

Neural Network Approximation

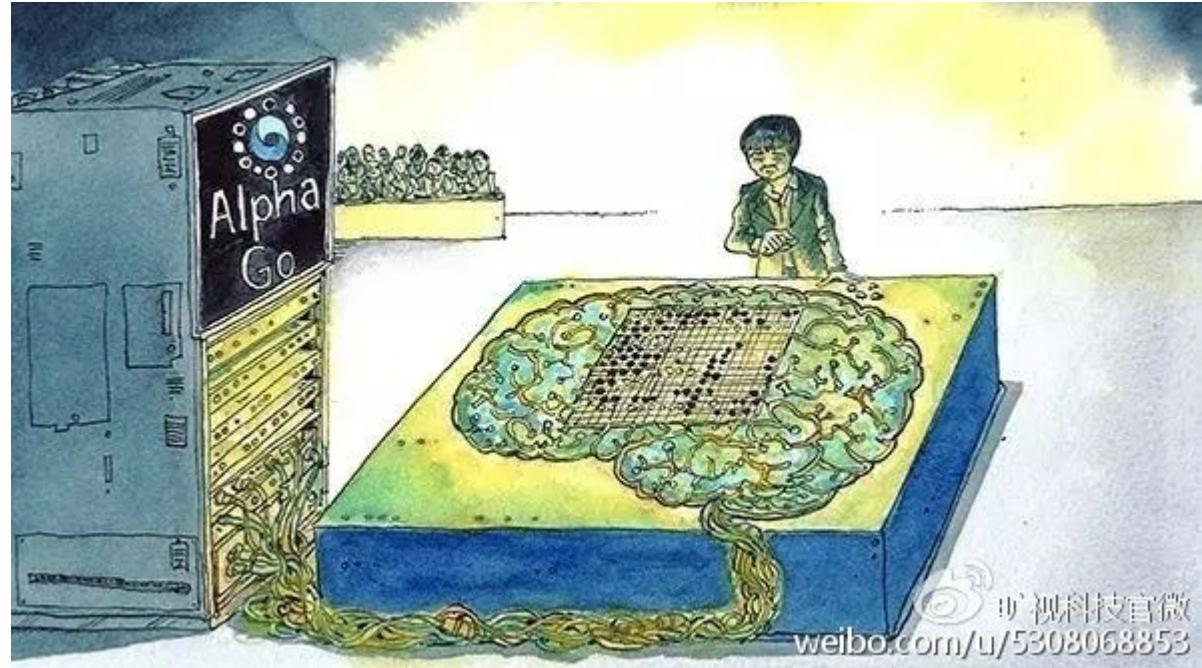
Low rank, Sparsity, and Quantization

zsc@megvii.com

Sep. 2019

Motivation

- Faster Inference
 - Latency critical scenarios
 - VR/AR, UGV/UAV
 - Saves time and energy
- Faster Training
 - Higher evolution speed
 - Saves life
- Lower Resource
 - storage size
 - memory footprint

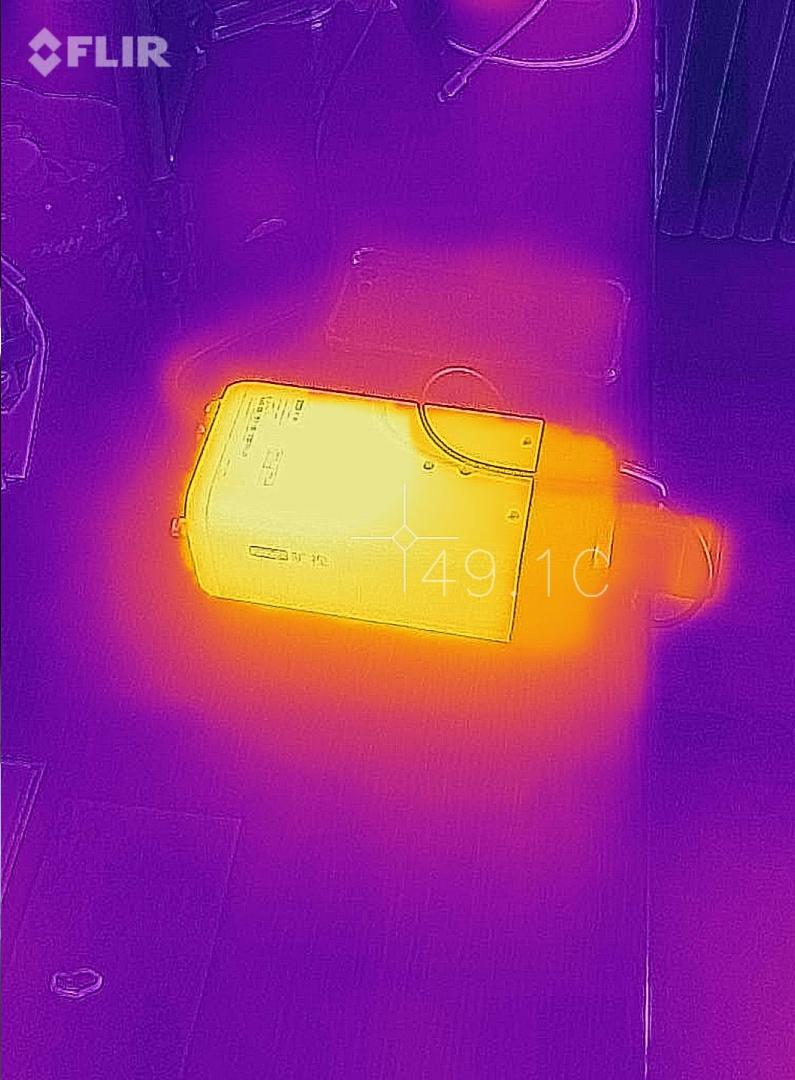


Lee Sedol vs. AlphaGo-Lee (48 TPU) vs. AlphaGo-Master(4 TPU)

83 Watt

1.9 kWatt (23x)

160 Watt (1.9x)



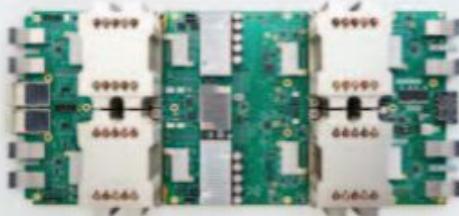
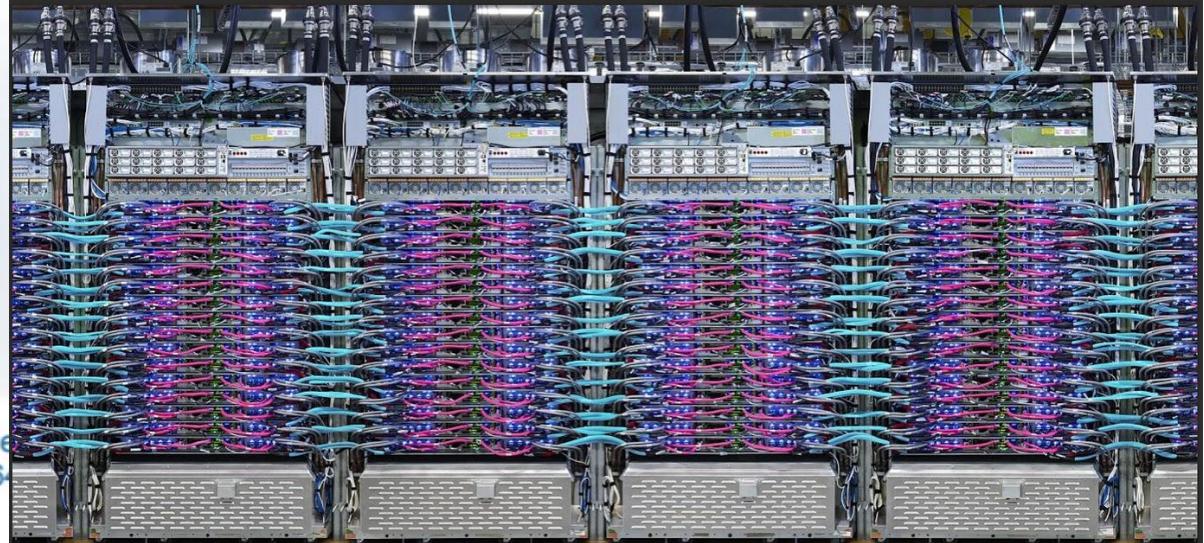


Figure 3. The second-generation Google TPU chips. The collection of racks includes 6



TPU v3 Pod: Revealed at Google I/O May 2017

52



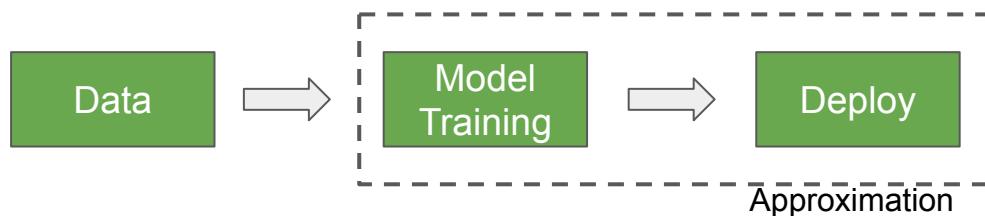


Approximation In Neural Network Lifetime

- Post-Training Approximation



- Approximation-Aware Training



Neural Network

Machine Learning as Optimization

- Supervised learning
 - θ is the parameter
 - \hat{y} is output, X is input
 - y is ground truth
 - d is the objective function
 - Unsupervised learning
 - some oracle function r: low rank, sparse, K
- $$\min_{\theta} d(y, \hat{y})$$
- $$\text{where } \hat{y} = f(X, \theta)$$
- $$\min_{\theta} r(\hat{y})$$
- $$\text{where } \hat{y} = f(X, \theta)$$

Machine Learning as Optimization

- Regularized supervised learning
 - $\min_{\theta} d(y, \hat{y}) + r(\hat{y})$
where $\hat{y} = f(X, \theta)$
- Probabilistic interpretation
 - d measures conditional probability
 - r measures prior probability
 - Probability approach is more constrained than the optimization-approach due to normalization problem
 - Not easy to represent uniform distribution over $[0, \infty]$

$$\begin{aligned}d(y, \hat{y}) &= -\log p(y|\hat{y}) \\r(\hat{y}) &= -\log p(\hat{y}) \\ \Rightarrow d(y, \hat{y}) + r(\hat{y}) &= -\log p(y)\end{aligned}$$

Gradient descent

$$\min_{\theta} d(y, \hat{y}) + r(\hat{y})$$

where $\hat{y} = f(X, \theta)$

- Can be solved by an ODE: $\dot{\theta} = -\frac{\partial C(\theta(t))}{\partial \theta}$
 - Discretizing with step length λ we get gradient descent with learning rate λ
 - $\dot{\theta} \approx \frac{\theta_{t+\lambda} - \theta_t}{\lambda}$  $\theta_{t+\lambda} = \theta_t - \lambda \frac{\partial C(\theta)}{\partial \theta}$

- Convergence proof

$$\frac{dC(\theta)}{dt} = \frac{\partial C(\theta)}{\partial \theta} \frac{d\theta}{dt} = -\left(\frac{\partial C(\theta)}{\partial \theta}\right)^2 \leq 0$$

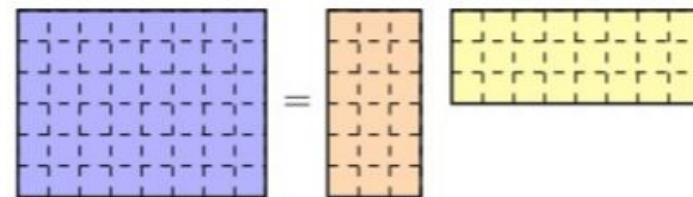
Linear Regression

- $\min_W \|y - Wx\|^2$
 - $\hat{y} = f(X, \theta) = WX$
 - $d(y, \hat{y}) = \|y - \hat{y}\|^2$
 - x is input, \hat{y} is prediction, y is ground truth.
 - W with dimension (m,n)
 - #param = $m n$, #OPs = $m n$
- $y \approx \hat{y} = Wx$

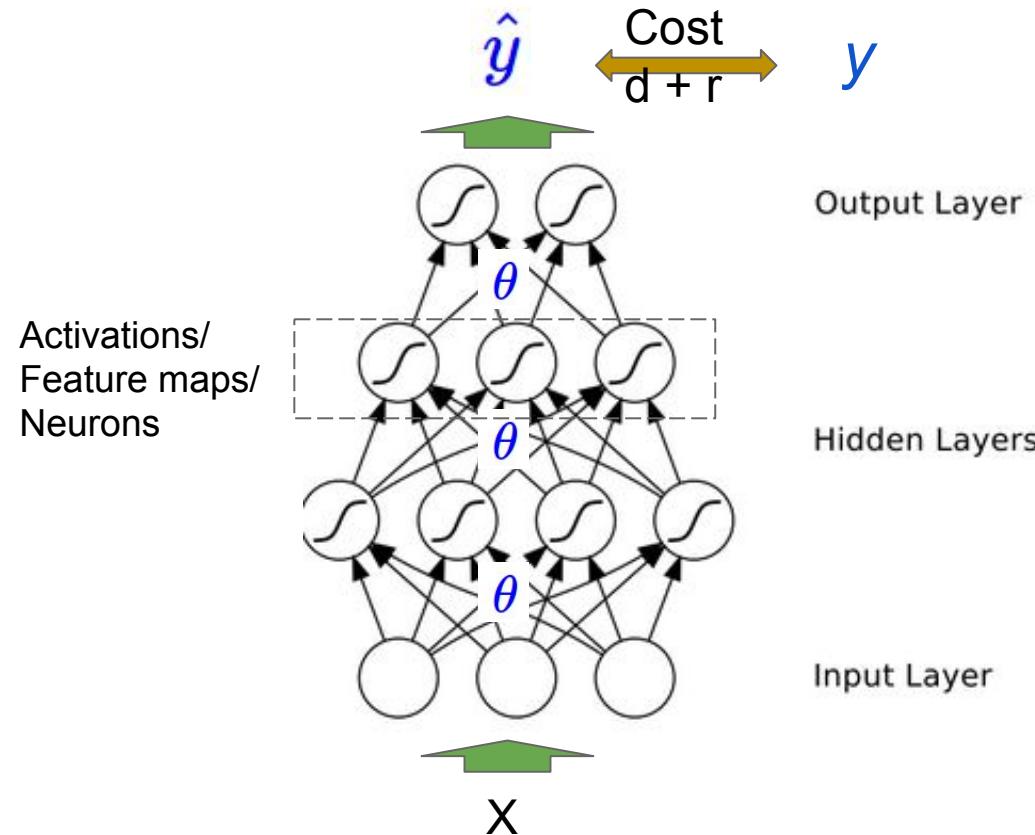


Fully-connected

- $y \approx W_2 f(W_1(x))$
 - In general, will use nonlinearity to increase “model capacity”.
 - Make sense if f is identity? I.e.
 $f(x) = x?$
 - Sometimes, if W_2 is m by r and W_1 is r by n , then $W_2 W_1$ is a matrix of rank r , which is different from a m by n matrix.
- $y \approx W_3(f(W_2 f(W_1(x))))$
 - Deep learning!



Neural Network



Gradient descent

$$\min_{\theta} d(y, \hat{y}) + r(\hat{y})$$

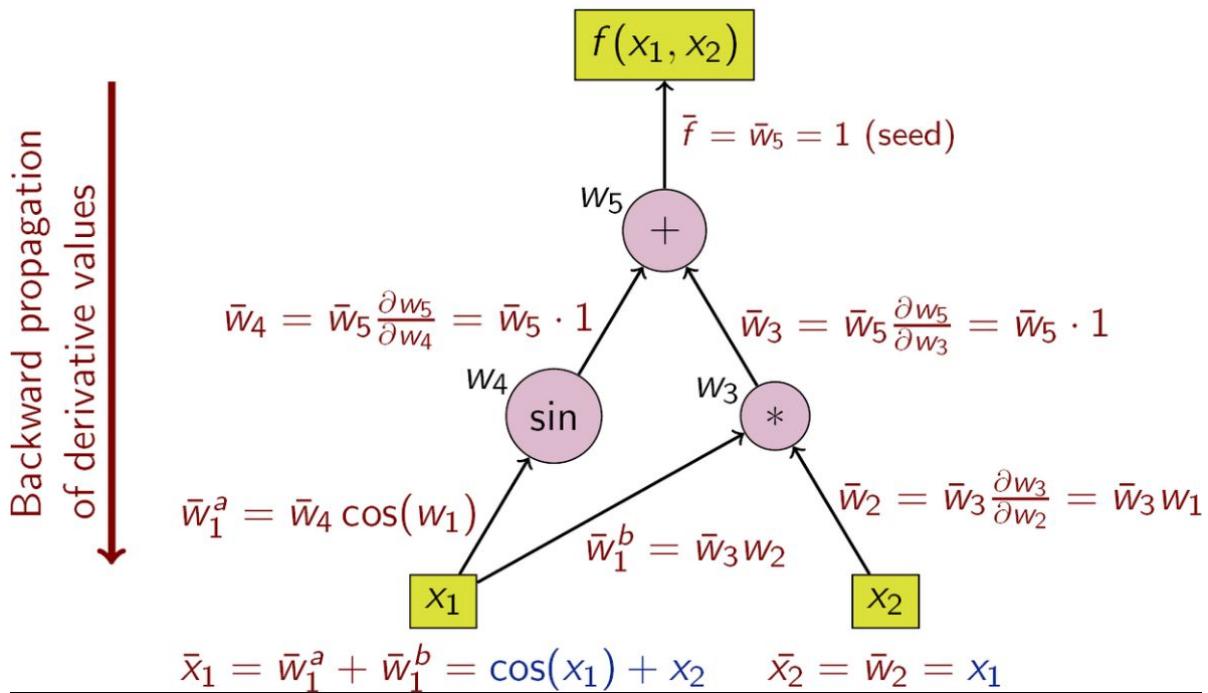
where $\hat{y} = f(X, \theta)$

- Can be solved by an ODE: $\dot{\theta} = -\frac{\partial C(\theta(t))}{\partial \theta}$
 - Discretizing with step length λ we get gradient descent with learning rate λ
 - $\dot{\theta} \approx \frac{\theta_{t+\lambda} - \theta_t}{\lambda}$  $\theta_{t+\lambda} = \theta_t - \lambda \frac{\partial C(\theta)}{\partial \theta}$
- Convergence proof

