tinyML® Research Symposium

Enabling Ultra-low Power Machine Learning at the Edge

March 27, 2023

TINY ML

www.tinyML.org

Automatic Network Adaptation for Ultra-Low Uniform-Precision Quantization

TinyML Research Symposium '2023

Seongmin Park, Beomseok Kwon, Jieun Lim, Kyuyoung Sim, Tae-Ho Kim and Jungwook Choi

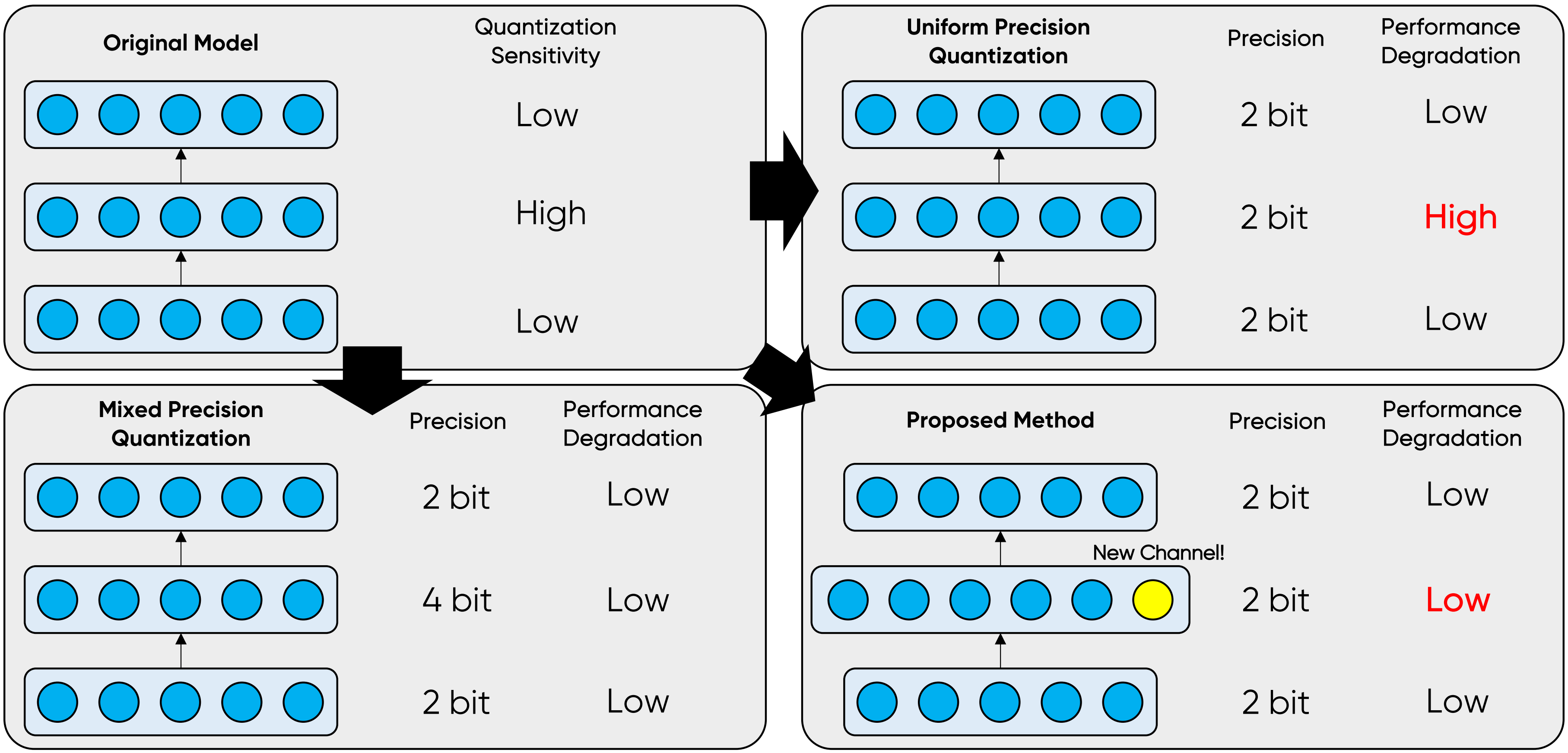
Presenter: Myungsu Chae, CEO, Nota AI

Summary

- Uniform-precision neural network quantization has gained popularity since it simplifies densely packed arithmetic unit for high computing capability.
- However, it ignores heterogeneous sensitivity to the impact of quantization errors across the layers, resulting in sub-optimal inference accuracy.
- This work proposes a novel neural architecture search called **neural channel expansion that adjusts the network structure to alleviate accuracy degradation from ultra-low uniform-precision quantization**.
- The proposed method selectively expands channels for the quantization sensitive layers while satisfying hardware constraints (e.g., FLOPs, PARAMs).
- We demonstrate that the proposed method can adapt several popular networks' channels to achieve superior 2-bit quantization accuracy on CIFAR10 and ImageNet.
- In particular, we achieve the best-to-date Top-1/Top-5 accuracy for 2-bit ResNet50 with smaller FLOPs and the parameter size.

Motivation of the Research

Compensating Performance Degradation of Ultra Low-precision Quantization



Related Work

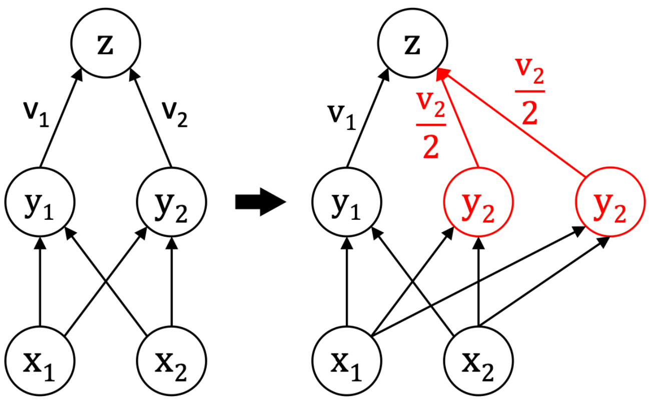
Neural-Net adaptation for accurate DNN quantization

