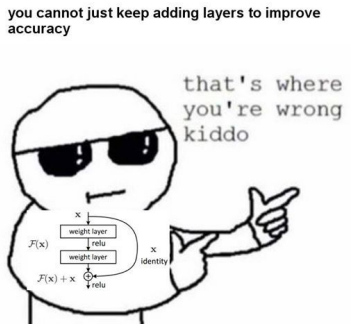
FINN:
Framework for Inference on Binarized Nets

Umuroglu, Fraser, Gambardella, Blott, Leong, Jahre, and Vissers

Skanda Koppula

6.888 Fall 2017 Paper Presentation



Outline

- Motivation
  - Accuracy experiments on BNNs
- Key Contributions
  - Estimating upper bound throughput
  - BNN HW optimizations: Max-Pool OR, Pop-count ACCUM, Batchnorm Thres.
  - Architecture: High level, MTVU, PE dataflow, SWU, Folding/Interleaving
- Evaluation
  - Throughput, power, and resource efficiency experiments
- Critique
- Extensions

Motivation

- Binarized neural networks → fewer computation and memory demands
- Comparable accuracy to full-precision networks!
- Using hardware designed for FP/uint32_t ops is wasteful...

Motivation

- Binarized neural networks → fewer computation and memory demands
- Comparable accuracy to full-precision networks!
- Using hardware designed for FP/uint32_t ops is wasteful...

FPGA-based accelerator for BNNs

binary weights $(-1, 1)$, binary activations, $\log_2$(neuron-fan-in) accum.

activation: `sign(x) = { +1 if x > 0 else -1 }`

Pre-HW BNN Experiments

Explore accuracy/precision trade-off:

- Smaller network with Fl. Pt. NN or larger network with a BNN?

Table 1: Accuracy results - BNN vs NN. (32-bit Fl. Pt.)

Neurons/layer	Binary Err. (%)	Float Err. (%)	# Params	Ops/frame
128	6.58	2.70	134,794	268,800
256	4.17	1.78	335,114	668,672
512	2.31	1.25	932,362	1,861,632
1024	1.60	1.13	2,913,290	5,820,416
2048	1.32	0.97	10,020,874	20,029,440
4096	1.17	0.91	36,818,954	73,613,312

Pre-HW BNN Experiments

Explore accuracy/precision trade-off:

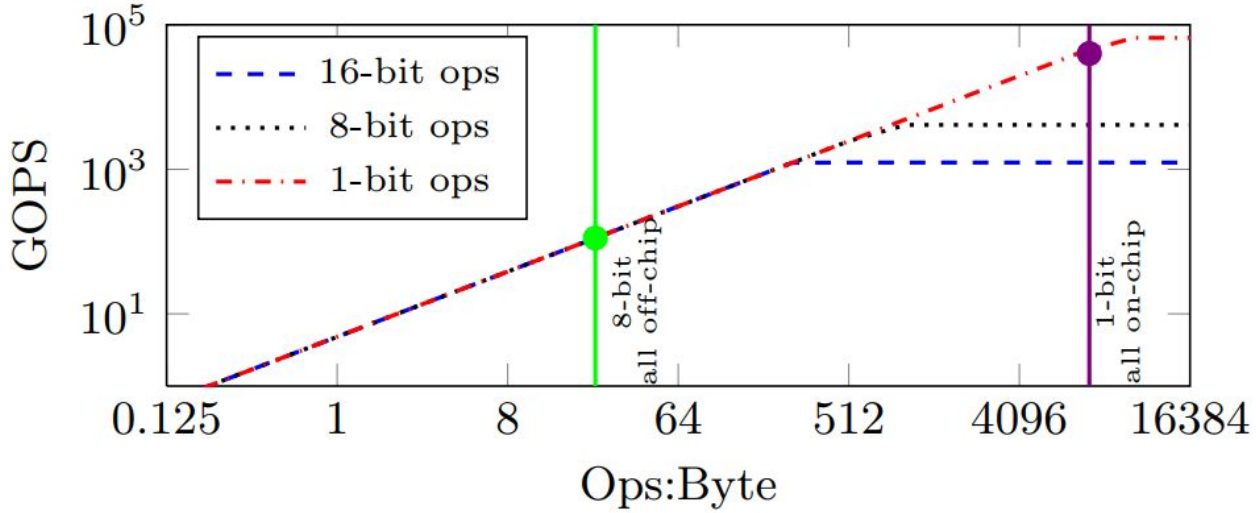
- Smaller network with Fl. Pt. NN or larger network with a BNN?