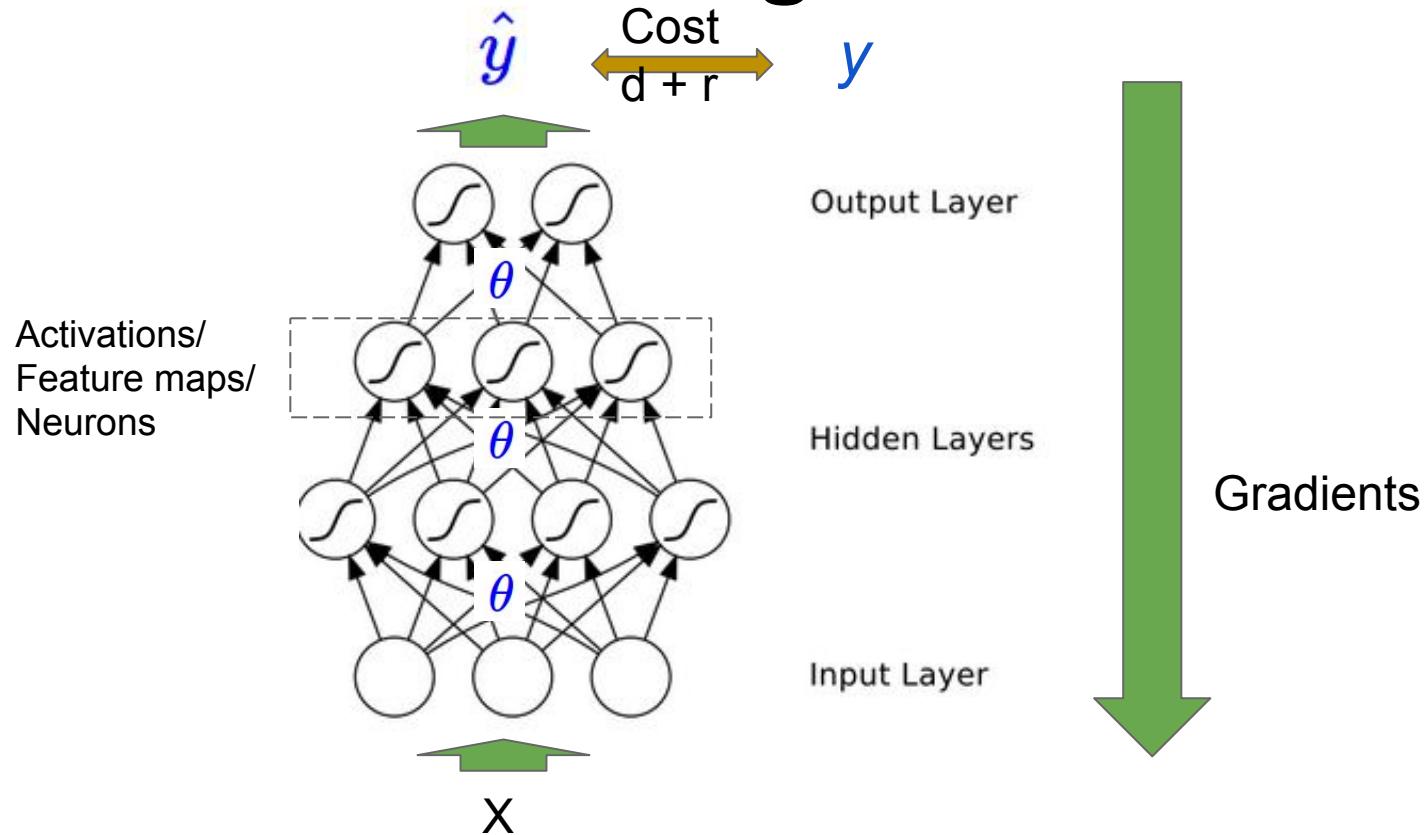
$$\frac{dC(\theta)}{dt} = \frac{\partial C(\theta)}{\partial \theta} \frac{d\theta}{dt} = -\left(\frac{\partial C(\theta)}{\partial \theta}\right)^2 \leq 0$$

Backpropagation

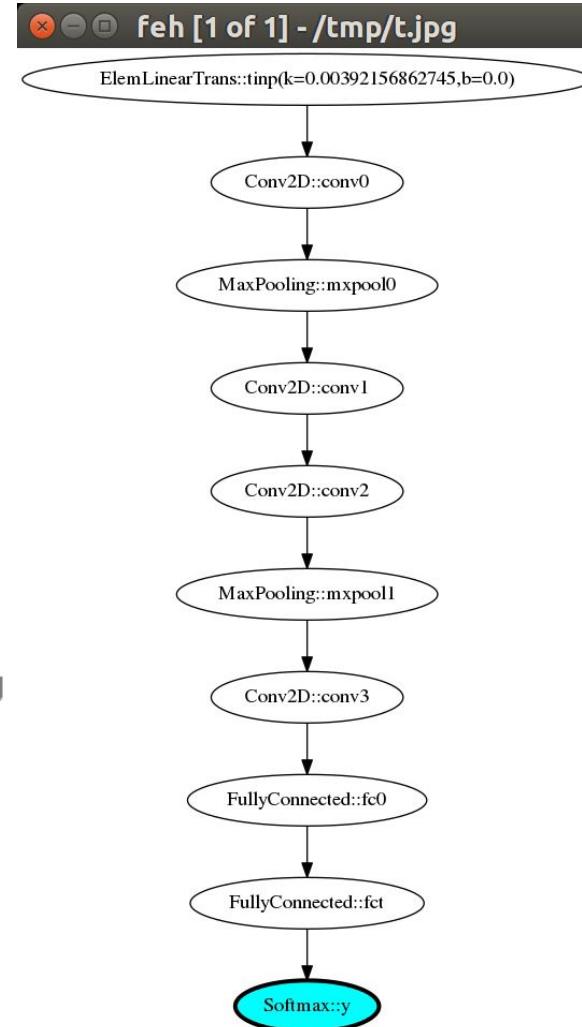
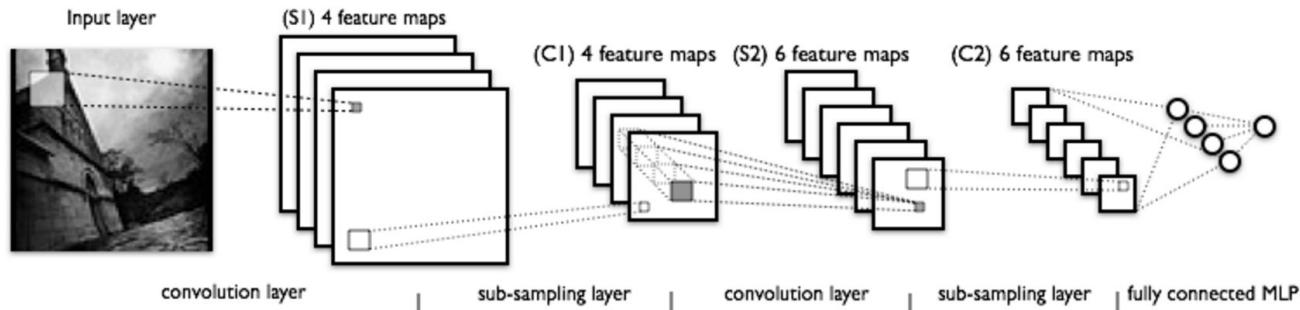
$$\frac{\partial y}{\partial x} = \frac{\partial y}{\partial w_1} \frac{\partial w_1}{\partial x} = \left(\frac{\partial y}{\partial w_2} \frac{\partial w_2}{\partial w_1} \right) \frac{\partial w_1}{\partial x} = \left(\left(\frac{\partial y}{\partial w_3} \frac{\partial w_3}{\partial w_2} \right) \frac{\partial w_2}{\partial w_1} \right) \frac{\partial w_1}{\partial x} = \dots$$



Neural Network Training



CNN: Alexnet-like



Method 1: Convolution as Product with Toeplitz Matrix

$$y = h * x = \begin{bmatrix} h_1 & 0 & \dots & 0 & 0 \\ h_2 & h_1 & \dots & \vdots & \vdots \\ h_3 & h_2 & \dots & 0 & 0 \\ \vdots & h_3 & \dots & h_1 & 0 \\ h_{m-1} & \vdots & \dots & h_2 & h_1 \\ h_m & h_{m-1} & \vdots & \vdots & h_2 \\ 0 & h_m & \dots & h_{m-2} & \vdots \\ 0 & 0 & \dots & h_{m-1} & h_{m-2} \\ \vdots & \vdots & \vdots & h_m & h_{m-1} \\ 0 & 0 & 0 & \dots & h_m \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_n \end{bmatrix}$$

Example: Convolution with Laplacian filter as Sparse Matrix Product

- $\text{vect}(M \star_{\text{conv}} \text{Laplacian}) = A \text{vect}(M)$
 - A is the sparse Poisson Matrix

0	-1	0
-1	4	-1
0	-1	0

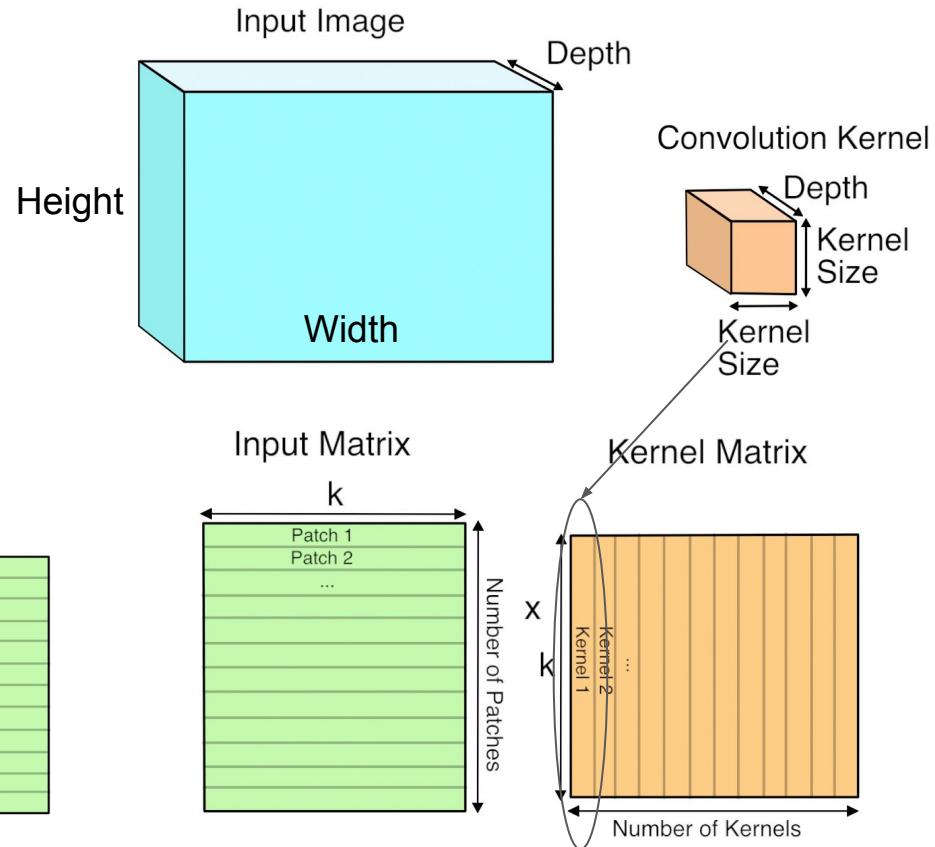
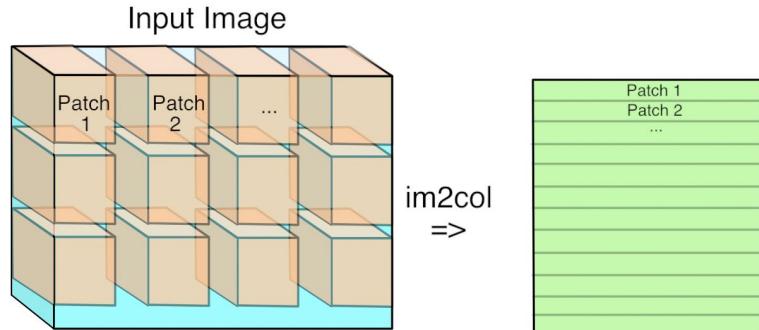
Laplacian

$$A = \begin{bmatrix} D & -I & 0 & 0 & 0 & \dots & 0 \\ -I & D & -I & 0 & 0 & \dots & 0 \\ 0 & -I & D & -I & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & \dots & 0 & -I & D & -I & 0 \\ 0 & \dots & \dots & 0 & -I & D & -I \\ 0 & \dots & \dots & \dots & 0 & -I & D \end{bmatrix}, D = \begin{bmatrix} 4 & -1 & 0 & 0 & 0 & \dots & 0 \\ -1 & 4 & -1 & 0 & 0 & \dots & 0 \\ 0 & -1 & 4 & -1 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & \dots & 0 & -1 & 4 & -1 & 0 \\ 0 & \dots & \dots & 0 & -1 & 4 & -1 \\ 0 & \dots & \dots & \dots & 0 & -1 & 4 \end{bmatrix},$$

Method 2: Convolution as matrix product

- Convolution
 - feature map $\langle N, C, H', W' \rangle$
 - weights $\langle K, C, H, W \rangle$
 - Convolution as FC
 - under proper padding, can extract patches
 - feature map $\langle N H' W', C H W \rangle$
 - weights $\langle C H W, K \rangle$

Kernel stride
determines
how much
overlap



Importance of Convolutions and FC

param_name	shape	#floats	size	perc
conv0_W	(24, 3, 5, 5)	1800	7.0 Kib	0.21%
conv0_b	(24,)	24	96.0 B	0.00%
conv1_W	(32, 24, 3, 3)	6912	27.0 KiB	0.80%
conv1_b	(32,)	32	128.0 B	0.00%
conv2_W	(32, 32, 3, 3)	9216	36.0 KiB	1.07%
conv2_b	(32,)	32	128.0 B	0.00%
conv3_W	(64, 32, 3, 3)	18432	72.0 KiB	2.14%
conv3_b	(64,)	64	256.0 B	0.01%
fc0_W	(1600, 512)	819200	3.1 MiB	95.11%
fc0_b	(512,)	512	2.0 Kib	0.06%
fct_W	(512, 10)	5120	20.0 KiB	0.59%
fct_b	(10,)	10	40.0 B	0.00%
total size		861354	3.3 MiB	

[tf-model-manip.py](#) 01-svhn/model.py info

Most storage size

Feature map size

layer_name	ispace	ospace	osize	#ops	readable	perc
tinp	(3, 40, 40)	(3, 40, 40)	4800	0	0.0	0.00%
conv0	(3, 40, 40)	(24, 36, 36)	31104	2332800	2.2 Mi	32.43%
mxpool0	(24, 36, 36)	(24, 18, 18)	7776	0	0.0	0.00%
conv1	(24, 18, 18)	(32, 16, 16)	8192	1769472	1.7 Mi	24.60%
conv2	(32, 16, 16)	(32, 14, 14)	6272	1806336	1.7 Mi	25.11%
mxpool1	(32, 14, 14)	(32, 7, 7)	1568	0	0.0	0.00%
conv3	(32, 7, 7)	(64, 5, 5)	1600	460800	450.0 Ki	6.41%
fc0	(64, 5, 5)	(512, 1, 1)	512	819712	800.5 Ki	11.39%
fct	(512, 1, 1)	(10, 1, 1)	10	5130	5.0 Ki	0.07%
y	(10, 1, 1)	(10, 1, 1)	10	0	0.0	0.00%
total OPs				7194250	6.9 MiOps	

Most Computation

The Matrix View of Neural Network

- Weights of FullyConnected and Convolutions layers
 - take up most computation and storage size
 - are representable as matrices
- Approximating the matrices approximates the network
 - The approximation error accumulates.

$$W_a \approx W \Rightarrow f(X, W_a) \approx f(X, W)$$

Low rank Approximation

Singular Value Decomposition

- Matrix decomposition view
 - $A = U S V^T$
 - Rows of U, V are orthogonal. S is diagonal.
 - $u, s, v^T = np.linalg.svd(x, \text{full_matrices}=0, \text{compute_uv}=1)$
 - The diagonals are non-negative and are in descending order.
 - $U^T U = I$, but $U U^T$ is not full rank

$$A_{m \times n} = \begin{bmatrix} & & \\ & & \\ & & \end{bmatrix}_{m \times k} \begin{bmatrix} & & \\ & & \\ & & \end{bmatrix}_{k \times k} \begin{bmatrix} & & \\ & & \\ & & \end{bmatrix}_{k \times n}$$

Truncated SVD

- Assume diagonals of S are in descending order
 - Always achievable
 - Just ignore the blue segments.

$$A_k = U_k \Sigma_k V_k^T$$

A_k = U_k Σ_k V_k^T

k? k? k? k?

SVD

$$M = s_1 u_1 \times v_1^T + s_2 u_2 \times v_2^T + \cdots + s_r u_r \times v_r^T$$

+

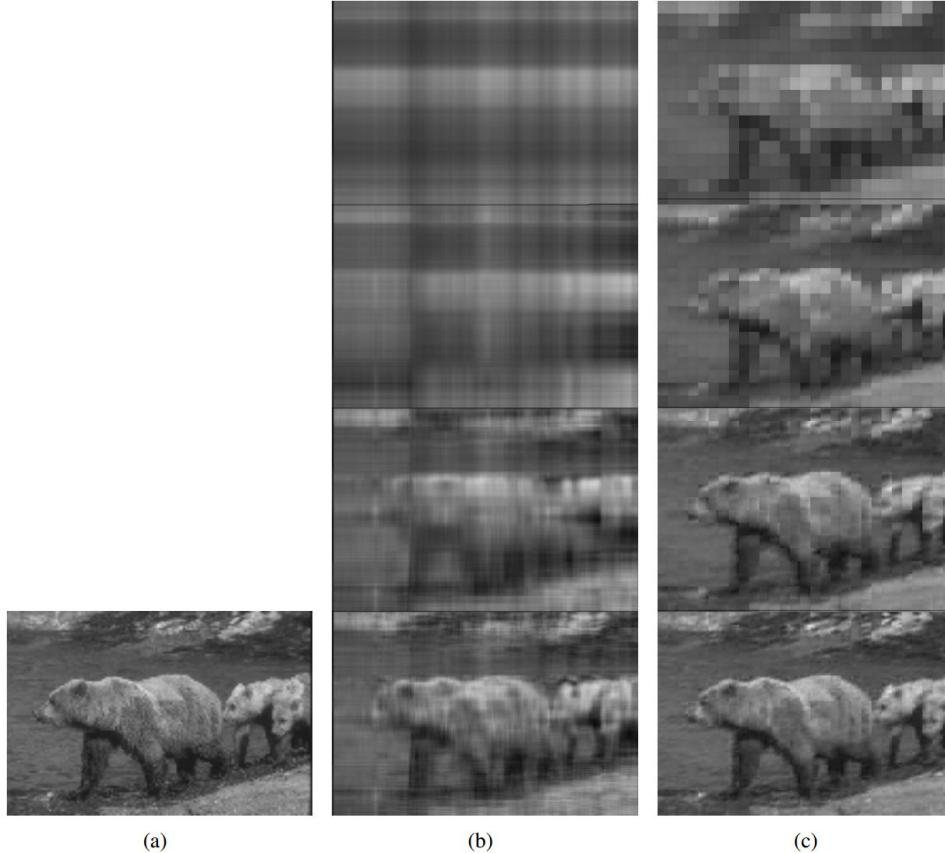
Kronecker
Product



$$\begin{matrix} a_{11}B & a_{12}B & a_{13}B \\ a_{21}B & a_{22}B & a_{23}B \end{matrix} = \begin{pmatrix} a_{1,1} & a_{1,2} & a_{1,3} \\ a_{2,1} & a_{2,2} & a_{2,3} \end{pmatrix} \otimes B$$

- KPSVD:

$$A \approx A_r = \sum_{i=1}^r \sigma_i U_i \otimes V_i.$$



(a)

(b)

(c)

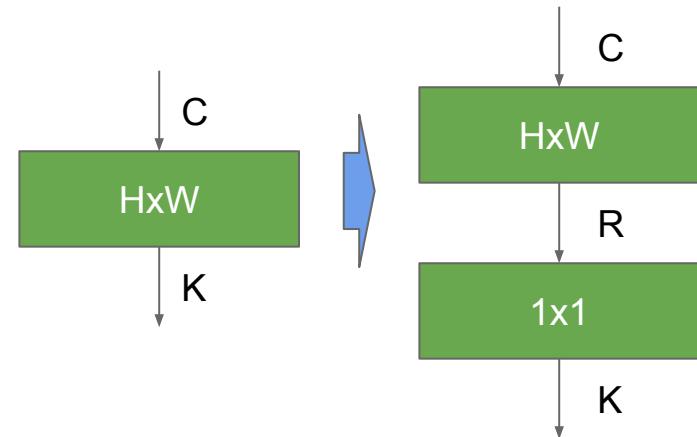
- 1: Comparison between approximations by outer product and Kronecker product for an image. The column (a) is the origin image of size 480×320 , selected from BSD500 dataset Arbelaez et al. (2011). The column (b) is the SVD approximations of (a) by outer product and the column (c) is the approximation based on Kronecker product Van Loan and Pitsianis (1993), with rank 1, 2, 5, 10 respectively from top to down. The shape of the right matrix in the Kronecker product is deliberately selected as 20×16 to make the number of parameters equal for each rank.

