

基于熵及随机擦除的针对目标检测物理攻击的防御

高红超¹, 周广治^{1,2}, 戴 娇¹, 李昭星¹, 韩冀中¹

¹中国科学院信息工程研究所 北京 中国 100093

²中国科学院大学 网络空间安全学院 北京 中国 100049

摘要 物理攻击通过在图像中添加受扰动的对抗块使得基于深度神经网络 (DNNs) 的应用失效, 对DNNs的安全性带来严重的挑战。针对物理攻击方法生成的对抗块与真实图像块之间的信息分布不同的特点, 本文提出了能有效避免现有物理攻击的防御算法。该算法由基于熵的检测组件 (Entropy-based Detection Component, EDC) 和随机擦除组件 (Random Erasing Component, REC) 两部分组成。EDC 组件采用熵值度量检测对抗块并对其灰度替换。该方法不仅能显著降低对抗块对模型推理的影响, 而且不依赖大规模的训练数据。REC 模块改进了深度学习通用训练范式。利用该方法训练得到的深度学习模型, 在不改变现有网络结构的前提下, 不仅能有效防御现有物理攻击, 而且能显著提升图像分析效果。上述两个组件都具有较强的可转移性且不需要额外的训练数据, 它们的有机结合构成了本文的防御策略。实验表明, 本文提出的算法不仅能有效的防御针对目标检测的物理攻击(在 Pascal VOC 2007 上的平均精度 (mAP) 由 31.3% 提升到 64.0% 及在 Inria 数据集上由 19.0% 提升到 41.0%), 并且证明算法具有较好的可转移性, 能同时防御图像分类和目标检测两种任务的物理攻击。

关键词 对抗样本; 物理攻击; 对抗块; 对抗防御; 目标检测

中图分类号 TP181 DOI号 10.19363/J.cnki.cn10-1380/tn.2023.01.09

Defense Against Physical Attacks on Object Detection Based on Entropy and Random Erasing

GAO Hongchao¹, ZHOU Guangzhi^{1,2}, DAI Jiao¹, LI Zhaoxing¹, HAN Jizhong¹

¹ Institute of Information Engineering, Chinese Academy of Sciences, Beijing 100093, China

² School of Cyber Security, University of Chinese Academy of Sciences, Beijing 100049, China

Abstract Existing physical attack techniques for deep learning models mislead the deep neural networks (DNNs) inference process by adding perturbed adversarial patches to the attacked image, thereby making the application which based on DNNs invalid to achieve the purpose of the attack. Such attack methods are easy to implement and highly transferable, which bring serious challenges to the security of DNNs. The quantities of information contained in the adversarial patches generated by existing physical attack methods is usually higher than that of real natural scene image patches. Using this phenomenon, this paper proposes a defense algorithm with strong versatility and obvious defense effect. This algorithm consists of an Entropy-based Detection Component (EDC) and a Random Erasing Component (REC). EDC component uses entropy measurement to detect the perturbed adversarial patches and replace it with gray patches. This component can not only significantly reduce the impact of adversarial patches on model inferencing, but also does not rely on large-scale training data. The REC component improves the general training paradigm of DNNs. The deep learning model trained by REC can not only effectively defend against existing physical attacks, but also improve the image analysis effect significantly without changing the network structure. The above two components have strong transferability and do not need additional training data. Furthermore, we propose an efficient and transferable defense algorithm through the organic combination of two components. The experimental results on different data of the two image analysis tasks show that the defense algorithm proposed in this paper can not only effectively defend against physical attacks against object detection (the average accuracy (mAP) on Pascal VOC 2007 is increased from 31.3% to 64.0%, and on the Inria dataset is increased from 19.0% to 41.0%), but also prove that the algorithm has good transferability, which can defend against physical attacks of both image classification and object detection tasks.

Key words adversarial examples; physical attacks; adversarial patch; adversarial defense; object detection

通讯作者: 韩冀中, 博士, 正研级高级工程师, hanjizhong@iie.ac.cn。

本课题得到科技创新 2030-“新一代人工智能”重大项目(No. 2020AAA0140000)资助。

收稿日期: 2020-01-16; 修改日期: 2020-03-13; 定稿日期: 2022-12-07

1 引言

深度神经网络 (DNNs) 已经广泛应用于安全领域的实际场景中, 包括自动驾驶^[1]、智能监控^[2]及人脸识别^[3-4]等。随着应用的普及, 其安全性也越来越受到学术界和工业界的关注。许多研究工作通过生成对抗样本实现对物理世界中的计算机视觉识别系统的攻击。比如, 自动驾驶系统将停止标识错误地识别为限速 45^[1]; 通过穿着特制衣服躲避智能监控系统^[2]; 戴着特制帽子^[3]或者眼镜^[4]欺骗最先进的人脸识别系统等。上述工作加深了人们对人工智能安全性和可靠性的担忧。

通过向样本 (图像) 中添加细微的、人眼不易察觉的像素级扰动, 即可成功欺骗原有的图像分析模型^[5], 这种扰动后的样本被称为“对抗样本”。目前, 针对计算机视觉中图像分类和目标检测两个基本任务, 存在两种对抗样本生成方式: 添加人眼不易察觉的像素级扰动^[5-8]和在图像固定区域添加对抗块的物理攻击^[1-2,4,9-10]。相对于物理攻击, 像素级扰动出现较早, 且广泛涉及图像分类^[5-7]及目标检测网络^[8]。而物理攻击的早期研究工作聚焦于图像分类网络, 如 Adversarial patch^[9]和 LaVAN^[10]。这类工作通过优化三元组 (误分类置信度, 目标类别及对抗块位置) 的形式来生成对抗块。最近, 针对目标检测的物理攻击不断涌现, 例如人脸识别^[3-4]及人体检测^[2]等。这些攻击对于特定的模型或应用已经取得很高的攻击成功率, 因此对相关防御方法提出了更大的挑战。

与攻击方法对应, 早期的防御方法主要聚焦于防御像素级扰动^[6,11-14], 近年来对于物理攻击的防御已受到越来越多的关注。现有防御算法主要提升图像分类网络^[15-17]的鲁棒和安全性, 并不能有效提升目标检测算法的安全性。针对图像分类任务的物理攻击方法通过生成小面积对抗块并融合到图像的背景区域。相应的防御策略首先检测对抗块区域, 然后直接进行灰度替换, 从而降低对抗块对分类网络的误导。然而, 对目标检测网络的物理攻击, 其生成的对抗块通常融合到目标区域内。若采用上述算法, 目标区域内关键特征信息随着灰度替换而消失, 直接降低目标分类和区域回归的准确率。因此, 针对目标检测物理攻击防御挑战更大。本文针对于目标检测网络的物理攻击, 提出了有效的防御算法。

图 1 展示了现有物理攻击方法^[24]利用不同数据集产生的对抗样本。从图可以看出, 图像中对抗块相比于真实图像颜色更为鲜亮且相邻像素变化更为明显。图 2 展示了对抗样本上基于熵值变化的信息分

布情况。分析可知, 对抗块包含的熵值远高于真实图像区域。因此, 基于对抗块特有的信息分布, 本文针对当前物理攻击提出一种有效的防御算法。首先, 考虑到对抗块含有较高且集中的信息分布, 本文提出了一个基于熵的检测组件 (Entropy-based Detection Component, EDC) 检测和定位图像中的对抗块并利用灰度块进行相应替换。上述操作不可避免的导致原有图像或目标替换区域的信息缺失, 进而降低目标检测性能。为解决该问题, 本文提出基于随机擦除组件 (Random Erasing Component, REC) 以促使深度学习模型利用图像或目标的全局信息进行推理分析, 克服由于采用 EDC 进行灰度替换带来的图像重要信息缺失的问题。相比于常规的训练流程, 基于 REC 训练得到模型的分析效果和鲁棒性显著提升。

