软件学报 ISSN 1000-9825, CODEN RUXUEW
2026,37(7):2989−3012 [doi: 10.13328/j.cnki.jos.007548] [CSTR: 32375.14.jos.007548]

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

PBAT: 基于代理分布的深度学习模型防御加固方法*

杨 波[1,2], 熊 倩[1], 徐 珞[3]

[1](北京林业大学 信息学院, 北京 100083)

[2](国家林业草原林业智能信息处理工程技术研究中心 (北京林业大学), 北京 100083)

[3](中国电子科技集团公司 信息科学研究院, 北京 100040)

通信作者: 杨波, E-mail: yangbo@bjfu.edu.cn

摘 要: 深度学习模型存在一定的安全隐患, 和这些模型关联的应用一旦发生安全事故, 很可能带来难以承受的后果. 为此, 需要对深度学习模型进行加固. 已有的加固方法包括针对对抗性噪声的对抗性训练方法、对抗噪声净化方法等. 对抗训练作为其中最常使用的方法具有较好的对抗攻击防御效果. 但是, 对抗训练后的模型容易出现鲁棒泛化不足的问题, 且会牺牲较多的原始精度. 基于此提出一种基于代理分布的对抗训练 (proxy-distribution-based adversarial training, PBAT) 方法. 该方法利用概率模型捕获数据样本分布规律, 以生成能够协调原始精度和鲁棒性能的增强训练样本. 然后通过调整训练过程实现对模型的加固. 利用 ResNet-20 和 GoogLeNet 在 2 个典型的数据集 CIFAR10 和 MNIST 上开展实验, 并进一步在 Faster R-CNN 模型和 PASCAL VOC 数据集上针对目标检测任务开展实验. 实验结果表明, PBAT 比其他 4 种典型的方法效果要好.

关键词: 深度学习模型; 模型加固; 模型防御; 对抗网络

中图法分类号: TP183

中文引用格式: 杨波, 熊倩, 徐珞. PBAT: 基于代理分布的深度学习模型防御加固方法. 软件学报, 2026, 37(7): 2989–3012. http://www.jos.org.cn/1000-9825/7548.htm

英文引用格式: Yang B, Xiong Q, Xu L. PBAT: Defense Reinforcement Method for Deep Learning Models Based on Proxy Distribution. Ruan Jian Xue Bao/Journal of Software, 2026, 37(7): 2989–3012 (in Chinese). http://www.jos.org.cn/1000-9825/7548.htm

PBAT: Defense Reinforcement Method for Deep Learning Models Based on Proxy Distribution

YANG Bo[1,2], XIONG Qian[1], XU Luo[3]

[1](School of Information Science and Technology, Beijing Forestry University, Beijing 100083, China)

[2](Engineering Research Center for Forestry-oriented Intelligent Information Processing of National Forestry and Grassland Administration (Beijing Forestry University), Beijing 100083, China)

[3](Information Science Academy, China Electronics Technology Group Co., Beijing 100040, China)

Abstract: Deep learning models face some security risks, and security breaches in their applications can lead to severe consequences. Enhancing the security of deep learning models is therefore necessary. Existing reinforcement methods include adversarial training against adversarial noise, adversarial noise purification methods, and others. Among them, adversarial training is the most widely used method and provides effective defense against adversarial attacks. However, models trained with adversarial training often suffer from insufficient robust generalization and considerable loss of original accuracy. To address these issues, this study proposes a new adversarial training method called proxy-distribution-based adversarial training (PBAT). The proposed method employs a probabilistic model to capture the distribution patterns of data samples and generate enhanced training samples that balance original accuracy and robustness. The resilience of the model is further enhanced through an adjusted training process. Experiments are conducted using ResNet-20 and GoogLeNet on two benchmark datasets, CIFAR10 and MNIST. Furthermore, experiments are conducted on the Faster R-CNN model and the PASCAL VOC

* 基金项目: 国家重点研发计划 (2023YFD2201805)
收稿时间: 2024-09-06; 修改时间: 2025-04-21, 2025-05-30; 采用时间: 2025-09-03; jos 在线出版时间: 2026-04-29
CNKI 网络首发时间: 2026-04-30

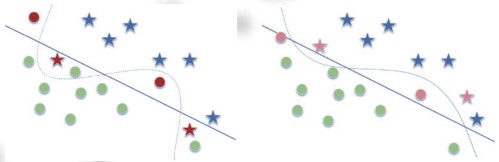
dataset for object detection tasks. The experimental results demonstrate that PBAT outperforms four representative methods.

Key words: deep learning model; model reinforcement; model defense; adversarial network

近年来, 深度神经网络 (deep neural network, DNN)[1]在各种机器学习任务中表现出了显著的性能, 通过在例如图像分类[2]、人脸识别[3]、目标识别[4]和自然语言处理[4]等应用领域使用深度学习模型, 获得了接近甚至超过人类的表现, 在一定程度上改变了人们的日常生活.

然而, 深度学习模型存在安全隐患. 与之关联的应用一旦发生安全事故, 可能造成重大经济损失. 例如: 自动无人驾驶汽车[5]、恶意软件检测[6]、医疗健康[7]等领域, 恶意设计的噪声可能避开人眼察觉, 却严重影响深度学习模型性能. 如 Szegedy 等人[8]在图像分类领域发现, 在原始样本上添加轻微扰动, 就能使基于深度学习模型的图像分类系统输出错误结果. 这种微小扰动还可能被不法分子利用, 带来严重安全后果[9,10].

针对对抗性噪声提升深度学习模型鲁棒性的方法中, 对抗训练表现较好且备受关注. 对抗训练通过在模型训练过程中添加对抗样本, 使模型学习对抗样本所在分布, 增强对输入数据变化的抵抗能力. 图 1(a) 用平面样例直观地展示了标准训练下非鲁棒模型的分类边界 (实线) 和对抗训练后鲁棒模型的分类边界 (虚线) 的表现差异. 其中五角星和圆形表示待区分的两类样本, 红色样本点为对抗样本. 可以发现, 非鲁棒决策边界下的模型错误地分类了对抗样本, 而对抗训练后, 获得了鲁棒决策边界 (虚线) 的模型虽然能正确分类对抗样本, 但在对抗样本方向上产生的对原始决策边界较大程度的推动, 也影响了模型原先学习到的数据中的重要判别性特征.



(a) 标准对抗训练 (b) PBAT

图 1 标准训练和对抗训练结果对比示例

在此基础上, 衍生出许多基于对抗训练[11–13]的防御方法. 但对抗训练也存在问题, 例如, 导致模型在标准训练模型输入空间的对抗性方向决策边际过度增加[14], 许多研究已经证明, 原始样本所在分布和对应的对抗样本所在分布截然不同[15–17], 最终加固后的模型会在原始数据集上的精度下降, 损失大量原始泛化性能. 研究进一步关注到模型鲁棒性和精度的权衡. 随着生成模型的发展, 一些研究选择尝试利用生成对抗网络 (generative adversarial network, GAN) 或扩散模型直接消除对抗性噪声, 达到不损害鲁棒模型的原始精度的目的[18–21]. 但净化模型的训练过程往往容易出现模式崩溃, 可能无法有效去除对抗性噪声, 甚至引入新噪声, 模型最终的鲁棒性表现通常落后于对抗训练方法[22]. 研究中使用的对比方法 ATLD (adversarial training with latent distribution)[23]从原始样本中诱发出潜在表示, 利用流形信息产生对抗性扰动进行训练, 并借助具有流形变换的推理方法 (IMT) 增强泛化性. CCAT (confidence-calibrated adversarial training) 方法[24]通过初始扰动和多次迭代的 PGD 攻击进行训练, 侧重于推测拒绝对抗样本的行为. TRADES (trade-off between robustness and accuracy) 方法[12]则聚焦于调节精度和鲁棒性之间的平衡, 通过调整优化目标来训练鲁棒模型. 这些方法虽然在一定程度上增强了鲁棒性和泛化能力, 但没能从样本多样性角度重复利用对抗样本, 在实际应用场景仍然有改进空间.

本文利用对抗训练提高模型对抗鲁棒性方面的优势, 借助概率生成模型分析对抗样本与原始样本之间的分布差异, 从而生成多样性对抗训练数据, 改善模型的鲁棒性. 为了说明这一新方法对模型决策边界产生的变化, 图 1(b) 展示了这种训练后模型更理想的决策边界 (渐变色曲线). 这条新的决策边界避免了先前在对抗样本方向上边界过度敏感的变化, 同时保留了模型的鲁棒性. 实现从图 1(a) 到图 1(b) 的调整效果, 关键是通过引入新数据 (粉色), 来帮助原始样本和对抗样本分布之间完成差异过渡. 考虑到概率生成模型在机器学习和人工智能领域的关键作用, 本文聚焦于利用其学习样本分布规律的能力, 构建一种基于样本分布规律的多样性数据构造方法, 以服务于模型的鲁棒训练.

具体而言, 本文提出了基于代理分布的对抗训练 (proxy-distribution-based adversarial training, PBAT) 方法. 该方法以对抗训练为基础框架, 借助概率模型学习原始样本和对抗样本的数据分布规律, 进而从对抗样本中衍生出新样本. 与传统的随机翻转、添加随机噪声、插值[25–27]等数据增强方法不同, PBAT 使用辅助模型深入学习对抗样本和原始样本的分布规律, 生成能平衡两者差异的代理样本, 从根本上改变了数据增强的方式, 从样本分布角度改善对抗性方向决策边际过度增加问题, 优化模型的对抗训练过程. PBAT 的核心在于通过有针对性的数据构造方法增强对抗鲁棒训练数据. 新生成的数据作为“协调者”, 在对抗训练中平衡模型的鲁棒性与精度, 降低对抗训练对模型泛化性能的负面影响.

本文的主要贡献在于以下 4 个方面.

(1) 提出一种数据增强方法, 利用概率模型学习对抗样本和原始样本之间分布的差异, 并指导生成逼近目标分布下的增强数据集, 是一种更有针对性, 适用于智能模型获取对抗鲁棒性的再训练过程中使用的多样性增强数据集.

(2) 构建新的基于代理分布的对抗训练框架, 该框架在图像分类和目标检测等多种任务以及不同模型结构上都展现出良好的效果, 提升了模型在多种攻击下的对抗鲁棒性, 拓展了深度学习模型在复杂安全环境下的应用范围.

(3) 在分类任务上, 使用两个典型的深度学习模型 (ResNet-20 和 GoogLeNet) 及两个典型图像分类数据集 (CIFAR10 和 MNIST) 进行实验, 实验结果表明, PBAT 方法可以在不同的模型之间进行转移, 并相较于其他加固方法更好地平衡了鲁棒性与精度, 提高鲁棒泛化能力, 并在多种攻击下普遍有效.

(4) 进一步将本方法应用范围扩展到更复杂的目标检测领域的任务上. 通过在经典的 PASCAL VOC 数据集上使用 PBAT 方法, 对 Faster R-CNN 模型进行加固, 实验检测精度结果进一步表明本文加固方法的有效性, 能够在不同计算机视觉领域任务中实现迁移, 灵活适用于各类场景.

本文第 1 节进行相关工作的介绍. 第 2 节详细说明 PBAT 方法, 包括训练中对数据的生成, 训练过程等内容. 第 3 节描述实验设置, 其中包括研究问题、实验数据集及模型、实验配置、评价指标和比较方法. 第 4 节对实验的结果进行展示与详细分析. 第 5 节对本文工作进行总结与展望.

1 深度学习模型加固相关工作

1.1 图像数据增强

图像数据增强是一种经常应用在计算机视觉领域的方法, 以往在模型训练时, 为了提高模型的泛化能力, 经常使用数据增强的方式. 数据增强已被证明可以减少标准训练的泛化误差. 数据增强的核心思想是通过生成合成数据集来提高训练数据的充分性和多样性. 基本的图像数据增强方法包括图像处理、图像擦除、图像混合.

图像处理方法大多使用随机变换包括旋转、翻转、缩放、裁剪、颜色变换等操作, 增加数据多样性, 可以使模型在训练过程中接触到更多的数据变化, 有助于模型学习到更丰富的数据特征, 从而提高模型的泛化能力, 易于实现. 但这类方法仅在假设现有训练数据服从于实际数据分布的情况下才是有意义的. 此时增强的数据可以看作是从一个接近真实分布的分布中提取出来的新数据, 大都存在填充效应, 会导致部分信息的浪费.

CutOut[25]是一种简单而有效的数据增强方法, 利用图像擦除的思想, 在训练过程中随机遮盖图像的一部分, 使模型无法直接看到被遮盖的区域, 从而鼓励模型学习更具有判别性的特征和全局上下文信息. 通过在训练过程中引入随机遮盖, CutOut 方法可以有效提高模型的泛化能力, 使其在测试阶段取得更好的性能. Random Erasing 与 CutOut 类似, 但区别在于它会随机选择一个矩形区域, 并用随机值 (而非平均值) 填充该区域. 这种变化使得模型需要在训练过程中学习如何从随机扰动中恢复真实信息, 进一步提高了模型的鲁棒性和泛化能力. de Jorge 等人[28]提出 N-FGSM, 将噪声用作数据增强的一种形式. 在足够强的噪声下进行数据增强和去除剪切步骤可以防止过拟合.

Mixup[26]的主要思想是图像混合, 在训练过程中将不同的样本及其标签以一定比例混合, 从而生成新的训练样本. 它为每个训练样本分配一个权重系数, 这些系数遵循均匀分布或其他预定义的概率分布. 根据权重系数, 将原始训练样本及其标签以一定比例混合. 具体来说, 对于两个样本 (x_1,y_1) 和 (x_2,y_2), 生成新的混合样本 (x',y'), 其中 $x'=\lambda x_1+(1-\lambda)x_2$, $y'=\lambda y_1+(1-\lambda)y_2$, λ 是权重系数. Mixup 方法的一个重要变体是 CutMix[29], 结合了 CutOut

和 Mixup 的特点, 它将一个样本的部分区域替换为另一个样本的对应区域, 同时保持混合后的标签不变. 调整了 CutOut 和 Mixup 两种方法的不足, 通过有效地利用训练像素并保留区域 dropout 的正则化效果, 在 ImageNet 弱监督定位任务上取得了不错的增强效果.

