计算机研究与发展
Journal of Computer Research and Development
DOI:10.7544/issn1000-1239.2018.20170649
55(6):1117-1142,2018

深度学习应用于网络空间安全的现状、趋势与展望

张玉清[1,2] 董 颖[1] 柳彩云[1] 雷柯楠[1,2] 孙鸿宇[1,2]
[1](中国科学院大学国家计算机网络入侵防范中心 北京 101408)
[2](西安电子科技大学网络与信息安全学院 西安 710071)
(zhangyq@nipc.org.cn)

Situation, Trends and Prospects of Deep Learning Applied to Cyberspace Security

Zhang Yuqing[1,2], Dong Ying[1], Liu Caiyun[1], Lei Kenan[1,2], and Sun Hongyu[1,2]
[1](*National Computer Network Intrusion Protection Center, University of Chinese Academy of Sciences, Beijing* 101408)
[2](*School of Cyber Engineering, Xidian University, Xi'an* 710071)

Abstract Recently, research on deep learning applied to cyberspace security has caused increasing academic concern, and this survey analyzes the current research situation and trends of deep learning applied to cyberspace security in terms of classification algorithms, feature extraction and learning performance. Currently deep learning is mainly applied to malware detection and intrusion detection, and this survey reveals the existing problems of these applications: feature selection, which could be achieved by extracting features from raw data; self-adaptability, achieved by early-exit strategy to update the model in real time; interpretability, achieved by influence functions to obtain the correspondence between features and classification labels. Then, top 10 obstacles and opportunities in deep learning research are summarized. Based on this, top 10 obstacles and opportunities of deep learning applied to cyberspace security are at first proposed, which falls into three categories. The first category is intrinsic vulnerabilities of deep learning to adversarial attacks and privacy-theft attacks. The second category is sequence-model related problems, including program syntax analysis, program code generation and long-term dependences in sequence modeling. The third category is learning performance problems, including poor interpretability and traceability, poor self-adaptability and self-learning ability, false positives and data unbalance. Main obstacles and their opportunities among the top 10 are analyzed, and we also point out that applications using classification models are vulnerable to adversarial attacks and the most effective solution is adversarial training; collaborative deep learning applications are vulnerable to privacy-theft attacks, and prospective defense is teacher-student model. Finally, future research trends of deep learning applied to cyberspace security are introduced.

Key words deep learning; cyberspace security; attacks and defenses; application security; network security

收稿日期:2017-09-06;**修回日期**:2018-01-17

基金项目:国家重点研发计划项目(2016YFB0800703);国家自然科学基金项目(61572460,61272481);信息安全国家重点实验室的开放课题(2017-ZD-01);国家发改委信息安全专项项目((2012)1424)

This work was supported by the National Key Research and Development Program of China (2016YFB0800703), the National Natural Science Foundation of China (61572460, 61272481), the Open Program of the State Key Laboratory of Information Security (2017-ZD-01), and the Special Program on Information Security of the National Development and Reform Commission of China ((2012)1424).

摘 要 近年来,深度学习应用于网络空间安全的研究逐渐受到国内外学者的关注,从分类算法、特征提取和学习效果等方面分析了深度学习应用于网络空间安全领域的研究现状与进展.目前,深度学习主要应用于恶意软件检测和入侵检测两大方面,指出了这些应用存在的问题:特征选择问题,需从原始数据中提取更全面的特征;自适应性问题,可通过 early-exit 策略对模型进行实时更新;可解释性问题,可使用影响函数得到特征与分类标签之间的相关性.其次,归纳总结了深度学习发展面临的十大问题与机遇,在此基础上,首次归纳了深度学习应用于网络空间安全所面临的十大问题与机遇,并将十大问题与机遇归为 3 类:1)算法脆弱性问题,包括深度学习模型易受对抗攻击和隐私窃取攻击;2)序列化模型相关问题,包括程序语法分析、程序代码生成和序列建模长期依赖问题;3)算法性能问题,即可解释性和可追溯性问题、自适应性和自学习性问题、存在误报以及数据集不均衡的问题.对十大问题与机遇中主要问题及其解决方案进行了分析,指出对于分类的应用易受对抗攻击,最有效的防御方案是对抗训练;基于协作性深度学习进行分类的安全应用易受隐私窃取攻击,防御的研究方向是教师学生模型.最后,指出了深度学习应用于网络空间安全未来的研究发展趋势.

关键词 深度学习;网络空间安全;攻击与防御;应用安全;网络安全

中图法分类号 TP393

近年来,硬件计算能力的强大和数据量的与日俱增,推动了深度学习(deep learning, DL)[1]的发展,使深度学习的实用性和普及性都有了巨大提升.深度学习是一种机器学习技术,其目的是通过经验和数据改进计算机系统,实现机器学习的原始目标:人工智能(artificial intelligence, AI).深度学习使用多个非线性特征变换,即多层感知机(multi-layer perceptron, MLP)构成的处理层来对数据进行表征学习(representation learning)[2-3].深度学习已经应用于计算机视觉[4-5]、语音识别[6-8]、自然语言处理[9]、生物医学[10]和恶意代码检测[11]等多个领域.

自 2015 年起,深度学习应用于网络空间安全(cyberspace security)的研究逐步涌现,引起学术界广泛关注.目前,深度学习主要应用于恶意软件检测和入侵检测两大网络空间安全领域,与传统的机器学习相比,深度学习提高了检测效率、降低了误报率.此外,深度学习算法摆脱了对特征工程的依赖,能够自动化智能化识别攻击特征,有助于发现潜在安全威胁.然而,现阶段学术界对于深度学习应用于网络空间安全的研究了解不够深入和全面,基于此,本综述对深度学习应用于网络空间安全的研究进行讨论,对其研究现状和研究发展趋势进行了分析,并对该领域的下一步研究方向进行了展望.

本文对深度学习应用于网络空间安全领域的研究进行了分类,针对恶意软件检测和入侵检测,主要从分类算法、特征提取和检测效果等方面总结并比较了这些应用.目前,这些应用采用的深度学习分类算法主要有 5 种,包括深度神经网络(deep neural network, DNN)、卷积神经网络(convolutional neural network, CNN)[12]、循环神经网络(recurrent neural network, RNN)[13]、深度信念网络(deep belief network, DBN)[14]和自编码器(autoencoder, AE)[15].同时,一些深度学习的网络空间安全应用也使用 RNN,AE,DBN 进行深度学习模型的特征提取.

目前,这些应用存在的最大问题是健壮性差[16],此问题广泛存在于使用深度学习算法进行分类的网络空间安全应用.比如,用于恶意软件检测和入侵检测的深度学习模型,这些模型基于深度学习算法进行分类检测,攻击者通过恶意构造对抗样本来对深度学习算法实施对抗攻击,使目标深度学习模型实现攻击者选择的特定输出[17].我们认为深度学习应用于网络空间安全的研究存在的第二大问题是机密性差,此问题存在于所有基于多方协作的深度学习模型(multi-party collaborative deep learning models)应用,协作性模型是由多个数据源提供方在保证数据私有的前提下,共同训练得到的更加准确的模型,然而,基于多方协作的深度学习模型易受隐私窃取攻击,即模型易被恶意的一方利用来还原其他数据源提供的数据.

本文的贡献有 4 个方面:

1) 将现有的深度学习应用于网络空间安全的研究进行了总结分类,主要针对恶意软件检测和入侵检测,从分类算法、特征提取算法和检测效果等方面分析了这些应用的研究进展,总结出存在的问题和后续研究方向:特征选择问题,需从原始数据(raw data)中提取更全面的特征;自适应性问题,可

通过 early-exit 策略[18]对模型进行实时更新;可解释性问题,可使用影响函数(influence functions)[19]得到特征与分类标签之间的相关性.现有的深度学习的研究大多基于图像处理,图像是网格数据,而安全领域的数据主要是序列数据,即离散型数据.对于深度学习的生成型模型而言,离散型数据不利于梯度的传递,因此,在解决网络空间安全领域所存在的问题时,需要借鉴现有深度学习的基于图像数据的处理方法,并对离散型数据的处理方法进行再设计与再创新.

2) 基于对本领域最新文献的深入调研,我们归纳总结了深度学习的研究面临的十大问题与机遇.在此基础上,我们首次归纳了深度学习应用于网络空间安全的研究面临的十大问题与机遇,按照这些问题的严重性从高到低,我们将这些问题分为 3 个层次:算法安全性(即算法脆弱性)、算法功能(即序列化模型相关问题)和算法性能.第 1 层是算法脆弱性问题,包括深度学习模型易受对抗攻击和隐私窃取攻击;第 2 层是序列化模型相关问题,包括程序语法分析、程序代码生成和序列建模长期依赖问题;第 3 层是算法性能问题,即可解释性和可追溯性问题、自适应性和自学习性问题、存在误报以及数据集不均衡的问题,为后续的研究工作指出方向.

3) 分析了深度学习应用于网络空间安全的研究存在的第一大问题,即基于深度学习进行分类的应用易受对抗攻击,并将现有的针对深度学习模型的对抗攻击及其防御措施进行了总结.我们指出了这些防御措施的局限性,并为更有效的防御方案指明方向:基于对抗训练的防御.

4) 分析了深度学习应用于网络空间安全的研究存在的第二大问题,即协作性模型的机密性差、易受隐私窃取攻击以及泄露训练数据或模型架构.总结与比较了现有的针对协作性模型的隐私窃取攻击的防御措施,并指出了更有效的防御措施的研究方向.

1 深度学习模型

深度学习是机器学习的一个分支,深度学习模型不同于传统机器学习模型:深度学习模型基于神经网络,通过训练调整神经网络的参数,得到每一层的权重值,每层代表一种对输入数据的表征,以此来将原始数据转换为最简单的表征[2].根据深度学习模型的使用场景,例如数据生成或数据识别,Deng 等人[3]将深度学习模型分为三大类:生成模型(generative models)、识别模型(discriminative models)和混合模型(hybrid models).生成模型对数据进行模式分析,得到数据之间的高阶相关性,用于生成新的数据,包括 RNN,DBN,AE 等;识别模型则具有模式分类和模式辨别能力,通常通过对标记数据的预测类的后验分布进行表征,如 CNN;混合模型是生成模型和识别模型的结合,基于生成模型来实现分类的目标,如 DNN.深度学习应用于网络空间安全的研究采用的分类算法主要有 DNN,CNN,RNN,DBN,AE,本节简要回顾这 5 种深度学习模型.

1.1 深度神经网络

深度神经网络(DNN)是典型的深度学习模型,其他深度学习模型都是在 DNN 的基础上扩展而来的.DNN 本质上是一个函数链,每个函数是一层,每层由神经元(neuron)组成.神经元之间由权重和偏差连接.在 DNN 的训练过程中,通过最小化训练数据集上的损失函数(loss function,error function 或 cost function)的值来确定权重和偏差[13],即优化(optimization)技术.正则化(regularization)技术[2]用来避免 DNN 过拟合[20],其目标是使训练的模型与真实的数据生成过程相匹配.

1.2 卷积神经网络

卷积神经网络(CNN 或 ConvNet)[12]是指在网络的至少一层中使用卷积运算来代替普通的矩阵乘法运算的神经网络[2].卷积是一种特殊的线性运算.例如图 1 所示的图像识别任务,每个卷积对应图像的不同特征.网络低层的卷积倾向于学习图像的简单属性,包括空间频率、边缘和颜色;高层卷积用于识别图像的复杂属性[21].卷积网络每层通常包括 3 级:卷积级、探测级(detector)和池化级(pooling)[3,22].经典的 CNN 模型有 LeNet[23],AlexNet[3],GoogleNet[24],VGG[25],ResNet[26]等.CNN 的变体包括递归卷积网络(RCN)[27]、叠加卷积 AE[28]和卷积深度信念网络(CDBN)[29]等.

1.3 循环神经网络

循环神经网络(RNN)用于处理序列数据,通常将序列划分为小批量(minibatch)来操作,并使序列的所有时间步共享相同的权重,即实现自循环.文献[30-32]相继提出并改进了基于 RNN 的 seq2seq(sequence-to-sequence)架构,用于将可变长度序列映射到另一个可变长度序列.双向 RNN(bidirectional recurrent neural network, BRNN)[33-34]用以同时学习当前时间点的未来和过去的序列数据,应用于手写

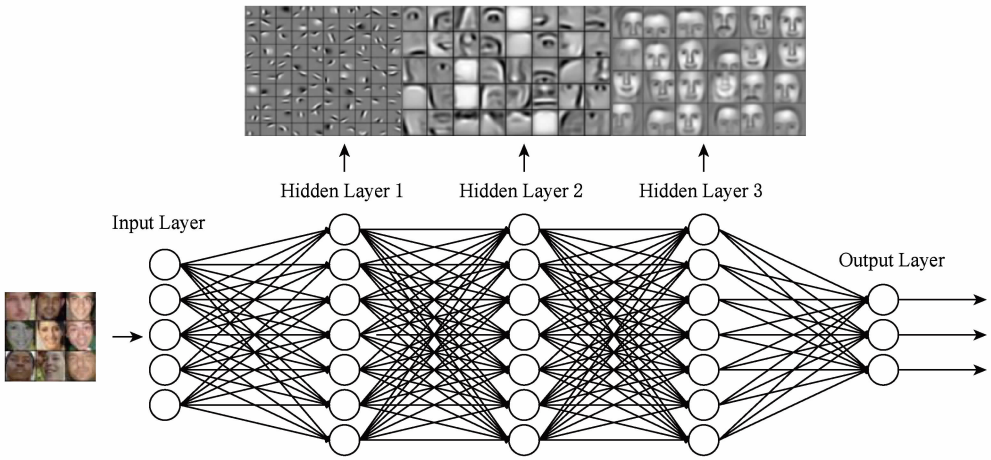

Fig. 1 A convolutional neural network for image classification

图 1 基于卷积神经网络的图像分类

识别[35-36]、语音识别[37-38]和生物信息学[39]等领域. 门控 RNN(gated recurrent neural network, GRNN)解决了原始 RNN 存在的梯度消失(vanishing gradient)与梯度爆炸(exploding gradient)问题,包括长短时记忆单元(long short-term memory, LSTM)和门控循环单元(gated recurrent unit, GRU)[40-46]. LSTM 使自循环的权重取决于上下文的特质,使得 LSTM 在手写识别与生成[36-38]、机器翻译[31]、语音识别[38,47]和图像标题生成[48-52]等领域成功应用.

1.4 深度信念网络

深度信念网络(DBN)由 Hinton 等人[14]提出,他们通过叠加若干限制玻尔兹曼机(restricted Boltzmann machine, RBM)[53]来构建 DBN. 普遍认为 RBM 是一种特殊的 Markov 随机场(Markov random field)[54-55],主要应用于特征学习[56-58]和数据生成[59-61]等方面. DBN 和 RBM 都没有层内连接,但是 DBN 具有多个隐藏层,单个隐藏层的隐藏单元之间存在连接. DBN 是一种混合模型[62-63],应用于图像处理和癌症预测[64-65]等诸多领域. 文献[66-69]优化了 DBN 在可扩展性、精度和灵活度等方面的性能.

1.5 自编码器

自编码器(AE)主要用于学习数据的有用信息,过滤无用信息[70]. 自编码器由一个编码器(encoder)和一个解码器(decoder)构成,可以在输入端进行维度降低[71-72]来提高学习效率. AE 的扩展有 k-sparse AE[73]、去噪 AE(denoising autoencoder)[74-78]、叠加去噪 AE(stacking denoising autoencoder)[79]、收缩 AE(contractive autoencoder)[80-81]和可分离深度 AE(separable deep autoencoder)[82]等.

