自动化学报
ACTA AUTOMATICA SINICA

针对身份证文本识别的黑盒攻击算法研究

徐昌凯 冯卫栋 张淳杰 郑晓龙 张辉 王飞跃

Research on Black-box Attack Algorithm by Targeting ID Card Text Recognition

XU Chang-Kai, FENG Wei-Dong, ZHANG Chun-Jie, ZHENG Xiao-Long, ZHANG Hui, WANG Fei-Yue

在线阅读 View online: https://doi.org/10.16383/j.aas.c230344

针对身份证文本识别的黑盒攻击算法研究

徐昌凯[1, 2] 冯卫栋[1, 2] 张淳杰[1, 2] 郑晓龙[3, 4, 5] 张 辉[6] 王飞跃[3, 4, 5]

摘 要 身份证认证场景多采用文本识别模型对身份证图片的字段进行提取、识别和身份认证, 存在很大的隐私泄露隐患. 并且, 当前基于文本识别模型的对抗攻击算法大多只考虑简单背景的数据 (如印刷体) 和白盒条件, 很难在物理世界达到理想的攻击效果, 不适用于复杂背景、数据及黑盒条件. 为缓解上述问题, 本文提出针对身份证文本识别模型的黑盒攻击算法, 考虑较为复杂的图像背景、更严苛的黑盒条件以及物理世界的攻击效果. 本算法在基于迁移的黑盒攻击算法的基础上引入二值化掩码和空间变换, 在保证攻击成功率的前提下提升了对抗样本的视觉效果和物理世界中的鲁棒性. 通过探索不同范数限制下基于迁移的黑盒攻击算法的性能上限和关键超参数的影响, 本算法在百度身份证识别模型上实现了 100% 的攻击成功率. 身份证数据集后续将开源.

关键词 对抗样本, 黑盒攻击, 身份证文本识别, 物理世界, 二值化掩码

引用格式 徐昌凯, 冯卫栋, 张淳杰, 郑晓龙, 张辉, 王飞跃. 针对身份证文本识别的黑盒攻击算法研究. 自动化学报, 2024, **50**(1): 103−120

DOI 10.16383/j.aas.c230344

Research on Black-box Attack Algorithm by Targeting ID Card Text Recognition

XU Chang-Kai[1, 2] FENG Wei-Dong[1, 2] ZHANG Chun-Jie[1, 2] ZHENG Xiao-Long[3, 4, 5]
ZHANG Hui[6] WANG Fei-Yue[3, 4, 5]

Abstract Identity card authentication scenarios often use text recognition models to extract, recognize, and authenticate ID card images, which poses a significant privacy breach risk. Besides, most of current adversarial attack algorithms for text recognition models only consider simple background data (such as print) and white-box conditions, making it difficult to achieve ideal attack effects in the physical world, and is not suitable for complex backgrounds, data, and black-box conditions. In order to alleviate the above problems, this paper proposes a black-box attack algorithm for the ID card text recognition model by taking into account the more complex image background, more stringent black-box conditions and attack effects in the physical world. By using the transfer-based black-box attack algorithm, the proposed algorithm introduces binarization mask and space transformation, which improves the visual effect of adversarial examples and the robustness in the physical world while ensuring the attack success rate. By exploring the performance upper limit and the influence of key hyper-parameters of the transfer-based black-box attack algorithm under different norm constraints, the proposed algorithm achieves 100% attack success rate on the Baidu ID card recognition model. The ID card dataset will be made publicly available in the future.

Key words Adversarial examples, black-box attack, ID card text recognition, physical world, binarization mask

Citation Xu Chang-Kai, Feng Wei-Dong, Zhang Chun-Jie, Zheng Xiao-Long, Zhang Hui, Wang Fei-Yue. Research on black-box attack algorithm by targeting ID card text recognition. *Acta Automatica Sinica*, 2024, **50**(1): 103−120

收稿日期 2023-06-08 录用日期 2023-10-16
Manuscript received June 8, 2023; accepted October 16, 2023
科技创新 2030 —— "新一代人工智能"重大项目 (2020AAA0108401), 北京市自然科学基金 (JQ20022), 国家自然科学基金 (62072026, 72225011), 中国人工智能学会 —— 昇腾 CANN 学术基金, OpenI 启智社区资助
Supported by Scientific and Technological Innovation 2030 - "New Generation Artificial Intelligence" of the Ministry of Science and Technology of China (2020AAA0108401), Beijing Natural Science Foundation (JQ20022), National Natural Science Foundation of China (62072026, 72225011), CAAI-CANN Open Foundation, and OpenI Community
本文责任编委 白翔
Recommended by Associate Editor BAI Xiang

1. 北京交通大学计算机与信息技术学院信息科学研究所 北京 100044 2. 现代信息科学与网络技术北京市重点实验室 北京 100044 3. 中国科学院自动化研究所多模态人工智能系统全国重点实验室 北京 100190 4. 中国科学院自动化研究所复杂系统管理与控制国家重点实验室 北京 100190 5. 中国科学院大学人工智能学院 北京 100049 6. 北京航空航天大学交通科学与工程学院 北京 100191

1. The Institute of Information Science, School of Computer and Information Technology, Beijing Jiaotong University, Beijing 100044 2. Beijing Key Laboratory of Advanced Information Science and Network Technology, Beijing 100044 3. State Key Laboratory of Multimodal Artificial Intelligence Systems, Institute of Automation, Chinese Academy of Sciences, Beijing 100190 4. State Key Laboratory for Management and Control of Complex Systems, Institute of Automation, Chinese Academy of Sciences, Beijing 100190 5. School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing 100049 6. School of Transportation Science and Engineering, Beihang University, Beijing 100191

近年来, 以深度学习为核心的人工智能技术得到了快速的发展, 广泛应用到了刷脸支付, 自动驾驶, 内容生成等多种生产生活场景. 自从 2012 年 Krizhevsky 等[1] 利用深度学习模型 AlexNet 赢得 ImageNet 竞赛冠军后, 对于深度学习模型的研究开始呈现井喷式增长, 在计算机视觉[2], 语音识别[3], 自然语言处理[4], 强化学习[5] 等领域不断刷新着性能表现. 如第一个击败人类职业选手的围棋高手 AlphaGo[6], 观测水平可以和人类利用冷冻电镜相媲美的蛋白质结构预测模型 AlphaFold[7], 以及当下非常火爆的生成式智能聊天机器人 ChatGPT[8]. 海量数据和强大算力支撑了深度学习模型在人工智能技术中的核心地位, 并深刻影响着人工智能技术的发展. 随着人工智能技术的发展, 各种检测和识别模型引起了人们对隐私问题的担忧. 各种包含隐私信息的图片在传输过程中很容易被其他人获取, 并借助深度学习模型迅速获取其中的隐私信息, 存在很大的隐私泄露隐患.

深度学习模型在不同领域和任务上取得了优异表现, 但是其在对抗场景下的脆弱性同样值得关注. 许多研究表明[9−14], 深度学习模型容易受到对抗样本的干扰, 从而给出错误的预测结果. 对抗样本是由攻击者在正常样本的基础上, 通过添加人类难以发现的细微对抗噪声所恶意构造的样本数据. 以图像分类模型为例, 如图 1 所示, 在正常样本上表现良好的模型, 在对抗样本上却给出了错误的预测结果. 对抗样本广泛存在于图像、文本、语音等不同模态, 且针对各个模态的具体模型, 如目标检测, 文本识别, 语音识别等都表现出较强的攻击性能. 不仅如此, 数字世界中的对抗样本还可以通过不同的方式 (如印刷、喷绘、3D 打印等) 转换到物理世界[15−17], 并且对环境因素的干扰具有一定的鲁棒性.

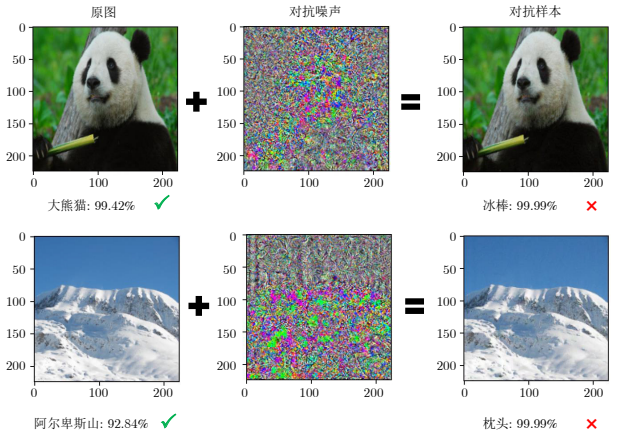


图 1 针对图像分类模型的对抗样本[14] (左: 正常样本, 中: 对抗噪声, 右: 对抗样本)

Fig. 1 Adversarial examples for the image classification model[14] (Left: normal samples, Middle: Adversarial noise, Right: Adversarial examples)

对抗样本是在特定的攻击条件下生成的, 目前大多数工作将其分为白盒条件和黑盒条件. 两者的区别在于攻击者掌握被攻击对象 (模型) 信息的多少, 白盒条件下攻击者可以获取被攻击对象的全部资料, 黑盒条件下只能获取部分资料. 黑盒条件较白盒条件的攻击难度更高, 且更贴近实际的攻击场景. 针对目标检测 (2D/3D), 文本识别, 语音识别等模型的对抗攻击算法已经涌现出了许多研究工作, 大多数工作都可以在白盒条件下达到较高的攻击成功率 (接近 100%), 但在黑盒条件下攻击成功率依然较低. 除此之外, 目前大多数工作集中于单模态的模型架构, 缺少对跨模态模型的探索. 已有的针对跨模态模型 (主要是光学字符识别模型) 的对抗攻击算法大多只考虑了简单背景 (如印刷体) 的数据集和白盒条件的攻击设定, 不适用于复杂背景的图像数据和更为严苛的黑盒攻击条件. 此外, 大多数工作在生成物理世界中的对抗样本时, 攻击设定都是白盒条件; 很少有工作同时考虑了物理世界和黑盒条件的攻击设定. 物理世界和黑盒条件的组合是一种攻击难度极高的攻击设定, 非常具有挑战性, 却也最贴近实际场景, 具有较高的理论与应用价值.

针对上述问题, 本文对身份证隐私保护方法进行研究, 提出了一种针对身份证文本识别模型的黑盒对抗攻击算法. 身份证文本识别模型是由图像到文本的跨模态模型, 同时涉及到图像特征提取和文本序列建模. 与已有的针对文本识别模型的研究工作相比: 1) 本文同时考虑黑盒条件和物理世界攻击设定, 并且在商用身份证识别模型 (百度身份证识别模型) 上达到 100% 攻击成功率; 2) 针对黑盒条件下对抗样本攻击成功率较低的问题, 本文探索了对抗样本在不同范数限制下的攻击能力上限, 并在保证攻击成功率的前提下使用二值化掩码来改善对抗样本的视觉效果; 3) 针对物理世界中环境因素干扰的问题, 本文在生成对抗样本的过程中引入了空间变换, 使得方法对于拍摄距离、光照等环境因素的干扰具有一定的鲁棒性.

1 相关工作

本节将介绍相关研究工作. 以图像分类模型为例, 阐述对抗样本的定义, 概述黑盒条件下的对抗攻击算法, 并梳理总结现有针对文本识别模型的对抗攻击算法.

