NJ176 #

icOPEN 2022
International Conference on Optical and Photonic Engineering
Nanjing, China
Nov. 24-27, 2022

An Energy-Efficient Quantized Inference Framework For Electro-Photonic Computing System

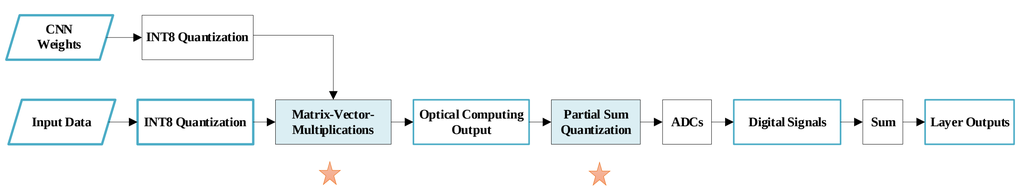
Zeyu Lin and Shenzhen International Graduate school, Tsinghua University
lzy20@mails.tsinghua.edu.cn

Introduction

- Electro-photonic computing system is one of the potential AI accelerators.
- The analog-to-digital converters(ADCs) in electro-photonic computing system have a crucial impact on both **model accuracy** and **energy consumption** of the entire system.
  - High-precision ADCs cause high energy consumption
    - The energy consumption of ADCs accounts for about half of the whole system
    - The energy consumption of ADCs increases exponentially with the precision
  - Low-precision ADCs destroy the optical computing accuracy
    - Using low-precision ADCs directly is equivalent to using the power of two quantization(POTQ), which will cause large quantization error

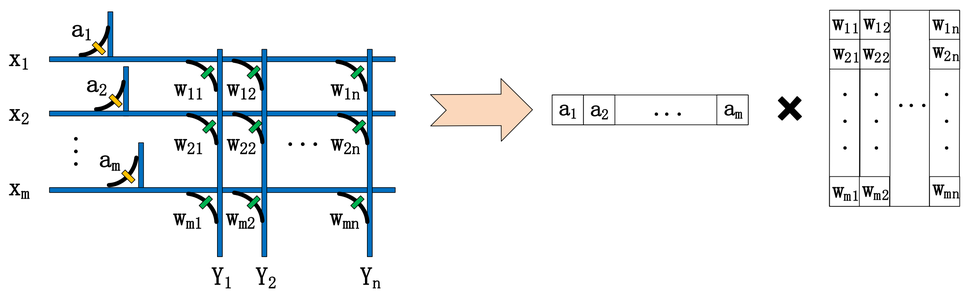
Objective

- An optical convolution operator, which can realize convolution calculation based on basic computing unit (Matrix-Vector-Multiplications) in optical computing system.
- A quantization scheme, which can balance the trade-off between the energy consumption of ADCs and model accuracy.