2 深度学习应用于网络空间安全的研究现状

机器学习在网络空间安全领域的应用已有 40 多年的历史,如贝叶斯、支持向量机(support vector machine, SVM)和逻辑回归等算法,对于恶意行为检测的研究具有重大贡献. 随着计算机硬件技术的发展,深度学习算法在多媒体处理方面带来突破性的成果,网络空间安全领域的研究者也在尝试将深度学习算法应用于恶意软件检测和入侵检测等领域,相对于传统的机器学习算法,深度学习提高了检测效率、降低了误报率,能够自动化智能化识别攻击特征,有助于发现潜在安全威胁. 通过对网络空间安全领域的文献进行深入全面的调研,本文发现应用深度学习的网络空间安全领域主要有恶意软件检测和入侵检测,其中具有代表性的工作分别是文献[11]和文献[83]. 本节分别对深度学习应用于恶意软件检测、入侵检测以及其他网络空间安全领域的研究进展进行分析总结.

2.1 恶意软件检测

根据表征方法的不同,用于恶意软件检测和分类的深度学习模型可以分为静态分析、动态分析和混合分析三大类. 静态分析技术从软件及其反编译后的代码中直接提取特征,而无需实际运行软件;动态分析技术在软件执行期间观察其行为;混合分析技术结合了静态分析和动态分析的特点,即检查软件代码特征,并观察其执行行为. 现有的基于深度学习的恶意软件检测方案,针对的恶意软件运行系统可分为 4 类: Generic, Windows, Android, Flash. 表 1 对这些方案的特征提取和分类算法进行了比较.

这些工作的分类算法可分为6类,即DNN,CNN,AE,RNN,DBN以及非深度学习方法(如统计方法或者其他机器学习算法等),下面按照分析技术和分类算法对这些工作进行分类介绍.

Table 1 Comparisons of Research on Malware Detection Using Deep Learning

表1 基于深度学习的恶意软件检测研究的比较

Analysis Techniques	Malware Types	Feature Extraction					Classifier					
		DNN	DBN	RNN	AE	RP	DNN	CNN	AE	RNN	DBN	Non-DL
Dynamic Analysis	Generic	×	×	×	×	√	Ref[11,84]					
		×	×	×	×	×		Ref[85]		Ref[85]		
	Windows	×	×	√	×	×		Ref[86]				Ref[87-88]
		×	×	√	√	×			Ref[89]			
		×	×	×	×	×		Ref[88]	Ref[90]	Ref[91]		
Static Analysis	Android	×	×	×	×	×		Ref[92]				
	Windows	×	×	×	×	×	Ref[93]			Ref[94]		
Hybrid Analysis	Android	×	√	×	×	×					Ref[95-96]	
		√	×	×	×	×						Ref[97]
	Flash	×	×	×	×	×	Ref[98]					

Note: "√" means the corresponding feature extraction algorithm is applied; "×" means the opposite of "√".

2.1.1 动态分析

根据恶意软件运行系统的不同,基于动态分析技术的恶意软件检测方案可分为2类:Generic和Windows.其中,对Generic类型的恶意软件检测的工作是文献[11,84-85].Dahl等人[84]首次将神经网络应用于恶意软件检测,他们提出了具有随机映射(random projection, RP)[99]功能的简单前馈神经网络,从可执行文件中提取出的特征集中学习.他们提取的特征包括API tri-gram,即3个连续的API调用序列、API调用参数以及从系统内存中得到的Nullterminated Objects,共179 000个初始特征,通过RP降维得到4 000个特征,使用DNN进行分类.二分类错误率是0.49%,多分类错误率(即识别恶意软件所属分类的错误率)是9.53%.他们表明DNN的准确率相比逻辑回归有大幅提升.随后,Huang等人[11]提出了MtNet,MtNet使用多任务学习(multi-task learning)和DNN,提取的特征与Dahl等人[84]相似.使用互信息进行特征提取,得到50 000个特征,再使用RP进行降维,得到4000个特征.他们使用的训练集和测试集大小分别是4 500 000和2 000 000,具有当前恶意软件检测领域研究所使用的最大数据规模.二分类错误率和多分类错误率分别是0.358%和2.94%.DNN的隐藏层使用整流线性单元(rectified linear unit, ReLU)作为激活函数(activation function),他们的研究表明,相比Sigmoid激活函数,使用ReLU时所需的训练次数(epochs)减少了一半,并且使用Dropout技术可防止过拟合和降低错误率.Kolosnjaji等人[85]构建了基于卷积层和循环层的神经网络,用于检测动态分析得到的恶意软件系统调用序列,在包含4 753个恶意软件样本的数据集上,该模型的检测准确率是89.4%.

对运行于Windows系统的恶意软件进行检测的深度学习方案有文献[86-91].其中,文献[86-88]均使用RNN进行特征提取.Tobiyama等人[86]同时使用RNN和CNN进行特征提取.Pascanu等人[87]首次将RNN应用于恶意软件检测,他们提出了一个基于动态分析的2层架构的恶意软件检测系统:第1层是RNN,用于学习API事件的特征表示;第2层是逻辑回归分类器,使用RNN学习的特征进行分类,然而误警率(false positive rate, FPR)较高(10%).Athiwaratkun等人[88]在文献[87]的基础上提出了用LSTM和GRU进行恶意软件检测,并提出了使用CNN的字符级别的检测方案.Wang等人[89]提取的特征是API调用序列,使用基于RNN-AE来实现数据降维,此外,他们使用了2个解码器,其中一个用于生成恶意软件的文件访问模式(file access pattern, FAP),另外一个用于恶意软件的分类.恶意软件的分类针对已知的恶意软件,通过生成所有已知恶意软件的FAP来检测未知恶意软件,恶意软件检测的错误率是1.6%.这2个解码器通过监督学习进行训练.该工作最大的贡献是,通过

FAP为实现深度学习模型的可解释性(interpretability)提供了可能:当判断出一个软件为恶意软件时,分类器会输出其判断依据,也就是该软件代码中具有恶意行为的API调用序列.然而,此方法只使用API调用序列作为特征,如何融合API调用参数和文件结构特征来改善模型,是一个技术难点.此外,生成FAP的解码器使用监督学习,需要大量的恶意软件的FAP来训练分类模型,如何产生这些恶意软件的FAP,也是需要解决的技术难点.文献[86,89]都使用RNN进行特征提取,然而二者不同的是,文献[86]采用RNN从日志文件中提取特征,日志文件中记录了软件的进程行为,包含API调用,还有操作结果、操作路径和操作描述等文本信息,而文献[89]仅使用了API调用序列.RNN训练完成后,文献[86]使用RNN提取的特征来生成特征图像,并使用CNN对特征图像进行分类,AUC值是96%.他们首次通过生成特征图像的方法使得CNN应用于恶意软件检测,然而数据量不够大,仅使用了81个恶意软件日志文件和69个合法软件的日志文件,该方法应用于大规模数据集时的有效性还有待验证.

Hardy等人[90]将叠加去噪自编码器(stacking autoencoder, SAE)应用于恶意软件检测,用SAE模型进行无监督的预训练,采用后向传播(back propagation)来调整模型顶层的参数.然而SAE的使用并没有显著提升检测效果,准确率是95.6%.作者指出,稀疏的SAE或许可以提升检测效果,因为特征矩阵是稀疏的.现有的深度学习的基于动态分析的恶意软件检测研究都是在软件运行结束后,对其运行行为进行分析,Rhode等人[91]提出在恶意软件运行的初期对其进行恶意行为的预测,他们使用RNN进行Windows PE文件的预测,数据集包含594个恶意样本和相同数量的正常样本.他们表明,根据软件前4 s的运行行为,RNN对恶意软件的预测准确率是91%,随着观察的运行时间的增长,RNN的预测准确率也随之提高,根据前19 s的运行行为,RNN的准确率达到98%,而其他传统机器学习算法达到的最高准确率仅有90%.

2.1.2 静态分析

使用深度学习作为静态分析工具来识别二进制文件中的函数,是许多二进制分析技术中非常关键的一步.此外,对于恶意软件检测、软件漏洞防御和逆向工程等技术,直接对程序的二进制文件进行分析往往是最有效的.文献[92-94]应用深度学习技术,研究了基于静态分析的恶意软件检测:文献[92]进行Android系统的恶意软件检测,文献[93-94]针对的是Windows系统的恶意软件检测.

Nix等人[92]使用CNN作为分类器,通过API调用序列来检测恶意软件,CNN的准确率是99.4%,高于主要用于序列建模的LSTM(89.3%),然而作者并未解释其中原因.同时,深度学习算法效果也领先于机器学习算法:基于n-gram的支持向量机和朴素贝叶斯的准确率分别是66%和82%.Saxe等人[93]提出针对Windows PE文件的静态恶意软件分类系统,分类模型是包含2个隐藏层的DNN,选取了PE文件的4个类型的特征:字节频率、二元字符频率、PE Import Table以及PE元数据特征.分值校准模型(score calibration model)对DNN的输出进行计算,求得每个软件的异常值.系统载荷小、轻量级,但是作者并未展示增加隐藏层数量之后的效果,无法确定该系统扩展成深度模型的可行性.2015年USENIX会议上,Shin等人[94]提出通过判断函数位置来进行恶意软件检测.他们使用RNN分析Windows二进制文件来检测函数的开始和结束位置,检测PE x86-64文件(portable execution file)的起止位置的$F1$值是99.38%.对于与训练集中的二进制代码相似的代码,该方案可以识别出其函数起止位置,但是对于与训练集差别很大的代码,该方案无效,因为他们所采用的RNN是对于整个序列对象建模,所以检测时观察的对象是整个代码序列.可能的解决方案是引入注意力机制(attention mechanism),使RNN只关注影响检测结果的那部分代码序列,即使被检测代码和训练集不相似,也不影响其检测结果.

2.1.3 混合分析

混合分析技术被用来进行Android系统恶意软件[95-97]和Flash恶意软件的检测[98].Droid-Sec[95]是一种基于深度学习的半监督的Android恶意软件检测系统,使用静态和动态分析提取202个特征,包括使用权限和API以及动态行为.使用DBN进行无监督的预训练,预先训练的DBN使用后向传播进行权值微调,准确率是96.5%.DroidDetector[96]改进了Droid-Sec,提供了在线的检测系统供用户测试,并且扩大了训练集,以提取更精确的特征.HADM[97]为每个特征向量集训练DNN,将DNN学习的特征与原始特征相结合,然后使用多核学习(multiple kernel learning, MKL)进行分类,准确率

达到93.8%.Jung等人[98]介绍了一种恶意Flash的检测方案,使用混合分析技术得到SWF文件的API调用序列等特征,实现了基于DNN的Flash恶意软件检测模块的原型,准确率是100%.此外,David等人[100]将深度学习应用于生成恶意软件签名,通过叠加去噪AE构造DBN,生成恶意软件行为的签名.

2.2 入侵检测

1992年,Debar等人[101]首次将神经网络应用于网络入侵检测;2014年,Creech等人[102]首次将神经网络应用于基于主机的入侵检测.在此基础上,应用深度学习模型的入侵检测系统兴起[103].应用深度学习进行网络入侵检测的工作大多基于KDD Cup 1999(KDD99)数据集[104].该数据集包含4 898 431条流量数据,每条数据包含协议类型、服务类型等41个特征,包含22种攻击,这些攻击可分为四大类:拒绝服务攻击(denial of service, DoS)、远程到本地的攻击(remote to local, R2L)、用户到远程的攻击(user to remote, U2R)和探测攻击(probing).为了解决KDD99数据集存在的问题,Tavallaee等人[105]在KDD99数据集的基础上提出了NSL-KDD[106],该数据集删除了KDD99中的一些冗余数据,其特征维度和攻击类型与KDD99数据集相同.表2比较了现有的基于深度学习的入侵检测方案的特征提取和分类算法,DBN和AE是主要的特征提取方法,而LSTM,DBN,AE是主要的分类算法.这些方案主要基于KDD99或者NSL-KDD数据集.

Table 2 Comparisons of Research on Intrusion Detection Using Deep Learning

表2 基于深度学习的入侵检测研究的比较

Datasets	Feature Extraction				Classifier					
	DBN	AE	CNN	Non-DL	LSTM	GRU	DBN	AE	CNN	Non-DL
KDD99	×	×	×	√	Ref[83,107]	Ref[108]	Ref[109]			
	×	√	×	×			Ref[110]			
NSL-KDD	√	×	×	×						Ref[111]
	×	√	×	×				Ref[112-113]		
	×	×	×	√			Ref[114]			
	×	×	×	×				Ref[115]		
Others	×	√	×	×				Ref[116-117]		
	×	×	√	×					Ref[118]	
	×	×	×	×	Ref[119]					
	×	×	×	√				Ref[120]		

Note: "√" means the corresponding feature extraction algorithm is applied; "×" means the opposite of "√".

在使用KDD99数据集的研究中,文献[83,107-109]使用的分类算法分别是LSTM,GRU,DBN.文献[110]先采用AE进行降维,然后采用DBN进行分类.Staudemeyer[83]首次使用LSTM进行网络入侵检测,输入特征是KDD数据集原有的41个特征,输出向量长度为5,包括4种攻击和正常请求.Kim等人[107]使用LSTM在KDD99数据集上进行网络入侵检测并进行参数选取,取得了较高的检测率(98.88%)和准确率(96.93%),然而,LSTM的误报率也偏高,达到了10.04%.作者猜想LSTM的初始权重值选取不当可能是导致误报率较高的主要因素之一.Putchala[108]提出将GRU应用于物联网领域的入侵检测,然而仅在KDD99数据集上进行实验,得到的准确率高于99%.Gao等人[109]首次将DBN作为分类模型应用于入侵检测,验证了DBN可以应用于入侵检测的分类.同时,他们表明,参数调试(fine-tuning)和预训练(pre-training)可大大提升DBN的检测效果,误报率是0.76%.Li等人[110]使用DBN进行特征提取的检测率比单独采用DBN进行分类时的检测率高.他们展示了网络架构、降维之后的特征个数、预训练次数等不同的参数对于检测效果的影响.

使用NSL-KDD的入侵检测方案有文献[111-115].Salama等人[111]首次将DBN作为生成模型应用于入侵检测中的数据降维,使用SVM对降维之后的数据进行分类.他们的结果显示,DBN-SVM结构比单独的DBN或者SVM进行分类得到的准确率更高,同时,当采用SVM进行分类时,相比于主成分分析(principal component analysis,PCA)、卡方检验等降维方法,采用DBN进行降维时的检测准

确率最高. Niyaz 等人[112]首先使用 1-to-n encoding 方法进行特征编码,得到 121 个特征,然后使用 Sparse AE 进行无监督的降维,最后通过 Softmax 回归来训练分类器. 将输出类别分为 3 种情况:异常检测(2 类,包括正常类和异常类)、攻击类别检测(5 类)和攻击检测(23 类). Abolhasanzadeh[113]介绍了一种使用 Bottleneck AE 架构进行特征降维的方法,训练时,该架构首先通过编码器将输入特征进行降维,生成 Bottleneck 特征,这些特征再通过解码器进行特征还原,重现输入特征. Bottleneck 特征的个数就是 Bottleneck 层神经元的个数. Bottleneck AE 降维效果优于 PCA,kernel PCA 以及因子分析. 然而,作者并未指出在衡量降维效果时所采用的分类算法. Alom 等人[114]也使用 DBN 进行入侵检测中的分类,通过对特征进行数字编码,并通过离差标准化(min-max normalization),得到 39 个特征,测试集上的准确率是 97.45%,优于 DBN-SVM 以及 DBN 的检测效果,但是作者并未对其结果的提升原因做出明确说明. Aygun 等人[115]提出随机去噪自编码器来进行入侵检测,准确率是 88.65%.

