

基于边界值不变量的对抗样本检测方法

严飞, 张铭伦, 张立强

(武汉大学国家网络安全学院空天信息安全与可信计算教育部重点实验室, 湖北 武汉 430072)

摘要: 目前, 深度学习成为计算机领域研究与应用最广泛的技术之一, 在图像识别、语音、自动驾驶、文本翻译等方面都取得良好的应用成果。但人们逐渐发现深度神经网络容易受到微小扰动图片的影响, 导致分类出现错误, 这类攻击手段被称为对抗样本。对抗样本的出现可能会给安全敏感的应用领域带来灾难性的后果。现有的防御手段大多需要对抗样本本身作为训练集, 这种对抗样本相关的防御手段是无法应对未知对抗样本攻击的。借鉴传统软件安全中的边界检查思想, 提出了一种基于边界值不变量的对抗样本检测防御方法, 该方法通过拟合分布来寻找深度神经网络中的不变量, 且训练集的选取与对抗样本无关。实验结果表明, 在 LeNet、vgg19 模型和 Mnist、Cifar10 数据集上, 与其他对抗检测方法相比, 提出的方法可有效检测目前的常见对抗样本攻击, 并且具有低误报率。

关键词: 深度神经网络; 边界检查; 不变量; 对抗样本检测

中图分类号: TP309.2

文献标识码: A

doi: 10.11959/j.issn.2096-109x.2020012

Adversarial examples detection method based on boundary values invariants

YAN Fei, ZHANG Minglun, ZHANG Liqiang

Key Laboratory of Aerospace Information Security and Trusted Computing, Ministry of Education, School of
Cyber Science and Engineering, Wuhan University, Wuhan 430072, China

Abstract: Nowadays, deep learning has become one of the most widely studied and applied technologies in the computer field. Deep neural networks(DNNs) have achieved greatly noticeable success in many applications such as image recognition, speech, self-driving and text translation. However, deep neural networks are vulnerable to adversarial examples that are generated by perturbing correctly classified inputs to cause DNN modes to misbehave. A boundary check method based on traditional programs by fitting the distribution to find the invariants in the deep neural network was proposed and it use the invariants to detect adversarial examples. The selection of training sets was irrelevant to adversarial examples. The experiment results show that proposed method can effectively detect the current adversarial example attacks on LeNet, vgg19 model, Mnist, Cifar10 dataset, and has a low false positive rate.

Key words: deep neuron network, boundary checking, invariant, adversarial examples detecting

收稿日期: 2019-09-11; 修回日期: 2020-02-02

通信作者: 张立强, zhanglq@whu.edu.cn

基金项目: 国家重点基础研究发展计划 (“973” 计划) 基金资助项目(No.2014CB340601); 国家自然科学基金资助项目 (No. 61272452)

Foundation Items: The National Basic Research Program of China (973 Program) (2014CB340601), The National Natural Science Foundation of China (No. 61272452)

论文引用格式: 严飞, 张铭伦, 张立强. 基于边界值不变量的对抗样本检测方法[J]. 网络与信息安全学报, 2020, 6(1): 38-45.

YAN F, ZHANG M L, ZHANG L Q. Adversarial examples detection method based on boundary values invariants [J]. Chinese Journal of Network and Information Security, 2020, 6(1): 38-45.

1 引言

深度神经网络(DNN)在许多应用中取得了非常明显的成功,如人脸识别^[1]、自动驾驶汽车^[2]、恶意软件分类^[3]和专用网络连接归属^[4]。但随着时间的推移,研究人员发现了DNN容易受到对抗样本的影响^[5],即攻击者可以干扰良性输入,加入扰动,使DNN出现分类错误的情况。目前对抗样本的攻击方法主要分为两种:基于梯度的方法和基于内容的方法。在基于梯度的方法中,攻击者将生成对抗样本视为优化问题,并进行梯度搜索以找到对抗样本^[6-10]。在基于内容的方法中,攻击者更追求现实世界中的可用性,通常为输入数据加入补丁,这些补丁与现实世界相符,如图像上的水印和由摄像机镜头上的污垢造成的黑点等^[11]。

目前对这类问题的解决思路主要包括两个方向:提升DNN系统自身稳健性和检测对抗样本。在提升DNN稳健性方面,比较典型的工作包括对抗训练^[7]和梯度遮掩^[12]。对抗训练主要是将对抗样本加入DNN的训练过程中,以此来增强DNN模型本身,这种技术对于已知攻击十分有效,对于未知攻击效果有限。梯度遮掩则是通过对于梯度进行遮掩,使攻击者难以利用梯度进行对抗样本生成。然而,攻击者已经开发了针对此类防御更高级的攻击手段。文献[13-16]不试图加强模型本身,而是在操作过程中检测对抗样本。例如, Ma等^[15]提出使用基于局部内在维度的异常检测度量来检测对抗样本; Xu^[16]等使用精心构造的滤波器来检查原始图像与转换后图像的预测的不一致性; MagNet^[17]和 HGD^[18]则通过训练编码器和解码器来去除对抗样本的附加噪声。

