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

Abstract

We propose a new method for creating computationally efficient and compact convolutional neural networks (CNNs) using a novel sparse connection structure that resembles a tree root. This allows a significant reduction in computational cost and number of parameters compared to state-of-the-art deep CNNs, without compromising accuracy, by exploiting the sparsity of inter-layer filter dependencies. We validate our approach by using it to train more efficient variants of state-of-the-art CNN architectures, evaluated on the CIFAR10 and ILSVRC datasets. Our results show similar or higher accuracy than the baseline architectures with much less computation, as measured by CPU and GPU timings. For example, for ResNet 50, our model has 40% fewer parameters, 45% fewer floating point operations, and is 31% (12%) faster on a CPU (GPU). For the deeper ResNet 200 our model has 25% fewer floating point operations and 44% fewer parameters, while maintaining state-of-the-art accuracy. For GoogLeNet, our model has 7% fewer parameters and is 21% (16%) faster on a CPU (GPU).

1. Introduction

This paper describes a new method for creating computationally efficient and compact convolutional neural networks (CNNs) using a novel sparse connection structure that resembles a tree root. This allows a significant reduction in computational cost and number of parameters compared to state-of-the-art deep CNNs without compromising accuracy.

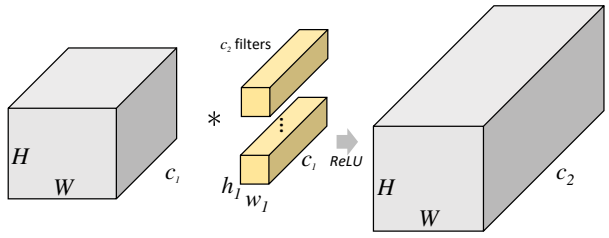
It has been shown that a large proportion of the learned weights in deep networks are redundant [1], a property that has been widely exploited to make neural networks smaller and more computationally efficient [2, 3]). It is unsurprising then that regularization is a critical part of training such networks using large datasets [4]. Without regularization deep networks are susceptible to over-fitting. Regularization may be achieved by weight decay or dropout [5]. Furthermore, a carefully designed sparse network connection structure can also have a regularizing effect. Convolutional Neural Networks (CNNs) [6, 7] embody this idea, using a sparse convolutional connection structure to exploit the locality of natural image structure. In consequence, they are easier to train.

With few exceptions, state-of-the-art CNNs for image recognition are largely monolithic, with each filter operating on the feature maps of all filters on a previous layer. Interestingly, this is in stark contrast to what we understand of biological neural networks, where we see "highly evolved arrangements of smaller, specialized networks which are interconnected in very specific ways" [8].

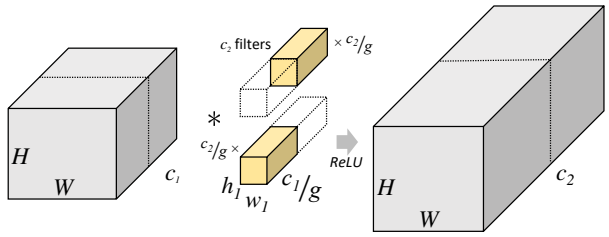
Recently, learning a low-rank basis for filters was found to improve generalization while reducing the computational complexity and model size of a CNN with only full rank filters [9]. However, this work addressed only the spatial extents of the convolutional filters (*i.e.* h and w in Fig. 1a). In this work we will show that a similar idea can be applied to the channel extents – *i.e.* filter inter-connectivity – by using *filter groups* [4]. We show that simple alterations to state-of-the-art CNN architectures can drastically reduce computational cost and model size without compromising accuracy.

2. Related Work

Most previous work on reducing the computational complexity of CNNs has focused on approximating convolutional filters in the spatial (as opposed to the channel) domain, either by using low-rank approximations [9–13], or Fourier transform based convolution [14, 15]. More general methods have used reduced precision number representations [16] or compression of previously trained models [17, 18]. Here we explore methods that reduce the computational impact of the large number of filter channels within state-of-the art networks. Specifically, we consider decreasing the number of incoming connections to nodes.



(a) Convolution.



(b) Convolution with filter groups.

Figure 1: **Filter Groups.** (a) Convolutional filters (yellow) typically have the same channel dimension c_1 as the input feature maps (gray) on which they operate. However, (b) with filter grouping, g independent groups of c_2/g filters operate on a fraction c_1/g of the input feature map channels, reducing 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.

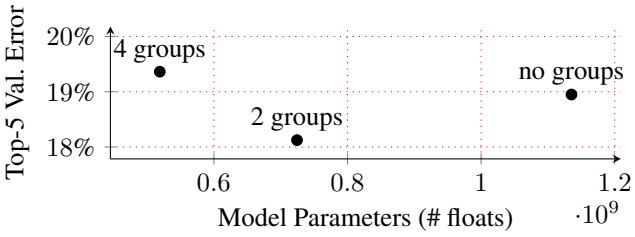


Figure 2: **AlexNet Filter Groups.** Model Parameters *vs.* top-5 error for variants of the AlexNet model on ILSVRC image classification dataset. Models with moderate numbers of filter groups have far fewer parameters, yet surprisingly maintain comparable error.

AlexNet Filter Groups. Amongst the seminal contributions made by Krizhevsky et al. [4] is the use of 'filter groups' in the convolutional layers of a CNN (see Fig. 1). While their use of filter groups was necessitated by the practical need to sub-divide the work of training a large network across multiple GPUs, the side effects are somewhat surprising. Specifically, the authors observe that independent filter groups learn a separation of responsibility (colour features vs. texture features) that is consistent over different random initializations. Also surprising, and not explicitly stated in [4], is the fact that the AlexNet network has approximately 57% fewer connection weights than the corresponding network without filter groups. This is due to the reduction in the input channel dimension of the grouped convolution filters (see Fig. 2). Despite the large difference in the number of parameters between the models, both achieve comparable accuracy on ILSVRC – in fact the smaller grouped network gets $\approx 1\%$ lower top-5 validation error. This paper builds upon these findings and extends them to state-of-the-art networks.

Low-dimensional Embeddings. Lin et al. [19] proposed a method to reduce the dimensionality of convolutional feature maps. By using relatively cheap '1×1' convolutional layers (i.e. layers comprising d filters of size $1 \times 1 \times c$, where $d < c$), they learn to map feature maps into lower-dimensional spaces, *i.e.* to new feature maps with fewer channels. Subsequent spatial filters operating on this lower dimensional input space require significantly less computation. This method is used in most state of the art networks for image classification to reduce computation [2, 20]. Our method is complementary.

GoogLeNet. In contrast to much other work, Szegedy et al. [2] propose a CNN architecture that is highly optimized for computational efficiency. GoogLeNet uses, as a basic building block, a mixture of low-dimensional embeddings [19] and heterogeneously sized spatial filters – collectively an 'inception' module. There are two distinct forms of convolutional layers in the inception module, low-dimensional embeddings (1×1) and spatial (3×3, 5×5). GoogLeNet keeps large, expensive spatial convolutions (*i.e.* 5×5) to a minimum by using few of these filters, using more 3×3 convolutions, and even more 1×1 filters. The motivation is that most of the convolutional filters respond to localized patterns in a small receptive field, with few requiring a larger receptive field. The number of filters in each successive inception module increases slowly with decreasing feature map size, in order to maintain computational performance. GoogLeNet is by far the most efficient state-of-the-art network for ILSVRC, achieving near state-of-the-art accuracy with the lowest computation/model size. However, we will show that even such an efficient and optimized network architecture benefits from our method.

Low-Rank Approximations. Various authors have suggested approximating learned convolutional filters using tensor decomposition [11, 13, 18]. For example, Jaderberg et al. [11] propose approximating the convolutional filters in a trained network with representations that are low-rank both in the spatial and the channel domains. This approach significantly decreases computational complexity, albeit at the expense of a small amount of accuracy. In this paper

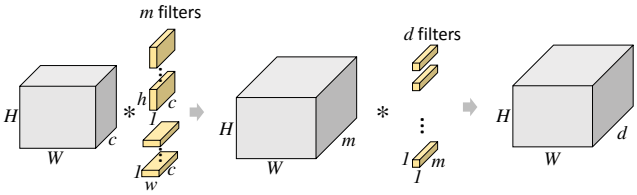


Figure 3: **Learning a (Spatial) Basis for Filters.** Learning a linear combination of mostly small, heterogeneously sized spatial filters [9]. Note that all filters operate on all c channels of the input feature map.

we are not approximating an existing model's weights but creating a new network architecture with explicit structural sparsity, which is then trained from scratch.