使用私有数据集或者其他公开数据集进行入侵检测的工作有文献[116-120]. Wang[116]使用企业私有流量数据集,基于叠加自编码器(stacked autoencoder, SAE)[79]对网络流量按照协议进行分类,用于检测未知协议的流量,即异常流量. Yu 等人[117]和 Wang 等人[118]从原始数据中提取特征来进行入侵检测与流量分类:Yu 等人[117]使用空洞卷积自编码器(dilated convolutional autoencoder, DCAE)来进行分类,准确率是 98.8%,他们并未与传统机器学习算法进行效果比较;Wang 等人[118]将流量的每个字节转换成像素,由此来把流量转换为图片,再将图片作为 CNN 的输入进行训练与分类,得到的二分类和多分类准确率分别是 100%和 99.17%,由此可见,使用原始特征的效果要优于使用 NSL-KDD 和 KDD99 数据集中人工提取的特征的效果. 然而,Yu 等人[117]和 Wang 等人[118]均使用私有数据集进行测试,无法与其他方案进行直接对比. 以上讨论的均是基于网络的入侵检测方案,Kim 等人[119]建立了基于主机的入侵检测系统,使用多个 LSTM 对系统调用建立语言模型(language model);再将多个 LSTM 组合,通过对这些 LSTM 得出的异常值求平均值来判断每个访问是否是攻击. Yu 等人[120]提出了基于 TCP,UDP,ICMP 会话的僵尸网络流量监测方案,使用叠加去噪自编码器的检测准确率是 99.48%.

2.3 其他应用

除了恶意软件检测和入侵检测,深度学习也应用于其他网络空间安全领域:

1) 程序分析与漏洞挖掘. Zaheer 等人[121]通过使用基于 LSTM 的语言模型实现了简单的静态分析器,该静态分析器的目的是检查程序中每个变量在调用之前是否被初始化,存在较高的误报率,无法实际应用. Godefroid 等人[122]首次将深度学习应用于程序漏洞挖掘,他们使用 seq2seq 架构来生成用于模糊测试(fuzzing test)的测试用例,进行 PDF 文件的漏洞挖掘. 然而,这种方案依赖于大量的测试用例作为训练集,包括合法输入和可以触发漏洞的输入. 触发漏洞的输入在实际中难以获取,是该方案最大的局限性.

2) 密码破解. Melicher 等人[123]使用 LSTM 实现了轻量级的密码破解器,作者将该密码破解器压缩并嵌入网页,提高了破解速度,可用于测试用户设置的密码强度,与传统的密码破解方案相比,基于 LSTM 的密码破解器可以在更短的时间内完成破解,然而,针对 LSTM 密码破解器的超参数的选取,依然缺乏充分的理论分析.

3) 针对恶意软件检测系统的对抗攻击与防御. 2.1 节介绍了将深度学习应用于恶意软件检测的诸多工作,文献[124-127]则利用了深度学习算法的脆弱性,实现了针对这些基于深度学习进行分类的恶意软件检测系统的对抗攻击;文献[128]提出了对抗攻击的防御,并将其应用于加固恶意软件检测的深度学习模型(见 4.1 节).

2.4 深度学习的安全应用研究小结

根据 2.1~2.3 节中深度学习应用于恶意软件检测和入侵检测等网络空间安全领域研究的分析,本节将该领域的研究进展从 3 方面进行总结.

2.4.1 特征选择问题

深度学习应用于恶意软件检测和入侵检测这 2 个领域的现有工作基本上都是使用现有数据集默认的特征来进行学习,如入侵检测的工作均使用 KDD 或者 NSL-KDD 数据集已提取的 41 个特征,恶意软件检测的工作使用 API 调用序列等作为特征. 这些特征不足以完全概括数据的全部特点,使用深度学习算法时,可以从原始数据入手对特征进行建模,以更好地利用深度学习算法调动硬件运算能力来提高学习效果.

2.4.2 特征学习问题

恶意软件检测和入侵检测的对象是序列化数据,RNN,DBN,AE 等算法被广泛应用于这些数据

的特征学习.然而现有工作更多关注于设计分类或者检测的结构,对特征学习方面的深入研究甚少.文献[89]对已知的恶意软件特征进行学习,但是并未实现特征的可解释性,即特征与攻击行为的相关联性.据调研,还未有网络空间安全领域的工作实现了特征的可解释性.目前较优秀的方案是采用影响函数[19](见3.1.1节),该方案在图像处理领域初有成效,实现深度学习在网络安全空间应用的可解释性,这方面研究的空白仍需填补.

2.4.3 自适应性问题

针对恶意软件检测和入侵检测领域的研究,采用了DNN,CNN,RNN,DBN,AE等深度神经网络架构,并且可以选取合适的优化算法和超参数,因此,深度学习算法效果相比传统机器学习算法有较大提升.应用于未知的恶意行为的检测时,提高了检测率并降低了误报率;应用于已知恶意行为的分类时,提高了分类准确率.然而,效果的提升带来的是训练以及测试时间的增长,不利于模型的及时更新.由于恶意代码和入侵行为会随着攻击者攻击技术的提升而更加难以检测,检测模型只有随着安全威胁的演变自适应性地进行更新,才可以在第一时间全面检测出新的安全威胁.深度学习模型具有训练和测试时间长的特点,如何才能在保证高准确率和低误报率的前提下,更加高效地训练和测试深度模型、实现检测模型的自适应性的问题,是深度学习应用于网络空间安全研究面临的重大难点,较有前景的方案是early-exit[18],一种自适应性的深度学习模型更新策略(见3.2.2节).

综上所述,现有的深度学习在这2个领域的应用研究相比传统机器学习有较大提升,但仍需在3方面进行改善:1)特征选择方面,从原始数据中提取更详尽全面的特征;2)特征学习方面,还需实现特征与分类标签相关性的可解释性;3)自适应性方面,需要在保证高准确率和低误报率的前提下,使模型可以自适应性地更新,来检测变化的安全威胁.此外,深度学习应用于网络空间安全的研究仍存在诸多问题,如算法脆弱性等问题,即分类模型易受对抗攻击、协作性模型易受隐私窃取攻击等,第3节分别对深度学习和深度学习应用于网络空间安全的研究面临的问题与相应的机遇进行分析.

3 问题与机遇

本文调研了2013年1月到2017年7月深度学习的高引论文、预印本数据库arXiv中深度学习应用于网络空间安全研究的论文,以及中国计算机学会CCF A类和CCF B类会议与期刊中该领域的论文,总计156篇,表3对这些论文进行了分类统计(注:由于一些论文涉及多个研究方向,故该统计存在重叠).基于调研的相关文献,我们归纳总结了深度学习发展面临的十大问题与机遇,在此基础上,首次归纳了深度学习应用于网络空间安全所面临的十大问题与机遇.

Table 3 Number of Research Papers Investigated

表3 本文调研的论文

Research Category	Research Content	#Papers
Cyberspace Security Applications of DL	DL applied to malware detection	20
	DL applied to intrusion detection	10
	DL applied to other cyberspace security fields	8
Problems of DL Applications in Cyberspace Security	Adversarial attacks against DL models	31
	Defenses against DL-based adversarial attacks	27
	Privacy risks in collaborative DL models	5
	Enhancement of collaborative DL models	11
DL Models	Deep neural networks	79
	Convolutional neural networks	82
	Recurrent neural networks	36
	Deep belief networks	25
	Autoencoders	31

3.1 深度学习发展面临的十大问题与机遇

如表4所示,深度学习发展面临的问题分为3类:1)神经网络训练技术难点,包括问题1~7;2)特征处理问题,即问题8;3)数据标签获取或者无监督学习技术难点,包括问题9~10.本节对这3类难点及其对应的问题和机遇展开分析.

3.1.1 神经网络训练技术难点

这类难点存在于不同的深度学习模型.问题1和问题3存在于所有判别型深度学习模型,问题1是模型对抗攻击测试问题,深度模型易受对抗攻击,即深度学习模型会被攻击者利用,以实现攻击者选择的特定输出或行为.因此,设计一个对抗攻击框架是有意义的,该框架可以结合各种对抗样本产生方法,以此来检验目标系统面对不同的对抗攻击时的健壮性,以更好地设计防御措施来保护目标系统.问题3是模型可解释性和可追溯性,深度学习模型输出分类结果时,其依据对用户往往是不可见的.普遍

Table 4 Top 10 Obstacles and Opportunities for Research on Deep Learning

表 4 深度学习的研究面临的十大问题与机遇

No	Obstacles	Opportunities
1	Adversarial attack tests for classification networks	Generating adversarial examples
2	Privacy risks in collaborative models	Teacher-student model
3	Low interpretability and traceability	Perturbations; Influence functions
4	Searching global minimum in optimization	Searching points with lower cost instead of the lowest cost
5	Improving initialization strategies	Regarding the range of initial weights as a hyperparameter
6	Memorizing facts related to concepts	Memory networks
7	Difficulty in reaching equilibrium in GAN	Separating generation process into various levels of details
8	Interpretability of extracted features	Visualizing features
9	Unsupervised deep learning for classification	DBN; RBM
10	Difficulty in obtaining massive labeled data	Distributed data source; Unsupervised learning

认为,深度模型的准确度与模型的可解释性和可追溯性成反比[129],在保障模型高准确率的前提下,如何提高模型的可解释性和可追溯性,使人类从机器的决策中学到知识,是这 2 年深度学习领域的重点解决问题. 2016 年 KDD 会议上 Ribeiro 等人[130]提出模型的局部解释性方案(local interpretable model-agnostic explanation, LIME),方法是:对样本在局部特征空间进行细微扰动,根据每次扰动之后的预测结果来得出特征与预测类别之间的关系. 2017 年 ICML 会议上 Koh 等人[19]实现了可追溯性,使用稳健统计学(robust statistics)中的影响函数,来得出训练集中对测试样本的预测类别影响最大的样本. 然而,这些方案的研究均处于起步阶段,具有运算量大、复杂性高的特点.

问题 2 存在于所有协作性深度学习模型,即协作性模型易受隐私窃取攻击、机密性差,目标模型的训练集或者架构参数等会被攻击者恶意获取(见第 6 节). 问题 4~5 存在于所有深度学习模型. 问题 4,即神经网络优化中的全局最小值点问题,在训练神经网络的过程中,由于损失函数往往非凸(nonconvex),容易使算法陷入局部最小值点,很难找到一个全局最小值点,使得损失函数值最低. 目前对于深度学习中的非凸优化问题,只有很少的理论分析,可以考虑选取使成本值尽可能低的点来解决非凸优化问题[2,131]. 问题 5,目前有关改进初始化策略的研究还不够完备,对于初始偏差的选取,大多数情况下设置为 0,对于初始权重的选取,传统的观念是采用较小的随机值,较小的初始权重虽然有利于正则化,但是不利于优化过程中的信息传递,目前人们仍然没有很好地理解初始权重值对模型泛化能力的影响. 可选的方案是将初始权重的数值范围设置为超参数,通过超参数搜索方法来确定这些参数的范围[2]. 问题 6 存在于生成型深度学习模型. 简单的人机对话已经实现,然而,为了让计算机的回答具有常识性,就需要深度学习模型能够记忆大量与概念相关的事实并进行描述,可选的方案是记忆网络(memory networks)[132-133]. 虽然 RNN 及其许多变体都具有记忆机制,但是 RNN 把状态及其权重压缩为一个低维向量,造成原始数据的信息损失,而记忆网络包含一个可以实现超长序列的记忆模块,其记忆能力优于 RNN.

问题 7 针对生成式对抗网络(generative adversarial network, GAN). GAN 由 Goodfellow 等人[134]在 2014 年提出,属于深度学习模型中的混合模型,包含一个生成器(generator)和一个识别器(discriminator),生成器用于学习真实的数据分布来生成逼真的数据,判别器用于判别数据是真实的还是生成器生成的. GAN 采用博弈论的纳什均衡思想,训练优化目的是实现生成器和识别器之间的纳什均衡. GAN 自从提出,其强大的生成新样本的能力引起了国内外学者广泛关注[135],出现了许多变体,在理论和应用上均对原始的 GAN 有所扩展,如 C-GAN[136], Semi-GAN[137], Bi-GAN[138], Info-GAN[139], AC-GAN[140], Seq-GAN[141], DC-GAN[142], LAP-GAN[143], LSTM-GAN[144], VAE-GAN[145], W-GAN (Wasserstein GAN)[146],其中最优秀的是 W-GAN,解决了原始 GAN 存在的诸多问题. 针对 GAN 中,生成器和识别器的纳什均衡状态实现困难的问题,可选方案是将生成器的生成过程层次化,逐层完成生成过程[147-148]. LAP-GAN[143]首次采用了层次化的思想,

他们将GAN与Laplacian pyramid的层次化表征相结合,先生成分辨率低的图片,再逐渐的向图像添加细节,提高图片的分辨率.

训练技术难点包含的问题中,前5个问题均与网络空间安全领域直接相关,因为这些问题涉及用于分类的深度学习模型,可以用于网络空间安全领域中的异常检测等;问题6~7涉及生成型深度学习模型,均与网络空间安全领域间接相关:在对深度学习的网络空间安全应用进行对抗攻击测试时,可以借助生成型深度学习模型来产生对抗样本.

3.1.2 特征处理技术难点

特征处理技术难点是问题8,即习得的特征解释性差.特征解释性指,对数据特征与数据所属类别内在关联之间的可解释性(见3.2.3节).在网络空间安全研究中,需要找出对恶意行为检测贡献较大的特征,也就是说,包含何种特征的软件是恶意软件的可能性比较大.

3.1.3 数据标签获取和无监督学习技术难点

问题9,无监督学习虽广泛应用于降低维度和聚类,其效果仍落后于监督学习,然而,无监督学习不依赖于样本标签的特点,使其依然对于深度学习的研究具有强大的诱惑力:在网络空间安全领域中,异常行为或者攻击行为出现的几率远远小于正常行为,因此网络空间安全领域的数据集具有正常样本和异常样本数量比例严重失衡的特点,在进行监督学习时,正负样本比例失衡会影响训练效果,采用无监督学习进行网络空间安全领域的异常检测不需要样本标签,避免了样本比例失衡带来的影响.现有的无监督深度学习的研究主要集中于生成型深度模型,应用无监督深度学习进行分类的理论分析较少,可以考虑在现有的无监督生成型深度模型(如DBN,RBM)上加以改进,实现判别型无监督深度模型.问题10,现有的针对深度学习的研究,主要集中于监督学习,监督学习依赖于海量的已标记的数据,虽然海量数据在现实中容易获取,然而其标签的获取却成为一个难点,可选的解决方案是使用不同的数据源提供的有标签的数据,即使用分布式数据源(见第5节).

