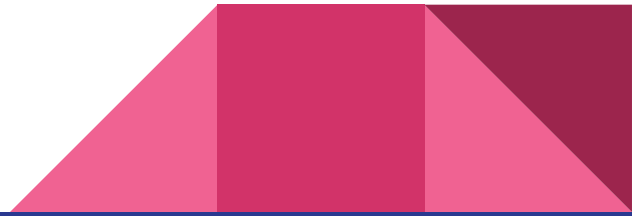
A review of Binary Neural Network by Courbariaux et. al.

Ramesh Sah

What is a Binary Neural Network (BNN)?

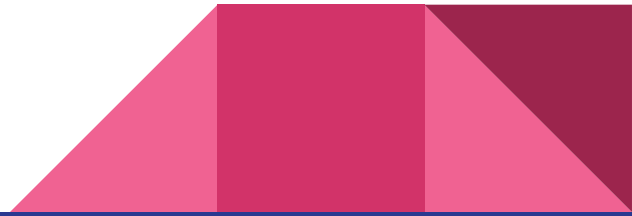
- A BNN is a neural network with binary weights and activations.
- At training-time binary weights and activations are used to compute the parameters gradient.



Why BNNs?

One of the best way to decrease the resource requirements of a machine learning model is to quantize the model - decrease the numerical resolution of the model.

BNN is the extreme form of model quantization and achieves drastic reduction in memory (32 times, compared to 32-bit float representation) and inference speed (many computation can be done using efficient bit-wise operations).



Training BNNs

$$x^b = \text{Sign}(x) = \begin{cases} +1 & \text{if } x \geq 0, \\ -1 & \text{otherwise,} \end{cases}$$

- Weights and activations can be either +1 or -1.
- Binarize a real-valued number x using the Sign function
- During training real-valued gradients of weights are accumulated for the SGD algorithm to work.
- Since the derivative of the sign function is zero, a Straight-Through Estimator is used to calculate the gradient.

$$\frac{\partial q(x)}{\partial x} = \begin{cases} 1 & |x| \leq 1 \\ 0 & |x| > 1 \end{cases}$$

Table 1. Classification test error rates of DNNs trained on MNIST (MLP architecture without unsupervised pretraining), CIFAR-10 (without data augmentation) and SVHN.

Data set	MNIST	SVHN	CIFAR-10
Binarized activations+weights, during training and test			
BNN (Torch7)	1.40%	2.53%	10.15%
BNN (Theano)	0.96%	2.80%	11.40%
Committee Machines' Array (Baldassi et al., 2015)	1.35%	-	-
Binarized weights, during training and test			
BinaryConnect (Courbariaux et al., 2015)	1.29± 0.08%	2.30%	9.90%
Binarized activations+weights, during test			
EBP (Cheng et al., 2015)	2.2± 0.1%	-	-
Bitwise DNNs (Kim & Smaragdis, 2016)	1.33%	-	-
Ternary weights, binary activations, during test			
(Hwang & Sung, 2014)	1.45%	-	-
No binarization (standard results)			
Maxout Networks (Goodfellow et al.)	0.94%	2.47%	11.68%
Network in Network (Lin et al.)	-	2.35%	10.41%
Gated pooling (Lee et al., 2015)	-	1.69%	7.62%

Figure 1. Training curves of a ConvNet on CIFAR-10 depending on the method. The dotted lines represent the training costs (square hinge losses) and the continuous lines the corresponding validation error rates. Although BNNs are slower to train, they are nearly as accurate as 32-bit float DNNs.

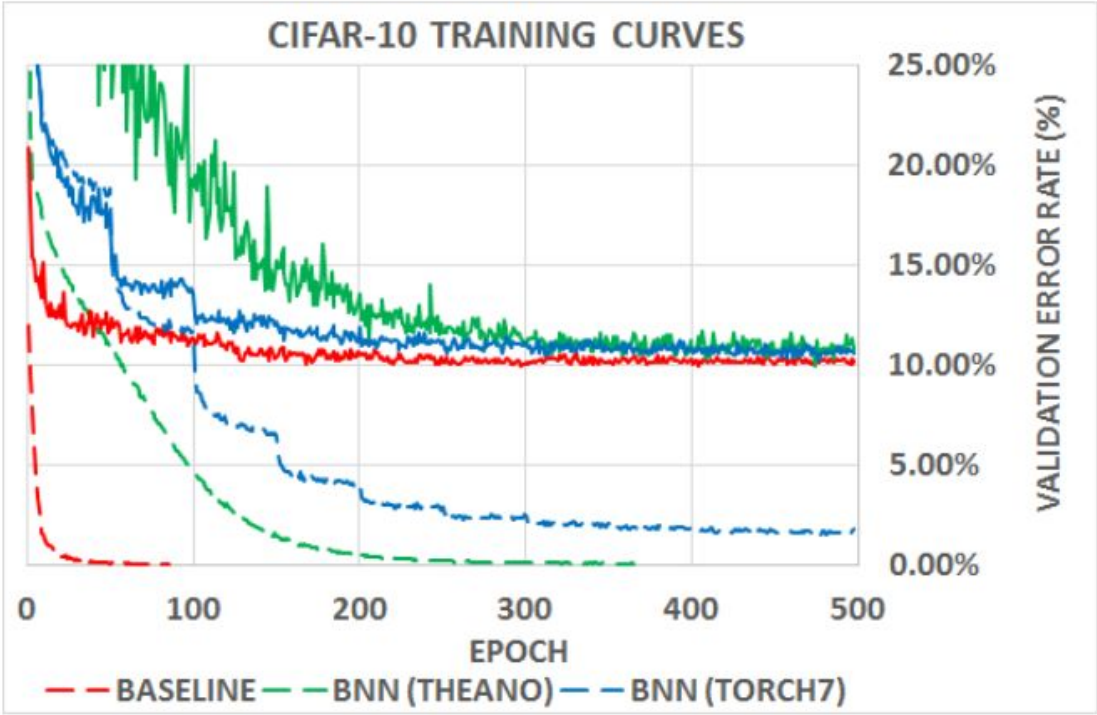


Table 2. Energy consumption of multiply-accumulations (Horowitz, 2014)

Operation	MUL	ADD
8bit Integer	0.2pJ	0.03pJ
32bit Integer	3.1pJ	0.1pJ
16bit Floating Point	1.1pJ	0.4pJ
32tbit Floating Point	3.7pJ	0.9pJ

Table 3. Energy consumption of memory accesses (Horowitz, 2014)

Memory size	64-bit memory access
8K	10pJ
32K	20pJ
1M	100pJ
DRAM	1.3-2.6nJ

Summary

- BNNs are neural networks with binary weights and activations.
- Use Sign function to binarize a real-value and approximation of Sign function to calculate gradients.
- Drastic improvement in memory and computation time.