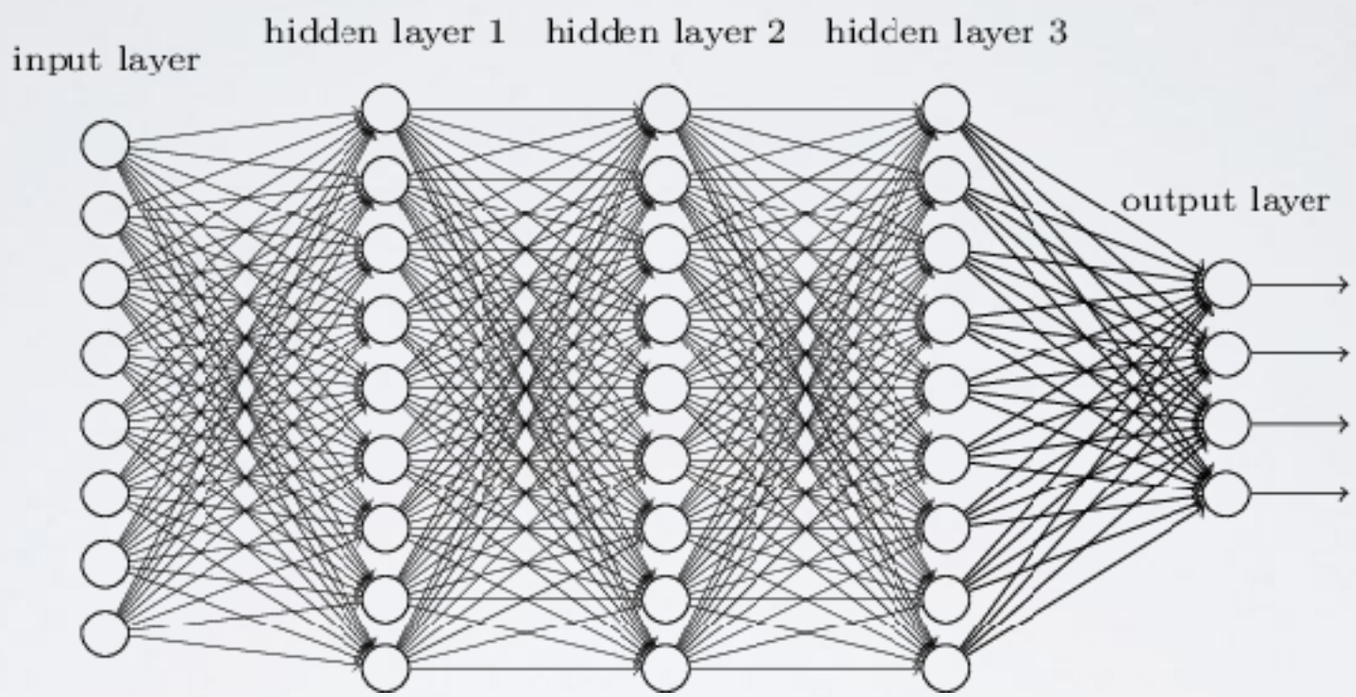
XNOR-Net

ImageNet Classification Using Binary Convolutional Neural Networks

Mohammad Rastegari
Vicente Ordonez
Joseph Redmon
Ali Farhadi

Presentation by Naveen

Deep Neural Networks are Complicated (And Huge!)