Matrix factorization => Convolution factorization

- Factorization into $H \times W$ followed by 1×1

- feature map ($N H' W', C H W$)
- first conv weights ($C H W, R$)
- feature map ($N H' W', R$)
- second conv weights (R, K)
- feature map ($N H' W', K$)

$$\begin{matrix} & K \\ \text{CHW} & \end{matrix} = \begin{matrix} R \\ \text{CHW} \end{matrix} \begin{matrix} & K \\ R & \end{matrix}$$



Matrix factorization => Convolution factorization

- $\text{conv}(W, F) = \text{reshape}(W) \text{ toeplitz}(F)$
- $\text{toeplitz}^{1 \times 1}(F) = F$
- $\text{conv}^{1 \times 1}(W_1, \text{conv}(W_2, F)) = \text{reshape}(W_1) \text{ toeplitz}_1(\text{reshape}(W_2) \text{ toeplitz}_2(F))$
 - When toeplitz_1 is a 1×1 toeplitz, RHS = $\text{reshape}(W_1) \text{ reshape}(W_2) \text{ toeplitz}_2(F)$, then the two conv's can be combined into one.

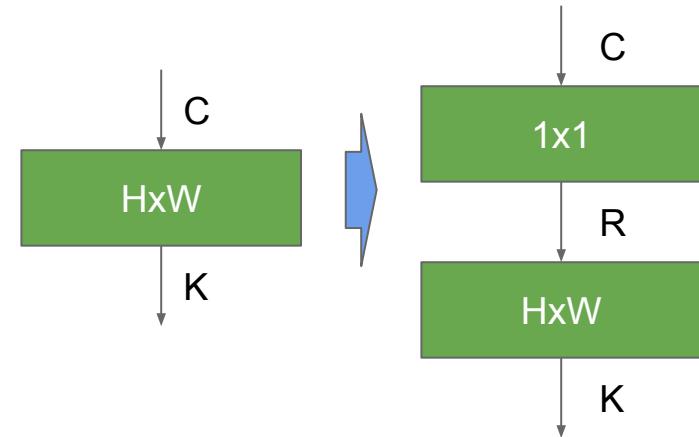
Approximating Convolution Weight

- $W \in \mathbb{R}^{K \times C \times H \times W}$ can be reshaped to a (CHW, K) matrix, etc.
- The approximation can be bound by F-norm

$$\|Mv - M_a v\| \leq \|M - M_a\|_{op} \|v\|$$
$$\|M - M_a\|_{op} = \sigma_{max} \leq \|\vec{\sigma}\|_2 = \|M - M_a\|_F$$
$$\|M - M_a\|_F = \|\text{reshape}(W) - \text{reshape}(W_a)\| = \|W - W_a\|_F$$

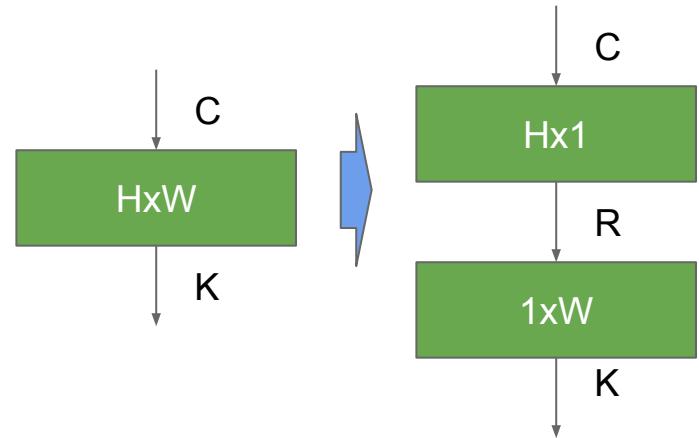
Matrix factorization => Convolution factorization

- Factorization into 1×1 followed by $H \times W$
 - feature map ($N H' W' H W, C$)
 - first conv weights (C, R)
 - feature map ($N H' W' H W, R$) = $(N H' W', R H W)$
 - second conv weights ($R H W, K$)
 - feature map ($N H' W', K$)
- Steps
 - Reshape (CHW, K) to (C, HW, K)
 - $(C, HW, K) = (C, R) (R, HW, K)$
 - Reshape (R, HW, K) to (RHW, K)



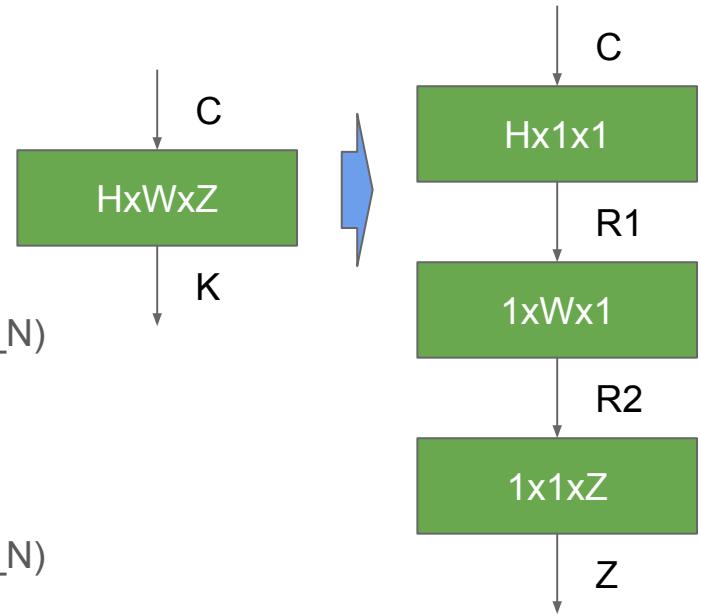
Horizontal-Vertical Decomposition

- Approximating with Separable Filters
- Original Convolution
 - feature map ($N H' W', C H W$)
 - weights ($C H W, K$)
- Factorization into $Hx1$ followed by $1xW$
 - feature map ($N H' W' W, C H$)
 - first conv weights ($C H, R$)
 - feature map ($N H' W' W, R$) = ($N H' W', R W$)
 - second conv weights ($R W, K$)
 - feature map ($N H' W', K$)
- Steps
 - Reshape (CHW, K) to (CH, WK)
 - $(CH, WK) = (CH, R) (R, WK)$
 - Reshape (R, WK) to (RW, K)



Factorizing N-D convolution

- Original Convolution
 - let dimension be N
 - feature map ($N D'_1 D'_2 \dots D'_Z, C D_1 D_2 \dots D_N$)
 - weights ($C D_1 D_2 \dots D_N, K$)
- Factorization into N number of $D_i \times 1$
 - $R_0 = C, R_Z = K$
 - feature map ($N D'_1 D'_2 \dots D'_Z, C D_1 D_2 \dots D_N$)
 - weights ($R_0 D_1, R_1$)
 - feature map ($N D'_1 D'_2 \dots D'_Z, R_1 D_2 \dots D_N$)
 - weights ($R_1 D_2, R_2$)
 - ...



Kronecker Conv

- $(C H W, K)$
- Reshape as $(C_1 C_2 H W, K_1 K_2)$
- Steps
 - Feature map is $(N C H' W')$
 - Extract patches and reshape $(N H' W' C_2, C_1 H)$
 - apply $(C_1 H, K_1 R)$
 - Feature map is $(N K_1 R H' W' C_2)$
 - Extract patches and reshape $(N K_1 H' W', R C_2 W)$
 - apply $(R C_2 W, K_2)$
- For rank efficiency, should have
 - $R C_2 \backslash \approx C_1$

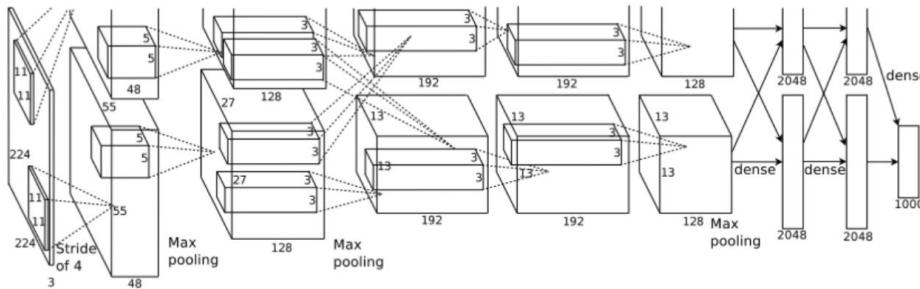
Exploiting Local Structures with the Kronecker Layer in Convolutional Networks [1512](#)

Methods	Configuration (r, o_1, c_1, h_1, w_1)	Validation Error	Speedup
Baseline	/	7.84%	/
KConv-a	1,128,24,9,1; 1,256,64,8,1	8.76%	3.3×
KConv-b	1,128,48,1,9; 1,512,64,1,8	8.69%	3.0×
KConv-c	2,64,24,9,1; 2,256,64,8,1	7.87%	2.9×

We have also experimented replacing the first convolutional layer with KConv layer. In this case, KConv with $(r, o_1, c_1, h_1, w_1) = (2, 12, 1, 1, 9)$, is found to outperform Jaderberg-style rank-1 filter with $(r, o_1, c_1, h_1, w_1) = (2, 96, 1, 1, 9)$ by 0.83%.

Shared Group Convolution is a Kronecker Layer

AlexNet partitioned a conv



The diagram shows the equivalence between a standard convolution/fully connected (Conv/FC) layer and a shared group convolution/fully connected (Shared Group Conv/FC) layer. On the left, a green square represents the input to a standard Conv/FC layer. An arrow points to the right, where the input is shown as a 4x4 grid of smaller green squares, representing it being partitioned into four groups. This is followed by an equals sign and a matrix multiplication operation. The matrix is a 4x4 identity matrix:

$$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \otimes \text{[green square]}$$

CP-decomposition and Xception

- Xception: Deep Learning with Depthwise Separable Convolutions [1610](#)
- CP-decomposition with Tensor Power Method for Convolutional Neural Networks Compression [1701](#)
- MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications [1704](#)
 - They submitted the paper to CVPR about the same time as Xception.

Layer/Modification	Million	Million
	Mult-Adds	Parameters
Convolution	462	2.36
Depthwise Separable Conv	52.3	0.27
$\alpha = 0.75$	29.6	0.15
$\rho = 0.714$	15.1	0.15

Table 5. Narrow vs Shallow MobileNet

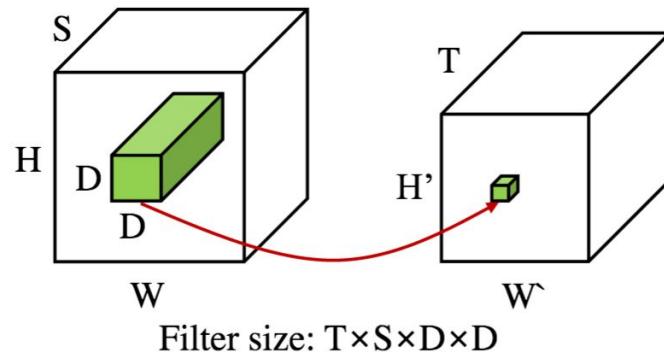
Model	ImageNet Accuracy	Million Mult-Adds	Million Parameters
0.75 MobileNet	68.4%	325	2.6
Shallow MobileNet	65.3%	307	2.9

Matrix Joint Diagonalization = CP

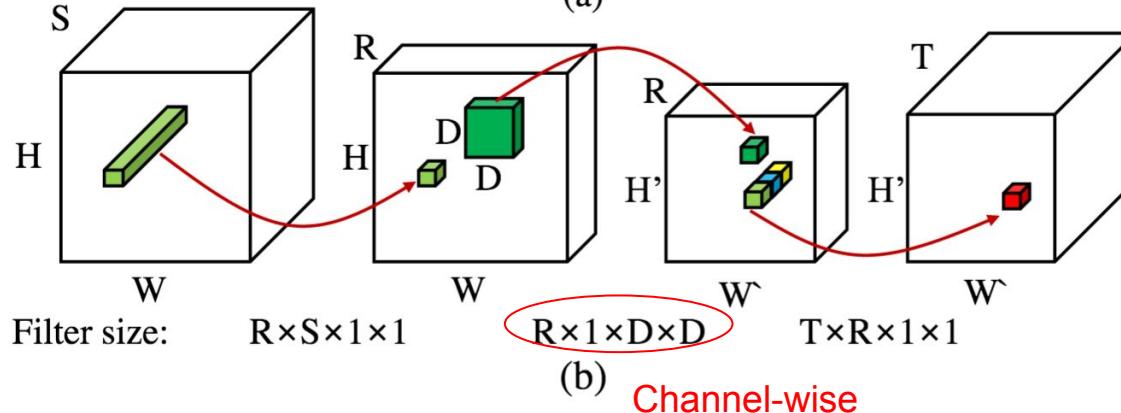
$$\text{CP} \quad \begin{matrix} \text{HxW} \\ \text{C} \quad \text{T} \\ \text{K} \end{matrix} = \text{C} \times \begin{matrix} \text{HxW} \\ \text{S} \end{matrix} \times \text{K}$$
$$\text{MJD} \quad \begin{matrix} \text{T}_1 \\ \vdots \end{matrix} = \text{C} \times \begin{matrix} \text{S} \\ \text{S}_1 \\ \text{S}_2 \\ \text{S}_3 \end{matrix} \times \text{K}$$

CP-decomposition with Tensor Power Method for Convolutional Neural Networks Compression 1701

Convolution



Xception



Tensor Train Decomposition

Oseledets, 2009:

TT-format

$$A(i_1, i_2, \dots, i_d) \approx \sum_{\alpha_1, \alpha_2, \dots, \alpha_{d-1}} G_1(i_1, \alpha_1) G_2(\alpha_1, i_2, \alpha_2) \dots G_d(\alpha_{d-1}, i_d)$$

$$f = x_1 + x_2 + \dots + x_d$$

Canonical rank: d

TT-rank: 2

$$f(x_1, \dots, x_d) = \sin(x_1 + x_2 + \dots + x_d)$$

FTT-decomposition has form

$$f = \begin{pmatrix} \sin x_1 & \cos x_1 \end{pmatrix} \begin{pmatrix} \cos x_2 & -\sin x_2 \\ \sin x_2 & \cos x_2 \end{pmatrix} \dots \dots$$

$$\begin{pmatrix} \cos x_{d-1} & -\sin x_{d-1} \\ \sin x_{d-1} & \cos x_{d-1} \end{pmatrix} \begin{pmatrix} \cos x_d \\ \sin x_d \end{pmatrix}.$$

$$f = \begin{pmatrix} x_1 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ x_2 & 1 \end{pmatrix} \dots \dots \begin{pmatrix} 1 & 0 \\ x_{d-1} & 1 \end{pmatrix} \begin{pmatrix} 1 \\ x_d \end{pmatrix}$$

Tensor Train Decomposition: just a few SVD's

Require: d -dimensional tensor \mathbf{A} required accuracy ε

Ensure: Cores G_1, \dots, G_d of the TT-approximation \mathbf{B} to \mathbf{A} in the TT-format with smallest possible compression ranks \hat{r}_k such that

$$\|\mathbf{A} - \mathbf{B}\|_F \leq \varepsilon \|\mathbf{A}\|_F,$$

{Initialization}

Compute truncation parameter $\delta = \frac{\varepsilon}{\sqrt{d-1}} \|\mathbf{A}\|_F$.

Temporary tensor: $\mathbf{C} = \mathbf{A}$.

$N = \text{numel}(\mathbf{A})$, $r_0 = 1$.

for $k = 1$ to $d - 1$ do

$\mathbf{C} := \text{reshape}(\mathbf{C}, [r_{k-1} n_k, \frac{N}{r_{k-1} n_k}]).$

Compute δ -truncated SVD: $\mathbf{C} = \mathbf{U}\mathbf{S}\mathbf{V}^\top + \mathbf{E}$, $\|\mathbf{E}\|_F \leq \delta$, $r_k = \text{rank}_\delta(\mathbf{C})$.

New core: $G_k := \text{reshape}(\mathbf{U}, [r_{k-1}, n_k, r_k]).$

$\mathbf{C} := \mathbf{S}\mathbf{V}^\top.$

$N := \frac{Nr_k}{n_k r_{k-1}}.$

end for

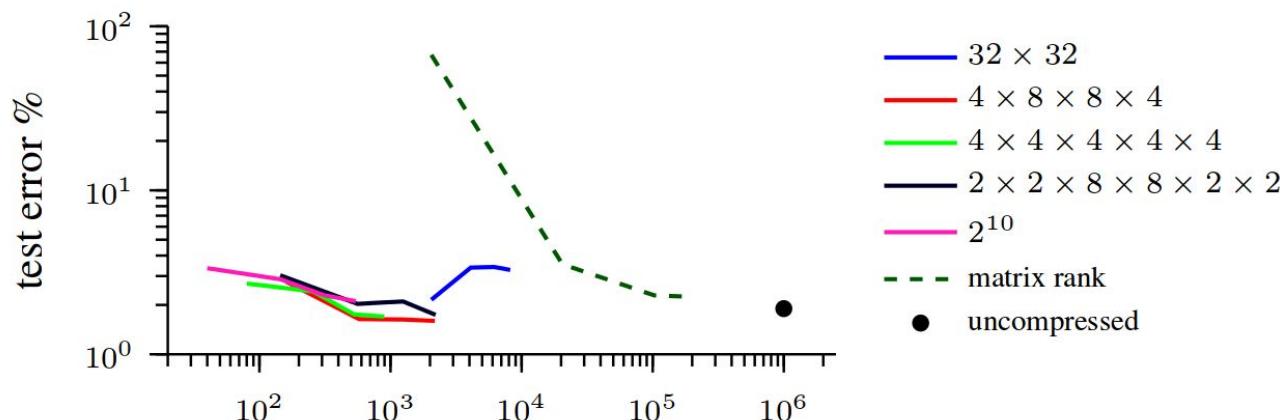
$G_d = \mathbf{C}$.

Return tensor \mathbf{B} in TT-format with cores G_1, \dots, G_d .

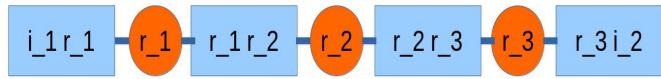
Tensor Train Decomposition on FC

Type	1 im. time (ms)	100 im. time (ms)
CPU fully-connected layer	16.1	97.2
CPU TT-layer	1.2	94.7
GPU fully-connected layer	2.7	33
GPU TT-layer	1.9	12.9

Table 3: Inference time for a 25088×4096 fully-connected layer and its corresponding TT-layer with all the TT-ranks equal 4. The memory usage for feeding forward one image is 392MB for the fully-connected layer and 0.766MB for the TT-layer.

