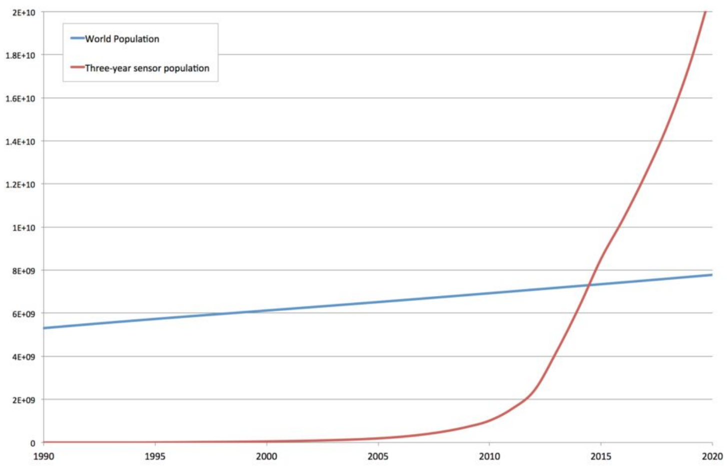
Smart Embedded Vision with Quantized Neural Networks

zsc@megvii.com

Face++ 旷视

Why Smart Embedded Vision?



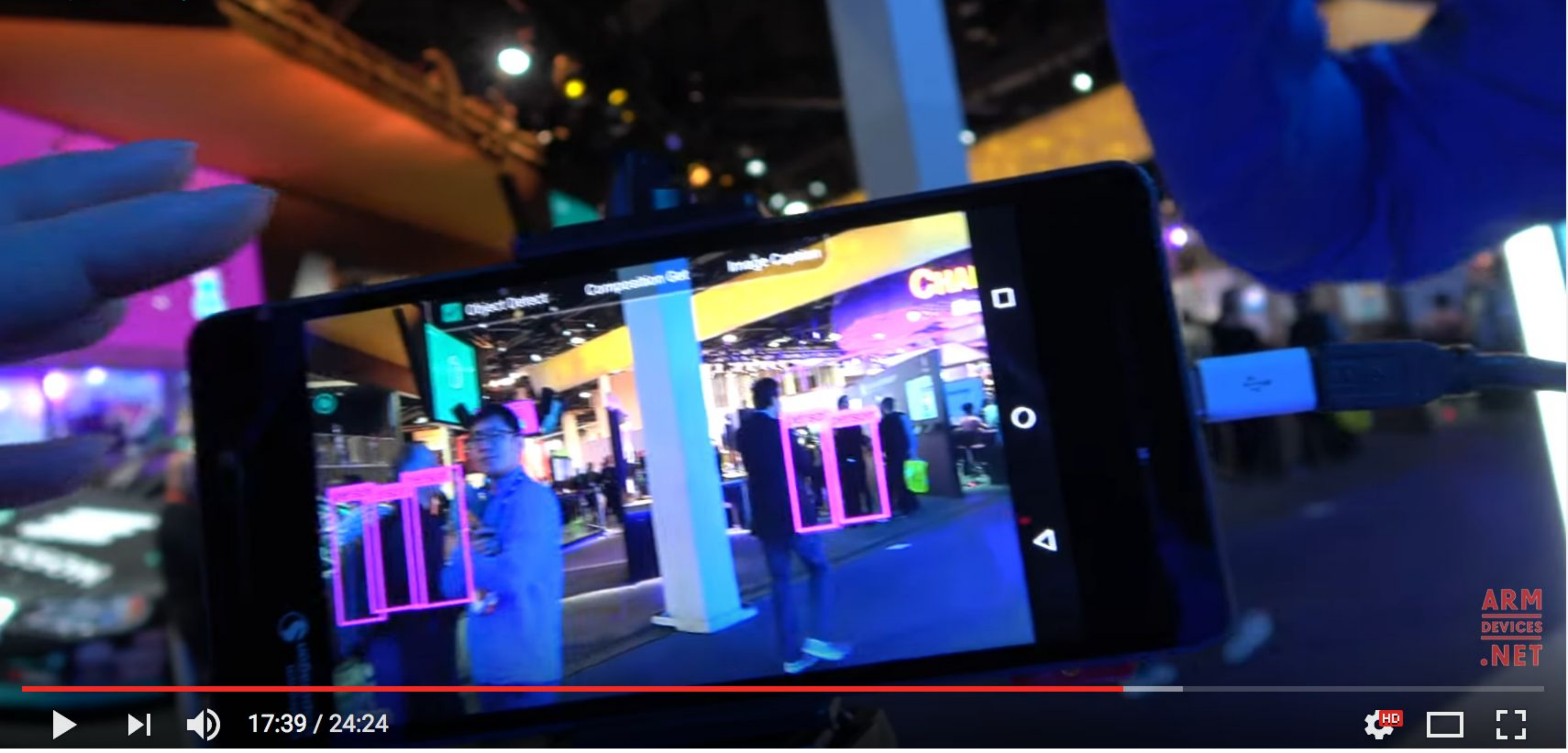
- Human Consumption
- Human Decision Making
- Autonomous Action