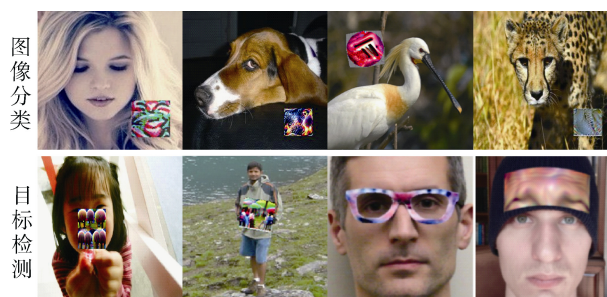


图 1 针对图像分类及目标检测的物理攻击产生的对抗样本

Figure 1 Adversarial examples generated by different physical attacks on image classification and object detection networks

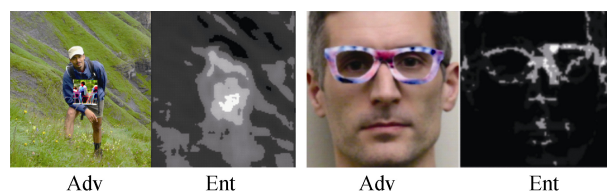


图 2 对抗图像及其基于熵的对抗区域检测结果
Figure 2 Adversarial images and their detection results based on entropy

现有物理攻击算法生成的对抗块形态各异且生成所需的模型训练周期长, 通过枚举对抗块针对性防御代价过高。本文基于传统图像处理技术提出 EDC 组件, 因此, 该算法不依赖大量的训练样本 (对抗块)。更重要的是, 结合本文提出的两个组件, 其他相关的防御算法、图像分类及目标检测网络的性能都可以进一步提升。

总的来说, 本文具有以下贡献:

(1) 提出基于熵的检测组件 (EDC) 度量图像中的

信息分布, 检测和定位图像中的对抗块, 降低对抗块对 DNNs 推理的误导。

(2) 提出基于随机擦除组件(REC) 的训练策略, 训练出的深度学习模型不仅能有效预防物理攻击, 而且能提升模型的图像分析效果。

(3) 基于以上组件提出高效且通用性强的针对图像分类和目标检测网络物理攻击的防御算法。

2 相关工作

攻击与防御是一个相互学习和促进的过程。图像块级的物理攻击和防御, 像素级的攻击和防御之

间可以相互借鉴。本文从图像块级和像素级两个方面介绍相关研究工作。

2.1 物理攻击

图 3 展示了现有物理攻击具有相似的攻击流程。首先, 随机初始化特定尺寸的图像块并贴附到图像的固定区域。然后, 设置损失函数通过迭代训练目标网络更新图像块。最后, 生成具有攻击效果的对抗块。这些攻击方法的主要工作在于设计不同的损失函数, 生成对抗块融合到图像中, 误导目标网络或应用推理过程, 实现攻击的目的。下面针对目标检测和图像分类的物理攻击进行介绍。

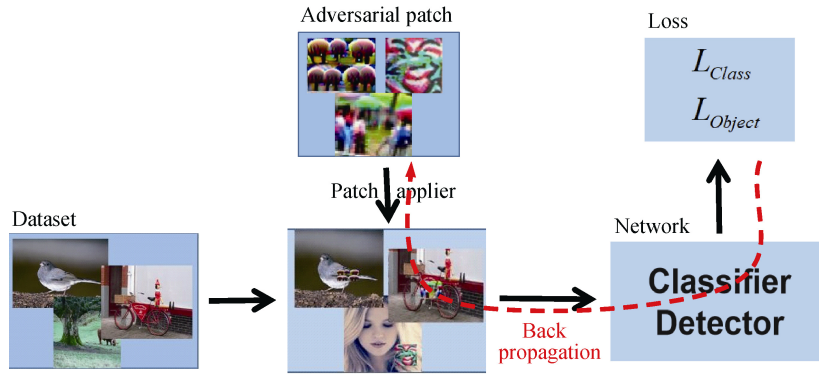


图 3 现有物理攻击方法的攻击流程

Figure 3 Overview of attack pipeline for existing physical attacks

2.1.1 针对目标检测的物理攻击

Fooling automated surveillance^[2] 该方法设计了一个能生成可打印的对抗块的系统。在物理世界中, 当将可打印的对抗块放置于人体特定位置上可显著降低目标检测结果, 从而躲避智能监控系统。在实际应用中具有很大的意义和警示性。为了达到可打印的效果, 引入可打印损失函数:

$$L_{nps} = \sum_{p_{patch} \in P} \min_{c_{print} \in C} |p_{patch} - c_{print}| \quad (1)$$

p_{patch} 是 $patch$ 中的像素, c_{print} 是可打印颜色集合 C 中的一种颜色值。为了增加对抗块与周围像素之间的平滑度, 使对抗块在人眼看起来与周围图像更加融合, 引入平滑损失函数:

$$L_{tv} = \sum_{i,j} \sqrt{((p_{i,j} - p_{i+1,j})^2 + (p_{i,j} - p_{i,j+1})^2)} \quad (2)$$

L_{tv} 值越低表示相邻像素之间的差值越小。最后, 为了达到隐藏目标的目的, 将其网络输出的最大目标分数作为第三部分损失。其总体损失函数如下:

$$L = \alpha L_{nps} + \beta L_{tv} + L_{obj} \quad (3)$$

L_{nps} 为不可打印分数, L_{tv} 为图像像素的总方差, L_{obj} 为网络输出图像的最大目标分数。

AdvHat^[3] 该方法可以生成可打印的对抗贴条(对抗块)。研究者带上附有对抗条贴的帽子就能实现欺骗先进的人脸识别系统 ArcFace。通过在帽子贴附训练时, 通过计算含有贴附对抗条图像的向量 e_x 与系统输出人脸的向量 e_a 之间的 \cosine 相似度, 并将其作为第一部分损失, 保证含有对抗贴条图像的预测结果与原始图像的预测结果不同, 从而达到欺骗系统的目的。损失函数设置如下:

$$L_{sim}(x, a) = \cos(e_x, e_a) \quad (4)$$

为了使对抗贴条更加平滑且对不同的图像变换更加鲁棒, 引入如下损失:

$$TV(x) = \sum_{i,j} ((x_{i,j} - x_{i+1,j})^2 + (x_{i,j} - x_{i,j+1})^2) \quad (5)$$

最终的损失函数设置如下:

$$L_{final}(x, a) = L_{sim}(x, a) + \lambda \cdot TV(x) \quad (6)$$

其中 λ 为 TV 损失的权重, 论文中设置的为 $1e-4$ 。

2.1.2 针对图像分类的物理攻击

LaVAN^[10] 该方法同样引入了一个新的损失函数。通过在图像局部位置添加固定扰动, 达到攻击原有分类网络的目的。具体的, 在每次迭代训练中, 优化算法都实现网络输出从源类别向目标类别攻击有

效, 从而更新图像块并实现通用攻击的目的。其损失函数为:

$$\begin{aligned} \max_{x'} &= F(\bar{y} | x') - F(y | x') \\ \text{s.t.} & \|x - x'\|_{\infty} \leq \varepsilon, 0 \leq \varepsilon \leq 1 \end{aligned} \quad (7)$$

x' 代表对抗样本, y 表示原始类别, \bar{y} 表示目标类别。

Adversarial Patch^[9] 是由 Brown 等人提出的一种物理攻击方法。该方法提出了一个块操作函数 $A(p, x, l, t)$ 生成对抗块 \hat{p} , 操作函数 A 将变换集 T (平移、旋转及缩放等) 应用到图像块 p , 然后将其放置于图像 x 的 l 位置, 通过迭代训练来更新 p 。该方法能较为灵活的选择对抗块, 从而确保添加扰动的普适性。其损失函数设置如下:

$$\hat{p} = \arg \max_p E_{x \sim X, t \sim T, l \sim L} [\log \Pr(\hat{y} | A(p, x, l, t))] \quad (8)$$

