基于主特征归因的对抗样本生成方法研究

王 硕,徐茹枝,关志涛
(华北电力大学控制与计算机工程学院,北京 102206)

摘 要: 为提高对抗训练的样本质量,本文对深度学习模型内部的特征识别过程进行了探究,并提出了一种基于主特征归因的迁移性对抗样本生成方法. 算法在提取样本的主要特征后对目标层神经元进行特征归因,并利用独立性假设简化梯度计算,通过抑制积极神经元的识别作用,更加高效地得到更具迁移性的对抗样本. 经过大量实验验证,相比于已有方法,在针对多模型的攻击中,本文算法的攻击成功率提高了5%以上,为后续研究如何提高模型的鲁棒性奠定了基础.

关键词: 深度学习;图像识别;特征提取;特征归因;对抗样本
基金项目: 国家自然科学基金(No.61972148)
中图分类号: TP391.4 **文献标识码:** A **文章编号:** 0372-2112(2023)11-3137-09
电子学报 URL:http://www.ejournal.org.cn **DOI:**10.12263/DZXB.20230383

Research on the Generation of Adversarial Samples Based on the Attribution of Principal Features

WANG Shuo, XU Ru-zhi, GUAN Zhi-tao
(*School of Control and Computer Engineering*, *North China Electric Power University*, *Beijing* 102206, *China*)

Abstract: In order to improve the quality of adversarial training samples, this article explores the feature recognition process within deep learning models and proposes a transferability adversarial sample generation method based on main feature attribution. After extracting the main features of the samples, the algorithm attributes the features of the target layer neurons, and simplifies the gradient calculation by using the independence assumption. By inhibiting the recognition of active neurons, the algorithm can more efficiently obtain more transferable adversarial samples. Through a large number of experiments, compared to existing methods, the success rate of our algorithm in attacks against multiple models has been improved by at least 5%, it lays a foundation for further research on how to improve the robustness of the model.

Key words: deep learning; image recognition; feature extraction; feature attribution; adversarial sample
Foundation Item(s): National Natural Science Foundation of China (No.61972148)

1 引言

近年来,深度神经网络(Deep Neural Networks,DNNs)的内部结构愈加复杂[1],神经元搭配更加灵活,模型训练效率逐渐提高[2]. 但已有研究表明,在正常样本上添加对抗性扰动生成的对抗样本,可以误导DNNs输出错误标签,使得深度神经网络模型在其应用领域的安全性受到了严重的威胁[3].

文献[4]中首先提出了快速梯度符号法(Fast Gradient Sign Method,FGSM),其沿逆梯度方向添加扰动快速生成对抗样本. 随之,更多基于梯度的攻击方法[5~8]被提出,进一步提高了对抗样本的攻击成功率. 此外,基于优化的攻击方法[9~11],也相继被提出,如文献[12]中,将扰动限制内部化,得到感知率更低且攻击成功率更高的对抗样本. 与白盒攻击不同,黑盒攻击无法得知模型的具体参数设置. 其中,基于访问的黑盒攻击[13,14],需要进行多次梯度近似估计来生成对抗样本,生成成本较大. 查询目标模型并利用获取的数据集训练生成替代模型是替代攻击的主要思路,但由于模型间的识别区域不同,所得样本的迁移性受到了目标模型的限制[15,16].

特征级的可迁移性攻击[17~19],通过修改深度神经网络模型目标层中的特征映射来生成对抗样本. Aditya等提出了特征破坏攻击方法(Feature Disruptive Attack,FDA)[20],利用平均激活值来区分神经元的重要性,文献[21]中提出了特征重要性感知攻击(Feature Important-aware Attack,FIA),利用神经元正向传播和反向梯度计算来进行神经元重要性度量,解决了反向梯

收稿日期:2023-04-28;修回日期:2023-09-21;责任编辑:覃怀银

度传播上存在的梯度饱和问题. 神经元属性攻击(Neuron Attribution-based Attack,NAA)[22]方法,进一步优化神经元归因方法,并在破坏积极特征的同时,促进消极特征,以此来生成一种可解释的且更具迁移性的对抗样本. 但在已有方法中仍存在归因特征的选取范围过大,神经元聚合次数过多,以及梯度计算步长较小等问题,限制了对抗样本迁移性和生成效率的提高.

为解决以上问题,本文首先通过主特征提取缩小特征归因的区域范围,进而细化目标层神经元对输出标签的作用,并提出近似计算方法用于快速特征归因,更加高效地得到强迁移性的对抗样本. 同时,本文还对神经元的特征归因价值进行了探究.

2 相关工作

2.1 对抗攻击与防御

对抗样本在产生威胁的同时,也为如何完善深度神经网络模型的内部缺陷,提高模型的鲁棒性提供了研究基础. 目前,针对对抗样本的恶意攻击,对抗训练(adversarial training)被认为是最有效的防御方法. 文献[23]中随机利用多种攻击方法生成对抗样本,并训练分类器对对抗样本进行正确分类,以此来达到防御对抗攻击的效果. Shafahi A 等提出了快速对抗训练方法(free adversarial training)[24],将旧对抗样本训练产生的梯度信息直接用于新对抗样本的产生,从而降低了生成对抗样本的计算开销. 在此基础之上,有学者相继提出多种优化方法[25~27],进一步提高了对抗训练的效率,以及针对多类对抗攻击的防御能力.

2.2 特征级攻击

特征级攻击并非通过对深度学习模型的访问,最小化损失函数来获得对抗样本,而是通过修改深度神经网络内部的特征映射,对目标层的神经元进行特征归因,依据神经元激活值的正负,将其分为积极与消极神经元,抑制积极神经元的同时,促进消极神经元发挥作用得到对抗样本. 如图 1 所示,特征归因后,积极神经元的识别作用受到抑制,原模型及对抗攻击的目标模型,对生成样本的识别区域转移至非主要特征区域,致使对抗样本输出错误标签.

