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

邹军华[1],段晔鑫[1,2],任传伦[3],邱俊洋[4],周星宇[1],潘志松[1]
(1. 陆军工程大学指挥控制工程学院,江苏南京 210007;2. 陆军军事交通学院镇江校区,江苏镇江 212003;3. 华北计算技术研究所,北京 100083;4. 江南计算所数字工程与先进计算国家重点实验室,江苏无锡 214083)

摘 要: 深度神经网络在多种模式识别任务上都取得了巨大突破,但相关研究表明深度神经网络存在脆弱性,容易被精心设计的对抗样本攻击. 本文以分类任务为着手点,研究对抗样本的迁移性,提出基于噪声初始化、Adam-Nesterov 方法和准双曲动量方法的对抗样本生成方法. 本文提出一种对抗噪声的初始化方法,通过像素偏移方法来预先增强干净样本的攻击性能. 同时,本文使用 Adam-Nesterov 方法和准双曲动量方法来改进现有方法中的 Nesterov 方法和动量方法,实现更高的黑盒攻击成功率. 在不需要额外运行时间和运算资源的情况下,本文方法可以和其他的攻击方法组合,并显著提高了对抗样本的黑盒攻击成功率. 实验表明,本文的最强攻击组合为 ANI-TI-DIQHM*(其中*代表噪声初始化),其对经典防御方法的平均黑盒攻击成功率达到 88.68%,对较为先进的防御方法的平均黑盒攻击成功率达到 82.77%,均超过现有最高水平.

关键词: 对抗样本; Adam-Nesterov 方法; 准双曲动量方法; 噪声初始化; 迁移性能
中图分类号: TP391 **文献标识码:** A **文章编号:** 0372-2112(2022)01-0207-10
电子学报 URL:http://www.ejournal.org.cn **DOI:**10.12263/DZXB.20200839

Perturbation Initialization, Adam-Nesterov and Quasi-Hyperbolic Momentum for Adversarial Examples

ZOU Jun-hua[1], DUAN Ye-xin[1,2], REN Chuan-lun[3], QIU Jun-yang[4], ZHOU Xing-yu[1], PAN Zhi-song[1]
(1. *Command and Control Engineering College, Army Engineering University of PLA, Nanjing, Jiangsu* 210007, *China*;
2. *Zhenjiang Campus, Army Military Transportation University of PLA, Zhenjiang, Jiangsu* 212003, *China*;
3. *North China Institute of Computer Technology, Beijing* 100083, *China*;
4. *Mathematical Engineering and Advanced Computing, Jiangnan Institute of Computing Technology, Wuxi, Jiangsu* 214083, *China*)

Abstract: Deep neural networks(DNNs) have made great breakthrough in many pattern recognition tasks. However, relevant research shows that the DNNs are vulnerable to adversarial examples. In this paper, we study the transferability of adversarial examples in the classification task, and propose perturbation initialization, the quasi-hyperbolic momentum iterative fast gradient sign method(QHMI-FGSM) and the adam-nesterov iterative fast gradient sign method(ANI-FGSM). We propose perturbation initialization method called pixel shift in adversarial attack. Furthermore, QHMI-FGSM and ANI-FGSM proposed in this paper are the improvements on the existing momentum iterative fast gradient sign method(MI-FGSM) and nesterov iterative fast gradient sign method(NI-FGSM). Additionally, perturbation initialization, QHMI-FGSM and ANI-FGSM are easily integrated into other existing methods, which can significantly improve the success rates of black-box attacks without additional running time and computing resources. Experimental results show that our best attack ANI-TI-DIQHM* can fool six classic black-box defense models with an average success rate of 88.68%, and fool four advance black-box defense models with an average success rate of 82.77%, which are higher than the state-of-the-art results.

Key words: adversarial examples; Adam-Nesterov method; quasi-hyperbolic momentum method; perturbation initialization; transferability

收稿日期:2020-08-04;修回日期:2021-01-22;责任编辑:王天慧
基金项目:国家自然科学基金(No.62076251)

1 引言

深度神经网络(Deep Neural Networks,DNNs)在图像分类[1]、目标检测[2]等领域取得了巨大突破,但相关研究表明DNNs存在着脆弱性,容易被精心设计的对抗样本[3]所攻击.进一步的研究表明,对抗样本具有迁移性[4],即针对某个DNN生成的对抗样本,同样可以让其他未知的DNNs输出错误结果.对抗样本还能威胁现实应用[5],因此大量研究致力于提高DNNs的防御能力,如对抗训练[6]、样本去噪声[7]、样本转换[8]和其他方法[9].综上所述,对于对抗样本迁移性的研究,有助于提高DNNs的鲁棒性,并使得现实应用更加可靠.

Foolbox[10]将对抗样本的生成方法分为3种:基于梯度的方法[11]、基于分数的方法[12]、基于输出的方法[13].其中基于梯度的生成方法主要依靠对抗样本的迁移性来实现对黑盒DNNs的攻击.本文主要研究对抗样本的迁移性,具体为分类任务中基于梯度的对抗样本生成方法.现有方法可以相互组合,形成更具迁移性能的攻击.例如,现有较强的攻击组合NI-TI-DIM由Nesterov算法[14]、动量算法[11]、样本多样化方法[15]和平移不变方法[16]组合而成.

目前,随机噪声初始化[10]是仅有的对抗噪声初始化方法.本文提出噪声初始化方法,通过像素偏移方法来预先增强干净样本的攻击性能.同时,本文提出基于Adam-Nesterov方法和准双曲动量方法的对抗样本生成方法,以对抗样本的迁移性能.现有的Nesterov算法[14]可理解为标准动量在求解梯度之前添加了一个临时的校正因子,但每次迭代中的Nesterov动量共享一个相同的学习率.而本文基于Adam-Nesterov方法的对抗样本生成方法,可以自适应地调整学习率,且Nesterov动量中的每个权值都有独立的学习率.此外,本文将准双曲动量算法用于对抗样本生成,取代常规动量算法[11].以NI-TI-DIM为例,对抗样本生成框架及本文方法所改进的位置如图1所示.本文在梯度计算前,将噪声初始化操作作为一个模块加入其中,并用准双曲动量算法和Adam-Nesterov方法分别取代动量方法[11]和Nesterov算法[14].实验表明,结合了本文方法的攻击组合能生成攻击成功率更高的对抗样本.同时,实验表明,3种方法都没有额外增加对抗样本生成所需的运行时间和运算资源.

