

大作业报告

张驰 2022010754 zhang_ch22@mails.tsinghua.edu.cn

一、模型的预训练

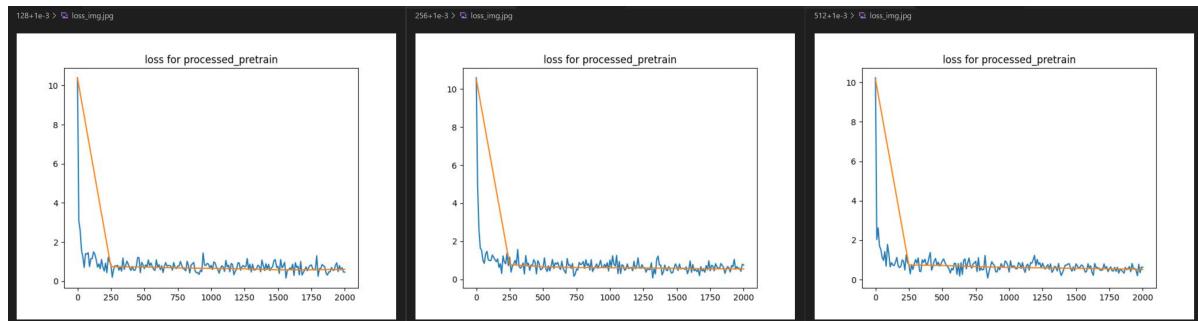
模型使用金庸《射雕英雄传》、《神雕侠侣》、《天龙八部》三本小说进行预训练。在预训练中，我尝试了不同的模型参数并比对了预训练效果。主要调整了batch-size, block-size, learning-rate, max-iterations。得到的结果如下：

1. block_size的影响

首先我测试了不同block_size对模型预训练的影响。下图分别为block_size=128/256/512的情况（其他参数完全相同）：

这些图并非最终版模型的实际曲线。

下三张图中loss下降到很低，这是因为在做这个测试的时候，预训练数据集只用了一本小说。当数据集扩大到三本的时候，loss最终会收敛到2左右。

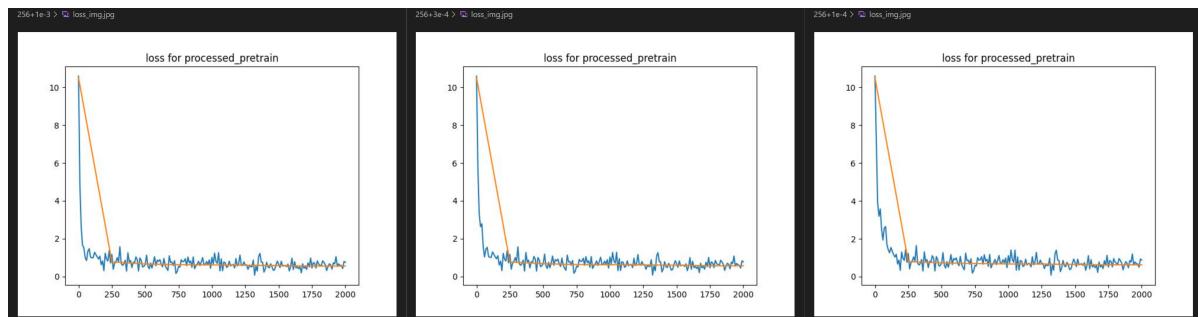


可以看到，增大block_size可以让模型的loss收敛得更慢，但最终收敛的数值却很接近。理论上讲，增大block_size可以让模型学到更长的语句，可能会增强模型的记忆能力。然而在实际测试中，除了block_size=128的模型续写能力略差之外，block_size=256/512/910的模型续写能力相当接近。由于后来在微调过程中，发现当微调的block_size与预训练的block_size不一致会让模型的问答能力变差，而256的block_size更接近微调数据中一般问答的长度（而且由于微调数据较多，若block_size太大会占用太多存储空间），所以最终的预训练模型采用block_size=256。

2. learning_rate的影响

其次我测试了不同learning_rate（和相应的min_lr=learning_rate/10）对模型预训练的影响。下图分别为learning_rate=1e-3/3e-4/1e-4的情况（其他参数完全相同）：

