

---

# Ternary Weight Networks

---

**Fengfu Li**

Institute of Applied Mathematics,  
AMSS, Chinese Academy of Sciences,  
Beijing, China 100090  
lifengful2@mailsucas.ac.cn

**Bin Liu**

Moshanghua Tech Co., Ltd.  
Beijing, China  
liubin@dress-plus.com

## Abstract

We introduce Ternary Weight Networks (TWNs) - neural networks with weights constrained to +1, 0 and -1. The L2 distance between the full (float or double) precision weights and the ternary weights along with a scaling factor is minimized. With the optimization, the TWNs own high capacity of model expression that is good enough to approximate the Full Precision Weight Networks (FPWNs) counterpart. Besides, the TWNs achieve up to 16x or 32x model compression rate and own much fewer multiplications compared with the FPWNs. Compared with recently proposed Binary Precision Weight Networks (BPWNs), the TWNs own nearly 38x more power of expression in a  $3 \times 3$  size filter, which is commonly used in most of the state-of-the-art CNN models like residual networks or VGG. Besides, the TWNs eliminate the singularity at zero and converge faster and more stably at training time. Benchmarks on MNIST, CIFAR-10, and the large scale ImageNet dataset show that TWNs achieve state-of-the-art performance which is only slightly worse than the FPWNs counterpart but outperforms the analogous BPWNs.

## 1 Introduction

Deep neural networks (DNN) have made significant improvements in lots of computer vision tasks such as object recognition [11,12,14] and object detection [16]. This motivates the interests to deploy the state-of-the-art DNN models to real world smart products such as virtual reality (VR) by Oculus [5], augmented reality (AR) by Magic Leap [21] and smart wearable devices. However, DNN models often own large number of weights which take away considerable storage requirement and computational power.

For example, the ResNet-18 network [4] has 11.6M parameters (45MB storage requirement) and performs 1.8B high precision operations to classify an image with size  $224 \times 224$ . These numbers are even much more higher for larger CNNs such as AlexNet [14] (249MB) and VGG-19 [11] (500MB). These large DNN models quickly overburden the limited storage, battery power, and computer capabilities of small devices like smart phones or other embedded devices. Thus, to deploy deep neural networks on mobile systems remains a challenge.

To address the limited storage and limited computational resources issues [9,17], we propose a simple but accurate approximation to the Full Precision Weight Networks (FPWNs) counterpart by shifting the weights to be ternary-valued. We optimize the L2 distance between the full precision weights and the ternary-precision weights along with a scaling factor. A threshold-based trinization function is proposed to get an approximate optimal solution that works well and can be fast computed. After the trinization, the resulted TWNs own high capacity of expression that is almost enough to approximate the FPWNs. Compared with FPWNs which use float (32 bit) or double (64 bit) weights, the TWNs only need 2 bit for each ternary weight unit and thus achieve up to 16x or 32x model compression rate. Besides, TWNs eliminate most of the multiplications in the for-

ward propagation. As a result, they own the potential of gaining significant benefits with specialized deep learning hardware by replacing many multiply-accumulate operations by simple accumulations [1,15].

## 2 Binary Weight Networks and Model Compression

Recently, several methods seek to binarize the weights and activations in deep neural networks for model compression as well as reducing computations. BinaryConnect [1] is one of the first attempts that use full resolution weights as a reference for the binarization process. The full resolution weights are updated, however, using the back propagated error without binarization in the update process. An extension of BinaryConnect is proposed by BinaryNet [2], where both weights and activations are binary-valued. Though BinaryConnect and BinaryNet achieved state-of-the-art results on small datasets such as MNIST, CIFAR-10 and SVHN, their convergence speed is very slow. Besides, these methods are not very successful on large-scale datasets [3] such as ImageNet.

Binary Weight Networks [3] and XNOR-Networks [3] make some optimization of the approximated binary-valued weights or activations over BinaryConnect and BinaryNet, respectively. The XNOR-Net achieved up to 64x memory saving for the parameters and 58x faster convolutional operations while shown relatively more accurate results than BinaryNet on the ImageNet dataset among several DNN models. However, there still exists a huge accuracy gap between the XNOR-Net and the full precision counterpart when dealing with large CNN such as ResNet-18 or GoogLeNet.

