基于注意力机制的序列模型在网络流量分类及其保护上的应用

(申请清华大学工学博士学位论文)

培 养 单 位 ： 计算机科学与技术系

研 究 生 ： 肖 文 韬

指 导 教 师 ： 导师姓名

二〇二一年十一月

Title

Thesis Submitted to

Tsinghua University

in partial fulfillment of the requirement

for the degree of

Doctor of Philosophy

in

by

Name of author

Thesis Supervisor: Name of supervisor

November, 2021

摘　要

不需要。

关键词：流量分类；注意力机制；序列模型；GAN

Abstract

No need.

Key Words: Recommendation System; click-through rate; deep learning; xDeepFM; AutoInt; FGCNN

目 录

插图和附表清单

第 1 章　研究背景和意义

1.1　研究背景

近十年来我国的互联网行业一直走在世界的前列，各式各样的信息科技创新模式和技术在我国互联网行业内生根发芽，且正以不可思议的速度发展。跟随互联网的发展和大数据科技的快速崛起，国内外的应用程序和网站呈现爆发式的增长。截至 2021 年 6 月[1]，我国三家基础电信企业的移动电话用户总数达 16.14 亿户，较 2020 年 12 月净增 1985 万户。其中，5G 手机终端用户达 3.65 亿户，较 2020 年 12 月增加 1.66 亿户。截至 2021 年 6 月，我国网民规模达 10.11 亿，其中手机网民规模达 10.07 亿，较 2020 年 12 月增长 2092 万，网民使用手机上网的比例为 99.6%。互联网给人们提供了越来越多的个性化资讯，并在工作、娱乐，生活中服务于大众，受到了人们的欢迎。在这一庞大的用户体量下，各种应用程序和网站的需求呈现爆发式增长。截至 2021 年 6 月，我国国内市场上监测到的应用程序数量为 302 万款，而且每天都会有一大批新的应用程序出现。截至 2021 年 6 月，我国网站数量为 422 万个。

科技是把双刃剑，互联网给人们生活带来便利的同时，也受到了恶意攻击者的关注。如今不管是应用程序还是网站大多存在安全隐患。以 Android 应用程序为例，根据移动互联网系统与应用安全国家工程实验室和爱加密移动应用大数据平台提供的数据，截止 2021 年 9 月底大数据平台共计收录 Android 移动应用程序 347 万款，其中 70% 以上存在高危漏洞威胁[2]，这类应用程序容易遭受不同形式的攻击。据调查，几乎所有（89%）的漏洞都可以被恶意软件利用[3]。正由于目前的应用程序漏洞总量特别庞大，这就有让黑客窃取个人密码、金融信息或隐私数据的可能。如多数应用程序存在滥用各类共享权限的情况，容易被不法分子利用，导致隐私的泄露。通过应用程序发起恶意攻击的攻势越来越猛烈。据统计，每年至少新增 150 万种恶意移动软件，至少造成超过 1600 万件的恶意移动软件攻击事件[4]。

2021 年 4 月，全球领先的网络安全解决方案提供商 Check Point® 软件技术有限公司发布了《2021 年移动安全报告》[5]。报告全面概述了移动恶意软件、设备漏洞及国家级网络攻击的趋势，并介绍了如何抵御当今和未来复杂的移动威胁。报告显示，2020 年，几乎每个组织都经历了至少一次移动恶意软件攻击。这些攻击中有 93% 源于互联网，它们试图诱骗用户通过受感染的网站或网址安装恶意应用，

或者窃取用户凭据。46% 的组织至少有一名员工下载了威胁网络和数据的恶意应用。由于芯片组的缺陷，全球至少有 40% 的移动设备存在固有的漏洞，需要紧急打补丁。目前全世界正处于新冠疫情的影响之下，这些恶意软件通常会隐藏在生成提供疫情相关信息的应用中。诸如此类的应用程序用户可能遭受不同形式的恶意攻击，包括流量分析攻击以及通过恶意应用程序攻击来窃取用户的个人隐私或者个人财产。

1.2　研究意义

第 2 章　国内外研究现状

2.1　网络流量分类

2.2　网络流量保护

第 3 章　研究目标和拟解决的关键问题

3.1　研究目标

整体的研究围绕网络流量安全展开, 总体分为两个方向:(1) 网络流量分类;(2) 网络流量保护。在本研究中，我们将探索基于注意力机制的序列模型在上述两个方向的应用。

