软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software, 2022,33(4):1218−1230 [doi: 10.13328/j.cnki.jos.006477]
©中国科学院软件研究所版权所有.
E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

双标签监督的几何约束对抗训练*

曹刘娟[1], 匡华峰[1], 刘 弘[2], 王 言[3], 张宝昌[4], 黄飞跃[5], 吴永坚[5], 纪荣嵘[1]

[1](厦门大学 信息学院 人工智能系 媒体分析与计算实验室, 福建 厦门 361005)

[2](National Institute of Informatics, Tokyo 101-8430, Japan)

[3](Pinterest, Seattle, WA 98101, USA)

[4](北京航空航天大学 人工智能研究院, 北京 100191)

[5](腾讯优图实验室, 上海 200030)

通信作者: 匡华峰, E-mail: skykuang@stu.xmu.edu.cn

摘 要: 近年来的研究表明, 对抗训练是一种有效的防御对抗样本攻击的方法. 然而, 现有的对抗训练策略在提升模型鲁棒性的同时会造成模型的泛化能力下降. 现阶段主流的对抗训练方法通常都是独立地处理每个训练样本, 而忽略了样本之间的关系, 这使得模型无法充分挖掘样本间的几何关系来学习更鲁棒的模型, 以便更好地防御对抗攻击. 因此, 重点研究如何在对抗训练过程中保持样本间的几何结构稳定性, 达到提升模型鲁棒性的目的. 具体而言, 在对抗训练中, 设计了一种新的几何结构约束方法, 其目的是保持自然样本与对抗样本的特征空间分布一致性. 此外, 提出了一种基于双标签的监督学习方法, 该方法同时采用自然样本和对抗样本的标签对模型进行联合监督训练. 最后, 分析了双标签监督学习方法的特性, 试图从理论上解释对抗样本的工作机理. 多个基准数据集上的实验结果表明: 相比于已有方法, 该方法有效地提升了模型的鲁棒性且保持了较好的泛化精度. 相关代码已经开源: https://github.com/SkyKuang/DGCAT.

关键词: 深度学习; 模型鲁棒性; 对抗训练; 几何约束; 双标签监督

中图法分类号: TP181

中文引用格式: 曹刘娟, 匡华峰, 刘弘, 王言, 张宝昌, 黄飞跃, 吴永坚, 纪荣嵘. 双标签监督的几何约束对抗训练. 软件学报, 2022, 33(4): 1218−1230. http://www.jos.org.cn/1000-9825/6477.htm

英文引用格式: Cao LJ, Kuang HF, Liu H, Wang Y, Zhang BC, Huang FY, Wu YJ, Ji RR. Towards Robust Adversarial Training via Dual-label Supervised and Geometry Constraint. Ruan Jian Xue Bao/Journal of Software, 2022, 33(4): 1218−1230 (in Chinese). http://www.jos.org.cn/1000-9825/6477.htm

Towards Robust Adversarial Training via Dual-label Supervised and Geometry Constraint

CAO Liu-Juan[1], KUANG Hua-Feng[1], LIU Hong[2], WANG Yan[3], ZHANG Bao-Chang[4], HUANG Fei-Yue[5], WU Yong-Jian[5], JI Rong-Rong[1]

[1](Media Analytics and Computing Laboratory, Department of Artificial Intelligence, School of Informatics, Xiamen University, Xiamen 361005, China)

[2](National Institute of Informatics, Tokyo 101-8430, Japan)

[3](Pinterest, Seattle, WA 98101, USA)

[4](Beihang University, Institute of Artificial Intelligence, Beijing 100191, China)

[5](Tencent YouTu Lab, Shanghai 200030, China)

* 基金项目: 国家杰出青年科学基金(62025603); 国家自然科学基金(U1705262, 62072386, 62072387, 62072389, 62002305, 61772443, 61802324 61702136); 广东省基础与应用基础研究基金(2019B1515120049); 中央高校基本科研业务费(20720200077, 20720200090, 20720200091)

本文由“面向开放场景的鲁棒机器学习”专刊特约编辑陈恩红教授、李宇峰副教授、邹权教授推荐.

收稿时间: 2021-05-30; 修改时间: 2021-07-16; 采用时间: 2021-08-27; jos 在线出版时间: 2021-10-26

Abstract: Recent studies have shown that adversarial training is an effective method to defend against adversarial example attacks. However, such robustness comes with a price of a larger generalization gap. To this end, existing endeavors mainly treat each training example independently, which ignores the geometry relationship between inter-samples and does not take the defending capability to the full potential. Different from existing works, this study focuses on improving the robustness of the neural network model by aligning the geometric information of inter-samples to make the feature spatial distribution structure between the natural and adversarial samples is consistent. Furthermore, a dual-label supervised method is proposed to leverage true and wrong labels of adversarial example to jointly supervise the adversarial learning process. The characteristics of the dual-label supervised learning method are analyzed and it is tried to explain the working mechanism of the adversarial example theoretically. The extensive experiments have been conducted on benchmark datasets, which well demonstrates that the proposed approach effectively improves the robustness of the model and still keeps the generalization accuracy. Code is available: https://github.com/SkyKuang/DGCAT.

Key words: deep learning; model robustness; adversarial training; geometry constraint; dual label supervised

近年来, 深度神经网络已经在各种领域任务中取得了巨大的成功, 如目标识别[1]、语音翻译[2]、强化学习[3]等. 然而, 研究表明, 基于传统训练方法的深度神经网络模型容易受到对抗样本的攻击[4−7]. 攻击者通过合成对抗样本来使神经网络模型预测出错, 使得模型无法提供正常的服务. 研究[8−10]表明: 基于深度神经网络模型的对抗缺陷是具有普遍性的, 它与模型的网络结构和任务无关. 这种缺陷严重限制了深度神经网络在一些需要较高安全性的场景中的应用. 因此, 如何通过机器学习方法获得一个良好鲁棒性的深度神经网络模型, 已成为当下亟需解决的重要问题.

为了抵御对抗样本的攻击, 研究者们已经提出了大量的防御策略和算法来提升模型的鲁棒性[11−14]. 然而, 大部分防御方法都无法做到一劳永逸, 每当新的防御方法被提出时, 攻击者可以通过分析相关的防御策略并提出对应的攻击算法, 进而导致新的防御方法被新的攻击算法所攻破. 截至目前, 对抗训练[11,15]防御策略被公认为是一种较完备的防御方法, 它能抵御目前主流的大多数攻击算法. 因此, 大量的研究者通过研究对抗训练来提升神经网络模型的鲁棒性.

当前, 在模型训练过程中, 对抗训练方法往往是相对独立地处理每个训练样本, 此做法忽略了样本间的几何结构关系. 本文认为: 样本间几何结构的稳定性, 是抵御对抗样本攻击的重要因素之一. 换而言之, 对于一个鲁棒性模型, 自然样本间的几何结构关系应该与其对应的对抗样本间的几何结构关系是一致的. 也就是, 合成的对抗样本应具有与自然样本相同的特征空间结构, 这能够保证数据分布的一致性. 因此, 本文的第一个核心贡献是探索并设计了两种新的空间几何度量约束, 即空间距离约束和相对角度约束, 旨在度量样本在特征空间中几何结构的匹配程度. 此外, 为了得到更好的模型泛化能力, 本文提出了一种双标签联合监督学习方法. 该方法充分利用自然样本的正确标签以及对抗样本的错误标签对模型训练过程进行联合监督, 使模型学到对抗样本中的鲁棒性特征和自然样本中的非鲁棒性特征(即泛化特征). 最后, 本文将几何度量约束和双标签监督学习方法融合到一个端到端的训练框架中, 称为双标签几何约束对抗训练, 整个框架如图 1 所示.

