华为昇腾平台模型迁移测试

华为平台

已测试的模型	推理	微调	改动量
Cerebrum-1.0-7b	√		极少
c4ai-command-r-v01	x		模型太大，需要单独做分配。
chatglm3-6b	√		极少
deepseek-vl-1.3b-base	√		极少
gpt2	√		极少
Mistral-7B-Instruct-v0.2	√	√	极少
moondream2	√		极少

llama3	√	√	极少（通过mindspore跑通）
superprompt-v1	√		极少
tinyllama 1.1B	√		极少
Yi-9B-200K	√		极少

模型训练后迁移回N卡平台

测试过将Mistral模型在华为平台上进行sft训练后，可以直接将模型权重迁移回A100上并且成功跑通了推理。

迁移过程如下

获取模型https://modelscope.cn/models/AI-ModelScope/Mistral-7B-Instruct-v0.2/summary

获取llama factory库https://github.com/hiyouga/LLaMA-Factory

修改好配置文件ds_config_zero2.json

```
{
 "train_batch_size": "auto",
 "train_micro_batch_size_per_gpu": "auto",
 "gradient_accumulation_steps": "auto",
 "gradient_clipping": "auto",
 "zero_allow_untested_optimizer": true,
 "fp16": {
 "enabled": "auto",
 "loss_scale": 0,
 "loss_scale_window": 1000,
 "initial_scale_power": 16,
```

```
 "hysteresis": 2,
 "min_loss_scale": 1
 },
 "bf16": {
 "enabled": "auto"
 },
 "zero_optimization": {
 "stage": 2,
 "allgather_partitions": true,
 "allgather_bucket_size": 5e8,
 "overlap_comm": true,
 "reduce_scatter": true,
 "reduce_bucket_size": 5e8,
 "contiguous_gradients": true,
 "round_robin_gradients": true
 }
}
```

进行训练

```
deepspeed --num_gpus 8 src/train_bash.py \
 --ddp_timeout 180000000 \
 --deepspeed /root/build_env/deepspeed_config/ds_config_zero2.json \
 --stage sft \
 --do_train \
 --model_name_or_path /root/build_env/AI-ModelScope/Mistral-7B-Instruct-v02 \
 --dataset alpaca_gpt4_en \
 --template default \
 --finetuning_type full \
 --lora_target q_proj,v_proj \
```

--output_dir path_to_sft_checkpoint \

--overwrite_cache \

--per_device_train_batch_size 4 \

--gradient_accumulation_steps 4 \

--lr_scheduler_type cosine \

--logging_steps 10 \

--save_steps 1000 \

--learning_rate 5e-5 \

--num_train_epochs 3.0 \

--plot_loss \

--fp16

训练完成后找到保存的模型权重将其传到A100服务器上

直接采用推理的代码即可跑通模型推理

遇到的问题

遇到的一个典型问题

TypeError: new() received an invalid combination of arguments - got (Tensor, requires_grad=bool), but expected one of:

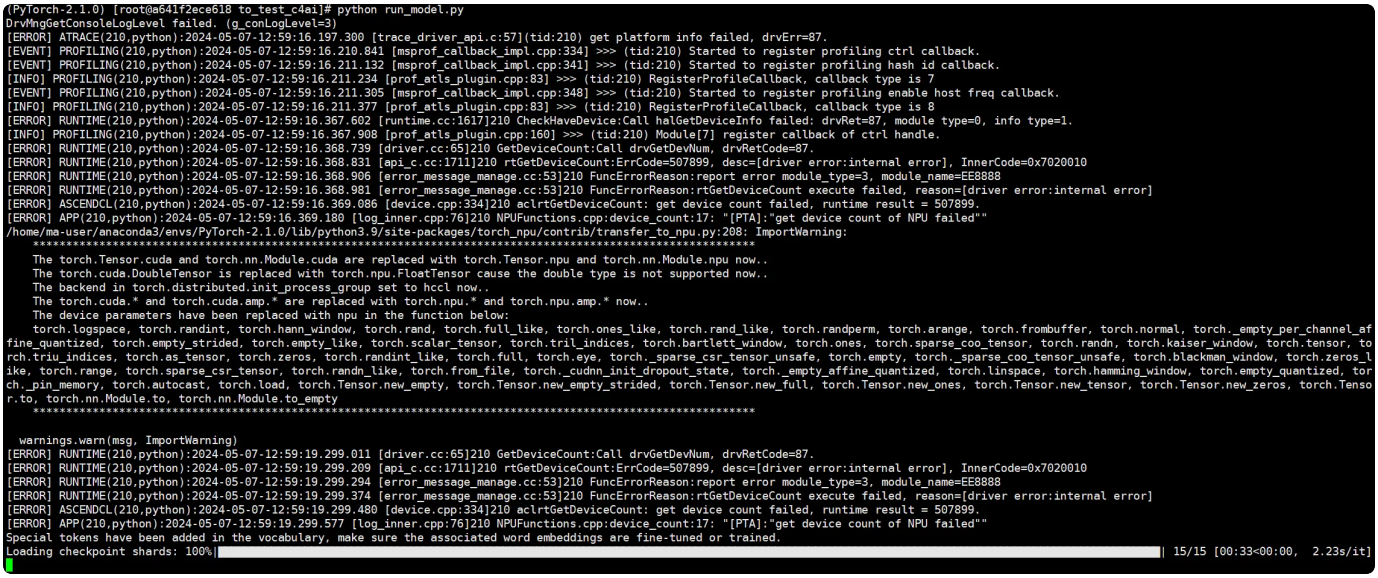
- (*, torch.device device)

didn't match because some of the keywords were incorrect: requires_grad

- (torch.Storage storage)
- (Tensor other)
- (tuple of ints size, *, torch.device device)
- (object data, *, torch.device device)

在对模型进行微调训练，以及TinyLlama进行推理时都遇到过（至少遇到过4，5次）。只要用到了accelerate库就很容易报这个错。判断大概率是平台对accelerate库支持不好导致，也只在华为社区看到这个issue的讨论。

跑c4ai推理时在导入checkpoint时卡住



此原因是因为模型权重太大，无法单卡推理。多卡推理需要单独设置切分策略。

跑完Mistral-7B-Instruct的lora微调后，merge权重时报错

TypeError: new() received an invalid combination of arguments - got (Tensor, requires_grad=bool), but expected one of:

- (*, torch.device device)

didn't match because some of the keywords were incorrect: requires_grad

- (torch.Storage storage)
- (Tensor other)
- (tuple of ints size, *, torch.device device)
- (object data, *, torch.device device)

accelerate相关库支持不到位导致。

训练llama3中遇到的问题

主要参考这篇教程实现
https://gitee.com/mindspore/mindformers/blob/dev/research/llama3/llama3.md#https://gitee.com/link?target=https%3A%2F%2Fhuggingface.co%2Fmeta-llama%2FMeta-Llama-3-8B

在权重转换以及训练步骤都遇到此报错：

ImportError: cannot import name 'swap_cache' from 'mindspore._c_expression'

应该是由于代码版本升级导致的bug，直接注释掉源码中的导入语句后解决（经测试并不影响后续的模型训练和推理）

可以参考此issue https://gitee.com/mindspore/mindformers/issues/I9NL4D

对于模型推理速度过慢的解答：

因为训练的权重用的bf16，而推理目前只支持f16，因此在进行推理前需要进行权重的转换。假如训练也用f16就不需要转换。同时，推理时将权重转换好后在多轮对话式推理速度正常。