调查发现目前多数针对对抗样本的防御,往往考虑的是正常图片与对抗样本之间基于图片像素之间的差异,通过这些差异进行对抗样本的检测防御。同时,大多数方法都需要对抗样本作为训练数据,由于新式对抗样本攻击方法层出不穷,因此并不能有效防御未知对抗样本攻击手段。本文借鉴软件安全中的边界检查概念,仅仅通过训练数据进行检测器的训练,提出了基于边界值不变量的对抗样本检测方法(简称BVI),该方法通过拟合分布来寻找深度神经网络中的不变量,具备对抗样本攻击

手法无关的检测能力。实验表明,该方法与同类工作相比,不仅可以检测未知新型对抗样本攻击,还具有误报率低的优势。

2 对抗样本简介及相关工作

本节将对深度神经网络以及常见的对抗样本攻击和防御手段进行介绍。

2.1 神经网络

神经网络通常可以被表示为一个 $F: X \rightarrow Y$ 的函数映射, X 表示输入空间, Y 表示输出空间。神经网络包括多个连接层,连接层之间通过权重 w_F 相连接。DNN的训练过程就是确定 w_F 值的过程。数据工程通过提供有监督的学习,提供大量 (x_i, y_i) 输入输出对,并且定义一个表示 $F(x)$ 的预期结果与对应真实标签 y^* 之间差异的损失函数或者代价函数 $J(F(x), y^*)$ 。训练过程是为了最小化损失函数,通过反向传播^[19]技巧不断更新参数。训练过程被训练速率等超参数控制。本文中由于模型已经给定,因此超参数是固定的。在测试阶段,未知输入组 X_t ,对于其中每一个输入 $x_t \in X_t$,分类模型对输入的分类标签表示为 $C(x_t) = \arg\max_i F_i(x_t)$, $F_i(x_t)$ 表示 x_t 被分类为 i 类别的概率。当 $C(x_t)$ 等于 x_t 的真实标签 $C^*(x_t)$ 时,分类结果正确。

本文集中于 m -class分类模型。对于 m -class分类模型,模型输出是一个 m 维的vector,每一维表示输入被分类为该类别的概率。定义一个神经网络,如式(1)所示。

$$y = F(x) = \text{softmax}Z(x) \quad (1)$$

$Z(x)$ 表示logits、softmax层的输入,倒数第二层的输出。softmax层负责归一化输入,将输入归一化为概率。最终的输出类别为概率最高的类别。

2.2 对抗样本分类介绍

DNN受到对抗样本的影响。对抗样本是在被正常分类的输入上加入一些扰动,使分类器对其分类错误。目前的工作主要有两种不同类别的生成对抗样本的方法:基于梯度的方法与基于内容的方法。

(1) 基于梯度的方法

基于梯度的方法是给定一个正确分类为 $C(x)$ 的输入 $x \in X$,如果满足下面的条件:

$$x' \in X \wedge C(x') \neq C(x) \wedge (x, x') \leq \epsilon \quad (2)$$

这种类型的样本被称作 **untargeted** 对抗样本。其中, ϵ 代表对抗强度阈值, 限制对抗样本对于原图片的修改程度。 x' 表示对抗输入, 与 x 很相似, 但分类器却给出了与 x 不相同的分类结果。另一类对抗样本被称作 **targeted** 对抗样本。这类攻击中, 对抗样本 x' 可以被指定分类成任意类别。给定一个正确分类的输入 $x \in X(C(x) = C^*(x))$ 和一个指定类别 $t \neq C(x)$, **targeted** 对抗样本 x' 满足:

$$x' \in X \wedge C(x') \neq C(x) \wedge C(x') = t \quad (3)$$

因此, 生成对抗样本可以被看作优化问题, **targeted** 攻击如式(4)所示, **untargeted** 攻击如式(5)所示。

$$\min(x, x') s.t. C(x') = t \wedge x' \in X \quad (4)$$

$$\min \Delta(x, x') s.t. C(x') \neq C(x) \wedge x' \in X \quad (5)$$

目前对抗样本通常使用 L_p 范数作为 $\Delta(\cdot)$ 函数。 $\Delta(\cdot)$ 代表距离度量函数, 度量两个输入之间的相似程度。总的来说: L_0 限制对抗样本修改的像素数目, 但不限制对于像素修改的大小, 因此, 这些对抗样本修改的扰动区域较小, 但扰动的幅度较大; L_∞ 限制每个像素修改的大小, 但不限制修改的数量; L_2 通过同时限制修改数量与大小, 达成一个平衡。