与上面类似，这里预训练集也只有一本小说，故loss收敛值较低。



如图可见，降低learning_rate会让模型的loss收敛地很慢。理论上讲，更低的learning_rate确实会让收敛变慢，而且可以让模型达到更精确的局部最优解。然而在实际测试中，比1e-3更低的learning_rate在实际续写中效果不佳。比如下文是learning_rate=1e-4时，以“段誉”为开头的续写内容：

段誉子向倒，当真拉住了他红，着他又捧父倒师和力，掌提上来。 萧峰摩智，阿朱微一♦
拗，使晃起，只见也不由得，只听见她兵晃公，心中一惊：「爹毛，不免身见帮主，老段誉父你的多的功，我宰亲约，事出去坐一躬之中，这双存了跟我，那若必有不便自己草，不以黑了，终必死知道：“全真请，却日掌萧和我心起的。”莫愁正要我在天白，说道：“尽，妈妈爹！” 段誉道：“你.....小子可要得对，自然不许，我要着你在过我不可是？” 摆头道：“你可是我.....这金么？”郭靖道：“你.....♦

而当learning_rate=1e-3时，效果便好了很多：

段誉一听大吃一惊，急忙回头，要待他走入内院，只见段誉远远传来脚步声响，脚步踉跄，又过一个时辰，段誉和王语嫣相距尚有十余丈，只听得蒙他叫道：“段公子，我手下留情，你自己老是不肯，老子明知出了你的毒手。”段誉道：“我自己自己要杀你，还是少了一下老贼？”王语嫣道：“我不杀你。”段誉道：“我不杀你！”段誉再也忍耐不住，喝道：“你自己杀了我，我才不会杀你。”王语嫣点点头，道：“我不会杀你。” 段誉心想此人说话时无论如何不休，不自禁的住口不说。见她刚一转念间，段誉一飘翻，身

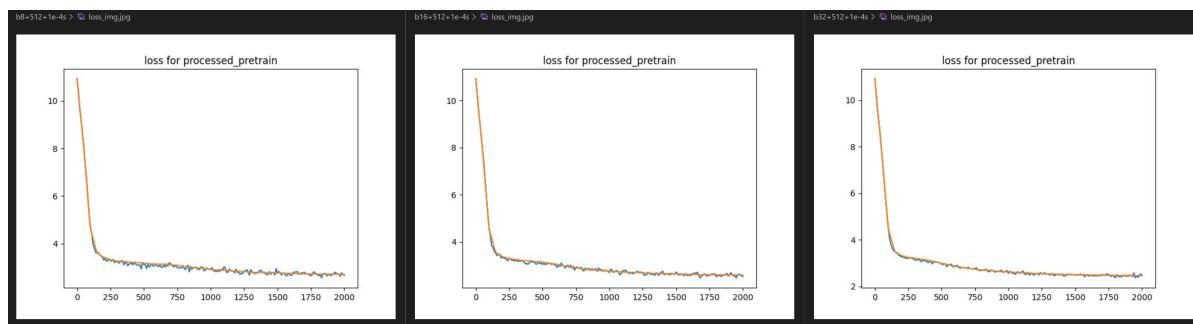
所以在最终版本的模型中，我采用了1e-3的learning_rate。

3. batch_size的影响

然后我测试了batch_size对模型预训练的影响。下图展示了batch_size=8/16/32时的情况（其他参数完全一样）：

可以注意到，这里的loss最终收敛值变高了，这是因为这里同时使用三本小说作为预训练数据集。同时，由于我把eval_interval从250降低到了50，使得黄色的validation_loss的曲线也变得平滑了。

