基于Transformer和GAN的对抗样本生成算法

刘帅威，李智，王国美，张丽
（贵州大学计算机科学与技术学院公共大数据国家重点实验室，贵州 贵阳 550025）

摘　要：对抗攻击与防御是计算机安全领域的一个热门研究方向。针对现有基于梯度的对抗样本生成方法可视质量差、基于优化的方法生成效率低的问题，提出基于Transformer和生成对抗网络（GAN）的对抗样本生成算法Trans-GAN。首先利用Transformer强大的视觉表征能力，将其作为重构网络，用于接收干净图像并生成攻击噪声；其次将Transformer重构网络作为生成器，与基于深度卷积网络的鉴别器相结合组成GAN网络架构，提高生成图像的真实性并保证训练的稳定性，同时提出改进的注意力机制Targeted Self-Attention，在训练网络时引入目标标签作为先验知识，指导网络模型学习生成具有特定攻击目标的对抗扰动；最后利用跳转连接将对抗噪声施加在干净样本上，形成对抗样本，攻击目标分类网络。实验结果表明：Trans-GAN算法针对MNIST数据集中2种模型的攻击成功率都达到99.9%以上，针对CIFAR10数据集中2种模型的攻击成功率分别达到96.36%和98.47%，优于目前先进的基于生成式的对抗样本生成方法；相比快速梯度符号法和投影梯度下降法，Trans-GAN算法生成的对抗噪声扰动量更小，形成的对抗样本更加自然，满足人类视觉不易分辨的要求。

关键词：深度神经网络；对抗样本；对抗攻击；Transformer模型；生成对抗网络；注意力机制
源代码链接：https://github.com/liushuaiwei/Trans-GAN
中图分类号：TP391　　**文献标志码：**A　　**DOI：**10.19678/j.issn.1000-3428.0067077

Adversarial Example Generation Algorithm Based on Transformer and GAN

LIU Shuaiwei, LI Zhi, WANG Guomei, ZHANG Li
（State Key Laboratory of Public Big Data, College of Computer Science and Technology, Guizhou University, Guiyang 550025, Guizhou, China）

【Abstract】Adversarial attack and defense is a popular research area in computer security. Trans-GAN, an adversarial example generation algorithm based on the combination of Transformer and Generate Adversarial Network (GAN), is proposed to address the problems of the poor visual quality of existing gradient-based adversarial example generation methods and the low generation efficiency of optimization-based methods. First, the algorithm utilizes the powerful visual representation capability of the Transformer as a reconstruction network for receiving clean images and generating adversarial noise. Second, the Transformer reconstruction network is combined with a deep convolutional network-based discriminator as a generator to form a GAN architecture, which improves the authenticity of the generated images and ensures the stability of training. Meanwhile, the improved attention mechanism, Targeted Self-Attention, is proposed to introduce target labels as a priori knowledge when training the network, which guides the network model to learn to generate adversarial perturbations with specific attack targets. Finally, adversarial noise is added to the clean examples using skip-connections to form adversarial examples. Experimental results demonstrate that the proposed algorithm achieves an attack success rate of more than 99.9% on both models used for the MNIST dataset and 96.36% and 98.47% on the two models used for the CIFAR10 dataset, outperforming the current state-of-the-art generative-based adversarial attack methods. The qualitative results show that compared to the Fast Gradient Sign Method (FGSM) and Projected Gradient Descent (PGD) algorithms, the generated adversarial noise of the Trans-GAN algorithm is less perturbed, and the formed adversarial examples are more natural and meet the requirements of human vision, which is not easily distinguished.

【Key words】deep neural network; adversarial example; adversarial attack; Transformer model; Generate Adversarial Network (GAN); attention mechanism

0　引言

深度神经网络在过去几年取得了巨大成功，并引起了学术界和产业界的广泛关注[1]。随着深度神经网络的快速发展和部署，其显露出的安全问题逐渐引起了社会的关注[2]。最近的研究发现，深度神经网络很容易受到对抗样本的影响，对抗样本通常是在合法样本上添加精心设计且难以察觉的扰

收稿日期：2023-03-03　　**修回日期：**2023-04-12
基金项目：国家自然科学基金（62062023）。
通信作者 E-mail：2870333101@qq.com

动[3]。利用对抗样本可以让攻击者在很多应用场景中对人工智能模型本身发起攻击，从而造成难以估计的危害，例如：基于人脸识别技术的手机解锁系统可能存在安全漏洞[4]，攻击者可以通过伪装等方式欺骗人脸解锁，从而进入他人手机；自动驾驶汽车的识别系统会被攻击者误导[5]，可能导致交通事故；不法分子利用对抗样本欺骗计算机辅助诊断系统修改其医疗诊断记录，从而实施骗保行为[6]。研究更先进的对抗攻击算法能评估深度神经网络的鲁棒性，进一步揭露深度学习的安全漏洞，有助于开发人员对系统的运行进行维护，从而提高系统的安全性和稳定性。

现有的对抗攻击方法根据攻击方式的不同大致分为 3 类：1）基于梯度的攻击；2）基于优化的攻击；3）基于生成式的攻击。文献[7]提出基于梯度的快速梯度符号法（FGSM），使用模型的梯度信息来计算对原始样本的扰动从而欺骗模型。文献[8]提出迭代式的投影梯度下降（PGD）攻击，在 FGSM 的基础上引入多次迭代，并加入投影步骤来限制对抗样本的扰动大小，以提高攻击的效果和稳定性。文献[9]提出优化式的 CW 攻击，通过最小化对抗样本与原始样本之间的距离来生成对抗样本。CW 攻击通过对模型输出的标签和置信度进行修改，使得模型更容易将对抗样本进行错误分类。

