软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2024,35(4):1861−1884 [doi: 10.13328/j.cnki.jos.006829]

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

基于可攻击空间假设的陷阱式集成对抗防御网络*

孙家泽[1,2,3], 温苏雷[1], 郑 炜[4], 陈 翔[5]

[1](西安邮电大学 计算机学院, 陕西 西安 710121)

[2](陕西省网络数据分析与智能处理重点实验室 (西安邮电大学), 陕西 西安 710121)

[3](西安市大数据与智能计算重点实验室 (西安邮电大学), 陕西 西安 710121)

[4](西北工业大学 软件学院, 陕西 西安 710072)

[5](南通大学 信息科学技术学院, 江苏 南通 226019)

通信作者: 温苏雷, E-mail: ccsk1wsl@stu.xupt.edu.cn

摘 要: 如今, 深度神经网络在各个领域取得了广泛的应用. 然而研究表明, 深度神经网络容易受到对抗样本的攻击, 严重威胁着深度神经网络的应用和发展. 现有的对抗防御方法大多需要以牺牲部分原始分类精度为代价, 且强依赖于已有生成的对抗样本所提供的信息, 无法兼顾防御的效力与效率. 因此基于流形学习, 从特征空间的角度提出可攻击空间对抗样本成因假设, 并据此提出一种陷阱式集成对抗防御网络 Trap-Net. Trap-Net 在原始模型的基础上向训练数据添加陷阱类数据, 使用陷阱式平滑损失函数建立目标数据类别与陷阱数据类别间的诱导关系以生成陷阱式网络. 针对原始分类精度损失问题, 利用集成学习的方式集成多个陷阱式网络以在不损失原始分类精度的同时, 扩大陷阱类标签于特征空间所定义的靶标可攻击空间. 最终, Trap-Net 通过探测输入数据是否命中靶标可攻击空间以判断数据是否为对抗样本. 基于 MNIST、K-MNIST、F-MNIST、CIFAR-10 和 CIFAR-100 数据集的实验表明, Trap-Net 可在不损失干净样本分类精确度的同时具有很强的对抗样本防御泛化性, 且实验结果验证可攻击空间对抗成因假设. 在低扰动的白盒攻击场景中, Trap-Net 对对抗样本的探测率高达 85% 以上. 在高扰动的白盒攻击和黑盒攻击场景中, Trap-Net 对对抗样本的探测率几乎高达 100%. 与其他探测式对抗防御方法相比, Trap-Net 对白盒和黑盒对抗攻击皆有很强的防御效力. 为对抗环境下深度神经网络提供一种高效的鲁棒性优化方法.

关键词: 深度神经网络; 对抗样本; 集成学习; 对抗防御; 鲁棒性优化

中图法分类号: TP18

中文引用格式: 孙家泽, 温苏雷, 郑炜, 陈翔. 基于可攻击空间假设的陷阱式集成对抗防御网络. 软件学报, 2024, 35(4): 1861–1884. http://www.jos.org.cn/1000-9825/6829.htm

英文引用格式: Sun JZ, Wen SL, Zheng W, Chen X. Trap-type Ensemble Adversarial Defense Network Based on Attackable Space Hypothesis. Ruan Jian Xue Bao/Journal of Software, 2024, 35(4): 1861–1884 (in Chinese). http://www.jos.org.cn/1000-9825/6829.htm

Trap-type Ensemble Adversarial Defense Network Based on Attackable Space Hypothesis

SUN Jia-Ze[1,2,3], WEN Su-Lei[1], ZHENG Wei[4], CHEN Xiang[5]

[1](School of Computer Science and Technology, Xi'an University of Posts and Telecommunications, Xi'an 710121, China)

[2](Shaanxi Key Laboratory of Network Data Analysis and Intelligent Processing (Xi'an University of Posts and Telecommunications), Xi'an 710121, China)

[3](Xi'an Key Laboratory of Big Data and Intelligent Computing (Xi'an University of Posts and Telecommunications), Xi'an 710121, China)

[4](School of Software, Northwestern Polytechnical University, Xi'an 710072, China)

[5](School of Information Science and Technology, Nantong University, Nantong 226019, China)

* 基金项目: 国家自然科学基金 (61876138, 62272387, 62141208); 国家重点研发计划 (2020YFC0833105Z1); 西安市重点产业链人工智能核心技术攻关项目 (2022JH-RGZN-0028); 陕西省重点研发计划 (2023-YBGY-030); 西安邮电大学创新基金 (CXJJZL2021007)
收稿时间: 2022-02-17; 修改时间: 2022-05-26; 采用时间: 2022-11-05; jos 在线出版时间: 2023-06-28
CNKI 网络首发时间: 2023-06-29

Abstract: Nowadays, deep neural networks (DNNs) have been widely used in various fields. However, research has shown that DNNs are vulnerable to attacks of adversarial examples (AEs), which seriously threaten the development and application of DNNs. Most of the existing adversarial defense methods need to sacrifice part of the original classification accuracy to obtain defense capability and strongly rely on the knowledge provided by the generated AEs, so they cannot balance the effectiveness and efficiency of defense. Therefore, based on manifold learning, this study proposes an origin hypothesis of AEs in attackable space from the feature space perspective and a trap-type ensemble adversarial defense network (Trap-Net). Trap-Net adds trap data to the training data based on the original model and uses the trap-type smoothing loss function to establish the seducing relationship between the target data and trap data, so as to generate trap-type networks. In order to address the problem that most adversarial defense methods sacrifice original classification accuracy, ensemble learning is used to ensemble multiple trap networks, so as to expand attackable target space defined by trap labels in the feature space and reduce the loss of the original classification accuracy. Finally, Trap-Net determines whether the input data are AEs by detecting whether the data hit the attackable target space. Experiments on MNIST, K-MNIST, F-MNIST, CIFAR-10, and CIFAR-100 datasets show that Trap-Net has strong defense generalization of AEs without sacrificing the classification accuracy of clean samples, and the results of experiments validate the adversarial origin hypothesis in attackable space. In the low-perturbation white-box attack scenario, Trap-Net achieves a detection rate of more than 85% for AEs. In the high-perturbation white-box attack and black-box attack scenarios, Trap-Net has a detection rate of almost 100% for AEs. Compared with other detection methods of AEs, Trap-Net is highly effective against white-box and black-box adversarial attacks, and it provides an efficient robustness optimization method for DNNs in adversarial environments.

Key words: deep neural network (DNN); adversarial example; ensemble learning; adversarial defense; robustness optimization

深度神经网络 (deep neural network, DNN) 作为人工智能最杰出的代表, 已被应用于图像分类、目标检测、语言识别等各个领域[1−3], 并且在这些领域展现出强大的性能. 然而研究表明, DNN 极易受到对抗样本的攻击. 对抗样本通过向干净样本添加特殊对抗攻击算法所生成的微小对抗扰动, 可以在不影响人类正常视觉辨别的同时使 DNN 产生分类错误, 是 DNN 安全领域的一大盲点. 因而提高 DNN 对对抗样本的防御能力, 提高 DNN 的对抗鲁棒性, 对 DNN 的后续研究和应用具有重大意义.

对抗样本的成因是对抗防御的一个重要前提. Szegedy 等人[4]认为训练数据不足导致 DNN 只能学习到目标数据流形的局部区域, 因而对抗样本所存在的数据流形的低概率区域未被模型正确划分是对抗样本存在的主要原因; Goodfellow 等人[5]认为 DNN 的脆弱性是由于模型在高维空间中存在的局部线性特征所导致; Gilmer 等人[6,7]认为 DNN 模型易受对抗样本攻击的主要原因在于目标数据流形的复杂高维几何结构. 综上, 本文认为对抗样本的成因, 关键在于目标数据流形与 DNN 特征空间的维度差异和训练数据本身所提供的特征信息不足而导致 DNN 的特征空间中暗藏大量对抗样本. 因此本文基于流形学习, 聚焦 DNN 的特征空间, 提出了可攻击空间对抗成因假设: DNN 通过利用训练数据所提供的信息对数据特征进行提取和分类, 以在广袤的 DNN 特征空间中定义对应的目标数据流形. 然而因为维度差异及空间不对等和缺乏特征信息等原因, 目标数据流形仅仅占据 DNN 特征空间中很少一部分, 而余留的未被训练数据所定义的广袤特征空间则是可能暗藏对抗样本的可攻击空间. 我们定义目标数据流形所占据的部分特征空间为 DNN 在具体事务中的特征敏感空间. 将可攻击空间中未被目标数据所影响的特征敏感空间和特征敏感空间之外的特征空间定义为背景可攻击空间. 将定义模糊的, 穿插在整体目标数据集流形之中的特征敏感空间定义为邻近可攻击空间.

从可攻击空间对抗成因的角度分析, 对抗训练作为目前最有效的对抗防御方法, 其通过向训练数据中增添对抗样本进行重训练对未定义的可攻击空间进行定义. 然而, Moosavi-Dezfooli 等人[8]指出, 无论添加多少对抗样本, 都存在新的对抗攻击样本可以再次欺骗网络. 这是因为对抗训练本身并没有向 DNN 中添加新的数据特征信息, 且其对抗防御的有效性依赖于现有的对抗样本所提供的信息, 所以在对抗防御泛化性方面有很大的缺陷. 我们根据可攻击空间假设, 提出一种新的对抗防御思路: 将暗藏对抗样本的可攻击空间标记而作为靶标, DNN 即可通过判定输入数据是否命中该靶标可攻击空间以区分输入数据是否为对抗样本.

综上, 为了提高深度神经网络的鲁棒性和对抗防御的泛化性, 本文基于可攻击空间对抗成因假设, 提出陷阱式集成对抗防御网络 (trap ensemble neural network against adversarial examples, Trap-Net). Trap-Net 通过向训练数据中添加目标数据类别之外的陷阱数据为网络模型提供更多的数据特征信息, 并使用这些新的数据特征为目标训练集数据未定义的可攻击空间赋予确切的身份类别, 从而消除未被模型认知而暗藏对抗样本的 DNN 特征空间. 同

时本文基于标签平滑[9]技术, 提出陷阱式平滑损失函数. 利用其作为诱导因子以加强目标数据和陷阱数据之间的联系, 诱使攻击算法所生成的对抗样本偏移至靶标可攻击空间. Trap-Net 通过集成学习的方式在保证不影响原目标数据分类精度的同时扩大靶标可攻击空间的大小, 最终通过判别输入样本是否命中靶标可攻击空间以区分输入样本是否为对抗样本.

本文的贡献可总结如下.

• 提出可攻击空间对抗成因假设. 因维度差异及空间不对等和训练数据不足导致 DNN 缺乏数据特征信息等原因, 导致 DNN 的特征空间中存在大量暗藏对抗样本的对应空间. 基于可攻击空间和目标数据流形之间的关系, 可攻击空间分为邻近可攻击空间和背景可攻击空间两大类.

• 基于可攻击空间对抗成因假设, 提出陷阱式集成对抗防御网络 Trap-Net. Trap-Net 能在保持目标数据分类精度的同时高效地探测输入样本是否为对抗样本. 与传统仅探测防御方法相比, Trap-Net 无需设计, 构建新的外部模块, 不依靠生成的对抗样本所提供的信息. Trap-Net 在白盒和黑盒攻击场景下都有极高的对抗样本探测率. 对各种程度的对抗扰动, 尤其是高扰动的对抗样本有极强的防御效力.

• 提出一种简单的集成学习子网络的扩充方式. 通过向同一模型结构的 DNN 中添加不同类别, 不同数目, 有别于目标数据类别的其他数据, 可以构建具有集成多样性的 DNN 子网络用于集成学习中的子网络扩充. 提出陷阱式平滑损失函数. 陷阱式平滑损失函数基于标签平滑技术, 于 Trap-Net 中作为诱导因子以增加靶标可攻击空间对攻击算法的吸引. 通过进一步诱导攻击算法所生成的对抗样本偏移至靶标可攻击空间以提高模型对对抗样本的探测效力.

• 在 5 个经典的深度学习数据集上进行了深入的研究, 验证了 Trap-Net 方法在不同数据集上的对抗防御有效性. 同时, 对 Trap-Net 的重要参数进行了实验与分析, 为模型的构建和相关参数的设置提出指导性意见. 为了方便拓展我们的研究工作, 我们对代码和实验结果进行了开源, 对应地址为: https://github.com/Ccsk-Xian/Trap-Net.

本文第 1 节介绍研究背景. 第 2 节详细介绍本文提出的陷阱式集成对抗防御方法. 第 3 节为实验展示及结果分析, 对 Trap-Net 的防御有效性和参数的选取, 陷阱数据的类别, 数目和陷阱数据与目标数据之间的关系进行实验与分析. 第 4 节介绍相关工作. 第 5 节对本文总结并对未来研究方向进行了展望.

1 研究背景

本节简要介绍 DNN 和对抗样本、流形学习、标签平滑和对抗攻击算法的相关背景知识. 这些背景知识是随后工作及实验的基础.

1.1 深度神经网络和对抗样本

深度神经网络 (DNN) 是一种从数据中提取特征并进行特征学习的多网络层数学经验模型. LeCun 等人[10]构建的卷积神经网络 (CNN) 是应用于图像分类领域的一种性能优越的 DNN. 本文使用的 ResNet 残差网络[11]是应用于图像分类和识别领域的一种易优化的 CNN. 本文定义 ResNet 为 $f(x_i,\theta): X \to Y$. 其中 $X \in R^d$ 为 d 维的输入样本特征集, $Y \in R^e$ 为 f 预测输出的 e 类输出分类向量. f 通过训练更新每层网络中的参数 θ. 当 f 训练完成后, 样本 $x_i \in X$ 输入 f 将输出其对应标签类别 $y_i \in Y$, 即 $f(x_i,y_i,\theta): g_F(g_{F-1}(\ldots(g_1(x_i)))) = y_i$. 其中, g_F 为 f 中的第 F 层网络, 包含隐藏层或卷积层以及激活函数和参数 θ 等.

传统计算机视觉的平滑假设认为 DNN 的输入 x 在被随机噪声所干扰时, 可以展现出极强的鲁棒性[12]. 然而不同于随机噪声, Szegedy 等人[4]证实 DNN 极易受到基于对抗攻击算法生成的对抗样本 x^* 的攻击. x^* 是攻击者通过向 x 添加精心制作的对抗扰动所生成的恶意攻击样本. x^* 无法从人类视觉的角度与 x 区分, 却能以极高的置信度使 DNN 分类错误, 是 DNN 安全领域的一大盲点.

