软件学报 ISSN 1000-9825, CODEN RUXUEW
2025,36(5):2114−2129 [doi: 10.13328/j.cnki.jos.007188] [CSTR: 32375.14.jos.007188]

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

面向卷积神经网络泛化性和健壮性权衡的标签筛选方法*

王益民, 龙显忠, 李 云, 熊 健

(南京邮电大学 计算机学院、软件学院、网络空间安全学院, 江苏 南京 210023)

通信作者: 龙显忠, E-mail: lxz@njupt.edu.cn

摘 要: 虽然卷积神经网络凭借优异的泛化性能被广泛应用在图像识别领域中, 但被噪声污染的对抗样本可以轻松欺骗训练完全的网络模型, 带来安全性的隐患. 现有的许多防御方法虽然提高了模型的健壮性, 但大多数不可避免地牺牲了模型的泛化性. 为了缓解这一问题, 提出了标签筛选权重参数正则化方法, 在模型训练过程中利用样本的标签信息权衡模型的泛化性和健壮性. 先前的许多健壮模型训练方法存在下面两个问题: 1) 大多通过增加训练集样本的数量或复杂度来提高模型的健壮性, 这不仅弱化了干净样本在模型训练过程中的主导作用, 也使得训练任务的工作量大大提高; 2) 样本的标签信息除了被用于与模型预测结果对比来控制模型参数的更新方向以外, 在模型训练中几乎不被另作使用, 这无疑忽视了隐藏于样本标签中的更多信息. 所提方法通过样本的正确标签和对抗样本的分类标签筛选出模型在分类该样本时起决定性作用的权重参数, 对这些参数进行正则优化, 达到模型泛化性和健壮性权衡的效果. 在 MNIST、CIFAR-10 和 CIFAR-100 数据集上的实验和分析表明, 提出的方法能够取得很好的训练效果.

关键词: 卷积神经网络; 对抗学习; 标签信息; 正则化

中图法分类号: TP18

中文引用格式: 王益民, 龙显忠, 李云, 熊健. 面向卷积神经网络泛化性和健壮性权衡的标签筛选方法. 软件学报, 2025, 36(5): 2114–2129. http://www.jos.org.cn/1000-9825/7188.htm

英文引用格式: Wang YM, Long XZ, Li Y, Xiong J. Label Screening Method for Generalization and Robustness Trade-off in Convolutional Neural Network. Ruan Jian Xue Bao/Journal of Software, 2025, 36(5): 2114–2129 (in Chinese). http://www.jos.org.cn/1000-9825/7188.htm

Label Screening Method for Generalization and Robustness Trade-off in Convolutional Neural Network

WANG Yi-Min, LONG Xian-Zhong, LI Yun, XIONG Jian

(School of Computer Science, Nanjing University of Posts and Telecommunications, Nanjing 210023, China)

Abstract: Although convolutional neural networks (CNNs) are widely used in image recognition due to their excellent generalization performance, adversarial samples contaminated by noise can easily deceive fully trained network models, posing security risks. Many existing defense methods improve the robustness of models, but most inevitably sacrifice model generalization. To alleviate this issue, a label-filtered weight parameter regularization method is proposed to balance the generalization and robustness of models using the label information of samples during model training. Many previous robust model training methods suffer from two main issues: 1) The robustness of models is mainly enhanced by increasing the quantity or complexity of training set samples, which not only diminishes the dominant role of clean samples in model training but also significantly increases the workload of training tasks. 2) The label information of samples is used only to compare with model predictions to control the direction of model parameter updates, neglecting the additional information hidden in sample labels. The proposed method selects weight parameters that play a decisive role in classifying samples by filtering the correct labels of samples and the classification labels of adversarial samples and optimizes these parameters regularly to

* 基金项目: 国家自然科学基金 (62371254, 61906098)
收稿时间: 2023-11-07; 修改时间: 2023-12-24, 2024-02-18; 采用时间: 2024-03-15; jos 在线出版时间: 2024-06-14
CNKI 网络首发时间: 2024-06-17

achieve a balance between model generalization and robustness. Experiments and analysis on the MNIST, CIFAR-10, and CIFAR-100 datasets demonstrate that the proposed method achieves good training results.

Key words: convolutional neural network (CNN); adversarial learning; label information; regularization

卷积神经网络 (convolutional neural network, CNN)[1]自提出起, 便在图像识别和分类领域展现出优异的效果, 深层和复杂的网络结构[2−6]也被提出以应对更加困难的图像处理任务. 然而, CNN 的健壮性已被证明具有一定的缺陷[7,8], 即训练好的模型极易受到对抗样本的攻击. 具体而言, 在图像上添加一个轻微的扰动噪声生成对抗样本, 该样本不会被人眼识别错误, 但可以欺骗泛化性能优秀的模型做出错误的判断. 现有的一些对抗攻击算法可以根据不同的需求来生成对抗样本[9], 如图 1 所示, 可以分别通过快速梯度符号法 (fast gradient sign method, FGSM)[10]、投影梯度下降法 (projected gradient descent, PGD)[11]、DeepFool[12]和 C&W[13]生成和干净样本相对应的对抗样本, 并使模型给出错误的预测结果. 其中, FGSM 能够迅速地生成噪声, 而 PGD、DeepFool 和 C&W 则可以生成更加细微的噪声. 对抗样本的存在给神经网络的理论研究和现实应用[14,15]带来了挑战, 设计防御策略保证模型的健壮性已经成为 CNN 训练阶段必须考虑的因素之一.

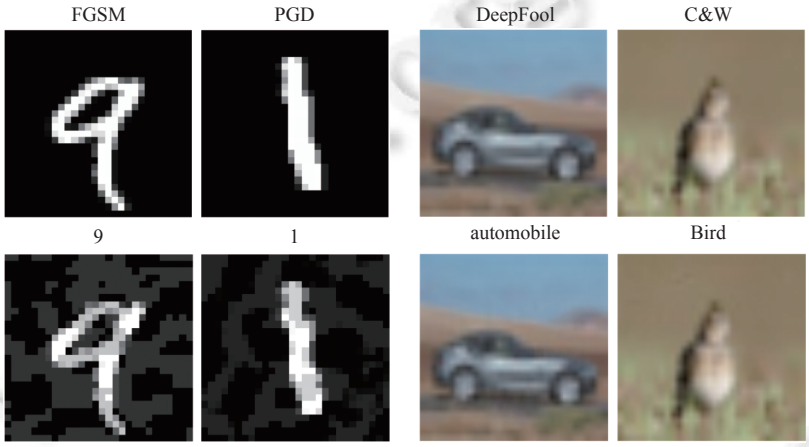


图 1 MNIST 和 CIFAR-10 数据集干净样本和不同攻击下的对抗样本的预测结果

现有的对抗防御算法大多分为正则化和数据增强两种, 从本质上来看, 两者都是为了防止模型对已知数据过度拟合和对未知数据缺乏泛化能力. 正则化技术通过在目标函数中添加新的正则项惩罚输出对于输入的变化程度, 如梯度正则化[16]、雅可比正则化[17,18]和曲率正则化[19]都在一定程度上使网络输出不受对抗扰动的影响. 数据增强[20,21]则尝试在将图像输入到网络模型前应用图像转换技术改变数据集的复杂度, 从而使模型获得健壮性, 对抗训练 (adversarial training, AT)[10,11]使用精心设计和挑选的对抗样本进行训练[22−24]可以获得更加健壮的模型, 也是目前最有效的数据增强防御策略.

然而高泛化性模型和高健壮性模型的训练目标是不同的, 所以 CNN 在获得对抗防御能力的同时难以保证其对干净样本的预测准确率[25], 因此设计一个兼顾泛化性能的防御策略成为对抗学习领域的一个研究热点. Zhang 等人[26]从理论层面上进行分析并设计了一种新的防御方法 TRADES (tradeoff-inspired adversarial defense via surrogate-loss minimization), 该方法利用正则化方式对模型泛化性和健壮性进行权衡. Co 等人[27]提出的雅可比集成法将雅可比正则化与模型集成技术结合起来, 在提高模型对扰动噪声的健壮性的同时也在一定程度上保证了模型的泛化性. 在对抗训练方面, 额外添加对抗样本进行训练大大提高了训练任务的时间开销, 而仅使用对抗样本训练会忽视干净样本的重要性, 导致模型泛化性下降. Grabinski 等人[28]实证分析了各种对抗训练模型, 探讨了不同的对抗训练策略对模型泛化性能的一系列影响. Zhang 等人[29]重新审视了各类对抗样本的噪声强度, 进而提出了

友好对抗训练方法; Sharma 等人[30]则利用健壮性分类器[31]筛选对抗训练输入的样本, 合理地平衡了训练模型的泛化性和稳健性.

我们发现样本的标签信息在深度学习和对抗学习任务中的作用过于单一. 对于有监督的深度学习任务, 标签信息最主要用于计算损失函数来控制模型参数更新的方向. 在对抗学习任务中, 标签信息仅用于生成噪声并判断对抗样本是否能够误导模型判断. 虽然现有的一些方法尝试从标签角度提高模型的泛化性或健壮性, 如 MixUp[32]尝试将数据集中任意两个图像按比例混合生成新的图像, 同时将它们的标签按相同比例混合成新图像的软标签来训练更具有泛化性的模型, 知识蒸馏[33]则先训练教师网络得到软标签, 再使用软标签来训练学生网络进而获得具有健壮性的最终模型. 然而, 这些方法仅重视优化网络模型的单一性质 (泛化性或健壮性), 很少有研究从标签角度对模型的泛化性和健壮性进行权衡. 那么, 能否将对抗样本的错误标签用于改进模型的健壮性?能否在训练健壮深度网络模型时利用干净样本的正确标签保证模型的泛化性?

为了解决这一问题, 本文从正则化防御的方式出发, 提出标签筛选权重参数正则化方法, 通过干净样本和其对抗样本的标签信息分别筛选出模型中决定该样本分类结果的参数, 并将它们作为正则化项进行优化, 前者用来保证模型的泛化性, 后者则用来提高模型的健壮性. 本文在 MNIST、CIFAR-10 和 CIFAR-100 这 3 个经典数据集上做了充分的实验和分析, 验证了该方法的有效性. 本文的创新点主要包括以下 3 个方面.

