Simplex 噪声区域中目标特征的对抗攻击算法

陈伟, 刘亚洲, 魏松杰*

(南京理工大学计算机科学与工程学院 南京 210094)
(swei@njust.edu.cn)

摘要: 针对当前黑盒环境中, 主流的图像对抗攻击算法在有限的目标模型访问查询次数条件下攻击准确率低的问题, 提出一种基于目标特征和限定区域采样的目标攻击算法. 首先根据原始图像和目标图像生成初始对抗样本; 然后在 Simplex-mean 噪声区域中进行扰动采样, 并根据对抗样本和原始图像差异度以及目标特征区域位置决定扰动大小; 最后将扰动作用于初始对抗样本中, 使新的对抗样本在保持对抗性的同时缩小与原始图像的差异度. 以常见的图像分类模型 InceptionV3 和 VGG16 等为基础, 在相同的目标模型访问查询, 以及与对抗样本和原始图像的 l_2 距离小于 55.89 的条件下, 采用 BBA 等算法对同一图像集和目标集进行攻击. 实验结果表明, 在同样的目标模型访问查询和 $l_2=55.89$ 的限制条件下, 不超过 5000 次目标查询时, 在 InceptionV3 模型上该算法的攻击准确率比同类攻击算法提升至少 50%.

关键词: 目标攻击; 图像分类; 黑盒环境; 区域采样; 对抗样本
中图分类号: TP391.41 **DOI:** 10.3724/SP.J.1089.2021.18365

Adversarial Attack Algorithm on Target Features in Simplex Noise Area

Chen Wei, Liu Yazhou, and Wei Songjie*

(School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing 210094)

Abstract: A target adversarial attack algorithm based on target features and limited area sampling is proposed to improve the low attack accuracy of the current adversarial attack algorithms when only limited target model access queries are allowed in the black-box scenario. Firstly, an initial adversarial example is generated by the original image and the target image. Then the disturbance is sampled in the Simplex-mean noise region and determined by the location of target feature region and the difference between the adversarial example and the original image. The disturbance is used in the initial adversarial example to keep the newly generated one adversarial and to reduce the difference between it and the original image. Based on the common image classification model InceptionV3 and VGG16, under the same target model access query and the l_2 distance between the adversarial example and the original image is less than 55.89. The experimental results using algorithms such as BBA to attack the same image set and target set show that the accuracy of the proposed algorithm is at least 50% higher than that of similar attack algorithms under the same target model access query and $l_2=55.89$, with no more than 5 000 target queries in InceptionV3 model.

Key words: target attack; image classification; black-box scenario; regional sampling; adversarial example

收稿日期: 2020-05-07; 修回日期: 2020-09-29. 基金项目: 国家自然科学基金(61472189, 61802186); 上海航天科技创新基金(SAST2019-033). 陈伟(1996—), 男, 硕士研究生, 主要研究方向为对抗攻击; 刘亚洲(1979—), 男, 博士, 副教授, 主要研究方向为计算机视觉、模式识别以及遥感影像处理; 魏松杰(1977—), 男, 博士, 副教授, CCF 会员, 论文通讯作者, 主要研究方向为网络与信息安全、安全风险评估.

深度神经网络广泛应用于众多数据处理和图像识别中,并取得了显著的效果,如图像分类^[1-4]、语义分割^[5-7]和目标检测。但是,深度神经网络的识别结果很容易受到恶意破坏者的攻击。最明显的是,它们容易受到对抗样本的影响。例如,仅仅在图像中添加人类难以察觉的扰动,就能够使最先进的分类器高概率分错类别或做出不正确的判断。

对抗样本对目标检测等关键应用构成了现实的安全威胁。在这种情况下,攻击者可以通过添加细微的扰动改变图像细节以引起控制故障。现有的大多数攻击算法,都是白盒攻击^[8-10]和黑盒攻击算法^[11-15],这些算法都是通过搜索整个图像空间来搜寻对抗扰动以生成对抗样本。白盒攻击可以由梯度信息引导,查询复杂度较低。但是,黑盒攻击没有这些数据引导,因此攻击更难成功。

