引用格式：曹瑞麒，杨雨龙，蔺琛皓，等. 基于结构化剪枝和对抗训练的自适应鲁棒优化方法[J]. 信息对抗技术，2025，4(5)：77-88.[CAO Ruiqi，YANG Yulong，LIN Chenhao，et al. Adaptive robust optimization method based on structured pruning and adversarial training[J]. Information Countermeasure Technology，2025，4(5)：77-88.(in Chinese)]

基于结构化剪枝和对抗训练的自适应鲁棒优化方法

曹瑞麒[1,2]，杨雨龙[1,2]，蔺琛皓[1,2*]，赵正宇[1,2]，李 前[1,2]，王 骞[3]，沈 超[1,2]

（1. 西安交通大学网络空间安全学院，陕西西安 710049；2. 智能网络与网络安全教育部重点实验室（西安交通大学），陕西西安 710049；3. 武汉大学国家网络安全学院，湖北武汉 430072）

摘 要 深度神经网络在资源受限设备部署时，面临存储与计算瓶颈。结构化剪枝技术通过移除冗余权重，可有效实现模型压缩与加速，但传统剪枝网络的对抗鲁棒性不足，制约其在安全敏感场景的应用。为兼顾模型轻量化需求与鲁棒性提升，提出一种结合对抗训练与结构化剪枝的迭代优化方法：在对抗训练过程中同步优化剪枝掩码，并创新设计基于“探索-利用”策略的自适应训练-剪枝频率调整机制，以实现超参数的动态优化。在 CIFAR-10 数据集和 ResNet-18 模型上的实验结果表明，该方法在 0.7 的稀疏度下，模型鲁棒准确率提升 10.32%；在稀疏度超过 0.9 的极端场景下，正常准确率与鲁棒准确率分别提升 4.76%和 15.52%；相较于固定频率策略，自适应机制进一步将正常准确率提升 0.80%～3.59%，鲁棒准确率提升 1.30%～8.50%，显著降低人工调参成本。该研究为深度神经网络在移动端安全高效部署提供有效技术方案。

关键词 结构化剪枝；对抗训练；模型压缩；对抗鲁棒性

中图分类号 TP 391 **文章编号** 2097-163X(2025)05-0077-12

文献标志码 A **DOI** 10.12399/j.issn.2097-163x.2025.05.006

Adaptive robust optimization method based on structured pruning and adversarial training

CAO Ruiqi[1]，YANG Yulong[1,2]，LIN Chenhao[1,2*]，ZHAO Zhengyu[1,2]，LI Qian[1,2]，WANG Qian[3]，SHEN Chao[1,2]

（1. School of Cyber Science and Engineering，Xi'an Jiaotong University，Xi'an 710049，China；2. Key Laboratory for Intelligent Networks and Network Security(Xi'an Jiaotong University)，Xi'an 710049，China；3. School of Cyber Science and Engineering，Wuhan University，Wuhan，430072，China）

Abstract Deep neural networks face storage and computational bottlenecks when deployed on resource-constrained devices. Structured pruning techniques can effectively achieve model compression and acceleration by removing redundant weights，but the adversarial robustness of traditional pruning networks is insufficient，limiting their application in security-sensitive scenarios. To balance the needs for model lightweighting and robustness enhancement，an iterative optimization method combining adversarial training and structured pruning was proposed：during the adversarial training process，the pruning mask is optimized synchronously，and an

收稿日期：2025-07-08 **修回日期**：2025-09-01

通信作者：蔺琛皓，E-mail：linchenhao@xjtu.edu.cn

基金项目：国家重点研发计划项目(2023YFE0209800)；国家自然科学基金资助项目(T2341003，62376210，62161160337，62132011，U24B20185，U21B2018，62206217)；陕西省重点研发计划项目(2023-ZDLGY-38)

adaptive training-pruning frequency adjustment mechanism based on the "exploration-exploitation" strategy was innovatively designed to realize the dynamic optimization of hyperparameters. Experimental results on the CIFAR-10 dataset and ResNet-18 model show that, under a sparsity of 0.7, the proposed method increases the model's robust accuracy by 10.32%; in extreme scenarios where sparsity exceeds 0.9, the normal accuracy and robust accuracy are improved by 4.76% and 15.52% respectively; compared with the fixed-frequency strategy, the adaptive mechanism further enhances the normal accuracy by 0.80%～3.59% and the robust accuracy by 1.30%～8.50%, significantly reducing the cost of manual hyperparameter tuning. This research provides an effective technical solution for the secure and efficient deployment of deep neural networks on mobile platform.

Keywords structured pruning;adversarial training;model compression;adversarial robustness

0 引言

近年来,深度神经网络(deep neural network,DNN)的突破性进展极大推动了人工智能技术革新,尤其在计算机视觉、自然语言处理[1-4]等关键领域取得革命性成果。作为模拟人脑信息处理机制的计算模型,DNN 通过多层次特征抽取实现对复杂数据的高效理解,在目标检测、语义分割等视觉任务中展现出超越传统方法的性能。这些突破直接催生自动驾驶、医疗影像诊断、工业质检等场景的智能化应用[5],深刻影响着现代社会的技术生态。

然而,DNN 性能提升伴随的模型膨胀问题日益凸显:典型模型中,ResNet-50[6]存储占用内存超 95 MB,包含 2 300 余万个可训练参数,并且需要 4 千兆浮点运算[7];Transformer 网络的 GPT-3 模型参数规模达 1 750 亿个[8],GPT-4 参数规模进一步扩大。当前神经网络规模持续扩大的趋势,将导致终端设备因资源限制无法部署标准视觉模型,严重制约技术落地。此矛盾推动模型压缩技术研究,其中网络剪枝作为主流方案,通过移除神经网络中冗余的连接、节点或层,降低网络的复杂度,实现模型轻量化目标。

尽管模型剪枝技术取得一定进展,但现有研究普遍忽视模型对对抗样本的防御能力——在原有图像上施加轻微扰动生成对抗样本,可以使模型输出不合理的结果[9-10]。针对此问题,本研究将对抗鲁棒性作为核心约束融入剪枝框架,构建"效率—精度—安全"三维优化范式,旨在加强 DNN 在安全敏感场景和资源受限环境下的表现。

本文的研究贡献与创新如下:

1) 揭示训练-剪枝更新频率对结构化剪枝鲁棒性的调控机制。研究发现,调整模型参数训练与剪枝更新频率,可显著提高结构化剪枝对抗鲁棒训练的性能。基于对抗训练的剪枝方法可以有效增强模型的鲁棒准确率,而在剪枝过程中加入不同频率参数微调,可补偿剪枝带来的性能损失,最终使模型在稀疏度、正常准确率、鲁棒准确率 3 个目标指标间达到最优平衡。

2) 提出基于"探索-利用"(exploration-exploitation,EE)策略的自适应训练-剪枝频率调整算法。针对迭代式结构化对抗鲁棒剪枝方法中"剪枝掩码更新和参数微调频率需人工调整、计算成本高"的问题,设计基于 EE 策略的自适应算法。通过设置 3 个初始参数比例,并根据训练表现动态调整比例进行探索,自适应地找到最适合当前数据集和模型的参数设置,从而提高了模型的效率和准确性。

