• 计算机工程与应用 •

基于决策边界搜索的对抗样本生成算法

刘欣刚*，江浩杨，苏 鑫，冯 晶

(电子科技大学信息与通信工程学院 成都 611731)

【摘要】神经网络模型已被广泛运用于人工智能领域，并取得了成功，然而当前神经网络面临着对抗样本攻击的困扰。对抗样本是一种人为构造的虚假数据，可使得神经网络输出错误的结果。故提出了一种基于神经网络决策边界搜索的对抗样本生成算法。首先，在两个真实样本之间使用二分搜索来找到一个初始攻击点。然后，计算神经网络在决策边界面上的法线向量，以找到神经网络最敏感的方向。最后，使用方向信息迭代找到更接近原始数据点的对抗样本，直到对抗样本收敛。在公开的数据集上，使用该算法进行对抗样本攻击实验，实验结果表明该算法能够生成对抗扰动更小的对抗样本，并且可以与其他攻击算法结合，达到较好的攻击效果。

关 键 词 对抗攻击； 对抗样本； 神经网络； 优化

中图分类号 TP391.4 文献标志码 A doi:10.12178/1001-0548.2021396

Adversarial ExamplesGeneration Algorithm Based on Decision Boundary Search

LIU Xingang*, JIANG Haoyang, SU Xin, and FENG Jing

(School of Information and Communication Engineering, University of Electronic Science and Technology of China Chengdu 611731)

Abstract The neural network model has been widely used in the fields of artificial intelligence, and has achieved great success. However, the current neural network is facing the problem of adversarial examples attack, which is artificially constructed fake data that can cause a neural network to output incorrect results. This paper proposes an adversarial examples generation algorithm based on searching the decision boundary of neural network. Firstly, weusebinary search between two real samples to find aninitialattacking point. And then,we calculate the normal vector of the neural network on the decision boundary surface, in order to find the most sensitive direction of the neural network. Finally, we usethe direction information to iteratively find the adversarialexample closer to the original data point until the adversarial example converges. By applying the proposed algorithm on the public data sets, the experimental results show that the algorithm can generate adversarial examples with smaller adversarial perturbations, and it can be combined with other attack algorithms to achieve a better attack result.

Key words adversarial attack; adversarial examples; neural networks; optimization

深度神经网络模型被广泛应用于各种机器学习领域，包括图像识别[1-2]、图像与视频目标检测领域[3-4]、音频数据处理[5]和自然语言处理领域[6-7]等。深度神经网络模型在许多任务上都取得了成功。

然而研究发现神经网络模型容易受到一种虚假样本的攻击[8]。这种样本通常是在真实数据样本上进行一定程度的轻微修改而生成的。当这种虚假样本输入到神经网络中，神经网络会产生与原始真实数据完全不同的输出。这样的虚假样本通常被称为对抗样本。

对抗样本被发现存在于许多领域。在图像处理领域，被轻微修改的图像可以使得神经网络对图像给出错误的分类[9-11]，也可以使得语义分割模型完全无法识别图像上的人物[12]；在自然语言处理领域，Seq2Sick 攻击可以生成对抗性的文本，使得基于序列的神经网络翻译模型无法正确理解文本的语义[13-15]；在音频数据领域，文献 [16] 提出了一种算法，将一个噪音嵌入到正常音频中，可以使得音频

收稿日期：2021－12－23；修回日期：2022－04－21
基金项目：国家自然科学基金 (61872404)
作者简介：刘欣刚 (1978－)，男，博士，教授，主要从事视频编码、图像视频处理、人工智能等方面的研究.
*通信作者：刘欣刚，E-mail：hanksliu@uestc.edu.cn

语义识别网络输出想要的任意结果。

对抗样本的生成技术也被广泛研究。文献 [17] 指出对抗样本的存在与神经网络的高度线性性质有关，并提出了一种快速梯度下降法 (fast gradient sign attack, FGSM) 来生成对抗样本，这种方法针对非鲁棒性模型非常有效，并且对抗样本的生成速度也非常快。文献 [10] 提出了一种有效生成最小化 L_2 度量距离的对抗样本技术，其所生成对抗样本的扰动相比 FGSM 更加隐蔽。文献 [9] 通过求解优化问题的方式找到对抗样本，该方法可以计算任意 p 范数的最小化 L_p 的对抗样本。文献 [18-19] 提出了寻找稀疏形式对抗样本的攻击，可以只改变少量的像素点即可达成攻击。文献 [20] 利用对抗生成网络框架，提出使用神经网络生成对抗样本，这种方法可以针对目标模型快速生成对抗样本。然而，这些方法大都使用类似梯度函数方向更新的方法，计算真实样本点附近的梯度，寻找可以使损失函数增大的对抗样本，而没有考虑沿着决策函数的边界进行搜索以寻找扰动最小的样本。

