TRAINED TERNARY QUANTIZATION

- Chenzhuo Zhu, Song Han, Huizi Mao, William J. Dally
- Tensorflow
- ICLR 2017

Overview

- **two quantization factors** for positive and negative weights
- trained quantization by back propagating **two gradients**

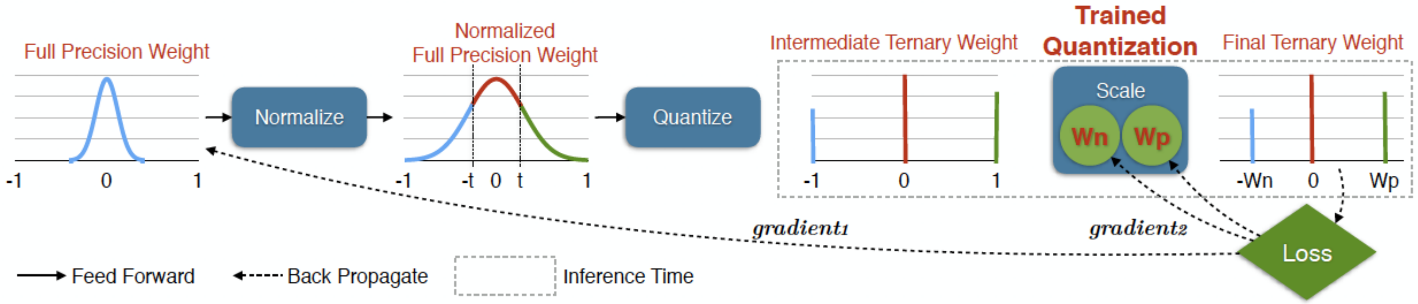


Figure 1: Overview of the trained ternary quantization procedure.

TTQ

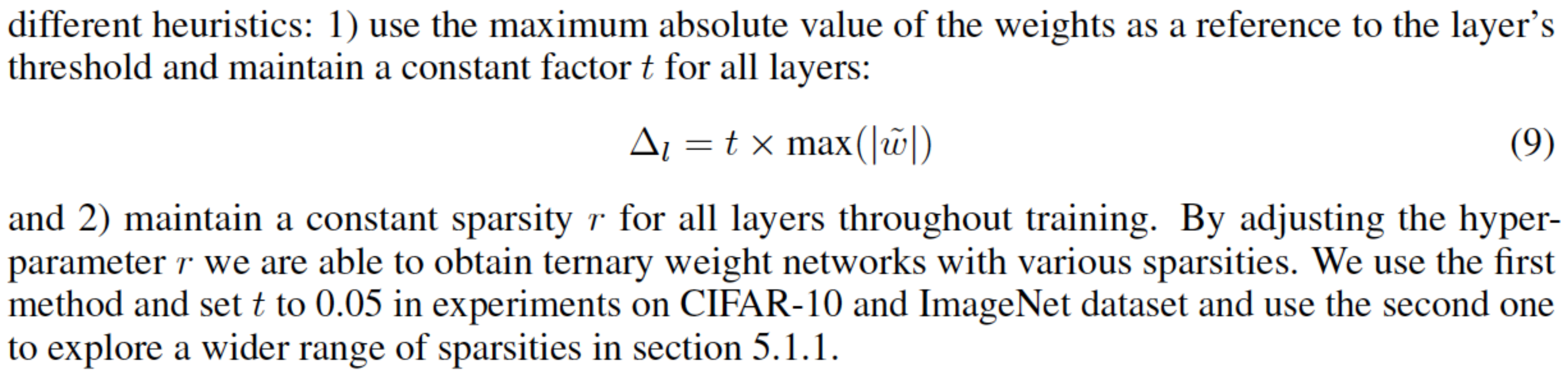
$$w_l^t = \begin{cases} W_l^p : \tilde{w}_l > \Delta_l \\ 0 : |\tilde{w}_l| \leq \Delta_l \\ -W_l^n : \tilde{w}_l < -\Delta_l \end{cases} \tag{6}$$

$$\frac{\partial L}{\partial W_l^p} = \sum_{i \in I_l^p} \frac{\partial L}{\partial w_l^t(i)}, \frac{\partial L}{\partial W_l^n} = \sum_{i \in I_l^n} \frac{\partial L}{\partial w_l^t(i)} \tag{7}$$

