软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2020,31(9):2756−2769 [doi: 10.13328/j.cnki.jos.005943]

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

利用特征融合和整体多样性提升单模型鲁棒性*

韦 璠, 宋云飞, 邵明莉, 刘 天, 陈小红, 王祥丰, 陈铭松

(上海市高可信计算重点实验室(华东师范大学),上海 200062)

通讯作者: 陈铭松, E-mail: mschen@sei.ecnu.edu.cn

摘 要: 使用深度神经网络处理物联网设备的急剧增加产生的海量图像数据是大势所趋,但由于深度神经网络对于对抗样本的脆弱性,它容易受到攻击而危及物联网的安全.所以,如何提高模型的鲁棒性,就成了一个非常重要的课题.通常情况下,组合模型的防御表现要优于单模型防御方法,但物联网设备有限的计算能力使得组合模型难以应用.为此,提出一种在单模型上实现组合模型防御效果的模型改造及训练方法:在基础模型上添加额外的分支;使用特征金字塔对分支进行特征融合;引入整体多样性计算辅助训练.通过在 MNIST 和 CIFAR-10 这两个图像分类领域最常用的数据集上的实验表明,该方法能够显著提高模型的鲁棒性.在 FGSM 等 4 种基于梯度的攻击下的分类正确率有 5 倍以上的提高,在 JSMA,C&W 以及 EAD 攻击下的分类正确率可达到原模型的 10 倍.同时,不干扰模型对干净样本的分类精度,也可与对抗训练方法联合使用获得更好的防御效果.

关键词: 物联网;特征融合;整体多样性;模型防御;鲁棒性;对抗样本

中图法分类号: TP183

中文引用格式: 韦璠,宋云飞,邵明莉,刘天,陈小红,王祥丰,陈铭松.利用特征融合和整体多样性提升单模型鲁棒性.软件学报, 2020,31(9):2756−2769. http://www.jos.org.cn/1000-9825/5943.htm

英文引用格式: Wei F, Song YF, Shao ML, Liu T, Chen XH, Wang XF, Chen MS. Improving adversarial robustness on single model via feature fusion and ensemble diversity. Ruan Jian Xue Bao/Journal of Software, 2020,31(9):2756−2769 (in Chinese). http://www.jos.org.cn/1000-9825/5943.htm

Improving Adversarial Robustness on Single Model via Feature Fusion and Ensemble Diversity

WEI Fan, SONG Yun-Fei, SHAO Ming-Li, LIU Tian, CHEN Xiao-Hong, WANG Xiang-Feng,

CHEN Ming-Song

(Shanghai Key Laboratory of Trustworthy Computing (East China Normal University), Shanghai 200062, China)

Abstract: It is an inevitable trend to use deep neural network to process the massive image data generated by the rapid increase of Internet of Things (IoT) devices. However, as the DNN is vulnerable to adversarial examples, it is easy to be attacked and would endanger the security of the IoT. So how to improve the robustness of the model has become an important topic. Usually, the defensive performance of the ensemble model is better than the single model, but the limited computing power of the IoT device makes the ensemble model difficult to apply. Therefore, this study proposes a novel model transformation and training method on a single model to achieve similar defense effect like ensemble model: adding additional branches to the base model; using feature pyramids to fuse features; and introducing ensemble diversity for training. Experiments on the common datasets, like MNIST and CIFAR-10, show that this method can significantly improve the robustness. The accuracy increases more than fivefold against four gradient-based attacks such as FGSM, and

* 基金项目: 国家重点研发计划(2018YFB2101300); 国家自然科学基金(61872147)

Foundation item: National Key Research and Development Program of China (2018YFB2101300); National Natural Science Foundation of China (61872147)

本文由“智能嵌入式系统”专题特约编辑王泉教授、吴中海教授、陈仪香教授、苗启广教授推荐.

收稿时间: 2019-07-01; 修改时间: 2019-08-18; 采用时间: 2019-11-02; jos 在线出版时间: 2020-01-13

CNKI 网络优先出版: 2020-01-14 11:27:05, http://kns.cnki.net/kcms/detail/11.2560.TP.20200114.1126.024.html

can be up to 10 times while against JSMA, C&W, and EAD. This method does not disturb the classification of clean examples, and could obtain better performance while combining adversarial training.

Key words: Internet of things; feature fusion; ensemble diversity; model defense; robustness; adversarial example

近年来,物联网设备的数量在不断飙升.据 IoT Analysis 网站发表的 2018 第一、第二季度的物联网发展状态统计[1],全球联网设备数量已达 170 亿,其中,物联网设备占 70 亿;并且该统计不包含固定电话、智能手机、平板以及笔记本电脑,也不包含过去加入到连接中但目前已不再使用的设备.尽管如此,这一数据也呈现出不断攀升的趋势,预计会在 2020 年达到 100 亿以及 2025 年达到 220 亿.大量的物联网设备会产生海量的数据,图像是其中一种重要的数据形式,使用深度神经网络处理海量图像数据[2−5]已是大势所趋.不可否认,深度神经网络经过近几十年的发展,在很多任务上都有了长足的进步,在图像分类领域的表现尤为突出.但这些高性能分类器在面对对抗样本攻击时却暴露了其脆弱性[6]:这些对抗样本只在原始图片上加了微小量的扰动,通常情况下,这些扰动是人眼无法发觉的,并不会对人的判断造成影响,但却能成功愚弄深度神经网络分类器,使其做出错误的分类判断.所以说,如何提高模型对对抗样本攻击的防御能力,就成为了一个非常重要的课题.

提高模型的鲁棒性,是目前应对对抗样本攻击的一个重要方法.模型鲁棒性是一个用于分析模型对于微小扰动的抵抗能力的评判标准,在相同扰动下,模型的判断准确率越高,鲁棒性越好.根据使用模型数量的不同,可以将现有工作分为两类:基于单模型的鲁棒性提升方法和基于组合模型的鲁棒性提升方法.根据对抗样本处理方式的不同,又可以将基于单模型的鲁棒性提升方法分为对抗样本检测和对抗样本防御.

- 对抗样本检测方法是在原模型之外构造一个单独的样本探测器,样本输入到目标模型之前必须先经过探测器判断,只有被判断为干净的样本才能输入到目标模型中[7,8].这不仅需要额外的资源消耗,而且样本探测器本身也存在受到攻击的风险.Carlini 等人[9]甚至通过研究指出:目前大多数的对抗样本探测器都是无效的,并且它们也能被轻易攻击;
- 对抗样本防御方法一般着眼于提升网络本身的鲁棒性,以使其能够正确分类对抗样本,相较于对抗样本检测方法而言更有效且资源消耗更低,更适用于物联网应用场景.

基于组合模型的鲁棒性提升方法则是综合评估多个模型的预测结果得到最终输出,通常情况下,相比于基于单模型的防御方式而言会有更好的防御表现[10].特别是清华团队最近提出的基于组合模型的提升模型整体多样性方法[11],通过分离组合成员的非极大预测的分布,能够在不影响模型准确率的情况下提升组合内各个成员之间的多样性,降低成员间的相似程度,从而使得对抗样本难以在各个成员之间迁移,更进一步地提高了组合模型的整体鲁棒性.尽管如此,考虑到物联网设备有限的资源配置、计算能力以及实时性要求等,基于组合模型的防御方式很难落实到物联网应用中.

