从频域角度重新分析对抗样本*

丁 烨[1],王 杰[1],宛 齐[1],廖 清[2]
(1.东莞理工学院 网络空间安全学院,广东 东莞 523820;
2.哈尔滨工业大学(深圳) 计算机科学与技术学院,广东 深圳 518055)

摘 要:目前在空间域上关于对抗样本的研究成果已经相当成熟,但是在频域上的相关工作却是十分缺乏。从频域的角度对对抗样本进行深入的研究,发现对抗样本在DCT域上表现出了高度可识别的伪影,并利用这些伪影信息训练了一个基于频域的对抗样本检测器CNN-DCT,结果表明,对于常见的对抗样本在数据集CIFAR-10和SVHN上都能达到98%的检测准确率。此外,针对对抗样本在频域上存在的伪影,也提出一种通用的改进算法IAA-DCT来解决。简而言之,本文不仅填充了对抗样本在频域上工作的缺少,也改进了对抗攻击算法在频域上存在伪影的弊端。

关键词:对抗样本;频域;DCT域;对抗攻击
中图分类号:TP391 **文献标识码**:A **DOI**:10.19358/j.issn.2096-5133.2022.05.009
引用格式:丁烨,王杰,宛齐,等.从频域角度重新分析对抗样本[J].信息技术与网络安全,2022,41(5):59-65,76.

Analysis of adversarial examples from frequency domain

Ding Ye[1], Wang Jie[1], Wan Qi[1], Liao Qing[2]
(1.School of Cyberspace Security, Dongguan University of Technology, Dongguan 523820, China;
2.School of Computer Science and Technology, Harbin Institute of Technology(Shenzhen), Shenzhen 518055, China)

Abstract: Research on adversarial examples in spatial domain is well studied, but related works in frequency domain is scarce. In this paper, we conduct thorough study of adversarial examples in frequency domain and find that adversarial examples exhibit highly identifiable artifacts in Discrete cosine transform(DCT) domain. Hence, a frequency domain-based adversarial example detector, CNN-DCT, is trained based on such artifact information, and the results achieve 98% detection accuracy for common adversarial examples on both CIFAR-10 and SVHN datasets. In addition, a general improved algorithm, IAA-DCT, is also proposed to address the artifacts that exist in the frequency domain for the adversarial examples. In conclusion, this paper not only provides studies of adversarial examples in frequency domain, but also improves the disadvantages of the adversarial attack algorithm with artifacts in the frequency domain.

Key words: adversarial example; frequency domain; discrete cosine transform(DCT) domain; adversarial attack

0 引言

对抗攻击通过在深度学习模型中加入人类视觉上无法察觉的扰动,被称为对抗样本[1]。对抗样本可以使模型受到干扰而产生错误的分类,从而导致错误类别的置信度大于正确类别的置信度。随着深度学习在不同的任务上取得优异性能,如人脸识别、自动驾驶、会议记录等,对人类社会进步带来了巨大的贡献。然而在许多的研究工作中,对抗攻击被证明可以在图像、视频、语音等领域的深度学习中执行恶意任务,从而造成重大的安全问题。

为了解决对抗攻击带来的影响,避免这种恶意的攻击,研究者们开始了对对抗攻击的防御工作。对抗防御主要分为两个方面,一个方面是直接改进模型而让现有的对抗攻击方法失效,如防御性蒸馏[2]。另外一个方面是进行对抗样本的检测。关于对抗检测的研究主要集中在图像域中对图片特征处理,如Xu等人[3]提出了一种基于特征压缩的对抗样本检测方法;Joel等人[4]在频谱上综合分析了现有的攻

*基金项目:国家自然科学基金面上项目(61976051);国家自然科学基金联合基金重点支持项目(U19A2067)

击方法和数据集，发现大部分的对抗样本在频域都出现了严重的伪影，并且在频域空间这些伪影数据可以分离，从而能够分类识别。

受到 Joel 等人[4]的启发，本文将对抗攻击后的图像和原始图像变换到 DCT 域[5]上进行频谱分析。通过对比，本文发现所有的对抗样本频谱图上都表现出了与原始图像频谱图明显的不同。由于攻击方法的方式迥异，本文更进一步分析了不同攻击方法产生的对抗样本和原始样本之间的 DCT 频谱，并表明对抗样本和原始样本的频谱图在高频上都出现了严重伪影的原因是来自于扰动的生成方法。

