LOSS-AWARE WEIGHT QUANTIZATION OF DEEP NETWORKS

Lu Hou, James T. Kwok
Department of Computer Science and Engineering
Hong Kong University of Science and Technology
Hong Kong
{lhouab, jamesk}@cse.ust.hk

ABSTRACT

The huge size of deep networks hinders their use in small computing devices. In this paper, we consider compressing the network by weight quantization. We extend a recently proposed loss-aware weight binarization scheme to ternarization, with possibly different scaling parameters for the positive and negative weights, and $m$-bit (where $m > 2$) quantization. Experiments on feedforward and recurrent neural networks show that the proposed scheme outperforms state-of-the-art weight quantization algorithms, and is as accurate (or even more accurate) than the full-precision network.

1 INTRODUCTION

The last decade has witnessed huge success of deep neural networks in various domains. Examples include computer vision, speech recognition, and natural language processing (LeCun et al., 2015). However, their huge size often hinders deployment to small computing devices such as cell phones and the internet of things. Many attempts have been recently made to reduce the model size. One common approach is to prune a trained dense network (Han et al., 2015; 2016). However, most of the pruned weights may come from the fully-connected layers where computations are cheap, and the resultant time reduction is insignificant. Li et al. (2017b) and Molchanov et al. (2017) proposed to prune filters in the convolutional neural networks based on their magnitudes or significance to the loss. However, the pruned network has to be retrained, which is again expensive.

Another direction is to use more compact models. GoogleNet (Szegedy et al., 2015) and ResNet (He et al., 2016) replace the fully-connected layers with simpler global average pooling. However, they are also deeper. SqueezeNet (Iandola et al., 2016) reduces the model size by replacing most of the $3 \times 3$ filters with $1 \times 1$ filters. This is less efficient on smaller networks because the dense $1 \times 1$ convolutions are costly. MobileNet (Howard et al., 2017) compresses the model using separable depth-wise convolution. ShuffleNet (Zhang et al., 2017) utilizes pointwise group convolution and channel shuffle to reduce the computation cost while maintaining accuracy. However, highly optimized group convolution and depth-wise convolution implementations are required. Alternatively, Novikov et al. (2015) compressed the model by using a compact multilinear format to represent the dense weight matrix. The CP and Tucker decompositions have also been used on the kernel tensor in CNNs (Lebedev et al., 2014; Kim et al., 2016). However, they often need expensive fine-tuning.

Another effective approach to compress the network and accelerate training is by quantizing each full-precision weight to a small number of bits. This can be further divided into two sub-categories, depending on whether pre-trained models are used (Lin et al., 2016a; Mellempudi et al., 2017) or the quantized model is trained from scratch (Courbariaux et al., 2015; Li et al., 2017a). Some of these also directly learn with low-precision weights, but they usually suffer from severe accuracy deterioration (Li et al., 2017a; Miyashita et al., 2016). By keeping the full-precision weights during learning, Courbariaux et al. (2015) pioneered the BinaryConnect algorithm, which uses only one bit for each weight while still achieving state-of-the-art classification results. Rastegari et al. (2016) further incorporated weight scaling, and obtained better results. Instead of simply finding the closest binary approximation of the full-precision weights, a loss-aware scheme is proposed in (Hou et al., 2017). Beyond binarization, TernaryConnect (Lin et al., 2016b) quantizes each weight to
\{−1, 0, 1\} \text{.} \) \cite{Li2016} and \cite{Zhu2017} added scaling to the ternarized weights, and DoReFa-Net \cite{Zhou2016} further extended quantization to more than three levels. However, these methods do not consider the effect of quantization on the loss, and rely on heuristics in their procedures \cite{Zhou2016,Zhu2017}. Recently, a loss-aware low-bit quantized neural network is proposed in \cite{Leng2017}. However, it uses full-precision weights in the forward pass and the extra-gradient method \cite{Vasilyev2010} for update, both of which are expensive.

In this paper, we propose an efficient and disciplined ternarization scheme for network compression. Inspired by \cite{Hou2017}, we explicitly consider the effect of ternarization on the loss. This is formulated as an optimization problem which is then solved efficiently by the proximal Newton algorithm. When the loss surface’s curvature is ignored, the proposed method reduces to that of \cite{Li2016}. Let \( l \) be a vector containing the diagonal of an approximate Hessian of \( \ell \). \( \{l\} \) is quantized to \( \{\Delta\} \), and is also related to the projection step of \cite{Leng2017}. Next, we extend it to (i) allow the use of different scaling parameters for the positive and negative weights; and (ii) the use of \( m \) bits (where \( m > 2 \)) for weight quantization. Experiments on both feedforward and recurrent neural networks show that the proposed quantization scheme outperforms state-of-the-art algorithms.

**Notations:** For a vector \( x \), \( \sqrt{x} \) denotes the element-wise square root (i.e., \( \sqrt{\|x\|_p} = \sqrt{\sum_i |x_i|^p} \) is its \( p \)-norm, and \( \text{Diag}(x) \) returns a diagonal matrix with \( x \) on the diagonal. For two vectors \( x \) and \( y \), \( x \odot y \) denotes the element-wise multiplication and \( x \odot y \) the element-wise division. Given a threshold \( \Delta \), \( \text{Diag}(x) \) returns a vector such that \( \text{Diag}(x)_i = 1 \) if \( x_i > \Delta \), \( -1 \) if \( x_i < -\Delta \), and 0 otherwise. \( \text{Diag}(x) \) considers only the positive threshold, i.e., \( \text{Diag}(x)_i = 1 \) if \( x_i > \Delta \), and 0 otherwise. Similarly, \( \text{Diag}(x)_i = -1 \) if \( x_i < -\Delta \), and 0 otherwise. For a matrix \( X \), \( \text{vec}(X) \) returns a vector by stacking all the columns of \( X \), and \( \text{diag}(X) \) returns a vector whose entries are from the diagonal of \( X \).

## 2 Related Work

Let the full-precision weights from all \( L \) layers be \( w = [w_1^T, w_2^T, \ldots, w_L^T]^T \), where \( w_l = \text{vec}(W_l) \), and \( W_l \) is the weight matrix at layer \( l \). The corresponding quantized weights will be denoted \( \hat{w} = [\hat{w}_1^T, \hat{w}_2^T, \ldots, \hat{w}_L^T]^T \).

### 2.1 Weight Binarized Networks

In BinaryConnect \cite{Courbariaux2015}, each element of \( w_l \) is binarized to \(-1 \) or \(+1 \) by using the sign function: \( \text{Binarize}(w_l) = \text{sign}(w_l) \). In the Binary-Weight-Network (BWN) \cite{Rastegari2016}, a scaling parameter is also included, i.e., \( \text{Binarize}(w_l) = \alpha_l b_l \), where \( \alpha_l > 0 \), \( b_l \in \{-1, 1\}\) and \( n_l \) is the number of weights in \( w_l \). By minimizing the difference between \( w_l \) and \( \alpha_l b_l \), the optimal \( \alpha_l \), \( b_l \) have the simple form: \( \alpha_l = \|w_l\|_1/n_l \), and \( b_l = \text{sign}(w_l) \).

Instead of simply finding the best binary approximation for the full-precision weight \( w_l^t \) at iteration \( t \), the loss-aware binarized network (LAB) directly minimizes the loss w.r.t. the binarized weight \( \alpha_l^t b_l^t \) \cite{Hou2017}. Let \( d_l^t \) be a vector containing the diagonal of an approximate Hessian of the loss. It can be shown that \( \alpha_l^t = \|d_l^t\|_1 \odot \|w_l^t\|_1 / \|d_l^t\|_1 \) and \( b_l^t = \text{sign}(w_l^t) \).

