计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.210800130

基于生成对抗网络的多目标类别对抗样本生成算法

李 建 郭延明 于天元 武与伦 王翔汉 老松杨
国防科技大学系统工程学院 长沙 410073
(li_jian@nudt.edu.cn)

摘 要 深度神经网络在很多领域表现出色,但是研究表明其很容易受到对抗样本的攻击。目前针对神经网络进行攻击的算法众多,但绝大多数攻击算法的攻击速度较慢,因此快速生成对抗样本逐渐成为对抗样本领域的研究重点。AdvGAN是一种使用网络攻击网络的算法,生成对抗样本的速度极快,但是当进行有目标攻击时,其要为每个目标训练一个网络,使攻击的效率较低。针对上述问题,提出了一种基于生成对抗网络的多目标攻击网络MTA,在进行攻击时MTA仅需要训练一次就可以完成多目标攻击并快速生成对抗样本。实验结果表明,MTA在CIFAR10和MNIST数据集上有目标攻击的成功率高于AdvGAN。文中还做了对抗样本的迁移实验和防御背景下的攻击实验,结果表明,MTA生成的对抗样本的迁移性比其他多目标攻击算法更强,而且在防御背景下攻击成功率更高。

关键词:神经网络;对抗攻击;生成对抗网络;多目标攻击;对抗样本
中图法分类号 TP183

Multi-target Category Adversarial Example Generating Algorithm Based on GAN

LI Jian,GUO Yan-ming,YU Tian-yuan,WU Yu-lun,WANG Xiang-han and LAO Song-yang
College of Systems Engineering,National University of Defense Technology,Changsha 410073,China

Abstract Although deep neural networks perform well in many areas,research shows that deep neural networks are vulnerable to attacks from adversarial examples. There are many algorithms for attacking neural networks,but the attack speed of most attack algorithms is slow. Therefore,the rapid generation of adversarial examples has gradually become the focus of research in the area of adversarial examples. AdvGAN is an algorithm that uses the network to attack another network,which can generate adversarial samples extremely faster than other methods. However,when carrying out a targeted attack,AdvGAN needs to train a network for each target,so the efficiency of the attack is low. In this article,we propose a multi-target attack network(MTA) based on the generative adversarial network,which can complete multi-target attacks and quickly generate adversarial examples by training only once. Experiments show that MTA has a higher success rate for targeted attacks on the CIFAR10 and MNIST datasets than AdvGAN. We have also done adversarial sample transfer experiments and attack experiments under defense. The results show that the transferability of the adversarial examples generated by MTA is stronger than other multi-target attack algorithms,and our MTA method also has a higher attack success rate under defense.

Keywords Neural network,Adversarial attack,Generative adversarial network,Multi-target attack,Adversarial example

1 引言

深度神经网络在目标识别、自然语言处理以及图像分割等领域的应用取得了显著的成就,但是研究表明[1-2]深度学习模型很容易受到对抗样本的攻击。Szegedy等[1]最早提出了对抗样本的概念,并且发现了对抗样本的可迁移性,以及通过对抗训练可以提高模型的鲁棒性等重要性质。图像领域的对抗样本,即对输入图像添加一个人眼几乎无法识别的扰动,但是经过扰动的输入可以使神经网络的结果发生巨大的变化。Goodfellow等[2]将对抗样本的存在归因于模型的高维线性,对输入添加的扰动在模型的前向传播过程中对运算结果的改变像滚雪球一样越来越大。由于神经网络在现实生活中应用广泛,因此研究对抗样本对于人工智能安全具有重大意义。

自从Goodfellow等对对抗样本的存在作出解释后,很多针对图像分类网络的攻击算法被相继提出,如FGSM[2],C&W[3],Deepfool[4],Zoo[5]等。为了使对抗样本具有更好的视觉效果,当前大多数算法在实施攻击时通常在像素空间进行迭代优化,但生成对抗样本的效率很低,如C&W和Opt生成一个对抗样本需要数十秒。Xiao等[6]提出了一种快速对抗扰动的生成方法AdvGAN,该方法通过一个自动编码器产生对抗扰动,使用鉴别器网络保证了生成对抗样本的真实性。此时AdvGAN不再进行迭代优化,网络训练好后无需访问被

到稿日期:2021-08-15 返修日期:2021-10-12
通信作者:郭延明(guoyanming@nudt.edu.cn)

攻击网络和进行梯度的反向传播，利用神经网络批量输入、显卡运算加速等特性可以快速地进行前向运算，生成对抗样本，其速度是 C&W 的数千倍。

如图 1 所示，AdvGAN 虽然能快速生成对抗样本，但是当进行有目标攻击时，需要针对每个目标类别分别训练一个模型，在攻击时根据攻击目标选择具体的模型实施攻击；在实验过程中我们发现，AdvGAN 在进行黑盒环境下的有目标攻击时的成功率较低。因此，本文提出一种基于生成对抗网络的多目标攻击对抗样本生成算法 MTA（Mutli-Target AdvGAN），该模型在训练时融入随机目标信息，生成使原始图像分类为目标类别的扰动。MTA 同样使用鉴别器来保证对抗样本的质量，使用生成器生成对抗噪声，模型训练结束后，攻击者为模型输入原始图像和攻击目标即可生成指定目标类别的对抗样本。MTA 使用比 AdvGAN 更短的训练时间就能达到与之相同甚至更优的攻击效果。本文在 CIFAR10 和 MNIST 数据集上针对不同的模型做了攻击实验，结果表明，MTA 有目标攻击的成功率都高于 AdvGAN。在 ImageNet 数据集上的实验进一步验证了我们方法的有效性。另外，在不同模型之间做了对抗样本的迁移实验和在防御背景下的攻击实验，结果表明，与 FGSM[2]，PGD[7] 等多目标攻击算法相比，本文方法具有更高的迁移成功率和防御背景下的攻击成功率。

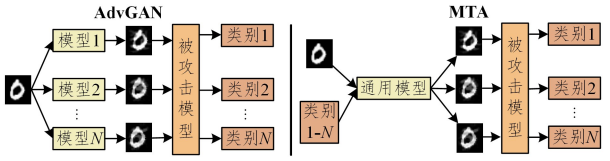


图 1 AdvGAN 和 MTA 生成对抗样本的过程

Fig. 1 Process of AdvGAN and MTA to generate adversarial examples

本文的主要贡献如下：

(1)MTA 方法在增加少量参数的情况下，只需要训练一个通用的模型即可在白盒和黑盒环境下进行任意类别的有目标攻击。在白盒环境下可以达到与 AdvGAN 相近甚至更高的攻击成功率；在黑盒环境下的攻击成功率也显著高于 AdvGAN。为此还做了进一步的实验，通过分析实验结果我们发现，代理模型与被攻击模型在结构上的差异对 AdvGAN 攻击成功率的影响较大，而 MTA 方法受此影响相对较小。

