

基于剪枝技术和鲁棒蒸馏融合的轻量对抗攻击防御方法

王滨^{1,2}, 李思敏¹, 钱亚冠¹, 张君³, 李超豪², 朱晨鸣³, 张鸿飞³

(1. 浙江科技学院, 浙江 杭州 310023; 2. 浙江省多维感知技术应用与安全重点实验室, 浙江 杭州 310052;
3. 浙江省电子信息产品检验研究院, 浙江 杭州 310007)

摘要: 对抗训练是一类常用的对抗攻击防御方法, 其通过将对抗样本纳入训练过程, 从而有效抵御对抗攻击。然而, 对抗训练模型的鲁棒性通常依赖于网络容量的提升, 即对抗训练所获得的网络为防御对抗攻击而大幅提升网络的模型容量, 对其可用性造成较大约束。因此, 如何在保证对抗训练模型鲁棒性的同时, 降低模型容量, 提出轻量对抗攻击防御方法是一大挑战。为解决以上问题, 提出一种基于剪枝技术和鲁棒蒸馏融合的轻量对抗攻击防御方法。该方法以对抗鲁棒准确率为优化条件, 在对预训练的鲁棒对抗模型进行分层自适应剪枝压缩的基础上, 再对剪枝后的网络进行基于数据过滤的鲁棒蒸馏, 实现鲁棒对抗训练模型的有效压缩, 降低其模型容量。在 CIFAR-10 和 CIFAR-100 数据集上对所提出的方法进行性能验证与对比实验, 实验结果表明, 在相同 TRADES 对抗训练下, 所提出的分层自适应剪枝技术相较于现有剪枝技术, 其剪枝所得到的网络结构在多种 FLOPs 下均表现出更强的鲁棒性。此外, 基于剪枝技术和鲁棒蒸馏融合的轻量对抗攻击防御方法相较于其他鲁棒蒸馏方法表现出更高的对抗鲁棒准确率。因此, 实验结果证明所提方法在降低对抗训练模型容量的同时, 相较于现有方法具有更强的鲁棒性, 提升了对抗训练模型在物联网边缘计算环境的适用性。

关键词: 对抗防御; 剪枝; 鲁棒蒸馏; 轻量网络

中图分类号: TP393

文献标志码: A

DOI: 10.11959/j.issn.2096-109x.2022074

Lightweight defense mechanism against adversarial attacks via adaptive pruning and robust distillation

WANG Bin^{1,2}, LI Simin¹, QIAN Yaguan¹, ZHANG Jun³, LI Chaohao², ZHU Chenming³, ZHANG Hongfei³

1. Zhejiang University of Science and Technology, Hangzhou 310023, China

2. Zhejiang Key Laboratory of Multi-dimensional Perception Technology, Application and Cybersecurity, Hangzhou 310052, China

3. Zhejiang Electronic Information Products Inspection and Research Institute, Hangzhou 310007, China

收稿日期: 2022-04-24; 修回日期: 2022-06-11

通信作者: 钱亚冠, qianyg@yeah.net

基金项目: 国家自然科学基金 (92167203); 浙江省自然科学基金 (LZ22F020007)

Foundation Items: The National Natural Science Foundation of China (92167203), The Natural Science Foundation of Zhejiang Province (LZ22F020007)

引用格式: 王滨, 李思敏, 钱亚冠, 等. 基于剪枝技术和鲁棒蒸馏融合的轻量对抗攻击防御方法[J]. 网络与信息安全学报, 2022, 8(6): 102-109.

Citation Format: WANG B, LI S M, QIAN Y G, et al. Lightweight defense mechanism against adversarial attacks via adaptive pruning and robust distillation[J]. Chinese Journal of Network and Information Security, 2022, 8(6): 102-109.

Abstract: Adversarial training is one of the commonly used defense methods against adversarial attacks, by incorporating adversarial samples into the training process. However, the effectiveness of adversarial training heavily relied on the size of the trained model. Specially, the size of trained models generated by the adversarial training will significantly increase for defending against adversarial attacks. This imposes constraints on the usability of adversarial training, especially in a resource-constraint environment. Thus, how to reduce the model size while ensuring the robustness of the trained model is a challenge. To address the above issues, a lightweight defense mechanism was proposed against adversarial attacks, with adaptive pruning and robust distillation. A hierarchically adaptive pruning method was applied to the model generated by adversarial training in advance. Then the trained model was further compressed by a modified robust distillation method. Experimental results on CIFAR-10 and CIFAR-100 datasets showed that our hierarchically adaptive pruning method presented stronger robustness under various FLOP than the existing pruning methods. Moreover, the fusion of pruning and robust distillation presented higher robustness than the state-of-art robust distillation methods. Therefore, the experimental results prove that the proposed method can improve the usability of the adversarial training in the IoT edge computing environment.