Table 1: Accuracy results - BNN vs NN. (32-bit Fl. Pt.)

Neurons/layer	Binary Err. (%)	Float Err. (%)	# Params	Ops/frame
128	6.58	2.70	134,794	268,800
256	4.17	1.78	335,114	668,672
512	2.31	1.25	932,362	1,861,632
1024	1.60	1.13	2,913,290	5,820,416
2048	1.32	0.97	10,020,874	20,029,440
4096	1.17	0.91	36,818,954	73,613,312

Same level accuracy ⇒ 2-11x more parameters and operations

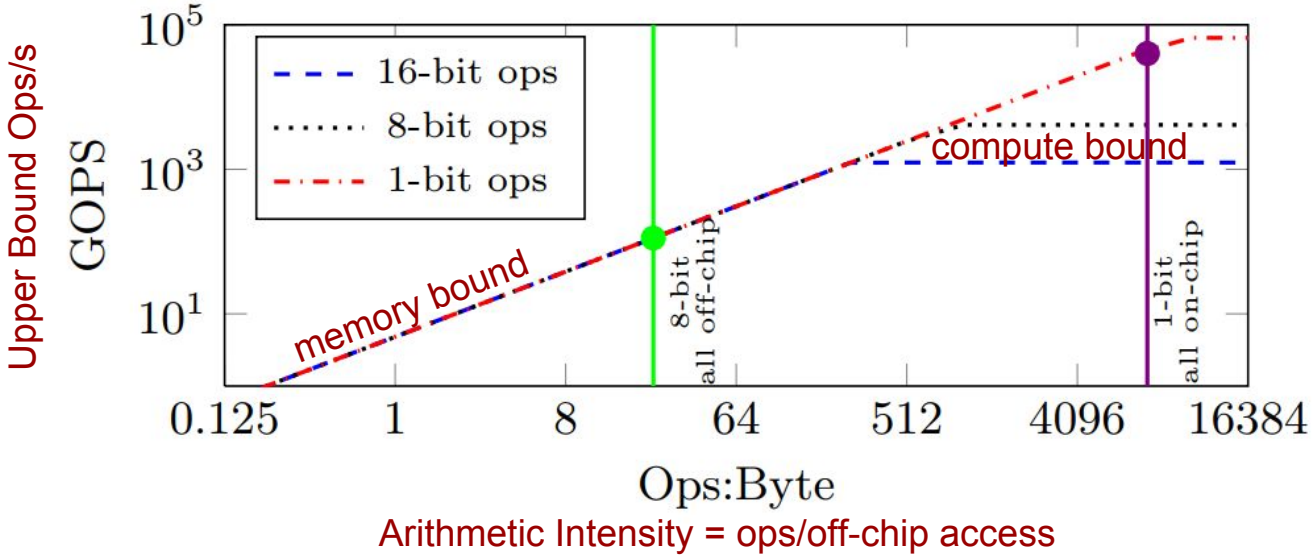
Estimating Upper Bound BNN Throughput



Roofline Model:

- Xilinx Zync ZC706, 350 MHz, 4.5GB/s off-chip bandwidth
- 2.5 LUTs/1-bit op; 40 LUTs/8-bit op; 8 LUTs and 1 DSP/16-bit op

Estimating Upper Bound BNN Performance



Roofline Model:

- We can reach the compute-bound limit b/c smaller params (!)
- We can expect higher throughput for BNN-based architectures!

Optimizations

Addition with -1 and +1 doesn’t require addition...

Signed Accum:
759 LUTs and 84 FFs

Pop-count Accum:
376 LUTs and 29 FFs

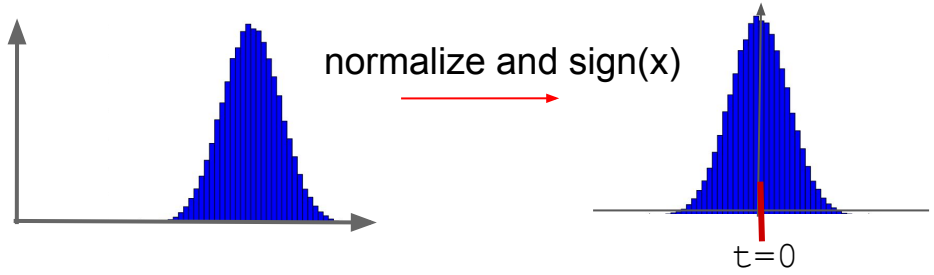
Optimizations

Addition with -1 and +1 doesn't require addition...

