FRanDI: Data-Free Neural Network Compression via Feature Regression and Deep Inversion

Konstantin Sobolev [1,2], Dmitry Ermilov [1], Nikolay Kozyrskiy [1], Anh-Huy Phan [1]

[1]Skoltech [2]AIRI

Skoltech
Skolkovo Institute of Science and Technology

AIRI

ICOMP 2024
INTERNATIONAL CONFERENCE ON COMPUTATIONAL OPTIMIZATION

Introduction

Modern post-training neural network compression methods effectively reduce model size and increase speed without significantly compromising performance. However, many of these techniques heavily depend on the original training dataset at various stages of the pipeline—during the evaluation of compression schemes, the compression process itself, and, most critically, during fine-tuning. However, in practical scenarios, access to training data may be limited due to privacy, security, licensing, or transmission issues.

In this work, we introduce **FRanDI**, an innovative framework that enables post-training neural network compression without the need for any data. The **FRanDI** framework consists of three components:

- **Synthetic data generation pipeline** that produces data mimicking the original training dataset;
- **Feature Regression** — a novel model recovery scheme that replaces fine-tuning when real data and labels are unavailable;
- **Output Discrepancy** — a new metric for evaluating model compression policies without the use of labels.

Method

Synthetic Data Generation pipeline optimizes input images to match original training data by reducing the distance between their feature distributions across multiple feature maps of a pre-trained model:

$$\mathcal{L}(\hat{x}) = \sum_{l=1}^{L} KL\left(\mathcal{N}\left(\hat{\mu}_l, \hat{\sigma}_l^2\right) \| \mathcal{N}\left(\mu_l^*, \sigma_l^{*2}\right)\right), \tag{1}$$

where statistics $\hat{\mu} = \mu(\hat{x})$ and $\hat{\sigma} = \sigma(\hat{x})$ are computed for image of $\hat{x}$ within the BN layer, μ^* and σ^* are original running estimates, L - number of layers.