3) 多数据集与多模型验证方法有效性。在 CIFAR-10[11]、CIFAR-100[11]、SVHN[12]数据集以及 VGG-16[13]、ResNet-18[6]和 MobileNetV1[14]模型上开展验证,结果表明:相比于人工设置更新频率超参数,所提自适应策略可将正常准确率提升 0.80%～3.59%,鲁棒准确率提升 1.30%～8.50%;另外,由于人工调参方法本身已优于基线方法,而自适应策略不弱于甚至优于人工调参方法,因此证明了该方法对不同的数据集和模型具有良好的适应性。

1 相关工作

剪枝神经网络在安全关键场景的部署,受限于其对抗鲁棒性的显著下降。因此,在保障高稀疏度的同时强化对抗鲁棒性,已成为当前该领域的核心挑战。近期研究初步验证了该方向的可行性[15-19],代表性工作包括:ℓ_1 filter[16]使用 ℓ_1 范

数评估滤波器重要性的剪枝方法，使用 ℓ_2 范数几何中位数进行重要性评估的 FPGM[17]，通过对抗训练优化参数重要性得分的 HYDRA[18]，以及基于对抗损失贡献评估滤波器重要性的 FRFP(FRE-based robustness-aware filter pruning, FRFP)[15]。

尽管已取得上述成果，但仍存在 2 大关键局限：其一，现有工作[18-19]主要聚焦非结构化剪枝，其不规则稀疏模式需依赖专用硬件支持实现加速，严重制约实际应用价值，而针对结构化剪枝与对抗鲁棒性的协同优化机制，尚未形成系统探索；其二，多数方法因引入复杂优化流程[15]产生巨额计算开销，难以适配移动设备等资源受限场景。

此外，学术界还尝试从其他维度探索提升神经网络的鲁棒性，主要方法包括梯度掩码、对抗去噪(含输入特征重建、压缩等)、附加网络 3 类。针对梯度掩码类方法，ATHALYE 等[20]在 ICML 2018 会议上证明，7 种发表于 ICLR 2018 的混淆梯度相关防御方法均存在安全漏洞，仅保留了投影梯度下降(projected gradient descent, PGD)相关防御的有效性，直接否定了混淆梯度方法在对抗样本防御中的可行性。针对对抗去噪类方法，JIA 等[21]提出了一种端到端的图像压缩模型 ComDefend，由压缩卷积神经网络(ComCNN)和重构卷积神经网络(RecCNN)组成：ComCNN 通过压缩原始图像消除对抗扰动，RecCNN 则对原始图像进行高质量的重建，以此实现对抗样本防御。针对附加网络的方法，以 MagNet[22]为例，该防御框架不依赖对抗样本及其生成过程，也不修改原始模型，仅利用输入数据的特征，由探测器和重组器组成。基于深度学习的流行假设，对抗样本远离或位于流行边界。探测器检测远离流行边界的对抗性样本，并拒绝对其分类，然后通过重组器来重构这些对抗样本，模型从而将接近流行边界的样本重构为原始样本进行分类。但是，上述方法都无法满足模型轻量化部署需求，仅实现了鲁棒性的提升。

2　结构化对抗鲁棒剪枝

2.1　问题定义

将 L 层神经网络看作由 θ 参数化的函数 $f(\cdot)$。为模拟训练过程中剪枝的效果，使用由 0－1 矩阵序列组成的张量 $\boldsymbol{M}=\{m_1,m_2,\cdots,m_l\}$ 作为滤波掩膜。第 l 层卷积层的滤波掩膜集合记为 $m_l=\{m_{l,j}\}_{j=1}^{C_{\text{out}}}=1$；对于第 i 层的第 j 个滤波器，将其对应的滤波器掩码记为 $m_{i,j}\in\mathbf{R}^{C_{\text{in}}\times K\times K}$。其中，$C_{\text{in}}$ 和 C_{out} 分别为第 l 个卷积层的输入和输出通道数，K 表示核的大小。结构化对抗鲁棒剪枝可建模为如下优化问题：

$$\begin{cases}\min\limits_{\theta} E_{(x,y)\sim D}\left[L_{\text{adv}}(f(x_{\text{adv}};\theta\odot\boldsymbol{M}),y)\right]\\ \text{s.t. }\|\theta\odot\boldsymbol{M}\|_0\leqslant\gamma,\boldsymbol{M}\in\{0,1\}^N\end{cases}\tag{1}$$

式中，θ 表示网络的参数；N 为整个网络中滤波器的总数；$\odot$ 表示元素级乘法；(x,y) 为从数据集 D 中采样的数据对和标签；x_{adv} 为由 x 生成的对抗样本；L_{adv} 为对抗训练损失；γ 为用户自定义的容量预算，用于表征剪枝后剩余参数的数量。

2.2　基本流程

针对鲁棒预训练神经网络，给出指定的稀疏度，基于滤波器鲁棒性评估(filter robustness estimation, FRE)的鲁棒感知滤波器剪枝 FRFP[15]，可在满足稀疏度要求的前提下，最大化模型的正常准确率和鲁棒准确率。该方法包括以下步骤：

步骤 1　对抗样本生成。在每一轮训练中，从数据集中随机抽取 1 个小批量样本，采用对抗样本生成技术构建对应的对抗样本。此步骤旨在使模型充分学习数据的各种变化和噪声。

步骤 2　对抗训练损失计算。将生成的对抗样本输入模型进行前向传播，计算对抗训练损失。该损失函数会鼓励模型在面对对抗样本时保持准确性，从而提高鲁棒性。

步骤 3　模型参数更新。基于对抗训练损失的梯度，执行后向传播并更新模型参数，不断优化模型性能。

步骤 4　FRE 评估。每间隔固定训练周期(设为 k 轮)，对模型所有滤波器计算其 FRE 分数，按分数排序。FRE 分数的定义为：

$$F_{\text{FRE}_{i,j}}=\frac{1}{\|\boldsymbol{\omega}_{i,j}\|_0}\|\omega_{i,j}\odot\nabla_{\omega_{i,j}}L_{\text{adv}}\|_2^2\tag{2}$$

式中，$\omega_{i,j}$ 表示网络中第 i 层的第 j 个滤波器的权重参数；$\odot$ 表示元素级乘法；L_{adv} 表示对抗训练损失。

步骤 5　滤波器剪枝。根据 FRE 分数排序结果，剪枝分数最低的滤波器，具体通过将对应卷积核的掩码置零来实现。

步骤 6　增量式稀疏度收敛。重复步骤 1～5，逐渐增加剪枝比例，直到达到所需的稀疏度水平。

本文将 FRFP 作为基线方法，通过对比所提自适应训练-剪枝频率调整方法与 FRFP 实验结果，验证所提方法在提升结构化鲁棒剪枝性能方面的优势。算法流程如图 1 所示。

