软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2024,35(7):3531−3550 [doi: 10.13328/j.cnki.jos.006937]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

基于生成对抗网络的目标检测黑盒迁移攻击算法*

陆宇轩[1], 刘泽禹[2], 罗咏刚[3], 邓森友[1], 江 天[3], 马金燕[3], 董胤蓬[1,2]

[1](北京瑞莱智慧科技有限公司, 北京 100089)

[2](清华大学 计算机科学与技术系, 北京 100084)

[3](重庆长安汽车软件科技有限公司, 重庆 400021)

通信作者: 董胤蓬, E-mail: dongyinpeng@mail.tsinghua.edu.cn

摘 要: 目标检测被广泛应用到自动驾驶、工业、医疗等各个领域. 利用目标检测算法解决不同领域中的关键任务逐渐成为主流. 然而基于深度学习的目标检测模型在对抗样本攻击下, 模型的鲁棒性存在严重不足, 通过加入微小扰动构造的对抗样本很容易使模型预测出错. 这极大地限制了目标检测模型在关键安全领域的应用. 在实际应用中的模型普遍是黑盒模型, 现有的针对目标检测模型的黑盒攻击相关研究不足, 存在鲁棒性评测不全面, 黑盒攻击成功率较低, 攻击消耗资源较高等问题. 针对上述问题, 提出基于生成对抗网络的目标检测黑盒攻击算法, 所提算法利用融合注意力机制的生成网络直接输出对抗扰动, 并使用替代模型的损失和所提的类别注意力损失共同优化生成网络参数, 可以支持定向攻击和消失攻击两种场景. 在 Pascal VOC 数据集和 MS COCO 数据集上的实验结果表明, 所提方法比目前攻击方法的黑盒迁移攻击成功率更高, 并且可以在不同数据集之间进行迁移攻击.

关键词: 对抗攻击; 目标检测; 黑盒迁移攻击; 生成对抗网络; 注意力损失

中图法分类号: TP391

中文引用格式: 陆宇轩, 刘泽禹, 罗咏刚, 邓森友, 江天, 马金燕, 董胤蓬. 基于生成对抗网络的目标检测黑盒迁移攻击算法. 软件学报, 2024, 35(7): 3531–3550. http://www.jos.org.cn/1000-9825/6937.htm

英文引用格式: Lu YX, Liu ZY, Luo YG, Deng SY, Jiang T, Ma JY, Dong YP. Black-box Transferable Attack Method for Object Detection Based on GAN. Ruan Jian Xue Bao/Journal of Software, 2024, 35(7): 3531–3550 (in Chinese). http://www.jos.org.cn/1000-9825/6937.htm

Black-box Transferable Attack Method for Object Detection Based on GAN

LU Yu-Xuan[1], LIU Ze-Yu[2], LUO Yong-Gang[3], DENG Sen-You[1], JIANG Tian[3], MA Jin-Yan[3], DONG Yin-Peng[1,2]

[1](Beijing RealAI Intelligent Technology Co. Ltd., Beijing 100089, China)

[2](Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China)

[3](Chongqing Changan Automobile Software Technology Co. Ltd., Chongqing 400021, China)

Abstract: Object detection is widely used in various fields such as autonomous driving, industry, and medical care. Using the object detection algorithm to solve key tasks in different fields has gradually become the main method. However, the robustness of the object detection model based on deep learning is seriously insufficient under the attack of adversarial samples. It is easy to make the model prediction wrong by adding the adversarial samples constructed by small perturbations, which greatly limits the application of the object detection model in key security fields. In practical applications, the models are black-box models. Related research on black-box attacks against object detection models is relatively lacking, and there are many problems such as incomplete robustness evaluation, low attack success rate of black-box, and high resource consumption. To address the aforementioned issues, this study proposes a black-box object detection attack algorithm based on a generative adversarial network. The algorithm uses the generative network fused with an attention mechanism to output the adversarial perturbations and employs the alternative model loss and the category attention loss to optimize the generated network parameters, which can support two scenarios of target attack and vanish attack. A large number of experiments are

* 收稿时间: 2022-11-27; 修改时间: 2023-01-14; 采用时间: 2023-03-17; jos 在线出版时间: 2023-08-23
CNKI 网络首发时间: 2023-08-28

conducted on the Pascal VOC and the MSCOCO datasets. The results demonstrate that the proposed method has a higher black-box transferable attack success rate and can perform transferable attacks between different datasets.

Key words: adversarial attack; object detection; black-box transferable attack; generative adversarial network (GAN); attention loss

深度学习在近几年蓬勃发展并成功应用到了计算机视觉领域[1], 这推动了众多视觉任务的进一步发展和落地应用, 例如图像分类[2−5]、目标检测[6−10]、图像分割[11]等. 其中目标检测算法应用领域广泛, 所涉及领域如自动驾驶[12,13], 工业检测[14]等. 同时这些应用与安全密切相关, 因此对算法的鲁棒性和安全性提出了更高的要求. 然而, 深度学习已被证明当面临对抗攻击时十分脆弱, 很容易受到对抗样本的干扰[15], 比如向输入图片中加入微小的扰动, 即可完全改变模型的输出结果, 因此围绕深度学习算法的对抗攻击和防御的研究受到了越来越多的关注[16−18]. 根据可获取的受害模型的信息可以将对抗攻击分为两类. 第 1 类为白盒攻击, 其假设受害模型完全公开, 攻击者可以获取模型的内部结构以及参数, 这样就可以针对性地构造对抗样本并对模型进行攻击. 然而, 白盒攻击场景较为简单, 当攻击者无法获取模型结构及参数时, 攻击的成功率会大幅降低. 第 2 类为黑盒攻击, 即攻击者无法获取模型或者训练数据集, 在信息受限的情况下进行对抗攻击. 黑盒攻击和实际应用更为接近, 对于评估真实应用中模型的鲁棒性具有重要研究意义. 为了成功攻破黑盒模型, 先前研究发现对抗样本具有迁移能力[19−21], 即对白盒模型生成的对抗样本也可以欺骗未知的黑盒模型, 这样就可以在不获取模型内部信息的情况下误导黑盒模型. 本文针对目标检测模型研究黑盒迁移攻击算法, 旨在提升黑盒攻击的成功率和效率.

目前, 针对图像分类的对抗攻击受到了广泛关注, 也产生了许多代表性的攻击算法, 如 FGSM[19]、DeepFool[22]、C&W[23]、MI-FGSM[24]等. 针对目标检测任务虽然也已产生一些对抗攻击算法, 如 DAG[25]、RAP[26]、UEA[27]、TOG[28]等, 但由于目标检测相比分类任务复杂度更高, 目前研究仍然存在不足. 图像分类模型预测图片中物体对应的类别, 因此对图像分类的攻击算法只能攻击图片所属类别. 目标检测算法的网络结构以及算法逻辑更加复杂, 通常预测图中多个目标物体的位置以及类别. 因此针对目标检测算法的对抗攻击研究可以从目标框、置信度、类别 3 个方面进行攻击, 攻击的形式更加多样. 现有的针对目标检测的攻击算法[19,25,26]都是在白盒模型中进行实验, 黑盒攻击的效果远不如白盒攻击. 而且这些算法并没有研究定向攻击场景, 无法全面衡量模型的鲁棒性.

本文在定向攻击和消失攻击两个方面对模型的鲁棒性进行评估, 旨在提升黑盒迁移攻击成功率. 定向攻击和消失攻击如图 1 所示, 定向攻击是将全图目标定向攻击为特定的类别, 消失攻击是让模型无法检测到图像中的物体. 在不依赖数据集中任何先验信息的情况下, 本文优化一个全图扰动, 将生成的扰动添加到原始图像上, 进行全图范围的攻击, 同时扰动在人眼看来十分隐蔽. 相比于目前研究较多的无定向随机攻击, 本文采取的两种攻击方式令生成的扰动更加具有指向性. 在现实应用中, 可以根据攻击者的意图进行具有目的性的攻击, 可以应用在靶向攻击的场景中. 通过本文这两种攻击方法, 可以更全面地评估目标检测模型在预测多个目标物时的鲁棒性. 现有的攻击算法[19,25−27]难以欺骗黑盒模型, 成功率较低, 导致对目标检测模型鲁棒性的评估不准确.

为了提升针对目标检测模型的定向攻击和消失攻击的黑盒攻击成功率, 本文提出基于生成对抗网络的目标检测黑盒攻击算法. 该算法采用一个生成网络学习对抗扰动的分布, 可以有效缓解已有基于梯度优化的攻击方法存在的过拟合问题, 提升黑盒攻击效果. 尽管在图像分类任务中已经有[28−32]等方法采用生成对抗网络进行黑盒攻击, 但是由于目标检测涉及多个目标物、决策机制更为复杂, 已有方法无法简单迁移到目标检测任务中. 对于目标检测任务而言, 目标物体在图像中所占区域较少, 存在大量的冗余信息, 会导致生成网络无法抓住图像中的重点区域, 而且利用替代模型的损失梯度去更新生成网络, 会使网络过拟合替代模型, 导致黑盒攻击效果不佳. 针对这些问题, 本文在生成网络内部加入注意力机制模块, 帮助生成网络将扰动添加到前景区域. 在网络的输出端采用了一个高斯平滑映射, 增加扰动的多样性, 使生成网络学习扰动的边界, 避免对当前替代模型过于敏感. 此外, 本文提出类别注意力损失, 通过计算特征的热力图, 帮助网络在攻击时能够针对性地进行攻击, 提升迁移能力. 即使网络结构不同, 但是会存在相似的注意力, 可以利用这种注意力指导生成网络的训练, 提升在其他黑盒模型的攻击成功率.

为了验证所提出的算法, 本文在第 3 节中进行大量实验. 实验选取了两个最常用的目标检测数据集: MS COCO[33]和 Pascal VOC[34]. 本文选取了 5 个对抗攻击的方法, 并以文献 [27] 作为本文的基准实验. 为了验证所提算法在黑

盒模型上的迁移攻击效果, 实验选取了 11 个不同目标检测网络进行测试. 在使用替代模型一致的前提下, 本文提出的方法比文献 [27] 的方法平均提升攻击成功率 9% (定向攻击) 和 15% (消失攻击). 本文尝试了在不同数据集之间定向攻击的迁移实验, 在数据集迁移攻击过程中由于类别分布存在差异, 会导致攻击成功率的下降. 因此本文选取了迁移后攻击成功率占迁移前攻击成功率的百分比作为评价标准, 本文提出的方法比文献 [27] 的方法平均提升 6.18%. 在进行类别注意力损失函数的实验中, 使用本文提出的攻击方法在加入类别注意力损失前后攻击成功率平均提升 1.57% (定向攻击), 0.37% (消失攻击).

原图预测　　定向攻击　　消失攻击

图 1　定向攻击与消失攻击

本文的贡献可以归纳为以下 3 点.

(1) 提出基于生成对抗网络的目标检测黑盒攻击算法. 该算法将生成网络引入对抗攻击中, 并引入了注意力机制以及高斯平滑映射, 用于生成对抗样本, 支持定向攻击以及消失攻击.

(2) 提出类别注意力损失. 使用该损失和替代模型损失共同优化生成网络, 进而提升攻击的成功率, 增强攻击算法的迁移能力, 并在黑盒模型验证了其有效性.

(3) 在 Pascal VOC 和 MS COCO 数据集上进行实验, 验证本文方法的普适性. 所提方法不仅在黑盒攻击表现优于目前的攻击方法, 而且在不同数据集之间可以进行有效的迁移攻击.

本文第 1 节介绍相关工作及研究. 第 2 节详细介绍本文提出的黑盒迁移攻击方法. 第 3 节介绍实验的设置及结果分析. 第 4 节为本文总结.

