基于深度神经网络的遗传算法对抗攻击

范海菊,马锦程,李名

(河南师范大学 计算机与信息工程学院;河南省教育人工智能与个性化学习重点实验室,河南 新乡 453007)

摘 要:深度神经网络(deep neural network,DNN)能够取得良好的分类识别效果,但在训练图像中添加微小扰动进行对抗攻击,其识别准确率会大大下降.在提出的方法中,通过遗传算法得到最优扰动后,修改图像极少的像素生成对抗样本,实现对 VGG16 等 3 个基于卷积神经网络图像分类器的成功攻击.实验结果表明在对 3 个分类模型进行单像素攻击时,67.92% 的 CIFAR-10 数据集中的自然图像可以被扰动到至少一个目标类,平均置信度为 79.57%,攻击效果会随着修改像素的增加进一步提升.此外,相比于 LSA 和 FGSM 方法,攻击效果有着显著提升.

关键词:卷积神经网络;遗传算法;对抗攻击;图像分类;信息安全

中图分类号:O413 **文献标志码**:A **文章编号**:1000-2367(2025)02-0082-09

计算机视觉与深度学习的结合是一个前景广阔的研究领域,目前常见的应用有图像分类识别、目标检测和图像风格迁移等.特别是在图像分类识别领域,深度神经网络通过模仿人脑的独立皮质神经元工作方式,充分利用像素之间的距离与其相似性的关系,克服了传统神经网络计算量大、适配性低的缺点,能够以较高的准确率执行分类问题,甚至达到了与人眼分类相媲美的程度[1].

然而,当前的多项研究表明,深度神经网络的输入向量是不连续的,对抗样本的出现更可能是数据维度较高和训练数据不足导致的,其预测结果对添加的扰动相当敏感.因此在对深度神经网络进行对抗攻击的过程中,通过向自然图像添加一定的扰动生成对抗样本[2-4],从而完成对神经网络的攻击.常见的生成对抗样本的方法需要修改较多的像素数量以达到攻击效果,因此无法保证修改后的图像达到未被察觉出来的效果.如图 1 所示,曾对汽车和鸟类等图像总像素的 4% 进行扰动[5],使得深度神经网络识别错误,但这种异常能够被专业技术轻松识别出来.

针对上述问题,本文提出了利用遗传算法生成对抗样本的方法,将修改像素的数量限制在尽可能少的范围内,而不是在理论上提出额外的约束或考虑其他的损失函数进行扰动.提出的方法将修改像素的数量作为衡量扰动强度的参考,并考虑仅修改一个像素的情况,以及与其他两种情况(即 3 和 5 像素)和其他方法进行比较,展示了所提方法的有效性.

本文第 1 节介绍相关工作,包括对抗攻击的研究现状和常见类型以及遗传算法的基础知识;第 2 节介绍本文提出的基于深度神经网络的遗传算法对抗攻击;第 3 节通过分析实验数据验证了所提方法的有效性;最后进行总结.

收稿日期:2023-09-21;**修回日期**:2023-10-10.

基金项目:国家自然科学基金(61602158);河南省科技攻关计划(222102210029);河南省高等学校重点科研项目(23A520009).

作者简介(通信作者):范海菊(1979—),女,河南新乡人,河南师范大学副教授,博士,研究方向为网络安全,E-mail:121064@htu.edu.cn.

引用本文:范海菊,马锦程,李名.基于深度神经网络的遗传算法对抗攻击[J].河南师范大学学报(自然科学版),2025,53(2):82-90.(Fan Haiju,Ma Jincheng,Li Ming.Genetic algorithm against attack based on deep neural network[J].Journal of Henan Normal University(Natural Science Edition),2025,53(2):82-90.DOI:10.16366/j.cnki.1000-2367.2023.09.21.0003.)

1　相关工作

卷积神经网络可以把图像当作向量输入到神经网络里，通过卷积、池化等提取特征，解决真实世界图像分类的问题.图像分类工作包含：将图像分解成部分重合的小图块输入小型神经网络、将输出结果保存到新的数列中、将数列中各个小方阵数列的异常值保留和做出预测 4 个部分.在一定范围内卷积层越多，神经网络就越能识别出复杂的特征.传统的卷积神经网络难以处理更多的训练数据，因此扩大神经网络，把节点一层一层的堆积起来形成了深度神经网络(DNN)，图 2 展示了深度神经网络的基本框架.

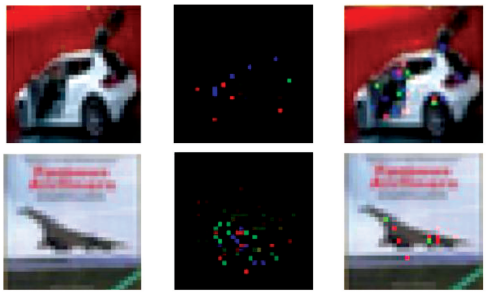


图1　修改4%的像素数量生成的对抗样本

Fig.1　Modifying 4% of the pixel count to generate the adversarial sample

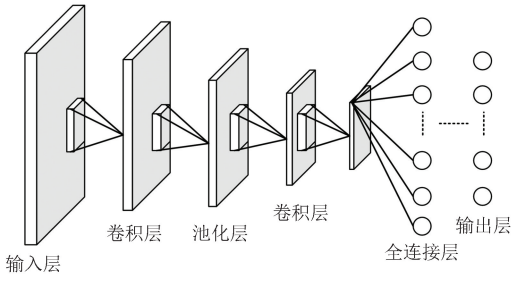


图2　深度神经网络基本框架

Fig.2　The basic framework of deep neural networks

当前的研究表明，DNN 容易受到对抗样本的恶意攻击，对图像数据添加人眼难以辨识到的细微扰动就能使模型产生误判，结合特定算法实施对抗攻击生成对抗样本，诱导模型做出错误的甚至是指定的分类结果.另外，通过裁剪图像和旋转图像也会产生对抗样本，此类对抗攻击方法也存在于语音识别、自然语言处理、恶意软件分类等领域.当前研究在不同的设想下提出了添加扰动的计算方法，如文献[6]在神经网络的敏感性是由于输入的高维和线性引起的假设下，提出了计算扰动的“快速梯度符号”算法，文献[7]通过决策边界的线性假设提出了一种贪婪扰动搜索方法.此外，文献[8]首先向良性样本中添加随机的噪声模拟物理世界的环境因素，并计算这些噪声样本上产生的梯度期望，进而实现物理世界的对抗攻击.文献[9]进一步考虑了掩膜和制造误差从而实现了交通标志的对抗性扰动，这些都验证了物理对抗样本的存在.

