结合自适应步长策略和数据增强机制提升对抗攻击迁移性

鲍 蕾[1],陶 蔚[2],陶 卿[1*]
(1. 中国人民解放军陆军炮兵防空兵学院信息工程系,安徽合肥 230031;2. 中国人民解放军军事科学院,北京 100091)

摘 要: 深度神经网络具有脆弱性,容易被精心设计的对抗样本攻击. 梯度攻击方法在白盒模型上攻击成功率较高,但在黑盒模型上的迁移性较弱. 基于Heavy-ball型动量和Nesterov型动量的梯度攻击方法由于在更新方向上考虑了历史梯度信息,提升了对抗样本的迁移性. 为了进一步使用历史梯度信息,本文针对收敛性更好的Nesterov型动量方法,使用自适应步长策略代替目前广泛使用的固定步长,提出了一种方向和步长均使用历史梯度信息的迭代快速梯度方法(Nesterov and Adaptive-learning-rate based Iterative Fast Gradient Method,NAI-FGM). 此外,本文还提出了一种线性变换不变性(Linear-transformation Invariant Method,LIM)的数据增强方法. 实验结果证实了NAI-FGM攻击方法和LIM数据增强策略相对于同类型方法均具有更高的黑盒攻击成功率. 组合NAI-FGM方法和LIM策略生成对抗样本,在常规训练模型上的平均黑盒攻击成功率达到87.8%,在对抗训练模型上的平均黑盒攻击成功率达到57.5%,在防御模型上的平均黑盒攻击成功率达到67.2%,均超过现有最高水平.

关键词: 对抗样本;迁移性;Nesterov型动量;自适应步长;线性变换不变性
基金项目: 国家自然科学基金(No.62076252,No.62106281)
中图分类号: TP391 **文献标识码:** A **文章编号:** 0372-2112(2024)01-0157-13
电子学报URL:http://www.ejournal.org.cn **DOI:**10.12263/DZXB.20220737

Boosting Adversarial Transferability Through Adaptive-Learning-Rate with Data Augmentation Mechanism

BAO Lei[1], TAO Wei[2], TAO Qing[1*]
(1. *Department of Information Engineering, PLA Army Academy of Artillery and Air Defense, Hefei, Anhui* 230031, *China*;
2. *PLA Academy of Military Science, Beijing* 100091, *China*)

Abstract: Deep neural networks are vulnerable to adversarial examples. Gradient based attacks exhibit weak transferability in the black-box setting, though perform well in the white-box situation. The Heavy-ball momentum and Nesterov momentum based attacks boost the transferability for the consideration of gradient history. To further take advantage of the gradient history information, we propose an iterative fast gradient method (NAI-FGM) on Nesterov momentum for its faster convergence property. As the commonly used constant step size is replaced by adaptive step size, NAI-FGM makes use of gradient history information both in step size and gradient direction. Additionally, we propose a new input transformation mechanism named linear-transformation invariant method (LIM). Experimental results demonstrate that NAI-FGM and LIM perform better than the same kind attacks. Besides, the integrated method LI-NAI-FGM could achieve an average rate of 87.8% on commonly trained models, 57.5% on adversarial trained models, 67.2% on defense models, which are higher than the state-of-the-art results.

Key words: adversarial examples; transferability; Nesterov momentum; adaptive step size; linear-transformation invariant
Foundation Item(s): National Natural Science Foundation of China (No.62076252, No.62106281)

收稿日期:2022-06-27;修回日期:2022-09-21;责任编辑:朱梅玉
*通讯作者:陶卿

1 引言

深度神经网络(Deep Neural Network,DNN)在图像分割[1]、目标检测[2]等计算机视觉任务中都取得了出色的表现.与此同时,DNN存在严重的安全隐患,即攻击者可以通过给干净样本添加精心设计的噪声而轻而易举地欺骗DNN模型,使其产生误判,并且通常不会被人类发现[3].为了提高DNN的防御能力,使其现实应用更加可靠,对抗攻击的研究受到越来越多的关注[4~6].

根据攻击者掌握的信息多少,可以简单地将对抗攻击方法分为两类[7]:白盒攻击和黑盒攻击.在白盒攻击模式下,模型的结构和参数均已知,研究者们通过最大化损失来生成对抗样本.因此,白盒攻击是一种典型的优化问题,可以采用优化算法进行求解.最常见的优化算法为梯度法.快速梯度符号方法(Fast Gradient Sign Method,FGSM)[3]是最早被提出的梯度攻击方法.FGSM方法简单有效,在图像攻击领域具有非常重要的地位.后续很多工作都基于这个方法展开.迭代快速梯度符号方法(Iterative Fast Gradient Sign Method,I-FGSM)[8]在FGSM方法的基础上,采用了多步攻击方式,提高了白盒攻击成功率.但是当目标模型的结构和参数未知时,无论FGSM还是I-FGSM方法在白盒模型上生成的对抗样本迁移性均较差,即攻击黑盒模型的成功率较低.

为了解决该问题,研究者们通过寻找更好的梯度计算方式来探索提高对抗样本迁移性的方法.研究表明,添加动量运算调整方向和自适应机制调整步长是累积历史梯度信息、提升梯度优化算法性能的两种主要技巧[9~11].对此,部分研究者在进行梯度计算时引入了动量算法,具有代表性的方法有基于动量的迭代快速梯度符号方法(Momentum Iterative Fast Gradient Sign Method,MI-FGSM)[12]和基于Nesterov的迭代快速梯度符号方法(Nesterov Iterative Fast Gradient Sign Method,NI-FGSM)[13].MI-FGSM方法首次引入了动量项,使用的是Heavy-ball型动量方法[14].通过累积之前梯度指数级衰减的移动平均,抑制梯度更新时上下震荡的情况,MI-FGSM方法能够更好地略过局部极小点,提高对抗样本的迁移性.不同于MI-FGSM,NI-FGSM方法使用了Nesterov加速梯度(Nesterov Accelerated Gradient,NAG)算法[15].由于利用了二阶导的信息,NAG算法填补了"一阶梯度方法在处理光滑凸函数只有$O(1/t)$收敛速率"与Nemirovski和Yudin所证明的"任何一阶优化算法都不可能得到比$O(1/t^2)$更快的收敛速率"之间的间隙[16].它可以将多种形式一阶梯度方法的收敛速率提升一个数量级,实现了凸光滑优化的最优收敛速率.NAG算法的提出在优化领域具有里程碑的意义,也成为动量优化方法发展的标志性事件.该动量算法在计算梯度之前,比Heavy-ball型动量多了一个本次梯度相对于上次梯度的变化量,即Nesterov项,本质上是对优化目标二阶导的近似.由于利用了二阶导的信息,采用了NAG算法的NI-FGSM方法相对于MI-FGSM能够更快、更精准地找到致使模型判断错误的对抗样本.无论是MI-FGSM还是NI-FGSM方法,虽然使用了多步攻击的方式,但是在迭代更新对抗样本时均使用了固定步长,没有从学习率的角度考虑历史梯度信息.

针对上述问题,本文采用NAG算法稳定梯度方向时,使用自适应步长代替固定步长,从学习率的角度累积历史梯度信息,提高黑盒模型的攻击成功率,提出了基于Nesterov和自适应步长的迭代快速梯度方法(Nesterov and Adaptive-learning-rate based Iterative Fast Gradient Method,NAI-FGM).图1所示为两组使用NAI-FGM方法攻击Inception-v3(Inc-v3)[17]模型生成的对抗样本.图1中噪声为差值图像,用于显示原始图像和对抗样本之间的差异.同时,我们给出了原始图像和对抗样本分别使用Inc-v3模型预测的标签及其概率.值得指出的是,文献[18]在生成对抗样本时,将AdaDelta算法和Nesterov项进行了结合,形成了基于AdaDelta-Nesterov的迭代快速梯度符号方法ADNI-FGSM;并且在ADNI-FGSM基础上,进一步提出了基于Adam-Nesterov方法的迭代快速梯度符号方法ANI-FGSM.通过组合使用多种数据增强技术,该工作在对抗训练模型和防御模型上均取得了非常好的黑盒攻击成功率.无论是ADNI-FGSM还是ANI-FGSM方法,在生成对抗样本时,均是在上一累积梯度上引入自适应步长;然后在计算NAG算法时,使用自适应步长代替传统的固定步长,预估梯度方向.因此,文献[18]本质上实现了NAG算法的固定步长到自适应步长的转换.不同于ANI-FGSM方法,本文提出的NAI-FGM首先在前一步的累积梯度上计算Nesterov项预估梯度方向,然后将指数衰减估计作用到更新后的累积梯度上,使用自适应机制优化对抗扰动,即在每次迭代过程中,为对抗样本不同位置点赋予与历史梯度相关的差异性步长,实现从动量和自适应步长两个角度综合考虑历史梯度信息更新对抗样本,提高黑盒攻击成功率.