1　相关工作及研究

1.1　对抗攻击

在原图上加入精心设计的扰动就可以产生一个对抗样本, 不仅可以逃过人眼, 也可以使模型出错. 在过去几年中, 已经有很多关于对抗攻击的工作. 可以将攻击算法分为 3 个方向: 基于优化方法, 基于修改注意力方法和基于平滑方法. 最早提出来的是基于优化的方法, 主要是针对白盒模型攻击的方法. 使用替代模型的梯度来优化目标函数, 例如最大化训练损失或最小化原始图像的得分或输出. 这些方法包括 FGSM[19]、BIM[35]、PGD[36]、DeepFool[22]. 基于修改注意力的方法利用了不同网络结构对相似特征具有相似注意力的原理, 通过修改模型的注意力结构来提高对抗的转移能力. 这些方法包括 AOA[37]、ATA[38]、ILA[39]、TAIG[40]. 由于在攻击黑盒模型时, 无法获取模型的参数梯度信息, 所以采用白盒模型作为替代模型. 但是扰动对于白盒模型会出现过拟合情况, 因此基

于平滑的方法就可以改善对白盒模型的决策面过拟合. 主要分为两个方向: 使用从决策曲面的多个点导出的平滑梯度; 修改梯度计算, 以估计其他曲面的梯度. 这些方法包括 TI[20]、SI[41]、DI[42]. 基于修改注意力的方法和基于平滑的方法多被用于黑盒攻击中. 目前很多方法将以上 3 种方法进行组合. 例如 SM^{2}I-FGSM[43], 该方法证明组合后的方法要比单一的方法攻击效果更强.

目前最热门的研究是黑盒攻击, 黑盒攻击表明攻击者不能访问模型的内部参数. 针对黑盒攻击的两种常见策略为: (1) 基于查询攻击方法[44−47]. 通常 AI 模型/算法是保密的, 攻击者虽然不能获取目标模型, 但是可以通过查询的方式估计出目标模型的运行方式. 以查询的结果为经验逐渐增强攻击的效果, 从而实现黑盒攻击. 这种方法虽然具有很高的成功率, 但需要进行大量次数地查询, 在现实生活中一般无法实现. 而且由于 AI 模型固有的高维属性, 也无法保证所查询信息的有效性. (2) 基于迁移攻击方法[19,28,29,38−40]. 目前很多对抗攻击的场景中, 例如安全支付等问题中, 攻击者无法进行大量的查询. 因此可以利用对抗样本的迁移性能, 对目标模型进行攻击. 采用一个白盒替代模型来模拟需要攻击的黑盒模型的情况, 并通过优化对于白盒模型的攻击成功率试图迁移到黑盒模型上. 这种攻击方式可以让攻击者不需要获取目标模型的结构参数, 直接使用对抗样本进行攻击就有可能取得成功. 但是如何提高对抗样本的迁移攻击能力仍然是当前的研究热点.

1.2 深度学习目标检测

目标检测是一个多任务的学习问题, 多任务包括前景背景区分, 多目标物区分以及目标物位置识别. 目标检测已经展现了它解决实际问题的能力. 目前应用最为广泛的检测算法可以分为: 基于区域提议的双阶段目标检测网络, 基于回归的单阶段目标检测网络以及 Transformer_Based 目标检测网络.

基于区域提议的双阶段网络主要以 R-CNN 系列为主: Fast R-CNN[48]、Faster R-CNN[8]、Mask R-CNN[49]. 与单阶段模型相比, 双阶段模型首先利用额外的网络确定物体框的大致位置, 然后再对可能的物体框位置进行回归确定物体的详细类别. 双阶段网络由于使用了额外的网络, 所以精度较高. 但额外的网络分支会造成整体结构复杂, 并且训练过程复杂, 速度较慢.

基于回归的单阶段网络有 YOLOv3[10]、SSD[9]. 单阶段网络将边界框和类别预测问题转换成了一个回归问题. 它对于给定的图像通过只看 1 次的方式, 可以直接预测边界框的坐标来联合估计目标物的坐标及类别信息. 单阶段网络通常结构较为简单, 在训练及推理过程中更加迅速. 但是由于没有辅助结构确定候选框的位置, 单阶段网络相比于双阶段网络会出现精度低的缺陷.

Transformer_Based 的网络代表算法有 DETR[50]、D-DETR[51]. 将 Transformer 引入到目标检测中, 是一种端到端目标检测算法. 算法抛开了预测锚框的束缚, 在训练和部署上均相比于传统方法有更为通用的提升. 但是 Transformer_Based 网络训练的时间成本高, 而且基于 Transformer 网络虽然可以获得更大的感受野, 但是也损失了在小目标上的检测性能.

1.3 目标检测攻击

目标检测算法的对抗攻击更为复杂, 可采用不同的优化函数. 通常利用替代检测模型的一个或多个损失: 类别损失, 置信度损失和检测框损失. 利用梯度优化扰动使得攻击算法能够向原始图像添加特定的扰动, 使得输入微小扰动, 将在被攻击检测器的整个向前传播过程中被放大, 足以改变模型的预测结果. 本文简要介绍 4 种代表性基于目标检测的攻击算法: DAG[25]、RAP[26]、UEA[19]、TOG[27]. DAG[25]是最早提出的攻击检测模型算法, 它通过操纵 RPN 网络生成大量错误建议区域, 以此来攻击受害者模型. 对于一个双阶段的目标检测模型, DAG[25]攻击过程可以分为两个阶段: 在第 1 阶段, 利用 RPN 网络生成可能包含目标的候选区域; 在第 2 阶段, 利用第 1 阶段得到的候选区域集合训练生成对抗扰动. DAG[25]为每个候选区域分配一个随机选择的标签, 然后执行迭代梯度攻击. RAP[26]直接攻击 RPN 网络, 使其无法返回前景对象, 攻击预测的检测框, 使其定位失效. 论文中尝试将多个攻击扰动进行累加, 可以产生一定的迁移攻击效果. 上述两种方法在每次更换攻击的标签时都需要重新训练, 训练成本随攻击类别数量而增加; 对于每张不同的输入都需要单独进行训练才可以得到对抗样本, 训练的效率较低; 两种算法主要针对双阶段的目标检测模型, 对于其他检测模型泛化效果不佳. UEA[19]解决了之前方法时间成本很高的问题,

通过引入生成网络替代了原本基于梯度的优化方法, 提高了攻击的成功率. UEA[19]改进了损失函数, 采用分类类别损失加上基于注意力机制的特征损失提高了模型迁移性. 该攻击算法在更换攻击标签时同样需要重新训练, 并且虽然该算法拥有迁移攻击的能力, 但是却受限于替代模型的梯度. TOG[27]是一种基于梯度迭代的通用攻击方法, 在训练时将损失进行反向传播就可以攻击所有不同类型的模型. 攻击方式包括: 无目标攻击, 消失攻击, 造假攻击, 愚弄标签攻击, 通用攻击, 区域攻击. 该算法利用了替代模型的损失函数以及梯度信息更新扰动. 但是 TOG[27]主要是为白盒模型攻击而设计的, 因此只开展了部分的黑盒模型评估工作, 并不能从多个方面对模型鲁棒性进行评估. 而且黑盒迁移攻击的能力很差. 因此, 本文为了更全面地评估目标检测黑盒模型的鲁棒性, 提出针对目标检测的黑盒攻击方法, 并进行定向攻击和消失攻击的研究.

2 基于生成对抗网络的黑盒攻击方法

在第 1 节中介绍了一些现有攻击方法, 但是在进行黑盒迁移攻击时, 大多数方法攻击成功率较低. 例如基于优化的攻击方法[20,22,36,52], 这些方法在黑盒迁移攻击中均没有达到在白盒攻击时的表现. 在本节中提出一个高效的黑盒迁移攻击算法, 并支持定向攻击和消失攻击. 该算法由一个生成网络构成, 在这个生成网络中加入注意力机制, 用以提升黑盒攻击的迁移能力. 在本节中提出一个支持多尺度类别注意力损失, 以此增加扰动的迁移能力. 本文将对抗样本的生成转变为了图像到图像的转换问题. 向网络中输入干净的图像, 输出带有针对扰动的对抗样本.

本节对基于生成网络的目标检测黑盒攻击算法进行介绍. 第 2.1 节介绍定向攻击和消失攻击的定义. 第 2.2 节介绍生成网络的结构. 第 2.3 介绍损失函数. 第 2.4 节介绍算法的流程.

2.1 定向攻击和消失攻击

本节定义原始图片 x_{benign}, 通过向 x_{benign} 中加入一个扰动 δ, 得到对抗样本 $x_{\text{adv}}=x_{\text{benign}}+\delta$, 并限制扰动 δ 的 ℓ_∞ 范数不超过 ϵ, 即 $x_{\text{benign}}-x_{\text{adv}_\infty}\leqslant\epsilon$. 本文的目标是令对抗样本能够攻击黑盒模型, 实现定向攻击任务以及消失攻击任务. 定义 M_ϕ 为目标检测网络, 网络输出多个目标位置类别信息的预测向量. 其中 M_{class} 代表网络预测的类别, M_{bbox} 代表网络预测的目标框. 在进行定向攻击时, 首先选定攻击类别 $C\in y_{\text{class}}$, 其中 y_{class} 为定向攻击目标类别的集合. 旨在定向攻击的对抗样本 x_{adv}^C 可以使受害者模型 M_ϕ 将所有前景目标识别为类别 C, 即 $M_{\text{class}}\left(x_{\text{adv}}^C\right)=C$. 在进行消失攻击时, 旨在消失攻击的对抗样本 $x_{\text{adv}}^{\text{vanish}}$ 可以使受害者模型 M_ϕ 无法识别到图中所有前景的目标, 只能预测为背景类 $C_{\text{background}}$. 即 $M_{\text{class}}\left(x_{\text{adv}}^{\text{vanish}}\right)=C_{\text{background}}$. 为了能够进行定向攻击, 目标函数旨在最小化模型预测类别和定向攻击类别之间的差异, 如公式 (1) 所示. 进行消失攻击时, 目标函数旨在最小化模型预测类别和消失攻击类别之间的差异, 如公式 (2) 所示.

$$\text{argmin}_{x_{\text{adv}}^C}\mathcal{L}(M_{\text{class}}(x_{\text{adv}}^C),C) \tag{1}$$

$$\text{argmin}_{x_{\text{adv}}^{\text{vanish}}}\mathcal{L}(M_{\text{class}}(x_{\text{adv}}^{\text{vanish}}),C_{\text{background}}) \tag{2}$$

其中, $\mathcal{L}$ 为攻击的目标函数, 具体形式将会在第 2.3 节中详细介绍.

定向攻击和消失攻击可以转化为对于攻击目标函数公式 (1) 和公式 (2) 的求解问题. 求解上述攻击目标函数, 可以和文献 [25−27] 这些方法一样, 利用梯度经过多次迭代优化扰动. 这种方式虽然在白盒攻击情况下有效, 但是十分依赖替代模型的结构. 因此在黑盒迁移攻击时, 生成的对抗样本往往没有攻击的能力, 不利于鲁棒性的评测. 目前还有一种方法, 利用生成式的方法如[19,28−30]. 利用生成网络可以自主学习到数据中的特征表示, 可以提升扰动的泛化能力. 因此利用生成网络生成的对抗样本更加鲁棒, 也有利于欺骗不同网络结构的深度学习模型, 增强不同模型的攻击迁移能力. 本文提出的攻击算法利用了生成网络生成扰动, 并且在定向攻击任务中, 支持多类别的攻击形式.

2.2 攻击网络介绍

2.2.1 生成网络结构

攻击算法包括由编解码器构成的生成网络、条件映射网络以及在输出端设计的高斯平滑映射. 网络结构如

图 2 所示.