Signed Accum:
759 LUTs and 84 FFs

Pop-count Accum:
376 LUTs and 29 FFs

Batchnorm + activation is one compare… (!)



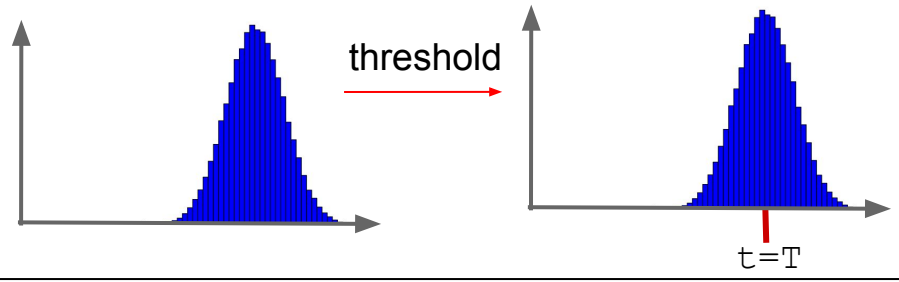
Optimizations

Addition with -1 and +1 doesn't require addition...

Signed Accum:
759 LUTs and 84 FFs

Pop-count Accum:
376 LUTs and 29 FFs

Batchnorm + activation is one compare… (!)



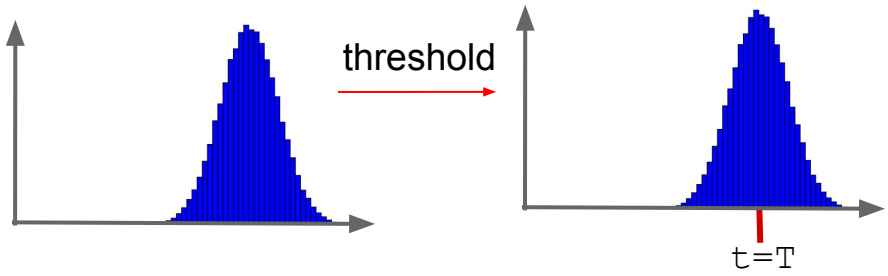
Optimizations

Addition with -1 and +1 doesn't require addition...

Signed Accum:
759 LUTs and 84 FFs

Pop-count Accum:
376 LUTs and 29 FFs

Batchnorm + activation is one compare… (!)



Max-Pool is Boolean OR

Pool and activation:

```
a = (max(a_1, a_2) > T)
```

is same as

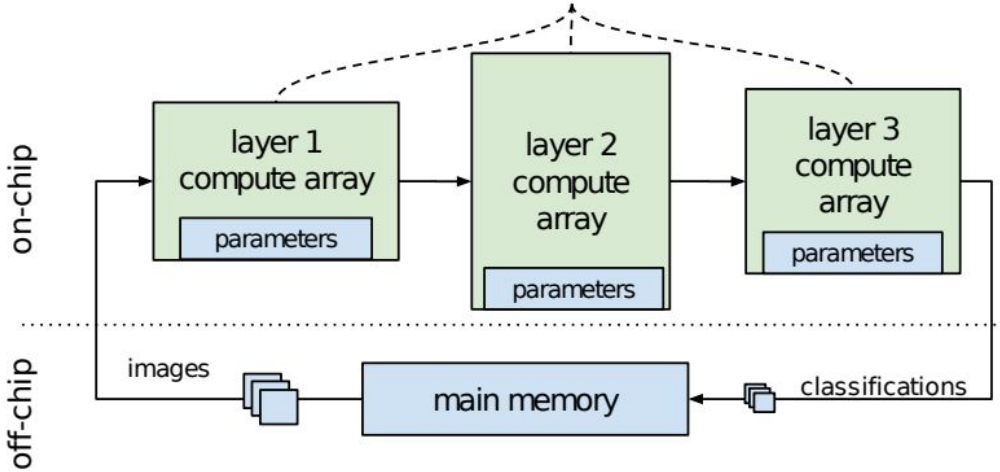
```
a = (a_1 > T || a_2 > T)
a = (b_1 || b_2)
```

