•人工智能安全•

DOI:10.15961/j.jsuese.202100165

本刊网刊

基于生成对抗网络的对抗样本集成防御

曹天杰[1,2], 余志坤[1,2], 祁韵妍[1,2], 杨　睿[1,2*], 张凤荣[1,2], 陈秀清[3]

(1.中国矿业大学 教育部矿山数字化工程研究中心, 江苏 徐州 221116; 2.中国矿业大学 计算机科学与技术学院, 江苏 徐州 221116;
3.徐州医科大学 医学信息与工程学院, 江苏 徐州 221004)

摘　要:针对现有对抗样本防御方法防御能力不足、时间消耗过高等问题,参考生成对抗网络与集成学习在对抗样本研究中的优势,本文提出一种基于生成对抗网络的对抗样本集成防御方法。该方法使用生成对抗网络训练多个能够消除对抗样本表面对抗扰动的生成器,使用集成学习方法将多个生成器进行集成作为最终的防御。该方法的生成对抗网络由生成器和判别器组成。生成器以对抗样本作为输入,其目的是消除对抗样本表面的对抗扰动;判别器以良性样本与消除对抗扰动后的样本作为输入,其目的是区分输入的样本;生成器与判别器交替训练,当判别器无法对输入的样本做出区分时,生成器达到最佳状态。集成防御使用平均法作为集成策略,通过平均多个生成器的防御结果,取长补短,提升单个防御的能力;通过预训练生成器来降低防御的时间消耗,通过集成多个生成器来提升单个生成器的防御能力。分别在MNIST数据集与CIFAR10数据集上,用本文的集成防御方法与其他防御方法对常见的对抗样本进行防御,以分类准确率作为评价防御能力的指标,并记录防御的时间消耗。实验结果表明,本文方法能以较低的时间消耗防御多种对抗样本,并且防御能力比已有的防御方法更好。

关键词:对抗样本;对抗样本防御;推理模型;生成对抗网络

中图分类号:TP391.4　　**文献标志码**:A　　**文章编号**:2096-3246(2022)02-0056-09

Ensemble Adversarial Example Defense Based on Generative Adversarial Network

CAO Tianjie[1,2], *YU Zhikun*[1,2], *QI Yunyan*[1,2], *YANG Rui*[1,2*], *ZHANG Fengrong*[1,2], *CHEN Xiuqing*[3]

(1.Mine Digitization Eng. Research Center of Ministry of Education, China Univ. of Mining and Technol., Xuzhou 221116, China;
2.School of Computer Sci. and Technol., China Univ. of Mining and Technol., Xuzhou 221116, China;
3.School of Medicine Info. and Eng., Xuzhou Medical Univ., Xuzhou 221004, China)

Abstract: Given the bottlenecks of existing adversarial example defense schemes, such as insufficient defense capability and high time consumption, an ensemble adversarial example defense scheme based on the generative adversarial network was proposed in this paper, by taking the advantages of the generative adversarial network and the ensemble learning in adversarial example research. In the scheme, a generative adversarial network was used to train multiple generators that can eliminate adversarial perturbations on the surfaces of adversarial examples, and the ensemble learning was used to integrate multiple generators as the final defense. The generative adversarial network was composed of generator and discriminator. While the generator takes adversarial examples as inputs and its purpose is to eliminate adversarial perturbations on the surface of adversarial examples, the discriminator takes benign examples and examples after eliminating the adversarial perturbations as inputs and its purpose is to distinguish them. The generator and discriminator were trained alternately, and the generator reaches to its best when the discriminator cannot distinguish them. The averaging method was adopted by the integration defense adopts as the integration strategy to learn from each other. Furthermore, the ability of a single defense is improved by averaging the defense results of multiple generators. The time consumption of defense was reduced by pre-training generators and the defense ability was improved by integrating multiple generators. Finally, the time consumption and

收稿日期:2021 – 02 – 26

基金项目:中国博士后科学基金项目(2020T130098ZX);江苏省博士后科研计划项目(1701061B);国家自然科学基金项目(61972400)

作者简介:曹天杰(1967—),男,教授. 研究方向:人工智能安全. E-mail: tjcao@cumt.edu.cn

* **通信作者**:杨　睿, E-mail: 2119344620@qq.com

网络出版时间:2022 – 03 – 14 10 : 53 : 03　　**网络出版地址**:https://kns.cnki.net/kcms/detail/51.1773.TB.20220312.1243.001.html

defense ability of the proposed scheme was verified on the MNIST and CIFAR10 dataset. With the classification accuracy as the evaluation index, the defense ability of the proposed scheme on six kinds of adversarial examples was verified, and compared with seven existing defense schemes. Results showed that the proposed scheme can defend against multiple adversarial examples with very low time consumption, and its defense ability is better than the existing defense schemes.

Key words: adversarial example; adversarial example defense; inference model; generative adversarial network

对抗样本是指在原始良性样本上通过添加细微的干扰所形成的恶意样本，导致推理模型以高置信度输出一个错误的结果。对抗样本的存在给人工智能模型的实际应用带来了潜在的安全威胁，例如：攻击者恶意篡改交通标志停止路牌，使得自动驾驶汽车将其识别成前进，造成交通事故；通过面部伪装，欺骗政府部门或是公司的人脸识别安全系统，侵入其内部，窃取机密等。因此，在推进人工智能模型部署的同时，迫切需要研究如何消除对抗样本的影响。

在对抗样本的防御研究中，主要包括两个方面。一方面，在样本输入推理模型之前，检测对抗样本；另一方面，通过提升模型自身的鲁棒性消除对抗样本的影响。基于检测的防御主要分为对抗样本分类器[1]、基于统计分析[2]、基于密度和不确定性预测[3]、基于修改损失和基于重建损失[4]。基于检测的防御的主要瓶颈在于不能有效检测出未知的对抗样本。另外，Athalye等[5]指出当前的基于检测的防御很难有效地对良性样本与对抗样本做出区分。基于提升模型鲁棒性的防御主要分为基于数据增强[6]、基于正则化[7]、基于随机化[8]和基于输入变换[9]。基于数据增强与正则化的防御需要重新训练推理模型，因此，这两类基于提升模型鲁棒性的防御方法时间消耗较高，且会降低推理模型对良性样本的分类准确率。基于随机化的防御主要利用推理模型或者输入样本的不确定性，其主要瓶颈在于不能有效消除对抗样本的影响。基于输入变换的防御主要是在样本输入推理模型之前进行一个预处理操作，其主要瓶颈也在于不能有效消除对抗样本的影响。因此，针对现有对抗样本防御方法防御能力不足、时间消耗过高等问题，迫切需要提出一种具有较低时间消耗且能有效防御多种类型对抗样本的防御方法。

