计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2023,32(3):150−156 [doi: 10.15888/j.cnki.csa.009029]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于图像识别的对抗鲁棒性评测系统①

章　威[1], 李辰琦[1], 胡逢法[1], 王　军[1], 钱　宸[1], 倪冰冰[2], 赵成龙[2]

[1](中电海康集团有限公司, 杭州 311100)
[2](上海交通大学 电子信息与电气工程学院, 上海 200240)
通信作者: 李辰琦, E-mail: lichenqi@cethik.com

摘　要: 深度神经网络的对抗鲁棒性研究在图像识别领域具有重要意义, 相关研究聚焦于对抗样本的生成和防御模型鲁棒性增强, 但现有工作缺少对其进行全面和客观的评估. 因而, 一个有效的基准来评估图像分类任务的对抗鲁棒性的系统被建立. 本系统功能主要为榜单评测展示、对抗算法评测以及系统优化管理, 同时利用计算资源调度和容器调度保证评测任务的进行. 本系统不仅能够为多种攻击和防御算法提供动态导入接口, 还能够从攻防算法的相互对抗过程中全方面评测现有算法优劣性.

关键词: 对抗样本; 防御模型; 对抗鲁棒性

引用格式: 章威,李辰琦,胡逢法,王军,钱宸,倪冰冰,赵成龙.基于图像识别的对抗鲁棒性评测系统.计算机系统应用,2023,32(3):150–156. http://www.c-s-a.org.cn/1003-3254/9029.html

Adversarial Robustness Evaluation System Based on Image Recognition

ZHANG Wei[1], LI Chen-Qi[1], HU Feng-Fa[1], WANG Jun[1], QIAN Chen[1], NI Bing-Bing[2], ZHAO Cheng-Long[2]

[1](Zhongdian Haikang Group Co. Ltd., Hangzhou 311100, China)
[2](School of Electronic Information and Electrical Engineering, Shanghai Jiao Tong University, Shanghai 200240, China)

Abstract: The adversarial robustness of deep neural networks is of great significance in the field of image recognition. Relevant studies focus on the generation of adversarial samples and the robustness enhancement of defense models but lack comprehensive and objective evaluation. Thus, an effective benchmark to evaluate the adversarial robustness of image classification tasks is developed. The main functions of this system are list display, adversarial algorithm evaluation, and system optimization management. At the same time, computing resource scheduling and container scheduling are applied to ensure the evaluation task. This system can not only provide a dynamic import interface for a variety of attack and defense algorithms but also evaluate the advantages and disadvantages of the existing algorithms from all aspects in the confrontation between attack and defense algorithms.

Key words: adversarial samples; defense models; adversarial robustness

自 20 世纪 50 年代以来, 人工智能技术飞速发展, 从最初的简单搜索和推理, 到自动定理证明, 再到现如今各种功能强大的智能机器, 人工智能技术极大影响了社会进程, 改变了人类的生活方式. 在图像识别学习领域, 研究者们综合利用人工智能、图像处理等技术形成了计算机视觉技术, 使得机器拥有了“视觉”, 各种相关应用层出不穷.

深度学习图像识别技术得到很大发展的同时, 其识别技术的鲁棒性容易受到各种对抗性样本的干扰而得出错误结果, 而这种干扰在人眼观察下通常是不可感知的. 因而, 研究者们聚焦于对抗样本技术的攻防, 不断衍生出各种高效的攻击算法和鲁棒的防御算法.