在这种环境下,当前主流对抗攻击算法在有限的目标模型访问查询条件下,对抗攻击准确率太低。受边界攻击(biased boundary attack, BBA)^[11]和 Simplex^[16]噪声的启发,本文提出一种 Simplex^[16]噪声区域中目标特征的对抗攻击算法(simplex-mean target feature attack, STFA)。

1 相关工作

虽然深度神经网络在许多研究和应用领域都表现不俗,但是 Szegedy 等^[4]于 2016 年第 1 次发现深度神经网络存在安全性漏洞,其通过在一幅图像上添加微小的扰动,就可以让图像以非常高的置信度被网络分类模型错误分类,甚至可以让图像被分类成指定的类别。2018 年, Ilyas 等^[13]提出一种自然进化策略(natural evolutionary strategies, NES)算法的变体,该算法以 NES 作为黑盒梯度估计技术,并使用带有估计梯度的映射梯度下降(project gradient descent, PGD)算法构造对抗样本,同时将 NES 与基于高斯分布的有限差分法联系起来。该算法可以在黑盒环境中有效地实施目标攻击。2018 年, Cheng 等^[17]提出一种将硬标签黑盒攻击转化为参数优化问题的对抗攻击算法,该算法可以将离散问题连续化,并且可以通过任何零阶优化算法实施对抗攻击,在实验中,该算法攻击效果优于随机优化算法。

2018 年, Brendel 等^[11]提出的边界攻击算法是一种基于决策的对抗攻击算法,它从较大的对抗性扰动开始,力求在保持对抗性的同时减少扰动

以降低图像失真。边界攻击^[11]需要使用已经是对抗的样本进行初始化,并从给定输入有效域的最大熵分布中采样。在目标攻击中,边界攻击^[11]从被分类为目标类别的对抗样本开始,但是在实际操作上,很难从上述分布中采样。因此, Brendel 等^[11]从高斯分布 $\eta_i^k \sim \phi(0,1)$ 中进行采样,并对其进行缩放,使新生成的对抗样本 $\hat{x}_i^k$ 在 $[0,255]$ 范围内;同时新产生的对抗扰动要小于设定的阈值。最后将对抗样本向原始图像方向移动,不断重复上述步骤,从而达到攻击的目的。

但是,原始的边界攻击^[11]算法从高斯噪声中采样,一些尖锐、高频、非自然的噪声中对抗样本所占比例较小,将这部分噪声剔除可以大大减少计算量,从而在有限的目标模型访问查询内获得更高的攻击准确率。

传统的白噪声充满了尖锐的噪声,与自然界的随机现象相差很远。而 Simplex^[16]噪声不仅含有随机的成分,而且相互之间有关联、平滑地进行变化,所以 Simplex^[16]噪声不像白噪声那么尖锐,而是非常类似于自然纹理。Simplex^[16]噪声的晶格为单形,单形就是 N 维空间中最简单紧凑的多边形。单形的顶点数目远少于 N 方体,因此计算梯度噪声时可以大大减少需要计算的顶点权重数目。但是 Simplex^[16]噪声目前主要用于表现自然界中无法用简单形状来表达的物体的形态。本文首次将其引入对抗攻击中。

2 STFA 对抗攻击算法

2.1 对抗攻击环境

鉴于上述应用环境和主流对抗攻击算法的缺点,本文将考虑具有以下参数的对抗攻击环境。

2.1.1 黑盒环境

在黑盒环境中,攻击者对被攻击模型的内部结构、训练参数、防御方法(如果有的话)等一无所知,只能通过输入输出与模型进行交互。而在白盒环境中,攻击者可以访问模型内部结构和参数。相比于白盒攻击,黑盒攻击要困难得多。

2.1.2 有限的目标模型访问查询

如果黑盒攻击需要对模型进行长时间、多频次的海量查询,需要数小时甚至数百万次尝试才能成功,则该攻击是不具有现实应用价值的。本文考虑了这种情况,限定攻击者必须在 5 000 次目标模型访问查询内攻击成功,即超过此查询阈值,则认

为攻击失败.

