DNN Thread 技术 paper

概述

DNN Thread 是使用原生 pascal 复刻而出的 cuda 核线程调度技术。

DNN Thread 可以支持任意 ZAI 计算引擎:GPU/MKL/X86/AIOT/JETSON/华为(AI Module 目前处于考察中,华为明确邮件答复内容,mindspore/grapengine 的技术战略并不是以支持三方 EDGE 设备公司为主,而是自己尝试着来做 AI 解决方案,复制 nv 运作模式).

DNN Thread 是纯正 pascal 实现的超算支持技术.

DNN Thread 可以在 FPC/Delphi 中编译和使用.

DNN Thread 可以工作与 arm/小序列 mips/jetson nano/huawei atlas 等等硬件设备中(需要 AI 计算引擎支持这类 IOT 设备,因为 edge 设备对于线程支持度很差,与 HPC 差距会非常大,因此 ZAI 也一直没有构建运行于 edge 设备中的计算引擎)

5G 移动网络+HPC 效果非常好,支持各种实时级应用.目前搭配一代 GPU 的 HPC 也很便宜,20K 内可以完成非常优良的 1 代 GPU 设备投资)

工作原理

GPU 是异构计算的硬件设备.GPU 的工作原理都是用传输(PCI)copy 给异构设备进行计算,然后取出结果来使用.这个过程叫 IO(Input/Output).

DNN Thread,内置是指令化模型,通过线程支持技术 TThreadProgressPost,向计算引擎 GPU/MKL 发送 command,计算引擎会将计算命令进行对应 GPU-SM/MKL-X86 的架构翻译,通过 PCI/北桥芯片传递给 GPU/MKL 异构设备,待计算完成后取出结果,形成 IO 过程.

结构说明

设计并没有使用工程化方式,直接面向 IO 机制定义出自动化结构,并没有任何多余设计.

DNN Thread 非常小巧,一共只有 10 个类,全部按自动化工作流程设计,分别如下

- TAI_DNN_Thread:基类,不能直接使用
- TAI_DNN_ThreadPool:线程池管理
- TAI_DNN_Thread_Metric:度量化网络支持
- TAI_DNN_Thread_LMetric:大规模度量化网络支持
- TAI_DNN_Thread_MMOD6L:通用检测器支持

- TAI_DNN_Thread_MMOD3L:高速的通用检测器支持
- TAI_DNN_Thread_RNIC:Resnet 图片分类器支持
- TAI_DNN_Thread_LRNIC:大规模 Resnet 图片分类器支持
- TAI_DNN_Thread_GDCNIC: Going Deeper with Convolutions 分类器支持
- TAI_DNN_Thread_GNIC: Gradient-based learning applied to document recognition 分类器支持
- TAI_DNN_Thread_SS:语义分割支持

依赖性说明

DNN Thread 调度核心依赖关系如下

1. TAI_DNN_Thread_MMOD6L (DNN Thread 的支持类)
2. TThreadProgressPost:(CoreProgressPost.inc,线程计算调度模块,属于 CoreClasses.pas 库)
3. TCompute:(CoreComputeThread.inc,线程调度模块, 属于 CoreClasses.pas 库)
4. TAI 引擎(zAI.pas 库)
5. TAI 引擎中 MMOD 的 API
6. AI 计算引擎(zAI_Cuda.dll 这类编译好的外部库)
7. cuda 驱动程序(例如,zAI1.32 需要对应 cuda10.2 sdk)
8. 异构硬件(GPU)

使用说明

初始化,默认按每 GPU 设备 2*IO 线程做计算构建,这里可以根据参数自己调整

```
DNNFaceDetectorPool := TAI_DNN_ThreadPool.Create;
DNNFaceDetectorPool.BuildPerDeviceThread(TAI_DNN_Thread_MMOD6L); // 异构线程池:人脸检测器
for i := 0 to DNNFaceDetectorPool.Count - 1 do
    TAI_DNN_Thread_MMOD6L(DNNFaceDetectorPool[i]).Open_Face; // 对每个实例发送一条加载内置人脸检测器模型的异步命令
```

使用方式,线程,循环,事件,直接按例程/Demo 调用即可,
下列范例,为人脸检测,在多卡 GPU 的 HPC 中可以支持了 200 路以上实时视频检测+识别.

```
        // 拾取一个负载量最小的dnn线程,让它执行视频解析工作
        th := DNNFaceDetectorPool.MinLoad_DNN_Thread as TAI_DNN_Thread_MMOD6L;
        th.ProcessP(nil, raster.Clone, True,
          procedure(thSender: TAI_DNN_Thread_MMOD6L; UserData: Pointer; Input: TMemoryRaster; output: TMMOD_Desc)
          // 检测完成以后,以异步方式触发该事件,不会卡GPU的处理队列
          begin
          end);
```

by qq600585
2020-7