生成对抗网络[10]是一种无监督生成模型，一些研究成果[11−12]已经将其应用到对抗样本的防御中。Kabkab等[11]提出了Defense–GAN防御对抗样本，该方法通过将靠近原始对抗样本的新的良性样本作为推理模型的输入来消除对抗样本的影响，其实验结果表明，Defense–GAN在单通道灰度图像上具有较好的表现，但在三通道彩色图像上并不能消除对抗样本的影响。Jin等[12]提出了APE–GAN来消除对抗样本的影响，该方法通过重建良性样本来消除对抗样本表面的对抗扰动，其实验结果表明，无论是单通道灰度图像还是三通道彩色图像，APE–GAN都能很好地消除对抗样本的影响，但APE–GAN的训练过程是不稳定的。集成学习常被用于提升推理任务的表现，将多个单推理模型进行集成以提升其在任务上的性能。目前，已有一些研究成果[13−14]将集成学习应用于对抗样本防御中，通过集成多个防御提升其性能，例如：Wei等[13]指出，不同的颜色空间能检测到图像数据某些自身明确的特征，因此，在同一推理模型中采用多个颜色空间来生成特征图；Gowda等[14]通过将基于不同输入转换的模型集成与不同的输出验证模型集成相结合来增强防御能力。

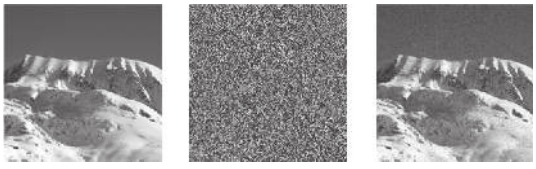
鉴于现有对抗样本防御方法存在的不足及生成对抗网络与集成学习在对抗样本防御中的表现，本文提出一种基于生成对抗网络的对抗样本集成防御方法。该方法是一种基于预处理的方法，通过提前训练生成器降低时间消耗，通过集成多个生成器提升防御能力。其生成对抗网络由生成器与判别器组成，生成对抗网络自身的损失采用WGAN–GP[15]中的损失函数以确保训练过程的稳定。生成器以对抗样本作为输入，其目的是通过重建良性样本消除对抗样本表面的对抗扰动；判别器以良性样本与重建的良性样本作为输入，其目的是对样本做出区分。生成器与判别器交替训练，相互博弈，当判别器无法对样本做出区分时，训练过程就达到了纳什平衡点。生成器的损失函数包括最小平方误差损失与生成对抗网络自身的损失，判别器的损失函数仅包括生成对抗网络自身的损失。在MNIST与CIFAR10数据集上验证了本文方法的性能。结果表明，本文集成防御方法能有效防御多种对抗样本，并且，具有较低的时间消耗。

1　相关基础

1.1　对抗样本

Szegedy等[16]发现了对抗样本的存在，并提出了对抗样本的概念。对抗样本是指可以使目标推理模型出错，人眼却能够正确推理的样本。图1是对抗样本的示例[17]，在推理模型以94.39%置信度推理为雪山的图片（图1（a））上添加一个人眼不可见的对抗扰动（图1（b）），生成对抗样本（图1（c））。虽然，人眼看到的对抗样本仍然是雪山，但是，目标推理模型会以99.99%的置信度将对抗样本推理为狗。将良性样本定义为x；对抗样本定义为x'；目标推理模型定义为f；

损失函数定义为$L(f(x),y)$，在分类任务中，损失函数通常为交叉熵；y为正确分类的类别；对抗扰动定义为$p_{\text{adv}}=x'-x$，通常使用L_P范数来量化对抗扰动$p_{\text{adv}}=\mathrm{L}_P(x',x)$，其中$P=0,1,2,\cdots,\infty$；对抗样本$x'$满足$\mathrm{L}_P(x',x)<\varepsilon\wedge f(x)\neq f(x')$，其中，$\varepsilon$为一个自定义的常数，用于控制对抗扰动$p_{\text{adv}}$的大小。



雪山：94.39% 置信度　　狗：99.99% 置信度

(a) 良性样本　(b) 对抗扰动　(c) 对抗样本

图 1　对抗样本示例[17]

Fig. 1　Instance of adversarial example [17]

1.2　对抗样本生成

对抗样本的生成方法包括白盒算法和黑盒算法。白盒算法分为基于梯度优化的方法和基于约束优化的方法。基于梯度优化的方法的代表是Goodfellow等[18]提出的FGSM。除此以外，基于梯度优化的常见方法还有BIM[19]、MI–FGSM[17]、DeepFool[20]、JSMA[21]等。其中，BIM为FGSM的改进方法，MI–FGSM为BIM的改进方法。基于约束优化的方法的代表是Szegedy等[16]提出的L–BFGS。除此以外，基于约束优化的常见方法还有L–BFGS的改进方法C&W[22]。黑盒算法分为基于搜索的方法[23]、基于进化算法的方法[24]、基于梯度估计的方法[25]和基于决策边界估计的方法[26]。黑盒算法与白盒算法相比，生成过程需要消耗大量的计算资源，并且攻击目标推理模型的成功率较低。实验中仅对白盒算法进行防御评估，因此，下面介绍实验中使用的白盒算法。

1.2.1　FGSM

FGSM[18]被用于快速找到给定良性样本x的对抗扰动p_{adv}的方向，从而使目标推理模型f的损失函数值$L(f(x),y)$增加，降低推理的置信度。虽然不能保证增加一定数量的损失函数值就会导致目标模型推理出错误的结果，但这仍然是一个合理的方向。FGSM通过计算损失函数$L(f(x),y)$相对于良性样本x的梯度$\nabla_x L(f(x),y)$，并通过一个自定义的常数ε乘以梯度$\nabla_x L(f(x),y)$的符号函数$S_{\text{sign}}=\text{sign}(\nabla_x L(f(x),y))$来产生对抗扰动$p_{\text{adv}}$，$\varepsilon$用于控制对抗扰动$p_{\text{adv}}$的大小，将对抗扰动$p_{\text{adv}}$叠加到良性样本$x$上生成对抗样本$x'$。生成公式为$x'=x+\varepsilon\cdot S_{\text{sign}}$。

