软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software [doi: 10.13328/j.cnki.jos.007172]

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

对抗鲁棒性评估的指标体系及其完备性*

石育澄[1,2], 韩亚洪[1,2]

[1](天津大学 智能与计算学部, 天津 300350)

[2](天津市机器学习重点实验室 (天津大学), 天津 300350)

通信作者: 韩亚洪, E-mail: yahong@tju.edu.cn

摘　要: 对抗鲁棒性评估需要结合对抗样本攻击能力与噪声幅度形成对深度学习模型噪声抵御能力的完整、准确的评测. 然而, 对抗鲁棒性评估评价指标缺乏完备性是现有对抗攻防方法的一个关键问题. 现有的对抗鲁棒性评估相关工作缺少对评价指标体系的分析与比较, 忽视了攻击成功率和不同范数对鲁棒性评估指标体系完备性的影响以及对攻防方法设计的限制. 从范数选择和度量指标两个维度展开对抗鲁棒性评价指标体系的讨论, 分别从评价指标定义域的包含关系、鲁棒性描述粒度以及鲁棒性评估序关系 3 个方面对鲁棒性评估指标体系完备性进行理论分析, 并得出以下结论: 使用均值等噪声统计量比使用攻击成功率等评价指标定义域更大且更全面, 同时能够保证任意两个对抗样本集合都能够进行比较; 使用 L_2 范数比使用其他范数在鲁棒性评估的描述上更具完备性. 在 6 个数据集上对 23 种模型及 20 种对抗攻击方法的大量实验验证了这些结论.

关键词: 对抗机器学习; 对抗鲁棒性; 对抗样本; 对抗攻击; 对抗扰动

中图法分类号: TP309

中文引用格式: 石育澄, 韩亚洪. 对抗鲁棒性评估的指标体系及其完备性. 软件学报. http://www.jos.org.cn/1000-9825/7172.htm

英文引用格式: Shi YC, Han YH. Metric System and Its Completeness of Adversarial Robustness Evaluation. Ruan Jian Xue Bao/Journal of Software (in Chinese). http://www.jos.org.cn/1000-9825/7172.htm

Metric System and Its Completeness of Adversarial Robustness Evaluation

SHI Yu-Cheng[1,2], HAN Ya-Hong[1,2]

[1](College of Intelligence and Computing, Tianjin University, Tianjin 300350, China)

[2](Tianjin Key Lab of Machine Learning (Tianjin University), Tianjin 300350, China)

Abstract: The assessment of adversarial robustness requires a complete and accurate evaluation of deep learning models' noise resistance by combining the attack ability and noise magnitude of adversarial samples. However, the lack of completeness in the adversarial robustness evaluation metric system is a key problem with the existing adversarial attack and defense methods. The existing work on adversarial robustness evaluation lacks analysis and comparison of the evaluation metric system. The impact of attack success rate and different norms on the completeness of the robustness evaluation metric system and the restrictions on designing attack and defense methods are neglected. In this study, the adversarial robustness evaluation metric system is discussed in two dimensions: norm selection and metric indicators. The theoretical analysis of robustness evaluation completeness is carried out from three aspects: the inclusion relation of the evaluation metric domain, robustness description granularity, and the order relationship of the robustness evaluation metric system. The following conclusions are drawn: using noise statistical quantities such as the mean results in a larger and more comprehensive definition domain of evaluation indicators compared to using attack success rates, while also ensuring that any two adversarial sample sets can be compared. Using the L_2 norm is more complete in the description of adversarial robustness evaluation compared to using other norms. Extensive experiments on 23 models and 20 adversarial attacks across 6 datasets validate these conclusions.

* 基金项目: 国家自然科学基金 (62376186, 61932009)
收稿时间: 2023-09-09; 修改时间: 2023-11-08, 2023-12-27; 采用时间: 2024-02-28; jos 在线出版时间: 2024-06-14
CNKI 网络首发时间: 2024-06-17

Key words: adversarial machine learning; adversarial robustness; adversarial sample; adversarial attack; adversarial perturbation

对抗机器学习[1]通过构建对抗样本探测与评估深度学习模型的鲁棒性, 即深度模型是否真正学习到了正确的语义概念, 以及该模型在何种情况下失效. 对抗攻击和对抗防御方法是对抗机器学习最重要的两方面研究内容, 而攻击和防御方法都围绕着针对深度学习模型的对抗鲁棒性评估展开. 其中, 对抗攻击的方法研究尝试在不同的场景下对图像[2]、视频[3]、音频[4]、文本[5]等数据构建对抗样本, 实现对目标模型可用性、隐私性和完整性等方面的评估与破坏. 而对抗防御的方法研究尝试构建对不同类型和目标的对抗攻击方法的反制措施, 在提升模型对抗鲁棒性的同时兼顾模型精度与效率问题. 因此, 对抗鲁棒性评估是对抗机器学习的核心问题. 除了可验证防御[6]等对模型鲁棒半径进行理论分析的方法外, 几乎所有的对抗攻防方法都需要直接或间接地在具体的模型上通过对抗鲁棒性评估的方式[7,8]对方法的性能进行检验.

全面、准确、完整地评估深度模型对抗鲁棒性的关键在于完备的评价指标体系, 但对抗鲁棒性评估指标体系缺乏完备性分析是现有对抗攻防方法的一个关键问题. 对抗样本攻击能力与对抗样本噪声幅度是对抗鲁棒性评估中两个核心的评价内容[9,10]. 对抗样本攻击能力即对抗样本影响和破坏目标模型的输出的能力, 如图像分类任务中目标模型的错误分类, 或目标检测任务中目标模型 mAP 的降低[11]. 对抗样本噪声幅度即对抗样本相对于原始数据的改变幅度, 例如图像分类任务中对图像噪声的范数的计算[12], 或结合人类的感知系统对噪声的不可察觉性的评价[13,14]. 在目前关于对抗鲁棒性的指标体系中, 评价一个攻击方法攻击能力较强的标准通常是在保证对目标模型攻击能力情况下减小对抗样本噪声幅度[15], 或在相同噪声幅度下尽可能提升对目标模型的攻击能力[16]. 在评价对抗防御方法时则相反, 多数情况下防御方法试图在相同的噪声幅度下减小对抗攻击对目标模型的影响[17], 或保证模型在相同的精度或性能下抵御更高的对抗噪声[18]. 对抗鲁棒性评价指标体系的选择在很大程度上决定了攻防方法的评估, 有助于发现和比较不同结构的深度学习模型在不同场景下的脆弱性. 构建完备的对抗鲁棒性评价指标体系不仅对攻防方法的发展方向有深远影响, 还对深度学习机理理解的加深, 以及提升深度学习模型鲁棒性也具有重要作用. 然而, 现有的关于对抗鲁棒性评估的研究主要针对特定的攻防方法或模型, 缺乏在不同范数下对鲁棒性评估完备性与鲁棒性指标体系的分析, 现有研究中也鲜有针对攻击成功率和不同范数对鲁棒性评估完备性影响的讨论. 同时, 目前对抗噪声幅度的度量标准不一, 不同范数下对抗攻击方法、对抗噪声分布、模型鲁棒半径都存在较大差异, 但目前尚缺乏有关不同范数对噪声搜索范围及对抗鲁棒性的影响的分析. 因此, 对抗鲁棒性评估的评价指标完备性的讨论, 对于改进当前对抗鲁棒性评估的方法, 对后续对抗攻防方法的评估与设计提供参考具有重要研究意义.

本文针对对抗机器学习中对抗鲁棒性评估的指标体系与完备性进行讨论. 其中, 对抗鲁棒性的指标体系包括范数选择和度量指标两个维度. 范数选择是评估单一对抗样本噪声幅度的方式, 常见的有 L_2 范数、 L_∞ 范数、 L_0 范数、 L_1 范数等. 范数选择反映了对抗鲁棒性指标体系对于噪声幅度大小的定义. 度量指标指的是评估多个对抗样本对目标模型攻击能力的方式. 目前最常见的两个度量指标分别是攻击成功率和一批对抗样本的噪声统计量, 如噪声幅度的均值/中位数. 另外还有模型的正确类别平均置信度, 以及扰动-准确率曲线等. 度量指标反映了对抗鲁棒性指标体系将一个对抗样本集合针对特定模型的攻击能力转化为一个可以直接比较的数值的方法的偏好. 范数选择与度量指标的组合形成了不同的对抗鲁棒性的完整评价指标. 目前对抗鲁棒性指标体系中最常见的组合有两类, 一类是在固定的范数选择下统计一批对抗样本的攻击成功率[16], 另一类是在保证一批对抗样本都被目标模型错分的情况下统计其噪声幅度的均值与中位数[15].

对抗鲁棒性评估作为检验对抗攻防方法性能的方式, 其不同的评价指标存在完备性方面的差异. 本文从评价指标定义域的包含关系、鲁棒性描述粒度以及鲁棒性评估序关系 3 个角度展开, 对常见的 4 种范数选择与 4 种度量指标组成的对抗鲁棒性评价指标体系的完备性进行分析. 首先, 从对抗鲁棒性评价指标定义域的包含关系出发, 本文证明了攻击成功率作为度量指标时参与对比的对抗样本集合都是噪声统计量的子集, 即噪声统计量是一种更全面的度量指标. 其次, 从鲁棒性描述粒度的角度, 本文证明了无论使用噪声统计量还是攻击成功率, 使用 L_2 范数可描述的模型对抗鲁棒性粒度都比其他范数粒度细. 当对比的相同目标模型上两个对抗样本集合相差较小时, 粒

度更细的范数能够更准确地对比二者对抗鲁棒性的细微区别, 因此 L_2 范数是一种更精细的范数选择. 另外, 关于鲁棒性评估序关系, 当设定的噪声幅度变化时, 目标模型即使在相同一批对抗样本上的成功率也可能发生变化. 类别置信度作为一种度量指标也无法正确反映模型在两批对抗样本上的鲁棒性对比关系. 但使用噪声统计量作为度量指标时, 可以对目标模型在任意两个对抗样本集合上对抗鲁棒性进行比较. 综合来看, 目前对抗鲁棒性指标体系下 L_2 范数+噪声统计量的组合是较为完备的评价指标.

为了验证上述有关对抗鲁棒性指标体系的分析, 本文在 6 个数据集上对 20 种不同设定下的对抗攻击方法生成的对抗样本进行了分析. 实验结果验证了 L_2 范数+噪声统计量较其他评价指标的完备性.

本文第 1 节介绍关于对抗鲁棒性评估及评价指标的研究现状. 第 2 节介绍对抗鲁棒性评估的指标体系, 包括单一对抗样本的范数选择和多个对抗样本的度量指标. 第 3 节从评价指标定义域的包含关系、鲁棒性描述粒度以及鲁棒性评估序关系 3 个方面对鲁棒性评估的完备性进行比较, 并讨论指标体系对攻防方法的反向影响. 第 4 节通过对比实验验证了提出的鲁棒性评估指标体系完备性相关理论. 最后总结全文.

1 对抗鲁棒性评估相关工作

随着深度学习方法广泛应用于计算机视觉、语音识别、自然语言处理等多个领域, 深度模型的鲁棒性引起了广泛关注. 对抗鲁棒性评估利用对抗攻击等方法, 通过理论分析与对抗样本生成的方式揭示并量化深度学习模型的脆弱性, 同时提升深度学习方法的可解释性. 现有的对抗鲁棒性评估主要包含 3 方面研究. 首先是从理论层面分析和比较不同深度模型鲁棒半径与鲁棒性边界, 例如对模型决策边界曲率与鲁棒性之间关系的分析[9]. 其次是从实证层面出发, 通过具体的攻击方法验证和量化深度模型鲁棒性, 例如通过将对抗噪声作为优化对象的优化攻击[20]评估深度模型能够抵御的最大噪声幅度, 或在不同的模型组合与参数设置下比较不同模型结构的对抗鲁棒性差异[21]. 另外, 对抗鲁棒性评估还可以为可验证防御[6]等非经验性的攻防方法设计提供理论依据和策略指导.

在对抗鲁棒性评估的评价指标方面, 首篇发现并提出对抗样本概念的工作[19]使用了模型的错误率 (error rate) 度量对抗样本在不同模型上的攻击能力, 并记录了不同对抗样本集合的基于 L_2 范数的改变幅度. 后续的工作[22]系统地整理了不同设定下攻击能力的评价指标, 将攻击能力细分为“置信度下降” “错分” “针对性错分” “源/目标针对性错分” 4 种情况. 目前对抗鲁棒性指标体系下不同的对抗攻防方法设计主要使用两种评价指标: 一类是固定 L_∞ 范数的阈值, 计算一个对抗样本集合在该阈值内攻击成功率, 常见于迁移攻击[23]、迭代攻击[24]和对抗训练[17]方法; 另一类是在确保所有对抗样本都错分情况下压缩对抗样本的 L_2 范数下的噪声, 常见于决策攻击[15]和零阶优化攻击[25]. 另外也有基于 L_0 范数[26]及 L_1 范数[27]的评价指标及针对性的攻防方法. 除了使用向量范数计算对抗样本噪声幅度之外, 结合人类感知系统设计适应性评价指标[14,28]也是一个研究方向.

目前有关对抗鲁棒性评估指标体系的完备性讨论仍处于探索阶段, 现有的分析主要包含 3 个方面. 首先是从鲁棒半径角度比较线性分类器在不同应用下 L_2 范数和 L_∞ 范数的优势[29], 其次是从图像输入空间中两个数据点之间距离及其相似度之间关系出发, 对不同范数对应的对抗鲁棒性评估方法的差异[12]展开讨论. 另外, 也有通过构建范数无关的对抗攻击[30]帮助改进当前对抗鲁棒性评估的方法. 然而, 现有的讨论主要针对特定的攻防方法或模型, 缺少系统的针对对抗鲁棒性指标体系及其完备性的分析.

2 对抗鲁棒性评估的指标体系

本文主要讨论分类任务. 假设 $F: X^N \to Y^C$ 是一个分类模型, 其中 X 表示输入空间, N 表示数据的维度. 这里假设输入空间的每个维度的取值范围相等, 都为 Gr . 如图像数据 $N = Width \times Height \times Channel$, 其中 $Width$ 、$Height$ 和 $Channel$ 分别表示图像宽、高和通道数, $Gr = [0,255]$. Y 表示分类任务的分类空间, C 为分类空间的类别数. 假设 $x \in X$ 为输入空间中任意数据, $F(x)$ 表示分类模型在 x 上预测的类别, 即 F 输出的 C 维置信度向量中最大值对应的类别. 下文在对抗机器学习语境下, 称分类模型为目标模型. 对于单一的原始数据 $x \in X$ 而言, 称在 x 上添加了噪声 z 的数据 $x' \in X$ 为对抗样本. 一个成功的对抗样本能够改变目标模型的预测结果, 或称使模型错分, 即

$F(x) \neq F(x')$. 注意, 在本文的定义中判定一个对抗样本是否成功根据是否改变目标模型在原始数据上的预测结果, 与原始数据本身的标签无关. 这是考虑到存在一些目标模型在原始数据上预测结果与标签不一致的情况.