3 主特征归因算法

3.1 算法设计

定义原始样本数据集为 D,x 表示数据集中的样本,其真实标签为 z,x_i 表示图像样本的第 i 个像素值,x_{adv} 表示对抗样本,x^0 为基准图像,$F(\cdot)$ 为深度学习模型,y 为深度学习模型第 y 层上的激活函数输出结果,y_i 为第 y 层上第 i 个神经元的激活值.

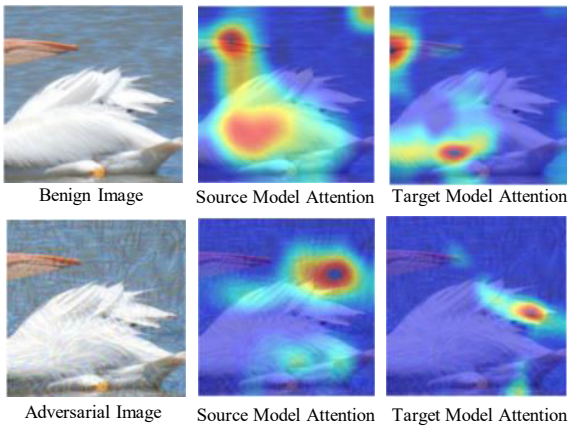

图 1 特征级对抗样本示例

3.1.1 主特征提取

算法首先对数据集 D 中样本进行去中心化处理,见式(1).

$$x_i = x_i - \frac{1}{n}\sum_{j=1}^{n} x_j \tag{1}$$

其次,计算样本的协方差矩阵 $\boldsymbol{Q}$,利用特征值分解法,将特征值从大到小进行排序,选取前 K 个特征值及其对应的特征向量构成向量 $\boldsymbol{u}=(u_1,u_2,\cdots,u_k)$,并将其分别作为行向量组成特征向量矩阵 $\boldsymbol{P}$,经投影变换后,得到初始样本 D'.

主特征提取流程如图 2 所示,原始样本经过去中心化处理后,将三维矩阵转换为一维矩阵,并利用计算得到的特征向量矩阵 $\boldsymbol{P}$ 进行投影变换,然后用零向量补齐矩阵,最后将矩阵还原为与原始图像大小一致的三维矩阵.

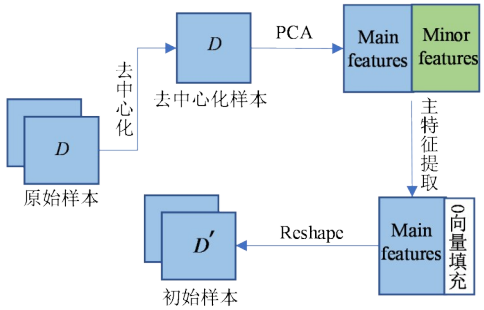

图 2 PCA 流程图

3.1.2 主特征归因

定义图像的主特征归因函数 AP,见式(2).

$$\mathrm{AP}:=\sum_{i=1}^{n^2}\left(x_i-x_i^0\right)\cdot\int_0^1 \Delta_i(x)\,\mathrm{d}\alpha \tag{2}$$

其中,$\Delta_i(x)=\dfrac{\partial F}{\partial x_i}\left(x^0+\alpha\cdot\left(x-x^0\right)\right)$ 为 $F(x)$ 对图像中第 i 个像素的偏微分,$\int_0^1 \Delta_i(x)\,\mathrm{d}\alpha$ 为对 $\Delta_i(x)$ 路径积分,用于求解图像在目标层的激活值. 根据路径积分定理,若

$F(x')\approx 0$ 则 $\mathrm{AP}\approx F(x)$,因此本文选取全黑图像,即 $x^0=0$,用于路径积分的近似计算.

将式(2)中的整体归因结果进一步细化至目标层的单个神经元,令 $x_\alpha=x^0+\alpha\cdot\left(x-x^0\right)$,更新 $\Delta_i(x)$ 函数,定义目标层神经元的归因函数为 AP_{y_j},见式(3):

$$\begin{aligned}\mathrm{AP}_{y_j}:&=\sum_{i=1}^{n^2}\left(x_i-x_i^0\right)\cdot\int_0^1\Delta_i\left(x_\alpha\right)\mathrm{d}\alpha\\ \Delta_i(x)&=\frac{\partial F}{\partial y_j}\left(y(x)\right)\frac{\partial y_j}{\partial x_i}(x)\\ \sum&\mathrm{AP}_{y_j}=\mathrm{AP}\end{aligned}\tag{3}$$

其中,$\Delta_i(x)$ 为目标层中第 j 个神经元对图像的第 i 个像素的偏导数,式(3)中 $\sum\mathrm{AP}_{y_j}=\mathrm{AP}$ 表明深度学习模型中目标层激活值与神经元激活值之和的一致性关系. 本文将积分路径划分为 n 份,令 $x_k=x^0+\frac{k}{n}\left(x-x^0\right)$,利用黎曼和积分对式(3)近似计算,见式(4):

$$\mathrm{AP}_{y_j}\approx\sum_{i=1}^{n^2}\left(x_i-x_i^0\right)\cdot\frac{1}{n}\cdot\sum_{k}^{n}\Delta_i\left(x_k\right)\tag{4}$$

