计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2023,32(7):171−178 [doi: 10.15888/j.cnki.csa.009177]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

应用引导积分梯度的对抗样本生成①

王正来, 关胜晓

(中国科学技术大学 信息科学技术学院, 合肥 230026)
通信作者: 关胜晓, E-mail: guanxiao@ustc.edu.cn

摘　要: 给图片添加特定扰动可以生成对抗样本, 误导深度神经网络输出错误结果, 更加强力的攻击方法可以促进网络模型安全性和鲁棒性的研究. 攻击方法分为白盒攻击和黑盒攻击, 对抗样本的迁移性可以借已知模型生成结果来攻击其他黑盒模型. 基于直线积分梯度的攻击 TAIG-S 可以生成具有较强迁移性的样本, 但是在直线路径中会受噪声影响, 叠加与预测结果无关的像素梯度, 影响了攻击成功率. 所提出的 Guided-TAIG 方法引入引导积分梯度, 在每一段积分路径计算上采用自适应调整的方式, 纠正绝对值较低的部分像素值, 并且在一定区间内寻找下一步的起点, 规避了无意义的梯度噪声累积. 基于 ImageNet 数据集上的实验表明, Guided-TAIG 在 CNN 和 Transformer 架构模型上的白盒攻击性能均优于 FGSM、C&W、TAIG-S 等方法, 并且制作的扰动更小, 黑盒模式下迁移攻击性能更强, 表明了所提方法的有效性.

关键词: 深度神经网络; 对抗攻击; 积分梯度; 引导路径; 迁移攻击

引用格式: 王正来,关胜晓.应用引导积分梯度的对抗样本生成.计算机系统应用,2023,32(7):171–178. http://www.c-s-a.org.cn/1003-3254/9177.html

Adversarial Sample Generation Applying Guided Integrated Gradients

WANG Zheng-Lai, GUAN Sheng-Xiao

(School of Information Science and Technology, University of Science and Technology of China, Hefei 230026, China)

Abstract: Adding specific perturbations to images can help generate adversarial samples that mislead deep neural networks to output incorrect results. More powerful attack methods can facilitate research on the security and robustness of network models. The attack methods are divided into white-box and black-box attacks, and the transferability of adversarial samples can be used to attack other black-box ones by the results generated by known models. Attacks based on linear integrated gradients (TAIG-S) can generate highly transferable adversarial samples, but they are affected by noise in the linear path, superimposing pixel gradients that are irrelevant to the prediction results, which limits the success rate of attacks. With guided integrated gradients, the proposed Guided-TAIG method uses adaptive adjustment to correct some pixel values with low absolute values on each segment of the integrated path calculation and finds the starting point of the next step within a certain interval, circumventing the accumulation of meaningless gradient noise. The experiments on the ImageNet dataset show that Guided-TAIG outperforms FGSM, C&W, and TAIG-S for white-box attacks on both CNN and Transformer architecture models, produces smaller perturbations, and has better performance for transferable attacks in the black-box mode. This demonstrates the effectiveness of the proposed method.

Key words: deep neural network (DNN); adversarial attack; integrated gradients; guided path; transferable attack

① 收稿时间: 2022-12-23; 修改时间: 2023-01-17; 采用时间: 2023-02-27; csa 在线出版时间: 2023-05-12
CNKI 网络首发时间: 2023-05-16

近年来人工智能技术不断发展, 从人工神经网络到卷积神经网络再到 Transformer 架构, 深度学习模型在生活中有着越来越广泛的运用. 但是, 研究者发现深度神经网络非常容易受到对抗样本的威胁[1,2], 大量研究表明, 通过给良性样本添加特定的微小扰动制作成对抗样本在不引起人警觉的情况下可以轻易地欺骗模型, 从而产生大相径庭的预测结果. 攻击者可以通过伪造人脸图片来欺骗人脸识别模型从而非法入侵个人账户, 通过发射激光就可以攻击基于 AI 视觉的自动驾驶汽车[3]. 制作攻击性更强、隐蔽性更高的对抗样本是对抗学习研究的一个重要方向, 有利于评估网络模型的鲁棒性和安全性, 并促进更加强大的防御方法的开发.

根据攻击过程中掌握目标模型知识的程度可以将现有的对抗攻击方法分为白盒攻击和黑盒攻击[4]. 在白盒攻击方式中, 攻击者具有目标模型的完整知识, 可以分为基于梯度的攻击, 如 FGSM[2]、PGD[5], 基于优化的方法 C&W[6] 和基于生成对抗网络的攻击. 而黑盒攻击有很大限制, 攻击者无法访问模型细节, 只能查询访问的结果, 攻击难度更高, 在现实中更普遍. 黑盒攻击的一种实现方式是基于对抗样本迁移性实现迁移攻击, 以白盒攻击方法攻击代理模型生成对抗样本来获得对黑盒模型的欺骗. 提高模型的迁移性的优化方向主要有 3 类, 包括优化标准目标 (FGSM、PGD)、修改注意力 (AOA[7]、ATA[8]) 和平滑梯度 (DI[9]、TI[10]).