(2) 基于内容的攻击

基于内容的攻击是利用输入内容并在原始输入上提供受约束的扰动来模拟真实世界场景。如由于相机镜头上的污垢, 图像可能会有黑点, 攻击者可以有意地向图像添加多个黑点来模拟这种情况。在这样的假设下, 攻击者不需要计算 $\Delta(\cdot)$ 距离值。因此, 许多这样的对抗样本是人类可识别的, 但更贴近现实世界情况。类似的攻击如文献[5,11,20-22]。

2.3 常见对抗样本攻击手段

本节将讨论针对 DNN 模型的 4 种现有的代表性攻击, 包括基于梯度的攻击和基于内容的攻击。虽然一般情况下机器学习模型存在对抗性攻击^[23], 但本文主要关注 DNN 模型上的对抗性样本。

(1) 快速梯度符号方法(FGSM, fast gradient sign method)

Goodfellow 等^[7]提出了使用快速梯度符号方法来生成对抗样本。DNN 通常使用分段线性激活

函数(如 ReLU), 这些激活函数易于优化, 方便训练 DNN。对于这样的模型, 输入的任何变化都可以传递到后面的隐藏层和输出层。因此, 输入上的错误会累积并导致错误分类。FGSM 攻击基于上述条件, 并假设所有维度具有相同的攻击强度, 因此 FGSM 是 L_∞ **untargeted** 攻击。形式上, 使用式(6)生成对抗样本。

$$x' = x + \epsilon \text{sign}(\nabla_x J(F(x))) \quad (6)$$

其中, ∇ 代表梯度, $J(\cdot)$ 是损失函数。FGSM 沿着梯度方向以固定比例同时改变所有像素。它可以快速生成对抗样本, 因为它只需要计算反向传播的梯度符号。FGSM 对许多像素进行了小的改动。

(2) DeepFool

Moosavi 等^[24]设计了 DeepFool 攻击, 首先假设模型是完全线性的。在这个假设下, 有一个多面体可以分离各个类。生成对抗样本成为一个更简单的问题, 因为类的边界是线性平面, 整个区域(对于这个类)是一个多面体。DeepFool 攻击通过使用 L_2 距离搜索特定区域内具有最小扰动的对抗样本, 采用几何中的方法来指导搜索对抗样本的过程。对于模型不完全线性的情况, 通过利用迭代线性化过程导出近似多面体, 并在找到真正的对抗样本时终止该过程。该攻击的变化是在原始对象附近。DeepFool 是 **untargeted** 攻击。

(3) 基于雅可比的显着性图攻击(JSMA, Jacobian-based saliency map attack)方法

Papernot 等^[10]提出了基于雅可比的显着性图攻击, JSMA 使用贪婪算法优化 L_0 距离。在每次迭代中, JSMA 选择修改最重要的像素来生成对抗样本。为了测量像素的重要程度, JSMA 引入了显着性图的概念, 首先, 该概率图使用雅可比矩阵计算修改单个像素对每个输出类标签的影响程度。然后, 攻击会修改最重要的像素。作为 L_0 攻击, JSMA 修改非常少的像素, 但扰动比 L_∞ 攻击(如 FGSM)要大。同时, JSMA 计算成本很高(由于需要计算雅可比矩阵), 使它对于高维输入(如来自 ImageNet 数据集的图像)的攻击不实际。

(4) Carlini and Wagner attacks(C&W)

Carlini 和 Wagner 提出了 3 种不同的基于梯度的攻击, 分别使用不同的 L_p 范数, 即 L_2 、 L_∞ 和 L_0 , 将它们分别称为 CW_2 、 CW_∞ 和 CW_0 攻击。在

CW_2 攻击中, Wagner 使用 L_2 范数作为距离测量来解决(目标)优化问题。但是, C&W 攻击运用一些新的设计手段, 使其非常有效。首先, 它们使用 logits $Z(\cdot)$ 代替损失函数中的最终预测 $F(\cdot)$, 这种设计对于绕过蒸馏方法的防御至关重要^[25]。其次, 它们引入了变量 α (也被称为最优常数) 来控制对抗样本的置信度, 使它们可以在预测值和距离值之间获得更好的权衡。另一个关键设计是将目标变量映射到 arctanh 空间, 然后解决优化问题。这种设计使攻击能够使用更新的优化算法, 如 Adam 算法^[26]来优化目标。这些技术使 CW_2 攻击非常有效地生成高质量的对抗样本。与其他 L_2 攻击相比, CW_2 扰乱像素数量更少, 且扰动值也更小。

2.4 现有的对抗样本防御与检测手段

目前, 主流防御手段主要分为模型本身防御与外置的检测手段。

对于现有的防御手段, 防御技术通过强化 DNN 模型以防止对抗样本攻击^[27-29]。Papernot 等^[30]全面研究了现有的防御机制, 并将其分为两大类: 对抗性训练和梯度掩蔽。