1.1 黑盒攻击算法

本节首先给出对抗样本的定义. 令 $f(x;\theta): x \in [0,1]^d \longrightarrow z \in \mathbf{R}^n$ 表示一个图像分类模型, 类别数为 n, 参数为 θ. 其中, x 表示一张维度为 d 的图像,

取值范围为 0 到 1, z 表示一个 n 维的概率矩阵. 假设 $f(x;\theta)$ 可以正确识别输入图像 x, 预测结果为图像 x 对应的标签 y. 则对抗样本 x_{adv} 可以定义为:

$$x_{adv} = x + \delta$$
$$\text{s.t. } f(x_{adv};\theta) \neq y,\ \ x_{adv} \in T(x,\epsilon) \tag{1}$$

$$T(x,\epsilon) = \left\{ x + \delta \in [0,1]^d \,|\, \|\delta\|_p \leq \epsilon \right\} \tag{2}$$

式 (1) 中, δ 表示对抗噪声. 在原始图像 x 上添加 δ 之后, 就生成了对抗样本 x_{adv}, 其必须满足两个限制条件: x_{adv} 必须导致分类器 f 输出错误的预测结果; x_{adv} 必须满足一定的数值范围 $T(x,\epsilon)$, 如式 (2) 所示. 首先限制了 x_{adv} 的取值范围为 0 到 1, 保证对抗样本和原始图像的数值范围相同; 其次规定了对抗噪声 δ 的 l_p 范数小于等于范数限制 ϵ, 目的是保证对抗噪声的幅值较小, 使得人类难以发现, 且不影响人类对于对抗样本的判断结果. 在目前大多数研究工作中, p 有三种取值, 分别为 0, 2 和无穷.

黑盒条件下, 攻击者掌握被攻击对象的资料信息有限, 只能与被攻击对象做特定形式的交互. 例如, 攻击者想要攻击某网站的图像识别模型, 攻击者无法获取模型结构, 参数等信息, 而只能上传图像或通过指定的程序接口 (Application programming interface, API) 调用其服务. 此时, 攻击者能得到的信息只有模型返回的识别结果. 黑盒条件是最不利于攻击者的条件, 但也是最贴近现实的攻击条件. 在黑盒条件下实现较高的攻击成功率, 其难度和挑战性远远大于白盒条件.

黑盒条件下的对抗攻击算法主要分为三类, 基于分数的对抗攻击算法, 基于决策的对抗攻击算法和基于迁移的对抗攻击算法, 分类的依据是被攻击对象输出信息的多少. 假设被攻击对象是图像分类模型, 其输出信息包含以下三种情况: 1) 分类结果和每一类的概率 (置信度); 2) 只有分类结果; 3) 只有分类结果, 且攻击者和被攻击对象的交互次数会受到客观限制, 例如付费识别时的经济成本. 这三种情况中, 攻击者可以获取到被攻击对象的信息逐渐减少, 攻击难度逐渐增加. 其中, 基于分数的算法适用于第一种情况, 代表性算法有 Zeroth order optimization (ZOO)[18], Natural evolution strategies gradient estimate (NES)[19], Simultaneous perturbation stochastic approximation (SPSA)[20] 和 Natattack[21]; 基于决策的算法适用于第二种情况, 代表性算法有 Boundary[22], Evolutionary[23] 和 Geometric decision-based attack (GeoDA)[24]; 基于迁移的算法适用于第三种情况, 代表性方法有 Momentum iterative fast gradient sign method (MI-FGSM)[25], Nesterov iterative fast gradient sign method (NI-FGSM)[26], Diverse inputs iterative fast gradient sign method (DI-FGSM)[27], Combination of translation-invariant method and the fast gradient sign method (TI-FGSM)[28], Variance tuning MI-FGSM (VMI-FGSM)[29].

这几种典型的基于迁移的黑盒攻击算法可以大致分为两类: 修改梯度更新过程, 增加输入变化来提升对抗样本的泛化性. 其中, MI-FGSM, NI-FGSM, VMI-FGSM 和 Variance tuning NI-FGSM (VNI-FGSM) 属于第一种. MI-FGSM 引入了动量项, 修改了梯度的累积过程; NI-FGSM 利用 Nesterov 来加速梯度的计算过程; VMI-FGSM 和 VNI-FGSM 分别在 MI-FGSM 和 NI-FGSM 的基础上, 通过采样一定邻域内样本的平均梯度, 对动量项做修正, 更改了动量的累积过程. Scale-invariant fast gradient sign method (SI-FGSM), TI-FGSM 和 DI-FGSM 属于第二种. SI-FGSM 增加了输入样本的幅值变化, 创建了 i 个输入样本的副本, 每个副本的幅值在原始幅值上缩小 2^i 倍; TI-FGSM 利用卷积神经网络的平移不变性, 引入了符合高斯分布的卷积核, 通过和梯度做卷积计算来增强梯度的泛化性; DI-FGSM 则通过输入空间变换, 即随机缩放和填充, 来增强输入样本的泛化性. 这几种典型的基于迁移的黑盒攻击算法都是在基于梯度的对抗攻击算法 I-FGSM 的基础上演化而来的, 并且可以通过超参数的调整互相转换, 例如, 当 VMI-FGSM 中的采样邻域范围 β 变为 0 时, VMI-FGSM 就转换为了 MI-FGSM; 当 MI-FGSM 中的动量累积项系数 μ 变为 0 时, MI-FGSM 就变为 I-FGSM. 同时, 这几种算法也可以加入一系列的输入变换方法, 达到更好的攻击性能. 这几种典型的对抗攻击算法关系图如图 2 所示. 其中, VM(N)I-FGSM 分别表示 VMI-FGSM 和 VNI-FGSM, D(T, S)I-FGSM 分别表示 DI-FGSM, TI-FGSM 和 SI-FGSM.

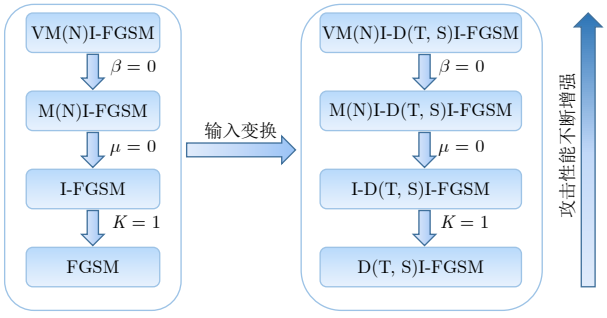


图 2　几种典型的黑盒迁移攻击算法关系图[29]

Fig. 2　Relationships among several transfer-based black-box adversarial attacks[29]

基于分数和基于决策的黑盒攻击算法虽然可以保证成功攻击目标模型, 但是其需要大量访问被攻

击对象的输出结果 (这一过程也称之为查询), 导致攻击成本不可控. 目前, 这两个方向的研究工作集中于在保证攻击性能的前提下, 提升查询效率, 降低攻击成本. 基于迁移的黑盒攻击算法攻击成本较低, 但之前的研究工作大多都假设替代模型和目标模型的训练集相同, 导致可能存在数据泄露的问题. 而且这样的设定与实际情况相悖, 导致其在实际应用中需要较强的对抗噪声才可以实现较为满意的攻击效果. 目前, 这一方向的研究集中于更加贴近实际情况的攻击设定和提升对抗样本的视觉效果.

1.2 针对文本识别模型的对抗攻击算法

文本识别模型属于光学字符识别 (Optical character recognition, OCR) 模型中的一个组件, 是由图像到文本的跨模态模型. OCR 模型分为两阶段和端到端两种设计架构. 两阶段的 OCR 模型包括文本检测和文本识别, 具体来说包括四个步骤: 文本检测, 检测框对齐, 方向矫正和文本识别. 其中, 文本检测用来检测图像中存在的文本位置; 检测框对齐将倾斜或者不规则排列的文本位置变为水平位置; 方向矫正可以保证字体的角度是 0 度, 防止出现字体颠倒的情况. 在保证字体位置水平且字体角度为 0 度之后, 将其送入文本识别模型进行识别. 端到端的 OCR 模型在一个网络中同时进行文本检测和文本识别, 相比两阶段的 OCR, 其模型更小, 速度更快, 但性能也会出现下降. 相比于端到端的 OCR 模型, 两阶段的 OCR 模型更加成熟, 在实际场景中应用更加广泛.

两阶段的 OCR 模型中, 基于 CRNN + CTC 架构[30] 的文本识别模型速度快, 准确率较高, 且由于 CTC 独特的解码方式, 这种模型架构可以识别任意长度的文字序列, 不会受到不定长文字序列的影响, 是一种主流的文本识别模型架构. 如图 3 所示, 模型架构主要包含三部分, 图像特征提取模块 CNN, 图像上下文信息提取模块 BLSTM 和 CTC 损失. 其中, CNN 一般使用主流的卷积神经网络, 如 ResNet (桌面端), MobileNet (移动端) 等. 文本图像中存在大量的上下文信息, 尤其是较长的文字序列. 由于卷积神经网络更关注于图像局部特征的建模, 缺乏全局特征和上下文建模的能力, 只使用卷积神经网络很难有效挖掘到文本图像中的上下文信息. 因此, CRNN 引入了双向长短时记忆网络 (Bidirectional long short-term memory, BLSTM), 不仅提供了上下文信息建模的能力, 还解决了传统循环神经网络中存在的长距离依赖问题. 得到 BLSTM 输出的特征序列后, 由于训练数据中文本图像的标签不包含每个字符在图像中的位置, 无法

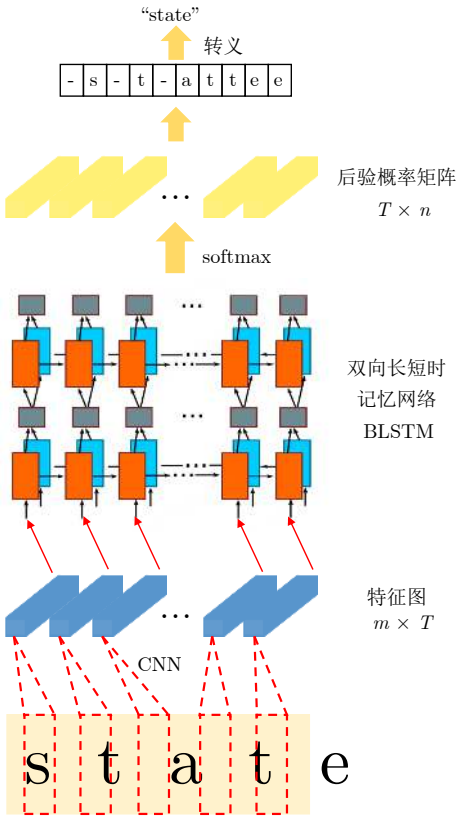


图 3 CRNN 文本识别模型架构图[30] (CNN 表示卷积神经网络. "-"表示空字符)

Fig. 3 CRNN text recognition model architecture diagram[30] (CNN stands for convolutional neural network. "-" indicates a null character)

将对应时刻的输出和 CNN 的感受野进行对齐, 不能使用常见的交叉熵损失函数训练模型. 为了解决这一问题, 引入了 CTC 损失, CTC 是一种不需要对齐的编码方法, 广泛应用于语音识别等领域[31].

近年来, 针对文本识别模型的对抗攻击算法涌现出了许多工作. 我们简要介绍几篇代表性的研究工作, 并总结分析已有工作与本工作的区别.

工作 1[32] 最先提出了针对文本识别模型的对抗攻击算法. 该算法采用基于优化的白盒攻击方法, 其损失函数采用了与 CW 算法相同的形式, 目的是在保证攻击成功率的前提下最小化对抗噪声, 使得生成的对抗样本较为隐蔽. 该工作探索了不同字体和字号对攻击成功率的影响, 并将对抗样本嵌入到文本段落中, 使其在其他下游任务 (如邮件识别和影评等) 展现攻击效果.