2.1.3 目标攻击

目标攻击是指当给出原始样本 (x, y) 时, 其中 x 为原始图像, $y \in \gamma$ 为 x 的真实标签, γ 是真实标签数据集; 攻击者旨在诱使分类器将被扰动的对抗样本 x' 分类为特定标记 $t \in \gamma$. 例如, 攻击者可能会通过伪装成金融公司的高度可信客户, 从而攻击该公司的信用评估模型.

2.2 参数定义

考虑一个在 C 个不同类别自然图像集上训练的分类器 K . 对每一幅原始输入图像 x , K 都有一个输出 $K(x) \in \{1, \dots, C\}$. 本文假定图像都被归一化到 $[0, 1]$ 范围内, 定义 $\mathfrak{S} \subset [0, 1]^n$ 代表自然图像输入空间, 假定 K 在自然图像上已经有很高的分类准确率, 即对大多数原始图像 $x \in \mathfrak{S}$, K 可以将其正确分类为标签 c_x , 因此 K 满足 $K(x) = c_x$. 定义 M_K 代表 K 的对抗样本空间, 分布于 M_K 中的图像

必须与原始图像类似, 同时能够使 K 分类错误. 因此, 对于每一个对抗样本 $\hat{x} \in M_K$, 存在 $x \in \mathfrak{S}$ 使 $d(\hat{x}, x)$ 很小, 同时满足 $K(\hat{x}) \neq c_x$; 其中, $d(\cdot, \cdot)$ 是一种衡量距离的尺度. 对于目标攻击, 定义目标类别为 $t (t \neq c_x)$. 定义 T_K 代表 K 的目标对抗样本空间, 分布于 T_K 中的图像 $m_t \in T_K$ 同样需要满足与原始图像类似, 即 $d(m_t, x)$ 很小, 并且使 K 分类错误, 即 $K(m_t) \neq t$.

2.3 算法架构

STFA 对抗攻击算法首先根据原始图像和目标类别获取最相似的目标图像, 其次根据目标图像生成目标特征热图, 然后通过线性搜索寻找分布于 M_K 的对抗样本空间中的初始对抗样本, 接着根据原始图像的平均值, 生成 Simplex-mean 噪声, 然后在 Simplex-mean 噪声中依据目标特征进行插值采样, 不断重复上述过程, 完成对抗攻击. 具体架构如图 1 所示.

图 1 本文算法架构图

2.4 目标特征算法

目前主要的对抗攻击算法 FGSM^[10], JSMA^[18], DeepFool^[9]都是将对扰扰动添加到整个图像空间. 不管是对抗性集中还是稀疏区域, 边界攻击^[11]的正交步骤都会均等地干扰所有像素. 在某些区域,

即使添加少量的扰动, 也会使对抗样本失去对抗性, 即不会被分类器 K 分类成目标 t . 而在其他区域添加大量的噪声, 对抗样本依然既可以保持对抗性, 又能够和原始图像不断逼近. 对特定图像 x , 不同区域对目标对抗攻击的影响差别很大. 而

在基于原始图像 x 和目标图像获得的非目标特征区域中采样形成的对抗样本 $m_i \in T_k$ 拥有最小的扰动. 具体步骤参考算法 1. 这是减少搜索空间的理想方法. 根据对抗样本和原始图像创建一个比例空间

$$Y = |m_i - x| \quad (1)$$

在每一次迭代中, 都会根据当前图像重新计算该比例空间, 并将其逐元素应用于边界攻击^[11]采样的对抗扰动

$$\begin{cases} \mathcal{G}_i^k = Y \odot \mathcal{G}_{i-1}^k \\ \mathcal{G}_i^k = \frac{\mathcal{G}_i^k}{\|\mathcal{G}_i^k\|} \end{cases} \quad (2)$$

其中, $\odot$ 表示 Hadamard 乘积, 即对应位置元素相乘. 每次迭代后, 对抗样本中和原始图像差异明显的区域大幅向原始图像靠拢, 即失真不断衰减. 定义对图像分类影响较大的区域为 r , 其他区域为 r_{other} . 定义比例参数为 p . 在边界攻击^[11]对对抗样本进行插值使其接近原始图像时, 对区域 r_{other} 的修改影响对抗性较小, 即修改大量的区域 r_{other} 仍然能使对抗样本保持对抗性. 本文引导算法在 r_{other} 区域大量采样. 最终的比例空间为

$$p_i^k = [p \odot r_{\text{other}} + (1-p) \odot r] \odot \mathcal{G}_i^k \quad (3)$$

算法 1. STFA 对抗攻击算法.