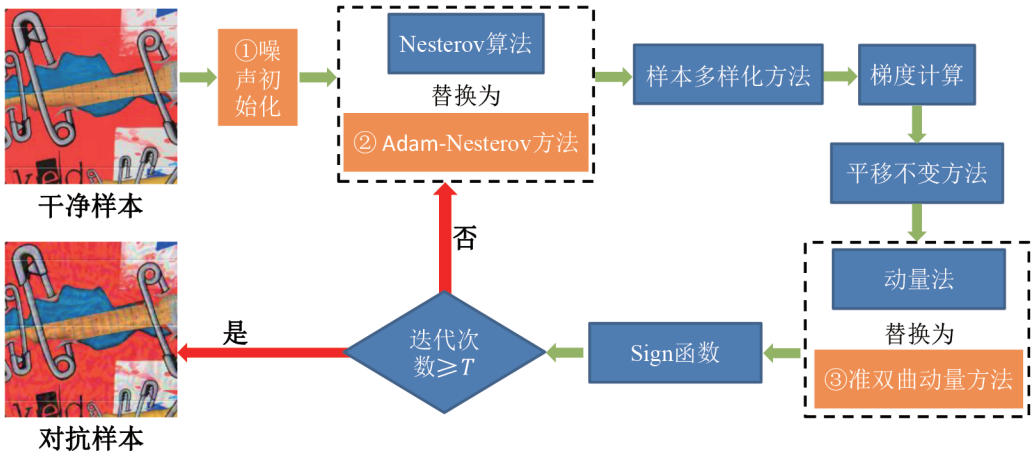

图1 本文方法框图

2 相关工作

2.1 对抗样本问题的定义

对于一个已知训练好的深度分类器$f(x)$:$x\in\mathcal{X}\rightarrow y\in\mathcal{Y}$,向其输入干净样本$x$,分类器输出正确的标签$y$.对抗攻击是在干净样本$x$邻域找出一个对抗样本$x^{\text{adv}}$,使得分类器输出错误的标签.对抗攻击分为无目标和有目标攻击,其中,无目标对抗样本能使得分类器的输出标签不等于正确标签,即$f(x^{\text{adv}})\neq y$,有目标对抗样本能使得分类器的输出标签等于目标错误标签y^{target},即$f(x^{\text{adv}})=y^{\text{target}}\neq y$.通常情况下,为了让干净样本$x$和对抗样本$x^{\text{adv}}$难以通过人眼进行区分,攻击者会将干净样本$x$和对抗样本$x^{\text{adv}}$之间的$L_p$距离限制在足够在足够小的范围$\varepsilon$内,即$\|x^{\text{adv}}-x\|_p\leqslant\varepsilon$,其中$p$可以是0,1,2或者$\infty$.本文主要关注$L_\infty$条件下的无目标攻击方法.

对抗样本具有迁移性,以无目标攻击为例,针对深度分类器$f_1(x)$生成的对抗样本x^{adv},不仅可以使$f_1(x)$输出错误的标签$f_1(x^{\text{adv}})\neq y$,还可以使其他未知模型$f_2(x),f_3(x),\cdots,f_n(x)$输出错误的标签$f_2(x^{\text{adv}})\neq y,f_3(x^{\text{adv}})\neq y,\cdots,f_n(x^{\text{adv}})\neq y$.

2.2 基于梯度的对抗样本生成方法

2.2.1 快速梯度符号方法

Goodfellow等人提出的快速梯度符号方法(Fast Gradient Sign Method, FGSM)[3]解决了对抗样本生成速度过慢的问题. FGSM通过最大化损失函数$J(x,y)$来找出相应的对抗样本:

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

其中,$\nabla_x J(x,y)$是损失函数对于x的梯度,ε是干净样本x和对抗样本x^{adv}的L_∞距离的限制阈值.

2.2.2 多次迭代的快速梯度符号方法

Kurakin等提出多次迭代的快速梯度符号方法(Iterative Fast Gradient Sign Method, I-FGSM)[17],解决了FGSM在白盒攻击中成功率过低的问题. I-FGSM以更小的步长α,通过T次迭代的方式重复快速梯度方法,从而找出白盒攻击能力更强的对抗样本.

$$x_0^{\text{adv}} = x \tag{2}$$

$$x_{t+1}^{\text{adv}} = \text{Clip}_{x,\varepsilon}\left\{x_t^{\text{adv}} + \alpha\text{sign}\left(\nabla_{x_t^{\text{adv}}} J\left(x_t^{\text{adv}},y\right)\right)\right\} \tag{3}$$

其中,α为步长. 对抗样本通过$\text{Clip}_{x,\varepsilon}\{\bullet\}$方程满足$L_\infty$限制条件,并限制对抗样本的每一个像素点于区间$[0,255]$内. $\text{Clip}_{x,\varepsilon}\{\cdot\}$的定义为

$$\text{Clip}_{x,\varepsilon}\left\{x^{\text{adv}}\right\} = \min\left\{255, x+\varepsilon, \max\left\{0, x-\varepsilon, x^{\text{adv}}\right\}\right\} \tag{4}$$

尽管I-FGSM在白盒攻击方面性能卓越,但在黑盒攻击方面却远差于FGSM.