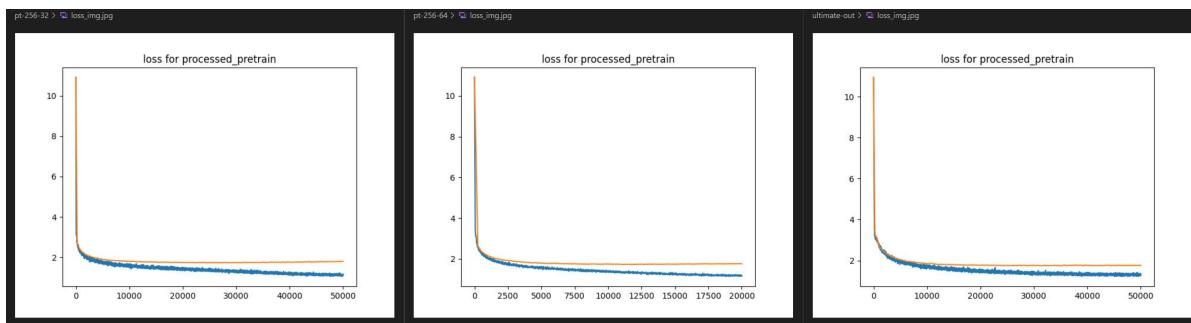
这三张图的learning_rate采用的是1e-4。



可以看到，三张图像非常接近。在实际测试中，三种batch_size值的模型续写能力也不相上下。可见batch_size对这个模型的预训练效果影响不大。

4. 最终版本预训练参数的选择

在准备最终版本预训练时，我增大了max_iterations以获得更好的预训练效果。在之前的多次尝试下，我最终做了以下三个版本的迭代次数较多的预训练。下三张图分别为block_size=256&batch_size=32；block_size=256&batch_size=64；block_size=910（这是在有限显存下能运行的最大block_size）&batch_size=16：



三张图像中，train_loss和validation_loss都得到了很好的收敛。而且，三种不同的训练结果在实测中都表现出了很强的续写能力。我在这三个模型的基础上尝试了进一步的微调。最终，第一张图
block_size=256&batch_size=32的模型基础上的微调取得了最好的效果，所以它也被我采用为最终版本的预训练模型。

二、微调数据的收集方法

本次大作业调用ChatGLM Pro的api，使用ChatGLM Pro模型生成数据。由四名同学合作，通过使用虚拟手机号注册智谱AI，获得了20多个ChatGLM Pro的api-key，从而可以在多个终端并行调用api生成。我们使用模型生成了十几个小时，一共得到309044条（约100M大小）有效数据，基本上每条数据长度在300字左右。

在使用ChatGLM Pro模型时我们发现，如果反复输入相同的prompt，可能会导致输出重复的问题。于是我们在每次的prompt中加入了上一次的输出结果，以减少输出的重复。调用api的具体代码如下：

```

import zhipuai
import json
import time
zhipuai.api_key = "xxxxxx"

text = ['' for i in range(12)]
now = time.time()

text[0] = '''请以以下格式，生成10条关于金庸《天龙八部》的问题和回答。要求回答至少100字。
{"Question": "乔峰的性格特点是什么？", "Answer": "乔峰性格豪迈勇猛，正直忠诚，具有强烈的正义感和民族气节。他在成长的过程中，始终坚守家国大义，面对强大的敌人，他从不退缩。他的人格魅力和英勇事迹为世人敬仰，成为了一位真正的英雄。"}'''

注意，不要生成'''+f"{now}"+'''之前生成过的问题。
注意，你需要用{"Question":""", "Answer":"""}的格式，注意每个问题和回答内部不要换行。'''

text[1] = '''请以以下格式，生成10条关于金庸《神雕侠侣》的问题和回答。要求回答至少100字。
{"Question": "杨过曾经用过哪些化名？", "Answer": "杨过曾经用过化名‘神雕侠’，在江湖中行侠仗义，也曾经化名‘西毒’欧阳锋，以参加华山论剑。"}'''

注意，不要生成'''+f"{now}"+'''之前生成过的问题。
注意，你需要用{"Question":""", "Answer":"""}的格式，注意每个问题和回答内部不要换行。'''

text[2] = '''请以以下格式，生成10条关于金庸《射雕英雄传》的问题和回答。要求回答至少100字。
{"Question": "欧阳锋的武功特点是什么？", "Answer": "《射雕英雄传》中，欧阳锋的武功特点是以毒为主，他的毒技独步武林，令人闻风丧胆。此外，他还擅长多种武学，如蛤蟆功、神龙摆尾等，是一位绝顶高手。"}'''

注意，不要生成'''+f"{now}"+'''之前生成过的问题。
注意，你需要用 {"Question":""", "Answer":"""} 的格式，注意每个问题和回答内部不要换行。'''

text[3] = '''请以以下格式，生成10条关于金庸《天龙八部》的问题和回答。要求回答至少100字。
{"Question": "乔峰的性格特点是什么？", "Answer": "乔峰性格豪迈勇猛，正直忠诚，具有强烈的正义感和民族气节。他在成长的过程中，始终坚守家国大义，面对强大的敌人，他从不退缩。他的人格魅力和英勇事迹为世人敬仰，成为了一位真正的英雄。"}'''