图 2 中, 定义原始图片为 X_{benign}, 定向攻击的类别 C, 生成网络 $\mathcal{G}_\theta$, 映射网络 $\mathcal{W}$, 替代模型 $\mathcal{M}_\varphi$. 数据从生成网络的编码端流向解码端. 该网络可以支持定向攻击和消失攻击, 两个任务在训练和计算损失时存在差异, 所以无法对两个任务训练一个特征提取器. 两个任务分别训练了两个独立的网络. 在网络结构的设计中, 可以分为网络输入, 特征提取以及网络输出 3 大部分. 两个任务拥有不同的网络输入, 但特征提取以及网络输出结构是一致的.

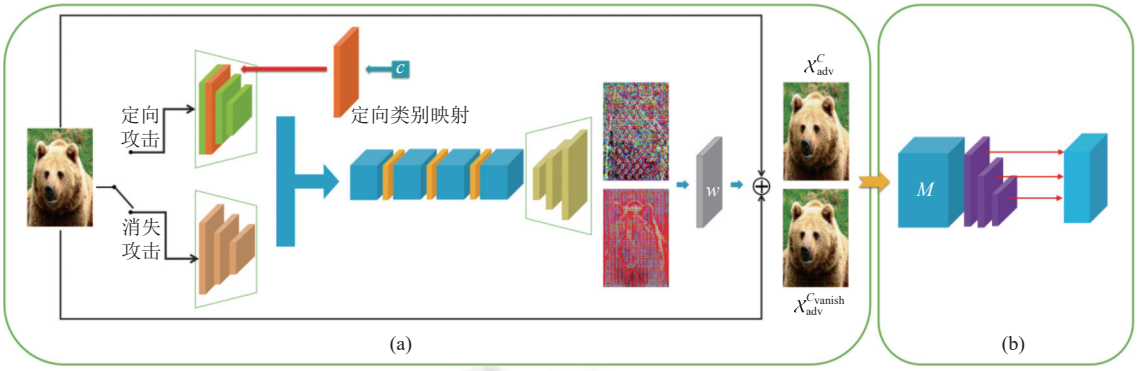


图 2 网络结构图

当进行消失攻击时, 只需要选择将数据流经图 2 中消失攻击分支, 将 X_{benign} 输入 $\mathcal{G}_\theta$ 的编码端. 而定向攻击时, 先进行类别映射, 之后将类别向量嵌入到编码网络中, 将数据流经图 2 中定向攻击分支. 在编码端首先通过引入定向攻击的目标类别 C, 使用映射网络 $\mathcal{W}$ 输出特定目标的隐式向量 w_C, 将向量沿高度和宽度方向展开, 得到带有类别信息的特征 $w_C \in \mathbb{R}^{C\times H\times W}$. 将 X_{benign} 输入 $\mathcal{G}_\theta$ 中, 经过第 1 层的卷积模块, 输出特征图 $m_\ell \in \mathbb{R}^{M\times H\times W}$, 将 w_C 和 m_ℓ 的特征图在通道维度进行拼接, 便可得到 $m_\ell^* \in \mathbb{R}^{(C+M)\times H\times W}$. 该方法通过拼接图片特征和目标标签的特征来生成混合特征. 此时将融合后的特征图 m_ℓ^* 输入到后续的网络中.

在编码端和解码端中间, 使用多个残差模块以及注意力机制模块, 提升生成模型的学习能力. 与图像分类任务不同, 目标检测任务对图中多个目标物进行检测. 图中的前景信息一般占比较小, 如果提取过多的背景信息特征, 将不利于网络的学习, 而注意力机制就是在帮助网络学习重要的前景特征. 受 CBAM[53]模块的启发, 本节在生成网络中加入了注意力机制, 帮助网络学习图像中的重点特征. 该模块是一个简单但有效的注意力机制的模块, 可以集成到大多数的网络架构中, 并且可以端到端的方式进行训练. 该模块将输入的特征图沿通道和空间两个独立维度依次推断注意力图, 然后将注意力图乘以输入的特征图, 以便执行自适应特征细化.

CBAM[53]模块结构如图 3 所示.

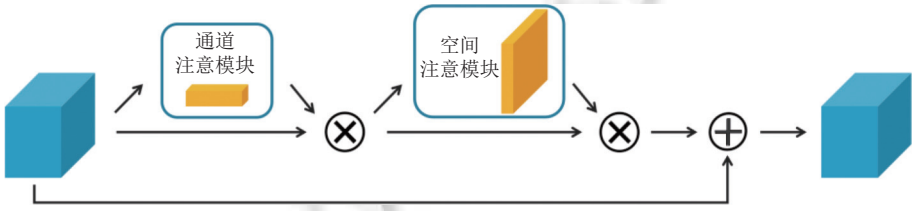


图 3 CBAM[53]模块结构图

根据 CBAM[53]中的实验, 在不同的分类和检测数据集上将模块集成到不同的模型中后, 模型的性能得到了很大的提高. 在对抗攻击中受到扰动的区域很大, 因此可以加入该模块提取注意力的区域, 帮助生成网络专注更重要的目标, 提升攻击的能力.

在解码端恢复特征图的分辨率, 对特征图 $m_{\text{last}} \in \mathbb{R}^{3\times H_{\text{input}}\times W_{\text{input}}}$ 进行了高斯平滑映射操作, 该操作包括一个平滑映

射和一个高斯卷积操作, 定义如下:

$$\mathcal{X}_{\text{out}} = \epsilon \cdot (\mathbb{W} * \tanh(m_{\text{last}})) \tag{3}$$

其中, $\mathbb{W}$ 是预定义的高斯卷积核, 设置的卷积核尺寸为 7, $\cdot$是乘法操作, *是卷积运算, tanh 是激活函数. 在 TIM 算法[20]使用高斯卷积核对空间域中的梯度进行平滑处理, 通过将未平移图像的梯度与预定义的核矩阵进行卷积来近似计算梯度, 增加梯度的多样性.

受到该算法的启发, 本节放弃了直接裁剪特征值的方式, 操作流程如图 4 所示. 首先利用 tanh 对网络输出 m_{last} 进行一个平滑映射, 然后使用高斯卷积操作在小幅度内移动图像. 该操作使模型对判别边界不敏感, 提高攻击其他模型的成功率, 使生成的结果在训练阶段获得自适应能力.

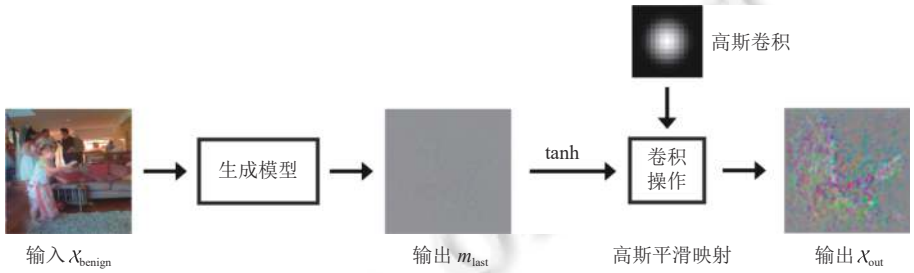


图 4 高斯平滑映射图

综上, 生成网络生成的对抗样本表示如下:

$$\mathcal{X}_{\text{adv}}^{C} = \mathcal{X}_{\text{benign}} + \epsilon \cdot (\mathbb{W} * \tanh(\mathcal{G}_{\theta}(\mathcal{X}_{\text{benign}}, C))) \tag{4}$$

$$\mathcal{X}_{\text{adv}}^{\text{vanish}} = \mathcal{X}_{\text{benign}} + \epsilon \cdot (\mathbb{W} * \tanh(\mathcal{G}_{\theta}(\mathcal{X}_{\text{benign}}))) \tag{5}$$

公式 (4) 代表定向攻击任务, 公式 (5) 代表消失攻击任务. 生成模型 $\mathcal{G}_{\theta}$ 可以通过公式 (4) 将原始输入 $\mathcal{X}_{\text{benign}}$ 和类别 C 的隐式向量转化为带有定向对抗扰动的输出图像 $\mathcal{X}_{\text{adv}}^{C}$, 生成模型可以合成多个目标的对抗图像; 通过公式 (5) 将原始输入 $\mathcal{X}_{\text{benign}}$ 转化为带有消失对抗扰动的输出图像 $\mathcal{X}_{\text{adv}}^{\text{vanish}}$. 该黑盒攻击方法可以通过最小化指定目标的损失函数去训练条件生成模型 $\mathcal{G}_{\theta}$. 但需要注意的是, 应该固定检测器 $\mathcal{M}_{\phi}$ 的参数, 之后通过不断的训练生成模型 $\mathcal{G}_{\theta}$, 去误导检测器 $\mathcal{M}_{\phi}$, 实现定向攻击和消失攻击.

2.2.2 定向多目标映射网络

图像分类任务是识别图像中的单目标的类别信息, 目标检测任务是识别图中多个目标物类别信息以及位置信息. 这比图像分类任务更加复杂, 尤其是需要考虑针对不同类别进行攻击时, 以往的算法以及攻击方式都存在不足. 利用生成网络也会存在一个弊端, 在进行消失攻击任务时, 可以使用一个生成网络进行学习. 但是在进行定向攻击任务时, 一个生成网络只能攻击一个类别, 如果需要攻击其他类别, 就需要重新训练一个新的模型, 训练成本和攻击类别成线性增加. 检测算法会识别数据集中的多个类别, 因此这种方式根本无法适用在针对检测算法定向攻击任务中.

MAN[31]和 C-GSP[32]通过训练一个生成模型, 进行定向类别攻击, 该模型可以指定类别进行攻击. 然而这些算法均是攻击分类算法. 根据 SENet[54]中不同特征的表示通道为不同的类捕获不同的特征, 使用通道来集成标签特征和图像特征. 映射网络结构如图 5 所示.

映射网络融合目标类别的隐式向量可以使攻击网络生成用于指定多目标的对抗性样本. 在这里本节使用了多层感知机和归一化层作为生成类别向量的网络 $\mathcal{W}$. 首先将定向攻击的类别 C 进行独热编码 ℓ_C, 此时将 ℓ_C 输入到映射网络 $\mathcal{W}$ 中, 映射网络通过使用独热编码在每个目标类别的隐式空间 $\mathbb{I}_c$ 中生成目标特定的向量:

$$w_C = \mathcal{W}(\ell_C) \tag{6}$$

该映射网络通过对不同类别的映射, 可以有效地学习不同目标的隐式向量. 这些含有类别信息的向量被存储在不同的通道中. 在融合操作中, 算法会将图片的特征以及类别向量编码进行通道维度的拼接融合.

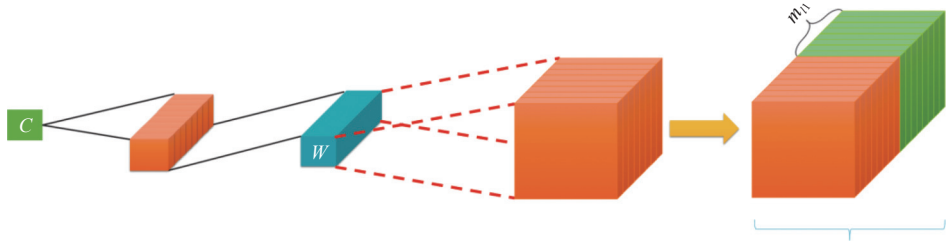


图 5 映射网络结构图

2.3 损失函数

2.3.1 替代模型损失函数

训练生成网络时, 本文计算白盒替代模型的损失函数, 利用白盒替代模型的梯度信息, 更新生成网络的参数. 在使用生成网络的情况下, 采用整体数据集 $\mathcal{X}$ 优化生成网络参数, 而不是采用单个数据 x_{benign} 进行优化. 因此公式 (1) 和公式 (2) 中的目标函数变为:

$$\text{argmin}_\theta \mathbb{E}_{(x_{\text{adv}}^C \sim \mathcal{X}, C \in y_{\text{class}})} \mathcal{L}(\mathcal{M}_{\text{class}}(x_{\text{adv}}^C), C) \tag{7}$$

$$\text{argmin}_\theta \mathbb{E}_{(x_{\text{adv}}^{\text{vanish}} \sim \mathcal{X}, C_{\text{background}})} \mathcal{L}(\mathcal{M}_{\text{class}}(x_{\text{adv}}^{\text{vanish}}), C_{\text{background}}) \tag{8}$$

和图像分类模型不同, 目标检测模型的损失函数更为复杂, 包括了对目标位置以及目标类别的损失函数. 因此本文在利用替代模型时, 计算了替代模型中所有的损失函数, 共同优化更新生成模型的参数, 多个损失函数能够使梯度信息更加丰富. 为了验证本文提出的算法具有普适性, 本文选择了两个白盒替代模型, 分别是单阶段目标检测网络 YOLOv3[10], 双阶段目标检测网络 Faster R-CNN[8].