1.2 流形学习

流形学习[13]用于形容无冗余的数据表示. 流形学习认为数据实际上是一种低维流形映射到高维空间的数据表示. 在高维空间中, 维度和数据的表示是不相关的, 数据在更低的维度中即可唯一表示. 而这种考虑了数据内部

特征的低维度模型表示为数据的流形. 流形学习认为在高维空间的距离度量中, 欧氏距离只适用于低维流形展开的空间, 而不能直接在高维的空间中进行度量. 直接使用欧氏距离在高维空间进行度量, 会丢失高维数据的绝大部分内部特征. 流形学习旨在刻画数据的本质, 这一点与深度学习利用模型提取并学习数据本质的特征相同.

1.3 标签平滑

DNN 的性能和损失函数的选取紧密相连. 传统 DNN 图像分类模型的损失函数为交叉熵损失函数 (CE), 其公式如下:

$$\mathcal{L}_{\mathrm{CE}}(x,y) = -\log(p_t) \tag{1}$$

其中, p_t 为目标类别的模型输出概率. DNN 通过训练使其预测的输出向量逐渐贴近正确标签的热独向量 (one-hot). 而 Szegedy 等人[9]提出的标签平滑技术修改了 CE 中的硬标签, 将目标类别的部分概率以均匀分布的方式分给其他标签, 使硬标签成为软标签, 从而更贴近人类对事物的判别逻辑. Müller 等人[14]对标签平滑技术的有效性进行了验证, 结果表明标签平滑可以提高模型的集束搜索能力, 泛化能力和修正能力, 同时减少模型对输出的过度“自信”, 在一定程度上防止模型过拟合, 标签平滑技术将 CE 损失函数修改为:

$$\mathcal{L}_{\mathrm{CE}}(x,y^{LS}) = -(1-\alpha)\log(p_t) - \alpha/(K-1)\sum_{i\neq t}\log(p_i) \tag{2}$$

其中, α 为标签平滑因子, 表示硬标签所分给其他类别的概率. K 为网络模型输出类别总个数, 即:

$$\hat{y}_i = \begin{cases} 1-\alpha+\alpha/K, & i = target \\ \alpha/K, & i \neq target \end{cases} \tag{3}$$

Müller 等人[14]对标签平滑在模型特征空间上的影响进行了实验. 结果表明标签平滑鼓励 DNN 所学习的不同类别的数据分布接近其对应的真实类别分布, 同时鼓励 DNN 所学习的不同类别的数据分布与其他类别的分布距离相等. 在其对细粒度类别图片进行的实验中, 相似类别的数据分布与差异性较大的数据分布呈现弧型包围状. 本文使用标签平滑对数据分布的优化特性, 提出陷阱式平滑损失函数. 将部分标签概率分给陷阱类数据以诱使对抗样本向陷阱类数据偏移的同时, 使用陷阱类数据将目标数据流形的外部特征空间包围, 从而进一步提高陷阱类别对目标数据流形之外特征空间的标记效率.

1.4 对抗攻击

对抗攻击通过从不同的攻击角度[15]生成高质量的攻击样本以揭露神经网络的脆弱性. 从攻击场景的角度, 对抗攻击可分为白盒攻击[5,16−19]和黑盒攻击[20,21].

• 白盒攻击: 攻击者可获取模型和训练数据的所有信息, 包括网络模型结构, 训练方式, 训练参数和训练过程中的梯度信息等.

• 黑盒攻击: 攻击者无法获取除输入数据和输出类别之外的其他信息.

根据对抗样本的生成原理, 对抗样本的攻击算法主要有以下 3 类.

• 基于梯度的对抗样本攻击算法[5,16,17]: 通过 DNN 的反向传播提取梯度数据信息生成对抗样本.

• 基于优化的攻击算法[18]: 通过特定的适应度函数进行对抗样本的搜索.

• 基于生成式对抗网络的攻击算法[19]: 通过生成对抗式网络的生成器和判别器之间的博弈, 生成高质量对抗扰动.

由于在白盒攻击场景下对抗样本攻击性更强, 本文将主要从白盒攻击的角度介绍以上 3 类攻击算法中的经典算法, 这些经典算法将在实验部分用于对防御方法的效力进行检测. 同时本文将基于代理模型生成对抗样本, 并通过对抗样本可迁移性对 Trap-Net 进行黑盒攻击的防御测试:

FGSM (fast gradient sign method) 由 Goodfellow 等人[5]在 2015 年提出. FGSM 是最简单的基于梯度的单步式对抗样本攻击算法. FGSM 的核心公式如下:

$$x^* = x + \varepsilon \cdot sign(\nabla_x J(\theta, x, y)) \tag{4}$$

FGSM 依据梯度方向的模型损失变化最大原理, 通过反向传播得到以 θ 为参数的模型损失 $J(\theta,x,y)$; 随后通过

sign 函数计算梯度方向, 并乘以扰动调节因子 ε 后生成对抗扰动; 最终原样本 x 添加对抗扰动生成对抗样本 x^*. FGSM 是最早的对抗攻击算法之一, 以 FGSM 算法衍生出了一大批攻击算法, 如 I-FGSM、PGD 等. FGSM 结构简单, 攻击力低, 是快速且廉价的对抗样本生成方式.

I-FGSM (iterative gradient sign method) 由 Kurakin 等人[16]在 2016 年提出. I-FGSM 又称基本迭代方法, 是 FGSM 的一种变体攻击算法. I-FGSM 使用迭代的方式, 沿着梯度增加的方向小步多次地生成扰动, 在相同步长下, I-FGSM 比 FGSM 有更强的攻击表现. 其核心公式如下:

$$x_{n+1}^* = x_n^* + clip_{\alpha,x}(\varepsilon \cdot sign(\nabla_{x_n^*} J(\theta, x_n^*, y))) \tag{5}$$

其中, x_{n+1}^* 代表经过一次迭代, 以更小的扰动因子 ε 生成的抗样本. I-FGSM 构造出的对抗扰动相较于 FGSM 更加精准. 但很显然, I-FGSM 攻击算法提高了对抗样本生成的计算量.

PGD (project gradient descent) 由 Madry 等人[17]在 2018 年提出. PGD 是一种基于 I-FGSM 的变体迭代式攻击算法. 与 I-FGSM 不同, PGD 攻击拥有更多的迭代次数, 并对输入噪声进行了随机初始化操作. 同时, 与 I-FGSM 直接在制定的范围内约束扰动大小不同, PGD 使用 l_∞ 范数映射予以替代. PGD 提出了最大最小化思想, 其公式如下:

$$\min_\theta \rho(\theta), \text{where } \rho(\theta) = E_{(x,y)\sim D}[\max_{\delta \in S} L(\theta, x+\delta, y)] \tag{6}$$

其中, 内部最大化旨在模型内部可以找出最强的对抗样本使得损失最大化. 而外部最小化旨在利用上一步生成的对抗样本进行对抗训练, 从而使得模型学习到更合适的参数来尽可能地降低数据分布上损失的期望. 与 I-FGSM 相比, PGD 拥有更强的攻击效果, 是目前最强的一阶攻击算法. 其核心公式为:

$$x_{t+1} = \prod_{x+S}(x_t + \varepsilon \cdot sign(\nabla_x J(x_t, y))) \tag{7}$$

其中, ε 为每次迭代的小扰动系数, S 为扰动的空间约束. 在每次迭代过程中, PGD 攻击方法都会将大于扰动阈值的扰动投影回扰动边界以保证扰动大小.

C&W (Carlini & Wagner) 由 Carlini 等人[18]在 2017 年提出. C&W 是一种基于优化的攻击算法. C&W 算法生成的对抗样本攻击性强, 扰动小. 但生成对抗样本的时间花销远大于其他攻击算法. 其生成核心公式为:

$$\begin{cases} r_n = \dfrac{1}{2}(\tanh(\omega_n)+1) - X_n \\ \min\limits_{\omega_n} \|r_n\| + c \cdot f\left(\dfrac{1}{2}\tanh(\omega_n)+1\right) \\ \text{where } f(x') = \max(\max\{Z(x')_i : i \neq t\} - Z(x')_t, -k) \end{cases} \tag{8}$$

其中, r_n 为对抗样本和干净样本的差值, c 为二进制搜索所选择的常数项, $Z(x)$ 为 Softmax 层输入向量, k 为对抗样本的置信度.

AdvGAN 由 Xiao 等人[19]在 2018 年提出. AdvGAN 是一种基于生成式对抗网络 (GAN) 的对抗样本生成方法. AdvGAN 由生成器 G, 判别器 D 和目标神经网络 C 构成. AdvGAN 通过 G 生成对抗扰动 $G(x)$, 并将对抗扰动添加到干净样本中生成对抗样本, 即 $x^* = x + G(x)$. D 将对 x^* 是否为对抗样本进行判别, 同时用 x^* 欺骗 C, 通过二者的反馈更新 G 的参数从而逐渐优化 $G(x)$ 的攻击性. AdvGAN 最终生成在视觉上与真实样本难以取分的对抗样本, 且生成的对抗样本相较于其他方法具有更强的迁移性[22].

2 陷阱式集成对抗防御网络

本节详细介绍陷阱式集成对抗防御方法. 图 1 展示了 Trap-Net 的结构图, 其中第 1 阶段基于流形学习, 从 DNN 特征空间的角度进行对抗成因的探讨, 提出可攻击空间对抗成因假设. 第 2 阶段根据可攻击空间对抗成因假设, 提出了一种针对对抗样本的陷阱式对抗防御思维. 陷阱式对抗防御思维的核心思想是使用可被探测的陷阱标记标注可能暗藏对抗样本的可攻击空间, 减少对抗样本对应的生成空间. 根据这种对抗防御思维, 构建了陷阱式集成对抗防御网络 Trap-Net. Trap-Net 基于集成学习的思想集成多个陷阱式网络, 可以在不损失原分类精度的同时扩大靶标可攻击空间的大小. Trap-Net 通过判断输入是否命中被陷阱类别标记的靶标可攻击空间以区分输入是否

为对抗样本. 第 3 阶段基于标签平滑技术, 提出陷阱式平滑损失函数以优化数据类型之间的关系, 并建立目标数据与陷阱数据之间的诱导关系, 使得各目标类别分布更加内聚且诱使对抗攻击方法生成偏向靶标可攻击空间的对抗扰动, 从而进一步提升 Trap-Net 的对抗防御效力.

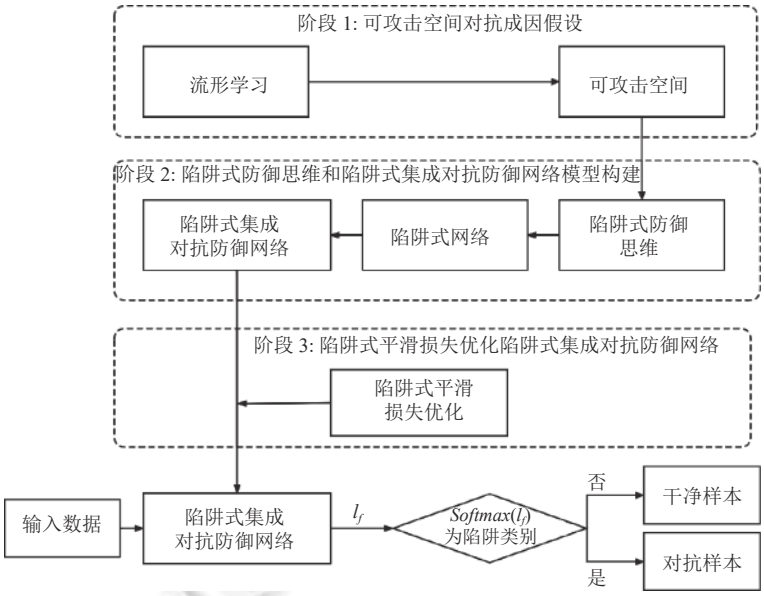


图 1 陷阱式集成对抗防御方法结构图

2.1 可攻击空间

基于流形学习和过往研究人员对对抗样本成因的假设[4−7], 本文认为 DNN 主要因为以下两点原因而存在对抗样本:

• 维度差异及空间不对等: 理论上, DNN 的特征空间是可以随着深度神经网络的结构和参数的扩增而无限增大的. 为了提高 DNN 的训练效率和防止训练过程中梯度消失及梯度爆炸现象的发生, 模型的训练数据需要预先标准化到一定范围之后再输入 DNN 进行训练. 这等同于在无限大的特征空间中规划出一片小范围的, 针对具体事务的特征敏感空间. 而由于目标数据流形所代表的特征敏感空间与总体模型特征空间之间的维度差异和空间大小不对等, 特征空间和特征敏感空间中存在大部分未被目标数据直接定义的特征空间. 这些未被目标数据直接定义的特征敏感空间和特征敏感区域之外的特征空间则存在暗藏对抗样本的可攻击空间.

• 缺乏特征信息: DNN 为了在不同事务中尽可能学习到目标数据的数据分布, 最直接的方式就是扩大 DNN 的广度和深度以谋求模型在特征空间中更佳的特征搜索能力. 与 DNN 优秀的特征搜索, 学习和分类能力相比, 训练数据所提供的数据特征信息却无法满足目标客观事物于特征敏感空间中所有区域的准确定义. 这就造成了当输入样本包含的特征信息并不包含于训练数据的数据流形时, 貌似高精度的 DNN 会被迫强制输出非确定的样本所属类别, 从而极易产生分类错误[5]. 这本身是一种因为数据特征信息不足与强大的模型数据分析能力不对等所造成的过拟合问题. 这将造成不同目标数据类别的流形之间存在大量定义模糊的, 暗藏对抗样本的特征敏感空间.

综上, 我们对可攻击空间有如下定义.

定义 1 (可攻击空间). 在 DNN 的特征空间中, 对不隶属于训练数据的数据流形之外的 DNN 特征空间定义为广义上的, 暗藏对抗样本的可攻击空间.

根据 Goodfellow 等人[5]对 DNN 特征空间中高维线性的验证与分析以及模型输入数据的预处理标准化操作, 训练数据的数据流形在网络模型所敏感的特征空间中只占据很少一部分. 除了目标数据流形所占据的小部分特征空间之外, DNN 存在广袤的, 暗藏对抗样本的可攻击空间. 我们根据目标数据流形与 DNN 特征空间之间的关系将

可攻击空间分别定义为邻近可攻击空间和背景可攻击空间.

