工程科学学报
Chinese Journal of Engineering

针对视频分类模型的共轭梯度攻击

戴扬 冯旸赫 黄金才

Adversarial attacks on videos based on the conjugate gradient method

DAI Yang, FENG Yanghe, HUANG Jincai

引用本文:
戴扬, 冯旸赫, 黄金才. 针对视频分类模型的共轭梯度攻击[J]. 北科大：工程科学学报, 2024, 46(9): 1630–1637. doi: 10.13374/j.issn2095–9389.2023.07.25.004
DAI Yang, FENG Yanghe, HUANG Jincai. Adversarial attacks on videos based on the conjugate gradient method[J]. *Chinese Journal of Engineering*, 2024, 46(9): 1630–1637. doi: 10.13374/j.issn2095–9389.2023.07.25.004

在线阅读 View online: https://doi.org/10.13374/j.issn2095–9389.2023.07.25.004

工程科学学报, 第 46 卷, 第 9 期: 1630−1637, 2024 年 9 月
Chinese Journal of Engineering, Vol. 46, No. 9: 1630−1637, September 2024
https://doi.org/10.13374/j.issn2095-9389.2023.07.25.004; http://cje.ustb.edu.cn

针对视频分类模型的共轭梯度攻击

戴　扬, 冯旸赫, 黄金才✉

国防科技大学系统工程学院, 长沙 410073
✉通信作者, E-mail: huangjincai@nudt.edu.cn

摘　要　基于深度神经网络的视频分类模型目前应用广泛, 然而最近的研究表明, 深度神经网络极易受到对抗样本的欺骗. 这类对抗样本含有对人类来说难以察觉的噪声, 而其存在对深度神经网络的安全性构成严重威胁. 尽管目前已经针对图像的对抗样本产生了相当多的研究, 针对视频的对抗攻击仍存在复杂性. 通常的对抗攻击采用快速梯度符号方法 (FGSM), 然而该方法生成的对抗样本攻击成功率低, 以及易被察觉, 隐蔽性不足. 为解决这两个问题, 本文受非线性共轭梯度下降法 (FR–CG) 启发, 提出一种针对视频模型的非线性共轭梯度攻击方法. 该方法通过松弛约束条件, 令搜索步长满足强 Wolfe 条件, 保证了每次迭代的搜索方向与目标函数损失值上升的方向一致. 针对 UCF-101 的实验结果表明, 在扰动上界设置为 3/255 时, 本文攻击方法具有 91% 的攻击成功率. 同时本文方法在各个扰动上界下的攻击成功率均比 FGSM 方法高, 且具有更强的隐蔽性, 在攻击成功率与运行时间之间实现了良好的平衡.

关键词　对抗样本; 深度学习安全性; 视频攻击; 白盒攻击; 共轭梯度算法
分类号　TG142.71

Adversarial attacks on videos based on the conjugate gradient method

DAI Yang, *FENG Yanghe*, *HUANG Jincai*✉

College of Systems Engineering, National University of Defense Technology, Changsha 410073, China
✉Corresponding author, E-mail: huangjincai@nudt.edu.cn

ABSTRACT　Deep neural network-based video classification models enjoy widespread use because of their superior performance on visual tasks. However, with its broad-based application comes a deep-rooted concern about its security aspect. Recent research signals highlight these models' high susceptibility to deception by adversarial examples. These adversarial examples, subtly laced with humanly imperceptible noise, escape the scope of human detection while posing a substantial risk to the integrity and security of these deep neural network constructs. Considerable research has been directed toward image-based adversarial examples, resulting in notable advances in understanding and combating such adversarial attacks within that scope. However, video-based adversarial attacks highlight a different landscape of complexities and challenges. The nuances of motion information, temporal coherence, and frame-to-frame correlation introduce a multidimensional battlefield, necessitating purpose-built solutions. The most straightforward implementation of adversarial attacks uses the fast gradient sign method (FGSM). Unfortunately, FGSM attacks lack several respects: the attack success rates are far from satisfactory, they are frequently easily identifiable, and their stealth measures do not pass muster in rigorous environments. Therefore, this study introduces a novel nonlinear conjugate gradient attack method inspired by the nonlinear conjugate gradient descent method. By relaxing the search step size constraints to comply with the strong Wolfe conditions, we aimed to maintain pace with the increasing loss value of our objective function. This critical enhancement helps maintain the trajectory of each iteration's search direction and the simultaneous increase in the loss value, thereby yielding more consistent results, which ensures that our attack method can achieve a high attack success rate and concealment after each iteration. Further invigorating testament to our approach's efficacy came

收稿日期: 2023−07−25
基金项目: 国家自然科学基金资助项目(62276272)

through experimental results on the UCF101 dataset, underlining an impressive 91% attack success rate when the perturbation upper limit is 3/255. Our method consistently and markedly outshone the FGSM in attack success rates across various perturbation thresholds—even as it offered superior stealth. More critically, it allowed us to strike an effective balance between the attack success rate and runtime, a potent recipe for a disruptive contribution to the fraternity of adversarial attacks in video classification models. This adversarial attack method considers generating video adversarial examples from an optimization perspective. This represents a step forward in the ongoing drive to develop robust, reliable, and efficient techniques to understand adversarial attacks, specifically for deep neural network-based video classification models.