现有算法迁移攻击效果普遍不佳, Huang 等人[11] 将积分梯度 (integrated gradient, IG)[12] 引入对抗攻击, 提出了目前最先进的 TAIG 算法, 证明其可以从以上 3 个方面同时进行攻击, 在黑盒模式下具有较强的迁移攻击成功率. TAIG-S 是 TAIG 算法的一个标准版本, 计算从基线到目标对象直线路径上的梯度的累积, 可以有效解决梯度饱和问题, 将注意力图作为反向更新方向来生成对抗样本. 但是其采用的直线路径未经过优化, 易进入与预测结果不相关的像素从而导致误差累积, 产生无效噪声干扰, 在一定程度上损失了迁移攻击的成功率. 为了解决上诉问题, 本文将引导积分梯度 (Guided-IG)[13] 引入对抗攻击, 提出了 Guided-TAIG 方法. 具体的, 如图 1 所示, 在 TAIG-S 算法基础上, 计算积分梯度时摒弃简单的直线计算, 沿着输入、基线和模型确定的自适应方向, 将计算生成的引导积分梯度作为对抗攻击过程中梯度的更新方向. 该算法可以过滤不相关像素的梯度误差累积, 更精细准确地找到模型最关注的特征区域, 进而生成攻击性和迁移性更强的对抗样本.

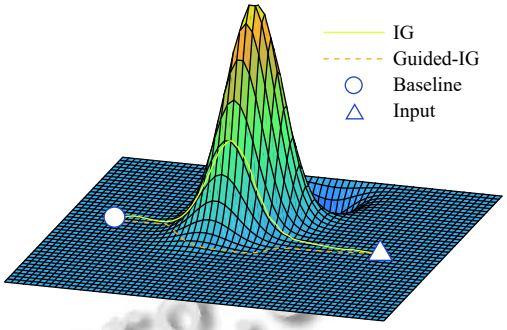


图 1 IG 与 Guided-IG 积分路径示意图

本文贡献主要有: (1) Guided-TAIG 将积分梯度的直线路径优化为自适应路径, 修正攻击过程的扰动方向, 精确攻击模型关注的区域; (2) 通过大量实验验证了所提算法在白盒和黑盒两种模式下都具有明显优越性.

1 相关工作

1.1 对抗攻击

对抗样本的概念最早由 Szegedy 等人[1] 于 2013 年提出, 发现在输入图片中添加微小的扰动成为对抗样本, 可以使图片以高置信度被误分为其他类别. 对抗样本的生成数学描述如下.

设图像分类器为函数$F:\mathbb{R}^n\to\mathbb{R}^k$, 其中, $x\in\mathbb{R}^n$为输入图像, $y\in\mathbb{R}^k$为分类得分向量, n表示输入图片的维度, k为分类器的类别数. 通过$Softmax$函数将输出归一化为: $\sum_{i=1}^{k}y_i=1$且$0\leqslant y_i\leqslant 1$, 其中$y_i$是$y$的第$i$个分量. 则图像预测结果为:

$$pred(x)=\arg\max_i Softmax(F(x)_i),\quad i=1,\cdots,k \tag{1}$$

假设图像的真实标签为y_{true}, 对抗攻击给原图添加微小噪声δ, 其中$\|\delta\|_p\leqslant\varepsilon$, ε为扰动约束, $\|\cdot\|_p$为L_p范数, 制作成可以使分类器分类错误的对抗样本x_{adv}, 使得:

$$pred(x_{\text{adv}})\neq y_{\text{true}},\quad \text{s.t.}\quad x_{\text{adv}}=x+\delta \tag{2}$$

1.2 TAIG-S 攻击算法

人类在做判断时往往选择性的关注信息的一部分, 忽略不重要的特征, 归因技术可以用来显示输入图片每个像素对最终结果的影响程度, TAIG-S 首次将积分梯度用于生成对抗样本, 积分梯度是神经网络归因分析的工具, 结合了直接梯度 DeepLift[14] 和基于反向传

播的归因技术 LRP[15], 满足敏感性和不变性, 可以表明每个像素对网络输出的重要性, 从而可以作为网络的注意力. 积分梯度是从参考图像$r \in \mathbb{R}^n$作为基线, 到输入图像x的直线路径梯度积分, 对输入图像的第i个像素的积分梯度的一个形式定义为:

$$IG_i(f,x,r) = (x_i - r_i) \times \int_{\eta=0}^{1} \frac{\partial f(r+\eta \times (x-r))}{\partial x_i} d\eta \quad (3)$$

其中, r_i是r的第i个像素. 基线r可以选取纯黑色、纯白色图片等. 完备性公理指出, $f(x)$与$f(r)$之间的差值等于$IG_i(f,x,r)$的和. TAIG-S 将计算生成的积分梯度结果作为对抗攻击反向更新的方向, 来擦除与预测结果最相关的特征.

