大学生创新训练项目计划申请书

|  |  |  |  |
| --- | --- | --- | --- |
| 项目编号 |  | | |
| 项目名称 | 基于深度学习的源代码缺陷定位研究 | | |
| 项目负责人 | 吴清晏 | 联系电话 | 19822606502 |
| 所在学院 | 吴健雄学院 | | |
| 学号 | 61522314 | 专业班级 | 计算机科学与技术(吴健雄学院) |
| 指导教师 | 李必信 | | |
| 申请日期 | 2023年10月23日 | | |
| 起止年月 | 2023年10月至2024年10月 | | |

**填写说明**

1. 本申请书所列各项内容均须实事求是，认真填写，表达明确严谨，简明扼要。
2. 申请人可以是个人，也可为创新团队，首页只填负责人。“项目编号”一栏不填。
3. 本申请书为大16开本（A4），左侧装订成册。可网上下载、自行复印或加页，但格式、内容、大小均须与原件一致。
4. 负责人所在学院认真审核, 经初评和答辩，签署意见后，将申请书（一式两份）报送项目管理办公室。

#### 一、 基本情况

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 项目名称 | 基于深度学习的源代码缺陷定位研究 | | | | | | |
| 项目级别 |  | | | | | | |
| 项目类型 | 创新训练项目 | | | | | | |
| 所属学科 | 学科一级门： 工学   学科二级类： 计算机类 | | | | | | |
| 申请金额 | 5000.00 元 | 起止年月 | | | | 2023年10月至2024年10月 | |
| 负责人 | 吴清晏 | 性别 | 男 | 民族 |  | 出生年月 | 年 月 |
| 学号 | 61522314 | 联系电话 | 宅： 手机： 19822606502 | | | | |
| 指导教师 | 李必信 | 联系电话 | 宅： 手机： 18602506179 | | | | |
| 项目简介 | | 开发者经常需要在软件质量的维护过程中定位缺陷代码，但手工定位缺陷困难且耗时，传统机器学习方法需要标注完善的数据集或耗时耗力的特征工程，且难以挖掘深层特征。而深度学习是一种强大的机器学习技术，它可以从无标记数据中学习高层次的特征，提高模型的预测能力。  本项目旨在探索基于深度学习的源代码缺陷定位方法，我们计划爬取GitHub上开源项目的issues信息，获取自然语言错误报告及修改前后代码并构建训练数据集。同时，我们将搭建深度学习模型，寻找合适的创新点。然后，我们将使用Accuracy@k等标准评估准确度，并在真实项目中进行检测。最后分析输出结果，通过transformer模型的自注意力机制实现一定程度上的可解释性。 | | | | | |
| 负责人曾经参与科研的情况 | | 在大一至大二的暑期学校中，开发了一款基于微信小程序的校园笔记共享平台。在开发过程中使用了Git版本管理工具和云数据库，对相关控制语句比较熟悉。 | | | | | |
| 指导教师承担科研课题情况 | | **软件测试和缺陷检测**：研究各种类型的软件缺陷检测和定位技术，保障开发出少缺陷、高质量、高可靠性、高安全性和可信的软件产品； | | | | | |
| 指导教师对本项目的支持情况 | | 选题指导，并在专利申请和论文发表方面可以给予资助等， Key Research and Development Program of Jiangsu Province: BE2021002-3 | | | | | |
| 项目组主要成员 | 姓名 | 学号 | 专业班级 | 所在学院 | 项目中的分工 | | |
| 吴清晏 | 61522314 | 计算机科学与技术(吴健雄学院) | 吴健雄学院 | 训练数据集建立中的爬虫编写和数据清洗，撰写项目任务书 | | |
| 张闻启 | 61522111 | 计算机科学与技术(吴健雄学院) | 吴健雄学院 | 负责模型评估和结果解释性分析，同时负责文档撰写中的数据集介绍、评估指标和对比方法、结果展示和解释等部分。 | | |
| 杨锦波 | 09022301 | 计算机科学与技术 | 计算机科学与工程学院 | 协助完成数据集的建立，参与模型的预训练和模型的优化，撰写创新点介绍和改进效果。 | | |
| 董子翔 | 71122207 | 软件工程 | 软件学院 | 负责预训练模型和模型构建。  同时负责划分数据集，及文档撰写中的模型介绍部分。 | | |

