控制与决策
Control and Decision

基于复合生成对抗网络的对抗样本生成算法研究

孔锐, 蔡佳纯, 黄钢, 张冰

引用本文:
孔锐,蔡佳纯,黄钢,张冰. 基于复合生成对抗网络的对抗样本生成算法研究[J]. 控制与决策, 2023, 38(2): 528−536.

在线阅读 View online: https://doi.org/10.13195/j.kzyjc.2021.0028

基于复合生成对抗网络的对抗样本生成算法研究

孔 锐[1,2], 蔡佳纯[2], 黄 钢[2], 张 冰[1,†]

(1. 暨南大学 智能科学与工程学院，广东 珠海 509070；2. 暨南大学 信息科学技术学院，广州 510632)

摘 要: 对抗样本能够作为训练数据辅助提高模型的表达能力,还能够评估深度学习模型的稳健性. 然而,通过在一个小的矩阵范数内扰乱原始数据点的生成方式,使得对抗样本的规模受限于原始数据. 为了更高效地获得任意数量的对抗样本,探索一种不受原始数据限制的对抗样本生成方式具有重要意义. 鉴于此,提出一种基于生成对抗网络的对抗样本生成模型(multiple attack generative adversarial networks, M-AttGAN). 首先,将模型设计为同时训练 2 组生成对抗网络,分别对原始数据样本分布和模型潜在空间下的扰动分布进行建模;然后,训练完成的 M-AttGAN 能够不受限制地高效生成带有扰动的对抗样本,为对抗训练和提高深度神经网络的稳健性提供更多可能性;最后,通过 MNIST 和 CIFAT-10 数据集上的多组实验,验证利用生成对抗网络对数据分布良好的学习能力进行对抗样本生成是可行的. 实验结果表明,相较于常规攻击方法,M-AttGAN 不仅能够脱离原始数据的限制生成高质量的对抗样本,而且样本具备良好的攻击性和攻击迁移能力.

关键词: 对抗攻击；对抗训练；生成式对抗网络；条件模型；样本生成

中图分类号: TP273 **文献标志码:** A

DOI: 10.13195/j.kzyjc.2021.0028

引用格式: 孔锐, 蔡佳纯, 黄钢, 等. 基于复合生成对抗网络的对抗样本生成算法研究 [J]. 控制与决策, 2023, 38(2): 528-536.

Research on generative adversarial example algorithm based on multiple GANs

KONG Rui[1,2], *CAI Jia-chun*[2], *HUANG Gang*[2], *ZHANG Bing*[1†]

(1. School of Intelligent Systems Science and Engineering，Jinan University，Zhuhai 509070，China；2. College of Information Science and Technology，Jinan University，Guangzhou 510632，China)

Abstract: Attack examples can not only be used as training data to improve the expressive ability of the model but also can be used to evaluate the robustness of the deep learning model. However, the size of the attack examples is limited to the original data by perturbing an existing data point within a small matrix norm. In order to obtain attack examples more efficiently, a multiple attack generative adversarial networks (M-AttGAN) is proposed, where the attackers are not restricted to original data. The proposed network is designed to train two pairs of GANs simultaneously to fit for the distribution of original data and the distribution of the perturbation in the GANs latent space. The trained model, can generate attack examples efficiently without restrictions, and provide more data for adversarial training and improve the robustness of neural networks. We adopt human evaluation and contrastive analysis with other state-of-the-art algorithms to prove that it is feasible to utilize GANs to attack example generation. Experimental results on the MNIST and CIFAR-10 dataset show that the proposed model not only generates high-quality attack examples breaking the limits of the original data, but also has good aggression and attack migration competence.

Keywords: adversarial attack；adversarial training；generative adversarial network；conditional models；example generation

0 引 言

现代深度神经网络(deep neural networks, DNN)在解决复杂问题方面取得了显著成果[1]. 但是研究表明,深度网络非常容易受到来自对抗样本的攻击[2]. 2013 年, Szegedy 等[3]注意到测试样本中不可察觉的扰动具有令神经网络误分类的可能性. 对抗攻击在

收稿日期: 2021-01-06；**录用日期:** 2021-11-26.
基金项目: 广东省自然科学基金项目(2020A151501718).
责任编委: 彭木根.
†通讯作者. E-mail: tzhangbing@jnu.edu.cn.

人类视觉系统无法察觉输入变化的同时,诱导模型得出完全偏离真实值的结果. 这些样本使得模型将其分类到攻击者指定的类别,或是与原始样本不同的类别. 2014年, Goodfellow等[4]提出了基于梯度的对抗样本生成算法(fast gradient sign method, FGSM),该算法通过寻找模型梯度变化最大的方向生成扰动,使得模型无法正确识别输入样本. FGSM操作简单、效果良好,继而出现了很多衍生算法[5-7]. 然而,此类方法需要在攻击中访问被攻击目标网络的架构和参数. Carlini等[8]提出C&W (Carlini and Wagner attacks),发现对抗扰动可以从不安全网络迁移至安全网络,可转移性意味着该算法同时适合进行黑盒攻击. 2018年, Xiao等[9]提出了基于GAN的攻击算法,通过训练好的生成器能够将输入样本转换为扰动并形成对抗样本,证明了基于无穷范式距离约束生成的对抗样本较基于最优化方程和简单像素空间的矩阵度量生成的样本更具真实性. Zhao等[10]在生成对抗网络的基础上通过输入原始样本构建图像的语义空间,将语义空间中的隐变量通过网络映射成对抗样本,使得生成的对抗样本更自然. 相较于全像素添加扰动,这种基于某些评判标准的部分添加扰动的策略更加注重选择的像素数量、代价与对抗性之间的关系[11].

对抗样本的存在表明模型倾向于依赖不可靠的特征以最大化性能,若特征受到干扰,则将造成模型误分类. 对抗样本存在的原因,归结于三方面: 1)深度模型的非线性导致的输入和输出映射的不连续性; 2)不充分的模型平均和不充分的正则化导致的过拟合[3]; 3)深度模型存在线性部分[4]. 从另一个角度而言,由于训练集是真实分布的抽样,训练出来的模型边界不可能完全拟合真实的决策边界,而对抗攻击算法就是要找到一种高效的方法生成这个对抗区域的样本,从而实现对模型的攻击[12].

