Created on	22/09/2017	Author	AImotive
Last modified	01/02/2018	Version	1.3
		Document type	Measurement Report

aiWare Benchmark Results

Measurement Report

CONFIDENTIAL

Created on	22/09/2017	Author	AImotive
Last modified	01/02/2018	Version	1.3
		Document type	Measurement Report

Contents

List of tables

Created on	22/09/2017	Author	AImotive
Last modified	01/02/2018	Version	1.3
		Document type	Measurement Report

1 INTRODUCTION

As part of its product portfolio, AImotive offers a unique, application-independent and general NN accelerator IP, aiWare, which is scalable from embedded solutions up to data centers.

aiWare offers a dedicated AI computation core. It is independent from the type of the used neural network (CNN, RNN, and so on). Underpinned by its unique and scalable architecture, aiWare can process a wide range of image input of quality beyond 8K. aiWare also offers fixed and variable precision based on the application need (compilation time). aiWare is accompanied by a complex SDK that can make the necessary transformations to run a network trained in floating point domain on a fixed point aiWare variant. Even in this case—due to AImotive's proprietary, hardware-level scaling technology—the quality degradation remains minimal.

To showcase the capabilities of aiWare NN accelerator, AImotive has prepared an FPGA-based aiWare evaluation kit. As this setup does not include a general-purpose DSP (GPDSP), it allows creating benchmarks for the layers that aiWare directly supports.

Note: As for the not covered layers, which do not require acceleration due to their scarce usage or their low computational needs, aiWare will also host a (GPDSP) in the planned HW prototype implementation.

This document contains the results of the benchmark measurements that AImotive carried out on generic NVIDIA GeForce GTX 1080 Ti GPU and on FPGA-based aiWare 1.0 evaluation kit.

2 ENVIRONMENT

2.1 HW AND SW ENVIRONMENT

Measurements were carried out in the following HW setup:

- GPU: NVIDIA GeForce GTX 1080 Ti, https://www.geforce.co.uk/hardware/10series/geforce-gtx-1080-ti/
- aiWare hardware IP 1.0 deployed on Nallatech 510T Compute Acceleration Card, http://www.nallatech.com/store/fpga-accelerated-computing/pcie-accelerator-cards/nallatech-510t-fpga-computing-acceleration-card/
- CPU: AMD A10-7870K Radeon R7, 12 Compute Cores 4C+8G: 1700 MHz, 64 bits
- Memory size: 30 GiB

The following table lists detailed specifications of used HW:

	GTX 1080 Ti			FPGA
GMAC capacity	5669,888		**GMAC capacity**	163,84000
GPU GFLOP capacity (GFLOP/s)	11339,776		**FPGA GOP capacity (GOP/s)**	327,68
CUDA cores	3584		**Number of Cores**	1
Clock frequency (MHz)	1582		**Clock frequency (MHz)**	160
Memory	11GB		**Memory**	8 GB
Memory bandwidth (GB/s)	484		**Memory bandwidth (GB/s)**	34

Table 1. Specifications of used hardware

The SW environment was made up of the following main components:

- Ubuntu 16.04.3 LTS
- NVIDIA Driver Version: 381.22
- aiWare Evaluation Kit 1.0; AImotive's publicly available NN evaluation SW for aiWare

- cuDNN v5.1
- Caffe 1.0.0
- CUDA Toolkit 8.0
- TensorFlow r1.2

2.2 USED NEURAL NETWORKS

The following public neural networks were used in the measurements:

- googlenet_base
- googlenet_seg
- vgg16_base
- vgg16_seg
- inception_v4_base
- inception_v4_seg
- resnet_50_base
- resnet_50_seg
- inception_resnet_v2_base
- inception_resnet_v2_seg

The base networks (e.g. googlenet_base) are the implementations of the classic convolutional networks but without the classification head, so that any kind of image processing network can be built upon them.

The segmentation networks (e.g. googlenet_seg) are examples of the usage of these network on the task of pixelwise semantic segmentation. In these examples, the base networks are used with a simple deconvolutional segmentation head.

We used both base network and segmentation network. We included segmentation network, as this can provide typical visual output of NN-based object detection. On the other hand, we included base network as that is flexible and allow attaching several kinds of extra layers for different kinds of output (classification, segmentation, bounding box, etc.). Their operations take up about 65-95% of NN operations, depending the architecture of the final network.

The descriptor files of actual neural networks used in this benchmark can be downloaded, see Appendix A – References.