```

{"Question": "《天龙八部》中的八部是指什么？", "Answer": "天龙八部包括：天人（或称天、提婆）、龙，夜叉，乾闥婆，阿修罗，迦楼罗，紧那罗，摩睺罗伽。他们是八类护法鬼神，随侍佛、菩萨、阿罗汉到世界各地说法，也会在人间护持修行人和寺院"}

注意，不要生成'''+f" {now}"+'''之前生成过的问题。

注意，你需要用{"Question":""", "Answer":"""}的格式，注意每个问题和回答内部不要换行。'''

text[4] = '''请以以下格式，生成10条关于金庸《神雕侠侣》的问题和回答。要求回答至少100字。

{"Question": "在《神雕侠侣》中，金轮寺的住持是谁？", "Answer": "在《神雕侠侣》中，金轮寺的住持是金轮法王，他是西藏密宗的高僧，武功高强，为金轮寺的第一高手。"}

注意，不要生成'''+f" {now}"+'''之前生成过的问题。

注意，你需要用{"Question":""", "Answer":"""}的格式，注意每个问题和回答内部不要换行。'''

text[5] = '''请以以下格式，生成10条关于金庸《射雕英雄传》的问题和回答。要求回答至少100字。

{"Question": "《射雕英雄传》中，杨康的生父是谁？", "Answer": "《射雕英雄传》中，杨康的生父是杨铁心。杨铁心在金庸武侠小说中是一个著名的悲剧人物，他因遭受陷害而家破人亡，最后在荒岛自杀。"}

注意，不要生成'''+f" {now}"+'''之前生成过的问题。

注意，你需要用{"Question":""", "Answer":"""}的格式，注意每个问题和回答内部不要换行。'''

text[6] = '''请以以下格式，生成10条关于金庸《天龙八部》的问题和回答。要求回答至少100字。

{"Question": "《天龙八部》中的爱情线索有哪些？", "Answer": "《天龙八部》中有许多爱情线索。其中包括乔峰和阿朱、段誉和王语嫣、虚竹和西夏公主的爱情纠葛以及其他角色之间的爱恨情仇。爱情线索为小说增添了浪漫和感人的色彩。"}

注意，不要生成'''+f" {now}"+'''之前生成过的问题。

注意，你需要用{"Question":""", "Answer":"""}的格式，注意每个问题和回答内部不要换行。'''

text[7] = '''请以以下格式，生成10条关于金庸《神雕侠侣》的问题和回答。要求回答至少100字。

{"Question": "《神雕侠侣》中杨过的武功是什么？", "Answer": "杨过所学武功博杂，涉猎古墓派武功、独孤求败的剑法、蛤蟆功、打狗棒法、弹指神通、玉箫剑法及一些九阴真经，最终自创黯然销魂掌。"}

注意，不要生成'''+f" {now}"+'''之前生成过的问题。

注意，你需要用{"Question":""", "Answer":"""}的格式，注意每个问题和回答内部不要换行。'''

text[8] = '''请以以下格式，生成10条关于金庸《射雕英雄传》的问题和回答。要求回答至少100字。

{"Question": "《射雕英雄传》中黄蓉的性格特点是什么？", "Answer": "黄蓉聪明机智，狡黠可爱，是一位具有极高智慧和武艺的女子。她性格独立，自尊心强，既有女性的柔情，又有江湖儿女的豪情。她对郭靖忠诚痴情，始终相伴左右，共同为家国大业奋斗。黄蓉的形象深入人心，被誉为金庸笔下最可爱的女子之一。"}