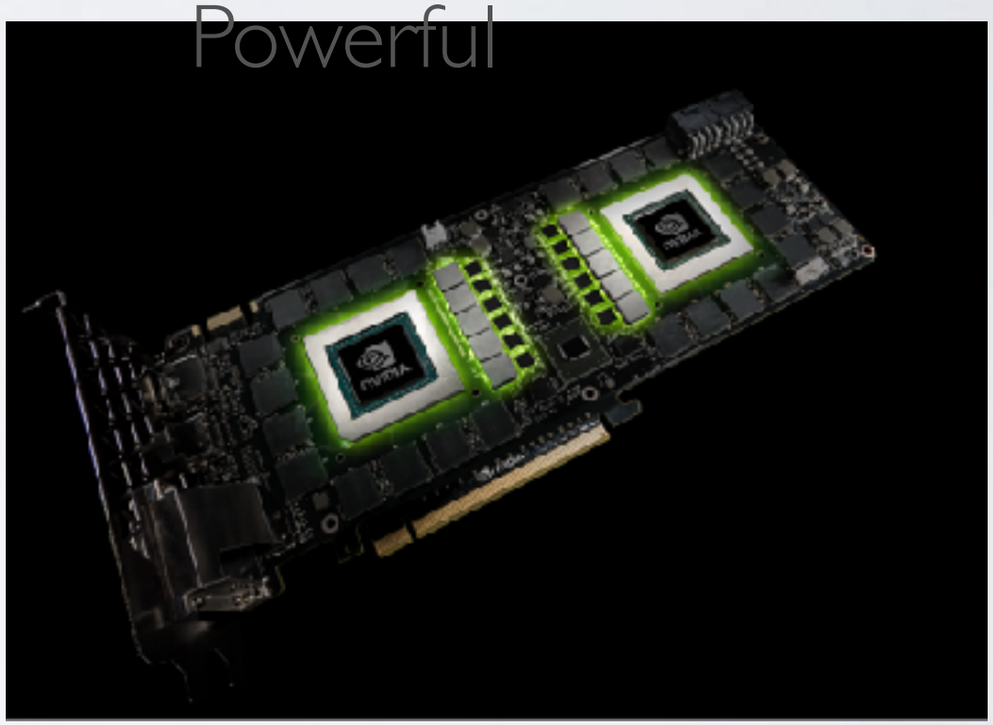


Remember **HW2** - Size of AlexNet?

