi 等人[149]在上述发现的基础上, 将研究焦点从经验现象推进到可解释的搜索准则. 他们基于神经正切核 (neural tangent kernel, NTK) 相关理论, 分析对抗训练的动力学特征, 并提出对抗性中奖彩票 (adversarial winning ticket, AWT), 即在初始化阶段寻找能够在动力学层面接近稠密对抗训练过程的稀疏子网络, 使其可以“从头对抗训练”并达到与稠密对抗训练相当的性能, 为对抗鲁棒的彩票提供了更强的理论与方法学支撑.

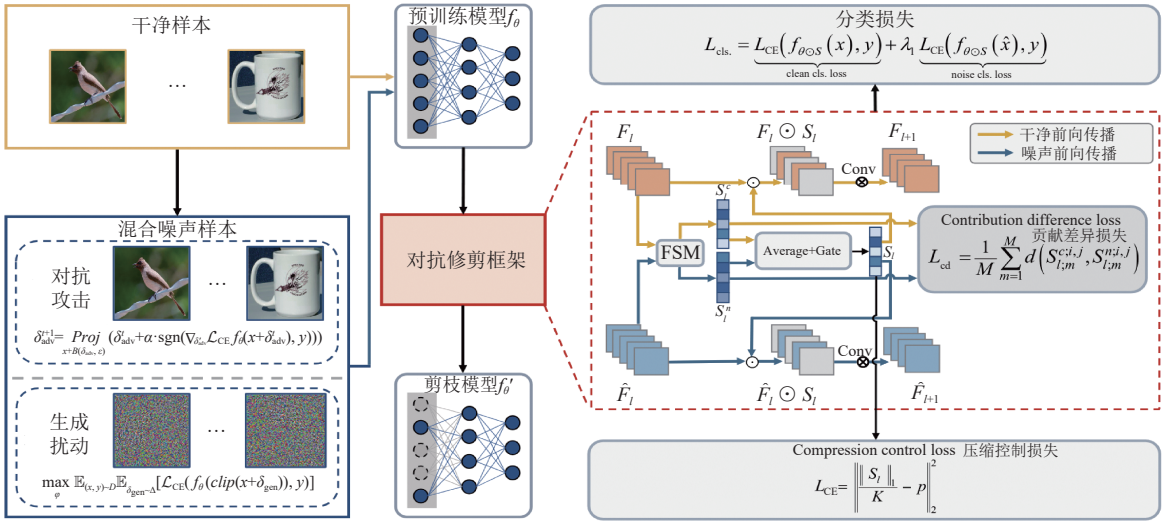


图 12 对抗性修剪方案示意图[4]

5.1.2 对抗鲁棒模型量化

模型量化已被广泛应用于资源受限的部署场景, 然而, 量化误差可能改变模型的数值稳定性和梯度特性, 从而影响其对输入扰动的敏感性; 同时, 数值离散化也可能使模型呈现出一定程度的鲁棒性变化. 现有研究表明, 量化是否提升对抗鲁棒性依赖于量化精度、策略以及攻击和评测设置, 相关结论并不一致. 因此, 近年来的工作逐步从现象分析转向在量化训练过程中显式引入鲁棒优化目标, 并通过系统化评测澄清不同方法和设置下的鲁棒性表现. 基于此, 本文从量化影响分析、鲁棒量化训练方法以及鲁棒性评测这 3 个方面对相关工作进行综述.

(1) 量化对对抗鲁棒性的影响分析

Liu 等人[150]较早研究了包括权重量化与哈希压缩在内的方法对对抗攻击的影响, 实验发现在部分设置下, 压缩模型相较于全精度模型对攻击略不敏感, 但这种现象难以作为普遍结论. 随后, Lin 等人[151]系统分析了低比特量化模型在对抗攻击下的行为, 发现常规量化在保持干净精度的同时, 往往显著削弱模型的对抗鲁棒性. 作者将这一现象归因于误差放大效应, 即量化引入的离散化误差会在深层网络中被逐层放大, 使得对抗扰动在中间特征空间中跨越量化区间, 从而加剧分类错误. Zhao 等人[115]进一步从攻击迁移性的角度研究了量化模型的对抗鲁棒性, 发现针对全精度模型生成的对抗样本在量化模型上仍具有较高的可迁移性. 尤其是在中等比特宽度设置下, 量化仅对梯度型攻击产生有限的削弱作用, 而无法从根本上阻断攻击迁移. 此外, Bernhard 等人[152]对低比特量化模型在不同攻击强度下的鲁棒性进行了系统评估, 指出当比特宽度持续降低时, 模型的对抗精度通常呈现下降趋势, 且这一趋势在强攻击设置下尤为明显. 该研究强调, 部分早期观察到的低比特更鲁棒现象, 往往源于评测攻击不足或梯

度掩蔽, 而非真实的防御能力.

Gorsline 等人[153]进一步通过统一攻击设置对量化神经网络进行了大规模评测, 发现量化并不会稳定提升模型的白盒或黑盒鲁棒性. 相反, 在多种标准攻击下, 量化模型往往表现出更大的性能波动, 其鲁棒性对量化精度和网络结构高度敏感. 最近, Li 等人[154]对现有研究中结论分歧的来源进行了系统分析, 指出不同工作在量化流程(PTQ 或 QAT)、是否结合鲁棒训练、量化阶段位置以及攻击强度选择上的差异, 是导致结论不一致的根本原因. 该工作表明, 在不引入鲁棒优化的情况下, 单独调整量化精度难以预测其对对抗鲁棒性的影响, 量化本身并不能被视为一种可靠的防御机制.

(2) 基于鲁棒优化的量化训练方法

这一研究方向将量化视为一种结构性或数值扰动, 并在训练阶段主动约束模型对该类扰动的敏感性, 从而同时兼顾效率与鲁棒性. Gui 等人[124]提出了统一的对抗鲁棒模型压缩框架, 在一个联合优化问题中同时引入对抗训练目标和量化约束. 在该框架中, 量化被纳入对抗鲁棒优化的约束集合, 与修剪和低秩分解共同作用. 实验结果表明, 在引入对抗训练的情况下, 量化模型能够在保持较高压缩率的同时显著提升对抗鲁棒性. 随后, Sen 等人[155]提出混合精度深度网络鲁棒集成方法 (ensembles of mixed precision deep networks for increased robustness, EMPIR), 如图 13 所示, EMPIR 模型由 M 个全精度 (full-precision, FP) 模型和 N 个低精度 (low-precision, LP) 模型组成. 全精度模型有助于提高整体模型的无扰动精度, 而低精度模型有助于提高稳健性. 所有的单个模型的输入相同, 并且它们的预测类别或概率在最后借助于集成技术被组合, 以确定 EMPIR 模型的最终预测. EMPIR 的低精度与全精度模型在对抗扰动下具有互补特性. 通过在训练与推理阶段协同利用不同量化精度, EMPIR 在不显著增加计算和存储开销的前提下, 提高了整体系统的对抗鲁棒性.

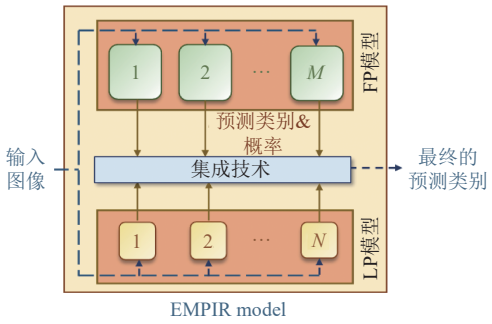


图 13 EMPIR 方案示意图[155]

在后续研究中, Song 等人[156]指出直接将标准对抗训练应用于量化模型往往会导致显著的量化损失. 为此, 他们提出了一种边界感知的重训练方法, 在训练过程中同时约束对抗扰动和量化误差的放大效应. 该方法在权重量化模型上有效恢复了对抗鲁棒性, 表明鲁棒训练策略需要针对量化特性进行专门设计. Yang 等人[157]则从理论层面对鲁棒优化与量化性能之间的关系进行了深入分析, 并在训练阶段通过正则化 Hessian 特征值来显式提升模型对参数扰动的稳定性, 从而在多种后训练量化精度下均表现出更稳定的性能. 在此基础上, Yang[122]系统地总结了上述研究思路, 并进一步从优化理论角度阐明了鲁棒性、泛化能力与量化稳定性之间的统一关系. 针对极低比特甚至深度量化场景, Ayaz 等人[158]提出在量化感知训练过程中引入 Jacobian 正则化, 以显式约束模型对输入扰动的敏感性. 该工作表明, 即使在极端低精度条件下, 合理的鲁棒优化仍可显著缓解量化带来的脆弱性. 在最新的研究中, Zakariyya 等人[159]对基于量化感知训练和 Jacobian 正则化的模型进行了定量分析, 验证了其在多种数据集和攻击场景下的鲁棒性优势. 该工作从实验层面确认了鲁棒优化在深度量化模型中的实际有效性, 同时指出其性能仍受限于模型结构和攻击类型.

(3) 量化模型对抗鲁棒性的评测与基准研究

随着量化模型在对抗鲁棒性研究中的结论逐渐呈现出明显分歧, 有研究者发现, 评测协议本身可能是导致结

论不一致的重要原因. 因此, 一类工作开始系统分析量化模型在不同攻击设置和应用场景下的鲁棒表现, 旨在澄清量化对对抗鲁棒性的真实影响. Wijayanto 等人[133]较早从整体视角系统评估了包括量化在内的压缩模型在白盒与黑盒攻击下的鲁棒性表现. 该工作强调, 在相同攻击强度下, 量化模型的鲁棒性高度依赖具体网络结构和攻击方式, 仅凭模型尺寸或精度变化难以得出可靠结论. Matachana 等人[135]则进一步从通用对抗扰动的角度分析了量化模型的鲁棒性与攻击迁移特性. 实验表明, 在部分设置下, 量化可能引入梯度掩蔽现象, 使模型在特定白盒攻击下表现出表面鲁棒, 但在迁移攻击和更强攻击设置下仍然高度脆弱. 作者明确指出, 仅基于单一攻击或弱攻击评测量化模型, 容易高估其安全性. Fukuda 等人[160]则将评测重点放在工业部署场景中, 系统评估了主流工具链 (如 Vitis-AI) 下量化模型对对抗样本的敏感性. 研究发现, 即使在相同比特宽度下, 不同部署环境中的鲁棒性结论也可能不一致. 这一结果表明, 量化鲁棒性的评测需要考虑实际部署条件, 不可仅停留在算法层面.

针对现有评测工作中攻击类型单一、数据集规模有限等问题, Xiao 等人[161]提出了系统性的量化鲁棒性评测框架, 对不同量化方法、比特宽度和网络结构在多种扰动下的表现进行了大规模对比. 研究发现, 低比特量化模型在某些对抗攻击下表现出更高的鲁棒性, 但同时对自然扰动和系统噪声更加敏感, 揭示了量化鲁棒性的多维权衡关系. 在此基础上, Xiao 等人[162]进一步构建了量化模型的鲁棒性基准, 对量化模型在对抗攻击、自然扰动和系统噪声这 3 类场景下进行统一评测. 该工作明确指出, 量化模型的鲁棒性不能仅以对抗精度衡量, 而应结合多种扰动源进行综合评估.

5.1.3 对抗鲁棒知识蒸馏

知识蒸馏通过在训练过程中利用教师模型的软监督信息, 引导学生模型学习更具判别性的表示. 因此, 其对对抗鲁棒性的作用机制更加间接, 也更依赖于教师模型属性和蒸馏策略的设计. 已有研究表明, 鲁棒性是否能够从教师传递到学生模型, 取决于教师的鲁棒属性、蒸馏信号的形式以及训练中的权衡机制. 因此, 近年来的工作从鲁棒教师蒸馏、蒸馏信号增强以及多教师与任务感知蒸馏等角度, 对对抗鲁棒知识蒸馏进行了系统探索.

(1) 基于鲁棒教师的对抗蒸馏方法

这类方法的核心思想是利用经过对抗训练的教师模型, 将其鲁棒决策边界或鲁棒预测行为通过蒸馏过程传递给结构更紧凑的学生模型. Arani 等人[163]较早探索了在知识蒸馏框架中引入鲁棒性因素的可能性, 通过在教师-学生协同训练过程中引入噪声和不确定性建模, 观察到学生模型在泛化性和一定程度上的鲁棒性方面均有所改善. 在此基础上, Goldblum 等人[164]提出了对抗鲁棒蒸馏 (adversarially robust distillation, ARD) 框架, 系统研究了鲁棒性在蒸馏过程中的可传递性问题, 指出即使仅使用干净样本进行蒸馏, 学生模型仍可在一定程度上继承鲁棒教师的对抗鲁棒性; 此外, 通过在蒸馏阶段引入对抗扰动, ARD 所得到的学生模型在相同架构下可在鲁棒性与自然精度上同时优于直接对抗训练的模型, 从而奠定了鲁棒教师蒸馏这一研究范式的基础. 随后, Shao 等人[165]指出, 标准知识蒸馏往往难以保持教师模型的鲁棒性, 并将原因归结为学生模型在梯度层面的行为偏差. 为此, 他们提出通过输入梯度对齐来约束学生模型, 使其在局部扰动下的响应与教师保持一致, 并在一定假设条件下给出了学生模型鲁棒性不劣于教师模型的理论保证. Li 等人[166]在之后的研究中进一步对基于鲁棒教师的蒸馏方法进行了经验性反思, 系统分析了教师模型的鲁棒精度、容量规模、输出分布以及蒸馏损失形式等因素对学生鲁棒性的影响. 其研究表明, 鲁棒教师并非越强越好, 教师与学生之间的容量匹配以及蒸馏信号的可靠性对鲁棒性传递起着关键作用.

