Journal of Computer Applications
计算机应用, 2022, 42(9): 2732 - 2741
ISSN 1001-9081
CODEN JYIIDU
2022-09-10
http://www.joca.cn

文章编号:1001-9081(2022)09-2732-10
DOI:10.11772/j.issn.1001-9081.2021071339

图像分类中的白盒对抗攻击技术综述

魏佳璇[1*],杜世康[1],于志轩[1,2],张瑞生[1]
(1. 兰州大学 信息科学与工程学院,兰州 730000; 2. 兰州大学第一医院,兰州 730000)
(*通信作者电子邮箱 weijx@lzu.edu.cn)

摘 要:在深度学习中图像分类任务研究里发现,对抗攻击现象给深度学习模型的安全应用带来了严峻挑战,引发了研究人员的广泛关注。首先,围绕深度学习中用于生成对抗扰动的对抗攻击技术,对图像分类任务中重要的白盒对抗攻击算法进行了详细介绍,同时分析了各个攻击算法的优缺点;然后,分别从移动终端、人脸识别和自动驾驶三个现实中的应用场景出发,介绍了白盒对抗攻击技术的应用现状;此外,选择了一些典型的白盒对抗攻击算法针对不同的目标模型进行了对比实验并分析了实验结果;最后,对白盒对抗攻击技术进行了总结,并展望了其有价值的研究方向。

关键词:对抗样本;白盒对抗攻击;深度学习;图像分类;人工智能安全
中图分类号:TP181 **文献标志码**:A

Review of white-box adversarial attack technologies in image classification

WEI Jiaxuan[1*], DU Shikang[1], YU Zhixuan[1,2], ZHANG Ruisheng[1]
(1. *School of Information Science and Engineering, Lanzhou University, Lanzhou Gansu* 730000, *China*;
2. *The First Hospital of Lanzhou University, Lanzhou Gansu* 730000, *China*)

Abstract: In the research of image classification tasks in deep learning, the phenomenon of adversarial attacks brings severe challenges to the secure application of deep learning models, which arouses widespread attention of researchers. Firstly, around the adversarial attack technologies for generating the adversarial perturbations, the important white-box adversarial attack algorithms in the image classification tasks were introduced in detail, and the advantages and disadvantages of different attack algorithms were analyzed. Then, from three realistic application scenarios: mobile application, face recognition and autonomous driving, the application status of the white-box adversarial attack technologies was illustrated. Additionally, some typical white-box adversarial attack algorithms were selected to perform experiments on different target models, and the experimental results were analyzed. Finally, the white-box adversarial attack technologies were summarized, and their valuable research directions were prospected.

Key words: adversarial example; white-box adversarial attack; deep learning; image classification; artificial intelligence security

0 引言

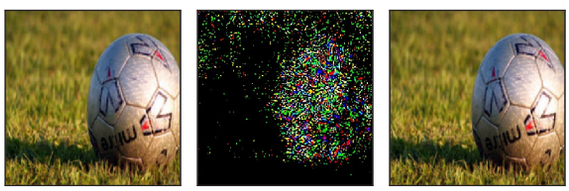
近年来,得益于图形处理器(Graphics Processing Unit, GPU)技术突飞猛进的发展以及计算机硬件升级带来的算力提升,深度学习取得了令人瞩目的发展成果。当前,深度学习技术不仅大量地应用在图像分类、语音识别和自然语言处理领域的常规任务中,更是在自动驾驶系统[1]、人脸识别系统[2]、恶意软件自动分类[3-4]和异常检测[5]等工业生产和生活领域的大量关键任务中发挥着重要作用。虽然深度学习技术在众多问题的解决上取得了一系列重要研究成果,但对于自身还存在一些关键的问题亟待解决,特别是深度学习模型存在的对抗样本问题。

2013年,Szegedy等[6]在研究图像分类任务时发现,在一个可以被深度学习模型正常分类的样本图片上添加噪声数据后,即使是分类准确率很高的深度学习模型也会以极高的置信度对该样本误分类。而添加在样本图片上的噪声数据是微小的,以人的肉眼几乎察觉不到在样本图像上进行的篡改,篡改后得到的输入样本被称为对抗样本。对抗样本的一个示例如图1所示,原始样本为ImageNet数据集[7]中图片,其真实的标签和在ImageNet上预训练的ResNet模型[8]的分类结果均为足球。添加了恶意的扰动数据后得到对抗样本,此时ResNet模型将该对抗样本误分类为橄榄球。

对抗样本问题揭示了深度学习模型存在严重的安全漏

收稿日期:2021-07-26;**修回日期**:2021-09-29;**录用日期**:2021-10-08。 **基金项目**:甘肃省自然科学基金资助项目(20YF8FA080)。
作者简介:魏佳璇(1983—),女,甘肃兰州人,工程师,博士,主要研究方向:人工智能、网络安全; 杜世康(1997—),男,甘肃武威人,硕士研究生,主要研究方向:对抗机器学习; 于志轩(1983—),男,甘肃兰州人,博士研究生,主要研究方向:机器学习、深度学习; 张瑞生(1962—),男,甘肃兰州人,教授,博士,主要研究方向:可解释机器学习、复杂网络分析、图像识别与分析、服务计算、化学信息学、生物信息学。

洞,给深度学习技术的普遍应用带来了严峻的安全挑战。在围绕对抗样本的研究过程中,主要以用于对抗样本生成的对抗攻击技术的研究为主,迭代发展出了多样性的对抗攻击算法。目前已知的大部分对抗攻击算法针对图像分类任务提出,并经过改造应用在诸如语义分割、目标检测等常见的计算机视觉任务上。经过改造的攻击算法甚至可以很好地推广到自然语言处理和语音识别等任务中。此外,对抗攻击现象不仅发生在数字图像空间,对于部署在真实应用场景下的深度学习模型[4,9-11]也能够带来安全威胁。



(a) 原始样本　(b) 扰动噪声　(c) 对抗样本

图1　对抗样本示例

Fig. 1　Examples of adversarial sample

对抗攻击技术由于其破坏力和潜在的应用前景,成为近年来深度学习学术界和工业界共同的研究热点。Carlini等[12]统计的2014年至今,arXiv网站发表的对抗样本相关论文的数量情况如图2所示。攻击者利用已有的对抗攻击算法可以在深度学习模型推理阶段对输入样本添加噪声数据,而达到恶意改变模型推理结果的目的。根据对抗攻击技术在生成对抗样本时是否需要了解目标模型的网络结构、参数设置、训练数据和方式等知识,对抗攻击技术可分为白盒攻击和黑盒攻击;根据攻击是否需要让模型输出指定的目标类别,又分为目标攻击和无目标攻击。目前针对对抗攻击技术的研究大多以白盒攻击方式为主,其攻击成功率大幅高于黑盒攻击方式,对深度学习模型带来的安全威胁也较为严峻。为及时评估对抗攻击技术给深度学习模型带来的安全风险,以便为深度学习技术的安全应用提供有益参考。本文围绕图像分类任务,对近年来研究人员提出的具有一定代表性的白盒对抗攻击技术进行全面阐述和分析总结。

图2　对抗样本论文数量

Fig. 2　Number of adversarial example papers

1　对抗攻击技术背景知识

对抗攻击是一种发生在深度学习模型推理阶段的攻击行为。在图像分类任务中,给定一个深度学习模型$f(\boldsymbol{x})=y$,$\boldsymbol{x}\in\mathbf{R}^m$为模型的输入,$y\in Y$为针对当前输入$\boldsymbol{x}$的模型输出。模型$f(\cdot)$一般还包含一组训练好的权重参数$\boldsymbol{\theta}$,为方便说明,对模型描述时省略该参数。对抗攻击技术描述为在针对目标模型$f(\cdot)$的输入$\boldsymbol{x}$上寻找一个小的噪声数据$\boldsymbol{r}$,当$\boldsymbol{r}$叠加在$\boldsymbol{x}$上输入目标模型后,使得$f(\boldsymbol{x}+\boldsymbol{r})\neq f(\boldsymbol{x})$;在目标攻击中,使得$f(\boldsymbol{x}+\boldsymbol{r})=y^t$,$y^t$为需要让模型输出的目标类目。

为了使噪声数据$\boldsymbol{r}$足够小,保证人眼察觉不到攻击者在图像上进行的篡改,大多数对抗攻击算法[6,13-14]会使用l_0、l_2或l_∞范数对扰动噪声$\boldsymbol{r}$的大小进行限制。用l_2范数来约束扰动大小时,在目标攻击中,对抗样本的生成问题可描述为如下的优化形式:

$$\text{Minimize}\ \|\boldsymbol{r}\|_2 \tag{1}$$

$$\text{s. t.}\ \ f(\boldsymbol{x}+\boldsymbol{r})=y^t;\ \boldsymbol{x}+\boldsymbol{r}\in\mathbf{R}^m$$