总体而言,设计一种能够快速产生高质量对抗样本的生成算法,为模型的研究提供足量的高质量、多样性好的对抗样本数据源,是有效推进机器学习研究的方向之一. 主流对抗攻击算法分为两种: 1)以原始样本作为输入,然后设计算法生成扰动,并将扰动叠加到原始样本,从而获得对抗样本; 2)以原始样本作为输入,然后设计算法直接生成对抗样本. 而大部分现有研究依赖于第1种方式[2],这意味着在一些数据源获取受限的场景中,现有算法产生的对抗样本的规模将无法满足诸如对抗训练[13]等需要大量对抗样本作为研究基础的项目. 本文基于第2种生成方法论,介绍一种更为通用的对抗样本生成算法. 该算法在训练后能够不受原始数据限制,高效地产生任意数量的对抗样本.

在各种生成模型中, 生成式对抗网络(generative adversarial networks, GAN) 由 Goodfellow 等[14] 于2014年提出, GAN的优化过程能够使生成器估测到数据样本的分布,从而产生真假难辨的数据样本. 2017年, Odena等[15]提出ACGAN (auxiliary classifier GAN),将目标函数设置为真实数据样本似然与正确分类标签似然的和,从而细分调节损失函数使得分类正确率更高,进一步地提高了网络的生成和判别能力. 此外, Arjovsky等[16]从损失函数入手,证明当使用JS散度作为目标分布与生成分布相近度的度量时,在目标分布与生成分布的重叠区域可忽略的情况下, JS散度为一常数,此时生成器的获得梯度为0,网络无法继续优化的基础上,提出了使用Earth-Mover (EM)距离作为相似度度量的Wasserstein GAN (WGAN),为解决GAN存在的训练困难、损失函数无法指导训练、生成样本缺乏多样性等问题指明了一个全新的方向. 2018年, Brock等[17]通过大Batch、大参数、截断技巧和大规模GAN训练稳定性控制等技巧实现了BigGAN,完成了超大规模生成架构的设计,同时基于对模型输入潜在空间的讨论,探究模型输入的先验噪声的选择对生成网络性能的影响. 生成对抗网络理论提出至今,其具有广泛的应用场景使得该理论支持下的衍生模型层出不穷,有效推动了图像生成、超分辨率、风格转换、图像修复等应用的进展[18].

鉴于GAN在数据表达和分布学习上的优势[19],所提出对抗样本生成方法被命名为M-AttGAN,其在生成式对抗网络原理的基础上,设计为2组生成对抗网络联合训练的结构,并提出对应的训练算法以训练生成器,这种网络架构不仅可以训练生成器拟合真实样本的分布,还使得生成样本具备攻击属性. 更重要的是, M-AttGAN生成的对抗样本不再依赖现有数据,训练完毕的M-AttGAN只需要输入噪声信号,便能够生成任意数量的带有扰动的对抗样本,无需再访问被攻击的模型,无需额外收集并清洗原始样本. 为了验证所提出方法的有效性,在实验部分,将从样本的生成效率、生成质量、攻击性和迁移攻击性等多方面对比所提出方法和现有的对抗样本生成方法.

1 M-AttGAN

通过对GAN和对抗攻击发展的研究,所提出的M-AttGAN在结构上设计为由2个生成器和2个判别器组成,同时在GAN训练方法的基础上借鉴C&W攻击算法,设计出一种新的训练算法,负责完成M-

AttGAN整体网络的训练.

对抗样本的基本思想是在样本上添加扰动,理想情况下,对抗样本x_{adv}在人眼能够正确识别的同时被训练好的分类器F误分类. 生成式对抗网络的全局最优表现为生成器的生成分布与目标分布的渐近一致,因此方法设计的关键为通过训练令生成器分布逼近目标分布,同时设计第2个生成器完成对扰动分布的逼近,最终得到一个不依赖原始数据的对抗样本生成器组.

基于以上分析,M-AttGAN如图1所示,训练框架包括2组生成式对抗网络共5个部分:生成器G_1负责逼真样本的生成,判别器D_1负责指导G_1训练,生成器G_2负责扰动的生成,判别器D_2负责指导G_2训练以及目标攻击网络F.

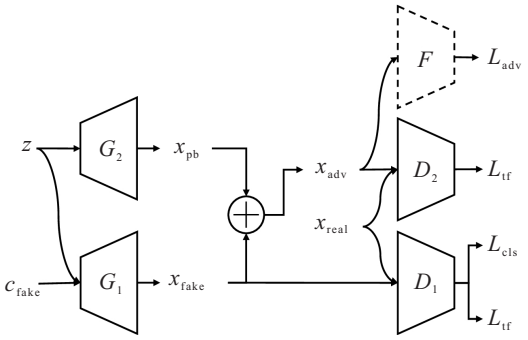

图 1 M-AttGAN架构

所提出模型的损失函数以WGAN-GP损失函数为基础进行设计,同时,为了生成具有指定语义的样本,M-AttGAN将类别信号引入生成器,从而引导模型训练;最后为了令生成样本具有攻击性,修改C&W攻击的优化函数作为对抗损失函数. 综上所述,由判别损失L_{tf}、分类损失L_{cls}、对抗损失L_{adv}共同训练M-AttGAN.

因此第1部分,G_1接受随机噪声z和类别信号c_{fake},生成假样本x_{fake};然后x_{fake}被D_1接收,D_1的任务是分辨真实样本x_{real}和生成样本x_{fake}的同时尽量对输入样本进行分类. 训练G_1和D_1由判别损失L_{tf}和分类损失L_{cls}完成,L_{tf}将引导x_{fake}与x_{real}不可区分,L_{cls}将使得D_1能够有效对数据进行分类从而令G_1能够根据类别信号c_{fake}生成指定样本.

