软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2020,31(1):67−81 [doi: 10.13328/j.cnki.jos.005884]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

对抗样本生成技术综述*

潘文雯, 王新宇, 宋明黎, 陈 纯

(浙江大学 计算机科学与技术学院,浙江 杭州 310027)

通讯作者: 宋明黎, E-mail: brooksong@zju.edu.cn

摘 要: 如今,深度学习已被广泛应用于图像分类和图像识别的问题中,取得了令人满意的实际效果,成为许多人工智能应用的关键所在.在对于模型准确率的不断探究中,研究人员在近期提出了"对抗样本"这一概念.通过在原有样本中添加微小扰动的方法,成功地大幅度降低原有分类深度模型的准确率,实现了对于深度学习的对抗目的,同时也给深度学习的攻方提供了新的思路,对如何开展防御提出了新的要求.在介绍对抗样本生成技术的起源和原理的基础上,对近年来有关对抗样本的研究和文献进行了总结,按照各自的算法原理将经典的生成算法分成两大类——全像素添加扰动和部分像素添加扰动.之后,以目标定向和目标非定向、黑盒测试和白盒测试、肉眼可见和肉眼不可见的二级分类标准进行二次分类.同时,使用 MNIST 数据集对各类代表性的方法进行了实验验证,以探究各种方法的优缺点.最后总结了生成对抗样本所面临的挑战及其可以发展的方向,并就该技术的发展前景进行了探讨.

关键词: 深度学习;对抗样本生成;扰动;目标定向;目标非定向;黑盒测试

中图法分类号: TP18

中文引用格式: 潘文雯,王新宇,宋明黎,陈纯.对抗样本生成技术综述.软件学报,2020,31(1):67−81. http://www.jos.org.cn/1000-9825/5884.htm

英文引用格式: Pan WW, Wang XY, Song ML, Chen C. Survey on generating adversarial examples. Ruan Jian Xue Bao/Journal of Software, 2020,31(1):67−81 (in Chinese). http://www.jos.org.cn/1000-9825/5884.htm

Survey on Generating Adversarial Examples

PAN Wen-Wen, WANG Xin-Yu, SONG Ming-Li, CHEN Chun

(School of Computer Science and Technology, Zhejiang University, Hangzhou 310027, China)

Abstract: Recently, deep learning has been widely used in image classification and image recognition, which has achieved satisfactory results and has become the important part of AI applications. During the continuous exploration of the accuracy of models, recent studies have proposed the concept of "adversarial examples". By adding small perturbations to the original samples, it can greatly reduce the accuracy of the original classifier and achieve the purpose of anti-deep learning, which provides new ideas for deep learning attackers, and also puts forward new requirements for defenders. On the basis of introducing the origin and principle of generating adversarial examples, this paper summarizes the research and papers on generating adversarial examples in recent years, and divides these algorithms into two categories: entire pixel perturbation and partial pixel perturbation. Then, the secondary classification criteria (targeted and not targeted, black-box test and white-box test, visible and invisible) were used for secondary classification. At the same time, the MNIST data set is used to validate the methods, which proves the advantages and disadvantages of the various methods. Finally, this paper summarizes the challenges of generating adversarial examples and the direction of their development, and also discusses the future of them.

Key words: deep learning; adversarial examples; perturbation; targeted; no targeted; black-box test

随着深度学习的概念被广泛应用于当今研究的诸多方面,各类相关算法层出不穷,如何提高其质量、降低

* 基金项目: 国家自然科学基金(61572426, 61572428)
Foundation item: National Natural Science Foundation of China (61572426, 61572428)
收稿时间: 2018-11-19; 修改时间: 2019-03-07, 2019-07-08; 采用时间: 2019-09-03; jos 在线出版时间: 2019-11-06
CNKI 网络优先出版: 2019-11-06 11:49:16, http://kns.cnki.net/kcms/detail/11.2560.TP.20191106.1148.007.html

其所需的时间和内存代价,也得到越来越广泛的关注.其中,对抗样本的概念应运而生.所谓对抗指的是对于深度学习的攻击;所谓对抗样本,就是能够使得深度学习出现错误的一类合成样本.

最早提出“对抗样本”这一概念的是 Szegedy 等人[1],文献使用在原有样本的像素上添加扰动的方法,促使包括卷积神经网络在内的深度学习模型出现显著的准确率降低现象.Nguyen 等人[2]提出:面对一些人类无法识别的样本,深度学习模型也可以将其以高置信度进行分类.这意味着深度学习模型具有极大的脆弱性,在理论上存在凭借垃圾样本通过识别分类系统的可能性.2016 年,Pedro 等人[3]提出,对抗样本的应用在图像方面占据了极大的空间.至此,深度网络和对抗样本成为研究热潮,而 Goodfellow 等人[4]在解释了深度网络的高维线性是对抗样本生成的原理后,围绕于生成对抗样本的迭代算法开始涌现,出现了各类流派.对抗样本的产生对深度学习的攻防两面都有很大的实际意义,适当应用能在信息安全和隐私保护等方面达到一定的效果.对于生成对抗样本各类算法的总结和分类,可以在找到基本生成规律的基础上发掘更多的创新点和应用价值.

本文重点对生成对抗样本的已有研究工作进行综述,主要选取了近年来有代表性的或取得比较显著效果的方法进行详细的原理介绍和优缺点分析.按照其生成方式和原理的不同,分为全像素添加扰动和部分像素添加扰动两类.在此基础上,根据目标是否定向、是否黑盒和是否肉眼可见这 3 个标准进行细分,将各类方法中的代表性算法在统一数据集(MNIST)上进行测试,验证并分析其优缺点,最终总结提出未来的发展前景.

本文第 1 节主要介绍对抗样本的基本概念和基础知识,包括对抗样本本身的定义、其延伸有关的相关概念以及基本操作流程.第 2 节则指出对抗样本是从深度学习中衍生出来的概念,同时介绍了对抗样本有效性的评估方法.第 3 节则介绍对抗样本的起源,说明了对抗样本的产生契机和原理解释.第 4 节介绍生成对抗样本的发展状况,以全像素添加扰动和部分像素添加扰动两大类进行算法说明,同时总结生成方法中常用的数据集.第 5 节是对第 4 节中代表方法的实验,结合对同一数据集的效果测试来说明各类方法的优缺点.通过这些优缺点,在第 6 节中讨论对抗样本生成技术面临的挑战和前景预测.

1 简 介

1.1 对抗样本的定义

对抗样本(adversarial example)是指在原数据集中通过人工添加肉眼不可见或在经处理不影响整体的肉眼可见的细微扰动所形成的样本,这类样本会导致训练好的模型以高置信度给出与原样本不同的分类输出[5].

1.2 相关概念

- 扰动(perturbation):对抗样本生成的重要部分.一般来说,扰动需要有两个方面的要求:一是要保证其微小性,达到添加后肉眼不可见或者肉眼可见但不影响整体的效果;二是将其添加到原有图像的特定像素上之后,所产生的新图像具有迷惑原有分类深度模型的作用;
- 对抗训练(adversarial training):指的是将按照一定的算法生成的对抗样本标注为原样本的类别,将这些对抗样本和原始样本混合在一起作为训练集,供分类器进行训练,是众多对抗防御方法中具有代表性的一类方法;
- 对抗性(adversarial):指的是对抗样本对原有分类器的迷惑程度,可以用分类器分类的准确率来衡量;
- 黑盒测试(black box test)[6]:未知模型内部结构与参数,从输入、输出数据的对应关系进行测试的方法;
- 白盒测试(white box test)[6]:在已知模型内部结构与参数的情况下进行测试的方法,与黑盒测试相对;
- 对抗样本的鲁棒性(robustness of adversarial examples)[7]:指的是对抗样本在经过复杂的光照、变形、去噪、转换或防御过程后,仍保持对模型攻击能力的一种性质.

1.3 基本操作流程

对抗样本的生成方法有多种,但是究其根本都有一定的操作流程[8].总体来说,对抗样本的生成与检测可以通过以下几个步骤完成,如图 1 所示.① 用正常的样本数据训练网络分类器;② 原始图片中加入扰动;③ 使用处理过的对抗样本输入到分类器中进行分类,得到分类误差;④ 对抗训练,得到分类误差而测试对抗样本的鲁

棒性;⑤ 重复进行第③步.

