

# 基于 RAdam 和颜色不变性的对抗样本可迁移性攻击<sup>\*</sup>

丁佳, 许智武

深圳大学 计算机与软件学院, 广东 深圳 518061

通讯作者: 许智武, E-mail: xuzhiwu@szu.edu.cn



**摘要:** 深度神经网络在物体检测、图像分类、自然语言处理、语音识别等众多领域上得到广泛应用。然而, 深度神经网络很容易受到对抗样本(即在原有样本上施加人眼无法察觉的微小扰动)的攻击, 而且相同的扰动可以跨模型、甚至跨任务地欺骗多个分类器。对抗样本这种跨模型迁移特性使得深度神经网络在实际生活的应用受到了很大限制。对抗样本对神经网络的威胁激发了研究者对对抗攻击的研究兴趣和发展。虽然研究者们已提出了不少对抗攻击方法, 但是大多数这些方法(特别是黑盒攻击方法)的跨模型的攻击能力往往较差, 尤其是对经过对抗训练、输入变换等的防御模型。为此, 本文提出了一种提高对抗样本可迁移性的方法, RLI-CI-FGSM。RLI-CI-FGSM 是一种基于迁移的攻击方法, 在替代模型上使用基于梯度的白盒攻击 RLI-FGSM 生成对抗样本, 同时使用 CIM 扩充源模型, 使 RLI-FGSM 能够同时攻击替代模型和扩充模型。具体而言, RLI-FGSM 算法将 RAdam 优化算法与迭代快速符号下降法相结合, 并利用目标函数的二阶导信息来生成对抗样本, 避免优化算法陷入较差的局部最优。基于深度神经网络具有一定的颜色变换不变性, CIM 算法通过优化对颜色变换图像集合的扰动, 针对防御模型生成更多可迁移的对被攻击的白盒模型不那么敏感的对抗样本。实验结果表明, 本文的方法在一般网络和对抗网络模型上都产生了更高的成功率。

**关键词:** 对抗样本; 对抗攻击; 黑盒攻击; 可迁移性; 基于迁移的攻击  
**中图法分类号:**

中文引用格式: 丁佳, 许智武. 基于 RAdam 和颜色不变性的对抗样本可迁移性攻击. 软件学报. <http://www.jos.org.cn/1000-9825/6589.htm>

英文引用格式: Ding J, Xu ZW. Improving the Transferability of Adversarial Examples with RAdam and Color Invariance. RuanJianXuBao/Journal of Software, 2022 (in Chinese). <http://www.jos.org.cn/1000-9825/6589.htm>

## Improving the Transferability of Adversarial Examples with RAdam and Color Invariance

DING Jia, XU Zhi-Wu

College of Computer Science and Software Engineering, Shenzhen University, Shenzhen 518061, China

**Abstract:** Deep neural networks have been widely used in object detection, image classification, natural language processing, speech recognition, and so on. However, deep neural networks are vulnerable to adversarial examples which could misclassify deep neural network classifiers by adding imperceptible perturbations to the input. Moreover, the same perturbation can deceive multiple classifiers across models and even across tasks. The cross-model transfer characteristic of adversarial examples limits the application of deep neural network in real life. The threat of adversarial examples to deep neural networks has aroused researchers' interest in adversarial attack. Recently, researchers have proposed several adversarial attacks, but the cross-model ability of adversarial examples generated by the existing attacks is often poor, especially for the defense models via adversarial training or input transformation. To improve the transferability of adversarial examples in black box environment, this paper proposes a method, namely, RLI-CI-FGSM. RLI-CI-FGSM is a transfer-based attack, which employs the gradient-based white-box attack RLI-

<sup>\*</sup> 基金项目: 国家自然科学基金(61836005, 61972260, 61772347); 广东省基础与应用基础研究基金项目(2019A1515011577); 深圳市高校稳定支持计划项目(20200810150421002).

收稿时间: 2021-09-05; 修改时间: 2021-10-14; 采用时间: 2022-01-10; jos 在线出版时间: 2022-01-28

FGSM to generate adversarial examples on the substitute model, as well as CIM to expand the substitute model so that RLI-FGSM is able to attack the substitute model and the extended model at the same time. Specifically, RLI-FGSM integrates the RAdam optimization algorithm into iterative fast gradient sign method, and makes use of the second-derivative information of objective function to generate adversarial examples, which prevents optimization algorithm from falling into poor local optimum. Based on the Color transformation-invariant property of deep neural networks, CIM optimizes the perturbations of the color transform image sets to generate adversarial examples that are less sensitive to the defense models. The experimental results show that the proposed method has a higher success rate in both normal and adversarial network models.

**Key words:** adversarial examples; adversarial attack; black-box attack; transferability; transfer-based attack

近年来, 神经网络凭借其能用较短时间以较高的准确性解决多种复杂问题的优势, 在物体检测<sup>[1]</sup>、图像分类<sup>[2]</sup>、自然语言处理<sup>[3]</sup>、语音识别<sup>[4] [5]</sup>等众多领域上取得了优异的成绩。然而, C. Szegedy 等人<sup>[6]</sup>发现了神经网络的一个致命性弱点——通过施加人眼无法察觉的微小扰动诱使神经网络分类器错误地分类。这些有意制作的能诱导神经网络分类器分类错误的输入样本被称为对抗样本, 施加在其上的扰动被称为对抗扰动。这个致命性弱点使得神经网络非常容易受到攻击 (即对抗攻击), 尤其在图像方面的网络。此外, 相同的图像扰动可以欺骗多个分类器 (本文中也称为模型)。这种由一个模型生成的对抗样本能够以一定的概率欺骗另一个不同模型的性质, 称为对抗样本的可迁移性。可迁移性使得对抗扰动可以跨模型甚至跨任务<sup>[7]</sup>转移, 更进一步的影响神经网络的安全性。这些发现大大影响了神经网络在商业领域, 尤其是在安全攸关重要的领域的应用, 比如自动驾驶<sup>[8]</sup>, 网络安全<sup>[9]</sup>等方面。

对抗样本对神经网络的威胁不仅激发了研究者对构建更具鲁棒性的神经网络的兴趣, 而且促发了如何构建有效健壮的对抗攻击的研究<sup>[10]</sup>。这两者之间不断博弈且互相促进。现有的对抗攻击工作大多数使用基于梯度的方法来生成对抗样本, 比如迭代梯度下降法(I-FGSM)<sup>[11]</sup>, 投影梯度下降(PGD)<sup>[12]</sup>, 动量迭代快速梯度符号法(MI-FGSM)<sup>[13]</sup>等。在白盒设置下, 利用现有模型的知识 (例如网络结构和网络参数), 这些攻击方法可以获得较高的成功率。然而, 在黑盒设置下, 由于无法获得模型的内部信息, 这些攻击方法的成功率往往较低。特别地, 在对经过对抗训练<sup>[14]</sup>或输入变换 (例如图像变形<sup>[15]</sup>/去噪<sup>[16]</sup>) 的防御模型进行攻击时, 这些攻击方法的成功率被进一步降低。

基于迁移的攻击<sup>[17][18]</sup> (transfer-based attack) 就是一种致力于提高黑盒攻击成功率的方法, 该方法先在替代模型 (源模型) 上使用白盒攻击生成对抗样本, 再用生成的对抗样本去攻击未知的目标模型。已有的基于迁移的对抗攻击通常使用基于梯度的方法来构造对抗样本, 从目前的研究进展而言, 仍然有较大的进度空间。本文旨在设计一种新的能够提高对抗样本可迁移性的算法, 换句话说, 构造一个健壮的基于迁移的攻击算法。