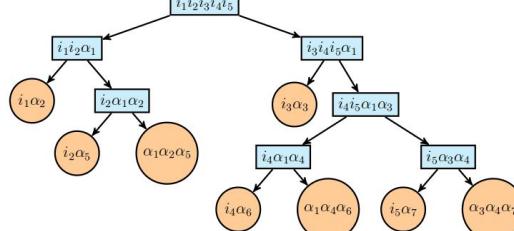


Graph Summary of SVD variants

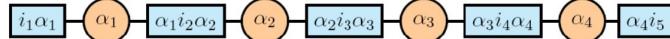


Matrix Product State

EXTENDED TT DECOMPOSITION



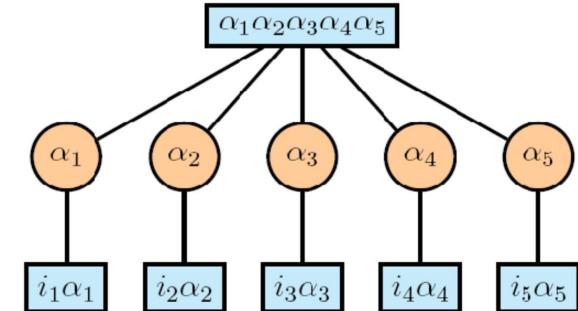
$$\text{NUMBER OF EXTENDED TT PARAMETERS} = dnr + (d - 2)r^3$$



$$a(i_1, \dots, i_d) = \sum_{\alpha_1, \dots, \alpha_{d-1}} g_1(i_1, \alpha_1) g_2(\alpha_1, i_2, \alpha_2) \dots$$

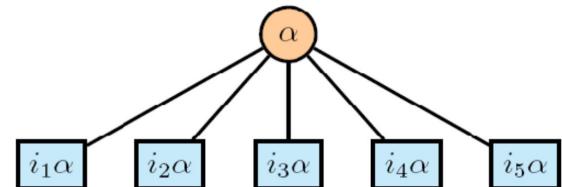
$$\dots g_{d-1}(\alpha_{d-2}, i_{d-1}, \alpha_{d-1}) g_d(\alpha_{d-1}, i_d)$$

TENSOR-TRAIN DECOMPOSITION



$$a(i_1, \dots, i_d) = \sum_{\alpha_1=1}^{r_1} \dots \sum_{\alpha_d=1}^{r_d} g(\alpha_1, \dots, \alpha_d) q_1(i_1, \alpha_1) \dots q_d(i_d, \alpha_d)$$

TUCKER DECOMPOSITION



$$a(i_1, \dots, i_d) = \sum_{\alpha=1}^R u_1(i_1, \alpha) \dots u_d(i_d, \alpha)$$

CANONICAL DECOMPOSITION (PARAFAC, CANDECOMP)

CNN layers as Multilinear Maps

Convolution: $A_{k,c} X_{n,c,h,w,h_k,w_k}$

FC: $A_{k,c} X_{n,c,h,w,1,1}$

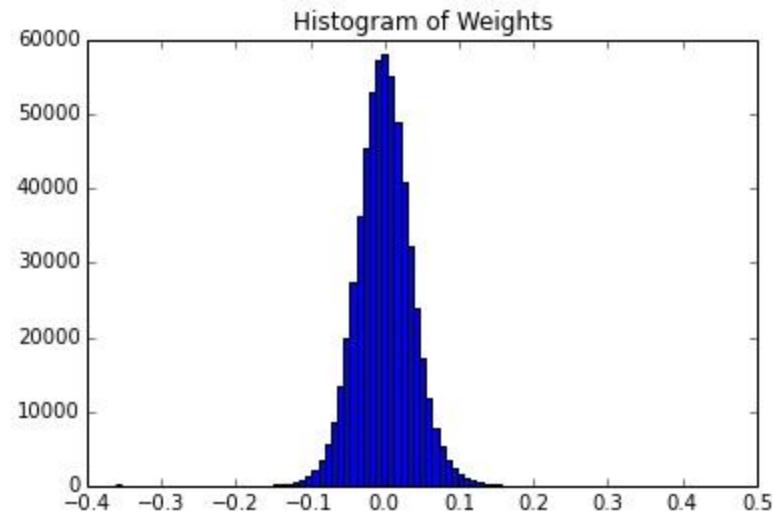
Xception: $A_{k,c} \delta_{k,c} X_{n,c,h,w,h_k,w_k}$

Kronecker: $A_{k_2,c_2} X_{n,c_1,c_2,h,w,h_k,w_k}$

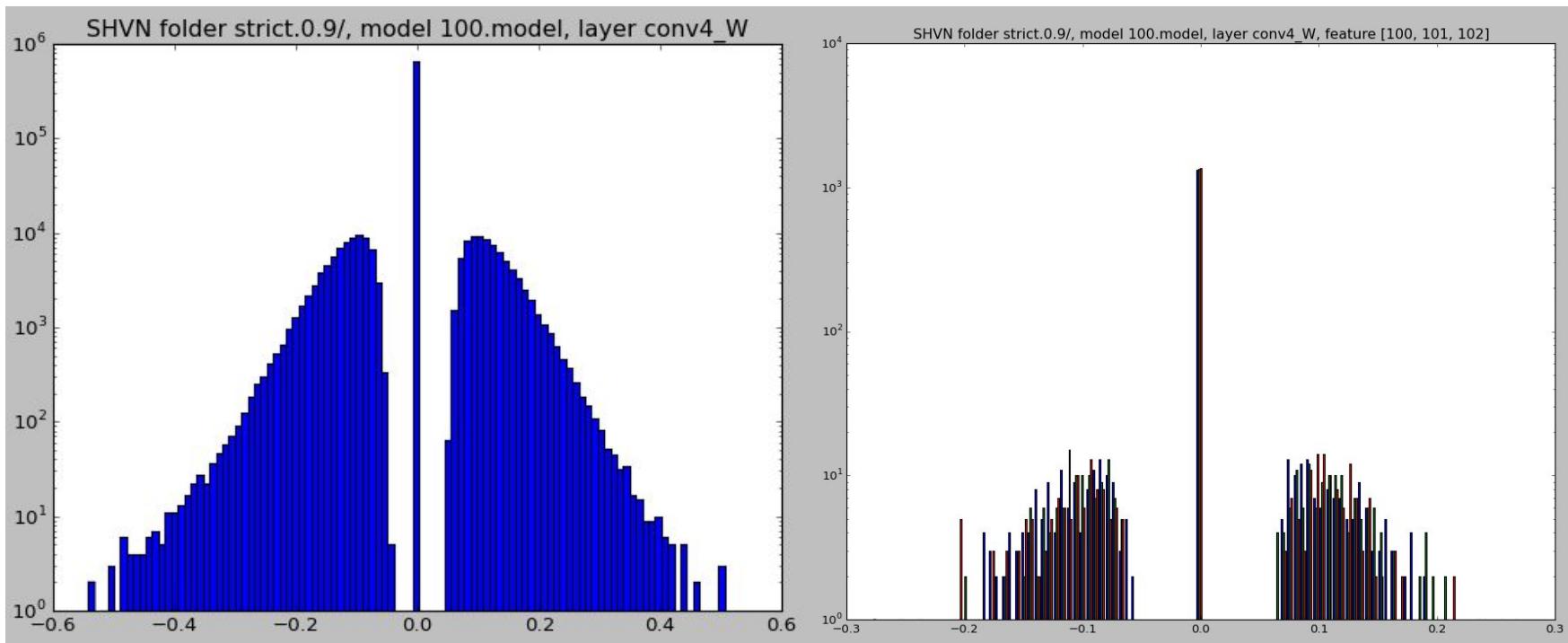
Sparse Approximation

Distribution of Weights

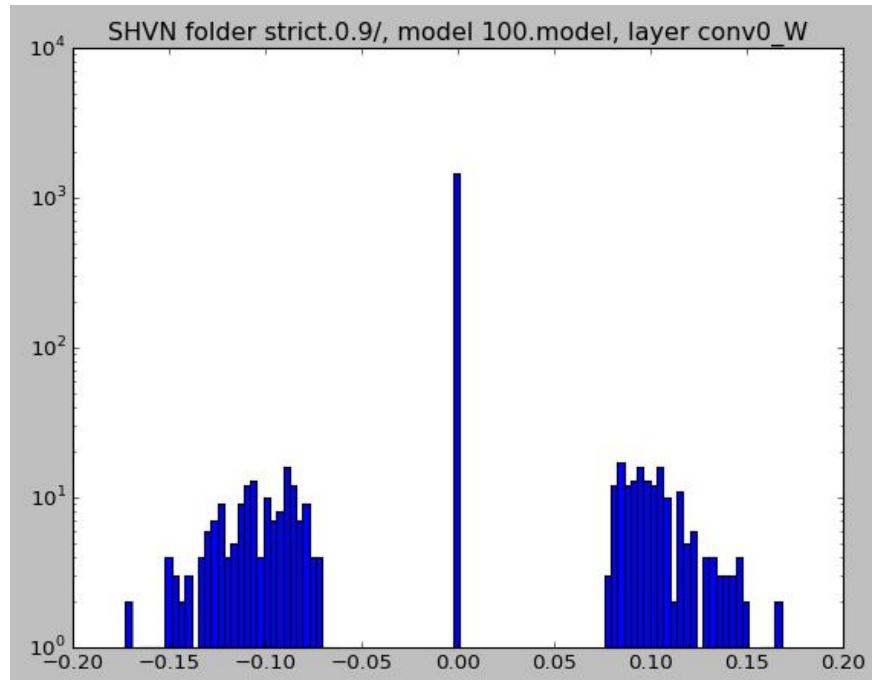
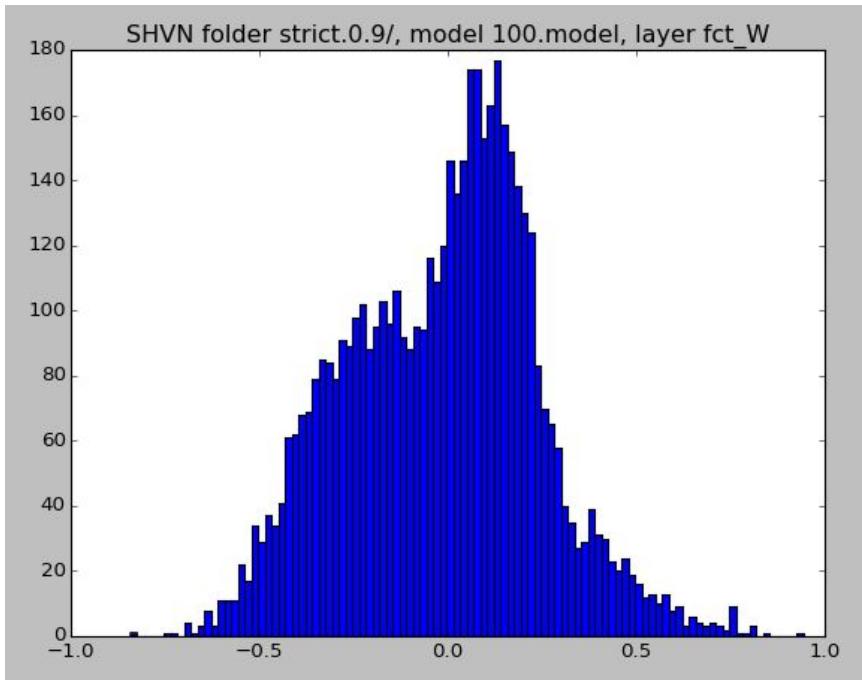
- Universal across convolutions and FC
- Concentration of values near 0
- Large values *cannot* be dropped



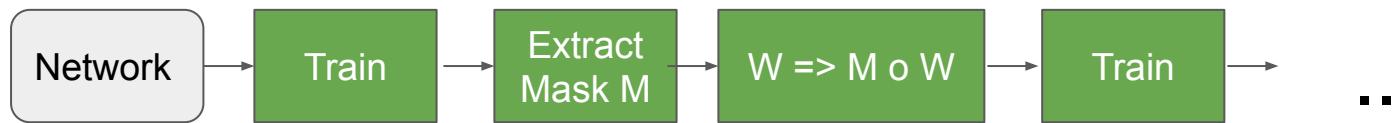
Sparsity of NN: statistics



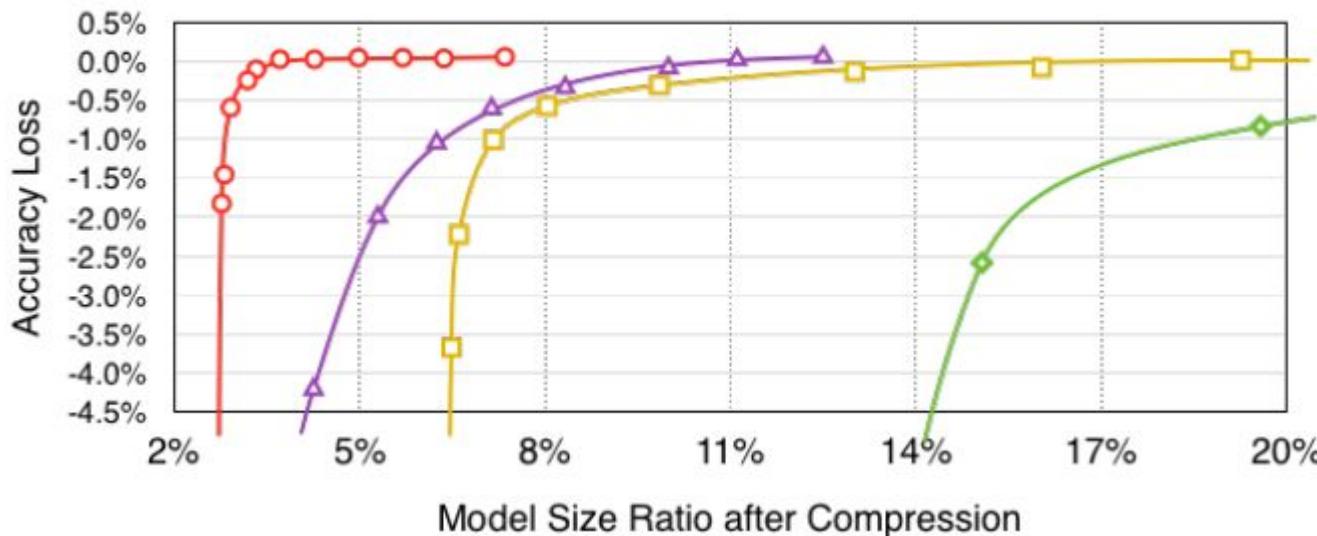
Sparsity of NN: statistics



Weight Pruning: from DeepCompression



◆ Pruning + Quantization ▲ Pruning Only □ Quantization Only ◇ SVD



The model has been trained with excessive #epoch.

Sparse Matrix at Runtime

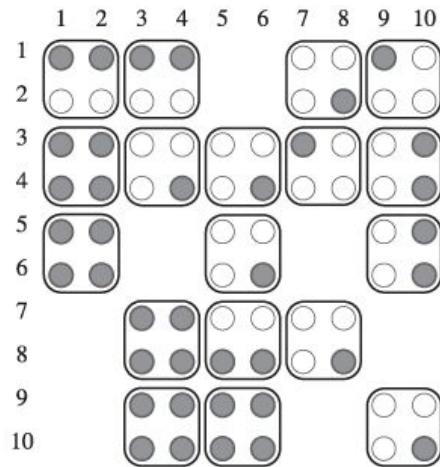
- Sparse Matrix = Discrete Mask + Continuous values
 - Mask cannot be learnt the normal way
 - The values have well-defined gradients
- The matrix value look up need go through a LUT
 - CSR format
 - A: NNZ values
 - IA: accumulated #NNZ of rows
 - JA: the column in the row

$$\begin{pmatrix} 0 & 0 & 0 & 0 \\ 5 & 8 & 0 & 0 \\ 0 & 0 & 3 & 0 \\ 0 & 6 & 0 & 0 \end{pmatrix}$$

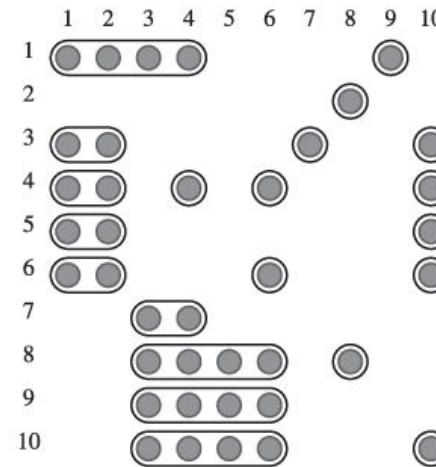
$$\begin{aligned} A &= [5 8 3 6] \\ IA &= [0 0 2 3 4] \\ JA &= [0 1 2 1] \end{aligned}$$

Burden of Sparseness

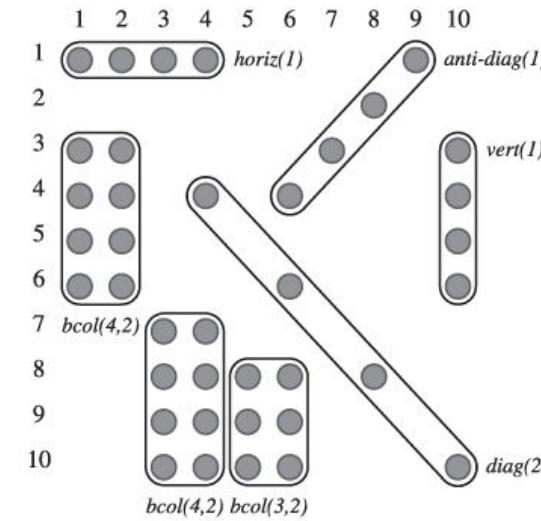
- *Lost of regularity of memory access and computation*
 - Need special hardware for efficient access
 - May need high zero ratio to match dense matrix
 - Matrices will less than 70% zero values, better to treat as dense matrices.



(a) BCSR.



(b) VBL.



(c) CSX.

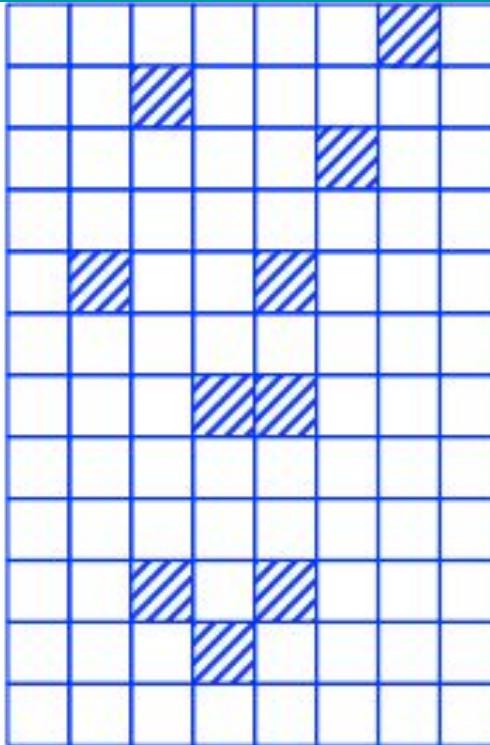
Convolution layers are harder to compress than FC

Table 5: Compression statistics for VGG-16. P: pruning, Q:quantization, H:Huffman coding.

