

# 深度学习模型鲁棒性研究综述

及安全攸关场景下的挑战和应对

## 1 摘要

深度学习方法在过去几年取得了飞速的发展，并应用到越来越多的业务场景，例如视觉障碍物感知、人脸识别、声音识别、智能家居等等。这篇综述将介绍近些年学术界、工业界在深度学习模型鲁棒性 [1] [2] [3] [4] [5] 的研究进展。我们从安全威胁着手，首先展示了各个具有代表性的业务场景在不同对抗样本攻击（例如数字空间白盒、物理世界白盒、云端黑盒等）以及正常的环境扰动时所面临的严重风险。其次，针对这些威胁，我们将介绍目前已有的模型鲁棒性的度量方法，以及提升模型鲁棒性的手段。同时我们也介绍了百度安全实验室在AI安全研究前沿的研究成果，包括物理世界与云端黑盒的AI模型攻击与检测，模型鲁棒性度量与提升等等。

## 2 简介

对抗样本攻击（Adversarial Example Attack）[5] [6]的出现，及其对深度学习模型预测的几乎100%的绕过率，使得深度学习模型在安全攸关(Safety-Critical)场景中的应用面临着严峻的挑战。目前业界通常把深度学习模型抵抗输入扰动并给出正确判断的能力称为鲁棒性 [2] [3] [4] [5]。本文包括了模型安全研究中如下几个方面的内容。

1. **细分AI模型面临的安全威胁。**一般而言有两类影响模型鲁棒性的威胁模型：a) 正常环境扰动对模型可靠性（Safety）的影响。以视觉模型为例，本文介绍了各种模拟环境变化的方法 [7]，比如光线亮度、对比度的变化，以及这些变换对模型预测所造成危害。b) 当有恶

Table 1: AI模型在不同应用场景下的安全威胁 数据集: ImageNet

Safety-正常环境概率扰动			
ResNet50	安全属性 雾化 抖动 亮度	扰动范围 人眼基本无影响 人眼基本无影响 人眼基本无影响	绕过率 100% 100% 68.8%
<hr/>			
Security-恶意攻击环境			
数字空间白盒			
VGG16	扰动方法 CW2	扰动范围 人眼基本无影响	绕过率 100%
<hr/>			
物理世界白盒			
YOLOv3	攻击手段 贴片攻击	扰动范围 贴片区域内	结果 使车辆消失
<hr/>			
云端黑盒			
Google Cloud Object Localization	攻击方法 Dispersion Attack Targeted Attack	黑盒访问次数 2 100 2 100	绕过率 33% 86% 16% 65%

意攻击者存在情况下对模型安全性（Security）的影响。按攻击者的攻击目标和攻击能力的大小的不同，产生的攻击效果也不一样，从而导致AI模型在不同威胁态势下对鲁棒性的要求也会有差异。从攻击目标的难易角度来讲，非定向攻击相对简单，而定向攻击比较复杂，后者所造成的攻击后果要比前者严重许多。从攻击者的能力来讲，越是对模型内部架构和参数了如指掌（白盒），攻击者越能容易的制造出产生严重后果的对抗样本；而相对的黑盒条件下，攻击者只能依赖对抗样本的攻击传递性 [8]对目标发起攻击，攻击难度会大幅度提升。

2. **揭示AI模型安全威胁对安全攸关场景造成的严重威胁。** Table 1展示的是AI模型在不同安全威胁下面临的安全风险。例如在正常环境的概率扰动下，只要少许的扰动即可对模型造成100%的绕过；同样在数字空间白盒情况下，欺骗模型所需的扰动可以是非常细微的。在物理世界里，也可以通过贴片式攻击，对目标检测模型造成漏检。我们在Blackhat Europe 2018的会议上演示了通过制造对抗样本使得物理世界的车从YOLOv3面前消失 [9]。而对于云端黑盒模型，我们只需要极少的访问数量即可造成黑盒模型的绕过，这一工作也在Blackhat Asia 2019大会上做了演示 [10]（详情参见Section 6）。另外，本文也讨论了其他安全攸关应用场景所面临的安全风险，如智能驾驶交通标识分类（LisaCNN [11]）、Google Voice声音识别和自然语言处理等场景。
3. **提出模型鲁棒性的标准度量和评测。** 模型安全威胁的多样性使得对生产线上的AI模型鲁棒性评估尤为必要。而目前的鲁棒性评估大多是围绕 $L_p$ 范式扰动的基于理论的衡量。德国的Tubingen大学提出的Robust Vision Benchmark [12]提供了公共的自动化平台，该平台聚集了各种已知的对抗攻击算法，为大部分开源的深度学习模型以及用户训练好后并上传的模型打分。而百度安全实验室提出的标准度量和评测工具Perceptron Robustness Benchmark [13]是一个更全面，更实用的一款评测工具，具备以下独特的优势：a) 统一接口：支持各类深度学习框架包括Tensorflow，Pytorch，PaddlePaddle等。b) 攻击代码透明：支持通用的攻击接口，不但为用户提供统一的方式接触不同攻击算法，而且为未来的攻击算法提供灵活的扩展。c) 任务多样化：支持包括图像分类，目标检测等多种视觉模型的评测。同时也支持不同威胁模型下（白盒、黑盒）的鲁棒性评估。d) 模型评估标准化：确立了不同视觉感知任务的鲁棒性度量的标准化。
4. **加强模型鲁棒性以及模型对抗防御的思路和挑战。** AI模型对抗的防御仍然是一个开放的研究领域，大多数公开的防御方法都是基于经验主义且非常容易被击败 [14]。我们也讨论了目前AI模型对抗防御的思路。首先是模型鲁棒性的形式化验证 [15] [16] [17]给出了鲁棒性上界的理论证明的完备性，解决了样本测试无法全面覆盖的难题。其次，在通过对抗训练提升

模型鲁棒性的同时，在训练过程中引入鲁棒性度量 [18] [19]。最后，对输入数据的合法性检测，利用对抗样本本身相对较弱的抗干扰能力，利用各种变换，并比较变换后的结果使得对抗样本可以更容易的被检测出 [20]。

越来越多的研究证明，AI的生态面临对抗攻击严峻的挑战，在日益普及的视觉、声音、自然语言模型都会对特定的对抗输入样本给出错误的预测或分类。我们希望通过这一篇对深度学习模型鲁棒性的综述，剖析其重要性，以及它带来的挑战，为研究人员提供一个铺路石。作为安全从业者，我们认为在安全攸关场景下，模型鲁棒性和模型准确性同等重要，并呼吁整个业界把鲁棒性作为评估模型除了准确性之外的一个新的维度，同时把模型鲁棒性评估标准化。

在接下来的文章里，Section 3细分了模型安全威胁；Section 4介绍目前深度学习鲁棒性研究的现状；Section 5为大家讨论抵御对抗攻击的方法和挑战；Section 6展示了百度安全实验室的研究成果。

### 3 深度学习模型面临的安全威胁

我们把深度学习模型安全威胁就两种不同的威胁模型分别进行讨论。1) **模型可靠性-Safety**（没有恶意攻击者存在的正常业务环境下）。在输入数据与训练数据的分布有偏离的时候，模型给出错误的判断所造成的安全威胁。2) **模型安全性-Security**（有恶意攻击者存在情况下）。攻击者可以通过白盒的方法了解模型架构和参数，使用梯度下降的方法对输入数据进行扰动，并生成对抗样本。该对抗样本可以误导模型做出攻击者设定的预测结果，从而导致严重的安全事故。另外，攻击者也可以在不需要了解模型的前提下实施黑盒攻击，达到同样的攻击效果。

#### 3.1 模型可靠性-Safety（没有恶意攻击者存在的正常业务环境下）

**图像分类场景。**几乎所有state-of-the-art的图像分类DNN模型（Table 2第4行所展示的Xception, VGG-16, VGG-19, Resnet-50, Mobilenet, Inception-V3）对正常的图像可以做出正确的分类（原始图片和正确分类在Table 2第一行中展示）。当对原图像进行一些与真实环境类似的模拟变换后 [7]（Table 2第三行所展示，例如遮蔽变换，光线明暗变化，缩放变换），模型会给出错误的分类结果（例如Table 2第2行所示，放药的橱柜被分类为微波炉，搬运车被分类为宇宙飞船，盛汤的碗被分类为冰激凌等等）。可以说通过简单的变换就可以削弱图像分类的DNN模型的可靠性。

**无人机视觉感知场景。**K. Pei et al. [21]探索了环境安全属性（例如光线明暗，对比度）对无人机视觉感知可靠性的重要程度。Figure 1a, 1b, 1c, 1d展示的是该论文中两个出

Table 2: 不同现实模拟变换对图像分类DNN模型影响



第一行的原始图片（正确分类在标识中显示）经过以下各自对应的变换

Occulation 遮蔽变换	Translation 空间变换	Brightness 光线明暗	Contrast 色彩对比度	Shear 剪切变换	Scale 缩放变换
↓	↓	↓	↓	↓	↓

把第二行的图片（变换后的结果）输入以下各自模型进行识别

Xception	VGG-16	VGG-19	Resnet-50	Mobilenet	Inception-V3
----------	--------	--------	-----------	-----------	--------------

模型预测错误的分类结果在第二行图片的标识中显示

自Nvidia和Udacity的End2End的无人车视觉深度学习模型[22]分别对输入图片中的行驶方向进行预测的结果。Figure 1a和1c展示的是这两个模型对正常的原始输入图片所分别做出正确的向左，向右的预测。而Figure 1b展示的是当对原始图片调暗光线，Nvidia Dave-2会做出向右的错误预测，而1d所展示的是Udacity Rambo模型在图片平滑后，会错误的给出直行的决策。如果这些决策是在真实行进中的车辆对摄像头实时采集的数据所采取的应对发生错误时，将极有可能造成的车毁人亡的严重后果。