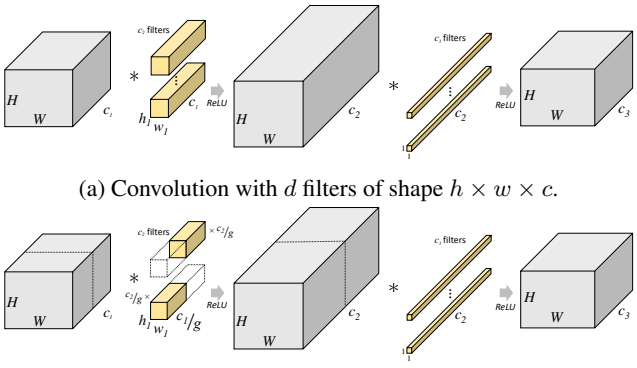
Learning a Basis for Filters Our approach is connected with that of Ioannou et al. [9] who showed that replacing 3×3×c filters with linear combinations of filters with smaller spatial extent (*e.g.* $1\times3\times c$, $3\times1\times c$ filters, see Fig. 3) could reduce the model size and computational complexity of state-of-the-art CNNs, while maintaining or even increasing accuracy. However, that work did not address the channel extent of the filters.

3. Root Architectures

In this section we present the main contribution of our work: the use of novel sparsely connected architectures resembling tree roots – to decrease computational complexity and model size compared to state-of-the-art deep networks for image recognition.

Learning a Basis for Filter Dependencies It is unlikely that every filter (or neuron) in a deep neural network needs to depend on the output of all the filters in the previous layer. In fact, reducing filter co-dependence in deep networks has been shown to benefit generalization. For example, Hinton et al. [5] introduced *dropout* for regularization of deep networks. When training a network layer with dropout, a random subset of neurons is excluded from both the forward and backward pass for each mini-batch. Furthermore, Cogswell et al. [21] observe a correlation between the covariance of hidden unit activations and overfitting. To explicitly reduce the covariance of hidden activations, they train networks with a loss function, based on the covariance matrix of the activations in a hidden layer.

Instead of using a modified loss, regularization penalty, or randomized network connectivity during training to prevent co-adaption of features, we take a much more direct approach. We use filter groups (see Fig. 1) to force the network to learn filters with only limited dependence on previous layers. Each of the filters in the filter groups is smaller



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

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

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

Figure 4: **Root Modules.** Root modules (b), (c) compared to a typical set of convolutional layers (a) found in ResNet and other modern architectures. Grey blocks represent the feature maps over which a layer's filters operate, while colored blocks represent the filters of each layer.

in the channel extent, since it operates on only a subset of the channels of the input feature map.

This reduced connectivity also reduces computational complexity and model size since the size of filters in filter groups are reduced drastically, as is evident in Fig. 4. Unlike methods for increasing the efficiency of deep networks by approximating pre-trained existing networks (see §2), our models are trained from random initialization using stochastic gradient descent. This means that our method can also speed up training and, since we are not merely approximating an existing model's weights, the accuracy of the existing model is not an upper bound on accuracy of the modified model.

Root Module The basic element of our network architecture, a *root module*, is shown in Fig. 4. A root module has a given number of filter groups, the more filter groups, the fewer the number of connections to the previous layer's outputs. Each spatial convolutional layer is followed by a low-dimensional embedding (1×1 convolution). Like in [9], this configuration learns a linear combination of the basis filters (filter groups), implicitly representing a filter of full channel depth, but with limited filter dependence.

Table 1: **Network-in-Network**. Filter groups in each convolutional layer.

Model	conv1			conv2			conv3		
	a	*b*	*c*	*a*	*b*	*c*	*a*	*b*	*c*
	5×5	*1×1*	*1×1*	*5×5*	*1×1*	*1×1*	*3×3*	*1×1*	*1×1*
Orig.	1	1	1	1	1	1	1	1	1
root-2	1	1	1	2	1	1	1	1	1
root-4	1	1	1	4	1	1	2	1	1
root-8	1	1	1	8	1	1	4	1	1
root-16	1	1	1	16	1	1	8	1	1

Table 2: **Network-in-Network CIFAR10**

Model	FLOPS $\times 10^8$	Param. $\times 10^5$	Accuracy	CPU (ms)	GPU (ms)
Orig.	2.22	9.67	0.9211	39.0	0.623
root-2	1.64	7.37	0.9209	31.2	0.551
root-4	1.23	4.55	0.9202	27.6	0.480
root-8	1.03	3.15	**0.9215**	24.4	0.482
root-16	**0.93**	**2.45**	0.9167	23.0	0.475

4. Results

Here we present image classification results obtained by replacing spatial convolutional layers within existing state-of-the-art network architectures with root modules (described in §3) .

4.1. Improving Network in Network on CIFAR-10

Network in Network (NiN) [19] is a near state-of-the-art network for CIFAR-10 [22]. It is composed of 3 spatial (5×5, 3×3) convolutional layers with a large number of filters (192), interspersed with pairs of low-dimensional embedding (1×1) layers. As a baseline, we replicated the standard NiN network architecture as described by Lin et al. [19] but used state-of-the-art training methods. We trained using random 32×32 cropped and mirrored images from 4-pixel zero-padded mean-subtracted images, as in [20, 23]. We also used the initialization of He et al. [24] and batch normalization [25]. With this configuration, ZCA whitening was not required to reproduce validation accuracies obtained in [19]. We also did not use dropout, having found it to have little effect, presumably due to our use of batch normalization, as suggested by Ioffe and Szegedy [25].

To assess the efficacy of our method, we replaced the spatial convolutional layers of the original NiN network with root modules (as described in §3). We preserved the original number of filters per layer but subdivided them into groups as shown in Table 1. We considered the first of the pair of existing 1×1 layers to be part of our root modules. We did not group filters in the first convolutional layer – since it operates on the three-channel image space, it is of limited computational impact compared to other layers. Results are shown in Table 2 and Fig. 5 for various network architectures[1]. Compared to the baseline architecture, the root variants achieve a significant reduction in computation and model size without a significant reduction in accuracy. For example, the root-8 architecture gives equivalent accuracy with only 46% of the floating point operations (FLOPS), 33% of the model parameters of the original network, and approximately 37% and 23% faster CPU and GPU timings (see §5 for an explanation of the GPU timing disparity).

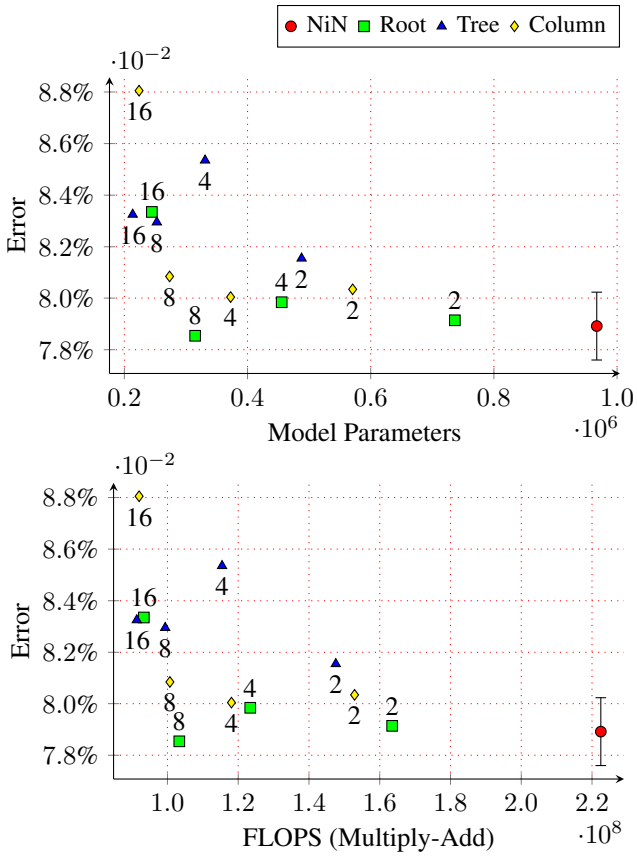
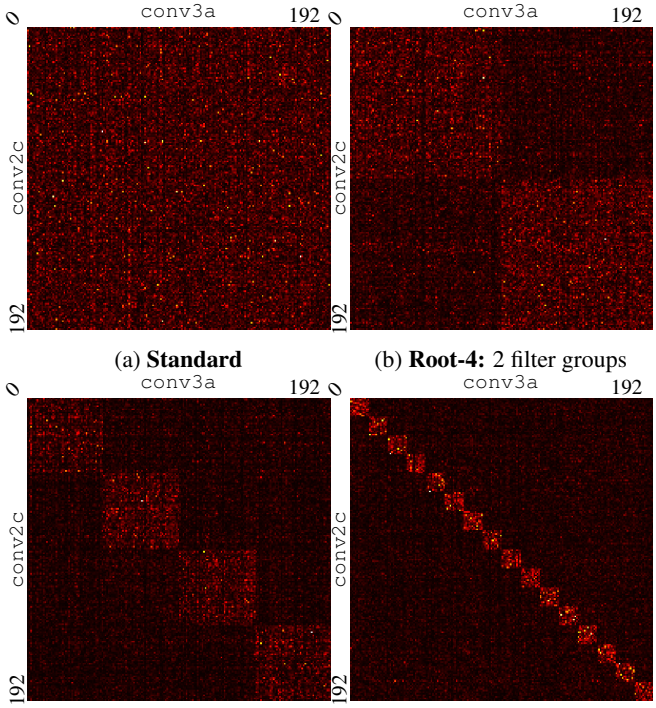