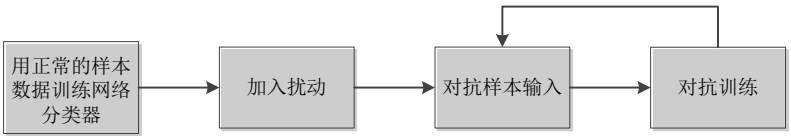


Fig.1 Schematic diagram of common process for generating adversarial examples

图 1 生成对抗样本及对抗训练常用流程示意图

2 前 传

2.1 机器学习在分类问题中的运用

机器学习的应用效果主要取决于训练数据集的属性和模型的特性.而在应用需求较广的图像分类方面,模型可以分为支持向量机和卷积神经网络两类.

- 支持向量机:将输入的图像作为一个向量进行输入,由于图像形成的像素向量横向维度较大,使用主成分分析进行一些无用像素的去除,达到降维的目的.输入的标签分为正例(+1)和负例(–1)两类,将图像向量和标签作为共同输入,SVM 的输出是寻找一个分离超平面,以达到将正例和负例的数据分离的实验效果;
- 卷积神经网络:神经网络是参考人类大脑构成和神经元信息传递模型而创建的一种算法,在图像分类问题中取得了很好的效果.卷积神经网络由输入层、卷积池化层和全连接层三大部分组成,使用特征提取器、权值共享和卷积核完成特征的提取和训练.通过一系列的降维操作,最终达到较好的分类效果.

2.2 深度学习在分类问题中的运用

深度学习在二分类和多分类的问题中取得了不错的实际效果[9],将深度学习运用到分类问题中可分为 3 个步骤[10]:用训练数据及其标签训练模型、输入验证数据和实际应用.深度学习的效果主要取决于训练数据的属性和模型的特性[11],在需求较广的图像分类方面,模型可以分为无监督学习模型和监督学习模型[12].

- 无监督学习模型[13]:是指在训练集数据缺少先验标签的条件下进行训练的机器学习方法,分为 3 种:基于限制玻尔兹曼机[14]的方法,利用能量函数拟合离散分布;基于自动编码器[15]的方法,通过输入经过编码映射到特征空间,特征经过解码映射回数据空间完成数据重建,进一步学习从输入到特征空间的映射关系;基于稀疏编码[16]的方法,借鉴神经学中大脑对视觉信号的处理方式,发掘良好过完备基向量;
- 监督学习模型[17]:监督学习模型所使用的数据集是带有相应标签的,比较有代表性的方法有多层感知器(MLP)和卷积神经网络(CNN)[18].其中,多层感知器[19]是一种引入多隐层结构的前馈神经网络,常用于模式分类;卷积神经网络[20]在第 2.1 节中已有描述.

目标物体分类是计算机视觉的基本问题[21],而深度学习能深层表达数据内部潜藏的复杂结构和规则.Andrew 等人[22]提出的有机组合递归神经网络在3D物体分类中取得了很好的效果,Karpathy 等人[23]和 Sanchze-Riera 等人[24]均使用基于 CNN 的模型分别在视频分类和手势分类中有所进展.随着深度学习优秀的表征表达能力被逐步发掘,深度学习在交通领域[25]、安防领域[26]和专业领域图像分类[27]等方面都有良好的应用.

2.3 评估方法

使用深度学习进行图像分类对效果有不同的评估标准,有准确率、精确率等多种[28],见表 1.*TP* 是原本为正判定为正的样本数,*FP* 代表原本为负判定为正的样本数,*FN* 代表原本为正判定为负的样本数,*TN* 代表原本为负被判定为负的样本数.

- 准确率(accuracy):是指分类深度模型分类正确的样本数和总样本数之比,计算方式如公式(1):

$$accurary = \frac{TP + TN}{TP + FP + FN + TN} \tag{1}$$

- 精确率(precision):也称查准率,计算方式如公式(2),精确率越高,则模型找准正类样本的能力越强:

$$precision=\frac{TP}{TP+FP} \tag{2}$$

- 召回率(recall):也称查全率,计算方式如公式(3),召回率越高,则模型找全所有正类的能力越强:

$$recall=\frac{TP}{TP+FN} \tag{3}$$

- F_1 值:精确率和召回率的调和均值,计算方式如公式(4),可以在准和全两方面找到较为综合的效果:

$$F_1=\frac{2\cdot precision\cdot recall}{precision+recall} \tag{4}$$

- 交叉熵损失函数:将交叉熵作为损失函数,可以很好地避免梯度消散,以得到比较好的实验结果.具体计算方式如公式(5),其中,$p(x)$为样本标签,而 $q(x)$为模型的预估:

$$H(p,q)=-\sum_x p(x)\log q(x) \tag{5}$$

- ROC 曲线:负正类率为横坐标,召回率为纵坐标,曲线覆盖面积越大,代表模型的综合分类效果越好.

Table 1 Schematic table of *TP*, *FP*, *FN*, *TN*

表 1 *TP*,*FP*,*FN*,*TN* 示意表

	正类	负类
检索到	*TP*	*FP*
未被检索到	*FN*	*TN*

3 起 源

3.1 首次发现

Szegedy 等人[1]在 2013 年首次提出对抗样本的概念,后被称为 L-BFGS.文献认为:单个神经元无法代表某个特征,特征的代表由整个空间描述,单个神经元并不具备语义信息.公式(6)描述了最后一个全连接层的神经元所包含的语义信息,而将特定神经元的输出最大化的结果和随机选取分量进行极大化的结果差别不大:

$$x'=\arg\max_{x\in Z}\langle\phi(x),e_i\rangle \tag{6}$$

其中,x 为输入图像,$\phi(x)$为某层的激活值,Z 是网络还没有训练过的数据分布的保留图像集,e_i 是第 i 个隐藏单元的基础向量.神经网络所学习到的输入输出映射函数不连续,具有高维非线性的特点,所以只需要对图像上进行微扰,就能使图像以高置信度被错误分类.文献实验发现:在图像中加入微小的扰动就可以对分类器进行巨大干扰,即找出符合公式(7)的ρ,使得原样本分类为最近的错误标签 l,具体实现方式如公式(7):

$$\min_{\rho}\|\rho\|_2 \quad \text{s.t. } C(x+\rho)=l;x+\rho\in[0,1]^m \tag{7}$$

其中,ρ是需要加入的扰动,x 是指原有的图像,C 是指网络分类器,l 则是需要网络误分类的分类结果.由于$\|\rho\|_2$的最小值实际计算起来比较困难,文献采用最小化损失函数的添加项,即改变公式(7)为公式(8):

$$\min_{\rho}C|\rho|+\ell(x+\rho,l) \quad \text{s.t. } x+\rho\in[0,1]^m \tag{8}$$

其中,$\ell(x+\rho,l)$是损失函数,一般可以通过交叉熵实现.Szegedy 等人[1]使用 L-BFGS 算法实现优化,以尽量减少内存负担,这种生成算法具有生成速度快、内存占用少的特点,但在对抗性方面还存在很大的提高空间.

3.2 基本原理

Goodfellow 等人[4]在 2014 年解释了对抗样本的基本原理,证明了高维网络实际的呈现状态为线性.文献认为:对抗样本之所以对于攻击分类器有显著的效果,并非传统所认为的网络高维非线性,而恰恰是网络高维线性导致的.假设原输入图像为 x,其对应的对抗样本为 $x'=x+\rho$,设置分类器的相关权重为ω,其中,限定$\|\rho\|_\infty<\varepsilon$,以保证在图像中所加的扰动$\eta$微小且肉眼不可见,则对抗样本进入分类器后完成如公式(9):

$$\omega^T x'=\omega^T x+\omega^T\eta \tag{9}$$

可以发现:当网络拥有高维属性的时候,扰动η所带来的改变$\omega^T\rho$越大,从而导致分类器的误分类.在此基础上,文献提出了一个基于梯度下降原理的对抗样本生成方法,后文称 FGSM.FGSM 通过在梯度方向上进行添加增量来诱导网络对生成的图片进行误分类,而梯度可以通过反向传播算法计算获得,如公式(10):

$$\rho=\varepsilon sign(\nabla\Im(\theta,x,y)) \tag{10}$$

其中,ρ是需要添加的扰动,θ为分类模型的参数,x 为模型的输入,y 是输入的正确标签,$\Im(\theta,x,y)$所求得的是训练神经网络的损失函数.FGSM 同样具有内存负担小的优点,通过公式(10),可有效地计算出对抗扰动.Miyato 等人[31]将公式(10)转换成公式(11),使用 l_2 范数归一化来计算梯度,最终得到加入扰动的具体值:

$$\rho=\varepsilon\frac{\nabla\Im(\theta,x,y)}{\|\nabla\Im(\theta,x,y)\|_2} \tag{11}$$

4 发 展

自 Szegedy 等人[1]提出对抗样本的概念以及 Goodfellow 等人[4]证明了神经网络高维度线性是导致对抗样本对抗性较好的根本原因后,逐渐出现一系列的对抗样本生成方法,基本上具有低生成成本和较好的效果.除 L-BFGS 和 FGSM 外,大多数的拓展衍生方法都基于迭代算法,以生成对抗性和鲁棒性更好的对抗样本.