**商用的视觉API场景。**具有代表性的Google Cloud Vision API为用户提供预先训练好的模型对图像进行分析，同时用户也可以使用其提供的AutoML Vision模块训练自定义的模型。即便是商用的预先训练好的模型同样存在类似的可靠性威胁。Figure 2展示的是把图像输入给Google Cloud Vision API得到的结果，当把旋转木马的图片（Figure 2a）旋转一定的角度后，Google Cloud Vision API返回的结果是一个未知的生物体(Figure 2b)。而把一个蜂巢的图片(Figure 2c)加了雾化处理后，Google Cloud Vision API 返回的结果是一个无脊椎动物(Figure 2d)。从这些例子看出，影响模型可靠性的变换可以是多种多样。我们把这些变换称作影响模型可靠性的变换属



Figure 1: 现实环境模拟变换对无人车视觉感知模型的影响, (a)和(b)使用的是Nvidia Dave-2 Self-driving car Platform, (c)和(d)使用的是Udacity self-driving car Rambo DNN model。



Figure 2: 现实环境模拟对Google Cloud Vision API的影响

性。Table 3 汇总了目前学术界已经尝试的12类不同的变换属性[7]。这些变换模拟了真实世界图像的失真、噪点、变形。而在高可靠性需求的应用场景中的视觉分类或感知系统必须对这些图像失真、噪点和变形做出正确的响应及判断。总而言之，这些变换大致可以分为三类。1) 基于卷积的变换，2) 基于像素点的变换，3) 基于几何空间的变换。百度安全实验室已经尝试并掌握了各种基于视觉的环境变换模拟，为加固并提升深度学习模型在正常业务环境下的鲁棒性提供有效的基础。

### 3.2 模型安全性-Security (有恶意攻击者存在情况下)

与模型可靠性相比，模型的安全性有着更高的安全标准。在恶意攻击者参与的情况下，模型本身被攻击的可能性大大的增加。我们从白盒和黑盒的两种情况来分别论述恶意攻击对模型的威胁。白盒是指攻击者可以获得模型的结构，参数甚至训练数据，并根据这些信息实施攻击。而黑盒则相反，攻击者在没有这些信息的时候，通过各种手段破坏模型安全性。最后我们还将讨论如何将理论上可行的攻击手段在真实的物理世界中复现的可行性及严重后果。

Table 3: 影响DNN可靠性的三种变换，以及12种安全变换属性

1) 基于卷积的变换 – 利用卷积核对原始图片实施变换			
Average Smoothing 平均平滑变换	Median Smoothing 中位平滑变换	Erosion 侵蚀变换	Dilation 膨胀变换
2) 基于像素点变换			
Contrast 色彩对比度		Brightness 光线明暗	
3) 基于几何空间变换			
Occlusion 遮蔽变换	Rotation 旋转变换	Reflection 对称变换	Shear 剪切变换
Scale 缩放变换	Translation 空间变换		

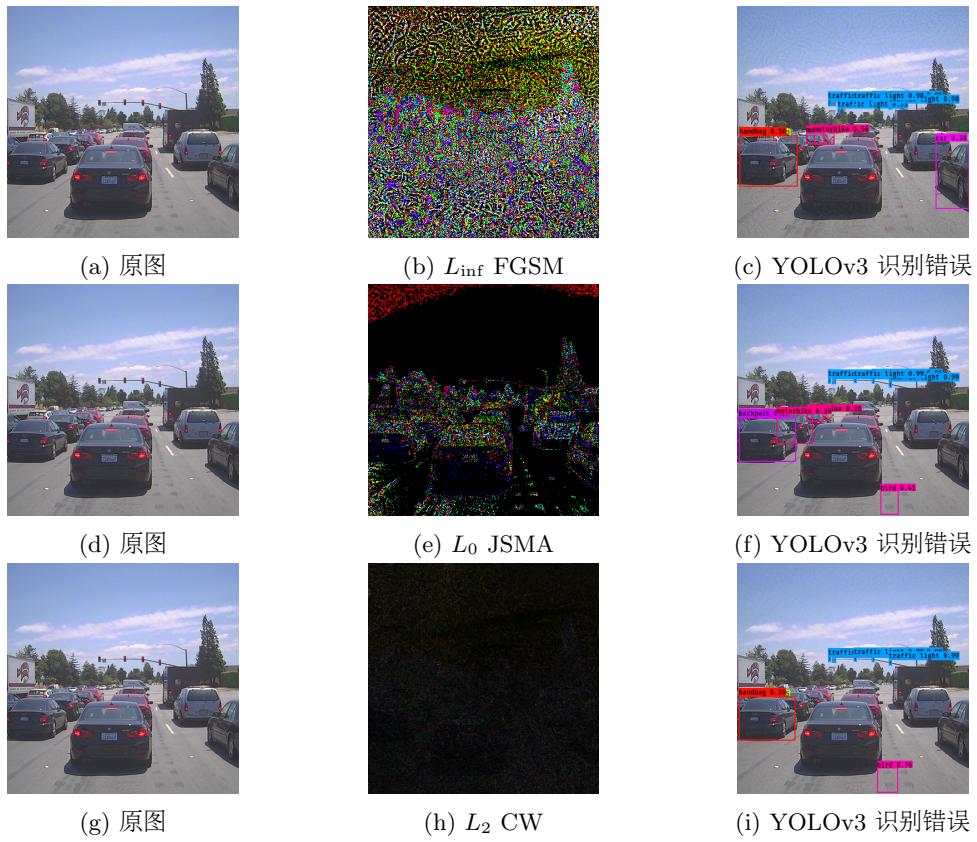
### 3.2.1 数字空间基于白盒的攻击威胁

当攻击者对DNN模型的架构和参数了如指掌的时候，他所要做的事情就变得相对简单。攻击者只需在目标神经网络所代表的高维空间里通过系统的优化方法在输入空间里寻找到一个对抗样本。这类对抗样本通常来说是不容易通过对输入数据的随机采样来获得，因为这些对抗样本只存在于高维空间里特定的分布区域。

从理论上，我们可以这样描述对抗样本的优化生成方法。假设函数  $f : \mathbb{R}^m \rightarrow \{1..k\}$  是一个分类器，它的输入是代表图像的像素向量，输出是离散化的分类结果集。同时我们假设函数  $f$  有一个与之对应的损失函数  $loss_f : \mathbb{R}^m \times \{1..k\} \rightarrow \mathbb{R}^+$  在给定一个输入图像  $x \in \mathbb{R}^m$ ，以及目标标记  $l \in \{1..k\}$ ，我们要解决以下的优化问题：  $\text{minimize} \|r\|_p$ , 并同时满足两个条件 1)  $f(x + r) = l$ , 2)  $x + r \in [0, 1]^m$ , 其中  $r$  是可以被最小化的距离。而一般来讲， $x + r$  是离输入图像  $x$  最近且最相像的，且被  $f$  分类为  $l$ 。当  $f(x) \neq f(x + r)$  时， $x + r$  即为生成的对抗样本。

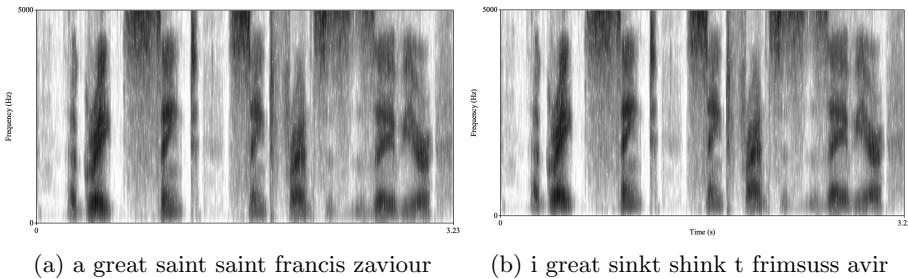
接下来，我们把白盒生成的对抗样本应用到以下两个具有代表性的场景中从而揭示其对业务场景的潜在威胁： 1) 障碍物识别， 2) 语音识别。

**障碍物识别场景** 障碍物识别的流行的方法之一是YOLOv3[23]。该模型训练数据用的是微软的COCO DATASET[24]，一共有80个不同的物体类别，其中包括与行车相关的物体。Figure 3 展

Figure 3: 用 $L_p$ norm生成的对抗样本对YOLOv3的威胁

		DNN	LR	SVM	DT	kNN
Source Machine Learning Technique	DNN	38.27	23.02	64.32	79.31	8.36
	LR	6.31	91.64	91.43	87.42	11.29
SVM	2.51	36.56	100.0	80.03	5.19	
DT	0.82	12.22	8.85	89.29	3.31	
kNN	11.75	42.89	82.16	82.95	41.65	

Figure 4: 不同模型之间的攻击样本可传递性



(a) a great saint saint francis zaviour      (b) i great sinkt shink t frimsuss avir

Figure 5: 声音图谱输出。转译目标为: A Great Saint Saint Francis Xavier.

