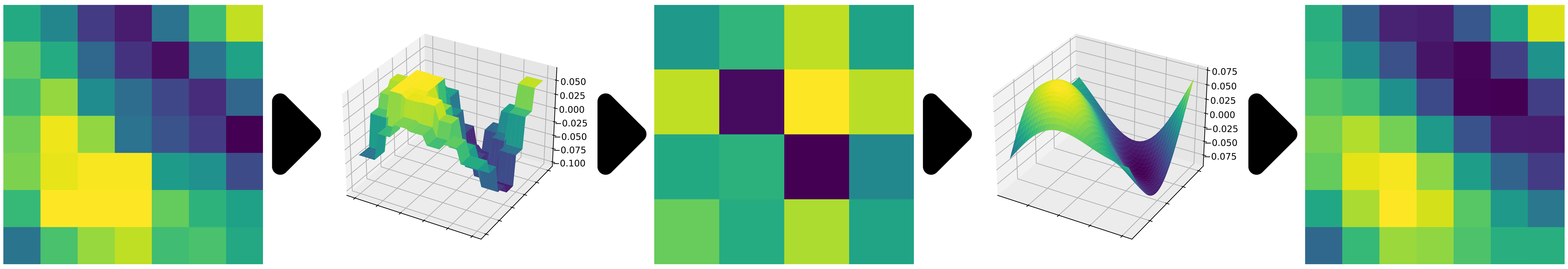
Approximating Continuous Convolutions for Deep Network Compression

T. W. Costain & V. A. Prisacariu
Active Vision Laboratory, University of Oxford
{costain,victor}@robots.ox.ac.uk

BMVC 2022

Method



We consider the discrete values of a given filter/channel as a continuous function

We approximate these continuous functions with a Cosine or Chebyshev series using fewer values than the original filter

We save these series coefficients, instead of the original filter

At inference time, the filter is reconstructed from these stored values

Unlike comparable methods [1, 2], our approach does not require training from scratch. Instead, we require only a small amount of fine-tuning (5 epochs).

Results

CIFAR 10:

		ResNet-20			ResNet-32				
	Layer	Pre Top 1	Post Top 1	Δ	Pre Top 1	Post Top 1	Δ	# Params	Δ %
	Conv2D	-	91.73	-	-	92.78	-	0.46M	0.0
	Fractional[2]	91.25	91.29	0.04	-	-	-	-	-
	FracSRF[1]	-	-	-	92.28	91.60	-1.18	0.16M	34.00
	FracSRF†	10.00	38.08	-53.65	9.82	37.91	-54.87	0.15M	33.43
CosConv	3,3,3,2	87.22	90.75	-0.98	87.74	91.51	-1.27	0.27M	57.74
	3,3,2,2	51.43	89.55	-2.18	75.62	90.99	-1.79	0.22M	47.35
	3,2,2,2	21.16	86.99	-4.74	16.52	88.4	-4.38	0.21M	44.57
	2,2,2,2	17.55	85.98	-5.75	15.39	87.87	-4.91	0.21M	44.48
ChebConv	3,3,3,2	87.22	90.76	-0.97	87.74	91.48	-1.30	0.27M	57.74
	3,3,2,2	51.43	89.83	-1.90	75.62	91.16	-1.62	0.22M	47.35
	3,2,2,2	21.16	87.79	-3.94	16.52	88.48	-4.30	0.21M	44.57
	2,2,2,2	17.55	86.66	-5.07	15.39	88.16	-4.62	0.21M	44.48

† indicates using our regression and retraining approach.

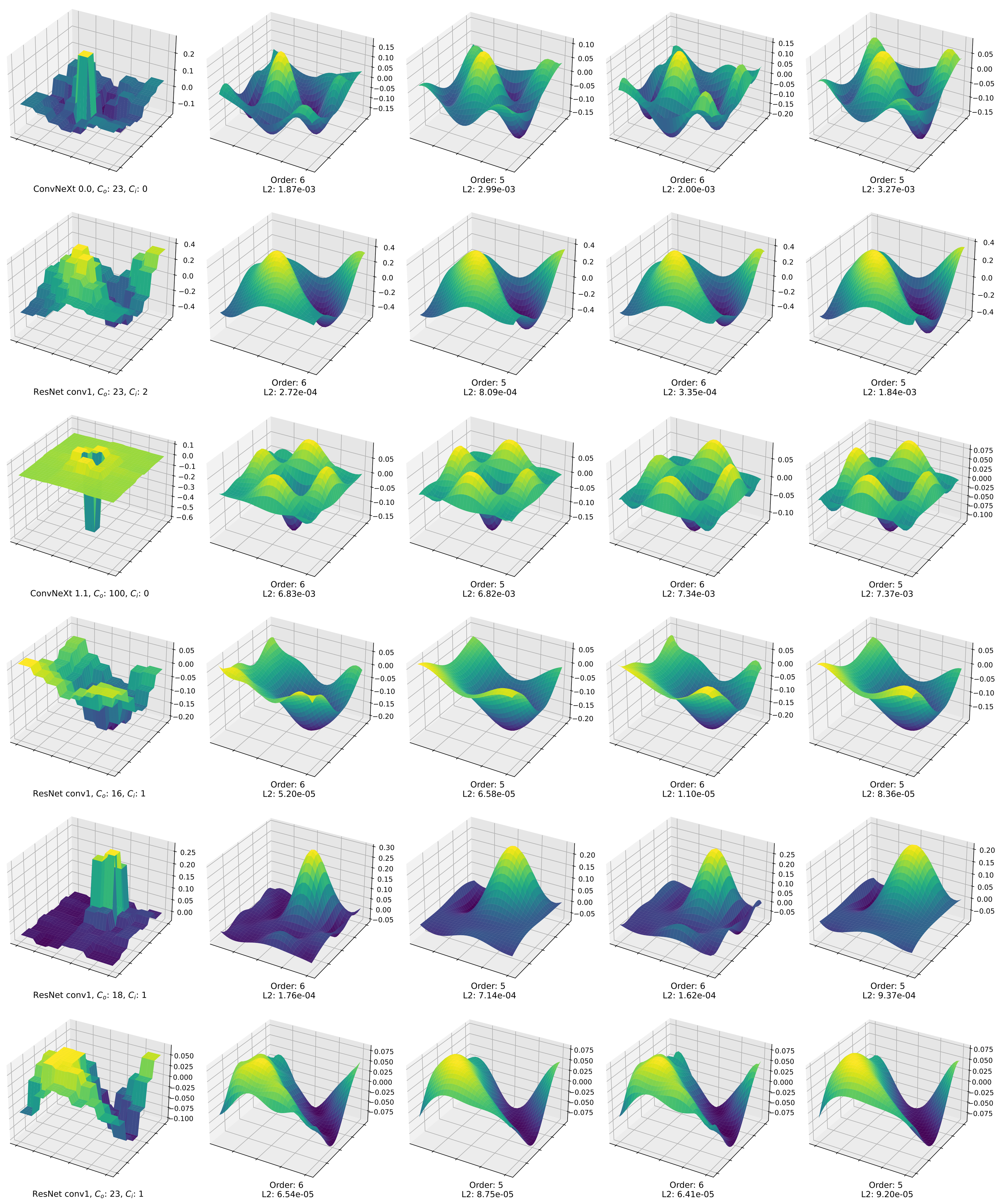
ResNet-18 on ImageNet:

	CosConv			ChebConv				
Config	Pre Top 1	Post Top 1	Δ %	Pre Top 1	Post Top 1	Δ %	# Params.	Δ %
Conv2D	-	69.76	0.0	-	69.76	0.0	11.68M	0.0
6,3,3,3,3	69.46	70.19	0.43	69.32	70.11	0.35	11.68M	99.98
6,3,3,3,2	64.33	68.97	-0.79	64.11	68.62	-1.14	7.09M	60.70
6,3,3,2,2	56.59	67.90	-1.86	56.35	67.23	-2.53	5.94M	50.88
5,3,3,3,3	66.71	69.80	0.04	65.79	69.74	-0.02	10.99M	94.10
5,3,3,3,2	60.81	68.53	-1.23	59.71	68.15	-1.61	7.09M	60.68
5,3,3,2,2	51.92	67.48	-2.28	13.58	65.44	-4.32	5.94M	50.86
4,3,3,3,2	56.21	68.05	-1.71	54.37	67.34	-2.42	7.09M	60.67

ConvNeXt-T on ImageNet:

	Pre Top 1	Post Top 1	Δ %	# Comp. Params.	Δ %
Conv2D	-	82.10		0.33M	0.0
$7_{\times 3}, 7_{\times 3}, 7_{\times 9}, 6_{\times 3},$	80.27	81.19	-0.91	0.30M	90.90
$7_{\times 3}, 7_{\times 3}, 7_{\times 4}, 6_{\times 5}, 6_{\times 3}$	79.10	80.87	-1.23	0.27M	83.32
$7_{\times 3}, 7_{\times 3}, 6_{\times 9}, 6_{\times 3}$	70.10	79.94	-2.16	0.25M	77.25
$7_{\times 3}, 7_{\times 3}, 5_{\times 9}, 5_{\times 3}$	62.40	79.28	-2.82	0.19M	58.01
$6_{\times 3}, 6_{\times 3}, 6_{\times 9}, 6_{\times 3}$	24.75	76.26	-5.84	0.24M	73.84
$5_{\times 3}, 5_{\times 3}, 5_{\times 9}, 5_{\times 3}$	14.94	77.29	-4.81	0.17M	51.71

Original Kernel | CosConv Approximation | ChebConv Approximation



Quantization

Applying simple quantisation on top of our method to ResNet-18 on ImageNet:

		CosConv			ChebConv				
bits	Config	Top1	Quant. Top1	Δ	Top1	Quant. Top1	Δ	# Params	Model size (MB)
8	Conv2D	69.76	69.74	-0.02	69.76	69.74	-0.02	11.68M	23.4
8	6,3,3,3,2	68.97	68.12	-0.85	67.23	67.39	0.16	7.09M	14.2
8	6,3,3,2,2	67.90	65.23	-2.67	65.63	65.20	-0.43	5.94M	11.9
4	Conv2D	69.76	69.28	-0.48	69.76	69.28	-0.48	11.68M	17.5
4	6,3,3,3,2	68.97	66.70	-2.27	67.23	66.50	-0.73	7.09M	10.6
4	6,3,3,2,2	67.90	64.10	-3.80	65.63	63.88	-1.75	5.94M	8.9

References

[1] N. Saldanha, S. L. Pintea, J. C. van Gemert, and N. Tomen. Frequency learning for structured cnn filters with gaussian fractional derivatives. In *BMVC*, 2021.

[2] J. Zamora, J. A. C. Vargas, A. Rhodes, L. Nachman, and N. Sundararajan. Convolutional filter approximation using fractional calculus. In *ICCV Workshops*, pages 383–392, October 2021.

Code and Paper:

HTTPS://APPROXCONV.ACTIVE.VISION