

基于噪声溶解的对抗样本防御方法

杨文雪^{1,2}, 吴非^{1,2}, 郭桐^{1,2}, 肖利民^{1,2}

(1.北京航空航天大学 软件开发环境国家重点实验室,北京 100191; 2.北京航空航天大学 计算机学院,北京 100191)

摘要: 深度神经网络在发展过程中暴露出的对抗攻击等安全问题逐渐引起了人们的关注和重视。然而,自对抗样本的概念提出后,针对深度神经网络的对抗攻击算法大量涌现,而深度神经网络自身的复杂性和不可解释性增大了防御攻击的难度。为了保证防御方法的普适性,以预处理方法为基本思路,同时结合对抗样本自身的特异性,提出一种新的对抗样本防御方法。考虑对抗攻击的隐蔽性和脆弱性,利用深度学习模型的鲁棒性,通过噪声溶解过程降低对抗扰动的攻击性和滤波容忍度。在滤波过程中,以对抗噪声贡献为依据自适应调整滤波范围及强度,有针对性地滤除对抗噪声,该方法不需要对现有深度学习模型进行修改和调整,且易于部署。实验结果表明,在ImageNet数据集下,该方法对经典对抗攻击方法L-BFGS、FGSM、Deepfool、JSMA及C&W的防御成功率均保持在80%以上,与JPEG图像压缩、APE-GAN以及图像分块去噪经典预处理防御方法相比,防御成功率分别提高9.25、14.86及14.32个百分点以上,具有较好的防御效果,且普适性强。

关键词: 深度神经网络;对抗样本;乘性噪声;类激活映射;自适应滤波

开放科学(资源服务)标志码(OSID):



中文引用格式:杨文雪,吴非,郭桐,等.基于噪声溶解的对抗样本防御方法[J].计算机工程,2022,48(4):158-164.

英文引用格式:YANG W X, WU F, GUO T, et al. Adversarial sample defense method based on noise dissolution[J]. Computer Engineering, 2022, 48(4): 158-164.

Adversarial Sample Defense Method Based on Noise Dissolution

YANG Wenxue^{1,2}, WU Fei^{1,2}, GUO Tong^{1,2}, XIAO Limin^{1,2}

(1.State Key Laboratory of Software Development Environment, Beihang University, Beijing 100191, China;

2.School of Computer Science and Engineering, Beihang University, Beijing 100191, China)

[Abstract] The security problems exposed in the rapid development of the Deep Neural Network(DNN) have gradually attracted our attention. However, since adversarial examples were first defined, many adversarial attacks on DNNs have been proposed, and the complexity and weak interpretability of DNNs increases their vulnerability to these attacks. To ensure the universality of our defense methods, in this paper, we propose a defense method against adversarial attacks based on the dissolution of noise. The proposed method takes pre-processing as the basic idea and combines it with the specificity of adversarial examples. Considering the stealthiness and vulnerability of adversarial attacks, we design the process of noise dissolution to destroy the aggressivity and the filtering tolerability of adversarial disturbance, taking advantage of the robustness of DNN. In the subsequent filtering process, we adaptively adjust the filtering range and intensity based on adversarial disturbance contribution and targeted filter adversarial noise. Our method is easy to deploy without modifying DNN. And the experiment results show that the defense success rate on the ImageNet dataset of our method against the classical adversarial attacks L-BFGS, FGSM, Deepfool, JSMA, and C&W is above 80%, and is 9.25, 14.86 and 14.32 percentage point higher than the classical pre-processing defense methods JPEG compression, APE-GAN, and D3, respectively. Our method has a good defense effect and strong universality.

[Key words] Deep Neural Network (DNN); adversarial examples; multiplicative noise; class activation mapping; adaptive filtering

DOI: 10.19678/j.issn.1000-3428.0061470

0 概述

近年来,随着数据规模的不断扩大和计算能力

极大提高,深度学习飞速发展,并在计算机视觉^[1]、自然语言处理^[2]等领域得到了大规模应用。在某些特定场景下,其性能已经超过了其他分类和识别算

基金项目:国家重点研发计划(2017YFB1010000);北京航空航天大学软件开发环境国家重点实验室基金(SKLSDE-2020ZX-15)。

作者简介:杨文雪(1998—),女,硕士研究生,主研方向为深度学习安全;吴非,博士研究生;郭桐,硕士研究生;肖利民,教授、博士。

收稿日期:2021-04-26 修回日期:2021-07-22 E-mail:nuomixuebing@foxmail.com

法,但其自身存在的安全问题也在发展过程中不断暴露出来。

在图像识别领域,根据深度神经网络在高维空间的线性性质,向图像添加精心设计的细微噪声生成对抗样本,这种微小的像素值改变在特征空间上的影响被层层放大,最终可以误导深度学习模型做出高置信度的错误判断。这使得基于深度学习的无人驾驶^[3]、人脸识别^[4]等应用面临严重威胁。目前对抗攻击依然是深度学习在可靠性领域应用和普及的较大阻碍。