1.1　对抗攻击

机器学习领域的对抗攻击指的是向输入的数值型向量添加微小扰动使分类器产生错误分类结果的过程，图 3 展示了向原始图像添加噪声的过程.无论是逻辑回归、决策树还是神经网络，几乎没有任何一种机器学习算法能够免受对抗性攻击，这可能是由于系统中的过度线性拟合造成的.虽然深度学习的每段架构都是基于激活函数的线性结构，但分类器的线性响应特别容易产生非线性的决策边界，即使是添加少量的噪声也可以影响决策边界的判断.对抗攻击是信息安全领域的主要风险之一，常见的攻击方法就是生成对抗样本输入模型，干扰模型的识别精度，从而使模型作出错误分类判断.因此，可以通过对抗学习研究对抗样本的生成和攻击模型的过程，增强神经网络对目标识别的可靠性和机器学习模型的抗欺骗、抗干扰能力.

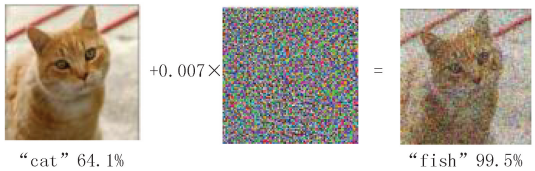


图3　对抗样本生成过程举例

Fig.3　Example of counter sample generation process

根据被攻击模型的分类结果，对抗攻击可以分为非目标攻击和目标攻击.一张图像可以被看作是一个具体的向量，假设在这个向量的每一维上都添加人眼无法识别的微小扰动，并将添加扰动后的图像输入神经网络中使得神经网络的分类结果出现错误甚至取得预计的分类结果的过程被称作是对抗攻击.未被添加扰动的图像作为原始图像，其类别被称作真实标签，添加扰动信息后的图像作为对抗样本，其类别被称作预测标签.根据攻击结果可以分为目标攻击和非目标攻击，在本文中将原始类别添加扰动后干扰到剩余的 9 个类别之一就完成了非目标攻击，在非目标攻击的基础上能够使得预测类别达到期望分类结果就完成了目标攻击.

根据对被攻击模型的内部结构了解程度,对抗攻击可以进一步分为黑盒攻击、灰盒攻击和白盒攻击.当前最常见的图像对抗攻击方法是对全图添加扰动,这种方法的缺点在于工作量大且容易被人眼识别出来,因此,NARODYTSKA 等[10]此前提出了修改一个像素改变图像对应类标签的方法,只将其作为一个起点,进一步推导出半黑盒攻击.白盒攻击中,攻击者可以拥有关于模型架构和参数的全部信息,例如在白盒攻击中 PGD 这一类基于梯度的方法[11],可能是由于在梯度信息存在的时候利用梯度进行基于优化的攻击是最有效.在灰盒攻击中,攻击者仅可获取模型的结构信息但无法获得模型参数,没有模型的查询权限.在黑盒攻击中,攻击者对攻击模型的内部结构、训练参数、防御方法等一无所知,只能通过输入输出以及预测结果与模型进行交互.根据预测结果是置信度还是 one-hot 向量又分为软标签和硬标签攻击,例如黑盒迁移攻击是对抗攻击中非常热门的一个研究方向,特别是基于动量梯度的黑盒攻击方法.此外,当前大部分研究主要通过在数据样本的尺寸、分布、规模、时序等方面来丰富梯度的多样性,使得生成的对抗样本在迁移到其他的模型攻击时,能够有更高的攻击成功率[12].

1.2 遗传算法

遗传算法(genetic algorithm,GA)是一种借鉴达尔文进化论中生物自然进化形成的优化求解方法,最初由文献[13]提出.遗传算法模拟染色体之间的交叉或变异,最终选择符合当前问题的最优种群,是一种高效的全局性自适应随机搜索方法.常见的对抗攻击算法,如 Fast Gradient Sign Method、Project Gradient Descent 和 Basic Iterative Method 都要求目标函数具有可导性,但遗传算法不要求优化对象的连续性和可微性,并具有良好的鲁棒性和规则学习能力.遗传算法从问题的可能解集出发构造一个种群,种群中的每一个个体对应问题的一个可能解,并通过适应度函数的值来反映每个个体能否作为当前最优解被保留,遗传算法流程图如图 4 所示.

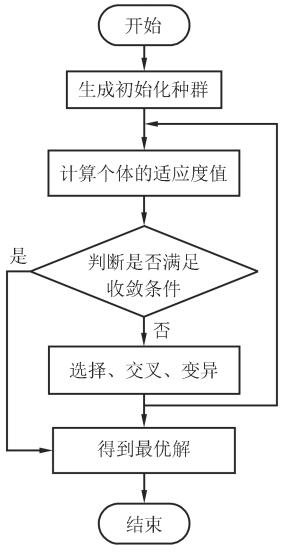


图4 遗传算法流程图

Fig.4 The genetic algorithm flow chart

现有的对抗方法常采用区域修改或多像素修改生成对抗样本,干扰分类模型从而输出错误分类,但本文提出了基于深度神经网络的遗传算法对抗攻击方法,通过对原始图像样本进行概率阈值筛选并根据遗传算法的特点控制攻击像素数量,从而实现了修改极少数像素特别是修改单个像素就能实现对神经网络的对抗攻击.该方法通过遗传算法生成扰动,既能求出全局最优解又避免了求解梯度.通过分析探究相似类别的图像之间实现对抗的难易程度和图像决策边界对生成对抗样本的影响可以在保护图像真实性、防止恶意攻击方面提供新的思路.