Figure 5: **Network-in-Network CIFAR10 Results.** Spatial filters (3×3, 5×5) are grouped hierarchically. The best models are closest to the origin. For the standard network, the mean and standard deviation (error bars) are shown over 5 different random initializations.

Figure 6 shows the inter-layer correlation between the adjacent filter layers `conv2c` and `conv3a` in the network

[1]Here (and subsequently unless stated otherwise) timings are per image for a forward pass computed on a large batch. Networks were implemented using Caffe (with CuDNN and MKL) and run on an Nvidia Titan Z GPU and 2 10-core Intel Xeon E5-2680 v2 CPUs.



(a) **Standard** (b) **Root-4:** 2 filter groups

(c) **Root-8:** 4 filter groups (d) **Root-32:** 16 filter groups

Figure 6: **Inter-layer Filter Correlation.** 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.

architectures outlined in Table 1 as evaluated on the CIFAR test set. The block-diagonalization enforced by the filter group structure (as illustrated in Fig. 1) is visible, more so with larger number of filter groups. This shows that the network learns an organization of filters such that the sparsely distributed strong filter relations, visible in 6a as brighter pixels, are grouped into a denser block-diagonal structure, leaving a visibly darker, low-correlated background. See §A.2 for more images, and an explanation of their derivation.

4.2. Grouping Degree with Network Depth

An interesting question concerns how the degree of grouping in our root modules should be varied as a function of depth in the network. For the NiN-like architectures described earlier, we might consider having the degree of grouping: (1) decrease with depth after the first convolutional layer, *e.g.* 1–8–4 ('root'); (2) remain constant with depth after the first convolutional layer, *e.g.* 1–4–4 ('column'); or (3) increase with depth, *e.g.* 1–4–8 ('tree').

To determine which approach is best, we created variants of the NiN architecture with different degrees of grouping per layer. Results are shown in Fig. 5 (numerical results are included in §A.1). The results show that the so-called root topology (illustrated in Fig. 7) gives the best performance, providing the smallest reduction in accuracy for a given reduction in model size and computational complexity. Similar experiments with deeper network architectures have delivered similar results and so we have reported results for root topologies. This aligns with the intuition of deep networks for image recognition subsuming the deformable parts model. If we assume that filter responses identify parts (or more elemental features), then there should be more filter dependence with depth, as more parts (filter responses) are assembled into complex concepts.

Table 3: **ResNet 50**. Filter groups in each conv. layer.

Model	conv1	res2{a–c}		res3{a–d}		res4{a–f}		res5{a–c}	
	7×7	*1×1*	*3×3*	*1×1*	*3×3*	*1×1*	*3×3*	*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

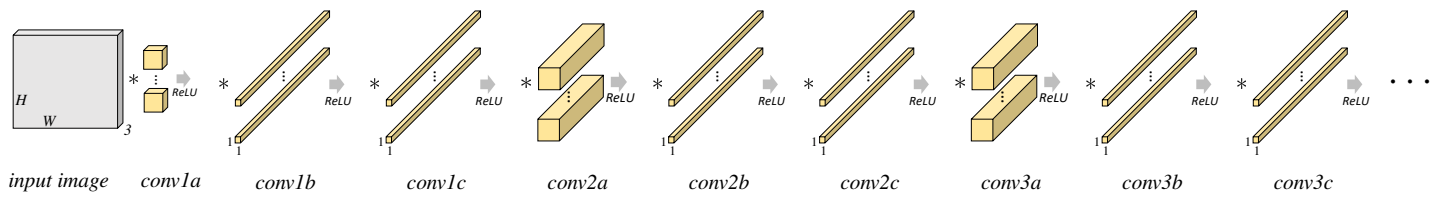
4.3. Improving Residual Networks on ILSVRC

Residual networks (ResNets) [20] are the state-of-the art network for ILSVRC. ResNets are more computationally efficient than the VGG architecture [26] on which they are based, due to the use of low-dimensional embeddings [19]. ResNets are also more accurate and quicker to converge due to the use of identity mappings.

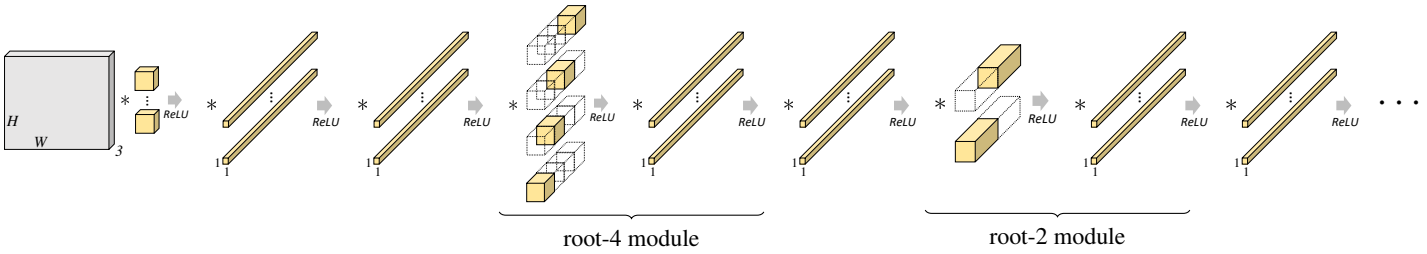
4.3.1 ResNet 50

As a baseline, we used the 'ResNet 50' model [20] (the largest residual network model to fit onto 8 GPUs with Caffe). ResNet 50 has 50 convolutional layers, of which one-third are spatial convolutions (non-1×1). We did not use any training augmentation aside from random cropping and mirroring. For training, we used the initialization scheme described by [24] modified for compound layers [9] and batch normalization [25]. To assess the efficacy of our method, we replaced the spatial convolutional layers of the original network with root modules (as described in §3). We preserved the original number of filters per layer but subdivided them into groups as shown in Table 3. We considered the first of the existing 1×1 layers subsequent to each spatial convolution to be part of our root modules.

Results are shown in Table 4 and Fig. 8 for various network architectures. Compared to the baseline architecture, the root variants achieve a significant reduction in compu-



(a) Standard



(b) Root-4 Architecture

Figure 7: **Network-in-Network Root Architecture.** The Root-4 architecture as compared to the original architecture for all the convolutional layers. Colored blocks represent the filters of each layer. Here we don't show the intermediate feature maps over which a layer's filters operate, or the final fully connected layer, out of space considerations (see Fig.4). The decreasing degree of grouping in successive root modules means that our network architectures somewhat resemble plant roots, hence the name root.

Table 4: **ResNet 50 Results.**

Model	FLOPS $\times 10^9$	Param. $\times 10^7$	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

tation and model size without a significant reduction in accuracy. For example, the best result (root-16) exceeds the baseline accuracy by 0.2% while reducing the model size by 27% and floating-point operations (multiply-add) by 37%. CPU timings were 23% faster, while GPU timings were 13% faster. With a drop in accuracy of only 0.1% however, the root-64 model reduces the model size by 40%, and reduces the floating point operations by 45%. CPU timings were 31% faster, while GPU timings were 12% faster.