本文把对抗样本的生成过程视为神经网络的训练过程, 把在替代模型上生成对抗样本的过程视为神经网络在训练数据上训练模型, 把用生成的对抗样本攻击目标模型的黑盒过程视为神经网络在测试数据上测试模型性能, 那么对抗样本的迁移能力就可以看成是神经网络的泛化能力。因此本文主要从三个方面提高对抗样本的迁移能力: (1) 优化基于梯度的白盒攻击方法; (2) 对替代模型进行数据扩充; (3) 集成 (ensemble) 方法。第一点本文通过改进优化算法避免其陷入较差的局部最优, 尽可能使算法搜索到更优的对抗样本, 更优的对抗样本往往具有更强的迁移性能; 第二点从本文使用提高泛化能力的角度看, 通过数据增强对图像进行几何变换, 丰富同类数据的表现形式, 这能够防止对抗样本对替代模型过拟合, 提高对抗样本在目标模型上的迁移成功率; 第三点, ensemble 方法利用多个模型的信息, 即所谓“群体的智慧”能够很好的提高的模型泛化能力。另一方面, 从集成攻击[19]的角度看, 攻击算法从只攻击一个源模型变成同时攻击源模型和扩充模型, 生成的对抗样本也从只能欺骗一个模型变成能够欺骗多个模型, 直觉和经验结果都表明能够同时欺骗多个模型的对抗样本更有可能成功迁移到其他的黑盒模型。

受 MI-FGSM、NI-FGSM<sup>[20]</sup>的启发, 优化算法的选择能够影响对抗样本的生成及其可迁移性, 因此本文采用了自适应矩估计优化算法 Adam 的变体 Rectified Adam (RAdam) 算法, 并结合迭代快速符号下降法和

利用目标函数（对抗样本生成对应的优化问题，具体见 1.1）的二阶导信息来生成对抗样本。该方法能够有效地避免其陷入较差的局部最优，提高对抗样本的可迁移性。

另一个能够提高对抗样本的可迁移性的方法是模型扩充。现有的数据扩充方法大致可以分为三类:空间变换<sup>[21]</sup>、颜色变换<sup>[22]</sup>和信息丢弃<sup>[23]</sup>。在空间变换方面，DIM<sup>[17]</sup>、TIM<sup>[18]</sup>、SIM<sup>[20]</sup>等工作通过对输入图像的空间变换提高了对抗样本的可迁移性；在信息丢弃方面，Xie<sup>[24]</sup>等人提出了基于输入的 dropout 方法；而在颜色变换方面，就作者所知，在对抗攻击中尚未有充分完整的研究内容。因此，本文还应用颜色变换方法来进一步提高对抗样本的可迁移性。

本文的主要贡献如下：

1.提出了基于 RAdam 优化算法的对抗攻击方法 RLI-FGSM。该方法将 RAdam 优化算法与迭代快速符号下降法相结合，并利用目标函数的二阶导信息来生成对抗样本。

2.提出了基于颜色变换的对抗攻击方法 CIM。该方法应用了颜色变换进行数据扩充，填补并完善了颜色变换在对抗攻击的应用。

3.通过在 ILSVRC 2012 验证集的一些实验验证了本文方法的有效性，且结果显示本文方法不仅在正常训练的模型上产生更高的成功率，而且还打破了其他强大的对抗网络的防御机制。

本文第 1 节定义了所用符号并介绍相关工作，第 2 节对本文提出的对抗样本生成算法进行详细描述，第 3 节为实验设置以及结果的分析讨论，第 4 节总结全文。

## 1 符号定义和相关工作

### 1.1 定义

表 1 列举了本文所需了解的相关术语和其解释。

**Table 1** Definition of Terms and Notations

**表 1** 术语和符号定义

名称和符号	描述
深度神经网络分类器 $f$	$f: \mathcal{X} \rightarrow \mathcal{Y}$ , 接受图像 $x \in \mathcal{X}$ 作为输入并输出某一个标签 $y \in \mathcal{Y}$ 。
输入图像 $x$	干净的、原始的图像
输出标签 $y$	分类器 $f$ 的预测标签
真实标签为 $y^{true}$	输入图像 $x$ 对应的真实标签
对抗样本 $x^{adv}$	在图像上施加人眼无法察觉的微小扰动所形成的能够使分类器分类错误的样本
对抗扰动 $\eta$	施加在原图像上的微小扰动, $\eta = x^{adv} - x$
$L_p$ 范数	$\ \theta\ _p = \left( \sum_{i=1}^n  \theta_i ^p \right)^{\frac{1}{p}}$
$L_\infty$ 范数	$\ \theta\ _\infty = \max( \theta_i )$
扰动阈值 $\epsilon$	用 $L_p$ 范数限制扰动的大小, 确保人眼无法察觉, 即 $\ \eta\ _p = \ x^{adv} - x\ _p \leq \epsilon$

根据表 1，对抗样本有两个条件：(i) 在原图像的微小抖动和 (ii) 分类器分类错误。本文使用  $L_\infty$  范数表示这个微小抖动。因此，这两个条件可以表示为

$$\begin{aligned} \min_{\eta} \|\eta\|_\infty &= \min_{x^{adv}} \|x^{adv} - x\|_\infty \\ s.t. f(x^{adv}) &= y, f(x) = y^{true}, y \neq y^{true} \end{aligned}$$

(1)

通过形式化定义可以发现对抗样本的生成过程是一个优化问题，即寻找导致错误分类的模型输入的最小扰动。对于上述约束优化问题，由于其求解的困难性，通常引入模型训练中使用的损失函数  $J$  来间接解决，其

中 $J$ 通常为交叉熵损失函数(*crossentropy loss*)。对于无目标攻击(*untargeted attack*), 即只要求 $y \neq y^{true}$ 的攻击而不指定输出标签目标 $y$ 的类别, 因此, 要求输出标签 $y$ 与真实标签 $y^{true}$ 的距离越大越好。从神经网络的角度来看, 这可以等价要求损失函数 $J(x^{adv}, y^{true})$ 越大越好。因此, 公式(1)可以转换成如下形式:

$$\begin{aligned} \arg \max_{x^{adv}} J(x^{adv}, y^{true}), s.t. \|x^{adv} - x\|_{\infty} \leq \varepsilon \text{ 或} \\ \arg \max_{\eta} J(x + \eta, y^{true}), s.t. \|\eta\|_{\infty} \leq \varepsilon \end{aligned} \quad (2)$$

公式(2)中的约束项可以通过 $Clip$ 函数对 $x$ 进行剪切来控制 $x$ 上的扰动大小而消除, 消除后的公式就是一个易于求解的无约束优化问题。对于公式(2), 本文采用 RAdam 优化算法和目标函数的二阶导信息进行求解。

## 1.2 对抗攻击

### (1) FGSM (Fast Gradient Sign Method)

Goodfellow 等人提出了快速梯度符号法(FGSM)<sup>[25]</sup>, 在梯度方向上通过一大步更新生成使损失函数最大化的扰动, 更新规则如下:

$$\begin{aligned} \eta &= \varepsilon \square sign(\nabla_x J(\theta, x, y^{true})) \\ x^{adv} &= x + \eta \end{aligned}$$

(3)

其中 $\nabla_x J$ 是 $x$ 的损失函数的梯度,  $\theta$ 是模型的参数。

### (2) I-FGSM (Iterative Fast Gradient Sign Method)

由于 FGSM 算法只涉及一次梯度更新, 有时一次更新不足以成功攻击且一次更新容易陷入局部最优值, Kurakin 等人<sup>[11]</sup>提出了基于 FGSM 的迭代快速梯度符号法(I-FGSM), 采用小步长 $\alpha$ 多次迭代。

$$x_{t+1}^{adv} = x_t^{adv} + \alpha \square sign(\nabla_x J(x_t^{adv}, y^{true})) \quad (4)$$

### (3) PGD (Projected Gradient Descent)

Madry 等人<sup>[12]</sup>提出了投影梯度下降(PGD), 这是一种比 I-FGSM 和 FGSM 更强大的梯度攻击。它在允许的标准球内的随机点初始化搜索对抗样本, 然后运行 I-FGSM 方法多次迭代。

$$x_{t+1}^{adv} = Clip_x^{\varepsilon} \left\{ x_t^{adv} + \alpha \square sign(\nabla_x J(x_t^{adv}, y^{true})) \right\}$$

(5)