Some other model compression methods focus on identifying a model with few parameters while preserving accuracy by taking an existing DNN model and compress it in a lossy way. SqueezeNet [20] is such a model that has 50x fewer parameters than AlexNet but maintains AlexNet-level accuracy on ImageNet. Deep Compression [17] is another most recently proposed method that uses pruning, trained quantization and Huffman coding for compressing neural networks. It reduced the storage requirement of AlexNet and VGG-16 by 35x and 49x, respectively, without loss of accuracy.

## 3 Ternary Weight Networks

In this section we seek to make a balance between the high rate of model compression and high model capacity with ternary-valued weight networks. On one hand, FPWNs own high model capacity but need too much storage as well as too much computational resources for float multiplications. On the other hand, the BPWNs go to another extremity that have very high model compression rate but sacrifice model capacities too much. A balance between the two extremity would be the TWNs that own both the merit of BPWNs and FPWNs but overcome the defects of them.

**Model compression:** BPWNs can achieve up to 32x or 64x model compression rate compared with the FPWNs counterpart. TWNs, though need 2x more storage compared with the BPWNs, can also achieve up to 16x or 32x model compression rate which is fair enough for compressing most of the current state-of-the-art DNN models.

**Model capacity:** In most recent network architectures such as VGG, GoogLeNet and residual networks, a most commonly used convolutional filter is of size  $3 \times 3$ . With binary weights, there is only  $2^{3 \times 3} = 512$  templates which means many filters would duplicate when the number of filters is larger than 512. However, a ternary filter with the same size owns  $3^{3 \times 3} = 19683$  templates, which is 38x more rich expressive than a binary counterpart.

**Computational requirement:** By adding a 0 in the ternary state, the computations keep unchanged (even less) compared with binary weights because the 0 terms would not be accumulated. Thus, it is also hardware-friendly for training large-scale networks with special DL hardware.

### 3.1 Problem formulation

To make the ternary weight networks perform well, we seek to optimize the L2 norm between the full precision weights and the ternary-valued counterpart. More specifically, weights of the TWNs own three states: -1, 0 and +1. A scaling factor  $\alpha$  is used to rescale the ternary weights  $\mathbf{W}^t$  to optimize the L2 distance between the full precision weights  $\mathbf{W}$  and the rescaled ternary-valued

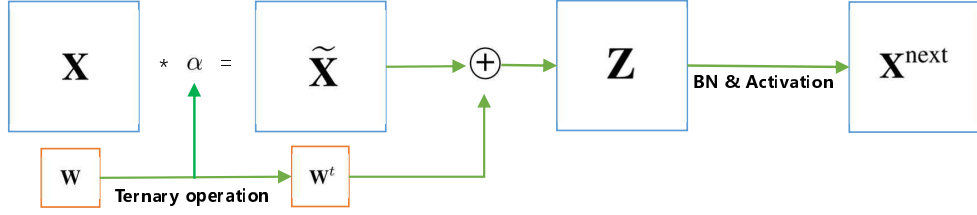


Figure 1: A basic block of forward propagation in ternary weight networks.

weights  $\alpha \mathbf{W}^t$ . The optimization problem is formulated as follows,

$$\begin{cases} \alpha^*, \mathbf{W}^{t*} = \underset{\alpha, \mathbf{W}^t}{\operatorname{argmin}} J(\alpha, \mathbf{W}^t) = \|\mathbf{W} - \alpha \mathbf{W}^t\|_2^2 \\ s.t. \quad \alpha \geq 0, \mathbf{W}_i^t \in \{-1, 0, 1\}, i = 1, 2, \dots, n. \end{cases} \quad (1)$$

Here  $n$  is the size of the filter.

With the approximation  $\mathbf{W} \approx \alpha \mathbf{W}^t$ , a basic block of forward propagation in a ternary weight network is as follows,

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

Here  $*$  is matrix-vector multiplication (inner product) or convolution operation and  $g$  is some kind of nonlinear activation function.  $\oplus$  indicates an inner product or a convolution without any multiplication. Fig. 1 shows the basic block of TWNs.

