

基于特征融合的对抗样本定向目标攻击可迁移性增强

凌海, 凌捷

(广东工业大学计算机学院, 广东 广州 510006)

摘要: 对抗样本可以在不知道黑盒模型内部的结构以及参数时利用代理模型进行迁移性攻击, 现有研究针对黑盒模型的定向目标攻击可迁移性都比较弱。提出一种基于特征融合增强图像定向目标攻击可迁移性的方法。通过模型集成攻击得到对抗样本, 以现有对抗样本的梯度方向为基准, 利用从原图提取出的干净特征作为干扰来微调现有的对抗样本, 以提高定向目标攻击的可迁移性。对于模型集成, 根据每个模型对整体对抗目标的贡献大小引入梯度自适应模块, 为减少不同模型之间的梯度差异, 提出梯度滤波器来同步控制梯度方向, 通过特征融合模块混合原图的干净特征对现有对抗样本的梯度方向进行微调以缓解过度关注特定特征的问题。在 ImageNet-Compatible 数据集上的对比实验结果表明, 所提方法对非鲁棒性训练模型, 相较 CFM(Clean Feature Mixup)方法平均攻击成功率提升了 7.7 个百分点, 对鲁棒性训练模型以及 Transformer 模型, 相较 CFM 方法平均攻击成功率提升了 5.3 个百分点, 验证了方法的有效性。

关键词: 深度学习; 对抗攻击; 对抗样本; 定向目标攻击; 可迁移性

中图分类号: TP391

文献标志码: A

DOI: 10.19678/j.issn.1000-3428.0069983

Transferability Enhancement of Adversarial Sample Directed Targeted Attack Based on Feature Fusion

LING Hai, LING Jie

(School of Computer, Guangdong University of Technology, Guangzhou 510006, Guangdong, China)

【Abstract】 Adversarial examples can be used to perform transferable attacks on black-box models using surrogates, without knowing the internal structure and parameters of the black-box model. Previous studies have reported relatively low transferability of targeted attacks on black-box models. This study proposes a method for enhancing the transferability of image-directed targeted attacks based on feature fusion. First, adversarial examples are generated via ensemble attacks. Subsequently, using the gradient direction of existing adversarial examples as a baseline, clean features extracted from the original image are used as perturbations to fine tune the existing adversarial examples for improving the transferability of targeted attacks. For model ensembling, a gradient adaptive module is introduced based on the contribution of each model to the overall adversarial objective. To reduce the gradient differences among different models, a gradient filter is proposed for synchronously controlling the gradient direction. Using the feature fusion module, the clean features of the original image are mixed to fine tune the gradient direction of the existing adversarial examples for mitigating the issue of overfocusing on specific features. Experiments on the ImageNet-Compatible dataset reveal that, compared to the Clean Feature Mixup (CFM) method, the proposed method improves the average attack success rate by 7.7 percentage points for non-robustly trained models and by 5.3 percentage points for robustly trained and Transformer models, demonstrating the effectiveness of the method.

【Key words】 deep learning; adversarial attacks; adversarial examples; directed targeted attacks; transferability

0 引言

深度神经网络在各种计算机视觉任务中表现出色, 已经达到甚至超过人眼的能力水平^[1], 如图像分类和对象检测, 但它们很容易受到对抗样本的恶意输入的影响。这些对抗样本通过优化不可察觉的扰动来误导模型, 使其产生不正确的预测。由于一些

精心设计的对抗样本可以跨模型传递, 因此由代理模型制作的对抗样本也可能干扰其他模型。对抗样本的这种特性被称为可迁移性, 即允许对手在不清楚目标模型内部的情况下攻击目标模型。由于模型决策边界的显著差异, 具有特定目标类别的定向对抗攻击的攻击成功率仍然较低。但是定向对抗攻击会欺骗模型输出特定的有害目标类别, 会带来严重

收稿日期: 2024-06-11 修回日期: 2024-07-17

基金项目: 广州市重点领域研发计划(202007010004)。

通信作者 E-mail: linghai12138@163.com

的风险。因此,研究定向对抗样本的生成以及如何提高其可迁移性非常重要,它可以帮助服务提供商为即将到来的对抗攻击准备防御模型,并评估模型的稳健性。

在代理模型和目标模型之间,代理模型与目标模型越相似,可以获得越高的迁移成功率。尽管利用代理模型生成对抗样本是有效的,但在具有巨大架构差异的模型之间迁移攻击时成功率比较低。一系列研究建议利用代理模型的集成来生成能够成功攻击所有代理模型的对抗样本,这是因为它可以捕获图像内在的可迁移的对抗性信息,做到同时欺骗几个差异很大的模型。

在图像分类中,深度学习模型从多层图像中提取各种特征,并对其进行综合评估,以计算每个类别的预测概率。由于许多特征可以对最终输出做出贡献,即使当两个图像被识别为同一类时,贡献的特征也可能显著不同。考虑到这一点,优化对抗样本以利用尽可能多的不同特征组合将有效地增强其可迁移性。

对抗样本通常会过度拟合源模型的特定架构和特征表示,从而导致对其他模型的定向目标攻击的可迁移性较低。本文的研究目标旨在提高定向对抗样本的可迁移性,因此提出了一种新的攻击方法。首先通过模型集成攻击得到对抗样本,接着利用从原图提取出的特征作为干扰来微调现有的对抗样本以获得更强的定向目标攻击可迁移性。本文认为模型集成攻击找到的梯度方向不是一个最优的方向,但是它的确能够缓解对源模型特定架构的过拟合问题,并且比随机的梯度更新方向收敛的可能性更大。因此本文方法在现有模型集成攻击的梯度方向上进行微调,从而更加快速地找到最优的梯度更新方向。这一定程度上借鉴了生成式对抗网络的思想,将模型集成后的对抗样本特征和原图特征作为特征融合的输入,但实现方式不同,两个训练阶段相互独立,训练参数更少。而把原图特征作为干扰进行微调,可以避免偏离原图特征太远从而导致扰动过大,也能缓解对源模型特征表示的过度拟合,增强了对抗样本向不同模型的可迁移性。