(2)MTA 算法通过网络学习如何添加一个更加普适的噪声，相较于其他基于人工策略的多目标攻击算法，其生成的对抗样本具有更高的迁移攻击成功率和防御背景下的攻击成功率。

2 相关工作

2.1 对抗样本

通过在原有图像中添加特殊的噪声扰动，使某个分类网络识别错误，此时添加了噪声的图像被称为对抗样本，添加到图像中的噪声被称为对抗扰动。对抗样本可以使分类网络分类错误，但在人类的视觉中原图像和对抗样本差别很小，甚至无法用肉眼区分。对抗样本的数学描述如下：对于某个被攻击的神经网络 F，给定一个图像 x 和该图像的标签 y_0，攻击算法的目的是寻找一个与原图像具有相同形状的噪声 δ，使得原图像加上该噪声后输入到被攻击网络中分类的标签变成 y_1，且 $y_0 \neq y_1$。P 是对对抗扰动的范数约束，通常 $P=0,2,\infty$，其中，0 范数约束的是对抗扰动中不等于 0 的像素的个数，2 范数约束的是对抗扰动的模长，∞ 范数约束的是对抗扰动的最大值。

$$\begin{aligned} &\min \ \|\delta\|_p \\ &\text{s.t.} \ \ y_0=F(x), y_1=F(x+\delta), y_0 \neq y_1 \end{aligned} \tag{1}$$

按照攻击者对被攻击模型的先验知识，可以将攻击划分为白盒攻击和黑盒攻击。按照攻击者的攻击目的，可以将攻击行为划分为无目标攻击和有目标攻击。下文将介绍白盒和黑盒环境下当前主流的几种有目标攻击方法。

2.2 白盒环境下的有目标攻击

白盒攻击指攻击者可以完全访问被攻击网络，得到网络预测的各个类别回归值并且经过反向传播计算梯度。白盒攻击大多直接或间接地利用梯度指导对抗噪声的添加，因此效率较高且扰动量较小。FGSM 算法由 Goodfellow 等[2]提出，是最直接的一阶梯度攻击，利用模型的近似梯度产生对抗样本。在进行有目标攻击时，给定被攻击样本 x 和目标类别 t，FGSM 首先计算损失值对输出结果 t 在图像上的梯度；然后让输入图像减去这个梯度，从而使分类结果向目标类别发生倾斜；最后进行单步攻击，产生具有 L_∞ 约束的对抗样本。

虽然 FGSM 产生对抗样本的速度较快，但是其攻击成功率并不高。Madry 等[8]和 Kurakin 等[9]分别在 FGSM 基础上改进并提出 PGD 攻击算法和 I-FGSM 算法，其原理都是在生成对抗样本的过程中加入迭代，改变 FGSM 单步攻击就生成对抗样本的方式，以较小的迭代步长多次添加噪声并进行剪裁，确保对抗样本位于原始图像的 ε 邻域内。使用迭代的 FGSM 算法能在相同攻击强度下提升攻击成功率，但是迭代进行反向传播添加噪声的行为又不可避免地降低了生成对抗样本的速度。

此外，为了获得更好的视觉效果，Carlini 等提出基于目标函数优化的 C&W[3] 算法。C&W 的目标是使 $\|\delta_p + c \cdot f(x+\delta)$ 最小，前半部分 $\|\delta\|_p$ 是对抗样本的范数约束，后半部分 $f(x+\delta)$ 是对抗样本的有效性约束，c 是两个目标之间的均衡。因为两个目标之间具有对抗性，所以权重因子 c 很难确定。C&W 在优化的同时，通过二分法搜索确定 c，且每次只能优化一个实例，导致其优化速度非常慢。因此，C&W 虽然被公认为当前生成对抗样本在视觉效果上最好的算法，但很难投入到实际应用中。其他类似的优化算法，如 Deepfool[4] 和 DDN[10] 等，虽然相比 C&W 有了较大改进，但是依然存在迭代优化、生成速度较慢的问题。

相比之下，我们提出的 MTA 方法使用网络的前向过程生成对抗样本而不进行任何优化，在相同的扰动强度下比 FGSM 的攻击成功率更高，且生成速度是 C&W 方法的数千倍。

2.3　黑盒环境下的有目标攻击

黑盒攻击指攻击者无法访问模型的参数，仅能通过向模型输入得到模型的预测结果，此结果为某一个具体的类别。现实中实施的攻击场景通常是黑盒，因此研究黑盒攻击更有意义。黑盒攻击大多通过训练代理模型，利用对抗样本的迁移特性或通过边界搜索的方法实施。Szegedy 等[1]的研究表明，对抗样本在不同模型之间具有迁移性，即对抗样本可以同时攻击多个在相同数据集上训练的模型。因此，使用对抗样本迁移攻击是黑盒的常用方法，首先使用白盒攻击方法攻击一个与被攻击模型使用相同数据集训练的代理模型，然后使用攻击代理模型产生的对抗样本去攻击黑盒模型。虽然此方法在理论上具有可行性，但是其攻击成功率很低。

具有代表性的边界搜索算法有 Boundary attack[11]，Opt[12]以及 Sign-Opt[13]，在进行有目标攻击时它们都分两步进行。第一步，通过查询从训练集中挑选出一个分类结果为目标类别的图像，然后从该图像出发通过二分法查找到被攻击图像与目标图像之间的决策边界。第二步，通过优化策略逐渐缩小决策边界与原始图像的距离。Boundary attack，Opt 以及 Sign-Opt 的第一步操作过程相同，但是优化过程采用了不同的策略。在相同 L2 约束的条件下，Sign-Opt 是三者中需查询次数最少的搜索算法。

通过边界搜索的方式能够实现较高的有目标攻击成功率，但是同样需要大量的查询和优化过程，且每次只能对一个实例进行运算，导致生成对抗样本的速度极慢。相比之下，MTA 在黑盒环境下无需进行优化就可以快速地生成对抗样本。

2.4　AdvGAN

虽然对抗攻击算法众多，但目前大多是基于某种设定好的策略迭代优化对抗扰动，这使得生成对抗样本的速度较慢。使用 GAN 生成逼真的样本是很多生成任务[6,7]中常用的方法，Xiao 等[6]提出使用前馈网络自主学习攻击网络的方法 AdvGAN，AdvGAN 是一个基于生成对抗网络的模型，能快速生成对抗样本且保持较高的攻击成功率。它针对每一个被攻击模型训练若干生成器网络，生成器为输入图像量身定做对抗扰动，并且为了使对抗样本更加真实，使用鉴别器监督生成器生成的噪声。模型训练结束后，直接利用生成器生成输入图像的对抗扰动。该对抗样本可以快速批量生成，而且生成过程中不需要访问被攻击网络。在进行黑盒攻击时，Xiao 等[6]使用动态蒸馏的方法，有效提升了黑盒攻击的成功率。但是，AdvGAN 在进行有目标攻击时，需要为每一个目标类别训练一个生成器，导致有目标攻击的灵活性较差；并通过实验发现，AdvGAN 在黑盒场景下有目标攻击的成功率较低。