1.2.2　BIM

BIM[19]是FGSM的众多改进方法之一，有时也被称为迭代FGSM或I–FGSM。BIM在对抗扰动的上范数界内多次使用FGSM。BIM的对抗样本生成过程如式(1)所示：

$$\begin{aligned}&x'_{i+1}=\text{Clip}(x'_i+\varepsilon\cdot\text{sign}(\nabla_{x'_i}L(f(x'_i),y))),\\&x'_0=x\end{aligned}\tag{1}$$

式中：$i=0,1,\cdots,n$，n为总的迭代次数，通常取值为8；Clip(·)为裁剪算子，用于将对抗样本值限定在有效的范围内。以某次迭代生成的图像对抗样本x'_i为例，Clip(·)表达式如式(2)所示：

$$\begin{aligned}\text{Clip}\left(x'_{i,(u,v,w)}\right)=&\min\{255,x_{(u,v,w)}+\\&\varepsilon\cdot\max\{0,x_{(u,v,w)}-\varepsilon,x'_{i,(u,v,w)}\}\}\end{aligned}\tag{2}$$

将图像对抗样本x'_i的三通道坐标(u,v,w)的取值控制在良性样本x的ε邻域内，也限制在可行的输入空间(如8位亮度值范围为[0,255])内。

1.2.3　MI–FGSM

MI–FGSM[17]为BIM的改进方法，在BIM迭代添加FGSM对抗扰动的过程中，加入动量概念，使生成的对抗扰动更小，但能够更有效地愚弄目标推理模型。MI–FGSM的对抗样本生成过程和动量对抗扰动更新分别如式(3)、(4)所示：

$$\begin{aligned}&x'_{i+1}=\text{Clip}(x'_i+\varepsilon\cdot\text{sign}(g_{i+1})),i=0,1,\cdots,n;\\&x'_0=x\end{aligned}\tag{3}$$

$$g_{i+1}=\kappa\cdot g_i+\frac{\nabla_{x'_i}L(f(x'_i),y)}{\left\|\nabla_{x'_i}L(f(x'_i),y)\right\|_1},g_0=0\tag{4}$$

式中：g_i为加入动量概念后的对抗扰动；κ用于控制对抗扰动的大小，是常量。

1.2.4　DeepFool

DeepFool[20]可以用于估计一个良性样本x到目标推理模型f的最近决策边界的距离。该距离既可以用于量化目标推理模型对对抗样本x'的鲁棒性，也可以作为一个最小的对抗扰动p_{adv}的方向。对于线性二分类模型，到决策边界的距离可以使用点到线的距离公式计算。对于线性多分类模型，该距离可以近似为良性样本所处的类中最接近决策边界的距离。对于非线性分类模型，DeepFool通过线性化模型的每个类的决策边界在当前设定值$x'_i(x'_0=x)$的邻域范围内迭代扰动良性样本x，目标类y是最接近线性决策边界的类，移动x'_i到估计的边界点。整个过程一直重复，直到$f(x'_i)$被误分类为目标类y。

1.2.5　JSMA

JSMA[21]是基于雅克比显著图的对抗样本生成方法，利用显著图生成对抗样本。显著图的概念最初是为了可视化输入x中对深度神经网络的输出最重要的特征。显著性映射根据输入x的变化(例如图像中的每个像素)对深度神经网络的输出的影响来寻找输入x中对深度神经网络的输出最重要的特征。

JSMA通过扰动一组输入特征导致错误分类来利用显著图的信息。这与FGSM等不同，FGSM修改了大部良性样本x的特征，JSMA往往会发现稀疏的扰动。给定推理模型的输出$f(x)$，c为推理模型f输出的类别，t为指定的攻击类别。在$\sum_{c=t}\nabla_x f(x)<0$或者$\sum_{c=t}\nabla_x f(x)>0$的情况下，显著图的计算公式为$S=-\sum_{c=t}\nabla_x f(x)\cdot\sum_{c\neq t}\nabla_x f(x)$；在其他情况下，显著图为零。

1.2.6　C&W

C&W[22]是L–BFGS[16]的改进方法，包含一系列的基于约束优化的方法来生成对抗样本x'，其中的不同之处在于分别使用L_0、L_2、L_∞范数量化对抗扰动。C&W将一种通用约束优化转换为一种无约束优化的损失函数$L_{C\&W}(\cdot)$，如式（5）所示：

$$L_{C\&W}(x',t)=\max(\max\{f_i(x')\}_{i\neq t}-f_t(x'),-K) \quad (5)$$

式中，t为指定的攻击类别，K为反映对抗样本的最小期望置信度的参数，$f_i(x')$、$f_t(x')$分别为目标推理模型输出的对抗样本x'的输出的第i、t个分量。基于L_2范数的C&W攻击满足：

$$\begin{aligned}&\underset{w}{\operatorname{argmin}}[(\|x'-x\|_2^2)+c\cdot L_{C\&W}(x',t)],\\&x'=\frac{1}{2}(\tanh(w)+1)\end{aligned} \quad (6)$$

式中：w为一个变量，用于将对抗样本x'控制在[0,1]取值范围内，根据样本取值范围的不同，这个取值区间可以进行调整；参数c是常量，最优值是通过外部优化循环程序（例如二分查找法）来选择的。

1.3　对抗样本防御

在对抗样本防御中，目前最有效的是对抗训练方法。对抗训练[1]是一种防御方法，通过在每次训练迭代时向训练集中注入对抗样本x'，进而对目标推理模型f再训练，其目标推理模型的损失函数满足：

$$\underset{\theta}{\operatorname{argmin}}[\alpha L(f(x),y)+(1-\alpha)L(f(x'),y)] \quad (7)$$

式中，θ为目标推理模型的参数，α用于平衡两边的损失函数值。Madry等[27]提出了一种对抗训练的变体，其目标推理模型的损失函数满足：

$$\underset{\theta}{\arg\min}\ \mathrm{E}\left[\max_{\|x'-x\|_\infty}L(f(x'),y)\right] \quad (8)$$