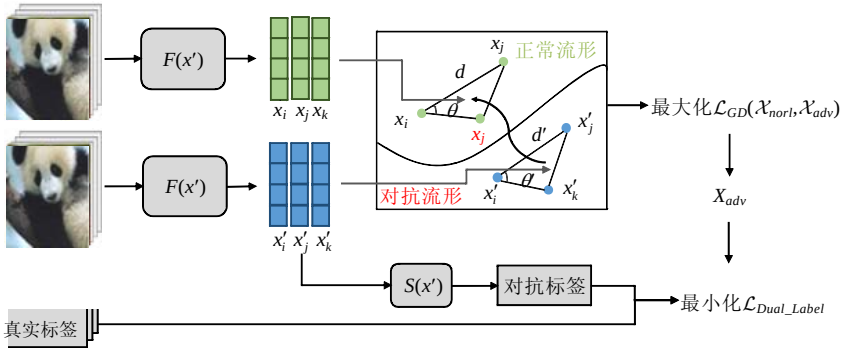


图 1 双标签几何约束对抗性训练流程图, 通过最大化自然样本和对抗样本之间的几何度量来生成对抗样本用于对抗训练; 通过最小化双标签损失来更新模型参数

本文第 1 节介绍传统的对抗训练算法. 第 2 节介绍基于几何约束的对抗训练方法, 重点介绍分析两个几何空间约束度量. 第3节介绍双标签监督学习方法, 并从理论上阐明双标签监督学习对模型特征的影响, 同时也阐述深度特征如何影响模型的鲁棒性. 第 4 节是实验结果与分析, 包含不同方法的性能比较、消融实验对比和实验结果分析. 第 5 节介绍相关工作, 包括对抗攻击、防御模型和对抗样本的可解释性. 第 6 节总结全文, 并指出未来工作的方向.

1 传统对抗训练

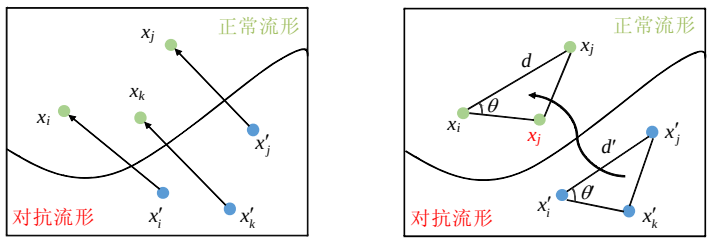
对抗训练由 Goodfellow 等人[16]首先提出, 然后经过 Madry 等人[11]的改善, 即通过使用 PGD[11]攻击算法产生对抗样本进行对抗训练, 从而极大地提高了模型的鲁棒性. 目前, 基于 PGD 的对抗训练成为主流的防御方法. 在对抗训练中, 首先通过已有的攻击算法产生对抗样本, 然后在模型训练过程中将对抗样本和正常样本一起作为训练数据送入模型进行监督训练. 此过程可以被表述为极小极大问题, 对抗训练的优化形式如下:

$$\min_{\theta} \mathbb{E}[\max_{x_{adv}} \mathcal{L}(F_\theta(x_{adv}), y)] \tag{1}$$

其中, F_θ是神经网络模型, 其参数为θ; $\mathcal{L}$是交叉熵损失函数; x_{adv} 为对抗样本. 整个优化目标是一种对抗的形式, 内部最大化进行对抗攻击, 产生对抗样本.即给定一个样本 x, 通过内部最大化, 在一定的范围内找到一个 x_{adv}, 使得训练损失函数最大化; 外部最小化过程则是使模型尽可能地拟合对抗样本, 使神经网络模型学习到对抗样本中的鲁棒性特征, 从而使神经网络模型具备鲁棒性. 传统的对抗训练只考虑单个样本对的信息, 从而忽略了数据分布的整体几何结构信息. 同时, 由于采用 PGD 攻击方式产生对抗样本, 导致训练耗时长, 且需要大量计算资源. 因此, 如何改善对抗训练、在有限的计算资源和时间条件下获得更高效、更鲁棒的神经网络模型, 是目前对抗训练研究中的热点问题.

2 几何约束对抗训练

几何约束对抗训练的核心在于约束样本特征空间的整体几何结构, 以保证数据分布的一致性, 而不是只针对当个样本进行约束训练. 为此, 我们提出了两种空间几何约束度量: 空间距离约束和相对角度约束. 通过在神经网络模型训练过程中增加几何空间约束, 使得模型利用样本的整体结构稳定性来防御对抗攻击, 从而提升模型的鲁棒性. 图 2 给出了传统对抗训练和几何约束对抗训练的对比示意图.



(a) 传统对抗训练　　(b) 几何对抗训练

图 2　传统的对抗训练只匹配单个自然样本和对抗样本之间的特征, 而几何约束对抗训练则是约束整个样本空间中的几何结构, 使得整个数据分布保持一致

为方便描述几何约束对抗训练算法, 相关符号定义如下: F_θ为神经网络模型, 其参数为θ; $\mathcal{L}$是交叉熵损失函数; f 为神经网络的逻辑输出层, 其中, $f_{norl}=F_\theta(x_{norl})$为自然样本的逻辑输出, $f_{adv}=F_\theta(x_{adv})$为对抗样本的逻辑输出; 定义$\mathcal{X}$为实例对的集合, $\mathcal{X}$中的每个实例包括一个自然样本和与之对应的对抗样本, 例如 $\mathcal{X}=\{x^i=(x^i_{norl},x^i_{adv}),x^j=(x^j_{norl},x^j_{adv}),...\}$. $\mathcal{X}_{norl}$ 表示只包含自然样本的集合, $\mathcal{X}_{adv}$ 表示只包含对抗样本的集合.

2.1 基于距离度量的约束

对于数据样本间的距离度量, 我们考虑训练数据中的一对自然样本$\langle x^i,x^j\rangle$, 并定义一个空间中的距离度量函数 D. 本文使用欧氏距离度量, 则这两个样本在特征空间中的距离计算如下:

$$\phi_D(f_{norl}^i, f_{norl}^j)=\frac{1}{\mu}\cdot\| f_{norl}^i - f_{norl}^j\|_2 \tag{2}$$

其中, μ是归一化因子. 为了计算样本对之间的相对距离, 将μ设置为批次训练数据的平均距离, μ的计算方式如下:

$$\mu=\frac{1}{|\mathcal{X}_B^2|}\cdot\sum_{(x^i,x^j)\in\mathcal{X}_B}\| f_{norl}^i - f_{norl}^j\|_2 \tag{3}$$

其中, B 表示一个批次的数据量大小. 同理, 可计算相应的两个对抗样本之间特征空间距离 $\phi_D(f_{adv}^i, f_{adv}^j)$. 我们认为: 对于鲁棒性模型, 两个自然样本间的特征空间距离, 在样本受到攻击后, 所得到的对抗样本之间的特征空间距离应该保持一致, 即 $\phi_D(f_{norl}^i, f_{norl}^j)=\phi_D(f_{adv}^i, f_{adv}^j)$, 从而能保证数据分布的一致性. 因此, 当获得自然样本对和对抗样本对之间的距离后, 便可以构建基于距离的约束函数, 函数定义如下:

$$\mathcal{L}_D=\sum_{(x^i,x^j)\in\mathcal{X}} l_\delta(\phi_D(f_{norl}^i, f_{norl}^j),\phi_D(f_{adv}^i, f_{adv}^j)) \tag{4}$$

其中, l_δ为ℓ_1 平滑函数[17], 它的定义如下:

$$l_\delta(x,y)=\begin{cases}\dfrac{1}{2}\cdot(x-y)^2, \text{ for } |x-y|\leqslant 1\\ |x-y|-\dfrac{1}{2}, \text{ otherwise}\end{cases} \tag{5}$$

在网络训练时, 我们将约束函数$\mathcal{L}_D$ 作为一个正则惩罚项, 从而在神经网络训练时, 能使模型对齐自然样本对和对抗样本对之间的特征距离, 保证样本间的特征距离一致.