本文的贡献总结如下:

1)提出特征融合模块,通过组合不同特征以防止对抗样本在攻击过程中只关注特定特征,缓解了优化过程中对源模型的过拟合问题。

2)提出梯度自适应模块来解决传统集成方法优化不足的问题,以控制优化过程中不同模型的权重。同时提出梯度滤波器来解决梯度优化上的方向同步问题,减小不同模型的梯度差异。

3)在 ImageNet-Compatible 数据集上以 16 个模型为黑盒目标进行了大量实验,与 CFM(Clean Feature Mixup)方法相比,本文方法对非鲁棒性训练模型平均攻击成功率提升了 7.7 百分点,对鲁棒性训练模型以及 Transformer 模型平均攻击成功率提升了 5.3 百分点,实验结果验证了本文方法的有效性。

1 相关工作

自从 SZEGEDY 等^[2]证明存在对抗样本以来,人们一直致力于强调深度神经网络的脆弱性。对抗攻击通常通过向原始输入图像 x 添加扰动来产生对抗样本,其目的是使模型的损失函数 $\mathcal{L}$ 最大化,即 $\arg\max_{x+\delta} \mathcal{L}(f(x+\delta), y)$ 。为了使扰动不可察觉,扰动 δ 受到 S 的约束, $S = \{\|\delta\|_p \leq \epsilon\}$, 其中, L_p 范数小于等于误差 ϵ 。

1.1 基于梯度的对抗攻击

为了优化攻击目标,通常使用梯度信息来最大化模型损失。GOODFELLOW 等^[3]在研究神经网络线性性质的基础上,设计了一种快速梯度符号法(FGSM)来生成对抗样本。刘梦庭等^[4]将优化梯度与 Adam 优化算法结合进行迭代优化。郑德生等^[5]通过对输入图片的通道进行拆分调整原始梯度的更新方向,避免出现局部最优。WANG 等^[6]在迭代过程中累加平均梯度,以稳定更新方向并避免较差的局部极大值。李哲铭等^[7]引入 Nadam 优化算法,稳定损失函数、更新梯度方向并动态调整学习率,提高损失函数收敛速度。虽然这些攻击可以在白盒模型上表现出较高的攻击成功率,但由于黑盒模型的梯度信息难以估计,这样的对抗样本对黑盒模型的攻击成功率通常较低。

1.2 基于变换的对抗攻击

为了提高可迁移性,现有的工作试图最大限度地对输入进行变换来提高对抗样本的攻击成功率。LIN 等^[8]利用图像缩放不变性这一特质,对输入图像进行不同尺度的缩放来计算平均梯度。DONG 等^[9]将 FGSM 与输入多样性和平移不变策略相结合,以产生用于生成对抗样本的不同输入模式。ZHANG 等^[10]提出了利用生成模型来学习对抗函数的方法,以破坏输入图像的低级特征。赫晓慧等^[11]利用权重分布的方法匹配感受野尺度,使用不同层级间的特征对全局先验信息进行表征,提高多尺度融合的准确性。尽管这些攻击相较于基于梯度的攻击可以实现可迁移性的改进,但它们很难迁移到比较复杂的深度神经网络架构。

1.3 基于模型集成的对抗攻击

模型集成攻击方法通常通过并行执行多个白盒攻击的加权线性求和来制作对抗样本。LIU 等^[12]直接对多个模型的预测进行平均,以获得应用基于梯度攻击的综合损失。DONG 等^[9]进一步融合了所有模型的 logits 损失。XIONG 等^[13]注意到集成模型之间的差异,提出了一种随机减少方差的集成攻击来提高攻击的泛化能力。这些研究虽然实现了一定程度的改进,但由于对每个模型个体优势的关注较少,因此总体仍未达到最佳状态。

1.4 对抗防御

对抗攻击的代表性防御方法之一是对抗训练^[14],它直接利用训练模型中的对抗样本。另一种有效的防御策略是构建一个单独网络的集合,对手应欺骗集合中的多个模型,而不是单个模型,从而提高鲁棒性。为此,通常使用交叉熵损失和一些正则化项来训练集合模型,以考虑单个网络之间的相互作用。PANG 等^[15]引入了自适应多样性促进正则化矩阵,该正则化矩阵增强了单个成员的非最大预测值之间的多样性,使对抗样本更难在它们之间迁移。赖妍菱等^[16]基于图像语义分割网络 U-Net 架构,通过压缩和重建方式进行特征提取和图像还原,

破坏对抗样本中的扰动结构,实施有效防御。

2 本文方法

2.1 方法概述

本文方法首先通过模型集成攻击得到对抗样本,接着利用从原图提取出的特征作为干扰来微调现有的对抗样本以获得更强的定向目标攻击可迁移性,框架示意图如图 1 所示(彩色效果见《计算机工程》官网 HTML 版,下同)。对于模型集成,引入梯度自适应模块(GA)对每个模型的输出进行自适应集成。但集成梯度可能与代理模型的个体梯度有很大不同,甚至相反,这已被证明与模型集成中的过拟合问题有关^[13]。因此,本文进一步提出梯度滤波器(GF)来减小各模型之间的梯度差异并同步更新方向,以达到增强对抗可迁移性的效果。最后利用模型集成生成的现有对抗样本作为输入与原始图像进行特征融合,融合方式是通过本文专门设计的特征融合模块(FM)来转换预先训练的源模型。FM 模块在每次推理时将原始图像的特征(即干净特征)与当前输入特征(即现有对抗样本的特征)随机混合。该过程可以微调对抗样本,防止其在定向目标攻击中只关注特定特征,有效地缓解对抗样本优化中的过拟合。下面将详细介绍本文方法。