### 3.2 Optimization with threshold-based trinization function

To solve the optimization problem (1), we first expand the cost function as follows,

$$\begin{aligned} J(\alpha, \mathbf{W}^t) &= \alpha^2 (\mathbf{W}^t)^T \mathbf{W}^t - 2\alpha (\mathbf{W}^t)^T \mathbf{W} + \mathbf{W}^T \mathbf{W} \\ &= \sum_{i=1}^n (\alpha^2 (\mathbf{W}_i^t)^2 - 2\alpha \mathbf{W}_i^t \mathbf{W}_i) + c \end{aligned} \quad (3)$$

where  $T$  is the transpose operation,  $c = \mathbf{W}^T \mathbf{W}$  is a constant;  $\mathbf{W}_i^t$  and  $\mathbf{W}_i$  are the  $i$ th elements of  $\mathbf{W}^t$  and  $\mathbf{W}$ , respectively. When there is no constraints, we can solve problem (1) by setting the derivatives of  $J(\alpha, \mathbf{W}^t)$  with respect to  $\alpha$  and  $\mathbf{W}_i^t$ s to 0. In this case, we will get

$$\alpha^* = \frac{\sum_{i=1}^n \mathbf{W}_i^{t*} \mathbf{W}_i}{\sum_{i=1}^n \mathbf{W}_i^{t*2}} \quad (4)$$

and

$$\mathbf{W}_i^{t*} = \frac{\mathbf{W}_i}{\alpha^*} \quad (5)$$

Due to  $\alpha^*$  and  $\mathbf{W}_i^{t*}$  are interdependent, there is no deterministic solution [18]. However, with the ternary-valued constraint, we try to seek an approximated optimal solution with the following form,

$$\mathbf{W}_i^t = f_t(\mathbf{W}_i | \Delta) = \begin{cases} +1, & \text{if } \mathbf{W}_i > \Delta \\ 0, & \text{if } |\mathbf{W}_i| \leq \Delta \\ -1, & \text{if } \mathbf{W}_i < -\Delta \end{cases} \quad (6)$$

Here  $\Delta$  is an adjustable threshold parameter whose value is greater than 0. With (6), the optimization problem is transformed into

$$\begin{aligned}
\alpha^*, \Delta^* &= \operatorname{argmin}_{\alpha \geq 0, \Delta > 0} \|\mathbf{W} - \alpha \mathbf{W}^t\|_2^2 \\
&= \operatorname{argmin}_{\alpha \geq 0, \Delta > 0} \left( \sum_{i: |W_i| > \Delta} |W_i - \alpha|^2 + \sum_{i: |W_i| \leq \Delta} W_i^2 \right) \\
&= \operatorname{argmin}_{\alpha \geq 0, \Delta > 0} \left( |I_\Delta| \alpha^2 - 2 \left( \sum_{i \in I_\Delta} |W_i| \right) \alpha + c \right)
\end{aligned} \tag{7}$$

Where  $I_\Delta = \{1 \leq i \leq n \mid |W_i| > \Delta\}$ , and  $|I_\Delta|$  denotes the number of elements in  $I_\Delta$ . The right side of (7) is a quadratic function of  $\alpha$ . Thus, for any given  $\Delta$ , the optimal  $\alpha$  can be computed as follows,

$$\alpha_\Delta^* = \frac{1}{|I_\Delta|} \sum_{i \in I_\Delta} |W_i|. \tag{8}$$

Given the computed optimal  $\alpha$ , the optimal  $\Delta$  can be computed by solving

$$\begin{aligned}
\Delta^* &= \operatorname{argmin}_{\Delta > 0} \|\mathbf{W} - \alpha_\Delta^* \mathbf{W}^t\|_2^2 \\
&= \operatorname{argmin}_{\Delta > 0} -\frac{1}{|I_\Delta|} \left( \sum_{i \in I_\Delta} |W_i| \right)^2 \\
&= \operatorname{argmax}_{\Delta > 0} \frac{1}{|I_\Delta|} \left( \sum_{i \in I_\Delta} |W_i| \right)^2
\end{aligned} \tag{9}$$

Problem (9) has no straightforward solutions. Though discrete optimization can be made to solve the problem (due to states of  $W_i$ s are finite), it is time consuming. Instead, we would like to find some approximate optimal solutions that can be easily and fast computed.

