

# 人工智能对抗环境下的模型鲁棒性研究综述

王科迪, 易平

上海交通大学网络空间安全学院 上海 中国 200240

**摘要** 近年来人工智能研究与应用发展迅速, 机器学习模型大量应用在现实的场景中, 人工智能模型的安全鲁棒性分析与评估问题已经开始引起人们的关注。最近的研究发现, 对于没有经过防御设计的模型, 攻击者通过给样本添加微小的人眼不可察觉的扰动, 可以轻易的使模型产生误判, 从而导致严重的安全性问题, 这就是人工智能模型的对抗样本。对抗样本已经成为人工智能安全研究的一个热门领域, 各种新的攻击方法, 防御方法和模型鲁棒性研究层出不穷, 然而至今尚未有一个完备统一的模型鲁棒性的度量评价标准, 所以本文总结了现阶段在人工智能对抗环境下的模型鲁棒性研究, 论述了当前主流的模型鲁棒性的研究方法, 从一个比较全面的视角探讨了对抗环境下的模型鲁棒性这一研究方向的进展, 并且提出了一些未来的研究方向。

**关键词** 对抗样本; 模型鲁棒性; 人工智能安全

**中图分类号** TP393.08 **DOI号** 10.19363/J.cnki.cn10-1380/tn.2020.05.02

## A Survey on Model Robustness under Adversarial Example

WANG Kedi, YI Ping

School of Cyber Science and Engineering, Shanghai Jiao Tong University, Shanghai 200240, China

**Abstract** In recent years, the research on artificial intelligence has developed rapidly. However, in order to apply machine learning model to real-world setting, we need to consider its security issues in particular. Recent studies have found that for unprotected models, attackers can easily fool the machine learning models by adding small, imperceptible disturbances to the samples, leading to serious security problems. Adversarial sample is a popular research direction nowadays. There are many researches on new attack methods, defense methods and robustness certifications, but there is no well-known and unified framework for certificating model's robustness. Our paper summarizes the research on model robustness in artificial intelligence adversarial setting. This paper describes the popular research methods of model robustness, discusses the research progress of model robustness in adversarial setting from a more comprehensive perspective, and puts forward some future research directions.

**Key words** adversarial examples; model robustness; artificial intelligence security

### 1 简介

近几年, 人工智能技术发展飞快, 并且在多个领域得到了大量的应用, 比如图像识别<sup>[1]</sup>, 自然语言处理<sup>[2]</sup>, 目标检测<sup>[3]</sup>, 机器翻译<sup>[4]</sup>等, 但是如果要将人工智能技术大量应用到现实的生活场景中, 我们必须要考虑它的安全性与可用性。虽然现阶段机器学习模型已经可以达到较高的准确性, 但是近期的研究表明通过对正常样本添加微小的扰动就可以使得模型产生误判, 而这些扰动基本上不会使得人眼产生任何误判<sup>[5-7]</sup>。这一类样本被称为对抗样本。现

阶段, 已经有许多针对对抗样本存在性原理以及模型的对鲁棒性的研究<sup>[8-11]</sup>。此外, 研究表明, 攻击者针对目标模型构造出的对抗样本, 有很大概率可以使得其他的机器学习模型也产生误判<sup>[12-13]</sup>, 这一对抗样本的迁移性使得在现实环境中保证模型在对抗环境下的鲁棒性变得尤为重要, 否则攻击者将可以在不了解模型详情的情况下, 也能轻易的实现攻击。以无人车自动驾驶中使用的图像视频识别技术为例: 在文献[14]中, 攻击者给一个“停止”的路牌加上了一些微小的扰动记号成功使得模型将“停止”误判成其他标记, 如下图1所示, 这些记号不影响人

**通讯作者:** 易平, 博士, 副教授, Email: yiping@sjtu.edu.cn。

本课题得到重点研发计划(No.2017YFB0802900)资助。

收稿日期: 2019-09-20; 修改日期: 2019-11-14; 定稿日期: 2020-04-27

眼对路牌的判断但是可以大概率使得机器产生误判,同时这些记号在不同的拍摄角度下都能实现攻击,这一类的误判可以导致非常严重的安全事故。机器学习分类模型将“停止”识别为“60km/h”的路牌,将“右转”识别为“停止”的路牌<sup>[14]</sup>

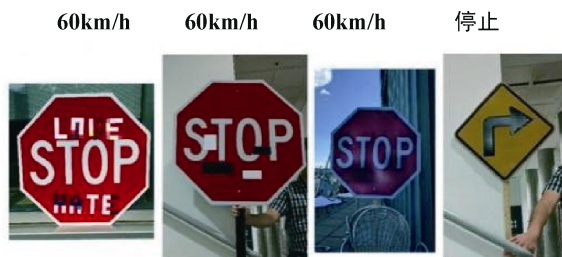


图 1 机器学习识别的“停止”路牌

Figure 1 Machine learning recognize STOP

在本篇综述中,我们将对这些对抗环境下模型鲁棒性的研究做一个完善的介绍,并分析它们的优势与缺陷,从一个比较全面的视角探讨模型鲁棒性这一研究方向的进展。

本文将首先介绍对抗样本的概念以及对抗环境下的安全性问题,并对对抗样本的存在性与原理进行阐述。接着文章将从两大部分介绍现阶段在对抗环境下对模型鲁棒性的研究:如何评估模型鲁棒性以及 how 提升模型鲁棒性。本文详细介绍了现阶段对于模型鲁棒性研究的进展,并在最后对该领域研究的未来发展方向提出了一些想法与建议。

## 2 对抗环境下的安全性问题

对于机器学习的模型而言,人工智能对抗环境下的安全性指的是存在恶意攻击者的场景下,机器学习模型的安全性。在算法理论研究的场景,针对机器学习分类算法的研究的主要目标是进一步提升模型性能,计算效率和模型准确率。但是在现实环境部署的场景下,任何算法都必须要考虑存在攻击者的场景下其算法的安全性。基于攻击者所能获知的算法系统的相关信息的程度以及最终攻击产生的效果,我们可以对机器学习模型抵抗攻击的能力做一个定性的度量:为了实现成功的攻击,攻击者所需要的模型信息越多,所实现的攻击效果越弱,说明模型的鲁棒性越强。

