人工智能的安全对抗技术

电子信息与电气工程学院： 沈xx 52002191xxxx

指导老师 电子信息与电气工程学院 xx

摘要

人工智能深度学习的概念源于人工神经网络的研究，由最早的感知机，到卷积神经网络、循环神经网络、生成式对抗网络等，通过多个复杂的网络层中大量数据的计算，得到输入数据的分布特征以处理不同的需求，在计算机科学，金融贸易，医药，工业，运输，游戏，音乐等诸多方面都有着重要的应用。

然而神经网络存在严重的安全问题。2014年，Christian Szegedyd等人分析了深度神经网络存在的问题，首次提出了对抗样本的概念[1]。Ian J. Goodfellow等人提出FGSM攻击算法[2],将一张被神经网络正确识别为“熊猫”加上人眼难以察觉的微小扰动后被同一个神经网络以99.3%的置信度识别为“长臂猿”。对抗样本的存在对于自动驾驶、安防等图像领域的应用方面产生了严重的威胁，因此人工智能的安全对抗技术是一个十分重要的研究方向。

本文介绍对抗样本相关概念，对一种最近提出的黑盒攻击算法模型Surfree[3]进行理解与复现，实验结果表明基于决策的黑盒攻击可在仅千次查询后就产生高质量的对抗样本。

关键词：人工智能安全，对抗样本，黑盒攻击

Abstract

The concept of artificial intelligence deep learning originated in the research of artificial neural network, from the earliest perceptron to convolutional neural networks, recurrent neural networks, generative adversarial networks and so on, which gain the distribution characteristics of the input data through complicated computations in several complex networks to handle different demands, thus playing an import role in computer science, finance and trade, medicine, industry, transportation, games, music and many other aspects.

However, there are serious security problems in neural networks. In 2014, Christian Szegedyd et al. analyzed the problems existing in deep neural networks and put forward the concept of adversarial samples for the first time[1]. Ian J. Goodfellow et al. proposed FGSM attack algorithm [2], which is able to make a picture, classified correctly as ‘Panda’ by a well-trained neural network, classified as ‘gibbon’ with 99.3% confidence by the same neural network after only applying a certain hardly perceptible perturbation. The existence of adversarial samples gives a serious threat to the application of image fields such as automatic driving. Therefore, artificial intelligence security technology is a very important research direction.

This paper introduces the concepts of adversarial samples, and replicates a recently proposed black box attack algorithm model Surfree[3]. The results show that the decision-based black box attack can produce high-quality adversarial samples within only a thousand queries.

**KEY WORDS:** artificial intelligence security, adversarial samples, black box attack

1. 绪论

1.1 深度神经网络的问题

Christian Szegedyd等人发现深度神经网络存在两个问题[1]，其一为在神经网络的高层中，包含语义信息的不是某个单元，而是整个空间；其二为神经网络学习的输入输出映射是相当不连续的，增加人眼难以感知的扰动就可以使网络产生错误的分类。Robert Geirhos等人发现卷积神经网络更倾向于识别图片的纹理而不是形状，这与人类的认知行为有极大的区别[4]。这些问题导致了对抗样本的存在。