### 3.2.1 Case analysis: uniform distribution

Suppose  $\mathbf{W}$  is uniformly distributed in  $[-a, a]$ , and  $\Delta$  lies in  $(0, a]$ . Then,  $|I_\Delta| \approx n(1 - \frac{\Delta}{a})$  and  $\sum_{i \in I_\Delta} |W_i| \approx \frac{a+\Delta}{2} \cdot |I_\Delta| \approx \frac{n}{2a}(a+\Delta)(a-\Delta)$ . Thus, an approximation of the optimal solution is

$$\begin{aligned}
\Delta^* &\approx \operatorname{argmax}_{0 < \Delta \leq a} \frac{\left( \frac{n}{2a}(a+\Delta)(a-\Delta) \right)^2}{n(1 - \frac{\Delta}{a})} \\
&= \operatorname{argmax}_{0 < \Delta \leq a} \frac{n}{4a} \cdot (-\Delta^3 - a\Delta^2 + a^2\Delta + a^3) \\
&= \frac{1}{3}a.
\end{aligned} \tag{10}$$

### 3.2.2 Case analysis: normal distribution

Suppose  $\mathbf{W}$  is generated from some normal distribution  $N(0, \sigma^2)$  ( $\sigma > 0$  is unknown). Then,  $|I_\Delta| \approx n \int_{|x| > \Delta} f_\sigma(x) dx = 2n(1 - F(\Delta/\sigma))$ , where  $f_\sigma(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp(-\frac{x^2}{2\sigma^2})$  is the probability density function of  $N(0, \sigma^2)$  and  $F(x)$  is the standard cumulative density function of  $N(0, 1)$ . Further more,  $\sum_{i \in I_\Delta} |W_i| \approx n \int_{|x| > \Delta} |x| f_\sigma(x) dx = \frac{2n\sigma}{\sqrt{2\pi}} \exp(-\frac{\Delta^2}{2\sigma^2})$ . Thus, an approximation of the optimal solution with the normally distributed  $\mathbf{W}$  is

$$\begin{aligned}
\Delta^* &\approx \operatorname{argmax}_{\Delta > 0} \frac{\left( \frac{2n\sigma}{\sqrt{2\pi}} \exp(-\frac{\Delta^2}{2\sigma^2}) \right)^2}{2n(1 - F(\Delta/\sigma))} \\
&= \operatorname{argmax}_{\Delta > 0} \frac{n\sigma^2}{\pi} \cdot \frac{\exp(-(\Delta/\sigma)^2)}{1 - F(\Delta/\sigma)} \\
&= 0.6\sigma.
\end{aligned} \tag{11}$$

### 3.2.3 A rule of thumb

In real applications, it is hard to estimate the distribution of  $\mathbf{W}$ . Instead, notice that the mean of  $|\mathbf{W}_i|$ s can be easily computed and it is a good estimator for  $E(|\mathbf{W}|)$ . Besides,  $\Delta^*/E(|\mathbf{W}|) \approx 0.7$  both for the uniform distribution ( $E(|\mathbf{W}|) = a/2$ ) and the normal distribution ( $E(|\mathbf{W}|) = 2\sigma/\sqrt{2\pi}$ ). Thus, as a rule of thumb, we would like to use

$$\Delta^* = 0.7 \cdot E(|\mathbf{W}|) \approx \frac{0.7}{n} \sum_{i=1}^n |\mathbf{W}_i| \quad (12)$$

as an approximation of the optimal  $\Delta$ .

### 3.3 Training ternary weight networks

To train ternary weight networks, ternary-valued weights are used during the forward and backward propagations but not during the parameters update. This is a commonly used strategy as in [1,3,15]. Besides, Batch Normalization [10] and learning rate scaling are two tricks that are used in all our experiments. In the forward propagation, the scaling factor  $\alpha$  could be transformed to the inputs according to (2). Algorithm 1 shows the overall training procedure for TWNs with the stochastic gradient descent (SGD) method.

---

#### Algorithm 1 SGD Training for the Ternary Weight Networks

---

**Input:** A minibatch of inputs ( $\mathbf{X}$ ) and targets ( $\mathbf{Y}$ ), current weights  $\mathbf{W}$  and learning rate  $\eta$ .