式(4)将计算得到目标层中每一个神经元的激活值,但每一次的计算均要通过 n 次求导近似,计算开销过大,对抗样本生成成本较高. 受文献[22]的启发,根据深度学习模型内部结构特点,每一层对图像的卷积处理可看作独立过程,因此本文假设正向传播与反向传播二者计算过程相互独立,并依据正向与反向传播过程进行整理,得神经元归因函数见式(5):

$$\mathrm{AP}_{y_j}\approx\frac{1}{n}\sum_{k=1}^{n}\left(\frac{\partial F}{\partial y_j}\left(y\left(x_k\right)\right)\right)\left(\sum_{i=1}^{n^2}\left(x_i-x_i^0\right)\cdot\frac{\partial y_j}{\partial x_i}\left(x_k\right)\right)\tag{5}$$

根据独立性假设性质可知,式(5)中正向与反向传播两部分线性无关,即协方差为 0 成立,满足 $\sum_{1}^{n}\left(x_i-\bar{x}\right)\left(y_i-\bar{y}\right)=0$. 其中,$\bar{x},\bar{y}$ 分别为 x_i,y_i 的平均值,展开可得 $\sum_{1}^{n}x_i\cdot y_i=\frac{1}{n}\sum_{1}^{n}x_i\cdot\sum_{1}^{n}y_i$,依据展开等式对式(5)进行展开后合并整理,结果见式(6):

$$\mathrm{AP}_{y_j}\approx\frac{1}{n}\sum_{k=1}^{n}\frac{\partial F}{\partial y_j}\left(y\left(x_k\right)\right)\frac{1}{n}\sum_{k=1}^{n}\sum_{i=1}^{n^2}\left(x_i-x_i^0\right)\frac{\partial y_j}{\partial x_i}\left(x_k\right)\tag{6}$$

由路径积分的微积分基本定理可知,$\frac{1}{n}\sum_{k=1}^{n}\sum_{i=1}^{n^2}\left(x_i-x_i^0\right)\frac{\partial y_j}{\partial x_i}\left(x_k\right)=y_i-y_i^0$,其中,$y_i^0$ 为基准图像在 y 层上神经元的激活值. 同时,根据主特征归因原理,定义正向神经元激活值部分为 $\mathrm{FA}\left(y_i\right):=y_i-y_i^0$;逆向梯度计算部分为 $\mathrm{BA}\left(y_i\right):=\frac{1}{n}\sum_{k=1}^{n}\frac{\partial F}{\partial y_j}\left(y\left(x_k\right)\right)$,整理后得主特征归因算法的表达式,见式(7):

$$\begin{aligned}\mathrm{AP}_{y_j}&=\mathrm{FA}\left(y_i\right)\cdot\mathrm{BA}\left(y_i\right)\\ \mathrm{AP}&=\sum\mathrm{AP}_{y_j}\end{aligned}\tag{7}$$

本文基于独立性假设,将归因函数简化为输入图像 x 与基准图像 x^0 激活函数上的差值,和预测标签对指定层上神经元的导函数的乘积. 整个计算过程将每一神经元均需对所有像素点进行逐一求导,化简至每次积分运算时,仅需进行一次神经元对输出标签结果的求导,在精准归因的基础之上简化了大量计算,进一步提高了对抗样本生成效率.

3.2　损失函数构造

本文在 $\left\|x-x_{\mathrm{adv}}\right\|_\infty\leqslant\varepsilon$ 无穷范数的约束条件下,通过抑制部分积极神经元的作用,在更小的范围中更改目标层的特征映射,获得可以误导多个深度神经网络模型的迁移性对抗样本. 为此,定义损失函数见式(8):

$$\mathrm{Loss}=\frac{1}{\mathrm{len}}\cdot\sum_{\mathrm{AP}_{y_j}>0}a\cdot\mathrm{AP}_{y_j}\tag{8}$$

其中,len 为深度学习模型中目标层 y 所包含的神经元总数,a 为超参数用于控制积极神经元的所占权重. Loss 计算得到对输出特征标签起积极作用的神经元权重数,与所有目标层的神经元总数比值作为损失函数.

3.3　算法流程

整体的主特征归因算法过程如算法 1. 首先对输入的原始样本进行主特征提取,得到初始化样本 D',然后根据设置的 batch 大小,依次进行主特征归因计算得到 AP,最后计算 Loss 值,并进行约束最小化处理获得具有迁移性的对抗样本 x_{adv}.

4　实验

4.1　实验设置

实验在 Python 3.6;TensFlow 1.14;Keras 2.2 的版本环境下,选取包含 1 000 个分类,共计 5 万张图片的 ILSVRC 2012 验证集作为实验数据集. 在数据集中随机抽取 1 000 张图片进行实验,选用动量迭代(MI-FSGM,MIM)算法来实现样本扰动的添加,并选取近三年提出的 4 种特征级攻击方法 NAA,FIA,FDA,NRDM(Neural Representation Distortion Method),分别在现有的 10 种无防御模型和防御模型上进行测试,计算 10 次实验结果的平均值,作为最终结果进行分析,模型具体信息如表 1.

4.2　评价指标

本文以原始样本的真实标签为基准,通过计算模型输出的对原始样本、对抗样本的预测标签准确数与样本总数的比值,得到原始图像识别准确率、对抗样本差错率. 通过计算对抗样本的错误预测标签数量,与样本总数的比值,得到对抗样本攻击成功率,从而客观完整地对比实验结果,具体评价指标如表 2,其中 n 为样本

算法1 主特征归因算法

输入:深度学习模型$F(\cdot)$;数据集$D=\{x_1,x_2,\cdots,x_n\}\in\mathbb{R}^{n\times n}$;基准图像$x^0$;扰动约束$\varepsilon$;迭代次数$n$;样本批处理数量batch;权重超参数$a$.

输出:具有迁移性的对抗样本x_{adv}.