输入. 原始图像 x , 神经网络分类器 f , 目标类别 t , 最大迭代步数 s_m , 目标图像 x_t , 阈值距离 d_{eps} , 正向区域 r_{other} , 反向区域 r , 比例参数 θ_{ratio} .

输出. 对抗样本 X_{adv} .

function GetAdvImage(x, f, t, s_m)

```
// 获取和原始图像最相似的目标图像
 $x_t \leftarrow \text{findClosestImg}(x)$ ,
// 获取目标图像特征热图
 $r_{\text{other}} \leftarrow \text{getTargetHeatMap}(x_t)$ ,
// 线性搜索, 优化出最接近原始图像的目标
// 图像
 $x_t \leftarrow \text{lineSearchToBoundary}(x, x_t)$ ,
// 设置对抗攻击起始点
 $x_0 \leftarrow x_t$ ,
// 初始化迭代次数
 $i \leftarrow 0$ ,
// 迭代攻击
while  $f(x_t) \neq t$  and  $i < s_m$  do
    // 获取当前图像和原始图像的差异
     $d_{\text{np}} \leftarrow |x_t - x|$ ,
    // 对差异做归一化处理
```

$$d_{\text{np}} \leftarrow \frac{d_{\text{np}}}{\max(d_{\text{np}})},$$

// 根据目标特征获取采样比例

$$d_{\text{rp}} \leftarrow \theta_{\text{ratio}} \times r_{\text{other}} + r \times (1.0 - \theta_{\text{ratio}}),$$

// 归一化采样比例

$$d_{\text{rp}} \leftarrow \frac{d_{\text{rp}}}{\max(d_{\text{rp}})},$$

// 生成综合采样比例

$$d_{\text{np}} \leftarrow d_{\text{np}} \times d_{\text{rp}},$$

// 在当前图像中沿着目标类别方向在

Simplex 噪声中采样

$$r_i \leftarrow g(x_t, d_{\text{np}}),$$

// 更新当前图像的值

$$x_{i+1} \leftarrow x_i + r_i,$$

// 递增迭代次数

$$i \leftarrow i + 1,$$

end while

// 返回攻击后的对抗样本

return x_i

end function

2.5 Simplex-mean 噪声算法

生成参数化低频模式的直接方案是使用 Simplex^[16]噪声. Simplex^[16]噪声最初是用做计算机图形的程序纹理生成器, 创建更加自然合理的外观低频噪声. 但是原始的 Simplex^[16]噪声用于边界攻击^[11]效果不是太理想, 因此本文做出相应的改进. 令 N 为输入空间的维数, 边界攻击^[11](如图 2a 所示)算法通过沿原始图像周围的超球面施加正交摄动 $\mathcal{G}^k$ 来逼近原始图像, 以更深度地进入对抗区域. 在原始边界攻击^[11]中, 可以在样本 $\mathcal{G}_i^k \sim \mathcal{P}(0,1)$ 中生成 $\mathcal{G}^k$ 的候选样本, 样本正交于原始方向投影. 为了在边界攻击^[11]中引入低频先验, 本文从 Simplex^[16]噪声分布中采样(如图 2b 所示). 首先用大小为 256 的排列向量参数化 Simplex^[16]噪声, 即

$$N_{\text{simplex}} = [i \text{ for } i \in (0, 256)] \quad (4)$$

然后生成 12 组以原点为中心的渐变点, 即

$$\begin{cases} g_{\text{grad}} = [g_0 = (0, 1, 1), g_1 = (0, 1, -1), g_2 = (0, -1, 1) \\ g_3 = (0, -1, -1), g_4 = (1, 0, 1), g_5 = (1, 0, 1) \\ g_6 = (-1, 0, 1), g_7 = (-1, 0, -1), g_8 = (1, 1, 0) \\ g_9 = (1, -1, 0), g_{10} = (-1, 1, 0), g_{11} = (-1, -1, 0)] \end{cases} \quad (5)$$

