知

首发于
片上神经网络

写文章 登录

给DNN处理器跑个分 - 指标篇

唐杉 · 4 个月前

> 对越来越多的DNN专用处理器设计（芯片和IP），我们很自然的需要解决一个问题“怎样对不同的DNN处理器设计做出公平的比较和评价？”能不能像手机跑分一样也让它们跑个分呢？这实际是个基准测试（Benchmarking）问题。正好最近看到两个在这方面的尝试，一

> 个是MIT Eyeriss团队给出的DNN Processor Benchmarking Metrics；另一个是百度DeepBench项目的更新。我们可以结合这两个项目讨论一下如何给DNN处理器“跑分”。

首先我们还是简单回顾一下Benchmark的背景知识。以下是Wiki的定义

> *In computing, a benchmark is the act of running a computer program, a set of programs, or other operations, in order to assess the relative performance of an object, normally by running a number of standard tests and trials against it.The term 'benchmark' is also mostly utilized for the purposes of elaborately designed benchmarking programs themselves.*

实际上，我们非常熟悉的手机跑分也是Benchmark的一种。Benchmark通常用于评估计算机硬件的性能特征,例如CPU的浮点运算性能,GPU的图像处理能力，存储系统访问的速度等等，有时也用于软件或者编译工具。总的来说，Benchmark提供了一种比较不同软硬件系统性能的方法。从另一个角度来说，Benchmark也可以在我们自己的设计过程中帮助我们评价不同版本优化改进的情况，这可以看作是一种纵向比较。

在我们研发的过程中，会用到各种各样的测试程序。Benchmark本质上也是测试程序，它的特殊性体现在：大家公认它可以评价（一般是定量的）被测试目标的某些特性。如果不是公认的基准测试，也就谈不上横向比较的作用了。

专用的DNN Processor（或者DNN/DL/ML Processor/Accelerator）的研究和应用还是最近几年的事情，目前还没有大家公认的基准测试方法和指标。一个原因是，专用硬件针对特定应用（范围比较窄），一般也进行了特殊的优化，设计能够相互比较的基准测试比较困难。但是，由于AI/Deep Learning的快速发展和广阔前景，目前很多研究机构和公司都投入到专用的DNN

处理器的研发当中，对比和竞争在所难免。而技术正是在对比和竞争中不断发展的，DNN处理器的基准测试已经越来越得到大家的关注。

> 1

MIT的Eyeriss团队无疑是DNN Processor界一股重要的力量。他们最近搞了一个“Tutorial on Hardware Architectures for Deep Neural Networks”[1]，非常全面，建议同学们好好看看。这个Tutorial中有一部分就是“Benchmarking Metrics”。

通常，我们说Benchmarking的时候一般包括两部分内容，一是测试程序（包括测试方法），即Benchmark本身；二是结果的表达（或者评价指标），即Metrics。例如，经典的Benchmark：Dhrystone，它的测试方法就是在目标处理器上运行一段精心设计的C程序：“*simple programs that are carefully designed to statistically mimic the processor usage of some common set of programs*”。运行这段程序实际会得到很多信息，而它采用的Dhrystone score的表达为“ *the number of Dhrystones per second (the number of iterations of the main code loop per second)*”。这里为什么不用MIPS (每秒能运行多少百万条指令)这种更常见表达呢？这是因为，对于RISC和CISC这两种不同的指令架构，在task层面统计结果要比在指令层面更为公平合理。此外，还可以用DMIPS (Dhrystone MIPS) 作为指标，即Dhrystone成绩除以1757（这是当年在VAX 11/780机器上获得的每秒Dhrystones的数量，名义上作为1 MIPS）；或者使用DMIPS/MHz，这样可以方便对比运行在不同时钟频率上的处理器。从Dhrystone的例子可以看出，对于一个成功的Benchmark来说，合理有效的测试程序和Metrics都是非常重要的。

我们回到Eyeriss的方案。应该说他们最主要的贡献在于Metrics。他们并没有专门设计“测试程序”，而是直接使用“*widely used state-of-the-art DNNs (e.g. AlexNet, VGG, GoogLeNet, ResNet) with input from well known datasets such as ImageNet*”。采用“经典”的DNN网络作为