目前多数研究主要从提高模型鲁棒性和预处理输入图像两个方面着手防御对抗攻击。文献[5]利用对抗样本进行对抗训练,通过减少深度神经网络的过拟合以提高鲁棒性。这种方法需要大量对抗样本用以训练,计算成本较高^[6],且出现新的对抗样本使该防御方法失效。文献[7]提出“蒸馏”作为对抗防御的新手段,利用“蒸馏”使深度学习模型更加平滑,提高模型的泛化性。文献[8]将标准攻击稍加修改后,成功攻破“蒸馏”防御。文献[9]提出在输出层前加入专门的网络用以除去非必要特征,以此提高深度学习网络的鲁棒性的DeepCloak防御机制。文献[10]以生成对抗网络(Generative Adversarial Network, GAN)为基础,提出以干净样本和对抗样本作为判别器和生成器输入,训练根据对抗样本构造“仿真”正常图像生成器的APE-GAN。文献[11]利用随机噪声输入训练Defense-GAN,模拟未被干扰图像的分布,并以此为依据为每个输入图像找到与之接近的不包含对抗干扰的输出,而生成对抗网络的训练和调试成为基于GAN防御方法的重点和难点。文献[12-13]提出的JPEG图像压缩和图像分块去噪方法对抗攻击的防御效果比较有限,在压缩图像的同时会降低正常样本的分类准确率。

基于对抗训练的防御方法可以提高模型对对抗噪声的容忍度和自身泛化性。但是,训练数据分布的局限性和深度学习模型的不可解释性,使得深度学习模型与理想分类模型的决策面总是存在一定差异,无法彻底消除对抗样本,导致训练投入和防御效果不成正比。利用预处理过程消除噪声对深度学习模型的干扰,实现正确识别的预处理方法相较于提高模型鲁棒性方法更为高效。但是,各类对抗攻击始终遵循尽可能减小对抗样本与原始图像差距的原则,大幅削弱了普通滤波过程对对抗噪声的敏感性,导致直接去噪的防御效果并不理想。同时,降噪过程中极易造成图像关键特征和边缘信息的丢失。

针对以上预处理防御方法的不足,本文提出一种基于噪声溶解的对抗样本防御方法。利用噪声溶解过程随机放大对抗扰动,降低对抗扰动对滤波过程的容忍程度,并将深度学习模型的识别特征应用于滤波器设计,经过区域自适应去噪过程得到平滑且可以被正确识别的去噪图像。

1 相关工作

本节将分析对抗攻击原理,并介绍实验中采用的对抗样本生成方法。

1.1 对抗攻击

对抗样本的存在性反映了深度学习模型的固有缺陷。如图1所示,分类决策面左侧的样本可以被正确识别为类别a,右侧样本识别为类别b。受到训练样本分布、规模及模型结构的限制,深度学习模型的分

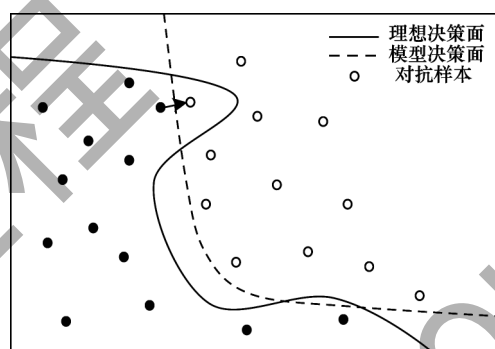


图1 对抗样本的存在性解释

Fig.1 Existence interpretation of adversarial examples

利用攻击算法计算出的细微扰动对a类样本加以修改,得到如式(1)所示的对抗样本:

$$\tilde{x} = x + \eta \quad (1)$$

对抗样本输入深度神经网络后,进行如式(2)所示的运算:

$$\omega^T \tilde{x} = \omega^T x + \omega^T \eta \quad (2)$$

其中: ω 是深度神经网络训练好的参数向量。设计 η 使其与 ω 方向一致,此时,即使 η 很小,经过多层计算后也会对激活值造成巨大的干扰,从而达到人类肉眼难以察觉但能够成功跨越模型决策面,误导神经网络将其识别为b类样本的效果。

1.2 对抗样本生成方法

对抗样本生成方法如下:

1) Box-constrained L-BFGS方法

文献[15]证明向图片中添加精心设计的微小扰动可以误导模型做出错误判断,并提出了利用Box-constrained L-BFGS最小化对抗扰动构造对抗样本的方法,如式(3)所示:

$$\text{minimize } c|r| + \text{loss}_f(x+r, l), x+r \in [0, 1]^m \quad (3)$$

其中: x 为原始图像; r 为对抗扰动; l 为目标标签。该方法求解得到的对抗扰动很小,很难被检测或清除。

2) 快速梯度符号方法

文献[16]提出的快速梯度符号方法(Fast Gradient Sign Method, FGSM)通过向模型梯度最大化的方向添加扰动生成对抗样本。在模型训练过程中,通常将损失作为衡量模型识别结果好坏的指标,损失值越小,识别正确的概率越大。反之,使损失反方向收敛,即可达

到攻击效果。对抗噪声如式(4)所示:

$$\eta = \varepsilon \text{sign}(\nabla_x J(\theta, x, y)) \quad (4)$$

其中: $J(\cdot)$ 为模型的损失函数; ∇ 为梯度; ε 为攻击步长,其大小决定了攻击强度。但由于FGSM为单次攻击,成功率不高,文献[17]提出了迭代FGSM,以小步长多次应用FGSM得到更加精准的对抗样本。

$$X_{N+1}^{\text{adv}} = \text{Clip}_{X, \varepsilon} \{X_N^{\text{adv}} + \alpha \text{sign}(\nabla_x J(X_N^{\text{adv}}, y_{\text{true}}))\} \quad (5)$$

该方法与L-BFGS相比,只需要进行反向传播梯度符号的计算,攻击效率高,但对抗性能稍差。

3) 基于雅可比矩阵的显著映射攻击

文献[18]提出的基于雅可比矩阵的显著映射攻击(Jacobian-based Saliency Map Attack, JSMA)方法不同于前几种使用损失函数梯度构造对抗样本的攻击方法,而是直接计算预测输出结果的梯度,用以代表每个输入特征对预测结果的影响并将其称为前向梯度。基于前向梯度使用雅可比矩阵构建对抗显著图,有针对性地找到对预测结果影响最大的输入特征对并修改,得到对抗样本。

4) DeepFool方法

文献[19]通过将二分类线性模型类比到复杂模型,提出了利用迭代线性计算的方法生成对抗扰动的DeepFool,如式(6)所示。该方法通过计算最短向量使原始图像朝着垂直于分类平面的方向前进最短距离,不断逼近分类平面,最终越过分类平面,实现错误分类。

$$\Delta(x; \hat{k}) = \min \|r\|_2, \hat{k}(x+r) \neq \hat{k}(x) \quad (6)$$

5) C&W方法

文献[20]提出通过限制 L_∞ 、 L_2 和 L_0 范数产生难以察觉的对抗扰动。C&W(Carlini and Wagner Attacks)攻击使用类别逻辑值代替损失函数中的最终预测值并引入了二分查找最优常数 C 来控制对抗样本的置信度,以平衡错误识别置信度和扰动添加值。C&W构造的对抗样本应满足以下2个条件:

(1)与对应的干净样本差距越小越好,如式(7)、式(8)所示:

$$r_n = \frac{1}{2} (\tan h(\omega_n) + 1) - x_n \quad (7)$$

$$\min_{\omega_n} \|r_n\| + C \cdot f\left(\frac{1}{2} (\tan h(\omega_n) + 1)\right) \quad (8)$$

(2)使得模型分类错误那类的概率越高越好,如式(9)所示:

$$f(x') = \max(\max \{Z(x'_i); i \neq t\} - Z(x'_t), -k) \quad (9)$$

其中: $f(\cdot)$ 表示目标函数; $Z(x'_i)$ 表示类别 i 的逻辑值;参数 k 用来控制错误分类的置信度,与对抗样本 x' 攻击的成功率呈正相关。

2 本文对抗样本防御方法

对抗样本是通过多次迭代像素级微小扰动而非依据语义改变得到的,在保证隐蔽性的同时具有很强的脆弱性^[21]。本文提出基于噪声溶解的对抗样本防御方法的防御流程如图2所示。

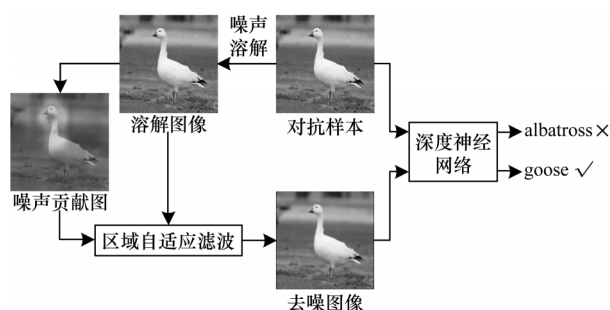


图2 本文方法防御流程

Fig.2 Defense procedure of proposed method

基于噪声溶解的对抗样本防御方法基本思路是利用基于自然噪声的噪声溶解过程,随机放大微小对抗扰动的同时溶解对抗扰动,破坏其攻击性。随后利用基于深度学习模型识别特征的区域自适应滤波有针对性地去除对抗扰动,得到可以重新被正确识别的去噪图像。