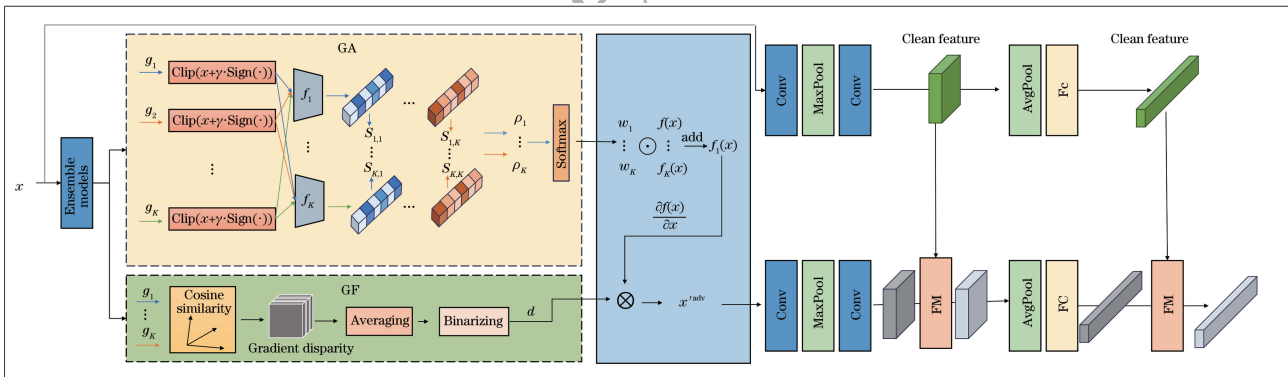


图 1 本文方法框架

Fig.1 Framework of the proposed method

2.2 梯度自适应方法

本文中采用的模型集成方式与传统方法有所不同,简单地对集成模型的输出进行平均忽略了每个模型的优势,因为从一个模型捕获的可迁移信息可以在融合过程中被另一个模型平滑,从而导致优化不足的结果。为了解决这个问题,本文通过梯度自适应对每个模型的输出进行自适应集成。如图 1 所示,输入图像经过各个代理模型 f_i 的前向传播之后得到输出 $f_i(x)$ 和梯度信息 g_i , 即 $g_i = \nabla_{x_{adv}} \mathcal{L}((x_{adv}), y)$ 。通过评估它们对对抗攻击目标

的贡献的大小来自适应地调整模型集合。具体来说,对于第 i 个集成模型 f_i , 通过计算由 g_i 生成的对抗样本 x_{adv} 对其他模型的攻击性能来评估 g_i 中潜在的对抗可迁移性,将其定义为对抗性比率,然后基于每个模型的对抗性比率来调整集合权重。在这里,本文首先通过以下公式计算损失值 $s_{k,i}$:

$$s_{k,i} = -1_y \cdot \ln(\text{Softmax}(P_k[x_{adv} + \alpha \text{sign}(g_i)])) \quad (1)$$

式中: α 表示学习率; x_{adv} 表示周期为 t 时生成的对

抗样本; $P_k[\cdot]$ 表示模型 f_k 的 logits 输出; 1_y 表示是真实的 logits; $s_{k,i}$ 可以看作是基于一模型 i 的梯度生成的对抗样本在第 k 个模型上的损失值。基于上述的 $s_{k,i}$, 本文把对抗性比率 ρ_i 定义为:

$$\rho_i = \frac{\beta}{K-1} \sum_{k=1, k \neq i}^K \frac{s_{k,i}}{s_{k,k}} \quad (2)$$

式中: K 是模型的集成数量; β 是控制模型集成加权的一个超参数, 将在消融实验部分(3.3 节)进一步讨论。注意, ρ_i 的值越高, 表示由 ρ_i 生成的对抗样本的可迁移性越好, 这意味着 ρ_i 包含更多可迁移的对抗性信息。通过上述计算找出哪个模型可以提供更通用的对抗性信息, 并自适应地分配更高的权重。因此, 根据每个模型的对抗性比率 ρ_i , 本文通过以下方式归一化每个模型的权重 w_i :

$$w_1, w_2, \dots, w_K = \text{Softmax}(\rho_1, \rho_2, \dots, \rho_K) \quad (3)$$

利用所获得的 w_i , 即第 i 个代理模型 f_i 的集合权重, 使得具有更强可迁移性的代理模型的输出在集合梯度中被放大, 从而促使对抗样本在黑盒模型上保持更高的迁移攻击成功率。

2.3 梯度滤波器设计

由于代理模型的梯度优化方向在很大范围内变化, 有时梯度会朝着彼此相反的方向移动, 导致对集成模型的过度拟合^[13]。为了解决这个问题并同步代理模型之间的梯度更新方向, 本文提出了梯度滤波器来减少代理模型之间的梯度方差。首先应用余弦相似性来评估代理模型中梯度的偏差, 并通过相似性得分求平均值来计算视差图 d_i , 可以描述如下:

$$d_i = \frac{1}{K-1} \sum_{k=1, k \neq i}^K \cos(g_i, g_k) \quad (4)$$

式中: $\cos(\cdot)$ 表示余弦相似性得分; g_i 和 g_k 表示从梯度中的同一位置 (p, q) 提取出来的梯度向量。最终通过对所有 d_i 求平均值来获得用于集成梯度的视差图 d 。然后通过使用滤波器 F 来消除集成梯度中的视差部分, 如下所示:

$$F(p, q) = \begin{cases} 0, & d_i \leq \eta \\ 1, & d_i > \eta \end{cases} \quad (5)$$