其中:y^t是要攻击的目标类别;$\boldsymbol{x}$为原始输入样本;$\boldsymbol{r}$表示扰动噪声;$\boldsymbol{x}+\boldsymbol{r}$表示得到的对抗样本,后文也用$\boldsymbol{x}^\wedge$表示。

下面简要说明和定义在本文中出现的对抗攻击技术相关术语。

模型　深度学习模型,文中一般指图像分类模型。

对抗样本(Adversarial Examples)　对抗样本概念由Szegedy等[6]提出,攻击者对原始输入样本添加轻微的噪声数据后能使深度学习模型推理错误,这类添加噪声数据后影响模型推理能力的样本称为对抗样本。

扰动/对抗扰动(Adversarial Perturbations)　在原始输入样本上添加噪声数据后能使深度学习模型推理错误,被添加的噪声数据称为扰动或对抗扰动。

通用对抗扰动(Universal Adversarial Perturbations)　不同于对抗扰动只针对特定的输入样本,通用对抗扰动添加在大部分输入样本上都会使得深度学习模型推理错误。

对抗扰动的迁移攻击性　针对目标模型A,在输入样本$\boldsymbol{x}$上生成对抗扰动$\boldsymbol{r}$。当扰动$\boldsymbol{r}$叠加在针对目标模型B的输入样本上后能使模型B推理错误的现象称为对抗扰动的迁移攻击性[6,14-15]。如果对抗扰动能够在目标模型B的大部分数据点上使模型推理错误,则称该对抗扰动具有很好的迁移攻击能力。利用对抗样本的迁移攻击性进行攻击是一种有效的对抗攻击方式。

模型鲁棒性　指模型的对抗鲁棒性,针对对抗攻击技术自身防御效果良好的模型,则称该模型具有较好的鲁棒性。

对抗训练　一种提升模型对抗鲁棒性的训练方式[14,16],在训练模型时,使用正常样本和对抗样本同时对模型进行训练的方式。

攻击成功率　利用对抗攻击技术生成对抗样本输入目标模型后,模型推理错误样本数占所有输入样本数的百分比,有时也可用模型对对抗样本的分类准确率代替表示,模型对对抗样本的分类准确率越低说明攻击者的攻击成功率越高。

2　白盒对抗攻击技术

就目前的研究来看,白盒对抗攻击技术是主要的对抗样本生成方式。大部分的白盒对抗攻击算法针对单个输入图像生成对抗扰动,但也有算法针对目标模型和整个数据集生成通用对抗扰动[17-18]。通过对大部分研究人员主要的成果[19-21]进行研究分析,发现大部分的白盒对抗攻击算法目前主要分为以下4种:1)基于直接优化的方法;2)基于梯度优化的方法;3)基于决策边界分析的方法;4)基于生成式神经网络生成的方法。其他方面,研究人员也利用差分进化、空

域变换等思路进行对抗样本的生成。本文依据上述分类，对在白盒条件设置下主要的对抗攻击算法进行全面分析和阐述，最后在表1中对攻击算法进行了比较和总结。

2.1 基于直接优化的攻击方法

基于直接优化的攻击方法是目前较为重要的一类对抗攻击技术，主要包含两种对抗攻击算法：基于Box-constrained L-BFGS（Box-constrained Limited-memory BFGS）的攻击算法[6]是第一个提出的对抗攻击算法，该算法也首次揭示了深度学习模型中存在的对抗样本问题；C&W（Carlini&Wagner）攻击[13]通过对基于Box-constrained L-BFGS的攻击算法的改进，能够生成对蒸馏防御网络[22]具有较好攻击能力的对抗样本。这类攻击方法通过算法对目标函数直接优化生成的对抗扰动相对较小，但存在优化时间长，算法需花费大量时间寻找合适超参数的问题。

2.1.1 基于Box-constrained L-BFGS的攻击

2013年，Szegedy等[6]首先提出了利用Box-constrained L-BFGS算法[23]直接优化求解的对抗样本生成方法。由于式（1）中的扰动限制条件难以直接优化，Szegedy等利用拉格朗日松弛法将其中的$f(\boldsymbol{x}+\boldsymbol{r})=y^t$限制条件简化为$loss_f(\boldsymbol{x}+\boldsymbol{r},y^t)$进行优化，$loss_f$表示交叉熵损失函数，最终得到的优化目标如下：

$$\text{Minimize } c\|\boldsymbol{r}\|_\infty + loss_f(\boldsymbol{x}+\boldsymbol{r},y^t) \tag{2}$$

$$\text{s.t. } \boldsymbol{x}+\boldsymbol{r}\in[0,1]^m$$

其中：输入图像被归一化在[0,1]，以满足凸优化方法中的箱型约束条件，使得上述目标可以利用L-BFGS算法进行求解。

利用Box-constrained L-BFGS算法生成对抗样本的方法是最早被设计的对抗攻击算法，该算法首次将生成对抗样本的过程抽象为一个凸优化的问题处理，是重要的基于优化方法的对抗攻击算法。该算法为目标攻击算法，使用该算法求解的思路是先固定超参数c来优化当前参数值下的最优解，再通过对c进行线性搜索即可找到满足$f(\boldsymbol{x}+\boldsymbol{r})=y^t$条件的最优对抗扰动$\boldsymbol{r}$，最终得到的对抗样本为$\boldsymbol{x}+\boldsymbol{r}$。

2.1.2 C&W攻击

为了攻破Papernot等[22]提出的蒸馏防御网络，Carlini和Wagner提出了著名的C&W攻击[13]。该攻击算法可以使用l_0、l_2或l_∞范数分别对扰动进行限制生成对抗样本，是目前较为强大的目标攻击算法之一。C&W攻击算法属于直接优化的攻击算法，是基于Box-constrained L-BFGS算法（式（2））的改进版，改进主要体现在两点：

1）基于Box-constrained L-BFGS的攻击中损失函数为交叉熵损失函数，而C&W攻击算法考虑了攻击目标类和其他类别之间的关系，选择了更好的损失函数[13]，如下所示：

$$loss_{f,t}(\boldsymbol{x}^{\mathrm{A}}) = \max(\max\{Z(\boldsymbol{x}^{\mathrm{A}})_i : i\neq t\} - Z(\boldsymbol{x}^{\mathrm{A}})_t, -k) \tag{3}$$

式中：$Z(\boldsymbol{x}^{\mathrm{A}})=\mathrm{Logits}(\boldsymbol{x}^{\mathrm{A}})$表示目标网络Softmax前一层的输出；$i$表示标签类别；$t$表示目标攻击的标签类；$k$表示对抗样本的攻击成功率，$k$越大，生成的对抗样本的攻击成功率越高。

2）去除了式（2）中的Box-constrained限定条件，使该优化问题转化为无约束的凸优化问题，方便利用梯度下降法，动量梯度下降法和Adam[24]等算法求解。为实现该目的，Carlini等[13]提供了两种有效方法：①采用投影梯度下降法的思路将每次迭代过程中得到的$\boldsymbol{x}+\boldsymbol{r}$裁剪在$[0,1]^m$内，以去除$\boldsymbol{x}+\boldsymbol{r}$的区间约束条件，但此方法在对$\boldsymbol{x}+\boldsymbol{r}$进行裁剪时会带来梯度信息的损失；②引入新的变量$\boldsymbol{\omega}$，令$\boldsymbol{\omega}\in[-\infty,+\infty]$，构造一个映射函数将$\boldsymbol{\omega}$从$[-\infty,+\infty]$映射到$[0,1]$，通过优化$\boldsymbol{\omega}$去掉方法①中由$\boldsymbol{x}+\boldsymbol{r}\in[0,1]^m$条件引起的优化误差，具体映射函数如下：

$$\begin{cases}\boldsymbol{\delta}_i = \dfrac{1}{2}(\tanh(\boldsymbol{\varpi}_i)+1) - \boldsymbol{x}_i \\ \boldsymbol{x}+\boldsymbol{r} = \boldsymbol{x}_i + \boldsymbol{\delta}_i\end{cases} \tag{4}$$

其中：$-1\leqslant\tanh(\boldsymbol{\varpi}_i)\leqslant 1$，故$0\leqslant\boldsymbol{x}+\boldsymbol{r}\leqslant 1$。

Carlini等[13]分别用方法①和方法②进行了实验分析，发现用投影梯度下降法处理Box-constrained限定条件时生成的对抗样本攻击能力较强，但引入变量进行优化的方法生成的对抗扰动较小。此外，在优化算法的选择上，梯度下降、动量梯度下降和Adam等优化算法都可以生成相同质量的对抗样本，但Adam算法的收敛速度要比其他两种快。C&W攻击算法生成的对抗样本针对蒸馏防御的模型攻击能力很好，是目前较为强大的白盒攻击算法，也是用于评估模型鲁棒性的主要测试算法之一。