工作 2[33] 采用基于梯度的白盒攻击算法和全局性水印, 通过计算梯度的权重确定水印的位置, 再将对抗噪声隐藏到水印中, 可以在保证攻击性能的前提下进一步提升对抗样本的隐蔽性. 该工作探索了单词级别和字符级别对抗样本的攻击性能, 并探

索了不同的水印形式对攻击成功率的影响. 值得关注的是, 工作 1 和工作 2 都只考虑了白色背景的图像数据和白盒条件, 且没有针对商用模型和物理世界进行试验. 这样的设定导致其攻击难度大大降低, 不利于算法在复杂背景数据和黑盒条件下的泛化.

工作 3[34] 提出了一种针对场景文本识别模型的对抗攻击算法, 可以自然地应用到基于 CTC 和基于注意力机制这两种不同的场景文本识别模型架构中. 与之前的工作相比, 该工作进一步考虑了背景更加复杂的场景文本图像, 而不是简单的白色背景图像. 但该算法只考虑了白盒条件, 没有考虑攻击难度更高的黑盒条件. 实验证明, 该算法在白盒条件下可以达到非常高的攻击成功率, 但在黑盒条件下其攻击成功率较低, 只有 20% 左右.

工作 4[35] 在工作 3 的基础上, 考虑了黑盒条件, 提出了一种针对场景文本识别模型的黑盒攻击算法, 旨在更改更少的扰动像素个数 (即 l_p 范数限制) 下提高对抗样本的攻击成功率. 为了确定被操纵像素的位置和值, 作者提供了一个有效的自适应−离散微分进化框架, 将连续搜索空间缩小到一个离散空间, 有效地提高了查询效率, 在本地的黑盒模型上实现了较高的攻击成功率. 但由于其需要黑盒模型的部分先验知识, 即需要输出每个字符和预测概率, 导致其在攻击商用模型时受限, 在百度的场景文本识别模型上只实现了 40% 左右的攻击成功率.

工作 5[36] 利用了对抗样本的攻击性和视觉友好性, 将其转换为鲁棒性的验证码, 用于攻击验证码破解系统. 具体地, 该算法针对验证码破解系统中的不同步骤, 设计了不同攻击方法组合的形式, 包括多目标攻击, 集成对抗训练, 可微逼近和期望顺序组合. 与之前的工作不同, 该工作在系统层面设计了对抗样本的生成算法, 而不仅仅在模型层面.

工作 6[37] 将对抗样本和图像隐写结合起来, 提出了一种基于对抗攻击的字符级隐写算法. 具体地, 该工作通过攻击局部模型来实现隐写消息嵌入, 通过捕获发送方和接收方局部模型识别的结果与参考模型不同的字符来提取信息. 该工作的重点在于图像隐写策略的创新性设计, 规定了发送方和接收方必须使用相同的 OCR 模型这一应用场景, 而不在于提高对抗样本的攻击成功率.

工作 7[38] 第一次将对抗攻击引入场景文本的识别, 研究其对基于 attention 的场景文本识别模型的影响. 具体地, 该工作通过将针对图像分类、语义分割和图像检索等非序列任务设计的目标函数调整为序列形式, 然后, 提出了一种新颖且有效的目标函数, 以进一步减少扰动量, 同时实现更高的攻击成功率.

工作 8[39] 提出了一种水印攻击方法, 利用水印的伪装来产生自然失真, 从而避开人眼的检测. 具体地, 此工作重点研究白盒, 有针对性的攻击, 生成自然的水印样式扰动. 水印不会妨碍文本的可读性, 使得对抗样本看起来更自然, 并获得与现有攻击方法相似的攻击性能.

工作 9[14, 40] 提出了一种高效的基于优化的对抗攻击方法. 具体地, 此工作对抗攻击目标定义为找到一个与普通样本非常相似 (人眼几乎不能发现差异) 但识别结果却差异很大的样本. 此工作通过优化的方式生成对抗样本, 并在一些主流的场景文本识别模型上达到了较高的攻击成功率.

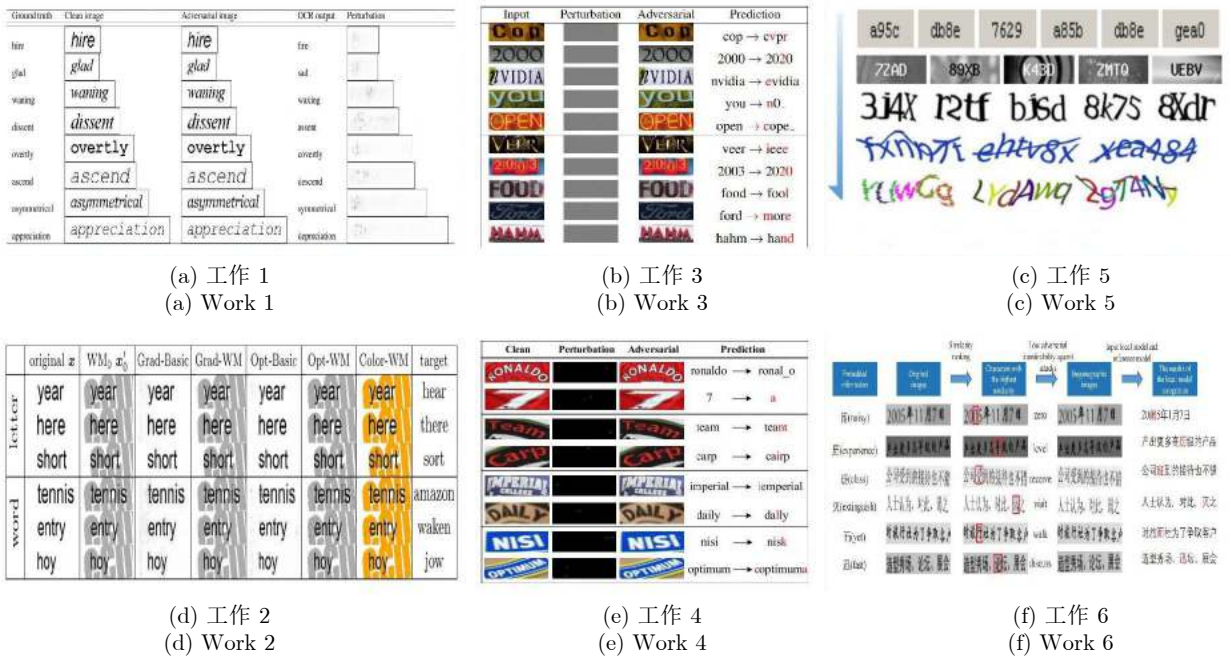
当前工作集中于一般类型的文本图像, 并没有针对具体 OCR 任务做优化, 且缺乏物理世界中的相关实验. 工作 4 考虑到了黑盒条件和商用模型, 但由于其需要部分黑盒模型的先验知识 (即每个字符的识别概率), 当被查询的商用模型不满足输出先验知识的要求或查询成本较高时, 其在实际场景中的应用就会受到限制, 泛化性较差. 物理世界和黑盒条件的组合攻击难度较高, 极具挑战性, 却也最贴近实际场景, 具有极高的应用价值. 本文在图 4 中展示了当前工作生成的对抗样本示例, 并且在表 1 中总结了当前工作和本工作的区别.

2 针对身份证识别模型的黑盒攻击算法

本节将详细介绍针对身份证识别模型的黑盒攻击算法. 首先, 在基于迁移的黑盒攻击算法框架下, 给出了问题定义; 其次, 同时考虑黑盒条件和物理世界攻击, 在算法中分别设计了二值化掩码和空间变换两个组件来提升对抗样本视觉效果和物理世界的鲁棒性; 最后, 通过在本地身份证模型和百度身份证模型上的攻击实验证明了本算法的有效性.

2.1 问题定义

本文将两阶段 OCR 模型中的文本识别模型作为攻击对象, 原因如下: 1) 文本检测模型属于目标检测模型的一个子类, 涉及的模态类型只有图像. 而文本识别模型需要同时处理图像和文字序列, 更加符合对于跨模态模型对抗攻击算法的探索. 2) 目前针对目标检测模型的对抗攻击算法需要较强的局部扰动 (常以对抗补丁的形式存在[41−43]) 才能达到较高的攻击成功率. 这种较强的局部扰动适用于图像中面积较大的一般物体, 但不适用于文字等结构性较强的对象. 例如, 在图像中某人的胸前生成扰动性很强的图案, 并不影响人类判断他是一个人; 但是在某个字的局部生成扰动性很强的图案, 就会干扰到人类对其的判断, 这与对抗样本的定义相违背. 3) 相



(a) 工作 1
(a) Work 1

(b) 工作 3
(b) Work 3

(c) 工作 5
(c) Work 5

(d) 工作 2
(d) Work 2

(e) 工作 4
(e) Work 4

(f) 工作 6
(f) Work 6

图 4 当前工作生成的对抗样本示例图[32–37]

Fig. 4 Adversarial examples generated by recent work[32–37]

表 1 本工作与已有工作的对比

Table 1 Comparison of this work with existing work

工作	攻击条件	图像数据	攻击方式	商用模型	物理世界
1	白盒	白色背景	基于优化	×	×
2	白盒	白色背景	基于梯度 + 水印	×	×
3	白盒	彩色背景	基于梯度 + 优化	✓	×
4	黑盒	彩色背景	基于查询	✓	×
5	白盒	白色背景	多方式集成	×	×
6	白盒	灰色背景	基于梯度	×	×
7	白盒	彩色背景	基于优化	×	×
8	白盒	灰色背景	基于生成	×	×
9	白盒	灰色背景	基于优化	✓	×
本文	黑盒	彩色背景	基于迁移 + 二值化掩码	✓	✓

比文本检测模型, 文本识别模型的架构更加统一, 在不同场景下的通用性更强. 因此, 针对文本识别模型的对抗攻击算法在不同的 OCR 模型间的泛化性更好.

下面给出问题定义: 假设有两个身份证文本识别模型 $f_1(\theta_1)$ 和 $f_2(\theta_2)$, 其参数分别为 θ_1 和 θ_2. 其中, f_1 的结构和参数 θ_1 是已知的, f_2 的结构和参数 θ_2 是未知的, 分别代表替代模型与目标模型. 此外, 攻击者只能与 $f_2(\theta_2)$ 做有限制次数的交互, 得到 $f_2(\theta_2)$ 的输出结果, 且结果中不包含每个字符识别的概率值. 针对输入样本 x 和对应的标签 y, 两个识别模型均可以将其正确识别. 本文目标是设计一种攻击算法, 求解对抗噪声 δ, 生成对抗样本 x_{adv}, 使得 $f_1(\theta_1)$ 和 $f_2(\theta_2)$ 均识别出错, 且限制对抗噪声的范数在一定范围之内, 如式 (3) 所示.

$$\begin{aligned} &x_{adv} = x + \delta \\ &\text{s.t. } f_1(x_{adv}) \neq y,\ f_2(x_{adv}) \neq y,\ x_{adv} \in T(x, \epsilon) \\ &T(x, \epsilon) = \left\{ x + \delta \in [0, 1]^d \,|\, \|\delta\|_p \leq \epsilon \right\} \end{aligned} \tag{3}$$

2.2 算法设计

2.2.1 二值化掩码

由于目标模型 $f_2(\theta_2)$ 的参数和结构都是未知的, 且攻击者只能与 $f_2(\theta_2)$ 做有限制次数的交互, 得到不包含概率值的识别结果, 这样的设定适用于基于迁移的黑盒攻击算法. 即先训练一个替代模型

$f_1(\theta_1)$, 使用白盒攻击算法生成针对 $f_1(\theta_1)$ 的对抗样本, 再使用基于迁移的黑盒攻击算法将其攻击性迁移到未知的 $f_2(\theta_2)$.