Here $I_l^p = \{i|\tilde{w}_l(i) > \Delta_l\}$ and $I_l^n = \{i|(i)\tilde{w}_l < -\Delta_l\}$.

$$\frac{\partial L}{\partial \tilde{w}_l} = \begin{cases} W_l^p \times \frac{\partial L}{\partial w_l^t} : \tilde{w}_l > \Delta_l \\ 1 \times \frac{\partial L}{\partial w_l^t} : |\tilde{w}_l| \leq \Delta_l \\ W_l^n \times \frac{\partial L}{\partial w_l^t} : \tilde{w}_l < -\Delta_l \end{cases} \tag{8}$$

QUANTIZATION HEURISTIC

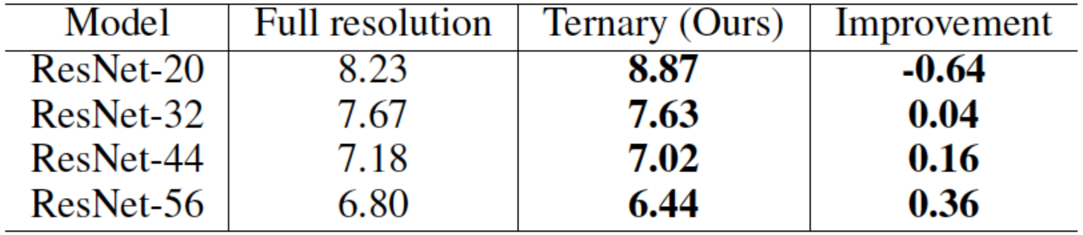


different heuristics: 1) use the maximum absolute value of the weights as a reference to the layer's threshold and maintain a constant factor t for all layers:

$$\Delta_l = t \times \max(|\tilde{w}|) \tag{9}$$

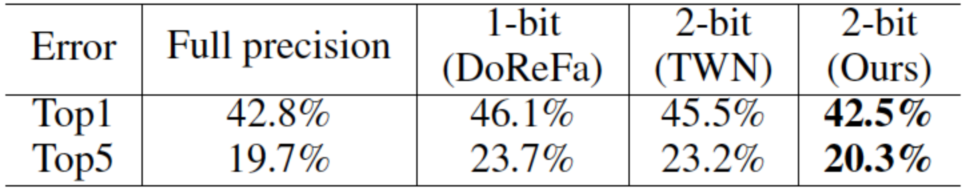
and 2) maintain a constant sparsity r for all layers throughout training. By adjusting the hyperparameter r we are able to obtain ternary weight networks with various sparsities. We use the first method and set t to 0.05 in experiments on CIFAR-10 and ImageNet dataset and use the second one to explore a wider range of sparsities in section 5.1.1.

Experiments



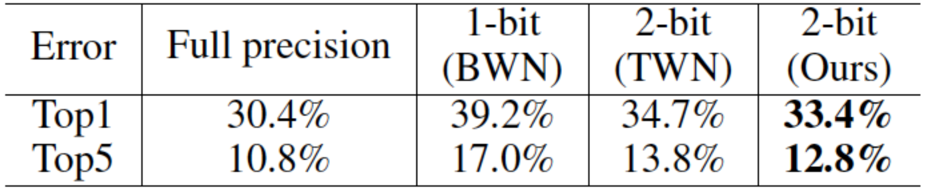
Model	Full resolution	Ternary (Ours)	Improvement
ResNet-20	8.23	**8.87**	**-0.64**
ResNet-32	7.67	**7.63**	**0.04**
ResNet-44	7.18	**7.02**	**0.16**
ResNet-56	6.80	**6.44**	**0.36**

Table 1: Error rates of full-precision and ternary ResNets on Cifar-10



Error	Full precision	1-bit (DoReFa)	2-bit (TWN)	2-bit (Ours)
Top1	42.8%	46.1%	45.5%	**42.5%**
Top5	19.7%	23.7%	23.2%	**20.3%**

Table 2: Top1 and Top5 error rate of AlexNet on ImageNet



Error	Full precision	1-bit (BWN)	2-bit (TWN)	2-bit (Ours)
Top1	30.4%	39.2%	34.7%	**33.4%**
Top5	10.8%	17.0%	13.8%	**12.8%**

Table 3: Top1 and Top5 error rate of ResNet-18 on ImageNet