

# 一种检测 C&W 对抗样本图像的盲取证算法

邓 康<sup>1</sup>, 罗盛海<sup>1</sup>, 彭安杰<sup>1,2</sup>, 曾 辉<sup>1,2</sup>, 黄晓芳<sup>1</sup>

<sup>1</sup>西南科技大学计算机科学与技术学院 绵阳 中国 621010

<sup>2</sup>中山大学广东省信息安全重点实验室 广州 中国 510275

**摘要** 对抗样本图像能欺骗深度学习网络, 亟待对抗样本防御机制以增强深度学习模型的安全性。C&W 攻击是目前较热门的一种白盒攻击算法, 它产生的对抗样本具有图像质量高、可转移、攻击性强、难防御等特点。本文以 C&W 攻击生成的对抗样本为研究对象, 采用数字图像取证的思路, 力图实现 C&W 对抗样本的检测, 拒绝对抗样本输入深度学习网络。基于对抗样本中的对抗扰动易被破坏的假设, 我们设计了基于 FFDNet 滤波器的检测算法。具体来说, FFDNet 是一种基于深度卷积网络 CNN 的平滑滤波器, 它能破坏对抗扰动, 导致深度学习模型对对抗样本滤波前后的输出不一致。我们判断输出不一致的待测图像为 C&W 对抗样本。我们在 ImageNet-1000 图像库上针对经典的 ResNet 深度网络生成了 6 种 C&W 对抗样本。实验结果表明本文方法能较好地检测 C&W 对抗样本。相较于已有工作, 本文方法不仅极大地降低了虚警率, 而且提升了 C&W 对抗样本的检测准确率。

**关键词** 深度学习; 对抗样本; 数字图像取证; 图像滤波

中图分类号 TP309.2 DOI 号 10.19363/J.cnki.cn10-1380/tn.2020.11.01

## Blind forensics of adversarial images generated by C&W algorithm

DENG Kang<sup>1</sup>, LUO Shenghai<sup>1</sup>, PENG Anjie<sup>1,2</sup>, ZENG Hui<sup>1,2</sup>, HUANG Xiaofang<sup>1</sup>

<sup>1</sup> School of Computer Science and Technology, Southwest University of Science and Technology, Mianyang 621010, China

<sup>2</sup> Guangdong Key Laboratory of Information Security Technology, Sun Yat-Sen University, Guangzhou 510275, China

**Abstract** Adversarial images which can fool Deep neural networks have attracted researchers to focus on how to harden DNNs against adversarial attacks. Among typical attack algorithms, the C&W attack is one of the strongest attacks, which ensures the attack success rates yet causes less adversarial perturbations on the original image, and is taken as a benchmark in defense attempts. In this paper, we employ the blind forensic methodology to detect C&W adversarial images, which aims to avoid adversarial inputs for deep neural networks. Supposing that the adversarial perturbations are easily damaged by some image processing operations, we proposed a detecting method by using the fast and flexible de-noising convolution neural network called FFDNet. Specially, we compare the model's prediction on the test image and its filtered version. If the original and filtered inputs produce substantially different outputs from the model, the test image is likely to be adversarial. We employ ResNet as the targeted network, and generate 6 kinds of C&W adversarial images on ImageNet-1000 database. Experimental results show that the proposed method is effective in the detection of C&W adversarial images, and outperforms state-of-the-arts in terms of false positive rates and true positive rates.

**Key words** deep learning; adversarial images; digital image forensics; image filtering

## 1 引言

近年来, 深度学习(Deep learning)在计算机视觉的相关任务中, 如物体分类与识别<sup>[1]</sup>, 语义分割<sup>[2]</sup>, 取得了令人瞩目的成就。然而, Szegedy 等发现深度学习网络易受对抗样本(Adversarial examples)攻击<sup>[3]</sup>。

对抗样本由攻击者往原始图像添加精心设计且不易察觉的对抗扰动(Adversarial perturbation, 也叫对抗噪声)生成, 其目的在于使深度学习系统产生误判, 即错误地判断原始干净图像的类别。对抗样本攻击阻挠了深度学习在人工智能领域的潜在应用, 特别是对安全敏感的应用, 如无人驾驶<sup>[4]</sup>, 机器人<sup>[5]</sup>等。

通讯作者: 彭安杰, 博士, 副教授, Email: penganjie200012@163.com。

本课题得到国家自然科学基金(No.61702429), 四川省科技厅基金(No.19yyjc1656), 四川省教育厅基金(No.17ZB0450)资助。

收稿日期: 2019-12-31; 修改日期: 2020-04-03; 定稿日期: 2020-09-22

自 2014 年 Szegedy 等<sup>[6]</sup>提出对抗样本以来, 对抗样本攻击算法发展迅猛, 按攻击目标可分为有目标攻击(即干净样本被判为特定类别)与无目标攻击(即干净样本被判为任一错误类别), 按攻击者拥有的资源可分为黑盒攻击与白盒攻击。典型攻击算法包括快速梯度符号法<sup>[7]</sup>(Fast gradient sign method, 简称 FGSM), 随机梯度符号法<sup>[8]</sup>(Randomized FGSM), 映射梯度下降法<sup>[9]</sup>(Projected gradient descent), 贾可比显著图攻击法<sup>[10]</sup>(Jacobian-based Saliency Map Attack), DeepFool 方法<sup>[11]</sup>, Carlini 和 Wagner 提出的方法<sup>[12]</sup>(简称 C&W 攻击)。相较于其他攻击, C&W 攻击具有攻击强度高, 对抗扰动小进而对抗样本图像质量高, 难以被防御, 可转移性(即攻击未知深度网络的能力)等特点。本文以 C&W 对抗样本为检测对象, 采用数字图像取证的方式尝试实现 C&W 对抗样本的盲检测。

