Training Multi-bit Quantized and Binarized Networks with A Learnable Symmetric Quantizer

Лукьянов Павел Александрович
МГУ им. М. В. Ломоносова

20 апреля 2021 г.

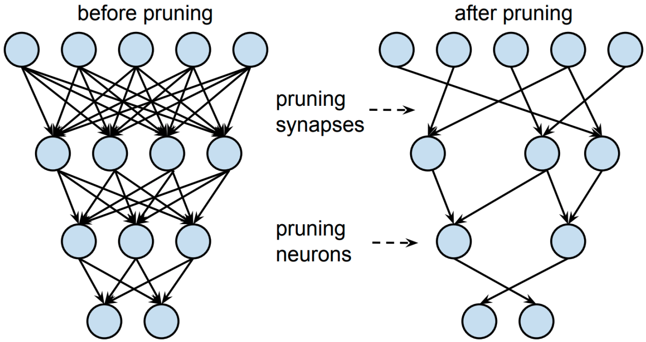
Neural network compression

Problem of the computational complexity and memory size.
Examples of methods for neural network compression:

- Pruning
- Quantization

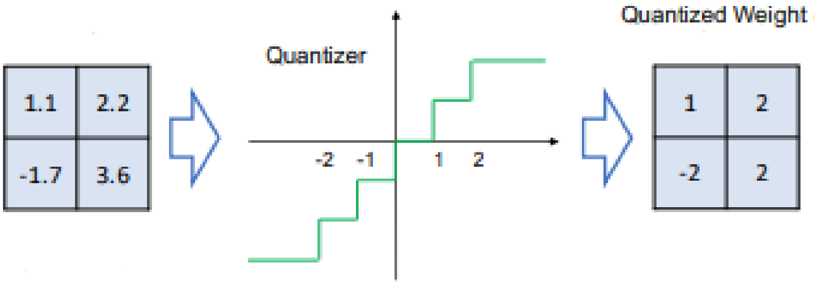
Pruning

Pruning removes redundant parameters or neurons that do not significantly contribute to the accuracy of results.



Quantization

Quantization is known as the process of approximating a continuous signal by a set of discrete symbols or integer values.



Learnable Symmetric Quantizer

The end points of the intervals are referred to as decision levels, the output is called the reconstruction level.
Δ (step size) - the length of the intervals, N - the total number of reconstruction levels.
$clip_N(x) = min(max(x, 0), N-1)$ - clip function

Uniform symmetric quantizer:
$Q_w(x) = [clip_N((x+\alpha)/\Delta)]\Delta - \alpha$, where $\alpha = \Delta(N-1)/2$

Q_w can be rewritten as:
$Q_w(x) = q_w(\Delta/2)$, where $q_w = 2[clip_N((x+\alpha)/\Delta)] - N + 1$;
q_w can be encoded into $[log_2(N)]$ bits using ± 1 encoding.

Learnable Symmetric Quantizer

The authors make the step size Δ a learnable parameter.
How to calculate gradient:

$$\frac{\partial Q_w(x)}{\partial \Delta} = \begin{cases} -\frac{x}{\Delta} + [\frac{x}{\Delta} - 0.5] + 0.5 & |x| < \alpha \\ sign(x)\alpha & \text{otherwise.} \end{cases}$$

Optimal MSE Initialization

Let X be the random variable for a quantizer input and its pdf is denoted by p(x). The optimal step size for Qw is defined in the mean squared error (MSE) sense by
$\Delta_w^* = argmin_\Delta D_w(\Delta)$, where
$D_w(\Delta) = E[(x - Q_w(x, \Delta))^2]$

$\frac{dD_w}{d\Delta} = -\sum_{i=1}^{N/2-1}(2i-1)\int_{(i-1)\Delta}^{i\Delta} 2(x - [\frac{2i-1}{2}]\Delta)p(x)dx$ -

- $(N-1)\int_{(N/2-1)\Delta}^{\infty} 2(x - [\frac{N-1}{2}]\Delta)p(x)dx$

ImageNet

Method	ResNet-18 (FP: 71.57)				ResNet-34 (FP: 75.11)				MobileNet-V2 (FP: 71.53)			
	Bit-width (W/A)											
	4/4	3/3	2/2	1/1	4/4	3/3	2/2	1/1	4/4	3/3	2/2	1/1
PACT [5]	69.2	68.1	64.4	-	-	-	-	-	61.4	-	-	-
DoReFa-Net [55]	68.1	67.5	62.6	-	-	-	-	-	-	-	-	-
DSQ [16]	69.6	68.7	65.2	-	72.8	72.5	70.0	-	-	-	-	-
QIL [14]	70.1	69.2	65.7	-	73.7	73.1	70.6	-	64.8	-	-	-
LSQ [11]	71.1	70.2	67.6	-	74.1	73.4	71.6	-	-	-	-	-
LSQ+ [2]	70.8	69.3	66.8	-	-	-	-		-	-	-	-
SAT [26]	70.3	69.3	65.5	-	-	-	-	-	-	-	-	-
QKD [29]	71.4	70.2	67.4	-	74.6	73.9	71.6	-	67.4	62.6	45.7	-
UniQ (Ours)	**71.5**	**70.5**	**67.8**	**60.5**	**75.0**	**74.2**	**72.1**	**65.8**	**68.2**	**65.0**	**50.5**	**23.2**

Рис. 1: Top-1 accuracy (%) on ImageNet dataset

Binarization

Network	Method	Acc(%)	Original
ResNet-18 (FP: 71.57)	ABC-Net [34]	42.7	
	XNOR-Net [42]	51.2	
	BNN+ [9]	53.0	✓
	DoReFa-Net [55]†	53.4	✓
	Bi-Real [36]	56.4	
	XNOR-Net++ [3]	57.1	
	IR-Net [41]	58.1	✓
	ProxyBNN [20]	58.7	✓
	RBNN [33]	59.9	✓
	BinaryDuo [28]	60.4	
	UniQ (Ours)	**60.5**	✓
ResNet-34 (FP: 75.11)	ABC-Net	52.4	
	Bi-Real	62.2	
	IR-Net	62.9	✓
	RBNN	63.1	✓
	UniQ (Ours)	**65.8**	✓

Рис. 2: Top-1 accuracy comparison to the existing stateof-the-art binarization methods on ImageNet.

Step Size Initialization

Bit-width	Step Size Initialization			
(W/A)	0.1	0.2	LSQ Init	Our Init
2/2	67.1	68.6	68.3	**69.3**
3/3	70.7	70.9	71.0	**71.4**

Рис. 3: Comparison of different methods for step size initialization.

References

- Training Multi-bit Quantized and Binarized Networks with A Learnable Symmetric Quantizer
  Phuoc Pham , Jacob A. Abraham, Jaeyong Chung
  https://arxiv.org/pdf/2104.00210.pdf
- A Survey of Quantization Methods for Efficient Neural Network Inference
  Amir Gholami, Sehoon Kim, Zhen Dong, Zhewei Yao, Michael W. Mahoney, Kurt Keutzer
  https://arxiv.org/pdf/2103.13630.pdf
- Pruning and Quantization for Deep Neural Network Acceleration: A Survey
  Tailin Liang , John Glossner , Lei Wang , Shaobo Shi
  https://arxiv.org/pdf/2101.09671.pdf