① 基金项目: 国家自然科学基金联合基金 (U20B2072)
收稿时间: 2022-08-09; 修改时间: 2022-09-27; 采用时间: 2022-10-21; csa 在线出版时间: 2022-12-09
CNKI 网络首发时间: 2022-12-13

例如, 防御蒸馏方法[1] 被提出以提高对抗鲁棒性, 但后来被证明对强攻击方法[2] 无效. 后续引入了很多方法, 如通过产生模糊梯度来建立鲁棒模型, 但自适应梯度的攻击方法又可以克服模糊梯度[3]. 因此, 确认这些攻防算法的影响、特性及适用范围是特别具有挑战性的. 此外, 当前的攻击和防御算法没有得到充分评估. 首先, 大多数防御算法只在有限的鲁棒模型下对一小部分攻击进行测试, 且许多攻击是在少数几个防御算法上评估的. 其次, 鲁棒性评价指标过于简单, 无法反映这些方法的性能. 对于给定的扰动预算和对抗性扰动的最小距离, 防御对攻击的准确率被用作主要的评估指标, 这往往不足以完全描述攻击和防御的行为, 使得研究者们无法全面了解这些方法的优势和局限性.

因此, 针对目前各种对抗攻击方法对各式防御鲁棒模型的威胁性、攻击脆弱性/鲁棒性缺乏系统化、标准化度量的问题, 构建一个泛在评测系统来全方位衡量多场景深度学习对抗攻防算法的鲁棒性, 既可以提升当前深度学习图像识别系统的安全与稳定性, 也可以为未来的研究提供基础.

1 相关工作

在构建泛在对抗算法评测系统前, 需要了解对抗领域中攻防算法主要方法、适用范围等基本概念, 为后续提出完善的系统奠定基础.

1.1 对抗攻击

考虑到对于目标模型的访问权限不同的情况, 攻击算法通常分为白盒攻击和黑盒攻击两类. 白盒攻击意味着攻击者不仅能够得到目标模型的反馈输出, 还能够掌握模型的具体结构及参数数值. 在已知内部结构的情况下, 基于梯度或者其他方式对其进行攻击, 直至输出错误.

目前大部分白盒对抗攻击算法基于梯度的方式生成对抗样本, 例如, 快速梯度下降法 (fast gradient sign method, FGSM)[4] 将输入空间中的损失函数线性化, 通过一步更新生成一个对抗性示例. 基本迭代法 (basic iterative method, BIM)[5] 通过迭代采用多个小梯度步骤扩展 FGSM. 与 BIM 类似, 投影梯度下降法 (gradient descent method, PGD)[6] 作为 FGSM 的迭代攻击版本, 针对非线性模型, PGD 算法在迭代过程中可以对上一次迭代得到的对抗样本进行随机噪声初始化, 以此避免优化过程中可能遇到的鞍点问题, 且 PGD 作为最强的一阶攻击, 常被用于评估模型鲁棒性的基准测试.

而有基于决策边界分析的白盒攻击方法, Moosavi-Dezfooli 等提出了 DeepFool[7] 来生成一个具有近似于最小扰动的对抗性示例, 推广至非线性决策边界的二分类问题, 可通过迭代过程近似得到针对数据点的最小扰动. 除此之外, 为了对抗防御蒸馏网络, Athalye 等人提出了 C&W 攻击方法[3], 该算法可以使用 Lp 范数分别对扰动进行限制生成对抗样本, 属于直接进行优化的攻击算法, 其相较于之前的直接优化算法, 对损失函数进行改进并将有约束的优化转化为无约束的凸优化问题, 方便梯度下降法的实施.

考虑现实场景下大部分的情况都适用于黑盒攻击的模式, 攻击者无法获取目标模型的内部结构和参数而发动的攻击. 黑盒模型又可细化为基于访问的攻击以及基于迁移的攻击.

相关研究者们提出了几种方法来提高可迁移性. 例如, 在 BIM 算法中引入动量的概念, 在攻击迭代过程中更新方向, 以此提出动量迭代法 (momentum iterative method, MIM)[8], 进一步提高了防御模型的可转移性. 而针对基于梯度优化的常用算法 FGSM, 其应用于迁移的黑盒攻击情况容易陷入过拟合, Dong 等人[8] 也将动量引入梯度迭代过程中来, 通过设置梯度累积量, 每次迭代的梯度都由当前梯度与之前的累积量相叠加所得, 从而提高了白盒攻击算法的迁移性.

