文章编号:1000-5862(2021)03-0285-07

二分类判别网络的对抗样本检测

曾利宏,张 巍*,滕少华
(广东工业大学计算机学院,广东 广州 510006)

摘要:在原始图像数据集中,添加特殊的细微扰动能形成对抗样本,经这类样本攻击的深度神经网络等模型可能以高置信度给出错误输出,然而当前大部分检测对抗样本的方法有许多前提条件,限制了其检测能力. 针对这一问题,该文提出一个二分类判别网络模型,通过多层卷积神经网络来提取样本数据的主要特征;应用特殊的判别目标函数,结合不同程度的噪声数据来训练并优化网络模型,以提高模型检测对抗样本的能力;模型采用端到端的方式,可直接部署到目标模型的源样本中来检测对抗样本的存在,亦可进行大规模应用. 实验结果表明:该模型的检测率优于其他相关模型.

关键词:二分类判别网络;深度神经网络;对抗样本;检测
中图分类号:TP 311 文献标志码:A DOI:10.16357/j.cnki.issn1000-5862.2021.03.10

0 引言

深度神经网络(Deep Neural Network,DNN)在许多计算机领域的应用中取得了巨大成功,特别是在图像分类领域中表现异常出色. 但 C. Szegedy 等[1]提出,对原始输入样本添加微量的扰动,会误导已经训练好的神经网络,使目标模型分类错误. 经过一些特定算法[2-3]攻击所生成的对抗样本[4]与原始输入几乎没有差异,人类无法察觉,但目标模型却以 99.99% 的置信度识别错误[5]. 在某些安全性至关重要的领域中[6-7],受对抗样本攻击后可能导致灾难性后果,例如配备了神经网络识别系统的行进自动驾驶汽车,若受到恶意制作出来的对抗样本攻击,则汽车可能失控而产生灾难性后果[8].

本文提出了一个新颖的检测对抗样本方法. 所提出的方法是基于如下假设:人为恶意地对原始输入添加对抗扰动,这个扰动被限制在一个较小的范围内,不易被人眼观察识别. 因此,扰动信息应远小于图像原始信息. 将这些扰动视为一种故意为之的人工噪声,为此提出并设计了一个二分类的判别网络来检测对抗样本. 在训练时,人为地添加随机噪声到原始输入中,然后将噪声数据和原始数据输入判别网络进行训练,应用特别设计的"判别目标函数"进行判别. 该目标函数对正常样本将反馈一个较高的评分值,而对噪声样本将反馈一个较低的评分值(评分值限制在 [0,1] 范围内). 训练完成后,判别网络对新输入的图像自动输出一个评分值. 为此,需要设置一个合适的阈值,大于该阈值的输入图像为正常样本,低于该阈值的输入图像为对抗样本.

1 相关工作

国内外已开展了许多关于对抗样本检测方面的研究,众多学者提出了不同类型的检测方法.

1.1 训练子网络

J. Metzen 等[9]提出了一种类似于对抗训练的检测方法,他们用一个小的"检测器"子网络来扩充深度神经网络,子网络用于区分真实数据和恶意数据. 这种检测方案需要修改已经训练好的目标模型,费时费力. D. Hendrycks 等[10]设计了一个小型"检测器"网络,通过使用大量的对抗样本来训练这个检测器识别未知的恶意样本. 这个检测方法直接或间接地使用了对抗样本的先验信息,导致训练成本

收稿日期:2020-05-17
基金项目:广东省重点领域研发计划(2020B010166006),国家自然科学基金(61972102),广东省教育厅课题(粤教高[2018]179 号,粤教高函[2018]1 号)和广州市科技计划(201903010107,201802030011,201802010026,201802010042,201604046017)资助项目.
通信作者:张 巍(1964—),女,江西南昌人,教授,主要从事大数据、数据挖掘和协同计算方面的研究. E-mail:weizhang@gdut.edu.cn

激增,检测模型的泛化性能也会大大降低.