其中 $Clip_x^{\varepsilon}$ 函数通过对 $x$ 进行剪切来控制 $x$ 上的扰动大小。

### (4) MI-FGSM(Momentum Iterative Fast Gradient Sign Method)

Dong 等人<sup>[13]</sup>提出了动量迭代快速梯度符号法(MI-FGSM), 动量法参数的更新过程中使用的是累积梯度 $g_t$ , 其中  $t$  表示在迭代次数。对于当前梯度与上一步梯度指向相同方向的维度,  $g_{t+1}$ 增加, 而对于当前梯度与上一步梯度指向不同方向的维度,  $g_{t+1}$ 减少。

$$\begin{aligned} g_{t+1} &= \mu \square g_t + \frac{\nabla_x J(x_t^{adv}, y^{true})}{\|\nabla_x J(x_t^{adv}, y^{true})\|_1} \\ x_{t+1}^{adv} &= Clip_x^{\varepsilon} \left\{ x_t^{adv} + \alpha \square sign(g_{t+1}) \right\} \end{aligned}$$

(6)

动量法得到了更快的收敛和减少振荡。因为用 I-FGSM 生成对抗样本时仍然容易陷入较差的局部最大值和过拟合模型, 这样生成的对抗样本是不太可能在模型之间转移的。MI-FGSM 将动量集成到迭代攻击中, 可以稳定更新方向, 有助于摆脱较差的局部极值, 减轻过拟合的影响, 因此 MI-FGSM 能够提高对抗样本的可迁移性。

#### (5) NI-FGSM(Nesterov Iterative Fast Gradient Sign Method)

Lin 等人<sup>[20]</sup>提出了 Nesterov 迭代快速梯度符号法(NI-FGSM), 将 Nesterov 加速梯度引入迭代攻击中, 在 MI-FGSM 的基础上, NI-FGSM 还利用了目标函数的二阶导信息, 使得攻击能够有效的向前看, 能够更容易、更快地摆脱局部极值差, 从而提高可迁移性。

$$x_t^{nes} = x_t^{adv} + \alpha \mu \square g_t$$

$$g_{t+1} = \mu \square g_t + \frac{\nabla_x J(x_t^{nes}, y^{true})}{\|\nabla_x J(x_t^{nes}, y^{true})\|_1}$$

(7)

$$x_{t+1}^{adv} = Clip_x^\epsilon \{x_t^{adv} + \alpha \square sign(g_{t+1})\}$$

#### (6) DIM (Diverse Inputs Method)

Xie 等人<sup>[17]</sup>提出了一种多样化输入方法 DIM 来改进对抗样本的可迁移性。DIM 随机应用一组保持标签的变换(例如, 调整大小、裁剪和旋转)来训练图像, 并将变换后的图像输入分类器进行梯度计算。

$$x_{t+1}^{adv} = Clip_x^\epsilon \left\{ x_t^{adv} + \alpha \square sign \left( \nabla_x J \left( T(x^{adv}; p), y^{true}; \theta \right) \right) \right\}$$

(8)

其中 $T(\cdot)$ 表示图像变换,  $p$ 表示进行变换的概率。

#### (7) TIM(Translation-Invariant Attack Method)

Dong 等人<sup>[18]</sup>提出了一种平移不变(TIM)攻击方法, 通过使用一组平移后的图像对对抗样本进行优化, 使对抗样本对被攻击的白盒模型的区分区域不那么敏感, 从而提高了对抗样本的可移动性。TIM 证明了对图像进行平移操作后求梯度等价于将梯度与平移矩阵的所有权值组成的核进行卷积:

$$x_{t+1}^{adv} = x_t^{adv} + \alpha \square sign(W \square \nabla_x J(x^{adv}, y^{true}))$$

(9)

其中  $W$  表示平移矩阵。

#### (8) SIM(Scale-Invariant Attack Method)

Lin 等人<sup>[20]</sup>还提出了另一种提高对抗样本可迁移性的方法 SIM。SIM 利用模型的尺度不变特性实现模型扩展, 提高对抗样本的可迁移性。

$$\arg \max_{x^{adv}} \frac{1}{m} \sum_{i=0}^m J(S_i(x^{adv}), y^{true}), s.t. \|x^{adv} - x\|_\infty \leq \epsilon$$

(10)

其中 $S_i(x) = x/2^i$ 表示输入图像 $x$ 的缩放副本, 缩放因子为 $1/2^i$ ,  $m$ 表示缩放副本的数量。

#### (9) RD-DE-RF-IFGSM(Resized-Diverse-Inputs,Diversity-Ensemble and Region Fitting Method)

Zou 等人<sup>[26]</sup>发现在不同输入的梯度中有许多垂直和水平条纹,可以用来缓解 TIM 造成的梯度信息的丢失。他们提出不同输入大小的方法(RDIM)可以与 TIM 相结合来发挥更好的攻击性能。此外,他们还提出了多样性集成方法(DEM),即 RDIM 的多尺度版本,以进一步提高对抗样本的可移植性。在前两个步骤之后,再通过迭代将值拟合转化为区域拟合。RDIM 和区域拟合不需要额外的运行时间,这三个步骤可以很好地集成到其他攻击中。

#### (10) UAP (Universal adversarial perturbations)

通用对抗扰动<sup>[27]</sup>指的是仅用一个小的图像扰动就能以高概率欺骗深度神经网络分类器,使分类器对大部分图像分类错误。换句话说,就是需要找到一个对抗扰动  $\eta$ ,这个扰动可以加到所有的样本点上,而且会以  $1-\delta$  的概率让对抗样本被分类错误。

$$P_{x \sim \mu} (f(x+\eta) \neq f(x)) \geq 1-\delta, s.t. \|\eta\|_{\infty} \leq \varepsilon$$

## 2 算法

本章将详细描述本文的攻击算法 RLI-CI-FGSM。第 2.1 节先简单介绍了 Adam 和 RAdam 算法,第 2.2 节介绍了基于 RAdam 的攻击方法 RLI-FGSM,第 2.3 节介绍了基于颜色变换不变的攻击方法 CIM。最后,第 2.4 节给出了这两者的合成攻击算法 RLI-CI-FGSM。

### 2.1 Adam和RAdam算法

本文所采用的 Rectified Adam (RAdam)优化算法,是 Adam 算法的一种变体,因此在介绍 RAdam 算法之前,先对 Adam 算法进行简单的说明。Adam(adaptive moment estimation)<sup>[28]</sup>是一种只需要一阶梯度和很少的内存需求的有效的随机优化方法,该方法根据梯度的一阶和二阶矩估计计算不同参数下的个体自适应学习速率,可以表示为:

$$m_t = \beta_1 \square m_{t-1} + (1 - \beta_1) \square g_t \quad (11)$$

$$v_t = \beta_2 \square v_{t-1} + (1 - \beta_2) \square g_t^2 \quad (12)$$

其中  $m_t$  和  $v_t$  分别表示梯度的一阶矩(表示梯度均值)和二阶矩(表示方差),  $\beta_1, \beta_2$  表示衰减率,  $g_t$  表示在迭代次数为  $t$  时的累加梯度。用平均移动(moving average)来对一阶矩和二阶矩进行估计时,因为  $m_t$  和  $v_t$  被初始化为 0 向量,导致矩估计偏向于 0,所以 Adam 使用偏差校正解决初始化偏差对移动平均的影响。

$$\square m_t = \frac{m_t}{1 - \beta_1^t} \quad (13)$$

$$\square v_t = \frac{v_t}{1 - \beta_2^t} \quad (14)$$

接着使用上述公式更新 Adam 的参数:

$$\theta_{t+1} = \theta_t - \frac{\eta}{\sqrt{v_t} + \epsilon} \bar{m}_t$$

(15)