Preliminary Research on Capacity Expansion[1]

Table 4: ResNet-34 top-1 validation accuracy % and compute cost as precision of activations (A) and weights (W) varies.

Width	Precision	Top-1 Acc. %	Compute cost
1x wide	32b A, 32b W	73.59	1x
	1b A, 1b W	60.54	0.03x
2x wide	4b A, 8b W	74.48	0.74x
	4b A, 4b W	74.52	0.50x
	4b A, 2b W	73.58	0.39x
	2b A, 4b W	73.50	0.39x
	2b A, 2b W	73.32	0.27x
	1b A, 1b W	69.85	0.15x
3x wide	1b A, 1b W	72.38	0.30x

Channel Splitting to Reduce Dynamic Range of Weights[2]



<u>However, above methods are not automatic and not plausible on the ultra-low precision.</u>

[1] Asit Mishra, Eriko Nurvitadhi, Jeffrey J Cook, and Debbie Marr. Wrpn: Wide reduced-precision networks. In International Conference on Learning Representations, 2018.
[2] RitchieZhao, YuweiHu, Jordan Dotzel, Chris De Sa, and Zhiru Zhang. Improving neural network quantization without retraining using outlier channel splitting. In International Conference on Machine Learning, pages 7543–7552, 2019.

Neural Channel Expansion

Algorithm 1: Neural Channel Expansion

Input:

Split the training set into two dis-joint sets: D_{weight} and D_{arch} $(n(D_{weight}) = n(D_{arch}))$

Search Parameter: $\{\alpha_1^l, \alpha_2^l, .., \alpha_n^l\} \in A^l$, $\{A^1, A^2, .., A^L\} \subset \mathbb{A}$, L =number of layer

Expand Threshold: T

Warm-up training

For Warm-up Epoch **do**
  Sample batch data D_w from D_{weight} and network from $\mathbb{A} \sim U(0,1)$
  Calculate $Loss_{weight}$ on D_w to update network weights
End for
For Search Epoch **do**