(2) 蒸馏信号增强与鲁棒信息重构方法

这一类方法更关注的是鲁棒性在蒸馏过程中通过什么信号被传递以及如何改造这些信号以避免信息失真. Zi 等人[167]首先从蒸馏视角重新审视多种对抗训练鲁棒蒸馏目标, 指出鲁棒软标签是提升学生鲁棒性的关键因素, 并据此提出鲁棒软标签对抗蒸馏 (robust soft label adversarial distillation, RSLAD): 用鲁棒教师在干净样本上的预测分布替代硬标签, 在外层与内层优化中都以 KL 对齐的方式引导学生学习, 从而让学生在对抗样本上更稳定地模仿教师的鲁棒行为. 该工作明确了蒸馏信号不应仅是软标签, 而应是来自鲁棒教师的、携带决策边界信息的概率分布, 并强调其在生成对抗样本中的作用. 在此基础上, Huang 等人[168]提出自适应对抗蒸馏 (adaptive adversarial distillation, AdaAD): 在内层优化中不再固定使用教师在原始样本点的输出, 而是让教师参与内层过程, 通过最大化师生预测差异来搜索“支持点”, 再在外层最小化该差异, 实现更具代表性的对齐. RSLAD 与 AdaAD 方法的对比

示意如图 14 所示. AdaAD 强调监督信号应与对抗邻域中的关键样本绑定, 其本质是对蒸馏信号的动态构造与自适应选择, 从而提升学生继承鲁棒性的效率与稳定性.

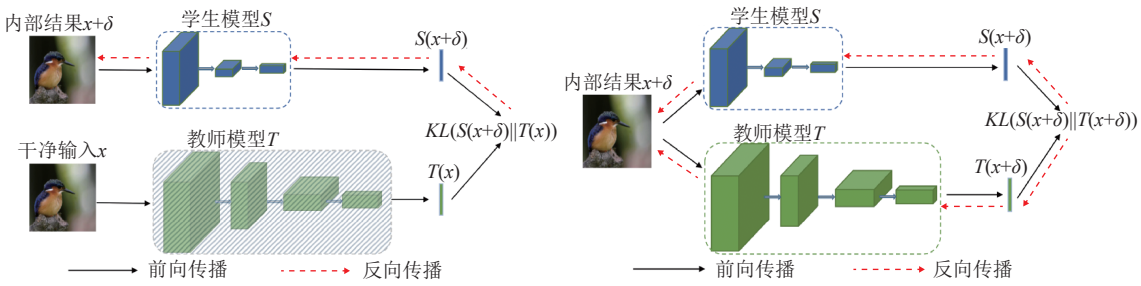


图 14 RSLAD 和 AdaAD 的内部优化对比示意图[168]

此外, Yue 等人[169]从鲁棒公平性角度指出: 既有鲁棒蒸馏方法往往只优化整体鲁棒精度, 可能导致不同类别鲁棒性差异增大, 出现“短板效应”. 为此, 他们提出公平对抗鲁棒蒸馏方案: 通过度量类别难度并进行类别级权重加权, 让蒸馏训练更关注困难类别, 从而显著改善最弱类别鲁棒性. 最近, Wu 等人[170]进一步指出, 主流鲁棒蒸馏过度依赖 logits 对齐, 且教师输出中不同类别的知识高度耦合, 噪声下更难提取对鲁棒分类真正关键的部分. 为此, 他们提出混合分解蒸馏方法: 在 logits 侧将教师预测拆分为目标相关与非目标相关知识并赋予不同权重; 在特征侧根据通道相关性将中间表示分解为更重要与次要成分, 进行选择性特征蒸馏, 实现 logits 与特征的双层重构与对齐. 与前述工作相比, 该方法能够更有效地使学生在对抗扰动下学习到更稳健的表征.

(3) 多教师与任务感知的对抗鲁棒蒸馏

为缓解对抗训练中普遍存在的准确率-鲁棒性权衡问题, 部分研究引入多教师或任务感知机制, 通过协同蒸馏不同类型的知识来增强学生模型的整体鲁棒性. Lee 等人[136]提出了一种面向安全敏感嵌入式系统的鲁棒模型压缩框架, 将对抗训练、模型修剪与知识蒸馏统一到同一优化目标中, 通过利用预训练教师模型在干净样本上的预测信息, 稳定压缩模型在对抗训练过程中的决策边界, 从而在保证模型紧凑性的同时提升对抗鲁棒性. 在更具体的任务场景中, Levi 等人[171]针对目标检测模型提出了结合对抗性调整的知识蒸馏方法, 通过构造教师-学生对并引入任务感知的多分量蒸馏损失, 使学生模型在面对对抗补丁时的预测行为与教师在干净输入上的输出保持一致. 进一步地, Zhao 等人[172]系统研究了多教师对抗蒸馏在缓解准确率-鲁棒性权衡中的潜力, 提出了平衡多教师对抗蒸馏框架, 同时引入干净教师和鲁棒教师并分别指导学生模型在自然样本与对抗样本上的学习. 该方法通过熵约束与损失归一化机制, 显式平衡不同教师之间的知识尺度与学习速度, 使学生模型能够在保持干净精度的同时获得更强的对抗鲁棒性.

5.1.4 对抗鲁棒低秩分解

低秩分解通过将高维权重矩阵分解为低秩因子, 在显著降低参数规模和计算复杂度的同时, 也会对模型的对抗鲁棒性产生直接影响. 已有研究注意到, 直接采用低秩近似可能在压缩过程中改变模型的谱结构和决策几何, 从而影响模型对输入扰动的敏感性. Gui 等人[124]较早在统一的对抗鲁棒模型压缩框架中系统引入低秩分解, 并将其与对抗训练联合建模. 实验结果表明, 在合理的秩设置下, 低秩分解并不会削弱鲁棒性, 反而可以与对抗训练形成互补, 为低秩压缩模型实现较好的效率-鲁棒性权衡提供了早期实证依据. 在此基础上, Yang[122]探讨了低秩分解对模型稳定性的影响并指出, 低秩近似改变了权重矩阵的奇异值分布, 不恰当的秩截断可能放大对扰动高度敏感的方向. 为此, 他们提出通过对低秩因子施加正交性与谱约束, 抑制不稳定方向, 从而在降低模型复杂度的同时提升模型对权重扰动和输入扰动的整体鲁棒性. 随后, Schotthöfer 等人[173]进一步从条件数和谱正则化角度对低秩模型的对抗鲁棒性进行了深入分析, 提出在动态低秩训练过程中引入谱正则项, 以显式控制低秩核心矩阵的奇异值分布, 使压缩模型在多种对抗攻击下保持更稳定的鲁棒性能. 最近, 面向大模型场景, Asante 等人[174]的研究结果表

明, 不同低秩分解策略在深度压缩条件下对对抗输入的敏感性差异显著, 但整体趋势显示, 经过合理设计的低秩压缩模型能够在维持模型效率优势的同时, 保持与全参数模型相当的鲁棒性表现. 这表明低秩分解在大模型鲁棒压缩中具有实际应用潜力.

5.2 隐私保护模型压缩

随着深度学习模型在隐私敏感场景中的广泛部署, 模型压缩不仅需要关注计算效率和存储开销, 还逐渐被寄予缓解隐私泄露风险的期望. 相较于对抗鲁棒性, 隐私风险更多源于模型对训练数据的记忆行为及其可被外部推断的程度, 而压缩操作通过改变模型容量、表示能力和训练过程, 可能对隐私泄露产生复杂影响. 因此, 近年来的工作开始从隐私保护视角重新审视知识蒸馏、模型修剪和模型量化等典型压缩技术, 探索其在成员推理等隐私攻击场景中的潜在优势与局限.

5.2.1 隐私保护知识蒸馏

知识蒸馏的安全性取决于蒸馏过程中信息共享的形式及所引入的隐私约束. 围绕隐私保护目标, 并基于蒸馏在隐私保护中所扮演的功能角色, 现有隐私保护知识蒸馏方法可归纳为两类: 一类方法面向知识迁移过程中的隐私约束, 通过差分隐私、样本筛选等方式限制蒸馏信号中的敏感信息传播; 另一类方法则主要面向联邦学习或多方协作场景, 从系统设计层面缩小隐私暴露面, 必要时辅以差分隐私机制进行增强. 下面将分别对这两类研究进行介绍.

(1) 面向隐私约束的知识蒸馏方法

这类方法的核心思想是在教师-学生蒸馏过程中限制敏感信息随蒸馏信号传播, 从而在压缩模型规模的同时降低训练数据隐私泄露风险. 一部分研究以差分隐私为形式化约束, 通过对教师模型输出或中间知识进行裁剪、扰动或聚合, 为蒸馏过程提供可证明的隐私保障. 例如, Wang 等人[175]提出隐私模型压缩框架, 该方法通过对教师模型输出的中间知识进行裁剪与噪声注入, 在蒸馏阶段满足 (ε, δ)-DP 约束. 在保证严格隐私预算的前提下, 该方法仍能获得较高的模型压缩率与较小的性能损失. 在此基础上, Gao 等人[176]进一步在教师预测结果上引入差分隐私噪声, 并分析隐私预算、蒸馏温度与模型性能之间的权衡关系, 强调对蒸馏信号进行精细控制的重要性. 与此同时, Lyu 等人[177]提出了差分隐私知识蒸馏框架, 将传统的教师集成隐私聚合 (private aggregation of teacher ensembles, PATE) 思想与模型蒸馏深度融合, 构建端到端差分隐私蒸馏框架, 使隐私预算能够在蒸馏过程中被系统追踪. Mireshghallah 等人[178]随后比较了不同压缩手段 (包括蒸馏、剪枝与量化) 在差分隐私训练下的表现, 指出蒸馏在隐私-效用权衡上通常优于直接参数扰动的方法, 并分析了噪声注入位置对隐私泄露风险的影响, 为 DP 蒸馏方法的设计提供了更统一的视角. Flemings 等人[179]进一步利用经 DP-SGD 微调的教师 LLM 生成差分隐私合成文本, 并结合合成文本标签与教师输出分布训练学生模型, 从而避免在学生蒸馏阶段再次执行 DP-SGD, 缓解差分隐私与模型压缩叠加带来的效用退化. Wang 等人[5]则从理论角度将蒸馏视为隐私安全的数据生成机制, 通过教师模型生成隐私保护的辅助数据以训练学生模型, 并给出收敛性与误差界分析. 除差分隐私噪声机制外, 近期也有研究从蒸馏过程本身出发设计隐私保护策略. Cui 等人[107]面向大语言模型知识蒸馏中的成员推理风险, 提出基于样本筛选和学生模型结构改造的隐私抑制方法. 该工作表明, 隐私保护蒸馏不仅可以依赖差分隐私噪声实现, 也可以通过蒸馏数据选择与学生结构设计降低成员信息泄露风险.

(2) 面向联邦与多方协作场景的隐私蒸馏方法

在联邦学习和多方协作场景中, 隐私风险不仅来源于模型对单个样本的记忆, 还与模型参数、梯度或预测在多轮通信中的反复暴露密切相关. 为此, 这类研究将知识蒸馏视为替代参数或梯度共享的隐私友好通信机制, 通过仅交换模型在公共数据上的预测结果或中间知识, 从系统层面缩小隐私暴露面. 例如, Gong 等人[180]提出的隐私保护联邦学习方法仅在公共数据上交换各本地模型的预测结果和注意力信息, 从而在设计上减少对私有数据分布的直接暴露. 通过集成注意力蒸馏, 该方法能够捕获不同本地模型的结构性知识, 在提升全局模型性能的同时, 降低了通信频率和潜在隐私泄露风险, 整体框架如图 15 所示. 随后, Sun 等人[181]指出直接对预测结果添加噪声会导致明显的性能退化. 为此, 他们提出了具有无噪声差分隐私的联邦模型蒸馏, 利用随机采样机制使模型预测在不显式注入噪声的情况下满足 (ε, δ)-DP.