由式（8）可知：$\max_{\|x'-x\|_\infty}L(f(x'),y)$最大化的目标是找到使损失函数$L(\cdot)$最大化的对抗样本$x'$；而$\underset{\theta}{\operatorname{argmin}}\ \mathrm{E}\left[\max_{\|x'-x\|_\infty}L(f(x'),y)\right]$最小化的目标是找到一组参数$\theta$，使最坏情况下损失函数$L(\cdot)$最小化。这与标准的对抗训练是不同的，标准的对抗训练在良性样本x和对抗样本x'上训练推理模型f，而在式（8）中只在对抗样本x'上训练推理模型f。集成对抗训练[6]是对抗训练的另一个变体，通过在其他的推理模型上生成对抗样本x'训练目标推理模型f。目标推理模型f和对抗样本x'的解耦克服了标准对抗训练中所观察到的过拟合现象。

1.4　生成对抗网络

2014年，Goodfellow等[10]提出了生成对抗网络，这是一种无监督的生成模型，因其强大的数据生成能力而受到广泛关注和研究。图2是生成对抗网络[10]的基本结构。

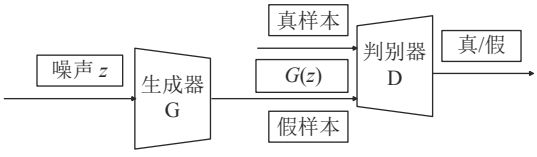


图 2　生成对抗网络架构[10]

Fig. 2　Architecture of the generative adversarial network[10]

由图2可知，生成对抗网络不是一个单一的网络，其有两个不同的网络，一个是生成器，另一个是判别器。生成器以随机噪声作为输入，输出为生成的假样本。判别器的目的是区分生成的假样本和现实中的真样本。生成对抗网络的训练采用了对抗博弈的方式，并且生成器的梯度更新信息来自于判别器，而不是数据。生成对抗网络的损失函数$L(D,G)$满足：

$$\begin{aligned}\min_G\max_D L(D,G)=&\mathrm{E}_{x\sim P_{\mathrm{data}}(x)}[\lg D(x)]+\\&\mathrm{E}_{z\sim P_z(z)}[\lg(1-D(G(z)))]\end{aligned} \quad (9)$$

式中，G为生成器函数，D为判别器函数，z为随机噪声，x为真实的样本，$G(z)$为生成的假样本，$P_{\mathrm{data}}(x)$为真实样本的分布，$P_z(z)$为生成的假样本的分布。生成器与判别器交替训练，生成器想要生成更加真实的假样本，判别器想要尽可能地区分真样本与假样本，从而相互博弈，达到纳什平衡点。最终，生成器可以生成以假乱真的假样本，判别器无法区分真样本与假样本。虽然生成对抗网络具有强大的数据生成能力，但原始的生成对抗网络也存在一些缺陷。主要问题是生成对抗网络的训练过程非常不稳定。Arjovsky等[28]分析生成对抗网络训练不稳定的原因，提出WGAN来保证训练过程的稳定，其最主要的改进是生成对抗网络的训练过程应该限制判别器的性能。Gulrajani等[15]提出WGAN–GP来解决WGAN在限制判别器性能上的不足。在WGAN–GP中，通过在生成对抗网络的损失函数中增加一个梯度惩罚项来限制判别器的性能。Wu等[29]对WGAN–GP中提出的梯度惩罚项，从数学上推导出梯度惩罚项的具体形式，提出WGAN–DIV的损失函数。本文在WGAN、WGAN–GP、

WGAN–DIV基础上建立损失函数，使得生成对抗网络的训练过程基本可以保持稳定。

2 基于生成对抗网络的对抗样本集成防御

2.1 防御原理

Kabkab等[11]证明了对抗样本位于良性样本的数据流形区域之外。因此，可以通过学习一个映射将对抗样本从对抗流形区域投影到良性流形区域以达到防御对抗样本的目的。生成对抗网络的优势在于可以很好地学习数据分布，并从所学到的分布中生成样本，即$x_{\text{fake}}=G(z)$，其中，z为随机噪声，x_{fake}为生成的样本。采用生成对抗网络学习一个输入为对抗样本、输出为良性样本的分布，通过学习该分布，生成器达到将对抗样本投影到良性流形区域的目的，即$x_{\text{benign}}=G(x_{\text{adversarial}})$，其中，$x_{\text{adversarial}}$为对抗样本，$x_{\text{benign}}$为良性样本。

2.2 基于WGAN–GP的对抗样本集成防御

图3为本文提出的基于生成对抗网络的对抗样本集成防御架构，其中，图3的上半部分为使用生成对抗网络训练多个将对抗样本投影到良性流形区域的生成器，图3的下半部分为集成多个生成器作为本文提出的集成防御。

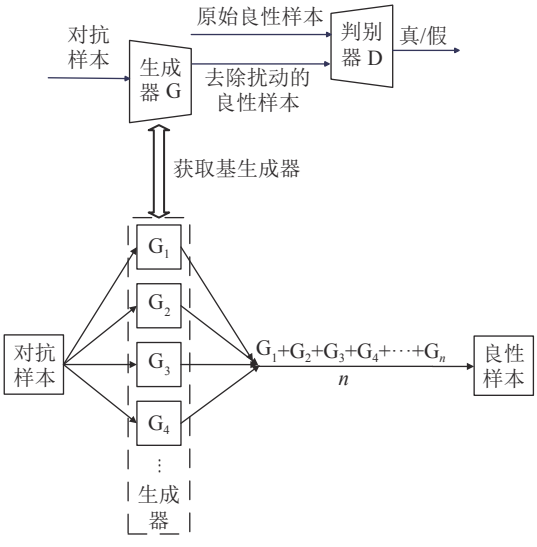


图 3 基于生成对抗网络的对抗样本集成防御架构

Fig. 3 Architecture of the ensemble defense scheme based on the generative adversarial network

为了确保生成对抗网络训练过程的稳定，使用Gulrajani等[15]提出的损失函数作为生成对抗网络自身的损失函数。生成器的网络结构采用卷积编码器–解码器架构，如图4所示。判别器的网络结构使用普通的卷积神经网络，如图5所示。

