清華大學

SEAM: Searching Transferable Mixed-Precision Quantization Policy through Large Margin Regularization

Chen Tang, Kai Ouyang, Zenghao Chai, Yunpeng Bai, Yuan Meng, Zhi Wang, Wenwu Zhu

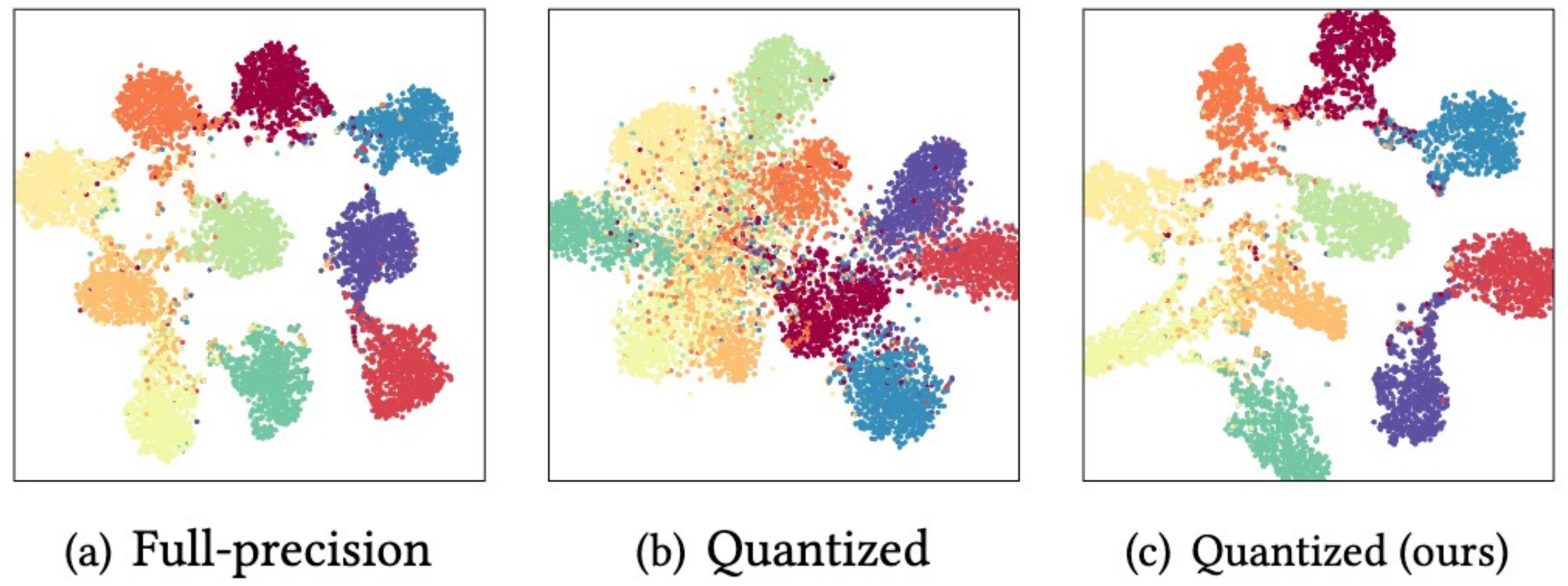
paper

our group

Challenge: Mixed-Precision Quantization Datasets Require Consistency with Model Training