对抗鲁棒性评估的指标体系包含范数选择和度量指标两个维度. 范数选择是指鲁棒性评估指标体系中对单一对抗样本噪声幅度评价方式. 在达成目标模型错分的前提下, 对抗样本相对于原始数据添加的噪声的量是对抗鲁棒性评估的关键. L_2 和 L_∞ 范数是目前最常用的两种关于单一对抗样本噪声的量的范数选择:

$$\|z\|_2 = \sqrt{\left(\sum_{i=1}^{N}(x_i' - x_i)^2\right)},\ \|z\|_\infty = \max_{1\leqslant i\leqslant N}|x_i' - x_i| \tag{1}$$

目前也有一些对抗鲁棒性评估的工作使用 L_0 范数[26]及 L_1 范数计算单一对抗样本噪声幅度:

$$\|z\|_0 = \sum_{i=1}^{N}\mathbb{I}(x_i' \neq x_i),\ \|z\|_1 = \sum_{i=1}^{N}\left|x_i' - x_i\right| \tag{2}$$

度量指标是指鲁棒性评估指标体系中评估多个对抗样本对目标模型攻击能力的方式. 在对抗攻防工作的实验部分, 通常不是在单一的原始数据上对目标模型进行对抗鲁棒性评估, 而是对整个测试集上所有的数据生成对抗样本[21]. 对于包含 N_d 个对抗样本的集合 $\{x_1', x_2', \ldots, x_{N_d}'\}$ 而言, 目前有 4 种常见的评估多个对抗样本对目标模型攻击能力的方式. 第 1 种是计算该集合的攻击成功率, 即计算集合中使目标模型错分的对抗样本的比例:

$$SR = \frac{\sum_{i=1}^{N_d}\mathbb{I}(F(x_i) \neq F(x_i'))}{N_d} \tag{3}$$

其中, $\mathbb{I}$ 表示指示函数. 第 2 种是统计集合中对抗噪声的性质, 如均值和中位数等:

$$STA_{\text{mean}} = \frac{\sum_{i=1}^{N_d}\|x_i' - x_i\|_p}{N_d} \tag{4}$$

$$STA_{\text{mid}} = \begin{cases} z'_{\frac{N_d+1}{2}} & N_d\text{为奇数} \\ \dfrac{1}{2}\left(z'_{\frac{N_d+1}{2}} + z'_{\frac{N_d}{2}+1}\right) & N_d\text{为偶数} \end{cases} \tag{5}$$

其中, $\|\cdot\|_p$ 表示 L_p 范数下的距离, z_j' 表示 N_d 个对抗样本的噪声按照降序排列后第 j 个噪声. 下文中, 对通过均值、中位数等对抗噪声集合性质代表的度量指标简称为噪声统计量.

第 3 种是计算模型输出的正确类别平均置信度[53], 即对抗样本输入模型后, 正确类别置信度的平均值:

$$ACTC = \frac{\sum_{i=1}^{N_d}P(F(x_i') = F(x_i))}{N_d} \tag{6}$$

其中, P 表示模型分类的置信度. 正确类别平均置信度越低, 表示对抗样本改变原始输出结果的能力越强, 模型在该对抗样本集合上的鲁棒性越低.

除以上度量指标外, 对抗鲁棒性评估工作中通常对多种不同噪声幅度下模型的分类准确率绘制曲线, 称作扰动-准确率曲线[21], 其中横坐标表示对抗攻击噪声幅度, 纵坐标表示模型的准确率. 若曲线随对抗噪声幅度下降较慢, 或更靠右, 表明模型的鲁棒性更强.

在对抗鲁棒性评估的实验验证中, 需要在指标体系下同时确定单一对抗样本的范数选择和多个对抗样本的度量指标. 基于上述范数选择和度量指标的定义, 对抗鲁棒性的指标体系可以选取上述 4 种范数选择中的一种与度量指标中的一种. 其中 L_∞ 范数+成功率、 L_2 范数+噪声统计量这两种是目前最常用的评价指标[15,16]. 参照文献 [9], 以 4 种评价指标为例展示对抗鲁棒性的定义.

定义 1. L_2 范数+成功率的对抗鲁棒性.

给定目标模型 F 和阈值 $Thr_2 \in [0, \sqrt{Gr_{\max}^2 N}]$, $Gr_{\max}$ 指取值范围的最大值, 定义 F 在 L_2 范数+成功率下的对抗

鲁棒性为: 在整个输入空间内与原始数据 L_2 距离不超过阈值的邻域内对抗样本占所有样本比例的期望,

$$\rho_{(L_2+SR)}=\mathbb{E}_{x\in X}\left[\frac{\sum\mathbb{I}(\|x'-x\|_2\leqslant Thr_2\wedge F(x')\neq F(x))}{\sum\mathbb{I}(\|x'-x\|_2\leqslant Thr_2)}\right] \tag{7}$$

其中, $\rho_{(L_2+SR)}$ 越低, 表示在相同的阈值下, 导致目标模型 F 在原始数据邻域内错分的对抗样本的比例的期望越小, 即对抗鲁棒性越高.

定义 2. L_∞ 范数+成功率的对抗鲁棒性.

给定目标模型 F 和阈值 $Thr_\infty\in Gr$, 定义 F 在 L_∞ 范数+成功率下的对抗鲁棒性为: 在整个输入空间内与原始数据 L_∞ 距离不超过阈值的邻域内对抗样本占所有样本比例的期望,

$$\rho_{(L_\infty+SR)}=\mathbb{E}_{x\in X}\left[\frac{\sum\mathbb{I}(\|x'-x\|_\infty\leqslant Thr_\infty\wedge F(x')\neq F(x))}{\sum\mathbb{I}(\|x'-x\|_\infty\leqslant Thr_\infty)}\right] \tag{8}$$

定义 3. L_2 范数+噪声统计量的对抗鲁棒性.

给定目标模型 F, 定义 F 在 L_2 范数+噪声统计量下的对抗鲁棒性为: 在整个输入空间内, 改变目标模型 F 在原始输入上输出结果所需最小 L_2 范数下噪声幅度的期望,

$$\rho_{(L_2+STA)}=\mathbb{E}_{x\in X}[\operatorname*{argmin}_{x'\in X}\|x'-x\|_2,\ \text{s.t.}\ \ F(x')\neq F(x)] \tag{9}$$

其中, $\rho_{(L_2+STA)}$ 越高, 表示使目标模型 F 错分所需添加的噪声幅度期望越高, 即对抗鲁棒性越高.

定义 4. L_∞ 范数+噪声统计量的对抗鲁棒性.

给定目标模型 F, 定义 F 在 L_∞ 范数+噪声统计量下的对抗鲁棒性为: 在整个输入空间内, 改变目标模型 F 在原始输入上输出结果所需最小 L_∞ 范数下噪声幅度的期望,

$$\rho_{(L_\infty+STA)}=\mathbb{E}_{x\in X}[\operatorname*{argmin}_{x'\in X}\|x'-x\|_\infty,\ \text{s.t.}\ \ F(x')\neq F(x)] \tag{10}$$

上述对抗鲁棒性的定义只停留在理论层面, 目标模型对抗鲁棒性的具体数值需要攻击方法生成对抗样本进行评估和逼近. 对于固定的目标模型和输入空间而言, 一种对抗攻击方法能够在阈值内生成攻击成功率更高的对抗样本集合, 或在保证错分情况下生成噪声幅度更小的对抗样本, 意味着这种方法能够更准确地反映目标模型真实的对抗鲁棒性, 因此称该攻击方法的攻击能力更强. 然而, 考虑到图像等高维数据输入空间庞大, 在实际的实验过程中难以在所有原始数据上生成所有可能的对抗样本以计算目标模型对抗鲁棒性, 通常使用 F 在一批噪声幅度小于或等于阈值的对抗样本组成的对抗样本集合上的错分率作为定义 1、定义 2 中对抗鲁棒性的统计量, 使用一批能够成功错分目标模型 F 的对抗样本的噪声幅度均值或中位数作为定义 3、定义 4 中对抗鲁棒性的统计量.

图 1 展示了由 5 个对抗样本组成的对抗样本集合在对抗鲁棒性指标体系下 4 种评价指标对应的结果, 这里假设 5 个对抗样本都能够使目标模型错分. 在度量指标为攻击成功率时, 黑色对抗样本表示落在噪声阈值内, 灰色表示噪声幅度大于噪声阈值. 当度量指标为噪声统计量时, 颜色越偏蓝/红表示对应的 L_2 / L_∞ 噪声幅度越大.

从集合与关系的角度出发, 指标体系下 4 种评价指标定义了输入空间 X 的幂集上 4 种不同的二元关系 $Re=\{<Adv_1,Adv_2>,|Adv_1,Adv_2\in\mathcal{P}(X)\}$, 其中 Adv_1 和 Adv_2 是两个定义在输入空间上的对抗样本集合, $\mathcal{P}(X)$ 表示由所有 X 的子集组成的集合. 进一步, 根据 4 种对抗鲁棒性的定义, 可给出对应的"优于或等于"关系 $\succcurlyeq$:

$$\succcurlyeq_{(L_2+SR)}=\{<Adv_1,Adv_2>|SR_1\geqslant SR_2\},\ SR=\frac{1}{N_d}\mathbb{I}(x'\in Adv\wedge\|x'-x\|_2\leqslant Thr_2\wedge F(x')\neq F(x)) \tag{11}$$

$$\succcurlyeq_{(L_\infty+SR)}=\{<Adv_1,Adv_2>|SR_1\geqslant SR_2\},\ SR=\frac{1}{N_d}\mathbb{I}(x'\in Adv\wedge\|x'-x\|_\infty\leqslant Thr_\infty\wedge F(x')\neq F(x)) \tag{12}$$

$$\succcurlyeq_{(L_2+STA)}=\{<Adv_1,Adv_2>|STA_{\text{mean}1}\leqslant STA_{\text{mean}2}\} \tag{13}$$

$$\succcurlyeq_{(L_\infty+STA)}=\{<Adv_1,Adv_2>|STA_{\text{mean}1}\leqslant STA_{\text{mean}2}\} \tag{14}$$

等式 (11)–(14) 中范数选择可根据等式 (2) 替换为 L_0 及 L_1 范数, 或根据等式 (6) 将度量指标替换为正确类别平均置信度.

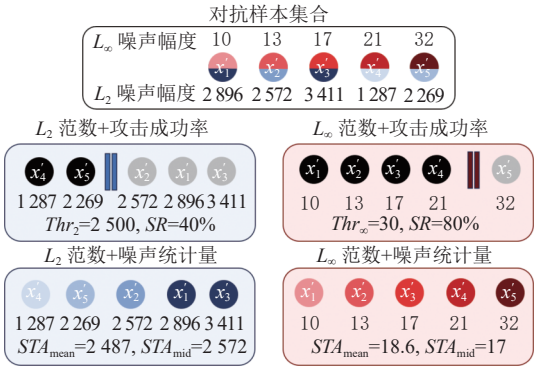

图 1　4 种对抗鲁棒性评价指标示意

3　对抗鲁棒性评估指标体系的完备性分析

对抗鲁棒性评估通过攻击方法生成对抗样本, 从而评估目标模型在不同分布和不同幅度的对抗噪声下抵御原始输出出现变化的能力. 其中, 对抗鲁棒性评估的指标体系完备性对评估过程中涉及的目标模型、攻击方法以及可能涉及的防御方法的性能比较至关重要. 首先, 对抗鲁棒性的载体应是完备的, 即评估对象应涵盖整个输入空间, 目标模型在任意对抗样本集合上都能得到对应的鲁棒性评估结果. 其次, 对抗鲁棒性的描述应是完备的, 即评估结果应尽可能准确地刻画目标模型的对抗鲁棒性, 尽可能避免出现相同的评估结果取值导致鲁棒性难以区分. 同时, 对抗鲁棒性评估的结果应保持独立性, 不依赖特定攻击方法或任何其他元素. 另外, 一个完备的对抗鲁棒性评估过程不应对攻防方法的设计造成反向的限制和束缚. 本节分别从评价指标定义域的包含关系、对抗鲁棒性描述粒度以及鲁棒性评估序关系 3 个层面分析对抗鲁棒性评估在载体、描述上的完备性以及评估结果的独立性, 并讨论不同评价指标对攻防方法设计的约束.

3.1　评价指标定义域的包含关系

在第 2 节关于对抗鲁棒性评估指标体系的定义中, 度量指标为成功率的定义 1 和定义 2 都包含一个阈值 Thr. 这是因为通常情况下分类任务类别数 $C>1$, 因此对任意的原始数据 $x_1 \in X$, 一定存在 $x_2 \in X$ 使得 $F(x_1) \neq F(x_2)$, 那么 x_2 就可以看作 x_1 的对抗样本. 如果不对噪声幅度的阈值加以限制, 对于任意的包含 N_d 个原始数据的集合 $Ori \subseteq X$, 一定可以找到对抗样本集合 $Adv \subseteq X$, 使得 $\forall 1 \leqslant i \leqslant N_d$, $F(x_i) \neq F(x_i')$, $x_i \in Ori$, $x_i' \in Adv$. 也就是说, 只要对原始数据添加大量噪声, 或直接选择与原始数据预测类别不同的数据作为对抗样本, 攻击成功率将轻易达到 100%. 因此, 添加阈值可以限制添加噪声的量, 只对集合中不超过噪声阈值的对抗样本计算成功率, 即仅比较对抗攻击方法在设定的噪声阈值内能够实现的攻击成功率. 评价指标本身可以看作是从对抗样本集合向鲁棒性评估结果的映射: 给定对抗样本集合 $Adv \subseteq X$, 输出对应的评估结果. 然而, 在度量指标为成功率的情况下引入噪声阈值的结果是, 定义 1 和定义 2 中评价指标的定义域不再是输入空间 X 的幂集.

定理 1. 对于 L_2, L_∞, L_0 和 L_1 范数 4 种不同的范数选择, 任意一种范数与成功率组成的评价指标定义域是该范数与噪声统计量组成的评价指标的子集, 即 $U_{L_\infty+SR} \subseteq U_{L_\infty+STA}$, $U_{L_2+SR} \subseteq U_{L_\infty+STA}$, $U_{L_0+SR} \subseteq U_{L_0+STA}$, $U_{L_1+SR} \subseteq U_{L_1+STA}$

证明: 将评价指标看作定义在对抗样本幂集 $\mathcal{P}(X)$ 上的函数, 记 U_{L_2+SR}, $U_{L_\infty+SR}$, U_{L_1+SR}, U_{L_0+STA}, $U_{L_\infty+STA}$, U_{L_2+STA}, U_{L_0+SR}, U_{L_1+STA} 分别是每个评价指标对应的定义域.