2.2.3 基于动量方法的多次迭代快速梯度符号方法

Dong等提出基于动量方法的多次迭代快速梯度符号方法(Momentum Iterative Fast Gradient Sign Method, MI-FGSM)[11],缓解I-FGSM迁移性能过低的问题. MI-FGSM将优化算法中的动量算法应用于对抗样本生成中,其更新过程为

$$x_0^{\text{adv}} = x, g_0 = 0 \tag{5}$$

$$g_{t+1} = \mu g_t + \frac{\nabla_{x_t^{\text{adv}}} J\left(x_t^{\text{adv}},y\right)}{\nabla_{x_t^{\text{adv}}} J\left(x_t^{\text{adv}},y\right)_1} \tag{6}$$

$$x_{t+1}^{\text{adv}} = \text{Clip}_{x,\varepsilon}\left\{x_t^{\text{adv}} + \alpha\text{sign}\left(g_{t+1}\right)\right\} \tag{7}$$

其中,g_{t+1}为前t次迭代中累加的梯度,μ为动量系数.

2.2.4 基于Nesterov算法的多次迭代快速梯度符号方法

Lin等提出基于Nesterov算法的多次迭代快速梯度符号方法(Nesterov Iterative Fast Gradient Sign Method, NI-FGSM)[14],增强了对抗样本的迁移性能. 初始化$x_0^{\text{adv}} = x, g_0 = 0$后,其过程为

$$x_t^{\text{nes}} = x_t^{\text{adv}} + \alpha\mu g_t \tag{8}$$

$$g_{t+1} = \mu g_t + \frac{\nabla_{x_t^{\text{nes}}} J\left(x_t^{\text{nes}},y\right)}{\nabla_{x_t^{\text{nes}}} J\left(x_t^{\text{nes}},y\right)_1} \tag{9}$$

$$x_{t+1}^{\text{adv}} = \text{Clip}_{x,\varepsilon}\left\{x_t^{\text{adv}} + \alpha\text{sign}\left(g_{t+1}\right)\right\} \tag{10}$$

其中,x_t^{nes}为Nesterov项,只参与梯度计算,不参与噪声叠加.

2.2.5 集成学习方法

Dong等通过集成学习联合多个模型共同生成对抗样本[11],其核心为融合所有K个模型的logits,并通过标签和融合的logits计算新的交叉熵损失.

$$l(x) = \sum_{k=1}^{K} w_k l_k(x) \tag{11}$$

$$J(x,y) = -1_y \cdot \log\left(\text{softmax}\left(l(x)\right)\right) \tag{12}$$

其中,$l(x)$表示第k个模型的logits,w_k表示集成系数,-1_y表示标签的独热编码. 集成学习方法能大大提升对抗样本的迁移性能,但也增加了对抗样本生成的时间和资源.

2.2.6 样本多样化方法

Xie等提出了样本多样化方法(Diverse Input Method, DIM)[15],在每次迭代中,提前对输入样本进行随机的多样化转换. 其过程为

$$T\left(x_t^{\text{adv}},p,s\right) = \begin{cases} T\left(x_t^{\text{adv}},s\right), \text{以概率}p\text{执行} \\ x_t^{\text{adv}}, \text{以概率}1-p\text{执行} \end{cases} \tag{13}$$

其中,s表示多样化转换后的样本大小,p表示执行转换的概率.

2.2.7 平移不变方法

Dong等提出了平移不变方法(Translation-Invariant Method, TIM)[16],在每次迭代中,通过集成多个平移单个像素的样本来提升对抗样本迁移性能. 同时,为了解决效率问题,Dong等将这种样本的集成等价为对梯度信息的高斯模糊. 梯度信息的高斯模糊过程为

$$\nabla^{\star}_{x_t^{\text{adv}}} J\left(x_t^{\text{adv}},y\right) \approx W * \nabla_{x_t^{\text{adv}}} J\left(x_t^{\text{adv}},y\right) \tag{14}$$

其中,W为一个预定义的高斯核.

2.2.8 尺度不变方法

Lin等提出尺度不变方法(Scale-Invariant Method, SIM)[14],这种方法相当于在每次迭代中对输入样本进行数据增强,然后进行数据集成,最后进行梯度计算. 这种方法大大提高了对抗样本生成所需的时间和资源,违背了快速梯度符号方法的样本快速生成初衷.

3 本文算法

3.1 对抗样本噪声初始化

深度学习中,对网络权重进行初始化有利于模型的收敛. 目前对抗噪声初始化方法仅有随机噪声初始

化,本文使用像素偏移方法对噪声进行初始化处理.

$$x' = \sum_{i,j} w_{ij} T_{ij}(x) \tag{15}$$

$$x_{\text{init}} = \text{Clip}_{x,\varepsilon}\{x'\} \tag{16}$$

其中,$T_{ij}(x)$表示将图像x位于(a,b)位置的像素值变换为$(a-i,b-j)$位置的像素值,且i,j取值范围为$\{-k,\cdots,0,\cdots,k\}$,w_{ij}为每次变换的权重,而$\text{Clip}_{x,\varepsilon}\{x'\}$限制$x'$的范围并令$x_{\text{init}}$满足$\left\| x_{\text{init}} - x \right\|_\infty \leqslant \varepsilon$.

3.2 基于准双曲动量方法的多次迭代快速梯度符号方法

Ma等在优化领域提出了准双曲动量方法(Quasi-Hyperbolic Momentum,QHM)[18],对比传统的动量方法,QHM引入了滑动平均系数v,其更新过程为

$$g_{t+1} \leftarrow \beta g_t + (1-\beta)\nabla \hat{L}_t(\theta_t) \tag{17}$$

$$\theta_{t+1} \leftarrow \theta_t - \alpha\left[(1-v)\nabla \hat{L}_t(\theta_t) + v g_{t+1}\right] \tag{18}$$

其中,β为动量系数.