**Output:** Updated weights  $\mathbf{W}^{new}$  and updated learning rate  $\eta^{new}$ .

---

- 1: **Forward propagation:**
  - 2: **for**  $l = 1$  to  $L$  **do**
  - 3:   Get the threshold value  $\Delta_l^*$  by (16) with the filters in the  $l$ th layer.
  - 4:   Compute the ternary weights  $\mathbf{W}_l^t$  by (6) for the  $l$ th layer.
  - 5:   Get the scaling factor  $\alpha_l^*$  by (8).
  - 6:   Rescale the inputs of  $i$ th layer  $\mathbf{X}^{l-1}$  with scaling factor  $\alpha_l^*$ , thus,  $\tilde{\mathbf{X}}^{l-1} = \alpha^* \mathbf{X}^{l-1}$ .
  - 7:   Compute  $\mathbf{Z}^l$  with input  $\tilde{\mathbf{X}}^{l-1}$  and the ternary-valued weights  $\mathbf{W}_l^t$ .
  - 8:   Perform Batch Normalization and non-linear activation to get  $\mathbf{X}^l$ .
  - 9: **end for**
  - 10: Compute the cost  $C$  with forward outputs  $\mathbf{X}^L$  and the targets  $\mathbf{Y}$ .
  - 11: **Backward propagation:**
  - 12: Initialize output layer's activation gradient  $\frac{\partial C}{\partial \mathbf{X}^L}$ .
  - 13: **for**  $l = L$  to 2 **do**
  - 14:   Compute  $\frac{\partial C}{\partial \mathbf{X}^{l-1}}$  knowing  $\frac{\partial C}{\partial \mathbf{X}^l}$  and  $\mathbf{W}_l^t$ .
  - 15: **end for**
  - 16: **Update parameters and learning rate:**
  - 17: **for**  $l = 1$  to  $L$  **do**
  - 18:   Compute  $\frac{\partial C}{\partial \mathbf{W}_l^t}$  according to  $\frac{\partial C}{\partial \mathbf{X}^{l+1}}$  and  $\mathbf{X}^l$ .
  - 19:    $\mathbf{W}_l^{new} \leftarrow \mathbf{W}_l - \eta \frac{\partial C}{\partial \mathbf{W}_l^t}$
  - 20: **end for**
  - 21: Update learning rate  $\eta^{new}$  according to any learning rate scaling method.
- 

### 3.4 Model compression and run time usage

As shown in Algorithm 1, there is no need to store the full precision weights for forward propagation. Thus, we only need to save the ternary-valued weights and the scaling factors for run time usage. This would results up to 16x or 32x model compression rate compared with float or double precision network counterpart, respectively.

## 4 Experiments

In this section, we benchmark the ternary weight networks with the binary precision weight networks and the full precision weight networks. The three methods are with same network architectures, the same regularization method (L2 weight decay), the same learning rate scaling procedure (multistep) and the same optimization method (SGD with momentum). The only difference is the precision of the weights. For FPWNs, the weights are float-valued; TWNs and BPWNs are ternary-valued and binary-valued weights, respectively.

A basic block used in the experiments is an inner product or convolution operation with full, ternary or binary precision, followed by a Batch Normalization layer and an ReLU activation layer as shown in Fig. 1. No other regularization methods such as dropout are used.

Table. 1 summarizes the overall accuracies of these methods on MNIST, CIFAR-10 and the ImageNet2012 dataset. More details are described in the subsequent subsections.

	MNIST	CIFAR-10	ImageNet (top-1)	ImageNet (top-5)
Ternary Weight Network	<b>99.35%</b>	<b>92.56%</b>	<b>61.8% / 65.3%</b>	<b>84.2% / 86.2%</b>
Binary Precision Counterpart	99.05%	90.18%	57.5% / 61.6%	81.2% / 83.9%
Full Precision Counterpart	99.41%	92.88%	65.4% / 67.6%	86.76% / 88.0%
BinaryConnect [1]	98.82%	91.73%	-	-
Binarized Neural Networks [2]	88.6%	89.85%	-	-
Binary Weight Networks [3]	-	-	60.8%	83.0%
XNOR-Net [3]	-	-	51.2%	73.2%
Maxout Networks [6]	90.06%	90.62%	-	-
Network in Network [7]	99.55%	91.19%	-	-
Deeply Supervised Network [8]	99.61%	92.03%	66.3%	86.9%

Table 1: Validation accuracy comparison of the three kinds of networks with different precisions. For the MNIST dataset, the LeNet-5 architecture is used. For the CIFAR-10 dataset, VGG7-128 architecture is used and all methods are with data preprocessing or data augmentation. For ImageNet dataset, ResNet-18 architecture is used and its counterpart ResNet-18B is used for comparison. The TWNs, BPWNs and FPWNs are trained without scale augmentation or color augmentation while Binary Weight Networks and XNOR-Net are trained with resized images.