(1) 提出了面向卷积神经网络泛化性和健壮性权衡的标签筛选方法, 从理论出发解释 CNN 模型预测过程并分析本文设计的正则项的合理性, 证明了标签信息除了能够控制模型权重参数更新的方向外, 还能用于筛选出模型中对预测样本更具重要性的权重参数并对它们针对性地更新, 增强了标签信息在模型训练中的作用.

(2) 利用标签信息统一正则化和数据增强, 不仅使用干净样本的正确分类标签, 还考虑了对抗样本的错误分类标签, 通过两种标签的联合作用, 弥补了卷积神经网络训练时泛化性和健壮性无法权衡的问题, 有效地平衡了两者此消彼长的关系.

(3) 受机器学习中没有免费午餐 (no free lunch, NFL) 定理的启发, 本文对所提方法涉及的参数进行了一系列的消融实验, 观察并分析了在参数变化时模型泛化性和健壮性的表现, 使得该方法可以通过简单地调参来应对不同的学习任务.

本文第 1 节详细介绍深度学习和对抗学习的概念及相关工作. 第 2 节通过数理推导分析标签筛选的权重参数对模型预测结果的影响并给出本文所提方法的具体思想. 第 3 节通过实验验证方法的有效性. 第 4 节通过消融实验分析所提方法取得成功的内在原因. 第 5 节得到相应的结论及展望.

1 相关概念及工作

卷积神经网络在多层感知机的基础上引入了卷积操作[1]. 如图 2 所示, CNN 利用卷积核的方式, 将原先感知机中每一层神经元之间的全连接改成局部连接, 使输出成为能够表征上一层局部感受野的特征图. CNN 在每个卷积层中使用尺寸相同但参数初始化不同的卷积核提取输入中不同的特征, 而不同的卷积层则采用不同大小的卷积核控制特征图所包含的信息量. CNN 通过多层卷积后的特征图能够更好地表示图像的局部抽象信息, 从而在图像识别领域表现优异. 同时, 每一层特征图中神经元的取值共享并依赖于卷积核的参数, 网络仅在最后使用全连接层输出与标签同维度的向量, 从而解决了神经网络参数量极多的缺点, 大大降低了模型复杂度和训练的时间开销, 使得 CNN 能够处理更加复杂的图像分类和识别任务.

在图 2 中, CIFAR-10 数据集中一张 32×32 的输入图像被拉伸成 1 024 维的向量, 经过卷积和全连接的操作后, 以 10 维向量的形式输出, 对应于 CIFAR-10 的 10 个分类预测概率, 其中最大值所在的索引便是网络模型对该图像的预测结果 (为方便解释及之后的证明, 本文省略了 CNN 中池化层的相关概念). 因此, CNN 从本质上可以被理解为一个将 d 维输入向量 $\mathbf{x}$ 转变为 k 维输出向量的映射函数 $F(\mathbf{x};\theta)$, 其中 θ 为网络模型的参数, 该映射函数可以详细表示为 $F(\mathbf{x};\theta)=[f_1(\mathbf{x}),f_2(\mathbf{x}),\ldots,f_k(\mathbf{x})]$, 函数中的 $f_n(\mathbf{x})$ 可被认为是模型预测输入图像为第 n 个类的概率值, 为满足 $\sum_{n=1}^{k}f_n(\mathbf{x})=1$ 的要求, $F(\mathbf{x};\theta)$ 中每一项在最终被输出前都会经过 Softmax 层进行归一化, 即:

$$f_n(\mathbf{x}) = \mathrm{e}^{f'_n(\mathbf{x})} \Big/ \sum_{i=1}^{k} \mathrm{e}^{f'_i(\mathbf{x})} \tag{1}$$

其中, $f'_i(\mathbf{x})$ 为归一化前的输出.

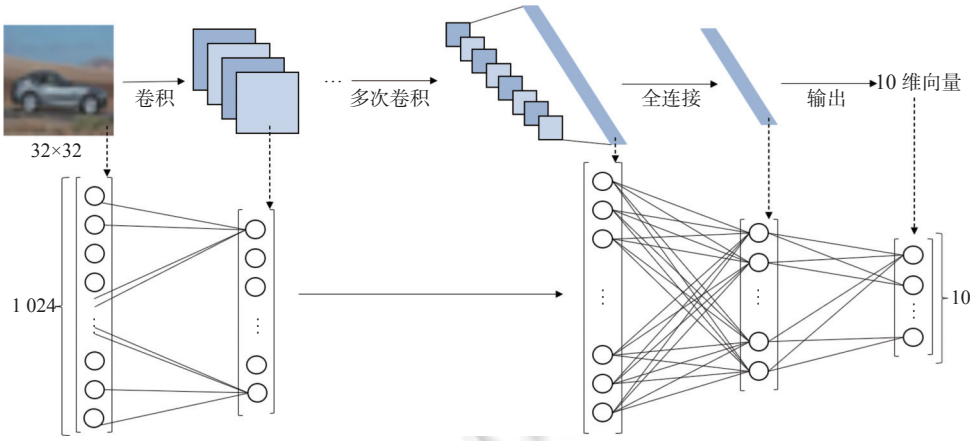


图 2 深度卷积神经网络

之后由损失函数 $\mathcal{L}(F(\mathbf{x};\theta);\mathbf{y})$ 计算该输出和正确标签 $\mathbf{y}$ 的差异值, 用于反向传导更新模型的参数. 目前最常用的损失函数为交叉熵损失, 具体形式如下:

$$\mathcal{L}_{\mathrm{CE}} = -\sum_{n=1}^{k} \mathbf{y}_n \log f_n(\mathbf{x}) \tag{2}$$

假定 $\hat{n}$ 为某一样本的正确类别所代表的索引且 $\mathrm{argmax}F(\mathbf{x};\theta) = \hat{n}$, 则模型预测该样本正确, 由 $\mathcal{L}_{\mathrm{CE}}$ 标准训练的模型能以很低的错误率轻松完成这一任务. 因此, 生成对抗样本的过程便是找到与输入同维度的噪声向量 $\mathbf{r}$, 使该模型无法正确分类对抗样本 $\mathbf{x}'$ ($\mathbf{x}' = \mathbf{x} + \mathbf{r}$), 即满足:

$$\min_{\mathbf{r}} \|\mathbf{r}\|_2 \quad \text{s. t. } \mathrm{argmax}F(\mathbf{x}+\mathbf{r}) \neq \mathrm{argmax}F(\mathbf{x}) \tag{3}$$

其中, $\|\mathbf{r}\|_2$ 为 $\mathbf{r}$ 的二阶范数, 即对抗样本和干净样本的欧氏距离.

1.1 对抗样本生成算法

Goodfellow 等人[10]提出的 FGSM 算法通过计算损失函数的反向梯度来生成噪声, 使得对抗样本的交叉熵损失增大, 从而达到欺骗模型的目的, 具体形式为:

$$r = -\epsilon \cdot sign(\nabla loss_{(F;\mathbf{y})}(\mathbf{x})) \tag{4}$$

其中, ϵ 决定添加的噪声强度, $sign(\cdot)$ 为符号函数, 对一个向量内负值取– 1, 正值取 1. 该算法仅仅利用损失函数即可完成对抗样本的生成, 无需多余的计算步骤, 因此可以快速完成攻击任务. 然而 FGSM 使用 ϵ 值一次性改变噪声中每一维的强度, 在优化过程中并没有对 $\mathbf{r}$ 的大小进行限制, 使得噪声过于明显, 即忽视了公式 (3) 的约束条件. 为解决这一问题, Madry 等人[11]采用多步迭代的 FGSM 算法, 并在每次迭代后将扰动 $\mathbf{r}$ 裁剪至固定范围内, 设计出了 PGD 攻击算法, 即:

$$r_{i+1} = clip_{\epsilon}(r_i + \alpha \cdot sign(\nabla loss_{(F;\mathbf{y})}(\mathbf{x}))) \tag{5}$$

相比于 FGSM 攻击, PGD 将 α 作为每一步攻击的强度, 用 ϵ 限制最终噪声的强度, 将 $\mathbf{r}$ 的大小严格控制在 ϵ 范围的球域内, 因此在相同的 ϵ 值下, PGD 相较于 FGSM 牺牲较多的时间生成了更微小且更具攻击性的噪声.

DeepFool[12]是一种基于优化的攻击, 该算法近似得到样本点到决策面的最短距离向量, 采用多步迭代的方式让样本在添加噪声后移动到决策面的另一侧, 同样生成了有效的对抗样本, 步数是衡量 DeepFool 攻击强度的标准, 每一步后的新样本 $\mathbf{x}_i$ 满足:

$$\mathbf{x}_i = \mathbf{x}_{i-1} - \mathbf{r}_{i-1} = \mathbf{x}_{i-1} - \frac{f_{\hat{n}}(\mathbf{x}_{i-1})}{\|\nabla f_{\hat{n}}(\mathbf{x}_{i-1})\|_2^2}\nabla f_{\hat{n}}(\mathbf{x}_{i-1}) \tag{6}$$

C&W[13]同样是基于优化角度的攻击, 该算法将公式 (3) 转化为如下的优化表达式:

$$\min D(\mathbf{x};\mathbf{x}+\mathbf{r}) + \lambda_{cw}\cdot(\max_{i\neq\hat{n}}(f_i(\mathbf{x}+\mathbf{r}) - f_{\hat{n}}(\mathbf{x}+\mathbf{r}))^+ \quad \text{s.t. } \mathbf{x}+\mathbf{r}\in[0,1]^n \tag{7}$$

其中, D 为计算干净样本与对抗样本差异性的函数, 如欧氏距离, $(\cdot)^+$为 $\max(\cdot;0)$ 的简写. 从该表达式可知 C&W 攻击通过参数 λ_{cw} 权衡攻击强度和噪声大小两个方面, λ_{cw} 的值越高意味着攻击越强.