本文将QHM用于对抗样本生成,取代原有的MI-FGSM,形成基于准双曲动量方法的多次迭代快速梯度符号方法(Quasi-Hyperbolic Momentum Iterative Fast Gradient Sign Method,QHMI-FGSM). QHMI-FGSM将式(6)、式(7)转化为

$$g_{t+1} = \beta g_t + (1-\beta)\frac{\nabla_{x_t^{\text{adv}}} J(x_t^{\text{adv}}, y)}{\nabla_{x_t^{\text{adv}}} J(x_t^{\text{adv}}, y)_1} \tag{19}$$

$$\tilde{g}_{t+1} = (1-v)\, g_{t+1} + v\frac{\nabla_{x_t^{\text{adv}}} J(x_t^{\text{adv}}, y)}{\nabla_{x_t^{\text{adv}}} J(x_t^{\text{adv}}, y)_1} \tag{20}$$

$$x_{t+1}^{\text{adv}} = \text{Clip}_{x,\varepsilon}\left\{x_t^{\text{adv}} + \alpha \text{sign}\left(\tilde{g}_{t+1}\right)\right\} \tag{21}$$

其中,$\tilde{g}_{t+1}$为准双曲动量方法对梯度信息处理后的输出.

3.3 基于Adam-Nesterov方法的多次迭代快速梯度符号方法

贾熹滨等在优化领域提出了AdaDelta-Nesterov动量方法[19],这种方法通过梯度的均方根(Root Mean Squared,RMS),对学习率进行了自适应约束,其过程为

$$E\left[\Delta\theta^2\right]_t = \rho E\left[\Delta\theta^2\right]_{t-1} + (1-\rho)\Delta\theta^2 \tag{22}$$

$$RMS[\theta]_t = \sqrt{E\left[\Delta\theta^2\right]_t + \epsilon} \tag{23}$$

$$\Delta\theta = \partial[\Delta\theta]_{t-1} - \text{RMS}[\theta]_{t-1}\nabla \hat{L}_t(\theta_t) \tag{24}$$

其中,$E\left[\Delta\theta^2\right]_t$表示前$t-1$次迭代所有梯度的平方和,$\text{RMS}[\theta]_t$表示前$t-1$次迭代所有梯度的均方根,$\rho$表示滑动平均系数,$\epsilon$表示极小值.

本文将AdaDelta-Nesterov方法应用于对抗样本生成中,形成基于AdaDelta-Nesterov多次迭代快速梯度符号方法(AdaDelta-Nesterov Iterative Fast Gradient Sign Method,ADNI-FGSM). ADNI-FGSM在NI-FGSM的基础上融入了自适应学习率,将式(8)优化为

$$E\left[g^2\right]_t = \rho E\left[g^2\right]_{t-1} + (1-\rho)\, g_t^2 \tag{25}$$

$$\text{RMS}\left[g\right]_t = \sqrt{E\left[g^2\right]_t + \epsilon} \tag{26}$$

$$x_t^{\text{nes}} = x_t^{\text{adv}} + \frac{\alpha}{\text{RMS}\left[g\right]_t} g_t \tag{27}$$

本文在ADNI-FGSM的基础上,进一步提出基于Adam-Nesterov多次迭代快速梯度符号方法(Adam-Nesterov Iterative Fast Gradient Sign Method,ANI-FGSM),用于生成对抗样本. 比较Adam[20]和AdaDelta[21],Adam在AdaDelta的基础上融入动量法,并修正一阶和二阶动量估计的偏差,ANI-FGSM可表示为

$$m_t = \beta_1 m_{t-1} + (1-\beta_1)\, g_t \tag{28}$$

$$v_t = \beta_2 v_{t-1} + (1-\beta_2)\, g_t^2 \tag{29}$$

$$\hat{m}_t = m_t/(1-\beta_1^t) \tag{30}$$

$$\hat{v}_t = v_t/(1-\beta_2^t) \tag{31}$$

$$g_t^* = \nabla_{x_t^{\text{adv}}} J\left(x_t^{\text{adv}} + \frac{\alpha}{\sqrt{\hat{v}_t} + \epsilon}\hat{m}_t, y\right) \tag{32}$$

其中,m_t和v_t分别为一阶和二阶动量估计,$\hat{m}_t$和$\hat{v}_t$分别为一阶和二阶动量估计的修正项,β_1和β_2分别为动量系数,β_1^t和β_2^t分别为动量项修正系数. 实验中,本文令$\beta_1 = \beta_1^t, \beta_2 = \beta_2^t$.

3.2节和3.3节的QHMI-FGSM和ANI-FGSM属于图1中对抗样本生成框架中的两个部分. QHMI-FGSM进行梯度运算后用于噪声叠加,而ANI-FGSM用于Nesterov项的生成.

3.4 对抗样本生成算法

本节以ANI-TI-DIQHM*(噪声初始化(3.1节)、ANI-FGSM(3.3节)、TIM、DIM、QHMI-FGSM(3.2节)的组合)为例,其详细过程如算法1所示.

算法1 ANI-TI-DIQHM*

输入:干净样本x,扰动量大小ε,步长α,迭代次数T,系数v,β,β_1和β_2,样本转换大小s,转换概率p,高斯核W,极小值ϵ,噪声初始化像素偏移范围k

输出:对抗样本x^{adv}

1. $\alpha = \varepsilon/T, g_0 = 0, m_0 = 0, v_0 = 0$
2. 噪声初始化生成x_{init},并令$x_0^{\text{adv}} = x_{\text{init}}$
3. FOR $t = 0$ TO $T-1$ DO
4. $m_t = \beta_1 m_{t-1} + (1-\beta_1)\, g_t$

5. $v_t = \beta_2 v_{t-1} + (1-\beta_2)\, g_t^2$

6. $\hat{m}_t = m_t / (1-\beta_1^t)$

7. $\hat{v}_t = v_t / (1-\beta_2^t)$

8. $x_t^{\text{nes}} = x_t^{\text{adv}} + \dfrac{\alpha}{\sqrt{\hat{v}_t} + \epsilon} \hat{m}_t$

9. $g_t^* = W * \nabla_{x_t^{\text{adv}}} J\left(T\left(x_t^{\text{nes}}, p, s\right), y\right)$