2.2 基于梯度优化的攻击方法

基于梯度优化的攻击方法是目前一种主要的对抗攻击技术。这类攻击方法的核心思想是在模型损失函数变化的方向上对输入样本进行扰动，来使模型误分类输入样本或使模型分类输入样本到指定的不正确目标类别上。这类攻击方法的优点是方法实现简单，且白盒对抗攻击成功率较高。该类方法以FGSM（Fast Gradient Sign Method）算法[14]为基础，衍生发展出了I-FGSM（Iterative FGSM）算法[25]、PGD（Projected Gradient Descent）算法[26]、动量迭代的MI-FGSM（Momentum Iterative FGSM）算法[27]以及多样性的梯度攻击算法[28-29]。

2.2.1 FGSM攻击

2014年，Goodfellow等[14]在其研究中认为，深度学习模型存在对抗样本是由于模型过于线性的特性导致，基于该观点提出了基于一步梯度计算的对抗样本生成算法FGSM。这项工作意义深远，受Goodfellow等启发，后来出现的基于梯度优化的大部分攻击算法都是FGSM算法的变种。

FGSM攻击算法的思想是使对抗扰动的变化量与模型损失梯度变化的方向保持一致。具体来说，在无目标攻击中，使模型损失函数关于输入$\boldsymbol{x}$的梯度在上升的方向上变化扰动达到让模型误分类的效果。假设$\boldsymbol{\theta}$为模型的参数，$\boldsymbol{x}$为模型的输入，y为输入$\boldsymbol{x}$对应的正确类别标签，$J(\boldsymbol{\theta},\boldsymbol{x},y)$为模型的损失函数，为交叉熵损失函数，$\nabla_x J(\boldsymbol{\theta},\boldsymbol{x},y)$为损失函数关于$\boldsymbol{x}$的梯度。FGSM算法描述为：

$$\boldsymbol{x}^{\mathrm{A}} = \boldsymbol{x} + \alpha\,\mathrm{sign}(\nabla_x J(\boldsymbol{\theta},\boldsymbol{x},y)) \tag{5}$$

其中：α为超参数，表示为一步梯度的步长；$\mathrm{sign}(\cdot)$为符号函数，故该方法生成的扰动为在l_∞范数约束下的对抗扰动。FGSM算法由于只计算一次梯度，其攻击能力有限，但生成的对抗样本具有较好的迁移攻击能力。

2.2.2　I-FGSM攻击

由于FGSM算法只经过一次梯度计算生成对抗样本,并且该方法成功应用的前提条件是损失函数的梯度变化方向在局部区间内是线性的。在非线性的优化区间内,沿着梯度变化方向进行大步长优化生成的对抗样本并不能保证攻击成功。针对该问题,Kurakin等[25]提出了迭代FGSM(I-FGSM)算法,通过把优化区间变小来使Goodfellow的线性假设[14]近似成立。I-FGSM算法的无目标攻击描述为:

$$\begin{cases}\boldsymbol{x}_0^{\mathrm{A}}=\boldsymbol{x}\\ \boldsymbol{x}_{i+1}^{\mathrm{A}}=\mathrm{Clip}(\boldsymbol{x}_i^{\mathrm{A}}+\alpha\,\mathrm{sign}(\nabla_x J(\boldsymbol{\theta},\boldsymbol{x}_i^{\mathrm{A}},y)))\end{cases}\tag{6}$$

其中:$\boldsymbol{x}_{i+1}^{\mathrm{A}}$为第$i$次迭代后得到的对抗样本;$\alpha$为超参数,表示为迭代过程中每步梯度的步长;Clip(·)操作把超过合法范围的$\boldsymbol{x}^{\mathrm{A}}$裁剪在规定的范围内。

I-FGSM算法较FGSM算法生成的对抗样本攻击能力更强[30],但其生成的对抗样本迁移攻击能力却不如FGSM算法。

此外,Kurakin等[25]还在I-FGSM算法的基础上,通过将攻击的目标类别y指定为原始样本在模型上输出置信度最低的类别标签y_l来进行针对置信度最低类别的目标攻击。在目标攻击过程中,扰动的变化方向与模型损失函数关于输入的梯度下降方向保持一致,其优化的目标形式为:

$$\begin{cases}\boldsymbol{x}_0^{\mathrm{A}}=\boldsymbol{x}\\ \boldsymbol{x}_{i+1}^{\mathrm{A}}=\mathrm{Clip}(\boldsymbol{x}_i^{\mathrm{A}}-\alpha\,\mathrm{sign}(\nabla_x J(\boldsymbol{\theta},\boldsymbol{x}_i^{\mathrm{A}},y_l)))\end{cases}\tag{7}$$

这种目标攻击方式生成的对抗样本使模型误分类到与正确类别差距很大的类,其攻击效果更具破坏性。

2.2.3　PGD攻击

2017年,Madry等[26]提出的PGD攻击算法是目前公认为最强的白盒攻击方法,也是用于评估模型鲁棒性的基准测试算法之一。PGD攻击本质上也是迭代的FGSM算法,与I-FGSM攻击类似。与I-FGSM算法相比,PGD算法的迭代次数更多,并在迭代过程中对上一版本得到的$\boldsymbol{x}^{\mathrm{A}}$随机地进行了噪声初始化,以此避免优化过程中可能遇到的鞍点[26]。使用PGD算法生成的对抗样本攻击能力比I-FGSM攻击能力强,但同样具有的迁移攻击能力弱的问题。

2.2.4　MI-FGSM攻击

在接下来的研究中,为了使对抗样本兼具强大的攻击能力和良好的迁移攻击能力,Dong等[27]提出了基于动量的迭代生成对抗样本的MI-FGSM算法。该算法在I-FGSM算法的迭代过程中引入动量技术[31-32],以此在损失梯度变化的方向上累计速度矢量以稳定梯度的更新方向,使得优化过程不容易陷入局部最优。MI-FGSM算法描述为:

$$\begin{cases}\boldsymbol{g}_{i+1}=\mu\boldsymbol{g}_i+\dfrac{\nabla_x J(\boldsymbol{x}_i^{\mathrm{A}},y)}{\|\nabla_x J(\boldsymbol{x}_i^{\mathrm{A}},y)\|_1}\\ \boldsymbol{x}_{i+1}^{\mathrm{A}}=\boldsymbol{x}_i^{\mathrm{A}}+\alpha\,\mathrm{sign}(\boldsymbol{g}_{i+1})\end{cases}\tag{8}$$

其中:$\boldsymbol{g}_{i+1}$表示在第i次迭代后累计的梯度动量;μ为动量项的衰减因子,当$\mu=0$,则上述形式为I-FGSM算法的形式。由于多次迭代中得到的梯度不在一个量级,将每次迭代中得到的当前梯度$\nabla_x J(\boldsymbol{x}_i^{\mathrm{A}},y)$通过其自身的$l_1$距离进行归一化。

MI-FGSM攻击算法生成的对抗样本在具有较好攻击能力的基础上还保留了一定的迁移攻击能力,是目前常用的白盒对抗攻击方法。

此外,为了更加有效提升基于梯度优化的方法生成对抗样本的迁移攻击能力,Xie等[28]提出了一种输入多样性的对抗攻击方式,采取数据增强的思路,在将图像输入到模型前,先对输入样本进行随机转化,如随机调整样本大小或随机填充给定的分布等。将转换后的图像输入至目标模型后,再应用I-FGSM[25]、MI-FGSM[27]等算法进行梯度计算生成对抗样本。另外,通过减轻对抗样本在不同模型间识别的敏感程度,也可以提高对抗样本的迁移攻击能力。Dong等[29]提出了基于梯度的平移不变攻击方式,通过将梯度与一个预先定义的核进行卷积来生成对大部分模型识别区域不太敏感的对抗样本。

2.3　基于决策边界分析的攻击方法

基于决策边界分析的攻击方法是一类特殊的对抗攻击方法。该方法最初由Moosavi-Dezfooli等[33]提出,分为针对单个输入图像生成对抗扰动的DeepFool攻击算法和针对目标模型和整个数据集生成通用对抗扰动的UAPs(Universal Adversarial Perturbations)攻击算法[17]。该类攻击方法的核心思想是通过逐步减小样本与模型决策边界的距离来使模型对该样本误分类,故其生成的对抗样本一般较小,但这类攻击方法不具备目标攻击能力。

