文章编号:1671-6833(2021)01-0050-06

基于 SmsGAN 的对抗样本修复

赵俊杰[1], 王金伟[1,2]

(1.南京信息工程大学 计算机与软件学院,江苏 南京 210044;2.中国科学院信息工程研究所 信息安全国家重点实验室, 北京 100093)

摘 要: 针对对抗样本给基于深度学习的检测模型带来的严重识别干扰问题,提出一种基于随机多滤波特征统计生成对抗网络(SmsGAN)的对抗样本修复方案。采用随机多滤波特征统计网络(SmsNet)构建了特征统计层,实现了对抗样本的高精度检测,并将每个卷积核输出的特征图直接送到特征统计层获取全局特征。随机多滤波特征统计生成对抗网络(SmsGAN)以 SmsNet 为判别器,生成器采用多尺度卷积核并行结构避免棋盘效应的产生。生成器的损失函数由判别损失和引导损失两部分组成,形成目标引导生成器。对抗样本经过下采样网络获取局部统计特征,再输入 SmsGAN 得到修复的样本。结果表明:采用 SmsGAN 修复对抗样本,样本修复率达到了 91.3%,*PSNR* 平均值达到 32 以上,视觉质量好于传统信号处理方法,达到了去除对抗扰动的目的。

关键词: 深度学习; 对抗样本; 图像取证; 样本修复; 生成对抗网络

中图分类号: TP183 **文献标志码**: A **doi**:10.13705/j.issn.1671-6833.2021.01.008

0 引言

深度卷积神经网络[1]由于性能优异、使用简单,在文本语义识别[2]、图像分类[3]、目标检测[4]等方面获得了广泛的应用。然而,近年来卷积神经网络受到了越来越多的质疑,最重要的原因之一是对抗样本的存在[5]。所谓的对抗样本指的是在自然样本上添加极小的、人类难以察觉的扰动噪声,使得机器学习模型以高置信度将其识别为错误类型。对抗样本具有巨大危害性,在医疗诊断、自动驾驶、司法证明等需要严格分类的重要场景中造成的损失更大[6]。因此,对抗样本的修复是一个非常重要的研究课题。

对抗样本在机器学习模型中普遍存在,但是深度卷积神经网络中的对抗样本生成方法简单且数量众多,因此对抗样本的研究多在深度学习领域展开。常见的对抗样本生成方法可分为有目标和无目标两大类型[6]。有目标类型指的是对抗样本可使识别网络将其判别为一个固定的错误类别;无目标类型指的是对抗样本使得识别网络将其判别为一个随机的错误类别。FGSM[7]、DeepFool[8]等方法是常见的无目标攻击方法,C&W[9]方法既可以进行无目标攻击,也可以进行有目标攻击。

目前,对抗样本的修复方案包含均值替换[10]、颜色位深缩减[11]、非局部均值替换[11]等。均值替换本质上为均值滤波;颜色位深缩减指的是减少存储单个像素的存储比特位数;非局部均值替换指的是在特定区域内寻找相似的块,并用这些块的均值来替换这些块中的内容。上述方案普遍存在着一个缺陷,即它们的操作过程不针对图像内容,不随着图像内容的改变而改变,处理粗糙,导致视觉质量较差。因此,本文所提方案依据图像内容,先对对抗样本进行下采样,获取局部统计特征,再将这些特征值送入 SmsGAN 完成图像的高视觉质量修复。

1 SmsGAN 修复方案

所提方案的整体结构如图 1 所示。对抗样本首先被送入下采样网络获取最大值、最小值、均值、方差等局部统计特征,再将这些特征值送入

收稿日期:2020-08-10;**修订日期**:2020-11-07

基金项目:国家自然科学基金资助项目(61772281, 61702235, 62072250)

通信作者:王金伟(1978—),男,内蒙古呼伦贝尔人,南京信息工程大学教授,博士,主要从事多媒体取证、人工智能以及多媒体水印研究,E-mail:wjwei_2004@163.com。

