计算机科学
COMPUTER SCIENCE

针对视频识别模型的边界黑盒对抗样本生成算法

荆瑜琳, 吴立军, 李志圆, 邓棋

引用本文

荆瑜琳, 吴立军, 李志圆, 邓棋. 针对视频识别模型的边界黑盒对抗样本生成算法[J]. 计算机科学, 2025, 52(10): 366-373.

JING Yulin, WU Lijun, LI Zhiyuan, DENG Qi. Boundary Black-box Adversarial Example Generation Algorithm on Video Recognition Models [J]. Computer Science, 2025, 52(10): 366-373.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.240700045

针对视频识别模型的边界黑盒对抗样本生成算法

荆瑜琳　吴立军　李志圆　邓　棋
电子科技大学计算机科学与工程学院　成都 611731
(jingyulin@std.uestc.edu.cn)

摘　要　随着深度学习的快速发展,神经网络在各个领域广泛应用。然而,当前神经网络仍然面临着对抗样本攻击的困扰。在所有类型的对抗样本攻击中,边界黑盒攻击只能获取被测试模型的最终分类标签,因此其最接近实际应用场景,被公认为最具有现实意义且最难实现的攻击,吸引了越来越多研究者的关注。但目前相关研究主要聚焦于图片识别模型,在视频识别模型方面的研究较少。为此,提出了一种基于边界的黑盒视频对抗样本生成算法 BBVA。BBVA 采用了一种渐进式探索机制生成视频对抗样本,有效提高了样本生成效率。实验表明,与最新的边界黑盒视频对抗样本生成算法 STDE 相比,BBVA 较好地权衡了噪声大小和模型访问次数,在视觉效果、优化距离和欺骗率等多项衡量指标中均达到了该研究领域目前最优水平;此外,在条件更为苛刻的情况下,BBVA 甚至优于一些最新的基于分数的黑盒视频对抗样本生成算法,如 EARL 和 VBAD。所提算法可用于提供对抗训练样本,从而提升视频模型的安全性。

关键词:对抗样本;视频识别;边界;黑盒;神经网络
中图分类号　TP391

Boundary Black-box Adversarial Example Generation Algorithm on Video Recognition Models

JING Yulin,WU Lijun,LI Zhiyuan and DENG Qi
School of Computer Science and Engineering,University of Electronic Science and Technology of China,Chengdu 611731,China

Abstract　With the rapid development of deep learning,neural networks are widely used in various fields. However,neural networks still face the problem of adversarial attacks. Among all types of adversarial attacks,the boundary black-box attack can only obtain the final classification label of the tested model,so it is closest to the actual application scenario,and is recognized as the most practical and difficult attacks,which has attracted more and more researchers to conduct related research. Nevertheless,current relevant research mainly focus on image recognition models,and with less research on video recognition models. To this end,this paper proposes a boundary black-box video adversarial example generation algorithm BBVA. BBVA uses a progressive exploration mechanism to generate adversarial videos,which effectively improves the efficiency of updating samples. Experiments show that compared with the state-of-the-art boundary black-box video adversarial example generation algorithm STDE,BBVA better balances the noise size and model queries,and gets the best results in this research field in many measurement indicators such as visual effect,optimization distance and fooling rate. In addition,under more severe conditions,BBVA even outperforms some state-of-the-art score-based black-box video adversarial example generation algorithms,such as EARL and VBAD. The proposed algorithm can be used to provide adversarial training samples to enhance video model security.

Keywords　Adversarial example,Video recognition,Boundary,Black-box,Neural networks

1　引言

近年来,深度神经网络(DNN)在视频识别[1-4]、视频字幕[5-6]和视频分段[7-8]等许多视频相关任务中表现优异。然而,相关研究表明,加入噪声的视频对抗样本会导致 DNN 误分类[9],这阻碍了其在安全性要求较高领域的应用。因此,部分研究开始聚焦于视频对抗样本(或对抗视频)生成领域[10-12]。

对抗样本生成起源于图像识别领域,故其在该领域的研究相对充分。根据算法可以获取被测试模型信息的程度,可以将对抗样本生成分为白盒、基于分数的黑盒和边界黑盒3种类型。白盒是分类中条件最宽松、最容易实现的类型,在该场景中,算法可以获取被测试模型的所有信息,包括其网络架构、网络参数、概率分数和分类标签等。在具体实施中,可将对抗样本生成问题转换为优化问题,然后通过正则化分类损失函数[13-14]或将对偶问题转换为约束优化问题来解决[15-17]。在基于分数的黑盒条件下,算法只能获取被测试模型的概率分数和分类标签,并通过有限差分(FD)[18-19]或自然

到稿日期:2024-07-08　返修日期:2025-03-01
通信作者:吴立军(wljuestc@sina.com)

进化算法(NES)[20]进行梯度估计。在边界黑盒条件下，算法只能获取被测试模型的分类标签，然后使用梯度估计算法[21]为对抗样本生成提供方向。在3种类型中，边界黑盒被公认为是最具有现实意义且最难实现的类型，因为在实际应用中，如MEGVII Face++，Microsoft Azure等商用模型只会为用户返回最终的分类标签，算法无法获取网络结构、网络参数和概率分数等与被测试模型相关的其他信息。

相较于图像领域，对抗样本生成在视频领域的研究相对较少，并且除STDE[22]外，大部分视频对抗样本生成相关研究都聚焦于白盒和基于分数的黑盒场景[23-24]。STDE是目前最新的边界黑盒视频样本生成算法，但其产生的对抗视频包含的噪声较大。为了弥补这一不足，本文提出了一种基于边界的黑盒视频对抗样本生成算法BBVA，其整体框架如图1所示。在初始阶段，BBVA选取1个源视频和1个目标视频。首先，BBVA将源视频作为初始对抗视频，利用投影算法得到对抗视频和目标视频之间的边界对抗视频。由于生成的边界对抗视频兼具原始对抗视频和目标视频的特征，对噪声较敏感，若加入随机噪声，则其分类标签会在对抗视频和目标视频之间随机转换。接着，在梯度估计阶段，利用边界对抗视频对噪声的敏感性，使用蒙特卡罗算法估计边界对抗视频的梯度。最后，利用估计的梯度和边界对抗视频生成新的对抗视频，新生成的对抗视频与目标视频在视觉上相似度更高。当迭代多轮后，对抗视频与目标视频越来越相似，直至人类视觉无法感知两者的差异。在进行算法设计时，面临着以下挑战：