本文提出了一种基于神经网络决策边界搜索的对抗样本生成算法。该算法首先使用线性搜索或二分搜索在数据空间中找到一个处于决策边界的数据点，并计算该点相对于决策函数的法向量方向，基于决策平面局部平滑的假设，利用法向量的正交空间，寻找一个更接近真实样本的数据点，通过多步迭代的方式最终找到最优对抗样本。本文将该过程进行数学建模，然后将该数学问题转化为一个标准的优化问题。为了求解该优化问题，首先推导出当神经网络决策函数为仿射函数时的解析解，然后给出在更一般情况下的迭代式求解算法。

1 对抗样本攻击建模

图 1 是典型的白盒攻击模式，一个训练良好的分类器 $F(\cdot)$ 对正常样本 $\boldsymbol{x}$ 进行预测，得到其正确的标签 y。攻击方通过观测分类器的预测过程，得到输入 $\boldsymbol{x}$ 对于分类器的梯度信息及最终的分类置信度等信息，根据这些信息生成对抗样本 $\boldsymbol{x}'$，如果分类器 $F(\cdot)$ 对输入 $\boldsymbol{x}'$ 做出错误的预测 y'，则攻击成功。

该过程可以用数学语言进行如下描述。

对给定的分类器 $F(\cdot)$ 和输入 $\boldsymbol{x}$，对应于该输入 $\boldsymbol{x}$ 的最优对抗样本 $\boldsymbol{x}'$ 可以通过求解以下优化问题获得：

$$\begin{aligned} &\min\|\boldsymbol{r}\|_p \quad \text{s.t.} \quad F(\boldsymbol{x}) \neq F(\boldsymbol{x}+\boldsymbol{r}) \\ &F(\boldsymbol{x}) = \underset{k}{\operatorname{argmax}} f_k(\boldsymbol{x}) \quad k = 1,2,\cdots,K \end{aligned} \tag{1}$$

式中，$f_k(\boldsymbol{x})$ 是网络对应第 k 类的输出；$\|\boldsymbol{r}\|_p$ 表示 L_p 距离，定义为：

$$\|\boldsymbol{v}\|_p = \left(\sum_{i=1}^{n} |v_i|^p\right)^{\frac{1}{p}} \tag{2}$$

在对抗样本攻击领域，通常研究生成最小化 L_2 与 L_∞ 的对抗样本。

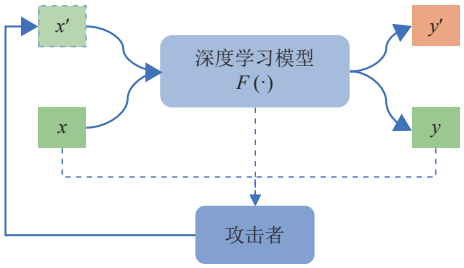


图 1 典型白盒攻击流程

由于原问题不属于标准的优化问题，需要将该问题转化为标准形式才能使用优化算法进行求解。

原问题可以转化为另一个等价的形式，即：

$$\begin{aligned} &\min\|\boldsymbol{r}\|_p \quad \text{s.t.} \quad L(\boldsymbol{x}+\boldsymbol{r}) < 0 \\ &L(\boldsymbol{x}) = f_i(\boldsymbol{x}) - f_j(\boldsymbol{x}) \\ &i = F(\boldsymbol{x}) \quad j \neq i \end{aligned} \tag{3}$$

式中，i 是输入 $\boldsymbol{x}$ 的原始标签；j 是任意其他标签。简化问题，令 j 为输入 $\boldsymbol{x}$ 最有可能出错的标签，即：

$$j = \underset{k}{\operatorname{argmax}} f_k(\boldsymbol{x}) \quad k = 1,2,\cdots,K \quad k \neq i \tag{4}$$

假设 $\lim\limits_{x \to x_0} L(\boldsymbol{x}) = L(\boldsymbol{x}_0)$，即 $L(\cdot)$ 是连续函数，且 $L(\boldsymbol{x}) > 0$，能够得出这个优化问题的解 $\boldsymbol{r}^{\text{opt}}$ 在函数 $L(\cdot)$ 的边界，即 $L(\boldsymbol{x}+\boldsymbol{r}^{\text{opt}}) = -\varepsilon$，$\varepsilon$ 为极小正数。因此，将原来的优化问题转化为如下标准形式：

$$\begin{aligned} &\min\|\boldsymbol{r}\|_p \\ &\text{s.t.} \quad L(\boldsymbol{x}+\boldsymbol{r}) = f_i(\boldsymbol{x}+\boldsymbol{r}) - f_j(\boldsymbol{x}+\boldsymbol{r}) - \varepsilon = 0 \\ &\quad\; i = F(\boldsymbol{x}) \\ &j = \underset{k}{\operatorname{argmax}} f_k(\boldsymbol{x}) \quad k = 1,2,\cdots,K \quad k \neq i \end{aligned} \tag{5}$$

