深度学习发展现状及其在安全领域方向的应用

组员：林修远 3220190962

郁裕磊 3220190992

张 博 3220190995

张家赫 3220190997

摘 要

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。深度学习是传统机器学习的一种延伸，深度学习的概念源于人工神经网络的研究，含多个隐藏层的多层感知器就是一种深度学习结构。深度学习通过组合低层特征形成更加抽象的高层表示属性类别或特征，以发现数据的分布式特征表示。研究深度学习的动机在于建立模拟人脑进行分析学习的神经网络，它模仿人脑的机制来解释数据，例如图像，声音和文本等。深度学习目前在自动文本翻译[1]、图像识别[2]、自动驾驶汽车[3]、智能城市[4]等许多领域发挥着重要作用。此外，各种框架，如numpy、tensorflow、pytorch现在都可用于深度学习应用的实现。目前，网络安全方向的研究人员也开始探索深度学习在网络安全领域中的应用，如入侵检测、恶意软件分类、android恶意软件检测、垃圾邮件和网络钓鱼检测以及二进制分析。本文概述了基于深度学习的针对各种网络安全方向的应用。

关键词：深度学习；二进制分析；恶意软件检测；入侵检测；

1. 引言

网络安全指保护关键数据和设备免受网络威胁。它保护着上至国家政治国防信息，下至企业与社交平台保存的用户个人数据的数据安全。随着网络上获取信息的成本日益下降，敏感数据受到的威胁也越来越大，网络犯罪的收益有时高达数十亿美元。网络安全是为保护计算机、网络、程序和数据免受攻击、未经授权的访问、更改或破坏而设计的一组技术和过程。这些系统由网络安全系统和主机安全系统组成，每一个系统都有一个防火墙、防病毒的计算机代码与相关的入侵检测系统。

近些年来，随着深度学习技术的蓬勃发展，网络安全的研究者们开始探索深度学习技术在二进制安全方向上的应用。二进制分析属于信息安全业界逆向工程中的一种技术，大多数情况下，研究人员无法获取软件程序的源代码，只能获取对应的二进制程序文件。通过利用可执行的机器代码（二进制）来分析应用程序的控制结构和运行方式，软件系统的数量和种类不断增多,而软件漏洞无法完全避免。软件漏洞的数量也逐年提升,及时检测出软件存在的缺陷以免被不法分子利用也变得越来越重要。因此,研究二进制程序的漏洞检测具有重要的实际意义。其有助于信息安全从业人员更好地分析各种漏洞、病毒以及恶意软件，从而找到相应的解决方案。二进制分析有许多重要的应用，如恶意软件检测和自动修复易受攻击的软件。

使用深度学习进行二进制分析，主要有以下几个原因：首先是效果足够好，在图像识别、语音识别、机器翻译等领域的效果都远远超过非深度学习的算法。第二是样本的性质合适，深度学习就是擅长处理单一类型的数据。对应我们的输入都是二进制文件。第三要求足够多的样本，而样本越多准确的也越高。涉及二进制安全的样本根据 Symantec 在2017年Q2的统计，单日样本可以达到 300万。第四，可以避免人工去选择特征，只要一开始就设计好网络结构，深度学习会自动学习到重要的特征。再对比静态分析，比如用 N-gram，特征数量会轻松突破百万，再乘以上面的样本数，机器学习是必然选择。总得来说，深度学习的优点（超多特征，无需人工特征，样本越多越好）都非常适合二进制病毒样本分类这个领域。本文总结了将深度学习技术应用于二进制网络安全的发展现状。

第二章 用深度学习进行二进制分析

2.1 二进制分析

在[5]提出应用人工神经网络来解决二进制分析中的关键问题。即函数识别的问题，这是许多二进制分析技术中至关重要的一步。过去的几年中机器学习在视觉对象识别、语言建模和语音识别等多个应用领域取得了突破性的成果，但是将这些技术应用于二进制分析中还是较少。文中利用已有的数据集，证明了递归神经网络比目前最先进的基于机器学习的方法更能准确有效地识别二进制函数，将训练模型的时长缩短了一个数量级。

在[6]中还附带了RNN网络，但在这项工作中，许多实验都是在机器代码片段上进行的。这些代码片段使用标记化方案在LLVM和MIPS二进制数据上执行。这有助于从低级二元结构中提取高级结构。

在[7]提出了一种利用递归神经网络从分解后的二进制码中恢复函数类型特征的新系统EKLAVYA，它可以通过自动学习指令、编译器约定、堆栈框架设置指令、写前使用模式以及直接从二进制文件中识别类型相关的操作之间的关系。文章使用clang和GCC编译的Linux二进制文件的评估中，对于两种不同的体系结构（x86和x64），EKLAVYA在函数参数计数和类型恢复任务上分别显示了大约84%和81%的准确性。并且系统高度集中化，不需要关于指令集、编译器或优化级别的专门知识。