Keywords: adversarial defenses; pruning; robust distillation; lightweight network

0 引言

随着深度学习技术在图像识别^[1]、语音识别^[2]、自然语言处理^[3]等领域的广泛应用,深度学习模型的安全直接决定了其所应用的系统或业务的安全。然而,研究指出了对抗样本攻击^[4,5]对于现有深度学习模型的攻击可行性,即通过在干净样本上添加微小的扰动,造成深度学习模型的功能错误。例如,CW(Carlina and Wagner)攻击可通过在图片上添加人眼不易察觉的对抗扰动,造成智能图像识别模型的分类错误^[5],如将停车标志识别为其他标志,可能会造成严重的安全事故^[6];同理,攻击者可通过在音频上叠加微小噪声,构造语音对抗样本,造成智能语音识别系统的错误识别,如将人耳所听的“你好”识别为“打开房门”等敏感指令^[7]。

为抵御来自对抗样本的安全威胁,对抗训练^[8-10]方法被提出并得到大量研究。对抗训练的核心是在模型训练阶段同时考虑干净样本和对抗样本,从而使训练获得的模型对对抗样本攻击具有较高的鲁棒性。然而,此类方法在提升训练模型的鲁棒性同时,通常会造成模型容量的大幅提升。此外,相关研究表明,模型容量与对抗训练效果成正相关性,即模型容量越大,对抗训练效果越好^[9]。然而,该特性对对抗训练的可用性造成较大的约束,如限制对抗训练在计算及存储资源相对受限的边缘设备(如智能手机、智能摄像

头等)上的应用部署。因此,亟须设计一种轻量对抗攻击防御方法,尤其在压缩对抗训练模型容量的同时,对模型的鲁棒性影响较小,从而提供轻量级的鲁棒模型,适应物联网边缘计算环境下的智能应用。

针对以上问题,研究者提出基于知识蒸馏(knowledge distillation)^[11]对抗训练模型优化方法,即将鲁棒知识从网络容量较大的教师网络迁移到轻量级的学生网络^[12-14]。例如,Goldblum等^[13]提出ARD(adversarially robust distillation)方法,将对抗样本参与模型的蒸馏,从而进一步提高模型的鲁棒性。在此基础上,Zi等^[14]提出了RSLAD(robust soft label adversarial distillation)方法,优化了蒸馏的损失函数,使模型的鲁棒性与经典对抗训练方法相比有所提高。但以上方法都是固定教师网络和学生网络的结构,无法进一步压缩网络容量以适应计算资源相对受限的边缘设备。

因此,本文提出一种基于剪枝技术和鲁棒蒸馏融合的轻量对抗攻击防御方法。首先基于分层自适应的剪枝策略对预训练的鲁棒对抗模型进行剪枝压缩,然后对剪枝后的网络进行基于数据过滤的鲁棒蒸馏,实现鲁棒对抗训练模型的有效压缩,降低其模型容量。

本文主要贡献如下。

1) 提出剪枝技术和鲁棒蒸馏融合的轻量对抗攻击防御方法,其在降低对抗训练模型网络容

量的同时,对于对抗训练模型的鲁棒性影响较小。

2) 改进现有剪枝技术,提出分层自适应剪枝技术,提升剪枝过后模型的鲁棒性。

3) 提出基于数据过滤的鲁棒蒸馏方法,通过对原始干净样本中被错误识别的样本进行过滤,提升鲁棒蒸馏的有效性。

1 相关工作

1.1 神经网络与对抗攻击

深度神经网络是关于 K 个类别数据集 $\mathcal{D} = \{X, Y\}$ 的映射函数 $\mathcal{F}: X \rightarrow Y$ 。 X 为样本集, Y 为与之对应的正确标签集合。网络可定义为 $\mathcal{F}(x; \theta)$, 可表达为

$$\mathcal{F}(x; \theta) = \text{softmax} \circ \mathcal{F}_n \circ \mathcal{F}_{n-1} \circ \cdots \circ \mathcal{F}_1 \quad (1)$$

其中, θ 代表网络的参数, $x \in X$ 为单个样本, $y_{\text{true}} \in Y$ 为单个样本所对应的正确分类标签。将数据集 $\mathcal{D} = \{(x_i, y_i)\}, i \in (1, 2, \dots, N)$ 中的样本 x_i 输入网络 $\mathcal{F}(x; \theta)$ 对其进行类别预测, 可得到

$$\mathcal{F}(x_i; \theta) = \arg \max_j [\text{softmax}(Z_j(x_i, \theta))] = y \quad (2)$$

其中, $Z(x_i, \theta)$ 代表除 softmax 层外, 网络关于输入 x_i 的输出, 称之为 logit。基于得到的 logit, 经过 softmax 变换后再得到网络对于输入的分类预测概率。当 $\mathcal{F}(x_i; \theta) = y = y_{\text{true}}$ 时, 样本类别预测正确; 当 $\mathcal{F}(x_i; \theta) = y \neq y_{\text{true}}$ 时, 样本类别预测错误。

