Deep Neural Network Quantization via Layer-Wise Optimization using Limited Training Data

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

School of Computer Science and Engineering
College of Engineering

Shangyu Chen, Wenya Wang, Sinno Jialin Pan

schen025@e.ntu.edu.sg, {wangwy, sinnopan}@ntu.edu.sg

Motivation

Most prevailing low bits compression methods relied on a heavy training process with large amount of labeled data.

Contribution

We propose a new layer-wise quantization method for limited training data scenario:

1. Layer-wise parameters are quantized using closed-form solution.
2. Preserves performance with theoretical guarantee.
3. A small portion of training data (1% in experiments) is required in the whole process.

Cascade Layer-Wise Quantization

Notations:

- Quantized weights: $\hat{\mathbf{\Theta}}$
- Un-Quantized weights: $\bar{\mathbf{\Theta}}$, $\bar{\mathbf{\Theta}}^{new}$
- Inputs to the l-th quantized layer: $\hat{\mathbf{Y}}^{l-1} = f(\mathbf{Y}^0; \hat{\mathbf{\Theta}}_{[1,\ldots,l-1]})$
- Origin inputs: $\mathbf{Y}^{l-1}$

Goal:

Divergence of final layer output before and after quantization is minimized:

$$\min_{\hat{\mathbf{\Theta}}_{[l,\ldots,L]}} ||f(\hat{\mathbf{Y}}^{l-1}; \hat{\mathbf{\Theta}}_{[l,\ldots,L]}) - f(\mathbf{Y}^{l-1}; \bar{\mathbf{\Theta}}_{[l,\ldots,L]})||_F^2,$$
$$\text{s.t.} \quad \hat{\mathbf{\Theta}}_{[l,\ldots,L]} \in \mathbf{\Omega}_{[l,\ldots,L]}.$$

Approximation:

$$\begin{aligned}
&||f(\hat{\mathbf{Y}}^{l-1}; \hat{\mathbf{\Theta}}_{[l,\ldots,L]}) - f(\mathbf{Y}^{l-1}; \bar{\mathbf{\Theta}}_{[l,\ldots,L]})||_F^2 \\
&\leq \underbrace{||f(\hat{\mathbf{Y}}^{l-1}; \hat{\mathbf{\Theta}}_{[l,\ldots,L]}) - f(\hat{\mathbf{Y}}^{l-1}; \bar{\mathbf{\Theta}}^{new}_{[l,\ldots,L]})||_F^2}_{\text{Quantization}} \\
&+ \underbrace{||f(\hat{\mathbf{Y}}^{l-1}; \bar{\mathbf{\Theta}}^{new}_{[l,\ldots,L]}) - f(\mathbf{Y}^{l-1}; \bar{\mathbf{\Theta}}_{[l,\ldots,L]})||_F^2}_{\text{Weights Update}},
\end{aligned}$$

Quantization

- Layer output with quantized weights: $\hat{\mathbf{Z}}^l = \hat{\mathbf{Y}}^{l-1} \cdot \hat{\mathbf{\Theta}}_l$
- Layer output with origin weights: $\mathbf{Z}^l_* = \hat{\mathbf{Y}}^{l-1} \cdot \bar{\mathbf{\Theta}}^{new}_l$
- $\delta\mathbf{\Theta}_l = \hat{\mathbf{\Theta}}_l - \bar{\mathbf{\Theta}}^{new}_l$

Error Function:

$$\begin{aligned}
E^l &= E(\hat{\mathbf{Z}}^l) = \frac{1}{n}\left\|\hat{\mathbf{Z}}^l - \mathbf{Z}^l_*\right\|_F^2 \\
&= \left(\frac{\partial E^l}{\partial \mathbf{\Theta}_l}\right)^\top \delta\mathbf{\Theta}_l + \frac{1}{2}\delta\mathbf{\Theta}_l{}^\top \mathbf{H}_l \delta\mathbf{\Theta}_l + O(\|\delta\mathbf{\Theta}_l\|_2^3)
\end{aligned}$$

Optimization Objective:

$$\min_{\hat{\mathbf{\Theta}}_l} f(\hat{\mathbf{\Theta}}_l) = \tfrac{1}{2}(\hat{\mathbf{\Theta}}_l - \bar{\mathbf{\Theta}}^{new}_l)^\top \mathbf{H}_l (\hat{\mathbf{\Theta}}_l - \bar{\mathbf{\Theta}}^{new}_l),$$
$$\text{s.t.} \ \hat{\mathbf{\Theta}}_l \in \mathbf{\Omega}_l,$$

Quantization with ADMM

By introducing an auxiliary parameter **G** to relax $\hat{\mathbf{\Theta}}$ to continuous:

$$\begin{aligned}
&L_\rho(\hat{\mathbf{\Theta}}, \mathbf{G}, \boldsymbol{\lambda}) \\
&= f(\hat{\mathbf{\Theta}}) + I_{\mathbf{\Omega}}(\mathbf{G}) + \frac{\rho}{2}\|\hat{\mathbf{\Theta}} - \mathbf{G} + \boldsymbol{\lambda}\|_2^2 - \frac{\rho}{2}\|\boldsymbol{\lambda}\|_2^2.
\end{aligned}$$