对抗样本检测是防御对抗攻击的一种重要途径。它通过事先检测输入图像是否为对抗样本, 避免对抗样本输入随后的深度学习模型。与对抗样本训练<sup>[13]</sup>, 蒸馏法<sup>[14]</sup>, 梯度遮罩<sup>[15]</sup>等相比, 对抗样本检测具有快速, 不修改网络模型结构等优点。Hendrycks<sup>[16]</sup>, Li 等<sup>[17]</sup>发现对抗样本与干净图像的主成分(PCA)间存在较大的差异, 设计了基于 PCA 的检测方法。Lu 等<sup>[18]</sup>分析了对抗样本与干净图像在深度模型最后一层激活函数 Relu 上的差异, 并基于这些差异提取特征, 使用高斯核支持向量机(RBF-SVM)检测对抗样本与干净图像。上述方法<sup>[16-18]</sup>能检测 FGSM 等基于梯度的攻击, 但不能有效地检测 C&W 攻击样本。Liang 等<sup>[19]</sup>将对抗扰动视为一类特殊噪声, 采用量化和空域滤波的方式处理图像, 通过分析滤波前后图像在深度网络中的输出差异检测对抗样本图像。一般而言, 对抗样本滤波前后的输出差异较大, 干净图像滤波前后的输出差异较小。类似地, Xu 等<sup>[20]</sup>使用减少色彩空间量化比特数, 中值滤波, 非局部滤波等方式, 提出了一种特征挤压(Feature squeezing)的检测算法。Guo 等<sup>[21]</sup>发现对抗样本与干净图像在不同深度模型上的预测输出有较大的差异(记为转移预测误差), 使用聚类的方式检测对抗样本。杨浚宇<sup>[22]</sup>使用自编码器把远离流形的对抗样本推回到流形周围, 重构输出, 然后通过重构前与重构后的预测差异检测对抗样本。上述方法能较好地检测 C&W 对抗样本, 但存在虚警率(False positive rate, 简称 FPR, 即将干净图像误认为对抗样本的比率)较高的缺点。

考虑到对抗样本检测与隐写分析的相似性, 部

分工作使用隐写分析策略检测对抗样本。对抗样本和隐写图像都可以看成通过向干净图像或原始图像添加不易被察觉的噪声生成。隐写分析使用二分类的方式检测原始图像(Cover)与隐写图像(Stego image)间的细微差异, 因而可以迁移隐写分析的方法检测对抗样本。Fan 等<sup>[23]</sup>使用集成检测器检测对抗样本。首先使用 SPAM(Subtractive pixel adjacency matrix)<sup>[24]</sup>特征区分对抗幅度较大的样本(主要为基于梯度攻击生成的样本)与对抗幅度较小的样本(包括 DeepFool、C&W 样本和干净样本), 然后使用添加高斯噪声的方法区分干净样本与 DeepFool、C&W 样本。类似于 Feature squeezing<sup>[20]</sup>, 该方法也是通过判断噪声添加前后版本的预测差异检测对抗样本。Liu 等<sup>[25]</sup>使用隐写分析经典特征 SPAM、SRM(Spatial rich model)<sup>[26]</sup>检测对抗样本。同时, 他们通过估计每个像素可能被攻击修改的概率生成修改概率图(Modification Probability Map), 在计算共生概率时赋予概率高的像素更大的权重, 形成了增强型特征 ESRM。隐写分析类方法集成检测器<sup>[23]</sup>与 ESRM<sup>[25]</sup>在检测对抗扰动大的梯度类对抗样本 FGSM、PGD 时取得了很好的检测效果(如对抗扰动为 8 比特时, 检测准确率大于 98.5%), 随着对抗扰动减小, 其检测准确率也下降。在检测对抗扰动更小的 C&W 对抗样本时, 隐写分析类方法的检测准确率下降的较多, 有待进一步提升。ESRM 方法的另一缺点是提取超高维 ESRM 特征(34671-D)比较费时。鉴于上述原因, 本文的出发点是针对 C&W 对抗样本设计一种快速, 有效的检测算法。需要强调的是, 考虑到设计一种能检测所有已知对抗样本类型的算法是困难的且新的未知对抗样本<sup>[30]</sup>不断演化, 本文检测方法只针对 C&W 对抗样本, 尽力提升 C&W 对抗样本的检测准确率, 以期与其他方法联合能检测更多类型的攻击。

本文使用数字图像取证的思路检测 C&W 对抗样本, 即仅依赖于干净图像与对抗样本间的差异, 无需事先嵌入额外的诸如水印、数字签名等认证信息, 适合于非受控情形下增强深度学习模型安全的场景。已有工作表明, 滤波降噪的方式<sup>[19-20]</sup>在对抗样本的检测中取得了较高的正阳性检测率(True positive rate, 简称 TPR, 即将对抗样本判为对抗样本的比率), 但虚警率控制的不够理想。究其原因在于, 论文[19-20]中的方法对原始干净样本的修改幅度较大。为此, 针对对抗扰动的特性, 我们采用基于噪声水平控制的深度卷积 CNN 模型<sup>[27]</sup>进行滤波。我们将待测图像及其滤波版本分别输入深度模型, 根据输出类别的差异判断待测图像类型。具体来说, 输出类

别不同的待测图像判为对抗样本, 输出类别相同的待测图像判为干净图像。ImageNet-1000 图像库上的实验结果表明, 对比于已有方法[19-20], 我们提出的方法明显地降低了虚警率 FPR, 同时提升了正阳性检测率 TPR。

## 2 背景知识

### 2.1 C&W 攻击方法简介

C&W 攻击<sup>[12]</sup>是一种较强的对抗样本攻击算法。与其他攻击算法一样, C&W 攻击通过最优化算式(1)求解对抗扰动  $\delta$ 。在公式(1)中, 对抗样本像素  $x_i + \delta_i$  的取值范围被限制在  $[0, 1]$  之间,  $F_\theta(x + \delta) = t$  表示深度模型  $F_\theta(\cdot)$  对对抗样本的预测输出类标为  $t$  ( $t$  不等于干净样本的类标  $y$ ),  $D(x + \delta, x)$  表示干净样本  $x$  与对抗样本  $x + \delta$  间的距离度量, C&W 攻击使用了常用的  $L_0, L_2, L_\infty$  等 3 种范数。

$$\begin{aligned} \arg \min_{\delta} D(x + \delta, x) \\ \text{s.t. } F_\theta(x + \delta) = t, \\ 0 \leq x_i + \delta_i \leq 1 \end{aligned} \quad (1)$$

