ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 36,No. 2,Mar. 2021,pp. 248－259
DOI:10. 16337/j. 1004-9037. 2021. 02. 006

http://sjcj. nuaa. edu. cn
E-mail:sjcj@nuaa. edu. cn
Tel/Fax:＋86-025-84892742

融合高斯噪声和翻转策略的对抗攻击

张　武[1]，段晔鑫[1,2]，邹军华[1]，潘志松[1]，周星宇[3]

(1. 陆军工程大学指挥控制工程学院,南京 210007;2. 陆军军事交通学院镇江校区,镇江 212001;3. 陆军工程大学通信工程学院,南京 210007)

摘　要：在对抗攻击研究领域,黑盒攻击相比白盒攻击更具挑战性和现实意义。目前实现黑盒攻击的主流方法是利用对抗样本的迁移性,然而现有大多数方法所得的对抗样本在黑盒攻击时效果不佳。本文提出了一种基于高斯噪声和翻转组合策略方法来增强对抗样本的迁移性,进而提升其黑盒攻击性能。同时,该方法可与现有基于梯度的攻击方法相结合形成更强的对抗攻击。本文在一个与ImageNet相容的数据集上做了大量实验,实验结果表明本文方法所得的对抗样本在黑盒攻击性能上有显著提升。并且,本文最佳攻击组合能以86.2%的平均成功率欺骗6种先进防御模型,相比目前最强攻击方法提升约8.0%。

关键词：黑盒攻击;对抗样本;迁移性;高斯噪声;翻转

中图分类号：TP181　　**文献标志码**:A

Adversarial Attacks with Gaussian Noise and Flipping Strategy

ZHANG Wu[1], DUAN Yexin[1,2], ZOU Junhua[1], PAN Zhisong[1], ZHOU Xingyu[3]

(1. Command and Control Engineering College, Army Engineering University of PLA, Nanjing 210007, China; 2. Zhenjiang Campus, Army Military Transportation University of PLA, Zhenjiang 212001, China; 3. Communication Engineering College, Army Engineering University of PLA, Nanjing 210007, China)

Abstract: For adversarial attacks, black-box attacks are more challenging and applicable than white-box attacks. Recently, black-box attacks based on the transferability of adversarial examples have become mainstream methods. However, the adversarial examples generated by most existing methods exhibit low efficiency in black-box attacks. In this paper, a combination strategy based on Gaussian noise and flipping is proposed to enhance the transferability of adversarial examples, thus achieving higher black-box attack success rates. Moreover, this strategy can be integrated into any gradient-based method to obtain stronger attacks. Extensive experiments on an ImageNet-compatible dataset show that our proposed method can generate more transferable adversarial examples. In addition, our best attack can fool six state-of-the-art defense models with an average success rate of 86.2%, and deliver 8.0% success rate increasement compared with the state-of-the-art gradient-based attack.

Key words: black-box attack; adversarial example; transferability; Gaussian noise; flipping

---

基金项目:国家自然科学基金(62076251)资助项目。
收稿日期:2020-08-23;**修订日期**:2020-12-14

引　言

对抗样本概念由Szegedy等[1]在2013年首次提出,即在原始图像中添加微小的扰动便可生成让神经网络模型高置信度错误分类的对抗样本。对抗样本在白盒攻击场景下生成较为容易,即攻击者能访问到目标模型的体系结构和参数。然而,在现实场景中,攻击者所遇到的目标模型大多为黑盒模型,即攻击者无法获取到其内部结构和参数,此时对抗样本生成较为困难。目前,对于黑盒模型的攻击方法,大致可分为基于查询的方法和基于模型迁移的方法[2]。其中,基于查询的方法需要大量访问黑盒模型的反馈结果,因此十分耗时且易被察觉。例如,Brendel等[3]提出的基于边界探索的黑盒攻击就是一种基于查询的方法,该攻击方式完全依赖于模型的反馈结果来生成对抗样本。除了基于查询的方法外,对抗样本的迁移性[4]为攻击者提供了另一种可行性,即针对白盒模型所生成的对抗样本往往能够成功攻击同一任务其他黑盒模型。

虽然研究者已提出许多方法来提高对抗样本的迁移性,但仍存在不足。例如,Dong等[5]和Xie等[6]提出的方法对于普通黑盒模型的攻击性较强,但对于防御黑盒模型的攻击性却较弱。相比一下,Dong等[7]在2019年提出的方法在攻击防御黑盒模型方面有了较大提升,但在攻击普通黑盒模型时相对变差。一般而言,迭代攻击比单步攻击更容易出现特定模型过拟合现象,所得对抗样本的迁移性较差[5]。受数据增强技术启发,该技术可以有效降低特定模型过拟合现象并提升泛化能力[8]。为此,本文在迭代攻击基础上利用高斯噪声和翻转组合策略方法来增强迁移对抗攻击,从而整体提升针对普通和防御模型的黑盒攻击成功率。本文在单模型和多模型条件下测试了所提出方法的攻击成功率。大量实验表明,本文方法既能使黑盒攻击性能优于基线方法,又能让白盒攻击保持较高的成功率。

1　相关研究

1.1　对抗样本生成描述

设输入样本为$x\in\mathbf{R}^d$,其对应的类别标签为$y\in\{1,2,3,\cdots,k\}$,深度神经网络分类器$f(x)$:$x\rightarrow y$是d维输入样本x到类别标签y的映射函数。攻击者通过添加小幅度的扰动δ来制作对抗样本$x^{\text{adv}}=x+\delta$,使得神经网络错误分类,即$f(x^{\text{adv}})\neq y$。为了生成对抗样本,既要最大化网络模型分类器的损失函数$J(x^{\text{adv}},y)$,又要使用L_∞范数将对抗样本x^{adv}约束在x附近,即扰动δ的无穷范数应小于阈值ε。因此,对抗样本生成过程可描述为

$$\arg\max_{x^{\text{adv}}}\left(J(x^{\text{adv}},y)\right)\quad \text{s.t. } \|\delta\|_\infty\leqslant\varepsilon \tag{1}$$