(1)FOR batch DO:

(2) 计算协方差矩阵$\boldsymbol{Q}$,求解前K个特征值及特征向量,组成投影变换矩阵$\boldsymbol{P}$;

(3) 经变换得到初始化样本D'

(4)END FOR

(5)FOR batch DO

(6) 正向求解激活值$\mathrm{FA}(y_i)$,逆向梯度求解$\mathrm{BA}(y_i)$

(7) 计算$\mathrm{AP}_{y_j}=\mathrm{FA}(y_i)\cdot\mathrm{BA}(y_i)$,并依据激活值的正负划分神经元的积极与消极作用.

(8)END FOR

(9)FOR 以batch为一组进行迭代n次 DO

(10) 计算$\mathrm{Loss}=\frac{1}{\mathrm{len}}\cdot\sum_{\mathrm{AP}_{y_j}>0}a\cdot\mathrm{AP}_{y_j}$;

(11) 更新梯度$g_{n+1}=\mu\cdot g_n+\frac{\nabla_x\mathrm{Loss}}{\|\nabla_x\mathrm{Loss}\|}$

(12) $x_{adv}=x_{adv}-\frac{\varepsilon}{n}\cdot\mathrm{sign}(g_{n+1})$

(13)END FOR

(14)RETURN x_{adv}

表1 模型信息

模型类别	模型名称	目标层
无防御模型	inception_v3[28]	Mixed 5b
	inception_v4[29]	Mixed 5e
	inception_resnet_v2[29]	Conv2d 4a
	resnet_v2_152[30]	block2 layer
	vgg_16[31]	Conv3_3
防御模型	adv_inception_v3	
	adv_inception_resnet_v2	
	ens3_adv_inception_v3	
	ens4_adv_inception_v3	
	ens3_adv_inception_resnet_v2	

总数,z为样本对应的真实标签,x为原始样本,x_{adv}为对抗样本,$F(x)$为模型输出的预测标签.

表2 评价指标

评价指标	计算方法
原始图像识别准确率	$\mathrm{sum}(F(x)==z)/n$
对抗样本差错率	$\mathrm{sum}(F(x_{adv})==z)/n$
对抗样本攻击成功率	$\mathrm{sum}(F(x_{adv})!=z)/n$

4.3 消融实验

主特征归因算法的提出为更大的扰动添加步长、更少的神经元聚合次数提供了的可能. 在消融实验中,本文选用inception_v3作为源模型生成对抗样本,在包括防御模型在内的5种模型上进行验证实验,现有针对每种模型的最高攻击成功率已用同样式横线绘出.

4.3.1 步长对样本迁移性的影响

本节探究添加扰动时步长对样本迁移性的影响,具体实验结果如图3. 随着步长的增加,基于主特征归因算法所生成的对抗样本的攻击成功率逐渐下降,当步长增长至10时,对抗样本在resnet_v2_152模型上的攻击成功率低于现有方法,但在防御模型上的攻击成功率仍优于已有方法. 相比于NNA方法选取的小步长2,本文的算法能以更大的步长获得更高的攻击成功率. 为了更高效地生成具有迁移性的对抗样本,本文选取步长为5进行后续实验.

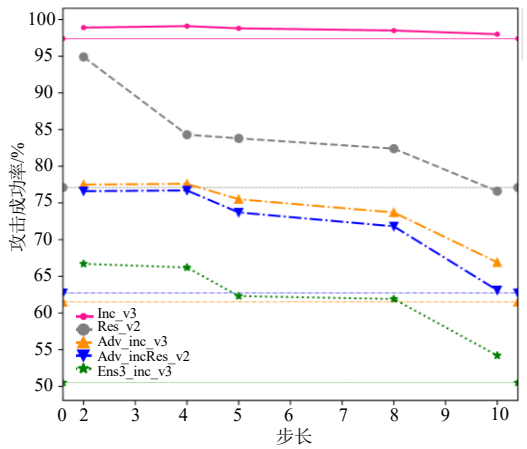

图3 步长对样本迁移性的影响

4.3.2 神经元聚合次数对样本迁移性的影响

神经元聚合次数即黎曼和积分计算中,神经元的近似求导次数. 本节探究神经元聚合次数对生成对抗样本的影响,具体结果如图4. 参照已有最优NAA方法

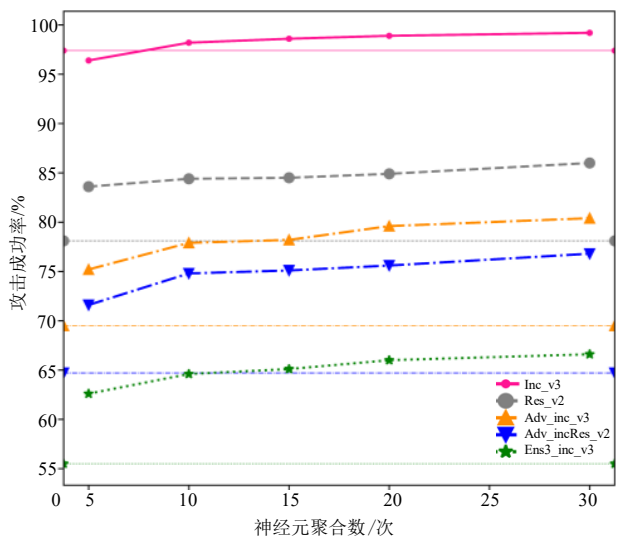

图4 神经元聚合次数对样本迁移性的影响

的神经元聚合次数30作为上限,随着神经元聚合次数的增多,所生成对抗样本的攻击成功率也逐渐增加. 当聚合次数下降为5时,对抗样本针对源模型的攻击成功率低于已有算法,但针对黑盒模型的攻击成功率仍优于已有算法,实验证明本文的算法能以更少的神经元聚合次数,获得更高的攻击成功率,进一步提高了对抗样本的生成效率. 结合实验结果,本文选取聚合次数为10进行后续对比实验.