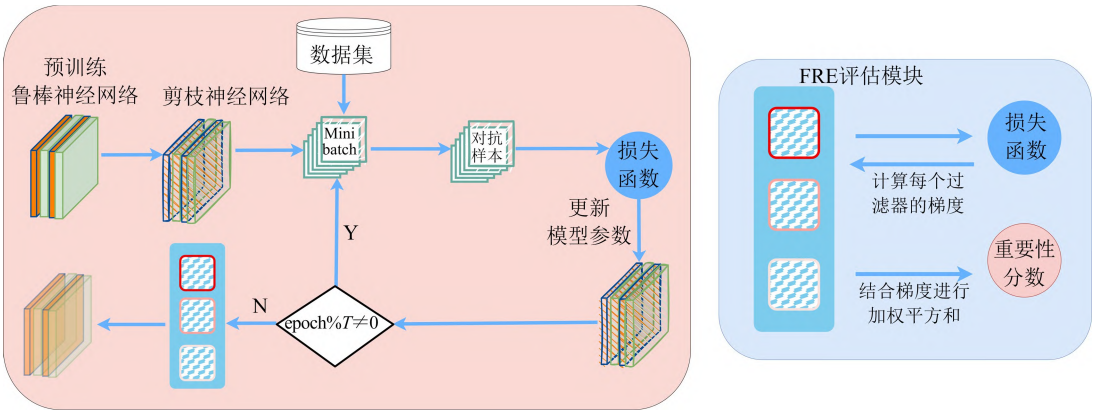


图 1 结构化对抗鲁棒剪枝算法流程图

Fig. 1 Flowchart of structured adversarial robust pruning algorithm

2.3 迭代式结构化对抗鲁棒剪枝方法

如前所述，在对抗鲁棒剪枝的过程中，既需要保持足够的稀疏度、准确率和鲁棒性，又要确保模型收敛，需要权衡才能达成这个目标。本节将详细描述该方法如何实现模型准确率和鲁棒性的权衡，具体见算法 1。

算法 1 迭代式结构化对抗鲁棒剪枝方法

1. 输入：训练数据集 D，预训练的鲁棒神经网络 $f(\cdot;\theta)$，目标稀疏度预算 τ，剪枝间隔 T，稀疏度步长 φ，随机梯度下降(stochastic gradient descent，SGD)，学习率 η

2. 输出：剪枝后的鲁棒神经网络 $f(\cdot;\theta\odot\boldsymbol{M})$

3. 初始化：epoch←0；稀疏度←0；$\boldsymbol{M}\leftarrow\{1\}^{N}$

4. 对于 epoch$<E$ 且稀疏度 $k<\tau$ 时：

1) 从 D 中抽取 1 个小批量样本 $\{(x_i,y_i)\}_{i=1}^{m}$

2) 生成对抗样本 $\{((x_{\text{adv}})_i,y_i)\}_{i=1}^{m}$

3) 前向传播：计算对抗训练损失 $L_{\text{adv}}(f((x_{\text{adv}})_i;\theta\odot\boldsymbol{M}),y)$

4) 后向传播：获取对抗训练损失的梯度

5) 计算重要性分数 FRE

6) 更新模型参数

7) 若 epoch$\%T=0$：

根据重要性对所有的滤波器进行排序

稀疏度←稀疏度$+\varphi$

通过更新 $\boldsymbol{M}$ 来修剪最不重要的滤波器

该算法的核心在于探寻配适模型和数据集的参数更新频率和剪枝掩码更新频率。本实验尝试采用 1∶1、1∶5、1∶2、2∶2、2∶1 等多种参数配比，具体实验效果评估将在 3.2 节中讨论。

2.4 基于 EE 的自适应结构化对抗鲁棒剪枝方法

对于 2.3 节所叙述的迭代式结构化对抗鲁棒剪枝方法，其核心挑战在于剪枝掩码更新和参数微调的频率需人工调整：为确定最优频率配比，需消耗大量计算资源，显著阻碍了该方法在不同神经网络模型与数据集间的迁移适配。为了能够自动寻找最优的频率配置，适应不同的模型和数据集特性，本文引入了 EE 机制，具体采用多臂老虎机(multi-armed bandit，MAB)范式。EE 的权衡是强化学习领域的核心问题[24-25]，简单来说，它描述的是一个学习智能体在尝试新事物(探索)和依据已有知识做决策(利用)之间如何平衡。

人工频率调参本质上属于离线网格搜索，计算成本高昂且无法适应训练动态。EE 机制的核心优势在于其在线学习和自适应平衡能力：无需预先进行大量实验，可以在模型训练过程中持续评估不同频率配置的效果，智能权衡“探索更好的新频率配置”和“利用当前最优频率配置”，避免陷入局部最优或错过更好的配置，这对处理未知模型/数据集尤其重要。

本文将剪枝/微调频率配置视为 MAB 问题中的“臂”。选择 MAB 框架的依据为：频率配置通常是离散且有限的(符合“臂”的概念)；MAB 的目标是快速、在线地找到能最大化累积训练收益(如模型性能提升)的臂；MAB 算法支持高效处理探索-利用权衡，计算开销相对较小，适合嵌入到模型训练循环中。相较于更复杂的强化学习算法(例如，需要学习值函数的 Q-learning)，MAB 可提供简洁高效的解决方案。此外，本文在初始化阶段设置多个配置点以加速探索过程，并降低对单一初始点的依赖。

具体来说，设计基于奖励驱动的 MAB 机制以自适应调整候选频率配置：

1) 动作空间。维护含 K 个(如 $K=3$)候选的剪枝/微调频率配置方案的集合 $C=\{c_1,c_2,$

…，c_k｝。每个配置 c_i 定义了参数更新频率和掩码更新频率的具体参数。

2）奖励函数。评估配置 c_i 优劣的关键是定义量化其近期“表现”的奖励 R_i。在每个剪枝间隔结束时计算奖励，采用配置 c_i 在最近 V 个 epoch 内的验证提升幅度作为奖励：

$$R_i=\frac{(A_{\text{Acc,current}}-A_{\text{Acc,baseline}})}{V} \tag{3}$$

式中，$A_{\text{Acc,current}}$ 为当前评估时刻的验证精度，$A_{\text{Acc,baseline}}$ 为 V 个 epoch 前的验证精度。除以 V 是为了粗略标准化，使奖励反映平均每 epoch 的精度增益。

3）策略更新。在每个评估点，根据收集的奖励更新配置选择策略。本文采用 ε-greedy 策略，具体为：以概率 $1-\varepsilon$，选择当前平均奖励最高的配置，即利用；以概率 ε，随机选择一个配置（含最优配置）进行尝试，即探索。ε 为固定值，随时间衰减（$\varepsilon=1/(\sqrt{t})$，t 为评估轮次）。

4）配置更新。ε-greedy 策略主要作用于配置选择，并不直接修改配置点本身的值。对配置点的调整方式为：定期对表现最优的配置施加小幅扰动生成新配置，将其加入候选集并替换长期表现最差的配置；对配置点的调整直至所有配置点满足 $c_k-c_m<\mu$（μ 为最小更新单位）或训练结束。