第2部分,G_2接受随机噪声z,生成扰动x_{pb},设置扰动阈值ε_{pb},有$|x_{\text{pb}}| \leqslant \varepsilon_{\text{pb}}$;然后$G_1$在固定参数后接收随机噪声$z$生成假样本$x_{\text{fake}}$,继而$x_{\text{adv}} = x_{\text{pb}} + x_{\text{fake}}$作为输入被送进判别器$D_2$. 此处$D_2$仅仅负责分辨$x_{\text{real}}$和攻击样本$x_{\text{adv}}$,其判别损失$L_{\text{tf}}$确保生成样本和扰动叠加后得到的$x_{\text{adv}}$也与真实样本不可区分. 同时为了保证攻击样本能够欺骗目标网络,将x_{adv}作为输入,F输出关于x_{adv}的对抗损失,该损失表示目标网络预测类别与真实类别的距离的负值或是预测类别与攻击期望的类别的距离,因此最小化L_{adv},能够引导G_2生成x_{pb}使得F无法正确分类x_{adv},同时添加范数损失L_{hinge}帮助训练使得扰动尽可能不被察觉. 最终为每个损失函数设置权重因子保证x_{adv}在逼近真实分布和保证攻击强度之间的平衡,策略上而言,判别损失L_{tf}的权重应该小于对抗损失L_{adv}的权重.

图2为所提出的基于GAN的对抗样本生成框架M-AttGAN的训练流程. 具体到每个网络和训练过程,损失函数设计如下.

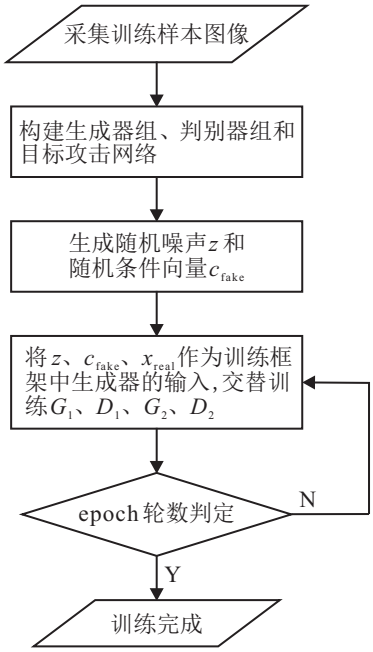

图 2 M-AttGAN训练流程

对于生成器G_1,输入为随机噪声z和类别信号c_{fake},输出生成样本x_{fake}. $L(G_1)$由两部分组成,分别为G_1的生成样本x_{fake}在D_1监督下的分类损失L_{cls}和判别损失L_{tf}. 有

$$L_{\text{tf}}(G_1) = -E_{z,c_{\text{fake}}}[D_1(G_1(z, c_{\text{fake}}))],$$

$$L_{\text{cls}}(G_1) = E_{z,c_{\text{fake}}}[L_{D_1}(c_{\text{fake}}|G_1(z, c_{\text{fake}}))],$$

$$L(G_1) = L_{\text{cls}}(G_1) + L_{\text{tf}}(G_1).$$

对于判别器D_1,输入来自G_1的x_{fake}、真实样本x_{real}和对应的类别c_{real}. $L(D_1)$由两部分组成,分别为D_1对真实样本和生成样本的判别损失$L_{\text{tf}}(D_1)$、对真实样本的分类损失$L_{\text{cls}}(D_1)$,此刻,$P_{\tilde{x}}$为从目标分布与当前模型分布P_{G_1}中的一对样本点之间沿直线均匀采样的分布[16]. 有

$$\begin{aligned} L_{\text{tf}}(D_1) = &-E_{x_{\text{real}}}[D_1(x_{\text{real}})]+ \\ &E_{z,c_{\text{fake}}}[D_1(G_1(z, c_{\text{fake}}))]+ \\ &\lambda E_{\tilde{x}}[(||\nabla_{\tilde{x}} D_1(\tilde{x})||_2 - 1)^2], \end{aligned}$$

$$L_{\text{cls}}(D_1) = E_{x_{\text{real}}}[L_{D_1}(c_{\text{real}}|x_{\text{real}})],$$
$$L(D_1) = L_{\text{tf}}(D_1) + L_{\text{cls}}(D_1).$$

对于生成器G_2,输入为随机噪声z,生成噪声x_{pb},得到攻击样本$x_{\text{adv}}=x_{\text{pb}}+x_{\text{fake}}$. $L(G_2)$由三部分组成,$L_{\text{tf}}(G_2)$为x_{adv}在D_1监督下的判别损失,$L_{\text{adv}}(G_2)$为将x_{adv}作为目标攻击网络F的输入得到的对抗损失,L_{hinge}为以L_2范数为基准的边界损失. 其中:η为用户定义的扰动的最大边界,φ和μ负责控制每个损失函数之间的相对重要程度. 有

$$L_{\text{tf}}(G_2) = -E_{z,x_{\text{fake}}}[D_2(x_{\text{fake}} + G_2(z))],$$
$$L_{\text{adv}}(G_2) = E_{z,c_{\text{fake}},x_{\text{fake}}}[L_F(c_{\text{fake}}|x_{\text{fake}} + G_2(z))],$$
$$L_{\text{hinge}}(G_2) = E_z\max[0, ||G_2(z)||_2 - \eta],$$
$$L(G_2) = L_{\text{tf}}(G_2) + \varphi L_{\text{adv}}(G_2) + \mu L_{\text{hinge}}(G_2).$$

对于判别器D_2,输入对抗样本x_{adv}和真实样本x_{real},判别损失函数为$L(D_2)$,此刻,$P_{\tilde{x}}$为从目标分布与当前模型分布$P_{G_1+G_2}$中的一对样本点之间沿直线均匀采样的分布. 有

$$L(D_2) = -E_{x_{\text{real}}}[D_2(x_{\text{real}})] + E_{z,x_{\text{fake}}}[D_2(x_{\text{fake}} + G_2(z))] + \lambda E_{\tilde{x}}[(||\nabla_{\tilde{x}} D_2(\tilde{x})||_2 - 1)^2].$$

对于超参数扰动阈值ε_{pb},数值设置得越大,噪声越明显,图片失真越严重,但攻击效果越显著,对于进行过标准化操作的训练数据,本文推荐阈值设置范围为$0.2 \leqslant \varepsilon_{\text{pb}} \leqslant 0.6$.