3.2 深度学习应用于网络空间安全的研究面临的十大问题与机遇

通过对深度学习应用于网络空间安全的研究相关的文献进行调研,我们总结出深度学习应用于网络空间安全的研究面临的十大问题与机遇,如表5所示.按照这些问题的严重性从高到低,本文将这些问题分为3个层次,即算法安全性、算法功能和算法性能.1)算法安全性.即算法脆弱性问题,网络空间安全应用对算法的安全性要求极为严苛,深度学习算法存在的脆弱性会使算法存在受到对抗攻击和隐私窃取攻击的潜在风险,影响模型的完整性、机密性和健壮性.因此,我们认为这是深度学习的网络空间安全应用首先应解决的基础性问题.2)算法功能.即序列化模型相关问题,在算法安全性得到保障的基础上,序列建模问题应该得到关注,因为基本上所有的网络空间安全数据都是序列化数据,所有的安全应用,如程序分析、漏洞挖掘和恶意代码检测等,均依赖于序列建模.3)算法性能.在算法安全性和算法功能实现的基础上,算法性能应得以关注,如算法自适应性、可解释性、特征选取、降低误报以及数据集均衡等问题.下面对这些问题其对应机遇展开分析.

Table 5 Top 10 Obstacles and Opportunities for Research on Deep Learning Applied to Cyberspace Security

表5 深度学习应用于网络空间安全的研究面临的十大问题与机遇

No	Obstacles	Opportunities
1	Difficulty in defending against adversarial attacks	Adversarial training
2	Privacy risks in collaborative models	Teacher-student model
3	Program grammar analysis	Memory networks;Generative models
4	Program vulnerability discovery	Memory networks
5	Long-time dependencies in sequential modeling	Gradient clipping and regularization terms
6	Low self-adaptability and self-learning ability	Modular training
7	Low interpretability and traceability	Weight mechanism in attention mechanism
8	Incomprehensive features	Modeling features from raw data
9	False alarms	Pre-training; Fine-tuning
10	Unbalanced dataset	Generating the minority class examples with GAN

3.2.1 算法脆弱性

深度学习属于机器学习的范畴,因此深度学习算法与机器学习算法具有相同的脆弱性[1-2],包括表5中的问题1～2.问题1,分类模型易受对抗攻击,即深度学习模型会被攻击者利用,以实现攻击者选择的特定输出或行为.可选的防御方案是对抗训练(见第4节).对抗攻击是深度学习领域近2年的研究热点,其研究热度呈现上升趋势,2017年NIPS会议新增了基于Kaggle平台的"对抗攻击与防御"的竞赛议程[149].问题2,协作性模型易受隐私窃取攻击,即目标模型的训练集或者架构参数等会被攻击者恶意重现或获取,可选的防御方案是教师学生模型(见第5节).

3.2.2 序列化模型相关问题

很多网络空间安全领域的数据都是序列化数据,如系统调用序列和网络请求数据载荷等,因此网络空间安全领域的数据分析需要借助序列化模型,与序列化模型相关的问题包括表5中的问题3～5.问题3,程序语法分析及程序语言生成,可以通过深度学习的序列生成模型实现,如记忆网络和DBN等.程序语法分析对于生成合法的程序至关重要.Zaheer等人[121]实现的基于LSTM的语言模型存在较高的误报率,无法实际应用.问题4,漏洞挖掘.Godefroid等人[122]提出基于seq2seq架构来生成用于模糊测试(fuzzing test)的测试用例,需要大量可以触发漏洞的测试用例,这是该方案最大的局限性.同时,对合法程序样本进行建模时,这些样本往往是超长序列,可使用能够对超长序列进行记忆的模型,如记忆网络.问题5,循环网络存在序列化建模长期依赖的问题:循环网络采用链式法则(chain rule)计算网络中每层的梯度,链式法则的本质是连乘操作,当连乘的梯度大部分都很小(比如小于1),乘积会衰减到0,造成梯度消失;当连乘的梯度大部分都很大(比如大于1),乘积会趋于无穷,造成梯度爆炸.梯度消失和梯度爆炸都会为优化损失函数造成困难,使学习陷入不稳定的状态.LSTM通过引入门(gate)机制解决了梯度消失问题,使用遗忘门(forget gate)、外部输入门(external input gate)和输出门(output gate)来控制信息的传递,使用相加操作来进行梯度计算,替换了原始RNN中使用的连乘操作,有效避免了梯度消失.最简单的梯度爆炸解决方案是梯度截断(gradient clipping)[150],然而梯度截断无法解决梯度消失,为了使非LSTM的网络也可以摆脱梯度消失的问题,可以使用正则化项[150],使得梯度在传递的过程中维持在一个幅度范围内.

3.2.3 算法性能问题

深度学习应用于网络空间安全的研究面临的算法性能问题包括表5中的问题6～10.问题6是深度学习的自适应性和自学习性问题,由于安全威胁会随着时间演变,使得一成不变的检测模型无法检测出最新的威胁,这就要求检测恶意行为的安全应用具有自适应性和自学习性,来适应变化的安全威胁和攻击技术.可选的措施是时间间隔性的模块化模型训练,2016年CVPR会议上,Andreas等人[151]指出,首次将模型训练完成之后,后续训练只需要对前期学习的特征和模型进行再利用,建立基于再利用和模块化程序的元学习系统,以此来提升模型训练效率.此外,2017年ICML会议上,Bolukbasi等人[18]指出,并不是所有的数据都需要完整的DNN进行分类,并提出一种自适应的early-exit策略,来决定对样本类别进行预测时,是否要绕过DNN中的某些层,此外,他们还提出一种对于复杂样本的网络选择策略.对于每个样本,通过层与层之间进行加权二分类来判断预测该样本时,应选取early-exit策略还是网络选择策略.问题7是可解释性和可追溯性差:深度学习模型对恶意数据段的解释性差,如恶意软件检测和入侵检测时,深度模型判断出一个软件或者访问请求是恶意时,该软件或请求很可能含有一些恶意的API调用或者恶意的字段,然而深度模型并没有给出任何与判断结果相关的恶意数据段信息.可能的解决方案是采用影响函数和注意力机制,注意力机制中的权重值机制使得不同的数据段对判别结果的重要性具有可见性(见3.1.1节).此外,深度学习应用于网络空间安全的研究还面临特征选择,存在误报以及正负样本比例失衡的问题.问题8,特征不够全面,入侵检测的工作均直接使用KDD或者NSL-KDD数据集已提取的41个特征,这些特征不足以全面获取数据的信息,因此,直接从原始数据中提取特征,会尽可能多地保留数据的信息,得到更精确的模型.问题9,误报问题,可通过预训练和参数微调来减少误报.问题10,正负样本比例失衡,即恶意样本数量远小于合法样本数量,可采用GAN来产生恶意样本[124-127].

在以上十大问题中,前2个问题与深度学习应用于网络空间安全的研究的安全性紧密相关,其余问题与深度学习算法的训练效果和性能相关.问题1广泛存在于使用深度学习算法进行分类的安全应用:这些应用均具有基于深度学习的分类模块,这

些分类模块易受对抗攻击,即分类模型会被攻击者利用,以实现攻击者选择的特定输出.问题2存在于所有使用协作性深度学习模型进行分类的安全应用,这些应用易受隐私窃取攻击,模型易被恶意的一方利用来还原其他数据源提供的数据.综上所述,问题1和问题2分别针对安全应用的健壮性和机密性,因此,本文认为前2个问题是深度学习的安全应用研究领域亟待解决的重大问题.第4节和第5节对这两大问题进行详尽分析,并讨论可能的解决方案.

4 问题1:易受对抗攻击

机器学习的局限性之一是用于分类的深度模型易受对抗样本(adversarial examples)的影响[152-155].作为机器学习的一个分支,深度学习继承了机器学习的易受对抗攻击的缺陷[17,156].对抗样本是由攻击者构造的,目的是使目标模型对其进行错误分类.对抗样本利用了机器学习的2个缺陷性:利用有限训练集训练得到的模型,具有未完全泛化(imperfect generalization)[69]的特性,即泛化能力差,以及学习模型组件的线性特质[157].图2中的对抗样本示例1和示例2展示了使用AlexNet[4,156]生成的对抗样本.原始图像(original)被攻击者修改,向原始图像添加微小的失真或扰动[158](distortion or perturbation),使得DNN将原始图中每个图都识别为鸵鸟(ostrich).对于人类来说,原始图像和攻击者伪造的对抗图像是无法用肉眼辨别的.无人驾驶汽车可能使用DNN来识别交通标志[159],如果攻击者伪造的"STOP"标志导致DNN错误分类,汽车则不会停止,容易导致交通事故.在网络空间安全领域,比如网络入侵检测系统使用DNN作为分类器时,若伪装成合法请求的恶意请求绕过了入侵检测系统,会使目标网络的安全性受到威胁.

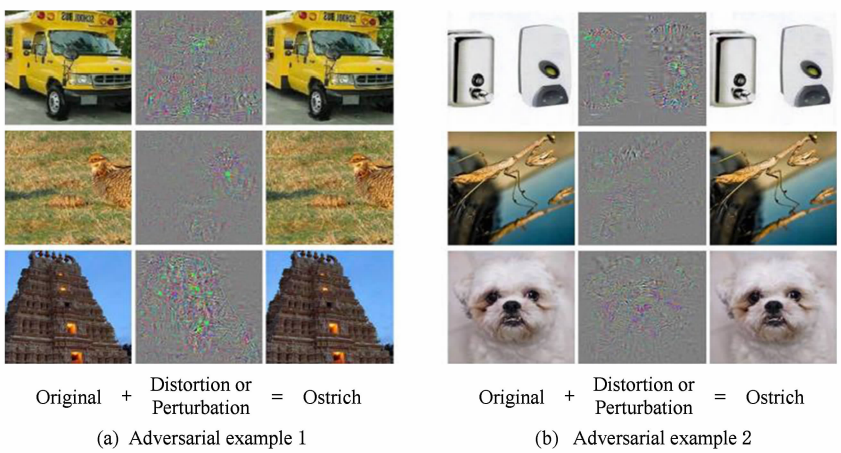

Fig. 2 Adversarial example crafting[156]

图2 对抗样本构造[156]

4.1 对抗攻击目标

对抗性深度学习(adversarial deep learning)是近2年的热点研究领域.对抗攻击目标的本质是造成目标模型进行错误分类.基于文献[160],我们将对抗攻击目标分成4类:1)非针对性输出(non-targeted output).使目标分类器输出的样本分类为与原始类不同的任意类.2)针对性输出(targeted output).使目标分类器输出的样本分类为与原始类不同的特定类.3)非针对性构造(non-targeted crafting).构造对抗样本,并且使目标分类器输出的这些对抗样本的分类为与原始类不同的任意类.4)针对性构造(targeted crafting).构造对抗样本,并且使目标分类器输出的这些对抗样本的分类为与原始类不同的特定类.不同的对抗攻击,需要攻击者掌握的关于目标深度模型的信息也不同,这些信息在文献[160]中称作对抗性知识(adversarial knowledge),包括模型架构(目标模型的参数、损失函数和激活函数等)、训练数据和模型查询(即目标模型对攻击者具有可得性,攻击者可对目标模型进行输入并得到相应输出).

4.2 对抗样本构造

攻击者利用已经训练完成的目标DNN来构造对抗样本,不会影响目标DNN的训练过程.目前有4种方法来构造针对DNN的对抗样本:

1) L-BFGS[156].L-BFGS是一种通过优化函数遍历由模型表示的流形(manifold),并在输入空间

(input space)中搜索对抗样本的方法. 通过对正确分类的输入图像添加细微扰动来获得对抗样本,从而使得目标 DNN 误判这些对抗样本.

2) Deepfool[161]. 使用 Deepfool 的理想前提是目标 DNN 是完全线性的,即存在可分割不同类别数据的超平面. Deepfool 分析得到简化分类问题的最优解,并在此基础上构建对抗样本. 然而,由于神经网络实际上不是线性的,只能朝着最优解逐步搜索,并重复该搜索过程. 当找到真正的对抗样本时,搜索终止.

3) Fast Gradient Sign[157]. 该方法得益于损失函数的梯度符号矩阵(sign matrix),在损失函数梯度变化的方向上寻找对抗样本,并用输入变化参数(input variation parameter)来控制损失函数梯度变化的幅度. 输入变化参数越大,构造的对抗样本被错误分类的可能性越大,但更容易被察觉. Kurakin 等人[162]改进了 Fast Gradient Sign 方法,他们并没有沿梯度符号的方向采用单个输入变化参数,而是采用多个较小的输入变化参数.

4) Jacobian-based Saliency Map[160]. 该方法使用 Jacobian 矩阵来评估模型对每个输入特征的敏感度;然后用 Adversarial saliency map 选择扰动,通过组合其 Jacobian 矩阵,将每个输入特征对误分类目标的贡献排序来获取对抗样本.

总之,Deepfool 类似于 L-BFGS,但产生的对抗样本的空间分布更紧凑. 最近提出的 Fast Gradient Sign 和 Jacobian-based Saliency Map 的共同点是:当模型对输入样本的变化十分敏感时,更容易计算得到导致错误分类的最小扰动,使得扰动后的样本不易被肉眼察觉,攻击更具有隐蔽性. Fast Gradient Sign 的特性是,可以快速构造具有较大扰动的对抗样本,却更容易被目标模型察觉.

4.3 对抗攻击

根据实施攻击是否需要了解目标模型的架构和参数,本文将针对 DNN 的对抗攻击(adversarial attacks)分为白盒攻击和黑盒攻击. 使用 4.2 节中介绍的 4 种方法来构造对抗样本,都需要攻击者了解目标模型的架构和参数,因此被视为白盒攻击.

Papernot 等人[163-164]提出使用 Fast Gradient Sign 和 Jacobian-based Saliency Map 这 2 种方法对未知的目标 DNN 模型进行黑盒攻击,攻击者无需了解目标模型的参数架构信息和训练数据,就可以使目标模型对输入进行误分类:攻击者通过查询目标模型来对刻意构造的样本进行标记,从而迭代生成一个标记的数据集,并使用该数据集训练替代模型(substitute model),通过替代模型来构造目标模型错误分类的对抗样本. 黑盒攻击依赖于对抗样本的跨模型传递性(cross-model transferability). 传递性[164-165]指的是,即使 2 个分类器具有不同的体系结构或者是在不相交的数据集上进行训练的,用其中一个分类器产生的对抗样本也可能导致另一个分类器也对该样本进行错误分类. 图 3 是一个跨模型传递率矩阵,第 i 行第 j 列的值代表使用分类器 i 构造的对抗样本中,使得分类器 j 错误分类的对抗样本所占的百分比,也称作传递率(transferability rate). 由图 3 得知,跨模型传递性并不局限于针对 DNN 构造的对抗样本,而是适用于多种机器学习技术;此外,逻辑回归的自身传递率最高,远远优于 DNN 对于对抗样本的自身传递率. 由文献[163]可知,传递率越高颜色越深,即传递率为 0 时是白色,传递率是 100%时是黑色.