-- Chris Rowen *MIPS, Tensilica*

Face++ 旷视

SnapDragon 835, CES 2017

https://www.youtube.com/watch?v=ZI0EY2rzlJI



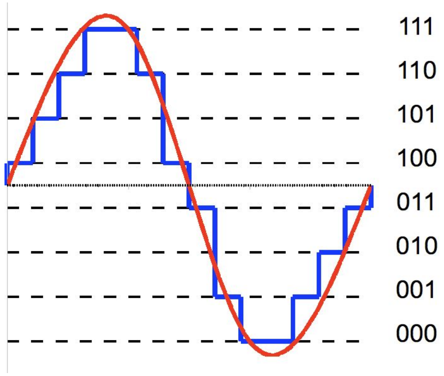
Challenges

- Vast amount of computations
- Power Consumption and Cooling

	Full load (TDP Watts)	FLOPS	Source
ARM (Snapdragon 835)	4.5 (2 (CPU) + 2 (GPU) + 0.5)	<0.06T (FP32), 4+4 cores	3rd party
TX1 (module)	15	1T (FP16)	NVidia
FPGA (7030)	7.5 (5.8 (chip) + 1 (DDR) + 0.7(power))	0.9T (2w2f)	Face++

Quantized Neural Networks (QNNs)

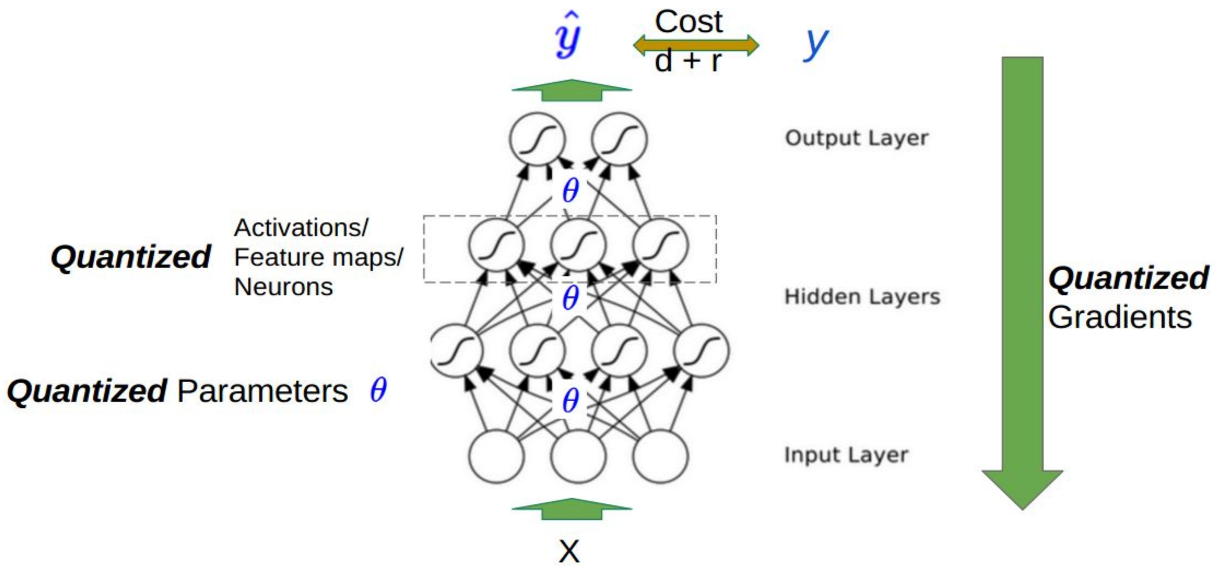
Parameters/Activations/Gradients are quantized to discrete values.