1.2　基于梯度攻击方法

快速梯度符号方法(Fast gradient sign method, FGSM)[9]是一种基于梯度的经典方法,只需要对输入样本进行一次梯度更新,就能快速生成对抗样本。该方法根据损失函数的梯度方向来对输入样本中的每个像素进行等值增加或者减少,其对抗样本生成公式为

$$x^{\text{adv}}=x+\varepsilon\cdot\text{sign}(\nabla_x J(x,y)) \tag{2}$$

式中:sign(·)为符号函数;$\nabla_x J(\cdot)$为计算损失函数的梯度;ε使扰动满足L_∞范数约束。

迭代快速梯度符号方法(Iterative fast gradient sign method, I-FGSM)[10]是FGSM的迭代版本,即将单步更新分解为多轮迭代更新,以实现添加更小幅度的扰动,其迭代过程可表示为

$$x_0^{\text{adv}}=x,\ x_{t+1}^{\text{adv}}=x_t^{\text{adv}}+\alpha\cdot\text{sign}(\nabla_{x_t^{\text{adv}}}J(x_t^{\text{adv}},y)) \tag{3}$$

式中:x_{t+1}^{adv}表示第$t+1$次迭代生成的对抗样本;T表示迭代次数;步长$\alpha=\varepsilon/T$确保所生成的对抗样本限制在x的ε领域内。相比于FGSM方法,I-FGSM方法在白盒攻击性能上有了提升,但黑盒攻击性能

相对变弱。

动量迭代快速梯度符号方法(Momentum iterative fast gradient sign method, MI-FGSM)[5]则是在I-FGSM基础上增加了动量因子,即在每一轮迭代中加入前面所有轮的梯度信息。因此,该方法在迭代过程中能稳定更新方向,有效避免陷入局部极值,其更新过程表示为

$$g_{t+1}=\mu\cdot g_t+\frac{\nabla_{x_t^{\mathrm{adv}}}J(x_t^{\mathrm{adv}},y)}{\left\|\nabla_{x_t^{\mathrm{adv}}}J(x_t^{\mathrm{adv}},y)\right\|_1} \tag{4}$$

$$x_{t+1}^{\mathrm{adv}}=x_t^{\mathrm{adv}}+\alpha\cdot\mathrm{sign}(g_{t+1}) \tag{5}$$

式中:g_{t+1}表示第$t+1$次迭代时的累积梯度;μ为动量衰减因子。

多样性输入迭代快速梯度符号方法(Diverse inputs iterative fast gradient sign method, DI^2-FGSM)[6]是在每次迭代时以一定的概率对输入图像进行随机调整大小及填充变换操作,从而生成更具迁移性的对抗样本。此方法可与MI-FGSM方法有效结合,从而形成更强的对抗攻击方法M-DI^2-FGSM。为了简洁起见,本文将M-DI^2-FGSM方法简称为DIM方法。

$$D(x_t^{\mathrm{adv}};p)=\begin{cases}D(x_t^{\mathrm{adv}}) & \text{with probability } p\\ x_t^{\mathrm{adv}} & \text{with probability } 1-p\end{cases} \tag{6}$$

式中:$D(\cdot)$表示变换函数;p为变换概率。

平移不变性迭代快速梯度符号方法(Translation-invariant iterative fast gradient sign method, TI-FGSM)[7]方法则使用预定义的高斯核卷积未平移输入图像的梯度来代替对一组已平移图像的梯度计算,从而大大提高了攻击的计算效率,并且生成的对抗样本黑盒攻击防御模型时效果很好。同样,将TI-FGSM方法与DIM方法进行融合,可形成目前最强黑盒攻击方法TI-DIM。

2 方法实现

本文提出将高斯噪声和翻转策略(Gaussian noise and flipping strategy, GF)应用到基于梯度的攻击方法中,生成可迁移性更强的对抗样本。其中,GF策略对抗攻击方法的整体框架如图1所示。

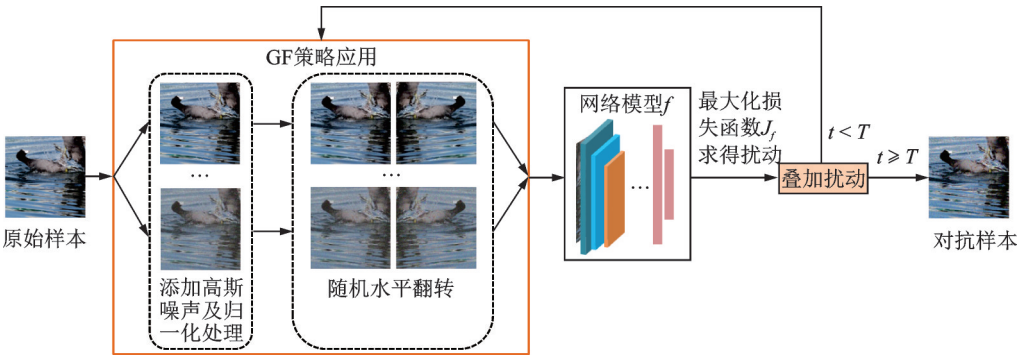


图1 GF策略对抗攻击框架图

Fig.1 Architecture graph of GF strategy adversarial attack

2.1 GF策略方法

高斯噪声是指其概率密度函数服从正态分布的一类噪声。高斯噪声注入作为一种数据增强技术,其主要把随机高斯噪声点添加到输入样本中以帮助模型降低过拟合现象。为此,本文采取类似方法,即将高斯噪声添加到样本中以降低所得对抗样本过度拟合于特定模型,从而提升其可迁移性。