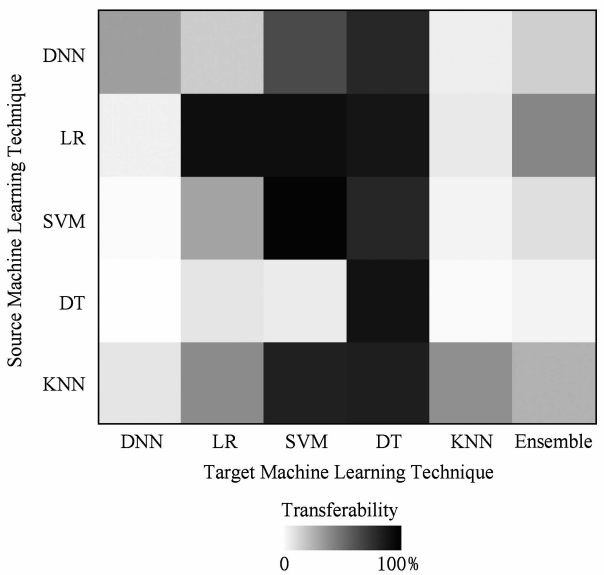

Fig. 3 Cross-model transferability[163]

图 3 跨模型传递性[163]

Moosavi-Dezfooli 等人[166]表明每个模型存在通用扰动(universal perturbations),他们使用 ImageNet[167]训练了不同的深度模型,发现使用通用扰动产生的对抗图像可以在这些模型之间传递. Liu 等人[168]全面研究了黑盒攻击的传递性,同时研究了非针对性和针对性的对抗样本,以及 3 种现有的基于单一模型搜索对抗样本的方法:基于优化的方法、Fast Gradient Sign 和 Fast Gradient. 他们表明,虽然很容易找到可传递的对抗样本,但使用现有方法生成的对抗样本几乎不会与其针对的标签一起

传递,也就是说,虽然不同的模型会对这些针对性对抗样本进行误判,但是无法统一误判得到的样本类别.他们还提出使用联合(ensemble)模型来生成可传递的对抗样本,这使得大部分针对性对抗样本首次在不同模型之间传递.Liu 等人[168]改进了此前 Papernot 等人[163-164]和 Moosavi-Dezfooli 等人[166]提出的黑盒攻击.与 Papernot 等人[163-164]提出的黑盒攻击相比,Liu 等人[168]实现的攻击使用联合分类器来生成对抗样本,不需要从目标模型查询标签,而 Papernot 等人[163-164]提出的攻击需要从目标模型查询标签.此外,Moosavi-Dezfooli 等人[166]仅研究了非针对性的对抗样本的可传递性,而 Liu 等人[168]同时研究了非针对性和针对性的对抗样本的传递性,测试了 ImageNet 数据集的在不同模型上的可传递性.

现有的对抗攻击的研究大多是基于网格数据的,主要是图像数据,最近,针对序列数据的对抗攻击研究逐步涌现,促进了网络空间安全领域的对抗攻击研究,如针对恶意软件检测系统的对抗攻击.Hu 等人[124]提出 MalGAN 来生成恶意软件的对抗样本.MalGAN 包括 3 个 DNN 模型,分别是生成器、判别器和替代模型.MalGAN 借助对抗样本的传递性,使用 GAN 来构造一个可以模拟判别器的替代模型,生成 API 调用序列构成的恶意软件的对抗样本,用来绕过恶意软件检测系统.在随后的工作中[125],他们用 RNN 代替了 MalGAN 中所使用的 3 个 DNN,与此类似的工作是文献[126].这些工作均实现的是黑盒攻击.Papernot 等人[169]使用 Fast Gradient Sign 和 Jacobian-based Saliency Map 算法,生成针对 LSTM 的对抗样本,Grosse 等人[127]使用与 Papernot 等人[169]相同的 2 种对抗样本生成方法,来生成安卓恶意软件的对抗样本.文献[127,169]需要攻击者了解目标模型参数,属于白盒攻击的范畴.

4.4 对抗攻击的防御

现有的针对对抗攻击的防御措施主要有 5 种:Defensive distillation[170]、预处理、正则化、对抗训练以及拒绝对抗样本,本节分别介绍这些防御措施:

4.4.1 Defensive distillation

2016 年 IEEE S&P 会议上,Papernot 等人[170]提出 Defensive distillation.Defensive distillation 是 Distillation[171]算法的扩展,为被保护的原始模型训练一个 Distilled 的模型,他们表示 Distilled 的模型将对抗样本的攻击成功率从 95%降低至 0.5%.训练 Distilled 的模型时,输入是训练原始模型所需的样本集合,样本标签是原始模型输出的这些样本的分类置信度(soft labels).2017 年 IEEE S&P 会议上,Carlini 等人[172]在他们之前工作[173]的基础上提出了 3 种对抗攻击,攻击者可以使用模型的一些参数来还原 Defensive distillation 的效果,使得 Defensive distillation 无效.Hosseini 等人[174]在黑盒模型中研究了 Defensive distillation 方法,并证实它不会提高分类器的健壮性.原因是当输入数据的一些特征被修改之后,Defensive distillation 便无效.最近,Papernot 等人[175]提出了扩展的 Defensive distillation:训练 Distilled 的模型时,训练数据的标签不仅包括了原始模型输出的这些样本的分类置信度,还有这些样本的分类标签(hard labels).他们指出,Defensive distillation 引发了 gradient masking 现象[176],即只是破坏攻击所需的损失函数梯度,并没有从实质上解决模型的错误,因此无法抵御 Carlini 等人[172]提出的攻击的原因,而扩展的 Defensive distillation 解决了这一问题.

4.4.2 预处理

2017 年 KDD 会议上 Wang 等人[128]提出 random feature nullification,在 DNN 的训练和测试阶段随机丢弃一些特征,使 DNN 具有不确定性,攻击者难以判断出保留下来的特征,则难以生成对抗样本.相对于原始模型,丢弃了一部分特征之后训练得到的模型的准确率有所降低.他们还提出另外一种防御对抗攻击的方案[177-178]:在 DNN 前面加一个不可逆的数据转换模块来进行数据降维,降维之后的数据是不可还原的,然后将这些数据输入 DNN 进行训练或者测试.这种对特征进行预处理的方法虽然有效,但是作者表示这种方法会损失模型的准确度.

4.4.3 正则化

Gu 等人[179]提出了深度收缩网络(deep contractive network, DCN)来抵御对抗攻击,他们采用的正则化方法使用平滑度惩罚(smoothness penalty)来进行训练,Ororbia 等人[180]采用扩展版的深度收缩自编码器来抵御对抗攻击.Lyu 等人[181]使用一组联合的正则化方法来规范模型的训练过程.除了这些工作以外,文献[182-183]也使用了正则化的方法来抵御对抗攻击,然而,正则化方法虽然有助于抵御对抗攻击,但是依然会使模型的效果(如准确度)略微变差.

4.4.4 对抗训练

Szegedy 等人[156]认为对抗训练可以使模型更安全,对抗训练的核心是构造虚拟对抗样本(virtual

adversarial example)[184],虚拟对抗样本指的是由训练完备的模型提供标签构造的对抗样本,对抗训练指的是把虚拟对抗样本注入训练集来训练要防御的模型.对抗训练使模型对合法样本和对抗样本的分类性能逐渐提高,得到的模型具有更高的健壮性,能更有效地防御对抗攻击.最近,Tramèr 等人[185]提出联合对抗训练(ensemble adversarial training),将多个预先训练好的模型产生的虚拟对抗样本注入训练集,Na 等人[186]也提出了与此相近的理念.文献[187-191]指出,对抗训练虽有效,但仍然不能完全抵挡对抗样本,主要因为在对抗训练过程中,需要把所有可能的虚拟对抗样本涵盖在训练集合中,然而这是不现实的,导致了对抗训练的局限性.

4.4.5 拒绝对抗样本

拒绝对抗样本指的是,通过对测试样本进行检测,检测器直接丢弃被判定为具有对抗性的测试样本,阻止对抗样本访问被保护的分类器.Hosseini 等人[174]提出了一种训练方法:训练模型区分合法样本和对抗样本,直接丢弃对抗样本,仅对合法样本进行分类预测,然而该方案容易产生误警(false positives).Metzen 等人[192]通过附加一个检测子网来观察原始分类网络的状态,以更全面地区分合法样本和对抗样本.Lu 等人[193]提出 SafetyNet,SafetyNet 由分类器和攻击检测器组成,如果检测器声明一个样本是对抗性的,则该样本被攻击者检测器拒绝,不再被分类器分类.

4.5 对抗攻击的防御措施研究进展小结

对抗性环境下的攻击分为白盒攻击和黑盒攻击,前者需要攻击者掌握目标系统的架构及参数信息(甚至训练数据),通过这些信息使用梯度下降算法构造对抗样本.现有的构造对抗样本的方法有 4 种:L-BFGS, Deepfool, Fast Gradient Sign, Jacobian-based Saliency Map.攻击者可使用这些方法,生成目标系统的对抗样本.黑盒攻击并不需要攻击者了解目标系统的信息,通过构建一个替代模型来模拟目标模型,并使用替代模型来构造对抗样本.黑盒攻击的出现,大大提高了防御难度.

我们将现有的对抗攻击防御措施存在的问题和下一步研究方向总结为 3 点:1)对目标模型参数的依赖问题.白盒模型使用的防御措施为,改变目标模型梯度传递过程,然而黑盒攻击使用替代模型来构造对抗样本,并不依赖于目标模型的梯度传递过程,使白盒模型的防御措施无效.2)拒绝对抗样本造成的误报.目标系统直接丢弃置信度较低的可疑对抗样本,不对其进行分类,使得攻击者无从下手.然而,目标模型也有可能拒绝合法样本,这样会造成目标系统的可用性降低.因此,选择一个合适的可识别对抗样本的方案是一个亟需解决的重点难点问题.3)对抗训练.根据本节的分析,我们得出的结论是,只有提高模型自身健壮性的方案,才可以最有效地防御对抗攻击.其他基于梯度的方案,如预处理、正则化和拒绝对抗样本等,都不能从根本上加固目标系统的健壮性,使目标系统依然存在安全隐患.对抗训练是目前最有效的提高模型健壮性,抵御对抗攻击的方案.对抗训练的主要方法是在训练过程中,利用训练完备的模型生成虚拟对抗样本,将虚拟对抗样本加入训练集,迭代这个过程,遍历所有存在的虚拟对抗样本.然而,在实际中,找到所有的虚拟对抗样本具有很大难度,是对抗攻击防御领域亟需攻克的难题.

5 问题 2:协作性模型易受隐私窃取攻击

在深度学习领域,普遍认为数据量越大,训练得到的模型越准确,由于网络安全活动的固有特征,即正常行为的数量远远超过恶意行为的数量,网络安全领域的数据集具有严重的不均衡性.用不均衡的数据集训练出的模型往往会拟合样本数量较多的样本类,也就是正常行为类,使模型产生严重偏倚.因此,为了使模型更精确,能够检测出更多类型的恶意行为,不同的数据提供方往往需要协作来扩大数据量,尤其是恶意行为数据量.对于工业界,目前企业的安全防控业务依赖于大数据,然而企业只处理自己私有的数据,不同企业之间并未对这些数据进行共享,造成了企业数据的封闭性.为了打破这种数据封闭性的状态,可选的 2 种方案是集中式数据源和分布式数据源.传统的集中式数据源,需要多个数据源的数据集中在某个数据操作方,这样会暴露很多数据源的隐私,集中式的数据源存在的隐私风险使其对于很多领域,尤其是网络空间安全领域不再适用,取而代之的是协作性数据源.2017 年 8 月底,Akamai, Cloudflare, Flashpoint, Google, Oracle Dyn, RiskIQ, Team Cymru 等公司的安全团队协作发现了一种新的僵尸网络 WireX[194].WireX 出现于 2017 年 8 月初,主要包含十几万台实施 DDoS 攻击的 Android 僵尸设备.这次的协作使得 WireX 在第一时间被检测出,将其带来的危害降至最低,协作研究对于安全威胁检测以及风险控制的作用可见一斑.

将协作性模型与深度学习相结合,通过多个数据源之间协作来训练更加精确的协作性深度学习模型,在保证不同企业之间数据独立性的基础上,不同的企业之间可以实现安全模型的共享,任何一个安全模型学习到的知识都可以在第一时间内共享给其他协作的企业,使所有企业的安全防御体系得以加强,共同防御风险.然而,协作性模型依然会受到模型反演攻击(model inversion attacks)[195]、污染攻击(poisoning attacks)[196]等隐私窃取攻击的威胁,网络空间安全领域对数据机密性的要求较为严苛.若遭受攻击,比如入侵检测中用到的流量数据,会暴露目标网络的一些重要的配置信息;网站的请求数据,会暴露网站架构以及用户的一些敏感隐私内容,因此安全厂商对于数据的机密性保护给予了高度重视.下面分析了针对基于深度学习的协作性模型面临的攻击,以及可能的防御措施.

5.1 隐私窃取攻击

现有的针对协作性模型的隐私窃取攻击分为3种:

1)模型反演攻击[195].机器学习法具有容易过拟合的缺陷,即模型会隐性记忆一些过拟合的训练数据,2015年CCS会议上,Fredrikson等人[195]提出的模型反演攻击就利用了机器学习算法这一缺陷.对于使用敏感数据训练的模型,模型反演攻击基于梯度下降技术,可以从深度学习模型中重现训练数据:针对人脸识别分类器,使用置信度值可重现训练数据中的人脸图像样本.一个可能的防御措施是降低从模型获取的梯度信息的精度.

2)污染攻击[196].在协作性深度学习环境中,多个用户协作生成更准确的模型,这些用户相互之间是不信任的.攻击者或者恶意的用户方恶意篡改这些用户的输入来误导模型,也就是对协作深度学习模型实施污染攻击.Shen等人[196]提出Auror来抵御污染攻击,Auror的核心思想是污染的训练数据影响了数据的模糊特征(masked features)的分布.协作的用户把原始数据构造得到的模糊特征提交给服务器,由服务器生成全局模型.服务器通过检查这些模糊特征来发现攻击.

3)使用GAN的信息窃取.Hitaj等人[197]提出了一种针对协作性深度学习模型隐私保护机制的攻击,攻击者通过训练GAN,生成与目标训练数据分布相同的无限逼近合法样本的假样本.

5.2 针对隐私窃取的防御

现有的深度学习模型的隐私保护方法是差分隐私(differential privacy)[198-199],差分隐私保证了使用2个不同版本的数据集(比如这2个数据集之间存在一个不相同的样本)训练深度学习模型时,模型的输出不会出现明显的统计差异,此处的样本代表用户隐私.差分隐私的目的是,允许研究者在不泄露单个样本信息的前提下,对一个数据集整体进行分析,比如了解数据集的均值、方差等统计学信息.其原理是,给数据集中注入扰动(也叫噪音)[200],扰动越大,对数据集整体的隐私保护效果就越好,然而,数据的可用性却越差.因此,扰动大小的选取,需要在隐私保护效果和数据可用性之间折中[201].

根据部署差分隐私阶段的不同,深度学习模型中实现差分隐私的方法可以分为基于测试阶段的防御方法和基于训练阶段的防御方法.Xie等人[202]提出了一种测试阶段的防御方法,他们使用同态加密(homomorphic encryption)[203]来加密数据,目的是允许神经网络处理数据的同时不对其进行解密,因此可以保护单个输入的机密性.主要的局限在于性能开销过大,以及同态加密对目标深度学习模型有额外约束.