10. $g_{t+1} = \beta g_t + (1-\beta) \dfrac{g_t^*}{g_t^*{}_1}$

11. $\tilde{g}_{t+1} = (1-v)\, g_{t+1} + v \dfrac{g_t^*}{g_t^*{}_1}$

12. $x_{t+1}^{\text{adv}} = \text{Clip}_{x,\varepsilon}\left\{x_t^{\text{adv}} + \alpha \text{sign}\left(\tilde{g}_{t+1}\right)\right\}$

13. END FOR

14. $x^{\text{adv}} = x_{t+1}^{\text{adv}}$

4　实验及结果分析

4.1　实验目标

如图1所示,基于本文方法,实验目标为以下4个方面:

(1)通过对比生成时间,验证本文方法对对抗样本生成效率的影响;

(2)通过实验,比较节3.3中2种方法ADNI-FGSM和ANI-FGSM的优劣;

(3)通过消融实验,验证本文方法对对抗样本迁移性能的影响;

(4)通过对比现有最好的攻击方法,验证本文方法的有效性.

4.2　实验设置

4.2.1　数据集

本文实验中使用的1000张样本取自ImageNet的测试集,同时也与NIPS 2017对抗大赛中使用的数据集相同. 实验中所有输入干净样本和输出对抗样本的大小均为299 × 299 × 3.

4.2.2　模型

实验共涉及13个模型,其中4个为Inception v3 (Inc-v3)[22], Inception v4 (Inc-v4), Inception ResNet v2 (IncRes-v2)[23]和ResNet v2-101 (Res-v2-101)[1],作为白盒模型用于生成对抗样本. 另外9个为Inc-v3ens3, Inc-v3ens4, IncResv2ens[6], NIPS 2017对抗大赛中排名前三的防御方法(HGD[7]、R&P[8]、NIPS-r3*), Feature Distillation[24], Comdefend[25]和Randomized Smoothing[26],作为黑盒模型用于测试对抗样本. 防御模型中, Inc-v3ens3, Inc-v3ens4, IncResv2ens, HGD、R&P和NIPS-r3是经典的防御方法,用于全部实验. 而Feature Distillation, Comdefend和Randomized Smoothing是目前较为先进的防御方法,用于测试实验中较强攻击.

4.2.3　攻击组合

通常情况下,对不同方法进行组合能增强对抗样本的迁移性能. 本文实验的组合都是节2.2中不同方法的组合,并与本文的三种方法进行横向比较. 各个攻击组合的具体解释如表1所示. 本文实验均在TI-DIM和NI-TI-DIM这两个较强攻击组合的基础上进行,通过对抗样本生成效率、消融实验、单模型黑盒攻击和多模型集成黑盒攻击这四个方面,比较不同攻击组合的运行时间和黑盒攻击成功率.

表1　攻击组合简称及其定义

攻击组合简称	定义
TI-DIM(基准方法)	TIM, DIM和MI-FGSM的组合
TI-DIQHM	TIM, DIM和QHMI-FGSM的组合
TI-DIQHM*	TIM, DIM, QHMI-FGSM和噪声初始化的组合
NI-TI-DIM(基准方法)	TIM, DIM, NI-FGSM和MI-FGSM的组合
ADNI-TI-DIQHM	TIM, DIM, ADNI-FGSM和QHMI-FGSM的组合
ANI-TI-DIQHM	TIM, DIM, ANI-FGSM和QHMI-FGSM的组合
ANI-TI-DIQHM*	TIM, DIM, ANI-FGSM, QHMI-FGSM和噪声初始化的组合
SI-NI-TI-DIM	SIM, TIM, DIM, NI-FGSM和MI-FGSM的组合

4.2.4　损失函数与超参数

实验中所有生成方法所采用的损失函数都是交叉熵损失函数. 所有实验设置最大扰动量ε为16,迭代次数T为10,步长$\alpha = \varepsilon/T$,高斯核W大小为15 × 15,转换概率p为0.7,图像转换大小为330 × 330. 本文方法的超参数则设置$\beta = 0.9$, $v = 0.1$, $\rho = 0.9$, $\beta_1 = 0.12$, $\beta_2 = 0.9$, $k = 3$.

4.3　攻击组合的生成效率

通常情况下,对抗样本生成存在运行时间和运算资源的限制,在同一条件下对比攻击组合的生成效率,具有现实意义. 本文比较表1中所有攻击组合的生成效率,实验设备使用的CPU为i7-6850K, GPU为GTX 1080 Ti,分别比较单模型黑盒攻击和多模型集成黑盒攻击. 各攻击组合的生成效率(s)如表2所示. 可以发

现,包含本文方法的攻击组合不会增加额外的运行时间,而包含SIM的攻击组合SI-NI-DIM和SI-NI-TI-DIM所需要的运行时间远超其他攻击组合.因此,单模型和多模型集成攻击实验中,将不包括SI-NI-DIM和SI-NI-TI-DIM.

4.4 消融实验

本节通过消融实验,验证本文3种方法对对抗样本迁移性能的影响.实验以NI-TI-DIM为基准方法,集成Inception v3,Inception v4,Inception ResNet v2和ResNet v2-101 4个白盒模型,逐步添加本文方法来生成对抗样本,并攻击Inc-v3ens3,Inc-v3ens4,IncResv2ens这3个黑盒防御模型.如表3所示,逐步添加本文提出的3种方法后,对抗样本对黑盒防御模型的攻击成功率逐步增加.实验表明,本文的3种方法都能提高对抗样本的迁移性.

