计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.210100079

# 多媒体模型对抗攻防综述

陈　凯　　魏志鹏　　陈静静　　姜育刚
复旦大学计算机科学技术学院　上海 201203
上海市智能信息处理重点实验室　上海 200433
(20210240012@fudan.edu.cn)

**摘　要**　近年来,随着以深度学习为代表的人工智能技术的快速发展和广泛应用,人工智能正深刻地改变着社会生活的各方面。然而,人工智能模型也容易受到来自精心构造的“对抗样本”的攻击。通过在干净的图像或视频样本上添加微小的人类难以察觉的扰动,就能够生成可以欺骗模型的样本,进而使多媒体模型在推理过程中做出错误决策,为多媒体模型的实际应用部署带来严重的安全威胁。鉴于此,针对多媒体模型的对抗样本生成与防御方法引起了国内外学术界、工业界的广泛关注,并出现了大量的研究成果。文中对多媒体模型对抗攻防领域的进展进行了深入调研,首先介绍了对抗样本生成与防御的基本原理和相关背景知识,然后从图像和视频两个角度回顾了对抗攻防技术在多媒体视觉信息领域的发展历程与最新成果,最后总结了多媒体视觉信息对抗攻防技术目前面临的挑战和有待进一步探索的方向。

**关键词**:对抗攻击;对抗防御;深度学习;图像对抗样本;视频对抗样本
**中图法分类号**　TP18

## Adversarial Attacks and Defenses on Multimedia Models:A Survey

CHEN Kai,WEI Zhi-peng,CHEN Jing-jing and JIANG Yu-gang
School of Computer Science,Fudan University,Shanghai 201203,China
Shanghai Key Laboratory of Intelligent Information,Shanghai 200433,China

**Abstract**　In recent years,with the rapid development and wide application of deep learning,artificial intelligence is profoundly changing all aspects of social life. However,artificial intelligence models are also vulnerable to well-designed “adversarial examples”. By adding subtle perturbations that are imperceptible to humans on clean image or video samples,it is possible to generate adversarial examples that can deceive the model,which leads the multimedia model to make wrong decisions in the inference process,and bring serious security threat to the actual application and deployment of the multimedia model. In view of this,adversarial examples generation and defense methods for multimedia models have attracted widespread attention from both academic and industry. This paper first introduces the basic principles and relevant background knowledge of adversarial examples generation and defense. Then,it reviews the recent progress on both adversarial attack and defense on multimedia models. Finally,it summarizes the current challenges as well as the future directions for adversarial attacks and defenses.

**Keywords**　Adversarial attack,Adversarial defense,Deep learning,Image adversarial sample,Video adversarial sample

## 1　引言

近年来,以深度学习为代表的新一代人工智能技术取得了突破性进展,在一系列任务上已经达到甚至超过人类水平,如人脸识别[1-2]、物体识别[3-4]、语音识别[5]、下围棋[6]等。因此,人工智能正在被广泛应用于实际生活中的各方面。自动驾驶、智能服务机器人、智慧医疗、智能安防、智能投顾等人工智能新产品、新业态层出不穷,深刻地改变着人类的生产生活,并对人类文明发展和社会进步产生了广泛且深远的影响。

然而,在人工智能带给人们巨大便利的同时,其本身也存在一些安全性的问题。最近的研究表明[7],深度学习模型容易受到以对抗样本为代表的异常数据的攻击,导致其出现错误的预测结果。对抗样本是一种由攻击者精心设计的特殊样本,对输入样本添加细微的人眼不可察觉的扰动就能误导智能系统,引发智能系统以高置信度输出错误的结果。例如,修改图像中的一个像素点的值,便能误导图像分类系统[8];在视频帧的少量帧上添加细微的噪音,便能误导视频分类系统[9]。

对抗样本的出现,给人工智能系统的安全性带来了极大

到稿日期:2021-01-10　返修日期:2021-02-03
基金项目:国家自然科学基金(62032006);上海市科委项目(20511101000)
This work was supported by the National Natural Science Foundation of China(62032006) and Science and Technology Commission of Shanghai Municipality(20511101000).
通信作者:姜育刚(ygj@fudan.edu.cn)

的威胁，严重影响了人工智能系统在一些安全性要求高的场景中的应用部署，如自动驾驶、安防监控、刷脸支付、网络内容监管等。针对自动驾驶，攻击者可以使用对抗样本让汽车按照攻击者的意愿行动，这对交通安全造成了极大的隐患；在安防监控中，攻击者可以通过穿戴特定花纹的饰品来干扰监控系统，躲避监控系统的追踪；针对刷脸支付，攻击者可以通过佩戴特定眼镜，诱导人脸识别系统以较高的置信度将目标识别成别人，严重地威胁到其他用户的财产安全；针对网络内容监管，如暴恐或者色情视频检测等，发布者可以通过给视频添加细微的扰动，来欺骗暴恐内容检测系统，为网络内容监管带来了极大的挑战。

目前，国内外的学者针对多媒体模型已经提出了多种有效的对抗攻击方法。根据攻击者是否可以获取目标模型的内部结构信息，可将对抗攻击分为白盒攻击和黑盒攻击。在白盒攻击中，攻击者可以使用反向传播计算目标模型的梯度来生成对抗样本；而在黑盒攻击中，攻击者无法获取目标模型的内部结构信息，只能通过查询获得目标模型对于输入样本的预测类别或概率。为了在黑盒设置下生成对抗样本，一种思路是利用对抗样本的迁移性，即针对一个模型生成的对抗样本通常也可以欺骗另一个模型；另一种思路是通过查询目标模型，使用零阶优化方法来估计目标模型的梯度。虽然目前已经有许多强大的对抗攻击方法被提出，但是对抗攻击领域仍然面临着一些挑战。具体来说，白盒攻击容易被防御模型误导，从而导致较高的虚假鲁棒性；基于迁移性的黑盒攻击难以训练替代模型以及难以保证生成对抗样本的迁移性；使用梯度估计的黑盒攻击则受到了输入样本维度的限制，在输入样本具有较高维度的情况下(例如视频)，需要大量的查询来实现一次有效的攻击。因此，未来关于对抗攻击的研究将主要集中在解决这些挑战上。

为了抵御对抗样本的威胁并提升多媒体模型的鲁棒性，大量关于对抗防御的研究工作已经被提出。总体而言，现有的对抗防御方法大致可以分为 3 种：对抗样本检测、对抗训练和去噪。其中，对抗样本检测为反应策略，在多媒体模型构建之后进行防御；对抗训练和去噪为主动策略，在对抗样本生成之前提升多媒体模型的鲁棒性。虽然目前针对对抗样本的防御已经取得了一些进展，但是仍然没有一种对抗防御方法能够很好地平衡性能和效率。一方面，就性能而言，对抗训练无疑是最有效的对抗防御方法，但其计算成本很高，无法用于像视频一样的高维度数据。另一方面，就效率而言，许多基于对抗样本检测和去噪的防御方法都极易部署且防御效率高，但最近的研究工作表明这些防御方法的防御效果一般。

本文将对对抗样本生成与防御方法进行讨论，对图像和视频领域中的最新进展进行系统和全面的概述。本文第 2 节介绍了对抗样本相关的背景知识；第 3 节介绍了图像领域对抗样本的生成和防御方法；第 4 节介绍了视频领域对抗样本的生成和防御方法；第 5 节对图像和视频领域对抗攻防的发展趋势进行了展望；最后总结全文。

## 2 背景知识

本节首先介绍本文使用的数学符号，接着介绍对抗攻防相关的基本概念。

### 2.1 符号与定义

我们定义神经网络为函数 $F(\boldsymbol{x})=\boldsymbol{y}$，其中输入为 $\boldsymbol{x}\in R^n$，输出为 $\boldsymbol{y}\in[0,1]^m$。$l$ 为输入 $\boldsymbol{x}$ 的真实标签，其中 $n$ 为像素数量，$m$ 为类别数量。我们使用 $softmax$ 函数计算神经网络的输出，以保证输出向量 $\boldsymbol{y}$ 满足 $0\leqslant \boldsymbol{y}_i\leqslant 1$ 且 $\boldsymbol{y}_1+\cdots+\boldsymbol{y}_m=1$。因此，输出向量 $\boldsymbol{y}$ 被视为概率分布，即 $\boldsymbol{y}_i$ 被视为输入 $\boldsymbol{x}$ 具有类别 $i$ 的概率。分类器将预测概率最大的类别 $C(\boldsymbol{x})=\arg\max_i F(\boldsymbol{x})_i$ 输出为 $\boldsymbol{x}$ 的标签。

我们定义 $F$ 为包含 $softmax$ 函数的完整神经网络，$Z(\boldsymbol{x})=\boldsymbol{z}$ 为除 $softmax$ 之外所有层的输出，并且：

$$F(\boldsymbol{x})=softmax(Z(\boldsymbol{x}))=\boldsymbol{y} \tag{1}$$

$softmax$ 函数的输入被称为 logits，因此 $Z(x)=\boldsymbol{z}$ 是 logits。

通常情况下，神经网络由多层组成，因此有：

$$F=softmax\circ F_n\circ F_{n-1}\circ\cdots\circ F_1 \tag{2}$$

其中：

$$F_i(\boldsymbol{x})=\sigma(\boldsymbol{\theta}_i\cdot\boldsymbol{x})+\hat{\boldsymbol{\theta}}_i \tag{3}$$

其中，$\sigma$ 是非线性激活函数，常见的激活函数包括 tanh[10]，sigmoid，ReLU[11] 或 ELU[12]；$\boldsymbol{\theta}_i$ 和 $\hat{\boldsymbol{\theta}}_i$ 分别是模型权重矩阵和模型偏置向量，它们共同构成模型参数。

在多媒体模型中，上文提到的神经网络输入 $\boldsymbol{x}$ 通常为图像或者视频。针对图像，一张 $h\times w$ 像素的灰度图像是一个矩阵 $\boldsymbol{x}\in R^{hw}$，其中 $\boldsymbol{x}_i$ 表示被缩放到[0,1]范围内的像素值。而彩色图像则是三维矩阵，即 $\boldsymbol{x}\in R^{3\times h\times w}$。对于视频，一段 $T$ 帧、每帧 $h\times w$ 像素的 RGB 视频可以表示为一个四维的张量，即 $\boldsymbol{x}\in R^{T\times h\times w\times 3}$。

### 2.2 对抗攻击

#### 2.2.1 主要概念

对抗攻击通过在干净样本上添加细微扰动来生成对抗样本，从而导致模型产生错误的分类结果。所添加的细微扰动通常称为对抗扰动，对抗扰动由攻击者精心构造，且通常不会影响人类的识别，但是含有对抗扰动的样本却可以欺骗分类器。

给定一个经过训练的深度学习模型 $F$ 和干净的输入数据样本 $\boldsymbol{x}$，生成对抗样本 $\boldsymbol{x}'$ 的过程通常可以形式化为一个边界约束的优化问题：

$$\begin{aligned}&\min_{\boldsymbol{x}'}\|\boldsymbol{x}'-\boldsymbol{x}\|\\&\text{s.t. } C(\boldsymbol{x}')=l'\\&\quad C(\boldsymbol{x})=l\\&\quad l\neq l'\\&\quad \boldsymbol{x}'\in[0,1]^n\end{aligned} \tag{4}$$