对抗样本的特点是在不影响人类判断的前提下, 干扰目标模型的判断. 上述基于迁移的黑盒攻击算法在生成对抗样本时, 大多基于 l_∞ 范数限制, 即采用全局的对抗噪声扰动, 这有利于提升对抗样本的攻击性能. 同时, 为了保证对抗样本的隐蔽性, 范数限制 ϵ 通常较小, 不会干扰到人类的判断. 针对身份证图像, 如果采用全局的对抗噪声, 那么关键字段上也会叠加噪声, 干扰人类的判断. 特别是当增大范数限制之后, 覆盖在关键字段上的噪声会更加明显; 而如果范数限制较小, 就会降低对抗样本的攻击性能. 如何平衡对抗样本的攻击性能和视觉效果, 特别是在保证攻击性能的前提下提升视觉效果, 是本文重点考虑的问题之一.

为了缓解上述问题, 更好地平衡对抗样本的攻击性能和视觉效果, 本文引入了二值化掩码 M, 并将其加入到对抗样本的生成过程中, 参与梯度的迭代计算. 具体地, 当输入样本 x 的维度为 (h, w, c) 时, 定义二值化掩码 M 如下:

$$M \in \{1, 0, 0, \cdots, 0, 1, 0\}^p, \ p = (h, w, c) \tag{4}$$

其中, (h, w, c) 表示输入样本的 (长, 宽, 通道数). 二值化掩码 M 的维度和输入样本的维度相同, 只包含 0 和 1, 其作用是分割身份证图像的前景和背景. 其中, 前景为关键字段对应的图像区域, 如姓名、性别和身份证号等, 背景是除了前景文字以外的全部区域. 由于身份证的前景和背景区域差异较为明显, 分割难度并不大, 二值化掩码的计算非常便捷, 可以通过自适应或者传统的阈值分割方法计算得到, 如自适应阈值分割算法或大津法等.

在得到二值化掩码 M 后, 将其加入到对抗样本的梯度计算过程中, 通过点乘操作“$\odot$”就可以将对抗噪声的覆盖区域限制在背景上, 避免全局的对抗噪声对关键字段的覆盖. 以 MI-FGSM 为例, 加入二值化掩码 M 后, 对抗样本 x_{adv} 的计算过程如下:

$$\begin{aligned} x_t^{adv} &= x_t^{adv} \odot M \\ g_{t+1} &= \mu \cdot g_t + \frac{\nabla_x \mathcal{L}\left(x_t^{adv}, y\right)}{\left\|\nabla_x \mathcal{L}\left(x_t^{adv}, y\right)\right\|_1} \\ x_{t+1}^{adv} &= \mathrm{Clip}_x^{\epsilon}\left\{x_t^{adv} + \alpha \cdot \mathrm{sign}\left(g_{t+1}\right)\right\} \end{aligned} \tag{5}$$

如式 (5) 所示, 在每一轮迭代开始之前, 生成的对抗样本 x_t^{adv} 先和二值化掩码 M 做点乘操作, 这样每次迭代起点就变为了只包含背景噪声的对抗样本. 如果只想将对抗噪声限制在背景区域上, 也可以在整个迭代过程结束后, 再将生成的对抗样本和二值化掩码做点乘操作. 这两种做法的差别在于二值化掩码是否影响了对抗噪声的生成. 将二值化掩码加入到梯度计算过程中, 梯度计算集中在损失函数到背景噪声上, 每一轮迭代生成的对抗样本只考虑了背景部分噪声对模型决策边界的干扰; 而只在对抗样本迭代过程结束后, 再将生成的对抗样本和二值化掩码做点乘操作, 并不能影响迭代过程中梯度的计算和每一轮迭代的起点, 这导致在迭代过程中, 梯度计算同时考虑了前景和背景部分, 最后将生成的对抗样本和二值化掩码做点乘操作, 简单地剔除前景部分的噪声, 必然会损失一部分攻击性能.

2.2.2　空间变换

物理世界中对抗样本的鲁棒性是指对抗样本在受到一定环境因素干扰后, 保持其攻击性能的能力. 在物理世界中, 对抗样本可能会受到多种环境因素的干扰, 如拍摄的角度、距离、光线的变化等, 此时, 对抗样本的攻击性能和其对抗鲁棒性成正比. 受到期望损失函数的启发[44], 本文通过在生成对抗样本的过程中引入一系列的输入变换, 增强了其在物理世界中的对抗鲁棒性. 具体地, 本文考虑了三种输入变换: 随机缩放, 亮度变化和高斯噪声. 随机缩放不仅可以模拟拍摄距离的远近, 还可以模拟摄像头变焦的效果. 由于 CRNN 模型架构对输入图像的长宽比例有一定的要求, 随机缩放需要限制缩放比例在一定范围内. 亮度变化模拟了拍摄环境光线的影响. 高斯噪声模拟了在拍摄时可能存在的噪点的影响. 这些输入变换以集合的形式存在, 即针对一个输入样本, 会以概率 p 将其顺序做这三种变换, 记为 $T(\cdot; p)$. 这些输入变换可以与基于迁移的黑盒攻击算法结合, 提升对抗样本在这些变换下的鲁棒性. 以 MI-FGSM 为例, 在加入二值化掩码 M 和输入变换 $T(\cdot\, p)$ 后, 对抗样本 x_{adv} 的计算过程如下:

$$\begin{cases} x_t^{adv} = T(x_t^{adv} \odot M; p) \\ g_{t+1} = \mu \cdot g_t + \dfrac{\nabla_x \mathcal{L}\left(x_t^{adv}, y\right)}{\left\|\nabla_x \mathcal{L}\left(x_t^{adv}, y\right)\right\|_1} \\ x_{t+1}^{adv} = \mathrm{Clip}_x^{\epsilon}\left\{x_t^{adv} + \alpha \cdot \mathrm{sign}\left(g_{t+1}\right)\right\} \end{cases} \tag{6}$$

由于身份证文本识别模型需要特定的前处理流程, 即在经过文本检测模型后, 会将待识别的图像区域通过仿射变换处理为水平的图像, 本文没有考虑旋转和平移这一类的空间变换. 除此之外, 由于文本图像经过旋转或平移后可能会导致上下文信息的丢失. 在这种情况下, 即使不加对抗噪声干扰, 模型也无法给出正确的判断结果, 此时生成的对抗样本是没有意义的.

综上, 针对身份证识别模型的黑盒对抗攻击算法流程图如图 5 所示. 以一张身份证图像为例, 在

图 5 针对身份证识别模型的黑盒对抗攻击算法流程图 (身份证信息为随机生成)

Fig. 5 Pipeline of the black-box adversarial attack algorithm for ID card recognition model (Information on ID card is randomly generated)

经过文本检测模型后, 会输出关键字段的图像区域对应的位置, 之后会将关键字段的文本图像裁剪下来. 在裁剪完关键字段的文本图像后, 会同时生成对应的二值化掩码, 一同送入本地的基于 CRNN 架构的文本识别模型 $f_1(\theta_1)$. 利用本文提出的攻击算法, 同时考虑迁移性和对抗鲁棒性, 生成只包含背景噪声的对抗样本, 具体流程见算法 1. 之后, 将生成的对抗样本复制到对应的关键字段区域. 最后利用生成的包含对抗样本字段的身份证攻击黑盒模型 $f_2(\theta_2)$. 图 5 中, 为了简洁表示, 只画出了姓名和身份证号字段的对抗样本, 其他字段的对抗样本同理.

算法 1. 针对身份证识别模型的黑盒对抗攻击算法

输入. 身份证识别模型 f, 损失函数 $\mathcal{L}$, 输入样本 x, 对应标签 y, 输入样本的维度 d, 二值化掩码 M.

参数. 对抗噪声范数限制 ϵ, 迭代次数 K, 动量因子 μ, 幅值缩放次数 m, 空间变化 $T(\cdot;p)$, 梯度卷积核 ω, 邻域采样数量 N, 邻域采样阈值 β.

输出. 对抗样本 x^{adv}.

$\alpha = \frac{\epsilon}{K}$;
$g_0 = 0,\ v_0 = 0,\ x_0^{adv} = x$;
for $k = 0$ to $K-1$ **do**
  $g = 0,\ g_{temp} = 0$;
  $x_k^{nes} = (x_k^{adv} + \alpha \cdot \mu \cdot g_k) \odot M$;
  for $i = 0$ to $m-1$ **do**
    $g = g + \nabla_x \mathcal{L}\left(f\left(T\left(S_i\left(x_k^{\text{nes}}\right); p\right)\right), y\right)$;
  end for
  $g_{new} = \frac{1}{m} \cdot g$;
  for $n = 0$ to $N-1$ **do**
    for $i = 0$ to $m-1$ **do**
      $r^i \sim \mathrm{U}\left[-(\beta \cdot \epsilon)^d, (\beta \cdot \epsilon)^d\right]$;
      $g = g + \nabla_x \mathcal{L}(f(T(S_i(x_k^{\text{nes}} + r^i); p)), y)$;
    end for
    $g_{temp} = g_{temp} + \frac{1}{m} \cdot g$;
  end for
  $g = (g_{new} + v_k) * \omega$;
  $g_{k+1} = \mu \cdot g_k + \frac{g}{\|g\|_1}$;
  $v_{k+1} = \frac{1}{N} \cdot g_{temp} - g_{new}$;
  $x_{k+1}^{adv} = \mathrm{Clip}_x^{\epsilon}\left\{x_t^{adv} + \alpha \cdot \mathrm{sign}\left(g_{k+1}\right)\right\}$;
end for
$x^{adv} = x_K^{adv}$;
返回对抗样本 x^{adv}.

2.3 实验设置与结果分析

2.3.1 数据集与实验设置

为了避免使用真实身份证数据带来的隐私问题, 本文自制了一批身份证数据用于模型的训练和攻击算法的测试. 身份证图像共包含姓名、性别、民族、出生日期、住址和身份证号 6 个关键字段信息, 这 6 个字段的信息包含中文, 数字和一个特殊字符 (身份证号中的 X) 三种字符类型. 本文按照表 2 中的标准生成关键字段信息所对应的文本图像.

如表 2 所示, 本文首先将关键字段信息分为两大类, 文字和数字 (包括特殊字符 X). 其中, 文字部分包括姓名和住址, 对应的字典长度为 6 270, 涵盖了绝大部分的常用汉字和生僻字. 姓名的长度为 2 到 4, 地址的长度为 1 到 11, 均为随机挑选. 由于民族和性别字段的字体, 字号和住址相同, 且长度也在 1 到 11 之间, 所以这两部分字段的信息可以由住址代替, 对文本识别模型的训练和算法测试没有影响. 数字部分包括出生日期和身份证号两部分, 对应的字典长度为 11 (数字 0 到 9 和一个特殊字符 X), 出生日期长度为 1 到 4, 身份证号长度为 18. 其中, 频率代表字典中的每个字符出现的次数, 确保数据集中字符的分布不会出现长尾效应. 训练集和测试集按照 10 比 1 的比例划分, 整个字段共包含 297 625 张训练图像和 29 763 张测试图像. 此外, 本文生成的关键字段信息符合身份证标准中字体, 字号和身份证号编码规则. 图 6 展示了 4 份身份证数据集样张.

