**深 圳 大 学 实 验 报 告**

**课程名称：­ 数据挖掘导论**

**实验项目名称： 从头训练大模型**

**学院： 计算机与软件学院**

**专业： 计算机科学与技术（创新班）**

**指导教师： 陈小军**

**报告人：何泽锋 学号：2022150221 班级： 高性能特色班**

**实验时间： 2025年4月1日——2025年4月14日**

**实验报告提交时间： 2025年4月14日**

**教务处制**

|  |
| --- |
| 实验目的与要求：  1.分析不同参数设置对训练minimind模型loss的影响。  2.设计测试数据并评估模型性能。 |
| 实验环境：  1. Anaconda3 开发环境  2. IDE是Cloud Studio(云端IDE)  3.大模型常用的库：PyTorch,Transformers等 |
| 实验内容及过程：  **1.创建实验环境**  考虑到本次实验训练时常较长，因此在服务器和本地都部署了环境，同时运行可以进行更多组数据的对比。  （1）搭建服务器环境  ①注册并登录Cloud Studio，本次实验选择了免费基础模型，硬件参数如下：   |  | | --- | | 显存：16GB  CPU：8核  GPU：Tesla T4  内存：32GB |     图1 配置服务器环境  ②按照实验指导文档从github上clone项目，本次实验采用的是名为MiniMind的项目，下载后的文件如下图所示，包含训练和使用的代码，以及预训练模型等。    图2 下载项目  ③使用conda构建实验解释器，注意到文件中包含requirements.txt，这是项目运行过程所依赖的包，需要配置到conda环境中。打开观察可以看到一些常用的包，例如torch、numpy、pandas、matplotlib等    图3 requirement.txt配置文件  （2）搭建本地实验环境  ①本地环境分析，对比服务器性能和本地性能，发现本地的性能会更好，除了显存较少可能导致batchsize大小受到限制，预期的训练速度会高于服务器，因此实验时可共同使用   |  | | --- | | 显存：12GB  CPU：32核  GPU：Nvidia 4080  内存：1T |   ②下载Anaconda用于配置解释器，同样可以使用requirments.txt安装依赖包，但后续实验时发现，直接配置的情况下pytorch安装的版本为cpu版本，训练速度非常慢，因此还需要到pytorch官网获取下载指令，最终安装的环境包（部分）如下：    图4 本地配置的包  （3）下载模型和数据集  采用git clone <https://huggingface.co/jingyaogong/MiniMind2>下载模型，然后在<https://www.modelscope.cn/datasets/gongjy/minimind_dataset/files>下载pretrain\_hq.jsonl + sft\_mini\_512.jsonl数据集。  （4）体验模型  使用指令运行预训练好的模型，此处采用[0]自动输入，会自动调用提前写好的提问语句与大模型进行对话    图5 体验模型  **2. 分析不同参数设置对训练minimind模型loss的影响**  （1）实验测试了9组不同的参数，接下来将分别分析学习率、批次大小和优化器对loss的影响。  表1 9组不同参数   |  |  |  |  | | --- | --- | --- | --- | | 实验编号 | 学习率 | 批次大小 | 优化器 | | A | 5e-4 | 64 | AdamW | | B | 5e-5 | 32 | AdamW | | C | 5e-4 | 32 | SGD | | D | 5e-4 | 32 | RMSprop | | E | 1e-3 | 32 | AdamW | | F | 1e-4 | 32 | AdamW | | G | 5e-4 | 64 | SGD | | H | 5e-5 | 32 | SGD |   ①学习率（lr）：  测试学习率对loss变化的影响，需要控制变量保持批次大小和优化器不变，因此采用的数据如下表所示：  表2 测试不同学习率下的loss   |  |  |  |  | | --- | --- | --- | --- | | 实验编号 | 学习率 | 批次大小 | 优化器 | | B | 5e-5 | 32 | AdamW | | E | 1e-3 | 32 | AdamW | | F | 1e-4 | 32 | AdamW |   训练得到的loss如下图所示，可以看到，随着学习率的增大，loss下降的更快，而下降的趋势具有一致性，在训练初期，loss下降的较快，随着训练的进行，下降速度越来越慢，最终趋于平缓（收敛）    图6 不同学习率下的loss结果  ②批次大小（batchsize）：  控制其他参数不变，调整batchsize的大小，测试loss的变化，采用的具体参数如下图所示：  表3 测试不同批次大小的loss   |  |  |  |  | | --- | --- | --- | --- | | 实验编号 | 学习率 | 批次大小 | 优化器 | | C | 5e-4 | 32 | SGD | | G | 5e-4 | 64 | SGD |   训练得到两组数据的loss变化如下图所示，可以看到，batchsize增大loss下降的速度显著增加，这是因为数据之间的联系更强。