1)相较于二维的静态图片，视频一般具有4个维度，其探索空间更大，这使得在迭代生成样本过程中寻找有效噪声变得更加困难。

2)目前，有一些梯度估计算法被应用于视频对抗样本生成领域，但它们全都聚焦于白盒和基于分数的黑盒场景，如何在基于边界的黑盒视频场景中高效进行梯度估计仍然是一个挑战。

3)由于视频维度较大，与图像对抗样本相比，利用估计的梯度生成对抗视频相对困难。

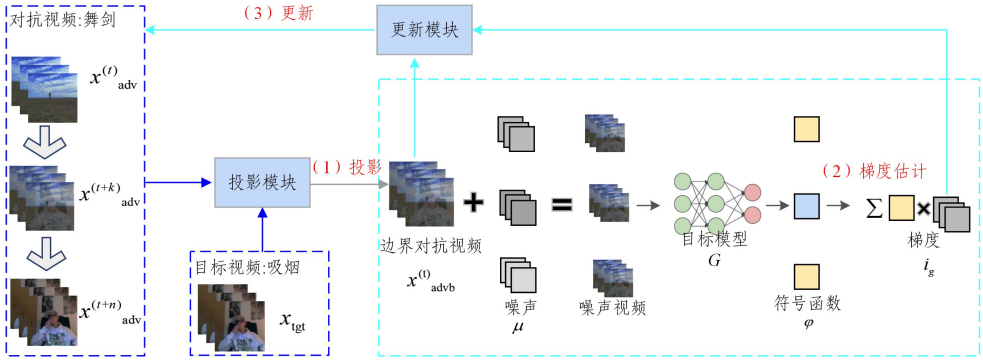

图1 BBVA架构

Fig. 1 Framework of BBVA

本文的主要贡献如下：

1)充分调研了边界黑盒视频对抗样本生成领域的研究现状，为改进现有算法的不足，针对性地提出了BBVA算法。实验表明，与最新的基于边界的黑盒视频对抗样本生成算法相比，BBVA较好地权衡了噪声大小和模型访问次数，其性能甚至优于一些最新的基于分数的黑盒视频对抗样本生成算法，如EARL，VBAD。

2)由于视频维度较大，与图像对抗样本相比，利用估计的梯度生成对抗视频相对困难。为了解决上述问题，采用了一种渐进式探索机制，有效提高了视频对抗样本生成的效率。

2 相关工作

对抗样本生成起源于图像识别领域，根据算法可以获取被测试模型信息的程度，可以将对抗样本生成分为白盒、基于分数的黑盒和边界黑盒3种类型。在白盒场景中，快速梯度符号方法(FGSM)[15]是一种一步生成算法，它通过沿着梯度方向增大分类损失函数的方式来产生对抗样本。与FGSM相同，投影梯度下降算法(PGD)[17]也是一种迭代生成方法，被公认为是最强大的一阶生成算法，其将加入到纯净样本的噪声限制在半径为ε的球形空间内。在基于分数的黑盒场景中，算法只能获取被测试模型的概率分数和分类标签，并用概率分数[18-19]进行梯度估计。其中，Bhagoji等[18]使用降维算法优化FD来提高梯度估计的速度；Chen等[19]使用有限差分(FD)算法进行梯度估计。迄今为止，自然进化策略(NES)[20]被认为是最快的梯度估计算法，其利用不同噪声对应的概率分数的变化进行梯度估计。在边界黑盒场景中，算法只能获取被测试模型的分类标签，然后使用梯度估计算法[21]为对抗样本生成提供方向。此外，还有一些提高噪声采样效率的算法，如QEBA[26]和NonLinear-BA[27]等。

相较于图像模型，对抗样本生成在视频模型领域的研究大多集中在白盒和基于分数的黑盒场景[10,12]，基于边界的黑盒场景的研究相对较少。其中，STDE[22]是最新的基于边界的黑盒视频对抗样本生成算法，它将目标视频作为原始噪声，并将其加入到通过时间差自适应算法选择的关键帧上，并在后续的过程中逐渐减小噪声的大小。VBAD[23]是首个基于分数的黑盒视频对抗样本生成算法，利用噪声在图像模型与视频模型之间的迁移性，用图像噪声初始化视频噪声；为了纠正噪声的偏差，VBAD将帧分成许多小块，然后使用NES算法

来估计这些小块的梯度。EARL也是基于分数的黑盒视频对抗样本生成算法[24]，其使用强化学习来选择视频的关键帧，并且只为这些关键帧添加噪声，从而降低了加入的噪声量并提升了生成效果。此外，一些算法研究噪声从图像模型到视频模型的迁移性。其中GIE[28]从整体和局部研究噪声在图像和视频间的通用性：整体方面，更加关注帧之间的相互关系；局部方面，打破了帧之间的相关性。GCMA[29]则通过时域一致性损失函数和特征差异损失函数来训练一个噪声采样网络，由该网络产生噪声，并提高噪声在图像模型和视频模型之间的迁移性。