其中，$l'$ 表示 $\boldsymbol{x}'$ 的目标标签；$\|\cdot\|$ 表示两个数据样本之间的距离；边界约束 $\boldsymbol{x}'\in[0,1]^n$ 意味着必须从有效边界内生成对抗样本。实际上，将每个像素值除以可达到的最大像素值(即 255)便能满足该边界约束，令 $\eta=\boldsymbol{x}'-\boldsymbol{x}$ 表示添加在 $\boldsymbol{x}$ 上的扰动。该优化问题的目标就是在最小化扰动的同时，使得模型将受到边界约束且添加了扰动的输入数据错误分类。

#### 2.2.2 威胁模型

精确定义威胁模型是进行对抗攻击的基础。基于不同的

场景和假设,威胁模型会确定攻击者能获取的信息以及攻击的目标,攻击者就可以在威胁模型的基础上部署特定的攻击方法。下文将从两个方面对威胁模型进行分类。

(1)攻击者能获取的信息

白盒攻击假设攻击者知道与训练的神经网络模型有关的所有信息,包括训练数据、模型架构、超参数、层数、激活函数和模型权重。通过计算模型的梯度便可以生成对抗样本。

黑盒攻击假设攻击者无法访问经过训练的神经网络模型,未掌握关于模型的任何信息,唯一有效的操作是对模型进行查询并获得相应的输出(预测类别概率或 top-1 预测类别)。因此,白盒攻击方法无法直接适用于黑盒设置。基于模型针对给定查询的输出,黑盒攻击又可分为基于分数和基于决策的攻击。

(2)攻击的目标

目标攻击将误导神经网络输出一个特定的类别,目标攻击通常发生在多类别分类问题中。例如,攻击者欺骗一个图像分类器,使其将所有的对抗样本预测为某个类别,如在人脸识别系统(Face Recognition System,FRS)中,攻击者试图伪装成已授权用户的面孔[13]。因此,对于特定输入,目标攻击通常会最大限度地增加模型对目标类别的预测概率。

非目标攻击不会给神经网络的输出指定一个特定的类别,输出的对抗类别可以是除真实标签之外的任意类别。例如,攻击者在人脸识别系统中将其面部误识别为任意面部,从而逃避检测[13]。与目标攻击相比,非目标攻击更易于实现,因为它有更多的选项和空间来重定向输出。

### 2.3　对抗防御

鉴于对抗样本带来的安全威胁,研究人员已经展开了大量研究,通过构建鲁棒的模型来实现对抗样本攻击的有效防御。与对抗攻击中的威胁模型类似,对抗防御也可以按照不同的维度进行分类。按照是否提高 DNN 的鲁棒性,可将对抗防御策略分为两种类型:1)反应式,在 DNN 构建后检测对抗样本;2)主动式,在攻击者生成对抗样本之前使 DNN 更加鲁棒。除此之外,对抗防御也可以按照是否经过理论证明分为两种类型:1)启发式,启发式防御的有效性仅仅通过实验验证,而没有从理论上得到证明,由于没有理论上的错误率保证,将来这些启发式防御可能会被新的攻击攻破;2)可证明,可证明的防御在明确定义的攻击类别下始终可以保持一定的准确性。本文将介绍反应式防御策略(如对抗样本检测)和主动式防御策略(如对抗训练、去噪),这几种对抗防御策略都属于启发式防御。

### 2.4　评价策略

#### 2.4.1　评价指标

(1)距离指标

在对抗样本的定义中,需要使用距离指标来度量对抗样本与干净样本的相似程度。目前,有 3 种广泛使用的距离度量可用于生成对抗样本,即$L_p$距离。

$L_p$距离被写为$\|\boldsymbol{x}'-\boldsymbol{x}\|_p$,其中 $p$ 范数$\|\cdot\|_p$被定义为:

$$\|\boldsymbol{v}\|_p=\left(\sum_{i=1}^{n}|\boldsymbol{v}_i|^p\right)^{\frac{1}{p}} \tag{5}$$

其中,$p$ 是实数,$n$ 是向量 $\boldsymbol{v}$ 的维度。

1)$L_0$距离测量的是满足$\boldsymbol{x}_i\neq\boldsymbol{x}_i'$的坐标 $i$ 的数量。因此,$L_0$距离对应于图像中被更改像素的数量。

2)$L_2$距离测量的是 $\boldsymbol{x}$ 和$\boldsymbol{x}'$之间的标准欧几里得距离。在最初的对抗攻击研究工作中使用了该距离度量[3]。

3)$L_\infty$距离测量的是所有像素的最大变化:

$$\|\boldsymbol{x}-\boldsymbol{x}'\|_\infty=\max(|\boldsymbol{x}_1-\boldsymbol{x}_1'|,\cdots,|\boldsymbol{x}_n-\boldsymbol{x}_n'|) \tag{6}$$

(2)准确率

准确率是图像分类等任务中常见的评价指标,在评估对抗样本时又被称为攻击成功率,它可以全面地评估模型的鲁棒性和对抗攻击的有效性。给定某种攻击方法生成的一组对抗样本,模型准确率越高,则该模型对于这种攻击越鲁棒;相应地,对抗样本的攻击成功率越低,说明该攻击方法的有效性就越差。

(3)攻击强度

攻击强度被定义为基于不同攻击方法的迭代次数或模型查询次数。它可以显示攻击的效率,以及模型对于攻击的抵抗能力。例如,针对 100 次迭代的攻击,准确率下降到零的防御模型被认为比另一种被 10 次迭代的相同攻击完全攻破的防御模型更能抵抗这种攻击,尽管两种防御模型的准确率都已下降为零。

值得一提的是,这 3 个指标是综合评估指标。通常,能够使用较少的迭代次数或模型查询次数生成较小的扰动,且极大地降低模型准确率的攻击方法更为强大。

#### 2.4.2　评估数据集

在图像领域中,MNIST[14],CIFAR-10[15]和 ImageNet[16]是 3 个广泛使用的图像数据集。MNIST 数据集用于手写数字识别,CIFAR-10 数据集和 ImageNet 数据集用于图像识别任务。CIFAR-10 数据集由来自 10 个类别的 6 万幅微小彩色图像(32×32)组成。ImageNet 数据集由来自 1 000 个类别的 14 197 122 幅图像组成。由于 ImageNet 数据集中有大量的图像,因此大多数对抗方法仅在一部分 ImageNet 数据集上进行评估。在视频领域中,UCF-101[17],HMDB-51[18]和 Kinetics-400[19]是 3 个广泛使用的视频动作识别数据集。其中,UCF-101 数据集是从 YouTube 中收集到的 13 320 个真实动作视频,包含 101 个类别;HMDB-51 数据集包含 51 个类别的 6 849 个视频片段;Kinetics-400 数据集由来自 400 个类别的 240 000 个视频构成。

## 3　图像领域对抗攻防

### 3.1　对抗攻击

本节按照攻击者能获取的信息对对抗攻击进行分类,从白盒攻击和黑盒攻击两个角度出发,概述图像领域中几种代表性的对抗攻击方法。

#### 3.1.1　白盒攻击

(1)L-BFGS 攻击

Szegedy 等[7]首次研究了针对深度神经网络的对抗攻击。他们提出了一种使用 L-BFGS 的方法,来寻找具有最小$L_p$范数的对抗扰动。他们将问题建模为有约束的最小化问题:

$$\begin{aligned}&\min\ \|\boldsymbol{x}-\boldsymbol{x}'\|_p\\&\text{s.t. } C(\boldsymbol{x}')=l'\\&\qquad\boldsymbol{x}'\in[0,1]^n\end{aligned} \tag{7}$$

其中，$\|\boldsymbol{x}-\boldsymbol{x}'\|_p$是对抗扰动的$L_p$范数。然而，这个优化问题难以解决，因此他们转而求解以下问题：

$$\begin{aligned}&\min c\cdot\|\boldsymbol{x}-\boldsymbol{x}'\|_p+loss(\boldsymbol{x}',l')\\&\text{s.t. }\boldsymbol{x}'\in[0,1]^n\end{aligned}\tag{8}$$

其中，$c$是超参数；$loss(\boldsymbol{x}',l')$用于计算分类器的损失，也叫对抗损失，常用的损失函数是 $softmax$ 交叉熵。在分类器具有凸损失函数的情况下，式(8)可以求得精确解，但 DNN 的损失函数一般是非凸的，因此只能求得一个近似解。Szegedy 等通过线性搜索 $c>0$，可以找到一个与干净样本 $\boldsymbol{x}$ 距离最小的 $\boldsymbol{x}'$，并同时欺骗图像分类器。

L-BFGS 攻击生成的对抗扰动具有较小的$L_p$范数，但是在对抗性方面还存在较大的提升空间。

(2)快速梯度符号方法

快速梯度符号方法(Fast Gradient Sign Method，FGSM)[20]是一种高效的非目标攻击方法，它在干净样本的$L_\infty$距离内生成对抗样本。FGSM 同时对所有像素沿对抗损失的梯度符号方向进行单步更新，更新过程如下：

$$\boldsymbol{x}'=\boldsymbol{x}+\varepsilon\cdot sign(\nabla_x loss(\boldsymbol{x},l))\tag{9}$$

其中，$\nabla_x loss(\boldsymbol{x},l)$计算对抗损失的梯度；$sign()$表示符号函数；$\varepsilon$是扰动的大小，限制了扰动的$L_\infty$范数。值得注意的是，FGSM 主要是为了快速地生成对抗样本，并不是为了生成具有最小对抗扰动的对抗样本。

FGSM 沿着对抗损失的梯度符号方向下降可以最大化目标类别的概率，从而扩展成一种目标攻击方法(目标 FGSM)[21]：

$$\boldsymbol{x}'=\boldsymbol{x}-\varepsilon\cdot sign(\nabla_x loss(\boldsymbol{x},l'))\tag{10}$$

FGSM 因为仅使用反向传播来计算扰动，所以生成对抗样本的速度比 L-BFGS 攻击更快。因此，FGSM 也可以满足需要生成大量对抗样本的需求。例如，在对抗训练中，可以使用 FGSM 为训练集中的所有样本生成对抗样本。此外，还可以在运行 FGSM 之前对干净样本进行随机扰动[22]，从而增强 FGSM 生成对抗样本的性能和多样性。

(3)基本迭代方法和投影梯度下降攻击

基本迭代方法(Basic Iterative Method，BIM)[23]通过沿对抗损失的梯度符号方向多次迭代运行更精细的优化来提高 FGSM 的性能。在每次迭代中，BIM 以更小的步长执行 FGSM，并将更新后的扰动图像的像素值裁剪到有效范围内以避免每个像素发生较大的变化。具体来说，首先做如下设置：

$$\boldsymbol{x}_0'=\boldsymbol{x}\tag{11}$$

然后，在每次迭代时：

$$\boldsymbol{x}_t'=Clip_\varepsilon(\boldsymbol{x}'_{t-1}+\alpha\cdot sign(\nabla_x loss(\boldsymbol{x}'_{t-1},l)))\tag{12}$$