### 2.2 Weight Ternarized Networks

In a weight ternarized network, zero is used as an additional quantized value. In TernaryConnect \cite{Lin2016}, each weight value is clipped to \([-1, 1]\) before quantization, and then a non-negative weight \( |w_l^t| \) is stochastically quantized to 1 with probability \( |w_l^t| \) (and 0 otherwise). When \( |w_l^t| \) is negative, it is quantized to \(-1 \) with probability \(-|w_l^t| \), and 0 otherwise.

In the ternary weight network (TWN) \cite{Li2016}, \( w_l^t \) is quantized to \( \hat{w}_l^t = \alpha_l^t \text{Diag}(\Delta_l^t(w_l^t)) \), where \( \Delta_l^t \) is a threshold (i.e., \( \hat{w}_l^t = \alpha_l^t \) if \( |w_l^t| > \Delta_l^t \), \( -\alpha_l^t \) if \( |w_l^t| < -\Delta_l^t \) and 0 otherwise). To obtain \( \Delta_l^t \) and \( \alpha_l^t \), TWN minimizes the \( \ell_2 \)-distance between the full-precision and ternarized
weights, leading to
\[
\Delta^l_t = \arg \max_{\Delta > 0} \frac{1}{\| \Delta \|_1} \left( \frac{1}{\alpha^l} \sum_{i: |w^l_i| > \Delta^l_t} |w^l_i| \right)^2, \quad \alpha^l_t = \frac{1}{\| \Delta^l_t \|_1} \sum_{i: |w^l_i| > \Delta^l_t} |w^l_i|.
\]
(1)
However, \( \Delta^l_t \) in (1) is difficult to solve. Instead, TWN simply sets \( \Delta^l_t = 0.7 \cdot E(|w^l_i|) \) in practice.

In TWN, one scaling parameter (\( \alpha^l_t \)) is used for both the positive and negative weights at layer \( l \). In the trained ternary quantization (TTQ) network (Zhu et al., 2017), different scaling parameters (\( \alpha^l_t \) and \( \beta^l_t \)) are used. The weight \( w^l_i \) is thus quantized to \( w^l_i = \alpha^l_t \frac{I_{\Delta^l_t}}{\Delta^l_t}(w^l_i) + \beta^l_t I_{\Delta^l_t}(w^l_i) \). The scaling parameters are learned by gradient descent. As for \( \Delta^l_t \), two heuristics are used. The first sets \( \Delta^l_t \) to a constant fraction of \( \max(|w^l_i|) \), while the second sets \( \Delta^l_t \) such that all layers are equally sparse.

### 2.3 Weight Quantized Networks

In a weight quantized network, \( m \) bits (where \( m \geq 2 \)) are used to represent each weight. Let \( Q \) be a set of \( (2k + 1) \) quantized values, where \( k = 2^m - 1 \). The two popular choices of \( Q \) are \( \{-1, -\frac{k-1}{2}, \ldots, -\frac{1}{2}, 0, \frac{1}{2}, \ldots, \frac{k-1}{2}, 1\} \) (linear quantization), and \( \{-1, -\frac{1}{2}, \ldots, -\frac{1}{2^n}, 0, \frac{1}{2}, \ldots, \frac{1}{2^n}, 1\} \) (logarithmic quantization). By limiting the quantized values to powers of two, logarithmic quantization is advantageous in that expensive floating-point operations can be replaced by cheaper bit-shift operations. When \( m = 2 \), both schemes reduce to \( Q = \{-1, 0, 1\} \).

In the DoReFa-Net (Zhou et al., 2016), weight \( w^l_i \) is heuristically quantized to \( m \)-bit, with \(^1\)
\[
|w^l_i| \in \{2 \cdot \text{quantize}_{m} \left( \frac{\tanh(|w^l_i|)}{2 \max(|\tanh(|w^l_i|)|)} + \frac{1}{2} \right) - 1 \}
\]
in \( \{-1, -\frac{2^m - 2}{2^{m-1}}, \ldots, -\frac{1}{2^n}, \frac{1}{2^n}, \ldots, \frac{2^m - 2}{2^{m-1}}, 1\} \), where \( \text{quantize}_{m}(x) = \frac{1}{2^{m-1}} \text{round}(2^m - 1 \cdot x) \). Similar to loss-aware binarization (Hou et al., 2017), Leng et al. (2017) proposed a loss-aware quantized network called low-bit neural network (LBNN). The alternating direction method of multipliers (ADMM) (Boyd et al., 2011) is used for optimization. At the \( t \)th iteration, the full-precision weight \( w^l_i \) is first updated by the method of extra-gradient (Vasilyev et al., 2010):
\[
\tilde{w}^l_i = w^{l-1} - \eta^l \nabla \mathcal{L}(w^{l-1}), \quad w^l_i = w^{l-1} - \eta^l \nabla \mathcal{L}(\tilde{w}^l_i),
\]
(2)
where \( \mathcal{L} \) is the augmented Lagrangian in the ADMM formulation, and \( \eta^l \) is the stepsize. Next, \( w^l_i \) is projected to the space of \( m \)-bit quantized weights so that \( \hat{w}^l_i \) is of the form \( \alpha l b_i \), where \( \alpha l > 0 \), and \( b_i \in \{-1, -\frac{1}{2}, \ldots, -\frac{1}{2^n}, 0, \frac{1}{2^n}, \ldots, \frac{1}{2} \} \).

### 3 Loss-Aware Quantization

#### 3.1 Ternarization using Proximal Newton Algorithm

In weight ternarization, TWN simply finds the closest ternary approximation of the full precision weight at each iteration, while TTQ sets the ternarization threshold heuristically. Inspired by LAB (for binarization), we consider the loss explicitly during quantization and obtain the quantization thresholds and scaling parameter by solving an optimization problem.

As in TWN, the weight \( w_l \) is ternarized as \( w_l = \alpha l b_l \), where \( \alpha l > 0 \) and \( b_l \in \{-1, 0, 1\}^{n_l} \). Given a loss function \( \ell \), we formulate weight ternarization as the following optimization problem:
\[
\min_{\hat{w}} \ell(\hat{w}) : \hat{w}_l = \alpha l b_l, \alpha l > 0, b_l \in Q^{n_l}, \ l = 1, \ldots, L,
\]
(3)
where \( Q \) is the set of desired quantized values. As in LAB, we will solve this using the proximal Newton method (Lee et al., 2014, Rakotomamonjy et al., 2016). At iteration \( t \), the objective is replaced by the second-order expansion
\[
\ell(\hat{w}^{t-1}) + \nabla \ell(\hat{w}^{t-1})^\top (\hat{w} - \hat{w}^{t-1}) + \frac{1}{2} (\hat{w} - \hat{w}^{t-1})^\top H^{t-1} (\hat{w} - \hat{w}^{t-1}),
\]
(4)
\(^1\)Note that the quantized value of 0 is not used in DoReFa-Net.
where $H^{t-1}$ is an estimate of the Hessian of $\ell$ at $\hat{w}^{t-1}$. We use the diagonal equilibration pre-conditioner (Dauphin et al. [2015]), which is robust in the presence of saddle points and also readily available in popular stochastic deep network optimizers such as Adam (Kingma & Ba [2015]). Let $D_l$ be the approximate diagonal Hessian at layer $l$. We use $D = \text{Diag}(\text{diag}(D_1), \ldots, \text{diag}(D_L))$ as an estimate of $H$. Substituting (4) into (3), we solve the following subproblem at the $t$th iteration:

$$\min_{w^t} \nabla \ell(\hat{w}^{t-1})^\top (w^t - \hat{w}^{t-1}) + \frac{1}{2} (w^t - \hat{w}^{t-1})^\top D^{t-1} (w^t - \hat{w}^{t-1})$$

s.t. $\hat{w}^t = \alpha_t b^t$, $\alpha_t > 0$, $b^t \in \mathcal{Q}_n^l$, $l = 1, \ldots, L$.