3 本文算法

用 $G(x,\theta):R^{N\times H\times W\times C}\rightarrow R^{K}$ 表示一个视频模型，其中 x 为输入的视频，其真实分类标签为 y，θ 为模型 G 的参数，K 为模型 G 的分类数，而 N,H,W,C 分别表示视频的帧数、帧高度、帧宽度和信道数。视频对抗样本生成的目的是在原始视频样本 x 的基础上加入一些人类视觉无法感知的噪声，从而产生一个可以让模型错误分类的对抗视频 x_{adv}，即 $G(x_{adv},\theta)=y_{adv}$。其中 $y_{adv}\neq y$。为了确保 x 和 x_{adv} 在视觉上无法区分，算法往往会将 x_{adv} 限制在以 x 为中心、ε_{adv} 为半径的球形空间内，即 $\|x_{adv}-x\|_{p}\leqslant\varepsilon_{adv}$。为了后续阐述方便，这里先给出区别函数 D 和符号函数 φ 的定义：

$$D(x)=S(x)_{y_{adv}}-\max[S(x)_{y\neq y_{adv}}] \tag{1}$$

$$\varphi(x)=\begin{cases}1, & D(x)\geqslant 0\\ -1, & D(x)<0\end{cases} \tag{2}$$

其中，$S(x)_{y}$ 表示样本 x 被识别为类别 y 时对应的概率分数。在基于分数的黑盒场景中，算法可同时获得区别函数 D 和符号函数 φ 的值，但是在基于边界的黑盒场景中，算法只能得到符号函数 φ 的值。

BBVA算法整体流程如图1所示，其中，$x^{(t)}{}_{adv}$ 表示算法在第 t 次迭代产生的对抗视频，$x^{(t)}{}_{advb}$ 是 $x^{(t)}{}_{adv}$ 投影到分类边界上的边界对抗视频，x_{tgt} 为目标视频。初始阶段，BBVA算法选择2个视频 x_{src} 和 x_{tgt}（其对应的标签分别为 y_{adv} 和 y），并将 x_{src} 作为 $x^{(0)}{}_{adv}$ 开始循环运行算法的3个步骤：投影、梯度估计和更新。每一次通过 $x^{(t)}{}_{advb}$ 更新 $x^{(t)}{}_{adv}$，都使 $x^{(t)}{}_{adv}$ 在保持分类标签 y_{adv} 不变的前提下与 x_{tgt} 在视觉上相似度越来越高，最后达到人类视觉无法感知的程度，如图1的2个视频所示，算法最终生成了看起来与“吸烟”一模一样，但最终分类为“舞剑”的对抗视频。

3.1 投影

在梯度估计算法中，样本对噪声越敏感，则估计的梯度越准确。而对神经网络而言，处于分类边界上的样本对噪声最敏感，微小的随机噪声即可使边界样本在不同的分类之间振荡。在本研究中，$x^{(t)}{}_{adv}$（将 x_{src} 作为初始 $x^{(0)}{}_{adv}$）和 x_{tgt} 分别为高维空间中的2个点，如果按照特定比例加权求出位于两点之间的中间样本，则该样本同时具有对抗视频和目标视频的特征，相应地，视频模型对其的分类结果可能为 y_{adv} 或 y。中间样本越靠近 $x^{(t)}{}_{adv}$，其特征与 $x^{(t)}{}_{adv}$ 越相似，越有可能被分类为 y_{adv}，反之则易被分类为 y。不难看出，这是一个从 $x^{(t)}{}_{adv}$ 到 x_{tgt} 渐进变化的过程，随着中间样本从 $x^{(t)}{}_{adv}$ 向 x_{tgt} 移动，其对应的特征也从 $x^{(t)}{}_{adv}$ 向 x_{tgt} 转变，分类标签也由 y_{adv} 逐渐变为 y，该过程具有明显单调性。而投影算法的意义就是，通过调整加权比例 α 找到2个类别分类边界上的中间样本 $x^{(t)}_{adv}$。由于 $x^{(t)}{}_{advb}$ 对噪声最敏感，因此，使用其进行梯度估计效果最好。相关研究表明，二分搜索算法可以用于寻找类似问题的边界样本[21]。故本文结合加权求和算法和二分搜索算法，找到 $x^{(t)}{}_{adv}$ 和 x_{tgt} 之间的边界对抗视频 $x^{(t)}{}_{advb}$：

$$x^{(t)}{}_{advb}=\alpha\cdot x^{(t)}{}_{adv}+(1-\alpha)\cdot x_{tgt} \tag{3}$$

其中，α 为二分搜索参数，如果是第一次执行算法，则将 x_{src} 赋值到 $x^{(0)}{}_{adv}$。与 $x^{(t)}{}_{adv}$ 相比，投影后的 $x^{(t)}{}_{advb}$ 更加不稳定，其对噪声的敏感程度更高。如果加入一些随机的噪声，那么其最终的分类标签也会在 y_{adv} 和 y 之间随机分布，这有助于找到准确的梯度方向。详细描述如算法1所示，其中，h 为二分搜索阈值。

算法1 投影算法(Project)

输入：$(x^{(t)}_{adv},x_{tgt},y_{adv},h,G)$

输出：$x^{(t)}_{advb}$

1. 初始化：high←1，low←0，α←0；
2. while(high−low)>h：
3. α←(high+low)/2
4. $x^{(t)}_{advb}\leftarrow\alpha\cdot x^{(t)}_{adv}+(1-\alpha)\cdot x_{tgt}$
5. if $G(x^{(t)}_{advb})=y_{adv}$：
6. high←α
7. else：
8. low←α
9. return $x^{(t)}_{advb}$

3.2 梯度估计

由于 $x^{(t)}{}_{advb}$ 对噪声的敏感程度较高，若加入一些随机采样的噪声，则其最终的分类标签也会在 y_{adv} 和 y 之间随机分布。基于此，本文使用蒙特卡罗算法在 $x^{(t)}{}_{advb}$ 附近进行梯度估计，从而得到 $x^{(t)}{}_{advb}$ 下一步的更新方向：

$$g=\frac{1}{N}\sum_{k=1}^{N}\varphi(x^{(t)}{}_{advb}+\delta\mu)\cdot\mu \tag{4}$$

$$i_{g}=\frac{g}{\|g\|} \tag{5}$$

其中，δ 为固定常数，μ 为采样的噪声。详细描述如算法2所示。

算法2 梯度估计(GraEst)