在 YOLOv3[10]中, 损失函数分为 3 个部分: 目标置信度损失 $\mathcal{L}_{\text{obj}}$, 目标类别损失 $\mathcal{L}_{\text{class}}$ 以及目标框损失 $\mathcal{L}_{\text{bbox}}$. 其中分类损失代表预测类别与标签类别的误差, 置信度损失代表预测置信度与标签置信度误差, 位置坐标损失代表预测坐标与标签坐标误差. YOLOv3[10]中的损失定义如下:

$$\mathcal{L}_{\mathcal{M}_\phi} = \alpha \mathcal{L}_{\text{class}} + \beta \mathcal{L}_{\text{obj}} + \gamma \mathcal{L}_{\text{bbox}} \tag{9}$$

其中, α, β, γ 分别代表每个损失部分的重要程度.

双阶段的损失函数在计算方式上略有不同, 在 Faster R-CNN[8]中, 需要考虑两个损失, RPN (region proposal network)[8]损失以及 Fast R-CNN[48]损失. Faster R-CNN[8]的损失函数可以定义为:

$$\mathcal{L}_{\mathcal{M}_\phi} = \tau \mathcal{L}_{\text{RPN}} + \eta \mathcal{L}_{\text{Fastrcnn}} \tag{10}$$

其中, τ, η 分别代表每个损失部分的重要程度.

2.3.2 类别注意力损失

如果只依赖替代模型的损失函数, 生成的对抗扰动会过拟合于替代模型, 这也是目前迁移攻击方法的一个弊病. 因此为了提升攻击的迁移能力, 除了使用替代模型的损失函数之外, 本文还额外增加了类别注意力损失. 不同的黑盒模型拥有不同的网络结构, 这也导致在迁移攻击时, 攻击效果很难保证. 但是有研究[37,38]发现, 虽然网络结构直接影响对抗扰动的分布, 但是黑盒模型网络对于目标物的注意力却是相似的. 模型在提取特征时, 会将注意力放在目标物上. 基于这一点, 本文尝试修改模型的注意力.

受到 CAM[55]、Grad-CAM[56]和 Grad-CAM++[57]的启发, 本节将 Grad-CAM[56]方法从图像分类任务迁移到了目标检测任务中. 将不同类别的热力图进行可视化后, 结果如图 6 所示. 在目标检测算法中, 属于相同类别的不同目标物具有极为相似的热力图, 如图 6 中类别为人的不同目标热力图, 在通道维度具有较为相似的重要程度; 而对于不同的类别而言, 热力图存在较大的差异, 如图 6 中不同类别间的热力图, 其特征通道间会有不同的重要程度.

原图 类别: 人 (目标 1) 类别: 人 (目标 2) 类别: 人 (目标 3) 类别: 蛋糕 (目标 1) 类别: 餐桌 (目标 1)

图 6 类别热力图

根据任务不同, 本文尝试增加和减小这些带有权重特征图之间的距离. 利用梯度信息计算通道的重要程度, 对于每一个类别而言其上一层的特征层的重要程度均会有偏差, 所以如果可以利用这个注意力权重, 便可以提升攻击的迁移能力. 首先将原始图片 $\mathcal{X}_{\text{benign}}$ 输入替代模型 $\mathcal{M}_{\phi}$ 中, 经过 NMS (non maximum suppression) 后得到预测结果, 在预测结果中找到定向类别 C, 便可以通过如下公式计算类别 C 的热力图. 令 $\mathcal{A}$ 为替代模型 $\mathcal{M}_{\phi}$ 中倒数第 2 层特征图, C 为类别标签, 给定一个输入 $\mathcal{X}_{\text{benign}}$, 即可计算类别 C 在模型 $\mathcal{A}$ 中的注意力图 $\mathcal{M}^C$, 计算攻击如下:

$$\mathcal{R}_{\mathcal{K}}^{C}=\frac{1}{Z}\sum_{i}\sum_{j}\frac{\partial \mathcal{P}^{C}}{\partial \mathcal{A}_{ij}^{\mathcal{K}}} \tag{11}$$

$$\mathcal{M}^{C}=\text{ReLU}\left(\sum_{k}\mathcal{R}_{\mathcal{K}}^{C}\cdot\mathcal{A}^{\mathcal{K}}\right) \tag{12}$$

其中, $\mathcal{R}_{\mathcal{K}}^{C}$ 是类别 C 在第 $\mathcal{K}$ 张特征图上的权重, Z 是特征图的像素数量, $\mathcal{P}^C$ 代表模型 $\mathcal{M}_{\phi}$ 对于类别 C 的预测分数, $\mathcal{A}_{\text{ij}}^{\mathcal{K}}$ 代表 $\mathcal{A}$ 中第 $\mathcal{K}$ 张特征图位置为 ij 的像素值. 利用梯度的反向传播, 通过公式 (11) 将梯度在宽度和高度维度上进行全局平均池化, 以此可以计算出类别 C 在第 $\mathcal{K}$ 张特征图中的权重 $\mathcal{R}_{\mathcal{K}}^{C}$. 之后使用一个 ReLU 函数计算加入权重的特征图 $\mathcal{M}^C$. 由于类别在通道上的权重存在差异, 所以在热力图中会产生不一样的热点区域. 在定向攻击和消失攻击时, 本文无需考虑不同类别间的差异, 只需利用生成网络输出扰动, 使其能够打乱并重新构建适配攻击的通道权重. 通过类别注意力损失, 增强扰动在黑盒模型上的攻击性.

2.3.3 定向攻击任务损失

在训练阶段, 定向攻击的损失函数包括两个部分, 类别注意力损失 $\mathcal{L}_{\text{att\_cls}}$ 以及替代模型损失 $\mathcal{L}_{\mathcal{M}_{\phi}}$. 其中类别注意力损失具体定义如下:

$$\mathcal{V}_{\mathcal{K}}^{C}=\frac{1}{N}\sum_{N}\mathcal{R}_{\mathcal{K}}^{C} \tag{13}$$

$$\mathcal{M}_{\ell}^{C}=\text{ReLU}\left(\sum_{k}\mathcal{V}_{\mathcal{K}}^{C}\cdot\mathcal{A}_{\ell}^{\mathcal{K}}\right) \tag{14}$$

$$\mathcal{L}_{\text{att\_cls}}=\sum_{\ell=1}^{L}\sqrt{(\mathcal{M}_{\ell}^{C}-\mathcal{H}_{\ell}^{C})^{2}} \tag{15}$$

计算时采用了文献 [56] 的方法, 但是该方法主要应用于分类任务, 而分类任务在一张图中只有一个类别, 但是检测任务却包含多个类别. 因此需要计算多个属于类别 C 的目标在特征层上权重的平均值. 公式 (13) 中的 N 代表在该图中属于定向攻击类别 C 的数量, L 则代表不同检测模型输出的多个尺度. 例如在 YOLOv3[10]中, 一共有 3 个尺度的输出. $\mathcal{A}_{\ell}^{\mathcal{K}}$ 则代表模型 $\mathcal{M}_{\phi}$ 第 ℓ 尺度下倒数第 2 层特征图中第 $\mathcal{K}$ 张特征图. $\mathcal{M}_{\ell}^{C}$ 代表原始图片得到的带有权重的特征图, $\mathcal{H}_{\ell}^{C}$ 代表对抗样本得到的带有权重的特征图. 定向攻击整体损失 $\mathcal{L}_{\text{tar}}$ 定义如下:

$$\mathcal{L}_{\text{tar}}=\mathcal{L}_{\mathcal{M}_{\phi}}+\mu[\mathcal{T}_{C}\geqslant 1]\mathcal{L}_{\text{att\_cls}} \tag{16}$$

其中, $[\mathcal{T}_{C}\geqslant 1]$ 表示, 如果当前图片中存在类别 C 的目标, 则该值为 1, 否则为 0. 在使用 YOLOv3[10]作为替代模型时, 本文设定 α=1, β=1, γ=0.2, μ=1×10^{-4}. 在使用 Faster R-CNN[8]作为替代模型时, 设定 τ=0.5, η=1, μ=1×10^{-4}.

2.3.4 消失攻击任务损失

在进行消失攻击时, 本文也尝试了加入类别注意力损失. 但是和定向攻击不同, 消失攻击无法针对特定类别计算损失, 因此选择图中目标最多的类别C, 如果目标数量都一致, 选择该图中预测得分最高的类别C, 计算类别C的热力图. 在训练阶段, 消失攻击的损失函数包括类别注意力损失$\mathcal{L}_{\text{att_cls}}$以及替代模型损失$\mathcal{L}_{\mathcal{M}_\phi}$, 损失函数$\mathcal{L}_{\text{vanish}}$定义如下:

$$\mathcal{L}_{\text{vanish}} = \mathcal{L}_{\mathcal{M}_\phi} - \mu[\mathcal{T} \geqslant 1]\mathcal{L}_{\text{att\_cls}} \tag{17}$$

注意力损失$\mathcal{L}_{\text{att_cls}}$和定向攻击定义一致, 但是在消失攻击时, 需要将类别注意力损失最大化. 因此和定向攻击不一样, $[\mathcal{T} \geqslant 1]$代表图中是否存在目标物, 如果存在则为 1, 否则为 0. 在进行消失攻击任务时, 将参数设置为: 在使用 YOLOv3[10]作为替代模型时, 设定$\alpha = 1 \times 10^{-2}$, $\beta = 1 \times 10^{-2}$, $\gamma = 1$, $\mu = 1 \times 10^{-4}$; 在使用 Faster R-CNN[8]作为替代模型时, 设定$\tau = 1$, $\eta = 0.5$, $\mu = 1 \times 10^{-4}$.

2.4 算法流程

设计后的攻击算法整体流程定义如算法 1.

算法 1. 攻击算法.

输入: 训练数据$\mathcal{X}$, 训练定向攻击的类别标签$\mathcal{Y}_{\text{class}}$, 定向攻击类别$C \in \mathcal{Y}_{\text{class}}$, 生成网络$\mathcal{G}_\theta$, 替代模型$\mathcal{M}_\phi$, 映射网络$\mathcal{W}$, 攻击任务 *T*, 训练最大轮次 *E*;
输出: 训练完成的生成模型$\mathcal{G}_\theta$.

```
for epoch in E do:
  //定向攻击任务
  if T:
    randomly sample target classes C in Y_class
    forward pass C to W , and compute the latent w_C
    use Eq.(4) get the X_adv^C
    forward pass X_adv^C to M_φ
  //消失攻击任务
  else:
    use Eq.(5) get the X_adv^vanish
    forward pass X_adv^vanish to M_φ
  //定向攻击任务
  if T:
    compute loss in Eq.(16)      //在定向攻击中, 将所有参与损失计算的标签定义为类别C
   //消失攻击任务
  else:
    compute loss in Eq.(17)      //在消失攻击时, 将所有参与损失计算的标签定义为背景类
  backward pass and update the G_θ
```

3 实 验

本节在两个目标检测数据集进行实验, 评估本文所提出的攻击算法. 第 3.1 节中介绍实验的数据集, 第 3.2 节介绍评价指标, 第 3.3 节介绍实验中用到的目标检测网络, 第 3.4 节介绍攻击算法及相关实验配置, 第 3.5 节进行多组实验, 包括定向攻击实验、消失攻击实验、数据集迁移攻击以及类别注意力损失实验.