除了从学习率的角度探索一种更好的梯度优化算法,我们还考虑组合数据增强技术进一步提高对抗样本的迁移性.对抗样本的生成作为一种优化问题进行求解,白盒攻击成功率高、黑盒攻击成功率低是典型的过拟合现象[19,20].数据增强是机器学习中一种经典且有效的防止模型过拟合的技术[21~23].观察发现,人类对样本进行判断时,样本整体变亮或者变暗,只要不是过亮或者过暗而超过人眼的感知范围,对结论就不会造成影响.即样本在像素空间同步调整时,存在线性变换

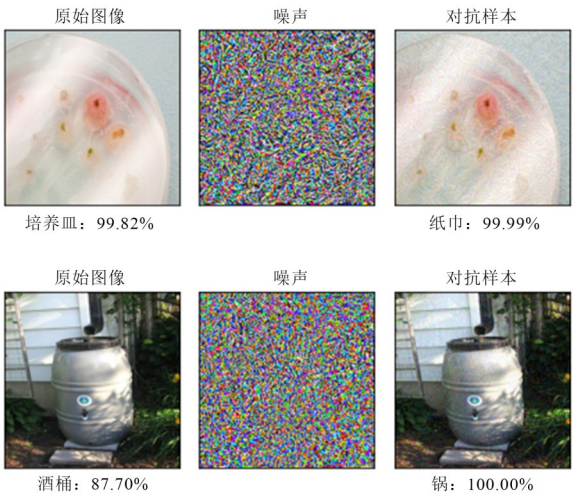

图1　NAI-FGM生成的对抗样本示例

不变性的特征. 模仿人类的这种视觉注意机制,我们将该特性应用到对抗攻击中,提出了线性变换不变性(Linear-transformation Invariant Method, LIM)的数据增强方法. 具体地,在每次迭代攻击时,对样本进行线性变换,以提高样本多样性来优化对抗扰动,提升对抗样本的迁移性能.

本文的主要贡献如下.

(1)提出了一种基于Nesterov和自适应步长的迭代快速梯度方法NAI-FGM,从优化梯度计算方式上提升黑盒攻击成功率. 该方法在NAG算法的基础上引入了自适应策略,既保持了从动量角度稳定梯度方向,又通过将固定步长转换为自适应步长,从学习率的角度将历史梯度信息运用到对抗样本的生成上,提高了迁移性.

(2)提出了一种线性变换不变性策略LIM,从组合数据增强技术上提升黑盒攻击成功率. 该策略在每次迭代过程中,通过对样本进行线性变换提高多样性来优化对抗扰动,提升对抗样本的黑盒攻击成功率.

(3)实验验证了本文方法的有效性. NAI-FGM方法相对于已有的梯度攻击方法具有更高的黑盒攻击成功率;LIM策略相对于常用的几种数据增强方法对样本迁移性的提升更加明显;将NAI-FGM方法和LIM策略进行组合,能够生成黑盒攻击成功率更高的对抗样本.

2　相关工作

给定训练好的分类器$f(\boldsymbol{x}):\boldsymbol{x}\in\boldsymbol{X}\to y\in Y$. 当输入$\boldsymbol{x}$为干净样本时,输出的$y$为正确的标签,即$f(\boldsymbol{x})=y$. 对抗攻击是在干净样本$\boldsymbol{x}$邻域找到一个对抗样本$\boldsymbol{x}^{\text{adv}}$,使分类器$f$输出错误的标签. 根据是否设定目标标签,可以将攻击方法分为无目标攻击和有目标攻击两类[24]. 无目标攻击方法是对干净样本添加噪声,获取的对抗样本能够误导分类器输出错误标签,即$f(\boldsymbol{x}^{\text{adv}})\neq y$. 有目标攻击方法则指定了错误标签,即满足$f(\boldsymbol{x}^{\text{adv}})=y^*$. 其中,$y^*$为攻击者指定的标签,且$y^*\neq y$. 为了让人类难以察觉在干净样本上添加了噪声,通常将干净样本$\boldsymbol{x}$和对抗样本$\boldsymbol{x}^{\text{adv}}$之间的$L_p$距离限制在足够小的扰动范围$\varepsilon$内,即$\|\boldsymbol{x}-\boldsymbol{x}^{\text{adv}}\|_p\leqslant\varepsilon$. 其中,$p$的取值可以设定为$0,1,2,\cdots,\infty$. 本文主要关注$L_\infty$范数下的无目标攻击方法.

2.1　梯度攻击方法

2.1.1　快速梯度符号方法

Goodfellow等人[3]通过最大化损失对输入样本的梯度来生成对抗样本,提出了快速梯度符号方法FGSM. 更新规则描述如下:

$$\boldsymbol{x}^{\text{adv}}=\boldsymbol{x}+\varepsilon\cdot\text{sign}\left(\nabla J\left(\boldsymbol{x},y\right)\right)\tag{1}$$

其中,$\boldsymbol{x}$为输入样本;y为标签;ε为扰动范围;sign(·)为符号函数;J为损失函数;$\nabla J(\boldsymbol{x},y)$为损失函数对输入$\boldsymbol{x}$的梯度;$\boldsymbol{x}^{\text{adv}}$为生成的对抗样本. FGSM方法是一种最为基础的梯度攻击方法,其生成的对抗样本满足L_∞范数约束,即$\|\boldsymbol{x}-\boldsymbol{x}^{\text{adv}}\|_\infty\leqslant\varepsilon$.

2.1.2　迭代快速梯度符号方法

Alexey等人[8]在FGSM的基础上,通过迭代若干次来生成对抗样本,提出了迭代快速梯度符号方法I-FGSM. 更新规则描述如下:

$$\boldsymbol{x}_0^{\text{adv}}=\boldsymbol{x}\tag{2}$$

$$\boldsymbol{x}_{t+1}^{\text{adv}}=\text{Clip}_{\boldsymbol{x},\varepsilon}\left\{\boldsymbol{x}_t^{\text{adv}}+\alpha\cdot\text{sign}\left(\nabla J\left(\boldsymbol{x}_t^{\text{adv}},y\right)\right)\right\}\tag{3}$$

其中,$\alpha=\varepsilon/T$为每次迭代时使用梯度符号对样本更新的步长;T为迭代次数;$\text{Clip}_{\boldsymbol{x},\varepsilon}\{\cdot\}$为裁剪函数,能够保证生成的对抗样本符合$L_\infty$范数约束. I-FGSM方法相对于单步攻击方法具有更高的白盒攻击成功率,但是迁移性相对于FGSM方法稍差.

2.1.3　基于Momentum的迭代快速梯度符号方法

Dong等人[12]首次将动量方法引入对抗样本的生成过程,提出了基于Heavy-ball型动量的迭代快速梯度符号方法MI-FGSM. 更新规则描述如下:

$$\boldsymbol{x}_0^{\text{adv}}=\boldsymbol{x},\boldsymbol{g}_0=0\tag{4}$$

$$\boldsymbol{g}_{t+1}=\mu\boldsymbol{g}_t+\frac{\nabla J(\boldsymbol{x}_t^{\text{adv}},y)}{\left\|\nabla J(\boldsymbol{x}_t^{\text{adv}},y)\right\|_1}\tag{5}$$

$$\boldsymbol{x}_{t+1}^{\text{adv}}=\text{Clip}_{\boldsymbol{x},\varepsilon}\left\{\boldsymbol{x}_t^{\text{adv}}+\alpha\cdot\text{sign}\left(\boldsymbol{g}_{t+1}\right)\right\}\tag{6}$$

其中,$\boldsymbol{g}_t$为前t次迭代中累加的梯度;μ为动量系数. MI-FGSM方法由于添加了动量项,累积了梯度,能够更好地跳过局部极小点,生成的对抗样本相对于I-FGSM方法具有更高的迁移性.

2.1.4　基于NAG算法的迭代快速梯度符号方法

Lin等人[13]将NAG算法引入对抗样本的生成过程,

提出了基于Nesterov型动量的迭代快速梯度符号方法NI-FGSM. 更新规则描述如下：

$$\boldsymbol{x}_0^{\text{adv}}=\boldsymbol{x},\boldsymbol{g}_0=0 \tag{7}$$

$$\boldsymbol{x}_t^{\text{nes}}=\boldsymbol{x}_t^{\text{adv}}+\alpha\cdot\mu\cdot\boldsymbol{g}_t \tag{8}$$

$$\boldsymbol{g}_{t+1}=\mu\boldsymbol{g}_t+\frac{\nabla J(\boldsymbol{x}_t^{\text{nes}},y)}{\left\|\nabla J(\boldsymbol{x}_t^{\text{nes}},y)\right\|_1} \tag{9}$$

$$\boldsymbol{x}_{t+1}^{\text{adv}}=\text{Clip}_{\boldsymbol{x},\varepsilon}\left\{\boldsymbol{x}_t^{\text{adv}}+\alpha\cdot\text{sign}\left(\boldsymbol{g}_{t+1}\right)\right\} \tag{10}$$

其中，$\boldsymbol{x}_t^{\text{nes}}$为Nesterov项. NAG算法相对于Heavy-ball型动量方法多了一个本次梯度相对上次梯度的变化量，即Nesterov动量项，本质上是对目标函数二阶导的近似. 因此，NAG算法具有更快的收敛速度，也就是说采用了Nesterov项的NI-FGSM方法能够更快地锁定导致分类器误判的对抗样本. 从式(6)和式(10)可以看出，MI-FGSM和NI-FGSM方法都采用了迭代的方式生成对抗样本，但是二者均使用了固定步长α，没有从学习率的角度考虑历史梯度信息.