当度量指标为噪声统计量时, $U_{L_\infty+STA}=\mathcal{P}(X)$. 当度量指标为成功率时, $U_{L_2+SR}=\mathcal{P}(\{x' \mid x' \in X \wedge \|x'-x\|_2 \leqslant Thr_2\})$. 记 $\widehat{X}=\{x' \mid x' \in X \wedge \|x'-x\|_2 \leqslant Thr_2\}$. 若 $Thr_2 < \sqrt{Gr_{\max}^2 N}$, 则 $\|x_2'-x_2\|_2=\sqrt{Gr_{\max}^2 N}$, $\exists x_2' \in X$. 根据定义 1, $x_2' \notin \hat{X}$, 因此 $\hat{X} \subsetneq X$. 若 $Thr_2=\sqrt{Gr_{\max}^2 N}$, 则 $\hat{X}=X$. 因此, $U_{L_\infty+SR} \subseteq U_{L_\infty+STA}=U_{L_2+STA}$. 其中, 第 1 个 $\subseteq$ 当且仅当 $Thr_2=\sqrt{Gr_{\max}^2 N}$ 时可取相等.

同理 $U_{L_2+SR} \subseteq U_{L_\infty+STA} = U_{L_2+STA}$, 当且仅当$Thr_\infty = Gr_{\max}$ 时第 1 个$\subseteq$ 可取相等.$U_{L_0+SR} \subseteq U_{L_0+STA}$, 当且仅当 $Thr_0 = N$ 时第 1 个$\subseteq$ 可取相等.$U_{L_1+SR} \subseteq U_{L_1+STA}$, 当且仅当$Thr_1 = Gr_{\max} N$ 时第 1 个$\subseteq$ 可取相等.

定理 1 说明, 当度量指标为噪声统计量时, 计算对抗鲁棒性时定义域没有任何限制, 因此所能描述的对抗样本范围更广. 输入空间内任意单一或多个对抗样本都包含在内, 对抗样本集合的全集即 $\mathcal{P}(X)$. 而当度量指标为成功率时, 超过噪声阈值的对抗样本不会参与评估对抗鲁棒性, 除非将噪声阈值设为最大, 否则参与评估的对抗样本只是整个输入空间的真子集, 所能描述的对抗样本范围更小. 也就是说, 成功率作为度量指标时评价指标的定义域包含于噪声统计量作为度量指标时评价指标的定义域. 但正如前文所述, 不对噪声幅度进行限制的情况下只要选取与原始数据在目标模型预测结果不同的输入即可构建对抗样本. 因此使用成功率作为度量指标时一定需要设置一个较小的噪声阈值. 例如在图像数据上, $Gr = [0,255]$, 当范数选择为 L_∞ 范数时常用的噪声阈值为 $Thr_\infty = 16$ [24]. 也就是说, 参与对抗鲁棒性评估的全集由整个输入空间缩小为不超过噪声阈值的子空间 $\widehat{X} = \{x' | x' \in X \wedge \|x' - x\|_\infty \leqslant 16\}$. 显然, 相对于整个输入空间而言, $\hat{X}$ 中对抗样本的数量远小于整个输入空间 X 中包含的数量. 这主要带来两个问题. 首先, 设定噪声阈值后, 对抗鲁棒性评估的目标会随之发生改变. 目标从“如何评估目标模型在整个输入空间内对抗样本的比例”改为“如何评估噪声幅度小于噪声阈值情况下对抗样本的比例”, 这可能导致对抗鲁棒性评估结果出现偏差, 这一点将在第 3.4 节中进一步讨论. 其次, 噪声阈值是由实验设计者主观设定的, 这影响了评估结果的独立性. 上述两个问题在噪声统计量作为度量指标时是不存在的.

定理 1 的结论解释了扰动-准确率曲线这种度量指标的存在. 由于不同的噪声阈值对应不同的攻击成功率, 实验中需要设定多种不同的噪声阈值才能相对全面地评估模型的对抗鲁棒性. 但扰动-准确率曲线只能在大量的阈值中的几点进行采样, 仍然无法解决改变攻击成功率能描述的对抗样本范围较小的问题. 另外, 正确类别平均置信度的定义域也是对抗样本集合的全集, 因为任何一组对抗样本都可以输入到模型中计算置信度.

综上, 无论采用哪种范数选择, 噪声统计量作为度量指标时对抗鲁棒性的评价指标定义域都大于或等于使用攻击成功率定义域. 噪声统计量评估的对象是整个输入空间中所有的对抗样本, 因此是一种更全面的度量指标. 使用噪声统计量时参与对抗鲁棒性评估的载体更加完备.

3.2 对抗鲁棒性描述粒度

对抗鲁棒性评估指标体系完备性的另一个重要层面是对抗鲁棒性描述粒度. 反映在具体的评估过程中, 粒度指面对所有可能的对抗样本集合时, 可能出现的不同对抗鲁棒性取值的总数. 数量越高则描述的粒度越细.

定义 5. 对抗鲁棒性描述粒度.

当使用噪声统计量作为度量指标时, 对抗鲁棒性描述粒度为:

$$Circ_{STA} = |S_{STA}|,\ S_{STA} = \left\{ \frac{\sum_{i=1}^{N_d} \|x_i' - x_i\|_p}{N_d} \middle| x_i \in X \right\} \tag{15}$$

其中, S_{STA} 表示由所有可能出现的噪声统计量组成的集合, $|S_{STA}|$ 表示该集合中元素总数.

当使用攻击成功率作为度量指标时, 使用对抗鲁棒性描述粒度为:

$$Circ_{SR} = |S_{Thr}| \cdot |S_{SR}|,\ S_{SR} = \left\{ \frac{\sum_{i=1}^{N_d} \mathbb{I}(F(x_i) \neq F(x_i'))}{N_d} \middle| x_i \in X \right\} \tag{16}$$

其中, S_{Thr} 表示所有可能出现的噪声阈值组成的集合, S_{SR} 表示所有可能出现的成功率组成的集合.

对于任意的对抗样本集合 $Adv_1, Adv_2 \subseteq X$, 评价指标体系中都能给出对应的成功率或噪声统计量, 同时根据第 2 节给出 Adv_1 和 Adv_2 之间对抗鲁棒性优于或等于的关系, 定义 5 是对特定评价指标下不同成功率或噪声统计量的取值总量的描述. 对于噪声统计量, $Circ_{STA}$ 描述在整个输入空间内任取 N_d 个对抗样本, 其噪声幅度均值可能

有多少种不同的取值. 对于攻击成功率, $Circ_{SR}$ 描述在不同的噪声阈值取值 S_{Thr} 下, 攻击成功率可能有多少种不同的取值. 由于攻击成功率只有在特定的噪声阈值下有意义且可比较, 因此每种可能的噪声阈值下对抗鲁棒性描述粒度的总和为总的对抗鲁棒性描述粒度. 根据 4 种不同的评价指标及定义 5, 可得出以下关系:

定理 2. 在 N 维输入空间 X 上, 若取值范围均匀且取值范围大于零, 则 4 种范数选择中 L_2 范数对应的对抗鲁棒性描述粒度大于或等于其他范数.

证明: 当使用噪声统计量作为度量指标时, 若 $N_d=1$, 在 L_∞ 范数下有 $\|x'-x\|_\infty \in Gr,\ \forall x', x \in X$. 若 $N_d \geqslant 1$, 则 $N_d \cdot Gr_{\min} \leqslant \sum_{i=1}^{N_d} \|x'_i - x_i\|_\infty \leqslant N_d \cdot Gr_{\max}$. 其中, $Gr_{\min}$ 为 Gr 中噪声幅度的最小值, $Gr_{\max}$ 为最大值. 因为取值范围均匀, 记 Gr 中所有元素降序排列时任意两个相邻元素间隔为 $gap = \dfrac{Gr_{\max} - Gr_{\min}}{|Gr|-1}$. 记 $Circ_{L_\infty+STA}$ 为使用噪声统计量作为度量指标时 L_∞ 范数下的描述粒度, 有 $Circ_{L_\infty+STA} = \dfrac{N_d(Gr_{\max}-Gr_{\min})}{gap} + 1 = N_d \cdot |Gr| - N_d + 1$.

若 $N_d=1$, 在 L_1 范数下, 噪声幅度可取从 $Gr_{\min}\cdot N$ 到 $Gr_{\max}\cdot N$ 范围内任意间隔为 gap 的值, 因此 $Circ_{L_1+STA} = |Gr|\cdot N - N + 1$. 在 L_0 范数下, 噪声幅度为非 0 元素个数, 因此 $Circ_{L_0+STA} = N$.

记 $Circ_{L_2+STA}$ 为使用噪声统计量作为度量指标时 L_2 范数下的描述粒度, 若 $N_d=1$, 在 L_2 范数下有 $Circ_{L_2+STA} \geqslant |Gr|$, s.t. $(x'-x)^{(i)} = (x'-x)^{(j)},\ \forall 1 \leqslant i, j \leqslant N$.

其中 $(x'-x)^{(i)}$ 表示对抗噪声 $x'-x$ 所有 N 个维度中第 i 个维度的值. 此时噪声幅度集合可表示为 $\{|gr|\cdot\sqrt{N} | gr \in Gr\}$. 记满足 $(x'-x)^{(i)} = Gr_{\min},\ \forall 1 \leqslant i \leqslant N,$ 的对抗样本为 $x'_{Gr_{\min}}$, 记对抗样本集合 Adv_{sel} 为 $Adv_{\text{sel}} = \{x'_{\text{sel}} | \mathbb{I}(x'^{(i)}_{\text{sel}} \neq x'^{(i)}_{Gr_{\min}}) = sel \wedge \mathbb{I}(x'^{(i)}_{\text{sel}} = Gr_{\min} + gap) = sel \wedge 1 \leqslant i \leqslant N-1 \wedge 1 \leqslant sel \leqslant N-1\}$.

此时集合 Adv_{sel} 的元素总数为 $N-1$, 满足 $\forall x' \in Adv_{\text{sel}},\ Gr_{\min}\cdot\sqrt{N} \leqslant \|x'-x\|_2 \leqslant (Gr_{\min}+gap)\cdot\sqrt{N}$.

因此, 若 $N_d=1$, 在 L_2 范数下满足 $Circ_{L_2+STA} \geqslant |Gr|\cdot N - N + 1$. 若 $N_d \geqslant 1$, 记对抗样本集合 Adv_1 为 $Adv_1 = \{x' | x'^{(i)} = gr,\ gr \in Gr,\ \forall i \in \mathcal{N}_+ \wedge i \leqslant N\}$.

记对抗样本集合 Adv_2 为:

$$Adv_2 = \{x' | \mathbb{I}(x'^{(i)} = Gr_{\min}) = j \wedge \mathbb{I}(x'^{(i)} = Gr_{\min} + gap) = N - j \wedge i, j \in \mathcal{N}_+ \wedge i \leqslant N \wedge j \leqslant N_d - 1\}.$$

集合 $\mathcal{P}(Adv_2)$ 中任意一个元素的噪声统计量满足:

$$\left(Gr_{\min} + \frac{gap}{N_d}\right)\cdot\sqrt{N} \leqslant STA_{\text{mean}} \leqslant (Gr_{\min}+gap)\cdot\sqrt{N},\ \forall adv \in \mathcal{P}(Adv_2),$$

且 $\mathcal{P}(Adv_2)$ 中所有元素的噪声统计量共有 N_d-1 种不同情况. 记对抗样本集合 Adv_3 为 $Adv_3 = Adv_2 \cup \{x'_{sel}\}$, $x'_{sel} \in Adv_{sel}$.

集合 $\mathcal{P}(Adv_3)$ 中任意一个元素的噪声统计量为:

$$STA_{\text{mean}} = \frac{\sqrt{N}}{N_d}(N_d \cdot Gr_{\min} + gap \cdot j - Gr_{\min} + \sqrt{(Gr_{\min}+gap)^2 \cdot sel + Gr_{\min}^2 \cdot (N-sel)}),$$

且 $\mathcal{P}(Adv_3)$ 中所有元素的噪声统计量共有 $N-1$ 种不同情况. 综上, 有:

$$Circ_{L_2+STA} \geqslant |Gr|\cdot N_d \cdot N - N_d \cdot N - 2|Gr|\cdot N + 2|Gr| + 2N - 1.$$

因此, 使用噪声统计量作为度量指标时, $Circ_{L_\infty+STA} \leqslant Circ_{L_2+STA}$, 当且仅当 $N=1$ 时两个不等式取等号.

对比 $N_d=1$ 时 L_2 和 L_1 范数的噪声统计量. 若 $N \geqslant 3$ 且 $|Gr| \geqslant 3$, 记对抗样本集合 Adv_4 为 $Adv_4 = \{x' \big| |x'^{(i)} - x'^{(j)}| = 0 \vee |x'^{(i)} - x'^{(j)}| = gap \wedge i, j \in \mathcal{N}_+ \wedge i, j \leqslant N\}$, L_1 范数和 L_2 范数上 Adv_4 的噪声统计量集合元素数量均为 $|Gr|\cdot N - N + 1$. 然而, 对于对抗样本集合 $Adv_5 = \{x' | \mathbb{I}(x'^{(i)} = Gr_{\min} + 2\cdot gap) = \mathbb{I}(x'^{(i)} = Gr_{\min}) = 1 \wedge \mathbb{I}(x'^{(i)} = Gr_{\min} + gap) = N-2 \wedge i \in \mathcal{N}_+ \wedge i \leqslant N\}$ 和 $Adv_6 = \{x' | \mathbb{I}(x'^{(i)} = Gr_{\min} + 2\cdot gap) = \mathbb{I}(x'^{(i)} = Gr_{\min} + gap) = 1 \wedge \mathbb{I}(x'^{(i)} = Gr_{\min}) = N-2 \wedge i \in \mathcal{N}_+ \wedge i \leqslant N\}$ 并集 $Adv_7 = Adv_5 \cup Adv_6$, 其中至少存在一个元素 x_{sp}', 其噪声统计量不属于 L_2 范数上 Adv_4 的噪声统计量集合中, 但属于 L_1 范数上 Adv_4 的噪声统计量集合. 因此 $N_d=1$ 时, 若 $N \geqslant 3$ 且 $|Gr| \geqslant 3$, $Circ_{L_1+STA} < Circ_{L_2+STA}$. 同理, $N_d > 1$ 时, 若 $N \geqslant 3$ 且 $|Gr| \geqslant 3$, $Circ_{L_1+STA} < Circ_{L_2+STA}$.

当使用攻击成功率作为度量指标时, 所有可能出现的成功率组成的集合满足 $|S_{SR}| = N_d + 1$. 在 L_∞ 范数下, 噪

声阈值取值总数有 $|S_{Thr_\infty}| = |Gr|$. 因此, 对抗鲁棒性描述粒度满足 $Circ_{L_\infty+SR} = |Gr|\cdot(N_d+1)$. 在 L_1 范数下, 噪声阈值取值总数为 $|S_{Thr_1}| = |Gr|\cdot N - N + 1$, 因此 $Circ_{L_1+SR} = (|Gr|\cdot N - N + 1)\cdot(N_d+1)$. 在 L_0 范数下, 噪声阈值取值总数为 $|S_{Thr_0}| = N$, 因此 $Circ_{L_0+SR} = N\cdot(N_d+1)$.