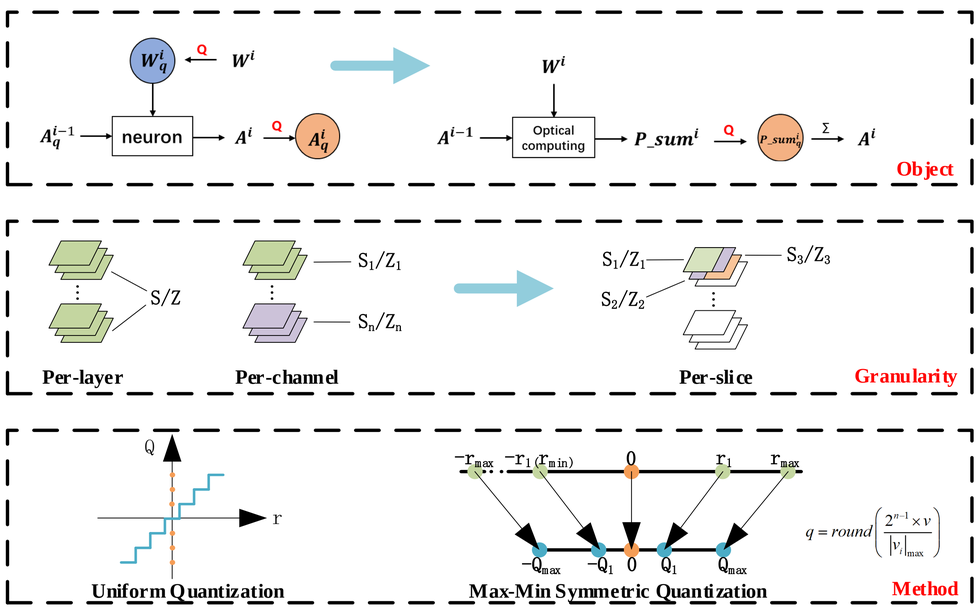


Method

- Optical convolution operator
  - The convolution can be realized by the basic computing unit (Matrix-Vector-Multiplications)



- Per-slice multi-scale quantization scheme
  - Quantization object: partial sum
  - Quantization granularity: Per-slice
  - Quantization method: max-min symmetric uniform quantization



- Quantization error in optical computing
  - The quantization error in optical computing can be divided into two parts:
    - Quantization ontology error, which is caused by the round function in the quantization
    - Quantization cumulative error, which is caused by the accumulating of the partial sum
  - The size of optical vector **K** is a pivotal factor for the quantization error of the partial sum. So selecting an appropriate optical vector size can reduce the quantization error and improve the quantization scheme

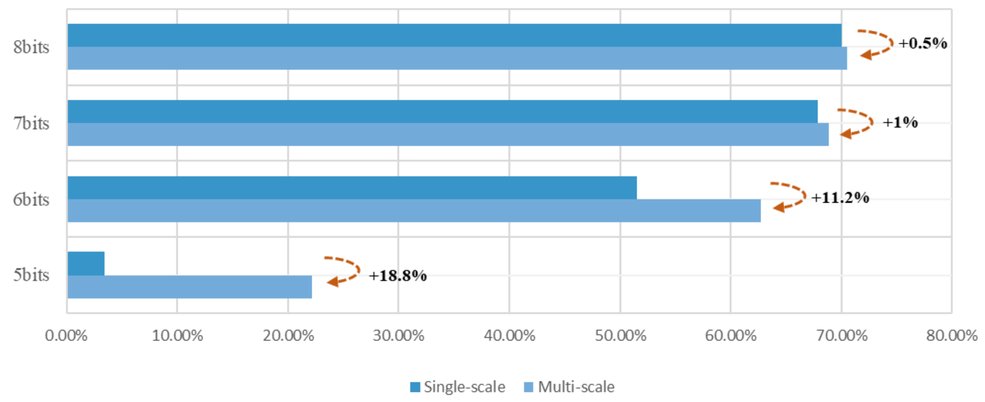
$$E_{p\_sum} = \sum_{i=1}^{\left\lceil \frac{C_{in}\times H_w \times W_w}{K} \right\rceil} \left(\frac{\left| \sum_{j=1}^{K} x_{ij} \times w_{ij} \right|_{\max}}{2^{n-1}} \times \Delta q' \right)$$

Results

- At present, most photonic accelerators directly use low-precision ADCs, which is equivalent to using the power of two quantization(POTQ) leading to a large quantization error. With our quantization scheme, the model accuracy is improved significantly.

Method	MobileNet_v2	ResNet50	Vgg16
POTQ	30%	0.5%	0.2%
Single-scale Quantization	**69.9%**	**68.2%**	**54.8%**
Multi-scale Quantization	**70.5%**	**68.9%**	**61.3%**

- Compared with single-scale method, the multi-scale method performs better, especially in the case of lower bits. We test the experiment on MobileNet-v2.



Conclusion

- This research proposed an energy-efficient quantized inference framework, consisting of **an optical convolution operator** and **a per-slice multi-scale quantization scheme**. What's more, we further improve the quantization scheme by analyzing the quantization error.
- The proposed inference quantization framework can reduce the ADC precision by **7 bits**, while ensuring that the model accuracy loss is less than **1%**, so that we can balance the trade-off between the energy consumption of ADCs and the model accuracy.