现今方法主要可以分为两大类:全像素添加扰动和部分添加像素扰动.全像素添加扰动是指在原图像的所有像素点均加上合适的扰动,部分像素扰动只对原图像的部分像素进行修改.全像素扰动注重对抗性和适应性的提高,即,是否能产生更高的误分类率和是否具有转移性等特性;部分像素扰动更加注重选择的像素数量、代价和对抗性之间的关系.以此分类,可突出各自不同的期望目标,即寻求质量提高还是寻求扰动微小.

可以发现:使用以上的标准分类后,方法种类还是偏多,故引进二级分类标准.2017 年,Kaggle 组织的 NIPS 大赛将攻击深度学习分为目标针对性和非目标针对性两类:目标针对性[29]是指对抗样本需要使得分类器误分类为某个特定的类别;非目标针对性[30]是指对抗样本使得分类器产生误分类即可,不需要指定特定的误分类类别.在对抗深度学习的方法模型构造时,黑盒测试还是白盒测试也是攻击者们关注的一点,原有分类器的参数是否已知是决定采用哪一种方式的最主要的因素.黑盒测试通常也被称为功能测试,其方法是将程序看作一个不知内部结构和属性的黑盒,直接使用接口进行测试,应用于对抗样本的测试,可以看作在不知道分类器的参数和内部结构的条件下直接测试.与之相对的白盒测试则是在已知分类器的参数和内部结构的情况下,使用相应的样本进行测试和调整.不同的对抗样本生成方式,可以通过是应用于黑盒测试还是应用于白盒测试这样的区别进一步加以分类.对于部分像素的方法,总体来说有两类:一种是添加肉眼不可见的扰动,一种是经过一系列的算法后添加肉眼可见、同时不影响整体的扰动.两类对于局部选择有不同的要求.

综上,本文将通过全像素添加扰动和部分像素添加扰动(一级分类),辅助目标针对性和非目标针对性、黑盒测试和白盒测试、肉眼可见和肉眼不可见(二级分类)对对抗样本的生成方法分别进行介绍.

4.1 分类方式及代表模型

4.1.1 全像素扰动的生成方法

(1) 非目标定向方法

(a) I-FGSM

在 Goodfellow 等人[4]证明神经网络高维度线性是导致对抗样本有较强对抗性的原因,且使用 FGSM 取得效果后,Alexey 等人[32]提出了基础迭代法,其基本思想是:通过多个小步增大损失函数的处理来优化一大步运算增大分类器的损失函数而进行图像扰动,以得到对抗性更好的对抗样本,主要计算过程如公式(12):

$$x'_0=x;x'_{N+1}=Clip_{x,\varepsilon}\{x'_N+\alpha sign(\nabla_x\Im(x'_N,y_{\text{true}}))\} \tag{12}$$

其中,x 表示原三维图像;y_{true} 代表图像 x 本来的标签类别;$\Im(x,y)$代表分类器分类结果交叉熵;$Clip$ 代表裁剪函数,对图像的每一个像素进行裁剪操作,此裁剪函数结果的具体实现如公式(13):

$$Clip_{x,\varepsilon}\{x'\}=\min\{255,x+\varepsilon,\max\{0,x-\varepsilon,x'\}\} \tag{13}$$

整个裁剪函数保证了裁剪结果维持ε-邻状态,且不会超过图像原先定义的 255 最大值.Alexey 等人在文献

[31]中将参数α设置为 1,即,在每一步中都改变一个像素的值来进行迭代生成.I-FGSM 通过迭代的方式极大地提高了对抗样本的对抗性和鲁棒性.

(b) DeepFool

同样,目标非针对性的方法还有 Seyed-Mohsen 等人[33]提出的 DeepFool,其对深度网络也有很强的对抗性和鲁棒性.假设分类器的分类函数 $f(x)=w^Tx+b$,根据分类函数,可知其仿射平面为$\Gamma=\{x:w^T+b=0\}$.当在某一点 x_0 加入扰动后垂直于平面Γ,则加入的扰动最小且可以符合迭代要求,如公式(14):

$$\rho_*(x_0):=\arg\min\|\rho\|_2 \quad \text{s.t. } sign(f(x_0+\rho))\neq sign(f(x_0))=-\frac{f(x_0)}{\|w\|_2^2} \tag{14}$$

在整体迭代过程中,对抗样本生成符合公式(15):

$$\arg\min_{r_*}\|r_*\|_2 \quad \text{s.t. } f(x_i)+\nabla f(x_i)^T r_*=0 \tag{15}$$

由公式(13)和公式(14),DeepFool 通过迭代来生成最小范数对抗扰动,每一步将位于分类边界内的像素一步步修改到边界外,直到最终出现分类错误为止.这种方法在保持与 FGSM 差不多的对抗性的同时,所产生的扰动更小.

(2) 目标定向方法

(a) ILCM

ILCM 是对 I-FGSM 的改进,由 Goodfellow 等人[32]提出,完成目标非针对性到针对性的转换.选择样本中对原图像分类置信度最低类别作为对抗样本的期望分类,即:对于训练好的网络,达到公式(16)的效果:

$$y_{LL}=\arg\min_y\{p(y\mid X)\} \tag{16}$$

为使对抗样本分类为 y_{LL},沿着 $sign\{\nabla_X\log p(y_{LL}|X)\}$的方向进行迭代,最大化 $\log p(y_{LL}|X)$,如公式(17):

$$x'_0=x;x'_{N+1}=Clip_{x,\varepsilon}\{x'_N-\alpha sign(\nabla_x\Im(x'_N,y_{LL}))\} \tag{17}$$

与 I-FGSM 方法进行对比,可以看出:公式(12)中的 y_{true} 改为 y_{LL},将误分类局限于特定类别,达到更有意义的结果.

(b) C&W attacks

C&W attacks 是由 Carlini 和 Wagner[34]在总结了 L-BFGS、FGSM 和 JSMA 几个对抗样本生成的方法后,提出了在范数 l_0、l_2 和 l_∞上均有较大改善的算法,是前 3 种方法的拓展.文中将对抗样本的生成方式进行了适当的改变,如公式(18)所示:

$$\min D(x,x+\rho)+c\cdot f(x+\rho) \quad \text{s.t.} \quad x+\rho\in[0,1]^n \tag{18}$$

其中,D 是距离度量;f 函数满足当 $C(x+\rho)=t$ 时,当且仅当 $f(x+\delta)\leqslant 0$;c 为常量;δ为添加的扰动.与传统表达不同的是:直接加入 $f(x+\rho)$项,使得两个最小化合并在一个公式中.将距离度量通过 l_p 范数进行实例化,则可将公式(18)改为公式(19).

$$\min\|\delta\|_p+c\cdot f(x+\rho) \quad \text{s.t.} \quad x+\rho\in[0,1]^n \tag{19}$$

对于 l_2 范数的攻击,确立 t 为需要误判的类别,得到公式(20):

$$\left.\begin{aligned}&\min\left\|\frac{1}{2}(\tanh(w)+1)-x\right\|_2^2+c\cdot f\left(\frac{1}{2}(\tanh(w)+1)\right)\\&f(x')=\max(\max\{Z(x')_i:i\neq t\}-Z(x')_i-\kappa\end{aligned}\right\} \tag{20}$$

其中,$\frac{1}{2}(\tanh(w)+1)-x$ 计算的是加入的扰动,保证了其肉眼的不可见性.f的定义设定如上,是实际实验中效果较好的,而对于整个公式,所需要得到的就是导致扰动的变量 w.

对于 l_0 范数的攻击,根据公式可以发现,与标准的梯度下降法有很大的相似性.文献中使用了迭代的算法,在每一次的迭代中,对结果影响不大的像素进行操作,同时使用 l_2 范数的攻击来判别哪些像素点是不重要的.对于在 l_2 范数攻击中得到的δ,计算其梯度 $g=\nabla f(x+\rho)$,对 i 进行调整(将 i 移除原有集合),并通过 $g_i\rho_i$ 可计算出第 i 个

像素产生的 f 函数下降幅度.

