石雪莹, 卢彦利, 杨伊雷,等. 基于数值微分的对抗样本生成算法 [J]. 智能计算机与应用,2025,15(11):9-14. DOI:10.20169/j. issn. 2095-2163. 24032302

基于数值微分的对抗样本生成算法

石雪莹[1], 卢彦利[1], 杨伊雷[1], 柳雪飞[1], 蒋正锋[1,2]
(1 广西民族师范学院 数学与计算机科学学院,广西 崇左 532200; 2 武汉大学 计算机学院, 武汉 430072)

摘 要: 针对深度神经网络黑盒攻击对抗中存在攻击效率低、扰动强度大以及难以生成高攻击性对抗样本等问题,本文提出了一种基于数值微分的对抗样本生成算法。该算法采用中心差分法计算梯度,并融合了高斯噪声初始化与数值微分梯度优化的复合扰动策略;通过梯度下降法迭代优化扰动,同时引入适应度函数以动态权衡攻击效果与扰动程度。手写数字识别实验结果表明,本文提出的算法能够有效生成对抗样本,通过适应度函数评估对抗样本的攻击效果与扰动程度,验证了算法的有效性。该算法为评估深度神经网络的鲁棒性提供有效的黑盒攻击工具,并为设计安全防御机制提供参考。

关键词: 深度神经网络; 对抗攻击; 对抗样本; 数值微分; 黑盒攻击
中图分类号: TP391 **文献标志码**: A **文章编号**: 2095-2163(2025)11-0009-06

Adversarial sample generation algorithm based on numerical differentiation

SHI Xueying[1], LU Yanli[1], YANG Yilei[1], LIU Xuefei[1], JIANG Zhengfeng[1,2]
(1 College of Mathematics and Computer Science, Guangxi Minzu Normal University, Chongzuo 532200, Guangxi, China; 2 School of Computer Science, Wuhan University, Wuhan 430072, China)

Abstract: To address the challenges of low attack efficiency, excessive perturbation magnitudes, and difficulties in generating highly aggressive adversarial examples in black-box attacks against deep neural networks, this paper develops an adversarial sample generation algorithm based on numerical differentiation. The proposed methodology employs the central difference method for gradient computation and integrates a composite perturbation strategy that combines Gaussian noise initialization with gradient optimization via numerical differentiation. Perturbations are iteratively refined through gradient descent optimization, while a fitness function dynamically balances attack effectiveness with perturbation intensity. Experimental validation on benchmark datasets for handwritten digit recognition demonstrates the algorithm's efficacy in generating adversarial samples. Quantitative evaluation through the fitness function, which assesses both attack success rates and perturbation levels, confirms the approach's effectiveness. This algorithm serves as a robust black-box attack tool for evaluating the vulnerability of deep neural networks and provides a theoretical foundation for designing secure defense mechanisms.

Key words: deep neural networks; adversarial attack; adversarial samples; numerical differentiation; black-box attack

0 引 言

随着计算机技术和深度学习的革新与普及,目标检测、计算机视觉、语音识别和自然语言理解等已深入人们的日常生活。然而,深度学习应用面临着安全性挑战。2015 年,Goodfellow 等[1]首次通过一个直观实验演示了"对抗攻击"的概念,该实验对一张清晰的大熊猫图像添加视觉上难以察觉的细微扰动,生成"对抗样本",该样本被以高置信度误分类为"长臂猿",而原始熊猫图像则被正确识别,揭示了深度学习模型面对精心设计的微小输入扰动时的脆弱性。深度神经网络在面对对抗性样本时表现脆弱,即便是微小的扰动也可能导致深度神经网络的判断失误。当在标准样本中植入精心设计的细微扰