1.2 对抗防御

由于对抗样本对现有深度学习网络的威胁, 人们对建立抵御对抗性攻击的鲁棒模型进行了广泛的研究. 本文将防御技术大致分为 5 类, 包括鲁棒训练、输入转换、随机化、模型集成和认证防御. 但某些防御方法并不单单属于其中一类, 有可能是多种策略的复合型.

提高模型对对抗样本的鲁棒性是对抗防御的既定策略, 其基本思路是使分类器对内部的小噪声具有鲁棒性. 而鲁棒训练最有效的一种方法是对抗性训练[4,6,9,10], 它通过生成对抗样本来扩充训练数据. 对抗训练指的是在模型训练过程中加入对抗样本, 其在训练过程中加入由对抗攻击算法产生的对抗样本使得网络变得更加鲁棒. 然而, 这种生成对抗样本的方式所需成本不菲, 由于其迭代生成再训练的过程, 深度神经网络模型的训练时间和计算资源消耗都大大增加, 并且这种训练的有效性通常还取决于所生成样本的有效性能够类似于所防御对象的对抗样本的攻击性, 因而这种非适应

性的防御训练仅局限于已知的对抗攻击技术.

对抗训练这种在数据层面增强模型的方法被认为最为有效, 除此之外, Madry 等人[6] 通过在训练中扩大模型容量, 提出了一种对抗性训练的变体, 这种对抗训练的思路是先在允许的范围内随机初始化搜索对抗样本, 然后进行计算生成对抗样本, 再利用生成的对抗样本作为训练数据集进行训练, 由于随机的初始点带来了更好的攻击效果, 从而使得对抗训练的防御更加有效.

Kannan 等人[11] 介绍了一种基于逻辑值配对的方法, 该方法是在对抗训练加入一个正则项进行扩展. 本文用逻辑值配对进行了 3 组实验, 分别对应于 3 种常用开源图像数据集. 实验分别测量了分类器对原始样本、白盒与黑盒场景下的对抗样本的识别准确率. 结果表明经过逻辑值配对方法的分类器具有更高的准确性.

同时模型压缩也是一种有效防御手段, Papernot 等人[1] 提出一种基于知识蒸馏的防御方法, 将大模型压缩成具有更平滑的决策表面的小模型, 在提高模型鲁棒性的同时保持预测精度不变. Guo 等人[12] 证明, 利用模型剪枝来适当提高非线性深度神经网络的稀疏性能提高其鲁棒性, 但过度稀疏的模型可能更难以抵抗对抗样例. Zhao 等人[13] 发现模型剪枝减少了网络的参数密度, 对于用原网络作出的攻击防御性不足, 对参数和激活函数的大幅度量化也能使攻击的迁移性变小.

2 总体框架

在了解对抗攻防算法相关基础后, 基于建设面向智能视频理解场景的图像数据与算法网络的安全性评测系统这一基本目的, 架构设计上主要分为榜单评测展示、对抗算法评测及系统优化管理等模块. 如图 1.

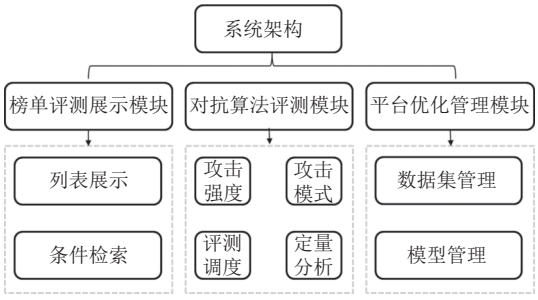

图 1 系统架构

图 1 中, 首页榜单显示提供显示评测榜单列表及对应检索操作; 可按创建时间顺序或者综合得分排列. 本系统在功能设计上, 将对抗算法评测作为主要功能, 其评测对象主要为特定条件下生成的对抗样本及鲁棒防御模型, 而衡量样本和模型效果的重要影响因素在系统中实现了计算指标从定性到定量的改进, 通过控制对抗样本的生成条件 (包含对抗攻击扰动的强度、攻击模式的种类及对抗扰动的制定规则等), 得出相应条件下的评测结果. 同时系统结合软件开发技术, 综合调度计算资源, 实现对评测任务并行评测.