Layer	#Weights	Weights% (P)	Weigh bits (P+Q)	Weight bits (P+Q+H)	Index bits (P+Q)	Index bits (P+Q+H)	Compress rate (P+Q)	Compress rate (P+Q+H)
conv1_1	2K	58%	8	6.8	5	1.7	40.0%	29.97%
conv1_2	37K	22%	8	6.5	5	2.6	9.8%	6.99%
conv2_1	74K	34%	8	5.6	5	2.4	14.3%	8.91%
conv2_2	148K	36%	8	5.9	5	2.3	14.7%	9.31%
conv3_1	295K	53%	8	4.8	5	1.8	21.7%	11.15%
conv3_2	590K	24%	8	4.6	5	2.9	9.7%	5.67%
conv3_3	590K	42%	8	4.6	5	2.2	17.0%	8.96%
conv4_1	1M	32%	8	4.6	5	2.6	13.1%	7.29%
conv4_2	2M	27%	8	4.2	5	2.9	10.9%	5.93%
conv4_3	2M	34%	8	4.4	5	2.5	14.0%	7.47%
conv5_1	2M	35%	8	4.7	5	2.5	14.3%	8.00%
conv5_2	2M	29%	8	4.6	5	2.7	11.7%	6.52%
conv5_3	2M	36%	8	4.6	5	2.3	14.8%	7.79%
fc6	103M	4%	5	3.6	5	3.5	1.6%	1.10%
fc7	17M	4%	5	4	5	4.3	1.5%	1.25%
fc8	4M	23%	5	4	5	3.4	7.1%	5.24%
Total	138M	7.5% (13x)	6.4	4.1	5	3.1	3.2% (31x)	2.05% (49x)

Dynamic Generation of Code

- CVPR'15: Sparse Convolutional Neural Networks
- Relies on compiler for
 - register allocation
 - scheduling
- Good on CPU



Input:

A: 8×12 dense matrix

B: 12×8 sparse matrix

Output:

$$C = A \times B$$

Operations:

$$c_7+ = a_1 \times b_{1,7}$$

$$c_3+ = a_2 \times b_{2,3}$$

$$c_6+ = a_3 \times b_{3,6}$$

$$c_2+ = a_5 \times b_{5,2}$$

$$c_5+ = a_5 \times b_{5,5}$$

$$c_4+ = a_7 \times b_{7,4}$$

$$c_5+ = a_7 \times b_{7,5}$$

$$c_3+ = a_{10} \times b_{10,3}$$

$$c_5+ = a_{10} \times b_{10,5}$$

$$c_4+ = a_{11} \times b_{11,4}$$

Channel Pruning

- Learning the Number of Neurons in Deep Networks [1611](#)

$$\min_{\Theta} \frac{1}{N} \sum_{i=1}^N \ell(y_i, f(\mathbf{x}_i, \Theta)) + r(\Theta)$$

- Channel Pruning for Accelerating Very Deep Neural Networks [1707](#)
 - Also exploits low-rankness of features

$$\arg \min_{\beta, W} \frac{1}{2N} \left\| Y' - \sum_{i=1}^c \beta_i X_i W_i^\top \right\|_F^2$$

subject to $\|\beta\|_0 \leq c'$

Sparse Communication for Distributed Gradient Descent 1704

Algorithm 1 Gradient dropping algorithm given gradient ∇ and dropping rate R .

```
function GRADDROP( $\nabla$ ,  $R$ )
     $\nabla_+ = residuals$ 
    Select threshold:  $R\%$  of  $|\nabla|$  is smaller
    dropped  $\leftarrow 0$ 
    dropped[ $i$ ]  $\leftarrow \nabla[i] \forall i : |\nabla[i]| > threshold$ 
    residuals  $\leftarrow \nabla - dropped$ 
    return sparse(dropped)
end function
```

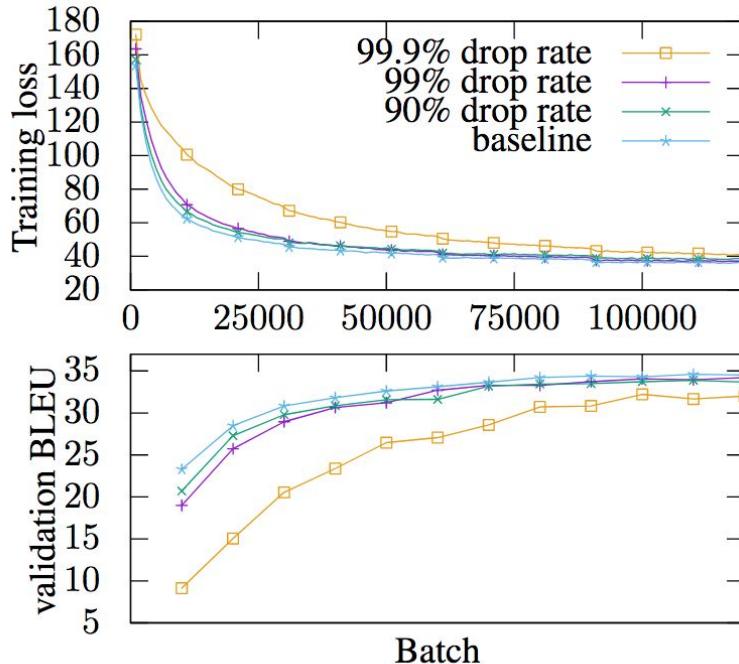
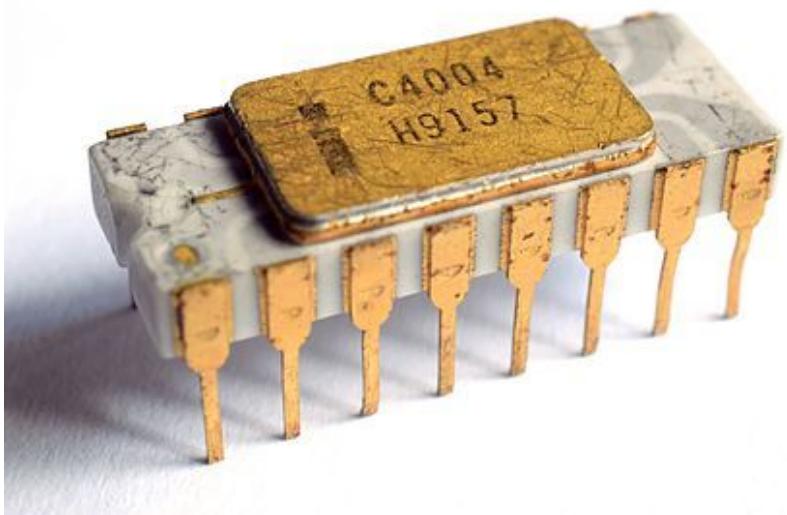


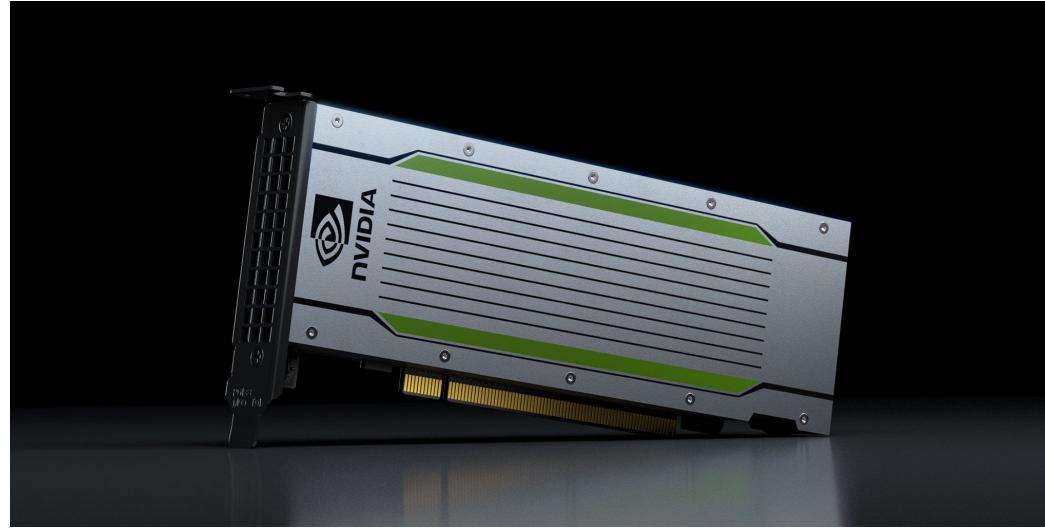
Figure 3: NMT: Training loss and validation BLEU for different dropping ratios.

Quantization

4-bit support in CPU & GPU



Intel® 4004, released 1971
1.1 Kops @ ??

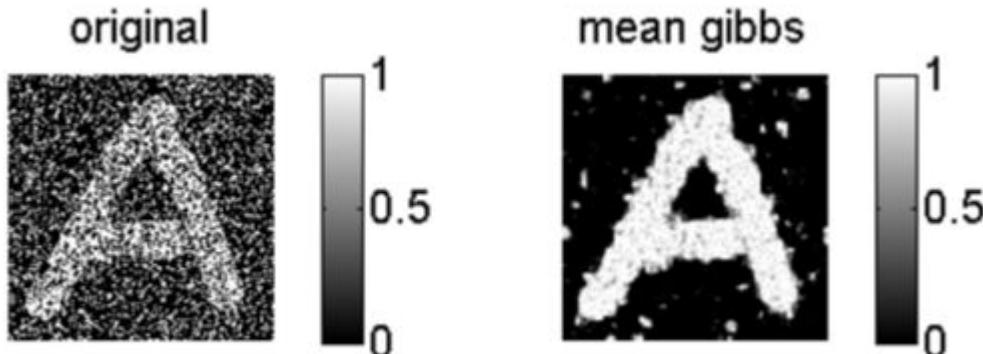


Tesla® T4, released 2018
260 Tops @ 75 Watt

Precursor: Ising model & Boltzmann machine

- Ising model
 - used to model magnetics
 - 1D has trivial analytic solution
 - 2D exhibits phase-transition
 - 2D Ising model can be used for denoising
 - when the mean signal is reliable
- Inference also requires optimization

$$H(\sigma) = - \sum_{\langle i,j \rangle} J_{ij} \sigma_i \sigma_j - \mu \sum_j h_j \sigma_j$$



Quantization

- Linear Quantization
 - Essentially rounding
 - Con: dynamic range of floating point numbers are problematic
- Logarithmic Quantization
 - Quantize $\log(x)$ instead of x
 - Achieve 12 bit quality with 8 bits
 - Con: addition not efficient on existing platforms

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

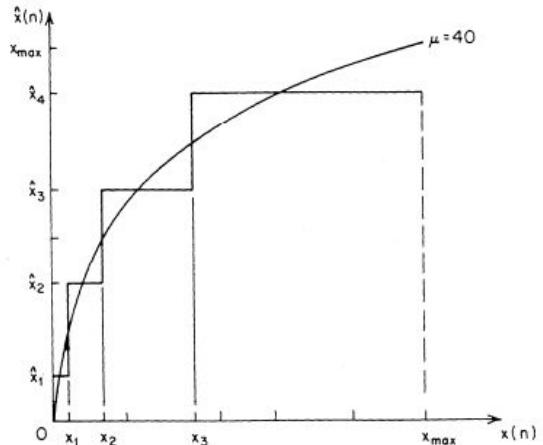
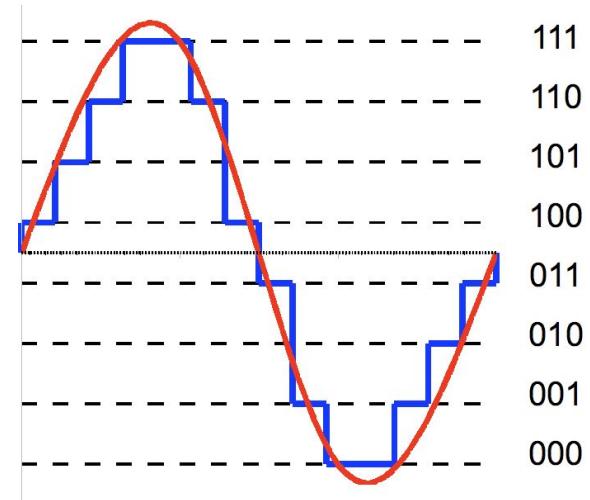


Fig. 5.16 Distribution of quantization levels for a μ -law 3-bit quantizer with $\mu = 40$.

Benefits of Quantized Neural Networks

- Reduce weight size
- Reduce feature size
 - larger feature library size
- Speedup computation (usually require hardware support)
 - $O(\text{Bit-width})$ if SIMD, $O(\text{Bit-width}^2)$ if special hardware
 - Saves bandwidth (intra-chip and off-chip), eases P & R
 - Dense computation
- Distributed training: reduces communication of weights

$$\mathbf{x} \cdot \mathbf{y} = \sum_{m=0}^{M-1} \sum_{k=0}^{K-1} 2^{m+k} \text{bitcount}[and(c_m(\mathbf{x}), c_k(\mathbf{y}))]$$

	INT8	INT4	Binary
Tesla T4	130 T	260 T	
Megvii Zynq 7020	0.05 T	0.2 T	3.2 T

Number format: Floating Point Problems

- Extreme values
 - overflow/underflow, denormal numbers
- Rounding rules
 - Rounding to nearest / directed rounding
- Breaks associativity
 - Kahan summation algorithm
 - Pairwise summation

```
function NeumaierSum(input)
    var sum = input[1]
    var c = 0.0          // A running compensation for lost low-order
bits.
    for i = 2 to input.length do
        var t = sum + input[i]
        if |sum| >= |input[i]| do
            c += (sum - t) + input[i] // If sum is bigger, low-order digits of
[i] are lost.
        else
            c += (input[i] - t) + sum // Else low-order digits of sum are lost
        sum = t
    return sum + c           // Correction only applied once in the very
end
```

Number format: Floating Point vs. Fixed Point

- FP32
- FP16
- “FP8”
- INT32
- INT16
- INT8
- “INT4”
- “INT2”
- Ternary
- Binary

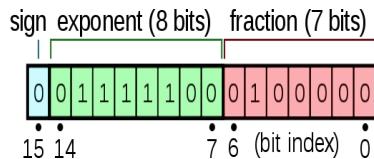
$$1.2345 = \underbrace{12345}_{\text{significand}} \times \underbrace{10^{-4}}_{\text{base}}^{\text{exponent}}.$$

Floating point

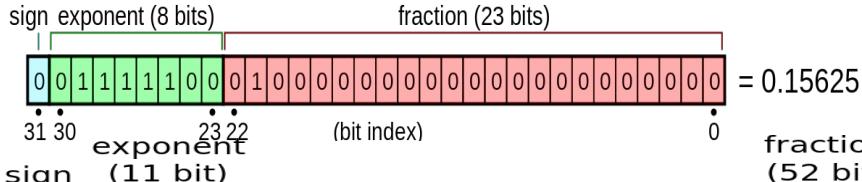
- Sharing exponent
 - Block Floating Point
- Power-of-2 exponent
 - DCT in JPEG

More Low Bit Formats

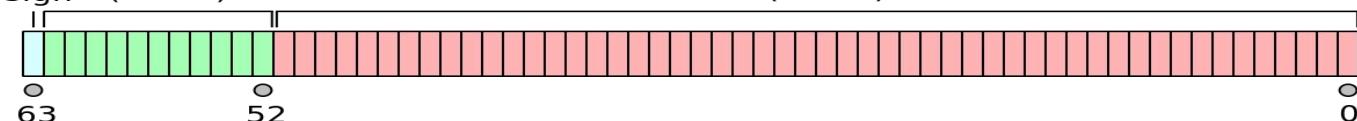
bfloat16



float32



float64



float8 ?

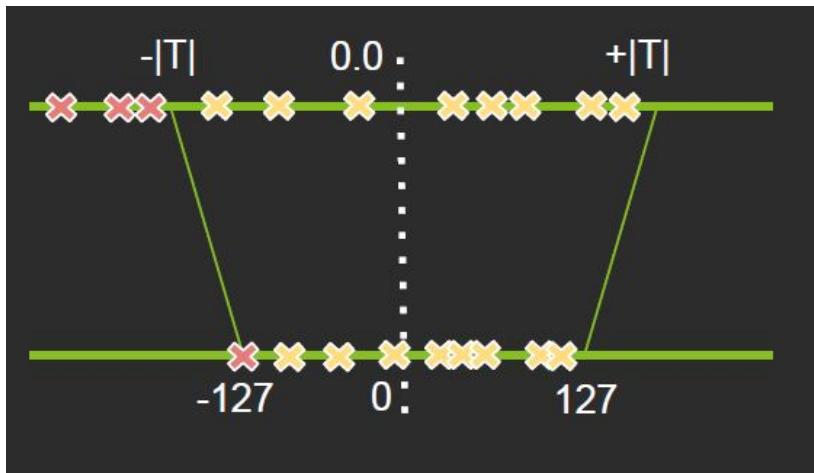
CSD

$x = \pm \sum_{r=1}^R a(r)2^{q(r)}$, 其中 $a(r) = 0, 1$ (unsigned power-of-2)

或 $x = \sum_{r=1}^R a(r)2^{q(r)}$, 其中 $a(r) = -1, 0, 1$ (signed power-of-2)

Posit

Inference Time Quantization



- Only a pre-processing step
- Use a calibration set to determine the distribution
- Quantization error is controlled by trial runs on real dataset

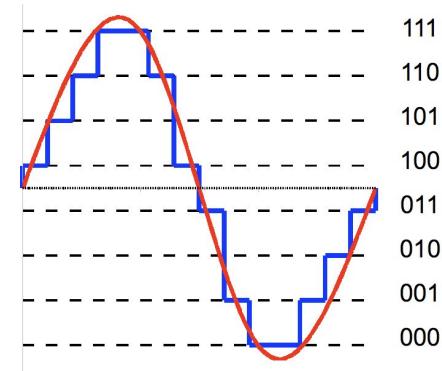
Quantization: During or After Training?

- After training quantization: a universal black-box quantization method

After training black-box quantization					
Claimed	Fine-tune free!	Small calibration set	Universal method	“Simple” quantization	Compress to save bandwidth
In effect	Try quantize for 100 times and pick the best	Difficulty of picking calibration set	Different quantizations for different task	Too complex to mimic in training	Unpredictable errors from intermediate compressions

Differentiable Quantization

- Bengio '13: Estimating or Propagating Gradients Through Stochastic Neurons for Conditional Computation
 - REINFORCE algorithm
 - Decompose binary stochastic neuron into stochastic and differentiable part
 - Injection of additive/multiplicative noise
 - Straight-through estimator



Gradient vanishes after quantization.