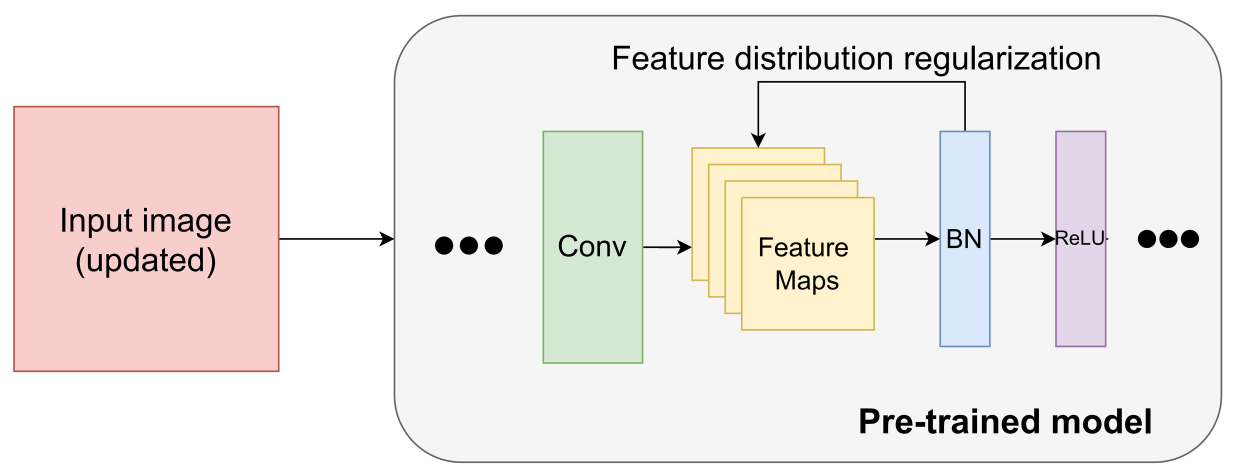


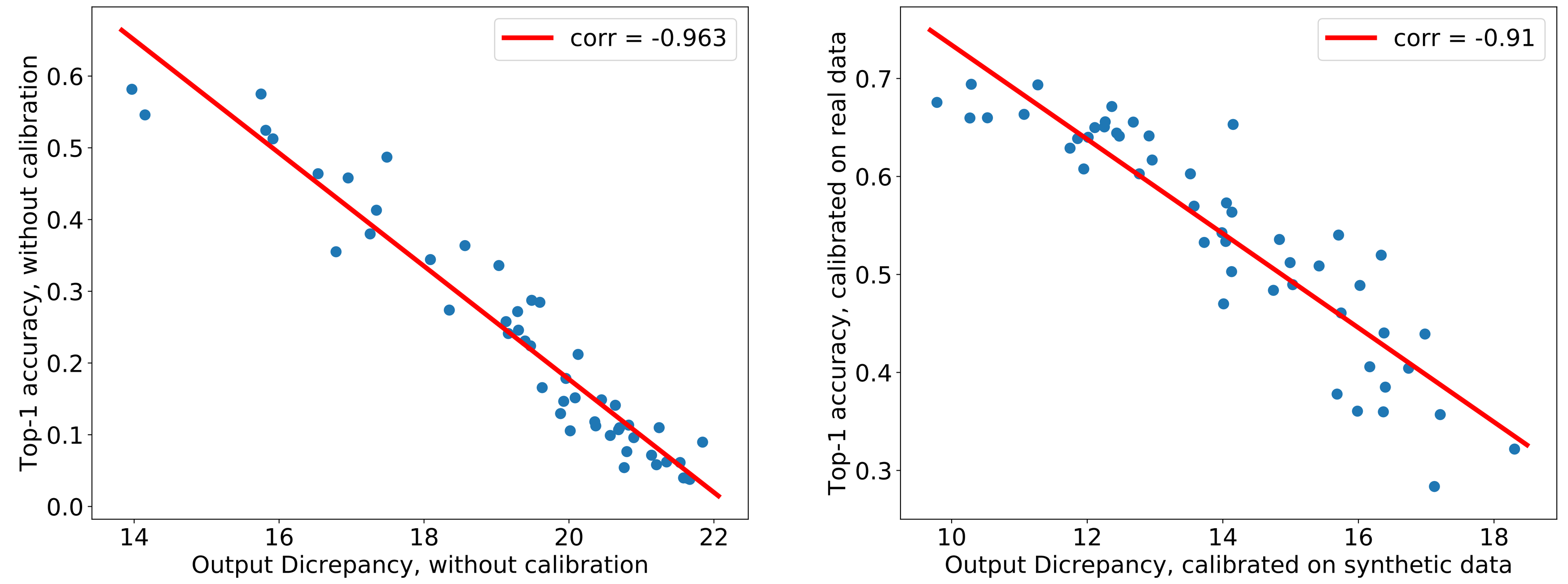
Figure 1. Synthetic data generation scheme overview.

Feature Regression utilizes a teacher-student approach on synthetic data and minimizes the *Feature Discrepancy* corresponding layers in the original and compressed models, reducing degradation after model compression:

$$\min_{\tilde{W}} L_{FR}(F_W(\hat{x}), F^*(\tilde{W})(\hat{x})) = \min_{\tilde{W}} \sum_{i=1}^{N} \left\| \frac{f_i^{w_i}(\hat{x})}{\|f_i^{w_i}(\hat{x})\|_2} - \frac{\tilde{f}_i^{\tilde{w}_i}(\hat{x})}{\|\tilde{f}_i^{\tilde{w}_i}(\hat{x})\|_2} \right\|_2, \tag{2}$$

Output Discrepancy is a new proxy metric that correlates with the target metric of the original model, enabling the evaluation of model compression policies, without use of the dataset and labels.

$$\textbf{OD}(F_W(\hat{x}), \tilde{F}_{\tilde{W}}(\hat{x})) = \left\| \frac{F_W(\hat{x})}{\|F_W(\hat{x})\|_2} - \frac{\tilde{F}_{\tilde{W}}(\hat{x})}{\|\tilde{F}_{\tilde{W}}(\hat{x})\|_2} \right\|_2 \tag{3}$$



(a) Without BN calibration (b) With BN calibration

Figure 2. Compressed model accuracy vs **OD** proxy metric.

Experiments

Table 1. Results for data-free unstructured pruning with magnitude-based approach. CR - compression ratio, ratio of non-zero parameters in the model.