上述 4 种攻击为常用的经典攻击算法, 也用来验证模型面对不同攻击时是否具有健壮性的标准. 本文将基于这些攻击算法验证所提方法的有效性.

1.2 对抗攻击防御策略

现有的防御策略大多从数据增强和正则化两个角度出发, 对抗训练是数据增强中效果最好的方法, 大多数数据增强方法也是基于对抗训练 (adversarial training, AT)[10,11]的改进. 一般来说, 对抗训练采用对抗样本进行模型训练, 旨在优化如下内部最大化外部最小化的问题:

$$\min_{\theta}\frac{1}{m}\sum_{j=1}^{m}\max_{\|\mathbf{r}_j\|_p\leqslant\epsilon}\mathcal{L}(F(\mathbf{x}_j+\mathbf{r}_j;\theta);\mathbf{y}_j) \tag{8}$$

该优化式内部求解满足 p 范数在 ϵ 域内的噪声, 使得生成的对抗样本的损失最大, 外部对数据集内所有 m 个对抗样本的损失求和并最小化, 从而达到训练健壮模型的目的. 对抗样本使数据集规模扩大了一倍, 如果将干净样本和对抗样本一起训练会产生更多的时间开销, 这在数据集规模庞大的模型训练任务中是不可取的. 因此大多数对抗训练策略仅使用对抗样本来训练健壮模型, 其有效性可以通过邻域风险最小化原则 (vicinal risk minimization, VRM)[34]来解释, 该原则要求训练数据为原样本邻近分布的新样本, 能在一定程度上缓解模型对干净数据的过拟合现象, 同时提高模型的健壮性. 然而, 仅使用对抗样本进行训练更侧重于使模型达到对抗稳健性, 不可避免地牺牲了预测精度.

正则化防御策略尝试改进模型训练的损失函数, 在基础的交叉熵损失上添加相关的正则项达到使模型健壮的目的. 雅可比正则化 (Jacobian regularization, JR)[17,18]将对抗样本的输出 $F(\mathbf{x}+\mathbf{r};\theta)$ 在样本点 $\mathbf{x}$ 处进行一阶泰勒展开, 得到如下展开式:

$$F(\mathbf{x}+\mathbf{r};\theta) = F(\mathbf{x};\theta) + \frac{\partial F(\mathbf{x};\theta)}{\partial\mathbf{x}}\mathbf{r} + O(\mathbf{r}^2) \tag{9}$$

由于 $\mathbf{r}$ 的二阶余项小到可以忽略不计, 因此对抗样本输出和干净样本输出的差距仅在第二项, 其系数 $\frac{\partial F(\mathbf{x};\theta)}{\partial\mathbf{x}}$ 为模型输出关于输入的雅可比矩阵, 是一个 k 行 d 列的矩阵, 只需优化该系数项即可缩小两个输出之间的差距, 故 JR 设计如下损失函数:

$$\mathcal{L}_{\text{JR}} = \mathcal{L}_{\text{CE}} + \frac{\lambda_{\text{JR}}}{2}\sum_{\{\mathbf{e}\}}\left[\frac{\partial\mathbf{e}F(\mathbf{x};\theta)}{\partial\mathbf{x}}\right]^2 \tag{10}$$

其中, λ_{JR} 控制雅可比正则项的优化程度, $\mathbf{e}$ 为 k 维空间的标准正交基, 这使得 JR 的时间开销随着分类任务的类别数 k 的增加而增大, 同时输出-输入雅可比矩阵的正则项过度优化了模型参数, 无法保证模型的准确性. 大多数基于雅可比正则化改进的方法, 如 Hoffman 等人[35]利用随机投影的概念将 JR 的训练时间缩减为原先的 1/k, Le 等人[36]将最优传输理论与 JR 结合实现强健壮性, 都忽视了雅可比正则化使模型准确性下降这一问题.

另一种正则化方法 TRADES[26]则侧重于权衡模型准确性与健壮性, 它设计的损失函数形式为:

$$\mathcal{L}_{\text{TRADES}} = \mathcal{L}_{\text{CE}} + \lambda_{\text{TRADES}}\max_{\mathbf{r}<\epsilon}KL(F(\mathbf{x};\theta)\|F(\mathbf{x}+\mathbf{r};\theta)) \tag{11}$$

其中, 正则项通过最大化干净样本和对抗样本的 KL 散度确定最具攻击性的噪声, 优化该项旨在提高模型的健壮性. TRADES 通过超参数 $\lambda_{\text{TRADES}}>0$ 控制该式中两项被优化的程度, 从而能够权衡模型准确性与健壮性, λ_{TRADES} 的值越高意味着模型越健壮.

本文所提方法同样将对模型准确性和健壮性进行权衡, 上述防御策略如对抗训练, 雅可比正则化和 TRADES

将被用于与所提方法进行比较.

2 标签筛选权重参数分析

在该部分中我们首先对卷积神经网络模型预测过程进行解释, 并以详细的公式推导展示这一过程, 之后对模型的泛化性和健壮性进一步分析, 在此基础上通过理论证明标签筛选权重参数方法的合理性.

在 CNN 的卷积操作中, 卷积核在图像上滑动并对覆盖区域的像素值加权求和, 其本质上可以看作是对输入向量的一次矩阵变换, 以图 2 中 32×32 的输入图像和 2×2 的卷积核为例, 卷积核将进行 900 次滑动, 每次滑动时卷积参数 c_1–c_4 与覆盖图像部分的像素值加权求和, 最后输出 30×30 的特征图, 整个过程可视为 1 024 维输入向量 $\mathbf{x}$ 经过 900×1024 的稀疏矩阵 $\mathbf{C}$ 变换为一个 900 维的向量 $\mathbf{z}$, 该操作用数学形式表达如下:

$$\mathbf{z}=\mathbf{C}\mathbf{x}^{\mathrm{T}}=\begin{bmatrix} c_1 & c_2 & 0 & \cdots & c_3 & c_4 & 0 & \cdots & 0 & 0 & 0 \\ 0 & c_1 & c_2 & \cdots & 0 & c_3 & c_4 & \cdots & 0 & 0 & 0 \\ & \ddots & & & & \ddots & & & & \ddots & \\ 0 & 0 & 0 & \cdots & c_1 & c_2 & 0 & \cdots & c_3 & c_4 & 0 \\ 0 & 0 & 0 & \cdots & 0 & c_1 & c_2 & \cdots & 0 & c_3 & c_4 \end{bmatrix}\begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \\ \vdots \\ \mathbf{x}_{d-1} \\ \mathbf{x}_d \end{bmatrix} \tag{12}$$

其中, 稀疏矩阵 $\mathbf{C}$ 由 2×2 卷积核参数 c_1–c_4 生成 (池化层的操作同样可被视为矩阵变换), 因此卷积操作仍为线性变换. 为使得输出关于输入非线性, CNN 在每层卷积后引入激活函数 $\sigma(\cdot)$ 对结果进行非线性变换, 本文所有模型训练过程中皆采用分段线性激活函数 *ReLU*, 即 $\sigma(\cdot)=ReLU(\cdot)=\max(\cdot;0)$. 若定义判断函数 $\mathbf{1}(\cdot)$, 当括号内条件成立则为 1, 反之为 0, 激活过程 $\sigma(\mathbf{z})$ 可表示为一个对角矩阵 $\mathbf{D}$ 变换, 其具体表达形式如下:

$$\sigma(z)=diag(\mathbf{1}(\mathbf{z}_1>0),\mathbf{1}(\mathbf{z}_2>0),\ldots,\mathbf{1}(\mathbf{z}_{900}>0))\,\mathbf{z}^{\mathrm{T}}=\mathbf{D}\mathbf{z}^{\mathrm{T}} \tag{13}$$

定义 1. $\boldsymbol{\theta}_s=\mathbf{D}_s\mathbf{C}_s$ 为第 s 层的卷积激活参数矩阵, 同时假设输入 $\mathbf{x}$ 经过模型 l 层卷积, t 层全连接后得到输出 $F(\mathbf{x};\theta)$, $\mathbf{W}=\mathbf{W}_t\ldots\mathbf{W}_1$ 为 t 层全连接矩阵的乘积, 则由公式 (12) 和公式 (13) 可将输出 $F(\mathbf{x};\theta)$ 表示为:

$$F(\mathbf{x};\theta)=Softmax((\mathbf{W}_t\cdots\mathbf{W}_1)(\mathbf{D}_l\mathbf{C}_1)(\mathbf{D}_{l-1}\mathbf{C}_{l-1})\ldots(\mathbf{D}_1\mathbf{C}_1\mathbf{x}))=Softmax(\mathbf{W}\boldsymbol{\theta}_l\boldsymbol{\theta}_{l-1}\ldots\boldsymbol{\theta}_1\mathbf{x}) \tag{14}$$

其中, 每一个矩阵的列数与后一个矩阵的行数一致, 即 $\mathbf{W}\in\mathbb{R}^{k\times h_l},\ldots,\boldsymbol{\theta}_s\in\mathbb{R}^{h_s\times h_{s-1}},\ldots,\boldsymbol{\theta}_1\in\mathbb{R}^{h_1\times d}$. 进一步矩阵相乘可知, CNN 模型的映射函数可被视为将 d 维输入转为 k 维输出的矩阵变换, 即:

$$F(\mathbf{x};\theta)=Softmax(\mathbf{V}^{\mathrm{T}}\mathbf{x}) \tag{15}$$

其中, $\mathbf{V}=[\mathbf{v}_1,\mathbf{v}_2,\ldots,\mathbf{v}_k]^{\mathrm{T}}$ 为模型的权重参数矩阵, 也是输出关于输入的雅可比矩阵 $\partial F(\mathbf{x};\theta)/\partial\mathbf{x}$.