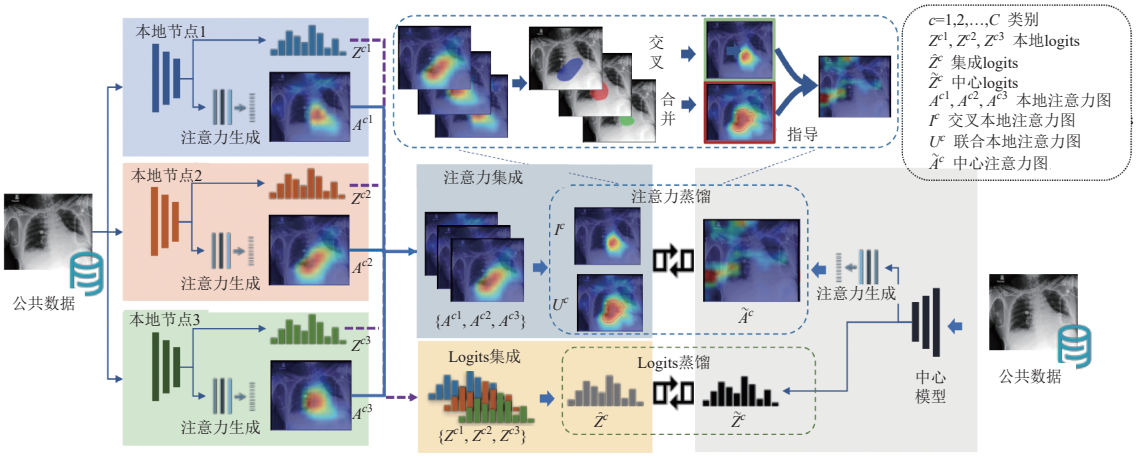


图 15 基于集成注意力蒸馏的隐私保护联邦学习框架示意图[180]

Huang 等人[182]则提出基于知识蒸馏的隐私保护数据分析框架, 通过让各参与方独立训练教师模型, 并在公共无标注数据上生成受限数量的伪标签, 再由学生模型进行学习, 避免了对私有数据或模型参数的直接共享. 实验结果表明, 在不引入复杂密码学机制的情况下, 蒸馏式协作能够在一定程度上兼顾隐私保护与模型性能. 在此基础上, Qi 等人[183]提出了隐私知识迁移 (private knowledge transfer, PrivateKT) 方法, 从知识迁移效率与隐私保护的角度进一步优化联邦蒸馏流程. PrivateKT 通过主动选择少量高信息量公共样本作为知识载体, 并在预测层面结合局部差分隐私机制, 实现了在严格隐私预算下的高效知识迁移. 最近, Wei 等人[184]对 Transformer 模型提出了隐私蒸馏框架, 分析了大模型蒸馏过程中可能引发的数据重构与反向推理风险. 该方法将知识蒸馏与差分隐私和安全多方计算相结合, 在多方协作场景下对蒸馏过程中的信息泄露进行系统防护. 该工作进一步表明, 在大模型与复杂协作环境中, 仅依赖单一隐私机制往往不足, 而蒸馏可作为连接多种隐私技术的关键中间层.

5.2.2 隐私保护模型修剪

模型修剪通过减少参数规模和模型容量, 被认为可能在一定程度上削弱模型对训练数据的记忆能力, 从而缓解隐私泄露风险. 近年来, 部分工作开始系统研究修剪与隐私保护之间的关系, 并尝试将修剪过程与差分隐私或安全目标相结合, 以实现兼顾模型压缩与隐私保护的训练框架. 早期研究主要从系统和部署角度探索隐私感知的修剪策略. 例如, Yan 等人[185]提出一种面向边缘计算场景的隐私保护压缩模型构建方法, 在云端通过差分隐私机制训练稠密模型后, 再对模型进行隐私感知的修剪与再训练, 从而生成可部署在边缘设备上的小规模模型. Mireshghallah 等人[178]则从更系统的角度研究了差分隐私约束下的模型压缩问题, 提出了将差分隐私训练与修剪过程协同设计的框架, 其将差分隐私随机梯度下降与迭代修剪相结合, 实验证明在中等稀疏率下可以获得性能接近原模型的私有稀疏子网络. 在联邦学习与分布式场景中, Lin 等人[186]通过在客户端侧进行稀疏化和安全聚合, 减少了模型更新中可能泄露的隐私信息. 该方法表明, 模型修剪能够在一定程度上缩小攻击者可利用的模型暴露面, 提升隐私安全性. 此外, Zhu 等人[187]从安全-性能协同优化的视角研究了模型修剪在隐私攻击场景中的作用, 提出了将成员推理攻击作为安全测试信号嵌入修剪流程的双目标优化框架. 该方法有效降低了成员推理攻击成功率, 表明隐私感知修剪需要在压缩阶段主动建模攻击风险, 而非依赖修剪带来的被动隐私收益.

5.2.3 隐私保护模型量化

相较于知识蒸馏与模型修剪, 围绕模型量化与隐私保护之间关系的研究仍然较为有限, 现有工作主要集中在差分隐私训练与量化部署的协同设计, 以及量化对模型隐私与可信属性的影响分析. 例如, Chu 等人[188]提出在差分隐私约束下联合优化剪枝与量化策略的方法. 该工作通过在模型更新与压缩阶段引入隐私预算分配机制, 使量化后的模型在显著降低通信与存储开销的同时, 维持可控的隐私泄露风险, 为隐私感知的模型量化提供了系统化设计思路. Deng 等人[189]进一步从大语言模型微调与部署的角度出发, 提出了强化 LLM 微调 (hardening LLM fine-

tuning, HardLLM) 框架, 系统分析了量化对隐私、鲁棒性等多维可信属性的影响. 研究表明, 若量化过程未加控制, 可能对模型的隐私与安全属性产生负面影响; HardLLM 框架在保持较高压缩率的同时, 能够有效缓解成员推理等隐私风险, 并在严格隐私预算下实现接近无损的性能表现. 这一工作明确指出, 隐私保护不应仅停留在训练阶段, 量化等部署级压缩同样需要纳入隐私威胁模型进行系统设计.

5.3 小　结

本节围绕对抗鲁棒性与隐私保护两个安全目标, 系统综述了不同模型压缩方法在安全约束下的研究进展. 表 3 和表 4 对相关工作进行了总结, 反映了已有方法的技术分布, 同时也揭示了当前安全模型压缩研究中的共性规律与局限性. 如表 3 所示, 对抗鲁棒模型压缩研究相对更成熟, 其中模型修剪相关方法最为丰富, 已形成从理论分析、训练期联合优化到训练后鲁棒性保持的较完整研究脉络; 相比之下, 模型量化在鲁棒场景下仍以影响分析与评测研究为主, 稳定有效的方法相对较少, 而知识蒸馏和低秩分解通常需要在训练阶段显式引入鲁棒目标. 如表 4 所示, 隐私保护模型压缩则主要集中于知识蒸馏和模型修剪, 其中蒸馏方法在联邦与多方协作场景中较为突出, 模型量化相关研究相对有限, 而低秩分解在隐私保护方面仍缺乏系统探索. 总体而言, 不同安全目标下各类压缩方法的发展并不均衡, 这表明未来仍需从统一视角深入理解压缩操作对安全属性的影响机理, 并进一步补足尚未充分展开的安全-压缩研究组合.

表 3　对抗鲁棒模型压缩相关文献分类总结

压缩手段	方法类别	核心设计思想	主要引入阶段	是否依赖对抗训练	主要特点	主要局限
模型修剪/稀疏化	理论分析与可证鲁棒稀疏化方法[113–122]	从谱性质、条件数等角度分析稀疏性对鲁棒性的影响, 避免病态权重分布	训练期/训练后	部分	揭示稀疏性与鲁棒性之间的内在关系, 为方法设计提供理论依据	可扩展性和在大规模模型上的验证相对不足
	对抗训练驱动的鲁棒稀疏化方法[123–132]	在对抗训练中联合优化稀疏约束, 使鲁棒性与压缩相协同	训练期	是	训练阶段显式兼顾稀疏性与鲁棒性, 通常比训练后直接修剪更稳定	训练成本较高, 且压缩率与鲁棒性之间仍存在明显权衡
	训练后鲁棒修剪与鲁棒性保持方法[133–140]	在已有模型上进行训练后修剪, 并通过补偿、蒸馏或集成恢复鲁棒性	训练后	部分	工程灵活性较强, 适合资源受限等场景	鲁棒性提升通常依赖额外补偿机制, 整体效果不如联合优化稳定
	鲁棒性指标与敏感度感知的修剪方法[4,141–146]	利用对抗损失、梯度、Hessian或特征敏感度指导修剪, 保留鲁棒相关结构	训练期/训练后	部分	直接针对鲁棒相关结构进行保留, 且与结构化修剪结合较紧密	指标设计对攻击形式和模型结构较为敏感
	鲁棒彩票与鲁棒子网络研究[147–149]	在稠密网络中寻找先天或可训练的鲁棒稀疏子网络	初始化/训练期	是	有助于理解子网络的鲁棒属性, 理论探索意义较强	稳定性和实际部署价值仍有待进一步验证
模型量化	量化对对抗鲁棒性的影响分析[118,150–154]	分析低比特量化对梯度、数值稳定性和攻击迁移性的影响	—	否	澄清鲁棒性与量化关系中的影响因素, 为后续方法提供认识基础	主要停留在现象分析层面, 难以直接形成稳定有效的鲁棒量化方法
	基于鲁棒优化的量化训练方法[122,124,155–159]	在量化感知训练中引入对抗鲁棒或参数鲁棒目标	训练期	是	在低精度约束下显式提升鲁棒性, 是目前实现鲁棒量化的主要路径	训练复杂度较高, 在不同量化设置下稳定性仍有限
	量化模型对抗鲁棒性的评测与基准研究[133,135,160–162]	系统评估不同量化设置、攻击与部署条件下的鲁棒性	—	否	统一评测协议并解释现有结论分歧, 对部署场景具有参考价值	评测结论受工具链和威胁模型影响较大
知识蒸馏	基于鲁棒教师的对抗蒸馏方法[163–166]	通过鲁棒教师将对抗鲁棒性迁移至学生模型	训练期	是	在精度与鲁棒性之间具有较好的平衡潜力	鲁棒性传递效果受教师质量和蒸馏目标影响较大
	蒸馏信号增强与鲁棒信息重构方法[167–170]	对鲁棒蒸馏信号进行分解、重加权或自适应构造	训练期	是	关注鲁棒知识本身的传递形式, 提高了鲁棒蒸馏的针对性和有效性	方法设计较复杂, 增加了训练成本和实现难度

表 3 对抗鲁棒模型压缩相关文献分类总结 (续)

压缩手段	方法类别	核心设计思想	主要引入阶段	是否依赖对抗训练	主要特点	主要局限
知识蒸馏	多教师与任务感知的对抗鲁棒蒸馏[136,171,172]	利用多教师或任务特定蒸馏缓解精度-鲁棒性权衡	训练期	是	适合复杂任务和高性能需求场景, 能灵活地平衡自然精度与鲁棒性	依赖多教师模型或任务定制设计, 训练与部署开销较大
低秩分解	低秩约束与鲁棒优化方法[122,124,173,174]	通过低秩约束与谱正则化控制扰动放大方向	训练期	是	在效率-鲁棒性权衡上具有一定潜力	相关工作数量较少, 方法体系尚不成熟

表 4 隐私保护模型压缩相关文献分类总结

压缩手段	方法类别	核心隐私机制	主要引入阶段	是否提供形式化DP	主要特点	主要局限
知识蒸馏	面向隐私约束的知识蒸馏方法[5,107,175–179]	对蒸馏信号或训练样本施加隐私约束, 抑制教师数据敏感信息向学生模型迁移	蒸馏阶段	部分	兼顾模型轻量化与隐私保护需求	对隐私预算、噪声强度、数据筛选策略等较敏感, 隐私-效用权衡明显
	面向联邦与多方协作场景的隐私蒸馏方法[180–184]	以蒸馏替代参数/梯度共享, 缩小攻击面	通信阶段	部分	从系统设计层面减少隐私暴露并降低通信开销	依赖公共数据或共享预测质量, 在复杂场景下需结合其他隐私机制
模型修剪	隐私约束驱动的模型修剪方法[178,185–187]	DP训练结合稀疏化, 或以隐私攻击为约束目标	训练期/压缩期	否	协同优化模型压缩与隐私约束	相关研究数量较少, 方法体系不够成熟
模型量化	隐私约束与量化协同优化方法[188,189]	DP训练与可信量化协同设计	训练期+部署期	是	在降低通信与存储开销的同时控制隐私泄露风险	现有研究有限, 尚缺乏系统化的隐私量化方法体系

6 挑战与未来研究展望

综合前文对压缩模型安全风险以及安全感知模型压缩方法的系统梳理可以看出, 模型压缩与安全、隐私之间并非简单的正向或负向关系. 压缩操作在提升模型部署效率的同时, 也会改变模型结构、参数分布、表示空间和训练动态, 从而影响模型的鲁棒性、隐私属性以及实际部署阶段的攻击面. 不同压缩策略、压缩强度、训练流程和威胁假设可能导致完全不同的安全结果, 甚至使压缩过程本身成为攻击触发条件或风险放大环节. 因此, 当前研究亟需进一步从评估规范、攻击防护、机理解释以及新兴部署场景等方面进行深入研究.

