基于逆扰动融合生成对抗网络的对抗样本防御方法

张世辉[1,2],张晓微[1],宋丹丹[1],杨永亮[1],左东旭[1]
(1. 燕山大学信息科学与工程学院,河北秦皇岛 066004; 2. 河北省计算机虚拟技术与系统集成重点实验室,河北秦皇岛 066004)

摘 要: 为了有效抵御对抗样本误导深度神经网络模型,提出一种基于逆扰动融合生成对抗网络的对抗样本防御方法(Inverse Perturbation Fusing Generative Adversarial Network,IP-GAN). 充分利用对抗样本中的对抗扰动信息,确定以逆扰动作为对抗样本防御方法的研究出发点,并从高维特征空间进行有效性分析. IP-GAN方法借鉴生成对抗网络思想,以生成器架构作为逆扰动构造模型,依据对抗样本构造相应的逆扰动用于获取重构样本,并引入深度神经网络模型指导逆扰动优化方向,最终将重构样本输入至深度神经网络模型获取正确分类结果. 实验结果表明,所构造的逆扰动可有效消除对抗扰动,辅助DNN模型正确识别并分类对抗样本,与现有最新防御方法相比,IP-GAN方法在MNIST和ImageNet数据集上防御成功率分别平均提高了0.86%和2.96%.

关键词: 对抗样本;生成对抗网络;逆扰动;对抗扰动消除;防御方法
基金项目: 中央引导地方科技发展资金(No.216Z0301G);国家自然科学基金(No.61379065);河北省自然科学基金(No.F2019203285)
中图分类号: TP391.4;TP183 **文献标识码:** A **文章编号:** 0372-2112(2023)04-0879-06
电子学报URL:http://www.ejournal.org.cn **DOI:**10.12263/DZXB.20220038

Adversarial Example Defense Method Based on Inverse Perturbation Fusing Generative Adversarial Network

ZHANG Shi-hui[1,2],ZHANG Xiao-wei[1],SONG Dan-dan[1],YANG Yong-liang[1],ZUO Dong-xu[1]
(1. *School of Information Science and Technology, Yanshan University, Qinhuangdao, Hebei* 066004, *China*;
2. *The Key Laboratory for Computer Virtual Technology and System Integration of Hebei Province, Qinhuangdao, Hebei* 066004, *China*)

Abstract: In order to effectively resist the misleading of the adversarial examples for deep neural network models, an inverse perturbation fusion generative adversarial network (IP-GAN) is proposed. This method makes full use of the adversarial perturbation information in adversarial examples, takes inverse perturbation as the starting point of the adversarial example defense method, and analyzes the effectiveness from the high-dimensional feature space. Drawing on the idea of the generative adversarial network, the generator architecture is used as a construction model to generate the corresponding inverse perturbation based on adversarial examples to obtain the reconstructed examples. Then, the deep neural network model is introduced to guide the direction of inverse perturbation optimization, and input the reconstruction examples into the deep neural network model to obtain the correct classification results. The experimental results show that the inverse perturbation constructed can eliminate adversarial perturbations effectively, and assist the DNN model to identify and classify adversarial examples correctly. Compared with the state-of-the-art defense methods, the defense success rates of the IP-GAN method on MNIST and ImageNet datasets are increased by 0.86% and 2.96%, respectively.

Key words: adversarial example; generative adversarial network; inverse perturbation; adversarial perturbation elimination; defense method
Foundation Item(s): Central Government Guided Local Funds for Science and Technology Development (No.216Z0301G); National Natural Science Foundation of China (No.61379065); Natural Science Foundation of Hebei Province (No.F2019203285)

1 引言

深度神经网络(Deep Neural Network,DNN)在图像分类、目标检测等复杂问题解决方面展现出惊人效果. 2014年,Szegedy[1]发现添加微小像素扰动的图像可误导DNN模型分类的现象,此扰动通常无法被人类视觉系统察觉,且该问题未得到相应的安全保障[2,3]. 所述

收稿日期:2022-01-05;修回日期:2022-09-06;责任编辑:李勇锋

微小像素扰动通常称为对抗扰动，添加对抗扰动的图像即为对抗样本.

现有对抗样本防御方法按照抵御方式，可分为检测性防御和鲁棒性防御两类[4]. 检测性防御在图像样本输入DNN模型之前，检测并筛除其中所包含的对抗样本，从而规避对抗样本. 相比检测性防御，鲁棒性防御通过现有技术保障DNN模型在受到对抗样本攻击时仍可以正确预测分类. 本文所提方法归属于鲁棒性防御.

2018年，Samangouei[5]提出Defense-GAN方法，采用随机噪声训练生成对抗网络[6]，通过模拟真实逆扰动来降低对抗扰动影响. 2019年，Jin[7]将对抗扰动消除定义为学习从对抗样本到真实样本间的流形映射问题，并提出APE-GAN防御方法，通过训练GAN模型来进行对抗扰动消除. 上述方法均基于GAN思想实现对抗样本的重构操作，但GAN模型学习具有高自由度，导致对抗扰动消除不佳且防御效果较差. 2020年，Hlihor[8]基于自动编码器，与DNN模型构成堆叠模型，通过图像去噪操作进行对抗样本防御. 与GAN模型相比，自动编码器在编码和解码的过程中，不可避免地造成数据损失，导致去噪后的图像可信度较低. 2021年，Chen[9]提出基于通用逆扰动的对抗攻击防御方法(Universal Inverse Perturbation Defense，UIPD)，通过对图像空间进行迭代强化来提取真实样本特征，从而生成通用逆扰动. 虽然UIPD方法的通用逆扰动可适用于整体数据集，但对于单一图像样本的针对性较弱. 2021年，Zheng[10]提出GRIP-GAN方法，通过GAN网络从真实样本中学习良性扰动，并根据输入随机噪声动态生成通用鲁棒逆扰动，最终消除潜在的对抗扰动，但该方法对较大幅度扰动的对抗样本的防御效果较差.

基于以上有关现有代表性研究工作的分析，在不增加数据样本规模和不修改DNN模型的前提下，本文借鉴生成对抗思想并引入DNN模型，通过构造逆扰动进行对抗扰动消除，从而达到防御对抗样本的目标. 本文主要贡献有如下两点.(1)给出逆扰动用于消除对抗扰动的防御思想. 充分利用对抗扰动信息，确定以消除对抗扰动作为对抗样本防御方法研究的出发点. 同时，本文从高维特征空间的角度分析其有效性.(2)提出一种基于逆扰动融合生成对抗网络的对抗样本防御方法(Inverse Perturbation Fusing Generative Adversarial Network，IP-GAN). 借鉴生成对抗网络思想，以生成器架构作为逆扰动构造模型，并引入具有分类能力的DNN模型指导逆扰动优化方向，依据对抗样本快速生成相应逆扰动，以此消除对抗扰动来防御对抗样本攻击. 所提方法有望为对抗样本防御方法研究提供新的思路.

2 总体概述

现有研究表明[11]，对抗样本的出现主要是由于DNN模型的线性特点所导致. 对于给定真实样本$\boldsymbol{x}^{\text{true}}$和对抗样本$\boldsymbol{x}^{\text{adv}}=\boldsymbol{x}^{\text{true}}+\boldsymbol{r}^{\text{adv}}$，则对抗样本$\boldsymbol{x}^{\text{adv}}$经过DNN模型神经元时的计算过程为

$$\boldsymbol{w}^{\mathrm{T}}\cdot\boldsymbol{x}^{\text{adv}}=\boldsymbol{w}^{\mathrm{T}}\cdot\boldsymbol{x}^{\text{true}}+\boldsymbol{w}^{\mathrm{T}}\cdot\boldsymbol{r}^{\text{adv}} \tag{1}$$

其中，$\boldsymbol{w}$表示权重向量，$\boldsymbol{r}^{\text{adv}}$表示对抗扰动. 由式(1)可知，对抗扰动使得激活函数$f(\boldsymbol{w}^{\mathrm{T}}\cdot\boldsymbol{x}^{\text{true}})$增加了$f(\boldsymbol{w}^{\mathrm{T}}\cdot\boldsymbol{r}^{\text{adv}})$. 即，输入样本中的微小扰动随着后向传播的进行呈增长趋势，最终导致模型误分类.