除了基本的人为设置数据增强手段, 更智能的数据增强方法是在搜索空间自动搜索增强方法. Cubuk 等人[30]提出的随机增强通过删除一个单独的搜索, 大大减少了数据增强的搜索空间还进一步提高了自动增强和 PBA[31]的性能. 扩展到模型的特征空间, Li 等人[32]提出了一种隐式数据增强方法特征增强, 通过利用模型中间的潜在张量增强模型训练过程. 最近的图像数据增强策略通常利用 GAN[33]模型生成新的样本. 其中典型的 StarGAN 系列[34]实现在一个数据集中进行多领域的图像转换任务, 可以合并包含不同标签的数据集, 对其中任意的标签属性灵活进行图像转换, 将输入图像转换到任何所需的目标域上. 然而, 这些方法主要关注提升标准训练下的泛化性能, 在改善对抗性方向决策边际过度增加问题上作用有限. 生成的数据与对抗样本分布缺乏紧密联系, 难以保证模型最终的对抗鲁棒性.

1.2 概率生成模型

概率生成模型是一种利用概率统计理论来生成数据的模型. 这类模型通过对数据的概率分布进行建模, 可以生成与真实数据具有相似特性的新数据样本. 例如生成对抗网络 (GAN)、变分自编码器 (variational autoencoder, VAE)、隐马尔可夫模型 (hidden Markov model, HMM) 等. 概率生成模型通常基于某种概率分布 (如正态分布、泊松分布等) 来描述数据的特性, 并通过学习得到模型的参数. 学习与训练过程通常涉及最大似然估计、贝叶斯推断等方法, 以求解模型的最优参数.

GAN 是一种利用生成器与鉴别器相互博弈实现训练的网络架构. 由于训练既不需要知道真实数据分布, 也不需要任何数学假设, 因而得以广泛应用. 最开始的 GAN 中, 生成模型和判别模型使用 MLP 进行定义. 后来, 由于卷积神经网络 (CNN) 在图像表示上优于 MLP 模型, 产生了深度卷积生成对抗网络 (deep convolutional generative adversarial network, DCGAN)[33]. 之后发展的 GAN 也大多基于 DCGAN 的架构. 使用自注意力机制的 SAGAN[35]允许对图像生成任务的注意进行驱动, 并对生成模型和判别模型都使用了光谱归一化技术, 来提高训练的效果. BigGAN[36]则在 SAGA 的基础上, 使用扩大了的网络结构并结合“截断技巧”, 提升了生成图像的质量和训练的稳定性. StyleGAN[37]使用一种基于风格的条件生成器和使用联合训练的鉴别器, 从而实现了对生成图像中特定尺度属性的控制.

最先进的生成模型能够建模当前的大规模图像数据集的分布, 可以隐式地学习给定数据集的概率分布, 即 $P(X)$. 因此, 它能够产生高保真的人工样本. GAN 架构由两个神经网络组成, 即生成器 G 和鉴别器 D. 其中生成器 G 从训练样本中学习数据的分布以生成尽可能相似的新的样本, 而鉴别器 D 则被训练来尽可能地区分出真实数据样本和合成新样本. 这些组件在一个对抗性的过程中同时进行训练. 在训练过程中, 首先从一个已知的概率分布 $N(Z)$ 中采样一个噪声向量 z, 通常是高斯分布, 并输入 G. 然后, G 将 z 从 $N(Z)$ 所在分布转换为 x 所在分布. 另一方面, D 通过比较 G 的输出与来自数据集的真实样本 x 来检查 G 的分布获取性能如何. 在训练过程中, 通过优化公式 (1) 实现模型的训练, 其中最为常用的损失函数如公式 (2) 所示:

$$\min_G \max_D L^{\mathrm{JS}}_{\mathrm{GAN}}(D, p'; p) \tag{1}$$

$$L^{\mathrm{JS}}_{\mathrm{GAN}}(D, p'; p) = E_{x\sim p(X)}[\log(D(x)] + E_{x\sim p'(X)}[\log(1 - D(x))] \tag{2}$$

其中, p' 是由固定的先验分布 $N(Z)$ 中采样后再由 G 所生成的分布. 为了使 GAN 训练的过程更加稳定, 生成更高质量的合成样本, 研究人员不断调整 L_{GAN} 的选择, 如 WGAN-GP[38]判别器损失函数如公式 (3) 所示, 利用 Wasserstein 距离代替了原先的 JS 距离, 并加入了梯度惩罚项: 公式前两项为基础的 WGAN 判别器损失函数. 最后一项是添加的梯度约束限制. $x\sim\hat{p}$ 是每一批次的生成样本和真实样本中加权和的采样, 即 $x\in\hat{p}=\varepsilon x\in p+(1-\varepsilon)x'\in p'$. $\|\nabla_x D(X)\|_2-1$ 是对权重的裁剪项.

$$L_{\mathrm{GAN}} = -E_{x\sim p}[D(x)] + E_{x\sim p'}[D(x)] + \lambda E_{x\sim\hat{p}}[(\|\nabla_x D(x)\|_2 - 1)^2] \tag{3}$$

概率生成模型能利用概率统计理论对数据分布建模, 生成与真实数据相似的新样本. 如 GAN 通过生成器和

鉴别器的对抗训练学习数据分布, 在防御加固中, GAN 可学习对抗样本和原始样本分布规律, 生成多样性数据, 辅助对抗训练优化模型性能, 减少模型对对抗样本的敏感性, 提升鲁棒性.

1.3 智能模型的鲁棒性

现实世界中攻击的普遍存在以及其对模型性能的损害要求我们训练出更牢固、高鲁棒性的智能模型以改善智能模型的脆弱性. 输入中的扰动噪声不会使本能做出正确决策的模型因为这类微小的扰动而决策错误. 即使模型判断错误, 也希望是在可接受范围内的. 以分类模型为例, 鲁棒模型应该满足公式 (4): 对于所有 $\|\delta\|_p \leqslant R$, 模型 f 均能将样本 x 判断为其所在正确类别, 其中 R 是对扰动的限制范围, 扰动将噪声添加到样本 x, 通常利用 L_p 范数对其大小进行约束; c_A 是样本所属类别.

$$f(x+\delta)=c_A \tag{4}$$

(1) 基于对抗训练的防御

针对对抗攻击这类特殊的扰动, 已经得到广泛证明效果最好的防御方法是利用攻击样本的对抗训练 (adversarial training), 这类方法是通过优化一个训练用的最小最大化公式 (见公式 (5)), 重新训练模型参数, 实现针对模型的加固, 提高模型抵御攻击的能力. 公式 (5) 中 x 表示从数据集 D 中采样得到的原始样本, ε 表示在扰动空间 Ω 上的扰动强度, 一般取很小的值. y 表示样本标签类且 $x \in R^n$, $y \in R$. $L(\cdot)$ 表示训练模型参数 θ 是所使用的损失函数 $\theta \in R^m$. 该公式的内部最大化部分产生用于加固的对抗样本, 而外部的最小化部分是用来训练模型最小化和对抗样本之间的损失, 拟合出鲁棒决策边界.

$$\min_{\theta} E_{(x,y)\sim D}[\max_{\varepsilon\in\Omega} L(x+\varepsilon,y;\theta)] \tag{5}$$

许多研究都在优化对抗训练上展开, Xie 等人[39]认为训练使用的 ReLU 激活函数由于其非平滑的性质, 显著削弱了对抗性训练的效果. 从而提出了平滑对抗训练 SAT 方法, 即用平滑近似函数来代替 ReLU 函数来加强对抗训练. Maini 等人[40]使用最陡梯度下降的对抗性训练方法, 将不同的扰动模型合并到一个统一的模型中, 用来解决寻找内部优化问题. 通过在每个预测的最陡梯度下降中最大化所有扰动模型的最坏损失来创建一个单一的对抗扰动, 以更好地实现鲁棒优化目标. Chen 等人[41]利用预训练扩散模型构建生成式分类器, 它净化对抗噪声的方式是将扩散模型直接转化为生成式分类器进行分类决策, 最大的特点是无需针对特定对抗攻击开展训练. 侧重于构建生成式分类器实现通用的对抗防御. 而 Sui 等人[42]则是通过生成对抗图像和对抗语义来增加对抗样本的多样性, 其侧重点在于挖掘图像语义层面的信息. 本文未将这两种方法纳入对比, 首先, 从研究重点来看, PBAT 专注于利用样本分布差异生成多样性样本改进对抗训练, 平滑决策边界. 研究重点的差异使得在相同实验条件下对比时, 难以精准突出 PBAT 在其核心改进方向上的优势. 结果难以直接反映 PBAT 在传统对抗训练框架下对样本利用和决策边界优化的独特价值. 此外, 将这两种方法纳入对比, 需要对现有实验流程, 计算资源和数据处理方式进行大幅调整, 会极大增加实验的复杂性和不确定性. 相比之下, 本研究中已选取的对比方法, 在技术复杂度以及实验环境适配性上与 PBAT 更具可比性. 最后, 考虑对比结果的可解释性, 在得出明确且有价值的研究结论的基础上选择了对比方法.

(2) 基于去噪的防御

一些方法使用去噪技术来消除或减少输入数据中的对抗噪声, 从而提高模型的鲁棒性和安全性. Jin 等人[43]提出使用 GAN 作为纯化模型提出了 APE-GAN 方法. 该方法在一个对抗性的环境下进行训练, 设计融合损失函数使对抗样本与原始样本流形保持高度一致. Zhou 等人[19]提出了一种自监督的对抗性训练机制来消除类激活特征空间中的对抗性噪声——CAFD 去噪器, 该去噪器以 RaGAN[44]为原型, 利用一个类激活特征损失和一个对抗性损失组成的混合损失函数, 再进行训练而得到. 然后, 通过最小化类激活特征空间中对抗性样本和原始样本之间的距离, 来实现消除对抗性噪声的效果. 最近, Ho 等人[20]提出了一种通过局部流形投影消除对抗性扰动的具有局部隐式函数的方法 DISCO. 该方法通过编码器和局部隐式函数实现对查询位置的原始图像 RGB 值的预测. 训练好的模型以局部的像素信息作为输入实现对抗性去除. 利用概率生成模型作为净化模型往往难以完全去除对抗噪声,

同时净化模型的性能不稳定, 对不同类型的对抗噪声去除效果参差不齐, 在复杂攻击场景下难以保障模型的鲁棒性, 无法从根本上平衡模型的鲁棒性和精度. 仅单独使用净化模型无法实际增强模型鲁棒性, 模型在面对多种攻击时防御能力不足.

(3) 鲁棒性与精确度的权衡

Zhang 等人[12]以实现鲁棒性和准确性之间的权衡作为指导原则提出了 TRADES 方法, 并给出了鲁棒误差与分类误差之间的上界. Dobriban 等人[45]研究了在 2、3 类高斯分类模型上鲁棒性与精度的权衡, 并在不平衡数据设置中推导出 l_2 和 l_∞ 范式的最优鲁棒分类器. Mehrabi 等人[46]研究了在对抗训练中标准风险和对抗风险之间的基本权衡, 认为在 l_2 范式的约束下, 标准风险和对抗风险之间, 存在一个保证最优标准风险以及最优对抗风险的模型. Liu 等人[47]提出 SimpleAT 作为一种通用的对抗性训练方法, 它使用了针对 PGD[48]和 FGSM[49]的统一训练协议, 在鲁棒性和精度之间保持良好的平衡, 并减少鲁棒性过拟合的风险, 还强调了数据增强在增强模型鲁棒性和泛化方面的重要性. Cui 等人[50]提出了可学习边界引导对抗训练 (LBGAT) 方法, 在尽可能保持自然数据精度的情况下提高模型的鲁棒性. 他们从模型边界引导的角度来理解, 认为对抗样本作为输入时, 模型输出的 logits 值应该与原始样本作为输入时相似. 这种 logits 值的指导使得鲁棒模型继承了更优秀的决策边界. Gong 等人[51]针对多种扰动提出一种节省参数的对抗训练 (PSAT) 方法, 并在最终推理阶段利用最低熵策略来选择最优的超网络, 加强了多扰动下的鲁棒性. Li 等人[52]认为提高对抗性训练的鲁棒泛化性需要数据增强尽可能多样化. 这些方法试图在保证一定鲁棒性的同时提高模型精度, 但在应对对抗性方向决策边界过度增加问题时, 仍存在局限性. 它们大多只是在损失函数或训练策略上进行调整, 未能从数据分布的本质出发, 由于对数据的理解与把握不足, 导致在复杂攻击和多样化数据场景下, 加固效果有待提高. 以上的研究是本文工作的基础.

2 本文方法 PBAT

PBAT 方法的整体框架如图 2 所示. PBAT 主要流程是: 利用攻击方法 $A(\cdot)$, 在待加固模型 f 上对原始样本 x 添加对抗噪声, 用于产生对抗样本 x'. 这些样本将用来训练概率生成模型 G, 得到能够生成目标分布下数据的代理分布模型. 在使用调整权重的损失函数进行优化训练时, 对抗样本数据将被模型 G 再利用, 生成代理分布样本 $\tilde{x}$. 并加入训练过程中, 增强训练数据集, 用于获得最终的鲁棒模型 f'.