1.2 集成检测器

Li Xin 等[11]构建了一个级联分类器,并将分类器与目标模型集成,通过检查原始模型的内部卷积层输出,从而确定一个输入样本是否是恶意的. Meng Dongyu 等[8]使用搭建多个外部检测器将输入图像分类为敌对图像或正常图像. 在训练时,框架的目标是学习正常图像的流形,在测试阶段,对远离正常图像流形的视为对抗样本并将其剔除. 但是这种检测技术也可以被较大的攻击扰动击败.

1.3 输出不一致检测

Xu Weilin 等[12]提出了一种特征压缩(Feature Squeezing)的方法进行对抗样本的检测. 特征压缩是将输入图像的每个像素值从 8 个比特(一般为 8 个比特)表示减小到更少比特,甚至于用 1 个比特来表示,然后结合空间过滤器平滑对应的图像,通过测量输入图像和被压缩或平滑处理图像的预测输出不一致来识别对抗样本; 实验表明: 在 MNIST 模型中检测对抗样本的存在时,该方法具有较高的性能. Liang Bin 等[13]提出先使用输入图像的熵作为阈值,然后结合标量量化和空间平滑滤波方法对输入进行相应的处理,最后根据目标模型不同的输出判断输入是否是对抗性的. 这些方法不需要预先训练检测模型,也不需要对抗样本的任何先验知识,但是这些方法比较难以确定一个准确的阈值,从而导致检测效果不佳或对正常样本的误检率过高.

本文所提出的方法,与第 1 类检测方法相比,无须修改目标网络; 并且克服了第 2 类方法需与目标模型集成这个困难,可以直接进行对抗样本的检测. 通过使用不同的噪声数据训练判别模型,并且模型的训练不需要对抗样本的任何先验知识,与第 3 类检测方法相比,能以更低的成本获取检测阈值.

2 二分类判别网络模型

二分类判别网络(Binary Discrimination Network)模型主要包括 2 个组成部分,其基础部分是设计一个二分类的判别网络 D. 在判别网络 D 上,使用原始数据和带有随机噪声的数据共同来训练网络模型,通过设置的判别目标函数持续对网络进行优化训练,提高判别网络的检测能力.

2.1 判别网络

判别网络需要完成图像的二分类任务,即判断输入图像是否属于对抗性图像. A. Krizhevsky 等[14]提出深度卷积神经网络和全连接神经网络对图像分类任务有显著效果. 因此,本文设计的判别网络主要基于卷积神经网络和全连接神经网络.

彩色图像与灰度图像通道数不一致,因此本文对不同通道的图像数据集分别设计了不同的网络结构. 相应的具体网络结构如表 1 所示.

表 1 判别网络结构

Layer#	网络结构(3 通道)	网络结构(单通道)
1	Conv2D(3,8,4,2,1), LeakyReLU(0.2)	Linear(784,1024), LeakyReLU(0.2), Dropout(0.3)
2	Conv2D(8,16,4,2,1), LeakyReLU(0.2)	Linear(1024,512), LeakyReLU(0.2), Dropout(0.3)
3	Conv2D(16,32,4,2,1), LeakyReLU(0.2)	Linear(512,256), LeakyReLU(0.2), Dropout(0.3)
4	Conv2D(32,64,4,2,1), Sigmoid()	Linear(256,1), Sigmoid()
5	Linear(256,32), LeakyReLU(True)	
6	Linear(32,1), Sigmoid()	

对于 3 通道的彩色图像,判别网络模型由 4 层卷积层 Conv2D 和 2 层全连接层 Linear 组成. 卷积层与卷积层、全连接层与全连接层之间使用 LeakyRelu(·)激活函数连接,卷积层与全连接层之间则采用 Sigmoid(·)激活函数. 4 层的卷积层用来降低图像分辨率,并对基本的局部特征进行编码以进行分类,在这里,卷积层主要负责提取输入图像的局部特征. 全连接层用于对卷积层选取的特征进行加权求和,实际上起到分类的作用. LeakyRelu(·)激活函数的输出可以有效防止负值输入致使其导数为 0,神经网络无法反向传播更新参数问题的出现. 模型的最后输出采用 Sigmoid 函数控制,使输出范围控制在 $[0,1]$.