定义 2 (邻近可攻击空间). 邻近可攻击空间位于目标数据流形所处的特征敏感空间之中. 图 2 展示了可攻击空间与 DNN 特征空间的示意图. 邻近可攻击空间位于特征敏感空间之内, 其穿插在虚线和实线所示的不同数据类别流形之间的高维断层空间之中, 是有目标对抗样本所存在的特征空间.

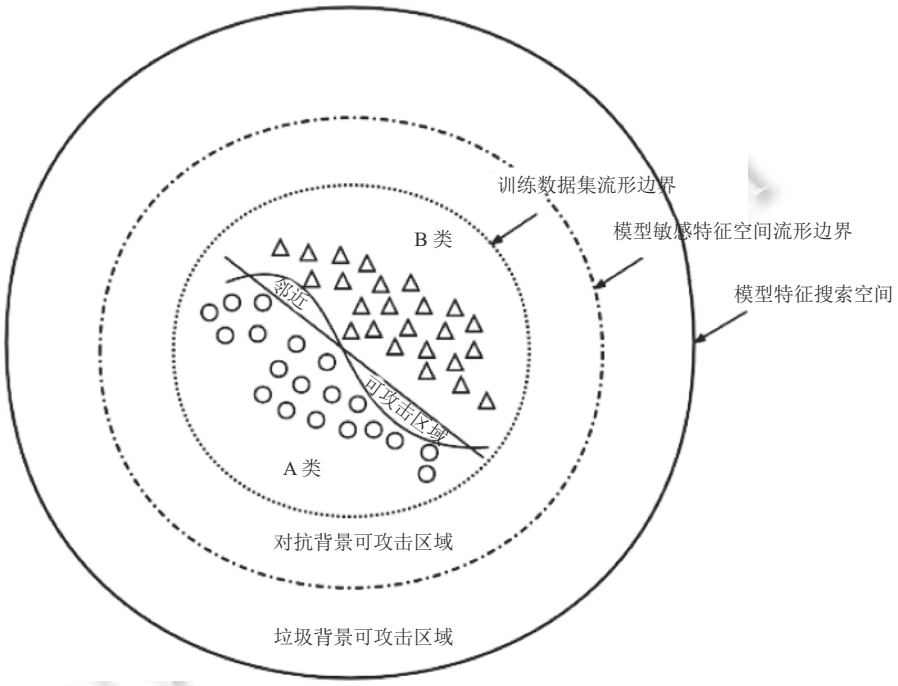


图 2 可攻击空间示意图

定义 3 (背景可攻击空间). 背景可攻击空间位于整体目标数据流形之外的广袤特征空间之中. 根据无目标攻击区域和垃圾图像区域可将背景可攻击空间细分为对抗背景可攻击空间和垃圾背景可攻击空间. 这两个区域分别代表无目标对抗样本的存在区域和垃圾图片的存在区域. 图 2 所示的训练数据集流形边界和模型敏感特征空间流形边界之间的对抗背景可攻击空间更贴近训练数据流形空间, 而模型敏感特征空间流形边界之外的特征空间被称为垃圾背景可攻击空间.

总体上, 可攻击空间的概念图如图 2 所示. 图中小圆代表 A 类数据, 三角形代表 B 类数据. 中心实线为真实决策边界, 中心虚线为模型拟合的决策边界. 邻近可攻击空间指两个不同类别的数据流形之间真实决策边界和模型拟合的决策边界为界所形成的高维特征空间. 图中点划线代表的是因模型数据预处理所导致的数据特征敏感空间. 其中, 虚线所代表的训练数据集流形边界是模型特征敏感空间中的一部分. 而对抗背景可攻击空间则存在于这两个边界之间. 垃圾背景可攻击空间因不符合模型数据预处理规则, 存在于特征敏感空间之外.

邻近可攻击空间和背景可攻击空间是平滑区域和对抗区域所混杂的特征空间. 平滑区域符合计算机视觉的传统平滑假设, 即数据样本增添微小扰动或噪声之后, 网络模型仍能以正确的类别输出. 而对抗区域是暗藏对抗样本的特殊特征空间, 需要以特定的攻击算法作为钥匙进行搜寻.

2.2 陷阱式集成对抗防御网络

传统训练模式中, DNN 没有被赋予对输入样本怀疑的权力以及针对对抗样本的判断逻辑. 所以 DNN 只能凭借已有的数据特征信息为高维特征空间中的每个点赋予不同类别对应的输出概率. 而可攻击空间的关键在于其所处的特征空间未被训练数据所直接定义. 这些未被训练数据所直接影响的可攻击空间无法在训练数据的帮助下对特征空间进行目标类别的分类和标记, 而这正为对抗样本的存在提供了基础条件. 如果赋予 DNN 对输入数据怀

疑的权力, 则可以在一定程度上避免模型因目标类别特征信息不足和输出类别受限而导致的被迫犯错. 模型可通过判别输入数据是否命中可攻击空间以判别输入数据是否为对抗样本. 我们将这种新的对抗防御思路称之为陷阱式防御思维.

Pang 等人[23,24]的研究表明, 对单个深度神经网络而言, 提升其从不同类别数据之间所学习到的特征的多样性, 能极大提升深度神经网络的对抗鲁棒性. 这表明可通过给予深度神经网络更多的数据特征信息以提升网络模型的鲁棒性. 这启发了我们可以从陷阱式防御思维的角度, 通过为 DNN 增添更多的数据特征信息以作为陷阱标签来标记未定义的可攻击空间, 以此赋予 DNN 对输入怀疑的权力. 通过判定输入样本是否命中靶标可攻击空间来分辨对抗样本与干净样本. 理论上, 当被标记的陷阱区域等于可攻击空间时, 可消除对抗样本对 DNN 鲁棒性的影响.

综上, 我们构建了陷阱式网络. 通过向 DNN 增添多个有别于训练数据类别, 且不影响原目标数据精确度的其他类别数据作为陷阱数据, 使用相同的深度神经网络训练过程生成陷阱式网络. 陷阱式网络通过提取陷阱数据中的特征信息进行学习, 利用所学习到的新的特征信息以标记原本未被定义的可攻击空间. 这些被陷阱数据类别所定义的可攻击空间被称为靶标可攻击空间. 因为目标数据与陷阱数据之间的差异性, 干净输入数据不会被分类为陷阱数据类别, 所以在应用中我们可以将被分类为陷阱类别的输入样本判定为对抗样本.

假设原干净样本 x 在特征空间中处于其自身的数据流形之中, 对抗攻击通过向 x 添加对抗扰动 ε 生成对抗样本 x^*, 从而使得 x 远离并偏移至可攻击空间中对抗样本所存在的特征空间. 理论上, Trap-Net 的有效性是因为以下 4 种情况.

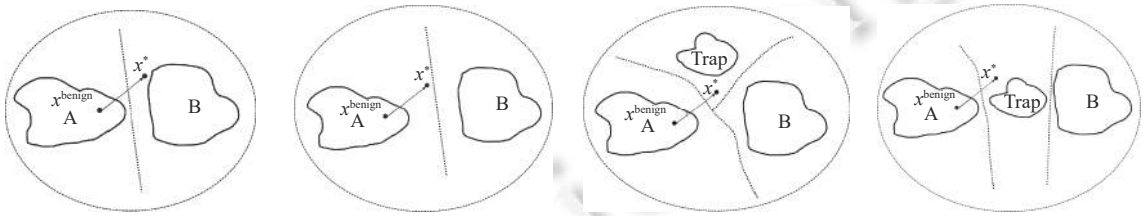
情况 1. 陷阱数据所带来的特征信息优化了深度神经网络的分类能力, 使得原目标数据类别之间的差异性变大, 不同类别间的数据流形距离变大, 导致被 ε 扰动后的 x^* 无法偏移至其他流形区域中.

情况 2. 在邻近可攻击空间中, 陷阱数据流形的加入改变了原本不同目标数据类别流形之间的决策边界, 从而使原本邻近可攻击空间被陷阱数据流形所定义, 最终截获了被 ε 扰动后的 x^*.

情况 3. 在邻近可攻击空间中, 陷阱数据的流形插入到了原本不同数据类别流形之间. 使得被 ε 扰动后的 x^* 只能偏移至陷阱数据流形之中.

情况 4. 陷阱数据定义了背景可攻击空间, 从而使得被 ε 扰动后的 x^* 偏移至已被定义的背景可攻击空间.

图 3(b)–图 3(d) 在二维空间的简单二分类模型中展现了上述前 3 种情况. 其中, 图 3(a) 展示干净样本 x^{benign} 添加对抗扰动 ε 后对抗样本 x^* 跨过决策边界, DNN 分类错误. 图 3(b) 展示陷阱数据提升了不同目标类别的数据流形之间的差异, 使得被 ε 扰动后的 x^* 留在了正确的特征空间区域. 图 3(c) 展示陷阱数据改变了原有的决策边界, 使得被 ε 扰动后的 x^* 被分类为陷阱类. 图 3(d) 展示陷阱数据流形插入不同类别的目标数据流形之间, 使得被 ε 扰动后的 x^* 被分类为陷阱类.



(a) 2D 子流形区域　(b) 扩大流形间距离　(c) 修改决策边界　(d) 插入流形之间

图 3　Trap-Net 在二维空间中的防御效力解释图

不同于传统仅探测防御方法基于已有的对抗样本, 通过训练额外的二分类神经网络[25], 或根据某种距离度量方式[26]以区分输入数据是否属于对抗样本. Trap-Net 不依靠已有的对抗样本所提供的信息, 其对抗防御效力取决于陷阱类别在特征空间所定义的靶标可攻击空间的大小. 为了进一步扩大网络模型的靶标可攻击空间, 我们利用集成学习的方式集成多个陷阱式网络, 从而在不影响原目标数据的分类精度的基础上尽可能扩大靶标可攻击空间. 最终生成对目标数据的分类结果鲁棒性更强且靶标可攻击空间所占特征空间更大的陷阱式集成对抗防御网络 Trap-Net.

图 4 展示了 Trap-Net 的模型结构图. Trap-Net 旨在利用集成学习的方式尽可能扩大靶标可攻击空间的大小, 在保持原目标数据集分类精度的同时加强对抗防御效力. 通过靶标可攻击空间的定义及对抗样本的探测逻辑, 对对抗样本进行后验式探测防御. 与传统仅探测式对抗防御方法相比, Trap-Net 基于已有的神经网络模型, 无需设计其他外部结构, 且不依靠并受限于已有的对抗样本所提供的信息.

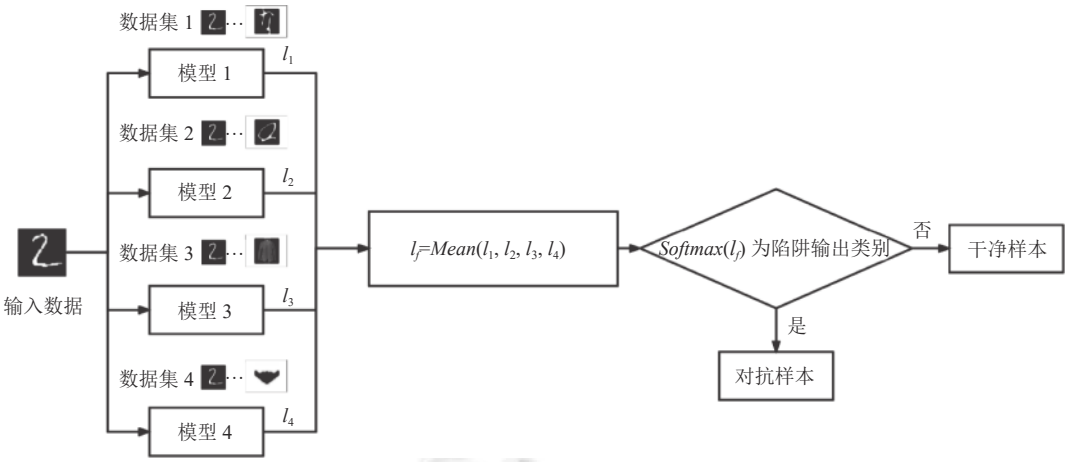


图 4 Trap-Net 模型结构图

考虑到当目标数据集过大时, 无法获取大量有效的陷阱数据集. Trap-Net 可将目标数据集自身的部分数据作为陷阱数据集进行陷阱式网络的训练. 当得到原始 DNN 的输出结果时, 根据输出类别进行不同陷阱式集成网络的验证. 当且仅当二者输出的分类类别相同时, 输入样本为干净样本. 这种结构的网络模型称为后验陷阱式集成网络模型, 后验陷阱集成网络模型结构图如图 5 所示.

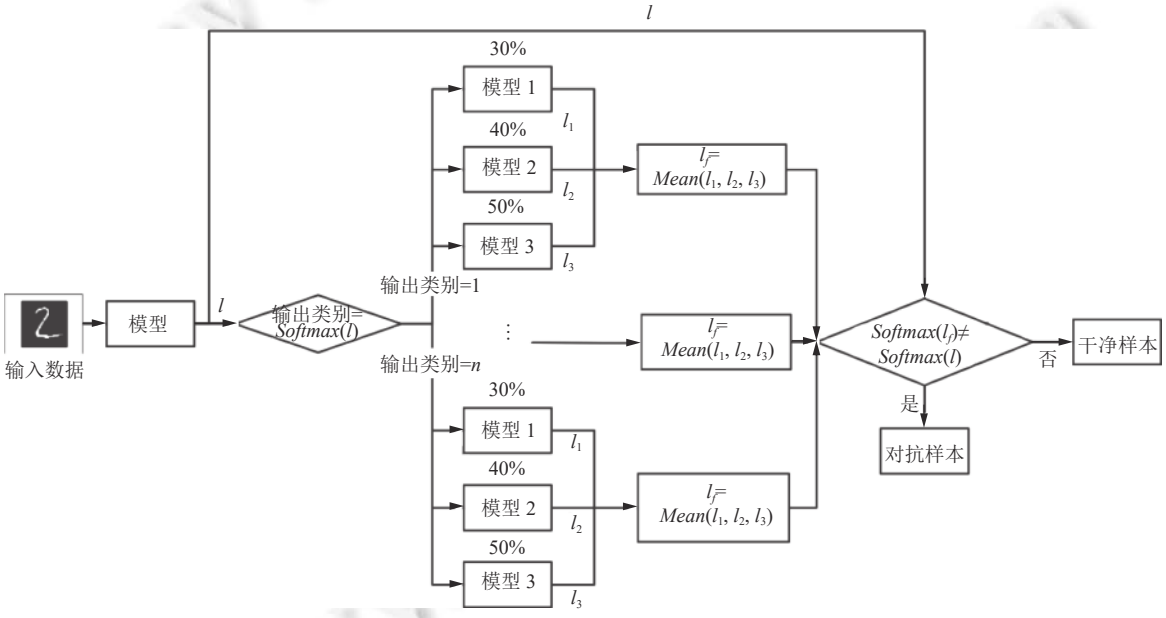


图 5 后验陷阱式集成网络结构图

图 5 中, {模型 1,…,模型 3}代表使用不同百分比的它类目标数据作为陷阱数据集, 用以训练该输出类别的后验陷阱式网络. 输入数据首先经过目标神经网络“模型”后得到预测标签输出类别, 根据输出类别的数值使用不同

的后验陷阱式集成网络进行后验式对抗样本验证. 其中, 每一个陷阱式集成网络由不包含目标标签类别的不同百分比数据作为陷阱数据类别训练组成. 通过对比陷阱式集成网络的输出是否与模型的输出类别一致以判断输入数据是否为对抗样本. 当该基准预测结果与目标模型的输入一致时, 输入样本为干净样本. 当该基准预测结果与目标模型的输出不一致时, 输入样本为对抗样本.