Model	Dataset	CR	BatchSize	Top-1 Accuracy, %		
				Original	Fine-tuned	Recovered
ResNet-18	Cifar-100	0.5	256	77.10	76.12	76.62
ResNet-18	ImageNet	0.8	256	69.76	69.16	69.20
ResNet-50	ImageNet	0.5	128	76.13	72.23	72.81

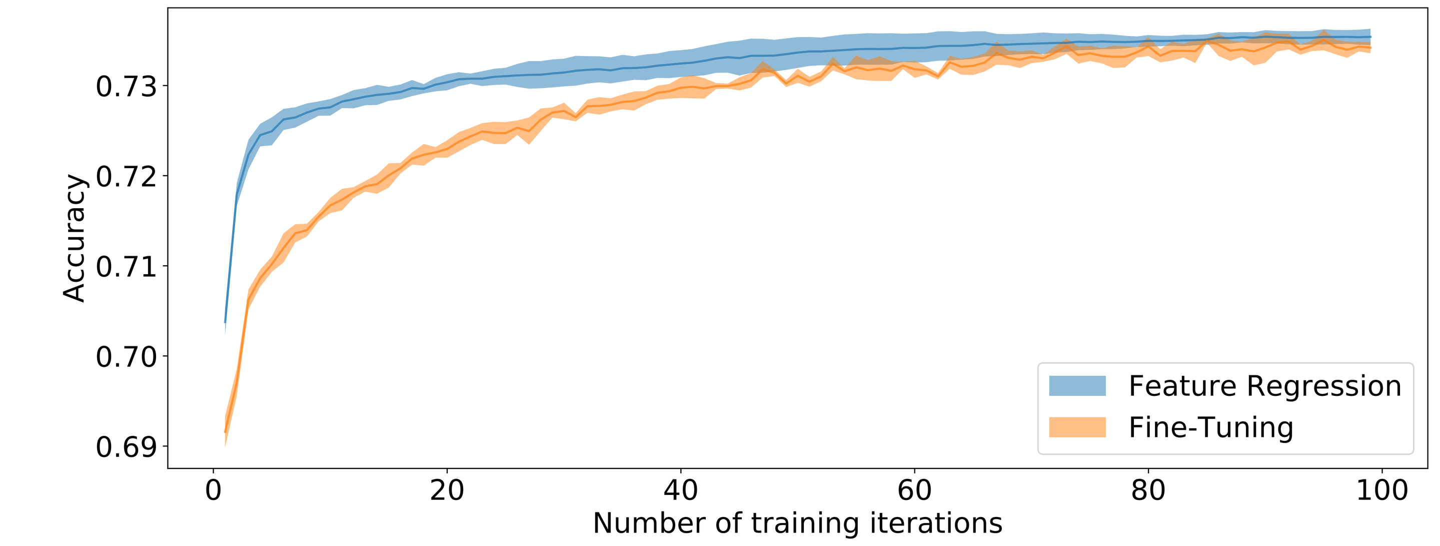


Figure 3. Feature Regression vs fine-tuning on real data, CIFAR-100 dataset.

Table 2. Data-free quantization methods comparison.

Method	Settings
Top-1 acc.	
ZeroQ [2]	70.25
GDFQ [5]	71.53
DFQ [4]	40.35
ACIQ [1]	54.73
ZAQ [3]	72.67
Ours	**75.90**

References

[1] Ron Banner, Yury Nahshan, Elad Hoffer, and Daniel Soudry. ACIQ: Analytical clipping for integer quantization of neural networks, 2019.

[2] Yaohui Cai, Zhewei Yao, Zhen Dong, Amir Gholami, Michael W Mahoney, and Kurt Keutzer. Zeroq: A novel zero shot quantization framework. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13169–13178, 2020.

[3] Yuang Liu, Wei Zhang, and Jun Wang. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2021.

[4] Markus Nagel, Mart van Baalen, Tijmen Blankevoort, and Max Welling. Data-free quantization through weight equalization and bias correction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

[5] Xu Shoukai, Li Haokun, Zhuang Bohan, Liu Jing, Cao Jiezhang, Liang Chuangrun, and Tan Mingkui. Generative low-bitwidth data free quantization. In *The European Conference on Computer Vision*, 2020.