对于分类任务中的第 n 类, 在权重参数矩阵 $\mathbf{V}$ 中, $\mathbf{v}_n\in\mathbb{R}^d$ 为该类别对应的权重参数向量, $\mathbf{v}_n^{\mathrm{T}}\cdot\mathbf{x}$ 则为输入图像经过模型后的该类别的预测值, 因此可将式 (1) 改写为:

$$f_n(\mathbf{x})=\mathrm{e}^{\mathbf{v}_n^{\mathrm{T}}\cdot\mathbf{x}}\Big/\sum_{i=1}^{k}\mathrm{e}^{\mathbf{v}_i^{\mathrm{T}}\cdot\mathbf{x}} \tag{16}$$

由于每个类别的输出值经过 Softmax 层归一化后的分母项一致, 最终的输出 $F(\mathrm{x};\theta)$ 中样本为正确类 $\hat{n}$ 的预测概率值仅由其分子项中的权重参数向量 $\mathbf{v}_{\hat{n}}$ 决定. 因此优化分类对应权重参数来改变干净样本或对抗样本的预测结果是一种在模型泛化性和健壮性之间进行权衡的可行策略.

2.1 干净样本标签筛选权重参数正则化

假设一个具有高泛化性的 CNN 模型, 对于数据集中任意样本 $\mathbf{x}$, 该模型映射函数的结果都满足 $\mathrm{argmax}F(\mathbf{x})=\hat{n}$, 即在样本输出结果 $F(\mathbf{x};\theta)$ 中, 对于任意非正确类 n 都满足 $f_{\hat{n}}(\mathbf{x})>f_n(\mathbf{x})$, 根据公式 (16) 可得:

$$\mathrm{e}^{\mathbf{v}_{\hat{n}}^{\mathrm{T}}\cdot\mathbf{x}}\Big/\sum_{i=1}^{k}\mathrm{e}^{\mathbf{v}_i^{\mathrm{T}}\cdot\mathbf{x}}>\mathrm{e}^{\mathbf{v}_n^{\mathrm{T}}\cdot\mathbf{x}}\Big/\sum_{i=1}^{k}\mathrm{e}^{\mathbf{v}_i^{\mathrm{T}}\cdot\mathbf{x}} \tag{17}$$

由于以自然常数 e 为底的指数函数单调递增且恒为正值, 该不等式两边的分母项相等, 同时约去分母项并对分子项进行简化可得到:

$$\mathbf{v}_{\hat{n}}^{\mathrm{T}} \cdot \mathbf{x} > \mathbf{v}_{n}^{\mathrm{T}} \cdot \mathbf{x} \tag{18}$$

即样本向量和权重参数向量点积值的大小关系, 根据向量点积公式 $\mathbf{a} \cdot \mathbf{b} = \|\mathbf{a}\|_2 \|\mathbf{b}\|_2 \cos\{\mathbf{a}, \mathbf{b}\}$, 其中 $\{\mathbf{a}, \mathbf{b}\}$ 为两向量之间的夹角, 则上式可进一步改写如下:

$$\|\mathbf{v}_{\hat{n}}^{\mathrm{T}}\|_2 \|\mathbf{x}\|_2 \cos\{\mathbf{v}_{\hat{n}}^{\mathrm{T}}, \mathbf{x}\} > \|\mathbf{v}_{n}^{\mathrm{T}}\|_2 \|\mathbf{x}\|_2 \cos\{\mathbf{v}_{n}^{\mathrm{T}}, \mathbf{x}\} \tag{19}$$

约去不等号两边的等值项 $\|\mathbf{x}\|_2$ 可得最终式:

$$\|\mathbf{v}_{\hat{n}}^{\mathrm{T}}\|_2 \cos\{\mathbf{v}_{\hat{n}}^{\mathrm{T}}, \mathbf{x}\} > \|\mathbf{v}_{n}^{\mathrm{T}}\|_2 \cos\{\mathbf{v}_{n}^{\mathrm{T}}, \mathbf{x}\} \tag{20}$$

由公式 (20) 可知当 CNN 模型完成训练之后, 其对某一样本的判断将不受该样本向量的二范数影响, 而由样本向量和正确类别对应权重参数向量的夹角值, 以及该权重参数向量的二范数所决定. 由于数据集样本空间的复杂性使得样本向量的方向难以确定, 同时模型权重参数向量的方向难以人为控制, 因此无法优化两个向量之间夹角余弦值. 我们考虑通过对权重参数向量的二范数取负然后进行优化以强化公式 (20) 中不等关系的成立, 使模型具有更高的泛化性能, 定义原始样本的标签为 $\mathbf{y}_T$, 设计正确标签筛选权重参数正则化 (true label screening weight parameter regularization, TLWR) 的损失函数如下:

$$\mathcal{L}_{\mathrm{TLWR}} = \mathcal{L}_{\mathrm{CE}} - \lambda_T \left[\frac{\partial \mathbf{y}_T F(\mathbf{x}; \theta)}{\partial \mathbf{x}} \right]^2 \tag{21}$$

该损失函数由两项组成, 第 1 项为标准的监督损失函数, 第 2 项为正确标签筛选权重参数的正则损失函数, 由独热编码的正确标签从输入输出的雅可比矩阵中筛选出决定样本正确分类的权重参数, 旨在稳定并增强模型的泛化能力, 正则项系数 λ_T 用于权衡标准项和正则项的优化程度, 由于优化该正则项使权重参数提高会导致梯度爆炸从而影响模型收敛, 因此该系数需要设置较小的值.

2.2 对抗样本标签筛选权重参数正则化

对于一个训练好的模型, 如果添加了噪声 $\mathbf{r}$ 的对抗样本能够使它判断错误, 即样本 $\mathbf{x}$+$\mathbf{r}$ 的模型预测结果都满足 $\mathrm{argmax} F(\mathbf{x}+\mathbf{r}) \neq \hat{n}$, 假设该对抗样本的预测结果类为 $\tilde{n}$, 可知 $f_{\tilde{n}}(\mathbf{x}+\mathbf{r}) > f_{\hat{n}}(\mathbf{x}+\mathbf{r})$, 同样由公式 (16) 可得:

$$\mathrm{e}^{\mathbf{v}_{\tilde{n}}^{\mathrm{T}} \cdot (\mathbf{x}+\mathbf{r})} \Big/ \sum_{i=1}^{k} \mathrm{e}^{\mathbf{v}_{i}^{\mathrm{T}} \cdot (\mathbf{x}+\mathbf{r})} > \mathrm{e}^{\mathbf{v}_{\hat{n}}^{\mathrm{T}} \cdot (\mathbf{x}+\mathbf{r})} \Big/ \sum_{i=1}^{k} \mathrm{e}^{\mathbf{v}_{i}^{\mathrm{T}} \cdot (\mathbf{x}+\mathbf{r})} \tag{22}$$

同第 2.1 节的分析和化简上式可得到:

$$\|\mathbf{v}_{\tilde{n}}^{\mathrm{T}}\|_2 \cos\{\mathbf{v}_{\tilde{n}}^{\mathrm{T}}, \mathbf{x}+\mathbf{r}\} > \|\mathbf{v}_{\hat{n}}^{\mathrm{T}}\|_2 \cos\{\mathbf{v}_{\hat{n}}^{\mathrm{T}}, \mathbf{x}+\mathbf{r}\} \tag{23}$$

可知模型的脆弱性是由对抗样本的错误分类结果对应的权重参数向量所决定, 即该向量的二范数以及它和对抗样本向量之间的夹角共同导致模型的错误判断. 本文根据对抗样本 $\mathbf{x}$+$\mathbf{r}$ 的分类结果进行独热编码, 并定义其为原始样本 $\mathbf{x}$ 的对抗标签 $\mathbf{y}_A$, 正则优化对抗标签筛选权重参数向量的二范数以降低其值可使得公式 (23) 的不等关系难以成立. 本节设计如下对抗标签筛选权重参数正则化 (adversarial label screening weight parameter regularization, ALWR) 的损失函数:

$$\mathcal{L}_{\mathrm{ALWR}} = \mathcal{L}_{\mathrm{CE}} + \lambda_A \left[\frac{\partial \mathbf{y}_A F(\mathbf{x}; \theta)}{\partial \mathbf{x}} \right]^2 \tag{24}$$

该损失函数与 $\mathcal{L}_{\mathrm{TLWR}}$ 的区别在于其用对抗标签筛选出模型中导致样本被误分类的权重参数, 正则项系数 λ_A 用于权衡标准项和正则项的优化程度.

2.3 标签筛选权重参数正则化

由前两个小节的简单推导可以得知, 公式 (21) 优化提高模型的正确标签筛选权重参数向量的二范数以达到提高模型泛化性的效果, 但忽视了模型的健壮性. 公式 (24) 优化降低模型的对抗标签筛选权重参数向量的二范数以提高模型的健壮性, 但忽视了模型的泛化性. 为了在深度学习训练过程中利用标签信息对模型的泛化性和健壮性进行权衡, 保证最终模型能在高健壮性的前提下具有稳定的泛化性, 本节结合 TLWR 和 ALWR 提出标签筛选权重参数正则化 (label screening weight parameter regularization, LWR).

具体而言, 对于一个样本 $\mathbf{x}$, 其正确标签和对抗标签大概率是不同的, 那么由这两种标签分别筛选出的模型权重参数也是不同的, 对两者的权重参数同时进行优化就不会产生冲突. 即使某个样本的两种标签相同, 相应的权重参数优化也会互相抵消. 因此我们结合上述两种正则化的方式, 在损失函数中同时优化这两个正则项, 以增大正确标签筛选权重参数向量二范数, 同时降低对抗标签筛选权重参数向量二范数, 即定义损失函数 $\mathcal{L}_{\text{LWR}}$ 为:

$$\mathcal{L}_{\text{LWR}} = \mathcal{L}_{\text{CE}} - \lambda_T \left[\frac{\partial \mathbf{y}_T F(\mathbf{x};\theta)}{\partial \mathbf{x}}\right]^2 + \lambda_A \left[\frac{\partial \mathbf{y}_A F(\mathbf{x};\theta)}{\partial \mathbf{x}}\right]^2 \tag{25}$$

通过设置系数 λ_T 和 λ_A 来控制两个正则项的优化程度, 即权衡模型最终的泛化性和健壮性. 可以发现, LWR 的损失函数和公式 (10) 所示的雅可比正则化损失函数都是对输出-输入雅可比矩阵的利用, 有一定的相似之处. 区别在于, 雅可比正则化从对抗样本的泰勒展开出发得出雅可比矩阵对模型训练的意义, 进而设计损失函数对整体的雅可比矩阵进行优化, 旨在仅提高模型的健壮性. 而 LWR 方法从模型预测过程出发得到雅可比矩阵中对预测结果起重要影响的部分参数, 突出雅可比矩阵的数学意义, 即向量函数中每个分量相对于其自变量偏导的意义. 因此只针对关键的而非整体的矩阵元素进行优化, 能够对模型的泛化性和健壮性进行权衡. 同时, LWR 利用标签进行筛选雅可比矩阵中的元素不会增加模型训练任务的复杂度, 在训练的时间开销上相比于雅可比正则化也更具优势.