在上述若干损失函数的指导下,交替训练G_1、D_1、G_2、D_2,优化器采用Adam优化损失函数.

2　实验与分析

本文在MNIST数据集和CIFAR-10数据集上进行实验分析,实验环境:计算机处理器为32 Intel (R) Xeon (R) CPU E5-2620 v4@2.10 GHz, 64 GB运行内存(RAM),两块NVIDIA Tesla P4 GPU, PyTorch框架.

1) MNIST数据集包含0～9共10类手写数字灰度样本,样本尺寸为(1, 28, 28),整个数据集有60 000个训练样本,10 000个测试样本.

2) CIFAR-10数据集包含了10类共60 000个样本尺寸为(3, 32, 32)的彩色样本,整个数据集由50 000个训练样本和10 000个测试样本组成.

2.1　MNIST 实验

对于MNIST数据集,其生成器和判别器网络分别如图3和图4所示,均以(1, 28, 28)大小的训练数据为例.

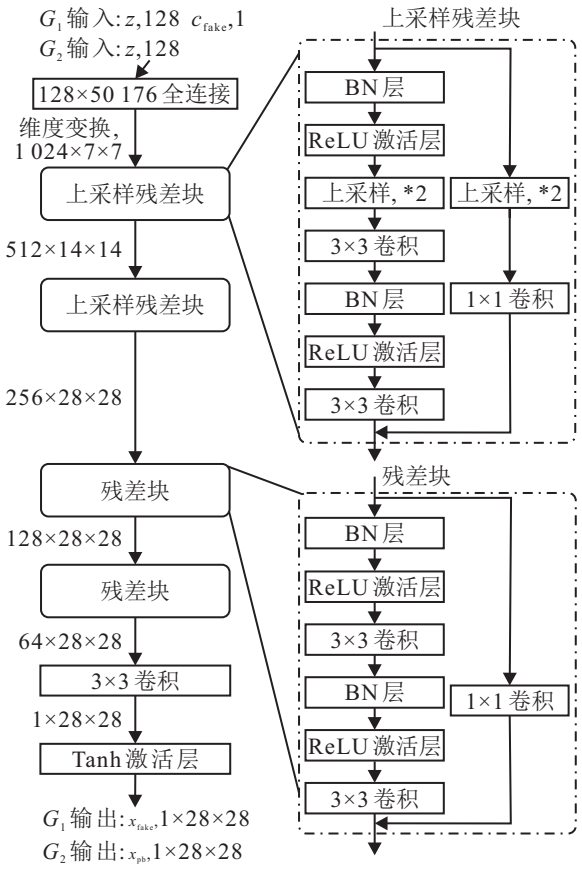

图3　生成器构架图

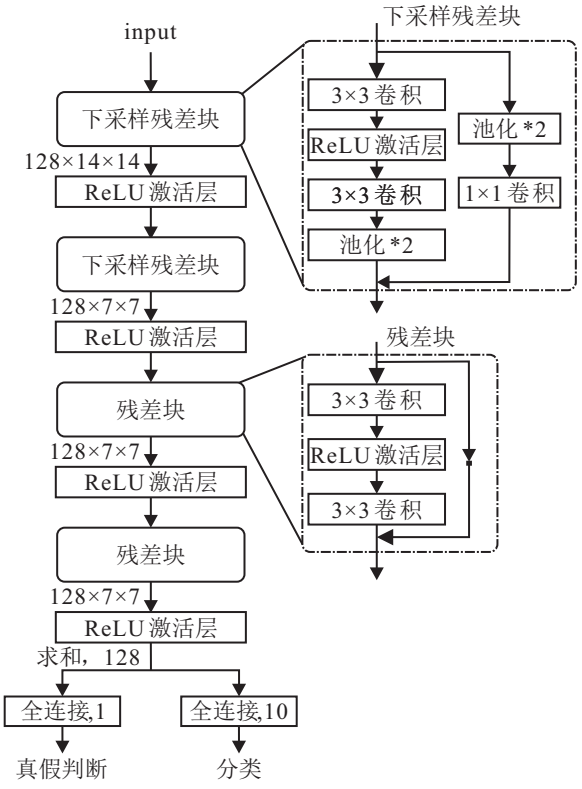

图4　判别器构架图

MNIST数据集中,超参数设置如下.

1) 训练迭代次数epoch设置为64轮;

2) 基于MNIST数据集是黑白手写数据集,样本相对简单,因此G_1和D_1的更新次数前期设置为2 : 5,

中后期设置为1 : 1;

3) G_2 和 D_2 判别器与生成器的更新次数设置为1 : 1;

4) Adam优化器的学习速率设置为0.000 2,一阶矩估计的指数衰减率为0.5,二阶矩估计的指数衰减率为0.9;

5)扰动阈值 ε_{pb} 设置为0.5;

6)损失函数 $L_{\text{adv}}(G_2)$ 和 $L_{\text{hinge}}(G_2)$ 的权重 $\varphi:\mu$ 设置为10 : 1;

7)用户定义的扰动的最大边界 η 设置为0.2.

采用以残差单元为基础的简单残差网络作为此次MNIST实验的生成器和判别器的架构,具体如下.

1)无论是判别器还是生成器,均去除反卷积,只保留普通卷积层;

2)卷积核的大小统一使用 3×3;

3)通过UpSampling 2D和AvgPooling 2D实现上采样和下采样;

4)生成器除了最后一层使用Tanh激活函数,其他层使用ReLU激活函数,判别器统一使用ReLU作为激活函数;

5)生成器模型输入由噪声 z 和类别信号 c_{fake} 组成,模型中Batch normalization (BN)层将对隐藏层的输入进行归一化,其参数 β 和 γ 将依赖于类别信号[20];

6)BN会引入同一个batch中不同样本的相互依赖关系,而本文模型需要对每个样本独立地施加梯度惩罚,因此判别器的模型架构中不使用BN.