本文首先在本地训练了两个身份证文本识别模型, 这两个模型均采用 CRNN 架构和 CTC 损失函数, 区别是应用了两种不同的卷积神经网络作为图像特征提取模块 (ResNet-34 和 MobileNet-v3). 为了简化表述, 本文将这两个模型分别记为 res34 和 mbv3. 其次, 本文在本地的黑盒条件下, 进行了两组实验. 第一组实验中, 替代模型为 res34, 黑盒模型为 mbv3; 第二组实验中, 替代模型为 mbv3, 黑盒模型为 res34. 两组实验互为对照, 可以更好地测试本文所提出的算法在不同实验设置中的性能. 为

表 2　身份证关键字段文本图像生成标准

Table 2　ID card key field text image generation standard

类别	字典长度	字段	长度	频率	训练集	测试集
文字	6270	姓名	[2, 4]	50	110875	11088
		住址	[1, 11]	150	156750	15675
数字	11	出生日期	[1, 4]	5000	20000	2000
		身份证号	18	18000	10000	1000

图 6　身份证数据集样张

Fig. 6　Examples of the ID card dataset

了验证本文提出方法对基于 attention 方法和基于 transformer 方法的文本识别模型的有效性, 在本地训练了一个基于 attention 方法的文本识别模型和一个基于 transformer 的文本识别模型, 以黑盒条件进行了第三组实验. 基于 attention 方法的文本识别模型将 res34 模型的解码器改为 seq2seq 结构, 并基于 attention 方法进行解码, 为了简化表述, 将该模型称为 res34-att. 基于 transformer 方法的文本识别模型采用 NRTR 方法, 记为 NRTR. 第三组实验, 黑盒模型分别为 res34-att 和 NRTR, 替代模型为 mbv3. 为了验证不同的基于迁移的黑盒攻击算法的性能, 本文共测试了 7 种基于迁移的黑盒攻击算法, 分别为 MI-FGSM, TMI-FGSM, SI-NI-TMI, DMI-FGSM, SI-NI-DMI, DI-TIM 和 VNI-SI-DI-TIM. 算法中涉及到的具体参数设置如下: 迭代次数 $K=10$, 动量因子 $\mu=1$, 幅值缩放次数 $m=5$, 空间变化 $T(\cdot;p)$ 的概率 $p=0.5$, 邻域采样数量 $N=20$, 邻域采样阈值 $\beta=1.5$. 之后, 本文针对梯度卷积核 ω 尺寸进行了消融实验, 用于选择合适的卷积核来平衡视觉效果和攻击成功率. 为了测试本算法在实际场景中的表现, 本文将百度身份证识别模型作为攻击对象, 从测试集中挑选了合适的样本作为新的攻击数据集, 并且针对二值化掩码进行了消融实验, 验证二值化掩码对于视觉效果的提升作用. 最后, 本文在物理世界中验证了空间变换对于提升鲁棒性的作用.

实验用的评价指标包括攻击成功率和编辑距离. 其中, 攻击成功的标准是模型输出和原始标签不同. 编辑距离是指两个字符串之间, 由一个字符串转换为另一个字符串所需要的最少操作次数. 可用的操作包括替换, 插入和删除. 它可以量化模型输出和原始标签之间的差异性, 编辑距离越大, 差异性越大, 编辑距离越小, 差异性越小. 攻击成功率 (Attack sueccess rate, ASR) 和平均编辑距离 (Average edit distance, AED) 的计算公式分别如式 (7) 和式 (8) 所示.

$$\mathrm{ASR}=\frac{\sum_{i=1}^{N}1(f(x_{adv})\neq y)}{N}\times 100\% \tag{7}$$

$$\mathrm{AED}=\frac{\sum_{i=1}^{N}\{d\left[f(x_{adv})\rightarrow y\right]/l(y)\}}{N} \tag{8}$$

式 (7) 和 (8) 中, N 代表样本总数, $d\left[f(x_{adv})\rightarrow y\right]$ 表示对抗样本的输出结果和原始标签之间的编辑距离, $l(y)$ 代表原始标签的长度. 此处, 平均编辑距离经过了归一化处理.

2.3.2　实验结果

经过模型训练, res34 和 mbv3 在测试集上分别达到了 99.99% 和 99.81% 的识别准确率, 满足攻击算法的测试要求. 本文将所采用的基于迁移的黑盒攻击算法 VNI-D(T, S)I-FGSM 记为 VNI-SI-DI-TIM. 为了更加全面地评估算法的性能, 实验中采用了不同大小的 l_∞ 范数限制, 对每种范数限制下的攻击成功率和平均编辑距离都进行了统计, 并且给出了这两个指标在不同范数限制下的均值. 白盒条件下的两组实验结果如表 3 和表 4 所示, 黑盒条件下的两组实验结果如表 5 和表 6 所示. 这四个实验中, 均采用了 7 种不同大小的范数限制, 从 4 到 80 (以输入图像范数为 0 到 255 为量纲).

由表 3 和表 4 可知, 在白盒条件下, 所测试算法的平均攻击成功率都达到了 97%, 平均编辑距离大于 0.6. 平均编辑距离和攻击成功率成正比关系, 攻击成功率越高, 平均编辑距离也越大. 即使是在较小的范数限制下, 如 4 或 8, 攻击成功率几乎都在 90% 以上. 通过白盒条件下的这两组实验, 充分

表 3 第一组实验中针对替代模型 res34 的白盒攻击结果

Table 3 White-box attack results against the surrogate model res34 in the first set of experiments

评价指标	白盒条件	范数限制							指标均值
	对抗攻击算法	4	8	16	32	48	64	80	
攻击成功率 (%)	MI-FGSM	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
	TMI-FGSM	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
	SI-NI-TMI	88.24	100.00	100.00	100.00	100.00	100.00	100.00	98.32
	DMI-FGSM	96.47	98.82	100.00	100.00	100.00	100.00	100.00	99.33
	SI-NI-DMI	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
	DI-TIM	90.59	92.94	98.82	100.00	100.00	100.00	100.00	97.48
	VNI-SI-DI-TIM	92.94	98.82	100.00	100.00	100.00	100.00	100.00	98.82
平均编辑距离	MI-FGSM	0.4612	0.6120	0.8203	0.9382	0.9705	0.9910	0.9939	0.83
	TMI-FGSM	0.4328	0.6399	0.8345	0.9190	0.9519	0.9169	0.9903	0.81
	SI-NI-TMI	0.3156	0.4200	0.5679	0.6580	0.7536	0.8083	0.8454	0.62
	DMI-FGSM	0.4148	0.5757	0.7379	0.8907	0.9375	0.9693	0.9869	0.79
	SI-NI-DMI	0.5038	0.6676	0.7998	0.8889	0.9716	0.9764	0.9887	0.83
	DI-TIM	0.3279	0.4115	0.5432	0.6689	0.7093	0.7685	0.8007	0.60
	VNI-SI-DI-TIM	0.4046	0.5194	0.6116	0.6876	0.7506	0.7730	0.8378	0.65

表 4 第二组实验中针对替代模型 mbv3 的白盒攻击结果

Table 4 White-box attack results against the surrogate model mbv3 in the second set of experiments

评价指标	白盒条件	范数限制							指标均值
	对抗攻击算法	4	8	16	32	48	64	80	
攻击成功率 (%)	MI-FGSM	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
	TMI-FGSM	98.82	100.00	100.00	100.00	100.00	100.00	100.00	100.00
	SI-NI-TMI	98.82	100.00	100.00	100.00	100.00	100.00	100.00	99.83
	DMI-FGSM	95.29	100.00	100.00	100.00	100.00	100.00	100.00	99.33
	SI-NI-DMI	98.82	100.00	100.00	100.00	100.00	100.00	100.00	100.00
	DI-TIM	90.59	98.82	100.00	100.00	100.00	100.00	100.00	98.49
	VNI-SI-DI-TIM	90.59	95.29	98.82	100.00	100.00	100.00	100.00	97.81
平均编辑距离	MI-FGSM	0.6540	0.6911	0.8557	0.9594	0.9895	0.9945	0.9975	0.88
	TMI-FGSM	0.6683	0.6930	0.8901	0.9652	0.9982	0.9991	0.9996	0.89
	SI-NI-TMI	0.5756	0.6348	0.7950	0.9263	0.9876	0.9927	0.9956	0.84
	DMI-FGSM	0.5227	0.6420	0.7592	0.9395	0.9701	0.9813	0.9859	0.83
	SI-NI-DMI	0.6478	0.7881	0.8962	0.9833	0.9941	0.9965	0.9990	0.90
	DI-TIM	0.5294	0.6257	0.7360	0.8430	0.8756	0.9136	0.9507	0.78
	VNI-SI-DI-TIM	0.4619	0.5740	0.6240	0.7392	0.8217	0.8887	0.9176	0.72

说明了基于迁移的黑盒攻击算法的有效性, 同时也揭示了身份证识别模型在白盒条件下的脆弱性.

由表 5 和表 6 可知, 在黑盒条件下, 所测试算法的平均攻击成功率在 35% 到 65% 之间, 平均编辑距离在 0.15 到 0.45 之间. 在两组实验中, VNI-SI-DI-TIM 的性能表现都是最好的, 平均攻击成功率分别为 56.83% 和 62.69%, 平均编辑距离分别为 0.28 和 0.43. 值得注意的是, 在较小的范数限制下, 如 4 和 8, 所测试算法的攻击成功率均较低, 都在 10% 以下. 而当范数限制增加到一定阈值时, 如 32 或 48, 攻击成功率才有了明显的提升.

由表 7 和表 8 可知, 在黑盒条件下, 所测试算法的平均攻击成功率在 83% 到 88% 之间, 平均编辑距离在 0.82 到 0.83 之间. 在两组实验中, VNI-SI-DI-TIM 的性能表现都是最好的, 平均攻击成功率分别为 87.65% 和 87.04%, 平均编辑距离分别为 0.8340 和 0.8243. 值得注意的是, ocr-att 模型和 NRTR 模型在身份证数据集上的鲁棒性表现弱于 CRNN 结

表 5　第一组实验中针对黑盒模型 mbv3 的黑盒攻击结果

Table 5　Black-box attack results against the black-box model mbv3 in the first set of experiments

评价指标	黑盒条件	范数限制							指标均值
	对抗攻击算法	4	8	16	32	48	64	80	
攻击成功率 (%)	MI-FGSM	7.06	7.06	18.82	41.18	62.35	87.06	90.59	44.87
	TMI-FGSM	7.06	7.06	12.94	35.29	47.06	71.76	90.59	38.82
	SI-NI-TMI	7.06	7.06	17.65	36.47	67.06	84.71	89.41	44.20
	DMI-FGSM	7.06	7.06	12.94	49.41	88.24	94.12	98.82	51.09
	SI-NI-DMI	7.06	9.41	16.47	63.53	87.06	95.29	100.00	54.12
	DI-TIM	7.06	9.41	21.18	37.65	64.71	88.24	89.41	45.38
	VNI-SI-DI-TIM	7.06	9.41	16.47	75.29	92.94	96.67	100.00	**56.83**
平均编辑距离	MI-FGSM	0.0316	0.0316	0.0446	0.1386	0.2062	0.3148	0.4156	0.17
	TMI-FGSM	0.0316	0.0316	0.0398	0.1152	0.1535	0.2815	0.3921	0.15
	SI-NI-TMI	0.0316	0.0316	0.0526	0.1161	0.233	0.3153	0.4265	0.17
	DMI-FGSM	0.0316	0.0316	0.0631	0.1545	0.362	0.4191	0.5942	0.24
	SI-NI-DMI	0.0316	0.0402	0.0616	0.2252	0.3862	0.4943	0.5695	0.26
	DI-TIM	0.0316	0.0350	0.0577	0.1390	0.2179	0.3289	0.4331	0.18
	VNI-SI-DI-TIM	0.0316	0.0378	0.0519	0.2609	0.4450	0.5263	0.6202	**0.28**

