­­

**哈尔滨工业大学（深圳）**

**大一年度项目中期检查报告**

**项目名称：**

**项目负责人：** **学 号：**

**联系电话：** **电子邮箱：**

**学 院：**

**指导教师：** **职 称：**

**联系电话：** **电子邮箱：**

**学 院：**

**填表日期： 年 月 日**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 姓名 | 性别 | 所在学院 | 学号 | 联系电话 | 本人签字 |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |

**一、项目基本信息**（包括项目负责人、按顺序）

**立项背景**（200字以内）

|  |
| --- |
|  |

**项目研究内容及实施方案**（400字以内）

|  |
| --- |
|  |

注：请根据项目立项报告，概述要点，简要叙述立项背景、项目研究内容及实施方案。

**二、项目研究中期报告**（请根据以下内容填写，字数在1500字左右）

|  |
| --- |
| （一）项目实施的进展情况及取得的成果  目前，项目正在开发一个NLP应用，使用Qt/C++编写了一个简陋的前端。虽然我们计划在后期重新设计前端，但我们已经学习了一些NLP算法，如RNN、LSTM、transformer等，以及基于transformer的一些模型，如BERT、T5等，并了解了一些基本原理。  我们的小组成员熟悉Linux环境，能够编写一些简单的Shell脚本，并能使用一些命令行工具，如Docker等，能够完成一些简单的项目部署。我们计划在后端开发完成之后，使用一些Web框架（例如Flask）重新设计前端，这个想法来源于目前很多应用，比如Jupyter/ALIST等。  我们小组成员是学会如何利用pytorch搭建神经网络，以及API文档编程，尽管在我们这个NLP项目中不会接触到这么底层的了。  在NLP应用中，Transformer已经成为了一种非常重要的模型。它是一种基于自注意力机制的神经网络，旨在解决序列到序列学习问题。Transformer的优点在于能够处理长序列数据，并且不需要像RNN那样依赖于序列顺序。同时，Transformer还可以并行计算，因此在训练和推理速度方面具有优势。  除了Transformer之外，RNN和LSTM已经逐渐被淘汰。这是因为它们无法有效地处理长序列数据，并且在训练和推理速度方面较慢。与此相比，Transformer更加高效和灵活。  在NLP应用中，BERT和GPT是两个非常重要的模型。BERT是一种基于Transformer的预训练语言模型，可用于各种NLP任务，如文本分类、命名实体识别等。GPT是一种基于Transformer的生成式语言模型，可以用于生成文本、对话系统等任务。这些模型在NLP领域具有广泛的应用前景。  总之，在NLP应用中，选择合适的模型非常重要。我们将继续努力学习和实践，为开发出更加优秀的NLP应用而努力。  （二）遇到的困难及下一步工作计划  走了一些弯路。学习了一些传统的机器学习的算法决策树/kmeans/XGBoost等，但事实上这些传统的机器学习算法更多的适用于结构化数据的分类或回归问题。以及网络上资源良莠不齐，以及一些书籍的落后。  以及学习反向传播，梯度下降，transformer的时候数学水平跟不上。  在自然语言处理领域，数据集的数量一直是一个普遍的难题。我们做的是一个消息分类，数据集不同很多现成的新闻分类，因此需要自制数据集。但即便是计科飞书群通知和哈工大官方通知，其数据集数量也仍然有限，或许只有千条左右，我们不能保证训练出来的效果如何。  当然，我们会在实现NLP项目时，通过一些技术手段来制作数据集。我们小组决定采用爬虫技术，从飞书和哈工大官网上爬取一些通知，并将这些数据进行打标记。当然，这个分类过程可能需要人工参与，我们计划使用GPT提供的API以及Doccano等技术来实现数据的标记。  在数据集制作完成后，接下来就是模型训练。在Hugging Face上，有许多现成的BERT模型可供使用。我们小组打算采用迁移学习等技术来实现模型的训练，这样可以极大地缩短训练时间，并提高训练出来的模型的准确率。由于我们只是做一个文本分类，因此使用BERT模型即可。  此外，我们的目标还包括文本摘要。事实上，文本摘要已经有了许多现成的训练好的模型，例如 [mT5\_multilingual\_XLSum](https://huggingface.co/csebuetnlp/mT5_multilingual_XLSum)，我们打算直接拿来使用。  我们计划在暑期完成后端的开发。  （三）结题预期目标   * 实现基于 flask 的 web 前端 * 实现文本分类的任务 * 实现将消息的摘要输出到对应日期的 ddl 日历里   （四）经费使用情况  暂无经费使用。 |

|  |
| --- |
|  |

**二、项目研究中期报告**

**三、指导教师意见**

签 名： 年 月 日

**四、项目专家组意见**

1．是否达到中期目标（在□内打√）：

□达到中期目标 □基本达到中期目标 □未达到中期目标

2．成绩评定（在□内打√）：

□合格

□改进后可继续执行

□不合格，项目实施 □提出警告、观察后再定继续执行或中止

□中止实施

3. 其他意见和建议：

组长签名： 年 月 日