6.1 面向压缩模型安全性的基准化评估框架构建

当前压缩模型安全研究仍缺乏统一和可比较的评估框架. 已有工作在威胁模型、压缩方法、攻击设置和评价指标等方面差异较大, 导致不同研究结论难以直接比较. 更重要的是, 模型压缩安全性并不只取决于攻击算法本身, 还与压缩发生的阶段、攻击者是否了解压缩流程以及最终部署环境等密切相关. 如果实验中未明确区分原始模型、压缩中模型和压缩后模型, 或未对压缩前后的安全变化进行充分对照, 就很难判断安全风险究竟是由压缩操作引入、放大或缓解, 还是由模型结构、攻击强度等其他因素造成. 因此, 围绕压缩过程建立可比较、可复现的评估规范, 是理解压缩模型安全性的基础问题.

未来研究有必要推动面向压缩模型安全性的基准化评估. 基准化评估并不要求所有研究必须采用完全相同的模型、压缩方法或攻击算法, 而是应在关键实验要素上形成清晰和可报告的基本规范. 具体而言, 后续研究应更加关注压缩过程在安全评估中的作用, 将压缩前后的模型变化、攻击假设和部署条件纳入同一分析环节中, 从而提高不同工作之间的可比性, 帮助研究者更准确地判断安全变化的来源. 后文表 5 给出了建议的面向压缩模型安全性的评估维度. 需要指出的是, 表 5 中内容并非所有可能的评估因素, 而是结合当前研究现状归纳出的若干可供参考的评估维度, 后续仍可根据具体任务、威胁模型和部署场景进一步补充与细化.

6.2 压缩触发型攻击的检测与缓解

压缩触发型攻击是压缩模型安全研究中需要重点关注的问题. 这类攻击的隐蔽性更强: 模型在原始未压缩状

态下可能表现正常, 甚至能够通过常规安全检测, 但在经过修剪、量化等压缩操作后, 恶意行为才被激活. 例如, 前文介绍的压缩伪装后门攻击表明, 攻击者可以在完整模型中植入未压缩状态下不可激活的后门, 而当模型经过修剪后, 该后门反而被激活并产生明显攻击行为. 其危险之处在于, 实际部署流程中压缩往往被视为模型优化步骤, 安全审查通常集中在原始模型或训练过程, 而压缩后的最终部署模型未必会被系统检测. 因此, 压缩触发型攻击容易绕过传统检测流程, 并在模型真正上线后暴露风险. 该问题的难点在于, 攻击行为通常与压缩误差、参数扰动、剪枝保留结构或蒸馏过程相互耦合, 导致风险来源难以直接定位.

表 5　面向压缩模型安全性的基准化评估维度建议

评估维度	建议报告内容	关注重点
压缩对象与压缩流程	原始模型、压缩后模型、压缩方法、压缩顺序、是否经过微调或再训练	明确安全变化发生在训练前、压缩中还是压缩后, 避免只比较最终结果
压缩强度与关键参数	剪枝率、剪枝粒度、量化位宽、量化方式、蒸馏温度、学生模型规模、低秩秩值	分析安全风险随压缩强度变化的趋势, 而不是只报告单一压缩配置
压缩前后安全对照	原始模型与压缩模型的干净精度、攻击成功率、鲁棒精度、隐私攻击指标对比	判断压缩是否引入、放大或缓解安全/隐私风险
攻击者对压缩过程的知识	攻击者是否知道压缩方法、压缩参数、压缩后模型结构, 是否可参与压缩或微调过程	区分普通攻击、自适应攻击和压缩触发型攻击, 明确威胁假设
部署形态与工具链	推理框架、硬件平台、量化/剪枝工具链、边缘端/本地化/联邦部署环境	评估压缩模型在真实部署环境中的安全差异, 而非仅在离线实验中评估

未来可从压缩前后行为一致性、压缩敏感参数定位和压缩后模型专门检测这 3 个角度开展研究. 首先, 可比较模型在压缩前后的输出分布、激活模式、关键神经元响应和决策边界变化, 识别仅在压缩后出现的异常行为. 其次, 可分析量化误差、剪枝掩码和蒸馏知识与攻击信号之间的关系, 定位对压缩触发风险贡献较大的参数区域或结构单元. 此外, 应将安全检测对象从原始模型扩展到压缩后的最终部署模型, 发展面向压缩模型的异常行为检测和恶意行为修复方法. 在缓解方面, 可以进一步探索带有安全约束的压缩过程, 例如在剪枝时避免保留相关子结构, 在量化时抑制异常舍入误差放大, 在蒸馏时削弱相关知识的迁移等, 从而降低压缩过程被恶意利用的可能性.

6.3　模型压缩影响安全与隐私的内在机理研究

当前关于压缩模型安全与隐私的研究仍以实验观察和现象总结为主, 缺乏对内在机理的系统解释. 已有工作已经发现, 模型压缩可能在不同条件下增强、削弱或改变模型的安全属性. 例如, 修剪可能移除冗余参数, 也可能保留或强化后门相关关键结构; 量化可能引入数值扰动并改变决策边界, 也可能在特定设置下影响成员推理风险. 然而, 现有研究往往围绕单一压缩技术展开, 不同压缩方法之间缺乏横向对比, 也缺少统一视角来解释压缩如何影响模型的鲁棒性、隐私记忆和攻击面分布.

未来研究需要从单点实验评估转向压缩影响安全与隐私机制的系统性分析. 一方面, 可以从参数空间、表示空间和决策空间这 3 个层次切入, 分析压缩如何改变权重分布、激活模式、梯度特征和样本记忆程度等, 从而解释攻击成功率变化的原因. 另一方面, 可以开展跨压缩技术的对比研究, 在相同模型、数据、攻击和评价指标下比较不同压缩技术对安全属性的不同影响, 区分共性机制与方法特有机制. 此外, 还可引入可解释性分析、因果分析和损失景观分析等工具, 辅助揭示压缩操作与安全风险之间的关系, 从而为安全感知压缩方法设计提供更可靠的理论和实验依据.

6.4　大模型压缩中的安全对齐保持与隐私记忆控制

随着大模型逐步向本地化、边缘化场景部署, 压缩技术被广泛用于降低显存占用、推理成本和服务延迟. 然而, 大模型压缩带来的安全问题不同于传统中小规模模型. 大模型通常经过安全对齐, 其拒答能力、越狱鲁棒性和内容安全边界可能依赖于特定参数区域和表示结构. 压缩操作在降低模型规模的同时, 也可能破坏这些安全相关结构, 使模型出现安全对齐退化、越狱风险上升或异常输出增加等问题. 同时, 大模型具有更强的训练数据记忆能力, 压缩过程可能改变模型对训练样本的记忆分布, 影响成员推理和隐私泄露等风险. 该问题的难点在于, 大模型

内部机制复杂, 安全对齐和隐私记忆并不集中于单一层或单一模块, 压缩强度、压缩位置和后续微调过程都可能共同影响安全结果.

未来在安全对齐保持方面, 可以分析不同量化位宽、剪枝位置和低秩适配模块等对拒答能力和越狱鲁棒性的影响, 识别安全对齐相关的关键模块或参数区域, 并在压缩过程中对这些区域设置保护约束. 在隐私记忆控制方面, 可评估压缩前后模型对成员样本、敏感样本和已记忆文本的响应变化, 研究压缩是否会抑制或放大训练数据记忆. 对于使用知识蒸馏的压缩技术, 可以进一步分析教师模型中的成员信息、记忆痕迹和后门行为如何通过中间特征或生成式输出迁移到学生模型, 并设计带有隐私约束或风险过滤的蒸馏策略. 对于低秩适应方法, 还需要关注 LoRA 等适配模块对安全对齐和成员推理风险的影响, 探索适配模块加载、合并和共享过程中的安全验证机制.

6.5 面向硬件-算法协同部署的压缩模型安全防护

模型压缩的实际安全风险不仅存在于算法层面, 也会在真实硬件和推理框架中被进一步放大. 随着量化、稀疏化和结构化修剪被广泛应用于边缘设备、专用加速器和云端多租户平台, 压缩模型的安全性开始受到存储结构、运行时调度和低比特计算机制的影响. 例如, 量化模型中的关键比特或缩放因子可能成为比特翻转攻击的敏感目标; 稀疏执行可能通过访存模式泄露模型结构; 低比特模型在共享硬件环境中也可能面临功耗侧信道等隐私风险. 该问题的难点在于, 传统模型安全评估通常只关注算法输出和攻击成功率, 而硬件-算法协同风险需要同时考虑压缩格式、硬件平台和物理执行行为等, 评估链路更长, 影响因素更多.

未来应面向真实部署环节开展压缩模型安全防护研究. 首先, 需要将量化工具链、稀疏执行机制、硬件平台、推理框架和运行时环境纳入威胁模型, 评估压缩模型在完整性、可用性和隐私泄露方面的风险. 其次, 可针对关键压缩参数开展敏感性分析, 例如识别对比特翻转最敏感的量化权重或稀疏索引, 并探索冗余存储、完整性校验等保护机制. 此外, 对于稀疏化和低比特部署, 可研究执行路径随机化和侧信道风险评估等方法, 降低模型结构和参数通过物理行为泄露的可能性. 最后, 还应将时延、能耗、显存占用和崩溃率等部署层指标纳入安全评估, 避免仅以干净精度或攻击成功率衡量压缩模型的可靠性.

7 结束语

模型压缩技术已成为深度学习模型高效部署与规模化应用的重要支撑手段, 而其与安全和隐私问题之间的相互作用, 也使模型压缩研究从单纯的效率优化逐渐扩展到安全与可信维度. 本文从安全视角出发, 系统梳理了模型压缩在带来效率优势的同时可能引发的安全与隐私风险, 总结了面向对抗鲁棒性与隐私保护的安全感知模型压缩方法, 并对该领域面临的主要挑战与未来研究方向进行了讨论. 综合现有研究可以发现, 模型压缩已不再只是性能优化工具, 而是深刻影响模型安全属性的重要因素, 其设计与应用需要在压缩率、模型性能以及安全与隐私目标之间进行系统权衡. 未来, 如何建立更可比较的评估规范、深入理解压缩影响安全与隐私的内在机理, 并在大模型和真实部署环境中实现安全可信的模型压缩, 仍是值得持续关注的问题.

References

[1] Deng L, Li GQ, Han S, Shi LP, Xie Y. Model compression and hardware acceleration for neural networks: A comprehensive survey. Proc. of the IEEE, 2020, 108(4): 485–532. [doi: 10.1109/JPROC.2020.2976475]

[2] Tian YL, Suya F, Xu FY, Evans D. Stealthy backdoors as compression artifacts. IEEE Trans. on Information Forensics and Security, 2022, 17: 1372–1387. [doi: 10.1109/TIFS.2022.3160359]

[3] Yuan XY, Zhang L. Membership inference attacks and defenses in neural network pruning. In: Proc. of the 31st USENIX Security Symp. Boston: USENIX Association, 2022. 4561–4578.

[4] Luo H, Zhuang ZW, Li YQ, Tan MK, Chen C, Zhang JL. Toward compact and robust model learning under dynamically perturbed environments. IEEE Trans. on Circuits and Systems for Video Technology, 2024, 34(6): 4857–4873. [doi: 10.1109/TCSVT.2023.3337538]

[5] Wang LX, Gu QQ. A knowledge transfer framework for differentially private sparse learning. In: Proc. of the 34th AAAI Conf. on Artificial Intelligence. New York: AAAI, 2020. 6235–6242.

[6] Pavlitska S, Grolig H, Zöllner JM. Relationship between model compression and adversarial robustness: A review of current evidence. In: Proc. of the 2023 IEEE Symp. Series on Computational Intelligence. Mexico City: IEEE, 2023. 671–676. [doi: 10.1109/SSCI52147.2023.10371946]

[7] Darvishzadeh A, Salloum H, Rasheed B, Pezer M, Mazzara M. Exploring novel perspectives on model compression techniques and their impact on adversarial robustness in deep learning: A comprehensive review and analysis. 2025. https://www.preprints.org/manuscript/202507.0074

[8] Gao H, Tian YL, Xu FY, Zhong S. Survey of deep learning model compression and acceleration. Ruan Jian Xue Bao/Journal of Software, 2021, 32(1): 68–92 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/6096.htm [doi: 10.13328/j.cnki.jos.006096]

[9] Blalock D, Ortiz JJG, Frankle J, Guttag J. What is the state of neural network pruning? In: Proc. of the 3rd Conf. on Machine Learning and Systems. Austin: mlsys.org, 2020. 1–18.