表 6　第二组实验中针对黑盒模型 res34 的黑盒攻击结果

Table 6　Black-box attack results against the black-box model res34 in the second set of experiments

评价指标	黑盒条件	范数限制							指标均值
	对抗攻击算法	4	8	16	32	48	64	80	
攻击成功率 (%)	MI-FGSM	7.06	7.06	27.06	58.82	89.41	91.76	98.82	54.28
	TMI-FGSM	7.06	9.41	21.18	55.29	83.53	89.41	95.29	51.60
	SI-NI-TMI	7.06	9.41	23.53	62.35	88.24	89.41	97.65	53.95
	DMI-FGSM	7.06	8.24	30.59	83.24	91.29	95.46	100.00	59.41
	SI-NI-DMI	5.88	7.06	36.47	87.06	94.12	100.00	100.00	61.51
	DI-TIM	7.06	9.41	23.53	62.35	88.24	90.59	100.00	54.45
	VNI-SI-DI-TIM	7.06	4.71	40.00	89.41	97.65	100.00	100.00	**62.69**
平均编辑距离	MI-FGSM	0.0182	0.0201	0.0810	0.2821	0.4629	0.5447	0.6466	0.29
	TMI-FGSM	0.0182	0.0210	0.0726	0.2441	0.4271	0.4797	0.5173	0.25
	SI-NI-TMI	0.0182	0.0254	0.0654	0.2833	0.4435	0.5106	0.5690	0.27
	DMI-FGSM	0.0182	0.0192	0.0815	0.4807	0.6509	0.7377	0.8433	0.40
	SI-NI-DMI	0.0173	0.0185	0.1024	0.4961	0.6931	0.7505	0.8771	0.42
	DI-TIM	0.0182	0.0232	0.0908	0.3027	0.4506	0.5138	0.6211	0.29
	VNI-SI-DI-TIM	0.0182	0.0169	0.1021	0.4745	0.6471	0.8197	0.9620	**0.43**

构的识别模型, 其具体表现为 ocr-att 模型和 NRTR 模型很容易将扰动噪声误解码.

为了更加直观地比较不同算法的攻击效果, 本文展示了不同范数限制下的各个算法的攻击成功率, 如图 7 所示. 由图 7 可知, 攻击成功率与范数限制成正比关系, 范数限制越大, 攻击成功率越高. VNI-SI-DI-TIM 与其他算法相比, 在较小的范数限制下, 如 4 和 8 时, 优势并不明显; 而在较大的范数限制下, 如 32 及以上, 才具有较为明显的攻击成功率优势. 由于范数限制和视觉效果成反比, 即范数限制越大, 对抗噪声就会越明显, 视觉效果越差. 在达到相似的攻击成功率时, VNI-SI-DI-TIM 需要的范数限制较小, 生成的对抗样本视觉效果较好.

为了测试梯度卷积核尺寸对于攻击成功率的影响, 本文进行了两组消融实验. 第一组实验中, 黑盒模型为 res34, 第二组实验中, 黑盒模型为 mbv3. 这两组实验均选取了大小为 5×5 和 15×15 的两种高斯卷积核, 所选择的基准方法为 TMI, 因为其只包

表 7 第三组实验中针对黑盒模型 res34-att 的黑盒攻击结果

Table 7 Black-box attack results against the black-box model res34-att in the third set of experiments

评价指标	黑盒条件	范数限制						指标均值
	对抗攻击算法	8	16	32	48	64	80	
攻击成功率 (%)	MI-FGSM	65.17	74.76	84.13	94.57	99.07	100.00	86.28
	TMI-FGSM	65.67	69.62	80.83	99.04	100.00	100.00	83.03
	SI-NI-TMI	70.82	70.56	85.03	95.74	100.00	100.00	87.02
	DMI-FGSM	65.60	75.52	80.39	99.45	99.81	100.00	86.80
	SI-NI-DMI	65.00	75.92	69.64	90.10	99.42	100.00	83.35
	DI-TIM	74.72	76.47	82.44	94.01	98.66	98.96	87.55
	VNI-SI-DI-TIM	70.34	65.74	89.82	100.00	100.00	100.00	**87.65**
平均编辑距离	MI-FGSM	0.7700	0.7800	0.7800	0.8000	0.8500	0.9600	0.8200
	TMI-FGSM	0.7693	0.7743	0.7842	0.8025	0.8618	0.9715	0.8270
	SI-NI-TMI	0.7701	0.7750	0.7823	0.8068	0.8741	0.9445	0.8257
	DMI-FGSM	0.7699	0.7745	0.7801	0.7972	0.8645	0.9639	0.8250
	SI-NI-DMI	0.7698	0.7745	0.7814	0.8000	0.8622	0.9652	0.8260
	DI-TIM	0.6847	0.8035	0.7840	0.7796	0.8458	0.9909	0.8148
	VNI-SI-DI-TIM	0.7690	0.7769	0.7841	0.8100	0.8756	0.9900	**0.8340**

表 8 第三组实验中针对黑盒模型 NRTR 的黑盒攻击结果

Table 8 Black-box attack results against the black-box model NRTR in the third set of experiments

评价指标	黑盒条件	范数限制						指标均值
	对抗攻击算法	8	16	32	48	64	80	
攻击成功率 (%)	MI-FGSM	59.29	74.93	84.05	94.59	99.08	100.00	85.32
	TMI-FGSM	65.11	69.33	79.45	99.32	99.65	99.62	85.41
	SI-NI-TMI	69.19	69.77	84.08	94.51	98.96	99.06	85.93
	DMI-FGSM	64.10	74.95	79.39	100.00	99.08	99.61	86.19
	SI-NI-DMI	64.39	74.37	69.06	88.89	99.01	99.27	82.50
	DI-TIM	74.76	76.73	80.57	93.26	96.98	98.51	86.80
	VNI-SI-DI-TIM	69.46	65.08	88.63	99.49	99.60	100.00	**87.04**
平均编辑距离	MI-FGSM	0.7493	0.7414	0.7270	0.7765	0.8302	0.9357	0.7933
	TMI-FGSM	0.7533	0.7602	0.7955	0.7506	0.8673	1.0099	0.8228
	SI-NI-TMI	0.7805	0.7618	0.8034	0.7646	0.8952	0.9310	0.8228
	DMI-FGSM	0.7612	0.7475	0.7929	0.7523	0.8245	0.9529	0.8052
	SI-NI-DMI	0.7781	0.7453	0.7928	0.8143	0.8072	0.9282	0.8110
	DI-TIM	0.7037	0.7494	0.7428	0.7576	0.8373	0.9509	0.8076
	VNI-SI-DI-TIM	0.7828	0.8020	0.7653	0.8293	0.8307	0.9357	**0.8243**

含对梯度的卷积这一项输入变化, 排除了其他输入变化的干扰. 实验结果如表 9 和表 10 所示.

由表 9 和表 10 可得, 卷积核尺寸由 5×5 增加到 15×15 后, 平均攻击成功率升高. 第一组实验中, TMI 的平均攻击成功率增加了 8.66%, VNI-SI-DI-TIM 增加了 5.85%; 第二组实验中, TMI 的平均攻击成功率增加了 7.37%, VNI-SI-DI-TIM 增加了 7.32%. 为了更加直观地观察到这一现象, 本文展示了不同卷积核尺寸下的攻击成功率趋势图, 如图 8 所示. 从图中可以看出, 在较大的范数限制下 (大于 16), 增大卷积核可以明显提升攻击成功率. 为了清晰地观察到不同卷积核尺寸下生成对抗样本的视觉效果, 本文以姓名为"康惫凭"的文本图像为例, 在图 9 中展示了大小为 5×5 和 15×15 的两组卷积核在不同范数限制下所生成的对抗样本. 可以观察到, 随着范数限制的增大, 对抗噪声的幅值增加, 颜色加深, 生成的对抗样本视觉效果变差. 而较大的卷积核尺寸, 使得生成的对抗噪声更加平滑, 提升

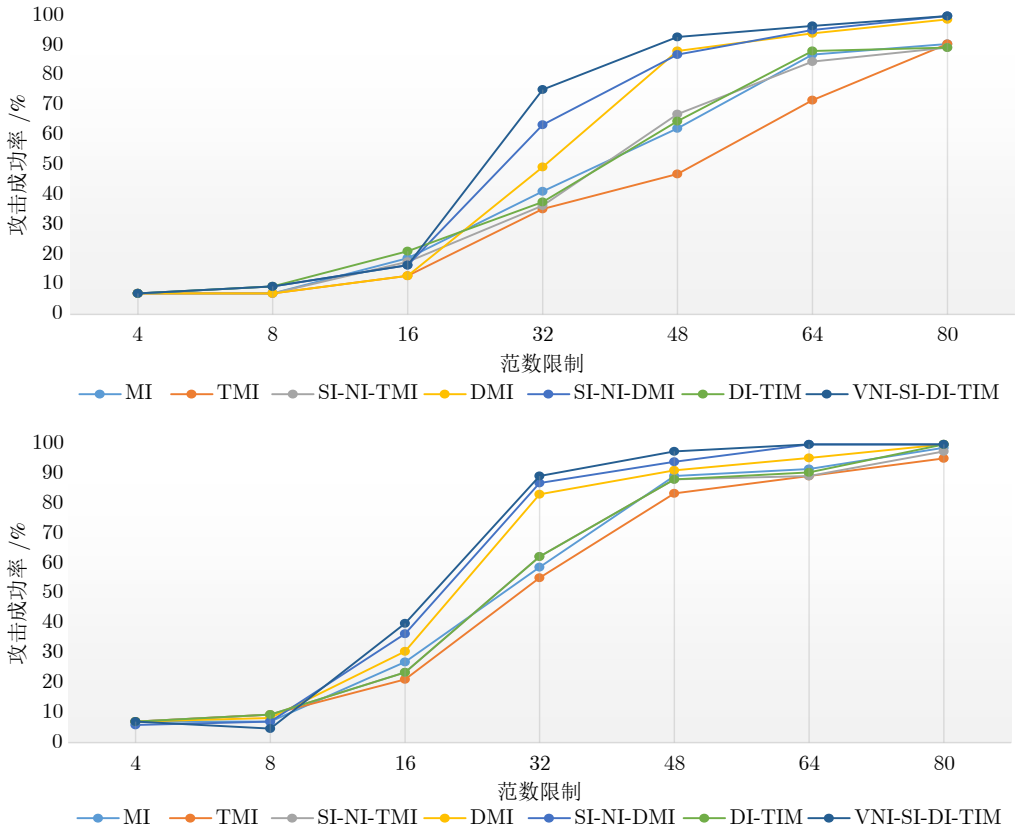


图 7　不同范数限制下的攻击成功率趋势图 (上: 针对 mbv3 的黑盒攻击实验, 下: 针对 res34 的黑盒攻击实验)

Fig. 7　Trend graph of attack success rate under different norm constraints (Top: Black-box attack experiment for mbv3; Bottom: Black-box attack experiment for res34)

表 9　卷积核尺寸消融实验, 第一组实验中针对黑盒模型 res34 的攻击结果

Table 9　Convolution kernel size ablation experiment, the attack results against the black-box model res34 in the first set of experiments

评价指标	黑盒条件		范数限制							指标均值
	卷积核	对抗攻击算法	4	8	16	32	48	64	80	
攻击成功率 (%)	5 × 5	TMI	7.06	7.06	8.21	23.97	31.34	55.49	77.98	30.16
		VNI-SI-DI-TIM	7.06	8.11	13.56	66.37	80.71	87.67	93.38	50.98
	15 × 15	TMI	7.06	7.06	12.94	35.29	47.06	71.76	90.59	38.82
		VNI-SI-DI-TIM	7.06	9.41	16.47	75.29	92.94	96.67	100.00	56.83

表 10　卷积核尺寸消融实验, 第二组实验中针对黑盒模型 mbv3 的攻击结果

Table 10　Convolution kernel size ablation experiment, the attack results against the black-box model mbv3 in the second set of experiments

评价指标	黑盒条件		范数限制							指标均值
	卷积核	对抗攻击算法	4	8	16	32	48	64	80	
攻击成功率 (%)	5 × 5	TMI	5.88	7.06	15.67	44.31	72.15	79.76	84.81	44.23
		VNI-SI-DI-TIM	5.88	7.06	29.86	78.03	85.69	88.96	92.11	55.37
	15 × 15	TMI	7.06	9.41	21.18	55.29	83.53	89.41	95.29	51.60
		VNI-SI-DI-TIM	7.06	4.71	40.00	89.41	97.65	100.00	100.00	62.69

了视觉效果. 基于此, 本文在后续实验中将高斯卷积核尺寸设置为 15×15.