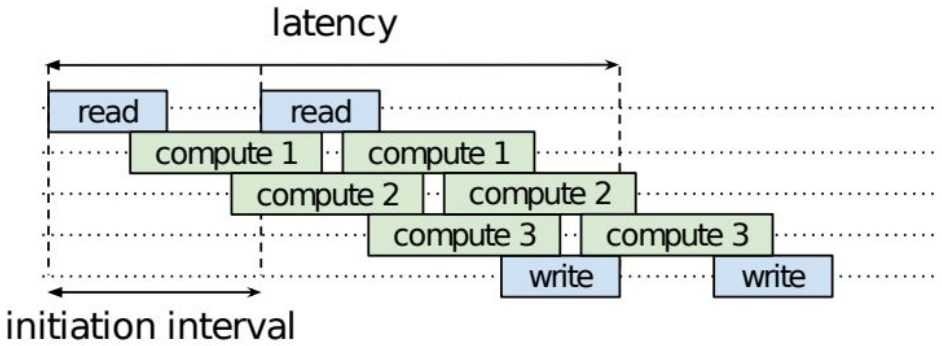
Architecture

Heterogeneous streaming architecture

→ compute on partial outputs

→ different sizing of compute arrays





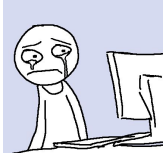
What is a ‘compute array’?

Each layer is evaluated with a “MVTU”

Weight matrix is divided among PEs

Conv becomes matrix multiply (!!!!)

- Sliding window transform on input



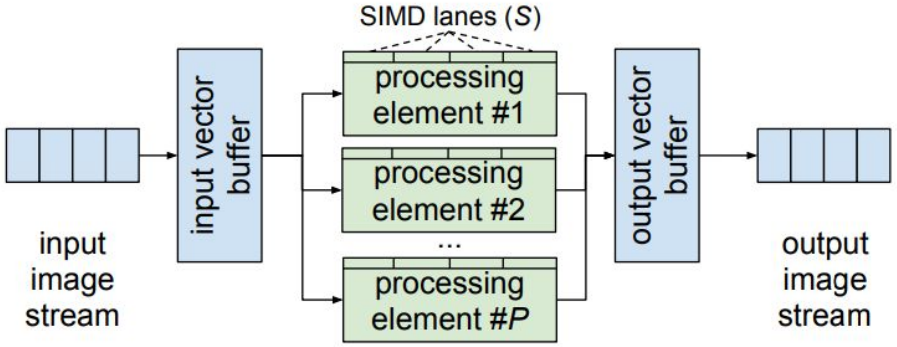
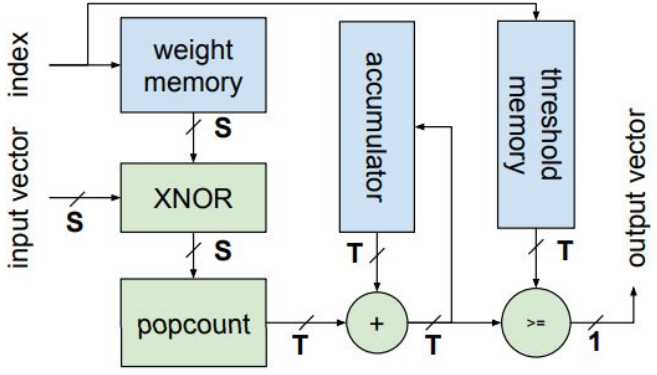


Figure 5: Overview of the MVTU. = compute array
= matrix multiplier



What is a ‘compute array’?

- Interleave channels in input for streaming gains
- Folding to change area/latency trade-off

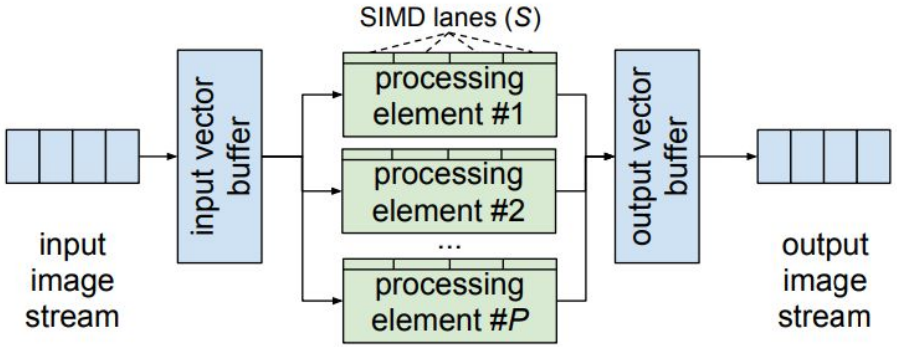
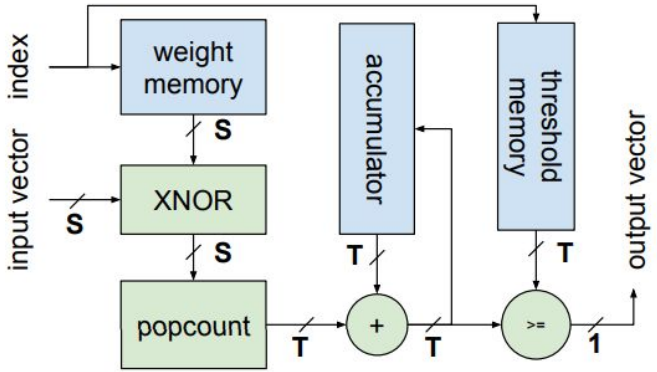