式中: η 是滤波器的阈值, 将在消融实验部分(第 3.3 节)进一步讨论。通过滤除集成梯度中的视差部分, 可以同步梯度优化方向。因此, 集成梯度可以用下面的公式来表示:

$$g_{t+1} = \nabla_{x_{adv}} \mathcal{L} \left(\sum_k w_k f_k(x_{adv}), y \right) \otimes F \quad (6)$$

式中: x_{adv} 表示按元素做乘积; x_{adv} 表示第 t 个周期生成的对抗样本; K 是代理模型的个数; k 是第 k 个代理模型 y 的集合权重; y 表示目标类别。利用梯度

滤波器 F 可以抑制代理模型之间的梯度差异。

2.4 特征融合

特征空间比图像空间要宽得多, 并且每个模型可以有不同体系结构, 这使得特征融合的实现具有挑战性。由于图像和特征之间存在明显的域间隙, 直接将传统的图像变换应用于特征映射可能会过度阻碍优化。因此, 本文通过将干净特征与输入特征随机混合来转换特征。这里把原图中提取的特征称之为干净特征, 把经过自适应集成之后的特征作为输入特征去随机混合。要实现这种特征融合, 需要先提取出原图的干净特征以便在模型推理时转换特征。因为深度学习神经网络结构并不统一, 所以本文设计了一个特征融合模块, 并将特征融合模块附加到卷积层与全连接层之间。原图的干净特征在模型进行第一次推理时被提取到内存中保存, 如图 1 所示。深度神经网络通常有几十到几百层, 在所有这些层中同时混合干净特征会过度破坏推理过程, 因此特征融合模块以概率 p 随机地应用混合特征, 如图 2 所示, 这样既保证了特征的随机性, 又减少了超参数调优所需的工作量。

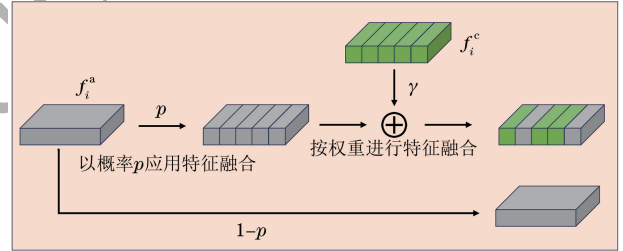


图 2 特征融合模块示意图

Fig.2 Schematic diagram of feature fusion module

本文提出的特征融合模块基于上述思路把干净特征与经过自适应集成后的特征进行随机混合, 从而找到具备可迁移性的特征进行进一步优化, 可以表示为以下公式:

$$f_{m,i} = (1 - \gamma_i) \odot f_{a,i} + \gamma_i \odot f_{c,i}, \quad i = 1, 2, \dots, B \quad (7)$$

式中: B 是 Batch size 的大小; i 表示第 i 批次图像; $f_{a,i}$ 是经过自适应集成后的对抗样本的特征; $f_{c,i}$ 是原图的干净特征。 $u(0, \gamma_{\max})$ 表示两种特征的混合权重, $u(0, \gamma_{\max})$ 的取值服从均匀分布 $u(0, \gamma_{\max})$, 其中, $0 \leq \gamma_{\max} \leq 1$, $\{\gamma_1, \gamma_2, \dots, \gamma_B\}$ 都是从每一次的推理过程中随机采样得来, 从而达到随机混合的效果。

首先, 每个特征融合模块通过线性插值混合干净特征 $f_{c,i}$ 和现有对抗样本的特征 $f_{a,i}$, 并且为了增加随机性, 每个通道的混合比例 $f_{c,i}$ 是随机采样的, 实现了更多样化的干扰。其次, 混合的干净特征

$f_{c,i}$ 如果对图像真实类别有积极作用,那么 $f_{c,i}$ 对于指定类别的定向目标攻击是起到抑制作用的,这将会促使对抗样本在优化过程中寻找对指定类别有利的特征,从而提高对抗样本的可迁移性。混合的干净特征 $f_{c,i}$ 如果对除指定类别外的其他类别有积极作用,那么 $f_{c,i}$ 对于指定类别的定向目标攻击也是起到抑制作用的,这将会促使对抗样本稳健地探索对抗性的特征扰动,从而在攻击中取得成功。总之,融合作为干扰的干净特征可以有效地缓解对抗样本的过拟合,防止它们在对模型进行定向目标攻击时专注于特定特征。

3 实验及结果分析

3.1 实验设置

1)数据集。参照之前的工作^[17-18],本文使用了常用的 ImageNet-Compatible 数据集,该数据集是为 NIPS 2017 对抗攻击挑战赛发布的。它包含了 1 000 张 299×299 像素的图像,以及它们的真实类别和目标类别,用于定向目标的对抗攻击。

2)参数设置。本文的大多数实验环境都遵循了最近的研究^[17-18]。具体来说,实验使用了 L_∞ 范式来约束扰动,其中, $\epsilon = 16/255$,并参照之前的工作迭代攻击设置步长 $\alpha = 2/255$ 。所有方法都基于 logit 损失来优化对抗样本,包括本文方法和基线方法。为了给出足够的迭代周期来优化对抗样本,这里将总迭代次数 T 设置为 $T = 300$,这与之前的工作^[17-18]中设置一样。

