DFFG: Fast Gradient Iteration for Data-free Quantization

Huixing Leng[1], Shuangkan Fang[1], Yufeng Wang[2], Zehao Zhang[1], Dacheng Qi[1], Wenrui Ding[2]

[1] School of Electronic and Information Engineering, Beihang University

[2] Institute of Unmanned System, Beihang University

BMVC 2023

Abstract

Model quantization is a technique that optimizes neural network computation by converting weight parameters and activation values from floating-point numbers to low-bit integers or fixed-point representations.

Currently, common quantization methods, such as QAT and PTQ, optimize quantization parameters using training data to achieve the best performance. However, in practical applications, there may be little or no data available for downstream model quantization due to restrictions such as privacy and security. This article proposes a data-free quantization technique called DFFG, based on fast gradient iteration, which uses information learned from the full-precision model, such as the BN layer, to recover the distribution of the original training data.

We propose, for the first time, using a momentum-assisted variant of the FGSM gradient iteration strategy to update the generated data. This approach enables quick perturbation of the optimized data while maintaining the diversity of the generated data through the manipulation of gradient variability. We also propose using intermediate data generated during the iteration process as a part of data for subsequent model quantization, greatly improving the speed of data generation. We have demonstrated the effectiveness of our proposed method through empirical evaluations.

Preliminary Formulation

Quantizer : Uniform quantizer

$$\theta^q = round\,(\theta \times S - Z)\,,\ S = \frac{2^n - 1}{u - l},\ Z = S \times l + 2^{n-1}$$