2 基于深度神经网络的遗传算法对抗样本生成

通过对抗样本进行对抗攻击可以使深度神经网络的分类出现错误,基于此本文提出了利用遗传算法生成扰动(微小噪声)将其与原始图像结合生成对抗样本.网络分类模型的输出包含预测标签和置信度两部分,置信度是 softmax 的输出结果,指图像作为某种类别的概率,预测标签是在数据集中相关置信度对应的数字标签.对抗攻击的原理是通过添加扰动,提高输入对抗图像预测标签的置信度,降低真实标签的置信度,从而使神经网络分类错误.为改进攻击成功率和置信度,本文提出基于遗传算法的对抗样本生成.首先,通过深度神经网络对原始图像进行分类,筛选 softmax 值低于 0.999 的图像得到较易实现对抗攻击的微处理图像.其次,利用遗传算法对扰动进行编码,经过选择和交叉操作生成最优扰动.最后,在微处理图像中嵌入扰动生成对抗样本,将对抗样本输入分类器中,查看分类结果,判断是否实现对抗攻击(包括非目标攻击和目标攻击),总体流程图如图 5 所示.

具体步骤如下.

1）输入原始图像：将原始图像送入深度神经网络中进行分类，网络模型通过卷积、池化等操作提取特征，最终输出原始图像的类别标签和置信度.

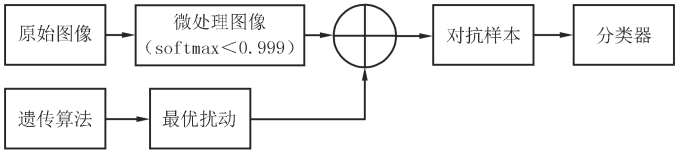


图5　实验流程图

Fig.5　The flow chart of the experiment

2）筛选图像：网络模型中的置信度值由 softmax 函数输出，选取置信度值小于 0.999 的图像进行有限个像素点的修改可以取得较好的攻击结果.

3）生成最优扰动：为方便嵌入原始图像，通过遗传算法将扰动编码为一个五元组并作为种群进化的候选解构成相关解集，在种群进化过程中比较子代和父代的适应度值，得出最优解.

4）生成对抗样本：根据遗传算法得到的最优扰动编码找到需要修改的像素点位置，修改原始图像的像素值，从而生成对抗样本.

5）判断是否攻击成功：将对抗样本输入分类器，判断原始图像的类别标签与对抗样本的类别标签是否相等，类别标签相等说明攻击失败，反之，说明攻击成功.

2.1　生成对抗样本

对抗样本生成过程包括扰动生成和对抗嵌入两个部分，通过遗传算法可以在解空间内找到可添加扰动的全局最优解，与优化算法相比跳出局部最优化的困境.对于实际问题，遗传算法通过编码生成候选解并构成相应的解集，把适应度函数值作为衡量可能解优劣程度的标准，为避免陷入局部最优而引入交叉，保证了随机性和灵活性，在每次进化中选择适应度值大的解构成新的解集，最终取得最优解集.

种群进化算法流程图如图 6 所示，实验中将种群的规模设置为 300，交叉因子选择为 0.5，最大迭代次数为 $T=50$.编码后的扰动组成的种群通过进化产生最优种群，P_i 表示当前种群，P_{i+1} 表示下一代种群，x 表示原始图像，$e(x)$ 表示添加的扰动，l、p 分别表示深度神经网络输出预测标签和概率，l_t 表示原始类别标签.在生成最优种群的过程中，为降低真实类别概率，在最大迭代次数范围内选择父代扰动或子代扰动进行进化：当父代扰动和子代扰动生成的对抗样本标签都不等于原始类别标签时，说明两代扰动都能攻击成功，选择概率大的扰动；当父代扰动和子代扰动生成的对抗样本标签相等时，说明两代扰动不能实现攻击，选择概率较小的扰动.

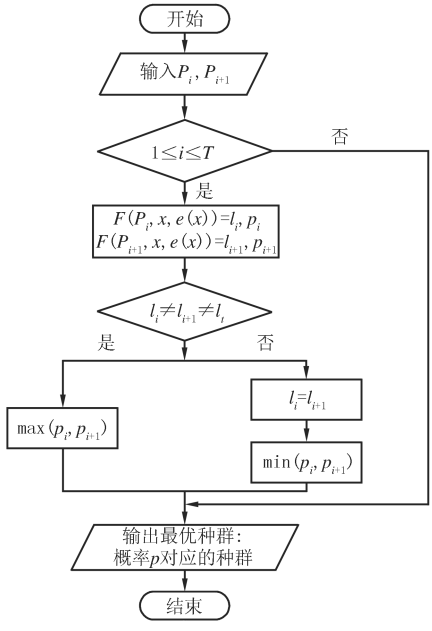


图6　种群进化算法流程图

Fig.6　The flow chart of population evolution algorithm

下面具体说明遗传算法产生扰动的步骤.

1）初始化.扰动被编码成一个包括 5 个元素的元组 $C=(x,y,R,G,B)$，分别为 x 坐标、y 坐标和 RGB 三通道的值，这样编码的目的是便于将扰动嵌入原始图像，噪声的嵌入过程也就是将 $C=(x,y,R,G,B)$ 覆盖在原始图像的像素点位置并修改原始图像的三通道值.该元组作为 GA 算法的可能解，每个解集都由固定数量的可能解组成，遗传算法从初始解集出发进行进化.