inefficient

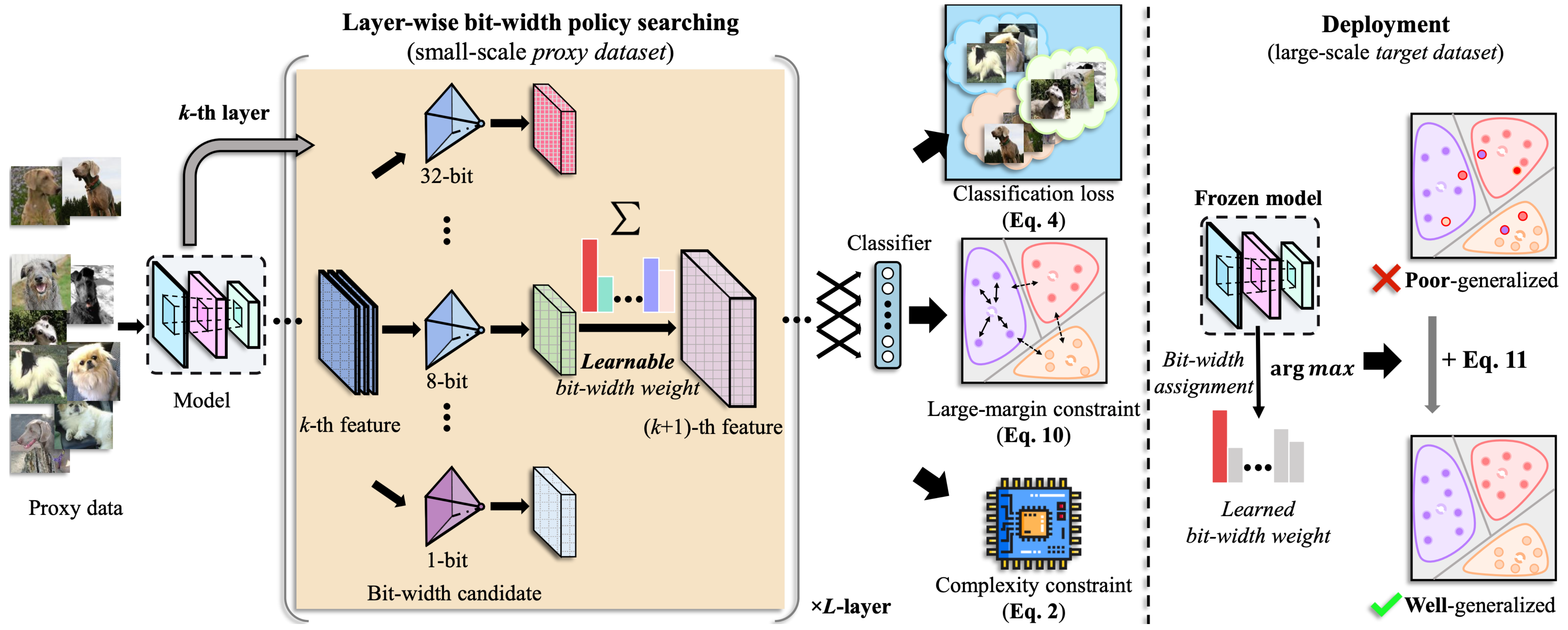
SEAM: decoupling the datasets for MPQ searching efficiency using class-level information


(a) Full-precision

(b) Quantized

(c) Quantized (ours)

Observation:
Improper mixed-precision policy diminishes class margins and produces ambiguous decision boundaries


Layer-wise bit-width policy searching
(small-scale *proxy dataset*)
k-th layer
Proxy data
Model
32-bit
8-bit
1-bit
k-th feature
Σ
Learnable bit-width weight
(*k*+1)-th feature
Bit-width candidate
×*L*-layer
Classifier
Classification loss (Eq. 4)
Large-margin constraint (Eq. 10)
Complexity constraint (Eq. 2)
Deployment
(large-scale *target dataset*)
Frozen model
Bit-width assignment
arg *max*
Learned bit-width weight
Poor-generalized
+ Eq. 11
Well-generalized

Proposed method:
Search on a small proxy dataset by identifying policies that uphold the discriminative nature of feature representations

$$\mathcal{L}_{inc} = \sum_{i=1}^{N} q_i = \sum_{i=1}^{N} d(\boldsymbol{g_i}, \boldsymbol{\mu_{y_i}})$$
$$= \sum_{i=1}^{N} -\log\ p(y_i)\mathcal{N}(\boldsymbol{g_i}; \boldsymbol{\mu_{y_i}}, \boldsymbol{\Sigma_{y_i}}).$$

Eq.4 &Eq. 10:
intra-class compactness

$$\min \mathcal{L}_{cls} = \min_{o} \sum_{i=1}^{N}\sum_{j=1}^{K} o_{i,j},$$

$$o_{i,j} = \begin{cases} -\log \dfrac{\exp(h_j(\boldsymbol{g_i};m))}{\sum_{k=1}^{K} \mathbb{1}(k \neq y_i)\exp(h_k(\boldsymbol{g_i};0)) + \exp(h_j(\boldsymbol{g_i};m))}, & \text{if } j = y_i \\ 0, & \text{otherwise,} \end{cases}$$

Eq.4 &Eq. 10:
inter-class separation

Results: ImageNet classification

Table 2: Accuracy and efficiency results for MobileNetv1. “Top-1/5” represents Top-1 and top-5 accuracy respectively.

