机器学习基础期末项目实验报告：ChatExcel

马千里 2000011005

张少彤 2000017739
衣智远 2000011093

2024 年 1 月 14 日

1 收集数据

ChatExcel 目前主要支持常用的统计特性及简单的字符操作等，比如简单的查询、排序、计算指令等，并且还要求表格内容顶格放置在左上角，必须上传带有表头内容的表格。

基于此，我们从两方面出发，开展数据收集工作。一方面，我们收集了多个构建数据集所基于的 Excel 表。我们按照“表格内容顶格放置在左上角，必须上传带有表头内容的表格”的要求，在课程所提供的民用汽车拥有量汇总.xlsx 之外，还搜集并整理出了农林牧渔业总产值.xlsx、资本存量核算.xlsx、地区生产总值统计.xlsx 等表格，用多套 Excel 表实现数据集的互现，为大模型创造更多的表格提问条件，使数据集更加丰富可靠，为大模型的训练和调试提供更多保障。

另一方面，我们从简单查询、统计计算和图表展示三大类指令出发，用自然语言对每类指令编制若干个问题，并编写对应的 python 可执行代码，形成多个 query-answer 对。在“简单查询”中，我们设计了多个与简单检索、查找、计算相关的问题；在“统计计算”中，我们的问题包含对多种基本统计量（如总和、增长率、平均值、标准差、方差、最值等）的计算，以及对表格中数据的排序；在“图表展示”中，我们从饼图、折线图、柱状图、柱线混合图、堆叠柱状图、堆叠面积图、热图、玫瑰图、雷达图、散点图等多种数据统计中常见图表类型出发，针对每个图表类型提出多个问题，其中包含对同一表格不同的 sheet 作对比分析，以及对多个表格的交叉分析。随后，并编写回答这些问题所需的 python 代码，最后在本地环境上运行调试，依据输出结果对效果进行核对，修改代码，保证代码的可执行性和执行结果的正确性，从而严格保证数据集的质量。最终，我们整理形成了完整的多个 query-answer 对，作为微调开源大模型的数据集。

后续，我们又依据模型特性，先后将数据集整合、调整为.py 和.jsonl 格式。上述数据集将被我们在训练模型时用作训练集的素材来源。再然后，我们依据 query-answer 对的形式，结合前述三大类指令要求，准备了三套测试集，用以调试模型。最终，数据集的总数达到约 200 条。

2 选择模型架构

我们对现有的 ChatGLM、Llama、baichuan 等开源大语言模型进行了衡量。

ChatGLM 拥有先进的深度学习技术与海量中文语料的训练成果。在自然语言理解与生成方面展现了出色的性能，为国内的自然语言处理研究与应用提供了强有力的支持，其最大的特点是卓越的自然语言处理能力。

Llama 具备具有高效的并行计算、良好的可扩展性和强大的图处理能力，偏重于对计算任务的图抽象和多级优化，其最大的特点是面向大规模数据处理和并行计算问题。

baichuan 则显现出面向用户、面向复杂自然语言任务的特点。

最终，我们综合了三者的优缺点，我们认为 ChatExcel 首要解决的问题是操作，主要面向特定场景下的垂直领域问题，以自然语言模型为前端，所面向的指令集的关键点也是落在自然语言处理上面。同时，ChatExcel 要求我们归纳符合人类各种提问的 query 部分，并且 query 需要具有明确的意义和指令，这也对我们所依托的大模型的自然语言处理能力提出了要求。此外，我们还综合考虑了模型发展历程、相关资料丰富程度、调节难易等问题，最终，我们选取了发展历程相对较久、相对容易入门，并且以自然语言处理能力为特点的的 ChatGLM3-6b 作为我们主要调节的模型架构。

3 训练并调试模型

3.1 提示词

ChatGPT 的难点，在于 Prompt（提示词）的编写。如果没有提示词，或者提示词编写得不合适，那么我们在训练模型时可能得不到想要的效果。

因此，我们对提示词进行规范的编写。

我们学习了参考资料中的《Prompt Engineering for Developer》教程，了解了大模型 Prompt 工程的提示原则与迭代过程，尝试了文本概括推断、转换和拓展等多种功能。同时也有意基于此规范优化本次训练中准备的提示词集，而在某种程度上，此项任务功能属于文本转换的范畴，可理解为从自然语言向 python 解释器理解的可执行 python 代码的“翻译”。

