中国科学院自动化研究所
INSTITUTE OF AUTOMATION
CHINESE ACADEMY OF SCIENCES

中国科学院大学
University of Chinese Academy of Sciences

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

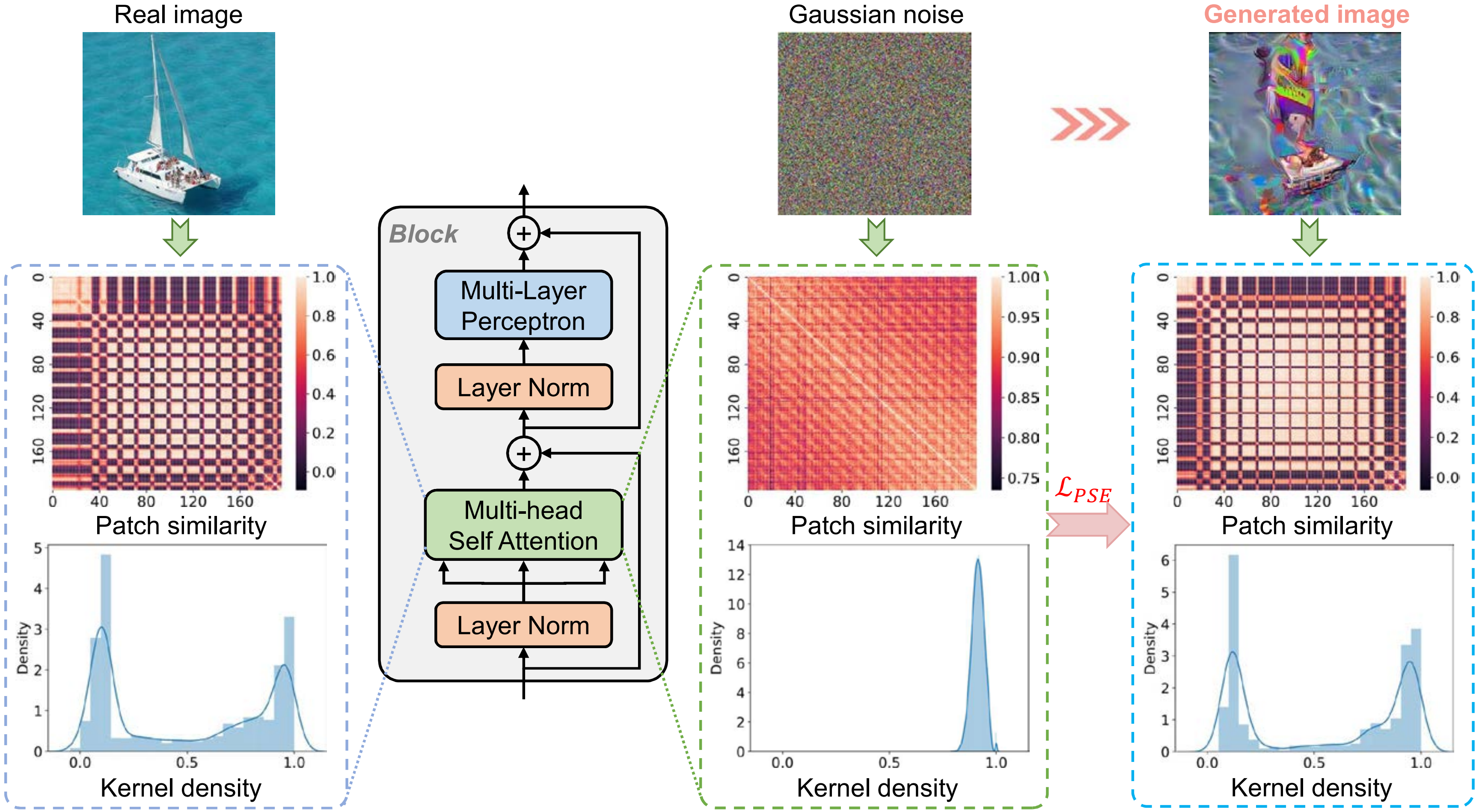
Motivation

- Quantization is important but suffers from **data privacy and security issues.**
  - ➢ **Data-free quantization** is a potential and practice solution.
  - ➢ However, existing BN regularization-based methods are **only designed for CNNs and inapplicable to ViTs.**

- In this paper, we perform **a first attempt** at data-free quantization for ViTs, with the following contributions:
  - ➢ A **general difference** in self-attention module's processing of Gaussian noise and real images, *i.e.*, **patch similarity**.
  - ➢ A **relative value metric** to optimize the Gaussian image to approximate the real images.