- **Proximal Step** optimize $\hat{\mathbf{\Theta}}$:

$$(\mathbf{H} + \mathrm{diag}(\rho))\hat{\mathbf{\Theta}}^{k+1} = \mathbf{H}\bar{\mathbf{\Theta}}^{new} + \mathrm{diag}(\rho)(\mathbf{G}^k - \boldsymbol{\lambda}^k).$$

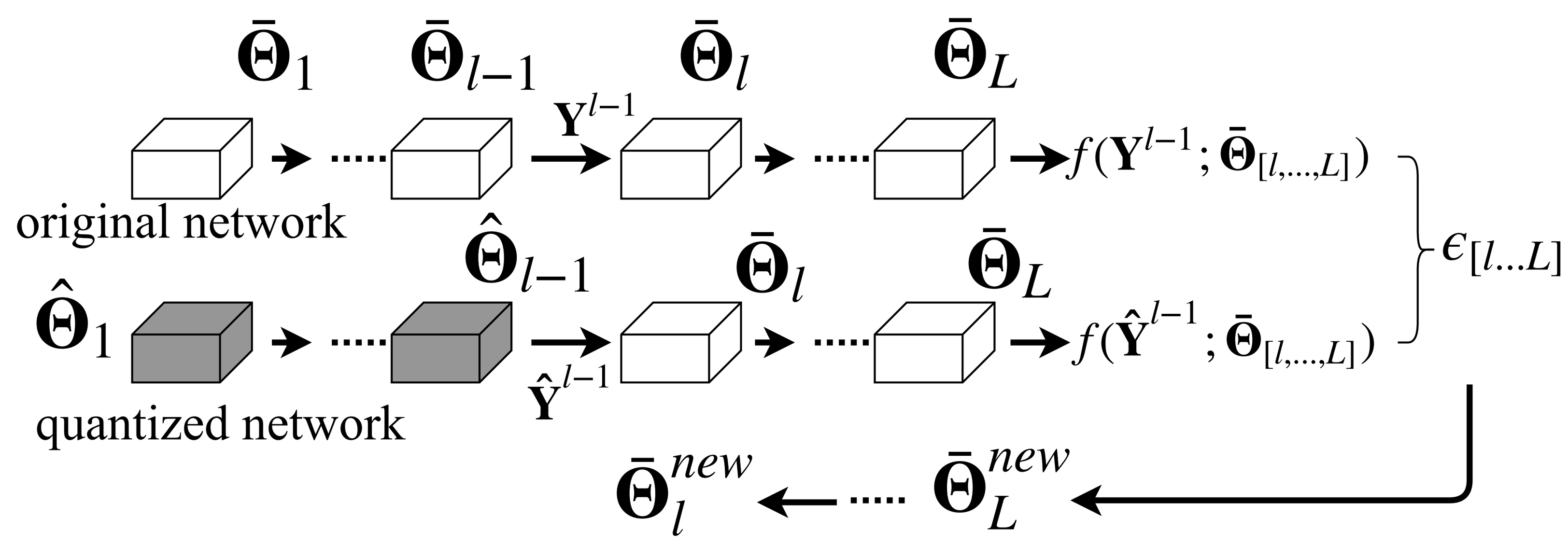
- **Projection Step** optimize **G** :

$$\min_{\mathbf{G}} \|\hat{\mathbf{\Theta}}^{k+1} - \mathbf{G} + \boldsymbol{\lambda}^k\|_2^2, \ \text{s.t.} \ \mathbf{G} \in \mathbf{\Omega}.$$

- **Dual Update Step** optimize $\boldsymbol{\lambda}$:

$$\boldsymbol{\lambda}^{k+1} = \boldsymbol{\lambda}^k + \hat{\mathbf{\Theta}}^{k+1} - \mathbf{G}^{k+1}.$$

Remaining Non-quantized Weights Update



Experiments

- Using only 1% of CIFAR10 and ImageNet dataset:

Dataset	Network	Method	bits	Improve(%)	Full-Precision
CIFAR10	ResNet20	TTQ	3	-77.25	91.77
		INQ	15	-48.48	90.02
		ExNN	3	-11.15	91.5
		VQ	3	-11.27	
		DQ	3	-19.92	
		L-DNQ	3	**-4.30**	
ImageNet	ResNet18	TTQ	3	-69.48/-88.49	69.6/89.2
		INQ	15	-61.27/-64.22	68.27/88.69
		ExNN	3	-43.53/-37.82	69.76/89.02
		VQ	3	-35.69/-29.08	
		DQ	3	-61.22/-65.64	
		L-DNQ	3	**-16.43/-10.67**	
		DQ	8	-56.78/-58.92	
			16	-13.81/-8.68	
			32	-2.82/-1.51	
		L-DNQ	9	**-2.73/-0.90**	

- Performance as number of instances climbs: