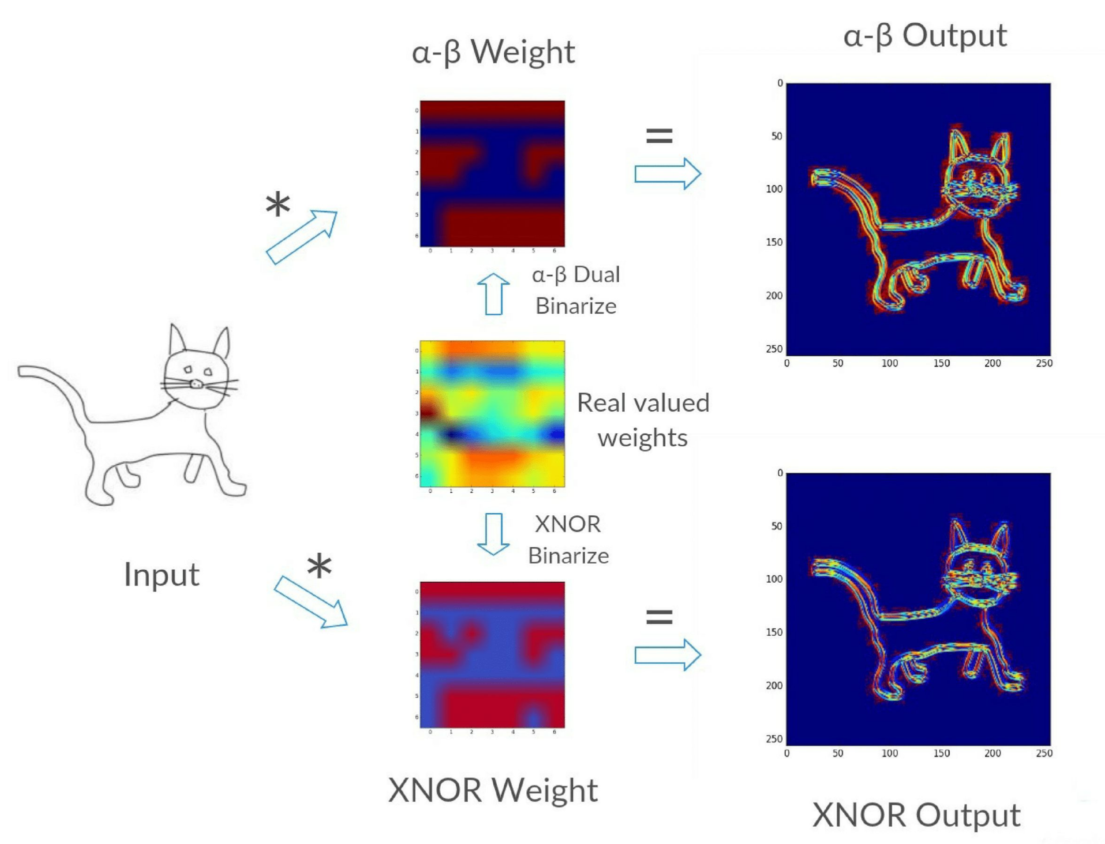
Distribution-Aware Binary Networks

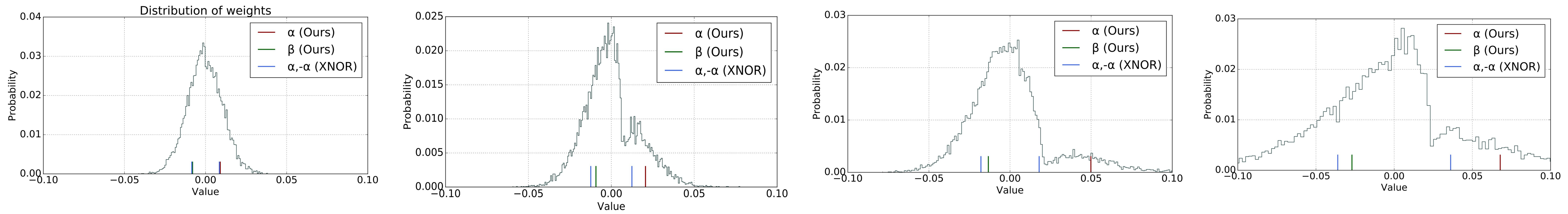
Ameya Prabhu, Vishal Batchu, Sri Aurobindo Munagala, Rohit Gajawada, Anoop Namboodiri
Center for Visual Information Technology, KCIS, IIIT Hyderabad, India

Introduction

- Network Compression aims to bring down the size of the neural network, and attempts to speed up computations - to achieve the goal of making DNNs easy to run on mobile devices.
- Network Quantization involves quantizing layer weights/activations from floats to a discrete set of values. Binarization is the extreme form of it - allowing for 58x computational speedups through XNOR-Popcount operations, 16x compression, and reduced power consumption.
- Classical network binarization techniques such as XNOR-Nets or BWNs (Binary-Weight Networks) cause significant accuracy drops.
- We explore a generalized binarization technique - binarize layer weights to alpha and beta rather than {-1, 1}. We outline the optimal values for alpha and beta in this representation, provide a fast Dynamic Programming algorithm to compute the optimal binarized vectors in a DNN, and derive an Alpha-Beta Binary layer, with improved accuracy.