其中 $\eta$ 表示步长,  $\theta_t$ 表示结果参数,  $\epsilon$ 是一个很小的值, 保证分母不为 0。

Liu 等人<sup>[29]</sup>发现由于训练早期样本的缺乏, 自适应学习率 $v_t$ 的方差过大, 导致陷入较差的局部最优。因此, Liu 等人提出了 RAdam 算法, 引入了修正项 $r_t$ 来修正其自适应学习率, 稳定其早期无边界的方差, 同时实验表明 RAdam 算法也能加速收敛和改进泛化性能。RAdam 算法更新规则如下, 其中 $m_t$ ,  $v_t$ 和 $\bar{m}_t$ 的计算同 Adam 一致, 不再重复:

$$\rho_\infty = \frac{2}{1 - \beta_2} - 1$$

(16)

$$\rho_t = \rho_\infty - 2t\beta_2^t / (1 - \beta_2^t)$$

(17)

当方差易处理时, 如 $\rho_t > 4$

$$l_t = \sqrt{(1 - \beta_2^t) / v_t}$$

(18)

$$r_t = \sqrt{\frac{(\rho_t - 4)(\rho_t - 2)\rho_\infty}{(\rho_\infty - 4)(\rho_\infty - 2)\rho_t}}$$

(19)

$$\theta_t = \theta_{t-1} - \alpha_t r_t \bar{m}_t l_t$$

(20)

当方差不易处理时, 如 $\rho_t \leq 4$

$$\theta_t = \theta_{t-1} - \alpha_t \bar{m}_t$$

(21)

其中 $\rho_\infty$ 表示近似 SMA (simple moving average) 的最大长度,  $\rho_t$ 表示近似 SMA 的长度,  $\alpha_t$ 表示步长,  $l_t$ 表示 RAdam 中的自适应学习率,  $r_t$ 表示方差校正项,  $\theta_t$ 表示结果参数。

具体来说, 在训练初期自适应学习率 $v_t$ 的方差过大甚至可以趋于无穷大, 这时 Adam 的更新方法将不再可靠, 所以 RAdam 在初期退化带动量的 SGD (随机梯度下降), 即如公式(21)所示不再使用自适应学习率。RAdam 使用了只与 $t$ 有关的参数 $\rho_t$ 控制算法是否退化。当 $\beta_2 = 0.999$  (Adam 推荐的参数值), 在 $t = 1 \sim 4$ 时 (即 $\rho_t \leq 4$ ), 方差过大不易处理, 方差校正项 $r_t$ 会出现一定的震荡, 甚至会出现对负数开根号的情况, 而在 $t > 4$ 时 (即 $\rho_t > 4$ ), 方差不再有趋于无穷大的风险易于处理,  $r_t$ 的更新趋于稳定。因此, RAdam 设置 $\rho_t \leq 4$ 时, 算法退化带动量的 SGD, 设置 $\rho_t > 4$ , 算法不退化, 使用方差校正项 $r_t$ 进一步稳定

算法。

## 2.2 RI-FGSM和RLI-FGSM方法

本文首先将 RAdam 算法集成到基于梯度的迭代攻击中, 构建一个健壮的对抗攻击, 称之为 RI-FGSM。从  $g_0 = 0$  开始, RI-FGSM 的更新过程形式化如下,  $\alpha_t, r_t, \widehat{m}_t, l_t$  的计算同 RAdam 一致, 在公式(21)后计算:

$$g_t = \mu \square g_{t-1} + \frac{\nabla_x J(x_{t-1}^{adv}, y^{true})}{\|\nabla_x J(x_{t-1}^{adv}, y^{true})\|_1} \quad (22)$$

当方差易处理时, 如  $\rho_t > 4$

$$x_t^{adv} = Clip_x^\varepsilon \left\{ x_{t-1}^{adv} + \alpha_t r_t \square sign(\widehat{m}_t \square l_t) \right\} \quad (23)$$

当方差不易处理时, 如  $\rho_t \leq 4$

$$x_t^{adv} = Clip_x^\varepsilon \left\{ x_{t-1}^{adv} + \alpha_t \square sign(\widehat{m}_t) \right\} \quad (24)$$

其中  $g_t$  表示在迭代次数为  $t$  时的累加梯度,  $\mu$  是  $g_t$  的衰减因子,  $sign$  是符号函数。

受 Nesterov 加速梯度的启发, 本文在 RI-FGSM 的基础上再利用目标函数的二阶导信息, 使得攻击能够有效的向前看(look ahead), 构造了 RLI-FGSM 算法。具体而言, RLI-FGSM 中的梯度不是根据当前参数位置  $x_t^{adv}$  计算的, 而是根据参数下一个位置的近似值  $x_t^{lookahead}$  计算的, 通过这个近似值粗略地知道参数下一步将在哪里从而有效地预测未来。这种预先的更新可以进一步帮助算法更容易、更快地摆脱较差的局部极大值, 从而提高生成的对抗样本的可迁移性。RLI-FGSM 算法与 RI-FGSM 相比, 前者的梯度多加了本次梯度相对于上次梯度的变化量。RLI-FGSM 的更新分为三步:

(1) 计算新参数近似位置  $x_t^{lookahead}$

当方差易处理时, 如  $\rho_t > 4$

$$x_t^{lookahead} = x_t^{adv} + \alpha_t r_t \square sign(\widehat{m}_t \square l_t) \quad (25)$$

当方差不易处理时, 如  $\rho_t \leq 4$

$$x_t^{lookahead} = x_t^{adv} + \alpha_t \square sign(\widehat{m}_t) \quad (26)$$

(2) 根据新参数近似位置更新梯度



$$g_{t+1} = \mu g_t + \frac{\nabla_x J(x_t^{lookahead}, y^{true})}{\|\nabla_x J(x_t^{lookahead}, y^{true})\|_1}$$

(27)

(3)更新  $x_{t+1}^{adv}$ 当方差易处理时, 如  $\rho_t > 4$ 

$$x_{t+1}^{adv} = x_t^{adv} + \alpha_{t+1} r_{t+1} \text{sign}(m_{t+1} \odot l_{t+1}) \quad (28)$$

当方差不易处理时, 如  $\rho_t \leq 4$ 

$$x_{t+1}^{adv} = x_t^{adv} + \alpha_{t+1} \text{sign}(m_{t+1}) \quad (29)$$

### 2.3 CIM方法

除了为对抗攻击考虑一个更好的优化算法外, 本文还考虑通过模型扩充来提高对抗样本的可迁移性。在同一模型上, 若原始图像和经变换后的图像的损失值是相似的, 则称这种变换为保损变换。通过保损变换, 本文可以从原模型推导出一个模型集合, 实现模型扩充。

为了得到保损变换, 本文发现深度神经网络具有颜色变换不变性并进行了实验验证, 即适量的改变深度神经网络的颜色特性, 若变化前后网络的损失值可以忽略不计。因此, 颜色变换可以作为一种模型扩充方法。在上述分析的驱动下, 本文提出了一种颜色变换不变方法 CIM。CIM 通过对模型的颜色扩充, 包含亮度, 色相, 对比度, 饱和度四个特征, 优化了对输入图像的对抗扰动。

令  $C_i(x)$  表示颜色变换不变性, CIM 方法描述如下:

$$\arg \max_{x^{adv}} J(C_i(x^{adv}), y^{true}), C = \{C_1, C_2, C_3, C_4\}$$

(30)

$$\|x^{adv} - x\|_{\infty} \leq \varepsilon$$

其中  $C_1, C_2, C_3, C_4$  分别表示亮度, 色相, 对比度, 饱和度的抖动变化函数。

(1)对于亮度特征, 本文改变整张 RGB 图片的亮度, 即在图片  $x$  上每一个像素值上添加一个范围在  $[-\max \delta_1, \max \delta_1]$  之间的调整因子  $\delta_1$ , 调整亮度后的图像为:

$$C_1(x) = x + \delta_1$$

(31)

(2)对于色相特征, 本文先把图片  $x$  从 RGB 颜色空间转换成 HSV 颜色空间, 向色相通道(H)添加偏移量  $\delta_2$  来调整图片的色相, 然后再从 HSV 颜色空间转换回 RGB。色相特征调整函数如下所示:

$$C_2(x) = RGB(HSV(x) + {}_H \delta_2)$$

(32)

