# Deep Roots
## Improving CNN Efficiency with Hierarchical Filter Groups

Yani Ioannou[1] Duncan Robertson[2] Roberto Cipolla[1] Antonio Criminisi[2]

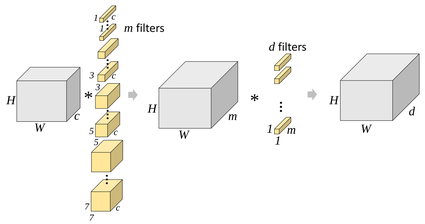
[1]University of Cambridge, [2]Microsoft Research, Cambridge

## Summary


- We train CNNs with 'root modules', a novel sparse connection structure based on filter groups.
- In effect our networks learn a *basis* for the channel extents of filters, based on compact filter groups
- Our models are faster and use fewer parameters, while maintaining or increasing accuracy
- ResNet 50, our model has 40% fewer parameters, 45% fewer FLOPS, and is 31% (12%) faster on a CPU (GPU)
- ResNet 200, our model has 48% fewer parameters and 27% fewer FLOPS
- GoogLeNet, our model has 7% fewer parameters and is 21% (16%) faster on a CPU (GPU)


## Previous Work: Learning a Basis for Filter Size

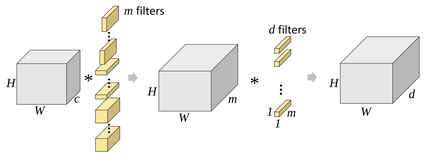
- In [1], linear combination of different sized filters is learned, *i.e.* a basis space for filters:




- Learns a limited number of large filters (7×7, 5×5), and a great number of small filters (3×3, 1×1)
- Motivation: expect most image correlations to be highly localized, *i.e.* many small filters. However, a few may require larger, more complex filters
- We can learn an effective filter of full size, but limited parameterization, by learning a basis – the 1×1 filters can learn a linear combination of the basis

## Previous Work: Learning a Low-Rank Basis for Filters

- In [2], linear combination of low-rank filters is learned, *i.e.* a basis space for filters:




- A set of filters of different shape (similar to 'Inception', but low-rank and of different orientation [1])
- On the following layer, use $d \times [1 \times 1 \times m]$ filters to linearly combine
- Only learning a low-rank filter in the spatial dimensions – filters maintain full channel depth

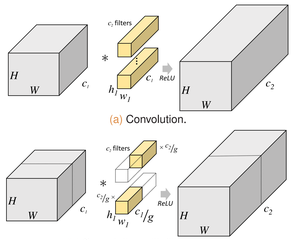
## Motivation: Learning a Basis for Inter-Layer Filter Dependence

- These previous works have learned a more compact representation by learning a basis, informed by our prior knowledge of the task
- However, they have only looked at the spatial extents of a convolutional filter – what about the 'channel' extents?

Idea:

- In a CNN, each filter learns a connection to every filter on the previous layer
- However, it has been shown that the learned weights in CNNs are sparse
- In a DPM/hierachical representation parts have sparse relationships
- Can we learn a more compact representation exploiting this sparsity?

## Review: Filter Groups

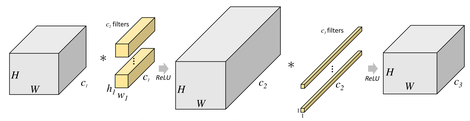



(a) Convolution.

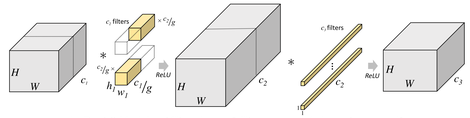
(b) Convolution with filter groups.

- Convolutional filters (yellow) typically have the same channel dimension ($c_1$) as the input feature maps (gray)
- With convolutional filter groups [3], $g$ independent groups of $c_2/g$ filters operate on a fraction $c_1/g$ of the input feature map channels
- This reduces filter dimensions from $h \times w \times c_1$ to $h \times w \times c_1/g$
- This change does not affect the dimensions of the input and output feature maps but significantly reduces computational complexity and the number of model parameters.