注意，不要生成'''+f" {now}"+'''之前生成过的问题。

注意，你需要用{"Question":""", "Answer":"""}的格式，注意每个问题和回答内部不要换行。'''

text[9] = '''请以以下格式，生成10条关于金庸《天龙八部》的问题和回答。要求回答至少100字。

{"Question": "《天龙八部》中的天山童姥和李秋水是什么关系？", "Answer": "天山童姥和李秋水是同门师姐妹，她们都是天山派的开山祖师，创立了天山折梅手和天山六阳掌等绝世武学。两人虽然师出同门，但性格迥异，最终走上了不同的道路。"}

注意，不要生成'''+f" {now}"+'''之前生成过的问题。

注意，你需要用{"Question":""", "Answer":"""}的格式，注意每个问题和回答内部不要换行。'''

text[10] = '''请以以下格式，生成10条关于金庸《神雕侠侣》的问题和回答。要求回答至少100字。

{"Question": "《神雕侠侣》中的“玉女心经”是谁所创？", "Answer": "玉女心经是由当年古墓派祖师林朝英独居古墓时创下。林朝英撰述玉女心经，虽是要克制全真派武功，但因其对王重阳始终情意不减，所以撰述到最后一篇玉女素心剑之时，林朝英幻想终有一日能与意中人并肩击敌，因之玉女心经的最后一章的武术特别地有转喻之意。"}

注意，不要生成'''+f" {now}"+'''之前生成过的问题。

注意，你需要用{"Question":""", "Answer":"""}的格式，注意每个问题和回答内部不要换行。'''

text[11] = '''请以以下格式，生成10条关于金庸《射雕英雄传》的问题和回答。要求回答至少100字。

{"Question": "《射雕英雄传》中郭靖在成吉思汗驾崩后做了什么事？", "Answer": "郭靖在成吉思汗驾崩后，帮助其子窝阔台稳定了蒙古帝国的政局，展示了他的智勇双全和忠诚担当。"}

注意，不要生成'''+f"{}+'''之前生成过的问答。

注意，你需要用{"Question":""", "Answer":"""}的格式，注意每个问题和回答内部不要换行。'''

```
count = 0
print(0)
while True:
    try:
        count+=1
        time1 = time.time()
        response = zhipuai.model_api.invoke(
            model="chatglm_pro",
            prompt=[
                {"role": "user", "content": text[count%12]},
            ],
            temperature=1,
            top_p=0.7,
            max_tokens=2000,
        )
        str = response['data']['choices'][0]['content']
        str = str.strip('\'')
        str = str.replace('\\n\\n', '\n')
        str = str.replace('\\n', '\n')
        str = str.replace('\\', '')
        str = str.strip()
        strlist = str.split('\n')

        now = time.time()
        if count%3==0: book="《天龙八部》"
        elif count%3==1: book="《神雕侠侣》"
        else: book="《射雕英雄传》"
        text[count%12] = '''请以以下格式，生成10条关于金庸'''+f"{{book}}"+'''的问题和回答。要求回答至少100字。
        ''' + strlist[0].strip() + '''
注意，不要生成'''+f"{}+'''之前生成过的问答。
注意，你需要用{"Question":""", "Answer":"""}的格式，注意每个问题和回答内部不要换行。'''

        for i in strlist:
            i = i.strip()
            with open('output1.jsonl', 'a', encoding="utf-8") as f:
                if i[-1]=='}':
                    if i[-2]!='\'': i[-2]='\''
                    f.write(i + '\n')
                    print("written", end=' ')
            print()

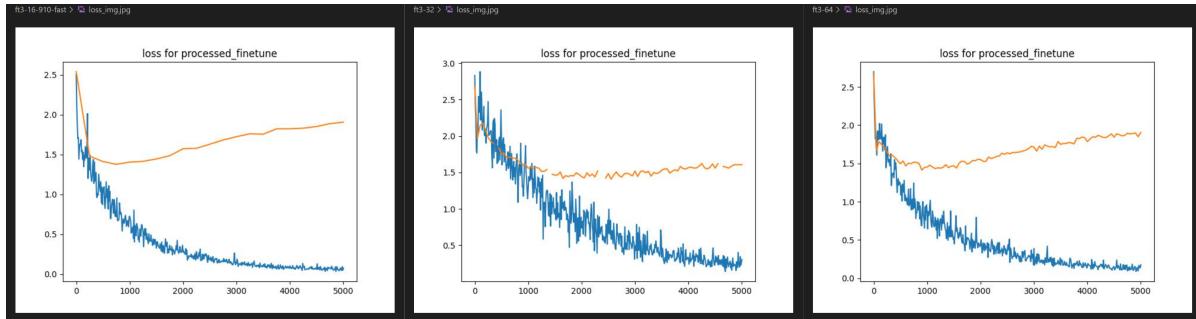
        time2 = time.time()
        print(count, f"Time used:{time2-time1}s")
    except:
        print(count, "Failed!")
```