2.3.1　DeepFool攻击

Moosavi-Dezfooli等[33]在对模型的决策边界分析后提出了一种精确计算对抗扰动的DeepFool方法。DeepFool算法生成的对抗扰动非常小,该扰动一般被认为近似于最小扰动。DeepFool算法具体描述为:首先,针对线性二分类问题,给定一个分类器$\hat{k}(\boldsymbol{x})=\mathrm{sign}(f(\boldsymbol{x}))$,$f(\boldsymbol{x})=\boldsymbol{\omega}\boldsymbol{x}+b$,分类器的决策边界用$F=\{\boldsymbol{x}:f(\boldsymbol{x})=0\}$表示,如图3所示。要使当前数据点$\boldsymbol{x}_0$被该模型误分类到决策边界另一边,其最小扰动对应于$\boldsymbol{x}_0$在$F$上的正交投影$r_*(\boldsymbol{x}_0)$:

$$r_*(\boldsymbol{x}_0)\equiv\arg\min_{\boldsymbol{r}}\|\boldsymbol{r}\|_2\tag{9}$$

根据$f(\boldsymbol{x})=\boldsymbol{\omega}\boldsymbol{x}+b$,推导得$r_*(\boldsymbol{x}_0)=-\dfrac{f(\boldsymbol{x}_0)}{\|\boldsymbol{x}_0\|_2^2}\boldsymbol{\omega}$,此为模型决策边界线性时计算得到的最小扰动值。

推广到非线性决策边界的二分类问题,可通过迭代的过程来近似得到针对数据点$\boldsymbol{x}_0$的最小扰动$r_*(\boldsymbol{x}_0)$。具体来说,在每次迭代过程中认为模型$f(\cdot)$近似线性,此时扰动后数据点$\boldsymbol{x}_0$对应于这次迭代的最小距离$r_i(\boldsymbol{x}_i)$为:

$$r_i(\boldsymbol{x}_i)=\arg\min_{\boldsymbol{r}_i}\|\boldsymbol{r}_i\|_2\tag{10}$$

s.t. $f(\boldsymbol{x}_i)+\nabla f(\boldsymbol{x}_i)^{\mathrm{T}}\boldsymbol{r}_i=0$

根据式(9)得到的闭解,继而推导得出针对非线性决策边界的最小扰动距离$r_i(\boldsymbol{x}_i)=-\dfrac{f(\boldsymbol{x}_i)}{\|\nabla f(\boldsymbol{x}_i)\|_2^2}\nabla f(\boldsymbol{x}_i)$,通过将每次迭代得到的扰动$\boldsymbol{r}_i$累加就可以得到针对当前数据点$\boldsymbol{x}_0$所需的最小扰动。

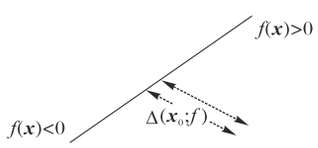


图3　线性分类器的决策边界

Fig. 3　Decision boundary of linear classifier

2.3.2 UAPs攻击

多数对抗攻击算法[14,25,33]针对单个输入样本生成对抗样本，从而达到攻击目的。而Moosavi-Dezfooli等[17]发现深度学习模型存在与输入样本无关的通用对抗扰动，这种扰动与目标模型结构和数据集特征相关。当通用对抗扰动叠加在数据集的输入样本上时得到对抗样本，得到的对抗样本大部分具有攻击能力。通用对抗扰动定义如下：假设μ表示数据集中数据的分布情况，δ表示分布μ上所有数据点希望攻击成功的比例，ξ用于度量扰动的大小，通用对抗扰动$\boldsymbol{\nu}$要满足如下两个条件：

$$\begin{aligned}&1)\ \|\boldsymbol{\nu}\|_p \leqslant \xi\\&2)\ \mathrm{P}(f(\boldsymbol{x}+\boldsymbol{\nu}) \neq f(\boldsymbol{x})) \geqslant 1-\delta\end{aligned} \tag{11}$$

Moosavi-Dezfooli等[17]提出的UAPs攻击算法通过在采样的少量数据点上迭代计算生成通用对抗扰动。每次迭代过程中，计算能够使当前数据点$\boldsymbol{x}_i$欺骗分类器的最小扰动$\Delta\boldsymbol{\nu}_i$，其优化目标描述为：

$$\Delta\boldsymbol{\nu}_i \leftarrow \arg\min_{\boldsymbol{r}} \|\boldsymbol{r}\|_2 \tag{12}$$

$$\text{s.t.}\quad f(\boldsymbol{x}_i+\boldsymbol{\nu}+\boldsymbol{r}) \neq f(\boldsymbol{x}_i)$$

最后，将在采样数据点上计算得到的扰动汇总到通用对抗扰动$\boldsymbol{\nu}$。为了保证汇总得到的通用对抗扰动满足$\|\boldsymbol{\nu}\|_p \leqslant \xi$，在每次迭代汇总时，对更新的扰动进行如下投影操作：

$$\mathcal{P}_{p,\xi}(\boldsymbol{\nu}) = \arg\min_{\nu'} \|\boldsymbol{\nu}-\boldsymbol{\nu}'\|_2 \tag{13}$$

$$\text{s.t.}\quad \|\boldsymbol{\nu}'\|_p \leqslant \xi$$

于是，ν的更新规则为$\boldsymbol{\nu} \leftarrow \mathcal{P}_{p,\xi}(\boldsymbol{\nu}+\Delta\boldsymbol{\nu}_i)$。直到满足预先定义的愚弄率后，算法停止迭代。愚弄率定义如下：

$$Err(X_v) \coloneqq \frac{1}{m}\sum_{i=1}^{m} 1_{f(\boldsymbol{x}_i+\boldsymbol{\nu}) \neq f(\boldsymbol{x}_i)} \geqslant 1-\delta \tag{14}$$

UAPs攻击算法在迭代过程中的扰动计算使用DeepFool算法[33]进行求解。最终，经过多次迭代后通过将数据点推送到模型决策边界另一边，达到对抗攻击的目的。

2.4 基于生成式神经网络生成的攻击方法

基于生成式神经网络生成的攻击方法利用自监督的方式，通过训练生成式神经网络来生成对抗样本。这类攻击方法的特点是一旦生成式模型训练完成，可非常高效地生成大量具有良好迁移攻击能力的对抗样本。典型的这类攻击方法有ATN(Adversarial Transformation Network)攻击[34]、UAN(Universal Adversarial Network)攻击[18]和AdvGAN攻击[35]。

2.4.1 ATN攻击

Baluja等[34]首次提出利用生成式神经网络生成对抗样本的ATN攻击方式，并设计了ATN用来生成对抗样本。ATN将一个输入样本转换为针对目标模型的对抗样本，ATN定义如下：

$$g_{f,\theta}(\boldsymbol{x}): \boldsymbol{x} \in X \rightarrow \boldsymbol{x}^{\mathrm{A}} \tag{15}$$

其中：$\boldsymbol{\theta}$表示神经网络g的参数，f表示为要攻击的目标网络。针对目标攻击问题，对ATN中参数$\boldsymbol{\theta}$的训练，可描述为如下的优化目标：

$$\arg\min_{\theta} \sum_{x_i \in x} \beta L_X(g_{f,\theta}(\boldsymbol{x}_i), \boldsymbol{x}_i) + L_Y(f(g_{f,\theta}(\boldsymbol{x}_i)), f(\boldsymbol{x}_i)) \tag{16}$$

其中：L_X为视觉损失函数，可用常见的l_2范数表示或者采用与文献[36]中类似的视觉感知相似性函数；L_Y为类别损失函数，定义为$L_Y = L_2(y', r(y,t))$，其中$y = f(\boldsymbol{x})$，$y' = f(g_f(\boldsymbol{x}))$，$t$是目标攻击的类别，$r(\cdot)$是重新排序函数[34]，它对$\boldsymbol{x}$进行修改，使$y_k \leqslant y_t, \forall k \neq t$。

ATN[34]可以训练为仅生成对抗扰动的P-ATN(Perturbations Adversarial Transformation Network)，这种情况下ATN一般选择残差的网络结构[8]就可有效地生成扰动。ATN还可以训练为直接生成对抗样本的AAE(Adversarial AutoEncoding)网络，这种情况下ATN结构采用自编码器可很好地生成对抗样本。通过这两种方法得到的对抗样本差异较大，AAE方法生成的对抗样本整体变化较为明显，而P-ATN方法生成的对抗样本扰动程度较小。总体而言，该方法生成对抗样本的速度较快，且其攻击能力较强，但迁移攻击能力较弱。