本章将对对抗样本及其存在性进行分析探讨,同时对对抗环境下的攻击手段进行简要介绍。

### 2.1 对抗样本简介

文献[15]首次提出了对抗样本的概念:对抗样本指的是通过对原样本添加微小的扰动,使得机器学

习模型产生误判,但同时并不会使得人眼产生误判的一类人工构造的样本。对抗样本的可视化如下图 2,左侧是正常图像,右侧是添加了对抗扰动的图像,两者对人眼几乎没有差异,但是对抗扰动可以使得深度神经网络产生误判。同时,Goodfellow 在文献[16]中提出了一种基于梯度快速构造出对抗样本的方法。研究表明,对抗样本在各类机器学习模型中普遍存在,如果不经有效的防御,几乎所有的机器学习模型都会被对抗样本所攻击。计算机视觉的分类模型在对抗样本上会产生戏剧性的错误。如下图,深度神经网络把左图识别为狗,右图识别为鸵鸟。<sup>[17]</sup>

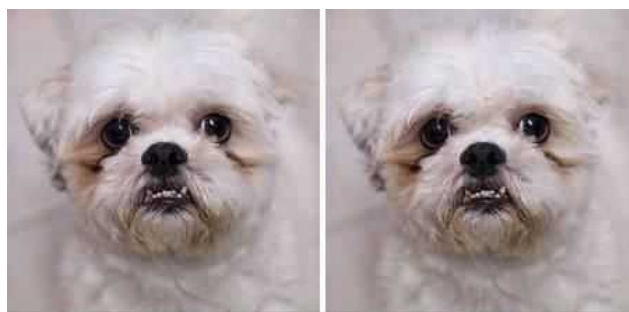


图 2 计算机视觉的分类模型在对抗样本上产生的错误

Figure 2 The classification model of computer vision will produce mistakes on adversarial samples

### 2.2 对抗样本的存在性分析

由于对抗样本在机器学习模型中普遍存在,为了研究在对抗环境下的模型鲁棒性,针对对抗样本存在性的研究必不可少。

在机器学习分类模型的训练中,常常包含着一个基本的假设:训练数据与测试数据两者应该是独立同分布的。因此,基于训练数据训练的模型可以泛化到测试数据集,从而能够正确预测测试集的结果。而对抗样本集本质上是偏离正常样本的集合,这就破坏了这一假设。从模型的决策边界的角度来看,我们的分类任务存在一个事实的决策边界,而模型通过训练数据会得到一个模型的决策边界,由于训练样本无法覆盖整个输入空间,模型的决策边界与事实的决策边界必然是无法完全重合的,而这些不重合的地方就被成为对抗区域,可能会产生对抗样本。并且这一特性在输入维度变高的时候,会变得更加明显,不重合的区域会更多。

对于对抗样本的存在原因,现阶段的研究提出了许多不同的观点。这些观点通常与研究人员在攻击或防御深层神经网络时所做的局部经验观察一致。然而,它们在泛化性方面往往存在一些缺陷。

### 2.2.1 损失函数的线性特性

Goodfellow 在文献[16]中提出损失函数的线性特性是對抗样本存在的原因之一。该特性也是 FGSM 攻击能够成功找到对抗样本的理论依据。

令神经网络的权重参数为  $w$ , 输入向量为  $x$ , 那么对于一个线性模型, 它的输出为:  $score = w^T x$ 。添加一个微小的扰动  $\eta$ , 生成一个新的输入向量为  $x' = x + \eta$ , 那么它的输出为:  $score' = w^T x + w^T \eta$ 。当输入的维数足够大的时候, 通过令  $\eta = sign(w)$  可以使得最终的输出产生巨大的变化, 从而使得神经网络产生误判。

实际应用中的, 深度神经网络虽然是非线性的, 但是它们的激活函数通常只是起值域压缩的作用, 设计得一般较为平滑, 如常用的 ReLU, sigmoid 等激活函数。这就导致了损失函数在输入样本  $X$  的域内较为平滑, 呈明显的线性。因此, 对抗样本的线性解释也可以被应用在深度神经网络中。以 FGSM 为例, 此方法只计算了样本点的损失函数梯度, 正是因为样本点的梯度方向上损失函数会持续线性上升, 才导致按这个方向总能找到对抗样本。

图 3 是 MNIST 数据集中一个样本的损失函数曲线。纵轴为这个样本被分到正确类别中的损失函数值, 横轴为对这个样本添加扰动的  $L_0$  范数, 扰动方向为该样本点损失函数的符号梯度方向。可以看到在中间部分, 损失函数具有相当的线性, 这也解释了为什么 FGSM 可以仅仅使用单步的攻击达到优越的攻击效果。

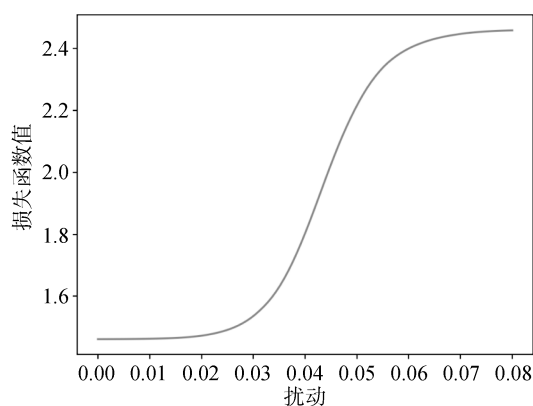


图 3 损失函数-扰动关系图

Figure 3 Loss function-disturbance

### 2.2.2 样本像素空间维度

Tabacof 和 Eduardo 从另一个角度解释了对抗样本的存在原因<sup>[18]</sup>。他们在 MNIST<sup>[19]</sup> 和 IMAGENET<sup>[20]</sup> 数据集上对浅层和深层网络分类器

生成了对抗样本, 并利用不同分布和强度的噪声对抗样本的空间进行了探索, 发现对抗样本所在的像素空间维度要高于正常样本的维度, 并且他们认为浅层的, 更线性的分类器和深层的更具有非线性能力的分类器一样容易受到对抗样本的影响。

Tramer 等人<sup>[21]</sup>提出了一种估计对抗样本空间的维度的方法, 并且认为对抗样本涵盖了一个连续的高维子空间。而正是因为这个原因导致了在高维空间中不同分类器的子空间可能会有相交, 这就产生了对抗样本的高可迁移性。