Goodfellow 等^[7]介绍了对抗性训练的思想。对抗训练扩展了训练数据集, 包括带有正确标签的对抗样本。但是, 对抗训练需要事先了解所有可能的攻击, 因此无法处理新的攻击方法。

梯度遮掩^[12]的基本思想是通过训练具有微小(如接近 0)梯度的模型来增强训练过程, 使模型对输入中的微小变化不敏感。然而, 实验表明梯度遮掩可能导致正常输入样本的准确性降低。Papernot 等^[31]引入防御性蒸馏来强化 DNN 模型。它通过平滑现有 DNN 的预测结果来训练模型。通过将 softmax 函数平滑, 来隐藏来自攻击者的梯度信息。然而, 此类模型可以通过高级攻击^[9-10, 32]来打破。Athalye 等^[33]研究表明, 通过梯度近似可以避免混淆梯度。Papernot 等^[30]得出结论, 由于对抗样本的可传递性, 在训练中控制梯度信息对防御对抗性攻击的影响有限, 传递性意味着从某模型生成的对抗样本可用于攻击不同的模型。

对于现有检测方面, 对抗样本检测用于确定特定输入样本是否是对抗样本。许多之前的研究^[14, 34-37]已经建立了检测系统。现有的检测手段可分为两

大类: 基于度量的方法与基于预测不一致的方法。

(1) 基于度量的方法

有关研究人员已经提出对输入(和激活值)进行统计测量以检测对抗样本。FEINMAN 等^[13]提出使用核密度估计(KD)和贝叶斯不确定性(BU)来识别对抗子空间, 以分离正常输入和对抗样本。Carlini 等^[38]表明这种方法可以被绕过, 但也认为这种方法是希望的方向。受到来自异常检测社区的想法的启发, Ma 等^[15]最近提出使用被称为局部内在维度(LID)的测量。对于给定的样本输入, 该方法估计 LID 值, 该 LID 值通过计算样本的距离分布和各个层邻居的数量来评估样本周围区域的空间填充能力。该文献表明, 对抗样本往往具有较大的 LID 值。他们的研究结果表明, LID 在对抗样本检测中优于 BU 和 KD, 目前代表了这类型探测器的最精确水平。这些技术的关键挑战是如何定义高质量的统计指标, 使该指标可以清楚地分辨正常样本和对抗样本之间的差异。Lu 等^[39]研究表明 LID 对攻击部署的置信参数很敏感, 并且容易受到对抗样本传递性的影响。

(2) 基于预测不一致的方法

许多其他工作^[6, 40-41]都基于预测不一致的方法, 即对抗样本具有扰动, 利用其他检测手段与原输出进行比较, 一致为正常样本, 不一致则为对抗样本。TAO 等^[41]提出通过测量原始神经网络和用人类可感知属性增强的神经网络之间的不一致来检测对抗性的例子, 然而, 这种方法需要人类定义的检测属性。最先进的检测技术 Feature Squeezing^[16]可以实现对各种攻击非常高的检测率, 文献指出, DNN 具有不必要的大输入特征空间, 这允许对手产生对抗样本。因此, 他们提出使用挤压技术(即减小图像的颜色深度和平滑图像), 以产生一些压缩图像。特征挤压限制了对手可用的自由度, 然后 DNN 模型获取所有压缩图像和原始种子图像, 并单独进行预测。通过测量原始种子输入的预测矢量和每个压缩图像之间的距离来检测对抗性样本。如果其中一个距离超过阈值, 则种子输入被视为恶意。然而, 根据文献[16], 该技术在 FGSM、BIM 和一些基于内容的攻击方法在 CIFAR 和 ImageNet 上表现不佳。这是因为其性能高度依赖于设计的挤压器的质量。

3 对抗样本检测器分析与实现

3.1 威胁模型

假设对手知道原始分类器的所有内容（包括训练的权重），以便攻击者可以构建强攻击，如 CW 攻击。同时检测器不知道用于生成对抗样本的方法。根据攻击者对于检测器的了解程度，可以分成多种场景。最弱的攻击场景是对手对探测器一无所知，在这种情况下，攻击者仅了解原始分类器。最强大的攻击场景是对手完全了解使用的探测器。

由于探测器本身也是一个分类器，这使它容易受到对抗性样本的影响^[42]。不过，此限制并非特定于本文的技术，因为其他现有检测技术也遇到相同的问题。在如此强大的威胁模型下，与其他技术相比，本文的技术具有更好的弹性。如第 3.2 节所述，本文采用的检测器基于多个子特征。在检测器的训练期间，在激活的神经元上分别拟合分布以产生多个分布函数，这能够灵活地生成多个检测器。在运行时，可以使用不同的检测器（或它们的组合）来检测对抗样本，这大大提高了产生对抗样本的难度。但是，完全防止对抗样本对于所有的 DNN 是几乎不可能的。本文的目标是建立一个通用而实用的解决方案来大幅提高攻击者的攻击门槛。