$$\theta' = \frac{\theta^q + Z}{S}$$

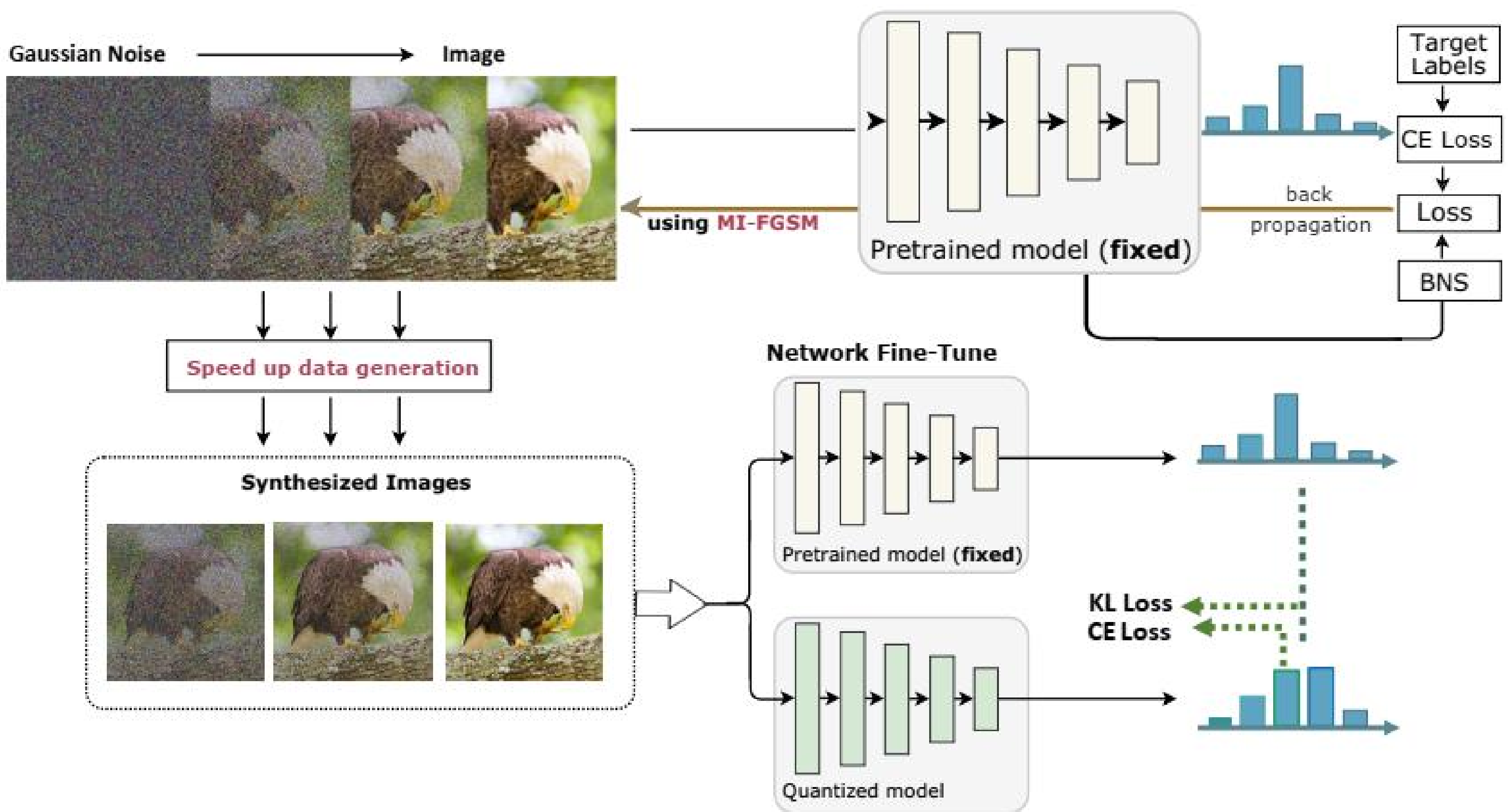
Optimizer : MI-FGSM

$$g_{t+1} = \mu \cdot g_t + \frac{J\left(x_t^*, y^*\right)}{\left\|\nabla_x J\left(x_t^*, y^*\right)\right\|_1}$$

$$x_{t+1}^* = x_t^* - \varepsilon \cdot \operatorname{sign}\left(g_{t+1}\right)$$

Method

DFFG : Data generation + Quantization(QAT)



Step1 : Data generation

Loss function :

$$\mathcal{L}_{total} = \mathcal{L}_{BNS} + \alpha_{\text{tv}}\mathcal{R}_{\text{TV}}(\mathbf{x}_i) + \boldsymbol{\alpha}_{\ell_2}\mathcal{R}_{\ell_2}(\mathbf{x}_i) + \boldsymbol{\beta}\mathcal{L}_{CE}$$

Optimizer : MI-FGSM

Speed up strategy :
save images multiple times during a complete iteration cycle

Step2 : Quantization(QAT)

Loss function :

$$\mathcal{L}^Q = \mathcal{L}_{\text{CE}}^Q + \alpha \cdot \mathcal{L}_{\text{KD}}^Q$$

Quantizer : uniform quantizer

Experiments & Results

Quantitative Results :

Dataset	Model	Bit width	Real Data	ZeroQ	DSG	ZAQ	Qimera	GDFQ	IntraQ	DFFG (ours)
CIFAR-10	ResNet-20 **(93.89)**	3w3a	87.94	69.53	48.99	-	-	71.1	77.07	**84.68**
		4w4a	91.52	89.66	88.93	**92.13**	91.26	90.25	91.49	91.63
		5w5a	-	-	-	93.36	**93.46**	93.38	-	93.30
CIFAR-100	ResNet-20 **(70.33)**	3w3a	56.26	26.35	43.42	-	-	43.87	48.25	**52.13**
		4w4a	66.8	63.97	62.62	60.42	65.1	63.58	64.98	**66.30**
		5w5a	-	-	-	68.7	69.02	67.52	-	**69.30**
ImageNet	ResNet-18 **(71.59)**	4w4a	67.89	63.38	63.11	52.64	63.84	60.6	66.47	**66.69**
		5w5a	70.31	69.72	69.53	64.54	69.29	66.82	69.94	**70.03**
	MobileNetV2 **(73.08)**	4w4a	67.9	60.15	60.45	0.1	61.62	51.3	65.10	**65.63**
		5w5a	72.01	70.95	70.87	62.35	70.45	68.14	71.28	**71.53**
	ResNet-50 **(77.76)**	4w4a	-	-	-	53.02	66.25	54.16	-	**69.47**
		5w5a	-	-	-	73.38	75.32	71.63	-	**75.83**

Visual generated images :

Ablation Studies :

a) quantization effect between Adam and MI-FGSM

model	Bit width	Adam	DFFG	Diff
ResNet-18 **(71.59)**	3w3a	37.68	**43.06**	**+5.38**
	4w4a	66.28	**66.69**	**+0.41**
	5w5a	**70.09**	70.03	-0.06
ResNet-50 **(77.76)**	4w4a	67.46	**69.47**	**+2.01**
	5w5a	75.52	**75.83**	**+0.31**

b) image CLIP similarity between Adam and MI-FGSM