因此，本文以消除对抗扰动为研究出发点，通过构造逆扰动消除对抗样本中添加的对抗扰动，重新构造对抗样本使其正确分类，如图1所示.

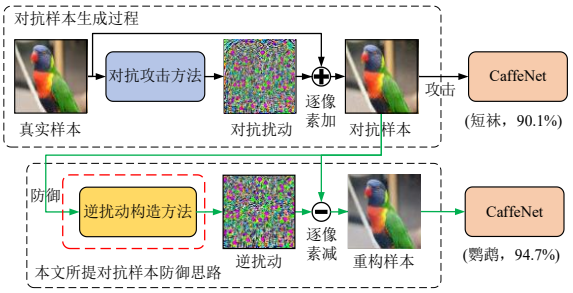

图1 逆扰动用于对抗样本防御的示意图

图1防御思路包含的优化问题形式化表示为

$$\min_{\boldsymbol{r}^{\text{IP}}} L_{\text{MSE}}\left(\boldsymbol{r}^{\text{IP}},\boldsymbol{r}^{\text{adv}}\right),\ \text{s.t.}M\left(\boldsymbol{x}^{\text{adv}}-\boldsymbol{r}^{\text{IP}}\right)=\boldsymbol{y}^{\text{true}} \tag{2}$$

其中，$\boldsymbol{r}^{\text{IP}}$表示近似对抗扰动的逆扰动，$M$表示深度神经网络模型，$L_{\text{MSE}}$表示均方误差损失.

3 基于逆扰动融合生成对抗网络的对抗样本防御方法

3.1 问题定义

设$(\boldsymbol{x}^{\text{true}},\boldsymbol{y}^{\text{true}})$为真实样本分布$P_{\text{data}}$中一组数据样本，$\boldsymbol{X}$为真实样本$\boldsymbol{x}^{\text{true}}$所组成的集合，$\boldsymbol{Y}$是真实样本$\boldsymbol{x}^{\text{true}}$对应标签$\boldsymbol{y}^{\text{true}}$所组成的集合. 对抗样本$\boldsymbol{x}^{\text{adv}}$是通过在真实样本上添加对抗扰动$\boldsymbol{r}^{\text{adv}}$所获得的可误导DNN模型$M:\boldsymbol{X}\rightarrow\boldsymbol{Y}$的数据样本，即$M(\boldsymbol{x}^{\text{adv}})\neq M(\boldsymbol{x}^{\text{true}})$.

给定真实样本集合$(\boldsymbol{X},\boldsymbol{Y})\sim P_{\text{data}}$，对抗样本生成方法将真实样本修改为对抗样本，不改变DNN模型，采用有效对抗样本防御方法，使得DNN模型准确分类对抗样本，即$M(\boldsymbol{x}^{\text{adv}})=M(\boldsymbol{x}^{\text{true}})$.

3.2 有效性分析

本节从高维特征空间出发，分析逆扰动用于对抗样本防御任务的有效性. 该防御过程不对DNN模型结构或参数进行修改，即决策边界固定不变，如图2所示.

对于DNN模型(以二分类器为例)，真实样本被分

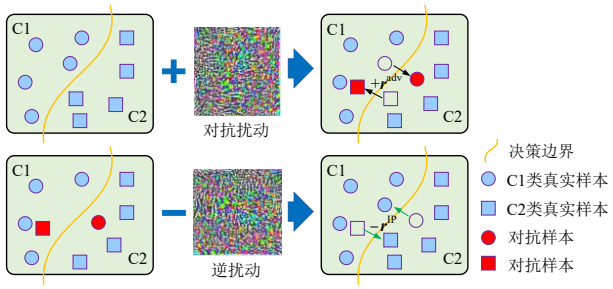

图2　基于高维特征空间的逆扰动有效性分析示意图

类为C1类(图2中决策边界左侧)和C2类(图2中决策边界右侧). 当对真实样本添加精心设计的对抗扰动后,真实样本从正确决策空间被引导至错误决策空间(图2右上所示),即对抗样本成功攻击DNN模型. 所述攻击过程形式化描述为

$$M(\boldsymbol{x}^{\text{true}}+\boldsymbol{r}^{\text{adv}})\neq M(\boldsymbol{x}^{\text{true}}) \tag{3}$$