3.2 基于边界值检查的不变量

本文做以下规定：在神经网络中，用 $N = \{n_1, n_2, \dots\}$ 表示 DNN 的一组神经元，用 $\phi(x, n)$ 代表给定输入 x ，返回神经元 n 的输出；假设 DNN 有 l 层，用 L_i 表示第 i 层神经元的集合， $1 \leq i \leq l$ ；给定 Tr 为训练集输入。

边界检查在程序设计中是指在使用某一个变量前，检查该变量是否处在一个特定范围之内。最常见的是数组的下标检查，防止下标超出数组范围而覆盖其他数据。若边界检查未能有效发现错误，最常见的结果是程序出现异常并终止运行，但也可能出现其他现象。将边界检查应用到神经网络中，每一个神经元经过训练集的训练后，都可以得到一组该神经元关于训练集的输出集合，如式(7)所示。

$$S_n = \left\{ \bigcup_{x \in Tr} \phi(x, n) \right\} \quad (7)$$

S_n 代表了神经元 n 在训练集 Tr 中的取值范

围。但是，神经网络与传统程序设计不同，神经网络具有泛化性能，即不确定性。传统程序设计可以精确地确定边界值的大小，而神经网络如果仅仅套用传统程序设计的方法，会使神经网络失去其泛化性，将决策边界构成一个一个样本，而非一个平面或曲面。因此，对于 S_n ，采用分布拟合的手段，即通过 S_n 拟合成一个分布 f ，判断输入的 $\phi(temp, n)$ 值是否符合该分布 f ，即判断输入对于该分布的置信度，以此来充当神经元 n 的边界检查。即在每一层，采用神经元作为粒度，每个神经元对应一个由训练集得出的分布，当输入进入时，就可以算出该输入对于每一个神经元的置信度，即该输入是否符合神经元的边界。

同时，由于 DNN 拥有多层结构与大量神经元，因此通过计算每个神经元的边界值开销过大。基于对抗样本是对于原图片扰动的假设，在神经网络开始的层扰动带来的影响更大。后续的层扰动经过传递逐渐减弱，每一层点的分布开始和正常样本趋于一致。

3.3 检测器的实现

对抗样本本质是在正常图片中加入扰动，该扰动造成传递过程中输出值的改变，从而最终影响分类结果。在神经网络中，低层往往提取的是最基本的特征。对抗样本为了造成分类错误，对于低层扰动更大，更可能造成其分布与正常样本不同，因此可以通过拟合前面几层的数据分布作为训练监测模型的特征。

本文仅使用第一个池化层后的输出作为训练输入，而这已经基本达到了目前最好的检测效果，同时具有最低的误报率。假设该层是第 L 层， N 代表 L 层的所有神经元， $N = \{n_1, n_2, n_3, \dots, n_m\}$ 。对于输入的训练集 Tr ，得到 Tr 在每一个神经元上的输出。对每一个神经元的输出集合使用 `scipy` 自带的 80 种分布进行拟合，去掉最终拟合 `error` 大于 10 的分布，去掉输出基本全为 0 的神经元，最后得到 $F = \{f_1, f_2, f_3, \dots, f_k | k \leq m\}$ ， f_i 代表神经元 i 关于训练集 Tr 的分布。 F 为 L 层所有合格神经元的关于训练集 Tr 拟合的分布集合。

对于输入 x ，需要了解 x 是否符合 F 中的所有分布。由于是对单个输入 x ，并且目标 F 中的分布不一定是正态分布，所以采用 K-S (Kolmogorov-

Smirnov) 检验。K-S 检验是以两位苏联数学家 Kolmogorov 和 Smirnov 的名字命名的, 它是一个拟合优度检验, 研究样本观察值的分布和设定的理论分布是否吻合, 通过对两个分布差异的分析确定是否有理由认为样本的观察结果来自所假定的理论分布总体。因此, 对于输入 x , 计算 x 关于所有分布 f_i 的 K-S 检验, 得到的 pvalue 值为输入 x 关于分布 f_i 的置信度, 也就是关于神经元 i 的边界检查置信度。该置信度值为后面训练检测器所使用的特征。

本文检测器的一大特点是仅使用良性样本, 也就是只使用训练样本作为检测器的训练集, 这使该检测器不同于那些基于对抗训练的防御方法以及需要对抗样本作为训练集的检测器。因此, 本文的检测器是与对抗样本攻击无关的, 不需要考虑具体的对抗样本攻击手段, 这样对于未知的对抗样本也具有很好的防御效果。