示的是通过三个具有代表性的攻击算法 (FGSM[3], JSMA[25], CW2[2]) 所生成的扰动在与原图叠加后的效果。FGSM是基于 $L_{\infty}$ 距离标准, 代表输入图片中每个像素值所能改动的最大变换范围。JSMA是基于 $L_0$ 距离标准, 代表着输入图片中改动后对分类结果影响最大的一组像素, CW2是基于 $L_2$ 距离标准, 攻击者对输入图片的任意像素做极小的改动保证改动前后的欧式距离足够小。不管是遵从哪一种距离标准所生成的对抗样本, 都可以让人眼几乎无法区分它们与原图的区别, 但这些叠加后的图却能使得YOLOv3对车辆做出错误的识别。

Table 4: 声音识别定向攻击

Manual Transcription	Adversarial Target	Adversarial Prediction
a great saint saint Francis Xavier	a green thank saint frenzier	a green thanked saint fredstus savia
no thanks I am glad to give you such easy happiness	notty am right to leave you soggy happiness	no to ex i am right like aluse o yve have misser

**语音识别场景** 自动语音识别的技术一开始是在不同的模块分别建模, 比如声音模型, 语言模型, 和发声模型等, 每一个部分都是一个独立的模型。近些年, 基于深度学习的端到端的语音识别直接把语音片段当作输入, 直接输出转译后的文字。Cisse, et al. [26]搭建了一个由两个卷积层, 七个双向长短期记忆循环神经网络, 和一个全连接层的深度学习网络模型。并以Connectionist Temporal Classification (CTC) [27]损失函数为目标进行优化。一般来说, 衡量语音识别准确率的标准是两个, Word Error Rate (WER) 和Character Error Rate (CER)。基本是计算段落中删除, 添加, 替换操作的次数和目标长度的比率。这个比率有一个专用名词叫Levenshtein距离 [28]。该模型在Librispeech数据集上[29]获得了12%的WER和1.5%的CER。该论文提到了两类攻击, 一个是不定向攻击, 另一个是定向攻击。Cisse et al. [26]提出不定向攻击使用Houdini损失函数的方法生成的对抗样本造成WER是原先的2.3倍, 而CER是原先的1.8倍。如图所示, Figure 5a是原始

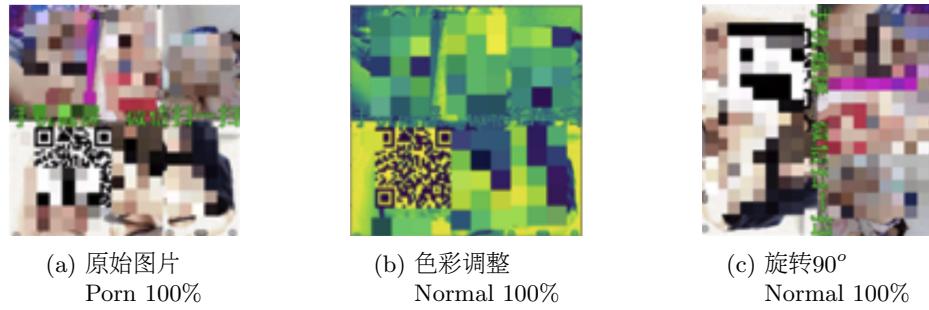


Figure 6: 色情图片经过仿射变换后成功欺骗鉴黄API

声音样本, Figure 5b是对抗样本。对抗样本的转译结果与原始结果有着明显的差异。与此同时对抗样本从声音图谱的角度来比较, 并不能明确的区分两者有多大的区别。定向攻击的目的是将输入的语音片段转译为攻击者指定的文字, Table 4展示的是定向攻击的结果。模型对于原始的声音输入可以完全准确的还原其文字, 对于音效很接近的对抗样本来说, 模型可以输出接近于攻击者预先设定的文字, 而对于音效差异太大的样本来说定向攻击就变得十分具有挑战。

### 3.2.2 基于黑盒的攻击威胁

黑盒攻击是指模型的架构和参数不被人所知的情况下, 攻击者对模型的欺骗。攻击者所依赖的是对抗样本有效攻击的可传递性属性。当对抗样本可以对一种已知的深度学习模型架构实施有效的白盒攻击的情况下, 该对抗样本依然也会对其他类似结构的深度学习模型同样具备攻击性。一般来说, 攻击者会针对被攻击的黑盒模型, 训练一个替代模型, 该替代模型的架构是由攻击者选择并往往是在应用场景中普遍使用的模型结构。然后按照白盒攻击的手段生成对抗样本。Papernot et al.[8]展示的是不同模型之间的攻击可传递性的结果 (Figure 4) 。y轴代表替代模型的算法, x轴代表被攻击的模型算法。表格里的每一个格子里的数字是所有生成的对抗样本成功欺骗被攻击的模型算法的百分比。可以看出同一个对抗样本, 对即使是不同的机器学习算法训练后的模型来说都有一定的攻击效果。接下来我们就两个常用的机器学习黑盒场景来讨论其各自的安全威胁。

**云端商用机器学习 API 的黑盒威胁** 目前Google, Microsoft, Amazon等AI公司都在其自己的云服务里提供了机器学习的服务。这些Machine Learning As a Service (MLaaS) 里基本囊括了大部分的算法及其应用的场景。例如人脸识别、图像识别、声音识别、自然语言处理等等。并且大多都提供了机器学习的预先训练好的模型供用户使用。其中Amazon会提供部分信息说明模型是用的什么架构算法, Google训练好的模型基本不会给出任何信息。我们把类似Google的Machine Learning API归类为黑盒模型。百度安全实验室经过研究发现, 通过简单的仿射变换能欺

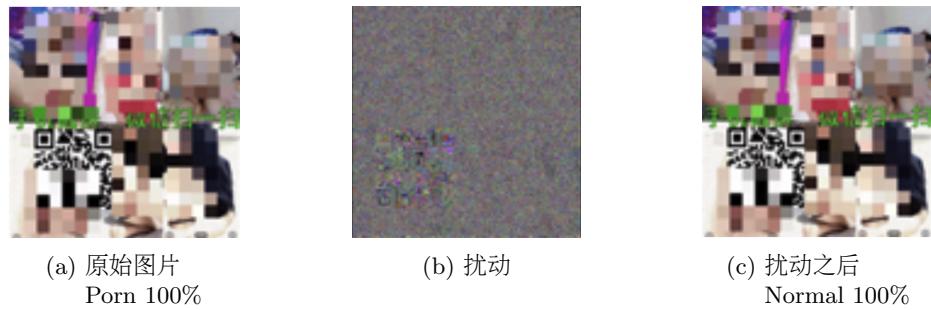


Figure 7: 色情图片经过扰动后成功欺骗鉴黄API

骗Google SafeSearch API，从而对输入图片是否包含色情内容做出错误的判断。另外和我们预期一样，通过添加扰动的方式，该API同样被轻易的欺骗。Figure6 和7展示的是我们的实验结果。在整个实验中，我们不需要了解API背后具体的模型信息。当我们把原始的色情图片的色彩做一番变化，或者简单的旋转90度，便能成功通过API的审查。同样的在原始色情图片中添加人眼几乎不能感知的扰动，也能够成功通过鉴黄API的审查。

**语音识别的黑盒威胁** 在语音识别转译的应用中，Google Voice接受用户的语音输入，并转译成文字。攻击者不需要知道其背后使用的是哪一类模型架构，及其参数，他可以利用已经开源的DeepSpeech2 [30]的模型生成与原始样本从声音上非常接近的对抗样本声音数据。这两个声音文件在普通人耳中没有区别，但是当它们分别在Android的APP上播放，并经过Google Voice的转译。那些原始的声音样本绝大多数被正确的转译，但是经过处理的对抗样本在转译中与原文有较大的差异。Figure8展示的是正常声音样本和对抗声音样本在经过Google Voice转译后的比较。（高亮显示的是它们之间的不同）可以看出原始正常的声音样本经过转译后和原始声音样本差别不大，但对抗样本转译后与原始样本的差别在50%以上。

### 3.2.3 物理世界基于白盒的攻击威胁

之前讨论的白盒场景下的基于视觉的障碍物识别和语音识别模型的对抗攻击是对数据本身直接进行修改，我们称之为数字攻击。而此类攻击往往假设攻击者具备图像或声音接收感应器（摄像头，麦克风）的控制能力，并能够修改和处理这些感应器接收后的数据。而通常情况下，攻击者是很少有这方面的能力。更多的情况下，攻击者往往需要对真实世界的物体做一些变化，使之可以被感应器接收到，并可以实施攻击。这种攻击的难度更高，受自然环境变换的影响更大。

**针对视觉图像感知的物理攻击。** 我们已经可以通过实验证明把生成好的对抗样本图像贴在目标物体上，并把它作为输入传递给摄像头后，可以成功欺骗后台的深度学习模型做出错误的

**Groundtruth Transcription:**  
The fact that a man can recite a poem does not show he remembers any previous occasion on which he has recited it or read it.

**G-Voice transcription of the original example:**  
The fact that a man can **decide** a poem does not show he remembers any previous occasion on which he has **work cited** or read it.

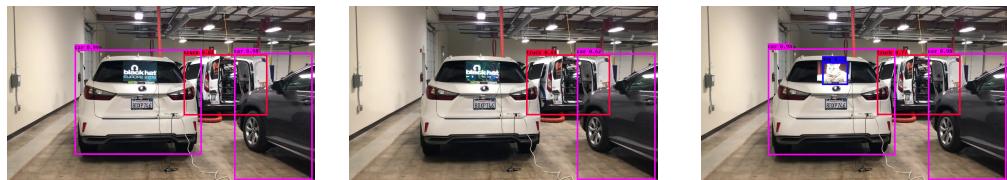
**G-Voice transcription of the adversarial example:**  
The fact that **I can rest I'm just not sure that you heard there is** any previous occasion **I am at he has your side** it or read it.

**Groundtruth Transcription:**  
Her bearing was graceful and animated she led her son by the hand and before her walked two maids with wax lights and silver candlesticks.

**G-Voice transcription of the original example:**  
The bearing was graceful **an** animated she **let** her son by the hand and before he walks two maids with wax lights and silver candlesticks.

**G-Voice transcription of the adversarial example:**  
**Mary was grateful then admitted** she **let** her son before **the walks** to Mays would like slice furnace filter count six.