Benchmark相当于用目标应用来作为测试程序，虽然比较直接，但也有很多问题。我们在本文“下篇”分析百度的DeepBench的时候再做详细讨论。

Eyeriss团队提出两个层次的Metrics。第一个是Metrics for DNN Algorithm，主要是算法层面的指标。我们重点看第二个，Metrics for DNN Hardware：

Measure energy and off-chip (e.g., DRAM) access relative to number of non-zero MACs and bit-width of MACs

- Account for impact of sparsity in weights and activations
- Normalize DRAM access based on operand size
- To compute the off-chip access, assume the DNN processor is a stand-alone chip. The off-chip access should account for all accesses needed to complete all the layers listed including initial inputs and final outputs from an off-chip device (e.g., DRAM). The goal is to compare the off-chip access at steady state, so accesses during ramp-up/ramp do not need to be included (e.g. loading configuration parameters, or loading weights *if* all weights can be stored on chip).

Energy Efficiency of Design

- pJ/non-zero MAC

External Memory Bandwidth

- Off-chip access (in Bytes)/non-zero MAC

Area Efficiency

- Total chip mm2/multiplier and storage capacity/multiplier
- Accounts for on-chip memory

这里列出的一些指标是非常有针对性的。它们和传统的指标，比如Run Time，Power，相结合，希望能够覆盖：Accuracy，Power，Throughput，Cost这些硬件相关的基本要素，并能够提供External memory bandwidth，Required on-chip storage，Utilization of cores这些重要信息。可以看出，很多指标都非常强调效率，Efficiency，特别是对于面积和功耗这些和成本相关的指标。另外，在对DNN处理器的核心MAC做相关测试时，强调了non-zero MAC（和稀疏性相关）以及bit-width of MAC（和精度相关）这两个条件，反映了DNN的特点。

> 2

说到这里，顺便提一下我刚看到的另一个指标，来自Intel/Nervana的Naveen Rao在“O'Reilly Artificial Intelligence Conference 2017”上的一个讲演。

他提出了Computational Capacity的概念，它由几个因素来决定：1）Memory Bandwidth（m）；2）Precision（量化的比特数 b）；3）Utilized OPs（每秒有效的操作 o）。这个概念的提出，是因为常用的FLOPS指标已经不能准确的评价处理器的DNN处理能力了，而综合上述三个因素才是更合理的表示方法。这个指标和上述Eyeriss提出的多个指标有

一定的类似之处，它的好处是比较简洁，可以用一个公式来表示，但它包含的信息也要少一些。具体可以参考他的文章，Comparing dense compute platforms for AI - Intel Nervana。

另外一个很有用的指标（或者叫性能评价模型）是Roofline Model[3]。通过这个model，既可以评估一个设计的效率，还能很容易看出你的设计倒底是computation-limited还是memory bandwidth-limited，可以帮助你确定进一步优化的思路.

比如Google在TPU的论文里采用这Roofline Model来和GPU，CPU进行了对比，如下图。

Roofline model很好的说明了，一个好的评价模型，可以很直观的给我们展示出最重要的信息。它的玩法很多，用好了也很有帮助，建议大家好好看看。。

> 3

再回到正题，下面是对Eyeriss处理器为例进行Benchmarking的结果。首先是处理器的spec和芯片整体的测试结果。这些需要指出的是，处理器的Spec在进行对比的时候也是很有必要的，是很多对比评价的基础信息。

然后是对Alexnet各个layer分解的测试结果。

最后还给出了用来评价FPGA的指标。

由于我们关注的是Benchmarking的指标设计，这里就不具体分析测试结果了。Eyeriss团队还设立了一个网站（rle.mit.edu/eems/dnn-be...），供大家提交自己的测试结果。在这个网站上可以下载相应的表格，也可以看到Eyeriss处理器的实例。

总的来说，由于DNN处理器的特殊性，比如对memory bandwidth的需求、DNN稀疏性的特点、MAC利用率问题等等，对于它做基准测试的时候很难简单的借用传统处理器的评价指标。我们需要综合性的或更有针对性的指标才能更好的表征DNN处理器的实际效能。Eyeriss团队提出一系列比较有针对性的指标，基本上能够覆盖了DNN处理器的各方面性能。不过这些指标能否以更简洁直观的形式表达，也是值得我们思考的问题。