2）交叉.实验提出的遗传算法通过交叉完成种群进化，迭代一次相当于进化一次，在每次迭代过程中都会产生新的解.交叉需要从上一代解集中选择两个可能解，根据交叉因子随机交叉.$P(i)$ 表示第 i 个种群，$P(j)$ 表示第 j 个种群，$i\neq j$，下标 t 表示种群代数之间的关系.θ 是交叉因子，$\theta=0.5$.

$$P_{t+1}(i)=P_t(i)+\theta P_t(j). \tag{1}$$

3）选择.提出的方法设置适应度函数值为深度神经网络输出的图像类别概率标签，攻击过程中不断降低图像真实类别分类概率，在进化过程中比较子代和父代扰动的适应度值，选择真实类别概率低的扰动进入下一

次迭代.

4)迭代.重复交叉和选择步骤,在最大迭代次数内真实类别概率最低的扰动作为最优扰动.

2.2 对抗攻击

2.2.1 对抗攻击原理

衡量基于遗传算法的图像对抗攻击效果,包括非目标攻击和目标攻击两个部分.假设一个输入图像可以用一个 n 维向量 $\boldsymbol{x}=(x_1,x_2,\cdots,x_n)$ 表示,其中每个标量元素代表一个像素,函数 F 表示图像分类器,t 为不同输入图像对应的分类,adv 为添加扰动后生成的对抗样本的预期分类,$F_t(\boldsymbol{x})$ 表示分类模型认为输入图像是类别 t 的概率,n 维向量 $\boldsymbol{e}(\boldsymbol{x})=(e_1,e_2,\cdots,e_n)$ 为在原图像 x_0 上添加的扰动,对抗攻击可以简化为一个带有约束条件的优化问题,该问题涉及两个条件:1) 哪些维度需要扰动;2) 每个维度对应的扰动强度.此外,在非目标攻击中,为达到攻击目的,期望加入扰动后分类器模型将图像分类为 t 类别的概率越小越好.在目标攻击中,为达到攻击目的,期望加入扰动后分类器模型将图像分类为目标类别的概率越大越好.在之前的研究中将最大修正 L 作为添加扰动 $e(\boldsymbol{x})$ 的限制性条件:

$$\begin{cases}\min F_t(\boldsymbol{x}),\max\limits_{e(x)^*} F_{\text{adv}}(\boldsymbol{x}+e(\boldsymbol{x})),\\ \|e(\boldsymbol{x})\|\leq L.\end{cases}\tag{2}$$

这种方法得到的对抗样本通常是通过修改所有维度的一部分进行扰动,并对累积修改强度进行总体约束来构造的[7].本文提出的单像素攻击与这种攻击思路恰恰相反,仅针对修改的像素数量 d 进行限制,但不限制修改的 RGB 值,在对应取值范围中搜索能改变类别属性的值,对抗攻击前先将图像送入分类器筛选,选择 softmax 输出结果低于 0.999 的图像添加扰动,如下所示:

$$\begin{cases}\min F_t(\boldsymbol{x}),\max\limits_{e(x)^*} F_{\text{adv}}(\boldsymbol{x}+e(\boldsymbol{x})),\\ \|e(\boldsymbol{x})\|\leq d, f(x_0)<0.999.\end{cases}\tag{3}$$

d 指修改像素点的数量,x 指原始图像表示的 n 维向量 $\boldsymbol{x}=(x_1,x_2,\cdots,x_n)$,$e(\boldsymbol{x})$ 指对原始图像添加的扰动向量.单像素攻击时 $d=1$,意味着在扰动的过程中只有一个维度被修改,也就是说,单像素攻击允许在 n 维向量 $\boldsymbol{x}=(x_1,x_2,\cdots,x_n)$ 中,以任意的强度向选定的 x_i 方向修改图像,同理可推导到三像素攻击和五像素攻击.

假设神经网络是一个内部参数固定的函数,将原始图像和对抗样本定义为 x_0 和 x_1,定义损失函数 L_1 为输出结果 $\boldsymbol{y}$ 与真实结果 $\boldsymbol{y}'$ 交叉熵的相反数,交叉熵的值越大损失函数越小,即 $\boldsymbol{y}$ 与 $\boldsymbol{y}'$ 差距越大,攻击效果越好.将 L_1 作为非目标攻击的损失函数:

$$L_1(\boldsymbol{x})=-e(\boldsymbol{y},\boldsymbol{y}').\tag{4}$$

在目标攻击中用 $\boldsymbol{y}_{\text{target}}$ 表示目标攻击结果,$\boldsymbol{y}$ 与 $\boldsymbol{y}_{\text{target}}$ 都是 one-hot 向量,定义损失函数 L_2 为输出结果 $\boldsymbol{y}$ 与目标结果 $\boldsymbol{y}_{\text{target}}$ 的交叉熵.非目标攻击的实现指的是将原始图像的分类扰动为其他分类,目标攻击是在非目标攻击的基础上进行的,它进一步将原始图像的分类扰动为指定分类,因此定义 L_3 为目标攻击的损失函数,即 L_1 与 L_2 的总和.

$$L_2(\boldsymbol{x})=e(\boldsymbol{y},\boldsymbol{y}_{\text{target}}),\tag{5}$$

$$L_3(\boldsymbol{x})=-e(\boldsymbol{y},\boldsymbol{y}')+e(\boldsymbol{y},\boldsymbol{y}_{\text{target}}).\tag{6}$$

2.2.2 攻击结果分析

对抗攻击结果根据要求不同分为非目标攻击和目标攻击,其中目标攻击的结果是在非目标攻击的基础上得出的,p 表示图像的输出概率,设图像数量为 100 张,l_i 表示原始图像类别标签,l'_i 表示非目标攻击对抗样本类别标签,l'_{adv} 表示目标攻击对抗样本类别标签.攻击结果算法流程图如图 7 所示.