同时我们还比对了不同的大模型生成问答的能力。可能由于训练语料问题，ChatGPT并不擅长生成金庸小说相关内容。中文大模型中，ChatGLM的能力又明显比文心一言要强很多，ChatGLM Pro又要比ChatGLM Std强一些。故它也被我们用为生成模型。

三、模型的微调

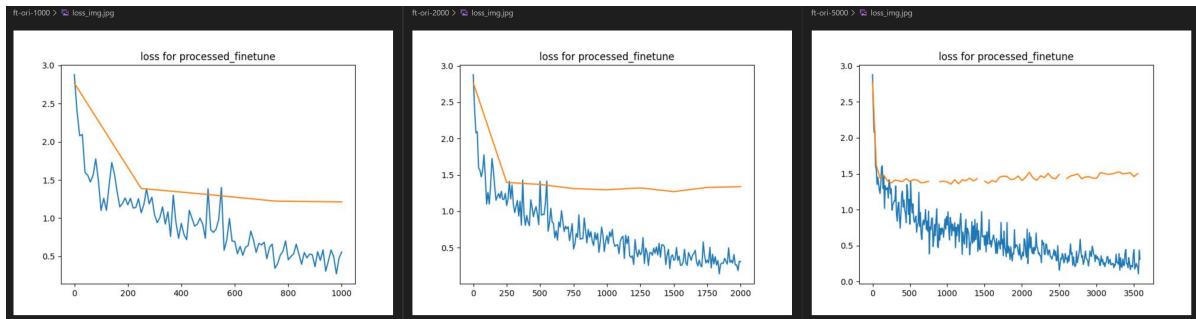
1. 小规模微调数据集微调效果

起初我并没有使用300000条的微调数据集，而是使用的1200条左右的小规模微调数据集，并根据block_size=910的预训练模型进行微调，采用不同的block_size和batch_size效果如下（三张图block_size和batch_size分别为910 * 16, 256 * 32, 256 * 64）：



可以看到，validation_loss都无法很好地收敛。而且在实际测试中，效果不佳。

我考虑到预训练和微调block_size不符可能导致模型适应出现问题，所以改使用block_size=256的预训练模型，效果如下（batch_size=32, max_iterations=1000/2000/3700）：

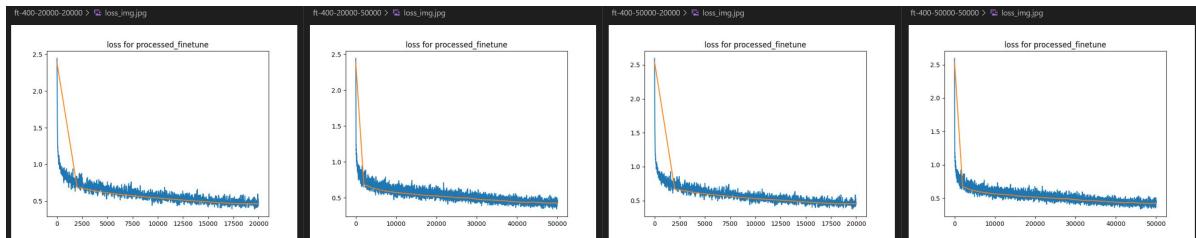


可以看到，validation_loss的收敛情况好了很多，黄线最终没有明显抬升，但较蓝线还有一定差距。然而在实际测试中，这些模型的效果都不佳。于是我才采用了更大规模的数据集（300000条）继续微调。

2. 大规模微调数据集的微调

（1）预训练模型基础

由于在预训练中我发现，虽然我的预训练一共有50000次迭代，但validation_loss最小时，迭代次数大概在20000左右。所以我尝试了不同的迭代次数的预训练模型进行微调。下面四张图分别展示了预训练迭代次数和微调迭代次数为20000+20000, 20000+50000, 50000+20000, 50000+50000的情况：