### 4.1 MNIST

MNIST is a basic benchmark image classification dataset. It consists of 60000 training images and 10000 testing images. All images are grey value images of size  $28 \times 28$  pixels, falling into 10 classes corresponding to the 10 digits. We split the full training set into a 50000 training set and a 10000 validation set. We use a similar LeNet-5 [13] architecture for benchmarking. The architecture is as follows,

$$32@5 \times 5 \text{ CONV} + 2 \times 2 \text{ MP} + 64@5 \times 5 \text{ CONV} + 2 \times 2 \text{ MP} + 512 \text{ FC} + \text{SVM}$$

Here “ $32@5 \times 5 \text{ CONV}$ ” means a convolutional block that owns 32 filters with size  $5 \times 5$ . “ $2 \times 2 \text{ MP}$ ” is a max-pooling layer with stride 2. “FC” is a fully connected block or inner product block. The top layer is a SVM classifier with 10 labels. Hinge loss is minimized with simple SGD and momentum 0.9. The only regularizer we use is L2-normalized weight decay whose value is set to be 0.0001. Batch normalization with a minibatch of size 50 is used to speed up the training. The learning rate starts at 0.01 and decays by 0.1 at epoch 15 and 25.

Fig. 2 shows the validation accuracy curves on this dataset. As shown in the figure, the full precision network converges fast and stays stable after epoch 16. The binary precision network converges slowly and vibrates a lot even when the learning rate is very small after epoch 25. The ternary weight network, though vibrates a lot with the initial learning rate, keeps stable when the learning rate decays. Besides, it achieves the state-of-the-art result compared with the full precision network within 20 epochs.

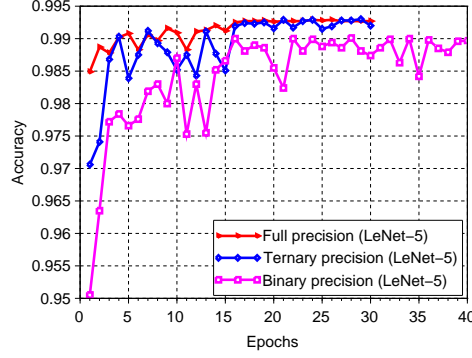


Figure 2: Validation accuracy curves on the MNIST dataset with different precisions.

## 4.2 CIFAR-10

CIFAR-10 is another image classification benchmark dataset. It contains images of size  $32 \times 32$  RGB pixels in a training set of 50000 and a test set of 10000. We use a VGG inspired CNN architecture [11] as follows,

$$2 \times (K-C3) + MP2 + 2 \times (2K-C3) + MP2 + 2 \times (4K-C3) + MP2 + 8K-FC + \text{Softmax}$$

We name this architecture as VGG7- $K$ , where  $K$  is the number of filters in the first convolution layer. Compared with the architecture adopted in [1], we delete the last full connection layer.

Softmax classifier is used in this architecture. We use Batch Normalization with a minibatch of size 100 to speed up training. The learning rate starts at 0.1 and divides by 10 at epoch 80, 120 and 150. The only regularizer we use is L2-normalized weight decay whose value is set to be 0.0001. SGD with momentum equals 0.9 is used as the solver. We follow the data augmentation in [4,8] for training: 4 pixels are padded on each side, and a  $32 \times 32$  crop is randomly sampled from the padded image or its horizontal flip. For testing, we only evaluate the single view of the original  $32 \times 32$  image.

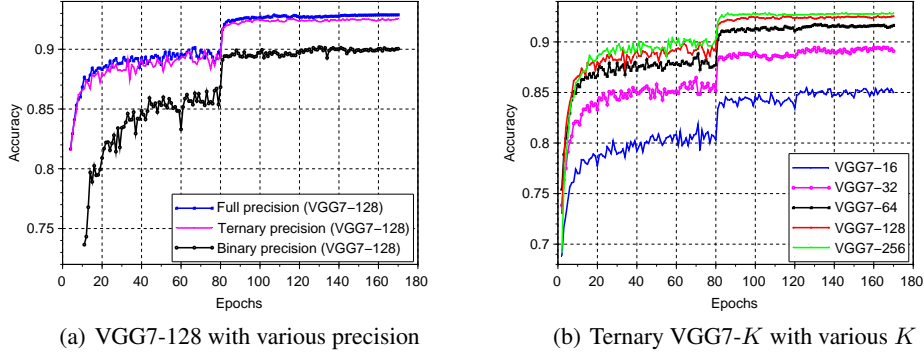


Figure 3: Validation accuracy curves on the CIFAR-10 dataset with the VGG7- $K$  architecture.

