Interleaved Group Convolutions for Efficient Deep Neural Network

Jingdong Wang
Senior Researcher
Microsoft Research, Beijing, China

Deep learning in the past decade

- Reducing the dimensionality of data with neural networks, Science, 2006
  - Fast learning algorithms for Restricted Boltzmann machine

Reducing the Dimensionality of Data with Neural Networks

G. E. Hinton* and R. R. Salakhutdinov

High-dimensional data can be converted to low-dimensional codes by training a multilayer neural network with a small central layer to reconstruct high-dimensional input vectors. Gradient descent can be used for fine-tuning the weights in such "autoencoder" networks, but this works well only if the initial weights are close to a good solution. We describe an effective way of initializing the weights that allows deep autoencoder networks to learn low-dimensional codes that work much better than principal components analysis as a tool to reduce the dimensionality of data.

Dimensionality reduction facilitates the classification, visualization, communication, and storage of high-dimensional data. A simple and widely used method is principal components analysis (PCA), which finds the directions of greatest variance in the data set and represents each data point by its coordinates along each of these directions. We describe a nonlinear generalization of PCA that uses an adaptive, multilayer "encoder" network

- ImageNet Classification with deep convolutional neural networks, NIPS, 2012
  - Dramatic performance improvement
  - ImageNet, GPU

ImageNet Classification with Deep Convolutional Neural Networks

Alex Krizhevsky
University of Toronto
kriz@cs.utoronto.ca

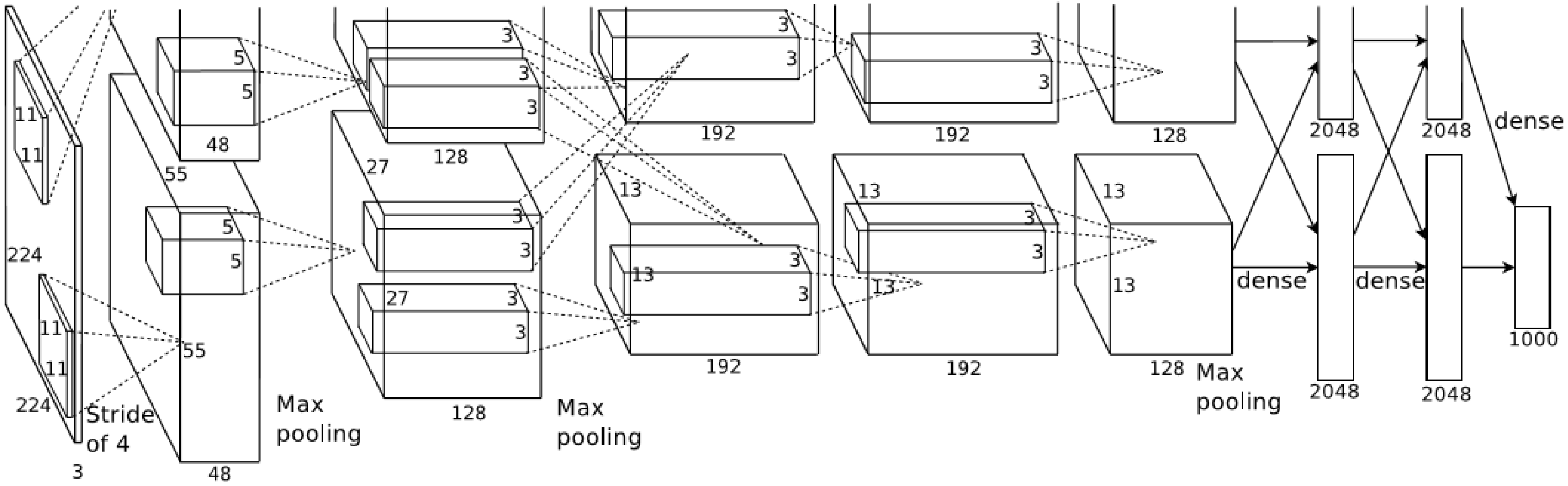
Ilya Sutskever
University of Toronto
ilya@cs.utoronto.ca

Geoffrey E. Hinton
University of Toronto
hinton@cs.utoronto.ca

Abstract

We trained a large, deep convolutional neural network to classify the 1.2 million high-resolution images in the ImageNet LSVRC-2010 contest into the 1000 different classes. On the test data, we achieved top-1 and top-5 error rates of 37.5% and 17.0% which is considerably better than the previous state-of-the-art. The neural network, which has 60 million parameters and 650,000 neurons, consists of five convolutional layers, some of which are followed by max-pooling layers, and three fully-connected layers with a final 1000-way softmax. To make training faster, we used non-saturating neurons and a very efficient GPU implementation of the convolution operation. To reduce overfitting in the fully-connected layers we employed a recently-developed regularization method called "dropout" that proved to be very effective. We also entered a variant of this model in the ILSVRC-2012 competition and achieved a winning top-5 test error rate of 15.3%, compared to 26.2% achieved by the second-best entry.