输入：$(x^{(t)}_{advb},\delta,\mu,G,N)$

输出：i_g

1. $x\leftarrow x^{(t)}_{advb}+\delta\mu$
2. $D(x)\leftarrow S(x)_{y_{adv}}-\max[S(x)_{y\neq y_{adv}}]$
3. if D(x)>0：
4. φ(x)←1
5. else：
6. φ(x)←−1
7. $g\leftarrow\frac{1}{N}\sum_{k=1}^{N}\varphi(x^{(t)}_{advb}+\delta\mu)\cdot\mu$

8. $i_g \leftarrow \frac{g}{\|g\|}$

9. return i_g

3.3 更新

根据估计的梯度移动$x^{(t)}_{advb}$,得到新的对抗视频$x^{(t+1)}_{adv}$:

$$x^{(t+1)}_{adv} = x^{(t)}_{advb} + \eta \cdot \boldsymbol{i}_g \tag{6}$$

其中,η为步长,$\boldsymbol{i}_g$为单位化后的梯度向量。由于视频维度较大,与图像对抗样本相比,利用估计的梯度生成视频对抗样本相对困难。为解决上述问题,采用了一种渐进式探索机制,有效提高了对抗视频生成的效率。更新后,新生成的$x^{(t+1)}_{adv}$比$x^{(t)}_{adv}$在视觉上更像x_{tgt},但其分类标签始终为y_{adv}。完整描述如算法 3 所示,其中η_{ini}为 η 的初始值,d 为调整参数。

算法 3 更新算法(Update)

输入:$(x^{(t)}_{advb}, \eta_{ini}, y_{adv}, d, G, \mathbf{i}_g)$

输出:$x^{(t+1)}_{adv}$

1. 初始化:$\eta \leftarrow \eta_{ini}$;
2. while true:
3. $\quad x^{(t+1)}_{adv} \leftarrow x^{(t)}_{advb} + \eta \cdot \mathbf{i}_g$
4. $\quad$ if $G(x^{(t+1)}_{adv}) = y_{adv}$:
5. $\quad\quad$ break
6. $\quad$ else:
7. $\quad\quad \eta \leftarrow \eta \cdot d$
8. return $x^{(t+1)}_{adv}$

综上所述,BBVA 整体算法如算法 4 所示。首先,使用投影算法 Project 找到边界对抗视频$x^{(t)}_{advb}$;接着,使用梯度估计算法 GraEst 在$x^{(t)}_{advb}$附近进行梯度估计,得到下一步$x^{(t)}_{advb}$的移动方向 i_g;最后,以 i_g 为基准,利用融合了渐进式探索机制的更新算法 Update 更新$x^{(t)}_{advb}$,生成新的对抗视频$x^{(t+1)}_{adv}$。

算法 4 BBVA 算法

输入:$(x^{(t)}_{adv}, x_{tgt}, y_{adv}, h, G, \delta, \mu, N)$

输出:$x^{(t+1)}_{adv}$

1. $x^{(t)}_{advb} \leftarrow \text{Project}(x^{(t)}_{adv}, x_{tgt}, y_{adv}, h, G)$
2. $i_g \leftarrow \text{GraEst}(x^{(t)}_{advb}, \delta, \mu, G, N)$
3. $\quad x^{(t+1)}_{adv} \leftarrow \text{Update}(x^{(t)}_{advb}, \eta_{ini}, y_{adv}, d, G, i_g)$
4. return $x^{(t+1)}_{adv}$

4 实验

4.1 目标模型和数据集

实验使用的数据集为 HMDB-51[30] 和 UCF-101[31]。HMDB-51 是视频识别模型领域最常用的数据集之一,其包含 51 个类别,共 7000 个视频,其中训练集占 70%,测试集占 30%。UCF-101 是从 YouTube 收集的动作识别数据集,它包含 13320 个视频,分为 101 个类别,其中训练集占 80%,测试集占 20%,UCF-101 也是视频识别模型常用的标准数据集。

此外,选择 C3D[32] 和 TSN[33] 作为被测试模型。C3D 通过 3D 卷积层学习视频的高维特征,TSN 则使用稀疏时间采样策略,提升了视频特征提取效率。这两个模型是视频识别领域常用的主流模型。

4.2 评价指标

借鉴经典黑盒对抗样本生成领域的实验对比基准[22-24],使用以下指标来评估不同算法的效果。

1)均方差(MSE):对抗视频与原始目标视频之间的平均方差,表示加入目标视频的噪声的大小。MSE 越小,对抗视频与目标视频越相似,效果越好。

2)欺骗率(FR):即成功率,表示在特定模型访问次数和 MSE 阈值限制下,成功使被测试模型误分类的对抗视频数与总视频数的比值。算法 STDE[22],VBAD[23] 和 EARL[24] 也有类似评价指标,但其欺骗率的限制条件只有模型访问次数,即在一定模型访问次数限制内使模型误分类就算成功。但本文的欺骗率同时受模型访问次数和 MSE 阈值两个条件限制,即使在特定模型访问次数内成功使模型误分类,但若生成的对抗视频的 MSE 超过设定的阈值,仍算失败,这也是在进行对比实验时上述 3 种算法性能相对较差的原因之一。一般而言,欺骗率越高,算法性能越好。

3)平均访问次数(AQN):生成对抗视频需要频繁访问被测试模型,AQN 表示所有成功的对抗视频的平均模型访问次数。AQN 越小,算法性能越好。

4)噪声平均面积比(AOA):借鉴 STDE[22] 的对比指标,将加入的噪声和原始的 2 个测试视频的差异矩阵转换成连续像素组成的噪声块,并计算两者的面积比。AOA 越小,加入的噪声越少,算法性能越好。