2.2 数据增强方法

2.2.1 样本多样化方法

Xie等人[25]通过在每次迭代过程中，对输入图像进行多样化转换来优化对抗扰动，主要包括随机尺寸变化和随机零填充，提出了样本多样化方法(Diverse Input Method，DIM). DIM能够与任意基于梯度的攻击方法组合，提高对抗样本的迁移性.

2.2.2 平移不变方法

Dong等人[26]通过在每次迭代过程中，集成多个平移单个像素的样本来提高样本多样性优化对抗扰动，提出了平移不变方法(Translation-Invariant Method，TIM). 为了提高效率，Dong等人[26]将多个样本的集成近似为对梯度信息的一次卷积过程. TIM能够很好地提升对抗样本在具有防御机制的黑盒模型上的攻击性能.

2.2.3 尺度不变方法

Lin等人[13]为了提高对抗样本的迁移性能，在每次迭代过程中，对输入样本的数值以2为倍数降低进行尺度变换，提出了尺度不变方法(Scale-Invariant Method，SIM). 与SIM类似，Wang等人[7]通过在每次迭代中，对输入样本的数值在一定范围内随机平移优化对抗扰动，提出了差值调整方法(Variance Tuning Method，VTM). 无论SIM还是VTM，虽然需要更多的时间和资源生成对抗样本，但是明显提高了对抗样本的白盒攻击成功率，其在黑盒模型上的迁移性能也得到了较大的提升.

3 本文算法

为了在方向和步长上均使用历史梯度信息，本文基于Nesterov动量提出了一种采用自适应步长迭代更新对抗样本的攻击方法NAI-FGM.

NAG算法的梯度更新由以下两个步骤组成：

$$\boldsymbol{v}_t=\gamma\boldsymbol{v}_{t-1}+\eta\nabla J\left(\boldsymbol{\theta}_t-\gamma\boldsymbol{v}_{t-1}\right) \tag{11}$$

$$\boldsymbol{\theta}_{t+1}=\boldsymbol{\theta}_t-\boldsymbol{v}_t \tag{12}$$

其中，$\boldsymbol{\theta}_t$为模型参数；η为学习率；γ为动量系数.

本文提出的NAI-FGM攻击方法首先将NAG算法应用于对抗样本的生成，通过计算Nesterov动量预估梯度方向；然后在迭代更新对抗样本的过程中，采用自适应步长代替人为设定的固定步长，从学习率的角度累积历史梯度信息，提升对抗样本的迁移性能. NAI-FGM攻击方法的更新规则如下：

$$\boldsymbol{x}_t^{\text{nes}}=\boldsymbol{x}_t^{\text{adv}}+\alpha\cdot\mu\cdot\boldsymbol{g}_t \tag{13}$$

$$\hat{\boldsymbol{g}}_t=\frac{\nabla J(\boldsymbol{x}_t^{\text{nes}},y)}{\left\|\nabla J(\boldsymbol{x}_t^{\text{nes}},y)\right\|_1} \tag{14}$$

$$\boldsymbol{m}_t=\beta_1\boldsymbol{m}_{t-1}+(1-\beta_1)\hat{\boldsymbol{g}}_t \tag{15}$$

$$\hat{\boldsymbol{m}}_t=\boldsymbol{m}_t/\left(1-\beta_1^t\right) \tag{16}$$

$$\boldsymbol{V}_t=\beta_2\boldsymbol{V}_{t-1}+(1-\beta_2)\hat{\boldsymbol{g}}_t\odot\hat{\boldsymbol{g}}_t \tag{17}$$

$$\hat{\boldsymbol{V}}_t=\boldsymbol{V}_t/\left(1-\beta_2^t\right) \tag{18}$$

$$\boldsymbol{g}_{t+1}=\frac{\hat{\boldsymbol{m}}_t}{\hat{\boldsymbol{V}}_t^{1/2}+\delta} \tag{19}$$

$$\boldsymbol{x}_{t+1}^{\text{adv}}=\text{Clip}_{\boldsymbol{x},\varepsilon}\left\{\boldsymbol{x}_t^{\text{adv}}+\alpha\cdot\boldsymbol{g}_{t+1}\right\} \tag{20}$$

其中，β_1和β_2为衰减系数；$\odot$为元素积；δ为极小值. 上述更新规则中，式(13)计算了Nesterov项，为NAG算法的核心，实现了在每次迭代过程中在前一步的累积梯度上预估对抗样本的更新方向. 式(15)~(18)计算了自适应步长，为Adam算法的核心，实现了对抗样本的更新由固定步长到自适应步长的转变. 从式(19)和式(20)可以看出，通过将指数衰减一阶矩$\hat{\boldsymbol{m}}_t$和二阶矩$\hat{\boldsymbol{V}}_t$的估计添加到更新过程中，实现了从学习率的角度融合历史梯度信息，更新对抗样本. 本文方法在迭代更新对抗样本时，通过裁剪函数$\text{Clip}_{\boldsymbol{x},\varepsilon}\{\cdot\}$来保证生成的对抗样本符合$L_\infty$范数约束，即$\|\boldsymbol{x}-\boldsymbol{x}^{\text{adv}}\|_\infty\leqslant\varepsilon$.

对比NI-FGSM方法[13]的更新规则，可以发现NI-FGSM计算了Nesterov项，提高了对抗样本的迁移性. 由于该方法在更新对抗样本上采用的是固定步长α，没有从学习率的角度累积梯度信息. 对比ANI-FGSM方法[17]的更新规则，可以发现ANI-FGSM方法首先在上一累积梯度上计算Adam的一阶矩和二阶矩(原文式(28)~(31))，然后将自适应步长作用在Nesterov项上预估梯度方向(原文式(32))，更新对抗样本.

不同于ANI-FGSM方法，本文提出的NAI-FGM攻击方法则是在上一累积梯度上首先计算Nesterov项预估梯度方向(见式(13))，然后在预估方向上计算Adam

的一阶矩和二阶矩(见式(15)~(18)),从学习率的角度融合历史梯度信息,即使用自适应步长更新对抗样本.

除了分别从方向和步长两个角度累积历史梯度外,我们还提出了线性变换不变性策略LIM进行数据增强,进一步提高黑盒攻击成功率. 本文使用了最为简单的线性变换函数实现LIM策略:

$$\boldsymbol{L}_{k,b}(\boldsymbol{x})=k\cdot\boldsymbol{x}+b \tag{21}$$

其中,$\boldsymbol{x}$为样本;k和b为线性变换参数. 组合LIM和NAI-FGM方法,构建了LI-NAI-FGM攻击方法,其详细过程如算法1所示.

算法1　LI-NAI-FGM

输入: 干净样本$\boldsymbol{x}$;标签y;损失函数J;扰动ε;最大迭代次数T;线性变换次数n;线性变换参数k,b

输出: 对抗样本$\boldsymbol{x}^{\text{adv}}$

$\boldsymbol{x}_0^{\text{adv}}=\boldsymbol{x}$; $\boldsymbol{g}_0=0$; $m_0=0$; $V_0=0$; $\alpha=\varepsilon/T$

FOR t=0 TO $T-1$ DO

　计算$\boldsymbol{x}_t^{\text{nes}}=\boldsymbol{x}_t^{\text{adv}}+\alpha\cdot\mu\cdot\boldsymbol{g}_t$

　更新梯度$\hat{\boldsymbol{g}}_t=\nabla J(\boldsymbol{x}_t^{\text{nes}},y)\big/\left\|\nabla J(\boldsymbol{x}_t^{\text{nes}},y)\right\|_1$

　FOR n=1 TO N DO

　　令$k=1/n$且$b\in[-\varepsilon\beta/n,\varepsilon\beta/n]$

　　对$\boldsymbol{x}_t^{\text{nes}}$进行线性变换$\boldsymbol{L}_{k,b}(\boldsymbol{x}_t^{\text{nes}})=k\cdot\boldsymbol{x}_t^{\text{nes}}+b$