### 2.2.3 全局通用扰动的存在性

Moosavi-Dezfooli 等人在文献[22]中证明了全局通用的对抗扰动的存在。通过给样本集中的所有样本添加一个统一的微小的对抗扰动, 攻击者可以很大概率使得原样本转变为对抗样本。基于此发现, Moosavi-Dezfooli 等人提出了对抗样本的存在是由于其利用了分类器决策边界之间的几何相关性。在文献[23]中, 他们进一步对全局通用的扰动进行了分析, 并证明了存在共同的方向(在数据点之间共享), 沿着这些方向, 分类器的决策边界可以高度正弯曲, 也就是可以使得样本点附近的损失函数可以沿着这些方向快速变大, 从而越过决策边界, 使得模型产生误判。

### 2.2.4 非鲁棒性特征

Andrew 等人在文献[24]中证明了对抗样本的存在可以归因到非鲁棒性特征在样本空间的普遍存在。非鲁棒性特征指的是一类对于机器学习模型具有高度的预测性, 但对人类来说是脆弱和不可理解的特征。模型利用的不仅仅是人眼可以理解的一些图像特征。这些人眼无法理解的特征称为非鲁棒性特征。过去我们常常会认为这些特征是训练样本的一些异常, 而模型会对此产生过拟合。论文中提出这些特征对于提高模型的准确率也不可或缺, 对于人的视觉无意义, 但是对于机器学习是有意义的。

Andrew 在文献[24]中将训练集数据的特征拆分为鲁棒性和非鲁棒性特征分别进行训练, 发现仅使用非鲁棒性特征进行训练也能得到较高的准确率。具体的训练算法如图 4 中所示。该研究认为对抗样本的存在可能是基于样本分布的特性, 并且不依赖于特定的分类模型。从这个角度也可以解释对抗样本的高可迁移性。

表 1 将各类对抗样本存在性的研究进行了分析和对比。

## 2.3 对抗攻击

对抗环境中的攻击者针对目标模型的攻击行为

被称为对抗攻击。这一部分包含许多的内容, 各种新的对抗攻击的方式也在被不断地研究, 本文侧重于对模型鲁棒性的介绍与讨论, 所以对于对抗攻击进行简单的介绍。在文献[25]中, 对对抗攻击和防御方式做了详细的介绍, 可以做进一步的了解。

表 1 对抗样本存在性研究的比较分析

Table 1 Comparative analysis of the research on the existence of adversarial samples

研究理论	理论简介	可靠性
损失函数的线性特性	损失函数的线性特性是对于对抗样本存在的原因之一, 也是 FGSM 攻击能够成功找到对抗样本的理论依据	***
样本像素空间维度	对抗样本涵盖了一个连续的高维子空间, 也就是说对抗样本所在的像素空间维度要高于正常样本的维度	****
全局通用扰动存在性	每个样本点的同一个方向均有一个曲度较大的决策边界, 可以使样本点附近的损失函数可以沿着这些方向快速变大	****
非鲁棒特性	数据分布中存在非鲁棒性的特征可以提升模型正确率但对人眼无法区分	****

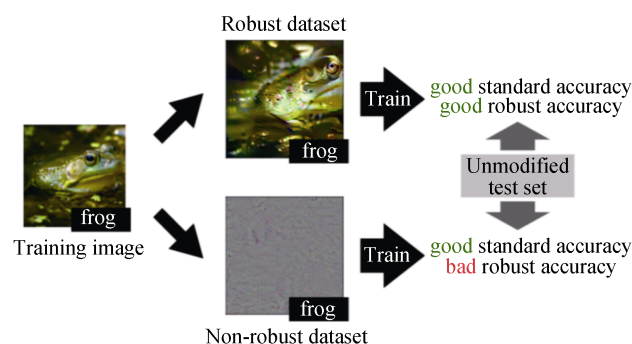


图 4 区分模型的鲁棒性<sup>[23]</sup>

Figure 4 Robustness of discriminant model<sup>[23]</sup>

基于文献[26]提出的概念, 我们可以将对抗攻击分为以下几类:

**逃避攻击** 这是在对抗环境下最常见的攻击方式。攻击者通过在测试阶段对测试样本添加扰动, 从而躲避或欺骗模型的检测。该方法不影响训练数据。常见的攻击方法包括: (Fast Gradient Sign Method, FGSM<sup>[16]</sup>), (Project Gradient Decent, PGD<sup>[27]</sup>), (Carlini Wagner Attack, CW<sup>[28]</sup>), (Jacobian-based Saliency Map Attack, JSMA<sup>[29]</sup>)等。

**毒化攻击** 这一类攻击主要影响模型的训练阶段, 通过在模型训练阶段投入恶意的样本, 使得训练得到的模型存在后门或漏洞, 从而可以被攻击者

所利用来实施攻击。如文献[30]中提出的 Deep-Confuse 算法就是通过微弱扰动数据库来彻底破坏对应的学习系统的性能。

**探测攻击** 这一类攻击的主要目的是获取目标系统的信息, 包括学习算法, 训练数据的模式等, 相对来说影响较小, 但也更隐蔽难以被发现。该类攻击可以盗取目标模型的相关信息, 为后续进一步攻击或破坏模型提供基础信息。

### 3 评估模型鲁棒性

#### 3.1 模型鲁棒性的定义

##### 3.1.1 基于 $L_p$ 范数的模型鲁棒性

对于一个鲁棒的模型, 当模型的输入产生微小的变化的时候, 不会导致模型的输出产生巨大的变化, 这保证了模型在预测结果时的稳定性。而在对抗样本的环境下, 对于模型鲁棒性的研究主要解决这样一个问题: 对于一个给定的机器学习模型, 给定一个输入样本点, 如果该样本点周围的空间的所有样本通过模型得到的预测结果都是一致的, 那么就认为该样本在它的局部空间内是鲁棒的。这里的局部空间可以通过  $L_p$  范数的扰动大小来定义。定义一个  $L_p$  范数限制的空间球体, 如果样本点在该空间内得到的预测结果都是一致的, 那么它就在该  $L_p$  球的空间内就是鲁棒的。具体的定义如下公式 1, 给定一个样本点  $x_0$ , 给定一个分类模型  $F$ , 扰动大小  $\epsilon$ , 对应的预测分类结果为  $y_0$ , 对于  $x \in \mathcal{B}(x_0, \epsilon)$ , 满足:

$$y_0 = F(x) \quad (1)$$

其中,  $\mathcal{B}(x, \epsilon)$  指的是以  $x$  为中心,  $\epsilon$  为半径的空间球体。

##### 3.1.2 对抗距离

基于  $L_p$  范数的模型鲁棒性定义, 我们可以发现对于一个样本点, 存在一个最小的扰动半径  $\epsilon$ , 使得该半径空间内的所有样本都可以被正确预测, 而大于该半径的空间存在使得模型误判的样本。该最小的扰动半径的大小就被定义为对抗距离<sup>[31]</sup>。

