模型部署资料

2021年12月13日　　16:10

TensorRT部分

1. TensorRT官方资料
    a. TensorRT GitHub： https://github.com/NVIDIA/TensorRT 里面有官方给出的例子
    b. TensorRT官方文档: https://docs.nvidia.com/deeplearning/tensorrt/archives/tensorrt-803/developer-guide/index.html
    c. TensorRT官方优化指南: https://docs.nvidia.com/deeplearning/tensorrt/archives/tensorrt-803/best-practices/index.html
2. TensorRT和pytorch
    a. https://github.com/NVIDIA-AI-IOT/torch2trt： torch2trt是一个PyTorch到TensorRT的转换器，它利用了TensorRT Python API
3. Tensorrt和ONNX
    a. ONNX Models zoo： https://github.com/onnx/models
    b. https://github.com/onnx/onnx-tensorrt
4. TensorRT和Tensorflow
    a. https://github.com/tensorflow/tensorrt 集成了tensorrt的Tensorflow
5. TensorRT和各种网络
    a. https://elinux.org/TensorRT/YoloV3
    b. https://elinux.org/TensorRT/YoloV4
    c. https://github.com/Megvii-BaseDetection/YOLOX
    d. https://github.com/CoinCheung/BiSeNet
    e. https://github.com/wang-xinyu/tensorrtx 提供了非常多的基于trt部署的网络代码实现
    f. https://github.com/NVIDIA/retinanet-examples
    g. https://github.com/GeekAlexis/FastMOT
    h. https://github.com/hunglc007/tensorflow-yolov4-tflite
    i. https://github.com/CaoWGG/TensorRT-CenterNet
    j. https://github.com/NVIDIA-AI-IOT/trt_pose
    k. https://github.com/shouxieai/tensorRT_Pro
    l. https://github.com/grimoire/mmdetection-to-tensorrt
    m. https://github.com/nvidia/tensorrt-laboratory/
    n. https://github.com/NVIDIA/TensorRT/tree/master/demo
6. TensorRT FAQ系列
    a. 官方FAQ： TensorRT Developer Guide#FAQs
    b. 常见问题解答:
        i. 你可以在这里找到关于使用TRT的一些常见问题的答案。请参考页面TensorRT/CommonFAQ
    c. TRT准确度常见问题:
        i. 如果你的FP16结果或Int8结果不符合预期，下面的页面可以帮助你解决精度问题。请参考TensorRT/AccuracyIssues

d. TRT性能常见问题：
    i. 如果用TRT做推理的性能没有达到预期，下面的页面可能会帮助你优化性能。请参阅TensorRT/PerfIssues
e. TRT Int8校准常见问题
    i. 下面的页面将介绍一些关于TRT Int8校准的常见问题。请参考页面TensorRT/Int8CFAQ
f. TRT插件常见问题
    i. 下面的页面将介绍一些关于TRT插件的常见问题。请参考TensorRT/PluginFAQ
g. 如何解决一些常见的错误
    i. 如果你在使用TRT时遇到一些错误，请在下面的页面中找到答案。请参阅TensorRT/CommonErrorFix。
h. 如何调试或分析
    i. 下面的页面将帮助你以某种方式调试你的推理。请参考TensorRT/How2Debug

CUDA部分

1. CUDA docker镜像： https://hub.docker.com/r/vistart/cuda
2. CUDA书籍
    a. 《CUDA并行程序设计-GPU编程指南》
    b. 《CUDA by Example》
    c. 《GPU高性能编程CUDA实战》
3. CUDA在线文档
    a. CUDA C++ Programming Guide
    b. CUDA C++ Best Practices Guide
    c. CUDA for Tegra
    d. https://docs.nvidia.com/cuda/#cuda-api-references 各种CUDA API：CUBLAS、CUFFT、NPP等
4. CUDA资源
    a. https://docs.opencv.org/4.1.2/d1/d1e/group__cuda.html
    b. https://developer.download.nvidia.cn/compute/cuda/1.1-Beta/x86_website/projects/reduction/doc/reduction.pdf Optimizing Parallel Reduction in CUDA
    c. CUDA 进阶学习
    d. 苹果妖—CUDA
    e. intro to parallel programming 视频：https://www.bilibili.com/video/BV1yt411w7h8
    f. CUDA Wiki
5. 官方博客
    a. CUDA Refresher: Reviewing the Origins of GPU Computing
    b. CUDA Refresher: Getting started with CUDA
    c. CUDA Refresher: The GPU Computing Ecosystem
    d. CUDA Refresher: The CUDA Programming Model
    e. How to Implement Performance Metrics in CUDA C++

f. How to Query Device Properties and Handle Errors in CUDA C++
g. How to Optimize Data Transfers in CUDA C++
h. How to Overlap Data Transfers in CUDA C++
i. How to Access Global Memory Efficiently in CUDA C++
j. Using Shared Memory in CUDA C++
k. An Efficient Matrix Transpose in CUDA C++
l. Finite Difference Methods in CUDA C++, Part 1
m. Finite Difference Methods in CUDA C++, Part 2
n. Accelerated Ray Tracing in One Weekend with CUDA

6. CUDA代码
   a. https://github.com/NVIDIA/cuda-samples
   b. libcu++，NVIDIA C++标准库，是整个系统的C++标准库。它提供了一个C++标准库的异构实现，可以在CPU和GPU代码中及之间使用。
7. CUDA工具
   a. CUDA Pro Tip: nvprof is Your Handy Universal GPU Profiler
   b. CUDA-GDB
   c. https://github.com/NVIDIA/nvbench cuda kernel benchmark
   d. https://github.com/PatWie/cuda-design-patterns 并没有讲述如何使用cuda，但是简化了一些workflow
   e. https://github.com/NVIDIA/thrust 对标C++ STL的并行编程库
   f. CUDA Occupancy Calculator CUDA利用率计算工具，用excel计算