系统优化管理即包含数据集管理、模型管理以及计算指标管理等功能. 管理模块主要涵盖上传、下载、修改及删除等功能, 为用户提供自定义接口以实现对本地算法的评测.

因而, 本文旨在建立一个全面的、客观的基准来评估对抗鲁棒性, 可以详细了解现有防御方法在不同场景、不同攻击强度下的效果, 希望能为未来的研究提供帮助. 特别强调的事, 我们关注了 Lp 范数威胁模型下的图像分类器的鲁棒性, 其中已经投入了大量的工作. 系统整合了许多典型的、最先进的攻击和防御方法来进行鲁棒性评估, 包括多种攻击方法和防御模型. 为了充分展示这些方法的性能, 我们采用多对多的攻防算法交叉评估实验来进行防御流程评测, 其评测结果记录在后端数据库中, 并实时显示在首页榜单.

经过算法与软件的适配开发, 我们开发了一个新的面向图像与深度网络的鲁棒性评测系统——HikDeepSec, 与现有的系统 (如 CleverHans[14] 和 Foolbox[15]) 相比, 本系统能够支持现有的大部分评测需求.

3 榜单评测展示模块

在榜单评测展示模块, 展示列表表头为评测名称、评测类型、算法测试类别 (黑盒/白盒)、创建人员、创建时间、攻击方法、防御模型、数据集、综合得分及操作的榜单记录列表; 用户可按创建时间顺序或者综合得分排列. 且针对所有用户, 评测记录都会开放展示.

因而, 在系统的软件设计方面, 本对抗样本评测系统采用 B/S (browser/server) 架构开发, 可以实现前后端分离, 提升开发效率. B/S 结构的优点是维护和升级相对简单, 而且成本较低. 防御系统的设计需要分成两部分: 前端页面展示和后端数据处理. 前端使用 Vue.js

开发, 并搭载了目前较为流行的 Element-UI 的界面组件, 后端使用 Django 开发.

前端页面使用 Vue.js 框架进行设计开发, 对比于其他前端大型框架, 使用 Vue.js 构建网页可视化界面, 主要是因为 Vue.js 在使用上更加灵活轻量, 适合 Web 开发, 可以进行自底向上增量开发, 在前后端分离的项目中能够较好地发挥其优势. 各个功能模块以组件形式存在. 一个功能组件是一个单独文件, 每个组件有自身的逻辑与样式, 减少不同的代码风格造成在页面样式上的冲突.

前端页面可视化可以实现用户注册登录, 新建评测过程中展示可用攻击算法及防御模型, 展示防御后的各防御方法生成的识别结果和置信度等指标. 后端处理主要完成与前端数据交互的任务, 数据交互主要包括用户身份信息验证, 图片的上传与下载, 相关算法参数的上传和下载. 前端页面通过对应的 (application programming interface, API) 向后台发送请求, 后台接收到请求后根据功能的实际需求返回相应的数据供前端展示.

4 对抗算法评测模块

在本系统中进行对抗方法评测, 重点关注对抗样本的生成效率及防御模型鲁棒性的提高, 模型效率提高可以通过优化模型结构、梯度正则化、批标准化、特征去噪等方式, 但这些方式都需要本地重新训练且耗时较多. 基于此原因, 对抗算法评测模块不提供训练资源, 仅对系统中的训练好的算法进行评测.

4.1 评测流程制定

为有效了解对抗算法的机理及应用价值, 对抗算法评测模块的评测流程首先需要根据面向的图像场景确立原始样本, 再根据任务类型及目标配置对抗攻击算法, 生成具有对抗扰动的对抗样本. 而针对对抗训练后所得的对抗防御模型, 将未对抗训练的原始模型与其进行模型组合, 分别适配于原始样本与对抗样本的数据组合, 根据系统提供的评测指标包得到对抗样本的攻击效果及防御模型的鲁棒性表现. 其流程如图 2.