4.3.3　损失函数中权重值对样本的影响

在确定了步长大小和神经元聚合次数的前提下,本节对损失函数中的权重进行了探究,具体实验结果如图5.

随着权重值的增加,对积极神经元的抑制作用在增加,对抗样本的攻击成功率逐渐提高,并在0.5处取得最大值,但当权重值超过0.5时,在resnet_v2_152和adv_inception_v3模型上的攻击成功率下降,说明在样本识别过程中,积极与消极神经元均发挥着同等重要的作用,证明对神经元的细化归因有利于生成更具迁移性的对抗样本. 为了权衡各模型之间的最优情况,本文选择权重值为0.5进行对比实验.

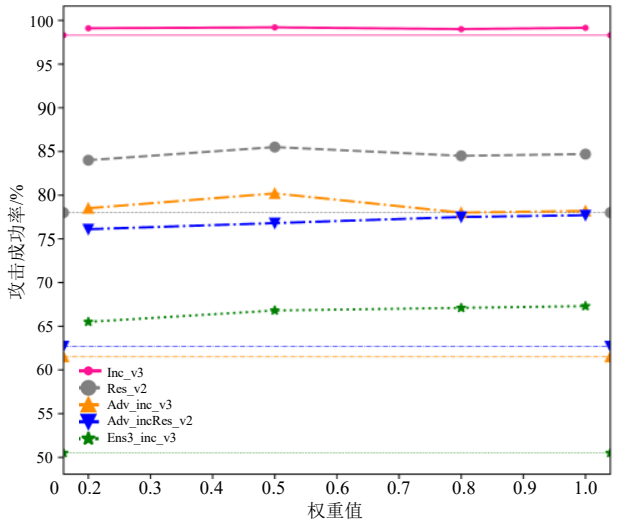

图5　权重对样本迁移性的影响

4.4　对比实验

4.4.1　主特征归因算法实验结果对比

本文选取inception_v3等5个基础模型作为源模型用以生成对抗样本,与现有先进的攻击方法进行黑白盒攻击的对比实验,以此来全面验证基于主特征归因算法生成的对抗样本具有更好的迁移性,具体实验结果如表3.

表中最优的实验结果已经加粗显示. 在白盒对比实验中,MIM算法在对5个源模型的攻击中均取得了最

表3　算法实验结果对比　　单位:%

源模型	攻击算法	Inc_v3	Inc_v4	IncRes_v2	Res_v2	Vgg_16	Inc_v3_{Adv}	$IncRes\_v2_{Adv}$	Inc_v3_{Ens3}	Inc_v3_{Ens4}	$IncRes\_v2_{Ens3}$
Inc_v3	MIM	**99.98**	40.50	40.00	32.60	38.60	23.10	20.10	16.00	16.50	8.10
	NRDM	98.89	62.40	54.10	51.10	49.98	25.70	19.20	9.90	10.50	5.20
	FDA	82.10	43.30	35.80	34.50	51.90	19.30	12.20	8.70	6.60	2.40
	FIA	98.30	82.70	80.00	72.20	71.40	52.40	51.00	35.60	37.00	20.00
	NAA	98.10	85.00	82.30	76.10	81.60	60.00	61.50	50.00	50.60	31.20
	Ours	98.30	**87.40**	**84.50**	**82.40**	**85.60**	**73.40**	**71.20**	**60.20**	**60.90**	**45.70**
Inc_v4	MIM	58.00	**99.60**	45.50	38.80	36.80	23.80	21.10	19.50	18.60	8.80
	NRDM	78.20	96.40	60.00	62.00	51.00	26.50	26.50	17.70	15.60	5.60
	FDA	83.80	99.50	71.00	68.80	56.40	28.00	26.00	19.50	17.10	7.40
	FIA	84.00	95.70	78.50	72.00	72.80	45.50	47.20	38.00	37.30	19.20
	NAA	85.80	96.60	80.00	75.60	80.60	52.20	56.00	50.50	49.20	31.60
	Ours	**89.90**	97.60	**85.40**	**84.00**	**86.20**	**71.20**	**72.00**	**66.30**	**64.80**	**51.50**
IncRes_v2	MIM	60.00	51.50	**99.20**	42.20	46.90	24.50	31.10	21.80	23.50	12.60
	NRDM	71.00	66.90	77.50	57.60	49.40	33.40	29.00	16.20	23.80	19.60
	FDA	69.20	67.60	77.40	55.60	53.50	37.60	29.60	16.20	22.10	18.00
	FIA	81.50	77.20	87.80	71.50	71.50	64.40	65.00	49.80	47.80	34.10
	NAA	82.00	78.00	92.60	74.00	80.50	65.00	67.10	60.00	57.60	47.50
	Ours	**89.60**	**85.90**	94.10	**83.90**	**86.30**	**73.70**	**75.90**	**65.40**	**63.20**	**55.60**
Res_v2	MIM	54.00	47.20	43.50	**99.40**	72.80	26.40	25.10	24.00	25.30	12.60
	NRDM	73.30	70.50	55.80	89.40	79.30	38.00	30.00	23.50	18.80	9.50
	FDA	84.00	84.10	71.80	88.80	75.00	51.10	41.10	28.10	23.50	11.50
	FIA	83.00	80.90	74.80	95.60	76.50	58.30	52.80	48.60	44.60	29.10
	NAA	84.80	83.80	76.70	96.00	79.70	60.20	58.00	51.20	48.80	35.00
	Ours	**88.50**	**84.40**	**81.10**	97.00	**84.70**	**67.20**	**63.90**	**56.80**	**53.50**	**40.80**
Vgg16	MIM	80.30	81.10	74.60	84.40	**99.99**	64.30	61.10	50.20	50.50	45.00
	NRDM	73.60	72.80	57.10	73.00	93.20	69.90	63.50	58.80	56.30	47.90
	FDA	76.10	76.70	64.00	78.70	95.70	73.70	66.20	64.80	67.80	52.10
	FIA	**95.70**	93.60	88.60	91.30	99.80	82.60	78.60	**85.60**	82.00	70.80
	NAA	93.60	93.40	90.70	91.20	99.90	86.40	82.00	84.20	83.60	69.40
	Ours	94.90	**94.90**	**91.30**	**91.90**	99.90	**86.50**	**86.80**	82.80	**83.90**	**74.20**