当 $\mathcal{F}(x; \theta) = y_{\text{true}}$ 时, 存在微小扰动 δ , 使 $\mathcal{F}(x + \delta; \theta) \neq y_{\text{true}}$, 则称 $x' = x + \delta$ 为对抗样本^[4-5]。对抗样本攻击的主要目标之一是在用户不易察觉的情况下, 造成深度神经网络输出错误, 即将扰动 δ 作为优化约束。常见的对抗样本生成方法包括 CW^[5]、FGSM^[8]、PGD^[9]等。

1.2 对抗训练

近年来, 对抗训练作为一种常用的对抗样本攻击防御方法得到广泛的研究与应用^[15]。Madry 等^[9]提出标准对抗训练 (SAT, standard adversarial training), 其将对抗训练建模成双层优化问题, 并基于 PGD 攻击方法生成对抗样本训练模型。除 SAT 对抗训练框架之外, Zhang 等^[16]基于鲁棒性和精度的权衡考虑提出了 TRADES 框架。另外, 还有一类工作是利用没有标签的数据进行对抗训

练, 进一步提升模型的鲁棒性^[17-18]。上述方法在 ResNet 和 WideResNet 等大容量深度神经网络上均取得了较高的鲁棒性。此外, 研究发现对抗训练通常可使容量较大的网络获得更好的鲁棒性^[9]。然而, 对于部署到边缘环境中的轻量级神经网络, 直接对其进行对抗训练所能取得的鲁棒性效果不如较大网络显著。

1.3 知识蒸馏

为了使深度神经网络能够运行在计算和存储资源相对受限的边缘设备上, 研究人员提出了一系列神经网络压缩方法。其中, 知识蒸馏是一种常用的模型容量压缩方法, 其核心思想是让一个轻量级的学生网络学习比它性能更强的教师网络的知识。知识蒸馏通过教师网络输出的软标签 (概率向量) 训练学生网络, 从而实现教师网络和学生网络之间的知识迁移。Papernot 等^[12]首先提出防御性蒸馏机制, 通过利用知识蒸馏来提高深度神经网络的鲁棒性。然而, 该方法要求教师网络和学生网络的结构保持一致, 难以应用到轻量级网络上。此外, 相关研究指出已有对抗攻击 (如基于 ℓ_0 、 ℓ_2 和 ℓ_∞ 范数的 CW 对抗攻击^[5]) 可实现对该机制的有效攻破。除了防御性蒸馏之外, 研究人员进一步提出将对抗训练与知识蒸馏进行结合, 在鲁棒的预训练教师网络指导下, 提升学生网络的鲁棒性。Goldblum 等^[13]提出了对抗鲁棒蒸馏 (ARD, adversarial robust distillation) 方法, 其利用对抗样本将教师网络中鲁棒的知识迁移到学生网络中, 从而提升学生网络的鲁棒性。

2 轻量对抗攻击防御方法实现

为兼顾对抗训练模型的鲁棒性与可用性, 本文提出一种基于剪枝技术和鲁棒蒸馏融合的轻量对抗攻击防御方法。首先, 本文方法基于分层自适应的剪枝策略对预训练的鲁棒对抗模型进行剪枝压缩。然后, 对经过剪枝后的网络进行鲁棒蒸馏。区别于现有鲁棒蒸馏机制, 本文方法通过对原始干净样本中被错误识别的样本进行过滤, 提升鲁棒知识的高效迁移。剪枝技术和鲁棒蒸馏两类技术的有机结合, 既可实现鲁棒对抗训练模型的有效压缩, 降低其模型容量, 又能降低对模型鲁棒性的影响。下面对以上所述方法进行详细介绍。

2.1 分层自适应剪枝

剪枝技术的核心是将深度神经网络中冗余的卷积核或者通道去除。传统剪枝策略一般分为两类：预定义剪枝策略、基于全局阈值的剪枝策略。首先，预定义剪枝策略通常依赖于人工设定的剪枝率，但往往达不到最优的剪枝效果。其次，基于全局阈值的剪枝策略在使用中具有一定的局限性，如该方法会造成层崩塌，即该层的卷积核几乎被剪完，从而使模型性能骤降。针对以上问题，本文提出一种分层自适应的剪枝策略，其创新点在于依据模型各层输出的软标签信息与模型最终的软标签输出进行比较，进而指导各层剪枝率的分配（如图1所示），使网络结构能够适应鲁棒性的需要。

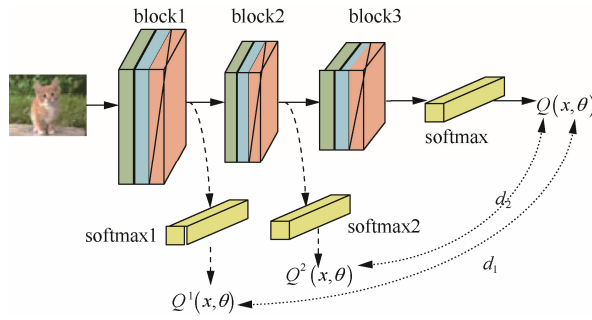


图1 剪枝分配准则
Figure 1 Pruning allocation guidelines

分层自适应剪枝策略的详细过程如下。

对于给定输入样本 x ， $Q(x, \theta)$ 为网络对输入 x 输出的最后一层的概率分布。将网络依照深度在不同层后面各添加一条计算概率向量的通路，并且分别记为 $Q^1(x, \theta)$ ， $Q^2(x, \theta)$ ， $\dots$ ， $Q^M(x, \theta)$ 。然后使用 KL 距离来衡量不同层输出概率分布差异，公式如下。