虽然将原优化问题转化成为了标准形式，但因为函数 $L(\cdot)$ 是一个高度非凸函数，无法使用如牛顿法及拉格朗日乘子法等对该问题求解。

2 决策边界搜索攻击算法

2.1 随机初始点搜索

$L(\boldsymbol{x}') > 0$ 表明输入 $\boldsymbol{x}'$ 与 $\boldsymbol{x}$ 是相同的类，$L(\boldsymbol{x}') \leqslant 0$ 表明输入 $\boldsymbol{x}'$ 与 $\boldsymbol{x}$ 分属不同类，对抗样本只可能存在于决策边界，且满足 $L(\boldsymbol{x}') = 0$。

通常而言，直接得到一个输入 $\boldsymbol{x}_0$ 满足 $L(\boldsymbol{x}_0) = 0$ 是困难的，但是利用函数 $L(\cdot)$ 的连续性质，可以有

效找到该初始点。

假设输入$\boldsymbol{x}$为正常样本，且满足$L(\boldsymbol{x})>0$。随机生成输入数据$\bar{\boldsymbol{x}}$，直至满足$L(\bar{\boldsymbol{x}})<0$。因为$L(\cdot)$为连续函数，则必定存在初始点$\boldsymbol{x}_0=\theta\bar{\boldsymbol{x}}+(1-\theta)\boldsymbol{x}$，满足$L(\boldsymbol{x}_0)=0$，其中，$0<\theta<1$。

实验发现，以固定概率分布随机生成的数据$\bar{\boldsymbol{x}}$，通常会以高概率被判别为某一类。因此，更有效的方法是直接在数据集中挑选一个标签不同的样本$\bar{\boldsymbol{x}}$。θ则可以使用线性搜索或者二分搜索来得到。使用二分搜索的初始点搜索由算法 1 给出。

算法 1 基于二分搜索法的初始攻击点查找算法

函数输入：原样本$\boldsymbol{x}$，决策函数$L(\cdot)$

从数据集中找到与$\boldsymbol{x}$不同类的样本$\bar{\boldsymbol{x}}$

$k\leftarrow 1$

$\theta\leftarrow 1$

while $k>0.001$

$k\leftarrow\dfrac{k}{2}$

if $L(\theta\bar{\boldsymbol{x}}+(1-\theta)\boldsymbol{x})<0$

$\theta\leftarrow\theta-k$

else

$\theta\leftarrow\theta+k$

return θ

搜索到的初始点$\boldsymbol{x}_0$通常与样本$\boldsymbol{x}$相差较大，因此，需要通过迭代的方式找到新的对抗样本$\boldsymbol{x}_i$，使得$\boldsymbol{x}_i$与$\boldsymbol{x}$越来越接近，即使得其对抗扰动$\boldsymbol{r}_i=\boldsymbol{x}_i-\boldsymbol{x}$的范数减小。

2.2　迭代攻击算法

1) 线性决策器。给定一个线性函数$L(\boldsymbol{x})=\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}+b$，和一个已知的起始扰动$\boldsymbol{r}_0$，满足$L(\boldsymbol{x}+\boldsymbol{r}_0)=0$。$\boldsymbol{r}_0$可以分解为两个部分，即$\boldsymbol{r}_0=\boldsymbol{r}^{\mathrm{opt}}+\boldsymbol{r}^{\mathrm{step}}$，其中$\boldsymbol{r}^{\mathrm{opt}}$存在闭式解，且$\boldsymbol{r}^{\mathrm{opt}}$为全局最优解：

$$\boldsymbol{r}^{\mathrm{opt}}=\frac{\boldsymbol{r}_0^{\mathrm{T}}\boldsymbol{w}}{\|\boldsymbol{w}\|^2}\boldsymbol{w} \tag{6}$$

$$w=\frac{\partial L(\boldsymbol{x}+\boldsymbol{r}_0)}{\partial(\boldsymbol{x}+\boldsymbol{r}_0)} \tag{7}$$

式中，$\boldsymbol{w}$为$L(\cdot)$在$\boldsymbol{r}_0$处的梯度，$\boldsymbol{w}$可以使用现有的神经网络学习框架计算得出。

