SynQ: Accurate Zero-shot Quantization by Synthesis-aware Fine-tuning

Minjun Kim
minjun.kim@snu.ac.kr

Jongjin Kim
j2kim@snu.ac.kr

U Kang
ukang@snu.ac.kr

Data Mining Laboratory
Seoul National University

ICLR

Paper GitHub

Summary

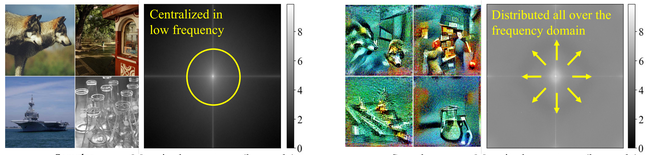
SynQ (**Syn**thesis-aware Fine-tuning for Zero-shot **Q**uantization)

- **TL;DR**: We clearly illustrate and address the three major challenges in Zero-shot Quantization (ZSQ)
- **Github**: https://github.com/snudm-starlab/SynQ

Challenges

Challenge 1. Noise in the Synthetic Dataset

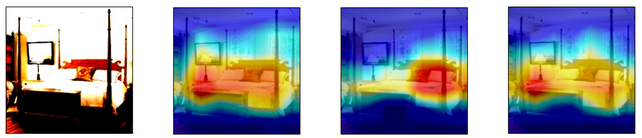
- Synthetic samples contain **high-frequency noise** unlike real images that concentrate on low frequencies



(a) ImageNet dataset (b) Synthetic dataset (TexQ)

Challenge 2. Predictions based on Off-target Patterns

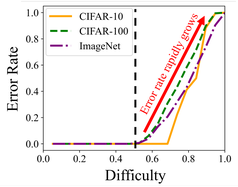
- Quantized models rely on **incorrect image patterns** for predictions



(a) Input (b) ResNet-18 (c) TexQ (d) **SynQ (Ours)**

Challenge 3. Misguidance by Erroneous Hard Labels

- Erroneous hard labels of **difficult samples** lead the quantized model into inaccuracy



Zero-shot Quantization

Problem. Zero-shot Quantization

- **Input**: a pre-trained model M and quantization bits B **(without any real data)**
- **Output:** an accurate B-bit quantized model M^q

Objective Function

$$\min_{\theta_q} \frac{1}{N}\sum_{i=1}^{N} KL\left(q(\mathbf{x}_i;\theta)||q(\mathbf{x}_i;\theta_q)\right) + \lambda_{CE} CE(q(\mathbf{x}_i;\theta_q), y_i)$$

Typical Solution: Two-step Scheme

Step 1: Dataset Synthesis

- Generate a synthetic dataset that resembles the original training dataset

Step 2: Model Quantization

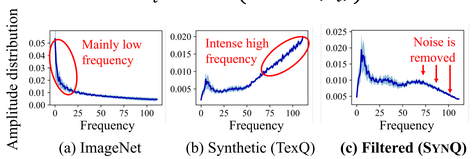
- **Step 2-1**. Quantization Phase (by RTN)
- **Step 2-2**. Fine-tuning Phase

Proposed Method

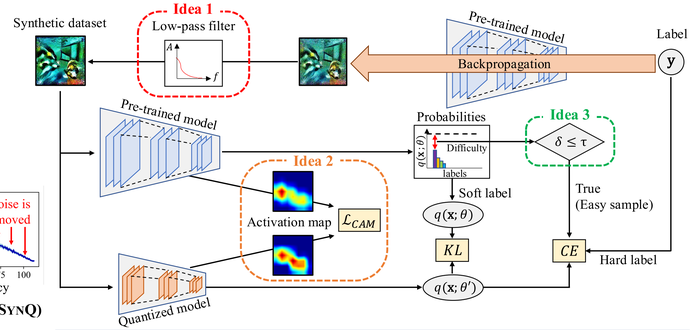
Idea 1. Low-pass Filter

- Exploit a **Gaussian low-pass filter**, performed in the frequency domain through Fourier transforms $\mathcal{F}$ and $\mathcal{F}^{-1}$

$$\mathbf{x}_i^F = \mathcal{F}^{-1}(\mathbf{G} \odot \mathcal{F}(\mathbf{x}_i))$$



(a) ImageNet (b) Synthetic (TexQ) **(c) Filtered (SynQ)**



Idea 2. Alignment of Class Activation Map (CAM)