生成器的目的是去除对抗样本的对抗扰动，因此，生成器的输入为多种扰动的对抗样本，输出为去除扰动的良性样本。判别器的目的是对原始良性样本与去除扰动的良性样本做出判别，因此，判别器的输入为原始良性样本与去除扰动的良性样本。为了控制生成器去除对抗扰动的能力，本文的集成防御方法在原始WGAN–GP损失函数的基础上增加原始良性样本与去除扰动的良性样本之间的平方误差损失，总的损失函数为$L=L_{\text{WGAN–GP–G}}+L_{\text{WGAN–GP–D}}+L_{\text{MSE}}$，其中：$L_{\text{MSE}}$为最小平方误差损失函数，如式(10)所示；$L_{\text{WGAN–GP–G}}$为生成器的WGAN–GP损失函数，如式(11)所示；$L_{\text{WGAN–GP–D}}$为判别器的WGAN–GP损失函数，如式(12)所示。

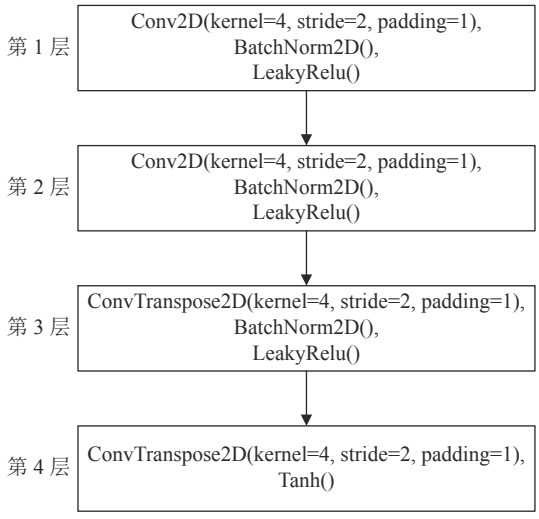


图 4 生成器的网络结构

Fig. 4 Network structure of the generator

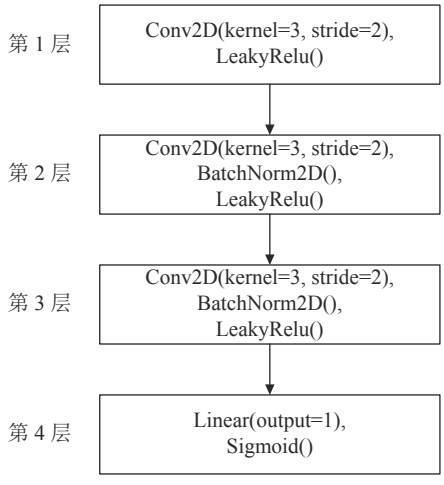


图 5 判别器的网络结构

Fig. 5 Network structure of the discriminator

$$L_{\text{MSE}}=\mathrm{E}\left(\left\|x_{\text{adversarial}}-x_{\text{benign}}\right\|_2^2\right) \tag{10}$$

$$L_{\text{WGAN–GP–G}}=-\mathrm{E}(D(G(x_{\text{adversarial}}))) \tag{11}$$

$$L_{\text{WGAN–GP–D}}=-\mathrm{E}(D(x_{\text{benign}}))+\mathrm{E}(D(G(x_{\text{adversarial}})))+\mathrm{GP} \tag{12}$$

式中，E(·)为求期望函数，$GP = E\left[(\|\nabla_x D(x)\|_2 - 1)^2\right]$为Gulrajani等[15]提出的损失函数中的梯度惩罚项，$\nabla_x D(x)$为对输入判别器的所有样本求梯度。

随着生成器与判别器交替训练，相互博弈，达到纳什平衡，生成器能够很好地去除对抗样本的扰动。考虑到所训练出的生成器对不同扰动的对抗样本会表现出不同的防御性能。因此，本文的集成防御方法将多个生成器进行集成作为最终的防御。本文方法通过提前训练生成器降低方法的时间消耗，通过集成多个生成器弥补方法的防御能力。

3　实验与分析

3.1　实验设置

使用MNIST与CIFAR10作为图像分类数据集。图6为目标推理模型在MNIST数据集上的网络结构。训练图6中的推理模型，得到分类准确率为98%。图7为目标推理模型在CIFAR10数据集上的网络结构。训练图7中的推理模型，得到分类准确率为83%。

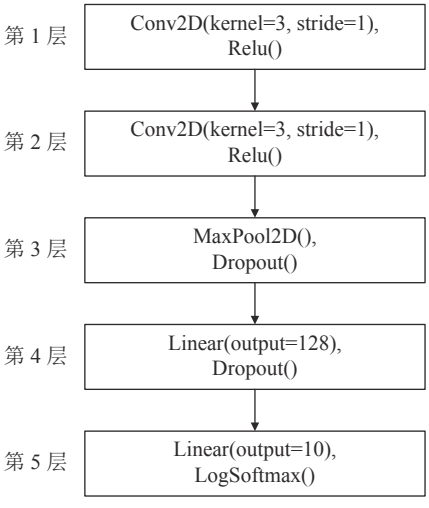


图 6　MNIST数据集上目标推理模型的网络结构

Fig. 6　Network structure of target model on MNIST dataset

对抗样本生成方法包括FGSM[18]、BIM[19]、DeepFool[20]、JSMA[21]、C&W[22]、MI–FGSM[17]。其中，FGSM、BIM、MI–FGSM在良性样本表面叠加全局的对抗扰动，DeepFool、JSMA、C&W修改良性样本局部的特征。

对抗样本防御方法包括APE–GAN[12]、Bit Depth[4]、TotalVarMin[30]、SpatialSmoothing[4]、JpegCompression[31]、FGSM对抗训练[1]、PGD对抗训练[27]，其中：APE–GAN、Bit Depth、TotalVarMin、SpatialSmoothing、JpegCompression是基于预处理的防御；FGSM对抗训练与PGD对抗训练是基于提升模型鲁棒性的防御，通过重新训练目标推理模型来防御对抗样本。