此外,在对抗样本生成过程中,样本的像素值是有范围限定的,而对于添加了高斯噪声的样本其部分像素值可能会发生越界情况。若此时直接采用像素值裁剪方式来处理越界像素值,会使得所添加的噪声信息部分丢失,最终影响到所生成的对抗样本的可迁移性。为此,本文采取像素值归一化方式来处理添加了高斯噪声的样本,这样既能保留所添加的噪声信息,又能保持原有样本数据分布。

翻转是一种简单又有效的几何空间变换增强技术,其中水平翻转比垂直翻转更为常见[11]。因此,文中选择将随机水平翻转(Random horizontal flipping, RHF)与高斯噪声相结合以生成迁移性更强的对抗样本。该组合策略可形式化表示为

$$GF(x)=\mathrm{RHF}\left(\frac{x+\mathrm{noise}}{\|x+\mathrm{noise}\|_\infty}\right)\quad \text{s.t. noise}\sim N(0,\mathrm{std}^2) \tag{7}$$

式中:RHF表示随机水平翻转输入;noise为高斯噪声点;std为高斯噪声标准差。

2.2　GF策略单模型攻击

由于FGSM白盒攻击成功率不高和I-FGSM黑盒攻击性能相对较弱,本文主要将GF策略整合到诸如MI-FGSM,DIM和TI-DIM较强基线方法中,从而衍生出GF-MI-FGSM、GF-DIM和GF-TI-DIM方法。根据3.3节结果可知,对抗攻击强度最强为GF-TI-DIM方法,其次是GF-DIM方法,最后为GF-MI-FGSM方法。为此,本文主要对GF-DIM和GF-TI-DIM方法进行阐述。

由于GF策略与随机调整大小填充属于两种不同的图像变换方式,将它们融合在一起可进一步缓解过度拟合现象,并形成攻击力更强的GF-DIM攻击方法,其融合公式为

$$x_t^{\mathrm{mix}}=\beta\cdot\mathrm{GF}(x_t^{\mathrm{adv}})+(1-\beta)\cdot D(x_t^{\mathrm{adv}};p) \tag{8}$$

$$g_{t+1}=\mu\cdot g_t+\frac{\nabla_{x_t^{\mathrm{adv}}}J(x_t^{\mathrm{mix}},y)}{\left\|\nabla_{x_t^{\mathrm{adv}}}J(x_t^{\mathrm{mix}},y)\right\|_1} \tag{9}$$

$$x_{t+1}^{\mathrm{adv}}=x_t^{\mathrm{adv}}+\alpha\cdot\mathrm{sign}(g_{t+1}) \tag{10}$$

式中:x_t^{mix}为x_t^{adv}经过GF策略和随机调整大小填充两种不同变换后的样本;β为用于权衡两种变换方式的滑动平均系数;$D(\cdot)$表示以一定概率p对输入进行随机调整大小填充。当$p=0$时,GF-DIM方法则将变为GF-MI-FGSM方法。

同样,若将GF策略整合到TI-DIM方法中,则可衍生成GF-TI-DIM攻击算法,其融合过程与GF-DIM攻击方法大致一样,唯一不同之处是要在式(9)中加入高斯核,即为

$$g_{t+1}=\mu\cdot g_t+\frac{W*\nabla_{x_t^{\mathrm{adv}}}J(x_t^{\mathrm{mix}},y)}{\left\|W*\nabla_{x_t^{\mathrm{adv}}}J(x_t^{\mathrm{mix}},y)\right\|_1} \tag{11}$$

式中W为预定义高斯核。

2.3　GF策略集成模型攻击

假如对抗样本能够欺骗多个模型,则意味着它对于其他模型具有较强迁移性[12]。因此,攻击一组模型可实现更强的黑盒攻击。本文采用Dong等[5]提出的logit集成方案来构造GF策略集成模型攻击方法。根据3.4节结果可知,黑盒攻击成功率最好的是MGF-TI-DIM算法,该算法是GF-TI-DIM集成攻击方法,其伪代码如下。

输入:k个分类模型,图像样本x,真实类别标签y,迭代次数T,高斯核W,随机调整大小填充概率p,衰减因子μ,w_i为第i个模型权重,最大扰动ε,则步长$\alpha=\varepsilon/T$。

输出：对抗样本 x^{adv}

步骤1 初始化参数

$$i=1, t=0, x_0^{\text{adv}}=x$$

步骤2 循环迭代 T 次

$$\text{while } t<T \text{ do}$$

步骤3 求解 k 个模型logit值

$$\text{while } i\leqslant k \text{ do}$$

$$l_i(x_t^{\text{mix}})=l_i(\beta\cdot \text{GF}(x_t^{\text{adv}})+(1-\beta)\cdot D(x_t^{\text{adv}};p))$$

$$\text{end while}$$

步骤4 将 k 个模型logit值乘以对应权重进行累加

$$l(x_t^{\text{mix}})=\sum_{j=1}^{k} w_j l_j(x_t^{\text{mix}})$$

步骤5 求解交叉熵损失函数

$$J(x_t^{\text{mix}}, y)=-l_y\cdot \log(\text{softmax}(l(x_t^{\text{mix}})))$$

步骤6 计算梯度

$$g_{t+1}=\mu\cdot g_t+\frac{W*\nabla_{x_t^{\text{adv}}} J(x_t^{\text{mix}}, y)}{\left\| W*\nabla_{x_t^{\text{adv}}} J(x_t^{\text{mix}}, y)\right\|_1}$$

步骤7 叠加扰动

$$x_{t+1}^{\text{adv}}=x_t^{\text{adv}}+\alpha\cdot \text{sign}(g_{t+1})$$

步骤8 迭代结束

$$\text{end while}$$

步骤9 输出对抗样本

$$\text{return } x^{\text{adv}}=x_T^{\text{adv}}$$

若把MGF-TI-DIM算法中第6步操作修改为式(9)，则MGF-TI-DIM算法就演变为MGF-DIM算法。同样，若在输入时将概率 p 设置为0，MGF-DIM算法就退化为MGF-MI-FGSM算法。详细转化关系如图2所示。