Figure 5: Overview of the MVTU. = compute array
= matrix multiplier



Evaluation

- Networks and Datasets:
  - **S**mall **FC**N: 3-deep, 256 wide, MNIST
  - **L**arge **FC**N: 3-deep, 1024 wide, MNIST
  - **CNV**: VGG-16, CIFAR-10, SVHN
    - “first and last layers” not binarized

- Two usage scenarios tested:
  - **max**: maximize throughput
  - **fix:** fixed throughput requirement, minimize area and, relatepower

Evaluation

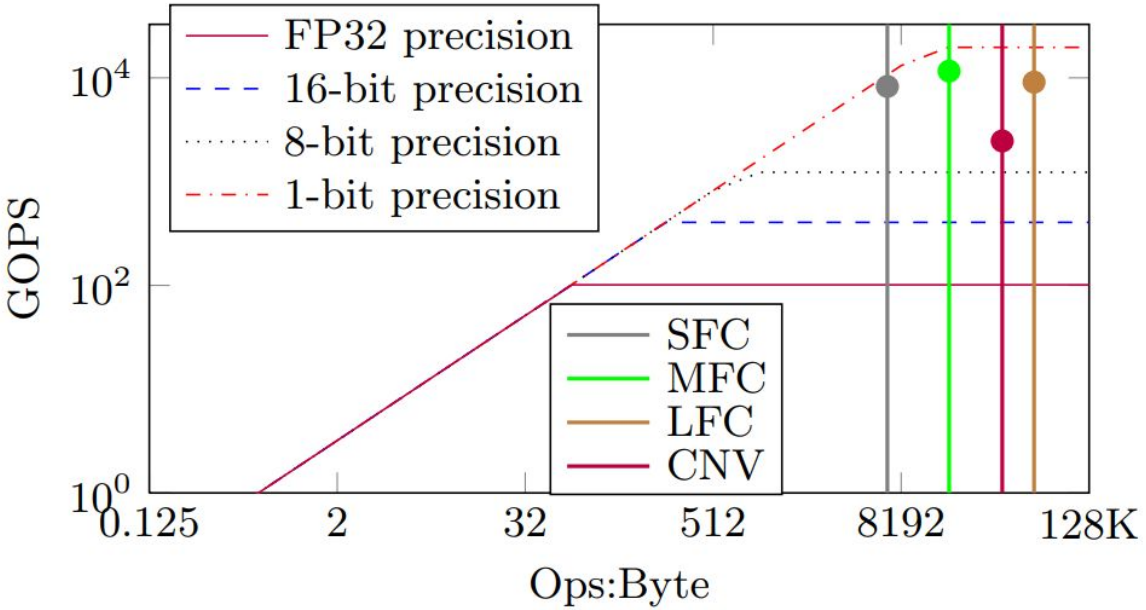
Table 3: Summary of results from FINN 200 MHz prototypes.

Name	Thr.put (FPS)	Latency (µs)	LUT	BRAM	P_{chip} (W)	P_{wall} (W)
SFC-max	12361 k	0.31	91131	4.5	7.3	21.2
LFC-max	1561 k	2.44	82988	396	8.8	22.6
CNV-max	21.9 k	283	46253	186	3.6	11.7
SFC-fix	12.2 k	240	5155	16	0.4	8.1
LFC-fix	12.2 k	282	5636	114.5	0.8	7.9
CNV-fix	11.6 k	550	29274	152.5	2.3	10

Evaluation

- Op measured = XNOR-popcount op

- SFC is memory bound
- LFC/MFC is close to compute bound
- CNV is architecture bounded….