Method	W-bits	A-bits	Top-1/5 (%)	BitOPs (G)	Cost (h)
PACT	4	4	62.4 / 82.2	9.68	-
LSQ	3	3	68.3 / 88.1	5.8	-
HMQ	3MP	4MP	69.3 / -	-	-
FracBits	3MP	3MP	68.7 / 88.2	5.78	237.2
LIMPQ	3MP	3MP	69.5 / 89.1	5.78	3.4
Ours-C	3MP	3MP	**69.9 / 89.3**	6.28	1.0
Ours-S	3MP	3MP	69.6 / 89.2	6.13	0.8

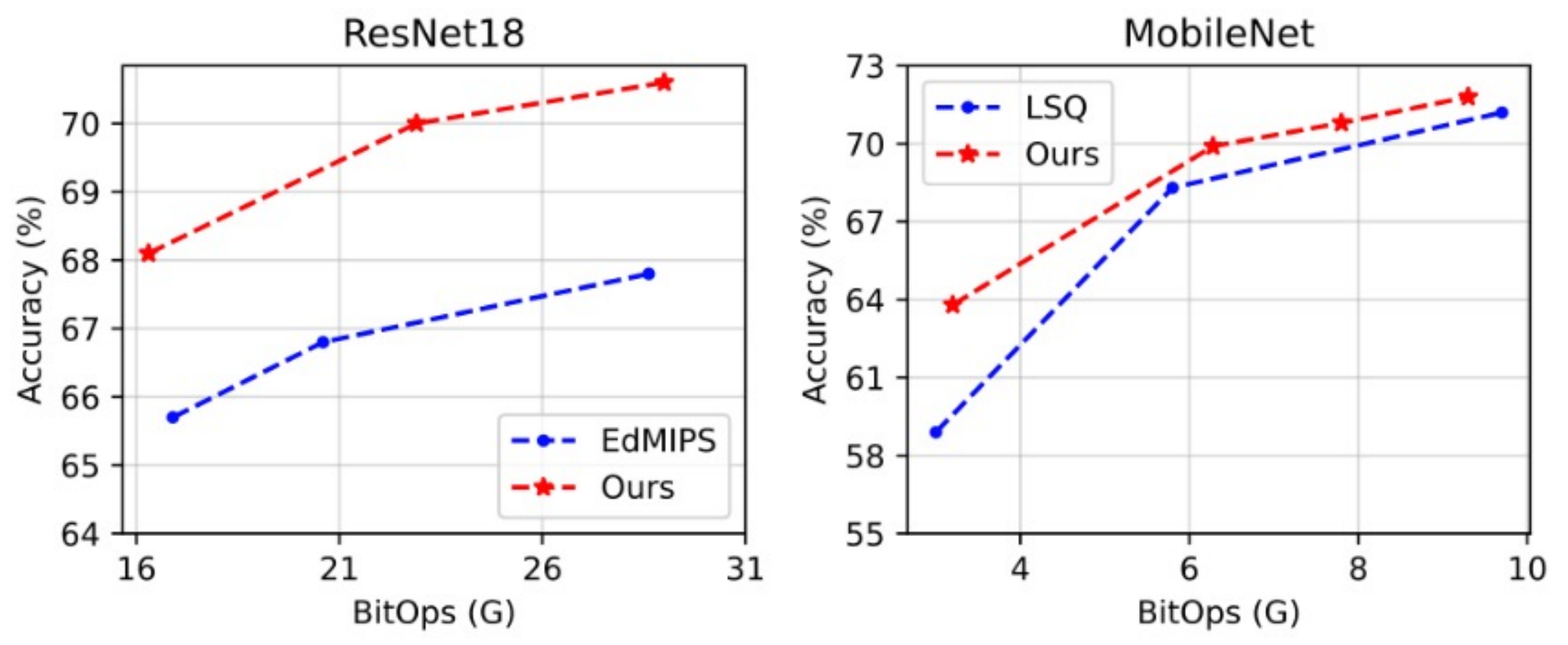
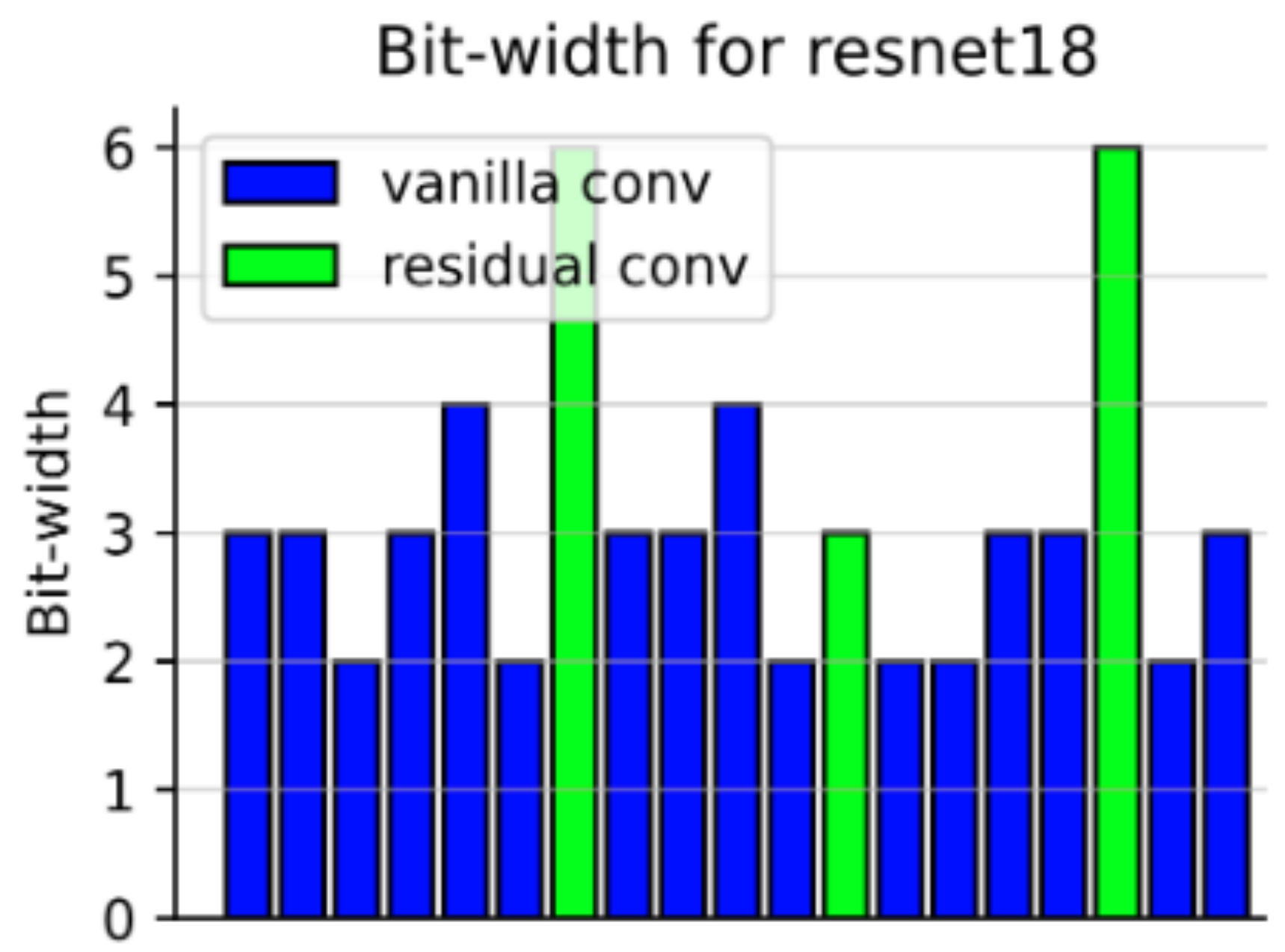
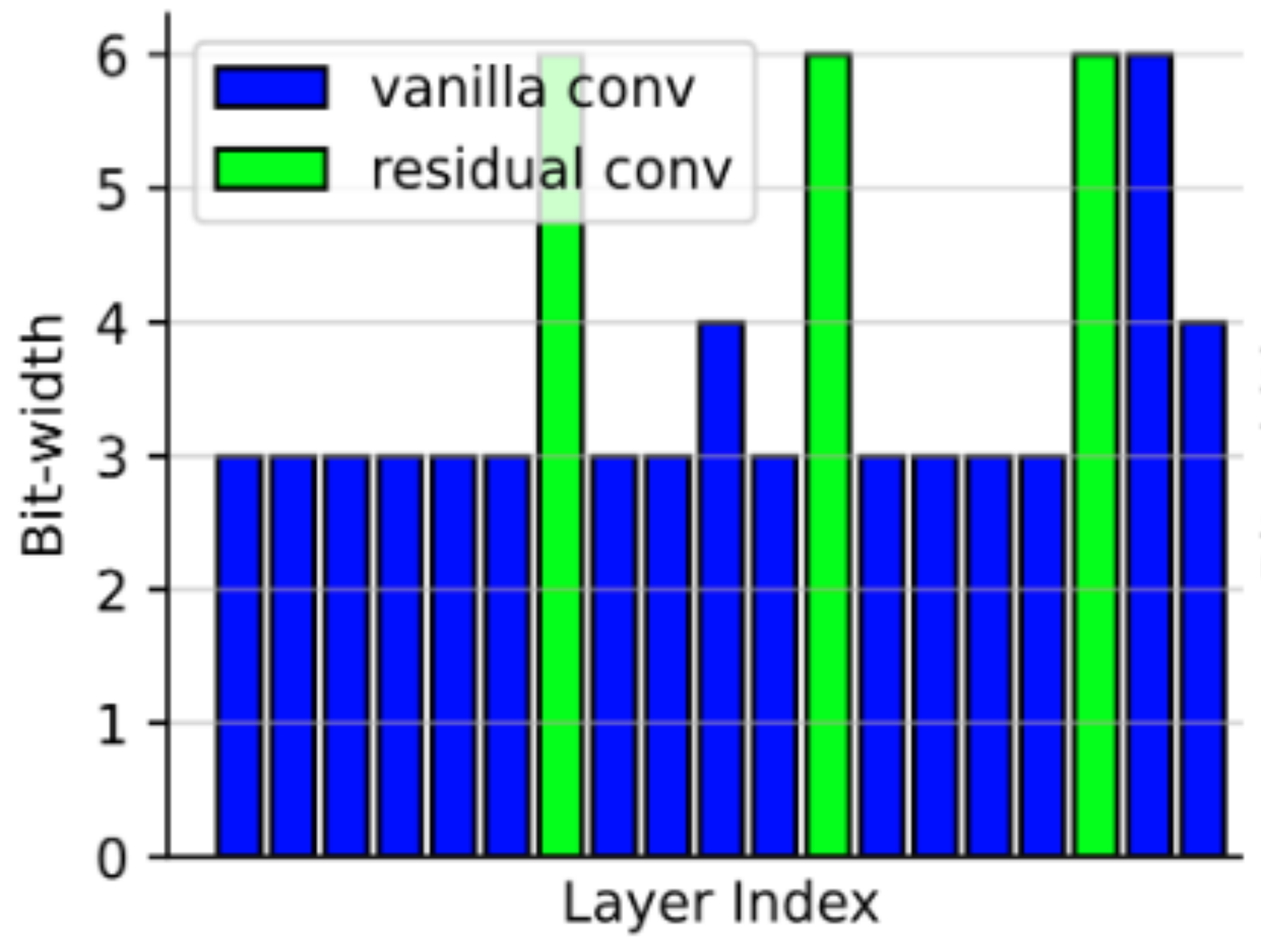