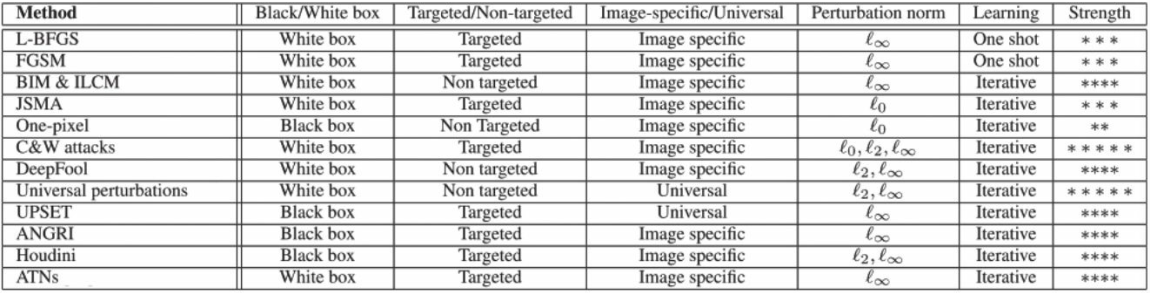
1.2 对抗样本

Christian Szegedy等人提出了对抗样本的概念[1]，即将输入样本故意添加细微的干扰形成新的输入样本，该样本可导致模型以高置信度给出一个错误的输出，且两个样本人眼无法区分。他们发现许多深度学习模型对于对抗样本都具有极高的脆弱性。有许多产生对抗样本的方式，也可称为对抗攻击的方式。

* 1. 攻击方式

根据攻击者对被攻击模型内部信息的获取程度，可以将对抗攻击大致分为两类：白盒攻击与黑盒攻击。前者可获取深度学习模型内部的所有信息和参数并借此进行攻击，例如基于给定模型的梯度生成对抗样本，而后者将神经网络结构视为黑箱，仅通过模型的输入和输出生成对抗样本。

表1 常见的对抗样本生成算法[5]



* + 1. 白盒攻击

白盒攻击的研究更加成熟，各类算法层出不穷，如FGSM[2]、Deepfool[6] 、CW[7]等。在正常训练一个网络模型时，把输入样本作为固定的量，把网络参数作为需要学习更新的量，通过损失函数对于网络参数的梯度更新网络参数，多次迭代后得到训练好的模型。而生成对抗样本的方式则正好相反，把网络的参数作为固定的量，根据损失函数对输入样本的梯度对输入样本进行修改，通过一次或多次迭代产生对抗样本。利用梯度信息是众多白盒攻击算法的手段，可在短时间内生成高质量的对抗样本。

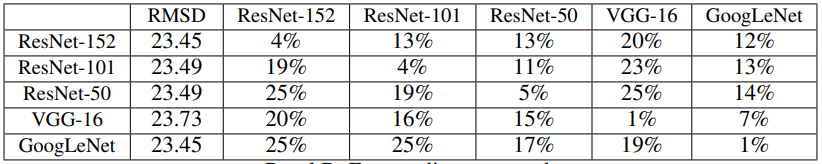
* + 1. 黑盒攻击

大多数现实世界的系统是“黑盒子”，白盒攻击并不适用。然而黑盒攻击对模型内部的结构细节不了解，仅能控制模型的输入，攻击难度明显上升。目前主要实现方法有基于迁移的攻击（transfer-based）和基于访问的攻击（query-based）。前者能够以极少的访问次数实现攻击，但不能保证个体级别的攻击成功，后者则依赖于模型的预测分数，以高访问量的代价来提升个体攻击的成功率。

1.3.2.1 基于迁移的攻击

迁移学习攻击算法的基本思想是基于结构相似的深度学习网络在面对相同的对抗样本的攻击时具有类似表现的特点。大量知名的机器视觉领域的深度学习模型都已经开源， 比如VGG、ResNet等， 这为迁移学习提供了大量的资源。Yanpei Liu等人对迁移学习攻击算法进行了系统的阐述[8]。

表2 对一模型使用FGSM生成的对抗样本被另一模型正确分类的比例[8]



1.3.2.2 基于访问的攻击

根据访问过程中所获信息的特点，基于访问的攻击可以细分为基于分数的攻击（score-based）和基于决策的攻击（decision-based）。前者获得的输出是模型预测的top-k概率而后者获得的输出是模型预测的top-k标签，对于后者若k = 1则难度更大因为可利用的信息更少。

Wieland Brendel等人开创基于决策的攻击算法[9]，不需要任何梯度或分数信息，可使得之前的众多防御方法失效，但需要百万次访问来生成一个对抗样本。随后Jianbo Chen等人提出HSJA算法[10]，将访问次数降低至万次； Ali Rahmati等人提出GeoDA算法[11]，将访问次数降低至几千次。