对于 l_∞范数的攻击,可得到公式(21):

$$\min\|\rho\|_\infty+c\cdot f(x+\rho) \tag{21}$$

为使得梯度下降会产生较好的结果,将上式演化为公式(22):

$$\min c\cdot f(x+\rho)+\sum_i[(\rho_i-\tau)^+] \tag{22}$$

对于每一次的迭代,如果对于所有的 i 都有$\rho_i<\tau$,则将τ进行 0.9 的递减迭代;否则,直接终止寻找.该方法完成了从黑盒测试到白盒测试的转换,使得在不知道分类网络相关参数的条件下,依旧实现误分类的效果.

(c) UPSET

UPSET 是由 Sayantan Sarkar[36]提出的一种具有定向攻击目标和适用黑盒测试两个特点的对抗样本生成方法.UPSET 主要运用了一个对抗生成网络 R,在选择好目标类别 t 后构建 $r_t=R(t)$,可用公式(23)表示:

$$x'=U(x,t)=\max(\min(s\times R(t)+x,1),-1) \tag{23}$$

其中,U 代表的是 UPSET 网络,像素值都被标准化到[0,1]的范围内;s 是保证 $s\times R(t)$项在范围[0,1]的参数.训练过程如图 2 所示:在确定需要误分类的类别 t 后,让 t 经过一个对抗网络 R,得到像素结果,调试后得到 $s\times R(t)$项,与原有的图像结合后得到 $s\times R(t)+x$ 项,经过裁剪后得到 $U(x,t)$.图 2 中的 L_F 代表保真度的损失函数,以保证添加的扰动微小;而 L_C 代表误分类的损失函数;图 2 中的 $C_1,\ldots,C_m$ 是训练好的待定分类网络.

综上,整体的损失函数可以表示为公式(24):

$$L(x,x',t)=L_C(x',t)+L_F(x,x')=-\sum_{i=1}^{m}\log(C_i(x')[t])+\omega\|x'-x\|_k^k \tag{24}$$

(d) ANGRI[36]

ANGRI 和 UPSET 是一起被提出来的,与 UPSET 不同的是,ANGRI 所生成的扰动不具有通用性,也就是说,输出依赖于输入图像的属性,训练过程如图 3 所示.图中 $A(x,t)$表示的是 ANGRI 网络,与 UPSET 的区别就是将 A_t 和 A_x 连接以后得到 A_c,以完成之后的操作.

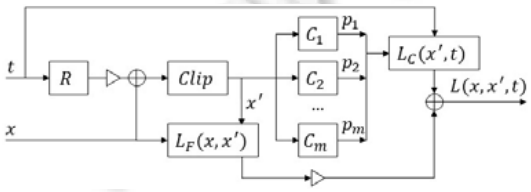


Fig.2 Training process chart for UPSET[35]

图 2 UPSET 训练过程图[35]

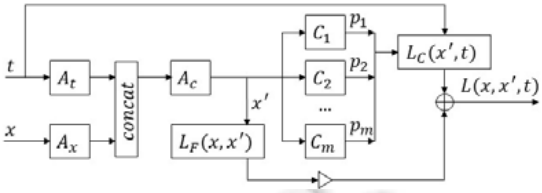


Fig.3 Training process chart for ANGRI[35]

图 3 ANGRI 训练过程图[35]

(e) Houdini

Houdini 方法是由 Yossi 等人[37]提出,针对于深度分类网络进行的一种对抗样本生成的方法.与一般方法不同的是,Houdini 将损失函数的计算改为公式(25):

$$\overline{l}_H(\theta,x,y)=P_{\gamma\sim N(0,1)}[g_\theta(x,y)-g_\theta(x,\hat{y})<\gamma]\cdot l(\hat{y},y) \tag{25}$$

其中,g_θ代表的是参数为θ的神经网络;而 l 为原有的损失函数,输出是得分最高的类别.Houdini 分为两个部分:一部分是随机极限,可影响模型预测的置信度;另一部分是任务损失,以完成最大化的处理.在经过反向传播之后,可以得到在经过 Houdini 方法处理的网络输出造成的拓展损失函数为公式(26):

$$\nabla_g[P_{\gamma\sim N(0,1)}[g_\theta(x,y)-g_\theta(x,\hat{y})<\gamma]\cdot l(\hat{y},y)]=\nabla_g\left[\frac{1}{\sqrt{2\pi}}\int_{\delta g(y,\hat{y})}^{\infty}\mathrm{e}^{-\frac{v^2}{2}}\mathrm{d}v\cdot l(\hat{y},y)\right] \tag{26}$$

该方法除了在图像分类的问题上取得效果之外,在语音识别网络也有很好的实验结果[38].

(3) 目标定向和目标非定向兼可

(a) ATNs

另一种目标针对性的白盒样本生成方法 ATNs[35]由 Shumeet 等人提出,可以在不同网络间转移,具有更强的

适应性.ATNs 实现了让对抗样本从定向攻击一个网络转移为成功攻击一系列网络的神经网络结构,理论上可以实现黑盒测试与白盒测试、目标定向与非定向.ATNs 可以总结为以下的模型,如公式(27):

$$g_{f,\theta}(x):x\in\chi\to x' \tag{27}$$

其中,θ是向量 g 的参数,f 为目标网络.为了找到合适的 $g_{f,\theta}(x)$,在训练过程中,需要实现公式(28)的优化:

$$\arg\min_{\theta}\sum_{x_i\in\chi}\beta L_{\chi}(g_{f,\theta}(x_i),x_i)+L_y(g_{f,\theta}(x_i),f(x_i)) \tag{28}$$

其中,L_{χ}是输入空间的损失函数,L_y 是输出空间中的损失函数,β是平衡 L_{χ}和 L_y 的参数.

(b) MI-FGSM

MI-FGSM 是由 Dong 等人[39]在借鉴 I-FGSM 和 ILCM 两种算法以后提出的基于定向目标的黑盒攻击方法.MI-FGSM 在 CAAD 攻防赛中也使用了该模型,取得了第一的成绩.与 I-FGSM 和 ILCM 相比,MI-FGSM 方法把梯度迭代部分用动量迭代来替代.基于梯度下降法的动量迭代方法可以用公式(29)表示:

$$x'_{t+1}=x'_t+\alpha\cdot\frac{g_{t+1}}{\|g_{t+1}\|_2} \tag{29}$$

其中,g_{t+1} 是根据梯度下降的方向进行更新;α被设定为ε/T,T 是整个迭代的总体数量.按照这个方法进行 l_∞范数和 l_2 范数的拓展,l_∞范数和 l_2 范数的拓展结果为公式(30):

$$l_\infty: x'_{t+1}=x'_t-\alpha\cdot sign(g_{t+1});l_2: x'_{t+1}=x'_t-\alpha\cdot\frac{g_{t+1}}{\|g_{t+1}\|_2} \tag{30}$$

(c) Curls & Whey[40]

Curls & Whey 在 MI-FGSM 上进行了改进,由 Shi 等人提出,是针对黑盒攻击设计的方法.根据在沿梯度上升方向单调地添加扰动所生成的迭代轨迹缺乏多样性和适应性的缺陷,以及容易添加过多扰动的问题,采用卷曲迭代和过滤扰动结合的解决算法.其中,卷曲迭代(curls iteration)以交叉熵的变化作为标准决定下一步是梯度上升或下降,Whey Optimization 利用对抗性扰动的鲁棒性,根据像素值将扰动分成若干组,对每一组的扰动进行滤除,再随机提取出对抗样本中的每个像素,删除多余扰动.