2.4.2 UAN攻击

Hayes等[18]提出了基于神经网络生成通用对抗扰动进行攻击的UAN攻击算法。UAN攻击通过训练一个简单的反卷积神经网络将一个在自然分布$\mathrm{N}(0,1)^{100}$上采样的随机噪声转换为通用对抗扰动。针对目标攻击问题，Hayes等[18]为反卷积神经网络的训练设计了如下的优化函数：

$$\begin{aligned}L_t = \max\Big\{&\max_{i \neq y^t} \log[f(\boldsymbol{\delta}'+\boldsymbol{x})]_i -\\&\log[f(\boldsymbol{\delta}'+\boldsymbol{x})]_t,\ -k\Big\} + \alpha\|\boldsymbol{\delta}'\|_p\end{aligned} \tag{17}$$

其中：模型损失函数选择了与C&W攻击[13]相同的损失函数，y^t为目标攻击选择的类别，α控制扰动大小优化项的相对重要程度。

在实际利用UAN攻击方式[18]进行攻击时，使用l_2或l_∞范数均可生成攻击能力较好的通用对抗扰动，其攻击能力强于先前提出的UAPs攻击[17]。

2.4.3 AdvGAN攻击

Xiao等[35]在基于神经网络生成的攻击算法中首次引入了生成式对抗网络[37]的思想，提出了包含生成器、鉴别器和攻击目标模型的AdvGAN。如图4所示，经过训练的AdvGAN可以将随机噪声转换为有效的对抗样本。

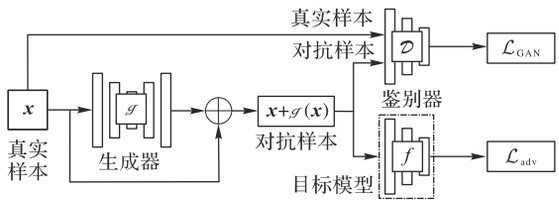


图4 AdvGAN架构

Fig. 4 Architecture of AdvGAN

与UAN攻击[18]中使用l_p范数对扰动大小的限制不同，AdvGAN利用生成式对抗网络中的“对抗损失”项，来保证对抗样本的真实性。AdvGAN中的对抗损失项采用与Goodfellow等[37]相同的定义：

$$\mathcal{L}_{\mathrm{GAN}} = \mathrm{E}_{\boldsymbol{x}} \log \mathcal{D}(x) + \mathrm{E}_{\boldsymbol{x}} \log(1-\mathcal{D}(\boldsymbol{x}+\mathcal{G}(\boldsymbol{x}))) \tag{18}$$

其中：生成器$\mathcal{G}(\cdot)$用于将输入噪声转化为对抗扰动，鉴别器$\mathcal{D}(\cdot)$的目的是尽可能使生成的对抗样本与原始输入样本具有较高的相似性。在目标攻击中，针对目标模型的误导损失项定义为：

$$\mathcal{L}_{\mathrm{adv}}^{\mathrm{f}} = \mathrm{E}_{\boldsymbol{x}}\, loss_f(\boldsymbol{x}+\mathcal{G}(\boldsymbol{x}), y^t) \tag{19}$$

其中：y^t为目标攻击的类别，$loss_f$为交叉熵损失函数。此外，为了明确量化扰动大小以及稳定GAN的训练过程[38]，对扰

动添加一个soft hinge损失项，如下所示：

$$\mathcal{L}_{\text{hinge}} = \mathrm{E}_{\boldsymbol{x}} \max\left(0, \left\| \mathcal{G}(\boldsymbol{x}) \right\|_2 - c\right) \tag{20}$$

其中c为指定的扰动大小。最终，针对AdvGAN训练的整体优化函数设计如下：

$$\mathcal{L} = \mathcal{L}_{\text{adv}}^{\text{f}} + \alpha\mathcal{L}_{\text{GAN}} + \beta\mathcal{L}_{\text{hinge}} \tag{21}$$

其中：α和β参数用来控制每个优化项的相对重要性，整体来说，$\mathcal{L}_{\text{GAN}}$损失项的目的是使生成的对抗扰动与原始样本相似，$\mathcal{L}_{\text{adv}}^{\text{f}}$损失项的目的是达到对抗攻击的效果。

2.5　其他的攻击方法

2.5.1　JSMA攻击

Papernot等[39]提出的JSMA（Jacobian-based Saliency Map Attack）算法是一种基于l_0范数约束下的攻击，通过修改图像中的几个像素来使模型对输入样本误分类。JSMA攻击算法利用显著图[40]表示输入特征对预测结果的影响程度，每次修改一个干净图像的像素，然后计算模型最后一层的输出对输入的每个特征的偏导。通过得到的前向导数，计算得出显著图[40]。最后利用显著图找到对模型输出影响程度最大的输入特征，通过修改这些对输出影响程度较大的特征点从而得到有效的对抗样本。

2.5.2　单像素攻击

在其他的攻击算法中，单像素攻击[41]是一种基于差分进化算法[42]的攻击算法。单像素攻击算法每次只修改样本数据点的1个像素值试图让模型误分类。实际应用中，这是一种极端的攻击方式。该方法对于简单的数据集有较好的攻击效果，比如MNIST数据集[43]。当输入图像的像素空间较大时，1个像素点的改变很难影响到分类结果，随着图像增大，算法的搜索空间也会迅速增大，使得算法性能下降。

2.5.3　stAdv攻击

Xiao等[44]提出了一种通过对图像样本进行空域变换来产生对抗样本的stAdv（spatially transformed Adversarial）攻击算法。该算法对局部图像特征进行平移、扭曲等操作实现针对输入样本的空域变换攻击。使用stAdv算法生成的对抗样本较于传统基于l_p范数距离度量生成的对抗样本更为真实，且针对目前采用对抗训练措施的模型具有很好的攻击效果。

2.5.4　BPDA攻击

破碎梯度策略[15]是一种用来针对FGSM[14]、I-FGSM[25]等基于梯度攻击方法的对抗防御方法。破碎梯度策略使用一个不可微的函数$g(\boldsymbol{x})$预处理输入样本，使训练得到的模型$f(g(\boldsymbol{x}))$在$\boldsymbol{x}$上不可微，使得攻击者计算不出用于对抗样本生成的梯度[15]。

Athalye等[45]针对破碎梯度策略，提出利用近似梯度生成对抗样本的BPDA（Backward Pass Differentiable Approximation）算法。BPDA算法在反向传播计算梯度时，使用一个可微的函数$h(\boldsymbol{x})$替代函数$g(\boldsymbol{x})$来近似获得梯度，生成对抗样本。

本文从扰动范数、攻击类型和攻击强度等角度对上述白盒对抗攻击算法进行了比较，总结分析了不同算法的优势及劣势。其中，对抗攻击类型分为单步迭代攻击和多步迭代攻击。单步迭代攻击算法生成对抗样本速度较快，而多步迭代攻击算法的攻击能力较强。对比分析的结果如表1所示，其中在攻击强度的对比结果中，*的数量代表攻击强度。

表1　对抗攻击算法总结

Tab. 1　Summary of adversarial attack algorithms

攻击算法	扰动范数	攻击类型	攻击强度	算法优势	算法劣势
L-BFGS[6]	l_2	单步	***	对抗样本有良好的迁移攻击能力，是第一个提出的对抗攻击算法	算法需要花费大量时间优化超参数c
C&W[13]	l_2、l_∞	迭代	*****	针对大部分蒸馏防御模型的攻击能力强且生成的扰动小	算法攻击效率低，需耗时寻找合适的超参数
FGSM[14]	l_∞	单步	***	生成效率非常高且扰动具有良好的迁移攻击能力	计算一次梯度生成对抗样本，对抗样本的扰动强度较大
I-FGSM[25]	l_∞	迭代	****	多步迭代生成对抗样本，攻击能力强	容易过拟合，迁移攻击能力较差
PGD[26]	l_∞	迭代	*****	比I-FGSM算法的攻击能力强	迁移攻击能力较差
MI-FGSM[27]	l_∞	迭代	****	兼具较好的攻击能力和迁移攻击能力，算法收敛速度更快	攻击能力比PGD算法差
DeepFool[33]	l_2、l_∞	迭代	****	精确计算得到的扰动更小	不具备目标攻击能力
UAPs[17]	l_2、l_∞	迭代	****	生成的通用对抗扰动具有较好的迁移攻击能力	无法保证对特定数据点的攻击成功率
ATN[34]	l_∞	迭代	****	可同时针对多个目标模型进行攻击，对抗样本更具多样性	训练生成网络寻找合适的超参数
UAN[18]	l_2、l_∞	迭代	****	生成扰动的速度快且攻击能力强于UAPs算法	生成模型的训练需花费一定时间
AdvGAN[35]	l_2	迭代	****	生成的对抗样本在视觉上与真实样本非常相似	对抗训练过程不稳定
JSMA[40]	l_0	迭代	***	生成的对抗样本与真实样本相似度高	生成的对抗样本不具备迁移攻击能力
单像素攻击[42]	l_0	迭代	**	可仅修改一个像素点进行攻击	计算量大，仅适用于尺寸较小的数据集
stAdv[44]	—	迭代	***	针对对抗训练防御有较好的攻击效果	只针对特定防御策略的模型攻击效果好
BPDA[45]	l_∞	迭代	***	可有效针对混淆梯度防御的模型进行攻击	只针对混淆梯度的防御进行攻击