3 实验结果

3.1 实验条件准备

1)实验环境:Windows10 专业版操作系统,AMD R7-4800 处理器,NVIDA GeForce RTX 2060 显卡,Python 解释器版本为 3.7.11,深度学习框架为 Pytorch1.9.0.

2）实验对象：实验将深度神经网络 All convolution network[14−16]、Network in Network[17]和 VGG16[18−20]作为图像分类器，向其中输入一张图像并根据神经网络的输出结果判断该图像的类别[21].

3）实验数据集：实验在 CIFAR-10 数据集上进行[17]，数据集图像尺寸大小为 32×32，包含 10 个不同类别的 *RGB* 彩色图像（飞机、汽车、鸟、猫、鹿、狗、青蛙、马、船、卡车），其对应的标签为 0 到 9.

3.2　实验结果及分析

在每次攻击中从 CIFAR-10 数据集随机选择 500 张图像，攻击方式为目标攻击和非目标攻击，在不同的攻击方式下分别完成单像素、三像素和五像素攻击.对于 1 张自然图像，通过添加扰动可以使它被错误分类到其他 9 个类，实际上非目标攻击的过程包含在目标攻击中，非目标攻击的有效性是根据目标攻击结果来评估的.也就是说，如果 1 幅图像能被扰动到 9 个类中至少 1 个目标类，那么对该图像的非目标攻击就成功了.

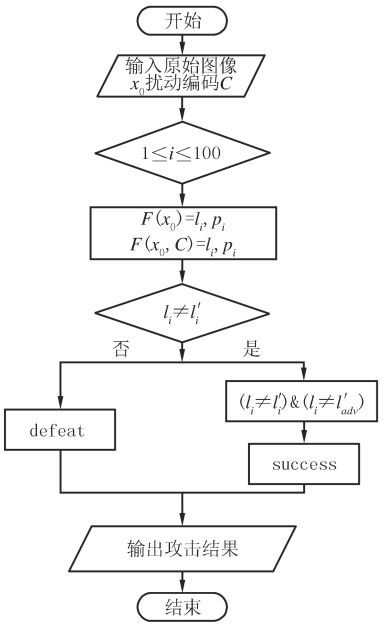


图7　攻击结果分析算法流程图

Fig. 7　The flow chart of attack result analysising algorithm

3.2.1　对抗图像视觉质量

在本文实验中，先将 CIFAR-10 数据集中的自然图像分别送入 All convolution network[14−16]、Network in Network[17]和 VGG16[18−20]3 个网络中进行训练，让网络能够准确给出图片分类预测结果，再通过遗传算法（GA）找到需要修改的像素点坐标和对应的 *RGB* 三通道数值，更改原始图像生成对抗样本，降低真实类标签的置信度，只关注输出结果与模型的交互作用而不考虑深度神经网络的内部结构（黑盒攻击），让图像分类器错误的分类结果.图 8 展示了对 3 种网络进行单像素目标攻击部分结果，每幅图像下方标注为原始图像类别－对抗样本类别－对抗样本置信度，可知仅仅修改 1 个像素点对图像的视觉呈现效果影响较小，每个图片下方展示了图像真实类别，对抗攻击后非真实类别以及置信度，神经网络对于对抗图像分类后输出的置信度越高对抗效果越好，其中多数置信度在 70％以上.实验结果表明，提出的方法对 3 种类型的深度神经网络均可达到攻击目的且效果明显.

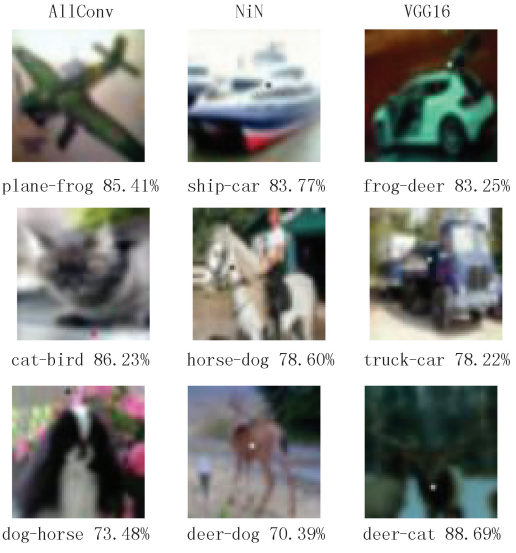


图8　3种网络进行单像素目标攻击部分结果

Fig. 8　Three kinds of networks carry out single-pixel target attack partial results

3.2.2　成功率和置信度

原始成功率被定义为深度神经网络对原始图像的分类识别准确率.在非目标攻击的情况下，成功率被定义为将图像扰动为任意其他类别时占总输入图像数量的比例.在目标攻击的情况下，成功率被定义为将图像扰动为一个特定目标类别时占输入图像数量的比例.置信度被定义为累加每次成功扰动为目标类别概率值与成功扰动总数的比值，表示的是对抗样本使图像分类器产生“误分类”的平均置信度.在 CIFAR-10 数据集上对不同神经网络进行单像素攻击的结果如表 1 所示.实验结果表明，非目标攻击成功率与原始成功率相差

不大,说明对抗图像对神经网络具有良好的攻击效果,目标攻击的成功率基本在 20%以上,进一步证明了所提出攻击方法的有效性.