高的攻击成功率,但主特征归因算法的攻击成功率,均要高于现有4种方法,并与MIM算法的最高攻击成功率最小相差仅0.09%. 在黑盒对比实验中,除Vgg16源模型外,主特征归因算法针对无防御模型的攻击成功率,相比已有的5种方法平均提高了5%以上;针对防御模型,主特征归因算法的攻击成功率相比于现有的方法均提高了10%以上.

综上,在对抗样本生成效率提高的前提下,本文所提出的算法在黑白盒攻击中,仍能保持较高的攻击成功率,且优于近三年的已有方法. 本文算法将特征归因区域集中在样本的主要特征区域,使得对抗样本的攻击成功率在防御模型上要比无防御模型提高更多. 实验证明了算法的有效性.

4.4.2 结合随机转换方法的实验结果对比

为了进一步验证算法的有效性,本节将主特征归因算法与随机转换方法(DI2-FGSM、PI-FGSM)结合进行验证实验. 整体上,初始样本经过包括调整大小、像素填充等随机处理方法,可将因增大步长而导致超过阈值的噪声投影到周围区域,使得攻击成功率整体上有所提高. 具体实验结果见表4.

表4 结合随机转换算法实验结果对比 单位:%

源模型	攻击算法	Inc_v3	Inc_v4	IncRes_v2	Res_v2	Vgg_16	Adv_Inc_v3	Adv_IncRes_v2	Ens3_Inc_v3	Ens4_Inc_v3	Ens3_IncRes_v2
Inc_v3	MIM^{PD}	**99.80**	70.50	67.60	53.60	60.60	31.00	28.00	21.20	22.00	9.30
	$NRDM^{PD}$	86.60	67.60	63.60	59.50	62.96	29.50	23.00	12.40	18.60	13.40
	FDA^{PD}	76.00	50.50	45.60	39.00	48.80	23.00	16.00	10.80	12.00	8.00
	FIA^{PD}	97.70	87.60	86.00	80.00	85.00	58.90	57.00	38.20	36.60	21.50
	NAA^{PD}	97.80	88.80	87.50	83.50	88.60	67.60	66.80	55.50	54.60	32.60
	Ours	98.70	**90.70**	**89.60**	**87.50**	**89.90**	**76.70**	**72.90**	**61.70**	**61.50**	**46.50**
Inc_v4	MIM^{PD}	81.40	**99.30**	72.00	59.30	55.80	30.60	28.80	23.90	24.40	12.50
	$NRDM^{PD}$	88.80	96.80	80.00	77.70	62.00	34.50	35.00	21.30	19.00	8.60
	FDA^{PD}	91.30	99.20	86.60	82.20	78.60	36.60	38.00	22.00	21.00	9.00
	FIA^{PD}	90.50	97.10	88.60	84.50	85.80	55.30	60.70	45.90	42.50	23.40
	NAA^{PD}	91.20	97.70	89.30	86.00	88.20	60.00	71.10	56.90	53.60	36.60
	Ours	**93.40**	98.40	**90.30**	**88.10**	**91.10**	**76.10**	**76.90**	**70.00**	**67.40**	**53.70**
IncRes_v2	MIM^{PD}	80.60	76.50	**98.10**	64.00	66.80	36.80	41.50	28.60	26.70	16.20
	$NRDM^{PD}$	76.60	74.50	79.10	66.20	64.60	41.00	33.20	18.80	30.50	26.00
	FDA^{PD}	78.50	75.00	80.00	66.20	68.20	42.10	36.50	17.70	30.00	25.50
	FIA^{PD}	85.10	78.00	90.00	75.60	72.60	67.00	66.70	49.80	45.00	32.00
	NAA^{PD}	85.50	80.50	90.80	79.00	78.50	71.00	68.00	52.40	50.90	40.40
	Ours	**89.60**	**84.90**	91.30	**81.80**	**86.40**	**73.50**	**70.70**	**57.60**	**57.20**	**45.10**
Res_v2	MIM^{PD}	81.60	76.60	75.70	**99.40**	72.60	42.00	44.40	36.30	34.20	18.00
	$NRDM^{PD}$	60.50	56.00	50.00	87.20	76.50	26.40	18.50	13.60	14.50	6.00
	FDA^{PD}	64.80	60.10	55.60	92.10	78.00	28.80	21.50	13.80	15.50	7.10
	FIA^{PD}	90.00	88.40	80.70	96.70	80.20	71.00	69.60	58.60	53.00	34.50
	NAA^{PD}	91.00	88.00	82.20	96.70	81.00	74.20	70.00	60.10	57.60	38.80
	Ours	**92.80**	**89.90**	**88.10**	97.30	**90.70**	**77.10**	**73.50**	**64.30**	**61.50**	**47.60**
Vgg16	MIM^{PD}	81.30	80.00	73.90	84.20	**99.99**	65.50	61.10	52.40	51.00	45.50
	$NRDM^{PD}$	75.60	76.80	55.80	72.00	93.00	70.10	63.90	60.20	55.60	48.00
	FDA^{PD}	86.00	79.60	66.20	79.90	95.60	74.50	66.60	66.80	66.90	58.60
	FIA^{PD}	92.60	91.20	86.00	91.30	99.80	84.20	78.00	85.60	80.00	68.80
	NAA^{PD}	95.70	95.70	**93.80**	92.30	99.80	89.90	86.00	**87.80**	83.60	74.20
	Ours	**96.20**	**96.00**	92.70	**93.00**	99.92	**90.00**	**89.80**	86.60	**86.50**	**77.20**