Figure 8: Google Voice 分别对正常声音数据和对抗样本数据的转译结果[26]



(a)  $t_0$  : 车后显示BH Europe Logo YOLOv3 正确辨识      (b)  $t_1$  : 车后显示扰动的图案, YOLOv3无法识别      (c)  $t_2$  :车后显示正常猫的图案, YOLOv3 正确辨识

Figure 9: 固定距离, 固定视角物理攻击



(a)  $t_0$  : 车后显示扰动后的图片, (b)  $t_1$  : 摄像头zoom in, 并横向位移, YOLOv3无法正确识别      (c)  $t_2$  : 摄像头继续zoom in, YOLOv3依然无法正确识别

Figure 10: 不同距离, 不同视角的物理攻击

Transformations	Parameters	Parameter ranges
Translation	$(t_x, t_y)$	(10, 10) to (100, 100) step (10, 10)
Scale	$(s_x, s_y)$	(1.5, 1.5) to (6, 6) step (0.5, 0.5)
Shear	$(s_x, s_y)$	(-1.0, 0) to (-0.1, 0) step (0.1, 0)
Rotation	$q$ (degree)	3 to 30 with step 3
Contrast	$\alpha$ (gain)	1.2 to 3.0 with step 0.2
Brightness	$\beta$ (bias)	10 to 100 with step 10
Averaging	kernel size	$3 \times 3, 4 \times 4, 5 \times 5, 6 \times 6$
Gaussian	kernel size	$3 \times 3, 5 \times 5, 7 \times 7, 3 \times 3$
Blur	Median	aperture linear size 3, 5
	Bilateral Filter	diameter, sigmaColor, sigmaSpace 9, 75, 75

Figure 11: 计算机图形学各类模拟变换方法

预测。Figure 9展示了固定距离，固定角度的针对YOLOv3的物理攻击。在时间 $t_0$ 的时候，当在车后显示器中显示正常logo时，YOLOv3可以正确识别目标车辆。而在 $t_1$ 时，我们切换到扰动后的图片时，它可以立刻让目标车辆在YOLOv3面前变得无法辨识。在 $t_2$ 时，我们切换回正常的图片，YOLOv3重新可以识别目标车辆。其次，我们进一步演示了在动态变化的情况下（多角度，不同距离如 Figure 10所示）的物理攻击，从 $t_0$ 到 $t_2$ 之间，摄像头经历了由远到近，由左到右的位移变化。在此时间段内，在车背后显示的扰动后的图片保持不变。在绝大多数时间内，攻击扰动样本让YOLOv3无法识别目标车辆。百度安全实验室在Blackhat Europe 2018会议上对基于视觉的目标检测模型物理攻击的挑战和方法做了深入的研讨，并成功展示了物理攻击的实例 [9]。

针对语音识别的物理攻击，特别是对原数据做扰动然后通过扬声器播放后被智能音箱捕获，并达到语义篡改的整个过程，目前尚未有成熟的物理攻击实例。但是其他针对智能音箱的物理攻击在学术界已经有深入的研究，例如合成一个新的语音音频数据，要么这个数据完全超出普通人的听觉频率 [31]，但可以被麦克风捕捉到并处理识别；或者这个音频数据对普通人来讲基本无法辨别 [32]，但可以被语音识别系统错误的音译为有效的提示音或是声音命令。此类攻击可以认为与模型对一个非法输入产生了一个具有较高置信度的预测结果的情况类似。我们在语音识别的黑盒攻击威胁的讨论中所涉及的对Google Voice的攻击，也是一个针对黑盒的物理攻击成功案例。

## 4 深度学习鲁棒性研究现状

### 4.1 深度学习模型鲁棒性的度量

目前对深度学习模型鲁棒性的度量主要有两方面：1) 针对正常业务场景下的模型可靠性。2) 针对

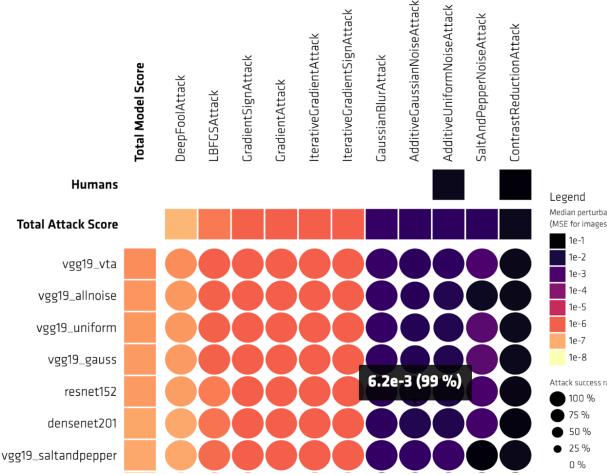


Figure 12: Robust Vision Benchmark

恶意攻击环境下的模型安全性。在安全攸关 (safety-critical) 的重要的场景，仅仅依靠传统的模型准确性评估远远不能达到令人放心的效果。在正常环境的概率扰动下，例如低能见度、光线过强等都是小概率事件，而且在训练测试集中可能占比很小，甚至不存在，但往往这些情况是能导致目标检测出现问题的。因此为了确保目标检测的安全性，还需要更多考虑正常环境小概率扰动情况的鲁棒性度量。同时，在一些以API或SDK形式提供服务的AI场景中，例如，Google Cloud SafeSearch API的图像内容审查功能、移动APP的人脸识别API等等，攻击者也可以出于各种目的绕过检测或认证，生成并向模型发送对抗样本。作为服务提供者，同时需要评估模型在恶意攻击环境下的安全性，保证模型不会轻易被对抗样本所欺骗。

在测量中，描述模型鲁棒性的指标通常是变换后的输入预测与原结果相比是否一致的百分比。目前针对深度学习在正常业务环境下的可靠性 (Safety) 度量，已有的方法主要是通过穷举现实中可能出现的输入扰动，并在实验环境中通过计算机视觉等技术模拟这些扰动，然后观察模型的输入是否在这些扰动下始终稳定。比如Yuchi et al.[33]举例了无人车场景的可能出现的几种环境因素的变化，如光照、能见度降低、摄像头的抖动等，并利用计算机图形学的一些方法去模拟这些真实的环境因素变化。如Figure 11所示，文中提出了使用图像的旋转、亮度、对比度、高斯模糊等技术，可以模拟许多无人驾驶场景中的不稳定因素，并用来测试模型的鲁棒性。在Section 6.2我们也会介绍有着更全面的鲁棒性度量指标的Perceptron Robustness Benchmark工具。

衡量恶意攻击环境下的安全性 (Security)，通常有两种不同的标准：1) 面对对抗样本模型预测的准确性。2) 对抗样本要达到欺骗目的所需要的最小扰动。Figure 12是德国 Tübingen 大学推

出的Robust Vision Benchmark对于几种常见的图像分类模型所做的安全性（Security）度量结果。横轴是测试中包含的对抗样本生成策略，包括6种白盒攻击以及5种黑盒攻击，而纵轴是被测模型。每个模型颜色的深浅，代表某种攻击想要成功欺骗该模型所需的最小扰动（平方差估计），而圆形的大小代表着攻击在生成一定数量的对抗样本的情况下成功率。用这种方法，可以横向的比较用不同网络结构，不同数据集，不同训练方法得到的模型在恶意攻击环境下的鲁棒性。

## 4.2 深度学习鲁棒性的形式化验证

由于深度学习模型的超高维特性和复杂的应用场景，如果使用样本测试的安全验证方法，我们无法确定选出的有限样本是否具有全覆盖的代表性，所以不能够有效的证明其鲁棒性。而单纯的遍历验证的方法在使用中也难以实现，于是形式化验证成为这方面研究的焦点。根据深度学习模型在不同环境下的不同应用，学术界从理论完备性和方法可实现性同时入手，沿着不同方向做了各种探索。

### 4.2.1 恶意攻击环境的鲁棒性验证

恶意攻击环境下的模型鲁棒性验证，从构造对抗性样本的角度考虑就是找出所需扰动的值域和下界，以下界数值的大小为度量来判定模型对恶意攻击的鲁棒性。对于扰动的大小一般放在 $L_p$ 测度的范畴内讨论， $L_p$ 测度的一般数学定义形式是： $\|x\|_p = (\sum_{i=1}^n |x_i|^p)^{1/p}$ 。在考虑对抗性扰动大小时常用的特殊 $L_p$ 测度包括：

- $L_0 = |x_1|^0 + |x_2|^0 + \dots + |x_n|^0$ : 代表两点对应维度值不相等的维度数量
- $L_2 = (\sum_{i=1}^n x_i^2)^{1/2}$ : 代表维度空间里的欧几里得距离
- $L_{\infty} = \max(|x_1|, |x_2|, \dots, |x_n|)$ : 代表两点对应维度值差别的最大值

目前对于这方面验证的研究，由于在多层的神经网络结构下无法推出扰动影响的准确形式表达，众多的研究工作都主要集中于使用不同的数值方法对扰动大小的范围进行估算，在本文中我们介绍一下最新的两种方法。

**方法一**[15]是在模型具有Lipschitz[34]连续性的基础上，使用Lipschitz常数限制出不改变模型输出的扰动范围，再运用统计概率的极值理论对对抗性扰动的下界进行估算。具体到从理论证明至数值计算可以分成这么几部分：

1. 从数学上对于任意的常用深度学习归类模型 $f$ 在输入样本 $x_0$ 的局部，构造函数 $g(x) = f_c(x) - f_j(x)$ ，这里的 $f_j$ 是 $f$ 对于类别 $j$ 的归类子函数， $c$ 是样本 $x_0$ 的类别 $c = \text{argmax}_j f_j(x_0)$ 。