为了测试本文的方法在实际场景中的攻击性能, 本文针对百度身份证识别模型进行了实验. 首

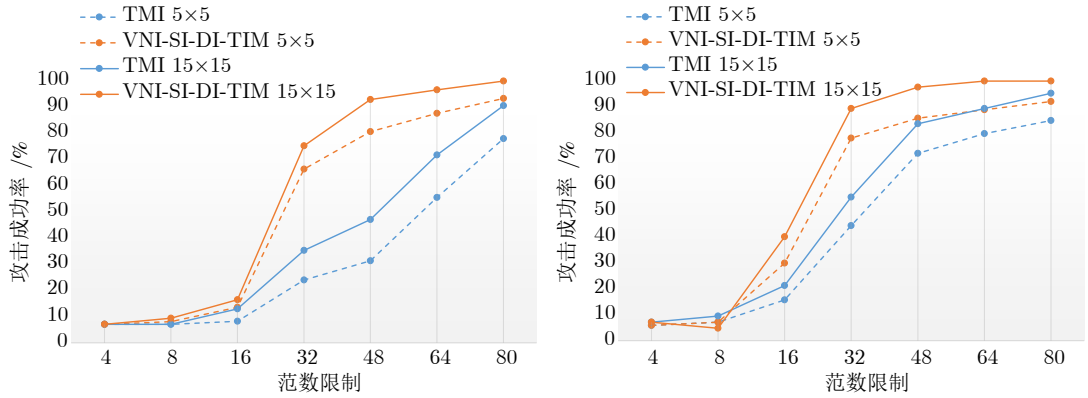


图 8 不同卷积核大小下的攻击成功率趋势图 (左: 针对 res34 的黑盒攻击结果, 右: 针对 mbv3 的黑盒攻击结果)

Fig. 8 Trend chart of attack success rate under different convolution kernel sizes (Left: Black-box attack results against res34; Right: Black-box attack results against mbv3)

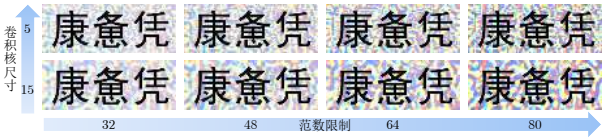


图 9 不同卷积核尺寸生成的对抗样本示例图

Fig. 9 Adversarial examples generated with different convolution kernel sizes

先, 本文在测试集中挑选了两种关键字段的文本图像, 姓名和身份证号, 每种类型的图像各 100 张, 还有 100 张相应的身份证图像, 作为攻击数据集. 挑选的标准是关键字段的文本图像可以被替代模型 mbv3 正确识别. 由于姓名和身份证号包含了几乎是最重要的个人信息 (出生, 性别, 住址的省市均可以由身份证号获取), 且同时含有文字, 数字和特殊字符 X, 选择这两部分字段的文本图像作为攻击数据集不仅可以全面地分析算法的性能, 而且可以控制百度身份证模型的调用成本. 之后, 利用 VNI-SI-DI-TIM 算法生成关键字段的对抗样本, 再将其替换回原有的关键字段区域, 生成身份证图像. 最后, 调用百度身份证识别模型接口进行识别. 由于在之前的实验中, 范数限制大小在 48 及以上时, 针对本地黑盒模型的攻击成功率才达到了 90% 以上, 本实验采用了 48 到 128 的 5 种范数限制. 此外, 为了测试二值化掩码对于视觉效果的提升作用, 同步设计了消融实验, 实验结果如表 11 所示. 其中, ✓ 表示加入二值化掩码, × 表示剔除二值化掩码.

由表 11 可得, 在较小的范数限制下, 如 48 和 64, 攻击成功率均较低. 当范数限制增加到 80 时, 攻击成功率有了明显的提高. 当范数限制达到 128 时, 不论是否加入二值化掩码, 攻击成功率均达到了 100%. 这说明, 范数限制大小是影响攻击成功率的主要因素. 范数限制越大, 对抗噪声的搜索空间越大, 攻击成功率越高. 然而, 增大范数限制会降低视觉效果, 如图 10 所示. 尤其对于文本图像而言, 全局的范数限制会导致对抗噪声覆盖在字体上, 严重干扰人类的判断, 这与对抗样本的定义相违背, 二值化掩码的加入可以极大地改善这一问题.

如图 10, 针对姓名字段, 加入二值化掩码的算法, 平均攻击成功率下降了 1.4%; 针对身份证号字段, 平均攻击成功率下降了 1.2%. 从实验结果看出, 通过加入二值化掩码, 对抗样本在视觉效果与攻击效果之间达到一个平衡, 在不显著影响攻击成功率的同时, 有效地增强视觉效果.

为了更加直观地观察到二值化掩码对视觉效果的影响, 本文在图 10 中展示了“康惫凭”这一姓名

表 11 针对百度身份证识别模型的攻击结果

Table 11 Attack results against Baidu ID card recognition model

评价指标	对抗攻击算法		字段	范数限制					指标均值
	VNI-SI-DI-TIM			48	64	80	96	128	
攻击成功率 (%)	二值化掩码	✓	姓名	1	12	45	84	100	48.4
			身份证号	2	9	44	80	100	47.0
		×	姓名	2	15	46	86	100	49.8
			身份证号	2	11	46	82	100	48.2

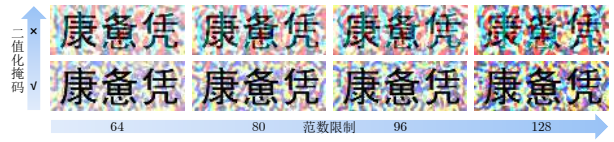


图 10　二值化掩码消融实验中生成的对抗样本示例图

Fig. 10　Adversarial examples generated in the ablation experiment of the binarized mask

图 11　身份证对抗样本及百度身份证模型识别结果 (第一行: 加入二值化掩码, 第二行: 剔除二值化掩码)

Fig. 11　ID card adversarial examples and Baidu ID card recognition results (The first line: Add binarization mask; The second line: Cancel binarization mask)

字段的对抗样本, 并且在图 11 中展示了其完整身份证图像和百度身份证识别模型的输出结果.

从图 10 可以观察到, 加入二值化掩码对于提升对抗样本的视觉效果有较为显著的作用. 尤其是在较大的范数限制下, 将对抗噪声限制在背景区域是非常有必要的. 从图中可以观察到, 二值化掩码是否存在, 影响了生成的对抗噪声的颜色和形状. 这说明, 加入二值化掩码并不是简单地将字体上的对抗噪声消除, 而是在背景区域产生相似攻击性的对抗噪声. 这样的实验结果也从侧面说明, 基于 CRNN 架构的文本识别模型, 其决策边界容易受到所学的背景特征的影响.

如图 11 所示, 百度身份证识别模型将这两张身份证图像中的姓名字段全部判断出错, 而身份证号字段由于没有检测到, 所以并没有返回识别结果, 其范数限制均为 128. 虽然这两张对抗样本身份证均可以攻击成功, 但剔除二值化掩码所生成的对抗样本, 已经显著影响到了人类的观察, 而加入二值化掩码可以有效地改善视觉效果, 降低对人类观察的影响. 对于百度身份证识别模型的攻击实验充分说明, 模型架构的差异对于攻击成功率的影响非常大. 由于本地的替代模型和训练用的数据集与百度的存在较大的差异性, 在本地范数限制 48 就可以达到 90% 以上攻击成功率的算法, 在百度身份证识别模型上需要几乎两倍的范数限制, 才能达到类似的攻击成功率. 这说明在实际应用场景中, 替代模型和黑盒模型之间的差异性对于基于迁移性的黑盒攻击算法影响非常大.

为了验证本方法在物理世界中的攻击性能和空间变换对于对抗鲁棒性的提升作用, 本文在物理世界中进行了实验. 首先, 本文在数字世界中挑选了十张范数限制为 128 的对抗身份证图像, 图像分辨率为 2 120 × 1 370 像素, 且这十张图像均可以成功攻击百度身份证识别模型. 之后, 将其保存为便捷文档格式 (Portable document format, PDF) 并利用打印机打印出来. 最后用手机分别在三组不同拍摄距离和光照条件下进行拍摄, 共得到 90 张对抗身份证图像, 利用其攻击百度身份证识别模型. 手机型号为小米 9, 镜头参数为默认长焦, 图像分辨率均为 4 000 × 2 250 像素. 图 12 展示了其中一张身份证图像对应的 9 张物理世界的对抗样本.

如图 12 所示, 在较强的亮度和较近的拍摄距离下, 对抗样本的视觉效果最好. 图中的 9 张对抗样本及全部的 90 张对抗样本均可以成功攻击百度身份证识别模型, 攻击结果是姓名和身份证号全部识别出错, 其他的关键字段均可以正确识别. 该实验充分证明了本方法在物理世界中的攻击性能, 并且说明了空间变换可以有效地增强对抗鲁棒性.

3　总结

本文提出了针对身份证文本识别模型的黑盒攻击算法, 同时考虑了黑盒条件和物理世界两种高难度的攻击设定, 在对比分析基于迁移的黑盒攻击算法的基础上, 引入了二值化掩码和空间变换, 改善了较大范数限制下对抗样本的视觉效果和其在物理世界中的鲁棒性. 二值化掩码的加入, 将对抗噪声的搜索空间限制在了背景区域, 只通过背景的改变就影响到了文本识别模型的决策边界, 从侧面揭示了文本识别模型对字体所处的背景特征具有较强的依赖性. 空间变换在数字世界建模了物理世界中光照, 噪声和拍摄距离的影响, 使得对抗样本对环境因素的干扰具有一定的鲁棒性. 在本地的黑盒身份证文本识别模型和商用的百度身份证文本识别模型上均达到了 100% 的攻击成功率. 与已有的针对文本识别模型的对抗攻击算法不同, 本算法在较复杂背景的数据集, 黑盒攻击条件, 物理世界三个方面均具有良好的泛化优势, 攻击设定更加贴近现实场景, 实用价值较高. 未来我们将在驾驶证, 社保卡和银行支票等自然场景文字识别应用上进一步将本文的结论进行扩展, 以适应更多的应用场景.