2.5　对抗防御

针对不同的攻击方法，研究者们提出了很多防御方法[14-22]，这些方法总体可以分为检测性防御和鲁棒性防御。检测性防御即只识别一个实例是否为对抗样本，如果是对抗样本就拒绝该样本访问网络，代表性的方法有 Feature Squeezing[14]和 MagNet[15]。鲁棒性防御即通过防御措施降低模型的受威胁程度，提升模型受到对抗样本攻击时的准确率，其目标是将对抗样本重新分类为正确的类别。鲁棒性防御最常用且高效的方法是对抗训练。为了抵御对抗样本的攻击，Goodfellow 等[2]首次提出了对抗训练（Adv），该方法使用原始训练集和对抗样本训练被攻击模型，增强模型的鲁棒性。模型的损失计算方法如式(2)所示：

$$Loss=\alpha\cdot l_f(x,y_{\text{true}})+(1-\alpha)\cdot l_f(x_{\text{adv}},y_{\text{true}}) \tag{2}$$

其中，对抗样本由 FGSM 方法生成；超参数 α 用来均衡正常样本和对抗样本的重要程度，通常取值 0.5。对抗训练不仅让模型学会了区分原始样本，还学会了正确地区分对抗样本。

Tramèr 等[16]提出聚合对抗训练（Ens），即通过聚合在相同训练集的多个模型上产生的对抗样本，加上原始训练集进行对抗训练。该模型在训练时的损失计算的方法如式(3)所示：

$$Loss=\alpha\cdot l_f(x,y_{\text{true}})+(1-\alpha)\cdot l_f(x^i_{\text{adv}},y_{\text{true}}) \tag{3}$$

其中，x^i_{adv} 表示攻击算法攻击第 i 个模型产生的对抗样本。

对抗训练虽然被公认为最简单有效的防御手段，但其效率低下。以 CIFAR10 数据集为例，在进行对抗训练时，每个 epoch 的训练过程中都要生成 5 万个对抗样本，如果训练 100 个 epoch，那么整个训练过程就有 500 万个对抗样本生成。因此研究一种快速生成对抗样本的方法，对于防御手段的应用也具有重大的意义。此外，使用预处理的方法对网络的输入进行去噪、滤波、重建，也能有效地抵御对抗样本的攻击，并且处理速度很快。Das 等[17]经过实验验证得出，使用 JPEG 压缩可以有效地防御有目标攻击。本文将在实验中验证 MTA 攻击方法在这些防御方法下的有效性。

3　MTA 攻击方法

如图 2 所示，MTA 网络主要由 3 个步骤组成，即特征提取、特征融合和样本生成。

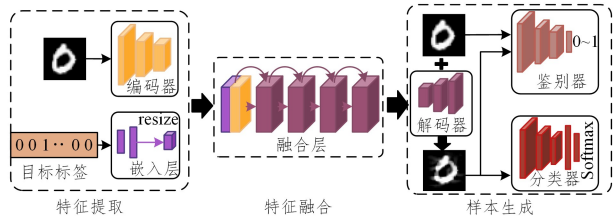


图 2　MTA 网络结构

Fig. 2　Network structure of MTA

(1)特征提取。给定输入图像 x，编码器通过 3 个卷积层得到图像的三维特征图 $F_{\text{img}}\in R^{c\times w\times h}$，其中 c，w，h 分别表示特征图的通道、宽度和高度。嵌入层通过全连接将目标类别的 one-hot 标签映射到更高维度的特征向量，之后对特征向量进行平铺得到三维的特征向量 $F_{\text{target}}\in R^{c\times w\times h}$。编码层和嵌入层的结构设置如表 1 所列，其中 $X_1=8$，$X_2=16$，$X_3=32$。在不同的数据集上嵌入层的编码长度是不同的，例如 CIFAR10 数据集上输入图像经过卷积后的特征图尺寸为 $32\times4\times4$，设置嵌入层的编码长度为 512；在 MNIST 数据集上特征图大小是 $32\times5\times5$，嵌入层的编码长度是 800，经过平铺后标签特征与图像特征大小相同。

表 1 MTA 网络的生成器参数

Table 1 Generator parameters of MTA network

Module	Layer	CIFAR10				MNIST			
		Parameter	Stride	Padding	Activation	Parameter	Stride	Padding	Activation
编码器	Convolution	$3\times3\times X_1$	2	1	ReLU	$3\times3\times X_1$	2	1	ReLU
	Convolution	$3\times3\times X_2$	2	1	ReLU	$3\times3\times X_2$	2	1	ReLU
	Convolution	$3\times3\times X_3$	2	1	ReLU	$3\times3\times X_3$	1	0	ReLU
嵌入层	Linear	10×512	—	—	—	10×800	—	—	—
融合层	ResNet Block	$3\times3\times X_4$	1	1	—	$3\times3\times X_4$	1	1	—
	ResNet Block	$3\times3\times X_4$	1	1	—	$3\times3\times X_4$	1	1	—
	ResNet Block	$3\times3\times X_4$	1	1	—	$3\times3\times X_4$	1	1	—
	ResNet Block	$3\times3\times X_4$	1	1	—	$3\times3\times X_4$	1	1	—
解码器	ConvTranspose	$2\times2\times X_3$	2	0	ReLU	$3\times3\times X_3$	1	0	ReLU
	ConvTranspose	$2\times2\times X_2$	2	0	ReLU	$2\times2\times X_2$	2	0	ReLU
	ConvTranspose	$2\times2\times X_5$	2	0	Tanh	$2\times2\times X_6$	2	0	Tanh

(2)特征融合。由于特征提取过程中得到的两个特征图具有相同的高度和宽度，因此首先将两个特征图在通道层进行拼接$F_{\text{img}} \| F_{\text{target}}$，得到新的特征图$F\in R^{2c\times w\times h}$；再对新的特征图使用 4 个残差卷积模块，在每一个残差卷积块中使用大小为 3×3、步长为 1 且 Padding 为 1 的 64 个卷积核，因此融合层只对两部分特征图进行解析而不改变特征图的大小和通道。