**Proposition 3.1** The objective in (5) can be rewritten as

$$\min_{w^t} \frac{1}{2} \sum_{l=1}^L \left( \sqrt{d_{l}^{t-1}} (w^t_l - \hat{w}^t_l) \right)^2,$$

where $d_{l}^{t-1} \equiv \text{diag}(D_{l}^{t-1})$, and

$$w^t_l = \hat{w}^{t-1}_l - \nabla_l \ell(\hat{w}^{t-1}) \odot d_{l}^{t-1}.$$

Obviously, this objective can be minimized layer by layer. Each proximal Newton iteration thus consists of two steps: (i) Obtain $w^t_l$ in (7) by gradient descent along $\nabla_l \ell(\hat{w}^{t-1})$, which is preconditioned by the adaptive learning rate $1 \odot d_{l}^{t-1}$ so that the rescaled dimensions have similar curvatures; (ii) Quantize $w^t_l$ to $\hat{w}^t_l$ by minimizing the scaled difference between $w^t_l$ and $\hat{w}^t_l$ in (6). Intuitively, when the curvature is low ($[d_{i}^{t-1}]_i$ is small), the loss is not sensitive to the weight and ternarization error can be less penalized. When the loss surface is steep, ternarization has to be more accurate.

Though the constraint in (6) is more complicated than that in LAB, interestingly the following simple relationship can still be obtained for weight ternarization.

**Proposition 3.2** With $Q = \{-1, 0, 1\}$, and the optimal $\hat{w}^t_l$ in (6) of the form $\alpha b$. For a fixed $b$, $\alpha = \frac{\|b \odot d_{l}^{t-1} \odot w^t_l\|_1}{\|b \odot d_{l}^{t-1}\|_1}$; whereas when $\alpha$ is fixed, $b = \mathbf{I}_{\alpha/2}(w^t_l)$.

Equivalently, $b$ can be written as $\Pi_Q(w^t_l/\alpha)$, where $\Pi_Q(\cdot)$ projects each entry of the input argument to the nearest element in $Q$. Further discussions on how to solve for $\alpha^t_l$ will be presented in Sections 3.1.1 and 3.1.2. When the curvature is the same for all dimensions at layer $l$, the following Corollary shows that the solution above reduces that of TWN.

**Corollary 3.1** When $D_{l}^{t-1} = \lambda I$, $\alpha_t^l$ reduces to the TWN solution in (4) with $\Delta_t^l = \alpha_t^l/2$.

In other words, TWN corresponds to using the proximal gradient algorithm, while the proposed method corresponds to using the proximal Newton algorithm with diagonal Hessian. In composite optimization, it is known that the proximal Newton algorithm is more efficient than the proximal gradient algorithm (Lee et al. [2014], Rakotomamonjy et al. [2016]). Moreover, note that the interesting relationship $\Delta_t^l = \alpha_t^l/2$ is not observed in TWN, while TTQ completely neglects this relationship.

In LBNN (Leng et al. [2017]), its projection step uses an objective which is similar to (6), but without using the curvature information. Besides, their $w^t_l$ is updated with the extra-gradient in (2), which doubles the number of forward, backward and update steps, and can be costly. Moreover, LBNN uses full-precision weights in the forward pass, while all other quantization methods including ours use quantized weights (which eliminates most of the multiplications and thus faster training).

When (i) $\ell$ is continuously differentiable with Lipschitz-continuous gradient (i.e., there exists $\beta > 0$ such that $\|\nabla \ell(u) - \nabla \ell(v)\|_2 \leq \beta \|u - v\|_2$ for any $u, v$); (ii) $\ell$ is bounded from below; and (iii) $|d_{k}^{t}|_k > \beta \forall l, k, t$, it can be shown that the objective of (3) produced by the proximal Newton algorithm (with solution in Proposition 3.2) converges (Hou et al. [2017]). In practice, it is important to keep the full-precision weights during update (Courbariaux et al. [2015]). Hence, we replace (7) by $w^t_l \leftarrow w^{t-1}_l - \nabla_l \ell(\hat{w}^{t-1}_l) \odot d_{l}^{t-1}$. The whole procedure, which is called Loss-Aware Ternarization (LAT), is shown in Algorithm 3 of Appendix B. It is similar to Algorithm 1 of LAB (Hou et al. [2017]), except that $\alpha^t_l$ and $b^t_l$ are computed differently. In step 4, following Li & Liu [2016], we first rescale input $x^{t-1}_l$ with $\alpha_l$, so that multiplications in dot products and convolutions become additions. Algorithm 3 can also be easily extended to ternarize weights in recurrent networks. Interested readers are referred to Hou et al. [2017] for details.
3.1.1 Exact Solution of $\alpha_f^1$

To simplify notations, we drop the superscripts and subscripts. From Proposition 3.2,

$$\alpha = \frac{\|b \odot d \odot w\|_1}{\|b \odot d\|_1}, \quad b = I_{\alpha/2}(w).$$  \hspace{1cm} (8)

We now consider how to solve for $\alpha$. First, we introduce some notations. Given a vector $x = [x_1, x_2, \ldots, x_n]$, and an indexing vector $s \in \mathbb{R}^n$ whose entries are a permutation of $\{1, \ldots, n\}$, $\text{perm}_a(x)$ returns the vector $[x_{s_1}, x_{s_2}, \ldots, x_{s_n}]$, and $\text{cum}(x) = [x_1, \sum_{i=1}^2 x_i, \ldots, \sum_{i=1}^n x_i]$ returns partial sums for elements in $x$. For example, let $a = [1, -1, -2]$, and $b = [3, 1, 2]$. Then, $\text{perm}_b(a) = [-2, 1, -1]$ and cum($a$) = $[1, 0, -2]$.

We sort elements of $|w|$ in descending order, and let the vector containing the sorted indices be $s$. For example, if $w = [1, 0, -2]$, then $s = [3, 1, 2]$. From (8),

$$\alpha = \frac{\|I_{\alpha/2}(w) \odot d \odot w\|_1}{\|I_{\alpha/2}(w) \odot d\|_1} = \frac{\text{cum}(\text{perm}_s((d \odot w)))_j}{\text{cum}(\text{perm}_s(d))_j} = 2c_j,$$

where $c = \text{cum}(\text{perm}_s((d \odot w))) \odot \text{cum}(\text{perm}_s(d)) \odot 2$, and $j$ is the index such that

$$\text{cum}(\text{perm}_s(|w|))_j > c_j > \text{cum}(\text{perm}_s(|w|))_{j+1}.$$  \hspace{1cm} (10)

For simplicity of notations, let the dimensionality of $w$ (and thus also of $c$) be $n$, and the operation find(condition($x$)) returns all indices in $x$ that satisfies the condition. It is easy to see that any $j$ satisfying (10) is in $S = \text{find}(\text{cum}(\text{perm}_s(|w|))_{1:(n-1)} - c_{1:(n-1)}) \odot (\text{cum}(\text{perm}_s(|w|))_{2:n} - c_{1:n-1}) < 0$, where $c_{1:(n-1)}$ is the subvector of $c$ with elements in the index range 1 to $n - 1$. The optimal $\alpha$ (= $2c_j$) is then the one which yields the smallest objective in (6), which can be simplified by Proposition 3.3 below. The procedure is shown in Algorithm 1.

**Proposition 3.3** The optimal $\alpha_f^1$ of (6) equals $2 \text{arg max}_{c_j \in S} c_j^2 \cdot \text{cum}(\text{perm}_s(d_i^{-1}))_j$.