X 表示输入图像集, p 表示初始块, \hat{y} 为目标类别, L 表示放置位置集及 T 表示对抗块的变换集合。

2.2 对抗防御

2.2.1 针对像素级扰动的防御

目前, 针对像素级扰动的防御方法主要分为三类。

1) 将生成的对抗样本重新加入训练集进行对抗训练^[6,12-13]。这类方法的原理是将攻击算法生成的对抗性样本合并到训练集中, 基于合成的数据集训练防御模型。这类方法实现简单和易于理解, 对于特定的对抗样本取得有效的防御效果。然而, 生成用于训练的对抗样本数量较少且只能针对于特定的攻击方法, 从而显著降低这类算法的防御效果和泛化能力。

2) 通过图像预处理实现对抗样本的检测和去除, 如 DefenseGAN^[19], PixelDefend^[10]等。主要通过对抗性样本的检测以及样本去噪使其失去原有的对抗性质。这类防御预处理机制对图像的处理可能会改变原始图像的语义信息, 从而降低原有模型的图像分析结果。

3) 通过改变训练流程或网络结构, 如标签平滑^[11,14], 正确类别的标签为 $1 - \varepsilon$, 其他错误类别的标签为 $\varepsilon / (N - 1)$ 。这类算法需要攻击应用的网络结构、训练数据和训练流程。神经网络依然被视为一个黑盒状态, 所以这一方向的进展比较缓慢。

总的来说, 现有这些防御方法对于特定的像素级别的攻击方法取得了有效的防御效果。然而, 防御方法的可迁移性差且防御成本高。

2.2.2 针对图像分类物理攻击的防御

现有针对物理攻击的防御算法主要聚焦于图像分类网络。该类攻击的防御方法和防御原理是相似的。首先, 根据分类结果创建对抗样本的显著映射。然后, 根据显著映射结果判断图像当中对抗块所在的位置。Simonyan 等人^[21]利用网络梯度生成图像热力图展示图像不同区域对于分类网络的重要程度。具体的, 当通过线性整流函数 (Rectified Linear Unit, ReLU) 回传影响信号时, 如果前向传递小于 0 则设置回传信号为 0。相似的, Zeiler 等人^[22]也创建了一个类似的显著映射, 当回传的时候如果信号是负的, 则输出为 0, 意味着忽略前向传递中通过 ReLU 单元的任何信息。

On Visible Adversarial Perturbations^[17] Jamie 等人发现对抗图像的显著性图通常在扰动位置周围具有密集簇。在对抗图像中, 图像分类结果几乎完全受这个小区域的影响, 而在自然图像中, 影响分类的像素分布则相对稀疏。基于如上发现, 通过建立图像的显著性图来检测图像当中的对抗区域。具体的, 将所有前向或后向经 ReLU 单元信号为负的全部归 0, 从而在网络进行图像分析时消除对抗区域信号的影响。

Local Gradients Smoothing^[15] Naseer 等人认为物理攻击在图像当中引入了高频噪声。因此, 算法通过保留对分类有重要意义的低频图像区域, 并抑制高频图像区域, 从而显著降低这种对抗性噪声的影响。具体地, 评估图像当中含有高噪声的区域并视其为对抗块区域, 然后在这些区域执行梯度平滑。通过公式 9 所示的梯度计算来评估图像当中的高频噪声区域:

$$\|\nabla x(a, b)\| = \sqrt{\left(\frac{\partial x}{\partial a}\right)^2 + \left(\frac{\partial x}{\partial b}\right)^2} \quad (9)$$

SentiNet^[16] Chou 等人主要基于 Grad-CAM (Gradient-weighted Class Activation Mapping)^[31]方法定位物理攻击的位置。首先根据输出及预测利用 Grad-CAM 提取相应 Mask, 然后结合类别提议的 Mask 来进一步获得更加精确的针对物理攻击的 Mask, 从而根据 Mask 来消除样本中对抗区域的影响。

3 方法

通过多轮迭代训练, 现有物理攻击方法生成的对抗块(可打印的贴画)能改变应用在物理世界中的基于 DNNs 的推理过程和结果。相比于真实自然场景图像, 该对抗块含有更多的高噪声且信息量更大。

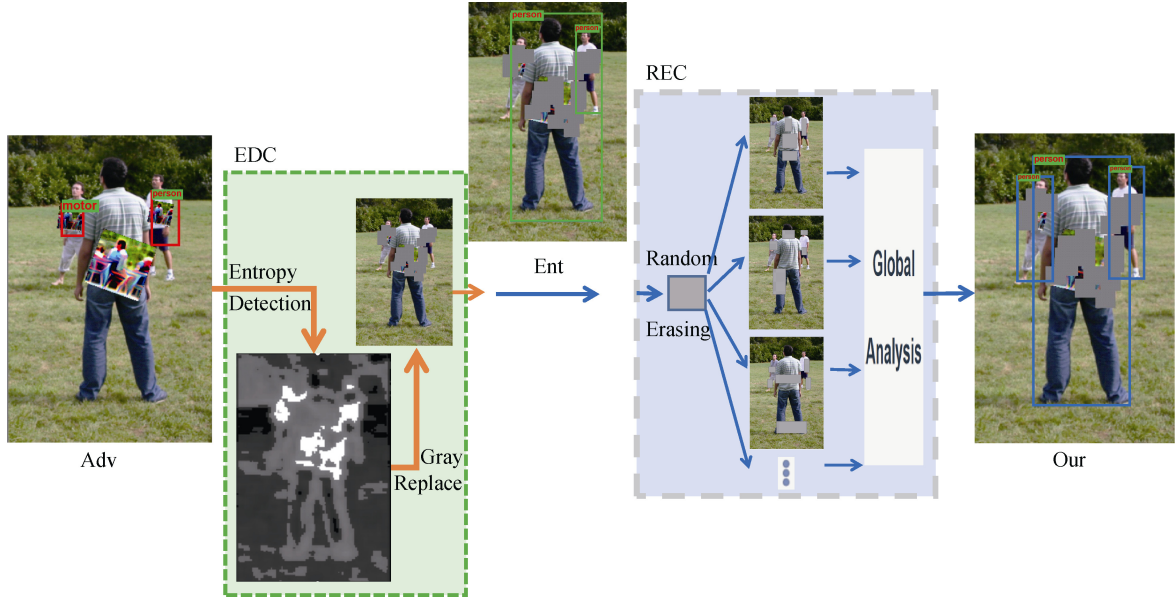


图 4 防御算法框架。对抗图像的目标检测结果与经过 EDC 和 REC 处理之后的目标检测结果

Figure 4 Overview of our proposed defense method. Object detection results on the adversarial image and the adversarial image processed by EDC and REC

基于该性质, 本文提出了一种针对该类物理攻击的防御算法, 算法结构如图 4 所示。

本文提出的防御算法包含两个组件: 基于熵的检测组件 (EDC) 及用于网络鲁棒训练的随机擦除组件 (REC)。给定一张对抗图像(Adv), 首先利用 EDC 去度量图像中的信息分布, 利用信息分布检测并定位到对抗区域, 并对检测到的对抗区域执行灰度值替换 (Ent) 操作。然后, 利用经 REC 鲁棒训练后的目标检测网络去检测灰度替换后的图像 (Our)。