另外，我们在编制数据集和训练集的过程当中，也注重从背景和指示两个主要角度，使我们的用语符合提示词的规范性。

我们的背景分为三个维度，即表头内容、横轴起止及含义、纵轴起止及含义。

我们的提示开头和结尾都放有定界词。一方面，在整个 query 中，“背景知识”和“query”本身都起到定界词的作用，这可以防止大模型混淆已知背景与未知问题；另一方面，query 中的 query 部分均以文件路径开头，问题句式力求明确，同样也起到定界词的作用。

3.2 整合并调节数据集

通过 message 中关于 system 和 user 的 content 进行区分，前者主要提供背景知识，即提前说明需要进行的操作对象和大致目标，而在后者即 user 的查询中，再详细注明本次查询的文件地址、内容格式和查询需求，以期待微调后的大模型能够实现不限于所给特定表格的泛化但又准确的“翻译”。

值得一提的是，在我们编写积累足够的查询语句和对应代码后，完全可以通过 chatgpt 的格式转换相对快速高效地准备正确 jsonl 格式的数据集。

我们希望，大模型可以进行结构化的输出，因此我们将数据集整合为 jsonl 格式，以便大模型读取和后续的进一步微调。ChatGLM3-6b 的 README.md 文档中已经对它所要求的 jsonl 格式进行了说明，在阅读了 README.md 后，我们确定的 jsonl 格式模板具体如下：

```
[
  {
    "conversations": [
```

```
      {
        "role": "system",
        "content": "<system prompt text>"
      },
      {
        "role": "user",
        "content": "<user prompt text>"
      },
      {
        "role": "assistant",
        "content": "<assistant response text>"
      },
    ]
  }
  // ...
]
```

最终，我们将数据集整合成 train-final.jsonl 和 test-final.jsonl，用于后续的训练和调试。

3.3 训练模型、结果简述与调试

随着大模型的发展，基础大模型基于文本训练数据，有着不错的预测接续词句的能力，但是如本任务一样，为了实现更加精准恰当的文本处理，我们在预训练的大模型上依照准备的数据集进行进一步训练和微调。查阅 OpenAI 的用户文档可知，准备好可微调的 API-KEY 和正确格式的数据集后，即可调用简单的函数创建训练任务，获得微调模型。

我们主要调节的广义参数有数据集的条数 data、迭代轮数 epoch、学习率 learning-rate 等、步数 step 等。我们先后依托学校提供的超算平台和阿里云服务器进行训练和调试。

我们先后进行了四次大的有效训练-调试过程，分别输出为 data=65-step=1000-lr=1e-2.out；data=113-step=1000-lr=1e-2.out；data=153-step=300-lr=5e-3.out；data=210-step=500-lr=1e-2.out。