[10] Molchanov P, Tyree S, Karras T, Aila T, Kautz J. Pruning convolutional neural networks for resource efficient inference. In: Proc. of the 5th Int'l Conf. on Learning Representations. Toulon: OpenReview.net, 2017.

[11] Zhu M, Gupta S. To prune, or not to prune: Exploring the efficacy of pruning for model compression. In: Proc. of the 6th Int'l Conf. on Learning Representations. Vancouver: OpenReview.net, 2018.

[12] Frankle J, Carbin M. The lottery ticket hypothesis: Finding sparse, trainable neural networks. In: Proc. of the 7th Int'l Conf. on Learning Representations. New Orleans: OpenReview.net, 2019.

[13] Li H, Kadav A, Durdanovic I, Samet H, Graf HP. Pruning filters for efficient ConvNets. In: Proc. of the 5th Int'l Conf. on Learning Representations. Toulon: OpenReview.net, 2017.

[14] Liu Z, Li JG, Shen ZQ, Huang G, Yan SM, Zhang CS. Learning efficient convolutional networks through network slimming. In: Proc. of the 2017 IEEE Int'l Conf. on Computer Vision. Venice: IEEE, 2017. 2755–2763. [doi: 10.1109/ICCV.2017.298]

[15] Fan A, Grave E, Joulin A. Reducing Transformer depth on demand with structured dropout. In: Proc. of the 8th Int'l Conf. on Learning Representations. Addis Ababa: OpenReview.net, 2020.

[16] Jordao A, Lie M, Schwartz WR. Discriminative layer pruning for convolutional neural networks. IEEE Journal of Selected Topics in Signal Processing, 2020, 14(4): 828–837. [doi: 10.1109/JSTSP.2020.2975987]

[17] He Y, Xiao LG. Structured pruning for deep convolutional neural networks: A survey. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2024, 46(5): 2900–2919. [doi: 10.1109/TPAMI.2023.3334614]

[18] Chen YF, Guo Y, Chen Q, Li ML, Zeng W, Wang YW, Tan MK. Contrastive neural architecture search with neural architecture comparators. In: Proc. of the 2021 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021. 9497–9506. [doi: 10.1109/CVPR46437.2021.00938]

[19] Chen LY, Chen YQ, Xi JT, Le XY. Knowledge from the original network: Restore a better pruned network with knowledge distillation. Complex & Intelligent Systems, 2022, 8(2): 709–718. [doi: 10.1007/s40747-020-00248-y]

[20] Muralidharan S, Sreenivas ST, Joshi R, Chochowski M, Patwary M, Shoeybi M, Catanzaro B, Kautz J, Molchanov P. Compact language models via pruning and knowledge distillation. In: Proc. of the 38th Int'l Conf. on Neural Information Processing Systems. Vancouver: Curran Associates Inc., 2024. 1299.

[21] Rokh B, Azarpeyvand A, Khanteymoori A. A comprehensive survey on model quantization for deep neural networks in image classification. ACM Trans. on Intelligent Systems and Technology, 2023, 14(6): 97. [doi: 10.1145/3623402]

[22] Tang W, Hua G, Wang L. How to train a compact binary neural network with high accuracy? In: Proc. of the 31st AAAI Conf. on Artificial Intelligence. San Francisco: AAAI Press, 2017. 2625–2631. [doi: 10.1609/aaai.v31i1.10862]

[23] Courbariaux M, Bengio Y, David JP. BinaryConnect: Training deep neural networks with binary weights during propagations. In: Proc. of the 29th Int'l Conf. on Neural Information Processing Systems. Montreal: MIT Press, 2015. 3123–3131.

[24] Deng X, Zhang ZF. Sparsity-control ternary weight networks. Neural Networks, 2022, 145: 221–232. [doi: 10.1016/j.neunet.2021.10.018]

[25] Jin Q, Ren J, Zhuang R, Hanumante S, Li ZG, Chen ZY, Wang YZ, Yang KY, Tulyakov S. F8Net: Fixed-point 8-bit only multiplication for network quantization. In: Proc. of the 10th Int'l Conf. on Learning Representations. OpenReview.net, 2022.

[26] Basat RB, Ben-Itzhak Y, Mitzenmacher M, Vargaftik S. Optimal and approximate adaptive stochastic quantization. In: Proc. of the 38th Int'l Conf. on Neural Information Processing Systems. Vancouver: Curran Associates Inc., 2024. 2990.

[27] Gong RH, Liu XL, Jiang SH, Li TX, Hu P, Lin JZ, Yu FW, Yan JJ. Differentiable soft quantization: Bridging full-precision and low-bit neural networks. In: Proc. of the 2019 IEEE/CVF Int'l Conf. on Computer Vision. Seoul: IEEE, 2019. 4851–4860. [doi: 10.1109/ICCV.2019.00495]

[28] Li YH, Dong X, Wang W. Additive powers-of-two quantization: An efficient non-uniform discretization for neural networks. In: Proc. of the 8th Int'l Conf. on Learning Representations. Addis Ababa: OpenReview.net, 2020.

[29] Jacob B, Kligys S, Chen B, Zhu ML, Tang M, Howard AG, Adam H, Kalenichenko D. Quantization and training of neural networks for efficient integer-arithmetic-only inference. In: Proc. of the 2018 IEEE Conf. on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 2704–2713.

[30] Shao RR, Liu YA, Zhang W, Wang J. A survey of knowledge distillation in deep learning. Chinese Journal of Computers, 2022, 45(8): 1638–1673 (in Chinese with English abstract). [doi: 10.11897/SP.J.1016.2022.01638]

[31] Gou JP, Yu BS, Maybank SJ, Tao DC. Knowledge distillation: A survey. Int'l Journal of Computer Vision, 2021, 129(6): 1789–1819. [doi: 10.1007/s11263-021-01453-z]

[32] Kullback S, Leibler RA. On information and sufficiency. The Annals of Mathematical Statistics, 1951, 22(1): 79–86. [doi: 10.1214/aoms/1177729694]

[33] Choudhary T, Mishra V, Goswami A, Sarangapani J. A comprehensive survey on model compression and acceleration. Artificial Intelligence Review, 2020, 53(7): 5113–5155. [doi: 10.1007/s10462-020-09816-7]

[34] Jaderberg M, Vedaldi A, Zisserman A. Speeding up convolutional neural networks with low rank expansions. In: Proc. of the 2014 British Machine Vision Conf. Nottingham: BMVA Press, 2014. 1–12. [doi: 10.5244/C.28.88]

[35] Hu EJ, Shen YL, Wallis P, Allen-Zhu Z, Li YZ, Wang SA, Wang L, Chen WZ. LoRA: Low-rank adaptation of large language models. In: Proc. of the 10th Int'l Conf. on Learning Representations. OpenReview.net, 2022.

[36] Ji SL, Du TY, Li JF, Shen C, Li B. Security and privacy of machine learning models: A survey. Ruan Jian Xue Bao/Journal of Software, 2021, 32(1): 41–67 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/6131.htm [doi: 10.13328/j.cnki.jos.006131]

[37] Biggio B, Nelson B, Laskov P. Poisoning attacks against support vector machines. In: Proc. of the 29th Int'l Conf. on Int'l Conf. on Machine Learning. Edinburgh: Omnipress, 2012. 1467–1474.

[38] Steinhardt J, Koh PW, Liang P. Certified defenses for data poisoning attacks. In: Proc. of the 31st Int'l Conf. on Neural Information Processing Systems. Long Beach: Curran Associates Inc., 2017. 3520–3532.

[39] Li YM, Jiang Y, Li ZF, Xia ST. Backdoor learning: A survey. IEEE Trans. on Neural Networks and Learning Systems, 2024, 35(1): 5–22. [doi: 10.1109/TNNLS.2022.3182979]

[40] Gu TY, Dolan-Gavitt B, Garg S. BadNets: Identifying vulnerabilities in the machine learning model supply chain. arXiv:1708.06733, 2017.

[41] Saha A, Subramanya A, Pirsiavash H. Hidden trigger backdoor attacks. In: Proc. of the 34th AAAI Conf. on Artificial Intelligence. New York: AAAI Press, 2020. 11957–11965. [doi: 10.1609/aaai.v34i07.6871]

[42] Bagdasaryan E, Shmatikov V. Blind backdoors in deep learning models. In: Proc. of the 30th USENIX Security Symp. USENIX Association, 2021. 1505–1521.

[43] Song CZ, Ristenpart T, Shmatikov V. Machine learning models that remember too much. In: Proc. of the 2017 ACM SIGSAC Conf. on Computer and Communications Security. Dallas: ACM, 2017. 587–601. [doi: 10.1145/3133956.3134077]

[44] Feldman V. Does learning require memorization? A short tale about a long tail. In: Proc. of the 52nd Annual ACM SIGACT Symp. on Theory of Computing. Chicago: ACM, 2020. 954–959. [doi: 10.1145/3357713.3384290]

[45] Carlini N, Ippolito D, Jagielski M, Lee K, Tramer F, Zhang CY. Quantifying memorization across neural language models. In: Proc. of the 11th Int'l Conf. on Learning Representations. Kigali: OpenReview.net, 2023.

[46] Szegedy C, Zaremba W, Sutskever I, Bruna J, Erhan D, Goodfellow IJ, Fergus R. Intriguing properties of neural networks. In: Proc. of the 2nd Int'l Conf. on Learning Representations. Banff: OpenReview.net, 2014.

[47] Goodfellow IJ, Shlens J, Szegedy C. Explaining and harnessing adversarial examples. In: Proc. of the 3rd Int'l Conf. on Learning Representations. San Diego: OpenReview.net, 2015.

[48] Zhang T, Yang KW, Wei JH, Liu Y, Ning YL. Survey on detecting and defending adversarial examples for image data. Journal of Computer Research and Development, 2022, 59(6): 1315–1328 (in Chinese with English abstract). [doi: 10.7544/issn1000-1239.20200777]

[49] Schwinn L, Dobre D, Günnemann S, Gidel G. Adversarial attacks and defenses in large language models: Old and new threats. In: Proc. of the 37th Conf. on Neural Information Processing Systems Workshop. New Orleans: PMLR, 2023. 239: 103–117.

[50] Parekh S, Shukla P, Shah D. Attacking compressed vision Transformers. In: Proc. of the 2023 Future of Information and Communication Conf. San Francisco: Springer, 2023. 743–758. [doi: 10.1007/978-3-031-28073-3_50]

[51] Shokri R, Stronati M, Song CZ, Shmatikov V. Membership inference attacks against machine learning models. In: Proc. of the 2017 IEEE Symp. on Security and Privacy. San Jose: IEEE, 2017. 3–18. [doi: 10.1109/SP.2017.41]

[52] Fredrikson M, Jha S, Ristenpart T. Model inversion attacks that exploit confidence information and basic countermeasures. In: Proc. of the 22nd ACM SIGSAC Conf. on Computer and Communications Security. Denver: ACM, 2015. 1322–1333. [doi: 10.1145/2810103.2813677]

[53] Orekondy T, Schiele B, Fritz M. Knockoff nets: Stealing functionality of black-box models. In: Proc. of the 2019 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 4949–4958. [doi: 10.1109/CVPR.2019.00509]

[54] Fan JX, Yan Q, Li MH, Qu GQ, Xiao Y. A survey on data poisoning attacks and defenses. In: Proc. of the 7th IEEE Int'l Conf. on Data Science in Cyberspace. Guilin: IEEE, 2022. 48–55. [doi: 10.1109/DSC55868.2022.00014]

[55] Koh PW, Liang P. Understanding black-box predictions via influence functions. In: Proc. of the 34th Int'l Conf. on Machine Learning. Sydney: JMLR.org, 2017. 1885–1894.

[56] Wang BL, Yao YS, Shan S, Li HY, Viswanath B, Zheng HT, Zhao BY. Neural cleanse: Identifying and mitigating backdoor attacks in neural networks. In: Proc. of the 2019 IEEE Symp. on Security and Privacy. San Francisco: IEEE, 2019. 707–723. [doi: 10.1109/SP.2019.00031]

[57] Zheng RK, Tang RJ, Li JZ, Liu L. Pre-activation distributions expose backdoor neurons. In: Proc. of the 36th Int'l Conf. on Neural Information Processing Systems. New Orleans: Curran Associates Inc., 2022. 1356.

[58] Wu DX, Wang YS. Adversarial neuron pruning purifies backdoored deep models. In: Proc. of the 35th Int'l Conf. on Neural Information Processing Systems. Curran Associates Inc., 2021. 1293.

[59] Goldblum M, Tsipras D, Xie CL, Chen XY, Schwarzschild A, Song D, Mądry A, Li B, Goldstein T. Dataset security for machine learning: Data poisoning, backdoor attacks, and defenses. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2023, 45(2): 1563–1580. [doi: 10.1109/TPAMI.2022.3162397]

[60] Madry A, Makelov A, Schmidt L, Tsipras D, Vladu A. Towards deep learning models resistant to adversarial attacks. In: Proc. of the 6th Int'l Conf. on Learning Representations. Vancouver: OpenReview.net, 2018.