3.1 基于熵的检测 (EDC) 组件

物理攻击产生的对抗块包含的信息明显高于图像当中真实块, 本文利用传统信息测量—离散熵^[23]度量图像中的信息分布。离散熵在图像处理中的作用已经被证明^[24]。

给定大小为 $M \times N$ 的图像, 利用式 10 和式 11 计算每个通道含有的信息量(熵值 E)大小:

$$f_i = s_i / (M \times N), i = 0, 1, \dots, 255 \quad (10)$$

$$E = -\sum_{i=0}^{255} f_i \log_2(f_i) \quad (11)$$

其中, s_i 表示灰度级为 i 的像素数量, f_i 表示灰度级为 i 的像素占图像总像素个数的比例。对于 RGB 类型图像, 其熵为三个通道熵的均值。

本文基于滑动窗口计算输入图像的熵, 根据设定的阈值 τ , 将熵值大于 τ 的窗口视为图像当中的对抗区域 (对抗块)。物理攻击的目标是仅需少量的对抗块即可达到欺骗网络且对抗块的熵值较自然场景

差别较大。通过自适应阈值的方法确定当前攻击的阈值 τ , 根据 τ 判定攻击区域。

图 5 展示了利用 EDC 在对抗样本上的检测效果。越亮的区域越大概率表示攻击区域。当检测并定位到图像中的对抗块后, 利用相同大小的灰度掩码块进行替换, 从而降低对抗块给目标检测网络推理的误导。

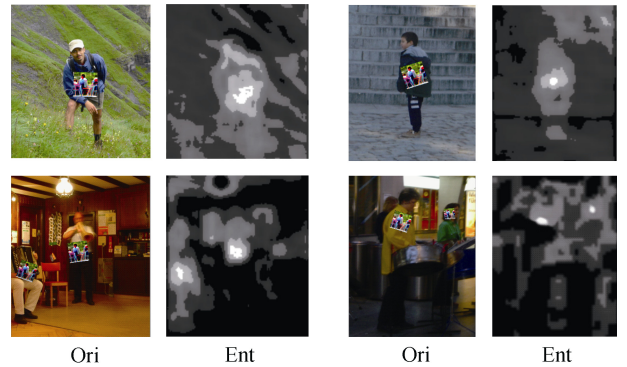


图 5 基于熵的对抗区域检测效果

Figure 5 The detection results of the adversarial regions under the entropy detection

3.2 随机擦除组件(REC)

通过 EDC 检测到图像当中的对抗块后, 使用相同大小的灰度块进行替换, 该操作必然会去除该区域的原有图像信息, 从而影响网络预测结果, 尤其是对于目标检测网络。图像分类的物理攻击中对抗块通常位于背景且相对较小, 而对于目标检测的物

理攻击, 其对抗块通常位于目标区域以内。EDC 灰度替换操作会去除真实目标区域内的关键信息, 进而显著降低检测任务的类别判断和目标框回归的准确率。本文提出的 REC 使深度学习模型利用整个目标区域的信息进行推理分析, 而不仅仅是目标区域的关键区域, 从而减弱目标区域内有效信息抹除对网络检测的影响。

在网络训练时, 对于每一张大小为 $H \times W$ 的图像 I , 随机擦除图像中的 N 个区域 ($N = \text{rand}(0, m)$)。本文用矩形框表示随机擦除区域。

首先, 根据图像的长宽, 设置每一个矩形框的长 h 宽 W 。

$$h = \alpha_h \times H$$

$$w = \alpha_w \times W$$

其中, α_h 和 α_w 为随机缩放系数

$$\alpha_h = \text{rand}(r1, r2), 0 < r1 < r2 < 1$$

$$\alpha_w = \text{rand}(r1, r2), 0 < r1 < r2 < 1$$

然后, 随机生成图像中位置, 并随机擦除 $h \times w$ 的矩形区域。

程序如算法 1 所示。不同于算法[30], 本文提出的擦除区域设置取决于输入图像尺寸, 且每张图像擦除区域的数量也是随机设置, 因此本文算法更加有效且易于实现。

算法 1. 随机擦除

输入: 图像 I

图像长 H 和宽 W

最大擦除矩形的数量 m

擦除矩形的长和宽所占比例范围($r1, r2$)

输出: 擦除后的图像 I^*

过程 1. 生成含有随机灰度块的图像

```

1:   $n = \text{Rand}(0, m)$ 
2:  FOR  $i = 0$  to  $n$  DO:
3:       $w = \text{Rand}(r1, r2) * W$ 
4:       $h = \text{Rand}(r1, r2) * H$ 
5:       $x = \text{Rand}(0, W-w)$ 
6:       $y = \text{Rand}(0, H-h)$ 
7:       $I[x:x+w, y:y+h] = 128$ 
8:       $I^* = I$ 
9:  END FOR
10: RETURN  $I^*$ 

```

3.2.1 图像分类网络的随机擦除训练

图像分类网络是通过分析输入的整张图像来输出预测类别。因此, 本文在整张输入图像上执行随机擦除用于鲁棒性训练。除了在每批次中的每一张图

像上执行随机擦除算法之外, 图像分类的网络以及其他训练设置与对比基准相同。本文使用随机擦除训练 ResNet18^[28]、ResNet50^[28]及 Inception V3^[29]图像分类网络。

3.2.2 目标检测网络的随机擦除训练

目标检测网络的任务是定位图像当中的每一个目标并进行识别。目标检测网络训练时, 每一个目标的边界框和目标类别都是给定的。因此, 本文利用 REC 在每一个目标框内执行随机擦除算法。具体的, 在训练目标检测网络时, 当前批次的所有图像上的每个目标框都执行 REC 操作, 其他的检测网络结构及训练设置与原有算法相同。本文使用随机擦除来训练 Tiny YOLO V2^[27]及 Darknet^[27]目标检测网络。

4 实验

为验证本文防御算法有效性和可迁移性, 本文进行如下实验设置。首先介绍基于熵的对抗块检测组件(EDC)及用于鲁棒训练的随机擦除组件(REC)所使用的数据集。然后, 分别验证每个组件的有效性和可转移性。最后, 验证本文提出的将两个组件合并后的防御算法的有效性。

设置 本文使用基于 Pascal VOC 数据集预训练的 Tiny YOLO 及 Darknet 网络模型用于目标检测任务, 使用基于 ImageNet 数据集预训练的 ResNet18、ResNet50 及 Inception V3 网络模型用于图像分类任务。所有的实验均基于 NVIDIA Tesla M40 GPU 并使用 Pytorch 实现。

4.1 数据集

4.1.1 EDC 所使用的数据集

Pascal VOC^[25]是目标检测的基准数据集, 其包含 20 个类别。本文从 Pascal VOC 2007(VOC 07) 和 Pascal VOC 2012 的训练集中选出包含有 “people” 类别的共 4015 张图像作为训练集 VOC_P 。测试阶段, 本文利用 Pascal VOC 2007(Ori)的测试集共 4952 张图像合成新的测试集(Adv)。具体地, 对于 Ori 中的每张图像粘贴对抗块合成新的图像。本文采用方法[2]利用训练集 VOC_P 生成对抗块。

Inria 数据集^[2]是针对于全身的行人, 其训练集包含 615 张图像及其测试集包含 288 张图像。本文利用其训练集产生对抗块。测试阶段, 利用其测试集合成对抗样本集。

4.1.2 REC 所使用的数据集

目标检测数据集选用 Pascal VOC 和 Inria 两种。对于 Pascal VOC, 本文利用 Pascal VOC 2007 和 Pascal VOC 2012 数据集。在训练阶段, 本文利用两