3　应用场景下的白盒对抗攻击技术

对抗攻击技术同时给部署在大部分应用场景下的深度学习系统带来了安全威胁，诸多研究[9-11,46]已表明这类系统面临的被对抗攻击技术误导的风险。与在第2章中介绍的白盒对抗攻击技术可以直接向深度学习模型输入对抗样本不同，真实的应用场景下并不能直接操作深度学习系统的输入。本章介绍几类发生在不同应用场景下的白盒对抗攻击，通过不同应用场景的白盒对抗攻击说明针对当前部署的深度学习系统的对抗照片攻击和对抗贴纸攻击技术。

3.1　针对移动终端应用的攻击

随着深度学习技术的发展，越来越多的人工智能技术应

用在诸如手机、平板电脑等智能化的移动终端设备上，而这类部署在移动终端设备上的深度学习系统也面临着被对抗攻击技术攻击的风险。Kurakin 等[25]对部署在手机上的图像分类应用进行了攻击测试，首次展示了对抗攻击技术给移动终端设备上部署的深度学习系统带来的危害。

Kurakin[47]攻击的目标模型是一个手机图像分类应用，该应用基于 Inception 分类模型[48]构建，可对手机相机拍摄得到的照片进行分类。Kurakin 利用 FGSM[14]和 I-FGSM 攻击算法[25]分别针对 Inception 分类模型[48]生成对抗样本后，把得到的对抗样本打印为照片，这种照片被称为对抗照片。最后，使用手机相机拍摄输入对抗照片后实现对手机图像分类应用的攻击。Kurakin 的攻击使图像分类应用误分类了大部分拍摄得到的对抗照片，但 Kurakin 在利用手机相机拍摄对抗照片时，采取了固定相机的拍摄距离、角度和光线等措施。在现实的攻击场景中，并不能完全具备这些条件，但 Kurakin 团队的工作首次验证了针对真实应用场景下深度学习系统进行对抗攻击来干扰其正常工作的可行性，提供了一种在该场景下对抗攻击的思路。

3.2 针对人脸识别系统的攻击

人脸识别系统是目前深度学习技术在现实生活中较为成功的应用，其广泛地部署在安检、考勤、支付等诸多身份核验场景。Sharif 等[10]针对部署在真实场景下的人脸识别系统进行了对抗攻击测试，并提出一种对抗贴纸的攻击方式。针对人脸识别系统，在 Kurakin 等[25]的工作基础上，鉴于无法直接对输入图像像素进行修改的限制，使用攻击算法生成对抗扰动后，通过将扰动打印在贴纸上，最后将贴纸张贴在眼镜框区域来达到对抗攻击的目的，如图 5 所示。

图 5 对抗眼镜

Fig. 5 Adversarial glasses

人脸识别系统本质上是一个多分类的深度学习模型，针对人脸识别系统的目标攻击，其对抗样本生成可描述为如下的优化问题：

$$\arg\min_{\boldsymbol{r}} \sum_{\boldsymbol{x}\in X} softmaxloss(f(\boldsymbol{x}+\boldsymbol{r}), l) \tag{22}$$

其中：$\boldsymbol{r}$ 表示要生成的对抗扰动，l 表示要攻击的类别，X 表示攻击者的人脸数据集。为了使得生成的扰动更加“平滑”和“自然”，保证攻击的隐蔽性。使用全变差约束方法对 $\boldsymbol{r}$ 进行约束，全变差约束函数的定义如下：

$$TV(\boldsymbol{r}) = \sum_{i,j}((\boldsymbol{r}_{i,j} - \boldsymbol{r}_{i+1,j})^2 + (\boldsymbol{r}_{i,j} - \boldsymbol{r}_{i,j+1})^2)^{1/2} \tag{23}$$

针对 Kurakin 等[25]在将对抗样本打印为“对抗照片”时，未考虑到打印设备带来的数字图像与打印输出之间的色域误差。Sharif 等[10]定义了色域误差值来明确这种误差，NPS（Non-Printability Score）值的定义如下：

$$NPS(\tilde{p}) = \prod_{p\in P} |\tilde{p} - p| \tag{24}$$

其中：p 表示打印机能打印出来的颜色值，$\tilde{p}$ 表示为数字图像中的颜色值。

综合考虑对抗贴纸的隐蔽性和打印设备带来的打印误差，最终针对人脸识别系统的对抗贴纸生成的优化问题描述为：

$$\arg\min_{\boldsymbol{r}} \left(\sum_{\boldsymbol{x}\in X} softmaxloss\,(f(\boldsymbol{x}+\boldsymbol{r}), l) + (\lambda_1 TV(\boldsymbol{r}) + \lambda_2 NPS(\tilde{\boldsymbol{r}})) \right) \tag{25}$$

通过常见的优化算法，求解上述优化问题，即可得到对抗贴纸。Sharif 等[10]利用梯度下降法求解得到的“对抗眼镜”使得人脸识别系统以高置信度将攻击者误识别为攻击目标人，达到了攻击目的。

Sharif 等[10]针对人脸识别系统的攻击，较为全面地考虑到了打印设备带来的打印误差，其生成的“对抗贴纸”较 Kurakin 等[25]生成的“对抗照片”的对抗鲁棒性有了一定提升，但该攻击方式依然缺乏在复杂物理因素影响下的攻击能力。

3.3 针对自动驾驶系统的攻击

近年来，依托于深度学习技术的自动驾驶系统取得了越来越大的进步。基于深度学习技术决策的自动驾驶系统已普遍搭载应用在无人驾驶车辆上，而对抗攻击技术的发展也为这类自动驾驶系统的安全性带来了严重的危害。Eykholt 等[11]针对自动驾驶系统的交通标志识别功能进行了攻击测试，展示了对抗攻击技术给自动驾驶系统带来的安全威胁。

由于大部分对抗攻击算法在数字图像空间中生成的对抗扰动由于打印设备[10]、相机输入[11]等过程带来的误差无法有效针对应用场景下部署的深度学习系统进行攻击。Eykholt 等[11]设计了一种针对复杂物理场景下的深度学习系统进行攻击的 RP2（Robust Physical Perturbations）算法。RP2 攻击算法[11]尽可能考虑到不同光照、视角、距离等物理因素的影响，采用类似数据增强的思路来生成在复杂物理因素影响下鲁棒的对抗扰动。在针对运动中自动驾驶车辆的道路交通标志识别进行攻击测试时，Eykholt 等利用相机拍摄要攻击的“STOP”交通标志在各种光照、视角、距离条件下得到的图像加入用于生成对抗扰动贴纸的数据集；同时，在定义扰动生成的目标优化函数时，引入 Sharif 等[10]定义的 NPS 误差，以此减少打印设备带来的误差；最后采用 Sharif 等[10]贴纸攻击的方法，将生成的对抗扰动打印后张贴在要攻击的道路交通标志上，以此达到自动驾驶系统无法识别正确该标志的目的，如图 6 所示。

(a) 原始标志 (b) 对抗样本

图 6 “STOP”交通标志

Fig. 6 “STOP” traffic sign

Sharif 等[11]提出的 RP2 攻击算法进一步提高了复杂应用

场景下对抗扰动的攻击能力和对抗鲁棒性。利用RP2攻击算法结合对抗贴纸进行攻击的手段是目前针对应用场景下深度学习系统的主要对抗攻击技术。

4　对抗攻击实验

选择第2章中所介绍的算法分别进行以下两组实验:1)基于CIFAR10数据集[49],介绍了不同种类的攻击算法对目标模型分类准确率的影响程度;2)基于MNIST数据集[44],介绍了基于梯度优化的攻击算法在不同扰动强度设置下对目标模型分类准确率的影响和其生成的对抗样本的扰动差异程度。实验中用到的MNIST数据集[43]包含6万张训练图像和1万张测试图像,每张图像为28×28像素的灰度图像;CIFAR10数据集包含5万张训练图像和1万张验证图像,每张图像为32×32像素的RGB图像。