基于上述分析，本文设计了一个基于频域信息进行分类的 CNN－DCT 模型，相比较以往基于空间域信息进行分类的 CNN 模型，大大提高了对抗样本的检测准确率。该模型在同一数据集上检测目前常见的对抗攻击方法，产生的对抗样本能够达到 98％的检测准确率，在 DCT 域上能极大程度区分开对抗样本和干净样本，可以为深度网络模型训练所需的数据集划分出干净的正常样本，从而提高模型的性能。该模型在物理世界里也具有较强的实用性，可以用来检测现实世界中“对抗样本”。例如在无人驾驶技术中，需要车载模型来识别道路路标，而一些路标容易被有意或无意地添加了扰动(对抗样本)，从而使车载模型判别错误，造成不必要的麻烦和潜在的危险。通过本文提出的 CNN－DCT 模型，可以判定该路标是否为干净的样本，提前进行风险规避。

值得注意的是，考虑到对抗样本在频域中存在伪影且可能被利用来检测的弊端，本文尝试去优化扰动生成的方法，通过一个低通滤波直接作用在不同攻击方法产生的对抗扰动上，使对抗样本与原始样本在频域上也尽可能表现一致。然而，这种改进的方法虽然十分简单，却严重降低了原始方法的攻击成功率。因此，本文提出了 IAA－DCT 算法，通过重新设定对抗样本的生成方式，利用启发式的方法去搜索在频域上能与原始样本一致的对抗样本，从而降低对抗样本的检测率。实验表明，本文的 IAA－DCT 算法明显降低了 CNN－DCT 模型关于对抗样本的检测率，并在一定程度保持或者提高了攻击的成功率。

本文的工作从频域的角度证明了对抗攻击存在的严重弊端，弥补了对抗样本在频域工作的不足。本文主要的贡献总结为以下几点：

(1)本文通过频域的角度发现了对抗样本在频谱上存在着与原始样本频域严重的不同。

(2)基于频域信息的不同，本文提出了一种在频域上的对抗样本检测模型 CNN－DCT，极大地提高了对抗样本检测率。

(3)针对对抗样本在频域上存在的伪影且易被利用来分类的弊端，本文设计了一种通用的优化算法 IAA－DCT，在保持攻击成功率的同时极大降低了对抗样本通过 CNN－DCT 的准确率。

1 图像域上的对抗样本

本节主要介绍图像域上对抗样本的通用生成方法以及几种经典对抗样本的原理，为下文分析对抗样本在频域上存在伪影以及提出改进算法 IAA－DCT 做介绍。

1.1 方法定义

在基于图像分类的对抗样本研究中，对抗攻击的目标是通过在自然图像上添加一个精心设计且难以察觉的扰动来干扰模型的推理结果。在形式上，它通过设计不同的优化问题来找到满足条件的扰动。给定一个分类模型 f 和一张图像 x，这个优化问题的一般数学表达如下：

$$\min\delta(r),\ \text{subject to:}\quad f(x+r)=y_t \ \text{ and } \ x+r\in[0,1]^m \tag{1}$$

其中 $\delta(\cdot)$ 表示的是干净样本和对抗样本之间的距离表达式，r 是添加到图像 x 上的扰动大小，常见的衡量方式有 L_1 范数 $||r||_1$、L_2 范数 $||r||_2$ 和 L_∞ 范数 $||r||_\infty$，而 $f(x)$ 输出的是 x 对应的分类结果，y_t 为指定的类别标签。

1.2 分类和方法

根据模型结构以及参数是否已知，可将对抗攻击分为白盒攻击和黑盒攻击。本文只讨论了白盒攻击方法下的两种主要实现手段。

一是基于梯度信息进行攻击，经典的研究工作有 FGSM[6]、BIM[7]、PGD[8]。文献[6]在 2014 年利用对抗样本的线性解释提出了一个快速产生对抗样本的方式，也即 Fast Gradient Sign Method(FGSM)方法。假定模型参数值为 θ，模型的损失函数为 $H(\theta,x,y)$。FGSM 方法在无穷范数限制下($||\eta||_\infty<\varepsilon$)添加扰动 $\eta=\varepsilon\text{sign}(\nabla x\,H(\theta,x,y))$，其中 ε 为限定扰动值大小的常量，$\text{sign}(\cdot)$表示为取变量值正负符号的函数，$\nabla x\,H(\cdot)$

表示的是损失函数 H 关于 x 的梯。用 x' 表示最后生成的对抗样本，则 FGSM 方法的完整公式如下：

$$x' = x + \varepsilon \text{sign}(\nabla x H(\theta, x, y)) \tag{2}$$

这是一种简单的单步攻击方法，存在噪声大、攻击率低的弊端。于是文献[7]基于之前的 FGSM 攻击方法做出了一部分改进，其中用迭代攻击代替单步攻击，于是提出了 Basic Iterative Methods(BIM)攻击方法，BIM 完整的攻击公式如下：

$$\begin{cases} x_0' = x \\ x_n' + 1 = \text{clip}(x_n' + \varepsilon \text{sign}(\nabla x H(\theta, x, y))) \end{cases} \tag{3}$$

初始化第一个对抗样本 x' 为原始样本 x，总共迭代 n 次，其中 Clip(·)函数是将每次迭代的扰动大小限制在一定的范围内。后来在文献[8]中指出这实际上等价于无穷范数版本的 Projected Gradient Descent(PGD)。

另外一种白盒攻击方法基于约束优化问题来实现。比如 C&W[9]方法，常规方法通过构造约束优化问题来创建对抗样本，具体见式(1)。但其中的方程约束很难推导，因此作者将该方程进行了如下变换：

$$f(x+r) = y_t \Rightarrow C(x+r) \leqslant 0 \tag{4}$$

文献[9]给出了 7 个目标函数 $C(\cdot)$，本文不作详细描述。

2 频域上的对抗样本

第 1 节介绍了图像域上对抗样本的相关工作以及几种经典对抗攻击方法，如 FGSM、BIM、PGD 和 C&W 的生成原理。它们都在图像域上对自然图像添加扰动而忽视了对频域的考虑。因此本节通过引入 DCT 域来分析对抗样本，提出一个基于 DCT 系数的对抗样本检测器 CNN-DCT。并且针对 DCT 频谱的差异性易被利用来分类对抗样本和原始样本，提出了改进算法 IAA-DCT。

2.1 DCT 域的定义

离散余弦变换(DCT)是一种与傅里叶变换相关的变换。对于二维功能(如图像)，DCT 允许视觉上的重要信息集中在一个小的信息上。因此，DCT 是针对 JPEG 压缩的国际标准有损算法的核心组成部分。它还可以将一个函数表示为不同振幅和频率的许多余弦函数的和，将信号从时空域转换为频域。1D-DCT 的一般公式如下：

$$F(u) = A(u) \sum_{x=0}^{N-1} f(x) \frac{(2x+1)u\pi}{2N} \tag{5}$$

$$A(u) = \begin{cases} \dfrac{1}{\sqrt{N}}, & u = 0 \\ \sqrt{\dfrac{2}{N}}, & u \neq 0 \end{cases} \tag{6}$$

其中，$F(u)$ 为余弦变换值，u 为广义频率变量，$u = 1, 2, \cdots, N-1$；$f(x)$ 为时域中 N 个点的序列 $x = 1, 2, \cdots, N-1$。

本文进行了利用 2D-DCT 将图像从空间域转换为 DCT 域的实验，给定一个 2D 图像 $X \in \mathrm{R}^{d \times d}$，定义一个基础函数：

$$\Psi_d(i, j) = \cos\left[\frac{\pi}{d}(i + \frac{1}{2})j\right] \tag{7}$$

对于 $1 \leqslant i, j \leqslant d$，2D-DCT 变换 $V = \text{DCT}(X)$ 具体公式展开如下：

$$V_{j_1, j_2} = N_{j_1} N_{j_2} \sum_{i_1=0}^{d-1} \sum_{i_2=0}^{d-1} X_{i_1, i_2} \Psi_d(i_1, j_2) \Psi_d(i_1, j_2) \tag{8}$$

$$N_j = \begin{cases} \dfrac{1}{\sqrt{d}}, & j = 0 \\ \sqrt{\dfrac{2}{d}}, & j \neq 0 \end{cases} \tag{9}$$

其中，N_{j_1}、N_{j_2} 是归一化项，以确保图变换是等距的，例如 $||X||_2 = ||\text{DCT}(X)||_2$。$V_{i,j}$ 项对应于 $\Psi_d(i, j)$ 波的幅值，低频率用低 i、j 表示。此外，DCT 是可逆的，逆 $X = \text{IDCT}(V)$，具体展开如下：

$$X_{i_1, i_2} = \sum_{j_1=0}^{d-1} \sum_{j_2=0}^{d-1} N_{j_1} N_{j_2} V_{j_1, j_2} \Psi_d(i_1, j_2) \Psi_d(i_1, j_2) \tag{10}$$

对于包含多个彩色通道的图像，DCT 和 IDCT 可以分别在通道上应用。

2.2 DCT 域上的对抗样本分析

为了更好地探究对抗样本在频域中的变化，本文对 CIFAR-10 数据集进行对抗攻击，该数据集共有 60 000 张分辨率大小为 32×32 的彩色图像，总共划分为 10 个类，每类 6 000 张图。本文将二维 DCT 变换后的 DCT 系数绘制为热力图(Heatmap)，如图 1 所示。每个 DCT 系数对应空间频率对图像的贡献比例。在实践中，本文对图像的每个通道分别进行行和列的 1D-DCT 变换，将它们相乘(对应于水平和垂直方向)，得到 2D-DC 变换后的系数，然后进行加权平均。热力图的左上区域对应图像的低频，右下区域对应的是图像的高频。当从低频观察高频时，可以注意到系数下降得特别快，因此本文在制作热图之前截取了 2.0~4.5 范围的系数。

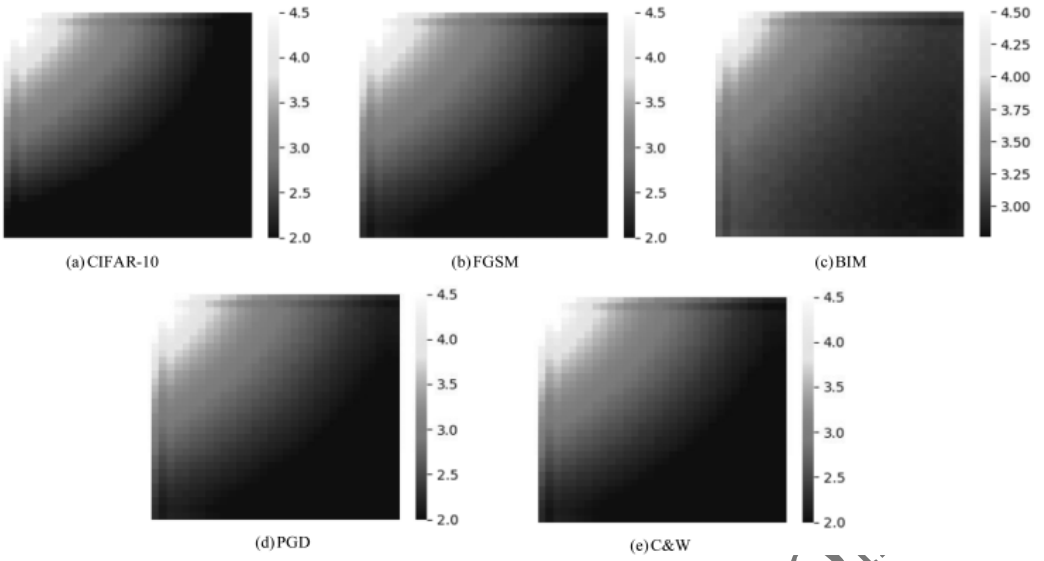

图 1 不同算法在 CIFAR-10 上攻击后的平均图谱结果

其中，图 1(a)是 CIFAR-10 数据集 10 000 个干净样本的平均频谱，平均频谱指的是样本经过 DCT 变换后求平均得到的频谱图。图 1(b)、图 1(c)、图 1(d)和图 1(e)分别是 FGSM[6]、BIM[7]、PGD[8]和 C&W[9]这几种经典对抗攻击方法在数据集 CIFAR-10 上产生的 10 000 个对抗样本的 DCT 平均图谱。与文献[10-11]相似，图 1(a)的结果表明，自然图像的频率主要集中在左上角低频部分，这部分的频率分量对图像的贡献最大，并随着往高频区域移动，低频对图像的贡献逐渐减小。主要的原因是因为图像中大部分相邻像素相互关联且变化平缓，因此可以用一个低频函数来接近完整的图像信息。研究发现这些对抗样本在图像域上看起来与原始图像十分接近，然而在DCT 域上与原始图像则有着明显的差别。从图 1(b)、图 1(c)、图 1(d)和图 1(e)可以明显地发现，对抗样本的频谱高频分量明显比图 1(a)增加了许多高频分量，频率从低往高发生了剧烈振荡变化，产生了明显的高频伪影。

2.3 基于 DCT 域的对抗样本检测器

针对 2.2 小节发现对抗样本在 DCT 域上存在高频伪影的问题，本文提出了一个基于 DCT 系数的训练的对抗样本检测器 CNN-DCT。相较于在图像域训练的分类器，都是对图像在空间域上的数据进行训练。而基于 DCT 系数训练的检测器，是在将所有的图像都归一化到区间[-1，1]，然后利用 2.1 小节介绍的 2D-DCT 将图像从空间域变换到 DCT 域，得到 DCT 系数后，把 DCT 系数作为输入进行训练。本文提出的对抗样本检测器 CNN-DCT 为一个简单的 8 层网络结构，如表 1 所示。

表 1 检测器网络结构

各层类型	具体参数
Conv2d	kernal_size:3×3，output:(30×30×96)
GroupNorm	(32，96)
ELU	N/A
Droupout2d	(0.2)
Conv2d	kernal_size:3×3，output:(28×28×96)
GroupNorm	(32，96)
ELU	N/A
Conv2d	kernal_size:3×3，output:(26×26×96)
GroupNorm	(32，96)
ELU	N/A
Droupout2d	(0.5)
Conv2d	kernal_size:3×3，output:(24×24×192)
GroupNorm	(32，192)
ELU	N/A
Conv2d	kernal_size:3×3，output:(22×22×192)
GroupNorm	(32，192)
ELU	N/A
Droupout2d	(0.5)
Conv2d	kernal_size:3×3，output:(20×20×256)
GroupNorm	(32，256)
ELU	N/A
Conv2d	kernal_size:1×1，output:(20×20×256)
GroupNorm	(32，192)
ELU	N/A
Conv2d	kernal_size:1×1，output:(20×20×256)
AvgPool2d	kernal_size:20×20，output:(1×1×2)

2.4 对抗攻击算法的改进

考虑到对抗样本高频伪影的存在，并且存在可能被利用来进行区分干净样本和对抗样本，一个直观的解决想法是直接在最后添加的扰动上增加一个低通滤波器，让对抗样本在频域上看起来与原始图像尽可能的相似。然而，这个简单的方法虽然减小了对抗样本和原始样本在频域上的差异，在一定程度上降低了 CNN－DCT 的检测率，但同时也增大了攻击的失败率。为此本文寻求一个改进方法，在不牺牲攻击成功率的前提下又能解决在频域上易被检测的问题。本文提出了一个改进的对抗攻击算法 IAA－DCT，可以将攻击空间限制在低频范围内，通过启发式的方法在低频空间搜索最优的扰动，从而使对抗样本在图像域和频域上都能最大程度上接近。

首先基于式(1)，本文进行了以下优化：

$$\min\ \delta(lf(r)),\ \text{subject to:}$$
$$f(x+lf(r))=y_t\ \text{ and }\ x+lf(r)\in[0,1]^m \tag{11}$$

其中 $lf(r)$ 表示的是将扰动 r 变换到 DCT 域后，去除一定的高频分量，保留了扰动 r 的低频组件，然后再通过 DCT 逆变换 IDCT 转换成图像域。具体展开如下式所示：

$$lf(r)=\text{IDCT}(\text{Mask}(\text{DCT}(r))) \tag{12}$$

本文通过对 DCT 变换后的扰动 $\text{DCT}(r)$ 应用掩模 Mask 去除扰动 r 中的高频分量。然后通过对掩模后的频率分量应用 IDCT 重构扰动。其中，掩模 $m=\{0,1\}^{d\times d}$ 是像素值分别为 0 和 1 的二维矩阵图像，掩模采用逐元素积的方式进行。

算法 1 详细描述了对抗攻击算法改进后的整个流程，通过设定预期的攻击成功率 η，保证在限定扰动 r 的频域大小同时，也能维持一定的攻击成功率。其中 ATK 表示的是在 N 个原始样本 x 上成功被攻击的样本数占所有被攻击原始样本的比例，$\text{randint}(K)$ 表示从 K 类别中随机选取一个类别。

算法 1：生成频域鲁棒的对抗样本算法

输入：原始样本 $X\in\text{R}^{N\times H\times W\times C}$，预训练的分类器 θ，预期的攻击成功率 η，掩膜 Mask

输出：频域不可见扰动 $lf(r)$，指定的标签 y_t

(1) 初始化 $lf(r)=0^{H\times W\times C}$，$\eta^{\text{best}}=\text{ATK}(X)$，$y_t=\text{randint}(K)$
(2) while $\eta^{\text{best}}<\eta$ do
(3)　　for $x_i\in X$ do
(4)　　if $\theta(x_i+lf(r))!=y_t$
(5)　　$lf(r)=\text{IDCT}(\text{Mask}(\text{DCT}(r)))$
(6)　　end if
(7)　　end for
(8)　　$X_{\text{adv}}=X+lf(r)$
(9)　　$y_t=\theta(X_{\text{adv}})$
(10)　　if $\text{ATK}(X_{\text{adv}})>\eta^{\text{best}}$
(11)　　$\eta^{\text{best}}=\text{ATK}(X_{\text{adv}})$
(12)　　$lf(r)^{\text{best}}=lf(r)$
(13)　　$y_t^{\text{best}}=y_t$
(14)　　end if
(15) end while
(16) return $lf(r)^{\text{best}}$，y_t^{best}

---

3 实验分析

3.1 实验设置

本文实验在 CIFAR－10[12] 数据集和 SVHN[13] 数据集上进行了验证。受攻击基准模型为 VGG－19[14] 和 ResNet－34[15]。实验从以下几个指标上进行观测：

(1) 被攻击后模型的鲁棒性准确率 ACC(Accuracy)；

(2) 对抗样本的检测率 AER(Adversarial Examples Rate)，AER 值越高则表示对抗样本检测率越高；

(3) 攻击成功率 ASR(Attack Success Rate)，值为成功使分类器分类错误的图像数量占全部图像总数的比例，ASR 值越高代表攻击成功率越高。

3.2 频域上的对抗样本检测

3.2.1 实验细节

在对抗样本检测上主要验证了几种经典的对抗样本，包括 FGSM[6]、BIM[7]、PGD[8] 和 C&W[9]。在 CIFAR－10 数据集上，分别利用这几种攻击方法在 CIFAR－10 数据集的训练集上各自随机生成 10 000 张对抗样本，其中 8 000 张样本用来训练，2 000 张样本用来测试。而在 SVHN 数据集上，则从训练集上随机筛选出 10 000 张生成对抗样本，用来训练和测试的比例与 CIFAR－10 数据集相同。本文使用交叉熵作为损失函数，SGD 作为优化器，其中学习率为 0.001，动量值为 0.9。

3.2.2 实验结果和总结

(1) 实验结果

本文利用 FGSM[6]、BIM[7]、PGD[8] 和 C&W[9] 这几个经典对抗攻击算法在数据 CIFAR－10 和 SVHN 上分别对 VGG－19 和 ResNet－34 进行攻击，并且利用得到的等比例混合对抗样本集的 DCT 系数训练得

到的检测器 CNN-DCT 和图像域上训练的 CNN 进行性能比较，其中 ACC 和 AER 表示的是在 DCT 域上实验得到的结果，ACC* 和 AER* 表示的是在图像域上操作的结果。结果(如表 2 所示)表明，CNN-DCT 取得了平均 97%以上的对抗样本检测率 AER，相对于在图像域训练得到的检测器 CNN 取得的 92%平均对抗样本检测率 AER* 提升了近 5%。同时发现，在面对不同网络和不同数据集时，对同一个数据集上不同网络或者同一个网络在不同数据集上进行攻击，CNN-DCT 得到的对抗样本检测结果和 ACC 相差不大。例如，FGSM 方法在数据集 CIFAR-10 和 SVHN 上分别对 VGG 和 ResNet 攻击后，通过 CNN-DCT 得到的对抗样本检测率几乎落在 97.5%左右，与 ACC 平均 98.6%的检测率相差不大。这说明改检测模型得到一个较高的假阳率，于是本文在 3.3 小节进行了检测模型的迁移性实验测试。

表 2 CNN-DCT 检测结果 (%)

		FGSM	BIM	PGD	C&W	Blend
CIFAR-10 & VGG	ACC	97.59	97.35	97.69	97.70	97.46
	AER	98.69	98.92	98.76	98.79	98.81
	ACC*	90.44	90.89	90.90	90.57	91.19
	AER*	92.74	92.70	92.74	92.67	92.70
CIFAR-10 & ResNet	ACC	97.76	97.66	97.31	97.45	97.36
	AER	98.84	99.14	98.85	98.67	98.96
	ACC*	91.24	90.67	90.60	91.41	90.89
	AER*	92.67	92.76	92.66	92.77	92.72
SVHN & VGG	ACC	97.32	91.66	98.69	97.08	96.51
	AER	98.64	93.56	99.05	98.59	97.33
	ACC*	92.90	88.98	93.72	93.57	92.16
	AER*	94.85	90.07	94.84	94.58	93.43
SVHN & ResNet	ACC	97.16	92.08	97.22	96.87	95.30
	AER	98.54	93.72	98.59	98.39	97.14
	ACC*	93.11	88.92	92.91	92.37	91.64
	AER*	94.80	90.39	94.94	94.09	93.53

(2)总结

对抗样本在图像域上看起来与干净样本几乎一致，而在频域上存在的高频伪影可以被有效利用。相比基于图像域信息训练的检测器，基于频域信息训练的检测器取得了更高更稳定的对抗样本检测率。

3.3 基于频域检测器的迁移性

(1)迁移性结果分析

为了得到 CNN-DCT 在面对新的数据集时的表现性能，本文将在 SVHN 数据集上进行对抗攻击生成的对抗样本训练得到检测器迁移到 CIFAR-10 数据集上进行检测。同时也进行了从 SVHN 数据集训练的检测器迁移到 CIFAR-10 数据集测试的验证实验。

本文分别对受攻击模型 VGG-19 和 ResNet-34 在数据集 CIFAR-10 和 SVHN 上进行迁移性性能测试。在被攻击模型 ResNet-34 上，从SVHN 数据集训练的 CNN-DCT 迁移到 CIFAR-10 数据集结果不足 80%。而从数据集 CIFAR-10 迁移到 SVHN 的检测结果都接近 90%，如表 3 和表 4 所示(其中 S 表示源数据集，T 表示目标数据集)。这对于未知模型或者未知数据集而言，是一个可观的表现结果，为以后对抗样本的检测提供了一个新奇的研究方向。

表 3 VGG-19 模型上迁移结果 (%)

S→T		FGSM	BIM	PGD	C&W	Blend
SVHN To CIFAR-10	ACC	80.17	79.94	80.87	81.01	81.25
	AER	82.85	82.44	82.85	82.56	82.75
CIFAR-10 To SVHN	ACC	88.35	86.34	89.15	88.32	87.55
	AER	90.20	88.06	90.52	90.67	89.50

表 4 ResNet-34 模型上迁移结果检测结果 (%)

S→T		FGSM	BIM	PGD	C&W	Blend
SVHN To CIFAR-10	ACC	71.02	70.78	71.22	70.87	71.23
	AER	72.33	72.03	72.33	72.65	72.26
CIFAR-10 To SVHN	ACC	87.65	86.21	89.42	87.73	87.65
	AER	89.15	87.90	90.07	89.25	89.23

(2)总结

由于高频伪影普遍存在于不同的数据集上，并且 CNN-DCT 在迁移性上表现优异，使得防御者即使在不了解攻击者攻击的数据集的情况下，可以很好地利用迁移学习进行对抗样本的检测。未来将继续展开深入的工作。

3.4 对抗攻击算法的改进评估

本文定义了攻击成功率 ASR(Attack Success Rate)，即成功使分类器在一定数量图像分类错误占全部图像的比例。为了更好地评估本文改进的算法，本文在 CIFAR-10 数据集图像域上训练了一个有 95%分类准确度的 CNN 模型作为基准模型。

如表 5 所示，ASR 和 AER 表示的是对上述基准

表 5 IAA-DCT 算法对比检测结果

(%)

	FGSM	BIM	PGD	C&W	Blend
ASR	97.89	97.30	98.26	97.83	98.55
AER	99.13	98.67	98.83	99.16	98.99
ASR*	76.35	46.70	43.29	75.38	58.95
AER*	51.62	50.19	51.09	51.46	50.97
ASR**	95.68	96.22	93.17	94.39	92.50
AER**	92.36	91.29	89.88	92.30	91.18

模型的攻击成功率以及对应对抗样本的检测率，ASR* 和 AER* 表示的是直接对扰动添加一个低通滤波后的结果，ASR** 和 AER** 表示用了本文的改进算法取得的结果。ASR 越高并且 AER 越低，证明攻击的成功率越好且不易在频域中被检测器检测到。在正常的对抗攻击下，本文可以利用频域信息很好地区分开对抗样本和干净样本，分辨度接近 100%。为了解决这个弊端，本文尝试在常规的方法攻击后方添加一个低通滤波器，将要添加的扰动筛选出一定的低频分量作为最后的扰动。虽然这个方法极大程度上降低基于频域检测器CNN-DCT 对对抗样本的正确检测率，同时也带了攻击失败的问题，它严重地降低了攻击的成功率，从 97%降到了不足 80%。为此，本文更近一步提出了另外一个优化改进算法 IAA-DCT，对抗样本的检测率从 99%降低到了 95%以下，并且还保持着 90%以上的攻击成功率。

4 结论

本文以频域的角度重新探索了对抗样本的相关工作。从对抗样本在 DCT 域上的平均图谱结果发现，即使在图像域上跟原始样本看起来完全一致的对抗样本，在 DCT 域上也表现出了与原始样本的巨大不同，普遍存在高频伪影。因此，本文以此为切入点，设计了一个基于 DCT 域信息的对抗样本检测器 CNN-DCT。结果表明，相对于直接在图像域上将对抗样本和原始样本进行分类，本文设计的检测器 CNN-DCT，在数据集 CIFAR-10 和 SVHN 上取得了近乎 98%的成功检测率，分类性能相对在图像域上训练的 CNN 有极大提升，同时在 DCT 域上能够极大程度区分开对抗样本和干净样本，从而提高训练模型的性能。迁移性实验结果表明，在未知攻击数据集的情况下，也可以利用迁移学习来进行对抗样本检测，未来将成为一个新的对抗防御方向。最后针对于上述对抗样本在频域上存在的高频伪影以及易被检测的缺陷，提出了改进算法 IAA-DCT，保证了对抗样本在图像域上的视觉一致，也让其和原始样本在频域上尽可能相似，在保持攻击成功率的同时，也极大程度上降低在频域上被检测的风险。

参考文献

[1] SZEGEDY C, ZAREMBA W, SUTSKEVER I, et al. Intriguing properties of neural networks[C]// International Conference on Learning Representations, 2014. arxiv: 1312.6199.

[2] PAPERNOT N, MCDANIEL P D, WU X, et al. Distillation as a defense to adversarial perturbations against deep neural networks[C]//IEEE Symposium on Security and Privacy(SP), 2016: 582-597.

[3] XU W L, EVANS D, QI Y J. Feature squeezing: detecting adversarial examples in deep neural Networks[C]// Network and Distributed System Security Symposium (NDSS), 2018: 1-15.