(3)样本生成。解码器的结构设置如表 1 所列，其中$X_5=3$，$X_6=1$。解码器将经过解析的特征图F作为输入，通过上采样生成一个与原始图像具有相同大小和通道数的噪声$G(x,t)$。为了控制噪声的强度，我们在得到噪声之后对噪声进行剪裁，与 AdvGAN 相同，将噪声的最大改变量控制在 0.3 以内。解码器生成的噪声加上原始图像，即得到对抗样本，对得到的对抗样本进行剪裁，以确保像素值在有效范围之内。特征提取、特征融合以及解码器生成噪声的过程共同组成了一个生成器，它能根据原始图像x和目标类别t生成一个特定的噪声。经过上述步骤得到的对抗样本将被分别送入鉴别器和分类器中，鉴别器以原始图像和对抗样本作为输入，输出图像为真实图像的置信度。鉴别器的目的是区分原始图像和对抗样本，通过对抗性训练迫使生成器生成的对抗样本更加逼真，以骗过鉴别器。分类器即要攻击的目标模型，若是在黑盒环境下的攻击，则分类器为代理模型。分类器仅以对抗样本作为输入，输出对抗样本在所有类别上的回归值计算损失。

对 MTA 模型的训练与普通生成对抗网络的不同之处在于：对生成器的约束不仅限于能骗过鉴别器，还要能骗过目标分类器。MTA 的训练分为两个连续的过程，首先是训练一个普通的 GAN，目标函数如式(4)所示：

$$\min_G \max_D V(D,G)=E_x\log D(x)+E_x\log\left(1-D(x+G(x,t))\right) \tag{4}$$

其中，$D(*)$表示鉴别器的输出，即鉴别器认为输入图像是真实图像的概率值。通过最小最大化目标函数来确保生成高质量的对抗样本。

其次还要对生成器的输出做有效性约束和范数约束。有效性约束即生成的对抗样本要在被攻击模型中被分类为输入的目标类别。我们使用损失函数来鼓励生成器欺骗目标模型，如式(5)所示：

$$L_{\text{adv}}=E_x\max_{i\neq t}\{k,\max\{Z(x+G(x,t))_i-Z(x+G(x,t))_t\}\} \tag{5}$$

其中，$Z(*)$为被攻击模型的回归值，t为目标类别，i为除了目标类别以外的其他所有类别。MTA 与 AdvGAN 的一个不同之处在于，AdvGAN 在训练时要固定t，因此要为每个目标类别训练一个模型；而 MTA 模型在训练时，t是在所有类别中随机采样且不等于真实类别的，当对抗样本在目标类别上的回归值比其他类别上最大的回归值大且超过阈值k时，该项损失为 0，这保证了生成器生成的对抗样本是我们指定的目标类别。

为了限制扰动的幅度，对对抗扰动进行范数约束，鼓励对抗样本在样本空间接近真实样本。Xiao 等[6]验证了如式(6)所示的 soft hinge loss 损失函数更适合他们的模型，其中c取 0.3。因此，我们仍沿用该损失函数。最后，用超参数β控制有效性约束和范数约束的权重。生成器的损失值计算公式如式(7)所示：

$$L_{\text{pert}}=E_x\max(0,\|G(x,t)\|_2-c) \tag{6}$$

$$Loss=L_{\text{adv}}+\beta*L_{\text{pert}} \tag{7}$$

4 实验

4.1 实验设置

4.1.1 数据集

本文使用 3 个常用于评价对抗样本工作的数据集：CIFAR10，MNIST，ImageNet。CIFAR10 数据集由 5 万张训练图像和 1 万张测试图像组成，每张图都是长和宽均为 32 像素的彩色图像，共包含 10 个生活中常见的物体类别。MNIST 数据集由 6 万张训练图像和 1 万张测试图像组成，其图像内容均是黑白的手写数字，包含了从 0 到 9 的 10 个类别，图像的长和宽均为 28 个像素。对于 ImageNet，我们使用了一个包含 10 个类别的子集，训练集中每个类别包含约 1 000 张图像，测试集中每个类别包含约 300 张图像。

4.1.2 模型及模型训练

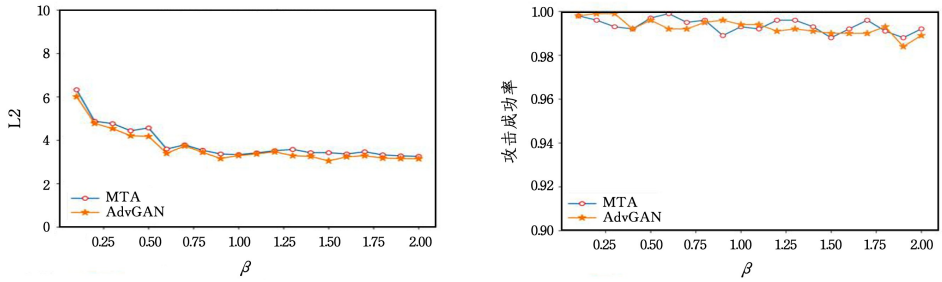
对模型的训练主要分为两个部分，一是训练被攻击模型，二是训练攻击模型。

(1)训练被攻击模型。我们首先训练了一批被攻击模型，模型训练结束后仅用于测试攻击模型。在 CIFAR10 数据集上，我们训练了常用的 VGG11，ResNet18 以及 WRN-28 模型。由于 MNIST 数据集中的图像较小，无法使用上述网络，因此在 MNIST 数据集上我们训练了常用的且有足够表达能力的网络 LeNet，AlexNet，C&Wmodel[3]（Carlini 和 Wagner 在验证 C&W 攻击方法时使用的 MNIST 分类模型）。训练过程中，设定批大小为 256，学习率为 0.001，每个模型使用 Adam 优化器训练了 100 个 epoch。在 ImageNet 数据集上，使用迁移学习的方法训练了一个具有高准确率且更适用于 10 个类别子集的模型 VGG16，用于验证本文方法在高分辨率数据集上的有效性。

(2)训练攻击模型。对于 MTA 和 AdvGAN 模型，在训练时，我们设置批大小为 128，使用 Adam 优化器以 0.001 的初始学习率训练 200 个 epoch，并分别在第 50 个和第 100 个 epoch 后将学习率降为之前的 1/10。

4.2　超参数设置

因为 MTA 方法和 AdvGAN 方法的损失函数中都包含超参数，所以本文对两部分约束的权重作了消融分析。我们使用 MTA 和 AdvGAN 攻击了 CIFAR10 数据集上的第一个网络 VGG11，在使用 AdvGAN 攻击时设置目标类别为 0。令 β 从 0.1 变化到 2，实验结果如图 3 所示。



(a)扰动的 L2 范数与 β 的关系　(b)攻击成功率与 β 的关系

图 3　参数 β 对两种攻击方法性能的影响

Fig. 3　The influence of parameter β on the performance of two attack methods

图 3(b)为攻击成功率与 β 的关系图，可以看出，两种方法的攻击成功率基本都保持在 99% 以上，且受参数 β 的影响较小。图 3(a)为对抗扰动的 L2 范数与 β 的关系图，可以看出，对抗扰动随着 β 的增加逐渐减小，但是当 β 大于 1 后，两者都趋于平稳，因此在接下来的所有实验中都设置 $\beta=1$，这样既保证了两种方法的攻击成功率，又确保了生成的对抗样本具有良好的视觉效果。