KEY WORDS　adversarial attack; the security of DNN; video attack; white-box attack; conjugate gradient algorithm

近年来, 深度学习技术得到了迅速发展, 在人物行为分类[1]、异常检测[2]、视频内容分割[3]等各个领域取得了优异成绩. 然而最近的研究表明, 神经网络非常脆弱, 并且在图像、视频等领域中, 极易被经过精心设计的样本欺骗. 这种被人为精心设计, 用于欺骗神经网络的样本一般被研究人员统称为对抗样本[4].

对抗样本是指在原始样本上添加人类难以、甚至无法察觉的扰动, 但可以使目标模型错误决策的样本. 其存在揭示了深度神经网络潜在的安全性, 严重影响了深度神经网络在实际应用的可靠性. 例如, 通过戴上特制的眼镜框, 就能让广泛应用的人脸识别系统进行错误识别[5]. 同时, 对抗样本具有可转移性, 即为攻击某一模型而设计的对抗样本, 可以直接用于攻击其他结构的模型[4]. 这些例子都说明深度神经网络面对对抗样本时是脆弱的. 因此, 工业界和学术界越来越关注深度神经网络的可解释性和安全性研究.

目前已有一些针对图像分类模型的攻击, 大致可以分为白盒攻击和黑盒攻击两类. 对于白盒攻击, Szegedy 等[4]在 2014 年首先通过 L-BFGS (Limited-memory Broyden–Fletcher–Goldfarb–Shanno) 算法生成了对抗样本, 其将对抗样本生成问题转化为了一个框约束问题进行求解. Goodfellow 等[6]在 2015 年提出了快速梯度法 (Fast gradient method, FGM) 和快速梯度符号法 (Fast gradient sign method, FGSM) 攻击方法, 该攻击方法实际开销很小, 而且行之有效.

在黑盒模型中, 攻击者无法获取目标模型的梯度, 但可通过查询来获得模型的预测结果及预测概率. 通常黑盒攻击可以通过估计梯度[7]、构建替代模型[8]、遗传算法[9]三种方法来生成对抗样本.

相较于图像分类模型, 针对视频分类模型的攻击还处于研究阶段. 视频数据相较于图像数据维度更高, 因此视频对抗样本的搜索空间更大. 同时由于视频数据的时序特性, 传统对抗样本生成方法容易造成“闪烁”现象, 隐蔽性不佳, 人眼更容易看出对抗样本攻击. 目前认为 Wei 等[10]首先提出了针对循环神经网络的白盒攻击方法, 并提出了稀疏攻击的思想, 保证了扰动的稀疏性和隐蔽性; Mu 等[11]提出了 DeepSava(Sparse adversarial video attacks) 攻击, 其采用了更合理的结构相似性 (Structure similarity index measure, SSIM) 而不是范数来度量原始样本和对抗样本间的距离, 并在视频攻击中第一个将加性扰动和空间变换扰动相结合. 即便目前已有一些视频攻击方法, 但是这些方法大多数都是从目标函数的角度来考虑的, 而对优化阶段的考虑不足.

本文提出了一种基于共轭梯度方法的视频对抗样本生成方法, 通过将对抗样本生成建模为优化过程, 并引入共轭梯度作为优化方向. 为了避免二阶优化方法计算量过大的问题, 加速对抗样本的生成, 本文使用了 FR–CG 迭代格式. 在 UCF-101 上的实验结果表明, 本文的视频对抗样本具有隐蔽性高、攻击成功率高的特点, 进一步说明了对抗样本带来的安全隐患.

1　共轭梯度法

优化问题一直是机器学习中的一个非常重要的领域, 假设目标函数为 f, $\boldsymbol{\theta}$为待优化的参数. 当 $f(\boldsymbol{\theta})$对其参数$\boldsymbol{\theta}$可微时, 可通过梯度下降法通过向当前位置梯度的负方向移动以实现最小化损失函数, 其更新公式为

$$\boldsymbol{\theta}_n = \boldsymbol{\theta}_{n-1} - \eta \cdot \nabla f(\boldsymbol{\theta}_{n-1}) \tag{1}$$

其中, $\boldsymbol{\theta}_n$表示第n轮批次更新的待优化参数, η为学习率, 需要人为给定. 若 $f(\boldsymbol{\theta})$为凸函数, 则梯度下降法可以得到全局最优解; 若 $f(\boldsymbol{\theta})$为非凸函数, 其可能趋于局部最优值.

最速下降法可以被视为梯度下降法的推广情况, 其区别在于前者的学习率一般通过线搜索算法计算得到, 而后者一般人为给定. 如果初始点情

况不好, 最速下降会出现迭代步数过多的问题, 这时候我们可以使用共轭梯度法来降低迭代步数.