基于以上情况并受到组合模型整体多样性的启发,本文提出了一种在单模型上使用组合模型防御方式的模型鲁棒性提升方法:依据分支网络(branchynet)[12]中浅层出口也可以达到较好的预测准确率理论,给原始模型添加额外的分支模拟组合模型效果;在分支之间加入特征金字塔[13]实现特征融合,消除了各分支输出的尺度差异;针对多分支单模型改进了整体多样性计算方式,提高了单模型内各个分支间的整体多样性,有效避免了对抗样本在各分支之间的迁移.本文对图像分类模型 Resnet-32[14]做了以上改造及训练,通过在 MNIST 和 CIFAR-10 数据集上使用目前流行的攻击方式进行防御实验,以证明本方法的可行性与有效性.

本文第 1 节主要介绍目前已有的在单模型上应对对抗样本攻击、提高模型防御能力的相关研究工作.第 2 节从模型改造、多分支单模型的整体多样性计算以及训练过程这 3 个方面详细阐述本文提出的模型鲁棒性提升方法.第 3 节通过在 MNIST 和 CIFAR-10 这两个图像分类领域最常用的数据集上对比目前流行的 7 种白盒攻击方式下的分类正确率,证明本方法能够在不影响原模型分类效果的前提下,显著提高模型的鲁棒性.第 4 节对本文工作做出总结并给出未来的工作展望.

1 相关工作

对抗样本防御是目前在单模型上应对对抗样本攻击的一个有效方法.对抗样本攻击是指使用经过恶意微小调整的样本输入到神经网络,使其给出完全不同于真实样本的错误结果.这些调整通常是人眼无法察觉的,但却能成功愚弄神经网络使其做出错误判断,这就会对机器学习模型支持的系统,特别是安全攸关的应用带来极大的安全威胁.对抗样本防御方法的主要思想是:提升模型自身的鲁棒性,以使其能够对对抗样本做出正确判断.根据 Rajeev 等人[15]的理论,可将其分为 3 类:对抗训练、对输入降噪预处理以及直接修改目标模型的网络结构或优化训练过程.

对抗训练是把对抗样本加入到训练集中对模型进行训练.例如 Miyato 等人提出的虚拟对抗训练方法(VAT)[16],他们基于虚拟对抗性损失提出了一种新的正则化方法,定义了没有标签信息的对抗方向,将对抗训练扩展到了半监督学习任务中;Kurakin 等人则成功地将对抗训练扩展到大型模型和数据集中[17],并解决了对抗训练过程中的"标签泄露"效应;在 Google Brain 组织的 NIPS 2017 竞赛中,也有许多参赛队伍使用了对抗训练方法并取得了不错的成绩[18].尽管如此,对抗训练却有其难以规避的显著缺点,即:这种训练方式下生成的模型的防御效果不具有普适性,意味着它只能应对训练过程中使用的攻击方式,当攻击方式发生变化时,模型的脆弱性就会再次体现,将其运用到物联网设备中也会遇到同样的问题.

对输入进行降噪预处理是在将样本输入到目标模型之前先去掉样本上的恶意扰动.例如 Samangouei 等人提出的 Defense-GAN 模型[19],它是一个利用生成模型的表达能力来保护深度神经网络免受对抗样本攻击的新框架,在将给定图像输入到目标模型推理之前,利用生成器找到该图像的一个不包含对抗性变化的相近输出,然后再把这个输出交由分类器处理;Guo 等人[20]也研究了这种在将输入样本馈送到系统之前转换输入来抵御图像分类系统上的对抗样本攻击策略,他们在把图像输入到卷积网络分类器之前,应用诸如比特深度缩减、JPEG 压缩、总方差最小化和图像缉缝之类的变换,通过在 ImageNet 数据集上的实验证明,总方差最小化和图像缉缝能够有效提高系统防御能力.但总体而言,这种解决方案实际上也需要在将输入样本输送到目标模型之前添加额外的预处理过程,增加了物联网设备的计算负担.

直接修改目标模型的网络结构或者优化训练过程,是另一种有效提高模型鲁棒性的方法.例如 Lamb 等人提出的强化网络[21],它是对现有网络的一个简单转化,通过识别隐藏状态在数据流中断开的时间,把这些隐藏状态映射回网络中运行良好的数据流部分,加强了深度网络的隐藏层,并通过实验证明了强化这些隐藏状态可以提高深度网络的鲁棒性.但是,强化网络只有在与对抗训练同时使用时才能提高鲁棒性,这对于迭代攻击而言是非常昂贵的.Rajeev 等人则认为,可以通过紧凑特征学习(compact feature learning)[15]来中和对抗攻击.他们认为,在一个封闭有界的区间里来学习特征能够提升网络的鲁棒性,并由此提出了一种创新的卷积方式——紧凑卷积,保证了每层的特征有界且彼此相似.最后,通过实验证明使用此方法构造的新的紧凑卷积网络能够抵御多种攻击方式,并且与 CNN 相比不会产生额外的训练开销,但这种改进方式也只对卷积神经网络有效.

通过对比以上 3 种基于单模型的对抗样本防御方法,本文决定采取第 3 种方式,直接修改目标模型的网络结构并优化训练过程.受到组合模型的防御表现一般优于单模型[10]的论断和基于组合模型的整体多样性[11]定义的启发,通过给目标模型添加分支并在训练过程中添加整体多样性计算,提出了一种在单模型上实现组合模型鲁棒性提升效果的方法.

2 提出的方法

本文提出了一种基于特征融合和整体多样性的单模型鲁棒性提升方法.本文认为,在单个模型中进行均化的多分支预测可以提高模型鲁棒性.该方法主要包含两个部分:一是将基础模型改造为均化多分支预测模型,二是提出针对在单模型中各分支长度不一致情况下的整体多样性计算方法.下面将分别从结构改造、基于多分支单模型的整体多样性计算以及模型训练这 3 个方面具体阐述.

2.1 结构改造

深度神经网络中:浅层次的特征图尺寸较大,蕴含了更多的信息,表达了输入的细节特征;深层次的特征图尺寸较小,内容更抽象,表达了输入的语义特征.由于语义特征可以更抽象、更彻底地描述一个物体,所以层次的加深使得模型可以更全面地认识输入的特征,达到更好的表现.但很多物体只需要使用部分细节特征或不那么抽象的语义特征就可以辨认,所以浅层次的特征图做预测亦可以达到较好的精度.一般来说,特征融合是指采用现有的多个特征集生成新的融合特征,它可以使两组不同层次的特征图信息互补,横向链接的特征融合(如图 1(c)所示)需要浅层特征图在不改变尺寸的同时增加到与深层相同的通道数,而深层的特征图需要使用上采样的方法将尺寸扩大到与浅层一致,这样可以使得浅层的最终特征图获得深层信息,提高自身的判断精度.网络通常为深层次的准确率会更好,同时也由于浅到深会有尺寸变换,细节特征还是会被抽象化,而深到浅的上采样则是均分扩大,不会将语义特征具体化,所以在本文中,特征融合只需将深层特征融合到浅层.据此,定义了鲁棒性提升方法的改造过程,如图 1 所示,具体过程如下.

1) 以现有模型(如图 1(a)所示)为基础,根据使用浅层特征图也能达到较好的预测精度理论,在模型浅层添加额外分支如 Pre1~Pre3,形成多预测结构的分支网络(如图 1(b)所示);
2) 对图 1(b)的分支网络的各个分支做特征融合,从最深的出口开始,将做出预测前的特征上卷积融合进相对浅层的出口,递归执行此步骤至最浅层出口处.在图 1(d)中的表现为 pre3 处特征融合更深处的全部特征,pre2 处特征获得融合后 pre3 处特征的上卷积,pre1 处特征获得融合后 pre2 处特征的上卷积,完成全部的融合过程后,形成图 1(d)所示的特征金字塔结构.最终,浅层特征就融合了相对其的全部深层特征,可以更有效地辅助判断,同时复杂化模型的线性过程;
3) 为了保证单模型的唯一出口,对图 1(d)中各个分支的预测做平均运算,获得最终单出口输出 Final Out;
4) 除了计算最终单出口输出 Final Out 的误差 L_{CE} 之外,也对图 1(d)中各个分支的预测做误差计算以提高各自的识别精度.为了降低分支之间预测的相似性,从而提高模型鲁棒性,使用第 2.2 节提出的基于多分支单模型的整体多样性计算公式计算 L_{ED};最终,根据误差结果完成模型参数优化.