$$d_i = \text{KL}(Q^i(x, \theta), Q(x, \theta)) \quad (3)$$

此外，定义 d_i 为第 i 层后 softmax 输出的 $Q^i(x, \theta)$ 与模型最后一层概率分布的差异，接着分别计算 $d_1, d_2, \dots, d_M$ 的值，并将其归一化至 0 到 1 区间。然后对归一化后的数据做如下转换。

$$I_i = \frac{d_i}{d_1 + d_2 + \dots + d_M}, I_i \in [0, 1] \quad (4)$$

$$p_i = \frac{1}{I_i}, P_i = \frac{p_i}{p_1 + p_2 + \dots + p_M}, P_i \in [0, 1] \quad (5)$$

若 d_i 的值相对越小，代表第 i 层及之后的卷

积核或者通道越不重要，需在该层加大剪枝率。因此，分层自适应剪枝策略的优势在于对浅层网络进行的剪枝操作较少；而对冗余较多的深层特征所进行的剪枝操作则较多。假设在特定的 FLOP 预算下，共计要剪 N 个卷积核，且各层剪去的卷积核或者通道数量记为 $N_i (i \in 1, \dots, M)$ 。本文根据式(6)合理分配各层的剪枝率：

$$N_i = P_i \cdot N \quad (6)$$

在确定好各层的剪枝率之后，剪枝操作将与蒸馏训练过程交替进行。

2.2 基于数据过滤的鲁棒蒸馏

知识蒸馏的核心是将教师网络学到的知识迁移到更小的学生网络中，从而使学生网络的概率输出分布接近教师网络的概率输出分布，其形式可表达为

$$\min_{\theta_s} \mathcal{L}_{KD}(\theta_s) \quad (7)$$

$$\mathcal{L}_{KD}(\theta_s) = E_{X \sim \mathcal{D}} \text{KL}(Q(X, \theta_t), Q(X, \theta_s)) \quad (8)$$

其中， θ_t 和 θ_s 代表教师网络和学生网络， $Q(\cdot)$ 为网络 softmax 层之后输出的包含各类别的概率向量， $\text{KL}(\cdot)$ 为两个概率向量之间的 KL 距离， t 为蒸馏温度，其具体表现形式如下。

$$Q_j(x_i, \theta) = \frac{\exp(z_j(x_i, \theta)/t)}{\sum_{k=0}^{K-1} \exp(z_k(x_i, \theta)/t)} \quad (9)$$

其中，蒸馏温度 t 的使用可以使概率分布更为平滑，且能够更多地保留各类别之间的差异信息。

现有鲁棒蒸馏方法在原有知识蒸馏方法的基础上，基于干净数据集 X 生成对应的对抗样本数据集 X' ，然后将 X' 输入学生网络进行训练，最终使学生网络对于对抗样本 X' 的输出逐渐逼近教师网络对于 X' 所对应的干净样本 X 的高置信度输出，进而提高分类正确率，即模型对对抗样本的防御能力。尽管如此，现有方法中通常并未考虑到教师网络对于原有干净样本的分类正确率。具体而言，尽管教师网络对于干净数据集 X 有着相对较高置信度的分类预测，但是其依然对其部分干净样本存在错误分类的情况。然而，现有鲁棒蒸馏方法并未对所有干净样本数据进行区分，且通常将全部样本直接进行训练。该做法易对模型的鲁棒性产生不利的影响，从而使其目标

损失逐渐偏离鲁棒目标的优化。针对以上问题, 本文对现有鲁棒蒸馏方法进行改进, 提出基于数据过滤的鲁棒蒸馏机制。具体而言, 本文通过对原始干净样本中被错误识别的样本进行过滤, 提高教师网络中鲁棒知识迁移的有效性, 其具体步骤如下。

第一步, 将干净样本 $x_i \in \mathcal{D}$ 输入鲁棒的教师网络 $\mathcal{F}_T$, 若 $\mathcal{F}_T(x_i) \neq y^{\text{true}}$, 则表明网络对该样本进行预测后得到的软标签信息并不可靠。因此, 本文首先将满足条件 $\mathcal{F}_T(x_i) = y^{\text{true}}$ 的干净样本进行筛选选择, 得到含有 M 个样本的新数据集 $\mathcal{D}' = \{(x_i, y_i^{\text{soft}})\}_{i=1}^M$, 其中 y_i^{soft} 为被网络正确分类的样本输出的软标签。因此, 数据集 $\mathcal{D}'$ 中的所有样本 x_i , 都能够被教师网络 $\mathcal{F}_T$ 正确分类。