(3)对于对比度特征, 本文对 RGB 图片的每个通道都计算通道中图像像素的均值, 然后根据均值调整每个像素的每个分量, 如下所示:

$$C_3(x_i) = (x_i - \text{mean}) * \delta_3 + \text{mean}$$

(33)

其中  $\delta_3$  为对比度调整因子。

(4) 对于饱和度特征, 本文先把图片  $x$  从 RGB 颜色空间转换成 HSV 颜色空间, 并将饱和度(S)通道乘以饱和度因子  $\delta_4$  和剪切来调整图像饱和度, 最后再将图像转换回 RGB。饱和度特征调整函数如下所示:

$$C_4(x) = \text{RGB}(\text{Clip}(\text{HSV}(x) * \delta_4))$$

(34)

具体的调整因子的选择在 3.2 节中给出, 利用 CIM, 本文不需训练一组模型来攻击, 而是通过模型扩充来有效地实现对多个模型的集成攻击。更重要的是, 它可以帮助避免对被攻击的白盒模型的过拟合, 并生成更多可迁移的对抗样本。

## 2.4 攻击算法

本文提出的 RLI-CI-FGSM 攻击算法, 是将基于优化算法的攻击方法 RLI-FGSM 和基于模型扩充的方法 CIM 结合起来的一个更强大的攻击方法, 具体伪代码如算法 1 所示。

---

### 算法 1 RLI-CI-FGSM

---

输入: 干净的样本  $x$  及其对应的真实标签  $y^{\text{true}}$ , 分类器  $f$ , 损失函数  $J$ ; 最大扰动  $\epsilon$ , 最大迭代次数  $T$ , 衰减系数  $\mu$ , 颜色变换扩充的模型数量  $i$ 。

输出: 对抗样本  $x^{\text{adv}}$

1.  $\alpha_0 = \epsilon$  // 初始步长
  2.  $g_0 = 0$ ;  $x_0^{\text{adv}} = 0$
  3. **for**  $t=0$  to  $T-1$  **do**
  4.  $g = 0$
  5. calculate  $x_t^{\text{lookahead}}$  by Eq.(25) or Eq.(26)
  6. **for**  $i=0$  to 4 **do** //  $i=0$  表示未经颜色变换,  $i=1,2,3,4$  表示经亮度不变性, 色相不变性, 对比度不变性, 饱和度不变性四种子方法变换
  7. calculate the gradients:  $\nabla_x J(C_i(x_t^{\text{lookahead}}), y^{\text{true}})$
  8. sum the gradients:  $g = g + \nabla_x J(C_i(x_t^{\text{lookahead}}), y^{\text{true}})$
  9. calculate average gradients:  $g = \frac{1}{5} \cdot g$
  10. update  $g_{t+1} = \mu \cdot g_t + \frac{g}{\|g\|_1}$
  11. update  $x_{t+1}^{\text{adv}}$  by Eq.(26)
  12. **Return**  $x^{\text{adv}}$
- 

## 3 实验

本章将介绍实验结果, 以验证本文所提方法的有效性。第 3.1 节介绍了实验设置, 包括数据集, 模型和超参数的设置。第 3.2 节验证了深度神经网络的颜色变换不变特性。第 3.3 节研究了初始步长对攻击成功率的影响, 挑选出了合适的初始步长。第 3.4 节在单一模型下分别将本文方法与基于优化的方法和基于模型扩充的方法进行了比较。第 3.5 节在多模型集合的环境下将本文方法与基线方法进行集成和比较。

### 3.1 实验设置

#### (1) 数据集

本文从 ILSVRC 2012 验证集的 1000 个类别中随机选择了 1000 张图片, 这些图片几乎被所有测试模型正确分类。ILSVRC 2012 是 ImageNet2012 竞赛的数据集, 是图像分类数据集中最常用的测试数据集和预训

练数据集之一。

### (2) 模型

在正常训练的模型方面, 本文考虑了 Inception-v3 (Inc-v3)<sup>[30]</sup>、Inception-v4 (Inc-v4)<sup>[31]</sup>、Inception-Resnet-v2 (IncRes-v2)<sup>[31]</sup> 和 Resnet-v2-101 (Res-101)<sup>[32]</sup>。

在对抗训练的模型方面, 本文考虑 Inc-v3<sub>ens3</sub>, Inc-v3<sub>ens4</sub> 和 IncRes-v2<sub>ens</sub><sup>[32]</sup>。

### (3) 超参数设置

本文设置最大扰动  $\epsilon = 16$ , 迭代次数  $T = 10$ , RLI-FGSM 的步长  $\alpha = \epsilon$ , 其余方法的步长  $\alpha = \epsilon / T$ 。对于 MI-FGSM, 本文采用默认的衰减因子  $\mu = 1.0$ ; 对于 DIM, 变换概率设为 0.5; 对于 TIM, 本文采用大小为  $7 \times 7$  的高斯核。对于 CIM, 亮度调整因子  $\delta_1 \in [-0.9, 0.9]$ , 色相调整因子  $\delta_2 = 0.1$ , 对比度调整因子  $\delta_3 = 0.2$ , 饱和度调整因子  $\delta_4 = 0.6$ 。

## 3.2 颜色变换不变特性

为了验证深度神经网络的尺度不变特性, 本文从 ImageNet 数据集中随机选取 1000 幅原始图像, 对图像的亮度, 色相, 对比度, 饱和度四个颜色特征分别进行调整。然后将原始图像和进行四种调整后的图像输入测试模型 Inc-v3 中, 每张输入图像的损失值由模型给出, 下图给出了不同输入下的 1000 张图片的平均损失值。

如下图 1 所示, 四个颜色特征的调整因子取值范围都不同, 根据观察, 取值范围的选择综合考虑了模型损失值和攻击目标模型的成功率。本文中, 对于亮度特征, 将调整因子大小保持在  $[0.1, 0.9]$  间时, 原始图像和变换图像的损失值相差不大。因此我们假设深度模型在此范围内有颜色变换不变属性; 对于对比度特征, 将调整因子大小保持在  $[0.2, 1.5]$  间时, 原始图像和变换图像的损失值相差不大。因此我们假设深度模型在此范围内有对比度变换不变属性; 对于色相特征, 将调整因子大小保持在  $[0, 0.1]$  间时, 原始图像和变换图像的损失值相差不大。因此我们假设深度模型在此范围内有色相变换不变属性; 对于饱和度特征, 将调整因子大小保持在  $[0.6, 3.2]$  间时, 原始图像和变换图像的损失值相差不大。因此我们假设深度模型在此范围内有饱和度变换不变属性。