2. 根据Lipschitz连续性定理我们知道在 $L_p$ 测度下存在Lipschitz常数 $L_q^j = \max\|\nabla g\|_q$ , 满足限制关系 $|g(x) - g(y)| \leq L_q^j \|x - y\|_p$ , 其中 $1/p + 1/q = 1$ 。
3. 由此可以证明在 $x_0$ 附近对抗性扰动的下界是  $\min_{j \neq c} (f_c(x_0) - f_j(x_0)) / L_q^j$ , 任何小于这个值的扰动都不能影响模型的归类结果, 这个下界就可以被用为衡量鲁棒性的指标。
4. 于是问题就在数学上转化为在以 $x_0$ 为中心的空间小球内数值估算 $\max\|\nabla g\|_q$ 。这时我们把 $\|\nabla g\|_q$ 视为一个随机概率的累积分布函数, 运用极值理论的Fisher-Tippett-Gnedenko[35]定理, 由于其最大值的存在, 可以得出结论其分布只可能是第三类的Reverse Weibull分布 (另两类分布没有最大值)。
5. 最后我们在空间小球内随机取所需样本值, 运用Maximum Likelihood Estimation, 就可以估算出分布, 再根据分布公式求得到这个最大值( $\max\|\nabla g\|_q$ )的数值估计, 从而得到了对抗性扰动的下界。

方法二是通过区间分析方法[16][36], 从初始加入扰动后的输入范围, 推导出模型输出结果的范围以及对结果的影响, 从而判断模型的鲁棒性。这个方法的数学理论基础主要是两个论证:

1. 首先把神经网络模型函数 $f$ 简单延拓成为区间域上的函数 $F$ , 使得延拓满足对于任何 $f$ 的输入值 $x$ ,  $F([x, x]) = f(x)$ 。那么 $F$ 的值域就是模型函数 $f$ 本身值域的超集, 这保证了区间运算得出的输出估计完全包含了原模型的值域。如果在这个估计上能验证鲁棒性, 也就是证明原模型的鲁棒性。
2. 其次是通过有限次细分输入区间得到的简单延拓函数的值域并集, 可以任意程度逼近原模型值域, 这样我们可以最大程度的在接近真实的结果上验证鲁棒性。

基于这两点论证我们就可以构造出一个区间计算循环递推的方法, 在给定输入区间上算出模型简单延拓函数的输出区间, 并对输出区间上进行判断。如果满足安全属性说明模型的鲁棒性对初始扰动免疫, 如果不满足就尝试随机寻找对抗性样本。如果未能找到对抗性样本就进入下一个循环, 继续细分输入区间, 从而使得到的输出区间更加缩小逼近于模型的实际值域, 然后再进行判断和样本寻找, 直至达到预设的最大循环数。这时由于没有能说明扰动免疫, 也没有确定的对抗性样本, 这表明该模型鲁棒性无法在规定时间尽可能逼近的范围内被有效形式化验证。

在上述两种方法之外, 其他研究者也做了很多不同的尝试。比如把形式化验证转换为Satisfiability Modulo Theory (SMT) [17]和大规模Linear Programming (LP) [37]的问题, 或者通过找出对抗性样本的数值特性来衡量鲁棒性。这些方法在具体实现的时候都遇到计算复杂的瓶

颈或者结果的值域过于放松的情况。对于这些问题上述的两种方法都做了较好的提高和解决。然而这些方法还远远没达到一个使安全性形式化验证通用并且实用的标准，其原因主要有以下三点：

1. 它们的论证过程中有对激活函数的是类似ReLU的限制性前提假设。
2. 运用极值理论估算使用了随机样本MLE求参数这本身的不确定性。
3. 区间分析方法在实现中对一些优化效果的依赖直接影响计算实现的复杂性。

#### 4.2.2 正常业务环境的鲁棒性验证

正常业务环境与恶意攻击环境不同，在正常业务环境下的鲁棒性研究主要关注在现实或物理上具有实际性和规律性的输入变动，比如光线和阴影对无人车物体识别的影响。所以鲁棒性验证主要是通过模拟各种实际业务场景中可能出现的不确定性，根据规则构造在定向范围内构造的变动，对模型的鲁棒性进行验证。比如Pei et al.[7]首先从理论上给出了一个通用的鲁棒性验证的形式化框架，然后针对智能图像识别的模型把输入变动限制在某些特定的图像转换上，比如边缘平滑，调光亮和对比度，旋转和剪切变换等等，并论证了由于像素的值域有限且离散，这些图像转换所得不同结果也是有限多个而且等同于多项式数量级，因此我们可以用有限个不同参数表示出所有可能的转换输出图像。于是研究者实现了一个用穷举方法进行鲁棒性验证的工具，并在实验中对常用的图像识别API和自训练模型进行了验证。根据所用的论证，这篇文章的方法在输入是离散和有界的其他机器学习模型上也可以实现类似的推广，然而这个方法的主要前提是验证的图像转换有完整的理解和数学表达，所以它在诸如物体识别等模型上是否可以广泛应用，主要的挑战在于我们能否完全解析应用场景而得到所有可能的图像转换或者输入变动的准确数学计算表达。

## 5 提高深度学习模型鲁棒性的途径

### 5.1 在模型训练过程中加入鲁棒性考量

提高模型鲁棒性的一个方法是在训练过程中加入对抗样本，而对抗样本是在每一个训练样本上用不同的对抗样本生成方法所生成的[18] [19]。加入对抗样本的目的是用更多的数据训练，增强模型的泛化能力。此外，还可以通过其他方法包括修改模型的激活函数或事损失函数， 使用network add-on，利用GAN，以及Defensive distillation等来提高模型的鲁棒性。具体可分为：

- 数据压缩：使用图像压缩的方法，减少对抗扰动对准确率的影响 [20]。

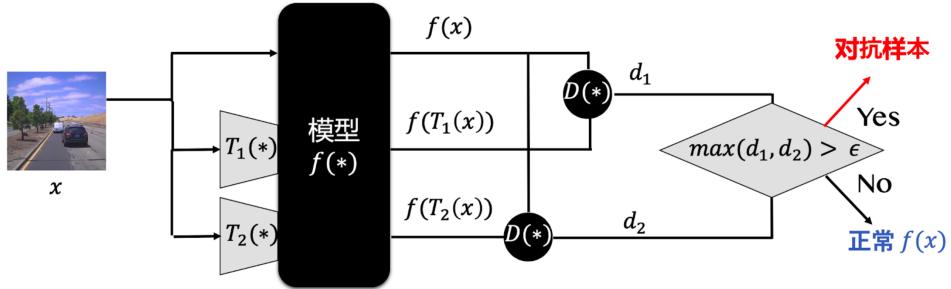


Figure 13: 常见的对抗样本检测流程,  $T$ 是对于输入所进行的变换,  $D$ 是预测结果不一致性的度量

- 深度压缩网络: 在网络中使用和压缩自编码器 (Contractive Auto Encoders) 类似的平滑度惩罚项 [38]。
- 基于 GAN 的防御: 用 Generative Adversarial Network 为基础训练网络来抵抗对抗攻击 [39]。

Table 5: 经过对抗样本训练后的攻击成功率

Dataset	输入维度	黑盒攻击成功率	白盒攻击成功率
MNIST	28x28	7.24%	11.44%
CIFAR10	32x32	36.61%	55.29%
ImageNet-2 classes	224x224	5%	4.4%

目前的攻击手段对未经鲁棒训练的模型有很高攻击成功率, 尤其是白盒攻击, 攻击成功率可达100%。基于鲁棒训练的方法在不同训练集上均使得训练模型鲁棒性有明显提升。如Table 5所示, 在MNIST [40], CIFAR10 [41], ImageNet-2classes [42]三种不同数据集的实验中, 不管白盒还是黑盒, 攻击成功率都有显著下降。尽管鲁棒训练可以降低攻击成功率, 攻击者仍然可以有足够的机会对模型造成严重的威胁, 所以在安全攸关的场景下直接使用这样的模型是不安全的。

## 5.2 对恶意攻击样本的检测

检测输入数据的合法性, 同样可以达到防御对抗样本攻击的效果。在基于深度学习的检测任务中, 一旦检测到对抗样本, 可以采取报告异常的方式, 防止其绕过。这在一些应用场景例如恶意软件的检测, 低俗内容的审查中具有重要意义。

Xu et al. [20]提出的“特征压缩”的方法利用了对抗样本的扰动通常过于精细, 导致过度拟合于



Figure 14: 特征压缩后的图片效果

用于生成的原始样本和目标模型的特点，在不影响人类观测者观测结果的前提下，将8个比特位表示的RGB图片用更低的比特位表示，大大降低了攻击者可以攻击的空间。如Figure 14所示，8个比特位表示的原图（最左），可以通过特征压缩，变成更粗粒度的图片（从左至右），但基本不会影响人类对图片的判断。这种压缩也通常可以使对抗样本的扰动无效，从而达到检测的目的。具体的检测方法如Figure 13所示，对于输入图片，模型提供商使用不同的变换方法 $T_1$ 、 $T_2$ 产生不同于原始输入的两个变换后的样本并交给模型进行预测。其中 $T_1$ 、 $T_2$ 可以是特征压缩变换，也可以是其他的变换。如果原始样本是正常样本，则三次的预测结果会高度一致。在这种情况下系统可以给用户返回正常的预测结果。而如果原始输入为对抗样本，则三次的结果会有比较大的差异。如果两两结果的差异的最大值高于某个阈值，则可断定其为对抗样本。

其它对抗样本检测的方法包括（1）基于局部本征维数Local Intrinsics Dimension（LID）[43] [44]的检测方法，利用对抗样本的LID值远高于正常样本的性质来识别对抗样本与正常样本。（2）模型可解释性方法Attacks Meets Interpretation(AMI) [45]，利用提高模型中对应人脸重要特征的神经元的权重的方法来检测对抗样本。以及（3）MagNet方法[46]，采用两步检测的措施，利用检测器检测扰动比较大的样本，直接丢弃；然后针对扰动量小的对抗样本，使用降噪将其转化成正常样本，最后再交由原模型识别。