- **Align the CAMs** of pre-trained and quantized models
- Saliency map $\mathbf{S}^{\theta}(\cdot)$ from Grad-CAM [ICCV '17]

$$\mathcal{L}_{CAM}(\mathbf{x}_i;\theta,\theta^q) = \left\|\mathbf{S}^{\theta}(\mathbf{x}_i) - \mathbf{S}^{\theta^q}(\mathbf{x}_i)\right\|_F^2$$

Idea 3. Soft Labels for Difficult Samples

- Exclude the cross-entropy loss from hard labels for difficult samples (classified based on threshold τ)
- Difficulty $\delta(\mathbf{x}_i,\theta) = 1 - q_{y_i}(\mathbf{x}_i;\theta)$

Objective Function

$$\mathcal{L}_{\text{SynQ}} = \frac{1}{N}\sum_{i=1}^{N} KL\left(q(\mathbf{x}_i^F;\theta)||q(\mathbf{x}_i^F;\theta_q)\right) + \lambda_{CAM}\mathcal{L}_{CAM}(\mathbf{x}_i^F;\theta,\theta_q) + \mathbf{1}_{\{\delta(\mathbf{x}_i^F,\theta)\le\tau\}}\lambda_{CE} CE(q(\mathbf{x}_i^F;\theta_q), y_i)$$

Experiments

1. SynQ achieves the state-of-the-art performance

- Regardless of model type, quantization bits, dataset, and quantization settings (PTQ / QAT)

2. SynQ is compatible with any ZSQ methods

- By all main ideas targeting **Step 2-2**, Fine-tuning Phase

* CNN Quantization

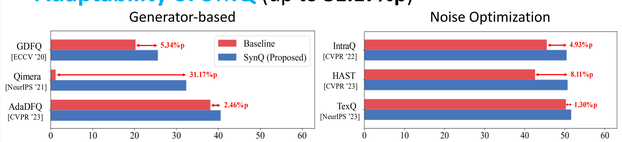
- Quantization-aware Training (QAT) setting (up to **1.74%p**)

Method	R-20 (CIFAR-10)		R-20 (CIFAR-100)		R-18 (ImageNet)		R-50 (ImageNet)		MV2 (ImageNet)	
	W4A4	W3A3	W4A4	W3A3	W4A4	W3A3	W4A4	W3A3	W4A4	W3A3
Full Precision (W32A32)	93.89		70.33		71.47		77.73		73.03	
AdaDFQ (Qian et al., 2023a)	92.31	84.89	66.81	52.74	66.53	38.10	68.38	17.63	65.41	28.99
HAST (Li et al., 2023a)	92.36	86.34	66.68	55.67	66.91	42.58	-	-	65.60	-
TexQ (Chen et al., 2023)	92.68	86.47	67.18	55.87	67.73	50.28	70.72	25.27	67.07	32.80
PLF (Fan et al., 2024)	92.47	88.04	66.94	57.03	67.02	-	68.97	-	-	-
SynQ (Proposed)	**92.76**	**88.11**	**67.34**	**57.28**	**67.90**	**52.02**	**71.05**	**26.89**	**67.27**	**34.21**

- Post-training Quantization (PTQ) setting (**0.64%p** in average)

Method	W2A2	W2A4	W3A3	W4A4	Average
Genie (Jeon et al., 2023b)	54.01	65.10	66.84	69.66	63.90
+ SynQ (Proposed)	**54.97 ± 0.35**	**65.88 ± 0.27**	**67.42 ± 0.21**	**69.88 ± 0.19**	**64.54**

* Adaptability of SynQ (up to **31.17%p**)



* ViT Quantization (up to **0.58%p** in average)

Bits	Method	DeiT-Tiny	DeiT-Small	Swin-Tiny	Swin-Small	Average
	Full Precision	72.21	79.85	81.35	83.20	79.15
W4A8	PSAQ-ViT (Li et al., 2022)	65.57 ± 0.10	72.04 ± 0.19	69.78 ± 1.67	75.03 ± 0.63	70.61
	SynQ (Proposed)	**65.90 ± 0.07**	**72.28 ± 0.34**	**70.76 ± 1.61**	**75.82 ± 0.54**	**71.19**
W8A8	PSAQ-ViT (Li et al., 2022)	71.56 ± 0.03	75.97 ± 0.20	73.54 ± 1.61	76.68 ± 0.53	74.44
	SynQ (Proposed)	**71.74 ± 0.03**	**76.16 ± 0.29**	**74.11 ± 1.82**	**77.32 ± 0.59**	**74.83**