基金项目: 国家级大学生创新创业项目(202210604038); 广西高校中青年教师科研基础能力提升项目(2024KY0787,2022KY0767);广西民族师范学院校级项目(2024YB121, 2024YB122, JGYB202458)。
作者简介: 石雪莹(2003—),女,本科生,主要研究方向:人工智能; 卢彦利(2001—),女,本科生,主要研究方向:人工智能; 杨伊雷(2003—),男,本科生,主要研究方向:机器学习; 柳雪飞(1984—),男,硕士,讲师,主要研究方向:大数据技术,智能优化算法。
通信作者: 蒋正锋(1979—),男,硕士,副教授,主要研究方向:最优化理论及其应用,机器学习,深度学习等。Email:jiangzhengfeng@ gxnun. edu. cn。
收稿日期: 2024-03-23

动时,深度神经网络往往以较高的置信度输出错误预测,这种经过特殊处理的对抗样本具有跨模型的泛化能力,能够同时使多个不同的深度神经网络产生误判。

根据攻击者对深度神经网络内部信息的掌握程度,攻击场景可划分为白盒攻击和黑盒攻击。在实际应用场景中,攻击者往往难以获取深度神经网络的完整技术细节,包括网络架构设计、权重参数配置及训练数据集等核心信息。这种信息不对称的条件下,黑盒攻击模式更贴合真实世界的对抗场景。现有研究已提出多种黑盒攻击范式,Su 等[2]提出一种只需对图片中少数的像素进行修改,便能导致模型出现分类错误的无目标黑盒攻击方法;Brendel 等[3]提出了一种基于边界的黑盒攻击方法,从生成较大的对抗扰动开始,并在保持对抗性的同时,根据一定的策略将该对抗样本向原始图像的方向移动,直至该对抗样本与原始图像的差异最小;Chen 等[4]提出了零阶优化(Zeroth Order Optimization, ZOO)攻击,仅利用输入和模型提供的相应置信度得分来估算模型的梯度,采用有限差分法,通过在添加微小扰动后评估图像坐标来估算每个坐标的梯度方向;黄立峰等[5]开发了一种基于进化计算和注意力机制的黑盒攻击策略,成功实现了高效的黑盒攻击;Bhagoji 等[6]提出了一种新型的黑盒攻击方法,依赖于梯度估计和目标模型的类别概率,并展现了针对不同攻击目标的通用性。

针对现有方法对梯度信息或概率输出的依赖性问题,本文提出一种基于数值微分的黑盒对抗样本生成算法。该算法通过迭代施加微小结构化扰动,结合损失函数约束机制,在无需掌握模型内部结构和梯度信息回传的条件下,实现深度神经网络响应的近似估计与对抗样本的生成,进一步探索黑盒攻击的可能性。

1 理论基础

1.1 数值微分

数值微分,通常指的是利用函数在特定离散点上的测量值(或观测值)来估算其近似导数的过程[7]。这一过程旨在解决一类典型的不适定问题,即由于离散数据的有限性和不连续性,微分运算的精确性可能会受到一定影响。当待求导的函数受到较小的扰动时,求导后产生的误差可能会变得非常大,这是一个典型的数值不稳定性问题[8]。为减少数值微分的误差,采用中心差分来实现数值差分,即估算的梯度 g_i:

$$g_i = (f(x + \varepsilon_i) - f(x - \varepsilon_i))/(2\varepsilon_i) \approx \frac{\partial f(x)}{\partial x_i} \tag{1}$$

其中,ε_i 是一个微小的值,$\frac{\partial f(x)}{\partial x_i}$ 表示关于 x_i 的偏导数。

梯度法是机器学习中最优化问题求解的常用方法,通过迭代更新模型参数以最小化损失函数,从而实现模型的优化和学习[9]。梯度法如下式所示:

$$x_{i\_new} = x_{i\_old} - lr \times \frac{\partial f(x)}{\partial x_i} \tag{2}$$

