TRT模型推理加速：

- 简介：在tensorflow2.0上使用tensorrt6加速
- 工作流程：

运行环境准备

- OS：Ubuntu18.0.4 LST
- python：3.6.5
- tensorflow：2.1.0
- tensorrt：6.0.1
- GPU：只支持GPU

执行步骤

- 安装Nvidia深度学习驱动，Cudnn和Cuda驱动：此部分从略。
- 下载安装TensorRT 6.0

保存模型

- save_models形式

转换成tensorrt图

```
from tensorflow.python.compiler.tensorrt import trt_convert as trt
params=trt.DEFAULT_TRT_CONVERSION_PARAMS
params._replace(precision_mode=trt.TrtPrecisionMode.FP32)
converter = trt.TrtGraphConverterV2(input_saved_model_dir='tf_savedmodel',conversion_params=params)
converter.convert()#完成转换,但是此时没有进行优化,优化在执行推理时完成
converter.save('trt_savedmodel')
```

优化并推理

```
import tensorflow as tf
from tensorflow.python.compiler.tensorrt import trt_convert as trt
from tensorflow.keras.datasets import mnist
import time

(x_train, y_train), (x_test, y_test) = mnist.load_data()
x_test = x_test.astype('float32')
x_test = x_test.reshape(10000, 784)
x_test /= 255

saved_model_loaded = tf.saved_model.load(
    "trt_savedmodel", tags=[trt.tag_constants.SERVING])#读取模型
graph_func = saved_model_loaded.signatures[
    trt.signature_constants.DEFAULT_SERVING_SIGNATURE_DEF_KEY]#获取推理函数,也可以使用saved_model_loaded.signatures['serving_default']
frozen_func = trt.convert_to_constants.convert_variables_to_constants_v2(
    graph_func)#将模型中的变量变成常量,这一步可以省略,直接调用graph_func也行

t=time.time()
output = frozen_func(tf.constant(x_test))[0].numpy()
print(time.time()-t)
print((output.argmax(-1)==y_test).mean())
```

可能遇到的错误

> 转换模型时python直接退出 查看tensorrt的调试输出，可能会看到名义上一个CUDA的一些库无法找到libcublas.so.10，这可能是因为CUDA版本不符合tensorrt6所需要的版本，更改CUDA版本之前推荐尝试我的方法： 我安装的是cuda10.0，CUDA / lib64下里面有libcublas.so.10.0而没有tensorrt所需要的libcublas.so.10，将libcublas.so.10.0复制为libcublas.so.10 对其他调试输出也进行类似修改即可。 突然莫名其妙打不开tensorrt 试试 import pycuda.autoinit 会不会报错,如果报错,我的解决方案是重启.

参考

- 在tensorflow2.0上使用tensorrt6加速：https://blog.csdn.net/zt1091574181/article/details/100972562
- TensorRT6.0.1安装指南：https://blog.csdn.net/zt1091574181/article/details/100972562
- TRT6下载地址：https://developer.nvidia.com/nvidia-tensorrt-6x-download