具体的算法设计如算法 2 所示。

算法 2　基于 EE 的自适应结构化对抗鲁棒剪枝方法

1. 输入：训练数据集 D，预训练的鲁棒神经网络 $f(\cdot;\theta)$，目标稀疏度预算 τ，剪枝间隔 T，稀疏度步长 φ，SGD 学习率 η

2. 输出：剪枝后的鲁棒神经网络 $f(\cdot;\theta\odot\boldsymbol{M})$

3. 初始化：epoch←0；稀疏度←0；$\boldsymbol{M}\leftarrow\{1\}^N$，设置 3 个初始参数比例 c_1,c_2,c_3

4. 任意选择一个参数比例 c_i 作为初始参数配置点

5. 当 epoch$<E$ 且稀疏度 $k<\tau$ 时：

1）从 D 中抽取一个小批量样本$\{(x_i,y_i)\}_{i=1}^m$

2）生成对抗样本$\{((x_{\text{adv}})_i,y_i)\}_{i=1}^m$

3）前向传播：计算对抗训练损失 $L_{\text{adv}}(f((x_{\text{adv}})_i;\theta\odot\boldsymbol{M}),y)$

4）后向传播：获取对抗训练损失的梯度

5）计算重要性分数

6）更新模型参数，更新 $R_i=\{R_1,R_2,R_3\}$

7）若 epoch$\%T=0$：

根据重要性对所有的滤波器进行排序：

稀疏度←稀疏度$+\varphi$

通过更新 $\boldsymbol{M}$ 来修剪最不重要的滤波器

8）若 epoch$\%V=0$ 且 $c_1\neq c_2\neq c_3$：

根据 R_i 选择下 V 轮的参数，具体策略为：

若 random()$<\varepsilon$(epoch)：

next_arm_idx＝random_int(1,3) # 随机选一个臂

否则：

next_arm_idx＝argmax(R) # 选当前平均奖励最高的臂

9）若 epoch$\%3V=0$：

根据 R_i 更新 c_1,c_2,c_3，具体策略为：

丢弃平均奖励最低的配置 c_m，用 $c_{\text{argmax(R)}}\pm\mu$ 代替 c_m

3　实验论证

3.1　实验设置

1）数据集。在 CIFAR-10[9]、CIFAR-100[9]、SVHN[10]数据集上展开实验。本文利用测试集中的正常样本生成对抗样本，并进行相应的结构化剪枝对抗训练。

2）模型结构。保证实验中的模型结构和基线方法文献中所采用的模型结构一致，本实验采用 VGG-16[11]，ResNet-18[14]和 MobileNetV1[12]结构进行算法实践。

3）攻击方法。使用 10 轮迭代的$\epsilon=8/255$ 的 PGD[8]作为攻击算法；在补充实验中，使用$\epsilon=8/255$ 的动量迭代快速梯度符号法（momentum iterative fast gradient sign method，MI-FGSM）[23]，多样化输入迭代快速梯度符号法（diverse input iterative FGSM，DI-FGSM）[24]，自动化攻击（AutoAttack）[25]，Carlini-Wagner（C&W）[26]攻击。

4）测试指标。① 模型准确率 Acc（accuracy）：由模型在测试集上的推断结果进行度量；② 稀疏度（sparsity）：模型被剪枝参数占模型全部参数的比例；③ 鲁棒准确率 Rac（robust accuracy）：通过利用$\epsilon=8/255$ 的 PGD-10 攻击对测试数据生成的对抗样本进行推断来度量。

3.2　迭代式结构化对抗鲁棒剪枝方法实验结果

3.2.1　CIFAR-10 数据集实验结果

在 CIFAR-10 数据集上进行实验以验证本文方法的有效性，实验结果见表 1 所列。由表 1 可知，本文方法在不同稀疏度的不同参数比例设置下均展现出良好效果。在表现最优的 ResNet-18 模型上，相比 baseline 方法，准确率和鲁棒性均有

提升。在 0.5 和 0.7 的稀疏度下，本文方法的 Rac 最多提升了 10.32%；在 0.9 的极端稀疏度下，重新训练的剪枝网络仍保持较高的鲁棒性，在 FRFP 网络上，Acc 提高了 4.76%，Rac 约提高 15.52%。在 VGG-16 网络上，基于 1∶1、1∶2 参数配比设计的实验结果也有小幅度提升，在 0.5 的常规稀疏度且 Acc 波动不大的前提下，Rac 最多提升了 5.24%的。

表 1 自适应结构化对抗鲁棒剪枝方法在 CIFAR-10 数据集上的实验结果

Tab. 1 Experimental results of adaptive structured adversarial robust pruning on CIFAR-10 /%

网络结构	方法	预训练模型		稀疏度=0.5		稀疏度=0.7		稀疏度=0.9	
		Acc	Rac	Acc	Rac	Acc	Rac	Acc	Rac
VGG-16	ℓ_1 filter	80.80	44.19	77.24	42.90	71.82	35.67	42.03	22.94
	HYDRA			75.56	40.87	51.03	27.15	12.56	3.80
	FPGM			75.03	40.16	45.32	24.08	13.20	9.28
	FRFP			79.43	45.79	79.33	46.08	**74.72**	41.19
	Ours(Manual)			79.60	51.03	79.20	**49.35**	72.60	42.12
	Ours(Adaptive)			**82.00**	**55.22**	**80.80**	48.60	72.40	**45.38**
ResNet-18	ℓ_1 filter	80.34	49.50	77.56	45.02	60.84	32.25	41.02	23.78
	HYDRA			77.23	44.88	57.89	30.82	50.87	26.85
	FPGM			76.33	42.53	48.21	24.57	15.66	8.09
	FRFP			80.70	48.54	79.00	46.25	69.22	37.24
	Ours(Manual)			81.33	**58.86**	79.10	53.86	**73.98**	**52.76**
	Ours(Adaptive)			**82.50**	58.30	**79.90**	**62.36**	73.60	52.13
MobileNetV1	ℓ_1 filter	75.07	40.94	70.10	37.22	60.24	30.26	31.15	16.73
	HYDRA			72.22	38.61	70.01	35.56	57.81	26.35
	FPGM			71.75	38.05	53.96	28.01	42.79	21.05
	FRFP			73.76	39.65	72.05	38.32	64.48	29.56
	Ours(Manual)			74.01	42.31	72.62	39.35	65.69	**34.05**
	Ours(Adaptive)			**74.36**	**42.90**	**72.76**	**40.03**	**66.26**	33.97

3.2.2 CIFAR-100 数据集实验结果

在 CIFAR-100 数据集上同样进行实验验证本文方法的有效性，实验结果见表 2 所列。可以看出，本文方法在不同稀疏度的不同参数比例设置下仍然展现良好效果。在 ResNet-18 模型上，仍然取得了优于 baseline 较多的表现，在 0.9 的极端稀疏度和 1∶2 的参数配比下，Rac 为 26.12%，较 FRFP 方法提升了 13.67%；此外，针对 VGG-16 网络也有不错的表现。

图 2～3 分别为迭代式结构化对抗鲁棒剪枝方法在 CIFAR-10 和 CIFAR-100 数据集上的实验结果。可以看出，不同模型结构与稀疏度下，参数更新的最优频率存在显著差异，难以确定一个适用于所有稀疏度和模型结构的统一更新频率，这表明设计一种自适应方法以自动探索最优参数配比具有重要研究价值。

