同行专家业内评价意见书编号：20240854228

附件1

浙江工程师学院（浙江大学工程师学院）
同行专家业内评价意见书

姓名： 吴至禹

学号： 22160277

申报工程师职称专业类别（领域）： 电子信息

浙江工程师学院（浙江大学工程师学院）制

2024年03月27日

一、个人申报

（一）基本情况【围绕《浙江工程师学院（浙江大学工程师学院）工程类专业学位研究生工程师职称评审参考指标》，结合该专业类别(领域)工程师职称评审相关标准，举例说明】

1.对本专业基础理论知识和专业技术知识掌握情况

作为一名计算机技术专业的学生，我的课程成绩为85分，专业实践的评分为86分，我掌握了计算机科学的核心概念和基本原理，包括计算机体系结构、操作系统、计算机网络、算法和数据结构等。我了解计算机科学的发展历程，从早期的机械计算机到现代计算机的各种形式和应用，以及计算机科学的未来发展方向。我还了解计算机科学的伦理和社会责任，以及计算机安全和隐私保护等方面的问题。

2.工程实践的经历

实践单位简介：杭州智语网络科技有限公司是一家专注于语音与数据人工智能产品研发的高科技企业。
实习实践内容：对恶意软件检测模型的安全性进行研究，包括恶意软件的对抗样本，数据投毒等
主要研究目标：展开面向基于深度学习的恶意软件检测的对抗攻击方法研究，通过启发式的优化算法不断修改恶意软件中的指令，使得被修改后的恶意软件规避检测，同时在现实世界的反病毒产品上对本方法进行评估，对成功规避检测器的样本进行分析，从而提升恶意软件检测器的准确率。
技术难点：在黑盒场景下构造能够规避恶意软件检测器检测的恶意软件对抗样本，且该样本的恶意性能不受到影响。
研究内容、方案及技术路线：

为了使得修改后的PE恶意软件依然能够执行恶意功能，首先定义一组无任何语义信息（Semantic-
Nop）的汇编指令。由于PE文件中存在未被使用的片段，将无任何语义的汇编指令插入到该片段中不会影响恶意软件的执行。此外，为了改变程序控制流图中的节点，可以将恶意软件中的部分代码转移到新添加的区段中。

对于基于深度学习模型恶意软件检测器，在白盒场景下，攻击者可以获得模型的内部参数以及模型对于给定样本的输出结果。在这种场景下，攻击者可以根据模型损失函数的梯度来对恶意软件的数字特征进行优化，直到生成的数字特征能够使得恶意软件检测器产生误判。最后，将生成的数字特征映射回对应的恶意软件样本即可。

在黑盒场景下，攻击者只能获得待测软件是否为恶意的结果。在现实部署的反病毒产品中，用户只能得到产品给出的结果，而无法接触到产品内部的技术细节。因此，对黑盒场景下对抗攻击方法的研究具有现实意义。本文将探索使用蒙特卡洛树搜索、遗传算法等启发式方法，在允许的时间范围内对恶意软件的修改操作进行搜索。同时，将对成功规避检测器的样本进行分析，从而提升恶意软件检测器的准确率。
团队分工：恶意软件的收集、恶意软件的预处理（包括分析恶意软件所属的家族、提取恶意软件的程序控制流图等）、基于深度学习的恶意软件检测器的训练和性能测试、商用杀毒软件的部署、攻击方法的实施与测试。
本人承担任务及完成情况：恶意软件的预处理（包括分析恶意软件所属的家族、提取恶意软

