计算机学院软件工程专业2022届毕业设计跟踪记录表

学生姓名：王小娅 指导老师：康松林 罗京

毕业设计题目：基于序列到序列架构的缺陷自动修复模型

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **第4周工作记录** | | | | |
| 填表人 | | **王小娅** | 周起始/结束日期 | 20220313-20210319 |
| 本周工作计划：  尝试其他的网络优化方法，如果在CodRep和Bugs2Fix上取得的验证精度和测试均分尚可，则再在Defects4J数据集上进行比对。 | | | | |
| 详情 | | | | |
| 星期日 | 休息 | | | |
| 星期一 | 进行理论学习，主要收获如下  1.RNN的时间步索引(**x**(t)中的t)未必表示现实中流逝的时间，也有可能仅表示序列中的位置  2.RNN展开计算过程的优点：(1)可以处理长度不同的序列，保持输入大小相同(2)在每个时间步都可以使用相同参数的转移函数  3.seq2seq架构的明显不足是编码器RNN输出的上下文C维度太小难以概括一个长序列，解决思路(1)让C成为变长序列(2)引入注意力机制  4.贪婪搜索为何不能保证得到最优输出序列(采用束搜索的原因)  5.注意力机制应用于解码器计算背景向量并更新隐藏层状态 | | | |
| 星期二 | 1.在本地win10配置miniconda虚拟环境 | | | |
| 星期三 |  | | | |
| 星期四 |  | | | |
| 星期五 |  | | | |
| 星期六 |  | | | |
| 本周任务执行情况： | | | | |
| 下周工作计划： | | | | |
| 本周工作体会(至少200字): | | | | |
| 指导教师意见： | | | | |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **第3周工作记录** | | | | |
| 填表人 | | **王小娅** | 周起始/结束日期 | 20220306-20210312 |
| 本周工作计划：   1. 继续阅读sequencer模型+OpenNMT框架代码。 2. 将sequencer模型的编解码器替换为门控神经单元(GRU)进行消融实验，收集数据。 | | | | |
| 详情 | | | | |
| 星期日 | 休息 | | | |
| 星期一 | 改Bug未成，改的过程中学到了一些理论： 1.自回归模型中自回归和序列生成的关系  2.Python类中的self类似C++中的this  3.embedding是将词转化成向量表示 | | | |
| 星期二 | 1.昨天的Bug仍然没能解决，卸载重装了更低版本的torchtext来避免用torchtext.legacy.data取代torch.data。  2.参加组会，汇报进度。 | | | |
| 星期三 | 安装更低版本的torchtext后，在torch==1.6.0，torchtext=0.7.0的环境下复现了sequencer在CodRep的1,2,3&5四个部分和Bugs2Fix的全部组成的数据集上的训练过程。 | | | |
| 星期四 | 1.解决复现sequencer在CodRep4测试集上的测试过程中的bug:  RuntimeError: Integer division of tensors using div or / is no longer supported, and in a future release div will perform true division as in Python 3. Use true\_divide or floor\_divide (// in Python) instead.  sequencer-test.sh done  解决方式：将 / 替换为 // | | | |
| 星期五 | 1.学习linux指令：将标准输入输出和错误重定向，和文件描述符  2.复现了sequencer在CodRep4测试集上的测试脚本sequencer-test.sh  结果：  3.尝试修改训练脚本里rnn的类型，即给data/train.sh添加以下内容  --rnn\_type GRU(在rnn\_size和global\_attention之间)  模型可以train起来，但是验证精度基本停在了17%，和之前的87%相差过大，说明此方案不可行，以下是训练进行到10000/20000 steps的截图。 | | | |
| 星期六 | 阅读《深度学习与神经网络》第七章-网络优化与正则化的第1、2、7节，学习了小批量随机梯度下降中的基本概念如epoch,iteration,batch\_size以及改变这些量带来的影响，还学到了正则化的一些基本方法如L1L2正则化、权重衰减、丢弃法(dropout)。 | | | |