显然地,  $F_\theta(\cdot)$  的非线性导致直接优化算式(1)及其困难。C&W 攻击首先使用目标函数  $f(x) = \max(\max\{Z(x)_j : j \neq t\} - Z(x)_t, -\kappa)$ , 当且仅当  $f(x) \leq 0$  时替换  $F_\theta(x + \delta) = t$ , 其中  $Z(x)_j$  表示深度网络倒数第2层(即 Softmax 前一层)的第  $j$  个输出(对于  $n$  分类而言,  $1 \leq j \leq n$ ), 参数  $\kappa$  控制攻击强度,  $\kappa$  越大攻击强度越大, 默认值为  $\kappa = 0$ 。然后使用拉格朗日法则将算式(1)转换为算式(2)所示的优化问题。

$$\begin{aligned} \arg \min_{\delta} D(x + \delta, x) + cf(x + \delta) \\ \text{s.t. } 0 \leq x_i + \delta_i \leq 1 \end{aligned} \quad (2)$$

在公式(2)中,  $c$  越大, 攻击成功率越高, 攻击耗时更长。C&W 攻击使用二分搜索寻找最优的参数  $c$ 。在  $L_2$  范数限制下, C&W 攻击使用多起始点梯度下降法搜索最佳的扰动。对于  $L_0$  范数限制, C&W 攻击迭代使用  $L_2$  攻击去寻找不重要的像素并固定不修改它们, 直到  $L_2$  攻击失败便停止迭代。  $L_\infty$  攻击也使用迭代法求解对抗扰动  $\delta$ 。

C&W 攻击成功率高, 能有效突破蒸馏法防御机制<sup>[14]</sup>, 且生成的对抗图像质量远优于 FGSM 等攻击。C&W 的缺点是攻击耗时, 对于 ImageNet 中尺寸为  $224 \times 224 \times 3$  的一副图像, Intel Xeon CPU 3.40GHz+

32G RAM+ 11G 1080Ti GPU 的配置下,  $L_0, L_2, L_\infty$  攻击分别耗时约 13min, 1min, 7min。

### 2.2 现有检测方法简介

Detection Filter<sup>[19]</sup>和 Feature squeezing<sup>[20]</sup>两种方法在检测 C&W 对抗样本时取得了较好的检测效果。本文方法与这两种方法的检测思路类似, 皆为预先对待测图像进行图像处理 P 操作, 然后根据处理前后图像的网络预测输出是否一致判断待测图像类型。Detection Filter<sup>[19]</sup>和 Feature squeezing<sup>[20]</sup>使用的 P 操作分别介绍如下。

Detection Filter<sup>[19]</sup>将对抗扰动看作作为一种特殊加性噪声。P 操作的目的是尽量去除对抗样本中的加性噪声, 使去噪后对抗样本的预测输出恢复为原始干净样本的类别。同时, P 操作还应尽量较少原始干净样本中的关键信息, 避免误认原始干净样本的类别。为此, P 操作由标量量化和空域平滑滤波构成, 关键信息由图像熵度量。具体地, 低熵和中等熵值图像的 P 操作仅采用标量量化, 步长分别为 128, 56。高熵值图像(即信息量较大的图像)的 P 操作先进行步长为 256/6 的量化, 然后再进行空域滤波, 最后对量化+滤波信息损失较大的像素进行校正, 即只保留量化操作。

Feature squeezing<sup>[20]</sup>认为原始图像作为神经网络的输入存在“冗余”信息。该冗余信息是针对神经网络分类而言, 例如, 即使将 8 比特灰度图像转换为 1 比特黑白图像, 人眼也能以较大概率识别黑白图像中的物体类别, 因而原文作者认为 8 比特灰度级图像存在冗余信息。Feature squeezing 使用 P 操作压缩图像的冗余信息, 减少对抗扰动存在的空间。P 操作包括减少图像灰度级, 中值滤波, 非局部滤波等 3 种操作。具体来说, 分别计算测试图像与每种操作图像在深度神经网络 softmax 输出的  $L_1$  距离。若最大的  $L_1$  距离超过给定阈值, 则测试图像判定为对抗样本。

Detection Filter<sup>[19]</sup>和 Feature squeezing<sup>[20]</sup>在检测 C&W 对抗样本时取得了较高的 TPR, 其缺点在于虚警率 FPR 较高。究其原因在于他们使用的 P 操作造成了原始图像信息的损失。我们在下文将使用 FFDNet 作为 P 操作, 尽力减少原始图像信息的损失, 以降低 FPR。

## 3 基于 FFDNET 的 C&W 对抗样本检测算法

### 3.1 检测模型

一般来说, 深度学习网络尽管不能抵抗对抗样

本的攻击, 但对于常规的图像后处理操作, 如图像缩放, 压缩, 平滑滤波, 随机噪声污染等, 具备一定程度的鲁棒性<sup>[28-29]</sup>。也就是说, 对于干净的原始样本, 深度学习网络的鲁棒性很大程度上能确保滤波前图像与滤波后图像的预测输出类别一致。然而, 对于对抗样本  $x + \delta$ , 其精心设计的对抗扰动  $\delta$  极有可能被平滑

滤波操作破坏, 如被过滤, 减弱或变成普通的随机噪声。因此, 我们假设平滑滤波后的对抗样本类似于经过常规后处理的干净图像。由此可得出, 对于对抗样本, 深度学习网络对于滤波前后版本的输出不一致。基于上述假设, 我们设计了图 1 所示的检测模型, 其中滤波器的设计最为关键。

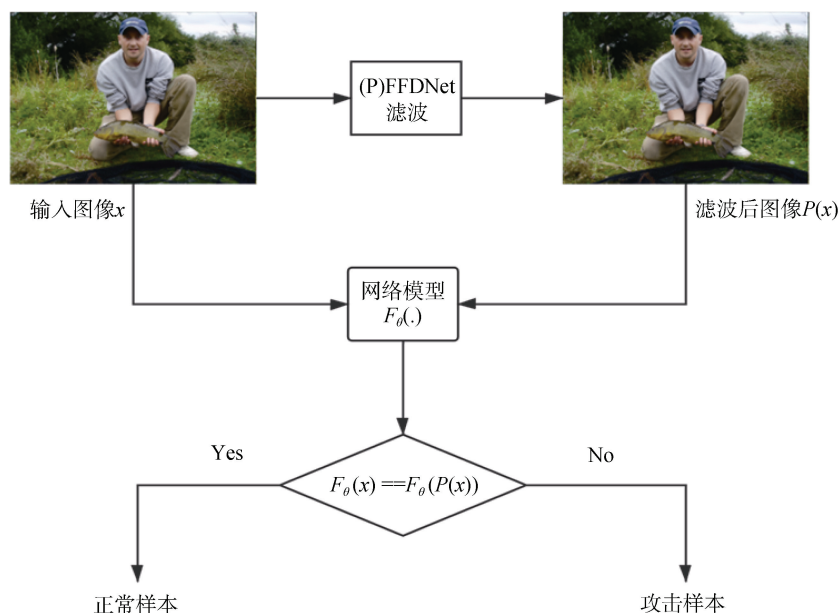


图 1 C&W 攻击样本检测流程

Figure 1 The detection process of C&W attack sample

论文[19-20]采用空域平滑滤波方法如中值滤波, 高斯滤波等在一定程度上实现了对抗样本的检测。但是这些方法的缺陷在于对某些干净图像的损伤较大, 如边缘失真, 导致深度学习模型误判干净图像滤波版本的类别, 即增加了检测方法的虚警率。因此, 我们需要一种模糊滤波方法既能破坏对抗样本的对抗扰动又能保持原始图像的关键信息(该关键信息是针对深度神经网络分类而言, 如轮廓, 边缘等信息), 实现检测准确率的最大化。

FFDNet(Fast and flexible denoising convolutional neural network)是一种基于 CNN 卷积网络的快速去噪算法<sup>[27]</sup>。它以噪声水平估计图作为输入, 能处理不同噪声强度下的滤波任务(对抗扰动对应强度较小的一类噪声), 具有速度快, 保持图像边缘等优点, 去噪效果优于 BM3D 等传统算法。本文基于 FFDNet 设计了图 1 所示的 C&W 对抗样本检测算法。若滤波后图像的预测输出  $F_\theta(P(x))$  等于输入图像的预测输出  $F_\theta(x)$ , 则输入图像判定为干净图像, 否则为对抗样本图像。

### 3.2 FFDNet 滤波器

FFDNet 用于图像去噪<sup>[27]</sup>。图像去噪本质上是退化图像的复原。对抗样本可视为添加加性噪声的退化图像<sup>[19]</sup>。从这一点上来说, 用 FFDNet 检测对抗样本和图像去噪有异曲同工之妙。





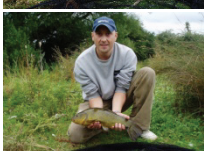
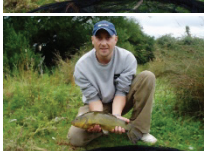
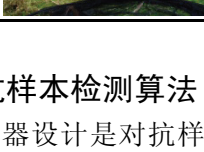
FFDNet 的网络结构如表 1 所示。首先, 为了加快处理速度, 输入图像被无损下采样(隔行隔列采样)为 4 个同等大小的图像块。小图像块连同噪声水平估计图作为后续卷积层的输入。噪声水平图由方差  $\sigma$  生成。 $\sigma$  越大, 噪声强度越强。对于尺寸为  $h \times w \times c$  ( $c$  为颜色通道数)的输入图像, 经过下采样后的输出大小为  $h/2 \times w/2 \times 4 \times (c+1)$ 。接下来是卷积操作, 共计 15 个卷积层, 前 14 个卷积层相继使用了卷积、归一化和 ReLU 激活函数操作, 注意未使用池化层, 每层使用了 96 个  $3 \times 3$  卷积核, 最后 1 个卷积层只进行卷积与归一化操作, 使用了 12 个  $3 \times 3$  卷积核。最后的输出图像由上采样小图像块重建生成。相较于传统方法, FFDNet 的优点在于较好地保持了图像质量。

表 1 FFDNet 网络结构, 以  $224 \times 224 \times 3$  RGB 图像为例说明

Table 1 Architecture of the FFDNet		
网络层	卷积参数	输出
下采样	--	$56 \times 56 \times 15$
Conv2d+BN+ReLU	$3 \times 3 \times 96$	$56 \times 56 \times 96$
...(13 个同样的层)	$3 \times 3 \times 96$	$56 \times 56 \times 96$
Conv2d+BN	$3 \times 3 \times 12$	$56 \times 56 \times 12$
上采样	---	$224 \times 224 \times 3$

表 2 给出了不同  $\sigma$  下 FFDNet 滤波图像与传统滤波图像的比较。从表格可以明显看出, 经过 FFDNet 处理之后的图像质量比较好, PSNR 峰值信噪比较高, 意味着失真较少。虽然随着  $\sigma$  增大, 图像的质量会变差, 但还是优于传统方法高斯滤波和中值滤波。

表 2 不同滤波方法处理后的图像质量比较  
Table 2 Comparison of image quality after different filtering methods

滤波方法		PSNR(dB)
FFDNet ( $\sigma = 4$ )		41.83
FFDNet ( $\sigma = 6$ )		
FFDNet ( $\sigma = 8$ )		
高斯滤波 ( $3 \times 3$ 滤波窗口)		37.65
中值滤波 ( $2 \times 2$ 滤波窗口)		35.12
		23.95
		22.36

### 3.3 C&W 对抗样本检测算法

FFDNet 滤波器设计是對抗样本检测算法的关键。考虑到 C&W 攻击添加的扰动较小, 为避免大幅

度地降低原始图像质量, 我们采用了  $\sigma$  较小的噪声水平图训练 FFDNet。同时, 为了处理多样化的自然图像, 使网络适应不同的噪声等级, 我们选择使用均匀分布的  $\sigma \in [1, 10]$ 。需要注意的是, 虽然训练采用的是高斯噪声, 并非表明 C&W 攻击添加的对抗扰动也为高斯噪声。事实上, 对抗扰动的分布难以估计<sup>[19]</sup>。由于我们使用 FFDNet 滤波的目的是破坏对抗扰动进而检测对抗图像, 而非去噪, 因而可认为选择  $\sigma \in [1, 10]$  构造训练图像是合理的。此外, 均匀分布的高斯噪声图像训练集使得 FFDNet 具备处理不同噪声图的能力, 即可看作一种依据噪声水平的“自适应滤波器”。FFDNet 训练使用的干净图像集由约 100 万张  $70 \times 70$  的图像块组成<sup>[27]</sup>, 噪声图像集由添加均匀分布的  $\sigma \in [1, 10]$  的高斯噪声生成。FFDNet 滤波器在 Intel Xeon CPU 3.40GHz+ 32G RAM+ 11G 1080Ti GPU 的配置下训练时间约为 38h。

在检测待测图像是否为对抗样本时, FFDNet 滤波器除了需要输入待测图像自身外, 还需要输入噪声水平估计图及其  $\sigma$ 。 $\sigma$  增大, FFDNet 滤波强度增加, 图像质量会变差, 有可能会影响原始干净图像的检测。参数  $\sigma$  由训练数据确定(请参见第 4.3 节)。我们通过判断滤波前后图像的输出差异检测 C&W 对抗样本。具体地, 若滤波后图像的预测输出  $F_\theta(P(x))$  等于输入图像的预测输出  $F_\theta(x)$ , 则输入图像判定为干净图像, 否则为对抗样本图像。

## 4 实验结果

我们选用  $L_0, L_2, L_\infty$  3 种距离度量, 对应于 C&W 有目标攻击与无目标攻击, 共计使用 6 种攻击方法进行实验。

### 4.1 实验设置

实验图像库来自 ImageNet ILSVRC2012 中的验证集。ILSVRC2012 验证集包含 1000 个类别, 每个类别有 50 张图像, 合计 50000 张图像。由于实验条件限制, 我们将图像库按标签升序排列, 选择前 100 个类别, 合计 5000 张图像用于实验。目前较为先进的 ResNet-50<sup>[1]</sup>作为攻击目标(网络结构如表 3 所示)。在预训练模型上, 对于所选择的 5000 张图像, ResNet-50 的 Top-1 分类准确率为 73.42%。

C&W 实验采用白盒攻击。我们仅选择分类正确的图像生成对抗样本。由于 C&W 攻击生成对抗样本速度较为缓慢, 我们分别随机选择 2000 张, 500 张, 100 张原始图像分别生成  $L_2$  对抗样本,  $L_\infty$  对抗样本和  $L_0$  对抗样本。C&W 攻击的参数  $\kappa$  均为默认值 0。



表 3 ResNet50 卷积神经网络结构图

Table 3 Architecture of the ResNet50

网络层结构	输出尺寸	50-layer
Conv1	112×112	7×7, 64, stride 2 3×3 max pool, stride 2†
Conv2_x	56×56	1×1.64 3×3.64×3 1×1.256
Conv3_x	28×28	1×1.128 3×3.128×4 1×1.512
Conv4_x	14×14	1×1.256 3×3.256×6 1×1.1024
Conv5_x	7×7	1×1.512 3×3.512×3 1×1.2048
	1×1	average pool, 1000-d fc, softmax

有目标攻击使用 next 方式, 即分类器将对抗样本的标签误认为相邻下一个标签。对抗样本生成代码由论文作者提供<sup>①</sup>。每种攻击图像的数量请参见表 4。

实验结果评估采用召回率 (Recall)、精确率 (Precision) 和 F1-score。其计算公式如(3)~(5)所示, TP 指检测器将对抗样本检测为对抗样本的数量, FN 指检测器将对抗样本检测为原始样本的数量, FP 指检测器将原始样本误认为为对抗样本的数量, TN 指检测器将原始样本检测为原始样本的数量。

表 4 C&amp;W 对抗样本图像数量

Table 4 Number of images used in C&amp;W attacks

攻击方法	目标攻击	图像数量
$L_0$	是	100
$L_0$	否	100
$L_2$	是	2000
$L_2$	否	2000
$L_\infty$	是	500
$L_\infty$	否	500

$$Recall = \frac{TP}{TP + FN} \quad (3)$$

$$Precision = \frac{TP}{TP + FP} \quad (4)$$

$$F1 = 2 * \frac{Recall * Precision}{Recall + Precision} \quad (5)$$

## 4.2 C&W 攻击效果

表 5 展示了 6 种方法生成的对抗样本的攻击成功率。测试图像来源于表 4。实验结果表明, C&W 攻击几乎能 100% 攻击成功, 即 ResNet-50 完全误判对抗样本的类别。

表 5 C&amp;W 对于 ResNet-50 的攻击成功率

Table 5 Success rate of the C&amp;W attack on the ResNet-50 classifier

攻击方法	目标攻击	攻击成功数	攻击成功率/%
$L_0$	是	100	100
$L_0$	否	100	100
$L_2$	是	1999	99.95
$L_2$	否	2000	100
$L_\infty$	是	500	100
$L_\infty$	否	500	100

图 2 展示了对抗样本的图像质量。从主观上看, C&W 攻击对图像质量的影响非常小, 人眼几乎无法发现正常图像与对抗图像的区别。客观上, 我们采用峰值信噪比 (PSNR) 评估图像质量, PSNR 越高说明对抗样本对原始图像的干扰越小, 图像质量越高。可以看出, C&W 攻击产生的对抗样本 PSNR 均位于较高的水平。总体上看, 无目标攻击比有目标攻击对原始图像的干扰强度小。正如 C&W 攻击<sup>[12]</sup>所示,  $L_2$ 、 $L_\infty$  和  $L_0$  攻击对原始图像的干扰强度依次增大。

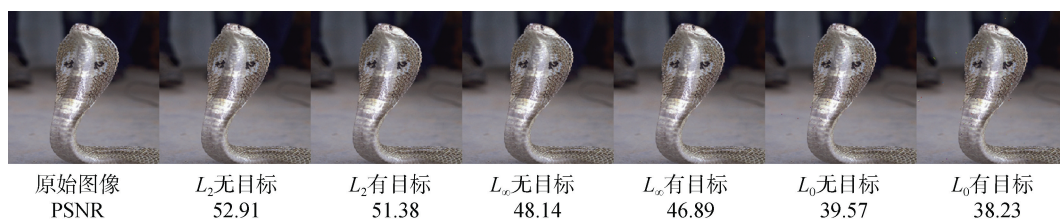


图 2 对抗样本及其峰值信噪比

Figure 2 Adversarial samples and their peak signal-to-noise ratio

① <https://github.com/mzweilin/EvadeML-Zoo>

### 4.3 参数 $\sigma$ 选择

参数  $\sigma$  控制了滤波图像的质量, 影响本文方法的检测准确率。由于在真实情况下不可能知道具体的攻击方法, 所以如何选择一个合适的  $\sigma$  均衡各种攻击方法之间的检测效果, 是一个比较复杂的工作。我们在本小节通过实验验证选择参数  $\sigma$ 。

表 6 展示了本文方法在不同  $\sigma$  取值下检测无目标攻击对抗样本的结果。在  $L_2$  攻击下, 所有 4 个  $\sigma$  值取得了近似的  $F1$  值。在  $L_0$  和  $L_\infty$  攻击下, 使用  $\sigma=8$  的方法取得了最好的  $F1$  值, 分别超第二名 3.75%, 2.03%。通过以上分析, 我们统一使用参数  $\sigma=8$ 。

表 6 本文方法在不同  $\sigma$  下的检测结果

Table 6 Detection results of the proposed method for various values of  $\sigma$

攻击方法	$\sigma$	对抗样本数量	$TP$	$FP$	$TN$	$FN$	Recall/%	Precision/%	$F1$ /%
$L_2$ 无目标	3	2000	1831	<b>31</b>	<b>1969</b>	169	91.55	<b>98.34</b>	94.82
$L_2$ 无目标	4	2000	1875	45	1955	125	93.75	97.66	<b>95.66</b>
$L_2$ 无目标	6	2000	1881	77	1923	119	94.05	96.07	95.05
$L_2$ 无目标	8	2000	<b>1882</b>	104	1896	<b>118</b>	<b>94.1</b>	94.76	94.43
$L_\infty$ 无目标	3	500	282	<b>9</b>	<b>491</b>	218	56.4	<b>96.91</b>	71.3
$L_\infty$ 无目标	4	500	328	13	487	172	65.6	96.19	78
$L_\infty$ 无目标	6	500	387	19	481	113	77.4	95.23	85.43
$L_\infty$ 无目标	8	500	<b>408</b>	25	475	<b>92</b>	<b>81.6</b>	94.23	<b>87.46</b>
$L_0$ 无目标	3	100	71	<b>3</b>	<b>97</b>	29	71	<b>95.95</b>	81.61
$L_0$ 无目标	4	100	77	4	96	23	77	95.06	85.08
$L_0$ 无目标	6	100	78	5	95	22	78	93.98	85.25
$L_0$ 无目标	8	100	<b>85</b>	6	94	<b>15</b>	<b>85</b>	93.41	<b>89</b>

### 4.4 对比实验

论文[19-20]的方法是目前较为先进的检测算法, 我们在这一小节将本文方法与它们对比。下文将论文[19-20]的方法分别称为 Detection filter 和 Feature squeezing。Detection filter 和 Feature squeezing 都使用了论文中提供的代码进行实验。Feature squeezing 进行了最佳阈值选择, 设为 0.78。表 7 展示了本文方法、Detection filter 以及 Feature squeezing 对 C&W 攻击对抗样本的检测结果。整体上看, 有目标攻击的对抗样本比无目标攻击的对抗样本更容易被检测出来, 三种检测方法均能在有目标攻击中取得较好的效果, 而无目标攻击的效果差一些, 原因可能是有目标攻击产生的干扰强度要高于无目标攻击产生的干扰强度, 干扰强度越大越容易被检测。对于 C&W 的 6 种攻击方式, 本文方法的  $F1$  值均超过 Detection filter 和 Feature squeezing。在检测  $L_2$  无目标攻击时, 本文方法的  $F1$  值比 Detection filter 的  $F1$  值高 9.02%。Detection filter 的  $FP$  远高于本文方法和 Feature squeezing, 可能是因为 Detection filter 使用的量化和滤波(滤波窗口为  $7 \times 7$ )对干净原始图像质量影响非常大, 导致检测器

无法正确检测原始图像。Feature squeezing 在  $L_0$  和  $L_\infty$  实验的结果比本文方法稍差一些, 但是在  $L_2$  实验上比本文方法差很多, 本文方法的  $F1$  值为 94.43%, 而 Feature squeezing 的  $F1$  值为 83.18%, 相差了 11.25%。究其原因在于 Feature squeezing 涉及的参数较多, 而且对数据集敏感, 不同的对抗样本数据需要选择不同的阈值, 当无法获得训练数据用于调参时, 检测效果往往较差。同时, 表 7 显示大多数情况下 Feature squeezing 的  $FN$  高于本文方法和 Detection filter, 表明 Feature squeezing 更容易漏检对抗样本。相较于 Detection filter 和 Feature squeezing, 本文方法参数少, 速度快, 通用性强。虽然  $TP$  有时低于 Detection filter, 但是本文方法没有将大量正常样本检测为对抗样本, 即虚警率低。

在某些实际情形下, 我们可能并不知道攻击的具体类型。为此, 我们综合所有攻击样本(5199 对抗样本, 5199 干净样本)去评估本文方法的检测能力。表 8 的结果表明, 本文方法, Detection filter 和 Feature squeezing 分别获得了  $F1=94.97\%$ ,  $87.71\%$ , 和  $88.72\%$ 。从总体结果来看, 本文方法的检测效果仍然优于其余 2 种对比方法。

表 7 Detection filter、Feature squeezing 和本文方法的检测结果

Table 7 Detection results of the detection filter, the feature squeezing and the proposed method

检测方法	攻击方法	对抗样本数量	TP	FP	TN	FN	Recall/%	Precision/%	F1/%
Detection filter[19]	$L_2$ 无目标	2000	1771	376	1624	229	88.55	82.49	85.41
Feature squeezing[20]	$L_2$ 无目标	2000	1538	160	1840	462	76.9	90.58	83.18
本文方法	$L_2$ 无目标	2000	1882	104	1896	118	94.1	94.76	<b>94.43</b>
Detection filter	$L_2$ 有目标	1999	1960	376	1623	39	98.05	83.9	90.42
Feature squeezing	$L_2$ 有目标	1999	1860	160	1839	139	93.05	92.08	92.56
本文方法	$L_2$ 有目标	1999	1984	104	1895	15	99.25	95.02	<b>97.09</b>
Detection filter	$L_\infty$ 无目标	500	429	101	399	71	85.8	80.94	83.3
Feature squeezing	$L_\infty$ 无目标	500	412	33	467	88	82.4	92.58	87.2
本文方法	$L_\infty$ 无目标	500	408	25	475	92	81.6	94.23	<b>87.46</b>
Detection filter	$L_\infty$ 有目标	500	494	101	399	6	98.8	83.03	90.23
Feature squeezing	$L_\infty$ 有目标	500	475	33	467	25	95	93.5	94.25
本文方法	$L_\infty$ 有目标	500	491	25	475	9	98.2	95.16	<b>96.65</b>
Detection filter	$L_0$ 无目标	100	88	22	78	12	88	80	83.81
Feature squeezing	$L_0$ 无目标	100	83	5	95	17	83	94.32	88.3
本文方法	$L_0$ 无目标	100	85	6	94	15	85	93.41	<b>89</b>
Detection filter	$L_0$ 有目标	100	98	22	78	2	98	81.67	89.09
Feature squeezing	$L_0$ 有目标	100	93	5	95	7	93	94.9	93.94
本文方法	$L_0$ 有目标	100	95	6	94	5	95	94.06	<b>94.53</b>

表 8 Detection filter、Feature squeezing 和本文方法的总体检测结果

Table 8 The overall detection results of the detection filter, the feature squeezing and the proposed method

检测方法	攻击方法	对抗样本数量	TP	FP	TN	FN	Recall/%	Precision/%	F1/%
Detection filter[19]	C&W	5199	4840	998	4201	359	93.09	82.91	87.71
Feature squeezing[20]	C&W	5199	4461	396	4803	738	85.81	91.85	88.72
本文方法	C&W	5199	4945	270	4929	254	95.11	94.82	<b>94.97</b>

我们在实验中发现了一个有意思的现象。本文方法、Detection filter 以及 Feature squeezing 方法都在  $L_2$  攻击检测上取得了最好的效果, 而  $L_2$  攻击样本的质量是最好的, 即添加的对抗扰动最小。按照隐写分析的观点, 隐写嵌入率越低, 嵌入信息越少, 隐写分析准确率应越低。文献[25]将 ESRM 用于检测对抗样本时也得到了类似隐写分析的结论, FGSM 攻击比 C&W 攻击添加了更大的对抗扰动, 因而 FGSM 更容易被检测。上述实验结果表明, 本文方法可以和隐写分析方法有效结合, 实现对不同对抗扰动等级的对抗样本检测, 这是我们未来的一个研究方向。

## 5 结论

针对 C&W 白盒攻击, 本文提出了基于 FFDNet 滤波器的对抗样本检测算法, 实现了对 C&W 对抗样本的有效检测。相对于已有工作, 本文方法不仅降低了虚警率, 而且实现了较高的 F1 值。本文方法在检

测 C&W  $L_2$  攻击时, 取得了较为满意的结果, 而对  $L_0$  和  $L_\infty$  攻击的检测仍需进一步提高。本文方法在虚警率 FPR 方面取得了较为满意的结果, 后续工作可通过提升正阳性检测率 TPR 进一步提升 F1 值。自适应地选取噪声水平估计图, 使 FFDNet 能进一步地破坏对抗扰动进而提升 TPR 是本文未来的研究方向。另一研究方向是结合隐写分析, 数字图像取证与本文方法检测更多的对抗攻击。

## 参考文献

- [1] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]. *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016: 770-778.
- [2] Garcia-Garcia A, Orts-Escobedo S, Oprea S, et al. A Review on Deep Learning Techniques Applied to Semantic Segmentation[EB/OL]. 2017: arXiv:1704.06857[cs.CV]. <https://arxiv.org/abs/1704.06857>.