1.3 引导积分梯度

积分梯度采用简单的直线路径, 生成的结果具有大量无效像素产生的非零梯度积分累计噪声, 这由于网络模型的曲面映射形状, 沿直线很容易穿过非均匀梯度区域. 引导积分梯度是一种自适应路径的方法, 如图 1 所示, 在积分梯度基础上, 舍弃固定的积分方向, 不是所有像素的增量都相等, 而是在每一步都进行选择, 选择一个像素子集, 这些像素具有不等于输入的最小绝对值, 然后下一步只移动子集中那些更接近输入图像强度的部分, 当所有像素的强度与输入一致时, 则路径更新结束.

2 基于 Guided-IG 的对抗样本生成方法

本文提出了一种基于引导积分梯度的对抗样本生成方法 Guided-TAIG, 下面详细介绍该方法的具体实现过程.

首先, 如第 1.1 节所描述, 定义一个映射函数$F: \mathbb{R}^n \to \mathbb{R}^k$, 其表示分类网络, 输入$n$维向量$X$, 并映射到含有$k$维的输出$Y$. 积分梯度的一个重要概念是基线值, 表示不含任何信息的量, 用X^B表示, 另一个概念是积分路径, 指连接基线值和输入值的参数曲线, 设$\gamma: \alpha \to \mathbb{R}^n, \alpha \in [0,1]$, 其中, $\gamma(\alpha=0)=X^B$, $\gamma(\alpha=1)=X^I$, 则对于输入X^I的每一维的积分梯度表示如下:

$$IG_i(X^I) = \int_{\alpha=0}^{1} \frac{\partial F(\gamma(\alpha))}{\partial \gamma_i(\alpha)} \frac{\partial \gamma_i(\alpha)}{\partial \alpha} d\alpha \quad (4)$$

积分梯度用路径上每一点的梯度累积来代替单独输入一点的梯度, 可以避免进入模型饱和区梯度为 0 的情况. 简单的, 当积分路径是直线, 即γ为一次函数, 那么$\partial\gamma_i(\alpha)/\partial\alpha$退变为常数, 实际计算时将 [0, 1] 区间分成m段, 每一段的路径为$\alpha_j \to \alpha_{j+1}$, 则其离散形式为:

$$IG(X^I) = \sum_{j=1}^{m} \sum_{i=1}^{n} \frac{\partial F(\gamma(\alpha_j)}{\partial \gamma_i(\alpha_j)} (\gamma_i(\alpha_j) - \gamma_i(\alpha_{j-1})) \quad (5)$$

原始的积分路径穿过非均匀的梯度区域, 路径上可能会存在着与输入无关的非零梯度方向, 随着积分的累积, 与预测结果不相关区域的梯度值不会相互抵消而会成为噪声, 由此在对抗攻击任务的过程中, 计算后的不准确的更新方向会降低生成的对抗样本的攻击性与迁移性. 而理想的积分路径γ^*, 文献 [13] 用式 (6)–式 (8) 来衡量:

$$\gamma^* = \arg\min\ l_{\text{noise}} + l_{\text{distance}} \quad (6)$$

$$l_{\text{noise}} = \sum_{i=1}^{n} \int_{\alpha=0}^{1} \left| \frac{\partial F(\gamma(\alpha))}{\partial \gamma_i(\alpha)} \frac{\partial \gamma_i(\alpha)}{\partial \alpha} \right| d\alpha \quad (7)$$

$$l_{\text{distance}} = \int_{\alpha=0}^{1} \left\| \gamma(\alpha) - \gamma^{\text{cal}}(\alpha) \right\| d\alpha \quad (8)$$

其中, l_{noise}是需要优化的噪声, 取值范围是无限的, 为了限制路径的取值, 引入有界限制l_{distance}, $\gamma^{\text{cal}}(\alpha)$是实际计算时的路径点.

在实际计算中, 采用离散化的思想进行近似, 将从基线X^B到输入X^I路径划分为m段, 每一段从起始点到目标点方向纠正的处理采用自适应的方法. 每一步中, 在每一点的梯度结果中找到一个特征子集S, 其是所有不等于输入X^I的像素中绝对值最小的一部分的集合, 选取比例为p. 具体的, 先得到集合S_{tmp}:

$$S_{\text{tmp}} = \{|\partial F(x)/\partial x_i|\},\ \text{其中},\ i \in \{j | x_j \neq x_j^I\} \quad (9)$$

将S_{tmp}从小到大排序, 选择前p部分的像素下标组成集合S, 则更新之后的路径方向修改为:

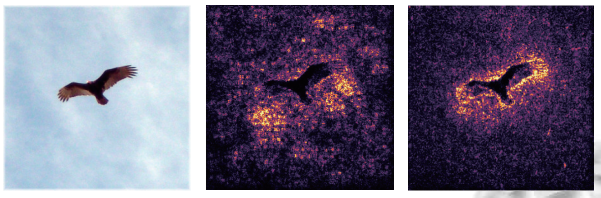
$$\left(\frac{\partial \gamma_i(\alpha)}{\partial \alpha}\right)^{\text{new}} = \begin{cases} x_i^I - x_i^B, \text{if } i \in S \\ 0, \text{其他} \end{cases} \quad (10)$$

