基于 Grad-CAM 的 Mask-FGSM 对抗样本攻击

余 莉 萍
(复旦大学计算机科学技术学院 上海 201203)

摘 要 深度学习缺乏可解释性,其容易受到对抗性样本的攻击。对此引入一种深度学习可解释性模型 Grad-CAM(Gradient-weighted Class Activation Mapping),通过神经网络输入和输出之间的映射关系得到输入的热力图,结合 FGSM(Fast Gradient Sign Method)引入一种高效的算法来生成对抗样本。实验证明,该算法能够挖掘潜在的最佳攻击位置,仅需要修改 3.821% 的输入特征,就能有效生成使得神经网络错误分类的对抗样本,充分验证了该算法的高效性。

关键词 深度学习 Grad-CAM FGSM 可解释性 对抗样本

中图分类号 TP183 **文献标志码** A DOI:10.3969/j.issn.1000-386x.2022.07.030

MASK-FGSM ADVERSARIAL SAMPLES ATTACK BASED ON GRAD-CAM

Yu Liping
(*School of Computer Science and Technology, Fudan University, Shanghai 201203, China*)

Abstract Deep learning lacks the interpretability, which makes it vulnerable to adversarial samples. This paper introduced the interpretable work Grad-CAM (gradient-weighted class activation mapping) of deep learning, which obtained the heat map of the input based on the mapping relationship between the DNN input and output. Combined with fast gradient sign method (FGSM) work, an efficient algorithm was introduced to generate adversarial samples. Experiments show that this algorithm can mine the potential best attack location. And only need to modify 3.821% of the input features, it can effectively generate adversarial samples that make the neural network misclassify, which fully verifies the efficiency of the algorithm.

Keywords Deep learning Grad-CAM FGSM Interpretability Adversarial samples

0 引 言

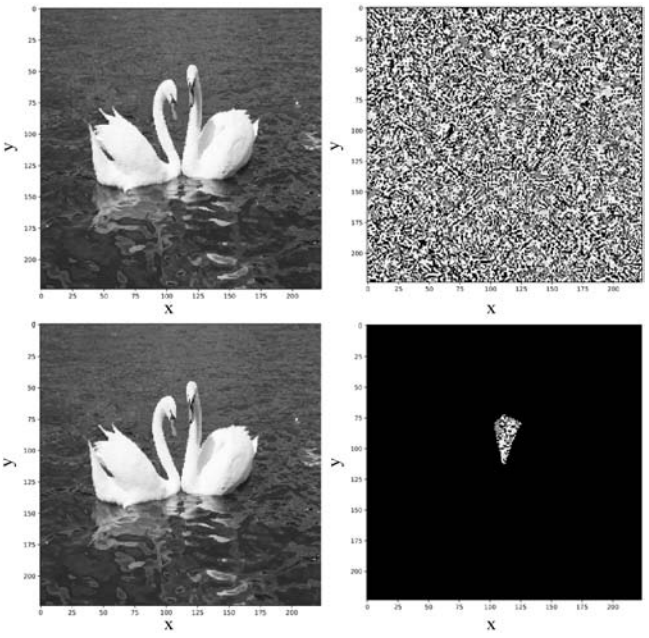
目前,基于深度学习算法的最新进展已经在很多任务上取得突破(例如图像分类[1]、自然语言处理[2]和语音处理[3]等领域)。但是,目前的方法通常以牺牲可解释性为代价来提升深度神经网络(DNN)模型的性能。如何直观地理解复杂的 DNN 的推理背后的依据具有挑战性,决策的可解释性是关键的先决条件,而简单的黑盒预测是不可信的。DNN 的另一个缺点是其固有的易受对抗性,恶意制作的样本可触发目标 DNN 失效[4-6],这将造成不可预测的模型行为并阻碍其在对安全敏感的领域中使用。在诸如自动驾驶、医疗和金融决策等高风险领域,利用深度学习进行重大决策时,往往需要知晓算法所给出结果的依据。因此,透明化深度学习的“黑盒子”,使其具有可解释性,具有重要意义。