其中，迭代次数 $T$ 由 $\min(\varepsilon+4,1.25\varepsilon)$决定；$\boldsymbol{x}_t'$表示第 $t$ 次迭代时的扰动图像；$Clip_\varepsilon(\boldsymbol{x}')=\min\{255,\boldsymbol{x}+\varepsilon,\max\{0,\boldsymbol{x}-\varepsilon,\boldsymbol{x}'\}\}$，限制了每次迭代时生成扰动图像的改变；$\alpha$ 表示步长(通常，$\alpha=1$)，且 $\alpha T=\varepsilon$。投影梯度下降(Projected Gradient Descent，PGD)[24]攻击可以被视为没有 $\alpha T=\varepsilon$ 约束的 BIM，而且 PGD 使用干净样本$L_\infty$距离内的任意一点进行随机初始化。

与将 FGSM 扩展为目标 FGSM 类似，BIM 也可以被扩展为迭代最不可能类别方法(Iterative Least-likely Class Method，ILCM)[23]。ILCM 选择图像分类器预测的概率最小的类别为目标类别，并最大化目标类别的概率：

$$\boldsymbol{x}_0'=\boldsymbol{x}\tag{13}$$

$$l'=\arg\min_i F(\boldsymbol{x})_i\tag{14}$$

$$\boldsymbol{x}_t'=Clip_\varepsilon(\boldsymbol{x}'_{t-1}-\alpha\cdot sign(\nabla_x loss(\boldsymbol{x}'_{t-1},l')))\tag{15}$$

BIM/PGD 攻击启发式地搜索干净样本 $\boldsymbol{x}$ 的$L_\infty$距离内损失值最大的样本。当扰动强度(扰动的$L_p$范数)受限时，这种对抗样本是最具攻击性且最可能欺骗图像分类器的样本点，找到这些对抗样本有助于发现 DNN 的弱点。

(4)基于雅可比的显著性图攻击

基于雅可比的显著性图攻击(Jacobian-Based Saliency Map Attack，JSMA)[25]可以用$L_0$范数较小的扰动来欺骗图像分类器。实际上，JSMA 的目标是仅仅修改图像中的几个像素，而不是扰动整个图像。JSMA 是一种贪婪的攻击方法，每次迭代时修改一个对输出影响最大的像素来增加目标类别的概率。该方法首先计算了 logits 输出对于 $\boldsymbol{x}$ 的雅可比矩阵：

$$J(\boldsymbol{x})=\frac{\partial Z(\boldsymbol{x})}{\boldsymbol{x}}=\left[\frac{\partial Z_j(\boldsymbol{x})}{\boldsymbol{x}_i}\right]_{i\in1,\cdots,n;j\in1,\cdots,m}\tag{16}$$

雅可比矩阵确定了 $\boldsymbol{x}$ 的每个像素如何影响不同类别的 logits 输出。根据雅可比矩阵，JSMA 计算了一个显著性图 $S(\boldsymbol{x},l')$来选择应该被扰动的像素，以获得预期的 logits 输出变化：

$$S_i(\boldsymbol{x},l')=\begin{cases}0, & J_{il'}(\boldsymbol{x})<0\text{ 或}\sum_{j\neq l'}J_{ij}(\boldsymbol{x})>0\\ J_{il'}(\boldsymbol{x})\left|\sum_{j\neq l'}J_{ij}(\boldsymbol{x}),\right. & \text{否则}\end{cases}\tag{17}$$

具体来说，JSMA 每次迭代时扰动$S_i(\boldsymbol{x},l')$值最大的元素 $\boldsymbol{x}_i$，以显著增加目标类别的 logits 输出(目标类别的可能性上升)并减少其他类别的 logits 输出(其他类别的可能性下降)。重复此过程，直至超过了设定的可修改像素的阈值或成功地改变了预测类别。因此，JSMA 可以通过仅扰动一小部分的像素来影响 $Z(\boldsymbol{x})$并欺骗 DNN。但是，由于计算雅可比矩阵需要极大的计算开销，JSMA 的运行速度较慢。

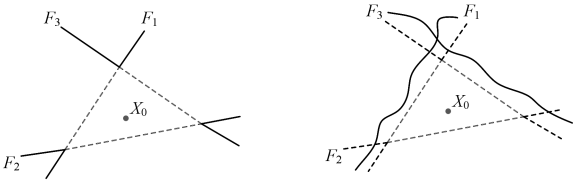
(5)DeepFool

DeepFool[26]是一种在$L_2$距离度量下生成对抗样本的非目标攻击方法。与前面介绍的 L-BFGS 攻击相比，DeepFool 可以更加高效地在$L_2$距离度量下生成与干净样本更接近的对抗样本。

对于仿射分类器 $F(\boldsymbol{x})=\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}+b$(其中 $\boldsymbol{w}$ 是仿射分类器的权重，$b$ 是仿射分类器的偏置)，改变干净样本 $\boldsymbol{x}$ 类别的最小扰动是从 $\boldsymbol{x}$ 到决策边界超平面 $\mathcal{F}=\{\boldsymbol{x}:\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}+b=0\}$的距离，即 $-\left(\frac{F(\boldsymbol{x})}{\|\boldsymbol{w}\|^2}\right)\boldsymbol{w}$。对于一般的可微二元分类器 $F$，为解决高维中的非线性，DeepFool 假设 $F$ 在 $\boldsymbol{x}'_t$周围是线性的并迭代地计算扰动$\eta_t$：

$$\arg\min_{\eta_t}\|\eta_t\|_2$$
$$\text{s.t. } F(\boldsymbol{x}_t')+\nabla_x F(\boldsymbol{x}_t')^{\mathrm{T}}\eta_t=0 \tag{18}$$

这个过程一直持续到 $F(\boldsymbol{x}_t')\neq F(\boldsymbol{x})$，并且最小扰动最终由 $\eta_t$ 的总和来近似。该方法也可以扩展到攻击一般的多元分类器，其中问题转化为计算从 $\boldsymbol{x}$ 到由所有类别之间的决策边界形成的凸多面体表面的距离，如图 1 所示[26]。




(a)线性模型　　(b)非线性模型

图 1　所有类别之间的决策边界形成的凸多面体[26]

Fig. 1　Convex polyhedron formed by decision boundaries between all classes[26]

DeepFool 每次迭代时生成 $L_2$ 范数最小的对抗扰动，将 $\boldsymbol{x}$ 一步步地移动到决策边界外，直到出现分类错误为止。与 FGSM 生成的对抗样本相比，DeepFool 生成的对抗样本在保持相当的对抗性的同时，具有更小的对抗扰动。

(6)Carlini 和 Wagner(C&W)攻击

C&W 攻击[27]可以在 $L_0$，$L_2$ 和 $L_\infty$ 距离度量下生成对抗样本，分别被称为 $CW_0$，$CW_2$ 和 $CW_\infty$ 攻击。与 L-BFGS 攻击类似，C&W 攻击求解的优化问题如下：

$$\min_{\boldsymbol{x}'}\|\boldsymbol{x}-\boldsymbol{x}'\|_p+c\cdot f(\boldsymbol{x}',l')$$
$$\text{s.t. } \boldsymbol{x}'\in[0,1]^n \tag{19}$$

其中，$c\cdot f(\boldsymbol{x}',l')$ 是反映对抗攻击成功与否的损失函数。C&W 攻击采用以下形式的 $f(\boldsymbol{x}',l')$ 进行有效的目标攻击：

$$f(\boldsymbol{x}',l')=\max(\max_{i\neq l'}(Z(\boldsymbol{x}')_i)-Z(\boldsymbol{x}')_{l'},-\mathscr{K}) \tag{20}$$

其中，$\max_{i\neq l'}(Z(\boldsymbol{x}')_i)-Z(\boldsymbol{x}')_{l'}\leqslant 0$ 表示对抗样本 $\boldsymbol{x}$ 对类别 $l'$ 的预测概率最高，因此目标攻击成功，而 $\max_{i\neq l'}(Z(\boldsymbol{x}')_i)-Z(\boldsymbol{x}')_{l'}>0$ 表示使用 $\boldsymbol{x}$ 进行目标攻击是不成功的；$\mathscr{K}\geqslant 0$ 是调整参数，$\mathscr{K}$ 的作用是确保 $Z(\boldsymbol{x}')_{l'}$ 和 $\max_{i\neq l'}(Z(\boldsymbol{x}')_i)$ 之间保持恒定的间隔，C&W 攻击在目标攻击时设置 $\mathscr{K}=0$，而在进行基于迁移的攻击时设置的 $\mathscr{K}$ 较大。

与 L-BFGS 攻击方法中使用边界约束的 L-BFGS 来限制 $\boldsymbol{x}'\in[0,1]^n$ 不同，C&W 攻击将 $\boldsymbol{x}$ 替换为 $\frac{1+\tanh(\boldsymbol{w})}{2}$ 来消除边界约束，其中 $\boldsymbol{w}\in R^n$。通过使用这样的变量替换，在确保优化过程中 $\boldsymbol{x}'$ 始终位于 $[0,1]$ 范围内的同时，将式(19)中的优化问题变成了优化 $\boldsymbol{w}$ 的无约束最小化问题。$CW_2$ 攻击可以表示为：

$$\min_{\boldsymbol{w}}\left\|\boldsymbol{x}-\frac{1+\tanh(\boldsymbol{w})}{2}\right\|_2+c\cdot f\left(\frac{1+\tanh(\boldsymbol{w})}{2},l'\right) \tag{21}$$

由于 $L_0$ 距离是不可微的，因此 $CW_0$ 攻击是迭代进行的。在每次迭代中，对于生成对抗样本影响不大的一些像素会被固定，而像素的重要性是由 $CW_2$ 攻击决定的。如果 $CW_2$ 攻击针对剩余的像素无法生成对抗样本，则迭代停止。

$CW_\infty$ 攻击也是一种迭代攻击，在每次迭代中用新的惩罚项来代替 $L_2$ 项：

$$\min_{\eta}\sum_i[(\eta_i-\tau)^+]+c\cdot f(\boldsymbol{x}+\eta,l') \tag{22}$$

在每次迭代时，如果所有的 $\eta_i<\tau$，则将 $\tau$ 减小 0.1。$CW_\infty$ 攻击中的 $\tau$ 被视为 $L_\infty$ 距离的估计。

C&W 攻击是目前最强的白盒攻击方法之一，它攻破了许多被证明是成功的对抗防御策略。

除了上述经典的白盒攻击方法之外，最近还有一些新的白盒攻击方法被提出以评估模型的鲁棒性，例如快速自适应边界攻击(Fast Adaptive Boundary Attack，FAB)[28]和自动攻击(AutoAttack，AA)[29]等。FAB 针对 $L_p$ 范数可以找到改变给定输入类别所需的最小扰动。虽然 FAB 使用了与 DeepFool 类似的模型线性近似和在决策超平面上进行投影的策略，但其具体实现过程比 DeepFool 更加复杂。AA 是由 4 种攻击方法组成的无参数集成攻击方法，具体包括两种具有不同损失函数的自动 PGD 攻击以及两种互补攻击(FAB 和方形攻击[30])。AA 已被证明是迄今为止最先进的攻击方法之一[29]。