同时可以看到epoch结束的更快，这是因为batchsize增大一倍的情况下，迭代所需次数减半，最终呈现的效果就是收敛的更快。理论上来说，batchsize增大更有利于训练，但是受到显存的影响，无法调整过大。    图7 不同批次大小下的loss结果  ③迭代器（optimizer）:  控制变量的方式于上述两个参数的测试一致，此处选择如下参数进行对比：  表4 测试不同优化器下的loss   |  |  |  |  | | --- | --- | --- | --- | | 实验编号 | 学习率 | 批次大小 | 优化器 | | C | 5e-4 | 32 | SGD | | D | 5e-4 | 32 | RMSprop |   训练得到的loss变化如下所示，可以明显看到SGD迭代器的loss更新非常缓慢，而RMSprop的loss下降很迅速，对于不同的网络需要采用不同的迭代器进行训练，对于大模型问题，通过实验结果可以看出来，SGD迭代器并不合适。    图8 不同迭代器下的loss结果  简单分析两种迭代器   |  | | --- | | **SGD迭代器：**  ·核心思想：每次迭代随机选择一个样本（或一个小批量）计算梯度，直接沿负梯度方向更新参数。  ·参数更新公式：  ·SGD的问题：  ·固定学习率难以同时适应所有参数：  ·学习率太小时，梯度小的参数更新过慢；  ·学习率太大时，梯度大的参数会振荡甚至发散。  ·容易陷入鞍点或平缓区域。 |  |  | | --- | | **RMSprop**  ·核心思想：自适应调整每个参数的学习率，通过对梯度平方的指数移动平均调整步长，缓解SGD学习率固定的问题。  ·特点：  ·自适应学习率：梯度大的参数步长减小，梯度小的步长增大。  ·缓解SGD的振荡问题，适合非平稳目标。 |   为了更清晰体现不同迭代器的效果，此处还进行了一组对比：  表5 测试另一组迭代器的比较   |  |  |  |  | | --- | --- | --- | --- | | 实验编号 | 学习率 | 批次大小 | 优化器 | | B | 5e-5 | 32 | AdamW | | H | 5e-5 | 32 | SGD |   可以看到AdamW优化器也优于SGD，理论来说，AdamW效果会比RMSprop更优，这是因为Adam已经结合了动量和RMSprop（自适应学习率），而AdamW是Adam的改进，将权重衰减独立作用于参数，不受自适应学习率影响，更适合大模型训练。    图9 不同迭代器下的loss结果  （2）总结最优参数  将9组数据显示在一个图中，观察loss下降情况，观察到E组在训练一个epoch后，loss下降最快，其参数为如下表所示，若是继续增大批次大小，loss下降理论上会更快，但实验未进行验证，因此在9组测试中最优解为E  表6 最优参数表   |  |  |  |  | | --- | --- | --- | --- | | 实验编号 | 学习率 | 批次大小 | 优化器 | | E | 1e-3 | 32 | AdamW |     图10 9组参数的loss比较  **3. 设计测试数据并评估模型**  （1）根据模型的训练任务，设计测试输入保存在custom\_prompts.txt中，涵盖不同类型和难度的样本，设计的问题如下图所示，包含基础常识、专业领域、逻辑推理、长文本生成、多轮对话，对每个类别设计了简单和复杂问题，进而更精确的辨别模型的性能。    图11 设计测试数据 |
| ①首先测试预训练参数下的模型，此时的结果如下图，可以看到，回答结果非常差，除了部分基础问题能够回答，其他内容都是答非所问，还出现了回答上的混乱。    图12 预训练模型的测试结果  根据回答效果可绘制如下表格：  表7 测试结果   |  |  |  | | --- | --- | --- | | **预训练模型** | **简单问题** | **复杂问题** | | 基础知识 | √ | × | | 专业领域 | × | × | | 逻辑推理 | × | × | | 长文本生成 | √ | × | | 多轮对话 | × | × |   ②测试训练时得到的loss最小的组合（E组）的回答效果，如下图所示：相对预训练结果，回答的长度有所提升，能够识别简单的问题，但是对于专业知识或者逻辑题目完全无法解答。    图13 E组参数下的测试结果  ③详细分析各个类别问题的效果：  （1）基础常识  ·简单题  ·问题：中国的首都是哪里？  