通过提供模型级别[7-9]或实例级别[10-13]的解释,这些方法的提出推动了 DNN 可解释性领域的深入研究。这种可解释性帮助用户理解 DNN 的内部工作原理,启用包括模型验证、模型诊断、辅助分析、知识发现等领域的应用。在本文中,引入可解释性工作 Grad-CAM[14],利用 Grad-CAM 生成热力图观察输出类别和输入的内在关系,图 1(a)所示为 Grad-CAM 的结果,原始输入分类结果为“68.7% goose”,热度越高的区域,表明该输入部分对于类别导向起到越关键的作用。本文定义该热力图为注意力视图,通过在注意力更加集中的地方引入噪声,可以更高效地生成对抗样本。如

收稿日期:2020-01-15。**余莉萍**,硕士生,主研领域:人工智能与认知科学。

图 1(b)所示,第一排为原始 FGSM 的对抗样本以及叠加的噪音,对抗并未成功并且需要扰动 100% 的输入。第二排为本文方法,仅仅需要扰动 1.13% 的像素便可达到攻击目的。实验验证了本文方法能够潜在地挖掘最佳攻击位置。实验证明,本文方法平均仅需扰动 3.821% 的输入特征就能达到攻击目的。

(a)



(b)

图 1 攻击效果对比

1 相关工作

1.1 可解释性工作

可解释性和辨别力是 DNN 的两个关键方面[15]。近年来,深度学习已经成功运用在计算机视觉、语音和自然语言处理等相关的特定领域。然而,这种“黑盒”模型在“端到端”的模式下,依赖数据驱动的工作机理,缺乏解释性。研究表明,深度学习的这种模式在数据集存在偏差的情况下依然能对“biased knowledge”进行编码,从而产生决策失误[9]。因此,通过可解释性的工作来可视化隐藏在卷积神经网络(CNN)内部的知识层具有重要意义。

近年来,出现了多种方法来探索 CNN 内部隐藏的语义[16-17]。已经提出了许多统计方法[18-20]来分析 CNN 功能的特征。CNN 中滤波器的可视化[15]是探索隐藏在神经单元内部的模式的最直接方法。上卷积网络[21]将学习到的特征映射转化为图像。相比之下,基于梯度的可视化[13,22-23]生成能够使得给定单元最大化类别置信度的图像,这更接近于理解 DNN 的内部机制。Zintgraf 等[24]通过可视化对 DNN 决策贡献最大的区域从而提供视觉解释性。CAM(Class Activation Mapping)[25]利用 GAP(Global Average Pooling)的作用,保留空间信息的同时并且达到定位的目的,但是也正是由于 GAP 的限制,导致在一个新网络的结构上需重新训练模型,在实际应用中受限。Grad-CAM[14]和 CAM 的基本思路一致,区别在于获取每个特征图的权重时,采用梯度的全局平均来计算权重,该方法可以达到与 CAM 一样的可解释性效果,并且不受限于网络结构。

1.2 对抗样本生成

尽管深度学习在许多领域的任务中已经取得重大突破,但由于“黑盒”性质,很难确切知道它背后的决策依据,其在安全敏感任务中实际应用饱受质疑。恶意构造的对抗样本可以轻易让 DNN 决策产生偏差或错误[4-6]。攻击任务一般分为两类:黑盒攻击和白盒攻击。在黑盒攻击中,攻击者无法知悉模型的结构信息,只有模型的输入和输出信息[26]。Papernot 等[27]利用模型蒸馏来拟合受攻击的黑盒模型的决策结果,从而完成从黑盒模型到代理模型的知识迁移,然后利用以后的攻击方法生成对抗样本对黑盒模型进行迁移攻击。Li 等[26]在文本攻击任务中,通过观察去掉某个词前后模型决策结果的变化来定位文本中的重要单词,进而利用人类无法感知的噪音进行扰动直到达到攻击目标。白盒攻击是黑盒攻击的重要基础,在此类攻击中,攻击者可以知悉受攻击模型的结构参数等信息。Goodfello 等[28]通过计算模型输入和输出的敏感性映射(FGSM),并朝着敏感方向添加噪声来生成对抗样本。Papernot 等[29]基于雅可比图攻击(JSMA)选择最重要的特征进行攻击。