2.2 基于角度度量的约束

在高维空间中, 只使用空间距离度量无法精确捕获两个样本之间的几何结构关系. 因此, 我们引入了样本间相对角度信息来进一步确定样本间的几何结构. 对于相对角度度量, 考虑训练数据中的一个三元组$\langle x^i,x^j,x^k\rangle$, 并定义一个角度度量函数$\phi_A$, 函数$\phi_A$可计算特征空间中三元组之间的相对角度信息, 计算方法如下:

$$\phi_A(f_{norl}^i, f_{norl}^j, f_{norl}^k)=\cos\angle ijk=\langle \boldsymbol{e}^{ij},\boldsymbol{e}^{kj}\rangle \tag{6}$$

其中, $\boldsymbol{e}^{ij}=\dfrac{f_{norl}^i - f_{norl}^j}{\| f_{norl}^i - f_{norl}^j\|_2}$, $\boldsymbol{e}^{kj}=\dfrac{f_{norl}^k - f_{norl}^j}{\| f_{norl}^k - f_{norl}^j\|_2}$. 同理, 也可以计算与之对应的对抗样本三元组之间的相对角度信息 $\phi_A(f_{adv}^i, f_{adv}^j, f_{adv}^k)$. 之后, 定义一个角度匹配函数$\mathcal{L}_A$. 通过函数$\mathcal{L}_A$, 可以计算自然样本三元组和相应的对抗样本三元组之间的角度误差, 从而通过构建一个角度约束函数来保证他们之间的角度一致.具体计算公式如下:

$$\mathcal{L}_A=\sum_{(x^i,x^j,x^k)\in\mathcal{X}} l_\delta(\phi_A(f_{norl}^i, f_{norl}^j, f_{norl}^k),\phi_A(f_{adv}^i, f_{adv}^j, f_{adv}^k)) \tag{7}$$

其中, l_δ为ℓ_1 平滑函数. 在网络训练时, 我们同样将角度约束函数$\mathcal{L}_A$ 作为一个正则惩罚项加入到模型训练中, 从而可以在网络训练时对齐自然样本三元组和对抗样本三元组之间的角度信息, 保证特征空间中样本相对角度的一致性, 从而更好地保证自然样本与对抗样本的特征结构一致性.

2.3 几何约束对抗训练

在第 2.1 节和第 2.2 节, 我们构建了基于距离和角度的特征几何约束, 现在定义基于几何约束的对抗训练. 传统的对抗训练都是通过 PGD 攻击方式产生对抗样本, 导致训练速度慢以及需要大量计算资源. 为了解决这一缺陷, 本文提出了一种新的对抗样本产生方式, 称为几何离散对抗样本. 它通过最大化训练样本间的特征

距离和相对角度来产生对抗样本, 即在高维特征空间中, 让对抗样本的特征空间结构与自然样本的特征空间结构越不匹配越好. 这种产生对抗样本的过程, 称为几何离散. 与 PGD 攻击算法不同是, 在产生对抗样本过程中, 几何离散不需要样本标签, 是一种无监督对抗攻击算法. 与传统的对抗训练方法相比, 本文的算法有效地杜绝了标签泄露的问题, 它的计算形式如下:

$$\mathcal{X}'_{adv}=\underset{\mathcal{X}_{adv}\in\mathcal{X}}{\arg\max}\,\mathcal{L}_{GD}(\mathcal{X}_{norl},\mathcal{X}_{adv}) \tag{8}$$

其中, $\mathcal{L}_{GD}=\mathcal{L}_D+\mathcal{L}_A$, 为产生的几何对抗样本. 公式(8)直观地解释为最大化自然样本和对抗样本之间的几何结构信息, 通过破坏样本间的空间结构稳定性来生成对抗样本. 因此, 与传统的没有考虑几何结构的对抗样本相比, 这样产生的对抗样本包含了数据分布的几何结构信息. 之后, 我们利用几何离散产生的对抗样本进行对抗训练, 其数学优化形式如下:

$$\min_{\theta}\mathbb{E}\left[\mathcal{L}\left(\underset{x_{adv}\in\mathcal{X}'_{adv}}{F_\theta(x_{adv})},y\right)\right] \tag{9}$$

基于几何结构约束的对抗训练可以被认为既学习到了样本的语义信息, 同时也学习到了整体数据分布的几何结构信息, 使得神经网络模型可以利用数据的整体几何结构稳定性来提升模型的鲁棒性.

3 双标签监督学习

在传统的神经网络训练中, 都是利用原有的数据标签对神经网络模型进行单一监督训练. 但在对抗攻击场景下, 对抗样本可以误导神经网络产生错误的预测结果, 而且对抗样本和自然样本的微小误差人眼无法区别. 因此我们猜测: 1) 自然样本中包含了对抗样本错误类别的相关特征; 2) 正常训练得到的神经网络模型只使用少量的特征来进行结果预测, 只要其中的一些特征被对抗噪声干扰, 那么神经网络就会输出错误的结果. 基于此设计了一种双标签监督学习方法, 利用对抗样本的真实标签以及被攻击后的错误标签对神经网络模型进行联合监督训练, 其算法形式如下:

$$\mathcal{L}_{DL}=\left[(1-\alpha)\cdot\mathcal{L}\left(\underset{x_{adv}\in X_{adv}}{F_\theta(x_{adv})},y_{true}\right)+\alpha\cdot\mathcal{L}\left(\underset{x_{adv}\in X_{adv}}{F_\theta(x_{adv})},y_{adv}\right)\right] \tag{10}$$

其中, y_{true} 表示输入样本的原始正确标签, y_{adv} 表示对抗样本的错误标签. α是一个超参数用以平衡正确标签和错误标签的置信度, 取值范围$\alpha\in[0,0.5]$. 当我们使用双标签监督时, 需要保证 y_{true} 与 y_{adv} 不同. 对于无目标攻击, 可以给 y_{adv} 分配一个随机标签; 对于目标攻击, 将 y_{adv} 设置为攻击目标标签.

需要注意的是, 我们提出的双标签监督算法与噪声标签和标签平滑有本质的不同. 其中: 噪声标签只是平滑或者替换了原始标签, 原则上还是单个标签; 而标签平滑是在其余所有类别的标签上都给予一个较小的值. 而双标签监督算法是从特征包含的角度出发, 由于对抗样本只是加了微小的扰动, 而且样本的生物信息特征并没有改变, 因此认为原始样本就包含了对抗样本错误类别的特征. 与标签平滑不同的是, 我们只考虑两个最大关联的标签, 而不是所有的标签信息都考虑. 因此, 双标签监督总是涉及到两个标签. 进一步, 我们提供了理论分析, 阐述了双标签监督算法如何影响模型的鲁棒性.