本文将不含对抗样本的训练任务建模为一个单分类任务。在单分类任务中, 大部分训练样本是正例, 在测试时包含所有种类输入 (如不同环境下的对抗样本攻击)。单分类任务在文献[37,43]中经过详细讨论。尽管单分类不如正负分类精确, 但在本文的场景中非常适合, 因为使用了多组不变量联合作为决策依据, 因此可以有效缓和单分类的不精确问题。

本文使用 One-class SVM^[37] 算法作为分类器。One-class SVM 的基本思想是在不同类别之间假设一个决策边界, 通过训练集来学习决策边界的参数。对于 OSVM 来说, 最常用到的是 RBF 核, 在本文的条件中, 大部分输入是不合法的 (如大多数随机图片是不真实的), 合法输入都聚集在一个小空间中, 因此使用 RBF 可以得到很好的效果。使用之前得到的输入关于 F 中分布的 K-S 检验置信度作为特征进行 OSVM 的训练, 核为 RBF 核, 由此得到本文的检测器。其正例输出为 1, 则为正常样本; 输出 -1, 则为对抗样本。

4 实验测试与分析

4.1 实验设置

1) 数据集。本文采用两个流行的图像数据集 MNIST^[44], CIFAR-10^[45]。MNIST 是用于手写数字识别的灰度数据图像。CIFAR-10 是用于对象识

别的彩色图像数据集。本文选择这两个数据集是因为它们是此任务中使用最广泛的数据集, 并且目前大部分攻击都是针对它们执行的。其中 MNIST 数据集具有 60 000 张训练图片, 10 000 张测试图片。CIFAR-10 具有 50 000 张彩色训练图片, 10 000 张测试图片。MNIST 和 CIFAR-10 均为多分类问题。

2) 对抗样本攻击手段。本文评估了第 2 节描述的 4 种攻击的检测。对于 FGSM、JSMA 攻击, 使用 Foolbox 库^[46]来生成对抗样本, 而对于其他攻击, 参考了文献[24-25]。对于这 4 种攻击, FGSM 采用 *untargeted* 攻击方式, 因为相比 *targeted* 方式, *untargeted* 修改的幅度更小, 更难被检测出来。对于 CW 和 JSMA 攻击, 采用两种攻击配置: 下一类别攻击 (记作 Next), 即使图片错误分类成它的下一个类别 (如将 2 分类成 3) 最小类别攻击 (记作 LL), 将图片错误分类成其差异最大的类别 (如将 1 分类成 8)。

3) 模型。本文在两种流行的模型上评估提出的技术。对于 MNIST 数据集, 使用 LeNet 家族中的 LeNet-5 模型^[45]; 对于 CIFAR-10 数据集, 采用 VGG19 模型^[47]。

4) 比较。同样和顶尖的检测器进行对比。基于度量的防御方法选取 LID^[15]; 降噪的防御方法选取 MagNet^[17]; 基于预测不一致的防御方法, 选取 Feature Squeezing^[16]。

4.2 实验结果

本文方法对于对抗样本检测的结果如表 1 所示。从结果可以看出, 除了 JSMA 攻击, 其余都基本全部检测出来。由于 JSMA 是 L_0 攻击, 修改的像素点较少, 因此边界检查即分布拟合的违反较小, 所以准确率相较于其他方法较低。

本文方法与其他检测器的检测结果的比较如表 2 所示。每一行代表一个检测器方法, 每一列代表一种攻击手段。从表 2 可以看出, 本文的检测手段基本达到了目前最高水准的准确率。同时, 具有目前最低的误报率, 这是基于测试集分布与训练集基本一样得到的, 这意味着本文的检测器模型最大程度上避免将良性样本分类成对抗样本, 因此本文的分类器分类结果最为可信。

表 1 本文方法对于对抗检测的结果

Table 1 Detection result of various adversarial examples attacks in this paper

数据集	模型	误报率	FGSM	DeepFool	JSMA Next	JSMA LL	CW2 Next	CW2 LL
MNIST	LeNet-5	0.0%	98.6%	98.8%	85.4%	85.2%	98.6%	98.6%
CIFAR-10	Vgg19	0.6%	97.6%	97.8%	82.2%	83.2%	97.2%	97.2%

表 2 本文方法与其他检测器的检测结果的比较

Table 2 Comparison of the detection results of the proposed method and other detectors

检测器	数据集	误报率	FGSM	DeepFool	JSMA Next	JSMA LL	CW2 Next	CW2 LL
本文 BVI	MNIST	0.0%	98.6%	98.8%	85.4%	85.2%	98.6%	98.6%
MagNet	MNIST	5.1%	100.0%	91.0%	84.0%	84.0%	85.0%	87.0%
LID	MNIST	4.4%	97.0%	92.0%	96.0%	96.0%	92.0%	91.0%
FS	MNIST	4.0%	98.6%	—	100.0%	100.0%	100.0%	100.0%
本文 BVI	CIFAR-10	0.6%	97.6%	97.8%	82.2%	83.2%	97.2%	97.2%
MagNet	CIFAR-10	6.4%	100.0%	87.0%	72.0%	74.0%	89.0%	91.0%
LID	CIFAR-10	5.6%	94.0%	84.0%	92.0%	92.0%	86.0%	88.0%
FS	CIFAR-10	4.9%	21.0%	77.0%	84.0%	89.0%	100.0%	100.0%