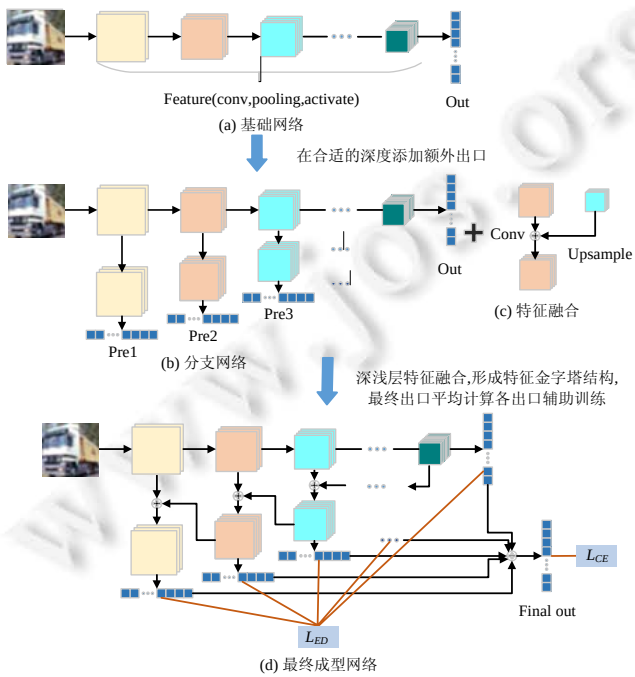


Fig.1 Structure transformation process

图 1 结构改造过程

2.2 基于多分支单模型的整体多样性计算

组合模型在训练过程中,会因为成员之间的信息交互导致各个成员的最终输出表征相似,这一问题在图1(d)所示的多分支单模型中表现更为明显.这是因为分支预测之间有共享层,在简单的训练下,分支之间的信息交互会比组合模型成员之间的更强.为了使训练出来的各个分支预测不趋同,提升分支之间的整体多样性就显得尤为重要.为此,本文设计了基于多分支单模型的整体多样性计算方法,分为3个部分.

- 第1部分为确保每个分支的准确率.

作为可以单独输出的预测,我们要保证分支预测的准确性,准确性的判断一般由交叉熵来衡量.每个分支的交叉熵在反向传播过程中都会影响其路径上的参数,而在单个模型中,这些路径存在重叠(这些重叠的部分称为共享层),因此会导致共享层的内部参数同时,受多个分支影响.同时,由于各分支的路径长度不一致,因此它们在训练中占有的比重也应有所不同.所以,使用一个带权重的组合交叉熵来保证各分支的预测准确率:

$$L_{BE}(P,y;\theta)=\sum_{n\in N}\omega_n L(p_n,y;\theta).$$

其中,$P=\{p_n\}$,$n\in N$ 是所有分支预测的集合;ω_n 是各分支的权重,一般由路径长度的比例决定,或者按照实践经验重新设定.

- 第2部分为提高预测的置信度.

在有标签的交叉熵计算中,标签是一个独热码(one-hot vector),所以计算结果省略了非正确标签的信息,而最终输出的置信度也是可以评价模型精度的一种度量.香农熵可以对信息量化,熵值越大,表示变量的不确定越大,使用其来作为置信度的计算,可以使训练过程考虑到输出的总体分布带来的不确定因素影响.虽然本文改进模型的最终输出是均化值,但分支长度的不同依旧需要考虑,所以此阶段的置信度计算对象结合了权重信息:

$$P^*=\frac{1}{N}\sum_{n\in N}\omega_n p_n\ .$$

- 第3部分为定义多样性的度量.

本方法为使整体多样性提升不影响正确标签的分布,采用修改非极大预测分布的策略,非极大预测分布即为分支预测分布去除正确标签对应位后剩余的错误分类方向分布情况.将所有分支的非极大预测分布整合,构成一个 n–1 维的向量组 $\tilde{M}_{\backslash y}$.同样考虑分支长度不一致,各分支在 n–1 空间内的指向在受到相同力量的分离时,相对与中心分离的角度会不一致,所以在组合向量组 $\tilde{M}_{\backslash y}$ 中引入各分支的权重,此处权重可定为与路径长度比例一致,欲达到最优效果,需要经过角速度的计算获得:

$$\tilde{M}_{\backslash y}=(\omega_1\tilde{F}^1_{\backslash y},...,\omega_n\tilde{F}^n_{\backslash y},...,\omega_N\tilde{F}^N_{\backslash y}),$$

$$L_{DS}=\log(\det(\tilde{M}^T_{\backslash y}\tilde{M}_{\backslash y})).$$

其中,$\tilde{F}^n_{\backslash y}$ 为第 n 个分支应用非极大预测策略后,经 L2 标准化的预测向量;$\tilde{M}^T_{\backslash y}$ 是 $\tilde{M}_{\backslash y}$ 的转置矩阵;det 是矩阵行列式计算;log 为对数计算.

最终,对多分支单模型的整体多样性计算为

$$L_{ED}=L_{BE}+\gamma\cdot H(P^*)+\mu\cdot L_{DS}.$$

其中:$H(\cdot)$为香农熵计算公式;γ和μ为超参,负责权衡多样性的提高程度和准确率,达到最优.

2.3 训练过程

一般的模型训练过程中,只会使用最终输出和训练标签来计算误差.但本文方法给原模型添加了额外分支,为了发挥各个分支的作用,还需要将分支预测结果添加到误差计算过程中.针对改进后的模型结构,我们设计了算法1.在模型没有完全拟合或者还没有到达轮数上限时执行以下操作.

1) 获取各个分支的预测结果,保存到张量 p_k 中;
2) 计算所有分支预测结果的平均值,保存到最终结果 y_f 中;
3) 计算出模型的整体多样性,输入为各分支预测 p_k、已设定的整体多样性超参γ,μ以及每个分支的权重;

4) 使用 y_f 和训练标签 y 计算交叉熵,以提高模型的预测准确率;

5) 累加 L_{ED} 和 L_{CE} 作为模型此次推理的误差,并求出模型所有训练参数的梯度;

6) 依据计算出的梯度,使用优化器更新模型参数.

7) 当前训练轮数的值属于预先设定的修改学习率轮数集 Q 时,将学习率降低为上一阶段的十分之一.

算法 1. 基于多分支单模型的训练算法.

输入:训练数据 X,训练标签 Y,批量大小 m,出口数量 K,优化器参数 α,轮数上限 N,需要修改学习率的轮数集 Q,整体多样性超参 γ,μ,ω;

输出:分类器 C.

While θ has not converged and $n<N$ **do**:
  $p_k \leftarrow F_k(\theta,x), k\in K$ //获取分支预测结果
  $y_f \leftarrow average(p_1,\ldots,p_K)$ //对各分支结果做平均计算
  $L_{ED} \leftarrow L_{ED}(y,p_1,\ldots,p_K)$ //计算整体多样性
  $L_{CE} \leftarrow L_{CE}(y,y_f)$ //计算交叉熵
  $g_\theta \leftarrow \nabla_\theta(L_{ED}+L_{CE})$ //计算训练参数梯度
  $\theta \leftarrow Optimizer(g_\theta,\theta,\alpha)$ //更新模型参数
  If n in Q:
    $\alpha \leftarrow \alpha/10$ //降低学习率
End while

3 实 验

为了验证本方法对单模型鲁棒性提高的有效性及可用性,需要回答如下几个问题.