首先考虑线性共轭梯度法, 对于二次凸问题的一般形式, 其待优化变量为 $\boldsymbol{x}\in\mathbf{R}^n$, $\boldsymbol{A}\in\mathbf{R}^{n\times n}$, $\boldsymbol{b}\in\mathbf{R}^n$为参数矩阵.

$$\min\ \frac{1}{2}\boldsymbol{x}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{x}-\boldsymbol{b}^{\mathrm{T}}\boldsymbol{x} \tag{2}$$

可以通过以下流程进行求解, 其中 x_k 表示第 k 轮批次更新的待优化变量.

$$\boldsymbol{x}_{k+1}=\boldsymbol{x}_k+\alpha_k\boldsymbol{p}_k \tag{3}$$

其中

$$\begin{cases}\boldsymbol{p}_0=-\boldsymbol{r}_0,\\ \boldsymbol{r}_0=\boldsymbol{A}\boldsymbol{x}_0,\\ \alpha_k=\dfrac{\boldsymbol{r}_k^{\mathrm{T}}\cdot\boldsymbol{r}_k}{\boldsymbol{p}_k^{\mathrm{T}}\boldsymbol{A}\boldsymbol{p}_k},\\ \boldsymbol{r}_k=\boldsymbol{r}_{k-1}+\alpha_{k-1}\boldsymbol{A}\boldsymbol{p}_{k-1},\\ \boldsymbol{p}_k=-\boldsymbol{r}_k+\beta_k\boldsymbol{p}_{k-1},\\ \beta_k=\dfrac{\boldsymbol{r}_k^{\mathrm{T}}\cdot\boldsymbol{r}_k}{\boldsymbol{r}_{k-1}^{\mathrm{T}}\cdot\boldsymbol{r}_k}\end{cases} \tag{4}$$

对于优化问题为非二次型的情况, 可以通过 FR–CG 进行求解. FR–CG 是 Fletcher 和 Reeves 在 1964 年提出的非线性共轭梯度法, 其直接基于共轭梯度法改进[12].

考虑无约束非线性优化问题, 其中 $f:\mathbf{R}^n\to\mathbf{R}$ 为目标函数

$$\min\ f(\boldsymbol{x}) \tag{5}$$

其迭代形式如下: $\boldsymbol{x}_{k+1}=\boldsymbol{x}_k+\alpha_k\boldsymbol{p}_k$. 对于步长 α_k, 通过线搜索算法进行确定, 对于迭代方向 $\boldsymbol{p}_k$, 通过下式进行更新

$$\boldsymbol{p}_{k+1}=-\nabla\boldsymbol{f}_{k+1}+\beta_{k+1}^{\mathrm{FR}}\boldsymbol{p}_k \tag{6}$$

其中

$$\beta_{k+1}^{\mathrm{FR}}=\frac{\nabla\boldsymbol{f}_{k+1}^{\mathrm{T}}\cdot\nabla\boldsymbol{f}_{k+1}}{\nabla\boldsymbol{f}_k^{\mathrm{T}}\cdot\nabla\boldsymbol{f}_k} \tag{7}$$

在非线性问题中, 共轭向量很难被构造出来, 且优化的精确步长很难求得. 因此通过线搜索算法来得到 $\boldsymbol{p}_k$ 方向上可接受的步长 α_k, 同时用目标函数的梯度 $\nabla\boldsymbol{f}_k$ 来替代残差 $\boldsymbol{r}_k$.

为了保证 FR–CG 每次迭代都能生成使得目标函数下降的方向, 在线搜索步长时, 需要令步长满足强 Wolfe 条件.

定义 1(强 Wolfe 条件) 如果任意步长 α_k 满足

$$\begin{aligned}&f(\boldsymbol{x}_k+\alpha_k\boldsymbol{p}_k)\leqslant f(\boldsymbol{x}_k)+c_1\alpha_k\nabla\boldsymbol{f}_k^{\mathrm{T}}\cdot\boldsymbol{p}_k,\\ &\left|\nabla f(\boldsymbol{x}_k+\alpha_k\boldsymbol{p}_k)^{\mathrm{T}}\cdot\boldsymbol{p}_k\right|\leqslant c_2\left|\nabla\boldsymbol{f}_k^{\mathrm{T}}\cdot\boldsymbol{p}_k\right|\end{aligned} \tag{8}$$

且 $0<c_1<c_2<1$ 任意给定, 则称它满足强 Wolfe 条件.

定理 1 设 $\{\boldsymbol{x}_k\}$ 为使用 FR–CG 格式得到得迭代序列. 强 Wolfe 条件中的系数满足 $0<c_1<c_2<0.5$, 那么搜索方向 $\boldsymbol{p}_k$ 和梯度方向的关系满足

$$-\frac{1}{1-c_2}\leqslant\frac{\nabla f(\boldsymbol{x}_k)^{\mathrm{T}}\cdot\boldsymbol{p}_k}{\|\nabla f(\boldsymbol{x}_k)\|^2}\leqslant\frac{2c_2-1}{1-c_2}<0 \tag{9}$$

证明: 使用数学归纳法, 当 $k=0$ 时, 初始搜索方向为负梯度方向, 易得结论成立. 假设其对于 k 成立, 对于 $k+1$ 时, 根据式 (6), 则有

$$\frac{\nabla f(\boldsymbol{x}_{k+1})^{\mathrm{T}}\cdot\boldsymbol{p}_{k+1}}{\|\nabla f(\boldsymbol{x}_{k+1})\|^2}=-1+\beta_{k+1}\frac{\nabla f(\boldsymbol{x}_{k+1})^{\mathrm{T}}\cdot\boldsymbol{p}_k}{\|\nabla f(\boldsymbol{x}_{k+1})\|^2} \tag{10}$$

