6、Profiling

本文采用AlexNet网络作为基准，评估Caffe的性能。AlexNet的网络结构如图【来源请求】，其主体由五层卷积层连接三层全连接层构成。为方便开发者测试网络性能，Caffe封装了time函数，对外以argument的形式开放。我们将time函数劫持出来，写到自定义源文件caffe-time.cpp的入口，放到/tools下随Caffe的其他源文件一起编译得到caffe-time.bin。编译完成后，我们采用这个命令作为性能评估基准：

./build/tools/caffe-time.bin --model=./models/bvlc\_alexnet/deploy.prototxt –iterations 50

caffe-time.bin代表含有time函数的可执行文件；model参数赋值为AlexNet的ProtoTXT格式网络描述；iterations为测试的迭代次数。由于第一次迭代开始之前，内存和CPU缓存是冷的，所以需要多次迭代测试取平均值来消除这种误差。“一次迭代”在这个语境下指对整个网络进行一次前向计算和反向传播。该命令等价于

./build/tools/caffe time --model=./models/bvlc\_alexnet/deploy.prototxt –iterations 50

执行命令之后，Caffe的输出范例如下【来源请求，Intel whitepaper】

I0101 01:08:22.013758 3686 caffe-time.cpp:56] Testing for 50 iterations.

I0101 01:08:25.120884 3686 caffe-time.cpp:84] Iteration: 1 forward-backward time: 3107 ms.

I0101 01:08:27.764709 3686 caffe-time.cpp:84] Iteration: 2 forward-backward time: 2643 ms.

I0101 01:08:30.477269 3686 caffe-time.cpp:84] Iteration: 3 forward-backward time: 2712 ms.

I0101 01:08:33.219909 3686 caffe-time.cpp:84] Iteration: 4 forward-backward time: 2742 ms.

I0101 01:08:35.858683 3686 caffe-time.cpp:84] Iteration: 5 forward-backward time: 2638 ms.

I0101 01:08:38.482145 3686 caffe-time.cpp:84] Iteration: 6 forward-backward time: 2623 ms.

……

I0101 01:10:31.874108 3686 caffe-time.cpp:84] Iteration: 47 forward-backward time: 2829 ms.

I0101 01:10:34.847990 3686 caffe-time.cpp:84] Iteration: 48 forward-backward time: 2973 ms.

I0101 01:10:37.622038 3686 caffe-time.cpp:84] Iteration: 49 forward-backward time: 2773 ms.

I0101 01:10:40.450104 3686 caffe-time.cpp:84] Iteration: 50 forward-backward time: 2827 ms.

I0101 01:10:40.450280 3686 caffe-time.cpp:87] Average time per layer:

I0101 01:10:40.450315 3686 caffe-time.cpp:90] dataforward: 0.00376 ms.

I0101 01:10:40.450361 3686 caffe-time.cpp:93] databackward: 0.00446 ms.

I0101 01:10:40.450405 3686 caffe-time.cpp:90] conv1forward: 158.396 ms.

I0101 01:10:40.450447 3686 caffe-time.cpp:93] conv1backward: 157.189 ms.

I0101 01:10:40.450488 3686 caffe-time.cpp:90] relu1forward: 14.4338 ms.

I0101 01:10:40.450694 3686 caffe-time.cpp:93] relu1backward: 0.00406 ms.

I0101 01:10:40.450765 3686 caffe-time.cpp:90] norm1forward: 61.1189 ms.

I0101 01:10:40.450791 3686 caffe-time.cpp:93] norm1backward: 69.6954 ms.

I0101 01:10:40.450812 3686 caffe-time.cpp:90] pool1forward: 104.729 ms.

I0101 01:10:40.450834 3686 caffe-time.cpp:93] pool1backward: 0.00524 ms.

I0101 01:10:40.450855 3686 caffe-time.cpp:90] conv2forward: 316.762 ms.

I0101 01:10:40.450878 3686 caffe-time.cpp:93] conv2backward: 306.659 ms.

I0101 01:10:40.450899 3686 caffe-time.cpp:90] relu2forward: 13.2367 ms.

I0101 01:10:40.450920 3686 caffe-time.cpp:93] relu2backward: 0.00414 ms.

I0101 01:10:40.450942 3686 caffe-time.cpp:90] norm2forward: 47.8888 ms.

I0101 01:10:40.450963 3686 caffe-time.cpp:93] norm2backward: 47.677 ms.

I0101 01:10:40.450985 3686 caffe-time.cpp:90] pool2forward: 64.9354 ms.

I0101 01:10:40.451006 3686 caffe-time.cpp:93] pool2backward: 0.00482 ms.