- 问题 1:本方法是否能够提升模型的防御能力?即:改进后的模型在受到攻击的情况下是否能够比以前的模型得到更好的分类正确率?防御能力是否只是针对某种特定攻击方式有效?
- 问题 2:本方法会否损害原模型的分类效果?即,改进后的模型是否仍能对原始的干净样本正确分类?
- 问题 3:本方法中,对分支的特征融合以及在训练过程中加入的整体多样性正则项是否必须同时满足?
- 问题 4:本方法是否能与其他的模型防御方法组合使用,并且不会影响防御效果?

3.1 实验设计

为了回答以上 4 个问题,我们进行了如下设计.

- 首先构造要对比的模型.

根据问题 1 和问题 2,只需要对比初始原模型(记为 B 模型)和使用本文方法改造后的模型(记为 F+D 模型);为了回答问题 3,需要额外设计只使用了特征融合的模型(记为 F 模型)和只使用了整体多样性的模型(记为 D 模型);为了回答问题 4,我们选择了一种常用的模型防御方法——对抗训练,它有助于使神经网络的函数从接近线性变化转化为局部近似恒定,从而可以灵活地捕获到训练数据中的线性趋势,同时学习抵抗局部扰动,是一种广泛使用的模型防御能力再提升方法.实验需要对比只使用本文方法改造后的模型(记为 F+D 模型)和综合使用本文方法与对抗训练的模型(记为 AdvT+F+D 模型).这些模型的具体构造方式如下:B 模型选择了 Resnet-32 模型,不对其做任何改动;F 模型是对 B 模型加入了额外分支和分支之间的特征金字塔,但在训练过程中不考虑分支的整体多样性;D 模型是在 B 模型中加入分支,并在训练过程中加入分支的整体多样性正则项,但不添加分支之间的特征融合部分;F+D 模型则是综合以上两种方法,对原模型加入额外分支出口以及分支间的特征融合,同时在训练过程中使用整体多样性的正则项;AdvT+F+D 模型是在 F+D 模型的基础上,综合使用对抗训练而得到的模型.

- 其次,选择攻击方式以及对应的参数.

白盒攻击是在攻击者已知分类器的全部信息,包括训练数据、模型结构和权重的情况下,生成对抗样本攻

击神经网络,相对于黑盒攻击而言具有更强的攻击能力.因此,使用白盒攻击来验证提高模型鲁棒性带来的防御能力具有更大的说服力.本文选择了 7 种流行的白盒攻击方式:快速梯度符号法(fast gradient sign method,简称 FGSM)[22]、基本迭代法(basic iterative method,简称 BIM)[23]、投影梯度下降(project gradient descent,简称 PGD)[24]、动量迭代法(momentum iterative method,简称 MIM)[25]、基于雅可比矩阵的显著图攻击(Jacobian-based saliency map attack,简称 JSMA)[26]、Carlini & Wagner(C&W)[27]、弹性网络攻击(elastic-net attack,简称 EAD)[28].此外,为了观察模型在不同攻击力度下防御表现的变化,每种攻击方式都设置了 3 种参数.

接下来选择实验使用的数据集.我们选取了图像分类领域两个最常用的数据集 MNIST 和 CIFAR-10.MNIST 是一个黑白手写数字数据集,包含 0~9 这 10 类来自 250 人的手写数字,图片尺寸为 28×28.其中,训练集图片数量为 60 000 张,测试集图片数量为 10 000 张;CIFAR-10 是一个更接近于普适物体的彩色图像数据集,包含飞机、汽车、鸟类、猫、鹿、狗、蛙类、马、船和卡车这 10 类数据,图片尺寸为 32×32.其中,训练集图片数量为 50 000 张,测试集图片数量为 10 000 张.

本文认为:使用这两种数据集,能成功验证本方法的可行性和有效性.

为了回答问题 1 和问题 3,我们设计了 2(数据集)×7(攻击方式)×3(攻击参数)×4(模型)=168 组对比实验,在两个数据集上分别使用上述 7 种攻击方式,对每种攻击方式设置 3 种攻击参数调节攻击力度,对比上述 4 种模型的分类正确率.

为了回答问题 2 和问题 3,我们设计了 2(数据集)×4(模型)=8 组对比实验,在两个干净的数据集上对比上述 4 种模型的分类准确率.同时,使用 T-SNE 视图[29]对各个模型的输出降维可视化,进一步佐证了本方法的可用性.

为了回答问题 4,我们设计了 2(数据集)×4(攻击方式)×3(攻击参数)×2(模型)=48 组对比实验,在两个数据集上分别使用 FSGM,BIM,MIM,PGD 这 4 种攻击方式.之所以使用这 4 种方式,是因为它们具有相似的攻击原理,同样,对每种攻击方式设置 3 种攻击参数,对比 F+D 模型和 AdvT+F+D 模型的防御表现.

3.2 实验过程

本次实验所使用的 CPU 型号为 Intel i7 9700k,使用的图形处理器型号为 Nvidia RTX 2080Ti,操作系统为 Linux 18.04,Python 版本为 3.7,机器学习平台为 Tensorflow v1.12[30]以及 Keras v2.4.

在训练模型之前,我们首先对数据集进行了归一化预处理:将所有的训练样本都归一化到 0-1 范围内.同时,为达到更好的训练效果并降低训练出来的模型的过拟合程度,在训练过程中也使用了数据增广技术,对原始图像样本分别进行水平翻转、水平平移和竖直平移操作.在平移过程中,使用常量 0 填充超出边界的部分.

实验选择的 B 模型为深度残差网络 Resnet-32,它的基本结构如图 2(a)所示,它包含 3 组通道数依次为 16,32,64 的残差块:第 1 组残差块由 5 个恒等残差块构成,第 2 组、第 3 组残差块均由 1 个卷积残差块和 4 个恒等残差块构成.恒等残差块和卷积残差块的差别在于块中残差分支是否做卷积操作,所以卷积残差块会改变特征图的大小,从而满足每个阶段特征图尺寸缩小的需求.这 3 个残差块对应的特征图尺寸分别为 32,16,8.最后,通过一个全连接层输出 10 分类结果.

F 模型的结构如图 2(b)所示,它是在 B 模型的基础上,在第 1 组和第 2 组残差块之后分别加入额外分支,算上原出口,改动后的模型存在 3 个出口.在模型中加入特征金字塔,实现特征融合的过程分为 3 步.

- 第 1 步在第 3 组残差块的特征图做上采样后,达到与第 2 组残差块的特征图同一尺寸,再采用横向连接的方法与第 2 组的特征图融合;
- 第 2 步对第 1 步得到的融合后特征图再做上采样到与第 1 组残差块的特征图同一尺寸,再次进行特征融合,并在融合过程中使用小卷积,以保证 3 部分特征图的通道数一致;
- 最后对 3 个分支都进行小尺寸卷积核的特征压缩,统一生成 8×8 尺寸 128 通道的特征图,然后对其进行全局平均池化,生成 1×128 的向量,分别经过全连接层得到最后的 10 分类预测(包含 softmax 变换),最终输出结果是 3 个预测结果的平均值.