表 1 3 种网络单像素攻击结果

Tab. 1 Three kinds of network single pixel attack results %

结果分析指标	AllConV[14-16]	NiN[17]	VGG16[18-20]	结果分析指标	AllConV[14-16]	NiN[17]	VGG16[18-20]
原始成功率	86.94	87.70	85.45	非目标攻击成功率	70.85	67.00	65.91
目标攻击成功率	21.33	18.33	25.89	平均置信度	78.24	82.80	77.67

3.2.3 原始目标类对

原始目标类对指的是原始图像类别及其对抗图像类别,附录图 S1~S4 展示了对神经网络进行攻击的热图矩阵,每个热图矩阵大约是对称的,热图矩阵横坐标表示的是原图像类别,热图矩阵纵坐标表示的是对抗图像类别,主对角线数值表示的是神经网络分类成功次数,其他小方框的数值表示的是对神经网络进行对抗攻击的成功次数.通过观察图 S1~S4 可知:1)一些特定的原始目标类对比其他类更容易受到攻击,说明脆弱目标的类别是由属于同一类的不同数据点共享的.例如,船类容易添加扰动生成对抗样本,但其他类别的图像很难生成船类对抗样本,这可能还与图像的决策边界有关.2)一些图像容易被扰动到其他类别,例如猫类图像和狗类图像容易修改像素点互相扰动,两类图像较为相似,这是由于神经网络本身提取特征的特性导致,说明了攻击效果与神经网络特性密切相关.3)攻击成功率与修改像素点的数量呈正相关,将像素的数量从 1 增加到 5 进行高维扰动时,热图矩阵中的数值明显增加.

3.2.4 适应度值变化

适应度值被设置为每个图像的真实类的概率标签,攻击的目标是使适应度值最小化.这里以 VGG16 攻击 100 张图像为例,分别展示单像素攻击和三像素攻击目标攻击后每张图像的适应度值变化来探究修改像素点数量与适应度值变化之间的关系.附录图 S5 中一半数量以上的图像适应度值低于 50%,附录图 S6 中大多数图像的适应度值低于 10%,适应度值与图 S5 相比明显下降.实验结果表明了随着修改像素点数量的增加,攻击的效果越好,为达到不同效果可以按要求更改像素点数量,也反映了提出方法的灵活性.

3.2.5 攻击方法对比

将实验中提出的攻击方法与随机攻击(random attack,RA)进行比较,评估遗传算法对 CIFAR-10 数据集进行单像素非目标攻击的有效性.对于每一幅自然图像,随机搜索重复 100 次,每次随机修改图像中随机 *RGB* 值的一个随机像素,试图改变其标签并将攻击对一幅图像的置信度设为 100 次攻击中概率最高的目标类标签.通过对表 2 进行比较发现,对于 All convolution network[14-16]、Network in Network[17] 和 VGG16[18-20]3 种深度神经网络,遗传算法的效率分别比随机搜索方法提高了 21.15%、25.82%和 50.34%,因此可以得出结论,相比随机攻击,GA 能显著提升攻击效率.

表 2 遗传算法与随机攻击结果比较

Tab. 2 The result comparing about GA and RA %

结果分析指标	AllConV[14-16]	NiN[17]	VGG16[18-20]	结果分析指标	AllConV[14-16]	NiN[17]	VGG16[18-20]
遗传算法成功率	70.85	67.00	65.91	随机攻击成功率	49.70	41.72	15.57
GA 置信度	78.24	82.80	77.67	RA 置信度	87.73	75.83	59.90

如表 3 所示,通过 GA 与 LSA[18] 和 FGSM[6] 两种方法进行非目标攻击的结果比较,GA 修改 1 个像素点生成对抗样本对 Network in Network[17] 和 VGG16[18-20] 进行攻击的成功率能达到 65%以上,GA 修改 3 个像素点生成对抗样本对 Network in Network 进行攻击的成功率与 LSA[18] 修改 30 多个像素点进行对抗攻击的成功率仅相差 10%左右,GA 修改 5 个像素点生成对抗样本对 VGG16 进行攻击的成功率比 FGSM[6] 修改 1 024 个像素点进行对抗攻击的成功率高 6%左右.

4 总 结

深度神经网络容易受到对抗攻击,在对现有的生成对抗样本方法进行对比和分析后,本文提出了基于深

度神经网络的遗传算法对抗攻击，仅仅修改一个像素就能生成攻击效果良好的对抗样本，结合遗传算法的特性可以灵活设置修改像素点的数量，通过对不同网络的攻击结果、适应度值变化和不同攻击方法对比，展示了本文提出的方法在成功率和置信度方面都有较大提升.扰动少量的像素值还可以看作对不同的网络结构和图像使用低维切片进行切割，这是探索高维 DNN 输入空间特征的一种方式[22]，这种方法进一步可以扩展到其他领域，如自然语言处理，语音识别.此外，进化算法还为解决对抗机器学习相关漏洞提供了一些有前景的方法，通过进化算法[23]来学习网络的拓扑结构，使用相对较小的初始候选解决方案集进行 GA 迭代，通过更多的迭代、更高级的算法或更大的初始候选解集，扰动成功率应该进一步提高，易受攻击的图像可能有助于生成人工对抗图像，扩大训练模型的数据集，这将有助于进一步提升模型的鲁棒性.

表 3　GA，LSA[18]和 FGSM 方法结果比较

Tab. 3　The results comparing about GA，LSA[18] and FGSM[6]

提出方法	成功率/%	置信度/%	像素数量	分类网络	提出方法	成功率/%	置信度/%	像素数量	分类网络
GA	67.00	82.80	1	NiN	LSA[18]	97.89	72.00	33	NiN
GA	65.91	77.67	1	VGG	LSA[18]	97.89	77.00	30	VGG
GA	86.72	85.61	3	NiN	FGSM[6]	93.67	93.00	1 024	NiN
GA	93.7	99.70	5	VGG	FGSM[6]	90.93	90.00	1 024	VGG

附录见电子版(DOI:10.16366/j.cnki.1000-2367.2023.09.21.0003).

参　考　文　献

[1] TAIGMAN Y，YANG M，RANZATO M，et al.DeepFace：closing the gap to human-level performance in face verification[C]//2014 IEEE Conference on Computer Vision and Pattern Recognition.Columbus：IEEE，2014.