令每一段的引导积分梯度为$Guided\text{-}IG_{j,j+1}$, 从而类比式 (5), 得到优化之后的积分梯度:

$$\begin{aligned} Guided\text{-}IG(X^I) &= \sum_{j=1}^{m} \sum_{i=1}^{n} Guided\text{-}IG_{j-1,j} \\ &= \sum_{j=1}^{m} \sum_{i=1}^{n} \frac{\partial F(\gamma(\alpha_j)}{\partial \gamma_i(\alpha_j)} \left(\frac{\partial \gamma_i(\alpha_j)}{\partial \alpha_j}\right)^{\text{new}} (\alpha_j - \alpha_{j-1}) \end{aligned} \quad (11)$$

图 2 展示了积分梯度与引导积分梯度生成的归因

图, 可以看出引导积分梯度过滤了原始梯度大量的无关噪声, 模型关注的区域更加集中. 引导积分梯度更能提取直接与网络模型预测结果相关的像素的强度, 反向更改这些像素点的值可以更改输入中包含的具体某一类别的信息, 这是对抗攻击可以进行的内部原因, 并可以直接应用于高质量对抗样本的生成过程. 类比 FGSM 和 I-FGSM[16] 类型的攻击方法, 可以有以下单步攻击和多步攻击形式.



(a) 原图 (b) IG (c) Guided-IG

图 2 IG 与 Guided-IG 归因图

对于输入图像X, 单步攻击定义如下:

$$X^* = X + \varepsilon \times sign(Guided\text{-}IG(X)) \tag{12}$$

其中, ε为扰动系数, 因为计算出的Guided-IG矩阵元素大小不均, 为了限制其无穷范数, 用$sign(\cdot)$函数提取积分梯度方向, 一步更新可以快速地生成对抗样本. 同时, 因为单步攻击只沿积分梯度方向更新一次, 跨越距离较大, 可以将一步扰动ε均分为多步小扰动形式, 设划分步数为T次, 则每次的扰动距离为ε/T. 在实际攻击过程中, 判断每一小步是否攻击成功, 如果不成功则再重新计算积分梯度进行下一小步的更新. 则所提出的 Guided-TAIG 多步攻击算法定义如下:

$$\begin{cases} X_0^* = X^I \\ X_{i+1}^* = Clip\left\{X_i^* + \dfrac{\varepsilon}{T} \times sign(Guided\text{-}IG(X_i^*))\right\} \end{cases} \tag{13}$$

其中, $Clip(\cdot)$为裁剪函数, 将第i步的更新结果限制在一定范围内. 算法伪代码如算法 1 所示.

3 实验分析

本节设计实验来验证和分析 Guided-TAIG 算法生成对抗样本的效果以及该方法的参数设计和对比研究. 分别介绍实验选用的数据集、白盒模式下的攻击效果、黑盒迁移的攻击性能以及算法参数分析.

3.1 实验设计

为了公正合理地评判不同攻击方法的优劣, 被测模型选择深度学习史上各阶段的经典模型 VGG-16[17], ResNet-50[18] , InceptionV3 (Inc-V3)[19], EfficentNet B3 (Effi-B3)[20]. 随着深度学习的发展, 近几年 Transformer 架构的模型逐步展现出比传统 CNN 架构模型更具优势, 所以也选择了最先进的基于 Transformer 架构的分类模型 Swin Transformer (Swin)[21] 和 MViTv2[22]. 并且对比了引入卷积的 Transformer 架构, convolutional Transformers (CvT)[23], 分别选取各种模型的版本为 Swin-B, MViTv2-B 和 CvT-13, 所有模型均在 ImageNet 数据集上进行训练.

本文使用的数据集为从 ImageNet[24] 挑选的 1 000 张图片, 并且保证所有分类模型都可以正确预测结果, 这样可以忽略不同模型分类准确度的差异, 只关注对抗样本生成算法本身在不同结构模型上的攻击效果.

算法 1. Guided-TAIG 流程

输入: 分类模型f, 图片x^I, 基线x^B, 标签y_{true}, 扰动距离ε, 更新步数 T, 离散路径分段m, 像素更新比例p.

输出: 对抗样本x_{adv}.