第二步, 利用这些鲁棒软标签 y_i^{soft} , 可以更好地引导学生网络往鲁棒方向优化, 整体损失函数如下:

$$\begin{aligned} \text{Loss} &= \alpha \cdot \text{Loss1} + (1 - \alpha) \cdot \text{Loss2} \\ \text{Loss1} &= t^2 \text{KL}(Q(x, \theta_T, t), Q(x', \theta_S, t)) \quad (10) \\ \text{Loss2} &= L_{\text{ce}}(Z(x, \theta_S), y^{\text{true}}) \end{aligned}$$

其中, $x \in \mathcal{D}'$ 是可以被教师网络正确分类的干净样本, L_{ce} 为交叉熵损失函数, α 用来控制损失项的比例, 用于平衡鲁棒准确率与干净样本准确率。算法 1 对鲁棒蒸馏算法的过程进行详细描述。

算法 1 鲁棒蒸馏算法

输入 干净样本训练集 $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^N$, 鲁棒的预训练教师网络 $\mathcal{F}_T(x; \theta_T)$, 待训练的学生网络 $\mathcal{F}_S(x; \theta_S)$, 训练 T 个 epochs, 总共 N 个 batch, 蒸馏温度 $\mathcal{D}' = \{(x_i, y_i^{\text{soft}})\}_{i=1}^M$

输出 新的训练集 $\mathcal{D}' = \{(x_i, y_i^{\text{soft}})\}_{i=1}^M$ 和鲁棒的轻量级学生网络 $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^N$

- 1) 将 $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^N$ 输入 $\mathcal{F}_T(x; \theta_T)$
- 2) if $\mathcal{F}_T(x_i; \theta_T) = y^{\text{true}}$ then
- 3) 将所有符合条件的干净样本 $\mathcal{D}' = \{(x_i, y_i^{\text{soft}})\}_{i=1}^M$, 组成新的训练集合 $\mathcal{D}' = \{(x_i, y_i^{\text{soft}})\}_{i=1}^M$
- 4) end if
- 5) 对学生网络 $\mathcal{F}_S(x; \theta_S)$ 进行相同的参数初始化
- 6) for epoch = 1, ..., T do

- 7) for batch = 1, ..., N do
- 8) 将 $\mathcal{D}' = \{(x_i, y_i^{\text{soft}})\}_{i=1}^M$ 中的样本 $\mathcal{F}_T(\cdot; \theta_T)$ 输入 $\mathcal{F}_T(x; \theta_T)$, 在蒸馏温度 $Q(x, \theta_T, t)$ 下得到软标签概率向量 $Q(x, \theta_T, t)$
- 9) 基于每一个样本 $x_i \in \text{batch}$ 构建对抗样本 $\mathcal{F}_S(\cdot; \theta_S)$, 并将其输入学生模型 $\mathcal{F}_S(x; \theta_S)$ 得到蒸馏温度 $Q(x', \theta_S, t)$ 下的类别预测的概率向量 $Q(x', \theta_S, t)$
- 10) 计算式(10)的 Loss, 并通过 SGD 更新学生模型 $\mathcal{F}_S(x; \theta_S)$ 参数
- 11) end for
- 12) end for

3 实验评估

3.1 实验设置

(1) 数据集和模型

1) 基准数据集: 本文实验采用两个基准数据集, 分别为 CIFAR-10 和 CIFAR-100。这两个数据集各包含 60 000 张 RGB 彩色图片, 其中训练集 50 000 张, 测试集 10 000 张, 图片大小为 32×32 像素。CIFAR-10 和 CIFAR-100 分别包含 10 类和 100 类图片数据。

2) 测试模型: 本文选取了 VGG-16^[19] 和 resnet56^[20] 作为待剪枝的教师网络模型。通过对测试模型进行剪枝和鲁棒蒸馏, 以提高模型在较低 FLOP 情况下的鲁棒性。

(2) 实验环境

本文实验所用硬件为 RTX 2080Ti GPU, 运行环境为 Ubuntu 16.04.6 LTS, Pytorch 版本为 1.4.0, CUDA 版本为 10.0, cuDNN 版本为 7.6.0, torchattacks 库的版本为 3.0.0。

(3) 评价指标

1) FLOP: 即浮点运算数, 用来衡量算法/模型的复杂度。FLOP 越小, 所需计算量也越小, 代表越有利于算法/模型在资源相对受限的边缘设备计算环境下进行应用部署。

2) 对抗鲁棒准确率: 代表算法/模型对于对抗样本的分类精度。本文使用目前较为常用的鲁棒性基准测试平台 AutoAttack^[21] (下面简称 AA 测试) 对模型的鲁棒性进行评估。