如图 2 所示，$\boldsymbol{r}^{\mathrm{opt}}$是$\boldsymbol{r}_0$在梯度方向$\nabla L(\boldsymbol{x}+\boldsymbol{r}_0)$的投影，$\boldsymbol{r}^{\mathrm{step}}$是$\boldsymbol{r}_0$在决策方向上的投影，注意到$\boldsymbol{r}^{\mathrm{step}}$和$\boldsymbol{r}^{\mathrm{opt}}$是相互正交的，所以对于任意$\boldsymbol{r}_1=\boldsymbol{r}^{\mathrm{opt}}+\alpha\boldsymbol{r}^{\mathrm{step}}$，$0\leqslant\alpha<1$，可得出：

$$\begin{aligned}\|\boldsymbol{r}_1\|_2&=\|\boldsymbol{r}^{\mathrm{opt}}+\alpha\boldsymbol{r}^{\mathrm{step}}\|_2\\&<\|\boldsymbol{r}^{\mathrm{opt}}+\boldsymbol{r}^{\mathrm{step}}\|_2=\|\boldsymbol{r}_0\|_2\end{aligned} \tag{8}$$

$$\begin{aligned}L(\boldsymbol{x}+\boldsymbol{r}_1)&=\boldsymbol{w}^{\mathrm{T}}(\boldsymbol{x}+\boldsymbol{r}_1)+b=\\&\boldsymbol{w}^{\mathrm{T}}\left(\boldsymbol{x}+\boldsymbol{r}^{\mathrm{opt}}+\alpha\boldsymbol{r}^{\mathrm{step}}\right)+b=\\&\boldsymbol{w}^{\mathrm{T}}\left(\boldsymbol{x}+\boldsymbol{r}_0+(\alpha-1)\boldsymbol{r}^{\mathrm{step}}\right)+b=\\&\boldsymbol{w}^{\mathrm{T}}(\boldsymbol{x}+\boldsymbol{r}_0)+(\alpha-1)\boldsymbol{w}^{\mathrm{T}}\boldsymbol{r}^{\mathrm{step}}+b=\\&L(\boldsymbol{x}+\boldsymbol{r}_0)+0=0\end{aligned} \tag{9}$$

以上推导证明了$(\boldsymbol{x}+\boldsymbol{r}_1)$同样是一个处于分类边界的数据点，同时$\boldsymbol{r}_1$相比于$\boldsymbol{r}_0$在二范数意义下更小。当$\alpha=0$时，$\boldsymbol{r}_1$退化为最优解$\boldsymbol{r}^{\mathrm{opt}}$。

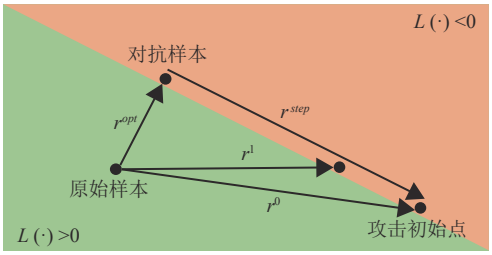


图 2　线性决策器

2) 非线性决策器。对于大多数的深度神经网络，$L(\boldsymbol{x})$是高度非凸函数，但可以从线性情况的推导中得出。每次迭代得到的扰动$\boldsymbol{r}_{i+1}$会比上一个扰动$\boldsymbol{r}_i$更优，因此，最后$\boldsymbol{r}_i$会收敛至一个较小的值。

迭代公式由下式得出：

$$\begin{aligned}\boldsymbol{r}_{i+1}&=\boldsymbol{r}_i^{\mathrm{opt}}+\alpha\boldsymbol{r}_i^{\mathrm{step}}=\\&\boldsymbol{r}_i^{\mathrm{opt}}+\alpha\left(\boldsymbol{r}_i-\boldsymbol{r}_i^{\mathrm{opt}}\right)=\\&\alpha\boldsymbol{r}_i+(1-\alpha)\boldsymbol{r}_i^{\mathrm{opt}}=\\&\alpha\boldsymbol{r}_i+(1-\alpha)\frac{\boldsymbol{r}_i^{\mathrm{T}}\boldsymbol{w}}{\|\boldsymbol{w}\|_2}\boldsymbol{w}\end{aligned} \tag{10}$$

$$\boldsymbol{w}=\frac{\partial L(\boldsymbol{x}+\boldsymbol{r}_i)}{\partial(\boldsymbol{x}+\boldsymbol{r}_i)} \tag{11}$$

式中，$(1-\alpha)$表示学习率，$0<(1-\alpha)<1$。学习率大可能导致不收敛；学习率小，则收敛速度慢。

图 3 给出了在非线性决策器形况下，该方法的几何说明。首先利用二分搜索法选择一个接近决策边界的起始点，然后使用迭代求解，直到$\|\boldsymbol{r}_i\|_2$收敛。