文献[204-208]介绍了使用训练阶段的差分隐私方法.2016年USENIX会议上,Ohrimenko等人[204]提出了一个隐私保护的多方机器学习系统(multi-party machine learning system).他们提出一种可应用于神经网络的数据遗忘算法(data-oblivious algorithm).数据遗忘算法用于防止由内存、磁盘和网络访问引起的间接隐私泄露.Shokri等人[205]提出隐私保护的分布式随机梯度下降(stochastic gradient descent,SGD)算法,描述了多方在联合输入端训练深度神经网络的方法.他们使用的假设是,各方训练自己的模型,彼此之间并不共享数据,仅在训练期间交换中间参数.他们指出深度模型可以通过扰动参数的多方计算进行训练,并提供差分隐私保证.然而,该技术提供的隐私保证范围是根据大量参数给出的,大量参数不利于提供有力的隐私保证.2016年CCS会议上,Abadi等人[206]提出通过引入时间计算(moments accountant)减小有噪音的SGD引起的隐私损失.然而,应用于文献[205-206]中协作深度学习模型的差分隐私无法抵御Hitaj等人[197]提出的基于GAN的攻击,因为差分隐私模型的训练只与某些特定特征相关联,差分隐私最多可以防止这些特征的重现.基于文献[206]中的时间计算技术,2017年ICLR会议上,Papernot等人[207]改进了保护训练数据隐私的机器

学习模型[208],提出了一种为训练数据提供强有力的隐私保证的方法.该方法首先学习一组用不同的敏感数据集训练的模型,这些模型不公开,被用作教师模型(teacher model).教师模型对公开的未标记的非敏感数据进行预测,用于训练学生模型(student model).此策略确保学生模型不依赖于任何敏感的训练数据集,即使攻击者可以获取学生模型的参数,仍然无法获取任何关于敏感训练数据的信息.然而,此方法的前提是存在可用的未标记非敏感数据,文献[205-206]无需此前提.

5.3 深度学习的隐私保护研究进展小结

深度学习的隐私保护已经出现多种不同的方案,这些方案都存在不同的局限性,现将这些局限性和下一步研究方向总结如下:1)弱化模型性能的问题.数据遗忘算法[204]和分布式 SGD 算法[205-206]都是训练阶段采用的隐私保护手段,主要原理是梯度下降的过程中对模型的参数进行扰乱.此方法虽然可以保护数据隐私,然而代价是大大弱化了深度学习的学习性能.2)负载高.现有的基于梯度的防御措施都需要存储大量参数,对硬件的存储要求较高,使得系统负载巨大.如何减少运算过程中所需的参数,仍然是个难点问题.3)对非敏感数据的依赖性.目前避免了以上 2 个问题的防御方案是教师学生模型[207],然而其用到半监督学习,需要大量的非敏感数据来训练学生模型,因此,如何使脱敏数据仍然保持原本的数据分布,是亟需解决的下一个挑战.

6 未来研究展望

深度学习应用于网络空间安全方面的研究是近 2 年来的研究热点,对于计算机系统和网络安全有非常重要的意义,受到广泛关注.然而将深度学习应用到诸如二进制数据分析和代码分析等网络空间安全领域的研究处于起步阶段,结合目前深度学习应用于网络空间安全的研究存在的问题,我们指出 4 个未来的研究方向:

1) 防御对抗攻击

现有的基于分类的深度学习模型易受对抗攻击,然而对抗攻击及其防御技术目前主要的研究领域是图像处理,据调研,深度学习应用于网络空间安全领域面临的对抗攻击及其防御措施的研究基本上处于空白.防御措施的研究方向是对抗训练,搜索到全部可能的虚拟对抗样本来扩充训练集,是对抗训练的难点问题.

2) 防御针对协作性模型的攻击

协作性深度学习模型通过多个数据源之间协作来训练更加精确的模型,通过加密措施,使不同数据源的数据对彼此是不可见的,不仅保护了数据隐私,也实现了扩大数据量的目的.然而,协作性深度学习模型依然受到模型反演攻击、污染攻击等隐私窃取攻击.可能的研究方向是教师学生模型,然而该模型需要大量的非敏感数据来训练学生模型,但是脱敏数据很容易破坏原始数据的分布情况,该难点也亟需克服.

3) 特征学习

深度学习模型对网络空间安全数据直接进行分类时的性能,相比于机器学习模型,并没有太大提高,因此,不提倡直接使用深度学习模型对网络空间安全数据进行分类,应首先智能化学习有效特征.此外,大规模的数据,如二进制代码等,进行手工特征提取也是不实际的.因此,将深度学习应用于提取有效特征也是一个值得探索的方向.

4) 可解释性

深度学习算法带来高准确率的代价是较差的可解释性,也就是说,深度模型判断出一个软件或者访问请求是恶意时,并不给出任何与判断结果相关的恶意数据段信息.可能的解决方案是采用影响函数和注意力机制,该机制中的权重值机制使得不同的数据段对判别结果的重要性具有可见性.

7 结束语

随着深度学习在不同领域的进步,攻击者正在寻求方案来绕过深度学习模型的限制,并利用深度模型实现其恶意目标.深度学习应用于网络空间安全的研究是近 2 年来的研究热点,仍然处于起步阶段.本文中,我们首先介绍了深度学习的网络空间安全研究涉及的深度模型;其次总结了深度学习在网络空间安全领域的应用,如恶意软件检测、入侵检测等,并归纳了其发展趋势以及存在的问题;随后,我们分别列出了深度学习以及深度学习应用于网络空间安全的研究面临的十大问题与机遇,并介绍了深度学习应用于网络空间安全的研究存在的 2 个最严重的问题,即分类模型易受对抗攻击和协作性模型易受隐私窃取攻击,以及相应的防御方案的研究方向.我们建议在部署深度学习的网络空间安全应用时,应全面检测该应用对于不同的对抗攻击和隐私窃取攻击的抵抗能力,结合多种方法来提高模型的健壮性和机密性.

致谢 感谢宾夕法尼亚州立大学邢新宇教授在本文综述书写过程中参与讨论并给出建议!

参 考 文 献

[1] LeCun Y, Bengio Y, Hinton G. Deep learning [J]. Nature, 2015, 521(7553): 436-444

[2] Goodfellow I, Bengio Y, Courville A. Deep Learning [M]. Cambridge, MA: MIT Press, 2016: 528-566

[3] Deng Li, Yu Dong. Deep learning: Methods and applications [J]. Foundations and Trends in Signal Processing, 2014, 7(3/4): 197-387

[4] Krizhevsky A, Sutskever I, Hinton G E. ImageNet classification with deep convolutional neural networks [J]. Communications of the ACM, 2012, 60(2): 2012-2025

[5] Taigman Y, Yang Ming, Ranzato M A, et al. Deepface: Closing the gap to human-level performance in face verification [C] //Proc of the 29th IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2014: 1701-1708

[6] Dahl G E, Deng Li, Yu Dong, et al. Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition [J]. IEEE Trans on Audio, Speech, and Language Processing, 2012, 20(1): 30-42

[7] Sainath T N, Vinyals O, Senior A, et al. Convolutional, long short-term memory, fully connected deep neural networks [C] //Proc of the 39th IEEE Int Conf on Acoustics, Speech and Signal. Piscataway, NJ: IEEE, 2015: 4580-4584

[8] Sak H, Senior A, Beaufays F. Long short-term memory based recurrent neural network architectures for large vocabulary speech recognition [J]. Computer Science, 2014, 9(3): 338-342

[9] Collobert R, Weston J. A unified architecture for natural language processing: Deep neural networks with multitask learning [C] //Proc of the 25th Int Conf on Machine Learning. New York: ACM, 2008: 160-167

[10] Cruz-Roa A, Ovalle J E A, Madabhushi A, et al. A deep learning architecture for image representation, visual interpretability and automated basal-cell carcinoma cancer detection [C] //Proc of the 9th Int Conf on Medical Image Computing and Computer-Assisted Intervention. Berlin: Springer, 2013: 403-410

[11] Huang Wenyi, Stokes J W. MtNet: A multi-task neural network for dynamic malware classification [C] //Proc of the 5th Int Conf on Detection of Intrusions and Malware, and Vulnerability Assessment. Berlin: Springer, 2016: 399-418

[12] LeCun Y, Jackel L D, Boser B, et al. Handwritten digit recognition: Applications of neural network chips and automatic learning [J]. IEEE Communications Magazine, 1989, 27(11): 41-46

[13] Rumelhart D E, Hinton G E, Williams R J. Learning representations by back-propagating errors [J]. Cognitive Modeling, 1988, 5(3): 533-536

[14] Hinton G E, Osindero S, Teh Y W. A fast learning algorithm for deep belief nets [J]. Neural Computation, 2006, 18(7): 1527-1554

[15] Hinton G E, McClelland J L. Learning representations by recirculation [C] //Proc of the 1st Int Conf on Neural Information Systems. Cambridge, MA: MIT Press, 1987: 358-366

[16] Papernot N, McDaniel P, Sinha A, et al. Towards the science of security and privacy in machine learning [OL]. 2016[2017-07-12]. https://arxiv.org/pdf/1611.03814.pdf

[17] Nguyen A, Yosinski J, Clune J. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images [C] //Proc of the 30th IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2015: 427-436

[18] Bolukbasi T, Wang J, Dekel O, et al. Adaptive neural networks for efficient inference [C] //Proc of the 34th Int Conf on Machine Learning. New York: ACM, 2017: 527-536

[19] Koh P W, Liang P. Understanding black-box predictions via influence functions [OL]. 2017[2017-08-11]. https://arxiv.org/pdf/1703.04730.pdf

[20] Srivastava N, Hinton G E, Krizhevsky A, et al. Dropout: A simple way to prevent neural networks from overfitting [J]. Journal of Machine Learning Research, 2014, 15(1): 1929-1958

[21] RSIP. Deep learning and convolutional neural networks: RSIP vision blogs [EB/OL]. 2016 [2017-08-18]. http://www.rsipvision.com/exploring-deep-learning/

[22] Deng Li. Three classes of deep learning architectures and their applications: A tutorial survey [J]. APSIPA Trans on Signal and Information Processing, 2012, 11(2): 1132-1160

[23] 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

[24] Szegedy C, Liu Wei, Jia Yangqing, et al. Going deeper with convolutions [C] //Proc of the 30th IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2015: 37-46

[25] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition [OL]. 2014[2017-08-02]. https://arxiv.org/pdf/1409.1556.pdf

[26] He Kaiming, Zhang Xiangyu, Ren Shaoqing, et al. Deep residual learning for image recognition [C] //Proc of the 31st IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2016: 770-778

[27] Eigen D, Rolfe J, Fergus R, et al. Understanding deep architectures using a recursive convolutional network [J]. Computer Science, 2014, 10(2): 38-55

[28] Masci J, Meier U, Cireşan D, et al. Stacked convolutional auto-encoders for hierarchical feature extraction [C] //Proc of the 20th Int Conf on Artificial Neural Networks. Berlin: Springer, 2011: 52-59

[29] Krizhevsky A, Hinton G. Convolutional deep belief networks on cifar-10 [J]. Unpublished Manuscript, 2010, 7(6): 1007-1020

[30] Cho K, Van Merriënboer B, Gulcehre C, et al. Learning phrase representations using RNN encoder-decoder for statistical machine translation [J]. Computer Science, 2014, 10(2): 187-205

[31] Sutskever I, Vinyals O, Le Q V. Sequence to sequence learning with neural networks [J]. IEEE Trans on Signal Processing, 2014, 4(2): 3104-3112

[32] Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate [J]. Computer Science, 2014, 7(2): 109-136

[33] Schuster M, Paliwal K K. Bidirectional recurrent neural networks [J]. IEEE Trans on Signal Processing, 1997, 45(11): 2673-2681

[34] Graves A. Supervised Sequence Labelling with Recurrent Neural Networks [M]. Berlin: Springer, 2012: 4-38

[35] Graves A, Liwicki M, Bunke H, et al. Unconstrained on-line handwriting recognition with recurrent neural networks [C] //Proc of the 29th Conf on Neural Information Processing Systems. Piscataway, NJ: IEEE, 2007: 458-64

[36] Graves A, Schmidhuber J. Offline handwriting recognition with multidimensional recurrent neural networks [C] //Proc of the 30th Int Conf on Neural Information Processing Systems. Piscataway, NJ: IEEE, 2008: 545-552

[37] Graves A, Schmidhuber J. Framewise phoneme classification with bidirectional LSTM and other neural network architectures [J]. Neural Networks, 2005, 18(5): 602-610

[38] Graves A. Generating sequences with recurrent neural networks [J]. Computer Science, 2013, 10(3): 30-45

[39] Baldi P, Brunak S, Frasconi P, et al. Exploiting the past and the future in protein secondary structure prediction [J]. Bioinformatics, 1999, 15(11): 937-946

[40] Gers F A, Schmidhuber J, Cummins F. Learning to forget: Continual prediction with LSTM [J]. Neural Computation, 2000, 12(10): 2451-247

[41] Cho K, Van Merriënboer B, Bahdanau D, et al. On the properties of neural machine translation: Encoder-decoder approaches [J]. Computer Science, 2014, 7(2): 103-111

[42] Chung J, Gulcehre C, Cho K H, et al. Empirical evaluation of gated recurrent neural networks on sequence modeling [OL]. 2014 [2017-08-11]. https://arxiv.org/pdf/1412.3555

[43] Chung J, Gulcehre C, Cho K, et al. Gated feedback recurrent neural networks [C] //Proc of the 32nd Int Conf on Machine Learning. New York: ACM, 2015: 2067-2075

[44] Jozefowicz R, Zaremba W, Sutskever I. An empirical exploration of recurrent network architectures [C] //Proc of the 32nd Int Conf on Machine Learning. New York: ACM, 2015: 2342-2350

[45] Chrupała G, Kádár A, Alishahi A. Learning language through pictures [J]. Computer Science, 2015, 8(2): 76-90

[46] Hochreiter S, Schmidhuber J. Long short-term memory [J]. Neural Computation, 1997, 9(8): 1735-1780

[47] Graves A, Jaitly N. Towards end-to-end speech recognition with recurrent neural networks [C] //Proc of the 31st Int Conf on Machine Learning. New York: ACM, 2014: 1764-1772

[48] Kiros R, Salakhutdinov R, Zemel R S. Unifying visual-semantic embeddings with multimodal neural language models [J]. Computer Science, 2014, 10(3): 137-152

[49] Vinyals O, Toshev A, Bengio S, et al. Show and tell: A neural image caption generator [C] //Proc of the 30th IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2015: 3156-3164

[50] Xu K, Ba J, Kiros R, et al. Show, attend and tell: Neural image caption generation with visual attention [C] //Prof of the 32nd Int Conf on Machine Learning. New York: ACM, 2015: 2048-2057

[51] Vinyals O, Kaiser Ł, Koo T, et al. Grammar as a foreign language [C] //Proc of the 28th Conf on Advances in Neural Information Processing Systems. Cambridge, MA: MIT Press, 2015: 2773-2781

[52] Pascanu R, Gulcehre C, Cho K, et al. How to construct deep recurrent neural networks [J]. Computer Science, 2013, 6(5): 90-109

[53] Salakhutdinov R, Mnih A, Hinton G. Restricted Boltzmann machines for collaborative filtering [C] //Proc of the 24th Int Conf on Machine Learning. New York: ACM, 2007: 791-798

[54] Salakhutdinov R, Hinton G. Deep Boltzmann machines [J]. Journal of Machine Learning Research, 2009, 5(2): 196-2006

[55] Rozanov Y A. Markov Random Fields [M]. Berlin: Springer, 1982: 55-102

[56] Elfwing S, Uchibe E, Doya K. Expected energy-based restricted Boltzmann machine for classification [J]. Neural Networks, 2015, 64(2): 29-38

[57] Mnih V, Larochelle H, Hinton G E. Conditional restricted Boltzmann machines for structured output prediction [C] // Proc of the 27th Conf on Uncertainty in Artificial Intelligence. Berlin: Springer, 2011: 514-522

[58] Taylor G W, Hinton G E, Roweis S T. Two distributed-state models for generating high-dimensional time series [J]. Journal of Machine Learning Research, 2011, 12(3): 1025-1068

[59] Hinton G. A practical guide to training restricted Boltzmann machines [J]. Momentum, 2010, 9(1): 926-937