$x_{\text{adv}} \leftarrow x^I$
$map \leftarrow zero(x^I)$ //shape 与x^I一致
for $t \leftarrow 1$ to T do
  $x \leftarrow x_{\text{adv}}$
  $grad \leftarrow \nabla f(x, y_{\text{true}})$ //输入模型计算梯度
  for $\mu \leftarrow 1$ to m do
    $\theta \leftarrow \infty$ //临时变量
    while $\theta > 1$
      $tmp \leftarrow x$
      $d_t \leftarrow \|x^B - x^I\|_1 (1-\mu/m)$
      $d_c \leftarrow \|x - x^I\|_1$
      寻找$grad$中所有不等于X^I的像素中升序排列前p范围的集合S
      $d_s \leftarrow \sum_{i \in S} \|x_i - X_i^I\|$
      $\theta \leftarrow (d_c - d_t)/d_s$
      $x_i \leftarrow X_i^I, \forall i \in S$
      $grad_i \leftarrow x_i - x_i^B, \forall i \in S$
      $grad_i \leftarrow 0, \forall i \notin S$
      $map_i \leftarrow map_i + (x_i - tmp_i) \times grad_i$
    end while
  end for
  $x_{\text{adv}} \leftarrow x_{\text{adv}} + (\varepsilon/T) \times sign(map)$
  $x_{\text{adv}} \leftarrow Clip(x_{\text{adv}})$ //裁剪到有效区间
end for
return x_{adv}

比较算法选择白盒攻击算法 FGSM, C&W, 以及黑盒迁移性较强的攻击算法 TAIG-S. 在白盒攻击中限制最大扰动像素大小ε, TAIG-S 与 Guided-TAIG 都设

置相同的m值, 并固定 Guided-TAIG 的p值.

评测指标包括: (1) 攻击成功率: 攻击成功样本数/总样本数量; (2) 平均攻击耗时: 每个样本的平均生成时间; (3) 平均噪声和: 单样本平均添加的扰动在 RGB 三通道绝对值的和.

实验代码采用 PyTorch 框架编写, 硬件平台采用两块 NVIDIA GeForce RTX 3090.

3.2 白盒单步攻击

首先将 Guided-TAIG 与流行的攻击方法进行比较, 实验限制为白盒单步攻击, 即直接攻击模型本身, 并且对原图只进行一轮扰动更新, 限定相同的最大扰动距离ε, 这样可以比较不同算法计算的更新方向的优劣. 而如果采用多步攻击, 则大多数算法都可以很容易的取得接近 100% 的攻击成功率. C&W 是一种优化攻击, 其损失函数为对抗样本的预测输出和其 one-hot 编码差值的二范数, 实验采用迭代的方式优化噪声矩阵来扩大其损失函数. FGSM、TAIG-S 和 Guided-TAIG 的损失函数都为交叉熵函数. TAIG-S 和 Guided-TAIG 的分段值m设置为 30. 在ε分别为 8, 16, 25 时, 比较不同算法分别单步攻击 VGG-16、ResNet-50、Inc-V3、Effi-B3、Swin、MViTv2-B 和 CvT-13, 即设置攻击步数 $T=1$, 令 $p=20\%$, 攻击成功率见表 1. 从表中可以得出, 在白盒模式下, Guided-TAIG 也可以取得很好的攻击成功率, 比 TAIG-S 平均成功率高出 2%–3%, 超过了 FGSM 和 C&W 算法. 基于 Transformer 架构的模型相比 CNN 架构更难攻击. 同时也注意到 FGSM 在扰动距离增大的时候部分模型的攻击成功率不升反降, 这说明其优化方向并不精确, 给图像添加范围更大的错误扰动在一定程度上会降低对抗样本的攻击性. 而 TAIG-S 和 Guided-TAIG 计算的扰动反向更新方向更为准确, 随着扰动距离的增加, 对抗样本的攻击力显著增强.

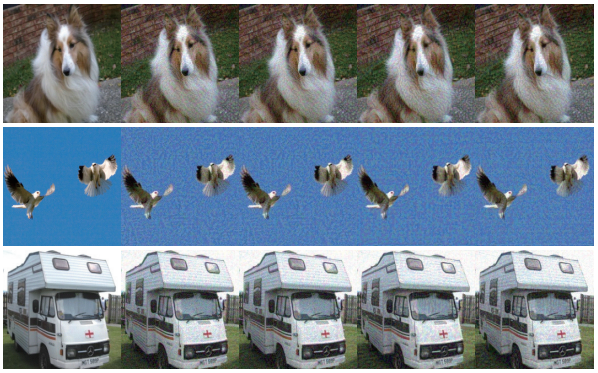
表 1 不同攻击算法单步攻击模型的成功率 (%)

算法	ε	VGG-16	ResNet-50	Inc-V3	Effi-B3	Swin-B	MViTv2-B	CvT-13
FGSM	8	**97.6**	**91.6**	82.4	68.1	**49.9**	**47.0**	61.8
	16	96.0	87.8	83.6	68.9	51.4	47.4	62.5
	25	95.6	85.5	82.4	70.8	53.3	46.7	67.1
C&W	8	96.8	89.6	82.4	69.2	49.0	45.3	63.5
	16	95.8	88.1	82.6	70.2	52.3	47.3	63.8
	25	95.1	86.0	81.4	71.1	51.1	47.0	65.7
TAIG-S	8	93.0	73.4	82.7	69.8	47.9	36.1	62.1
	16	96.4	86.5	92.3	79.3	**58.1**	39.2	72.4
	25	98.4	92.7	94.5	85.2	**64.0**	41.8	77.6
Guided-TAIG	8	93.2	76.6	**84.1**	**72.3**	44.9	42.3	**62.7**
	16	**96.8**	**89.6**	**92.8**	**79.6**	53.9	**47.8**	**72.8**
	25	**99.0**	**95.5**	**94.6**	**85.5**	58.6	**52.1**	**79.2**