表2 生成效率/s

攻击组合	Inc-v3	Inc-v4	IncRes-v2	Res-v2-101	多模型集成
TI-DIM	171.3	266.7	275.1	237.9	766.9
TI-DIQHM	169.2	259.3	279.5	239.5	761.4
TI-DIQHM*	180.9	273.4	290.7	231.7	789.3
NI-TI-DIM	192.5	239.7	289.1	239.5	813.4
ADNI-TI-DIQHM	181.6	231.0	285.9	241.5	805.9
ANI-TI-DIQHM	189.3	224.8	273.2	251.6	823.7
ANI-TI-DIQHM*	194.2	231.5	285.3	263.8	832.6
SI-NI-TI-DIM	602.7	1083.5	1158.3	1091.1	3491.4

表3 消融实验成功率/%

	QHMI-FGSM	ANI-FGSM	初始化	Inc-v3ens3	Inc-v3ens4	IncResv2ens
NI-TI-DIM				85.5	84.9	79.9
	√			88.0	86.0	81.4
	√	√		89.6	87.7	84.3
	√	√	√	91.1	88.7	84.5

4.5 单模型黑盒攻击

本节通过对比黑盒攻击成功率,验证QHMI-FGSM和ANI-FGSM分别替换MI-FGSM和NI-FGSM的有效性,同时验证噪声初始化的有效性,以及比较ADNI-FGSM和ANI-FGSM.单模型黑盒攻击中,对比实验分别以Inception v3,Inception v4,Inception ResNet v2和ResNet v2-101为目标模型,通过2组不同的攻击组合生成对抗样本,并攻击6个不同的黑盒防御模型.

2组攻击组合在单模型黑盒攻击中的成功率如表4和表5所示,在不增加运行时间和运算资源的前提下,与MI-FGSM、NI-FGSM相比,单模型黑盒攻击中QHMI-FGSM和ANI-FGSM能和其他方法更好地组合,实现更高的黑盒攻击成功率,即生成的对抗样本具有更好的迁移性能.同时,噪声初始化能在此基础上,实现更高的黑盒攻击成功率.此外,ANI-FGSM在单模型对抗样本生成中要优于ADNI-FGSM.

表4 TI-DIM,TI-DIQHM和TI-DIQHM*单模型黑盒攻击成功率/%

	攻击组合	Inc-v3ens3	Inc-v3ens4	IncResv2ens	HGD	R&P	NIPS-r3
Inc-v3	TI-DIM	46.7	47.1	38.6	38.3	36.2	41.5
	TI-DIQHM	50.3	50.7	38.9	38.5	37.2	43.6
	TI-DIQHM*	54.4	54.0	39.6	40.1	39.5	45.6
Inc-v4	TI-DIM	48.3	47.7	39.4	40.7	39.1	41.3
	TI-DIQHM	52.9	52.2	40.8	42.3	41.9	43.1
	TI-DIQHM*	56.2	57.1	45.5	46.8	45.7	48.4
IncRes-v2	TI-DIM	60.5	59.3	59.3	58.4	57.5	61.4
	TI-DIQHM	66.0	62.4	62.4	61.9	59.3	63.9
	TI-DIQHM*	70.6	69.2	66.6	65.4	63.8	69.3
Res-v2-101	TI-DIM	56.3	55.5	49.1	51.3	50.6	52.1
	TI-DIQHM	59.8	58.6	51.1	52.9	51.5	54.4
	TI-DIQHM*	64.0	62.4	55.4	55.9	54.1	59.5

4.6 多模型集成黑盒攻击

在多模型集成黑盒攻击条件下,本节验证QHMI-FGSM和ANI-FGSM分别替换MI-FGSM和NI-FGSM的有效性,同时验证噪声初始化的有效性,以及比较ADNI-FGSM和ANI-FGSM.实验以Inception v3,Inception v4,Inception ResNet v2和ResNet v2-101的集成模型为目标模型,通过不同的攻击组合生成对抗样本,并攻击个不同的黑盒防御模型.

表5　NI-TI-DIM,ADNI-TI-DIQHM,ANI-TI-DIQHM和ANI-TI-DIQHM*单模型黑盒攻击成功率/%

	攻击组合	Inc-v3ens3	Inc-v3ens4	IncResv2ens	HGD	R&P	NIPS-r3
Inc-v3	NI-TI-DIM	49.2	49.1	37.1	37.9	35.6	41.3
	ADNI-TI-DIQHM	50.1	49.8	37.6	38.5	36.9	42.1
	ANI-TI-DIQHM	53.0	**52.2**	37.1	39.2	37.9	42.8
	ANI-TI-DIQHM*	**53.5**	51.4	**37.7**	**39.5**	**38.1**	**43.7**
Inc-v4	NI-TI-DIM	49.6	51.0	37.2	37.1	36.5	41.3
	ADNI-TI-DIQHM	50.3	51.5	38.4	38.8	37.2	42.9
	ANI-TI-DIQHM	53.8	51.9	42.1	41.6	40.3	43.5
	ANI-TI-DIQHM*	**54.2**	**54.2**	**42.4**	**42.5**	**41.9**	**44.2**
IncRes-v2	NI-TI-DIM	64.7	63.9	61.7	62.1	60.9	64.5
	ADNI-TI-DIQHM	66.5	64.9	62.5	63.4	62.9	65.2
	ANI-TI-DIQHM	68.7	66.5	66.7	65.1	64.0	67.5
	ANI-TI-DIQHM*	**68.9**	**67.6**	**67.0**	**66.9**	**65.8**	**68.3**
Res-v2-101	NI-TI-DIM	59.2	58.7	50.0	51.2	49.7	57.6
	ADNI-TI-DIQHM	60.0	60.5	52.9	53.9	51.3	58.7
	ANI-TI-DIQHM	**65.1**	62.1	54.0	55.9	52.1	59.6
	ANI-TI-DIQHM*	64.0	**63.7**	**55.6**	**56.6**	**53.2**	**60.4**

多模型黑盒攻击中的成功率如表6和表7所示,与MI-FGSM和NI-FGSM比较,本文方法QHMI-FGSM和ANI-FGSM在多模型集成黑盒攻击中,能和其他攻击方法更好地组合,实现更高的黑盒攻击成功率. 而本文提出的噪声初始化能提高黑盒攻击成功率. 此外,ANI-FGSM在多模型集成攻击中要优于ADNI-FGSM. 最强攻击组合ANI-TI-DIQHM*对经典防御方法和较为先进的防御方法的平均黑盒攻击成功率分别为88.68%和82.77%,均超过现有最高水平.