4.3 对比算法

为了全面地评估 BBVA,在选择对比算法时,同时选择了主流的基于边界的黑盒对抗样本生成算法和基于分数的黑盒对抗样本生成算法。

1) 基于边界的黑盒对抗样本生成算法。将其与 STDE[22] 和 PSBA[25] 进行比较。STDE 是目前最新的基于边界的黑盒视频对抗样本生成算法,其首先使用时间差自适应算法选择关键帧,然后将目标视频作为原始噪声加入到关键帧上,并在后续的过程中逐渐减小噪声的大小。PSBA 也是一种基于边界的黑盒对抗样本生成算法,其主要通过优化采样噪声来提高效率。为了对比准确,在实验过程中使用与 BBVA 相同的视频预处理过程。

2) 基于分数的黑盒对抗样本生成算法。目前,主流的针对视频识别模型的基于分数的黑盒对抗样本生成算法有 VBAD[23],EARL[24] 等。将 BBVA 与 EARL 和 VBAD 进行比较,与这类算法作对比对于 BBVA 来说是一个挑战,因为基于分数的算法可以同时获取被测试模型的概率分数和分类标签,而 BBVA 只能获取分类标签。

4.4 实验设置

为了更好地凸显算法的性能差异,实验中将最大模型访问次数设置为 20000,即如果算法在 20000 次访问内未能生成成功的对抗视频,则该算法失败。据我们所知,这是截至目前在黑盒视频对抗样本生成领域最严格的测试条件。大多数算法的阈值都远大于 20000,例如,EARL 和 VBAD 分别将最大模型访问次数设置为 300000 和 600000,分别是本文算法

所设次数的 15 倍和 30 倍。

4.5 结果分析

为了评估 BBVA 的性能，汇总了对比算法在不同模型访问次数下所有成功对抗视频的平均 MSE 和 AOA，结果如表 1 和表 2 所列。可以看出，在测试 C3D 时，BBVA 的最优 MSE 和 AOA 分别为 29.4%和 10.69%，两者仅为 STDE 的 30%，PSBA 的 20%。在测试 TSN 时，BBVA 的最优 MSE 和 AOA 分别为 22.3%和 22.37%，均不到 STDE 和 PSBA 的 50%。表中没有 EARL 和 VBAD 的相关数据，是因为它们属于基于分数的算法，只会在算法结束时产生一个对抗视频。为了公平地与 VBAD 和 EARL 进行比较，汇总了 5 种算法生成成功对抗视频的 MSE，AOA 和 AQN，结果如表 3 所列。可以看出，BBVA 以最少的模型访问次数达到了最好的性能。具体来讲，在测试 C3D 时，在相同 AQN 的条件下，BBVA 的 MSE 和 AOA 比 STDE 小 68%，比 PSBA 小 80%。测试 TSN 时，BBVA 只需要 STDE 49% 的 AQN，就可以达到比 STDE 小 69%的 MSE 和 AOA；相较于 PSBA，BBVA 只需要 PSBA 48% 的 AQN 就可以达到比 PSBA 小 75%的 MSE 和 AOA。此外，在相同 AQN 的条件下，BBVA 的 MSE 和 AOA 均比 VBAD 小 14%。

表 1 不同访问次数下成功的对抗视频与目标视频的平均 MSE

Table 1 Average MSE between the successful adversarial video and the target video under different model queries

模型	算法	模型访问次数									
		0	2000	4000	6000	8000	10000	12000	14000	16000	18000	20000
C3D	BBVA	275	160	118.4	**90.1**	**68.2**	**54.9**	**47.4**	**40.2**	**35.1**	**32.0**	**29.4**
	STDE	275	**126**	**106.6**	101.2	99.2	98.0	97.1	95.3	94.4	94.0	93.3
	PSBA	275	222.1	192.3	172.9	164.6	160.7	157.8	155.9	154.3	153.2	152.5
TSN	BBVA	99.7	57.2	47.8	**40.6**	**35.5**	**31.9**	**28.5**	**26.5**	**24.6**	**23.2**	**22.3**
	STDE	99.7	**45.0**	**44.9**	44.9	44.9	44.9	44.9	44.9	44.9	44.9	44.9
	PSBA	99.7	84.1	74.4	68.5	63.3	58.9	55.7	53.8	52.4	51.3	50.6

注：加粗字体为最优值。

表 2 不同访问次数下成功的对抗视频与目标视频的平均 AOA

Table 2 Average AOA between the successful adversarial video and the target video under different model queries

(%)

模型	算法	模型访问次数									
		0	2000	4000	6000	8000	10000	12000	14000	16000	18000	20000
C3D	BBVA	0	58.18	43.05	**32.76**	**24.80**	**19.96**	**17.24**	**14.62**	**12.76**	**11.64**	**10.69**
	STDE	0	**45.82**	**38.76**	36.80	36.07	35.64	35.31	34.65	34.33	34.18	33.93
	PSBA	0	80.76	69.93	62.87	59.85	58.44	57.37	56.69	56.11	55.71	55.45
TSN	BBVA	0	57.37	47.94	**40.72**	**35.61**	**32.0**	**28.59**	**26.58**	**24.67**	**23.27**	**22.37**
	STDE	0	**45.14**	**45.04**	45.04	45.04	45.04	45.04	45.04	45.04	45.04	45.04
	PSBA	0	84.35	74.62	68.71	63.49	59.08	55.87	53.95	52.55	51.44	50.74

注：加粗字体为最优值。

表 3 成功的对抗视频的 AQN、平均 AOA 和平均 MSE

Table 3 AQN, average AOA and average MSE of successful adversarial videos

模型	算法	AQN	AOA/%	MSE
C3D	BBVA	**19910**	**10.76**	**29.6**
	STDE	19984	33.93	93.3
	EARL	—	—	—
	VBAD	—	—	—
	PSBA	19974	54.95	151.1
TSN	BBVA	**9662**	**12.54**	**12.5**
	STDE	19825	40.12	40.0
	EARL	—	—	—
	VBAD	9682	14.64	14.6
	PSBA	19865	49.65	49.5

