**毕业设计（论文）中期检查表**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **学生姓名** | 曾旭阳 | | **学 号** | | 162140212 | |
| **专 业** | 物联网工程 | | **指导教师** | | 蔡淼 | |
| **题目名称** | 大模型KVCache优化方法研究 | | | | | |
| **学 生**  **精力投入** | 日平均工  作时数 | 4 | | 周平均  工作时数 | | 20 |
| **教 师**  **辅导情况** | 指导方式 | 线上 | | 每周指导  次数 | | 1 |
| 每次指导  时数 | 1 | | | | |
| **毕业设计（论文）工作进度：** | | | | | | |
| **已完成的**  **主要内容** | 1. 测试gpt-2模型在wikitext-2、Penn Treebank (PTB)、imdb数据集上的推理速度； 2. 测试bert-base-uncased模型在imdb数据集上的推理速度； 3. 完成kvcache manager核心模块的设计并实现代码。 | | | | | |
| **已完成的内容占总任务的** % | | | | | |
| **待完成的**  **主要内容** | 1. 测试并优化kvcache manager模块的功能； 2. 测试使用了kvcache manager模块后，gpt-2模型在相同数据集上的推理速度； 3. 测试gpt-2模型在更多数据集上的推理速度； 4. 毕业论文初稿。 | | | | | |
| **存在的问题和解决办法** | （1）如何减轻推理过程中GPU显存压力，提高推理速度？  采用kvcache manager，将利用率低的kvcache搬运到cpu进行运算，利用率高的kvcache留在GPU。  （2）对模型在数据集上的推理速度测试  部署在云服务器上进行推理速度测试，根据测试结果，优化代码。 | | | | | |
| **学生签名：**曾旭阳 | | | | | |
| **指导教师**  **意见** | **学生已完成毕设工作占总任务的 %。**  **学生所填内容是否属实：**  **是否通过中期检查：**  **其他建议：**    （说明：教师在系统中填写或勾选后，学生才能打印）  **指导教师签名：** | | | | | |

南京航空航天大学教务处制表