在基于生成式的攻击方法中：文献[10]提出 AdvGAN 方法，其主要思想是训练一个生成器和一个鉴别器，使用生成器生成具有误导性的对抗样本，使用鉴别器区分对抗样本和原始样本；文献[11]提出 AdvGAN 的改进版 AdvGAN++，向生成器中加入一定的高斯噪声，使得生成的对抗样本更具随机性和多样性，提高了攻击的鲁棒性和隐蔽性；文献[12]提出 AI-GAN 方法，其设计灵感来自于人类的攻击行为，即在攻击者拥有攻击目标背景知识的情况下，攻击者可以针对目标进行精准攻击。AI-GAN 利用这种攻击思路，首先通过背景知识收集对目标模型有益的信息，然后利用这些信息生成对抗样本。

虽然上述攻击方法在对抗样本生成方面取得了一定成果，但是它们也存在一些缺点，例如：基于梯度的攻击方法生成的对抗样本有一定的规律性，易于被发现；基于优化的 CW 方法生成的对抗样本迁移性差，在多个模型中攻击效果不一致，且计算成本较高；基于生成式的方法生成的对抗样本在可接受性和可解释性方面有待提高。虽然 AI-GAN 可以针对特定目标进行攻击，生成的对抗样本更具针对性和欺骗性，但是 AI-GAN 也存在一些缺点，如需要额外的辅助数据集和背景知识，攻击效率较低。上述不足都限制了对抗攻击在实际中的应用，因此，还需要进一步研究和改进对抗攻击方法。

本文结合先进的深度学习模型 Transformer[13]和生成对抗网络（GAN）[14]，提出基于生成式的攻击算法 Trans-GAN。该算法使用 Transformer 作为重构网络，利用其强大的视觉表征能力增强对抗样本的可接受性和可解释性；算法同时训练基于 Transformer 的生成器和基于深度卷积神经网络（DCNN）[15]的鉴别器，通过对鉴别器的对抗训练来提高生成图像的真实性以及攻击的鲁棒性和欺骗性。此外，本文提出改进的注意力机制 Targeted Self-Attention，修改网络输入，同时接收干净图像和攻击目标，针对特定目标进行攻击，生成更具针对性和欺骗性的对抗样本。

1 相关工作

1.1 对抗样本

考虑由正常图像训练得到的目标分类模型 f，正常的输入图像是 x，攻击者试图找到一个对抗扰动 δ，使得 $x'=x+\delta$，其中，对抗扰动 δ 使得 x 跨越了分类模型 f 的决策边界导致 $f(x)\neq f(x')$，x' 即为对抗样本。对抗样本的形式化定义如下：

$$\text{Find AP } \delta$$
$$\text{s.t. } f(x')\neq f(x), x'=x+\delta, \|\delta\|_p<\varepsilon \tag{1}$$

下面对相关术语进行介绍：

1）对抗攻击和对抗防御。对抗攻击表示通过一定的算法在原输入图像上加入攻击噪声得到攻击图像，使分类器结果出错。攻击任务的难点在于攻击成功率和扰动大小之间的平衡，一般而言，扰动越大，攻击成功率越高。对抗防御表示构建足够鲁棒的分类器或防御模型，使其在输入攻击图像时也能够正确分类。

2）白盒攻击（white-box attack）和黑盒攻击（black-box attack）。根据攻击者对目标模型先验知识的掌握情况，对抗攻击可以分为白盒攻击和黑盒攻击。还有一种只在训练时利用目标模型先验知识的攻击，称为半白盒攻击。

3）有目标攻击（targeted attack）和无目标攻击（non-targeted attack）。根据对抗攻击是否设置目标结果，攻击被分为有目标攻击和无目标攻击。

4）评价指标。目前攻击算法的评价指标主要采用 L_p 距离（一般也称为 L_p 范数），如式（2）所示，其中，v_i 表示像素点 i 的像素值变化大小，n 表示像素点个数，v 一般通过攻击图像减原图像得到，表示两张图的差值，也就是扰动。目前常用的 L_p 距离包括 L_0、L_2 和 L_∞。

$$\|v\|_p=\left(\sum_{i=1}^{n}|v_i|^p\right)^{\frac{1}{p}} \tag{2}$$

1.2 Transformer 模型

ViT（Vision Transformer）[16]将纯 Transformer 架构直接应用到一系列图像块上进行分类任务，取得

了优异的结果，引发了 Transformer 在计算机视觉应用中的热潮。Transformer 强大的表达能力主要来源于自注意力机制的应用。计算机视觉中的自注意力机制借鉴自 NLP 的思想，保留了 Query、Key 和 Value 等名称。图 1 中的自注意力机制结构自上而下分为 3 个分支，分别是 Query、Key 和 Value（彩色效果见《计算机工程》官网 HTML 版，下同）。i 表示输入的卷积特征图，o 表示输出的自注意力特征图。计算时通常分为 3 步：第 1 步令 Query 和每个 Key 进行相似度计算得到权重；第 2 步使用 Softmax 归一化这些权重；第 3 步将归一化权重和 Key 相应的 Value 进行加权求和，得到最后的注意力特征图。输出的自注意力特征图如下（忽略了 1×1 卷积操作）：

$$o=\mathrm{Softmax}(f(i)\otimes g(i))\otimes h(i) \tag{3}$$

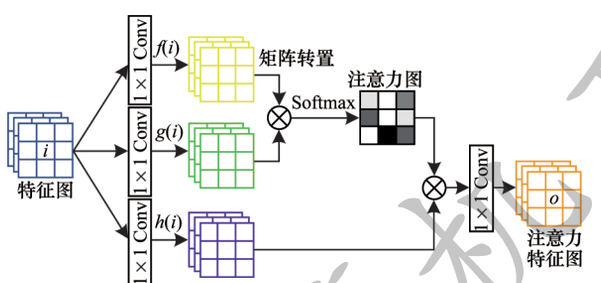


图 1 计算机视觉中的自注意力机制

Fig.1 Self-attention mechanism in computer vision

自注意力模块有助于模拟跨越图像区域的长距离、多层依赖关系。通过在原始特征图上添加加权的 attention 来获得特征图中任意 2 个位置的全局依赖关系。生成器可以利用自注意力模块来绘制细节充分协调的图像。