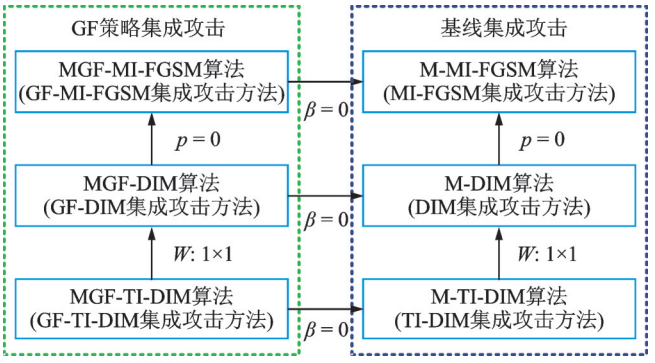


图2 集成攻击算法转化图

Fig.2 Conversion graph of ensemble-based attack algorithm

3 实验结果与分析

3.1 实验设置

3.1.1 数据集与超参数设置

本文采用了一个与ImageNet相兼容的数据集。该数据集包含1 000张大小为299像素×299像素×3像素的图像，且在NIPS 2017对抗竞赛所用。在超参数设置方面，本文设置最大扰动 $\varepsilon=16$，适用于输入样本像素值在[0, 255]范围内计算，以及总迭代次数 $T=10$ 和动量衰减因子 $\mu=1$。对于DIM方法，变换概率 $p=0.7$，而对于MI-FGSM方法，则变换概率 $p=0$。对于TI-DIM方法，高斯核 W 大小设置为15×15。本文在GF策略单模型攻击方法和集成模型攻击方法上作了对比实验，每种GF策略攻击方法与其相应的基线攻击方法在超参数设置方面相同。

3.1.2　评估指标

目前,在对抗攻击研究领域里,用于评价攻击效果的指标主要是攻击成功率(Attack success rate,ASR),即被攻击模型的错误分类率。然而,对于不同类型攻击,ASR定义有所不同。本文攻击方法属于一种无目标攻击方式,即生成的对抗样本使得分类模型预测结果与真实类别不一致就已达到攻击效果,其相应的ASR可定义为

$$\mathrm{ASR}=\frac{\text{样本错误分类个数}}{\text{样本总个数}}\times 100\% \tag{12}$$

此外本文还考虑了10个模型作为攻击对象,用于评价本文方法的黑盒攻击性能。其中4个是普通模型,分别是Inception-v3(Inc-v3)[13]、Inception-v4(Inc-v4)[14]、Inception-Resnet-v2(IncRes-v2)[14]和Resnet-v2-152(Res-152)[15]。其余6个为防御模型,分别是经过对抗训练的集成模型Inc-v3ens3、Inc-v3ens4和IncRes-v2ens[16]以及NIPS 2017防御竞赛中排名前三的模型HGD[17],R&P[18]和NIPS-r3。集成防御模型是利用其他模型生成的对抗样本来扩充训练数据,使得对抗样本与特定模型脱钩,从而提高防御性能。HGD模型则使用降噪网络作为防御措施来去除对抗性噪声。R&P模型和NIPS-r3模型是先将图像经过随机变换后再传递给卷积神经网络进行分类,以减轻对抗效果。

3.2　消融分析

3.2.1　标准差取值

本文翻转策略是通过对输入进行随机水平翻转方式来实现。虽然随机水平翻转在提高对抗样本的黑盒攻击性能方面低于高斯噪声,但它可与高斯噪声有效组合生成黑盒攻击能力更强的对抗样本。然而,高斯噪声的标准差值选择对于提高对抗样本的攻击性起着关键作用。如果标准差设置为0,则高斯噪声和翻转组合策略方法就将变为只有随机水平翻转,所得对抗样本的攻击性能将会变弱。同样,如果标准差std设置过大,则会导致原有样本失真严重,从而影响所得对抗样本的攻击性能,为此进行消融研究以找到合适的高斯噪声标准差值。

本节设置标准差值变化范围为0.0~0.25,并分别使用GF-MI-FGSM、GF-DIM和GF-TI-DIM方法白盒攻击Inc-v3模型生成相应的对抗样本。为了精确地测量标准差效果,开始阶段变化间隔设置为0.02,随着趋于稳定时变化间隔设置为0.01。图3显示了生成的对抗样本对于4种普通模型和4种防御模型的攻击成功率。从图3中可以看到,在白盒攻击保持较高成功率的同时,对于所有黑盒攻击,普通模型要比防御模型稍微较早趋于稳定。此外,当防御模型攻击成功率趋于最大值时普通模型开始轻微下降。为此,本文综合考虑8个模型平均攻击成功率,选择使平均攻击成功率达到最大的标准差值,即std=0.2。因此在以下实验中将高斯噪声标准差值设置为0.2。