Curls Iteration 调整梯度上升或下降如公式(31)所示:

$$\begin{cases}x'_0=x,x'_1=Clip_{x,\varepsilon}\{x'_0-\alpha\cdot\nabla J_{sub}(x'_0)\}\\ g_{t+1}=\begin{cases}-\nabla J_{sub}(x'_t), & J(x'_t)<J(x'_{t-1})\\ \nabla J_{sub}(x'_t), & J(x'_t)\geqslant J(x'_{t-1})\end{cases}\\ x'_{t+1}=Clip_{x,\varepsilon}\{x'_t+\alpha\cdot g_{t+1}\}\end{cases} \tag{31}$$

其中,x 代表原图像,x'_t 为第 t 步迭代后的对抗样本,$\nabla J_{sub}(x'_t)$ 和 $J(x'_t)$ 代表在替代模型和目标模型上 x'_t 的交叉熵,g 为调整后的梯度.每次迭代更新时,选择所有对抗样本的平均梯度方向 $\bar{R}$ 为更新方向,如公式(32):

$$\bar{R}=\frac{1}{K}\sum_{i=1}^{K}x' \text{ s.t. } N(x)\neq N(x') \tag{32}$$

其中,N 为目标模型.再使用二分搜索的方法进行优化,如公式(33):

$$BS(L,R)=\left\{\begin{array}{c}L=x,R=x',\\ \begin{cases}BS\left(L,\dfrac{L+R}{2}\right), & N(x)\neq N((L+R)/2)\\ BS\left(\dfrac{L+R}{2},R\right), & N(x)=N((L+R)/2)\end{cases}\end{array}\right\} \tag{33}$$

Whey Optimization 采用了对抗样本鲁棒性的特点,首先将产生的对抗噪声分组,如公式(34)所示:

$$\left.\begin{array}{c}z_0=x'-x,\\ \rho_{t+1}^{whc}=\dfrac{\rho_t^{whc}}{2},\ \text{s.t. } \rho_t^{whc}=L(V(\rho_0),t)\end{array}\right\} \tag{34}$$

其中,ρ为添加的扰动;$L(V,t)$代表的是在像素集合 V 中第 t 大的绝对值,使用公式(35)进行扰动挤压,其中,*mask* 和

ρ大小一样,δ为每个像素的概率设定值:

$$\left.\begin{aligned}\rho_{t+1}&=\rho_t\cdot mask,\\ mask^{whc}&=\begin{cases}0, & random(\cdot)\leqslant\delta\\ 1, & \text{else}\end{cases}\end{aligned}\right\} \tag{35}$$

4.1.2 部分像素扰动的生成方法

(1) 肉眼不可见类

(a) JSMA

JSMA[41]由 Nicolas 提出,是一种针对于深度神经网络类型进行对抗样本生成的方法,利用前向导数来具体实现.前向导数的生成,使用的是训练好的网络中功能函数的 Jacobian 矩阵,如公式(36)所示:

$$\nabla F(X)=\frac{\partial F(X)}{\partial F}=\left[\frac{\partial F_j(X)}{\partial x_i}\right]_{i\in1,\ldots,M,\,j\in1,\ldots,N} \tag{36}$$

其中,F 为网络函数,x_i 为不同的维度.使用前向导数的方法来计算梯度和后向转播算法有一定的相似性,但前项导数直接使用的是网络的导数而非代价函数,同时,更多地依赖于输入的特征而非网络参数,从而可以在输出中获得更显著的结果.使用公式(37),递归地区分隐藏层:

$$\frac{\partial H_k(X)}{\partial x_i}=\left[\frac{\partial f_{k,p}(W_{k,p}\cdot H_{k-1}+b_{k,p})}{\partial x_i}\right]_{p\in1,\ldots,m_k} \tag{37}$$

其中,$H_k(X)$是隐藏层的输出向量,$f_{k,p}$是第 k 层中第 j 个神经元的激励函数.对每一个前向导数和输出标签进行显著性映射操作,得到的结果作为对抗样本调整的依据.其中,显著性映射如公式(38):

$$S(X,t)[i]=\begin{cases}0, & \dfrac{\partial F_t(X)}{\partial X_i}<0 \text{ or } \sum_{j\neq t}\dfrac{\partial F_t(X)}{\partial X_i}>0\\ \left(\dfrac{\partial F_t(X)}{\partial X_i}\right)\left|\sum_{j\neq t}\dfrac{\partial F_j(X)}{\partial X_i}\right|, & \text{others}\end{cases} \tag{38}$$

其中,i 是输入特征.映射要求 t 所带来的 $\dfrac{\partial F_t(X)}{\partial X_i}$ 为正数,其他点产生负面影响时,不符合相关显著性条件的点设为 0.最后,取使得所有的显著值中最大的输入特征来调整样本,与原来的值相减后得到干扰值.

(b) ONE-PIXEL

由 Su 等人提出的 ONE-PIXEL[42]是通过只改变一个像素来生成对抗样本的方法.这是一种基于目标非定向性的黑盒方法,同时也实现了只需改变少量像素就导致误分类.ONE-PIXEL 方法可用公式(39)表示出来:

$$\max f_{adv}(x+e(x))\quad \text{s.t.}\quad \|e(x)\|_0\leqslant d \tag{39}$$

与其他方法改变整张图片的全部或者部分像素不同,ONE-PIXEL 只改变一个像素,故将式中的 d 设定为 1.改变一个像素造成扰动实际上是沿着平行于 n 维中的一个轴方向进行数据点的干扰,每次对扰动包括 5 个元素:x 坐标、y 坐标和扰动的 RGB 值.对每个像素进行如公式(40)的迭代操作:

$$x_i(g+1)=x_{r_1}(g)+F(x_{r_2}(g)+x_{r_3}(g)),r_1\neq r_2\neq r_3 \tag{40}$$

其中,x_i是候选解决方案的元素;r_1,r_2,r_3是随机的数字;F是尺度参数,设定为0.5;g是当前迭代的索引数.在每次迭代中,候选解决方案的结果如果优于父结果,则候选结果进入下一次迭代;如果没有优于父结果,则父结果进入迭代,以此选出最好的一个像素扰动样本结果.

(2) 肉眼可见类

(a) Adversarial Patch[43]

Adversarial Patch 是由 Brown 提出的一种添加局部像素扰动的方法,因为只需要改变图像中的 patch,所以可以达到灵活添加和局部扰动的效果.该方法通过 *mask* 来调整 patch 的大小和形状,随机让 patch 在图像上进行平移、缩放和旋转;与此同时,使用梯度下降的方法进行优化.定义一个 patch 选择器 $A(p,x,l,t)$,p 为相应的 patch,l 为 patch 的位置,x 为图像,t 为转换操作.先使用选出的 p 转换相应的对抗结果,再将此结果应用于相应的位置上.

在 patch 训练时的优化函数如公式(41)所示:

$$\hat{p} = \arg\max_{p} E_{x\sim X,t\sim T,l\sim L}[\log \Pr(\hat{y} \mid A(p,x,l,t))] \tag{41}$$

其中,X 为图像训练集,T 为 patch 的转换集,L 为 patch 的位置集.这种方法对于 patch 的选择具有很高的灵活性,从而使得这种添加扰动的方法有普遍性.

(b) LaVAN[44]

LaVAN 是由 Karmon 提出的另一种部分像素添加扰动的方法,该方法在设定噪声可见的情况下,在图像的局部位置添加扰动,以产生较好的对抗样本.首先,设定置信阈参数$\mathcal{K}$,$mask$ 用于调整大小 m、图像 x、模型 f,从而计算出初始扰动,如公式(42)所示,⊙为像素乘积:

$$(1-m)\odot x+m\odot p \tag{42}$$

更新扰动,如公式(43)所示,其中,$\partial f(x)|_{\hat{y}}$ 为目标类别输出,$\partial f(x)|y$ 为原图像输出:

$$-\varepsilon\cdot\left(\frac{\partial f(x)|_{\hat{y}}}{\partial x}-\frac{\partial f(x)|_{y}}{\partial x}\right) \tag{43}$$

(c) PS-GAN[45]

PS-GAN 是对于 Adversarial Patch 的改进,针对攻击力的增强和逼真程度的提高,提出的一种感知敏感生成对抗网络.为提高视觉逼真度,PS-GAN 将 patch 的生成转化为一个 patch 到另一个 patch 的翻译,从而输出与被攻击图像具有高度感知相关性的类似对抗 patch.为增强对抗样本的攻击能力,在对抗样本的生成中引入 attention 机制,预测出合适的攻击区域作为 patch,进而产生更真实、更有攻击性的对抗样本.PS-GAN 生成的对抗样本如公式(44)所示:

$$x'=x+_{M(x)}G(\rho) \tag{44}$$

其中,$G(\delta)$是更有攻击性的 patch,$M(x)$为 patch 在图像中的位置.为产生更逼真的结果,loss 的设置如公式(45):

$$L_{GAN}(G,D)=E_x[\log D(\rho,x)]+E_{x,z}[\log(1-D(\delta,x+_{M(x)}G(\rho)))] \tag{45}$$