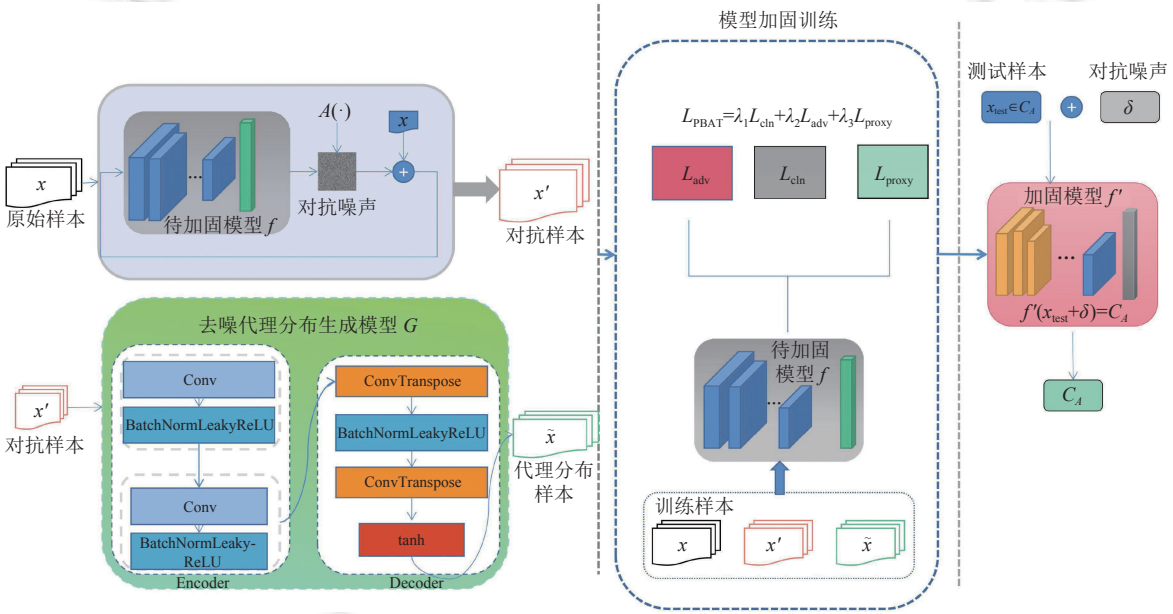


图 2 PBAT 方法架构

本文方法主要包括 3 个步骤: (1) 基于概率模型的智能模型数据样本分析; (2) 对抗鲁棒性增强的多样性数据构造; (3) 智能模型鲁棒决策边界训练. 下面将详细介绍 3 个部分的细节.

2.1 基于概率模型的智能模型数据样本分析

为构造出引言中提到的新数据, 需要先掌握数据样本的分布规律. 由于很难直接得到真实的分布情况, 就需要建立一个概率模型来描述智能模型的数据样本分布. 通过对样本分布的精准分析与潜在规律的把握, 为后续新数据的构造提供依据. 这个模型必须能准确“捕获”数据样本所具有的内在统计规律, 即模型的概率 $\hat{P}(X)$ 是真实概率分布 $P(X)$ 的近似. 估计数据集的概率密度或概率分布. 这个模型还应该能够捕捉数据的分布特性, 包括均值、方差以及可能存在的聚类现象. 在机器学习和人工智能领域, 概率模型也是核心工具之一, 因其能够量化不确定性, 广泛应用于各种任务上.

概率模型的基本思想是在给定一些可观测变量的情况下, 去估计或者预测另一些不可观测变量. 在分类问题中, 最常用的概率模型之一是朴素贝叶斯分类器, 它假设各个特征之间相互独立, 通过计算后验概率来确定样本的类别. 另外, 高斯混合模型 (Gaussian mixture model, GMM) 也被广泛用于概率聚类分析, 其中每个观测点被假定为从一组高斯分布中的一个分布生成的. 然而, 许多概率模型都假设数据符合某种特定的分布, 这在实际应用中往往无法满足. 而生成对抗网络 (GAN) 通过一个独特的对抗博弈过程来学习所需样本的分布. 在理想情况下, 经过足够的训练时间, 生成器能够捕捉到训练数据的整体分布, 以至于判别器无法区分生成数据和真实数据. 此时, 生成器便掌握了样本的分布规律. 然后就可以从这个新分布中采样, 获得所需的新数据. GAN 生成的数据是直接从数据本身学习到一种经验分布. 这使得 GAN 在捕捉复杂数据分布方面特别有优势. 图 3 描述了这种分布的变化过程, 最终的代理分布模型应该尽可能地捕获目标分布的规律. 学习到这种分布后的代理分布模型 G 能把在分布 $Q(X)$ 下采样的样本 x 转换到分布 $G(X)$ 下. 而 $G(X)$ 与真实的目标分布 $R(X)$ 近似. 在 $G(X)$ 分布下采样的样本将可以近似视为在 $R(X)$ 分布下采样的样本.

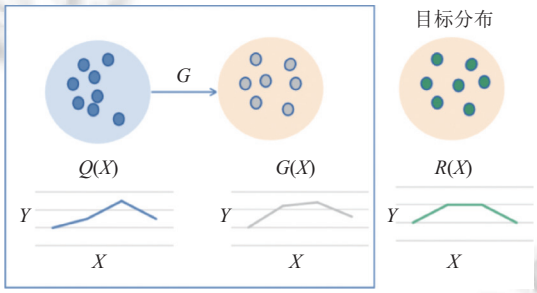


图 3 通过概率生成模型获得目标分布的变化过程

该如何确定这个目标分布呢? 这个分布下采样的数据应该介于对抗样本与原始样本分布之间. 用来过渡两个分布的差异. 假设原始样本分布为 $P(X)$, 对抗样本分布为 $Q(X)$. 所需的目标分布 $R(X)$ 可以表示为两个分布的加权和, 通过对原始样本和对抗样本分布的分析与组合来构建目标分布. 这种构建方式不会造成真实分布与原始分布的混淆, 因为它明确基于原始样本和对抗样本的分布特性进行加权组合, 如公式 (6), 旨在生成更有利于模型加固的样本分布, 而非对真实分布和原始分布进行混淆或错误关联. 并且, 对抗训练时仍会保留部分原始样本和对抗样本数据供模型学习.

$$R(X)=\alpha P(X)+(1-\alpha)Q(X) \tag{6}$$

其中, α 是一个权重参数, 取值范围为 (0, 1). 当 α 接近 0 时, 新样本分布更接近对抗样本分布; 当 α 接近 1 时, 新样本的分布更接近原始样本分布. PBAT 利用概率生成模型来学习对抗样本和原始样本之间分布的差异. 采用生成对抗网络作为概率模型的一种. 目标分布 $R(X)$ 被定义为原始样本分布 $P(X)$ 和对抗样本分布 $Q(X)$ 的加权和, 通过对抗过程, 生成器逐渐掌握对抗样本和原始样本分布的差异, 进而生成符合目标分布的增强训练样本. 例如, 设置攻击为 PGD 攻击获取对抗样本, 将其作为生成器的输入, 训练生成器使其输出接近目标分布的数据, 这个过程中

生成器不断学习对抗样本与原始样本分布的不同特征, 从而实现对两者分布差异的学习.

2.2 对抗鲁棒性增强的多样性数据构造

分析数据样本后, 则需要利用上述对目标分布的分析和模型工具, 来构造所需的多样性数据. 已知概率生成模型能够学习分布规律. 结合对抗性纯化模型的思路, 样本的对抗性扰动消除可以定义为从对抗样本到原始样本的映射问题, 可以从对抗样本中恢复出原始样本. 那么, 如果使用对抗样本和原始样本, 构造能将对抗样本映射到目标分布上的概率生成模型, 就能提供目标分布下的数据了.

首先, 需要采集对抗样本作为输入的 $Q(X)$ 分布下采样的样本. 设置攻击 *Attack* 为较强的 PGD (projected gradient descent) 攻击, 因为它采取了多步梯度下降策略, 相比仅执行单步梯度下降的一次性攻击, 如 FGM (fast gradient sign method), PGD 通常能够产生更鲁棒的对抗样本. 它的基本思想是迭代地更新对抗样本, 使其既能最大限度地欺骗网络, 同时又不会超出某个预定的扰动限制 (通常表示为 L 范数的约束). 公式 (7) 和 (8) 体现了 PGD 的攻击算法. 在每次迭代都使用梯度方向更新对抗样本, 同时将更新后的对抗样本投影回原始样本的邻域内, 确保扰动不超过预定的限制. 重复若干次, 最终得到的对抗样本应该能够极大地迷惑智能模型, 促使它做出错误的预测. 此外, PGD 攻击的投影步骤有助于保证每次迭代后的对抗样本都满足预定的扰动限制, 这在实践中是非常重要的, 因为对抗样本往往需要满足一定的可感知质量标准. 而在对抗性训练中, PGD 攻击也常被用作一种有效的手段来提升模型的鲁棒性, 通过在训练阶段引入该攻击下的对抗样本, 模型可以表现得更加鲁棒.

$$Attack_{\mathrm{PGD}}(x \in P(X)) = x' \in Q(X) \tag{7}$$

$$x_{t+1} = \prod_{x+S}[(x_t + \varepsilon \cdot \mathrm{sign}(\nabla_x J(x_t, y))] \tag{8}$$

之后, 代理分布概率生成网络 G 结构首先采取了广泛使用的 DCGAN 网络中的生成器 G 设计规范, 使用 encoder-decoder 卷积结构. 这也是去噪方法 APE-GAN[19]中使用到的模型结构. 训练捕获 $R(X)$ 的概率生成模型 G 也是方法中的关键一步. 后续实验中还会使用不同的模型结构并分析其效果. 设从 $Q(X)$ 分布中采样, 训练 G 到分布 $P(X)$, 则 GAN 训练所用损失函数可改为公式 (9) 的形式:

$$\min_G \max_D E_{x\sim P(X)} \log D(x) + E_{x\sim G(x'\in Q(X))} \log(1-(D(G(x')))) \tag{9}$$

此时训练的 G 将尽可能地将对抗样本 x' 所在分布投影到原始样本所在分布, 以欺骗鉴别器 D. PBAT 需要对抗样本不再只是投影到干净样本分布上, 而是投影到公式 (6) 中的分布. 为了使 G 能学习从 $Q(X)$ 到 $R(X)$ 的分布, 需要通过调整模型权重参数 $\hat{\theta}_G$ 优化一个指定的损失函数而得到 G, 具体如公式 (10)–(12):

$$\hat{\theta}_G = \arg\min_{\theta_G} \frac{1}{N} \sum_{i=1}^{N} L(G_{\theta_G}(x'_i), (\alpha x_i + (1-\alpha)x'_i)) \tag{10}$$

$$\min_G \max_D E_{x\sim P(X), x'\sim Q(X)} \log D(\alpha x + (1-\alpha)x') + E_{x\sim G(x'\in Q(X))} \log(1-(D(G(x')))) \tag{11}$$

$$\begin{aligned} L &= \eta_1 L_{\mathrm{mse}} + \eta_2 L_G \\ &= \eta_1 \left(\frac{1}{WH} \sum_{i=1}^{W} \sum_{j=1}^{H} (\tilde{x}_{i,j} - G(x')_{i,j}) \right)^2 + \eta_2 \left(\sum_{k=1}^{N} (1 - \log(D(G(x'_k)))) \right) \\ &= \eta_1 \left(\frac{1}{WH} \sum_{i=1}^{W} \sum_{j=1}^{H} ((\alpha x_{i,j} + (1-\alpha)x'_{i,j}) - G(x')_{i,j}) \right)^2 + \eta_2 \left(\sum_{k=1}^{N} (1 - \log(D(G(x'_k)))) \right) \end{aligned} \tag{12}$$

其中, $\tilde{x}_i$ $(i = 1, 2, \ldots, N)$ 是服从分布 $R(X)$ 的样本. 新样本分布来源于对抗样本, 并作为“协调者”, 被期望在对抗训练中平衡模型的鲁棒性与精度. 在训练过程中, 为了让模型学习到更有效的特征, 优化决策边界, 会将新数据分布当作一种待学习分布, 引导模型学习新数据所蕴含的信息, 以改进加固过程.

公式 (10) 旨在使概率生成模型 G 学习从对抗样本分布到目标分布的映射, 通过调整模型权重参数优化指定的损失函数得到. 公式 (11) 是基于公式 (10), 在训练 GAN 时的最大最小博弈公式. 它结合了公式 (10) 中模型的映射目标, 并将目标分布作为最终学习分布. 公式 (12) 是训练使用的损失函数, 为 MSE 损失和 GAN 的博弈损失加

权和. 它综合考虑了生成样本与目标样本之间的均方误差 (MSE 损失) 以及生成器和鉴别器之间的对抗博弈关系 (GAN 的博弈损失). 使得模型在学习目标分布时保证生成样本的质量. 公式 (10) 确定了模型的学习目标, 公式 (11) 基于此构建了对抗博弈框架, 公式 (12) 则定义了具体的损失函数.

到此, 就可以训练该代理分布模型, 图 4 展示了整个代理分布模型的训练过程, $A(\cdot)$ 在此表示用于生成 X' 的 PGD 攻击. 在公式 (12) 的指导下, 通过欺骗 D 辨别网络的判断达成对 G 的训练.