在具体的模型训练策略上, 考虑到模型刚开始训练时没有学习到任何先验知识, 此时在损失函数中添加正则项很难保证监督损失能够收敛到使模型拥有良好的泛化性能的值上, 进而影响最终模型的泛化性能. 同时, 对抗样本需要针对模型参数来生成, 无法准确预测干净样本的模型, 其生成的对抗样本的攻击性较弱, 相应的对抗标签并不一定是干净样本最容易被误判的分类. 因此, 在训练策略上, 我们仍然使用标准的监督损失 $\mathcal{L}_{\text{CE}}$, 在固定 epochs 下对网络进行预训练, 在得到稳定的监督损失之后再使用 $\mathcal{L}_{\text{LWR}}$ 进行训练. 可以认为 LWR 为一种后处理方法, 即在已经具有一定泛化性能的模型参数基础上优化正确标签和对抗标签分别对应的正则损失, 使模型在训练过程中能够维持泛化性并提高其健壮性. 算法 1 给出了该训练策略的具体流程, 其中第 9 行为对抗标签的生成方式, 判断对抗样本预测结果中的每一个元素, 值最大的元素设为 1, 其余元素设为 0, 保证对抗标签为独热编码.

算法 1. 标签筛选权重参数正则化 (LWR) 训练模型.

输入: 初始化网络 $F(\cdot;\theta)$, 训练集 $\mathcal{D} = (\mathbf{x}_s, \mathbf{y}_s)_{s=1}^N$, 学习率 η, 预训练 epochs 数 T_{pre}, 后处理 epochs 数 T_{post}, PGD 攻击强度 ϵ, 每一批次训练样本个数 m, 正则项系数 λ_T, λ_A;

输出: 最终网络 $F(\cdot;\theta)$.

1. **for** epoch = 1, 2, …, T_{pre} **do**
2. **for** 每个批次 $\{(\mathbf{x}_s, \mathbf{y}_s)\}_{s=1}^m$ **do**
3. $\theta \leftarrow \theta - \eta/m \sum_{s=1}^{m} \nabla_\theta \mathcal{L}_{\text{CE}}(F((\mathbf{x}_s);\theta), \mathbf{y}_s)$
4. **end for**
5. **end for**
6. **for** epoch = 1, 2, …, T_{post} **do**
7. **for** 每个批次 $\{(\mathbf{x}_s, \mathbf{y}_s)\}_{s=1}^m$ **do**
8. $\mathbf{x}'_s = clip_\epsilon(\mathbf{r}_s + \alpha \cdot sign(\nabla loss_{(F;\mathbf{y}_s)}(\mathbf{x}_s)))$
9. $\mathbf{y}'_s = \mathbf{1}_{\text{argmax } f(\mathbf{x}'_s;\theta)}(f_{i=1}^k(\mathbf{x}'_s;\theta))$
10. $\theta \leftarrow \theta - \eta/m \sum_{i=1}^{m} \nabla_\theta \mathcal{L}_{\text{LWR}}(F((\mathbf{x}_i);\theta), \mathbf{y}_i)$
11. **end for**
12. **end for**

3 实验分析

3.1 实验设置

本文选用公开的数据集 MNIST[1]、CIFAR-10 和 CIFAR-100 来对所提方法进行实验评估. MNIST 为包含 0–9 的 10 类手写字符数据集, 共有 60 000 张训练图像和 10 000 张测试图像, 每张图像规模为 28×28×1 的单通道灰度图, 因此 MNIST 凭借数据简单的特点而被广泛应用于机器学习算法的验证. CIFAR-10 为包含多个动物和交通工具的 10 类现实物体图像数据集, 共有 50 000 张训练图像和 10 000 张测试图像, 每张图像规模为 32×32×3 的三通道彩图, 因此 CIFAR-10 更接近物理世界且数据复杂, 在深度学习的模型评估中成为最常用的数据集. CIFAR-100 数据集规模大小与 CIFAR-10 类似, 但分类类别包含 100 类现实物体图像, 因此每个类别仅有 500 张训练图像和 100 张测试图像, 因此其分类难度高于 CIFAR-10.

针对这 3 种复杂度不同的数据集, 本文采用 3 种不同结构的网络模型分别对它们进行训练和测试. 对于 MNIST 数据集, 采用 LeNet-5[1]浅层模型, 其结构由 2 个卷积层和 2 个池化层交错连接后, 再添加 2 个全连接层得到最终输出, 尽管网络深度较浅结构简单, 但在识别 MNIST 数据集上能够达到极高的准确率, 同时训练的时间开销短, 能够方便迅速地得到用于评估的模型. 对于 CIFAR-10 数据集, 采用 VGG19[3]深度模型, 采用小卷积核进行特征提取, 在 16 个卷积层和 5 个池化层后再添加 3 个全连接层得到最终输出, 这种更深度的模型结构由于参数较多, 训练速度也相对较慢, 但能够完成更加复杂的数据集分类任务. 对于 CIFAR-100 数据集, 采用 ResNet18[4]残差深度神经网络, 该网络使用残差块连接有效地缓解了深层网络存在的梯度消失和梯度爆炸的问题, 提高了更深层模型面对复杂任务的识别准确率.

LWR 模型的训练策略参考算法 1, 采用 60 个 epochs 的预训练和 40 个 epochs 的后处理. 在参数选取上后处理采用噪声强度 ϵ 为 0.2 的 PGD 攻击获取对抗样本及相应的对抗标签, 每个训练批次含样本数 m 为 128 个. 对于 MNIST 数据集下的 LeNet-5 网络, 选取正则项系数 λ_T 为 $5e^{-4}$, λ_A 为 0.05 训练 LWR 模型. 对于 CIFAR-10 数据集下的 VGG19 网络, 选取正则项系数 λ_T 为 $5e^{-5}$, λ_A 为 $1e^{-2}$ 训练 LWR 模型. 对于 CIFAR-100 数据集下的 ResNet18 网络, 选取正则项系数 λ_T 为 $1e^{-3}$, λ_A 为 $1e^{-1}$ 训练 LWR 模型.

本文所有实验都在一块 Nvidia GTX 3090 GPU 上进行, 皆以 Python 语言及 PyTorch 框架实现, 此外验证健壮性所用的四种对抗攻击算法都由该框架下的对抗攻击库 Torchattacks 提供.

3.2 结果及分析

本节首先对提出的单一正则项 TLWR 和 ALWR 方法分别进行实验, 探讨它们在提高模型泛化性或健壮性上是否具备应有的作用. 对于 TLWR 方法, 我们在 CIFAR-10 数据集上用 VGG19 的网络结构, 采用后处理训练策略, 使用标准监督损失 $\mathcal{L}_{\mathrm{CE}}$ 训练 60 个 epochs 的预训练模型, 再选取不同正则项系数的 $\mathcal{L}_{\mathrm{TLWR}}$ 分别后处理训练 40 个 epochs 后得到多个 TLWR 模型. 表 1 给出了这些模型最终的自然准确度, 其中 Base 是在预训练模型基础上仅用 $\mathcal{L}_{\mathrm{CE}}$ 进行同步训练 40 个 epochs 之后的基准模型, 用于与 TLWR 模型进行对比, 同时用 λ_T 值为后缀区分不同的 TLWR 模型. 对最佳的自然准确度值加粗表示.

表 1 Base 模型与不同正则系数下 TLWR 模型的泛化性对比 (%)

模型	自然准确度
Base	90.51
TLWR-$5e^{-7}$	90.31
TLWR-e^{-7}	**91.33**
TLWR-$5e^{-8}$	90.92
TLWR-e^{-8}	90.76
TLWR-$5e^{-9}$	90.40

从表 1 中可知, TLWR 训练能够获得比 Base 模型更好的自然准确度, 且 λ_T 为 e^{-7} 时所得的模型具有最佳的泛化性能. 此外自然准确度的变化与 λ_T 的取值不呈线性关系, 过大或过小的系数甚至会降低模型的泛化能力.

对于 ALWR 方法, 同样在 CIFAR-10 数据集上以 VGG19 的网络结构, 采用相同的预训练模型和后处理训练策略用 $\mathcal{L}_{\text{ALWR}}$ 训练多个模型, 并使用第 1.1 节中介绍的 4 种攻击来检验这些模型的健壮性表现, 具体结果如表 2 所示. 其中 Base 模型与表 1 中使用的相同, 用 λ_A 值为后缀区分不同的 ALWR 模型, 每种攻击采用括号内的固定强度, 且各列中最高的准确值都加粗表示.