为提高 patch 对输入图像的上下文相关性,使用 GAN 进行 loss 的优化,如公式(46)所示:

$$L_{patch}(\delta)=E_x\|G(\rho)-\rho\|_2 \tag{46}$$

其中,取 l_2 范数进行 L_{patch} 的确立.再引入 attention 机制,对抗 loss 可表示为公式(47):

$$L_{adv}(G,F)=E_{x,\rho}[\log P_F(x')] \tag{47}$$

最终的生成公式如公式(48),结合以上 3 个部分的 loss 得到:

$$\min_G \max_D L_{GAN}+\lambda L_{patch}+\gamma L_{adv} \tag{48}$$

(d) Printable Adversarial Patches[46]

Printable Adversarial Patches 是一种肉眼明显可见的对抗样本生成方法,着力于在实际应用方面.该方法生成一个局部可打印出来的对抗样本,使得混入对抗因素后实现欺骗检测网络.例如摄像头下原本可以检测出来的人,在手拿包含此类可打印的图案后无法被网络检测,实际应用中有很大的意义和警示性.为了达到可以打印的效果,引入了一个相关的打印损失 L_{nps},以实现被打印图取代的可能性,如公式(49)所示:

$$L_{nps}=\sum_{p_{patch}\in p}\min_{c_p\in C}|p_{patch}-c_p| \tag{49}$$

其中,p_{patch} 是 patch 中的像素,c_p 是可以打印出来的颜色集合 C 的颜色值.损失函数的第 2 个组成部分是有生成对抗样本的 patch 与周围像素的平滑度,如公式(50)所示:

$$L_{tv}=\sum_{i,j}\sqrt{((p_{i,j}-p_{i+1,j})^2+(p_{i,j}-p_{i,j+1})^2)} \tag{50}$$

如果周围的像素相似,则 L_{tv} 值偏低;如果周围的像素差距较大,则 L_{tv} 值偏高.损失函数的第 3 个组成部分是类误差 L_{obj},网络的训练目标是隐藏图像中的人,故需要最小化测出的类误差.将以上 3 部分结合后得到最终的误差函数,如公式(51),以此作为训练优化标准,从而达到完成生成可打印局部对抗样本的效果:

$$L=\alpha L_{nps}+\beta L_{tv}+\gamma L_{obj} \tag{51}$$

4.2 常用数据集

MNIST[47]:在几乎所有提出方法的文献中,MNIST 都被用来作为结果对比.MNIST 数据集是一个手写数字的数据集,共计 7 万组图片数据.数据集由 4 部分组成.(1) 训练集图片:47MB,60 000 张;(2) 训练集图片标签:60KB,60 000 个;(3) 测试集图片:7.8MB,10 000 张;(4) 测试集图片标签:10KB,10 000 个.

MNIST 数据集来自美国国家标准与技术研究所,由 250 个人的手写数字组成,样本图片以字节的形式存储.

ImageNet dataset:ImageNet dataset[48]是由美国斯坦福的李飞飞模拟人类视觉识别系统建立的数据库,是目前世界上图像识别中最大的数据库,图像内容是具体的物体,目前包含 1 400 多万的样本个数,涵盖 2 万多个类别(超过百万有类别和位置两方面的标注),关于这个数据集的大赛在视觉识别方面也有很大的关注度.

CIFAR-10:CIFAR-10 是带标签的图像数据集,共有 60 000 张彩色图像,像素大小为 32×32.60 000 张图像分为 10 个类,每个类 6 000 张图,设定其中 50 000 张为训练数据,构成 5 个训练批次,每一批 10 000 张图;设定剩下的 10 000 张为测试数据.数据集以字典结构的形式进行存储,分为数据和标签两个部分.数据部分中,图像以 numpy 数组的形式保存,每一行储存 32×32 大小的彩色图像,其中,每 1 024 个数字代表颜色通道.标签部分中,是一个范围在 0~9 的含有 10 000 个数的列表,第 *i* 个数就是第 *i* 个图像的类标签.

5 实验结果对比

将在第 4 节中所介绍的几种方法在同一个数据集(MNIST)进行实验,通过对于实验结果的对比和鲁棒性的测试,可证明不同方法的优缺点.实验设置训练集为原有设置的前 60 000 张图片及其标签,设置测试集为原有设置的 10 000 张图片及其标签,设置 *epoch* 值为 6,批处理数量为 128,学习率定为 0.001.

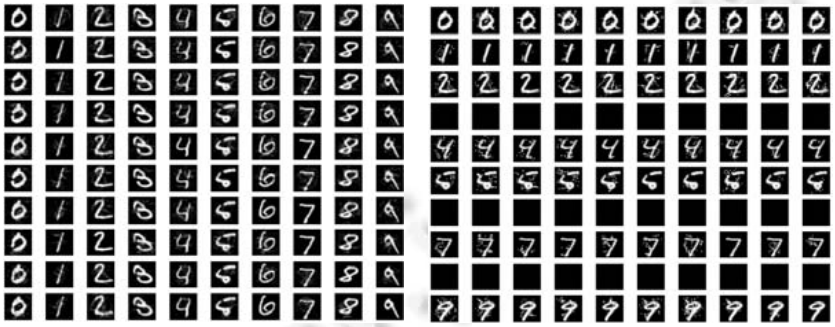
首先对样本数据不作改变训练分类器,以其分类准确率作为标准与几种不同的生成方法进行比较,若生成的对抗样本经过训练好的分类器的准确率越低,说明此种生成方法的效果越好;对抗训练后,再经过分类器的准确率越低,此种生成方法的鲁棒性越好.正常分类器得到的分类训练结果见表 2,随着 *epoch* 的训练,得到的分类准确率越好.虽然 *epoch* 设定的值为 6,但准确率基本在 99%以上,已经达到了后续实验的要求.

Table 2 Normal classification training results of classifiers

表 2 分类器正常分类训练结果表

	*Epoch*1	*Epoch*2	*Epoch*3	*Epoch*4	*Epoch*6	*Epoch*7
准确率	0.9882	0.9904	0.9918	0.9920	0.9922	0.9929

以最为基础的 L-BFGS 为例,加入微小扰动之后所产生的对抗样本图例如图 4 所示.



(a) L-BFGS (b) JSMA

Fig.4 Examples of generating adversarial examples

图 4 生成对抗样本实例

可以看出,加入扰动之后的图像与原图的肉眼可见差距并不大.按照前文所述的方法进行逐个类别的对抗样本生成,方法的基本属性及其具体得到的实验结果见表 3.其中,universal 表示对抗样本生成方法是否针对整个数据集.

Table 3 Attribute and performance table of representative methods (l_∞)

表 3 代表方法属性及效果表(l_∞)

方法	Black/White box	If targeted	If universal	是否只改变少量像素点	Attack frequency	准确率
L-BFGS	W	Y	N	N	Iterative	0.930 0
FGSM	W	Y	N	N	One shot	0.510 0
I-FGSM	W	N	N	N	Iterative	0.005 8
JSMA	W	Y	N	Y	Iterative	0.177 8
ONE-PIXEL	B	N	N	Y	Iterative	0.732 2
C&W	W	Y	N	N	Iterative	0.990 0
DEEPFOOL	W	N	N	N	Iterative	0.100 0
MI-FGSM	B	Y	N	N	Iterative	0.005 6
UPSET	B	Y	Y	N	Iterative	0.345 0

从表 3 中可以看出:在之前所选择的几种比较有代表性的对抗样本生成方法中,适应于黑盒测试的有 ONE-PIXEL、MI-FGSM 和 UPSET,而适应于白盒测试的有 L-BFGS、FGSM、I-FGSM、JSMA 和 DeepFool,从白盒到黑盒的转变可以从 I-FGSM 转变为 MI-FGSM 看出.而在一般的分类标准中,目标定向性和目标非定向性也是很重要的一个依据,可用来衡量所生成的样本的意义.其中,生成的目标具有定向性的有 L-BFGS、FGSM、JSMA、C&W、MI-FGSM 和 UPSET,不具有定向性的有 I-FGSM、ONE-PIXEL 和 DeepFool.而 Madry 等人提出了一种将目标非定向转化为目标定向的算法,即在生成的时候选择原有标签中最少的一类作为定向的标签,从而生成具有目标定向性的对抗样本生成方法.