D 模型的结构如图 2(c)所示,它同样在 B 模型的第 1 组和第 2 组残差块之后加入额外分支,与 F 模型不同的是:它并不会对各个出口输出的特征图做特征融合,而是直接通过全连接层得到一个 10 分类预测结果.为了

削减共享层的影响,使训练出来的各个分支的预测结果不过于趋同,我们添加了整体多样性计算:使用了一个带权重的组合交叉熵来保证了各分支的预测准确率,根据出口深度,由浅至深权重依次设置为 1~3;最后,使用第 3.2 节提出的单模型内多分支预测的整体多样性 L_{ED} 的计算公式为输出结果添加了整体多样性计算,公式中,超参数γ和μ的值分别设为 1 和 0.01.

F+D 模型的结构如图 2(d)所示,它综合了 F 模型与 D 模型的改造方法,对各个出口输出的特征图做特征融合之后再输入到全连接层,得到一个 10 分类预测结果;然后,对 3 个出口的预测结果进行整体多样性计算,参数设置与 D 模型相同.AdvT+F+D 模型的结构与 F+D 模型一致,它与 F+D 模型的区别只在于训练过程中动态地生成对抗样本用作数据扩容,设置对抗样本和正常样本的比例为 1:1,计算模型参数梯度的误差由对抗样本误差和正常样本误差累加获得.

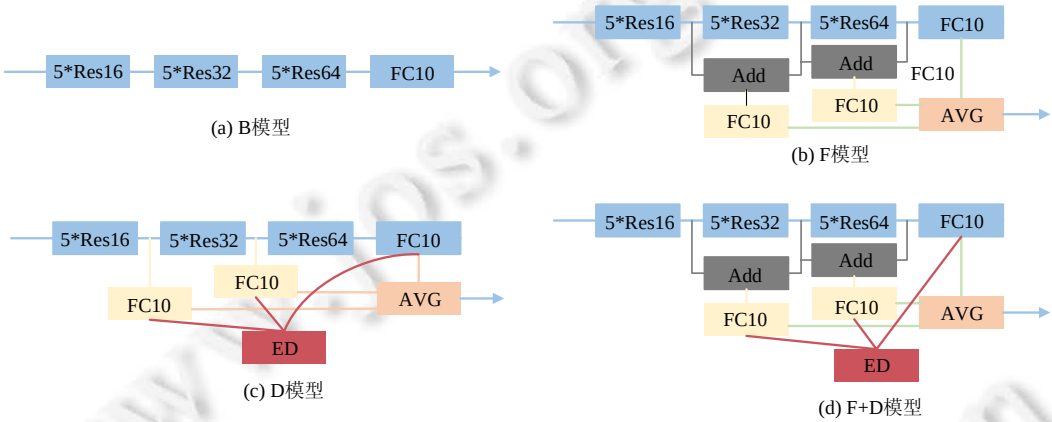


Fig.2 Four models used in the experience

图 2 实验中使用的 4 种模型

在上述模型的训练过程中,我们将初始学习率α设置为 0.001.指数衰减率β_1,β_2 分别控制之前的时间步的梯度动量和梯度平方动量的影响情况.为了使对比结果更公平,将它们设置为领域内默认值.其中,β_1 设置为 0.9,β_2 设置为 0.999.在 MNIST 训练集上的训练轮数为 40,但会在 20 轮之后,将学习率降低为初始的十分之一,直至降到 $10e^{-4}$ 数量级;在 CIFAR-10 数据集上训练轮数为 180,分别在 80,120,160 轮之后降低学习率为前一时刻的十分之一,直至降到 $10e^{-6}$ 数量级.批量(batchsize)设置为 128.此外,由于在训练过程中损失函数表示的误差不可能归为 0,若出现,则意味着模型过拟合了,所以我们使用了基于验证集准确率的模型保存机制.

实验中,对抗样本生成使用框架为 cleverhans v2.1.0[31],cleverhans 是谷歌基于 Tensorflow 开发的,集成了大多数现有对抗样本生成方法.攻击过程为针对要攻击的模型图结构,cleverhans 按照所需攻击方法和设定的攻击力度,生成对应的数据流图,输入原始样本后,输出生成对抗样本.防御实验中,CIFAR-10 数据集上 FGSM 扰动力度为 0.01~0.04;BIM 等 3 种迭代方法扰动力度设计为 0.01~0.03;JSMA 设计扰动力度为 0.1,而攻击像素点占比为 5%~15%;C&W 攻击力度为 0.001,0.01 和 0.1;EAD 下,$L1$ 正则的超参为 0.01,攻击力度为 0.1,1,5.MNIST 数据集上,FGSM 扰动力度为 0.1~0.3;BIM 等 3 种迭代方法扰动力度设计为 0.05~0.15;JSMA 设计扰动力度为 0.2,而攻击像素点占比为 10%~40%;C&W 攻击力度为 0.1,1 和 5;EAD 下,$L1$ 正则的超参为 0.01,攻击力度为 1,5,10.

3.3 实验结果分析

3.3.1 对于对抗样本的处理能力

为回答问题 1 和问题 3,首先在 CIFAR-10 数据集和 MNIST 数据集上测试 B 模型、F 模型、D 模型以及 F+D 模型对于 7 种常用对抗样本生成方法下白盒攻击的防御效果.除每种攻击方法的扰动参数设定之外,设置 BIM,MIM 和 PGD 这 3 种攻击的迭代次数为 10,设置 C&W 和 EAD 攻击的迭代次数为 1 000,且学习率为 0.01.实验结果记录在表 1 中.

Table 1 Comparison of classification accuracy against adversarial examples on CIFAR-10 and MNIST (%)

表 1 CIFAR-10 和 MNIST 数据集上对于对抗样本分类正确率比较 (%)

数据集	攻击类型	攻击参数	B 模型	F 模型	D 模型	F+D 模型
CIFAR-10	FGSM	ε=0.01	20.62	33.28	46.51	**64.32**
		ε=0.02	13.64	24.21	32.52	**60.28**
		ε=0.04	9.84	19.25	18.12	**49.44**
	BIM	ε=0.01	6.5	9.35	18.79	**42.26**
		ε=0.02	5.76	5.74	10.46	**32.42**
		ε=0.03	5.75	5.48	8.95	**27.58**
	MIM	ε=0.01	7.5	10.98	24.74	**47.82**
		ε=0.02	5.8	5.84	11.47	**38.07**
		ε=0.03	5.76	5.52	8.93	**32.33**
	PGD	ε=0.01	7.78	12.01	22.82	**43.96**
		ε=0.02	5.37	5.8	10.43	**31.57**
		ε=0.03	4.84	5.01	7.47	**24.14**
	JSMA	θ=0.1, γ=0.05	11.1	18	38.6	**45.3**
		θ=0.1, γ=0.1	3.1	7.8	17.2	**32.4**
		θ=0.1, γ=0.15	2.3	7.1	8.6	**26.4**
	C&W	c=0.001	38.2	28.3	**69.05**	66.3
		c=0.01	5.75	5.6	**48.75**	47.9
		c=0.1	5.5	5.4	23.1	**30.8**
	EAD	β=0.01, c=0.1	73.7	36.9	88	**89.9**
		β=0.01, c=1	5.55	4.75	61.8	**69.2**
		β=0.01, c=5	2.3	2.75	15.65	**37.15**
MNIST	FGSM	ε=0.1	49.69	71.61	26.11	**94.84**
		ε=0.2	13.21	20.68	11.03	**65.46**
		ε=0.3	5.42	11.65	9.77	**20.68**
	BIM	ε=0.05	91.4	95.42	80.18	**95.76**
		ε=0.1	21.99	54.6	15.67	**87.8**
		ε=0.15	1.28	10.88	7.47	**72.84**
	MIM	ε=0.05	92.46	95.7	84.41	**96.38**
		ε=0.1	32.84	62.83	17.68	**90.89**
		ε=0.15	4.3	17.84	9.41	**79.84**
	PGD	ε=0.05	91.74	96.31	69.02	**95.98**
		ε=0.1	7.31	51.02	7.17	**75.89**
		ε=0.15	0.18	8.04	1.82	**38.61**
	JSMA	θ=0.2, γ=0.1	71.4	78.2	54.2	**86.8**
		θ=0.2, γ=0.2	30.6	52.4	32.4	**62.7**
		θ=0.2, γ=0.4	15.6	28.8	16.2	**30.6**
	C&W	c=0.1	60.9	89.4	93.1	**97.4**
		c=1	0.55	2.8	30.05	**87.55**
		c=5	0.55	0.8	3.25	**38.35**
	EAD	β=0.01, c=1	77.1	82.65	98.3	**98.8**
		β=0.01, c=5	0.65	6.45	69.35	**95.35**
		β=0.01, c=10	0.55	2.4	36.55	**93.5**

表 1 的第 1 部分为在 CIFAR-10 数据集上的实验结果.