3)源模型和目标模型。本文采用了 10 个预先训练的神经网络作为目标网络: VGG-16, ResNet-18 (RN-18), ResNet-50 (RN-50), DenseNet-121 (DN-121), Xception (Xcep), MobileNet-v2 (MBv2), EfficientNet-B0 (EF-B0), Inception ResNetv2 (IR-v2), Inception-v3 (Inc-v3), Inceptionv4 (Inc-v4)。此外还包括了一个对抗训练的 RN-50 网络 (adv-RN-50),该网络使用 L_2 范数约束的对抗样本, $\|\delta\|_2 \leq 0.1$,因为用它作为源模型时,可以有效地提高对抗攻击迁移成功率^[19]。实验还添加了 5 个基于 Transformer 的分类器: ViT (Vision Transformer), LeViT, ConViT, Twins, PiT (Pooling-based Vision Transformer)。

4)基线方法。本文使用 8 种现有技术的不同组合组成了基线方法: RDI^[20], MI^[21], TI^[9], SI^[8], VT^[22], AdMix^[23], ODI^[18], CFM^[24]。实验将 MI 和 TI 技术应用于所有攻击方法,所以本文在表示它们时省略了 MI-TI。其中,本文将 CFM 看作主要基

线方法。对于 RDI、TI 和 ODI 的详细设置遵循 CFM。具体而言,对于 RDI,图像大小的缩放因子为 $1 \sim 340/299$ 。将 TI 的卷积核大小设置为 5×5 ,将 MI 的衰减因子 μ 设置为 $\mu = 1.0$ 。对于 VT 和 SI,将样本数和缩放比例设置为 5,将 VT 的 σ 设置为 $\sigma = 1.5$ 。对于 AdMix,将混合权重 w 设置为 $w = 0.2$,将混合的图像数量 N 设置为 $N = 3$ (即 AdMix 中 $m_2 = 3$)。原始的 AdMix 设置在其内部循环中使用具有 5 个缩放副本的 SI。然而,在 AdMix 的内部循环中使用 SI 很难直接比较 AdMix 中的图像级混合和 CFM 中的特征级混合的性能改进。因此,实验将 AdMix 中 SI 的缩放副本数设置为 1 (即 AdMix 中 $m_1 = 1$)。对于 AdMix 和 CFM 都使用了 20 个批次大小进行公平比较。对于 CFM,将通道混合比 γ 设置为从 $u(0, 0.75)$ 中随机采样。实验将 ImageNet 数据集的混合概率 p 设置为 $p = 0.1$ 。此外,由于 CFM 消耗一个推理来存储干净的特征,因此将可用的剩余迭代次数减少到 299 次,以进行严格公平的比较。

5)本文方法的设置。实验使用了在 L_∞ 约束下迭代 20 次的 I-FGSM 作为先验攻击方向,为了公平比较,将可用的剩余迭代次数减少到 279 次,并在对抗样本生成过程中设置 $\epsilon = 16/255$ 和 $\alpha = 2/255$ 。对于超参数,在 FM 中设置通道混合比 γ 为从 $u(0, 0.75)$ 中随机采样,同 CFM 一样。在 GF 中设置 $\eta = -0.3$,在 GM 中设置 $\beta = 10$ 。所有实验都是在 NVIDIA A100 GPU 处理器上使用 PyTorch 实现的, GPU 内存为 32 GB。

3.2 结果分析

本文使用 ImageNet-Compatible 数据集进行了定向对抗攻击实验。实验使用预训练的 RN-50 和 Inc-v3 作为源模型,分别对 10 个非鲁棒性目标模型以及 1 个对抗训练模型和 5 个 Transformer 模型的定向目标攻击成功率进行了评估。

表 1 和表 2 显示了将 RN-50 和 Inc-v3 作为源模型分别生成的对抗样本对 10 个非鲁棒性模型的定向对抗攻击的成功率 (加粗数据表示最优值,下同)。可以看出,本文方法在所有实验中得到了最高的平均迁移攻击成功率 (77.6% 和 43.7%),在所有源模型中都以明显的优势优于其他基线方法。特别是以 Inc-v3 为源模型时,所提方法获得了较好的性能提升,相较于 CFM 提高了 11.5 个百分点。基于 RN-50 和 Inc-v3 两个源模型进行对抗攻击得到的平均攻击成功率相较 CFM 方法提高了 7.7 百分点,明显地表现了这种基于特征融合微调梯度的方法的先进性。

表 1 在 10 个非鲁棒性模型上进行定向目标攻击的攻击成功率(RN-50 为源模型)

Table 1 Attack success rate of targeted attacks on 10 non-robust models (RN-50 is source model) %

Method	Target model										Avg
	VGG-16	RN-18	RN-50	DN-121	Xcep	MB-v2	EF-B0	IR-v2	Inc-v3	Inc-v4	
RDI	65.4	71.8	98.0	81.3	13.1	46.6	46.6	16.8	30.7	23.9	49.4
SI-RDI	70.5	79.8	98.8	88.9	29.5	56.2	66.2	37.9	56.4	43.6	62.8
VT-RDI	68.8	78.7	98.2	82.5	27.9	54.5	56.1	32.8	45.8	37.9	58.3
Admix-RDI	74.2	80.7	98.7	86.8	20.9	59.4	56.1	26.7	42.7	34.1	58.0
ODI	78.3	77.1	97.6	87.0	43.8	67.3	70.0	49.5	65.9	55.4	69.2
CFM-RDI	84.7	88.4	98.4	90.3	51.1	81.5	78.8	48.0	65.5	59.3	74.6
Ours	87.1	88.8	98.3	91.5	56.3	83.6	81.9	53.2	70.1	64.8	77.6