表4中最优的实验结果也已经加粗显示. 同样在白盒的对比实验中,MIMPD算法取得了最高的攻击成功率,但主特征归因算法的攻击成功率也均高于了已有方法,并且与MIMPD算法的最高值相差在1%之内. 在黑盒对比实验中,针对无防御模型,本文的方法,相比于已有算法均取得了最高的攻击成功率,并且最高提高了4%以上. 针对防御模型,其攻击成功率也均有所提高. 例如,基于inception_v3模型生成的对抗样本,其在inception_v3模型训练的防御模型上的攻击成功率相较其他方法,均提高了5%以上,在非inception_v3上训练的防御模型,均提高了15%以上.

综上,在结合随机转换方法后,算法生成的对抗样本也得到了更高的攻击成功率,证明了本文主特征提取后的样本保留了合适的特征部分,并为特征归因效率提升奠定了基础. 经过大量的实验证明,所提出的主特征归因算法可以更高效地生成更具迁移性的对抗样本.

4.4.3 结合输入转换防御方法的实验结果对比

本节将现有的防御模型,结合输入转换防御方法进行实验验证. 实验以inception_v3作为源模型生成对抗样本,然后进行像素平滑处理,再对目标模型进行攻击,计算10次实验结果的平均值,作为最后的实验结果,最优的实验结果已经加粗显示.

整体上,对抗样本经过像素平滑处理后,会减少扰动对样本的影响,从而达到对抗防御的目的. 如表5,对抗样本针对目标模型的攻击成功率,相比于未结合输入转换防御方法的有所降低. 但相比于已有算法,本文方法的白盒攻击成功率取得了最高值. 在对其他模型的迁移性黑盒攻击上,也取得了最高的攻击成功率. 实验证明,本文算法在针对多防御方法的攻击中,仍具有较好的有效性.

表5　结合输入转换防御方法的实验结果对比表　　单位:%

源模型	攻击算法	Inc_v3	Inc_v4	IncRes_v2	Res_v2	Vgg_16	Adv_Inc_v3	Adv_IncRes_v2	Ens3_Inc_v3	Ens4_Inc_v3	Ens3_IncRes_v2
Inc_v3	FDA	95.80	74.50	61.80	45.20	56.10	20.20	21.00	16.80	16.00	8.50
	FIA	95.20	77.20	76.00	70.00	70.00	55.60	52.00	32.60	29.60	20.00
	NAA	96.60	80.60	77.60	72.60	72.90	59.30	60.20	50.00	48.40	30.00
	Ours	**97.70**	**85.20**	**81.80**	**77.60**	**84.80**	**70.20**	**68.00**	**60.00**	**57.50**	**43.80**

4.4.4　主特征归因算法时间效率对比

本文在64 GB内存的i5-1240P处理器,GeForce RTX 3080Ti显卡的环境下,将inception_v3作为源模型,在与所有对比实验相同参数设置下,记录生成对抗样本的时间,及对应白盒攻击成功率,结果见表6.

表6　算法效率对比

方法	攻击成功率/%	样本生成时间/h
FDA	82.10	15.65
FIA	98.30	13.08
NAA	98.10	10.43
Ours	**98.30**	**5.86**

本文算法在取得最高攻击成功率的同时,拥有小的时间成本,证明算法中主特征提取部分的时间成本,远远小于未使用独立性假设进行近似计算的时间成本. 同时证明更少的神经元聚合次数以及更大的扰动添加步长,在保证样本迁移性的同时,提高了生成效率,证明了主特征归因算法的有效性.

5　结论

经过实验证明,所生成的对抗样本在黑白盒攻击的对比实验中,均表现出了出色的攻击成功率. 在效率方面,通过独立性假设对算法的简化,以及对损失函数的优化,主特征归因算法可以用更少地迭代次数、更少的神经元聚合数得到迁移性更强的对抗样本. 该算法的提出为后续研究如何提高深度学习模型的鲁棒性究奠定了基础.

参考文献

[1] 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)

[2] 张思思, 左信, 刘建伟. 深度学习中的对抗样本问题[J]. 计算机学报, 2019, 42(8): 1886-1904.
ZHANG S S, ZUO X, LIU J W. The problem of the adversarial examples in deep learning[J]. Chinese Journal of Computers, 2019, 42(8): 1886-1904. (in Chinese)

[3] 李盼, 赵文涛, 刘强, 等. 机器学习安全性问题及其防御技术研究综述[J]. 计算机科学与探索, 2018, 12(2): 171-184.
LI P, ZHAO W T, LIU Q, et al. Security issues and their countermeasuring techniques of machine learning: A survey[J]. Journal of Frontiers of Computer Science and Technology, 2018, 12(2): 171-184. (in Chinese)

[4] GOODFELLOW I, SHLENS J, SZEGEDY C .Explaining and harnessing adversarial examples[EB/OL]. (2014-12-19)[2023-09-27]. https://arxiv.org/pdf/1412.6572.pdf.

[5] KURAKIN A, GOODFELLOW I J, BENGIO S. Adversarial examples in the physical world[M]//First Edition.Artificial Intelligence Safety and Security. Boca Raton: CRC Press/Taylor & Francis Group, 2018: 99-112.