MNIST实验中采用VGG 11作为目标网络,其预训练识别MNIST样本的准确率为99 %+. 图5为MNIST数据集下目标分类器 F 识别生成样本 x_{fake} 的准确率. 如图5所示,随着训练的进行,G_1 生成的假样本 x_{fake} 在目标分类器 F 中的识别率越来越高并逐渐接近真实样本 X_{real} 在目标分类器中的识别率,可以判断 G_1 的分布逐渐拟合了目标分布. 图6为MNIST数据集下目标分类器 F 识别对抗样本 x_{adv} 的准确率. 如图6所示,假样本叠加了扰动后成为攻击样本 x_{adv},目标分类器 F 对 x_{adv} 的识别率呈下降趋势,这意味着攻击样本的攻击性随着训练的持续而逐渐增强. 最终,模型训练时常约为20个小时,总的趋势表现为,假样本越来越真,同时假样本攻击性越来越强,对于简单数据而言,本文方法 G_1 能够很快地完成对目标分布的拟合,因而更加需要关注 G_2 的学习效果,即扰动强度是否能够成功攻击分类器.

为分析所提出算法M-AttGAN生成的攻击样本质量,选择将M-AttGAN生成的攻击样本与FGSM[4]、

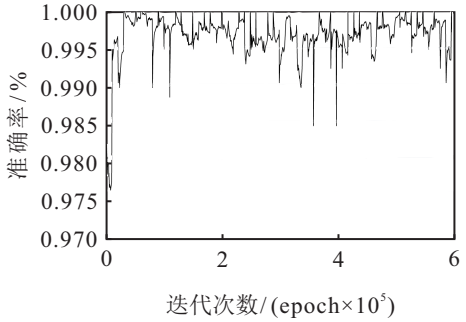

图5 MNIST数据集下目标分类器 F 识别生成样本 x_{fake} 的准确率

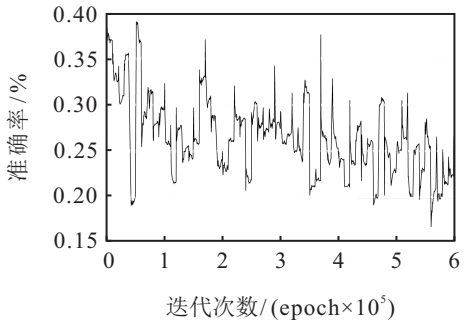

图6 MNIST数据集下目标分类器 F 识别对抗样本 x_{adv} 的准确率

BIM[5]、C&W[8]三者进行比较,从而突出所提出方法相较于最常用的梯度方法和优化方法的优势. 目前为止尚未有被学界广泛接受的关于GAN的质量评价方法,其博弈和收敛机制背后的数学分析仍有待建立[21],因此对样本的生成质量的评估方法采用人工评估的方法,分别使用4种攻击算法随机抽样产生500张攻击样本(500张图片均能够导致目标网络误分类). 本文实验将相同的攻击样本分配给10个志愿者,每位志愿者需要为每张图片选择正确的标签,并统计出人类对不同方法生成的攻击样本的识别率;与此同时,实验中还提取出500张原始样本(Orig.images)供志愿者识别并进行类别选择,从而计算人类对原始样本的识别率.

实验结果如表1所示,原始图片的平均识别率为98.26 %,是本文实验人眼识别的基准数据. 所提出方法的平均识别率为98.76 %,高于原始图片的平均识别率,显著高于对照组算法生成的攻击样本识别率,这个结果意味着M-AttGAN生成的样本质量最好,识别M-AttGAN的样本比识别原始样本还要容易. 其原因为在原始数据集中本身存在一些难以辨别的样本,这些样本影响了人眼的判断,而基于GAN的方法侧重拟合原始数据的通用特征,这使得M-AttGAN在训练过程中能够忽略样本集的噪声特征,最终的生成样本的识别率超过了原始样本的识别率,整体实验数据表明了M-AttGAN生成的数据质量最优.

表1　MNIST数据集中不同攻击算法下的攻击样本的人类准确率　　%

	1	2	3	4	5	6	7	8	9	10	**Avg.**
Orig.	99.2	97.8	**97**	99.2	97.2	97.8	97.2	98.8	99.2	**99.2**	98.26
FGSM	94	93.4	94	97.2	96.4	93	93	93.4	94.4	96	94.48
BIM	96.8	93.8	94.8	98	94.2	95.6	96.8	96.8	97.4	97.6	96.18
C&W	96	94.8	95.6	98.2	96.2	96.6	97.8	96	96	96.4	96.36
M-AttGAN	**99.2**	**99.2**	96.8	**99.6**	**99**	**97.8**	**99**	**99.2**	**99.8**	98	**98.76**

图7为MNIST数据集中不同攻击算法下的攻击样本采样图片,以直观的方式展示了MNIST数据集中各算法攻击产生的攻击样本,将所提出方法与原始图片、FGSM[4]、BIM[5]、C&W[8]方法进行对比. 由图7可见,着重效率的基于梯度的攻击方法在样本质量整体视觉上不如着重质量的基于优化的攻击方法,FGSM生成的样本噪声非常大,BIM产生可视噪声方面略优于FGSM,C&W攻击方法的样本质量最纯净,噪声最小. 通过进一步观察发现,所提出方法生成的样本纯净度方面比梯度攻击方法好,稍逊于基于优化的攻击方法.

(a) 原始图片　(b) FGSM　(c) BIM　(d) C&W　(e) M-AttGAN

图 7　MNIST数据集中不同攻击算法下的攻击样本采样图片

实验中,在被分类器识别正确的图片中添加对抗扰动从而获得攻击样本,分类器对攻击样本进行预测,并将其预测错误的样本数占攻击样本总数的比例定义为攻击成功率.