在 L_2 范数下, 噪声阈值取值总数有 $|S_{Thr_2}| \geqslant |Gr|\cdot N - N + 1$. 因此, 对抗鲁棒性描述粒度满足:

$$Circ_{L_2+SR} \geqslant (|Gr| + N - 1)(N_d+1).$$

因此, $Circ_{L_\infty+SR} \leqslant Circ_{L_2+SR}$, $Circ_{L_0+SR} \leqslant Circ_{L_2+SR}$ 当且仅当 $N = 1$ 时两个不等式取等号. 根据上述分析, $Circ_{L_1+SR} < Circ_{L_2+SR}$.

定理 2 说明, 在数据维度 $N > 1$ 的情况下, L_2 是一种对抗鲁棒性描述粒度更细的范数选择. 对抗鲁棒性描述粒度越细, 比较接近的模型间对抗鲁棒性的细微差别就越可能被区分. 可以看出, 当使用 L_2 作为范数选择时, 无论使用何种度量指标, 对抗鲁棒性描述粒度都是与数据维度 N 正相关的. 这是因为数据维度越高, 单一对抗样本各维度的噪声幅度平方和可能的取值就越多. 但无论维度是多少, 使用 L_∞ 范数时单一对抗样本的取值总数都恒等于输入空间的取值总数. 如果同时存在的对抗样本集合数量大于 $|Gr|\cdot(N_d+1)$, 那么至少存在两组对抗样本在 L_∞ 范数+成功率的评价指标下反映的模型对抗鲁棒性是没有差别的. 使用 L_0 范数时单一对抗样本的取值总数都恒等于输入空间的维数. 同时, 存在一些对抗样本, 其 L_2 范数可以区分但 L_1 范数相等的情况, 因此 L_2 范数能够描述的对抗鲁棒性情况最多.

对于其他两类度量指标来说, 扰动-准确率曲线的描述粒度等于相同范数选择下攻击成功率的描述粒度, 并不影响定理 2 的结论. 而正确类别平均置信度是一个连续值, 因此理论上其粒度是无限细的, 但正确类别平均置信度的问题在于鲁棒性评估序关系, 这一点将在第 3.3 节讨论.

无论使用噪声统计量还是成功率, 使用 L_2 范数可描述的对抗样本鲁棒性粒度都比其他 3 种范数粒度细. 当两个对抗样本集合相差较小时, L_2 范数能够更准确地区分二者反映的目标模型对抗鲁棒性的细微区别, 出现相同评估结果取值的可能性越低. 作为一种更精细的范数选择, 使用 L_2 范数时对抗鲁棒性评估的描述更加完备.

3.3 鲁棒性评估序关系

对抗鲁棒性是深度学习模型的本质属性, 与攻击方法性能高低相互独立, 因此不同模型对抗鲁棒性评估结果反映的不同攻击方法的性能高低应是较为一致的. 一个完备的鲁棒性评估结果应保持足够的独立性, 不依赖特定攻击方法或任何其他元素. 同时, 任意两个对抗鲁棒性评估结果之间都应该能够进行比较. 然而, 从对抗鲁棒性评估序关系的角度来说, 使用攻击成功率和使用噪声统计量作为度量指标时“优于或等于”关系的性质是不同的:

定理 3. 只有当噪声统计量作为度量指标时, 对抗鲁棒性评估满足完全性的预序关系.

证明: 对于 $\succcurlyeq_{(L_p+STA)}$, $p \in \{0,1,2,\infty\}$, 有 $STA_{\text{mean}} = STA_{\text{mean}}$, $\forall Adv \in \mathcal{P}(X)$. 因此, 有 $Adv \succcurlyeq_{(L_p+STA)} Adv$, $p \in \{0,1,2,\infty\}$, 故 $\succcurlyeq_{(L_p+STA)}$, $p \in \{0,1,2,\infty\}$ 满足自反性.

对于 $\succcurlyeq_{(L_p+STA)}$, $p \in \{0,1,2,\infty\}$, 有

$$Adv_1 \succcurlyeq_{(L_p+STA)} Adv_2 \wedge Adv_2 \succcurlyeq_{(L_p+STA)} Adv_3 \Rightarrow STA_{\text{mean1}} \leqslant STA_{\text{mean2}} \leqslant STA_{\text{mean3}},\ \forall Adv_1, Adv_2, Adv_3 \in \mathcal{P}(X).$$

因此 $Adv_1 \succcurlyeq_{(L_p+STA)} Adv_3$, 故 $\succcurlyeq_{(L_p+STA)}$ 满足传递性.

对于 $\succcurlyeq_{(L_p+STA)}$, 因为有理数上的 $\langle \mathbb{Q}, \leqslant \rangle$ 满足完全性, 所以 $STA_{\text{mean1}} \leqslant STA_{\text{mean2}} \vee STA_{\text{mean2}} \leqslant STA_{\text{mean1}}$.

因此 $Adv_1 \succcurlyeq_{(L_p+STA)} Adv_2 \vee Adv_2 \succcurlyeq_{(L_p+STA)} Adv_1$, 故 $\succcurlyeq_{(L_2+STA)}$ 满足完全性.

综上, $\succcurlyeq_{(L_p+STA)}$ 是满足完全性的预序关系.

对于 $\succcurlyeq_{(L_\infty+SR)}$, 当阈值为 $Thr_\infty \in Gr$ 且共有 N_d 个对抗样本时, 令

$$N_{d1} = \mathbb{I}(x' \in Adv \wedge \|x' - x\|_\infty \leqslant Thr_\infty),\ N_{d2} = \mathbb{I}(x' \in Adv \wedge \|x' - x\|_\infty \leqslant Thr_\infty \wedge F(x') \neq F(x)),$$

其中, N_d, N_{d1}, N_{d2} 三者关系满足 $0 \leqslant N_{d2} \leqslant N_{d1} \leqslant N_d$. 当阈值 $Thr'_\infty \neq Thr_\infty$ 时, 有 $N'_{d1} \neq N_{d1} \wedge \dfrac{N_{d2}}{N_d} \neq \dfrac{N'_{d2}}{N_d} \Rightarrow Adv \not\succcurlyeq_{(L_\infty+SR)} Adv$.

因此 $\succcurlyeq_{(L_\infty+SR)}$ 不满足自反性. 同理可证 $\succcurlyeq_{(L_p+SR)}$, $p \in \{0,1,2\}$ 不满足自反性.

对于 $Adv_1, Adv_2, Adv_3 \in \mathcal{P}(X)$, 假设阈值 $Thr_\infty = Gr_{\max}$, 有:

$$0 < SR_1 < SR_2 < SR_3 \Rightarrow Min(Adv_1) < Max(Adv_1) < Min(Adv_2) < Max(Adv_2) < Min(Adv_3) < Max(Adv_3),$$

其中, $Min(Adv)$, $Max(Adv)$ 分别表示对抗样本集合 Adv 的最大和最小噪声幅度.

当噪声阈值取 $Thr_\infty = Max(Adv_1)$ 时, $0 = SR_3 < SR_1$.

因此 $\succcurlyeq_{(L_p+SR)}$, $p \in \{0,1,2,\infty\}$ 不满足传递性. 这种需要在特定阈值下讨论和比较成功率的特性也决定了 $\succcurlyeq_{(L_p+SR)}$, $p \in \{0,1,2,\infty\}$ 不满足完全性.

当使用正确类别平均置信度作为度量指标时, 当 $ACTC_1, ACTC_2 < 0.5$ 时, 可能存在 $ACTC_1 < ACTC_2$ 但 $ACTC_1$ 对应对抗样本集合全部错分, $ACTC_2$ 对应对抗样本集合全部分类正确的情况. 因此正确类别平均置信度对应的“优于或等于”关系不满足完全性.

另外值得注意的是, 由于对抗样本集合与成功率或噪声统计量之间都不是双射, 指标体系中 4 种关系都不满足反对称性. 噪声统计量或攻击成功率相等的两个对抗样本集合可能是由完全不同的对抗样本组成的. 因此 $\succcurlyeq_{(L_2+STA)}$ 和 $\succcurlyeq_{(L_\infty+STA)}$ 并不是偏序关系.

定理 3 说明, 当使用噪声统计量作为度量指标时, 任意两个对抗样本集合反映的模型对抗鲁棒性都可进行比较. 但使用攻击成功率作为度量指标时, 如果设定的噪声阈值发生变化, 即使是相同的对抗样本集合的成功率也可能发生变化, 因此由攻击成功率反映的对抗鲁棒性是不能直接进行对比的. 攻击成功率对噪声阈值的依赖影响了其对抗鲁棒性评估结果的独立性. 反映在扰动-准确率曲线上, 两种攻击方法在相同目标模型上可能出现曲线交叉的情况. 同时, 由于正确类别平均置信度反映的只是不同类别置信度的相对情况, 而非噪声的绝对攻击能力, 因此只能作为参考值, 例如对高置信度错分的对抗样本进行筛选. 综上, 使用噪声统计量作为度量指标时任意两个对抗样本集合之间都是可比的, 同时鲁棒性评估的结果满足独立性.

3.4 对抗鲁棒性评价指标约束下的攻防方法设计

上述分析分别从评价指标定义域的包含关系、鲁棒性描述粒度以及鲁棒性评估序关系这 3 方面分析了不同对抗鲁棒性评价指标间的差异. 这些差异只针对评价指标本身, 并不针对具体的对抗攻防方法. 然而, 作为对抗鲁棒性评估和提升的手段, 攻击和防御方法分别在不同的对抗鲁棒性评价指标下受到显著的规范与约束作用.

首先, L_∞ 范数+成功率的评价指标会诱导迭代攻击生成各维噪声绝对值逼近噪声阈值的对抗样本. 根据上文的讨论, 使用 L_∞ 作为范数意味着对单一对抗样本只考虑噪声绝对值最大的维度, 使用成功率作为度量指标意味着需要指定噪声阈值. 在这种情况下, 以 IFGSM 和 PGD[31]为代表的迭代攻击方法通常会在生成对抗样本时直接使用符号函数处理反向传播得到的梯度. 例如, PGD 对对抗样本单步的更新为:

$$z^{t+1} = \prod_{\varepsilon}(z^t + \alpha \cdot sign(\nabla \ell(h(x+z^t), y))) \tag{17}$$

其中 z^t 表示经过 t 步更新的对抗噪声, α 表示迭代攻击的步长, $sign()$ 表示符号函数, $\nabla \ell(h(x+z^t))$ 表示当原始输入为 x, 对抗噪声为 z^t 时, 对目标模型最后一层 h 的输出计算损失函数经过反向传播后得到的梯度值. $\prod_{Thr_\infty} z$ 表示将 z 投影到半径为 Thr_∞ 的 N 维球面上的操作:

$$\prod_{Thr_\infty} z_i = \begin{cases} \varepsilon \cdot sign(z_i) & |z_i| > Thr_\infty \\ z_i & |z_i| \leqslant Thr_\infty \end{cases} \tag{18}$$

由等式 (18) 可以看出, 对抗样本单步的更新只与梯度值的符号有关, 与梯度值的数值无关. 也就是说, 无论单步计算得到的梯度值是多少, 在投影前对抗样本每个维度的更新的值都是 α, $-\alpha$ 或 0. 梯度值本身反映了损失函数关于输入数据不同维度上的变化率. 梯度较高的维度可能相对敏感, 添加相同的量的噪声更有可能实现错分. 然而, 在 L_∞ 范数+成功率的评价指标下, 攻击方无需考虑噪声敏感性的差异, 只要保证噪声绝对值最大的元素不超过噪声阈值即可. 这种情况下攻击方法性能提升的关键不再是如何找到数据中敏感的元素, 在较少的噪声幅度下实现错分, 而是如何在限定的噪声阈值内尽可能多加噪声, 保证目标模型错分. 由定理 1 可知, 这时 L_∞ 范数+成功率的评价指标定义域进一步缩小为 $\mathcal{P}(\hat{X}), \hat{X} = \{x' \mid x' \in X \wedge \|x'-x\|_\infty = Thr_\infty\}$. 当迭代步数 $t=1$ 时, $z = Thr_\infty \cdot sign$

$(\nabla\ell(h(x),y)))$, 即每个维度的噪声幅度都等于噪声阈值. 为了尽可能提升攻击成功率, 攻击方法的最优策略是保持每个对抗样本的噪声幅度都维持或接近噪声阈值. 这严重限制了攻击方法的设计, 迭代过程中每一步可以添加噪声的方向的数量由 $|Gr|^N$ 显著降低为 3^N.

其次, L_∞ 范数+成功率的对抗鲁棒性评价指标同样会约束防御方法, 尤其是以对抗训练[17]为代表的主动防御方法的设计. 对抗训练本质上是将对抗样本加入目标模型训练数据, 从而扩展目标模型以原始数据为中心的鲁棒半径[6]. 然而在 L_∞ 范数+成功率的评价指标下生成的对抗样本集中在 $\|x'-x\|_\infty = Thr_\infty$ 的局部区域, 这使得对抗训练无法有效覆盖其他输入空间[32], 进而导致对抗训练后的模型只对特定 L_∞ 范数+成功率的评价指标下的对抗攻击方法表现出较强的对抗鲁棒性. 另外, 由于用于对抗训练的对抗样本噪声幅度与噪声阈值 Thr_∞ 直接相关, 在一种噪声阈值下对抗训练后的目标模型缺乏对另一种噪声阈值下生成的对抗样本的对抗鲁棒性[22]. 因此, 使用基于 L_∞ 范数+成功率评价指标的对抗样本进行对抗训练, 并在相同噪声阈值下的基于 L_∞ 范数+成功率的攻击方法上进行评测, 无法准确且全面地评估对抗训练方法对目标模型在面对潜在的对抗样本时对抗鲁棒性的提升.

另外, 使用攻击成功率作为度量指标还会给对抗鲁棒性评估的实验验证带来额外的困难. 为了全面比较两个对抗攻击或对抗防御方法的性能, 实验者需要比较不同方法在多种不同的噪声阈值下的攻击成功率. 考虑到对抗鲁棒性评估的实验验证中还有迭代步数、查询次数、步长等多个不同的超参数, 充分比较两个方法需要付出的计算量和展示难度都显著增加了. 因此, 从对抗鲁棒性评估计算效率角度而言, 使用噪声统计量作为度量指标可避免使用成功率导致的对不同噪声阈值下评估结果的重复计算, 减小鲁棒性评估的计算成本.