1.3 生成对抗网络

GAN 是一种无监督生成网络，包含生成、判别和对抗 3 个部分，利用对抗的思想交替训练生成器 G 和鉴别器 D。生成器 G 负责根据随机向量 z 生成尽可能真实的样本 $G(z)$，鉴别器负责判别接收的内容是否真实。鉴别器会输出一个概率，概率接近 1 表示鉴别器认为输入样本符合真实数据集的数据分布，接近 0 表示输入样本不符合真实数据分布。对抗是指 GAN 交替训练生成器 G 和鉴别器 D 的过程。以图片生成为例，生成器学习真实数据的分布，生成一些假样本，期望能够骗过鉴别器，然后鉴别器学习区分真样本和假样本，交替这一动态博弈过程，直至达到纳什均衡点，即鉴别器对任何图片的预测概率都接近 0.5，无法判别图片的真假，此时表示已经训练好生成器，可以停止训练。GAN 的优化目标如式（4）所示：

$$\min_G \max_D V(D,G)=\mathbb{E}_{x\sim P_{\mathrm{data}}(x)}[\log_a(D(x))]+\mathbb{E}_{z\sim P_z(z)}[\log_a(1-D(G(z)))] \tag{4}$$

其中：$P_{\mathrm{data}}(x)$ 表示真实样本数据的分布；$P_z(z)$ 表示随机向量的分布；$D(x)$ 表示 x 是真实图像的概率；$G(z)$ 表示从输入噪声 z 产生的生成图像。GAN 的优点是可以产生更加清晰、逼真的样本，缺点是训练不稳定，容易出现模式崩溃问题，即使长时间地训练生成器，生成效果依然可能很差。

2 本文算法设计与实现

受 Transformer 和 GAN 的启发，本文提出一种深度学习模型 Trans-GAN，利用 Trans-GAN 生成对抗样本，实现有目标攻击。

2.1 网络结构

如图 2 所示，Trans-GAN 网络结构包括生成模型 G、判别模型 D 和目标分类模型 f，其中，生成模型使用 Transformer 代替原始 GAN 中简单的卷积神经网络。网络以原始实例图像 x 和目标标签 t 作为输入，在模型中进行端对端的训练以生成对抗实例。

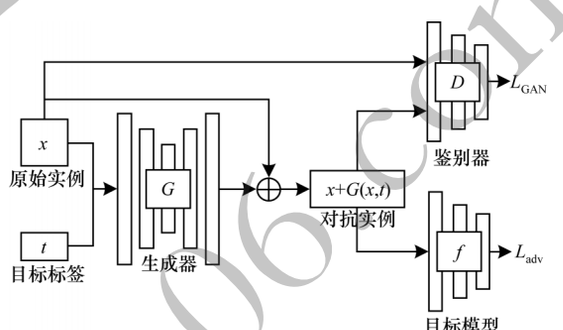


图 2 Trans-GAN 网络结构

Fig.2 Trans-GAN network structure

本文使用基于 Transformer 的 Restormer 模型[17]作为生成器网络结构，其在高分辨率图像复原领域取得了较高的性能。为了实现有目标攻击，将目标标签作为辅助信息参与到重建任务中，提出自注意力机制的修改版本 Targeted Self-Attention，改进后的生成器模型架构如图 3 所示。在图 3 中，生成器模型[图 3（a）]采用编码-解码结构，由若干个 Transformer Block 组成。模型接收干净样本 x 进行下采样，在编码器的每一层，同时接收由目标标签 t 变换到同等尺寸大小的目标特征图 t_n（n 取 1、2、3、4）作为输入，在充分提取到图像特征和目标信息后经解码器进行逐步上采样，恢复原始图像大小，输出噪声信息。下采样的每个 Transformer Block 中使用修改的自注意力机制 Targeted Self-Attention[图 3（c）]，上采样中每个 Transformer Block 使用原有的自注意力机制[图 3（b）]。在修改的注意力机制中，V 不再由 x 通过卷积操作得到，而是由输入的目标特征 t_n 产生，t_n 表示目标标签扩张维度后的特征，其尺寸大小和 x 相同，其余部分不变，这种简单操作将目标标签作为辅助信息，能有效指导生成器进行有目标攻击。