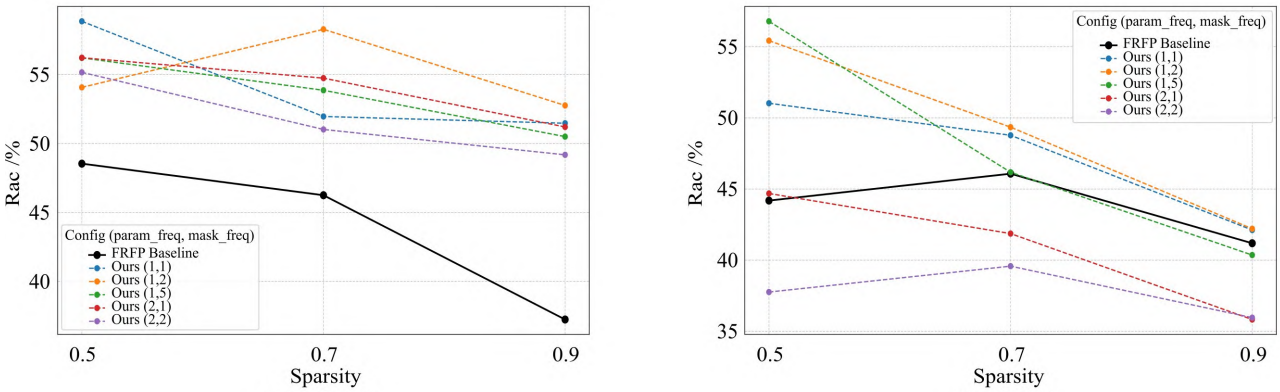
表 2 自适应结构化对抗鲁棒剪枝方法在 CIFAR-100 数据集上的实验结果

Tab. 2 Experimental results of adaptive structured adversarial robust pruning on CIFAR-100 /%

网络结构	方法	预训练模型		稀疏度=0.5		稀疏度=0.7		稀疏度=0.9	
		Acc	Rac	Acc	Rac	Acc	Rac	Acc	Rac
VGG-16	ℓ_1 filter	51.88	22.16	47.2	18.93	32.15	14.07	15.33	8.06
	HYDRA			1.05	0.65	1.11	0.83	1.02	0.91
	FPGM			43.22	20.03	30.84	15.21	18.28	8.33
	FRFP			51.02	22.65	49.28	21.17	35.83	15.39
	Ours(Manual)			52.20	26.81	48.70	21.03	35.00	16.43
	Ours(Adaptive)			**52.80**	**27.07**	**50.70**	**25.88**	**36.00**	**18.57**

续表

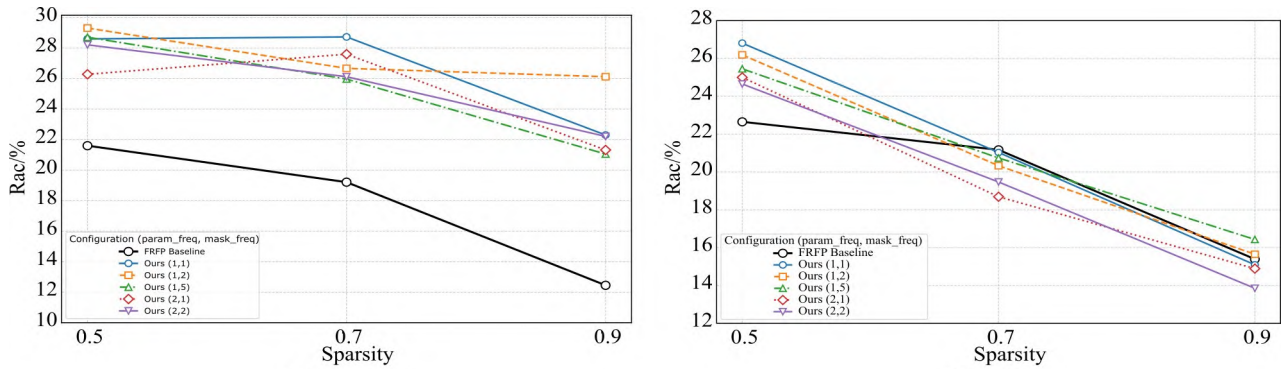
网络结构	方法	预训练模型		稀疏度＝0.5		稀疏度＝0.7		稀疏度＝0.9	
		Acc	Rac	Acc	Rac	Acc	Rac	Acc	Rac
ResNet-18	ℓ_1 filter	55.20	25.91	47.92	21.01	35.96	13.88	6.02	3.11
	HYDRA			49.03	20.85	35.57	15.01	18.02	8.36
	FPGM			39.75	20.65	33.12	17.64	19.07	9.12
	FRFP			49.98	21.60	45.31	19.21	29.96	12.45
	Ours(Manual)			54.10	28.60	50.31	28.73	**41.67**	**26.12**
	Ours(Adaptive)			**57.00**	**34.05**	**53.90**	**30.03**	35.78	22.80
MobileNetV1	ℓ_1 filter	47.36	18.82	35.61	14.33	30.05	13.71	19.22	9.06
	HYDRA			39.12	15.17	32.25	11.05	20.04	8.20
	FPGM			40.01	14.36	35.06	13.68	28.67	11.05
	FRFP			41.96	15.05	36.51	13.07	30.03	10.94
	Ours(Manual)			42.15	15.87	36.68	13.24	**31.00**	11.04
	Ours(Adaptive)			**44.06**	**16.11**	**38.61**	**13.78**	30.96	**11.55**



(a) ResNet-18　　(b) VGG-16

图 2　迭代式结构化对抗鲁棒剪枝方法在 CIFAR-10 数据集上的实验结果

Fig. 2　Experimental results of iterative structured adversarial robust pruning on CIFAR-10 dataset



(a) ResNet-18　　(b) VGG-16

图 3　迭代式结构化对抗鲁棒剪枝方法在 CIFAR-100 数据集上的实验结果

Fig. 3　Experimental results of iterative structured adversarial robust pruning on CIFAR-100 dataset

3.3　基于 EE 的自适应结构化对抗鲁棒剪枝方法实验结果

3.3.1　CIFAR-10 数据集实验结果

针对自适应结构化对抗鲁棒剪枝方法，其在 CIFAR-10 数据集上的实验结果见表 1 所列。其中，Ours(Manual)表示通过人工调参确定最优实验配比得到的结果，Ours(Adaptive)表示自适应结构化对抗鲁棒剪枝的实验结果。实验结果表明，自适应方法在 VGG-16，ResNet-18 和 MobileNetV1 这 3 种网络模型上的性能均与原方法最优实验结果相近，在 ResNet-18 网络 0.7 稀疏度下，该方法的 Rac 较人工调参方法提高了 8.50%。

3.3.2 CIFAR-100 数据集实验结果

在 CIFAR-100 数据集上进行测试，结果见表 2 所列。在 ResNet-18 网络上实验表明，该方法相较人工调参方法性能显著提升：在 0.5 稀疏度下，Rac 甚至提升至 34.05%，升幅达 5.45%，充分说明了自适应方法的有效性。

3.3.3 SVHN 数据集实验结果

在 SVHN 数据集上进行测试，结果见表 3 所列。实验表明，该方法在 ResNet-18、VGG-16 和 MobileNetV1 3 种模型上的性能不仅优于人工调参方法，较其他常见剪枝方法也有较大的提升，充分说明了本文方法的有效性。