件的程序控制流图等）、基于深度学习的恶意软件检测器的训练和性能测试、商用杀毒软件的部署、攻击方法的实施与测试。目前均已完成。

3.在实际工作中综合运用所学知识解决复杂工程问题的案例（不少于1000字）

在专业实践中，我一直在努力提升我们的恶意软件检测系统的性能。在我加入公司后，我发现我们的系统在检测新兴和未知的恶意软件方面存在一定的问题。这是因为我们的系统主要依赖于特征码数据库，而这种方法对新兴和未知的恶意软件的检测能力有限。因此，我决定利用我在学习过程中掌握的知识，尝试改进我们的恶意软件检测系统。首先，我对公司的恶意软件检测系统进行了深入的分析，了解了其工作原理和存在的问题。我发现我们的系统主要使用基于特征码的静态检测方法，这种方法虽然在检测已知的恶意软件上效果很好，但对于新兴和未知的恶意软件，其检测能力有限。因此，我决定引入基于深度学习的恶意软件检测方法，以提升我们系统的检测能力。我首先从VirusTotal上收集了大量的恶意软件样本，然后使用这些样本来训练我们的深度学习模型。我选择了基于图神经网络的magic模型和malgraph模型，以及基于卷积神经网络的malconv模型。这些模型都已经在恶意软件检测方面取得了很好的效果。
在训练完深度学习模型后，我将这些模型集成到了我们的恶意软件检测系统中。通过测试，我发现使用深度学习模型后，我们的系统在检测新兴和未知的恶意软件上的能力大大提升。

同时，我也意识到，随着深度学习技术在恶意软件检测方面的应用，恶意软件的制作者可能会使用对抗攻击方法来规避我们的检测。因此，我决定研究和开发一种黑盒场景下面向恶意软件检测的攻击方法。然而，现有的对抗攻击方法面临着两大挑战。第一，大多数方法仅生成能规避模型检测的恶意软件数值特征，而不能保证恶意软件的功能完整性；第二，现有攻击方法无法在黑盒场景下高效搜索能够规避检测的对抗样本。为了解决上面的两个挑战，我选择了基于蒙特卡洛树搜索的启发式方法，这种方法能够在允许的时间范围内对恶意软件的修改操作进行搜索，从而找到能够规避检测的恶意软件样本。结果表明，在部署了7种不同的商用杀毒软件后，我们的对抗攻击方法能够在商用杀毒软件上取得超过70%的攻击成功率。此外，我还搭建了一个沙箱环境，将成功规避恶意软件检测器的样本输入沙箱系统，验证了这些样本是否还具有与修改前等价的恶意功能。结果表明，超过90%的恶意软件依然具有原来的恶意功能，能够在沙箱中完成与原来相同的恶意活动。

此外，我将这些样本进行分析，以找出它们是如何规避我们的检测器的，然后对我们的检测器进行改进，以提升其对攻击的防御能力。总的来说，通过运用我在学习过程中掌握的知识，我成功地改进了恶意软件检测系统，提升了其在检测新兴和未知的恶意软件方面的能力，同时也提高了其面对对抗攻击的防御能力。这个经验让我深刻地理解了理论知识在解决实际问题中的重要性，也提升了我的工程能力和抗挫能力。

（二）取得的业绩（代表作）【限填3项，须提交证明原件（包括发表的论文、出版的著作、专利证书、获奖证书、科技项目立项文件或合同、企业证明等）供核实，并提供复印件一份】					
1. 公开成果代表作【论文发表、专利成果、软件著作权、标准规范与行业工法制定、著作编写、科技成果获奖、学位论文等】					
成果名称	成果类别 [含论文、授权专利（含发明专利申请）、软件著作权、标准、工法、著作、获奖、学位论文等]	发表时间/授权或申请时间等	刊物名称/专利授权或申请号等	本人排名/总人数	备注
一种基于控制流图的恶意软件变换方法及系统	发明专利申请	2023年10月30日	申请号：CN202311436384.7	2/5	
一种基于蒙特卡洛树搜索的深度学习模型版权保护方法	发明专利申请	2023年02月14日	申请号：CN202310110901.5	2/4	
AdvParams: An Active DNN Intellectual Property Protection Technique via Adversarial Perturbation Based Parameter Encryption	TOP期刊	2022年12月27日	IEEE Transactions on Emerging Topics in Computing	2/5	

2.其他代表作【主持或参与的课题研究项目、科技成果应用转化推广、企业技术难题解决方案、自主研发设计的产品或样机、技术报告、设计图纸、软课题研究报告、可行性研究报告、规划设计方案、施工或调试报告、工程实验、技术培训教材、推动行业发展中发挥的作用及取得的经济社会效益等】