4 评估指标体系完备性的实证验证

4.1 实验设置

本节针对第 3 节中对抗鲁棒性评估的指标体系完备性分析结果, 在 4 个不同尺寸的公开图像分类数据集、1 个无线信号室内定位数据集和 1 个语音识别数据集上, 对 23 个模型和 20 种不同的对抗攻击方法进行验证. 表 1 展示了 4 个图像分类数据集和 1 个语音识别数据集的具体信息, 包括数据尺寸、类别数、数据量及目标模型数量. 所有 6 个数据集上的 17 个模型及其编号对应关系如表 2 所示. 针对图像分类任务, 使用的数据集为 CIFAR-10[33]、Tiny-ImageNet[34]、ImageNet[35]及 ImageNet-21K[36]. 在针对图像分类任务使用的 17 个模型中, 10 个模型结构基于卷积神经网络, 7 个模型结构基于 ViT[37]. Tiny-ImageNet 数据集上的 VGG19_adv 表示该模型训练过程中使用了对抗训练方法. 图像每个像素的取值范围为 $Gr=[0,255]$. 另外, 为了方便查看与比较, 在计算 L_2 范数下的噪声幅度时统一将像素取值范围等比映射到 $Gr=[0,1]$, 并不影响评价指标定义域或鲁棒性评估序关系.

实验部分同时考虑白盒和黑盒两种攻击场景. 攻击方法可分为两类. 第 1 类是基于迁移的攻击方法, 即使用白盒攻击在替代模型上生成对抗样本, 并在目标模型上进行评估. 如果替代模型和目标模型一致即为白盒攻击, 否则为黑盒攻击. 第 2 类是基于决策的攻击方法, 这类方法不需要替代模型, 直接通过查询目标模型生成对抗样本并压缩对抗噪声. 第 1 类攻击方法包括 FGSM[16], MIFGSM[23], IFGSM[24], VRIGSM[38], DDN[39], DeepFool[40], EAD[41], NEWTON[42]和 CW[20], 第 2 类攻击方法包括 SIGN_OPT[43], HSJA[44], WHEY[45], BOUNDARY[15], PAR[46], BBA[47], Evolutionary (EVO)[48]和 CISA[49].

表 1 实验数据集

数据集名称	数据尺寸	类别数	数据量及来源	模型数量
CIFAR-10	32×32×3	10	测试集 10000张	2
Tiny-ImageNet	64×64×3	200	测试集 2000张	4
ImageNet	224×224×3	1000	验证集 10000张	7
ImageNet-21K	224×224×3	21000	测试集 21000张	4
Mozilla Common Voice	9×16000	N/A	验证集 150000段	2

表 2 数据集模型编号与结构对应关系

数据集名称	模型编号及名称						
CIFAR-10	1 VGG16				2 ResNet18		
Tiny-ImageNet	1 ResNet-101		2 Inception v3		3 Inception-ResNet v2		4 VGG19_adv
ImageNet	1 ResNet-101	2 DenseNet-161	3 VGG19	4 SENet-154	5 vit-small-r26-s32-224	6 vit-small	7 vit-tiny-PAR16-224
ImageNet-21K	1 vit-small-r26-s32-224		2 vit-large-PAR16-224		3 vit-tiny-PAR16-224		4 r50-s32
Mozilla Common Voice	1 WaveNet				2 Mozilla DeepSpeech		
UJIIndoorLoc	1 CNN		2 DNN		3 CiFi		4 Pixeldp_cnn

为了验证第 3 节的定理在不同数据与任务上的有效性, 本文在 Mozilla Common Voice 语音识别数据集[50]上进一步进行了验证. 音频数据采样后 Waveform 数据取值范围为 $Gr=[-32768,32768]$. 为了方便查看与比较, 在计算 L_2 范数下的噪声幅度时统一将像素取值范围等比映射到 $Gr=[0,1]$. 针对语音识别的对抗攻击使用 CW[20], Universal Attack[51], Qin-I[52]和 Qin-R[52], 使用了 WaveNet 和 Mozilla DeepSpeech 两种目标模型, 如表 2 所示. 语音识别任务攻击成功的标准依 Universal Attack 的通用标准, 即转录为文本后 CER 超过 50% 即为攻击成功[51]. 另外, 为了进一步验证第 3 节的定理在 L_0 和 L_1 范数上的有效性, 本文在 UJIIndoorLoc 无线信号室内定位数据集[44]上进一步进行了验证. 数据取值范围 $Gr=[-104,100]$, 数值范围等比映射到 $Gr=[0,1]$. 针对信号定位的攻击方法采用 CW[20]和 L-BFGS, 分别使用 L_0 和 L_1 范数作为攻击的优化目标.

本文针对“L_2 范数+成功率”“L_∞ 范数+成功率”“L_2 范数+噪声统计量”“L_∞ 范数+噪声统计量”4 种不同的对抗鲁棒性评价指标进行比较. 实验结果统计了每种攻击方法在不同的替代模型与目标模型组合上生成的对抗样本集合的成功率及对应范数上的噪声统计量. 需要注意的是, 由于对抗攻击方法初始设计的差异, 不同的攻击方法适用的评价指标是不同的. 例如, IFGSM 可以通过调整梯度计算与截断的方式适用于所有 4 种评价指标, 但 Boundary 不适用 L_∞ 范数, 且无法设定噪声阈值, 因此只适用于“L_2 范数+噪声统计量”一种评价指标. 为了攻击方法在指标体系下公平比较, 实验部分根据是否设定噪声阈值将评估方式分为两种. 在不设定噪声阈值的情况下, 对抗样本的空间即为原始输入空间, 所有攻击方法可对一张原始图像至多查询目标模型 1 000 次并将噪声幅度最小的对抗样本作为最终输出. 这时除 PAR[46]和 CISA[49]有特殊的噪声初始化方法外, 其余所有基于决策的攻击方法都是用相同的基于高斯噪声的对抗样本进行初始化. 在设定噪声阈值的情况下, 攻击方法无法对目标模型进行查询, 且需要在对应范数选择上设定的噪声阈值内生成对抗样本. 最终对每种攻击方法在整个数据集上生成的所有对抗样本的成功率与噪声统计量 (均值与中位数) 进行记录. 下面所有实验结果根据等式 (3)–等式 (5) 分别计算攻击成功率、噪声均值以及中位数. 在计算噪声均值和中位数时, 对于对抗样本集合中没有错分的对抗样本使用一个较大的噪声幅度作为惩罚项. 在 L_2 范数下攻击失败的对抗样本噪声幅度计为 100, L_∞ 范数下计为 80 (对语音识别数据, L_2 范数惩罚项为 800, L_∞ 范数惩罚项为 600). 这两个值远大于对应范数下攻击成功的对抗样本的噪声幅度, 可用于惩罚攻击失败的情况.

4.2 对抗鲁棒性评估载体完备性分析

表 3–表 13 展示了在设定/不设定噪声阈值的两种评估方式下, 不同的攻击方法在 L_2 / L_∞ 范数下对 6 个数据集生成对抗样本的实验结果. 其中表 3–表 7 以及表 11 不设定噪声阈值, 表 8–表 10 以及表 12 设定噪声阈值.表 3、表 4、表 8、表 9 为 L_2 范数下的对抗样本, 其余为 L_∞ 范数下的对抗样本. 表 13 展示的是 L_0 和 L_1 范数. 表格中每条数据包括攻击方法、替代模型编号、中位数与均值两种噪声统计量以及成功率. 表 3–表 7 以及表 11 中每条数

据标注了目标模型, 表 8 中目标模型编号为 7, 表 9 中目标模型编号为 4, 表 10 中目标模型编号为 4. 受限于篇幅, 实验部分并未展示所有数据集、所有范数、所有目标模型与替代模型组合的结果, 但表 3–表 13 展示的数据已经有足够的代表性. 第 4.3 与第 4.4 节展示的涉及鲁棒性描述粒度与鲁棒性评估序关系的元数据是在所有数据集、所有范数、所有目标模型与替代模型的所有组合中整理得到的.

表 3　不设定噪声阈值, L_2 范数下 CIFAR-10 数据集实验结果

攻击方法	目标模型	替代模型	中位数	均值	成功率 (%)	攻击方法	目标模型	替代模型	中位数	均值	成功率 (%)
FGSM	1	1	0.275	1.804	99.4	FGSM	2	1	1.233	3.972	98.7
FGSM	1	2	1.699	11.377	90.9	FGSM	2	2	0.187	7.609	93.3
MIFGSM	1	1	0.166	0.192	100.0	MIFGSM	2	1	0.479	0.626	100.0
MIFGSM	1	2	0.563	0.697	100.0	MIFGSM	2	2	0.13	0.143	100.0
IFGSM	1	1	0.139	0.19	100.0	IFGSM	2	1	0.422	0.566	100.0
IFGSM	1	2	0.523	1.013	100.0	IFGSM	2	2	0.112	0.334	100.0
VRIGSM	1	1	0.162	0.188	100.0	VRIGSM	2	1	0.438	0.544	100.0
VRIGSM	1	2	0.504	0.594	100.0	VRIGSM	2	2	0.128	0.14	100.0
DDN	1	1	0.136	0.139	100.0	DDN	2	1	0.47	0.54	100.0
DDN	1	2	0.565	0.642	100.0	DDN	2	2	0.106	0.111	100.0
EAD	1	1	0.153	0.175	100.0	EAD	2	1	0.408	0.51	100.0
EAD	1	2	0.495	0.597	100.0	EAD	2	2	0.114	0.126	100.0
NEWTON	1	1	0.295	0.287	100.0	NEWTON	2	1	100	60.871	39.3
NEWTON	1	2	100	79.974	20.1	NEWTON	2	2	0.272	0.245	100.0
CW	1	1	0.133	0.155	100.0	QEBA	1	N/A	1.488	1.727	100.0
SIGN_OPT	1	N/A	2.369	2.7	100.0	SIGN_OPT	2	N/A	2.245	2.686	99.8
HSJA	1	N/A	1.746	2.024	100.0	HSJA	2	N/A	1.591	2.012	99.8
WHEY	1	N/A	1.284	1.489	100.0	WHEY	2	N/A	1.16	1.452	99.8
BOUNDARY	1	N/A	0.972	1.184	100.0	BOUNDARY	2	N/A	0.851	1.178	99.8
PAR	1	N/A	0.586	0.665	100.0	PAR	2	N/A	0.529	0.783	99.8
BBA	1	N/A	0.89	1.1	100.0	BBA	2	N/A	0.706	1.037	99.8
EVO	1	N/A	2.044	2.181	100.0	EVO	2	N/A	1.96	2.349	99.8
CISA	1	1	0.208	0.239	100.0	CISA	2	1	0.344	0.558	99.8
CISA	1	2	0.415	0.574	100.0	CISA	2	2	0.16	0.313	100.0

表 4　不设定噪声阈值, L_2 范数下 ImageNet 数据集实验结果

攻击方法	目标模型	替代模型	中位数	均值	成功率 (%)	攻击方法	目标模型	替代模型	中位数	均值	成功率 (%)
FGSM	2	1	9.995	19.584	99.2	FGSM	4	3	12.546	23.254	96.6
FGSM	6	5	38.063	49.794	100.0	FGSM	1	7	16.785	23.872	100.0
MIFGSM	2	1	3.538	4.954	100.0	MIFGSM	4	3	7.237	10.919	100.0
MIFGSM	6	5	9.797	15.992	100.0	MIFGSM	1	7	10.168	12.725	100.0
IFGSM	2	1	3.138	4.151	100.0	IFGSM	4	3	6.41	10.132	100.0
IFGSM	6	5	9.107	12.931	100.0	IFGSM	1	7	7.71	9.646	100.0
VRIGSM	2	1	2.94	4.277	100.0	VRIGSM	4	3	5.868	8.928	100.0
VRIGSM	6	5	8.081	12.559	100.0	VRIGSM	1	7	8.076	10.431	100.0
DDN	2	1	3.9	6.12	100.0	DDN	4	3	9.452	13.083	100.0
DDN	6	5	12.995	24.034	100.0	DDN	1	7	11.043	14.326	100.0
DeepFool	2	1	100	64.56	35.8	DeepFool	4	3	100	77.408	23.0
DeepFool	6	5	100	81.647	18.9	DeepFool	1	7	100	58.71	46.3
EAD	2	1	2.843	4.184	100.0	EAD	4	3	7.531	9.847	100.0
EAD	6	5	9.953	14.711	100.0	EAD	1	7	7.308	10.634	100.0
NEWTON	2	1	100	54.119	46.5	NEWTON	4	3	100	75.023	25.3
NEWTON	6	5	100	86.84	13.2	NEWTON	1	7	100	79.235	21.1

表 4 不设定噪声阈值, L_2 范数下 ImageNet 数据集实验结果 (续)

攻击方法	目标模型	替代模型	中位数	均值	成功率 (%)	攻击方法	目标模型	替代模型	中位数	均值	成功率 (%)
CW	2	1	6.392	50.445	50.2	QEBA	1	7	27.339	29.056	100.0
SIGN_OPT	2	N/A	35.846	40.355	99.9	SIGN_OPT	4	N/A	52.066	49.825	100.0
SIGN_OPT	6	N/A	60.838	56.536	98.3	SIGN_OPT	1	N/A	36.25	36.958	100.0
HSJA	2	N/A	28.663	34.216	99.9	HSJA	4	N/A	44.255	43.859	100.0
HSJA	6	N/A	52.093	51.796	98.3	HSJA	1	N/A	30.003	32.126	100.0
WHEY	2	N/A	22.15	35.075	99.9	WHEY	4	N/A	41.716	44.846	100.0
WHEY	6	N/A	53.293	53.31	98.3	WHEY	1	N/A	23.526	30.449	100.0
BOUNDARY	2	N/A	16.352	22.686	99.9	BOUNDARY	4	N/A	27.782	30.332	100.0
BOUNDARY	6	N/A	27.999	34.028	98.3	BOUNDARY	1	N/A	16.984	20.799	100.0
PAR	2	N/A	4.095	8.25	99.9	PAR	4	N/A	6.089	9.907	100.0
PAR	6	N/A	7.041	17.826	98.3	PAR	1	N/A	4.564	7.411	100.0
BBA	2	N/A	12.214	19.423	99.9	BBA	4	N/A	21.549	25.27	100.0
BBA	6	N/A	25.409	32.322	98.3	BBA	1	N/A	11.337	15.705	100.0
EVO	2	N/A	14.068	17.591	99.9	EVO	4	N/A	22.149	21.847	100.0
EVO	6	N/A	28.155	29.621	98.3	EVO	1	N/A	15.201	16.893	100.0
CISA	2	1	2.393	4.003	99.9	CISA	4	3	5.576	8.467	100.0
CISA	6	5	8.064	12.042	98.1	CISA	1	7	5.101	7.739	100.0