表 2 Base 模型与不同正则系数下 ALWR 模型的健壮性对比 (%)

模型	自然准确度	FGSM (ϵ =0.01)	PGD (ϵ =0.01)	DeepFool (steps=3)	C&W (c=0.1)
Base	**90.51**	40.84	20.99	26.89	20.34
ALWR-e^{-1}	84.15	**63.33**	**61.06**	**38.03**	**74.24**
ALWR-e^{-2}	88.18	53.44	45.15	35.44	43.50
ALWR-$5e^{-3}$	88.50	48.74	38.79	33.92	40.65
ALWR-e^{-3}	89.68	42.23	27.71	23.80	33.48
ALWR-$5e^{-4}$	89.79	41.92	25.17	19.39	30.07

由表 2 可知, ALWR 在一定程度上牺牲了模型的自然准确度, 但显著提高了模型在任一攻击下的健壮性, 且自然准确度的提升效果和正则项系数大小成反比, 健壮性的提升效果和正则项系数大小成正比. Base 模型在 C&W 攻击下具有最弱的表现, 然而正则系数为 0.1 的 ALWR 模型在 C&W 攻击下具有最高的健壮准确度, 因此我们猜想 ALWR 方法能很好地抵御 C&W 这一基于优化的对抗攻击算法. 此外, ALWR-e^{-3} 和 ALWR-$5e^{-4}$ 模型的自然准确度没有大幅降低, 同时它们在各种攻击下的健壮准确度相比于 Base 模型没有大幅提高, 可见 ALWR 的正则项系数过小并不能带来模型泛化性和健壮性的显著改变.

对于 LWR 方法, 按照第 2.3 节的算法 1 和第 3.1 节的实验设置进行模型训练, 同时分别在 3 个数据集上使用其他模型, 包括: 雅可比正则化 (JR), TRADES 和对抗训练 (AT), 以及表现最佳的 TLWR 和 ALWR 模型, 如 CIFAR-10 上选取表 1 中的 TLWR-e^{-7} 和表 2 中的 ALWR-e^{-1} 模型. 比较 LWR 和它们在自然准确度和健壮准确度表现上的差别. 表 3–表 5 分别展示了 3 个数据集上的实验结果, 最优的自然准确度和健壮准确度加粗显示.

表 3 MNIST 数据集上各模型自然准确度及健壮准确度表现 (%)

模型	自然准确度	FGSM		PGD		DeepFool		C&W	
		ϵ = 0.05	ϵ = 0.2	ϵ = 0.05	ϵ = 0.2	steps = 1	steps = 5	λ_{cw} = 0.1	λ_{cw} = 10
Base	99.03	94.27	53.38	92.39	7.93	84.16	57.40	93.92	3.28
JR	97.86	95.09	60.97	94.59	34.47	89.18	64.72	97.45	19.16
TRADES	98.45	96.84	52.79	95.81	44.37	91.68	61.70	97.57	40.29
AT	97.32	96.72	63.77	96.05	45.56	91.87	55.59	96.44	4.41
TLWR	**99.15**	93.65	51.83	91.17	5.17	85.80	60.38	97.05	7.20
ALWR	97.89	**97.53**	**66.75**	**96.76**	**46.29**	**92.93**	**67.81**	98.89	**58.73**
LWR	99.12	97.07	63.73	96.59	45.18	91.97	64.45	**99.09**	56.86

在泛化性能上, TLWR 模型在 3 个数据集上都有最佳的表现, 而 JR、AT 和 ALWR 模型未考虑健壮性和泛化性之间的权衡, 自然准确度明显低于其他模型, 同时 LWR 相比于 ALWR 具有更高的准确率也证明了添加干净样本标签筛选权重参数正则项能够提高模型的泛化性. 此外, 在 3 个数据集上, LWR 模型总能表现出比 JR、AT 和 TRADES 模型更好的自然准确性, 表明本文所提方法相比于其他健壮训练策略在稳定模型泛化性上具有一定的优势.

在健壮性能上, 本文对于每种攻击都选取了高低两种不同强度的噪声来判断各个模型的健壮准确度. 首先从不同攻击角度上看, 在基于梯度的攻击算法 FGSM 和 PGD 下, 无论攻击的强弱, ALWR 在 3 个数据集中都有最佳的健壮性表现, 仅在 CIFAR-10 的低强度 PGD 攻击下的健壮准确度比 LWR 低 0.13 个百分点. 在基于优化的攻击

算法 DeepFool 和 C&W 下, ALWR 和 LWR 的健壮性表现相似, 在 CIFAR-10 和 CIFAR-100 上 LWR 有更好的健壮准确度, 因此我们认为 LWR 对不同样本标签筛选权重参数的权衡优化能够在更复杂的数据集和网络结构中对此类攻击展现出优势. 从实验结果可知, JR 模型在大多数攻击下都表现出比其他模型更弱的健壮性, 但在强噪声的 DeepFool 攻击下比 TRADES 和 AT 模型表现更好, 符合文献 [35] 得出的 JR 使得模型决策域更广且决策边界更平滑的结论. TRADES 模型在基于梯度的攻击下健壮性表现略差于 AT 模型, 但其自然准确度相比于 Base 模型下降得更少, 这也体现 TRADES 方法能够权衡模型泛化性和健壮性的原理. AT 模型相比于前两个模型有更好的健壮性, 而梯度攻击下健壮性表现更为明显, 这可能是因为使用 PGD 算法生成的对抗样本进行对抗训练的原因. 在本文所提的方法中, TLWR 与 Base 模型在健壮性表现上远不如其他健壮防御策略, 在 CIFAR-100 的实验中都能被高强度 PGD 完全攻击, 但 TLWR 在基于优化的攻击下相比于 Base 模型表现出更好的健壮性. ALWR 模型和 LWR 模型在任意攻击下都具有相似的健壮性表现, ALWR 在 MNIST 数据集上表现得更好, 而 LWR 在 CIFAR-10 和 CIFAR-100 上对基于优化角度的攻击更健壮, 猜测这可能是因为 LWR 相比于 ALWR 添加了干净样本标签筛选权重参数正则项.

表 4 CIFAR-10 数据集上各模型自然准确度及健壮准确度表现 (%)

模型	自然准确度	FGSM		PGD		DeepFool		C&W	
		$\epsilon = 0.002$	$\epsilon = 0.02$	$\epsilon = 0.002$	$\epsilon = 0.02$	steps = 1	steps = 5	$\lambda_{cw} = 0.1$	$\lambda_{cw} = 10$
Base	90.51	77.88	28.42	76.84	14.09	40.92	25.77	20.71	5.20
JR	85.00	80.91	40.50	78.91	32.47	42.90	34.40	46.00	42.27
TRADES	86.90	81.83	41.35	79.45	18.11	38.67	30.94	51.92	35.58
AT	84.53	82.63	41.65	80.77	32.96	54.21	25.16	41.89	7.06
TLWR	**91.33**	78.39	31.81	77.07	18.56	46.60	29.97	39.18	20.46
ALWR	84.15	**83.38**	**43.21**	81.25	**33.71**	**57.35**	36.90	74.24	40.92
LWR	88.36	82.62	42.00	**81.38**	33.62	57.28	**39.22**	**74.94**	**41.59**

表 5 CIFAR-100 数据集上各模型自然准确度及健壮准确度表现 (%)

模型	自然准确度	FGSM		PGD		DeepFool		C&W	
		$\epsilon = 0.002$	$\epsilon = 0.02$	$\epsilon = 0.002$	$\epsilon = 0.02$	steps = 1	steps = 5	$\lambda_{cw} = 0.1$	$\lambda_{cw} = 10$
Base	72.64	44.03	11.34	44.54	0	39.58	25.40	33.10	9.91
JR	62.38	53.73	19.87	53.21	16.49	40.38	26.09	55.13	29.68
TRADES	68.02	54.67	21.78	54.63	18.99	46.60	29.09	49.37	21.43
AT	64.77	59.85	23.17	55.83	15.32	52.34	43.33	58.65	36.44
TLWR	**73.69**	45.59	12.83	44.96	0	46.27	39.92	36.46	10.91
ALWR	64.11	**60.59**	**24.89**	**59.76**	**20.68**	**60.21**	46.72	60.20	38.15
LWR	68.15	57.90	21.96	59.43	18.85	56.04	**49.49**	**61.25**	**39.02**

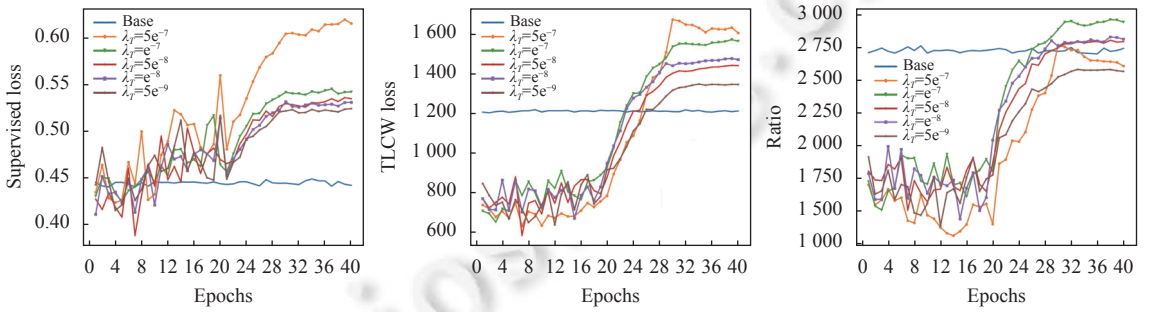
综上对比可以发现, 在本文所提方法中, TLWR 模型仅表现出优异的泛化性, 并不具备可靠的健壮性. 而 ALWR 模型与之相反, 牺牲了泛化性但获得更好的健壮性. 结合了两种正则项的 LWR 模型兼具了 TLWR 的高泛化性能和 ALWR 的高健壮性能, 即相比于其他健壮训练策略, LWR 总能够在保证模型泛化性的前提下提高模型的健壮性, 而两者之间的权衡可以通过调整 LWR 中两个正则项前的系数来实现.

4 消融分析

为了进一步了解不同标签筛选的权重参数以及不同正则项系数对模型最终性能的具体影响, 本节进行了如下的消融实验分析.