对于图像特征比较简单的单通道灰度图像数据集 MNIST 和 FASHION-MNIST,本文设计 4 个全连接层 Linear 组成判别模型. 模型的全连接层之间使用 LeakyRelu(·)激活函数; 由于过多的全连接层会导致神经网络出现过拟合的情况,在所有全连接层之间添加 Dropout 层,通过设置 0.3 的 Dropout 系数防止网络出现过拟合.

2.2 网络结构设计与分析

判别网络本质上是分类网络. 在图像分类领域中,全连接网络和卷积网络是分类精度比较高的神

经网络.

输入图像在进行前向传播时,假设卷积层有 A 个输出通道和 B 个输入通道,卷积核大小为 $E \times F$,每个输出通道的特征图大小均为 $M \times N$,则计算量记为 $C_j = A \times B \times E \times F \times M \times N$,其所要学习的参数量为 $R = E \times F \times A \times B$. 定义计算量与参数量之比为 $C_R = C_j / R = M \times N$,可以看出若卷积层的输出特征图尺寸越大,则 C_R 值越大,参数的重复率越高.

卷积层在输出特征图维度上实现了权值共享,可降低参数量,卷积层的局部连接也大幅度减少参数量,但是计算量仍然较大. 因此,对于数据量较多的 3 通道彩色图像数据集,设计 4 个卷积层相连,然后使用 2 层全连接层连接,可在保证分类性能的前提下减少计算量. 由于是 3 通道数据,卷积层输入维度为 3,随后的输出维度可自定义,一般以 $2n$ 为原则,本文设计为 $3 > 8 > 16 > 32 > 64$,输出维度依次增大,更加有利于图像关键特征的提取. 卷积层的其余参数,神经网络会根据给出的维度变化自动确定,在此不再赘述.

在全连接层中数据样本前向传播计算量为 $C_j = V \times G$,其中 G 表示输入节点组成向量的维度,V 表示输出节点组成向量的维度,其参数量为 $C_R = V \times G$. 由此可见,全连接层的权值重复率较低,与输入输出维度无关,但计算量比卷积层大,因此适用于数据量小的单通道灰度图像. 本文对灰度图像使用 4 层全连接层网络相连,由于 MNIST 和 FASHION-MNIST 图像大小均为 28×28,按照 $2n$ 原则,网络结构设计由大到小依次为 $784 > 1\,024 > 512 > 256 > 1$. 因为全连接层容易发生参数冗余,所以在网络层之间添加 Dropout 层,防止网络出现过拟合的情况,提高网络的分类能力.

2.3　应用噪声数据训练

本文的检测方法的出发点是基于 Liang Bin 等[13] 的假设,将添加到原始输入的扰动看成是一种噪声. 应用噪声数据可以模拟对抗样本数据的特征分布,提高模型的判别能力.

在训练过程中,判别模型接受 2 种数据样本的输入. 一种是原始训练集样本数据,另一种是被添加了随机噪声 φ_i 的噪声图像. 判别网络会根据本文描述的目标函数进行自动地训练.

理论上,若把对正常图像所添加的扰动看成是一种噪声,则只要在训练神经网络过程中,神经网络遍历所有噪声系数的噪声数据,就可以检测出所有噪声样本(设检测模型所面对的检测数据只有正常样本和对抗样本,在已有的研究中一般只检测原始数据和恶意样本),但这在计算上是不可行的,并且也会在实际应用中造成一定的误检. 因此,应选取合适的噪声系数.

在生成对抗样本中,攻击系数表示为对正常样本的噪声影响. 根据对对抗样本多种攻击方法[1-4] 的研究,绝大部分的攻击不会对正常样本造成太大的扰动,但随着攻击系数的增大,数据样本的关键语义特征破坏程度亦会逐渐升高. 为了最大限度地保留样本的重要特征,一般的攻击方法的攻击系数 ε 设定在(0, 0.30] 范围内. 依据这个特点,本文选取噪声系数 δ 在 [0.01, 0.30] 之间,由于噪声系数为连续值,因此以 0.05 为间隔,选取 7 个数值作为训练判别模型的噪声系数 δ,向原始图像添加相应的噪声.

$$z_i = \delta_i \varphi_i + x_i, \tag{1}$$

$$\text{s.t.} \quad z_i = c_{lip}(z_i, [0,1]), \tag{2}$$

其中 x_i 为原始数据样本(原始数据已进行归一化处理),z_i 为噪声数据样本,通过对原始图像 x_i 添加不同噪声系数 δ_i 的随机噪声数据 φ_i,得到噪声样本 z_i.