4.3　白盒攻击

本文在不同的分类模型上评估了 MTA，首先在 MNIST 数据集上攻击了经过预训练的 LeNet，AlexNet 以及 C&Wmodel 模型。为了与 AdvGAN 作对比，使用 AdvGAN 分别为 3 个被攻击模型训练了 10 个生成器，除了不使用嵌入层外，AdvGAN 与 MTA 使用相同的网络结构。在进行攻击能力评估时，仅以被攻击模型分类正确的图像作为输入，且攻击目标选择非正确标签。若经过生成器添加扰动后的对抗样本在被攻击模型中的分类结果为我们的指定目标类别，则认为此次攻击有效。

我们以各个被攻击模型在测试集中的前 100 张分类正确的图像作为输入来生成 900 个对抗样本，由于 AdvGAN 每个目标都需要一个模型，因此文中列出了 AdvGAN 在每个目标类别上的攻击成功率，最后取平均值与 MTA 方法进行对比。攻击结果如表 2 所列，表格中第二列为模型的分类准确率，虽然 AdvGAN 每次仅向 1 个目标攻击，使训练的模型更具有针对性，但是由于本文方法在生成对抗扰动时有目标标签的特征信息进行指导，因此 MTA 方法仅训练一个模型就可以达到比 AdvGAN 更高的攻击成功率。另外我们发现，被攻击网络的鲁棒性越高就越难被攻击，其中 C&Wmodel 的准确率最高，因此无论使用哪一种攻击方法，在攻击 C&Wmodel 时的成功率都比其他两个分类模型低。此外，鲁棒性越高的网络，MTA 方法的优越性越明显，例如在攻击 AlexNet 网络时，与 AdvGAN 相比，MTA 仅提升了约 0.6% 的成功率，但在攻击 C&Wmodel 时，却提升了约 3.8% 的成功率。

表 2　MNIST 数据集上的白盒攻击成功率

Table 2　The white box attack success rate on the MNIST dataset

（单位：%）

Model	*Acc*	AdvGAN										MTA
		0	1	2	3	4	5	6	7	8	9	Mean	
LeNet	98.7	96.6	92.9	98.2	98.2	98.2	97.6	96.9	97.8	98.4	99.0	97.38	**98.8**
AlexNet	98.6	98.7	98.1	98.6	98.6	99.4	99.1	98.9	98.9	99.6	98.4	98.83	**99.4**
C&Wmodel	99.3	94.9	88.8	65.7	99.1	97.6	87.1	94.4	98.9	99.4	98.9	92.48	**96.2**

我们进一步在 CIFAR10 数据集上攻击了 VGG11，ResNet18 以及 WRN-28，其攻击成功率如表 3 所列。可以看出，MTA 取得了与 AdvGAN 相似的攻击效果，但是鉴于 MTA 在进行有目标攻击时只需要训练一个模型，而 AdvGAN 需要为每个目标单独训练模型，因此我们认为 MTA 比 advGAN 更具实用性。

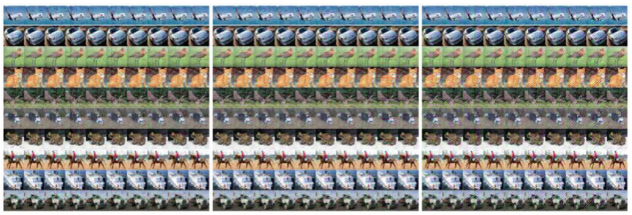
表 3 CIFAR10 数据集上的白盒攻击成功率

Table 3 The white box attack success rate on the CIFAR10 dataset

(单位:%)

Model	Acc	AdvGAN										MTA
		Plane	Car	Bird	Cat	Deer	Dog	Frog	Horse	Ship	Truck	Mean	
VGG11	87.6	99.1	99.6	99.7	99.7	99.7	99.6	99.7	99.0	99.7	99.4	**99.52**	99.3
ResNet18	89.2	99.8	99.8	99.6	99.4	99.8	99.1	99.5	99.5	99.5	99.8	99.58	**99.6**
WRN-28	83.5	99.4	99.3	99.4	99.7	99.2	99.7	99.6	99.1	99.1	99.6	99.40	**99.6**

图 4 为白盒环境下 MTA 方法在 CIFAR10 数据集上生成的对抗样本图,我们从 10 个类别的原始图像中随机各选出一张,并输入到 MTA 模型中生成 10 个目标类别的对抗样本,为了便于对比,每个子图的对角线上显示的是原始图像,其中图 4(a)、图 4(b)、图 4(c)分别为攻击 VGG11,ResNet18,WRN-28 模型的结果。由对抗样本图可以看出,在鉴别器和 L_{pert} 损失函数的约束下,MTA 生成的不同目标对抗样本保持了与对角线上的原始图像有相似的视觉效果。



(a)攻击 VGG11 产生的对抗样本 (b)攻击 ResNet18 产生的对抗样本 (c)攻击 WRN-28 产生的对抗样本

图 4 MTA 方法在白盒环境下产生的对抗样本

Fig. 4 Adversarial examples generated by the MTA method in white-box environment

4.4 黑盒攻击

在进行黑盒攻击时,使用与 AdvGAN 相同的动态蒸馏方法,即在攻击的过程中训练代理模型,要求代理模型不仅要在原数据上的分类结果与被攻击模型保持一致,而且在对抗样本上的分类结果也要与被攻击模型保持一致。在 MNIST 数据集上,分别使用 MTA 和 AdvGAN 方法攻击模型 LeNet,AlexNet 以及 C&Wmodel,当一个模型被攻击时,其余模型轮流作为代理模型,攻击成功率的计算方法与白盒相同。MNIST 数据集上的攻击成功率如表 4 所列,在黑盒环境下,MTA 方法相比 AdvGAN,其在攻击成功率上的优越性更加明显,例如,当以 AlexNet 作为代理模型攻击 LeNet 模型时,使用 MTA 方法的攻击成功率比 AdvGAN 高出约 24%(94.8% vs 70.44%),以 LeNet 作为代理模型攻击 C&Wmodel 时,MTA 方法的攻击成功率高出约 22%(94.1% vs 72.51%)。虽然 AdvGAN 在黑盒模型上的攻击成功率不高,但其攻击难度与 MTA 方法相当,例如,无论使用 MTA 还是 AdvGAN,对于黑盒模型 LeNet 使用 C&Wmodel 作为代理模型都比使用 AlexNet 的成功率高,其他黑盒模型也出现了同样的结果。这可能与被攻击模型和代理模型之间的结构差异有关,但也证明了 MTA 方法在攻击能力上优于 AdvGAN 方法。