表 5 不设定噪声阈值, L_∞ 范数下 Tiny-ImageNet 数据集实验结果

攻击方法	目标模型	替代模型	中位数	均值	成功率 (%)	攻击方法	目标模型	替代模型	中位数	均值	成功率 (%)
FGSM	2	1	11	15.657	100.0	FGSM	3	2	13	14.848	100.0
FGSM	4	3	23	21.764	100.0	FGSM	1	4	4	7.833	100.0
MIFGSM	2	1	14	26.089	100.0	MIFGSM	3	2	17	24.304	100.0
MIFGSM	4	3	40	51.887	100.0	MIFGSM	1	4	4	6.265	100.0
IFGSM	2	1	10	16.195	100.0	IFGSM	3	2	12	14.81	100.0
IFGSM	4	3	26	29.066	100.0	IFGSM	1	4	3	4.682	100.0
VRIGSM	2	1	9	19.396	100.0	VRIGSM	3	2	12	19.266	100.0
VRIGSM	4	3	33	43.33	100.0	VRIGSM	1	4	4	5.682	100.0
DeepFool	2	1	19	39.4	80.5	DeepFool	3	2	55	43.884	59.5
DeepFool	4	3	80	60.39	24.5	DeepFool	1	4	3	9.236	93.9
HSJA	1	N/A	4	10.629	100.0	HSJA	2	N/A	12	18.056	100.0
HSJA	3	N/A	20	21.601	100.0	HSJA	4	N/A	21	23.399	100.0
SIGN_OPT	1	N/A	19	44.028	100.0						

表 6 不设定噪声阈值, L_∞ 范数下 ImageNet-21K 数据集实验结果

攻击方法	目标模型	替代模型	中位数	均值	成功率 (%)	攻击方法	目标模型	替代模型	中位数	均值	成功率 (%)
FGSM	2	1	10	16.581	100.0	FGSM	3	2	6	10.388	100.0
FGSM	4	3	8	15.677	100.0	FGSM	1	4	17	25.638	100.0
MIFGSM	2	1	10	19.081	100.0	MIFGSM	3	2	3	6.318	100.0
MIFGSM	4	3	6	7.181	100.0	MIFGSM	1	4	10	16.362	100.0
IFGSM	2	1	9	12.651	100.0	IFGSM	3	2	2	5.424	100.0
IFGSM	4	3	4	5.906	100.0	IFGSM	1	4	8	12.524	100.0
VRIGSM	2	1	7	15.163	100.0	VRIGSM	3	2	3	5.247	100.0
VRIGSM	4	3	4	6.197	100.0	VRIGSM	1	4	7	13.486	100.0
DeepFool	2	1	80	70.783	11.6	DeepFool	3	2	4	24.302	75.3
DeepFool	4	3	13	40.48	51.2	DeepFool	1	4	80	74.146	7.6
HSJA	1	N/A	17	27.58	100.0	HSJA	2	N/A	8	15.676	100.0
HSJA	3	N/A	21	31.388	100.0	HSJA	4	N/A	14	27.853	100.0

表 7　不设定噪声阈值, L_∞ 范数下 ImageNet 数据集实验结果

攻击方法	目标模型	替代模型	中位数	均值	成功率 (%)	攻击方法	目标模型	替代模型	中位数	均值	成功率 (%)
FGSM	2	1	14	20.486	100.0	FGSM	4	3	19	23.947	100.0
FGSM	6	5	52	48.15	100.0	FGSM	1	7	20	22.131	100.0
MIFGSM	2	1	8	13.68	100.0	MIFGSM	4	3	22	37.37	100.0
MIFGSM	6	5	41	53.512	100.0	MIFGSM	1	7	29	42.723	100.0
IFGSM	2	1	7	9.637	100.0	IFGSM	4	3	15	21.381	100.0
IFGSM	6	5	31	33.079	100.0	IFGSM	1	7	19	24.38	100.0
VRIGSM	2	1	7	11.018	100.0	VRIGSM	4	3	16	28.809	100.0
VRIGSM	6	5	33	40.654	100.0	VRIGSM	1	7	25	33.23	100.0
DeepFool	2	1	27	41.716	50.7	DeepFool	4	3	80	61.032	25.5
DeepFool	6	5	80	70.26	19.7	DeepFool	1	7	80	76.335	52.6
HSJA	1	N/A	32	35.719	100.0	HSJA	2	N/A	36	38.039	100.0
HSJA	3	N/A	20	23.691	100.0	HSJA	7	N/A	27	31.872	100.0

表 8　设定噪声阈值, L_2 范数下 ImageNet 数据集实验结果

攻击方法	Thr_2	替代模型	中位数	均值	成功率 (%)	替代模型	中位数	均值	成功率 (%)	替代模型	中位数	均值	成功率 (%)
MIFGSM	1	1	0.910	0.982	99.4	2	0.910	1.144	99.1	3	0.896	1.047	99.4
		4	0.932	24.427	64.4	5	0.911	11.435	83.8	6	0.917	8.758	87.8
	4	1	3.792	2.378	100.0	2	3.810	2.375	100.0	3	3.747	3.166	99.4
		4	3.972	6.667	94.9	5	3.876	3.489	99.0	6	3.758	2.465	100.0
	16	1	14.604	9.242	100.0	2	14.487	9.174	100.0	3	14.624	10.588	100.0
		4	15.672	11.296	100.0	5	15.041	10.725	100.0	6	14.459	9.427	100.0
IFGSM	1	1	0.914	0.984	99.4	2	0.913	1.146	99.1	3	0.903	1.051	99.4
		4	0.919	22.726	66.9	5	0.911	8.737	87.9	6	0.921	8.526	88.1
	4	1	3.813	2.372	100.0	2	3.818	2.370	100.0	3	3.785	3.158	99.4
		4	3.868	3.420	99.2	5	3.842	2.707	100.0	6	3.755	2.453	100.0
	16	1	14.476	9.092	100.0	2	14.418	9.056	100.0	3	14.483	10.346	100.0
		4	15.363	11.128	100.0	5	15.102	10.740	100.0	6	14.543	9.358	100.0
VRIGSM	1	1	0.916	0.985	99.4	2	0.919	0.957	99.4	3	0.904	1.052	99.4
		4	0.919	21.595	68.6	5	0.910	6.714	90.9	6	0.915	7.126	90.2
	4	1	3.803	2.369	100.0	2	3.799	2.367	100.0	3	3.796	3.160	99.4
		4	3.800	2.743	100.0	5	3.849	2.715	100.0	6	3.770	2.458	100.0
	16	1	14.151	8.877	100.0	2	14.288	8.887	100.0	3	14.177	10.133	100.0
		4	15.037	10.897	100.0	5	15.101	10.699	100.0	6	14.421	9.287	100.0

表 9　设定噪声阈值, L_2 范数下 ImageNet-21K 数据集实验结果

攻击方法	Thr_2	替代模型	中位数	均值	成功率 (%)	替代模型	中位数	均值	成功率 (%)	替代模型	中位数	均值	成功率 (%)
MIFGSM	1	1	0.962	4.375	94.9	2	0.972	0.552	100.0	3	0.971	18.224	76.7
	4	1	3.698	2.676	99.4	2	3.966	2.193	100.0	3	3.892	9.004	91.4
	16	1	14.263	8.881	100.0	2	15.636	8.788	100.0	3	14.927	9.110	99.4
IFGSM	1	1	0.918	1.490	98.7	2	0.985	0.552	100.0	3	0.950	16.357	79.1
	4	1	3.733	2.135	100.0	2	3.907	2.166	100.0	3	3.792	5.064	96.3
	16	1	14.490	8.804	100.0	2	15.353	8.526	100.0	3	14.571	8.623	100.0
VRIGSM	1	1	0.907	1.974	98.1	2	0.967	0.553	100.0	3	0.958	14.961	81.0
	4	1	3.760	2.120	100.0	2	3.911	2.161	100.0	3	3.761	4.539	96.9
	16	1	14.871	8.809	100.0	2	15.171	8.375	100.0	3	14.778	8.584	100.0

表 10 设定噪声阈值, L_∞ 范数下 Tiny-ImageNet 数据集实验结果

攻击方法	Thr_∞	替代模型	中位数	均值	成功率 (%)	替代模型	中位数	均值	成功率 (%)
MIFGSM	8	1	6	3.814	100.0	2	7	4.032	99.4
		3	7	5.453	93.9	4	5	2.781	99.4
	16	1	12	6.872	100.0	2	12	6.985	100.0
		3	12	7.245	97.8	4	8	4.771	99.7
	32	1	24	14.030	100.0	2	25	13.847	100.0
		3	25	13.619	99.6	4	18	10.426	100.0
IFGSM	8	1	6	3.946	100.0	2	7	4.317	99.4
		3	8	5.736	94.8	4	5	2.778	99.4
	16	1	12	7.727	100.0	2	14	8.353	100.0
		3	14	8.486	99.1	4	9	4.889	99.7
	32	1	28	16.547	100.0	2	30	16.849	100.0
		3	30	16.809	100.0	4	19	10.904	100.0
VRIGSM	8	1	6	4.109	100.0	2	7	4.712	99.4
		3	8	5.478	96.1	4	5	2.757	99.4
	16	1	13	8.106	100.0	2	14	8.691	100.0
		3	15	9.164	99.6	4	9	5.105	99.7
	32	1	30	18.022	100.0	2	31	17.951	100.0
		3	30	17.387	100.0	4	19	11.207	100.0

表 11 不设定噪声阈值, Mozilla Common Voice 数据集实验结果

攻击方法	目标模型	替代模型	中位数	均值	成功率 (%)	攻击方法	目标模型	替代模型	中位数	均值	成功率 (%)
Universal	1	2	2.550	4.783	100.0	Universal	1	2	134	127.683	100.0
Universal	2	1	4.942	8.183	100.0	Universal	2	1	302	258.778	100.0
Carlini	1	2	2.704	3.878	100.0	Carlini	1	2	80	91.397	100.0
Carlini	2	1	5.504	9.049	100.0	Carlini	2	1	211	210.604	100.0
Qin-I	1	2	3.129	6.228	100.0	Qin-I	1	2	181	172.386	100.0
Qin-I	2	1	8.148	13.298	100.0	Qin-I	2	1	305	255.642	100.0
Qin-R	1	2	3.39	5.513	100.0	Qin-R	1	2	174	159.931	100.0
Qin-R	2	1	6.385	9.851	100.0	Qin-R	2	1	392	337.815	100.0

从评估载体完备性的角度, 表 8–表 10、表 12 的实验结果和其他表格的评价指标定义域存在明显差异. 由于表 3–表 7 以及表 11、13 不设定噪声阈值, 因此可以认为 L_∞ 范数下噪声阈值为取值范围最大值 $Gr_{\max}$, L_2 范数下噪声阈值为 $\sqrt{Gr_{\max}^2 N}$. 由于表 8–表 10、表 12 中每条实验结果都设定了噪声阈值, 因此其中攻击成功率为 100% 结果的噪声统计量都低于对应的噪声阈值. 根据第 3.1 节中的讨论, 在使用成功率作为评价指标时, 如果噪声阈值低于最大值, 则实际是在整个输入空间的一个子集对比不同对抗样本集合的成功率. 对比相同数据集、相同范数选择下设定与不设定噪声阈值的结果, 可以观察到以下几方面现象.

(1) 选择成功率作为评价指标给对抗鲁棒性评估带来了额外的困难. 由于不同的噪声阈值对应不同的输入空间子集, 在一个噪声阈值上生成的对抗样本只能作为反映方法在对应范围内的攻击能力的载体, 噪声阈值不同时相同原始数据生成的对抗样本的攻击成功率也是不可比的, 因此不同攻击方法需要在多个不同的噪声阈值上才能进行系统、全面的对比, 而每个噪声阈值都需要攻击方法完整生成一遍对抗样本. 这展示了使用噪声统计量作为度量指标的一个额外优势, 即无需在不同噪声阈值下对不同的攻击方法进行重复计算, 可显著降低计算量并提升对抗鲁棒性评估的效率.

(2) 并非所有攻击方法都具备在指定输入空间内生成对抗样本的能力. 多数决策攻击都无法主动设定噪声阈值. 由于这些攻击方法的初始对抗噪声多来自高斯噪声, 因此如果使用成功率较低的噪声阈值进行评价, 会出现大量对抗样本噪声幅度高于噪声阈值的情况. 设定噪声阈值以计算攻击成功率的评价指标难以兼顾不同攻击方法间

的公平性与适用性.

(3) 从表 8–表 10、表 12 中可以看出, 噪声阈值影响的不仅仅是噪声幅度的上限, 同时会影响噪声统计量. 各组对抗样本的噪声幅度中位数都较接近噪声阈值. 根据第 3.4 节的分析, 这意味着攻击方法在噪声阈值内尽可能多加噪声. 这导致评价指标定义域进一步向噪声阈值邻域缩小, 作为鲁棒性评估的载体, 每组对抗样本成功率的代表性与对抗鲁棒性评估的完备性进一步减弱.

(4) 如果不看噪声统计量, 只有在特殊的噪声阈值下表 8–表 10 中不同对抗样本的攻击成功率才出现明显差异. 如果选择的噪声阈值过大, 不同攻击方法都能实现接近 100% 的成功率, 无法对对抗鲁棒性高低进行准确且完备的评价和比较.

表 12　设定噪声阈值, Mozilla Common Voice 数据集实验结果

攻击方法	Thr_2	目标模型	中位数	均值	成功率 (%)	攻击方法	Thr_∞	目标模型	中位数	均值	成功率 (%)
Universal	1	1	0.410	14.008	86.2	Universal	100	1	98	398.060	17.0
		2	0.428	3.641	96.6			2	97	313.590	62.5
	2	1	1.629	2.449	98.5		200	1	192	341.460	21.2
		2	1.387	3.117	97.7			2	199	287.805	70.5
	4	1	1.629	2.449	98.5		400	1	378	409.470	53.3
		2	1.629	2.947	98.0			2	395	416.570	78.2
Carlini	1	1	0.407	25.313	74.8	Carlini	100	1	96	459.220	11.9
		2	0.408	26.595	73.5			2	99	342.308	46.2
	2	1	0.410	14.664	85.5		200	1	197	349.900	16.2
		2	0.411	14.334	85.9			2	196	339.806	51.1
	4	1	3.970	4.041	99.0		400	1	386	504.850	32.3
		2	4.871	3.832	99.0			2	377	441.748	70.8
Qin-I	1	1	0.439	34.618	65.5	Qin-I	100	1	99	517.480	11.5
		2	0.431	33.585	66.5			2	95	245.444	46.8
	2	1	0.405	25.305	74.8		200	1	200	366.310	21.4
		2	1.637	1.975	99.0			2	196	320.390	65.6
	4	1	3.298	3.138	97.9		400	1	362	470.820	36.0
		2	3.200	2.588	98.1			2	375	399.563	70.4
Qin-R	1	1	0.406	25.307	74.8	Qin-R	100	1	93	536.89	14.2
		2	0.220	33.042	67.1			2	92	340.473	48.6
	2	1	1.379	36.983	63.6		200	1	198	388.45	23.9
		2	1.377	36.079	64.4			2	193	341.572	52.0
	4	1	3.445	13.616	87.1		400	1	395	444.66	40.5
		2	3.605	12.373	98.5			2	373	409.845	68.6

表 13　不设定噪声阈值, UJIIndoorLoc 数据集实验结果

攻击方法 L_0	目标模型	替代模型	中位数	均值	成功率 (%)	攻击方法 L_1	目标模型	替代模型	中位数	均值	成功率 (%)
CW	1	2	64	41	99.85	CW	1	2	8846	6667	99.70
CW	2	3	25	34	95.80	CW	2	3	8796	6818	97.10
CW	3	4	23	26	99.80	CW	3	4	8757	6429	99.70
CW	4	1	63	39	99.40	CW	4	1	9041	6923	99.50
L-BFGS	1	2	25	36	100.0	L-BFGS	1	2	8954	6842	99.85
L-BFGS	2	3	24	36	94.64	L-BFGS	2	3	8872	7004	99.80
L-BFGS	3	4	68	43	99.90	L-BFGS	3	4	9283	7500	99.40
L-BFGS	4	1	32	35	99.30	L-BFGS	4	1	9007	8095	98.60

4.3　对抗鲁棒性评估描述完备性分析

从鲁棒性描述粒度的角度, 表 3–表 13 的实验结果表明, 选择 L_2 范数对同一个数据集给出的不同取值结果总

数较 L_∞ 范数存在明显优势. 图 2 展示了在 5 个数据集上 L_2 / L_∞ 范数数据总量与不同取值结果的统计. 其中浅蓝/浅红条柱表示 L_2 / L_∞ 范数下一种度量指标比较的对抗样本集合数量, 深蓝/深红条柱表示 L_2 / L_∞ 范数下该度量指标 (中位数、均值、成功率) 共有多少个互不相同的取值结果. 深色条柱在浅色条柱中所占比例越大, 说明该评价指标在对应数据集和范数选择下通过对抗样本集合对鲁棒性的描述的粒度越细.