3.2　拟解决的关键问题

- **包级别和流级别的高精度网络流量分类模型**：网络流量分析技术在网络监控与管理、用户行为分析等领域具有重要应用，区分流量所属的网站和应用是流量分析的首要关键步骤。如何从原始流量字节中自动学习特征表示，以及如何高精度地对各种网站和应用的流量进行分类是至关重要的问题。
- **基于深度学习的网络流量分类模型的可解释性**：虽然基于深度学习的方法在许多网络安全的研究中取得了可观的进展，但是基于深度学习的方法仍然饱受可解释性差的问题，这使得这类方法的实际部署存在很多的困难。为基于深度学习的模型提供解释，对于它们在安全敏感领域的使用至关重要。可解释性帮助用户了解基于深度学习的方法的的内部工作原理：深度学习模型是如何对给定的输入作出具体决定的？可解释性能够通过让人类（用户）参与决策过程来提供一种安全感。因为如果无法解释这类方法的运行原理，方法的使用者无法评估和控制这些方法的误差上界和行为。
- **针对基于网络流量分类的攻击的网络流量保护算法**：通过网络流量分类，攻击者可以监听用户的流量窥探出用户访问的应用和网站。这导致用户隐私的泄漏。即使用户使用虚拟专用网络（VPN）和洋葱网络（Tor）等手段也无法避免被攻击。这促使学术界开始研究针对基于网络流量分类的攻击的网络流量保护算法，而已有的方法在存在着无法实际部署、资源消耗过大、保护效果差等问题。
- **高吞吐量**：因为在实际应用中，我们需要将网络流量分类和保护的算法部署到真实设备中，并服务于大量的用户流量。这要求我们提出的模型达到满足实时处理的模型执行速度，即高吞吐量。

3.3　预期创新点

针对上述研究目标和关键问题，我们提出的下一代智能网络流量分类及保护技术在实际应用方面和理论研究方面都有创新，预期在以下几个方面达到创新：

- 高精度、高吞吐量、高可解释性的网络流量分类方法。
- 针对基于网络流量分类的攻击的网络流量保护算法。
- 基于注意力机制的序列模型在网络流量分类及其保护上的理论研究。
- 在真实数据集上的充分实验证明我们提出的方法的有效性。

第 4 章　研究内容和方法

4.1　研究框架和研究内容

4.2　网络流量分类

4.3　网络流量保护

第 5 章　现有工作基础

- **课程情况**：已完成培养计划里除专业实践、文献综述与选题报告外的所有课程。
- 已完成文献阅读与调研。
- **网络流量分类**：
  - 实现提出的方法的代码。
  - 收集网站识别任务的数据集，处理应用识别任务的数据集。
  - 在两个数据集上进行充分的实验，对超参数和模型架构进行实验。
- **网络流量保护**：
  - 实验提出的方法的大部分代码。
  - 初步实验结果。
- **已接收论文成果**：
  - （清华 A 类期刊学生一作）X. Xiao, W. Xiao, R. Li, X. Luo, H. -T. Zheng and S. -T. Xia, ”EBSNN: Extended Byte Segment Neural Network for Network Traffic Classification,” in IEEE Transactions on Dependable and Secure Computing, doi: 10.1109/TDSC.2021.3101311.

第 6 章　预期研究成果

- 完成系统设计及实验测试。
- 总结创新点及实验结果，预期高水平安全方向或网络方向论文一篇，专利一篇。
- 撰写硕士学位论文。

第 7 章　研究工作计划

时间	任务安排
2021.04 ~ 2021.09	文献调研
2021.09 ~ 2022.03	复现相关文献，模型算法的初步设计
2022.03 ~ 2023.01	改进模型算法，撰写学术论文，学术论文投稿
2023.01 ~ 2023.05	总结工作成果，硕士学位论文撰写、修改、定稿
2023.05 ~ 2023.06	硕士论文答辩

参考文献

[1] CNNIC 中国互联网络信息中心. 第 48 次中国互联网络发展状况统计报告 [Z]. 2021.

[2] 北京智游网安科技有限公司（爱加密），移动互联网系统与应用安全国家工程实验室. 全国移动 App 风险监测评估报告》（2021 年 3 季度版）[Z]. 2021.

[3] 中关村在线. 调查显示：大量移动 App 易受恶意软件攻击 [EB/OL]. 2019. https://safe.zol.com.cn/720/7200414.html.

[4] 爱加密. 全国移动 App 安全性研究报告：约 702019. https://www.freebuf.com/articles/paper/202843.html.

[5] 网络安全那点事. 《2021 年移动安全报告》恶意软件攻击遍布全球 [EB/OL]. 2021. https://baijiahao.baidu.com/s?id=1697081313109424120&wfr=spider&for=pc.