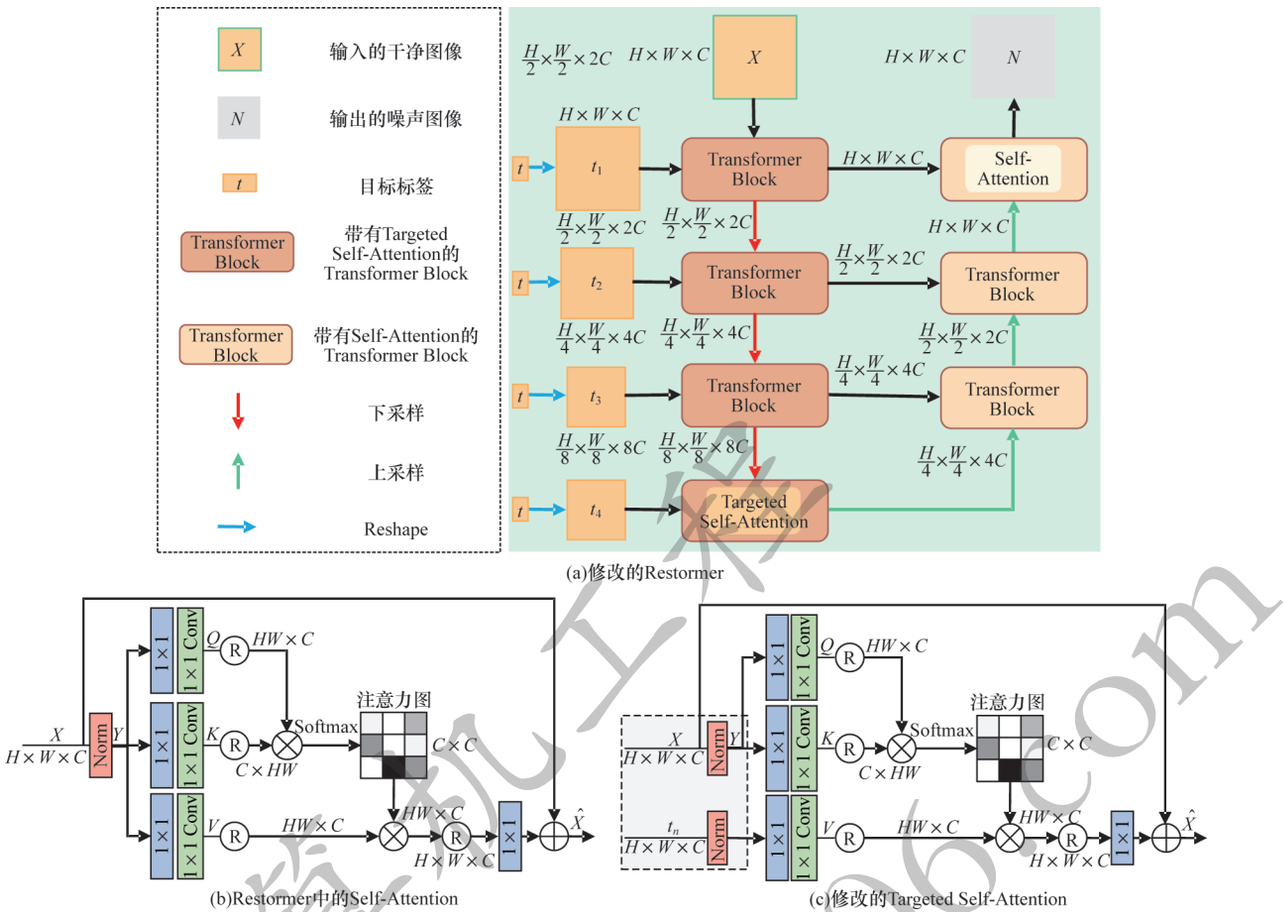


图 3 生成器模型架构

Fig.3 Generator model architecture

本文使用CNN作为鉴别器网络。表1所示为用于CIFAR10数据集的鉴别器架构(用于MNIST数据集的架构由于输入图片尺寸不同而稍有改变,具体详见公开代码)。在表1中,Conv表示卷积层,BN表示批归一化,激活函数使用斜率为0.2的LeakyReLU。鉴别器架构默认使用核窗口大小为4、步幅为2的卷积操作进行下采样,最后使用Sigmoid函数输出鉴别器的置信分数。

表 1 用于CIFAR10数据集的鉴别器架构

Table 1 Discriminator architecture for CIFAR10 dataset

层数	鉴别器架构
Layer 1	Conv(32,4,4)+LeakyReLU
Layer 2	Conv(64,4,4)+BN+LeakyReLU
Layer 3	Conv(128,4,4)+BN+LeakyReLU
Layer 4	Conv(1,4,4)+Sigmoid

2.2 损失函数

在训练过程中,Transformer接收原始干净图像和攻击目标类别后生成噪声信息,将残差连接与原始干净图像相结合生成扰动实例。判别模型和生成模型联合训练,使用L_{GAN}损失函数来提高模型的稳定性,L_{GAN}损失设计如式(5)所示,其中,G表示生成器,D表示鉴别器,x表示干净样本,t表示目标标签。

$$L_{\mathrm{GAN}}=\mathbb{E}_x\log_a D(x)+\mathbb{E}_x\log_a(1-D(x+G(x,t))) \tag{5}$$

生成模型和分类模型联合训练,使用L_{adv}^f损失函数来指导生成具有攻击性的对抗样本,L_{adv}^f损失设计如式(6)所示,其中,f表示目标分类器,t表示攻击目标类别。

$$L_{\mathrm{adv}}^f=\mathbb{E}_x\ell_f(x+G(x,t),t) \tag{6}$$

此外,通过在噪声信息上添加L_2约束$L_{\mathrm{perturbation}}$来控制所生成扰动的大小幅度,如式(7)所示,其中,$\|\cdot\|_2$表示L_2范数,即求欧氏距离。

$$L_{\mathrm{perturbation}}=\mathbb{E}_x\|G(x,t)\|_2 \tag{7}$$

为了提高生成模型的能力,改善对抗样本的视觉质量,采用结构相似度损失L_{ssim}[18]和感知损失$L_{\mathrm{perception}}$[19]进行联合训练,分别如式(8)、式(9)所示:

$$L_{\mathrm{ssim}}=1-\mathrm{SSIM}(x,a)=1-\frac{(2\mu_x\mu_a+c_1)(2\sigma_{x,a}+c_2)}{(\mu_x^2+\mu_a^2+c_1)(\sigma_x^2+\sigma_a^2+c_2)} \tag{8}$$

$$L_{\mathrm{perception}}^{\phi,j}(x,a)=\frac{1}{C_jH_jW_j}\|\phi_j(x)-\phi_j(a)\|_2 \tag{9}$$

其中:a表示生成的对抗样本,即$x+G(x,t)$;μ_x、μ_a、σ_x、σ_a和$\sigma_{x,a}$分别表示2个图像、图像的均值、方差以及协方差;c_1和c_2为常数,通常取$c_1=(K_1\cdot L)^2$,$c_2=(K_2\cdot L)^2$,一般地,K_1取0.01,K_2取0.03;L表示灰度级,取255;ϕ表示VGG网络[20];j表示网络的第j层;$\phi_j(\cdot)$和

$C_jH_jW_j$表示第j层的特征图和尺寸。

网络总体损失函数如式(10)所示，α、β、γ和λ为各部分损失函数的平衡参数，为了增强模型对扰动量的约束能力，上述参数默认设置分别为1、10、1和1，以提高所生成对抗样本的可视质量。

$$L=L_{\mathrm{GAN}}+\alpha L_{\mathrm{adv}}+\beta L_{\mathrm{perturbation}}+\gamma L_{\mathrm{ssim}}+\lambda L_{\mathrm{perception}} \quad (10)$$

