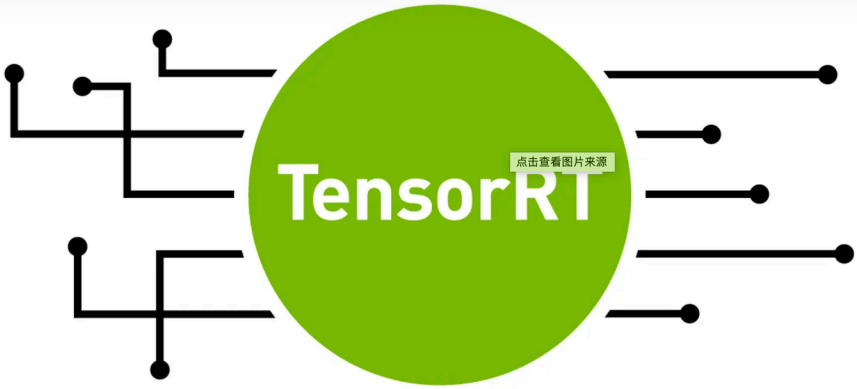


模型部署-TensorRT笔记-1.TensorRT安装教程

老苏聊AI
做一名会摄影的NLPer
关注他

1、背景

由于最近线上相关业务的需要，需要将Bert模型应用到搜索场景，主要的通过文本的语义相似度，来召回语义上相近的文本，这里面遇到的一个非常棘手的问题就是耗时的问题，需要将NLP的BERT模型部署到线上，所以最近专门研究了一下模型部署+的相关问题，所以在这里做一个专门的记录

2、 目的

本小节，我们先来说一下TensorRT的安装

3、环境介绍

3-1、基本环境

这里物理机的环境为

- Ubuntu 20.04.3 LTS
- cuda，11.4.3
- cudnn，8.2.4
- python，3.7

我们针对这个部署的环境，使用docker环境

具体的docker从下面这个地址下载

registry.hub.docker.com...

下载的docker具体的版本是

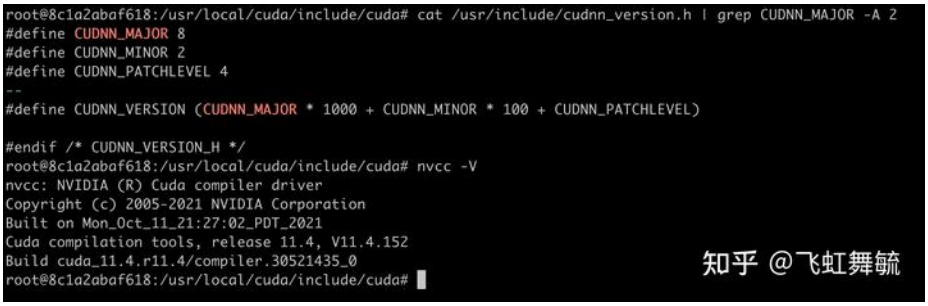
```
docker pull nvidia/cuda:11.4.3-cudnn8-devel-ubuntu20.04
```

启动docker之后，可以查看相关的版本

查看cudnn的版本

```
cat /usr/include/cudnn_version.h | grep CUDNN_MAJOR -A 2
```

```
nvcc -V
```



CUDA和Cudnn版本

3-2、TensorRT下载

developer.nvidia.com/nv...

具体的版本我们选择，8.2-GA版本

下载文件如下

下载相关文件

TensorRT-8.2.1.8.Linux.x86_64-gnu.cuda-11.4.cudnn8.2.tar.gz

将相关文件上传到服务器解压

```
tar -zvxf TensorRT-8.2.1.8.Linux.x86_64-gnu.cuda-11.4.cudnn8.2.tar.gz
```

解压之后得到相关的文件夹，TensorRT-8.2.1.8

3-3、环境变量+设置

```
vim ~/.bashrc

export LIBRARY_PATH=/xxxx/TensorRT-8.2.1.8/lib:$LIBRARY_PATH
export CUDA_HOME=/usr/local/cuda
export PATH=$CUDA_HOME/bin:$PATH
export LD_LIBRARY_PATH=/xxxxxx/TensorRT-8.2.1.8/lib:$CUDA_HOME/lib64:$LD_LIBRARY_PATH
```

3-4、组件安装

3-4-1、python组件安装

TensortRT，有C++和Python两个接口，如果python需要使用，需要安装相关的组件

一共有4个文件，需要按装python的版本安装

```
TensorRT-8.2.1.8/python
python/tensorrt-8.2.1.8-cp36-none-linux_x86_64.whl
python/tensorrt-8.2.1.8-cp37-none-linux_x86_64.whl
python/tensorrt-8.2.1.8-cp38-none-linux_x86_64.whl
python/tensorrt-8.2.1.8-cp39-none-linux_x86_64.whl

python -m pip install python/tensorrt-8.2.1.8-cp37-none-linux_x86_64.whl

In [1]: import tensorrt

In [2]: tensorrt.__version__
Out[2]: '8.2.1.8'
```