如果我们计算出目标模型的对抗距离, 那么对抗距离将可以作为一个合理的模型鲁棒性评估的指标, 越大的对抗距离表示模型具有更强的鲁棒性。但是对抗距离的计算被证明是一个 NP-完全的问题<sup>[32-33]</sup>。因此 形式化的分析验证的方式即使在非常小型的网络中也需要极大的计算量, 几乎是不可用的。因此有许多研究专注于估计对抗距离的大小, 从而允许我们可以量化的评估模型的鲁棒性, 包括对抗距离上边界的评估以及下边界的评估。



### 3.2 对抗距离上边界评估

对抗距离的上边界可以理解为对于实际对抗距离大于这个上边界的样本, 存在一种扰动使得其可以变为一个对抗样本。这一类的评估通常通过设计算法去构造扰动更小的对抗样本来实现, 因此大部分是攻击相关的评估方式。理想情况下, 模型鲁棒性的评估应该与特定的攻击无关, 但是现阶段由于对抗环境下模型鲁棒性的研究不完备, 依然有大量的实验基于这类特定攻击的指标进行评估。

#### 3.2.1 LP formulation

Bastani 等人在文献[34]中提出了一个构造对抗样本的方法: LP formulation。该方法可以找到比基于梯度的攻击算法如 L-BFGS<sup>[15]</sup>等更小的扰动, 并以该方法生成的扰动大小作为模型鲁棒性评估的一个指标。但是该方法只能在小型的网络中使用, 无法应用到 ImageNet 等更为复杂的分类任务中。

#### 3.2.2 DeepFool

Moosavi-Dezfooli 等人在文献[31]中提出了一个可以构造出细微扰动的攻击算法 DeepFool, 图 5 展示了 DeepFool 算法构造的对抗样本的效果。

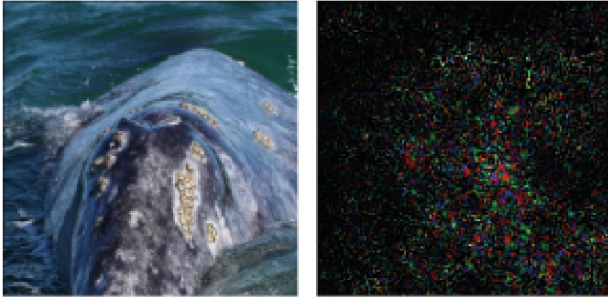


图 5 第一行为 DeepFool 生成的对抗样本图像和对应的扰动图像

Figure 5 First row is the adversarial samples generated by DeepFool.

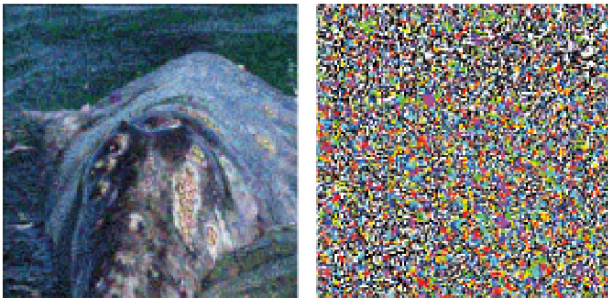


图 6 第二行为 FGSM 生成的对抗样本图像和对应的扰动图像

Figure 6 Next row is the adversarial samples generated by FGSM

同时文献[31]首次形式化的提出了一个针对模型鲁棒性评估的指标。

给定一个样本  $x$ , 一个分类模型  $F$  和对抗扰动大小  $r$ :

$$\Delta(x; F) := \min_r \|r\|_2 \text{ s.t. } F(x+r) \neq F(x)$$

定义  $\Delta(x; F)$  为  $F$  在  $x$  点的鲁棒性, 分类器  $F$  的鲁棒性定义如下:

$$\rho_{\text{adv}}(F) = \mathbb{E}_x \frac{\Delta(x; F)}{\|x\|_2}$$

其中,  $\mathbb{E}_x$  是整个分布的数据的期望。

论文中提出使用  $\rho_{\text{adv}}(F)$  作为鲁棒性评估的指标, 但是对于最小扰动的计算, 依然依赖与特定的对抗攻击算法, 因此依然有比较大的局限性, 要通过对抗距离的上边界去不断近似对抗距离相对来讲是一个难以优化的问题。

### 3.3 对抗距离下边界评估

受限于对抗距离上边界的评估通常依赖于特定的攻击方式, 有许多研究把目标转向了对于对抗距离下边界的评估。对抗距离下边界的评估目标是寻找一个下边界, 使得小于该边界距离的扰动都无法使得原样本被转化为对抗样本。可证明的下边界不依赖于特定的攻击算法, 因此更具普适性, 也更适合作为鲁棒性评估的指标。本节将对相关的研究进行简介。

#### 3.3.1 CLEVER score

TW Weng 等人在文献[35]中提出了一个基于李普希兹约束的对抗样本下边界的评估方式, 称为 CLEVER, 全称为 Cross Lipschitz Extreme Value for nEtnetwork Robustness。CLEVER 是首个攻击无关的模型鲁棒性评估指标, 并且可以被应用到任何神经网络模型中, 同时该方法使用极值理论来估计李普希兹常数从而极大减少了计算量使得可以很好的在大型的网络中工作, 比如针对 ImageNet 的分类模型。基于 CLEVER 评估指标, 在文献[36]中, TW Weng 等人进一步扩展了 CLEVER, 提出了二阶-CLEVER score, 使得该扩展后的指标可以被二次微分, 同时说明了 CLEVER 可以应对梯度遮掩的情况<sup>[37]</sup>。

#### 3.3.2 Fast-Lin 和 Fast-Lip

TW Weng 等人在文献[38]中提出了两种针对使用 ReLU 激活函数的网络的对抗距离可证明下边界的计算方式。Fast-Lin 使用合适的线性函数作为边界限制, 而 Fast-Lip 使用李普希兹常数作为边界限制。通过利用 ReLU 结构的信息, 这两种方法可以提供更接近真实对抗距离的可证明下边界评估, 同时有更高的计算速度。