Thibault Maho等人在今年的CVPR中提出的基于决策的黑盒攻击算法Surfree，该算法仅需模型的top-1输出标签，且较其他基于决策的黑盒攻击算法收敛速度较快，需要更少的访问次数[3]。本文对于Surfree模型算法进行理解、复现与实验并给出简单的复现代码，在千次访问后即可生成高质量的对抗样本。

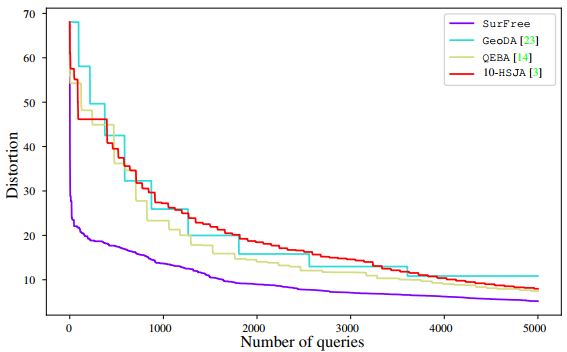


图1 多种基于决策的黑盒攻击算法效率对比[3]

1. 研究内容及方法

笔者在阅读理解了Surfree算法原理后重新整理阐述，思路与原论文[3]略有区别，且补充了相关数学问题的证明。

* 1. 算法目的

设输入样本为维向量, 共个类别，神经网络分类器为,，对于待攻击的输入样本,对应的对抗样本集合为，需要寻找到与原输入样本差距最小的最优对抗样本,这里使用距离,即

* 1. 算法原理

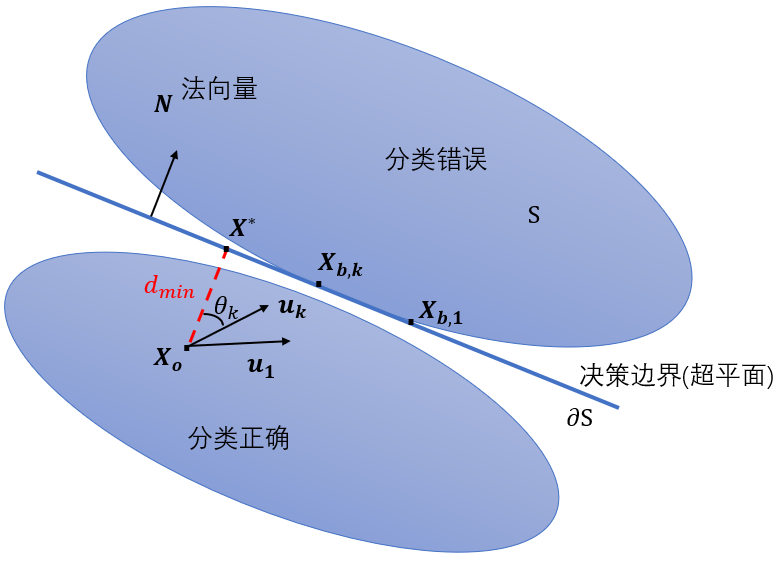


图2 算法原理的几何模型

设,,找到任意一组的正交补空间的基，记为，即满足,因此可作为样本空间的一组标准正交基

令其中为未知常数，可知

令

可知

令且**,** 即

令 , 则

由图4：

下用数学归纳法证明

k=1时,成立

假设k=m时成立,即,

则k=m+1时， 也成立，证毕

由(4)(5)式:

将(1)(6)式代入(3)式有

由(7)式进行递推，最终得到 **,** 且

对于向量组，并非预先生成，而是在迭代的过程中生成，生成的方式为：对已经相互正交的及**、**D维随机向量组成的向量组进行施密特正交化，由得到，具体计算方式如下：

施密特正交化过程中，由于已经相互正交且标准化，故只对**、**进行操作

由施密特正交化结论可知，与正交，同时由(8)式可知是的线性组合，故与正交。又由递推式(1)可知是的线性组合，故与也正交。所以由此种方式生成的向量组符合最初的定义条件。