2.1 噪声溶解

如图3所示,深度学习分类模型对自然噪声有很强的鲁棒性^[22]。利用深度学习模型对自然噪声的鲁棒性,向图像中引入图像传输过程中因信道干扰而在图像上产生的乘性噪声,如式(10)所示。乘性变换随机放大了图像中的对抗扰动数值,提高了对抗扰动对后续滤波过程的敏感性,使其更容易被滤波去除。

$$g = I + n \times I \quad (10)$$

其中: I 为待处理图像; n 为均值为0的符合均匀分布的随机噪声, n 的方差用以控制乘性噪声的添加强度。

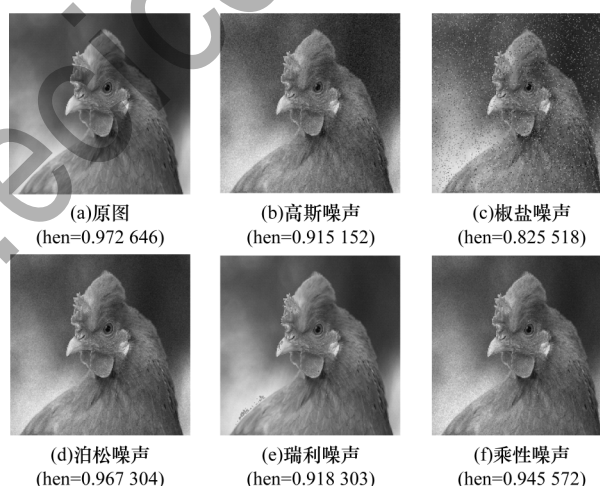


图3 自然噪声对分类结果的影响

Fig.3 Influence of natural noise on classification results

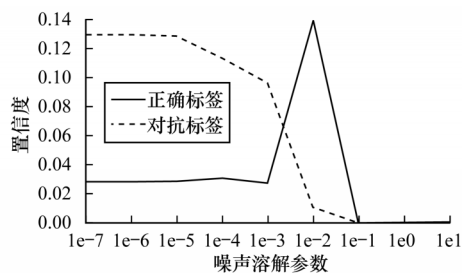
另一方面,精心设计的对抗扰动的整体性结构具有较强的脆弱性。如图4和图5所示,向对抗样本添加方差在一定范围内的乘性噪声后带来的随机共振效应^[23],提高了对抗样本被深度学习模型重新识别为正确标签的概率,即在一定程度上破坏了对抗扰动的攻击性,使对抗扰动的影响更趋于自

然噪声。同时,尽管干净样本被识别正确的概率有稍微下降,但依然可以保证图像被正确分类。因

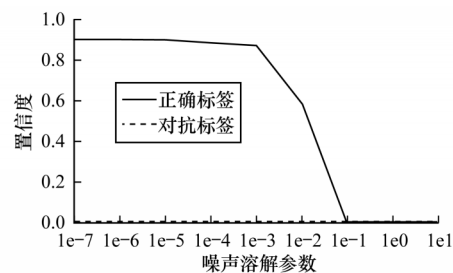
此,可以认为该过程在对抗样本修复上整体利大于弊。



(a)正确标签(spider monkey)与对抗标签(patas)



(b)对抗样本



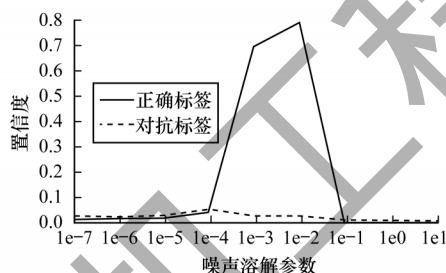
(c)原图

图4 乘性噪声对分类结果的影响1

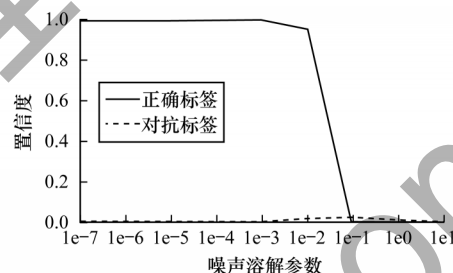
Fig.4 Influence of multiplicative noise on classification results 1



(a)正确标签(cypripediumcalceolus)与对抗标签(banana)