可解释性本身和攻击是一对攻防对象,可解释性为攻击者提供了对类别敏感的输入特征信息,而这一点正为进一步的研究提供攻击方向的关注焦点。本文提出一种基于 Grad-CAM 生成类别相关的热力图,在 FGSM 的基础上仅仅需要少量的噪声扰动就能达到高效的攻击。

2　基于 Grad-CAM 的对抗样本生成

2.1　Grad-CAM

CAM 可以轻松获取 DNN 结构中对于相关类别的粗定位,Grad-CAM 基于 CAM,用全局平均池化(GAP)替换全连接层,最后一个卷积层的输出通道数设置为待分类类别的个数。假设倒数第二层生成 K 个宽 μ、高 ν 的特征图 $A^k \in \mathbf{R}^{\mu\times\nu}$。这些特征图通过全局平均池化然后通过一个线性组合产生每个类别 c 的置信度得分 y^c, $y^c = \sum_k \omega_k^c \frac{1}{z}\sum_i\sum_j A_{ij}^k$, 其中 ω 为网络系数。

为了产生与类别 c 相关的定位图 $L_{\text{CAM}}^c \in \mathbf{R}^{\mu\times\nu}$, CAM 使用学习到的最后一层的权重计算最终特征图的线性组合,得到:

$$L_{\text{CAM}}^c = \sum_k \omega_k^c A^k \tag{1}$$

最后将其归一化到 0 - 1 从而达到可视化的目的。但是为了应用 CAM 需要将全连接层替换为卷积层,并重新训练网络,这是 CAM 的局限所在。

Grad-CAM:在 Grad-CAM 方法中,直接通过特征图导数来获取特征激活图,首先让类别输出 y^c 对卷积层的输出特征图 A 求导得到 $\frac{\partial y^c}{\partial A_{ij}^k}$,通过计算得到类似于 GAP 求出的权重 a_k^c:

$$a_k^c = \frac{1}{z}\sum_i\sum_j \frac{\partial y^c}{\partial A_{ij}^k} \tag{2}$$

将 a_k^c 与特征图累加得到类似于 CAM 的可视化结果,权重评估了类别 c 对特征的重要性。由于只关注特征图中的正值对分类的影响,因此需要对权重加权的结果再用 ReLU 去除负值干扰,结果为:

$$L_{\text{Grad-CAM}}^c = \text{ReLU}\left(\sum_k a_k^c A^k\right) \tag{3}$$

2.2　基于 FGSM(快速梯度下降法)的噪音图生成

快速梯度下降法。在已知模型结构的情况下,通过求模型对输入的导数,利用符号函数得到具体的梯度方向,可以得到“扰动”后的输入从而得到 FGSM 攻击下的样本。设 θ 为模型参数,x 为输入,y 为对应的标签,训练损失为 $J(\theta,x,y)$,那么叠加的噪音为:

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

式中:ε 为学习率;∇_x 表示对 x 求偏导。

2.3　基于 Mask-FGSM 的对抗样本生成

如图 2 所示,基于 Grad-CAM 可以得到对于输入图像扰动的方向,越是对于类别重要的特征,受到攻击越敏感,利用这样的结果本文算法可以对原图施以微弱的扰动,便可进行有效攻击。利用 Grad-CAM 得到输出样本的热力图,作为掩码 Mask,与 FGSM 生成的噪音图进行叠加,得到最终的对抗样本:

$$x' = x + F(S_{\text{Grad-CAM}}, P_{\text{th}}) \cdot \eta \tag{5}$$

式中:P_{th} 为施加在掩码上的阈值。

$F(S_{\text{Grad-CAM}}, P_{\text{th}})$ 的计算如式(6)所示。

$$F(S_{\text{Grad-CAM}}, P_{\text{th}}) = \begin{cases} S_{\text{Grad-CAM}} & S_{\text{Grad-CAM}_{ij}} > P_{\text{th}} \\ 0 & \text{其他} \end{cases} \tag{6}$$