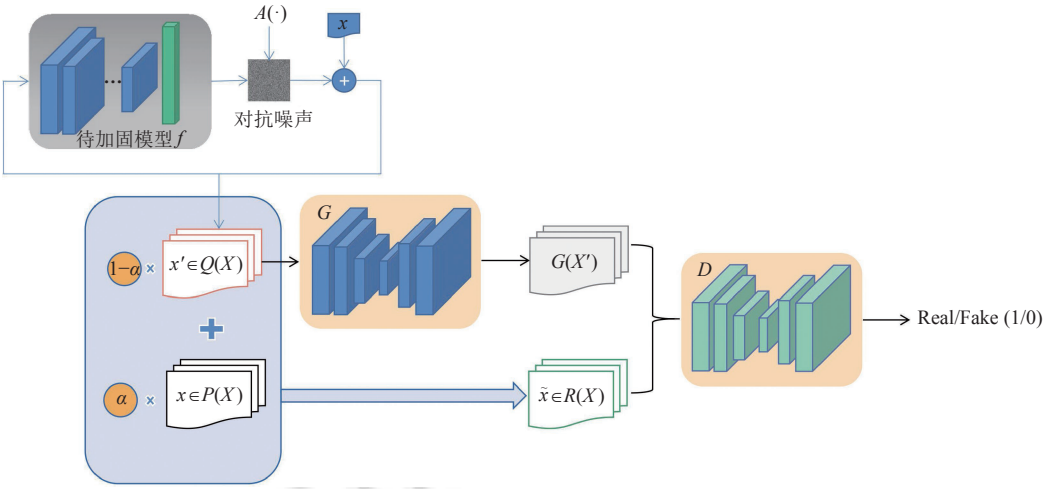


图 4 通过对抗博弈学习训练捕获目标分布 $R(X)$ 的代理分布模型

2.3 智能模型鲁棒决策边界训练

现有的对抗训练方法简单地将对抗样本加入训练过程中, 而忽略了决策边界在训练使用攻击下的过拟合. 从而损失了大量原始精度和泛化性. PBAT 利用混合重组的多样性数据重构训练过程, 不仅考虑了对抗训练的基础鲁棒性优势, 还考虑了平衡对抗样本和原始样本. 新训练所用的损失函数如公式 (13) 所示, 总体加固训练损失由 3 部分组成. 其中 L_{cln}、L_{adv}、L_{proxy} 分别为原始样本、对抗样本和从代理分布中采样样本的损失. 权重系数 λ_i ($i=1,2,3$) 用来调整各训练损失的比例.

$$L_{\text{PBAT}} = \lambda_1 L_{\text{cln}} + \lambda_2 L_{\text{adv}} + \lambda_3 L_{\text{proxy}} \tag{13}$$

在简单数据集上 (如 MNIST) 进行对抗训练时, PGD 下产生的对抗样本可能与原始样本在高维空间中有很大的分布差异, 因此模型在学习了这些样本之后通常会发生鲁棒性过拟合问题, 即模型在对抗样本决策上表现得过于完美, 以至于失去了对干净样例的泛化能力. 因而产生了一个奇怪的现象, 那就是在对抗性攻击下的对抗性样本分类精度高于原始样本的分类精度. 为了解决鲁棒过拟合的问题, 本文进一步调整损失函数来控制训练过程, 新的损失函数如公式 (14) 所示:

$$L_{\text{PBAT}} = (L_{\text{cln}} \mid\sim epoch \,\%\, \omega_1 = 0) \cap (L_{\text{cln}} \times \min(\omega_2, |\cos(epoch)|) + L_{\text{proxy}} \times \max(\omega_3, 1-|\cos(epoch)|) \mid epoch \,\%\, \omega_1 = 0) \tag{14}$$

我们利用 cos(·) 函数特性来平衡比例: 在训练初期, 模型如同余弦函数的起始阶段, 逐步探索数据特征; 到了中期, 类似余弦函数峰值附近, 学习速度和效果达到一个相对高点; 后期又类似余弦函数的回落阶段, 逐渐收敛. 此外, |cos(*epoch*)|的值域为 [0, 1], 随着 *epoch* 变化平滑波动, 使得权重调整不是突变的. 前期以固定参数 ω_2 为主, 后期随着|cos(*epoch*)|变化, 平滑地改变原始样本权重, 避免权重突然大幅变动对模型稳定性和学习效果产生不良影响. *epoch* 为当前训练轮数, $epoch \,\%\, \omega_1$ 是一个条件, 当余数不为 0 时, 只进行 L_{cln} 相关计算. 此时只学习原始数据. 起到了控制模型学习节奏的作用, $epoch \,\%\, \omega_1$ 余数为 0 的情况出现时, 模型再逐渐引入对抗样本和代理分布样本进行学习, 此时模型已具备一定的基础, 能够更好地融合不同样本的信息, 优化决策边界, 提升对多种样本的分类能力, 有效避免鲁棒性过拟合问题, 提高模型的整体性能和泛化能力. 参数 ω_2 和 ω_3 分别来限制原始样本和代理分

布样本的最大和最小边界. ω_2 设定了一个权重上限 (最大边界). 在训练过程中, |cos(*epoch*)|会随 *epoch* 变化, 但其值不会超过 ω_2. 如 $\omega_2 = 0.7$, 无论|cos(*epoch*)|如何变化, 原始样本损失权重最大就是 0.7. 这就限制了原始样本损失在总损失中占比的上限, 防止模型过度依赖原始样本学习, 避免在对抗训练中过于注重原始精度而忽视鲁棒性提升. 如果没有这个限制, 模型可能会在原始样本上过度拟合, 过于注重原始数据的分类精度, 而忽略了对对抗样本或其他复杂情况的学习, 导致模型的泛化能力和鲁棒性不足平衡学习. 此外, 确保在训练的不同阶段, 原始样本损失在总损失中的占比不会过高, 从而为代理分布样本损失等其他部分留出合理的比重, 使模型能够平衡地学习不同方面的特征, 提升综合性能. 设置 $\omega_2 = 0.3$, 那么代理分布样本损失权重最小为 0.3. 这保证了在训练过程中, 代理分布样本损失在总损失中始终有一定占比, 使模型不会完全忽略代理分布样本 (包含对抗样本相关特性) 的学习, 有助于提升模型的鲁棒性. 限制 ω_3 保证鲁棒性学习, 可以稳定训练过程, 使模型在训练过程中对代理分布样本的学习保持一定的稳定性. 避免因为权重过小而导致在某些训练阶段对代理分布样本的学习不足, 进而影响模型整体的训练效果和性能提升. 在简单数据集上, 对抗样本与原始样本分布差异显著, 若固定高比例对抗样本参与训练, 易导致模型过度适应对抗扰动, 产生鲁棒过拟合. 实验结果也表明了将 ω_3 设置为 0.3 的合理性. 公式 (13) 是 PBAT 的基础损失公式, 公式 (14) 是针对简单数据集上解决鲁棒过拟合的优化版本. 实验部分未特别说明时, 默认使用公式 (13); 仅在 MNIST 数据集上为解决鲁棒过拟合, 才启用公式 (14).

最后, 具体的训练过程如算法 1 的流程所示, 输入是待加固的原始模型与原始样本集合 X, 输出是训练后得到的, 获得了新的决策边界的模型参数文件. 轮次训练方法考虑了模型在训练过程中可能出现的鲁棒过拟合问题. 例如, 在简单数据集 (如 MNIST) 上, PGD 攻击下产生的对抗样本与原始样本在高维空间分布差异大, 容易导致模型出现鲁棒过拟合. PBAT 调整损失函数, 利用 cos(·) 函数特性平衡比例. 随着训练轮数 *epoch* 的变化, 通过调整相关参数来控制模型对不同样本的学习程度. 在训练初期, 模型可能更多地学习原始数据, 随着训练进行, 逐渐增加对代理分布样本的学习.

算法 1. PBAT.

输入: 待加固的神经网络模型 f_θ, 原始样本集合 X;
输出: 训练后得到的, 获得了新的决策边界的模型参数文件.

1. 初始化模型参数 θ, 设置超参数 λ_i $(i = 1,2,3)$, ω_i, 学习率 η
2. **for** $epoch = 1,2,\ldots,N$ **do**
3. **for** $batch\ B \subset X\{x_i,y_i\}_{i=1}^m$ **do**
4. **for** $i = 1,2,\ldots,m$ **do**
5. 生成对抗样本 $x_i' \leftarrow pgd\_generation(x_i)$;
6. 通过 G 生成代理分布样本 $\tilde{x}_i \leftarrow gan\_generation(x_i')$;
7. $L_{\text{cln}} = cross\_entropy(f_\theta(x_i),y_i)$;
8. $L_{\text{adv}} = cross\_entropy(f_\theta(x_i'),y_i)$;
9. $L_{\text{proxy}} = cross\_entropy(f_\theta(\tilde{x}_i),y_i)$;
10. 反向传播, 使用自适应学习率优化器 Adam 更新参数 $\theta \leftarrow \theta - \eta\sum_{i=1}^m \nabla_\theta(\lambda_1 L_{\text{cln}} + \lambda_2 L_{\text{adv}} + \lambda_3 L_{\text{proxy}})/m$;
11. **end for**
12. **end for**
13. **end for**

3 实 验

为了验证 PBAT 方法的有效性, 在多个数据集上进行了实验. 实验结果表明, 通过概率模型分析数据样本并基于样本分布构建多样性数据, 可以显著提高智能模型鲁棒性, 并适用于不同任务领域.

3.1 研究问题

- RQ1: 从加固后的模型对对抗样本的决策结果的角度, PBAT 与其他的防御方法比较是怎样的?
- RQ2: 从模型所生成的对抗样本质量的角度分析, PBAT 与其他的防御方法比较是怎样的?
- RQ3: 在使用了不同的代理分布模型时, PBAT 方法防御效果如何?
- RQ4: 在不同的超参数 α、λ_i、ω_i $(i=1,2,3)$ 的设置下, 防御效果变化情况又如何?
- RQ5: 在目标检测任务领域, PBAT 方法的效果如何?

3.2 数据集及模型

图像分类和目标检测是计算机视觉领域的两个核心任务, 需要算法具备理解图像中不同对象之间关系的能力. 通过大量数据的训练, 智能模型能够学习到越来越复杂的特征表示. 本文针对这两个核心任务下的典型模型进行实验, 用以说明方法的普遍有效性.

具体的, 针对分类任务, 将用 MNIST[53]和 CIFAR10[54]这 2 个数据集进行实验. 其中, MNIST 数据集由 70 000 张从 0 到 9 的手写数字的灰度图像组成, 尺寸为 28×28 像素, 其中 60 000 张作为训练集和 10 000 张作为测试集. 不同于 MNIST 手写数据集, CIFAR10 数据集由 10 种类别 (如卡车、飞机、马、青蛙等) 的 60 000 张 32×32 像素的三通道彩色图像组成, 其中 50 000 张为训练集, 10 000 张为测试集.

针对目标检测任务, 将在 PASCAL VOC2007 以及 PASCAL VOC2012[55]上进行, PASCAL VOC2007 中包含 9 963 张标注过的图片, 由训练集、验证集和测试集这 3 部分组成, 共标注出 24 640 个物体. PASCAL VOC2012 中包括 20 类物体, 训练集和验证集中有 11 530 张图片, 共有 27 450 个目标检测标签. 本文后续实验数据将 PASCAL VOC2007 和 PASCAL VOC2012 中的数据进行合并, 统一分析具体的实验结果.

在模型的选择上, 实验使用了 3 种广泛使用的模型架构: ResNet-20[56]、GoogLeNet[57]和 Faster R-CNN[58].

3.3 实验配置

所有实验均在 Intel Xeon Processor (Skylake, IBRS) 服务器进行, GPU 为 A100. 其操作系统为 Ubuntu 18.04.6 LTS, 内存为 125 GB. 测试代码使用 Python 语言及 PyTorch 框架实现.

(1) 对抗攻击设置

为了更好地评估 PBAT, 在实验中使用几种对抗攻击算法来对训练后的模型进行攻击, 以对其防御加固效果进行评测. 实验总共选取了 6 种应用广泛且强度不同的攻击方法: FGSM[59]、PGD[60]、AutoPGD[22]、DeepFool[61]、CW[62]和 DAG[63].

FGSM 攻击基于对抗样本存在的线性解释, 并利用损失函数、梯度和半径进行计算. 实验中, 使用 $\varepsilon=0.31$ 作为扰动参数; PGD 攻击使用鞍点公式的迭代攻击来寻找一个强扰动. 设置了攻击参数 $\varepsilon=2\times8/255$, 步长 0.01, 并迭代 40 次; 对于 AutoPGD 攻击同样设置了 $\varepsilon=2\times8/255$ 并迭代了 40 次; DeepFool 攻击是一种迭代攻击并能找到一个输入的最小扰动, 选取最后的扰动向量所乘的一个常数 (1+*overshoot*) 中的 *overshoot* 为 0.02 并设置最大迭代次数为 40 次; 随后, CW 攻击作为最强大的攻击之一, 使用了 3 种不同的向量规范 (L_0, L_2, L_∞). 使用平滑裁剪梯度下降方法并显示低失真 L_2 攻击规范的 CW 攻击, 设置其对每层查找迭代的学习率为 8/255, 迭代 20 次. 最后, DAG 作为针对目标检测模型的特殊攻击方法. 通过在多源的任务上进行约束达到攻击效果, 使模型在分类和定位上决策错误. 实验将在设置该攻击扰动迭代 50 次且强度参数 $\gamma=0.2$ 上开展.

(2) 评测指标

为了解决实验前所提出的问题, 针对问题的需要从不同角度考虑多种用于评测的指标来评估 PBAT 的性能, 并在每个数据集的测试集中采样了 1 000 张样本作为测试时的评估对象.

具体评测指标与说明如表 1 所示. 使用 ACC/ACC_CCAT 和 ACTC 指标来回答 RQ1; 使用 ALDp 指标来回答 RQ2; 使用 ACC 和 COS 指标来回答 RQ3 和 RQ4; 使用 AP 指标来回答 RQ5.

几种评测指标中, 最基础的衡量标准是希望模型的精度 ACC/AP 能够始终保持较高水平. 进一步的, ACTC 作为用来观察模型对攻击成功样本的分类置信度的指标. 可以通过该指标, 将目光锁定到被模型分类错误的样本当

中, 通过观察模型对这些无法决策正确样本的正确类别所在的置信度水平变化, 来判断模型最终的决策边界是否趋于理想化——实现了更好加固效果的模型应该减少之前对错误类别过高置信度的情况, 提升对正确类别的置信度. 我们可以认为, 使用更优的加固方法所加固后的模型, 其 ACTC 指标应该能有所提高, 因为它代表着模型对正确类别的分类置信度情况. 其次, 使用 ALDp 指标作为对攻击后样本失真度的评价, 可以作为评测所采用的攻击强度的直观数值表现. 通过该指标来观察对模型施加在 L_p 约束下的攻击后, 不同攻击作用在样本上的扰动效果. 较大的指标值说明对应的攻击扰动范围较广, 以突出本文方法在面对强攻击下防御的有效性.

表 1 评测指标与说明

对应问题	评测指标	指标说明
RQ5	AP	目标检测平均精准度
RQ1、RQ3、RQ4	ACC	分类准确率
	ACC_CCAT	针对CCAT拒绝对抗样本检测防御方法分类拒绝的样本百分比 (ACC_CCAT=拒绝的对抗样本数/总样本数)
	ACTC	正确分类类别的平均预测置信度 (攻击成功的样本, ACTC=样本正确的类模型给出的预测值之和/攻击成功个数)
RQ2	ALDp	样本平均L_p失真度 (L_0, L_2, L_∞)
RQ3、RQ4	COS	JS散度来衡量两个比较模型对攻击成功的样本输出概率的相似性

ACC_CCAT 将不参与几组 ACC 的比较中, 因为 CCAT 方法本身拒绝检测的性质使得该比较将不能公平展示模型自身的防御加固效果. 我们在本文中引用此指标仅作为对拒绝检测方法实际效果的表示, 以便说明该方法存在的缺陷.