| 本周任务执行情况：   1. 在torch==1.6.0，torchtext=0.7.0的环境下复现了sequencer在CodRep的1,2,3&5四个部分和Bugs2Fix的全部组成的数据集上的训练过程以及在CodRep4测试集上的测试过程。 2. 尝试将sequencer的LSTM替换为GRU进行训练，得到的验证精度很低，说明此方法不可行，或者需要结合一些别的网络优化方法。 | | | | |
| 下周工作计划：  尝试其他的网络优化方法，如果在CodRep和Bugs2Fix上取得的验证精度和测试均分尚可，则再在Defects4J数据集上进行比对。 | | | | |
| 本周工作体会(至少200字):  1.配环境踩坑收官。安装pytorch和torchtext时我过于迷信官网提供的shell命令以及网上的各种教程，最终发现我做的工作不必用到torchaudio和torchvision，只需要保证torchtext和pytorch版本相匹配即可，cuda选择了向下兼容，终于复现了sequencer在CodRep和Bugs2Fix组成的数据集上的训练和测试。经过20000个steps的训练，得到了87.1992的验证精度和1.65342的困惑度，算是比较满意的结果。在测试集上得到-0.16的平均预测得分(还不理解是什么)和1.1735的困惑度。  2.我找到关键的配置网络结构的地方(一个shell脚本)，加入了--rnn\_type GRU，将编码器和解码器的rnn都从lstm换成了gru，其他配置不变，但是训练到1/4，精度几乎停在了17%，说明了这个想法不具备可行性。下一步我会阅读《神经网络与深度学习》的第七章网络优化与正则化，系统地了解一下优化网络的思路，再进行合理的改进尝试。  3.在实验的过程中我有很多茫然的时候，感觉这也不清楚，那也不清楚，不知道该从哪个了解起。发现因为没有对自己的知识储备正确认识，开题报告写的进度计划可行性不强。 | | | | |
| 指导教师意见： | | | | |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **第2周工作记录** | | | | |
| 填表人 | | **王小娅** | 周起始/结束日期 | 20220227-20220305 |
| 本周工作计划：  掌握OpenNMT-py这一神经网络机器翻译框架，初步实现序列到序列缺陷修复模型，训练后能完成基本的单行修复。 | | | | |
| 详情 | | | | |
| 星期日 | 配置用于运行深度学习模型的cuda\_11.5.2\_495.29.05\_l。 | | | |
| 星期一 | 发现实验室服务器已经具有10.2版本的cuda环境，于是安装了基于python3的miniconda3，用conda安装和cuda匹配版本的pytorch。 | | | |
| 星期二 | 1.调试sequencer模型开源代码，解决bug：TypeError: \_\_reduce\_ex\_\_() takes exactly one argument (0 given)。  解决办法：用miniconda创建python3.6虚拟环境。  2.运行Open-NMT的demo，英语-德语机器翻译。测试集上结果PRED AVG SCORE: -2.4961, PRED PPL: 12.1354。 | | | |
| 星期三 | 1.发现之前将服务器cuda版本弄错(应为11.2)，在python3.6虚拟环境下重装pytorch 1.10.1和对应版本的torchtext 0.9.1。  2.解决bug：RuntimeError: [enforce fail at inline\_container.cc:222] . file not found: archive/data.pkl。  解决办法：删除被损坏的final.train.0.pt重新训练生成一个。  3.解决bug：RuntimeError: Subtraction, the `-` operator, with a bool tensor is not supported. If you are trying to invert a mask, use the `~` or `logical\_not()` operator instead。  解决办法：将1-mask改为~mask。 | | | |
| 星期四 | 训练sequencer模型，时长约1.5小时。 | | | |