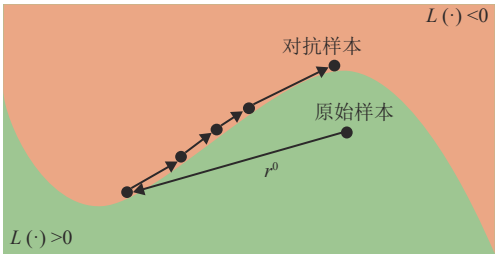


图 3　非线性决策器

2.3　针对无穷范数指标的攻击优化

上述推导中给出了针对L_2范数的对抗样本攻击

迭代公式，本文所提出的攻击方法也可以应用于最小化L_∞范数的对抗样本攻击。对于最小化L_∞范数的对抗样本生成，$\boldsymbol{r}_i$的迭代公式为：

$$\boldsymbol{r}_i \leftarrow \alpha \boldsymbol{r}_i + (1-\alpha)\frac{\boldsymbol{r}_i^{\mathrm{T}}\boldsymbol{w}}{\|\boldsymbol{w}\|_1}\mathrm{sign}(\boldsymbol{w}) \tag{12}$$

2.4 整体攻击流程

算法 2 展示了决策边界搜索攻击算法的攻击流程：首先，需要通过线性 (或二分) 方法定位到模型的决策边界；然后，计算该点相对分类模型的梯度信息，利用梯度信息更新对抗样本。新的对抗样本依然处于模型的决策边界，且更接近真实的数据点。当所找到的对抗样本满足范数要求或者对抗扰动收敛时，停止迭代并输出最后一个找到的对抗样本。

算法 2 基于决策边界搜索的对抗样本生成算法

函数输入：原样本$\boldsymbol{x}$，决策函数$L(\cdot)$

使用使用二分搜索寻找初始点$\boldsymbol{r}_0$

$i \leftarrow 0$

do

$\boldsymbol{w} \leftarrow \dfrac{\partial L(\boldsymbol{x}+\boldsymbol{r}_i)}{\partial(\boldsymbol{x}+\boldsymbol{r}_i)}$

$\boldsymbol{r}_i \leftarrow \alpha \boldsymbol{r}_i + (1-\alpha)\dfrac{\boldsymbol{r}_i^{\mathrm{T}}\boldsymbol{w}}{\|\boldsymbol{w}\|^2}\boldsymbol{w}$

$i \leftarrow i+1$

while $\|\boldsymbol{r}_i\|_2 - \|\boldsymbol{r}_{i-1}\|_2 > \sigma$

return $\boldsymbol{r}_i$

函数结束

3 实验与分析

为了测试本文算法的有效性，将所提出的算法应用于 3 个图像数据集与 3 种图像分类模型架构上，测试在受到攻击的状态下，5 个图像分类神经网络的分类性能下降程度。

3.1 实验设置

1) 数据集与分类模型。将使用 3 个图像数据集与 3 种图像分类模型用于受攻击测试。

MNIST[21]：手写体数字图像数据集，数据集中包含 70 000 张图像，其中 10 000 张为测试图像。对于该数据集，使用两个分类模型，分别为一个多层全连接神经网络与一个两层结构的 LeNet 卷积神经网络架构。

FashionMNIST[22]：MNIST 数据集的一个变体，包括 10 类不同商品的图像，数据格式与 MNIST 完全一致。对于该数据集，考虑使用两个分类模型，即一个两层结构的 LeNet 卷积神经网络架构和一个 ResNet 架构模型。

CIFAR10[23]：普适物体的图像数据集，包括飞机、车辆、船只等一共 10 种类别。对于该数据集，使用一个 ResNet 架构模型进行分类。

其中，MNIST 与 FashionMNIST 数据集的数据格式完全一致，因此所使用的 LeNet 网络架构完全一致。

本文实验使用 Pytorch 深度学习框架[24] 训练图像分类模型，表格 1 列出 5 个图像分类模型对 3 个数据集的分类精度。

表 1 受攻击模型参数

数据集	分类模型	模型分类精度/%
MNIST	全连接网络	98.1
	LeNet	99.1
Fashion-MNIST	LeNet	89.8
	ResNet	92.0
CIFAR10	ResNet	88.1

2) 最大可允许扰动ρ与对抗样本分类率。一个有效的对抗样本$\boldsymbol{x}'$被定义为：其预测标签与原数据$\boldsymbol{x}$的真实标签y不同，且$\boldsymbol{x}$与$\boldsymbol{x}'$必须足够相似，即$F(\boldsymbol{x}') \neq y$且$\|\boldsymbol{x}-\boldsymbol{x}'\|_p < \rho$。

对一个对抗样本攻击算法而言，越大的ρ表示可允许的攻击范围越大，代表其可允许的攻击强度越大。因此，分类模型对其生成的对抗样本的分类率越低。对多个攻击算法进行对比，在相同的ρ下实施攻击，分类模型对攻击算法所生成的对抗样本的分类率越低，说明该攻击算法越有效。

