Zero-shot Quantization: A Comprehensive Survey

Minjun Kim*, Jaehyeon Choi*, Jongkeun Lee, Wonjin Cho, and U Kang†

{minjun.kim, jaehyeon_choi, jklee2, chowonjin0627, ukang}@snu.ac.kr

Seoul National University, Seoul, South Korea

*: Equal Contribution, †: Corresponding Author

IJCAI 2025 MONTREAL

Paper

GitHub

Summary

A Comprehensive Survey on Zero-shot Quantization

- **TL;DR**: We survey Zero-shot Quantization (ZSQ), a data-free model compression paradigm
- **GitHub**: https://github.com/snudm-starlab/ZSQ-Survey

Zero-shot Quantization

Problem. Zero-shot Quantization

- **Input**: a pre-trained model θ on task $\mathcal{T}$ and quantization bits B **(without any real data)**
- **Output**: an accurate B-bit quantization model θ_q

Main Challenges of ZSQ

Challenge 1: Knowledge transfer from the pre-trained model

- How do we transfer knowledge without real data?

Challenge 2: Discrepancy between real and synthetic datasets

- How do we generate real-like synthetic images from a pre-trained model?

Challenge 3: Diversity of the problem settings

- How can we design algorithms that generalize across diverse settings?

Overview

Synthesis-free ZSQ

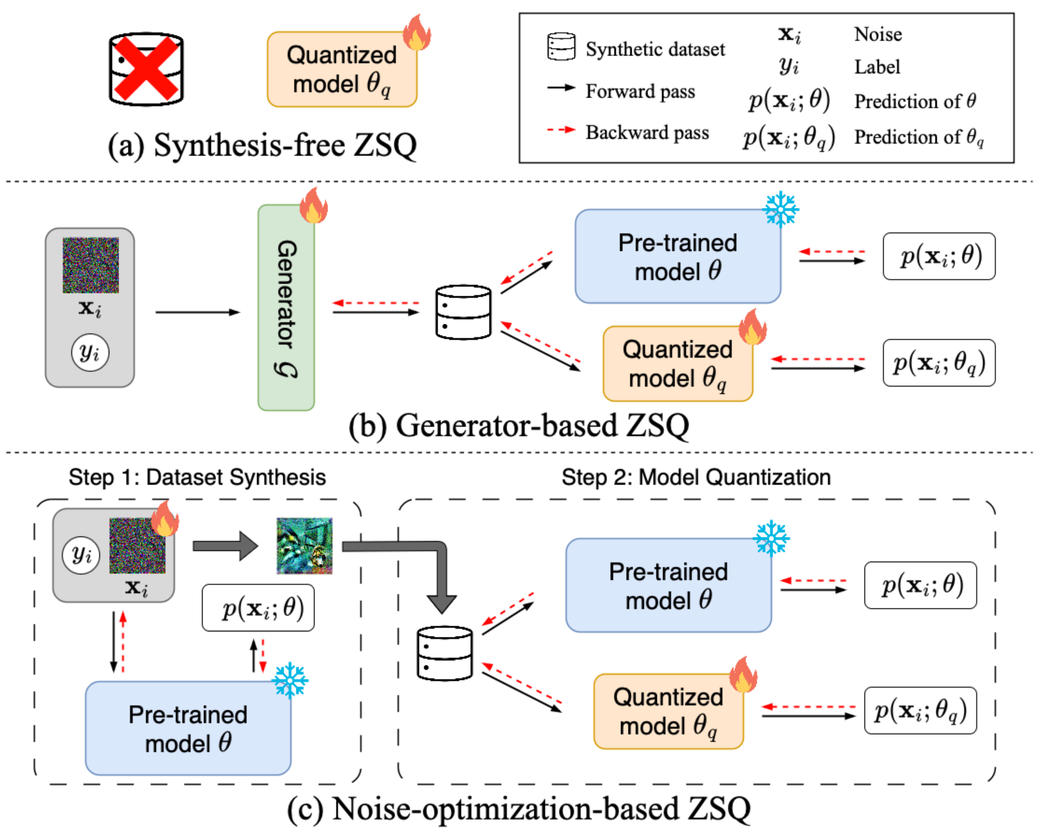
- Quantize models without generating any synthetic data

Generator-based ZSQ

- Train an additional generator $\mathcal{G}$ to produce synthetic data

Noise-optimization-based ZSQ

- Directly optimize noise inputs to make synthetic data



Synthesis-free ZSQ

ZSQ without generating any synthetic data

- Leverage structural properties or theoretical foundations to mitigate performance degradation
- e.g., Diagonal Hessian approximation, Bias correction

