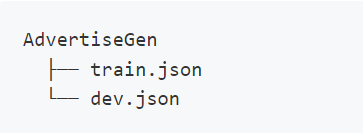
# ChatGLM2微调

1. cp /user/config/nbstart\_hccl.json /home/ma-user/work/mindformers 将生成RANK\_TABLE\_FILE的json复制粘贴在mindformers文件下。
2. wget <https://ascend-repo-modelzoo.obs.cn-east-2.myhuaweicloud.com/XFormer_for_mindspore/glm2/glm2_6b.ckpt> 下载使用MindFormers提供的已转换权重。
3. wget <https://drive.google.com/file/d/13_vf0xRTQsyneRKdD1bZIr93vBGOczrk/view?usp=sharing> 下载处理好的ADGEN数据集，目录结构为：



1. wget [https://ascend-repo-modelzoo.obs.cn-east-2.myhuaweicloud.com/XFormer\_for\_mindspore/glm2/tokenizer.model 下载](https://ascend-repo-modelzoo.obs.cn-east-2.myhuaweicloud.com/XFormer_for_mindspore/glm2/tokenizer.model%20%20下载)文件tokenizer.model
2. 修改configs/glm2/run\_glm2\_6b.yaml中*===dataset config===*为

（1）dataset\_dir: "/home/ma-user/work/mindformers/AdvertiseGen/train.json"

（2）eval\_dataset中dataset\_dir的路径dataset\_dir: "/home/ma-user/work/mindformers/AdvertiseGen/dev.json"

（3）vocab\_file: "/home/ma-user/work/mindformers/tokenizer.model"

6. pip install –r requirments.txt 安装运行支持的库

7. 单机多卡微调：

cd scripts

bash run\_distribute.sh ./../jobstart\_hccl.json ./../configs/glm2/run\_glm2\_6b.yaml [0,8] finetune

打开日志

tail -f ./output/log/rank\_0/info.log

8. 训练结束后，合并权重。参考llama分布式权重合并文档。

9. 推理：

提供一个模型推理样例脚本infer.py

from mindformers import AutoConfig, AutoModel, AutoTokenizer

import mindspore as ms

ms.set\_context(mode=ms.GRAPH\_MODE, device\_target="Ascend", device\_id=6)

config = AutoConfig.from\_pretrained("glm2\_6b")

# 可以在此使用下行代码指定自定义权重进行推理，默认使用自动从obs上下载的预训练权重

# config.checkpoint\_name\_or\_path = "/path/to/glm2\_6b\_finetune.ckpt"

config.use\_past = True

model = AutoModel.from\_config(config)

tokenizer = AutoTokenizer.from\_pretrained("glm2\_6b")

while True:

user\_question=input("输入:")

print(type(user\_question))

# print(f'{user\_questions}')

if user\_question=="退出":

print('结束')

break

inputs = tokenizer(tokenizer.build\_prompt(f'{user\_question}'))["input\_ids"]

outputs = model.generate(inputs, max\_length=128)

print(tokenizer.decode(outputs))

进行单卡推理：

运行 python infer.py

即可实现对话。