接下来推导双标签监督算法关于神经网络模型鲁棒性的相关性质. 我们定义$f\in\mathcal{R}^{N\times1}$为神经网络的逻辑层输出, 其中, N 为 f 中包含特征的数量. 定义神经网络中最后一层的权重矩阵为 $\boldsymbol{W}\in\mathcal{R}^{N\times K}$, 其中, K 为样本类别数量. 正确标签 y_{true} 在 $\boldsymbol{W}$ 中对应的权值为 $W_t\in\mathcal{R}^{N\times1}$, 则预测正确标签出现的概率为 $p_t=\mathcal{S}(f^{\mathrm{T}}W_t)$, 其中, $\mathcal{S}(P)_i=\dfrac{\exp(p_i)}{\sum_{i=0}^{K}\exp(p_i)}$, $\mathcal{S}$表示 *softmax* 函数. 错误标签y_{adv}在 $\boldsymbol{W}$中对应权值为 $W_a\in\mathcal{R}^{N\times1}$, 则预测出错误标签的概率为: $p_a=\mathcal{S}(f^{\mathrm{T}}W_a)$.

性质 1. 当使用双标签对神经网络进行监督训练时, 可以得到:

$$f^{\mathrm{T}}W_t-f^{\mathrm{T}}W_a=\log\frac{1-\alpha}{\alpha\cdot(1+c)} \tag{11}$$

其中，$c=\sum_{j\neq t,\,j\neq a}^{K-2}\mathrm{e}^{f^{\mathrm{T}}W_j-f^{\mathrm{T}}W_t}$. 从上面公式可以导出以下相关性质.

(1) 当$\alpha\to 0$时，此时可以看作只有单个标签进行监督训练，可以得到$f^{\mathrm{T}}W_t-f^{\mathrm{T}}W_a=\infty$;

(2) 当$\alpha\to 0.5$时，此时两个标签同时对网络进行监督，可以得到$f^{\mathrm{T}}W_t-f^{\mathrm{T}}W_a=0$;

(3) 当α介于 0 和 0.5 之间时，即$\alpha\in(0,0.5)$，此时虽然有两个标签对网络进行监督训练，但正确标签占主导作用，有$f^{\mathrm{T}}W_t-f^{\mathrm{T}}W_a=C$, C是一个与α相关的常数.

证明: 首先，对$\mathcal{L}_{DL}$进行展开:

$$\begin{aligned}\mathcal{L}_{DL}&=-(1-\alpha)\cdot\log(p_t)-\alpha\cdot\log(p_a)X=f^{\mathrm{T}}W_t-f^{\mathrm{T}}W_a\\&=(\alpha-1)\left(f^{\mathrm{T}}W_t-\log\left(\sum_{j=1}^{K}\mathrm{e}^{f^{\mathrm{T}}W_j}\right)\right)-\alpha\cdot\left(f^{\mathrm{T}}W_a-\log\left(\sum_{j=1}^{K}\mathrm{e}^{f^{\mathrm{T}}W_j}\right)\right)\\&=\alpha\cdot(f^{\mathrm{T}}W_t-f^{\mathrm{T}}W_a)-f^{\mathrm{T}}W_t+\log\left(\sum_{j=1}^{K}e^{f^{\mathrm{T}}W_j}\right)\\&=\alpha\cdot(f^{\mathrm{T}}W_t-f^{\mathrm{T}}W_a)+\log\left(\sum_{j=1}^{K}\mathrm{e}^{f^{\mathrm{T}}W_j-f^{\mathrm{T}}W_t}\right)\\&=\alpha\cdot(f^{\mathrm{T}}W_t-f^{\mathrm{T}}W_a)+\log\left(1+\sum_{j\neq t,\,j\neq a}^{K}\mathrm{e}^{f^{\mathrm{T}}W_j-f^{\mathrm{T}}W_t}+\mathrm{e}^{f^{\mathrm{T}}W_a-f^{\mathrm{T}}W_t}\right)\end{aligned}\tag{12}$$

定义$c=\sum_{j\neq t,\,j\neq a}^{K-2}\mathrm{e}^{f^{\mathrm{T}}W_j-f^{\mathrm{T}}W_t}$，其中, $c\in(0,K-2]$, $\alpha\in[0,0.5]$. 然后可以得到:

$$\mathcal{L}_{DL}=\alpha\cdot X+\log(1+c+\mathrm{e}^{-X})\tag{13}$$

对$\mathcal{L}_{DL}$关于X求导:

$$\nabla_X\mathcal{L}_{DL}=\alpha+\frac{-\mathrm{e}^{-X}}{1+c+\mathrm{e}^{-X}}\tag{14}$$

当F_θ是一个理想的鲁棒分类器时，那么有$\mathcal{L}_{DL}=0$. 此时$\nabla_X\mathcal{L}_{DL}=0$, 可以得到$X=\log\frac{1-\alpha}{\alpha\cdot(1+c)}$，即:

$$f^{\mathrm{T}}W_t-f^{\mathrm{T}}W_a=\log\frac{1-\alpha}{\alpha\cdot(1+c)}\tag{15}$$

(1) 当$\alpha\to 0$, 则整个优化过程相当于单个标签进行监督，可计算得$f^{\mathrm{T}}W_t-f^{\mathrm{T}}W_a=\infty$;

(2) 当$\alpha\to 0.5$, 则整个优化过程有两个标签联合监督，且两个标签占的置信度一样，对于任意的$t,a\in K$, 都有$f^{\mathrm{T}}W_t-f^{\mathrm{T}}W_a=0$, 此时$c\to 0$;

(3) 当$\alpha\in(0,0.5)$, 则整个优化过程仍然有两个标签联合监督，但真实标签的置信度占据主导. 此时计算可以得到$f^{\mathrm{T}}W_t-f^{\mathrm{T}}W_a=C$, C是一个常数，它只与α有关.

通过上述性质可知: 当α接近 0 时，$f^{\mathrm{T}}W_t-f^{\mathrm{T}}W_a=\infty$, 这意味着$W_t-W_a=\infty$. 如果交换正确标签和错误标签的位置，有$W_t-W_a=\infty$. 当考虑整个数据集时，假设数据集有$K$个类别，那么对任意的$i,j\in K$且$i\neq j$时，$W_i$和$W_j$需要满足$W_i-W_j=\infty$. 如果$\boldsymbol{W}$能满足以上条件，那么$\boldsymbol{W}$矩阵的每列中将平均只有$N/K$个有效值. 也就是说，在神经网络进行推理时，最后计算预测概率时只用到了N/K个f中的特征. 因此，单标签监督的优点是:当输入是自然样本时，模型根据少量特征就可以给出较高的置信度. 然而其缺陷是: 由于只用了少量特征，一旦这些特征被对抗攻击干扰，神经网络便会输出错误的结果.