又由式 (7), 得到

$$-1+\beta_{k+1}\frac{\nabla f(\boldsymbol{x}_{k+1})^{\mathrm{T}}\cdot\boldsymbol{p}_k}{\|\nabla f(\boldsymbol{x}_{k+1})\|^2}=-1+\frac{\nabla f(\boldsymbol{x}_{k+1})^{\mathrm{T}}\cdot\boldsymbol{p}_k}{\|\nabla f(\boldsymbol{x}_k)\|^2} \tag{11}$$

根据式 (8), 则得到

$$\left|\nabla f(\boldsymbol{x}_{k+1})^{\mathrm{T}}\cdot\boldsymbol{p}_k\right|\leqslant-c_2\nabla f(\boldsymbol{x}_k)^{\mathrm{T}}\cdot\boldsymbol{p}_k \tag{12}$$

从而代入式 (11) 得到

$$\begin{aligned}&-1+c_2\frac{\nabla f(\boldsymbol{x}_k)^{\mathrm{T}}\cdot\boldsymbol{p}_k}{\|\nabla f(\boldsymbol{x}_k)\|^2}\leqslant-1+\frac{\nabla f(\boldsymbol{x}_{k+1})^{\mathrm{T}}\cdot\boldsymbol{p}_k}{\|\nabla f(\boldsymbol{x}_k)\|^2}\leqslant\\ &-1-c_2\frac{\nabla f(\boldsymbol{x}_k)^{\mathrm{T}}\cdot\boldsymbol{p}_k}{\|\nabla f(\boldsymbol{x}_k)\|^2}\end{aligned} \tag{13}$$

而根据归纳假设有

$$\begin{cases}-1+c_2\dfrac{\nabla f(\boldsymbol{x}_k)^{\mathrm{T}}\cdot\boldsymbol{p}_k}{\|\nabla f(\boldsymbol{x}_k)\|^2}\geqslant-1-c_2\cdot\dfrac{1}{1-c_2},\\ -1-c_2\dfrac{\nabla f(\boldsymbol{x}_k)^{\mathrm{T}}\cdot\boldsymbol{p}_k}{\|\nabla f(\boldsymbol{x}_k)\|^2}\leqslant-1-c_2\cdot\dfrac{2c_2-1}{1-c_2}\end{cases} \tag{14}$$

从而结论成立.

2 对抗样本生成方法

2.1 L-BFGS 攻击

在图像、视频等输入数据中增加一些人类很难分辨出来的扰动, 能够导致神经网络分类错误, 甚至分类为攻击者需要的结果. 类似这样增加了扰动的输入数据, 我们一般称之为对抗样例, 或对抗样本.

Szegedy 等[4] 在 2014 年给出的对抗样本的定义是通过对测试集的图片样本添加难以察觉的非随机扰动, 可以任意改变神经网络的预测结果. 通过优化输入样本来最大化预测误差, 可以生成这些扰动. 广义上来说, 对抗样本是任何能够使得机器学习模型犯错的有效输入.

具体而言, 对抗样本生成问题是一个约束优化问题, 对于分类任务中的数据样本 $\boldsymbol{x}\in\mathbf{R}^n$, 令其

真实标签为$y\in\{1,\cdots,C\}$. 假定有分类器函数$F:\mathbf{R}^n\to\mathbf{R}$能够使得对于任意的$x\in\mathbf{R}^d$, 有$F(\boldsymbol{x})=y$. 则生成对抗样本的约束优化问题如下

$$\begin{aligned}&\max\ l(F(\boldsymbol{x}_0+\boldsymbol{\varepsilon}),y_0),\\&\text{s.t. }\|\boldsymbol{\varepsilon}\|\leqslant\epsilon,\ \boldsymbol{x}_0+\boldsymbol{\varepsilon}\in[0,1]^n\end{aligned}\tag{15}$$

其中, $\boldsymbol{\varepsilon}$为目标扰动, ϵ为给定的扰动上界, $(\boldsymbol{x}_0,y_0)$为样本点, $l(\cdot)$为损失函数.

一般来说, 这样的问题难以解决, 因此 Szegedy 等将其松弛为下面的框约束问题

$$\begin{aligned}&\min\ c\cdot\|\boldsymbol{\varepsilon}\|_2-l(F(\boldsymbol{x}_0+\boldsymbol{\varepsilon}),y_0),\\&\text{s.t. }\boldsymbol{x}_0+\boldsymbol{\varepsilon}\in[0,1]^n\end{aligned}\tag{16}$$

其中$c>0$为超参数. 对于这个框约束问题, 可以使用 L-BFGS-B 优化算法进行求解, 并使用线性搜索的方式来决定c的值.