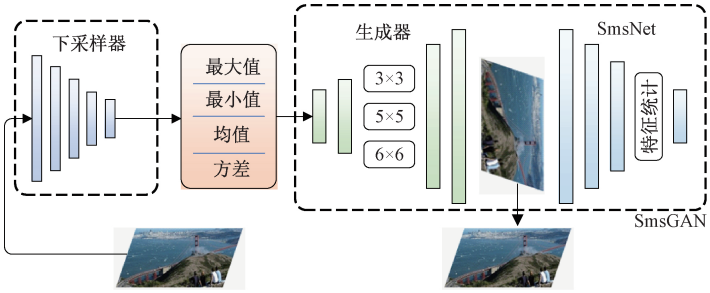


图 1　所提方案整体框架

Figure 1　The overall framework of the proposed method

SmsGAN 中进行样本修复。SmsGAN 主要包含两部分,即判别器和生成器。

1.1　判别器 SmsNet

为有效识别对抗样本,SmsNet 与普通的卷积神经网络主要有两个方面的不同:①提出一个新的神经网络组件,称为特征统计层;②网络的连接方式使用随机多滤波特征统计连接(SmsConnection)。

1.1.1　特征统计层

传统的卷积神经网络中,只有最后一个卷积层输出的特征图被送到全连接层。还有一种全卷积神经网络(FCN)[12-13],没有全连接层,但并不是主流的网络形式。全局池化层的出现减少了网络参数量,从而有效避免过拟合现象。与池化层类似,全局池化层的功能被认为是下采样。全局池化层可以分为两种类型:平均全局池化层和最大全局池化层,它们对应着不同的采样模式。从另一个角度看,均值和最大值都是数值统计特征。扩展该统计特征,得到一个新的网络组件,即特征统计层。特征统计层和全局池化层的对比如图 2 所示。在图 2 中,左侧的彩色平行四边形均表示卷积核输出的特征图。图 2(a)中下方的长方体表示全局池化层,右侧的彩色方块表示采样结果。图 2(b)中下方的长方体表示特征统计层,右侧的彩色方块表示特征统计的结果。

与全局池化层相比,特征统计层可以获得更多的统计信息。统计特征引导着卷积核的优化,多种统计特征的引导导致了卷积核的优化方向的多元化,这使得网络具有更好的检测微小扰动的能力。

1.1.2　SmsConnection 连接方式

在传统的分类网络中,只有最后一个卷积层输出的特征图被传输到全局池化层或全连接层。当网络用于取证时,并不希望网络关注图像的语义信息。卷积神经网络的语义信息由高卷积层获取[14],低卷积层基本不包含语义信息。然而,实验

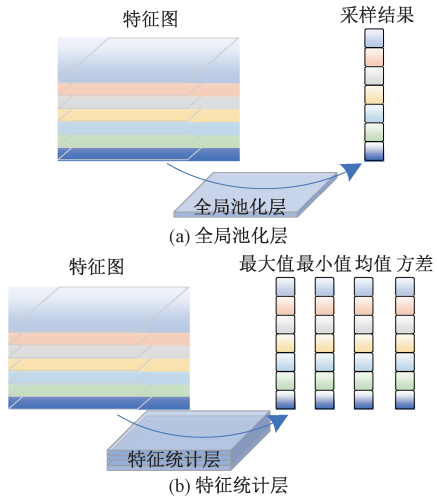


图 2　全局池化层与特征统计层

Figure 2　Global pooling layer and feature statistical layer

表明:高卷积层和低卷积层输出的特征对于取证都是有用的。为此设计了 SmsConnection 连接方式。SmsConnection 的结构如图 3 所示,每个卷积层输出的特征图都被直接送到特征统计层。这种结构打破了传统神经网络简单的高低层结构——从特征统计层往下看,每一个卷积层都是最高层。