Fig. 2(a) shows the validation accuracy curves of VGG7-128 architecture under different precisions. The TWNs own almost the same performance as with the full precision counterpart. Both of the TWNs and FPWNs outperform the BPWNs a lot. Furthermore, the BPWNs is more volatile than the TWNs. Fig. 2(b) shows the validation accuracy curves with different model size  $K$ . With relatively larger size (64, 128 or 256), the capacity of VGG7- $K$  is enough to model the CIFAR-10 dataset. However, when  $K$  is too small (16 or 32), the model would be underfit for the dataset.

### 4.3 ImageNet2012

ImageNet has 1.2M train images from 1K categories and 50K validation images. The images in this dataset are natural images with reasonably high resolution in contrast to CIFAR-10 and MNIST dataset, which have relatively small images.

We use the recent proposed residual network architecture [4] to benchmark the performance of ternary weight network on this dataset. For simplicity, the smallest model (ResNet-18) among all of the residual networks is adopted. Besides, to address the issue of model size, we also benchmark another enlarged counterpart whose number of filters in each layer is 1.5x of the original one. This means the number of filters in the first layer becomes 96. We name this counterpart model as ResNet-18B.

In each training iteration, images are randomly cropped with  $224 \times 224$  size. We do not use any resize tricks [3] or any color augmentation, yet. (In latter versions, we would make a more comprehensive comparison over the augmentation issues.) The learning rate starts at 0.1 and is divided by 10 at epoch 30, 40 and 50. We adopt batch normalization with a minibatch of size 64 to speed up the training, and accumulate 4 iterations of gradients for parameters updating<sup>1</sup> (which means 256 images are processed for one updating of parameters). The weight decay is set to 0.0001 and SGD with momentum 0.9 is served as the solver.

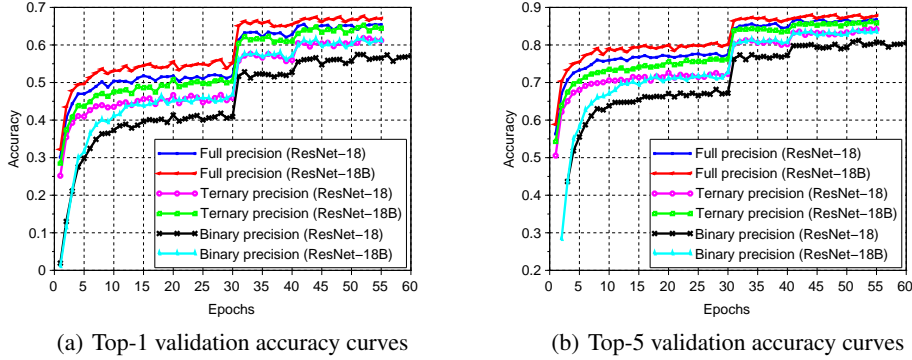


Figure 4: Validation accuracy curves on the ImageNet2012 dataset with ResNet-18 architecture.

Fig. 3 shows the top-1 and top-5 validation accuracy curves on the ImageNet dataset. We get the following observations: 1). The ternary weight networks outperform the binary weight networks a lot (about 4% gap, top-1) both with the ResNet-18 and ResNet-18B. 2). With ResNet-18, the full precision network outperforms the ternary weight network with 3.6% gap. However, the gap reduced to 2.3% with ResNet-18B. This means model size has larger impact to TWNs than to the FPWNs counterpart. 3). The differences among top-5 accuracies are not that obvious compared with that of top-1 accuracy.

We also use the method described in Sec. 3.4 to compress the float-valued ResNet-18 and ResNet-18B model. We pack 16 ternary-valued weight units into a single fixed32 ProtoBuffer unit [19]. All other terms such as biases, scaling factors, and parameters of the Batch Normalization layers are stored in float values. With this procedure, we compress the float-valued ResNet-18 (45MB) and ResNet18B (99MB) to ternary-valued models whose storage requirement are 2.9MB and 6.4MB, respectively. The compression rates are about 15.52x and 15.47x, respectively.