个数据集的训练集总共 8218 张图像。

图像分类: 本文利用 ImageNet^[26]数据集, 该数据集包含 1000 个类别及训练集和测试集分别包含 130 万及 5 万张图像。

4.2 EDC 的有效性和可转移性

4.2.1 有效性

为了验证 EDC 的有效性, 采用目标检测算法评价指标 mAP 进行度量。相比未受攻击(Ori)图像上的测试结果, mAP 下降程度越小, 证明攻击效力越低, EDC 的对抗块检测越准。

对 Pascal VOC 和 Inria 两个数据集分别按照如下流程进行处理: 首先利用物理攻击算法[2]基于训练集生成对抗块, 把对抗块融合到测试集(Ori)中的每张图像上构成攻击测试集 Adv。采用 EDC 进行对抗块的检测和消除, 得到测试集 Ent。

如表 1 所示, 对抗样本经 EDC 处理之后 (Ent) 其检测结果具有较大提升。例如, 在 Inria 数据集上, Darknet 网络的 mAP 从 19.0% (Adv) 提升到 35.9% (Ent)。Tiny YOLO 网络的 mAP 从 17.5%提升到了 31.1%。类似的, 本文算法在 VOC 07 数据集上检测结果也可以获得有效的提升。结果表明 EDC 能够有效的防御此类型的物理攻击, 一定程度上证明了, EDC 检测对抗块的准确性和鲁棒性。另外, 从表中可知参数量大的检测网络 Darknet 相比较小的检测网络 Tiny YOLO 网络更加鲁棒。

表 1 不同处理流程下的目标检测结果(mAP)

Table 1 Object detection results (mAP) of different processing

Dataset	Network	Ori	Adv	Ent
Inria	Darknet	61.4	19.0	35.9
	Tiny YOLO	53.3	17.5	31.1
Pascal VOC 07	Darknet	69.1	31.3	48.1
	Tiny YOLO	56.8	20.3	27.2

图 6 展示了利用预训练的 Darknet 检测网络在对抗样本(Off-Adv)和 EDC 处理后的图像上的检测效果(Off-Ent)。可以发现, 经过 EDC 处理后的图像 (Off-Ent), Darknet 在每一类的检测结果相比于对抗样本(Off-Adv)都具有有效的提升。尤其对于目标尺寸比较大或者特征比较明显的类别, 如对于 “dog” 和 “horse” 类别其检测结果有近 30%的提升。

为验证图像窗口大小、步长及阈值三个超参数对 EDC 检测的影响, 我们通过可视化不同窗口大小、步长及不同熵阈值下的对抗区域检测效果, 如图 7 和图 8 所示。从图可以得知, 对抗块的熵值明显高

于真实图像块, 熵值度量在一定范围内对不同窗口和步长较为鲁棒, 然而熵阈值的选取对检测结果有一定程度的影响。需要针对不同的攻击方法设计特定的熵阈值。本文利用自适应阈值的方法确定当前攻击的阈值 σ 范围, 在范围内进行阈值的确定。在本文算法中, 设置窗口大小 20*20、步长 3*3 及熵的阈值 7.1。

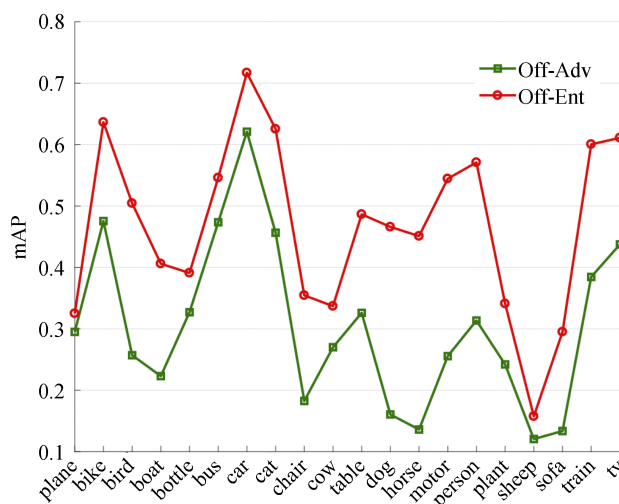


图 6 不同处理流程下每类的目标检测结果(mAP)

Figure 6 Object detection results (mAP) of each category under different processing

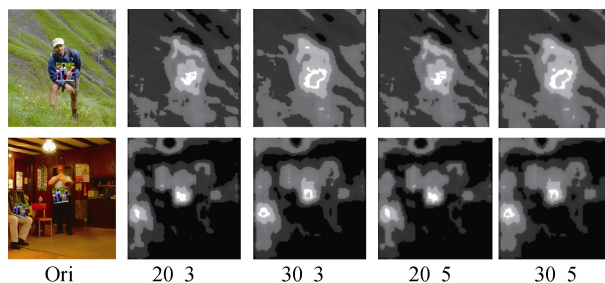


图 7 不同滑动窗口大小(20 或 30)及不同步长(3 或 5)下对抗区域检测效果

Figure 7 Detection results of the adversarial region at different sliding window sizes (20 or 30) and strides(3 or 5)

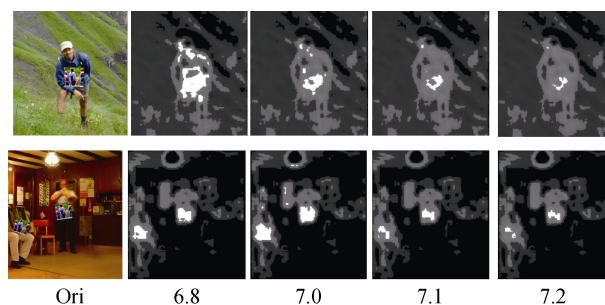


图 8 不同熵阈值下对抗区域的检测效果图

Figure 8 Detection results of the adversarial region under different entropy thresholds

4.2.2 可转移性

本节验证 EDC 对于不同物理攻击防御的可转移性。图 9 展示了基于攻击方法^[4]产生的对抗样本和基于 EDC 组件的对抗区域检测结果。越亮的区域, 意味着越大概率地表示对抗区域。可以看出本文算法能检测出图像中所有的对抗区域。



图 9 对抗样本^[4]及其对抗区域检测结果

Figure 9 Detection results of the adversarial regions on the adversarial images^[4]

图 10 展示了对攻击 ArcFace 人脸识别系统^[3]中对抗块的检测效果。可知, 对抗块区域的熵值明显高于图像中其他区域。且在对抗块的不同区域含有不同的信息分布, 可产生不同的攻击效果。本文仅需要找到具有最大信息分布的区域并执行灰度块替换, 便可降低物理攻击的成功率。



图 10 对抗样本^[3]及其对抗区域检测结果

Figure 10 Detection result of the adversarial region on the adversarial image^[3]

对于含有特定身份标识的区域, 如眼睛、嘴角等, 由于其熵值比较高, 存在较高的被 EDC 组件检测并灰度替换的风险。关键目标区域内关键信息的抹除, 不可避免的降低目标检测算法的检测性能。然而, 经过 REC 组件训练得到的深度学习模型能利用图像中其他区域的信息辅助推理, 一定程度上降低了由 EDC 错误检测带来的风险。

4.3 REC 的有效性

本节在目标检测及图像分类两个任务上展示基于 REC 训练模型的原始测试集上以及对抗图像上的分析结果, 验证基于 REC 训练的有效性和鲁棒性。

4.3.1 目标检测

为了验证 REC 对于目标检测网络的有效性, 分别基于 Tiny YOLO^[27]及 Darknet^[27]网络, 对比基准模型(Official)及经 REC 训练后的模型在 Pascal VOC 和