(3) 比较方法及参数设置

为了进行对比分析, 复现了 4 种不同角度的防御方法. 分别为原始对抗训练的方法、ATLD (adversarial training with latent distribution) 方法、CCAT (confidence-calibrated adversarial training) 方法以及 TRADES (trade-off between robustness and accuracy) 方法. 在实验分析中还将通过图像化展示检测结果的方式来表现本文方法的有效性. 用来对比的加固方法以及 PBAT 方法的参数设置具体情况如下.

① 原始对抗训练 (projection-based adversarial training, PAT) 方法采用投影梯度下降 (projected gradient descent, PGD) 攻击生成对抗样本. 设定 PGD 攻击强度 $\varepsilon = 16/255$, 步长为 0.01, 迭代次数 40 次, 训练时将对抗样本与原始样本按 1:1 的比例混合, 以增强模型对对抗扰动的适应性. 针对目标检测模型, 选用经典的 DAG 攻击算法, 通过 50 次迭代生成对抗样本, 模拟真实攻击场景下的模型挑战.

② ATLD (adversarial training with latent distribution) 方法聚焦底层流形对抗训练, 通过诱发原始样本的潜在表示生成对抗扰动. 为探究其核心训练机制效果, 实验仅采用底层流形对抗训练后的模型作为对比对象, 暂不应用流形变换推理方法 (inference with manifold transformation, IMT). 辨别器深度设为 28, 加宽系数为 10; 攻击强度 $\varepsilon = 8.0/255\times4$, 迭代步数 *num_steps*=2, 步长 *step_size* = $8.0/255\times2$, 确保在统一条件下对比其与本文方法的差异.

③ CCAT (confidence-calibrated adversarial training) 方法着重于对抗样本的拒绝推测, 训练时采用初始扰动 ε=0.3、40 次迭代的 PGD 攻击. 测试环节以 99% 真阳性率 (TPR) 下的分类准确率为核心指标. 鉴于其与本文非鲁棒模型再训练的思路存在本质区别, 在对比分析中作为参考方法, 辅助剖析不同防御策略的技术特点.

④ TRADES (trade-off between robustness and accuracy) 方法设定扰动步长为 40, $\varepsilon = 0.3$, 并在 L_∞ 范数约束下开展训练. 该设置有助于验证其在权衡模型性能时的效果, 与本文方法形成多维度对比, 凸显研究创新点的优势. 而 RDC (robust diffusion classifier) 利用预训练扩散模型构建生成式分类器, 净化对抗噪声是将扩散模型直接转化为生成式分类器进行分类决策, 不需要针对特定的对抗攻击进行训练. ISDAT (image-semantic dual adversarial training framework) 通过生成对抗图像和对抗语义来增加对抗样本多样性侧重于挖掘图像语义层面的信息. 由于本文核心在于借助概率模型学习对抗样本和原始样本的分布规律, 生成兼具两者特性的多样性样本, 以此改进对抗训练过程, 实现决策边界的平滑优化, 提升模型的鲁棒性与精度. 这种差异使本文未对这两种方法进行对比, 从

技术复杂度和与实验环境的适配性上选择了与 PBAT 更具可比性的 4 种方法.

⑤ PBAT 默认使用 $\lambda_1:\lambda_2:\lambda_3=1:1:2$ 的固定权重比例分配样本损失, $\alpha=0.5$、$\omega_1=2$、$\omega_2=0.7$ 以及 $\omega_3=0.3$ 的超参数设置, 以平衡模型对原始精度和对抗鲁棒性的学习. 在 MNIST 的实验中, 引入公式 (14) 的动态权重调节机制.

4 实验结果与分析

为了回答提出的 5 个问题, 开展了两大类实验, 分别是基于分类模型的加固实验和基于目标检测模型的加固实验, 结果在第 4.1–4.4 节中展示. 后续实验结果表格中的第 1 行表示测试使用的 5 种攻击方法, 加固方法中, PAT 代表使用了标准 PGD 攻击生成对抗样本的对抗训练方法, ATLD、CCAT 和 TRADES 作为对比多种加固防御实现角度所复现的其他方法. PBAT 则为本文提出的使用代理分布模型利用新数据改进的对抗训练加固方法; 所使用的测试指标说明请参考表 1 中的内容. 特别地, CCAT 方法拒绝对抗样本角度来实现防御的特性, 将不再针对该方法在对抗攻击下的分类准确率 (ACC) 做出评测, 而是使用对抗样本拒绝率作为评价指标. 即上文中提到的单独设置的 (ACC_CCAT), 并不计入最终比较. 后续表格中用加粗表示最优结果.

基于目标检测模型的加固实验结果将在第 4.5 节介绍, 相关表格的训练方法列中, PBAT 为本文提出的使用代理分布模型利用新数据改进的对抗训练加固方法; 由于目标检测不同于分类任务, 每张样本不存在唯一一个类别标签. 因此在实验时, 选择了训练时无需提供标签的 APE-GAN 和 DISCO 去噪模型作为代理分布模型的选择. 第 1 行表示采用的传统对抗训练方法加固的模型在对抗样本上对模型的检测精度指标 AP 性能指标. 未将 ATLD、CCAT 和 TRADES 方法纳入对比, 主要是因为这 3 种方法在目标检测任务上存在适用性问题.

4.1 对于 RQ1 的回答

从表 2 所示的使用 ResNet-20 模型在 MNIST 数据集下的 ACC 指标实验结果可知, PBAT 方法展现出了显著优势. 在面对 FGSM 攻击时, PBAT 的分类准确率在几种对比方法中最高, 与原始的 PAT 对抗训练方法相比, 优势明显. 这表明 PBAT 能够更精准地将样本分类到正确类别, 意味着其决策边界在该攻击场景下更靠近真实的类别划分边界. 虽然在 PGD 攻击下未达最高鲁棒分类精度, 但考虑到 PGD 攻击在简单数据集上易引发鲁棒过拟合现象, PBAT 未过度偏向该攻击且仍保持较高准确率, 说明它能合理调整决策边界, 避免因特定攻击导致决策边界过度扭曲. 同时, 在训练未涉及的 DeepFool 攻击下, PBAT 的鲁棒性提高, 这意味着模型的决策边界具有更好的泛化性, 能够适应多种攻击场景, 而不是局限于训练时的攻击类型. 这种在不同攻击下的良好表现, 充分体现了 PBAT 可以有效改善决策边界, 更具适应性. 在更为复杂的 CIFAR10 数据集上, PBAT 方法依旧表现出色. 在各项攻击下, PBAT 的指标均优于对抗训练方法. 虽然在 AutoPGD 和 DeepFool 攻击下, PBAT 的 ACC 值略低于 TRADES 方法, 这可能是由于 AutoPGD 有一个动态调整步长的机制, 可以根据目标模型的特点和攻击需求选择合适的范数, 增加了攻击的灵活性和有效性. 当损失函数出现振荡时, 会将步长缩小. 这种自适应的步长调整能让攻击在不同阶段都能更有效地探索对抗样本空间. 且支持不同范数, 根据不同的范数类型采取不同的对抗样本生成方式. 此外, AutoPGD 会进行多次重启攻击每次重启都会重新初始化对抗样本, 增加了找到有效对抗样本的概率. 这种多轮重启策略可以在不同的初始条件下探索对抗样本空间, 避免陷入局部最优解. 而 PBAT 防御方法采用的是固定步长和固定范式, 没有考虑到多次重启攻击的情况, 难以适应 AutoPGD 这种动态的攻击节奏, 导致在攻击过程中无法及时应对. 但综合多种攻击下的整体防御性能, PBAT 方法最优. 尤其是面对 CW 攻击时, TRADES 加固后的模型无法防御实验所设置的攻击, 这是因为 TRADES 方法是在 L_∞ 约束下进行加固, 而 CW 攻击设定在 L_2 约束下, 两种约束衡量的扰动类型和程度不同, 针对一种约束训练的模型难以有效防御另一种约束下的攻击, 这也反映出 TRADES 方法在面对未知攻击时存在缺陷. 这些指标变化充分说明, PBAT 方法在复杂数据集上同样适用, 相较于原始对抗训练, PBAT 训练后的模型决策边界得到有效调节. 在训练中加入代理分布样本后, 加固模型的分类边界更接近理想的鲁棒决策边界, 不会因攻击类型的变化而出现大幅波动, 增强了模型防御多种类型攻击的能力.

从表 3 中的 ACTC 指标结果可见, PAT 下的加固模型在遭受几种攻击后, 对攻击成功样本的正确类别置信度较低. 而采用 PBAT 方法训练的模型, 该指标大多有一定程度的提高, PBAT 能提升该指标, 说明模型在决策时对

正确类别的判断更加自信, 进一步证明了其对决策边界的优化作用, 模型的决策结果更趋向于正确方向. 此外, CCAT 方法确实显著降低了攻击成功的对抗样本的错误类别分类置信度, 实现了一定的拒绝率. 然而, 同样存在明显缺陷: 会拒绝部分原本能够正确分类的样本. 这不仅大大损害了模型的分类精度, 而且在实际应用中, 由于数据通常具有较高价值, 难以直接舍弃样本, 导致方法在应用中存在较大局限.

表 2　对于 RQ1, 使用 ResNet-20 模型在不同数据集下的 ACC 指标实验结果 (%)

数据集	加固方法	FGSM	PGD	AutoPGD	DeepFool	CW
MNIST	PAT (对抗训练)	43.0	96.7	96.7	1.1	12.4
	ATLD	8.9	8.9	8.6	5.5	8.9
	CCAT	100.0	100.0	85.4	90.9	87.4
	TRADES	53.6	**98.6**	**98.6**	0.7	12.3
	PBAT	**93.1**	95.5	65.5	**21.4**	6.4
CIFAR10	PAT (对抗训练)	38.5	22.9	6.5	9.2	**10.3**
	ATLD	26.8	22.0	5.3	12.7	9.0
	TRADES	17.4	18.9	**15.3**	**15.4**	0.0
	PBAT	**48.7**	**34.9**	11.7	14.6	10.0

表 3　对于 RQ1, 使用 ResNet-20 模型在不同数据集下的 ACTC 指标实验结果

数据集	加固方法	FGSM	PGD	AutoPGD	DeepFool	CW
MNIST	PAT (对抗训练)	0.10773	0.25251	0.31052	0.37747	0.03659
	ATLD	0.02687	0.02506	0.03911	0.17636	0.02831
	CCAT	0.11833	0.0998	0.10461	0.14398	6.7592E–06
	TRADES	0.1105	**0.27708**	**0.33207**	0.39024	0.0537
	PBAT	**0.13741**	0.176477	0.1239	**0.40333**	**0.27217**
CIFAR10	PAT (对抗训练)	0.08167	0.06429	0.07189	**0.45379**	0.04531
	ATLD	0.10516	0.09831	0.1125	0.20273	0.08057
	TRADES	0.1063791	**0.10704**	**0.15576**	0.35316	**0.12553**
	PBAT	**0.11703**	0.10115	0.12455	0.38372	0.06821

进一步地, 实验更换了更为复杂的模型结构, 得到了表 4、表 5 中的数据. 分析实验结果数据可知, 使用更复杂的 GoogLeNet 模型, 本文提出的方法在 5 种攻击下同样获得了优于对抗训练的效果, 且获得了最好的 5 种攻击下的平均精度.

表 4　对于 RQ1, 使用 GoogLeNet 模型在不同数据集下的 ACC 指标实验结果 (%)

数据集	加固方法	FGSM	PGD	AutoPGD	DeepFool	CW
MNIST	PAT (对抗训练)	15.1	77.3	78.8	2.9	1.7
	ATLD	9.0	9.0	8.0	3.7	8.8
	TRADES	9.2	**97.8**	**97.8**	3.6	**11.9**
	PBAT	**90**	94.2	84.8	**9.7**	8.2
CIFAR10	PAT (对抗训练)	33.9	27.3	**10.3**	14.0	10.3
	ATLD	24.8	21	4.2	17.6	10.0
	TRADES	26.2	21	5.3	**17.4**	0
	PBAT	**41.6**	**29.3**	7.7	14.0	**11.2**

特别地, 实验在 MNIST 数据集上, 利用了公式 (16) 中的防止鲁棒过拟合的训练损失函数, 实现了简单数据集上的高鲁棒性能的表现, 再一次体现出 PBAT 针对一阶攻击的良好加固性能. 加固后的模型在保证不损失大量原始精度的同时获得了对多种攻击下的高鲁棒性. 这表明 PBAT 能够在复杂模型中有效平衡鲁棒性和精度, 其决策边界能够在抵抗多种攻击时更准确地对样本进行分类, 避免因追求鲁棒性而过度牺牲原始精度, 进一步说明了 PBAT 对决策边界的优化效果. 结果中对抗训练方法下部分低鲁棒精度的结果获得了高 ACTC 指标可能与模型的复杂程度的提高有关. 更复杂的 GoogLeNet 模型其决策空间维度更大, 决策行为的复杂程度更高, 此时 ACTC 指

标预测值的变化范围更大. 模型在错误类别同样拥有更大的置信度水平, 更能反映模型的决策边界波动大, 容易产生极端值. 而 TRADES 方法对不同类型的攻击的普遍防御性能则逊色于本文的方法. PBAT 加固后的模型能更好地防御对抗攻击, 整体性能最好.

表 5　对于 RQ1, 使用 GoogLeNet 模型在不同数据集下的 ACTC 指标实验结果

数据集	加固方法	FGSM	PGD	AutoPGD	DeepFool	CW
MNIST	PAT (对抗训练)	0.05759	0.16409	0.22058	0.40543	0.12974
	ATLD	0.04220	0.04017	0.04915	0.21041	0.04206
	TRADES	0.011917	**0.270894**	**0.31174**	0.44210	0.12935
	PBAT	**0.09513**	0.109418	0.09732	**0.47415**	**0.26966**
CIFAR10	PAT (对抗训练)	0.10581	0.09928	**0.12953**	**0.36150**	0.07359
	ATLD	0.09399	0.09166	0.11732	0.18897	0.08785
	TRADES	0.09665	**0.10222**	0.07912	0.10159	**0.20224**
	PBAT	**0.11625**	0.09699	0.12573	0.356575	0.05942

综合来看, PBAT 方法在不同模型 (ResNet-20 和 GoogLeNet) 和数据集 (MNIST 和 CIFAR10) 上, 通过在各种攻击下提高分类准确率 (ACC)、提升正确分类类别的平均预测置信度 (ACTC) 以及展现出良好的平均防御效果, 充分证明了其能够有效改善决策边界, 使模型的决策边界更趋向于理想的鲁棒决策边界, 增强了模型对多种攻击的防御能力.