Weights update

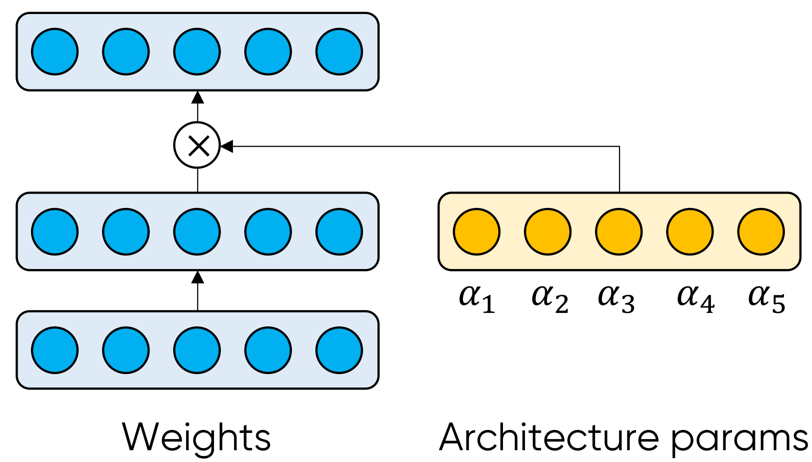
  Sample batch data D_w from D_{weight} and network from $Softmax(\mathbb{A})$
  Calculate $Loss_{weight}$ on D_w to update network weights

Architecture Params update

  Sample batch data D_a from D_{arch} and network from $Softmax(\mathbb{A})$
  Calculate $Loss_{arch}$ on D_a to update $\mathbb{A}$

If the architecture param of max channel exceeds the threshold, expand the search space.

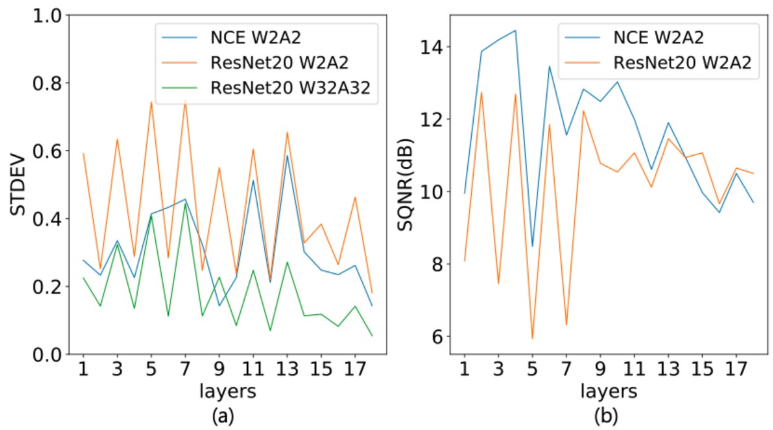
  For layer **do**
    $j \leftarrow \#A^l$
    If $Softmax(\alpha_j^l; \{\alpha_k^l\}_{k \in j}) \geq T$ **do**
      Expand search space(α_{j+1}^l)
      $\alpha_{j+1}^l \leftarrow \alpha_j^l$ # copy search parameter
    End if
  End for
End for
Derive the searched network from $\mathbb{A}$
Randomly initialize the searched network and optimize it on the training set

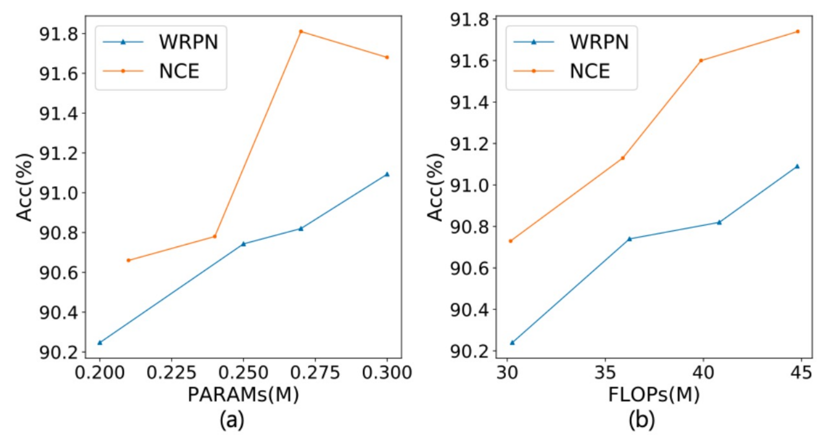


Analysis on the Impact of Channel Expansion to Quantization

Quantization affects to the dynamic range of the weights

- Quantization applied to a given network substantially increases the dynamic range of activation, hindering successful DNN quantization
- Unlike straightforward capacity to all layers as described in WRPN, we selectively expand layers so we can keep lower PARAMs while preserving accuracy.