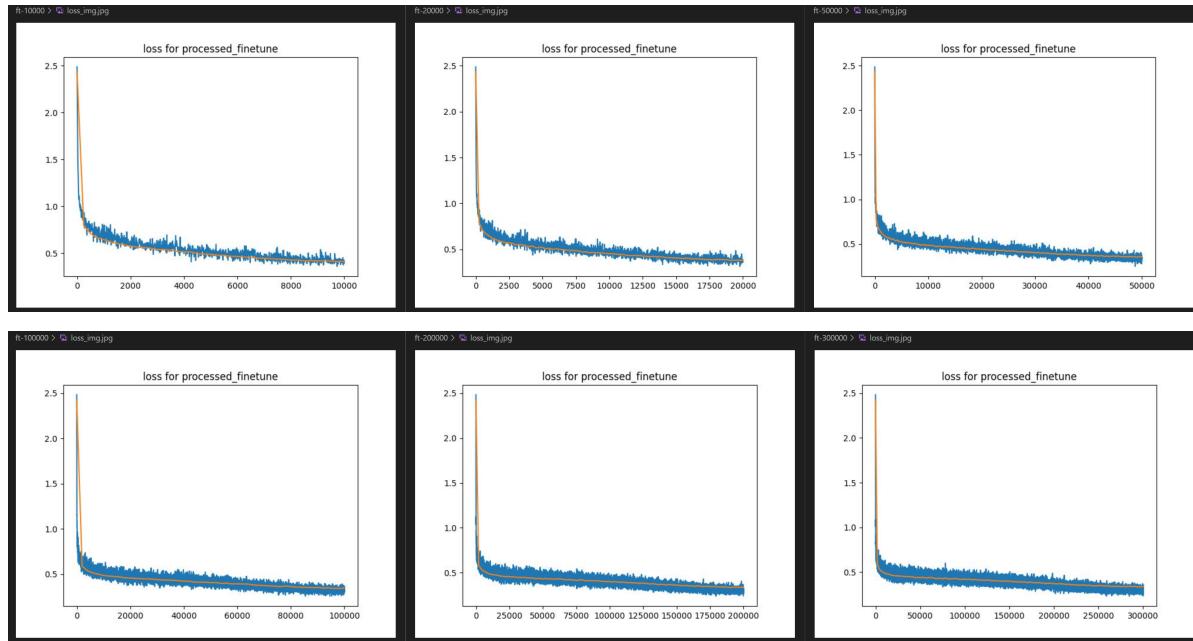


可以看到，与小规模微调数据相比，大规模微调数据的训练明显能让validation_loss更好地收敛。一直训练到50000条都没有出现黄线抬升的情况。

然而在实际比对中我发现，虽然预训练迭代次数大概在20000左右validation_loss最小，但是迭代次数20000预训练模型微调后的回答问题的能力明显不如迭代次数50000的模型。所以最后采用了max_iterations=50000的模型作为我的预训练模型，在此基础上尝试不同的微调方案。

(2) 微调迭代次数的影响

在微调模型时我发现，如果使用300000条的数据，不论迭代多少次，validation_loss的黄线几乎不会反弹。但是在实际测试中，并非迭代次数越多的模型效果越好。我尝试了10000、20000、50000、100000、200000、300000次迭代的微调效果，如下：



可见，黄线始终在下降，标志着模型应该没有过拟合到train_data上。然而在实际测试中，我发现不同的模型各有自己的优缺点。

10000次迭代的模型能力并不足够强。可能是因为10000次迭代不足以充分覆盖300000条的微调数据，经常会有莫名其妙的问答，如下：

天龙八部中谁最有智谋？《天龙八部》中最具智慧的人物是虚竹，他本是一个少林寺的和尚，因误打误撞成为少林寺的和尚，而被误认为他是一个少林寺的和尚；另一方面，他却被误认为是一个和尚，最终在少林寺中找到了真正的平安宁。

