**LORA**

**“低秩适应”（LoRA）是一种“参数高效微调”（PEFT）的形式，它允许使用少量可学习参数对大型模型进行微调** 。LoRA改善微调的几个点：

1. 将微调视为学习参数的变化(▲w)，而不是调整参数本身(w)。
2. 通过删除重复信息，将这些变化压缩成较小的表示。
3. 通过简单地将它们添加到预训练参数中来“加载”新的变化。

ChatGLM微调，使用ChatGLM-Finetuning-master项目

### Freeze 方法

即参数冻结，对原始模型部分参数进行冻结操作，仅训练部分参数。

### PT 方法

即 P-Tuning 方法，参考 ChatGLM 官方代码，是一种针对于大模型的 soft-prompt 方法。

### Lora 方法

即在大型语言模型上对指定参数增加额外的低秩矩阵，并在模型训练过程中，仅训练而外增加的参数。

git clone <https://github.com/liucongg/ChatGLM-Finetuning.git>

cd ChatGLM-Finetuning

pip install -r requirements.txt

数据准备（医疗领域

covid-数据集

ChatGLM单卡训练

CUDA\_VISIBLE\_DEVICES=0 deepspeed --master\_port 520 train.py \

--train\_path "/home/devin/project/2205A-CV/zmq/covid-数据集/train.json" \

--model\_name\_or\_path "/home/devin/project/2205A-CV/zmq/my\_Linly-Talker2/chatglm-6b-int4" \

--per\_device\_train\_batch\_size 1 \

--max\_len 1560 \

--max\_src\_len 1024 \

--learning\_rate 1e-4 \

--weight\_decay 0.1 \

--num\_train\_epochs 2 \

--gradient\_accumulation\_steps 4 \

--warmup\_ratio 0.1 \

--mode glm \

--train\_type lora \

--lora\_dim 16 \

--lora\_alpha 64 \

--lora\_dropout 0.1 \

--lora\_module\_name "query\_key\_value" \

--seed 1234 \

--ds\_file ds\_zero2\_no\_offload.json \

--gradient\_checkpointing \

--show\_loss\_step 10 \

--output\_dir ./output-glm

服务器驱动掉了未运行成功