

昇思 MindSpore 模型开发挑战赛

【模型微调赛题第二阶段】-2024-年起科南-复赛文档

一.赛题介绍

本赛题要求基于中英文选择题数据集，跑通 baseline，并对 MindFormers 中 InternLM-7B 模型进行微调（LoRA 或其他微调算法）。微调后的模型在原有能力不丢失的前提下（需保持在原能力的 90%及以上），回答数学运算准确率相对 baseline 有提升，按照低参比例及准确率进行综合排名。

二.相关文件 obs 路径

- OBS 桶下目录结构

名称	存储类别	大小	最后修改时间	操作
ckpt	--	未统计	--	统计 复制路径 更多
config	--	未统计	--	统计 复制路径 更多
mmlu	--	未统计	--	统计 复制路径 更多
r4_a16_3epoch_v2	--	未统计	--	统计 复制路径 更多
cmmlu-csv2json.py	标准存储	2.11 KB	2024/10/31 11:45:18 GMT+08:00	下载 复制路径 更多
eval_squad.log	标准存储	6.20 MB	2024/10/31 11:29:03 GMT+08:00	下载 复制路径 更多
mindformers.zip	标准存储	80.57 MB	2024/10/31 11:20:41 GMT+08:00	下载 复制路径 更多
test1_mmlu_alpaca_format.json	标准存储	1.36 MB	2024/10/30 20:21:49 GMT+08:00	下载 复制路径 更多
train-2.log	标准存储	18.76 MB	2024/10/31 08:43:35 GMT+08:00	下载 复制路径 更多

- 项目代码

MindFormers 及项目代码

<https://llm-gm-sub.obs.cn-southwest-2.myhuaweicloud.com/mindformers.zip>

解押后遵循 tutorial 中进行项目环境配置

- 配置文件:

主要修改为 lora_rank=4 和 lora_alpha=16

Plain Text

```
# 验证原有能力
https://llm-qm-sub.obs.cn-southwest-2.myhuaweicloud.com/config/eval.yaml

# 预测
https://llm-qm-sub.obs.cn-southwest-2.myhuaweicloud.com/config/predict_r4.yaml









# 训练
https://llm-qm-sub.obs.cn-southwest-2.myhuaweicloud.com/config/train_r4.yaml
```

权重

```
# 最终权重(lora_rank=4)
https://llm-qm-sub.obs.cn-southwest-2.myhuaweicloud.com/ckpt/r4_a16_5epoch_v2.ckpt

# 训练中间权重目录
https://llm-qm-sub.obs.cn-southwest-2.myhuaweicloud.com/r4_a16_3epoch_v2/<中间权重名>

# 权重名见下图，其中保存了前 3 个 epoch 和后两个 epoch 的部分中间权重
```

<input type="checkbox"/>	 internlm_7b_lora_rank_0-42075_2.ckpt	标准存储	13.68 GB	2024/10/31 08:53:04 GMT+08:00	下载 复制路径 更多 ▾
<input type="checkbox"/>	 internlm_7b_lora_rank_0-42500_2.ckpt	标准存储	13.68 GB	2024/10/31 08:53:04 GMT+08:00	下载 复制路径 更多 ▾
<input type="checkbox"/>	 internlm_7b_lora_rank_0-42795_2.ckpt	标准存储	13.68 GB	2024/10/31 08:53:05 GMT+08:00	下载 复制路径 更多 ▾
<input type="checkbox"/>	 internlm_7b_lora_rank_0-62900_2.ckpt	标准存储	13.68 GB	2024/10/31 08:53:05 GMT+08:00	下载 复制路径 更多 ▾
<input type="checkbox"/>	 internlm_7b_lora_rank_0-63325_2.ckpt	标准存储	13.68 GB	2024/10/31 08:53:05 GMT+08:00	下载 复制路径 更多 ▾
<input type="checkbox"/>	 internlm_7b_lora_rank_0-63750_2.ckpt	标准存储	13.68 GB	2024/10/31 08:56:20 GMT+08:00	下载 复制路径 更多 ▾
<input type="checkbox"/>	 internlm_7b_lora_rank_0-64175_2.ckpt	标准存储	13.68 GB	2024/10/31 08:56:24 GMT+08:00	下载 复制路径 更多 ▾
<input type="checkbox"/>	 internlm_7b_lora_rank_0-64192_2.ckpt	标准存储	13.68 GB	2024/10/31 08:56:24 GMT+08:00	下载 复制路径 更多 ▾

日志

Plain Text

```
# 微调日志
https://llm-qm-sub.obs.cn-southwest-2.myhuaweicloud.com/train-2.log

# 原有能力验证日志
https://llm-qm-sub.obs.cn-southwest-2.myhuaweicloud.com/eval_squad.log
```