### 5.3 现有模型安全防御的挑战

现有的深度学习模型安全防御存在着许多不足。上一章提到的鲁棒性的度量，验证，以及鲁棒性训练目前都有着各自的缺陷。

**首先，目前鲁棒性度量标准与真实场景仍存在差距。**现有的用于模拟正常业务场景可能出现的扰动的技术，通常与现实情境还是存在差别。比如单纯用增加RGB值的方法模拟现实中的光照变换忽略了不同颜色的区域对于光照反应的不同；或是用图片的左右旋转模拟摄像头的抖动并没有考虑到前后方向的抖动。因此，目前学术界提出的一些度量标准难以获得令人信服的鲁棒性测试结果。

**其次，现有的防御方案都不实际有效。**Table 6展示的是学术界近些年提出的一些防御的方

Table 6: ICLR 2018发表的对抗防御方法的无效。

防御	数据集	距离 (评测基准)	准确率
Buckman et al (2018) [47]	CIFAR	0.031 ( $L_{inf}$ )	0%
Ma et al. (2018) [43]	CIFAR	0.031 ( $L_{inf}$ )	5%
Guo et al. (2018) [48]	ImageNet	0.05 ( $L_2$ )	0%
Dhillon et al. (2018) [49]	CIFAR	0.031 ( $L_{inf}$ )	0%
Xie et al. (2018) [50]	ImageNet	0.031 ( $L_{inf}$ )	0%
Song et al. (2018) [51]	CIFAR	0.031 ( $L_{inf}$ )	9%
Samangouei et al. (2018) [52]	MNIST	0.005 ( $L_2$ )	55%
Madry et al. (2018) [18]	CIFAR	0.031 ( $L_{inf}$ )	47%
Na et al. (2018) [53]	CIFAR	0.015 ( $L_{inf}$ )	15%

法，以及这些方法在Athalye et al., [14]更有力的攻击下的失效。虽然在各自的数据集上，以及各自的评测标准下实验证明有一定的效果，但是严格来讲，这些方案都不具备完整性，且无效。考量模型鲁棒性应当解决的是安全威胁的worst case，而不是针对某一数据集，或是某一测量标准。在如何评估对抗攻击的防御机制是否有效，Carlini et al., [54]给出了指导性的意见。

**第三，目前鲁棒性验证方法难以拓展到复杂的模型。**深度学习模型的本质是大量的矩阵乘法与加法运算。因此，模型的验证本质上就是对原始样本的输入区间，以及一个扰动的范围，对该输入进行值域分析。准确的值域分析需要用到线性规划，而为了应对神经网络中的非线性单元，如ReLU单元，需要在每一个非线性运算时进行分段讨论，这给计算带来了很大的开销。因此，在最新的一些方案中，有人提出用放宽约束的方法，减少需要分段讨论的情况，从而加快验证速度，不过同时也会引入误报（false positive）。由于计算时间开销的限制，目前的模型验证的极限在约10万神经元的网络，大约9层，通常也只局限于简单的图像识别任务。

**最后，目前鲁棒性训练的方法在高维度输入上效果不佳。**鲁棒性训练的原理是在训练的过程中对每个样本生成若干个对抗样本，并一起加入到训练中，达到扩展输入空间，提升模型鲁棒性的目的，目前在MNIST数据集上，通过鲁棒性训练得到的模型，已经对基于像素扰动的攻击达到了超过80%的正确率。然而，想要在更高维度的输入，比如ImageNet数据集上训练鲁棒的模型，目前的训练方法还存在着一些问题。首先，相对于 $28 \times 28$ 的MNIST图片数据， $224 \times 224$ 的ImageNet图片有更高的输入维度，这也给攻击者提供了更大的空间去制造扰动。因此，鲁棒性训练过程中生成的对抗样本无法仍不足以达到给模型提供足够对抗样本的目的。如Table 5所示，目前在稍小的Cifar-10数据集上的鲁棒性训练模型仅能达到45%左右的准确性，而ImageNet上鲁棒性训练还没达到理想的结果。



Figure 15: 图像分类的物理攻击

## 6 百度安全实验室的相关研究

### 6.1 安全攸关 (Safety-Critical) 感知模型的物理攻击风险研究

当前障碍物识别的感知模型大多是基于视觉感知，即通过摄像头实时捕捉前方的图像，并通过深度学习模型完成物体识别并分类。我们针对视觉感知开源模型YOLOv3做了深入的剖析，就对抗样本是否能够对视觉感知场景做到真实的物理有效攻击进行了实质性的探索和实验，并证实了物理威胁的真实存在以及对模型鲁棒性的威胁。首先是对目标检测场景图像的简单数字扰动，在Section 3里，Figure 3所展示的是针对YOLOv3的数字攻击成功实例。我们分别使用了FGSM, JSMA, CW的方法对同一张原始图片进行扰动。基于像素点的改动的 $\Delta$ 在中间一列显示，而最右边一列显示的是叠加后的结果。经过数字扰动后的图片完全可以欺骗YOLOv3对图片内车辆的识别能力。

其次，我们成功实现针对图像分类模型(*LisaCNN*)的物理攻击。如Figure 15所示，通过在Stop Sign内部添加少量扰动，使得模型错误地将该标志识别为Speed Limit 45。在整个的演示过程中，约71.3%的帧被错误地识别为非Stop Sign。

为了实施对目标识别的物理攻击，我们选择了YOLOv3作为实时目标检测的工具和算法代表。其识别帧率和准确率在同类算法和方案里最接近Safety-Critical目标识别场景的需求。Figure 9, Figure 10中所展示的是百度安全实验室在Blackhat Europe 2019上展示的针对YOLOv3的物理攻击效果 [9]。我们实现了在不同位移不同角度中，对真实车辆实施贴图攻击，使得YOLOv3对眼前的车辆失去识别能力。

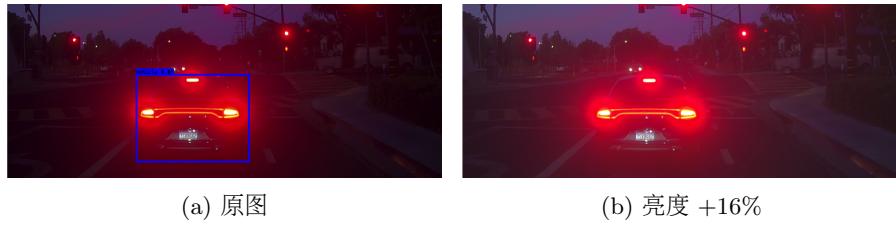


Figure 16

## 6.2 模型鲁棒性基准测试平台Perceptron Robustness Benchmark

目前市面上对于深度学习模型的鲁棒性缺乏标准化的评估手段。同时针对现有的AI安全工作大都集中在图像分类任务上这一现状，我们开发了将图像分类，目标检测等一系列视觉感知相关的任务囊括在内的基准测试平台—Perceptron Robustness Benchmark。它相对于其他对抗样本生成工具如 Google的Cleverhans [55]，以及IBM的ART [56]有着独特的优势，例如：a) 统一接口：支持各类深度学习框架包括Tensorflow，Pytorch，PaddlePaddle等。b) 攻击代码透明：支持通用的攻击接口，不但为用户提供统一的方式接触不同攻击算法，而且为未来的攻击算法提供灵活的扩展。c) 任务多样化：支持包括图像分类，目标检测等多种视觉模型的评测。同时也支持不同威胁模型下（白盒、黑盒）的鲁棒性评估。d) 模型评估标准化：确立了不同视觉感知任务的鲁棒性度量的标准。

Table 7: 安全可靠性属性1：(0.1, 10) 的亮度变化内，模型识别结果不变 mAP &gt; 0.8

	YOLOv3	YOLOv2	ResNet-50
Highway	82.1%	81.6%	31.2%
Local	100.0%	100.0%	69.2%
Dark	83.6%	80.5%	14.5%

目前Perceptron Robustness Benchmark已在百度内部开源，我们也已使用其对多种不同的图像识别与目标检测模型进行了初步的鲁棒性度量。例如Table 7，我们把图像识别最具代表性模型之一的Resnet50与目标检测模型YOLOv2、YOLOv3做了比较。在对于亮度鲁棒性的测试中，在不同的路况条件下（高速、市区、夜间）我们发现Resnet50对于光照变换的鲁棒性要远差于目标检测模型。还有原本可以被检测到的车（见Figure 16a），在亮度细微变换后，无法检测出目标物体（见Figure 16b）。我们现在正在进一步将更多的正常业务场进行的变换，以及恶意攻击方法添加进Perceptron Robustness Benchmark。

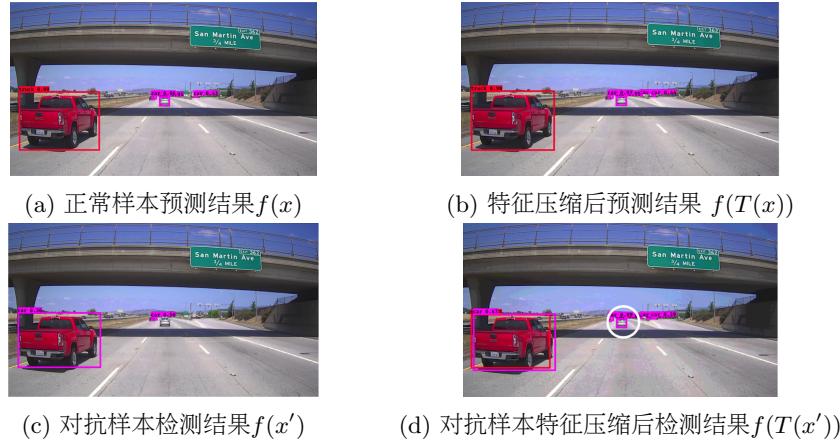


Figure 17: 特征压缩后的图片目标检测结果

### 6.3 安全攸关 (Safety-Critical) 场景的对抗样本攻击检测

在安全攸关 (Safety-Critical) 场景中，如果系统能够及时检测到对抗样本，可以避免后续的控制逻辑受到恶意样本的影响。因此，对抗样本的检测，可以作为提升模型鲁棒性的一种手段。对抗样本的检测主要利用了其与正常样本间局部平滑性 (local smoothness) 上的区别，这在之前的章节有详细讨论。