Generator-based ZSQ

Employs an independent generator $\mathcal{G}$ to produce datasets

- Generally, train a GAN-based generator from scratch
- Recent works leverage pre-trained diffusion models

Representative method. AdaSG [AAAI 2023]

- Adversarial sample generation: Reformulating ZSQ into a zero-sum game between the generator $\mathcal{G}$ and the quantized model θ_q on reward $\mathcal{R}(\cdot)$

Noise-optimization-based ZSQ

Optimizes noise to generate the dataset from iterative updates

- They universally follow a two-step scheme
- 1. Dataset synthesis ➔ 2. Model quantization

Representative method. HAST [CVPR 2023]

- Hard sample generation: Produce more samples difficult for both original and quantized models to perform better on difficult images

Taxonomy

Method	Training Requirement	Scope of Contribution	Architecture	# Images	Accuracy (FP = 71.47)	
					W4A4	W3A3
DFQ [2019]	PTQ	Q	CNN	0	55.78	-
SQuant [2022]	PTQ	Q	CNN	0	66.14	25.74
UDFC [2023]	PTQ	Q	CNN	0	63.49	-
GDFQ [2020]	QAT	S, Q	CNN	1.28M	60.60	20.23
ZAQ [2021]	QAT	S, Q	CNN	1.28M	52.64	-
ARC [2021]	QAT	S, Q	CNN	1.28M	61.32	23.37
Qimera [2021]	QAT	S, Q	CNN	1.28M	63.84	1.17
ARC + AIT [2022]	QAT	Q	CNN	1.28M	65.73	-
AdaSG [2023b]	QAT	S, Q	CNN	1.28M	66.50	37.04
AdaDFQ [2023a]	QAT	S, Q	CNN	1.28M	66.53	38.10
Causal-DFQ [2023]	QAT	S, Q	CNN	1.28M	68.11	-
RIS [2024]	QAT	S	CNN	1.28M	67.75	-
GenQ [2024b]	PTQ / QAT	S	CNN	1K§	69.77§	-
DeepInversion [2020]	QAT	S	CNN	32	70.27*	64.28†
IntraQ [2022]	QAT	S, Q	CNN	5.12K	66.47	45.51
HAST [2023]	QAT	S, Q	CNN	5.12K	66.91	51.15
TexQ [2023]	QAT	S, Q	CNN	5.12K	67.73	50.28
PLF [2024]	QAT	Q	CNN	5.12K	67.02	-
SynQ [2025b]	QAT	Q	CNN / ViT	5.12K	67.90	52.02
ZeroQ [2020]	PTQ	S, Q	CNN	1K	26.04	-
KW [2020]	PTQ	S, Q	CNN	1K	69.08	-
DSG [2021]	PTQ	S	CNN	1K	34.53	-
MixMix [2021b]	PTQ / QAT	S	CNN	1K§	69.46§	-
PSAQ-ViT [2022]	PTQ	S	ViT	32	71.56*	65.57†
Genie [2023b]	PTQ	S, Q	CNN	1K	69.66	66.89
SADAG [2024]	PTQ	S, Q	CNN	1K	69.72	67.10
SMI [2024]	PTQ	S	ViT	32	70.13*	64.04†
CLAMP-ViT [2024]	PTQ	S, Q	ViT	32	72.17*	69.93†

- Scope of Contribution: S (data synthesis), Q (network quantization)
- Accuracy: ZSQ accuracy [%] of ResNet-18 pre-trained on ImageNet

Future Research Directions

More principled analysis on synthetic datasets

- Most studies fix individual features instead of their root causes
- Deeper analysis may yield fundamental improvements

Broader application to various tasks and domains

- Most research sets task $\mathcal{T}$ as image classification, with a few work on object detection ➔ Extending research to various tasks is crucial

Theoretical exploration of ZSQ

- ZSQ lacks formal understanding such as error bounds
- Mathematical principles would guide towards robust algorithms

Faster generation of synthetic datasets

- Increasing the size of synthetic datasets enhances the performance of quantized models ➔ How can we reduce the generation time?

Combining other model compression techniques

- Current ZSQ algorithms achieve competitive results in 4-bit regime, but struggle in 3-bit or lower-bits ➔ integrating with other methods may achieve a higher compression rate while maintaining accuracy

Evaluating practical impact on real-world scenarios

- The importance of ZSQ lies in its applications for handling real-world scenarios with limited data
- However, current ZSQ methods present experimental results solely on benchmark datasets and models

Diverse problem settings

- Extending ZSQ to real-time quantization and edge-device deployments (e.g. few-instance quantization or pre-trained diffusion models)