4.2　对 RQ2 的回答

针对 RQ2, 从对抗样本失真度角度设计实验, 结果如表 6 所示. 3 个指标结果值分别代表 L_0、L_2 和 L_∞ 失真度. L_0、L_2 距离用于衡量对抗样本与原始样本在特征空间的差异, 反映视觉欺骗性和细微改变程度, 距离越小表明对抗样本越接近原始样本, 越易绕过图像识别系统检测; L_∞ 值则体现最大扰动幅度, 在目标模型鲁棒性较强时, 需增大 L_∞ 值以确保对抗样本成功攻击; ALDp 指标直观量化输入扰动程度, 数值越大代表扰动越强, 间接反映模型鲁棒性水平.

表 6　使用 ResNet-20 模型在 MNIST 和 CIFAR10 数据集下的 ALDp 指标结果

数据集	方法	FGSM	PGD	AutoPGD	DeepFool	CW
MNIST	PAT (对抗训练)	(0.99671, 0.95372, 1.0)	(0.99667, 0.95374, 1.0)	(0.96870, 0.92696, 0.97155)	(0.99999, 0.09949, 0.64579)	(1.0, 1.0, 1.0)
	ATLD	(0.99877, 0.96346, 1.0)	(0.99884, 0.96346 1.0)	(0.88749, 0.85676, 0.88840)	(0.99997, 0.26221, 2.07590)	(1.0, 1.0, 1.0)
	CCAT	(0.99542, 0.97512, 1.0)	(0.70845, 0.07170, 0.03137)	(0.78142, 0.06577, 0.03068)	(0.99950, 0.00277, 0.00244)	(0.25859, 1.0, 0.99284)
	PBAT	(0.99721, 0.96226, 1.0)	(0.99763, 0.96110, 1.0)	(0.97431, 0.93594, 0.97637)	(1.0, 0.10696, 0.66620)	(1.0, 1.0, 1.0)
CIFAR10	PAT (对抗训练)	(1.0, 0.77685, 2.16694)	(1.0, 0.76514, 2.16527)	(0.84171, 0.63349, 1.81952)	(0.99996, 0.02092, 0.18030)	(1.0, 1.0, 2.48444)
	ATLD	(0.99600, 0.74181, 0.90931)	(0.99648, 0.73484, 0.91005)	(0.75378, 0.547104, 0.69160)	(0.999994, 0.06613, 0.25967)	(1.0, 1.0, 0.96346)
	CCAT	(0.99384, 0.06285, 0.03137)	(0.99720, 0.04554, 0.03137)	(0.88902, 0.11988, 0.08899)	(0.99932, 0.00089, 0.00308)	(0.99761, 1.0, 0.940446)
	PBAT	(1.0, 0.78568, 2.17947)	(1.0, 0.77587, 2.16681)	(0.66931, 0.50271, 1.44745)	(0.99999, 0.02448, 0.22013)	(1.0, 1.0, 2.50250)

实验数据表明, 相较于原始对抗训练 (PAT) 方法, PBAT 生成的对抗样本在 ALDp 指标上显著提升. 这一结果意味着 PBAT 生成的对抗样本需施加更强扰动才能突破模型防御, 揭示了 PAT 的关键差异, 即, 传统对抗训练模型防御边界脆弱, 易被轻微扰动突破; 而 PBAT 需施加更强扰动才能成功攻击模型. 结合 RQ1 中 PBAT 在分类准确率 (ACC) 和正确类别置信度 (ACTC) 上的优势, 双重验证了相较于 PAT 的核心改进, 实现了训练范式升级.

4.3 对 RQ3 的回答

针对 RQ3, 使用 CAFD (以 RaGAN 为原型, 通过由类激活特征损失和对抗性损失组成的混合损失函数训练, 最小化类激活特征空间中对抗样本与原始样本的距离来消除对抗噪声), APE-GAN (利用融合损失函数使对抗样本与原始样本流形保持一致, 去除对抗扰动), DISCO (通过编码器和局部隐式函数预测原始图像 RGB 值, 以局部像素信息输入实现对抗性去除) 这 3 种模型结构作为代理分布模型的选择, 代理模型是指用于学习样本分布规律并生成新数据 (代理分布样本) 的概率生成模型. 并对待加固模型进行了加固训练, 之后再对加固效果进行比较. 实验的结果如表 7 所示. 使用的指标为 COS 值, 即对攻击成功的样本使用了 JS 散度值来衡量在使用不同代理分布模型时与原始对抗训练方法加固后模型输出概率的相似性之间的差别. 以此来说明在采用不同代理模型结构后对模型的决策结果是否有显著性的区别. 对于 MNIST 数据集和 ResNet-20 模型, 模型对攻击成功的样本输出概率的相似度普遍高于 CIFAR10 数据集和 GoogLeNet 模型下的相似度. 猜测是因为在简单模型和简单数据集上, 模型的学习空间维度较小, 模型最终行为存在的差异并不大. 而代理模型的选择和数据集存在较大的相关性. 在 MNIST 数据集上 DISCO 表现最优, 而在 CIFAR10 数据集上 CAFD 作为代理模型表现最优. 但基于 APE-GAN 在两类数据集上表现较为均衡, 整体都有较好的效果. 之后将作为本文默认使用的代理模型结构的选择. 在第 4.1 和 4.2 节则根据代理模型在任务上的适用性动态在三者之间进行择优选择. 进一步分析 COS 指标发现, 不同加固方法后模型的输出相似性都不太高, 说明本文的方法与一般的对抗训练方法确实对模型的决策行为有明显的差异, 突出本文方法和对抗训练的不同之处.

表 7 PBAT 分别使用 APE-GAN、CAFD 与 DISCO 作为代理分布模型的实验结果

数据集	模型结构	加固方法	测试指标	FGSM	PGD	AutoPGD	DeepFool	CW
MNIST	ResNet-20	APE-GAN+对抗训练	ACC	87	94.6	**94.5**	2.2	**12.2**
			COS	46.6	46.9	45.8	33.4	45.8
		CAFD+对抗训练	ACC	8.9	8.9	8.5	3.0	8.9
			COS	64.5	68.3	29.3	47.7	68.6
		DISCO+对抗训练	ACC	**93.1**	**95.5**	65.5	**21.4**	6.4
			COS	13.5	**12.8**	33.7	43.6	4.7
	GoogLeNet	APE-GAN+对抗训练	ACC	55.2	93.2	**93.2**	1.9	8.3
			COS	45.5	34.2	30.6	30.3	30
		CAFD+对抗训练	ACC	11.4	10.6	8.7	3.1	**8.5**
			COS	57.5	57.6	49.8	30.5	60
		DISCO+对抗训练	ACC	**90**	**94.2**	84.8	**9.7**	8.2
			COS	21.6	**26.5**	52.5	45.4	5.5
CIFAR10	ResNet-20	APE-GAN+对抗训练	ACC	34.5	21.9	4.5	**15.7**	**10.3**
			COS	16.7	17.7	19.2	24.0	12.0
		CAFD+对抗训练	ACC	**48.7**	**34.9**	**11.7**	14.6	10.0
			COS	18.8	19.7	19.4	21.2	12.2
		DISCO+对抗训练	ACC	34.5	26.6	9.2	13.2	**10.3**
			COS	16.8	**17.1**	16.9	18.9	10.6
	GoogLeNet	APE-GAN+对抗训练	ACC	41.6	29.3	7.7	14.0	10.0
			COS	11.7	12.8	11.1	10.9	2.1
		CAFD+对抗训练	ACC	**58.9**	**38.6**	7.0	**17.4**	**11.2**
			COS	18.8	22.6	21.8	24.0	19.5
		DISCO+对抗训练	ACC	30.0	26.1	**9.2**	16.5	10.3
			COS	8.0	**8.2**	8.4	11.2	3.2

通过对结果 ACC 值进行参数检验. 如图 5 所示, 使用 Kruskal-Wallis 检验统计量进行分析发现, 不同加固方法对于 5 种攻击方法不会表现出显著性 (p>0.05). 但使用 DISCO 和 APE-GAN 作为代理模型时在不同模型和数据集上都有较好的表现, 加固效果整体更好.

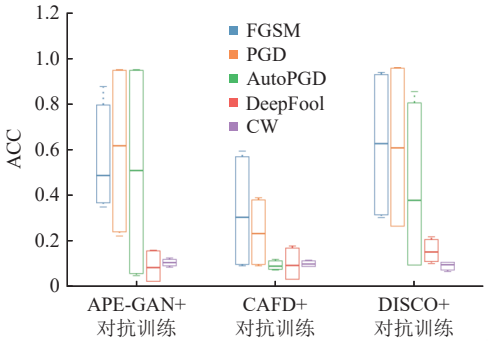


图 5　5 种攻击方法在不同组别间对于 ACC 的结果箱线图

4.4　对 RQ4 的回答

为了观察代理模型学习目标分布时对抗样本占比的变化对代理模型效果的影响, 对超参数 α 设置在 0.25、0.5、0.75 这 3 组设置下进行了实验, α 用于调整目标分布各类样本的权重, 其取值影响生成样本的特性. 当前设置的 3 个值涵盖了从偏向对抗样本到偏向原始样本的不同情况, 初步展示了 α 对模型的影响范围. 在当前研究中选择这 3 个值, 是因为它们能够体现出明显的结果差异, 帮助研究人员初步了解不同样本比重对效果的影响规律. 进一步的, 通过评测鲁棒精度 ACC 指标变化情况, 分析不同参数 α 下训练出的代理模型对后续加固的影响情况. 再针对参数 λ_i 和 ω_i $(i=1,2,3)$ 对加固训练的影响, 依次设计并实施了几组消融实验来确定对参数的选择. 其中, 表 8 和表 9 为设置的 4 组 λ_i 和 ω_i $(i=1,2,3)$ 的值. 表 8 中第 1 组属于平均分配 3 种类型的训练样本的比例, 第 2 组设置了一半的训练样本为原始样本, 第 3 组设置了一半的训练样本为对抗样本, 第 4 组则设置了一半的训练样本为代理分布样本, 而其他两个类型的样本平分剩下的权重. 表 9 中针对超参数 ω_i 分别设计为每两个 *epoch* 使用到代理分布样本, 每 3 个 *epoch* 后使用到代理分布样本, 每 4 个 *epoch* 后使用到代理分布样本, 以及每 5 个 *epoch* 后使用到代理分布样本. 原始样本与代理分布样本的比例基线为 0.7 和 0.3. 这是因为从实验中经验发现在这个比例下能更好地避免 MNIST 数据集的鲁棒过拟合现象, 因此将这个比例作为实验的默认设置. 对于每个 λ 设置中, 使用 4 组 $\omega_1=\{2,3,4,5\}$ 进行实验后将最好的结果作为对应实验结果. ω_2和ω_3 为当前基于经验的设置, 用于保障训练稳定性. 模型在训练初期更侧重于学习原始样本特征, 这符合深度学习中“先基础后复杂”的学习规律. 原始样本蕴含数据的基础语义信息, 优先学习这些信息可使模型构建稳定的基础特征提取能力, 避免过早受到对抗样本复杂扰动的干扰, 维持训练初期的稳定性. 确保模型在学习新特征的同时, 不会丢失对原始数据分布的适应性, 使训练过程在探索新特征与保持原有能力之间达到平衡. 表 10–表 12 则展示了以上消融实验的结果. 对每次的实验结果使用攻击下的鲁棒分类精度 ACC 作为比较评测指标.

表 8　对于 RQ4, 消融实验中训练过程损失函数的超参数 λ_i $(i=1,2,3)$ 的设置

实验组别	λ_1	λ_2	λ_3
1	1/3	1/3	1/3
2	1/2	1/4	1/4
3	1/4	1/2	1/4
4	1/4	1/4	1/2

表 9　对于 RQ4, 消融实验中训练过程损失函数的超参数 ω_i $(i=1,2,3)$ 的设置

实验组别	ω_1	ω_2	ω_3
1	2	0.7	0.3
2	3	0.7	0.3
3	4	0.7	0.3
4	5	0.7	0.3

表 10、表 11 与表 8 的实验组别对应, ω_i 只作为控制鲁棒过拟合时使用的参数. 由表 10 和表 11 的数据可知, 分析在 MNIST 数据集上表现最优的 DISCO 代理模型, 发现在两种待加固模型上, 把训练集的一半设置为代理分布样本时, 得到的加固模型在 5 种攻击下实现了更好的防御效果. 其平均鲁棒精度表现在对应的其他 3 组中最高. 在训练数据集中代理分布样本的高占比帮助模型获得了更好的鲁棒性这一结果, 表现出本方法使用新增数据样本加固训练过程的有效性. 进一步分析在 CIFAR10 数据集上表现最好的 CAFD 代理模型, 发现在保持一半原始训练

样本不变的情况下, 替换一半的对抗样本为代理分布样本时, 得到的加固模型在 5 种攻击下实现了更好的防御效果. 其平均鲁棒精度表现在对比的其他 3 组中最高. 与简单 MNIST 数据集不同, 复杂数据集上的加固训练原始样本有必要保持一定的参与度, 而不是仅使用对抗样本. 它能有效调控对抗训练过程, 对平衡原始精度与鲁棒性起到帮助. 同时, 代理分布样本作为部分对抗样本的代替者, 训练中使用后提升了加固效果. 最终的鲁棒加固模型表现优于仅使用对抗样本时对抗训练后的模型. 最终可以认为, 保持一半的原始样本不变, 并将再训练使用到的一半对抗样本替换为代理分布样本的数据分布设置, 更有利于在更广泛的模型结构和多种数据集上推广. 我们将这个比例设置为默认比例. λ_3 占比 50% 时, 模型在数据集上的平均鲁棒精度最优, 说明代理分布样本的高参与度可有效平衡原始精度与对抗鲁棒性, 验证了新生成的多样性数据对决策边界优化的关键作用. 而 ω 的核心作用是改进简单数据集上的鲁棒过拟合问题, 本文实验并未强调对它的细化分析, 而是为具体使用时提供经验调整方案. $\omega_1 = 2$ 结合 $\omega_2 = 0.7$ 和 $\omega_3 = 0.3$ 的边界限制, 模型在 PGD 攻击下的分类精度 (ACC) 达 95.5%, 原始样本精度未显著下降, 验证了该参数对平衡鲁棒性与原始精度的有效性.