在[8]中描述了一种名为Hecate的新兴模式，它利用动态执行和追踪构建一个定制的自我包含的程序，从而减少潜在的攻击面。它可以自动地根据二进制代码，通讯特征确定程序特征，裁剪和利用这些特征创造一个符合用户需求定制化的程序二进制程序，这个过程是一个完全无监督的过程。Hecate 的关键特征是它利用了深度学习来辨别程序和通讯相关的特征。利用数据集提取不同程序特征，应用这些从完整指令序列中提取动态行路径片段进行拼接，并将这其与程序中功能的二进制代码相对应，最终通过程序的组成功能辨别程序特征。Hecate 将程序功能的辨认看作一个多级分类问题，每一个功能看作是一个类别标签，二进制代码是分类的样本，从动态指令追踪得到的执行路径是训练样本。通过RNN进行语义级别的向量嵌入，训练一个多级分类的CNN网络确定特征构成函数Hecate 的原型包含两个部分，特征辨别和特征裁剪，通过记录每个指令的虚拟地址，通过内存排列出每个二进制模型，而后由这两部分信息就能将动态执行路径独一无二地映射到静态代码中。

在[9]中提出了一种模糊测试模型，NeuFuzz，这种模糊测试模型使用了深度神经网络在灰盒模糊测试中智能化知道种子文件的选择从而减少无用种子的时间和空间的开销。特别是利用深度神经网络学习大量有缺陷和没有缺陷的路径来学习程序中隐藏的脆弱函数的模式，通过这种这种模式对被测程序的路径进行分类，判断其是否属于脆弱函数。

在[10]中提出了SAFE(Self-Attentive Function Embeddings for Binary Similarity)，这是一种新型的基于self-attention 神经网络二进制函数embeddings架构。它直接在反汇编的程序中运行，并不需要手动提取特征，在计算上这比现在的方法更加有效率（因为不用提取指令，建立和调整控制流图，减少了计算的开销），并且这个工具可以运行在多个架构上。经过实验和分析，该工具的效率较现有的工具都有了一定的提升。

2.2 恶意软件检测

使用现有的深度学习方面的成就，来对恶意软件进行分析，是现在恶意软件检测的主流方向。使用的数据预处理方式是静态和动态结合的分析方法，同时使用了机器学习和深度学习方法，这是因为随着互联网的不断发展，不仅仅各种软件和APP在不断涌现，恶意软件的类型和个数也在不断的增加，所以我们必须改变以前手工检测恶意软件的方法，将恶意软件和深度学习方法结合起来，自动进行学习分类，并将其形成自动的系统，这样就会不断扩展其应用范围，为恶意软件检测方面的发展提供不断的助力。

在[11]中提出了一种用来通过控制流攻击来检测恶意文档的系统Barnum。利用该系统收集硬件执行痕迹信息的来构建深度学习模型并检测控制流异常。Barnum是一个端到端系统，主要由三个主要组件组成：跟踪收集；行为建模；通过二进制分类进行异常检测。这个系统利用intelPT 来进行低开销的跟踪执行，并对跟踪信息所确定的基本块序列应用深度学习模式来训练一个正常的程序行为模式。根据模型的预测精度，Barnum会确定一个决策边界来区分良性和恶意执行文档程序。

在[12]中使用基于控制流图的方式对物联网恶意软件进行检测，文章使用了CFG图来表示恶意样本特征，对于CFG图的特征，包括图的节点数，边数、度中心性、直径、半径、最短路径分布等多种特征，对这些特征进行学习分类，可以得到这两种文件之间比较显著的差别。对于这些特征的统计，需要得出其对应结果的折线图，并得出能够对两种恶意软件进行区分的特征，在这之后，使用机器学习和深度学习方法对所得结果进行分析，包括LR，SVM，RF，CNN等方法，最后CNN的训练结果相比较而言比较高。

恶意软件样本的动态分析是恶意软件检测的重要方法。在[13]提出了一种结合了机器学习和深度学习的恶意软件检测体系结构。本文采用递归神经网络模型提取抽象特征，并为了祛除冗余程序，研究了集中序列预处理方法，本文结合LSTM深度学习模型和随机森林模型，提出了一种恶意软件检测体系结构ASSCA，检测体系结构采用API序列特征和统计特征。文章使用Cuckoo sandbox对恶意软件进行沙箱信息收集，对API序列进行分析使用N-gram方法，并移除相同的连续的API模式，然后首先进行深度学习模型的训练，然后进行机器学习模型的选择，文章选择了随机森林算法来进行处理深度学习特征。