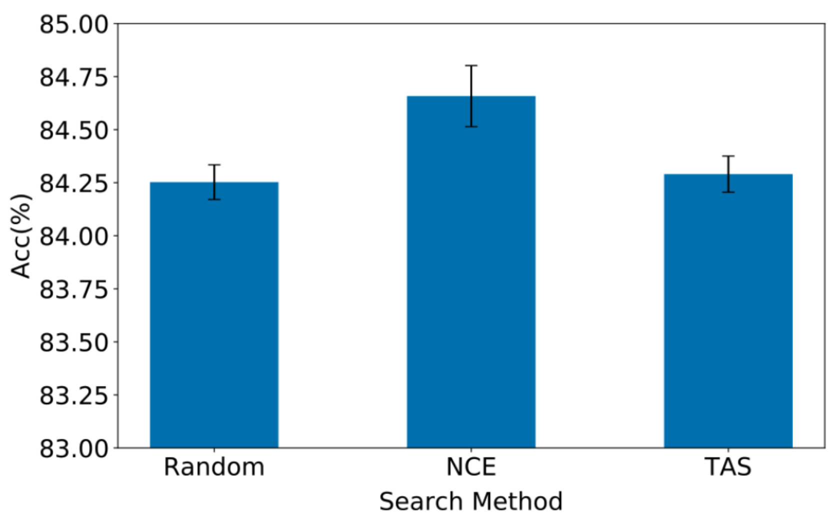




Analysis on Search Space of Neural Channel Expansion

The flexible search space of NCE can lead efficient search compared to TAS

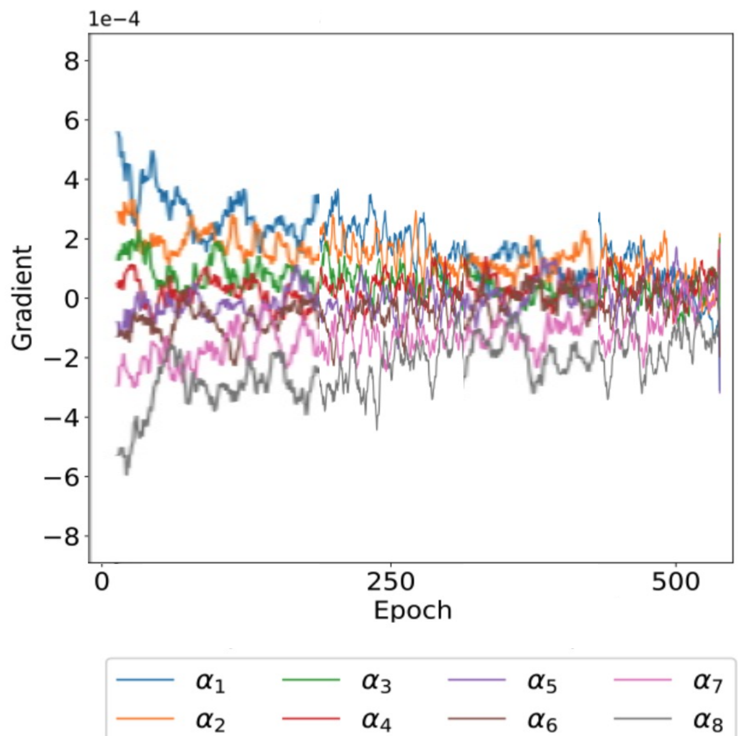
- When we sample the network from the search space randomly, architecture searched from NCE and TAS, then the accuracy of the networks are compared. NCE shows the higher accuracy, it means that search space and the method is superior to others.



Analysis on Channel Selection Preference of NCE

Gradient shows the preference on the larger channel numbers

- Architecture parameters associated to large channel get negative gradient at the early stage of training, but the demand is decreased as training goes.