我们针对视觉目标检测模型设计了对抗样本的实时检测系统。Figure 17展示的是该方法检测对抗样本的效果。其中Figure 17a展示的是原始图片在目标检测模型的检测结果。Figure 17b展示的是通过特征压缩的转换后目标检测模型的检测结果。可以看到，这两者的检测结果是基本一致。Figure 17c展示的对原始图片做扰动后的图像在目标检测模型下的检测结果，其中明显的漏检了中间的车辆。而通过特征压缩对扰动的图片进行处理后（见Figure 17d），目标检测模型依然可以正确检测出中间的车辆。

在我们设计的预测结果相似度的度量下，我们在BDD [57]公开的数据集上单帧的对抗样本检测达到了低漏报率 ( $FN = 5\%$ ) 与低误报率 ( $FP = 2.5\%$ )。该结果已经超越了目前学术界在图像分类这样更简单的任务上的对抗样本检测的最优结果。另外我们还尝试了通过引入连续帧之预测结果之间的一致性检测进一步提高准确性，在测试中达到了0漏报，0误报的良好结果。实验证明生产环境中的实时对抗样本检测具有可行性。

### 6.4 针对AI云服务黑盒攻击的风险研究

多家大公司如Google, Amazon, Microsoft等都将AI能力作为一种云服务，以API的形式提供给用

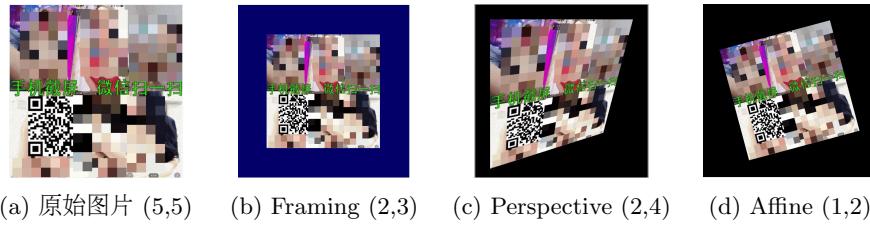


Figure 18: Google SafeSearch API 返回结果 (adult, racy)  
1-Very Unlikely, 2-Unlikely, 3-Possible, 4-Likely, 5-Very Likely

户或合作伙伴使用。这些服务于众多业务的模型通常在云端以黑盒的形式部署，仅向用户返回调用的结果。我们发现攻击者仍可以利用有限的信息构造对抗样本，从而绕过AI模型的检测。

我们实现了多种黑盒攻击的技术，并提出了一种对黑盒模型进行“指纹攻击”的方式，即根据极少的请求结果推测出模型的结构，并针对性的构造对抗样本的方法。通过对深度学习模型的特征提取层中的最后一层的神经元的离散值的最小化，从而使得目标分类的置信度降低，通过搜集的14个不同的公开模型（VGG16, VGG19, RESNET50, MobileNET等），分别构造输入样本使得对应的模型特征提取层的最后一层神经元的离散值最小化，并把该构造后的样本以API方式输入云端黑盒模型，并观察最后分类层的输出结果。最后，选取API返回结果中置信度最低的样本，并把生成该样本的网络架构作为云端模型的架构。当攻击者知晓云端模型特征提取层的架构之后，他就可以按照白盒的方式精确的构造对抗样本，从而对云端模型进行定向和非定向攻击。这些技术目前也已被集成进Perceptron Robustness Benchmark，并已被用来测试Google的AI平台，并发现了多种高危问题。

在对Google SafeSearch API的测试中发现，鉴黄API对图像的一些恶意的仿射与透射变换防御性相对比较弱，Figure 18展示的是攻击者可以很方便的使用这些变化达到绕过API的目的。针对该弱点，我们设计了数据增强与对抗训练的方法，通过在训练中添加对抗样本的方法去提升模型对这类攻击的抵抗力。

我们同时使用Perceptron Robustness Benchmark的黑盒攻击的技术测试了Google Cloud AI提供的Object Localization 服务，发现通过使用我们的“指纹攻击”方法，可以轻易达到欺骗这些API的目的，从而影响Google搜索结果，以及目标检测结果的鲁棒性。我们于2019年3月在新加坡的Blackhat Asia展示了我们的攻击成果 [10]。

## 7 结论

本综述揭示了在各个场景中广泛应用的深度学习模型的攻击实例。从Safety角度出发，正常环境的概率扰动可能造成严重的错误结果；从Security角度出发，作为白盒模型的目标检测模型YOLOv3，物理对抗样本可以使它无法有效探测面前的车辆；端到端的声音识别模型，例如Google Voice也会遭到语音对抗样本的误导；最具代表性的云端黑盒模型，只要对输入做变换便可绕过图像内容审查，对于更复杂的云端目标检测模型，也面临着迁移攻击的威胁。深度学习在安全攸关场景中应用的安全性受到严峻的挑战。我们希望这篇综述为广大AI研究人员和工程应用人员提供一个模型鲁棒性的阶段总结，并呼吁业界将模型鲁棒性作为AI模型预测准确性之外的一个重要测评标准，同时制定有效的模型安全标准和规范，为深度学习在安全攸关场景中的应用打下坚实的基础。

## References

- [1] O. Bastani, Y. Ioannou, L. Lampropoulos, D. Vytiniotis, A. V. Nori, and A. Criminisi, “Measuring neural net robustness with constraints,” in *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS’16, (USA), pp. 2621–2629, Curran Associates Inc., 2016.
- [2] D. W. Nicholas Carlini, “Towards evaluating the robustness of neural network,” in *Proceedings of the 38h IEEE Symposium on Security and Privacy*, 2017.
- [3] I. J. Goodfellow, J. Shlens, and C. Szegedy, “Explaining and Harnessing Adversarial Examples,” *arXiv e-prints*, p. arXiv:1412.6572, Dec 2014.
- [4] S. Gu and L. Rigazio, “Towards Deep Neural Network Architectures Robust to Adversarial Examples,” *arXiv e-prints*, p. arXiv:1412.5068, Dec 2014.
- [5] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, “Intriguing properties of neural networks,” in *International Conference on Learning Representations*, 2014.
- [6] B. Biggio, I. Corona, D. Maiorca, B. Nelson, N. Šrndić, P. Laskov, G. Giacinto, and F. Roli, “Evasion attacks against machine learning at test time,” in *Machine Learning and Knowledge*

- Discovery in Databases* (H. Blockeel, K. Kersting, S. Nijssen, and F. Železný, eds.), (Berlin, Heidelberg), pp. 387–402, Springer Berlin Heidelberg, 2013.
- [7] K. Pei, Y. Cao, J. Yang, and S. Jana, “Towards practical verification of machine learning: The case of computer vision systems,” *CoRR*, vol. abs/1712.01785, 2017.
  - [8] N. Papernot, P. McDaniel, and I. Goodfellow, “Transferability in Machine Learning: from Phenomena to Black-Box Attacks using Adversarial Samples,” *arXiv e-prints*, p. arXiv:1605.07277, May 2016.
  - [9] Z. Zhong, W. Xu, Y. Jia, and T. Wei, “Perception Deception: Physical Adversarial Attack Challenges and Tactics for DNN-based Object Detection,” *BlackHat Europe Briefing 2018*, Dec 2018.
  - [10] Y. Jia, Z. Zhong, Y. Zhang, Q. Feng, T. Wei, and Y. Lu, “The Cost of Learning from the Best: How Prior Knowledge Weakens the Security of Deep Neural Networks,” *BlackHat Asia Briefing 2019*, Mar 2019.
  - [11] K. Eykholt, I. Evtimov, E. Fernandes, B. Li, A. Rahmati, C. Xiao, A. Prakash, T. Kohno, and D. Song, “Robust Physical-World Attacks on Deep Learning Visual Classification,” in *Computer Vision and Pattern Recognition (CVPR)*, 2018.
  - [12] J. Rauber and W. Brendel, “Robust vision benchmark.” "<https://github.com/bethgelab/robust-vision-benchmark>", 2017.
  - [13] B. X-Lab, “Baidu perceptron robustness benchmark.” "<https://github.com/advboxes/perceptron-benchmark>", 2019.
  - [14] A. Athalye, N. Carlini, and D. Wagner, “Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples,” in *Proceedings of the 35th International Conference on Machine Learning* (J. Dy and A. Krause, eds.), vol. 80 of *Proceedings of Machine Learning Research*, (Stockholmsmässan, Stockholm Sweden), pp. 274–283, PMLR, 10–15 Jul 2018.
  - [15] T.-W. Weng, H. Zhang, P.-Y. Chen, J. Yi, D. Su, Y. Gao, C.-J. Hsieh, and L. Daniel, “Evaluating the robustness of neural networks: An extreme value theory approach,” in *International Conference on Learning Representations*, 2018.