当使用双标签进行监督训练时，有$W_t-W_a=C$, 其中, C是一个与α相关的常数. 此时，对任意的$i,j\in K$且$i\neq j$, 都满足$W_i-W_j=C$. 此时，$\boldsymbol{W}$中有更多的有效值在网络推理时被应用. 也就是说，在双标签监督下得到的模型在推理时利用了更多的高层特征. 之前的研究[16]认为，对抗样本是由于神经网络的高维线性所导致的. 结合我们对于双标签监督算法的分析，由于正常训练的神经网络是高维线性的，而且使用了少量的特征进行推理，从而造成了神经网络对于对抗扰动的脆弱性. 此外，有工作将对抗样本看成是一种特殊的特征. Ilyas 等人[18]

认为，神经网络学习的特征可以分为鲁棒性特征和非鲁棒性特征. 根据我们的理论，可以很好地解释这一观点. 当使用单个标签对网络进行监督训练时，神经网络模型主要学到的是 N/K 个非鲁棒性特征,而鲁棒性特征则是在双标签监督下学习到的 N/K 个特征之外的其他特征. 除了正式的分析和实验外，我们进一步提供了直观的解释：扰乱神经网络模型中的特征是对抗攻击的关键所在，因为攻击者试图修改最少数量的特征以改变模型输出，而防御者则不断地增强特征的鲁棒性，并且在模型推理时，尽可能地使用多个鲁棒性特征. 因此，如果神经网络模型使用更多的特征进行推理，那么攻击者想要成功攻击网络就需要花费更多的攻击成本来查询模型信息. 而双标签监督训练策略通过在训练中考虑对抗标签来加强多个特征的学习，从而有效地干扰了攻击者在攻击网络时的特征选择以及增加攻击者的查询代价.

使用双标签监督的另一个好处是：在训练阶段，即使神经网络对训练集已经进行了完全正确的拟合，网络仍然会有一定的损失函数值来更新网络权重，这使得网络能够从正确标签类别中学到更深层次的特征，从而更好地提高模型的泛化能力和鲁棒性. 最后，将基于几何约束的对抗训练和双标签监督学习算法整合到一个端到端的训练框架中，我们称为双标签几何约束对抗训练(dual-label geometry constraint adversarial training, DGCAT). 算法 1 中总结了 DGCAT 的一般训练过程.

算法 1. 双标签几何约束对抗训练算法.

输入：数据集$\mathcal{X}$，批数量 m，外部训练轮次 E，内部迭代轮数 T，步长δ，扰动大小ε，学习率γ，平衡参数α，网络模型 F_θ；

输出：鲁棒网络模型 F_θ.

1. 随机初始化网络 F_θ
2. **for** e=1,…,E **do**
3. **for** random batch x form $\mathcal{X}$ **do**
4. $X_{adv} \leftarrow X_{norl}+\mathcal{N}(0,\boldsymbol{I})$ //$\mathcal{N}(0,\boldsymbol{I})$为标准高斯分布
5. **for** t=1,…,T **do**
6. $X_{adv} \leftarrow \Pi(X_{adv}+\delta \cdot sign(\nabla_{x_{adv}} \mathcal{L}_{GD}(X_{norl}, X_{adv})))$ //Π为投影函数
7. **end for**
8. $\Delta = (1-\alpha)\cdot \sum_{i=1}^{m} \nabla \mathcal{L}(F_\theta(x_{adv}^i), y_{true}^i) + \alpha \cdot \sum_{i=1}^{m} \nabla \mathcal{L}(F_\theta(x_{adv}^i), y_{adv}^i)$
9. $\theta \leftarrow \theta - \frac{\gamma}{m} \cdot \Delta$
10. **end**
11. **end**

4 实验与分析

4.1 实验设置

为了验证本文所提方法的有效性，我们将本文提出的方法与现在主流的基准方法进行了比较，主要的基准方法包括标准对抗训练[11](Madry)、TRADES[15]和 BAT[19]. 使用白盒和黑盒两类攻击方法对我们的方法进行攻击测试. 白盒攻击包括快速梯度攻击[16](FGSM)和迭代投影梯度下降攻击[11](PGD)，黑盒攻击主要使用迁移攻击[20−22]. 为了验证方法的泛化性，我们使用多个基准数据集进行实验验证，包括 CIFAR-10, CIFAR-100[23]以及 SVHN[24]. 我们采用的网络结构为 Wide-ResNet[25]，这与之前的基准方法[11,15,19]保持一致.

实现细节：

- 在训练阶段，初始化学习率为 0.1；整个训练过程中，学习率的变化为[0.1,0.01,0.001]. 对于 CIFAR-10 和 CIFAR-100 数据集，分别在训练的第 80 轮和 120 轮调整学习率，总共训练共 200 轮；而对于 SVHN 数据集，分别在第 40 轮和第 60 轮调整学习率，总共训练 80 轮. 仿照之前基准工作[11]的参数设置，我

们设置扰动大小为 8.0/255. 但在 CIFAR-10 和 CIFAR-100 数据集上, 本文方法在生成对抗样本时反向梯度迭代只需要 1 次, 在 SVHN 数据集上反向梯度迭代只需 3 次, 这使得方法在训练时要快于传统的对抗训练方法. 相比于 Madry 的 7 次和 TRADES 的 10 次, 我们的算法可以大大地节省训练时间和计算资源;

- 在测试阶段, 通过测量模型在不同对抗攻击下的正确精度, 近似计算测试集上的鲁棒性上界来评估模型的鲁棒性. 对于实验中的超参数, 设置双标签监督算法中的α=0.4, 代码基于 PyTorch 框架, 相关代码已经开源: https://github.com/SkyKuang/DGCAT.

4.2 白盒攻击

为了验证模型的鲁棒性, 首先在 CIFAR-10 数据集上对提出的算法以及基准方法进行了鲁棒性评估. CIFAR-10 数据集被认为是对抗训练鲁棒性评估的基准数据集, 它包含 10 个类别, 拥有 5 万张训练图像和 1 万张测试图像. 使用不同的攻击算法在 1 万张测试集上生成对抗样本, 然后测试我们模型在对抗样本上的准确率. 采用 FGSM 和多步 PGD 攻击算法进行攻击, PGD 步长分别为 20, 40 和 100. 实验结果如表 1 所示. 可以看到, 标准训练的模型在白盒攻击下基本全军覆没. 使用标准对抗训练的 Madry 方法极大地提高了模型鲁棒性, 在 20 步 PGD 攻击下取得了 46.4%的准确度. 其他的基准方法 TRADES 和 BAT 进一步提升了模型的鲁棒性, 分别取得了 55.7%和 57.2%的准确度. 而提出的 DGCAT 算法使模型的鲁棒性大幅度提升, 超过了所有的基准方法, 达到了 66.3%的准确度, 领先 Madry 方法 19.9 个百分点. 相比于 TRADES 和 BAT, 也都高出了大约 10 个百分点. 此外, 对于更强的 100 步 PGD 攻击算法, 我们的方法同样取得了 63.7%的准确度, 超过了所有的基准方法.

表 1 在 CIFAR-10 数据集上模型鲁棒性 (%)

方法名称	干净测试样本	FGSM 攻击	PGD 攻击		
			PGD-20	PGD-40	PGD-100
Standard	**95.5**	32.7	0.0	0.0	0.0
Madry	86.7	54.2	46.4	46.1	44.7
TRADES	87.4	64.5	55.7	54.1	53.3
BAT	91.1	70.7	57.5	56.3	55.2
DGCAT	90.4	**77.1**	**66.3**	**64.6**	**63.7**

在第 3 节中我们提出了神经网络模型的脆弱性的原因之一是标准训练的网络模型只用了少量特征进行预测推理, 而鲁棒性模型则使用更多的特征来进行预测推理. 为了验证这一结论, 用 L1 正则约束来使网络权重变得稀疏, 从而强迫网络使用较少的特征进行推理. 实验结果表明, 基于 L1 约束的模型在 20 步 PGD 攻击下只取得了 57.3%的准确度. 相比于不用 L1 约束的模型, 鲁棒性严重下降, 这也侧面证明了我们结论的正确性.