[6] 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.

[7] 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.

[8] MADRY A, MAKELOV A, L SCHMIDT, et al. Towards deep learning models resistant to adversarial attacks[EB/OL]. (2014-12-19)[2023-09-27]. https://arxiv.org/pdf/1412.6572.pdf.

[9] SZEGEDY C, ZAREMBA W, SUTSKEVER I, et al. In-

triguing properties of neural networks[C/OL]//The 2nd International Conference on Learning Representations. Washington DC: ICLR, 2014. [2023-09-27]. https://www.mendeley.com/catalogue/76b4e01d-6cc1-31dd-a3a5-df1354e1cf8d/.

[10] CARLINI N, WAGNER D. Towards evaluating the robustness of neural networks[C]//2017 IEEE Symposium on Security and Privacy (SP). Piscataway: IEEE, 2017: 39-57.

[11] LIU X Z, LI L, WANG X Y, et al. Adversarial examples generated from sample subspace[J]. Computer Standards & Interfaces, 2022, 82: 103634.

[12] XU Q L, TAO G H, ZHANG X Y. Bounded adversarial attack on deep content features[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2022: 15182-15191.

[13] LI P C, YI J F, ZHANG L J. Query-efficient black-box attack by active learning[C]//2018 IEEE International Conference on Data Mining (ICDM). Piscataway: IEEE, 2018: 1200-1205.

[14] CHEN P Y, ZHANG H, SHARMA Y, et al. ZOO: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models[C]//Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security. New York: ACM, 2017: 15-26.

[15] PAPERNOT N, MCDANIEL P, I GOODFELLOW, et al. Practical black-box attacks against deep learning systems using adversarial examples[EB/OL]. (2016-02-08)[2023-09-27]. https://arxiv.org/pdf/1602.02697v2.pdf.

[16] DUAN M X, LI K L, DENG J Y, et al. A novel multi-sample generation method for adversarial attacks[J]. ACM Transactions on Multimedia Computing, Communications, and Applications, 2022, 18(4): 1-21.

[17] ZHOU W, HOU X, CHEN Y J, et al. Transferable adversarial perturbations[C]//Computer Vision - ECCV 2018. Cham: Springer International Publishing, 2018: 471-486.

[18] HUANG Q A, KATSMAN I, GU Z Q, et al. Enhancing adversarial example transferability with an intermediate level attack[C]//2019 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway: IEEE, 2020: 4732-4741.

[19] NASEER M, KHAN S H, RAHMAN S, et al. Task-generalizable adversarial attack based on perceptual metric[EB/OL]. (2018-1122)[2023-09-27].https://arxiv.org/pdf/1811.09020.pdf.

[20] GANESHAN A, VIVEK B S, RADHAKRISHNAN V B. FDA: Feature disruptive attack[C]//2019 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway: IEEE, 2020: 8068-8078.

[21] WANG Z B, GUO H C, ZHANG Z F, et al. Feature importance-aware transferable adversarial attacks[C]//2021 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway: IEEE, 2022: 7619-7628.

[22] ZHANG J P, WU W B, HUANG J T, et al. Improving adversarial transferability via neuron attribution-based attacks[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2022: 14973-14982.

[23] 孔锐, 蔡佳纯, 黄钢. 基于生成对抗网络的对抗攻击防御模型[J/OL]. 自动化学报, 2020. DOI: 10.16383/j.aas.2020.c200033.
KONG R, CAI J C, HUANG G. Defense to adversarial attack with generative adversarial network[J/OL]. Acta Automatica Sinica, 2020. DOI: 10.16383/j.aas.2020.c200033. (in Chinese)

[24] SHAFASHI A, NAJIBI M, GHIASI M A, et al. Adversarial training for free! [C]//Neural Information Processing Systems 32. San Diego: NIPS, 2019: 3358-3369.

[25] ZHU C, CHENG Y, GAN Z, et al. FreeLB: Enhanced adversarial training for natural language understanding[C]// International Conference on Learning Representations. New Orleans: Ithaca, Computational and Biological Learning Society, 2019: 1-14.

[26] ZHANG D, ZHANG T, LU Y, et al. You only propagate once: Accelerating adversarial training via maximal principle[C]//Neural Information Processing Systems 32. San Diego: NIPS, 2019: 227-238.

[27] LI T, WU Y W, CHEN S Z, et al. Subspace adversarial training[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2022: 13399-13408.

[28] 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.

[29] SZEGEDY C, IOFFE S, VANHOUCKE V, et al. Inception-v4, inception-ResNet and the impact of residual connections on learning[EB/OL]. (2016-02-23)[2023-04-28]. https://arxiv.org/abs/1602.07261.

[30] 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.

[31] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[C/OL]//The 3rd International Conference on Learning Representations (ICLR2015). Ithaca: Computational and Biological Learning Society, 2015. [2023-09-27]. https://www.mendeley.com/catalogue/949ce8a0-2d23-32b0-9dbb-b86394a92c62/.

作者简介

王　硕　男,1999年出生于河北省沧州市. 现为华北电力大学硕士研究生. 主要研究方向为人工智能安全、对抗防御.
E-mail: 120212227097@ncepu.edu.cn

徐茹枝　女,1966年出生于江西省上饶市. 现为华北电力大学教授、硕士生导师. 主要研究方向为人工智能安全、智能电网.
E-mail: xuruzhi@ncepu.edu.cn

关志涛(通讯作者)　男,1979年出生于辽宁省沈阳市. 现为华北电力大学教授、博士生导师. 主要研究方向为人工智能安全、区块链、隐私计算.
E-mail: guan@ncepu.edu.cn