3.1 数据集

本节在两个公开数据集上进行迁移攻击实验, 分别是 COCO[33]以及 VOC[34]. COCO 数据集包含 91 个类别, 用于目标检测的有 80 个类别, VOC 中有 20 个类别用于目标检测. 在测试时, 本节采用了各自的测试集, 其中 COCO 测试集共 5 000 张, VOC 测试集共 4 952 张.

在实验中, 本文没有对所有类别进行实验. 为了测试定向攻击任务在迁移攻击的表现, 本文在 COCO 和 VOC 数据集共有类别中, 选取了 6 个类别. 对表 1 中的类别进行实验, 验证本文所提出的对抗方法在数据集之间的迁移能力, 类别及对应编号如表 1 所示.

表 1 实验数据集

class_name	class_num (COCO)	class_num (VOC)
boat	9	4
bus	6	6
car	3	7
dog	17	12
person	1	15
train	7	19

3.2 计算指标

为了验证本文提出算法的攻击性能, 本文采用了两种评价指标, 分别是攻击成功率 ASR (attack success rate) 和最常用的目标检测评价标准 mAP (mean average precision). 这两种指标在其他研究中[19,25,26,35]被广泛应用为评价标准.

在定向攻击时, 本文使用攻击成功率来评估攻击算法在黑盒模型的攻击性能. 在定向攻击任务中, 只需要关心黑盒模型是否能够将目标物错误识别为定向攻击的目标标签. 在实验中会首先根据标签剔除属于定向攻击类别的目标物, 将剩下的目标物用于计算成功率. 判断黑盒模型是否能够将剩余目标物预测成定向类别的依据是, 保证真实框和预测框的交并比大于 0.5. 如果满足则攻击成功, 否则攻击失败.

在消失攻击时, 本文同时采用 ASR 以及 mAP 作为评价标准. 但是和文献 [27] 不同, 为了更严格地评估本文方法在攻击不同模型时的鲁棒性, 选取了一个较低的置信度阈值: 单阶段网络采用 0.2, 双阶段以及 Transformer_Based 网络则采用 0.4. 如果被攻击模型能够在对抗样本上识别出原目标物类别, 并且真实框和预测框的交并比大于 0.5, 则攻击失败, 否则攻击成功.

3.3 目标检测网络

在测试攻击算法对黑盒模型的迁移能力时, 本文选取了 11 个目标检测模型. 模型结构及模型的性能如表 2 所示.

表 2 目标检测算法表

模型编号	模型	骨架网络	mAP (%)	模型属性
N1	YOLOv3[10]	Darknet	32.7	one-stage
N2	YOLOv3-spp[10]	Darknet	35.6	one-stage
N3	YOLOv5x[58]	CSPDarknet53	50.4	one-stage
N4	YOLOv5l[58]	CSPDarknet53	48.2	one-stage
N5	SSD300[9]	VGG16	25.1	one-stage
N6	RetinaNet[59]	ResNet50_FPN	36.4	one-stage
N7	Faster R-CNN[8]	ResNet50_FPN_COCO	37.0	two-stage
N8	Faster R-CNN[8]	MobileNet_FPN_COCO	32.8	two-stage
N9	Mask R-CNN[49]	ResNet50_FPN_COCO	37.9	two-stage
N10	DETR[50]	ResNet50	42.0	Transformer_Based
N11	DETR[50]	ResNet101	43.5	Transformer_Based

3.4 攻击算法配置

本文经过对攻击算法的分析后, 选取了 PGD[36]、TIM[20]、DAG[25]、RAP[19]、TOG[27]这 5 种常用攻击算法和本文的方法进行对照试验. 在定向攻击任务和消失攻击任务中, 采用了不同的替代模型进行训练. 为了探究注意力机制对于扰动的影响, 本文选择了 SKNet[60]中 SKCONV 模块作为 CBAM[53]模块的对照试验. 本文利用 CBAM[53]验证通道注意力和空间注意力对于扰动的影响; 利用 SKNet[60]中 SKCONV 验证在感受野上自适应注意力对于扰动的影响.

攻击方法设置如表 3 所示.

表 3 攻击算法配置表

攻击算法编号	攻击方法	替代模型	备注
A1	PGD[36]	N1	$\epsilon = 16$
A2	PGD[36]	N7	$\epsilon = 16$
A3	TIM[20]	N1	$\epsilon = 16$
A4	TIM[20]	N7	$\epsilon = 16$
A5	DAG[25]	N7	—
A6	RAP[19]	N7	—
A7	TOG[27]	N7	mislabel
A8	TOG[27]	N7	vanish
A9	(ours)	N1	$\epsilon = 16$ CBAM
A10	(ours)	N7	$\epsilon = 16$ CBAM
A11	(ours)	N1	$\epsilon = 16$ SKCONV

针对攻击算法本文选取两种基于优化梯度的攻击算法和 3 种针对目标检测的攻击算法, 为了公平起见, 本论文选取各个攻击算法迭代 50 次之后的攻击结果作为评估依据. 在设计生成模型时, 本文参考了文献 [30] 的网络结构. 在训练生成网络时, 选用 Adam (adaptive moment estimation)[61]作为优化器, 学习率设置为 1×10^{-5}, 每一批次的数量设置为 16.

3.5 实验结果与分析

为了评估本文提出的黑盒攻击算法的有效性, 本节将实验分为两个部分. 首先评估本方法在不同模型和不同数据集上的迁移能力, 其次评估本方法当中的类别注意力损失在模型框架中的有效性, 同时在进行消融实验时采用了表 3 中的对抗算法作为对照实验, 选择文献 [27] 为基准.

3.5.1 定向攻击实验

在定向攻击实验中, 对 COCO 中的 6 个类别分别进行定向攻击实验. 对每一组实验, 将 6 个类别的 ASR 取平均值, 来代表该攻击算法在某个模型中的攻击结果, 实验结果如表 4 所示.

表 4 定向攻击实验结果 (%)

攻击方法	N1	N2	N3	N4	N5	N6	N7	N8	N9	N10	N11
A1	93.75	69.48	54.28	54.87	10.77	49.34	17.96	16.18	17.53	3.87	3.25
A2	40.97	43.94	44.17	45.66	12.97	89.28	88.40	28.36	88.46	6.21	5.33
A3	92.10	74.97	74.25	74.73	13.59	49.62	22.53	18.57	22.01	3.59	3.47
A4	61.42	59.80	62.13	62.94	19.85	87.43	90.62	28.20	91.92	5.45	5.24
A5	35.85	34.86	39.08	39.10	10.34	71.08	88.54	14.99	82.66	2.23	2.21
A6	33.18	34.97	44.00	38.19	7.34	81.25	79.88	19.24	73.59	6.77	4.54
A7	68.99	65.25	77.29	65.45	25.70	**91.67**	94.55	40.84	90.93	9.45	**10.60**
A9	97.96	**97.12**	**97.09**	**96.60**	29.61	52.65	57.35	51.59	44.88	8.81	8.29
A10	88.43	89.25	89.02	88.87	**32.23**	89.82	**95.82**	**54.40**	**91.35**	**10.16**	10.42
A11	**98.72**	96.94	95.19	95.81	24.65	48.75	54.47	49.34	45.08	5.52	6.49

通过以上实验可以发现, 所有攻击算法在攻击白盒替代模型时, 均有一个很好的攻击效果, 成功率最低79.88% (A6–N7), 白盒的攻击效果都符合预期. 但是对黑盒模型攻击时, 却得到了不同的结果. A2, A4, A5, A6, A7, A10 均是以 N7 作为白盒替代模型. A1, A3, A9, A11 均是以 N1 作为白盒替代模型. 通过攻击结果可以发现, 在攻击具有类似骨架网络的黑盒模型时, 攻击成功率会比其他黑盒模型有很大提升. 例如 A1 在攻击 YOLO 系列网络时, 攻击成功率最低为 54.28% (N3). 在攻击 RCNN 系列网络时, 最高攻击成功率仅为 17.96% (N7); A2 在攻击 YOLO 系列网络时, 攻击成功率最高为 45.66% (N4), 在攻击 RCNN 系列网络时最低攻击成功率为 28.36% (N8), 这是因为 N8 的骨架网络是 MobileNet, 但是 N6 (89.28%), N7 (88.40), N9 (88.46%) 均是以 ResNet 作为骨架网络. 由此可以看出, 白盒替代模型的结构直接影响迁移攻击的结果.

对比 A10 和 A7, 可以发现本文提出的方法在黑盒迁移攻击能力上要优于文献 [27] 的方法. A7 虽然是目前目标检测攻击能力较好的攻击方法, 但是在对相似的网络结构进行攻击时, 才能得到一个较高的攻击成功率. 在对网络结构差异较大的黑盒模型时, 例如在攻击 N5 和 N8 时, A10 可以达到 32.23% 和 54.4% 的攻击成功率, 而 A7 只能达到 25.70% 和 54.40% 的攻击成功率. A9 和 A11 在对于 YOLO 系列的网络攻击成功率达到最高, 虽然在攻击以 ResNet 为骨架的网络时略有不足, 但是 A9 和 A11 在 N5 和 N8 的迁移攻击能力上也略胜于 A7. A9 和 A11 利用了不同的注意力机制, 通过攻击成功率可以发现: 利用 SKCONV[60]时, A11 在白盒的攻击效果最优, 感受野上的注意力可以增加扰动的白盒攻击能力; 利用 CBAM[53]时, A9 在黑盒迁移攻击效果更优. 在本节的实验中, 空间与通道上的注意力是更有助于增加扰动的迁移攻击能力, 而感受野上的注意力则是更加有助于扰动在白盒模型的攻击能力.

为了更直观地展示攻击效果, 将 A10 在 N5 的攻击结果进行了可视化, 结果如图 7 所示.

原始图片　原始图片预测　第 1 轮预测结果　第 5 轮预测结果　第 10 轮预测结果

图 7　网络预测结果

将生成模型第 1 轮, 第 5 轮和第 10 轮的预测结果进行可视化. 可以发现模型在第 1 轮时, 降低了图中所有目标物的置信度, 在多经过轮次的训练后, 通过加入生成的扰动, 逐步将图中的目标物攻击为 Boat 类.

通过对本节实验的分析, 在定向攻击中, 所有攻击方法均无法在 N5, N8, N10, N11 的模型中取得较高的成功率. 这些模型在本文的攻击实验中具有较强的鲁棒性. 一部分因素是由于攻击算法选取的替代模型骨架网络不相似, 例如 N5, N8 采用了不同的骨架网络. 另一部分因素则是由于网络本身的结构设计, 例如 N10, N11 都是基于 Transformer 结构. 从实验的最后两列可以发现, 上述攻击方法都无法在 Transformer_Based 的模型上取得一个很高的攻击成功率, 最高只能达到 10.60%. 这是因为 Transformer_Based 的网络结构与目前检测网络结构差异过大, 导致迁移攻击的效果大打折扣.

通过对定向攻击实验结果分析, 可以发现不同类别的攻击成功率存在差异, 因此本节汇总了所有模型的定向攻击的结果, 然后统计定向攻击类别平均攻击成功率, 统计结果如图 8 所示.

