FGFP: A Fractional Gaussian Filter and Pruning for Deep Neural Networks Compression

Kuan-Ting Tu*, Po-Hsien Yu*, Yu-Syuan Tseng, Shao-Yi Chien

Introduction

Network compression techniques have become increasingly important in recent years because the loads of Deep Neural Networks (DNNs) are heavy for edge devices in real-world applications. While many methods compress neural network parameters, deploying these models on edge devices remains challenging. To address this, we propose the fractional Gaussian filter and pruning (FGFP) framework, which integrates fractional-order differential calculus and the Gaussian function to construct fractional Gaussian filters (FGFs). To reduce the computational complexity of fractional-order differential operations, we introduce Grünwald-Letnikov fractional derivatives to approximate the fractional-order differential equation. The number of parameters for each kernel in FGF is minimized to only seven. Beyond the architecture of Fractional Gaussian Filters, our FGFP framework also incorporates Adaptive Unstructured Pruning (AUP) to achieve higher compression ratios. Experiments on various architectures and benchmarks show that our FGFP framework outperforms recent methods in accuracy and compression. On CIFAR-10, ResNet-20 achieves only a 1.52% drop in accuracy while reducing the model size by 85.2%. On ImageNet2012, ResNet-50 achieves only a 1.63% drop in accuracy while reducing the model size by 69.1%.

Overview

$$f^{(n)}(x) = \frac{d^{(n)}f}{dx^{(n)}} = \lim_{\Delta x \to 0} \frac{f^{(n-1)}(x+\Delta x) - f^{(n-1)}(x)}{\Delta x}$$

$$f^{(n)}(x) = \frac{d^{(n)}f}{dx^{(n)}} = \lim_{h \to 0} \frac{1}{h^n} \sum_{r=0}^{n} (-1)^r \binom{n}{r} f(x - rh)$$

n is replaced by α
calculate by Γ function

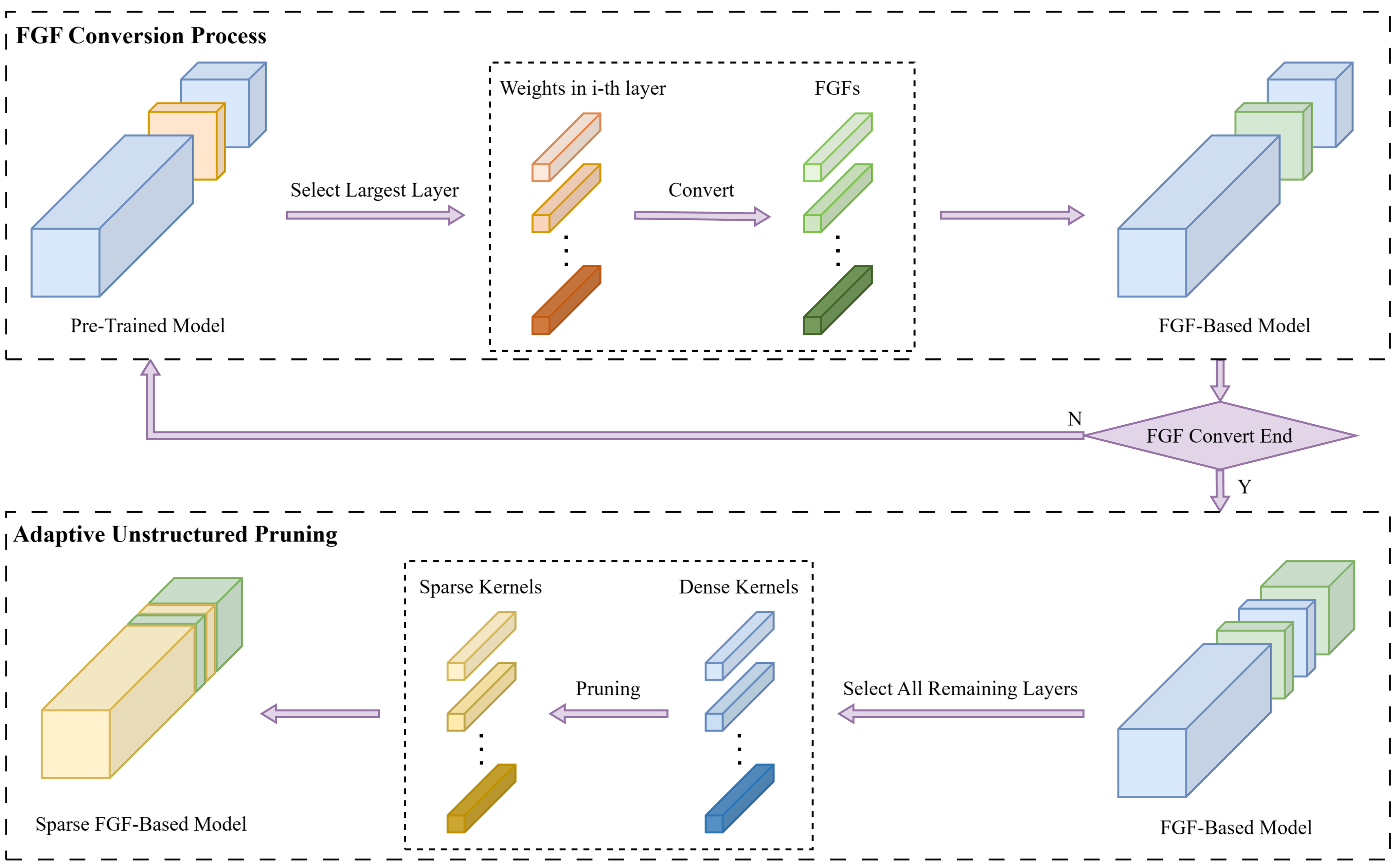
$$D^{\alpha}_{G-L} f(x) \approx f(x) + (-\alpha) f(x-1) + \frac{(-\alpha)(-\alpha+1)}{2} f(x-2) + \cdots + \frac{\Gamma(-\alpha+1) f(x-n)}{n!\,\Gamma(-\alpha+n+1)}$$