4.3.2 ResNet 200

To show that the method applies to deeper architectures, we also applied our method to ResNet 200, the deepest network

Table 5: **ResNet-200 Results**

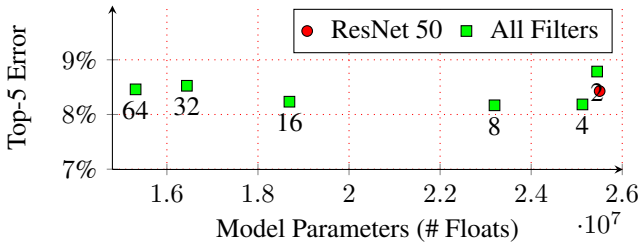
Model	FLOPS $\times 10^{12}$	Param. $\times 10^7$	Top-1 Err.	Top-5 Err.
Orig.	5.65	6.25	**0.2196**	**0.0623**
root-8	4.84	4.91	0.2205	0.0626
root-32	**4.23**	**3.51**	0.2207	0.0630

for ILSVRC 2012. To provide a baseline we used code implementing full training augmentation to achieve state-of-the-art results[2]. Table 5 shows the results of these experiments, top-1 and top-5 error are for center cropped images. The models trained with roots have comparable error to the baseline network, with fewer parameters and less computation. The root-32 model has 25% fewer FLOPS and 44% fewer parameters than ResNet 200.

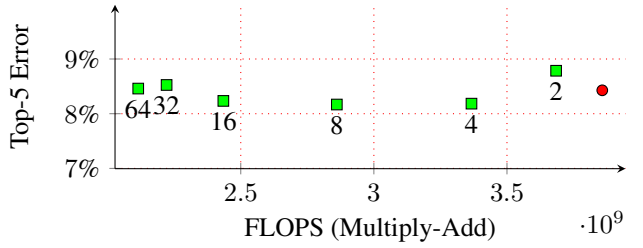
4.4. Improving GoogLeNet on ILSVRC

We replicated the network as described by Szegedy et al. [2], with the exception of not using any training augmentation aside from random crops and mirroring (as supported by Caffe [27]). To train we used the initialization of [24] modified for compound layers [9] and batch normalization without the scale and bias [25]. At test time we only evaluate the center crop image.

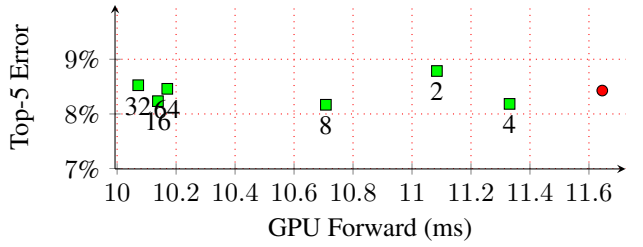
[2] https://github.com/facebook/fb.resnet.torch



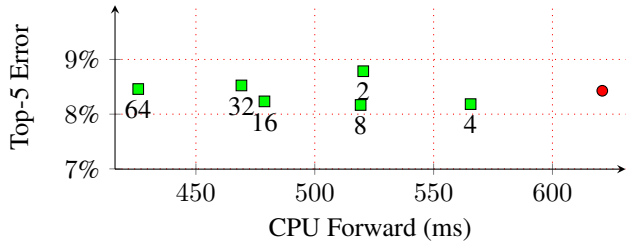
(a) **Model Parameters *vs*. Top-5 Error.**



(b) **FLOPS (Multiply-Add) *vs*. Top-5 Error.**



(c) **GPU Forward Time *vs*. Top-5 Error.**



(d) **CPU Forward Time *vs*. Top-5 Error.**

Figure 8: **ResNet-50 Results.** Models with filter groups have fewer parameters, and less floating point operations, while maintaining error comparable to the baseline.

While preserving the original number of filters per layer, we trained networks with various degrees of filter grouping, as described in Table 7. While the inception architecture is relatively complex, for simplicity, we always use the same number of groups within each of the groups of different filter sizes, despite them having different cardinality. For all of the networks, we only grouped filters within each of the 'spatial' convolutions (3×3, 5×5).

As shown in Table 6, and plotted in Fig. 9, our method shows significant reduction in computational complexity – as measured in FLOPS (multiply-adds), CPU and GPU timings – and model size, as measured in the number of floating point parameters. For many of the configurations the top-5 accuracy remains within 0.5% of the baseline model. The highest accuracy result, is 0.1% off the top-5 accuracy of the baseline model, but has a 0.1% higher top-1 accuracy – within the error bounds resulting from training with different random initializations. While maintaining the same accuracy, this network has 9% faster CPU and GPU timings. However, a model with only 0.3% lower top-5 accuracy than the baseline has much higher gains in computational efficiency – 44% fewer floating point operations (multiply-add), 7% fewer model parameters, 21% faster CPU and 16% faster GPU timings.

While these results may seem modest compared to the results for ResNet, GoogLeNet is by far the smallest and fastest near state-of-the-art model ILSVRC model. We believe that more experimentation in using different cardinalities of filter grouping in the heterogeneously-sized filter groups within each inception module will improve results further.

Table 6: **GoogLeNet Results.**

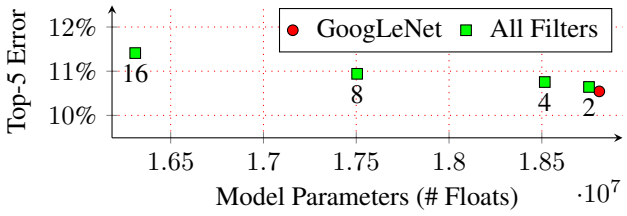
Model	FLOPS $\times 10^9$	Param. $\times 10^7$	Top-1 Acc.	Top-5 Acc.	CPU (ms)	GPU (ms)
Orig.	1.72	1.88	0.694	**0.894**	315	4.39
root-2	1.54	1.88	**0.695**	0.893	285	4.37
root-4	1.29	1.85	0.693	0.892	273	4.10
root-8	0.96	1.75	0.691	0.891	246	3.72
root-16	**0.76**	**1.63**	0.683	0.886	207	3.59

Table 7: **GoogLeNet**. Filter groups in each convolutional layer and Inception module (*incp.*)

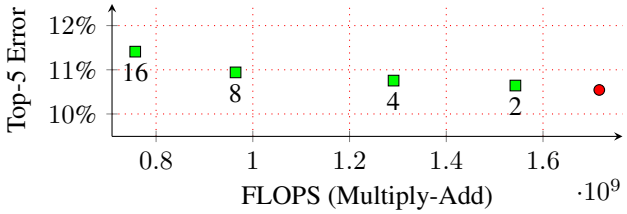
Model	conv1	conv2		incp. 3{a,b}			incp. 4{a–e}			incp. 5{a,b}		
	7×7	*1×1*	*3×3*	*1×1*	*3×3*	*5×5*	*1×1*	*3×3*	*5×5*	*1×1*	*3×3*	*5×5*
Orig.	1	1	1	1	1	1	1	1	1	1	1	1
root-2	1	1	2	1	1	1	1	1	1	1	1	1
root-4	1	1	4	1	2	2	1	1	1	1	1	1
root-8	1	1	8	1	4	4	1	2	2	1	1	1
root-16	1	1	16	1	8	8	1	4	4	1	2	2

5. GPU Implementation

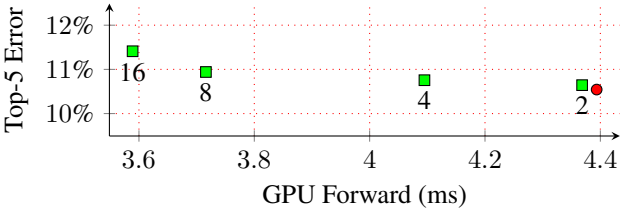
Our experiments show that our method can achieve a significant reduction in CPU and GPU runtimes for state-of-the-art CNNs without compromising accuracy. However, the reductions in GPU runtime were smaller than might have been expected based on theoretical predictions of computational complexity (FLOPs). We believe this is largely a consequence of the optimization of Caffe for existing net-



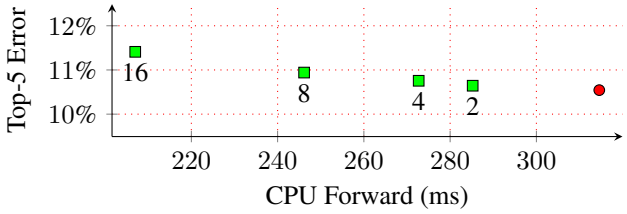
(a) **Model Parameters *vs*. Top-5 Error.**



(b) **FLOPS (Multiply-Add) *vs*. Top-5 Error.**



(c) **GPU Forward Time *vs*. Top-5 Error.**



(d) **CPU Forward Time *vs*. Top-5 Error.**

Figure 9: **GoogLeNet Results.** Models with filter groups have fewer parameters, and less floating point operations, while maintaining error comparable to the baseline.

work architectures (particularly AlexNet and GoogLeNet) that do not use a high degree of filter grouping.