- 
- [16] S. Wang, K. Pei, J. Whitehouse, J. Yang, and S. Jana, “Formal security analysis of neural networks using symbolic intervals,” in *Proceedings of the 27th USENIX Conference on Security Symposium*, SEC’18, (Berkeley, CA, USA), pp. 1599–1614, USENIX Association, 2018.
  - [17] R. Ehlers, “Formal Verification of Piece-Wise Linear Feed-Forward Neural Networks,” *arXiv e-prints*, p. arXiv:1705.01320, May 2017.
  - [18] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, “Towards deep learning models resistant to adversarial attacks,” in *International Conference on Learning Representations*, 2018.
  - [19] E. Wong, F. Schmidt, J. H. Metzen, and J. Z. Kolter, “Scaling provable adversarial defenses,” in *Advances in Neural Information Processing Systems 31* (S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, eds.), pp. 8400–8409, Curran Associates, Inc., 2018.
  - [20] W. Xu, D. Evans, and Y. Qi, “Feature squeezing: Detecting adversarial examples in deep neural networks,” *arXiv preprint arXiv:1704.01155*, 2017.
  - [21] K. Pei, Y. Cao, J. Yang, and S. Jana, “Deepxplore: Automated whitebox testing of deep learning systems,” in *Proceedings of the 26th Symposium on Operating Systems Principles*, SOSP ’17, (New York, NY, USA), pp. 1–18, ACM, 2017.
  - [22] M. Bojarski, D. D. Testa, D. Dworakowski, B. Firner, B. Flepp, P. Goyal, L. D. Jackel, M. Monfort, U. Muller, J. Zhang, X. Zhang, J. Zhao, and K. Zieba, “End to end learning for self-driving cars,” 2016.
  - [23] J. Redmon and A. Farhadi, “Yolov3: An incremental improvement,” *arXiv*, 2018.
  - [24] T.-Y. Lin, M. Maire, S. Belongie, L. Bourdev, R. Girshick, J. Hays, P. Perona, D. Ramanan, C. L. Zitnick, and P. Dollár, “Microsoft COCO: Common Objects in Context,” *arXiv e-prints*, p. arXiv:1405.0312, May 2014.
  - [25] N. Papernot, P. McDaniel, S. Jha, M. Fredrikson, Z. Berkay Celik, and A. Swami, “The Limitations of Deep Learning in Adversarial Settings,” *arXiv e-prints*, p. arXiv:1511.07528, Nov 2015.

- [26] M. M. Cisse, Y. Adi, N. Neverova, and J. Keshet, “Houdini: Fooling deep structured visual and speech recognition models with adversarial examples,” in *Advances in Neural Information Processing Systems 30* (I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, eds.), pp. 6977–6987, Curran Associates, Inc., 2017.
- [27] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, “Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks,” in *Proceedings of the 23rd International Conference on Machine Learning*, ICML ’06, (New York, NY, USA), pp. 369–376, ACM, 2006.
- [28] V. I. Levenshtein, “Binary Codes Capable of Correcting Deletions, Insertions and Reversals,” *Soviet Physics Doklady*, vol. 10, p. 707, Feb. 1966.
- [29] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, “Librispeech: an ASR corpus based on public domain audio books,” in *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2015.
- [30] D. Amodei, R. Anubhai, E. Battenberg, C. Case, J. Casper, B. Catanzaro, J. Chen, M. Chrzanowski, A. Coates, G. Diamos, E. Elsen, J. Engel, L. Fan, C. Fougner, T. Han, A. Hannun, B. Jun, P. LeGresley, L. Lin, S. Narang, A. Ng, S. Ozair, R. Prenger, J. Raiman, S. Satheesh, D. Seetapun, S. Sengupta, Y. Wang, Z. Wang, C. Wang, B. Xiao, D. Yogatama, J. Zhan, and Z. Zhu, “Deep Speech 2: End-to-End Speech Recognition in English and Mandarin,” *arXiv e-prints*, p. arXiv:1512.02595, Dec 2015.
- [31] G. Zhang, C. Yan, X. Ji, T. Zhang, T. Zhang, and W. Xu, “Dolphinattack: Inaudible voice commands,” in *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, CCS ’17, (New York, NY, USA), pp. 103–117, ACM, 2017.
- [32] N. Carlini, P. Mishra, T. Vaidya, Y. Zhang, M. Sherr, C. Shields, D. Wagner, and W. Zhou, “Hidden voice commands,” in *25th USENIX Security Symposium (USENIX Security 16)*, (Austin, TX), pp. 513–530, USENIX Association, 2016.
- [33] Y. Tian, K. Pei, S. Jana, and B. Ray, “Deeptest: Automated testing of deep-neural-network-driven autonomous cars,” in *Proceedings of the 40th International Conference on Software Engineering*, ICSE ’18, (New York, NY, USA), pp. 303–314, ACM, 2018.

- [34] R. Paulavičius and J. Žilinskas, “Analysis of different norms and corresponding lipschitz constants for global optimization,” *Technological and Economic Development of Economy*, vol. 12, pp. 301–306, 01 2006.
- [35] L. de Haan and A. Ferreira, *Extreme Value Theory: An Introduction (Springer Series in Operations Research and Financial Engineering)*. Springer, 1st edition. ed., 2010.
- [36] R. E. Moore, R. B. Kearfott, and M. J. Cloud, *Introduction to Interval Analysis*. Philadelphia, PA, USA: Society for Industrial and Applied Mathematics, 2009.
- [37] X. Huang, M. Kwiatkowska, S. Wang, and M. Wu, “Safety Verification of Deep Neural Networks,” *arXiv e-prints*, p. arXiv:1610.06940, Oct 2016.
- [38] D. Meng and H. Chen, “Magnet: a two-pronged defense against adversarial examples,” in *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, pp. 135–147, ACM, 2017.
- [39] G. K. Santhanam and P. Grnarova, “Defending against adversarial attacks by leveraging an entire gan,” *arXiv preprint arXiv:1805.10652*, 2018.
- [40] Y. LeCun and C. Cortes, “MNIST handwritten digit database,” 2010.
- [41] A. Krizhevsky, V. Nair, and G. Hinton, “Cifar-10 (canadian institute for advanced research),”
- [42] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, “ImageNet: A Large-Scale Hierarchical Image Database,” in *CVPR09*, 2009.
- [43] X. Ma, B. Li, Y. Wang, S. M. Erfani, S. Wijewickrema, G. Schoenebeck, M. E. Houle, D. Song, and J. Bailey, “Characterizing adversarial subspaces using local intrinsic dimensionality,” in *International Conference on Learning Representations*, 2018.
- [44] P.-H. Lu, P.-Y. Chen, and C.-M. Yu, “On the limitation of local intrinsic dimensionality for characterizing the subspaces of adversarial examples,” 2018.
- [45] S. Ma, Y. Liu, G. Tao, W.-C. Lee, and X. Zhang, “Nic: Detecting adversarial samples with neural network invariant checking,” in *Proceedings of the 26th Network and Distributed System Security Symposium, NDSS’19*, 2019.

- [46] D. Meng and H. Chen, “Magnet: A two-pronged defense against adversarial examples,” in *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security, CCS ’17*, (New York, NY, USA), pp. 135–147, ACM, 2017.
- [47] J. Buckman, A. Roy, C. Raffel, and I. Goodfellow, “Thermometer encoding: One hot way to resist adversarial examples,” in *International Conference on Learning Representations*, 2018.
- [48] C. Guo, M. Rana, M. Cisse, and L. van der Maaten, “Countering adversarial images using input transformations,” in *International Conference on Learning Representations*, 2018.
- [49] G. S. Dhillon, K. Azizzadenesheli, J. D. Bernstein, J. Kossaifi, A. Khanna, Z. C. Lipton, and A. Anandkumar, “Stochastic activation pruning for robust adversarial defense,” in *International Conference on Learning Representations*, 2018.
- [50] C. Xie, J. Wang, Z. Zhang, Z. Ren, and A. Yuille, “Mitigating adversarial effects through randomization,” in *International Conference on Learning Representations*, 2018.
- [51] Y. Song, T. Kim, S. Nowozin, S. Ermon, and N. Kushman, “Pixeldefend: Leveraging generative models to understand and defend against adversarial examples,” in *International Conference on Learning Representations*, 2018.
- [52] P. Samangouei, M. Kabkab, and R. Chellappa, “Defense-GAN: Protecting classifiers against adversarial attacks using generative models,” in *International Conference on Learning Representations*, 2018.
- [53] T. Na, J. H. Ko, and S. Mukhopadhyay, “Cascade adversarial machine learning regularized with a unified embedding,” in *International Conference on Learning Representations*, 2018.
- [54] N. Carlini, A. Athalye, N. Papernot, W. Brendel, J. Rauber, D. Tsipras, I. J. Goodfellow, A. Madry, and A. Kurakin, “On evaluating adversarial robustness,” *CoRR*, vol. abs/1902.06705, 2019.
- [55] N. Papernot, F. Faghri, N. Carlini, I. Goodfellow, R. Feinman, A. Kurakin, C. Xie, Y. Sharma, T. Brown, A. Roy, A. Matyasko, V. Behzadan, K. Hambardzumyan, Z. Zhang, Y.-L. Juang, Z. Li, R. Sheatsley, A. Garg, J. Uesato, W. Gierke, Y. Dong, D. Berthelot, P. Hendricks, J. Rauber, and R. Long, “Technical report on the cleverhans v2.1.0 adversarial examples library,” *arXiv preprint arXiv:1610.00768*, 2018.

- [56] M.-I. Nicolae, M. Sinn, M. N. Tran, A. Rawat, M. Wistuba, V. Zantedeschi, N. Baracaldo, B. Chen, H. Ludwig, I. Molloy, and B. Edwards, “Adversarial robustness toolbox v0.5.0,” *CoRR*, vol. 1807.01069, 2018.
- [57] F. Yu, W. Xian, Y. Chen, F. Liu, M. Liao, V. Madhavan, and T. Darrell, “BDD100K: A Diverse Driving Video Database with Scalable Annotation Tooling,” *arXiv e-prints*, p. arXiv:1805.04687, May 2018.