最后计算 8 个方向的梯度并得到每个顶点的噪声贡献度为

$$N = (r^2 - |d|^2)^4 \times \text{dot}(d, g_{\text{grad}}[N_{\text{simplex}}]) \quad (6)$$

其中, d 是输入点到渐变点的距离向量, 由于 $r^2 = 0.5$ 时可以保证没有不连续的间断点, 所以实验中取 $r^2 = 0.5$.

同时为了减少对抗样本和原始图像的差异, 本文将 Simplex^[16]噪声的平均值归一化到原始图像平均值的尺度, 形成 Simplex-mean 噪声模式, 即

$$N = \frac{N \times \text{avg}(\mathbf{O})}{\max(N \times \text{avg}(\mathbf{O}))} \quad (7)$$

由于 Simplex^[16]的低频特性使得样本集中在低频方向上, 这有效地提高了对抗攻击的效率. 因此, 与高斯随机噪声相比, Simplex^[16]噪声低频分布所包含的对抗方向集中度更高, 从而在同样的目标模型访问查询下, 有效地提高了对抗攻击准确率. 图 2a 和图 2b 都可以让分类器分类为指定类别, 但是图 2b 和原始图像的差异更小.

图 2 对抗样本及噪声

3 实验结果与分析

本文提出的 STFA 对抗攻击算法和对比实验均基于 Windows 10 下的 Python 3.6 以及 Tensorflow 1.10, Pytorch 1.1.0, Keras 2.2.0 等深度学习框架实现, 具体硬件参数为 Intel® Core™ i7-8700K CPU, 32 GB 内存. 为了加快攻击速度, 实验中大部分算法选用 NVIDIA GTX1080 GPU 加速完成. 本文根据 ImageNet 数据集上的 InceptionV3^[4], VGG16 等分类器评估上述算法, 对比研究以评估每种算法的有效性. 同时将本文的实验结果与一些最近提出的黑盒对抗攻击算法进行比较.

3.1 数据集

本文的数据集采用 ImageNet 数据集, 该数据集包含 299×299 像素的彩色图像, 具有 1 000 个类别. 实验中使用 Google 已经训练好的 InceptionV3^[4], VGG16 等分类模型作为攻击对象, 其中 Inception-V3^[4]分类模型可实现 78% 的 Top1 准确性, VGG16 分类模型可实现 74% 的 Top1 准确性.

本实验的侧重点是评估在相同的目标模型访问查询, 以及对抗样本和原始图像的 l_2 距离小于 55.89 的条件下, 不同算法的攻击成功率. 所以在实验中, 本文随机选择 100 幅原始图像, 以这 100 幅原始图像为输入数据集, 用 STFA 算法和边界攻击(BBA)^[11]等算法生成对抗本来评估这几种对抗攻击算法的攻击效率.

3.2 评价指标

实验中使用 STFA 和 BBA^[11]等攻击算法对输入数据集的 100 幅图像生成对抗本来做攻击测试. 对抗样本相对于原始图像的扰动低于阈值数量越多, 则攻击算法越有效.

评估攻击算法的距离度量公式为

$$d(x, \hat{x}) = \sqrt{(x - \hat{x})^2} \quad (8)$$

其中, x 表示原始图像; $\hat{x}$ 表示生成的对抗样本. 通过多个志愿者辨别不同 l_2 距离的对抗样本和原始图像, 发现在原始图像和对抗样本的距离等于 55.89 时, 大多数志愿者无法区分出原始图像和对抗样本, 所以本文在对抗扰动的 l_2 范数上定义一个阈值 $l_2 = 55.89$, 扰动小于此阈值才表示攻击成功. 经过反复实验, 不断尝试, 发现在实验中设置最小对抗扰动 $e = 0.5$, 比例参数 $p = 0.658$ 有最优的效果, 所以本文实验也采用上述参数.