其中,x_{i_old} 表示原来的值;x_{i_new} 为更新后的值;lr 是学习率。

1.2 深度学习模型

在构建神经网络时,输入层神经元的数量直接对应于数据集中样本的维度。对于 sklearn 库中 datasets 模块的手写数字数据集中 8×8 二维图像样本,当转化为一维向量时,包含 64 个灰度值,因此在构建针对该数据集的神经网络模型时,设定输入层神经元的数量为 64;第 2 层(即隐藏层 1)包含 128 个神经元,这些神经元与输入层的所有元素(即输入数据的特征)实现全连接,通过一组特定的权重和偏置项,每个神经元对输入数据进行线性组合,通过一个非线性激活函数的处理,引入非线性特性,从而生成该神经元的输出;第 3 层(即隐藏层 2)则包含 64 个神经元,其与第 2 层(隐藏层 1)的所有神经元实现全连接,每个神经元通过权重和偏置项同样进行线性组合,并通过非线性激活函数处理,产生该层的输出。这一设计通过减少神经元的数量(从 128 减少到 64),以降低过拟合的风险,并进一步提炼和整合从前面层次学习到的特征。MNIST 数据集包含的是手写体数字 0~9,共计 10 个分类,因此第 4 层(即输出层)的神经元数量设定为 10,这些神经元与第 3 层(隐藏层 2)的所有神经元实现全连接。在输出层,采用 Softmax 激活函数,将原始输出转化为概率分布,从而反映输入数据属于每个类别的可能性。

深度神经网络的结构设计,包括网络深度、各层神经元的数量以及激活函数的选择,均根据具体任务需求和实验结果调整[10-11]。不同的设计可能会导致神经网络在性能和效果上有所差异。

2 基于数值微分的对抗样本生成算法

在机器学习模型的安全性评估中,对抗样本的

生成起到关键作用。无目标对抗样本生成算法的核心在于寻找一种能够导致深度神经网络对处理后的样本进行任意错误分类的扰动，而无需预设特定的目标类别。这种算法主要关注评估深度神经网络的整体脆弱性，而非针对某一特定类别的攻击。通过逐步引入细微的结构化扰动，并结合损失函数限制机制，可以在不获取模型内部结构与梯度信息的情况下，近似计算出所需的梯度，进而指导扰动的生成。本文提出的基于数值微分的无目标对抗样本生成算法伪代码描述如下：

算法 1　基于数值微分的对抗样本生成算法

输入　损失函数 fitness(X)，训练好的模型 mymodel，输入原始样本 X 及其标签 Y，原始样本数量 NumSample

输出　对抗样本 X_α

Step 1　设置算法参数（候选对抗样本 NumAdvSample，扰动大小 ε，迭代次数 T）

Step 2　while(i←1<NUM)

Step 2.1　for $n \leftarrow 1$ to NumAdvSample do

Step 2.1.1　产生与原始样本 i 形状相同的高斯噪声，生成含有高斯噪声扰动的初始化样本 ad_i_n

Step 2.1.2　for $t \leftarrow 1$ to T do

按公式(1)计算样本 ad_i_n 的梯度信息

按公式(2)更新样本 ad_i_n

end for

end for

Step 2.2　按公式(6)计算 NumAdvSample 个对抗样本的的损失值

Step 2.3　根据损失值对 NumAdvSample 个对抗样本进行排序，选择损失值最小的对样样本为样本 i 生成的对抗样本

end while

Step 3　输出 NUM 个对抗样本 X_α

3　实验与分析

3.1　实验设置与数据集

实验在 Windows 10 操作系统下进行，采用 Jupyter Notebook 作为交互式编程平台，实验使用 Python 语言开发。实验环境的硬件配置包括 Intel (R) Core (TM) i7 - 6700HQ CPU，运行频率为 2.60 ~ 2.59 GHz，并配备 8 GB 内存，确保实验的稳定性和性能。