Inria 测试集(Ori)上的目标检测结果。为验证 REC 对物理攻击鲁棒性。首先, 利用训练集生成对抗块。然后, 对 Ori 中的每张图像粘贴生成的对抗块制作数据集(Adv); 最后, 在 Adv 数据集上通过目标检测准确率下降幅度验证 REC 防御的有效性。

表 2 展示两个模型在 Pascal VOC 2007 上的检测结果。可知, 经过 REC 鲁棒性训练后的网络在原始图像(Ori)上的检测准确率具有较大的提升。更重要的是, 该网络在对抗图像(Adv)上检测结果提升更大。例如, 在 Darknet 网络上, 本文模型的检测结果分别在原始图像和对抗图像上提升 7.6% 和 28.7%。证明 REC 不仅提升检测网络在原始图像上的检测性能, 同时对于对抗图像也具备有效的防御能力。同样的如表 3 所示, 在 Inria 数据集上也可以验证上述结论。

表 2 不同检测模型在 Pascal VOC 2007 测试集上的目标检测结果(mAP)

Network	Source	Ori	Adv
Darknet	Official	69.1	31.3
	REC	76.7	60.0
Tiny YOLO	Official	56.8	20.3
	REC	67.4	33.1

表 3 不同检测模型在 Inria 测试集上的目标检测结果(mAP)

Network	Source	Ori	Adv
Darknet	Official	61.4	19.0
	REC	61.8	20.0
Tiny YOLO	Official	53.3	17.5
	REC	58.6	23.0

图 11 和图 12 分别展示了 REC 对于原始未被攻击的图像 (Ori) 及对抗图像 (Adv) 中每一类的提升效果。如图 11 的“chair”和“table”类别, 其在真实环境当中容易被遮挡, 经 REC 鲁棒训练之后, 其检测结果提升近 15%。同样对于不易遮挡的“tv”类别, 其检测结果的提升也超过 10%。类似的, 对于图 12 中的对抗图像, 经 REC 鲁棒训练后的模型相比于官方模型 (Off), 其每一类的检测准确率都具有显著提升。上述结果表明, REC 不仅能提升目标检测的检测准确率, 而且能增强目标检测模型的鲁棒性, 降低模型被攻击的风险。

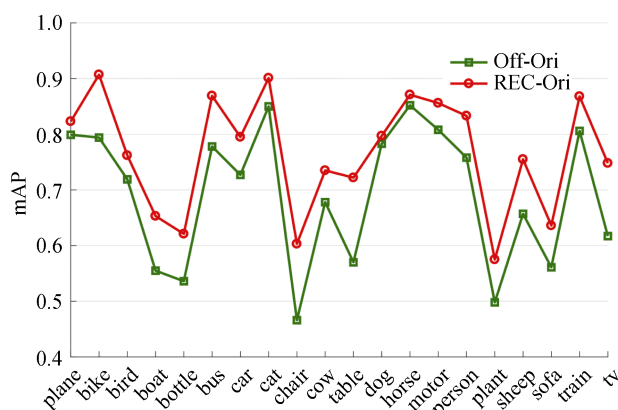


图 11 REC 在原始图像中每类的目标检测结果

Figure 11 The detection results of REC for each category in the original image

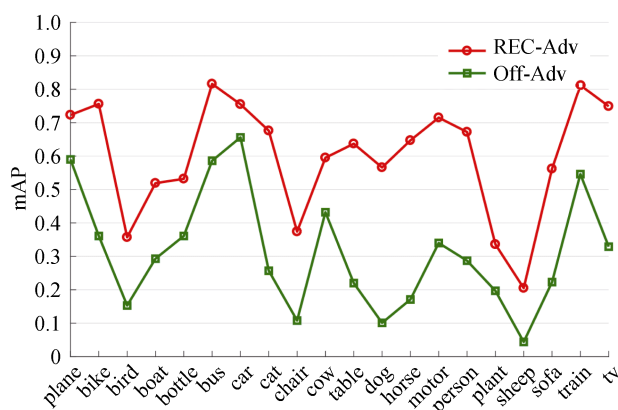


图 12 REC 在对抗图像中每类的目标检测结果

Figure 12 The object detection results of REC for each category in the original image

4.3.2 图像分类

为验证 REC 在图像分类网络上的有效性, 本文分别基于 ResNet18^[28], ResNet50^[28]及 InceptionV3^[29]网络, 对比官方预训练模型(Official)及经 REC 训练后的模型在相应测试集上的图像分类结果。

如表 4 所示, REC 对于不同分类网络的分类结果都具有一定的提升, 例如 ResNet50 的 Top-1 及 Top-5 分别提升了 2.9% 和 1.3%。同时表 5 展示了 REC 对于物理攻击^[9]的鲁棒性, 经 REC 训练后的模型, 相比于原模型物理攻击的成功率下降了 56.9%。证明了本文的 REC 不仅可以增强原有分类网络的图像推理能力, 同时对于图像分类物理攻击也具有一定的鲁棒性。

通过如上对于不同的目标检测网络、图像分类网络及相应的物理攻击在不同的数据集下所得出的实验结果, 证明了 REC 对于图像分析网络的性能提升以及对于相关物理攻击的鲁棒性。

表 4 不同图像分类模型在原始图像上的分类结果

Table 4 The classification results of different image classification models on the original image under different networks

Network	Source	Top 1	Top 5
ResNet18	Official	66.2	87.4
	REC	68.3	88.6
ResNet50	Official	72.2	91.2
	REC	75.1	92.5
Inception V3	Official	75.8	92.8
	REC	76.1	92.8

表 5 针对于 Inception V3 网络的有目标物理攻击^[9], 其不同模型下的攻击成功率Table 5 Targeted physical attack on Inception V3 network^[9], its attack success rate under different models

Network	Source	Attack
Inception V3	Official	67.1
	REC	10.2

4.4 本文算法的有效性(EDC+REC)

本节验证集成 EDC 和 REC 两个组件的防御算法针对目标检测任务展开的物理攻击的有效性。具体的, 给定一副图像, 首先利用 EDC 组件对其进行对抗区域检测并执行灰度替换。然后, 将替换后的图像输入到经 REC 鲁棒训练后的目标检测网络进行检测。根据测试集上检测准确率验证防御算法的有效性。

表 6 和表 7 展示了在 VOC 07 测试集上, 每一类在不同网络及处理流程下的目标检测结果。可知, 在添加对抗块后的测试集(Adv)上, 模型的平均检测准确率(mAP)有显著的下降。具体到每一类, 下降幅度差别较大, 例如在大目标“plane”上, 下降幅度低于 25%, 然而在小目标“sheep”类上的检测精度幅度达 50%以上下降到 4.4% 以下, 尤其是在小网络 Tiny YOLO 上。采用本文防御算法, 平均检测精度提升明显, 例如大检测网络 Darknet mAP 由 31.3% 提升到 64.0%。尤其是对于小尺寸的目标或特征比较明显的“dog”、“horse”以及“sheep”等类上的检测结果的提升超过 50%。Darknet 网络对“sheep”类检测准确率由 4.4% 提升到 65.8%, 并优于未攻击图像上的检测效果。表 8 展示了 Inria 上大小两个检测网络的检测准确率。从表中可知, 本文防御算法在对抗图像上检测结果的提升同样超过了 20%。

上述结果表明, 本文防御算法能够有效的防御

表 6 基于 Darknet 网络, 本文算法在对抗图像中每一类的检测效果

Table 6 Based on the Darknet network, the defense effect of our method for each category of adversarial image											
Darknet	mAP	plane	bike	bird	boat	bottle	bus	car	cat	chair	cow
Ori	69.1	79.9	79.4	71.9	55.5	53.6	77.8	72.7	85.0	46.4	67.8
Adv	31.3	59.0	36.1	15.3	29.3	36.1	58.6	65.6	25.7	10.8	43.2
Our(EDC+REC)	64.0	74.5	69.6	62.9	40.6	20.9	84.0	79.2	75.6	48.3	67.8
		table	dog	horse	motor	person	plant	sheep	sofa	train	tv
		57.0	78.3	85.2	80.8	75.8	49.7	65.7	56.1	80.6	61.7
		22.0	10.1	17.1	34.0	28.7	19.7	4.4	22.3	54.6	32.9
		55.2	61.8	79.6	79.0	75.7	36.0	65.8	57.5	78.4	67.6