在推理过程中，本文算法不再需要目标分类模型的参与，直接输入干净图像和攻击标签就能生成相对应的对抗样本，能够攻击目标分类器，使其分类错误。

2.3 算法流程

Trans-GAN算法训练流程如算法1所示。生成器根据干净样本x和目标标签t生成对抗扰动，计算出对抗样本a；将a输入目标分类器f中计算对抗损失，提升攻击能力；将a和x一起输入D中计算GAN损失，提升生成的对抗样本和干净样本的区分度；对x和a计算相似度损失、感知损失和L_2损失，进一步提高对抗样本的视觉效果。

算法1 Trans-GAN算法

输入 干净样本x，目标标签t

输出 对抗样本a

1.从训练集的干净样本中采样出一个mini-batch的数据x，并为x中的每个样本随机生成不同于其真实类别的目标标签t。

2.将x和t通过生成网络G得到对抗扰动，通过x+G(x, t)计算出对抗样本a。

3.根据式(10)计算损失，反向传播，更新生成器G和鉴别器D。

4.重复步骤1~步骤3直至算法收敛。

在测试阶段，不需要鉴别器D和目标分类器f的参与，只需向生成器G中输入测试集的干净样本以及想要攻击的目标类别标签，即可计算得出相应的对抗样本，从而实施相应的攻击。

3 实验结果与分析

本文所有实验均运行在Linux操作系统上，系统版本为Ubuntu 18.04 LTS，所使用的GPU资源为2块RTX 2080Ti，采用PyTorch 1.9.0作为深度学习框架，CUDA版本号为11.1。实验设置的总训练轮数为100个Epoch，采用Adam作为优化器，学习率设为0.000 1，权重衰减系数设为0.000 5。

首先评估所提算法在没有防御时对MNIST[21]和CIFAR10[22]的攻击效果，然后评估所提算法在有对抗训练防御下的攻击效果。此外，通过对ImageNet数据集[23]执行半白盒攻击，评估所提算法的可扩展性。实验过程中对抗性扰动的评价指标采用L_2和L_∞范式。MNIST的扰动幅度限制在0.3以内，CIFAR10和ImageNet的扰动使用L_2范式进行限制。使用不同攻击方法生成对抗性例子，以进行公平的比较。

从表2可以看出，在计算效率方面：FGSM生成对抗样本的运行时间为0.06 s；PGD在FGSM的基础上执行迭代攻击，多次计算梯度信息，因此运行时间是FGSM的数十倍；基于优化的攻击算法CW超过3 h，运行时间最长；本文所提算法和AdvGAN一样属于生成式攻击，运行时间远小于基于梯度的攻击算法和基于优化的攻击算法，在训练好模型参数后，只需进行一次简单的前向传播即可快速生成对抗样本。此外，FGSM和优化类算法只能进行白盒攻击，本文所提算法可以在半白盒设置下进行攻击，只在训练时需要模型参与。

表2 不同攻击算法的攻击效率比较

Table 2 Comparison of attack efficiency of different attack algorithms

攻击算法	运行时间
FGSM	0.06 s
PGD	0.7 s
CW	>3 h
AdvGAN	<0.01 s
本文算法	<0.01 s

与AdvGAN算法相比，本文算法接收目标标签作为输入，只需训练1个模型即可实现任意目标类别的有目标攻击，而AdvGAN的攻击目标在训练时的损失函数中指定，训练好的模型只能实现特定目标的攻击。如果将AdvGAN应用到ImageNet这种具有1 000个物体类别的数据集中，就需要针对每个目标类别共训练1 000个模型，而本文算法只需训练1个模型就能使目标分类器分错至任意目标类别，因此，本文算法具有更强的应用价值。

3.1 没有防御系统下的攻击评估

首先评估本文所提算法在没有任何防御设置下对MNIST和CIFAR10的攻击能力。对于MNIST数据集，在所有实验中为2个模型生成对抗性例子，2种模型网络结构如表3所示，模型A是文献[24]使用的架构，模型B是文献[9]用于评估基于优化策略的目标网络体系结构。对于CIFAR10数据集，选择ResNet-32[25]和Wide ResNet-34(WRN-34)[26]作为目标模型。

表3 用于MNIST数据集的2种分类模型结构

Table 3 Structure of two classification models for MNIST dataset

层数	模型A	模型B
Layer 1	Conv(64,5,5)+ReLU	Conv(32,3,3)+ReLU
Layer 2	Conv(64,5,5)+ReLU	Conv(32,3,3)+ReLU
Layer 3	Dropout(0.25)	MaxPooling(2,2)
Layer 4	FC(128)+ReLU	Conv(64,3,3)+ReLU
Layer 5	Dropout(0.5)	Conv(64,3,3)+ReLU
Layer 6	FC(10)+Softmax	MaxPooling(2,2)
Layer 7		FC(200)+ReLU
Layer 8		FC(10)+ReLU

表4中展示了在MNIST和CIFAR10测试数据集上使用AI-GAN和本文算法对各个目标类别进行攻击的实验结果，表中数值表示攻击成功率(ASR)，最优结果加粗标注。从表4可以看出，在不同数据集上，本文算法的攻击能力在大多数攻击类别上都要强于AI-GAN，并都取得了最高的平均攻击成功