### 3.3.3 CROWN 框架

H Zhang 等人在文献[39]中提出了一个针对任意激活函数的通用鲁棒性证明的框架 CROWN。该框架意在提供一个更接近真实对抗距离的可证明的下边界, 是现阶段效果最好的一个算法之一, 并且该方法对激活函数的类型不作任何限制, 通用性更好。相比于 CLEVER, 该方法提供了可证明的下边界, 相比于 Fast-Lin 和 Fast-Lip, 该方法的通用性更好, 不只局限于 ReLU 激活函数。

### 3.4 其他鲁棒性评估指标

现阶段, 对抗环境下大部分模型的鲁棒性评估指标都是基于  $L_p$  范数的, 并且大部分攻击模型也是基于  $L_p$  范数来实现攻击的, 但是对于图像中  $L_p$  范数的扰动大小与人眼感知到的图片差异的大小的一致性依然是无法保证的。对抗样本的概念指的是使得机器学习模型误判而人眼可以正确识别的一类样本, 那么如何保证人眼可以正确识别呢? 针对该问题, 一些非  $L_p$  范数的指标也被提出用来评估模型的鲁棒性。

#### 3.4.1 平均结构相似度 (ASS)

文献[40]提出了一个指标来比较图像之间的结构相似度 SSIM。通过大量的实验发现, 相比于  $L_p$  范数, SSIM 的效果和人眼的视觉有更好的一致性。因此文献[41]提出使用平均结构相似度 ASS 作为一个评估模型鲁棒性的指标, ASS 的定义如下:

$$ASS = \frac{1}{n} \sum_{i=1}^n SSIM(X_i^a, X_i)$$

#### 3.4.2 扰动敏感性距离 (PSD)

基于对比度遮蔽理论<sup>[42]</sup>, 文献[43]提出使用扰动敏感性距离来评估对人眼视觉的扰动大小。对于对抗样本而言, 越小的 PSD 数值代表对人眼越不可察觉。实验表明, 该指标相比  $L_p$  范数能更好的评估图片扰动对人眼扰动的程度。

#### 3.4.3 样本间 Wasserstein 距离

文献[44]提出了一类区别于  $L_p$  范数的对抗样本, 它使用 Wasserstein 距离代替  $L_p$  范式来衡量正常样本和对抗样本之间的差异。因此样本间的 Wasserstein 距离也可以被用来作为模型鲁棒性的评估指标。基于该指标训练得到的防御后的模型被证明可以抵御  $L_p$  范数的攻击, 但是该指标与人眼分类的一致性关系依然并不清楚。

表 2 为各类评估指标的对比分析。

## 4 提升模型鲁棒性

提升模型鲁棒性的目标是使得模型可以抵御对

抗环境下攻击者的攻击, 相比于模型鲁棒性的评估指标, 提升模型鲁棒性的方法更加的多样化。为了提升模型的鲁棒性, 现在主流的研究大致分为三个方向:

(1) 修改模型输入数据, 包括在训练阶段修改训练数据以及在测试阶段修改输入的样本数据。

(2) 修改网络结构, 比如添加更多的网络层数, 改变损失函数或激活函数等方法。

(3) 添加外部模块作为原有网络模型的附加插件, 提升网络模型的鲁棒性。

现阶段很难找到一个可以抵御任何对抗样本的防御方法, 也就是在对抗环境下, 并没有一个可证明的完全鲁棒的模型存在。防御对抗样本的困难, 主要可以归结为以下两个原因<sup>[45]</sup>:

(1) 构造对抗样本过程的理论模型很难建立。对于大部分机器学习模型而言, 构造对抗样本是一个复杂的非线性的优化过程, 现在并没有合适的理论分析工具来对这一过程进行理论分析与证明, 所以对于一个鲁棒性的模型, 很难证明它可以将任何对抗样本的正确分类。

(2) 机器学习模型被要求为任何可能的输入提供合理的输出, 而为了实现鲁棒性的模型, 这一目标可能就会无法实现, 必须要拒绝一部分非法的输入样本。

在本节, 我们将对常见的一些提升模型鲁棒性的方式进行介绍。

表 2 鲁棒性评估算法的比较分析

Table 2 Comparative analysis of robustness evaluation algorithms

评估算法	算法优势	算法劣势
上边界评估	评估方式实现相对简单, 计算量较小	依赖于特定攻击, 通用性较差, 无法达到理论最优
下边界评估	攻击无关, 通用性强, 有完善的理论基础	逼近确界的计算复杂度很高
其他评估方式	攻击无关, 发展空间巨大	没有完善的研究框架支持, 性能验证较难实现

## 4.1 修改模型输入数据

### 4.1.1 对抗训练

修改模型输入数据方式主要的提出依据是由于模型在训练过程中训练数据的不完备。如下图 7 所示, 图 7A 中展示了该分类任务实际的决策边界, 以及提供的训练样本数据。基于这些训练样本数据, 对于一个理想的分类器而言是很难学到真实的决策边界的, 如图 7B 所示, 深色的直线是模型的决策边界,

它已经可以做到完整正确的对样本集数据进行分类。但是对于不属于该样本集合分布的数据, 该模型的性能表现就会急剧下降。通过添加更多合适的样本集合, 我们可以构建出更接近真实的决策边界, 如图 7C 所示。更进一步的讲, 通过将得到的对抗样本也加入训练, 可以进一步提升模型的鲁棒性, 使得两者的决策边界更加接近, 如图 7D 所示, 其中红色的点是使得原模型分类错误的是对抗样本。

