数据驱动的网络安全——最近发展以及未来发展新视角

郭嘉

摘要：

数据驱动的网络安全是主动防御的重要手段。有许多综述论文从不同的角度总结了不同方向的发展历史，而本综述注重于大范围覆盖近年来数据驱动的网络安全出现的一些新技术和新问题，在总结了数据驱动的的网页指纹安全、数据驱动的网页内容安全、数据驱动的 URL 和域名安全、数据驱动的网络流安全、数据驱动的恶意软件安全、数据驱动的物联网设备安全近年的发展后，本文将给出未来可能能够使数据驱动的网络安全取得进一步发展的思考，希望能够通过新的视角打开研究者进一步研究的视野。

一、背景介绍

信息技术的广泛应用和网络空间的兴起发展和普及，极大程度改变了人类生产生活的方式，促进了社会文化的传播以及经济的发展与繁荣。然而，网络毕竟是人类设计的系统，百密一疏，存在着很多安全隐患。网络空间安全也成为世界各国关注的重点问题，全世界多个国家政府发布了网络空间安全的战略。2016 年，国家互联网信息办公室发布《国家网络空间安全战略》[1]，战略中指出网络空间安全是政治安全、经济安全、文化安全、社会安全的重要保障、网络空间的国际竞争方兴未艾、网络空间机遇挑战并存。

机器学习是当前广泛使用的数据分析技术，在图像识别、语音处理、自然语言处理、强化学习、推荐系统、药物分析、自动驾驶、交通流量分析、垃圾邮件过滤、量化交易、政策效果分析等方面都有非常重要的应用。在大数据时代、计

算机性能不断发展的时代，利用机器学习处理问题成为越来越常用的策略。而网络空间恰恰是数据不断产生和流动的场所，也是计算机之间交流信息的方式，因此利用机器学习解决网络空间安全问题是一个再自然不过的想法。

目前有许多综述分析了机器学习在网络空间安全方面的应用。[2]主要专注于入侵检测的综述，该文对常用的数据类型分成两类，一类是包级别的数据，一类是 NetFlow 数据，还总结了几个用于分析网络入侵检测的数据集。该文根据所用机器学习方法以及任务类别两个维度对基于机器学习的入侵检测进行了分类，并分析了不同算法的时间复杂度，总结了影响入侵检测系统的因素。该文总结的数据集注重于入侵检测，没有关注与其他网络空间问题有关的数据集.[3]分析了机器学习算法与深度学习算法在网络空间安全应用上的区别，并从不同算法的层面上分类总结了不同机器学习与深度学习的在网络空间安全分析中具体应用方法，并指出未来需要建立大规模、多种类覆盖、类别平衡的数据集，还指出需要解决在线学习的问题。而且该文的分类方法不够精细，仅仅从模型方面进行分类，然而特征的提取，任务的类别也是非常重要的方面。[4]总结了有关基于机器学习的恶意软件检测的论文，该文对不同论文的目标问题进行了细分，包括恶意软件检测、恶意软件相似性检测（演化检测、类别检测、相似检测、相异检测）、恶意软件类型检测。对特征工程所用方法进行了分类，包括静态检测方法、动态检测方法、混合方法。该文还对常用特征类型进行了详尽的分类：字节序列特征、API 和系统调用特征、操作码特征、软件对网络影响特征、对文件系统影响特征、对 CPU 寄存器的使用特征、文件自身特性、文件所含字符串。并将不同论文使用方法分成基于监督学习的方法、基于无监督学习的方法、基于半监督学习的方法。该文总结了恶意软件检测所面临的一些挑战，包括如何应对反检测技术、如

何构建好的数据集，该文建议新的数据集应该专门针对不同的任务进行收集、数据的分布应该与现实世界场景相似，即正常软件的比例应当远远多于恶意软件的数量、数据集应当动态地更新以保持与恶意软件的更新迭代保持同步，并且标注这些恶意软件是什么时候被首次发现的。