Patch Similarity Aware Sample Generation

- 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.



Patch Similarity Metric Calculation

1. Calculate the cosine similarity between each subspace vector in the patch dimension, specifying the data range at [-1, 1], as follows:

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

2. Calculate the continuous probability density function of Γ_l using kernel density estimation as follows:

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

3. Calculate the differential entropy to measure the diversity of patch similarity as follows:

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

4. Sum the differential entropy of each layer to account for the diversity of patch similarity across all layers as follows:

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

Experimental Results

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

Table 1. Quantization results on ImageNet dataset.

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**

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

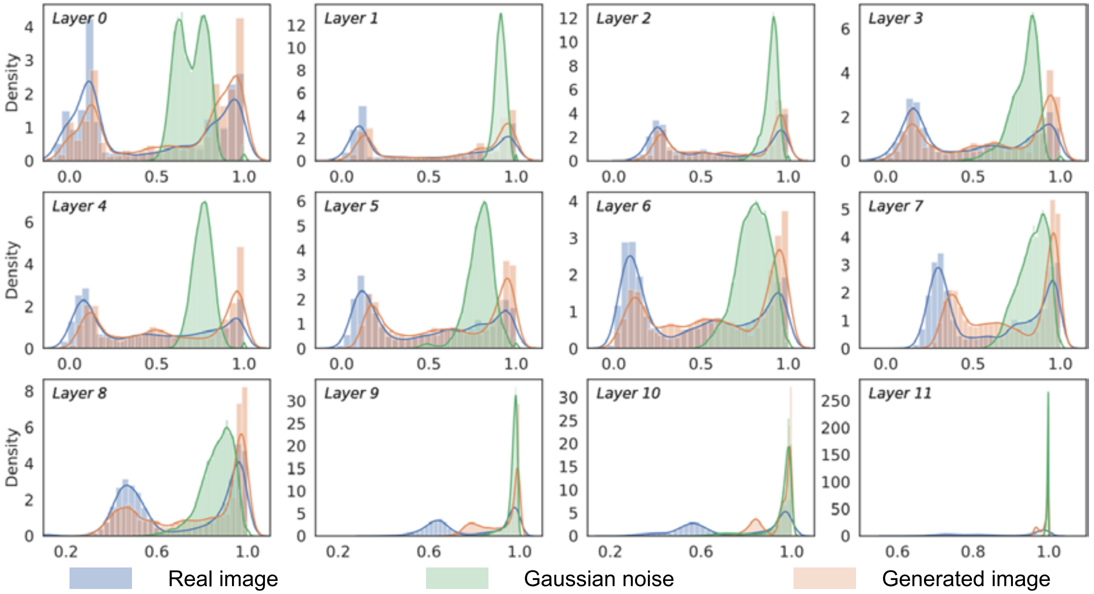


Fig. 3. Comparison of the kernel density curves of the patch similarity for each layer

Conclusions

- We propose PSAQ-ViT, a Patch Similarity Aware data-free Quantization framework for Vision Transformers.
  - PSAQ-ViT achieves high accuracy **without any access to training/testing data** during the quantization process.
  - Thanks to the positive feedback effect of the generated images, PSAQ-ViT can even **outperform the real-data-driven methods** at the same settings.

- An enhanced version has made it **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).

中国科学院自动化研究所
INSTITUTE OF AUTOMATION
CHINESE ACADEMY OF SCIENCES

中国科学院大学
University of Chinese Academy of Sciences

Patch Similarity Aware Data-Free Quantization for Vision Transformers

Thank you !