东南大学校/院级SRTP项目中期检查表

院系名称：

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 项目名称 | 基于深度学习的源代码错误定位研究 | | | | |
| 项目编号 | 202461011 | 学生负责人 | 吴清晏 | 指导教师 | 李必信 |
| 1. 项目进展状况：   根据任务书，我们计划于4月前完成项目目标确立和创新点选取，研究方法和技术选择，并基于选择的技术路线完成训练数据集的搭建。  目前，我们已经明确了项目目标：比较基于抽象语法树(AST)的CodeBERT，基于数据流图(DFG)的GraphCodeBert和同时训练代码处理和自然语言处理的UniXcoder这三种预训练模型在源代码缺陷定位任务上的表现。我们的创新点在于，此前还没有研究关注最新的NL-PL双模态大模型在错误定位(Fault-localization)上的应用。我们的选择的技术路线为静态错误定位，研究方法类似自然语言处理领域的预训练+微调模式。  数据集方面，我们原计划通过爬虫获取Github错误报告的形式构建，但在研究过程中，我们发现大型开源项目往往会使用类似Bugzilla的错误跟踪系统和类似git的版本控制系统，通过调用对应的api就可以获取错误报告和软件更新历史。此外，为了与其他静态错误定位研究进行横向对比，我们还需要选择更为通用的基准数据集。因此，我们从此前研究常用的数据集(Buglocator的数据集，Ye等人研究LR技术提出的数据集，Bench4FL评估框架等)出发，构建了一个Java数据集，该数据集通过Bugzilla获取错误报告，通过git获取代码修改信息和更新日志。我们选择的预训练模型已经在CodeSearchNet代码搜索数据集上进行了预训练，并可以针对特定编程语言进一步迁移训练。我们计划把预训练模型使用的Java语言代码搜索数据集与我们此前构建的Java语言错误报告数据集进行混合，但目前还没有完成格式的统一。  模型训练方面，我们申请了东南大学大数据共享平台账号，并使用云计算节点完成了训练数据集的预处理(使用云CPU)，并尝试使用云GPU对预训练模型进行迁移训练。训练结果如图所示，其中灰线是云服务器上训练时损失率随着训练轮数的变化曲线，其他线是在本机上用数据集中的一小部分测试代码时的结果，可以看到随着训练轮数的增加，损失函数的值呈收敛趋势，且数据量大的情况下曲线更稳定。    最后，在研究的过程中，我们共同完成了一篇研究基于机器学习的静态错误定位技术的综述报告，报告在附件中。 | | | | | |
| 1. 项目存在问题及解决方案   项目目前在工作沟通上存在一定问题。经过成员变更，本小组共4名组员，2名来自吴健雄学院，1名来自计算机学院，1名来自软件学院，很难有机会线下沟通，而线上很难及时发现分工中的问题。目前解决方案是提前一周确定组会时间，尽量让所有人都能到场。此外，组会上大家参与度不高，后续会进一步优化分工，保证每个人都能充分加入到研究中。此外，我们还存在课业压力大，组员时间有限的问题。我们的2名来自吴健雄学院的组员均有自己的SRTP小组，相对来说精力有限，另外两名组员也有参加社团。后续我会进一步优化时间安排，尽量给组员足够的时间调整工作计划。  除了沟通问题外，组员知识差异也一度阻碍了项目进行。在项目刚开始时，一半的组员从未接触过Python，所有人都未接触过机器学习。在学长和导师的帮助下，大家现在已经能熟练使用Python语言完成机器学习任务。  最后，项目还面临硬件设备不足和经费不足的问题。尽管我们已经申请了学校的云服务器，但为了保证不会因为代码错误浪费项目经费，需要在真正上传训练任务前在本机上调试代码，但由于数据集较大，组员电脑性能和存储空间有限，无法正常运行代码，需要消耗额外的时间调整参数。后续计划把闲置的硬盘制作成共享盘，将训练数据放到共享盘中，保证组员不会因为硬件原因阻碍实验。 | | | | | |
| 三、指导教师意见：  签字： 日期 | | | | | |
| 四、院系中期检查小组结论及意见：  □ 优秀 □ 良好 □通过 □不通过    推荐排序：  签字： 日期 | | | | | |

注：表头及第一、第二栏由学生负责人填写。（可附页）