由于上述综述都是较早的综述，本文希望总结机器学习与深度学习近几年在网络空间安全上的应用，并且希望做到囊括面尽量广，每个方面选取近年来的一些文章，或往年具有代表性的论文，而非每个方面都进行历年所有工作的总结。本文使用的研究方法是，尽可能收集近几年 CCS、S&P、USENIX Security Symposium 以及 NDSS 四大安全领域会议中有关机器学习和深度学习在网络空间安全方面应用的论文，并补充以上综述中所总结的一些开创性论文。本文将所有相关论文根据所解决的安全问题进行分类，而不是根据所使用模型分类。

本文将从数据驱动的的网页指纹安全、数据驱动的网页内容安全、数据驱动的 URL 和域名安全、数据驱动的网络流安全、数据驱动的恶意软件安全、数据驱动的物联网设备安全六个方面介绍近几年的工作，所选取的材料只求覆盖面广、发布时间新。在二至七节回顾完近些年的这些进展后，作者将通过新模型、新视角给出基于数据驱动的网络安全未来发展的可能方向，希望能够打开该方向研究的视野。

二、数据驱动的网页指纹安全

基于网页指纹信息的攻击是指攻击者仅仅通过网络流量的信息就可以在加密信道上判断用户在浏览什么网页。近年来的一些新工作将单网页流量检测延伸到多流量，或利用相似的思想解决 MPEG-DASH 的视频流指纹。

[5]指出很多这种基于流量的攻击仅仅适用于用户仅访问一个网页，但这往往是不切实际的，因为用户常常是同时访问多个网页，当多个网页访问的流量重合在一起时，很多攻击方法都会失效。该文提出了一种基于 XGBoost 机器学习模型的分析方法，利用该模型进行下一网页访问初始时刻的判断。在基于 SSH 加密的信道上，他们达到了 92.58%的 TPR，在基于 Tor 加密的信道上，他们达到了 64.94%的 TPR。[6]指出 MPEG-DASH 的视频流传输标准会导致信息泄密，尽管视频流已经被加密，攻击者仍然可以根据流量中一些突然激增的时间点分布的规律判断用户在访问什么视频，因为这些突然激增的时间点分布规律会根据用户访问视频的不同而变化，该文发现在使用卷积神经网络的情况下，这些规律能被很有效地捕捉，从而导致了用户隐私安全问题。

三、数据驱动的网页内容安全

近年来出现了一些基于图模型的第三方广告屏蔽、基于神经网络的恶意网页内容的检测工作，图模型对于关系的建模、神经网络对于深层特征与联系的建模是他们成功的原因。

一些网站利用第三方广告进行用户信息的收集，这使用户的隐私受到了侵犯，因此自动屏蔽掉这些第三方广告非常重要。一种常用的屏蔽手段是用户把访问到的第三方广告进行反馈，利用用户的反馈信息建立一个大型的黑名单，利用这个黑名单进行第三方广告的过滤。另一些较为智能的方法是通过网站的 URL、代码结构、DOM 结构进行自动判别，而[7]提出一个基于图网络的机器学习模型，他们把各种网站利用一个大规模的图联系在一起，然后进行图表征学习，由于该方法结合了各种来源的信息，它对混淆攻击等具有一定的鲁棒性。他们在 Alexa 排

名前一万的网站上测试了他们的模型，准确率达到了 90.9%。

恶意的网页内容成为当前网络安全的一个重要问题。[8]利用深度学习的方法自动检测具有恶意内容的网页。不同于以往仅仅使用语意分析和对 HTML 和 Javascript 动态仿真来提取特征的方法。该文直接从静态的 HTML 中提取语言无关的符号流，从而避免了耗时的语义分析和仿真代码。不同于传统的忽略了局部特征的词袋模型，该文提出的神经网络能够有效捕获局部的信息。这种方法能够需要高频数据流处理的应用场景下取得效果，该方法的高准确率和低时延使得其能够被部署于终端、防火墙、网页代理。