利用同样的方法,我们在 CIFAR10 数据集上攻击了 VGG11,ResNet18 以及 WRN-28 这 3 个模型。攻击成功率如表 5 所列,可以看出,MTA 方法始终优于 AdvGAN,值得一提的是,当 WRN-28 作为代理模型时,MTA 方法较 AdvGAN 的攻击成功率提升了约 20%。

表 4 MNIST 数据集上的黑盒攻击成功率

Table 4 Black box attack success rate on the MNIST dataset

(单位:%)

Model	Acc	Agent Model	AdvGAN										MTA
			0	1	2	3	4	5	6	7	8	9	Mean	
LeNet	98.7	AlexNet	65.0	52.8	83.9	67.3	59.8	62.1	72.7	78.3	84.9	77.6	70.44	**94.8**
		C&Wmodel	91.8	82.2	89.3	94.9	94.2	95.9	92.4	90.4	95.3	96.4	92.28	**98.9**
AlexNet	98.6	LeNet	66.0	69.2	83.7	61.6	85	85.1	52.9	79.8	65.7	65.4	71.44	**89.4**
		C&Wmodel	88.3	88.4	85.3	85	85.1	68.2	74.9	94.4	84.3	84.2	83.81	**95.7**
C&Wmodel	99.3	LeNet	77.1	58.7	83.7	79.7	72.4	81.2	58.2	73.1	73.7	68.8	72.51	**94.1**
		AlexNet	58.0	44.2	84.8	83.8	74.0	78.4	36.8	74.8	78.3	74.1	68.72	**89.8**

表 5 CIFAR10 数据集上的黑盒攻击成功率

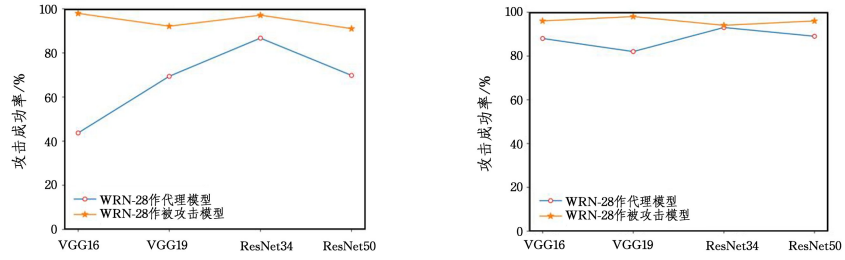
Table 5 Black box attack success rate on the CIFAR10 dataset

(单位:%)

Model	Acc	Agent Model	AdvGAN										MTA
			Plane	Car	Bird	Cat	Deer	Dog	Frog	Horse	Ship	Truck	Mean	
VGG11	87.6	ResNet18	98.8	98.1	98.7	96.6	98.2	98.6	98.3	98.7	96.8	98.9	98.17	**99.7**
		WRN-28	65.6	66.7	90.0	71.0	97.3	76.2	69.9	97.8	71.6	82.4	78.85	**98.1**
ResNet18	89.2	VGG11	99.1	93.1	99.1	98.7	99	98.8	99.1	98.6	98.8	99.1	98.34	**99.2**
		WRN-28	85.2	65.6	75.4	93.0	77.7	81.9	93.9	65.3	72.8	73.4	78.42	**98.8**
WRN-28	83.5	VGG11	99.4	97.2	98.8	98.8	99.6	97.9	99.3	98.7	99.4	99.2	98.83	**99.4**
		ResNet18	99.7	91.2	98.6	97.8	99.1	98.8	99.2	98.3	91.1	99.0	97.28	**99.7**

此外,对于 AdvGAN 模型而言,选择不同的被攻击模型和代理模型,可能会对攻击效果产生非常大的影响。例如,当

我们选择 WRN-28 作为代理模型时，攻击成功率（约为 78%）较 VGG11 和 ResNet18 会有较大幅度下降；而对于 MTA 模型，模型的差异并不会对攻击效果造成非常大的影响，即使选择 WRN-28 作为代理模型，仍能取得非常高的攻击成功率（约为 98%）。为了进一步验证 MTA 对不同模型之间的鲁棒性，在 CIFAR10 上进行了实验，新增了模型 VGG16，VGG19，ResNet34 以及 ResNet50，使用 WRN-28 作为代理模型进行黑盒攻击，另外，在攻击 WRN-28 时依次使用上述 4 个模型作为代理模型。图 5 为两种攻击方法的攻击成功率，从图 5(a)可以看出，当使用 WRN-28 作为代理模型时，AdvGAN 攻击成功率很低，但是当 WRN-28 作为被攻击模型而其他 4 个模型作为代理模型时，AdvGAN 的攻击成功率却很高。这个结果表明，AdvGAN 在不同结构模型间的鲁棒性较差。从图 5(b)中 MTA 的攻击结果来看，当 WRN-28 分别作为代理模型和被攻击模型时的攻击成功率相差很小，而且比 AdvGAN 的攻击成功率高出 20%～50%，且都保持在 98%以上，这说明不同模型结构上的差异对 MTA 的影响较小。



(a) AdvGAN　　(b) MTA

图 5　黑盒环境下 WRN-28 作为被攻击模型和代理模型时的攻击成功率

Fig. 5　Success rate of AdvGAN and MTA when WRN-28 is used as the attacked model and agent model in black-box respectively

图 6 为黑盒环境下 MTA 方法在 CIFAR10 数据集上实施有目标攻击的对抗样本图。其中横轴为被攻击的黑盒模型，纵轴为黑盒的代理模型。对角线上为白盒攻击的结果，可以看出，黑盒攻击与白盒攻击一样仍然在视觉效果上保持了与原图像较高的相似度。

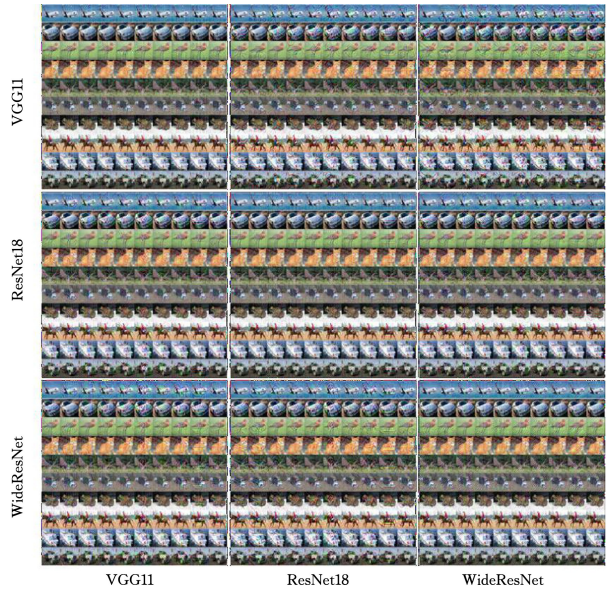


图 6　MTA 方法在黑盒环境下产生的对抗样本

Fig. 6　Adversarial examples generated by the MTA method in black box environment

4.5　高分辨率的对抗样本

为了验证 MTA 方法生成高分辨率对抗样本的能力，我们攻击了经过预训练的模型 VGG16。VGG16 模型在 ImageNet 子集测试集上的准确率高达 97.4%，是一个鲁棒性较强的分类器。