根据实验发现:最初没有经过迭代方法而直接生成对抗样本的 FGSM 最后得到的对抗效果较差一些,为 0.51,但较之于原来的 99.3%以上,已经有了很好的对抗特性.只改变图像中的少量像素为代表的 JSMA 和 ONE-PIXEL 取得了很不错的效果,分别为 0.177 8 和 0.732 2.ONE-PIXEL 的结果由于只改变了一个像素,导致对抗性的牺牲,但两种方法,尤其是 JSMA 所达到的对抗性已经拥有了迷惑正常分类器的能力.而在对于全图像像素都进行扰动改变的方法中,除却两种非迭代的理论方法,I-FGSM 获得 0.005 8 的结果,C&W 获得 0.990 0 的结果,DeepFool 获得 0.100 0 的结果,MI-FGSM 获得 0.005 6 的结果.其中以 MI-FGSM 的初始效果最好,说明通过动量来对标准的梯度下降方法进行改良是一个很好的改善思想.

在进行对抗训练之后,再进行一次对抗样本的输入,从而可以测试出对抗样本的鲁棒性,得到的实验结果见表 4.可以看出,经过迭代后的 I-FGSM 方法相比于非迭代方法有较强的迭代性.

Table 4 Comparison of accuracy after adversarial training

表 4 对抗训练后的准确率对比

方法	对抗训练之前准确率	对抗训练之后准确率
FGSM	0.121 8	0.946 5
FGSM(定向处理后)	0.005 4	0.926 3
I-FGSM	0.005 8	0.906 6
MI-FGSM	0.005 6	0.895 0

6 面临挑战与前景预测

(1) 可迁移性和鲁棒性

可以发现:除了在最初的 L-BFGS 和 FGSM 外,其余的文献所采用的都是迭代方法,以保证更高的正确率.而由于 L-BFGS 和 FGSM 两篇文献本身更着重于介绍对抗样本的发现和对抗样本产生的基本原理,所以在对抗样本产生的质量方面有所欠缺.在之后的样本方法的生成中,应该会继续使用迭代的方式.

在适用黑盒测试和白盒测试的考量上,涉及黑盒测试和白盒测试方法的文献基本上有相同的数量,在这两方面的研究都比较热门.但是按照实际情况和意义上来看,适用于黑盒测试的样本生成方法有更好的前景,更适用于实际需要的对抗需求.

在近期的 NIPS 2017 和 CAAD CTF 大赛中,都把对抗样本的生成分成目标定向和目标非定向、黑盒测试和白盒测试两大类,说明目前比较关注目的和条件这两个方面,对于实际应用有比较大的意义.同时,在大赛中,

参赛的各个团队对于对抗样本的可迁移性和鲁棒性都有很大的关注.可迁移性[49]一般情况下指的是一种方法所生成的对抗样本在不同的分类模型下保持一定对抗性的能力.可迁移性影响着生成的对抗样本的适用范围,可以作为今后衡量对抗样本的重要指标.以现有的文献可以发现,使用了生成对抗网络(GAN)[50]的生成方法具有较好的可迁移性,故可总结为加入对抗网络的生成方法将有更大的应用价值.

(2) 攻防特性

对于对抗样本这个概念,可以从攻和防两个方面作进一步的研究.

- 在攻击方面,生成的对抗样本是否具有足够的欺骗性,是生成对抗样本的基本问题,而这个欺骗性可以分为对分类器的欺骗性和对人眼的欺骗性:对于分类器的欺骗性,表现在生成样本的对抗性上,目前主要趋势是通过迭代和梯度下降的方法加以完善,同时也有动量等方式;而对于人眼的欺骗性,主要体现在加入的扰动所具有的微小性[51],这主要可以在部分像素添加扰动这类操作中体现,使用尽量少的像素进行扰动改变来生成对抗样本,但是只改变少量像素,势必存在减少对抗性和增加生成时间的代价.综上,分类器欺骗性和人眼欺骗性这两方面的综合考量是攻方的关键,选取适量的评价标准尤为重要;
- 在防御方面,对抗样本在隐私保护中亦或可以起到比较重要的作用.对于需要公开的图像等资料进行添加扰动的操作,以做出相应的“对抗样本”,可以实现对于真实数据的隐藏目的.

(3) 模型本身的扰动添加

对抗样本生成这种样本添加扰动进行混淆干扰的思路,可以应用于模型本身.众所周知,分类器的分类准确性就根本而言是模型和样本两方面决定的.样本所带来的混淆效果可以由对抗样本生成,模型也可以通过添加扰动的方式进行对于分类准确率的改变.Liao 等人[52]提出了对于模型加扰动的相关算法,通过产生扰动植入和训练前植入或训练更新内植入的方式进行对抗模型的形成,最终获得了 90%的攻击成功率.

(4) 在 OCT 上的应用价值

OCT[53]是光学相干断层扫描技术(optical coherence tomography)的缩写,被广泛应用于指纹识别技术中.假指纹的攻击问题是 OCT 技术的重要挑战,分为 3 个等级:(a) 传感器层面(根据真实指纹制作假指纹);(b) 数据库层面(根据指纹数据制作假指纹);(c) 指纹识别算法层面(无须指纹数据先验制作假指纹).

对于以上 3 点,对抗样本在攻击和防御两个角度都有比较好的应用价值:在攻的方面,着重于假指纹的制作,可以运用目标非定向性的对抗样本生成方法,在原本的图像中加入扰动,使之被分类器误分类为其他有意义的类别,即可以把一张标签为 a 的指纹识别为有意义的 b;在防的方面,主要涉及数据库的保护,将所储存的图像加入相关扰动,使得即使数据库发生泄露,也能达到保护相应标签的目的.

7 结 语

本文主要介绍了生成对抗样本的概念,就前传、起源和发展这 3 个部分进行论述.从深度学习在分类问题中的显著应用,发掘出生成对抗样本的潜在价值,并以 L-BFGS 和 FGSM 为代表进行了理论方面的解释.而在发展部分,本文将各类对抗样本生成方法分类为全像素添加扰动和部分像素添加扰动两类,从目标定向和目标非定向、黑盒测试和白盒测试、肉眼可见和不可见将全像素添加扰动进行二次分类,对各方法的对抗性进行了分析和对比.而对于部分像素添加扰动这类,更多地注重像素改变量的减少和对抗性的保持两者的综合效果.同时,使用相同的数据集(mnist)进行各类实验,进行具体结果的对比,以反映相关方法的优劣.根据实验结果,总结了生成对抗样本所面临的可转移性和鲁棒性的挑战,在攻防、模型本身扰动和 OCT 应用方面分析其发展前景.

References:

[1] Szegedy C, Zaremba W, Sutskever I, *et al*. Intriguing properties of neural networks. Computer Science, 2013. https://arxiv.org/abs/1312.6199

[2] Nguyen A, Yosinski J, Clune J. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR). Ohio: IEEE, 2014. 427–436. [doi: 10.1109/CVPR.2015.7298640]

[3] Tabacof P, Valle E. Exploring the space of adversarial images. 2016. 426–433. [doi: 10.1109/IJCNN.2016.7727230]

[4] Goodfellow IJ, Shlens J, Szegedy C. Explaining and harnessing adversarial examples. Computer Science, 2014. https://arxiv.org/abs/1412.6572

[5] Akhtar N, Mian A. Threat of adversarial attacks on deep learning in computer vision: A survey. IEEE Access, 2018,6: 14410−14430. [doi: 10.1109/ACCESS.2018.2807385]

[6] Zhao C. Analysis of black box testing and white box testing. Silicon Valley, 2010,(11):39 (in Chinese).

[7] Yan Z, Guo Y, Zhang C. Deep defense: Training DNNs with improved adversarial robustness. arXiv:1803.00404, 2018.

[8] Hu WW, Tan Y. Generating adversarial malware examples for black-box attacks based on GAN. arXiv:1702.05983, 2017.

[9] He Q, Li N, Luo WJ, *et al*. A survey of machine learning algorithms for big data. Pattern Recognition and Artificial Intelligence. 2014,(4) (in Chinese with English abstract).

[10] Zhong SH, Liu Y. Bilinear deep learning for image classification. In: Proc. of the 19th ACM Int'l Conf. on Multimedea. Arizona: ACM Press, 2011. 343−352. [doi: 10.1145/2072298.2072505]

[11] Schmidhuber J. Deep learning in neural networks: An overview. Neural Networks, 2014,61:85−117. [doi: 10.1016/j.neunet.2014.09.003]

[12] Lecun Y, Bengio Y, Hinton G. Deep learning. Nature, 2015,521(7553):436. [doi: 10.1038/nature14539]

[13] Alom MZ, Taha TM, Yakopcic C, *et al*. The history began from AlexNet: A comprehensive survey on deep learning approaches. arXiv:1803.01164, 2018.

[14] Mcclelland J. Information Processing in Dynamical Systems: Foundations of Harmony Theory. MIT Press, 1986. 194−281.

