Benchmarking Metrics for DNN Hardware

ISCA Tutorial (2017)

Website: http://eyeriss.mit.edu/tutorial.html

Joel Emer, Vivienne Sze, Yu-Hsin Chen

Metrics Overview

- How can we compare designs?
- Target Metrics
  - Accuracy
  - Power
  - Throughput
  - Cost
- Additional Factors
  - External memory bandwidth
  - Required on-chip storage
  - Utilization of cores

Download Benchmarking Data

- Input (http://image-net.org/)
  - Sample subset from ImageNet Validation Dataset

- Widely accepted state-of-the-art DNNs (Model Zoo: http://caffe.berkeleyvision.org/)
  - AlexNet
  - VGG-16
  - GoogleNet-v1
  - ResNet-50

Metrics for DNN Algorithm

- Accuracy
- Network Architecture
  - # Layers, filter size, # of filters, # of channels
- # of Weights (storage capacity)
  - Number of non-zero (NZ) weights
- # of MACs (operations)
  - Number of non-zero (NZ) MACS

Metrics of DNN Algorithms

Metrics	AlexNet	VGG-16	GoogLeNet (v1)	ResNet-50
Accuracy (top-5 error)*	19.8	8.80	10.7	7.02
Input	227x227	224x224	224x224	224x224
# of CONV Layers	**5**	**16**	**21**	**49**
Filter Sizes	3, 5,11	3	1, 3 , 5, 7	1, 3, 7
# of Channels	3 - 256	3 - 512	3 - 1024	3 - 2048
# of Filters	96 - 384	64 - 512	64 - 384	64 - 2048
Stride	1, 4	1	1, 2	1, 2
# of Weights	2.3M	14.7M	6.0M	23.5M
# of MACs	666M	15.3G	1.43G	3.86G
# of FC layers	**3**	**3**	**1**	**1**
# of Weights	58.6M	124M	1M	2M
# of MACs	58.6M	124M	1M	2M
Total Weights	**61M**	**138M**	**7M**	**25.5M**
Total MACs	**724M**	**15.5G**	**1.43G**	**3.9G**

*Single crop results: https://github.com/jcjohnson/cnn-benchmarks

Metrics of DNN Algorithms

Metrics	AlexNet	VGG-16	GoogLeNet (v1)	ResNet-50
Accuracy (top-5 error)*	19.8	8.80	10.7	7.02
# of CONV Layers	**5**	**16**	**21**	**49**
# of Weights	2.3M	14.7M	6.0M	23.5M
# of MACs	666M	15.3G	1.43G	3.86G
# of NZ MACs**	**394M**	**7.3G**	**806M**	**1.5G**
# of FC layers	**3**	**3**	**1**	**1**
# of Weights	58.6M	124M	1M	2M
# of MACs	58.6M	124M	1M	2M
# of NZ MACs**	**14.4M**	**17.7M**	**639k**	**1.8M**
Total Weights	**61M**	**138M**	**7M**	**25.5M**
Total MACs	**724M**	**15.5G**	**1.43G**	**3.9G**
# of NZ MACs**	**409M**	**7.3G**	**806M**	**1.5G**

*Single crop results: https://github.com/jcjohnson/cnn-benchmarks

**# of NZ MACs computed based on 50,000 validation images

Metrics of DNN Algorithms

Metrics	AlexNet	AlexNet (sparse)
Accuracy (top-5 error)	19.8	19.8
# of Conv Layers	**5**	**5**
# of Weights	2.3M	2.3M
# of MACs	666M	666M
# of NZ weights	**2.3M**	**863k**
# of NZ MACs	**394M**	**207M**
# of FC layers	**3**	**3**
# of Weights	58.6M	58.6M
# of MACs	58.6M	58.6M
# of NZ weights	**58.6M**	**5.9M**
# of NZ MACs	**14.4M**	**2.1M**
Total Weights	**61M**	**61M**
Total MACs	**724M**	**724M**
# of NZ weights	**61M**	**6.8M**
# of NZ MACs	**409M**	**209M**

# of NZ MACs computed based on 50,000 validation images

Metrics for DNN Hardware

- **Measure energy and 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**
- **Energy Efficiency of Design**
  - **pJ/(non-zero weight & activation)**
- **External Memory Bandwidth**
  - **DRAM operand access/(non-zero weight & activation)**
- **Area Efficiency**
  - **Total chip mm^2/multi (also include process technology)**
  - **Accounts for on-chip memory**

ASIC Benchmark (e.g. Eyeriss)

ASIC Specs	
Process Technology	65nm LP TSMC (1.0V)
Clock Frequency (MHz)	200
Number of Multipliers	168
Total core area (mm^2) /total # of multiplier	0.073
Total on-Chip memory (kB) / total # of multiplier	1.14
Measured or Simulated	Measured
If Simulated, Syn or PnR? Which corner?	n/a

ASIC Benchmark (e.g. Eyeriss)

Layer by layer breakdown for AlexNet CONV layers

Metric	Units	L1	L2	L3	L4	L5	Overall*
Batch Size	#	4					
Bit/Operand	#	16					
Energy/ non-zero MACs (weight & act)	pJ/MAC	16.5	18.2	29.5	41.6	32.3	21.7
DRAM access/ non-zero MACs	Operands/ MAC	0.006	0.003	0.007	0.010	0.008	0.005
Runtime	ms	20.9	41.9	23.6	18.4	10.5	115.3
Power	mW	332	288	266	235	236	278

* Weighted average of CONV layers

Website to Summarize Results

- http://eyeriss.mit.edu/benchmarking.html
- Send results or feedback to: eyeriss@mit.edu

ASIC Specs	Input
Process Technology	65nm LP TSMC (1.0V)
Clock Frequency (MHz)	200
Number of Multipliers	168
Core area (mm^2) / multiplier	0.073
On-Chip memory (kB) / multiplier	1.14
Measured or Simulated	Measured
If Simulated, Syn or PnR? Which corner?	n/a

Metric	Units	Input
Name of CNN	Text	AlexNet
# of Images Tested	#	100
Bits per operand	#	16
Batch Size	#	4
# of Non Zero MACs	#	409M
Runtime	ms	115.3
Utilization vs. Peak	**%**	**41**
Power	mW	278
Energy/non-zero MACs	**pJ/MAC**	**21.7**
DRAM access/non-zero MACs	**operands /MAC**	**0.005**

Implementation-Specific Metrics

Different devices may have implementation-specific metrics

Example: FPGAs

Metric		Units	AlexNet
Device		Text	Xilinx Virtex-7 XC7V690T
Utilization	DSP	#	2,240
	BRAM	#	1,024
	LUT	#	186,251
	FF	#	205,704
Performance Density		GOPs/slice	8.12E-04