An example convolution using our approach vs the naive binarization approach (XNOR)



Variation of α and β across a filter's weights during training

Theory

- The theorem given below shows that binary networks can approximate any given polynomial. $p(x)$ is a multivariate polynomial where n is the input dimension, and k being number of layers. We define $B_K(p, \sigma)$ as the minimum number of binary neurons required to approximate p.

Theorem 1 *For $p(\mathbf{x})$ equal to the product $x_1 x_2 \cdots x_n$, and for any σ with all nonzero Taylor coefficients, we have one construction of a binary neural network which meets the condition*

$$B_k(p,\sigma) = \mathcal{O}\left(n^{(k-1)/k} \cdot 2^{n^{1/k}}\right). \qquad (1)$$

and Rolnick et Al.'s conjecture is stated:

Conjecture *Let $p(\mathbf{x})$ equal to the product $x_1 x_2 \cdots x_n$, and suppose that σ has all nonzero Taylor coefficients. Then, we have:*

$$m_k^{uniform}(p) = 2^{\Theta(n^{1/k})}$$

If this conjecture is true, this would imply that weight-binarized networks have the same representational power as full-precision networks, since the network that was essentially used to prove the above theorem was a binary network.

- Assume a weight vector $\mathbf{W}$, which we attempt to binarize to a vector of form $[\alpha\alpha...\beta\alpha\beta]$. $\mathbf{e}$ is a vector such that $\mathbf{e} \in \{0,1\}^n \ni \mathbf{e} \neq 0$ and $\mathbf{e} \neq 1$. We define K as $\mathbf{e}^T\mathbf{e}$, or the number of 1s in $\mathbf{e}$. The optimization problem is modelled as:

$$\widetilde{\mathbf{W}}^* = \underset{\widetilde{W}}{argmin} \mid\mid \mathbf{W} - \widetilde{\mathbf{W}} \mid\mid^2$$

- Upon solving the above, we get optimal weight vector in terms of α, β as:

$$\widetilde{\mathbf{W}}^* = \alpha\mathbf{e} + \beta(\mathbf{1} - \mathbf{e}) \ where$$

$$\alpha = \frac{\mathbf{W}^T\mathbf{e}}{K}, \ \beta = \frac{\mathbf{W}^T(\mathbf{1}-\mathbf{e})}{n-K}$$

- Where the following expression is calculated for every possible value of K, for the top K weights or the bottom K weights:

$$\mathbf{e}^* = \underset{e}{argmax}\left(\frac{\mid\mid \mathbf{W}^T\mathbf{e} \mid\mid^2}{K} + \frac{\mid\mid \mathbf{W}^T(\mathbf{1}-\mathbf{e}) \mid\mid^2}{n-K}\right)$$

Results

- We compare accuracies of DAB-Net for WBin and FBin models on the TU-Berlin and Sketchy datasets.
- Note that DAB-Net FBin models consistently perform better than their XNOR-Net counterparts.
- DAB-Net WBin models achieve similar accuracies to BWN counterparts, because WBin models already achieve accuracies close to FPrec.
- The proposed binary representation takes into account the distribution of weights, unlike previous binarization approaches - and we showed how this representation can be computed efficiently in O(n.logn) time using Dynamic Programming, enabling efficient training on larger datasets and outperforming classical binarization techniques.

Models	Method	Accuracies	
		TU-Berlin	Sketchy
Sketch-A-Net	FPrec	72.9%	85.9%
	WBin (BWN)	73.0%	85.6%
	FBin (XNOR-Net)	59.6%	68.6%
	WBin DAB-Net	72.4%	84%
	FBin DAB-Net	60.4%	70.6%
Improvement	XNOR-Net vs DAB-Net	+0.8%	+2.0%
ResNet-18	FPrec	74.1%	88.7%
	WBin (BWN)	73.4%	89.3%
	FBin (XNOR-Net)	68.8%	82.8%
	WBin DAB-Net	73.5%	88.8%
	FBin DAB-Net	71.3%	84.2%
Improvement	XNOR-Net vs DAB-Net	+2.5%	+1.4%
GoogleNet	FPrec	75.0%	90.0%
	WBin (BWN)	74.8%	89.8%
	FBin (XNOR-Net)	72.2%	86.8%
	WBin DAB-Net	75.7%	90.1%
	FBin DAB-Net	73.7%	87.4%
Improvement	XNOR-Net vs DAB-Net	+1.5%	+0.6%

Comparison of DAB-Net accuracies across datasets

Link: http://bit.do/ameyap

Email: ameya.prabhu@research.iiit.ac.in