CPU vs GPU

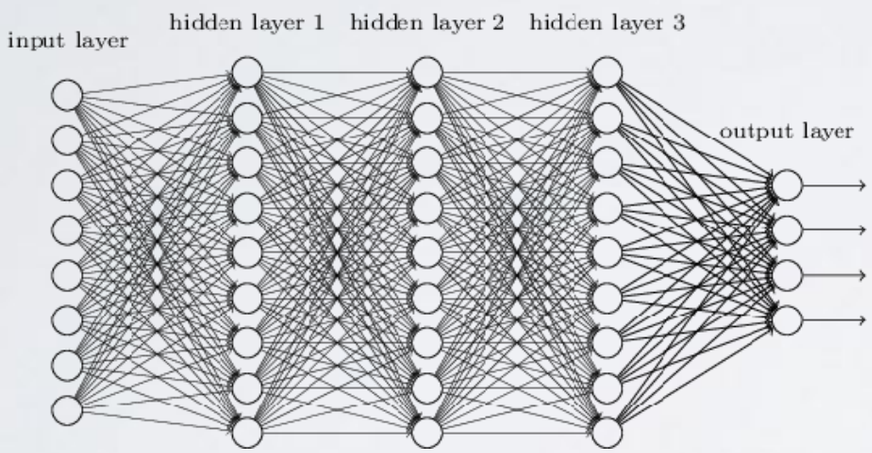
Small
Weak
Scrawny

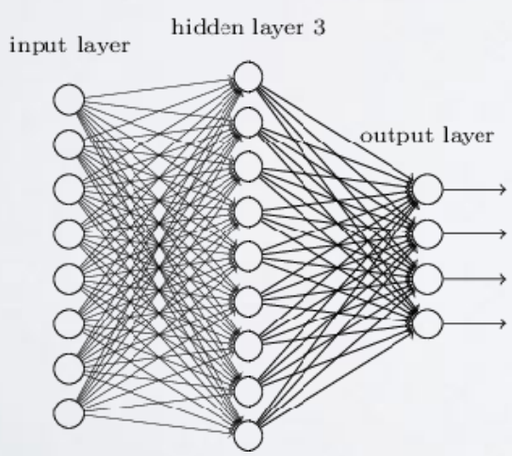
Big
Strong
Powerful



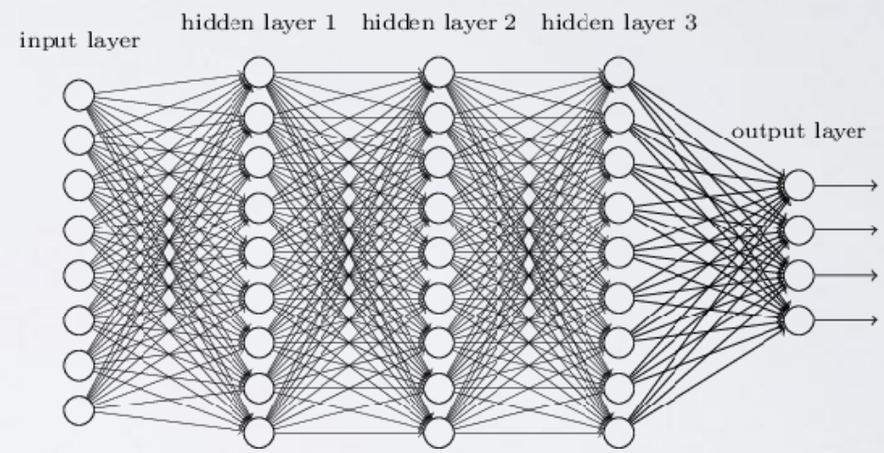
Possible Approaches

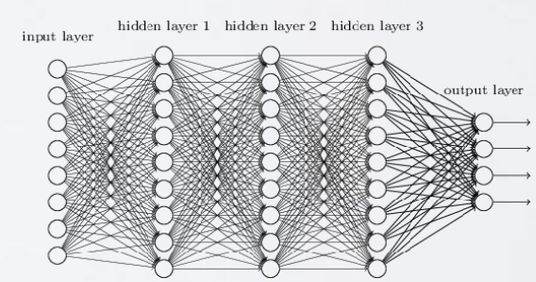
Shallow Approximation





Compression



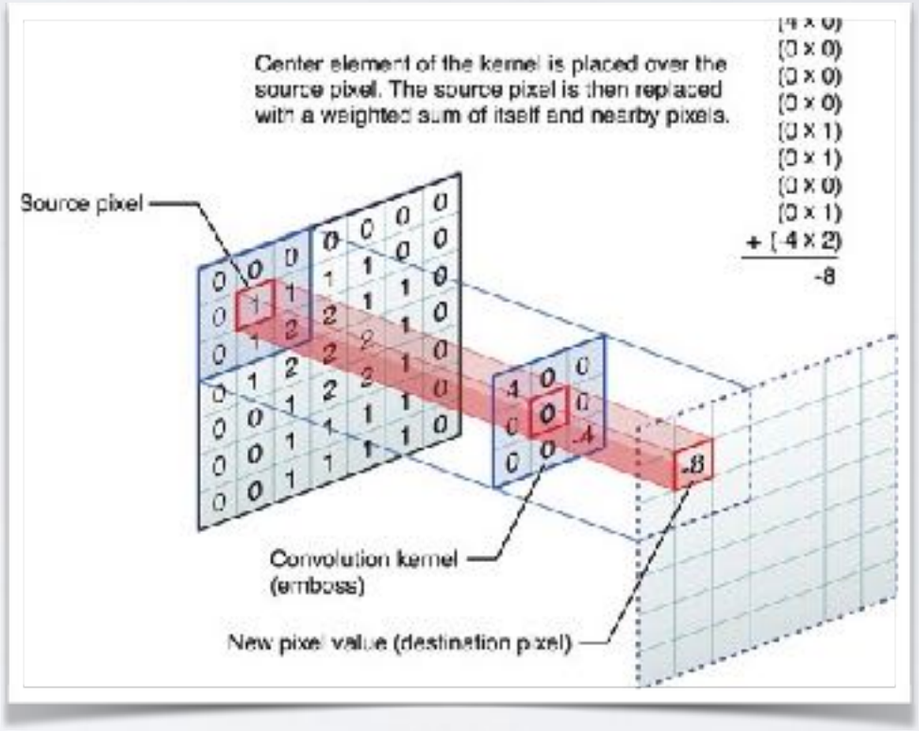


Binary-Weights Network

Basic Idea: Too much information in each convolutional layer. Can we store less?