2.3 METHODOLOGY

Measurements were taken with the following limits:

- Input was available in accelerator's local memory.
- Runtime was measured until the result arrived in the accelerator's local memory.
- Only runtime was measured, no power efficiency was considered.

The following operation requirements were applied:

- FPGA GOP requirement = ((int8) multiplications + (int8) additions + (int8) comparisons)/1000000000. GMAC requirement = GOPS/2
- GPU GFLOP requirement = ((float32) multiplications + (float32) additions + (float32) comparisons)/1000000000. GMAC requirement = GFLOPS/2

  Int8 and float32 operations are compared due to the limitations of GPUs; the performance of the NNs are substantially the same on both precisions. Int8 operations result only in a minimal precision loss compared to float32.
- MAC utilization = MAC_busy_time/total_inference_time = NN_MAC_requirement / (total_inference_time * MAC_capacity)

The runtime measurements were done on the above NNs with random weights and input data. The input data and weights were prepared in PC memory before runtime measurement; the output data of the

network was not read back to keep the measure clean from acceleration-independent data movement times. The runtime measurements were done with PC timer on high number of network inferences (50x at GPU with Caffe and 50x at aiWare). The variance of the runtime values was less than 0.3%.

The aim of this benchmark is to compare aiWare and GPU hardware architectures for rendering NN inferences. aiWare is designed for latency constrained, real-time, embedded AI for computer-vision applications, while GPUs were originally designed for real-time 3D graphics rendering.

This benchmark does not include the original resolution of NNs. NNs were run on different input sizes ranging from 640x480 to 4096x2176 pixels (all using 3 input channels). MAC utilization is shown in the results table as it has key importance in embedded applications. MAC utilization is directly proportional with performance and affects power efficiency independently from target technology. MAC utilization is also a good measure on how optimal a hardware architecture is for a given task; which is rendering NN inferences in this benchmark.

2.3.1 GPU

The runtime measurements were done with Caffe as it has provided the highest performance amongst software used in the wider community (we have measured Caffe and TensorFlow).

The performance related settings of the GPU were set for maximum performance (i.e. Prefer Maximum Performance, Fan speed fixed to 100%).

2.3.2 AIWARE

The original Caffe models were converted automatically to an aiWare-proprietary format. The conversion and the runtime measurement were done with aiWare SDK. The aiWare hardware IP was deployed on an off-the-shelf Intel FPGA-based PCIe development board.

3 BENCHMARK

3.1 GENERAL CONSIDERATIONS

The aiWare IP 1.0 on the actual FPGA has less performance then the “aiWare IP 1.0 @5,6TMAC” represented in this benchmark. The FPGA implementation of the aiWare IP has lower clock frequency, lower memory bandwidth and fewer cores than scaled “aiWare IP 1.0 @5,6TMAC” referred in this benchmark.

For “aiWare IP 1.0 @5,6TMAC”, we have selected the aiWare clock frequency and core number to match the GPU’s GMAC capacity (5,6 TMAC in this benchmark). These attributes of the aiWare hardware IP can be linearly scaled to the values used in this benchmark. A high-performance ASIC implementation of the aiWare hardware IP will be available in 2018 for internal testing purposes.

Created on	22/09/2017	Author	AImotive
Last modified	01/02/2018	Version	1.3
		Document type	Measurement Report

3.2 RESULTS

The following table summarizes the main results of our benchmarking:

			GeForce GTX 1080 Ti (5.6TMAC)		aiWare IP			aiWare/ GPU
			Caffe		aiWare IP 1.0 in FPGA @0.16 TMAC		aiWare IP 1.0 @5.6 TMAC	
Network name	Size	GMAC requirement	Measured runtime (ms/ inference)	MAC Utilization	Measured runtime (ms/ inference)	MAC Utilization	AIWARE ESTIMATE runtime (ms/ inference)	MAC Utilization ratio
googlenet base	3x4096x2176	282.09	108.977	45.65%	1789.9	96.19%	51.72	210.70%
googlenet base	3x2560x1920	155.56	60.5815	45.29%	982.84	96.61%	28.40	213.31%
googlenet base	3x2048x1536	99.56	39.5447	44.40%	623.3	97.49%	18.01	219.56%
googlenet base	3x2048x1024	66.37	28.2706	41.41%	413.94	97.87%	11.96	236.35%
googlenet base	3x1920x1056	64.17	28.0511	40.35%	428.26	91.45%	12.38	226.67%
googlenet base	3x1280x704	28.52	15.2601	32.96%	185.08	94.05%	5.35	285.33%
googlenet base	3x800x576	14.58	11.2712	22.82%	106.02	83.96%	3.06	367.91%
googlenet base	3x704x480	10.69	9.72185	19.40%	74	88.21%	2.14	454.64%
googlenet base	3x640x480	9.72	9.47351	18.10%	65.96	89.97%	1.91	497.03%
googlenet seg	3x4096x2176	286.26	-*	-*	1961.34	89.08%	56.68	-*
googlenet seg	3x2560x1920	157.86	86.7387	32.10%	1080.38	89.18%	31.22	277.84%
googlenet seg	3x2048x1536	101.03	56.5917	31.49%	684.38	90.10%	19.78	286.16%
googlenet seg	3x2048x1024	67.36	39.8759	29.79%	454.5	90.45%	13.13	303.62%
googlenet seg	3x1920x1056	65.12	39.1588	29.33%	468.84	84.77%	13.55	289.04%
googlenet seg	3x1280x704	28.94	20.5055	24.89%	202.9	87.06%	5.86	349.74%
googlenet seg	3x800x576	14.80	14.18	18.41%	115.36	78.30%	3.33	425.38%
googlenet seg	3x704x480	10.85	12.0495	15.89%	80.92	81.86%	2.34	515.31%
googlenet seg	3x640x480	9.87	11.6927	14.88%	72	83.64%	2.08	562.00%
vgg16 base	3x4096x2176	2726.61	-*	-*	16871.6	98.64%	487.53	-*
vgg16 base	3x2560x1920	1503.64	-*	-*	9277.64	98.92%	268.09	-*
vgg16 base	3x2048x1536	962.33	179.39	94.61%	5943.86	98.82%	171.76	104.44%
vgg16 base	3x2048x1024	641.55	122.54	92.34%	3961.68	98.84%	114.48	107.04%
vgg16 base	3x1920x1056	620.25	123.40	88.65%	3893.8	97.22%	112.52	109.68%
vgg16 base	3x1280x704	275.67	53.12	91.52%	1708.92	98.46%	49.38	107.57%
vgg16 base	3x800x576	140.97	29.83	83.35%	894.84	96.15%	25.86	115.36%
vgg16 base	3x704x480	103.38	20.96	87.00%	655.58	96.24%	18.94	110.62%
vgg16 base	3x640x480	93.98	19.82	83.65%	592.66	96.78%	17.13	115.71%
vgg16 seg	3x4096x2176	2730.69	-*	-*	17044.06	97.79%	492.51	-*
vgg16 seg	3x2560x1920	1505.89	-*	-*	9371.22	98.08%	270.80	-*
vgg16 seg	3x2048x1536	963.77	198.50	85.63%	6003.62	97.98%	173.48	114.42%
vgg16 seg	3x2048x1024	642.52	135.01	83.94%	4003.42	97.96%	115.68	116.70%

Created on	22/09/2017	Author	AImotive
Last modified	01/02/2018	Version	1.3
		Document type	Measurement Report