三.数据处理及分析

数据文件操作: 在原有数据集的基础上, 添加一定比例从原有数据中抽取的部分类别, 将补充数据的题干修改为如下内容(避免直接重复复制):

```
the correct answer is one of the options A/B/C/D. Please use the knowledge about {domain} to select the correct option and answer the question with 'The right option is'.",
+ str(item['A']) + '\nB.' + str(item['B']) + '\nC.' + str(item['C']) + '\nD.' + str(item['D']),
```

```
SQL
... "Please use the knowledge about {domain} to select the correct
option" ...
```

相关脚本: <https://llm-qm-sub.obs.cn-southwest-2.myhuaweicloud.com/cmmlu-csv2json.py>

修改后数据 mindrecord 路径:

https://llm-qm-sub.obs.cn-southwest-2.myhuaweicloud.com/mmlu/r4_supply.mindrecord

https://llm-qm-sub.obs.cn-southwest-2.myhuaweicloud.com/mmlu/r4_supply.mindrecord.db

四. 模型微调方法与记录

本次微调采用 lora 方式在单机进行。

下面记录中 rxx_axx_nepoch: 表示 lora_rank, lora_alpha 以及 epoch 的不同超参数调试

模型训练记录	参数量	准确率	得分	低参比例得分	准确率得分		备注
r128_a64_4epoch	134217728	97.76%	0.929320	0.245000	0.68432	>100%	
r64_a64_4epoch	67108864	98.06%	0.958920	0.272500	0.68642	>100%	
r32_a64_5epoch	33554432	97.42%	0.968190	0.286250	0.68194	98.55%	
r32_a64_6epoch		97.76%					r32 的训练第 6 个

							epoch 仍有提升
r4_a16_5 epoch	4194304	96.83%	0.976091	0.298281	0.67781		
r4_a16_6 epoch		96.32%					r4 原有数据第六个 epoch 有所下降
r4_a16_6 epoch		96.41%					使用补充数据准确率有一点上升
r4_a16_3 epoch		93.11%					
r4_a8_3epoch		90.58%					
r4_a16_5 epoch_v2		96.92%					针对我们的测试集有所上升且综合低参比例最佳

最终提交 r4_a16_5epoch_v2 版本。

五. 低参量

4194304

```
2024-10-29-11:33:22,851--mindformers[mindformers/trainer/base_trainer.py:543]--INFO--Network Parameters: 4194304.
2024-10-29-11:33:22,853--mindformers[mindformers/trainer/base_trainer.py:678]--INFO--Build Optimizer For Train.....
2024-10-29-11:33:22,853--mindformers[mindformers/trainer/base_trainer.py:426]--INFO--Build Optimizer From Config.....
2024-10-29-11:33:22,853--mindformers[mindformers/trainer/base_trainer.py:459]--INFO--Build LR Schedule From Config.....
2024-10-29-11:33:22,862--mindformers[mindformers/trainer/optimizer_grouped_parameters.py:74]--WARNING--dynamic_lr_schedule will be reset
2024-10-29-11:33:22,867--mindformers[mindformers/trainer/optimizer_grouped_parameters.py:113]--INFO--Param groups:={
```

六. 运行脚本

需要更换配置文件, 权重文件, tokenizer.model 等文件路径

```
SQL
# 微调训练
cd /home/ma-user/work/mindformers/
bash scripts/msrun_launcher.sh "python
research/internlm/run_internlm.py --run_mode finetune --
use_parallel False --config
```

```
~/work/mindformers/research/internlm/train_r4.yaml --  
load_checkpoint ~/work/internlm.ckpt --auto_trans_ckpt False --  
train_dataset /home/ma-user/work/mmlu/r4_supply.mindrecord" 1
```

原有能力验证脚本

```
cd /home/ma-user/work/mindformers/research/internlm python  
run_internlm.py \  
--config eval.yaml \  
--run_mode eval \  
--load_checkpoint ~/work/tmp.ckpt \  
--use_parallel False \  
--eval_dataset /home/ma-user/work/squad8192.mindrecord >  
eval_squad.log 2>&1 &
```

推理最终题目

```
cd /home/ma-user/work/mindformers/research/internlm python  
run_internlm.py \  
--config ~/work/mindformers/research/internlm/predict_r4.yaml \  
--run_mode predict \  
--use_parallel false \  
--load_checkpoint ~/work/tmp.ckpt \  
--auto_trans_ckpt false \  
--input_dir ~/work/test1_mmlu_alpaca_format.json > predict2000.log  
2>&1 &
```