> 4

Eyeriss团队总结的Benchmarking Metrics，对于评价DNN处理器，甚至设计DNN处理器都很有启发。但是，他们使用几种DCNN网络（AlexNet，VGG16，GoogleNet，ResNet-50）作为Benchmark的方法是否合理有效呢？

首先，一个实际的问题是，要得到所有这几个网络在目标硬件上运行的数据，是一项巨大的工程，特别是对于小规模的研究团队来说。我们看到即使是Eyeriss处理器也只给出了AlexNet和VGG16的结果。实际上，这几种DCNN网络还是有不少相似之处的，用它们作为Benchmark，是否是做了很多重复和冗余的测试？

第二个问题是，这几种网络是否真正覆盖了各种DNN的需求。虽然Eyress更关注卷积层的处理，但FC/RNN/LSTM/GRU这类网络的应用也很广泛，在很多新出现非常有效的模型中，FC/RNN类型的层和卷积层经常是结合在一起使用的。对于这种情况，只用这几种以CNN为主的网络作为Benchmark是否有足够的代表性呢？如果还是采取把实际网络用作Benchmark的思路，我们就需要不断的扩大这个Benchmark的集合。显然这也是不可取的。

解决上述问题的一个思路就是“设计”新的Benchmark，就像Dhrystone这样的“Synthetic Benchmark”一样。今天就先到这里，下次我们结合Baidu的DeepBench讨论一下Benchmark的设计问题。

T.S.

> Reference: [1] “Tutorial on Hardware Architectures for Deep Neural Networks”, eyeriss.mit.edu/tutoria... [2] "Benchmarking DNN Processors" , eyeriss.mit.edu/benchma... [3] Williams, Samuel; Waterman, Andrew; Patterson, David (2009-04-01). "Roofline: An Insightful Visual Performance Model for Multicore Architectures". Commun. ACM. 52 (4): 65–76. ISSN 0001-0782. doi:10.1145/1498765.1498785

欢迎关注我的微信公众号:StarryHeavensAbove

题图来自网络，版权归原作者所有

推荐阅读

自己动手设计专用处理器！

当我们设计一个专用处理器的时候我们在干什么？（指令集）

当我们设计一个专用处理器的时候我们在干什么？（微结构）

深度神经网络的模型·硬件联合优化

追求极限性能的芯片设计方法（一）

追求极限性能的芯片设计方法（二）

追求极限性能的芯片设计方法（三）

追求极限性能的芯片设计方法（四）

「真诚赞赏，手留余香」

赞赏

还没有人赞赏，快来当第一个赞赏的人吧！

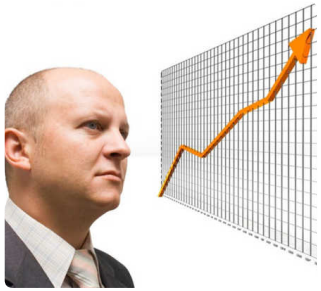
深度学习（Deep Learning） 芯片设计

处理器

劳务派遣中三方主体的权利义务

自2008年《劳动合同法》实施以来，许多企业为规避用工风险，多采取劳务派遣的形式——通过第三方人力资源公司的介入，阻隔其与劳动者之间的关系。值得注意的是，劳务派遣与劳务外包具有本质… 查看全文 >

夏心 · 2 个月前 · 编辑精选



我为什么要在Excel和R之间徘徊——数据分析者的基本修养

这两天兴致上头，暂时把数据分析的学习抛诸脑后，竟然去写Calligraphy的文章去了。但是，我… 查看全文 >

Still · 1 年前 · 编辑精选

关于《英雄联盟》音乐节，这里应该有你想知道的一切

文：Paolo图：一村、网络11月4日，2017全球总决赛的最终总冠军争夺将在北京鸟巢体育馆展开。… 查看全文 >

PentaQ刺猬电竞社 · 14 天前 · 编辑精选

不注意这4点，新股东也可能变成双刃剑！

对公司而言，引入新股东无疑将为公司带来新的生命力，但是，股东之间纠纷时有发生，如果… 查看全文 >

信之源律师事务所 · 10 天前 · 编辑精选