BSQ: Exploring Bit-Level Sparsity for Mixed-Precision Neural Network Quantization

Huanrui Yang, Lin Duan, Yiran Chen and Hai Li
Duke University

ICLR 2021

The Need of Mixed-Precision Fixed-point Quantization

- Less bits, less memory consumption, less energy cost
  - Add: **30x** less energy; Mult: **18.5x** less energy (8-bit fixed-point vs. 32-bit float)
- Some layers are more important -> mixed-precision quantization
- **Key problem: find optimal MP quantization scheme**
- MP quantization introduced a large and discrete design space: exponential to #layers
  - Search with NAS
  - Rank layers/filters with saliency and manually decide precision

- **Goal: get MP quantization schemes directly with differentiable regularizer, efficiently explore performance-model size tradeoff**

A Bit-level View of Quantization

- For a fixed-point quantized matrix, when can its precision be reduced?
  - MSB=0 for all elements: precision can reduce directly

$$\begin{bmatrix} 6 \\ 3 \end{bmatrix} \equiv \begin{bmatrix} 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 \end{bmatrix}_2 \equiv \begin{bmatrix} 1 & 1 & 0 \\ 0 & 1 & 1 \end{bmatrix}_2$$

  - LSB=0 for all elements: precision can be reduced with scaling factor 2

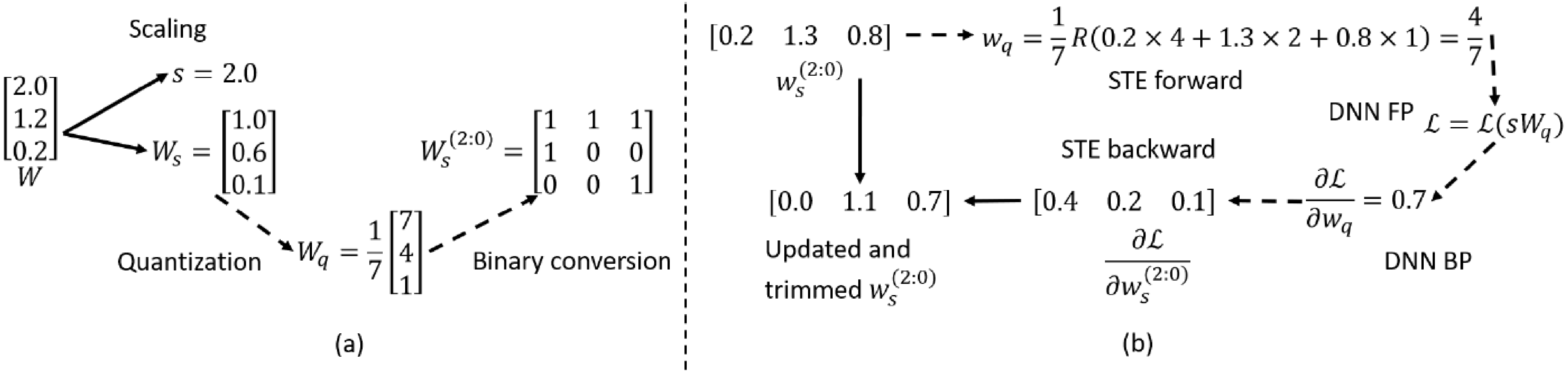
$$\begin{bmatrix} 10 \\ 4 \end{bmatrix} \equiv \begin{bmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix}_2 \equiv 2 \times \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix}_2$$

- MP quantization scheme can be explored by inducing **structural bit-level sparsity ->** We name this method **BSQ**

DNN Training under Bit Representation

Bit representation conversion

FP/BP loop: each bit as **float** trainable variable



STE formulation

$$\textbf{Forward: } W_q = \frac{1}{2^n - 1} Round\left[\sum_{b=0}^{n-1} W_s^{(b)} 2^b\right]; \quad \textbf{Backward: } \frac{\partial \mathcal{L}}{\partial W_s^{(b)}} = \frac{2^b}{2^n - 1} \frac{\partial \mathcal{L}}{\partial W_q}$$

BSQ Training Pipeline

- Start with 8-bit quantized model
  - Most models can keep original accuracy under 8-bit quantization
- Bit-level group LASSO

$$B_{GL}(W^g) = \sum_{b=0}^{n-1} \left\| \left[W_p^{(b)}; W_n^{(b)} \right] \right\|_2$$