[4] FRANK J, EISENHOFER T, SCHÖNHERR L, et al. Leveraging frequency analysis for deep fake image recognition[C]//International Conference on Machine Learning(ICML), 2020: 3247-3258.

[5] ER M J, CHEN W, WU S. High-speed face recognition based on discrete cosine transform and RBF neural networks[J]. IEEE Trans. Neural Networks, 2005, 16(3): 679-691.

[6] GOODFELLOW I J, SHLENS J, SZEGEDY C. Explaining and harnessing adversarial examples[C]// International Conference on Learning Representations, 2015: 1-11.

[7] KURAKIN A, GOODFELLOW I J, BENGIO S. Adversarial examples in the physical world[C]// International Conference on Learning Representations, 2017: 1-10.

[8] MADRY A, MAKELOV A, SCHMIDT L, et al. Towards deep learning models resistant to adversarial attacks[C]// International Conference on Learning Representations, 2017. arXiv: 1706.06083.

[9] CARLINI N, WAGNER D. Adversarial examples are not easily detected: Bypassing ten detection methods[C]// In ACM Workshop on Artifificial Intelligence and Security, 2017: 3-14.

[10] BURTON G J, MOORHEAD I R. Color and spatial structure in natural scenes[J]. Applied Optics, 1987,

(下转第 76 页)

of the 2018 on Audio/Visual Emotion Challenge and Workshop, 2018: 3-13.

[10] XING X, CAI B, ZHAO Y, et al. Multi-modality hierarchical recall based on gbdts for bipolar disorder classification[C]//Proceedings of the 2018 on Audio/Visual Emotion Challenge and Workshop, 2018: 31-37.

[11] DU Z, LI W, HUANG D, et al. Bipolar disorder recognition via multi-scale discriminative audio temporal representation[C]//Proceedings of the 2018 on Audio/Visual Emotion Challenge and Workshop, 2018: 23-30.

[12] HINTON G E, SALAKHUTDINOV R R. Reducing the dimensionality of data with neural networks[J]. Science, 2006, 313(5786): 504-507.

[13] KING D E. Dlib-ml: a machine learning toolkit[J]. The Journal of Machine Learning Research, 2009(10): 1755-1758.

[14] CAO Q, SHEN L, XIE W, et al. Vggface2: a dataset for recognising faces across pose and age[C]//2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG2018). IEEE, 2018: 67-74.

[15] BARSOUM E, ZHANG C, FERRER C C, et al. Training deep networks for facial expression recognition with crowd-sourced label distribution[C]//Proceedings of the 18th ACM International Conference on Multimodal Interaction, 2016: 279-283.

(**收稿日期**: 2022-02-12)

作者简介:

穆家宝(1996-),**男**,**硕士研究生**,**主要研究方向**:**抑郁症检测、情感计算**。

(**上接第 65 页**)

26(1): 157-170.

[11] TOLHURST D, TADMOR Y, CHAO T. Amplitude spectra of natural images[J]. Ophthalmic and Physiological Optics, 1992, 12(2): 229-232.

[12] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. Comparison of regularization methods for imagenet classification with deep convolutional neural networks[C]//Conference and Workshop on Neural Information Processing Systems, 2012: 1097-1105.

[13] NETZER Y, WANG T, COATES A, et al. Reading digits in natural images with unsupervised feature learning[C]//Advances in Neural Information Processing Systems, 2011: 1-10.

[14] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[C]//International Conference on Learning Representations, 2014: 1-9.

[15] HE K, ZHANG X, REN S, et al. Deep residual Learning for image recognition[C]//Computer Vision and Pattern Recognition, 2016: 770-778.

(**收稿日期**: 2022-03-04)

作者简介:

丁烨(1987-),**男**,**博士**,**副教授**,**主要研究方向**:**大数据分析、人工智能**。

王杰(1995-),**男**,**硕士研究生**,**主要研究方向**:**人工智能**。

廖清(1988-),**通信作者**,**女**,**博士**,**教授**,**主要研究方向**:**大数据安全、人工智能安全**。E-mail: liaoqing@hit.edu.cn。

版权声明

经作者授权，本论文版权和信息网络传播权归属于《信息技术与网络安全》杂志，凡未经本刊书面同意任何机构、组织和个人不得擅自复印、汇编、翻译和进行信息网络传播。未经本刊书面同意，禁止一切互联网论文资源平台非法上传、收录本论文。

截至目前，本论文已经授权被中国期刊全文数据库（CNKI）、万方数据知识服务平台、中文科技期刊数据库（维普网）、JST 日本科技技术振兴机构数据库等数据库全文收录。

对于违反上述禁止行为并违法使用本论文的机构、组织和个人，本刊将采取一切必要法律行动来维护正当权益。

特此声明！

《信息技术与网络安全》编辑部

中国电子信息产业集团有限公司第六研究所