式中:$S_{\text{Grad-CAM}}$ 为利用 Grad-CAM 得到输出样本的热力图。

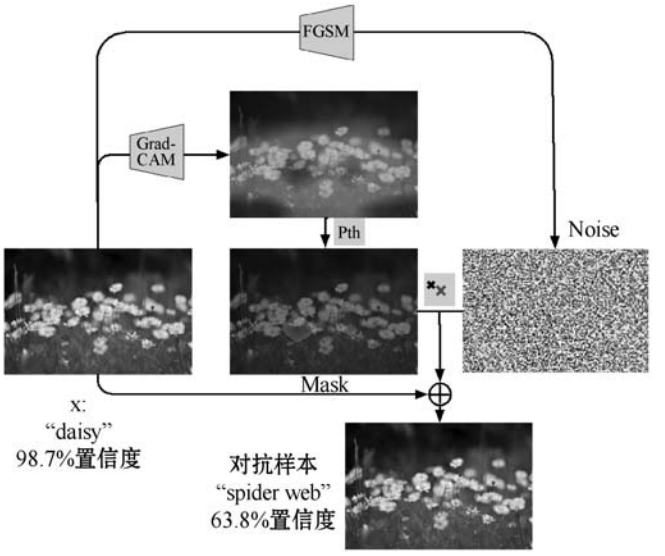


图 2　对抗样本生成

2.4　评价指标

值得注意的是,控制对抗样本和原图的最大的 L_0 距离,理论上给出任意距离下的对抗样本。

$$\|x\|_0 = \#(i) \text{ with } x_i \neq 0 \tag{7}$$

一般而言,L_0 距离越小,扰动越小。但是,本文在保证同一个 L_0 距离下,生成更符合人类视觉感知的扰动,探寻潜在高效的攻击方向。

SSIM(Structural Similarity)结构相似性是一种全参考的图像质量评价指标,它分别从亮度、对比度和结构三方面度量图像相似性。SSIM 取值范围为[0,1],值越大,表示图像失真越小。因此,本文引入图像的质量评价指标 SSIM,计算式为:

$$SSIM(X,X^*) = \frac{(2\mu_X\mu_{X^*} + C_1)(2\sigma_{XX^*} + C_2)}{(\mu_X^2 + \mu_{X^*}^2 + C_1)(\sigma_X^2 + \sigma_{X^*}^2 + C_2)} \tag{8}$$

式中:C_1、C_2 是为了避免当分母为 0 时造成的不稳定

问题引入的常数;μ_X、σ_X、μ_{X^*}、σ_{X^*} 和 σ_{XX^*} 分别是输入图像 X 的亮度均值、亮度标准差、对抗图像 X^* 的亮度均值和亮度标准差,以及它们的相关系数。

$$\sigma_{XX^*} = \frac{1}{N-1}\sum_{i=1}^{N}(X_i - \mu_X)(X_i^* - \mu_{X^*}) \tag{9}$$

原输入样本是 X,目标网络输出是 Y,F 是网络在训练期间学习的函数,η 是针对特征所做的扰动,τ 是最大扰动 L_0 距离,P_{th} 是过滤掉热力图里过小的像素值。利用算法 1 产生对抗样本。