5 结束语

本文基于传统软件安全领域的边界检查，通过拟合输出的分布，构建了一种不依赖于攻击方式的对抗样本防御手段。本文的防御手段只使用了训练集和图像本身的信息，而不需要任何特定对抗样本攻击知识。实验结果表明，采用边界不变量的检测手段达到了目前最好的检测效果级别，同时拥有极低的误报率，可以更有效地进行对抗样本检测。

参考文献:

- [1] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. Imagenet classification with deep convolutional neural networks[C]//Advances in Neural Information Processing Systems. 2012: 1097-1105.
- [2] BOJARSKI M, Del TESTA D, DWORAKOWSKI D, et al. End to end learning for self-driving cars[J]. arXiv preprint arXiv:1604.07316, 2016.
- [3] DAHL G E, STOKES J W, DENG L, et al. Large-scale malware classification using random projections and neural networks[C]//2013 IEEE International Conference on Acoustics, Speech and Signal Processing. 2013: 3422-3426.
- [4] MIRSKY Y, DOITSHMAN T, ELOVICI Y, et al. Kitsune: an ensemble of autoencoders for online network intrusion detection[J]. arXiv preprint arXiv:1802.09089, 2018.
- [5] SZEGEDY C, ZAREMBA W, SUTSKEVER I, et al. Intriguing properties of neural networks[J]. arXiv preprint arXiv:1312.6199, 2013.
- [6] DHILLON G S, AZIZZADENESHELI K, LIPTON Z C, et al. Stochastic activation pruning for robust adversarial defense[J]. arXiv preprint arXiv:1803.01442, 2018.
- [7] GOODFELLOW I J, SHLENS J, SZEGEDY C. Explaining and harnessing adversarial examples[J]. arXiv: preprint arXiv: 1412.6572, 2014.
- [8] KURAKIN A, GOODFELLOW I, BENGIO S. Adversarial examples in the physical world[J]. arXiv preprint arXiv:1607.02533, 2016.
- [9] CARLINI N, WAGNER D. Defensive distillation is not robust to adversarial examples[J]. arXiv preprint arXiv:1607.04311, 2016.
- [10] 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.
- [11] LIU Y, MA S, AAFER Y, et al. Trojaning attack on neural networks[C]// Network and Distributed System Security Symposium. 2018.
- [12] GU S, RIGAZIO L. Towards deep neural network architectures robust to adversarial examples[J]. arXiv preprint arXiv:1412.5068, 2014.
- [13] FEINMAN R, CURTIN R R, SHINTRE S, et al. Detecting adversarial samples from artifacts[J]. arXiv preprint arXiv:1703.00410, 2017.
- [14] GROSSE K, MANOHARAN P, PAPERNOT N, et al. On the (statistical) detection of adversarial examples[J]. arXiv preprint arXiv:1702.06280, 2017.
- [15] MA X, LI B, WANG Y, et al. Characterizing adversarial subspaces using local intrinsic dimensionality[J]. arXiv preprint arXiv: 1801.02613, 2018.
- [16] XU W, EVANS D, QI Y. Feature squeezing: detecting adversarial examples in deep neural networks[J]. arXiv preprint arXiv: 1704.01155, 2017.
- [17] MENG D, CHEN H. Magnet: a two-pronged defense against adversarial examples[C]//The 2017 ACM SIGSAC Conference on Computer and Communications Security. 2017: 135-147.