实验采用 sklearn 库中 datasets 模块的手写数字数据集，该数据集包含 1 797 张 8×8 像素的灰度数字图像，划分为 1 707 个训练样本和 90 个测试样本，每个样本表示为 64 维特征向量。

3.2　评价指标

在评估对抗样本的攻击效果时，本实验采用攻击性和扰动程度作为评价指标。攻击性指标用于衡量生成的对抗样本对深度神经网络的攻击效果，具体可分为目标对抗性攻击和无目标对抗性攻击两种类型[12]，公式如下：

$$\mathrm{ADV}(\mathrm{adv\_sample}) = J(\mathrm{mymodel}(\mathrm{adv\_sample}), y\_\mathrm{target}) \tag{3}$$

$$\mathrm{ADV}(\mathrm{adv\_sample}) = (\mathrm{score\_true})/(\mathrm{rank}(\mathrm{adv\_sample}_i)) \tag{4}$$

其中，ADV(adv_sample) 表示对抗样本 adv_sample 的攻击性；交叉熵函数 $J(\cdot)$ 用于衡量预测值与真实标签之间的差异；score_true 是真实标签的置信度；rank(adv_sample) 表示真实标签的置信度得分排名。

扰动程度是量化原始样本与其对应对抗样本之间差异的度量，通常使用 L_p 范数(L_0、L_2、L_∞)表示，其值越小意味着扰动越隐蔽[13]。$L_0(p=0)$ 范数统计被修改特征的数量(如图像中改变的像素数量)，反映扰动的稀疏性；$L_2(p=2)$ 范数衡量对抗样本与原始样本之间的差异，即扰动量；$L_\infty(p=\infty)$ 范数则标识所有特征中最大的绝对值变化，体现了单维度上的最大扰动量。L_p 范数的数学定义如下式：

$$L_p = \|\mathrm{adv\_sample}\|_p = \left(\sum_{i=1}^{n} |\mathrm{adv\_sample}_i|^p\right)^{\frac{1}{p}} \tag{5}$$

其中，$\|\mathrm{adv\_sample}\|_p$ 表示对抗样本 adv_sample 的 L_p 范数，$\mathrm{adv\_sample}_i$ 是对抗样本 adv_sample 在第 i 个维度上的分量值。

3.3　适应度函数设计

在对抗攻击场景中，有效的对抗样本需同时满足两个关键条件：高攻击性即成功误导深度神经网络产生错误输出和低扰动程度即所添加的扰动微小到难以被人类察觉[14]。这两个条件构成了一个多目标优化问题。因此，本文设计的损失函数综合考虑了攻击性和扰动程度这两个因素，计算公式为：

$$\mathrm{loss\_Funtion}(\mathrm{adv\_sample}) = \mathrm{ADV}(\mathrm{adv\_sample}) + \mathrm{alpha} \times \|\mathrm{adv\_sample}\|_2 \tag{6}$$

$$\|\mathrm{adv\_sample}\|_2 = \|\mathrm{adv\_sample} - \mathrm{ori\_sample}\|_2 \tag{7}$$

其中，loss_Funtion(adv_sample) 表示对抗样本

adv_sample 的适应度函数；ADV(adv_sample)表示对抗样本 adv_sample 的攻击性；$\|\text{adv\_sample}\|_2$ 表示对抗样本 adv_sample 的 L_2 范数；alpha 表示多目标优化中约束目标函数程度的惩罚系数，本文取值 alpha=0.01；$\|\text{adv\_sample}-\text{ori\_sample}\|_2$ 表示对抗样本 adv_sample 与原始样本之间的差异程度即扰动程度。