2.2　FGSM 攻击

FGSM 攻击的思路为沿着损失函数梯度方向添加扰动, 具体为取度量为L_∞范数, 求解如下约束问题

$$\max\nolimits_{\boldsymbol{\varepsilon}:\|\boldsymbol{\varepsilon}\|_\infty\leqslant\epsilon}l(F(\boldsymbol{x}_0+\boldsymbol{\varepsilon}),y_0)\tag{17}$$

假设分类器和损失函数在样本点附近是线性的, 将目标函数在样本点$(\boldsymbol{x}_0,y_0)$附近泰勒展开, 得到

$$l(\widehat{\boldsymbol{\varepsilon}})=l(F(\boldsymbol{x}_0),y_0)+\nabla l(F(\boldsymbol{x}_0),y_0)\cdot\boldsymbol{\varepsilon}\tag{18}$$

从而优化问题变为如下形式

$$\begin{aligned}&\max\nolimits_{\boldsymbol{\varepsilon}:\|\boldsymbol{\varepsilon}\|_\infty\leqslant\epsilon}l(F(\boldsymbol{x}_0),y_0)+\nabla l(F(\boldsymbol{x}_0),y_0)\cdot\boldsymbol{\varepsilon}\Rightarrow\\&\max\nolimits_{\boldsymbol{\varepsilon}:\|\boldsymbol{\varepsilon}\|_\infty\leqslant\epsilon}\nabla l(F(\boldsymbol{x}_0),y_0)\cdot\boldsymbol{\varepsilon}\end{aligned}\tag{19}$$

由不等式

$$\begin{aligned}&\nabla l(F(\boldsymbol{x}_0),y_0)\cdot\boldsymbol{\varepsilon}\leqslant\|\nabla l(F(\boldsymbol{x}_0),y_0)\|\cdot\\&\|\boldsymbol{\varepsilon}\|\cos(\nabla l(F(\boldsymbol{x}_0),y_0),\varepsilon)\end{aligned}\tag{20}$$

得到满足约束条件的最大扰动为

$$\boldsymbol{\varepsilon}=\epsilon\cdot\nabla l(\theta,\boldsymbol{x},y)\tag{21}$$

对于任意L_p范数下的 FGSM 攻击来说, 类似可得对抗样本生成公式为

$$\boldsymbol{\varepsilon}=\epsilon\text{sign}(\nabla l(F(\boldsymbol{x}),y))\left(\frac{|\nabla l(F(\boldsymbol{x}),y)|}{\|\nabla l(F(\boldsymbol{x}),y)\|_{p^*}}\right)^{\frac{1}{p-1}}\tag{22}$$

其中, ϵ为给定的扰动上界, $\|\cdot\|_p$为给定的L_p范数, $p^*=\dfrac{p}{p-1}$.

3　共轭梯度 FGSM 攻击

本文基于FR–CG 迭代思想, 提出了共轭梯度FGSM攻击.

考虑优化问题为

$$\begin{aligned}&\min\nolimits_{\boldsymbol{\varepsilon}}\|\boldsymbol{\varepsilon}\|_2-c\cdot l_1(F(\boldsymbol{x}+\boldsymbol{\varepsilon}),y),\\&\text{s.t. }\ x_i+\varepsilon_i\in[\min(x_i),\max(x_i)],\ \ i=1,2,\cdots,n\end{aligned}\tag{23}$$

其中$\min(x_i)$和$\max(x_i)$分别代表输入视频第i帧的最小值和最大值, 即保证攻击后视频帧像素点强度不超过原视频帧的边界. $l_1(\cdot)$为衡量预测结果和真实标签的损失函数, 本文考虑在各类攻击算法中被广泛使用的交叉熵损失[13]. 为了能够使用 FR–CG 迭代格式, 将约束条件松弛为每次迭代将所有溢出边界的像素点强制设置为$\min(x_i)$和$\max(x_i)$, 从而将框约束优化问题转化为了无约束的标准优化问题, 并使用 FR-CG 迭代格式进行求解.

令$F(\boldsymbol{\varepsilon})=\|\boldsymbol{\varepsilon}\|_2-c\cdot l_1(F(\boldsymbol{x}+\boldsymbol{\varepsilon}),y)$, 则得到下式

$$\begin{aligned}&\min\nolimits_{\boldsymbol{\varepsilon}}F(\boldsymbol{\varepsilon}),\\&\text{s.t. }\ x_i+\varepsilon_i\in[\min(x_i),\max(x_i)],i=1,2,\cdots,n\end{aligned}\tag{24}$$

则得到算法流程如下所列:

算法3: 共轭梯度FGSM攻击算法

输入: 视频识别模型F, 原始视频样本x, 正确类别y.

输出: 对抗样本x_{adv}.

参数: 超参数c, 扰动上界$\boldsymbol{\epsilon}$, 最大迭代次数N, 最大步长搜索次数 max_iter, 强Wolfe系数c_1,c_2.

$c_1\leftarrow0.3$, $c_2\leftarrow0.3$ % 给定强Wolfe条件中系数的值

Def linesearch$(c_1,c_2,F,\alpha_k,\boldsymbol{p}_k,\boldsymbol{\varepsilon})$

　while i ⩽ max_iter

　　$\alpha_k\leftarrow\alpha_k/N$

　　if $F(\boldsymbol{x}_k+\alpha_k\boldsymbol{p}_k)\leqslant F(x_k)+c_1\alpha_k\nabla F_k^{\mathrm{T}}\boldsymbol{p}_k$ **and** $\left|\nabla F(\boldsymbol{x}_k+\alpha_k\boldsymbol{p}_k)^{\mathrm{T}}p_k\right|\leqslant c_2\left|\nabla F_k^{\mathrm{T}}\boldsymbol{p}_k\right|$

　　break

　　else $\alpha_k\leftarrow\text{serach}(\alpha_k-(0.5\alpha_k)/N,\alpha_k+(0.5\alpha_k)/N,10)$

　　end if

　end while

$\boldsymbol{r}_0\leftarrow\nabla F(x_0)$, $\boldsymbol{p}_0\leftarrow-\boldsymbol{r}_0$, $k\leftarrow0$

while $r_k\neq\mathbf{0}$ **and** $k<N$

　$\alpha_k\leftarrow\text{linesearch}(\alpha_k,\boldsymbol{p}_k,\boldsymbol{\varepsilon})$ % 使用线搜索算法确定非精确步长