- B 模型在这 7 种攻击方式下的分类表现都受到了很大的影响,在扰动较低的情况下,准确率大幅下降;扰动较高的情况下,准确率甚至只有个位数水平;
- F 模型仅在 FGSM 和 JSMA 这两种攻击方式下,分类准确率略有提升;但是对于其他的攻击方式,防御效果并不明显.在 C&W 和 EAD 这两种比较相似的攻击方式下,准确率下降的跨度甚至超过了 B 模型;
- D 模型相比于前两种模型而言,对所有类型对抗样本都有提高.在 7 种攻击方式下,准确率都达到了 B 模型的两倍以上.其中:对 JSMA 和 C&W 攻击的防御表现提升了 3~4 倍,对高扰动的 EAD 攻击防御效果甚至达到了 B 模型的 7 倍左右;
- 最后一列记录了应用本文提出的单模型鲁棒性提高方法后形成的 F+D 模型的防御结果.在前 4 种攻击方法下,该模型准确率相比于 D 模型都有成倍的提高;同时,3 种不同扰动值间的下降幅度也远小于 D 模型;后 3 种攻击方式下的防御表现完美继承了 D 模型的优势,JSMA 和 EAD 攻击下,面对各种扰动值都进一步提高,面对 EAD 高扰动攻击的分类准确率更是达到了 B 模型的 15 倍以上;对 C&W 两种小扰

动的防御表现略差于 D 模型,但是实验中的最高扰动下出现了反超,表明 F+D 模型在扰动提高时防御表现的下降趋势比较平缓.

表 1 中,第 2 部分为 MNIST 数据集上实验结果.在进行此部分实验时,由于 MNIST 数据集的样本结构比较简单,所以各种攻击方式的扰动范围较于 CIFAR-10 数据集上有大幅提高,而迭代次数和 C&W 的学习率和前部分实验相同.在如此高的扰动范围下.

- B 模型对对抗样本的分类准确率都大幅下降,特别在实验中设定的第 1 种、第 2 种扰动值间,出现了 3~4 倍的极速下滑;甚至在 C&W 和 EAD 这两种有目标的攻击方式下,受较高扰动攻击后的分类准确率下降到了 1 以下;
- F 模型的防御表现在前 5 种攻击方式下都有较大提升,特别是面对扰动幅度变大情况,下降趋势相对平缓;而对 C&W 和 EAD 两种相似攻击的防御表现只有略微提高;
- D 模型对 FGSM 等前 5 种对抗样本的防御表现虽然略高于 B 模型,却比 F 模型又有下降;而在 C&W 和 EAD 两种攻击下表现良好,抑制住了不同幅度间快速下降的趋势;
- 而本文方法得到的 F+D 模型在所有攻击方式下的防御表现都得到了非常大的提升,多种攻击下的表现提高了超过 60 个点;甚至在 C&W 和 EAD 这两种 B 模型表现极差的情况下,几乎保持了对原始样本的分类准确率,不同扰动幅度间的下降趋势也更加平缓.

观察整张表发现:面对 C&W 和 EAD 攻击,整体多样性可以提供更好的防御效果;而 FGSM 等前 5 种攻击方法则会受到样本复杂程度的影响.在相对复杂的 CIFAR-10 数据集上,整体多样性带来的提升高于特征融合;而在相对简单的 MNIST 数据集上,特征融合会提供更好的帮助.本文方法结合特征融合和整体多样性,最终达到了 1+1 大于 2 的优秀表现.在整个实验中,一直保持较优的防御表现,完美解答了问题 3.在两种测试集下进行对 4 种模型的白盒攻击防御实验中,回答了问题 1,证明本文提出的方法可以大幅提升模型的鲁棒性,对常见对抗样本攻击方法均可做出有效防御.

3.3.2 对于干净样本的处理能力

为回答问题 2,实验评估了 B 模型、F 模型、D 模型以及 F+D 模型在干净测试集上面的表现,实验结果记录在表 2 中.其中,

- B 模型在 MNIST 和 CIFAR-10 的识别率分别为 99.59%和 91.17%,达到了现有的深度神经网络分类器的基本水平;
- F 模型在两个数据集上的分类准确率则是 99.65%和 91.41%,可以看出:特征融合使得最终分类可以同时考虑语义特征和细节特征,模型能达到更好的精度;
- D 模型没有加入特征融合,考虑了整体多样性,但由于是在一个模型内,各分支路线有共享层,可以预见,分类准确率会受到一定的影响,最终结果 99.53%和 89.14%,也符合预期;
- F+D 模型作为本文方法改进并训练的模型,在 CIFAR-10 数据集上的分类准确率达到了 91.05%,比 D 模型表现优秀,较于 B 模型也基本没有准确率的下降,甚至在 MNIST 数据集上的表现超过了 F 模型,达到了 99.7%.上述结果表明,本文方法改进并训练的模型仍能保证对原始样本的分类精度.

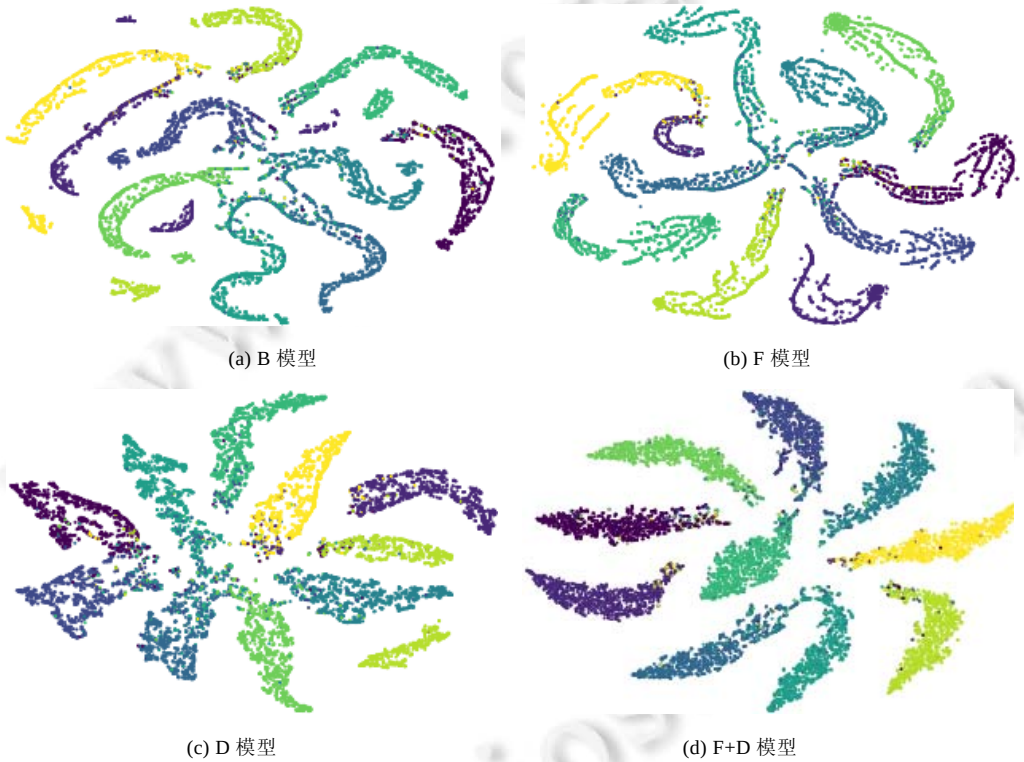
Table 2 Comparison of classification accuracy on clean examples from MNIST and CIFAR-10 (%)