Quantization also at Train time

- Neural Network can adapt to the constraints imposed by quantization
- Exploits “Straight-through estimator” (Hinton, Coursera lecture, 2012)

$$\textcolor{blue}{x} \approx \hat{x}$$

\Rightarrow

$$\frac{\partial}{\partial x} \approx \frac{\partial}{\partial \hat{x}}$$

- Example

Forward: $q \sim \text{Bernoulli}(p)$ $\textcolor{blue}{q} \approx \mathbb{E}[q] = p$

Backward: $\frac{\partial c}{\partial p} = \frac{\partial c}{\partial q}.$

STE Alternative: Local Reparameterization Trick

Learning Discrete Weights using the local reparameter trick(Shayer, ICLR 2018)

$$\nabla L(W) = E_{W \in S} \left[\sum_{i=1}^N l(f(x_i, W), y_i) \right]$$

- local reparameterization trick (Kingma&Welling, 2014)

$$E_{p(x)}[f(x)] \quad x = g(\epsilon, \theta) \quad \text{usually Gaussian}$$

$$\nabla_{\theta} E_{p(\epsilon)}[f(g(\epsilon, \theta))] \approx \sum_{i=1}^m \nabla_{\theta} f(g(\epsilon_i, \theta)) \quad \text{Monte Carlo approximation}$$

- sampling weights -> sampling pre-activations

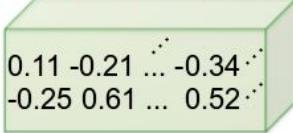
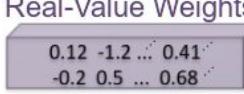
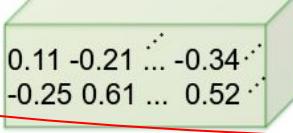
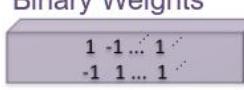
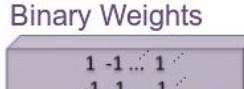
$$W_{ij,l} \sim N(\mu_{ij,l}, \sigma_{ij,l}) \quad z_{i,l} \sim N\left(\sum_j \mu_{ij,l} h_{j,l-1}, \sum_j \sigma_{ij,l}^2 h_{j,l-1}^2\right)$$

Bit Neural Network

- Matthieu Courbariaux et al. BinaryConnect: Training Deep Neural Networks with binary weights during propagations. <http://arxiv.org/abs/1511.00363>
- Itay Hubara et al. Binarized Neural Networks <https://arxiv.org/abs/1602.02505v3>
- Matthieu Courbariaux et al. Binarized Neural Networks: Training Neural Networks with Weights and Activations Constrained to +1 or -1. <http://arxiv.org/pdf/1602.02830v3.pdf>
- Rastegari et al. XNOR-Net: ImageNet Classification Using Binary Convolutional Neural Networks <http://arxiv.org/pdf/1603.05279v1.pdf>
- Zhou et al. DoReFa-Net: Training Low Bitwidth Convolutional Neural Networks with Low Bitwidth Gradients <https://arxiv.org/abs/1606.06160>
- Hubara et al. Quantized Neural Networks: Training Neural Networks with Low Precision Weights and Activations <https://arxiv.org/abs/1609.07061>

Binarizing AlexNet

Theoretical

	Network Variations	Operations used in Convolution	Memory Saving (Inference)	Time Saving on CPU (Inference)	Accuracy on ImageNet (AlexNet)
Standard Convolution	<p>Real-Value Inputs</p>  <p>Real-Value Weights</p> 	+ , - , ×	1x	1x	%56.7
Binary Weight	<p>Real-Value Inputs</p>  <p>Binary Weights</p> 	+ , -	~32x	~2x	%53.8
BinaryWeight Binary Input (XNOR-Net)	<p>Binary Inputs</p>  <p>Binary Weights</p> 	XNOR , bitcount	~32x	~58x	%44.2

Scaled binarization

$$\min_{\Lambda \in \text{diagonal}, B \in \{1, -1\}^{m \times n}} \|\Lambda B - W\|_F^2$$

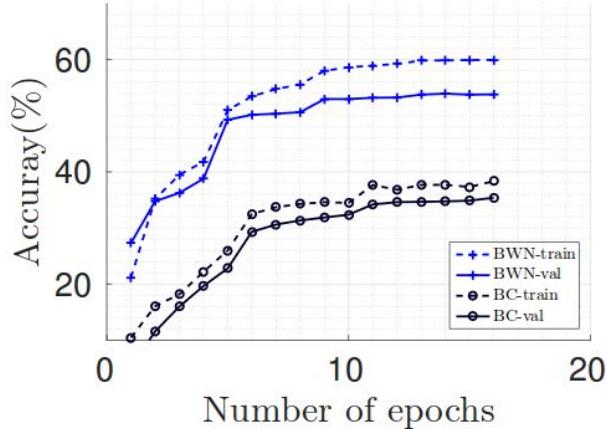
- Sol:

$$\begin{aligned}\|\Lambda B - W\|_F^2 &= \|(\Lambda B) \circ B - W \circ B\|_F^2 \\ &= \|\Lambda(B \circ B) - W \circ B\|_F^2 \\ &= \|\Lambda \mathbf{1} - W \circ B\|_F^2\end{aligned}$$

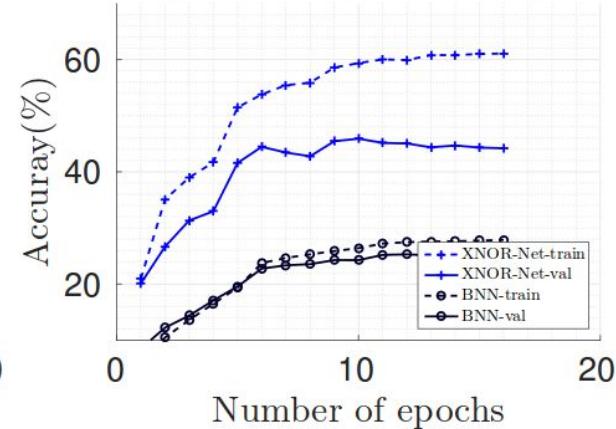
= Variance of rows of $W \circ B$

XNOR-Net

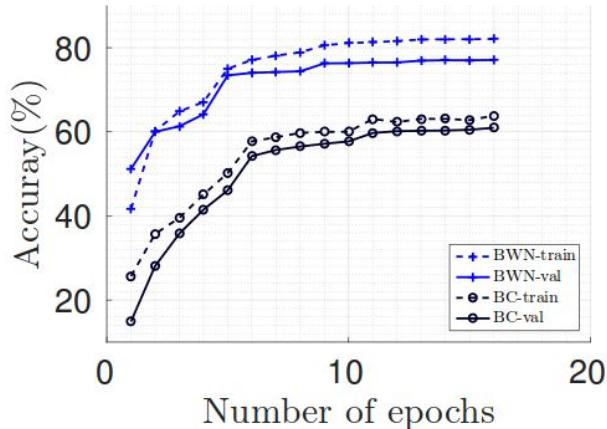
Top-1, Binary-Weight



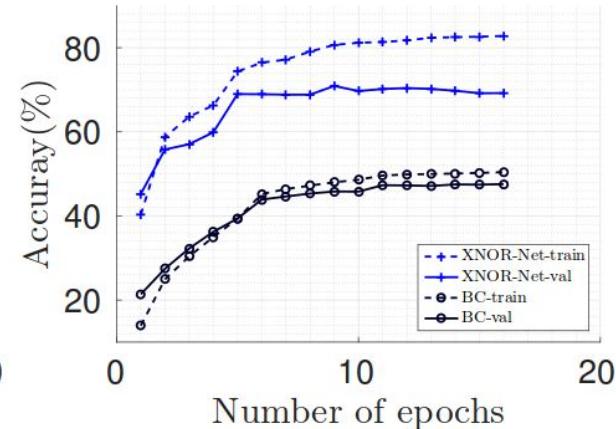
Top-1, Binary-Weight-Input



Top-5, Binary-Weight



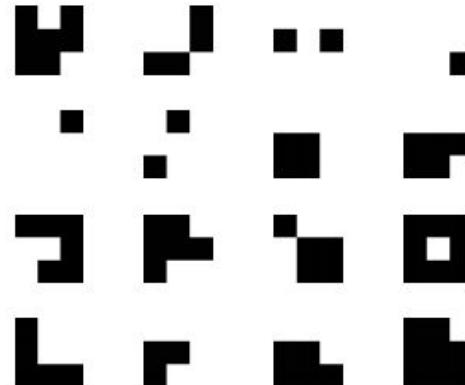
Top-5, Binary-Weight-Input



Binary weights network

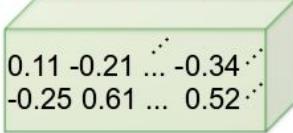
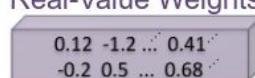
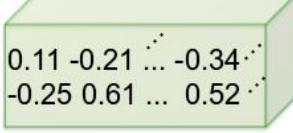
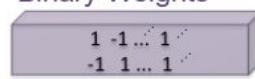
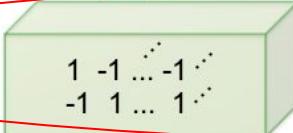
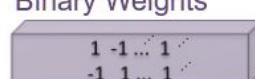
- Filter repetition
 - 3x3 binary kernel has only 256 patterns modulo sign.
 - 3x1 binary kernel only has only 4 patterns modulo sign.
 - Not easily exploitable as we are applying CHW as filter

Figure 2: Binary weight filters, sampled from of the first convolution layer. Since we have only 2^{k^2} unique 2D filters (where k is the filter size), filter replication is very common. For instance, on our CIFAR-10 ConvNet, only 42% of the filters are unique.



Binarizing AlexNet

Theoretical

	Network Variations	Operations used in Convolution	Memory Saving (Inference)	Time Saving on CPU (Inference)	Accuracy on ImageNet (AlexNet)
Standard Convolution	Real-Value Inputs  Real-Value Weights 	+ , - , ×	1x	1x	%56.7
Binary Weight	Real-Value Inputs  Binary Weights 	+ , -	~32x	~2x	%53.8
Binary Weight Binary Input (XNOR-Net)	Binary Inputs  Binary Weights 	XNOR , bitcount	~32x	~58x	%44.2

Scaled binarization is no longer exact and not found to be useful

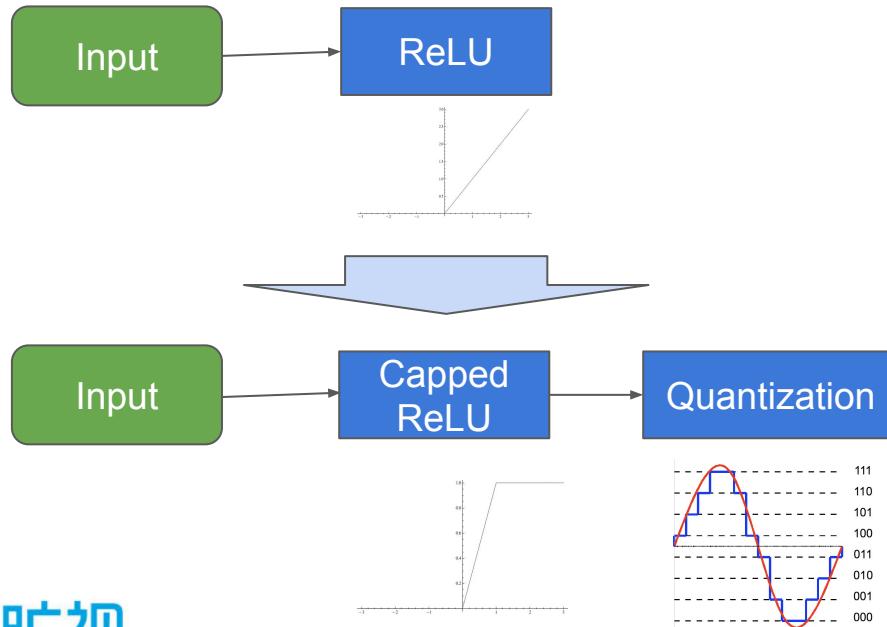
$$\alpha^*, \mathbf{B}^*, \beta^*, \mathbf{H}^* = \operatorname{argmin}_{\alpha, \mathbf{B}, \beta, \mathbf{H}} \|\mathbf{X}^\top \mathbf{W} - \beta \alpha \mathbf{H}^\top \mathbf{B}\|$$

The solution below is quite bad, like when $\mathbf{Y} = [-4, 1]$

$$\mathbf{C}^* = \operatorname{sign}(\mathbf{Y}) = \operatorname{sign}(\mathbf{X}^\top) \operatorname{sign}(\mathbf{W}) = \mathbf{H}^{*\top} \mathbf{B}^*$$

Quantization of Activations

- XNOR-net adopted STE method in their open-source our code

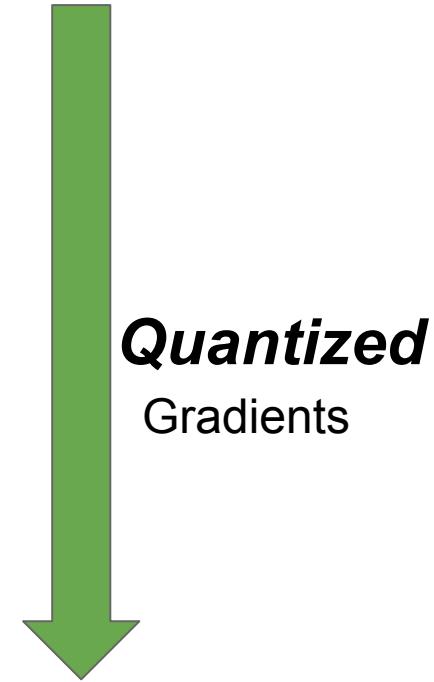
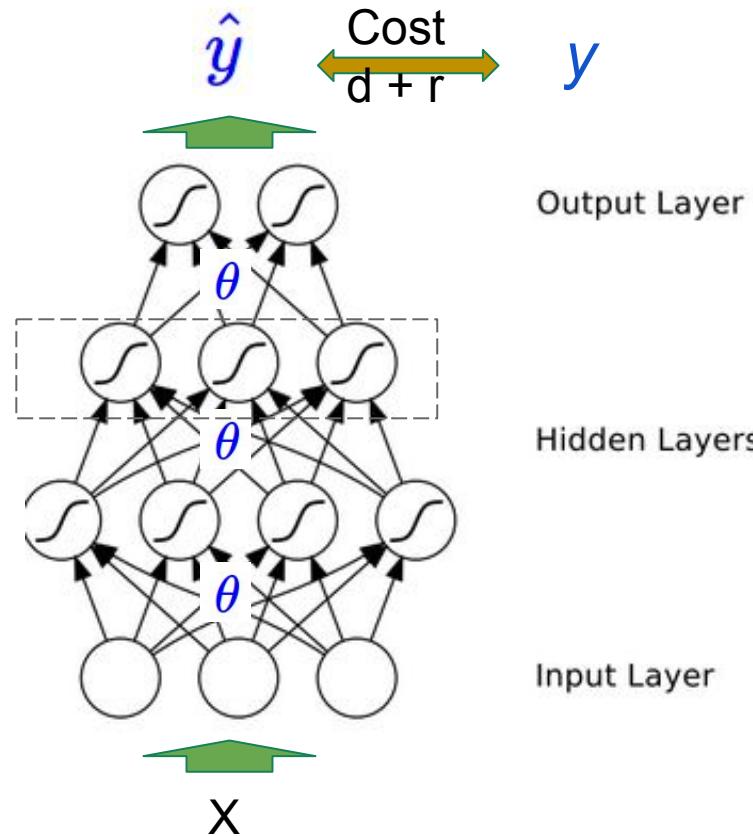


Neural Network Training

Quantized

Activations/
Feature maps/
Neurons

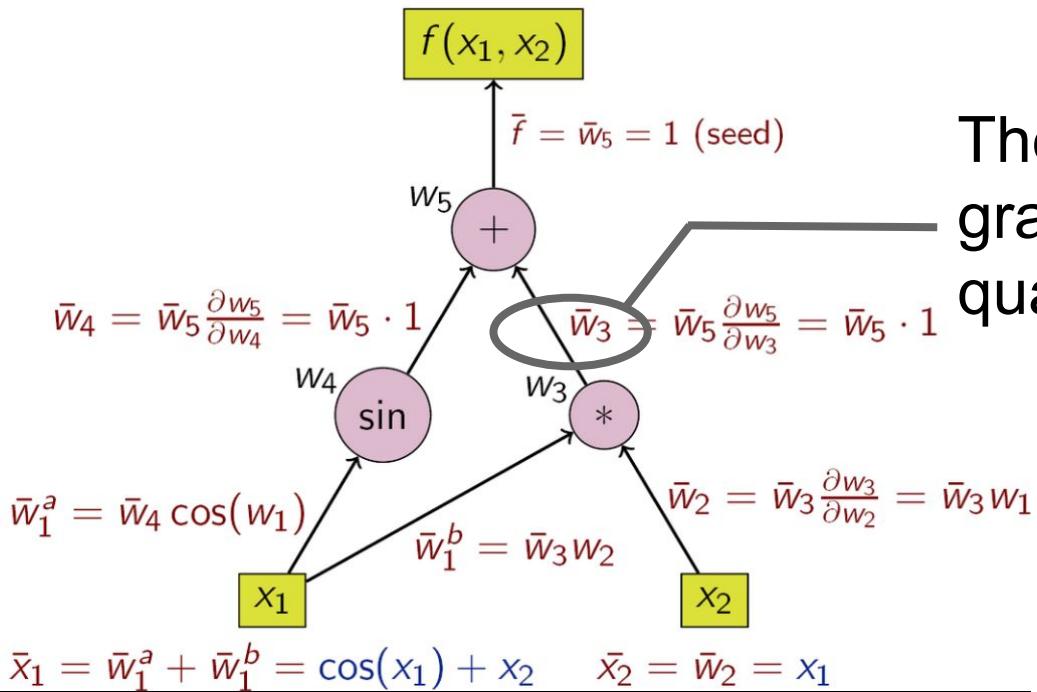
Quantized θ



Backpropagation

$$\frac{\partial y}{\partial x} = \frac{\partial y}{\partial w_1} \frac{\partial w_1}{\partial x} = \left(\frac{\partial y}{\partial w_2} \frac{\partial w_2}{\partial w_1} \right) \frac{\partial w_1}{\partial x} = \left(\left(\frac{\partial y}{\partial w_3} \frac{\partial w_3}{\partial w_2} \right) \frac{\partial w_2}{\partial w_1} \right) \frac{\partial w_1}{\partial x} = \dots$$

Backward propagation
of derivative values



There will be no gradient flow if we quantize somewhere!

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
- Future work
 - BN requires FP computation during training
 - Require FP weights for accumulating gradients

$$\text{uniform-quant}_k(\mathbf{W}) \stackrel{\text{def}}{=} 2 \max(|\mathbf{W}|) \text{quant}_k\left(\frac{\mathbf{W}}{2 \max(|\mathbf{W}|)}\right)$$



Floating Point
Scaling Factor

Integer

Table 1: Comparison of prediction accuracy for SVHN with different choices of Bit-width in a DoReFa-Net. W , A , G are bitwidths of weights, activations and gradients respectively. When bitwidth is 32, we simply remove the quantization functions.

W	A	G	Training Complexity	Inference Complexity	Storage Relative Size	Model A Accuracy	Model B Accuracy	Model C Accuracy	Model D Accuracy
1	1	2	3	1	1	0.934	0.924	0.910	0.803
1	1	4	5	1	1	0.968	0.961	0.916	0.846
1	1	8	9	1	1	0.970	0.962	0.902	0.828
1	1	32	-	-	1	0.971	0.963	0.921	0.841
1	2	2	4	2	1	0.909	0.930	0.900	0.808
1	2	3	5	2	1	0.968	0.964	0.934	0.878
1	2	4	6	2	1	0.975	0.969	0.939	0.878
2	1	2	6	2	2	0.927	0.928	0.909	0.846
2	1	4	10	2	2	0.969	0.957	0.904	0.827
1	2	8	10	2	1	0.975	0.971	0.946	0.866
1	2	32	-	-	1	0.976	0.970	0.950	0.865
1	3	3	6	3	1	0.968	0.964	0.946	0.887
1	3	4	7	3	1	0.974	0.974	0.959	0.897
1	3	6	9	3	1	0.977	0.974	0.949	0.916
1	4	2	6	4	1	0.815	0.898	0.911	0.868
1	4	4	8	4	1	0.975	0.974	0.962	0.915
1	4	8	12	4	1	0.977	0.975	0.955	0.895
2	2	2	8	4	1	0.900	0.919	0.856	0.842
8	8	8	-	-	8			0.970	0.955
32	32	32	-	-	32	0.975	0.975	0.972	0.950



A	B	C	D
---	---	---	---

A has two times as many channels as B.
B has two times as many channels as C.

...

Table 1: Comparison of prediction accuracy for SVHN with different choices of Bit-width in a DoReFa-Net. W , A , G are bitwidths of weights, activations and gradients respectively. When bitwidth is 32, we simply remove the quantization functions.