2.3 陷阱式平滑损失函数

基于 Pang 等人[27]对集成网络中子网络多样性的探究和 Müller 等人[14]对标签平滑在神经网络领域有效性的验证实验, 本文提出陷阱式平滑损失函数. 陷阱式平滑损失函通过将目标训练数据硬标签中的部分概率平分给陷阱数据集中的各个陷阱数据类别, 建立两个数据集流形的平滑诱导关系, 以诱使对抗样本更易偏向生成于被标记的靶标可攻击空间.

我们对陷阱式平滑损失有如下定义.

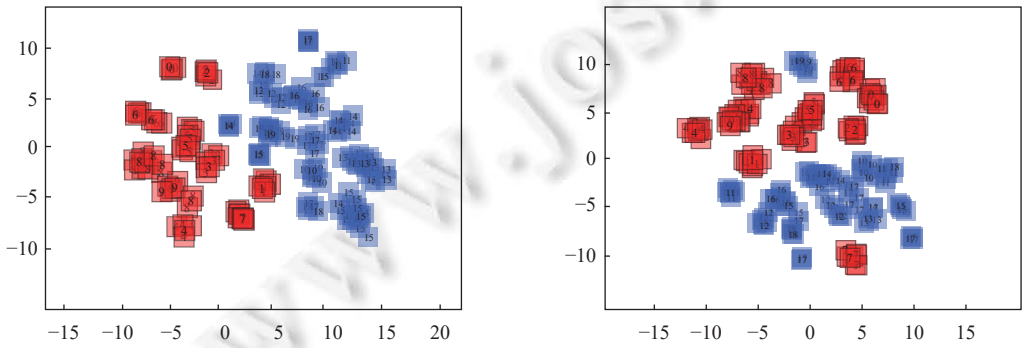
定义 4 (陷阱式平滑). 定义 K 为目标数据集的类别数, α 为陷阱诱导因子, T 为陷阱数据集的类别数. 陷阱式平滑将目标类别的概率以均匀分布的方式分散给陷阱类别, 陷阱式平滑损失函数的公式如下:

$$L_{\mathrm{CE}}(x,y^{LS}) = -(1-\alpha)\log(\rho_t) - (\alpha/K)\sum_{i=K+1}^{K+T}\log(\rho_i) \tag{9}$$

基于 Müller 等人[14]使用标签平滑在 ImageNet 细粒度类别上进行的对比实验, 修改后的陷阱式平滑损失函数将目标数据集作为一个整体与其他细粒度陷阱类数据进行标签平滑, 可使陷阱式平滑损失函数在鼓励目标数据分类为正确类别的同时, 促使目标数据中各类数据和陷阱数据中各类数据的距离相同. 理论上, 最终各类陷阱数据将呈圆球状包围目标数据集流形. 陷阱式平滑损失函数对目标数据流形和陷阱数据流形之间关系的优化, 极大地提高了陷阱类数据对暗藏对抗样本的可攻击空间, 尤其是背景可攻击空间和垃圾背景可攻击空间的陷阱类别标记, 从而进一步扩大了靶标可攻击空间的大小, 进而极大地提高了 Trap-Net 的对抗防御效力.

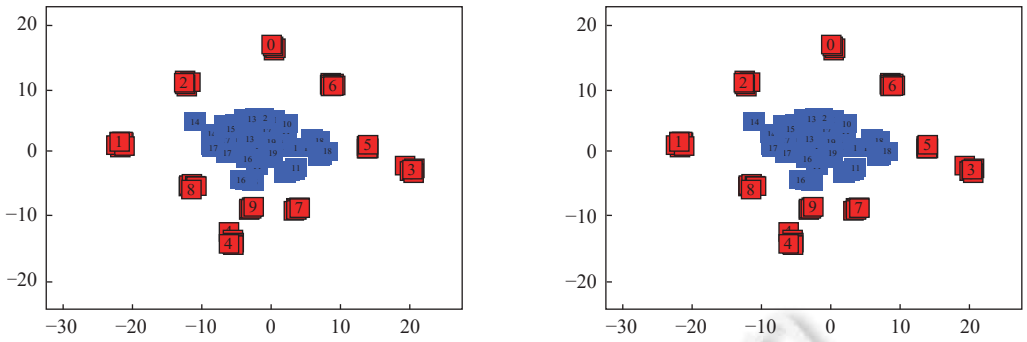
本文使用 T-SNE 降维可视化技术[28]对陷阱式网络使用陷阱式标签平滑作为陷阱诱导因子进行模型数据分布的二维和三维可视化展示.

如图 6 所示, 带有边框的类别{0,…,9}为目标数据集, 无边框的类别{10,…,19}为陷阱数据集. 图 6(a) 为以 CE 为损失函数训练的陷阱式网络. 当损失函数为 CE 时, DNN 最后一层全连接层中目标数据与陷阱数据的数据分布之间没有明显的交集, 陷阱数据对目标数据流形的影响有限. 图 6(b)–图 6(d) 为以陷阱式标签平滑为损失函数训练的网络. 陷阱类数据逐渐穿入目标数据的数据流形之中. T-SNE 类别之间的距离反应类别之间服从 t 分布的相似度. 添加陷阱数据集并充分训练后, 由目标数据集所直接定义的 DNN 特征空间变大, 原目标数据的不同类别与陷阱数据集整体的距离相同. 这反应出添加陷阱式标签平滑作为诱导因子之后, 陷阱数据穿插进入目标数据集流形中, 将目标数据流形中的邻近可攻击空间以陷阱类别定义. 陷阱类别数据在陷阱式平滑损失函数的影响下对外部背景可攻击空间标记的实验及分析, 将在第 3 节实验部分的 RQ2 进行讨论.



(a) 目标数据与陷阱数据明确区分　　(b) 陷阱数据穿入目标数据流形

图 6　DNN 数据分布 T-SNE 视图



(c) 目标数据类别与陷阱数据集等距离包围　　(d) 陷阱数据成功插入目标数据流形中

图 6　DNN 数据分布 T-SNE 视图 (续)

3　实验及评估

3.1　实验设置

3.1.1　实验数据

本文使用标准数据集 MNIST, K-MNIST, F-MNIST, CIFAR-10 和 CIFAR-100 进行 Trap-Net 的验证及实验分析, 以证明陷阱式集成对抗防御方法的有效性. 其中, 以 MNIST 为主体目标数据, 使用 K-MNIST, F-MNIST, CIFAR-10 和 CIFAR-100 为陷阱数据进行 Trap-Net 的构建以及对抗防御效力验证. MNIST (D1), K-MNIST (D2), F-MNIST (D3) 属于不同图片类型, 大小为 28×28 单通道灰度图片. CIFAR-10 (D4) 与 CIFAR-100 (D5) 是大小为 32×32 的三通道彩色图片. 表 1 给出了数据集的具体参数及介绍.

表 1　数据集介绍

ID	数据集	训练集	测试集	种类	规格	数据集描述
D1	MNIST	60000	10000	10	28×28	灰度图片—手写数字图片
D2	K-MNIST	60000	10000	10	28×28	灰度图片—平假名图片
D3	F-MNIST	60000	10000	10	28×28	灰度图片—衣物图片
D4	CIFAR-10	50000	10000	10	32×32	RGB图片—动物, 交通工具图片等
D5	CIFAR-100	50000	10000	20/100	32×32	RGB图片—20大超类, 各含5子类

3.1.2　研究问题

为了验证 Trap-Net 的对抗防御有效性以及模型参数对防御效力的影响, 我们从以下问题进行研究, 并进行相应的实验与分析.

● RQ1: 陷阱式网络能否作为一种简单, 有效的集成网络中子网络的扩充方式以提高集成多样性?

对抗集成网络防御方法通过集成多个子网络模型以构成鲁棒性更强的集成网络. 由于对抗样本存在可迁移性, 集成网络中子网络的特征多样性对集成网络整体的鲁棒性显得尤为重要[23]. 陷阱式网络旨在通过向 DNN 添加新的陷阱数据类别, 为 DNN 特征空间增添新的数据特征信息. 理论上, 陷阱式网络有利于为集成网络提升整体的特征多样性. 因此在本问题研究中, 我们从对抗防御效力的角度进行评估, 与现有的对抗集成网络方法进行对比, 探究陷阱式网络能否可以在不影响 DNN 应对高斯噪声等传统鲁棒性的同时, 作为一种简单有效的集成网络子网络扩充方式.

● RQ2: Trap-Net 能否对对抗样本的迁移性进行有效的防御?

对抗样本已被证明有极强的可迁移性[29]. 攻击者可通过代理模型生成对抗样本或对抗扰动, 并利用对抗样本的可迁移性有效攻击目标黑盒 DNN. 在本问题的研究中, 我们使用基于梯度的对抗攻击方法以及可迁移性更强的

基于生成式对抗网络的对抗攻击方法对 Trap-Net 进行黑盒攻击, 以验证 Trap-Net 能否对对抗样本的迁移性进行有效的防御.

● RQ3: Trap-Net 的相关参数对防御方法的防御效力影响如何?

DNN 是一种数据驱动的数学经验模型, 各种 DNN 的超参数直接影响 DNN 最终的性能表现. Trap-Net 最大特色即引入陷阱数据集对特征空间中的未定义可攻击空间进行陷阱类别标记. 因此, 在本问题的研究中, 我们围绕陷阱数据类别个数和陷阱数据与目标数据间的相似性对 Trap-Net 对抗防御效力的影响进行实验探究与分析.

① RQ3.1: 陷阱数据类别个数对防御效力的影响如何?

Trap-Net 通过向 DNN 中增添不同特征类型的陷阱数据进行靶标可攻击空间的定义. 陷阱数据类别个数的多少直接影响着防御成本与防御效力的好坏. 在本问题研究中, 我们围绕陷阱类数据类别个数的多少对 Trap-Net 的防御效力的影响进行实验分析.

② RQ3.2: 陷阱数据集与目标数据间相似性大小对防御效力的影响如何?

在本问题研究中, 我们承接陷阱数据类别个数对 Trap-Net 对抗防御效力的影响, 从陷阱数据与目标数据之间的数据相似性角度对陷阱数据对 Trap-Net 的对抗防御效力影响进行实验分析.

● RQ4: 陷阱式平滑损失函数能否优化目标数据与陷阱数据之间的数据分布, 从而进一步提高 Trap-Net 的对抗防御效力?

不同于传统集成对抗防御依靠模型之间的特征差异性, 以集成多个子网络鲁棒性的方式形成一个更加鲁棒的对抗集成网络. Trap-Net 的集成性质在于利用不同陷阱式网络的多样性扩充靶标可攻击空间, 从探测的角度抵御对抗样本的攻击. 为了凸显 Trap-Net 的这种特点, 我们提出并使用了陷阱式平滑损失函数. 直观上, 陷阱式平滑损失函数能建立目标数据流形和陷阱数据流形之间的诱导关系. 使得靶标可攻击空间更易被对抗攻击算法所攻击. 从而达到双赢的效果, 即攻击者制作的对抗样本可成功将原数据类别进行变换, 然而实际上对抗样本被诱导指向靶标可攻击空间, 从而被 Trap-Net 所识别判定为对抗样本. 在本研究问题中, 我们对陷阱类别数据在陷阱式平滑损失函数的影响下对外部背景可攻击空间的标记进行实验探究. 与在 CE 损失函数下的 Trap-Net 的防御效力进行对比, 以验证陷阱式平滑损失函数的有效性, 并对陷阱式诱导因子的大小对防御效力的影响进行实验探究.

● RQ5: Trap-Net 相较于其他类似的对抗防御方法的防御效力如何?

为了验证 Trap-Net 的有效性. 在本问题研究中, 我们使用核密度和贝叶斯不确定性估计法, 特征压缩对抗防御方法以及频谱对抗防御方法与 Trap-Net 进行比较. Feinman 等人[26]所提出的核密度和贝叶斯不确定性估计法是一种仅探测式对抗样本防御方法, 这种方法通过计算正常样本和对抗样本的核密度以及贝叶斯不确定性分数, 训练生成对抗样本的逻辑回归预测模型以进行对抗样本的检测. 然而核密度和贝叶斯不确定性估计法十分依赖于生成的对抗样本所提供的信息. 所以我们同时选用了 Xu 等人[30]提出的特征压缩对抗防御方法进一步进行对比实验. 特征压缩对抗防御方法无需依赖生成的对抗样本所提供的信息, 其通过对输入样本进行色位压缩和中值平滑后破坏对抗样本的攻击性, 通过比对变化前后的 *Softmax* 向量差是否超过阈值以判断输入样本是否为对抗样本. 最后, 我们选用 Harder 等人[31]提出的频谱对抗防御方法与 Trap-Net 进行比较. 频谱对抗防御方法通过傅里叶变换发掘图片中的频谱信息, 并以此分辨干净样本与对抗样本. 我们对以上 3 种防御方法与 Trap-Net 进行了对比实验, 并对这几种方法之间的特点和差异进行了分析.