　$\boldsymbol{x}_{k+1}\leftarrow\boldsymbol{x}_k+\alpha_k\boldsymbol{p}_k$

　$\left\{x_{k+1}^i\,\middle|\,\forall i,x_{k+1}^i>\max(x_i)\right\}\leftarrow\max(x_i)$ % 防止扰动超出上界

　$\left\{x_{k+1}^j\,\middle|\,\forall j,x_{k+1}^j<\min(x_i)\right\}\leftarrow\min(x_i)$ % 防止扰动超出下界

　$\boldsymbol{r}_{k+1}\leftarrow\nabla F(\boldsymbol{x}_{k+1})$

　$\beta_{k+1}\leftarrow\dfrac{\nabla F(x_{k+1})^{\mathrm{T}}\nabla F(x_{k+1})}{\nabla F(x_k)^{\mathrm{T}}\nabla F(x_k)}$

　$\boldsymbol{p}_{k+1}\leftarrow-\boldsymbol{r}_{k+1}+\beta_{k+1}\boldsymbol{p}_k$

　$k\leftarrow k+1$

end while

return $\boldsymbol{x}_{\text{adv}}\leftarrow\boldsymbol{x}_k$

4　实验结果

4.1　实验准备

所有的实验都在一台 CPU 为 Intel(R) Xeon(R) Platinum 8255, GPU 为 GeForce GTX 2080Ti 的计算机上进行, Python 版本为 3.7, Pytorch 版本为 1.7.0.

数据集准备

选择 UCF-101 数据集[14] 来评估提出的视频模型攻击方法在动作识别任务中的表现, 该数据集是在许多研究工作[1, 3, 11] 中广泛使用的公开动作识别数据集, 包含在不受约束的环境中录制的 101 个类别共 13320 个 YouTube 视频, 分属于“运动”、“仅限身体运动”、“人和物互动”、“人和人互动”、“乐器演奏”五大类型. 每个类别有 25 组视频, 每组包含了 4 ~ 7 个视频片段, 同一组的视频片段通常是相似的, 具有相同的主体和拍摄场景, 视频片段的平均长度为 7.21 s.

为了对齐视频分类模型的输入, 首先需要对数据集中的样本进行预处理, 具体操作如表 1 所示.

表 1　预处理设置

Table 1　Preprocessing settings

Preprocessing name	Specific operation	Description
Split	1	Use the official first partition file to split the training and test sets.
Size adjustment	(320, 240)→(256, 256)	Resize frames
Regularization	Mean = [123.675, 116.28, 103.53] std = [58.395, 57.12, 57.375]	Normalize the data using the RGB mean, where the mean and variance are the mean and variance of the test set.

视频分类模型

采用在视频分类领域表现极佳的模型, 分别是 TSN[15] 和 TSM[16−17] 双流模型. 两种模型的参数数值的设定与实现来自 MMaction2, 下面给出了各模型的基本情况.

TSN 模型: 以 Resnet-50 作为主干网络, 在 ImageNet 数据集上进行了预训练, 并在 UCF-101 训练集上训练了 75 个 epochs, 其报告的 GPU 显存占用为 8332 M. 测试方案为输入视频样本中的 25 帧并做 3-crops 数据增强. 最终汇报的 top-1 准确率为 83.03%, top-5 准确率为 96.78%.

TSM 模型: 以 Resnet-50 作为主干网络, 在 Kinetics-400 数据集上进行了预训练, 并在 UCF-101 训练集上训练了 25 个 epochs, 其报告的 GPU 显存占用为 10389 M. 测试方案为输入视频样本中的 13 帧, 不做数据增强. 最终汇报的 top-1 准确率为 94.58%, top-5 准确率为 99.37%.

在经过实际测试后, 得到 TSM 和 TSN 模型在 UCF-101 测试集上的实际准确率如表 2 所示.

表 2　目标模型 TSM 和 TSN 在测试集上的实际准确率

Table 2　Actual accuracy of the target models TSM and TSN on the test set

Target model	Accuracy rate/%
TSN	82.84
TSM	94.48

4.2　评估指标

与之前的工作类似, 本文报告攻击前后目标模型的分类准确率来表明提出攻击方法的攻击能力, 并通过 L_p 范数量化对抗样本和正常样本之间的相似性.

攻击能力

基于攻击前后目标模型的分类准确率的下降情况, 可以衡量对抗样本攻击能力水平, 具体计算公式如下

$$\text{Success Rate} = \frac{\text{Acc}_0 - \text{Acc}_1}{\text{Acc}_0} \tag{25}$$

其中, Acc_0 表示攻击前目标模型的准确率, Acc_1 表示攻击后目标模型的准确率.

隐蔽性指标

在对抗样本的生成过程中, 需要令最终的对抗样本和原始样本差异尽可能小, 从而保证人眼难以察觉对抗样本和原始样本的区别, 因而需要使用度量方法来约束对抗扰动的大小.