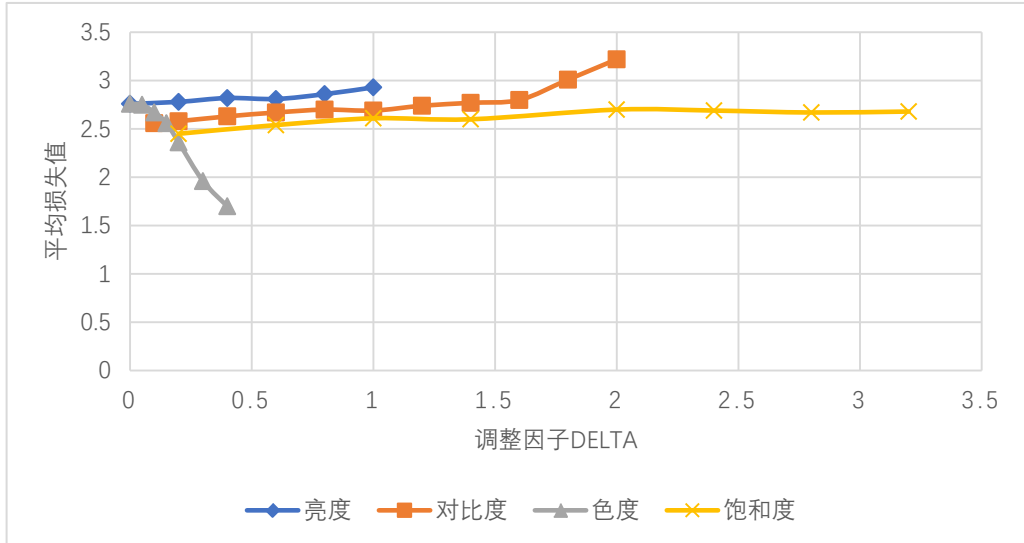


图 1 Inc-v3 模型在亮度, 色相, 对比度, 饱和度四个颜色特征影响下在 1000 张图片上的平均损失值

### 3.3 步长

接下来研究初始步长对攻击成功率的影响, 本节在不同步长下以 Inc-v3 为源模型生成对抗样本攻击 Inc-v3、Inc-v4、IncRes-v2、Res-101 模型, 结果如图 2 所示。不同于前文提到的其他基于梯度的攻击, RLI-FGSM 的步长是自适应变化的, 因此 RLI-FGSM 需要通过实验选择一个合适的初始步长。本节中初始步长的选择范围从  $\epsilon/T$  到  $15*\epsilon/T$ , 间隔为  $\epsilon/T$ 。如图二所示可以很容易地观察到, 白盒攻击在初始步长  $\alpha = 3 * \epsilon / T$  时攻击成功率已经接近 100%。而三条黑盒攻击的曲线在初始步长  $\alpha = 10 * \epsilon / T$  到  $\alpha = 12 * \epsilon / T$  时具有最高的攻击成功率, 在  $\alpha > 12 * \epsilon / T$  时攻击成功率不再继续增加。因此本文选择  $\alpha = 10 * \epsilon / T$  作为初始步长, 又因迭代次数  $T = 10$ , 最后  $\alpha = \epsilon$ 。

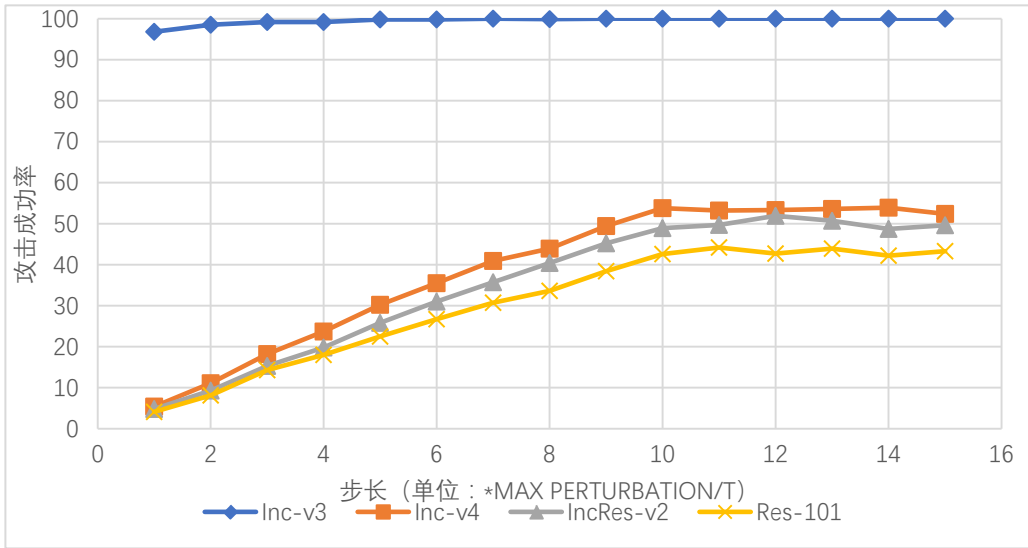


图 2 不同初始步长设置下以 Inc-v3 为源模型攻击 Inc-v3、Inc-v4、IncRes-v2、Res-101 模型的攻击成功率

### 3.4 攻击单个模型

本节介绍在单个模型上的对抗攻击实验。本文只在 Inc-v3、Inc-v4、IncRes-v2、Res-101 模型上生成对抗样本, 并在正常训练和对抗训练的 7 个模型上测试它们。表 2 给出了 Inc-v3、Inc-v4、IncRes-v2、Res-101 这 4 个模型生成的对抗样本对所有 7 个模型的攻击成功率, 其中\*表示白盒攻击, 无\*则表示黑盒攻击。从表 2 可以看出, 本文提出的方法 RLI-FGSM 能够有效生成迁移性较高的对抗样本。RLI-FGSM 应用在 Inc-v3、Inc-v4、IncRes-v2、Res-101 这 4 个模型生成的对抗样本在正常训练的 4 个模型的平均攻击成功率分别为 62.4%、65.2%、66.2%、67.9%, 在对抗训练的 3 个模型的平均攻击成功率分别为 11.1%、13.4%、16.0%、19.0%。而且, 与已有的基于优化方法 (即 I-FGSM、MI-FGSM 和 NI-FGSM) 比较, RLI-FGSM 具有更高的攻击成功率。具体地说, RLI-FGSM 在本实验中的平均攻击成功率分别比 I-FGSM、UAP、MI-FGSM 和 NI-FGSM 高 22.7%、11.6%、6.1%、1.1%。此外, 本文的集成攻击方法 RLI-CI-FGSM 在正常训练的模型上比 RLI-FGSM 提升了 10%~20%的成功率, 在对抗训练的模型上比 RLI-FGSM 提升了 0%~20%的成功率。

表 2 在单模型设置下不同攻击方法在 Inc-v3、Inc-v4、IncRes-v2、Res-101 模型生成的对抗样本对 7 种模型的对抗攻击成功率(%), \*表示白盒攻击

模型	攻击	Inc-v3	Inc-v4	IncRes-v2	Res-101	Inc - v3 <sub>ens3</sub>	Inc - v3 <sub>ens4</sub>	IncRes - v2 <sub>ens</sub>
	I-FGSM	100.0*	22.8	19.9	16.2	7.5	6.4	4.1

Inc-v3	UAP	78.2	22.4	20.3	17.9	9.5	9.8	4.9
	MI-FGSM	100.0*	46.4	42.6	36.0	14.6	12.5	6.1
	NI-FGSM	100.0*	52.0	48.9	39.8	12.7	12.7	6.4
	RLI-FGSM	100.0*	55.1	51.1	43.6	13.8	13.2	6.4
	RLI-CI-FGSM	<b>99.9*</b>	<b>72.3</b>	<b>71.0</b>	<b>61.8</b>	<b>24.4</b>	<b>24.2</b>	<b>12.1</b>
Inc-v4	I-FGSM	22.0	99.9*	13.2	10.9	3.2	3.0	1.7
	UAP	23.4	73.3	22.5	13.1	5.3	5.1	3.3
	MI-FGSM	51.1	99.9*	39.4	33.7	11.2	10.7	5.3
	NI-FGSM	62.8	<b>100.0*</b>	51.4	46.2	15.7	13.0	7.3
	RLI-FGSM	63.0	99.4*	53.2	45.3	16.9	15.0	8.2
	RLI-CI-FGSM	<b>79.5</b>	99.3*	<b>71.2</b>	<b>63.8</b>	<b>30.4</b>	<b>28.4</b>	<b>16.9</b>
IncRes-v2	I-FGSM	22.2	17.7	97.9*	12.6	4.6	3.7	2.5
	UAP	24.0	19.3	72.5	17.6	12.8	12.8	5.2
	MI-FGSM	53.5	45.9	98.4*	37.8	15.3	13.0	8.8
	NI-FGSM	62.0	54.2	<b>98.6*</b>	45.7	18.1	14.8	10.1
	RLI-FGSM	62.8	55.8	97.2*	48.1	20.5	16.5	11.0
	RLI-CI-FGSM	<b>75.4</b>	<b>70.8</b>	96.4*	<b>65.5</b>	<b>37.9</b>	<b>31.3</b>	<b>25.3</b>
Res-101	I-FGSM	26.7	22.7	21.2	98.6*	9.3	8.9	6.2
	UAP	29.0	23.4	21.8	75.3	11.8	10.2	5.5
	MI-FGSM	53.6	48.9	44.7	98.5*	22.1	21.7	12.9
	NI-FGSM	64.5	58.2	56.2	99.4*	23.4	21.1	11.4
	RLI-FGSM	63.7	57.8	56.9	99.2*	24.0	20.9	12.2
	RLI-CI-FGSM	<b>74.5</b>	<b>69.4</b>	<b>68.4</b>	<b>99.5*</b>	<b>36.3</b>	<b>21.7</b>	<b>36.4</b>

