



ZeroQ: A Novel Zero Shot Quantization Framework

**Yaohui Cai¹, Zhewei Yao², Zhen Dong²,
Amir Gholami², Michael W. Mahoney², Kurt Keutzer²**

¹Peking University, ²University of California, Berkeley



Motivation

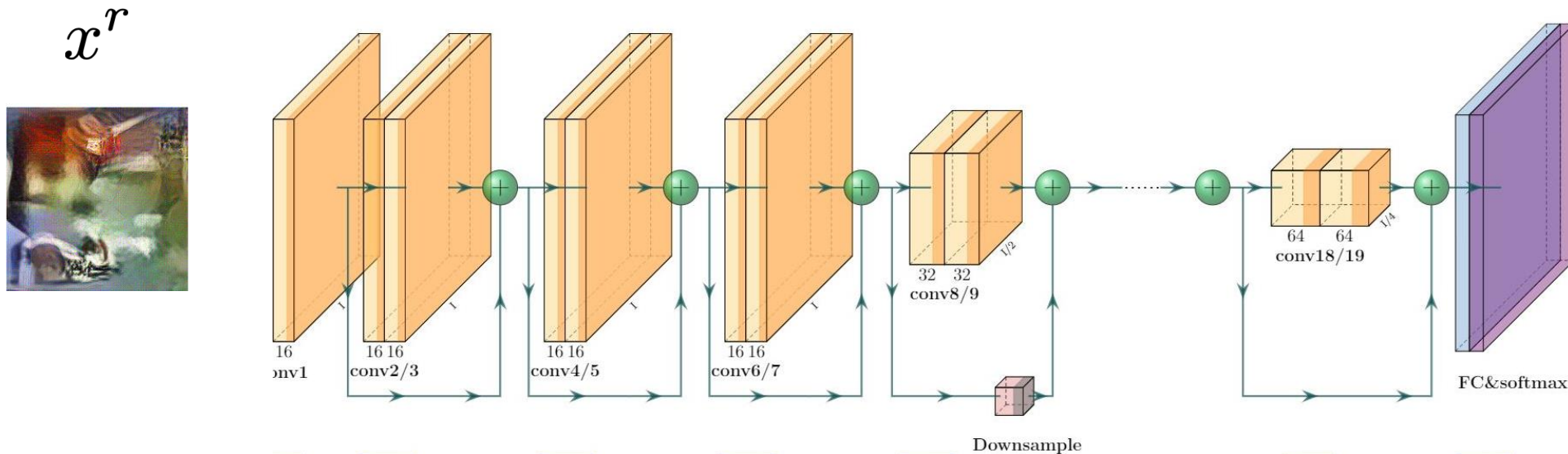
- Quantization is important but has following limitations:
 - Data might not be accessible due to privacy and security concerns
 - Fine-tuning (retraining) can be time-consuming
- We propose **ZeroQ** which addresses the above challenges by incorporating:
 - A novel data distillation method
 - A new fast zeroth-order sensitivity analysis