[2] MOOSAVI-DEZFOOLI S M，FAWZI A，FROSSARD P.DeepFool：a simple and accurate method to fool deep neural networks[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition(CVPR).Las Vegas：IEEE，2016.

[3] BARRENO M，NELSON B，SEARS R，et al.Can machine learning be secure?[C]//Proceedings of the 2006 ACM Symposium on Information，Computer and Communications Security.[s.l.]：ACM，2006.

[4] SZEGEDY C，ZAREMBA W，SUTSKEVER I，et al.Intriguing properties of neural networks[EB/OL].[2023-09-13].http://arxiv.org/abs/1312.6199v4.

[5] GROSSE K，PFAFF D，SMITH M T，et al.The Limitations of Model Uncertainty in Adversarial Settings[EB/OL].[2024-05-04].https://arxiv.org/abs/1812.02606.

[6] GOODFELLOW I J，SHLENS J，SZEGEDY C，et al.Explaining and harnessing adversarial examples[EB/OL].[2023-09-13].http://arxiv.org/abs/1412.6572v3.

[7] MOOSAVI-DEZFOOLI S M，FAWZI A，FAWZI O，et al.Robustness of classifiers to universal perturbations：a geometric perspective[EB/OL].[2023-08-17].http://arxiv.org/abs/1705.09554v2.

[8] ELSAYED G F，SHANKAR S，CHEUNG B，et al.Adversarial examples that fool both computer vision and time-limited humans[EB/OL].[2023-08-17].http://arxiv.org/abs/1802.08195v3.

[9] EYKHOLT，ROBERT G.Driving and complicating features of the electrokinetic treatment of contaminated soils[EB/OL].[2024-05-04].https://mr.mbd.baidu.com/r/1iWLAncvrlm?f=cp&u=f6760da97f71bee7.

[10] NARODYTSKA N，KASIVISWANATHAN S.Simple black-box adversarial attacks on deep neural networks[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops(CVPRW).Honolulu：IEEE，2017.

[11] 吴子斌，闫巧.基于动量的映射式梯度下降算法[J].计算机科学，2022，49(S1)：178-183.
WU Z B，YAN Q.Projected gradient descent algorithm with momentum[J].Computer Science，2022，49(S1)：178-183.

[12] ZHANG C N，BENZ P，LIN C G，et al.A survey on universal adversarial attack[EB/OL].[2023-08-17].http://arxiv.org/abs/2103.01498v2.

[13] CLARKE D K，DUARTE E A，MOYA A，et al.Genetic bottlenecks and population passages cause profound fitness differences in RNA viruses[J].Journal of Virology，1993，67(1)：222-228.

[14] SPRINGENBERG J T，DOSOVITSKIY A，BROX T，et al.Striving for simplicity：the all convolutional net[EB/OL].[2023-08-22].http://arxiv.org/abs/1412.6806v3.

[15] 李云波.基于全卷积神经网络的多级阈值图像分割算法[J].微型电脑应用，2023，39(6)：145-147.

LI Y B.Multistage threshold image segmentation algorithm based on full convolutional neural network[J].Microcomputer Applications,2023,39(6):145-147.

[16] 付鹏飞,许斌.全卷积神经网络仿真与迁移学习[J].软件,2019,40(5):216-221.
FU P F,XU B.All convolution neural network simulation and transfer learning[J].Software,2019,40(5):216-221.

[17] LIN M,CHEN Q,YAN S.Network In Network[EB/OL].[2024-05-04].https://arxiv.org/abs/1312.4400.

[18] SIMONYAN K,ZISSERMAN A.Very deep convolutional networks for large-scale image recognition[EB/OL].[2024-05-04].https://arxiv.org/abs/1409.1556.

[19] 童占北,钟建伟,李祯维,等.基于 VGG16 图像特征提取和 SVM 的电能质量扰动分类[J].电工电气,2023(7):7-13.
TONG Z B,ZHONG J W,LI Z W,et al.Power quality disturbance classification based on VGG16 image feature extraction and SVM[J].Electrotechnics Electric,2023(7):7-13.

[20] 方犇,童强.基于 VGG16 网络的参数优化方法研究[J].湖北师范大学学报(自然科学版),2023,43(1):44-50.
FANG B,TONG Q.Research on parameter optimization method based on VGG16 network[J].Journal of Hubei Normal University(Natural Science),2023,43(1):44-50.

[21] KRIZHEVSKY A,HINTON G.Learning multiple layers of features from tiny images[J].Handbook of Systemic Autoimmune Diseases,2009,1(4):60.

[22] HAMDIPOOR V,MOON J,KIM Y.Stability margin of undirected homogeneous relative sensing networks:a geometric perspective[J].Systems & Control Letters,2021,156:105027.

[23] Michalewicz Z,Schoenauer M.Evolutionary Algorithm[EB/OL].[2024-05-04].https://arxiv.org/abs/1805.11014.