由于未知，可通过搜索的方式寻找合适的使得通过(6)式生成的

* 1. 算法过程

对于待攻击的输入样本，取任意,,二分使

由递推得的算法如下(,为超参数):

①由(2)式生成，生成D维随机向量，由(8)(9)(10)(11)式计算得到这里向量组可仅保留后个

②按的顺序搜索，根据(7)式更新,使得,若所有均失败则重新生成此次的与

③在的范围内二分搜索(限制搜索次数为),使得

为最优的对抗样本。

然而很多时候更注重生成对抗样本的效率，因此不需要迭代至最后一次，而是在访问次数到达某一阈值后就停止。由于这种算法收敛速度快，可以保证每次迭代后样本的失真情况()都会迅速变小，因此提前结束迭代也可以生成高质量的对抗样本。

1. 研究结果及讨论

3.1 实验设定

实验使用pytorch1.7框架，mnist与cifar10数据集中的样本均将像素数据归一化至[0,1]范围内。在mnist数据集上使用自定义的CNN网络，在cifar10数据集上使用预训练的resnet18模型，均训练5次，达到90%以上的准确率。在选择待攻击的样本时会先确保模型正确分类此样本。

表3 实验数据集与网络模型

|  |  |  |
| --- | --- | --- |
| 数据集 | mnist | cifar10 |
| 样本维度 |  |  |
| 神经网络模型 | 两层卷积池化层的CNN | 预训练的resnet18 |
| 测试集top-1准确率 | 98% | 91% |

超参数的选择与原论文[3]一致，如下表所示。

表4 实验超参数

|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  |  |
| 100 | 3 |  | 10 |

3.2 实验结果

对于mnist数据集，,即完成783次迭代即可得到此算法下的最优样本，在实验设定的超参数下大约访问次数为17000次。而对于cifar10数据集, ，若完成所有的次迭代大约需要百万次的访问次数，效率太低。因此选择对两个数据集的所有类别样本均使用Surfree攻击算法访问17000次得到对抗样本。

表5 访问17000次生成的对抗样本与原样本对比

|  |  |  |
| --- | --- | --- |
| mnist | 原样本 |  |
| 对抗样本 |  |
| cifar10 | 原样本 |  |
| 对抗样本 |  |

由mnist数据集生成的对抗样本可见，数字的形状、轮廓均与原样本一致，但在背景中还是有较明显的噪声存在，人眼可以识别出这些数字，但也能看出扰动的痕迹。笔者认为这样的原因在于mnist数据集中像素点仅有0，1两种取值，而算法中进行的运算是连续的，因此完全消除噪声十分困难；其次，由于mnist数据集样本较简单(仅有784维)，CNN网络识别能力较强，理论上最小的失真程度可能也会被人识别出。

而cifar10数据集避免了上述两个问题，在17000次访问后产生的对抗样本几乎与原样本无差别，在物体、背景、轮廓均一致的情况下并无可识别出的扰动痕迹存在，可见此攻击算法对于一般的图片(非按某种特定规则生成)能力较强。

接着在cifar10数据集上实验，减少访问次数至1000次，并观察样本失真程度随访问次数增加的变化情况。

表6 不同访问次数下的对抗样本

|  |  |  |  |
| --- | --- | --- | --- |
| 访问次数  原样本 | 100 | 500 | 1000 |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

由表6可看出，在100次访问后，生成的对抗样本在整张图的范围内都具有非常明显的噪声，失真较严重；在500次访问后，对抗样本较原图的噪声情况改善很多，但仍然可以一眼看出噪声的存在；在1000次访问后，对抗样本的质量具有明显改善，在不放大仔细观察的情况下与原图十分相似。在本实验条件下，使用1个gpu，访问1000次得到一个对抗样本的时间为8秒左右。由此可见Surfree攻击算法十分高效，可将生成一张高质量对抗样本的访问次数降低至一千次以内。