20000次的模型类似。

然而当迭代次数过大时，模型也会出现奇怪的输出。对于“介绍一下《天龙八部》的主角？”这个问题，50000次和100000次迭代的模型可以很好地给出回答：

介绍一下《天龙八部》的主角？《天龙八部》的主角是萧峰、虚竹、段誉、王语嫣、阿朱和阿紫，他们分别代表了天龙八部中的天、龙、夜叉三部，各自身负绝世武功，共同经历了许多传奇故事。

介绍一下《天龙八部》的主角？《天龙八部》的主角有三位，分别是段誉、虚竹和萧峰。他们分别代表了大理、少林和丐帮，是故事中的重要角色。

譬如在迭代次数为300000时，会有如下不符合小说真实内容的问答：

介绍一下《天龙八部》的主角？黄蓉是《天龙八部》的主角之一，她的主角之一。在《天龙八部》的故事中，黄蓉已经成为了丐帮帮主，并且在武林中享有盛名。她的聪明才智和武功，使得她在武林中有着极高的声望。

而200000次迭代的模型更是语无伦次，完全无法正常回答。

然而，在有些问题上，更高次迭代的模型展现出了优势，比如在询问“《射雕英雄传》中有哪些有名的比赛情节时？”时，50000次的模型虽然能大致理解“比赛情节”这个词语，但回答有些偏题：

讲一下射雕英雄传中有名的比赛情节？《射雕英雄传》中有许多比赛的情节，如郭靖和黄蓉的爱情故事、黄药师的桃花岛、桃花岛、黄药师的爱情等。这些情节都展示了江湖中的情节和人物性格特点，为故事增色不少。

100000的模型表现稍好：

讲一下射雕英雄传中有名的比赛情节？《射雕英雄传》中有许多比赛的情节，如郭靖在蒙古大漠与拖雷等。他们在江湖上各有建树，共同抵御外敌，守护家国。这些情节展示了江湖儿女的豪情壮志，也反映了他们在江湖中的地位。

然而300000的模型表现得非常出色，能够精确地给出“华山论剑”这一个情节，并做出详细的描述。

经过多重实际测试，我发现50000次迭代的模型在大多问题上的回答效果都尚可，综合能力最强。除上述两个问题外，模型对其他测试的回答如下：

小龙女和杨过的关系和故事？小龙女和杨过在《神雕侠侣》中有着深厚的感情。尽管他们曾因误会而分离，但最终还是化解误会，重新成为了朋友。他们的爱情故事充满了悲欢离合，感人至深。

神雕大侠是谁？神雕大侠是杨过的称号。他在江湖上以神雕大侠的名号闻名，是因为他饲养了一只神雕，并在神雕的帮助下，练就了无人能及的武功。他为了报答神雕的恩情，决定继承神雕大侠的名号，行侠仗义，为民除害。

介绍一下天龙八部的主要角色。《天龙八部》中的主要角色是阿朱，她是萧峰的恋人，也是一位武林中的女侠。她的性格特点是非常聪明，善良，但也有着深沉的父爱。她在故事中与萧峰、虚竹等人有着复杂的感情纠葛，最终她选择了与萧峰共度余生。

段誉的父亲是谁？段誉的父亲是段正淳，他是大理国的一位王子，也是一位武林高手。他因为误会而离开了大理，后来在江湖上四处游历，寻找自己的真实身份。

神雕大侠是谁？神雕大侠是杨过的称号。他在江湖上以神雕大侠的名号闻名，是因为他饲养了一只神雕，并在神雕的帮助下，练就了无人能及的武功。他为了报答神雕的恩情，决定继承神雕大侠的名号，行侠仗义，为民除害。

鸠摩智有哪些武功？鸠摩智是吐蕃国的国师，他精通多种武学，尤其擅长火焰刀和少林七十二绝技。他的武功高强，是江湖中一位传奇人物。

最终在验收时，我使用也是该模型。

四、实验感想

最大的体会就是：语言模型还是需要足够大的数据规模。当使上300000条的数据集时，训练效果一下子好了太多太多（某种程度上说，这也是现在AI“力大砖飞”的体现吧……）另外，在调整参数试图改善模型效果的过程中，我经常兜兜转转，尝试半天，最后发现给的默认参数效果最好。看来调试模型的经验还需要进一步积累。

训练中也遇到了一些波折，很长一段时间模型都只会蹦乱码。不过，在模型最后终于能“说人话”的时候，还是很有成就感的。