　　对梯度进行累加$\hat{\boldsymbol{g}}_t=\hat{\boldsymbol{g}}_t+\nabla J\left(\boldsymbol{L}_{k,b}\left(\boldsymbol{x}_t^{\text{nes}}\right),y\right)$

　END FOR

　计算平均梯度$\hat{\boldsymbol{g}}_t=\hat{\boldsymbol{g}}_t/n$,更新梯度$\hat{\boldsymbol{g}}_t=\hat{\boldsymbol{g}}_t\big/\left\|\hat{\boldsymbol{g}}_t\right\|_1$

　计算一阶矩$\boldsymbol{m}_t=\beta_1\boldsymbol{m}_{t-1}+\left(1-\beta_1\right)\hat{\boldsymbol{g}}_t$

　计算二阶矩$\boldsymbol{V}_t=\beta_2\boldsymbol{V}_{t-1}+\left(1-\beta_2\right)\hat{\boldsymbol{g}}_t\odot\hat{\boldsymbol{g}}_t$

　偏差修正$\hat{\boldsymbol{m}}_t=\boldsymbol{m}_t\big/\left(1-\beta_1^t\right)$和$\hat{\boldsymbol{V}}_t=\boldsymbol{V}_t\big/\left(1-\beta_2^t\right)$

　计算梯度$\boldsymbol{g}_{t+1}=\dfrac{\hat{\boldsymbol{m}}_t}{\hat{\boldsymbol{V}}_t^{1/2}+\delta}$

　更新对抗样本$\boldsymbol{x}_{t+1}^{\text{adv}}=\text{Clip}_{\boldsymbol{x},\varepsilon}\left\{\boldsymbol{x}_t^{\text{adv}}+\alpha\cdot\boldsymbol{g}_{t+1}\right\}$

END FOR

$\boldsymbol{x}^{\text{adv}}=\boldsymbol{x}_T^{\text{adv}}$

4　实验及结果分析

4.1　实验目标

本文实验目标为以下几个方面:

(1)通过对比梯度攻击方法,验证本文提出的NAI-FGM攻击方法的有效性;

(2)通过对比组合攻击方法,验证本文提出的LIM数据增强方法的有效性;

(3)通过对比SI-NI-FGSM和VNI-FGSM方法,验证本文提出的LI-NAI-FGM攻击方法的有效性;

(4)通过对比不同攻击方法获取的对抗样本,验证本文方法的有效性.

4.2　实验设置

4.2.1　数据集

本文实验使用的数据来自ILSVRC2012验证集[27]中随机抽取的1 000张图片,同时也与NI-FGSM[13]方法使用的数据集相同. 用于生成对抗样本的干净样本的尺寸为299×299×3.

4.2.2　模型

本文实验共涉及12个模型. 其中,4个常规训练模型包括:Inc-v3,Inception-v4(Inc-v4),Inception-ResNet-v2(IncRes-v2)[28]和ResNet-v2-101(Res-101)[29],作为白盒模型用于生成对抗样本;3个对抗训练模型包括:Inc-v3$_{ens3}$,Inc-v3$_{ens4}$,IncResv2$_{ens}$[30],以及5个防御模型包括:ComDefend[31],R&P[32],FD[33],Bit-Red[34],JPEG[35],作为黑盒模型用于测试对抗样本的迁移性能.

4.2.3　对抗样本生成方法

为了验证NAI-FGM方法的有效性,本文实验对比了基于动量的攻击方法MI-FGSM[12]和NI-FGSM[13],基于自适应NAG的攻击方法ANI-FGSM[24]. 为了分析自适应策略对黑盒攻击成功率的影响,本文设计并比较了基于Adam的迭代快速梯度方法Adam-FGM和基于AMSGrad的迭代快速梯度方法AMSGrad-FGM. 其中,Adam是一种自适应学习率的算法,不但使用动量引导更新方向,而且可以自适应调整学习率[36]. 针对Adam可能不收敛及可能错过全局最优解的问题,Reddi等人[37]使用过去平均梯度的最大值来更新参数,提出了AMSGrad方法. 为了验证LIM策略的有效性,实验对比了DIM[25],TIM[26],SIM[13]数据增强策略. 为了验证LI-NAI-FGM方法的有效性,实验对比了SI-NI-FGSM[13],VNI-FGSM[7]攻击方法. 为了验证本文方法生成的对抗样本满足最大扰动量的限制,实验对比了MI-FGSM,NI-FGSM,ANI-FGSM,SI-NI-FGSM,VNI-FGSM攻击方法.

4.2.4　超参设置

实验中参与对比的已有的攻击方法,超参均采用默认设置. 本文提出的NAI-FGM,LI-NAI-FGM方法,以及参与对比的Adam-FGM,AMSGrad-FGM方法,超参的设置均参照了文献[13],即最大扰动量ε为16,迭代次数T为10,步长α为ε/T,衰减因子μ为1. 计算自适应步长时涉及的衰减系数β_1为0.9,β_2为0.999,极小值δ为1×10^{-6},均为机器学习过程中使用Adam算法进行优化的默认设置. 实验中,LIM策略的超参设置参照了文献[7],即k为$1/n$,b为$-\varepsilon\beta/n$与$\varepsilon\beta/n$之间的随机值,n=1,2,…,N(N=20),$\varepsilon=16$,$\beta=1.5$.

4.3　梯度攻击方法

为了验证NAI-FGM方法的有效性,本节对比了不同梯度攻击方法的黑盒攻击成功率. 进行比较的MI-

FGSM，NI-FGSM均为使用动量优化算法生成对抗样本的方法；ANI-FGSM是已有的基于自适应NAG的攻击方法；Adam-FGM，AMSGrad-FGM均为本文设计的用于验证自适应策略在黑盒攻击中的有效性的方法，算法流程如算法2和算法3所示.

对比实验分别以Inc-v3，Inc-v4，IncRes-v2，Res-101为目标模型攻击其他黑盒模型，仿真结果如表1所示.

算法2 Adam-FGM

输入：干净样本$\boldsymbol{x}$；标签y；损失函数J；扰动ε；最大迭代次数T

输出：对抗样本$\boldsymbol{x}^{\text{adv}}$

1. $\boldsymbol{x}_0^{\text{adv}}=\boldsymbol{x}$; $\boldsymbol{g}_0=0$; $\boldsymbol{m}_0=0$; $\boldsymbol{V}_0=0$; $\alpha=\varepsilon/T$
2. FOR t=0 TO $T-1$ DO
3. 更新梯度$\hat{\boldsymbol{g}}_t=\boldsymbol{g}_t/\left\|\boldsymbol{g}_t\right\|_1$
4. 计算一阶矩$\boldsymbol{m}_t=\beta_1\boldsymbol{m}_{t-1}+(1-\beta_1)\hat{\boldsymbol{g}}_t$
5. 计算二阶矩$\boldsymbol{V}_t=\beta_2\boldsymbol{V}_{t-1}+(1-\beta_2)\hat{\boldsymbol{g}}_t\odot\hat{\boldsymbol{g}}_t$
6. 偏差修正$\hat{\boldsymbol{m}}_t=\boldsymbol{m}_t/\left(1-\beta_1^t\right)$和$\hat{\boldsymbol{V}}_t=\boldsymbol{V}_t/\left(1-\beta_2^t\right)$
7. 更新对抗样本$\boldsymbol{x}_{t+1}^{\text{adv}}=\text{Clip}_{\boldsymbol{x},\varepsilon}\left\{\boldsymbol{x}_t^{\text{adv}}+\alpha\cdot\dfrac{\hat{\boldsymbol{m}}_t}{\hat{\boldsymbol{V}}_t^{1/2}+\delta}\right\}$
8. END FOR
9. $\boldsymbol{x}^{\text{adv}}=\boldsymbol{x}_T^{\text{adv}}$

算法3 AMSGrad-FGM

输入：干净样本$\boldsymbol{x}$；标签y；损失函数J；扰动ε；最大迭代次数T

输出：对抗样本$\boldsymbol{x}^{\text{adv}}$

1. $\boldsymbol{x}_0^{\text{adv}}=\boldsymbol{x}$; $\boldsymbol{g}_0=0$; $\boldsymbol{m}_0=0$; $\boldsymbol{V}_0=0$; $\hat{\boldsymbol{V}}_0=0$; $\alpha=\varepsilon/T$
2. FOR t=0 TO $T-1$ DO
3. 对梯度进行更新$\hat{\boldsymbol{g}}_t=\boldsymbol{g}_t/\left\|\boldsymbol{g}_t\right\|_1$
4. 计算一阶矩$\boldsymbol{m}_t=\beta_1\boldsymbol{m}_{t-1}+(1-\beta_1)\hat{\boldsymbol{g}}_t$
5. 计算二阶矩$\boldsymbol{V}_t=\beta_2\boldsymbol{V}_{t-1}+(1-\beta_2)\hat{\boldsymbol{g}}_t\odot\hat{\boldsymbol{g}}_t$
6. 偏差修正$\hat{\boldsymbol{V}}_t=\max\left(\hat{\boldsymbol{V}}_{t-1},\boldsymbol{V}_t\right)$
7. 更新对抗样本$\boldsymbol{x}_{t+1}^{\text{adv}}=\text{Clip}_{\boldsymbol{x},\varepsilon}\left\{\boldsymbol{x}_t^{\text{adv}}+\alpha\cdot\dfrac{\boldsymbol{m}_t}{\hat{\boldsymbol{V}}_t^{1/2}+\delta}\right\}$
8. END FOR
9. $\boldsymbol{x}^{\text{adv}}=\boldsymbol{x}_T^{\text{adv}}$

其中，“*”表示白盒攻击，加粗数据为最优结果.