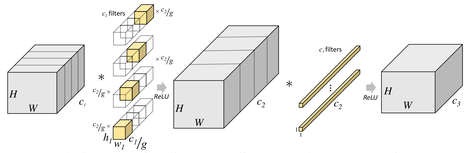
## Contribution: Root Module




(a) Convolution with $d$ filters of shape $h \times w \times c$.



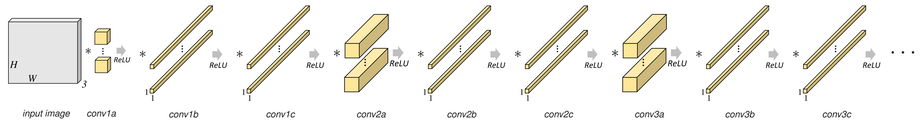

(b) Root-2 Module: $d$ filters in $g = 2$ filter groups, of shape $h \times w \times c/2$.



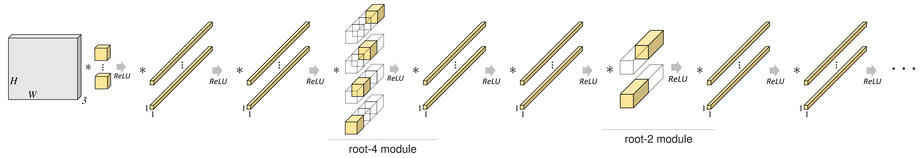

(c) Root-4 Module: $d$ filters in $g = 4$ filter groups, of shape $h \times w \times c/4$.

- (a) shows the typical set of conv. layers found in ResNet and other modern architectures
- Root modules (b), (c) have a given number of filter groups, with fewer connections to the previous layer's outputs
- Spatial convolutional layer is followed by a 1×1 convolution. Like in [2], this learns a linear combination of the basis filters (filter groups)
- Learns a filter of full channel depth, but with limited filter dependence

## Network-in-Network with Root Modules

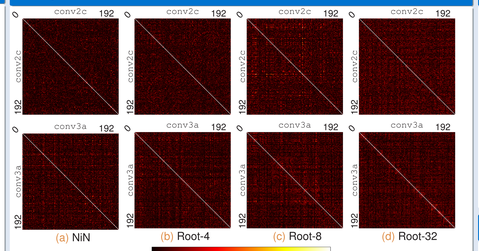



(a) Standard Network-in-Network Architecture



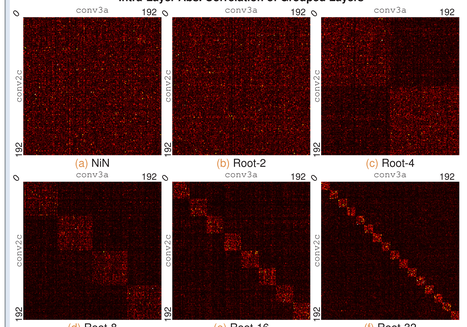

(b) Root-4 Architecture

## Analysis: Network-in-Network




(a) NiN (b) Root-4 (c) Root-8 (d) Root-32

Intra-Layer Abs. Correlation of Grouped Layers



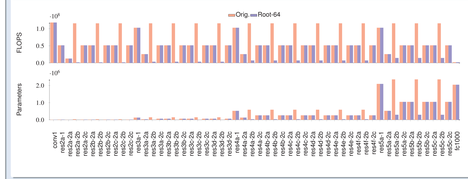

(a) NiN (b) Root-2 (c) Root-4

(d) Root-8 (e) Root-16 (f) Root-32

Inter-layer Covariance of conv2c and conv3a

- The block-diagonal sparsity learned by a root module is visible in the correlation of filters on layers `conv3a` and `conv2c` in the NiN network
- Rather than the salt and pepper like structure of dependence seen in (a), root modules have more structured dependence, where well correlated filters are grouped close to the related filters on the previous layer

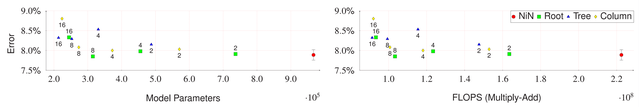
## Efficiency: ResNet 50 Layer-wise FLOPS/Parameters




## Later Work: ResNeXt

- "Aggregated Residual Transformations" [4] ≡ Root Modules [5]
- Explores a different compute/accuracy trade-off than our paper. We focus on decreasing computation while they focus on increasing accuracy

## Results: Network-in-Network (CIFAR10)




- Spatial filters (3×3, 5×5) are grouped hierarchically. The best models are closest to the origin. For the standard network, the mean and standard deviation (error bars) are shown over 5 different random initializations.