(b)对抗样本



(c)原图

图5 乘性噪声对分类结果的影响2

Fig.5 Influence of multiplicative noise on classification results 2

2.2 区域自适应滤波

深度学习模型本身具有很强的定位能力^[24],即对于不同的图像特征会产生不同强度的响应,添加到不同区域的对抗噪声为误导神经网络做出的贡献差别很大。图6以热力图的形式展示了图像各区域像素的修改程度,其中,第1列从上往下分别为L-BFGS、FGSM、JSMA、DeepFool、C&W方法,第2、3列分别为对抗样本和溶解图像。与图像类激活映射对比,对抗攻击对原始图像的修改主要集中在被识别图像的核心区域。因此,识别核心区域对实现对抗攻击做出的贡献更多且对抗噪声强度更大,而经过噪声溶解过程后得到的溶解图像的核心区域的变化极其微小。基于以上分析,本文提出以图像不同区域的噪声贡献为依据,自适应调节滤波强度的区域自适应滤波器,具体滤波步骤如图7所示。文献[25]提出的基于梯度加权的类激活映射(Grad-CAM)利用模型识别结果对最后一个卷积层特征图梯度的全局平均池化表征特征权重,以特征权重为依据,对特征图进行加权求和后以热力图的形式可视化特征对图像识别的影响程度,如式(11)、式(12)所示。该特征图同时反映了对抗扰动对误导神经网络做出的贡献强度,称该图为噪声贡献图。

$$\alpha_k^C = \frac{1}{Z} \sum_i \sum_j \frac{\partial y^C}{\partial A_{ij}^k} \quad (11)$$

$$L_{\text{Grad-CAM}}^C = \text{Relu} \left(\sum_k \alpha_k^C A^k \right) \quad (12)$$

其中: Z 为特征图的像素数; y^C 为 C 的分类分数; A_{ij}^k 为

第 k 个特征图在 (i,j) 处的像素值; $\text{Relu}(\cdot)$ 用于滤除未对分类为 C 做出贡献的特征。

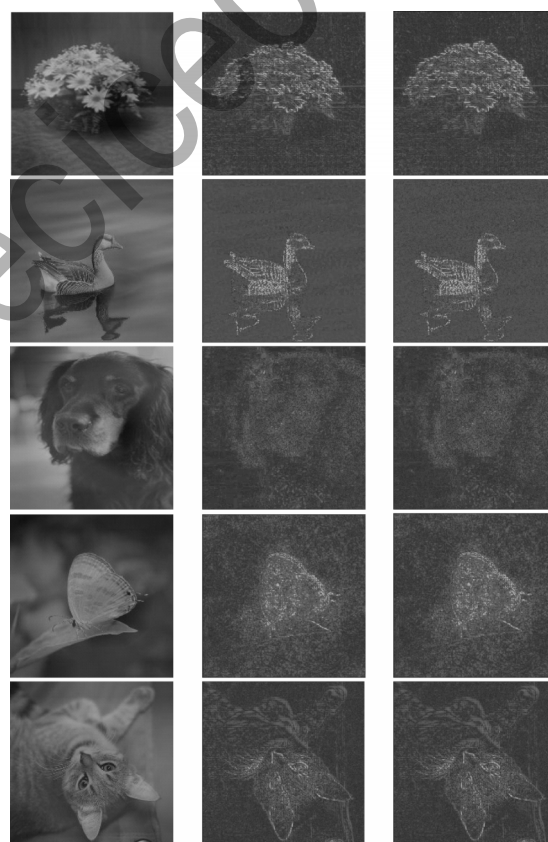


图6 噪声热力图

Fig.6 Heatmap of noise

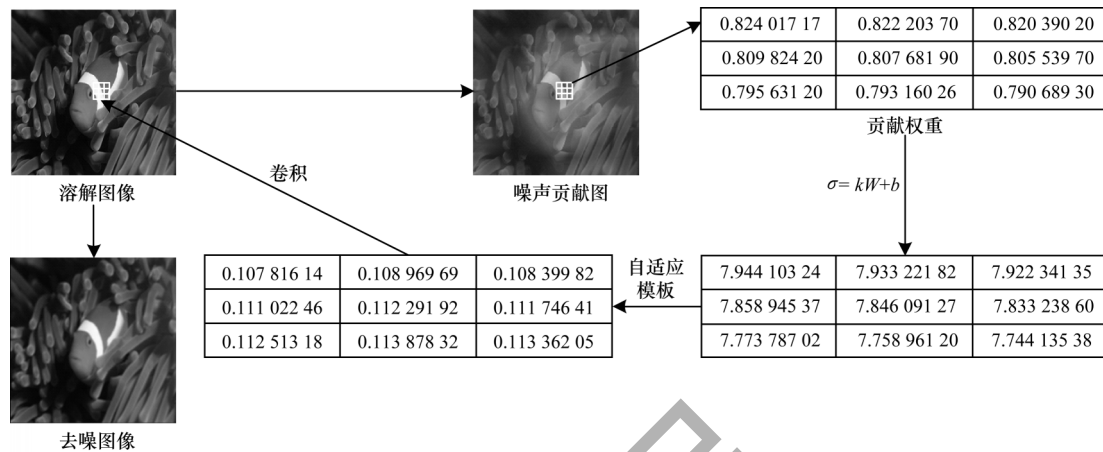


图7 区域自适应滤波

Fig.7 Regional adaptive filter

噪声贡献图每点的值表征对应像素对模型决策结果的影响,将其作为该像素的噪声贡献权重 W ,用放缩后的 W 决定各区域滤波强度。利用式(13)逐个求得图像每个像素对应的改进区域自适应滤波模板。最后,将模板与图像卷积,得到滤除噪声的去噪图像。