4.1　不同种类算法攻击模型的分类准确率对比

在介绍的4种不同类型的攻击算法中,分别选择C&W[13]、PGD[26]、DeepFool[33]和UAN[18]算法在CIFAR10数据集上进行攻击实验。实验中,攻击算法的扰动范数值统一l_∞范数值。攻击的目标模型分别为CNN模型、ResNet34模型和VGG19模型,训练后的目标模型在验证集上的分类准确率分别为81.97%、93.50%和92.03%。算法的攻击能力以其生成的对抗样本在目标模型上的分类准确率表示,分类准确率越低,则说明该算法的攻击能力越强。实验结果如表2所示,从结果中可以看到PGD算法的攻击能力较强,而ResNet34模型相比CNN模型和VGG19模型更容易被对抗样本攻击。

表2　CIFAR10验证集上的对抗样本分类准确率　单位:%

Tab. 2　Accuracy of adversarial examples classification on CIFAR10 validation set　unit:%

算法	CNN	ResNet34	VGG19
C&W	17.16	14.98	31.27
PGD	1.50	0.00	6.39
DeepFool	9.94	4.50	5.64
UAN	16.38	13.65	28.35

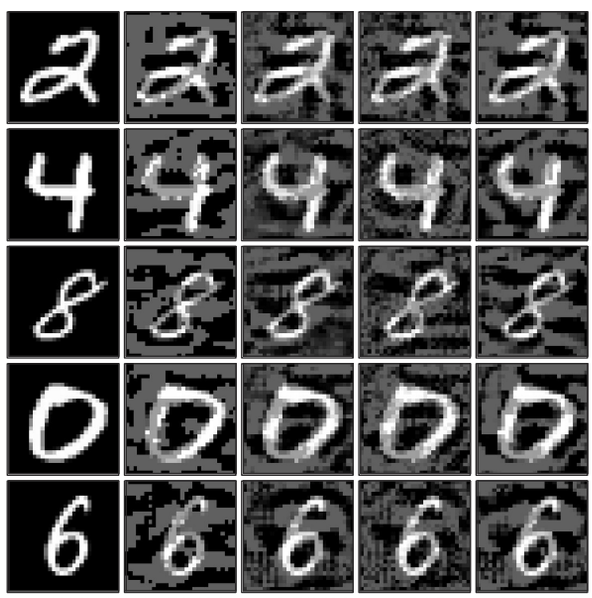
4.2　基于梯度优化算法攻击模型的扰动强度对比

选择基于梯度优化的FGSM[14]、I-FGSM[25]、PGD[26]和MI-FGSM[27]算法在MNIST数据集上进行攻击实验。实验中,攻击算法的扰动范数设置为l_∞范数,*eps*表示扰动强度。Goodfellow等[14]中将*eps*值设置为0.07以保证攻击成功率的同时限制扰动的大小。在实验中,为了比较不同扰动强度对模型分类准确率的影响程度,本文将*eps*值分别设置为0.05、0.1、0.2和0.3进行实验。攻击的目标模型为CNN模型,训练后的该模型在验证集上可以达到99.04%的分类准确率。实验结果如表3所示,可以看到随着扰动强度值(*eps*)的不断增大,攻击算法的攻击能力不断提高,生成的对抗样本能够被正确分类的可能性越小。4种攻击算法在*eps*值设置为0.3的情况下生成的对抗样本如图7所示,图中的每张字图表示一个原始输入样本或其对应的对抗样本。

表3　MNIST验证集上的对抗样本分类准确率　单位:%

Tab. 3　Accuracy of adversarial examples classification on MNIST validation set　unit:%

算法	*eps*值			
	0.05	0.1	0.2	0.3
FGSM	82.16	45.59	18.74	9.80
I-FGSM	65.12	9.70	0.78	0.69
PGD	63.71	9.15	0.74	0.63
MI-FGSM	66.55	12.83	1.07	0.77



(a) 原始样本　(b) FGSM算法　(c) I-FGSM算法　(d) PGD算法　(e) MI-FGSM算法

图7　MNIST数据集上的对抗样本(*eps* = 0.3)

Fig. 7　Adversarial examples on MNIST dataset (*eps* = 0.3)

5　结语

随着互联网技术的不断发展以及人工智能理论研究、相关方法的推广应用,有关对抗样本生成的方法将会层出不穷。研究对抗攻击技术,不但能促进深度学习模型可解释性研究的发展,进一步使对抗攻击防御技术得到完善,而且还可以利用对抗攻击技术促进一些相关领域的研究。目前,图像分类任务中针对白盒对抗攻击技术的研究已经取得了显著的成果。本文以图像分类任务作为切入点,对白盒对抗攻击技术进行了全面的回顾和研究。主要研究分析了当前白盒对抗攻击的几类方法,同时结合实际的应用场景,介绍了对抗攻击方法给深度学习模型带来的巨大影响和安全威胁。

本文通过对所调研的文献进行研究,按照对抗攻击算法生成对抗样本原理的不同,将主要的对抗攻击算法分为4类,按照分类对算法进行了全面的分析和阐述。其中,基于直接优化的算法生成的扰动较小,但存在寻找合适超参数耗时较长的问题;基于梯度优化的算法是目前对抗攻击算法中主要的一类算法,该类算法大多通过多步迭代计算梯度生成高质量对抗样本,其特点是针对无防御措施的模型攻击能力较强;基于决策边界分析的算法通过精确计算得到的扰动更小,但不具备目标攻击的能力;基于神经网络生成的算法是一种特殊的对抗攻击技术,这类算法通过训练一个生成模型

来生成对抗样本，一旦生成模型训练完成，在对抗样本生成阶段可非常高效地生成大量对抗样本。

未来，随着对抗攻击技术在自然语言处理[50-51]、语音识别[52-53]等任务上的推广应用，人工智能系统将面临更加严峻的安全挑战。为实现真正安全的深度学习应用，对抗攻击技术的研究将会受到长期的重视。其中的白盒对抗攻击技术依然会是重要的研究课题之一，其研究目标将朝着生成高隐蔽性、高鲁棒性和高迁移攻击能力的对抗样本进行；同时，白盒对抗技术在其他类型的任务中进行推广应用也将是一个可能的发展方向，而与之相对应的，针对黑盒对抗攻击技术的研究也颇受关注。面对黑盒对抗攻击技术给深度学习系统带来的危害，也将在接下来的研究中予以持续的关注。

参考文献（References）

[1] TABELINI L, BERRIEL R, PAIXÃO T M, et al. PolyLaneNet: lane estimation via deep polynomial regression[C]// Proceedings of the 25th International Conference on Pattern Recognition. Piscataway: IEEE, 2021: 6150-6156.

[2] SUN Y F, XU Q, LI Y L, et al. Perceive where to focus: learning visibility-aware part-level features for partial person re-identification [C]// Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2019: 393-402.

[3] DAHL G E, STOKES J W, DENG L, et al. Large-scale malware classification using random projections and neural networks [C]// Proceedings of the 2013 IEEE International Conference on Acoustics, Speech and Signal Processing. Piscataway: IEEE, 2013: 3422-3426.

[4] GROSSE K, PAPERNOT N, MANOHARAN P, et al. Adversarial perturbations against deep neural networks for malware classification [EB/OL]. (2016-06-16) [2021-06-15]. https://arxiv. org/pdf/1606. 04435. pdf.

[5] DU M, JIA R X, SONG D. Robust anomaly detection and backdoor attack detection via differential privacy [EB/OL]. (2019-11-16) [2021-06-15]. https://arxiv. org/pdf/1911. 07116. pdf.

[6] SZEGEDY C, ZAREMBA W, SUTSKEVER I, et al. Intriguing properties of neural networks [EB/OL]. (2014-02-19) [2021-06-15]. https://arxiv. org/pdf/1312. 6199. pdf.

[7] DENG J, DONG W, SOCHER R. ImageNet: a large-scale hierarchical image database [C]// Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2009: 248-255.

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

[9] CARLINI N, MISHRA P, VAIDYA T, et al. Hidden voice commands [C]// Proceedings of the 25th USENIX Security Symposium. Berkeley: USENIX Association, 2016: 513-530.

[10] SHARIF M, BHAGAVATULA S, BAUER L, et al. Accessorize to a crime: real and stealthy attacks on state-of-the-art face recognition [C]// Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security. New York: ACM, 2016: 1528-1540.

[11] EYKHOLT K, EVTIMOV I, FERNANDES E, et al. Robust physical-world attacks on deep learning models [C]// Proceedings of the 2018 IEE/CVFE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2018: 1625-1634.

[12] CARLINI N. A complete list of all (arXiv) adversarial example papers [EB/OL]. (2019-06-15) [2021-06-15]. https://nicholas. carlini. com/writing/2019/all-adversarial-example-papers. html.

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

[14] GOODFELLOW I J, SHLENS J, SZEGEDY C. Explaining and harnessing adversarial examples [EB/OL]. (2015-03-20) [2021-06-15]. https://arxiv. org/pdf/1412. 6572. pdf.