给图片添加绝对值更小的扰动可以使攻击更加隐蔽, 表 2 比较了不同算法平均生成的每张对抗样本中修改像素 3 个颜色通道绝对值的和, 记为$|\widetilde{noise}|$, 并且比较了不同算法在所有模型生成样本的平均耗时. 设置同样的扰动距离$\varepsilon=16$, 迭代步数$T=1$, Guided-TAIG 攻击所有模型得到的$|\widetilde{noise}|$都是最小的, 因为像素的更新大小都是相同的, 在扰动的像素数量最少的情况下, 却得到了更优的攻击性能. 但是在耗时上 Guided-TAIG 处于劣势, 平均要花 1 s 多时间生成一张对抗样本, 并且分段数越多耗时显然会增加. 图 3 展示了攻击 CvT-13 模型的部分对抗样本. 第 1 列为原图, 第 2–5 列分别为不同算法生成的对抗样本示例, 从中可以看出 Guided-TAIG 制作的对抗样本噪声更加柔和, 在现实生活更有隐蔽性.

表 2 不同攻击算法单步攻击平均噪声和平均耗时

算法	ε	VGG-16	ResNet-50	Inc-V3	Effi-B3	Swin-B	MViTv2-B	CvT-13	耗时 (ms)
FGSM	16	2 273 271	2 273 039	2 214 500	2 273 340	2 273 335	2 273 569	2 273 009	**47**
C&W	16	2 273 283	2 273 031	2 212 800	2 273 327	2 273 327	2 273 573	2 273 019	49
TAIG-S	16	2 266 838	2 266 575	2 205 953	2 263 241	2 266 471	2 266 434	2 267 425	384
Guided-TAIG	16	**2 265 623**	**2 266 142**	**2 205 182**	**2 262 020**	**2 266 349**	**2 266 397**	**2 266 348**	1 125



(a) 原图 (b) FGSM (c) C&W (d) TAIG-S (e) Guided-TAIG

图 3 攻击 CvT-13 生成的对抗样本

3.3 黑盒迁移攻击

迁移攻击是黑盒攻击常用的攻击方式, 主要利用了对抗样本具有迁移性, 通过攻击代理模型来生成对未知模型的攻击, 不同模型都偏向于利用图片的相似特征. 固定$\varepsilon = 16$, $m = 30$, $T = 1$, 分别用 FGSM, C&W, TAIG-S 和 Guided-TAIG 算法对每个代理模型进行攻击, 然后将生成的对抗样本作为数据集来攻击其他模型, 得到的迁移攻击成功率见表 3. 表中每一行用同一种代理模型来迁移攻击其余黑盒模型, 相比 TAIG-S, 所提算法迁移攻击成功率提升了 0.4%–10%.

3.4 参数分析

这部分具体检验 Guided-TAIG 算法中不同参数对算法性能的影响.

首先固定 Guided-TAIG 积分梯度划分段$m = 30$, 扰动距离$\varepsilon = 16$, $p = 20\%$, 比较在不同的迭代步数T下白盒攻击目标模型的成功率, 见图 4. 横坐标为迭代的轮数, 纵坐标为攻击成功率, 随着T的增加, 攻击效果明显增强, 在经过 4–5 轮的迭代攻击后, 所有模型都达到了 95% 及以上的攻击成功率. 纯 CNN 架构的模型最容易攻击, 纯 Transformer 的模型攻击成功率最低, 而结合两种架构为一体的 CvT-13 的被攻击难度介于两者之间. 当然, 这并不能直接说明 Transformer 架构就比 CNN 更鲁棒, 因为模型具有更多的参数量, 训练过程中采用了更优秀的损失函数和更大的数据集, 额外需要更多的实验来验证.

表 3 单步攻击代理模型生成对抗样本的迁移攻击成功率 (%)