Uniform Quantization

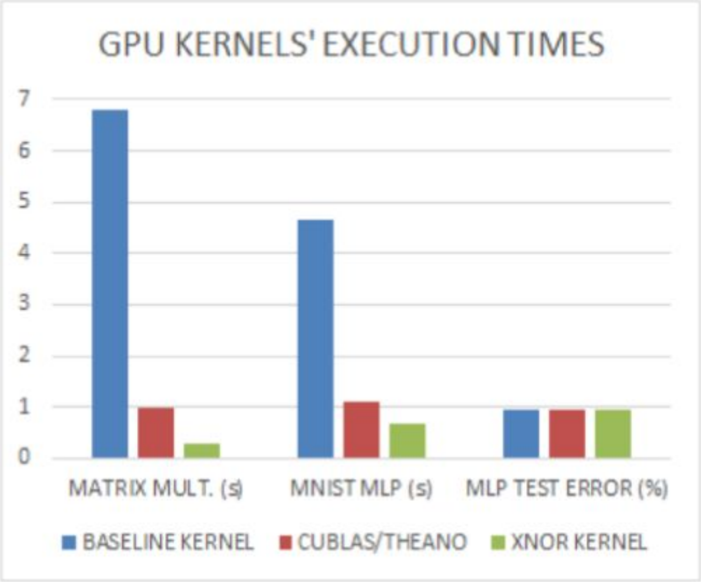
$$Q(x) = \Delta \cdot \left\lfloor \frac{x}{\Delta} + \frac{1}{2} \right\rfloor = \Delta \cdot \text{floor}\left(\frac{x}{\Delta} + \frac{1}{2}\right)$$

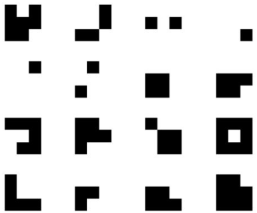
Quantized Neural Network Training



Impact of Quantization

- Pro
  - Can exploit bitwise operations for speeding up computations.
    - XNOR-popcnt kernel instead of multiply-add
  - Smaller storage size and memory footprint
- Con
  - Often less accurate in predictions, especially when bitwidth less than 4





3-by-3 filters in a QNN, the weights are 1-bit hence black and white.

Figures from Hubara et. al. https://arxiv.org/pdf/1609.07061.pdf

QNN at Megvii (Face++)

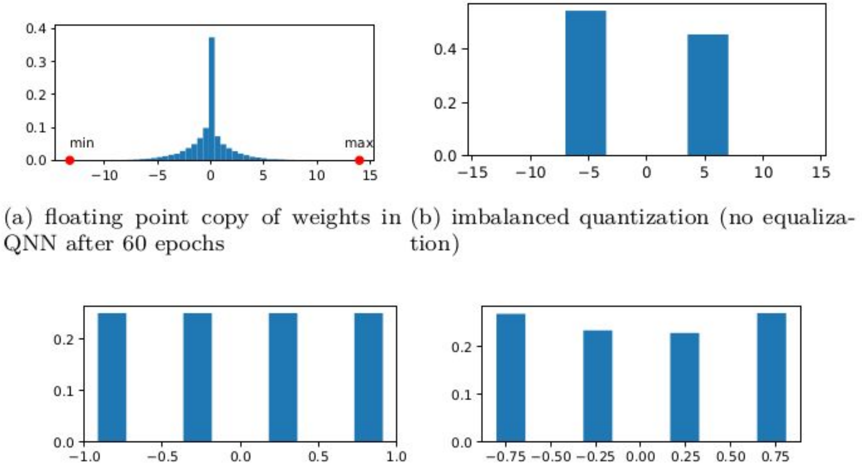
- DoReFa-net https://arxiv.org/abs/1606.06160
  - Stochastic Quantization of Gradients for ImageNet
- Quantization of RNN https://arxiv.org/abs/1611.10176
- Quantization of FCN http://cn.arxiv.org/pdf/1612.00212v1
- Balanced Quantization https://arxiv.org/abs/1706.07145
  - State-of-the-art in 4-bit quantization of GoogleNet/ImageNet and RNN/PTB