注：加粗字体为最优值，“—”表示对应算法无法在 20000 次模型访问次数内生成成功的对抗视频。

当 MSE 阈值为 25 时，不同模型访问次数下各算法的欺骗率如表 4 所列(部分算法欺骗率一直为 0，表示在 20000 次模型访问次数内无法针对对应模型生成成功的对抗视频)。不难看出，在测试 C3D 时，STDE 可在 6000 次模型访问次数内将欺骗率提高到 30.1%，但在后续过程中，STDE 很难继续优化对抗视频，其最终欺骗率维持在 31.8%；相比之下，BBVA 可以随着模型访问次数的增加不断优化对抗视频，其最终欺骗率达到了 65.1%，是 STDE 的 2 倍；虽然 PSBA 也可以逐渐优化对抗视频，但其欺骗率较低，仅为 BBVA 的 25%。在测试 TSN 模型时，STDE 可在 2000 次模型访问次数内将欺骗率提高到 12.7%，但在后续测试过程中其欺骗率提升较小，最终为 16.4%；PSBA 虽然在测试过程中欺骗率逐渐提升，但其变化较小，最终欺骗率也相对较低；相比之下，BBVA 可随着模型访问次数的增加不断高效地优化对抗视频，其最终欺骗率达到了 80.1%，分别为 STDE 和 PSBA 的 4.8 倍和 6.5 倍；VBAD 在测试过程中虽然也可以不断优化对抗视频，但其欺骗率明显低于 BBVA。为了进一步评估 BBVA 在不同模型访问次数和 MSE 阈值下的性能，进行了对比实验，结果如表 5 所列。不难看出，在测试 C3D 时，只有 STDE，PSBA 和 BBVA 能够成功；当模型访问次数阈值为10000 时，BBVA 在大多数测试中都优于其他算法(当 MSE 为 5 时，STDE 略优于 BBVA)；当模型访问次数阈值大于10000 时，BBVA 的欺骗率是 STDE 的 1.5～2 倍，是 PSBA 的 4～6 倍。在测试 TSN 时，除

EARL外,其他4种算法均能成功生成对抗视频。具体来讲,当MSE小于10时,只有BBVA可成功生成对抗视频;当MSE大于10时,BBVA也远优于其他算法,其欺骗率甚至是STDE的4～11倍。

表4 在MSE阈值为25的条件下不同访问次数对应的FR

Table 4 Fooling rate under different model queries when the MSE threshold is 25

(%)

模型	算法	模型访问次数										
		0	2000	4000	6000	8000	10000	12000	14000	16000	18000	20000
C3D	BBVA	0	9.1	14.0	25.4	**33.1**	**44.8**	**53.4**	**55.3**	**65.1**	**65.1**	**65.1**
	STDE	0	**21.3**	**25.7**	**30.1**	30.1	30.9	30.9	30.9	31.8	31.8	31.8
	EARL	0	0	0	0	0	0	0	0	0	0	0
	VBAD	0	0	0	0	0	0	0	0	0	0	0
	PSBA	0	2.4	5.8	9.3	12.5	14.1	14.8	14.8	16.4	16.4	16.4
TSN	BBVA	0	10.3	**18.3**	**43.4**	**54.1**	**59.8**	**62.7**	**65.1**	**70.2**	**75.3**	**80.1**
	STDE	0	**12.7**	14.2	14.2	15.3	16.4	16.4	16.4	16.4	16.4	16.4
	EARL	0	0	0	0	0	0	0	0	0	0	0
	VBAD	0	5.6	11.4	20.4	30.6	31.6	34.4	34.7	49.7	58.1	59.8
	PSBA	0	3.1	6.3	7.6	8.8	10.2	10.9	10.9	12.4	12.4	12.4

注:加粗字体为最优值。

表5 不同MSE阈值和模型访问次数对应的FR

Table 5 Fooling rate under different MSE thresholds and model queries

模型	MSE阈值	最大模型访问次数=10000/%					最大模型访问次数=15000/%					最大模型访问次数=20000/%				
		BBVA/	STDE/	EARL/	VBAD/	PSBA/	BBVA/	STDE/	EARL/	VBAD/	PSBA/	BBVA/	STDE/	EARL/	VBAD/	PSBA/
C3D	25	**44.8**	30.9	0	0	14.1	**64.2**	31.8	0	0	15.2	**65.1**	31.8	0	0	16.4
	20	**39.4**	29.8	0	0	13.3	**52.8**	29.8	0	0	13.8	**63.8**	30.2	0	0	13.8
	15	**34.8**	25.3	0	0	9.4	**44.3**	25.6	0	0	9.4	**63.2**	25.7	0	0	10.3
	10	**23.7**	19.7	0	0	0	**37.6**	20.9	0	0	0	**39.1**	24.6	0	0	0
	5	14.2	**16.4**	0	0	0	**25.1**	16.4	0	0	0	**32.6**	17.2	0	0	0
TSN	25	**59.8**	16.4	0	31.6	10.2	**70.2**	16.4	0	44.8	12.4	**80.1**	16.4	0	59.8	12.4
	20	**48.9**	16.4	0	30.9	7.7	**67.7**	16.4	0	44.5	7.7	**73.6**	16.4	0	59.8	8.9
	15	**35.3**	4.8	0	25.5	2.2	**53.4**	4.8	0	41.2	3.1	**68.2**	5.9	0	41.8	3.1
	10	**13.3**	0	0	0	0	**44.6**	0	0	0	0	**58.5**	0	0	0	0
	5	**9.7**	0	0	0	0	**14.1**	0	0	0	0	**15.3**	0	0	0	0

注:加粗字体为最优值。

接下来对部分实验结果进行分析。

1)为什么EARL和VBAD的效果相对较差?