3.4 实验结果与分析

对抗样本是通过向原始样本引入特定扰动生成的。本文采用 4 类对抗样本，分别评估其对深度神经网络的攻击效果：

(1)第一类样本：未引入任何扰动的原始样本，作为基准对照组；

(2)第二类样本：通过叠加随机高斯噪声生成的扰动样本；

(3)第三类样本：利用梯度计算生成定向扰动的数值微分优化样本；

(4)第四类样本：融合高斯噪声与数值微分策略的复合扰动样本。

首先，使用包含 1 707 个手写数字样本的训练集对深度神经网络进行训练，该模型在测试集上实现了 94.4% 的分类准确率；其次，通过对测试集原始样本施加扰动，生成了 4 类对抗样本，包括原始样本、高斯噪声扰动样本、数值微分优化样本、高斯噪声和数值微分复合扰动样本；最后，使用这些对抗样本对训练完成的深度神经网络进行攻击测试。为了全面评估攻击效果，设置对抗样本的数量为 10，并控制其他参数不变，分别进行迭代次数为 2、3、4、5、10 和 15 的攻击对抗实验，实验结果见表 1。

表 1 不同迭代次数攻击成功率

Table 1 Attack success rate of different iterations

对抗样本	迭代次数					
	2	3	4	5	10	15
第一类样本	0.056	0.056	0.056	0.056	0.056	0.056
第二类样本	0.067	0.133	0.122	0.067	0.067	0.122
第三类样本	0.589	0.589	0.580	0.578	0.600	0.578
第四类样本	0.960	0.944	0.900	0.967	0.911	0.944

由表 1 可见，在攻击效果方面，采用复合扰动策略的第四类样本攻击效果最好，其在第 5 次迭代时达到了 96.7% 的攻击成功率，相比使用数值微分优化方法的第三类样本，平均优势为 37.8%；而第二类样本与第一类样本的攻击成功率差异维持在 8.7% 以内，说明随机噪声对模型抗攻击能力的削弱作用较为有限；第四类样本在第 5 次迭代时达到峰值 96.7%，但在第 15 次迭代时降至 94.4%，表明过度迭代可能导致扰动效果过度优化；第三类样本攻击成功率的波动幅度维持在±1.1%之间，说明梯度攻击具有较好的稳定性；第二类样本在第 3 次迭代时出现异常峰值 13.3%，凸显了随机扰动的不确定性。

根据表 2 所示的实验结果，当迭代次数为 5 时，第四类样本呈现出最佳的攻击效果，其成功率达到 96.7%。基于此，本文进一步探讨了对抗样本数量对攻击效果的影响。在保持迭代次数为 5 的条件下，本文设计 4 类不同数量的对抗样本实验，样本数量从 1 到 15，对深度神经网络进行攻击，4 类对抗样本攻击成功率见表 2。

表 2 不同对抗样本数量的攻击成功率

Table 2 Attack success rate under varying numbers of adversarial examples

对抗样本数量	对抗样本			
	第一类样本	第二类样本	第三类样本	第四类样本
1	0.056	0.111	0.544	0.633
2	0.056	0.133	0.556	0.744
3	0.056	0.156	0.556	0.800
4	0.056	0.189	0.567	0.833
5	0.056	0.167	0.578	0.856
6	0.056	0.189	0.578	0.911
7	0.056	0.222	0.589	0.900
8	0.056	0.244	0.589	0.911
9	0.056	0.278	0.589	0.944
10	0.056	0.289	0.578	0.900
11	0.056	0.233	0.578	0.944
12	0.056	0.256	0.589	0.922
13	0.056	0.267	0.589	0.944
14	0.056	0.333	0.589	0.944
15	0.056	0.256	0.589	0.956

表 2 实验结果表明，在相同对抗样本数量下，4 类样本的攻击成功率存在明显差异，例如：当对抗样本数量为 10 时，4 类对抗样本对深度神经网络攻击的成功率分别为 0.056、0.289、0.578 和 0.900。整体而言，增加对抗样本数量有助于提升攻击成功率，这一趋势具有普遍性，其中第二类和第四类样本尤为明显。值得注意的是，仅需少量对抗样本即可实现较高攻击成功率，突显深度神经网络在面对第四类对抗样本时存在的严重安全隐患。