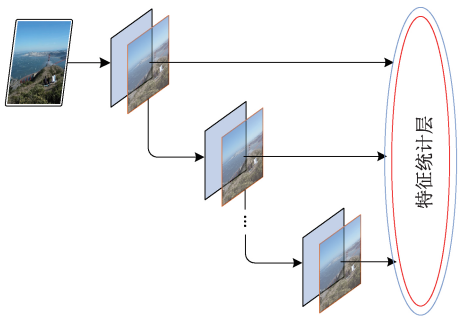


图 3　SmsConnction

Figure 3　SmsConnection

1.1.3　SmsNet 特性分析

除输出更多的特性外,SmsNet 结构上的突破带来两大特性。首先,它建立了网络宽度与深度之间的联系。卷积神经网络的宽度通常是由卷积

层中卷积核的数目决定的。这是因为在最高卷积层卷积核的数目直接决定了输入全连接层的特征数目。然而,这并不符合 SmsNet 中的情况。因此,网络的宽度被重新定义为输入特征统计层的特征值的个数。在此定义下,SmsNet 的宽度不仅与卷积核的数目成正比,而且与网络的深度成正比。相比之下,传统的深度卷积神经网络的网络宽度和深度之间没有直接联系。此外,反向传播中的梯度优化过程也发生了改变。这种变化的原因主要有两个:一是由 SmsConnection 决定,从特征统计层往下看,每个卷积层都处于最高层,这意味着从全连接层传回的梯度信息只需要求导一次就可以到达每个卷积核。求导计算次数的减少意味着更新步长变大,从而使得网络收敛速度更快。二是由特征统计层产生,尽管全局池化层相较直接摊平有很多优势,但它只能将卷积层向单特征有效方向引导。特征统计层引导卷积层向多特征有效方向优化,多方向优化又使得特征统计层获得的特征值更有效。

1.2 生成器

1.2.1 多尺度卷积核并行

样本在进入生成器前经历了下采样过程,因此必须经过上采样才能恢复原始尺寸。在 SmsGAN 中反卷积层被用来完成这项任务。然而反卷积过程也带来了一些问题,其中一个最明显的反作用是棋盘效应。棋盘效应是指反卷积过程中卷积核平移时输出块之间产生重叠从而产生阴影的现象。

SmsGAN 的生成器采用多尺度卷积并行来解决棋盘效应的问题。多尺度卷积核阴影位置相互交错,再经卷积处理,使得阴影最终被消除。如图 4 所示,3 个分支的卷积核大小分别为 3×3、5×5、6×6。在图 4 中,反卷积和卷积矩形右侧的×6 表示有连续 6 个反卷积或卷积层。3 个分支之间没有任何连接,他们输出的特征图在最高维度上连接起来,由最后一个卷积层卷积输出 3 个通道,得到最终的归一化图像。

1.2.2 目标引导生成器

GAN 通常被认为是一种典型的无监督学习网络,这是因为它的生成器没有固定的生成方向。然而,SmsGAN 的目的是修复对抗样本。因此,在损失函数中加入与原始样本之间的损失项来引导优化方向。设 $image_{\text{generated}}$ 表示生成器重建的样本;$pred_{\text{fake}}$ 表示鉴别器中重新构造的样本获得的标签;$image_{\text{real}}$ 表示原始的自然样本;

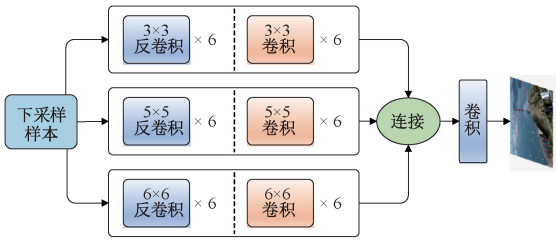


图 4 SmsGAN 生成器结构

Figure 4 The generator of SmsGAN

$label_{\text{real}}$ 表示全为 0 的标签;$F_{\text{cri}}(\ ,\ )$ 表示欧氏距离的计算。生成器的损失函数 L_{cri} 可以用式(1)表示:

$$L_{\text{cri}} = F_{\text{cri}}(pred_{\text{fake}}, label_{\text{real}}) + F_{\text{cri}}(image_{\text{generated}}, image_{\text{real}})。 \quad (1)$$