（三）在校期间课程、专业实践训练及学位论文相关情况	
课程成绩情况	按课程学分核算的平均成绩： 85 分
专业实践训练时间及考核情况(具有三年及以上工作经历的不作要求)	累计时间： 1 年（要求1年及以上） 考核成绩： 86 分（要求80分及以上）
本人承诺	
个人声明：本人上述所填资料均为真实有效，如有虚假，愿承担一切责任，特此声明！ **申报人签名：**	

二、日常表现考核评价及申报材料审核公示结果

日常表现考核评价	非定向生由德育导师考核评价、定向生由所在工作单位考核评价： ☑优秀 ☐良好 ☐合格 ☐不合格 德育导师/定向生所在工作单位分管领导签字（公章）：2020年3月17日
申报材料审核公示	根据评审条件，工程师学院已对申报人员进行材料审核（学位课程成绩、专业实践训练时间及考核、学位论文、代表作等情况），并将符合要求的申报材料在学院网站公示不少于5个工作日，具体公示结果如下： ☐通过 ☐不通过（具体原因：　　） 工程师学院教学管理办**公室**审核签字（公章）：　　年　月　日

浙 江 大 学 研 究 生 院

攻读硕士学位研究生成绩表

学号：22160277	姓名：吴至禹	性别：男	学院：工程师学院	专业：计算机技术	学制：2.5年
毕业时最低应获：24.0学分		已获得：24.0学分		入学年月：2021-09	毕业年月：2024-03
学位证书号：1033532024602231			毕业证书号：103351202402600457		授予学位：电子信息硕士

学习时间	课程名称	备注	学分	成绩	课程性质	学习时间	课程名称	备注	学分	成绩	课程性质
2021-2022学年秋季学期	知识图谱导论		2.0	88	专业选修课	2021-2022学年春夏季学期	研究生英语		2.0	81	公共学位课
2021-2022学年秋冬季学期	中国特色社会主义理论与实践研究		2.0	84	公共学位课	2021-2022学年夏季学期	物联网信息安全技术与应用基础		2.0	87	专业学位课
2021-2022学年秋冬季学期	研究生论文写作指导		1.0	92	专业学位课	2021-2022学年夏季学期	自然辩证法概论		1.0	75	公共学位课
2021-2022学年秋冬季学期	电子与信息工程技术管理		2.0	90	专业学位课	2021-2022学年夏季学期	工程伦理		2.0	88	公共学位课
2021-2022学年冬季学期	物联网操作系统与边缘计算		2.0	89	专业选修课	2021-2022学年夏季学期	大数据与人工智能工程应用		2.0	88	专业学位课
2021-2022学年春季学期	数学建模		2.0	95	专业选修课	2021-2022学年夏季学期	移动互联网智能设备应用设计与实践		3.0	83	专业学位课
2021-2022学年夏季学期	研究生英语基础技能		1.0	81	公共学位课						

说明：1. 研究生课程按三种方法计分：百分制，两级制（通过、不通过），五级制（优、良、中、及格、不及格）。

2. 备注中“*”表示重修课程。

学院成绩校核章：

成绩校核人：张梦依

打印日期：2024-04-02

国 家 知 识 产 权 局

310013

浙江省杭州市西湖区古墩路 701 号紫金广场 C 座 1506 室 杭州求是专利事务所有限公司

邱启旺(0571-87911726-808)