[60] Tang Yichuan, Salakhutdinov R, Hinton G. Robust Boltzmann machines for recognition and denoising [C] //Proc of the 27th Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2012: 2264-2271

[61] Li Guoqi, Deng Lei, Xu Yi, et al. Temperature based restricted Bzoltzmann machines [J]. Scientific Reports, 2016, 6(2): 191-210

[62] Nair V, Hinton G E. 3D object recognition with deep belief nets [C] //Proc of the 22nd Conf on Advances in Neural Information Processing Systems. Cambridge, MA: MIT Press, 2009: 1339-1347

[63] Indiveri G, Liu S. Memory and information processing in neuromorphic systems [J]. Proceedings of the IEEE, 2015, 103(8): 1379-1397

[64] Liao Bin, Xu Jungang, Lü Jintao, et al. An image retrieval method for binary images based on DBN and softmax classifier [J]. IETE Technical Review, 2015, 32(4): 294-303

[65] Abdel-Zaher A M, Eldeib A M. Breast cancer classification using deep belief networks [J]. Expert Systems with Applications, 2016, 46(2): 139-144

[66] Deng Li, Yu Dong. Deep convex net: A scalable architecture for speech pattern classification [C] //Proc of the 12th Conf of the Int Speech Communication Association. Florence, Italy: ISCA, 2011: 2285-2288

[67] Hinton G E, Salakhutdinov R R. Using deep belief nets to learn covariance kernels for Gaussian processes [C] //Proc of the 1st Conf on Advances in Neural Information Processing Systems. Cambridge, MA: MIT Press, 2008: 1249-1256

[68] Arel I, Rose D C, Karnowski T P. Deep machine learning-a new frontier in artificial intelligence research [J]. IEEE Computational Intelligence Magazine, 2010, 5(4): 13-18

[69] Bengio Y. Learning deep architectures for AI [J]. Foundations and Trends in Machine Learning, 2009, 2(1): 1-127

[70] Alain G, Bengio Y. What regularized autoencoders learn from the data-generating distribution [J]. The Journal of Machine Learning Research, 2014, 15(1): 3563-3593

[71] Hinton G E, Salakhutdinov R R. Reducing the dimensionality of data with neural networks [J]. Science, 2006, 313(5786): 504-507

[72] Schmidhuber J. Deep learning in neural networks: An overview [J]. Neural Networks, 2015, 61(9): 85-117

[73] Makhzani A, Frey B. K-sparse autoencoders [OL]. 2013 [2017-06-14]. https://arxiv.org/pdf/1312.5663

[74] Vincent P. A connection between score matching and denoising autoencoders [J]. Neural Computation, 2011, 23(7): 1661-1674

[75] Vincent P, Larochelle H, Bengio Y, et al. Extracting and composing robust features with denoising autoencoders [C] //Proc of the 25th Int Conf on Machine Learning. New York: ACM, 2008: 1096-1103

[76] Ling Zhenhua, Kang Shiyin, Zen H, et al. Deep learning for acoustic modeling in parametric speech generation: A systematic review of existing techniques and future trends [J]. IEEE Signal Processing Magazine, 2015, 32(3): 35-52

[77] Kamyshanska H, Memisevic R. The potential energy of an autoencoder [J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 2015, 37(6): 1261-1273

[78] Bengio Y, Courville A, Vincent P. Representation learning: A review and new perspectives [J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 2013, 35(8): 1798-1828

[79] Vincent P, Larochelle H, Lajoie I, et al. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion [J]. Journal of Machine Learning Research, 2010, 11(12): 3371-3408

[80] Rifai S, Vincent P, Muller X, et al. Contractive autoencoders: Explicit invariance during feature extraction [C] //Proc of the 28th Int Conf on Machine Learning. New York: ACM, 2011: 833-840

[81] Rifai S, Mesnil G, Vincent P, et al. Higher order contractive autoencoder [C] //Proc of the 24th European Conf on Machine Learning and Knowledge Discovery in Databases. Berlin: Springer, 2011: 645-660

[82] Sun Meng, Zhang Xiongwei, Zheng T F. Unseen noise estimation using separable deep auto encoder for speech enhancement [J]. IEEE/ACM Trans on Audio, Speech, and Language Processing, 2016, 24(1): 93-104

[83] Staudemeyer R C. Applying long short-term memory recurrent neural networks to intrusion detection [J]. South African Computer Journal, 2015, 56(1): 136-154

[84] Dahl G E, Stokes J W, Deng Li, et al. Large-scale malware classification using random projections and neural networks [C] //Proc of the 38th Int Conf on Acoustics, Speech and Signal Processing. Piscataway, NJ: IEEE, 2013: 3422-3426

[85] Kolosnjaji B, Zarras A, Webster G, et al. Deep learning for classification of malware system call sequences [C] //Proc of the 30th Australasian Joint Conf on Artificial Intelligence. Berlin: Springer, 2016: 137-149

[86] Tobiyama S, Yamaguchi Y, Shimada H, et al. Malware detection with deep neural network using process behavior [C] //Proc of the 8th Int Conf on Computer Software and Applications. Piscataway, NJ: IEEE, 2016: 577-582

[87] Pascanu R, Stokes J W, Sanossian H, et al. Malware classification with recurrent networks [C] //Proc of the 40th Int Conf on Acoustics, Speech and Signal Processing. Piscataway, NJ: IEEE, 2015: 1916-1920

[88] Athiwaratkun B, Stokes J W. Malware classification with LSTM and GRU language models and a character-level CNN [C] //Proc of the 42nd Int Conf on Acoustics, Speech and Signal Processing. Piscataway, NJ: IEEE, 2017: 2482-2486

[89] Wang Xin, Yiu S M. A multi-task learning model for malware classification with useful file access pattern from API call sequence [OL]. 2016 [2017-08-17]. https://arxiv.org/pdf/1610.05945

[90] Hardy W, Chen Lingwei, Hou Shifu, et al. DL4MD: A deep learning framework for intelligent malware detection [C] //Proc of the 16th Int Conf on Data Mining. Piscataway, NJ: IEEE, 2016: 61-68

[91] Rhode M, Burnap P, Jones K. Early stage malware prediction using recurrent neural networks [OL]. 2017 [2017-06-14]. https://arxiv.org/pdf/1708.03513

[92] Nix R, Zhang Jian. Classification of Android apps and malware using deep neural networks [C] //Proc of the 17th Int Joint Conf on Neural Networks. Piscataway, NJ: IEEE, 2017: 1871-1878

[93] Saxe J, Berlin K. Deep neural network-based malware detection using two-dimensional binary program features [C] //Proc of the 10th Int Conf on Malicious and Unwanted Software. Piscataway, NJ: IEEE, 2015: 11-20

[94] Shin E C R, Song D, Moazzezi R. Recognizing functions in binaries with neural networks [C] //Proc of the 24th USENIX Security Symp. Berkely, CA: USENIX Association, 2015: 611-626

[95] Yuan Zhenlong, Lu Yongqiang, Wang Zhaoguo, et al. Droid-Sec: Deep learning in Android malware detection [J]. ACM SIGCOMM Computer Communication Review, 2014, 44(4): 371-372

[96] Yuan Zhenlong, Lu Yongqiang, Xue Yibo. DroidDetector: Android malware characterization and detection using deep learning [J]. Tsinghua Science and Technology, 2016, 21(1): 114-123

[97] Xu Lifan, Zhang Dongping, Jayasena N, et al. HADM: Hybrid analysis for detection of malware [C] //Proc of the 3rd SAI Intelligent Systems Conf. Berlin: Springer, 2016: 702-724

[98] Jung W, Kim S, Choi S. Poster: Deep learning for zero-day flash malware detection [C] //Proc of the 36th IEEE Symp on Security and Privacy. Piscataway, NJ: IEEE, 2015: 32-34

[99] Li Ping, Hastie T J, Church K W. Very sparse random projections [C] //Proc of the 12th ACM SIGKDD Int Conf on Knowledge Discovery and Data Mining. New York: ACM, 2006: 287-296

[100] David O E, Netanyahu N S. Deepsign: Deep learning for automatic malware signature generation and classification [C] //Proc of the 12th Int Joint Conf on Neural Networks. Piscataway, NJ: IEEE, 2015: 76-84

[101] Debar H, Becker M, Siboni D. A neural network component for an intrusion detection system [C] //Proc of the 23rd Computer Society Symp on Research in Security and Privacy. Piscataway, NJ: IEEE, 1992: 240-250

[102] Creech G, Hu Jiankun. A semantic approach to host-based intrusion detection systems using contiguous and discontiguous system call patterns [J]. IEEE Trans on Computers, 2014, 63(4): 807-819

[103] Fiore U, Palmieri F, Castiglione A, et al. Network anomaly detection with the restricted Boltzmann machine [J]. Neurocomputing, 2013, 122(3): 13-23

[104] University of California. KDD Cup 99 [EB/OL]. 1999 [2017-08-18]. http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html

[105] Tavallaee M, Bagheri E, Lu Wei, et al. A detailed analysis of the KDD CUP 99 data set [C] //Proc of the 2nd IEEE Symp on Computational Intelligence for Security and Defense Applications. Piscataway, NJ: IEEE, 2009: 1-6

[106] Canadian Institute for Cybersecurity. NSL-KDD dataset [EB/OL]. 2017 [2017-08-18]. http://www.unb.ca/cic/research/datasets/nsl.html

[107] Kim J, Kim J, Thu H L T, et al. Long short-term memory recurrent neural network classifier for intrusion detection [C] //Proc of the 22nd Int Conf on Platform Technology and Service. Piscataway, NJ: IEEE, 2016: 49-54

[108] Putchala M K, Deep learning approach for intrusion detection system (IDS) in the Internet of things (IoT) network using gated recurrent neural networks (GRU)[D]. Dayton, Ohio, USA: Wright State University, 2017

[109] Gao Ni, Gao Ling, Gao Quanli, et al. An intrusion detection model based on deep belief networks [C] //Proc of the 12th Int Conf on Advanced Cloud and Big Data. Piscataway, NJ: IEEE, 2014: 247-252

[110] Li Yuancheng, Ma Rong, Jiao Runhai. A hybrid malicious code detection method based on deep learning [J]. International Journal of Software Engineering & Its Applications, 2015, 9(5): 205-216

[111] Salama M A, Eid H F, Ramadan R A, et al. Hybrid intelligent intrusion detection scheme [G] //Soft Computing in Industrial Applications. Berlin: Springer, 2011: 293-303

[112] Niyaz Q, Javaid A, Sun W, et al. A deep learning approach for network intrusion detection system [C] //Proc of the 9th EAI Int Conf on Bio-inspired Information and Communications Technologies. New York: ACM, 2016: 21-26

[113] Abolhasanzadeh B. Nonlinear dimensionality reduction for intrusion detection using autoencoder bottleneck features [C] //Proc of the 7th Conf on Information and Knowledge Technology. Piscataway, NJ: IEEE, 2015: 26-31

[114] Alom M Z, Bontupalli V R, Taha T M. Intrusion detection using deep belief networks [C] //Proc of the 9th Conf on Aerospace and Electronics. Piscataway, NJ: IEEE, 2015: 339-344

[115] Aygun R C, Yavuz A G. Network anomaly detection with stochastically improved autoencoder based models [C] //Proc of the 4th Int Conf on Cyber Security and Cloud Computing. Piscataway, NJ: IEEE, 2017: 193-198

[116] Wang Zhanyi. The applications of deep learning on traffic identification [EB/OL]. 2015 [2017-07-15]. https://www.blackhat.com/docs/us-15/materials/us-15-Wang-The-Applications-Of-Deep-Learning-On-Traffic-Identification-wp.pdf

[117] Yu Yang, Long Jun, Cai Zhiping. Network intrusion detection through stacking dilated convolutional autoencoders [J]. Security and Communication Networks, 2017, 2(3): 212-225

[118] Wang Wei, Zhu Ming, Zeng Xuewen, et al. Malware traffic classification using convolutional neural network for representation learning [C] //Proc of the 1st Int Conf on Information Networking. Piscataway, NJ: IEEE, 2017: 712-717

[119] Kim G, Yi H, Lee J, et al. LSTM-based system-call language modeling and robust ensemble method for designing host-based intrusion detection systems [OL]. 2016 [2017-08-02]. https://arxiv.org/pdf/1611.01726

[120] Yu Yang, Long Jum, Cai Zhiping. Session-based network intrusion detection using a deep learning architecture [C] // Proc of the 14th Conf on Modeling Decisions for Artificial Intelligence. Berlin: Springer, 2017: 144-155

[121] Zaheer M, Tristan J B, Wick M L, et al. Learning a static analyzer: A case study on a toy language [EB/OL]. 2016 [2017-08-17]. https://openreview.net/references/pdf?id=ry54RWtxx

[122] Godefroid P, Peleg H, Singh R. Learn&-fuzz: Machine learning for input fuzzing [OL]. 2017 [2017-08-17]. https://arxiv.org/pdf/1701.07232

[123] Melicher W, Ur B, Segreti S M, et al. Fast, lean, and accurate: Modeling password guessability using neural networks [C] //Proc of the 25th USENIX Security Symp. Berkely, CA: USENIX Association, 2016: 175-191

[124] Hu Wei, Tan Ying. Generating adversarial malware examples for black-box attacks based on GAN [OL]. 2017 [2017-06-14]. https://arxiv.org/pdf/1702.05983

[125] Hu Weiwei, Tan Ying. Black-box attacks against RNN based malware detection algorithms [OL]. 2017 [2017-08-02]. https://arxiv.org/pdf/1705.08131

[126] Rosenberg I, Shabtai A, Rokach L, et al. Generic black-box end-to-end attack against RNNs and other API calls based malware classifiers [OL]. 2017 [2017-06-14]. https://arxiv.org/pdf/1707.05970

[127] Grosse K, Papernot N, Manoharan P, et al. Adversarial perturbations against deep neural networks for malware classification [OL]. 2016 [2017-08-19]. https://arxiv.org/pdf/1606.04435

[128] Wang Qinglong, Guo Wenbo, Zhang Kaixuan, et al. Adversary resistant deep neural networks with an application to malware detection [C] //Proc of the 23rd ACM SIGKDD Int Conf on Knowledge Discovery and Data Mining. New York: ACM, 2017: 1145-1153

[129] DARPA. Explainable artificial intelligence (XAI) [EB/OL]. 2016 [2017-08-18]. https://www.darpa.mil/program/explainable-artificial-intelligence

[130] Ribeiro M T, Singh S, Guestrin C. Why should I trust you? Explaining the predictions of any classifier [C] //Proc of the 22nd ACM SIGKDD Int Conf on Knowledge Discovery and Data Mining. New York: ACM, 2016: 1135-1144

[131] Goodfellow I J, Vinyals O, Saxe A M. Qualitatively characterizing neural network optimization problems [OL]. 2014 [2017-08-19]. https://arxiv.org/pdf/1412.6544

[132] Weston J, Chopra S, Bordes A. Memory networks [OL]. 2014 [2017-07-15]. https://arxiv.org/pdf/1410.3916

[133] Kumar A, Irsoy O, Ondruska P, et al. Ask me anything: Dynamic memory networks for natural language processing [C] //Proc of the 33rd Int Conf on Machine Learning. New York: ACM, 2016: 1378-1387

[134] Goodfellow I, Pouget-Abadie J, Mirza M, et al. Generative adversarial nets [C] //Proc of the 27th Conf on Advances in Neural Information Processing Systems. Cambridge, MA: MIT Press, 2014: 2672-2680