#### 二、 立项依据（可加页）

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| （1）研究目的 探索基于深度学习的源代码缺陷定位方法，借鉴和改进transformer模型，并结合传统缺陷定位方法和机器学习方法，如结合transformer模型与对抗学习技术；分析和评估模型在预测和定位缺陷方面的效果，并与其他现有方法进行对比分析；通过注意力机制实现一定程度的可解释性。 研究内容  1. 利用ScraPy框架构建Python爬虫，从GitHub选取开源项目的issues信息，提取错误描述及修改前后代码，构建训练数据集； 2. 学习transformer并将其与软件缺陷定位的传统方法相结合，在此基础上结合传统机器学习方法如对抗神经网络，实现更强大的无监督模型。 3. 评估模型结果并解释模型输出  （3）国、内外研究现状和发展动态 1. 数据集构建  国外利用开源项目的错误报告和修改前后代码，构建了一些规模大、质量高、覆盖多种编程语言和领域的软件缺陷数据集，如[Codeflaws](http://jos.org.cn/html/2023/2/6696.htm" \t "_blank)、[Big-Vul](http://jst.tsinghuajournals.com/article/2021/4302/20211104.htm" \t "_blank)、[DeepBugs](http://www.jos.org.cn/jos/article/abstract/6696" \t "_blank)等。  国内参考国外的数据集构建方法，利用开源项目或商业项目的错误报告和修改前后代码，构建了一些适合国内软件开发环境和需求的软件缺陷数据集，如[CodeDefectAI](https://www.chinaai.org.cn/newsinfo/5563202.html)、[VulDeePecker](http://www.jos.org.cn/jos/article/pdf/6696" \t "_blank)等 。  2. 模型设计  国外借鉴和改进了图像处理、自然语言处理等领域的深度学习模型，设计了一些适用于源代码缺陷检测的神经网络模型，如基于卷积神经网络 (CNN)、循环神经网络 (RNN)、长短时记忆网络 (LSTM)、注意力机制 (Attention)、图神经网络 (GNN)等  国内设计了一些针对特定类型或领域的源代码缺陷检测模型，如基于对抗生成网络 (GAN)、变分自编码器 (VAE)、强化学习 (RL)等 。 （4）创新点与项目特色  * 此前不存在使用transformer结构的源代码缺陷定位项目。 * 此前采用深度学习的源代码缺陷检测项目往往在可解释性上表现不佳。 * 此前的研究没有很好地结合传统缺陷定位方法与深度学习方法。  （5）技术路线、拟解决的问题及预期成果 技术路线：   * 数据获取：   + 利用scrapy框架搭建python爬虫从github获取开源项目的错误报告。   + 清洗数据，区分错误报告与修改前后代码，建立训练数据集。 * 模型建立：   + 使用所有数据进行模型预训练   + 分别构建模型处理自然语言部分和错误代码部分。   + 引入transformer模型处理代码上下文信息。   + 使用对抗神经网络训练获取最终可疑度排名。 * 模型评估：   + 选取真实项目进行实战检测，计算准确度并与同类型项目比较。   + 尝试分析输出结果，实现一定程度上的可解释性。   拟解决的问题：   * 如何结合传统缺陷定位方法与深度学习方法。 * 如何清洗数据，保留代码的语法特征。 * 如何提供足够深度学习模型预训练的庞大数据集。 * 如何利用更合适的指标和方法，评估模型在预测和定位缺陷方面的效果。 * 如何利用更有效的技术和手段，分析模型给出的缺陷定位结果和原因，实现更高程度上的可解释性。   预期成果：   * 一个覆盖多种编程语言和领域的软件缺陷数据集，为后续研究提供数据支持 * 一个基于深度学习的源代码缺陷定位模型，具有较高的预测和定位缺陷的效果 * 一篇结果分析论文报告，展示模型在预测和定位缺陷方面效果，给出模型的缺陷定位结果和原因解释，并与同类型其他现有方法进行对比分析。  （6）项目研究进度安排  * 第1月：确定项目目标和创新点 * 第2月：选择研究方法和技术 * 第3-4月：构建训练数据集 * 第5-7月：掌握模型搭建方法 * 第8-9月：复现热门模型，优化模型效果 * 第10月：撰写研究报告，展示项目成果并准备结项材料  （7）已有基础 **1.与本项目有关的研究积累和已取得的成绩**  在暑期学校中掌握了使用Git版本管理工具进行团队开发。  参加Python语言与数据科学通选课，掌握Request，NumPy与Pandas库的使用  **2.已具备的条件，尚缺少的条件及解决方法**  当前条件：   |  |  | | --- | --- | | 已具备的条件 | 尚缺少的条件 | | Python爬虫技术，能够从网络上获取大量的软件代码作为训练数据 | 深度学习和神经网络技术，能够构建和训练有效的模型来识别和修复软件缺陷 | | Python数据科学技术，能够使用NumPy库完成基本的数据操作 | 软件缺陷方面的知识，能够对软件缺陷进行合理的定义和分类 |   解决方法：   * 深度学习和神经网络方面，参加《动手学习深度学习》系列课程，通过网站上配套的代码巩固所学知识；学习《Transformers快速入门》课程 * 软件缺陷方面，根据《Introduction to Software Testing》了解传统软件缺陷定位技术。 |

#### 三、 经费预算

| **开支科目** | **预算经费（元）** | **主要用途** | **阶段下达经费计划（元）** | |
| --- | --- | --- | --- | --- |
| **前半阶段** | **后半阶段** |
| 预算经费总额 | 5000.00 | 无 | 2500.00 | 2500.00 |
| 1. 业务费 | 5000.00 | 无 | 2500.00 | 2500.00 |
| （1）计算、分析、测试费 | 3000.00 | 租用云GPU  训练深度学习模型 | 1500.00 | 1500.00 |
| （2）能源动力费 | 0.00 | 无 | 0.00 | 0.00 |
| （3）会议、差旅费 | 0.00 | 无 | 0.00 | 0.00 |
| （4）文献检索费 | 1000.00 | 获取相关课程  和电子书 | 1000.00 | 0.00 |
| （5）论文出版费 | 1000.00 | 争取在国内期刊  如《软件学报》上  出版论文 | 0.00 | 1000.00 |
| 2. 仪器设备购置费 | 0.00 | 无 | 0.00 | 0.00 |
| 3. 实验装置试制费 | 0.00 | 无 | 0.00 | 0.00 |
| 4. 材料费 | 0.00 | 无 | 0.00 | 0.00 |
| **学校批准经费** |  | | | |

#### 四、 指导教师意见

|  |
| --- |
| 选题新颖，内容合适，思路清晰，方法可行，结果可期。  **导师（签章）：**  **年        月        日** |

#### 五、 院系意见

|  |
| --- |
| **学院（盖章）：**  **年        月        日** |

#### 六、 学校意见

|  |
| --- |
| **学校（盖章）：**  **年        月        日** |

|  |  |
| --- | --- |
|  | 8 / 9 |