[15] PAPERNOT N, McDANIEL P, GOODFELLOW I, et al. Practical black-box attacks against machine learning [C]// Proceedings of the 2017 ACM Asia Conference on Computer and Communications Security. New York: ACM, 2017: 506-519.

[16] TRAMEÈR F, KURAKIN A, PAPERNOT N, et al. Ensemble adversarial training: attacks and defenses [EB/OL]. (2020-04-26) [2021-06-15]. https://arxiv. org/pdf/1705. 07204. pdf.

[17] MOOSAVI-DEZFOOLI S M, FAWZI A, FAWZI O, et al. Universal adversarial perturbations [C]// Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2017: 86-94.

[18] HAYES J, DANEZIS G. Learning universal adversarial perturbations with generative models [C]// Proceedings of the 2018 IEEE Symposium on Security and Privacy Workshops. Piscataway: IEEE, 2018: 43-49.

[19] 陈岳峰，毛潇锋，李裕宏，等. AI安全——对抗样本技术综述与应用[J]. 信息安全研究，2019，5(11)：1000-1007.（CHEN Y F, MAO X F, LI Y H, et al. AI security — Research and application on adversarial example [J]. Journal of Information Security Research, 2019, 5(11): 1000-1007.）

[20] 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.）

[21] 张玉清，董颖，柳彩云，等. 深度学习应用于网络空间安全的现状、趋势与展望[J]. 计算机研究与发展，2018，55(6)：1117-1142.（ZHANG Y Q, DONG Y, LIU C Y, et al. Situation, trends and prospects of deep learning applied to cyberspace security [J]. Journal of Computer Research and Development, 2018, 55(6): 1117-1142.）

[22] PAPERNOT N, McDANIEL P, WU X, et al. Distillation as a defense to adversarial perturbations against deep neural networks [C]// Proceedings of the 2016 IEEE Symposium on Security and Privacy. Piscataway: IEEE, 2016: 582-597.

[23] LIU D C, NOCEDAL J. On the limited memory BFGS method for large scale optimization [J]. Mathematical Programming, 1989, 45 (1/2/3): 503-528.

[24] KINGMA D P, BA J L. Adam: a method for stochastic optimization [EB/OL]. (2017-01-30) [2021-06-15]. https://arxiv. org/pdf/1412. 6980. pdf.

[25] KURAKIN A, GOODFELLOW I J, BENGIO S. Adversarial examples in the physical world [EB/OL]. (2017-02-11) [2021-06-15]. https://arxiv. org/pdf/1607. 02533. pdf.

[26] MĄDRY A, MAKELOV A, SCHMIDT L, et al. Towards deep learning models resistant to adversarial attacks [EB/OL]. (2019-09-04) [2021-06-15]. https://arxiv. org/pdf/1706. 06083. pdf.

[27] DONG Y P, LIAO F Z, PANG T Y, et al. Boosting adversarial attacks with momentum[C]// Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2018: 9185-9293.

[28] XIE C H, ZHANG Z S, ZHOU Y Y, et al. Improving transferability of adversarial examples with input diversity[C]// Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2019: 2725-2734.

[29] DONG Y P, PANG T Y, SU H, et al. Evading defenses to transferable adversarial examples by translation-invariant attacks[C]// Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2019: 4307-4316.

[30] KURAKIN A, GOODFELLOW I, BENGIO S. Adversarial machine learning at scale[EB/OL]. (2017-02-11)[2021-06-15]. https://arxiv. org/pdf/1611. 01236. pdf.

[31] POLYAK B T. Some methods of speeding up the convergence of iteration methods [J]. USSR Computational Mathematics and Mathematical Physics, 1964, 4(5): 1-17.

[32] SUTSKEVER I, MARTENS J, DAHL G, et al. On the importance of initialization and momentum in deep learning[C]// Proceedings of the 30th International Conference on Machine Learning. New York: JMLR. org, 2013: 1139-1147.

[33] 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. Piscataway: IEEE, 2016: 2574-2582.

[34] BALUJA S, FISCHER I. Adversarial transformation networks: learning to generate adversarial examples[EB/OL]. (2017-03-28)[2021-06-15]. https://arxiv. org/pdf/1703. 09387. pdf.

[35] XIAO C, LI B, ZHU J, et al. Generating adversarial examples with adversarial networks [C]// Proceedings of the 27th International Joint Conference on Artificial Intelligence. California: ijcai. org, 2018: 3905-3911.

[36] JOHNSON J, ALAHI A, LI F F. Perceptual losses for real-time style transfer and super-resolution[C]// Proceedings of the 2016 European Conference on Computer Vision, LNCS 9906. Cham: Springer, 2016: 694-711.

[37] GOODFELLOW I, POUGET-ABADIE J, MIRZA M, et al. Generative adversarial networks [C]// Proceedings of the 27th International Conference on Neural Information Processing Systems. Cambridge: MIT Press, 2014: 2672-2680.

[38] ISOLA P, ZHU J Y, ZHOU T H, et al. Image-to-image translation with conditional adversarial networks[C]// Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2017: 5967-5976.

[39] 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. Piscataway: IEEE, 2016: 372-387.

[40] SIMONYAN K, VEDALDI A, ZISSERMAN A. Deep inside convolutional networks: visualising image classification models and saliency maps[EB/OL]. (2014-04-19)[2021-06-15]. https://arxiv. org/pdf/1312. 6034. pdf.

[41] SU J W, VARGAS D V, SAKURAI K. One pixel attack for fooling deep neural networks [J]. IEEE Transactions on Evolutionary Computation, 2019, 23(5): 828-841.

[42] STORN R, PRICE K. Differential evolution — a simple and efficient heuristic for global optimization over continuous spaces [J]. Journal of Global Optimization, 1997, 11(4): 341-359.

[43] LECUN Y, CORTES C, BURGES C J C. The MNIST database of handwritten digits[DB/OL]. [2021-06-15]. http://yann. lecun. com/exdb/mnist/.

[44] XIAO C W, ZHU J Y, LI B, et al. Spatially transformed adversarial examples [EB/OL]. (2018-01-09)[2021-06-15]. https://arxiv. org/pdf/1801. 02612. pdf.

[45] ATHALYE A, CARLINI N, WAGNER D. Obfuscated gradients give a false sense of security: circumventing defenses to adversarial examples[C]// Proceedings of the 35th International Conference on Machine Learning. New York: JMLR. org, 2018: 274-283.

[46] GUO Y, WEI X X, WANG G Q, et al. Meaningful adversarial stickers for face recognition in physical world[EB/OL]. (2021-04-14)[2021-06-15]. https://arxiv. org/pdf/2104. 06728. pdf.

[47] KURAKIN A. Objects detection machine learning TensorFlow demo [EB/OL]. [2021-06-15]. https://play. google. com/store/apps/details? id=org. tensorflow. detect&hl=en.

[48] SZEGEDY C, VANHOUCKE V, IOFFE S, et al. Rethinking the inception architecture for computer vision[C]// Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2016: 2818-2826.

[49] KRIZHEVSKY A. Learning multiple layers of features from tiny images[EB/OL]. (2009-04-08)[2021-06-15]. https://www. cs. toronto. edu/~kriz/learning-features-2009-TR. pdf.

[50] LE T, WANG S H, LEE D. MALCOM: generating malicious comments to attack neural fake news detection models [C]// Proceedings of the 2020 IEEE International Conference on Data Mining. Piscataway: IEEE, 2020: 282-291.

[51] NIE Y X, WILLIAMS A, DINAN E, et al. Adversarial NLI: a new benchmark for natural language understanding [C]// Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA: Association for Computational Linguistics, 2020: 4885-4901.

[52] ŻELASKO P, JOSHI S, SHAO Y W, et al. Adversarial attacks and defenses for speech recognition systems[EB/OL]. (2021-03-31)[2021-06-15]. https://arxiv. org/pdf/2103. 17122. pdf.

[53] CHEN Y X, ZHANG J S, YUAN X J, et al. SoK: a modularized approach to study the security of automatic speech recognition systems[J]. ACM Transactions on Privacy and Security, 2022, 25(3): No. 17.

This work is partially supported by Natural Science Foundation of Gansu Province (20YF8FA080).

WEI Jiaxuan, born in 1983, Ph. D., engineer. Her research interests include artificial intelligence, network security.

DU Shikang, born in 1997, M. S. candidate. His research interests include adversarial machine learning.

YU Zhixuan, born in 1983, Ph. D. candidate. His research interests include machine learning, deep learning.

ZHANG Ruisheng, born in 1962, Ph. D., professor. His research interests include interpretable machine learning, complex network analysis, image recognition and analysis, service computing, chemoinformatics, bioinformatics.