本文还将 RLI-CI-FGSM 分别与 TIM 和 DIM 集成在一起, 并将集成后的方法 RLI-CI-TIM 和 RLI-CI-DIM 分别与 TIM 和 DIM 进行比较。比较结果如表 3 和表 4 所示。由表 3 和表 4 可知, RLI-CI-FGSM 的集成对攻击成功率的提升效果显著。具体地说, 在 Inc-v3、Inc-v4、IncRes-v2、Res-101 模型上, RLI-CI-TIM 生成的对抗样本比 TIM 分别提升了平均 20.6%、19.1%、16.6%、15.6% 的成功率。RLI-CI-DIM 生成的对抗样本比 DIM 分别提升了平均 29.0%、32.2%、29.9%、22.8% 的成功率。

由以上可知, RLI-CI-FGSM 能够有效地提高对抗样本的可迁移性。

**表 3** 在单模型设置下针对 7 种模型进行对抗攻击的成功率(%), 对抗样本分别使用 TIM 和 RLI-CI-TIM 在 Inc-v3、Inc-v4、IncRes-v2 模型上制作, \*表示白盒攻击。

模型	攻击	Inc-v3	Inc-v4	IncRes-v2	Res-101	Inc - v3 <sub>ens3</sub>	Inc - v3 <sub>ens4</sub>	IncRes - v2 <sub>ens</sub>
Inc-v3	TIM	100.0*	47.8	42.8	39.5	24.0	21.4	12.9
	RLI-CI-TIM	<b>100*</b>	<b>73</b>	<b>71.3</b>	<b>61.8</b>	<b>40.7</b>	<b>38.6</b>	<b>26.3</b>
Inc-v4	TIM	58.5	<b>99.6*</b>	47.5	43.2	25.7	23.3	17.3
	RLI-CI-TIM	<b>76.8</b>	98.6*	<b>70.0</b>	<b>62.8</b>	<b>44.9</b>	<b>42.2</b>	<b>33.3</b>
IncRes-v2	TIM	62.0	56.2	<b>97.5*</b>	51.3	32.8	27.9	21.9
	RLI-CI-TIM	<b>75.6</b>	<b>71.1</b>	95.9*	<b>67.2</b>	<b>51.9</b>	<b>43.9</b>	<b>41.7</b>
Res-101	TIM	59.0	53.6	51.8	99.3*	36.8	32.2	23.5
	RLI-CI-TIM	<b>73.5</b>	<b>69.3</b>	<b>69.3</b>	<b>99.5*</b>	<b>51.4</b>	<b>46.8</b>	<b>37.5</b>

**表 4** 在单模型设置下针对 7 种模型进行对抗攻击的成功率(%), 对抗样本分别使用 DIM 和 RLI-CI-DIM 在 Inc-v3、Inc-v4、IncRes-v2、Res-101 模型上制作, \*表示白盒攻击。

模型	攻击	Inc-v3	Inc-v4	IncRes-v2	Res-101	Inc - v3 <sub>ens3</sub>	Inc - v3 <sub>ens4</sub>	IncRes - v2 <sub>ens</sub>
Inc-v3	DIM	<b>99.9*</b>	35.5	27.8	21.4	5.5	5.2	2.8
	RLI-CI-DIM	99.2*	<b>73.5</b>	<b>71.4</b>	<b>63.2</b>	<b>25.9</b>	<b>25.4</b>	<b>12.5</b>
Inc-v4	DIM	43.3	<b>99.7*</b>	28.9	23.1	5.9	5.5	3.2
	RLI-CI-DIM	<b>82.3</b>	98.6*	<b>74.7</b>	<b>66.7</b>	<b>31.3</b>	<b>30.1</b>	<b>18.2</b>
IncRes-v2	DIM	46.5	40.5	95.8*	28.6	8.2	6.6	4.8
	RLI-CI-DIM	<b>77.3</b>	<b>72.6</b>	<b>95.9*</b>	<b>66.7</b>	<b>38.9</b>	<b>32.8</b>	<b>26.5</b>
Res-101	DIM	51.2	43.7	41.9	98.6*	15.7	14.0	8.9
	RLI-CI-DIM	<b>75.9</b>	<b>70.3</b>	<b>71.7</b>	<b>98.7*</b>	<b>38.0</b>	<b>33.9</b>	<b>22.7</b>

3.5 攻击模型集合

虽然表 2 的结果表明, RLI-FGSM 和颜色变换方法 CIM 可以显著提高对抗样本的可迁移性, 但黑盒设置下在攻击对抗训练网络时, 它们仍然相对较弱。因此, 本节考虑同时攻击多个模型来提高所有方法的攻击成功率, 进一步展示我们方法的性能。攻击模型集合能提高成功率的原因是——一个样本是多个模型的对抗样本, 比是单个模型的对抗样本更困难, 对样本的要求更高, 那么它成功转移到另一个黑盒模型上的概率就越高。

本节使用 TIM, RLI-CI-TIM, DIM, RLI-CI-DIM, TI-DIM 和 RLI-CI-TI-DIM 来攻击正常训练模型 Inc-v3, Inc-v4, IncRes-v2、Res-101 的集合。如表 5 所示, 本文提出的方法提高了所有基线攻击的攻击成功率, RLI-CI-TI-DIM 攻击能够以平均 95.6%的成功率欺骗正常训练模型, 能够以平均 72.0%的成功率欺骗对抗训练模型。

表 5 在多模型设置下针对 7 种模型进行对抗攻击的成功率(%), 对抗样本分别使用 RLI-FGSM,RLI-CI-FGSM,TIM, RLI-CI-TIM, DIM, RLI-CI-DIM, TI-DIM, RLI-CI-TI- DIM 在 Inc-v3、Inc-v4、IncRes-v2、Res-101 模型的集成模型上制作, \*表示白盒攻击。

攻击	Inc-v3*	Inc-v4*	IncRes-v2*	Res-101*	Inc-v3 <sub>ens3</sub>	Inc-v3 <sub>ens4</sub>	IncRes-v2 <sub>ens</sub>
RLI-FGSM	99.9	98.3	95.0	<b>99.9</b>	37.4	32.2	21.3
RLI-CI-FGSM	<b>99.9</b>	<b>99.3</b>	<b>96.4</b>	99.5	<b>53.1</b>	<b>46.7</b>	<b>40.4</b>
TIM	69.9	67.9	64.1	51.7	36.3	35.0	30.4
RLI-CI-TIM	<b>98.9</b>	<b>95.9</b>	<b>93.2</b>	<b>95.9</b>	<b>74.1</b>	<b>69.2</b>	<b>60.7</b>
DIM	99.9	98.3	94.6	99.9	31.1	27.1	19.0
RLI-CI-DIM	<b>98.9</b>	<b>95.6</b>	<b>94.7</b>	<b>96.2</b>	<b>59.1</b>	<b>54.0</b>	<b>39.9</b>
TI-DIM	<b>99.7</b>	<b>97.0</b>	93.1	<b>99.5</b>	55.0	48.6	37.3
RLI-CI-TI- DIM	98.7	95.2	<b>93.3</b>	95.4	<b>75.7</b>	<b>70.7</b>	<b>69.8</b>

4 总结

本论文提出了一种新的基于转移的对抗攻击方法 RLI-CI-FGSM。RLI-CI-FGSM 由基于 RAdam 迭代快速梯度符号法(RLI-FGSM)和颜色变换不变攻击法(CIM)共同组成。RLI-FGSM 的目标是在基于梯度的攻击中采用 RAdam 优化算法, 并利用目标函数的二阶导信息, 与迭代快速梯度符号法相结合。CIM 的目标是利用模型的颜色变换不变特性实现模型扩充。本文利用上述算法提高了对抗样本的可迁移性。此外, RLI-CI-FGSM 通过与 TIM 和 DIM 攻击的集成, 可以进一步构造一个更强大的基于转移的攻击, 提高对抗样本的可迁移性。实验结果表明, 本文的方法不仅在正常训练的模型上产生更高的成功率, 而且还打破了其他强大的对抗网络的防御机制。

References:

[1] Redmon J, Farhadi A. Yolov3: An incremental improvement. arXiv preprint arXiv:1804.02767, 2018.