Evaluation

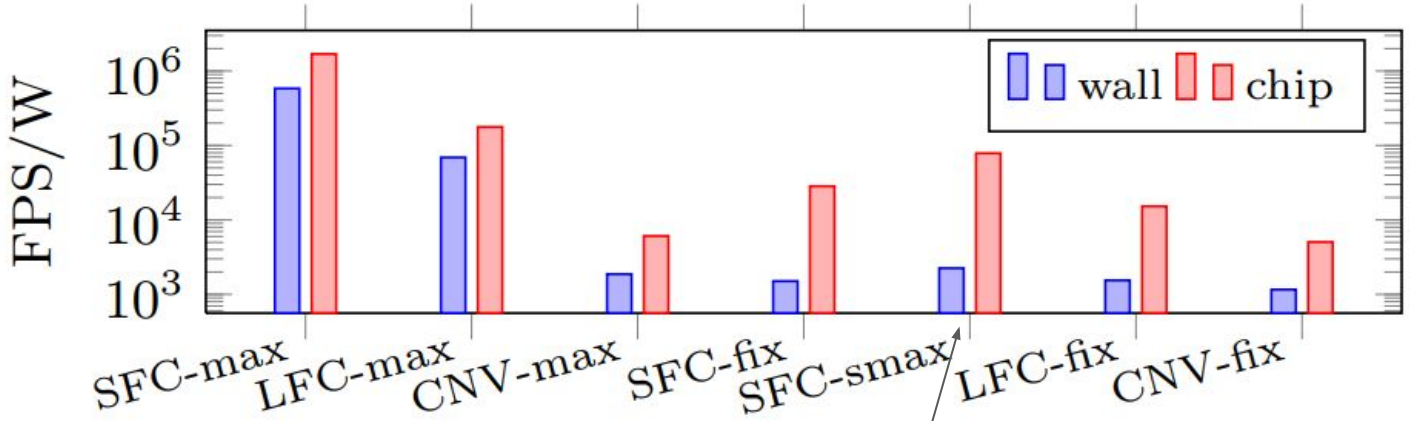


Figure 10: Prototype energy efficiency.

same area as SFC-max, clocked low enough to match SFC-fix throughput

Should we prefer highly parallel design at low clock frequency?
Or prefer less parallel design at high clock frequency?

Strengths of Paper (and Summary)

- You can make some effective algorithmic optimizations for speedy inference when you are hanging out in a two-bit universe
- Demonstrated **very high** throughput inference accelerator for neural networks
  - Apparently, fastest inference on a single FPGA of comparable LUT/FF count
- Real (not simulated) power numbers!

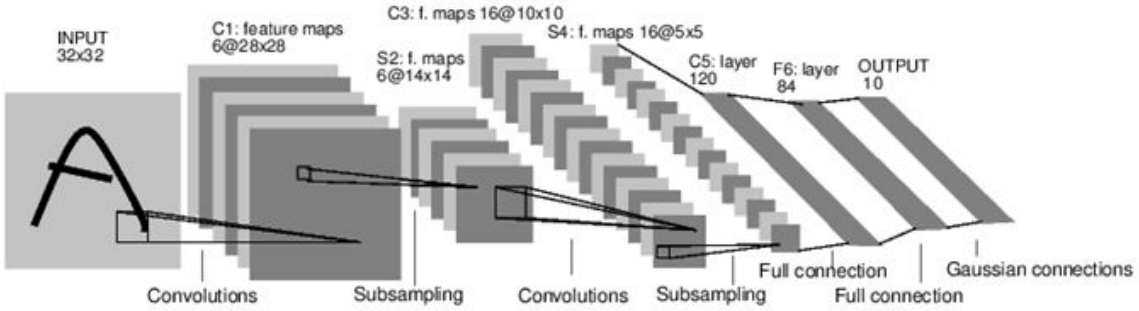
Weaknesses and Possible Extensions

- Did not address how to pad with -1/1
- How do they support first and last non-binarized layers!?
- What are the baseline 4/8/16-bit Fx. Pt. / Fl. Pt. FPS/Watt?
  - Should be fixing target accuracy, and comparing throughput...
- What are the benchmarks on other real world networks?
- No concrete accuracy metrics of end-to-end system
- Re-use existing HW to overload instructions on unsigned ints/bit vectors?
- Lowering Conv to Matrix Multiply was a poor choice
  - “Sliding Window Unit” speculated bottleneck
  - CNV is “architecture bound” performance

Questions?

deep learning is truly in vogue

(Capsule Nets are REAL!!!!)



A Full Convolutional Neural Network (LeNet)