

赞同 1

分享

模型部署-TensorRT笔记-2.ONNX-TensorRT安装教程



老苏聊AI

做一名会摄影的NLPer

关注他

1 人赞同了该文章

1、背景

上一节我们安装了TensorRT，这一节我们来记录一下onnx到tensorrt的相关环境

因为我们模型的优化路线是，Pytorch-onnx-TensorRT或者是Tensorflow-onnx-TensorRT

所以我们需要安装onnx到tensorRT的相关环境

2、目的

本小节，我们安装一下onnx到tensorRT转化的相关环境

3、环境安装

- gcc, 8.4.0
- g++, 8.4.0
- [cmake](#)⁺, 3.16.3
- protobuf, 3.9.2
- onnx, 1.8.0
- [pycuda](#)⁺, 2021.1
- tensorrt, 8.2.1.8

3-1、相关依赖安装

安装相关的依赖包

```
apt-get+ install gcc-8 g++-8
apt install cmake
apt install make
apt-get install libprotobuf-dev protobuf-compiler
# pycuda version 2021.1
pip install pycuda -i https://mirror.baidu.com/pypi/simple
```

3-2、protobuf安装

下载的地址如下

我们选择的版本是，3.9.2

protobuf-all-3.9.2.tar.gz

```
# 解压文件
tar -zxvf+ protobuf-all-3.9.2.tar.gz

cd protobuf-3.9.2/
./autogen.sh
./configure

# 编译，安装
make -j 12
make -j 12 check
make install

# 刷新共享库
ldconfig

# 验证版本
protoc --version
libprotoc 3.9.2
```

3-3、onnx-tensorrt安装

具体下载地址

[github.com/onnx/onnx-te...](https://github.com/onnx/onnx-tensorrt)

这里我们不要选择master分支，选择8.2-GA分支

```
git clone https://github.com/NVIDIA/TensorRT.git
git checkout -b 8.2-GA origin/8.2-GA

git submodule update --init --recursive
```

修改CMakeLists.txt文件

添加，6，7，8，9行

```
# SPDX-License-Identifier: Apache-2.0

cmake_minimum_required+(VERSION 3.13)
project(onnx2trt LANGUAGES CXX C)

include_directories(/xxx/TensorRT-8.2.1.8/include)
include_directories(/usr/local/cuda/include)
link_directories(/xxx/TensorRT-8.2.1.8/lib)
set(TENSORRT_ROOT /xxx/TensorRT-8.2.1.8)
set(ONNX2TRT_ROOT ${PROJECT_SOURCE_DIR})
```

接下来开始编译，安装

```
mkdir build && cd build

cmake .. -DProtobuf_PROTOC_EXECUTABLE=/usr/local/bin/protoc -DProtobuf_INCLUDE_DIRS=/x

make -j12
```

安装成功显示如下

```
root@8c1a2abaf618:/tensorrt_work/github/onnx-tensorrt/build# make install
[ 2%] Built target gen_onnx_proto
[ 13%] Built target onnx_proto
[ 27%] Built target nvonnxparser_static
[ 31%] Built target getSupportedAPITest
[ 83%] Built target onnx
[ 86%] Built target onnx2trt
[100%] Built target nvonnxparser
Install the project...
-- Install configuration: "Release"
-- Installing: /usr/local/lib/libnvonnxparser.so.8.2.1
-- Installing: /usr/local/lib/libnvonnxparser.so.8
-- Set runtime path of "/usr/local/lib/libnvonnxparser.so.8.2.1" to ""
-- Installing: /usr/local/lib/libnvonnxparser.so
-- Installing: /usr/local/lib/libnvonnxparser_static.a
-- Installing: /usr/local/bin/onnx2trt
-- Set runtime path of "/usr/local/bin/onnx2trt" to ""
```

知乎 @飞虹舞毓

安装

3-4、python 模块安装

具体的安装方法，如下

Python Modules

Python bindings for the ONNX-TensorRT parser are packaged in the shipped `.whl` files. Install them with

```
python3 -m pip install <tensorrt_install_dir>/python/tensorrt-8.x.x-cp<python_ver>-none-linux_x86_
```

TensorRT 8.2.1.8 supports ONNX release 1.8.0. Install it with:

```
python3 -m pip install onnx==1.8.0
```

The ONNX-TensorRT backend can be installed by running:

```
python3 setup.py install
```

知乎 @飞虹舞毓

安装python模块

测试，下列代码导入没有相关的报错

```
In [1]: import onnx
...: import onnx_tensorrt.backend as backend
```

4、总结

总结下来，这个搭建环境的过程真的有很多的坑，整体的耗时大概有2天左右的事件，主要的问题就是各个版本的对应问题，强烈大家使用docker环境。


希望这个资料对大家后续搭建相关的环境有所帮助

5、参考资料

github.com/onnx/onnx-te...

github.com/protocolbuff...

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



发布



还没有评论，发表第一个评论吧

推荐阅读



【TensorRT系列】2.ONNX-TensorRT安装教程

Tramac



用TensorRT优化Mobilenetv2踩坑记录

无心睡眠



TensorRT组件编译+ONNX部署+CMake实战

Jack Sigmoid

TensorRT ONNX 学习

TensorRT 安装windows软件提供python支持 如果使用python api, 需要安装pycuda,pip install pycuda>=2019.1.1 tensorflow 6.0.1支持cuda9.0, 10.0, 10.2.NVIDIA TensorRT设置

闪电侠的右... 发表于计算机视觉