本文在 CIFAR-10 和 CIFAR-100 两个数据集

上测试模型的鲁棒性。具体设置默认的扰动大小 ε 为 8/255。所有的鲁棒性测试结果均为运行 3 次后的平均结果。除此之外, 对比的剪枝方法包含: L1-norm^[22]剪枝、Slimming^[23]剪枝以及 CHIP^[24]剪枝。其中, L1-norm 剪枝为基于人为预定义剪枝比例的剪枝方法, 通过在待剪枝的卷积层中计算卷积核的 L1 范数并进行排序, 对那些范数值较小的卷积核进行剔除; Slimming 剪枝是一种全局的剪枝方法, 其将通道 BN 层的 γ 系数作为重要性度量的指标, 对于那些 γ 系数较小的通道进行剪枝处理以压缩模型; CHIP 剪枝方法则提出用通道独立性的指标来度量特征图之间的相关性, 独立性越差的特征图所对应的卷积核被认为是冗余的, 可以被剪掉。

3.2 CIFAR-10 实验结果

本节基于 CIFAR-10 数据集, 选取 VGG-16 和 resnet56 作为待剪枝的教师网络, 且都是经过对抗训练的鲁棒网络。在此基础上, 实验分别从对抗训练和鲁棒蒸馏两个角度, 对比不同剪枝方法得到的模型结构对鲁棒性的影响。图 2 给出了模型在不同剪枝方法以及在不同训练方法下的鲁棒性对比。

在对抗训练实验中, 本文选择 TRADES 框架^[16]作为对抗训练的方法。从图 2(a)可以观察到, 本文所提的分层自适应剪枝方法得到的模型在相同 TRADES 对抗训练条件下, 相比于其他剪枝方法, 几乎在各个 FLOP 层次上均表现出更高的鲁棒性 (AA 测试)。这充分表明分层自适应剪枝得到的模型结构更能适应模型鲁棒性的需要, 即模型结构更为合理。为了公平对比, 图 2 和图 3 所有数据点均是模型在实验训练中取得的最好结果。

在鲁棒蒸馏实验中, 为更合理地进行对比, 所有测试方案均采用剪枝技术与鲁棒蒸馏融合的形式。测试剪枝技术包括 L1-norm、Slimming 与 CHIP。实验中除本方案外, 其余剪枝操作后使用的鲁棒蒸馏方法均为 RSLAD 鲁棒蒸馏方法^[14]。该方法是现有较为先进的开源鲁棒蒸馏方法。首先, 图 2(b)给出了不同方案在 VGG-16 教师网络上的性能对比。结果表明本文所提出的基于剪枝技术和鲁棒蒸馏融合的轻量对抗攻击防御方法与其余剪枝与鲁棒融合方法相比, 在相同的 FLOP 层次上具有更高的对抗鲁棒准确率; 在相同的对

抗鲁棒准确率上需要更少的 FLOP。因此, 本文所提方法在综合性能上表现更好, 尤其是高剪枝率、低 FLOP 情况下更为明显。此外, 实验将相同 FLOP 条件下的 VGG-11 模型和 VGG-13 模型参与横向对比。结果表明, 本文所提方法是一种有效的压缩模型手段, 且比预设模型结构进行训练的效果更好。以上的高性能主要得益于分层自适应剪枝策略与鲁棒蒸馏前的数据过滤预处理方法带来网络结构的优化。

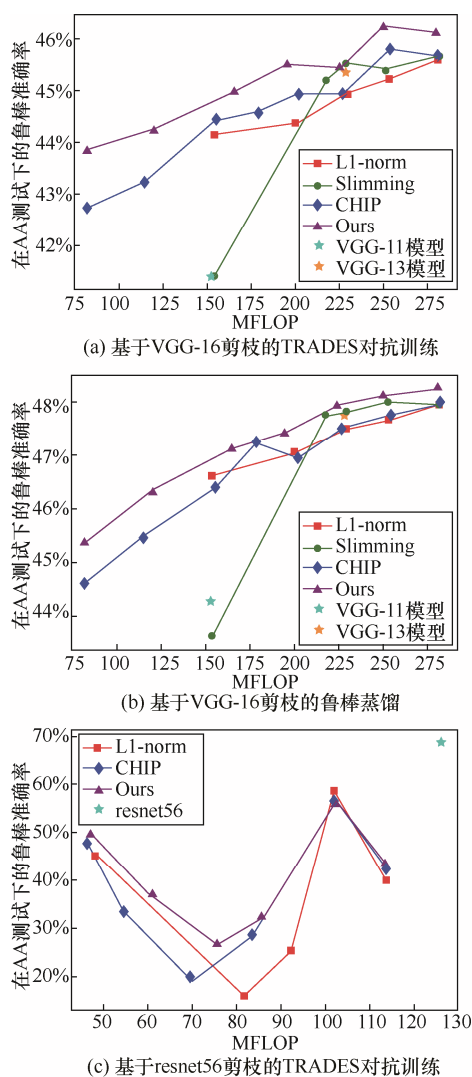


图2 模型在不同剪枝方法以及在不同训练方法下的鲁棒性对比
Figure 2 Comparison of the model robustness under different pruning methods and various training methods on CIFAR-10