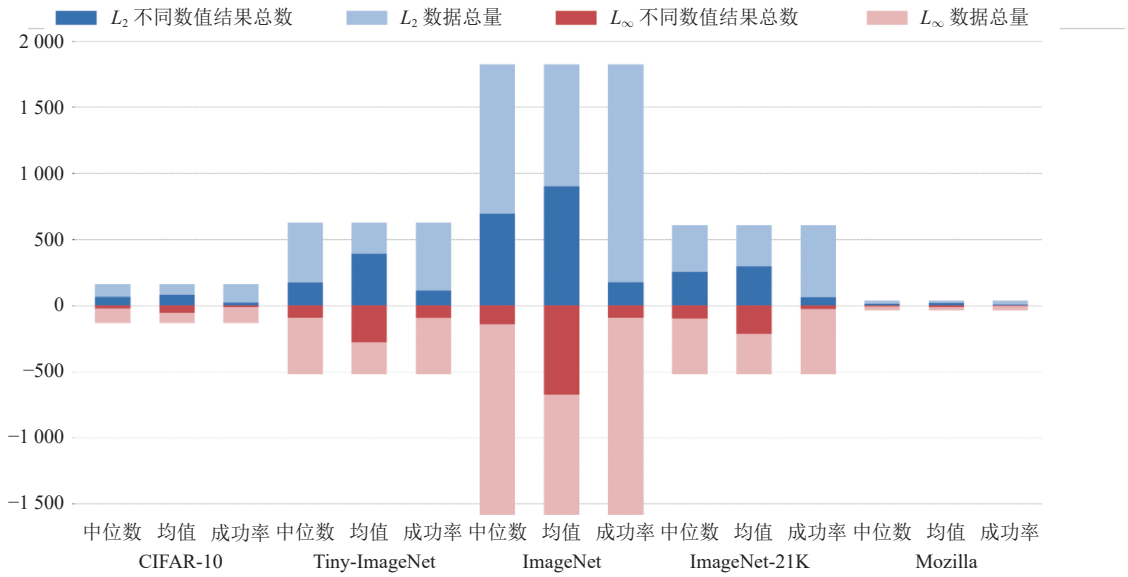

图 2 L_2 与 L_∞ 范数在不同数据集上鲁棒性描述粒度对比

4.4 评估结果与攻击方法独立性分析

在对抗鲁棒性评估结果与攻击方法之间的独立性分析方面, 本文引入逆序数 Inv 用于对比不同评价指标在鲁棒性评估序关系方面的差异. 对于相同一组原始数据, 同一种攻击方法使用同样的评价指标可能在不同的替代模型与目标模型组合上得到不同的结果. 如表 9 中所示, IFGSM 在范数选择为 L_2 、数据集为 ImageNet-21K、噪声阈值为 1 时, 若替代模型为 1 (vit-small-r26-s32-224), 在目标模型 4 (r50-s32) 上取得 98.7% 的成功率, 高于同设定下 VRIGSM 的 98.1%. 然而, 当替代模型改为 3 (vit-tiny-PAR16-224), 其他条件不变时, IFGSM 成功率为 79.1%, 小于 VRIGSM 的 81.0%. 使用相同的范数选择、度量指标、噪声阈值和目标模型, 两个方法在不同的替代模型上性能排序的结果相反, 这给实验验证及方法比较带来了困难. 设 π_1,π_2 是在 2 组不同的替代模型与目标模型组合上根据相同评价指标对所有对抗样本生成方法性能的 2 个排列, 对于编号为 i, j 的任意两种方法, 若两个排列中两种方法的顺序存在 $\pi_1(i)<\pi_2(j)\wedge\pi_1(j)<\pi_2(i)$, 称 (i, j) 是 π_1,π_2 间的一组逆序对, 称 π_1,π_2 间所有逆序对的总数为二者的逆序数. 逆序数评价了在相同的评价指标下各种攻击方法在不同模型上对抗鲁棒性评估结果排序的不符程度. 逆序数越低, 表示不同攻击方法使用该评价指标在两组不同的替代模型-目标模型组合上对抗鲁棒性评估结果的排序越一致, 反之则越混乱.

图 3 展示了各度量指标在 5 个数据集上逆序数的统计. 由于逆序数针对两组排列计算, 而每种评价指标在每个替代模型-目标模型组合上都可得到一个排列, 因此本文以每个排列为基准, 计算该排列与其他所有排列逆序数的和, 以此代表该替代模型-目标模型组合的逆序数. 图 3 中每个数据集上每种评价指标对应的逆序数是一个条柱, 条柱的左端对应逆序数和最小的排列, 右端对应逆序数和最大的排列, 圆点表示所有排列的平均逆序数和. 也就是说, 条柱整体越偏左, 表示该数据集的这一评价指标对不同方法性能高低的评价越趋于一致.

从图 3 中可以看出, 在几乎所有数据集和范数选择上, 使用噪声统计量作为度量指标的逆序数都显著低于成功率. 值得注意的是, 逆序数计算过程中不统计数值相等的逆序对. 也就是说, 如果以成功率为度量指标, 而在某 2

组替代模型-目标模型组合上所有攻击方法都取得 100% 的成功率, 则两个排列的逆序数为 0, 尽管这时事实上无法通过不同的攻击方法区分不同目标模型对抗鲁棒性的高低. 结合第 4.3 节与图 2 可知, 成功率中存在大量重复的数值. 在鲁棒性描述粒度较粗的情况下, 攻击成功率作为一种度量指标仍然在每个数据集上产生了较噪声统计量更高的逆序数. 这意味着攻击成功率不仅给出了大量攻击方法“性能相等”的评价, 而且有较高概率在不同的模型组合上给出不同方法自相矛盾的评价, 这对鲁棒性评估结果与攻击方法的独立性造成了严重的负面影响. 这种情况很大程度上是由攻击成功率在单一对抗样本上二元的评价导致的. 攻击成功率只考虑在对抗样本是否落在噪声阈值内以及是否错分, 并不对噪声幅度的大小进行区分. 但事实上同样是错分, 不同的噪声幅度反映出模型对抗鲁棒性的高低是存在差异的. 攻击成功率作为一种度量指标忽视了这种差异, 这是其评价结果中出现大量相同或矛盾结果的根源. 这也验证了第 3.3 节中的结论, 即噪声统计量作为一种度量指标时评估结果与攻击方法间的独立性.

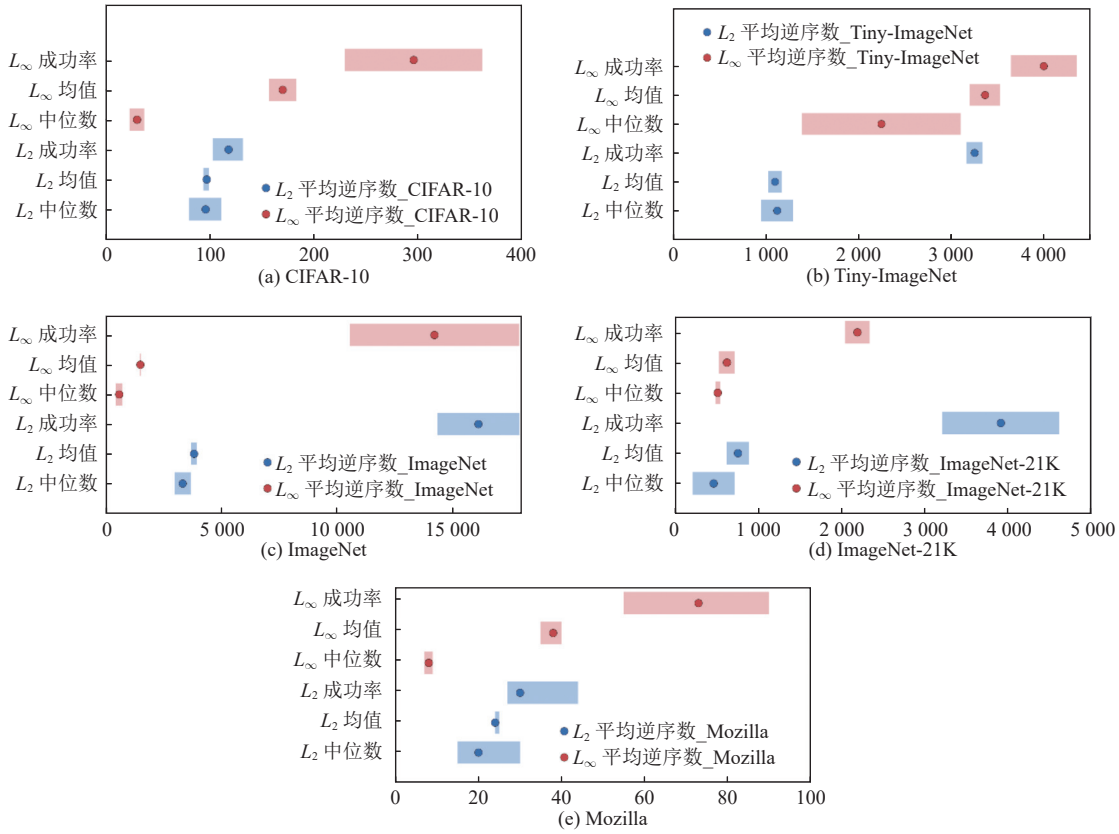

图 3　各度量指标下实验结果逆序数比较

4.5　对抗鲁棒性评估指标体系完备性讨论

通过上述 CIFAR-10、Tiny-ImageNet、ImageNet、ImageNet-21K、Mozilla Common Voice 及 UJIIndoorLoc 数据集上的实验结果与分析可知, 一个完备的对抗鲁棒性评估指标体系需要同时满足载体与描述方面的完备性, 以及评估结果的独立性. 第 4.2 节验证了定理 1 中评价指标定义域的包含关系, 即使用成功率作为评价指标时, 若噪声阈值不设为最大则对抗鲁棒性评估载体并不完备. 由于攻击成功率只在输入空间的一个子集内比较不同的对抗样本集合, 其反映的目标模型对抗鲁棒性本身是局部且片面的. 第 4.3 节对定理 2 中评价指标描述粒度的关系进行了验证, 即 L_∞ 范数鲁棒性评估描述的完备性远小于 L_2 范数. 由于 L_∞ 范数下鲁棒性评估结果的描述粒度与数据维度无关, 因此在高维数据上的实验结果存在大量相同的取值, 导致 L_∞ 范数下相当部分本来存在差异的鲁棒性

评估结果无法区分. 第 4.4 节验证了定理 3 中鲁棒性评估的序关系, 即噪声统计量作为度量指标时鲁棒性评估的结果不依赖特定攻击方法或任何其他元素. 攻击成功率对单一对抗样本的二元评价以及对噪声阈值的依赖导致评估结果独立性极差, 出现了大量不可比或相互矛盾的评价结果. 综上所述, 在本文讨论的由两种范数选择与两种度量指标组成的指标体系中, L_2 范数+噪声统计量是同时满足载体、描述与独立性三个方面的完备性的对抗鲁棒性评价指标.

另外需要注意的是, 本文关于对抗鲁棒性评估指标体系完备性的结论在几乎所有的可使用 L_2 范数与 L_∞ 范数进行度量的输入空间内都是满足的, 即 L_2 范数+噪声统计量的完备性在不同的场景下普适的. 未来有关对抗鲁棒性评估的研究中, 基于 L_2 范数+噪声统计量设计的评价指标在通用性与完备性方面将会有一定的优势.

在本文的基础上, 后续工作可以对其他类型的任务、评价指标以及鲁棒性评估的场景展开更加深入的分析, 进而设计和构建更加完备的对抗鲁棒性评估的评价指标.

5 总 结

对抗鲁棒性评估是使用攻击方法对模型进行鲁棒性评估的标准和规范. 本文对比了鲁棒性评估指标体系中常见的 4 种范数选择 (L_2 , L_∞ , L_0 和 L_1 范数) 与 4 种度量指标 (攻击成功率, 噪声统计量, 平均置信度, 扰动-准确率曲线) 组成的不同的鲁棒性评价指标, 从评价指标定义域的包含关系, 鲁棒性描述粒度, 以及鲁棒性评估序关系等角度针对鲁棒性评估的指标体系完备性展开理论分析. 本文通过比较不同评价指标对应的定义域 (U_{L_2+SR}, $U_{L_\infty+SR}$, U_{L_1+SR}, U_{L_0+STA}, $U_{L_\infty+STA}$, U_{L_2+STA}, U_{L_0+SR}, U_{L_1+STA}) , 证明了噪声统计量较攻击成功率更全面; 通过计算不同评价指标的鲁棒性描述粒度, 证明了 L_2 是更精细的范数选择; 通过分析不同度量指标鲁棒性评估序关系, 证明噪声统计量的鲁棒性评估结果的独立性比攻击成功率更高. 实验部分在多个数据集上展开, 在大量模型及攻击方法组合上的实验结果验证了对抗鲁棒性评估指标体系完备性方面的理论分析. 本文的结论可对后续对抗攻防方法的评估与设计提供参考.