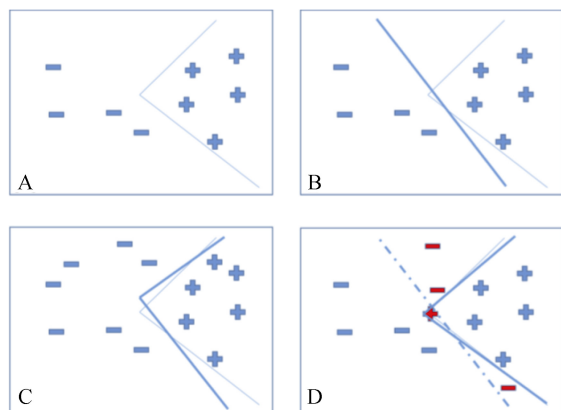


图 7 添加额外的数据来提升模型鲁棒性

Figure 7 Add additional data to improve model robustness

对抗训练就是基于这个思路, 将利用 FGSM 攻击生成的对抗样本放入原模型网络中进行训练, 从而提升模型的鲁棒性<sup>[16]</sup>。实验表明, 该方法能在一定程度上提升模型的鲁棒性, 并且基于经典的对抗训练, 衍生出许多优化的对抗训练方式, 如集成对抗训练<sup>[46]</sup>, 层叠对抗训练<sup>[47]</sup>。集成对抗训练使用了多个相似模型生成的对抗样本放入目标模型中进行训练, 从而提高对抗样本的泛化能力, 同时减少训练时的计算量。层叠对抗学习则使用了经过对抗学习训练的模型生成的对抗样本, 从而进一步提升用于训练的对抗样本的质量来保证训练得到一个更具有鲁棒性的机器学习模型。

#### 4.1.2 数据压缩

对抗训练修改训练样本集来提升模型的鲁棒性, 而数据压缩算法主要修改测试阶段数据的输入, 比如对于图像输入, 利用图像压缩的算法对图像进行预处理在输入模型进行预测, 从而去除对抗样本的对抗扰动的影响, 提升模型的鲁棒性。文献[48]研究发现了 JPG 压缩算法可以大概率除去 FGSM 攻击构造的扰动对模型分类的影响, 文献[49]进一步以 JPG 图像压缩算法为基础, 构建了针对 FGSM 和 DeepFool 的防御方式, 但是该类数据压缩的方式被证明无法防御更强的攻击算法如 CW 攻击等。该类

算法的优点是计算量较小, 实现简单, 但是防御的性能不佳, 同时会对正常图像的分类准确率产生一定的影响。

## 4.2 修改网络结构

### 4.2.1 梯度遮掩

大部分对抗样本的攻击都是基于梯度的, 因此一个自然的防御想法就是遮掩模型的梯度, 使得对抗攻击的算法无法找到合适的扰动方向<sup>[46]</sup>。比如如果一个模型是不可微分的, 那么诸如 FGSM 之类的算法就无法在该模型上构造对抗样本。但是由于对抗样本的高可迁移性, 攻击者通过训练替代模型<sup>[12]</sup>的方式依然可以成功的对模型进行攻击, 构造出对抗样本, 从而影响模型的鲁棒性。

### 4.2.2 防御性蒸馏

蒸馏的概念是由 Hinton 等人在文献[50]中提出的。它是一种知识迁移的方法, 可以将大型神经网络学到的信息迁移到小型网络中。而 Papernot 等人在文献[51-52]中提出将该方法用于提升模型鲁棒性中。在对抗样本领域, 该方法曾经是非常流行的一种方法, 不过在之后的研究中被证明也是一种利用了梯度遮掩的技术。防御性蒸馏算法的核心是将原模型输出的概率分布向量再次输入相同的模型中进行学习, 使得最终学习到的模型的分类边界更加平滑, 从而防御常见的对抗攻击方法。但是 Carlini 等人在文献[28]中提出, 防御性蒸馏不能很好的防御 CW 攻击。

### 4.2.3 添加对抗样本分类

Grosse 等人在文献[53]中提出在分类网络中添加一个额外的类别标签来判断输入的样本是否是抗样本, 从而使得网络拥有检测对抗样本的能力, 从而提升模型的鲁棒性。Hosseini 等人在文献[54]中也使用了一类相似的策略来检测黑盒环境下的攻击。

## 4.3 添加外部模块

该类方法的核心是不影响原有模型的正常工作, 通过添加前置或者后端的模块增强整个系统的鲁棒性。

### 4.3.1 防御全局扰动

在上文对抗样本的存在性分析中, 我们提到了针对模型的全局扰动的存在。Akhtar 等人在文献[55]中提出了一种针对全局扰动的防御方式。该方法在目标模型之前, 添加了一个额外的处理层, 对该层网络进行训练, 使得该层网络拥有还原全局扰动的对抗样本的能力。整个训练过程与原模型完全独立, 不影响原模型的分类。该方法存在一定的局限性, 只能防御全局的扰动, 而对其他类型的攻击无能为力。

### 4.3.2 特征压缩

特征压缩的方法在文献[56]中被提出。该方法在原有的分类模型前添加了一个分类器而不影响原有模型的结构。特征压缩算法的实现主要包含两类特征压缩的方式, 第一类是减小每个像素的深度, 从原有的 256 个取值范围进行压缩, 减小取值范围。第二类是减小空间维度上的差异, 通过一些像素平滑技术, 如中值模糊等去除相邻像素之间的相关性。该方法可以有效的提升模型的鲁棒性, 但同时也会在一定程度上导致模型的准确率下降。

表 3 为各类提升模型鲁棒性方式的分析与对比。

表 3 提升模型鲁棒算法的比较分析

Table 3 Comparative analysis of robust algorithms for lifting model

算法类别	算法优势	算法劣势
修改模型输入	实现相对简单, 无需改动网络结构	缺少理论性与形式化的研究
修改网络结构	理论性能最优, 实现方式多元化	算法实现复杂, 训练难度较高, 计算量较大
添加外部模块	较轻量, 易于部署, 高可扩展性	单独使用性能不佳, 通常要与其他方式一起使用

## 5 结论

本文着重介绍了对抗环境下模型鲁棒性的相关研究, 包括对抗样本的存在性原理, 模型鲁棒性的评估方式, 以及提升模型鲁棒性的方式, 同时也对常见的对抗攻击的算法做了简单的介绍。通过本文的介绍可以发现, 对抗环境下模型鲁棒性的各类研究涉及的面非常的广, 并且这是一个非常活跃的领域, 不断有新的想法被提出, 我们认为人工智能对抗环境下的模型鲁棒性研究在如下几个方面, 可以有进一步的发展:

(1) 我们可以发现现阶段大部分的研究都是基于  $L_p$  范数的扰动的, 但是这与实际人眼的判断不是强一致的, 对于特定的图片, 微小  $L_p$  范数的扰动也可能导致人产生大概率的误判, 所以寻找更合适的鲁棒性评价的指标是未来的研究的一个重要方向。

