Ternary weight networks

- Fengfu Li, Bo Zhang, Bin Liu
- Caffe
- 2016

Overview

- introduce ternary weight networks (TWNs) - neural networks with weights constrained to +1, **0** and -1
- achieve up to 16x or 32x model compression rate
- gains 38x more stronger expressive abilities than the binary counterpart
- slightly worse than the full precision counterparts but outperforms the analogous binary precision counterparts a lot.
- need fewer multiplications compared with the full precision counterparts (like binary network)

Ternary weight networks

- Compared with the BPWNs, TWNs own an extra 0 state. But 0 terms need not be accumulated for any multiple operations. Thus, the multiply-accumulate operations in TWNs keep unchanged compared with binary precision counterparts.

$$\begin{cases} \mathbf{Z} & = & \mathbf{X} * \mathbf{W} \approx \mathbf{X} * (\alpha \mathbf{W}^t) = (\alpha \mathbf{X}) \oplus \mathbf{W}^t \\ \mathbf{X}^{\text{next}} & = & g(\mathbf{Z}) \end{cases} \tag{2}$$

$$\mathrm{W}_i^t = f_t(\mathrm{W}_i|\Delta) = \begin{cases} +1, & \text{if } \mathrm{W}_i > \Delta \\ 0, & \text{if } |\mathrm{W}_i| \le \Delta \\ -1, & \text{if } \mathrm{W}_i < -\Delta \end{cases}$$

$$\alpha^*, \Delta^* = \underset{\alpha \ge 0, \Delta > 0}{\arg\min} \left(|\mathbf{I}_\Delta| \alpha^2 - 2 \left(\sum |\mathrm{W}_i| \right) \alpha + c_\Delta \right) \tag{4}$$

where $\mathbf{I}_\Delta = \{i \mid |\mathrm{W}_i| > \Delta\}$ and $|\mathbf{I}_\Delta|$ denotes the number of elements in $\mathbf{I}_\Delta$; $c_\Delta = \sum_{i \in \mathbf{I}_\Delta^c} \mathrm{W}_i^2$ is a α-independent constant. Thus, for any given Δ, the optimal α can be computed as follows,

$$\alpha_\Delta^* = \frac{1}{|\mathbf{I}_\Delta|} \sum_{i \in \mathbf{I}_\Delta} |\mathrm{W}_i|. \tag{5}$$

By substituting α_Δ^* into (4), we get a Δ-dependent equation, which can be simplified as follows,

$$\Delta^* = \underset{\Delta > 0}{\arg\max} \frac{1}{|\mathbf{I}_\Delta|} \left(\sum_{i \in \mathbf{I}_\Delta} |\mathrm{W}_i| \right)^2 \tag{6}$$

由于 Problem (6) has no straightforward solutions. 作如下操作

equals to $0.75 \cdot \mathrm{E}(|\mathbf{W}|)$. Thus, we can use a rule of thumb that $\Delta^* \approx 0.7 \cdot \mathrm{E}(|\mathbf{W}|) \approx \frac{0.7}{n} \sum_{i=1}^n |\mathrm{W}_i|$ for fast and easy computation.

Experiments

Table 1: Network architecture and parameters setting for different datasets.

	MNIST	CIFAR-10	ImageNet
network architecture	LeNet-5	VGG-7	ResNet-18(B)
weight decay	1e-4	1e-4	1e-4
mini-batch size of BN	50	100	64 ($\times$4)[2]
initial learning rate	0.01	0.1	0.1
learning rate decay[3] epochs	15, 25	80, 120	30, 40, 50
momentum	0.9	0.9	0.9

Table 2: Validation accuracies (%). Results on ImageNet are with ResNet-18 / ResNet-18B.

	MNIST	CIFAR-10	ImageNet (top-1)	ImageNet (top-5)
TWNs	**99.35**	**92.56**	**61.8 / 65.3**	**84.2 / 86.2**
BPWNs	99.05	90.18	57.5 / 61.6	81.2 / 83.9
FPWNs	99.41	92.88	65.4 / 67.6	86.76 / 88.0
BinaryConnect	98.82	91.73	-	-
Binarized Neural Networks	88.6	89.85	-	-
Binary Weight Networks	-	-	60.8	83.0
XNOR-Net	-	-	51.2	73.2