- [18] LIAO F, LIANG M, DONG Y, et al. Defense against adversarial attacks using high-level representation guided denoiser[C]//The IEEE Conference on Computer Vision and Pattern Recognition. 2018: 1778-1787.
- [19] LECUN Y, BOTTOU L, BENGIO Y, et al. Gradient-based learning applied to document recognition[J]. Proceedings of the IEEE, 1998, 86(11): 2278-2324.
- [20] BROWN T B, MANÉ D, ROY A, et al. Adversarial patch[J]. arXiv preprint arXiv:1712.09665, 2017.
- [21] EYKHOLT K, EVTIMOV I, FERNANDES E, et al. Robust physical-world attacks on deep learning models[J]. arXiv preprint arXiv:1707.08945, 2017.
- [22] PEI K, CAO Y, YANG J, et al. Deepxplore: automated whitebox testing of deep learning systems[C]//The 26th Symposium on Operating Systems Principles. 2017: 1-18.
- [23] BIGGIO B, ROLI F. Wild patterns: ten years after the rise of adversarial machine learning[J]. Pattern Recognition, 2018, 84: 317-331.
- [24] MOOSAVI-DEZFOOLI S M, FAWZI A, FROSSARD P. DeepFool: a simple and accurate method to fool deep neural networks[C]//IEEE Conference on Computer Vision and Pattern Recognition. 2016: 2574-2582.
- [25] CARLINI N, WAGNER D. Towards evaluating the robustness of neural networks[C]//2017 IEEE Symposium on Security and Privacy (SP). 2017: 39-57.
- [26] KINGMA D P, BA J. Adam: a method for stochastic optimization[J]. arXiv preprint arXiv:1412.6980, 2014.
- [27] ROUHANI B D, SAMRAGH M, JAVAHERIPIM, et al. Deepfense: online accelerated defense against adversarial deep learning[C]//IEEE/ACM International Conference on Computer-Aided Design (ICCAD). 2018: 1-8.
- [28] SONG Y, KIM T, NOWOZIN S, et al. Pixeldefend: leveraging generative models to understand and defend against adversarial examples[J]. arXiv preprint arXiv:1710.10766, 2017.
- [29] XIE C, WANG J, ZHANG Z, et al. Mitigating adversarial effects through randomization[J]. arXiv preprint arXiv:1711.01991, 2017.
- [30] PAPERNOT N, MCDANIEL P, SINHA A, et al. Towards the science of security and privacy in machine learning[J]. arXiv preprint arXiv:1611.03814, 2016.
- [31] 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.
- [32] PAPERNOT N, MCDANIEL P, GOODFELLOW I, et al. Practical black-box attacks against machine learning[C]//ACM on Asia Conference on Computer and Communications Security. 2017: 506-519.
- [33] ATHALYE A, CARLINI N, WAGNER D. Obfuscated gradients give a false sense of security: circumventing defenses to adversarial examples[J]. arXiv preprint arXiv:1802.00420, 2018.
- [34] BHAGOJI A N, CULLINA D, MITTAL P. Dimensionality reduction as a defense against evasion attacks on machine learning classifiers[J]. arXiv preprint arXiv:1704.02654, 2017.
- [35] GONG Z, WANG W, KU W S. Adversarial and clean data are not twins[J]. arXiv preprint arXiv:1704.04960, 2017.
- [36] HENDRYCKS D, GIMPEL K. Early methods for detecting adversarial images[J]. arXiv preprint arXiv:1608.00530, 2016.
- [37] TAX D M J, DUIN R P W. Support vector domain description[J]. Pattern Recognition Letters, 1999, 20(11-13): 1191-1199.
- [38] CARLINI N, WAGNER D. Adversarial examples are not easily detected: bypassing ten detection methods[C]//The 10th ACM Workshop on Artificial Intelligence and Security. 2017: 3-14.
- [39] LU P H, CHEN P Y, YU C M. On the limitation of local intrinsic dimensionality for characterizing the subspaces of adversarial examples[J]. arXiv preprint arXiv:1803.09638, 2018.
- [40] GUO C, RANA M, CISSE M, et al. Countering adversarial images using input transformations[J]. arXiv preprint arXiv:1711.00117, 2017.
- [41] TAO G, MA S, LIU Y, et al. Attacks meet interpretability: attribute-steered detection of adversarial samples[C]//Advances in Neural Information Processing Systems. 2018: 7717-7728.
- [42] GILMER J, METZ L, FAGHRI F, et al. Adversarial spheres[J]. arXiv preprint arXiv:1801.02774, 2018.
- [43] PERERA P, PATEL V M. Learning deep features for one-class classification[J]. IEEE Transactions on Image Processing, 2019, 28(11): 5450-5463.
- [44] TAX D M J, DUIN R P W. Data domain description using support vectors[C]//ESANN. 1999, 99: 251-256.
- [45] KRIZHEVSKY A, HINTON G. Learning multiple layers of features from tiny images[R]. Technical Report, University of Toronto, 2009.
- [46] RAUBER J, BRENDEN W, BETHGE M. Foolbox: a Python toolbox to benchmark the robustness of machine learning models[J]. arXiv preprint arXiv:1707.04131, 2017.
- [47] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[J]. arXiv preprint arXiv:1409.1556, 2014.

[作者简介]



严飞 (1980—)，男，湖北武汉人，武汉大学副教授、硕士生导师，主要研究方向为系统安全、可信计算、系统安全验证与形式化分析和移动目标防御。



张铭伦 (1995—)，男，江苏连云港人，武汉大学硕士生，主要研究方向为人工智能系统本身的安全防护问题。



张立强 (1979—)，男，黑龙江哈尔滨人，武汉大学讲师，主要研究方向为系统安全、可信计算和安全测评。