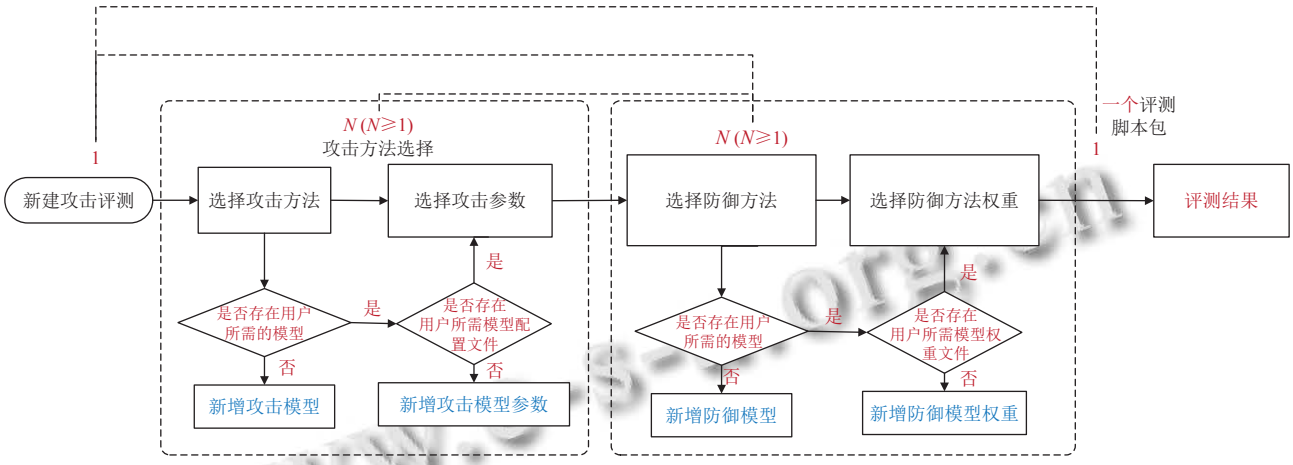

图 2 对抗算法评测流程

这种评测流程的制定有利于后续确定目标图像识别模型的优化目标和优化方式, 增加目标图像识别模型针对攻击样本 (类似于上述对抗样本) 的免疫性.

4.2 评测指标定量分析

从数值定量分析进行评估对抗样本和防御模型的鲁棒性具有重要意义, 不仅能够准确比较能力强弱, 还可以综合判断各种防御策略的优劣势, 衡量维度更加丰富全面. 本系统中, 一次评测任务的评测脚本包需要涵盖多种指标定义, 综合衡量模型鲁棒性, 系统拟从以下 4 种指标对模型进行评测.

(1) 分类准确度方差 (classification accuracy variance, CAV) 为评估深度学习模型性能的最重要指标, 其计算结果为防御前后模型对测试用例的分类精度插值, 定义如式 (1):

$$CAV = ACC(F^D, T) - ACC(F, T) \tag{1}$$

其中, $ACC(F, T)$ 表示模型 F 在数据集 T 下的识别准确率, F^D 表示添加防御性能的模型. 式 (1) 表明 CAV 指标的数值越大, 模型添加防御策略后的防御效果越强,

但该数值仅能反应分类准确差值, 对于分类置信分布没有进行统计.

(2) 分类置信方差 (classification confidence variance, *CCV*): 对原始模型增加防御, 可能不会影响评估准确率, 但是正确分类样本的预测可信度可能会降低, 因此引入 *CCV*, 测量防御增强模型引起的置信方差, 定义如式 (2):

$$CCV = \frac{1}{n}\sum_{i=1}^{n}\left|P(X_i)_{y_i} - P^D(X_i)_{y_i}\right| \tag{2}$$

其中, $P(X_i)$ 为第 i 个样本被分类为对应类的概率, y_i 表示第 i 个样本被正确分类为 y_i 类. n 表示在原始模型及防御增强后模型全部分类准确的样本数量.