发文日：

2023 年 11 月 01 日

申请号：202311436384.7　　**发文序号：**2023110100937040

专 利 申 请 受 理 通 知 书

根据专利法第 28 条及其实施细则第 38 条、第 39 条的规定，申请人提出的专利申请已由国家知识产权局受理。现将确定的申请号、申请日等信息通知如下：

申请号：2023114363847

申请日：2023 年 10 月 30 日

申请人：浙江大学

发明人：吴春明,吴至禹,凌祥,唐馨,黄泸明

发明创造名称：一种基于控制流图的恶意软件变换方法及系统

经核实，国家知识产权局确认收到文件如下：

权利要求书 1 份 3 页,权利要求项数 ： 7 项

说明书 1 份 6 页

说明书附图 1 份 1 页

说明书摘要 1 份 1 页

专利代理委托书 1 份 2 页

发明专利请求书 1 份 5 页

实质审查请求书 文件份数：1 份

申请方案卷号：邱-231-322-吕

提示：

1.申请人收到专利申请受理通知书之后，认为其记载的内容与申请人所提交的相应内容不一致时，可以向国家知识产权局请求更正。

2.申请人收到专利申请受理通知书之后，再向国家知识产权局办理各种手续时，均应当准确、清晰地写明申请号。

国家知识产权局 专利申查业务章 1101081359734

审 查 员：赵燕

联系电话：010-62356655

审查部门：初审及流程管理部

200101 2022.10

纸件申请，回函请寄：100088 北京市海淀区蓟门桥西土城路 6 号 国家知识产权局专利局受理处收

电子申请，应当通过专利业务办理系统以电子文件形式提交相关文件。除另有规定外，以纸件等其他形式提交的文件视为未提交。

国 家 知 识 产 权 局

310013

浙江省杭州市西湖区古墩路 701 号紫金广场 C 座 1506 室 杭州求是专利事务所有限公司

邱启旺(0571-87911326-808)

发文日：

2023 年 02 月 14 日

申请号：202310110901.5　　**发文序号：**2023021401221470

专 利 申 请 受 理 通 知 书

根据专利法第 28 条及其实施细则第 38 条、第 39 条的规定，申请人提出的专利申请已由国家知识产权局受理。现将确定的申请号、申请日等信息通知如下：

申请号：2023101109015

申请日：2023 年 02 月 14 日

申请人：浙江大学

发明人：吴春明,吴至禹,黄泸明,唐馨

发明创造名称：一种基于蒙特卡洛树搜索的深度学习模型版权保护方法

经核实，国家知识产权局确认收到文件如下：

权利要求书 1 份 3 页,权利要求项数 ： 6 项

说明书 1 份 6 页

说明书附图 1 份 1 页

说明书摘要 1 份 1 页

专利代理委托书 1 份 2 页

发明专利请求书 1 份 5 页

实质审查请求书 文件份数：1 份

申请方案卷号：邱-231-32-吕

提示：

1.申请人收到专利申请受理通知书之后，认为其记载的内容与申请人所提交的相应内容不一致时，可以向国家知识产权局请求更正。

2.申请人收到专利申请受理通知书之后，再向国家知识产权局办理各种手续时，均应当准确、清晰地写明申请号。

国家知识产权局 专利申查业务章 1101081359734

审 查 员：自动受理

联系电话：010-62356655

审查部门：初审及流程管理部

纸件申请，回函请寄：100088 北京市海淀区蓟门桥西土城路 6 号 国家知识产权局专利局受理处收

电子申请，应当通过专利业务办理系统以电子文件形式提交相关文件。除另有规定外，以纸件等其他形式提交的文件视为未提交。

IEEE TRANSACTIONS ON
EMERGING TOPICS
IN COMPUTING

Received 23 September 2021; revised 28 October 2022; accepted 19 December 2022.
Date of publication 27 December 2022; date of current version 6 September 2023.

Digital Object Identifier 10.1109/TETC.2022.3231012

AdvParams: An Active DNN Intellectual Property Protection Technique via Adversarial Perturbation Based Parameter Encryption

MINGFU XUE, (Senior Member, IEEE), ZHIYU WU, YUSHU ZHANG, (Member, IEEE), JIAN WANG, AND WEIQIANG LIU, (Senior Member, IEEE)

Mingfu Xue, Yushu Zhang, and Jian Wang are with the College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing 211106, China
Zhiyu Wu is with the Polytechnic Institute, Zhejiang University, Hangzhou 310058, China
Weiqiang Liu is with the College of Electronic and Information Engineering, Nanjing University of Aeronautics and Astronautics, Nanjing 211106, China
CORRESPONDING AUTHOR: MINGFU XUE (mingfu.xue@nuaa.edu.cn)

This work was supported by the National Natural Science Foundation of China under Grants 61602241 and 62022041, and in part by the CCF-NSFOCUS Kun-Peng Scientific Research Fund under Grant CCF-NSFOCUS 2021012.