Caffe presently parallelizes over filter groups by using multiple CUDA streams to run multiple CuBLAS matrix multiplications simultaneously. However, with a large degree of filter grouping, and hence more, smaller matrix multiplications, the overhead associated with calling CuBLAS from the host can take approximately as long as the matrix computation itself. To avoid this overhead, CuBLAS provides batched methods (*e.g.* `cublasXgemmBatched`), where many small matrix multiplications can be batched together in one call. Jhurani and Mullowney [28] explore in depth the problem of using GPUs to accelerate the multiplication of very small matrices (smaller than 16×16), and show it is possible to achieve high throughput with large batches, by implementing a more efficient interface than that used in the CuBLAS batched calls. We have modified Caffe to use CuBLAS batched calls, and achieved significant speedups for our root-like network architectures compared to vanilla Caffe without CuDNN, e.g. a 25% speed up on our root-16 modified version of the GoogleNet architecture. However, our optimized implementation still is not as fast as Caffe with CuDNN (which was used to generate the results in this paper), presumably because of other unrelated optimizations in the (proprietary) CuDNN library. Therefore we suggest that direct integration of CuBLAS-style batching into CuDNN could improve the performance of filter groups significantly.

6. Future Work

In this paper we focused on using homogeneous filter groups (with a uniform division of filters in each group), however this may not be optimal. Heterogeneous filter groups may reflect better the filter co-dependencies found in deep networks. Learning a combined spatial [9] and channel basis, may also improve efficiency further.

7. Conclusion

We explored the effect of using complex hierarchical arrangements of filter groups in CNNs and show that imposing a structured decrease in the degree of filter grouping with depth – a 'root' (inverse tree) topology – can allow us to obtain more efficient variants of state-of-the-art networks without compromising accuracy. Our method appears to be complementary to existing methods, such as low-dimensional embeddings, and can be used more efficiently to train deep networks than methods that only approximate a pre-trained model's weights.

We validated our method by using it to create more efficient variants of state-of-the-art Network-in-network, GoogLeNet, and ResNet architectures, which were evaluated on the CIFAR10 and ILSVRC datasets. Our results show similar accuracy with the baseline architecture with fewer parameters and much less compute (as measured by CPU and GPU timings). For Network-in-Network on CIFAR10, our model has 33% of the parameters of the original network, and approximately 37% (23%) faster CPU (GPU) timings. For ResNet 50, our model has 27% fewer parameters, and was 24% (11%) faster on a CPU (GPU). Even for the most efficient of the near state-of-the-art ILSVRC network, GoogLeNet, our model uses 7% fewer parameters and is 21% (16%) faster on a CPU (GPU).

References

[1] M. Denil, B. Shakibi, L. Dinh, M. Ranzato, and N. de Freitas, "Predicting Parameters in Deep Learning,"

in *Neural Information Processing Systems (NIPS)*, 2013, pp. 2148–2156. arXiv:1306.0543 (cit. on p. 1).

[2] C Szegedy, W Liu, Y Jia, and P Sermanet, "Going deeper with convolutions," in *arXiv preprint arXiv: 1409.4842*, 2014, ISBN: 9781467369640. arXiv:1602.07360 (cit. on pp. 1, 2, 6).

[3] E. Denton, W. Zaremba, J. Bruna, Y. LeCun, and R. Fergus, "Exploiting Linear Structure Within Convolutional Networks for Efficient Evaluation," in *arXiv*, 2014, pp. 1–11. arXiv:1404.0736 (cit. on p. 1).

[4] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," in *Advances In Neural Information Processing Systems*, P. L. Bartlett, F. C. N. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, Eds., 2012, pp. 1–9, ISBN: 9781627480031. arXiv:1102.0183 (cit. on pp. 1, 2).

[5] G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. R. Salakhutdinov, *Improving neural networks by preventing co-adaptation of feature detectors*, 2012. arXiv:1207.0580 (cit. on pp. 1, 3).

[6] K Fukushima, "Neocognitron: A self-organizing neural network model for a mechanish of pattern recognition unaffected by shifts in position," *Biological Cybernetics*, vol. 36, pp. 193–202, 1980 (cit. on p. 1).

[7] Y Lecun, L Bottou, Y Bengio, and P Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998, ISSN: 0018-9219 (cit. on p. 1).

[8] M. Minsky and S. Papert, *Perceptrons*. MIT press, 1988 (cit. on p. 1).

[9] 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 (cit. on pp. 1, 3, 5, 6, 8).

[10] F. Mamalet and C. Garcia, "Simplifying convnets for fast learning," in *Artificial Neural Networks and Machine Learning–ICANN 2012*, Springer, 2012, pp. 58–65 (cit. on p. 1).

[11] M. Jaderberg, A. Vedaldi, and A. Zisserman, "Speeding up Convolutional Neural Networks with Low Rank Expansions.," in *British Machine Vision Conference*, 2014 (cit. on pp. 1, 2).

[12] A. Sironi, B. Tekin, R. Rigamonti, V. Lepetit, and P. Fua, "Learning separable filters," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 1, pp. 94–106, 2015, ISSN: 01628828 (cit. on p. 1).

[13] V. Lebedev, Y. Ganin, M. Rakhuba1, I. Oseledets, and V. Lempitsky, "Speeding-Up Convolutional Neural Networks Using Fine-tuned CP-Decomposition," *International Conference on Learning Representations (ICLR)*, vol. abs/1412.6, pp. 1–10, 2015. arXiv:arXiv:1412.6553v2 (cit. on pp. 1, 2).

[14] M. Mathieu, M. Henaff, and Y LeCun, "Fast Training of Convolutional Networks through FFTs," *International Conference on Learning Representations (ICLR)*, pp. 1–9, 2014. arXiv:arXiv:1312.5851v5 (cit. on p. 1).

[15] O. Rippel, J. Snoek, and R. P. Adams, "Spectral Representations for Convolutional Neural Networks," *Advances in Neural Information Processing Systems 28*, pp. 2440–2448, 2015, ISSN: 10495258. arXiv:1506.03767 (cit. on p. 1).

[16] S. Gupta, A. Agrawal, K. Gopalakrishnan, and P. Narayanan, *Deep Learning with Limited Numerical Precision*, 2015. arXiv:1502.02551 (cit. on p. 1).

[17] W. Chen, J. T. Wilson, S. Tyree, K. Q. Weinberger, and Y. Chen, "Compressing Neural Networks with the Hashing Trick," in *Proceedings of The 32nd International Conference on Machine Learning*, F. R. Bach and D. M. Blei, Eds., ser. JMLR Proceedings, vol. 37, JMLR.org, 2015, pp. 2285–2294, ISBN: 9781510810587. arXiv:1504.04788 (cit. on p. 1).

[18] Y.-D. Kim, E. Park, S. Yoo, T. Choi, L. Yang, and D. Shin, "Compression of Deep Convolutional Neural Networks for Fast and Low Power Mobile Applications," in *International Conference on Learning Representations (ICLR)*, 2016, pp. 1–16. arXiv:1511.06530 (cit. on pp. 1, 2).

[19] M. Lin, Q. Chen, and S. Yan, "Network In Network," *arXiv preprint*, vol. abs/1312.4, p. 10, 2013. arXiv:1312.4400 (cit. on pp. 2, 4, 5).

[20] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," *Arxiv.Org*, vol. 7, no. 3, pp. 171–180, 2015, ISSN: 1664-1078. arXiv:1512.03385 (cit. on pp. 2, 4, 5).