[135] Wang Kunfeng, Gou Chao, Duan Yanjie, et al. Generative adversarial networks: The state of the art and beyond [J]. Acta Automatica Sinca, 2017, 43(3): 321-332 (in Chinese) (王坤峰, 苟超, 段艳杰, 等. 生成式对抗网络 GAN 的研究进展与展望[J]. 自动化学报, 2017, 43(3): 321-332)

[136] Mirza M, Osindero S. Conditional generative adversarial nets [OL]. 2014 [2017-08-02]. https://arxiv.org/pdf/1411.1784

[137] Odena A. Semi-supervised learning with generative adversarial networks [OL]. 2016 [2017-07-15]. https://arxiv.org/pdf/1606.01583

[138] Donahue J, Krähenbühl P, Darrell T. Adversarial feature learning [OL]. 2016 [2017-08-19]. https://arxiv.org/pdf/1605.09782

[139] Chen Xi, Duan Yan, Houthooft R, et al. InfoGAN: Interpretable representation learning by information maximizing generative adversarial nets [C] //Proc of the 29th Conf on Advances in Neural Information Processing Systems. Cambridge, MA: MIT Press, 2016: 2172-2180

[140] Odena A, Olah C, Shlens J. Conditional image synthesis with auxiliary classifier GANs [OL]. 2016 [2017-08-13]. https://arxiv.org/pdf/1610.09585

[141] Yu Lantao, Zhang Weinan, Wang Jun, et al. SeqGAN: Sequence generative adversarial nets with policy gradient [C] //Proc of the 31st Conf on Artificial Intelligence. Menlo Park, CA: AAAI, 2017: 2852-2858

[142] Radford A, Metz L, Chintala S. Unsupervised representation learning with deep convolutional generative adversarial networks [OL]. 2015 [2017-07-16]. https://arxiv.org/pdf/1511.06434

[143] Denton E L, Chintala S, Fergus R. Deep generative image models using a Laplacian pyramid of adversarial networks [C] //Proc of the 28th Conf on Advances in Neural Information Processing Systems. Cambridge, MA: MIT Press, 2015: 1486-1494

[144] Larsen A B L, Sønderby S K, Larochelle H, et al. Autoencoding beyond pixels using a learned similarity metric [OL]. 2015 [2017-07-16]. https://arxiv.org/pdf/1512.09300

[145] Im D J, Kim C D, Jiang Hui, et al. Generating images with recurrent adversarial networks [OL]. 2016 [2017-08-13]. https://arxiv.org/pdf/1602.05110

[146] Arjovsky M, Chintala S, Bottou L. Wasserstein GAN [OL]. 2017 [2017-07-25]. https://arxiv.org/pdf/1701.07875

[147] Salimans T, Goodfellow I, Zaremba W, et al. Improved techniques for training GANs [C] //Proc of the 29th Conf on Advances in Neural Information Processing Systems. Cambridge, MA: MIT Press, 2016: 2234-2242

[148] Goodfellow I. NIPS 2016 tutorial: Generative adversarial networks [OL]. 2016 [2017-07-24]. https://arxiv.org/pdf/1701.00160

[149] NIPS. Non-targeted adversarial attack [EB/OL]. 2017 [2017-07-24]. https://www.kaggle.com/nips-2017-adversarial-learning-competition

[150] Pascanu R, Mikolov T, Bengio Y. On the difficulty of training recurrent neural networks [C] //Proc of the 30th Int Conf on Machine Learning. New York: ACM, 2013: 1310-1318

[151] Andreas J, Rohrbach M, Darrell T, et al. Neural module networks [C] //Proc of the 31st IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2016: 39-48

[152] McDaniel P, Papernot N, Celik Z B. Machine learning in adversarial settings [J]. IEEE Security & Privacy, 2016, 14(3): 68-72

[153] Huang Ling, Joseph A D, Nelson B, et al. Adversarial machine learning [C] //Proc of the 4th ACM Workshop on Security and Artificial Intelligence. New York: ACM, 2011: 43-58

[154] Biggio B, Corona I, Maiorca D, et al. Evasion attacks against machine learning at test time [C] //Proc of the 23rd Joint European Conf on Machine Learning and Knowledge Discovery in Databases. Berlin: Springer, 2013: 387-402

[155] Biggio B, Fumera G, Roli F. Pattern recognition systems under attack: Design issues and research challenges [J]. International Journal of Pattern Recognition and Artificial Intelligence, 2014, 28(7): 146-158

[156] Szegedy C, Zaremba W, Sutskever I, et al. Intriguing properties of neural networks [OL]. 2013 [2017-08-02]. https://arxiv.org/pdf/1312.6199

[157] Goodfellow I J, Shlens J, Szegedy C. Explaining and harnessing adversarial examples [OL]. 2014 [2017-07-27]. https://arxiv.org/pdf/1412.6572

[158] Warde-Farley D, Goodfellow I, Hazan T, et al. Perturbations, Optimization, and Statistics [M]. Cambridge, MA: MIT Press, 2016: 1-32

[159] Cireşan D, Meier U, Masci J, et al. Multi-column deep neural network for traffic sign classification [J]. Neural Networks, 2012, 32(Special Issue): 333-338

[160] Papernot N, McDaniel P, Jha S, et al. The limitations of deep learning in adversarial settings [C] //Proc of the 1st European Symp on Security and Privacy. Piscataway, NJ: IEEE, 2016: 372-387

[161] Moosavi-Dezfooli S M, Fawzi A, Frossard P. Deepfool: A simple and accurate method to fool deep neural networks [C] //Proc of the 31st IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2016: 2574-2582

[162] Kurakin A, Goodfellow I, Bengio S. Adversarial examples in the physical world [OL]. 2016 [2017-07-27]. https://arxiv.org/pdf/1607.02533

[163] Papernot N, McDaniel P, Goodfellow I, et al. Practical black-box attacks against machine learning [C] //Proc of the 12th ACM Asia Conf on Computer and Communications Security. New York: ACM, 2017: 506-519

[164] Papernot N, McDaniel P, Goodfellow I. Transferability in machine learning: From phenomena to black-box attacks using adversarial samples [OL]. 2016 [2017-07-19]. https://arxiv.org/pdf/1605.07277

[165] Tramèr F, Papernot N, Goodfellow I, et al. The space of transferable adversarial examples [OL]. 2017 [2017-07-19]. https://arxiv.org/pdf/1704.03453

[166] Moosavi-Dezfooli S M, Fawzi A, Fawzi O, et al. Universal adversarial perturbations [C] //Proc of the 32nd IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2017: 893-901

[167] Russakovsky O, Deng Jia, Su Hao, et al. ImageNet large scale visual recognition challenge [J]. International Journal of Computer Vision, 2015, 115(3): 211-252

[168] Liu Yanpei, Chen Xinyun, Liu Chang, et al. Delving into transferable adversarial examples and black-box attacks [OL]. 2016 [2017-08-02]. https://arxiv.org/pdf/1611.02770

[169] Papernot N, McDaniel P, Swami A, et al. Crafting adversarial input sequences for recurrent neural networks [C] //Proc of the 35th Military Communications Conf. Piscataway, NJ: IEEE, 2016: 49-54

[170] Papernot N, McDaniel P, Wu Xi, et al. Distillation as a defense to adversarial perturbations against deep neural networks [C] //Proc of the 37th IEEE Symp on Security and Privacy. Piscataway, NJ: IEEE, 2016: 582-597

[171] Hinton G, Vinyals O, Dean J. Distilling the knowledge in a neural network [OL]. 2015 [2017-08-14]. https://arxiv.org/pdf/1503.02531

[172] Carlini N, Wagner D. Towards evaluating the robustness of neural networks [C] //Proc of the 38th IEEE Symp on Security and Privacy. Piscataway, NJ: IEEE, 2017: 39-57

[173] Carlini N, Wagner D. Defensive distillation is not robust to adversarial examples [OL]. 2016 [2017-08-09]. https://arxiv.org/pdf/1607.04311

[174] Hosseini H, Chen Yize, Kannan S, et al. Blocking transferability of adversarial examples in black-box learning systems [OL]. 2017 [2017-08-09]. https://arxiv.org/pdf/1703.04318

[175] Papernot N, McDaniel P. Extending defensive distillation [OL]. 2017 [2017-08-09]. https://arxiv.org/pdf/1705.05264

[176] Brendel W, Bethge M. Comment on "biologically inspired protection of deep networks from adversarial attacks"[OL]. 2017 [2017-08-09]. https://arxiv.org/pdf/1704.01547

[177] Wang Qinglong, Guo Wenbo, Zhang Kaixuan, et al. Learning adversary-resistant deep neural networks [OL]. 2016 [2017-08-16]. https://arxiv.org/pdf/1612.01401

[178] Wang Qinglong, Guo Wenbo, Ororbia I I, et al. Using non-invertible data transformations to build adversary-resistant deep neural networks [OL]. 2016 [2017-08-16]. https://arxiv.org/pdf/1610.01934

[179] Gu Shixiang, Rigazio L. Towards deep neural network architectures robust to adversarial examples [OL]. 2014 [2017-08-16]. https://arxiv.org/pdf/1412.5068

[180] Ororbia I I, Alexander G, Giles C L, et al. Unifying adversarial training algorithms with flexible deep data gradient regularization [OL]. 2016 [2017-06-06]. https://arxiv.org/pdf/1601.07213

[181] Lyu C, Huang Kaizhu, Liang Haining. A unified gradient regularization family for adversarial examples [C] //Proc of the 15th Int Conf on Data Mining. Piscataway, NJ: IEEE, 2015: 301-309

[182] Zhao Qiyang, Griffin L D. Suppressing the unusual: Towards robust cnns using symmetric activation functions [OL]. 2016 [2017-08-14]. https://arxiv.org/pdf/1603.05145

[183] Rozsa A, Gunther M, Boult T E. Towards robust deep neural networks with BANG [OL]. 2016 [2017-06-06]. https://arxiv.org/pdf/1612.00138

[184] Miyato T, Maeda S, Koyama M, et al. Distributional smoothing with virtual adversarial training [OL]. 2015 [2017-06-06]. https://arxiv.org/pdf/1507.00677

[185] Tramèr F, Kurakin A, Papernot N, et al. Ensemble adversarial training: Attacks and defenses [OL]. 2017 [2017-06-04]. https://arxiv.org/pdf/1705.07204

[186] Na T, Ko J H, Mukhopadhyay S. Cascade adversarial machine learning regularized with a unified embedding [OL]. 2017 [2017-06-04]. https://arxiv.org/pdf/1708.02582

[187] Shaham U, Yamada Y, Negahban S. Understanding adversarial training: Increasing local stability of neural nets through robust optimization [OL]. 2015 [2017-09-14]. https://arxiv.org/pdf/1511.05432

[188] Huang Ruitong, Xu Bing, Schuurmans D, et al. Learning with a strong adversary [OL]. 2015 [2017-09-17]. https://arxiv.org/pdf/1511.03034

[189] Nøkland A. Improving back-propagation by adding an adversarial gradient [OL]. 2015 [2017-09-12]. https://arxiv.org/pdf/1510.04189

[190] Demyanov S, Bailey J, Kotagiri R, et al. Invariant backpropagation: How to train a transformation-invariant neural network [OL]. 2015 [2017-08-02]. https://arxiv.org/pdf/1502.04434

[191] Grosse K, Manoharan P, Papernot N, et al. On the (statistical) detection of adversarial examples [OL]. 2017 [2017-09-12]. https://arxiv.org/pdf/1702.06280

[192] Metzen J H, Genewein T, Fischer V, et al. On detecting adversarial perturbations [OL]. 2017 [2017-09-12]. https://arxiv.org/pdf/1702.04267

[193] Lu Jiajun, Issaranon T, Forsyth D. SafetyNet: Detecting and rejecting adversarial examples robustly [OL]. 2017 [2017-08-02]. https://arxiv.org/pdf/1704.00103

[194] Cloudflare. The wireX botnet: How industry collaboration disrupted a DDoS attack [EB/OL]. 2017 [2017-09-02] https://blog.cloudflare.com/the-wirex-botnet/

[195] Fredrikson M, Jha S, Ristenpart T. Model inversion attacks that exploit confidence information and basic countermeasures [C] //Proc of the 22nd ACM SIGSAC Conf on Computer and Communications Security. New York: ACM, 2015: 1322-1333

[196] Shen Shiqi, Tople S, Saxena P. Auror: Defending against poisoning attacks in collaborative deep learning systems [C] //Proc of the 32nd Annual Conf on Computer Security Applications. New York: ACM, 2016: 508-519

[197] Hitaj B, Ateniese G, Perez-Cruz F. Deep models under the GAN: Information leakage from collaborative deep learning [OL]. 2017 [2017-09-01]. https://arxiv.org/pdf/1702.07464

[198] Dwork C, Roth A. The algorithmic foundations of differential privacy [J]. Foundations and Trends in Theoretical Computer Science, 2014, 9(3/4): 211-407

[199] Dwork C. Differential privacy: A survey of results [C] // Proc of the 5th Int Conf on Theory and Applications of Models of Computation. Berlin: Springer, 2008: 42-61

[200] Dwork C, McSherry F, Nissim K, et al. Calibrating noise to sensitivity in private data analysis [J]. Journal of Privacy and Confidentiality, 2016, 7(3): 265-284

[201] Zhu Tianqing, Li Gang, Zhou Wanlei, et al. Differentially private data publishing and analysis: A survey [J]. IEEE Trans on Knowledge and Data Engineering, 2017, 29(8): 1619-1638

[202] Xie Pengtao, Bilenko M, Finley T, et al. Crypto-nets: Neural networks over encrypted data [OL]. 2014 [2017-09-01]. https://arxiv.org/pdf/1412.6181

[203] Rivest R L, Adleman L, Dertouzos M L. On data banks and privacy homomorphisms [J]. Foundations of Secure Computation, 1978, 4(11): 169-180

[204] Ohrimenko O, Schuster F, Fournet C, et al. Oblivious multi-party machine learning on trusted processors [C] // Proc of the 25th USENIX Security Symp. Berkely, CA: USENIX Association, 2016: 619-636

[205] Shokri R, Shmatikov V. Privacy-preserving deep learning [C] //Proc of the 22nd ACM SIGSAC Conf on Computer and Communications Security. New York: ACM, 2015: 1310-1321

[206] Abadi M, Chu A, Goodfellow I, et al. Deep learning with differential privacy [C] //Proc of the 23rd ACM SIGSAC Conf on Computer and Communications Security. New York: ACM, 2016: 308-318

[207] Papernot N, Abadi M, Erlingsson ú, et al. Semi-supervised knowledge transfer for deep learning from private training data [OL]. 2016 [2017-08-02]. https://arxiv. org/pdf/1610.05755

[208] Hamm J, Cao P, Belkin M. Learning privately from multi-party data [C] //Proc of the 33rd Int Conf on Machine Learning. New York: ACM, 2016: 555-563

Zhang Yuqing, born in 1966. PhD, professor, PhD supervisor in the University of Chinese Academy of Sciences. His main research interests include computer networks and information security.

Dong Ying, born in 1991. PhD candidate in the University of Chinese Academy of Sciences. Her main research interests include Web security and machine learning.

Liu Caiyun, born in 1992. Master candidate in the University of Chinese Academy of Sciences. Her main research interests include data mining and information security.

Lei Kenan, born in 1992. Master candidate in Xidian University. Her main research interests include network and information security.

Sun Hongyu, born in 1992. Master candidate in Xidian University. His main research interests include network and information security.