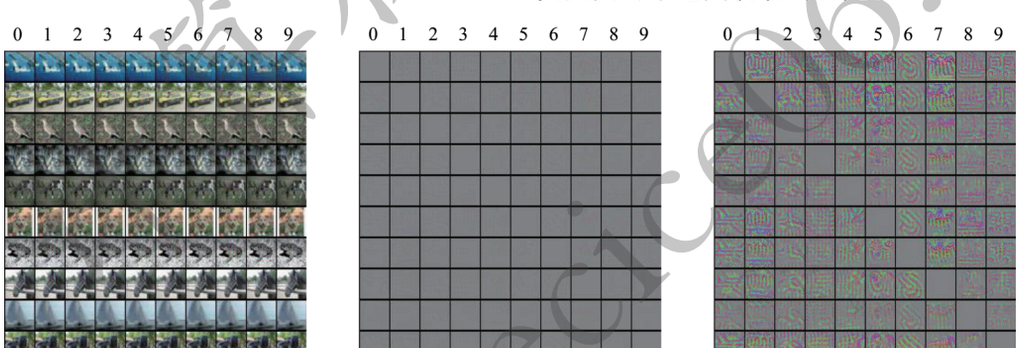
率。具体来说，在 MNIST 数据集上，本文算法的平均攻击成功率达到了 99.9% 以上，在 CIFAR10 数据集上平均攻击成功率也分别达到了 96.36% 和 98.47%，这主要得益于作为生成器的 Transformer 网络具有很强的学习能力。

表 4　在不同类别上攻击成功率的比较结果

Table 4　Comparison results of attack success rates on different categories　%

目标类别	MNIST				CIFAR10			
	模型 A		模型 B		ResNet-32		WRN-34	
	AI-GAN	本文算法	AI-GAN	本文算法	AI-GAN	本文算法	AI-GAN	本文算法
类别 1	98.71	**99.96**	99.45	**99.97**	95.90	**97.40**	90.70	**98.55**
类别 2	97.04	**99.92**	98.53	**99.60**	**95.20**	95.06	88.91	**98.26**
类别 3	**99.94**	**99.94**	98.14	**99.95**	95.86	**96.19**	93.20	**98.82**
类别 4	**99.96**	99.94	96.26	**99.95**	95.63	**95.99**	98.20	**98.94**
类别 5	99.47	**99.94**	99.14	**99.95**	94.34	**96.08**	96.56	**98.62**
类别 6	99.80	**99.90**	99.35	**99.90**	**95.90**	95.27	95.86	**97.30**
类别 7	97.41	**99.97**	99.34	**99.91**	95.20	**97.79**	**98.44**	97.93
类别 8	99.85	**99.93**	98.62	**99.98**	95.31	**95.64**	**98.83**	98.30
类别 9	99.38	**99.93**	98.50	**99.96**	95.74	**96.46**	98.91	**99.00**
类别 10	99.83	**99.90**	97.67	**99.94**	94.88	**97.75**	98.75	**98.96**
平均	99.14	**99.93**	98.50	**99.91**	95.39	**96.36**	95.84	**98.47**

针对 CIFAR10 数据集随机选择样本，生成的对抗性示例如图 4 所示。使用训练好的攻击算法从 CIFAR10 的 10 个类别中随机选择一张图像对其他类别进行攻击。图 4(a)表示生成的对抗样本示例，其中对角线上为原始干净图像，“0~9”表示攻击的目标类别，可以看出，从肉眼上基本看不出干净图像和对抗样本的区别。将生成器生成的对抗噪声进行可视化，如图 4(b)所示，可以看出，生成的扰动非常小，看不出任何细节。将噪声图放大 10 倍后如图 4(c)所示，从中可以看出明显的扰动细节。此外，在对不同图像进行相同目标的攻击时，生成器产生了纹理极为相似的噪声扰动，猜测这是由于在训练过程中同时接收了干净图像和目标标签作为输入，而针对 CIFAR10 数据集而言，训练集图像有 50 000 张，而目标类别只有 10 种，在训练过程中，目标标签容易占据主导地位，指导生成器生成与攻击类别相对应的特定扰动，这样的扰动施加在干净图像上可以促使目标分类器将关注点聚焦于对抗样本中的特定扰动，从而产生错误分类。上述这一现象可能推动相关人员对于通用扰动的研究。



(a)对抗样本　(b)对抗噪声　(c)对抗噪声放大10倍

图 4　CIFAR10 数据集的对抗样本示例

Fig.4　Examples of adversarial samples of CIFAR10 dataset

3.2　防御系统下的攻击评估

面对不同类型的攻击策略，学者们提出了不同的防御手段，其中，对抗训练被广泛认为是最有效的方法。文献[7]首先提出对抗性训练作为提高深度神经网络鲁棒性的有效方法，文献[24]将其扩展到集成对抗性学习，文献[8]也提出针对更强攻击方法的鲁棒网络。本文选择 3 种流行的对抗性训练方法来提高目标模型的鲁棒性：1）使用 FGSM 进行对抗性训练；2）集成对抗性训练；3）使用 PGD 进行迭代式对抗性训练。假设攻击者并不知道防御模型，并使用白盒设置中的普通目标模型作为目标，直接尝试攻击原始的学习模型，在这种情况下，如果攻击者仍然可以成功攻击模型，就表明攻击策略具有很强的鲁棒性。

首先不考虑任何防御，应用不同的攻击方法在原始模型的基础上生成对抗性样本，然后应用不同的防御方法来直接防御这些对抗性的实例以得到鲁棒模型。将本文所提算法与 FGSM、CW 攻击、PGD 攻击、AdvGAN、AI-GAN 一起对这些防御方法进行攻击，定量比较结果如表 5 所示。从表 5 可以看出，本文所提算法具有最高的攻击成功率，优于所有其他算法。与传统的攻击方法和 AdvGAN 相比，AI-GAN 和本文算法在 2 个数据集的不同模型、不同防御下都取得了更高的攻击成功率。与 AI-GAN 相比，本文所提算法攻击成功率高出 10.74%~55.14%，展现出其良好的攻击能力。

表5 在防御模型上不同算法的攻击性能比较

Table 5 Comparison of attack performance of different algorithms on defense models %