在[14]中为了在通用性和性能之间取得平衡，文章探索了使用新的机器学习技术来将恶意程序分类为控制流图（CFGs），构建了新的使用了深度图卷积神经网络（DGCNN）将CFGs固有的结构信息嵌入其中，并开发了一个名为MAGIC的新系统，这是一个端到端的恶意软件防御系统，实现了高效的恶意软件分类。其使用了超过20k的恶意软件样本进行训练模型。实验结果表明，可以对CFG表示的恶意软件程序进行分类，其性能可与手工制作的恶意软件特性上应用最先进的方法向媲美。文章最后提出，将Magic系统放到云端，可以实现上传文件，系统可以自动获取文件CFG图，并对控制流图进行分析，并得到分类结果，通过这种方式，可以提高magic对一些没有进行训练的类型进行分析，可以进行提高训练准确率。

2.3 入侵检测

通过侵检测系统能够检测环境中的各种网络攻击。IDS通过分析收集到的数据包、向计算机用户发出的警报来检测恶意网络活动，并阻止来自入侵方的恶意连接[15]。它还与防火墙连接，作为网络安全的基本技术。将网络的特性评估为基于主机的入侵检测（hids），hids包括对系统内部数据包的放置和监控。为了执行入侵检测，hids从其系统调用、os审计跟踪、应用程序日志等收集数据。基于网络的入侵检测检测网络流量中的恶意活动。入侵检测算法一般分为误用检测（基于签名）和异常检测两种方法。基于签名的入侵检测是一种查找网络中一系列恶意的字节或序列的技术。基于异常检测有助于识别异常和系统中的严重事件，并纠正网络中的异常流量模式。为了解决这两种检测方法的缺点，提出了一种将异常的复杂度和问题检测系统结合起来的新的框架。

在文[16]中，提出了一种新的网络入侵检测系统深度学习方法模型，将深度学习和浅层学习相结合，这有助于分析非对称深度自动编码器技术（NDAE）上的网络流量问题。在文献[17]、[18]、[19]中，简单的研究说明了LSTM、递归神经网络、卷积神经网络与其他机器学习算法相比在IDS系统中表现良好。并且简要介绍了n-gram技术的cnn，以及cnn、cnn-递归神经网络、CNN长短期记忆和CNN门控递归单元等混合网络，这些技术有助于识别网络连接中的恶意网络id。在文章[17]的提取过程中，CNN有能力从低层特征集获得高层特征表示。基于文章[18]、[20]，[17]在针对基于异常的入侵检测系统方向，提出了基于LSTM算法的系统调用建模方法。系统调用建模有助于捕获网络上每个调用和关系的语义。集成方法主要关注符合IDS设计的虚警率。该方法有助于在较小空间中存储参数，该方法被认为是在序贯矩阵应用中快速有效的方法之一。

第三章 目前存在问题

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问题主要分为３个层次，即算法安全性、算法功能和算法性能。

1.1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。而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。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。解决算法脆弱性问题，较好的一种办法是进行对抗训练。

1.2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。

1.3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。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。

第四章 未来研究方向

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，未来可能有如下几个发展方向。

1.1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，对抗性深度学习是近两年的热点研究领域，对抗攻击目标的本质是造成目标模型的错误分类，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。

1.2防御针对协作性模型的攻击

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，使模型产生严重偏移，因此为了使模型更准确，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。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，该难点也需克服。

1.3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。安全系统属于机器学习的下游，我们不能直接照搬机器学习的特征提取方式。因此，直接使用深度学习模型对网络空间安全数据进行分类并不是一个好的选择，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。

1.4可解释性

由深度学习算法带来高准确率同时，相应的代价是其判断结果有较差的可解释性，这也就是说，当深度学习模型判断出一个软件或者访问请求是恶意的时候，并不给出任何与判断结果相关的恶意数据段相关信息。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。

参考文献

[1] Y.Wu,M.Schuster,Z.Chen, Q.V.Le,M.Norouzi,W.Macherey,M.Krikun, Y. Cao,Q. Gao, K. Macherey, J. Klingner, A. Shah, M. Johnson, X. Liu, L. Kaiser, S. Gouws, Y. Kato,T. Kudo, H. Kazawa, K. Stevens, G. Kurian, N. Patil, W. Wang, C. Young, J. Smith, J. Riesa,A. Rudnick, O. Vinyals, G. Corrado, M. Hughes, and J. Dean, “Google’s neural machine translation system: Bridging the gap between human and machine translation,” CoRR, vol. abs/1609.08144, 2016.

[2] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” *Communications of the ACM*, vol. 60, no. 6, pp. 84–90, May 2017.

[3] M. Bojarski, D. D. Testa, D. Dworakowski, B. Firner, B. Flepp, P. Goyal, L. D. Jackel, M. Mon- fort, U. Muller, J. Zhang, X. Zhang, J. Zhao, and K. Zieba, “End to end learning for self-driving cars,” *CoRR*, vol. abs/1604.07316, 2016.

[4] M. Mohammadi, A. Al-Fuqaha, M. Guizani, and J. Oh, “Semisupervised deep reinforcement learning in support of iot and smart city services,” *IEEE Internet of Things Journal*, vol. 5, no. 2, pp. 624–635, 2018.

[5] E. C. R. Shin, D. Song, R. Moazzezi, Recognizing functions in binaries with neural networks., in: USENIX Security Symposium, 2015, pp. 611–626.

[6] D. S. Katz, J. Ruchti, E. Schulte, Using recurrent neural networks for decompilation, in: 2018 IEEE 25th International Conference on Software Analysis, Evolution and Reengineering (SANER), IEEE, 2018, pp. 346–356.

[7] Z. L. Chua, S. Shen, P. Saxena, Z. Liang, Neural nets can learn function type signatures from binaries, in: Proceedings of the 26th USENIX Conference on Security Symposium, Security, Vol. 17, 2017.

[8] Xue, Hongfa & Chen, Yurong & Venkataramani, Guru & Lan, Tian. (2019).

Hecate: Automated Customization of Program and Communication Features to

Reduce Attack Surfaces.

[9] Wang Y , Wu Z , Wei Q , et al. NeuFuzz: Efficient Fuzzing with Deep Neural Network[J]. IEEE Access, 2019:1-1.

[10] Massarelli L , Di Luna G A , Petroni F , et al. SAFE: Self-Attentive Function Embeddings for Binary Similarity[J]. 2018.

[11] Barnum: Detecting Document Malware via Control Flow Anomalies in Hardware Traces

[12] Alasmary H , Khormali A , Anwar A , et al. Analyzing, Comparing, and Detecting Emerging Malware: A Graph-based Approach[J]. 2019.

[13] [Lu XF](http://apps.webofknowledge.com/OutboundService.do?SID=8CnYP2DH4oHEewe3bdj&mode=rrcAuthorRecordService&action=go&product=WOS&daisIds=25606) , [Zhou X](http://apps.webofknowledge.com/OutboundService.do?SID=8CnYP2DH4oHEewe3bdj&mode=rrcAuthorRecordService&action=go&product=WOS&daisIds=17044022) , [Jiang FS](http://apps.webofknowledge.com/OutboundService.do?SID=8CnYP2DH4oHEewe3bdj&mode=rrcAuthorRecordService&action=go&product=WOS&daisIds=14120315) , [Yi SW](http://apps.webofknowledge.com/OutboundService.do?SID=8CnYP2DH4oHEewe3bdj&mode=rrcAuthorRecordService&action=go&product=WOS&daisIds=2430093) ,  [Sha, J](http://apps.webofknowledge.com/OutboundService.do?SID=8CnYP2DH4oHEewe3bdj&mode=rrcAuthorRecordService&action=go&product=WOS&daisIds=1250229) . ASSCA: API based Sequence and Statistics features Combined malware detection Architecture,2018 , pp. 248-256

[14] J. Yan, G. Yan and D. Jin, "Classifying Malware Represented as Control Flow Graphs using Deep Graph Convolutional Neural Network," 2019 49th Annual IEEE/IFIP International Conference on Dependable Systems and Networks (DSN), Portland, OR, USA, 2019, pp. 52-63.

[15] R. Vinayakumar, K. Soman, P. Poornachandran, Evaluating effectiveness of shallow and deep networks to intrusion detection system, in: Advances in Computing, Communications and Informatics (ICACCI), 2017 International Conference on, IEEE, 2017, pp. 1282–1289.

[16] N. Shone, T. N. Ngoc, V. D. Phai, Q. Shi, A deep learning approach to network intrusion detection, IEEE Transactions on Emerging Topics in Computational Intelligence 2 (1) (2018) 41–50.

[17] R. Vinayakumar, K. Soman, P. Poornachandran, Applying convolutional neural network for network intrusion detection, in: Advances in Computing, Communications and Informatics (ICACCI), 2017 International Conference on, IEEE, 2017, pp. 1222–1228.

[18] G. Kim, H. Yi, J. Lee, Y. Paek, S. Yoon, Lstm-based systemcall language modeling and robust ensemble method for designing host-based intrusion detection systems, arXiv preprint arXiv:1611.01726.

[19] R. Nix, J. Zhang, Classification of android apps and malware using deep neural networks, in: Neural Networks (IJCNN), 2017 International Joint Conference on, IEEE, 2017, pp. 1871–1878.

[20] R. C. Staudemeyer, C. W. Omlin, Evaluating performance of long short-term memory recurrent neural networks on intrusion detection data, in: Proceedings of the South African Institute for Computer Scientists and Information Technologists Conference, ACM, 2013, pp. 218–224.