依据逆扰动,针对对抗样本进行相应的对抗扰动消除,该样本便引导回到正确决策空间(图2右下所示),从而获得正确预测结果,即成功防御对抗样本. 所述防御过程形式化描述为

$$M(\boldsymbol{x}^{\text{adv}}-\boldsymbol{r}^{\text{IP}})=M(\boldsymbol{x}^{\text{true}}) \tag{4}$$

综上可知,在高维特征空间中,经逆扰动处理所得重构样本$\boldsymbol{x}^{\text{RE}}=\boldsymbol{x}^{\text{adv}}-\boldsymbol{r}^{\text{IP}}$可被引导回正确决策空间,即逆扰动用于对抗样本防御任务是有效的.

3.3　逆扰动构造方法的确定

逆扰动构造方法主要借鉴于生成对抗网络,通过生成器模型G^{IPCM}模拟对抗扰动分布,由此获得重构样本,如式(5)所示.

$$\boldsymbol{x}^{\text{RE}}=\boldsymbol{x}^{\text{adv}}-G^{\text{IPCM}}\left(\boldsymbol{x}^{\text{adv}}\right) \tag{5}$$

即逆扰动具体表示为$\boldsymbol{r}^{\text{IP}}=G^{\text{IPCM}}\left(\boldsymbol{x}^{\text{adv}}\right)$.

逆扰动构造模型训练过程如图3所示. 生成器模型G^{IPCM}作为逆扰动构造模型,用于模拟对抗扰动分布,获得重构对抗样本所需逆扰动;判别器模型D的输入分别为重构样本和真实样本,通过判断输入样本是否来自真实样本分布,指导生成器模型G^{IPCM}构造有效逆扰动;DNN模型M对重构样本进行预测分类,引导生成器模型G^{IPCM}在正确类别方向构造有效逆扰动.

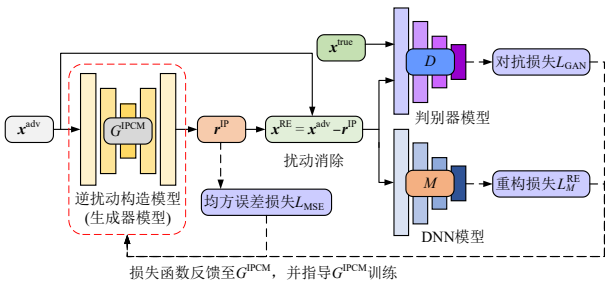

图3　逆扰动构造模型训练过程

图3所示训练过程需要对两个网络模型进行参数优化,即逆扰动构造模型G^{IPCM}和判别器模型D. 由于DNN模型M已具备成熟的样本预测分类能力,因此训练过程不包含M的参数优化.

(1) 对于G^{IPCM}的损失函数设计为

$$L_{G^{\text{IPCM}}}=\alpha L_{\text{GAN}}+\beta L_M^{\text{RE}}+\gamma L_{\text{MSE}} \tag{6}$$

其中,α、β、γ为超参数,用于控制各个损失项的占比权重;L_{GAN}为对抗损失,表示重构样本对D的欺骗能力;L_M^{RE}为重构损失,表示重构样本可被M正确分类的能力;L_{MSE}为均方误差损失,表示逆扰动与对抗扰动之间的相似程度.

(a) 经G^{IPCM}所获取的重构样本$\boldsymbol{x}^{\text{RE}}$期望误导$D$,使其认为$\boldsymbol{x}^{\text{RE}}$来自真实样本分布,即$D(\boldsymbol{x}^{\text{RE}})=1$. 针对此优化目标,对抗损失$L_{\text{GAN}}$表示为

$$L_{\text{GAN}}=E_{\boldsymbol{x}^{\text{adv}}\sim P_{\text{fake}}}\left[\log\left(1-D\left(\boldsymbol{x}^{\text{RE}}\right)\right)\right] \tag{7}$$

其中,P_{fake}为对抗样本分布. 由式(7)可知,当D对于重构样本$\boldsymbol{x}^{\text{RE}}$的预测分类结果为1,即$D$判定$\boldsymbol{x}^{\text{RE}}$来自真实样本分布时,$L_{\text{GAN}}$趋于最优值,表示经$G^{\text{IPCM}}$所得重构样本近似于真实样本.