3.3 灵敏度分析

灵敏度分析是研究与分析一个模型的状态或输出变化对模型参数变化的敏感程度的方法. 本文将对攻击算法 STFA 与原生的 BBA^[11]算法做了对比实验, 使用每种算法在不同的最大目标模型访问查询次数条件下生成的 100 个对抗样本. 统计出原始图像与对抗样本的 l_2 距离, 并评估其攻击成功率. 表 1 结果显示, 本文研究的每种对抗攻击算法都会提高攻击成功率. 而且这 2 种对抗攻击算法不会互相干扰, 相反它们结合使用会提高攻击成功率. 当 Simplex-mean 噪声和目标特征 2 种算法都使用时, 只需要 5 000 次查询, STFA 算法的攻击成功率在 InceptionV3 分类模型上就能达到 85%, 而在 VGG16 分类模型上能达到 95%, 这也

表 1 在 ImageNet 数据集上对 2 种模型
灵敏度分析(目标攻击)

被攻击模型	是否使用算法		最大目标模型访问查询次数			
	Simplex-mean 噪声	目标特征区域	500	1 000	2 500	5 000
InceptionV3	否	否	0.20	0.27	0.49	0.67
	是	否	0.32	0.46	0.56	0.60
	否	是	0.24	0.38	0.58	0.69
	是	是	0.38	0.57	0.75	0.85
VGG16	否	否	0.33	0.43	0.70	0.84
	是	否	0.66	0.77	0.83	0.89
	否	是	0.35	0.43	0.63	0.85
	是	是	0.51	0.79	0.93	0.95

是本文研究的攻击成功率最高的对抗攻击算法。上述算法在最大目标模型访问查询次数为 5 000 时 l_2 距离大小如图 3 所示, 其中图 3a 为仅使用 Simplex-mean 和原生的 BBA^[11] 算法比较, 图 3b 为仅使用目标特征和原生的 BBA^[11] 算法比较, 图 3c 为使用 Simplex-mean 和目标特征和原生的 BBA^[11] 算法比较, 图 3d 为 3 种算法整体的比较。根据图 3 显示, STFA 对抗攻击算法在同样的目标

模型访问查询下可以有效降低原始图像和对抗样本的 l_2 距离。

图 4 所示为 STFA 算法在 500, 1 000, 2 500 和 5 000 次最大目标模型访问查询原始图像和对抗样本的 l_2 距离结果对比, 从图 4 中可以发现, 较大的最大目标模型访问查询对抗攻击后原始图像和对抗样本的 l_2 距离明显比较小的最大目标模型访问查询小, 从图 4 中还可以看出, 有部分点在较小最大目标模型访问查询次数的条件下, 原始图像和对抗样本的 l_2 距离反而更小, 这是由于插值采样的随机性导致的, 随着迭代次数的增多, 这种现象越来越少。图 4 中 STFA 算法在最大目标模型访问查询次数为 5 000 时所有样本 l_2 距离都比 500 时小说明了这一点。实验结果整体仍然符合最大目标模型访问查询次数越大, 原始图像和对抗样本的 l_2 距离越小的客观规律。

3.4 与其他攻击算法进行比较

本文使用 BAALQI^[13], OPT-ATTACK^[17] 和 BBA^[11] 算法对 STFA 进行基准测试, 各算法攻击成功率结果对比如表 2 所示。STFA 算法能够在同样的目标模型访问查询和 $l_2=55.89$ 的限制条件下,

图 3 最大目标模型访问查询次数为 5 000 时 4 种算法的 l_2 距离

图 4 最大目标模型访问查询次数不同时
本文算法 l_2 距离对比

在不多于 5 000 次目标查询时, 攻击成功率比同类攻击算法在 InceptionV3 模型上至少提升 50%.