ABSTRACT The construction of Deep Neural Network (DNN) models requires high cost, thus a well-trained DNN model can be considered as intellectual property (IP) of the model owner. To date, many DNN IP protection methods have been proposed, but most of them are watermarking based verification methods where model owners can only verify their ownership passively after the copyright of DNN models has been infringed. In this article, we propose an effective framework to actively protect the DNN IP from infringement. Specifically, we encrypt a small number of model's parameters by perturbing them with well-crafted adversarial perturbations. With the encrypted parameters, the accuracy of the DNN model drops significantly, which can prevent malicious infringers from using the model. After the encryption, the positions of encrypted parameters and the values of the added adversarial perturbations form a secret key. Authorized user can use the secret key to decrypt the model on Machine Learning as a Service, while unauthorized user cannot use the model. Compared with the existing DNN watermarking methods which passively verify the ownership after the infringement occurs, the proposed method can prevent infringement in advance. Moreover, compared with few existing active DNN IP protection methods, the proposed method does not require additional training process of the model, thus introduces low computational overhead. Experimental results show that, after the encryption, the test accuracy of the model drops by 80.65%, 81.16%, and 87.91% on Fashion-MNIST (DenseNet), CIFAR-10 (ResNet), and GTSRB (AlexNet) datasets, respectively. Moreover, the proposed method only needs to encrypt an extremely low number of parameters. The proportion of the encrypted parameters in all the model's parameters is as low as 0.000205%. Experimental results also indicate that, the proposed method is robust against model fine-tuning attack, model pruning attack, and the adaptive attack where attackers know the detailed steps of the proposed method and all the parameters of the encrypted model.

INDEX TERMS Artificial intelligence security, deep neural networks, intellectual property protection, active authorization control, adversarial perturbation

I. INTRODUCTION

Deep Neural Networks (DNNs) are extensively deployed in commercial applications. Many companies use their trained high-performance DNN models to provide services for the public, which is called Machine Learning as a Service (MLaaS) [1]. The users can only obtain the model's predictions, but have no access to the internal parameters. Training a DNN model with high accuracy is a costly and time-consuming task. However, some malicious infringers may illegally duplicate or abuse the well-trained model, and obtain benefits from it, which greatly infringes the intellectual property (IP) of the model owner. IP protection for DNN models is an emerging problem.

AdvParams: An Active DNN Intellectual Property Protection Technique via Adversarial Perturbation Based Parameter Encryption

Publisher: IEEE

Mingfu Xue ; Zhiyu Wu ; Yushu Zhang ; Jian Wang ; Weiqiang Liu All Authors

Abstract:

The construction of Deep Neural Network (DNN) models requires high cost, thus a well-trained DNN model can be considered as intellectual property (IP) of the model owner. To date, many DNN IP protection methods have been proposed, but most of them are watermarking based verification methods where model owners can only verify their ownership passively after the copyright of DNN models has been infringed. In this article, we propose an effective framework to actively protect the DNN IP from infringement. Specifically, we encrypt a small number of model's parameters by perturbing them with well-crafted adversarial perturbations. With the encrypted parameters, the accuracy of the DNN model drops significantly, which can prevent malicious infringers from using the model. After the encryption, the positions of encrypted parameters and the values of the added adversarial perturbations form a secret key. Authorized user can use the secret key to decrypt the model on Machine Learning as a Service, while unauthorized user cannot use the model. Compared with the existing DNN watermarking methods which passively verify the ownership after the infringement occurs, the proposed method can prevent infringement in advance. Moreover, compared with few existing active DNN IP protection methods, the proposed method does not require additional training process of the model, thus introduces low computational overhead. Experimental results show that, after the encryption, the test accuracy of the model drops by 80.65%, 81.16%, and 87.91% on Fashion-MNIST (DenseNet), CIFAR-10 (ResNet), and GTSRB (AlexNet) datasets, respectively. Moreover, the proposed method only needs to encrypt an extremely low number of parameters. The proportion of the encrypted parameters in all the model's parameters is as low as 0.000205%. Experimental results also indicate that, the proposed method is robust against model fine-tuning attack, model pruning attack, and the adaptive attack where attackers know the de...