本文使用 MTA 方法以 0.05 的小扰动范围，生成大小为 224×224 像素的对抗样本，并在测试集上达到了 99.33%的攻击成功率。我们在测试集中随机挑选部分原始图像和对抗样本，并对它们进行组合可视化，如图 7 所示。其中左侧为原始图像，右侧为对抗样本。可以看出，MTA 方法在高分辨率数据集上仍然能以较高的攻击成功率完成有目标攻击，并且能够保持与原始图像相似的视觉效果。

(a) Tench　　(b) Springer

(c) Chainsaw　　(d) Parachute

图 7　MTA 方法在 ImageNet 上生成的对抗样本

Fig. 7　Adversarial examples generated by the MTA method on ImageNet dataset

4.6　对抗样本的迁移攻击

本节将评估 MTA 生成的对抗样本在不同模型之间的迁移性能。在 CIFAR10 数据集上将其与 FGSM 和 PGD 等多目标攻击方法的对抗样本在迁移成功率上进行对比。为公平起见，FGSM 和 PGD 的攻击强度都设置为 0.3，PGD 的迭代次数为默认的 40，步长为 0.01。我们攻击每个模型并生成

900 个对抗样本，然后将对抗样本迁移至其他使用同一个数据集训练的模型上，若攻击仍然能够成功，则认为对抗样本迁移性有效。其他攻击算法的实现，我们借助了在 PyTorch 框架下的对抗样本工具箱 Advertochbox。迁移结果如表 6 所列，纵轴为生成对抗样本的源模型，横轴为被攻击模型，源模型和被攻击模型相同时，表格中的成功率表示各攻击方法的白盒攻击成功率。可以看出，FGSM 的攻击成功率相对较低；PGD 算法作为目前最强的一阶梯度攻击算法，可以达到 100%的攻击成功率；而 MTA 以比 PGD 更快的速度在 3 个模型上分别达到 99.3%，99.8%，99.6%的攻击成功率，与 PGD 几乎相同。非对角线上为对抗样本在不同模型之间的迁移攻击成功率，MTA 与 FGSM，PGD 相比，在迁移成功率上有较大的提升，例如，从 VGG11 到 WRN-28，MTA 方法比 PGD 高出约 26%(61.52% vs 35.56%)，比 FGSM 高出约 41%(61.52% vs 20.31%)，从 WRN-28 到 ResNet18，MTA 方法比 PGD 高出约 37%(51.34% vs 14.44%)。

表 6 对抗样本迁移攻击成功率

Table 6 Success rate of transfer attack

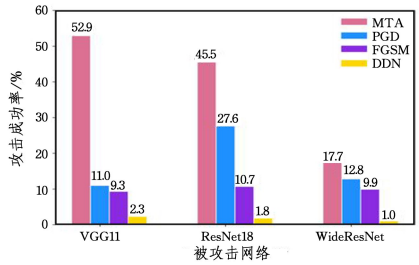
(单位：%)

Model	Method	VGG11	ResNet18	WRN-28
VGG11	MTA	99.3	**74.94**	**61.52**
	FGSM[2]	94.1	18.42	20.31
	PGD[7]	**100**	62.22	35.56
ResNet 18	MTA	**39.17**	99.8	**37.39**
	FGSM[2]	19.57	97.1	22.31
	PGD[7]	28.89	**100**	35.56
WRN-28	MTA	**23.21**	**51.34**	99.6
	FGSM[2]	12.33	12.33	96.4
	PGD[7]	17.78	14.44	**100**

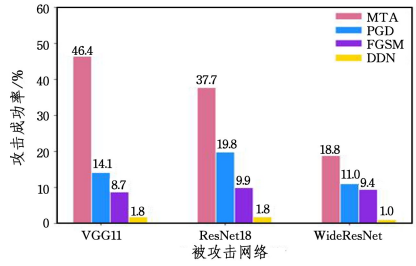
4.7 防御背景下的攻击

防御背景下的攻击指假设攻击者不知道模型经过防御，仍然攻击原模型，如果此时攻击原模型产生的对抗样本仍然可以攻击防御后的模型，则证明该攻击有效。本文使用 Adv 和 Ens 两种对抗训练方法测试网络的性能，在对抗训练过程中，设置损失函数中干净样本与对抗样本的比例 α 为 0.5，学习率为 0.001，使用 Adam 优化器训练 100 个 epoch，保存在测试集上具有最高准确率的模型。

Madry 等[8]经过验证提出 PGD 方法是最强的一阶梯度攻击，如果有防御方法能对 PGD 有较好的防御效果，那么对其他攻击也应该一样。图 8 为在 CIFAR10 数据集上的 3 个模型在防御背景下使用 MTA，PGD，FGSM 以及 DDN 算法的攻击成功率。总体而言，相比使用 Adv 对抗训练，使用 Ens 对抗训练时模型有着更高的攻击成功率，证明 Adv 对抗训练方法能有效提高模型的对抗鲁棒性。无论采用哪种对抗训练，使用 MTA 方法时总能保持较高的攻击成功率，例如，在使用 Ens 对抗训练后攻击 VGG11 模型，PGD 有 11%的攻击成功率，FGSM 和 DDN 更低，而 MTA 有 52.9%的攻击成功率，高出 PGD 方法 4 倍。使用 Adv 对抗训练后攻击 VGG11，MTA 的攻击成功率为 46.4%，仍比 PGD 的 14.1%高出 3 倍。其他情况下，MTA 显著领先 PGD，FGSM，DDN，这充分证明了 MTA 方法具有较高的防御鲁棒性。



(a) Ens 对抗训练



(b) Adv 对抗训练

图 8 防御背景下 MTA 及其他多目标攻击方法的攻击结果

Fig. 8 Attack results of MTA and other multi-target attack methods under defense

此外，为了验证 MTA 方法对预处理防御方法的鲁棒性，我们使用了包括 JPEG 压缩、图像位深度缩减、平均滤波、中值滤波以及二值化滤波在内的 5 种预处理方法处理对抗样本后，再次对目标网络实施有目标攻击。若经预处理后的对抗样本仍然能够误导目标网络输出目标类别，则认为攻击有效。其中，JPEG 压缩的图像质量设置为 Advertorchbox 默认的 75%，位深度缩减方法的位深度设置为 2，平均滤波的 kernel 大小设置为 3。各种预处理防御方法下，MTA 与其他对比方法的有目标攻击成功率如表 7 所列。