(b) IP-GAN方法的主要目标是通过G^{IPCM}辅助M正确分类对抗样本,即$M(\boldsymbol{x}^{\text{adv}}-\boldsymbol{r}^{\text{IP}})=\boldsymbol{y}^{\text{true}}$. 因此,在训练过程中引入具有分类能力的DNN模型M,通过对重构样本$\boldsymbol{x}^{\text{RE}}$进行预测分类来指导$G^{\text{IPCM}}$的训练过程,重构损失$L_M^{\text{RE}}$表示为

$$L_M^{\text{RE}}=E_{\boldsymbol{x}^{\text{adv}}\sim P_{\text{fake}}}\left[L_{\text{ce}}\left(M\left(\boldsymbol{x}^{\text{RE}}\right),\boldsymbol{y}^{\text{true}}\right)\right] \tag{8}$$

其中,$L_{\text{ce}}(\cdot)$表示交叉熵损失. 经过最小化L_M^{RE}可令$\boldsymbol{x}^{\text{RE}}$被$M$正确分类,由此引导逆扰动优化方向.

(c) 在保证重构样本$\boldsymbol{x}^{\text{RE}}$可被DNN模型正确分类的基础上,本文期望最大程度地兼顾重构样本的视觉感知效果,IP-GAN方法将约束逆扰动修改幅度作为优化目标,尽可能地近似对抗扰动. 均方误差损失L_{MSE}借鉴MSE数学公式并定义为

$$L_{\text{MSE}}=\frac{1}{C\times W\times H}\sum_{c=1}^{C}\sum_{w=1}^{W}\sum_{h=1}^{H}\left[\boldsymbol{r}^{\text{IP}}(c,w,h)-\boldsymbol{r}^{\text{adv}}(c,w,h)\right]^2 \tag{9}$$

其中,C、W、H分别表示扰动图像的图层数、宽、高,(c,w,h)表示像素点坐标. L_{MSE}引导G^{IPCM}构造的逆扰动近似对抗样本中分布的对抗扰动,进而保障重构样本与真实样本之间的相似程度.

(2) 对于判别器模型D.损失函数表示为

$$L_D=E_{\boldsymbol{x}^{\text{true}}\sim P_{\text{data}}}\log D\left(\boldsymbol{x}^{\text{true}}\right)+E_{\boldsymbol{x}^{\text{adv}}\sim P_{\text{fake}}}\log\left[1-D\left(\boldsymbol{x}^{\text{RE}}\right)\right] \tag{10}$$

通过求解min-max问题获得用于对抗样本防御任务的逆扰动构造模型G^{IPCM},如式(11)所示

$$\arg\min_{G^{\text{IPCM}}}\max_{D}\left(L_{G^{\text{IPCM}}}+L_D\right) \tag{11}$$

3.4 IP-GAN方法用于对抗样本防御

首先,根据3.3节所述方法及损失函数训练逆扰动构造模型G^{IPCM};其次,将对抗样本输入至G^{IPCM},得到逆扰动;再次,根据逆扰动进行对抗扰动消除,得到近似真实样本的重构样本;最后,重构样本输入至DNN模型,得到与真实样本相同的预测分类结果. 所述防御流程如图4所示.

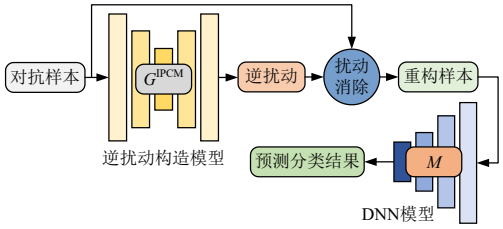

图4 IP-GAN方法的对抗样本防御流程

4 实验及分析

4.1 实验基本设置

(1)实验环境及数据集

实验环境:Intel Xeon(R) Gold-5118@2.30 GHz(CPU),NVIDIA RTX 2080Ti(GPU),Ubuntu 14.04.5(OS),CUDA9.0.176,Python2.7.13,Pytorch1.3.1.