BAALQI^[13]是一种梯度估计攻击算法. 该算法的攻击成功率在少数查询的限制条件下大大降低. 在 5 000 个查询内没有产生有效的对抗样本, 即对抗攻击成功率为 0. 而 OPT-ATTACK^[17]算法是将参数设置重构为一个参数优化问题. 在最大目标模型访问查询次数为 5 000 时在 InceptionV3 模型上仅有 30%的攻击成功率, 即使在 VGG16 模

型上也仅有 45%的攻击成功率. 从图 5 中可以看出, STFA 算法相比于其他算法具有收敛速度快的优点. 相对于现有技术而言, 对于查询量非常低的情况, STFA 的攻击成功率要高得多. 基于上述比较, 本文算法优于其他最新攻击算法.

3.5 对抗样本可视化结果分析

本文对比了 STFA 攻击算法和 BBA^[11]等算法生成的对抗样本, 如图 6 所示. 从图 6 可以看出, Simplex-mean 算法生成的对抗样本噪声无论在 InceptionV3 还是 VGG16 模型上都比 BBA^[11]算法生成的对抗样本更加平滑、自然. 而目标特征算法

表 2 4 种算法在 InceptionV3, VGG16 模型下的比较结果

被攻击模型	算法	最大目标模型访问查询次数			
		500	1 000	2 500	5 000
InceptionV3	BAALQI ^[13]	0.00	0.00	0.00	0.00
	OPT-ATTACK ^[17]	0.18	0.19	0.24	0.30
	BBA ^[11]	0.20	0.27	0.49	0.67
	本文	0.38	0.57	0.75	0.85
VGG16	BAALQI ^[13]	0.00	0.00	0.00	0.00
	OPT-ATTACK ^[17]	0.21	0.30	0.38	0.45
	BBA ^[11]	0.33	0.43	0.63	0.85
	本文	0.51	0.79	0.93	0.95

a. 最大目标模型访问查询次数为 500

b. 最大目标模型访问查询次数为 1 000

c. 最大目标模型访问查询次数为 2 500

d. 最大目标模型访问查询次数为 5 000

图 5 最大目标模型访问查询次数不同时 4 种算法的 l_2 距离

生成的对抗样本相比于 BBA^[11]算法生成的对抗样本扰动区域更小. STFA 算法扰动比 BBA^[11]算法只在图中的部分目标特征区域加上了对抗扰动, 图像中其他区域仍然基本与原图保持不变. STFA 算法扰动明显比 OPT-ATTACK^[17]算法扰动更小. OPT-ATTACK^[17]算法在较小最大目标模型访问查

询次数的条件下还保留肉眼可见的目标图像轮廓. 在图 7 中, 其他算法在攻击的同时均对图像整体产生较大的扰动, 而本文算法能够精准地在部分主要区域添加扰动, 同时能够有效地隐藏游艇的轮廓细节, 其他算法则不能较好地去除目标图像的轮廓. 更多对抗攻击的例子如图 8 和图 9 所示.

图 6 不同算法对抗样本及原图展示 1

图 7 不同算法对抗样本及原图展示 2

图 8 不同算法对抗样本及原图展示 3

图9 不同算法对抗样本及原图展示 4

4 结 语

本文提出的 STFA 算法是一种高效的黑盒对抗攻击算法,其引入 Simplex-mean 算法和目标特征对 BBA^[11]算法进行改进. STFA 算法在攻击时首先从噪声较大的原始对抗样本出发,然后高概率地在低频区域和非目标特征区域采样,以缩小对抗样本和原始图像的差异.实验结果表明,基于先验知识可以大大提高黑盒对抗攻击成功率.在 ImageNet 数据集上,使用相同的最大目标模型访问查询次数,STFA 算法在攻击成功率方面优于其他算法.在不多于 5 000 次目标查询时,STFA 算法对抗攻击准确度比同类攻击算法至少提升 50%.只需要 5 000 次查询,STFA 算法的攻击成功率在 InceptionV3 模型上就能达到 85%,在 VGG16 模型上能达到 95%的攻击成功率.这种选择性在图像对抗方向稀疏的区域进行采样的方法,让对抗攻击样本和原始图像快速逼近,从而在较小的最大目标模型访问查询次数条件下具有更好的伪装效果.

本文提出的基于目标特征和限定区域采样的 STFA 算法,有效地提高了 BBA 算法的对抗攻击成功率,进一步揭示出图像分类的安全脆弱性,为进一步建立有效的防御机制奠定了基础.