此外, 我们还在 CIFAR-100 数据集上进行了相关实验. CIFAR-100 拥有 100 个类别图像, 其中包含 5 万张训练图像和 1 万张测试图像, 由于种类的变多, 在 CIFAR-100 上取得较好的性能会更具挑战性. 实验结果如表 2 所示. 可以看出: 我们的方法无论在 FGSM 还是 PGD 攻击下都取得了最优的结果, 比基准模型高出了近 10 个百分点. 这说明提出的算法在多种类(种类大于 10)数据集上仍然具有较强的鲁棒性.

表 2 在 CIFAR-100 数据集上模型鲁棒性 (%)

方法名称	干净样本	FGSM 攻击	PGD 攻击		
			PGD-20	PGD-40	PGD-100
Standard	**79.1**	10.1	0.8	0.3	0.1
Madry	59.1	28.5	23.6	23.1	22.2
TRADES	64.3	32.4	26.2	25.8	24.3
BAT	68.2	60.8	26.7	26.2	25.3
DGCAT	73.1	**68.9**	**37.3**	**36.1**	**31.2**

为了验证模型的泛化能力, 除了在 CIFAR 类数据集上进行测试之外, 我们还在 SVHN 数据集上进行了鲁棒性评估. SVHN 是一个拥有 10 类房号标记的数据集, 包含 73 257 张训练图像和 26 032 张测试图像. 在 SVHN 数据集上, 使用 3 次迭代产生训练所需要的对抗样本, 每次的步长为 4.0/255. 整个实验结果见表 3. 可以看出:

相比于基准方法, 我们的方法不仅取得了最高的鲁棒性准确度, 而且在干净样本上同样保持了较高的精度. 这充分证明我们的方法可以泛化到多个数据集, 体现出本文所提算法不仅能提升模型的鲁棒性, 同时还能保证模型较强的泛化能力.

表 3 在 SVHN 数据集上模型鲁棒性 (%)

方法名称	干净样本	FGSM 攻击	PGD 攻击		
			PGD-20	PGD-40	PGD-100
Standard	**97.3**	41.1	0.4	0.1	0.0
Madry	93.7	66.5	47.8	47.1	46.3
TRADES	94.3	68.3	52.3	50.9	48.8
BAT	94.1	69.8	53.9	52.7	50.3
DGCAT	96.8	**95.9**	**83.4**	**80.7**	**76.7**

4.3 黑盒攻击

为了进一步验证模型在黑盒攻击下的鲁棒性, 选取了主流的黑盒攻击方法来测试本文所提方法的鲁棒性. 在黑盒攻击中, 最常见以及应用最广泛的方法为迁移攻击[21,22], 迁移攻击通过一个代理模型产生对抗样本, 然后用这些对抗样本攻击目标模型. 为了测试模型对于迁移攻击的鲁棒性, 我们训练了两个不同类型的代理模型, 一个普通训练的模型和一个使用本文提出的方法对抗训练得到的模型. 所有模型的训练扰动为 ε=8.0/255, 使用的攻击方法为 FGSM 和 PGD, 实验结果见表 4. 从表中可知: 我们的方法在多个基准数据集都取得较高的鲁棒性, 表明模型在黑盒攻击下同样具有较好的防御效果.

表 4 黑盒攻击下不同数据集上的模型鲁棒性 (%)

方法名称	正常代理模型		鲁棒代理模型	
	FGSM	PGD-20	FGSM	PGD-20
CIFAR-10	88.8	89.4	81.8	76.4
CIFAR-100	67.4	70.3	72.1	64.7
SVHN	85.4	87.3	91.7	79.5

4.4 消融实验

为了进一步研究本文提出方法中的几个组件对模型鲁棒性的影响, 我们做了大量消融实验. 为了保证消融实验的公平性, 所有实验都在 CIFAR-10 数据集上进行, 并且所有除了需要对比的参数不同外, 其他超参数全部保持一致.

(1) 几何约束对模型鲁棒性的影响.

通过比较不同对抗训练方法的鲁棒性来验证本文提出的几何约束对抗训练的优势. 首先, 构建了不同训练模型的方式, 分别为: a) 标准训练(Standard); b) 基于随机噪声的对抗训练(Random); c) 基于 FGSM 攻击的对抗训练(FGSM); d) 基于迭代 PGD 攻击的对抗训练(Madry); e) 只是用距离约束的对抗训练(Geometry-D); f) 只是使用角度约束的对抗训练(Geometry-A); g) 同时使用距离和角度约束的对抗训练(Geometry). 整个实验结果见表 5. 从结果可以看出, 使用几何约束训练的模型的鲁棒性超过了所有的对比方法. 同时也发现: 只有同时使用距离约束和角度约束时效果达到最好, 而使用单一的约束效果并不明显. 我们认为: 这是由于在高维空间中, 单一的距离度量或者角度度量无法精确地计算出两个样本之间的几何关系, 只有当两个样本间的距离和角度都确定时, 才可以精确度量两者的几何关系.

此外, 当结合距离约束和角度约束时, 模型在 FGSM 攻击下的鲁棒性有所下降. 我们认为:这是由于 FGSM 攻击只进行了 1 次梯度计算, 如公式(16), 所以 FGSM 的梯度信息相对简单, 易于被拟合; 而我们的方法为了加快训练速度, 同样使用 1 次迭代. 由于单一的距离约束或者角度约束不够强时, 可能导致产生的对抗样本与 FGSM 攻击产生的对抗样本相似, 从而导致模型更好地拟合 FGSM 攻击算法产生的对抗样本.从实验结果中也可以看出: 仅仅使用随机噪声进行对抗训练的模型对于 FGSM 攻击的防御就能达到 43.7%, 但对于较强的 PGD 攻击, 基于 FGSM 的训练方法和随机噪声的训练方法的防御效果都很差; 相反, 基于更强约束(同时约束距离和角度)的对抗训练在面对更强的攻击算法时表现得更好.

表 5 不同对抗训练模型在白盒攻击下的鲁棒性 (%)

方法名称	干净样本	FGSM	PGD-20	PGD-40
Standard	95.5	32.7	0.0	0.0
Standard+DL	**95.8**	71.4	10.9	2.8
Random	95.4	43.7	0.2	0.0
Random+DL	95.2	79.8	29.2	11.3
FGSM	89.4	**98.6**	0.4	0.0
FGSM+DL	92.8	95.3	31.8	28.5
Madry	86.7	54.2	46.4	44.7
Madry+DL	85.9	64.7	**50.1**	**48.8**
Geometry-D	90.2	89.1	4.9	1.2
Geometry-D+DL	93.4	89.2	45.0	27.2
Geometry-A	91.4	**90.8**	7.5	1.5
Geometry-A+DL	**94.1**	88.5	44.7	28.8
Geometry	91.5	64.7	37.7	36.6
Geometry+DL	90.4	77.1	**66.3**	**63.7**

注: 后缀“DL”表示该模型使用双标签监督算法训练

为了进一步分析几何离散产生的对抗样本与PGD系列算法产生的对抗样本的区别, 我们测试不同方法产生对抗样本的攻击能力. PGD 使用交叉熵损失函数产生对抗样本, TRADES 使用 KL 散度产生对抗样本, Geometry 使用几何离散产生对抗样本. 实验结果见表 6. 从表中数据可以看出: 几何离散方式产生的对抗样本的攻击能力并没有 PGD 和 TRADES 的强; 同样的, TRADES 基于 KL 散度产生的对抗样本的攻击能力也弱与 PGD 方法. 但从防御结果来看, 基于几何离散和 KL 算法的对抗训练方法的防御能力都强于基于 PGD 的对抗训练. 这说明在对抗训练中, 模型最终的鲁棒性不取决于对抗训练中对抗样本的攻击能力, 而更多地在于使用的数据样本信息.