分类输出稳定性 (classification output stability, *COS*): 使用 JS 散度来衡量原始模型和防御增强模型输出概率相似性, 即分类输出的稳定性. 对所有正确分类的测试实例, 定义如式 (3):

$$COS = \frac{1}{n}\sum_{i=1}^{n} JSD(P(X_i)\|P^D(X_i)) \tag{3}$$

其中, n 表示在原始模型及防御增强后模型全部分类准确的样本数量, *JSD* 表示计算 JS 散度的函数. *COS* 值越低, 两个模型的差距越小.

(3) 分类校正 (classification rectify, *CRR*)/补偿比率 (sacrifice ratio, *CSR*): 为了评估防御对模型在测试集上预测结果的影响, 将 *CRR* 定义为原始模型错误分类, 但增加防御后, 模型正确分类的测试实例的百分比. 与此相反, *CSR* 表示原始模型正确分类, 但增加防御后, 模型错误分类的测试实例的百分比. 定义如式 (4) 和式 (5):

$$CRR = \frac{1}{N}\sum_{i=1}^{N} count(F(X_i) \neq y_i \ \&\ F^D(X_i) = y_i) \tag{4}$$

$$CSR = \frac{1}{N}\sum_{i=1}^{N} count(F(X_i) = y_i \ \&\ F^D(X_i) = y_i) \tag{5}$$

同时 $CAV=CRR-CSR$. 其中 F 表示原始模型预测结果, F^D 表示防御后模型预测结果, &符号表示同时满足. 返回值越小, 两个模型的差距越小. 说明 *CRR* 越大, 防御效果越好; *CSR* 越大, 防御效果越差.

5 系统优化管理模块

考虑到系统现有算法并不能对所有对抗攻击和对抗防御算法进行评估, 系统优化管理模块中则可以实现用户评测新算法的需求. 根据实际使用及软件设计经验, 管理模块主要包含数据集管理、攻击算法管理、防御模型管理. 管理操作即添加、修改、上传及下载等, 但依照管理员和普通用户设计权限, 确保系统的稳定性和可操作性.

攻击算法管理中涉及算法脚本及参数配置, 分别开放 API 接口, 根据系统引导, 用户可在基础攻击算法类的基础上继承该类方法创建新的攻击方法且对抗样本的生成可以实现在系统上, 无需本地上传对抗样本即可实现新的攻击算法的评测. 且攻击算法参数可根据用户需求自行设定, 满足用户对该类算法参数设置合理性及有效性的判断.

防御模型管理则有算法脚本和权重文件的对应接口, 防御算法脚本同攻击一样, 继承基础防御模型类; 权重文件格式遵照系统格式及结构, 用户需本地训练, 系统仅提供调用接口. 这种模式下, 可以减轻系统所消耗的计算负担, 同时实现了动态调用模型的需求. 如图 3 所示.

图 3 防御模型上传

数据集管理遵照图像数据文件夹和标签文本的组合方式, 上传及下载过程皆为压缩包格式, 且修改及删除权限仅限管理员所有. 用户如要对多种场景的数据进行评估, 则可上传对应的压缩包, 但普通用户所上传数据集仅限当前用户使用, 这样设计保证了用户数据的安全性.

6 对抗评测结果及分析

6.1 评测指标定量分析

本系统基于 Docker 调度算法后台, 一次评测任务启动一个 Docker 容器, 评测结束即关闭容器. 这样保证了算法环境与其他软件环境的相互隔离, 避免因环境冲突引发系统运行失败.

在两个图像基准数据集 CIFAR10 和 SVHN 上训练防御模型并使用多种对抗攻击算法进行鲁棒性评估. CIFAR10 基准数据集有 5 万张训练图像和 1 万张测试图像以及 10 个标签类. SVHN 数据集本文提取单个数字作为图像源, 其总共有 26 032 张图像以及 10 个标签类, 训练方式为随机抽取 90% 作为训练集, 其余为测试集.