[21] M. Cogswell, F. Ahmed, R. B. Girshick, L. Zitnick, and D. Batra, "Reducing Overfitting in Deep Networks by Decorrelating Representations.," in *International Conference on Learning Representations*, 2016 (cit. on p. 3).

[22] A. Krizhevsky, "Learning Multiple Layers of Features from Tiny Images," Univ. Toronto, Technical Report, 2009, pp. 1–60. arXiv:arXiv:1011.1669v3 (cit. on pp. 4, 11).

[23] I. J. Goodfellow, D. Warde-Farley, M. Mirza, A. Courville, and Y. Bengio, "Maxout Networks," in *Proceedings of the 30th International Conference on Machine Learning (ICML)*, vol. 28, 2013, pp. 1319–1327. arXiv:1302.4389 (cit. on p. 4).

[24] K. He, X. Zhang, S. Ren, and J. Sun, "Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification," in *IEEE Conference on Computer Vision and Patern Recognition (ICCV)*, IEEE, 2015, pp. 1026–1034, ISBN: 978-1-4673-8391-2. arXiv:1502.01852 (cit. on pp. 4–6).

[25] S. Ioffe and C. Szegedy, "Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift.," in *Proceedings of the 32 nd International Conference on Machine Learning, Lille, France, 2015*, 2015 (cit. on pp. 4–6).

[26] K Simonyan and A Zisserman, "Very deep convolutional networks for large-scale image recognition," in *eprint ar{X}iv:arXiv:1409.1556v5*, 1409 (cit. on p. 5).

[27] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, "Caffe: Convolutional Architecture for Fast Feature Embedding," *ACM International Conference on Multimedia*, pp. 675–678, 2014, ISSN: 10636919. arXiv:1408.5093 (cit. on p. 6).

[28] C. Jhurani and P. Mullowney, "A GEMM interface and implementation on NVIDIA GPUs for multiple small matrices," *Journal of Parallel and Distributed Computing*, vol. 75, pp. 133–140, 2015 (cit. on p. 8).

A. Appendices

A.1. Full Network-in-Network Results

Table 8: **Network-in-Network CIFAR10**

Model	FLOPS $\times 10^8$	Param. $\times 10^5$	Accuracy	CPU (ms)	GPU (ms)
Orig.	2.22	9.67	0.9211	39.0	0.623
root-2	1.64	7.37	0.9209	31.2	0.551
root-4	1.23	4.55	0.9202	27.6	0.480
root-8	1.03	3.15	**0.9215**	24.4	0.482
root-16	0.93	2.45	0.9167	23.0	0.475
tree-2	1.48	4.88	0.9185	31.4	0.541
tree-4	1.15	3.31	0.9147	29.1	0.535
tree-8	0.99	2.53	0.9171	25.7	0.500
tree-16	**0.91**	**2.14**	0.9168	20.6	0.512
col-2	1.53	5.71	0.9197	28.8	0.568
col-4	1.18	3.73	0.9200	26.1	0.536
col-8	1.01	2.73	0.9192	23.0	0.475
col-16	0.92	2.24	0.9120	22.8	0.494

Table 8 shows the full results for the Network-in-Network experiments on CIFAR10 on various hierarchical network topologies.

A.2. Inter-Layer Covariance

To show the relationships between filters between adjacent convolutional layers, as illustrated in Fig. 1 , we calculate the covariance of the responses from two adjacent featuremaps, the outputs of convolutional layers with c_1 and c_2 filters.

Let $X_i = [\mathbf{x}_{i,1}; \mathbf{x}_{i,2}; \ldots; \mathbf{x}_{i,N}]$ be the matrix of N samples $\mathbf{x}_{i,n}$ from the c_i dimensional featuremap for layer i. We consider each pixel across the two featuremaps to be a sample, and thus each vector $\mathbf{x}_{i,n}$ is a single pixel filter response of dimension c_i. If two featuremaps have different spatial dimensions, due to pooling, we up-sample the smaller featuremap (with nearest neighbor interpolation) such that there are the same number of pixels (and thus samples) in each featuremap.

Given two samples X_1, X_2 with zero mean (*i.e.* mean subtracted) for two adjacent featuremaps, we calculate the inter-layer covariance,

$$cov(X_1, X_2) = E\left[X_1 X_2^{\mathrm{T}}\right], \quad (1)$$

$$= \frac{1}{N-1} X_1 X_2^{\mathrm{T}}. \quad (2)$$

While this shows the covariance between layers, it is conflated with the inherent covariances within X_1 and X_2 from the data (as shown in Fig. 10a). We can more clearly show the covariance between layers by first whitening (using ZCA [22]) the samples in X_1 and X_2. For a covariance matrix,

$$cov(X, X) = \frac{1}{N-1} X X^{\mathrm{T}}, \quad (3)$$

The ZCA whitening transformation is given by,

$$W = \sqrt{N-1}\left(X X^{\mathrm{T}}\right)^{-\frac{1}{2}}. \quad (4)$$

Since the covariance matrix is symmetric, it is easily diagonalizable (*i.e.* PCA),

$$cov(X, X) = \frac{1}{N-1} X X^{\mathrm{T}}, \quad (5)$$

$$= \frac{1}{N-1} P D P^{\mathrm{T}}, \quad (6)$$

$$(7)$$

where P is a orthogonal matrix and D a diagonal matrix. This diagonalization allows a simplified calculation of the whitening transformation (see the derivation in Appendix A of [22]),

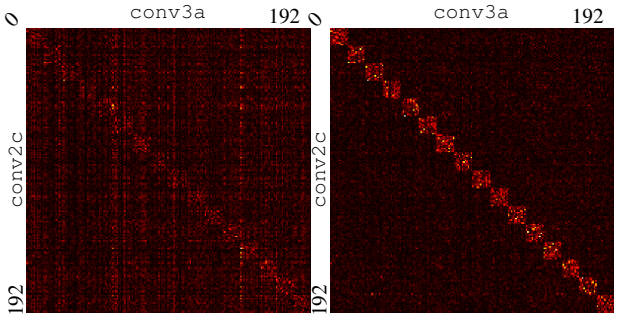
$$W = \sqrt{N-1} P D^{-\frac{1}{2}} P^{\mathrm{T}}, \quad (8)$$

where $D^{-\frac{1}{2}}$ is simply D with an element-wise power of $-\frac{1}{2}$.