表 6 不同方法产生的对抗样本的攻击能力 (%)

数据集	FGSM	PGD	TRADES	Geometry
CIFAR10	77.1	66.3	68.4	75.1
CIFAR100	68.9	37.3	48.9	56.6
SVHN	95.9	83.4	62.5	87.9

(2) 双标签监督算法对模型鲁棒性的影响.

传统的神经网络训练都是使用单个标签监督训练, 为了验证本文提出的双标签监督训练算法的有效性, 我们对比了多个训练模型在使用单标签训练和双标签训练时的鲁棒性. 实验结果见表 5. 表格中加 DL 结尾的表示使用双标签监督算法训练得到的模型. 可以看到: 双标签监督算法相对于单标签监督算法无论是在干净样本上的准确度, 还是对抗样本上的鲁棒性都有显著的提升. 这充分证明了我们提出的双标签监督算法的有效性.

(3) 超参数α对模型鲁棒性的影响.

为了确定在双标签监督下, α对模型鲁棒性的影响, 我们对超参数α的不同取值进行对比实验. α表示错误标签的置信度. 设置α从 0.1 到 0.5, 测试不同取值下训练得到的神经网络的鲁棒性, 实验结果见表 7. 实验发现: 在α=0.4 时, 模型取得最好的鲁棒性.

表 7 不同α取值下模型的鲁棒性 (%)

α取值	干净测试样本	FGSM	PGD-20	PGD-40
0.1	90.6	74.8	60.3	56.1
0.2	**90.8**	75.4	58.8	54.4
0.3	90.1	76.5	62.5	59.5
0.4	90.4	**77.1**	**66.3**	**63.7**
0.5	90.3	76.8	64.6	61.9

5 相关工作

5.1 对抗攻击

Sezgedy 等人[4]首次提出对抗样本这一概念, 即: 在原始的自然样本上添加人类肉眼无法感知的噪声, 使扰动后的输入样本造成神经网络预测出错. 在这一概念被提出之后, 一系列的对抗攻击方法相继被提出[11,16], 其中使用最广的为 FGSM[16]攻击算法和 PGD[11]攻击算法. FGSM 方法使用模型梯度信息产生对抗样本, 它将在固定的扰动范围内寻找最具有攻击性的扰动定义成一个优化问题, 其数学形式如下:

$$x_{adv}=x_{norl}+\varepsilon\cdot sign(\nabla_x\mathcal{L}(F_\theta(x_{norl}),y)) \tag{16}$$

其中, (x_{norl},y)为自然样本和对应的标签; x_{adv}为对应的对抗样本; F_θ为神经网络模型, 参数为θ; $\mathcal{L}$是交叉熵损失函数; ε为扰动大小, 且满足$\|x_{adv}-x_{norl}\|_\infty\leqslant\varepsilon$. Madry 等人[11]进一步改善了 FGSM 算法, 提出了 PGD 迭代攻击算法. PGD 通过多步迭代投影的方式产生较强的对抗样本, 它的整个数学形式如下:

$$x_k=\Pi(x_{k-1}+\alpha\cdot sign(\nabla_x\mathcal{L}(F_\theta(x_{k-1}),y))) \tag{17}$$

其中, α为迭代步长, Π为投影函数, x_k为迭代 k 次后产生的对抗样本. 目前, PGD 是使用最广泛的攻击算法.

除了以上这种需要访问模型梯度信息的攻击算法(白盒攻击), 现实情况下, 更多的攻击者使用的是黑盒攻击算法[21,22]. Wu 等人[21]研究表明, 由一个模型产生的对抗样本同样可以高成功率攻击另外的模型. 这使得攻击者可以用对抗迁移的性质对目标模型进行黑盒攻击. 本文对提出的方法进行了白盒攻击和黑盒攻击测试. 实验表明, 我们的方法能够有效地防御这两类算法的攻击.

5.2 对抗防御

随着一系列攻击方法被提出来, 大量的防御策略也相继被研发出来抵御对抗攻击. 比如: Papernot 等人[26]提出通过知识蒸馏的方式来增加模型的鲁棒性; Samangouei 等人[27]提出通过数据流行投影的方式来防御对抗攻击; Guo 等人[28]提出通过对输入样本进行预处理来消除对抗样本的破坏性能力; Raghunathan 等人[29]提出可验证的防御模型, 使得模型能防御在一定扰动内的所有对抗样本; Goodfellow 等人[16]提出了著名的对抗训练防御策略. 有研究[30,31]表明: 大部分防御方法都无法有效地防御基于梯度掩码的自适应对抗攻击, 只有对抗训练防御策略是目前被证实唯一有效的防御手段. 对抗训练旨在将对抗样本加入到模型训练种进行训练, 它可以被定义成一个极小极大的优化问题, 如公式(1)所示: 内部最大化过程可以通过 FGSM[16]或 PGD[11]算法近似求得, 外部通过梯度下降法进行极小值优化.

在对抗训练基础上, Zhang 等人[15]提出了一种平衡模型准确度和鲁棒性的防御算法, 并命名为 TRADES. 它的整个优化过程定义如下:

$$\mathcal{L}(F_\theta(x_{norl}),y)+\lambda\cdot\mathcal{KL}(P(\cdot|x_{norl})\|P(\cdot|x_{adv})) \tag{18}$$

其中, $\mathcal{KL}$为 KL 散度函数, λ是超参数, 用于控制对抗如鲁棒性和标准精度之间的平衡关系. 此外, 一些新颖的对抗训练方法也相继被提出来, Wang 等人[19]提出一种快速对抗训练方法 BAT, 它通过同时扰动原始输入图像和标签来进行对抗训练. 本文提出一种基于双标签监督的几何约束对抗训练算法, 在实验部分对比了之前的基准对抗训练方法. 实验结果表明, 我们提出的算法具有更强的鲁棒性.

5.3 对抗可解释性

对抗样本的反常现象引起了很多研究者对其内在机理的研究兴趣, 如何解释这一现象, 成为了近些年的研究热点. Szegedy 等人[4]认为, 对抗样本是高维流行空间中的低概率集, 神经网络很难识别这一低概率样本; GoodFellow 等人[16]研究表明, 对抗样本是由高维输入空间的线性性造成的; 而 Ilyas 等人[18]则提出对抗样本是一种特殊的特征, 他们把神经网络学到的特征分为鲁棒性特征和非鲁棒性特征, 神经网络模型的脆弱性主要是由于模型无法学习到鲁棒性特征所造成的, 因此只要能让神经网络模型学习到鲁棒性特征, 就可以防御对抗样本的攻击. 本文通过对双标签监督算法的分析, 基于鲁棒性特征和非鲁棒性特征的假设下, 本文提出的算法可以很好地解释神经网络的脆弱性, 双标签监督算法使神经网络模型学习到了更多的鲁棒性特征, 而

单标签监督的算法只能学习到非鲁棒性特征.