2.3.1　目标函数　目标函数是神经网络模型训练成功的关键. 二分类判别网络的目标函数为

$$\max_D E_{x \sim Pdata}(\log D(x)) + E_{z \sim p_z}(\log(1 - D(z))), \tag{3}$$

其中 D 表示判别神经网络,x、z 分别为干净的原始数据和被添加了随机噪声的噪声数据,$D(x)$ 表示 x 来自原始数据集的概率,$D(z)$ 表示 z 来自噪声数据集的概率. 通过训练,判别网络 D 对输入的正常样本输出较高的数值,而对于噪声样本输出较低的数值. 为了实现式(3)的目标,使用本文描述的损失函数来完成.

2.3.2　损失函数　对于式(3),希望尽可能最大化 $D(x)$ 和最小化 $D(z)$,通过使用交叉熵损失函数解决这个问题.

对于原始和噪声这 2 种输入样本,本文设置了不同的标签. 原始样本的网络输出,设置为真实标签,对于噪声样本的输出,则设为错误标签. 2 种样本输出都是通过交叉熵损失函数 $B_{CELoss}(\cdot)$ 来达到最小化,其计算方法为

$$B_{CELoss}(D(x), P) = -\mu_i(P\log D(x_i) + (1-P) \cdot \log(1 - D(x_i))), \tag{4}$$

其中 x_i 为原始数据样本,$D(x_i)$ 为原始样本的网络输出,P 为正确标签,μ_i 为权重,通过最小化式(4)不断增大 $D(x)$.

$$B_{CELoss}(D(z), Q) = -\omega_i(Q\log D(z_i) + (1-Q) \cdot \log(1 - D(z_i))), \tag{5}$$

其中 z_i 为噪声数据样本,Q 为错误标签,ω_i 为权重,通过最小化式(5)尽可能减小 $D(z)$. μ_i、ω_i 由系统随机初始化.

2.4 训练判别网络

如图 1 所示,首先将原始样本或者噪声样本输入判别网络 D,网络产生输出 $D(x)$,对于正常样本数据设置正确的标签 P,通过最小化式(4)定义的交叉熵损失函数 $B_{CELoss}(\cdot)$,训练判别网络对正常样本产生较高则设置错误的标签 Q,通过最小化式(5),使得网络输出较小的数值.

在训练过程中,原始样本和噪声样本交替输入,对不同的数据样本输出设置不同的标签,并分别使用对应的损失函数,不断训练优化网络. 训练完成的判别网络对一个给定的输入将会产生一个相应的数值,通过与设定的阈值相比较,可以检测输入样本的对抗性与否.

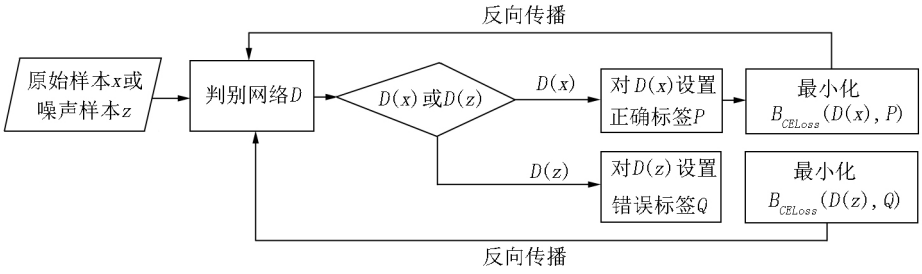

图 1 判别网络训练过程

2.5 检测对抗样本

综合考虑对抗样本的检测率和正常样本的误检率情况,判别模型的误检率不宜设置过高. 因此,对已经训练完成的二分类判别网络 D,取 MNIST、FASHION-MNIST(FMNIST)以及 CIFAR-10 数据集的训练集,作为正常样本集 x,输入对应的判别网络 D(Binary Discrimination Network),得到所有的输出值 $D(x)$,从小到大排列,选取数据的 5% 所对应位置表示的数值作为检测阈值 T. 如 MNIST 训练集 60 000张图像输入判别网络 D,得到 60 000 个数值,依序排列,第 3 000($60\,000 \times 0.05 = 3\,000$)个数值即为检测阈值.