从表1可以看出，NAI-FGSM方法生成的对抗样本具有更好的迁移性. 无论是在常规训练模型上还是对抗训练模型上，相对于其他已有的方法（MI-FGSM，NI-FGSM，ANI-FGSM）或者自定义的自适应攻击方法（Adam-FGM，AMSGrad-FGM）均具有明显更高的黑盒攻击成功率. 从仿真实验的角度佐证了本文在NAG算

表1 基于梯度的黑盒攻击成功率 单位：%

目标模型	攻击方法	Inc-v3	Inc-v4	IncRes-v2	Res-101	Inc-v3$_{ens3}$	Inc-v3$_{ens4}$	IncRes-v2$_{ens}$
Inc-v3	MI-FGSM	100*	44.1	43.1	36.7	13.1	12.8	6.4
	NI-FGSM	100*	52.1	49.2	40.4	12.2	12.6	5.7
	ANI-FGSM	100*	47.6	45.3	39.4	13.3	13.3	7.2
	Adam-FGM	100*	45.1	41.6	35.6	13.9	12.5	7.2
	AMSGrad-FGM	100*	43.6	41.5	34.4	12.9	12.7	5.9
	NAI-FGM	100*	**59.7**	**56.2**	**46.4**	**15.2**	**15.4**	**7.7**
Inc-v4	MI-FGSM	55.6	99.8*	45.6	41.3	16.4	14.8	7.4
	NI-FGSM	64.1	100*	50.6	45.6	15.6	14.2	6.9
	ANI-FGSM	58.7	100*	48.5	42.2	16.4	14.4	7.0
	Adam-FGM	55.0	100*	46.1	40.3	16.4	15.3	7.9
	AMSGrad-FGM	55.3	99.9*	46.1	40.4	15.9	15.2	7.7
	NAI-FGM	**70.3**	100*	**57.6**	**50.3**	**18.4**	**17.6**	**8.2**
IncRes-v2	MI-FGSM	57.9	51.1	98.0*	44.4	22.7	16.5	11.2
	NI-FGSM	61.7	54.3	98.9*	45.9	20.0	16.2	10.2
	ANI-FGSM	60.3	53.1	98.4*	45.7	21.3	15.8	10.9
	Adam-FGM	59.2	50.7	97.9*	43.5	21.7	15.5	11.9
	AMSGrad-FGM	59.4	50.2	98.0*	44.9	22.6	17.0	11.4
	NAI-FGM	**71.4**	**64.5**	98.3*	**53.5**	**25.0**	**19.0**	**11.9**
Res-101	MI-FGSM	57.5	50.4	49.1	99.3*	24.4	22.1	12.9
	NI-FGSM	63.6	57.7	56.3	99.4*	24.0	20.6	12.5
	ANI-FGSM	57.8	53.6	50.0	99.3*	24.7	20.4	12.8
	Adam-FGM	57.9	50.1	47.9	99.2*	24.8	21.4	12.8
	AMSGrad-FGM	58.1	49.9	47.7	99.2*	24.7	21.7	13.4
	NAI-FGM	**71.4**	**65.5**	**63.6**	99.3*	**28.0**	**24.1**	**15.6**

法基础上引入自适应策略更新对抗样本,能够兼顾方向和学习率两个角度累积历史梯度信息的优势,构建的对抗样本具有更好的黑盒攻击成功率.

此外,对比NI-FGSM和MI-FGSM两种采用了动量算法的攻击方法,NI-FGSM方法的迁移性整体上明显优于MI-FGSM方法. 进一步观察发现,NI-FGSM方法在常规训练模型上的黑盒攻击成功率更高,但是在对抗训练模型上的表现略低于MI-FGSM方法. 该现象的普遍存在证实了Nesterov项的使用,稳定了梯度方向的同时,容易牺牲模型的泛化能力.

4.4　组合攻击方法

为了验证LIM策略的有效性,本节对比了梯度攻击方法与LIM策略组合前与组合后的黑盒攻击成功率. 我们分别以MI-FGSM,NI-FGSM,ANI-FGSM,Adam-FGM,AMSGrad-FGM方法为基准,添加LIM策略,构建了LI-MI-FGSM,LI-NI-FGSM,LI-ANI-FGSM,LI-Adam-FGM,LI-AMSGrad-FGM方法. 对比实验分别以Inc-v3,Inc-v4,IncRes-v2,Res-101为目标模型攻击其他黑盒模型,仿真结果如表2所示. 为了更好地说明LIM策略提升梯度攻击方法黑盒攻击成功率的能力,对在不同目标模型上获取的黑盒攻击成功率进行了平均,仿真结果如表3所示,加粗数据表示最优结果.

从表2可以看出,LIM策略能够与任意梯度攻击方法进行组合. 从表3可以看出,LIM策略能够有效提升梯度攻击方法的黑盒攻击成功率. 其中,LIM策略对MI-FGSM方法平均提升了30%,对NI-FGSM方法平均提升了32%,对ANI-FGSM方法平均提升了35%,对Adam-FGM方法平均提升了36%,对AMSGrad-FGM方法平均提升了36%.

对比NI-FGSM方法和其他几种采用了自适应步长的攻击方法,整体上NI-FGSM方法能够获取更高的黑盒攻击成功率. 此外,我们发现NI-FGSM在常规训练模型上具有更好的黑盒攻击成功率,但是添加了自适应策略的攻击方法均能够在对抗训练模型上获取高于NI-FGSM的黑盒攻击成功率. 这说明了在优化对抗扰动过程中采用自适应策略从学习率的角度累积梯度,能够一定程度上保障对抗样本生成方法的泛化性.

表2　组合LIM策略的黑盒攻击成功率　　单位:%

目标模型	攻击方法	Inc-v3	Inc-v4	IncRes-v2	Res-101	Inc-v3$_{ens3}$	Inc-v3$_{ens4}$	IncRes-v2$_{ens}$
Inc-v3	LI-MI-FGSM	100*	52.5	48.0	44.8	25.7	24.4	14.0
	LI-NI-FGSM	100*	82.9	80.7	77.1	48.7	44.9	29.8
	LI-ANI-FGSM	100*	82.0	79.5	75.6	48.2	46.8	30.4
	LI-Adam-FGM	100*	82.5	79.7	73.7	48.5	46.4	32.3
	LI-AMSGrad-FGM	100*	81.4	78.9	73.0	49.9	47.2	32.0
Inc-v4	LI-MI-FGSM	88.7	99.7*	84.4	81.5	64.5	60.7	46.9
	LI-NI-FGSM	89.3	100*	83.8	79.6	59.5	55.9	40.2
	LI-ANI-FGSM	89.5	99.9*	84.0	78.7	62.0	59.2	44.1
	LI-Adam-FGM	88.7	99.7*	83.7	79.5	64.1	59.4	45.4
	LI- AMSGrad-FGM	88.7	99.7*	83.3	78.4	62.7	59.9	45.5
IncRes-v2	LI-MI-FGSM	90.2	86.8	99.0*	84.4	72.9	65.2	63.3
	LI-NI-FGSM	89.1	85.8	99.4*	82.1	64.7	56.2	51.9
	LI-ANI-FGSM	90.5	87.7	99.2*	84.9	71.1	63.4	59.9
	LI-Adam-FGM	90.0	87.0	99.1*	84.9	72.2	64.8	64.4
	LI- AMSGrad-FGM	89.3	86.5	98.9*	84.0	71.6	64.4	62.5
Res-101	LI-MI-FGSM	80.3	77.0	77.4	99.8*	62.5	55.4	43.6
	LI-NI-FGSM	82.3	75.3	75.7	99.9*	55.6	49.0	38.8
	LI-ANI-FGSM	81.1	74.7	76.2	99.9*	57.9	53.3	39.9
	LI-Adam-FGM	80.4	74.4	76.1	99.9*	57.7	53.5	42.9
	LI- AMSGrad-FGM	80.2	75.1	74.6	99.8*	57.5	54.8	43.1

为了验证NAI-FGM方法的组合攻击能力,本文对比了NAI-FGM与不同数据增强策略组合前和组合后的黑盒攻击成功率. 本文分别使用DIM,TIM,SIM策略与NAI-FGM方法进行组合,构建了DIM-NAI-FGM,TIM-NAI-FGM,SIM-NAI-FGM方法. 对比实验分别以Inc-v3,Inc-v4,IncRes-v2,Res-101为目标模型攻击其他黑盒模型,仿真结果如表4所示.