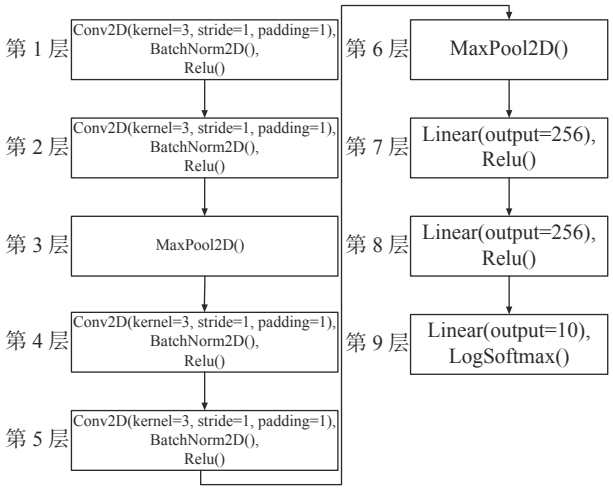


图 7　CIFAR10数据集上目标推理模型的网络结构

Fig. 7　Network structure of target model on CIFAR10 dataset

3.2　实验评估

实验步骤包括训练集成防御与防御对抗样本。集成防御由在数据集上生成的FGSM对抗样本训练得到。实验评估的内容包括在MNIST与CIFAR10数据集上使用各防御方法处理对抗样本得到目标推理模型的分类准确率和时间消耗对比。

3.2.1　MNIST数据集实验

MNIST数据集上，本文提出的集成防御方法与无防御（原始）及6种对比防御方法处理不同对抗样本得到目标推理模型的分类准确率如表1所示，其中，FGSM、BIM、MI–FGSM的对抗扰动设置为0.3。由于MNIST数据集的数据内容是灰度图片，所以对比的防御方法中不包括JpegCompression防御方法。

表 1　MNIST数据集上各防御方法处理不同对抗样本得到的分类准确率

Tab. 1　Classification accuracy of different types of adversarial examples processed by various defenses schemes on MNIST dataset

%

防御模型	FGSM	BIM	MI–FGSM	JSMA	DeepFool	C&W
无防御	10.18	1.12	1.28	52.34	51.37	54.69
本文提出的集成防御	96.02	95.89	98.22	80.37	97.41	96.56
APE–GAN	76.73	72.72	68.85	88.44	91.21	84.38
Bit Depth	83.27	78.85	73.90	68.48	94.82	55.47
TotalVarMin	33.37	14.42	11.11	70.22	80.86	57.81
SpatialSmoothing	18.90	1.19	10.36	37.61	70.07	70.31
FGSM对抗训练	67.51	8.91	3.16	53.26	75.70	14.06
PGD对抗训练	81.37	82.36	90.46	74.81	96.91	93.18

从表1可以看出：本文提出的集成防御方法在FGSM、BIM、MI–FGSM、DeepFool及C&W对抗样本上的分类准确率明显高于无防御（原始）及其他6种对比的防御方法的分类准确率。本文的集成防御方

法在JSMA对抗样本上的分类准确率略低于APE–GAN防御方法的分类准确率，其原因是本文的集成防御方法会去除JSMA对抗样本表面的良性特征。

另外，实验还在MNIST数据集上探究了对抗扰动设置与不同防御方法的目标推理模型分类准确率的关系。以FGSM对抗样本为例，当设置对抗扰动分别为0.1、0.3、0.5、0.7时，不同防御方法处理后的目标推理模型的分类准确率如图8所示。从图8可以看出：当对抗扰动设置为0.5时，本文提出的集成防御方法的分类准确率低于PGD对抗训练防御方法的分类准确率，其原因是本文的集成防御方法不能消除过大的对抗扰动。当对抗扰动设置为0.7时，本文提出的集成防御方法的分类准确率高于对抗扰动设置为0.5时的本文集成防御方法的分类准确率，其原因是超出图像数据值范围的对抗扰动被截断。

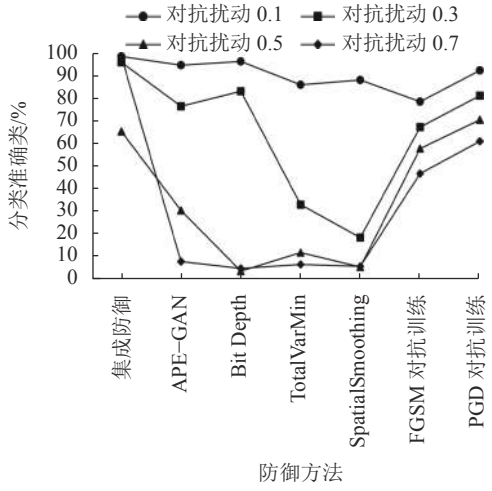


图 8 MNIST数据集上FGSM对抗样本的不同对抗扰动与分类准确率的关系

Fig. 8 Relationship between different adversarial perturbations of FGSM adversarial example and classification accuracy of various defense schemes on MNIST dataset

表2为本文提出的集成防御方法与6种对比防御方法处理10 000幅FGSM对抗样本图片的时间消耗。从表2可以看出，本文的集成防御方法的时间消耗小于TotalVarMin、FGSM对抗训练、PGD对抗训练防御方法的时间消耗，等于Bit Depth防御方法的时间消耗，但略大于APE–GAN防御方法的时间消耗，且差值平均到每幅图片后是相当小的，可以看作是系统误差。

综上所述，在对抗扰动不太大的情况下，本文的集成防御方法在5种对抗样本上的分类准确率高于其他对比的防御方法的分类准确率；仅在JSMA对抗样本上的分类准确率略低于APE–GAN防御方法。在忽略系统误差的情况下，本文的集成防御方法的时间消耗小于或者等于其他对比的防御方法的时间消耗。

表 2 MNIST数据集上7种防御方法的时间消耗

Tab. 2 Time consumption of seven defense schemes on MNIST dataset

防御模型	时间消耗
本文提出的集成防御	1.4 s
APE–GAN	1.3 s
Bit Depth	1.3 s
TotalVarMin	22 min
SpatialSmoothing	1.4 s
FGSM对抗训练	2 h
PGD对抗训练	6 h