数据集	目标模型	防御模型	FGSM	CW	PGD	Adv-GAN	AI-GAN	本文算法
MNIST	模型A	Adv.	4.30	4.60	20.59	8.00	23.85	**36.92**
		Ens.	1.60	4.20	11.45	6.30	12.17	**29.11**
		Iter.Adv	4.40	3.00	11.08	5.60	10.90	**21.64**
	模型B	Adv.	2.70	3.00	10.67	18.70	20.94	**64.28**
		Ens.	1.60	2.20	10.34	13.50	10.73	**65.87**
		Iter.Adv	1.60	1.90	9.90	12.60	13.12	**63.91**
CIFAR10	ResNet-32	Adv.	5.76	8.35	9.22	10.19	9.85	**31.56**
		Ens.	10.09	9.79	10.06	8.96	12.48	**48.34**
		Iter.Adv	1.98	0.02	11.41	9.30	9.57	**29.10**
	WRN-34	Adv.	0.10	8.74	8.09	9.86	10.17	**37.29**
		Ens.	3.00	12.93	9.92	9.07	11.32	**26.58**
		Iter.Adv	1.00	0.00	9.87	8.99	9.91	**25.20**

3.3 ImageNet实验评估

为了评估本文所提算法生成高分辨率对抗样本的能力，针对复杂数据集进行相关实验，研究所提算法的有效性和可扩展性。从ImageNet-1000k数据集中随机选择100个类别，每个类别含有600张图像（输入大小为224×224像素）进行对抗攻击实验。为了减轻训练压力，在进行有目标攻击时，从100个目标中随机选择10个目标进行训练。在训练好攻击模型参数后，为测试数据集中的每个样本随机选择一个不同于其真实类别的目标标签，然后采用FGSM、PGD、本文算法进行有目标攻击测试，统计攻击所需总时长以及攻击成功率，实验结果如表6所示。从表6可以看出：虽然FGSM作为经典且简单有效的对抗攻击算法，在MNIST和CIFAR10数据集上有着较高的攻击能力和生成速率，但对于ImageNet这种较大尺寸和多目标类别的数据集，其攻击性能明显下降，有目标攻击的成功率只有28.76%；相比于FGSM，基于迭代的PGD攻击具有更强的攻击能力，但其生成对抗样本的时间也成倍增加；本文算法相比于PGD攻击不仅具有更高的攻击成功率，同时具有更快的生成速率。

表6 在ImageNet数据集上不同算法的攻击性能比较

Table 6 Comparison of attack performance of different algorithms on ImageNet dataset

攻击算法	攻击成功率/%	时间/s
FGSM	28.76	26.25
PGD	96.36	67.25
本文算法	97.33	53.62

本文算法的攻击性能不仅体现在高攻击成功率上，还体现在生成的对抗样本具有高可视质量。如图5所示，图中第1行表示从测试集中随机选取的干净样本，第2行表示使用本文算法生成的对抗扰动的可视图像，第3行表示施加了对抗扰动的对抗样本。由于训练过程中相似度损失函数和感知损失函数的约束，模型学习到了更高级的对抗扰动，能够改变原始图像的亮度（如第1列和第2列示例），或者改变原始图像的颜色（如第3列和第4列示例），或者改变原始图像的纹理特征（如第5列和第6列示例）。传统方法生成的对抗扰动通常表现为无规律的杂乱无章的噪声点，会严重影响干净样本的清晰度，而本文算法针对干净样本和目标标签生成的对抗样本看起来更加自然，更符合人类视觉的观感特性。

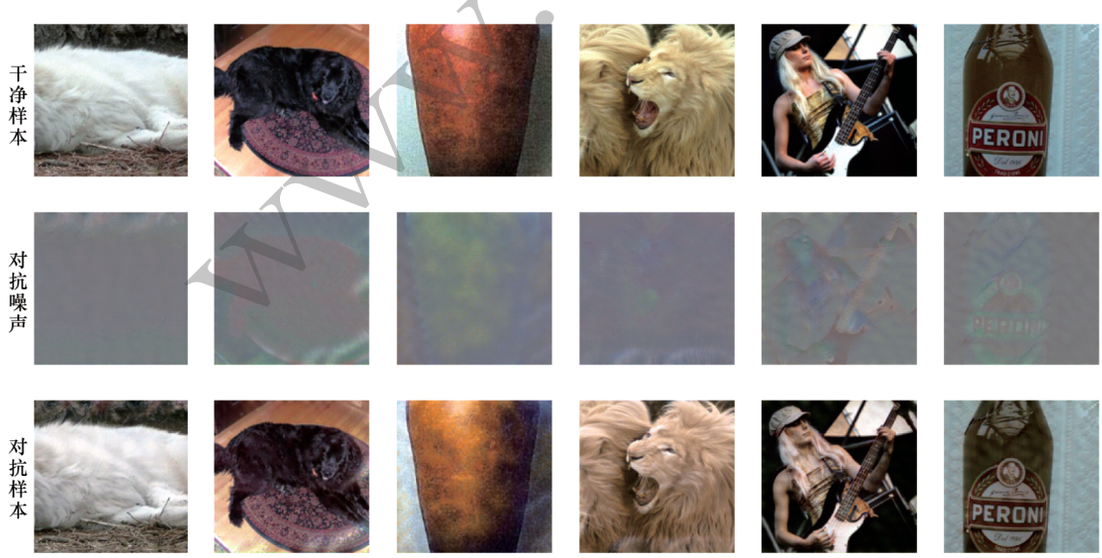


图5 ImageNet数据集的对抗样本示例

Fig.5 Example of adversarial samples of ImageNet dataset

4 结束语

本文提出一种基于 Transformer 和 GAN 的对抗样本生成算法 Trans-GAN。利用 Transformer 架构作为 GAN 的生成器，同时接收干净图像和目标标签作为输入，提出改进的注意力机制 Targeted Self-Attention，有针对性地生成对抗性扰动，并施加到干净图像上生成对抗样本。将生成网络与鉴别器进行联合训练，以实现快速收敛，同时使得对抗样本和干净样本不可区分。实验结果表明，相比 FGSM、PGD 等算法，Trans-GAN 算法生成的对抗样本可视质量更高，图像更为自然，且在各个攻击目标类别上具有较高的攻击成功率，有更好的迁移性能，能够满足复杂应用场景的需求。但是，通过定性实验结果发现，本文算法针对同一目标类别生成的对抗扰动具有相似的纹理特征，因此，下一步将研究如何快速有效地生成通用扰动。