首先,两者都是基于分数的黑盒视频对抗样本生成算法,其梯度估计都是以概率分数为基础。与分类标签相比,概率分数值域更广,算法必须频繁访问被测试模型才可能计算出准确的梯度[23-24]。因此,在实验时,通常为这类算法设置较大的模型访问次数,EARL和VBAD分别将最大模型访问次数设置为300000和600000。为了突出BBVA的效果,将该值设置为20000,这给VBAD和EARL带来了较大挑战,导致其欺骗率较低。其次,为了估计EARL成功所需的最小模型访问次数,逐渐放宽阈值限制,发现其至少需要36463次模型访问次数才可成功生成对抗视频,并且其最小MSE为21.9。在STDE的原始论文[22]中,也可以得到和本文类似的实验结果。最后,除算法本身原因外,本文设计的对比指标欺骗率也相对苛刻,一定程度上导致EARL和VBAD表现相对较差,详细阐述见4.2节。

2)为什么BBVA的效果优于STDE?

STDE是基于补丁的算法[22],而BBVA是基于像素的算法,因此BBVA可以更准确地为像素添加噪声,STDE则使用目标视频作为初始补丁,并将其添加到源视频中,然后逐步优化并减小补丁的大小。此外,STDE是基于关键帧的算法,其只在选择的关键帧上加入噪声,而BBVA则将所有视频帧作为整体,根据梯度估计情况动态优化噪声。因此,在实验初始阶段,STDE加入的噪声相对较少,其MSE和欺骗率略优于BBVA。但STDE的噪声优化能力不如BBVA,这也是其在后续整体的测试过程中表现不如BBVA的原因。为了形象化阐述该问题,汇总了BBVA和STDE的部分成功的对抗视频,如图2所示。第1－4列为原始纯净视频,其对应的标签分别为"舞剑""拳击""微笑""笑"。第5－8列是成功的对抗视频,其中,第5和第6列的分类标签为"拳击",第7和第8列的分类标签为"笑"。第5列为STDE访问被测试模型20000次生成的成功对抗视频,其MSE为218.2,第6列是BBVA访问被测试模型10061次生成的成功对抗视频,其MSE为29.7。显然,相较于BBVA,STDE加入的噪声很容易被人类视觉识别出,并且BBVA只用STDE一半的模型访问次数就可以达到比STDE小86%的MSE。第7列是STDE访问被测试模型20000次生成的成功对抗视频,其MSE为8.7。第8列是BBVA访问被测试模型2350次生成的成功对抗视频,其MSE为8.1。由此可见,BBVA的模型访问次数仅为STDE的12%就达到了相同的效果。此外,当添加相同大小的噪声时,BBVA可以实现更好的视觉隐藏性。在STDE的原始论文[22]中,也可以得到和本文类似的MSE数据。除算法本身原因外,本文设计的对比指标欺骗率也相对苛刻,一定程度上导致STDE表现相对较差,详细阐述见4.2节。

(a)原始视频
标签:舞剑

(b)原始视频
标签:拳击

(c)原始视频
标签:微笑

(d)原始视频
标签:笑

(e)对抗视频
标签:拳击
算法:STDE
MSE:218.2

(f)对抗视频
标签:拳击
算法:BBVA
MSE:29.7

(g)对抗视频
标签:笑
算法:STDE
MSE:8.7

(h)对抗视频
标签:笑
算法:STDE
MSE:8.1

注:(1)第 5—8 列成功的对抗视频对应的模型访问次数依次为 20000,10061,20000 和 2350;(2)STDE 的对抗视频模型访问次数多,噪声易被人类视觉感知,BBVA 的对抗视频模型访问次数少,噪声隐藏性高。

图 2 BBVA 和 STDE 的效果对比

Fig. 2 Effects comparison of BBVA and STDE

3)为什么 VBAD 在测试不同模型时表现不同?

从实验结果可以看出,相较于 C3D 模型,VBAD 在测试 TSN 时表现更好,这是因为:TSN 模型利用稀疏帧进行预测,因此每帧的图像噪声影响较大[33];与 TSN 模型不同,C3D 主要使用稠密帧进行预测,每帧噪声的影响相对较小[32];VBAD 使用图像噪声来初始化帧噪声,相关研究证明[13],这种方式可以提高效率。因此,VBAD 在测试 TSN 时表现更好。

4)为什么 BBVA 的效果优于 PSBA?

相较于 PSBA,BBVA 采用了一种更高效的样本更新策略,可以随着测试过程的进行动态调整。相关研究证明[34-36],精确、动态的调整策略可提升样本更新效率,故 BBVA 优于 PSBA。

结束语 本文提出了一种基于边界的黑盒视频对抗样本生成算法 BBVA。该算法采用了一种渐进式探索机制生成对抗视频,有效提高了样本更新效率。实验表明,与最新的边界视频对抗样本生成算法相比,BBVA 较好地权衡了噪声大小和模型访问次数,甚至优于一些最新的基于分数的黑盒视频对抗样本生成算法,如 EARL 和 VBAD。总体来讲,BBVA 的性能显著优于其他对比算法。

参 考 文 献

[1] KARPATHY A, TODERICI G, SHETT Y, et al. Large-Scale Video Classification with Convolutional Neural Networks[C]// IEEE Conference on Computer Vision and Pattern Recognition. 2014:1725-1732.

[2] CARREIR A, JOÅ O, ZISSERMA N, et al. Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset [C]// IEEE Conference on Computer Vision and Pattern Recognition. 2017:4724-4733.

[3] WU Z X, JIANG Y G, WANG X, et al. Multi-Stream Multi-Class Fusion of Deep Networks for Video Classification [C]// Proceedings of the 24th ACM International Conference on Multimedia. 2016:791-800.

[4] ZHANG X, WU Z X, WENG Z J, et al. VideoLT: Large-Scale Long-Tailed Video Recognition [C]// Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). 2021:7960-7969.

[5] YANG Z W, HAN Y H, WANG Z, et al. Catching the Temporal Regions-of-Interest for Video Captioning[C]// Proceedings of the 25th ACM International Conference on Multimedia. 2017:146-153.

[6] LIU S, REN Z, YUAN J S, et al. SibNet: Sibling Convolutional Encoder for Video Captioning[C]// IEEE Transactions on Pattern Analysis and Machine Intelligence. 2021:3259-3272.