算法 1 对抗样本生成

输入:X,Y,F,τ,P_{th}。

1. $X^* \leftarrow X$;
2. **while** $F(X^*) \neq Y$ and $\|\delta_X\|_0 < \tau$ *do*
   //$\|\delta_X\|_0$是最大扰动的 L_0 距离
3. $S = Grad_{CAM}(F(X^*), X^*, Y)$;
4. $\eta = FGSM(J(X^*, Y))$;
5. $S[S < P_{\text{th}}] = 0$;
6. 用噪声 η 修改 $X_{ij}^* = X_{ij}^* + S_{ij} \times \eta_{ij}$;
7. $\delta_X \leftarrow X^* - X$;
8. **end while**
9. **return** X^*

3 实验与结果分析

3.1 实验设置及数据集

以 Densenet161 作为模型结构,ImageNet 作为训练集。实验采集了来自 ILSVRC2014、网络图像等数据集一共 1 万幅图片作为测试集,来验证攻击效果。

3.2 验证攻击方向

为了验证本文方法的高效性,即验证 Grad-CAM 热力中心的攻击效果是否优于非热力中心。如图 3 所示,通过随机放置噪声块的位置,从而探究攻击位置与热力中心的关系。

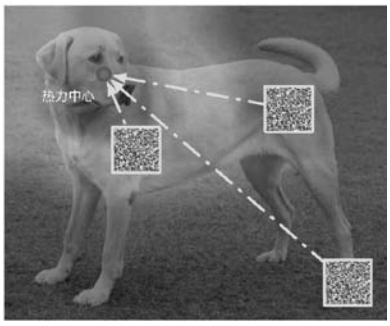


图 3 噪声块与热力中心的相对位置

通过实验验证,发现同样大小的一块噪声叠加在原图的攻击效果与距离热力中心的距离呈现图 4 所示关系,其中:虚线以上表示攻击不成功,虚线以下表示攻击成功。噪声块距离热力中心越近,则攻击效果越好,表现为模型对于错误预测的类别的置信度的绝对值越高,噪声块距离热力中心越远,则攻击效果越差,表现为模型对于正确预测的类别的置信度的绝对值越高。因此,实验验证了本文方法的高效性和有效性,该方法能挖掘潜在高效的攻击方向。

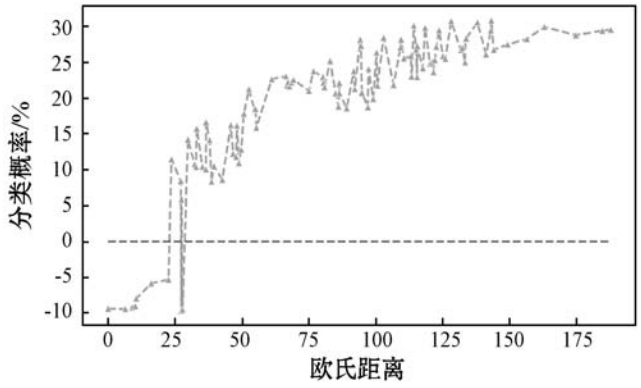
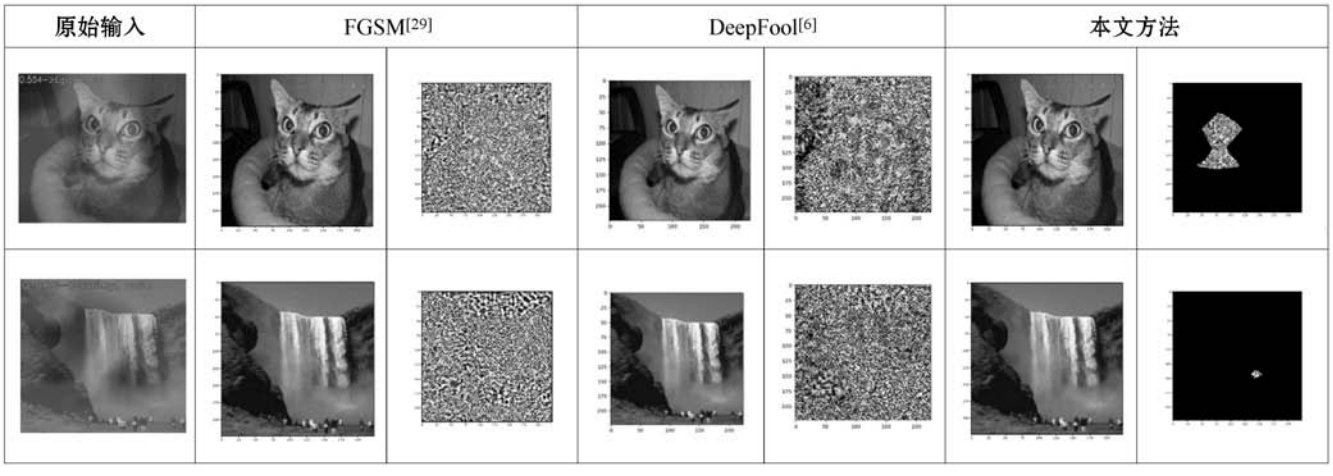