表 2 MNIST 和 CIFAR-10 上干净样本分类准确率比较 (%)

数据集	B 模型	F 模型	D 模型	F+D 模型
CIFAR-10	91.17	**91.41**	89.14	91.05
MNIST	99.59	99.65	99.53	**99.7**

为更好地展现整体多样性的效果,实验打印了各模型最终输出的 T-SNE 视图.T-SNE 利用条件概率表示相似性,使用相对熵训练,可将高维分布的点映射到低维空间中,明确地显示出输入的聚类状况.图 3 绘制了对比的 4 种模型最终输出在低维的分布情况.图中每一种颜色代表一种分类,此实验在 CIFAR-10 验证集(10 000 张)上进行,所以颜色数目为 10.从图中可以看出:在同类的点中间参杂其他颜色的点,表示这些点是分类出错的部分.

因为本实验输入是原始样本,可以看出,图中的错误点仅是少数.图 3(a)是 B 模型输出的 T-SNE 视图,可以看出:每种分类并没有完全聚集在一起,会有部分零散分布在其他位置,总体分布得杂乱无章,甚至有几类出现了交融.从这里可以认为 B 模型对对抗样本很敏感,符合前一实验的结果.图 3(b)是 F 模型输出的视图,可以看出:在同时利用了语义特征和细节特征之后,各类自己的聚集相对基本模型已经有了改善,但是依旧存在不同类交叉的问题,分离程度不足.图 3(c)是 D 模型输出的视图,可以看出:各分类之间已经有了分离的趋势,但错误的点也明显增加.推测是由于在一个模型内,因为各分支存在共同层,多样性的加入可能影响了拟合,所以准确率有所下降.图 3(d)则是本文方法改进并训练的模型,同时引入了特征金字塔和改进的整体多样性,相比于前几种的结果,同类之间的聚集程度得到了极大的提高;且不同类之间有明显的分离,错误的点基本没有增加,符合表 2 中的实验结果.这部分同样证明了本文方法成功地提高了模型的鲁棒性,并且基本没有影响模型对原始样本精度.



(a) B 模型 (b) F 模型

(c) D 模型 (d) F+D 模型

Fig.3 T-sne views of the final output from each model on CIFAR-10

图 3 CIFAR-10 测试集上各模型最终输出的 T-SNE 视图

3.3.3 与对抗训练方法组合使用的效果

为回答问题 4,测试了 F+D 模型以及额外加入了对抗训练后的 AdvT+F+D 模型在受到相同攻击下的分类正确率,结果记录在表 3 中.CIFAR-10 和 MNIST 上的对抗模型训练都使用 PGD 方法,对抗训练过程中,CIFAR-10 数据集下设置 PGD 的扰动值为 0.01~0.05 随机采样,MNIST 数据集下设置 0.05~0.2 随机采样.随后测试了与前部分相同参数的 FGSM,BIM,MIM 和 PGD 这 4 种攻击.实验结果表明:在使用对抗训练之后,模型的防御表现进一步提高.其中,使用 PGD 作为训练中的扩容方式时提高最为明显:CIFAR-10 下,准确率都提升了近 1 倍;而 MNIST 下,对于 0.15 扰动攻击,更是近 3 倍的提高.BIM 和 MIM 除基本原理相似外,与 PGD 同样使用迭代方法,测试中,防御表现都有一定程度的提高.

- MNIST 下,基本达到了对干净样本的辨识水平;FGSM 攻击下,对于前两种较小的扰动情况都有提高;
- 但是同时,在 CIFAR-10 和 MNIST 下,对第 3 种较大扰动出现准确率下降的情况.我们认为:相比于迭代

的攻击方式,FGSM 在大扰动下对图片的破坏情况相对严重;而使用 PGD 攻击方式做对抗训练,模型达到的局部恒定比较适合图片未被严重破坏的情况.

总体上,本文方法改进并训练的模型在对抗训练前后防御表现有提高,可证明本文方法不与对抗训练冲突.

Table 3 Comparison of classification accuracy against corresponding adversarial examples before and after using adversarial training with PGD (%)

表 3 PGD 对抗训练前后对相似对抗样本的分类正确率的比较 (%)

数据集	攻击类型	攻击参数	F+D 模型	AdvT+F+D 模型
CIFAR-10	FGSM	ε=0.01	64.32	**75.82**
		ε=0.02	60.28	**62.79**
		ε=0.04	**49.44**	46.17
	BIM	ε=0.01	42.26	**75.17**
		ε=0.02	32.42	**58.81**
		ε=0.03	27.58	**45.62**
	MIM	ε=0.01	47.82	**75.53**
		ε=0.02	38.07	**60.23**
		ε=0.03	32.33	**48.33**
	PGD	ε=0.01	43.96	**78.4**
		ε=0.02	31.57	**66.33**
		ε=0.03	24.14	**54.61**
MNIST	FGSM	ε=0.1	94.84	**98.89**
		ε=0.2	65.46	**97.59**
		ε=0.3	**20.68**	12.36
	BIM	ε=0.05	95.76	**99.05**
		ε=0.1	87.8	**98.85**
		ε=0.15	72.84	**98.16**
	MIM	ε=0.05	96.38	**99.05**
		ε=0.1	90.89	**98.85**
		ε=0.15	79.84	**98.2**
	PGD	ε=0.05	95.98	**99.09**
		ε=0.1	75.89	**98.96**
		ε=0.15	38.61	**98.69**

4 结论与展望

针对深度神经网络对于对抗样本的脆弱性问题,本文提出了一种基于特征融合和整体多样性的单模型鲁棒性提升方法.该方法受组合模型防御效果优于单模型的启发,依据分支网络中浅层出口也可以达到较好的预测准确率理论,在现有模型基础上添加额外的分支模拟组合模型效果,同时在分支之间加入特征融合实现特征金字塔,并引入改进后的多分支单模型整体多样性计算辅助训练,以提高模型鲁棒性,使其具有更好的防御能力.通过在 MNIST 和 CIFAR-10 两种数据集上的实验结果表明:本文方法改进并训练的模型防御效果显著,对对抗样本的防御能力比改进前的原模型在 FGSM 等 4 种基于梯度的攻击下有 5 倍以上的提高,JSMA,C&W 以及 EAD 攻击下可达到 10 倍的提升;同时不干扰对干净样本的分类精度,也与对抗训练方法不抵触,可以联合使用,获得更好的防御效果.证明了本文提出的提升鲁棒性方法是可行且有效的.此外,实验中还发现:在不同复杂度的样本上,特征融合和整体多样性带来的鲁棒性影响不同.在今后的工作中,我们会对此方面做深入的研究,以改进本文提出的方法,获得更好的效果.