表6　多模型黑盒攻击对经典防御方法的成功率/%

攻击组合	Inc-v3ens3	Inc-v3ens4	IncRes-v2ens	HGD	R&P	NIPS-r3	平均
TI-DIM	83.9	83.2	78.4	81.9	81.2	83.6	82.03
TI-DIQHM	86.5	84.8	80.9	83.1	82.5	84.1	83.65
TI-DIQHM*	89.0	86.8	83.5	87.1	85.4	88.9	86.78
NI-TI-DIM	86.4	84.9	81.5	83.7	82.9	84.1	83.95
ADNI-TI-DIQHM	89.2	86.6	83.8	86.9	85.7	88.8	87.17
ANI-TI-DIQHM	89.6	87.7	84.3	87.3	86.3	89.8	87.50
ANI-TI-DIQHM*	91.1	88.7	84.5	88.9	87.7	91.2	88.68

4.7　对抗样本扰动量

扰动量大小是对抗样本的一个重要衡量指标,尽管对抗样本满足$\|x^{adv}-x\|_\infty \leqslant \varepsilon$的约束,但本文的目标是在维持扰动量大小的前提下,令本文方法生成的对抗样本具有更高的黑盒攻击成功率. 因此,本节通过比较不同方法所生成的对抗样本的平均扰动量,以及比较针对9个黑盒防御模型的平均成功率,来说明本文方法的有效性. 不同方法所生成对抗样本的平均扰动量和针对9个黑盒防御模型的平均成功率比较如图2所示($\varepsilon=16$).

由图2可以发现,对比TI-DIM和NI-TI-DIM,本文

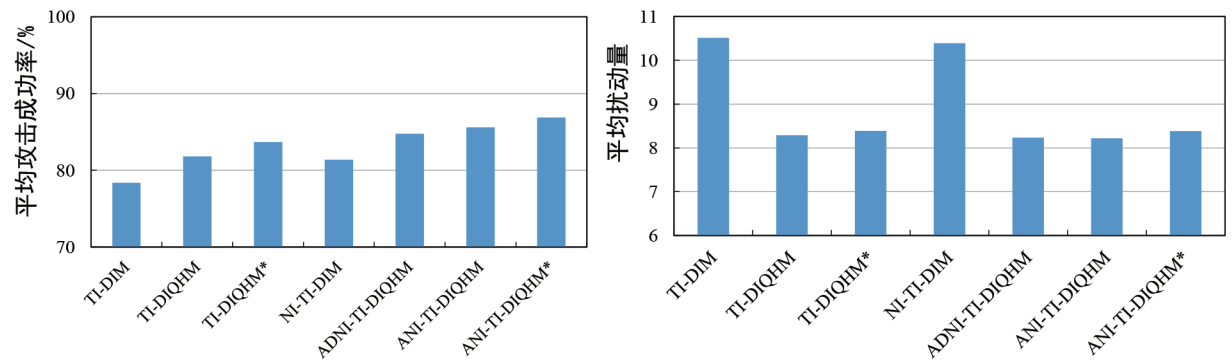

图2　多模型集成黑盒攻击对抗样本的平均成功率和平均扰动量

表7 多模型黑盒攻击对较为先进的防御方法的成功率/%

攻击组合	Feature Distillation	Comdefend	Randomized Smoothing	平均
TI-DIM	83.1	78.2	49.9	70.40
TI-DIQHM	84.3	86.9	59.2	76.80
TI-DIQHM*	89.9	88.1	63.1	80.37
NI-TI-DIM	82.1	84.7	58.6	75.13
ADNI-TI-DIQHM	88.9	85.8	62.9	79.20
ANI-TI-DIQHM	90.3	88.5	64.5	81.10
ANI-TI-DIQHM*	91.2	89.7	67.4	82.77

方法不仅能提高对抗样本的黑盒攻击成功率，还能将对抗样本的平均扰动量降低10%以上.

4.8 对抗样本对比

为了验证图2的结果，本节对比不同方法生成的对抗样本($\varepsilon = 16$). 由图3可以发现，本文方法所生成的对抗样本与TI-DIM，NI-TI-DIM所生成的对抗样本相比，由于平均扰动量更低，其对抗噪声形成的条纹更淡. 然而无论是现有方法TI-DIM和NI-TI-DIM，还是本文方法，对比干净样本，对抗样本上的条纹都较为明显. 显然，通过$\left\| x^{\text{adv}} - x \right\|_\infty \leqslant \varepsilon$去限制对抗样本的扰动量是不够的，平均扰动量可以作为参考指标之一. 在接下来的工作中，维持黑盒攻击成功率，降低对抗样本的平均扰动量，使得对抗样本更具有威胁，是一项有意义的研究.

5 结论及展望

本文针对基于梯度的对抗样本生成方法，提出基于噪声初始化、Adam-Nesterov方法和准双曲动量方法的对抗样本生成方法. 本文对对抗噪声初始化进行研究，通过像素偏移方法来预先增强干净样本的攻击性能. 同时，本文使用Adam-Nesterov方法和准双曲动量方法来改进现有生成方法中的Nesterov方法和动量方法，实现更高的黑盒攻击成功率. 在不需要额外运行时间和运算资源的情况下，本文方法可以和其他的攻击

图3 不同方法生成的对抗样本

方法组合,并显著提高了对抗样本的黑盒攻击成功率.实验表明,本文的最强攻击组合为ANI-TI-DIQHM*,其对经典防御方法的平均黑盒攻击成功率达到88.68%,对较为先进的防御方法的平均黑盒攻击成功率达到82.77%,均超过现有最高水平.

参考文献

[1] HE K M, ZHANG X G, REN S Q, et al. Identity mappings in deep residual networks[C]//2016 14th European Conference on Computer Vision. Amsterdam, NEP: Springer, 2016: 630-645.