W	A	G	Training Complexity	Inference Complexity	Storage Relative Size	Model A Accuracy	Model B Accuracy	Model C Accuracy	Model D Accuracy
1	1	2	3	1	1	0.934	0.924	0.910	0.803
1	1	4	5	1	1	0.968	0.961	0.916	0.846
1	1	8	9	1	1	0.970	0.962	0.902	0.828
1	1	32	-	-	1	0.971	0.963	0.921	0.841
1	2	2	4	2	1	0.909	0.930	0.900	0.808
1	2	3	5	2	1	0.068	0.064	0.034	0.078
1	2	4	6						
2	1	2	6						
2	1	4	10						
1	2	8	10						
1	2	32	-						
1	3	3	6						
1	3	4	7	3	1	0.974	0.974	0.959	0.897
1	3	6	9	3	1	0.977	0.974	0.949	0.916
1	4	2	6	4	1	0.815	0.898	0.911	0.868
1	4	4	8	4	1	0.975	0.974	0.962	0.915
1	4	8	12	4	1	0.977	0.975	0.955	0.895
2	2	2	8	4	1	0.900	0.919	0.856	0.842
8	8	8	-	-	8			0.970	0.955
32	32	32	-	-	32	0.975	0.975	0.972	0.950

Accuracy \propto #bit operations

= #input-channel * #output-channel *
input-bitwidth * output-bitwidth



A has two times as many channels as B.
B has two times as many channels as C.

...

Quantization Methods

- Deterministic Quantization

$$Q_k^{det}(\mathbf{X}) = \alpha Q_k(\tilde{\mathbf{X}}) + \beta \approx \mathbf{X}$$

- Stochastic Quantizaiton

$$Q_k^{stoc}(\mathbf{X}) = \alpha Q_k(\tilde{\mathbf{X}} + \frac{\xi}{2^k - 1}) + \beta \approx \mathbf{X}$$

$$\xi \sim U(-\frac{1}{2}, \frac{1}{2})$$

Injection of noise realizes the sampling.

$$\tilde{\mathbf{X}} = \frac{\mathbf{X} - \beta}{\alpha}$$

$$\alpha = \max(\mathbf{X}) - \min(\mathbf{X})$$

$$\beta = \min(\mathbf{X})$$

Quantization of Weights, Activations and Gradients

- A half #channel 2-bit AlexNet (same bit complexity as XNOR-net)

Quantization Method	Balanced Deterministic	Unbalanced Deterministic	Stochastic
Weights	0.469	0.346	0.120
Activations	0.315	0.469	diverge
Gradients	diverge	diverge	0.469
XNOR-net		0.442*	
BNN		0.279*	

Quantization Error measured by Cosine Similarity

- W_{b_sn} is n-bit quantizaiton of real W
- x is Gaussian R. V. clipped by \tanh

W_{b_s1}	W_{b_s2}	W_{b_s3}	W_{b_s4}	W_{b_s6}	W_{b_s8}	W
0.684	0.684	0.699	0.795	0.882	0.888	0.888 x_b
0.742	0.742	0.758	0.862	0.957	0.964	0.964 x_2
0.763	0.763	0.780	0.887	0.985	0.991	0.992 x_3
0.768	0.768	0.785	0.893	0.991	0.998	0.998 x_4
0.769	0.770	0.786	0.894	0.993	0.999	1.000 x_6
0.769	0.770	0.786	0.895	0.993	1.000	1.000 x_8
0.769	0.770	0.786	0.895	0.993	1.000	1.000 x



Saturates

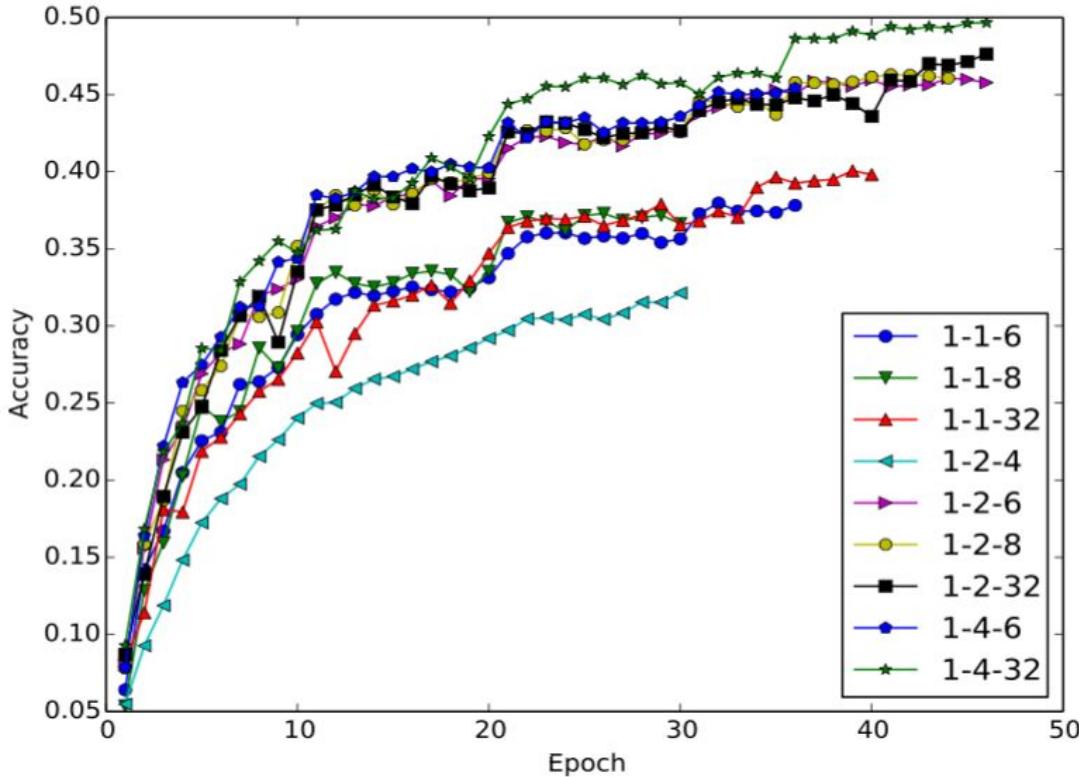
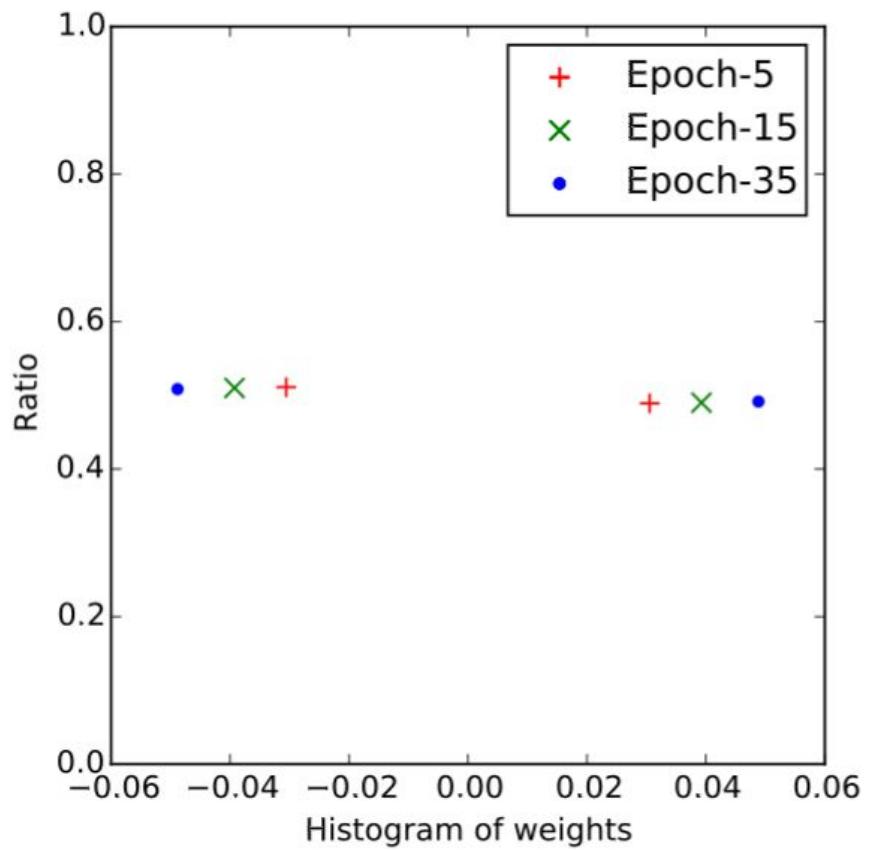
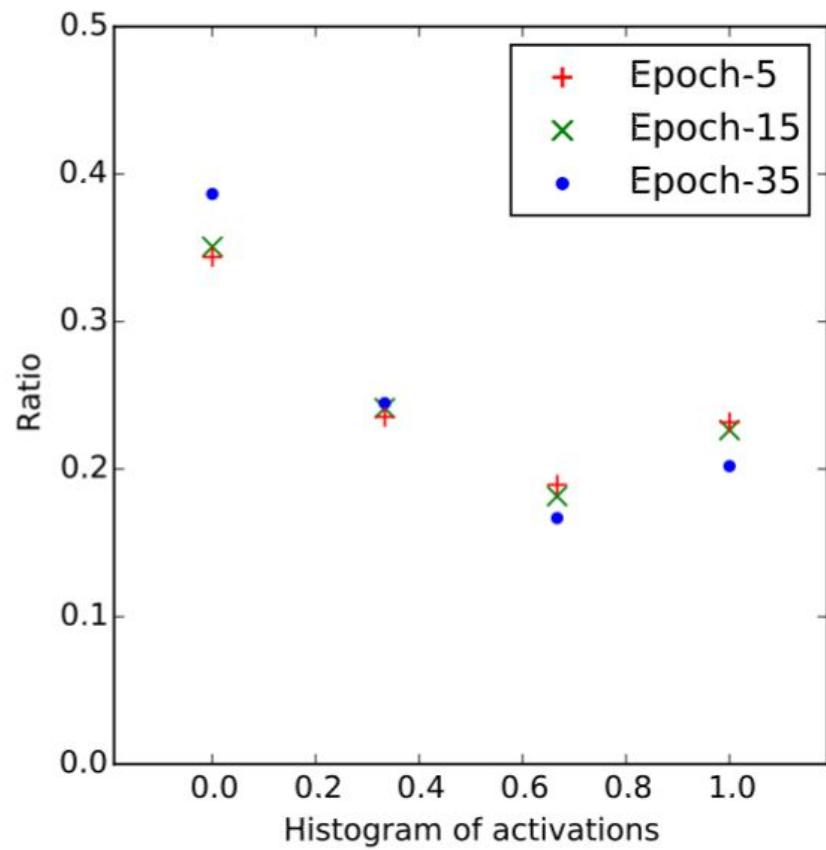


Figure 1: Prediction accuracy of AlexNet variants on Validation Set of ImageNet indexed by epoch number. “W-A-G” gives the specification of bitwidths of weights, activations and gradients. E.g., “1-2-4” stands for the case when weights are 1-bit, activations are 2-bit and gradients are 4-bit. The figure is best viewed in color.



(a)



(b)

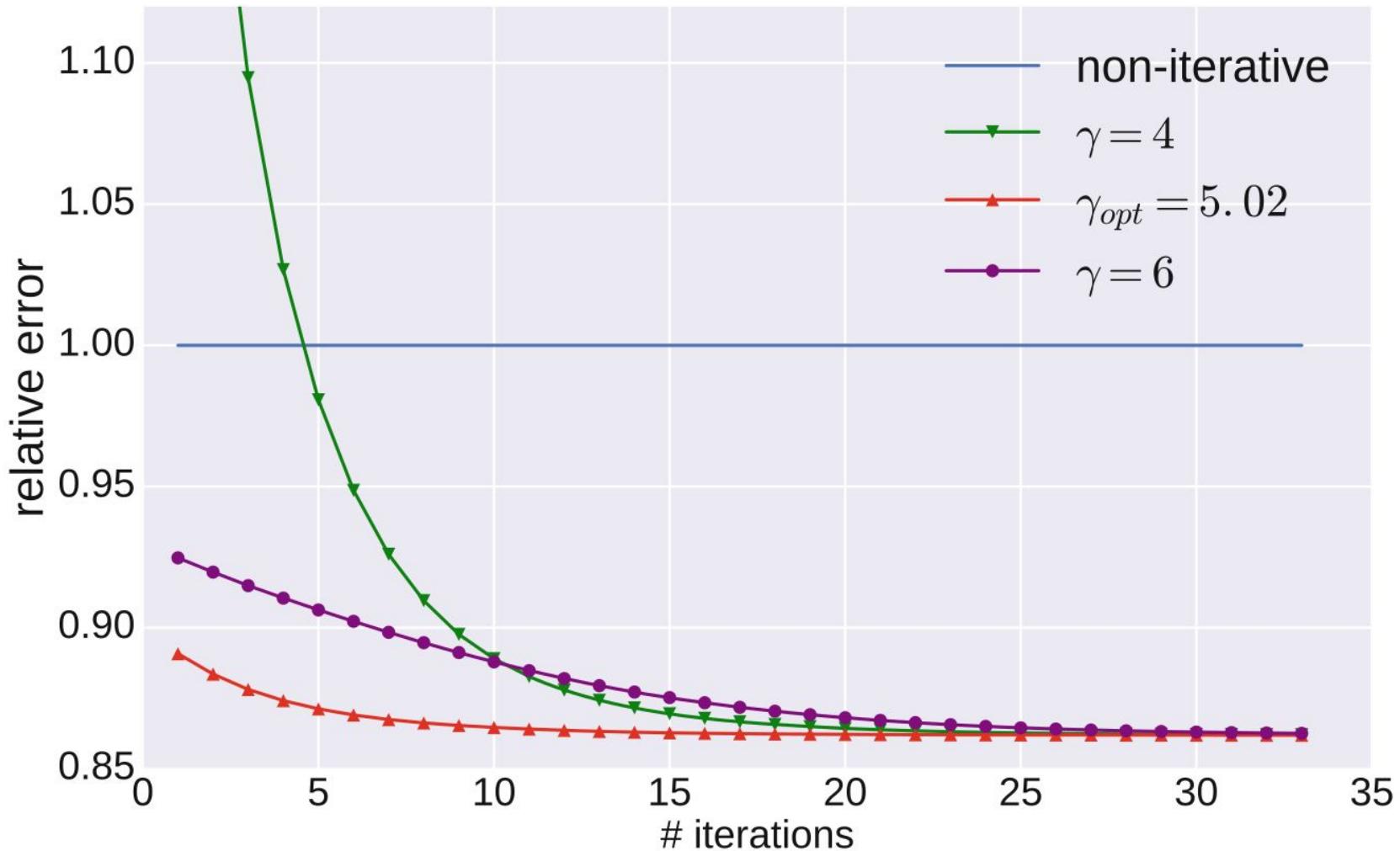
Figure 3: (a) is histogram of weights of layer “conv3” of “1-2-6” AlexNet model at epoch 5, 15

Optimizing the Scaling Factor

IQNN: Training Quantized Neural Networks with Iterative Optimizations

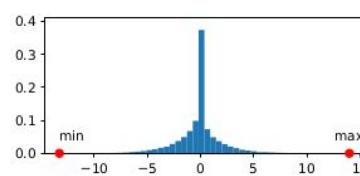
$$\min_{\Lambda, Q} \|\Lambda Q - W\|_F$$

- Previous methods approximate this
 - DoReFa-net: 2 max ($|W|$)
 - Balanced Quantization: scaled mean($|W|$) or median($|W|$) when 2-bit (more complex when 4-bit)

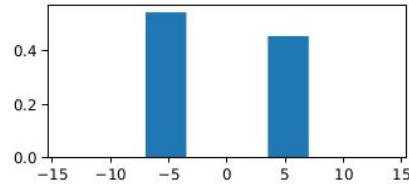


Effective Quantization Methods for Recurrent Neural Networks 2016

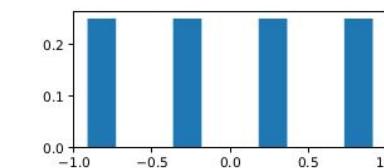
Model	weight-bits	activation-bits	PPW	
			balanced	unbalanced
LSTM	2	2	152	164
LSTM	2	3	142	155
LSTM <small>(Hubara et al., 2016a)</small>	2	3	220	
LSTM <small>(Hubara et al., 2016a)</small>	4	4	100	



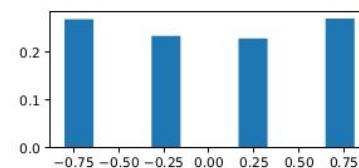
(a) floating point copy of weights in QNN after 60 epochs



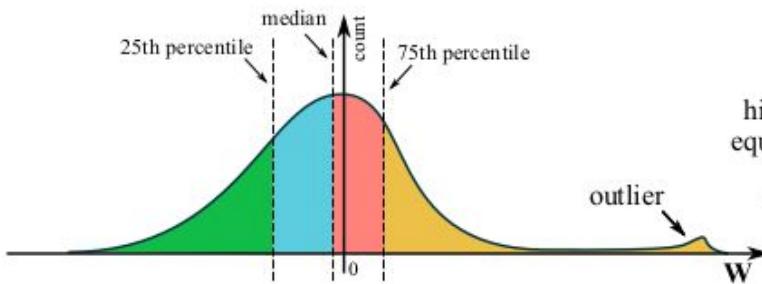
(b) imbalanced quantization (no equalization)



(c) balanced quantization with mean-dian (before matching value range)



(d) balanced quantization with mean-dian (before matching value range)



(a) histogram 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.