为了更好地说明不同数据增强策略对NAI-FGM方

表3 与LIM策略组合前后的平均黑盒攻击成功率 单位:%

攻击方法	Inc-v3	Inc-v4	IncRes-v2	Res-101	Inc-v3$_{ens3}$	Inc-v3$_{ens4}$	IncRes-v2$_{ens}$
MI-FGSM	57.0	48.5	45.9	40.8	19.2	16.6	9.5
LI-MI-FGSM	**86.4**	**72.1**	**69.9**	**70.2**	**56.4**	**51.4**	**42.0**
NI-FGSM	63.1	54.7	52.0	43.9	18.0	15.9	8.8
LI-NI-FGSM	**86.9**	**81.3**	**80.1**	**79.6**	**57.1**	**51.5**	**40.2**
ANI-FGSM	58.9	51.4	47.9	42.4	18.9	16.0	9.5
LI-ANI-FGSM	**87.0**	**81.5**	**79.9**	**79.7**	**59.8**	**55.7**	**43.6**
Adam-FGM	57.4	48.6	45.2	39.8	19.2	16.2	10.0
LI-Adam-FGM	**86.4**	**81.3**	**79.8**	**79.4**	**60.6**	**56.0**	**46.3**
AMSGrad-FGM	57.6	47.9	45.1	39.9	19.0	16.7	9.6
LI-AMSGrad-FGM	**86.1**	**81.0**	**78.9**	**78.5**	**60.4**	**56.6**	**45.8**

表4 DIM,TIM,SIM策略与NAI-FGM方法组合的黑盒攻击成功率 单位:%

目标模型	攻击方法	Inc-v3	Inc-v4	IncRes-v2	Res-101	Inc-v3$_{ens3}$	Inc-v3$_{ens4}$	IncRes-v2$_{ens}$
Inc-v3	DI-NAI-FGM	100*	83.8	81.8	72.8	36.7	33.8	19.4
	SI-NAI-FGM	100*	84.1	81.1	72.3	36.9	34.5	19.8
	TI-NAI-FGM	100*	62.3	56.8	49.1	28.1	26.4	16.7
Inc-v4	DI-NAI-FGM	80.8	99.2*	70.7	62.0	21.8	21.9	12.5
	SI-NAI-FGM	88.9	100*	85.0	77.4	49.9	45.8	30.6
	TI-NAI-FGM	72.9	99.7*	60.6	51.3	31.3	28.1	21.3
IncRes-v2	DI-NAI-FGM	77.0	71.5	94.7*	64.6	31.5	24.8	17.4
	SI-NAI-FGM	90.5	86.9	99.2*	81.2	59.1	50.2	43.1
	TI-NAI-FGM	72.5	66.6	98.0*	56.8	39.0	31.8	28.3
Res-101	DI-NAI-FGM	81.5	76.6	75.5	98.5*	37.9	33.5	21.3
	SI-NAI-FGM	85.0	81.5	80.1	99.8*	49.5	44.9	30.4
	TI-NAI-FGM	70.0	65.4	66.1	99.4*	42.9	39.0	29.5

法黑盒攻击成功率的提升能力,对在不同目标模型上获取的黑盒攻击成功率进行了平均,同时对比了未采用任何数据增强策略进行组合攻击的NAI-FGM方法,以及组合了LIM策略的LI-NAI-FGM方法,仿真结果如表5所示.

表5 NAI-FGM方法与不同数据增强方法组合的平均黑盒攻击成功率 单位:%

攻击方法	Inc-v3	Inc-v4	IncRes-v2	Res-101	Inc-v3$_{ens3}$	Inc-v3$_{ens4}$	IncRes-v2$_{ens}$
NAI-FGM	71.0	63.2	59.1	50.1	21.7	19.0	10.9
DI-NAI-FGM	79.8	77.3	76.0	66.5	32.0	28.5	17.7
SI-NAI-FGM	88.1	84.2	82.1	77.0	48.9	43.9	31.0
TI-NAI-FGM	71.8	64.8	61.2	52.4	35.3	31.3	24.0
LI-NAI-FGM	**91.4**	**88.4**	**86.4**	**84.9**	**65.6**	**59.5**	**47.3**

从表4可以看出,DIM,TIM,SIM策略均能与NAI-FGM方法进行组合攻击.从表5可以看出,不同数据增强策略均能提高NAI-FGM方法的黑盒攻击成功率.并且,本文提出的LIM策略对NAI-FGM方法的提升效果最好,在常规训练模型上平均提升了27%(高于DIM的14%,SIM的22%,TIM的2%),在对抗训练模型上平均提升了40%(高于DIM的9%,SIM的24%,TIM的13%).

4.5 LI-NAI-FGM方法

为了验证LI-NAI-FGM方法的有效性,本节对比了该方法与SI-NI-FGSM,VNI-FGSM方法的黑盒攻击成功率.进行比较的SI-NI-FGSM,VNI-FGSM方法均取得了当前最优的黑盒攻击成功率.对比实验分别以Inc-v3,Inc-v4,IncRes-v2,Res-101为目标模型,直接攻击普通训练模型和常规训练模型,获取黑盒攻击成功率,其仿真结果如表6所示,加粗数据表示最优结果.为了更好地说明LI-NAI-FGM方法的有效性,对在不同目标模型上获取的黑盒攻击成功率进行了平均,仿真结果如表7所示,加粗数据表示最优结果.

从表6可以看出,本文构建的组合攻击方法LI-NAI-FGM明显具有更高的黑盒攻击成功率.从表7可

以看出,LI-NAI-FGM攻击方法在常规训练模型上的平均黑盒攻击成功率为87.8%(高于SI-NI-FGSM的77.2%,VNI-FGSM的74.8%),在对抗训练模型上的平均黑盒攻击成功率为57.5%(高于SI-NI-FGSM的35.3%,VNI-FGSM的35.7%).

为了获取不同攻击方法在防御模型上的黑盒攻击成功率,本文分别使用LI-NAI-FGM,SI-NI-FGSM,VNI-FGSM方法在目标模型Inc-v3上生成对抗样本,然后使用防御模型ComDefend,R&P,FD,Bit-Red,JPEG对生成的对抗样本进行净化,转化为干净样本后再去攻击其他黑盒模型,获取黑盒攻击成功率,其仿真结果如表8所示,加粗数据为最优结果. 从表8可以看出,本文构建的组合攻击方法LI-NAI-FGM在防御模型上的黑盒攻击成功率均高于SI-NI-FGSM和VNI-FGSM方法.

表6　常规训练模型和对抗训练模型的黑盒攻击成功率　　单位:%

目标模型	攻击方法	Inc-v3	Inc-v4	IncRes-v2	Res-101	Inc-v3$_{ens3}$	Inc-v3$_{ens4}$	IncRes-v2$_{ens}$
Inc-v3	SI-NI-FGSM	100*	76.7	73.9	66.9	33.5	29.9	16.6
	VNI-FGSM	100*	77.4	74.6	65.6	34.0	33.5	18.4
	LI-NAI-FGM	100*	**90.0**	**86.7**	**83.0**	**55.4**	**52.0**	**32.9**
Inc-v4	SI-NI-FGSM	84.8	100*	79.1	72.7	43.9	40.2	24.7
	VNI-FGSM	83.3	100*	75.6	66.3	39.6	36.7	24.4
	LI-NAI-FGM	**93.1**	100*	**88.1**	**83.9**	**67.7**	**62.8**	**47.6**
IncRes-v2	SI-NI-FGSM	87.2	81.5	99.5*	75.4	51.4	42.6	34.8
	VNI-FGSM	81.3	77.0	98.3*	68.9	47.4	40.1	34.7
	LI-NAI-FGM	**93.9**	**91.3**	99.3*	**87.7**	**72.9**	**64.7**	**62.1**
Res-101	SI-NI-FGSM	79.9	74.7	74.1	99.8*	42.3	38.3	25.6
	VNI-FGSM	79.0	74.8	73.4	99.6*	46.3	42.5	31.0
	LI-NAI-FGM	**87.3**	**84.0**	**84.3**	99.8*	**66.4**	**58.6**	**46.5**