首先, 针对表 1 中 TLWR 方法训练所得的相关模型, 我们记录了各个模型在后处理训练的 40 个 epochs 中监督损失值和正则损失值, 并在图 3(a) 和 (b) 中分别可视化展示了它们的变化趋势. Base 模型的监督损失和正则损失在训练过程中没有呈现大的波动, 而 5 个 TLWR 模型的监督损失与正则损失具有相似的上升规律, 并在某一个

epoch 后超过 Base 模型的相应损失, 且最终趋于平坦, 证明由公式 (21) 训练的模型增大了正确标签筛选的网络权重参数的二范数值, 且正则项的添加导致模型的监督损失也在上升. 在深度学习中, 虽然更高的监督损失意味着更低的自然准确度, 但在 TLWR 方法中, 优化正则损失, 即增大干净样本标签筛选的权重参数向量的二范数能够削弱这一影响. 同时, 监督损失的变化关于正则项系数是敏感的, 调高系数很容易造成监督损失过高, 进而导致模型无法收敛的问题, 虽然极小的系数能够起到训练效果, 但通过大量的调参实验来选择合适的系数会让训练任务变得繁重. 因此我们进一步计算同一个 epoch 下正则损失和监督损失的比值, 如图 3(c) 所示, 实验表明在 TLWR 训练过程的后期, 正则损失比监督损失增长的速度更快, 最后两者的比值趋于稳定, 结合表 1 可以推测, 该比值越高得到的模型泛化性越好, 而较低的比值可能导致比基准模型更差的泛化性. 因此, 深度学习模型训练过程中的监督损失和干净样本标签筛选权重参数二范数共同影响模型的泛化能力, 控制两者损失之间的比值是 TLWR 方法维持和提高模型泛化性能的关键.



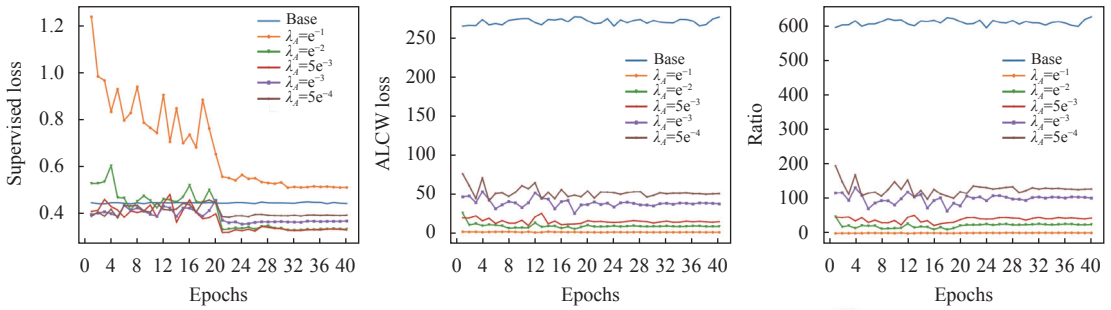
(a) 监督损失变化 (b) TLWR 正则损失变化 (c) TLWR 正则损失与监督损失比值变化

图 3 TLWR 模型训练过程中监督损失和正则损失的变化

同样地, 我们记录表 2 中 ALWR 各个模型在后处理训练的 40 个 epochs 中监督损失值和正则损失值, 将它们的变化情况可视化于图 4(a) 和 (b) 中. 实验表明 ALWR 模型的自然准确度与最终的监督损失值密切相关, 过大的监督损失导致自然准确度大幅下降, 保持比 Base 模型更低的监督损失更能维持模型的泛化性能. 图 4(b) 中正则损失明显低于 Base 模型, 证明 $\mathcal{L}_{\mathrm{ALWR}}$ 训练能够显著降低对抗标签筛选权重参数的二范数, 结合表 2 模型在攻击下的表现可知降低这一值能够起到提高模型健壮性的效果, 且更低的二范数值带来更健壮的模型, 符合我们的理论分析. 与 TLWR 的正则损失与监督损失比值变化不同, ALWR 这一比值变化的趋势平缓, 这可能是监督项与正则项同时优化的结果. 因此, 在对抗样本标签筛选权重参数正则化过程中, 正则项系数的选取过大极易导致模型自然准确度的大幅下降, 选取过小则无法训练出健壮的模型. 因此利用 ALWR 方法虽然能训练出健壮的模型, 但仍无法忽视其对模型泛化性能的影响.

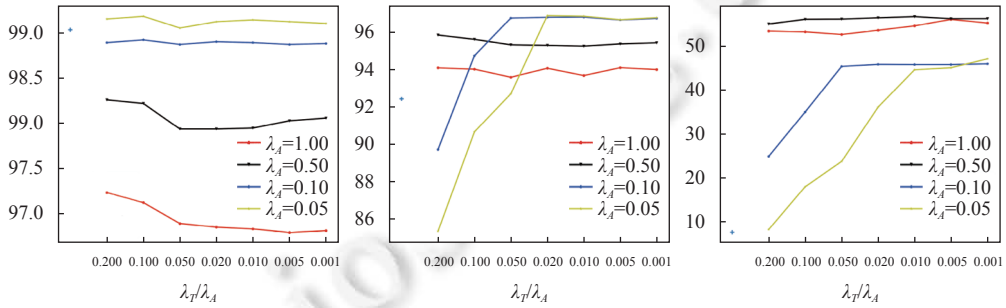
之后, 我们针对 LWR 的两个正则项的系数进行了消融实验, 进一步了解在同时优化干净样本和对抗样本标签筛选权重参数的过程中, λ_T 和 λ_A 的相互作用. 图 5 给出了 LWR 在不同正则参数下训练所得模型在 MNIST 数据集上的自然准确度和在两个不同强弱的 PGD 攻击下的健壮准确度. 其中每条线代表在固定 λ_A 值下仅改变 λ_T 值的准确率变化, 为方便得出两个参数的相互影响, 横坐标代表 λ_T 与 λ_A 的比值, 如图 5(a) 中黑线的第 1 个点为 $\lambda_A=0.5$, $\lambda_T=0.1$ 的模型自然准确度. 同时每张子图最左侧的+号点代表 Base 模型的各准确率表现.

可以发现, λ_A 取值对模型最终的自然准确度有决定性作用, 且值越大所得的模型泛化性越差, λ_T 值的减小同样造成自然准确度的下降, 然而当 λ_T 非常小时, 这一下降趋势将减缓, 即当 λ_T 与 λ_A 比值较大且 λ_A 值较小时模型泛化性最好. 在弱噪声攻击下, 越高的 λ_A 值并不能代表越好的模型健壮性, 而在强噪声攻击下, 高 λ_A 值模型健壮性表现更好. 特别的是, 在高 λ_A 值下, 变化的 λ_T 值几乎没有影响模型健壮性, 而低 λ_A 值下, 模型健壮性随 λ_T 值减小而增强, 最后趋于平缓. 我们认为 LWR 可被视作模型微调的正则化方法, 小的参数选取能够很好地保证模型的泛化性和对基于梯度的攻击具有一定的健壮性, 但面对强的梯度攻击时, 可以适当提高 λ_A 值来获得更加健壮的模型.



(a) 监督损失变化 (b) ALWR 正则损失变化 (c) ALWR 正则损失与监督损失比值变化

图 4 ALWR 模型训练过程中监督损失和正则损失的变化



(a) 自然准确率 (b) PGB(ϵ=0.05)健壮准确率 (c) PGB(ϵ=0.20)健壮准确率

图 5 MNIST 上不同参数的 LWR 模型自然准确度和 PGD 健壮准确度

图 6 给出了 LWR 在不同正则参数下的训练所得模型在 CIFAR-10 数据集上的自然准确度和在两个不同强弱的 C&W 攻击下的健壮准确度. 考虑到 VGG19 结构更加复杂且 LWR 的微调作用, 我们取较小的 0.01 为 λ_A 值, 然后改变 λ_T 的值, 观察各准确率的变化情况, 仍以 λ_T 与 λ_A 的比值为横坐标, 最左侧的+号点为 Base 模型各准确率表现. 发现自然准确度由 λ_A 值决定, 且随着 λ_T 的减小而减小, 与 MNIST 数据集上展现出的结果一致. 在面对基于优化的攻击时的健壮性表现上, 无论噪声的强弱, 准确率都随 λ_T 的减小而表现出明显的提升.

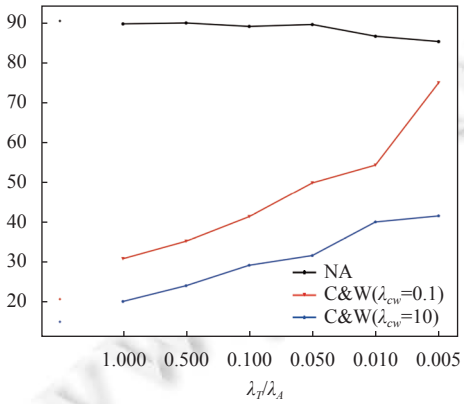


图 6 CIFAR-10 上不同参数的 LWR 模型自然准确度和 C&W 健壮准确度

综上分析, 结合没有免费午餐定理 (no free lunch, NFL), 我们认为获得一个泛化性能优异且对任何类型的攻击都具有很好健壮性的模型是困难的. 针对不同的学习任务采用不同的学习策略是必要的, 对于泛化要求高且不会受对抗攻击的任务, 可以使用较小的 λ_T 值并将 λ_A 设置为 0 即训练 TLWR 模型, 对于模型健壮性要求较高的任务,

可以选取较高的 λ_A 值并将 λ_T 设置为 0, 即训练 ALWR 模型. 对于需要权衡泛化性和健壮性的任务, 我们建议 λ_A 值在 0.1 左右同时控制 λ_T/λ_A 值在 0.01–0.1 之间.

5 总结与展望

训练一个兼具泛化性和健壮性的模型一直是对抗学习中的重要目标. 本文提出的标签筛选权重参数正则化方法, 利用干净样本和对抗样本的标签信息来筛选模型判断过程中起重要因素的权重参数, 并以正则化的方式对这些权重参数进行优化达到提升模型性能的目的. 通过理论分析和实验证明, 该方法相比于其他健壮模型训练方法, 在权衡模型的泛化性和健壮性上具有更好的表现. 同时通过一系列消融实验, 证明了这一方法可以针对待定的学习任务进行具体分析, 从而训练出更适合目标场景的模型. 在未来的工作中, 将更详细地探究所提方法中参数的变化情况, 根据各正则项值的联系动态地调整超参数, 同时尝试把所提方法与其他模型进行结合, 并在更复杂的数据集上进行实验, 给出更合理的, 具有高效性和实用性的模型性能权衡方案.