Genetic algorithm against attack based on deep neural network

Fan Haiju, Ma Jincheng, Li Ming

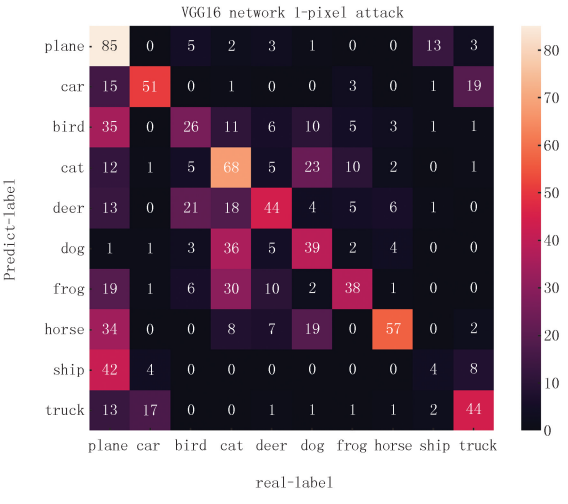
(College of Computer and Information Engineering; Henan Provincial Key Laboratory of Educational Artificial Intelligence and Personalized Learning, Henan Normal University, Xinxiang 453007, China)

Abstract: Deep Neural Network(DNN)can achieve good classification and recognition effect, but the recognition accuracy will be greatly reduced when adding small disturbance to the training image to counter the attack. This paper, by proposing a method a small number of pixels on the image are modified to generate adversarial samples after the optimal disturbance is obtained by genetic algorithm. Different convolutional neural networks are attacked as image classifiers, and parameters such as the number of images processed in each batch and the number of modified pixels are adjusted. The experimental results show that 67.92% of the natural images in CIFAR-10 data set can be disturbed to at least one target class, with an average confidence of 79.57%, and the attack effect will be further improved with the increase of modified pixels. In addition, compared with LSA and FGSM methods, the attack effect is significantly improved.

Keywords: convolutional neural network; genetic algorithm; adversarial attack; image classification; information security

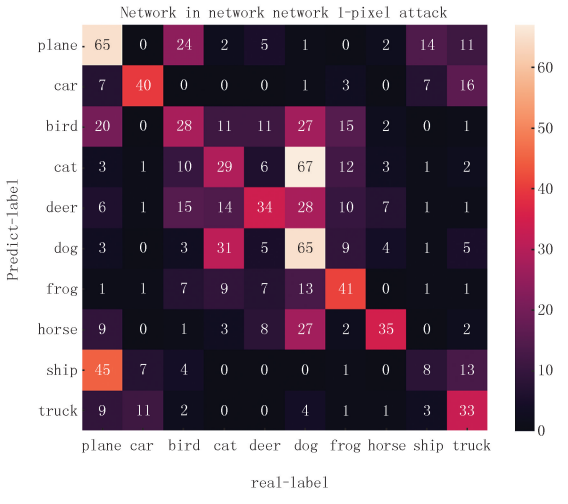
[责任编校 陈留院 杨浦]

附　录



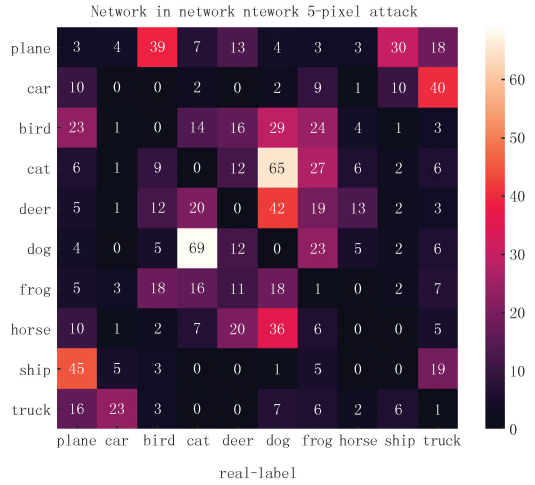
图S1　VGG16单像素攻击热图矩阵

Fig. S1　The heat map of VGG-single pixel attack



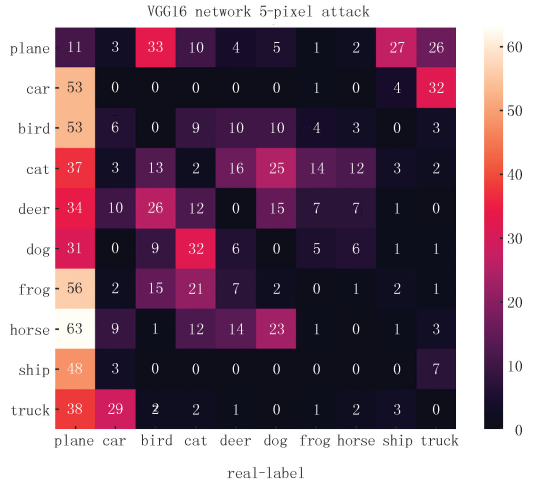
图S2　NiN单像素攻击热图矩阵

Fig. S2　The heat map of NiN-single pixel attack



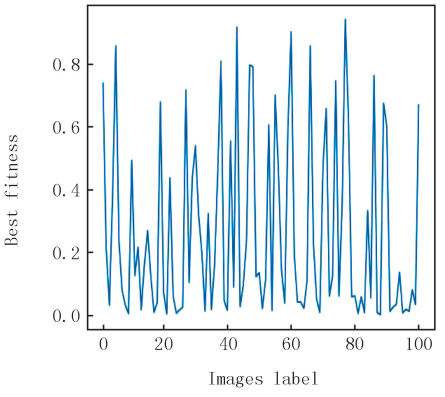
图S3　VGG16五像素攻击热图矩阵

Fig. S3　The heat map of VGG-five pixels attack



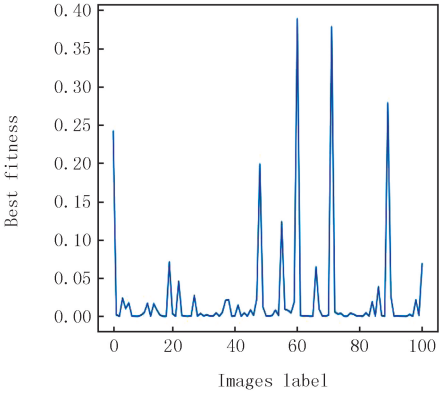
图S4　NiN五像素攻击热图矩阵

Fig. S4　The heat map of NiN-five pixels attack



图S5　单像素目标攻击适应度值曲线

Fig. S5　The fitness curve of single pixel attack



图S6　三像素目标攻击适应度值曲线

Fig. S6　The fitness curve of three pixels attack