3) 对比方案。在 5 个训练良好的图像分类模型上，使用本文所提出的对抗样本攻击算法，分别针对L_2与L_∞范数生成对抗样本。在多个最大可允许扰动ρ值下，生成对抗样本，然后使用分类模型对对抗样本进行分类，记录在不同攻击强度下的对抗样本分类精度。本文使用 3 个经典的攻击算法进行对比。

1) 快速梯度方向算法 (fast gradient sign method, FGSM) 和快速梯度算法 (fast gradient method, FGM)[17]：FGSM 是经典的对抗样本生成方法，主要针对于生成最小化L_∞范数的对抗样本，FGM 为 FGSM 的最小化L_2范数版本。

2) 梯度投影下降算法 (projected gradient descent, PGD)[25]：FGSM 的多次迭代版本，并在此基础上加入了初始点随机化以增加攻击成功率。

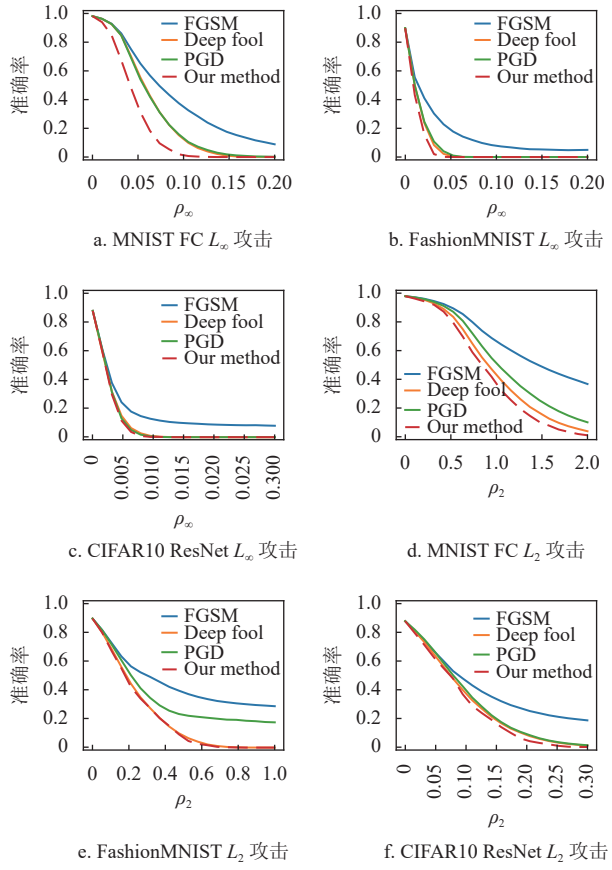
3) DeepFool[10]：基于超平面分类决策器假设的

攻击算法，可以生成对抗扰动较小的对抗样本。

除了本文外算法，其余的攻击算法由通用对抗样本攻击工具箱 Foolbox[26] 实现。

3.2　实验结果

图 4 展示了在 3 个网络模型 (FC-MNIST, LeNet-FashionMNIST, ResNet-CIFAR10) 上使用 4 种攻击算法在不同攻击强度下的攻击效果。当最大可允许扰动ρ为0时，纵坐标值表示该模型对正常图像的分类精度。图 4 表明，这 4 种攻击方法针对L_2与L_∞指标所生成的对抗样本，都可以有效地使得神经网络分类错误。



a. MNIST FC L_∞ 攻击　b. FashionMNIST L_∞ 攻击

c. CIFAR10 ResNet L_∞ 攻击　d. MNIST FC L_2 攻击

e. FashionMNIST L_2 攻击　f. CIFAR10 ResNet L_2 攻击

图 4　对抗样本攻击算法在 3 个模型上的攻击结果

从实验结果中可以看出，本文所提出的对抗样本生成算法的攻击效果最佳。在每一组实验中，在同一个最大可允许扰动ρ下，图像分类模型对本文算法生成的对抗样本的分类率都是最小的。而 DeepFool 算法与 PGD 算法效果在大多数情况下相近，而在 MNIST 与 FashionMNIST 数据集中，DeepFool 算法相比 PGD 算法生成的最小化L_2对抗样本更优。FGSM 为 4 种算法中计算效率最高的，但攻击效果最差。

本实验所使用的 3 个数据集都有 10 个图像类别，可以认为对于一个完全无分类能力的模型 (随机分类模型) 而言，其图像分类精度应为 10%。表 2～表 4 分别记录了使用 4 种攻击方式对 5 个图像分类模型进行攻击，图像分类模型对对抗样本的分类精度达到 10% 时，每种攻击方法所需要的最大可允许扰动值。