在图像和视频领域, L_1, L_2, L_∞ 是三种较为常用的基于 L_p 范数的对抗扰动度量方式. L_p 范数的计算公式如下所示.

$$\|\Delta \boldsymbol{x}\|_p = \sqrt[p]{\sum_{i=1}^{n} |x'_i - x_i|^p} \tag{26}$$

其中, $\Delta \boldsymbol{x}$ 表示添加的对抗扰动, x'_i 和 x_i 分别是 n 维视频样本对应的向量的第 i 个元素. p 的取值为整数, 可以是 1, 2, ∞等. 当 $p = \infty$ 时, 上式转化为

$$\|\boldsymbol{x}\|_{\infty}\max\left(\left|x'_1-x_1\right|,\cdots,\left|x'_n-x_n\right|\right) \tag{27}$$

4.3 实验结果

共轭梯度 FGSM 攻击结果与 FGSM 攻击结果对比如图 1 所示.

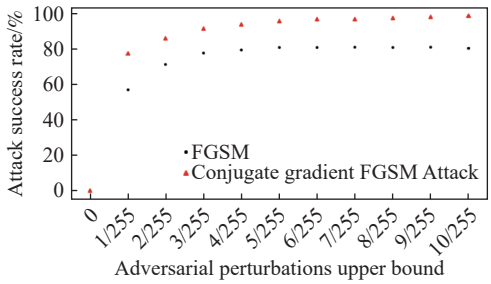

图 1　共轭梯度方法对 TSN 模型的攻击成功率

Fig.1　Attack success rate of the FGSM method and Conjugate Gradient FGSM method on the TSN model

从图 1 可以看出, 共轭梯度 FGSM 攻击相较于 FGSM 方法, 攻击成功率显著上升, 在扰动上界设置为 3/255 时即可取得 91% 的攻击成功率. 这也就是说, TSN 模型在被攻击前具有 82.84% 的准确率, 在遭受共轭梯度 FGSM 攻击后, 面对视频对抗样本的分类准确率只有 1.50%.

表 3 展示了共轭梯度攻击和 FGSM 攻击结果, 可以看出共轭梯度 FGSM 生成的对抗扰动成功率更高, 且具有更小的 L_1 和 L_2 度量.

图 2 是攻击 TSN 模型时, 在扰动上界设置为 25/255 时, 对抗扰动的可视化结果, 此时共轭梯度 FGSM 攻击的成功率已接近 100%. 图 2(a) 四张图为 FGSM 攻击得到的对抗扰动可视化情况, 图 2(b) 四张图为共轭梯度 FGSM 攻击的对抗扰动可视化情况, 可以看出共轭梯度攻击生成的扰动更难被人眼察觉.

图 3 展示了对抗样本攻击的隐蔽性, 其中受害者模型为 TSN 模型, 扰动上界设置为 3/255, 此时共轭梯度 FGSM 攻击成功率为 91%, 在遭受共轭梯度 FGSM 攻击后受害者模型仅有 1.5% 的准确率. 其中图 3(a) 中的四张图是原始样本 1 ~ 4 帧的可视化情况, 图 3(b) 中的四张图是 FGSM 攻击后的对抗样本的可视化情况, 该对抗样本被错误分类为了锤击 (Hammer throw), 图 3(c) 是共轭梯度 FGSM 攻击后对抗样本的可视化情况, 该对抗样本被错误分类为了投掷铁饼 (Throw discus). 图 3 说明了对抗样本与原始样本的差异极小, 但是却能够使得神经网络模型错误识别.

表 3　攻击结果对比

Table 3　Comparison of the attack results

Model	Attack method	Success rate/%	L_1	L_2	L_∞	Time/s
TSN	FGSM	79.85	577038	150	0.039	2.97
	Ours	98.19	236135	61	0.047	22.93
TSM	FGSM	77.83	46323	43	0.039	0.20
	Ours	88.52	20860	24	0.047	2.31

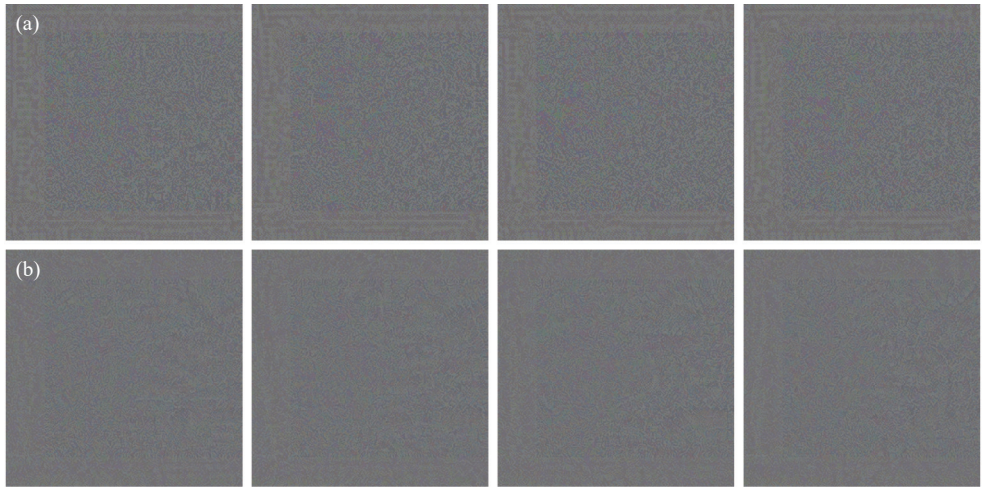

图 2　对抗扰动的可视化情况. (a) FGSM 攻击生成的对抗扰动; (b) 共轭梯度攻击生成的对抗扰动

Fig.2　Visualization against perturbations: (a) adversarial perturbations generated by FGSM attacks; (b) adversarial disturbance generated by Conjugate Gradient FGSM attack