$$G(x,y)=\frac{1}{2\pi(kW+b)^2}e^{-\frac{x^2+y^2}{2(kW+b)^2}} \quad (13)$$

其中:参数 k 、 b 用于噪声贡献权重 W 的放缩,保证滤波强度控制在合理的范围。

3 实验结果与分析

本节将验证防御方法的有效性和可行性。首先介绍实验的数据来源、实验设置、评价指标及参数选择,然后验证本文方法对对抗攻击的防御效果及与其他预处理方法的对比。

3.1 数据来源

文献[26]提出的 ImageNet 是计算机图像识别领域最大、应用最广的自然图像数据集,包含分属于 1 000 个类的超过一百万个被手动标注的彩色识别样本。实验以 ImageNet 数据集为基准,随机选取能够被目标深度学习模型正确识别的干净图像,分别使用 L-BFGS、FGSM、Deepfool、JSMA 以及 C&W 方法以 InceptionV3^[27]、VGG16^[28]、ResNet 50^[29] 为攻击目标生成 10 000 张对抗样本。

3.2 实验设置

目前很难对不同的防御方法进行直观的比较。一方面,大部分攻击方法和防御技术主要作用于低分辨率图像,而对高分辨率图像攻击和防御的研究相对较少;另一方面,不同文献对于测试图像选择、扰动量级等实验参数设置的标准不统一,增大了各类防御方法的效果对比难度。

为了使防御效果的对比更加公正合理,对量化指标做出如下规定:1)可以被深度学习模型误判视

为攻击成功,生成的对抗样本即可用于后续实验;2)对抗样本处理后可以被正确分类的情况即视为防御有效。同时,为了避免实验结果的偶然性,进行了多次实验来计算识别准确率的平均值。

将噪声溶解线性变换过程中服从均匀分布的随机噪声的方差设置为 $1e-04$,滤波器滤波模板尺寸设置为 3×3 ,将噪声贡献放缩参数 k 设为 6, b 设为 0.5。由于噪声溶解过程具有一定的随机性,设置实验测试次数为 100 次。

用于对比的防御方法选择了普通滤波以及迁移性较高的 3 种预处理防御方法,即 JPEG 图像压缩、APE-GAN 和图像分块去噪。其中,普通滤波选择高斯滤波器,高斯核尺寸设置为 3×3 ,标准差设置为 5, JPEG 图像压缩的质量因子设置为 0.8。

实验采用的硬件配置:CPU 为 i7-7820k,内存为 16 GB DDR4,显卡为 RTX2080ti $\times 2$ 。

3.3 结果分析

预处理方法可能会造成图像细节尤其是边缘信息的丢失,从而影响识别效果。因此,本文基于噪声溶解的对抗样本防御方法对干净样本识别的影响进行评估。表 1 所示为本文防御方法在不同深度识别网络下对干净样本识别准确率的影响。

表1 干净样本识别准确率

Table 1 Identification accuracy of clean samples %

模型	准确率
Inception V3	99.36
VGG16	98.97
ResNet50	98.29

该方法使干净样本的识别准确率出现了轻微下降,但仍均保持在 98% 以上。因此,该方法在防御对抗攻击的同时极大程度上避免了图像信息的丢失及将原始干净样本分类错误的情况。

本文方法在不同的深度识别网络下对不同对抗攻击的识别准确率如表 2 所示,表中数字表示最差/均值/最优。根据 100 次实验结果绘制的箱线图如图 8 所示,用于衡量提出方法的稳定性。

表 2 不同方法的对抗样本识别准确率
Table 2 Adversarial sample recognition accuracy of different methods %

方法	Inception V3	VGG16	ResNet50
L-BFGS	88.78/89.80/91.71	81.63/82.65/85.59	84.39/85.71/86.69
FGSM	88.49/89.95/92.26	83.26/85.75/88.45	84.82/86.13/88.67
JSMA	81.98/83.05/85.15	81.25/82.61/83.06	83.31/84.42/86.50
DeepFool	89.53/91.72/93.03	88.47/89.53/91.16	90.71/93.25/94.94
C&W	90.62/92.57/93.49	84.38/85.42/86.21	89.67/90.11/92.03