图 2(c)则给出了不同方案在 resnet56 教师网络上的性能对比。实验结果表明本文所提方法剪枝得到的结构经过相同的对抗训练后, 其性能曲

线均位于其他方法上方, 即其鲁棒性相较于其他方法在整体上具有优势。此外, 所提方法的鲁棒准确率在 resnet56 与 VGG-16 两种模型上的趋势存在差异, 即图 2(c) 中存在拐点, 该现象主要是源于两种模型的内在结构差异。VGG 是一种基于直连结构的模型, 各层之间具有不同的语义信息。只要未对 VGG 模型中间的某些层进行过度剪枝, 模型表达能力和性能不会有明显的骤降, 且训练后的表现较为稳定。因此, 本文所提方法的鲁棒准确率在 VGG 模型上与模型的 FLOP 大致呈现正相关的趋势。resnet56 则是通过特征融合, 采取跳层连接结构。这种内在结构的不同, 可能会使剪枝操作优先关注某些语义表达丰富的特定层, 并且这些层在剪掉相当一部分卷积核之后, 出现过渡态的非最优结构, 进而出现层与层之间特征表达及语义信息不匹配的情况, 造成模型的表达能力受限, 即 FLOP 处于 70~80 MFLOP 的低谷点。随着迭代剪枝的进行, 该方法会动态地优化剩下的模型结构, 使其前后跨通道的特征融合更优, 即前后层的结构更匹配、更互相适应。

3.3 CIFAR-100 实验结果

图 3 展示了基于 CIFAR-100 数据集以及 VGG-16 模型的实验结果。结果表明, 如同 CIFAR-10 的表现, 本文所提方法不论在相同的 TRADES 对抗训练下, 还是对比其他的鲁棒蒸馏方法, 综合鲁棒性要表现更好。

另外, 尽管 Slimming 方法是基于全局阈值方法的剪枝, 但是由该方法剪枝得到的模型在高剪枝条件下会导致模型鲁棒性降。因此该方法存在较大缺陷, 无法有效部署于边缘环境中。

4 结束语

针对物联网边缘环境对轻量对抗攻击防御方法的需求, 本文提出一种基于剪枝技术和鲁棒蒸馏融合的轻量对抗攻击防御方法, 通过分层自适应剪枝技术与基于数据过滤的鲁棒蒸馏方法, 实现鲁棒对抗训练模型的有效压缩, 即在降低其模型容量的同时, 对模型的鲁棒性影响较小。实验结果证明, 本文所提方法不仅能使相同 FLOP 条件下的轻量级网络经过同等对抗训练更具鲁棒

性, 并且在相同鲁棒对抗准确率下, 使压缩的模型具有更低的 FLOP。未来将对剪枝技术和鲁棒蒸馏做进一步的研究拓展, 继续提高模型的压缩率同时保持较高的鲁棒性。

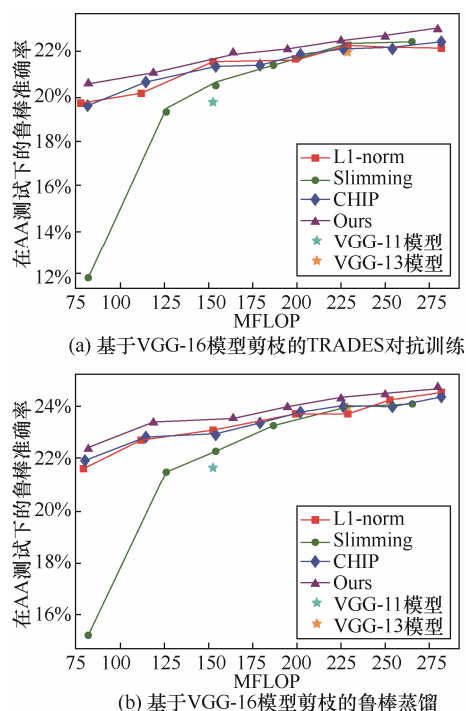


图 3 模型在不同剪枝方法以及在不同训练方法下的鲁棒性对比
Figure 3 Comparison of the model robustness under different pruning methods and various training methods on CIFAR-100

参考文献:

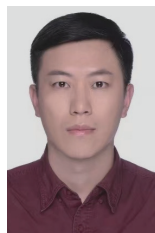
- [1] 郑远攀, 李广阳, 李晔. 深度学习在图像识别中的应用研究综述[J]. 计算机工程与应用, 2019, 55(12): 20-36.
ZHENG Y P, LI G Y, LI Y. Survey of application of deep learning in image recognition [J]. Computer Engineering and Applications, 2019, 55(12): 20-36.
- [2] 鱼昆, 张绍阳, 侯佳正, 等. 语音识别及端到端技术现状及展望[J]. 计算机系统应用, 2021, 30(3): 14-23.
YU K, ZHANG S Y, HOU J Z, et al. Survey of speech recognition and end-to-end techniques[J]. Computer Systems & Applications, 2021, 30(3): 14-23.
- [3] 王睿怡, 罗森林, 吴舟婷, 等. 深度学习在汉语语义分析的应用与发展趋势[J]. 计算机技术与发展, 2019, 29(9): 110-116.
WANG R Y, LUO S L, WU Z T, et al. Application and development trend of deep learning in Chinese semantic analysis[J]. Computer Technology and Development, 2019, 29(9): 110-116.
- [4] SZEGEDY C, ZAREMBA W, SUTSKEVER I, et al. Intriguing properties of neural networks[C]//Proceedings of 2nd International Conference on Learning Representations (ICLR 2014). 2014.
- [5] CARLINI N, WAGNER D. Towards evaluating the robustness of neural networks[C]//Proceedings of 2017 IEEE Symposium on Se-