图 3　原始样本与对抗样本的对比. (a) 原始样本; (b) FGSM 攻击生成的对抗样本; (c) 共轭梯度 FGSM 攻击生成的对抗样本

Fig.3　Comparison between original and adversarial examples: (a) original example; (b) adversarial example generated by FGSM attack; (c) adversarial example generated by Conjugate Gradient FGSM attack

5　结论

本文提出了用于视频识别模型的共轭梯度 FGSM 攻击. 在 UCF-101 数据集上对 TSN 和 TSM 模型进行了实验, 结果表明本文生成的对抗样本能够很好地欺骗视频识别模型, 且人眼难以察觉. 这说明对抗样本带来了严重的安全隐患, 深度神经网络具有潜在的安全问题.

本文提出的攻击算法均为白盒攻击算法, 但在实际中, 模型的内部信息一般是不公开的. 对于这种情况, 可以尝试使用合适的白盒模型替代目标模型. 我们之后将以目前的方法为基础, 通过构建替代模型进一步研究黑盒攻击.

最后, 尽管本文通过实验证明视频识别模型和图像分类模型一样, 非常容易被对抗样本攻击, 具有潜在的安全问题. 但是如何消除对抗样本带来的安全隐患仍是一个尚未解决的问题. 目前流行的研究思路是, 使用集成攻击的思想生成对抗样本, 并将对抗样本加入模型的训练集中进行对抗训练, 以提高模型的鲁棒性和安全性. 但是这种方法成本较高, 之后我们将继续研究如何更好地消除对抗样本带来的安全隐患.

参　考　文　献

[1] Feichtenhofer C, Fan H, Malik J, et al. Slowfast networks for video recognition // *Proceedings of the IEEE/CVF International Conference on Computer Vision*. Seoul, 2019: 6202

[2] Xia X, Pan X, Li N, et al. GAN-based anomaly detection: A review. *Neurocomputing*, 2022, 493: 497

[3] Gao M, Zheng F, Yu J J Q, et al. Deep learning for video object segmentation: a review. *Artificial Intelligence Review*, 2023, 56(1): 457

[4] Szegedy C, Zaremba W, Sutskever I, et al. Intriguing properties of neural networks // 2*nd International Conference on Learning Representations*. Toulon, 2014

[5] Sharif M, Bhagavatula S, Bauer L, et al. Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition // *Association for Computing Machinery*. New York, 2016: 1528

[6] Goodfellow I J, Shlens J, Szegedy C. Explaining and harnessing adversarial examples // *International Conference on Learning Representations*. San Diego, 2015

[7] 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 // *Proceedings of the* 10*th ACM Workshop on Artificial Intelligence and Security*. San Francisco, 2017: 15

[8] Juuti M, Szyller S, Marchal S, et al. PRADA: Protecting against DNN model stealing attacks // 4*th Proceedings of IEEE European*

Symposium on Security and Privacy. Stockholm, 2019: 512

[9] Alzantot M, harma Y, Chakraborty S, et al. GenAttack: practical black-box attacks with gradient-free optimization // *Proceedings of the Genetic and Evolutionary Computation Conference*. Stockholm, 2019: 1111

[10] Wei X X, Zhu J, Su H. Sparse adversarial perturbations for videos//*Proceedings of the AAAI Conference on Artificial Intelligence*, 2018, 33

[11] Mu R H, Ruan W, Marcolino L S, et al. Sparse adversarial video attacks with spatial transformations [J/OL]. *arXiv preprint* (2021-11-10)[2021-11-10] https://arxiv.org/abs/2111.05468, 2021-11-10

[12] Fletcher R, Reeves C M, Fletcher R, Reeves CM. Function minimization by conjugate gradients. *Comput J*, 1964, 7: 149.

[13] Chen P Y, Sharma Y, Zhang H, Yi J, et al. EAD: elastic-net attacks to deep neural networks via adversarial examples // *Proceedings of the ThirtySecond AAAI Conference on Artificial Intelligence*. New Orleans, 2018

[14] Soomro K, Zamir A R, Shah M. UCF101: a dataset of 101 human actions classes from videos in the wild. *Comput Sci*, 2012, 3(12): 2

[15] Wang L, Xiong Y, Wang Z, et al. Temporal segment networks: towards good practices for deep action recognition [J/OL]. *arXiv preprint* (2016-08-02) [2016-08-02]. https://arxiv.org/abs/1608.00859

[16] Ji H, Teng G, Yu J, et al. Efficient aggressive behavior recognition of pigs based on temporal shift module. *Animals*, 2023, 13(13): 2078

[17] Lin J, Gan C, Han S. Tsm: Temporal shift module for efficient video understanding // *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2019: 7083