3.2.2 CIFAR10数据集实验

表3为CIFAR10数据集上本文提出的集成防御方法与无防御（原始）和7种对比防御方法处理不同对抗样本生成方法得到目标推理模型的分类准确率，其中，FGSM、BIM、MI–FGSM的对抗扰动设置为0.03。由于CIFAR10数据集的数据内容是彩色图片，所以所对比的防御方法中还包括JpegCompression防御方法。与表1相比，本文提出的集成防御方法在CIFAR10数据集上的表现弱于其在MNIST数据集上的表现。这主要是因为CIFAR10数据集比MNIST数据集更复杂，需要具有更复杂网络结构的生成器去防御对抗样本。从表3可以看出：本文提出的集成防御方法在JSMA、DeepFool对抗样本上的分类准确率明显高于无防御（原始）及其他对比的防御方法的分类准确率。本文提出的集成防御方法在FGSM、BIM、MI–FGSM对抗样本上的分类准确率低于PGD对抗训练防御方法的分类准确率，其原因是PGD对抗训练是通过提升模型的鲁棒性来防御对抗样本，不受数据集的影响。本文提出的集成防御方法在C&W对抗样本上的分类准确率略低于APE–GAN对抗训练防御方法的分类准确率，但该差值可以考虑为正常的计算误差。

表 3 CIFAR10数据集上各防御方法处理不同对抗样本得到的分类准确率

Tab. 3 Classification accuracy of different types of adversarial examples processed by various defense schemes on CIFAR10 dataset

%

防御模型	FGSM	BIM	MI–FGSM	JSMA	DeepFool	C&W
无防御	10.17	9.37	9.37	1.80	44.87	7.03
本文提出的集成防御	62.96	63.97	68.24	67.77	69.97	71.09
APE–GAN	58.23	59.93	58.56	59.38	64.84	71.88
Bit Depth	20.28	12.24	24.26	35.60	57.81	54.69
TotalVarMin	35.24	35.62	34.16	45.31	49.22	47.66
SpatialSmoothing	20.19	11.07	28.49	41.41	66.41	65.63
JpegCompression	12.05	9.37	30.15	39.84	66.41	68.75
FGSM对抗训练	41.54	12.55	64.23	7.50	37.00	28.13
PGD对抗训练	66.59	65.71	75.32	46.71	55.91	51.66

实验还在CIFAR10数据集上探究了对抗扰动设置与不同防御方法处理后的目标推理模型分类准确率的关系。以FGSM对抗样本为例，当设置对抗扰动为0.01、0.03、0.05、0.07时，不同防御方法的目标推理模型的分类准确率如图9所示。从图9可以看出，随着对抗扰动的增加，本文提出的集成防御方法的分类准确率始终低于PGD对抗训练防御方法的分类准确率，其原因是PGD对抗训练是通过提升模型的鲁棒性来防御对抗样本，不受数据集的影响。

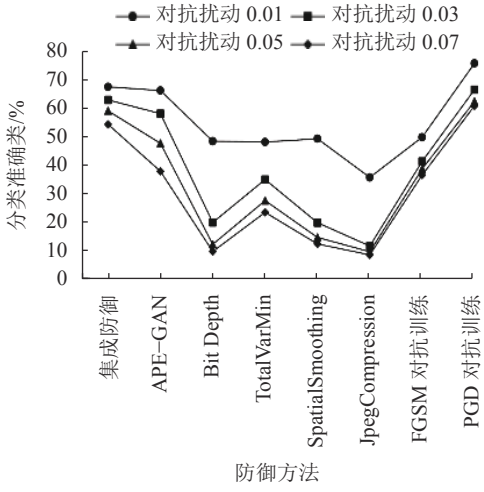


图 9　CIFAR10数据集上FGSM对抗样本的不同对抗扰动与不同防御方法的分类准确率的关系

Fig. 9　Relationship between different adversarial perturbations of FGSM adversarial example and classification accuracy of various defense schemes on CIFAR10 dataset

表4为本文提出的防御方法与7种对比防御方法处理10 000幅FGSM对抗样本图片的时间消耗。从表4中可以看出，本文的集成防御方法的时间消耗小于TotalVarMin、FGSM对抗训练、PGD对抗训练防御方法的时间消耗，大于APE–GAN、Bit Depth、JpegCompression防御方法的时间消耗，但该差值平均到每幅图片后也是很小的，可以看作是系统误差。

表 4　CIFAR10数据集上8种防御方法的时间消耗

Tab. 4　Time consumption of eight defense schemes on CIFAR10 dataset

防御模型	时间消耗
本文提出的集成防御	1.6 s
APE–GAN	1.4 s
Bit Depth	1.4 s
TotalVarMin	28 min
SpatialSmoothing	1.4 s
JpegCompression	1.4 s
FGSM对抗训练	4 h
PGD对抗训练	8 h

综上所述，本文的集成防御方法在JSMA、DeepFool对抗样本上的分类准确率高于其他对比的防御方法的分类准确率，在其他对抗样本上的分类准确率略低于个别防御方法。在忽略系统误差的情况下，本文的集成防御方法的时间消耗小于或者等于其他对比的防御方法的时间消耗。

4　结　论

针对现有对抗样本防御方法防御能力不足、时间消耗过高等问题，参考生成对抗网络与集成学习在对抗样本研究中的优势，本文提出一种基于生成对抗网络的对抗样本集成防御方法。该方法通过提前训练生成器来降低方法的时间消耗，通过集成多个生成器来弥补方法的防御能力。在MNIST与CIFAR10数据集上验证了本文提出的集成防御方法在目标推理模型上的分类准确率与时间消耗。实验结果表明本文的集成防御方法能以较低的时间消耗防御多种对抗样本，并且在目标模型上的分类准确率比其他对比防御方法更高。下一步研究，将对现有的生成对抗网络的架构及损失函数进行改进，期望进一步提升方法的防御能力。另外，计划在大尺寸图片、语音、视频等更加复杂的数据集上验证该方法的通用性。

参考文献：

[1] Metzen J H,Kumar M C,Brox T,et al.Universal adversarial perturbations against semantic image segmentation[C]//Proceedings of the 2017 IEEE International Conference on Computer Vision.Venice:IEEE,2017:2774–2783.

[2] Hendrycks D,Gimpel K.A baseline for detecting misclassified and out-of-distribution examples in neural networks[EB/OL].[2021–02–23].https://arxiv.org/abs/1610.02136.

[3] Feinman R,Curtin R R,Shintre S,et al.Detecting adversarial samples from artifacts[EB/OL].[2021–02–23].https://arxiv.org/abs/1703.00410.

[4] Xu Weilin,Evans D,Qi Yanjun.Feature squeezing:detecting adversarial examples in deep neural networks[C]//Proceedings 2018 Network and Distributed System Security Symposium.San Diego:NDSS,2018:1–15.