References:

[1] Lueth KL. State of the IoT 2018: Number of IoT devices now at 7B—Market accelerating. IOT ANALYTICS. https://iot-analytics.com/state-of-the-iot-update-q1-q2-2018-number-of-iot-devices-now-7b/

[2] Dourado Jr CM, da Silva SP, da Nóbrega RV, Barros AC, Rebouças Filho PP, de Albuquerqu VH. Deep learning IoT system for online stroke detection in skull computed tomography images. Computer Networks, 2019,152:25–39.

[3] Mookherji S, Sankaranarayanan S. Traffic data classification for security in IoT-based road signaling system. In: Proc. of the Soft Computing in Data Analytics. 2019. 589–599.

[4] Rodrigues JD, Rebouças Filho PP, Peixoto Jr E, Kumar A, de Albuquerque VH. Classification of EEG signals to detect alcoholism using machine learning techniques. Pattern Recognition Letters, 2019,125:140−149.

[5] Zhang Y, Li PS, Wang XH. Intrusion detection for IoT based on improved genetic algorithm and deep belief network. IEEE Access, 2019,7:31711−31722.

[6] Athalye A, Carlini N, Wagner D. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In: Proc. of the Int'l Conf. on Machine Learning (ICML). 2018. 274−283.

[7] Lu J, Issaranon T, Forsyth D. Safetynet: Detecting and rejecting adversarial examples robustly. In: Proc. of the 2017 IEEE Int'l Conf. on Computer Vision (ICCV). 2017. 446−454.

[8] Metzen JH, Genewein T, Fischer V, Bischoff B. On detecting adversarial perturbations. In: Proc. of Int'l Conf. on Learning Representations (ICLR). 2017.

[9] Carlini N, Wagner D. Adversarial examples are not easily detected: Bypassing ten detection methods. In: Proc. of the 10th ACM Workshop on Artificial Intelligence and Security. 2017. 3−14.

[10] Liao FZ, Liang M, Dong YP, Pang TY, Zhu J, Hu XL. Defense against adversarial attacks using high-level representation guided denoiser. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR). 2018. 1778−1787.

[11] Pang TY, Xu K, Du C, Chen N, Zhu J. Improving adversarial robustness via promoting ensemble diversity. In: Proc. of Int'l Conf. on Machine Learning (ICML). 2019. 4970−4979.

[12] Teerapittayanon S, McDanel B, Kung H. BranchyNet: Fast inference via early exiting from deep neural networks. In: Proc. of the IEEE Int'l Conf. Pattern Recognition (ICPR). 2016. 2464−2469.

[13] Lin TY, Dollár P, Girshick R, He K, Hariharan B, Belongie S. Feature pyramid networks for object detection. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR). 2017. 936−944.

[14] He KM, Zhang XY, Ren SQ, Sun J. Deep residual learning for image recognition. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR). 2016. 770−778.

[15] Ranjan R, Sankaranarayanan S, Castillo CD, Chellappa R. Improving network robustness against adversarial attacks with compact convolution. arXiv preprint arXiv:1712.00699, 2017.

[16] Miyato T, Maeda SI, Koyama M, Ishii S. Virtual adversarial training: A regularization method for supervised and semi-supervised learning. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2018,41(8):1979−1993.

[17] Kurakin A, Goodfellow I, Bengio S. Adversarial machine learning at scale. In: Proc. of the Int'l Conf. on Learning Representations (ICLR). 2017.

[18] Kurakin A, Goodfellow I, Bengio S, Dong YP, Liao FZ, Liang M, Pang TY,Zhu J, Hu, XL, Xie CH, *et al*. Adversarial attacks and defences competition. In: Proc. of the NIPS 2017 Competition: Building Intelligent Systems. Cham: Springer-Verlag, 2018. 195−231.

[19] Samangouei P, Kabkab M, Chellappa R. Defense-Gan: Protecting classifiers against adversarial attacks using generative models. In: Proc. of the Int'l Conf. on Learning Representations (ICLR). 2018.

[20] Guo C, Rana M, Cisse M, Van Der Maaten L. Countering adversarial images using input transformations. In: Proc. of the Int'l Conf. on Learning Representations (ICLR). 2018.

[21] Lamb A, Binas J, Goyal A, Serdyuk D, Subramanian S, Mitliagkas I, Bengio Y. Fortified networks: Improving the robustness of deep networks by modeling the manifold of hidden representations. arXiv preprint arXiv:1804.02485, 2018.

[22] Goodfellow IJ, Shlens J, Szegedy C. Explaining and harnessing adversarial examples. In: Proc. of the Int'l Conf. on Learning Representations (ICLR). 2015.

[23] Kurakin A, Goodfellow IJ, Bengio S. Adversarial examples in the physical world. In: Proc. of the Int'l Conf. on Learning Representations (ICLR) Workshop. 2017.

[24] Madry A, Makelov A, Schmidt L, Tsipras D, Vladu A. Towards deep learning models resistant to adversarial attacks. In: Proc. of the Int'l Conf. on Learning Representations (ICLR). 2018.

[25] Dong YP, Liao FZ, Pang TY, Su H, Hu XL, Li JG, Zhu J. Boosting adversarial attacks with momentum. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR). 2018. 9185−9193.

[26] Papernot N, McDaniel P, Jha S, Fredrikson M, Celik ZB, Swami A. The limitations of deep learning in adversarial settings. In: Proc. of the 2016 IEEE European Symp. on Security and Privacy (EuroS&p). 2016. 372−387.

[27] Carlini N, Wagner D. Towards evaluating the robustness of neural networks. In: Proc. of the 2017 IEEE Symp. on Security and Privacy (S&P). 2017. 39−57.

[28] Chen PY, Sharma Y, Zhang H, Yi JF, Hsieh CJ. Ead: Elastic-net attacks to deep neural networks via adversarial examples. In: Proc. of the AAAI Conf. on Artificial Intelligence (AAAI). 2018. 10−17.

[29] Maaten L, Hinton G. Visualizing data using t-SNE. Journal of Machine Learning Research, 2008,9:2579−2605.

[30] Abadi M, Barham P, Chen J, Chen Z, Davis A, Dean J, Devin M, Ghemawat S, Irving G, Isard M, Kudlur M. Tensorflow: A system for large-scale machine learning. In: Proc. of the 12th USENIX Symp. on Operating Systems Design and Implementation (OSDI). 2016. 265−283.

[31] Papernot N, Faghri F, Carlini N, Goodfellow I, Feinman R, Kurakin A, Xie C, Sharma Y, Brown T, Roy A, Matyasko A. Technical report on the cleverhans v2.1.0 adversarial examples library. arXiv preprint arXiv:1610.00768, 2016.

韦璠(1996－),男,硕士生,CCF 学生会员,主要研究领域为深度学习,对抗样本防御.

宋云飞(1994－),男,硕士生,CCF 学生会员,主要研究领域为人工智能安全.

邵明莉(1997－),女,硕士生,CCF 学生会员,主要研究领域为深度学习.

刘天(1988－),男,博士生,CCF 学生会员,主要研究领域为新型非易失性存储,嵌入式系统,机器学习.

陈小红(1982－),女,博士,副教授,CCF 专业会员,主要研究领域为需求工程,形式化方法.

王祥丰(1987－),男,博士,副教授,CCF 专业会员,主要研究领域为分布式优化,多智能体强化学习.

陈铭松(1982－),男,博士,教授,博士生导师,CCF 高级会员,主要研究领域为信息物理融合系统设计自动化,计算机体系结构,物联网技术,形式化方法.