			GeForce GTX 1080 Ti (5.6TMAC)		aiWare IP			aiWare/ GPU
			Caffe		aiWare IP 1.0 in FPGA @0.16 TMAC		aiWare IP 1.0 @5.6 TMAC	
Network name	Size	GMAC requirement	Measured runtime (ms/ inference)	MAC Utilization	Measured runtime (ms/ inference)	MAC Utilization	AIWARE ESTIMATE runtime (ms/ inference)	MAC Utilization ratio
vgg16 seg	3x1920x1056	621.18	137.68	79.57%	3934.5	96.36%	113.69	121.10%
vgg16 seg	3x1280x704	276.08	58.92	82.64%	1726.6	97.59%	49.89	118.10%
vgg16 seg	3x800x576	141.18	33.08	75.27%	904.32	95.28%	26.13	126.60%
vgg16 seg	3x704x480	103.53	23.47	77.79%	662.08	95.44%	19.13	122.69%
vgg16 seg	3x640x480	94.12	22.03	75.34%	598.68	95.95%	17.30	127.36%
inception_v4_base	3x4096x2176	1451.05	-*	-*	9577.14	92.48%	276.75	-*
inception_v4_base	3x2560x1920	800.21	-*	-*	5277.62	92.54%	152.50	-*
inception_v4_base	3x2048x1536	512.13	-*	-*	3354.02	93.20%	96.92	-*
inception_v4_base	3x2048x1024	341.42	-*	-*	2235.58	93.21%	64.60	-*
inception_v4_base	3x1920x1056	330.09	-*	-*	2353.74	85.60%	68.01	-*
inception_v4_base	3x1280x704	146.71	85.54	30.25%	1026.9	87.20%	29.67	288.27%
inception_v4_base	3x800x576	75.02	57.35	23.07%	602	76.06%	17.40	329.70%
inception_v4_base	3x704x480	55.01	53.11	18.27%	412.04	81.49%	11.91	446.04%
inception_v4_base	3x640x480	50.01	50.93	17.32%	361.06	84.54%	10.43	488.12%
inception_v4_seg	3x4096x2176	1455.32	-*	-*	9751.28	91.09%	281.78	-*
inception_v4_seg	3x2560x1920	802.57	-*	-*	5373.24	91.16%	155.27	-*
inception_v4_seg	3x2048x1536	513.64	-*	-*	3415.5	91.79%	98.70	-*
inception_v4_seg	3x2048x1024	342.43	-*	-*	2276.7	91.80%	65.79	-*
inception_v4_seg	3x1920x1056	331.06	-*	-*	2394.02	84.40%	69.18	-*
inception_v4_seg	3x1280x704	147.14	90.92	28.54%	1044.78	85.96%	30.19	301.16%
inception_v4_seg	3x800x576	75.24	60.58	21.90%	611.54	75.09%	17.67	342.83%
inception_v4_seg	3x704x480	55.18	55.20	17.63%	419	80.37%	12.11	455.90%
inception_v4_seg	3x640x480	50.16	53.57	16.51%	367.5	83.31%	10.62	504.47%
resnet_50_base	3x4096x2176	687.06	-*	-*	4751.64	88.25%	137.31	-*
resnet_50_base	3x2560x1920	378.89	-*	-*	2623.74	88.14%	75.82	-*
resnet_50_base	3x2048x1536	242.49	105.73	40.45%	1661.66	89.07%	48.02	220.19%
resnet_50_base	3x2048x1024	161.66	73.46	38.82%	1106.5	89.17%	31.97	229.74%
resnet_50_base	3x1920x1056	156.29	72.09	38.24%	1182.34	80.68%	34.17	211.02%
resnet_50_base	3x1280x704	69.46	32.97	37.16%	512.16	82.78%	14.80	222.78%
resnet_50_base	3x800x576	35.52	20.92	29.95%	308.16	70.35%	8.90	234.88%
resnet_50_base	3x704x480	26.05	16.31	28.16%	206.25	77.09%	5.96	273.73%
resnet_50_base	3x640x480	23.68	15.46	27.02%	187.3	77.17%	5.41	285.57%
resnet_50_seg	3x4096x2176	691.42	-*	-*	4926.68	85.66%	142.36	-*

Created on	22/09/2017	Author	AImotive
Last modified	01/02/2018	Version	1.3
		Document type	Measurement Report