向已经训练完成的网络输入一张待检测的图像 x,对应的判别网络 D 会产生一个输出 $D(x)$,将输出值 $D(x)$ 与设定的阈值 T 相比较,大于 T 的判断为正常图像,小于 T 的则视为对抗样本. 完整的检测过程如图 2 所示.

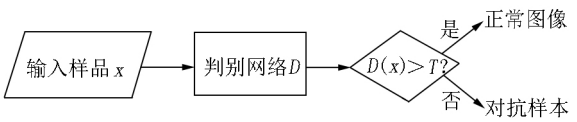

图 2 对抗样本检测过程

3 实验

本文实验采用的硬件设备主要为 Intel 公司生产的 i7-5500U CPU,主频 2.40 GHz,内存 16 GB. 操作系统为 Windows10,以 3.7.9 版本 Python 为开发语言,使用 Opencv 4.4.0 对数据集图像进行处理,基于 PyTorch 深度学习框架完成网络模型的搭建.

3.1 数据集与目标模型

本文在 MNIST、FASHION-MNIST(FMNIST)以及 CIFAR-10 这 3 个标准数据集上开展此次实验. 在对抗样本检测领域中,这 3 个数据集都是众多学者常用的实验数据集. MNIST 数据集都是“0 ~ 9”的手写数字图像,其中训练集 60 000 张,测试集10 000张. FMNIST 数据集与 MNIST 类似,但图像内容不再是手写数字,而是换成了“牛仔裤”“T 恤”“凉鞋”“外套”等服饰. CIFAR-10 是一个包含有10 个类别的 RGB 彩色图片,共有 50 000 张训练图片和10 000张测试图片. 本文采用各个数据集的训练集训练目标模型,用测试集来测试其识别精度.

MNIST 数据集的目标模型为经典的 LeNet-5[15] 网络,FMNIST 使用的网络为 CNN[16] 模型,CIFAR-10 则采用 ResNet[17]. 目标模型的测试皆采用 Top-1 精度作为评判标准,各个数据集相对应的目标模型的测试精度,如表 2 所示.

表 2 目标模型的测试精度

数据集	目标模型	测试精度/%
MNIST	LeNet-5	99.21
FMNIST	CNN	93.08
CIFAR-10	ResNet	93.69

3.2 产生对抗样本

针对单通道的灰度图像数据集,采用 FGSM[4]、DDN[18]、PGD[19] 这 3 种攻击方法,对于 3 通道的彩色图像数据集,则使用 BIM[2]、JSMA[3]、EAD[20] 这 3 种攻击方法产生相应的对抗样本.

在实验中,设置 FGSM 和 BIM 这 2 种攻击方法的攻击系数 $\varepsilon = 0.1$,因为太大的攻击系数会导致生

成的对抗图像噪声过大,人眼容易观察得知,而失去了隐蔽攻击的意义. 其他攻击方法则直接采用了已有文献的实现方法.

对所采用的 3 个数据集的测试集进行攻击,平均攻击成功率分别为 100.00%、100.00%、99.85%.

表 3 是各种攻击方法在 MNIST、FMNIST 和 CIFAR-10 这 3 个数据集上攻击其对应的目标模型的详细结果.