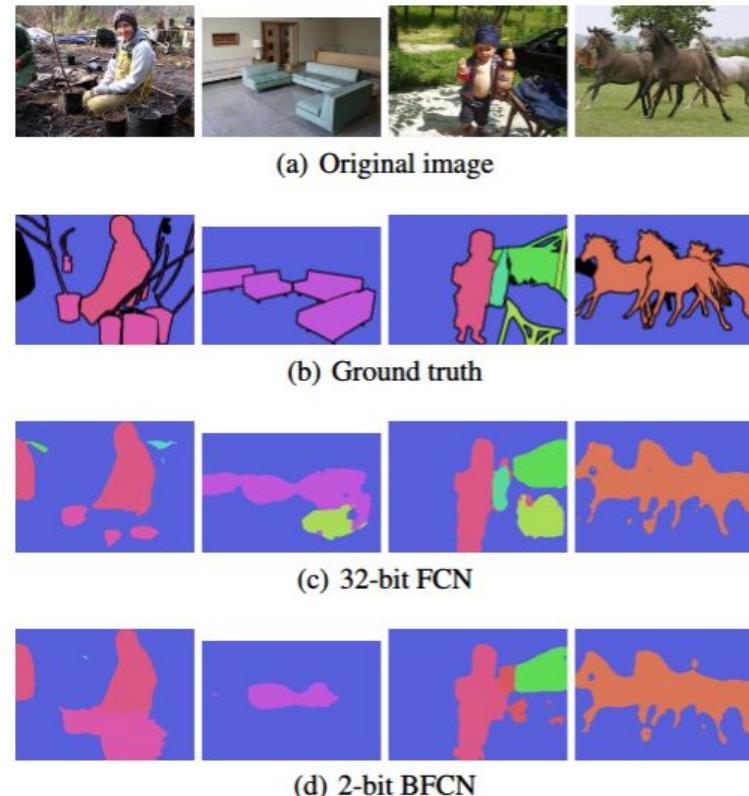
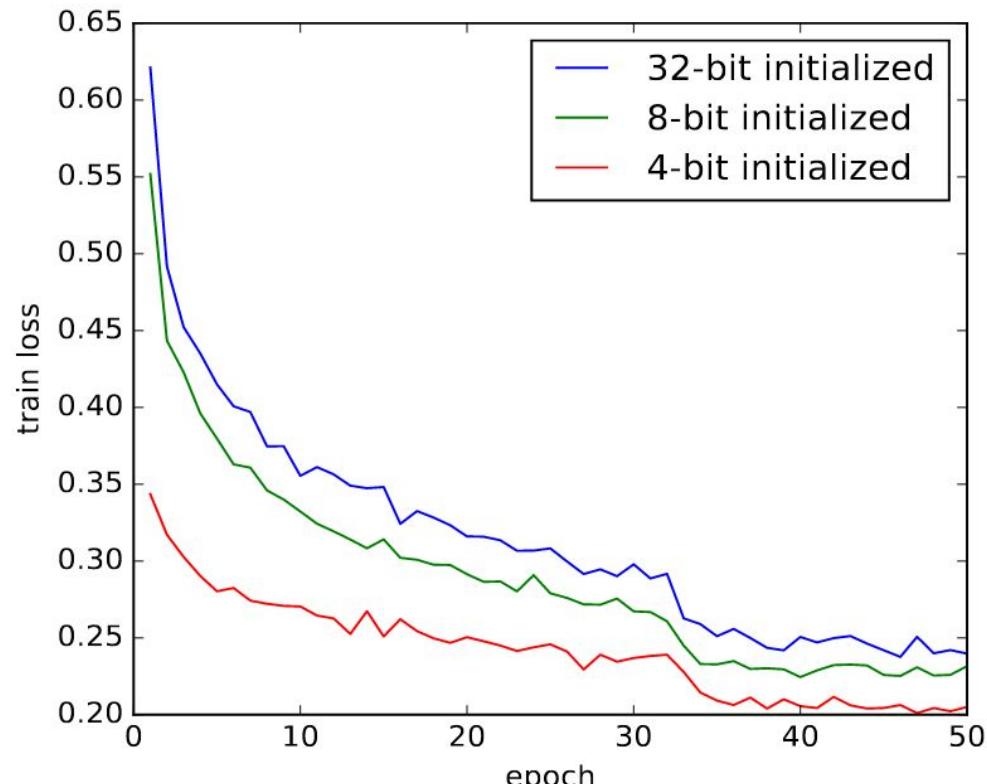
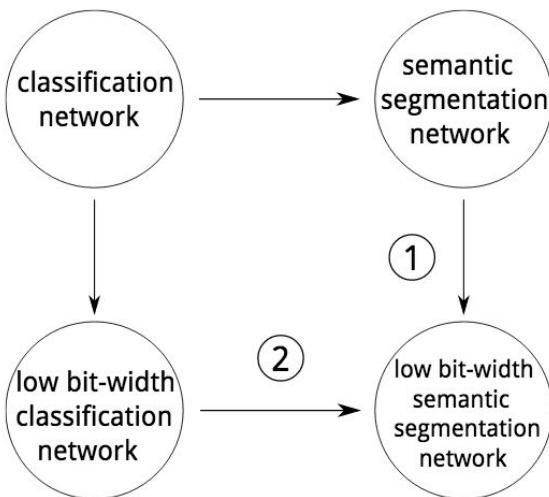
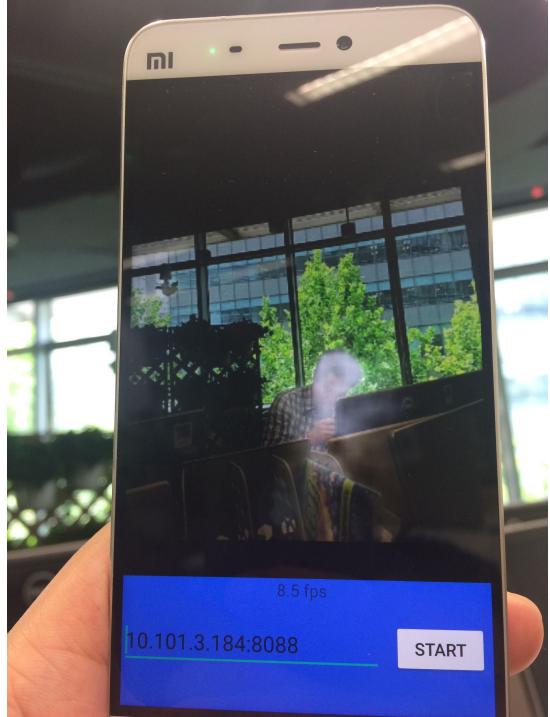
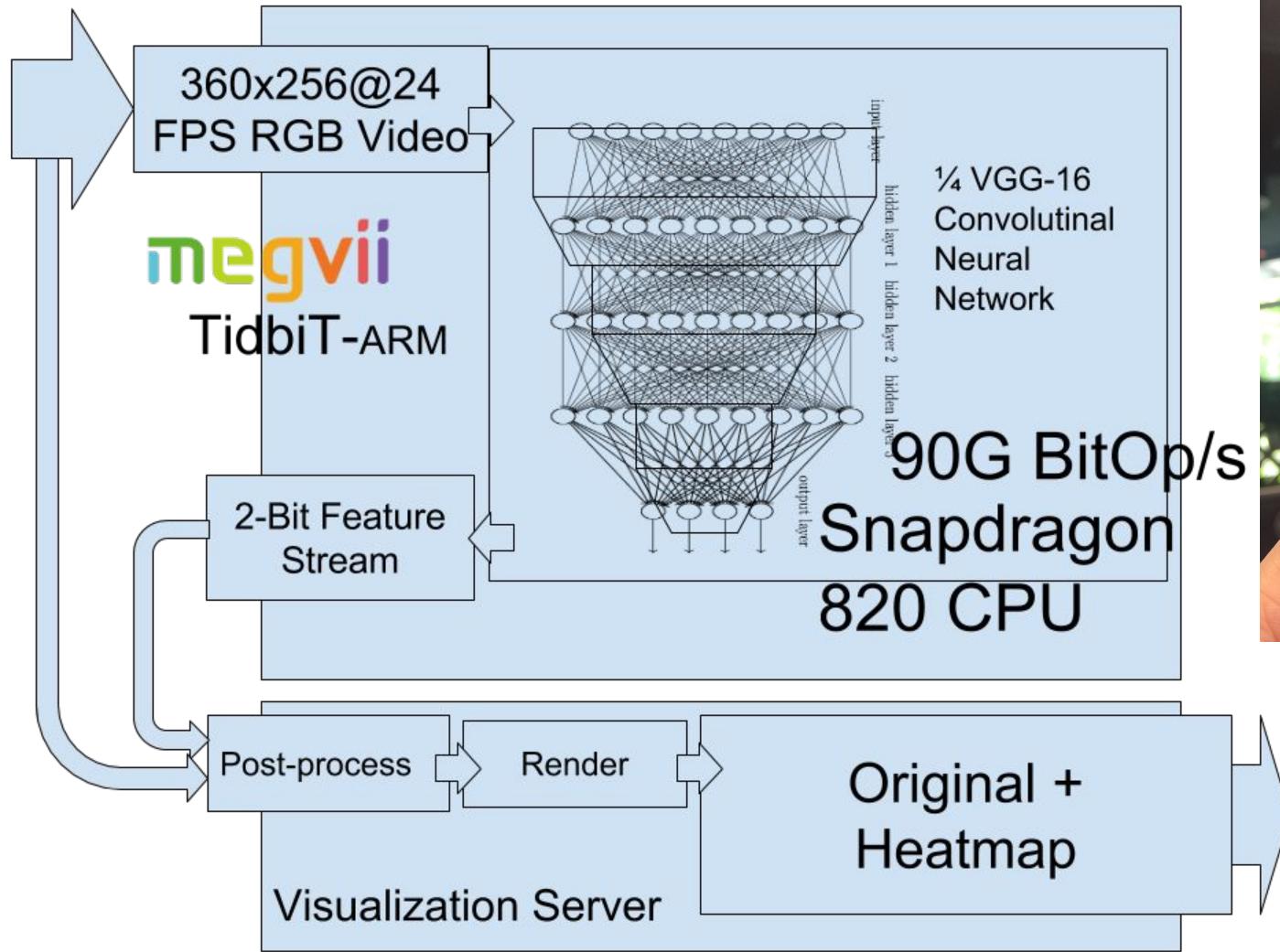


Figure 4: Examples on PASCAL VOC 2012.

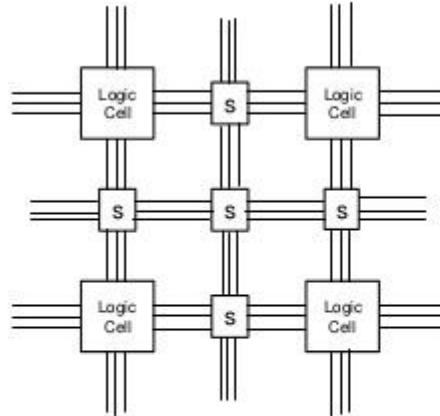
Quantization tricks: Bitwidth decay



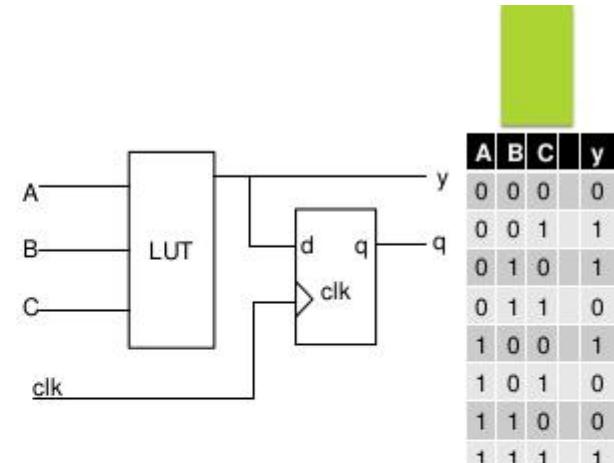


FPGA is made up of many LUT's

FPGA

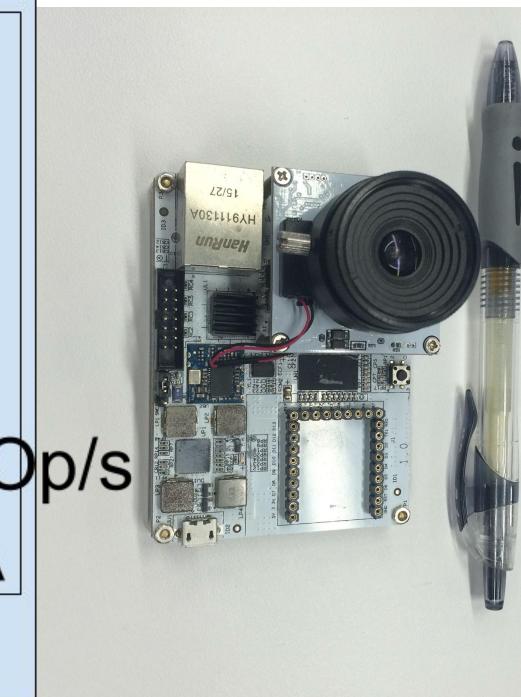
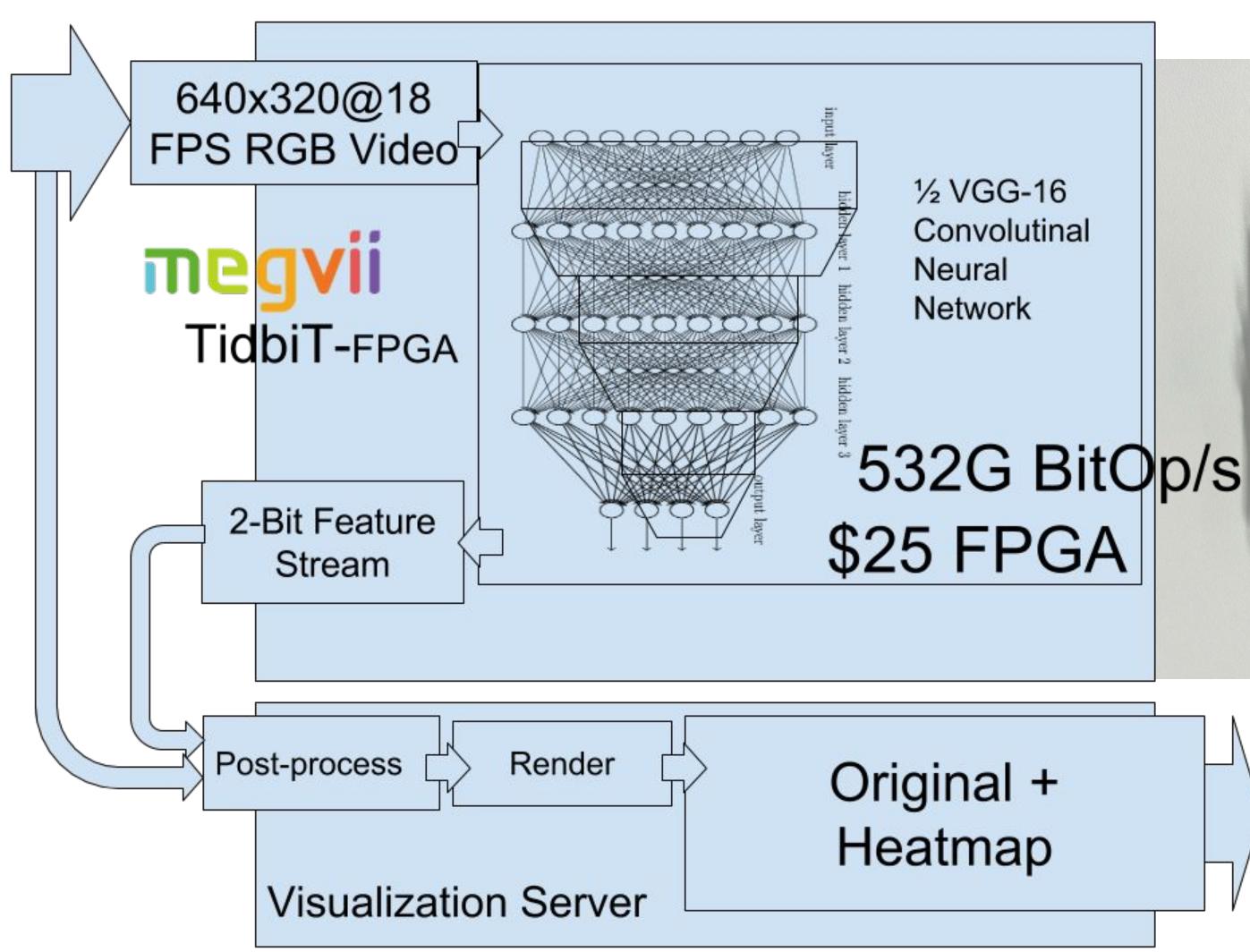


a) Conceptual structure of an FPGA device.



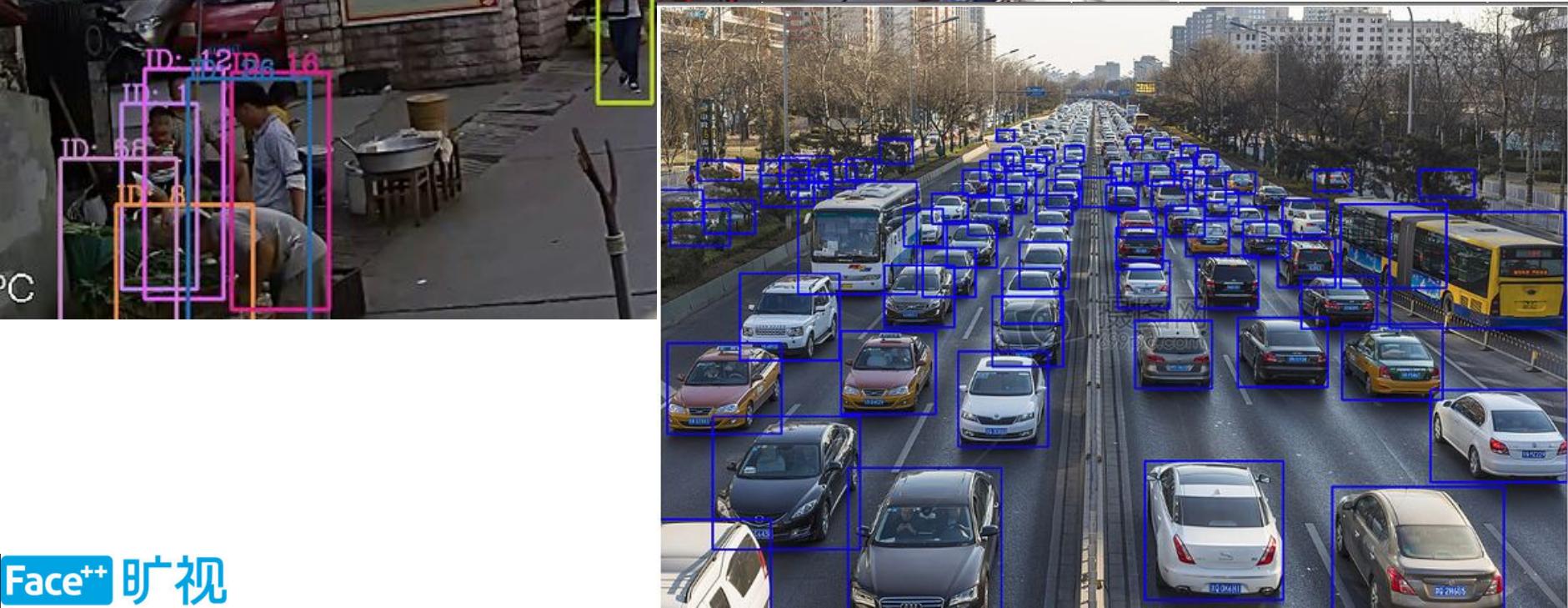
b) Three-input LUT-based logic cell

Flow



Edge Computing





More References

- Xiangyu Zhang, Jianhua Zou, Kaiming He, Jian Sun: Accelerating Very Deep Convolutional Networks for Classification and Detection. IEEE Trans. Pattern Anal. Mach. Intell. 38(10): 1943-1955 (2016)
- ShuffleNet: An Extremely Efficient Convolutional Neural Network for Mobile Devices
<https://arxiv.org/abs/1707.01083>
- Aggregated Residual Transformations for Deep Neural Networks
<https://arxiv.org/abs/1611.05431>
- Convolutional neural networks with low-rank regularization <https://arxiv.org/abs/1511.06067>
- Optimizing Neural Networks with Kronecker-factored Approximate Curvature
<https://arxiv.org/pdf/1503.05671.pdf>

Backup after this slide

Slide also available at my home page:

<https://zsc.github.io/>



THE #1 PROGRAMMER EXCUSE FOR LEGITIMATELY SLACKING OFF: "MY MODEL'S TRAINING"



TernGrad: Ternary Gradients to Reduce Communication in Distributed Deep Learning [1705](#)

- Weights and activations not quantized.

Table 2: Accuracy comparison for *AlexNet*.

base LR	mini-batch size	workers	iterations	gradients	weight decay	DR [†]	top-1	top-5
0.01	256	2	370K	floating	0.0005	0.5	57.33%	80.56%
				<i>TernGrad</i>	0.0005	0.2	57.61%	80.47%
				<i>TernGrad-noclip</i> [‡]	0.0005	0.2	54.63%	78.16%
0.02	512	4	185K	floating	0.0005	0.5	57.32%	80.73%
				<i>TernGrad</i>	0.0005	0.2	57.28%	80.23%
0.04	1024	8	92.5K	floating	0.0005	0.5	56.62%	80.28%
				<i>TernGrad</i>	0.0005	0.2	57.54%	80.25%

[†] DR: dropout ratio, the ratio of dropped neurons. [‡] *TernGrad* without gradient clipping.

Low-rankness of Activations

- Accelerating Very Deep Convolutional Networks for Classification and Detection