● RQ6: Trap-Net 与其他对抗防御方法的兼容性如何?

邻近可攻击空间中的未定义可攻击空间标记难度大, 造成 Trap-Net 在有目标攻击场景下面对不同大小的对抗扰动攻击时, 对抗防御效力呈的“V”型结构. 因此在本问题研究中, 我们希望通过与其他对抗防御方法合作的方式解决这一问题. 我们选择主流公认最有效的对抗训练防御方法对 Trap-Net 进行与其他对抗防御方法的兼容性测试.

3.1.3 评测指标

受试者工作特征曲线 (receiver operating characteristic curve, ROC 曲线). ROC 曲线在不同的阈值大小下以同一刺激信号构建以假阳性率 (FPR) 为横坐标, 召回率 (TPR) 为纵坐标的评价曲线. 其中, ROC 曲线的下面积 AUC

用于评估模型的性能. AUC 取值范围在 0.5–1 之间, 一般认为 AUC 值为 0.5–0.7 之间时, 模型性能低; 为 0.7–0.9 之间时, 模型性能较好; 为 0.9 以上时, 模型性能优越.

探测后分类准确率. 与传统分类准确率不同, 本文将成功判定为对抗样本的输入也定义为分类正确. 依据混淆矩阵的原理, 将陷阱集成网络模型正确分类目标数据的部分定义为真阳性 (TP), 将正确分类为对抗样本的对抗样本输入分类为真阴性 (TN), 当所有分类结果以 All 表示时, 探测后分类准确率用公式 (10) 表示:

$$Accuracy = \frac{TP + TN}{All} \tag{10}$$

3.1.4 对比方法

核密度和贝叶斯不确定性估计法由 Feinman 等人[26]在 2017 年提出. 其假设对抗样本和训练数据处在不同的流形之中, 提出使用核密度估计法和贝叶斯不确定性组合方法进行对抗样本的检测. 该方法利用神经网络最后一层全连接层的逻辑输出, 通过计算正常样本和对抗样本的核密度以及贝叶斯不确定性分数, 训练生成对抗样本的逻辑回归预测模型以进行对抗样本的检测. 这种方法虽然简单有效, 但十分依赖于已有生成的对抗样本的所提供的信息, 对新的对抗攻击方法所生成的对抗样本, 其防御效力将大大降低.

特征压缩法由 Xu 等人[30]在 2018 年提出. 其认为特征空间过大导致生成对抗样本的机会增多. 因此提出了基于特征压缩的对抗样本检测方法. 他们利用色位压缩和中值平滑两种方式进行输入样本的特征压缩. 该方法通过对比压缩和非压缩样本的输出结果之间的 L_1 范数差是否大于阈值 T 以区分输入样本是否为对抗样本. 这种方法可以有效地嵌入各种神经网络以检测对抗样本. 但该方法会以牺牲部分目标数据的分类精度为代价, 具有较高的误报率, 且无法有效防御高扰动的对抗样本攻击.

频谱对抗防御方法由 Harder 等人[31]在 2021 年提出. 其通过使用傅里叶变换, 分别从干净样本和对抗样本的样本本身及其深度神经网络各层不同的特征图中提取正常样本所得不到的频谱信息, 并利用这些频谱信息训练生成针对对抗样本检测的回归预测模型. 根据频谱信息的来源不同, 其制定了两种频谱对抗防御方法. 其中, 频谱信息源于样本本身的防御方法称为高量傅里叶频谱对抗防御方法 (MFS); 频谱信息来源于深度神经网络各层特征图的被称为阶段性傅里叶频谱对抗防御方法 (PFS). 频谱对抗防御方法对已知的同类对抗攻击方法具有很好的鲁棒性, 但依赖于已知对抗样本所提供的数据信息, 且在面对不同类型的对抗攻击的攻击时, 防御效力有一定程度的降低.

3.2 方法实现和实验运行环境

实验使用 Python 3.7.3 编码完成, 基于 PyTorch 1.7.0 进行深度神经网络模型的构建以及相关数据集的测试工作. 在实验过程中, 本文通过使用白盒攻击和黑盒攻击方法生成对抗样本以验证 Trap-Net 对抗防御效力的有效性. 根据攻击原理不同, 本文基于深度神经网络框架 sklearn[32]和 cleverhans[33]和 DeepRobust[34]等对抗样本框架, 使用基于梯度的对抗样本生成方法 FGSM, PGD、基于优化的对抗样本生成方法 C&W 以及基于生成式对抗网络的攻击方法 AdvGAN 对 Trap-Net 于白盒和黑盒的攻击场景下进行鲁棒性实验及评估. 我们首先对数据集进行了归一化预处理: 将所有的训练样本归一化到区间 [−1, 1] 之中. 同时, 对数据大小不同的陷阱数据集, 我们采用了中心裁剪的方式以完成其与目标数据集的合成. 我们将设置初始学习率为 0.005, 并根据测试集分类准确率动态减小学习率大小. 在模型训练阶段, 图片批量大小为 256; 而在对抗样本生成阶段, 图片批量大小为 150. 我们使用不同的对抗扰动大小, 有目标攻击及无目标攻击对 Trap-Net 的对抗防御效力进行全面的测试. 在基于梯度的对抗攻击方法中, 我们设置对抗扰动大小为 0.1–0.7. 在基于优化的对抗攻击方法 C&W 中, 我们设置代表攻击力度的初始化常数 c 为 1, 10; 并使用二分查找的方式动态变化 c 以提高攻击性. 在基于生成式对抗网络的对抗攻击方法 AdvGAN 中, 我们设置扰动大小为 0.1, 0.3.

本文选择使用 ResNet 残差网络和 VGG16 作为目标 DNN 进行 Trap-Net 对抗防御有效性验证. 其中, 在针对 RQ1 的实验中将使用 ResNet-18, ResNet-34 和 ResNet-50 进行陷阱式网络特征提取多样性的说明. 在与频谱对抗防御的对比实验中将使用 VGG16. 而在其余实验中, 将使用 ResNet-50 为主体模型进行实验.

所有实验均使用图形处理器 NVIDIA GeForce RTX 2070 Super 进行处理, 其中 CUDA 版本号为 11.4, 内存大小为 24 GB.

3.3 结果分析

● RQ1: 陷阱式网络能否作为一种简单, 有效的集成网络中子网络的扩充方式以提高集成多样性?

为了评估陷阱式网络能否作为一种有效的对抗集成网络中子网络的扩充方式, 本文通过对比 Strauss 等人[35]提出的集成防御网络构架方式, 验证陷阱式网络作为对抗集成网络中子网络扩增方式的有效性. 实验以 MNIST 为目标数据集, 以 ResNet-50 残差网络为基准原始模型. 使用传统集成网络构成方式构建 $M1$ 和 $M2$ 集成网络. 其中, $M1$ 集成 3 个不同初始化参数的 ResNet-50 残差网络; 集成不同结构网络 ResNet-18, ResNet-34 和 ResNet-50 为集成网络 $M2$. 为了验证假设, 以 CIFAR-10 作为陷阱数据构成的输出类别数为 20 的陷阱式网络 $M3$ 并以输出类别为 16, 18 和 20 的陷阱式集成网络 $M4$. 实验使用基于梯度的对抗攻击方法 FGSM 和 PGD 进行对抗防御效力验证, 同时为了验证陷阱式网络是否会影响相对于传统高斯噪声的鲁棒性, 实验设置不同扰动幅度的高斯噪声对模型的传统鲁棒性进行验证. 最终通过对比 $M1$, $M2$, $M3$ 和 $M4$ 的实验结果以评估陷阱式网络的性能表现.

表 2 展示不同网络模型在面对不同攻击方法及不同大小的对抗扰动攻击时的分类准确率. 其中, 第 1 列为对抗攻击方法, 第 2 列显示扰动参数, 第 3 列为原始模型 ResNet-50 的分类准确率. 第 4–7 列为 $M1$–$M4$ 模型的分类准确率. 通过表 2 可知, 在不同幅度的高斯噪声攻击场景下, 集成类网络 $M1$, $M2$ 和 $M4$ 的传统鲁棒性皆有一定的提升, 这是集成网络在该方面天然的优势. 值得注意的是, 陷阱式网络 $M3$ 相较于基准原始模型, 其探测前后模型的准确率变化体现出被陷阱类标记的靶标可攻击空间有利于提升面对传统噪声的鲁棒性. 在不同扰动的白盒攻击场景下, $M3$ 在低扰动 (ε=0.3) 条件下表现出与 $M1$, $M2$ 近似的对抗防御效力. 而在高扰动 (ε=0.6) 条件下, $M3$ 的防御效力明显优于后者. 且通过分析 $M3$ 和 $M4$ 的实验结果可以看出, $M4$ 的干净目标分类准确率 99.50% 优于传统集成对抗防御网络 99.38%, 同时集成后的 $M4$ 在探测逻辑判断后的准确率比 $M3$ 整体提升了 15% 左右. 这表明 Trap-Net 通过集成学习的方式能在保证原目标数据的分类精度的同时, 扩大靶标可攻击空间以提高探测效力.

表 2 Trap-Net 对集成对抗网络进行子网络扩充后与 $M1$ 和 $M2$ 防御效力对比 (%)

攻击方法	参数	原始模型	$M1$	$M2$	$M3$		$M4$	
					探测前	探测后	探测前	探测后
无	—	98.86	99.38	99.25	98.04	98.04	**99.50**	**99.50**
高斯噪声	$\sigma=0.2$	96.75	98.49	99.00	96.58	97.69	99.19	**99.42**
高斯噪声	$\sigma=0.8$	53.81	75.39	86.66	52.62	74.86	77.15	**90.67**
FGSM	$\varepsilon=0.3$	25.56	69.27	62.50	32.16	64.08	47.33	**79.26**
FGSM	$\varepsilon=0.6$	8.47	33.64	12.56	2.12	73.56	3.64	**92.65**
PGD	$\varepsilon=0.3$	2.76	0.79	30.21	1.56	41.23	6.53	**65.73**
PGD	$\varepsilon=0.6$	0.51	0.00	1.51	0.00	80.50	3.12	**93.64**

小结: (1) 子网络的集成多样性对对抗集成网络防御方法的防御效力至关重要, 提升子网络的集成多样性可进一步提升对抗集成网络的鲁棒性. (2) 陷阱式网络有利于提升针对高斯噪声等传统噪声的鲁棒性. (3) 不同陷阱数据类别所构成的陷阱式网络与不同结构的 DNN 具有相似的集成多样性, 陷阱式网络可作为一种简单有效的对抗集成网络中子网络的扩增方式.

● RQ2: Trap-Net 能否对对抗样本的迁移性进行有效的防御?

为了评估 Trap-Net 能否对对抗样本的迁移性进行有效防御, 设计如下的对抗样本黑盒攻击实验. 我们使用 ResNet-50 作为黑盒攻击场景下对抗攻击的代理模型. 被测试的 Trap-Net 由以 CIFAR-10 为陷阱数据, 输出类别个数为 16, 18, 20 的陷阱式网络所构成. 通过在代理模型 ResNet-50 上使用基于梯度的对抗攻击方法 FGSM, PGD 方法和迁移攻击性更强的基于生成式对抗网络 AdvGAN 于不同大小的对抗扰动下进行有目标攻击和无目标攻击生成对抗样本, 并将在代理模型上生成的对抗样本输入目标 Trap-Net 网络模型进行对抗样本的迁移性防御测试.

Trap-Net 的黑盒防御效力如表 3 所示, 第 1 列为对抗攻击方法, 第 2 列为关键的超参数设置, 第 3 列为代理模型 ResNet-50 的分类准确率, 第 4 列为 Trap-Net 的探测前后分类准确率. 其中在探测逻辑判断前准确率表明, 黑盒

攻击的对抗样本并不隶属于目标模型的目标数据流形之内, 即能有效攻击传统网络模型. 而在探测逻辑判断后, Trap-Net 的准确率近乎 100% 表明 Trap-Net 对可攻击空间的陷阱类靶标定义能有效地对对抗样本的黑盒攻击进行防御.

表 3 Trap-Net 黑盒防御效力 (%)

攻击方法	参数	有目标攻击	原始模型	Trap-Net	
				探测前	探测后
无	—	—	99.13	99.46	99.46
FGSM	$\varepsilon = 0.3$	否	0.46	11.36	100
FGSM	$\varepsilon = 0.3$	是	0.00	7.89	100
PGD	$\varepsilon = 0.3$	否	0.00	38.56	99.84
PGD	$\varepsilon = 0.3$	是	0.46	36.38	100
AdvGAN	$\varepsilon = 0.3$	否	0.29	0.98	100

小结: (1) 对抗样本的迁移性极大增强了对抗样本于现实世界中的攻击性, 因此设计能有效抵御对抗样本迁移性攻击, 即黑盒攻击下生成的对抗样本具有重要的实际意义. (2) Trap-Net 对可攻击空间的陷阱类靶标定义, 使得其具有极强的对抗防御泛化性. 实验表明, Trap-Net 能有效抵御对抗样本的迁移性攻击.

● RQ3: Trap-Net 的相关参数对防御方法的防御效力影响如何?

为了验证 Trap-Net 的相关参数对防御方法的防御效力影响, 从陷阱数据类别个数以及陷阱数据集类别对防御效力的影响这一问题进行实验与分析. 针对陷阱数据类别个数对防御效力的影响这一问题, 我们通过训练不同陷阱数据类型个数的陷阱式网络以探寻陷阱数据个数对陷阱式集成网络的防御效力的影响, 并对其在 FGSM 和 PGD 的有目标以及无目标攻击场景下进行检测. 针对陷阱数据类别对防御效力影响这一问题, 我们首先对陷阱数据与目标数据的相似性进行计算, 以 K-MNIST, F-MNIST, CIFAR-10 和 CIFAR-100 为陷阱数据集进行陷阱式网络的训练测试. 对于计算陷阱数据集与目标数据集相似度的大小这一问题, 首先通过逐个遍历的方式使用 Wasserstein 距离度量方式[36]对不同陷阱数据集与目标数据集最后一层全连接层的输出向量进行距离度量. Wasserstein 距离是一种可以在数据分布不相交情况下仍能进行有效度量的距离度量方式. 我们使用公式 (11) 表示不同数据集的相似度:

$$S_{(\text{true,trap})} = \left(1 - \frac{W_{(\text{true,trap})} - W_{(\text{true,true})}}{W_{\max} - W_{\min}}\right) \cdot 100\% \tag{11}$$

其中, $W_{\max}$ 和 $W_{\min}$ 指 K-MNIST, F-MNIST, CIFAR-10 和 CIFAR-100 中与 MNIST 距离最远和最近的 Wasserstein 距离. $W_{(\text{true,true})}$ 指 MNIST 数据流形内部的 Wasserstein 距离. $W_{(\text{true,trap})}$ 指当前被测陷阱数据集与 MNIST 数据流形之间的 Wasserstein 距离. $S_{(\text{true,trap})}$ 表示当前被测陷阱数据集与 MNIST 数据集之间的相似度. $S_{(\text{true,trap})}$ 值越大, 表示两个数据集越相似.