(2) 现阶段对各种攻击, 防御方法性能的评估是不完善的, 每篇论文都会用不同的评价指标, 不同的数据集, 不同的评判模型, 不同的实验参数来论证方法的合理性, 没有一个标准的方法, 论文之间的互相比较也很困难, 这也导致了对分类模型做系统性的鲁棒性分析比较困难, 所以未来需要一个系统化的框架来对模型在对抗环境下的鲁棒性做分析, 以保证其在现实环境下应用时的安全性。

(3) 对抗样本的存在原因和机器学习模型脆弱的原因需要更多理论性的研究, 现阶段相关的研究相互之间并无法保证完全的兼容, 也无法匹配所有场景的情况, 所以依然有很多提升的空间。比如损失函数的线性特征可能是对抗样本存在的原因之一, 但不可能是唯一的因素。更多针对性的理论分析模型需要被提出和改进。

## 参考文献

- [1] D. A. Forsyth, J. Ponce, Computer vision: a modern approach[C]. Prentice Hall Professional Technical Reference, 2002: 123-132.
- [2] Bates M. Models of Natural Language Understanding[J]. *Proceedings of the National Academy of Sciences of the United States of America*, 1995, 92(22): 9977-9982.
- [3] A. Krizhevsky, I. Sutskever, G. E. Hinton, Imagenet classification with deep convolutional neural networks[C]. *Advances in neural information processing systems (NIPS)*, 2012: 1097-1105.
- [4] I. Sutskever, O. Vinyals, Q. V. Le, Sequence to sequence learning with neural networks[C]. *Advances in neural information processing systems (NIPS)*, 2014: 3104-3112.
- [5] Nguyen A, Yosinski J, Clune J. Deep Neural Networks are Easily Fooled: High Confidence Predictions for Unrecognizable Images[C]. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015: 427-436.
- [6] A. Kurakin, I. Goodfellow, S. Bengio. Adversarial machine learning at scale[BT/OL]. 2016:arXiv preprint arXiv:1611.01236.
- [7] Corneanu C A, Madadi M, Escalera S, et al. What does it Mean to Learn in Deep Networks? And, how does one Detect Adversarial Attacks?[C]. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019: 4757-4766.
- [8] C.-Y. Ko, Z. Lyu, T.-W. Weng, et al. Popqorn: Quantifying robustness of recurrent neural networks[BT/OL]. 2019: arXiv preprint arXiv:1905.07387.
- [9] Wicker M, Kwiatkowska M. Robustness of 3D Deep Learning in an Adversarial Setting[C]. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019: 11767-11776.
- [10] Arnab A, Miksik O, Torr P H S. On the Robustness of Semantic Segmentation Models to Adversarial Attacks[C]. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018: 888-897.
- [11] H. Zhang, T.W. Weng, P.Y. Chen et al. Efficient neural network robustness certification with general activation functions[C]. *Advances in Neural Information Processing Systems (NIPS)*, 2018: 4939-4948.
- [12] Papernot N, McDaniel P, Goodfellow I, et al. Practical Black-Box Attacks Against Machine Learning[C]. *Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Se-*



- curity, 2017: 506-519.
- [13] Xie C H, Zhang Z S, Zhou Y Y, et al. Improving Transferability of Adversarial Examples with Input Diversity[C]. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019: 2730-2739.
- [14] K. Eykholt, I. Evtimov, E. Fernandes, et al. Robust physical-world attacks on deep learning models[BT/OL]. 2017: arXiv preprint arXiv:1707.08945.
- [15] C. Szegedy, W. Zaremba, I. Sutskever, al. Intriguing properties of neural networks[BT/OL]. 2013: arXiv preprint arXiv:1312.6199.
- [16] I. J. Goodfellow, J. Shlens, C. Szegedy. Explaining and harnessing adversarial examples[BT/OL]. 2014: arXiv preprint arXiv:1412.6572.
- [17] Nguyen A, Yosinski J, Clune J. Deep Neural Networks are Easily Fooled: High Confidence Predictions for Unrecognizable Images[C]. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015: 427-436.
- [18] Tabacof P, Valle E. Exploring the Space of Adversarial Images[C]. *2016 International Joint Conference on Neural Networks (IJCNN)*, 2016: 426-433.
- [19] LeCun Y, Boser B, Denker J S, et al. Backpropagation Applied to Handwritten Zip Code Recognition[J]. *Neural Computation*, 1989, 1(4): 541-551.
- [20] J. Deng, W. Dong, R. Socher, et al. Imagenet: A large-scale hierarchical image database[C]. *IEEE conference on computer vision and pattern recognition (CVPR)*. IEEE, 2009: 248-255.
- [21] F. Tramer, N. Papernot, I. Goodfellow, et al. The space of transferable adversarial examples[BT/OL]. 2017: arXiv preprint arXiv:1704.03453.
- [22] Moosavi-Dezfooli S M, Fawzi A, Fawzi O, et al. Universal Adversarial Perturbations[C]. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017: 1765-1773.
- [23] S.-M. Moosavi-Dezfooli, A. Fawzi, et al. Analysis of universal adversarial perturbations[ET/OL]. 2017: arXiv preprint arXiv:1705.09554.
- [24] A. Ilyas, S. Santurkar, D. Tsipras, et al. Adversarial examples are not bugs, they are features[ET/OL]. 2019: arXiv preprint arXiv:1905.02175.
- [25] A. Chakraborty, M. Alam, V. Dey, et al. Adversarial attacks and defences: A survey[ET/OL]. 2018: arXiv preprint arXiv:1810.00069.
- [26] Biggio B, Fumera G, Roli F. Security Evaluation of Pattern Classifiers under Attack[J]. *IEEE Transactions on Knowledge and Data Engineering*, 2014, 26(4): 984-996.
- [27] A. Madry, A. Makelov, L. Schmidt, et al. Towards deep learning models resistant to adversarial attacks[C]. *International Conference on Learning Representations (ICLR)*, 2018. [Online]. Available: <https://openreview.net/forum?id=rJzIBfZAb>
- [28] Carlini N, Wagner D. Towards Evaluating the Robustness of Neural Networks[C]. *2017 IEEE Symposium on Security and Privacy (SP)*, 2017: 39-57.
- [29] Papernot N, McDaniel P, Jha S, et al. The Limitations of Deep Learning in Adversarial Settings[C]. *2016 IEEE European Symposium on Security and Privacy (EuroS&P)*, 2016: 372-387.
- [30] J. Feng, Q.-Z. Cai, Z.-H. Zhou. Learning to confuse: Generating training time adversarial data with auto-encoder[ET/OL]. 2019: arXiv preprint arXiv:1905.09027.
- [31] Moosavi-Dezfooli S M, Fawzi A, Frossard P. DeepFool: A Simple and Accurate Method to Fool Deep Neural Networks[C]. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016: 2574-2582.
- [32] Katz G, Barrett C, Dill D L, et al. Reluplex: An Efficient SMT Solver for Verifying Deep Neural Networks[M]. *Computer Aided Verification*. Cham: Springer International Publishing, 2017: 97-117.
- [33] A. Sinha, H. Namkoong, J. Duchi. Certifiable distributional robustness with principled adversarial training[C]. *International Conference on Learning Representations (ICLR)*, [Online]. Available: <https://openreview.net/forum?id=Hk6kPgZA->
- [34] O. Bastani, Y. Ioannou, L. Lampropoulos, et al. Measuring neural net robustness with constraints[C]. *Advances in neural information processing systems (NIPS)*, 2016: 2613-2621.
- [35] T.-W. Weng, H. Zhang, P.-Y. Chen, et al. Evaluating the robustness of neural networks: An extreme value theory approach[ET/OL]. 2018: arXiv preprint arXiv:1801.10578.
- [36] Weng T W, Zhang H, Chen P Y, et al. On Extensions of Clever: A Neural Network Robustness Evaluation Algorithm[C]. *2018 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, 2018: 1159-1163.
- [37] I. Goodfellow, Gradient masking causes clever to overestimate adversarial perturbation size[ET/OL]. 2018: arXiv preprint arXiv:1804.07870.
- [38] T.-W. Weng, H. Zhang, H. Chen, et al. Towards fast computation of certified robustness for relu networks[ET/OL]. 2018: arXiv preprint arXiv:1804.09699.
- [39] H. Zhang, T.-W. Weng, P.-Y. et al. Efficient neural network robustness certification with general activation functions[C]. *Advances in neural information processing systems (NIPS)*, 2018: 4939-4948.
- [40] Wang Z, Bovik A C, Sheikh H R, et al. Image Quality Assessment: From Error Visibility to Structural Similarity[J]. *IEEE Transactions on Image Processing*, 2004, 13(4): 600-612.
- [41] Ling X, Ji S L, Zou J X, et al. DEEPSEC: A Uniform Platform for Security Analysis of Deep Learning Model[C]. *2019 IEEE Sym-*