The covariance between the whitened featuremap responses is then,

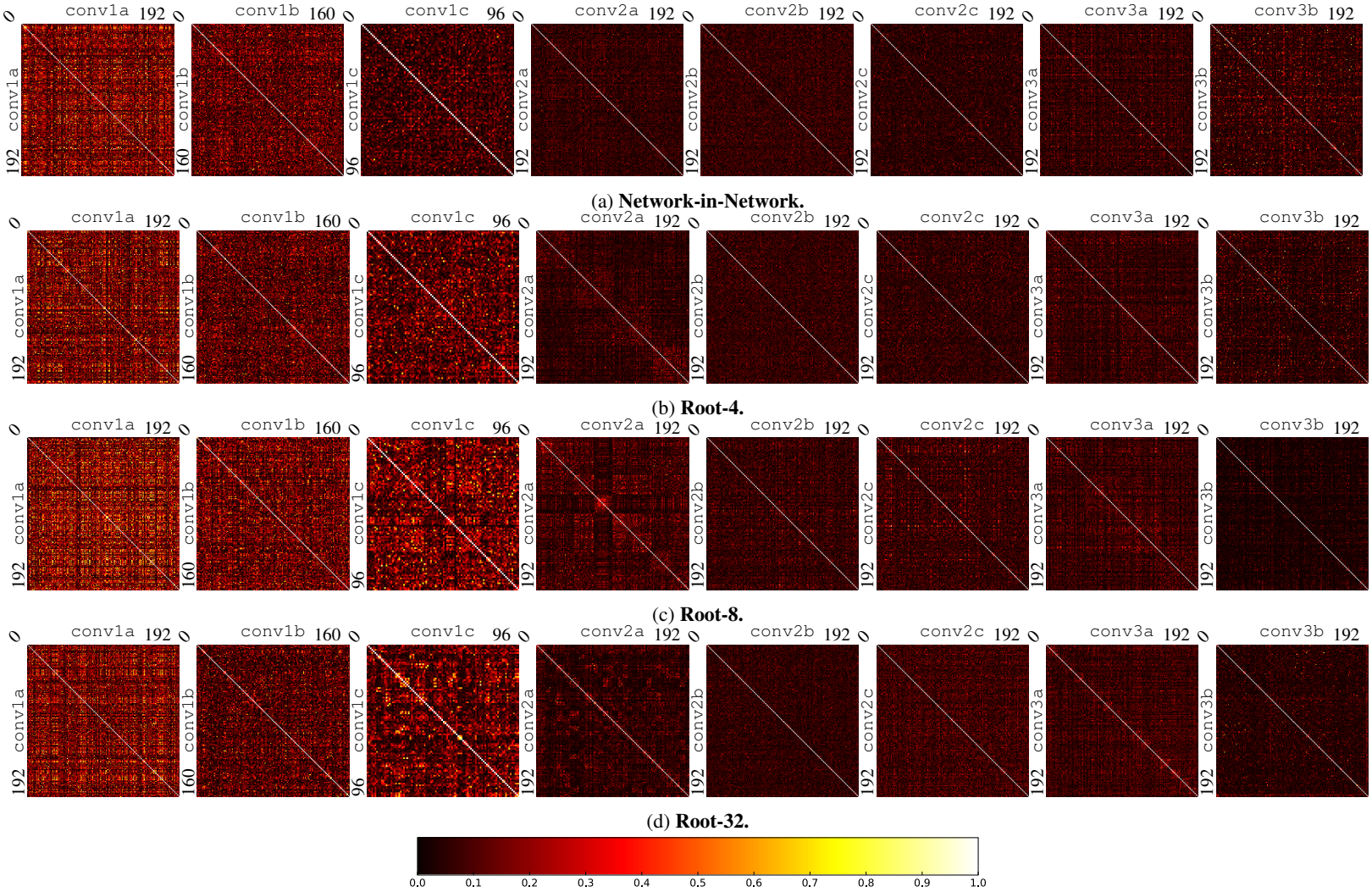
$$cov(W_1 X_1, W_2 X_2) = E\left[(W_1 X_1)(W_2 X_2)^{\mathrm{T}}\right]. \quad (9)$$

Figure 11 shows the per-layer (intra-layer) filter correlation. This shows the correlation of filters is more structured in root-networks, filters are learned to be linearly combined into useful filters by the root module, and thus filters are often grouped together with other filters they correlate strongly with.



(a) Non-whitened responses (b) Whitened responses

Figure 10: Covariance for between two layers in the root-32 Network-in-Network model with and without whitened responses



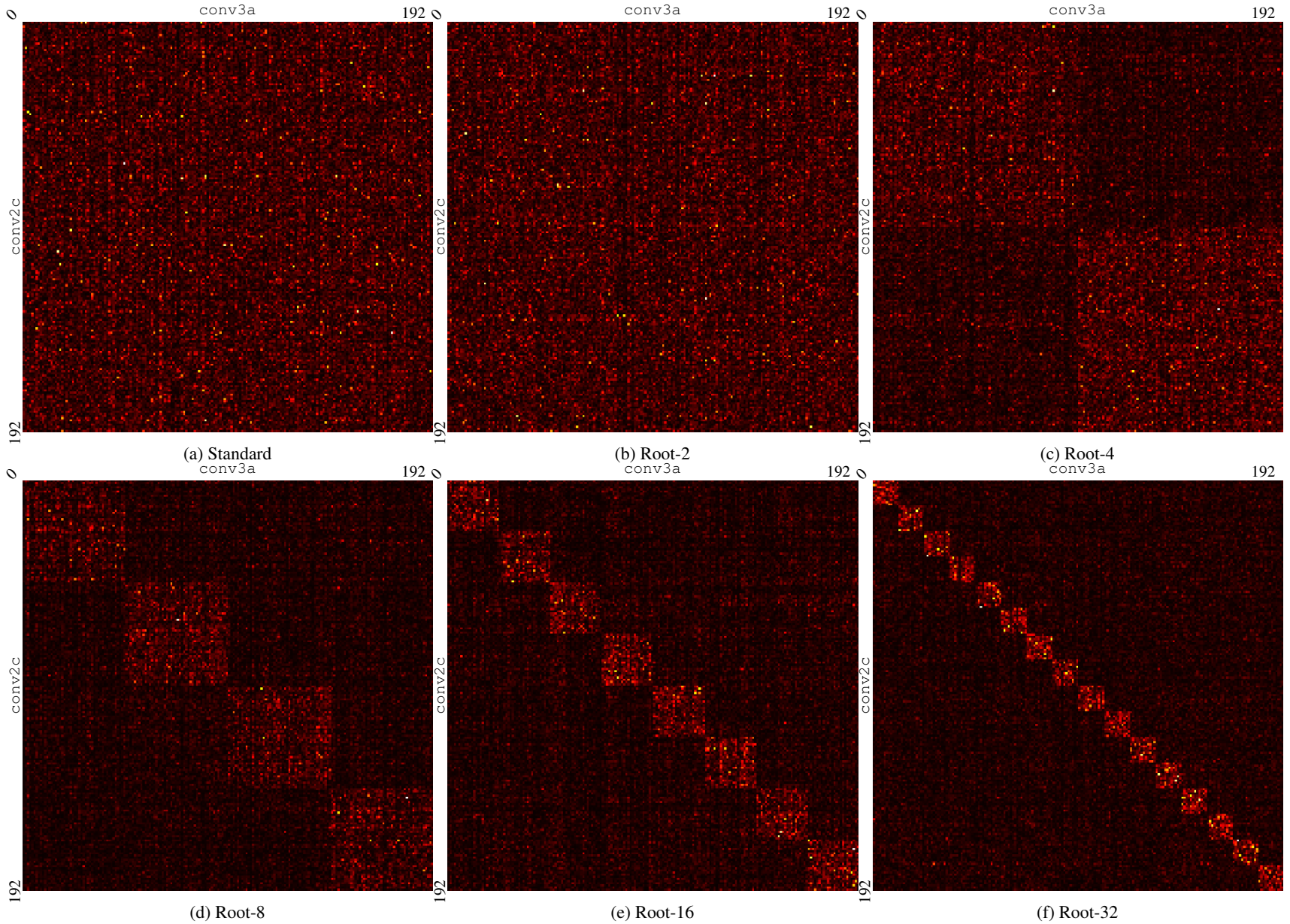
(a) **Network-in-Network.**

(b) **Root-4.**

(c) **Root-8.**

(d) **Root-32.**

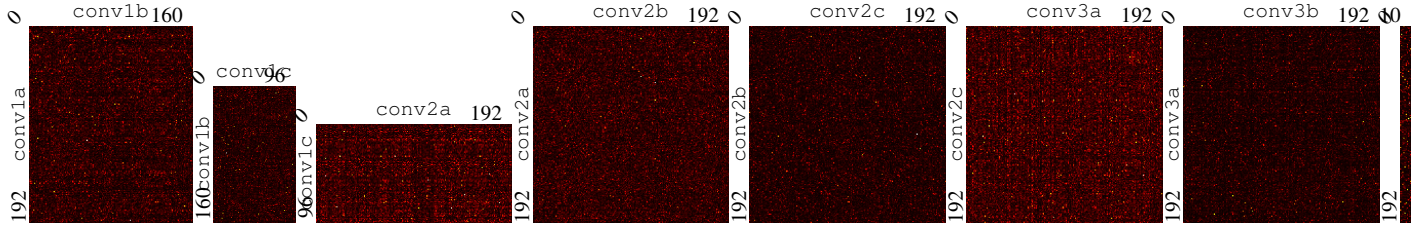
Figure 11: **Network-in-Network Intra-Layer Correlation.** Absolute Correlation of filters within each layer of a NiN model variant.