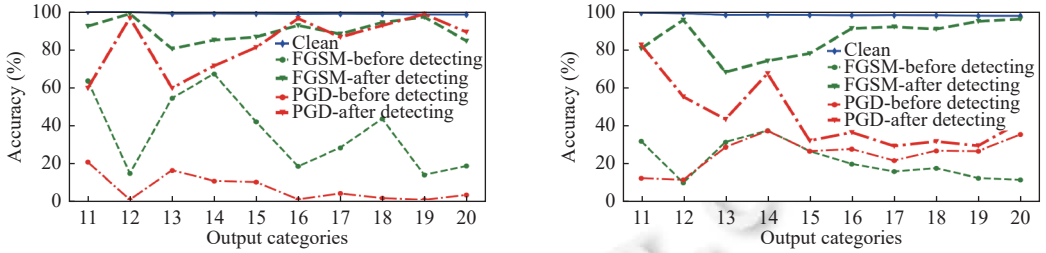
其次, 分别使用 K-MNIST, F-MNISTCIFAR-10 和 CIFAR-100 作为陷阱数据集, 以 MNIST 为目标数据集进行 Trap-Net 对抗防御效力实验及分析.

结果分析如下.

RQ3.1: 陷阱数据类别个数对防御效力的影响如何?

图 7 展示了无目标和有目标攻击场景下陷阱数据类别个数对 Trap-Net 防御效力的影响. 其中, 点划线代表 PGD 攻击, 虚线代表 FGSM 攻击. 标记点为圆形代表探测前陷阱式网络准确率. 标记点为三角形代表探测后 Trap-Net 准确率. 在图 7(a) 所示的无目标攻击场景下, 随着陷阱数据个数类别的增加, 探测前准确率和探测后准确率之间的差值不断增加, 这体现出 Trap-Net 的对抗样本探测防御效力逐步增强. 然而在图 7(b) 所示的迭代式有目标攻击场景下, 低陷阱数据类别个数的陷阱式网络却体现出更强的防御效力. 这是因为低陷阱数据类别时, DNN 内部特征空间不易形成陷阱数据流形, 所以陷阱数据于特征中能更好地插入邻近可攻击空间. 而当陷

阱数据类别增多时, DNN 内部特征空间会形成特定的陷阱数据集流形, 同时因为陷阱式平滑损失函数的效用, 此时的陷阱数据更偏向于定义目标数据流形之外的背景可攻击空间, 从而导致防御效力的相对下降. 综上, 低陷阱数据类型个数的陷阱式网络和高陷阱数据类型个数的陷阱式网络偏向于定义不同的可攻击空间, 所以在进行网络构建时, 应同时考虑低陷阱数据类型的陷阱式网络和高陷阱数据类型的陷阱式网络, 以提高 Trap-Net 的防御效力和泛化能力.



(a) 无目标攻击 (b) 有目标攻击

图 7 陷阱数据类别个数对 Trap-Net 防御效力影响

RQ3.2: 陷阱数据集与目标数据间相似性大小对防御效力的影响如何?

表 4 展示陷阱数据集与目标数据集的相似度, 其中 K-MNIST 与 MNIST 的相似度最高, 为 92.74%. 而 CIFAR-100 与 MNIST 相似度最低, 为 0.

表 4 陷阱数据集与目标数据集相似度 (%)

相似度	MNIST	K-MNIST	F-MNIST	CIFAR-10	CIFAR-100
$S_{(\text{true,trap})}$	100	92.74	59.35	4.29	0

本节展示单一数据集作为陷阱数据进行 Trap-Net 的构建时, 使用不同扰动大小的 FGSM, PGD, C&W 和 AdvGAN 对抗攻击方法在有目标和无目标白盒攻击场景下对各种陷阱数据集于 Trap-Net 的防御效力的关系进行测试. 表 5 展示了实验结果, 其中第 3 列表示对抗攻击是否设置目标攻击类别, 且有目标攻击场景下, 固定设置目标攻击类别设置为 2. 第 4 列 M_{initial} 代表 ResNet-50 残差网络的分类准确率, 第 5–8 列为不同陷阱数据所训练的 Trap-Net 在面对不同对抗攻击时的准确率. $\{M_{\text{mk}},\ldots,M_{\text{mc}}\}$ 代表以 D2, D3, D5 和 D4 为陷阱数据集构建的 Trap-Net. 相较于 M_{initial} , 4 种 Trap-Net 皆不影响原始分类精度. 通过对比实验结果, 陷阱数据与目标数据相似度较高的 M_{mf} 和相似度较低的 M_{mc} 陷阱式集成防御网络的防御效力最好. 这表明不同相似度的陷阱数据带来的防御效力是不同的. 据此我们又将不同的陷阱数据集进行组合实验, 实验结果表明, 多种不同陷阱数据集组合的陷阱式集成网络的防御效力并没有明显的提升. 虽然 M_{mfc} 的防御效力 (MNIST 为目标数据集, F-MNIST 和 CIFAR-10 为陷阱式数据集) 在多个攻击算法下略微优于 M_{mf} 和 M_{mc}. 但在陷阱数据类别数目固定的情况下, 随着陷阱数据集类别的增多, 增添不同的陷阱数据集类别反而会因为不同陷阱数据集类别较少, 无法形成其各自的陷阱数据集流形, 从而造成防御效力的降低.

值得注意的是, Trap-Net 根据对抗扰动的大小, 其防御效力呈“V”型结构. 特别是当陷阱式集成网络遭遇有目标攻击时, 这种“V”型尤为明显. 这在一定程度上验证了本文所提出的可攻击空间对抗成因假设. 直观上, 造成这种现象的原因是邻近可攻击空间需在考虑不影响原始目标分类精度的前提下进行陷阱类靶标定义, 因此无法以在背景可攻击空间内的同等标记效率进行标记. 当对抗扰动较小时, Trap-Net 依靠集成各子网络的平滑特性以保持模型整体的鲁棒性. 而当对抗扰动较大时, 对抗样本被诱导到靶标可攻击空间从而被 Trap-Net 成功探测. 而当对抗扰动位于“V”型拐点时, DNN 本身的平滑鲁棒性以及陷阱类别标记的邻近可攻击空间因防御机制不同, 造成各自防御效力一定程度上的降低. 所以如何兼容两种防御特性以及进一步优化邻近可攻击空间的标记是我们未来重要的研究点.

表 5 陷阱数据与目标数据间相似度大小对防御效力的影响 (%)

攻击方法	参数	有目标攻击	$M_{initial}$	M_{mk}		M_{mf}		M_{mc_100}		M_{mc}	
				探测前	探测后	探测前	探测后	探测前	探测后	探测前	探测后
无	—	—	98.24	98.89		99.67		99.28		99.44	
FGSM	$\varepsilon=0.1$	否	76.56	67.44	97.00	86.45	95.33	32.24	98.58	73.89	**98.89**
FGSM	$\varepsilon=0.3$	否	48.67	2.62	99.49	13.64	97.95	1.26	99.78	0.26	**99.85**
FGSM	$\varepsilon=0.1$	是	84.51	74.92	97.52	92.25	96.51	82.34	96.25	81.58	**98.33**
FGSM	$\varepsilon=0.5$	是	39.48	2.15	100	4.07	98.59	20.68	**97.34**	6.02	**100**
PGD	$\varepsilon=0.1$	否	10.43	27.00	96.22	45.89	92.22	4.34	98.56	10.56	**98.22**
PGD	$\varepsilon=0.3$	否	3.33	0.27	99.48	0.21	99.23	0.00	**100**	0.10	99.85
PGD	$\varepsilon=0.1$	是	40.67	90.76	97.53	94.51	96.08	85.35	96.48	83.92	**97.50**
PGD	$\varepsilon=0.5$	是	11.70	32.59	55.78	32.52	69.63	43.74	68.56	28.37	**77.04**
C&W	$C=1$	否	0.00	90.68	96.67	93.34	93.34	76.67	100	96.28	**100**
C&W	$C=10$	否	0.00	33.67	83.52	26.67	90.26	0.00	100	20.25	**100**
C&W	$C=1$	是	28.33	98.37	98.37	96.67	96.67	98.54	98.54	99.24	**99.46**
C&W	$C=10$	是	16.78	33.34	33.34	39.56	39.56	29.32	29.32	63.34	**68.43**
AdvGAN	$\varepsilon=0.1$	否	71.88	30.48	99.18	45.32	99.45	46.54	99.89	32.66	**99.64**
AdvGAN	$\varepsilon=0.3$	否	32.81	0.00	99.26	0.00	98.44	0.00	**100**	0.00	**100**

小结: (1) 陷阱数据类型个数与 Trap-Net 的集成多样性相关, 在构建 Trap-Net 时, 应同时考虑集成低类别和高类别个数的陷阱式网络, 以提高 Trap-Net 的对抗防御效力和泛化能力. (2) Trap-Net 的对抗防御效力与陷阱数据与目标数据的相似度相关, 陷阱数据集在选取时应同时兼顾与目标数据类别相似度高的陷阱数据以及相似度低的陷阱数据, 且在陷阱类别阈值个数较少时, 不可选择过多的不同类别数据集作为陷阱数据集.

● RQ4: 陷阱式平滑损失函数能否优化目标数据与陷阱数据之间的数据分布, 从而进一步提高 Trap-Net 的对抗防御效力?

为了验证陷阱式平滑损失函数对 Trap-Net 的对抗防御效力影响, 我们针对以下 3 个问题展开研究. 首先, 针对陷阱类别数据在陷阱式平滑损失函数的影响下对外部背景可攻击空间的标记这一问题. 本文通过向 MNIST 数据集以逐步递增的方式添加不同大小的扰动, 从目标数据流形出发, 逐渐穿越敏感特征空间, 通过观察最后一层全连接层所输出的逻辑向量以分析 Trap-Net 对特征空间的解析与标记. 其次, 针对陷阱式平滑损失函数对 Trap-Net 的对抗防御效力是否有优化这一问题, 本文通过在相同的 Trap-Net 模型结构下分别使用 CE 和陷阱式平滑损失函数作为损失函数训练 Trap-Net, 并在相同的对抗样本生成方式下进行对抗防御效力测试. 通过实验对比以验证陷阱式平滑损失函数对 Trap-Net 对抗防御效力的有效性. 最后, 针对陷阱式诱导因子大小对对抗防御效力的影响这一问题, 本文使用相同结构的 Trap-Net 网络模型, 利用不同大小的陷阱式诱导因子进行训练. 并通过在有目标和无目标攻击场景下使用相同的对抗样本生成方法进行攻击, 以评估不同大小的陷阱式诱导因子在不同的攻击场景下对 Trap-Net 防御效力的影响.

在对陷阱类别数据在陷阱式平滑损失函数的影响下对外部背景可攻击空间的标记进行实验探究中, 首先我们使用高斯随机噪声对模型的鲁棒性进行探究, 当对抗扰动 $\sigma>0.25$ 时, 高斯随机噪声可以使得干净样本发生错误. 而当 σ 足够大时, 所有样本的输出皆为标签 2. 这证明了 DNN 对于其所不敏感的特征空间以泛化的方式所标注. 图 8 展示了在无目标攻击场景下, 使用基于梯度式的攻击算法进行无扰动大小限制攻击后, 不同大小的对抗扰动下各目标类别于最后一层全连接层的输出的向量大小. 其中数据的正确类别为以点线表示的标签 2, 以点横线表示的为陷阱类标签 16. 通过更换不同的扰动阈值进行实验发现, DNN 只有在很小的扰动空间中, 代表正确类别的向量值为最大. 当扰动逐渐增大, 点划线所代表的类别 16 的向量值为最大. 这表明当扰动较大, 数据偏移出目标数据流形所处的特征空间后, DNN 因为缺乏外部空间的特征定义而将外部特征空间全部标记为类别 16. 即陷阱类别 16 已通过定义外部特征敏感空间和垃圾背景可攻击空间的方式将目标数据流形所在的特征敏感空间包围.

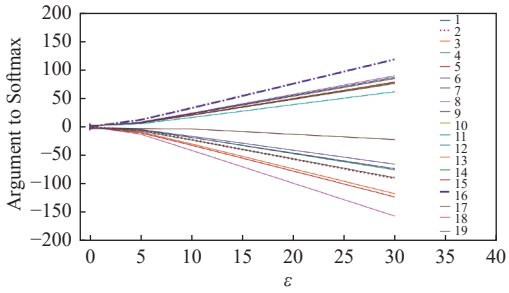


图 8　不同大小扰动下各数据类别的 DNN 最终逻辑输出

针对陷阱式平滑损失函数对 Trap-Net 的对抗防御效力是否有优化这一问题, 表 6 展示了 Trap-Net 分别使用 CE 和陷阱式标签平滑作为损失函数后, 面对不同的对抗攻击方法所表现出的对抗防御效力. 其中, *M*4 以 *CE* 为损失函数, 以 CIFAR-10 为陷阱数据集的输出类别为 16, 18 和 20 的 Trap-Net. *M*5 以陷阱式平滑为损失函数, 其模型结构与 *M*4 相同.