其中,L_{cri} 是两项相加的结果,前者通过 SmsNet 来衡量修复样本与自然图像的近似程度,称之为判别损失;后者直接衡量修复样本与原始样本之间的近似程度,称之为引导损失。目标引导生成器带来的好处不仅是可以决定样本生成的方向,由于 SmsGAN 需要生成器和判别器协同训练,任意部分不能正常收敛,训练过程就不能正常进行。从理论上讲,用式(1)作为损失函数是没有问题的。然而,在 SmsGAN 训练的早期,生成器很难找到正确的收敛方向,这会导致网络长时间振荡,浪费大量计算资源。为解决这一问题,当生成器损失函数的值较大时,需要对其损失值进行进一步的放大,以加速收敛过程;当损失值逐渐减小时,梯度应变化更加缓慢,以避免越过最优点,指数函数 $y=\exp x$ 当 x 越大导数越大,x 越小时导数越小,满足上述要求。此外,发现判别损失与引导损失之间采用乘法关系比加法关系更稳定。用 $\exp\{\cdot\}$ 表示指数函数的计算,最终的损失函数 $L_{\exp(\text{cri})}$ 可以用式(2)表示:

$$\begin{aligned} L_{\exp(\text{cri})} &= \exp\{F_{\text{cri}}(pred_{\text{fake}}, label_{\text{real}})\} \times \exp\{F_{\text{cri}}(image_{\text{generated}}, image_{\text{real}})\} = \\ &\exp\{F_{\text{cri}}(pred_{\text{fake}}, label_{\text{real}}) + F_{\text{cri}}(image_{\text{generated}}, image_{\text{real}})\}。 \end{aligned} \quad (2)$$