表 3　不同攻击方法在 3 个数据集的攻击成功率

数据集	目标模型	攻击方法	攻击成功率/%	平均攻击成功率/%
MNIST	LeNet-5	FGSM	100.00	100.00
		DDN	100.00	
		PGD	100.00	
FMNIST	CNN	FGSM	100.00	100.00
		DDN	100.00	
		PGD	100.00	
CIFAR-10	ResNet	FGSM	100.00	99.85
		JSMA	99.55	
		BIM	100.00	

3.3　评估指标

为了更好地评估所提出的检测模型的有效性以及与其他相关的检测方案进行公平的比较,采用机器学习领域二分类问题常用的召回率 r、准确率 p 和 F_1 评分[21] 这 3 个评估指标来量化检测性能,F_1 值越高表明一个检测模型对对抗样本的整体检测性能越好. 从模型误检方面考虑,还引入了一个新的度量 F_{PR}(False Positives Rate) . F_{PR}表示一个检测器对正常样本的误检情况,F_{PR}值越低表明检测器发生误判的可能性越小. 它们的计算公式分别为

$$r = T_P/(T_P + F_N),\ p = T_P/(T_P + F_P),\ F_1 = 2rp/(r + p),\ F_{PR} = F_P/C_E,$$

其中 T_P 是被正确检测到的对抗样本的数量(True Positives) ,F_N 是没有被检测到的恶意样本的数量(False Negatives) ,F_P 是被检测到的原始图像的数量(False Positives) ,C_E 为原始样本的数量(Clean Examples) .

3.4　实验结果及分析

根据二分类判别网络模型的检测结果,对于 MNIST、FMNIST、CIFAR-10 这 3 个数据集,使用 3 个表详细列出每种攻击方法在每个数据集上的具体检测结果.

由表 4 ~ 表 5 可知,FGSM、DDN 和 PGD 这 3 种攻击方法产生的对抗样本,根据检测评估标准,在 FMNIST 数据集上,判别网络的检测方法 F_1 值达到了 99% 以上,对于 MNIST,则超过了 97%. 结果充分表明本文检测模型是有效的.

从表 6 可知,对 CIFAR-10 数据集的检测,模型的检测对 BIM、JSMA 攻击方法展现出了良好的效果,F_1 值分别达到了 98.25%、92.59%.

对于原始样本的误检率 F_{PR},本文的检测方法也有出色的表现,在 3 个数据集中的误检率都低于 5%,在 FMNIST 上,只出现了 0.65% 的误检情况,这表明本文提出的模型对正常样本产生误检的可能性非常低.

表 7 列举了本文与 Xu Weilin 等[12]、Liang Bing 等[14] 的比较实验结果. 在比较实验中,只关注 F_1 和误检率 F_{PR}这 2 个评判标准的数值结果,因为 F_1 综合了召回率和准确率的结果,可以展现一个模型对恶意样本的检测性能,而 F_{PR}则能体现一个检测模型对正常样本的误检情况.

表 4　对 MNIST 数据集的检测结果

攻击方法	F_P/幅	C_E/幅	F_{PR}/%	T_P/幅	F_N/幅	召回率/%	准确率/%	F_1/%
FGSM	494	10 000	4.94	10 000	0	100.00	95.29	97.59
DDN	494	10 000	4.94	9 951	49	99.51	95.27	97.34
PGD	494	10 000	4.94	9 955	45	99.55	95.27	97.36

表 5　对 FMNIST 数据集的检测结果

攻击方法	F_P/幅	C_E/幅	F_{PR}/%	T_P/幅	F_N/幅	召回率/%	准确率/%	F_1/%
FGSM	65	10 000	0.65	10 000	0	100.00	99.35	99.68
DDN	65	10 000	0.65	9 951	49	99.51	99.35	99.43
PGD	65	10 000	0.65	9 955	45	99.55	99.35	99.45

表 6　对 CIFAR-10 数据集的检测结果

攻击方法	F_P/幅	C_E/幅	F_{PR}/%	T_P/幅	F_N/幅	召回率/%	准确率/%	F_1/%
BIM	355	10 000	3.55	9 999	1	99.99	96.57	98.25
JSMA	355	10 000	3.55	8 888	1 067	89.28	96.16	92.59
EAD	355	10 000	3.55	1 186	8 751	11.94	76.96	20.67

由表 7 可知,检测模型在 FMNIST 数据集上,误检率仅有 0.65%,优胜于对比方,对于 3 种不同的攻击也显示出了优异的检测性能,达到了 99% 以上,Xu Weilin 等[12]、Liang Bin 等[13]的检测方案则由于误检率偏高,导致 F_1 值偏低,降低了模型的检测性能. 在 MNIST 数据集上,本文检测模型的 F_1 值最高达到 97.59%,对其他攻击方法的检测和对比方案基本持平,平均相差不到 1%.