			GeForce GTX 1080 Ti (5.6TMAC)		aiWare IP			aiWare/ GPU
			Caffe		aiWare IP 1.0 in FPGA @0.16 TMAC		aiWare IP 1.0 @5.6 TMAC	
Network name	Size	GMAC requirement	Measured runtime (ms/ inference)	MAC Utilization	Measured runtime (ms/ inference)	MAC Utilization	AIWARE ESTIMATE runtime (ms/ inference)	MAC Utilization ratio
resnet_50_seg	3x2560x1920	381.30	-*	-*	2720.34	85.55%	78.61	-*
resnet_50_seg	3x2048x1536	244.03	123.03	34.98%	1723.42	86.42%	49.80	247.05%
resnet_50_seg	3x2048x1024	162.69	85.04	33.74%	1147.68	86.52%	33.16	256.43%
resnet_50_seg	3x1920x1056	157.29	83.44	33.25%	1222.88	78.50%	35.34	236.12%
resnet_50_seg	3x1280x704	69.90	38.57	31.97%	530	80.50%	15.32	251.83%
resnet_50_seg	3x800x576	35.75	24.54	25.69%	317.92	68.63%	9.19	267.16%
resnet_50_seg	3x704x480	26.21	18.67	24.76%	213.04	75.10%	6.16	303.34%
resnet_50_seg	3x640x480	23.83	17.70	23.75%	193.78	75.06%	5.60	316.01%
inception_resnet_v2_base	3x4096x2176	1577.70	-*	-*	10572.4	91.08%	305.51	-*
inception_resnet_v2_base	3x2560x1920	870.06	-*	-*	5828.76	91.11%	168.43	-*
inception_resnet_v2_base	3x2048x1536	556.84	-*	-*	3701.14	91.83%	106.95	-*
inception_resnet_v2_base	3x2048x1024	371.22	-*	-*	2463.7	91.97%	71.19	-*
inception_resnet_v2_base	3x1920x1056	358.90	-*	-*	2650.18	82.66%	76.58	-*
inception_resnet_v2_base	3x1280x704	159.51	154.56	18.20%	1166.88	83.43%	33.72	458.37%
inception_resnet_v2_base	3x800x576	81.57	102.32	14.06%	701.88	70.93%	20.28	504.49%
inception_resnet_v2_base	3x704x480	59.82	87.47	12.06%	478	76.38%	13.81	633.30%
inception_resnet_v2_base	3x640x480	54.38	84.35	11.37%	425.92	77.93%	12.31	685.36%
inception_resnet_v2_seg	3x4096x2176	1581.97	-*	-*	10746.2	89.85%	310.53	-*
inception_resnet_v2_seg	3x2560x1920	872.41	-*	-*	5924.02	89.88%	171.18	-*
inception_resnet_v2_seg	3x2048x1536	558.34	-*	-*	3762.6	90.57%	108.73	-*
inception_resnet_v2_seg	3x2048x1024	372.23	-*	-*	2504.34	90.72%	72.37	-*
inception_resnet_v2_seg	3x1920x1056	359.87	-*	-*	2690.44	81.64%	77.74	-*
inception_resnet_v2_seg	3x1280x704	159.94	160.50	17.58%	1184.56	82.41%	34.23	468.88%
inception_resnet_v2_seg	3x800x576	81.79	104.10	13.86%	711.04	70.21%	20.55	506.65%
inception_resnet_v2_seg	3x704x480	59.98	89.63	11.80%	484.94	75.49%	14.01	639.65%
inception_resnet_v2_seg	3x640x480	54.53	85.73	11.22%	432	77.04%	12.48	686.75%

Table 2. Benchmarking results

* Insufficient memory on GPU

Created on	22/09/2017	Author	AImotive
Last modified	01/02/2018	Version	1.3
		Document type	Measurement Report

4 CONCLUSION

Measurements have shown that AImotive's aiWare solution beats even the most powerful commercially available GPUs in terms of performance and efficiency, with ratios ranging from about 104% to even 686%.

Our dedicated NN accelerator HW provides exceptional results both in terms of performance and MAC utilization, due its dedicated architecture providing support for NN layers that comprises 80-99% of processing power requirements of a typical NN.

This benchmark shows the viability of AImotive's aiWare solution for NN acceleration.

Created on	22/09/2017	Author	AImotive
Last modified	01/02/2018	Version	1.3
		Document type	Measurement Report

5 APPENDIX A – REFERENCES

- The actual NN descriptor files used in this benchmark in Caffe `.prototxt` format can be downloaded from here.

Created on	22/09/2017	Author	AImotive
Last modified	01/02/2018	Version	1.3
		Document type	Measurement Report

6 APPENDIX B – TERMS AND ABBREVIATIONS

The following terms and abbreviations are used in this document:

Terms	Description
ASIC	Application-Specific Integrated Circuit
CNN	Convolutional Neural Network
FPGA	Field-Programmable Gate Array
GFLOP	Giga Floating Point Operations
GMAC	Giga Multiply-Accumulate Operations
GOP	Giga Operations
GPDSP	General-Purpose Digital Signal Processor
GPU	Graphics Processing Unit
RNN	Recurrent Neural Network

Created on	22/09/2017	Author	AImotive
Last modified	01/02/2018	Version	1.3
		Document type	Measurement Report

7 APPENDIX C – DOCUMENT HISTORY

Version	Date	Modified by	Description of the modification
1.0	04/10/2017	Tamas Forgacs Andras Juhasz	Initial public version
1.1	04/12/2017	Tamas Forgacs Andras Juhasz	Minor corrections in Table 2. Benchmarking results
1.2	12/12/2017	Tamas Forgacs Andras Juhasz	Measurements in Table 2. Benchmarking results extended with new NNs
1.3	01/02/2018	Tamas Forgacs Andras Juhasz	Measurements in Table 2. Benchmarking results extended with new NNS

Created on	22/09/2017	Author	AImotive
Last modified	01/02/2018	Version	1.3
		Document type	Measurement Report

Disclaimer