表 7 预处理防御下各攻击方法的攻击成功率

Table 7 Success rate of attack methods under pre-processing defense

(单位：%)

Defense	Model	PGD	FGSM	DDN	MTA
JPEG 压缩	VGG11	43.4	11.8	2.4	**86.6**
	ResNet18	47.6	13.2	2.8	**87.3**
	WRN-28	39.6	11.0	3.0	**82.4**
位深度缩减	VGG11	81.9	12.6	7.0	**89.2**
	ResNet18	82.4	14.6	6.2	**94.9**
	WRN-28	68.2	11.6	5.8	**96.2**
平均滤波	VGG11	20.9	21.8	5.8	**34.0**
	ResNet18	22.9	20.9	6.3	**35.1**
	WRN-28	11.4	13.4	6.0	**19.1**
中值滤波	VGG11	29.3	14.1	3.6	**46.3**
	ResNet18	23.6	14.8	3.8	**37.0**
	WRN-28	12.0	10.4	4.3	**25.8**
二值化滤波	VGG11	29.0	13.4	8.4	**41.3**
	ResNet18	26.9	12.8	8.7	**43.6**
	WRN-28	20.6	13.3	8.0	**45.0**

虽然预处理的方法较为简单，但是其防御有目标攻击的效果明显。如表 7 所列，几个滤波方法都能有效地降低各攻击方法的有目标攻击成功率。这是因为在有目标攻击过程中，各方法产生的对抗噪声针对性较强；而滤波方法使得噪声相对平滑，因此对抗性噪声的针对性降低。虽然经过滤波的

对抗样本仍然有可能误导模型，但是模型的预测类别已经不是我们设定的目标，从而使有目标攻击失效。尽管如此，在所有的防御设定下，MTA 方法的攻击成功率仍然优于其他几个多目标攻击方法，从而证明了 MTA 方法对预处理的防御仍然具有较高的鲁棒性。

结束语　本文提出了一种基于生成对抗网络的多目标攻击对抗样本生成网络 MTA，通过将图像和分类标签编码成相同维度，从而融合两者的特征，进而解码出在视觉上与输入图像相似但是在被攻击网络中的分类结果为输入标签的对抗样本。相比 AdvGAN，MTA 在仅增加少量参数的情况下，只需要经过一次训练即可快速地生成多目标类别的对抗样本，并且在白盒和黑盒环境下都具有较高的攻击能力；与 PGD，FGSM，DDN 等多目标攻击方法相比，MTA 生成的对抗样本具有更好的迁移性能和防御背景下的攻击能力。但无论是 MTA 还是 AdvGAN，由于要使用动态蒸馏，使模型在黑盒环境下的训练速度均较慢，因此在接下来的工作中，我们将进一步考虑更高效的蒸馏方式。

参 考 文 献

[1] SZEGEDY C, ZARENBA W, SUTSKEVER I, et al. Intriguing properties of neural networks[C]// International Conference on Learning Representations. 2014.

[2] GOODFELLOW I J, SHLENS J, SZEGEDY C. Explaining and harnessing adversarial examples[C]// International Conference on Learning Representations. 2015.

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

[4] MOOSAVIDEZFOOLI S M, FAWZI A, FROSSARD P. Deepfool: A simple and accurate method to fool deep neural networks [C]// Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2016: 2574-2582.

[5] CHEN P Y, ZHANG H, SHARMA Y, et al. Zoo: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models[C]// Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security. 2017: 15-26.

[6] XIAO C, LI B, ZHU J Y, et al. Generating Adversarial Examples with Adversarial Networks[C]// Proceedings of the 27th International Joint Conference on Artificial Intelligence. 2018: 3905-3911.

[7] LI B, XIE J Z. Study on the Prediction of Imbalanced Bank Customer Churn Based on Generative Adversarial Network [J]. Journal of Chongqing University of Technology (Natural Science), 2021, 35(8): 136-143.

[8] MADRY A, MAKELOV A, SCHMIDT L, et al. Towards Deep Learning Models Resistant to Adversarial Attacks[C]// International Conference on Learning Representations. 2017.

[9] KURAKIN A, GOODFELLOW I, BENGIO S. Adversarial examples in the physical world[C]// International Conference on Learning Representations Workshop. 2017.

[10] RONY J, HAFEMANN L G, OLIVEIRA L S, et al. Decoupling direction and norm for efficient gradient-based l2 adversarial attacks and defenses[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. IEEE, 2019: 4322-4330.

[11] BRENDEL W, RAUBER J, BETHGE M. Decision-based adversarial attacks: Reliable attacks against black-box machine learning models[C]// International Conference on Learning Representations. 2018.

[12] CHENG M, LE T, CHEN P Y, et al. Query-efficient hard-label black-box attack: An optimization-based approach[C]// International Conference on Learning Representations. 2019.

[13] CHENG M, SINGH S, CHEN P, et al. Sign-opt: A query-efficient hard-label adversarial attack[C]// International Conference on Learning Representations. 2020.

[14] XU W, EVANS D, QI Y. Feature squeezing: Detecting adversarial examples in deep neural networks[C]// Network and Distributed System Security Symposium. 2018.

[15] MENG D, CHEN H. Magnet: a two-pronged defense against adversarial examples[C]// Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security. 2017: 135-147.

[16] TRAMÈR F, KURAKIN A, PAPERNOT N, et al. Ensemble adversarial training: Attacks and defenses[C]// International Conference on Learning Representations. 2018.

[17] DAS N, SHANBHOGUE M, CHEN S T. Keeping the Bad Guys Out: Protecting and Vaccinating Deep Learning with JPEG Compression[J]. arXiv: 1705.02900, 2017.

[18] RAFF E, SYLVESTER J, FORSYTH S, et al. Barrage of Random Transforms for Adversarially Robust Defense[C]// Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2019: 6521-6530.

[19] JEDDI A, SHAFIEE M J, KARG M, et al. Learn2perturb: an end-to-end feature perturbation learning to improve adversarial robustness[C]// Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2020: 1241-1250.

[20] JIANG Z, CHEN T, CHEN T, et al. Robust Pre-Training by Adversarial Contrastive Learning[C]// Advances in Neural Information Processing Systems. 2020.

[21] KIM M, TACK J, HWANG S J, et al. Adversarial self-supervised contrastive learning [C]// Advances in Neural Information Processing Systems. 2020.

[22] BAI Y, ZENG Y, JIANG Y, et al. Improving adversarial robustness via channel-wise activation suppressing [C]// International Conference on Learning Representations. 2021.

LI Jian, born in 1996, postgraduate. His main research interests include computer vision and deep learning.

GUO Yan-ming, born in 1989, associate professor. His main research interests include computer vision, natural language processing and deep learning.

（责任编辑：李亚辉）