Patch Similarity Aware Data-Free Quantization for Vision Transformers

Zhikai Li[1,2] , Liping Ma[1] , Mengjuan Chen[1] , Junrui Xiao[1,2] , and Qingyi Gu[1*]

[1] *Institute of Automation, Chinese Academy of Sciences*

[2] *School of Artificial Intelligence, University of Chinese Academy of Sciences*

Code: https://github.com/zkkli/PSAQ-ViT

ECCV TEL AVIV 2022

1. Background

- **Motivation:** Data-free quantization is a potential and practice scheme to address **data privacy and security** issues. However, existing BN regularization-based methods **are only designed for CNNs and inapplicable to ViTs**.
- **Insight:** Since there is no elegant ***absolute value metric*** like BN statistics, we intend to investigate the **general difference** in model inference when the input is Gaussian noise and a real image, and then accordingly design a ***relative value metric*** to optimize the noise.

2. Contributions

- From an in-depth analysis of the self-attention module, we reveal a **general difference** in its processing of Gaussian noise and real images, **patch similarity**, which provides some insights for sample generation.
- With the above insights, we propose PSAQ-ViT, where we **reduce the general difference** to optimize the Gaussian noise to approximate the real images and then utilize them to calibrate the quantization parameters. **To the best of our knowledge, this is the first work to quantify ViTs without access to any real-world data.**

3. Methodology

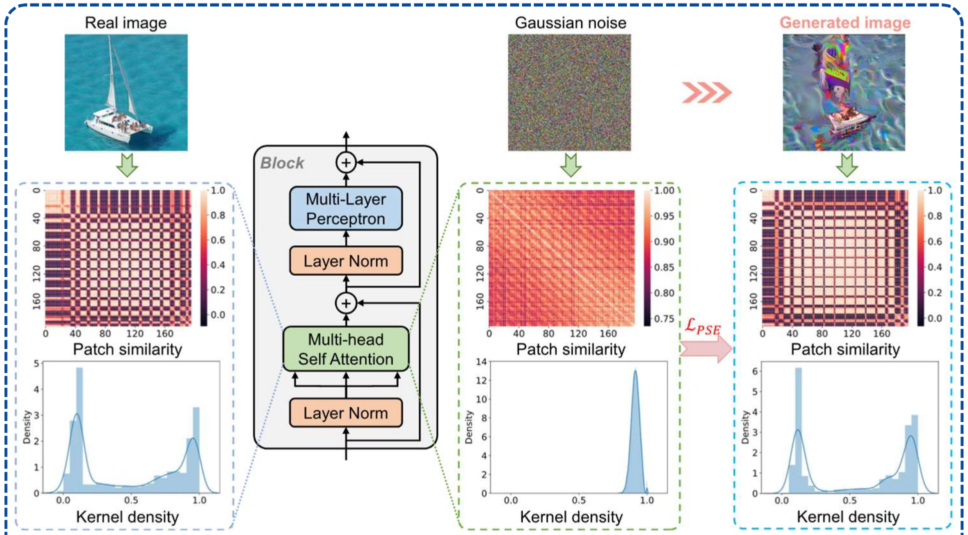

Fig. 1. Illustration of the proposed sample generation approach. Our generated image can potentially represent the real-image features, producing diverse patch similarity and a bimodal kernel density curve, where the left and right peaks describe inter- and intra-category similarity, respectively.