**Algorithm 1** Exact solver of (6)

1: **Input:** full-precision weight $w_f^t$, diagonal entries of the approximate Hessian $d_i^{t-1}$.
2: $s = \text{arg sort}(|w_f^t|)$;
3: $c = \text{cum}(\text{perm}_s((d_i^{-1} \odot w_f^t))) \odot \text{cum}(\text{perm}_s(d_i^{-1})) \odot 2$;
4: $S = \text{find}((\text{cum}(\text{perm}_s(|w_f^t|))_{1:(n-1)} - c_{1:(n-1)}) \odot (\text{cum}(\text{perm}_s(|w_f^t|))_{2:n} - c_{1:n-1}) < 0)$;
5: $\alpha_f^1 = 2 \text{arg max}_{c_j \in S} c_j^2 \cdot \text{cum}(\text{perm}_s(d_i^{-1}))_j$;
6: $b_f^t = I_{\alpha_f/2}(w_f^t)$;
7: **Output:** $\hat{w}_f^t = \alpha_f^1 b_f^t$.

3.1.2 Approximate Solution of $\alpha_f^t$

In case the sorting operation in step 2 is expensive, $\alpha_f^t$ and $b_f^t$ can be obtained by alternating the iteration in Proposition 3.2 (Algorithm 2). Empirically, it converges very fast, usually in 5 iterations.

**Algorithm 2** Approximate solver for (6)

1: **Input:** $b_f^{t-1}$, full-precision weight $w_f^t$, diagonal entries of the approximate Hessian $d_i^{t-1}$.
2: **Initialize:** $\alpha = 1.0$, $\alpha_{\text{old}} = 0.0$, $b = b_f^{t-1}$, $\epsilon = 10^{-6}$;
3: while $|\alpha - \alpha_{\text{old}}| > \epsilon$ do
4: $\alpha_{\text{old}} = \alpha$;
5: $\alpha = \frac{\|b \odot d_i^{t-1} \odot w_f^t\|_1}{\|b \odot d_i^{t-1}\|_1}$;
6: $b = I_{\alpha/2}(w_f^t)$;
7: **end while**
8: **Output:** $\hat{w}_f^t = \alpha b$. 

Published as a conference paper at ICLR 2018
3.2 Extension to Ternarization with Two Scaling Parameters

As in TTQ (Zhu et al., 2017), we can use different scaling parameters for the positive and negative weights in each layer. The optimization subproblem at the $t$th iteration then becomes:

$$\min_{\mathbf{w}_t} \nabla \ell (\mathbf{w}_t^{-1})^\top (\mathbf{w}_t - \mathbf{w}_t^{-1}) + \frac{1}{2} (\mathbf{w}_t - \mathbf{w}_t^{-1})^\top \mathbf{D}_t^{-1} (\mathbf{w}_t - \mathbf{w}_t^{-1})$$  \hspace{1cm} (11)

s.t. \hspace{.2cm} \mathbf{w}_t \in \{-\beta_t^2, 0, \alpha_t^2\}^n, \quad \alpha_t^2 > 0, \quad \beta_t^2 > 0, \quad l = 1, \ldots, L.

**Proposition 3.4** The optimal $\hat{\mathbf{w}}_t^i$ in (11) is of the form $\hat{\mathbf{w}}_t^i = \alpha_t^i \mathbf{p}_t^i + \beta_t^i \mathbf{q}_t^i$, where $\alpha_t^i = \frac{||\mathbf{p}_t^i \circ \mathbf{d}_t^{-1} \circ \mathbf{w}_t||_1}{||\mathbf{p}_t^i \circ \mathbf{d}_t^{-1}||_1}$, $\mathbf{p}_t^i = \mathbf{I}_{\alpha_t^2/2}(\mathbf{w}_t)$, $\beta_t^i = \frac{||\mathbf{q}_t^i \circ \mathbf{d}_t^{-1} \circ \mathbf{w}_t||_1}{||\mathbf{q}_t^i \circ \mathbf{d}_t^{-1}||_1}$, and $\mathbf{q}_t^i = \mathbf{I}_{\beta_t^2/2}(\mathbf{w}_t)$.

The exact and approximate solutions for $\alpha_t^i$ and $\beta_t^i$ can be obtained in a similar way as in Sections 3.1.1 and 3.1.2. Details are in Appendix C.

3.3 Extension to Low-Bit Quantization

For $m$-bit quantization, we simply change the set $\mathcal{Q}$ of desired quantized values in (3) to one with $k = 2^m - 1$ quantized values. The optimization still contains a gradient descent step with adaptive learning rates like LAT, and a quantization step which can be solved efficiently by alternating minimization of $(\alpha, \mathbf{b})$ (similar to the procedure in Algorithm 2) using the following Proposition.

**Proposition 3.5** Let the optimal $\hat{\mathbf{w}}_t^i$ in (10) be of the form $\alpha \mathbf{b}$. For a fixed $\mathbf{b}$, $\alpha = \frac{||\mathbf{b} \circ \mathbf{d}_t^{-1} \circ \mathbf{w}_t||_1}{||\mathbf{b} \circ \mathbf{d}_t^{-1}||_1}$, whereas when $\alpha$ is fixed, $\mathbf{b} = \Pi_{\mathcal{Q}}(\frac{\mathbf{w}_t}{\alpha})$, where $\mathcal{Q} = \{-1, -\frac{k-1}{k}, \ldots, -\frac{1}{k}, 0, \frac{1}{k}, \ldots, \frac{k-1}{k}, 1\}$ for linear quantization and $\mathcal{Q} = \{-1, -\frac{1}{2}, \ldots, -\frac{1}{2^{m-1}}, 0, \frac{1}{2^{m-1}}, \ldots, \frac{k-1}{k}, 1\}$ for logarithmic quantization.

4 Experiments

In this section, we perform experiments on both feedforward and recurrent neural networks. The following methods are compared: (i) the original full-precision network; (ii) weight-binarized networks, including BinaryConnect (Courbariaux et al., 2015), Binary-Weight-Network (BWN) (Rastegari et al., 2016), and Loss-Aware Binarized network (LAB) (Hou et al., 2017); (iii) weight-ternarized networks, including Ternary Weight Networks (TWN) (Li & Liu, 2016), Trained Ternary Quantization (TTQ) (Zhu et al., 2017), the proposed Loss-Aware Ternarized network with exact solution (LATe), approximate solution (LATa), and with two scaling parameters (LAT2e and LAT2a); (iv) $m$-bit-quantized networks (where $m > 2$), including DoReFa-Netm (Zhou et al., 2017), the proposed loss-aware quantized network with linear quantization (LAQm(linear)), and logarithmic quantization (LAQm(log)). Since weight quantization can be viewed as a form of regularization (Courbariaux et al., 2015), we do not use other regularizers such as dropout and weight decay.

4.1 Feedforward Networks

In this section, we perform experiments with the multilayer perceptron (on the MNIST data set) and convolutional neural networks (on CIFAR-10, CIFAR-100 and SVHN). For MNIST, CIFAR-10, and SVHN, the setup is similar to that in Courbariaux et al. (2015) and Hou et al. (2017). Details can be found in Appendix D. For CIFAR-100, we use 45,000 images for training, another 5,000 for validation, and the remaining 10,000 for testing. The testing errors are shown in Table 1.

**Ternarization:** On MNIST, CIFAR100 and SVHN, the weight-ternarized networks perform better than weight-binarized networks, and are comparable to the full-precision networks. Among the weight-ternarized networks, the proposed LAT and its variants have the lowest errors. On CIFAR-10, LATa has similar performance as the full-precision network, but is outperformed by BinaryConnect.

Figure 1(a) shows convergence of the training loss for LATa on CIFAR-10, and Figure 1(b) shows the scaling parameter obtained at each CNN layer. As can be seen, the scaling parameters for the weighted networks are better than those for the weight-binarized networks.
Table 1: Testing errors (%) on the feedforward networks. Algorithm with the lowest error in each group is highlighted.