表 7 基于 Tiny YOLO 网络, 本文算法在对抗图像中每一类的检测效果

Table 7 Based on the Tiny YOLO network, the defense effect of our method for each category of adversarial image											
Tiny YOLO	mAP	plane	bike	bird	boat	bottle	bus	car	cat	chair	cow
Ori	56.8	71.7	67.1	56.1	40.8	33.0	78.0	68.7	69.2	34.6	48.2
Adv	20.3	44.6	22.0	15.4	21.7	14.6	35.4	39.4	7.5	10.7	25.9
Our(EDC+REC)	48.2	61.0	50.0	45.0	26.1	23.2	68.9	63.5	50.8	31.8	45.4
		table	dog	horse	motor	person	plant	sheep	sofa	train	tv
		48.9	61.5	67.9	66.6	62.6	36.3	51.7	45.9	69.1	57.7
		18.5	10.7	10.8	25.2	16.8	13	1.1	9.8	23.6	39.2
		38.8	39.2	59.9	51.1	55.7	22.9	46.8	52.0	64.5	66.8

表 8 Inria 测试集上, 本文算法在不同网络模型下在对抗图像的检测效果

Table 8 Based on Inria test dataset, the detection results of our method on adversarial images under different networks			
Network	Ori	Adv	Our
Darknet	61.4	19.0	41.0
Tiny YOLO	53.3	17.5	35.8

针对目标检测网络的物理攻击, 在小目标上的防御效果尤其明显。

进一步的, 我们在未被攻击的图片上验证本文防御策略的作用。具体地, 我们在未被攻击的原始 Pascal VOC 2007(Ori)测试集上, 通过 Darknet 及 Tiny YOLO 网络的检测平均准确率验证防御策略可转移性。

表 9 列举了在未被攻击的数据集上, 利用基准网络^[27] Darknet 及 Tiny YOLO 网络 (Off) 的准确率, 以及本文防御算法(Our)得到的平均准确率。从表 9 可知, 相比于官方模型的检测结果 (Ori), 本文算法在两个网络上分别提升了 6.4%和 5.5%。证明了本文防御算法中 EDC 可能抹除掉信息量较大的区域, 但经过 ERC 训练后的模型能通过其他区域信息进行目标位置的回归和类别鉴定。图 13 展示了表 9 中 Darknet 网络对每类的检测准确率。可知, 在 20 个类别中 17 个的检测准确率要优于文献[27]。上述结果

证明了本文防御算法能显著提升目标检测网络的检测准确率。更为重要的是, 本文提出的算法不仅可以有效防御现有物理攻击而且能提升未被攻击图像上的检测准确率。

表 9 Pascal VOC 2007 测试集上, 本文算法在原始图像检测结果

Table 9 Object detection results on the Pascal VOC 2007 test dataset clean images		
Network	Off-Ori	Our-Ori
Darknet	69.1	75.5
Tiny YOLO	56.8	62.3

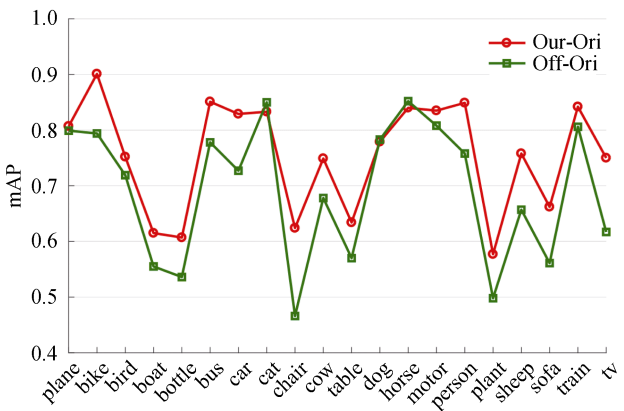


图 13 本文算法在原始图像中每一类检测效果
Figure 13 The detection result of our method on each category in the original image

图 14 实例展示本文算法和基准网络^[27]在 4 种场景下的检测结果。可知, 基准网络^[27]在未被攻击的图像(Ori)能准确检测到目标, 然而对抗块的图像(Adv)中检测准确率显著下降, 尤其是第 2 行图像完全失效。利用 EDC 检测对抗块并进行灰度替换后, 降低了对抗块对网络的干扰, 从而提升基准检测网络的检测准确率。但是, EDC 也可能丢失目标检测关键信息, 导致检测精度的下降。比如, 在第 2 行图上, 相比于未被攻击的图片, 目标“人”有两个漏检。基于 REC 训练得到的网络可以借助其他区域的信息辅助检测, 降低关键信息丢失对网络推理的影响。

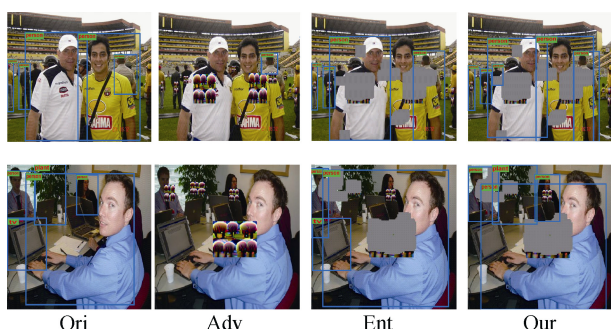


图 14 不同处理流程下目标检测结果, 前 3 列展示文献^[27]的检测检测结果, 最后一类展示本文算法的检测结果

Figure 14 Object detection results of images under different processing. The first three columns show the detection results of [27], and the last one shows the results of Ours

5 总结

本文提出一个针对目标检测物理攻击的防御算法。验证了现有物理攻击方法产生的对抗块与真实图像区域包含的信息分布是不同的。防御算法由基于熵的检测组件 (EDC) 及随机擦除组件 (REC) 组成。其中 EDC 用于检测和替换给定图像中物理攻击算法产生的对抗块。EDC 基于传统的图像处理技术进行图像预处理, 不依赖大量的训练数据。EDC 检测对抗块并对其进行灰度替换。该操作不可避免的造成图像原有关键信息的缺失从而影响网络的推理过程。为解决上述问题, 提出 REC 负责训练更加鲁棒的图像分析网络。实验表明 REC 不仅可以有效的避免现有的物理攻击, 而且能增强原有网络的图像分析能力。

针对深度学习网络的物理攻击和防御是相辅相成的两种技术手段。现有物理攻击的方式更加隐匿化, 比如, 最先进的攻击方法产生的对抗块分布更加平滑, 意味着信息熵更接近于真实图像分布。该攻