[7] NILSSON D, SMINCHISESCU C. Semantic Video Segmentation by Gated Recurrent Flow Propagation [C]// IEEE Conference on Computer Vision and Pattern Recognition. 2018:6819-6828.

[8] WANG W G, SONG H M, ZHAO S Y. Learning Unsupervised Video Object Segmentation Through Visual Attention [C]// IEEE Conference on Computer Vision and Pattern Recognition. 2019:3059-3069.

[9] WEI X X, ZHU J, YUAN S, et al. Sparse Adversarial Perturbations for Videos [C]// AAAI Conference on Artificial Intelligence. 2019:1101.

[10] WEI Z P, CHEN J J, WU Z X, et al. Boosting the Transferability of Video Adversarial Examples via Temporal Translation [C]// AAAI Conference on Artificial Intelligence. 2021:239016118.

[11] WEI Z, CHEN J, WU Z, et al. Cross-Modal Transferable Adversarial Attacks from Images to Videos [C]// IEEE Conference on Computer Vision and Pattern Recognition. 2022:15044-15053.

[12] LI S S, AJAYA N, PAUL S, et al. Adversarial Perturbations Against Real-Time Video Classification Systems [J]. arXiv:

1807.00458,2018.
[13] CHRISTIAN S,WOJCIECH Z,ILYA S,et al. Intriguing properties of neural networks [C]// International Conference on Learning Representations. 2014.
[14] CARLINI N,WAGNER D. Towards Evaluating the Robustness of Neural Networks [C]// IEEE Symposium on Security and Privacy. 2017:2375-1207.
[15] GOODFELLOW I J,JONATHON S,CHRISTIAN S,et al. Explaining and Harnessing Adversarial Examples [C]// International Conference on Learning Representations. 2015.
[16] ALEXEY K,GOODFELLOW I J,SAMY B,et al. Adversarial Machine Learning at Scale [C]// International Conference on Learning Representations. 2017.
[17] ALEKSANDER M,ALEKSANDAR M,LUDWIG S,et al. Towards Deep Learning Models Resistant to Adversarial Attacks [C]// International Conference on Learning Representations. 2018.
[18] BHAGOJI A N,HE W,LI B,et al. Practical Black-Box Attacks on Deep Neural Networks Using Efficient Query Mechanisms [C]// ECCV. 2018:158-174.
[19] CHEN P Y,ZHANG H,SHARMA Y 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.
[20] ILYAS A,ENGSTROM L,ATHALYE A,et al. Black-box Adversarial Attacks with Limited Queries and Information [C]// International Conference on Machine Learning. 2018:5046541.
[21] CHEN J B,JORDAN M I,WAINWRIGHT M. HopSkipJumpAttack:A Query-Efficient Decision-Based Attack [C]// IEEE Symposium on Security and Privacy(SP). 2020:1277-1294.
[22] JIANG K,CHEN Z,HUANG H,et al. Efficient Decision-based Black-box Patch Attacks on Video Recognition [C]// International Conference on Computer Vision. 2023:4356-4366.
[23] JIANG L X,MA X J,CHEN S X,et al. Black-Box Adversarial Attacks on Video Recognition Models [C]// ACM International Conference on Multimedia. 2019:864-872.
[24] YAN H Q,WEI X X. Efficient Sparse Attacks on Videos Using Reinforcement Learning [C]// ACM International Conference on Multimedia. 2021:2326-2334.
[25] ZHGNA J,LI L,LI H,et al. Progressive-scale boundary black-box attack via projective gradient estimation [C]// International Conferenceon Machine Learning. 2021:235417051.
[26] LI H C,XU X J,ZHANG X L,et al. QEBA:Query-Efficient Boundary-Based Blackbox Attack [C]// IEEE Conference on Computer Vision and Pattern Recognition. 2020:1218-1227.
[27] LI H C,LI L Y,XU X J,et al. Nonlinear Projection Based Gradient Estimation for Query Efficient Blackbox Attacks [C]// International Conference on Artificial Intelligence and Statistics. 2021.
[28] WANG R K,GUO Y F,WANG Y H,et al. Global-local characteristic excited cross-modal attacks from images to videos [C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2023:2635-2643.
[29] CHEN K,WEI Z P,CHEN J J,et al. GCMA:Generative Cross-Modal Transferable Adversarial Attacks from Images to Videos [C]// ACM International Conference on Multimedia. 2023:698-708.
[30] KUEHNE H,JHUANG H,GARROTE E,et al. HMDB:A large video database for human motion recognition [C]// International Conference on Computer Vision. 2011:2556-2563.
[31] KHURRAM S,AMIR ROSHAN Z,MUBARAK S,et al. UCF101:A Dataset of 101 Human Actions Classes From Videos in The Wild [J]. arXiv:1212.0402,2012.
[32] HARA K,KATAOKA H,SATOH Y,et al. Can Spatiotemporal 3D CNNs Retrace the History of 2D CNNs and ImageNet [C]// Conference on Computer Vision and Pattern Recognition. 2018:6546-6555.
[33] WANG L M,XIONG Y J,WANG Z,et al. Temporal Segment Networks:Towards Good Practices for Deep Action Recognition [C]// International Conference on Computer Vision. 2016:20-36.
[34] DIEDERIK P K,BA L J. Adam:A Method for Stochastic Optimization [C]// International Conference on Learning Representations. 2015.
[35] SASHANK J R,SATYEN K,SANJIV K,et al. On the Convergence of Adam and Beyond [C]// International Conference on Learning Representations. 2018.
[36] ZHANG M R,LUCAS J,HINTON G,et al. Lookahead optimizer:k steps forward,1 step back [C]// Proceedings of the 33rd International Conference on Neural Information Processing Systems. 2019:9597-9608.

JING Yulin, born in 1989, postgraduate. His main research interests include artificial intelligence security and computer vision.

WU Lijun, born in 1965, professor. His main research interests include artificial intelligence and information security.

(责任编辑:何杨)