表 10　对于 RQ4, 在不同组别下 MNIST 数据集上的鲁棒分类精度 (ACC) 结果 (%)

模型结构	加固方法	实验组别	FGSM	PGD	AutoPGD	DeepFool	CW
ResNet-20	APE-GAN+对抗训练	1	50	90.3	90.3	3.4	4.1
		2	**87**	94.6	94.5	**2.2**	**12.2**
		3	35.7	93	92.9	1.4	3.3
		4	61.9	**95.1**	**95.2**	1.7	0
	CAFD+对抗训练	1	8.9	**8.9**	**8.6**	2.7	8.9
		2	8.9	**8.9**	8.5	2.9	8.9
		3	8.9	8.0	7.0	**19.5**	**11.6**
		4	**9.6**	8.6	4.5	7.5	**11.6**
	DISCO+对抗训练	1	88.3	94.1	76.1	12.1	3.5
		2	**93.1**	**95.5**	65.5	**21.4**	6.4
		3	89.7	95.1	69.1	13.4	5.1
		4	84.3	93.6	**87**	7.4	**10.4**
GoogLeNet	APE-GAN+对抗训练	1	**55.2**	**93.2**	**93.2**	1.9	**8.3**
		2	55	89.1	89.2	3	4.2
		3	45.2	92.8	93.1	**3.2**	0.4
		4	41.4	91.1	91.2	2.1	4
	CAFD+对抗训练	1	**17.1**	**17.0**	**15.9**	3.1	**12.6**
		2	11.4	10.6	8.7	3.1	8.5
		3	12.7	12.3	10.7	**3.2**	8.5
		4	10.2	10.0	8.7	2.4	11.0
	DISCO+对抗训练	1	89.4	93.3	80	**12.6**	**11.2**
		2	**90.3**	95.2	73.8	10.1	8.8
		3	**90.3**	**95.8**	74.2	9.4	8.3
		4	90	94.2	**84.8**	9.7	8.2

表 11　对于 RQ4, 在不同组别下 CIFAR10 数据集上的鲁棒分类精度 (ACC) 结果 (%)

模型结构	加固方法	实验组别	FGSM	PGD	AutoPGD	DeepFool	CW
ResNet-20	APE-GAN+对抗训练	1	36.1	19.6	4.2	14.3	**10.3**
		2	34.5	21.9	4.5	**15.7**	**10.3**
		3	**38.4**	22.8	6.0	14.9	**10.3**
		4	35.5	**23.8**	**8.2**	13.7	9.0
	CAFD+对抗训练	1	45.8	33.7	12.7	12.6	**10.3**
		2	**48.7**	34.9	11.7	**14.6**	10.0
		3	43.4	33.6	**13.5**	12.7	9.0
		4	46.6	**35**	13.2	11.0	10.3

表 11 对于 RQ4, 在不同组别下 CIFAR10 数据集上的鲁棒分类精度 (ACC) 结果 (%) (续)

模型结构	加固方法	实验组别	FGSM	PGD	AutoPGD	DeepFool	CW
ResNet-20	DISCO+对抗训练	1	**36.4**	**27.8**	**9.7**	**14.6**	**10.3**
		2	34.5	26.6	9.2	13.2	**10.3**
		3	31.8	22.8	7.6	11.9	**10.3**
		4	35.7	26.4	9.5	13.0	**10.3**
GoogLeNet	APE-GAN+对抗训练	1	**47.0**	**37.7**	**37.8**	14.3	10.6
		2	41.6	29.3	7.7	14.0	10.0
		3	**47.0**	35.3	13.9	13.1	**10.3**
		4	40.9	31.6	10.5	**16.8**	**10.3**
	CAFD+对抗训练	1	45.2	35.0	13.8	12.6	10.3
		2	**58.9**	**38.6**	7.0	**17.4**	**11.2**
		3	34.4	28.8	**14.3**	13.4	10.3
		4	45.6	31.8	12.6	12.4	10.3
	DISCO+对抗训练	1	37.4	31.0	10.4	**17.2**	**10.6**
		2	30.0	26.1	9.2	16.5	10.3
		3	**41.3**	**32.1**	11.0	14.7	9.0
		4	29.0	25.1	**11.3**	16.2	10.3

表 12 展示了针对参数 α 的不同设置下, 所训练出的代理模型对加固效果的影响情况. 从表 12 的数据可知, 使用 50% 的对抗样本作为训练代理模型时对抗样本的参与程度, 所获得的代理模型将更有利于后续模型的加固过程. 最后, 可以认为, 较为平衡的数据占比下, 使得训练的代理模型生成的新数据样本更有利于模型的加固过程. 此时代理模型学习到的数据分布规律能有效提高鲁棒训练时使用的数据集的质量, 帮助获得更优鲁棒精度的加固模型. 而 0.75 的 α 设置下, 对后续的加固表现相对较差, 猜测是由于原始数据分布占比较重, 训练的代理模型所生成的新数据对待加固模型获得更优的鲁棒性影响较少, 新数据的分布对加固训练的调整能力较小导致的.

表 12 对于 RQ4, 使用不同参 α 训练出的代理模型对鲁棒分类精度 (ACC) 的影响 (%)

模型结构	α	FGSM	PGD	AutoPGD	DeepFool	CW
ResNet-20	0.25	84.9	**95.9**	**95.9**	2.5	10.1
	0.5	**87**	94.6	94.5	2.2	**12.2**
	0.75	23.7	76.1	76.7	**3.4**	1.2
GoogLeNet	0.25	73.8	78.3	77.5	**3.6**	**10.8**
	0.5	**91.7**	**94.9**	**94.8**	3.1	4.2
	0.75	33.5	82.3	81.9	3.1	0.9

4.5 对 RQ5 的回答

为了弄清本文的方法在其他计算机视觉领域是否有效, 实验进一步在更复杂的目标检测领域任务上进行了探索. 使用 PASCAL VOC 数据集进行了一系列针对目标检测模型的加固实验. 将加固后的模型和原始模型进行比较, 得到表 13 中的最终实验结果. 可以发现, 在 PBAT 加固方法下模型的 AP 指标得到了显著的提高. 该结果进一步说明本文方法的普遍有效性, 可以有效帮助各类深度学习模型获得良好的对抗鲁棒性, 并适用于各类图像任务.

为了清楚地展示本文方法的优势, 图 6–图 8 中的例子可视化地表现出模型最终检测结果所发生的改变. 可以发现, 未经加固方法的原始模型在对抗样本上性能损失严重: (a)、(b) 两张子图在肉眼上并无明显差别, 但却导致模型难以检测出原先可以正确识别的“bottle”和“diningtable”, 甚至还错误地将“bottle”识别成“person”. 正确识别出的“person”的置信度也从 95% 降低到了 61% (图 6). 而对抗训练后可以明显发现模型在对抗样本上 (图 7(b)) 的检测结果得到了改善, 能够正确识别出相应的物品, 但在原始图片上 (图 7(a)), 模型对“person”的识别效果出现一定损伤, 出现了一些不准确的候选框. 最后, 展示了在 PBAT 方法下加固的模型, 模型在对抗样本上对图中的物品均实现了高置信度的正确检测 (后文图 8(b)), 且在原始样本上的检测结果也没有出现对抗训练后出现的偏差情况 (后文

图 8(a)). 结果共同体现出模型在本文方法下得到了更有效的加固.

表 13 目标检测模型在 PASCAL VOC 数据集上对于 DAG 攻击的 AP 指标情况

模型结构	方法	AP	AP50	AP75
Faster R-CNN-FPN	原始模型	0.001 8	0.008 4	0.000 4
	对抗训练	0.004 8	0.018 2	**0.002 2**
	PBAT (GAN)	0.006 3	0.028 3	0.001 8
	PBAT (DISCO)	**0.007 5**	**0.029 1**	0.000 7
Faster R-CNN-C4	原始模型	0.013 8	0.046 4	0.002 5
	对抗训练	0.014 0	0.057 7	0.001 0
	PBAT (GAN)	**0.037 4**	**0.286 5**	0.000 0
	PBAT (DISCO)	0.012 9	0.052 3	**0.004 0**

(a) 原始样本 (b) 对抗样本

图 6 未加固的 Faster R-CNN 模型的检测结果

(a) 原始样本 (b) 对抗样本

图 7 对抗训练后 Faster R-CNN 模型的检测结果

4.6 实验结果进一步分析

PBAT 在各组实验中均获得了最优的表现, 说明使用本文提出的代理分布模型去作为学习样本数据的分布规律的工具, 并构造新的训练数据, 丰富了鲁棒训练数据集, 比对抗训练方法更有利于模型获得对抗鲁棒性. 并广泛适应多种攻击方法. 使用新分布下的数据改进了标准对抗训练的防御效果, 在不同计算机视觉领域, 以及多个数据集和模型结构上, 获得了更好的训练效果, 特别是对于一阶攻击有着出色的防御表现.

从实验结果来看, 在不同数据集和模型上, 这种轮次训练方法有效提高了模型的鲁棒性和泛化能力, 侧面证明了其对训练稳定性的积极作用. 这种动态调整的方式有助于避免模型对对抗样本分布的过度拟合, 使得训练过程更加稳定, 而不是造成训练不稳定.

(a) 原始样本 (b) 对抗样本

图 8 PBAT 训练后 Faster R-CNN 模型的检测结果

从根本原因上分析, PBAT 降低了对抗训练后模型发生鲁棒过拟合的可能性, 从正确类别的分类置信度有所提高这一表现可以看出, 模型的决策边界过度偏向对抗样本这一对抗训练方法中存在的典型问题得到了一定的减轻. 鲁棒决策边界更加合理. 从在多种类型的攻击下具有最好的平均鲁棒精度这一表现来看, 提高了模型对广泛攻击的普遍防御性. 体现出新数据的加入确实有利于再训练这一加固过程. 而对于不同的参数进行的一系列消融实验, 进一步确定了较为均衡的训练数据的设置将更有利于鲁棒性的获得, 达到调整加固训练的效果.

本文提出的 PBAT 方法默认使用了对抗生成模型结构, 让代理分布模型学习目标分布, 生成帮助实现原先两个分布之间的平滑过渡的新数据——代理分布样本. 利用代理分布样本这一增强数据, 帮助模型更好地获得鲁棒边界, 从而训练出在对抗攻击下有更好表现的加固模型. 同时, 还进一步提出了为防止在简单 MNIST 数据集上会出现的鲁棒过拟合现象而设计的新损失函数, 实验结果也证明了使用该损失函数后能有效地阻止这一现象发生.

5 总结与展望

在本文中, 提出了一种基于代理分布的对抗训练方法 PBAT, 在典型的图像领域的智能模型 ResNet-20、GoogLeNet 和 Faster R-CNN, 3 个典型的数据集 CIFAR10、MNIST 和 PASCAL VOC 上, 以及图像分类和目标检测两大领域任务上开展了大量实验. 实验结果表明, 基于代理分布的对抗训练方法可以在不同的模型及任务之间进行转移, 并相较于其他方法具有较好的表现, 有利于模型把握数据分布, 降低模型鲁棒风险, 提高泛化, 帮助模型获得更好的安全性, 为解决智能模型安全领域问题提供一个新的思路.

本文认为, PBAT 方法有利于智能模型获得更好的对抗鲁棒性, 是一个优于对抗训练的有效的模型加固方法. 在未来, 将继续探索更优质的概率模型和数据多样性技术, 进一步改善训练数据集的分布构成, 在更新更广泛的领域开展实验, 以解决现存的问题, 进一步提高在面对更强更新的迭代攻击时的防御性能, 保证智能模型的安全. 另外, 未来还将在更多实际项目以及智能模型应用领域继续开展实验, 以不断完善成熟这项技术.

References

[1] Hinton GE, Salakhutdinov RR. Reducing the dimensionality of data with neural networks. Science, 2006, 313(5786): 504–507. [doi: 10.1126/science.1127647]

[2] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. In: Proc. of the 3rd Int'l Conf. on Learning Representations. San Diego, 2015. 1–14.

[3] Taigman Y, Yang M, Ranzato M, Wolf L. DeepFace: Closing the gap to human-level performance in face verification. In: Proc. of the 2014 IEEE Conf. on Computer Vision and Pattern Recognition. Columbus: IEEE, 2014. 1701–1708. [doi: 10.1109/CVPR.2014.220]

[4] Han D, Im D, Park G, Kim Y, Song S, Lee J, Yoo HJ. A DNN training processor for robust object detection with real-world environmental adaptation. In: Proc. of the 4th IEEE Int'l Conf. on Artificial Intelligence Circuits and Systems (AICAS). Incheon: IEEE, 2022. 501–501. [doi: 10.1109/AICAS54282.2022.9869954]

[5] Paigwar A, Erkent Ö, Sierra-Gonzalez D, Laugier C. GndNet: Fast ground plane estimation and point cloud segmentation for autonomous

vehicles. In: Proc. of the 2020 IEEE/RSJ Int'l Conf. on Intelligent Robots and Systems (IROS). Las Vegas: IEEE, 2020. 2150–2156. [doi: 10.1109/IROS45743.2020.9340979]

[6] Zhang X, Zhang Y, Zhong M, Ding D, Cao Y, Zhang Y, Zhang M, Yang M. Enhancing state-of-the-art classifiers with API semantics to detect evolved Android malware. In: Proc. of the 2020 ACM SIGSAC Conf. on Computer and Communications Security. 2020. ACM, 2020. 757–770. [doi: 10.1145/3372297.3417291]

[7] Chiu YC, Chen HIH, Zhang TH, Zhang SY, Gorthi A, Wang LJ, Huang YF, Chen YD. Predicting drug response of tumors from integrated genomic profiles by deep neural networks. BMC Medical Genomics, 2019, 12(S1): 18. [doi: 10.1186/s12920-018-0460-9]

[8] Szegedy C, Zaremba W, Sutskever I, Bruna J, Erhan D, Goodfellow I, Fergus R. Intriguing properties of neural networks. arXiv:1312.6199, 2013.