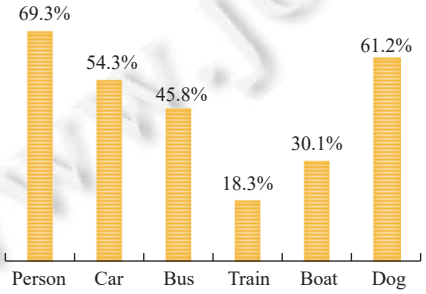


图 8　各类攻击成功率

图 8 中 person, dog 和 car 是相对容易攻击的类别, 平均攻击成功率均在 50% 以上, 而其他类别的攻击成功率却低于 50%. Train 是最难攻击的类别, 攻击成功率仅有 18.3%. 在不同类别上呈现了不同的攻击成功率. 这种情况可能是模型对于某些类别鲁棒性较低, 因此再定向攻击时, 添加的扰动不足将原有分布攻击到鲁棒性差的类别分布上, 导致很多情况下无法攻击成功, 而且类别之间的影响也不利于攻击.

通过上述的实验结果, 可以发现在定向攻击实验中, 利用生成网络的攻击方法在迁移能力上要优于基于梯度优化的方法. 网络中加入的高斯平滑映射, 有助于生成的对抗样本在不同模型间的迁移. 本文提出的算法使用训练集训练, 使用测试集测试. 通过实验结果可以发现生成网络学习的扰动有效的从训练集迁移到了验证集, 在迁移能力方面均优于实验中的其他攻击算法. 在进行推理时只需运行一次神经网络, 因此本方法在攻击时运行速度快, 解决目前基于多步梯度迭代方法效率较低的问题.

3.5.2 消失攻击实验

除了在定向攻击任务上进行实验, 本节还增加了消失攻击实验. 不仅统计了攻击的成功率 ASR, 也统计了 mAP 的结果. 表 5 统计了攻击成功率的结果, 表 6 统计了 mAP的结果.

表 5 消失攻击成功率结果表 (%)

攻击方法	N1	N2	N3	N4	N5	N6	N7	N8	N9	N10	N11
A1	82.35	17.22	6.10	5.91	1.68	0.94	1.69	3.00	1.70	10.19	8.43
A2	3.53	3.29	2.98	3.16	1.70	17.09	69.69	4.04	49.61	14.02	11.01
A3	89.39	8.10	6.56	7.48	2.69	1.87	3.68	4.68	3.50	10.98	10.52
A4	2.92	2.54	2.43	2.45	2.73	3.24	25.20	4.94	10.84	11.12	10.89
A5	1.56	1.45	1.36	1.27	0.83	9.61	86.23	9.88	72.35	18.68	19.62
A6	2.64	1.85	0.89	1.13	1.88	8.65	85.11	7.54	78.97	20.37	**22.16**
A8	2.98	2.11	1.24	0.92	2.51	10.74	**91.34**	14.23	**83.14**	17.62	19.44
A9	92.80	76.56	**66.41**	**74.19**	**19.63**	16.66	24.29	23.16	21.87	19.38	18.93
A10	46.70	44.23	39.41	41.97	14.59	**25.80**	77.32	**28.79**	50.65	**21.29**	20.54
A11	**94.15**	**80.25**	54.23	53.18	15.24	17.18	21.24	19.37	18.41	18.06	16.63

表 6 消失攻击 mAP 结果表 (%)

攻击方法	N1	N2	N3	N4	N5	N6	N7	N8	N9	N10	N11
ori	32.7	35.6	50.4	48.2	25.1	37	32.8	36.4	37.9	42	43.5
A1	6.7	33.8	49.1	47.1	24.9	36.8	32.1	35.0	37.1	39.3	40.1
A2	31.6	34.4	49.6	45.3	24.2	31.4	17.8	35.1	19.6	37.9	39.4
A3	5.6	33.6	49.2	47.3	23.8	36.5	31.4	35.2	36.3	38.7	39.6
A4	32.4	34.9	48.5	47.9	23.9	35.8	28.9	34.9	29.6	39.1	39.4
A5	32.3	35.2	49.8	47.2	25.0	35.4	5.8	31.3	13.2	36.7	36.1
A6	32.1	35.3	50.2	48.0	24.8	34.1	6.0	33.7	11.3	35.4	35.8
A8	32.4	35.1	49.0	47.9	24.5	34.7	**3.9**	30.4	**9.7**	37.0	35.4
A9	1.2	5.3	**3.1**	**2.8**	18.4	33.1	28.1	29.7	25.1	36.2	35.6
A10	22.4	26.0	36.3	39.1	21.8	**26.5**	17.5	**27.8**	21.3	**34.8**	**34.1**
A11	**1.0**	**4.7**	4.2	3.5	**18.3**	32.7	29.3	31.0	28.4	38.9	39.5

从表 5 中可以看出 A1 和 A3 在 N1 的攻击成功率达到了 82.35% 和 89.39%, 这是因为 A1, A3 的替代模型均是 N1. 但是在对其他黑盒模型进行攻击时, 最高的攻击成功率仅为 17.22%. A2 和 A4 的攻击情况和 A1, A3 近似. A5, A6, A8 和 A10 均是采用 N7 作为替代模型, 但是在对其他模型的迁移攻击时, A10 的综合表现是最好的, 虽然也会对替代模型产生过拟合的情况, 但是对于 N6, N8, N10 的迁移攻击结果是最优的, 分别达到了 25.8%, 28.79%, 21.29%. A9 和 A11 在 N5, N6, N8 的攻击实验中, 攻击成功率要比 A1–A8 的成功率高, 在 N1–N4 上的攻击表现也是较优的. A6 在 N11 的攻击结果是最优的, 达到 22.16%. 但是从 mAP 的结果看却不是下降最大的, 下降 7.7. 通过对比定向攻击和消失攻击, A9 和 A11 的加入注意力的方法在攻击方面都比较有效. 但是 A9 的侧重点是黑盒攻

击, A11 的侧重点则是白盒攻击.

从表 6 中可以看到, mAP 的结果和 ASR 的结果是存在线性关系的.

通过这些实验结果可以发现, 本文提出的生成式攻击方法, 相比于其他攻击方法, 在黑盒迁移攻击能力上有一定的提升. 本节通过 ASR 和 mAP 共同评价消失攻击任务在黑盒迁移攻击的表现. 本文提出的网络可以降低扰动对于替代模型的敏感度, 由于深度神经网络可以自主学习到数据中的特征表示, 泛化能力具有一定提升, 因此其所生成的对抗样本更加本质, 也有利于欺骗不同网络结构的深度学习模型, 增强不同模型的攻击迁移能力. 而基于优化的迭代算法, 会严重过拟合于替代模型的网络结构, 不利于鲁棒性的评测. 在消失攻击的实验中, 上述攻击方法在攻击 N5, N6, N8, N10, N11 的模型中, 均无法取得较高的攻击成功率. 在本文的实验中, 这些模型拥有较强的鲁棒性. N6 在定向攻击中鲁棒性较差, 但是在消失攻击中保留了较强的鲁棒性, 这也说明 N6 能够较准确地区分前景和背景, 但是 RetinaNet[59]属于单阶段检测网络, 并没有较为鲁棒的识别具体类别的能力.

通过对定向攻击和消失攻击的整体分析, 可以发现大多数单阶段的网络更容易被攻击, 因为单阶段网络对于前景和背景的判断远不及双阶段网络的精度. 基于 Transformer 的 DETR[50]网络则很难被攻击, 因为该网络具有不同的网络结构, 计算每一个像素和其他像素的相关性, 并不会受到局部感受野造成的影响, 扰动的攻击性被大大降低, 因此相对而言较为鲁棒.

3.5.3 数据集迁移实验

在本节的实验中, 进行了数据集迁移攻击的实验. 在两个数据集上, 使用第 3.1 节选取的 6 个类别进行定向攻击, 并对比了 A2, A7, A9, A10 在 N1 和 N7 上的表现. 其中 N1_COCO 代表 N1 在 COCO 预训练的模型, A2_COCO 代表利用 A2 的方法在 COCO 数据集上生成对抗样本. 实验结果如表 7 所示.

表 7 数据迁移攻击结果 (%)

数据集	N1_COCO	N7_COCO	N1_VOC	N7_VOC
A2_COCO	40.97	88.40	8.45	20.62
A7_COCO	68.99	91.67	6.25	18.92
A2_VOC	5.94	8.27	64.32	92.76
A7_VOC	3.22	10.24	70.43	96.91
A9_COCO	97.96	57.35	34.98	22.31
A10_COCO	88.43	95.82	31.12	32.76
A9_VOC	20.54	13.31	99.16	68.58
A10_VOC	18.12	22.21	89.54	97.40

利用 A2 在 COCO 中生成的扰动迁移攻击 VOC 时, 攻击成功率平均损失占比 78.0% (N1 损失占比 79.37%, N2 损失占比 76.67%). A7 在 COCO 中生成的扰动迁移攻击 VOC 时, 攻击成功率平均会损失占比 85.1% (N1 损失占比 90.94%, N2 损失占比 79.36%). 反之, A2 在 VOC 中生成的扰动迁移攻击 COCO 时, 攻击成功率平均会损失占比 90.92% (N1 损失占比 90.76%, N2 损失占比 91.08%), A7 在 VOC 中生成的扰动迁移攻击 COCO 时, 攻击成功率平均会损失占比 92.42% (N1 损失占比 95.42%, N2 损失占比 89.43%). 相比较而言, 本文提出的生成式的方法在数据集迁移攻击上, 保持一定的攻击效果, 分析发现利用 A9, A10 在 COCO 向 VOC 迁移实验中, 攻击成功率损失占比分别是 62.69%, 65.20%. 反之 A9, A10 在 VOC 向 COCO 迁移实验中, 攻击成功率损失占比分别是 79.93%, 78.47%. 通过对比可以发现本文提出的迁移攻击方法, 要比现有的方法更优.

实验结果可以发现, 在 COCO 中生成的扰动迁移到 VOC 中会存在一定的损失. 在 VOC 生成的扰动迁移到 COCO 数据集中, 这种损失会增大. 这是因为 COCO 中共包含 80 个类别, 而 VOC 只有 20 个类别, 共有的类别会呈现相似的分布, 而不相关的类别则分布差异较大. 添加的扰动会破坏原来目标的语义信息. A2, A7 均无法实现语义层面的迁移攻击, 而本文提出的生成网络则可以捕捉不同类别的语义信息, 添加对应的扰动. 因此可以在数据集的迁移攻击中保留一定的攻击能力, 但是由于两个数据集会存在各自独有的类别, 这些类别分布不同, 所以攻击并没有出现较高的攻击成功率.

3.5.4 注意力损失实验

在第 2.3.2 节的图 6 中通过 Grad-CAM[56]方法可视化了不同类别及相同类别的热力图. 通过热力图的形式反映出在通道维度上, 不同类别和相同类别与通道的关联关系. 为了更好地展示类别与通道重要程度之间的关系, 本节使用了 Pearson correlation coefficient[62]算法对公式 (11) 中的 $\mathcal{R}^C$ 进行了相关性分析, $\mathcal{R}^C$ 代表了类别 C 在通道维度的权重. 计算并展示了不同类别在通道维度上重要程度的相关程度.

在进行相同类别的计算时, 选取了 6 张均包含 person 类别的图片. 计算每一张图片中所有 person 类的通道权重并取平均值, 通过 Pearson correlation coefficient[62]算法计算通道权重的相关性, 结果如图 9 所示.