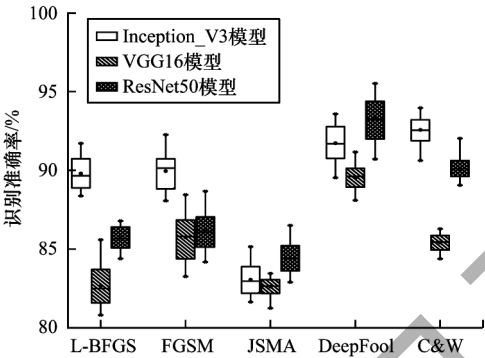


图 8 不同方法的防御效果

Fig.8 Defense effects of different methods

在 ResNet 50 模型下,面对 DeepFool 攻击,本文方法取得了最好的防御效果,识别准确率最高达到了 94.94%。而对抗扰动相对更微小的 L-BFGS 和 JSMA,对噪声溶解和去噪过程的敏感度更低,导致识别准确率略低,但也均维持在 80% 以上,达到了较优的防御效果。

表 3 所示为在 InceptionV3 模型下,本文方法与普通高斯滤波、JPEG 图像压缩、APE-GAN 和图像分块去噪 3 种预处理防御方法的防御表现对比。

表 3 不同方法的防御效果对比
Table 3 Comparison of defense effects of different methods %

方法	普通滤波	JPEG	APE-GAN	图像分块去噪	本文方法
L-BFGS	16.10	72.90	61.71	69.29	89.80
FGSM	18.52	65.63	50.91	65.31	89.95
JSMA	19.21	73.80	68.19	68.73	83.05
DeepFool	18.73	71.88	54.15	64.42	91.72
C&W	16.09	71.85	73.92	70.57	92.57

对抗噪声足够细微的性质使其对普通滤波过程不敏感,在直接进行普通滤波后,对抗样本识别准确率在 20% 以下,防御效果很差。在 FGSM 攻击下,本文方法识别准确率相比 JPEG 图像压缩、APE-GAN 和图像分区去噪分别提高了 24.32、39.04 和 24.64 个百分点。在面对其他攻击时,本文方法的防御效果相对于其他防御方法,同样取得了大幅提高。

综合以上分析,基于噪声溶解的对抗样本防御方法相对于其他预处理防御方法,将自然噪声添加

应用于预处理过程中,并将模型识别特征与滤波过程相结合,更有针对性破坏和滤除对抗噪声,达到较优的防御效果。

4 结束语

本文针对对抗扰动对预处理过程的脆弱性,提出基于噪声溶解的对抗样本防御方法。利用噪声溶解过程放大对抗扰动并破坏其攻击性,使用区域自适应滤波更有针对性的滤除噪声,改善图片质量。该方法脱离对对抗样本的依赖,普适性更强,同时无需进行对抗训练,简化了防御流程。实验结果表明,本文方法与同类预处理方法相比,可以更加有效地达到对对抗攻击的防御效果。本文在实验过程中暴露出噪声溶解过程对极微小对抗噪声的放大效果不够,下一步将根据单个图像的特征进行噪声溶解参数的自适应调整,以构建更为有效的面向多场景的对抗攻击防御方法。

参考文献

[1] LECUN Y, DENKER J S, HENDERSON D, et al. Handwritten digit recognition with a back-propagation network [C] // Proceedings of IEEE Advances in Neural Information Processing Systems. Washington D. C. , USA : IEEE Press, 1990 : 396-404.

[2] COLLOBERT R, WESTON J, BOTTOU L, et al. Natural language processing (almost) from scratch [J]. Journal of Machine Learning Research, 2011, 12 (8) : 2493-2537.

[3] SAADNA Y, BEHLOUL A. An overview of traffic sign detection and classification methods [J]. International Journal of Multimedia Information Retrieval, 2017, 6 (3) : 193-210.

[4] HUANG G B, LEE H, LEARNED-MILLER E. Learning hierarchical representations for face verification with convolutional deep belief networks [C] // Proceedings of 2012 IEEE Conference on Computer Vision and Pattern Recognition. Washington D. C. , USA : IEEE Press, 2012 : 2518-2525.

[5] ZHENG S, SONG Y, LEUNG T, et al. Improving the robustness of deep neural networks via stability training [C] // Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Washington D. C. , USA : IEEE Press, 2016 : 4480-4488.

[6] MOOSAVI-DEZFOOLI S M, FAWZI A, FAWZI O, et al. Universal adversarial perturbations [C] // Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Washington D. C. , USA : IEEE Press, 2017 : 86-94.

[7] PAPERNOT N, MCDANIEL P, WU X, et al. Distillation as a defense to adversarial perturbations against deep neural networks [C] // Proceedings of 2016 IEEE Symposium on Security and Privacy. San Jose, USA : IEEE Press, 2016 : 582-597.

[8] CARLINI N, WAGNER D. Defensive distillation is not robust to adversarial examples [EB/OL]. [2021-07-22]. <https://arxiv.org/abs/1607.04311>.