攻击设置如下: 本文只讨论白盒攻击 (攻击者知道模型的所有参数, 直接攻击模型产生对抗样本), 使用 1 种经典的单步攻击和 2 种先进的迭代攻击来生成对抗样本, 单步攻击为 FGSM 攻击, 迭代攻击包括 BIM 攻击和 PGD 攻击. 对于 BIM 和 PGD 攻击, 迭代次数为 5, 扰动大小 ε 取值为 1/255, 2/255, 3/255, 对应的迭代扰动步长为 α=0.004.

模型训练参数设置如下: 所有模型基于 ResNet34 网络采用 SGD 优化器进行训练, 训练集批次大小为 128, 训练次数 epoches 为 200, 学习率为 0.01.

同时, 根据本文在第 4.2 节中对于评测结果的多重量化分析, 本文既采用分类准确率进行评估, 也会根据防御前后的指标变化进行综合评估, 其中分类准确率越高, 能够直观判断出当前防御模型的鲁棒性较好.

6.2 基于数据集评测结果分析

首先, 使用不同扰动大小的对抗攻击算法对基于 CIFAR10 数据集的 ResNet34 防御模型进行评测.

如表 1 所示, 随着对抗扰动的增强, 防御模型也会随着初始模型对对抗样本的识别率轻微下降, 但是其分类准确率方差 CAV 则会有较大程度的提升, 说明即使扰动强烈, 防御模型相较于普通模型仍然有较强的鲁棒性. 且在多种攻击条件下, CCV 指标都没有较大变化且数值保持在 0.3 以内, 说明防御模型没有过多降低正确样本的分类置信度.

同理, 如表 2 所示, 使用不同扰动大小的对抗攻击算法对基于 SVHN 数据集的 ResNet34 防御模型进行评测.

表 1 基于 CIFAR10 的 ResNet34 防御模型评测

指标	FGSM			PGD			BIM		
	1/255	2/255	3/255	1/255	2/255	3/255	1/255	2/255	3/255
ACC	**0.7475**	**0.7404**	**0.7318**	**0.7477**	**0.7394**	**0.7313**	**0.7477**	**0.7388**	**0.7385**
CAV	0.153	0.2026	0.2306	0.1828	0.2577	0.3249	0.1854	0.232	0.234
CCV	0.2849	0.136	0.1351	0.1353	0.1429	0.1464	0.135	0.1407	0.1405
COS	0.2823	0.2986	0.3008	0.2995	0.3305	0.3429	0.2993	0.3191	0.3194
CRR	0.2108	0.2629	0.2662	0.2326	0.2924	0.352	0.2342	0.2701	0.2717
CSR	0.0578	0.051	0.0356	0.0498	0.0347	0.0271	0.0488	0.0381	0.0377

表 2 基于 SVHN 的 ResNet34 防御模型评测

指标	FGSM			PGD			BIM		
	1/255	2/255	3/255	1/255	2/255	3/255	1/255	2/255	3/255
ACC	0.7789	0.7665	0.7497	0.7796	0.7698	0.7599	0.78	0.7669	0.7665
CAV	0.1623	0.2108	0.2285	0.1882	0.2704	0.342	0.1911	0.2433	0.245
CCV	0.1118	0.1161	0.1196	0.1157	0.1247	0.1322	0.1152	0.1224	0.1219
COS	0.2281	0.2404	0.2401	0.2509	0.2839	0.3002	0.2497	0.2685	0.2686
CRR	0.1842	0.2256	0.2403	0.2082	0.2825	0.3503	0.2109	0.2563	0.2579
CSR	0.0219	0.0148	0.0118	0.0201	0.0121	0.0082	0.0199	0.013	0.0128

对比 CIFAR10 数据集上的统计结果, 可以看出基于 SVHN 数据集的防御模型有普遍较高准确率. 基于 ResNet34 网络模型面对多种攻击算法时, 无论是否有进行鲁棒训练, 其识别结果具有一定鲁棒性.