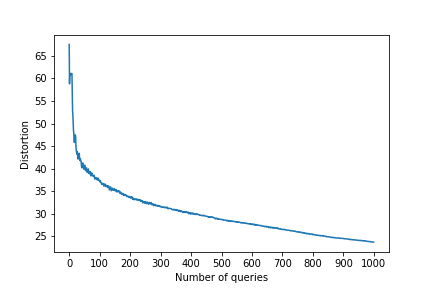


图3 在cifar10数据集下对抗样本失真情况随访问次数的变化

本实验采用距离衡量对抗样本相对于原样本的失真情况，即

由算法原理部分可知，随着迭代次数的增加，一定单调减小，但每次迭代包含若干次访问，这些访问是为了寻找决策边界上的新一轮的对抗样本，因此在此过程中并不一定单调减小。这导致总体来看呈减小趋势，但并不单调，会有不少局部上下浮动的情况，且由于初始对抗样本的随机性，每次的范围也不尽相同。因此选择进行100次实验，每次进行1000次访问并记录得到的对抗样本与原样本的距离即值，取平均值后得到图3的结果。

根据图3可以发现，Surfree算法可使得对抗样本的失真情况在100次访问之内就迅速下降，此后保持稳定的下降速率，在1000次访问后值仅有最初的左右。

1. 结论

本文详细阐述了Surfree攻击算法[3]的数学原理，相较于原论文的推导更简洁易懂，同时对该算法进行复现、实验。实验表明了Surfree算法生成对抗样本的良好效果以及其高效的特点，证明了在基于决策的黑盒攻击算法中少于1000次访问得到高质量对抗样本的可行性。未来可学习更多的黑盒攻击算法并与此算法进行对比，补充更多的实验。

5. 参考文献

[1] Christian Szegedy, Wojciech Zaremba, et al., ‘Intriguing properties of neural networks’, in International Conference on Learning Representations, 2014

[2] Ian J. Goodfellow, Jonathon Shlens, et al., ‘Explaining and Harnessing Adversarial Examples’, in International Conference on Learning Representations,2015

[3] Thibault Maho, Teddy Furon, et al., ‘SurFree: a fast surrogate-free black-box attack’, in IEEE Conference on Computer Vision and Pattern Recognition, 2021

[4] Robert Geirhos, Patricia Rubisch, et al., ‘ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness’, in International Conference on Learning Representations, 2019

[5] 兜哥，AI安全之对抗样本入门，机械工业出版社，2019，109

[6] Seyed-Mohsen Moosavi-Dezfooli,et al.,‘DeepFool: a simple and accurate method to fool deep neural networks’ in IEEE Conference on Computer Vision and Pattern Recognition, 2016

[7] Nicholas Carlini, David Wagner,‘Towards Evaluating the Robustness of Neural Networks’, in IEEE Symposium on Security and Privacy, 2017

[8] Yanpei Liu, Xinyun Chen, el al., ‘Delving into Transferable Adversarial Examples and Black-box Attacks’, in International Conference on Learning Representations, 2017

[9] Wieland Brendel, Jonas Rauber, et al., ‘Decision-based adversarial attacks: Reliable attacks against black-box machine learning models’, in International Conference on Learning Representations, 2018

[10] Jianbo Chen, Michael I Jordan, et al., ‘HopSkipJumpAttack: A query-efficient decision-based attack’, in IEEE Symposium on Security and Privacy, 2020

[11] Ali Rahmati, Seyed-Mohsen Moosavi-Dezfooli, et al., ‘GeoDA: a geometric framework for black-box adversarial attacks’, in IEEE Conference on Computer Vision and Pattern Recognition, 2020