2 实验与分析

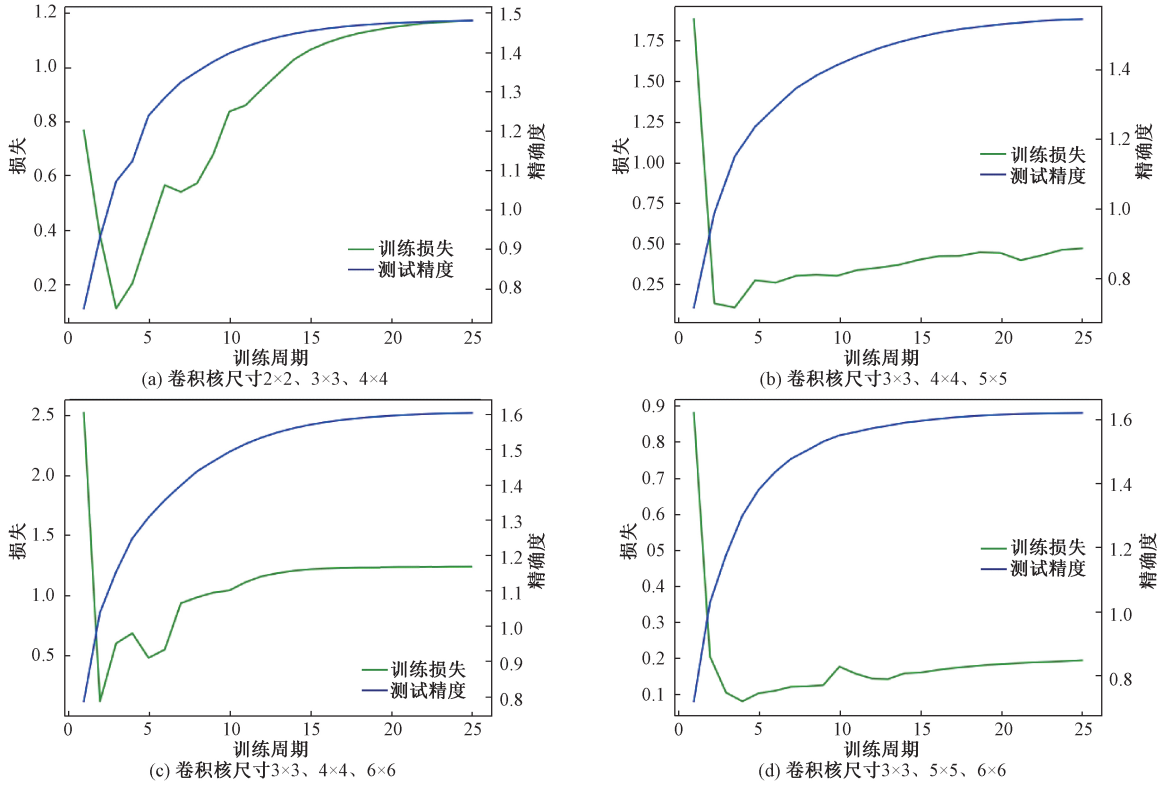
2.1 数据集和实验环境说明

数据集基于 ImageNet 2012 制作。笔者从中随机挑选了 15 000 张图片,并使用 C&W 方法生成对应的对抗样本。整个数据集共 30 000 个样本,其中 20 000 个作为训练集,10 000 个作为测试集。训练过程在 2080Ti GPU 上进行,显存为 11 G。

2.2　生成器卷积核尺寸对比

卷积核的尺寸对生成器的生成效果有很大影响。由于 SmsGAN 中的生成器具有多尺度卷积核并行结构,卷积核的大小对其影响更大。比较了由不同尺寸卷积核组成的生成器的训练效果,其训练过程如图 5 所示。

将图 5(a)的最终精确度与图 5(b)~5(d)对比可以发现,过小的卷积核使得网络拟合能力较弱。图 5(a)中生成器的损失值最先发散,而图 5(b)、图 5(d)中的损失值均未呈现明显的扩散趋势。而图 5(b)最终达到的精确度较低,这与小尺寸卷积核的可学习参数较少有关。图 5(c)的精确度接近于图 5(d),生成器的损失值略有发散现象。但是从测试精度上来说,图 5(d)所展示的结构达到了最大值。综合来说,采用尺寸为 3×3、5×5、6×6 的卷积核的生成器效果最好。



(a) 卷积核尺寸2×2、3×3、4×4　(b) 卷积核尺寸3×3、4×4、5×5

(c) 卷积核尺寸3×3、4×4、6×6　(d) 卷积核尺寸3×3、5×5、6×6

图 5　不同卷积核尺寸训练过程对比

Figure 5　Comparison of the training process of different convolution kernel sizes

值得注意的是,对于生成对抗网络来说,生成器损失值的上升并不一定意味着图像质量的下降。这是因为随着判别器训练效果的提升,生成器的损失值呈上升趋势。从图中可以看出,3×3、5×5、6×6 的卷积核尺寸组合在测试集上的精确度最高,训练后期损失值也没有大幅增加。

2.3　对比与分析

图 6 展示了原始样本与经过窗口大小分别为 5×5、7×7 的均值滤波器[10]处理后的样本,将普通 24 位色深样本压缩为 16 位的样本[11]以及 SmsGAN 修复的样本。之所以将色深压缩为 16 位,是因为图片常用存储格式除 24 色位外最高为 16 色位。对比图 6(a)~6(c)可以看出,采用均值滤波的方法去除对抗样本的对抗特性会对图像内容的细节造成较大破坏,均值滤波器尺寸越大,这种损失也越严重。对比图 6(a)、图 6(d),图像位深的缩减造成过多的细节丢失,同时颜色发生了巨大变化,对人眼视觉造成较大障碍。使用 SmsGAN 修复得到的样本则获得了较高的视觉质量,甚至在某些细节部分,例如图 6 中第 1 行图像的右下角树叶部分,细节上的表现超过了原始样本,边缘轮廓更为清晰。