[2] HE K M, GKIOXARI G, DOLLÁR P, et al. Mask R-CNN[C]// 2017 IEEE International Conference on Computer Vision. Venice, Italy: IEEE Computer Society, 2017: 2980-2988.

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

[4] LIU Y P, CHEN X Y, LIU C, et al. Delving into transferable adversarial examples and black-box attacks[C]//2017 5th International Conference on Learning Representations. Toulon, France: Conference Track Proceedings, 2017: 1-24.

[5] ATHALYE A, ENGSTROM L, ILYAS A, et al. Synthesizing robust adversarial examples[C]//2018 35th International Conference on Machine Learning. Stockholmsmässan, SWE: Proceedings of Machine Learning Research, 2018: 284-293.

[6] TRAMÈR F, KURAKIN A, PAPERNOT N, et al. Ensemble adversarial training: Attacks and defenses[C]//2018 6th International Conference on Learning Representations. Vancouver, CAN: Conference Track Proceedings, 2018: 1-22.

[7] LIAO F Z, LIANG M, DONG Y P, et al. Defense against adversarial attacks using high-level representation guided denoiser[C]//2018 IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE Computer Society, 2018: 1778-1787.

[8] XIE C H, WANG J Y, ZHANG Z S, et al. Mitigating adversarial effects through randomization[C]//2018 6th International Conference on Learning Representations. Vancouver, CAN: Conference Track Proceedings, 2018: 1-16.

[9] RAGHUNATHAN A, STEINHARDT J, LIANG P. Certified defenses against adversarial examples[C]//2018 6th International Conference on Learning Representations. Vancouver, CAN: Conference Track Proceedings, 2018: 1-15.

[10] RAUBER J, BRENDEL W, BETHGE M. Foolbox v0.8.0: A python toolbox to benchmark the robustness of machine learning models[J/OL]. [2020]. https://arxiv.org/abs/1707.04131.

[11] DONG Y P, LIAO F Z, PANG T Y, et al. Boosting adversarial attacks with momentum[C]//2018 IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE Computer Society, 2018: 9185-9193.

[12] NARODYTSKA N, KASIVISWANATHAN S P. Simple black-box adversarial perturbations for deep networks [J/OL]. (2016-12-19) [2020]. https://arxiv. org/abs/1612.06299.

[13] CHEN J B, JORDAN M I. Boundary attack++: Query-efficient decision-based adversarial attack[J/OL]. [2020]. https://arxiv.org/abs/1904.02144.

[14] LIN J D, SONG C B, HE K, et al. Nesterov accelerated gradient and scale invariance for improving transferability of adversarial examples[J/OL]. [2020]. https://arxiv.org/abs/1908.06281.

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

[16] DONG Y P, PANG T Y, SU H, et al. Evading defenses to transferable adversarial examples by translation-invariant attacks[C]//2019 IEEE Conference on Computer Vision and Pattern Recognition. Long Beach, USA: Computer Vision Foundation, 2019: 4312-4321.

[17] KURAKIN A, IAN J. GOODFELLOW, SAMY BENGIO. Adversarial examples in the physical world[C]//2017 5th International Conference on Learning Representations. Toulon, France: Conference Track Proceedings, 2017: 1-14.

[18] MA J, YARATS D. Quasi-hyperbolic momentum and adam for deep learning[C]//2019 7th International Conference on Learning Representations. New Orleans, USA: Conference Track Proceedings, 2019: 1-38.

[19] 贾熹滨,史佳帅. Ada_Nesterov动量方法——种具有自适应学习率的Nesterov动量方法[J]. 计算机科学与应用,2019,9:351-358.
JIA X B, SHI J S. Ada_Nesterov momentum algorithm—the nesterov momentum algorithm with adaptive learning rate[J]. Computer Science and Application, 2019, 9: 351-358. (in Chinese)

[20] KINGMA D P, BA J. Adam: A method for stochastic optimization [C]//2015 3rd International Conference on

Learning Representations. San Diego, USA: Conference Track Proceedings, 2015: 1-15.
[21] DUCHI, JOHN, HAZAN, et al. Adaptive subgradient methods for online learning and stochastic optimization[J]. The Journal of Machine Learning Research, 2011, 12: 2121-2159.
[22] 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. Las Vegas, USA: IEEE Computer Society, 2016: 2818-2826.
[23] SZEGEDY C, IOFFE S, VANHOUCKE V, et al. Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning[C]//2017 31st AAAI Conference on Artificial Intelligence. San Francisco, USA: AAAI Press, 2017: 4278-4284.
[24] LIU Z H, LIU Q, LIU T, et al. Feature distillation: DNN-Oriented JPEG compression against adversarial examples[C]//2019 IEEE Conference on Computer Vision and Pattern Recognition. Long Beach, USA: Computer Vision Foundation, 2019: 860-868.
[25] JIA X J, WEI X X, CAO X C, et al. ComDefend: An efficient image compression model to defend adversarial examples[C]//2019 IEEE Conference on Computer Vision and Pattern Recognition. Long Beach, USA: Computer Vision Foundation, 2019: 6084-6092.
[26] COHEN J M, ROSENFELD E, KOLTER J Z. Certified adversarial robustness via randomized smoothing[C]// 2019 36th International Conference on Machine Learning ICML. Long Beach, USA: Proceedings of Machine Learning Research, 2019: 1310-1320.

作者简介

邹军华 男,1991 年 12 月出生,广东河源人. 2017 年获解放军理工大学硕士学位. 现为陆军工程大学在读博士. 研究方向为对抗学习.
E-mail:278287847@qq.com

潘志松(通信作者) 男,1973 年 3 月出生,江苏南京人. 2003 年获南京航空航天大学博士学位. 现为陆军工程大学教授,博士生导师. 研究方向为人工智能、模式识别.
E-mail:panzs@nuaa.edu.cn