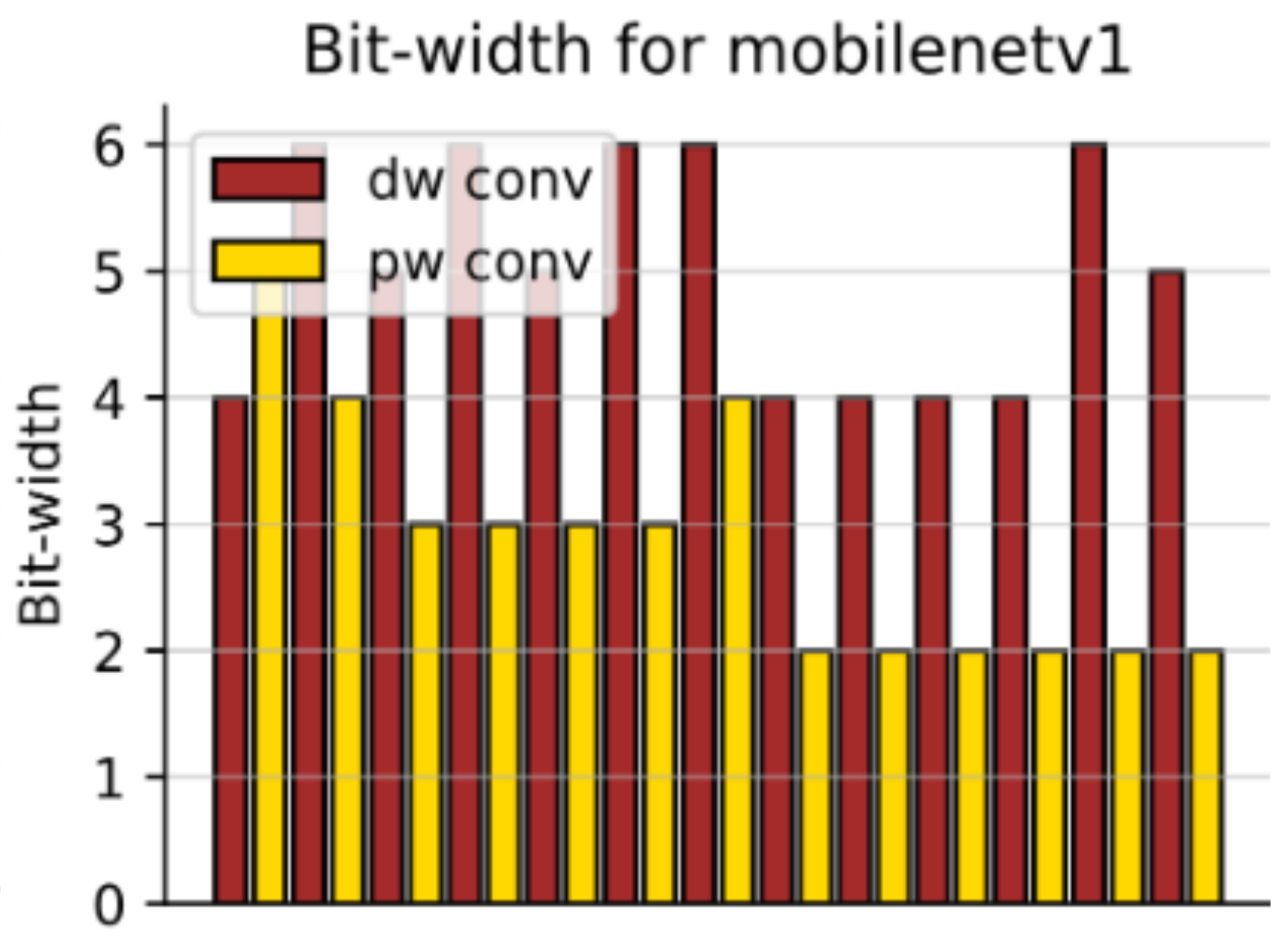
Figure 3: Complexity-accuracy trade-off for ResNet18 and MobileNet.