- curity and Privacy (SP). 2017.
- [6] SONG D, EYKHOLT K, EVTIMOV I, et al. Physical adversarial examples for object detectors[C]//12th USENIX Workshop on offensive technologies (WOOT 18). 2018.
- [7] CARLINI N, WAGNER D. Audio adversarial examples: targeted attacks on speech-to-text[C]//2018 IEEE Security and Privacy Workshops (SPW). 2018: 1-7.
- [8] GOODFELLOW I, SHLENS J, SZEGEDY C. Explaining and harnessing adversarial examples[C]//Proceedings of 3rd International Conference on Learning Representations (ICLR 2015). 2015.
- [9] MADRY A, MAKELOV A, SCHMIDT L, et al. Towards deep learning models resistant to adversarial attacks[C]//Proceedings of 6th International Conference on Learning Representations (ICLR 2018). 2018.
- [10] GOWAL S, QIN C, UESATO J, et al. Uncovering the limits of adversarial training against norm-bounded adversarial examples[J]. arXiv:2010.03593, 2020.
- [11] GOU J, YU B, MAYBANK S J, et al. Knowledge distillation: a survey[J]. International Journal of Computer Vision, 2021, 129(6): 1789-1819.
- [12] PAPERNOT N, MCDANIEL P, WU X, et al. Distillation as a defense to adversarial perturbations against deep neural networks[C]//IEEE Symposium on Security and Privacy, 2016: 582-597.
- [13] GOLDBLUM M, FOWL L, FEIZI S, et al. Adversarially robust distillation[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2020: 3996-4003.
- [14] ZI B, ZHAO S, MA X, et al. Revisiting adversarial robustness distillation: robust soft labels make student better[C]//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision, 2021:16423-16432.
- [15] BAI T, LUO J, ZHAO J, et al. Recent advances in adversarial training for adversarial robustness[C]//Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, 2021: 4312-4321.
- [16] ZHANG H, YU Y, JIAO J, et al. Theoretically principled trade-off between robustness and accuracy[C]//International Conference on Machine Learning. 2019: 7472-7482.
- [17] ALAYRAC J B, UESATO J, HUANG P S, et al. Are labels required for improving adversarial robustness[C]//Advances in Neural Information Processing Systems. 2019: 12192-12202.
- [18] CARMON Y, RAGHUNATHAN A, SCHMIDT L, et al. Unlabeled data improves adversarial robustness[C]//Advances in Neural Information Processing Systems. 2019: 11190-11201.
- [19] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[J]. arXiv:1409.1556, 2014.
- [20] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition[C]//Proceedings of 2016 IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2016: 770-778.
- [21] 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. 2020: 2206-2216.
- [22] LI H, KADAV A, DURDANOVIC I, et al. Pruning filters for efficient convnets[C]//Proceedings of 5th International Conference on Learning Representations (ICLR 2017). 2017.

[23] LIU Z, LI J, SHEN Z, et al. Learning efficient convolutional networks through network slimming[C]//Proceedings of 2017 IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2017: 2755-2763.

[24] SUI Y, YIN M, XIE Y, et al. CHIP: channel Independence-based pruning for compact neural networks[C]//Advances in Neural Information Processing Systems. 2021: 24604-24616.

[作者简介]



王滨 (1978-), 男, 山东泗水人, 博士, 浙江省多维感知技术应用与安全重点实验室研究员、博士生导师, 主要研究方向为物联网安全、人工智能安全、密码学。



李思敏 (1997-), 男, 浙江义乌人, 浙江科技学院硕士生, 主要研究方向为深度学习、人工智能安全和模型压缩。



钱亚冠 (1976-), 男, 浙江嵊州人, 博士, 浙江科技学院理学院教授, 主要研究方向为深度学习、人工智能安全。

张君 (1980-), 女, 河南内乡人, 浙江省信息化发展中心高级工程师, 主要研究方向为网络与信息安全。



李超豪 (1995-), 男, 浙江温州人, 博士, 浙江省多维感知技术应用与安全重点实验室副主任, 主要研究方向为物联网安全、人工智能安全、感知对抗安全、数据隐私保护。

朱晨鸣 (1981-), 男, 浙江杭州人, 浙江省电子信息产品检验研究院高级工程师, 主要研究方向为网络与信息安全。

张鸿飞 (1984-), 男, 浙江嘉兴人, 浙江省电子信息产品检验研究院工程师, 主要研究方向为网络与信息安全。