[15] Hinton GE, Zemel RS. Autoencoders, minimum description length and Helmholtz free energy. In: Proc. of the Int'l Conf. on Neural Information Processing Systems. Morgan Kaufmann Publishers Inc., 1993. 3−10.

[16] Olshausen BA, Field DJ. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. Nature, 1996,381(6583):607−609. [doi: 10.1038/381607a0]

[17] Dong X, Guan Y. Dependency parsing model based on supervised learning: A survey. Intelligent Computer & Applications, 2013,(2).

[18] Liu D, Li S, Cao ZD. State-of-the-art on deep learning and its application in image object classification and detection. Computer Science, 2016,43(12):13−23.

[19] Hinton GE. Learning distributed representations of concepts. In: Proc. of the 8th Conf. of the Cognitive Science Society. 1989.

[20] Zhou FY, Jin LP, Dong J. A review of convolutional neural networks. Journal of Computer Science, 2017,40(6):1229−1251.

[21] Kılıça K, Boyacıa İH, Köksela H, *et al*. A classification system for beans using computer vision system and artificial neural networks. Journal of Food Engineering, 2007,78(3):897−904.

[22] Socher R, Huval B, Ng AY, *et a1*. Semantic compositionality through recursive matrix-vector spaces. In: Proc. of the Advances in NeuralInformation Processing Systems (NIPS). Tahoe: IEEE, 2012. 665−673.

[23] Karpathy A, Toderici G, Shetty S, *et al*. Large-scale video classification with convolutional neural networks. In: Proc. of the IEEE Conf. on Computer Vision andPattern Recognition (CVPR). Ohio: IEEE, 2014. 1725−1732. [doi: 10.1109/CVPR.2014.223]

[24] Sanchezriera J, Hsiao YS, Lim T, *et al*. A robust tracking algorithm for 3D hand gesture with rapid hand motion through deep learning. In: Proc. of the IEEE Conf. on Multimedia and Expo Workshops (ICMEW). Zhejiang: IEEE, 2014. 1−6.

[25] Sun N, Han G, Du K, *et al*. Person/Vehicle classification based on deep belief networks. In: Proc. of the IEEE Int'l Conf. on Natural Computation (ICNC). Xiamen: IEEE, 2014. 1928−1934. [doi: 10.1109/ICNC.2014.6975819]

[26] Yu B, Lane I. Multi-task deep learning for image understanding. In: Proc. of the Soft Computing and Pattern Recognition. IEEE, 2015. 37−42. [doi: 10.1109/SOCPAR.2014.7007978]

[27] Sawada Y, Kozuka K. Transfer learning method using multi-prediction deep Boltzmann machines for a small scale dataset. In: Proc. of the IAPR Int'l Conf. on Machine Vision Applications. IEEE, 2015. [doi: 10.1109/MVA.2015.7153145]

[28] Zhou ZH. Machine Learning. Beijing: Tsinghua University Press, 2016 (in Chinese).

[29] Dong G, Gao J, Du R, *et al*. Robustness of network of networks under targeted attack. Physical Review E Statistical Nonlinear & Soft Matter Physics, 2013,87(5):052804. [doi: 10.1103/PhysRevE.87.052804]

[30] Avila G, Withers T, Holwell G. Retrospective risk assessment reveals likelihood of potential non-target attack and parasitism by Cotesia urabae (Hymenoptera: Braconidae): A comparison between laboratory and field-cage testing results. Biological Control, 2016,103:108−118. [doi: 10.1016/j.biocontrol.2016.08.008]

[31] Miyato T, Maeda SI, Ishii S, *et al*. Virtual adversarial training: A regularization method for supervised and semi-supervised learning. IEEE Trans. on Pattern Analysis & Machine Intelligence, 2017,PP(99):1. [doi: 10.1109/TPAMI.2018.2858821]

[32] Kurakin A, Goodfellow I, Bengio S. Adversarial examples in the physical world. arXiv:1607.02533, 2016.

[33] Moosavidezfooli SM, Fawzi A, Frossard P. DeepFool: A simple and accurate method to fool deep neural networks. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR). Las Vegas: IEEE, 2016. 2574−2582. [doi: 10.1109/CVPR.2016.282]

[34] Carlini N, Wagner D. Towards evaluating the robustness of neural networks. In: Proc. of the 2017 IEEE Symp. on Security and Privacy (SP). 2017. [doi: 10.1109/SP.2017.49]

[35] Baluja S, Fischer I. Adversarial transformation networks: Learning to generate adversarial examples. arXiv:1703.09387v1, 2017.

[36] Sarkar S, Bansal A, Mahbub U, *et al*. UPSET and ANGRI: Breaking high performance image classifiers. arXiv:1707.01159, 2017.

[37] Cisse M, Adi Y, Neverova N, *et al*. Houdini: Fooling deep structured prediction models. arXiv:1707.05373, 2017.

[38] Amodei D, Anubhai R, Battenberg E, Case C, Casper J, Catanzaro B, Chen J, Chrzanowski M, Coates A, Diamos G. Deepspeech 2: End-to-end speech recognition in English and mandarin. arXiv Preprint, 2015.

[39] Dong Y, Liao F, Pang T, *et al*. Boosting adversarial attacks with momentum. arXiv:1710.06081, 2017.

[40] Shi Y, Wang S, Han Y. Curls & Whey: Boosting black-box adversarial attacks. arXiv:1904.01160, 2019.

[41] Papernot N, Mcdaniel P, Jha S, *et al*. The limitations of deep learning in adversarial settings. In: Proc. of the IEEE European Symp. on Security and Privacy. IEEE, 2016. 372−387. [doi: 10.1109/EuroSP.2016.36]

[42] Su J, Vargas DV, Kouichi S. One pixel attack for fooling deep neural networks. arXiv:1710.08864, 2017.

[43] Brown TB, Mané D, Roy A, *et al*. Adversarial patch. arXiv:1712.09665, 2017.

[44] Karmon D, Zoran D, Goldberg Y. LaVAN: Localized and visible adversarial noise. arXiv:1801.02608, 2018.

[45] Liu AS, Liu XL, Fan JX, *et al*. Perceptual-sensitive GAN for Generating Adversarial Patches. 2019. [doi: doi:10.1609/aaai.v33i01.33011028]

[46] Thys S, Van Ranst W, Goedemé T. Fooling automated surveillance cameras: Adversarial patches to attack person detection. arXiv:1904.08653v1, 2019.

[47] Deng L. The MNIST database of handwritten digit images for machine learning research [Best of the Web]. IEEE Signal Processing Magazine, 2012,29(6):141−142. [doi: 10.1109/msp.2012.2211477]

[48] Deng J, Dong W, Socher R, *et al*. ImageNet: A large-scale hierarchical image database. In: Proc. of the IEEE Conf. on Computer Vision andPattern Recognition (CVPR). Florida: IEEE, 2009. 248−255. [doi: 10.1109/CVPR.2009.5206848]

[49] Papernot N, Mcdaniel P, Goodfellow I. Transferability in machine learning: From phenomena to black-box attacks using adversarial samples. arXiv:1605.07277, 2016.

[50] Goodfellow I. NIPS 2016 Tutorial: Generative adversarial networks. arXiv:1701.00160, 2016.

[51] Krizhevsky A. Learning multiple layers of features from tiny image. Technical Report, University of Toronto, 2009.

[52] Liao C, Zhong H, Squicciarini A, *et al*. Backdoor embedding in convolutional neural network models via invisible perturbation. arXiv:1808.10307, 2018.

[53] Hee MR, Izatt JA, Swanson EA, *et al*. Optical coherence tomography of the human retina. Arch Ophthalmol, 1995,113(3):325−332. [doi: 10.1001/archopht.1995.01100030081025]

附中文参考文献:

[6] 赵宸.浅析黑盒测试与白盒测试.硅谷,2010,(11):39.

[9] 何清,李宁,罗文娟,等.大数据下的机器学习算法综述.模式识别与人工智能,2014,(4).

潘文雯(1995－),女,浙江杭州人,硕士,主要研究领域为人工智能.

宋明黎(1976－),男,博士,教授,博士生导师,CCF 专业会员,主要研究领域为计算机视觉,图像增强,模式识别,人工智能.

王新宇(1979－),男,博士,教授,博士生导师,CCF 专业会员,主要研究领域为软件工程.

陈纯(1955－),男,博士,教授,博士生导师,CCF 会士,主要研究领域为计算机软件与理论,大数据实时智能处理技术.