表 2 在 10 个非鲁棒性模型上进行定向目标攻击的攻击成功率(Inc-v3 为源模型)

Table 2 Attack success rate of targeted attacks on 10 non-robust models (Inc-v3 is source model) %

Method	Target model										Avg
	VGG-16	RN-18	RN-50	DN-121	Xcep	MB-v2	EF-B0	IR-v2	Inc-v3	Inc-v4	
RDI	3.5	3.8	4.0	7.0	3.1	3.0	5.9	6.3	98.7	7.1	14.2
SI-RDI	4.0	5.2	5.7	11.0	6.3	4.6	8.2	11.6	98.8	12.1	16.8
VT-RDI	5.9	8.9	9.4	13.2	7.4	5.9	9.8	12.3	98.7	14.7	18.6
Admix-RDI	6.3	6.5	8.8	12.8	6.0	6.1	10.9	12.2	98.7	13.6	18.2
ODI	14.3	14.9	16.7	32.3	20.3	13.7	25.3	26.4	95.6	31.6	29.1
CFM-RDI	22.9	26.8	26.2	39.1	34.1	27.1	38.6	36.2	95.9	44.8	39.2
Ours	27.0	30.3	32.5	44.8	38.6	28.6	43.2	44.2	97.5	50.0	43.7

表 3 和表 4 显示了基于 2 个源模型(RN-50 和 adv-RN-50)对 1 个对抗训练模型和 5 个基于 Transformer 的模型的定向目标攻击成功率。可以看出,本文方法超过了所有的基线方法。特别是由 RN-50 生成的对抗样本达到了 34.0%的平均定向攻击成功率,相较于 CFM 提高了 10.4 百分点。基于 RN-50 和 adv-RN-50 生成的对抗样本在对抗训练模型以及 Transformer 模型上表现良好,

表 3 在鲁棒性模型以及 Transformer 模型上进行定向目标攻击的攻击成功率(RN-50 为源模型)

Table 3 Attack success rate of directed targeted attacks on robustness models and Transformer models (RN-50 is source model) %

Method	Target model						Avg
	adv-RN-50	ViT	LeViT	ConViT	Twins	PiT	
RDI	34.8	0.7	13.1	1.9	5.9	6.8	10.5
SI-RDI	59.9	2.9	29.4	6.3	15.5	17.9	22.0
VT-RDI	64.2	2.9	28.1	5.2	15.0	14.0	21.6
Admix-RDI	52.4	1.3	22.5	2.5	8.5	8.4	15.9
ODI	64.7	5.1	37.0	10.7	20.1	29.1	27.8
CFM-RDI	75.5	4.3	46.1	8.9	25.2	24.7	30.8
Ours	78.4	5.2	49.6	11.7	29.7	29.2	34.0

表 4 在鲁棒性模型以及 Transformer 模型上进行定向目标攻击的攻击成功率(adv-RN-50 为源模型)

Table 4 Attack success rate of directed targeted attacks on robustness models and Transformer models (adv-RN-50 is source model) %

Method	Target model						Avg
	adv-RN-50	ViT	LeViT	ConViT	Twins	PiT	
RDI	98.8	10.8	49.5	19.9	29.4	35.8	40.7
SI-RDI	98.7	19.4	57.6	35.3	35.2	52.1	49.7
VT-RDI	98.5	10.6	46.3	20.0	27.1	34.4	39.5
Admix-RDI	98.9	12.1	55.5	23.1	32.4	38.9	43.5
ODI	97.3	22.2	57.7	38.8	40.0	54.9	51.8
CFM-RDI	98.3	29.5	69.8	41.8	52.7	59.8	58.6
Ours	98.6	29.6	68.7	43.1	52.4	60.1	58.7

平均攻击成功率超出了 CFM 方法 5.3 百分点,表明了本文提出的可迁移性攻击具有较好的鲁棒性。

图 3 显示了以 RN-50 为源模型、MB-v2 为目标

模型以及以 Inc-v3 为源模型、Xcep 为目标模型两种情况下的攻击成功率,可以看出本文方法收敛时间与作为对比的 CFM 近似,但攻击成功率明显优于其他基线方法。

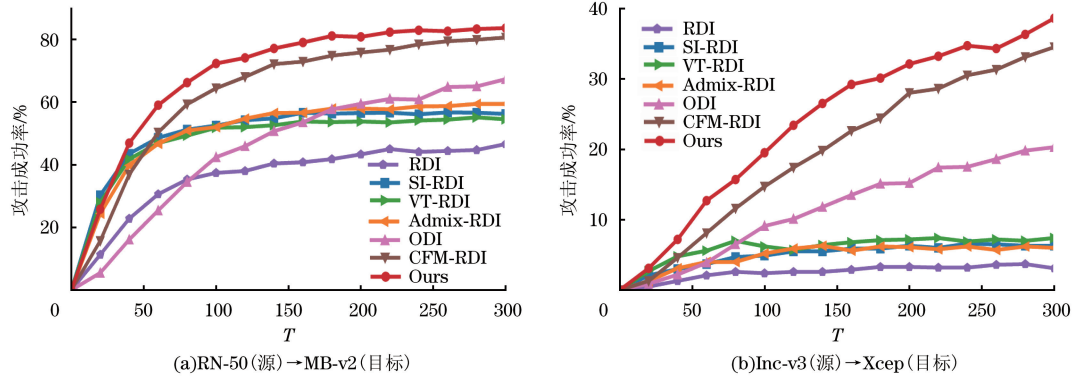


图 3 基于优化的定向目标攻击成功率

Fig.3 Directed targeted attack success rate based on optimization

3.3 消融实验

通过消融实验研究了梯度自适应权重因子 β 、梯度滤波器二叉阈值 η 的最佳取值,以及本文方法内部各个模块之间的影响。在这些消融实验中以 adv-RN-50 为源模型,选择了 10 个较难受到干扰的目标模型。