[61] Rice L, Wong E, Kolter JZ. Overfitting in adversarially robust deep learning. In: Proc. of the 37th Int'l Conf. on Machine Learning. JMLR.org, 2020. 749.

[62] Wang YS, Ma XJ, Bailey J, Yi JF, Zhou BW, Gu QQ. On the convergence and robustness of adversarial training. In: Proc. of the 36th Int'l Conf. on Machine Learning. Long Beach: PMLR, 2019. 6586–6595.

[63] Xu YC, Sun YC, Goldblum M, Goldstein T, Huang FR. Exploring and exploiting decision boundary dynamics for adversarial robustness. In: Proc. of the 11th Int'l Conf. on Learning Representations. Kigali: OpenReview.net, 2023.

[64] Hu HS, Salcic Z, Sun LC, Dobbie G, Yu PS, Zhang XY. Membership inference attacks on machine learning: A survey. ACM Computing Surveys, 2022, 54(11s): 235. [doi: 10.1145/3523273]

[65] Abadi M, Chu A, Goodfellow I, McMahan HB, Mironov I, Talwar K, Zhang L. Deep learning with differential privacy. In: Proc. of the 2016 ACM SIGSAC Conf. on Computer and Communications Security. Vienna: ACM, 2016. 308–318. [doi: 10.1145/2976749.2978318]

[66] Tan JX, Zhong N, Qian ZX, Zhang XP, Li S. Deep neural network watermarking against model extraction attack. In: Proc. of the 31st ACM Int'l Conf. on Multimedia. Ottawa: ACM, 2023. 1588–1597. [doi: 10.1145/3581783.3612515]

[67] Phan H, Xie Y, Liu J, Chen YY, Yuan B. Invisible and efficient backdoor attacks for compressed deep neural networks. In: Proc. of the 2022 IEEE Int'l Conf. on Acoustics, Speech and Signal Processing. Singapore: IEEE, 2022. 96–100. [doi: 10.1109/ICASSP43922.2022.9747582]

[68] Phan H, Shi C, Xie Y, Zhang TF, Li ZH, Zhao TM, Liu J, Wang Y, Chen YY, Yuan B. RIBAC: Towards robust and imperceptible backdoor attack against compact DNN. In: Proc. of the 17th European Conf. on Computer Vision. Tel Aviv: Springer, 2022. 708–724. [doi: 10.1007/978-3-031-19772-7_41]

[69] Yin ZY, Yuan Y, Guo PF, Zhou P. Backdoor attacks on federated learning with lottery ticket hypothesis. arXiv:2109.10512, 2021.

[70] Pan XD, Zhang M, Yan YF, Yang M. Understanding the threats of trojaned quantized neural network in model supply chains. In: Proc. of the 37th Annual Computer Security Applications Conf. ACM, 2021. 634–645. [doi: 10.1145/3485832.3485881]

[71] Hong S, Panaitescu M, Kaya Y, Dumitraș D. Qu-ANTI-zation: Exploiting quantization artifacts for achieving adversarial outcomes. In: Proc. of the 35th Int'l Conf. on Neural Information Processing Systems. Curran Associates Inc., 2021. 712.

[72] Ma H, Qiu HM, Gao YS, Zhang Z, Abuadbba A, Xue MH, Fu AM, Zhang JL, Al-Sarawi SF, Abbott D. Quantization backdoors to deep learning commercial frameworks. IEEE Trans. on Dependable and Secure Computing, 2024, 21(3): 1155–1172. [doi: 10.1109/TDSC.2023.3271956]

[73] Huynh T, Tran A, Doan KD, Pham T. Data poisoning quantization backdoor attack. In: Proc. of the 18th European Conf. on Computer Vision. Milan: Springer, 2024. 38–54. [doi: 10.1007/978-3-031-72907-2_3]

[74] Zhu YF, Peng HB, Fu AM, Yang W, Ma H, Al-Sarawi SF, Abbott D, Gao YS. Towards robustness evaluation of backdoor defense on quantized deep learning models. Expert Systems with Applications, 2024, 255: 124599. [doi: 10.1016/j.eswa.2024.124599]

[75] Li BH, Cai YS, Li HW, Xue F, Li ZF, Li YM. Nearest is not dearest: Towards practical defense against quantization-conditioned backdoor attacks. In: Proc. of the 2024 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Seattle: IEEE, 2024. 24523–24533. [doi: 10.1109/CVPR52733.2024.02315]

[76] Ge YJ, Wang Q, Zheng BL, Zhuang XL, Li Q, Shen C, Wang C. Anti-distillation backdoor attacks: Backdoors can really survive in knowledge distillation. In: Proc. of the 29th ACM Int'l Conf. on Multimedia. ACM, 2021. 826–834. [doi: 10.1145/3474085.3475254]

[77] Tang L, Shlomi T, Cai A. Learning the wrong lessons: Inserting trojans during knowledge distillation. In: Proc. of the 2023 Workshop on Backdoor Attacks and Defenses in Machine Learning. Kigali: OpenReview.net, 2023.

[78] Chen JY, Cao ZQ, Chen RX, Zheng HB, Li X, Xuan Q, Yang X. Like teacher, like pupil: Transferring backdoors via feature-based knowledge distillation. Computers & Security, 2024, 146: 104041. [doi: 10.1016/j.cose.2024.104041]

[79] Baras A, Zolfi A, Elovici Y, Shabtai A. QuantAttack: Exploiting quantization techniques to attack vision Transformers. In: Proc. of the 2025 IEEE/CVF Winter Conf. on Applications of Computer Vision. Tucson: IEEE, 2025. 6730–6740. [doi: 10.1109/WACV61041.2025.00655]

[80] Yang YL, Lin CH, Li Q, Zhao ZY, Fan HR, Zhou DW, Wang NN, Liu TL, Shen C. Quantization aware attack: Enhancing transferable adversarial attacks by model quantization. IEEE Trans. on Information Forensics and Security, 2024, 19: 3265–3278. [doi: 10.1109/TIFS.2024.3360891]

[81] Chu TS, Fang K, Yang J, Huang XL. Improving the adversarial robustness of quantized neural networks via exploiting the feature diversity. Pattern Recognition Letters, 2023, 176: 117–122. [doi: 10.1016/j.patrec.2023.10.024]

[82] Lian X, Huang ZQ, Wang C. AKD: Using adversarial knowledge distillation to achieve black-box attacks. In: Proc. of the 2023 Int'l Joint Conf. on Neural Networks. Gold Coast: IEEE, 2023. 1–7. [doi: 10.1109/IJCNN54540.2023.10191087]

[83] Savostianova D, Zangrando E, Tudisco F. Low-rank adversarial PGD attack. In: Proc. of the 14th Int'l Conf. on Learning Representations. Rio de Janeiro: OpenReview.net, 2026.

[84] Han G, Li Z, Guo SQ. PRJack: Pruning-resistant model hijacking attack against deep learning models. In: Proc. of the 2024 Int'l Joint Conf. on Neural Networks. Yokohama: IEEE, 2024. 1–8. [doi: 10.1109/IJCNN60899.2024.10650019]

[85] Yi X, Li Y, Zheng SF, Wang LL, Wang XL, He L. Unified attacks to large language model watermarks: Spoofing and scrubbing in unauthorized knowledge distillation. Knowledge-based Systems, 2025, 329: 114295. [doi: 10.1016/j.knosys.2025.114295]

[86] Shang J, Wang J, Wang KL, Liu JQ, Jiang N, Armanuzzaman M, Zhao ZM. Defending against membership inference attacks on iteratively pruned deep neural networks. In: Proc. of the 32nd Annual Network and Distributed System Security Symp. San Diego: Internet Society, 2025.

[87] Kowalski C, Famili A, Lao YJ. Towards model quantization on the resilience against membership inference attacks. In: Proc. of the 2022 IEEE Int'l Conf. on Image Processing. Bordeaux: IEEE, 2022. 3646–3650. [doi: 10.1109/ICIP46576.2022.9897681]

[88] Aubinais E, Formont P, Piantanida P, Gassiat E. Membership inference risks in quantized models: A theoretical and empirical study. arXiv:2502.06567, 2025.

[89] Jagielski M, Nasr M, Lee K, Choquette-Choo C, Carlini N, Tramèr F. Students parrot their teachers: Membership inference on model distillation. In: Proc. of the 37th Int'l Conf. on Neural Information Processing Systems. New Orleans: Curran Associates Inc., 2023. 1921.

[90] Galichin AV, Pautov M, Zhavoronkin A, Rogov OY, Oseledets I. GLiRA: Closed-box membership inference attack via knowledge distillation. IEEE Trans. on Information Forensics and Security, 2025, 20: 3893–3906. [doi: 10.1109/TIFS.2025.3550068]

[91] Li N, Gao YS, Hu HS, Kuang BY, Fu AM. CompLeak: Deep learning model compression exacerbates privacy leakage. arXiv: 2507.16872, 2025.

[92] Xu N, Liu Q, Liu T, Liu ZH, Guo XC, Wen WJ. Stealing your data from compressed machine learning models. In: Proc. of the 57th ACM/IEEE Design Automation Conf. San Francisco: IEEE, 2020. 1–6. [doi: 10.1109/DAC18072.2020.9218633]

[93] Shang J, Wang J, Wang KL, Jiang N, Liu JQ. MSG: Stealing data from pruned neural networks via malicious sparsity guidance. Neural Networks, 2026, 193: 108036. [doi: 10.1016/j.neunet.2025.108036]

[94] Yang DQ, Nair PJ, Lis M. HuffDuff: Stealing pruned DNNs from sparse accelerators. In: Proc. of the 28th ACM Int'l Conf. on Architectural Support for Programming Languages and Operating Systems. Vancouver: ACM, 2023. 385–399. [doi: 10.1145/3575693.3575738]

[95] Chen XX, Chen TL, Zhang ZY, Wang ZY. You are caught stealing my winning lottery ticket! Making a lottery ticket claim its ownership. In: Proc. of the 35th Int'l Conf. on Neural Information Processing Systems. Curran Associates Inc., 2021. 137.

[96] Kuang WX, Liang QZ, Sun P, Fu W, Hu Q, Hu YP. Unveiling the pruning risks on privacy vulnerabilities of deep neural networks. In: Proc. of the 2025 IEEE Int'l Conf. on Acoustics, Speech and Signal Processing. Hyderabad: IEEE, 2025. 1–5. [doi: 10.1109/ICASSP49660.2025.10889113]

[97] Wei BY, Huang KX, Huang YSB, Xie TH, Qi XY, Xia MZ, Mittal P, Wang MD, Henderson P. Assessing the brittleness of safety alignment via pruning and low-rank modifications. In: Proc. of the 41st Int'l Conf. on Machine Learning. Vienna: JMLR.org, 2024. 2156.

[98] Egashira K, Vero M, Staab R, He JX, Vechev M. Exploiting LLM quantization. In: Proc. of the 38th Int'l Conf. on Neural Information Processing Systems. Vancouver: Curran Associates Inc., 2024. 1319.

[99] Dong PR, Li HW, Guo S. Durable quantization conditioned misalignment attack on large language models. In: Proc. of the 13th Int'l Conf. on Learning Representations. Singapore: OpenReview.net, 2025.

[100] Hussain A, Zarkouei SAK, Rabin RI, Alipour MA, Lin S, Xu BW. Capturing the effects of quantization on Trojans in code LLMs. arXiv:2505.14200, 2025.

[101] Cheng PZ, Wu ZR, Ju TJ, Du W, Zhang ZS, Liu GS. Transferring backdoors between large language models by knowledge distillation. arXiv:2408.09878, 2024.

[102] Li MJ, Si WM, Backes M, Zhang Y, Wang YS. SaLoRA: Safety-alignment preserved low-rank adaptation. In: Proc. of the 13th Int'l Conf. on Learning Representations. Singapore: OpenReview.net, 2025.

[103] Liu HY, Zhong SC, Sun XT, Tian MH, Hariri M, Liu ZR, Tang RX, Jiang ZM, Yuan JY, Chuang YN, Li L, Choi SH, Chen R, Chaudhary V, Hu X. LoRATK: LoRA once, backdoor everywhere in the share-and-play ecosystem. In: Findings of the Association for Computational Linguistics: EMNLP 2025. Suzhou: ACL, 2025. 23009–23047. [doi: 10.18653/v1/2025.findings-emnlp.1253]

[104] Ni CJ, Wang ZP, Bao RX, Gao SQ, Zhang YF. Controllable memorization in LLMs via weight pruning. In: Proc. of the 2025 Conf. on Empirical Methods in Natural Language Processing. Suzhou: ACL, 2025. 15130–15145. [doi: 10.18653/v1/2025.emnlp-main.765]

[105] Haque MN, Yang H, Yang Z, Xu BW. How quantization impacts privacy risk on LLMs for code? arXiv:2508.00128, 2025.

