2019.4.27

TVM:手机[gpu|cpu]上运行 Tensorflow 模型

目的

描述在手机上运行 tvm 的过程,分 gpu/cpu 两种情况.
手机上运行 tvm 的话，考虑到要支持 android 和 ios 两个平台，用 c++来提高可移植性.

手机 cpu

用 c++调用 tvm，参考本文末尾的参考2和参考3就可以了.

手机 gpu

测试一下你的手机是否支持 gpu

在手机上安装参考5(Opencl-Z)，能查看手机 gpu 信息。
华为 mate7根本不支持 gpu 运算.
华为麦芒5看起来支持 gpu 运算，但 tvm 会运行出错.
华为 Mate10/P10/小米6就可以运行 tvm，华为麦芒5/小米5就不行。

准备 TVM C++代码

要在手机上运行 gpu，需要参考本文末尾的参考4才行.
参考4和参考2/3的区别是：device_type 从 kDLCPU 变成了 kDLGPU.

为了在手机上运行 gpu，参考4内的 kDLGPU，要修改成 kDLOpenCL;

如果直接操作了 gpu 数据，opencl 会报告错误:
OpenCL Error, code=-38: CL_INVALID_MEM_OBJECT

参考4和参考2/3还有一个重要的区别，就是对输入/输出数据的处理。

为了把数据复制进 gpu,必须调用 TVMArrayCopyFromBytes。
为了把数据从 gpu 复制出来，必须调用 TVMArrayCopyToBytes。

Tvm device_type

Pc 上使用 nvida gpu 的话，编程接口用 cuda.
手机上使用 gpu 的话，编程接口需要用 OpenCL 库.
在 tvm 内，要使用 cpu 的话，运算 device_type 是 kDLCPU(1)。
要使用手机 opencl gpu 的话，运算 device_type 是 kDLOpenCL(4)。
要使用 pc nvida gpu 的话，运算 device_type 是 kDLGPU(2)

准备 opencl.h 和.so 文件

编译 tvm，需要 opencl 的头文件.
从参考6下载 opencl.h 头文件.
手机上编译 tvm，还需要 libOpenCL.so 文件.
可以直接从你的手机内，找到 libOpenCL.so，下载下来，放到你的 ndk 目录内去.
我的手机和 ndk 目录位置:

手机指令集	从手机上下载	复制到 NDK 目录
Armeabi-v7	/vendor/lib/libOpenCL.so	C:\Users\<your name>\AppData\Local\Android\Sdk\android-ndk-r19\ndk-bundle\toolchains\llvm\prebuilt\windows-x86_64\lib64\clang\8.0.2\lib\linu
Arm64-v8	/vendor/lib64/libOpenCL.so	C:\Users\<your name>\AppData\Local\Android\Sdk\android-ndk-r19\ndk-bundle\toolchains\llvm\prebuilt\windows-x86_64\lib64\clang\8.0.2\lib\linu

准备手机 UI

准备好以上的 c++代码,opencl.h/.so,就可以编译出 tvm so 库文件了。
为了在手机上运行起来，需要 java UI 界面。
可以用 tvm 自带的 android demo(参考7)，编译后运行起来。
Tvm 自带的代码会从网络下载数据，并且由于调用了 linux 命令，只能在 linux 下编译。

运行 tvm

编译出 android demo app 后，就能在手机上测试了。

手机 gpu 运行速度测试

我们使用 gpu 的目的是为了提高速度，但是据说是 opencl 的一个 bug，导致 tvm gpu 在手机上失去了速度优势。
问题1是：tvm 第一次加载运行模型，opencl 调用会非常慢，然后后面就会快起来。对手机 app 来讲，启动慢，是不可忍受的。

问题2是：tvm 使用 gpu 的话，推理(run)确实很快，但是要取得推理(run)的结果（get_output），却非常慢。
问题2参见参考8。

考虑到这两个问题，在手机上，还是用 cpu 吧，gpu 还没法实用。

参考

名词解释:

名词1	t TVM：优化模型运算速度，生成能在手机上运行的二进制代码。
名词2	GPU：比 cpu 运算快，专门用于大数据运算.
名词3	Tensorflow: google 推的人工智能模型训练框架.

参考链接:

参考1	tvm 网站 https://tvm.ai/
参考2	在 CPP 下使用 TVM 来部署 mxnet 模型（以 Insightface 为例） https://zhuanlan.zhihu.com/p/55996985?utm_source=wechat_timeline&utm_medium=social&utm_oi=882
参考3	一步一步解读神经网络编译器 TVM(二)——利用 TVM 完成 C++端的部署 https://zhuanlan.zhihu.com/p/60981432
参考4	手机上运行 tvm gpu c++ tvm_deploy_gpu_sample.cpp *https://gist.github.com/masahi/d6ad36890d087f866de185f19aac3814*
参考5	Opencl-Z app 检测 android 是否支持 opencl(GPU) https://stackoverflow.com/questions/26795921/does-android-support-opencl

	You can use OpenCL-Z Android to check the available and capabilities of OpenCL on Andr OpenCL-Z apk: https://www.allfreeapk.com/opencl-z,1359401/download.html
参考6	下载 opencl.h https://stackoverflow.com/questions/29082524/android-studio-fatal-error-cl-cl-h-no-such-file-or-directory
参考7	Tvm 的 android demo(linux) Apps/android_deploy https://github.com/dmlc/tvm/tree/master/apps/android_deploy
参考8	How to make the GPU to CPU memory copy faster? #979 https://github.com/dmlc/tvm/issues/979?from=timeline clEnqueueReadBuffer is too slow https://stackoverflow.com/questions/32190761/clenqueuereadbuffer-is-too-slow?from=time