图 4 类别置信度与噪声块距离热力中心距离的关系

3.3 攻击效果

本文方法和原始 FGSM[29] 方法以及 DeepFool[6] 的生成对抗样本的实验对比如图 5 所示。从实验结果可以看出,本文方法仅仅需要扰动极为少量的输入便可以达到攻击目的。



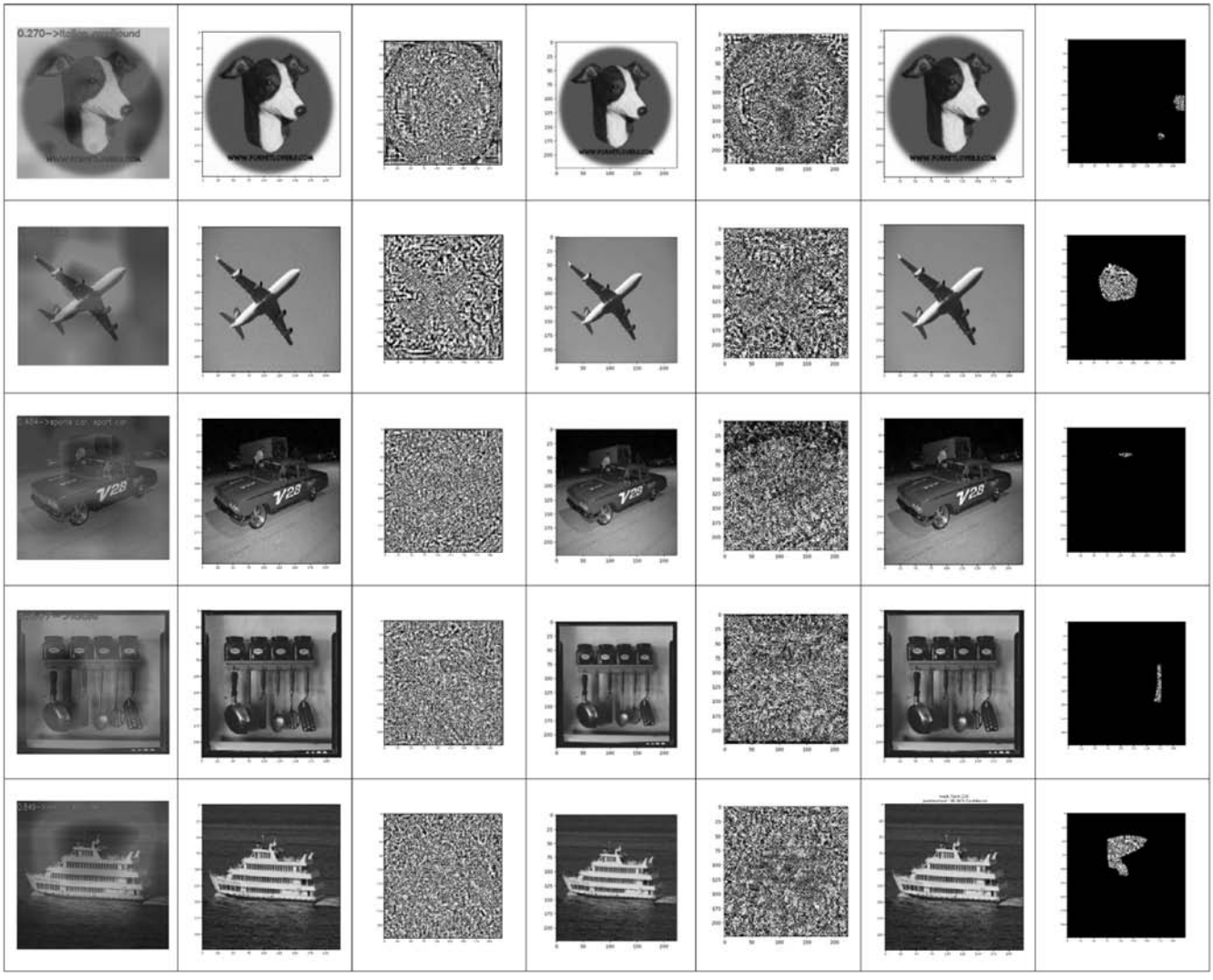


图 5　攻击效果对比(第一列:Grad-CAM 结果以及原输入分类结果;第二列第三列:原始 FGSM 攻击结果以及叠加的噪声扰动;第四列第五列:DeepFool 对抗结果以及叠加的噪声扰动;第六列第七列:本文的方法以及叠加的噪声扰动)

实验表明本文方法仅仅需要扰动极为少量的元素便可以达到攻击目的,表 1 给出了本文方法与 FGSM 以及目前典型的对抗样本攻击方法比较结果。可以看出,本文方法无论是在 L_0 距离还是 SSIM 评价指标上均取得最佳效果。

表 1　本文方法效果与经典方法对比

方法	扰动比例(L_0距离)/%	SSIM
FGSM[29]	100	0.978
JSMA[30]	4.520	0.991
DeepFool[6]	74.370	0.990
本文	3.821	0.997

4　结　语

本文引入深度学习可解释性的模型 Grad-CAM,针对深度神经网络(DNN)的结构并基于对 DNN 输入和输出之间映射的关系,结合 FGSM 方法,平均仅仅需要扰动 3.821% 的输入便可达到攻击目的。通过与目前已有的经典方法进行实验结果对比,充分验证了本文方法的高效性。本文结合了可解释性领域的成果,将其成功应用在对抗样本领域,实验结果表明本文方法效果显著,发掘了潜在的攻击方向,能够以更少的扰动成本达到攻击目的。此外,本文方法具有良好的普适性,可以进一步推广出更多的攻击思路,具有良好的应用前景。

参　考　文　献

[1] He K, Gkioxari G, Dollár P, et al. Mask R-CNN[C]// IEEE International Conference on Computer Vision, 2017: 2961 – 2969.

[2] Young T, Hazarika D, Poria S, et al. Recent trends in deep learning based natural language processing[J]. IEEE Com-