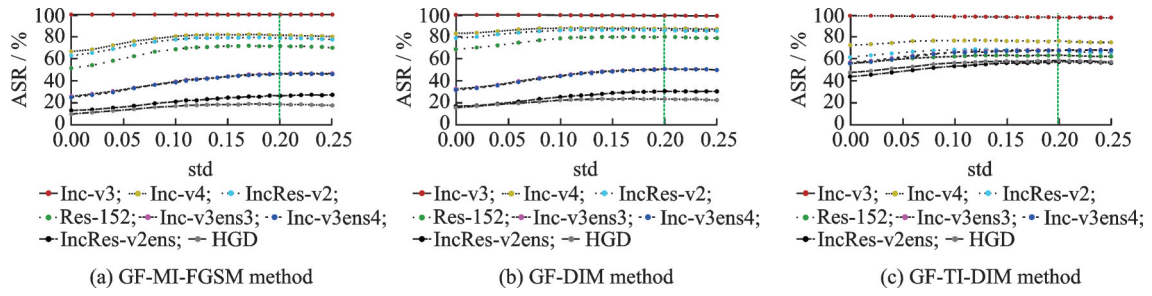


图3　高斯噪声标准差值对ASR影响

Fig.3　Influence of Gaussian noise standard deviation on ASR

3.2.2 滑动平均系数取值

滑动平均系数β用于权衡GF策略和随机调整大小填充两种变换方式影响。如果$\beta=0$,则GF-DIM降级为DIM,GF-TI-DIM降级为TI-DIM。因此,本节进行消融实验以找到适当的β值,设置滑动平均系数变化范围从0.0到1.0,变化间隔为0.1,并分别使用GF-DIM和GF-TI-DIM方法白盒攻击Inc-v3模型生成相应的对抗样本。图4显示了生成的对抗样本针对4种普通模型及和4种防御模型的攻击成功率。同样,实验中综合考虑8个模型平均攻击成功率,当β为0.6时获得最大值。因此,以下实验中将β设置为0.6。此外,$\beta=0.6$也表明了在提高对抗攻击性能方面GF策略所占比重更大,即GF策略发挥的作用要比随机调整大小填充变换方式更强。

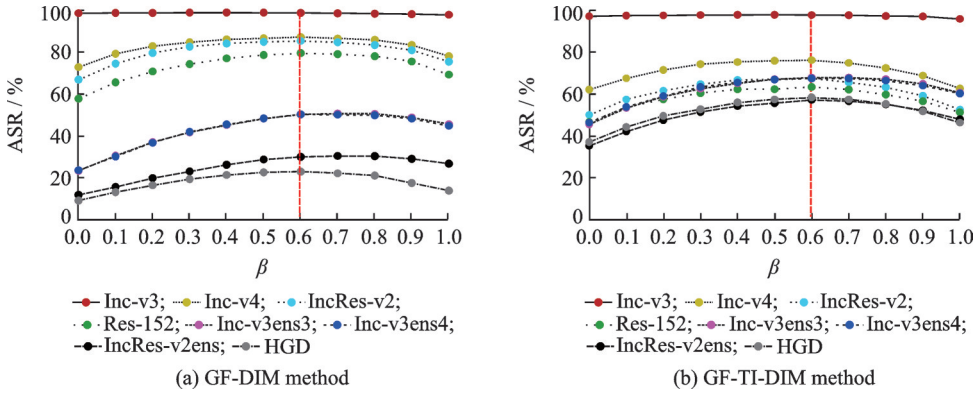


图4 滑动平均系数对ASR影响

Fig.4 Influence of moving average factor on ASR

3.2.3 高斯噪音与翻转有效性

GF策略是一种组合策略,包括了高斯噪声添加和随机水平翻转两种操作。为此,本节进行消融分析各自在提高对抗样本的黑盒攻击性能方面的有效性。首先,分别只考虑把随机水平翻转策略或者高斯噪声策略整合到诸如MI-FGSM、DIM和TI-DIM基线方法,从而形成F-MI-FGSM、F-DIM、F-TI-DIM、G-MI-FGSM、G-DIM和G-TI-DIM方法。然后,分别使用上述6种方法以及GF-MI-FGSM、GF-DIM和GF-TI-DIM方法白盒攻击Inc-v3模型生成相应的对抗样本。表1中显示了生成的对抗样本对于3种普通模型和6种防御模型的黑盒攻击成功率。从表1可以看到,对于所有黑盒攻击,高斯噪声策略要比随机水平翻转策略更强,这表明高斯噪声策略在提高对抗样本的黑盒攻击性能方面要优于随机水平翻转策略。此外,高斯噪声和翻转组合策略方式均比只考虑高斯噪声策略或者随机水平翻转策略表现得更好,从而说明高斯噪声能够与随机水平翻转有效融合,进一步增加了输入的多样性,从而得到更具迁移性的对抗样本。

3.2.4 归一化处理有效性

本文使用像素值归一化方式来处理添加了高斯噪声的样本,既能保留所添加的噪声信息,又能满足像素值限定范围。为此,本节进行消融分析像素值归一化方式在提高对抗样本的黑盒攻击性能方面的有效性。分别在像素值未归一化与归一化两种方式下采用G-MI-FGSM、G-DIM和G-TI-DIM方法白盒攻击Inc-v3模型生成相应的对抗样本。从表2可以看到,对于所有黑盒攻击,经过像素值归一化的方法均要比未经过像素值归一化的方法更强。例如,经过像素值归一化处理的G-MI-FGSM、G-DIM和G-TI-DIM方法的黑盒攻击平均成功率分别为45.1%、53.1%和61.9%。然而,未经过像素值归一化处理的G-MI-FGSM、G-DIM和G-TI-DIM方法的黑盒攻击平均成功率则为40.3%、46.9%和60.2%,

表1　高斯噪声与翻转策略对黑盒攻击成功率影响

Table 1　Influence of Gaussian noise and flipping strategy on black-box attack success rate　%

攻击方法	Inc-v4	IncRes-v2	Res-152	Inc-v3 ens3	Inc-v3 ens4	IncRes_v2ens	HGD	R&P	NIPS-r3	平均成功率
F-MI-FGSM	66.7	62.9	51.5	26.1	25.1	13.3	10.0	12.9	17.9	31.8
G-MI-FGSM	78.8	76.2	67.8	42.5	42.7	23.6	16.7	22.7	34.7	45.1
GF-MI-FGSM	**81.3**	**78.8**	**71.5**	**46.4**	**46.4**	**26.5**	**19.0**	**25.3**	**37.8**	**48.1**
F-DIM	83.1	78.9	68.8	32.9	31.7	17.1	15.8	17.6	25.7	41.3
G-DIM	87.2	84.6	78.0	47.1	47.8	27.9	21.1	28.1	40.3	51.3
GF-DIM	**87.8**	**85.9**	**80.1**	**50.9**	**50.8**	**30.6**	**23.5**	**30.2**	**42.1**	**53.5**
F-TI-DIM	72.5	61.4	55.7	56.6	55.8	43.7	47.4	43.7	48.6	53.9
G-TI-DIM	74.6	65.2	61.4	66.2	65.8	55.1	56.1	53.7	58.8	61.9
GF-TI-DIM	**76.3**	**67.6**	**63.4**	**67.9**	**67.5**	**57.3**	**58.4**	**55.5**	**60.5**	**63.8**