$$D^{\alpha}_{G-L} f(x) \approx f(x) - \alpha f(x-1) + \frac{\alpha(\alpha-1)}{2} f(x-2)$$

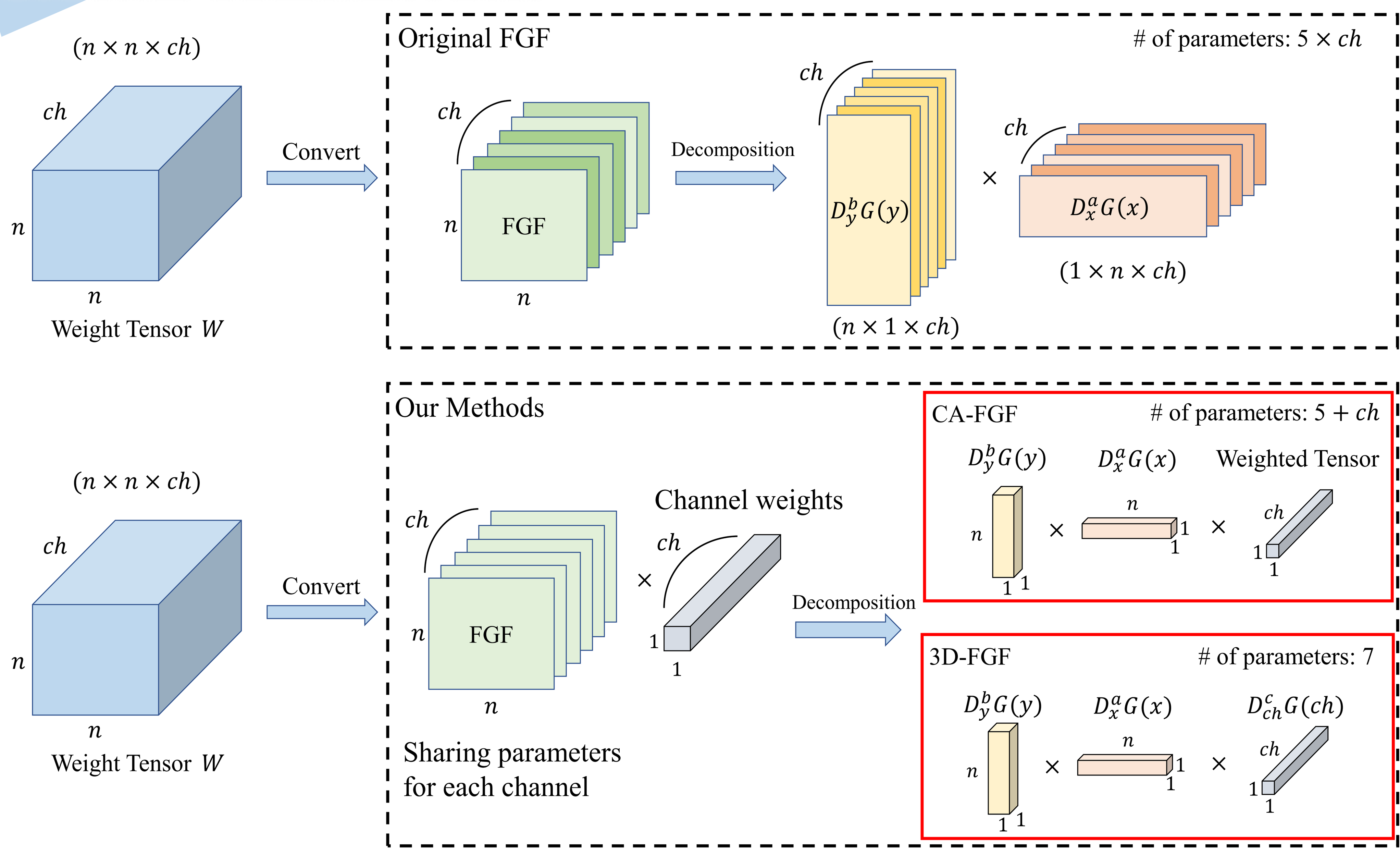
Contributions

1. Proposed the fractional Gaussian filter and pruning (FGFP) framework, which combines the fractional Gaussian filter (FGF) and adaptive unstructured pruning (AUP) to reduce the number of parameters significantly.
2. Used the channel-attention mechanism to design two forms of the fractional Gaussian filter (FGF): CA-FGF & 3D-FGF
3. Conducted comprehensive experiments with recent methods on two benchmarks, CIFAR-10 and ImageNet2012.