6.3 系统可视化数据

基于 Web 页面的可视化数据分析, 系统实现了对多种攻击、防御模型的鲁棒性指标评测, 不同对抗样本对同一防御模型或是不同防御模型框架对同一对抗样本的攻击性进行评测则验证了系统对不同攻防算法的评测有效性.

同时, 在启动多个评测任务时, 评测系统能够监测各个评测任务的状态, 容器异常及数据问题会及时抛

出, 评测正常的任务返回评测时长及评测结果. 因而本系统能够很好地完成容器和计算资源调度, 确保评测任务的正常进行. 如图 4 所示.

图 4 系统计算任务框

7 结论与展望

本文提出的系统建立了一个全面、严格、连贯的基准来评估图像分类器的对抗鲁棒性. 且根据评估结果, 一些重要的发现既为本系统提供了理论依据, 也会对未来的研究有帮助. 但系统仅提供图像识别领域的对抗评估, 对于其他任务有待开发, 还有针对黑盒问询攻击的评测也是一大难点.

参考文献

1 Papernot N, McDaniel P, Wu X, *et al*. Distillation as a defense to adversarial perturbations against deep neural networks. Proceedings of the 2016 IEEE Symposium on Security and Privacy (SP). San Jose: IEEE, 2016. 582–597.

2 Carlini N, Wagner D. Towards evaluating the robustness of neural networks. Proceedings of the 2017 IEEE Symposium on Security and Privacy (SP). San Jose: IEEE, 2017. 39–57.

3 Athalye A, Carlini N, Wagner D. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. Proceedings of the 35th International Conference on Machine Learning. Stockholmsmässan: PMLR, 2018. 274–283.

4 Goodfellow IJ, Shlens J, Szegedy C. Explaining and harnessing adversarial examples. Proceedings of the 3rd International Conference on Learning Representations (ICLR). San Diego: ICLR, 2015.

5 Kurakin A, Goodfellow IJ, Bengio S. Adversarial examples in the physical world. Proceedings of the 5th International Conference on Learning Representations. Toulon: OpenReview.net, 2017.

6 Madry A, Makelov A, Schmidt L, *et al*. Towards deep learning models resistant to adversarial attacks. arXiv: 1706.06083, 2017.

7 Moosavi-Dezfooli SM, Fawzi A, Frossard P. Deepfool: A simple and accurate method to fool deep neural networks. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016. 2574–2582.

8 Dong YP, Liao FZ, Pang TY, *et al*. Boosting adversarial attacks with momentum. Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 9185–9193.

9 Tramèr F, Kurakin A, Papernot N, *et al*. Ensemble adversarial training: Attacks and defenses. arXiv:1705. 07204, 2017.

10 Zhang HY, Yu YD, Jiao JT, *et al*. Theoretically principled trade-off between robustness and accuracy. Proceedings of the 36th International Conference on Machine Learning. Long Beach: PMLR, 2019. 7472–7482.

11 Kannan H, Kurakin A, Goodfellow I. Adversarial logit pairing. arXiv:1803.06373, 2018.

12 Guo YW, Zhang C, Zhang CS, *et al*. Sparse DNNs with improved adversarial robustness. Proceedings of the 32nd International Conference on Neural Information Processing Systems. Montréal: Curran Associates Inc., 2018. 240–249.

13 Shumailov I, Zhao Y, Mullins R, *et al*. To compress or not to compress: Understanding the interactions between adversarial attacks and neural network compression. Proceedings of Machine Learning and Systems, 2019. 230–240.

14 Papernot N, Faghri F, Carlini N, *et al*. Technical report on the CleverHans v2.1.0 adversarial examples library. arXiv:1610.00768v6, 2016.

15 Rauber J, Brendel W, Bethge M. Foolbox: A Python toolbox to benchmark the robustness of machine learning models. arXiv:1707.04131v3, 2017.

(校对责编: 牛欣悦)