表7　常规训练模型和对抗训练模型的平均黑盒攻击成功率　　单位:%

攻击方法	Inc-v3	Inc-v4	IncRes-v2	Res-101	Inc-v3$_{ens3}$	Inc-v3$_{ens4}$	IncRes-v2$_{ens}$
SI-NI-FGSM	84.0	77.6	75.7	71.7	42.8	37.8	25.4
VNI-FGSM	81.2	76.4	74.5	66.9	41.8	38.2	27.1
LI-NAI-FGM	**91.4**	**88.4**	**86.4**	**85.0**	**65.6**	**59.5**	**47.3**

表8　防御模型的黑盒攻击成功率　　单位:%

防御模型	攻击方法	Inc-v4	IncRes-v2	Res-101	Inc-v3$_{ens3}$	Inc-v3$_{ens4}$	IncRes-v2$_{ens}$	平均
ComDefend	SI-NI-FGSM	62.2	59.1	62.1	54.1	52.5	42.4	55.4
	VNI-FGSM	63.5	60.6	60.1	54.9	55.2	43.3	56.3
	LI-NAI-FGM	**76.8**	**71.8**	**75.7**	**67.2**	**68.1**	**56.1**	**69.3**
R&P	SI-NI-FGSM	71.4	70.3	66.4	32.1	31.8	17.4	48.2
	VNI-FGSM	71.2	70.1	62.7	35.9	32.3	20.2	48.7
	LI-NAI-FGM	**84.9**	**83.3**	**80.9**	**55.2**	**52.5**	**34.7**	**65.3**
FD	SI-NI-FGSM	67.9	67.3	64.3	52.3	51.6	41.7	57.5
	VNI-FGSM	68.2	66.5	61.5	49.7	51.1	40.7	56.3
	LI-NAI-FGM	**82.4**	**80.5**	**79.4**	**68.9**	**67.0**	**58.3**	**72.8**
Bit_Red	SI-NI-FGSM	53.2	53.3	52.8	29.8	32.0	23.1	40.7
	VNI-FGSM	56.4	54.2	50.4	28.4	28.5	22.9	40.1
	LI-NAI-FGM	**66.5**	**64.2**	**65.3**	**40.0**	**41.7**	**32.1**	**51.6**
JPEG	SI-NI-FGSM	76.4	72.7	66.6	57.3	54.2	39.6	61.1
	VNI-FGSM	77.1	73.8	64.2	57.1	53.7	40.7	61.1
	LI-NAI-FGM	**88.6**	**85.5**	**82.4**	**76.2**	**73.2**	**57.0**	**77.2**

4.6　对抗样本对比

为了验证NAI-FGM,LI-NAI-FGM方法的有效性,本节直观比较了NAI-FGM方法与非组合攻击方法MI-FGSM,NI-FGSM,ANI_FGSM生成的对抗样本,LI-NAI-FGM方法与组合攻击方法SI-NI-FGSM,VNI-FGSM生成的对抗样本. 对比实验以Inc-v3为目标模型,分别对背

景单一的样本和纹理复杂的样本进行攻击，生成的对抗样本如图2所示．图2中我们同时给出了对抗样本的图像质量评价SSIM和PSNR值，用于定量评估攻击过程中产生的对抗噪声．其中，SSIM（Structural SIMilarity）为结构相似性，是一种量化图像间结构相似性的指标．SSIM指标是一个0到1之间的数，越大表示进行对比的图像与原始图像的差距越小，即图像失真越小．PSNR（Peak Signal to Noise Ratio）为峰值信噪比，它是一种衡量噪声水平的度量标准．PSNR指标是进行对比的图像与原始图像之间的均方误差相对于$\left(2^n-1\right)^2$的对数值．所以均方误差越小，PSNR则越大，代表图像质量越好．

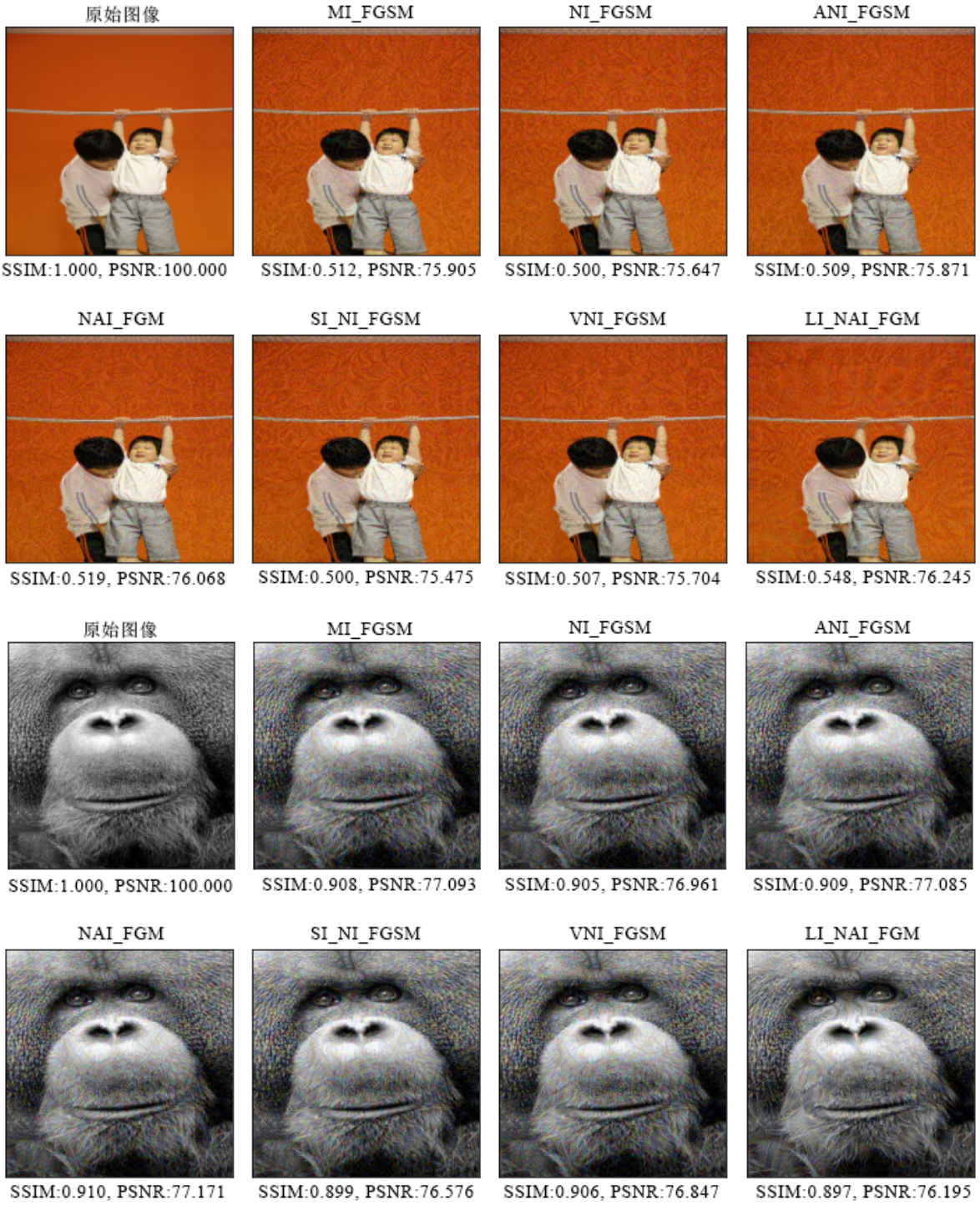

图2 不同攻击方法生成的对抗样本对比

从图2可以看出,无论是背景单一的样本还是纹理复杂的样本,NAI-FGM,LI-NAI-FGM方法的SSIM和PSNR值与其他已有的攻击方法均较为接近.具体地,NAI-FGM方法相对于MI_FGSM,NI_FGSM,ANI-FGSM攻击方法,在背景单一和纹理复杂的样本上生成的对抗样本均具有稍高的SSIM和PSNR值.LI-NAI-FGM方法相对于SI_NI_FGSM,VNI_FGSM两种组合方法,在背景单一的样本上生成的对抗样本具有稍高的SSIM和PSNR值,在纹理复杂的样本上生成的对抗样本的SSIM和PSNR值则稍低.

综合以上分析,我们可以得到以下结论:本文提出的方法相对于已有的攻击方法,在保证SSIM和PSNR指标的同时,能够获取更高的黑盒攻击成功率.

5　结论及展望

本文针对梯度攻击方法,从提高迁移性的角度出发,研究了梯度优化算法,提出了基于Nesterov和自适应步长的对抗样本生成方法NAI-FGM.为了进一步提高黑盒攻击成功率.本文探索了数据增强技术,提出了线性变换不变性策略LIM.NAI-FGM方法作为一种梯度攻击方法,能够和数据增强技术组合,提高黑盒攻击成功率.LIM策略作为一种数据增强技术,能够作用于梯度攻击方法,生成迁移性更强的对抗样本.实验表明,NAI-FGM方法相对于其他梯度攻击方法具有更高的黑盒攻击成功率,LIM策略相对于其他数据增强技术对黑盒攻击成功率的提升更加明显.组合攻击方法LI-NAI-FGM在常规训练模型上的平均黑盒攻击成功率达到87.8%,在对抗训练模型上的平均黑盒攻击成功率达到57.5%,在防御模型上的平均黑盒攻击成功率达到67.2%,均超过现有最高水平.