Overview



Fractional Gaussian Filter



Results

Method	Post-Trained Model Type	Top-1 Accuracy (%)			Parameter CR (%)
		Baseline	Compressed	Δ ↓	
ResNet-20					
SCOP (Tang et al., 2020)	Sparse	92.22	90.75	1.47	56.3
Hinge (Li et al., 2020)	Low-Rank + Sparse	92.54	91.84	0.70	55.5
FGFP(CA-FGF) (ours)	Fractional Filter + Sparse	91.34	90.77	**0.57**	59.3
FGFP(3D-FGF) (ours)	Fractional Filter + Sparse	91.34	90.34	1.00	**66.7**
PSTRN-M (Li et al., 2022)	Low-Rank	91.25	89.30	1.95	85.2
ELRT (Sui et al., 2024)	Low-Rank	91.25	89.64	1.61	83.4
TDLC (Liu et al., 2024)	Low-Rank	91.25	88.58	2.65	80.5
FGFP(CA-FGF) (ours)	Fractional Filter + Sparse	91.34	90.20	**1.14**	81.5
FGFP(3D-FGF) (ours)	Fractional Filter + Sparse	91.34	89.82	1.52	**85.2**
ResNet-32					
SCOP (Tang et al., 2020)	Sparse	92.66	92.13	0.53	56.2
PSTRN-S (Li et al., 2022)	Low-Rank	92.49	91.43	1.06	60.9
FGFP(CA-FGF) (ours)	Fractional Filter + Sparse	92.64	92.11	**0.53**	**76.1**
FGFP(3D-FGF) (ours)	Fractional Filter + Sparse	92.64	91.92	0.72	**76.1**
PSTRN-M (Li et al., 2022)	Low-Rank	92.49	90.59	1.90	80.4
ELRT (Sui et al., 2024)	Low-Rank	92.49	91.21	1.28	80.4
FGFP(CA-FGF) (ours)	Fractional Filter + Sparse	92.64	91.85	**0.79**	**80.4**
FGFP(3D-FGF) (ours)	Fractional Filter + Sparse	92.64	91.80	0.84	**80.4**
WRN-28-10					
GrowEfficient (Yuan et al., 2021)	Sparse	96.20	95.30	0.90*	90.7
BackSparse (Zhou et al., 2021)	Sparse	96.20	95.60	0.60*	91.6
FGFP(CA-FGF) (ours)	Fractional Filter + Sparse	94.78	93.68	1.10*	91.6
FGFP(3D-FGF) (ours)	Fractional Filter + Sparse	94.78	94.24	**0.54***	**96.8**

Method	Post-Trained Model Type	Top-1 Accuracy (%)			Parameter CR (%)
		Baseline	Compressed	Δ ↓	
ResNet-18					
FR (Chu & Lee, 2021)	Low-Rank	69.76	69.04	0.72	57.0
LRPET (Guo et al., 2024)	Low-Rank	69.76	67.87	1.89	50.3
FGFP(3D-FGF) (ours)	Fractional Filter + Sparse	69.30	68.61	**0.69**	60.1
FGFP(3D-FGF) (ours)	Fractional Filter + Sparse	69.30	68.28	1.02	**74.7**
ResNet-50					
EDP (Ruan et al., 2024)	Low-Rank + Sparse	75.90	75.34	0.56	43.9
ARPruning (Yuan et al., 2024)	Sparse	76.15	72.31	3.84	56.8
SFI-FP (Yang et al., 2024)	Sparse	76.15	75.21	0.94	57.3
CORING (Pham et al., 2024b)	Sparse	76.15	75.55	0.60	56.7
FGFP(3D-FGF) (ours)	Fractional Filter + Sparse	76.16	75.64	**0.52**	**57.4**
Stable (Phan et al., 2020)	Low-Rank	76.15	74.68	1.47	60.2
CC (Li et al., 2021)	Low-Rank + Sparse	76.15	74.54	1.61	58.6
FGFP(3D-FGF) (ours)	Fractional Filter + Sparse	76.16	75.42	**0.74**	**62.7**
AHC-A (Wang et al., 2024)	Sparse	76.20	74.70	1.50	63.4
LRPET-S (Guo et al., 2024)	Low-Rank	76.15	73.72	2.43	64.0
FGFP(3D-FGF) (ours)	Fractional Filter + Sparse	76.16	74.82	**1.34**	**66.8**
NORTON (Pham et al., 2024a)	Low-Rank + Sparse	76.15	74.00	2.15	68.8
FGFP(3D-FGF) (ours)	Fractional Filter + Sparse	76.16	74.53	**1.63**	**69.1**

Scan Me!!!