- sium on Security and Privacy (SP)*, May 19-23, 2019. San Francisco, CA, USA. Piscataway, NJ: IEEE, 2019: 256-261.
- [42] Liu A M, Lin W S, Paul M, et al. Just Noticeable Difference for Images with Decomposition Model for Separating Edge and Textured Regions[J]. *IEEE Transactions on Circuits and Systems for Video Technology*, 2010, 20(11): 1648-1652.
- [43] B. Luo, Y. Liu, L. Wei, et al. Towards imperceptible and robust adversarial example attacks against neural networks[C]. *Thirty-Second AAAI Conference on Artificial Intelligence (AAAI)*, 2018:65-75.
- [44] E. Wong, F. R. Schmidt, J. Z. Kolter. Wasserstein adversarial examples via projected sinkhorn iterations[ET/OL]. 2019: arXiv preprint arXiv:1902.07906.
- [45] I. Goodfellow, N. Papernot, S. Huang, et al. Attacking machine learning with adversarial examples. *OpenAI*. <https://blog.openai.com/adversarial-example-research>, 2017.
- [46] F. Tramer, A. Kurakin, N. Papernot, et al. Ensemble adversarial training: Attacks and defenses[ET/OL]. 2017: arXiv preprint arXiv:1705.07204.
- [47] T. Na, J. H. Ko, S. Mukhopadhyay. Cascade adversarial machine learning regularized with a unified embedding[ET/OL]. 2017: arXiv preprint arXiv:1708.02582.
- [48] G. K. Dziugaite, Z. Ghahramani, D. M. Roy. A study of the effect of jpg compression on adversarial images[ET/OL]. 2016: arXiv preprint arXiv:1608.00853.
- [49] N. Das, M. Shanbhogue, S.T. Chen, et al. Keeping the bad guys out: Protecting and vaccinating deep learning with jpeg compression[ET/OL]. 2017: arXiv preprint arXiv:1705.02900.
- [50] G. Hinton, O. Vinyals, J. Dean. Distilling the knowledge in a neural network[ET/OL]. 2015: arXiv preprint arXiv:1503.02531.
- [51] Papernot N, McDaniel P, Wu X, et al. Distillation as a Defense to Adversarial Perturbations Against Deep Neural Networks[C]. *2016 IEEE Symposium on Security and Privacy (SP)*, 2016: 582-597.
- [52] N. Papernot, P. McDaniel. Extending defensive distillation[ET/OL]. 2017: arXiv preprint arXiv:1705.05264.
- [53] K. Grosse, P. Manoharan, N. Papernot, et al. On the (statistical) detection of adversarial examples[ET/OL]. 2017: arXiv preprint arXiv:1702.06280.
- [54] H. Hosseini, Y. Chen, S. Kannan, et al. Blocking transferability of adversarial examples in black-box learning systems[ET/OL]. 2017: arXiv preprint arXiv:1703.04318.
- [55] Akhtar N, Liu J, Mian A. Defense Against Universal Adversarial Perturbations[C]. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018: 3389-3398.
- [56] W. Xu, D. Evans, Y. Qi. Feature squeezing: Detecting adversarial examples in deep neural networks[ET/OL]. *CoRR*, vol. abs/1704.01155, 2017. [Online]. Available: <http://arxiv.org/abs/1704.01155>



**王科迪** 于2017年在上海交通大学信息安全专业获得学士学位。现在上海交通大学电子与通信工程专业攻读硕士研究生学位。研究领域为计算机视觉和人工智能安全。研究兴趣包括: 人工智能安全和对抗样本。Email: [the\\_title@sjtu.edu.cn](mailto:the_title@sjtu.edu.cn)



**易平** 于2005年复旦大学计算机应用专业获得博士学位。现任上海交通大学网络空间安全学院副教授。研究领域为网络对抗。研究兴趣包括: 人工智能安全。Email: [yiping@sjtu.edu.cn](mailto:yiping@sjtu.edu.cn)