不失一般性,本文基于对抗样本防御领域常用的MNIST和ImageNet数据集进行实验分析. 随机挑选ImageNet中10个类别的图像,每个类别1 300张,70%作为训练集,30%作为测试集.

(2)实验评价指标

本文采用分类准确率(Classification Acuracy,ACC)衡量DNN模型对于真实样本的分类能力;攻击成功率(Attack Succes Rate,ASR)衡量对抗样本生成方法对于DNN模型的欺骗能力;防御成功率(Defense Success Rate,DSR)衡量对抗样本防御方法对于对抗样本的防御能力;结构相似性(SSIM)衡量不同样本之间的相似程度.

$$\begin{cases}\mathrm{ACC}=n_{\mathrm{right}}^{\mathrm{true}}/n^{\mathrm{true}}\times 100\%\\ \mathrm{ASR}=n_{\mathrm{fake}}^{\mathrm{adv}}/n_{\mathrm{right}}^{\mathrm{true}}\times 100\%\\ \mathrm{DSR}=n_{\mathrm{right}}^{\mathrm{adv}}/n_{\mathrm{fake}}^{\mathrm{adv}}\times 100\%\end{cases}\tag{12}$$

其中,n^{true}表示真实样本数量,$n_{\mathrm{right}}^{\mathrm{true}}$表示分类正确的真实样本数量,$n_{\mathrm{fake}}^{\mathrm{adv}}$表示成功欺骗DNN模型的对抗样本数量,$n_{\mathrm{right}}^{\mathrm{adv}}$表示防御方法作用下分类正确的对抗样本数量.

4.2 逆扰动的有效性验证

基于MNIST数据集,图5给出验证IP-GAN方法的对抗扰动消除实验结果,对抗样本生成方法为PGD-L_∞,目标网络模型为LeNet. IP-GAN方法构造的逆扰动可有效消除对抗扰动,且重构样本与真实样本的SSIM值均位于0.9以上,即结构相似度显著提高. 同时,LeNet模型对于重构样本以平均98.27%的高置信度进行正确类别分类.

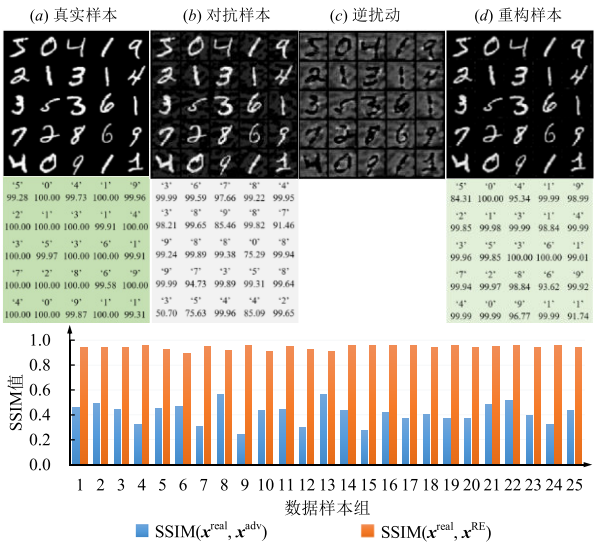

图5 IP-GAN方法的对抗扰动消除实验结果

基于MNIST和ImageNet数据集,图6给出不同扰动阈值下相关实验结果. 实验结果表明,经IP-GAN方法所得重构样本中像素扰动显著减少,且对MNIST和ImageNet数据集可分别保持在80%和70%以上的防御成功率.