参考文献(References):

- [1] Krizhevsky A, Sutskever I, Hinton G E. ImageNet classification with deep convolutional neural networks[C] //Proceedings of Advances in Neural Information Processing Systems. Cambridge: MIT Press, 2012: 1097-1105
- [2] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2015: 1-14
- [3] Yang Zhiming, Li Yawei, Yang Bing, *et al.* Cervical cell features based multi-stream convolutional neural networks classification method[J]. Journal of Computer-Aided Design & Computer Graphics, 2019, 31(4): 531-540(in Chinese)
(杨志明, 李亚伟, 杨冰, 等. 融合宫颈细胞领域特征的多流卷积神经网络分类算法[J]. 计算机辅助设计与图形学学报, 2019, 31(4): 531-540)
- [4] Szegedy C, Vanhoucke V, Ioffe S, *et al.* Rethinking the inception architecture for computer vision[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2016: 2818-2826
- [5] Li Yuyan, Zhang Yinhui, He Zifen. Semantic segmentation of tennis scene based on series atrous convolution neural network[J]. Journal of Computer-Aided Design & Computer Graphics, 2020, 32(4): 606-615(in Chinese)
(李逾严, 张印辉, 何自芬. 基于串联空洞卷积神经网络的网球场语义分割[J]. 计算机辅助设计与图形学学报, 2020, 32(4): 606-615)
- [6] Badrinarayanan V, Kendall A, Cipolla R. SegNet: a deep convolutional encoder-decoder architecture for image segmentation[J]. IEEE Transactions on Pattern Analysis And Machine Intelligence, 2017, 39(12): 2481-2495
- [7] Chen L C, Papandreou G, Kokkinos I, *et al.* Semantic image segmentation with deep convolutional nets and fully connected CRFs[OL]. [2020-05-07]. <https://arxiv.org/abs/1412.7062>
- [8] Carlini N, Wagner D. Towards evaluating the robustness of neural networks[C] //Proceedings of the IEEE Symposium on Security and Privacy. Los Alamitos: IEEE Computer Society Press, 2017: 39-57
- [9] 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. Los Alamitos: IEEE Computer Society Press, 2016: 2574-2582
- [10] Goodfellow I J, Shlens J, Szegedy C. Explaining and harnessing adversarial examples[C] //Proceedings of International Conference on Learning Representations. Los Alamitos: IEEE Computer Society Press, 2015: 1050-1070

- [11] Brendel W, Rauber J, Bethge M. Decision-based adversarial attacks: reliable attacks against black-box machine learning models[C] //Proceedings of International Conference on Learning Representations. Los Alamitos: IEEE Computer Society Press, 2018: 1055-1066
- [12] 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. New York: ACM Press, 2017: 15-26
- [13] Ilyas A, Engstrom L, Athalye A, *et al.* Black-box adversarial attacks with limited queries and information[C] //Proceedings of International Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2018: 2137-2146
- [14] Ilyas A, Engstrom L, Madry A. Prior convictions: black-box adversarial attacks with bandits and priors[C] //Proceedings of International Conference on Learning Representations. Los Alamitos: IEEE Computer Society Press, 2019: 1206-1230
- [15] Tu C C, Ting P S, Chen P Y, *et al.* AutoZOOM: autoencoder-based zeroth order optimization method for attacking black-box neural networks[C] //Proceedings of the Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2019, 33: 742-749
- [16] Hart J C. Perlin noise pixel shaders[C] //Proceedings of the ACM SIGGRAPH/Eurographics Workshop on Graphics Hardware. New York: ACM Press, 2001: 87-94
- [17] Cheng M H, Le T, Chen P Y, *et al.* Query-efficient hard-label black-box attack: an optimization-based approach[OL]. [2020-05-07]. <https://arxiv.org/abs/1807.04457>
- [18] Papernot N, McDaniel P, Jha S, *et al.* The limitations of deep learning in adversarial settings[C] //Proceedings of the IEEE European Symposium on Security and Privacy. Los Alamitos: IEEE Computer Society Press, 2016: 372-387