表 6　陷阱式平滑对抗防御效力的影响 (%)

攻击方法	参数	原始模型	*M*4		*M*5	
			探测前	探测后	探测前	探测后
无	—	98.97	99.24	99.24	99.32	99.32
FGSM	$\varepsilon=0.3$	25.56	48.26	80.77	**0.84**	**98.75**
FGSM	$\varepsilon=0.6$	8.47	3.22	92.45	**0.00**	**100**
PGD	$\varepsilon=0.3$	2.76	5.89	69.53	**0.12**	**99.59**
PGD	$\varepsilon=0.6$	0.51	2.59	94.27	**0.00**	**100**
C&W	$C=1$	0.00	0.00	3.33	**92.34**	**96.68**

从实验结果分析可知, *M*5 其探测前准确率明显低于 *M*4 , 如在 ε=0.3 的 FGSM 攻击场景下, *M*4 的探测前准确率为 48.26%, 而 *M*5 为 0.84%. 这表明陷阱式平滑使得陷阱式网络更易被“攻击”. 然而二者的探测后准确率对比可知, *M*5 的对抗样本探测效力大大增强, 在各种扰动大小的攻击场景下, 都有优异的防御效力表现. 这体现出陷阱式平滑对对抗样本的诱导作用及双赢思想下防御效力的增强. 值得注意的是, *M*4 无法有效防御 C&W 攻击, 而添加了陷阱式平滑损失函数的 *M*5 可以对 C&W 攻击有效防御. 这表明陷阱式平滑可极大程度上影响基于优化的对抗攻击方法针对对抗扰动的搜索.

针对陷阱式诱导因子大小对对抗防御效力的影响这一问题, 本文利用 FGSM 和 PGD 攻击算法, 通过选择不同大小的陷阱诱导因子以探测陷阱式集成网络在有目标和无目标攻击之下的防御效力.

如图 9 所示, 在 PGD 有目标攻击场景下, 当陷阱式诱导因子 $\alpha<0.2$ 时, 陷阱式集成防御效力低下. 这是因为陷阱数据穿插于目标数据流形之间的邻近可攻击空间的难度大于定义背景可攻击空间. 同时从对干净数据的分类精度可知, 随着 α 的增大, 干净数据集的分类精度将会缓慢下降. 具体地, 在 MNIST 数据集的分类场景下, 当 $\alpha<0.6$ 时, 陷阱式平滑对原分类精度的下降影响不超过 0.2%. 当 $\alpha>0.8$ 时, 陷阱式平滑对原分类精度的下降影响为 1% 左右. 通过后续实验我们发现, 在以 CIFAR-10 为主体, CIFAR-100 为陷阱数据集的 Trap-Net 实验中, 当 $\alpha>0.6$ 时, Trap-Net 将会变得不稳定. 综上, Trap-Net 对 α 的选取应在考虑陷阱数据对原目标数据流形的靶标可攻击空间定义效力的同时, 考虑 α 对原目标数据分类精度的影响. 我们建议 α 的选取区间为 [0.35, 0.6]. 后续的实验中将随机选取 $\alpha=0.4$ 进行 Trap-Net 的相关实验. 同时 Trap-Net 的防御效力与 α 未成单调性关系反映出通过在同一网络结构中使用不同大小 α 进行陷阱式网络的训练也是一种 Trap-Net 靶标可攻击空间的扩大方式.

小结: (1) 损失函数是评估 DNN 于具体事务环境下性能表现的关键指标, 完善优化损失函数对于提升 DNN 的分类性能及鲁棒性具有重要意义. (2) 陷阱式平滑损失函数能优化陷阱数据类别与目标数据类别之间的诱导关

系, 促使对抗扰动指向陷阱类别标记的靶标可攻击空间, 从而进一步提升了 Trap-Net 的对抗防御效力. (3) 陷阱式诱导因子应在不影响目标分类精度的基础上选取最大值.

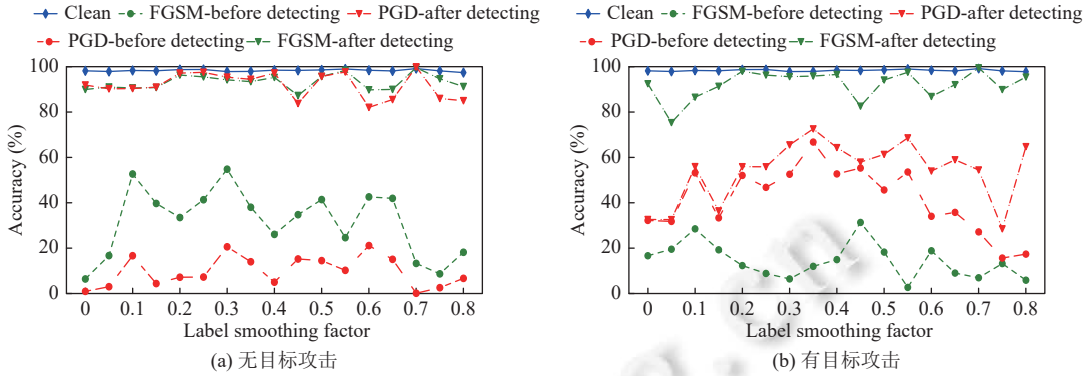


图 9 陷阱式诱导因子大小对防御效力影响

● RQ5: Trap-Net 相较于其他类似的对抗防御方法的防御效力如何?

为了进一步证明 Trap-Net 的对抗防御效力, 我们在相同的对抗攻击场景下, 对 Trap-Net 与核密度和贝叶斯不确定性估计法, 特征压缩对抗防御方法和频谱对抗防御方法进行防御效力的对比. 在与核密度和贝叶斯不确定性估计法的对比实验中, 我们以 Feinman 等人[26]的实验方式进行对抗样本和干净样本的探测比对. 将生成的对抗样本和干净样本各自标记为独立的一类, 使用两种方法对其进行对抗样本的探测实验. 同时根据原实验评价标准, 使用 AUC 值对模型的性能进行评估. 在与特征压缩的对比实验中, 我们以 7% 的误诊率计算阈值, 并使用 FGSM, PGD 和 C&W 在不同扰动大小及有目标和无目标攻击场景下进行对抗防御方法的防御效力评估. 在与频谱对抗防御方法的对比实验中, 我们使用 VGG16 网络模型, 以 CIFAR-10 为目标主体, CIFAR-100 为陷阱数据构建 Trap-Net. 将其与频谱对抗防御方法中所提出的高量傅里叶频谱防御方法 (MFS) 和阶段性傅里叶频谱防御方法 (PFS) 在不同扰动大小的 FGSM, BIM, PGD 和 DeepFool 对抗攻击方法下进行模型对抗防御效力的评估.

在与核密度和贝叶斯不确定性估计法的对比实验中, 实验结果如图 10 所示的 ROC 曲线图. 其中, 实线为核密度和贝叶斯不确定性探测法的 ROC 曲线. 点线为 Trap-Net 的 ROC 曲线. 由图可知, 点线所代表 Trap-Net 的 AUC 值优于实线所代表的核密度和贝叶斯不确定性估测法. 这表示 Trap-Net 相较于核密度和贝叶斯不确定性估测法能更好地分辨对抗样本与干净样本. 同时因为 Trap-Net 不依赖于生成的对抗样本所提供的信息, 所以 Trap-Net 有更强的防御泛化性.

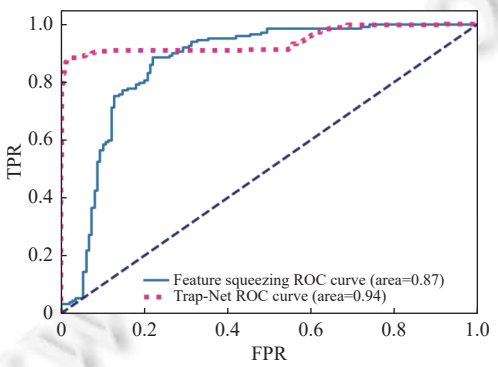


图 10 Trap-Net 与核密度和贝叶斯不确定性估计法防御效力对比图

在与特征压缩的对比实验中, 我们使用作者提供的最佳参数即 1-bit 色位压缩以及 2×2 卷积核进行中值平滑. 二者的对比实验结果如表 7 所示. 从表 7 中 Trap-Net 与特征压缩对抗防御方法的对比可看出, 在面对低扰动对抗

样本时, 二者的准确率皆可到达 90% 以上. 然而因为 Trap-Net 的集成特性, Trap-Net 能在不损失, 甚至提高原目标数据分类精度的基础上进行对抗样本的防御. 相比之下, 特征压缩对抗防御方法需要牺牲部分目标分类精度, 且因为其防御有效性取决于阈值的选取, 其探测结果需以较高的误诊率为代价所换取. 在面对高扰动攻击时, 特征压缩后的对抗样本仍以高置信度指向错误分类, 因此防御失效.

表 7 Trap-Net 与特征压缩对抗防御效力对比 (%)

攻击方法	参数	有目标攻击	原始模型	Feature squeezing		Trap-Net	
				探测前	探测后	探测前	探测后
无	—	—	99.13	96.73	100	**99.46**	**99.46**
FGSM	$\varepsilon = 0.3$	否	1.68	78.15	89.38	6.19	99.44
FGSM	$\varepsilon = 0.3$	是	2.86	86.34	92.37	6.63	99.64
FGSM	$\varepsilon = 0.7$	否	0.00	0.02	16.37	0.09	99.82
FGSM	$\varepsilon = 0.7$	是	0.72	12.34	34.51	0.37	99.98
PGD	$\varepsilon = 0.3$	否	0.00	80.34	93.67	0.13	99.94
PGD	$\varepsilon = 0.3$	是	0.56	83.48	94.48	31.58	86.27
PGD	$\varepsilon = 0.7$	否	0.00	0.00	0.00	0.00	100
PGD	$\varepsilon = 0.7$	是	0.00	10.34	13.48	11.93	97.37
C&W	$C = 1$	否	0.00	81.37	92.34	90.12	98.64
C&W	$C = 1$	是	3.46	78.63	79.34	97.10	97.10

在与频谱对抗防御方法的对比实验中, 我们使用了 Harder 等人[31]所提供的对抗样本生成代码以及相关的评判标准, 并使用 VGG16 作为原始模型. 以 CIFAR-10 作为目标数据集, 使用 FGSM 攻击方法, 以对抗扰动为 $\varepsilon = 0.03$ 的对抗样本进行频谱信息的提取, 并利用该频谱信息进行探测对抗样本的回归预测模型的训练. Trap-Net 使用经过挑选的 CIFAR-100 中的 10 大超类为陷阱数据集构建集成网络. 实验对比的整体实验结果如表 8 所示. 特别的, 因为其在评判模型防御力有效性时使用的为挑选后攻击成功的对抗样本, 所以表 8 的第 2 列中原始模型的预测准确率为 0. 同时我们预设基于梯度的对抗攻击方法 FGSM, BIM 和 PGD 的对抗扰动集合为 $\varepsilon = [0.003, 0.3, 0.7]$.

表 8 Trap-Net 与频谱对抗防御效力对比 (%)

攻击方法	原始模型	MFS	PFS	Trap-Net	
				探测前	探测后
无	92.13	92.13	92.13	92.64	92.64
FGSM	0.0	**98.14**	97.00	12.64	97.84
BIM	0.0	93.58	**98.06**	2.58	97.25
PGD	0.0	70.43	96.9	1.86	**97.12**
DeepFool	0.0	58.06	86.1	1.64	**96.48**

如表 8 所示, 与频谱对抗防御方法相比, Trap-Net 在同类型对抗攻击的检测防御方面略有不足. 频谱对抗防御在如 FGSM, BIM 和 PGD 等同类型的对抗样本检测方面拥有令人赞叹的性能以及相应的黑盒防御鲁棒性. 然而在面对如基于优化的对抗样本生成方法 DeepFool 等对抗样本的攻击时, 频谱对抗防御的对抗效力则会在一定程度上降低. 这在一定程度上表现出了频谱对抗防御方法对现有对抗样本所提供的数据信息的强依赖. 而 Trap-Net 则能很好地避免这一点, Trap-Net 在面对不同类别的对抗样本攻击时, 依旧能保持良好的对抗防御能力.

小结: (1) Trap-Net 相较于核密度和贝叶斯不确定性估计法具有更强的对抗样本探测能力, 具有更强的对抗防御泛化能力, 同时因为 Trap-Net 不依赖于生成的对抗样本所提供的信息, Trap-Net 相较于频谱对抗防御方法, 在面对不同类别的对抗攻击时, 具有更稳定的对抗防御能力. (2) 在低扰动的对抗攻击场景下, Trap-Net 相较于特征压缩对抗防御方法具有等同的对抗样本探测能力, 而在较大的对抗扰动攻击场景下, 特征压缩方法无法对损失的数

据特征进行修复, 进而无法通过阈值变化进行对抗样本的检测. 同时, 相较于特征压缩对抗防御方法, Trap-Net 无需损失原始目标数据的分类精度, 是一种鲁棒性更强的对抗防御方法.

● RQ6: Trap-Net 与其他对抗防御方法的兼容性如何?

为了进一步优化 Trap-Net 在对抗防御效力上的 V 型缺陷, 我们对 Trap-Net 与主流的对抗防御方法即对抗训练进行了兼容性测试. Trap-Net 的防御的核心是对 DNN 特征空间中未定义的可攻击空间予以陷阱类别标记, 从而通过针对对抗样本的探测有效识别对抗样本. 而作为主流的对抗防御方法, 对抗训练的核心思路是通过学习对抗样本的分布, 从而迭代式地优化 DNN 于特征空间中的数据分布, 进而提高模型的鲁棒性. 针对 Trap-Net 在有目标攻击场景中, 防御效力呈现 V 型结构. 我们希望通过对抗训练, 对邻近可攻击区域, 即目标数据流形周围的未定义可攻击区域进行优化. 我们收集成功攻破 Trap-Net 对抗防御, 且最终分类类别为有效目标类别的对抗样本进行 PGD 对抗训练, 以期在不影响 Trap-Net 防御机制的前提下, 通过对抗训练进一步优化敏感特征空间中的目标数据流形. 从而弥补 Trap-Net 在有目标攻击场景下面对某些扰动区间的对抗扰动攻击时, 对抗防御效力较差的情况.

表 9 展示了 Trap-Net 在以标签 2 为目标攻击类别的有目标攻击场景下, Trap-Net 与 PGD 对抗训练的对抗防御兼容性测试. 其中第 4 列为单一 Trap-Net 防御的分类准确率, 第 5 列为 Trap-Net 经过 PGD 对抗训练后的分类准确率. 分析表 9 可知, 对抗训练后的 Trap-Net 探测前及探测后的准确率基本一致, 这表示无论依次对单个 Trap-Net 子网络进行对抗训练或是直接对整个 Trap-Net 进行对抗训练, 其优化结果都会破坏原目标数据集和陷阱数据集基于陷阱式平滑建立的陷阱机制. 虽然与对抗训练兼容后的 Trap-Net 提升了低扰动有目标攻击场景下的对抗防御鲁棒性, 但对抗训练会破坏陷阱式集成网络的诱导和探测机制, 使得其在高扰动条件下的探测效力降低. 我们未来将尝试更多的对抗训练策略进行邻近可攻击空间定义的优化, 这也是我们未来的研究方向之一.