References:

[1] 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]

[2] Ma YK, Wu LF, Jian M, Liu FH, Yang Z. Algorithm to generate adversarial examples for face-spoofing detection. Ruan Jian Xue Bao/Journal of Software, 2019, 30(2): 469–480 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/5568.htm [doi: 10.13328/j.cnki.jos.005568]

[3] Wei XX, Zhu J, Yuan S, Su H. Sparse adversarial perturbations for videos. In: Proc. of the 33rd AAAI Conf. on Artificial Intelligence. Honolulu: AAAI, 2019. 8973–8980. [doi: 10.1609/aaai.v33i01.33018973]

[4] Yuan TH, Ji SH, Zhang PC, Cai HB, Dai QY, Ye SJ, Ren B. Adversarial example generation method for black box intelligent speech software. Ruan Jian Xue Bao/Journal of Software, 2022, 33(5): 1569–1586 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/6549.htm [doi: 10.13328/j.cnki.jos.006549]

[5] Wang WQ, Wang R, Wang LN, Tang BX. Adversarial examples generation approach for tendency classification on Chinese texts. Ruan Jian Xue Bao/Journal of Software, 2019, 30(8): 2415–2427 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/5765.htm [doi: 10.13328/j.cnki.jos.005765]

[6] Raghunathan A, Steinhardt J, Liang P. Certified defenses against adversarial examples. In: Proc. of the 6th Int'l Conf. on Learning Representations (ICLR 2018), 2018. http://doi.org/10.48550/arXiv.1801.09344

[7] Chen SH, Shen HJ, Wang R, Wang XZ. Relationship between prediction uncertainty and adversarial robustness. Ruan Jian Xue Bao/Journal of Software, 2022, 33(2): 524–538 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/6163.htm [doi: 10.13328/j.cnki.jos.006163]

[8] Li XJ, Wu GW, Yao L, Zhang WZ, Zhang B. Progress and future challenges of security attacks and defense mechanisms in machine learning. Ruan Jian Xue Bao/Journal of Software, 2021, 32(2): 406–423 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/6147.htm [doi: 10.13328/j.cnki.jos.006147]

[9] 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. 1632–1640.

[10] Papernot N, McDaniel P, Jha S, Fredrikson M, Celik ZB, Swami A. The limitations of deep learning in adversarial settings. In: Proc. of the 2016 IEEE European Symp. on Security and Privacy (EuroS&P). Saarbruecken: IEEE, 2016. 372–387. [doi: 10.1109/EuroSP.2016.36]

[11] Xie CH, Wang JY, Zhang ZS, Zhou YY, Xie LX, Yuille A. Adversarial examples for semantic segmentation and object detection. In: Proc. of the 2017 IEEE Int'l Conf. on Computer Vision. Venice: IEEE, 2017. 1378–1387. [doi: 10.1109/ICCV.2017.153]

[12] Sharif M, Bauer L, Reiter MK. On the suitability of lp-norms for creating and preventing adversarial examples. In: Proc. of the 2018 IEEE/CVF Conf. on Computer Vision and Pattern Recognition Workshops. Salt Lake City: IEEE, 2018. 1686–16868. [doi: 10.1109/CVPRW.2018.00211]

[13] Hendrycks D, Zhao K, Basart S, Steinhardt J, Song D. Natural adversarial examples. In: Proc. of the 2021 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021. 15257–15266. [doi: 10.1109/CVPR46437.2021.01501]

[14] Elsayed GF, Shankar S, Cheung B, Papernot N, Kurakin A, Goodfellow I, Sohl-Dickstein J. Adversarial examples that fool both computer vision and time-limited humans. In: Proc. of the 32nd Int'l Conf. on Neural Information Processing Systems. Montréal: Curran Associates Inc., 2018. 3914–3924.

[15] Brendel W, Rauber J, Bethge M. Decision-based adversarial attacks: Reliable attacks against black-box machine learning models. In: Proc. of the 6th Int'l Conf. on Learning Representations. Vancouver: ICLR, 2018. http://doi.org/10.48550/arXiv.1712.04248

[16] Goodfellow IJ, Shlens J, Szegedy C. Explaining and harnessing adversarial examples. In: Proc. of the 3rd Int'l Conf. on Learning Representations. San Diego: ICLR. 2015. http://doi.org/10.48550/arXiv.1412.6572

[17] Tramèr F, Kurakin A, Papernot N, Goodfellow IJ, Boneh D, McDaniel PD. ENSEMBLE adversarial training: Attacks and defenses. In: Proc. of the 6th Int'l Conf. on Learning Representations. Vancouver: ICLR, 2018. http://doi.org/10.48550/arXiv.1705.07204

[18] Li X, Li FX. Adversarial examples detection in deep networks with convolutional filter statistics. In: Proc. of the 2017 IEEE Int'l Conf. on Computer Vision. Venice: IEEE, 2017. 5775–5783. [doi: 10.1109/ICCV.2017.615]

[19] 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: ICLR, 2014. http://doi.org/10.48550/arXiv.1312.6199

[20] 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]

[21] Dong YP, Fu QA, Yang X, Pang TY, Su H, Xiao ZH, Zhu J. Benchmarking adversarial robustness on image classification. In: Proc. of the 2020 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020. 318–328. [doi: 10.1109/CVPR42600.2020.00040]

[22] Zhang H, Chen HG, Song Z, Boning DS, Dhillon IS, Hsieh CJ. The limitations of adversarial training and the blind-spot attack. In: Proc. of the 7th Int'l Conf. on Learning Representations. New Orleans: ICLR, 2019. http://doi.org/10.48550/arXiv.1901.04684

[23] Dong YP, Liao FZ, Pang TY, Su H, Zhu J, Hu XL, Li JG. Boosting adversarial attacks with momentum. In: Proc. of the 2018 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 9185–9193. [doi: 10.1109/CVPR.2018.00957]

[24] Kurakin A, Goodfellow IJ, Bengio S. Adversarial examples in the physical world. In: Proc. of the 5th Int'l Conf. on Learning Representations. Toulon: ICLR, 2017. http://doi.org/10.48550/arXiv.1607.02533

[25] Chen PY, Zhang H, Sharma Y, Yi JF, Hsieh CJ. ZOO: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models. In: Proc. of the 10th ACM Workshop on Artificial Intelligence and Security. Dallas: ACM, 2017. 15–26. [doi: 10.1145/3128572.3140448]

[26] Chen JL, Cao JN, Liang ZX, Cui XH, Yu LQ, Li W. STPD: Defending against ℓ_0-norm attacks with space transformation. Future Generation Computer Systems, 2022, 126: 225–236. [doi: 10.1016/j.future.2021.08.009]

[27] Seck I, Loosli G, Canu S. L_1-norm double backpropagation adversarial defense. In: Proc. of the 2019 European Symp. on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges, 2019. http://doi.org/10.48550/arXiv.1903.01715

[28] Liang SY, Wei XX, Cao XC. Generate more imperceptible adversarial examples for object detection. In: Proc. of the ICML 2021 Workshop on a Blessing in Disguise: The Prospects and Perils of Adversarial Machine Learning. 2021.

[29] Fawzi A. Robust image classification: Analysis and applications. Lausanne: EPFL, 2016.

[30] Pintor M, Roli F, Brendel W, Biggio B. Fast minimum-norm adversarial attacks through adaptive norm constraints. In: Proc. of the 35th Int'l Conf. on Neural Information Processing Systems. 2021. 20052–20062.

[31] 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: ICLR. 2018. http://doi.org/10.48550/arXiv.1706.06083

[32] Song CB, He K, Wang LW, Hopcroft JE. Improving the generalization of adversarial training with domain adaptation. In: Proc. of the 7th

Int'l Conf. on Learning Representations. New Orleans: ICLR, 2019.

[33] Krizhevsky A. Learning multiple layers of features from tiny images. Toronto: University of Toronto, 2009.

[34] Brendel W, Rauber J, Kurakin A, Papernot N, Veliqi B, Salathé M, Mohanty SP, Bethge M. Adversarial vision challenge. arXiv preprint arXiv:1808.01976, 2018.

[35] Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma SA, Huang ZH, Karpathy A, Khosla A, Bernstein M, Berg AC, Fei-Fei L. Imagenet large scale visual recognition challenge. Int'l Journal of Computer Vision, 2015, 115(3): 211–252. [doi: 10.1007/s11263-015-0816-y]

[36] Deng J, Dong W, Socher R, Li LJ, Li K, Fei-Fei L. ImageNet: A large-scale hierarchical image database. In: Proc. of the 2009 IEEE Conf. on Computer Vision and Pattern Recognition. Miami: IEEE, 2009. 248–255. [doi: 10.1109/CVPR.2009.5206848]

[37] Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai XH, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S, Uszkoreit J, Houlsby N. An image is worth 16x16 words: Transformers for image recognition at scale. In: Proc. of the 9th Int'l Conf. on Learning Representations. ICLR, 2021. http://doi.org/10.48550/arXiv.2010.11929

[38] Wu L, Zhu ZX, Tai C, E WN. Understanding and enhancing the transferability of adversarial examples. arXiv:1802.09707, 2018.

[39] Rony J, Hafemann LG, Oliveira LS, Ayed IB, Sabourin R, Granger E. Decoupling direction and norm for efficient gradient-based L2 adversarial attacks and defenses. In: Proc. of the 2019 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 4317–4325. [doi: 10.1109/CVPR.2019.00445]

[40] 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]

[41] Chen PY, Sharma Y, Zhang H, Yi JF, Hsieh CJ. EAD: Elastic-net attacks to deep neural networks via adversarial examples. In: Proc. of the 32nd AAAI Conf. on Artificial Intelligence. New Orleans: AAAI. 2018. 10–17. [doi: 10.1609/aaai.v32i1.11302]

[42] Jang U, Wu X, Jha S. Objective metrics and gradient descent algorithms for adversarial examples in machine learning. In: Proc. of the 33rd Annual Computer Security Applications Conf. Orlando: ACM, 2017. 262–277. [doi: 10.1145/3134600.3134635]

[43] Cheng MH, Singh S, Chen PH, Chen PY, Liu SJ, Hsieh CJ. Sign-OPT: A query-efficient hard-label adversarial attack. In: Proc. of the 8th Int'l Conf. on Learning Representations. Addis Ababa: ICLR. 2020. arXiv:1909.10773

[44] Chen JB, Jordan MI, Wainwright MJ. HopSkipJumpAttack: A query-efficient decision-based attack. In: Proc. of the 2020 IEEE Symp. on Security and Privacy. San Francisco: IEEE, 2020. 1277–1294. [doi: 10.1109/SP40000.2020.00045]

[45] Shi YC, Wang SY, Han YH. Curls & whey: Boosting black-box adversarial attacks. In: Proc. of the 2019 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 6512–6520. [doi: 10.1109/CVPR.2019.00668]

[46] Shi YC, Han YH, Tan YA, Kuang XH. Decision-based black-box attack against vision transformers via patch-wise adversarial removal. In: Proc. of the 36th Int'l Conf. on Neural Information Processing Systems (NeurIPS 2022). 2022. 12921–12933.

[47] Brunner T, Diehl F, Le MT, Knoll A. Guessing smart: Biased sampling for efficient black-box adversarial attacks. In: Proc. of the 2019 IEEE/CVF Int'l Conf. on Computer Vision. Seoul: IEEE, 2019. 4957–4965. [doi: 10.1109/ICCV.2019.00506]

[48] Dong YP, Su H, Wu BY, Li ZF, Liu W, Zhang T, Zhu J. Efficient decision-based black-box adversarial attacks on face recognition. In: Proc. of the 2019 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 7706–7714. [doi: 10.1109/CVPR.2019.00790]

[49] Shi YC, Han YH, Hu QH, Yang Y, Tian Q. Query-efficient black-box adversarial attack with customized iteration and sampling. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2023, 45(2): 2226–2245. [doi: 10.1109/TPAMI.2022.3169802]

[50] Hannun A, Case C, Casper J, Catanzaro B, Diamos G, Elsen E, Prenger R, Satheesh S, Sengupta S, Coates A, Ng AY. Deep speech: Scaling up end-to-end speech recognition. arXiv:1412.5567, 2014.

[51] Neekhara P, Hussain S, Pandey P, Dubnov S, McAuley JJ, Koushanfar F. Universal adversarial perturbations for speech recognition systems. In: Proc. of the 20th Annual Conf. of the Int'l Speech Communication Association. Graz: Interspeech, 2019. 481–485.

[52] Qin Y, Carlini N, Cottrell GW, Goodfellow IJ, Raffel C. Imperceptible, robust, and targeted adversarial examples for automatic speech recognition. In: Proc. of the 36th Int'l Conf. on Machine Learning. Long Beach: ICML, 2019. 5231–5240.

[53] Ling X, Ji SL, Zou JX, Wang JN, Wu CM, Li B, Wang T. DEEPSEC: A uniform platform for security analysis of deep learning model. In: Proc. of the 2019 IEEE Symp. on Security and Privacy (SP). San Francisco: IEEE, 2019. 673–690. [doi: 10.1109/SP.2019.00023]

[54] Torres-Sospedra J, Montoliu R, Martínez-Usó A, Avariento JP, Arnau TJ, Benedito-Bordonau M, Huerta J. UJIIndoorLoc: A new multi-building and multi-floor database for WLAN fingerprint-based indoor localization problems. In: Proc. of the 2014 Int'l Conf. on Indoor Positioning and Indoor Navigation (IPIN). Busan: IEEE, 2014. 261–270. [doi: 10.1109/IPIN.2014.7275492]

附中文参考文献:

[1] 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]

[2] 马玉琨, 毋立芳, 简萌, 刘方昊, 杨洲. 一种面向人脸活体检测的对抗样本生成算法. 软件学报, 2019, 30(2): 469–480. http://www.jos.org.cn/1000-9825/5568.htm [doi: 10.13328/j.cnki.jos.005568]

[4] 袁天昊, 吉顺慧, 张鹏程, 蔡涵博, 戴启印, 叶仕俊, 任彬. 针对黑盒智能语音软件的对抗样本生成方法. 软件学报, 2022, 33(5): 1569–1586. http://www.jos.org.cn/1000-9825/6549.htm [doi: 10.13328/j.cnki.jos.006549]

[5] 王文琦, 汪润, 王丽娜, 唐奔宵. 面向中文文本倾向性分类的对抗样本生成方法. 软件学报, 2019, 30(8): 2415–2427. http://www.jos.org.cn/1000-9825/5765.htm [doi: 10.13328/j.cnki.jos.005765]

[7] 陈思宏, 沈浩靖, 王冉, 王熙照. 预测不确定性与对抗鲁棒性的关系研究. 软件学报, 2022, 33(2): 524–538. http://www.jos.org.cn/1000-9825/6163.htm [doi: 10.13328/j.cnki.jos.006163]

[8] 李欣姣, 吴国伟, 姚琳, 张伟哲, 张宾. 机器学习安全攻击与防御机制研究进展和未来挑战. 软件学报, 2021, 32(2): 406–423. http://www.jos.org.cn/1000-9825/6147.htm [doi: 10.13328/j.cnki.jos.006147]

石育澄(1994－), 男, 博士, 讲师, 主要研究领域为对抗机器学习, 人工智能安全.

韩亚洪(1977－), 男, 博士, 教授, 博士生导师, CCF 高级会员, 主要研究领域为多媒体内容理解, 人工智能安全.