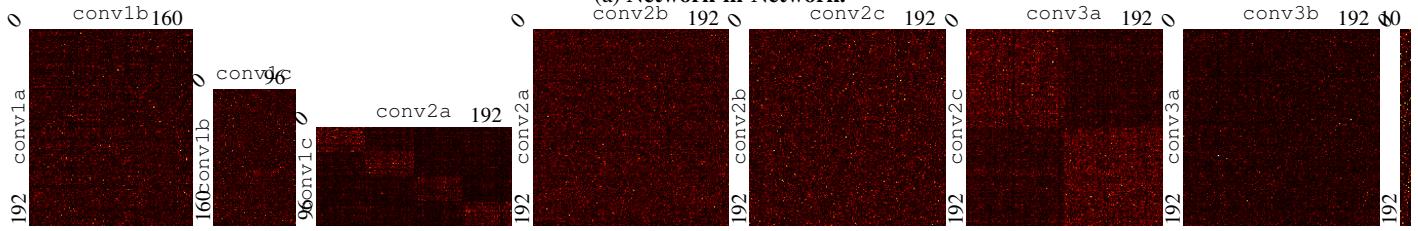
(a) Standard (b) Root-2 (c) Root-4

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

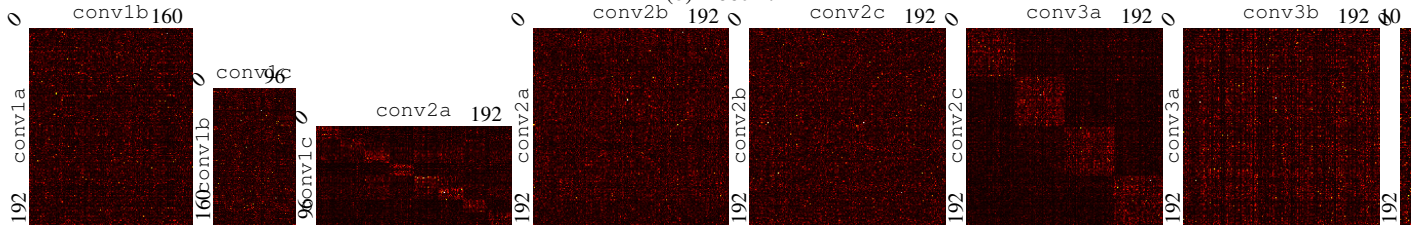
Figure 12: **Filter Inter-layer Covariance conv2c–conv3a.** The block-diagonal sparsity learned by a root-unit is visible in the correlation of filters on layers `conv3a` and `conv2c` in the NiN network.



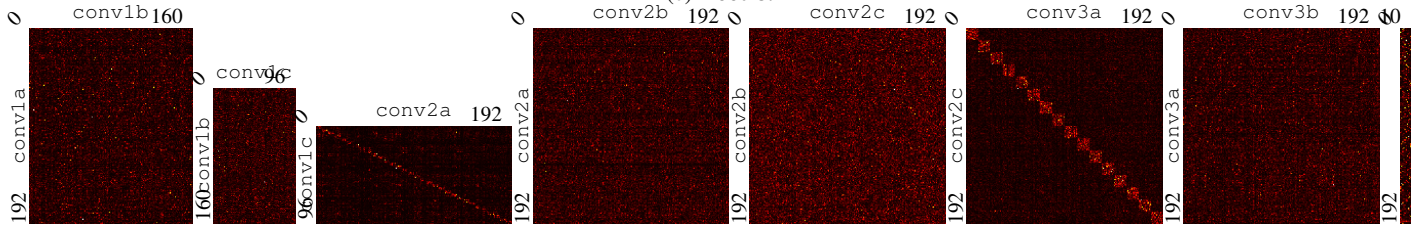
(a) **Network-in-Network.**



(b) **Root-4.**

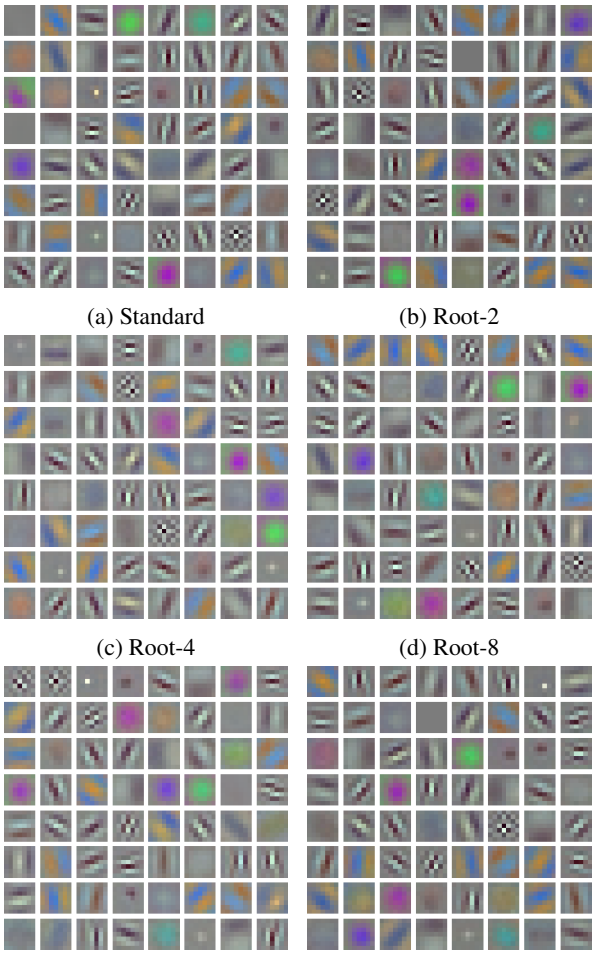


(c) **Root-8.**



(d) **Root-32.**

Figure 13: **Network-in-Network Inter-layer Absolute Covariance.** The inter-layer covariance for all layers in variants of the NiN network.



(a) Standard (b) Root-2

(c) Root-4 (d) Root-8

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

Figure 14: **ResNet 50 `conv1` filters.** With filter groups directly after `conv1`, in `conv2`, some of the organization of filters can be directly observed, and give us intuition as to what is happening in root networks.

Figure 12 shows the complete, enlarged version of Fig. 6, showing the inter-layer filter covariances between layers `conv3a` and `conv2c`. Figure 13 shows the full set of inter-layer covariances between all convolutional layers in the NiN models. Block-diagonal sparsity is visible on the layers with filter groups, `conv2a` and `conv3a`. This block-diagonal is shown for all variants in more detail in Fig. 13.

A.3. The Affect on Image-level Filters of Root Modules

In the ResNet root models, filter groups are used in `conv2`, directly after the image level filters of `conv1` some of the organization of filters can be directly observed, and give us intuition as to what is happening in root networks. Figure 14 shows the `conv0` filters learned for each of the ResNet 50 models. It is apparent that the filters learned in these networks are very similar to those learned in the original model, although sometimes inverted or with a different ordering. This ordering is somewhat consistent in models with filter groups however, even with different random initializations. This is because filter groups cause filters with strong mutual information to be grouped adjacent to each other.

For example, in the root-8 network (Fig. 14d), each row of filters corresponds to the input of an independent filter group in `conv2`. We can see that the first row primarily is composed of filters giving various directions of the same color gradient. These filters can be combined in the next layer to produce color edges easily. Due to the shortcut layer and the learned combinations of filters however, not all filter groupings are so obvious.

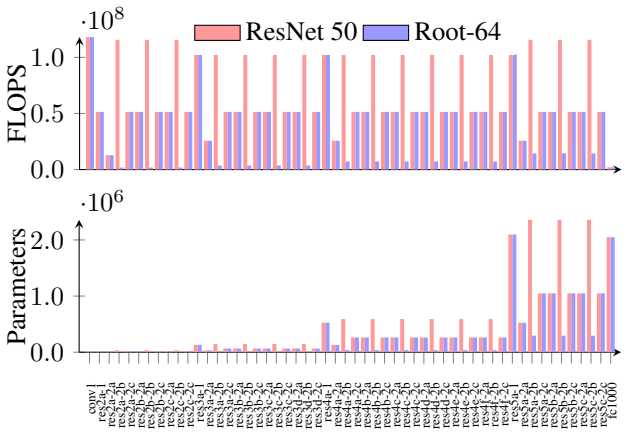


Figure 15: **ResNet 50 Layer-wise FLOPS/Parameters.**

A.4. Layer-wise Compute/Parameter Savings

Figure 15 shows the difference in compute and parameters for each layer in a standard ResNet-50 model and a root-64 variant. The layers in the original networks with the highest computational complexity are clearly the spatial convolutional layers, *i.e.* layers with 3×3 spatial filters. When instead a root-module is used, the computational complexity of these layers is reduced dramatically. While the low dimensional embedding layers (1×1) are not changed, these have less than half the compute of the spatial convolution layers. The number of parameters in spatial convolution layers with large numbers of input channels, which increase towards the end of the network, are similarly reduced.