表 2 的实验结果表明，增加对抗样本数量有助于提升攻击成功率。因此，本文在设定对抗样本数量为 15 个、攻击迭代次数为 5 次的条件下，对第二

类、第三类及第四类对抗样本进行了多维性能对比实验。评估指标采用攻击成功率、扰动程度和适应度函数值，其中适应度函数值越低，表明攻击效率越高，该指标用于综合评估攻击效果，实验结果见表 3。

表 3　对抗样本多维度性能对比实验结果

Table 3　Experimental results of multidimensional performance comparison for adversarial examples

对抗样本	攻击成功率	扰动程度	适应度函数值
第二类样本	0.244	0.080	0.162
第三类样本	0.578	0.057	0.152
第四类样本	0.922	0.112	0.011

根据表 3 数据，第四类样本的攻击成功率最高，但其扰动程度也最大；第三类样本的扰动程度最小，攻击成功率中等；第二类样本的表现最弱。第四类样本的综合效率最佳，第三类次之，第二类最差，表明攻击效率与扰动程度之间存在权衡：第四类样本以高扰动换取高效攻击，第三类样本在隐蔽性与效果之间取得平衡，第二类样本则需要进一步优化。在实际应用中，第四类样本适合强攻击场景，而第三类样本则适合隐蔽性需求较高的场景。

为可视化原始样本与对抗样本之间扰动的动态变化特征，选择生成第四类对抗样本，从测试集中随机选取了 3 个原始样本作为初始样本，分别设置了 2、3、4、5、10 和 15 次迭代，生成了一系列对应的对抗样本，从原始样本到最终对抗样本的动态演变过程如图 1 所示。

图 1 直观地呈现了扰动的累积过程，为理解攻击机理及深度神经网络的脆弱性提供了依据。随着迭代次数的增加，施加于原始样本的微小扰动逐步累积，使对抗样本与原始样本的差异逐渐显著，攻击效果也随之增强，这一动态过程揭示了对抗样本的生成是一个逐步优化的过程，有助于深入理解攻击方法的内在机理。

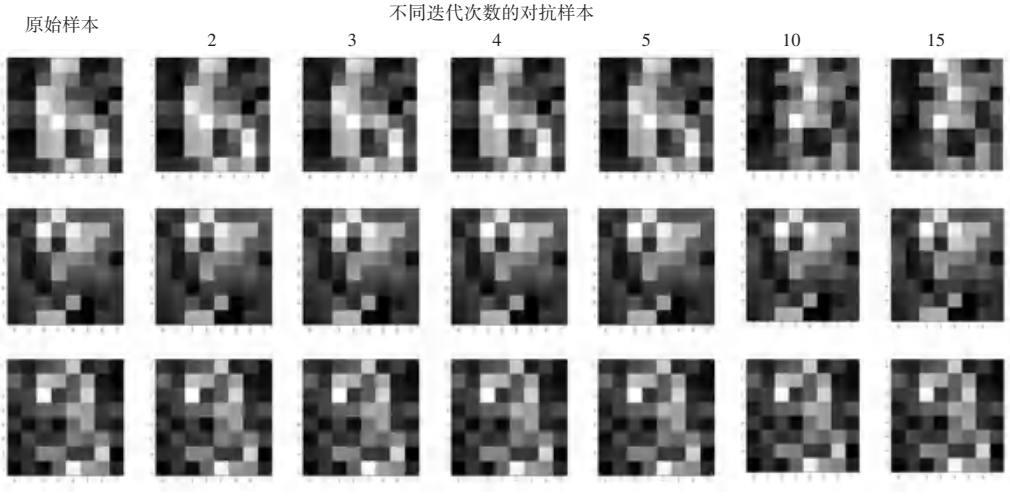

图 1　原始样本到最终对抗样本的动态演变过程

Fig. 1　Dynamic evolution process from original samples to final adversarial samples