太阳系中离太阳最近的行星是哪个？  ·结果：准确率高（首都、行星事实性问题回答正确）  ·复杂题  ·问题：为什么说“光合作用是地球上生命的基础”？  列举三次科技革命的主要发明，并说明其对社会的推动作用。  ·结果：光合作用解释逻辑混乱，出现重复表述和无关内容（如"古代中国农业"），科技革命回答完全偏离（错误聚焦区块链）  （2）专业领域  ·简单题  ·问题：（医学）人体最大的器官是什么？  （编程）Python中如何定义一个函数？  ·结果：医学基础题错误（最大器官应为皮肤，却答"肌肉关节"）；Python函数定义未给出代码示例  ·复杂题  ·问题：（医学）请解释CRISPR-Cas9技术的原理及其在基因编辑中的应用。  （编程）如何用动态规划解决背包问题？请给出代码示例和复杂度分析。  ·结果：CRISPR技术原理描述模糊，缺乏关键机制；动态规划问题未提供算法，回答内容完全无关  （3）逻辑推理  ·简单题  ·问题：数列2, 4, 8, 16的下一个数字是什么？  ·结果：基础数列题错误（2,4,8,16应输出32，却答9）  ·复杂题  ·问题：有三个人住店，每人付了10元，老板优惠5元让服务员退还。服务员私藏2元，退给每人1元。最终每人付了9元，共27元，加上服务员的2元是29元，剩下的1元去哪了？  ·结果：住店问题未识别27元已包含服务员2元的关键逻辑，回答呈现重复无效语句  （4）长文本生成  ·简单题  ·问题：写一段100字左右的介绍，描述你最喜欢的季节。  ·结果：能提供简答的文字描述，但季节介绍混用"春季"与"金秋酷暑"等表述  ·复杂题  ·问题：以“人工智能的未来”为题，写一篇500字的议论文，要求包含利弊分析。  ·结果：论文未达500字要求，利弊分析未能深入解释  （5）多轮对话  ·简单题：  ·问题：第一轮：你喜欢什么颜色？  第二轮：为什么喜欢这个颜色？  第三轮：这个颜色在文化中有什么象征意义？  ·结果：对问题避而不谈，没有关注问题的关键点，而且未能体现上下文之间的联系  ·复杂题：  ·问题：第一轮：如何评价秦始皇的功过？  第二轮：他的焚书坑儒政策对后世文化有何影响？  第三轮：对比秦始皇和汉武帝的统治策略。  ·结果：事实错误，语句没有逻辑，出现同一个句子的重复输出，答案完全不合理  ④分析结果的准确性、流畅性和多样性  ·准确性  ·部分基础事实类问题回答正确，但复杂问题存在严重事实错误或答非所问。  ·逻辑推理题完全错误，科技伦理类回答内容空洞重复。  ·流程性：  ·语言通顺，句式结构出现一定问题，存在大量无意义重复。  ·部分回答呈现“学术废话”风格（如人工智能的论文）  ·多样性  ·能生成不同长度的文本，但技术类问题倾向于套用固定模板、  ·表达不符合实际，历史问题出现虚构人物背景。  ⑤总结模型优缺点  表8 模型优缺点   |  |  |  | | --- | --- | --- | | **维度** | **优点** | **缺点** | | **知识覆盖** | 能处理部分简单事实问题 | 复杂主题错误率高，专业领域知识混乱 | | **语言生成** | 语法基本正确，但是句式会出现错误 | 内容重复率高，缺乏实质性信息 | | **逻辑能力** | 回答没有逻辑 | 完全无法处理逻辑推理和数学证明 | | **创造性** | 能生成较长文本，但大部分毫无意义 | 虚构内容与事实混淆 | | **实用性** | 不具有可靠性，缺乏实用性 | 技术/学术问题产生误导性信息 |   ⑥改进模型的方法  实验过程采用的是小数据集，但对于大语言模型需要大量的数据支撑，否则训练出来的模型缺乏可靠性和逻辑思维等。对于专业性的知识，更需要有相应的预训练数据来提高准确率。实验过程仅进行了一个epoch的训练，这是远远不够的，需要大量的训练才能得到较好的模型。 |

|  |
| --- |
| 实验收获：  通过本次实验，获得如下的收获：  （1）掌握大模型训练参数调整  实验过程对比不同学习率、批次大小和优化器的组合，了解了参数对模型训练效果的影响。  ·学习率较高时，模型收敛更快；  ·增大批次大小可加速训练，但受显存限制；  ·自适应优化器（AdamW、RMSprop）在单个epoch训练的情况下显著优于SGD，尤其适合大模型训练。  （2）实验设计与分析能力提升  ·掌握了控制变量法对比参数效果的实验设计方法，并通过可视化（如Loss曲线）直观分析训练动态。  ·认识到单次迭代的局限性，需长期训练和多轮验证才能稳定模型性能。  （3）模型评估与问题诊断  ·当前模型能回答个别基础问题，但在专业领域和逻辑推理上存在严重缺陷，需增强预训练数据的专业性。  ·生成文本会出现重复、虚构事实等问题，需引入多样性惩罚机制和事实校验来提高模型精度。 |
| 指导教师批阅意见：  成绩评定：  指导教师签字：  年 月 日 |
| 备注： |

注：1、报告内的项目或内容设置，可根据实际情况加以调整和补充。

2、教师批改学生实验报告时间应在学生提交实验报告时间后10日内。