3.1.2　黑盒攻击

(1)基于迁移的攻击

在黑盒设置下，攻击者无法访问目标模型内部，只能通过输入 $\boldsymbol{x}$ 从目标模型获取输出(预测类别概率或 top-1 预测类别)。为了实现对抗攻击，一种解决方案是利用对抗样本的迁移性。迁移性是由于不同的模型在同一个数据点周围学习到了相似的决策边界，使得针对一个模型生成的对抗样本对其他模型也有效。

基于迁移的攻击首先训练一个替代模型，然后针对替代模型使用白盒攻击方法来生成对抗样本，生成的对抗样本由于迁移性可能会对目标模型保持对抗性。但是，大多数白盒攻击方法生成对抗样本的迁移性较差，下面将概述几种已经提出的用于提高对抗样本迁移性的方法。这些方法都在保持白盒攻击方法的白盒攻击成功率的同时，显著提高了黑盒攻击的成功率，而且它们还能组合使用，从而进一步提高了对抗样本的迁移性。

1)基于多模型的集成攻击

集成的概念可以应用于对抗攻击，Liu 等[31]发现如果一个对抗样本能够对多个模型保持对抗性，那么它也更有可能迁移到其他模型。基于多模型的集成攻击的基本思想便是通过攻击多个替代模型来生成对抗样本，它通过求解以下优化问题：

$$\min_{\boldsymbol{x}'}-\log((\sum_{i=1}^{k}\omega_i F_i(\boldsymbol{x}')\cdot\mathbf{1}_{l'})+c\cdot\|\boldsymbol{x}-\boldsymbol{x}'\|_2 \tag{23}$$

其中，$k$ 是替代模型的数量；$\omega_i$ 是集成权重($\omega_i\geqslant 0$，且 $\sum_{i=1}^{k}\omega_i=1$)；$F_i(\boldsymbol{x}')$ 是第 $i$ 个替代模型在给定输入时的预测概率；$\mathbf{1}_{l'}$ 是 $l'$ 的独热(one-hot)编码。值得注意的是，式(23)是目标攻击，非目标攻击也可以类似地实现。除了融合多个替代模型的预测概率之外，还可以融合多个替代模型的损失和 logits[32]。为了方便，我们将其分别称为预测融合、损失融合和 logits 融合。具体来说，$k$ 个模型的 logits 融合为：

$$\sum_{i=1}^{k}\omega_i Z_i(\boldsymbol{x}') \tag{24}$$

其中，$Z_i(\boldsymbol{x}')$为第 $i$ 个替代模型的 logits。$k$ 个模型的损失融合为：

$$\sum_{i=1}^{k}\omega_i loss_i(\boldsymbol{x}',l') \tag{25}$$

其中，$loss_i(\boldsymbol{x}',l')$为第 $i$ 个替代模型的损失。不同的融合方式会导致不同的攻击性能，实验结果表明 logits 融合的性能优于预测融合和损失融合。

2)动量方法

与 FGSM 相比，BIM 在每次迭代时沿对抗损失的梯度符号方向“贪婪地”移动对抗样本，使对抗样本很容易陷入异常的难以迁移到其他模型的决策区域(对应于优化过程中较差的局部极大值)并“过拟合”模型，从而导致生成的对抗样本的迁移性较差。

受到动量方法[33]的启发，动量迭代快速梯度符号方法(Momentum Iterative Fast Gradient Sign Method，MI-FGSM)[32]将动量项整合到 BIM 攻击过程中，通过在每次迭代时沿对抗损失的梯度方向积累速度向量来摆脱较差的局部极小值或极大值。MI-FGSM 的更新过程如下：

$$\boldsymbol{g}_0=0,\boldsymbol{x}_0'=\boldsymbol{x} \tag{26}$$

$$\boldsymbol{g}_t=\mu\cdot\boldsymbol{g}_{t-1}+\frac{\nabla_x loss(\boldsymbol{x},l)}{\|\nabla_x loss(\boldsymbol{x},l)\|_1} \tag{27}$$

$$\boldsymbol{x}_t'=Clip_{\varepsilon}(\boldsymbol{x}_{t-1}'+\alpha\cdot sign(\boldsymbol{g}_t)) \tag{28}$$

其中，$\boldsymbol{g}_t$是前 $t$ 次迭代的积累梯度；$\mu$ 是$\boldsymbol{g}_t$的衰减因子。如果 $\mu$ 等于 0，MI-FGSM 便退化为 BIM；为了消除每次迭代时因当前梯度$\nabla_x loss(\boldsymbol{x},l)$大小不同而造成的影响，将其通过自身的 $L_1$ 距离(任何距离度量都是可行的)进行归一化。除了动量方法之外，Nesterov 加速梯度也可以被用来提高对抗样本的迁移性[34]。

3)多样化输入方法

从不同的角度来看，Dong 等[32]认为生成一个对抗样本和训练一个模型类似，而对抗样本的迁移性也类似于模型的泛化能力。动量方法和集成方法的引入提高了对抗样本的迁移性，这在一定程度上证明了对抗样本的生成和模型的训练是类似的，且可以使用其他提高模型泛化能力的方法来提高对抗样本的迁移性。

受到数据增强方法[35-37]的启发，多样化输入方法(Diverse Input Method，DIM)[38]通过创建多样化输入模式来提高对抗样本的迁移性，将其与 BIM 相结合，可得到多样化输入迭代快速梯度符号方法(Diverse Inputs Iterative Fast Gradient Sign Method，DI$^2$-FGSM)。DI$^2$-FGSM 的更新过程类似于 BIM，只是将式(12)替换为：

$$\boldsymbol{x}_t'=Clip_{\varepsilon}(\boldsymbol{x}_{t-1}'+\alpha\cdot sign(\nabla_x loss(T(\boldsymbol{x}_{t-1}';p),l))) \tag{29}$$

从式(29)可以发现，DI$^2$-FGSM 在每次迭代时，以概率 $p$ 将随机变换函数 $T(\boldsymbol{x}_{t-1}';p)$应用于输入图像。其中，随机变换函数 $T(\boldsymbol{x}_{t-1}';p)$的表达式如下：

$$T(\boldsymbol{x}_{t-1}';p)=\begin{cases}T(\boldsymbol{x}_{t-1}'), & \text{概率 } p\\ \boldsymbol{x}_{t-1}', & \text{概率 } 1-p\end{cases} \tag{30}$$

具体来说，DI$^2$-FGSM 使用的随机变换包括随机调整大小(将输入图像的大小调整为随机大小)和随机填充(以随机方式在输入图像边缘填充 0 元素)。随机变换函数中的随机变换概率 $p$ 平衡了 DI$^2$-FGSM 的白盒攻击性能和黑盒攻击性能。如果 $p=0$，DI$^2$-FGSM 就会退化为 BIM，从而导致生成的对抗样本过拟合被攻击的模型。如果 $p=1$，则只将随机变换后的输入图像用于对抗攻击，虽然这样可以显著提高黑盒攻击的性能，但同时也会使得白盒攻击的性能大幅下降。除了数据增强方法之外，提高模型泛化能力的 Dropout 技术也可以被用来提高对抗样本的迁移性[39]。

基于迁移攻击中的替代模型对于攻击者完全透明，因此可以使用快速、强大的白盒攻击方法来攻击替代模型以生成对抗样本。但是，在目标攻击中，针对替代模型生成的对抗样本很难迁移到目标模型，因此黑盒攻击的成功率较低。

(2)基于分数的攻击

在此黑盒设置下，攻击者可以提供任意输入 $\boldsymbol{x}$，来查询所有类别的预测概率 $\boldsymbol{y}$。基于分数的攻击通过迭代查询目标模型近似出梯度的估计值，然后使用白盒攻击方法和梯度估计值来生成对抗样本以欺骗目标模型。下面将概述几种通过查询高效估计梯度的算法，然后说明如何使用估计的梯度来生成对抗样本。

1)基于零阶优化的攻击

基于零阶优化(Zeroth Order Optimization，ZOO)的攻击[40]通过观察当调整 $\boldsymbol{x}$ 的像素值时预测类别概率的变化来估计梯度。受到 C&W 攻击的启发，ZOO 将 C&W 攻击中的 $f(\cdot)$修改为一种新的损失函数：

$$f(\boldsymbol{x}',l')=\max(\max_{i\neq l'}(\log[F(\boldsymbol{x}')]_i)-\log[F(\boldsymbol{x}')]_{l'},-\mathcal{K}) \tag{31}$$

并使用对称差商[41]来估计梯度和 Hessian：

$$\frac{\partial f(\boldsymbol{x})}{\partial \boldsymbol{x}_i}\approx\frac{f(\boldsymbol{x}+h\boldsymbol{e}_i)-f(\boldsymbol{x}-h\boldsymbol{e}_i)}{2h} \tag{32}$$

$$\frac{\partial^2 f(\boldsymbol{x})}{\partial \boldsymbol{x}_i^2}\approx\frac{f(\boldsymbol{x}+h\boldsymbol{e}_i)-2f(\boldsymbol{x})+f(x-h\boldsymbol{e}_i)}{h^2} \tag{33}$$

其中，$\boldsymbol{e}_i$表示第 $i$ 个分量为 1 的标准基向量；$h$ 是一个较小的常数。ZOO 使用 Adam[42]和牛顿方法来生成对抗样本。

通过利用梯度和 Hessian 的估计，ZOO 不需要访问目标模型内部就可以进行攻击，且攻击性能与 C&W 的白盒攻击性能相当。尽管 ZOO 之后使用了降维、分层攻击和重要性采样等加速技术，但坐标梯度估计方法仍然需要对目标模型进行过多的查询，使查询效率陷入瓶颈。

2)限制查询攻击

为了高效地估计梯度，限制查询攻击[43]使用了自然进化策略[44](Natural Evolutionary Strategies，NES)，这是一种基于搜索分布思想的无导数优化方法。NES 不是直接最大化目标函数 $f(\boldsymbol{x})$，而是最大化在搜索分布 $\pi(\boldsymbol{\theta}|\boldsymbol{x})$下损失函数的期望值。与有限差分方法相比，NES 可以通过更少的查询次数来进行梯度估计。对于损失函数 $f(\cdot)$和当前一组参数 $\boldsymbol{\theta}$，有：

$$E_{\pi(\boldsymbol{\theta}|\boldsymbol{x})}[f(\boldsymbol{\theta})]=\int f(\boldsymbol{\theta})\pi(\boldsymbol{\theta}|\boldsymbol{x})\mathrm{d}\boldsymbol{\theta} \tag{34}$$

$$\nabla_x E_{\pi(\boldsymbol{\theta}|\boldsymbol{x})}[f(\boldsymbol{\theta})]=E_{\pi(\boldsymbol{\theta}|\boldsymbol{x})}[f(\boldsymbol{\theta})\nabla_x\log(\pi(\boldsymbol{\theta}|\boldsymbol{x}))] \tag{35}$$