参考文献

[1] LANG C B, CHENG G, TU B F, et al. Learning what not to segment: A new perspective on few-shot segmentation [C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2022: 8047-8057.

[2] TIAN Z, SHEN C H, CHEN H, et al. FCOS: Fully convolutional one-stage object detection[C]//2019 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway: IEEE, 2020: 9626-9635.

[3] GOODFELLOW I J, SHLENS J, SZEGEDY C. Explaining and harnessing adversarial examples[C]//International Conference on Learning Representations. San Diego: ICLR, 2015: 1-11.

[4] LI Y D, LI L J, WANG L Q, et al. NATTACK: Learning the distributions of adversarial examples for an improved black-box attack on deep neural networks[C]//2019 International Conference on Machine Learning. Long Beach: ICML, 2019: 3866-3876.

[5] HUANG Y, KONG A W. Transferable adversarial attack based on integrated gradients[C]//International Conference on Learning Representations. Virtual Conference: Ithaca, 2022: 1-25.

[6] 纪守领, 杜天宇, 邓水光, 等. 深度学习模型鲁棒性研究综述[J]. 计算机学报, 2022, 45(1): 190-206.
JI S L, DU T Y, DENG S G, et al. Robustness certification research on deep learning models: A survey[J]. Chinese Journal of Computers, 2022, 45(1): 190-206. (in Chinese)

[7] WANG X S, HE K. Enhancing the transferability of adversarial attacks through variance tuning[C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2021: 1924-1933.

[8] KURAKIN A, GOODFELLOW I J, BENGIO S. Adversarial examples in the physical world[C]//International Conference on Learning Representations, Workshop Track Proceedings. Toulon: Ithaca, 2017: 1-14.

[9] RUDER S. An overview of gradient descent optimization algorithms[EB/OL]. (2016-09-15)[2022-06-27]. https://arxiv.org/abs/1609.04747.

[10] DEFAZIO A, JELASSI S. Adaptivity without compromise: A momentumized, adaptive, dual averaged gradient method for stochastic optimization[EB/OL]. (2021-01-26)[2022-06-27]. https://arxiv.org/abs/2101.11075.

[11] 陇盛, 陶蔚, 张泽东, 等. 基于AdaGrad的自适应NAG方法及其最优个体收敛性[J]. 软件学报, 2022, 33(4): 1231-1243.
LONG S, TAO W, ZHANG Z D, et al. Adaptive NAG methods based on AdaGrad and its optimal individual convergence[J]. Journal of Software, 2022, 33(4): 1231-1243. (in Chinese)

[12] DONG Y P, LIAO F Z, PANG T Y, et al. Boosting adversarial attacks with momentum[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2018: 9185-9193.

[13] LIN J D, SONG C B, HE K, et al. Nesterov accelerated gradient and scale invariance for adversarial attacks[C]// International Conference on Learning Representations. Virtual Conference: Ithaca, 2020: 1-12.

[14] POLYAK B T. Some methods of speeding up the convergence of iteration methods[J]. USSR Computational Mathematics and Mathematical Physics, 1964, 4(5): 1-17.

[15] NESTEROV Y E. A method for solving the convex programming problem with convergence rate $O(1/k^2)$[J]. Proceedings of the USSR Academy of Sciences, 1983, 269: 543-547.

[16] NEMIROVSKIĬ A S, IŪDIN D B. Problem Complexity and Method Efficiency in Optimization[M]. New York: Wiley, 1983.

[17] SZEGEDY C, VANHOUCKE V, IOFFE S, et al. Rethinking the inception architecture for computer vision [C]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2016: 2818-2826.

[18] PANG R, SHEN H, ZHANG X Y, et al. A tale of evil twins: Adversarial inputs versus poisoned models[C]// 2020 ACM SIGSAC Conference on Computer and Communications Security. New York: ACM, 2020: 85-99.

[19] ZHOU W, HOU X, CHEN Y J, et al. Transferable adversarial perturbations[C]//European Conference on Computer Vision. Cham: Springer, 2018: 471-486.

[20] WU W B, SU Y X, CHEN X X, et al. Boosting the transferability of adversarial samples via attention[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2020: 1158-1167.

[21] GONG C Y, WANG D L, LI M, et al. KeepAugment: A simple information-preserving data augmentation approach[C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2021: 1055-1064.

[22] Raphael G, Sylvia S, Ekin D C, et al. Tradeoffs in data augmentation: An empirical study[C]//International Conference for Learning Representations. Virtual Conference: Ithaca, 2021: 1-27.

[23] XIE S F, LV A, XIA Y C, et al. Target-side input augmentation for sequence to sequence generation[C]//International Conference for Learning Representations. Virtual Conference: Ithaca, 2022: 1-18.

[24] 邹军华, 段晔鑫, 任传伦, 等. 基于噪声初始化、Adam-Nesterov方法和准双曲动量方法的对抗样本生成方法[J]. 电子学报, 2022, 50(1): 207-216.
ZOU J H, DUAN Y X, REN C L, et al. Perturbation initialization, Adam-Nesterov and quasi-hyperbolic momentum for adversarial examples[J]. Acta Electronica Sinica, 2022, 50(1): 207-216. (in Chinese)

[25] XIE C H, ZHANG Z S, ZHOU Y Y, et al. Improving transferability of adversarial examples with input diversity [C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2020: 2725-2734.

[26] DONG Y P, PANG T Y, SU H, et al. Evading defenses to transferable adversarial examples by translation-invariant attacks[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2020: 4307-4316.

[27] RUSSAKOVSKY O, DENG J, SU H, et al. ImageNet large scale visual recognition challenge[J]. International Journal of Computer Vision, 2015, 115(3): 211-252.

[28] SZEGEDY C, IOFFE S, VANHOUCKE V, et al. Inception-v4, inception-ResNet and the impact of residual connections on learning[C]//2017 AAAI Conference on Artificial Intelligence. New York: ACM, 2017: 4278-4284.

[29] HE K M, ZHANG X Y, REN S Q, et al. Deep residual learning for image recognition[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2016: 770-778.

[30] TRAMÈR F, KURAKIN A, PAPERNOT N, et al. Ensemble adversarial training: Attacks and defenses[C]//International Conference on Learning Representations. Vancouver: Ithaca, 2018: 1-14.

[31] JIA X J, WEI X X, CAO X C, et al. ComDefend: An efficient image compression model to defend adversarial examples[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2020: 6077-6085.

[32] XIE C H, ZHANG Z SH, YUILLE A L. Mitigating adversarial effects through randomization[C]//International Conference on Learning Representations. Vancouver: Ithaca, 2018: 1-16.

[33] LIU Z H, LIU Q, LIU T, et al. Feature distillation: DNN-oriented JPEG compression against adversarial examples [C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2020: 860-868.

[34] XU X L, EVANS D, QI Y J. Feature squeezing: Detecting adversarial examples in deep neural networks[C]//Network and Distributed System Security Symposium. San Diego: Internet Society, 2017: 1-15.

[35] GUO C, RANA M, CISSE M, et al. Countering adversarial images using input transformations[C]//International Conference on Learning Representations. Vancouver:

Ithaca, 2018: 1-12.

[36] 陇盛, 陶蔚, 张泽东, 等. 非光滑强凸情形 Adam 型算法的最优收敛速率[J]. 电子学报, 2022, 50(9): 2049-2059.
LONG S, TAO W, ZHANG Z D, et al. The optimal convergence rate of Adam-type algorithms for non-smooth strongly convex cases[J]. Acta Electronica Sinica, 2022, 50(9): 2049-2059. (in Chinese)

[37] REDDI S J, KALE S, KUMAR S. On the convergence of adam and beyond[C]//International Conference on Learning Representations. Vancouver: Ithaca, 2018: 1-23.

作者简介

鲍　蕾　女, 1987 年 2 月生, 安徽芜湖人. 博士. 现为陆军炮兵防空兵学院讲师. 主要研究领域为机器学习、计算机视觉.
E-mail: baolei1219@sina.cn

陶　蔚　男, 1991 年生, 安徽合肥人. 博士. 现为中国人民解放军军事科学院助理研究员. 主要研究领域为机器学习.
E-mail: wtao_plaust@163.com

陶　卿　男, 1965 年生. 安徽合肥人. 博士. 现为陆军炮兵防空兵学院教授, 博士生导师. 主要研究领域为机器学习、模式识别、应用数学.
E-mail: qing. tao@ia.ac.cn