[9] GAO J, WANG B, LIN Z, et al. DeepCloak: masking deep neural network models for robustness against adversarial

- samples [EB/OL]. [2021-07-22]. <https://arxiv.org/abs/1702.06763>.
- [10] JIN G Q, SHEN S W, ZHANG D M, et al. APE-GAN: adversarial perturbation elimination with GAN [C]//Proceedings of 2019 IEEE International Conference on Acoustics, Speech and Signal Processing. Washington D. C., USA: IEEE Press, 2019: 3842-3846.
- [11] SAMANGOU EI P, KABKAB M, CHELLAPPA R. Defense-GAN: protecting classifiers against adversarial attacks using generative models [EB/OL]. [2021-07-22]. <https://arxiv.org/abs/1805.06605>.
- [12] DZIUGAITE G K, GHAMRANI Z, ROY D M. A study of the effect of JPG compression on adversarial images [EB/OL]. [2021-07-22]. <https://arxiv.org/abs/1608.00853>.
- [13] MOOSAVI-DEZFOOLI S M, SHRIVASTAVA A, TUZEL O. Divide, denoise, and defend against adversarial attacks [EB/OL]. [2021-07-22]. <https://arxiv.org/abs/1802.06806>.
- [14] TANAY T, GRIFFIN L. A boundary tilting perspective on the phenomenon of adversarial examples [EB/OL]. [2021-07-22]. <https://arxiv.org/abs/1608.07690>.
- [15] SZEGEDY C, ZAREMBA W, SUTSKEVER I, et al. Intriguing properties of neural networks [EB/OL]. [2021-07-22]. <https://arxiv.org/abs/1312.6199>.
- [16] GOODFELLOW I J, SHLENS J, SZEGEDY C. Explaining and harnessing adversarial examples [EB/OL]. [2021-07-22]. <https://arxiv.org/abs/1412.6572>.
- [17] KURAKIN A, GOODFELLOW I, BENGIO S. Adversarial examples in the physical world [EB/OL]. [2021-07-22]. <https://arxiv.org/abs/1607.02533>.
- [18] PAPERNOT N, MCDANIEL P, JHA S, et al. The limitations of deep learning in adversarial settings [C]//Proceedings of IEEE European Symposium on Security and Privacy. Berlin, Germany: Springer, 2016: 372-387.
- [19] MOOSAVI-DEZFOOLI S M, FAWZI A, FROSSARD P. DeepFool: a simple and accurate method to fool deep neural networks [C]//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA: IEEE Press, 2016: 2574-2582.
- [20] CARLINI N, WAGNER D. Towards evaluating the robustness of neural networks [C]//Proceedings of IEEE Symposium on Security and Privacy. San Jose, USA: IEEE Press, 2017: 39-57.
- [21] 李康. 人工智能系统实现中的安全风险 [C]//2018 第八届中国人工智能与安全专题论文集. 成都: [出版者不详], 2018: 231-242.
- LI K. Security risks in implementation of artificial intelligence systems [C]//Proceedings of the 8th Symposium on Artificial Intelligence and Security. Chengdu, China: [s. n.], 2018: 231-242. (in Chinese)
- [22] DIAMOND S, SITZMANN V, BOYD S, et al. Dirty pixels: optimizing image classification architectures for raw sensor data [EB/OL]. [2021-07-22]. <https://arxiv.org/abs/1701.06487>.
- [23] 杨雪, 李婷, 杨超琼, 等. 随机共振在图像处理中的研究综述 [J]. 图像与信号处理, 2015, 4(4): 132-138.
- YANG X, LI T, YANG C Q, et al. Review of research on image processing using stochastic resonance [J]. Journal of Image and Signal Processing, 2015, 4(4): 132-138. (in Chinese)
- [24] ZHOU B L, KHOSLA A, LAPEDRIZA A, et al. Learning deep features for discriminative localization [C]//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA: IEEE Press, 2016: 2921-2929.
- [25] SELVARAJU R R, COGSWELL M, DAS A, et al. Grad-CAM: visual explanations from deep networks via gradient-based localization [J]. International Journal of Computer Vision, 2020, 128(2): 336-359.
- [26] DENG J, DONG W, SOCHER R, et al. ImageNet: a large-scale hierarchical image database [C]//Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Washington D. C., USA: IEEE Press, 2009: 248-255.
- [27] SZEGEDY C, VANHOUCHE V, IOFFE S, et al. Rethinking the inception architecture for computer vision [C]//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA: IEEE Press, 2016: 2818-2826.
- [28] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition [EB/OL]. [2021-07-22]. <https://arxiv.org/abs/1409.1556>.
- [29] HE K M, ZHANG X Y, REN S Q, et al. Deep residual learning for image recognition [C]//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA: IEEE Press, 2016: 770-778.