Binary-Weights Network

I = Input Tensor
W = Weights
a = scaling factor



$$\mathbf{I} * \mathbf{W} \approx (\mathbf{I} \oplus \mathbf{W})\, \mathbf{a}$$

Training

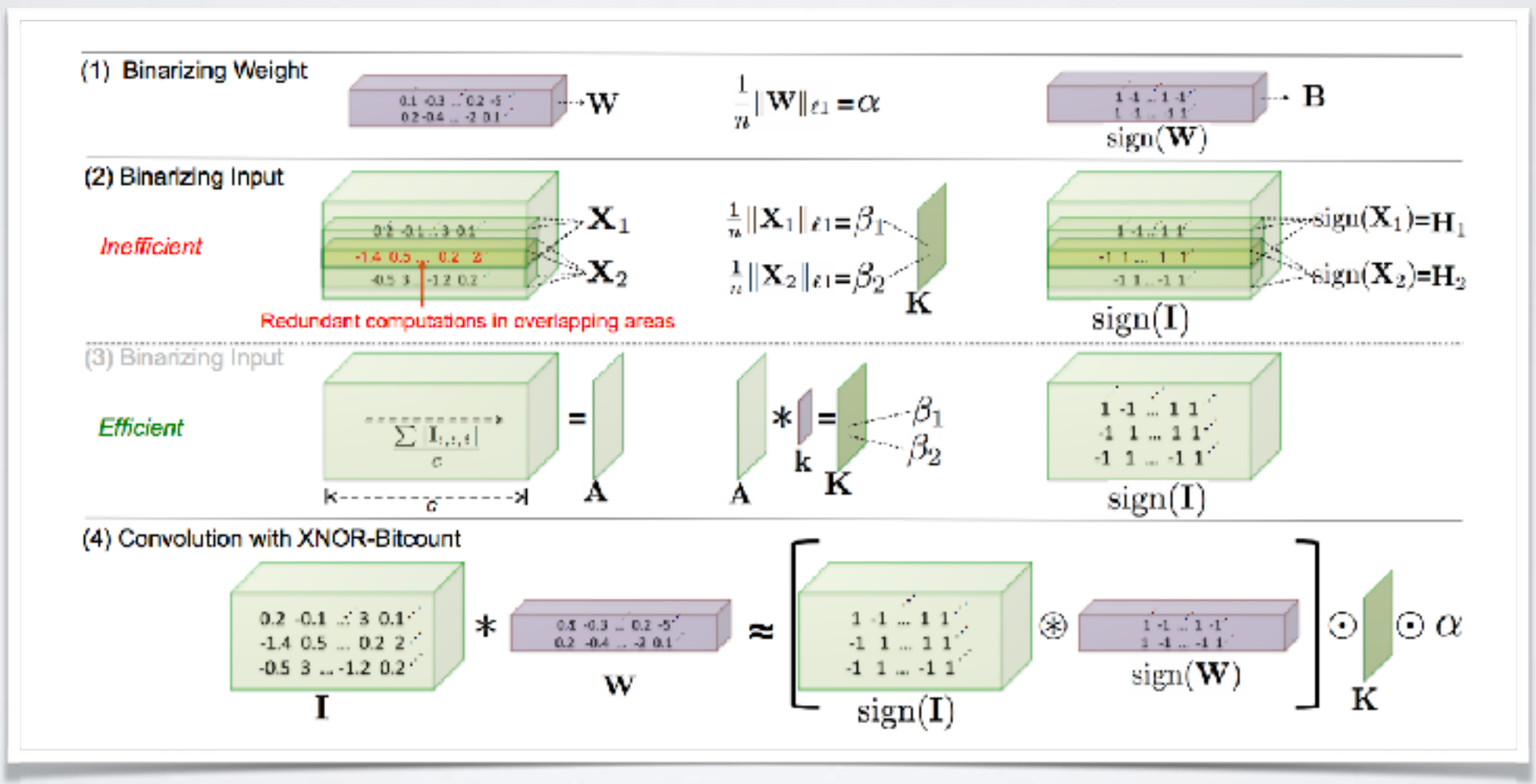
Binarize weights in forward pass and backward propagation

Use real valued weights in gradient descent (Why?)

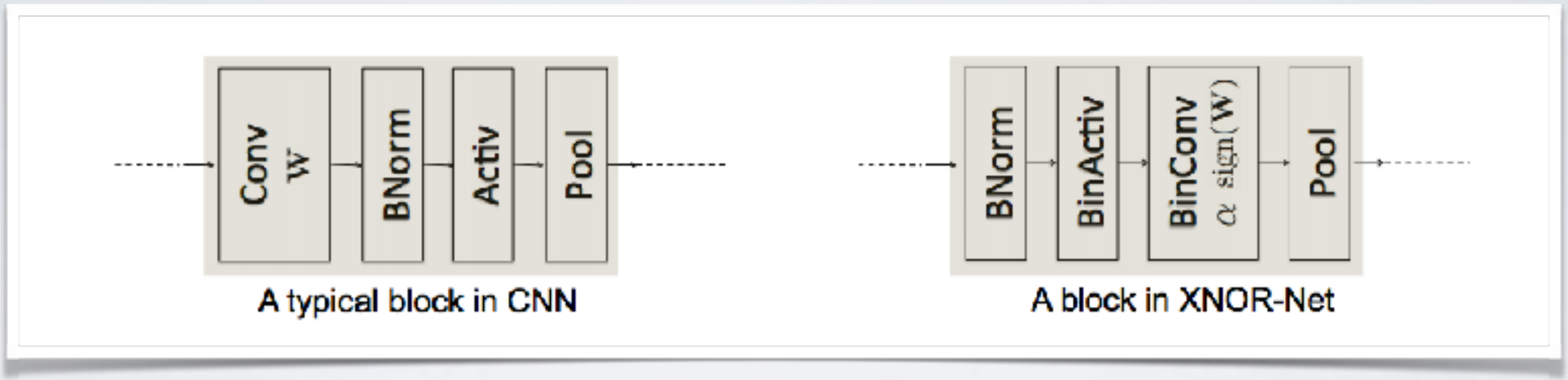
Also, if we are using real valued weights somewhere, what's the point?!