表 3 自适应结构化对抗鲁棒剪枝方法在 SVHN 数据集上的实验结果

Tab. 3 Experimental results of adaptive structured adversarial robust pruning on SVHN /%

网络结构	方法	预训练模型		稀疏度=0.5		稀疏度=0.7		稀疏度=0.9	
		Acc	Rac	Acc	Rac	Acc	Rac	Acc	Rac
VGG-16	ℓ_1 filter	90.72	54.64	87.06	51.02	86.38	46.82	16.71	12.28
	HYDRA			66.79	24.06	61.97	27.32	59.98	19.31
	FPGM			88.05	53.23	87.11	47.54	70.06	28.31
	FRFP			90.40	55.57	90.58	55.05	88.74	51.82
	Ours(Manual)			90.66	55.33	90.62	55.16	**89.13**	52.94
	Ours(Adaptive)			**91.03**	**55.68**	**90.79**	**55.53**	89.09	**53.11**
ResNet-18	ℓ_1 filter	94.72	54.23	91.97	34.78	90.26	42.55	60.21	18.33
	HYDRA			86.81	44.36	85.13	41.63	73.19	32.76
	FPGM			79.22	40.63	66.80	33.18	52.73	23.36
	FRFP			94.65	51.88	93.39	**54.24**	93.15	47.13
	Ours(Manual)			94.35	50.97	93.63	50.26	92.86	**48.33**
	Ours(Adaptive)			**94.81**	**51.36**	**93.61**	51.07	**93.35**	48.06
MobileNetV1	ℓ_1 filter	88.75	52.16	85.18	46.26	84.07	43.38	70.61	33.85
	HYDRA			83.67	45.38	82.07	41.15	77.22	38.60
	FPGM			79.23	43.86	69.37	40.67	60.82	35.02
	FRFP			85.87	47.01	84.62	44.65	79.36	42.06
	Ours(Manual)			86.02	**48.25**	85.30	46.21	80.13	43.22
	Ours(Adaptive)			**86.31**	47.94	**85.36**	**46.55**	**81.26**	**43.57**

3.4 消融实验及补充实验

3.4.1 剪枝幅度消融实验

本文进一步探索了调整增加剪枝幅度(即每次执行剪枝算法时剪枝掉的卷积核数)的频率对于本文方法的影响，相关结果如图 4 所示。特别地，当剪枝幅度调整频率设为 5 时，在 0.7 稀疏度、参数配比 1∶2 的条件下，得到 58.3%的鲁棒准确率，为所有设置相同稀疏度下的最高鲁棒 性；当剪枝幅度调整频率设为 8 时，在 0.9 极端稀疏度、参数配比 1∶1 的条件下，模型的鲁棒准确率达 54.5%，同样为所有设置相同稀疏度下的最高鲁棒性。由此可知，增加剪枝幅度的频率对实验结果有显著影响。

3.4.2 自适应方法消融实验

本文还探索了自适应方法对于实验结果的影响，如图 5 所示，特别突出了自适应方法在关键稀疏度(0.7)下的显著优势，充分说明了自适应方法的有效性。

3.4.3 鲁棒性攻击补充实验

本文进一步测试了不同剪枝方法在各种对抗攻击下的性能表现，结果如图 6 所示，分别用 MI-FGSM、DI-FGSM、AutoAttack、C&W 和 PGD 对各种防御模型进行攻击，结果表明，本文所提出的方法对 PGD 攻击抵抗力最强，并且在各稀疏度下表现较优，在 0.9 高稀疏度下，对所有攻击，Rac 均大于 40% ，鲁棒性最优，这充分说明了本文方法的有效性和全面泛化性。

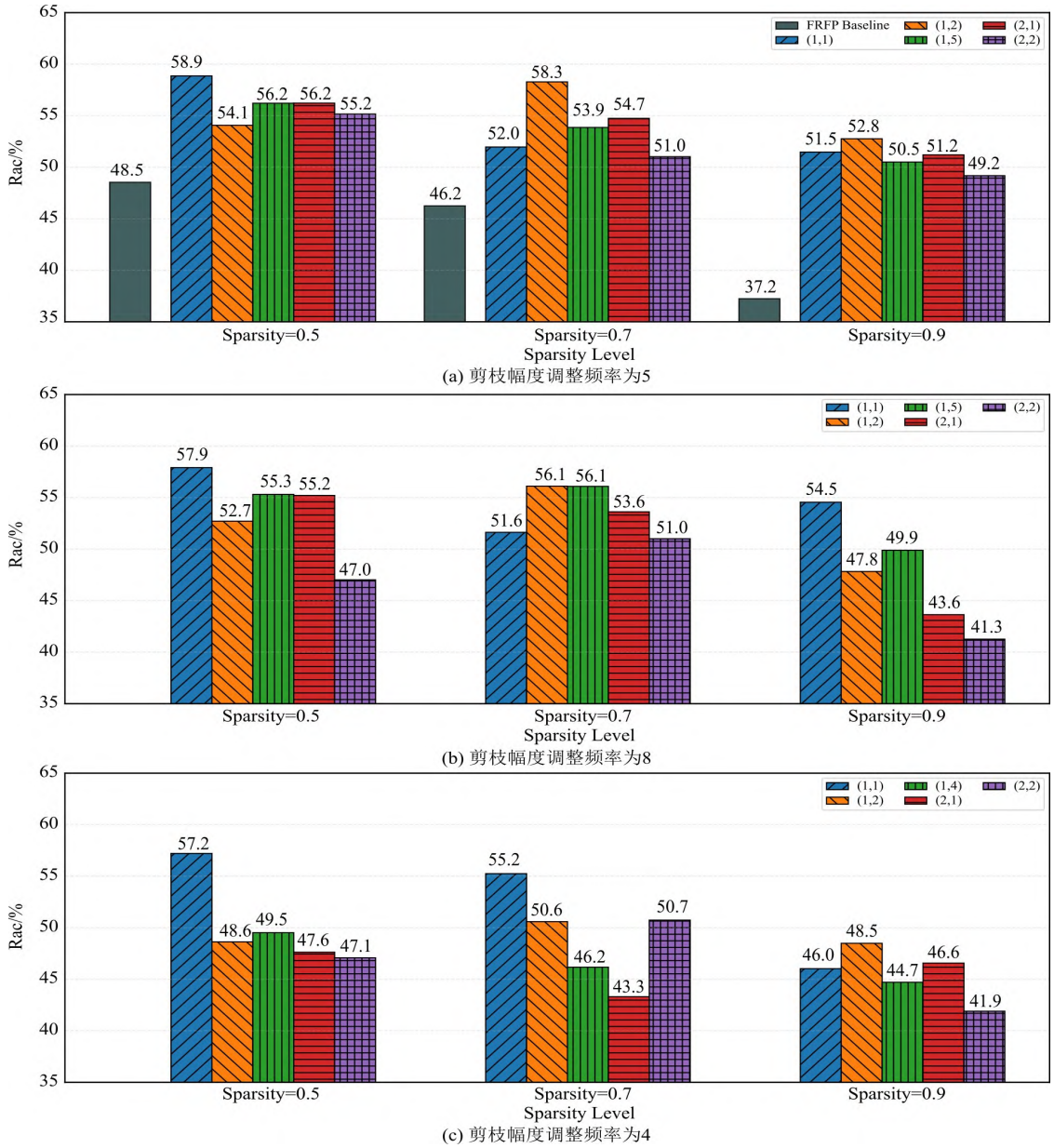


图 4　在 ResNet-18 上关于增加剪枝幅度频率的消融实验

Fig. 4　Ablation experiments on increasing of pruning magnitude frequency on ResNet-18