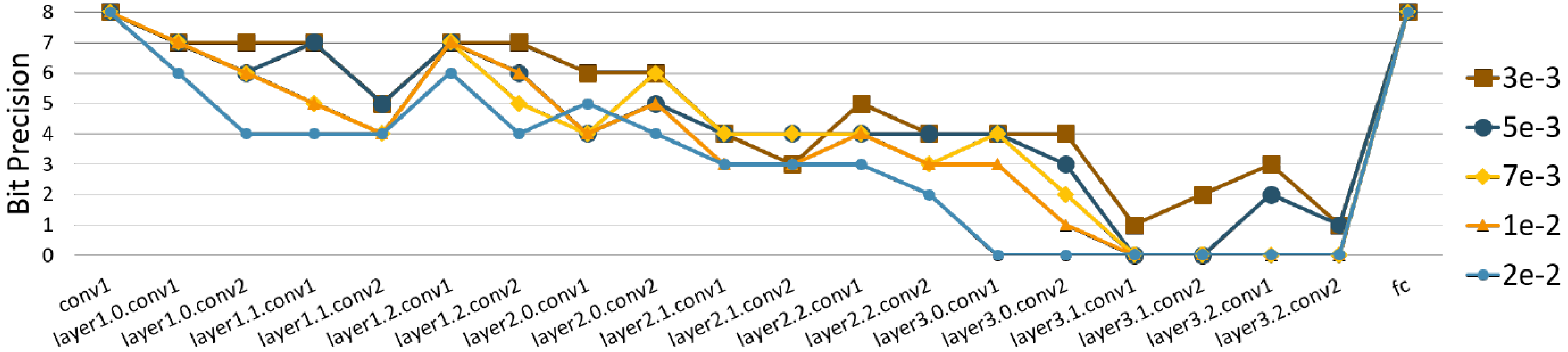
- Overall training objective

$$\mathcal{L} = \mathcal{L}_{CE}(W_q^{(1:L)}) + \alpha \sum_{l=1}^{L} \frac{\#Para(W^l) \times \#Bit(W^l)}{\#Para(W^{(1:L)})} B_{GL}(W^l)$$

- Apply periodic **re-quantization** and **precision adjustment** throughout the training process
- Finalize MP scheme and finetune quantized weight

Accuracy-#Bits Tradeoff

- BSQ achieves further compression with a larger regularization strength



- Better than training with the same quantization scheme from scratch

Strength α	3e-3	5e-3	7e-3	1e-2	2e-2
#Bits per Para / Comp ($\times$)	3.02 / 10.60	2.25 / 14.24	1.66 / 19.24	1.37 / 23.44	0.87 / 36.63
BSQ acc before / after FT (%)	91.30 / 92.60	90.98 / 92.32	90.42 / 91.48	90.35 / 91.16	85.77 / 89.49
Train from scratch acc (%)	91.72	91.45	91.12	89.57	89.14

Comparing with SOTA Methods

Larger compression rate under similar accuracy

Table 2: Quantization results of ResNet-20 models on the CIFAR-10 dataset. BSQ is compared with DoReFa-Net (Zhou et al., 2016), PACT (Choi et al., 2018), LQ-Net (Zhang et al., 2018), DNAS (Wu et al., 2019) and HAWQ (Dong et al., 2019). "MP" denotes mixed-precision quantization.

Benchmarks					BSQ		
Act. Prec.	Method	Weight Prec.	Comp ($\times$)	Acc (%)	α	Comp ($\times$)	Acc (%)
32-bit	Baseline	32	1.00	92.62			
	LQ-Nets	3	10.67	92.00	5e-3	14.24	92.77
	DNAS	MP	11.60	92.72	7e-3	19.24	91.87
	LQ-Nets	2	16.00	91.80			
4-bit	HAWQ	MP	13.11	92.22	5e-3	14.24	92.32
3-bit	LQ-Nets	3	10.67	91.60	2e-3	11.04	92.16
	PACT	3	10.67	91.10	5e-3	16.37	91.72
	DoReFa	3	10.67	89.90			
2-bit	LQ-Nets	2	16.00	90.20			
	PACT	2	16.00	89.70	5e-3	18.85	90.19
	DoReFa	2	16.00	88.20			

More results available in our paper

Thanks

More details can be found at:

Paper: https://openreview.net/pdf?id=TiXl51SCNw8

Code: https://github.com/yanghr/BSQ

This work was supported in part by NSF CCF-1910299 and NSF CNS-1822085.