四、数据驱动的 URL 和域名安全

利用 RNN，LSTM，CNN 等序列建模模型对随机生成域名、恶意 URL 的检测是近几年数据驱动的 URL 和域名安全的特点。

多种类型的恶意软件利用域名生成算法（DGA）生成大量伪随机域名以连接 C&C 服务器。为了阻挡 DGA C&C 流量，安全组织必须找到对恶意软件进行逆向工程的算法，然后利用这点根据随机种子生成一系列的域名，然后把这些域名预先放入黑名单，这种方法不仅非常枯燥，而且会被恶意软件制作者通过生成大量具有多变量重现性质的随机种子或使用动态种子生成的方法破解。[9]提出一种新的方法，他们的方法介入 DNS 查询以预测哪些域名是 DGA 生成的，该文利用 LSTM 预测 DGA 及他们相应的类别且不需要预先进行特征提取的工作。这种方法可以警示网络管理员网络中恶意软件的存在以及他们的类别。

恶意的 URL 检测是防御垃圾邮件、恶意广告、钓鱼网站等攻击的主要手段。传统的基于黑名单、正规表达、标志匹配的方法面对当前各种各样、日新月异的

恶意 URL 是非常低效的。尽管基于机器学习的方法可以一定程度上解决这个问题，然而机器学习需要繁重的特征工程，且这些特征工程方法需要基于日新月异的恶意 URL 做出一定程度的更新。[10]使用深度学习的方法，对 URL 中的单词和字母都进行了词嵌入，使用 CNN 卷积网络获取 URL 的特征。而[11]仅仅对 URL 中的每个字母进行信息嵌入，通过神经网络中的隐藏层获取 URL 的特征向量，通过非线性激活层推断 URL 为恶意的概率。[12]利用 CNN 卷积网络获取字母层次的特征，同时利用基于注意力机制的 RNN 网络提取单词层次的特征，然后将这些特征利用三层 CNN 网络获取了 URL 的有效特征，从而有效判别钓鱼网站的 URL。

五、数据驱动的网络流安全

近年来数据驱动的网络流安全较少使用深度学习，而注重特征工程，有的工作使用较为新颖的机器学习模型，如多示例学习。

[13]基于多示例学习算法对 HTTP 网络流进行了建模。该文对结构性的文件如 JSON、XML 或 ProtoBuffer 进行了建模，提取了其中的语意特征使其在机器学习模型中更具解释性，并将用于检测 HTTP 流量中被感染的计算机，该方法将经过特征工程提取的特征送入 CNN 卷积网络，该文能在面对新域名、新恶意软件的场景下有较好的表现。

分析加密网络流量中潜在的威胁是具有挑战性且重要的工作因为网络流量中可能存在恶意软件。对于加密的流量，特征比对的方法不再适用，[14]该文基于一些网络流量特征进行分析，这些特征包括 TLS 握手数据、与加密流量联系在一起的 DNS 背景流量、来自相同 IP 地址的 HTTP 背景流量的报文头部。为了

观察对于包含恶意信息的流量和良好流量之间的区别，该文首先分析了可以观察到的上百万条流量的 TLS、DNS、HTTP 信息，基于这些分析进行最具判别力的特征的选取。

六、数据驱动的恶意软件安全

6.1 数据驱动的恶意软件安全算法

近年来数据驱动的恶意软件安全算法有从模型上进行革新，如不必经过特征工程而使用深度学习直接处理二进制序列，也有从目的上进行革新，如在安装过程中就判断某软件是否为恶意软件。

[15]避免了特征工程，该文强调特征工程十分耗时耗力，且需要一定的专业知识进行特征的选择。该文做出大胆尝试，利用深度学习模型直接处理未经任何处理的二进制序列，以检测恶意软件。该文从字节的词向量嵌入开始分析了不同分辨下学习到的特征，该文分析了这些二进制反编译后获取的可理解的特征与网络获取的特征的关系，利用这个方法他们发现了深度学习模型学习到的一些与基于特征工程的传统机器学习学到的特征的关系。

[16]指出现在恶意软件的类别越来越多，数量也快速增长，设计一个能过够将全部类别的恶意软件一网打尽的模型或算法不切实际。政府和企业往往是通过对恶意软件进行分门别类，然后找出那些对他们影响最大的恶意软件类型，然而这样的工作十分耗时耗力，往往占据了安全研究人员 20%以上的工作时间。为了解决这个问题，他们提出 spotlight 模型进行恶意软件的分类。该算法首先从大型的恶意软件数据集中挑出那些已知的恶意软件，然后对于剩下的恶意软件进行聚类，然后根据他们对企业业务的影响进行了优先级排序。他们在 67M 个恶

意软件上测试了他们模型的表现，实验结果表明他们的模型能达到 99%的聚类纯净度。他们还将模型应用于现实场景广告欺诈恶意软件的检测，结果表明安全人员能够快速地将三个大型的用于广告欺诈的僵尸网络检测出来。

[17]指出在软件的安装过程中就可以分析其是否为恶意软件，并且该文指出在某些场景下软件的完整性和非恶意性必须在其安装过程中就了解清楚。作者提出用深度学习模型解决该问题，他们利用系统调用作为模型输入通过图 LSTM 模型作为编码器，多层神经网络作为解码器，该模型输出为某软件的安装过程是否表现得像恶意软件以及该安装表现出恶意得时间点。该模型在 625 个现实的恶意软件上达到了 96%的检测准确率，该模型可以运行在在多个平台和操作系统上并且对训练数据的污染以及对抗性攻击表现出了一定的鲁棒性。

6.2 数据驱动的恶意软件安全数据集

近年有两个新的 PE 数据集发布，其中一个在数量上占据优势，另一个在标注完整性上更胜一筹。

[18]发布了一个用于恶意 windows PE（portable executable 可移植可执行程序）检测的大型数据集。这个数据集包括了 1.1M 个从二进制文件提取的特征，其中 300K 个恶意、300K 个良好、300K 个未标记的用于训练，100K 个恶意，100K 个良好的用于测试。他们也开源了特征提取的代码。同时他们比较了用 LightGBM 训练的梯度提升决策树和不利用任何人工提取的特征作为输入深度学习模型 Malconv 的表现，发现还是前者表现更加出色，这也强调人工提取特征的重要性。

[19]发布了一个最新的可移植可执行程序数据集，该文对之前已有的一些相关数据集进行了一系列分析，发现这些数据集的一些恶意程序比较古老，同时这

些数据集没有对恶意程序的类别作出明确的区分，他们认为这阻碍了对于概念迁移、以及恶意软件更新迭代分析的进一步研究。为了解决这一问题，他们收集了一个大型数据集，其中包含了在 2019 年 8 月至 2020 年 9 月收集的 57293 个恶意软件、77142 个良好软件。他们通过一些分析说明了这个数据集为什么能促进对恶意软件概念迁移、更新迭代的进一步研究。

七、数据驱动的物联网设备安全

IoT 设备的安全是最近的一个新问题，他们的区别主要在于模型的选择，有的使用深度学习而有的使用传统机器学习，有的深度学习模型使用 RNN 作为深度特征提取器，有的深度学习模型使用 DAE 作为深度特征提取器。

越来越多的物联网（IoT）设备接入到互联网中，由于这些设备很多都有安全隐患，这直接导致互联网更容易受到攻击。一些僵尸网络如 Mirai 正是利用了 IoT 设备进行 DDoS 攻击，这使得自动检测网络中异常的 IoT 流量变得越来越重要。[20]设计了机器学习算法以解决这个问题。他们利用物联网的一些专有特性如：终端数量有限、包之间的时间间隔有一定的规律，这些特性有助于进行特征提取，他们的实验结果表明家庭路由器或其他网络中间件可以使用低时延的机器学习算法进行基于流量的、与协议无关的以 IoT 为发动源头的 DDoS 检测。[21]提出利用深度学习进行的以 IoT 为发动源头的 DDoS 检测，他们发现 BLSTM-RNN 即双向长短时记忆网络检测精度较好，但有一定的时延，同时他们还发布了一个相关的数据集。[22]使用自动编码器 DAE 作为深度学习模型进行 IoT 为发动源头的 DDoS 检测，他们利用 Mirari 和 BASHLITE 两类基于 IoT 的僵尸网络攻击感染了实验内 9 个商用 IoT 设备，利用此进行模型有效性的衡量。[23]不

采用深度学习模型，为了提高检测速度，他们使用单类支持向量机(OCSVM)，由于 OCSVM 依赖超参数的选取、核函数的选取、超参数的选择以及有效的特征选取，他们利用无监督的进化算法：灰狼优化算法（GWO）进行超参数的选取以及那些能过够很好描述 IoT 僵尸网络特征的自动选取。

八、未来展望

通过近几年数据驱动的网络空间安全的新工作，我们看到了许多新的建模方法、许多新的任务、新的数据采集方式、新的数据集。在看到这些可喜的进步之外，我们还要认识到目前存在的一些问题，下面将阐述数据驱动的网络空间安全未来仍需要解决的一些问题，以及作者认为的能够从其他机器学习模型上能借鉴的数据驱动的网络空间安全的新发展方向。

8.1 不同方法的统一比较

目前许多工作采用自己采集的数据集，这样给不同方法的横向比较带来了挑战，我们需要一个统一的数据集来进行有效的比较，根据此来判断不同方法的优劣。然而攻击方式、网络协议等随时间不断变化，要收集一个能与时俱进并保留历史版本的大型数据集是一个巨大的挑战，但如果能达成这样的目标，不同方法的比较就有了共同的土壤，也减轻了研究者的工作量。

8.2 连续深度学习

当前越来越多基于深度学习出现在基于数据的网络安全工作中，深度学习能减轻研究者进行繁杂特征工程的工作量，也能提取到一些手工提取难以注意到的特征。然而深度学习有其固有的弱点，即依赖大量的数据以及面对全新数据时难尽人意的泛化能力。连续学习是克服深度学习固有缺点的有效方法，连续不断的

网络流也为连续学习提供了土壤。如何设计有效的连续学习算法，使得深度网络能够与新的攻击方式、新的协议、新的防御任务保持同步更新是研究者面临的下一个挑战。

8.3 数据增强

数据增强是计算机视觉中提高模型泛化能力的方法，包括对图像进行旋转、裁剪、缩放等操作，这使模型能分析到更多未见过的情况。而数据增强在数据驱动的网络安全中仍未有深入的研究，通过更改流量中的部分头部数据、通过往网络流数据中增加人工噪声等或许是提高模型性能的一个方向。

8.4 多任务学习

目前数据驱动的安全模型往往注重于解决单一的任务，这不利于模型学习到根植于不同任务中的一些共同特征，也增加了模型过拟合的风险。采用多任务学习不仅能够使安全模型能同时对抗多种攻击，同时也能提高模型的特征提取能力和泛化能力。

8.5 迁移学习

迁移学习是计算机视觉中常用的学习技巧。在图像分类任务上训练的模型参数在下游任务如目标检测、图像分割等任务上进行微调能大大加快训练的速度。然而数据驱动的网络安全模型往往对于所有任务都是从头开始训练，忽略了各种安全任务内部的联系，忽视了不同任务所训练出的模型通过迁移加快训练的可能性，通过迁移学习提高模型性能或许是加快训练速度，发现不同任务联系的新方式。

8.6 强化学习

强化学习在许多任务上取得了令人难以想象的进步，模型通过与环境进行互

动能玩 MOBA 游戏、下围棋。这不禁让人对利用强化学习使安全防御模型不断适应新型攻击有了希望，防御模型不断与时刻潜藏威胁的网络流数据进行互动，如果模型没能认出这些威胁或过滤掉了无威胁的信息就给予惩罚，否则给予奖励，这种模型或许是未来新型网络防御模型的发展新方向。

九、总结

本文通过总结了近几年数据驱动的的网页指纹安全、数据驱动的网页内容安全、数据驱动的 URL 和域名安全、数据驱动的网络流安全、数据驱动的恶意软件安全、数据驱动的物联网设备安全的发展，并对利用新的机器学习范式改进目前方法给出了一些建议。希望能使研究者对近几年相关工作有全面的了解，并提出使研究者用新视角看待基于数据驱动的安全的方法。

十、参考文献

[1] 《国家网络空间安全战略》 http://www.cac.gov.cn/2016-12/27/c_1120195926.htm

[2] Buczak A L, Guven E. A survey of data mining and machine learning methods for cyber security intrusion detection[J]. IEEE Communications surveys & tutorials, 2015, 18(2): 1153-1176.

[3] Xin Y, Kong L, Liu Z, et al. Machine learning and deep learning methods for cybersecurity[J]. Ieee access, 2018, 6: 35365-35381.

[4] Ucci D, Aniello L, Baldoni R. Survey of machine learning techniques for malware analysis[J]. Computers & Security, 2019, 81: 123-147.

[5] Xu Y, Wang T, Li Q, et al. A multi-tab website fingerprinting attack[C]//Proceedings of the 34th Annual Computer Security Applications Conference. 2018: 327-341.

[6] Schuster R, Shmatikov V, Tromer E. Beauty and the burst: Remote identification of encrypted video streams[C]//26th {USENIX} Security Symposium ({USENIX} Security 17). 2017: 1357-1374.

[7] Kargaran A H, Akhondzadeh M S, Heidarpour M R, et al. Wide-AdGraph: Detecting Ad Trackers with a Wide Dependency Chain Graph[C]//13th ACM Web Science Conference 2021. 2021: 253-261.

[8] Saxe J, Harang R, Wild C, et al. A deep learning approach to fast, format-agnostic detection of malicious web content[C]//2018 IEEE Security and Privacy Workshops (SPW). IEEE, 2018: 8-14.

[9] Woodbridge J, Anderson H S, Ahuja A, et al. Predicting domain generation algorithms with long short-term memory networks[J]. arXiv preprint arXiv:1611.00791, 2016.

[10] Le H, Pham Q, Sahoo D, et al. URLNet: Learning a URL representation with deep learning for malicious URL detection[J]. arXiv preprint arXiv:1802.03162, 2018.

[11] KP S, Alazab M. Malicious URL Detection using Deep Learning[J]. 2020.

[12] Huang Y, Yang Q, Qin J, et al. Phishing URL detection via CNN and attention-based hierarchical RNN[C]//2019 18th IEEE International

Conference On Trust, Security And Privacy In Computing And Communications/13th IEEE International Conference On Big Data Science And Engineering (TrustCom/BigDataSE). IEEE, 2019: 112-119.

[13] Pevny T, Dedic M. Nested Multiple Instance Learning in Modelling of HTTP network traffic[J]. arXiv preprint arXiv:2002.04059, 2020.

[14] Anderson B, McGrew D. Identifying encrypted malware traffic with contextual flow data[C]//Proceedings of the 2016 ACM workshop on artificial intelligence and security. 2016: 35-46.

[15] Coull S E, Gardner C. Activation analysis of a byte-based deep neural network for malware classification[C]//2019 IEEE Security and Privacy Workshops (SPW). IEEE, 2019: 21-27.

[16] Kaczmarczyck F, Grill B, Invernizzi L, et al. Spotlight: Malware Lead Generation at Scale[C]//Annual Computer Security Applications Conference. 2020: 17-27.

[17] Han X, Yu X, Pasquier T, et al. {SIGL}: Securing Software Installations Through Deep Graph Learning[C]//30th {USENIX} Security Symposium ({USENIX} Security 21). 2021.

[18] Anderson H S, Roth P. Ember: an open dataset for training static pe malware machine learning models[J]. arXiv preprint arXiv:1804.04637, 2018.

[19] Yang L, Ciptadi A, Laziuk I, et al. BODMAS: An Open Dataset for Learning based Temporal Analysis of PE Malware[C]//Proceedings of

Deep Learning and Security Workshop (DLS), in conjunction with IEEE Symposium on Security and Privacy (IEEE SP). 2021.

[20] Doshi R, Apthorpe N, Feamster N. Machine learning ddos detection for consumer internet of things devices[C]//2018 IEEE Security and Privacy Workshops (SPW). IEEE, 2018: 29-35.

[21] McDermott C D, Majdani F, Petrovski A V. Botnet detection in the internet of things using deep learning approaches[C]//2018 international joint conference on neural networks (IJCNN). IEEE, 2018: 1-8.

[22] Meidan Y, Bohadana M, Mathov Y, et al. N-baiot—network-based detection of iot botnet attacks using deep autoencoders[J]. IEEE Pervasive Computing, 2018, 17(3): 12-22.

[23] Al Shorman A, Faris H, Aljarah I. Unsupervised intelligent system based on one class support vector machine and Grey Wolf optimization for IoT botnet detection[J]. Journal of Ambient Intelligence and Humanized Computing, 2020, 11(7): 2809-2825.