I0101 01:10:40.451028 3686 caffe-time.cpp:90] conv3forward: 209.004 ms.

I0101 01:10:40.451049 3686 caffe-time.cpp:93] conv3backward: 211.287 ms.

I0101 01:10:40.451070 3686 caffe-time.cpp:90] relu3forward: 3.98946 ms.

I0101 01:10:40.451089 3686 caffe-time.cpp:93] relu3backward: 0.00468 ms.

I0101 01:10:40.451109 3686 caffe-time.cpp:90] conv4forward: 156.937 ms.

I0101 01:10:40.451132 3686 caffe-time.cpp:93] conv4backward: 157.686 ms.

I0101 01:10:40.451151 3686 caffe-time.cpp:90] relu4forward: 3.49588 ms.

I0101 01:10:40.451172 3686 caffe-time.cpp:93] relu4backward: 0.00434 ms.

I0101 01:10:40.451191 3686 caffe-time.cpp:90] conv5forward: 107.479 ms.

I0101 01:10:40.451212 3686 caffe-time.cpp:93] conv5backward: 107.145 ms.

I0101 01:10:40.451232 3686 caffe-time.cpp:90] relu5forward: 2.47902 ms.

I0101 01:10:40.451253 3686 caffe-time.cpp:93] relu5backward: 0.00286 ms.

I0101 01:10:40.451272 3686 caffe-time.cpp:90] pool5forward: 14.9739 ms.

I0101 01:10:40.451293 3686 caffe-time.cpp:93] pool5backward: 0.00368 ms.

I0101 01:10:40.451313 3686 caffe-time.cpp:90] fc6forward: 203.838 ms.

I0101 01:10:40.451333 3686 caffe-time.cpp:93] fc6backward: 74.5167 ms.

I0101 01:10:40.451355 3686 caffe-time.cpp:90] relu6forward: 0.41766 ms.

I0101 01:10:40.451375 3686 caffe-time.cpp:93] relu6backward: 0.00292 ms.

I0101 01:10:40.451395 3686 caffe-time.cpp:90] drop6forward: 0.44218 ms.

I0101 01:10:40.451415 3686 caffe-time.cpp:93] drop6backward: 0.0051 ms.

I0101 01:10:40.451435 3686 caffe-time.cpp:90] fc7forward: 83.049 ms.

I0101 01:10:40.451457 3686 caffe-time.cpp:93] fc7backward: 34.9812 ms.

I0101 01:10:40.451476 3686 caffe-time.cpp:90] relu7forward: 0.26226 ms.

I0101 01:10:40.451497 3686 caffe-time.cpp:93] relu7backward: 0.00266 ms.

I0101 01:10:40.451517 3686 caffe-time.cpp:90] drop7forward: 0.47262 ms.

I0101 01:10:40.451538 3686 caffe-time.cpp:93] drop7backward: 0.00318 ms.

I0101 01:10:40.451557 3686 caffe-time.cpp:90] fc8forward: 21.9401 ms.

I0101 01:10:40.451578 3686 caffe-time.cpp:93] fc8backward: 10.1705 ms.

I0101 01:10:40.451598 3686 caffe-time.cpp:90] probforward: 0.45548 ms.

I0101 01:10:40.451619 3686 caffe-time.cpp:93] probbackward: 0.1202 ms.

I0101 01:10:40.451649 3686 caffe-time.cpp:98] Average Forward pass: 1591.12 ms.

I0101 01:10:40.451672 3686 caffe-time.cpp:100] Average Backward pass: 1177.4 ms.

I0101 01:10:40.451692 3686 caffe-time.cpp:102] Average Forward-Backward: 2768.74 ms.

I0101 01:10:40.451713 3686 caffe-time.cpp:104] Total Time: 138437 ms.

I0101 01:10:40.451732 3686 caffe-time.cpp:105] \*\*\* Benchmark ends \*\*\*

可以从输出中得知：

·指定次数迭代中，每次迭代使用总时间、每次迭代使用的平均时间；

·多次迭代中各层的平均耗费时间；

·多次迭代中总的前向反向计算时间；

·总共所用时间。

我们对Armeabi-v7a和Arm64-v8a两种ARM架构进行测试。其中，Armeabi-v7a对应的测试设备为ZTE Nubia Z7；【来源请求，性能参数】Arm64-v8a对应的测试平台为Qualcomm 820开发板。【来源请求，性能参数】两个平台使用【commit号版本】的caffe-android-lib脚本进行交叉编译。考虑到Arm64-v8a的兼容性，使用NDK版本为【NDK r11c】