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>MNIST</th>
<th>CIFAR-10</th>
<th>CIFAR-100</th>
<th>SVHN</th>
</tr>
</thead>
<tbody>
<tr>
<td>no binarization</td>
<td>1.11</td>
<td>10.38</td>
<td>39.06</td>
<td>2.28</td>
</tr>
<tr>
<td><strong>binarization</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BinaryConnect</td>
<td>1.28</td>
<td>9.86</td>
<td>46.42</td>
<td>2.45</td>
</tr>
<tr>
<td>BWN</td>
<td>1.31</td>
<td>10.51</td>
<td>43.62</td>
<td>2.54</td>
</tr>
<tr>
<td>LAT</td>
<td><strong>1.18</strong></td>
<td><strong>10.50</strong></td>
<td><strong>43.06</strong></td>
<td><strong>2.35</strong></td>
</tr>
<tr>
<td><strong>ternarization</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 scaling</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TWN</td>
<td>1.23</td>
<td>10.64</td>
<td>43.49</td>
<td>2.37</td>
</tr>
<tr>
<td>LATe</td>
<td>1.15</td>
<td>10.47</td>
<td>39.10</td>
<td>2.30</td>
</tr>
<tr>
<td>LATa</td>
<td><strong>1.14</strong></td>
<td><strong>10.38</strong></td>
<td><strong>39.19</strong></td>
<td><strong>2.30</strong></td>
</tr>
<tr>
<td>2 scaling</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LAT12e</td>
<td>1.20</td>
<td>10.45</td>
<td>39.01</td>
<td>2.34</td>
</tr>
<tr>
<td>LAT12a</td>
<td>1.19</td>
<td>10.48</td>
<td><strong>38.84</strong></td>
<td><strong>2.35</strong></td>
</tr>
<tr>
<td><strong>3-bit quantization</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DoReFa-Net3</td>
<td>1.31</td>
<td>10.54</td>
<td>45.05</td>
<td>2.39</td>
</tr>
<tr>
<td>LAQ3(linear)</td>
<td>1.20</td>
<td>10.67</td>
<td>38.70</td>
<td>2.34</td>
</tr>
<tr>
<td>LAQ3(log)</td>
<td><strong>1.16</strong></td>
<td><strong>10.52</strong></td>
<td><strong>38.50</strong></td>
<td><strong>2.29</strong></td>
</tr>
</tbody>
</table>

first and last layers (conv1 and linear3, respectively) are larger than the others. This agrees with the finding that, to maintain the activation variance and back-propagated gradients variance during the forward and backward propagations, the variance of the weights between the \(l\)th and \((l+1)\)th layers should roughly follow \(2/(n_l+n_{l+1})\) (Glorot & Bengio, 2010). Hence, as the input and output layers are small, larger scaling parameters are needed for their high-variance weights.

![Image](image1.png)

(a) Training loss.

![Image](image2.png)

(b) Scaling parameter \(\alpha\).

**Figure 1:** Convergence of the training loss and scaling parameter by LATa on CIFAR-10.

**Using Two Scaling Parameters:** Compared to TTQ, the proposed LAT2 always has better performance. However, the extra flexibility of using two scaling parameters does not always translate to lower testing error. As can be seen, it outperforms algorithms with one scaling parameter only on CIFAR-100. We speculate this is because the capacities of deep networks are often larger than needed, and so the limited expressiveness of quantized weights may not significantly deteriorate performance. Indeed, as pointed out in (Courbariaux et al., 2015), weight quantization is a form of regularization, and can contribute positively to the performance.

**Using More Bits:** Among the 3-bit quantization algorithms, the proposed scheme with logarithmic quantization has the best performance. It also outperforms the other quantization algorithms on CIFAR-100 and SVHN. However, as discussed above, more quantization flexibility is useful only when the weight-quantized network does not have enough capacity.

### 4.2 Recurrent Networks

In this section, we follow (Hou et al., 2017) and perform character-level language modeling experiments on the long short-term memory (LSTM) (Hochreiter & Schmidhuber, 1997). The training objective is the cross-entropy loss over all target sequences. Experiments are performed on three
data sets: (i) Leo Tolstoy’s *War and Peace*; (ii) source code of the *Linux Kernel*; and (iii) *Penn Treebank* Corpus [Taylor et al., 2003]. For the first two, we follow the setting in [Karpathy et al., 2016; Hou et al., 2017]. For *Penn Treebank*, we follow the setting in [Mikolov & Zweig, 2012]. In the experiment, we tried different initializations for TTQ and then report the best. Cross-entropy values on the test set are shown in Table 2.

Table 2: Testing cross-entropy values on the LSTM. Algorithm with the lowest cross-entropy value in each group is highlighted.

<table>
<thead>
<tr>
<th></th>
<th>War and Peace</th>
<th>Linux Kernel</th>
<th>Penn Treebank</th>
</tr>
</thead>
<tbody>
<tr>
<td>no binarization</td>
<td>1.268</td>
<td>1.326</td>
<td>1.083</td>
</tr>
<tr>
<td>Binarization</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BinaryConnect</td>
<td>2.942</td>
<td>3.532</td>
<td>1.737</td>
</tr>
<tr>
<td>BWN</td>
<td>1.313</td>
<td>1.307</td>
<td><strong>1.078</strong></td>
</tr>
<tr>
<td>LAB</td>
<td><strong>1.291</strong></td>
<td><strong>1.305</strong></td>
<td>1.081</td>
</tr>
<tr>
<td>1 scaling</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TWN</td>
<td>1.290</td>
<td>1.280</td>
<td>1.045</td>
</tr>
<tr>
<td>LATe</td>
<td>1.248</td>
<td><strong>1.256</strong></td>
<td>1.022</td>
</tr>
<tr>
<td>LATa</td>
<td>1.253</td>
<td>1.264</td>
<td>1.024</td>
</tr>
<tr>
<td>TTQ</td>
<td>1.272</td>
<td>1.302</td>
<td>1.031</td>
</tr>
<tr>
<td>2 scaling</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LAT2e</td>
<td>1.239</td>
<td>1.258</td>
<td>1.018</td>
</tr>
<tr>
<td>LAT2a</td>
<td><strong>1.245</strong></td>
<td>1.258</td>
<td><strong>1.015</strong></td>
</tr>
<tr>
<td>3-bit quantization</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DoReFa-Net3</td>
<td>1.349</td>
<td>1.276</td>
<td>1.017</td>
</tr>
<tr>
<td>LAQ3(linear)</td>
<td>1.282</td>
<td>1.327</td>
<td>1.017</td>
</tr>
<tr>
<td>LAQ3(log)</td>
<td><strong>1.268</strong></td>
<td><strong>1.273</strong></td>
<td><strong>1.009</strong></td>
</tr>
<tr>
<td>4-bit quantization</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DoReFa-Net4</td>
<td>1.328</td>
<td>1.320</td>
<td>1.019</td>
</tr>
<tr>
<td>LAQ4 (linear)</td>
<td>1.294</td>
<td>1.337</td>
<td>1.046</td>
</tr>
<tr>
<td>LAQ4 (log)</td>
<td><strong>1.272</strong></td>
<td><strong>1.319</strong></td>
<td><strong>1.016</strong></td>
</tr>
</tbody>
</table>

**Ternarization**: As in Section 4.1, the proposed LATe and LATa outperform the other weight ternarization schemes, and are even better than the full-precision network on all three data sets. Figure 2 shows convergence of the training and validation losses on *War and Peace*. Among the ternarization methods, LAT and its variants converge faster than both TWN and TTQ.

![Figure 2: Convergence of the training and validation losses on War and Peace.](image)

**Using Two Scaling Parameters**: LAT2e and LAT2a outperform TTQ on all three data sets. They also perform better than using one scaling parameter on *War and Peace* and *Penn Treebank*.