Distilled Data Computation

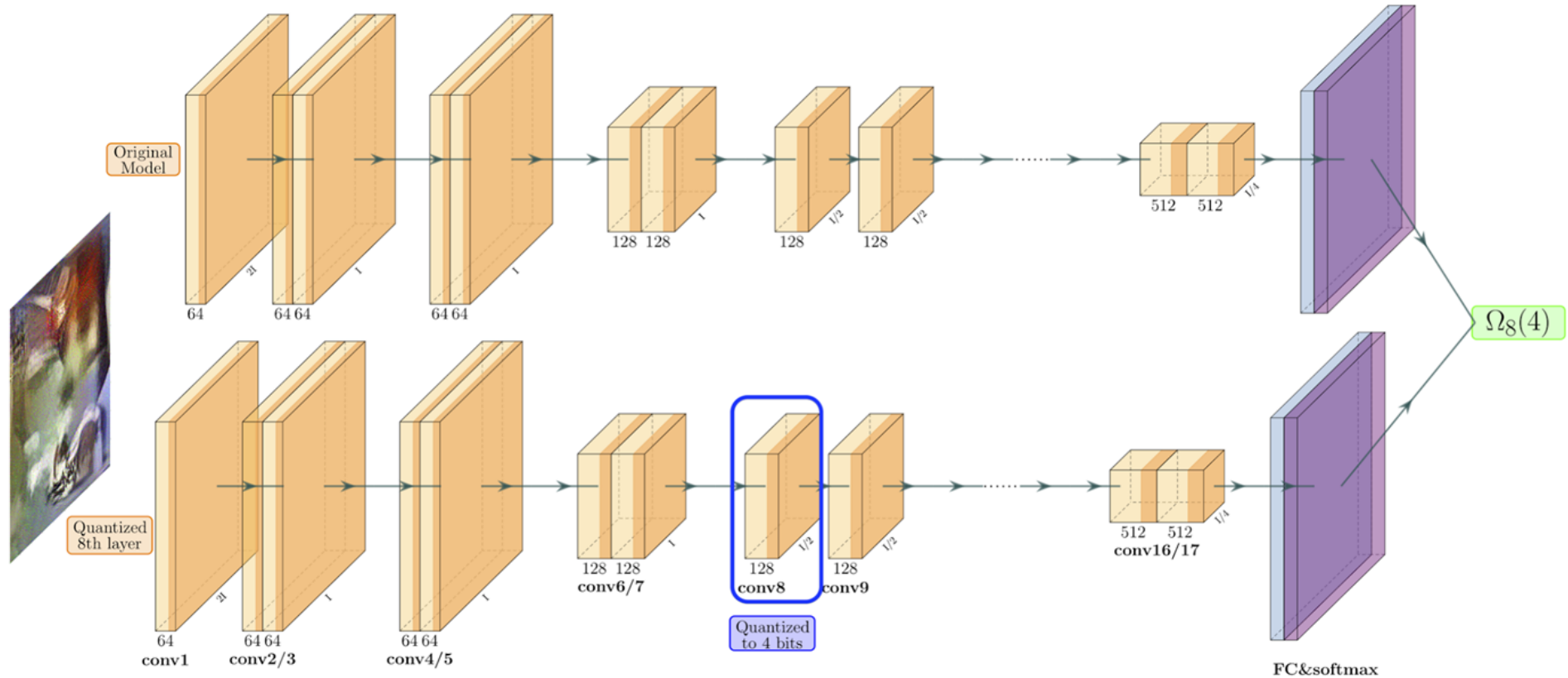
- After solving the optimization problem, we will have a set of distilled data that can be used to determine the range of quantization for each layer, and also perform zeroth-order sensitivity analysis.

$$\min_{x^r} \sum_{i=0}^L \|\tilde{\mu}_i^r - \mu_i\|_2^2 + \|\tilde{\sigma}_i^r - \sigma_i\|_2^2,$$



Zeroth-Order Sensitivity Analysis

- Using the distilled data, we measure the zeroth order sensitivity analysis of each layer to quantization.



Zeroth order sensitivity analysis of conv8 is performed by using distilled data and computing the perturbation with/without quantization of that layer.

Results

- All ZeroQ results are achieved **without any hyper-parameter tuning**.

(a) ResNet50

Method	No D	No FT	W-bit	A-bit	Size (MB)	Top-1
Baseline	–	–	32	32	97.49	77.72
OMSE [19]	✓	✓	4	32	12.28	70.06
OMSE [19]	✗	✓	4	32	12.28	74.98
PACT [5]	✗	✗	4	4	12.19	76.50
ZEROQ	✓	✓	MP	8	12.17	75.80
ZEROQ [†]	✓	✓	MP	8	12.17	76.08
OCS [44]	✗	✓	6	8	18.46	74.80
ZEROQ	✓	✓	MP	6	18.27	77.43
ZEROQ	✓	✓	8	8	24.37	77.67



Results

RetinaNet-ResNet50 Quantization

Method	W-bit	A-bit	Size (MB)	mAP
Baseline	32	32	145.10	36.4
ZEROQ	8	8	36.25	36.4
FQN	4	4	18.13	32.5
ZEROQ	MP	8	18.13	33.7

ZeroQ quantization only takes 2 minutes on 4 V100 GPUs for the end-to-end computations



Conclusions

- We introduced ZeroQ, a novel quantization framework that achieves state-of-the-art quantization results for:
 - Image Classification
 - Object Detection
- ZeroQ significantly outperforms existing methods **without any hyper-parameter tuning**
- ZeroQ achieves high accuracy **without any access to training/testing data** for quantization
- ZeroQ is very efficient and has low computational overhead
 - RetinaNet quantization on MS COCO takes less than 2 minutes on a 4 GPU cluster
- The code for ZeroQ is open sourced and available at:
 - <https://github.com/amirgholami/ZeroQ>