putational Intelligence Magazine,2018,13(3):55 – 75.
[3] Nicolis M, Nadolski A F. Text-to-speech processing with emphasized output audio: U. S. Patent 10319365[P]. 2019 – 06 – 11.
[4] Hayes J, Melis L, Danezis G, et al. LOGAN: Membership inference attacks against generative models[J]. Proceedings on Privacy Enhancing Technologies,2019(1):133 – 152.
[5] Nguyen A, Yosinski J, Clune J. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images[C]//IEEE Conference on Computer Vision and Pattern Recognition,2015:427 – 436.
[6] 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.
[7] Sabour S, Frosst N, Hinton G E. Dynamic routing between capsules[C]//Advances in Neural Information Processing Systems,2017:3856 – 3866.
[8] Karpathy A, Johnson J, Li F F. Visualizing and understanding recurrent networks[EB]. arXiv:1506.02078,2015.
[9] Zhang Q, Wu Y N, Zhu S C. Interpretable convolutional neural networks[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition,2018:8827 – 8836.
[10] Dabkowski P, Gal Y. Real time image saliency for black box classifiers[C]//Advances in Neural Information Processing Systems,2017:6967 – 6976.
[11] Fong R C, Vedaldi A. Interpretable explanations of black boxes by meaningful perturbation[C]//IEEE International Conference on Computer Vision,2017:3429 – 3437.
[12] Ribeiro M T, Singh S, Guestrin C. Why should i trust you?: Explaining the predictions of any classifier[C]//22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,2016:1135 – 1144.
[13] Simonyan K, Vedaldi A, Zisserman A. Deep inside convolutional networks: Visualising image classification models and saliency maps[EB]. arXiv:1312.6034,2013.
[14] Selvaraju R R, Cogswell M, Das A, et al. Grad-cam: Visual explanations from deep networks via gradient-based localization[C]//IEEE International Conference on Computer Vision,2017:618 – 626.
[15] Bau D, Zhou B, Khosla A,et al. Network dissection: Quantifying interpretability of deep visual representations[C]// IEEE Conference on Computer Vision and Pattern Recognition, 2017:6541 – 6549.
[16] Zhang Q, Cao R, Zhang S, et al. Interactively transferring CNN patterns for part localization[EB]. arXiv:1708.01783, 2017.
[17] Zhang Q, Zhu S C. Visual interpretability for deep learning: A survey[J]. Frontiers of Information Technology & Electronic Engineering,2018,19(1):27 – 39.
[18] Szegedy C, Zaremba W, Sutskever I, et al. Intriguing properties of neural networks[EB]. arXiv:1312.6199,2013.
[19] Yosinski J, Clune J, Bengio Y, et al. How transferable are features in deep neural networks? [C]//Advances in Neural Information Processing Systems,2014:3320 – 3328.
[20] Aubry M, Russell B C. Understanding deep features with computer-generated imagery[C]//IEEE International Conference on Computer Vision,2015:2875 – 2883.
[21] Dosovitskiy A, Brox T. Inverting visual representations with convolutional networks[C]//IEEE Conference on Computer Vision and Pattern Recognition,2016:4829 – 4837.
[22] Zeiler M D, Fergus R. Visualizing and understanding convolutional networks[C]//European Conference on Computer Vision,2014:818 – 833.
[23] Mahendran A, Vedaldi A. Understanding deep image representations by inverting them[C]//IEEE Conference on Computer Vision and Pattern Recognition,2015:5188 – 5196.
[24] Zintgraf L M, Cohen T S, Adel T, et al. Visualizing deep neural network decisions: Prediction difference analysis [EB]. arXiv:1702.04595,2017.
[25] Zhou B, Khosla A, Lapedriza A, et al. Learning deep features for discriminative localization[C]//IEEE Conference on Computer Vision and Pattern Recognition,2016:2921 – 2929.
[26] Li X, Ji S, Han M, et al. Adversarial examples versus cloud-based detectors: A black-box empirical study[EB]. arXiv:1901.01223,2019.
[27] Papernot N, McDaniel P, Goodfellow I, et al. Practical black-box attacks against machine learning[C]//2017 ACM on Asia Conference on Computer and Communications Security,2017: 506 – 519.
[28] Goodfellow I J, Shlens J, Szegedy C. Explaining and harnessing adversarial examples[EB]. arXiv:1412.6572,2014.
[29] 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.

(上接第 133 页)

[12] 苗夺谦,李道国. 粗糙集理论、算法与应用[M]. 北京:清华大学出版社,2008.
[13] 郭晓敏,申闫春. 基于 Unity/Vuforia 的 AR 导览系统研究[J]. 计算机仿真,2019,36(8):165 – 169.
[14] 张国梅. 基于 Vuforia SDK 开发的 AR 家居装修系统设计与开发[J]. 电脑知识与技术,2020,16(19):80 – 81.