对于 CIFAR-10 数据集,与竞争模型相比,判别模型在对 EAD 攻击的检测上,F_1 值仅为 20.67%,EAD 攻击对原始图像的扰动比较小,所以模型对其的检测性能会偏低,但是模型误检率要远低于对比方,与 Liang Bin 等[13]相比低 17%,并且在对 BIM、JSMA 攻击的检测也得到了较高的 F_1 分数,Liang Bin 等[13]对于 JSMA 的攻击,也仅取得了 20.74% 的 F_1 分数. Xu Weilin 等[12]的模型则表现得比较稳定,F_1 值平均保持在 80% 左右,但是模型误检率偏高,导致整体检测性能的下降.

表 7 与其他检测方案的比较 %

数据集	检测方法	F_{PR}	F_1					
			DDN	PGD	FGSM	BIM	JSMA	EAD
MNIST	文献[12]	6.35	96.89	96.89	96.71	–	–	–
	文献[13]	0.79	98.24	98.58	97.42	–	–	–
	本文	4.94	97.34	97.36	97.59	–	–	–
FMNIST	文献[12]	42.49	82.14	82.11	75.74	–	–	–
	文献[13]	52.58	70.71	70.80	51.26	–	–	–
	本文	0.65	99.43	99.45	99.68	–	–	–
CIFAR-10	文献[12]	24.04	–	–	–	79.07	83.29	85.78
	文献[13]	20.58	–	–	–	65.95	20.74	81.48
	本文	3.55	–	–	–	98.25	92.59	20.67

从模型误检率的结果来看,本文所设计的判别网络模型在多个数据集上的误检率均维持在一个较低的数值,优胜于竞争模型或基本与之持平.

对于整体检测性能,本文所提出的检测模型对于大部分的攻击方法具有较高的检测率并且检测率优于竞争模型.

4 结论与展望

对抗样本的攻击会对深度神经网络模型产生严重的安全性问题. 二分类判别网络检测模型以端到端的方式,无须设置其他限制条件,能从源样本中直接探测到对抗样本的存在,该模型通过设置合适的阈值,在多个数据集上都取得了良好的效果. 与其他的检测方案相比,判别网络模型对于恶意样本攻击的有效检测,在一定指标上性能优于竞争模型.

同时在不同数据集上的实验也表明: 本文所设计的检测模型对于大部分攻击方法都可以实现检测,但是,也会存在对个别攻击方法检测指标较低的情况,这需要进一步地改善. 在未来的工作中,希望能够更好地改进此判别网络的结构,以训练出性能更加优异的检测器,提升检测精度和效率.

5 参考文献

[1] Szegedy C, Zaremba W, Sutskever I, et al. Intriguing properties of neural networks [EB/OL]. [2020-06-17]. https://arxiv.org/abs/1312.6199.

[2] Kurakin A, Goodfellow I, Bengio S. Adversarial examples in the physical world [EB/OL]. [2020-06-17]. https://arxiv.org/abs/1607.02533v4.

[3] Papernot N, McDaniel P, Jha S, et al. The limitations of deep learning in adversarial settings [EB/OL]. [2020-06-17]. https://arxiv.org/abs/1511.07528.

[4] Goodfellow I J, Shlens J, Szegedy C. Explaining and harnessing adversarial examples [EB/OL]. [2020-06-17]. https://arxiv.org/pdf/1412.6572.pdf.

[5] 张钹,朱军,苏航. 迈向第三代人工智能 [J]. 中国科学: 信息科学,2020,50(9):1281-1302.

[6] 蒲元芳,张巍,滕少华,等. 基于决策树的协同网络入侵检测 [J]. 江西师范大学学报: 自然科学版,2010,34(3):302-307.