击将导致基于熵值的对抗块检测组件失效。我们下一步考虑利用目标检测和图像分割任务检测上述物理攻击生成的对抗块, 提升检测准确率和稳定性。

参考文献

- [1] Chen C Y, Seff A, Kornhauser A, et al. DeepDriving: learning affordance for direct perception in autonomous driving[C]. *2015 IEEE International Conference on Computer Vision*, 2016: 2722-2730.
- [2] Thys S., Van Ranst W., Goedeme T. Fooling automated surveillance cameras: adversarial patches to attack person detection[C]. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2019: 49-55.
- [3] Komkov S, Petiushko A. AdvHat: Real-World Adversarial Attack on ArcFace Face ID System[EB/OL]. 2019: arXiv: 1908.08705. <https://arxiv.org/abs/1908.08705>
- [4] Sharif M, Bhagavatula S, Bauer L, et al. Accessorize to a Crime: Real and Stealthy Attacks on State-of-the-Art Face Recognition[C]. *The 2016 ACM SIGSAC Conference on Computer and Communications Security*, 2016: 1528-1540.
- [5] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, et al. Intriguing properties of neural networks[M]. Computer Science, 2013.
- [6] Goodfellow I J, Shlens J, Szegedy C. Explaining and Harnessing Adversarial Examples[EB/OL]. 2014: arXiv: 1412.6572. <https://arxiv.org/abs/1412.6572>
- [7] Moosavi-Dezfooli S M, Fawzi A, Frossard P, et al. DeepFool: A simple and accurate method to fool deep neural networks[C]. *2016 IEEE Conference on Computer Vision and Pattern Recognition*, 2016: 2574-2582.
- [8] Chen S T, Cornelius C, Martin J, et al. ShapeShifter: Robust Physical Adversarial Attack on Faster R-CNN Object Detector[EB/OL]. 2018: arXiv: 1804.05810. <https://arxiv.org/abs/1804.05810>
- [9] Brown T B, Mané D, Roy A, et al. Adversarial Patch[EB/OL]. 2017: arXiv: 1712.09665. <https://arxiv.org/abs/1712.09665>
- [10] Karmon D, Zoran D, Goldberg Y. LaVAN: Localized and Visible Adversarial Noise[EB/OL]. 2018: arXiv: 1801.02608. <https://arxiv.org/abs/1801.02608>
- [11] Papernot N, McDaniel P, Wu X, et al. Distillation as a defense to adversarial perturbations against deep neural networks[C]. *2016 IEEE Symposium on Security and Privacy*, 2016: 582-597.
- [12] Tramèr F, Kurakin A, Papernot N, et al. Ensemble Adversarial Training: Attacks and Defenses[EB/OL]. 2017: arXiv: 1705.07204. <https://arxiv.org/abs/1705.07204>
- [13] Tsipras D, Santurkar S, Engstrom L, et al. Robustness may be at Odds with Accuracy[EB/OL]. 2018: arXiv: 1805.12152. <https://arxiv.org/abs/1805.12152>
- [14] Hazan T, Papandreou G, Tarlow D. Perturbations, optimization, and statistics[M]. Cambridge, Massachusetts: The MIT Press, 2016
- [15] Naseer M, Khan S, Porikli F, et al. Local gradients smoothing: Defense against localized adversarial attacks[C]. *2019 IEEE Winter Conference on Applications of Computer Vision*, 2019: 1300-1307.

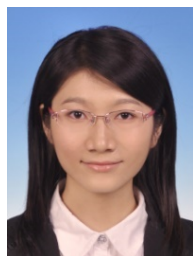
- [16] Chou E, Tramèr F, Pellegrino G. SentiNet: Detecting Localized Universal Attacks Against Deep Learning Systems[EB/OL]. 2018: arXiv: 1812.00292. <https://arxiv.org/abs/1812.00292>
- [17] Hayes J, Processing C A. On visible adversarial perturbations & digital watermarking[C]. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 2018: 1678-16787.
- [18] Yingqi Liu, Shiqing Ma, Yousra Aafer, et al. Trojaning attack on neural networks. Technical Report 17-002, Department of Computer Science, 2017.
- [19] Samangouei P, Kabkab M, Chellappa R. Defense-GAN: Protecting Classifiers Against Adversarial Attacks Using Generative Models[EB/OL]. 2018: arXiv: 1805.06605. <https://arxiv.org/abs/1805.06605>
- [20] Song Y, Kim T, Nowozin S, et al. PixelDefend: Leveraging Generative Models to Understand and Defend Against Adversarial Examples[EB/OL]. 2017: arXiv: 1710.10766. <https://arxiv.org/abs/1710.10766>
- [21] Simonyan K, Vedaldi A, Zisserman A. Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps[EB/OL]. 2013: arXiv: 1312.6034. <https://arxiv.org/abs/1312.6034>
- [22] Zeiler M D, Fergus R. Visualizing and Understanding Convolutional Networks[M]. Computer Vision - ECCV 2014. Cham: Springer International Publishing, 2014: 818-833.
- [23] Gray R M. Entropy and information theory[M]. 2nd ed. New York: Springer, 2011
- [24] Min B S, Lim D K, Kim S J, et al. A Novel Method of Determining Parameters of CLAHE Based on Image Entropy[J]. International Journal of Software Engineering and Its Applications, 2013, 7(5): 113-120.
- [25] Everingham M, Van Gool L, Williams C K I, et al. The Pascal Visual Object Classes (VOC) Challenge[J]. International Journal of Computer Vision, 2010, 88(2): 303-338.
- [26] Deng J, Dong W, Socher R, et al. ImageNet: A large-scale hierarchical image database[C]. 2009 IEEE Conference on Computer Vision and Pattern Recognition, 2009: 248-255.
- [27] Redmon J, Farhadi A, Processing C A. YOLO9000: better, faster, stronger[C]. 2017 IEEE Conference on Computer Vision and Pattern Recognition, 2017: 6517-6525.
- [28] He K M, Zhang X Y, Ren S Q, et al. Deep residual learning for image recognition[C]. 2016 IEEE Conference on Computer Vision and Pattern Recognition, 2016: 770-778.
- [29] Szegedy C, Vanhoucke V, Ioffe S, et al. Rethinking the inception architecture for computer vision[C]. 2016 IEEE Conference on Computer Vision and Pattern Recognition, 2016: 2818-2826.
- [30] Zhong Z, Zheng L, Kang G L, et al. Random Erasing Data Augmentation[EB/OL]. 2017: arXiv: 1708.04896. <https://arxiv.org/abs/1708.04896>
- [31] Selvaraju R R, Cogswell M, Das A, et al. Grad-CAM: Visual explanations from deep networks via gradient-based localization[C]. 2017 IEEE International Conference on Computer Vision, 2017: 618-626.



高红超 于 2020 年毕业于中国科学院信息工程研究所, 获计算机系统结构博士学位。现在中国科学院信息工程研究所助理研究员。研究领域为计算机视觉、数据挖掘。研究兴趣包括: 场景文字识别、人工智能安全等。Email: gaohongchao@iie.ac.cn



周广治 于 2017 年毕业于山东理工大学, 获得计算机科学与技术学士学位。现就读于中国科学院信息工程研究所, 攻读计算机技术硕士学位。研究领域为多媒体信息处理、人工智能安全。研究兴趣包括: 对抗样本、目标检测等。Email: zhouguangzhi@iie.ac.cn



戴娇 于 2019 年在中国科学院(计算机系统结构)专业获得博士学位。现任中国科学院信息工程研究所高级工程师。研究领域为多媒体信息处理、人工智能安全。研究兴趣包括: 图像检索、图像深度伪造、对抗防御。Email: daijiao@iie.ac.cn



李昭星 于 2015 年毕业于大连理工大学, 获软件工程硕士学位。现在中国科学院信息工程研究所第五工程部助理研究员。研究领域为计算机视觉、多媒体大数据处理。研究兴趣包括: 大规模图像检索、人工智能安全等。Email: lizhaoxing@iie.ac.cn



韩冀中 于 2001 年在中国科学院计算技术研究所大学计算机系统结构专业获得博士学位。现在中国科学院信息工程研究所任正高级工程师。研究领域为人工智能、计算机系统结构等。研究兴趣包括: 大数据存储与管理、多媒体信息智能化处理等。Email: hanjizhong@iie.ac.cn