表 1 展示了使用均值滤波器[10],颜色位深缩减[11]与 SmsGAN 进行对抗样本修复的对比。修复率指的是修复后正确分类的样本数量与修复总量的比。SmsGAN 修复的样本峰值信噪比(*PSNR*)高于均值滤波修复,且修复率较高。颜色位深缩减修复率最高,但 *PSNR* 值与结构相似性(*SSIM*)值均过小。综上所述,使用 SmsGAN 修复对抗样本,不仅修复率较高,且视觉质量良好。

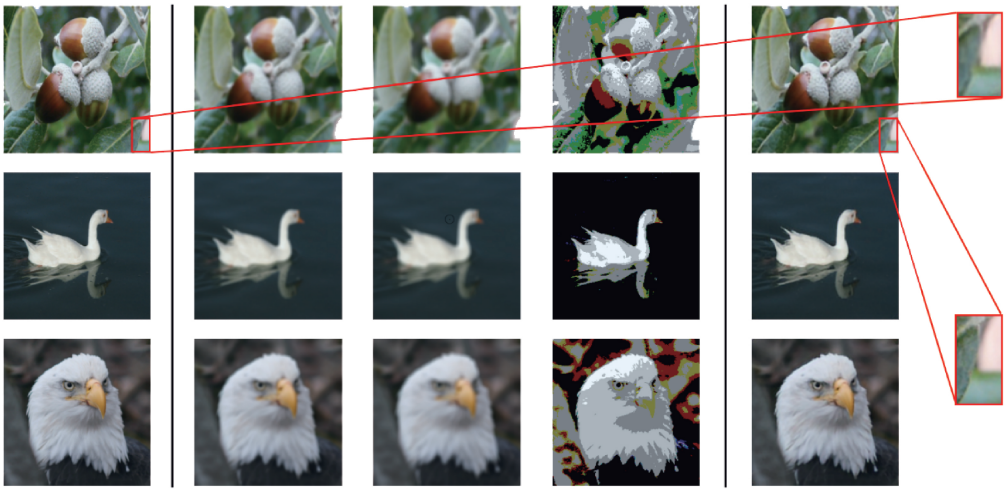


图 6 均值滤波、颜色位深缩减与 SmsGAN 修复效果对比

Figure 6 Comparison of recovery effect of mean filter, color bit depth reduction, and SmsGAN

表 1 不同修复方案结果对比

Table 1 Results of different recovery methods

修复方案	修复率/%	*PSNR*	*SSIM*
3×3 均值滤波	84.40	31.5	0.902
5×5 均值滤波	79.10	27.7	0.788
7×7 均值滤波	77.00	25.9	0.707
色深缩减	95.70	17.3	0.525
SmsGAN	91.30	32.2	0.893

3 结论

(1)提出 SmsGAN 修复方法,其判别器采用可高精确度识别对抗样本的 SmsNet,生成器采用目标引导生成器。特征统计层与 SmsConnection 的配合使得 SmsNet 能够高精确度识别对抗样本,目标引导生成器引导修复样本在内容上与原始样本匹配。

(2)使用 SmsGAN 修复对抗样本,样本修复率达到了 91.3%,*PSNR* 平均值达到 32 以上,视觉质量好于传统信号处理方法,达到了去除对抗扰动的目的。

(3)虽然对抗样本的攻击性很强,但鲁棒性较弱,常见的信号处理方式有明显的修复效果,且针对性明显,通常只对特定的网络模型有效。在下一步工作中,将进一步提升修复样本的视觉质量,并着力增强对抗样本的鲁棒性和跨网络适应性。

参考文献:

[1] 罗荣辉,袁航,钟发海,等.基于卷积神经网络的道路拥堵识别研究[J].郑州大学学报(工学版),2019,40(2):18-22.

[2] 刘发升,徐民霖,邓小鸿. 结合注意力机制和句子排序的情感分析研究[J]. 计算机工程与应用,2020,56(13):12-19.