Experimental Results

Benchmark with other SOTA models

ResNet18				
Method	Top-1 Acc(%)	Top-5 Acc(%)	FLOPs	PARAMs
Full precision	*70.56*	*89.88*	1.814G	11.69M
LSQ	67.6	87.6		
QIL	65.7	-		
PACT	64.4	85.6		
EdMIPS	65.9	86.5		
w/o NCE(Ours)	64.08	86.47		
w/ NCE(Ours)	**66.18**	**86.75**	**1.897G**	**10.94M**

ResNet50				
Method	Top-1 Acc(%)	Top-5 Acc(%)	FLOPs	PARAMs
Full precision	*76.82*	*93.33*	4.089G	25.56M
LSQ	73.7	91.5		
PACT	72.2	90.5		
EdMIPS	72.1	90.6		
w/o NCE(Ours)	72.36	90.81		
w/ NCE(Ours)	**74.03**	**91.63**	**3.932G**	**17.66M**

Impact of Threshold

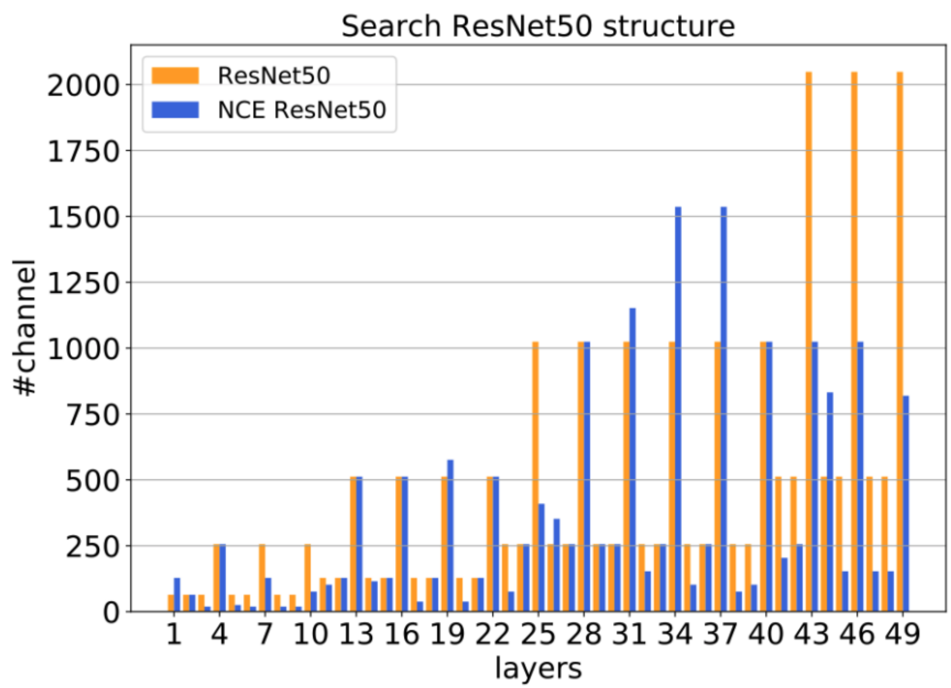
Threshold T	0.30	0.25	0.20	0.15
Accuracy	91.60%	91.44%	91.40%	90.77%

On the various precision

ResNet20-CIFAR10		Accuracy	FLOPs
Original structure (w/o NCE)	W32A32	92.88%	40.81M
	W3A3	92.45%	
	W4A4	92.69%	
NCE	W3A3	92.66%	39.07M
	W4A4	92.75%	37.07M

Analysis on Comparison of Network Structures

NCE balances the reduction and expansion of the ResNet50



Conclusion

- In this work, we propose a novel approach that explores the neural network structure to achieve robust inference accuracy while using the simple uniform-precision arithmetic operations.
- Our novel differentiable neural architecture search, called neural channel expansion, employs the search space that can shrink and expand the channels.
- More sensitive layers can be equipped with more channels while the overall resource requirements (e.g., FLOPs and PARAMs) are maintained.
- We demonstrate that the proposed method can achieve superior performance in ultra-low uniform-precision quantization for CIFAR10 and ImageNet networks.

Nota AI × HANYANG UNIVERSITY 한양 1939

Copyright Notice