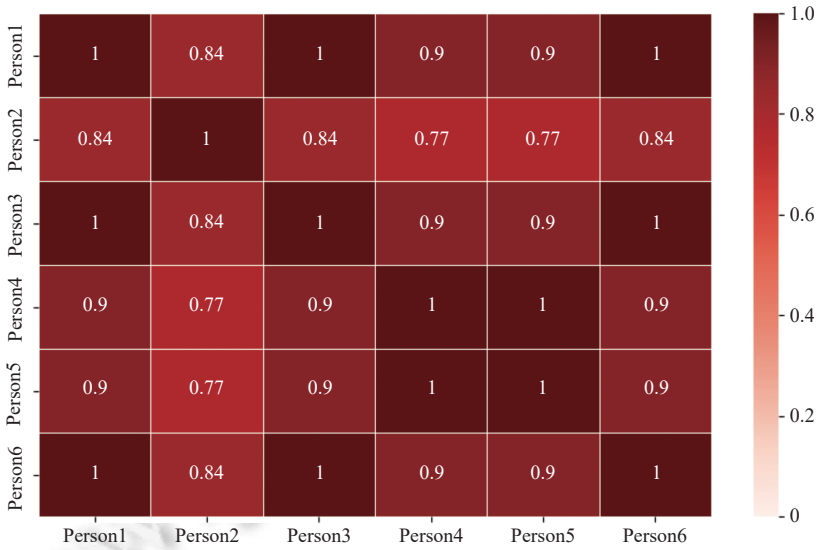


图 9 同类别通道重要程度相关性分析

根据图 9 中的结果可以发现, 相同类别在通道维度上权重具有相似的重要程度. 在进行不同类别的计算时, 每个类别选取了 20 张均包含当前类别的图片. 计算每个类别下所有选取的图片的通道权重并取平均值, 通过 Pearson correlation coefficient[62]算法计算通道权重的相关性, 结果如图 10 所示.

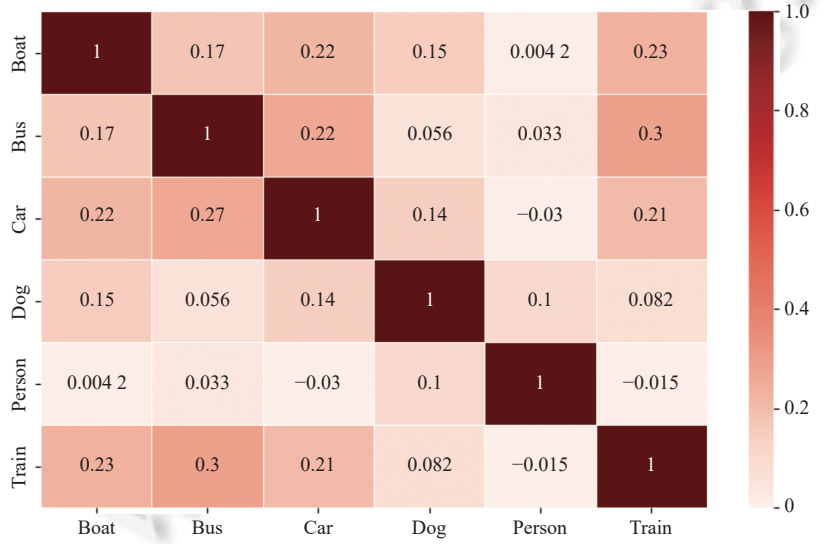


图 10 不同类别通道重要程度相关性分析

根据图 10 中的结果可以发现, 不同类别在通道维度上的重要程度均不相同, 并没有展示出较强的相关性. 本文提出的类别注意力损失帮助扰动在定向攻击和消失攻击时, 打乱并重新构建不同的通道权重, 增强扰动在黑盒模型上的攻击性.

为了验证本文所提出的类别注意力损失在黑盒攻击时的表现, 增加了对比实验. 在定向攻击和消失攻击任务中进行实验对比, 实验结果如表 8 所示.

表 8 类别注意力损失实验结果 (%)

类别	N1	N7	N10
A9_with_tar	97.96	57.35	8.29
A9_without_tar	95.62	55.17	8.10
A9_with_untar	92.80	24.29	18.93
A9_without_untar	92.41	23.83	18.68

本节使用了 A9 作为攻击算法, 并在 N1, N7, N10 上进行攻击测试. 从表 8 中的结果可以看出, 加入类别注意损失后, 在定向攻击任务和消失攻击任务中, 攻击成功率均有提升. 在定向攻击任务中, 分别提升了 2.34%, 2.18%, 0.19%; 在消失攻击任务中分别提升了 0.39%, 0.46%, 0.25%.

通过分析实验结果可以发现, 本文提出的类别注意力损失能够在一定程度上提升攻击的迁移能力. 该损失在指导生成网络训练时, 帮助网络抓住重要特征, 增强在多个黑盒模型上的迁移攻击能力. 能够提升攻击的迁移能力的主要原因是不同模型在对相同类别的注意力是相似的, 如果改变共有的注意力, 就可以实现在不用模型之间的迁移攻击, 这也是增强黑盒迁移攻击的方法之一. 但是也可以发现在白盒攻击的提升要高于在其他黑盒攻击提升, 这是因为本文的方法尽管缓解对于替代模型的依赖, 但依旧靠替代模型的梯度信息. 这也是目前黑盒迁移攻击所共有的弊病.

4 总 结

本文提出了基于生成对抗网络的目标检测黑盒攻击算法, 利用一个生成网络可以进行定向攻击以及消失攻击. 为了提升攻击的效率以及迁移攻击的能力, 本文还提出了类别注意力损失. 为了验证本文提出的算法, 在 COCO 和 VOC 两个数据集上进行了多组实验, 实验结果表明文中所提出的方法较优于其他攻击算法. 本文的实验弥补了在数字世界中, 对于目标检测定向攻击研究领域的空白, 算法具有较强的黑盒迁移攻击能力, 能够运用到不同黑盒攻击中, 具有一定的普适性.

当然算法也有不足, 在迁移攻击时还是会依赖替代模型的梯度信息. 并且在定向攻击时, 由于多个类别的相互影响, 会损失部分的迁移能力. 在未来的工作中, 我们会继续探索黑盒迁移攻击领域, 并尝试将本文所提出的方法迁移到物理世界中, 提出更加鲁棒的黑盒迁移攻击方法, 推动深度学习对抗攻击的发展.

References:

[1] Li JZ, Su H, Zhu J, Wang SY, Zhang B. Textbook question answering under instructor guidance with memory networks. In: Proc. of the 2018 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 3655–3663. [doi: 10.1109/CVPR.2018.00385]

[2] Gong ZQ, Zhong P, Yu Y, Hu WD, Li ST. A CNN with multiscale convolution and diversified metric for hyperspectral image classification. IEEE Trans. on Geoscience and Remote Sensing, 2019, 57(6): 3599–3618. [doi: 10.1109/TGRS.2018.2886022]

[3] Gong ZQ, Zhong P, Hu WD. Statistical loss and analysis for deep learning in hyperspectral image classification. IEEE Trans. on Neural Networks and Learning Systems, 2021, 32(1): 322–333. [doi: 10.1109/TNNLS.2020.2978577]

[4] Albert A, Kaur J, Gonzalez MC. Using convolutional networks and satellite imagery to identify patterns in urban environments at a large scale. In: Proc. of the 23rd ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. Halifax: Association for Computing Machinery, 2017. 1357–1366. [doi: 10.1145/3097983.3098070]

[5] Pritt M, Chern G. Satellite image classification with deep learning. In: Proc. of the 2017 IEEE Applied Imagery Pattern Recognition Workshop (AIPR). Washington: IEEE, 2017. 1–7. [doi: 10.1109/AIPR.2017.8457969]

[6] Zhao ZQ, Zheng P, Xu ST, Wu XD. Object detection with deep learning: A review. IEEE Trans. on Neural Networks and Learning Systems, 2019, 30(11): 3212–3232. [doi: 10.1109/TNNLS.2018.2876865]

[7] Joseph KJ, Khan S, Khan FS, Balasubramanian VN. Towards open world object detection. In: Proc. of the 2021 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021. 5826–5836. [doi: 10.1109/CVPR46437.2021.00577]

[8] Ren SP, He KM, Girshick R, Sun J. Faster R-CNN: Towards real-time object detection with region proposal networks. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2017, 39(6): 1137–1149. [doi: 10.1109/TPAMI.2016.2577031]

[9] Liu W, Anguelov D, Erhan D, Szegedy C, Reed S, Fu CY, Berg AC. SSD: Single shot multibox detector. In: Proc. of the 14th European Conf. on Computer Vision. Amsterdam: Springer, 2016. 21–37. [doi: 10.1007/978-3-319-46448-0_2]

[10] Redmon J, Farhadi A. YOLOv3: An incremental improvement. arXiv:1804.02767, 2018.

[11] Yuan XH, Shi JF, Gu LC. A review of deep learning methods for semantic segmentation of remote sensing imagery. Expert Systems with Applications, 2021, 169: 114417. [doi: 10.1016/j.eswa.2020.114417]

[12] Eykholt K, Evtimov I, Fernandes E, Li B, Rahmati A, Xiao CW, Prakash A, Kohno T, Song D. Robust physical-world attacks on deep learning visual classification. In: Proc. of the 2018 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 1625–1634. [doi: 10.1109/CVPR.2018.00175]

[13] Grigorescu S, Trasnea B, Cocias T, Macesanu G. A survey of deep learning techniques for autonomous driving. Journal of Field Robotics, 2020, 37(3): 362–386. [doi: 10.1002/rob.21918]

[14] Hu Y, Yang A, Li H, Sun YY, Sun LM. A survey of intrusion detection on industrial control systems. Int'l Journal of Distributed Sensor Networks, 2018, 14(8): 1–14. [doi: 10.1177/1550147718794615]

[15] Szegedy C, Zaremba W, Sutskever I, Bruna J, Erhan D, Goodfellow IJ, Fergus R. Intriguing properties of neural networks. In: Proc. of the 2nd Int'l Conf. on Learning Representations. Banff, 2013.

[16] Jia XJ, Zhang Y, Wu BY, Wang J, Cao XC. Boosting fast adversarial training with learnable adversarial initialization. IEEE Trans. on Image Processing, 2022, 31: 4417–4430. [doi: 10.1109/TIP.2022.3184255]

[17] Bai JW, Chen B, Li YM, Wu DX, Guo WW, Xia ST, Yang EH. Targeted attack for deep hashing based retrieval. In: Proc. of the 16th European Conf. on Computer Vision. Glasgow: Springer, 2020. 618–634. [doi: 10.1007/978-3-030-58452-8_36]

[18] Jia XJ, Zhang Y, Wu BY, Ma K, Wang J, Cao XC. LAS-AT: Adversarial training with learnable attack strategy. In: Proc. of the 2022 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. New Orleans: IEEE, 2022. 13388–13398. [doi: 10.1109/CVPR52688.2022.01304]

[19] Wei XX, Liang SY, Chen N, Cao XC. Transferable adversarial attacks for image and video object detection. In: Proc. of the 28th Int'l Joint Conf. on Artificial Intelligence. Macao: AAAI Press, 2019. 954–960.

[20] Dong YP, Pang TY, Su H, Zhu J. Evading defenses to transferable adversarial examples by translation-invariant attacks. In: Proc. of the 2019 IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR). Long Beach: IEEE, 2019. 4307–4316. [doi: 10.1109/CVPR.2019.00444]

[21] Goodfellow IJ, Shlens J, Szegedy C. Explaining and harnessing adversarial examples. In: Proc. of the 3rd Int'l Conf. on Learning Representations. San Diego, 2015.