**Using More Bits**: The proposed LAQ always outperforms DoReFa-Net when 3 or 4 bits are used. As noted in Section 4.1, using more bits does not necessarily yield better generalization performance, and ternarization (using 2 bits) yields the lowest validation loss on *War and Peace* and *Linux Kernel*. Moreover, logarithmic quantization is better than linear quantization. Figure 3 shows distributions of the input-to-hidden (full-precision and quantized) weights of the input gate trained after 20 epochs using LAQ3(linear) and LAQ3(log) (results on the other weights are similar). As can be seen, distributions of the full-precision weights are bell-shaped. Hence, logarithmic quantization can give finer resolutions to many of the weights which have small magnitudes.
Quantized vs Full-precision Networks: The quantized networks often perform better than the full-precision networks. We speculate that this is because deep networks often have larger-than-needed capacities, and so are less affected by the limited expressiveness of quantized weights. Moreover, low-bit quantization acts as regularization, and so contributes positively to the performance.

5 CONCLUSION

In this paper, we proposed a loss-aware weight quantization algorithm that directly considers the effect of quantization on the loss. The problem is solved using the proximal Newton algorithm. Each iteration consists of a preconditioned gradient descent step and a quantization step that projects full-precision weights onto a set of quantized values. For ternarization, an exact solution and an efficient approximate solution are provided. The procedure is also extended to the use of different scaling parameters for the positive and negative weights, and to $m$-bit (where $m > 2$) quantization. Experiments on both feedforward and recurrent networks show that the proposed quantization scheme outperforms the current state-of-the-art.

ACKNOWLEDGMENTS

This research was supported in part by the Research Grants Council of the Hong Kong Special Administrative Region (Grant 614513). We thank the developers of Theano [Theano Development Team, 2016], Pylearn2 [Goodfellow et al., 2013] and Lasagne. We also thank NVIDIA for the gift of GPU card.

REFERENCES


A Proofs

A.1 Proof of Proposition 3.1

With $w_i^t$ in (7), the objective in (5) can be rewritten as
\[
\nabla \ell(\mathbf{w}^t-\mathbf{w}^t) + \frac{1}{2}(\mathbf{w}^t - \mathbf{w}^t)^T \mathbf{D}^t - \mathbf{w}^t + \frac{1}{2} \left( \sqrt{d_i^t} \mathbf{w}_i^t - (\mathbf{w}_i^t - \mathbf{D}^t \mathbf{w}_i^t) \right)^2 + c_1
\]
\[
= \frac{1}{2} \sum_{i=1}^{L} \left( \sqrt{d_i^t} \mathbf{w}_i^t - (\mathbf{w}_i^t - \nabla \ell(\mathbf{w}_i^t) \circ \mathbf{D}_i^t) \right)^2 + c_1
\]
\[
= \frac{1}{2} \sum_{i=1}^{L} \left( \sqrt{d_i^t} \mathbf{w}_i^t - \mathbf{w}_i^t \right)^2 + c_1
\]
\[
= \frac{1}{2} \sum_{i=1}^{L} \left( \sqrt{d_i^t} \mathbf{w}_i^t - \mathbf{d}_i^t \right)^2 + c_1
\]
\[
= \frac{1}{2} \sum_{i=1}^{L} \sum_{t=1}^{n} (\mathbf{d}_i^t)_{1} |(\alpha_t^{i} \mathbf{b}_i^t - \mathbf{w}_i^t)^2 + c_1,
\]
where $c_1 = -\frac{1}{2} (\sqrt{d_i^t})^2 (\nabla \ell(\mathbf{w}_i^t) \circ \mathbf{D}_i^t)^2$ is independent of $\alpha_t^i$ and $\mathbf{b}_i^t$.

A.2 Proof of Proposition 3.2

To simplify notations, we drop the subscript and superscript. Considering one particular layer, problem (6) is of the form:
\[
\min_{\alpha, \mathbf{b}} \quad \frac{1}{2} \sum_{i=1}^{n} d_i (\alpha b_i - w_i)^2
\]
\[
s.t. \quad \alpha > 0, \quad b_i \in \{-1, 0, 1\}.
\]
When $\alpha$ is fixed,
\[
b_i = \arg \min_{b_i} \frac{1}{2} d_i (\alpha b_i - w_i)^2 = \frac{1}{2} d_i \alpha^2 (b_i - w_i / \alpha)^2 = I_{\alpha/2}(w_i).
\]
When $\mathbf{b}$ is fixed,
\[
\alpha = \arg \min_{\alpha} \frac{1}{2} \sum_{i=1}^{n} d_i (\alpha b_i - w_i)^2
\]
\[
= \arg \min_{\alpha} \frac{1}{2} \| \mathbf{b} \odot \mathbf{b} \odot \mathbf{d} \|_1 \alpha^2 - \| \mathbf{b} \odot \mathbf{d} \|_1 \alpha + c_2,
\]
\[
= \arg \min_{\alpha} \frac{1}{2} \| \mathbf{b} \odot \mathbf{b} \odot \mathbf{d} \|_1 \left( \alpha - \frac{\| \mathbf{b} \odot \mathbf{d} \odot \mathbf{w} \|_1}{\| \mathbf{b} \odot \mathbf{b} \odot \mathbf{d} \|_1} \right)^2 - \frac{1}{2} \| \mathbf{b} \odot \mathbf{d} \odot \mathbf{w} \|_1^2 + c_2
\]
\[
= \frac{\| \mathbf{b} \odot \mathbf{d} \odot \mathbf{w} \|_1}{\| \mathbf{b} \odot \mathbf{d} \|_1}
\]
\[
= \frac{\| \mathbf{b} \odot \mathbf{d} \odot \mathbf{w} \|_1}{\| \mathbf{b} \odot \mathbf{d} \|_1}.
\]

A.3 Proof of Corollary 3.1

When $\mathbf{D}_i^{t-1} = \lambda \mathbf{I}$, i.e., the curvature is the same for all dimensions in the $l$th layer, From Proposition 3.2
\[
\alpha_t^i = \frac{\| \mathbf{b} \odot \mathbf{d}_i^{t-1} \odot \mathbf{w}_i^t \|_1}{\| \mathbf{b} \odot \mathbf{d}_i^{t-1} \|_1} = \frac{\| I_{\Delta_t^i / 2}(\mathbf{w}_i^t) \odot \mathbf{w}_i^t \|_1}{\| I_{\Delta_t^i / 2}(\mathbf{w}_i^t) \|_1} = \frac{1}{\| I_{\Delta_t^i}(\mathbf{w}_i^t) \|_1} \sum_{i:|w_i^t| > \Delta_t^i} |w_i^t|_1,
\]
\[ \Delta_i^T = \frac{1}{2} \left\| I_{\alpha/2} \odot w_i^T \right\|_1 = \arg \max_{\Delta > 0} \frac{1}{\left\| I_\Delta (w_i^T) \right\|_1} \left( \sum_{i: \left\| w_i^T \right\|_1 > \Delta} \left\| w_i^T \right\|_1 \right)^2. \]

This is the same as the TWN solution in [1].

A.4 Proof of Proposition 3.3

For simplicity of notations, we drop the subscript and superscript. For each layer, we have an optimization problem of the form

\[ \arg \min_{\alpha} (\sqrt{d}^T (\alpha b - w))^2 \]

\[ = \arg \min_{\alpha} \frac{1}{\left\| I_{\alpha/2}(w) \odot d \right\|_1} \left( \alpha - \frac{\left\| I_{\alpha/2}(w) \odot d \right\|_1}{\left\| I_{\alpha/2}(w) \odot d \right\|_1} \right)^2 \]

\[ = \arg \min_{\alpha} \frac{1}{\left\| I_{\alpha/2}(w) \odot d \right\|_1} \left( \alpha - \frac{\left\| I_{\alpha/2}(w) \odot d \right\|_1}{\left\| I_{\alpha/2}(w) \odot d \right\|_1} \right)^2 \]

\[ = \arg \min_{\alpha} -\frac{2\beta}{\left\| I_{\alpha/2}(w) \odot d \right\|_1} \]

\[ = \frac{\left\| I_{\alpha/2}(w) \odot d \right\|_1}{\left\| I_{\alpha/2}(w) \odot d \right\|_1} \cdot \left\| I_{\alpha/2}(w) \odot d \right\|_1 \cdot \left\| I_{\alpha/2}(w) \odot d \right\|_1 \]

\[ = -2\alpha_2 \cdot 2\alpha_2 \cdot \left[ \text{cum} (\text{perm}_a (d)) \right]_j \]

Thus, the \( \alpha \) that minimizes \( (\sqrt{d}^T (\alpha b - w))^2 \) is \( \alpha = 2 \arg \max_{\alpha_1, \alpha_2} \left\{ 2 \alpha_2 \cdot \left[ \text{cum} (\text{perm}_a (d)) \right]_j \right\} \).