限制查询攻击在当前图像 $\boldsymbol{x}$ 周围选择随机高斯噪声的搜索分布，即 $\boldsymbol{\theta}=\boldsymbol{x}+\sigma\boldsymbol{\delta}$，其中 $\boldsymbol{\delta}\sim\mathcal{N}(0,I)$。它在估计梯度时使用 $n$ 个对偶采样[45]的$\boldsymbol{\delta}_i$，来提高 NES 的性能：

$$\nabla E[f(\boldsymbol{\theta})]\approx\frac{1}{\sigma n}\sum_{i=1}^{n}\boldsymbol{\delta}_i f(\boldsymbol{\theta}+\sigma\boldsymbol{\delta}_i) \tag{36}$$

最后，限制查询攻击基于 NES 梯度估计值使用有动量项的 PGD 来生成对抗样本。限制查询攻击的查询效率比 ZOO 高了 2～3 个数量级。

3）基于自动编码器的零阶优化方法

为了提高查询效率，基于自动编码器的零阶优化方法（Autoencoder-based Zeroth Order Optimizati-on Method，AutoZOOM）[46]使用随机无梯度（Random Gradient-Free，RGF）方法[47]进行$\nabla_x f(\boldsymbol{x})$的缩放随机全梯度估计：

$$\boldsymbol{g}\approx b\cdot\frac{f(\boldsymbol{x}+\beta\boldsymbol{u})-f(\boldsymbol{x})}{\beta}\cdot\boldsymbol{u} \tag{37}$$

其中，$\beta>0$ 是平滑参数；$\boldsymbol{u}$ 是从单位欧几里得球体均匀随机采样的单位长度向量；$b$ 是可调整的缩放参数，用于权衡梯度估计误差的偏差和方差。AutoZOOM 使用平均随机梯度估计，以有效地控制梯度估计中的误差。其中，梯度估计值是在 $q$ 个随机方向$\{\boldsymbol{u}_j\}_{j=1}^{q}$上的平均值：

$$\bar{\boldsymbol{g}}=\frac{1}{q}\sum_{j=1}^{q}\boldsymbol{g}_j \tag{38}$$

其中，$\boldsymbol{g}_j$是式(37)中定义的梯度估计，$\boldsymbol{u}=\boldsymbol{u}_j$。

AutoZOOM 使用自适应策略来选择 $q$，以平衡生成对抗样本时的查询次数和扰动大小。具体来说，它先使用 $q=1$（尽可能少的模型查询），以粗略的梯度估计来快速地完成初始攻击；随后使用 $q>1$，以更精确的梯度估计来微调图像质量，减小扰动。

为了提高查询效率和加速算法收敛，AutoZOOM 从更小的维度$n'<n$ 进行随机梯度估计。添加到 $\boldsymbol{x}$ 上的对抗扰动实际上是从降维空间生成的，然后利用解码器从低维空间表示中重构出高维对抗扰动。通常选择的解码器为在不同于训练数据的未标记数据上训练的自动编码器（autoencoder，AE），或者双线性图像大小调整器（bilinear image resizer，BiLIN）。

与 ZOO 相比，AutoZOOM 在保持相似攻击成功率的同时，减少了平均查询的次数。由于限制查询攻击没有考虑降维的影响，因此与限制查询攻击相比，AutoZOOM 进一步提高了查询效率。除此之外，AutoZOOM 还可以有效地微调图像扰动，以保持与干净图像较高的视觉相似性。

与基于迁移的攻击相比，基于分数的攻击通过查询目标模型进行梯度估计，不需要训练替代模型，解决了针对替代模型生成的对抗样本的迁移性较差的问题，从而提高了针对目标模型的黑盒攻击成功率。但其需要较多的迭代查询次数用于估计目标模型的梯度，在攻击过程中引入了额外的时间。

（3）基于决策的攻击

在此黑盒设置下，攻击者无法获得预测类别概率，相反，只能通过查询获得相应的最终决策，即 top-1 预测类别。基于决策的攻击仅依赖于目标模型的 top-1 预测类别进行直接攻击。下面将概述几种基于决策的攻击方法。

1）边界攻击

边界攻击[48]需要使用对抗样本（例如目标攻击中属于目标类别的样本）来进行初始化，然后沿着对抗区域和非对抗区域之间的决策边界执行随机游走，以在每次迭代时保持更新后的样本仍停留在对抗区域中，并且逐渐接近原始样本。边界攻击的核心就是使用合适的候选分布进行拒绝采样，并结合步长动态调整策略在决策边界随机游走，最后根据选择的对抗标准找到逐渐减小的对抗扰动。

边界攻击的基本运行原理（即从较大的扰动开始，逐渐减小）非常简单，颠覆了以前所有对抗攻击的基本逻辑。在最小化扰动的大小方面，边界攻击与白盒攻击方法相当。

2）仅标签攻击

仅标签攻击[43]的关键思想是，在没有分数输出的情况下，找到另一种方法来表征对抗攻击的成功。它提出了对抗样本的离散分数 $R(\boldsymbol{x}^{(t)})$，仅基于目标标签$l'$是否为 top-1 预测类别，就可以量化每次迭代时的图像对抗性：

$$R(\boldsymbol{x}^{(t)})=1-rank(l'\mid\boldsymbol{x}^{(t)}) \tag{39}$$

其中，若目标标签$l'$为$\boldsymbol{x}^{(t)}$的 top-1 预测类别，则函数 $rank(l'\mid\boldsymbol{x}^{(t)})$为 0，反之为 1。仅标签攻击使用替代了 $softmax$ 概率的离散分数来量化对抗性，并使用蒙特卡洛近似估计该分数：

$$S(\boldsymbol{x}^{(t)})=E_{\boldsymbol{\delta}\sim\mathcal{N}[-\mu,\mu]}[R(\boldsymbol{x}^{(t)}+\boldsymbol{\delta})] \tag{40}$$

$$\hat{S}(\boldsymbol{x}^{(t)})=\frac{1}{n}\sum_{i=1}^{n}R(\boldsymbol{x}^{(t)}+\mu\boldsymbol{\delta}_i) \tag{41}$$

其中，$\hat{S}(\boldsymbol{x}^{(t)})$被视为输出概率$F(\boldsymbol{x})_{l'}$的代替品。

仅标签攻击首先使用目标标签的图像初始化，之后在每次迭代时使用回溯线性搜索来找到可以将目标标签保持在 top-1 预测类别的$\varepsilon_t$，最后利用 NES 估计的梯度值$\nabla_x\hat{S}(\boldsymbol{x}^{(t)})$进行 PGD 攻击。

虽然边界攻击和仅标签攻击都可以找到与白盒攻击具有类似扰动的对抗样本，但它们都需要大量查询来探索高维空间，并且无法保证其收敛性。

3）基于优化的攻击

基于优化的攻击（Opt-attack）[49]提供了一种新的思路——将基于决策的攻击重新形式化为一个实值优化问题，其可以使基于决策攻击的查询效率更高。Opt-attack 首先根据攻击的类型定义目标函数 $f$：

非目标攻击的目标函数如下：

$$f(\boldsymbol{\theta})=\min_{\lambda>0}\lambda\quad\text{s.t. }C\left(\boldsymbol{x}+\lambda\frac{\boldsymbol{\theta}}{\|\boldsymbol{\theta}\|_2}\right)\neq l \tag{42}$$

目标攻击的目标函数如下：

$$f(\boldsymbol{\theta})=\min_{\lambda>0}\lambda\quad\text{s.t. }C\left(\boldsymbol{x}+\lambda\frac{\boldsymbol{\theta}}{\|\boldsymbol{\theta}\|_2}\right)=l' \tag{43}$$

其中，$\boldsymbol{\theta}$ 表示搜索方向；$f(\boldsymbol{\theta})$表示 $\boldsymbol{x}$ 沿 $\boldsymbol{\theta}$ 方向到最近对抗样本的距离，将输入方向映射到连续的实值输出。非目标攻击中 $f(\boldsymbol{\theta})$也表示 $\boldsymbol{x}$ 沿 $\boldsymbol{\theta}$ 方向到决策边界的距离。

Opt-attack 没有搜索对抗样本，而是搜索方向 $\boldsymbol{\theta}$ 来最小化扰动 $f(\boldsymbol{\theta})$，这就得到了以下优化问题：

$$\min_{\boldsymbol{\theta}}f(\boldsymbol{\theta}) \tag{44}$$

最后，通过$\boldsymbol{x}'=\boldsymbol{x}+f(\boldsymbol{\theta}^*)\frac{\boldsymbol{\theta}^*}{\|\boldsymbol{\theta}^*\|}$可以生成对抗样本，其中 $\boldsymbol{\theta}^*$ 是优化问题(44)的最优解。从图 2[49]中可以看出，$\boldsymbol{\theta}$ 发生小的改变通常会导致 $f(\boldsymbol{\theta})$发生较大的改变。

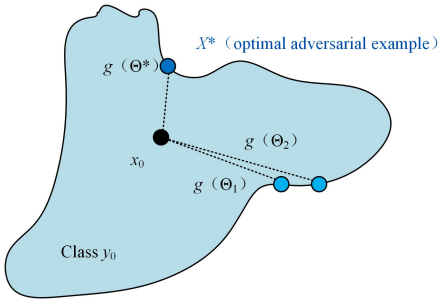



图 2 Opt-attack 的示意图[49]

Fig. 2 Illustration of Opt-attack[49]

对于归一化后的 $\boldsymbol{\theta}$，Opt-attack 通过查询目标模型的 top-1 预测类别，使用粗粒度搜索和二分搜索来计算 $f(\boldsymbol{\theta})$ 函数值。为了优化问题(44)，Opt-attack 使用 RGF 方法来获得梯度估计值 $\hat{\boldsymbol{g}}$，然后用步长 $\alpha$ 通过 $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} - \alpha \hat{\boldsymbol{g}}$ 进行更新。

当 $f(\boldsymbol{\theta})$ 是 Lipschitz 平滑时，Opt-attack 可以显著减少查询次数，并且保证收敛性。此外，Opt-attack 还可以攻击除 DNN 以外的其他离散、不连续、不可微的机器学习模型，例如梯度增强决策树等。

与基于分数的攻击相比，基于决策的攻击在现实世界应用中更为适用，现实世界中的模型不会给用户输出预测类别概率。但是，由于没有预测类别概率的信息，基于决策的攻击需要对目标模型进行额外的查询；与基于迁移的攻击相比，基于决策的攻击不需要训练替代模型，对目标模型的攻击更加直接。除此之外，与其他两类的黑盒攻击相比，基于决策的攻击对于对抗防御方法也更加鲁棒。

## 3.2 对抗防御

本节将概述图像领域中的几种对抗防御方法，包括一种反应式防御策略(对抗样本检测)和两种主动式防御策略(对抗训练以及去噪)。

### 3.2.1 对抗样本检测

对抗样本检测是防御对抗攻击的一种主流方法，这种方法不是直接预测模型的输入，而是先区分输入是干净样本还是对抗样本。如果检测到输入是对抗样本，深度模型将不会预测其类别。对抗样本检测的策略可分为 3 类：样本统计、训练检测器和预测不一致性。