实验选取文献[2]提及的常用对抗攻击算法,FGSM[4]、DeepFool[7]、BIM[5]、L-BFGS(limited-memory BFGS)[3]、C&W[8]、advGAN[9]、Uni. Perturbations(universal adversarial perturbations)[22]、ATNs(adversarial transformation networks)[23]和所提出方法M-AttGAN分别产生500张图片对目标网络进行攻击,其攻击结果如表2所示. 主流算法的生成机制均是在原始样本上进行扰动添加[2],M-AttGAN则是利用网络对原始样本分布进行学习,学习完成后利用网络生成假样本并叠加扰动从而形成攻击样本. 由表2可见,所提出方法的攻击成功率与主流攻击算法相当,从而验证了所提出方法机制的可行性. 同时,对比数据可以发现,所提出方法的攻击成功率高于梯度攻击算法,略差于基于优化的方法.

表2　各算法下攻击样本的攻击成功率(500张攻击样本)

攻击方法	攻击成功率/%
FGSM	75.2
DeepFool	68.8
BIM	82.4
L-BFGS	75.6
C&W	92.2
advGAN	90
ATNs	79
Uni. perturbations	77.4
M-AttGAN	84

对比评估训练完毕后M-AttGAN生成框架的样本生成效率,选取了不同策略下最常用的攻击方法作为代表与所提出方法进行比较. 与传统算法不同的是,M-AttGAN需要额外的训练时间训练生成器,不过一旦训练完毕,M-AttGAN在生成样本效率方面具有显著优势. 表3为不同攻击算法生成500张能够成功误导目标分类器F手写数字图片需要的时间,相同生成任务强度下,M-AttGAN比梯度攻击方法快5倍以上,而且远远快于基于优化的攻击方法.

表3　各算法下生成攻击样本所需时间(500张攻击样本)

攻击方法	运行时间
FGSM	≈ 0.04 s
BIM	≈ 0.09 s
C&W	> 4 h
M-AttGAN	< 0.01 s

由表1和表2可见,M-AttGAN能够有效兼顾攻击效果和生成质量,由表3可见,样本的生成效率方面M-AttGAN最佳. 在本文实验中,M-AttGAN攻击成功率和生成效率均比梯度方法要好;而与基于优

化的方法相比,M-AttGAN攻击成功率虽然稍微次于基于优化的方法,但是其生成效率却远大于基于优化的方法. 因此综合实验指标,M-AttGAN比现有的传统攻击算法有优势,而且由于生成策略并不直接依赖于原始样本,意味着M-AttGAN能够无限生成新的攻击样本,这是传统方法不具备的.

2.2 CIFAR-10 实验

在CIFAR-10实验中,着重探究攻击样本的迁移攻击能力. 模型结构方面,生成器和判别器的结构整体采用MNIST实验中的结构. 训练过程中,超参数设置如下.

1) 训练迭代次数epoch设置为60轮;

2) G_1 和 D_1 负责对真实分布进行学习从而生成假样本,为了保持对抗平衡,判别器和生成器的更新次数设置为2 : 5;

3) G_2 和 D_2 负责对对抗空间进行学习从而生成扰动,判别器和生成器的更新次数设置为1 : 1;

4) Adam优化器的学习速率设置为0.000 1,一阶矩估计的指数衰减率为0.5,二阶矩估计的指数衰减率为0.9;

5) 扰动阈值 ε_{pb} 设置为0.2;

6) 损失函数 $L_{\text{adv}}(G_2)$ 和 $L_{\text{hinge}}(G_2)$ 的权重 $\varphi:\mu$ 设置为5 : 1;

7) 用户定义的扰动的最大边界 η 设置为0.5.

类似地,对于CIFAR-10数据集而言,模型训练时长约为33个小时,在训练过程中,采用VGG 19网络作为目标网络,预训练模型的准确率为87 %+. 图8为CIFAR-10数据集下目标分类器 F 识别生成样本 x_{fake} 的准确率. 如图8所示,随着训练的进行,生成器 G_1 生成的假样本 x_{fake} 在目标分类器 F 中的识别率越来越高并逐渐接近原始样本在目标分类器中的识别率,这表明假样本逐渐真实. 继而,假样本叠加扰动后作为攻击样本 x_{adv} 攻击 F,随着训练进行,攻击效果如图9所示,与MNIST实验中目标分类器 F 对 x_{adv} 的识别率逐渐下降的趋势不同,CIFAR-10实验下,F 对 x_{adv} 的识别率一直保持在低位水平. 这是因为训练数据本身的分布导致的:MNIST数据集的样本为(1, 28, 28)的灰度样本,而CIFAR-10数据集的样本为(3, 32, 32)的彩色样本,CIFAR-10数据分布的复杂程度远大于MNIST,对于越复杂的数据,分类器越难以学习数据的完整分布. 这意味着,对于攻击算法而言,在CIFAR-10的分类器上更容易找到对抗空间,因此在CIFAR-10实验中,作为负责扰动生成的 G_2 对对抗空间分布的学习难度也更小,该逻辑应用于MNIST实验表现为需要更多轮数的训练,其攻击效果也更不容易体现. 因此,对于复杂数据而言,所提出方法更加需要关注 G_1 的生成效果,即假样本是否足够真.