A.5 Proof for Proposition 3.4

For simplicity of notations, we drop the subscript and superscript, and consider the optimization problem:

\[ \min_{\alpha, b} \frac{1}{2} \sum_{i=1}^n d_i (\hat{w}_i - w_i)^2 \]

\[ \text{s.t. } \hat{w}_i \in \{-\beta, 0, +\alpha\}. \]

Let \( f(\hat{w}_i) = (\hat{w}_i - w_i)^2 \). Then, \( f(\alpha) = (\alpha - w_i)^2 \), \( f(0) = w_i^2 \), and \( f(-\beta) = (\beta + w_i)^2 \). It is easy to see that (i) if \( w_i < \alpha/2 \), \( f(\alpha) \) is the smallest; (ii) if \( w_i < -\beta/2 \), \( f(-1) \) is the smallest; (iii) if \( -\beta/2 \leq w_i \leq \alpha/2 \), \( f(0) \) is the smallest. In other words, the optimal \( \hat{w}_i \) satisfies

\[ \hat{w}_i = \alpha I_{\alpha/2}(w_i) + \beta I_{\beta/2}(w_i), \]

or equivalently, \( \hat{w} = \alpha p + \beta q \), where \( p = I_{\alpha/2}(w) \), and \( q = I_{\beta}(w) \).

Define \( w^+ \) and \( w^- \) such that \( |w^+| = \left\{ \begin{array}{ll} w_i & w_i > 0 \\ 0 & \text{otherwise}, \end{array} \right. \) and \( |w^-| = \left\{ \begin{array}{ll} w_i & w_i < 0 \\ 0 & \text{otherwise} \end{array} \right. \). Then,

\[ \frac{1}{2} \sum_{i=1}^n d_i (\hat{w}_i - w_i)^2 = \frac{1}{2} \sum_{i=1}^n d_i (\alpha p_i - w^+_i)^2 + \frac{1}{2} \sum_{i=1}^n d_i (\beta q_i - w^-_i)^2. \]

(12)

The objective in (12) has two parts, and each part can be viewed as a special case of the ternarization step in Proposition 3.1 (considering only with positive or negative weights). Similar to the proof for Proposition 3.2, we can obtain that the optimal \( \hat{w} = \alpha p + \beta q \) satisfies

\[ \alpha = \frac{\|p \odot d \odot w\|_1}{\|p \odot d\|_1}, \quad p = I_{\alpha/2}(w), \]

\[ \beta = \frac{\|q \odot d \odot w\|_1}{\|q \odot d\|_1}, \quad q = I_{\beta}(w). \]
A.6 Proof of Proposition 3.5

For simplicity of notations, we drop the subscript and superscript. Since \( \frac{1}{2}(\sqrt{d}^T(\alpha b - w))^2 = \frac{1}{2} \sum_{i=1}^{n} d_i(\alpha b_i - w_i)^2 \) for each layer, we simply consider the optimization problem:

\[
\min_{\alpha, b} \frac{1}{2} \sum_{i=1}^{n} d_i(\alpha b_i - w_i)^2
\]

s.t. \( \alpha > 0, b_i \in Q \).

When \( \alpha \) is fixed,

\[
b_i = \arg \min_{b_i} \frac{1}{2} d_i(\alpha b_i - w_i)^2 = \frac{1}{2} d_i \alpha^2 (b_i - w_i/\alpha)^2 = \Pi_Q \left( \frac{w_i}{\alpha} \right).
\]

When \( b \) is fixed,

\[
\alpha = \arg \min_{\alpha} \frac{1}{2} \sum_{i=1}^{n} d_i(\alpha b_i - w_i)^2 = \arg \min_{\alpha} \frac{1}{2} \| b \odot b \odot d \|_1 \alpha^2 - \| b \odot d \odot w \|_1 \alpha + c_2
\]

\[
= \arg \min_{\alpha} \frac{1}{2} \| b \odot b \odot d \|_1 \left( \alpha - \frac{1}{\| b \odot b \odot d \|_1} \right)^2 - \frac{1}{2} \| b \odot d \odot w \|_1^2
\]

\[
= \frac{\| b \odot d \odot w \|_1}{\| b \odot b \odot d \|_1}
\]

B Loss-Aware Ternarization Algorithm (LAT)

The whole procedure of LAT is shown in Algorithm 3.

C Exact and Approximate Solutions for Ternarization with Two Scaling Parameters

Let there be \( n_1 \) positive elements and \( n_2 \) negative elements in \( w_i \). For a \( n \)-dimensional vector \( x = [x_1, x_2, \ldots, x_n] \), define inverse(\( x \)) = \( [x_n, x_{n-1}, \ldots, x_1] \). As is shown in (12), the objective can be separated into two parts, and each part can be viewed as a special case of ternarization step in Proposition 3.1 dealing only with positive or negative weights. Thus the exact and approximate solutions for \( \alpha_t^l \) and \( \beta_t^l \) can separately be derived in a similar way as that of using one scaling parameter. The exact and approximate solutions for \( \alpha_t^l \) and \( \beta_t^l \) for layer-\( l \) at the \( t \)th time step are shown in Algorithms 4 and 5.

D Experimental Details

D.1 Setup for Feedforward Networks

The setup for the four data sets are as follows:

1. MNIST: This contains 28 \( \times \) 28 gray images from 10 digit classes. We use 50,000 images for training, another 10,000 for validation, and the remaining 10,000 for testing. We use the 4-layer model:

\[
784 \text{FC} - 2048 \text{FC} - 2048 \text{FC} - 2048 \text{FC} - 10 \text{SVM},
\]

where \( FC \) is a fully-connected layer, and \( SVM \) is a \( \ell_2 \)-SVM output layer using the square hinge loss. Batch normalization with a minibatch size 100, is used to accelerate learning. The maximum number of epochs is 50. The learning rate starts at 0.01, and decays by a factor of 0.1 at epochs 15 and 25.
Algorithm 3 Loss-Aware Ternarization (LAT) for training a feedforward neural network.

Input: Minibatch \(\{(x_0^t, y^t)\}\), current full-precision weights \(\{w_i^t\}\), first moment \(\{m_i^{t-1}\}\), second moment \(\{v_i^{t-1}\}\), and learning rate \(\eta^t\).