表 2　MNIST 数据集攻击实验

分类模型	攻击类型	FGSM/FGM	PGD	Deep-Fool	本文算法
FC	L_∞	0.192	0.107	0.105	**0.073**
	L_2	3.219	2.017	1.683	**1.496**
Le-Net	L_∞	0.244	0.116	0.122	**0.082**
	L_2	–	2.117	1.655	**1.557**

表 3　Fashion-MNIST 数据集攻击实验

分类模型	攻击类型	FGSM/FGM	PGD	Deep-Fool	本文算法
Le-Net	L_∞	0.083	0.033	0.030	**0.025**
	L_2	–	–	0.473	**0.467**
Res-Net	L_∞	0.309	0.031	0.032	**0.020**
	L_2	–	0.731	0.425	**0.353**

从表中可以看出，在每组实验中本文算法所需要的最大可允许扰动ρ都最小。并且，在 MNIST 与 FashionMNIST 实验中，本文所提出的算法相较于其他 3 种有较大提升，而在 CIFAR10 的实验中，本文算法与 PGD，DeepFool 算法效果相近，但是，FGSM 在多个攻击实验中无法使得图像分类模型的对抗样本分类精度下降至 10%。

表 4　CIFAR10 数据集攻击实验

分类模型	攻击类型	FGSM/FGM	PGD	Deep-Fool	本文算法
Res-Net	L_∞	0.014	0.005	0.005	**0.005**
	L_2	0.497	0.196	0.193	**0.175**

实验分析认为本文算法在 3 个数据集上的性能不同的主要原因是数据集的数据分布存在差异。MNIST 与 FashionMNIST 的图像为单通道灰度图，图像尺寸较小。而 CIFAR10 数据集为三通道图像，尺寸相对 MNIST 稍大，图像内容更加丰富。因此，在 CIFAR10 数据集上，使用线性搜索寻找攻击初始点，与有效对抗样本存在的空间距离较远，后续的迭代无法有效找到最优的对抗样本。

设计实验对该假设进行验证。在所提出的攻击算法的第一步中，将线性搜索的方法替换为使用

DeepFool 算法生成初始攻击点，以此为基础进行对抗样本搜索。表 5 中列出了改进的算法在攻击模型至 10% 精度所需要的ρ值。

表 5 使用 DeepFool 作为初始点攻击效果

数据集	分类模型	攻击类型	随机攻击点	DeepFool攻击起点
MNIST	FC	L_∞	0.073	0.061
		L_2	1.496	1.228
	LeNet	L_∞	0.082	0.064
		L_2	1.557	1.191
Fashion-MNIST	LeNet	L_∞	0.025	0.019
		L_2	0.467	0.368
	Res-Net	L_∞	0.020	0.018
		L_2	0.353	0.285
CIFAR-10	Res-Net	L_∞	0.005	0.004
		L_2	0.175	0.149

可以看出相较使用随机初始点搜索的方法，使用 DeepFool 攻击所找到的对抗样本作为初始点，在所有对比实验中攻击图像分类模型至 10% 时所需要的ρ值都要更小。

改进后的攻击算法本质上是在使用 DeepFool 搜索到一个有效的对抗样本后，使用迭代的方法进一步缩小对抗样本与原样本的距离。这表明了使用本文算法与其他对抗攻击算法结合，可以找到扰动更小的对抗样本。

4 结 束 语

本文提出了一种新的对抗攻击算法，即基于分类模型决策函数边界的对抗样本搜索算法。该算法是基于现有的神经网络图像分类模型的全局连续性与可导性，使用多步迭代的方式在分类模型的决策边界寻找一个与原数据点距离相近的对抗样本。

实验证明在数据分布简单的数据集上，本文方法可以取得最优的攻击效果。而在复杂数据集上，需要使用更有效的方式找到攻击初始点，以保证整体的攻击性能。这表明本文算法可以有效地优化其他对抗攻击算法所生成的对抗样本。因此，在后续设计其他对抗样本生成技术时，可以将本文提出的迭代的搜索方法作为优化手段，提升其他攻击算法的性能。

参 考 文 献

[1] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. ImageNet classification with deep convolutional neural networks[J]. Communications of the ACM, 2017, 60(6): 84-90.

[2] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition[C]//IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE Computer Society, 2016: 770-778.

[3] GIRSHICK R B. Fast R-CNN[C]//IEEE International Conference on Computer Vision. Santiago: IEEE Computer Society, 2015: 1440-1448.

[4] ZHU X, WANG Y, DAI J, et al. Flow-guided feature aggregation for video object detection[C]//IEEE International Conference on Computer Vision. Venice: IEEE Computer Society, 2017: 408-417.

[5] SAK H, SENIOR A W, RAO K, et al. Fast and accurate recurrent neural network acoustic models for speech recognition[C]//Conference of the International Speech Communication Association. Dresden: Interational Speech Communication Association, 2015: 1468-1472.