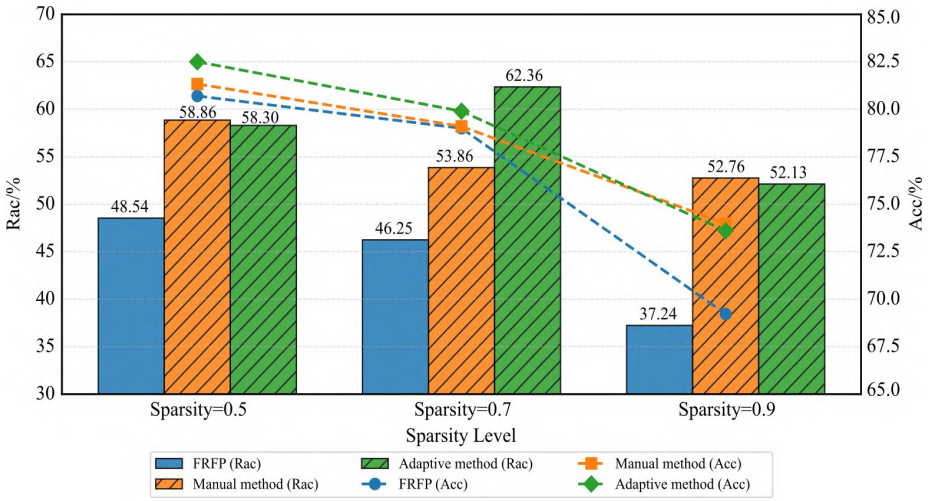


图 5　在 ResNet-18 上关于自适应方法的消融实验

Fig. 5　Ablation experiments on adaptive method on ResNet-18

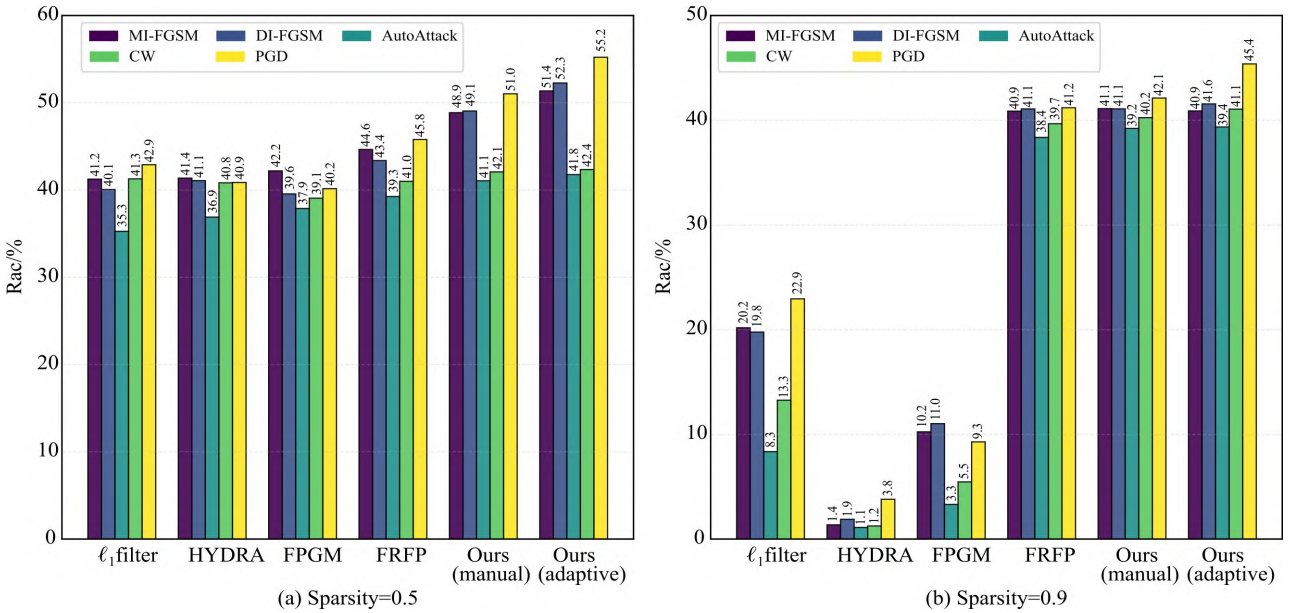


图 6 在 VGG-16 上关于鲁棒性攻击的补充实验

Fig. 6 Supplementary experiments on robust attacks on VGG-16

3.5 方法对比

迭代式结构化对抗鲁棒剪枝采用固定频率交替更新模型参数和剪枝掩码，其优势在于逻辑简洁、易于实现；但该方法需手动预设参数配比（如 1∶1 或 2∶1），且依赖大量实验调参。这不仅导致计算成本高，还使其在迁移到新模型或数据集时需重复调优，灵活性较低。因此，该方法更适合以下场景：

1）资源充足的固定任务。如实验室环境下对单一模型（如 CIFAR-10 数据集上的 ResNet-18）进行深度优化。

2）快速原型验证。需在短时间内验证基础剪枝效果，且对方法泛化性无明确要求。

3）人工调参经验丰富的场景。可凭经验快速定位较优配比，降低调参成本。

基于 EE 策略的自适应结构化对抗鲁棒剪枝方法，通过强化学习中的 EE 机制，动态优化参数配比。具体而言，该方法初始设置 3 种配比方案，随后定期根据精度表现淘汰次优方案，最终收敛至最优配比。其优势在于自动化程度高，能显著减少人工干预，且跨任务适应性强；但存在实现复杂度较高、初始方案选择对收敛效率存在影响的局限，因此更适合以下场景：

1）工业级部署场景。需适配多种模型架构（如移动端轻量模型与云端大模型）或动态更新的数据集。

2）长期自动化流水线。如 AI 平台需自动输出剪枝模型，以减少人工调参的依赖。

3）计算资源受限但需平衡效果与效率的场景。因其可避免遍历性实验，在资源受限下实现性能与效率的兼顾。

4 结束语

本文在深度神经网络剪枝和对抗鲁棒性研究方面取得了进展，为深度学习模型在安全敏感场景和资源受限环境中的部署应用提供了新的思路和方法。尽管所提方法已通过实验验证其有效性，但仍需在不同场景下进一步开展研究和验证，以认证其有效性、通用性和有效性。未来，随着深度神经网络技术的不断发展和完善，其实际应用价值日益凸显，有望为解决现实世界中的复杂问题提供更加有效的方案。

参 考 文 献

[1] LECUN Y, BENGIO Y, HINTON G. Deep learning [J]. Nature, 2015, 521(7553): 436-444.

[2] GALLIFANT J, FISKE A, LEVITES Y A, et al. Peer review of GPT-4 technical report and systems card[J]. PLOS Digital Health, 2024, 3(1): e0000417.