6 总结与展望

本文提出了一种基于双标签监督的几何约束对抗训练算法, 该方法通过约束神经网络高维空间特征的几何关系, 包括空间距离约束和相对角度约束, 来保证自然样本和对抗样本的特征分布一致性. 其次, 我们提出一种双标签监督训练方法, 利用对抗样本的真实标签和被攻击后的错误标签共同监督神经网络的训练, 并通过对双标签监督算法的分析, 在一定程度上揭示了对抗样本产生的内在机理. 为了验证提出算法的有效性, 我们在多个基准数据集上进行了实验分析. 实验结果表明: 本文提出的方法不仅提高了神经网络模型的鲁棒性, 还保持了模型一定的泛化能力, 且节省计算资源和训练时间, 是一种可行的对抗训练策略.

References:

[1] Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. In: Proc. of the Advances in Neural Information Processing Systems, Vol.25. 2012. 1097−1105.

[2] Hinton G, Deng L, Yu D, *et al*. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. IEEE Signal Processing Magazine, 2012, 29(6): 82−97.

[3] Mnih V, Badia AP, Mirza M, *et al*. Asynchronous methods for deep reinforcement learning. In: Proc. of the Int'l Conf. on Machine Learning. PMLR, 2016. 1928−1937.

[4] Szegedy C, Zaremba W, Sutskever I, *et al*. Intriguing properties of neural networks. In: Proc. of the Int'l Conf. on Learning Representations. 2014.

[5] Moosavi-Dezfooli SM, Fawzi A, Frossard P. Deepfool: A simple and accurate method to fool deep neural networks. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2016. 2574−2582.

[6] Cisse M, Adi Y, Neverova N, *et al*. Houdini: Fooling deep structured prediction models. arXiv: 1707.05373, 2017.

[7] Carlini N, Wagner D. Audio adversarial examples: Targeted attacks on speech-to-text. In: Proc. of the IEEE Security and Privacy Workshops (SPW). IEEE, 2018. 1−7.

[8] Akhtar N, Mian A. Threat of adversarial attacks on deep learning in computer vision: A survey. IEEE Access, 2018, 6: 14410−14430.

[9] Chakraborty A, Alam M, Dey V, *et al*. Adversarial attacks and defences: A survey. arXiv: 1810.00069, 2018.

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

[11] Madry A, Makelov A, Schmidt L, *et al*. Towards deep learning models resistant to adversarial attacks. In: Proc. of the Int'l Conf. on Learning Representations. 2018.

[12] Dhillon G, Azizzadenesheli K, Lipton Z, *et al*. Stochastic activation pruning for robust adversarial defense. In: Proc. of the Int'l Conf. on Learning Representations. 2018.

[13] Yang Y, Zhang G, Katabi D, *et al*. ME-Net: Towards effective adversarial robustness with matrix estimation. In: Proc. of the Int'l Conf. on Machine Learning. PMLR, 2019. 7025−7034.

[14] Song C, He K, Wang L, *et al*. Improving the generalization of adversarial training with domain adaptation. In: Proc. of the Int'l Conf. on Learning Representations. 2018.

[15] Zhang H, Yu Y, Jiao J, *et al*. Theoretically principled trade-off between robustness and accuracy. In: Proc. of the Int'l Conf. on Machine Learning. PMLR, 2019. 7472−7482.

[16] Goodfellow I, Shlens J, Szegedy C. Explaining and harnessing adversarial examples. In: Proc. of the Int'l Conf. on Learning Representations. 2015.

[17] Ren S, He K, Girshick R, *et al*. Faster R-CNN: Towards real-time object detection with region proposal networks. In: Proc. of the Advances in Neural Information Processing Systems, Vol.28. 2015. 91−99.

[18] Ilyas A, Santurkar S, Engstrom L, *et al*. Adversarial examples are not bugs, they are features. In: Proc. of the Advances in Neural Information Processing Systems. 2019.

[19] Wang J, Zhang H. Bilateral adversarial training: Towards fast training of more robust models against adversarial attacks. In: Proc. of the IEEE/CVF Int'l Conf. on Computer Vision. 2019. 6629−6638.

[20] Tramèr F, Kurakin A, Papernot N, *et al*. Ensemble adversarial training: Attacks and defenses. In: Proc. of the Int'l Conf. on Learning Representations. 2018.

[21] Wu L, Zhu Z, Tai C. Understanding and enhancing the transferability of adversarial examples. arXiv: 1802.09707, 2018.

[22] Dong Y, Pang T, Su H, *et al*. Evading defenses to transferable adversarial examples by translation-invariant attacks. In: Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition. 2019. 4312−4321.

[23] Krizhevsky A, Hinton G. Learning multiple layers of features from tiny images. Technical Report, Citeseer, 2009.

[24] Netzer, Y., Wang, T., Coates, A. *et al*. Reading digits in natural images with unsupervised feature learning. In: Proc. of the 2011 NIPS Workshop on Deep Learning and Unsupervised Feature Learning. 2011.

[25] Zagoruyko S, Komodakis N. Wide residual networks. In: Proc. of the British Machine Vision Conf. 2016. British Machine Vision Association, 2016.

[26] Papernot N, McDaniel P, Wu X, *et al*. Distillation as a defense to adversarial perturbations against deep neural networks. In: Proc. of the IEEE Symp. on Security and Privacy (SP). IEEE, 2016. 582−597.

[27] Samangouei P, Kabkab M, Chellappa R. Defense-GAN: Protecting classifiers against adversarial attacks using generative models. In: Proc. of the Int'l Conf. on Learning Representations. 2018.

[28] Guo C, Rana M, Cisse M, *et al*. Countering adversarial images using input transformations. In: Proc. of the Int'l Conf. on Learning Representations. 2018.

[29] Raghunathan A, Steinhardt J, Liang P. Semidefinite relaxations for certifying robustness to adversarial examples. In: Proc. of the Int'l Conf. on Neural Information Processing Systems. 2018. 10900−10910.

[30] Athalye A, Carlini N, Wagner D. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In: Proc. of the Int'l Conf. on Machine Learning. PMLR, 2018. 274−283.

[31] Athalye A, Carlini N. On the robustness of the cvpr 2018 white-box adversarial example defenses. arXiv: 1804.03286, 2018.

附中文参考文献:

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

曹刘娟(1983－), 女, 博士, 副教授, CCF 专业会员, 主要研究领域为机器学习, 模式识别.

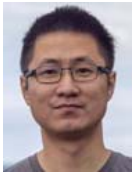
张宝昌(1976－), 男, 博士, 研究员, CCF 专业会员, 主要研究领域为视觉感知, 边缘计算.

匡华峰(1994－), 男, 博士生, 主要研究领域为对抗学习, 机器学习.

黄飞跃(1979－), 男, 博士, 高级工程师, 专业会员, 主要研究领域为机器学习与计算机视觉.

刘弘(1989－), 男, 博士, 主要研究领域为机器学习, 哈希检索.

吴永坚(1982－), 男, 博士, 研究员, CCF 专业会员, 主要研究领域为机器学习, 计算机视觉.



王言(1988－), 男, 博士, 工程师, 主要研究领域为多媒体内容检索, 机器视觉.

纪荣嵘(1983－), 男, 博士, 教授, 博士生导师, CCF 专业会员, 主要研究领域为计算机视觉, 机器学习.