表2　像素值归一化对黑盒攻击成功率影响

Table 2　Influence of pixel value normalization on black-box attack success rate　%

攻击方法	Inc-v4	IncRes-v2	Res-152	Inc-v3 ens3	Inc-v3 ens4	IncRes_v2ens	HGD	R&P	NIPS-r3	平均成功率
未归一化 G-MI-FGSM	74.9	71.8	62.8	37.5	36.0	19.8	11.5	19.0	29.8	40.3
归一化G-MI-FGSM	**78.8**	**76.2**	**67.8**	**42.5**	**42.7**	**23.6**	**16.7**	**22.7**	**34.7**	**45.1**
未归一化G-DIM	84.7	82.1	74.4	41.6	41.8	23.2	14.4	23.9	36.0	46.9
归一化G-DIM	**87.2**	**84.6**	**78.0**	**47.1**	**47.8**	**27.9**	**21.1**	**28.1**	**40.3**	**53.1**
未归一化G-TI-DIM	73.4	63.5	60.0	65.2	64.3	52.9	53.6	52.1	56.9	60.2
归一化G-TI-DIM	**74.6**	**65.2**	**61.4**	**66.2**	**65.8**	**55.1**	**56.1**	**53.7**	**58.8**	**61.9**

这表明像素值归一化方式有效提升了对抗样本的黑盒攻击性能。

3.3　单模型攻击

本文将GF策略攻击方法与基线攻击方法做了单模型攻击的对比实验,即分别针对4个普通模型Inc-v3、Inc-v4、IncResv2和Res-152生成相应的对抗样本,测试其对于所有10个模型的攻击成功率。从表3～5可以看出,GF策略方法可有效改进现有基于梯度攻击方法的攻击性能,使得改进后的对抗攻击方法在保持较高的白盒攻击成功率前提下整体提升其黑盒攻击成功率。例如,GF-TI-DIM方法在Inc-v3ens3模型和Inc-v3模型上的攻击成功率分别为67.9%和97.9%。然而,相应的基线方法TI-DIM则分别为45.5%和97.2%。

虽然本文方法对样本进行了高斯噪声添加处理,但是对于HGD这种采用降噪网络从输入中消除对抗性噪声的防御方法仍然有效。从表3～5可以看出,对于HGD防御模型,该方法的成功率大大超过了基线方法,这表明所生成的对抗样本仍对HGD防御模型保持很强攻击性。

此外,本文还采用直观的方式对比了GF策略攻击方法与基线攻击方法所生成的对抗样本。由图5可知,本文方法所生成的对抗样本相比于基线攻击方法并没有增加对抗噪声尺度。

表3 MI-FGSM和GF-MI-FGSM单模型ASR对比

Table 3 Comparison of single-model ASR of MI-FGSM and GF-MI-FGSM %

模型	攻击方法	Inc-v3	Inc-v4	IncRes-v2	Res-152	Inc-v3ens3	Inc-v3ens4	IncRes_v2ens	HGD	R&P	NIPS-r3
Inc-v3	MI-FGSM	**100**	51.5	48.5	40.7	21.1	19.2	10.3	6.5	10.0	13.7
	GF-MI-FGSM	**100**	**81.3**	**78.8**	**71.5**	**46.4**	**46.4**	**26.5**	**19.0**	**25.3**	**37.8**
Inc-v4	MI-FGSM	68.8	**99.9**	54.8	49.1	23.4	22.0	12.6	10.4	13.3	17.0
	GF-MI-FGSM	**88.4**	**99.9**	**82.3**	**75.4**	**52.9**	**52.5**	**34.9**	**29.6**	**35.6**	**45.3**
IncRes-v2	MI-FGSM	74.4	63.8	**100**	54.7	30.2	27.3	19.5	19.3	19.0	22.0
	GF-MI-FGSM	**89.2**	**87.2**	99.8	**80.2**	**62.7**	**58.6**	**50.0**	**46.2**	**47.8**	**56.7**
Res-152	MI-FGSM	56.4	49.6	47.4	**99.5**	26.4	26.1	15.8	18.0	15.5	18.8
	GF-MI-FGSM	**77.8**	**73.4**	**72.6**	99.4	**53.1**	**51.9**	**37.8**	**39.6**	**37.6**	**45.8**

表4 DIM和GF-DIM单模型ASR对比

Table 4 Comparison of single-model ASR of DIM and GF-DIM %

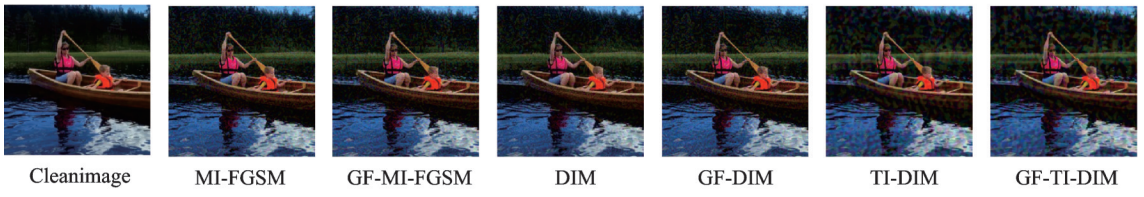
模型	攻击方法	Inc-v3	Inc-v4	IncRes-v2	Res-152	Inc-v3ens3	Inc-v3ens4	IncRes_v2ens	HGD	R&P	NIPS-r3
Inc-v3	DIM	**99.4**	72.5	67.1	57.7	24.1	23.7	12.4	9.7	13.5	19.7
	GF-DIM	99.3	**87.8**	**85.9**	**80.1**	**50.9**	**50.8**	**30.6**	**23.5**	**30.2**	**42.1**
Inc-v4	DIM	81.7	**99.6**	73.1	63.6	26.8	26.8	16.0	15.2	17.3	23.8
	GF-DIM	**91.3**	99.4	**87.6**	**82.4**	**57.1**	**55.7**	**38.9**	**35.4**	**41.3**	**50.9**
IncRes-v2	DIM	85.5	83.5	**98.4**	73.0	40.5	37.4	25.5	29.8	29.7	36.3
	GF-DIM	**91.3**	**90.1**	**98.4**	**85.0**	**67.4**	**62.9**	**54.4**	**53.4**	**54.9**	**63.3**
Res-152	DIM	82.0	79.8	77.7	98.5	41.7	39.2	26.0	33.9	29.3	36.2
	GF-DIM	**88.0**	**85.6**	**84.4**	**99.2**	**62.2**	**60.5**	**46.9**	**51.8**	**48.8**	**58.0**