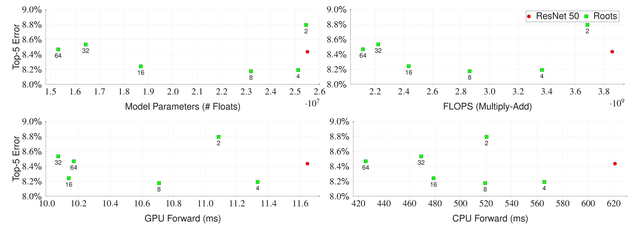
## Results: ResNet 50 (Imagenet)

(a) **Filter groups per layer**

| Model | conv1 7×7 | res2{a–c} 1×1 3×3 | res3{a–d} 1×1 3×3 | res4{a–f} 1×1 3×3 | res5{a–c} 1×1 3×3 |
|---|---|---|---|---|---|
| Orig. | 1 | 1 1 | 1 1 | 1 1 | 1 1 |
| root-2 | 1 | 1 2 | 1 1 | 1 1 | 1 1 |
| root-4 | 1 | 1 4 | 1 2 | 1 1 | 1 1 |
| root-8 | 1 | 1 8 | 1 4 | 1 2 | 1 1 |
| root-16 | 1 | 1 16 | 1 8 | 1 4 | 1 2 |
| root-32 | 1 | 1 32 | 1 16 | 1 8 | 1 4 |
| root-64 | 1 | 1 64 | 1 32 | 1 16 | 1 8 |

(b) **Results**

| Model | FLOPS ×10^9 | Param. ×10^7 | Top-1 Acc. | Top-5 Acc. | CPU (ms) | GPU (ms) |
|---|---|---|---|---|---|---|
| Orig. | 3.86 | 2.55 | 0.730 | 0.916 | 621 | 11.6 |
| root-2 | 3.68 | 2.54 | 0.727 | 0.912 | 520 | 11.1 |
| root-4 | 3.37 | 2.51 | **0.734** | 0.918 | 566 | 11.3 |
| root-8 | 2.86 | 2.32 | 0.734 | **0.918** | 519 | 10.7 |
| root-16 | 2.43 | 1.87 | 0.732 | 0.918 | 479 | 10.1 |
| root-32 | 2.22 | 1.64 | 0.729 | 0.915 | 469 | 10.1 |
| root-64 | **2.11** | **1.53** | 0.732 | 0.915 | 426 | 10.2 |




## Results: ResNet 200 (Imagenet)


## References


[1] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going Deeper with Convolutions", in *Computer Vision and Pattern Recognition (CVPR)*, 2015.

[2] Y. Ioannou, D. P. Robertson, J. Shotton, R. Cipolla, and A. Criminisi, "Training CNNs with Low-Rank Filters for Efficient Image Classification", in *International Conference on Learning Representations*, 2016.

[3] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks", in *Advances In Neural Information Processing Systems*, 2012. arXiv: 1102.0183.

[4] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, "Aggregated Residual Transformations for Deep Neural Networks", in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2017, Honolulu, Hawaii, 2017*. arXiv: 1611.05431 [cs.CV].

[5] Y. Ioannou, D. Robertson, R. Cipolla, and A. Criminisi, "Deep Roots: Improving CNN efficiency with hierarchical filter groups", in *IEEE Conference on Computer Vision and Patterrn Recognition (CVPR) 2017, Honolulu, Hawaii, 2016*. arXiv: 1605.06489.




yai20@cam.ac.uk