[2] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556, 2014.

[3] Ilya Sutskever, Oriol Vinyals, Quoc V. Le:Sequence to Sequence Learning with Neural Networks. In: Zoubin G, ed. Advances in Neural Information Processing Systems 27.Montreal: NIPS 2014: 3104-3112

[4] W. Xiong, J. Droppo, X. Huang, F. Seide, M. Seltzer, A. Stolcke, D. Yu, and G. Zweig. Achieving human parity in conversational speech recognition. arXiv preprint arXiv:1610.05256, 2016.

[5] Z. Zhang, J. Geiger, J. Pohjalainen, A. E.-D. Mousa, W. Jin, and B. Schuller. Deep learning for environmentally robust speech recognition: An overview of recent developments. ACM Transactions on Intelligent Systems and Technology (TIST), 9(5):49, 2018. [doi:10.1145/3178115]

[6] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus. Intriguing properties of neural networks. In: Yoshua Bengio, ed. 2nd International Conference on Learning Representations. Banff: ICLR 2014.

- [7] Lu, Y., et al. Enhancing Cross-Task Black-Box Transferability of Adversarial Examples With Dispersion Reduction. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: Computer Vision Foundation / {IEEE}, 2020(CVPR) IEEE, 2020. 937-946.
- [8] Ren, K., Wang, Q., Wang, C., Qin, Z., Lin, X.: The security of autonomous driving: Threats, defenses, and future directions. *Proceedings of the IEEE* 108(2), 357-372, 2019.
- [9] Ibitoye, O., Abou-Khamis, R., Matrawy, A., Shafiq, M.O.: The threat of adversarial attacks on machine learning in network security-a survey. *arXiv preprint arXiv:1911.02621*, 2019.
- [10] Pan WW, Wang XY, Song ML, Chen C. Survey on generating adversarial examples. *Ruan Jian Xue Bao/Journal of Software*, 2020,31(1):67-81 (in Chinese). <http://www.jos.org.cn/1000-9825/5884.htm>
- [11] Alexey Kurakin, Ian J. Goodfellow, Samy Bengi. Adversarial examples in the physical world. In: 5th International Conference on Learning Representations. Toulon: OpenReview.net, 2017.
- [12] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In: 6th International Conference on Learning Representations. Vancouver: OpenReview.net, 2018.
- [13] Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li. Boosting adversarial attacks with momentum. In: 2018 {IEEE} Conference on Computer Vision and Pattern Recognition. Salt Lake City: Computer Vision Foundation / {IEEE} Computer Society, 2019. 9185--9193
- [14] Florian Tramer, Alexey Kurakin, Nicolas Papernot, Dan Boneh, and Patrick McDaniel. Ensemble adversarial training: Attacks and defenses. In: 6th International Conference on Learning Representations. Vancouver: OpenReview.net, 2018.
- [15] Cihang Xie, Jianyu Wang, Zhishuai Zhang, Zhou Ren, and Alan Yuille. Mitigating adversarial effects through randomization. In: 6th International Conference on Learning Representations. Vancouver: OpenReview.net, 2018.
- [16] Fangzhou Liao, Ming Liang, Yinpeng Dong, Tianyu Pang, Xiaolin Hu, and Jun Zhu. Defense against adversarial attacks using high-level representation guided denoiser. In: 2018 {IEEE} Conference on Computer Vision and Pattern Recognition. Salt Lake City: Computer Vision Foundation / {IEEE} Computer Society, 2018. 1778-1787.
- [17] Cihang Xie, Zhishuai Zhang, Yuyin Zhou, Song Bai, Jianyu Wang, Zhou Ren, and Alan L Yuille. Improving transferability of adversarial examples with input diversity. In: {IEEE} Conference on Computer Vision and Pattern Recognition. Long Beach: Computer Vision Foundation / {IEEE}, 2019. 2730-2739.
- [18] Yinpeng Dong, Tianyu Pang, Hang Su, and Jun Zhu. Evading defenses to transferable adversarial examples by translation-invariant attacks. In: {IEEE} Conference on Computer Vision and Pattern Recognition. Long Beach: Computer Vision Foundation / {IEEE}, 2019. 4312-4321.
- [19] Hang J , Han K , Chen H , et al. Ensemble Adversarial Black-Box Attacks against Deep Learning System. *Pattern Recognition*, 2019, 101:107184.
- [20] Jiadong Lin, Chuanbiao Song, Kun He, Liwei Wang, John E. Hopcroft: Nesterov Accelerated Gradient and Scale Invariance for Adversarial Attacks. In: 8th International Conference on Learning Representations. Addis Ababa: OpenReview.net, 2020.
- [21] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In: Peter L. Bartlett, ed. 26th Annual Conference on Neural Information Processing Systems. Lake Tahoe, 2012. 1106-1114
- [22] Andrew G. Howard: Some Improvements on Deep Convolutional Neural Network Based Image Classification. In: Yoshua Bengio, ed. 2nd International Conference on Learning Representations. Banff, 2014.
- [23] Pengguang Chen, Shu Liu, Hengshuang Zhao, Jiaya Jia: GriMask Data Augmentation. *CoRR abs/2001.04086* (2020).
- [24] Pengfei Xie, Linyuan Wang, Ruoxi Qin, Kai Qiao, Shuhao Shi, Guoen Hu, Bin Yan: Improving the Transferability of Adversarial Examples with New Iteration Framework and Input Dropout. *CoRR abs/2106.01617* (2021) .
- [25] Ian J. Goodfellow, Jonathon Shlens, Christian Szegedy: Explaining and Harnessing Adversarial Examples. In: Yoshua Bengio, ed. 3rd International Conference on Learning Representations. San Diego, 2015.
- [26] Zou, J., Pan, Z., Qiu, J., Liu, X., Rui, T., Li, W.: Improving the Transferability of Adversarial Examples with Resized-Diverse-Inputs, Diversity-Ensemble and Region Fitting. In: *European Conference on Computer Vision*. pp. 563{579. Springer (2020).

- [27] Moosavi-Dezfooli, S. M. , et al. Universal adversarial perturbations. In: 2017 {IEEE} Conference on Computer Vision and Pattern Recognition. Honolulu: {IEEE} Computer Society,2017. 86-94.
- [28] Diederik P. Kingma, Jimmy Ba:Adam: A Method for Stochastic Optimization. In: Yoshua Bengio, ed. 3rd International Conference on Learning Representations. San Diego,2015.
- [29] Liyuan Liu, Haoming Jiang, Pengcheng He, Weizhu Chen, Xiaodong Liu, Jianfeng Gao, Jiawei Han:On the Variance of the Adaptive Learning Rate and Beyond. In : 8th International Conference on Learning Representations. Addis Ababa: OpenReview.net,2020.
- [30] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna.Rethinking the inception architecture for computer vision. In : 2016 {IEEE} Conference on Computer Vision and Pattern Recognition. Las Vegas: {IEEE} Computer Society, 2016.2818-2826.
- [31] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In: Satinder P. Singh, ed. Proceedings of the Thirty-First {AAAI} Conference on Artificial Intelligence. San Francisco: {AAAI} Press,2017. 4278-4284.
- [32] K. He, X. Zhang, S. Ren, and J. Sun. Identity mappings in deep residual networks. In: Bastian Leibe, ed. 14th European Conference. Amsterdam: Springer,2016. 630-645.

#### 附中文参考文献:

- [10] 潘文雯,王新宇,宋明黎,等. 对抗样本生成技术综述[J]. 软件学报,2020,31(1):67-81. DOI:10.13328/j.cnki.jos.005884.