表5 TI-DIM和GF-TI-DIM单模型ASR对比

Table 5 Comparison of single-model ASR of TI-DIM and GF-TI-DIM %

模型	攻击方法	Inc-v3	Inc-v4	IncRes-v2	Res-152	Inc-v3ens3	Inc-v3ens4	IncRes_v2ens	HGD	R&P	NIPS-r3
Inc-v3	TI-DIM	97.2	61.7	50.5	47.3	45.5	46.2	34.9	37.2	37.5	40.4
	GF-TI-DIM	**97.9**	**76.3**	**67.6**	**63.4**	**67.9**	**67.5**	**57.3**	**58.4**	**55.5**	**60.5**
Inc-v4	TI-DIM	69.1	**98.7**	56.0	49.9	48.6	48.1	39.0	41.0	40.7	43.0
	GF-TI-DIM	**79.5**	98.3	**70.2**	**62.5**	**67.1**	**67.5**	**58.7**	**59.9**	**58.1**	**61.3**
IncRes-v2	TI-DIM	73.9	71.0	**95.7**	61.2	60.6	60.5	59.2	59.2	63.1	61.9
	GF-TI-DIM	**80.9**	**79.1**	95.6	**70.3**	**75.1**	**75.0**	**75.0**	**71.6**	**73.4**	**74.2**
Res-152	TI-DIM	64.7	60.3	57.1	95.8	58.4	59.0	51.8	53.8	53.2	55.2
	GF-TI-DIM	**71.5**	**67.2**	**63.9**	**97.4**	**69.8**	**70.6**	**64.9**	**64.4**	**64.0**	**66.5**

(a) Adversarial example 1



(b) Adversarial example 2

(c) Adversarial example 3

图5　对抗样本

Fig.5　Adversarial examples

3.4　集成模型攻击

尽管生成的对抗样本的黑盒攻击成功率得到了提升,但在攻击防御模型方面仍然相对较弱。为此本文采用了集成模型攻击方法来进一步提升对抗样本的攻击性能。首先,将4个普通模型Inc-v3、Inc-v4、Res-152和IncRes-v2融合为集成模型。然后,采用GF策略攻击方法与基线攻击方法进行集成模型攻击对比。在实验中,集成模型里4个模型的权重值设置为相同。

从表6中可以看出,GF策略集成攻击算法在攻击防御模型的效果上均要优于基线集成攻击算法。同时,MGF-TI-DIM算法所生成的对抗样本黑盒攻击性能是所有集成攻击算法中最强的,能以86.2%的平均成功率欺骗6种先进防御模型。相比于目前最强黑盒攻击方法M-TI-DIM算法,MGF-TI-DIM算法在攻击成功率上提升约8.0%。

表6　集成模型ASR对比

Table 6　Comparison of ensemble-model ASR　%

攻击方法	白盒平均成功率	Inc-v3ens3	Inc-v3ens4	IncRes_v2ens	HGD	R&P	NIPS-r3	黑盒平均成功率
M-MI-FGSM	**99.70**	51.30	48.60	32.60	39.90	32.90	44.00	41.60
MGF-MI-FGSM	99.30	**83.90**	**80.60**	**71.10**	**71.50**	**72.10**	**80.10**	**76.60**
M-DIM	98.30	64.30	60.10	43.20	54.60	51.20	64.40	56.30
MGF-DIM	**98.60**	**85.80**	**82.30**	**73.50**	**74.90**	**75.90**	**82.10**	**79.10**
M-TI-DIM	94.50	79.70	78.90	75.00	78.30	78.20	79.20	78.20
MGF-TI-DIM	**96.20**	**88.30**	**87.90**	**84.90**	**85.60**	**84.30**	**86.10**	**86.20**

此外,为了表明在对抗攻击性能方面GF策略要比其他数据增强技术攻击方法[6-7]更强,本文做了单模型攻击和集成模型攻击下的攻击成功率对比。如图6所示,DIM代表文献[6]方法,TI-MI-FGSM代表文献[7]方法,GF-MI-FGSM则为本文方法。由图6可以看出,本文方法在单模型和集成模型下黑盒攻击性能均优于文献[6-7]方法,从而进一步验证GF策略方法的有效性。