4　结束语

针对深度神经网络在黑盒攻击场景下面临的安全性挑战，本文提出了一种基于数值微分的对抗样本生成算法。该算法融合了高斯噪声初始化与数值微分梯度优化的复合扰动策略，并结合梯度下降法进行迭代优化扰动，能够在无需掌握模型内部结构或依赖梯度回传的条件下，生成具有高攻击性的对抗样本；通过引入适应度函数，算法动态平衡了攻击效果与扰动程度，使得生成的对抗样本既具备攻击能力，又最大程度维持了视觉隐蔽性。在手写数字识别实验中，对比分析 4 种不同扰动策略生成的对抗样本，发现融合高斯噪声与数值微分的复合策略攻击效果最佳，优于仅采用随机高斯噪声或单一数值微分优化的方法。此外，实验探讨了对抗样本数量与迭代次数对攻击效果的影响，揭示了对抗样本在生成过程中的动态演变规律，结果表明：随着对抗样本数量的增加和迭代次数的优化，攻击成功率得到提升，为理解对抗攻击机理及深度神经网络的脆弱性提供了重要依据。

综上所述，本文所提出的基于数值微分的对抗样本生成算法为评估深度神经网络鲁棒性提供了一种有效的黑盒攻击工具，并为设计安全防御机制提供了理论参考与实践指导。未来研究将进一步探索该算法在不同数据集和模型架构下的普适性及其与其他防御策略相结合的可能性，旨在推动深度神经

网络安全领域的研究与发展。

参考文献

[1] GOODFELLOW I J, SHLENS J, SZEGEDY C. Explaining and harnessing adversarial examples[J]. arXiv preprint arXiv, 1412. 6572, 2014.

[2] SU J W, VARGAS D V, SAKURAI K. One pixel attack for fooling deep neural networks [J]. IEEE Transactions on Evolutionary Computation, 2019, 23(5): 828-841.

[3] BRENDEL W, RAUBER J, BETHGE M. Decision-based adversarial attacks: Reliable attacks against black - box machine learning models[J]. arXiv preprint arXiv,1712. 04248, 2017.

[4] 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 ACM Workshop on Artificial Intelligence and Security. New York: ACM, 2017: 15-26.

[5] 黄立峰,庄文梓,廖泳贤,等. 一种基于进化策略和注意力机制的黑盒对抗攻击算法[J]. 软件学报, 2021, 32(11): 3512-3529.

[6] BHAGOJI A N, HE W, LI B, et al. Exploring the space of black-box attacks on deep neural networks[J]. arXiv preprint arXiv,1712. 09491, 2017.

[7] 邱淑芳,叶智群,胡彬. 近似函数高阶导数的高精度积分逼近方法[J]. 数学的实践与认识,2021,51(6):217-224.

[8] 杨帆. 三类不适定问题的正则化方法研究[D]. 兰州:兰州大学,2014.

[9] 朱志广,王永. 基于高斯噪声扰动的随机梯度法的设计与应用[J]. 电子技术,2021,50(8):4-7.

[10] 蒋正锋,廖群丽. 基于多参数融合优化的深度神经网络设计研究[J]. 现代计算机,2021,27(31):13-24.

[11] 杨益暄,田益民,崔圆斌,等. 基于深度学习方法的手写文本行提取综述[J]. 智能计算机与应用,2020,10(11):154-157.

[12] 陈晋音,陈治清,郑海斌,等. 基于 PSO 的路牌识别模型黑盒对抗攻击方法[J]. 软件学报,2020,31(9):2785-2801.

[13] 黄路路,唐舒宇,张伟,等. 基于 Lp 范数的非负矩阵分解并行优化算法[J]. 计算机科学,2024,51(2):100-106.

[14] 刘梦庭,凌捷. 优化梯度增强黑盒对抗攻击算法[J]. 计算机工程与应用,2023,59(18):260-267.