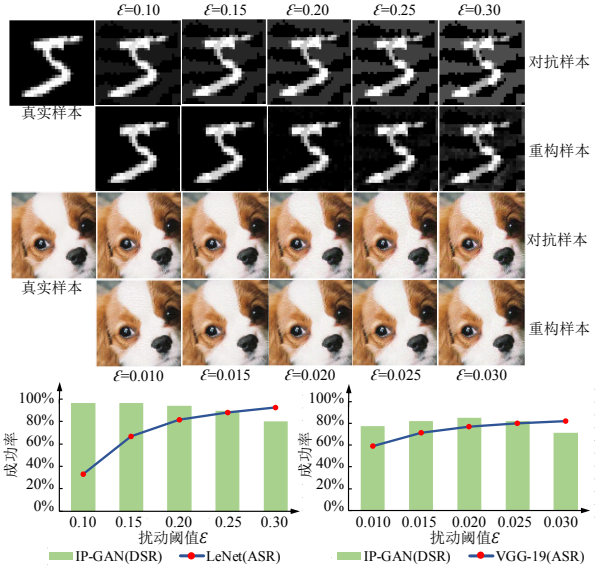

图6 IP-GAN方法对于扰动幅度变化的实验结果

4.3 不同防御方法的防御成功率比较

基于MNIST和ImageNet数据集,IP-GAN方法与现有代表性防御方法的防御成功率(DSR)对比实验结果如表1所示,最优值加粗显示. 本文所提IP-GAN方法的DSR值均高于已有防御方法.

IP-GAN方法通过训练逆扰动构造模型,附加于DNN模型之前进行对抗样本的重构操作. 与同样需要训练附加网络的Defense-GAN、APE-GAN和DAE方法相比,IP-GAN方法的防御成功率显著提高. 这是由于IP-GAN方法在训练过程中引入目标网络指导逆扰动构造的正确方向,从而避免了生成器模型学习自由度高的问题.

UIPD方法与IP-GAN方法的研究出发点都是基于逆扰动. UIPD方法通过迭代强化方法针对所有样本构造通用逆扰动. 相比UIPD方法的通用性,IP-GAN方法更侧重于特定性表达,通过生成器模型针对特定样本构造特定逆扰动,使得IP-GAN方法构造的逆扰动针对性更强.

GRIP-GAN方法通过向真实样本中注入随机噪声,训练生成器模型构造通用鲁棒逆扰动. 相比随机噪声,IP-GAN方法直接采用对抗样本训练生成器模型,更有效地学习了真实条件下对抗样本中的噪声分布,消除对抗扰动使其近似真实样本分布. 由表1可知,IP-GAN方法在MNIST数据集中对基于梯度和基于优化的对抗样本的DSR指标分别平均提高了0.55%和1.16%,在ImageNet数据集中对基于梯度和基于优化的对抗样本的DSR指标分别平均提高了3.16%和2.76%.

表1　不同防御方法针对基于梯度的对抗样本生成方法的防御成功率比较结果

对抗样本生成方法	对抗样本防御方法	MNIST			ImageNet		
		AlexNet	M_CNN[9]	LeNet	ResNet-50	VGG-19	Inc-V3
基于梯度	Defense-GAN[5]	72.40%	70.31%	68.26%	40.87%	51.04%	38.01%
	APE-GAN[7]	83.40%	82.36%	80.71%	54.10%	57.88%	51.80%
	DAE[8]	84.54%	85.68%	85.25%	65.94%	59.31%	64.54%
	UIPD[9]	88.92%	87.45%	86.89%	—	59.91%	—
	GRIP-GAN[10]	97.10%	98.25%	97.79%	—	86.35%	93.88%
	Ours	**97.83%**	**98.91%**	**98.04%**	**86.48%**	**87.23%**	**95.32%**
基于优化	Defense-GAN[5]	82.43%	86.13%	80.34%	42.08%	43.10%	41.96%
	APE-GAN[7]	82.46%	85.14%	85.01%	73.90%	49.28%	70.90%
	DAE[8]	83.66%	86.88%	84.17%	69.41%	51.61%	63.25%
	UIPD[9]	87.92%	87.54%	85.22%	—	52.91%	—
	GRIP-GAN[10]	89.34%	93.58%	92.69%	—	94.38%	86.73%
	Ours	**95.76%**	**94.77%**	**94.57%**	**97.99%**	**96.37%**	**90.25%**

5　结论

本文提出一种基于逆扰动融合生成对抗网络的对抗样本防御方法(IP-GAN). 该方法以附加网络的形式进行对抗样本的扰动消除操作,不需要基于大量数据样本,也不需要修改深度神经网络模型. 实验结果表明,所提方法构造的逆扰动可有效消除对抗扰动,提高对抗样本与真实样本之间的结构相似度. 与现有代表性对抗样本防御方法相比,IP-GAN方法的防御成功率得到有效提高. 后续研究工作将对损失函数作进一步优化或集成一种检测性防御方法加强IP-GAN方法对于真实样本的处理效果.