[3] YIN Q, WANG J, LUO X, et al. Quaternion convolutional neural network for color image classification and forensics[J]. IEEE access, 2019, 7: 20293-20301.

[4] 蒋弘毅,王永娟,康锦煜. 目标检测模型及其优化方法综述[J/OL].自动化学报,2020:1-26(2020-03-03)[2020-08-01]. https://doi.org/10.16383/j.aas.c190756.DOI:10.16383/j.aas.c190756.

[5] 何英哲,胡兴波,何锦雯,等. 机器学习系统的隐私和安全问题综述[J]. 计算机研究与发展,2019,56(10):1-22.

[6] LIU J Y, HOU D D, ZHANG W M, et al. Reversible adversarial examples [EB/OL]. (2018-11-01)[2020-08-01]. https://arxiv.org/abs/1811.00189.

[7] GOODFELLOW I J, SHLENS J, SZEGEDY C, et al. Explaining and harnessing adversarial examples [EB/OL]. (2014-12-20)[2020-08-01]. https://arxiv.org/abs/1412.6572.

[8] MOOSAVI-DEZFOOLI S M, FAWZI A, FROSSARD P. DeepFool: a simple and accurate method to fool deep neural networks[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. New York:IEEE, 2016: 2574-2582.

[9] CARLINI N, WAGNER D. Towards evaluating the robustness of neural networks[C]//2017 IEEE Symposium on Security and Privacy (SP). New York:IEEE, 2017: 39-57.

[10] LI X, LI F. Adversarial examples detection in deep networks with convolutional filter statistics[C]//Proceedings of the IEEE International Conference on Com-

puter Vision. New York:IEEE, 2017: 5764-5772.
[11] XU W, EVANS D, QI Y J. Feature squeezing: detecting adversarial examples in deep neural networks[EB/OL].(2017-04-04)[2020-08-01].https://arxiv.org/abs/1704.01155.
[12] LONG J, SHELHAMER E, DARRELL T. Fully convolutional networks for semantic segmentation[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. New York: IEEE, 2015: 3431-3440.
[13] 曾安,王烈基,潘丹,等. 基于 FCN 和互信息的医学图像配准技术研究[J]. 计算机工程与应用,2020,56(18):202-208.
[14] NGUYEN H H, TIEU T N D, NGUYEN-SON H Q, et al. Modular convolutional neural network for discriminating between computer-generated images and photographic images[C]//Proceedings of the 13th International Conference on Availability, Reliability and Security.New York:ACM, 2018: 1-10.

Recovery of Adversarial Examples Based on SmsGAN

ZHAO Junjie[1], WANG Jinwei[1, 2]

(1.School of Computer and Software, Nanjing University of Information Science and Technology, Nanjing 210044, China;2.State Key Laboratory of Information Security, Institute of Information Engineering, Chinese Academy of Sciences, Beijing 100093, China)

Abstract: Due to adversarial examples′ serious interference to the detection models based on deep learning, a recovery method of adversarial examples based on stochastic multifilter statistical generative adversarial network (SmsGAN) was proposed in this work. To achieve high-precision forensics of adversarial examples, this paper proposed the feature statistical layer in the stochastic multifilter statistical network (SmsNet). The feature map output from each convolution layer was directly transferred to the feature statistical layer to get global feature values. Stochastic multifilter statistical generative adversarial network (SmsGAN) used SmsNet as its discriminator, and its generator used a multi-scale convolution kernel parallel structure to avoid checkerboard artifacts. The generator′s loss function consisted of two parts, discriminative loss and guidance loss, to form a target guidance generator. The adversarial examples entered the down-sampling network to obtain local statistical features, and then these features were sent into SmsGAN for reconstruction to get denoised examples. Using SmsGAN to recover the adversarial examples, the recovery rate reached 91.3%, and the average *PSNR* reached more than 32. The visual quality was better than the traditional signal processing method, and the purpose of removing the anti-disturbance was achieved.

Key words: deep learning; adversarial example; image forensics; example recovery; GAN(generative adversarial network)