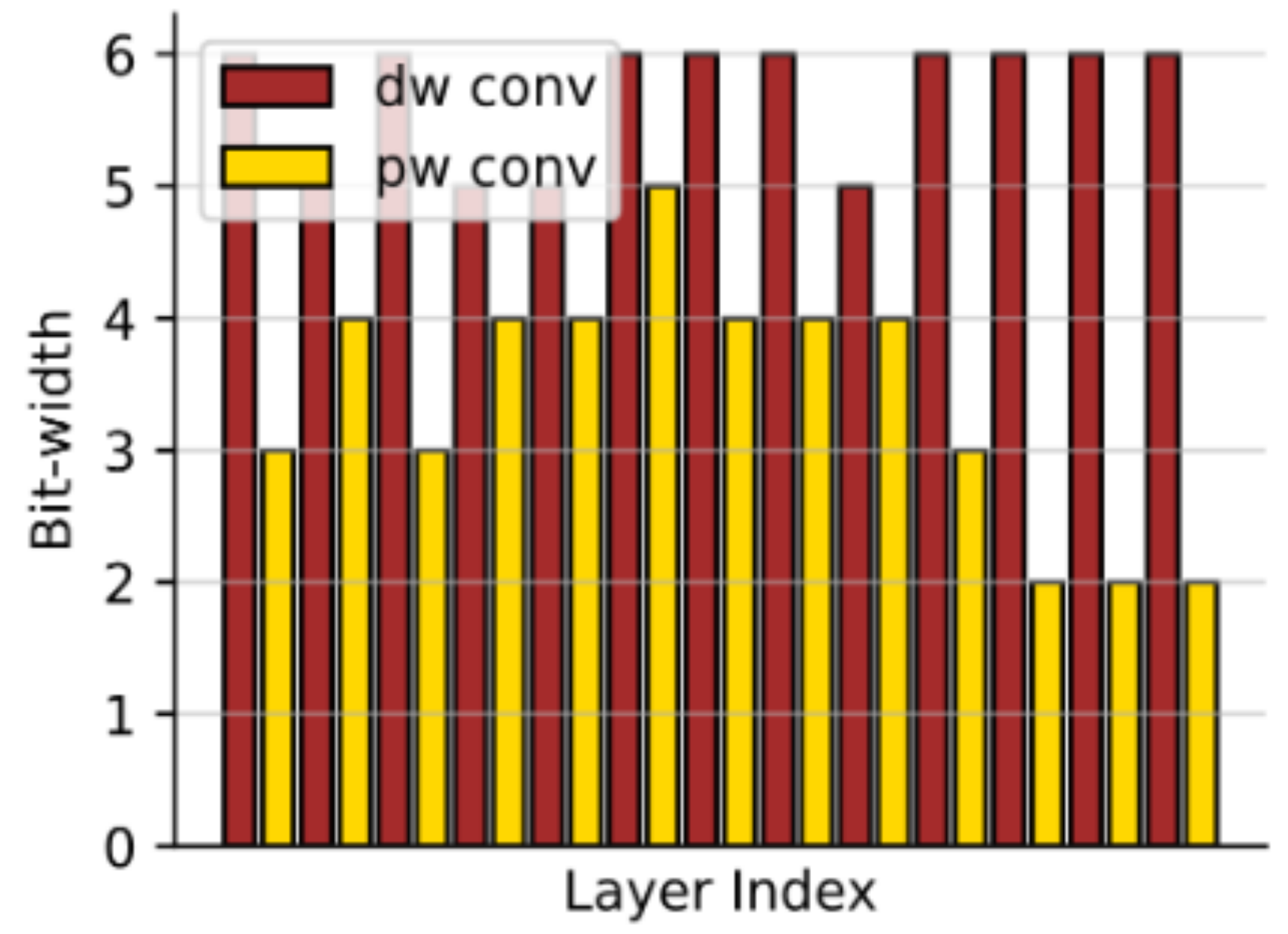
weights



activations



weights



activations