(1)样本统计

使用最大平均差异的统计测试[50]可以被用来检测对抗样本，它将能量距离作为统计距离来度量。该方法需要大量的对抗样本和干净样本，并且无法检测出个别的对抗样本。除此之外，核密度估计[51]也被提出用于检测对抗扰动，它使用来自 DNN 的一些中间层的表示来衡量未知样本和一组干净样本之间的距离。该方法的计算成本很高，并且只能检测出远离干净样本分布的对抗样本。由于对抗样本在本质上是不可感知的，因此使用样本统计将对抗样本与干净样本分开似乎不太可能有效。与其他对抗样本检测方法相比，依赖样本统计策略的检测性能较差。

(2)训练检测器

与对抗训练类似，对抗样本可以被用来训练检测器。然而，这种策略需要大量的对抗样本，代价较高并且容易发生过拟合，下面将介绍两种代表性的方法。一种方法是在原始 DNN 的中间层附加一个基于 CNN 的分支检测器[52]。检测器输出两个类别，并使用对抗样本和干净样本进行训练。检测器在训练的同时固定了原始 DNN 的权重，因此不会降低干净样本的分类精度。另一种方法在 DNN 的最后一层增加了一个新的“对抗”类别[50]。修改后的模型使用干净样本和对抗样本进行训练，由于模型体系结构的改变，该方法会降低干净样本的准确性。

(3)预测不一致性

预测不一致性的基本思想是在预测未知样本时测量多个模型输出的不一致性，因为一个对抗样本可能不会欺骗所有的 DNN。贝叶斯神经网络不确定性[51]借鉴了 Dropout 的思想，它使用了 Dropout 层的“训练”模式来为测试时的输入样本生成许多预测。该方法可行的原因是子模型的预测之间产生分歧在干净样本上很少见，而在对抗样本上很常见。除此之外，还可以使用输入样本本身来检查一致性[53]。对于每个输入样本，减小它们的图像的颜色深度也可以检测对抗样本。例如，每个像素具有 256 个可能值的 8 位灰度图像变为每个像素具有 128 个可能值的 7 位灰度图像。对于自然图像，减小颜色深度不会改变预测结果，但是对于对抗样本，预测结果会发生变化。

### 3.2.2 对抗训练

对抗训练是一种针对对抗样本的直观防御方法，通过对 DNN 进行对抗训练来提高 DNN 的鲁棒性，将已知的对抗样本和相应的真实标签引入训练中。理想情况下，模型将学习如何从对抗扰动中还原真实标签，并对未知的对抗样本表现鲁棒。

Goodfellow 等[20]首先在训练阶段加入了对抗样本，在训练的每个步骤中使用 FGSM 生成了对抗样本，并将其注入训练集中。实验结果表明，经过训练的模型对 FGSM 生成的对抗样本更加鲁棒，但是仍然容易受到迭代/基于优化的对抗攻击。因此，许多工作进一步研究了使用更强对抗攻击(例如 BIM/PGD 攻击)的对抗训练。

广泛的实验评估表明，PGD 攻击可能是通用的一阶 $L_\infty$ 攻击。如果是这样，则提高针对 PGD 攻击的模型鲁棒性意味着可以抵抗多种一阶 $L_\infty$ 攻击。基于这种推测，PGD 对抗训练[24]被提出用于训练鲁棒的 DNN。实验结果表明，基于 PGD 的对抗训练显著提高了 DNN 在 MNIST 数据集上针对多种不同攻击的鲁棒性。然而，其在 CIFAR-10 数据集上的结果较差。

随机初始化的 FGSM[21]也被用来在 ImageNet 数据集上进行对抗训练，以避免 PGD 对抗训练带来的巨大计算开销。然而，经过对抗训练的模型容易受到基于迁移的攻击。为了解决这个问题，集成对抗训练[22]使用了从多个预训练模型中迁移过来的对抗样本。直观上，集成对抗训练增加了用于对抗训练的对抗样本的多样性，从而增强了对于从其他模型迁移过来的对抗样本的鲁棒性。实验结果表明，集成对抗训练模型对于在其他模型上由单步/迭代攻击生成的对抗样本，仍

具有较高的鲁棒性。在某些情况下，集成对抗训练对黑盒攻击的表现优于 PGD 对抗训练。

3.2.3　去噪

就减轻对抗扰动而言，去噪是一种非常直接的方法。以往的研究指出了两个设计方向：输入去噪和特征去噪。第一个方向试图从输入中部分或完全消除对抗扰动，第二个方向试图减轻对抗扰动对 DNN 学习到的高级特征的影响。

第一个方向的代表性工作是一种被称为 MagNet 的防御系统[54]，其包括一个检测器和一个重构器。MagNet 中使用自动编码器来学习干净样本的多种流形。检测器根据样本与学习的流形之间的关系来区分干净样本和对抗样本，重构器被设计用于将对抗样本修正为干净样本。然而，MagNet 被证明容易受到 C&W 攻击生成的迁移对抗样本的攻击。

第二个方向的代表性工作是一种高级表示引导的去噪器（High-level Representation Guided Denoiser，HGD）[55]，可以完善受对抗扰动影响的特征。HGD 没有使用像素级去噪，而是使用特征级损失函数来训练去噪 u-net[56]，以最小化干净样本和对抗样本之间的特征级差异。尽管 HGD 在黑盒设置下有效，但在白盒设置下却可以受到 PGD 的攻击[57]。为了修正 DNN 中间层学习到的特征，DNN 被设计配备了使用多种去噪操作的模块[58]。尽管实验取得了巨大的成功，但特征去噪模块对模型鲁棒性的贡献还是不能和 PGD 对抗训练相比。

## 4　视频领域对抗攻防

### 4.1　对抗攻击

本节将从白盒攻击和黑盒攻击两个角度出发，介绍视频领域的对抗攻击方法。由于视频具有额外的时间维度，因此目前针对视频领域对抗攻击的研究主要关注如何在时序上攻击视频模型，以及如何降低攻击的空间维度。

4.1.1　白盒攻击

（1）稀疏对抗攻击方法

稀疏对抗攻击方法[9]基于帧间扰动的传播性生成针对视频识别模型的对抗样本。帧间扰动传播性指在当前帧上增加的对抗扰动可以通过模型的时序关联性传递到其他帧上，因此只需要在视频的部分帧上增加对抗扰动就可以使模型产生错误分类。稀疏对抗攻击方法利用基于$L_{2,1}$范数的优化算法来针对某一视频类别生成通用的视频对抗扰动：

$$\arg\min_{\eta}\lambda\|\boldsymbol{M}\cdot\eta\|_{2,1}-\frac{1}{N}\sum_{i=1}^{N}loss(\boldsymbol{x}_i+\boldsymbol{M}\cdot\eta,l_i) \tag{45}$$

其中，$\lambda$ 是平衡因子；$\boldsymbol{M}$ 表示时序掩码，用于控制仅在部分帧中增加对抗扰动；$\|\eta\|_{2,1}=\sum_{t}^{T}\|\eta_t\|_2$表示 $\eta$ 的$L_{2,1}$范数，用于衡量对抗扰动的大小，$L_{2,1}$范数在帧间应用$L_1$范数，以确保生成扰动的稀疏性；$N$ 是视频总数。

稀疏对抗攻击方法对 CNN＋RNN 视频模型进行攻击，使用 Adam 优化器来优化式(45)，通过指定时序掩码 $\boldsymbol{M}$，以仅在 $\boldsymbol{M}$ 指定区域内生成稀疏对抗扰动 $\eta$。结果表明，对抗扰动在 CNN＋RNN 结构的视频模型中具有传播性。此外，该方法生成的对抗扰动在具有不同 RNN 结构的模型以及视频之间具有良好的迁移性。

（2）针对实时视频分类系统的攻击方法

针对实时视频分类系统的攻击方法[59]利用生成对抗网络[60]（Generative Adversarial Network，GAN）来针对实时视频分类系统生成对抗扰动。GAN 在离线情况下生成的通用对抗扰动，会在视频分类系统工作时与干净视频相融合，最终导致视频分类系统决策错误。该方法通过反向传播来优化 GAN，其目标函数如下：

$$\min_{G}\sum_{i=1}^{N}-\log[1-F(\boldsymbol{x}_i+G(\boldsymbol{s}))\cdot\mathbf{1}_{l_i}] \tag{46}$$

其中，$G$ 表示 GAN，$\boldsymbol{s}$ 表示随机噪声，$G(\boldsymbol{s})$表示 GAN 生成的对抗扰动。通过优化目标函数(46)来对 GAN 的参数进行更新，逐步降低真实标签的概率值，导致目标模型分类错误。

该方法对视频片段中的时序结构进行了多角度分析，以人类难以察觉的对抗扰动，成功实现了针对实时视频分类模型的对抗攻击。

（3）闪烁对抗攻击方法

闪烁对抗攻击方法[61]通过对视频中每帧增加统一的 RGB 扰动，来模拟现实物理场景下光线等条件的变化，构造一个时序性的对抗模式，实现对视频分类模型的对抗攻击。当连续帧之间增加的统一 RGB 扰动相差较大时，人眼会感受到闪烁，因此闪烁对抗攻击方法在降低对抗扰动大小的同时，也会限制连续帧之间扰动变化的大小和幅度，从而提升对抗扰动的不可察觉性。该方法的目标函数为：

$$\arg\min_{\eta}\lambda\sum_{j}\beta_jD_j(\eta)+\frac{1}{N}\sum_{i=1}^{N}loss(\boldsymbol{x}_i+\eta,l_i) \tag{47}$$

其中，函数$D_j(\cdot)$表示正则化函数；$\beta_j$表示每个正则化函数的权重值；$\lambda$ 平衡了对抗性和正则化项的相对重要性。正则化项分为厚度正则化（Thickness Regularization）和粗糙正则化（Roughness Regularization），厚度正则化用来控制每帧的统一 RGB 对抗扰动的大小，其表达式为：

$$D_1(\eta)=\frac{1}{3T}\|\eta\|_2^2 \tag{48}$$

由于每一帧在 RGB 空间中分别有一个固定的扰动值，因此需要除以 $3T$。粗糙正则化用于控制连续帧之间扰动变化的大小和幅度，其表达式为：

$$D_2(\eta)=\frac{1}{3T}\left\|\frac{\partial\eta}{\partial t}\right\|_2^2+\left\|\frac{\partial^2\eta}{\partial^2 t}\right\|_2^2 \tag{49}$$

其中，第一项用来控制两个连续帧之间扰动差异的大小；第二项用来控制对抗扰动的趋势。

该方法通过正则化项约束了连续帧之间统一 RGB 扰动值的变化程度，实现了微小且平滑的对抗扰动，在满足人类观察者难以察觉的前提下，成功地对视频识别模型进行了对抗攻击。

（4）附加对抗帧方法

附加对抗帧方法[62]取代了直接在视频上增加对抗扰动的方法，而是在视频末尾增加额外帧（例如包含“感谢观看”文字的结尾帧），并在额外帧上增加对抗扰动来实现针对视频模型的对抗攻击，这种方法叫作附加对抗帧方法（Appending Adversarial Frames Method，$A^2FM$）。该方法通过对添加在额外帧上的对抗扰动进行优化来实现对抗攻击：

$$\arg\min_{\eta} \lambda \|\eta\|_p - loss(\hat{\boldsymbol{x}}, \mathbf{1}_l) \tag{50}$$

其中，$\hat{\boldsymbol{x}}$表示与额外对抗帧拼接起来的视频。此外，该方法还针对视频及视频模型生成了通用的额外对抗帧。

该方法充分利用视频的时间维度，通过增加额外对抗帧的方式来对视频模型发起攻击。与在视频帧内增加对抗扰动相比，在视频帧末尾增加对抗帧的操作更易实现。

4.1.2 黑盒攻击

(1)图像指导的黑盒攻击方法

图像指导的黑盒攻击方法[63]首次在黑盒设置下提出了基于图像模型扰动迁移的视频攻击方法(V-BAD)。V-BAD将视频中的每一帧输入到图像模型中，得到相应的特征图，并使当前视频对抗帧特征图逐渐逼近目标类别视频帧特征图，最终获得从图像模型迁移而来的初始扰动方向。V-BAD没有直接使用NES估计扰动方向的方法，而是在得到初始扰动方向的基础上，使用NES估计初始扰动方向的修正矩阵。此外，为了提升NES的效率，V-BAD将视频帧在空间上均等划分为多个块，其中同一块共享相同的修正方向。最后，V-BAD使用修正后的扰动方向来对视频迭代使用PGD。

V-BAD方法填补了视频黑盒攻击的空白，利用图像模型向视频提供了初始的扰动方向，提升了攻击效率，并通过均等划分块方式进一步提升了梯度估计的效率。

(2)启发式黑盒攻击方法

启发式黑盒攻击方法[64]基于启发式规则评估了视频各个帧以及帧中不同区域的重要性，在黑盒条件下生成了具有时空稀疏性的对抗样本。在空间维度上，启发式黑盒攻击方法使用显著区域检测算法，仅在显著区域上增加对抗扰动。在时间维度上，启发式黑盒攻击方法从目标类别的视频样本出发，迭代删除某一帧的对抗扰动，通过此时视频模型的输出结果对该帧进行打分。在满足对抗性的前提下，删除该帧后的模型输出概率值越高，该帧的得分就越低。之后，启发式黑盒攻击方法在满足对抗性的前提下，根据各帧的得分，由低到高逐渐删除该帧的扰动。这样就可以得到具备时空稀疏性的对抗扰动。最终，启发式黑盒攻击方法利用Opt-attack完成对稀疏对抗扰动的优化。

该方法在黑盒条件下，探索了对抗扰动的稀疏性，并证明了稀疏性在不影响成功率的同时也可以提升攻击效率。

(3)基于强化学习的稀疏黑盒攻击方法

基于强化学习的稀疏黑盒攻击方法[65]将黑盒攻击方法与强化学习相结合，通过视频模型的反馈来逐步调整关键帧的位置。在该方法的强化学习中，使用顶部为全连接层的双向LSTM作为智能体(agent)，智能体同时起到关键帧选择和视频攻击的作用。此外，该方法将使用视频模型作为强化学习中的环境(environment)，将视频模型返回的预测概率值以及视频关键帧之间的内容差异作为奖赏(reward)，将智能体调整帧选择策略以及执行攻击作为动作(action)。智能体将在可以提供奖励和更新其行为的环境中进行交互，通过最大化整体期望来选择关键帧。在特征选择阶段，智能体利用LSTM得到每帧的概率值进行关键帧选择。由于每帧对应于两种动作，因此对于$T$帧的视频共有$2^T$种选择。该方法设计了两种奖赏来反映每种动作的质量，来自视频关键帧本身及目标模型的输出概率值。最后，该方法使用REINFORCE[66]算法最大化奖赏的期望。此外，该方法使用类似于V-BAD的梯度估计方法来进行梯度估计。

该方法利用强化学习策略进行关键帧选择，并设计了一套有效的奖励机制，从视频本身以及视频模型两个角度来评估动作行为，有效提升了攻击效率。

(4)动作激发采样攻击方法

动作激发采样攻击方法[67]在梯度估计阶段是一种有效的动作激发噪声采样策略，可以在较少的查询数目下完成对视频模型的对抗攻击。在黑盒设置下，往往需要在当前视频上增加随机噪声，再根据视频模型的输出改变情况进行梯度估计，最终利用估计梯度对当前视频样本进行更新。而在该方法中，将间隔选取视频帧作为视频模型的输入数据，并使用间隔帧的累积光流信息或动作向量来对输入数据中的每帧进行分类。由于累积光流信息及动作向量在时间维度上体现了不同像素的移动信息，该方法利用移动信息对不同的像素进行分组，并对同一组内的像素增加相同的噪声。最终，该方法通过有限差分的方式来估计梯度信息，并对输入视频进行更新。

该方法通过间隔帧中的累积光流或动作信息在时间和空间维度上对像素进行分组，相比V-BAD，攻击效率得到了进一步的提升。

### 4.2 对抗防御

本节将概述视频领域中的对抗防御方法。由于对抗扰动在视频的不同帧中不具有一致性，因此目前的对抗防御工作主要关注于如何利用时间不一致性来实现对抗帧检测。

(1)对抗帧鉴别器

对抗帧鉴别器[68]针对多种类型的攻击方法，利用视频的时间一致性属性来实现视频中的对抗帧的有效识别(AdvIT)。在测试过程中，AdvIT首先计算之前的视频帧与当前视频帧的光流信息，再利用之前的视频帧与所得到的光流信息重建当前视频帧。之后将重建的视频帧与当前视频帧放入视频模型中分别得到预测结果。当预测结果相差不大时，说明可以满足时间一致性，而当预测结果相差较大时，当前帧会被识别为对抗帧。

AdvIT方法基于视频的连续性，利用光流信息对当前视频帧进行重建，通过重建帧与当前帧在视频模型的输出差异来对帧进行对抗帧检测。此外，该方法对于之前视频帧是否为对抗帧不敏感，但是该方法仅限于对抗帧检测，没有提出针对对抗帧的防御方法。

(2)基于时间一致性的对抗视频抵御方法

基于时间一致性的对抗视频抵御方法[69]分为两个阶段。第一阶段，利用视频的时间一致性进行对抗视频检测；第二阶段，在时间和空间两个维度上抑制对抗扰动。与AdvIT类似，该方法基于时间一致性进行对抗帧检测，并根据对抗帧在视频中的占比将该攻击类型分为稠密攻击和稀疏攻击。针对对抗视频，该方法提出了空间防御和时间防御两种防御措施。在空间防御中，利用端到端的压缩模型(图像压缩和图像重建

模块）针对对抗扰动进行去噪，减轻对抗扰动对视频模型的干扰。在时间防御中，以第一阶段的对抗帧检测结果为基础，利用干净帧重建对抗帧。该方法使用空间防御来抵御稠密攻击，使用时间防御来抵御稀疏攻击。

该方法首次探究了对抗视频的防御，针对现有攻击的特点，在确定攻击方法类型后选择相应的防御方法。该方法在提升视频模型性能的同时，可以有效抵御视频攻击方法。

(3)批标准化防御方法

批标准化防御方法[70]利用对抗视频检测器识别不同的攻击类型，并将对抗视频送入相对应的批标准化层（Batch Normalizations，BNs）进行处理，最终实现对抗视频的正确分类，提升视频模型的鲁棒性。该方法通过对抗训练来提升针对$L_p$范数及物理攻击的防御性能，其目标函数为：

$$\arg\min_{\boldsymbol{\theta}} E_{(\boldsymbol{x},l)\sim\mathbb{D}}\left[loss(\boldsymbol{x},l;\boldsymbol{\theta}^{c},\boldsymbol{\theta}_0^{b})+\sum_{i=1}^{N}\max_{\eta_i\in\mathbb{S}_i} loss(\boldsymbol{x}+\eta_i,l;\boldsymbol{\theta}^{c},\boldsymbol{\theta}_i^{b})\right] \tag{51}$$

其中，$\mathbb{D}$为训练集；$\boldsymbol{\theta}^{c}$为视频模型卷积参数；$\boldsymbol{\theta}_i^{b}$为第$i$种数据类型的BNs参数；$\boldsymbol{\theta}=\boldsymbol{\theta}^{c}+\sum_{i=1}^{N}\boldsymbol{\theta}_i^{b}$表示所有模型的参数；$N$表示除干净样本外的其他不同攻击类型的数目。这里，由于多种攻击类型数据之间的分布不同，BNs在训练中容易受到分布不一致的影响，因此，该方法为包含干净样本在内的数据类型分别提供了一个独立BNs分支来进行对抗训练。此外，为了输入样本可以在推断阶段进入指定的BNs分支，该方法应用端到端训练的3D CNN结构对数据类型进行分类。

与其他防御方法相比，该方法利用BNs分支实现了多种攻击类型的共同防御机制，实验结果也表明该方法在同种攻击类型、不同攻击方法下同样具备良好的防御性能。

## 5　未来研究方向

本节将针对图像、视频领域的对抗攻防方法，探讨其在未来的研究发展方向。

### 5.1　图像领域

近年来，图像领域的对抗攻防研究得到了迅速发展。目前的攻击方法很可能会被新的防御方法抵御，同样，这些防御方法也会被新的攻击方法规避，未来的对抗攻防方法会此消彼长，互相促进。现有的对抗攻防工作主要集中在图像单一模态任务中，其危害性、广泛性有限。未来需要从多模态联合的角度出发，研究基于多模态数据的高维、冗余、互补等特点的对抗攻防技术。此外，未来的对抗攻防会朝向特定任务（例如，人脸识别、目标检测及语义分割等）发展，以适应实际需求。

### 5.2　视频领域

与图像领域的研究相比，视频领域的对抗攻防研究仍旧较少。现有的面向视频领域的对抗攻防方法中，基于视频具有时间维度这一特点的研究方法较多。针对白盒攻击，现主要利用对抗扰动在帧间的传递性来降低对抗扰动的面积或大小，从而提升对抗样本的质量。未来需要从应对已有的防御策略出发，研究基于时间一致性的对抗扰动生成方法；从应用角度出发，研究基于视频水印的对抗性水印生成方法。针对黑盒攻击，目前主要利用视频的时间一致性来降低攻击维度，提升对抗样本的生成效率，但与图像相比，其生成效率仍旧较低。未来需要进一步提升攻击效率，研究基于帧间光流信息的对抗扰动传播方法，使得在单个帧上生成的对抗扰动可以依据帧间关系向其他帧传播。针对对抗防御，由于视频帧的连续性，目前主要利用视频的时间一致性来检测对抗帧。未来需要研究在攻击者知晓防御策略的条件下，仍能针对多种攻击方法进行防御的方法，提升视频模型的鲁棒性。

**结束语**　在推动人类社会进步的同时，快速发展的深度学习技术也带来了新的安全隐患——对抗样本。这将在多媒体视觉领域的应用过程（互联网视频内容监管、视频监控、自动驾驶以及医疗影像等）中带来严重危害，危及社会安全。为了应对深度学习带来的安全隐患，推动深度学习安全的研究，本文介绍了深度学习在多媒体视觉领域中的多种对抗攻防方法，从图像和视频两个角度表述了对抗攻防的研究现状及未来的研究方向。对抗攻防技术仍然存在一些挑战，如对抗训练成本高、攻防方法通用性低等。但是，随着对抗攻防的发展，这些问题与挑战终将被克服，对抗攻防定会为深度学习的应用夯实基础。

## 参 考 文 献


[1] SUN Y，WANG X，TANG X. Deep learning face representation from predicting 10 000 classes[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2014：1891-1898.

[2] TAIGMAN Y，YANG M，RANZATO M A，et al. Deepface：Closing the gap to human-level performance in face verification[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2014：1701-1708.

[3] REDMON J，FARHADI A. YOLO9000：better，faster，stronger[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017：7263-7271.

[4] REN S，HE K，GIRSHICK R，et al. Faster r-cnn：Towards real-time object detection with region proposal networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence，2016，39(6)：1137-1149.

[5] SAON G，KUO H K J，RENNIE S，et al. The IBM 2015 English Conversational Telephone Speech Recognition System[C]// Sixteenth Annual Conference of the International Speech Communication Association. 2015：3140-3144.

[6] SILVER D，SCHRITTWIESER J，SIMONYAN K，et al. Mastering the game of go without human knowledge[J]. Nature，2017，550(7676)：354-359.

[7] SZEGEDY C，ZAREMBA W，SUTSKEVER I，et al. Intriguing properties of neural networks[J]. arXiv：1312.6199，2013.

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

[9] WEI X，ZHU J，YUAN S，et al. Sparse adversarial perturbations for videos[C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2019，33：8973-8980.

[10] MISHKIN D, MATAS J. All you need is a good init[J]. arXiv: 1511.06422, 2015.

[11] MAAS A L, HANNUN A Y, NG A Y. Rectifier nonlinearities improve neural network acoustic models[C]// Proc. ICML. 2013, 30(1): 3.

[12] CLEVERT D A, UNTERTHINER T, HOCHREITER S. Fast and accurate deep network learning by exponential linear units (elus)[J]. arXiv: 1511.07289, 2015.

[13] 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. 2016: 1528-1540.

[14] 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.

[15] KRIZHEVSKY A, HINTON G. Learning multiple layers of features from tiny images[J]. Handbook of Systemic Autoimmune Diseases, 2009, 1(4).

[16] RUSSAKOVSKY O, DENG J, SU H, et al. Imagenet large scale visual recognition challenge[J]. International Journal of Computer Vision, 2015, 115(3): 211-252.

[17] SOOMRO K, ZAMIR A R, SHAH M. UCF101: A dataset of 101 human actions classes from videos in the wild[J]. arXiv: 1212.0402, 2012.

[18] KUEHNE H, JHUANG H, GARROTE E, et al. HMDB: a large video database for human motion recognition[C]// 2011 International Conference on Computer Vision. IEEE, 2011: 2556-2563.

[19] KAY W, CARREIRA J, SIMONYAN K, et al. The kinetics human action video dataset[J]. arXiv: 1705.06950, 2017.

[20] GOODFELLOW I J, SHLENS J, SZEGEDY C. Explaining and harnessing adversarial examples[J]. Stat, 2015, 1050: 20.

[21] KURAKIN A, GOODFELLOW I, BENGIO S. Adversarial machine learning at scale[J]. arXiv: 1611.01236, 2016.

[22] TRAMER F, KURAKIN A, PAPERNOT N, et al. Ensemble adversarial training: attacks and defenses[J]. Stat, 2018, 1050: 22.

[23] KURAKIN A, GOODFELLOW I J, BENGIO S. Adversarial examples in the physical world[J]. arXiv: 1607.02533, 2016.

[24] MDRY A, MAKELOV A, SCHMIDT L, et al. Towards Deep Learning Models Resistant to Adversarial Attacks[J]. Stat, 2017, 1050: 9.

[25] PAPERNOT N, MCDANIEL P, JHA S, et al. The limitations of deep learning in adversarial settings[C]// 2016 IEEE European Symposium on Security and Privacy (EuroS&P). IEEE, 2016: 372-387.

[26] MOOSAVI-DEZFOOLI S M, FAWZI A, FROSSARD P. Deepfool: a simple and accurate method to fool deep neural networks[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016: 2574-2582.

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

[28] CROCE F, HEIN M. Minimally distorted adversarial examples with a fast adaptive boundary attack[C]// International Conference on Machine Learning. PMLR, 2020: 2196-2205.

[29] CROCE F, HEIN M. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks[C]// International Conference on Machine Learning. PMLR, 2020: 2206-2216.

[30] ANDRIUSHCHENKO M, CROCE F, FLAMMARION N, et al. Square attack: a query-efficient black-box adversarial attack via random search[C]// European Conference on Computer Vision. Springer, Cham, 2020: 484-501.

[31] LIU Y, CHEN X, LIU C, et al. Delving into transferable adversarial examples and black-box attacks[J]. arXiv: 1611.02770, 2016.

[32] DONG Y, LIAO F, PANG T, et al. Boosting adversarial attacks with momentum[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 9185-9193.

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

[34] LIN J, SONG C, HE K, et al. Nesterov accelerated gradient and scale invariance for adversarial attacks[J]. arXiv: 1908.06281, 2019.

[35] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. Imagenet classification with deep convolutional neural networks[J]. Communications of the ACM, 2017, 60(6): 84-90.

[36] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[J]. arXiv: 1409.1556, 2014.

[37] HE K, ZHANG X, REN S, et al. Identity mappings in deep residual networks[C]// European Conference on Computer Vision. Springer, Cham, 2016: 630-645.

[38] XIE C, ZHANG Z, ZHOU Y, et al. Improving transferability of adversarial examples with input diversity[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2019: 2730-2739.

[39] LI Y, BAI S, ZHOU Y, et al. Learning transferable adversarial examples via ghost networks[C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2020: 11458-11465.

[40] 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.

[41] LAX P D, TERRELL M S. Calculus with applications[M]. New York: Springer, 2014.

[42] KINGMA D P, BA J. Adam: A method for stochastic optimization[J]. arXiv: 1412.6980, 2014.

[43] ILYAS A, ENGSTROM L, ATHALYE A, et al. Black-box Adversarial Attacks with Limited Queries and Information[C]// Proceedings of the 35th International Conference on Machine Learning (ICML 2018). 2018: 2137-2146.

[44] WIERSTRA D, SCHAUL T, PETERS J, et al. Natural evolu-

tion strategies[C]// 2008 IEEE Congress on Evolutionary Computation (IEEE World Congress on Computational Intelligence). IEEE,2008:3381-3387.

[45] SALIMANS T,HO J,CHEN X,et al. Evolution strategies as a scalable alternative to reinforcement learning[J]. arXiv:1703.03864,2017.

[46] TU C C,TING P,CHEN P Y,et al. Autozoom:Autoencoder-based zeroth order optimization method for attacking black-box neural networks[C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2019,33:742-749.

[47] NESTEROV Y,SPOKOINY V. Random gradient-free minimization of convex functions[J]. Foundations of Computational Mathematics,2017,17(2):527-566.

[48] BRENDEL W,RAUBER J,BETHGE M. Decision-based adversarial attacks:Reliable attacks against black-box machine learning models[J]. arXiv:1712.04248,2017.

[49] CHENG M,LE T,CHENP Y,et al. Query-efficient hard-label black-box attack: An optimization-based approach[J]. arXiv:1807.04457,2018.

[50] GROSSE K,MANOHARAN P,PAPERNOT N,et al. On the (statistical) detection of adversarial examples[J]. arXiv:1702.06280,2017.

[51] FEINMAN R,CURTIN R R,SHINTRE S,et al. Detecting adversarial samples from artifacts[J]. arXiv:1703.00410,2017.

[52] METZEN J H,GENEWEIN T,FISCHERV,et al. On detecting adversarial perturbations[J]. Stat,2017,1050:21.

[53] XU W,EVANS D,QI Y. Feature squeezing:Detecting adversarial examples in deep neural networks[J]. arXiv:1704.01155,2017.

[54] 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.

[55] LIAO F,LIANG M,DONG Y,et al. Defense against adversarial attacks using high-level representation guided denoiser[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018:1778-1787.

[56] RONNEBERGER O,FISCHER P,BROX T. U-net:Convolutional networks for biomedical image segmentation[C]// International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer,Cham,2015:234-241.

[57] ATHALYE A,CARLINI N. On the robustness of the cvpr 2018 white-box adversarial example defenses[J]. arXiv:1804.03286,2018.

[58] XIE C,WU Y,MAATEN L,et al. Feature denoising for improving adversarial robustness[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2019:501-509.

[59] LI S,NEUPANE A,PAUL S,et al. Adversarial perturbations against real-time video classification systems[J]. arXiv:1807.00458,2018.

[60] GOODFELLOW I,POUGET-ABADIE J,MIRZA M,et al. Generative adversarial nets[C]// Advances in Neural Information Processing Systems. 2014:2672-2680.

[61] NAEH I,PONY R,MANNOR S. Flickering Adversarial Attacks on Video Recognition Networks[J]. arXiv:2002.05123,2020.

[62] CHEN Z,XIE L,PANG S,et al. Appending adversarial frames for universal video attack[C]// Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. 2021:3199-3208.

[63] JIANG L,MA X,CHEN S,et al. Black-box adversarial attacks on video recognition models[C]// Proceedings of the 27th ACM International Conference on Multimedia. 2019:864-872.

[64] WEI Z,CHEN J,WEI X,et al. Heuristic black-box adversarial attacks on video recognition models[C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2020:12338-12345.

[65] YAN H,WEI X,LI B. Sparse black-box video attack with reinforcement learning[J]. arXiv:2001.03754,2020.

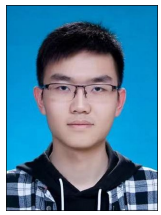
[66] WILLIAMS R J. Simple statistical gradient-following algorithms for connectionist reinforcement learning[J]. Machine Learning,1992,8(3/4):229-256.

[67] ZHANG H,ZHU L,ZHU Y,et al. Motion-Excited Sampler:Video Adversarial Attack with Sparked Prior[C]// European Conference on Computer Vision. Springer,Cham,2020:240-256.

[68] XIAO C,DENG R,LI B,et al. Advit:Adversarial frames identifier based on temporal consistency in videos[C]// Proceedings of the IEEE International Conference on Computer Vision. 2019:3968-3977.

[69] JIA X,WEI X,CAO X. Identifying and resisting adversarial videos using temporal consistency[J]. arXiv:1909.04837,2019.

[70] LO S Y,PATELV M. Defending against multiple and unforeseen adversarial videos[J]. arXiv:2009.05244,2020.



**CHEN Kai**, born in 1998, postgraduate. His main research interests include image and video adversarial attack.

**JIANG Yu-gang**, born in 1981, Ph. D, professor, Ph.D supervisor, is a member of China Computer Federation. His main research interests include multimedia content analysis, computer vision and robust & trustworthy AI.