References:

[1] Lecun Y, Bottou L, Bengio Y, Haffner P. Gradient-based learning applied to document recognition. Proc. of the IEEE, 1998, 86(11): 2278–2324. [doi: 10.1109/5.726791]

[2] Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. Communications of the ACM, 2017, 60(6): 84–90. [doi: 10.1145/3065386]

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

[4] 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. Las Vegas: IEEE, 2016. 770–778. [doi: 10.1109/CVPR.2016.90]

[5] Lu HY, Zhang M, Liu YQ, Ma SP. Convolution neural network feature importance analysis and feature selection enhanced model. Ruan Jian Xue Bao/Journal of Software, 2017, 28(11): 2879–2890 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/5349.htm [doi: 10.13328/j.cnki.jos.005349]

[6] Bai C, Huang L, Chen JN, Pan X, Chen SY. Optimization of deep convolutional neural network for large scale image classification. Ruan Jian Xue Bao/Journal of Software, 2018, 29(4): 1029–1038 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/5404.htm [doi: 10.13328/j.cnki.jos.005404]

[7] Szegedy C, Zaremba W, Sutskever I, Bruna J, Erhan D, Goodfellow IJ, Fergus R. Intriguing properties of neural networks. In: Proc. of the 2nd Int'l Conf. on Learning Representations. Banff, 2014. 1–10.

[8] Nguyen A, Yosinski J, Clune J. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In: Proc. of the 2015 IEEE Conf. on Computer Vision and Pattern Recognition. Boston: IEEE, 2015. 427–436. [doi: 10.1109/CVPR.2015.7298640]

[9] Pan WW, Wang XY, Song ML, Chen C. Survey on generating adversarial examples. Ruan Jian Xue Bao/Journal of Software, 2020, 31(1): 67–81 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/5884.htm [doi: 10.13328/j.cnki.jos.005884]

[10] Goodfellow IJ, Shlens J, Szegedy C. Explaining and harnessing adversarial examples. In: Proc. of the 3rd Int'l Conf. on Learning Representations. San Diego, 2015. 1–11.

[11] 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. 1–28.

[12] Moosavi-Dezfooli SM, 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. Las Vegas: IEEE, 2016. 2574–2582. [doi: 10.1109/CVPR.2016.282]

[13] Carlini N, Wagner D. Towards evaluating the robustness of neural networks. In: Proc. of the 2017 IEEE Symp. on Security and Privacy. San Jose: IEEE, 2017. 39–57. [doi: 10.1109/SP.2017.49]

[14] Wei XX, Guo Y, Yu J. Adversarial sticker: A stealthy attack method in the physical world. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2023, 45(3): 2711–2725. [doi: 10.1109/TPAMI.2022.3176760]

[15] Wu T, Tong L, Vorobeychik Y. Defending against physically realizable attacks on image classification. In: Proc. of the 8th Int'l Conf. on Learning Representation. Addis Ababa: OpenReview.net, 2020. 1–10.

[16] Lyu C, Huang KZ, Liang HN. A unified gradient regularization family for adversarial examples. In: Proc. of the 2015 IEEE Int'l Conf. on

Data Mining. Atlantic City: IEEE, 2015. 301–309. [doi: 10.1109/ICDM.2015.84]

[17] Jakubovitz D, Giryes R. Improving DNN robustness to adversarial attacks using Jacobian regularization. In: Proc. of the 15th European Conf. on Computer Vision. Munich: Springer, 2018. 525–541. [doi: 10.1007/978-3-030-01258-8_32]

[18] Chan A, Tay Y, Ong YS, Fu J. Jacobian adversarially regularized networks for robustness. In: Proc. of the 8th Int'l Conf. on Learning Representation. Addis Ababa: OpenReview.net, 2020. 1–13.

[19] Moosavi-Dezfooli SM, Fawzi A, Uesato J, Frossard P. Robustness via curvature regularization, and vice versa. In: Proc. of the 2019 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 9070–9078. [doi: 10.1109/CVPR.2019.00929]

[20] Guo C, Rana M, Cissé M, van der Maaten L. Countering adversarial images using input transformations. In: Proc. of the 6th Int'l Conf. on Learning Representations. Vancouver: OpenReview.net, 2018. 1–12.

[21] Rebuffi SA, Gowal S, Calian DA, Stimberg F, Wiles O, Mann TA. Data augmentation can improve robustness. In: Proc. of the 35th Conf. on Neural Information Processing Systems. 2021. 29935–29948.

[22] Wang YS, Zou DF, Yi JF, Bailey J, Ma XJ, Gu QQ. Improving adversarial robustness requires revisiting misclassified examples. In: Proc. of the 8th Int'l Conf. on Learning Representations. Addis Ababa: OpenReview.net, 2020. 1–13.

[23] Chen ZM, Xue W, Tian WW, Wu YH, Hua B. Toward deep neural networks robust to adversarial examples, using augmented data importance perception. Journal of Electronic Imaging, 2022, 31(6): 063046. [doi: 10.1117/1.JEI.31.6.063046]

[24] Jin GJ, Yi XP, Wu DY, Mu RH, Huang XW. Randomized adversarial training via Taylor expansion. In: Proc. of the 2023 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Vancouver: IEEE, 2023. 16447–16457. [doi: 10.1109/CVPR52729.2023.01578]

[25] Tsipras D, Santurkar S, Engstrom LG, Turner AM, Madry A. Robustness may be at odds with accuracy. In: Proc. of the 7th Int'l Conf. on Learning Representations. New Orleans, 2019. 1–24.

[26] Zhang HY, Yu YD, Jiao JT, Xing E, El Ghaoui L, Jordan M. Theoretically principled trade-off between robustness and accuracy. In: Proc. of the 36th Int'l Conf. on Machine Learning. Long Beach: PMLR, 2019. 12907–12929.

[27] Co KT, Martinez-Rego D, Hau Z, Lupu EC. Jacobian ensembles improve robustness trade-offs to adversarial attacks. In: Proc. of the 31st Int'l Conf. on Artificial Neural Networks. Bristol: Springer, 2022. 680–691. [doi: 10.1007/978-3-031-15934-3_56]

[28] Grabinski J, Gavrikov P, Keuper J, Keuper M. Robust models are less over-confident. In: Proc. of the 36th Int'l Conf. on Neural Information Processing Systems. New Orleans: Curran Associates Inc., 2022. 2831.

[29] Zhang JF, Xu XL, Han B, Niu G, Cui LZ, Sugiyama M, Kankanhalli M. Attacks which do not kill training make adversarial learning stronger. In: Proc. of the 37th Int'l Conf. on Machine Learning. PMLR, 2020. 11278–11287.

[30] Sharma A, Narayan A. Soft adversarial training can retain natural accuracy. In: Proc. of the 14th Int'l Conf. on Agents and Artificial Intelligence. SciTePress, 2022. 1–7. [doi: 10.5220/0010871000003116]

[31] Gehr T, Mirman M, Drachsler-Cohen D, Tsankov P, Chaudhuri S, Vechev M. AI2: Safety and robustness certification of neural networks with abstract interpretation. In: Proc. of the 2018 IEEE Symp. on Security and Privacy. San Francisco: IEEE, 2018. 3–18. [doi: 10.1109/SP.2018.00058]

[32] Takase T. Feature combination mixup: Novel mixup method using feature combination for neural networks. Neural Computing and Applications, 2023, 35(17): 12763–12774. [doi: 10.1007/s00521-023-08421-3]

[33] Papernot N, McDaniel P, Wu X, Jha S, Swami A. Distillation as a defense to adversarial perturbations against deep neural networks. In: Proc. of the 2016 IEEE Symp. on Security and Privacy. San Jose: IEEE, 2016. 582–597. [doi: 10.1109/SP.2016.41]

[34] Dong NQ, Wang JY, Voiculescu I. Revisiting vicinal risk minimization for partially supervised multi-label classification under data scarcity. In: Proc. of the 2022 IEEE/CVF Conf. on Computer Vision and Pattern Recognition Workshops. New Orleans: IEEE, 2022. 4211–4219. [doi: 10.1109/CVPRW56347.2022.00466]

[35] Hoffman J, Roberts DA, Yaida S. Robust learning with Jacobian regularization. In: Proc. of the 2020 Int'l Conf. on Learning Representations. 2020. 1–21.

[36] Le BM, Tariq S, Woo SS. OTJR: Optimal Transport Meets Optimal Jacobian Regularization for Adversarial Robustness. In: Proc. of the 23rd IEEE Conf. on Computer Vision and Pattern Recognition. 2023. 7551–7562. [doi: 10.48550/arXiv.2303.11793]

附中文参考文献:

[5] 卢泓宇, 张敏, 刘奕群, 马少平. 卷积神经网络特征重要性分析及增强特征选择模型. 软件学报, 2017, 28(11): 2879–2890. http://www.jos.org.cn/1000-9825/5349.htm [doi: 10.13328/j.cnki.jos.005349]

[6] 白琮, 黄玲, 陈佳楠, 潘翔, 陈胜勇. 面向大规模图像分类的深度卷积神经网络优化. 软件学报, 2018, 29(4): 1029–1038. http://www.jos.org.cn/1000-9825/5404.htm [doi: 10.13328/j.cnki.jos.005404]

[9] 潘文雯, 王新宇, 宋明黎, 陈纯. 对抗样本生成技术综述. 软件学报, 2020, 31(1): 67–81. http://www.jos.org.cn/1000-9825/5884.htm [doi: 10.13328/j.cnki.jos.005884]

王益民(1998－), 男, 硕士生, 主要研究领域为对抗机器学习.

李云(1974－), 男, 博士, 教授, 博士生导师, CCF 高级会员, 主要研究领域为对抗机器学习, 数据挖掘, 自然语言处理, 模式识别.

龙显忠(1985－), 男, 博士, 副教授, CCF 专业会员, 主要研究领域为自监督学习, 对抗机器学习.

熊健(1986－), 男, 博士, 副教授, 主要研究领域为视频图像编码, 计算机视觉, 机器学习.