FPGA-based Neural Network Accelerators

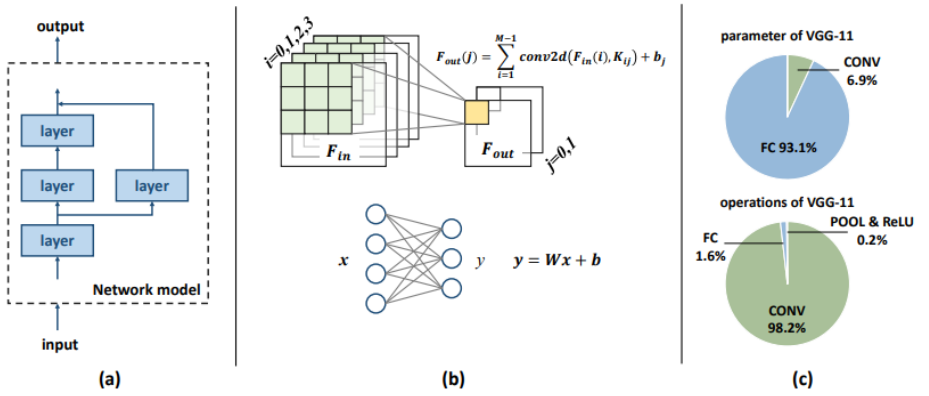
2020

Ando Ki, Ph.D.
adki@future-ds.com

Table of contents

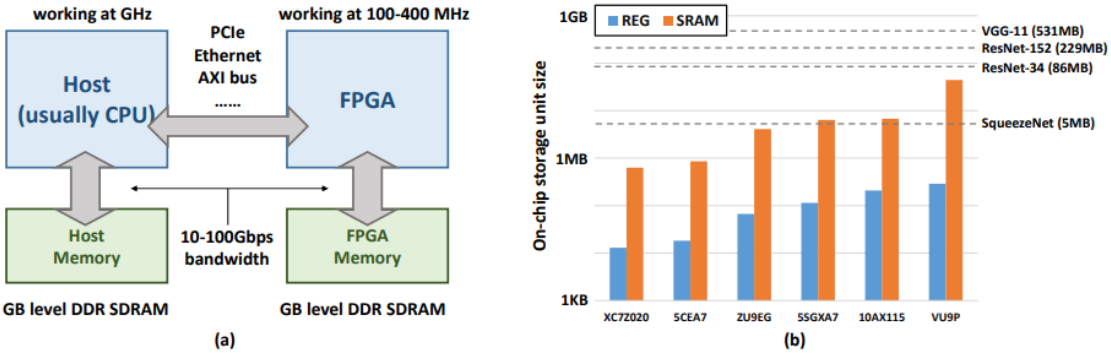
Neural Network

- (a) Computation graph of a neural network model.
- (b) CONV (convolution) and FC (fully connected) layers in NN (neural network) model.
- (c) CONV and FC layers dominate the computation and parameter of a typical NN model: VGG11.



FPGA memory issues

- (a) A typical structure of an FPGA-based NN accelerator.
- (b) Gap between NN model size and the storage unit size on FPGAs



FPGA resource issues

- Pure logic case v.s. DSP-utilized case
- multiply & add: A*B+C
- Note fp32:fixed32 for logic case
- By compressing 32-bit floating-point number to 8-bit fixed-point number, the multiplier and the adder are scaled down to about 1/10 and 1/50 respectively.

	Xilinx Logic				Xilinx DSP			Altera DSP	
	multiplier		adder		multiply & add			multiply & add	
	LUT	FF	LUT	FF	LUT	FF	DSP	ALM	DSP
fp32	708	858	430	749	800	1284	2	1	1
fp16	221	303	211	337	451	686	1	213	1
fixed32	1112	1143	32	32	111	64	4	64	3
fixed16	289	301	16	16	0	0	1	0	1
fixed8	75	80	8	8	0	0	1	0	1
fixed4	17	20	4	4	0	0	1	0	1

Xilinx XCKU060
Altera Arria 10 GX1150

Model compress approaches

- Data quantization to reduce bit-width of weights and activations
  - Lower precision floating-point
  - Fixed-point
  - Binarized neural network (BNN)
- Weight reduction by reducing the number of weights
  - Decomposition of matrix operation
  - Pruning, e.g, removing the zeros in weights or small absolute values
  - Normalization of weight during training

Efficient architecture

- Low bit-width computation unit
  - fixed-point
  - binalized neural network
- Fast convolution method
  - Discrete Fourier transformation based
- Frequency optimization
  - faster operating frequency of FPGA
- Explore parallelism
  - Loop unrolling
  - Pipelining

Benefits/Advantages of FPGA on Deep Learning

- FPGAs have product lifecycles of 15 years.
- High performance per Watt and low latency make it suitable for real-time embedded applications.
- The FPGA logic can be shaped to match any network architecture.
- Performance, cost and power will define the FPGA of choice.
- Future proof and scalable solution as the FPGA architecture can be re-configured for future neural networks.
- The deep learning core can be easily integrated with other CPU's, vision functionality and connectivity.

8

References

- A Survey of FPGA-Based Neural Network Inference Accelerator, arXiv:1712.08934v3 [cs.AR] 6 Dec 2018
- Optimizing FPGA-based Accelerator Design for Deep Convolutional Neural Networks