XNOR-Net

$$\mathbf{I} * \mathbf{W} \approx (\text{sign}(\mathbf{I}) \circledast \text{sign}(\mathbf{W})) \odot \mathbf{K}\alpha$$



Training

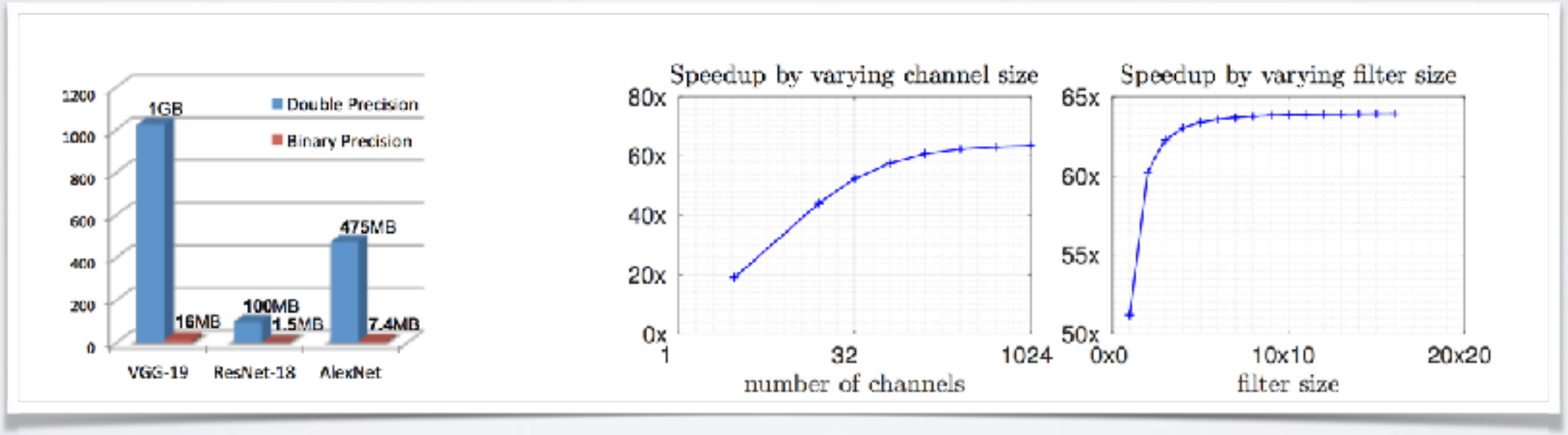


BinActiv
Computes the **K** and sign(**I**)

BinConv
Perform earlier Binary Convolution

Experiments

Efficiency



58x CPU Speedups

Experiments

Accuracy

Cifar-10

Binary-Weight Network: **9.88% Error**

XNOR-Net: **10.17% Error**

Experiments

Accuracy

Classification Accuracy(%)									
Binary-Weight				Binary-Input-Binary-Weight				Full-Precision	
BWN		BC[11]		XNOR-Net		BNN[11]		AlexNet[1]	
Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5
56.8	**79.4**	35.4	61.0	**44.2**	**69.2**	27.9	50.42	56.6	80.2

	ResNet-18		GoogLenet	
Network Variations	top-1	top-5	top-1	top-5
Binary-Weight-Network	60.8	83.0	65.5	86.1
XNOR-Network	51.2	73.2	N/A	N/A
Full-Precision-Network	69.3	89.2	71.3	90.0

Questions?