[106] Zhang ZQ, Shahin Shamsabadi A, Lu HX, Cai YF, Haddadi H. Membership and memorization in LLM knowledge distillation. In: Proc. of the 2025 Conf. on Empirical Methods in Natural Language Processing. Suzhou: ACL, 2025. 20074–20084. [doi: 10.18653/v1/2025.emnlp-main.1015]

[107] Cui ZY, Zhang MX, Pei J. On membership inference attacks in knowledge distillation. arXiv:2505.11837, 2025.

[108] Liang Z, Ye QQ, Wang YY, Zhang S, Xiao YX, Li RH, Xu JL, Hu HB. "Yes, My LoRD." Guiding language model extraction with locality reinforced distillation. In: Proc. of the 63rd Annual Meeting of the Association for Computational Linguistics. Vienna: ACL, 2025. 1441–1465. [doi: 10.18653/v1/2025.acl-long.73]

[109] Ran DL, He XL, Cong TS, Wang AY, Li Q, Wang XY. LoRA-Leak: Membership inference attacks against LoRA fine-tuned language models. arXiv:2507.18302, 2025.

[110] Zhang YD, Huang L, Gao PF, Song F, Sun J, Dong JS. Verification of bit-flip attacks against quantized neural networks. Proc. of the ACM on Programming Languages, 2025, 9(OOPSLA1): 115. [doi: 10.1145/3720471]

[111] Cheng GY, Fei YS. One flip away from chaos: Unraveling single points of failure in quantized DNNs. In: Proc. of the 2024 IEEE Int'l Symp. on Hardware Oriented Security and Trust. Tysons Corner: IEEE, 2024. 332–342. [doi: 10.1109/HOST55342.2024.10545351]

[112] Dubey A, Karabulut E, Awad A, Aysu A. High-fidelity model extraction attacks via remote power monitors. In: Proc. of the 4th IEEE Int'l Conf. on Artificial Intelligence Circuits and Systems. Incheon: IEEE, 2022. 328–331. [doi: 10.1109/AICAS54282.2022.9869973]

[113] Wang LY, Ding GW, Huang RT, Cao YS, Lui YC. Adversarial robustness of pruned neural networks. In: Proc. of the 6th Int'l Conf. on Learning Representations. Vancouver: OpenReview.net, 2018.

[114] Guo YW, Zhang C, Zhang CS, Chen YR. Sparse DNNs with improved adversarial robustness. In: Proc. of the 32nd Int'l Conf. on Neural Information Processing Systems. Montréal: Curran Associates Inc., 2018. 240–249.

[115] Zhao YR, Shumailov I, Mullins R, Anderson R. To compress or not to compress: Understanding the interactions between adversarial attacks and neural network compression. In: Proc. of the 2nd Conf. on Machine Learning and Systems. Stanford: mlsys.org, 2019.

[116] Merkle F, Samsinger M, Schöttle P. Pruning in the face of adversaries. In: Proc. of the 21st Int'l Conf. on Image Analysis and Processing. Lecce: Springer, 2022. 658–669. [doi: 10.1007/978-3-031-06427-2_55]

[117] Özdenizci O, Legenstein R. Training adversarially robust sparse networks via Bayesian connectivity sampling. In: Proc. of the 38th Int'l Conf. on Machine Learning. PMLR, 2021. 8314–8324.

[118] Wei X, Xu YY, Huang YH, Lv HR, Lan H, Chen MS, Tang X. Learning extremely lightweight and robust model with differentiable constraints on sparsity and condition number. In: Proc. of the 17th European Conf. on Computer Vision. Tel Aviv: Springer, 2022. 690–707. [doi: 10.1007/978-3-031-19772-7_40]

[119] Feng YQ, Lin SHJ, Gao BY, Wei X. Lipschitz constant meets condition number: Learning robust and compact deep neural networks. arXiv:2503.20454, 2025.

[120] Phan H, Yin M, Sui Y, Yuan B, Zonouz S. CSTAR: Towards compact and structured deep neural networks with adversarial robustness. In: Proc. of the 37th AAAI Conf. on Artificial Intelligence. Washington: AAAI Press, 2023. 2065–2073. [doi: 10.1609/aaai.v37i2.25299]

[121] Vora B, Patwari K, Hafiz SM, Shafiq Z, Chauh CN. Benchmarking adversarial robustness of compressed deep learning models. In: Proc. of the 2nd AdvML Frontiers Workshop at the 40th Int'l Conf. on Machine Learning. Honolulu: PMLR, 2023. 1–14.

[122] Yang HR. Towards efficient and robust deep neural network models [Ph.D. Thesis]. Durham: Duke University, 2022.

[123] Rakin AS, He ZZ, Yang L, Wang YZ, Wang LQ, Fan DL. Robust sparse regularization: Defending adversarial attacks via regularized sparse network. In: Proc. of the 2020 on Great Lakes Symp. on VLSI. ACM, 2020. 125–130. [doi: 10.1145/3386263.3407651]

[124] Gui SP, Wang HT, Yang HC, Yu C, Wang ZY, Liu J. Model compression with adversarial robustness: A unified optimization framework. In: Proc. of the 33rd Int'l Conf. on Neural Information Processing Systems. Vancouver: Curran Associates Inc., 2019. 115.

[125] Sehwag V, Wang SQ, Mittal P, Jana S. Towards compact and robust deep neural networks. arXiv:1906.06110, 2019.

[126] Ye SK, Xu KD, Liu SJ, Cheng H, Lambrechts JH, Zhang H, Zhou AJ, Ma KS, Wang YZ, Lin X. Adversarial robustness vs. model compression, or both? In: Proc. of the 2019 IEEE/CVF Int'l Conf. on Computer Vision. Seoul: IEEE, 2019. 111–120. [doi: 10.1109/ICCV.2019.00020]

[127] Xie HD, Qian LX, Xiang XS, Liu NJ. Blind adversarial pruning: Towards the comprehensive robust models with gradually pruning against blind adversarial attacks. In: Proc. of the 2021 IEEE Int'l Conf. on Multimedia and Expo. Shenzhen: IEEE, 2021. 1–6. [doi: 10.1109/ICME51207.2021.9428149]

[128] Kwon J, Lee S. Improving the robustness of model compression by on-manifold adversarial training. Future Internet, 2021, 13(12): 300. [doi: 10.3390/fi13120300]

[129] Vemparala MR, Fasfous N, Frickenstein A, Sarkar S, Zhao Q, Kuhn S, Frickenstein L, Singh A, Unger C, Nagaraja NS, Wressnegger C, Stechele W. Adversarial robust model compression using in-train pruning. In: Proc. of the 2021 IEEE/CVF Conf. on Computer Vision and Pattern Recognition Workshops. Nashville: IEEE, 2021. 66–75. [doi: 10.1109/CVPRW53098.2021.00016]

[130] Kundu S, Nazemi M, Beerel PA, Pedram M. DNR: A tunable robust pruning framework through dynamic network rewiring of DNNs. In: Proc. of the 26th Asia and South Pacific Design Automation Conf. Tokyo: ACM, 2021. 344–350. [doi: 10.1145/3394885.3431542]

[131] Kundu S, Fu Y, Ye B, Beerel PA, Pedram M. Toward adversary-aware non-iterative model pruning through dynamic network rewiring of DNNs. ACM Trans. on Embedded Computing Systems, 2022, 21(5): 52. [doi: 10.1145/3510833]

[132] Li YZ, Zhao P, Ding RY, Zhou T, Fei YS, Xu XL, Lin X. Neural architecture search for adversarial robustness via learnable pruning. Frontiers in High Performance Computing, 2024, 2: 1301384. [doi: 10.3389/FHPCP.2024.1301384]

[133] Wijayanto AW, Choong JJ, Madhawa K, Murata T. Towards robust compressed convolutional neural networks. In: Proc. of the 2019 IEEE Int'l Conf. on Big Data and Smart Computing. Kyoto: IEEE, 2019. 1–8. [doi: 10.1109/BIGCOMP.2019.8679132]

[134] Yan YS, Pei QQ. A robust deep-neural-network-based compressed model for mobile device assisted by edge server. IEEE Access, 2019, 7: 179104–179117. [doi: 10.1109/ACCESS.2019.2958406]

[135] Matachana AG, Co KT, Muñoz-González L, Martinez D, Lupu EC. Robustness and transferability of universal attacks on compressed models. In: Proc. of the 2021 AAAI Conf. on Towards Robust, Secure, and Efficient Machine Learning. AAAI, 2021. 1–10.

[136] Lee J, Lee S. Robust CNN compression framework for security-sensitive embedded systems. Applied Sciences, 2021, 11(3): 1093. [doi: 10.3390/app11031093]

[137] Weiss JO, Alves T, Kundu S. Hardening DNNs against transfer attacks during network compression using greedy adversarial pruning. In: Proc. of the 4th IEEE Int'l Conf. on Artificial Intelligence Circuits and Systems. Incheon: IEEE, 2022. 324–327. [doi: 10.1109/AICAS54282.2022.9869910]

[138] Li JW, Lei Q, Cheng W, Xu DK. Towards robust pruning: An adaptive knowledge-retention pruning strategy for language models. In: Proc. of the 2023 Conf. on Empirical Methods in Natural Language Processing. Singapore: ACL, 2023. 1229–1247. [doi: 10.18653/v1/2023.emnlp-main.79]

[139] Jung Y, Song BC. Two is better than one: Efficient ensemble defense for robust and compact models. In: Proc. of the 2025 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Nashville: IEEE, 2025. 9696–9706. [doi: 10.1109/CVPR52734.2025.00906]

[140] Kraidia I, Ghenai A, Belhaouari SB. Defense against adversarial attacks: Robust and efficient compressed optimized neural networks. Scientific Reports, 2024, 14(1): 6420. [doi: 10.1038/s41598-024-56259-z]

[141] Sehwag V, Wang SQ, Mittal P, Jana S. HYDRA: Pruning adversarially robust neural networks. In: Proc. of the 34th Int'l Conf. on Neural Information Processing Systems. Vancouver: Curran Associates Inc., 2020. 1649.

[142] Wei X, Zhu Y, Xia ST. Batch normalization assisted adversarial pruning: Towards lightweight, sparse and robust models. In: Proc. of the 2021 Int'l Joint Conf. on Neural Networks. Shenzhen: IEEE, 2021. 1–8. [doi: 10.1109/IJCNN52387.2021.9534077]

[143] Lee BK, Kim J, Ro YM. Masking adversarial damage: Finding adversarial saliency for robust and sparse network. In: Proc. of the 2022 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. New Orleans: IEEE, 2022. 15105–15115. [doi: 10.1109/CVPR52688.2022.01470]

[144] Zhuang XL, Ge YJ, Zheng BL, Wang Q. Adversarial network pruning by filter robustness estimation. In: Proc. of the 2023 IEEE Int'l Conf. on Acoustics, Speech and Signal Processing. Rhodes Island: IEEE, 2023. 1–5. [doi: 10.1109/ICASSP49357.2023.10095291]

[145] Zhong SC, You ZC, Zhang JM, Zhao S, LeClaire Z, Liu ZR, Zha DC, Chaudhary V, Xu S, Hu X. One less reason for filter-pruning: Gaining free adversarial robustness with structured grouped kernel pruning. In: Proc. of the 37th Int'l Conf. on Neural Information Processing Systems. New Orleans: Curran Associates Inc., 2023. 2712.

[146] Han C, Wang LY, Li DY, Cui WJ, Yan B. A pruning method combined with resilient training to improve the adversarial robustness of automatic modulation classification models. Mobile Networks and Applications, 2025, 30(1-2): 274–290. [doi: 10.1007/s11036-024-02333-9]

[147] Cosentino J, Zaiter F, Pei D, Zhu J. The search for sparse, robust neural networks. In: Proc. of the 33rd Conf. on Neural Information Processing Systems Workshop. Vancouver: PMLR, 2019. 1–14.

[148] Fu YG, Yu QX, Zhang Y, Wu S, Ouyang X, Cox D, Lin YY. Drawing robust scratch tickets: Subnetworks with inborn robustness are found within randomly initialized networks. In: Proc. of the 35th Int'l Conf. on Neural Information Processing Systems. Curran Associates Inc., 2021. 1000.

[149] Shi XP, Zheng PF, Ding AA, Gao Y, Zhang WZ. Finding dynamics preserving adversarial winning tickets. In: Proc. of the 25th Int'l Conf. on Artificial Intelligence and Statistics. Valencia: PMLR, 2022. 510–528.