通过改变梯度自适应模块中 β 的值与梯度滤波

器中 η 的值来评估迁移对抗攻击成功率的变化。如图 4(a)所示,随着权重因子 β 值增大,对抗样本的可迁移性在逐渐降低,这表明自适应集成后的集成梯度作用在减弱,当 $\beta=10$ 时达到峰值。如图 4(b)所示,随着二叉阈值 η 的增大,集成梯度的方差在减少,对抗样本的可迁移性有一定的提升,但 η 持续增大,会导致迁移性下降,当 $\eta=-0.3$ 时达到峰值。

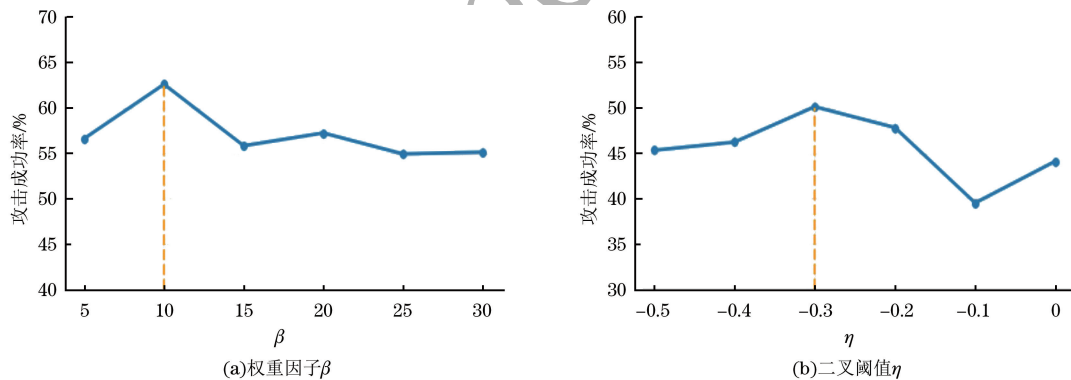


图 4 针对梯度自适应权重因子 β 以及梯度滤波器二叉阈值 η 的消融实验

Fig.4 Ablation experiments on gradient adaptive weight factor β and gradient filter binary threshold η

通过分解本文方法的 3 个内部模块来评估攻击成功率的变化。本次消融实验结果如表 5 所示,可

以看出每个内部模块都有助于提高对抗样本的可迁移性。仅仅使用梯度自适应与梯度滤波器进行模型

表 5 消融实验的攻击成功率

Table 5 The attack success rate in ablation experiments

Ablation			Target model										Avg
FM	GA	GF	Xcep	MB-v2	EF-B0	IR-v2	Inc-v4	ViT	LeViT	ConViT	Twins	PiT	
✓	✓	✓	67.6	83.7	84.0	64.5	68.9	29.6	68.7	43.6	52.4	62.1	62.5
	✓	✓	41.0	67.3	70.0	46.0	47.9	10.1	49.3	20.3	30.4	36.4	41.9
✓			62.0	83.8	80.2	60.3	68.7	25.1	71.1	42.9	52.9	60.7	60.8
✓	✓		66.9	82.9	83.1	65.2	67.7	26.2	71.5	43.4	53.6	61.5	62.2
✓		✓	66.8	82.2	83.8	63.7	67.7	27.4	69.8	42.6	52.6	60.8	61.7

集成而不使用特征融合意味着没有利用原图的干净特征对现有的对抗样本进行微调,实验的平均攻击成功率大幅降低。而如果只是使用特征融合而不应用模型集成,意味着混合的特征只是原图的特征,是图像本身的干净特征,特征表示过于单一,因此平均攻击成功率也有所下降。如果在特征融合中应用梯度自适应或者梯度滤波器其中的一个模块,可以看到平均攻击成功率均有所提升。这表明 3 个模块对提高对抗样本定向目标攻击的可迁移性都具有重要作用。

4 结束语

针对定向目标攻击可迁移性较弱问题,本文提出一种基于特征融合增强图像的定向目标攻击可迁移性的对抗攻击方法,首先通过模型集成攻击得到对抗样本,接着以现有对抗样本的梯度方向为基准,利用从原图提取出的干净特征作为干扰来微调现有的对抗样本以增强定向目标攻击的可迁移性。实验结果表明,本文方法能够生成可迁移性更强的对抗样本并且能够较好地与现有攻击方法兼容。在非鲁棒性模型上与 CFM 相比,所提方法的定向目标攻击成功率提升较明显,最低达到 4.0 百分点,最高达到 11.5 百分点。在鲁棒性模型以及 Transformer 模型上,所提方法相较 CFM 定向目标攻击成功率最低提高了 0.2 百分点,最高提高了 10.4 百分点。下一步将改进所提方法的时间复杂度,在保持较高攻击成功率的前提下加快对抗样本的生成速度。