图 12 不同拍摄距离和光照下的身份证对抗样本

Fig. 12 ID card adversarial examples under different shooting distances and lighting conditions

References

1 Krizhevsky A, Sutskever I, Hinton G E. ImageNet classification with deep convolutional neural networks. In: Proceedings of the 25th International Conference on Neural Information Processing Systems. Lake Tahoe, USA: Curran Associates Inc., 2012. 1097−1105

2 Liu Z, Mao H Z, Wu C Y, Feichtenhofer C, Darrell T, Xie S N. A ConvNet for the 2020s. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans, USA: IEEE, 2022. 11966−11976

3 Bahdanau D, Chorowski J, Serdyuk D, Brakel P, Bengio Y. End-to-end attention-based large vocabulary speech recognition. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Shanghai, China: IEEE, 2016. 4945−4949

4 Afkanpour A, Adeel S, Bassani H, Epshteyn A, Fan H B, Jones I, et al. BERT for long documents: A case study of automated ICD coding. In: Proceedings of the 13th International Workshop on Health Text Mining and Information Analysis (LOUHI). Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, 2022. 100−107

5 Ouyang L, Wu J, Jiang X, Almeida D, Wainwright C L, Mishkin P, et al. Training language models to follow instructions with human feedback. In: Proceedings of the 36th Conference on Neural Information Processing Systems (NeurIPS). 2022. 27730−27744

6 Silver D, Huang A, Maddison C J, Guez A, Sifre L, van den driessche G, et al. Mastering the game of Go with deep neural networks and tree search. *Nature*, 2016, **529**(7587): 484−489

7 Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Highly accurate protein structure prediction with AlphaFold. *Nature*, 2021, **596**(7873): 583−589

8 Sallam M. ChatGPT utility in healthcare education, research, and practice: Systematic review on the promising perspectives and valid concerns. *Healthcare*, 2023, **11**(6): Article No. 887

9 Wang J K, Yin Z X, Hu P F, Liu A S, Tao R S, Qin H T, et al. Defensive patches for robust recognition in the physical world. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans, USA: IEEE, 2022. 2446−2455

10 Yuan X Y, He P, Zhu Q L, Li X L. Adversarial examples: Attacks and defenses for deep learning. *IEEE Transactions on Neural Networks and Learning Systems*, 2019, **30**(9): 2805−2824

11 Wang B H, Li Y Q, Zhou P. Bandits for structure perturbation-based black-box attacks to graph neural networks with theoretical guarantees. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans, USA: IEEE, 2022. 13369−13377

12 Jia X J, Zhang Y, Wu B Y, Ma K, Wang J, Cao X C. LAS-AT: Adversarial training with learnable attack strategy. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans, USA: IEEE, 2022. 13388−13398

13 Li T, Wu Y W, Chen S Z, Fang K, Huang X L. Subspace adversarial training. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans, USA: IEEE, 2022. 13399−13408

14 Xu C K, Zhang C J, Yang Y W, Yang H Z, Bo Y J, Li D Y, et al. Accelerate adversarial training with loss guided propagation for robust image classification. *Information Processing & Management*, 2023, **60**(1): Article No. 103143

15 Chen Z Y, Li B, Xu J H, Wu S, Ding S H, Zhang W Q. Towards practical certifiable patch defense with vision transformer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans, USA: IEEE, 2022. 15127−15137

16 Suryanto N, Kim Y, Kang H, Larasati H T, Yun Y, Le T T H, et al. DTA: Physical camouflage attacks using differentiable transformation network. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans, USA: IEEE, 2022. 15284−15293

17 Zhong Y Q, Liu X M, Zhai D M, Jiang J J, Ji X Y. Shadows can be dangerous: Stealthy and effective physical-world adversarial attack by natural phenomenon. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans, USA: IEEE, 2022. 15324−15333

18 Chen P Y, Zhang H, Sharma Y, Yi J F, Hsieh C J. ZOO: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models. In: Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security. Dallas, USA: ACM, 2017. 15−26

19 Ilyas A, Engstrom L, Athalye A, Lin J. Black-box adversarial attacks with limited queries and information. In: Proceedings of

the 35th International Conference on Machine Learning (ICML). Stockholm, Sweden: PMLR, 2018. 2137−2146

20 Uesato J, O′donoghue B, Kohli P, Oord A. Adversarial risk and the dangers of evaluating against weak attacks. In: Proceedings of the 35th International Conference on Machine Learning (ICML). Stockholm, Sweden: PMLR, 2018. 5025−5034

21 Li Y D, Li L J, Wang L Q, Zhang T, Gong B Q. NATTACK: Learning the distributions of adversarial examples for an improved black-box attack on deep neural networks. In: Proceedings of the 36th International Conference on Machine Learning (ICML). Long Beach, USA: PMLR, 2019. 3866−3876

22 Brendel W, Rauber J, Bethge M. Decision-based adversarial attacks: Reliable attacks against black-box machine learning models. In: Proceedings of the 6th International Conference on Learning Representations (ICLR). Vancouver, Canada: OpenReview.net, 2018.

23 Dong Y P, Su H, Wu B Y, Li Z F, Liu W, Zhang T, et al. Efficient decision-based black-box adversarial attacks on face recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, USA: IEEE, 2019. 7706−7714

24 Rahmati A, Moosavi-Dezfooli S M, Frossard P, Dai H Y. GeoDA: A geometric framework for black-box adversarial attacks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA: IEEE, 2020. 8443−8452

25 Dong Y P, Liao F Z, Pang T Y, Su H, Zhu J, Hu X L, et al. Boosting adversarial attacks with momentum. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE, 2018. 9185−9193

26 Lin J D, Song C B, He K, Wang L W, Hopcroft J E. Nesterov accelerated gradient and scale invariance for adversarial attacks. In: Proceedings of the 8th International Conference on Learning Representations (ICLR). Addis Ababa, Ethiopia: OpenReview.net, 2020.

27 Xie C H, Zhang Z S, Zhou Y Y, Bai S, Wang J Y, Ren Z, et al. Improving transferability of adversarial examples with input diversity. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, USA: IEEE, 2019. 2725−2734

28 Dong Y P, Pang T Y, Su H, Zhu J. Evading defenses to transferable adversarial examples by translation-invariant attacks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, USA: IEEE, 2019. 4307−4316

29 Wang X S, He K. Enhancing the transferability of adversarial attacks through variance tuning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville, USA: IEEE, 2021. 1924−1933

30 Shi B G, Bai X, Yao C. An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017, **39**(11): 2298−2304

31 Graves A, Fernández S, Gomez F, Schmidhuber J. Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks. In: Proceedings of the 23rd International Conference on Machine Learning. Pittsburgh Pennsylvania, USA: ACM, 2006. 369−376

32 Song C Z, Shmatikov V. Fooling OCR systems with adversarial text images. arXiv preprint arXiv: 1802.05385, 2018.

33 Jiang H, Yang J T, Hua G, Li L X, Wang Y, Tu S H, et al. FAWA: Fast adversarial watermark attack. *IEEE Transactions on Computers*, DOI: 10.1109/TC.2021.3065172

34 Xu X, Chen J F, Xiao J H, Gao L L, Shen F M, Shen H T. What machines see is not what they get: Fooling scene text recognition models with adversarial text images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA: IEEE, 2020. 12301−12311

35 Xu Y K, Dai P W, Li Z K, Wang H J, Cao X C. The best protection is attack: Fooling scene text recognition with minimal pixels. *IEEE Transactions on Information Forensics and Security*, 2023, **18**: 1580−1595

36 Zhang J M, Sang J T, Xu K Y, Wu S X, Zhao X, Sun Y F, et al. Robust CAPTCHAs towards malicious OCR. *IEEE Transactions on Multimedia*, 2021, **23**: 2575−2587

37 Ding K Y, Hu T, Niu W N, Liu X L, He J P, Yin M Y, et al. A novel steganography method for character-level text image based on adversarial attacks. *Sensors*, 2022, **22**(17): Article No. 6497

38 Yang M K, Zheng H T, Bai X, Luo J B. Cost-effective adversarial attacks against scene text recognition. In: Proceedings of the 25th International Conference on Pattern Recognition (ICPR). Milan, Italy: IEEE, 2021. 2368−2374

39 Chen L, Xu W. Attacking optical character recognition (OCR) systems with adversarial watermarks. arXiv preprint arXiv: 2002.03095, 2020.

40 Xu Chang-Kai. Research on Adversarial Example Defense and Generation Algorithm Based on Deep Learning [Master thesis], Beijing Jiaotong University, China, 2023.
(徐昌凯. 基于深度学习的对抗样本防御与生成算法研究 [硕士学位论文], 北京交通大学, 2023.)

41 Hu Z H, Huang S Y, Zhu X P, Sun F C, Zhang B, Hu X L. Adversarial texture for fooling person detectors in the physical world. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans, USA: IEEE, 2022. 13297−13306

42 Hu Y C T, Chen J C, Kung B H, Hua K L, Tan D S. Naturalistic physical adversarial patch for object detectors. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). Montreal, Canada: IEEE, 2021. 7828−7837

43 Huang L F, Gao C Y, Zhou Y Y, Xie C H, Yuille A L, Zou X Q, et al. Universal physical camouflage attacks on object detectors. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA: IEEE, 2020. 717−726

44 Athalye A, Engstrom L, Ilyas A, Kwok K. Synthesizing robust adversarial examples. In: Proceedings of the 35th International Conference on Machine Learning (ICML). Stockholm, Sweden: PMLR, 2018. 284−293

徐昌凯 2023 年获得北京交通大学计算机与信息技术学院硕士学位. 主要研究方向为计算机视觉与对抗攻防技术.
E-mail: 20120320@bjtu.edu.cn
(**XU Chang-Kai** He received his master degree from the School of Computer and Information Technology, Beijing Jiaotong University in 2023. His research interest covers computer vision and adversarial attack-defense techniques.)

冯卫栋 北京交通大学硕士研究生. 2022 年获得中国地质大学 (北京) 学士学位. 主要研究方向为计算机视觉与对抗攻防技术.
E-mail: wdfeng@bjtu.edu.cn
(**FENG Wei-Dong** Master student at Beijing Jiaotong University. He received his bachelor degree from China University of

Geosciences (Beijing) in 2022. His research interest covers computer vision and adversarial attack-defense techniques.)

张淳杰 北京交通大学计算机与信息技术学院教授. 主要研究方向为图像处理与理解, 计算机视觉和多媒体数据处理与分析. 本文通信作者.
E-mail: cjzhang@bjtu.edu.cn
(**ZHANG Chun-Jie** Professor at the School of Computer and Information Technology, Beijing Jiaotong University. His research interest covers image processing and understanding, computer vision and multimedia data processing and analysis. Corresponding author of this paper.)

郑晓龙 中国科学院自动化研究所研究员. 主要研究方向为大数据与社会计算和多模态数据感知与理解.
E-mail: xiaolong.zheng@ia.ac.cn
(**ZHENG Xiao-Long** Professor at the Institute of Automation, Chinese Academy of Sciences. His research interest covers big data and social computing, and multimodal data sensing and understanding.)

张 辉 北京航空航天大学交通科学与工程学院教授. 主要研究方向为车辆动力学及其控制, 鲁棒控制, 网络控制系统以及信号处理.
E-mail: huizhang285@buaa.edu.cn
(**ZHANG Hui** Professor at the School of Transportation Science and Engineering, Beihang University. His research interest covers vehicle dynamics and control, robust control, networked control systems, and signal processing.)

王飞跃 中国科学院自动化研究所研究员, 复杂系统管理与控制国家重点实验室主任, 中国科学院大学中国经济与社会安全研究中心主任, 澳门科技大学特聘教授. 主要研究方向为社会计算, 平行智能与知识自动化.
E-mail: feiyue.wang@ia.ac.cn
(**WANG Fei-Yue** Professor at the Institute of Automation, Chinese Academy of Sciences, director of the State Key Laboratory for Management and Control of Complex Systems, director of China Economic and Social Security Research Center at University of Chinese Academy of Sciences, distinguished professor at Macau University of Science and Technology. His research interest covers social computing, parallel intelligence, and knowledge automation.)