[150] Liu Q, Liu T, Liu ZH, Wang YZ, Jin YE, Wen WJ. Security analysis and enhancement of model compressed deep learning systems under adversarial attacks. In: Proc. of the 23rd Asia and South Pacific Design Automation Conf. Jeju: IEEE, 2018. 721–726. [doi: 10.1109/ASPDAC.2018.8297407]

[151] Lin J, Gan C, Han S. Defensive quantization: When efficiency meets robustness. In: Proc. of the 7th Int'l Conf. on Learning Representations. New Orleans: OpenReview.net, 2019.

[152] Bernhard R, Moellic PA, Dutertre JM. Impact of low-bitwidth quantization on the adversarial robustness for embedded neural networks. In: Proc. of the 2019 Int'l Conf. on Cyberworlds. Kyoto: IEEE, 2019. 308–315. [doi: 10.1109/CW.2019.00057]

[153] Gorsline M, Smith J, Merkel C. On the adversarial robustness of quantized neural networks. In: Proc. of the 2021 Great Lakes Symp. on VLSI. ACM, 2021. 189–194. [doi: 10.1145/3453688.3461755]

[154] Li Q, Meng Y, Tang C, Jiang JC, Wang Z. Investigating the impact of quantization on adversarial robustness. In: Proc. of the 5th Workshop on Practical ML for Limited/Low Resource Settings. Vienna: OpenReview.net, 2024.

[155] Sen S, Ravindran B, Raghunathan A. EMPIR: Ensembles of mixed precision deep networks for increased robustness against adversarial attacks. In: Proc. of the 2020 Int'l Conf. on Learning Representations. Addis Ababa: OpenReview.net, 2020.

[156] Song C, Fallon E, Li H. Improving adversarial robustness in weight-quantized neural networks. In: Proc. of the 2021 AAAI Conf. on Artificial Intelligence Workshop. AAAI, 2021. 1–10.

[157] Yang HR, Yang XX, Gong NZ, Chen YR. HERO: Hessian-enhanced robust optimization for unifying and improving generalization and quantization performance. In: Proc. of the 59th ACM/IEEE Design Automation Conf. San Francisco: ACM, 2022. 25–30. [doi: 10.1145/3489517.3530678]

[158] Ayaz F, Zakariyya I, Cano J, Keoh SL, Singer J, Pau D, Kharbouche-Harrari M. Improving robustness against adversarial attacks with deeply quantized neural networks. In: Proc. of the 2023 Int'l Joint Conf. on Neural Networks. Gold Coast: IEEE, 2023. 1–8. [doi: 10.1109/IJCNN54540.2023.10191429]

[159] Zakariyya I, Ayaz F, Kharbouche-Harrari M, Singer J, Keoh SL, Pau D, Cano J. Quantitative analysis of deeply quantized tiny neural networks robust to adversarial attacks. arXiv:2503.08973, 2025.

[160] Fukuda Y, Yoshida K, Fujino T. Evaluation of model quantization method on Vitis-AI for mitigating adversarial examples. IEEE Access, 2023, 11: 87200–87209. [doi: 10.1109/ACCESS.2023.3305264]

[161] Xiao Y, Zhang T, Liu S, Qin H. Benchmarking the robustness of quantized models. In: Proc. of the 2023 IEEE/CVF Conf. on Computer Vision and Pattern Recognition Workshops. Vancouver: IEEE, 2023. 1–4.

[162] Xiao YS, Liu AS, Zhang TY, Qin HT, Guo JY, Liu XL. RobustMQ: Benchmarking robustness of quantized models. Visual Intelligence, 2023, 1(1): 30. [doi: 10.1007/s44267-023-00031-w]

[163] Arani E, Sarfraz F, Zonooz B. Improving generalization and robustness with noisy collaboration in knowledge distillation.

arXiv:1910.05057v1, 2019.

[164] Goldblum M, Fowl L, Feizi S, Goldstein T. Adversarially robust distillation. In: Proc. of the 34th AAAI Conf. on Artificial Intelligence. New York: AAAI Press, 2020. 3996–4003. [doi: 10.1609/aaai.v34i04.5816]

[165] Shao RL, Yi JF, Chen PY, Hsieh CJ. How and when adversarial robustness transfers in knowledge distillation? In: Proc. of the 17th European Conf. on Computer Vision. Tel Aviv: Springer, 2022. 1–13.

[166] Li JW, Yang J. Revisiting the benefits of knowledge distillation against adversarial examples. In: Proc. of the 6th Int'l Conf. on Data Science and Information Technology. Shanghai: IEEE, 2023. 305–310. [doi: 10.1109/DSIT60026.2023.00054]

[167] Zi BJ, Zhao SH, Ma XJ, Jiang YG. Revisiting adversarial robustness distillation: Robust soft labels make student better. In: Proc. of the 2021 IEEE/CVF Int'l Conf. on Computer Vision. Montreal: IEEE, 2021. 16423–16432. [doi: 10.1109/ICCV48922.2021.01613]

[168] Huang B, Chen MY, Wang Y, Lu JD, Cheng MH, Wang W. Boosting accuracy and robustness of student models via adaptive adversarial distillation. In: Proc. of the 2023 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Vancouver: IEEE, 2023. 24668–24677. [doi: 10.1109/CVPR52729.2023.02363]

[169] Yue XL, Mou NP, Wang Q, Zhao LC. Revisiting adversarial robustness distillation from the perspective of robust fairness. In: Proc. of the 37th Int'l Conf. on Neural Information Processing Systems. New Orleans: Curran Associates Inc., 2023. 1323.

[170] Wu YL, Lao MR, Guo YM, Chen DM, Yua TY. Boosting adversarial robustness distillation via hybrid decomposed knowledge. In: Proc. of the 2024 IEEE Int'l Conf. on Acoustics, Speech and Signal Processing. Seoul: IEEE, 2024. 4715–4719. [doi: 10.1109/ICASSP48485.2024.10447411]

[171] Levi YY, Grolman E, Yankelev I, Giloni A, Hofman O, Shimizu T, Shabtai A, Elovici Y. KDAT: Inherent adversarial robustness via knowledge distillation with adversarial tuning for object detection models. In: Proc. of the 39th AAAI Conf. on Artificial Intelligence. Philadelphia: AAAI Press, 2025. 4598–4606. [doi: 10.1609/aaai.v39i5.32485]

[172] Zhao SJ, Wang XZ, Wei XX. Mitigating accuracy-robustness trade-off via balanced multi-teacher adversarial distillation. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2024, 46(12): 9338–9352. [doi: 10.1109/TPAMI.2024.3416308]

[173] Schotthöfer S, Yang HL, Schnake S. Dynamic low-rank training with spectral regularization: Achieving robustness in compressed representations. In: Proc. of the 2025 Workshop on Methods and Opportunities at Small Scale. Vancouver: OpenReview.net, 2025.

[174] Asante DA, Chowdhury M, Li Y. Decomposed trust: Privacy, adversarial robustness, ethics, and fairness in low-rank LLMs. arXiv:2511.22099v5, 2026.

[175] Wang J, Bao WD, Sun LC, Zhu XM, Cao BK, Yu PS. Private model compression via knowledge distillation. In: Proc. of the 33rd AAAI Conf. on Artificial Intelligence. Honolulu: AAAI Press, 2019. 1190–1197. [doi: 10.1609/aaai.v33i01.33011190]

[176] Gao D, Zhuo C. Private knowledge transfer via model distillation with generative adversarial networks. In: Proc. of the 24th European Conf. on Artificial Intelligence. Santiago de Compostela: IOS, 2020. 1794–1801. [doi: 10.3233/FAIA200294]

[177] Lyu LJ, Chen CH. Differentially private knowledge distillation for mobile analytics. In: Proc. of the 43rd Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval. ACM, 2020. 1809–1812. [doi: 10.1145/3397271.3401259]

[178] Mireshghallah F, Backurs A, Inan HA, Wutschitz L, Kulkarni J. Differentially private model compression. In: Proc. of the 36th Int'l Conf. on Neural Information Processing Systems. New Orleans: Curran Associates Inc., 2022. 2137.

[179] Flemings J, Annavaram M. Differentially private knowledge distillation via synthetic text generation. In: Findings of the Association for Computational Linguistics: ACL 2024. Bangkok: ACL, 2024. 12957–12968. [doi: 10.18653/v1/2024.findings-acl.769]

[180] Gong X, Sharma A, Karanam S, Wu ZY, Chen T, Doermann D, Innanje A. Ensemble attention distillation for privacy-preserving federated learning. In: Proc. of the 2021 IEEE/CVF Int'l Conf. on Computer Vision. Montreal: IEEE, 2021. 15056–15066. [doi: 10.1109/ICCV48922.2021.01480]

[181] Sun LC, Lyu LJ. Federated model distillation with noise-free differential privacy. In: Proc. of the 30th Int'l Joint Conf. on Artificial Intelligence. Montreal: IJCAI.org, 2021. 1563–1570. [doi: 10.24963/ijcai.2021/216]

[182] Huang D, Ye XC, Sakurai T. Knowledge distillation-based privacy-preserving data analysis. In: Proc. of the 2022 Conf. on Research in Adaptive and Convergent Systems. ACM, 2022. 15–20. [doi: 10.1145/3538641.3561482]

[183] Qi T, Wu FZ, Wu CH, He L, Huang YF, Xie X. Differentially private knowledge transfer for federated learning. Nature Communications, 2023, 14(1): 3785. [doi: 10.1038/s41467-023-38794-x]

[184] Wei LT, Zhang ZM, Zheng JH. Privacy distillation: Collaborative privacy protection for Transformer. In: Proc. of the 2nd Int'l Conf. on Generative Artificial Intelligence and Information Security. Hangzhou: ACM, 2025. 43–49. [doi: 10.1145/3728725.3728733]

[185] Yan YS, Pei QQ, Li HN. Privacy-preserving compressive model for enhanced deep-learning-based service provision system in edge computing. IEEE Access, 2019, 7: 92921–92937. [doi: 10.1109/ACCESS.2019.2927163]

[186] Lin S, Wang CH, Li HJ, Deng JR, Wang YZ, Ding CW. ESMFL: Efficient and secure models for federated learning. In: Proc. of the

34th Conf. on Neural Information Processing Systems. Vancouver: PMLR, 2020. 1–7.

[187] Zhu J, Wang LY, Han X. Safety and performance, why not both? Bi-objective optimized model compression toward AI software deployment. In: Proc. of the 37th IEEE/ACM Int'l Conf. on Automated Software Engineering. Rochester: ACM, 2022. 88. [doi: 10.1145/3551349.3556906]

[188] Chu TY, Yang MW, Laoutaris N, Markopoulou A. PriPrune: Quantifying and preserving privacy in pruned federated learning. ACM Trans. on Modeling and Performance Evaluation of Computing Systems, 2025, 10(2): 11. [doi: 10.1145/3702241]

[189] Deng ZH, Sun RX, Xue MH, Ma WL, Wen S, Nepal S, Xiang Y. Hardening LLM fine-tuning: From differentially private data selection to trustworthy model quantization. IEEE Trans. on Information Forensics and Security, 2025, 20: 7211–7226. [doi: 10.1109/TIFS.2025.3581103]

附中文参考文献

[8] 高晗, 田育龙, 许封元, 仲盛. 深度学习模型压缩与加速综述. 软件学报, 2021, 32(1): 68–92. http://www.jos.org.cn/1000-9825/6096.htm [doi: 10.13328/j.cnki.jos.006096]

[30] 邵仁荣, 刘宇昂, 张伟, 王骏. 深度学习中知识蒸馏研究综述. 计算机学报, 2022, 45(8): 1638–1673. [doi: 10.11897/SP.J.1016.2022.01638]

[36] 纪守领, 杜天宇, 李进锋, 沈超, 李博. 机器学习模型安全与隐私研究综述. 软件学报, 2021, 32(1): 41–67. http://www.jos.org.cn/1000-9825/6131.htm [doi: 10.13328/j.cnki.jos.006131]

[48] 张田, 杨奎武, 魏江宏, 刘扬, 宁原隆. 面向图像数据的对抗样本检测与防御技术综述. 计算机研究与发展, 2022, 59(6): 1315–1328. [doi: 10.7544/issn1000-1239.20200777]

作者简介

商婧, 博士生, 主要研究领域为人工智能安全, 隐私保护机器学习, 轻量化模型的安全性.

王健, 博士, 教授, 博士生导师, 主要研究领域为密码与隐私计算, 人工智能系统安全, 量子智能计算.

王凯崙, 博士生, 主要研究领域为人工智能安全, 隐私保护机器学习, 大模型智能体安全.

杨磊, 硕士生, 主要研究领域为人工智能安全, 隐私保护机器学习, 轻量化模型的安全性.

姜楠, 博士, 教授, 博士生导师, CCF 专业会员, 主要研究领域为量子机器学习, 计算智能, 内容安全.

姜强, 博士生, 主要研究领域为软件开发安全, 人工智能安全.

刘吉强, 博士, 教授, 博士生导师, CCF 专业会员, 主要研究领域为隐私保护, 可信计算, 云计算安全, 区块链.