参考文献

[1] SZEGEDY C, ZAREMBA W, SUTSKEVER I, et al. Intriguing properties of neural networks[C]//Proceedings of the International Conference on Learning Representations. Banff: ICLR, 2014: 1-10.

[2] IRFAN M M, ALI S, YAQOOB I, et al. Towards deep learning: A review on adversarial attacks[C]//2021 International Conference on Artificial Intelligence. Islamabad: IEEE, 2021: 91-96.

[3] 邹军华, 段晔鑫, 任传伦, 等. 基于噪声初始化、Adam-Nesterov方法和准双曲动量方法的对抗样本生成方法[J]. 电子学报, 2022, 50(1): 207-216.
ZOU Jun-hua, DUAN Ye-xin, REN Chuan-lun, et al. Perturbation initialization, Adam-Nesterov and quasi-hyperbolic momentum for adversarial examples[J]. Acta Electronica Sinica, 2022, 50(1): 207-216. (in Chinese)

[4] ZHANG J L, LI C. Adversarial examples: Opportunities and challenges[J]. IEEE Transactions on Neural Networks and Learning Systems, 2020, 31(7): 2578-2593.

[5] SAMANGOUEI P, KABKAB M, CHELLAPA R. Defense-GAN: Protecting classifiers against adversarial attacks using generative models[C]//Proceedings of the International Conference on Learning Representations. Vancouver: ICLR, 2018: 1-12.

[6] GOODFELLOW I, POUGET-ABADIE J, MIRZA M, et al. Generative adversarial nets[J].Advances in Neural Information Processing Systems, 2014, 3: 2672-2680.

[7] JIN G Q, SHEN S W, ZHANG D M, et al. APE-GAN: Adversarial perturbation elimination with GAN[C]//Proceedings of the IEEE Conference on International Conference on Acoustics, Speech, and Signal Processing. Brighton: IEEE, 2019: 3842-3846.

[8] HLIHOR P, VOLPI R, MALAGÒ L. Evaluating the robustness of defense mechanisms based on autoencoder reconstructions against Carlini-Wagner adversarial attacks [C]//Proceedings of the Northern Lights Deep Learning Workshop. UiT The Arctic University of Norway: Septentrio Academic Publishing, 2020: 1-6.

[9] 陈晋音, 吴长安, 郑海斌, 等. 基于通用逆扰动的对抗攻击防御方法[J/OL]. 自动化学报, 2021: 1-20. DOI: 10.16383/j.aas.c201077.
CHEN JIN-YIN, WU CHANG-AN, ZHENG HAI-BIN, et al. Universal inverse perturbation defense against adversarial attacks[J/OL]. Acta Automatica Sinica, 2021: 1-20. DOI:10.16383/j.aas.c201077. (in Chinese)

[10] ZHENG H B, CHEN J Y, HANG D, et al. GRIP-GAN: An attack-free defense through general robust inverse perturbation[J/OL]. IEEE Transactions on Dependable and Secure Computing, 2021: 1-18. DOI:10.1109/TDSC.2021.3124337.

[11] KURAKIN A, GOODFELLOW I J, BENGIO S. Adversarial examples in the physical world[C]//Proceeding of the International Conference on Learning Representations. Toulon: ICLR, 2019: 1-13.

作者简介

张世辉 男, 1973 年 12 月生, 河北赞皇人, 燕山大学信息科学与工程学院教授, 博士生导师. 主要研究方向为视觉信息处理、模式识别.
E-mail: sshhzz@ysu.edu.cn

张晓微(通讯作者) 男, 1997 年 8 月生, 河北邢台人, 燕山大学硕士研究生. 主要研究方向为对抗样本攻防和计算机视觉.
E-mail: xwzhang0724@163.com