[3] RUSSAKOVSKY O, DENG J, SU H, et al. Imagenet large scale visual recognition challenge[J]. International journal of computer vision, 2015, 115(3): 211-252.

[4] LIN T Y, MAIRE M, BELONGIE S, et al. Microsoft COCO: common objects in c ontext[C]//Proceedings

of 2014 European Conference on Computer Vision. [S. l.]: Springer, 2014: 740-755.

[5] LITJENS G, KOOI T, BEJNORDI B E, et al. A survey on deep learning in medical image analysis[J]. Medical Image Analysis, 2017, 42: 60-88.

[6] HE K M, ZHANG X Y, REN S Q, et al. Deep residual learning for image recognition[C]//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. [S. l.]: IEEE, 2016: 770-778.

[7] YOU H R, LI C J, XU P F, et al. Drawing early-bird tickets: towards more efficient training of deep networks [C]//Proceedings of 2020 International Conference on Learning Representations. [S. l. : s. n.], 2020.

[8] CAO Y H, LI S Y, LIU X Y, et al. A survey of AI-generated content(AIGC)[J]. ACM Computing Surveys, 2025, 57(5): 1-38.

[9] SZEGEDY C, ZAREMBA W, SUTSKEVER I, et al. Intriguing properties of neural networks [C]// Proceedings of the 2nd International Conference on Learning Representations. [S. l. : s. n.], 2014: 1-10.

[10] MADRY A, MAKELOV A, SCHMIDT L, et al. Towards deep learning models resistant to adversarial attacks[C]//Proceedings of International Conference on Learning Representations. [S. l. : s. n.], 2018: 4138-4161.

[11] KRIZHEVSKY A, HINTONG. Learning multiple layers of features from tiny images[J]. Handbook of Systemic Autoim-mune Diseases, 2009, 1(4): 1-60.

[12] NETZER Y, WANG T, COATES A, et al. Reading digits in natural images with unsupervised feature learning[C]//Proceedings of NIPS Workshop on Deep Learning and Unsupervised Feature Learning. [S. l. : s. n.], 2011: 1-9.

[13] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition [C]// Proceedings of the 3rd International Conference on Learning Representations. [S. l. : s. n.], 2015: 301-307.

[14] SANDLER M, HOWARD A, ZHU M L, et al. MobileNetV2: inverted residuals and linear bottlenecks [C]//Proceedings of 2018 IEEE Conference on Computer Vision and Pattern Recognition. [S. l.]: IEEE, 2018: 4510-4520.

[15] ZHUANG X L, GE Y J, ZHENG B L, et al. Adversarial network pruning by filter robustness estimation[C]//Proceedings of 2023 IEEE International Conference on Acoustics, Speech and Signal Processing. [S. l.]: IEEE, 2023: 1-5.

[16] LI H, KADAV A, DURDANOVIC I, et al. Pruning filters for efficient convNets [C]//Proceedings of the 5th International Conference on Learning Representations. [S. l. : s. n.], 2017: 1683-1696.

[17] HE Y, LIU P, WANG Z W, et al. Filter pruning via geometric median for deep convolutional neural networks acceleration [C]//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. [S. l.]: IEEE, 2019: 4340-4349.

[18] SEHWAG V, WANG S Q, MITTAL P, et al. HYDRA: pruning adversarially robust neural networks [J]. Advances in Neural Information Processing Systems, 2020, 33: 19655-19666.

[19] MADAAN D, SHIN J, HWANG S J. Adversarial neural pruning with latent vulnerability suppression [C]//Proceedings of the 37th International Conference on Machine Learning. [S. l. : s. n.], 2020: 6575-6585.

[20] ATHALYE A, CARLINI N, WAGNER D. Obfuscated gradients give a false sense of security: circumventing defenses to adversarial examples[C]//Proceedings of the 35th International Conference on Machine Learning. [S. l. : s. n.], 2018: 274-283.

[21] JIA X J, WEI X X, CAO X C, et al. ComDefend: an efficient image compression model to defend adversarial examples[C]//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. [S. l.]: IEEE, 2019: 6084-6092.

[22] MENG D Y, CHEN H. MagNet: a two-pronged defense against adversarial examples[C]//Proceedings of 2017 ACM SIGSAC Conference on Computer and Communications Security. New York: ACM, 2017: 135-147.

[23] HAO J Y, YANG T P, TANG H Y, et al. Exploration in deep reinforcement learning: from single-agent to multiagent domain[J]. IEEE Transactions on Neural Networks and Learning Systems, 2024, 35(7): 8762-8782.

[24] ARULKUMARAN K, DEISENROTH M P, BRUNDAGE M, et al. Deep reinforcement learning: a brief survey [J]. IEEE Signal Processing Magazine, 2017, 34(6): 26-38.

[25] DONG Y P, LIAO F Z, PANG T Y, et al. Boosting adversarial attacks with momentum[C]//Proceedings of 2018 IEEE Conference on Computer Vision and Pattern Recognition. [S. l.]: IEEE, 2018: 9185-9193.

[26] XIE C H, ZHANG Z S, ZHOU Y Y, et al. Improving transferability of adversarial examples with input diversity[C]//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. [S. l.]: IEEE, 2019: 2730-2739.

[27] CROCE F, HEIN M. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks [C]//Proceedings of the 37th International Conference on Machine Learning. [S. l. : s. n.], 2020: 2206-2216.

[28] CARLINI N, WAGNER D. Towards evaluating the robustness of neural networks [C]//Proceedings of 2017 IEEE Symposium on Security and Privacy. [S. l.]: IEEE, 2017: 39-57.

作 者 简 介

曹瑞麒

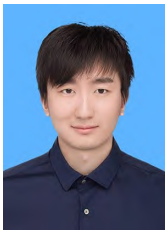
男，2002 年生，硕士研究生，研究方向为可信人工智能

E-mail: crq2002@stu. xjtu. edu. cn

杨雨龙

男，2000 年生，博士研究生，研究方向为对抗机器学习

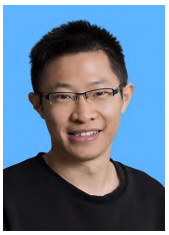
E-mail: yulongyang@stu. xjtu. edu. cn



蔺琛皓

男，1989 年生，教授，博士研究生导师，研究方向为人工智能安全、智能身份安全和 AI4Science

E-mail: linchenhao@xjtu. edu. cn



赵正宇

男，1992 年生，教授，博士研究生导师，研究方向为人工智能安全对抗

E-mail: zhengyu. zhao@xjtu. edu. cn

李 前

男，1992 年生，副教授，博士研究生导师，研究方向为可信人工智能与智能安全对抗

E-mail: qianlix@xjtu. edu. cn

王 骞

男，1980 年生，教授，博士研究生导师，研究方向为人工智能安全、云计算安全与隐私、无线系统安全、应用密码学

E-mail: qianwang@whu. edu. cn

沈 超

男，1985 年生，教授，博士研究生导师，研究方向为智能系统安全与控制、人工智能可信与安全、软硬件智能测试、大数据关联计算、人机交互行为分析

E-mail: chaoshen@mail. xjtu. edu. cn

责任编辑 董 莉