DoReFa-Net: Training Low Bitwidth Convolutional Neural Networks with Low Bitwidth Gradients

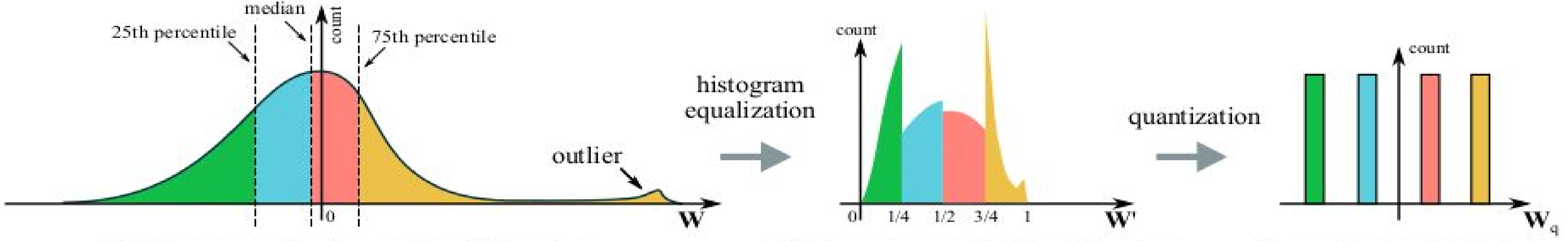
- Uniform stochastic quantization of gradients
  - 6 bit for ImageNet, 4 bit for SVHN
- Simplified scaled binarization: only scalar
  - Forward and backward multiplies the bit matrices from different sides.
  - Using scalar binarization allows using bit operations
- Floating-point-free inference even when with BN
  - Comparison with floating point thresholds can be scaled to be comparison with integers
- Future work
  - BN requires FP computation during training
  - Require FP weights for accumulating gradients

Effective Quantization Methods for Recurrent Neural Networks 2016
Balanced Quantization: An Effective and Efficient Approach to Quantized Neural Networks 2017

Model	weight-bits	activation-bits	PPW	
			balanced	unbalanced
LSTM	2	2	152	164
LSTM	2	3	142	155
LSTM (Hubara et al., 2016a)	2	3		220
LSTM (Hubara et al., 2016a)	4	4		100



(a) floating point copy of weights in QNN after 60 epochs (b) imbalanced quantization (no equalization)

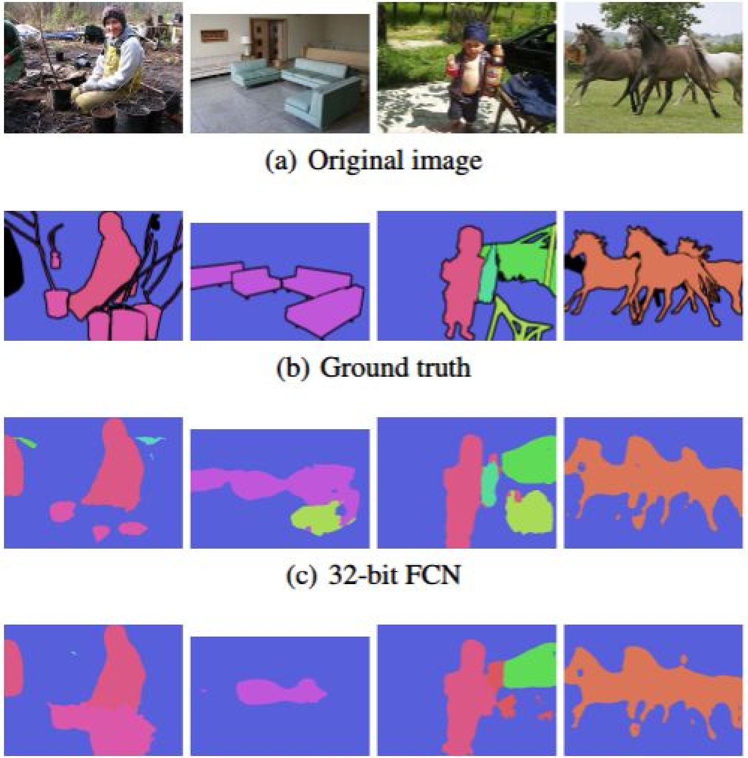


(a) historgram of floating-point weight values (b) histogram-equalized weight values (c) quantized weight values