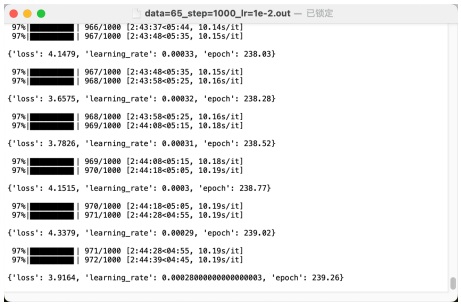

图 3.1: data=65,step=1000,lr=1e-2

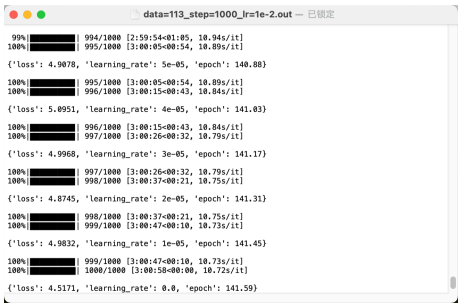

图 3.2: data=113,step=1000,lr=1e-2

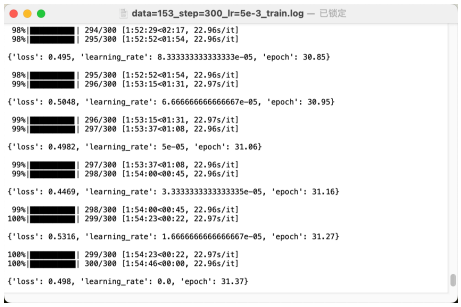

图 3.3: data=153,step=300,lr=5e-3

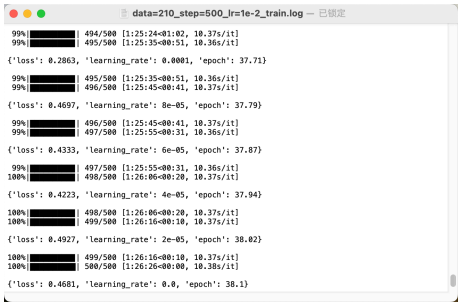

图 3.4: data=210,step=500,lr=1e-2

4 结果与思考

最初我们进行尝试性调试时，所得到的 loss 高达 0.2-0.4，train-accuracy 达到 0.95（见 65-32-step-metrics.csv）且随着训练过程的推进，loss 呈现出先降低、后回升的特点，峰值甚至上探到 0.5 左右（如下图）。经过分析，我们发现，在最初的几次调试中，所采用的参数呈现出数据集条数较少（data=65，113），step，learning-grate 较大的特点，loss 回升的原因，一是数据量相对较少，二是二是代码输出可能不止一种结果，比如同一种操作可能会存在远多于一种的代码解决方案，所以模型输出和预期之间并不是完全一对一契合的状态，三是可能存在过拟合现象。

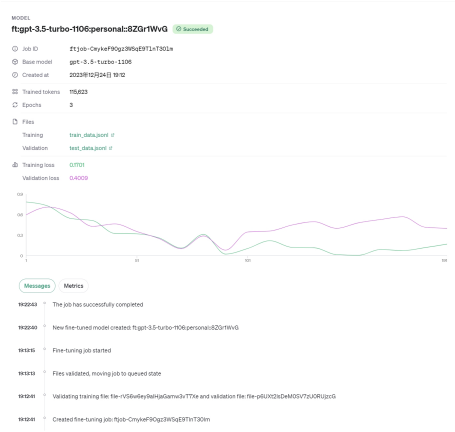

图 4.1: 最初的尝试性调试所得到的 loss 效果不理想

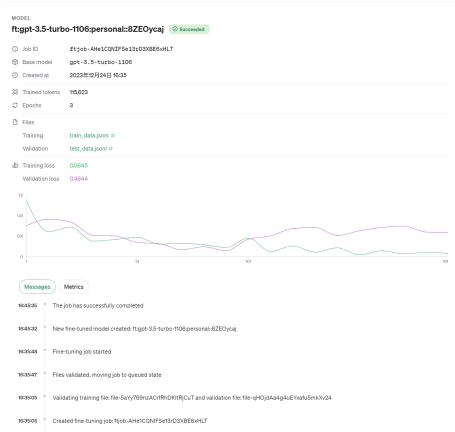

图 4.2: 最初的尝试性调试所得到的 loss 效果不理想

因此，我们对所用参数进行了有针对性的调节，一方面，我们将数据集的条数扩大到 210，另一方面，我们降低了 step，并对 learning-grate 进行了动态调节。在 ChatGLM3-6b 开源大模型上进行微调之后，loss 回落至 0.003-0.004，最低下探至 0.0028，相较于调节之前有显著降低，train-accuracy 升至 0.96，且趋稳速度有所改善。

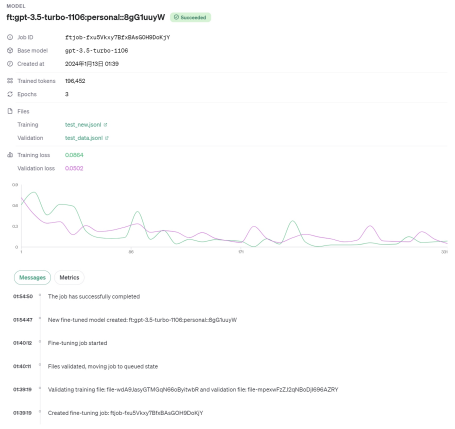

图 4.3: 只调整数据量后，loss 即有显著改善

5 小组分工

马千里：“图表展示”数据集、训练集制作，模型架构选择，总数据集、测试集整合，主持模型微调

张少彤：“简单查询”数据集、训练集制作，参与模型微调、实验报告撰写

衣智远：“统计计算”数据集、训练集制作，参与模型微调，中期 PPT 制作与汇报，实验报告撰写