代理模型	算法	VGG-16	ResNet-50	Inc-V3	Effi-B3	Swin-B	MViTv2-B	CvT-13
VGG-16	FGSM	96.0	50.3	44.2	31.7	**23.5**	**17.8**	29.5
	TAIG-S	96.4	60.3	58.4	42.6	20.1	16.2	34.7
	Guided-TAIG	**96.8**	**63.0**	**61.7**	**42.9**	21.1	17.5	**36.3**
ResNet-50	FGSM	65.7	87.8	47.5	36.1	22.5	19.6	32.8
	TAIG-S	74.5	86.5	63.1	45.0	21.8	18.8	32.9
	Guided-TAIG	**74.9**	**89.6**	**67.0**	**48.5**	**24.3**	**20.6**	**36.5**
Inc-V3	FGSM	56.8	44.7	83.6	36.5	19.4	18.7	26.8
	TAIG-S	69.8	60.1	92.3	53.1	22.0	19.5	34.3
	Guided-TAIG	**70.4**	**61.0**	**92.8**	**53.8**	**23.8**	**20.9**	**35.6**
Effi-B3	FGSM	59.1	45.8	47.6	68.9	30.1	29.9	36.6
	TAIG-S	72.6	65.8	71.8	79.3	**36.0**	31.8	42.2
	Guided-TAIG	**73.3**	**66.1**	**72.6**	**79.6**	34.6	**32.4**	**45.2**
Swin-B	FGSM	47.0	35.0	36.8	30.6	51.4	29.8	33.2
	TAIG-S	**56.4**	42.5	**47.6**	37.3	**58.1**	**30.8**	**40.9**
	Guided-TAIG	55.0	**43.6**	45.3	**37.8**	53.9	29.5	38.6
MViTv2-B	FGSM	46.0	36.0	35.7	28.4	30.3	47.4	32.0
	TAIG-S	44.7	34.4	38.6	28.7	29.2	39.2	33.1
	Guided-TAIG	**53.2**	**44.5**	**44.2**	**36.9**	**35.6**	**47.8**	**38.2**
CvT-13	FGSM	52.8	40.2	40.0	34.0	33.0	31.9	62.5
	TAIG-S	66.2	**56.6**	**57.4**	49.3	46.4	41.4	72.4
	Guided-TAIG	**67.4**	55.6	56.8	**49.6**	**46.8**	**42.3**	**72.8**

Guided-TAIG 采用离散自适应的方式来计算积分梯度, 固定$\varepsilon = 16$, $T = 1$, $p = 20\%$, 下面比较不同分段数m下白盒攻击的成功率, 见图 5. 横坐标为m, 纵坐标为攻击成功率, 分别计算了m等于 5、10、20、30、40 下的攻击效果. 随着分段数划分的越多, 需要的计算量更高, 攻击成功率逐渐趋近于饱和, $m = 20$时已经达到了可接受的结果.

在更新路径时, 选择的最小部分比例p对结果也有

影响, 不同取值的攻击成功率见图 6. 固定$\varepsilon = 16$, $T = 1$, $m = 20$, p取值范围为 10%–60%, 攻击成功率大致先上升再下降, 在 20%–30% 之间取得最大值.

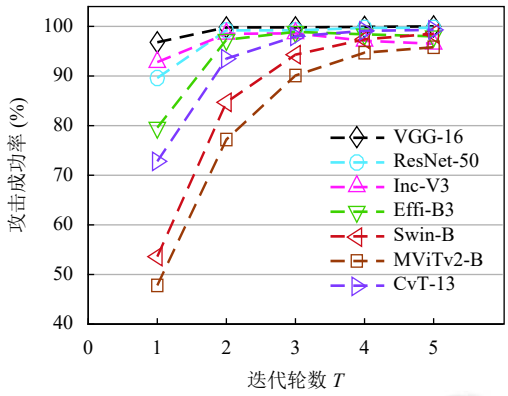


图 4 不同迭代轮数 T 攻击成功率

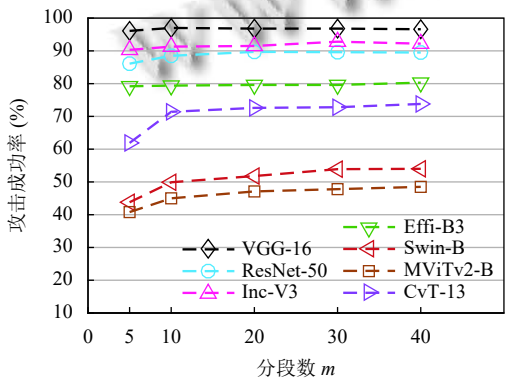


图 5 不同分段数m攻击成功率

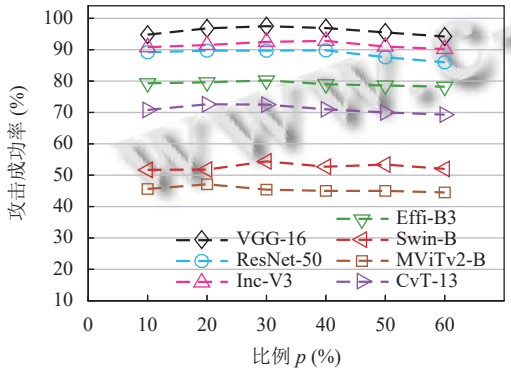


图 6 不同比例p攻击成功率

4 结论与展望

本文改进了 TAIG-S 算法, 引入引导积分梯度到对抗攻击, 提出了一种 Guided-TAIG 算法, 限制了更新路径不相关的梯度累积并分别给出了 Guided-TAIG 单步攻击和多步攻击的形式. 实验表明该方法对不同架构的模型都具有较强的白盒攻击成功率, 生成的对抗样本噪声更小, 更具隐蔽性. 在黑盒迁移攻击实验中取得了最好的效果, 在实际攻击中相比 TAIG-S 提升了 0.4%–10% 的成功率. 随着迭代步数的增加对实验模型较容易的达到 95% 以上的攻击成功率. 劣势是该算法精确计算每一小段方向, 耗时较多. 后续计划将对积分路径进行更加细致化的改进, 增强对抗样本的攻击性能.