表 9 Trap-Net 与对抗训练兼容前后的防御效力对比 (%)

攻击方法	参数	有目标攻击	对抗训练前		对抗训练后	
			探测前	探测后	探测前	探测后
PGD	ε=0.1	是	10.32	99.12	94.69	94.69
PGD	ε=0.3	是	47.56	87.21	92.21	92.34
PGD	ε=0.5	是	32.51	80.32	90.48	91.48
PGD	ε=0.8	是	0.00	100	76.34	84.46

小结: (1) 由于现有的主流对抗样本防御方法大多为启发式方法. 所以如何有效地对各种对抗防御方法进行兼容具有重要的研究意义. (2) 对抗训练可解决 Trap-Net 于低扰动有目标场景下防御鲁棒性不足的缺点, 但无法在不影响 Trap-Net 探测防御机制的同时优化邻近可攻击空间中的目标数据流形.

4 相关工作

4.1 对抗样本成因探究

自 2014 年 Szegedy 等人[4]通过 L-BFGS 对抗攻击方法生成对抗样本, 揭露了对抗样本这一 DNN 鲁棒性安全盲点以来, 对抗样本的成因至今仍是一个开放性问题. 对抗成因对对抗攻击和对抗防御方法的提出和优化有指导作用. 现有的对抗成因[4−7]大多基于流形学习, 聚焦于目标数据流形的低概率区域欠拟合, 高维线性, 高维几何结构等. Szegedy 等人[4]认为训练数据不足导致 DNN 只能学习到目标数据流形的局部区域, 对抗样本所存在的数据流形的低概率区域未被模型正确划分是对抗样本存在的主要原因; Goodfellow 等人[5]认为 DNN 的脆弱性是由于模型在高维存在局部线性特征所导致; Gilmer 等人[6,7]认为 DNN 模型易受对抗样本攻击的主要原因在于目标数据流形的复杂高维几何结构. 本文基于流形学习思想, 聚焦于 DNN 特征空间, 通过解析目标数据流形与 DNN 特征空间之间的关系, 提出了可攻击空间对抗成因假设.

4.2 对抗集成网络防御

对抗集成网络防御旨在通过集成学习的方式, 集成多个子网络的特征信息以生成鲁棒性更强的集成网络. 其

中, Strauss 等人[35]证明了通过集成多个不同初始化参数的 DNN, 或集成多个不同网络结构的 DNN, 可有效提升原目标数据的测试精度并提升 DNN 的鲁棒性. Abbasi 等人[37]发现数据的不同类别在对抗攻击中错误分类为其他类别的概率不同, 于是通过利用 FGSM 计算出网络模型对应的混淆矩阵, 并根据该混淆矩阵进行数据集的分类. 通过在不同的数据集训练 DNN 并最终组成鲁棒的集成网络. 结果表明, 这种方法可以对输入样本进行干净样本和对抗样本的区分, 从输入端提升网络模型的鲁棒性. Pang 等人[27]通过定义集成网络中子网络的多样性, 创建了一种类似于标签平滑的集成网络训练方式. 该方法通过衡量并提高各个子网络中非正确类别输出向量的正交程度, 以衡量并提高各个集成网络中子网络所提取和关注的特征信息的多样性, 最终通过提升集成网络整体所学习的特征信息以提升整体的鲁棒性. 然而, 对抗集成网络防御具有很大的防御效力上限. 本文提出的 Trap-Net 防御方法在不损失原目标分类精度的同时, 以对抗集成网络防御的思想, 通过扩大被陷阱类别定义的靶标可攻击空间, 从而进一步提高对抗样本的防御效力.

5 总结与展望

本文基于 DNN 的特征空间, 提出可攻击空间对抗成因假设. 可攻击空间是 DNN 特征空间中有别于目标数据流形的其他特征空间区域. 根据可攻击空间与目标数据流形的关系将可攻击空间分为邻近可攻击空间与背景可攻击空间. 其中, 邻近可攻击空间是隐藏有目标攻击对抗样本的特征空间, 而背景可攻击空间是暗藏无目标攻击区域和垃圾样本的特征空间. 基于可攻击空间对抗成因假设, 本文提出陷阱式对抗防御思维. 陷阱式对抗防御思维通过赋予 DNN 更多的数据特征信息和对输入样本怀疑的权力以对原本未定义的可攻击空间进行标记, 从而消除暗藏对抗样本的未定义特征空间. 理论上, 在不影响原分类精度的基础上, 命中该靶标可攻击空间的输入样本为对抗样本. 在模型实现方面, 本文提出陷阱式集成对抗防御网络 Trap-Net 用于对抗防御. Trap-Net 在 DNN 的训练阶段添加陷阱数据集, 并通过陷阱式平滑损失函数建立原目标数据和陷阱数据间的特殊诱导关系以诱使攻击算法所生成的对抗样本偏向靶标可攻击空间. 实验验证了 Trap-Net 通过集成网络的方式在不影响原目标数据分类精度的同时, 通过扩大靶标可攻击空间的空间大小以提高整体陷阱式集成网络的防御效力. Trap-Net 对白盒和黑盒对抗攻击都能进行有效的防御. 遗憾的是, Trap-Net 在面对有目标攻击时, 其防御有效性呈“V”型结构. 且对抗训练无法在保持目标数据流形与陷阱数据之间诱导关系的同时优化目标数据的流形分布, 这也是我们未来的研究方向之一.

References:

[1] Huang G, Liu Z, Van Der Maaten L, Weinberger KQ. Densely connected convolutional networks. In: Proc. of the 2017 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR). Honolulu: IEEE, 2017. 2261–2269. [doi: 10.1109/CVPR.2017.243]

[2] 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]

[3] Zhang Y, Pezeshki M, Brakel P, Zhang SZ, Laurent C, Bengio Y, Courville A. Towards end-to-end speech recognition with deep convolutional neural networks. In: Proc. of the 2016 Interspeech. ISCA, 2016. 410–414. [doi: 10.21437/Interspeech.2016-1446]

[4] Szegedy C, Zaremba W, Sutskever I, Bruna J, Erhan D, Goodfellow I, Fergus R. Intriguing properties of neural networks. In: Proc. of the 2nd Int'l Conf. on Learning Representations (ICLR). Banff: OpenReview.net, 2014

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

[6] Gilmer J, Metz L, Faghri F, Schoenholz S, Raghu M, Wattenberg M, Goodfellow I. Adversarial spheres. In: Proc. of the 6th Int'l Conf. on Learning Representations (ICLR). Vancouver: OpenReview.net, 2018.

[7] Gilmer J, Metz L, Faghri F, Schoenholz SS, Raghu M, Wattenberg M, Goodfellow I. The relationship between high-dimensional geometry and adversarial examples. arXiv:1801.02774, 2018.

[8] Moosavi-Dezfooli SM, Fawzi A, Fawzi O, Frossard P. Universal adversarial perturbations. In: Proc. of the 2017 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR). Honolulu: IEEE, 2017. 86–94. [doi: 10.1109/CVPR.2017.17]

[9] Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z. Rethinking the inception architecture for computer vision. In: Proc. of the 2016 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR). Las Vegas: IEEE, 2016. 2818–2826. [doi: 10.1109/CVPR.2016.308]

[10] LeCun Y, Bengio Y. Convolutional networks for images, speech, and time-series. The Handbook of Brain Theory and Neural Networks.

Cambridge: MIT Press, 1995.

[11] 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]

[12] Fawzi A, Moosavi-Dezfooli SM, Frossard P. Robustness of classifiers: From adversarial to random noise. In: Proc. of the 30th Int'l Conf. on Neural Information Processing Systems. Barcelona: Curran Associates Inc., 2016, 29: 1632–1640.

[13] Bengio Y, Courville A, Vincent P. Representation learning: A review and new perspectives. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2013, 35(8): 1798–1828. [doi: 10.1109/TPAMI.2013.50]

[14] Müller R, Kornblith S, Hinton G. When does label smoothing help? In: Proc. of the 33rd Int'l Conf. on Neural Information Processing Systems. Vancouver: Curran Associates Inc., 2019. 422.

[15] Cai XX, Du HM. Survey on adversarial example generation and adversarial attack method. Journal of Xi'an University of Posts and Telecommunications, 2021, 26(1): 67–75 (in Chinese with English abstract). [doi: 10.13682/j.issn.2095-6533.2021.01.011]

[16] Kurakin A, Goodfellow IJ, Bengio S. Adversarial examples in the physical world. In: Proc. of the 5th Int'l Conf. on Learning Representations. Toulon: OpenReview.net, 2018. 99–112.

[17] 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.

[18] 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]

[19] Xiao CW, Li B, Zhu JY, He W, Liu MY, Song D. Generating adversarial examples with adversarial networks. In: Proc. of the 27th Int'l Joint Conf. on Artificial Intelligence. Stockholm: AAAI Press, 2018. 3905–3911. [doi: 10.24963/ijcai.2018/543]

[20] Huang LF, Zhuang WZ, Liao YX, Liu N. Black-box adversarial attack method based on evolution strategy and attention mechanism. Ruan Jian Xue Bao/Journal of Software, 2021, 32(11): 3512–3529 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/6084.htm [doi: 10.13328/j.cnki.jos.006084]

[21] Papernot N, McDaniel P, Goodfellow I, Jha S, Celik ZB, Swami A. Practical black-box attacks against machine learning. In: Proc. of the 2017 ACM Asia Conf. on Computer and Communications Security. Abu Dhabi: ACM, 2017. 506−519. [doi: 10.1145/3052973.3053009]

[22] 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]

[23] Pang TY, Du C, Dong YP, Zhu J. Towards robust detection of adversarial examples. In: Proc. of the 32nd Int'l Conf. on Neural Information Processing Systems. Montréal: Curran Associates Inc., 2018. 4584–4594.

[24] Pang TY, Du C, Zhu J. Max-mahalanobis linear discriminant analysis networks. In: Proc. of the 35th Int'l Conf. on Machine Learning. Stockholm: PMLR, 2018. 4016–4025.

[25] Metzen JH, Genewein T, Fischer V, Bischoff B. On detecting adversarial perturbations. In: Proc. of the 5th Int'l Conf. on Learning Representations. Toulon: OpenReview.net, 2017.

[26] Feinman R, Curtin RR, Shintre S, Gardner AB. Detecting adversarial samples from artifacts. arXiv:1703.00410, 2017.

[27] Pang TY, Xu K, Du C, Chen N, Zhu J. Improving adversarial robustness via promoting ensemble diversity. In: Proc. of the 36th Int'l Conf. on Machine Learning. Long Beach: PMLR, 2019. 4970–4979.

[28] Van Der Maaten L, Hinton G. Visualizing Data using t-SNE. Journal of Machine Learning Research, 2008, 9(86): 2579–2605.

[29] Liu YP, Chen XY, Liu C, Song D. Delving into transferable adversarial examples and black-box attacks. In: Proc. of the 5th Int'l Conf. on Learning Representations. Toulon: OpenReview.net, 2017.

[30] Xu WL, Evans D, Qi YJ. Feature squeezing: Detecting adversarial examples in deep neural networks. In: Proc. of the 25th Annual Network and Distributed System Security Symp. Washington: The Internet Society, 2018. 15–26.

[31] Harder P, Pfreundt FJ, Keuper M, Keuper J. SpectralDefense: Detecting adversarial attacks on CNNs in the Fourier domain. In: Proc. of the 2021 Int'l Joint Conf. on Neural Networks. Shenzhen: IEEE, 2021. 1–8. [doi: 10.1109/IJCNN52387.2021.9533442]

[32] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, VanderPlas JT, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay É. Scikit-learn: Machine learning in Python. The Journal of Machine Learning Research, 2011, 12: 2825–2830.

[33] Papernot N, Faghri F, Carlini N, Goodfellow I, Feinman R, Kurakin A, McDaniel P. Adversarial Spheres. arXiv:1801.02774, 2018.

[34] Li YX, Jin W, Xu H, Tang JL. DeepRobust: A platform for adversarial attacks and defenses. In: Proc. of the 35th AAAI Conf. on Artificial Intelligence. Palo Alto: AAAI Press, 2021. 16078–16080. [doi: 10.1609/aaai.v35i18.18017]

[35] Strauss T, Hanselmann M, Junginger A, Ulmer H. Ensemble methods as a defense to adversarial perturbations against deep neural networks. In: Proc. of the 6th Int'l Conf. on Learning Representations. Vancouver: ICLR, 2018.

[36] Arjovsky M, Chintala S, Bottou L. Wasserstein generative adversarial networks. In: Proc. of the 34th Int'l Conf. on Machine Learning (ICML). Sydney: JMLR.org, 2017. 214–223.

[37] Abbasi M, Gagné C. Robustness to adversarial examples through an ensemble of specialists. In: Proc. of the 5th Int'l Conf. on Learning Representations. Toulon: OpenReview.net, 2017.

附中文参考文献:

[15] 蔡秀霞, 杜慧敏. 对抗攻击及对抗样本生成方法综述. 西安邮电大学学报, 2021, 26(1): 67–75. [doi: 10.13682/j.issn.2095-6533.2021.01.011]

[20] 黄立峰, 庄文梓, 廖泳贤, 刘宁. 一种基于进化策略和注意力机制的黑盒对抗攻击算法. 软件学报, 2021, 32(11): 3512–3529. http://www.jos.org.cn/1000-9825/6084.htm [doi: 10.13328/j.cnki.jos.006084]

[22] 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]

孙家泽(1980－), 男, 博士, 教授, CCF 高级会员, 主要研究领域为软件测试, 智能优化, 机器学习.

郑炜(1975－), 男, 博士, 副教授, CCF 高级会员, 主要研究领域为软件测试, 软件安全漏洞分析与检测.

温苏雷(1994－), 男, 博士生, 主要研究领域为对抗学习, 智能测试.

陈翔(1980－), 男, 博士, 副教授, CCF 高级会员, 主要研究领域为软件缺陷预测, 软件缺陷定位, 回归测试, 组合测试.