1: **Forward Propagation**
2:   for \(l = 1\) to \(L\) do
3:     compute \(a_i^t\) and \(b_i^t\) using Algorithm 1 or 2.
4:     rescale the layer-\(l\) input: \(\hat{x}_i^{t-1} = a_i^t \hat{x}_i^{t-1}\).
5:     compute \(z_i^t\) with input \(\hat{x}_i^{t-1}\) and binary weight \(b_i^t\).
6:     apply batch-normalization and nonlinear activation to \(z_i^t\) to obtain \(x_i^t\).
7:   end for
8: compute the loss \(\ell\) using \(x_L^t\) and \(y^t\).
9: **Backward Propagation**
10: initialize output layer’s activation’s gradient \(\frac{\partial \ell}{\partial x_L^t}\).
11: for \(l = L\) to \(2\) do
12:     compute \(\frac{\partial \ell}{\partial x_{l-1}^t}\) using \(\frac{\partial \ell}{\partial x_L^t}\), \(a_i^t\) and \(b_i^t\).
13: end for
14: Update parameters using Adam
15: for \(l = 1\) to \(L\) do
16:     compute gradients \(\nabla_i \ell(\hat{w}^t)\) using \(\frac{\partial \ell}{\partial x_{l-1}^t}\) and \(x_{l-1}^t\).
17:     update first moment \(\hat{m}_i^t = \beta_1 \hat{m}_i^{t-1} + (1 - \beta_1) \nabla_i \ell(\hat{w}^t)\).
18:     update second moment \(\hat{v}_i^t = \beta_2 \hat{v}_i^{t-1} + (1 - \beta_2) (\nabla_i \ell(\hat{w}^t) \circ \nabla_i \ell(\hat{w}^t))\).
19:     compute unbiased first moment \(m_i^t = \hat{m}_i^t/(1 - \beta_1^t)\).
20:     compute unbiased second moment \(v_i^t = \hat{v}_i^t/(1 - \beta_2^t)\).
21:     compute current curvature matrix \(d_i^t = \frac{1}{\eta^t} (\epsilon + \sqrt{v_i^t})\).
22: update full-precision weights \(w_i^{t+1} = w_i^t - m_i^t \odot d_i^t\).
23: update learning rate \(\eta^{t+1} = \text{UpdateLearningrate}(\eta^t, t + 1)\).
24: end for

Algorithm 4 Exact solver for \(\hat{w}_i^t\) with two scaling parameters.

1: **Input:** full-precision weight \(w_i^t\), diagonal entries of the approximate Hessian \(d_i^{t-1}\).
2: \(s_1 = \arg\text{sort}(w_i^t)\);
3: \(c_1 = \text{cum}(\text{perm}_{s_1}(d_i^{t-1} \odot w_i^t)) \odot \text{cum}(\text{perm}_{s_1}(d_i^{t-1})) \odot 2\);
4: \(S_1 = \text{find}([-\text{perm}_{s_1}(w_i^t)]_{1:(n_1-1)} - [c_1]_{1:(n_1-1)}) \odot [-\text{perm}_{s_1}(w_i^t)]_{2:n_1} - [c_1]_{1:n_1-1}) < 0\);
5: \(\alpha_i^t = 2 \arg\max_{c_i \in S_1} [c_i]_i^2 \cdot [\text{cum}(\text{perm}_{s_1}(d_i^{t-1}))];\)
6: \(p_i^t = I_{\alpha/\alpha_i^t}(w_i^t)\);
7: \(s_2 = \text{inverse}(s_1)\);
8: \(c_2 = \text{cum}(\text{perm}_{s_2}(d_i^{t-1} \odot w_i^t)) \odot \text{cum}(\text{perm}_{s_2}(d_i^{t-1})) \odot 2\);
9: \(S_2 = \text{find}([-\text{perm}_{s_2}(w_i^t)]_{1:(n_2-1)} - [c_2]_{1:(n_2-1)}) \odot ([-\text{perm}_{s_2}(w_i^t)]_{2:n_2} - [c_2]_{1:n_2-1}) < 0\);
10: \(\beta_i^t = 2 \arg\max_{c_i \in S_2} [c_i]_i^2 \odot [\text{cum}(\text{perm}_{s_2}(d_i^{t-1}));\)
11: \(q_i^t = I_{\beta_i^t}(w_i^t)\);
12: **Output:** \(\hat{w}_i^t = \alpha_i^t p_i^t + \beta_i^t q_i^t\).

2. **CIFAR-10:** This contains \(32 \times 32\) color images from 10 object classes. We use 45,000 images for training, another 5,000 for validation, and the remaining 10,000 for testing. The images are preprocessed with global contrast normalization and ZCA whitening. We use the VGG-like architecture:

\[(2 \times 128C3) - M P2 - (2 \times 256C3) - MP2 - (2 \times 512C3) - MP2 - (2 \times 1024FC) - 10SV M,\]

where \(C3\) is a \(3 \times 3\) ReLU convolution layer, and \(MP2\) is a \(2 \times 2\) max-pooling layer. Batch normalization with a minibatch size of 50, is used. The maximum number of epochs
Algorithm 5 Approximate solver for $\hat{w}_t^i$ with two scaling parameters

1. **Input**: $b_t^i$, full-precision weight $w_t^i$, and diagonal entries of approximate Hessian $d_t^i$.
2. **Initialize**: $\alpha = 1.0, \alpha_{old} = 0.0, \beta = 1.0, \beta_{old} = 0.0, b = b_t^i, p = I^b_0(b), q = I^b_0(b), \epsilon = 10^{-6}$.
3. **while** $|\alpha - \alpha_{old}| > \epsilon$ and $|\beta - \beta_{old}| > \epsilon$ **do**
   4. $\alpha_{old} = \alpha, \beta_{old} = \beta$;
   5. $\alpha = \frac{1}{\|p \odot d_t^i \odot w_t^i\|_1}$;
   6. $p = I^{1/2}_{\alpha}(w_t^i)$;
   7. $\beta = \frac{1}{\|q \odot d_t^i \odot w_t^i\|_1}$;
   8. $q = I^{1/2}_{\beta}(w_t^i)$;
9. **end while
10. **Output**: $\hat{w}_t^i = \alpha p + \beta q$.

is 200. The learning rate for the weight-binarized network starts at 0.03 while for all the other networks starts at 0.002, and decays by a factor of 0.5 after every 15 epochs.

3. **CIFAR-100**: This contains $32 \times 32$ color images from 100 object classes. We use 45,000 images for training, another 5,000 for validation, and the remaining 10,000 for testing. The images are preprocessed with global contrast normalization and ZCA whitening. We use the VGG-like architecture:

   $$(2 \times 128C3) - MP2 - (2 \times 256C3) - MP2 - (2 \times 512C3) - MP2 - (2 \times 1024FC) - 100SVM.$$  

Batch normalization with a minibatch size of 100, is used. The maximum number of epochs is 200. The learning rate starts at 0.0005, and decays by a factor of 0.5 after every 15 epochs.

4. **SVHN**: This contains $32 \times 32$ color images from 10 digit classes. We use 598,388 images for training, another 6,000 for validation, and the remaining 26,032 for testing. The images are preprocessed with global and local contrast normalization. The model used is:

   $$(2 \times 64C3) - MP2 - (2 \times 128C3) - MP2 - (2 \times 256C3) - MP2 - (2 \times 1024FC) - 10SVM.$$  

Batch normalization with a minibatch size of 50, is used. The maximum number of epochs is 50. The learning rate starts at 0.001 for the weight-binarized network, and 0.0005 for the other networks. It then decays by a factor of 0.1 at epochs 15 and 25.

### D.2 Setup for Recurrent Networks

The setup for the three data sets are as follows:

1. **Leo Tolstoy’s *War and Peace***: It consists of 3258K characters of almost entirely English text with minimal markup and a vocabulary size of 87. We use the same training/validation/test set split as in [Karpathy et al., 2016; Hou et al., 2017].

2. The source code of the **Linux Kernel**: This consists of 621K characters and a vocabulary size of 101. We use the same training/validation/test set split as in [Karpathy et al., 2016; Hou et al., 2017].

3. The **Penn Treebank** data set ([Taylor et al., 2003]): This has been frequently used for language modeling. It contains 50 different characters, including English characters, numbers, and punctuations. We follow the setting in [Mikolov & Zweig, 2012], with 5,017K characters for training, 393K for validation, and 442K characters for testing.

We use a one-layer LSTM with 512 cells. The maximum number of epochs is 200, and the number of time steps is 100. The initial learning rate is 0.002. After 10 epochs, it is decayed by a factor of 0.98 after each epoch. The weights are initialized uniformly in $[0.08, 0.08]$. After each iteration, the gradients are clipped to the range $[-5, 5]$. All the updated weights are clipped to $[-1, 1]$ for binarization and ternarization methods, but not for $m$-bit (where $m > 2$) quantization methods.