Algorithm 1: The PSAQ-ViT Pipeline

Input: A pre-trained FP ViT P with parameters θ^p.
Output: A quantized ViT Q with parameters θ^q.
Initialize the quantized model Q by Eq. (2);
Randomly produce Gaussian noise $I_G \sim \mathcal{N}(0, 1)$;
\# **Stage 1: Sample generation**
for $t = 1, 2, \ldots$ **do**
  Input I_G into the pre-trained FP model P;
  Calculate $\mathcal{L}_{PSE}$ by Eq. (6);
  Calculate $\mathcal{L}_{OH}$ and $\mathcal{L}_{TV}$ by Eq. (7) and Eq. (8);
  Combine three losses to obtain $\mathcal{L}_G$ by Eq. (9);
  Update I_G by back-propagation of $\mathcal{L}_G$;
end
\# **Stage 2: Quantization parameter calibration**
Get the generated samples $I = I_G$;
Input I into the quantized model Q;
Determine the clipping values of the activations in Q;

Calculation of $\mathcal{L}_{PSE}$

1. Cosine similarity

$$\Gamma_l(u_i, u_j) = \frac{u_i \cdot u_j}{\| u_i \| \| u_j \|}$$

2. Kernel density

$$\hat{f}_h(x) = \frac{1}{M} \sum_{m=1}^{M} K_h(x - x_m)$$

3. Differential entropy

$$H_l = -\int \hat{f}_h(x) \cdot \log[\hat{f}_h(x)] dx$$

4. Summation

$$\mathcal{L}_{PSE} = -\sum_{l=1}^{L} H_l$$

4. Experimental Results

- PSAQ-ViT consistently achieves superior results on various models, even **better than the real-data-driven Standard**.

Model	Method	No Data	Prec.	Top-1(%)	Prec.	Top-1(%)
ViT-S (81.39)	Standard	×	W4/A8	19.91	W8/A8	30.28
	Gaussian noise	✓	W4/A8	15.60	W8/A8	25.22
	PSAQ-ViT(ours)	✓	W4/A8	**20.84**	W8/A8	**31.45**
ViT-B (84.53)	Standard	×	W4/A8	24.76	W8/A8	36.65
	Gaussian noise	✓	W4/A8	19.45	W8/A8	31.63
	PSAQ-ViT(ours)	✓	W4/A8	**25.34**	W8/A8	**37.36**
DeiT-T (72.21)	Standard	×	W4/A8	65.20	W8/A8	71.27
	Gaussian noise	✓	W4/A8	7.80	W8/A8	10.55
	PSAQ-ViT(ours)	✓	W4/A8	**65.57**	W8/A8	**71.56**
DeiT-S (79.85)	Standard	×	W4/A8	72.10	W8/A8	76.00
	Gaussian noise	✓	W4/A8	13.30	W8/A8	18.16
	PSAQ-ViT(ours)	✓	W4/A8	**73.23**	W8/A8	**76.92**
DeiT-B (81.85)	Standard	×	W4/A8	76.25	W8/A8	78.61
	Gaussian noise	✓	W4/A8	11.09	W8/A8	14.72
	PSAQ-ViT(ours)	✓	W4/A8	**77.05**	W8/A8	**79.10**
Swin-T (81.35)	Standard	×	W4/A8	70.16	W8/A8	74.22
	Gaussian noise	✓	W4/A8	0.52	W8/A8	0.62
	PSAQ-ViT(ours)	✓	W4/A8	**71.79**	W8/A8	**75.35**
Swin-S (83.20)	Standard	×	W4/A8	73.33	W8/A8	75.19
	Gaussian noise	✓	W4/A8	5.43	W8/A8	5.66
	PSAQ-ViT(ours)	✓	W4/A8	**75.14**	W8/A8	**76.64**

Table 1. Quantization results on ImageNet dataset.

Fig. 2. Generated class-conditional samples (224×224 pixels), given only a pre-trained ViT-B model.

5. Future Works

- More **accurate** (8-bit lossless compression) and **general** (detection and segmentation applications), see [1] for further details.

[1] Li zhikai, et al. PSAQ-ViT V2: Towards Accurate and General Data-Free Quantization for Vision Transformers. *arXiv preprint arXiv:2209.05687* (2022).

Acknowledgements: This work was supported by the Scientific Instrument Developing Project of the Chinese Academy of Sciences under Grant YJKYYQ20200045.