- [3] Szegedy C, Zaremba W, Sutskever I, et al. Intriguing properties of neural networks[J]. 2013: arXiv preprint arXiv:1312.6199.
- [4] Bringing Big Neural Networks to Self-Driving Cars, Smartphones, and Drones. <https://spectrum.ieee.org/computing/embedded-systems/bringing-big-neural-networks-to-selfdriving-cars-smartphones-and-drones>. Mar. 2016.
- [5] Mnih V, Kavukcuoglu K, Silver D, et al. Human-level Control through Deep Reinforcement Learning[J]. *Nature*, 2015, 518(7540): 529-533.
- [6] Yuan X, He P, Zhu Q, et al. Adversarial examples: Attacks and defenses for deep learning[J]. *IEEE transactions on neural networks and learning systems*, 2019, 30(9): 2805-2824.
- [7] Goodfellow I J, Shlens J, Szegedy C. Explaining and harnessing adversarial examples[J]. 2014: arXiv preprint arXiv:1412.6572.
- [8] Tramèr F, Kurakin A, Papernot N, et al. Ensemble adversarial training: Attacks and defenses[J]. 2017: arXiv preprint arXiv:1705.07204.
- [9] Madry A, Makelov A, Schmidt L, et al. Towards deep learning models resistant to adversarial attacks[J]. 2017: arXiv preprint arXiv:1706.06083.
- [10] Papernot N, McDaniel P, Jha S, et al. The Limitations of Deep Learning in Adversarial Settings[EB/OL]. 2015: arXiv:1511.07528 [cs.CR]. <https://arxiv.org/abs/1511.07528>.
- [11] Moosavi-Dezfooli S M, Fawzi A, Frossard P. DeepFool: A Simple and Accurate Method to Fool Deep Neural Networks[EB/OL]. 2015: arXiv:1511.04599[cs.LG]. <https://arxiv.org/abs/1511.04599>.
- [12] Carlini N, Wagner D. Towards evaluating the robustness of neural networks[C]. *2017 IEEE Symposium on Security and Privacy (SP)*. IEEE, 2017: 39-57.
- [13] Goodfellow I J, Shlens J, Szegedy C. Explaining and harnessing adversarial examples[J]. 2014: arXiv preprint arXiv:1412.6572.
- [14] Papernot N, McDaniel P, Wu X, et al. Distillation as a Defense to Adversarial Perturbations Against Deep Neural Networks[EB/OL]. 2015: arXiv:1511.04508[cs.CR]. <https://arxiv.org/abs/1511.04508>.
- [15] Papernot N, McDaniel P, Goodfellow I, et al. Practical Black-Box Attacks Against Machine Learning[C]. *the 2017 ACM on Asia Conference on Computer and Communications Security*, 2017: 506-519.
- [16] Hendrycks D, Gimpel K. Early Methods for Detecting Adversarial Images[EB/OL]. 2016: arXiv:1608.00530[cs.LG]. <https://arxiv.org/abs/1608.00530>.
- [17] Li X, Li F. Adversarial examples detection in deep networks with convolutional filter statistics[C]. *Proceedings of the IEEE International Conference on Computer Vision*. 2017: 5764-5772.
- [18] Lu J J, Issarano T, Forsyth D. SafetyNet: Detecting and Rejecting Adversarial Examples Robustly[EB/OL]. 2017: arXiv:1704.00103 [cs.CV]. <https://arxiv.org/abs/1704.00103>.
- [19] Liang B, Li H C, Su M Q, et al. Detecting Adversarial Image Examples in Deep Neural Networks with Adaptive Noise Reduction[J]. *IEEE Transactions on Dependable and Secure Computing*, 2019: 1.
- [20] Xu W, Evans D, Qi Y. Feature squeezing: Detecting adversarial examples in deep neural networks[J]. 2017: arXiv preprint arXiv:1704.01155.
- [21] Guo F, Zhao Q J, Li X, et al. Detecting Adversarial Examples via Prediction Difference for Deep Neural Networks[J]. *Information Sciences*, 2019, 501: 182-192.
- [22] Yang J Y. IDAE: Iterative Denoising Autoencoder Based Deep Learning Model Enhancement Mechanism Against Adversarial Examples[J]. *Journal of Cyber Security*, 2019, 4(6): 34-44. (杨俊宇. 基于迭代自编码器的深度学习对抗样本防御方案[J]. *信息安全学报*, 2019, 4(6): 34-44.)
- [23] Fan W Q, Sun G L, Su Y Y, et al. Integration of Statistical Detector and Gaussian Noise Injection Detector for Adversarial Example Detection in Deep Neural Networks[J]. *Multimedia Tools and Applications*, 2019, 78(14): 20409-20429.
- [24] Pevny T, Bas P, Fridrich J. Steganalysis by Subtractive Pixel Adjacency Matrix[J]. *IEEE Transactions on Information Forensics and Security*, 2010, 5(2): 215-224.
- [25] Liu J Y, Zhang W M, Zhang Y W, et al. Detection Based Defense Against Adversarial Examples from the Steganalysis Point of View[EB/OL]. 2018: arXiv:1806.09186[cs.CV]. <https://arxiv.org/abs/1806.09186>.
- [26] Fridrich J, Kodovsky J. Rich Models for Steganalysis of Digital Images[J]. *IEEE Transactions on Information Forensics and Security*, 2012, 7(3): 868-882.
- [27] Zhang K, Zuo W M, Zhang L. FFDNet: Toward a Fast and Flexible Solution for CNN-Based Image Denoising[J]. *IEEE Transactions on Image Processing*, 2018, 27(9): 4608-4622.
- [28] Goodfellow I, Lee H, Le Q V, et al. Measuring invariances in deep networks[C]. *Advances in neural information processing systems*. 2009: 646-654.
- [29] LeCun Y, Boser B, Denker J S, et al. Backpropagation applied to handwritten zip code recognition[J]. *Neural computation*, 1989, 1(4): 541-551.
- [30] Athalye A, Carlini N, Wagner D. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples[J]. 2018: arXiv preprint arXiv:1802.00420.



**邓康** 现就读于西南科技大学。主要研究兴趣为深度学习安全, 多媒体安全与取证。Email: limitedin@163.com



**彭安杰** 2015 年于中山大学通信与信息系统专业或博士学位, 现为西南科技大学计算机科学与技术学院副教授。研究方向包括多媒体安全与取证, 机器学习, 人工智能安全等。Email: penga-jie200012@163.com。



**黄晓芳** 于 2010 年获得博士学位, 现为西南科技大学教授, 硕士生导师。主要研究领域为区块链、身份认证及公钥密码学方面的应用研究。Email: xf.swust@qq.com



**罗胜海** 现就读于西南科技大学。主要研究兴趣为深度学习安全, 机器学习。Email: lsh599194771@163.com



**曾辉** 2016 年于中山大学, 获得通信与信息系统专业博士学位。现为西南科技大学副研究员, 主要研究兴趣为视频/图像取证, 博弈论。Email: zengh5@mail2.sysu.edu.cn.