## 5 Conclusion

In this paper, we introduced a simple but accurate approximation to full precision deep neural networks using ternary-valued weight networks. The TWNs make a balance between the high rate of model compression and high model capacity. On MNIST and CIFAR-10, the TWNs achieve state-

<sup>1</sup>We use 4 GPUs to speed up the training in Caffe [19].



of-the-art result that is only slightly worse than FPWNs but outperforms the BPWNs a lot. On the large scale ImageNet dataset, the TWNs again outperforms BPWNs a lot. Some improvements that should be consider in the future work would be: 1). Low precision activations ( $\mathbf{X}$ ) could be considered along with the ternary weights. This would reduce the training time memory requirement and pave way for training much more large networks. 2). There is still considerable gap between the TWNs and the FPWNs counterpart when dealing with large scale datasets such as ImageNet. The gap would be reduced by enlarging the bits of ternary precision weights to half-char precision weights according to the previous benchmarks.

## References

- [1] Courbariaux, M., Bengio, Y., and David, J. P. (2015). BinaryConnect: Training Deep Neural Networks with binary weights during propagations. *Advances in Neural Information Processing Systems*. pp. 3105-3113.
- [2] Courbariaux, M., et al. (2016) Binarized neural networks: Training deep neural networks with weights and activations constrained to +1 and -1. *arXiv:1602.02830*.
- [3] Rastegari, M., Ordonez, V., Redmon, J., and Farhadi, A. (2016). XNOR-Net: ImageNet Classification Using Binary Convolutional Neural Networks. *arXiv:1603.05279*.
- [4] He, K., Zhang, X., Ren, S., and Sun, J. (2015). Deep Residual Learning for Image Recognition. *arXiv:1512.03385*.
- [5] Desai, P. R., Desai, P. N., Ajmera, K. D., and Mehta, K. (2014). A review paper on oculus rift-a virtual reality headset. *arXiv:1408.1173*.
- [6] Goodfellow, I., Warde-Farley, D., Mirza, M., Courville, A., and Bengio, Y. (2013). Maxout Networks. *Proceedings of The 30th International Conference on Machine Learning*. pp. 1319-1327.
- [7] Lin, M., Chen, Q., and Yan, S. (2013). Network in network. *arXiv:1312.4400*.
- [8] Lee, C. Y., Xie, S., Gallagher, P., Zhang, Z., and Tu, Z. (2014). Deeply-supervised nets. *arXiv:1409.5185*.
- [9] Esser, S. K., et al. (2016). Convolutional Networks for Fast, Energy-Efficient Neuromorphic Computing. *arXiv:1603.08270*.
- [10] Ioffe, S., and Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv:1502.03167*.
- [11] Simonyan, K., and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv:1409.1556*.
- [12] Szegedy, C., et al. (2015). Going deeper with convolutions. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. pp. 1-9.
- [13] LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11), pp. 2278-2324.
- [14] Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*. pp. 1097-1105.
- [15] Lin, Z., Courbariaux, M., Memisevic, R., and Bengio, Y. (2015). Neural Networks with Few Multiplications. *arXiv:1510.03009*.
- [16] Ren, S., He, K., Girshick, R., and Sun, J. (2015). Faster r-cnn: towards real-time object detection with region proposal networks. *arXiv:1506.01497*.
- [17] Han, S., Mao, H., and Dally, W. J. (2015). Deep Compression: Compressing Deep Neural Networks with Pruning, Trained Quantization and Huffman Coding. *arXiv:1510.00149*.
- [18] Hwang, K., and Sung, W. (2014). Fixed-point feedforward deep neural network design using weights +1, 0, and -1. *IEEE Workshop on Signal Processing Systems(SiPS)*. pp. 1-6.
- [19] Jia, Y., et al. (2014). Caffe: Convolutional architecture for fast feature embedding. *Proceedings of the ACM International Conference on Multimedia*. pp. 675-678.
- [20] Iandola, F. N., et al. (2016). SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <1MB model size. *arXiv:1602.07360*.
- [21] Tobak, S. (2014). Google leads \$542m round in virtual reality startup magic leap.