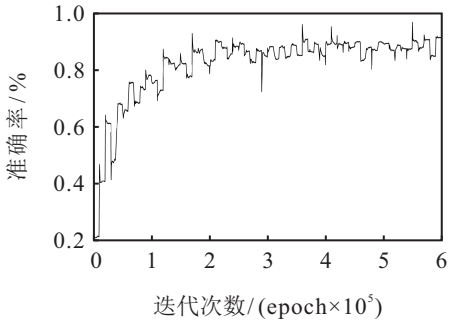

图 8 CIFAR-10数据集下目标分类器 F 识别生成样本 x_{fake} 的准确率

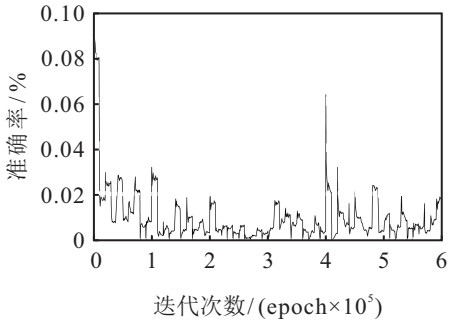

图 9 CIFAR-10数据集下目标分类器 F 识别对抗样本 x_{adv} 的准确率

本文利用攻击目标网络产生的攻击样本以验证对抗样本的迁移攻击性,选取3种不同的梯度攻击方法FGSM、BIM和MI-FGSM (momentum iterative FGSM)[24] 作为基准方法进行比较,将这些方法产生的样本输入至ResNet 18、ResNet 101、DenseNet 121、Inception-v4这4种预训练模型中得到准确率,其中Orig.为原始图片,实验结果如表4所示.

根据目标分类器 F 对攻击样本的分类准确率,可以推算得到攻击样本的攻击成功率,由表4可见,M-AttGAN是一个攻击效果显著的对抗样本生成算法,

表4 各CIFAR-10分类器对不同攻击算法以VGG 19作为目标网络产生的攻击样本的分类准确率 %

	VGG 19*	ResNet 18	ResNet 101	DenseNet 121	Inception-v4
Orig.	87.5	90.2	92.1	94.05	93.6
FGSM	10.5	35.3	37.2	40.5	37.3
BIM	3	40.4	43.2	50.5	48.1
MI-FGSM	0	33.6	33.1	37.5	29.2
M-AttGAN	5.6	27.5	30.3	32.1	29.6

对目标模型VGG 19的攻击成功率达94.4%,略弱于BIM和MI-FGSM,在迁移攻击方面,本文方法优于对照算法,效果最佳. 同时为了凸显所提出方法的有效性和说服力,再以ResNet 18作为目标分类器,选用FGSM、BIM和MI-FGSM作为基准方法进行比较. 将这些方法产生的样本输入至VGG 19、 ResNet 101、DenseNet 121、Inception-v4这4种预训练模型中得到分类准确率,其中Orig.为原始图片,实验结果如表5所示.

表5　各CIFAR-10分类器对数据集中不同攻击算法以ResNet18作为目标网络产生的攻击样本的分类准确率　　%

–	ResNet 18*	VGG 19	ResNet 101	DenseNet 121	Inception-v4
Orig.	90.2	87.5	92.1	94.05	93.6
FGSM	17.2	31.2	34.8	36.2	30.3
BIM	9.2	38.2	47	54.6	48.6
MI-FGSM	2.4	30.4	34.4	37	24.8
M-AttGAN	10.2	25	34.8	30.1	24.7

由表4和表5可见,所提出方法的结果与主流算法相当,但M-AttGAN的优势之一在于算法的输入不与原始数据强耦合,在生成过程中只能利用生成器将噪声转换为攻击样本,而不是依赖原始数据. 进一步而言,所提出方法在保证对目标网络充分的攻击强度的同时,能够有效地将攻击特性迁移到其他模型上,具体分析有两个要点: 1) 逼真的假样本是由基于学习真实样本后的假样本生成器产生,因此假样本基于假样本生成器分布而不是基于原始数据分布; 2) 负责扰动生成的扰动生成器在训练中并不依赖目标网络的内部信息,仅仅是依据对目标网络输入输出,这使得训练完毕的扰动生成器产生的扰动是基于其对对抗空间的学习得到的分布. 最终,假样本和扰动叠加得到的攻击样本表现为在保证攻击性的同时表现出强的迁移性.

图10和图11分别为M-AttGAN生成的对抗样本和原始样本. M-AttGAN能够生成多样性和可辨识性高的样本. 由于假样本生成器负责拟合真实样本的分布,单纯依靠假样本生成器产生的假样本无法完全欺骗分类器,还需要一个扰动生成器在假样本上产生攻击扰动,最终在双方同时作用下实现攻击样本的生成,因此在视觉上,攻击样本能够与真实样本类似,同时导致分类器误分类. 由于实验资源的局限性,生成样本的视觉效果还有提升空间. 通过原理分析可以肯定的是,在所提出生成方法论的基础上,通过选择如PG-GAN (progressive growing GAN)[25]、SA-GAN (self-attention GAN)[26]、胶囊GAN[27]等更为复杂的生成器网络,能够确保生成质量更为优秀,同时保证逼真样本的攻击强度.

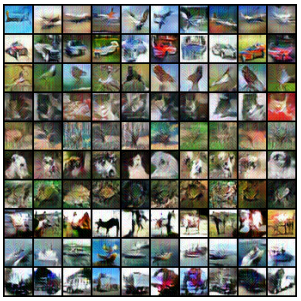

图 10　CIFAR-10数据集下生成的攻击样本

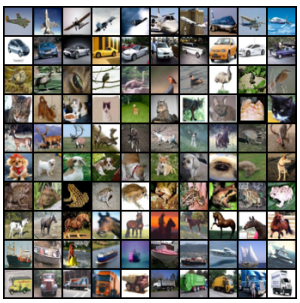

图 11　CIFAR-10数据集原始样本

3　结　论

本文利用生成对抗网络在分布学习和数据生成上的优势,提出了攻击样本生成算法M-AttGAN. 新方法在MNIST数据集和CIFAR-10数据集上进行实验均取得优异的实验结果. 在MNIST数据集上着重探究了M-AttGAN方法在生成效率、样本质量和攻击能力三方面,通过对比实验,M-AttGAN效率高、质量好,同时攻击强度与对照算法相当; 而在CIFAR-10的实验结果表明,M-AttGAN在攻击迁移性方面效果最显著. 综合而言,所提出方法将攻击样本的生成过程与原始数据解耦,在保证攻击强度的同时,提高了攻击样本的生成可能性,丰富了攻击样本的生成多样性和保证了攻击样本的强迁移性,这三方面的优势将为深度神经网络的对抗训练和稳健性评估提供充足有效的对抗样本数据源,继而为深度学习系统克服对抗攻击提供有效的帮助.