[5] Athalye A,Carlini N,Wagner D.Obfuscated gradients give a false sense of security:Circumventing defenses to adversarial examples[EB/OL].[2021–02–23].https://arxiv.org/abs/1802.00420v1.

[6] Tramèr F,Kurakin A,Papernot N,et al.Ensemble adversarial training:Attacks and defenses[EB/OL].[2021–02–23].https://arxiv.org/abs/1705.07204.

[7] Hinton G,Vinyals O,Dean J.Distilling the knowledge in a neural network[EB/OL].[2021–02–23].https://arxiv.org/abs/1503.02531.

[8] Dhillon G S,Azizzadenesheli K,Lipton Z C,et al.Stochastic activation pruning for robust adversarial defense[EB/OL].

[2021–02–23].https://arxiv.org/abs/1803.01442.

[9] Song Yang,Kim T,Nowozin S,et al.PixelDefend:Leveraging generative models to understand and defend against adversarial examples[EB/OL].[2021–02–23].https://arxiv.org/abs/1710.10766.

[10] Goodfellow I J,Pouget-Abadie J,Mirza M,et al.Generative adversarial nets[C]//Proceedings of the 27th International Conference on Neural Information Processing System.Cambridge:MIT Press,2014:2672–2680.

[11] Kabkab M,Samangouei P,Chellappa R.Task-aware compressed sensing with generative adversarial networks[EB/OL].[2021–02–23].https://arxiv.org/abs/1802.01284.

[12] Jin Guoqing,Shen Shiwei,Zhang Dongming,et al.APE-GAN:Adversarial perturbation elimination with GAN[C]//Proceedings of the ICASSP 2019-2019 IEEE International Conference on Acoustics,Speech and Signal Processing.Brighton:IEEE,2019:3842–3846.

[13] Wei Wenqi,Liu Ling,Loper M,et al.Cross-layer strategic ensemble defense against adversarial examples[C]//Proceedings of the 2020 International Conference on Computing,Networking and Communications (ICNC).Big Island:IEEE,2020:456–460.

[14] Gowda S N,Yuan Chun.StegColNet:Steganalysis based on an ensemble colorspace approach[M]//Lecture Notes in Computer Science.Cham:Springer International Publishing,2021:313–323.

[15] Gulrajani I,Ahmed F,Arjovsky M,et al.Improved training of Wasserstein GANs[C]//Proceedings of the 31st International Conference on Neural Information Processing Systems.Red Hook:Curran Associates Inc.,2017:5769–5779.

[16] Szegedy C,Zaremba W,Sutskever I,et al.Intriguing properties of neural networks[EB/OL].[2021–02–23].https://arxiv.org/abs/1312.6199v1.

[17] Dong Yinpeng,Liao Fangzhou,Pang Tianyu,et al.Boosting adversarial attacks with momentum[C]//Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition.Salt Lake City:IEEE,2018:9185–9193.

[18] Goodfellow I J,Shlens J,Szegedy C.Explaining and harnessing adversarial examples[EB/OL].[2021–02–23].https://arxiv.org/abs/1412.6572.

[19] Kurakin A,Goodfellow I,Bengio S,et al.Adversarial attacks and defences competition[M]//The NIPS'17 Competition:Building Intelligent Systems.Cham:Springer,2018:195–231.

[20] Moosavi–Dezfooli S M,Fawzi A,Frossard P.DeepFool:A simple and accurate method to fool deep neural networks[C]//Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition.Las Vegas:IEEE,2016:2574–2582.

[21] Papernot N,McDaniel P,Jha S,et al.The limitations of deep learning in adversarial settings[C]//Proceedings of the 2016 IEEE European Symposium on Security and Privacy (EuroS&P).Saarbruecken:IEEE,2016:372–387.

[22] Carlini N,Wagner D.Adversarial examples are not easily detected:Bypassing ten detection methods[C]//AISec'17:Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security.New York:ACM,2017:3–14.

[23] Rauber J,Zimmermann R,Bethge M,et al.Foolbox Native:Fast adversarial attacks to benchmark the robustness of machine learning models in PyTorch,TensorFlow,and JAX[J].Journal of Open Source Software,2020,5(53):2607.

[24] Su Jiawei,Vargas D V,Sakurai K.One pixel attack for fooling deep neural networks[J].IEEE Transactions on Evolutionary Computation,2019,23(5):828–841.

[25] Chen Pinyu,Zhang Huan,Sharma Y,et al.ZOO:Zeroth order optimization based black-box attacks to deep neural networks without training substitute models[C]//AISec'17:Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security.New York:ACM,2017:15–26.

[26] Zhang Yang,Foroosh H,David P,et al.CAMOU:Learning physical vehicle camouflages to adversarially attack detectors in the wild[C]//Seventh International Conference on Learning Representations.New Orleans:ICLR,2019:1–20.

[27] Madry A,Makelov A,Schmidt L,et al.Towards deep learning models resistant to adversarial attacks[EB/OL].[2021–02–23].https://arxiv.org/abs/1706.06083.

[28] Arjovsky M,Chintala S,Bottou L.Wasserstein generative adversarial networks[C]//Proceedings of the 34th International Conference on Machine Learning.Sydney:JMLR,2017:214–223.

[29] Wu Jiqing,Huang Zhiwu,Thoma J,et al.Wasserstein divergence for GANs[M]//Computer Vision – ECCV 2018.Cham:Springer,2018:673–688.

[30] Guo Chuan,Rana M,Cisse M,et al.Countering adversarial images using input transformations[EB/OL].[2021–02–23].https://arxiv.org/abs/1711.00117.

[31] Das N,Shanbhogue M,Chen S T,et al.SHIELD:Fast,practical defense and vaccination for deep learning using JPEG compression[C]//KDD'18:Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining.New York:ACM,2018:196–204.

（编辑 赵 婧）

引用格式：Cao Tianjie,Yu Zhikun,Qi Yunyan,et al.Ensemble adversarial example defense based on generative adversarial network[J].Advanced Engineering Sciences,2022,54(2):56–64.[曹天杰,余志坤,祁韵妍,等.基于生成对抗网络的对抗样本集成防御[J].工程科学与技术,2022,54(2):56–64.]