参考文献

[1] 姜妍，张立国. 面向深度学习模型的对抗攻击与防御方法综述[J]. 计算机工程，2021，47(1)：1-11.
JIANG Y，ZHANG L G. Survey of adversarial attacks and defense methods for deep learning model[J]. Computer Engineering，2021，47(1)：1-11. (in Chinese)

[2] 白祉旭，王衡军. 基于改进遗传算法的对抗样本生成方法[J]. 计算机工程，2023，49(5)：139-149.
BAI Z X，WANG H J. Adversarial example generation method based on improved genetic algorithm[J]. Computer Engineering，2023，49(5)：139-149. (in Chinese)

[3] SZEGEDY C，ZAREMBA W，SUTSKEVER I，et al. Intriguing properties of neural networks[EB/OL]. [2023-02-05]. https://arxiv. org/abs/1312. 6199.

[4] DONG Y P，SU H，WU B Y，et al. Efficient decision-based black-box adversarial attacks on face recognition[EB/OL]. [2023-02-05]. https://arxiv. org/abs/1904. 04433.

[5] CARRARA F，FALCHI F，AMATO G，et al. Detecting adversarial inputs by looking in the black box[EB/OL]. [2023-02-05]. https://www. researchgate. net/publication/336809813_Detecting_Adversarial_Inputs_by_Looking_in_the_black_box.

[6] MA X J，NUI Y H，GU L，et al. Understanding adversarial attacks on deep learning based medical image analysis systems[J]. Pattern Recognition，2021，110：107332.

[7] GOODFELLOW I J，SHLENS J，SZEGEDY C. Explaining and harnessing adversarial examples[EB/OL]. [2023-02-05]. https://arxiv. org/abs/1412. 6572.

[8] MADRY A，MAKELOV A，SCHMIDT L，et al. Towards deep learning models resistant to adversarial attacks[EB/OL]. [2023-02-05]. https://arxiv. org/abs/1706. 06083. pdf.

[9] CARLINI N，WAGNER D. Towards evaluating the robustness of neural networks[C]//Proceedings of 2017 IEEE Symposium on Security and Privacy. Washington D. C.，USA：IEEE Press，2017：39-57.

[10] XIAO C，LI B，ZHU J，et al. Generating adversarial examples with adversarial networks[EB/OL]. [2023-02-05]. https://arxiv. org/abs/1801. 02610. pdf.

[11] JANDIAL S，MANGLA P，VARSHNEY S，et al. AdvGAN++：harnessing latent layers for adversary generation[C]//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision Workshop. Washington D. C.，USA：IEEE Press，2019：2045-2048.

[12] BAI T，ZHAO J，ZHU J L，et al. AI-GAN：attack-inspired generation of adversarial examples[C]//Proceedings of 2021 IEEE International Conference on Image Processing. Washington D. C.，USA：IEEE Press，2021：2543-2547.

[13] VASWANI A，SHAZEER N，PARMAR N，et al. Attention is all you need[C]//Proceedings of the 31st International Conference on Neural Information Processing Systems. New York，USA：ACM Press，2017：6000-6010.

[14] GOODFELLOW I，POUGET-ABADIE J，MIRZA M，et al. Generative adversarial networks[J]. Communications of the ACM，2020，63(11)：139-144.

[15] KRIZHEVSKY A，SUTSKEVER I，HINTON G E. ImageNet classification with deep convolutional neural networks[C]//Proceedings of International Conference on Neural Information Processing Systems. New York，USA：ACM Press，2012：1097-1105.

[16] DOSOVITSKIY A，BEYER L，KOLESNIKOV A，et al. An image is worth 16×16 words：transformers for image recognition at scale[EB/OL]. [2023-02-05]. https://arxiv. org/abs/2010. 11929. pdf.

[17] ZAMIR S W，ARORA A，KHAN S，et al. Restormer：efficient Transformer for high-resolution image restoration[C]//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Washington D. C.，USA：IEEE Press，2022：5718-5729.

[18] WANG Z，BOVIK A C，SHEIKH H R，et al. Image quality assessment：from error visibility to structural similarity[J]. IEEE Transactions on Image Processing，2004，13(4)：600-612.

[19] JOHNSON J，ALAHI A，LI F F. Perceptual losses for real-time style transfer and super-resolution[EB/OL]. [2023-02-05]. https://arxiv. org/abs/1603. 08155.

[20] SIMONYAN K，ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[EB/OL]. [2023-02-05]. https://arxiv. org/abs/1409. 1556. pdf.

[21] LECUN Y，BOTTOU L，BENGIO Y，et al. Gradient-based learning applied to document recognition[J]. Proceedings of the IEEE，1998，86(11)：2278-2324.

[22] KRIZHEVSKY A. Learning multiple layers of features from tiny images[EB/OL]. [2023-02-05]. https://www. semanticscholar. org/paper/Learning-Multiple-Layers-of-Features-from-Tiny-Krizhevsky/5d90f06bb70a0a3dced62413346235c02b1aa086.

[23] DENG J，DONG W，SOCHER R，et al. ImageNet：a large-scale hierarchical image database[C]//Proceedings of 2009 IEEE Conference on Computer Vision and Pattern Recognition. Washington D. C.，USA：IEEE Press，2009：248-255.

[24] TRAMÈR F，KURAKIN A，PAPERNOT N，et al. Ensemble adversarial training：attacks and defenses[EB/OL]. [2023-02-05]. https://arxiv. org/abs/1705. 07204. pdf.

[25] HE K M，ZHANG X Y，REN S Q，et al. Deep residual learning for image recognition[C]//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Washington D. C.，USA：IEEE Press，2016：770-778.

[26] ZAGORUYKO S，KOMODAKIS N. Wide residual networks[EB/OL]. [2023-02-05]. https://arxiv. org/abs/1605. 07146. pdf.

编辑 吴云芳