参考文献

- [1] SADAK F, SAADAT M, HAJIYAVAND A M. Real-time deep learning-based image recognition for applications in automated positioning and injection of biological cells[J]. *Computers in Biology and Medicine*, 2020, 125: 103976.
- [2] SZEGEDY C, ZAREMBA W, SUTSKEVER I, et al. Intriguing properties of neural networks[EB/OL]. [2024-05-02]. <https://arxiv.org/abs/1312.6199v4>.
- [3] GOODFELLOW I J, SHLENS J, SZEGEDY C, et al. Explaining and harnessing adversarial examples[EB/OL]. [2024-05-02]. <https://arxiv.org/abs/1412.6572v3>.
- [4] 刘梦庭, 凌捷. 优化梯度增强黑盒对抗攻击算法[J]. *计算机工程与应用*, 2023, 59(18): 260-267.
LIU M T, LING J. Optimized gradient boosting black-box adversarial attack algorithm[J]. *Computer Engineering and Applications*, 2023, 59(18): 260-267. (in Chinese)
- [5] 郑德生, 陈继鑫, 周静, 等. 基于输入通道拆分的对抗攻击迁移性增强算法[J]. *计算机工程*, 2023, 49(1): 130-137.
ZHENG D S, CHEN J X, ZHOU J, et al. Adversarial attack transferability enhancement algorithm based on input channel splitting[J]. *Computer Engineering*, 2023, 49(1): 130-137. (in Chinese)
- [6] WANG X S, LIN J D, HU H, et al. Boosting adversarial transferability through enhanced momentum[EB/OL]. [2024-05-02]. <https://arxiv.org/abs/2103.10609v1>.
- [7] 李哲铭, 王晋东, 侯建中, 等. 基于显著区域优化的对抗样本攻击方法[J]. *计算机工程*, 2023, 49(9): 246-255, 264.
LI Z M, WANG J D, HOU J Z, et al. Adversarial example attack method based on salient region optimization[J]. *Computer Engineering*, 2023, 49(9): 246-255, 264. (in Chinese)
- [8] LIN J D, SONG C B, HE K, et al. Nesterov accelerated gradient and scale invariance for adversarial attacks[EB/OL]. [2024-05-02]. <https://arxiv.org/abs/1908.06281v5>.
- [9] DONG Y P, PANG T Y, SU H, et al. Evading defenses to transferable adversarial examples by translation-invariant attacks[C]//*Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Long Beach, USA: IEEE Press, 2019: 4307-4316.
- [10] ZHANG Q L, LI X D, CHEN Y F, et al. Beyond ImageNet attack: towards crafting adversarial examples for black-box domains[EB/OL]. [2024-05-02]. <https://arxiv.org/abs/2201.11528v4>.
- [11] 赫晓慧, 宋定君, 李盼乐, 等. 融合多尺度特征的遥感影像道路提取方法[J]. *计算机工程*, 2022, 48(8): 196-205.
HE X H, SONG D J, LI P L, et al. Remote sensing image road extraction method combined with multiscale features[J]. *Computer Engineering*, 2022, 48(8): 196-205. (in Chinese)
- [12] LIU Y, CHEN X, LIU C, et al. Delving into transferable adversarial examples and black-box attacks[EB/OL]. [2024-05-02]. <https://arxiv.org/abs/1611.02770>.
- [13] XIONG Y F, LIN J D, ZHANG M, et al. Stochastic variance reduced ensemble adversarial attack for boosting the adversarial transferability[C]//*Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. New Orleans, USA: IEEE Press, 2022: 14963-14972.
- [14] MADRY A, MAKELOV A, SCHMIDT L, et al. Towards deep learning models resistant to adversarial attacks[EB/OL]. [2024-05-02]. <https://arxiv.org/abs/1706.06083v4>.
- [15] PANG T, XU K, DU C, et al. Improving adversarial robustness via promoting ensemble diversity[C]//*Proceedings of International Conference on Machine Learning*. [S. l.]: PMLR, 2019: 4970-4979.
- [16] 赖妍菱, 石峻峰, 陈继鑫, 等. 基于 U-Net 的对抗样本防御模型[J]. *计算机工程*, 2021, 47(12): 163-170.
LAI Y L, SHI J F, CHEN J X, et al. Adversarial example defense model based on U-Net[J]. *Computer Engineering*, 2021, 47(12): 163-170. (in Chinese)
- [17] ZHAO Z, LIU Z, LARSON M. On success and simplicity: a second look at transferable targeted attacks[J]. *Advances in Neural Information Processing Systems*, 2021, 34: 6115-6128.
- [18] BYUN J, CHO S, KWON M J, et al. Improving the transferability of targeted adversarial examples through object-based diverse input[C]//*Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. New Orleans, USA: IEEE Press, 2022: 15223-15232.
- [19] SPRINGER J, MITCHELL M, KENYON G. A little robustness goes a long way: leveraging robust features for targeted transfer attacks[J]. *Advances in Neural Information Processing Systems*, 2021, 34: 9759-9773.
- [20] ZOU J, PAN Z, QIU J, et al. Improving the transferability of adversarial examples with resized-diverse-inputs, diversity-ensemble and region fitting[C]//*Proceedings of European Conference on Computer Vision*. Berlin, Germany: Springer International Publishing, 2020: 563-579.
- [21] DONG Y P, LIAO F Z, PANG T Y, et al. Boosting adversarial attacks with momentum[C]//*Proceedings of the*

- IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE Press, 2018: 9185-9193.
- [22] WANG X S, HE K. Enhancing the transferability of adversarial attacks through variance tuning[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville, USA: IEEE Press, 2021: 1924-1933.
- [23] WANG X S, HE X R, WANG J D, et al. Admix: enhancing the transferability of adversarial attacks [C] // Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). Montreal, Canada: IEEE Press, 2021: 16138-16147.
- [24] BYUN J, KWON M J, CHO S, et al. Introducing competition to boost the transferability of targeted adversarial examples through clean feature mixup[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Vancouver, Canada: IEEE Press, 2023: 24648-24657.

编辑 金胡考

计算机工程
www.ecice06.com