[22] Moosavi-Dezfooli SM, Fawzi A, Frossard P. DeepFool: A simple and accurate method to fool deep neural networks. In: Proc. of the 2016 IEEE Conf. on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016. 2574–2582. [doi: 10.1109/CVPR.2016.282]

[23] Carlini N, Wagner D. Towards evaluating the robustness of neural networks. In: Proc. of the 2017 IEEE Symp. on Security and Privacy. San Jose: IEEE, 2017. 39–57. [doi: 10.1109/SP.2017.49]

[24] Dong YP, Liao FZ, Pang TY, Su H, Zhu J, Hu XL, Li JG. Boosting adversarial attacks with momentum. In: Proc. of the 2018 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 9185–9193. [doi: 10.1109/CVPR.2018.00957]

[25] Xie CH, Wang JY, Zhang ZS, Zhou YY, Xie LX, Yuille A. Adversarial examples for semantic segmentation and object detection. In: Proc. of the 2017 IEEE Int'l Conf. on Computer Vision. Venice: IEEE, 2017. 1378–1387. [doi: 10.1109/ICCV.2017.153]

[26] Li YZ, Tian D, Chang MC, Bian X, Lyu S. Robust adversarial perturbation on deep proposal-based models. In: Proc. of the 2018 British Machine Vision Conf. Newcastle: BMVA Press, 2018. 231.

[27] Chow KH, Liu L, Loper M, Bae J, Gursoy ME, Truex S, Wei WQ, Wu YZ. Adversarial objectness gradient attacks in real-time object detection systems. In: Proc. of the 2nd IEEE Int'l Conf. on Trust, Privacy and Security in Intelligent Systems and Applications (TPS-ISA). Atlanta: IEEE, 2020. 263–272. [doi: 10.1109/TPS-ISA50397.2020.00042]

[28] Baluja S, Fischer I. Adversarial transformation networks: Learning to generate adversarial examples. arXiv:1703.09387, 2017.

[29] Poursaeed O, Katsman I, Gao BC, Belongie S. Generative adversarial perturbations. In: Proc. of the 2018 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 4422–4431. [doi: 10.1109/CVPR.2018.00465]

[30] Naseer M, Khan S, Khan MH, Khan FS, Porikli F. Cross-domain transferability of adversarial perturbations. In: Proc. of the 33rd Int'l Conf. on Neural Information Processing Systems. Vancouver: Curran Associates Inc., 2019. 1156.

[31] Han JF, Dong XY, Zhang RM, Chen DD, Zhang WM, Yu NH, Luo P, Wang XG. Once a MAN: Towards multi-target attack via learning

multi-target adversarial network once. In: Proc. of the 2019 IEEE/CVF Int'l Conf. on Computer Vision. Seoul: IEEE, 2019. 5157–5166. [doi: 10.1109/ICCV.2019.00526]

[32] Yang X, Dong Y, Pang T, *et al*. Boosting transferability of targeted adversarial examples via hierarchical generative networks. In: Proc. of the 17th European Conf. on Computer Vision. Tel Aviv: Springer, 2022. 725–742. [doi: 10.1007/978-3-031-19772-7_42]

[33] Lin TY, Maire M, Belongie S, Hays J, Perona P, Ramanan D, Dollár P, Zitnick CL. Microsoft COCO: Common objects in context. In: Proc. of the 13th European Conf. on Computer Vision. Zurich: Springer, 2014. 740–755. [doi: 10.1007/978-3-319-10602-1_48]

[34] Everingham M, Van Gool L, Williams C, Winn J, Zisserman A. The PASCAL visual object classes challenge 2007 (VOC2007) results. 2007. http://host.robots.ox.ac.uk/pascal/VOC/voc2007/

[35] Kurakin A, Goodfellow IJ, Bengio S. Adversarial examples in the physical world. Artificial Intelligence Safety and Security. Chapman and Hall/CRC, 2018. 99–112.

[36] Madry A, Makelov A, Schmidt L, Tsipras D, Vladu A. Towards deep learning models resistant to adversarial attacks. In: Proc. of the 6th Int'l Conf. on Learning Representations (ICLR). Vancouver: OpenReview.net, 2018.

[37] Chen SZ, He ZB, Sun CJ, Yang J, Huang XL. Universal adversarial attack on attention and the resulting dataset DAmagenet. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2022, 44(4): 2188–2197. [doi: 10.1109/TPAMI.2020.3033291]

[38] Wu WB, Su YX, Chen XX, Zhao SL, King I, Lyu MR, Tai YW. Boosting the transferability of adversarial samples via attention. In: Proc. of the 2020 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020. 1161–1170. [doi: 10.1109/CVPR42600.2020.00124]

[39] Huang Q, Katsman I, Gu ZQ, He H, Belongie S, Lim SN. Enhancing adversarial example transferability with an intermediate level attack. In: Proc. of the 2019 IEEE/CVF Int'l Conf. on Computer Vision. Seoul, Korea (South): IEEE, 2019. 4732–4741. [doi: 10.1109/ICCV.2019.00483]

[40] Huang Y, Kong AWK. Transferable adversarial attack based on Integrated Gradients. In: Proc. of the 10th Int'l Conf. on Learning Representations. OpenReview.net, 2022.

[41] Lin JD, Song CB, He K, Wang LW, Hopcroft JE. Nesterov accelerated gradient and scale invariance for adversarial attacks. In: Proc. of the 8th Int'l Conf. on Learning Representations. Addis Ababa: OpenReview.net, 2020.

[42] Xie CH, Zhang ZS, Zhou YY, Bai S, Wang JY, Ren Z, Yuille AL. Improving transferability of adversarial examples with input diversity. In: Proc. of the 2019 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 2725–2734. [doi: 10.1109/CVPR.2019.00284]

[43] Wang GQ, Wei XX, Yan HQ. Improving adversarial transferability with spatial momentum. arXiv:2203.13479, 2022.

[44] Bhagoji AN, He W, Li B, Song D. Practical black-box attacks on deep neural networks using efficient query mechanisms. In: Proc. of the 15th European Conf. on Computer Vision. Munich: Springer, 2018. 158−174. [doi: 10.1007/978-3-030-01258-8_10]

[45] Chen PY, Zhang H, Sharma Y, Yi JF, Hsieh CJ. Zoo: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models. In: Proc. of the 10th ACM Workshop on Artificial Intelligence and Security. Dallas: ACM, 2017. 15−26. [doi: 10.1145/3128572.3140448]

[46] Tu CC, Ting PS, Chen PY, Liu SJ, Zhang H, Yi JF, Hsieh CJ, Cheng SM. AutoZOOM: Autoencoder-based zeroth order optimization method for attacking black-box neural networks. In: Proc. of the 33rd AAAI Conf. on Artificial Intelligence and the 31st Innovative Applications of Artificial Intelligence Conf. and the 9th AAAI Symp. on Educational Advances in Artificial Intelligence. Honolulu: AAAI Press, 2019. 92. [doi: 10.1609/aaai.v33i01.3301742]

[47] Ilyas A, Engstrom L, Athalye A, Lin J. Black-box adversarial attacks with limited queries and information. In: Proc. of the 35th Int'l Conf. on Machine Learning. Stockholm: PMLR, 2018. 2142−2151.

[48] Girshick R. Fast R-CNN. In: Proc. of the 2015 IEEE Int'l Conf. on Computer Vision. Santiago: IEEE, 2015. 1440–1448. [doi: 10.1109/ICCV.2015.169]

[49] He KM, Gkioxari G, Dollár P, Girshick R. Mask R-CNN. In: Proc. of the 2017 IEEE Int'l Conf. on Computer Vision. Venice: IEEE, 2017. 2980–2988. [doi: 10.1109/ICCV.2017.322]

[50] Carion N, Massa F, Synnaeve G, Usunier N, Kirillov A, Zagoruyko S. End-to-end object detection with transformers. In: Proc. of the 16th European Conf. on Computer Vision. Glasgow: Springer, 2020. 213–229. [doi: 10.1007/978-3-030-58452-8_13]

[51] Zhu XZ, Su WJ, Lu LW, Li B, Wang XG, Dai JF. Deformable DETR: Deformable transformers for end-to-end object detection. In: Proc. of the 9th Int'l Conf. on Learning Representations. OpenReview.net, 2021.

[52] Huang LF, Zhuang WZ, Liao YX, Liu N. Black-box adversarial attack method based on evolution strategy and attention mechanism. Ruan Jian Xue Bao/Journal of Software, 2021, 32(11): 3512–3529 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/6084.htm [doi: 10.13328/j.cnki.jos.006084]

[53] Woo S, Park J, Lee JY, Kweon IS. CBAM: Convolutional block attention module. In: Proc. of the 15th European Conf. on Computer

Vision (ECCV). Munich: Springer, 2018. 3–19. [doi: 10.1007/978-3-030-01234-2_1]

[54] Zhang HT, Zhou WG, Li HQ. Contextual adversarial attacks for object detection. In: Proc. of the 2020 IEEE Int'l Conf. on Multimedia and Expo (ICME). London: IEEE, 2020. 1–6. [doi: 10.1109/ICME46284.2020.9102805]

[55] Zhou BL, Khosla A, Lapedriza A, Oliva A, Torralba A. Learning deep features for discriminative localization. In: Proc. of the 2016 IEEE Conf. on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016. 2921–2929. [doi: 10.1109/CVPR.2016.319]

[56] Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: Visual explanations from deep networks via gradient-based localization. In: Proc. of the 2017 IEEE Int'l Conf. on Computer Vision. Venice: IEEE, 2017. 618–626. [doi: 10.1109/ICCV.2017.74]

[57] Chattopadhay A, Sarkar A, Howlader P, Balasubramanian VN. Grad-CAM++: Generalized gradient-based visual explanations for deep convolutional networks. In: Proc. of the 2018 IEEE Winter Conf. on Applications of Computer Vision. Lake Tahoe: IEEE, 2018. 839–847. [doi: 10.1109/WACV.2018.00097]

[58] Ultralytics. YOLOv5. 2020. https://github.com/ultralytics/yolov5

[59] Lin TY, Goyal P, Girshick R, He KM, Dollár P. Focal loss for dense object detection. In: Proc. of the 2017 IEEE Int'l Conf. on Computer Vision. Venice: IEEE, 2017. 2999–3007. [doi: 10.1109/ICCV.2017.324]

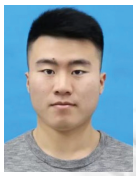
[60] Li X, Wang WH, Hu XL, Yang J. Selective kernel networks. In: Proc. of the 2019 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 510–519. [doi: 10.1109/CVPR.2019.00060]

[61] Kingma DP, Ba J. Adam: A method for stochastic optimization. In: Proc. of the 3rd Int'l Conf. on Learning Representations. San Diego, 2015.

[62] Benesty J, Chen JD, Huang YT, Cohen I. Pearson correlation coefficient. In: Cohen I, Huang YT, Chen JD, Benesty J, eds. Noise Reduction in Speech Processing. Berlin: Springer, 2009. 1–4. [doi: 10.1007/978-3-642-00296-0_5]

附中文参考文献:

[52] 黄立峰, 庄文梓, 廖泳贤, 刘宁. 一种基于进化策略和注意力机制的黑盒对抗攻击算法. 软件学报, 2021, 32(11): 3512–3529. http://www.jos.org.cn/1000-9825/6084.htm [doi: 10.13328/j.cnki.jos.006084]



陆宇轩(1997－), 男, 硕士, 主要研究领域为机器学习, 计算机视觉, 对抗攻击.

刘泽禹(2001－), 男, 本科生, 主要研究领域为机器学习, 计算机视觉.

罗咏刚(1992－), 男, 博士, 主要研究领域为人工智能.

邓森友(1990－), 男, 硕士, 主要研究领域为机器学习, 计算机视觉, 对抗攻击.

江天(1987－), 女, 硕士, 主要研究领域为计算机视觉, 算法安全性.

马金燕(1996－), 女, 硕士, 主要研究领域为计算机视觉, 目标检测.

董胤蓬(1995－), 男, 博士, 主要研究领域为机器学习, 计算机视觉, 对抗攻击.