[6] DEVLIN J, CHANG M W, LEE K, et al. BERT: Pre-training of deep bidirectional transformers for language understanding[C]//Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Minneapolis: Association for Computational Linguistics, 2019: 4171-4186.

[7] 奚雪峰, 周国栋. 面向自然语言处理的深度学习研究[J]. 自动化学报, 2016, 42(10): 1445-1465.
XI X F, ZHOU G D. A survey on deep learning for natural language processing[J]. Acta Automatica Sinica, 2016, 42(10): 1445-1465.

[8] SZEGEDY C, ZAREMBA W, SUTSKEVER I, et al. Intriguing properties of neural networks[C]//International Conference on Learning Representations. Banff: ICLR, 2014: 1-10.

[9] CARLINI N, WAGNER D A. Towards evaluating the robustness of neural networks[C]//IEEE Symposium on Security and Privacy. San Jose: IEEE Computer Society, 2017: 39-57.

[10] MOOSAVI-DEZFOOLI S M, FAWZI A, FROSSARD P. DeepFool: A simple and accurate method to fool deep neural networks[C]//IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE Computer Society, 2016: 2574-2582.

[11] 徐明, 蒋奔驰. 基于纹理和颜色感知距离的对抗样本生成算法[J]. 电子科技大学学报, 2021, 50(4): 558-564.
XU M, JIANG B C. Adversarial examples generation method based on texture and perceptual color distance[J]. Journal of University of Electronic Science and Technology of China, 2021, 50(4): 558-564.

[12] FISCHER V, KUMAR M C, METZEN J H, et al. Adversarial examples for semantic image segmentation[C]//International Conference on Learning Representations. Toulon: OpenReview, 2017: 1-4.

[13] CHENG M, YI J, CHEN P Y, et al. Seq2Sick: Evaluating the robustness of sequence-to-sequence models with adversarial examples[C]//AAAI Conference on Artificial Intelligence. New York: AAAI Press, 2020: 3601-3608.

[14] 仝鑫, 王斌君, 王润正, 等. 面向自然语言处理的深度学习对抗样本综述[J]. 计算机科学, 2021, 48(1): 258-267.
TONG X, WANG B J, WANG R Z, et al. Survey on

adversarial sample of deep learning towards natural language processing[J]. Computer Science, 2021, 48(1): 258-267.

[15] 杜小虎, 吴宏明, 易子博, 等. 文本对抗样本攻击与防御技术综述[J]. 中文信息学报, 2021, 35(8): 1-15.
DU X H, WU H M, YI Z B, et al. Adversarial text attack and defense: A review[J]. Journal of Chinese Information Processing, 2021, 35(8): 1-15.

[16] CARLINI N, WAGNER D A. Audio adversarial examples: Targeted attacks on speech-to-text[C]//IEEE Security and Privacy Workshops. San Francisco:. IEEE Computer Society, 2018: 1-7.

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

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

[19] MODAS A, MOOSAVI-DEZFOOLI S M, FROSSARD P. SparseFool: A few pixels make a big difference[C]//IEEE Conference on Computer Vision and Pattern Recognition. Long Beach: Computer Vision Foundation IEEE, 2019: 9087-9096.

[20] POURSAEED O, KATSMAN I, GAO B, et al. Generative adversarial perturbations[C]//IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City: Computer Vision Foundation IEEE, 2018: 4422-4431.

[21] LECUN Y, BOTTOU L. Gradient-based learning applied to document recognition[J]. Proceedings of the IEEE, 1998, 86(11): 2278-2324.

[22] XIAO H, RASUL K, VOLLGRAF R. Fashion-MNIST: A novel image dataset for benchmarking machine learning algorithms[EB/OL]. (2017-09-15). https://doi.org/10.48550/arXiv.1708.07747

[23] KRIZHEVSKY A. Learning multiple layers of features from tiny images[J]. University of Toronto, 2012, 1(4): 1-58.

[24] PASZKE A, GROSS S, MASSA F, et al. PyTorch: An imperative style, high-performance deep learning library[C]//Advances in Neural Information Processing Systems. Vancouver: MIT Press, 2019: 8024-8035.

[25] MADRY A, MAKELOV A, SCHMIDT L, et al. Towards deep learning models resistant to adversarial attacks[C]//International Conference on Learning Representations. Vancouver: ICLR, 2018: 1-23.

[26] RAUBER J, ZIMMERMANN R, BETHGE M, et al. Foolbox native: Fast adversarial attacks to benchmark the robustness of machine learning models in PyTorch, TensorFlow, and JAX[J]. J Open Source Softw, 2020, 5(53): 2607.

编　辑　刘飞阳