[9] Liu AS, Liu XL, Fan JX, Ma YQ, Zhang AL, Xie HY, Tao DC. Perceptual-sensitive GAN for generating adversarial patches. In: Proc. of the 33rd AAAI Conf. on Artificial Intelligence. Honolulu: AAAI Press, 2019. 1028–1035. [doi: 10.1609/aaai.v33i01.33011028]

[10] Finlayson SG, Bowers JD, Ito J, Zittrain JL, Beam AL, Kohane IS. Adversarial attacks on medical machine learning. Science, 2019, 363(6433): 1287–1289. [doi: 10.1126/science.aaw4399]

[11] Tramèr F, Kurakin A, Papernot N, Goodfellow I, Boneh D, McDaniel P. Ensemble adversarial training: Attacks and defenses. In: Proc. of the 6th Int'l Conf. on Learning Representations. Vancouver: OpenReview.net, 2018. 1–18.

[12] Zhang HY, Yu YD, Jiao JT, Xing EP, Ghaoui LE, Jordan MI. Theoretically principled trade-off between robustness and accuracy. In: Proc. of the 36th Int'l Conf. on Machine Learning. Long Beach: PMLR, 2019. 7472–7482.

[13] Chen JH, Cheng Y, Gan Z, Gu QQ, Liu JJ. Efficient robust training via backward smoothing. In: Proc. of the 36th AAAI Conf. on Artificial Intelligence. AAAI Press, 2020. 6222–6230. [doi: 10.1609/aaai.v36i6.20571]

[14] Rade R, Moosavi-Dezfooli SM. Helper-based adversarial training: Reducing excessive margin to achieve a better accuracy vs. robustness trade-off. In: Proc. of the 2021 ICML Workshop on a Blessing in Disguise: The Prospects and Perils of Adversarial Machine Learning. 2021.

[15] Cheng Z, Zhu F, Zhang XY, Liu CL. Adversarial training with distribution normalization and margin balance. Pattern Recognition, 2023, 136: 109182. [doi: 10.1016/j.patcog.2022.109182]

[16] Dong JH, Moosavi-Dezfooli SM, Lai JH, Xie XH. The enemy of my enemy is my friend: Exploring inverse adversaries for improving adversarial training. In: Proc. of the 2023 IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR). Vancouver: IEEE, 2023. 24678–24687. [doi: 10.1109/CVPR52729.2023.02364]

[17] Tsipras D, Santurkar S, Engstrom L, Turner A, Madry A. Robustness may be at odds with accuracy. In: Proc. of the 7th Int'l Conf. on Learning Representations. New Orleans: OpenReview.net. 1–23.

[18] Yang R, Chen XQ, Cao TJ. APE-GAN++: An improved APE-GAN to eliminate adversarial perturbations. IAENG Int'l Journal of Computer Science, 2021, 48(3): 827.

[19] Zhou DW, Wang NN, Peng CL, Gao XB, Wang XY, Yu J, Liu TL. Removing adversarial noise in class activation feature space. In: Proc. of the 2021 IEEE/CVF Int'l Conf. on Computer Vision (ICCV). Montreal: IEEE, 2021. 7858–7867. [doi: 10.1109/ICCV48922.2021.00778]

[20] Ho CH, Vasconcelos N. DISCO: Adversarial defense with local implicit functions. In: Proc. of the 36th Annual Conf. on Neural Information Processing Systems. New Orleans: NeurIPS, 2022. 1–20.

[21] Nie WL, Guo B, Huang YJ, Xiao CW, Vahdat A, Anandkumar A. Diffusion models for adversarial purification. In: Proc. of the 39th Int'l Conf. on Machine Learning. Baltimore: PMLR, 2022. 16805–16827.

[22] Croce F, Hein M. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In: Proc. of the 37th Int'l Conf. on Machine Learning. JMLR.org, 2020. 2206–2216.

[23] Qian Z, Zhang SF, Huang KZ, Wang QF, Zhang R, Yi XP. Improving model robustness with latent distribution locally and globally. arXiv:2107.04401, 2021.

[24] Stutz D, Hein M, Schiele B. Confidence-calibrated adversarial training: Generalizing to unseen attacks. In: Proc. of the 37th Int'l Conf. on Machine Learning. JMLR.org, 2019. 849.

[25] DeVries T, Taylor GW. Improved regularization of convolutional neural networks with cutout. arXiv:1708.04552, 2017.

[26] Zhang H, Cissé M, Dauphin Y, Lopez-Paz D. Mixup: Beyond empirical risk minimization. In: Proc. of the 6th Int'l Conf. on Learning Representations. Vancouver: OpenReview.net, 2018. 1–13.

[27] Zhong Z, Zheng L, Kang GL, Li SZ, Yang Y. Random erasing data augmentation. In: Proc. of the 34th AAAI Conf. on Artificial Intelligence. New York: AAAI Press, 2020. 13001–13008. [doi: 10.1609/aaai.v34i07.7000]

[28] de Jorge P, Bibi A, Volpi R, Sanyal A, Torr PH, Rogez G, Dokania PK. Make some noise: Reliable and efficient single-step adversarial

training. In: Proc. of the 36th Annual Conf. on Neural Information Processing Systems. New Orleans: NeurIPS, 2022. 1–13.

[29] Yun S, Han D, Chun S, Oh SJ, Yoo Y, Choe J. CutMix: Regularization strategy to train strong classifiers with localizable features. In: Proc. of the 2019 IEEE/CVF Int'l Conf. on Computer Vision (ICCV). Seoul: IEEE, 2019. 6022–6031. [doi: 10.1109/ICCV.2019.00612]

[30] Cubuk ED, Zoph B, Shlens J, Le QV. Randaugment: Practical automated data augmentation with a reduced search space. In: Proc. of the 2020 IEEE/CVF Conf. on Computer Vision and Pattern Recognition Workshops. Seattle: IEEE, 2020. 702–703. [doi: 10.1109/CVPRW50498.2020.00359]

[31] Ho D, Liang E, Chen X, Stoica I, Abbeel P. Population based augmentation: Efficient learning of augmentation policy schedules. In: Proc. of the 36th Int'l Conf. on Machine Learning. Long Beach: PMLR, 2019. 2731–2741.

[32] Li BY, Wu F, Lim SN, Belongie S, Weinberger KQ. On feature normalization and data augmentation. In: Proc. of the 2021 IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR). Nashville: IEEE, 2021. 12378–12387. [doi: 10.1109/CVPR46437.2021.01220]

[33] Radford A, Metz L, Chintala S. Unsupervised representation learning with deep convolutional generative adversarial networks. In: Proc. of the 4th Int'l Conf. on Learning Representations. San Juan: OpenReview.net, 2016. 1–16.

[34] Choi Y, Choi M, Kim M, Ha JW, Kim S, Choo J. StarGAN: Unified generative adversarial networks for multi-domain image-to-image translation. In: Proc. of the 2018 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 8789–8797. [doi: 10.1109/CVPR.2018.00916]

[35] Zhang H, Goodfellow I, Metaxas D, Odena A. Self-attention generative adversarial networks. In: Proc. of the 36th Int'l Conf. on Machine Learning. Long Beach: PMLR, 2019. 7354–7363.

[36] Brock A, Donahue J, Simonyan K. Large scale GAN training for high fidelity natural image synthesis. In: Proc. of the 7th Int'l Conf. on Learning Representations. New Orleans: OpenReview.net, 2019. 1–35.

[37] Karras T, Laine S, Aila T. A style-based generator architecture for generative adversarial networks. In: Proc. of the 2019 IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR). Long Beach: IEEE, 2018. 4396–4405. [doi: 10.1109/CVPR.2019.00453]

[38] Gulrajani I, Ahmed F, Arjovsky M, Dumoulin V, Courville A. Improved training of Wasserstein GANs. In: Proc. of the 31st Int'l Conf. on Neural Information Processing Systems. Long Beach: Curran Associates Inc., 2017. 5769–5779.

[39] Xie CH, Tan MX, Gong BQ, Yuille A, Le QV. Smooth adversarial training. arXiv:2006.14536, 2020.

[40] Maini P, Wong E, Kolter Z. Adversarial robustness against the union of multiple perturbation models. In: Proc. of the 37th Int'l Conf. on Machine Learning. PMLR, 2020. 6640–6650.

[41] Chen HR, Dong YP, Wang ZY, Yang X, Duan CQ, Su H, Zhu J. Robust classification via a single diffusion model. In: Proc. of the 41st Int'l Conf. on Machine Learning. Vienna: PMLR, 2024. 6643–6665.

[42] Sui CH, Wang A, Wang HP, Liu H, Gong QT, Yao J, Hong DF. ISDAT: An image-semantic dual adversarial training framework for robust image classification. Pattern Recognition, 2025, 158: 110968. [doi: 10.1016/j.patcog.2024.110968]

[43] Jin GQ, Shen SW, Zhang DM, Dai F, Zhang YD. APE-GAN: Adversarial perturbation elimination with GAN. In: Proc. of the 2019 IEEE Int'l Conf. on Acoustics, Speech and Signal Processing (ICASSP). Brighton: IEEE, 2017. 3842–3846. [doi: 10.1109/ICASSP.2019.8683044]

[44] Jolicoeur-Martineau A. The relativistic discriminator: A key element missing from standard GAN. In: Proc. of the 7th Int'l Conf. on Learning Representations. New Orleans: OpenReview.net, 2019. 1–26.

[45] Dobriban E, Hassani H, Hong D, Robey A. Provable tradeoffs in adversarially robust classification. IEEE Trans. on Information Theory, 2023, 69(12): 7793–7822. [doi: 10.1109/TIT.2022.3205449]

[46] Mehrabi M, Javanmard A, Rossi RA, Rao A, Mai T. Fundamental tradeoffs in distributionally adversarial training. In: Proc. of the 38th Int'l Conf. on Machine Learning. PMLR, 2021. 7544–7554.

[47] Liu H. Rethinking adversarial training with a simple baseline. arXiv:2306.07613, 2023.

[48] 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. 1–23.

[49] 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. 1–11.

[50] Cui JQ, Liu S, Wang LW, Jia JY. Learnable boundary guided adversarial training. In: Proc. of the 2021 IEEE/CVF Int'l Conf. on Computer Vision. Montreal: IEEE, 2021. 15701–15710. [doi: 10.1109/ICCV48922.2021.01543]

[51] Gong HH, Dong MJ, Ma SQ, Camtepe S, Nepal S, Xu C. Parameter-saving adversarial training: Reinforcing multi-perturbation robustness via hypernetworks. arXiv:2309.16207, 2023.

[52] Li L, Spratling MW. Data augmentation alone can improve adversarial training. In: Proc. of the 11th Int'l Conf. on Learning

Representations. Kigali: OpenReview.net, 2023. 1–23.
[53] LeCun Y, Bottou L, Bengio Y, Haffner P. Gradient-based learning applied to document recognition. Proc. of the IEEE, 1998, 86(11): 2278–2324.
[54] Netzer Y, Wang T, Coates A, Bissacco A, Wu B, Ng AY. Reading digits in natural images with unsupervised feature learning. In: Proc. of the 25th Annual Conf. on Deep Learning and Unsupervised Feature Learning. Granada: NeurIPS, 2011.
[55] Everingham M, van Gool L, Williams CKI, Winn J, Zisserman A. The PASCAL visual object classes (VOC) challenge. Int'l Journal of Computer Vision, 2010, 88(2): 303–338. [doi: 10.1007/s11263-009-0275-4]
[56] He KM, Zhang XY, Ren SQ, Sun J. Deep residual learning for image recognition. In: Proc. of the 2016 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR). Las Vegas: IEEE, 2016. 770–778. [doi: 10.1109/CVPR.2016.90]
[57] Szegedy C, Liu W, Jia YQ, Sermanet P, Reed S, Anguelov D, Erhan D, Vanhoucke V, Rabinovich A. Going deeper with convolutions. In: Proc. of the 2015 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR). Boston: IEEE, 2015. 1–9. [doi: 10.1109/CVPR.2015.7298594]
[58] Ren SQ, He KM, Girshick R, Sun J. Faster R-CNN: Towards real-time object detection with region proposal networks. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2017, 39(6): 1137–1149. [doi: 10.1109/TPAMI.2016.2577031]
[59] Wong E, Rice L, Kolter JZ. Fast is better than free: Revisiting adversarial training. In: Proc. of the 8th Int'l Conf. on Learning Representations. Addis Ababa: OpenReview.net, 2020. 1–17.
[60] Gowal S, Uesato J, Qin CL, Huang PS, Mann TA, Kohli P. An alternative surrogate loss for PGD-based adversarial testing. arXiv:1910.09338, 2019.
[61] Moosavi-Dezfooli S, Fawzi A, Frossard P. DeepFool: A simple and accurate method to fool deep neural networks. In: Proc. of the 2016 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR). Las Vegas: IEEE, 2016. 2574–2582. [doi: 10.1109/CVPR.2016.282]
[62] Carlini N, Wagner D. Towards evaluating the robustness of neural networks. In: Proc. of the 2017 IEEE Symp. on Security and Privacy (SP). San Jose: IEEE, 2017. 39–57. [doi: 10.1109/SP.2017.49]
[63] Zhang HC, Wang JY. Towards adversarially robust object detection. In: Proc. of the 2019 IEEE/CVF Int'l Conf. on Computer Vision. San Jose: IEEE, 2019. 421–430. [doi: 10.1109/ICCV.2019.00051]

作者简介

杨波, 博士, 教授, CCF 专业会员, 主要研究领域为深度学习, 软件测试, 软件故障定位, 软件需求分析.
熊倩, 硕士生, 主要研究领域为人工智能, 攻击与防御, 安全性与可靠性优化.
徐珞, 博士, 研究员级高级工程师, 博士生导师, 主要研究领域为人工智能, 智能体系总体, 软件工程.