此版本是基于horovod的tensorflow1.14版本，可能需要配置nccl等，需要查看是否可以调用nvcc

需要下载**BERT-wwm-ext, Chinese** [**https://github.com/ymcui/Chinese-BERT-wwm**](https://github.com/ymcui/Chinese-BERT-wwm)

**放在目录./code/horovod\_bert/pretrain\_model下**

建立镜像bash build.sh

docker run --it /local/path:/container/path horovod/horovod:0.16.4-tf1.14.0-torch1.1.0-mxnet1.4.1-py3.6 bash 进入到容器，/local/path是本地路径，/container/path是容器内部路径

环境需要设置下，安装vim， 然后vim ~/.vimrc 设置以下格式可以看中文了

set encoding=utf-8

set fileencodings=ucs-bom,utf-8,cp936

set fileencoding=gb2312

set termencoding=utf-8

实验分为以下几个：

1. Python extract\_data.py 提取gaokao.txt里的数据并进行处理，得到"train\_data.json"和"eval\_data.json"， sample\_text.txt 为提取的answer的内容，因为question没法做next sentence的操作，bert预训练是两个模式
2. 基于预训练，预训练跑以下代码：

基于gaokao.txt生成预训练数据 bash create\_pretrain\_data.sh 生成tf\_record格式的预训练数据

1. Bash run\_pretrain.sh 开始预训练，预训练得到的模型存储在 --output\_dir=/tmp/pretraining\_output
2. 训练bash train.sh

增加了几个参数

Shots （1 的话每个类只有1个样本参与训练）

Loss\_class（normal 和 focal\_loss）

这里可以做各种实验尝试

1. 训练next sentence， 基于规则是正确qa为正例，根据q随机选择a为负例做二分类

得到新的模型

运行 bash test.sh

1. 基于5步骤得到的bert训练模型，来训练5分类的qa历史集合，修改部分代码

Bash train\_1.sh