3-4-2、uff组件安装

安装uff组件，需要tensorflow，这里安装2.4.0版本

```
pip install tensorflow-gpu==2.4.0 -i https://mirror.baidu.com/pypi/simple

python -m pip install uff/uff-0.6.9-py2.py3-none-any.whl

# 验证
In [1]: import uff
2022-04-20 21:24:48.454240: I tensorflow/stream_executor/platform/default/dso_loader.

In [2]: uff.__version__
Out[2]: '0.6.9'

In [3]:
```

3-4-3、graphsurgeon组件安装

```
python -m pip install graphsurgeon/graphsurgeon-0.4.5-py2.py3-none-any.whl
```

3-4-3、onnx_graphsurgeon组件安装

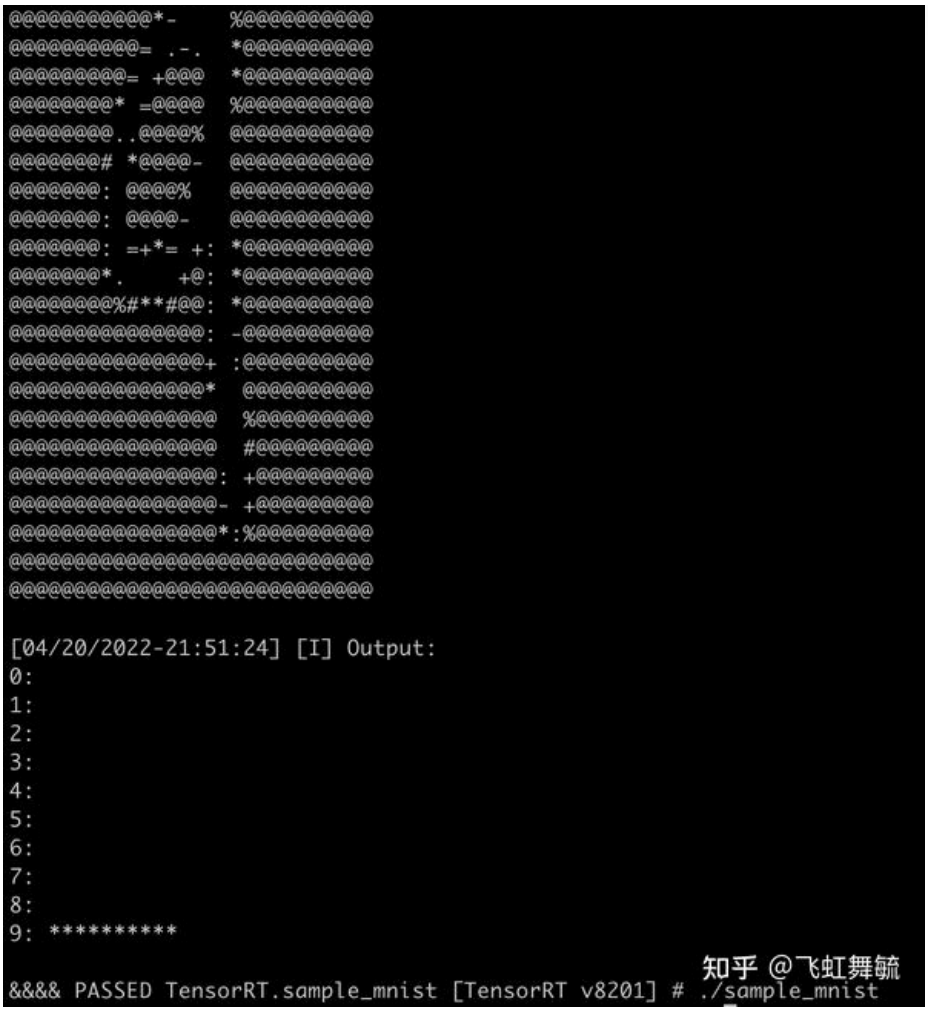
```
python -m pip install onnx_graphsurgeon/onnx_graphsurgeon-0.3.12-py2.py3-none-any.whl
```

4、测试

可以进入相关的目录进行测试

```
cd TensorRT-8.2.1.8/samples/sampleMNIST
make

cd TensorRT-8.2.1.8/bin
./sample_mnist
```



结果测试

5、总结

TensorRT的相关安装工作到此已经结束，下一节我们在聊一下onnx的安装，这里有几个点，首先我们这里在docker里来安装相关的环境，我们也建议大家使用docker，因为这个容易确保，模型训练和推理时，相关的环境保持一致

编辑于 2022-04-20 21:54

内容所属专栏

NLP模型部署
有关模型部署的相关的工作
订阅专栏

TensorFlow 学习 TensorRT 模型

理性发言，友善互动

1 条评论 默认 最新

CUICUI
写的很详细
2022-04-21 回复 1