Training Bit Fully Convolutional Network for Fast Semantic Segmentation 2016

bit-width (W / A)	mean IoU	Complexity
32 / 32	69.8%	-
8 / 8	69.8%	64
4 / 4	68.6%	16
3 / 3	67.4%	9
2 / 2	65.7%	4
1 / 4	64.4%	4
4 / 1	diverge	4
1 / 2	62.8%	2

Table 5: Results of different bit-width allocated to weight and activation on PASCAL VOC 2012 val set.



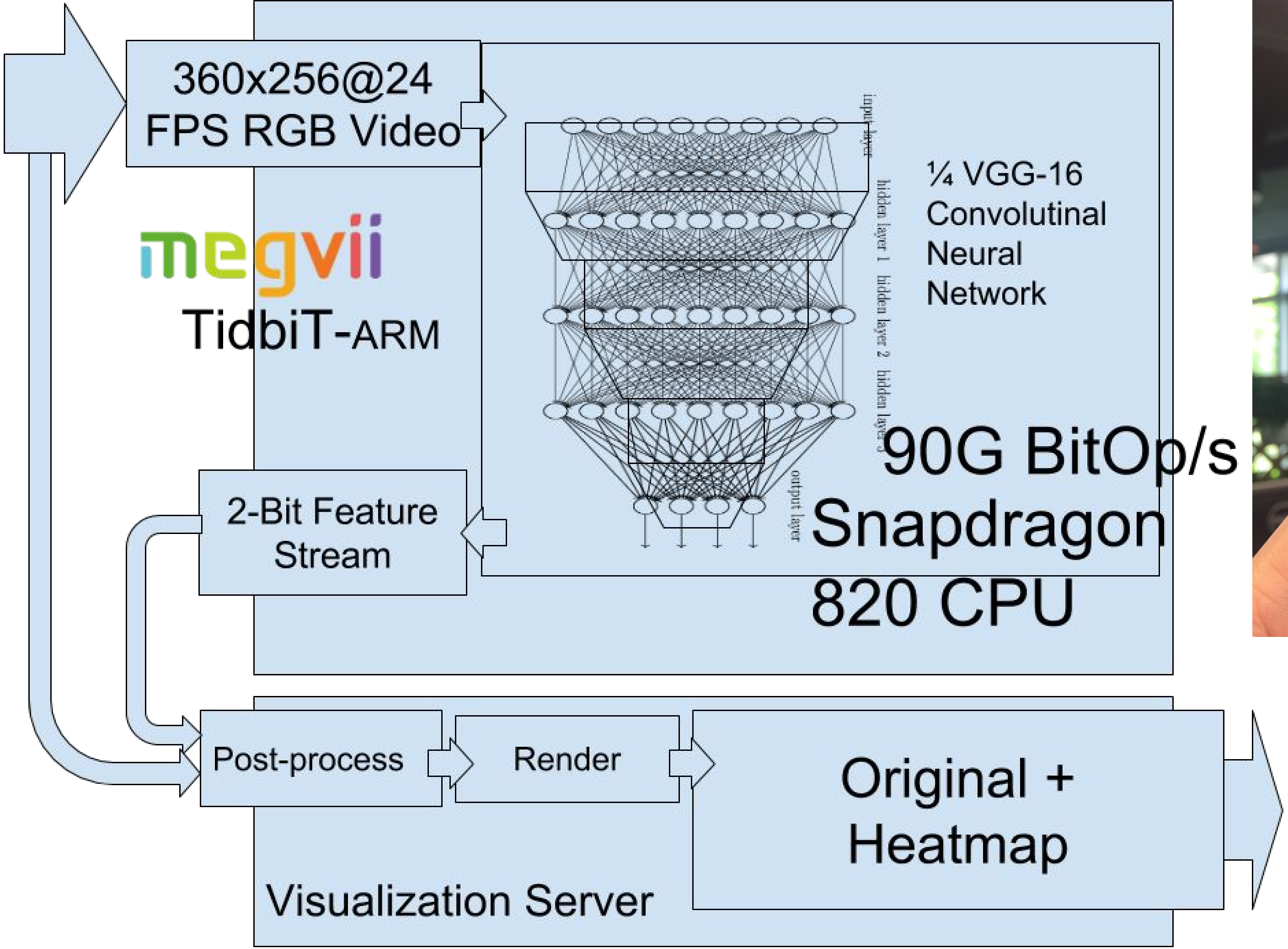
(a) Original image

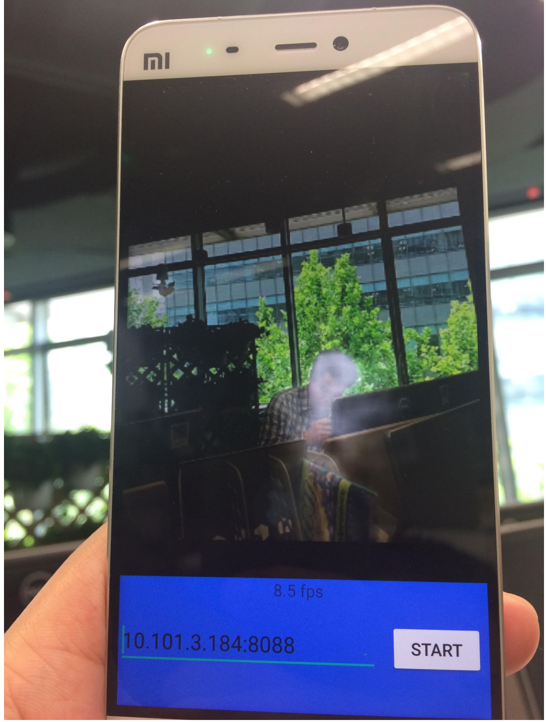
(b) Ground truth

(c) 32-bit FCN

(d) 2-bit BFCN

Figure 4: Examples on PASCAL VOC 2012.

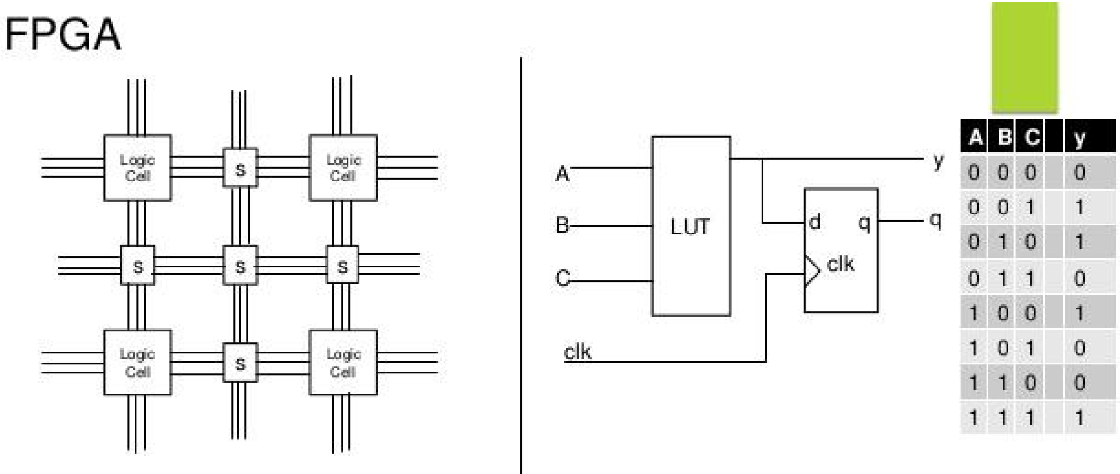




Demo at CVPR 2016

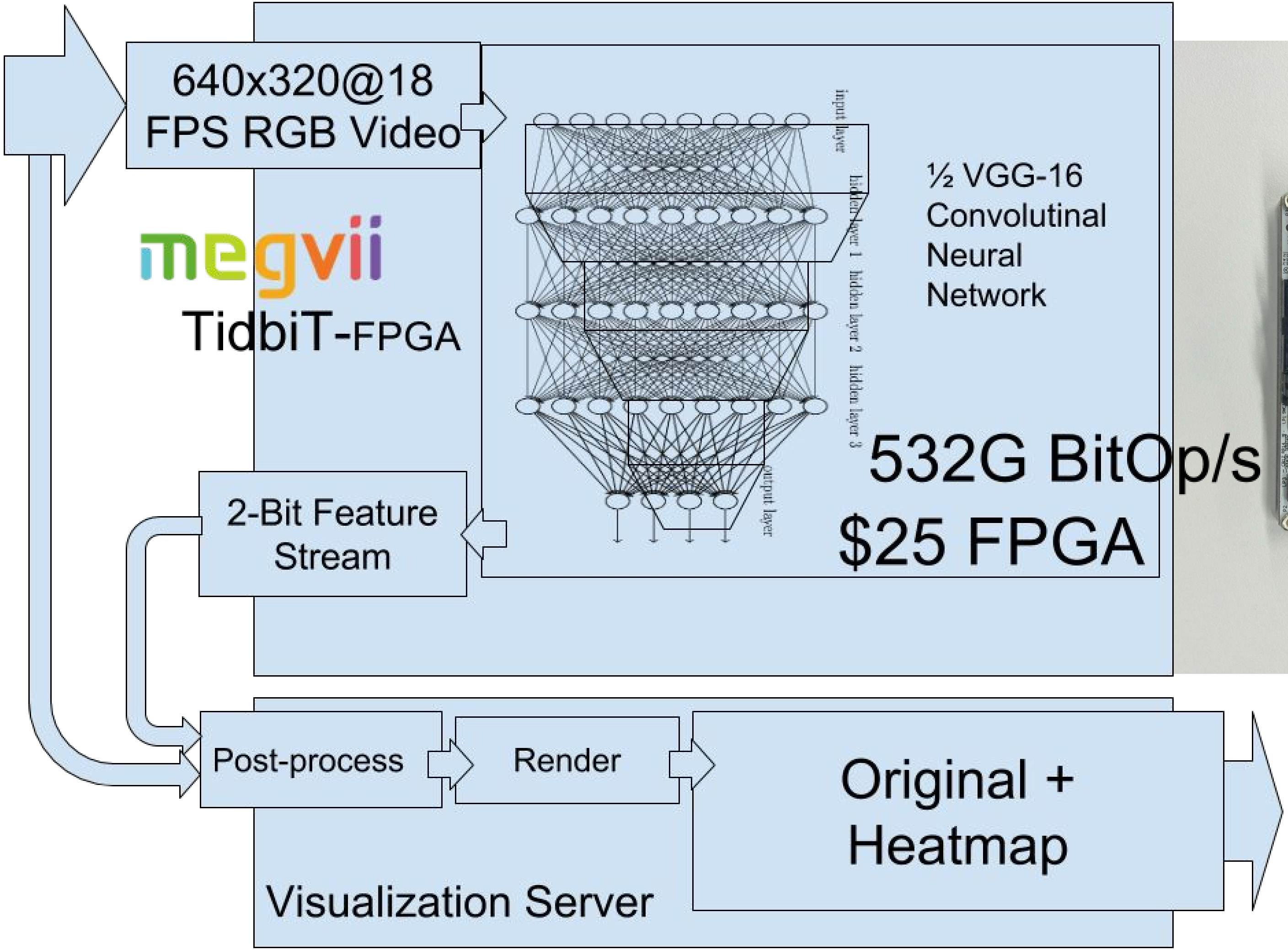
QNN on FPGA

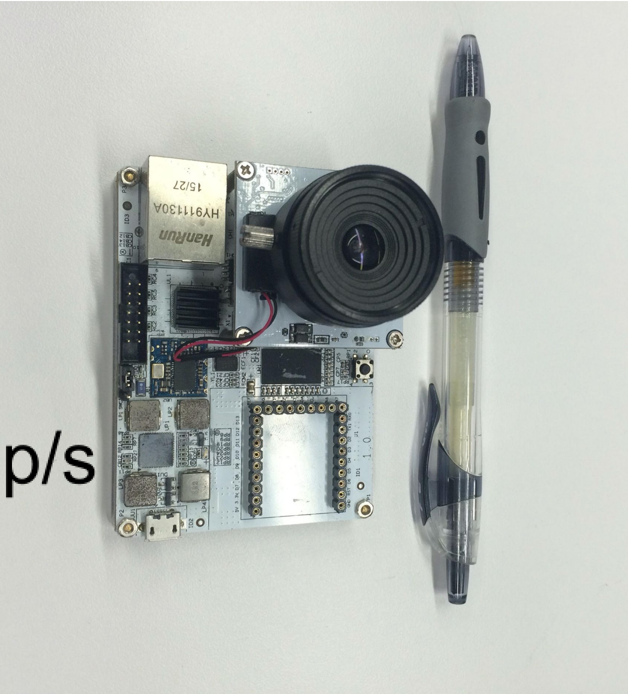
- FPGA is made up of LUT’s
- Bit-convolution kernel can be implemented by LUT



a) Conceptual structure of an FPGA device.

b) Three-input LUT-based logic cell



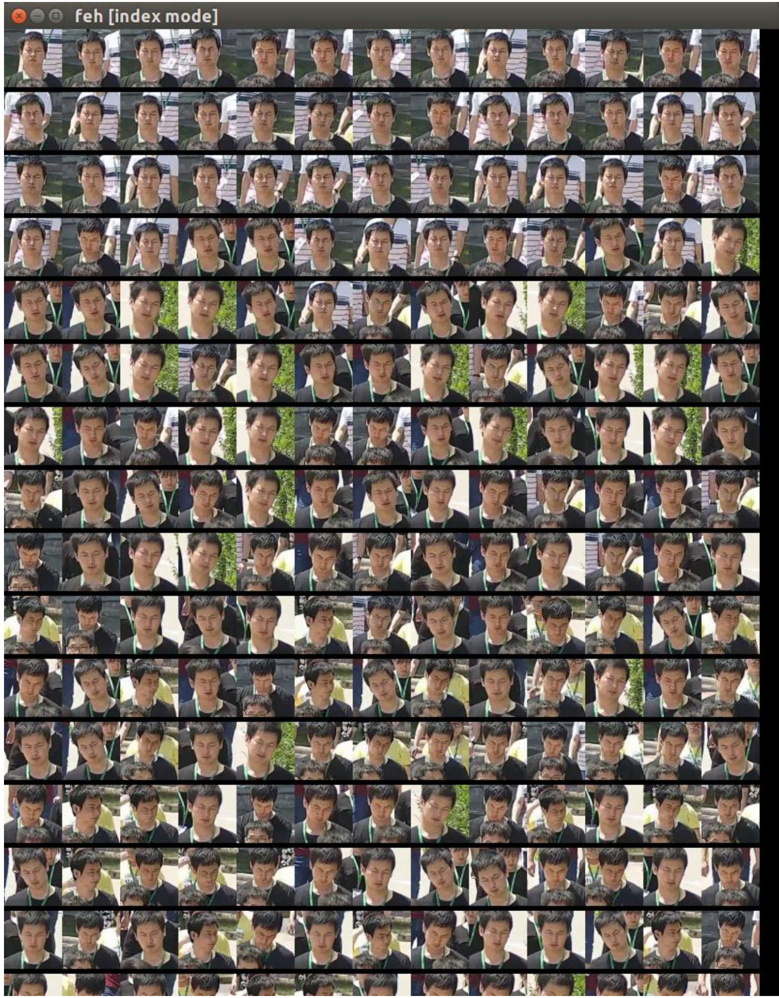


Demo at CVPR 2016

Smart Camera

- Benefits
  - Local processing
    - Low latency and high availability
  - High Frame Rate Conditional Capture
    - Less storage and bandwidth
    - High FPS = larger candidate set

Sort by Clearness



Sort by pose (frontal face)


2015-11-27 16:19:53
30
111
89
113
92
90
65
108
78
62
79
70
75
66
49

Backup after this slide

zsc@megvii.com

job@megvii.com

Face++ 旷视