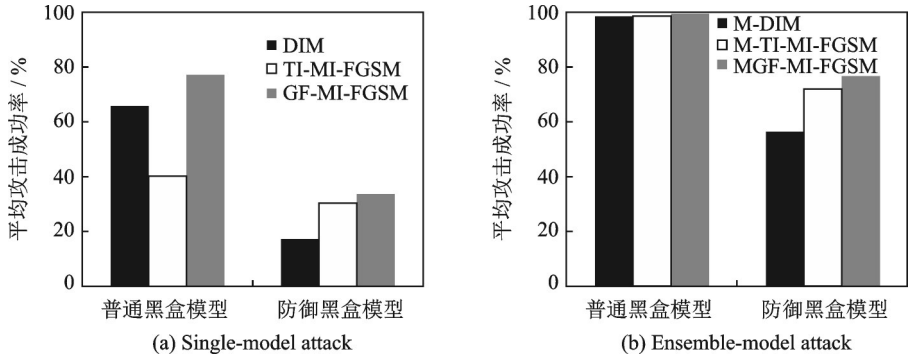


图6 攻击成功率对比

Fig.6 Comparison of attack success rate

4 结束语

本文针对如何整体提升普通和防御模型的黑盒攻击成功率问题,提出一种基于高斯噪声和翻转组合策略方法来增强对抗样本的可迁移性,从而提升对抗攻击能力。在NIPS 2017对抗竞赛的ImageNet数据集做了对比实验,实验结果表明所生成的对抗样本对于普通和防御黑盒模型均具有更强的攻击性,并且在白盒攻击方面仍能保持较高的成功率。同时,为了进一步增强算法在黑盒攻击中的攻击性能,本文采用了集成模型攻击方法,实现了以86.2%的平均成功率欺骗6种先进黑盒防御模型。下一步将继续研究如何将高斯噪声和翻转组合策略增强技术应用到诸如基于生成网络和基于优化的攻击方法中。

参考文献:

[1] SZEGEDY C, ZAREMBA W, SUTSKEVER I, et al. Intriguing properties of neural networks[EB/OL]. (2013-12-23)[2020-05-21]. https://arxiv.org/abs/1312.6199.

[2] 王伟, 董晶, 何子文,等. 视觉对抗样本生成技术概述[J]. 信息安全学报, 2020, 5(2): 39-48.
WANG Wei, DONG Jing, HE Ziwen, et al. A brief introduction to visual adversarial samples[J]. Journal of Cyber Security, 2020, 5(2): 39-48.

[3] BRENDEL W, RAUBER J, BETHGE M. Decision-based adversarial attacks: Reliable attacks against black-box machine learning models[C]//Proceedings of the 6th International Conference on Learning Representations. [S.l.]: OpenReview.net, 2018: 87-89.

[4] PAPERNOT N, MCDANIEL P, GOODFELLOW I J, et al. Practical black-box attacks against deep learning systems using adversarial examples[EB/OL]. (2016-02-08)[2020-06-25]. https://arxiv.org/abs/1602.02697.

[5] DONG Y P, LIAO F Z, PANG T Y, et al. Boosting adversarial attacks with momentum[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.[S.l.]: IEEE, 2018: 9185-9193.

[6] XIE C H, ZHANG Z S, ZHOU Y Y, et al. Improving transferability of adversarial examples with input diversity[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.[S.l.]: IEEE, 2019: 2730-2739.

[7] 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 Conference on Computer Vision and Pattern Recognition.[S.l.]: IEEE, 2019: 4312-4321.

[8] KRIZHEVSKY A, SUTSKEVER I, HINTON G. ImageNet classification with deep convolutional neural networks[C]//Advances in Neural Information Processing Systems. [S.l]: [s.n.], 2012: 1106-1114.

[9] GOODFELLOW I J, SHLENS J, SZEGEDY C. Explaining and harnessing adversarial examples[EB/OL]. (2014-12-20)[2020-05-25]. https://arxiv.org/abs/1412.6572.

[10] KURAKIN A, GOODFELLOW S I J, BENGIO S. Adversarial examples in the physical world[EB/OL]. (2016-07-08)[2020-06-02]. https://arxiv.org/abs/1607.02533.

[11] SHORTEN C, KHOSHGOFTAAR T M. A survey on image data augmentation for deep learning[J]. Journal of Big Data, 2019, 6(1): 60.

[12] LIU Y P, CHEN X Y, LIU C, et al. Delving into transferable adversarial examples and black-box attacks[EB/OL]. (2016-11-08)[2020-06-10]. https://arxiv.org/abs/1611.02770.

[13] SZEGEDY C, VANHOUCKE V, IOFFE S, et al. Rethinking the inception architecture for computer vision[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. [S.l.]: IEEE, 2016: 2818-2826.

[14] SZEGEDY C, IOFFE S, VANHOUCKE V, et al. Inception-v4, inception-resnet and the impact of residual connections on learning[C]//Proceedings of the 31st AAAI Conference on Artificial Intelligence. [S.l.]: AAAI Press, 2017: 4278-4284.

[15] HE K M, ZHANG X Y, REN S Q, et al. Identity mappings in deep residual networks[C]//Proceedings of European Conference on Computer Vision. [S.l.]: Springer, 2016: 630-645.

[16] TRAMER F, KURAKIN A, PAPERNOT N, et al. Ensemble adversarial training: Attacks and defenses[EB/OL]. (2017-05-19)[2020-06-18]. https://arxiv.org/abs/1705.07204.

[17] LIAO F Z, LIANG M, DONG Y P, et al. Defense against adversarial attacks using high-level representation guided denoiser [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. [S.l.]: IEEE, 2018: 1778-1787.

[18] XIE C H, WANG J Y, ZHANG Z S, et al. Mitigating adversarial effects through randomization[EB/OL]. (2017-11-06)[2020-06-15]. https://arxiv.org/abs/1711.01991.

作者简介:

张武(1990-),男,硕士研究生,研究方向:对抗机器学习,E-mail:18921132158@sina.cn。

段晔鑫(1987-),男,博士研究生,研究方向:对抗机器学习与图像识别,E-mail:duanyexin0713@163.com。

邹军华(1991-),男,博士研究生,研究方向:计算机视觉、对抗机器学习,E-mail:278287847@qq.com。

潘志松(1973-),**通信作者**,男,教授,研究方向:模式识别与机器学习,E-mail:hotpzs@hotmail.com。

周星宇(1985-),男,讲师,研究方向:计算机视觉与对抗样本,E-mail:universe-zhou@sina.cn。

(编辑:刘彦东)