| 星期五 | 1.解决bug：RuntimeError: index\_select(): Expected dtype int32 or int64 for index  解决办法：将prev\_k改成prev\_k. type(torch.IntTensor)。  2.调试预测模型的脚本，用昨天训练好的模型进行预测(未调试好)。  3.完善调研报告的*摘要*和*结论*部分。 | | | |
| 星期六 | 1.阅读截断操作python代码-truncate.py。  2.了解单行缺陷数据集1)CodRep 2)Bugs2Fix。 | | | |
| 本周任务执行情况：   1. 配置好miniconda+PyTorch环境。 2. 复现了模型sequencer的训练过程(解决了不少bugs)。 3. 阅读了部分的源码，重读了部分sequencer论文。 | | | | |
| 下周工作计划：   1. 继续阅读sequencer模型+OpenNMT框架代码 2. 将sequencer模型的编解码器替换为门控神经单元(GRU)进行消融实验，收集数据。 | | | | |
| 本周工作体会(至少200字):  本周搭建深度学习环境踩了很多坑，主要是各个包之间版本匹配的问题，网络上的信息良莠不齐，用conda,pip还是wget下载安装包来安装会导致不同结果，还要注意下载的提示，conda可能会顺便更新已安装包的版本。  在这个过程中用熟了一些linux的常用指令，如echo $查看环境变量的值，cat -n 查看加入行号的文件，知道vim编辑器如何定位到某一行，nohup在后台运行一个指令，了解了运行.sh脚本的多个命令。  复现sequencer模型的过程也并不顺利，由于不清楚论文作者做实验的具体环境，更改了很多包版本带来的错误，最后运行成功了训练sequencer的脚本sequencer-train.sh，sequencer-test有一个NoneType错误没有解决。  下周我会继续结合论文和代码深入了解sequencer的实验细节，争取跑通替换LSTM为GRU的训练模型。 | | | | |
| 指导教师意见： | | | | |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **第1周工作记录** | | | | |
| 填表人 | | **王小娅** | 周起始/结束日期 | 20220220-20210226 |
| 本周工作计划：  调研毕业设计选题的研究现状，确定研究的问题、方案和难点，完成一篇相关领域的外文文献翻译。 | | | | |
| 详情 | | | | |
| 星期日 | 翻译论文<Attention is all you need>  [1]Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[J]. Advances in neural information processing systems, 2017, 30. | | | |
| 星期一 | 调研缺陷自动修复技术的研究现状，重点了解了基于机器学习的缺陷自动修复技术近年来提出的重要模型，如DeepFix，CoCoNuT，完成调研报告的*绪论*部分。 | | | |
| 星期二 | 确定完成课题需要的相关技术基础，包括缺陷上下文抽象、序列生成模型、补丁推理、补丁准备，形成调研报告的*相关技术基础*部分。 | | | |
| 星期三 | 查阅资料确定本课题的重点研究内容三项，设计相应的实验方案，形成调研报告的*研究内容及实验方案*部分。 | | | |
| 星期四 | 阅读相近的论文，了解本课题可能遇到的主要问题，查阅已有的解决方案并结合相关知识设想初步解决方案，形成调研报告的难点及初步解决办法部分。 | | | |
| 星期五 | 确定预计的研究成果，安排好进度，形成调研报告的*预期成果及时间安排*部分。 | | | |
| 星期六 | 完成调研报告的*结论*，*摘要*部分，调整好格式。  根据老师的建议修改调研报告。 | | | |
| 本周任务执行情况：  基本完成了工作计划，调研报告仍需小修。 | | | | |
| 下周工作计划：  掌握OpenNMT-py这一神经网络机器翻译框架，初步实现序列到序列缺陷修复模型，训练后能完成基本的单行修复。 | | | | |
| 本周工作体会(至少200字):  本周的调研中我详细地了解了基于机器学习的缺陷自动修复中这一分支领域的发展现状，相信将发展正热门的机器学习有关技术和缺陷修复自身的成功经验进行结合，我可以完成一个表现尚可的缺陷修复模型。下周起就开始阅读代码，一边学习机器翻译框架，一边着手编写模型。  进度之外我还有一些方法论上的体会，即应该以目标为驱动地去学习，如果总想着做好充分准备之后再开始输出，只会导致漫无目的地学习而浪费大量时间精力。只有去做才能分辨虚假的困难和实际的问题。 | | | | |
| 指导教师意见： | | | | |