参考文献(References)

[1]　LeCun Y, Bengio Y, Hinton G. Deep learning[J]. Nature, 2015, 521(7553): 436-444.

[2] Akhtar N, Mian A. Threat of adversarial attacks on deep learning in computer vision: A survey[J]. IEEE Access, 2018, 6: 14410-14430.

[3] Szegedy C, Zaremba W, Sutskever I, et al. Intriguing properties of neural networks[J/OL]. 2013, arXiv: 1312.6199.

[4] Goodfellow I J, Shlens J, Szegedy C. Explaining and harnessing adversarial examples[J/OL]. 2014, arXiv: 1412.6572.

[5] Kurakin A, Goodfellow I J, Bengio S. Adversarial examples in the physical world[J/OL]. 2016, arXiv: 1607.02533.

[6] Tramèr F, Kurakin A, Papernot N, et al. Ensemble adversarial training: Attacks and defenses[J/OL]. 2017, arXiv: 1705.07204.

[7] 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. Las Vegas: IEEE, 2016: 2574-2582.

[8] Carlini N, Wagner D. Towards evaluating the robustness of neural networks[C]. IEEE Symposium on Security and Privacy. Piscataway, IEEE, 2017: 39-57.

[9] Xiao C, Li B, Zhu J Y, et al. Generating adversarial examples with adversarial networks[J/OL]. 2018, arXiv: 1801.02610.

[10] Zhao Z L, Dua D, Singh S. Generating natural adversarial examples[J/OL]. 2017, arXiv: 1710.11342.

[11] 潘文雯, 王新宇, 宋明黎, 等. 对抗样本生成技术综述[J]. 软件学报, 2020, 31(1): 67-81.
(Pan W W, Wang X Y, Song M L, et al. Survey on generating adversarial examples[J]. Journal of Software, 2020, 31(1): 67-81.)

[12] 易平, 王科迪, 黄程, 等. 人工智能对抗攻击研究综述[J]. 上海交通大学学报, 2018, 52(10): 1298-1306.
(Yi P, Wang K D, Huang C, et al. Adversarial attacks in artificial intelligence: A survey[J]. Journal of Shanghai Jiao Tong University, 2018, 52(10): 1298-1306.)

[13] 孔锐, 蔡佳纯, 黄钢. 基于生成对抗网络的对抗攻击防御模型[J]. 自动化学报, 2020, 41: 1-17.
(Kong R, Cai J C, Huang G. Defense to adversarial attack with generative adversarial network. Acta Automatica Sinica, 2020, 41: 1-17)

[14] Goodfellow I, Pouget-Abadie J, Mirza M, et al. Generative adversarial nets[J]. Advances in Neural Information Processing Systems, 2014: 2672-2680.

[15] Odena A, Olah C, Shlens J. Conditional image synthesis with auxiliary classifier GANs[C]. International conference on machine learning. Sydney, 2017: 2642-2651.

[16] Arjovsky M, Chintala S, Bottou L. Wasserstein GAN[J/OL]. 2017, arXiv: 1701.07875.

[17] Brock A, Donahue J, Simonyan K. Large scale GAN training for high fidelity natural image synthesis[J/OL]. 2018, arXiv: 1809.11096.

[18] 王坤峰, 苟超, 段艳杰, 等. 生成式对抗网络 GAN 的研究进展与展望[J]. 自动化学报, 2017, 43(3): 321-332.
(Wang K F, Gou C, Duan Y J, et al. Generative adversarial networks: The state of the art and beyond[J]. Acta Automatica Sinica, 2017, 43(3): 321-332.)

[19] 黄钢. 生成对抗算法及其在对抗样本生成上的应用研究[D]. 广州: 暨南大学, 2020.
(Huang G. Research and application of a generative adversarial algorithm and its impact on adversarial examples[D]. Guangzhou: Jinan University, 2020.)

[20] Miyato T, Koyama M. cGANs with projection discriminator[J/OL]. 2018, arXiv: 1802.05637.

[21] 林懿伦, 戴星原, 李力, 等. 人工智能研究的新前线: 生成式对抗网络[J]. 自动化学报, 2018, 44(5): 775-792.
(Lin Y L, Dai X Y, Li L, et al. The new frontier of AI research: Generative adversarial networks[J]. Acta Automatica Sinica, 2018, 44(5): 775-792.)

[22] Moosavi-Dezfooli S M, Fawzi A, Fawzi O, et al. Universal adversarial perturbations[C]. IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2017: 86-94.

[23] Baluja S, Fischer I. Adversarial transformation networks: Learning to generate adversarial examples[J/OL]. 2017, arXiv: 1703.09387.

[24] Dong Y P, Liao F Z, Pang T Y, et al. Boosting adversarial attacks with momentum[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018: 9185-9193.

[25] Karras T, Aila T, Laine S, et al. Progressive growing of gans for improved quality, stability, and variation[J/OL]. 2017, arXiv: 1710.10196.

[26] Zhang H, Goodfellow I, Metaxas D, et al. Self-attention generative adversarial networks[C]. International Conference on Machine Learning. Long Beach, 2019: 7354-7363.

[27] 孔锐, 黄钢. 基于条件约束的胶囊生成对抗网络[J]. 自动化学报, 2020, 46(1): 94-107.
(Kong R, Huang G. Conditional generative adversarial capsule networks[J]. Acta Automatica Sinica, 2020, 46(1): 94-107.)

作者简介

孔锐 (1964−), 男, 教授, 博士, 从事模式识别、图像处理等研究, E-mail: tkongrui@jnu.edu.cn;

蔡佳纯 (1995−), 女, 硕士生, 从事生成对抗网络的研究, E-mail: gptcjc1126@163.com;

黄钢 (1994−), 男, 硕士生, 从事生成对抗网络的研究, E-mail: hhhgggpps@gmail.com;

张冰 (1965−), 女, 高级工程师, 从事计算机视觉的研究, E-mail: tzhangbing@jnu.edu.cn.

(责任编辑: 魏 冰)