[7] 易倩,滕少华,张巍. 基于马氏距离的 K 均值聚类算法的入侵检测 [J]. 江西师范大学学报: 自然科学版,2012,36(3):284-287.

[8] Meng Dongyu, Chen Hao. Magnet: a two-pronged defense against adversarial examples [EB/OL]. [2020-06-17]. https://arxiv.org/abs/1705.09064v2.

[9] Metzen J, Jan H, Genewein T, et al. On detecting adversarial perturbations [EB/OL]. [2020-06-17]. https://arxiv.org/abs/1702.04267.

[10] Hendrycks D, Gimpel K. Early methods for detecting adversarial images [EB/OL]. [2020-06-17]. https://arx-

iv. org/pdf/1608. 00530v2. pdf.

[11] Li Xin, Li Fuxin. Adversarial examples detection in deep networks with convolutional filter statistics [EB/OL]. [2020-06-17]. https://arxiv. org/abs/1612. 07767.

[12] Xu Weilin, Evans D, Qi Yanjun. Feature squeezing: detecting adversarial examples in deep neural networks [EB/OL]. [2020-06-17]. https://arxiv. org/pdf/1704. 01155. pdf.

[13] Liang Bin, Li Hongcheng, Su Miaoqiang, et al. Detecting adversarial image examples in deep networks with adaptive noise reduction [EB/OL]. [2020-06-17]. https://arxiv. org/pdf/1705. 08378. pdf.

[14] Krizhevsky A, Sutskever I, Hinton G. Image net classification with deep convolutional neural networks [J]. Communications of the ACM, 2017, 60(6): 84-90.

[15] Lecun Y, Bottou L, Bengio Y, et al. Gradient-based learning applied to document recognition [J]. Proceedings of the IEEE, 1998, 86(11): 2278-2324.

[16] Ashmeet Lamba. CNN for Fashion MNIST Dataset [EB/OL]. [2020-06-17]. https://github. com/ashmeet13/FashionMNIST-CNN.

[17] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition [EB/OL]. [2020-06-17]. https://ieeexplore. ieee. org/document/7780459.

[18] Rony J, Hafemann L G, Oliveira L S, et al. Decoupling direction and norm for efficient gradient-based l2 adversarial attacks and defenses [EB/OL]. [2020-06-17]. https://arxiv. org/pdf/1811. 09600. pdf.

[19] Madry A, Makelov A, Schmidt L, et al. Towards deep learning models resistant to adversarial attacks [EB/OL]. [2020-06-17]. https://arxiv. org/pdf/1706. 06083. pdf.

[20] Chen Pinyu, Sharma Y, Zhang Huan, et al. Ead: elastic-net attacks to deep neural networks via adversarial examples [EB/OL]. [2020-06-17]. https://arxiv. org/pdf/1709. 04114. pdf.

[21] Powers D M. Evaluation: from precision, recall and *F*-measure to ROC, informedness, markedness and correlation [EB/OL]. [2020-06-17]. https://arxiv. org/abs/2010. 16061v1.

The Adversarial Samples Detection with a Binary Discrimination Network

ZENG Lihong, ZHANG Wei*, TENG Shaohua

(School of Computers, Guangdong University of Technology, Guangzhou Guangdong 510006, China)

Abstract: The deep neural network is vulnerable to the attack of adversarial samples that are generated by adding small but special perturbations to the original datasets, resulting in the network model giving error output with high confidence. Additionally, most of the detection methods of adversarial samples need to have many preconditions when detecting, and the whole detection ability is limited. Therefore, a binary discrimination network is proposed to effectively improve the detection rate of the adversarial samples, which extracts the main features of the sample data in the way of multi-layer convolution, trains the network with different levels of noise data, and continuously optimizes the network model with unique discriminant objective function. The model can be directly deployed to the source data of the target model to detect the presence of adversarial samples, and can be used on a large scale by an end-to-end way. Experimental results show that the detection rate of this model is better than that of other comparison models.

Key words: binary discrimination network; deep neural network; adversarial samples; detection

(责任编辑: 冉小晓)