参考文献

1 Szegedy C, Zaremba W, Sutskever I, *et al*. Intriguing properties of neural networks. arXiv:1312.6199, 2013.

2 Goodfellow IJ, Shlens J, Szegedy C. Explaining and harnessing adversarial examples. arXiv:1412.6572, 2014.

3 Duan RJ, Mao XF, Qin AK, *et al*. Adversarial laser beam: Effective physical-world attack to DNNs in a blink. Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021. 16057–16066. [doi: 10.1109/CVPR46437.2021.01580]

4 Ren K, Zheng TH, Qin Z, *et al*. Adversarial attacks and defenses in deep learning. Engineering, 2020, 6(3): 346–360. [doi: 10.1016/j.eng.2019.12.012]

5 Madry A, Makelov A, Schmidt L, *et al*. Towards deep learning models resistant to adversarial attacks. arXiv: 1706.06083, 2017.

6 Carlini N, Wagner D. Towards evaluating the robustness of neural networks. Proceedings of the 2017 IEEE Symposium on Security and Privacy (SP). San Jose: IEEE, 2017. 39–57. [doi: 10.1109/SP.2017.49]

7 Chen SZ, He ZB, Sun CJ, *et al*. Universal adversarial attack on attention and the resulting dataset DAmageNet. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 44(4): 2188–2197. [doi: 10.1109/TPAMI.2020.3033291]

8 Wu WB, Su YX, Chen XX, *et al*. Boosting the transferability of adversarial samples via attention. Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020. 1158–1167. [doi: 10.1109/CVPR42600.2020.00124]

9 Xie CH, Zhang ZS, Zhou YY, *et al*. Improving transferability of adversarial examples with input diversity. Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 2725–2734.

[doi: 10.1109/CVPR.2019.00284]

10 Dong YP, Pang TY, Su H, *et al*. Evading defenses to transferable adversarial examples by translation-invariant attacks. Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 4307–4316. [doi: 10.1109/CVPR.2019.00444]

11 Huang Y, Kong AWK. Transferable adversarial attack based on integrated gradients. arXiv:2205.13152, 2022.

12 Sundararajan M, Taly A, Yan QQ. Gradients of counterfactuals. arXiv:1611.02639, 2017.

13 Kapishnikov A, Venugopalan S, Avci B, *et al*. Guided integrated gradients: An adaptive path method for removing noise. Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021. 5048–5056. [doi: 10.1109/CVPR46437.2021.00501]

14 Shrikumar A, Greenside P, Kundaje A. Learning important features through propagating activation differences. Proceedings of the 34th International Conference on Machine Learning. Sydney: JMLR.org, 2017. 3145–3153.

15 Bach S, Binder A, Montavon G, *et al*. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. PLoS One, 2015, 10(7): e0130140. [doi: 10.1371/journal.pone.0130140]

16 Kurakin A, Goodfellow IJ, Bengio S. Adversarial examples in the physical world. arXiv:1607.02533, 2018.

17 Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. arXiv: 1409.1556, 2014.

18 He KM, Zhang XY, Ren SQ, *et al*. Deep residual learning for image recognition. Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016. 770–778. [doi: 10.1109/CVPR.2016.90]

19 Szegedy C, Vanhoucke V, Ioffe S, *et al*. Rethinking the inception architecture for computer vision. Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016. 2818–2826. [doi: 10.1109/CVPR.2016.308]

20 Tan MX, Le QV. EfficientNet: Rethinking model scaling for convolutional neural networks. Proceedings of the 36th International Conference on Machine Learning. Long Beach: PMLR, 2019. 6105–6114.

21 Liu Z, Lin YT, Cao Y, *et al*. Swin Transformer: Hierarchical vision Transformer using shifted windows. Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision. Montreal: IEEE, 2021. 9992–10002. [doi: 10.1109/ICCV48922.2021.00986]

22 Li YH, Wu CY, Fan HQ, *et al*. MViTv2: Improved multiscale vision Transformers for classification and detection. Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans: IEEE, 2022. 4794–4804. [doi: 10.1109/CVPR52688.2022.00476]

23 Wu HP, Xiao B, Codella N, *et al*. CvT: Introducing convolutions to vision Transformers. Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision. Montreal: IEEE, 2021. 22–31. [doi: 10.1109/ICCV48922.2021.00009]

24 Deng J, Dong W, Socher R, *et al*. ImageNet: A large-scale hierarchical image database. Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition. Miami: IEEE, 2009. 248–255. [doi: 10.1109/CVPR.2009.5206848]

(校对责编: 孙君艳)