Deep convolutional neural networks



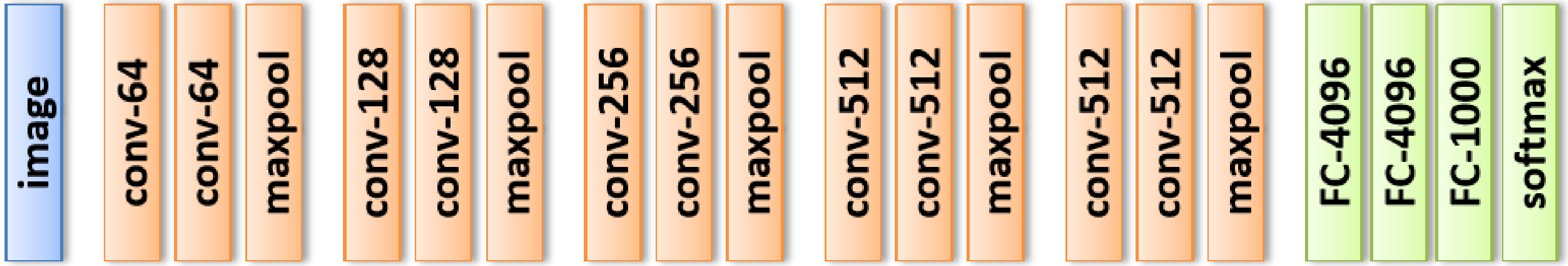
8 layers

AlexNet, 2012

Two architecture design paths

- Going deeper
  - Stack multiple blocks: VGG
  - Improve information flow by skip connections
    - GoogleNet, Highway, ResNet, Deeply-Fused Nets, FractalNets, DenseNets, Merge-and-run

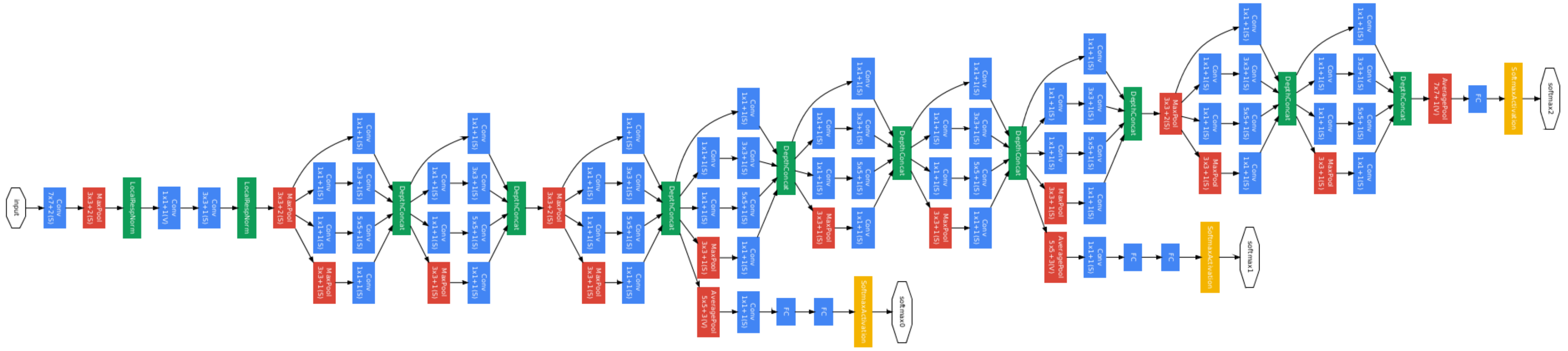
Stack multiple blocks



19 layers

VGGNet, 2014

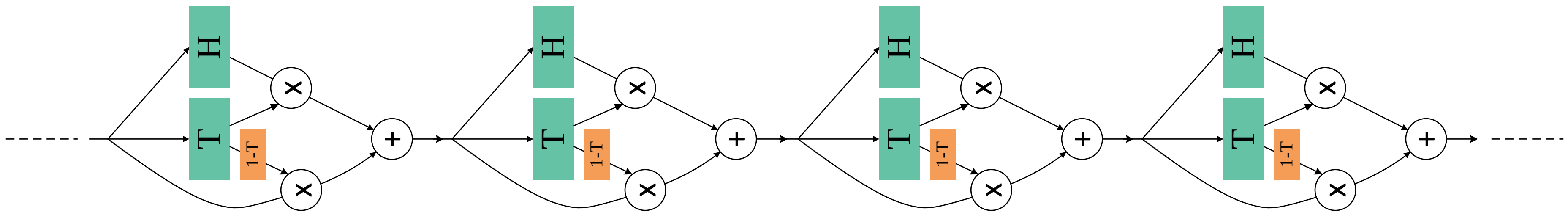
Improve information flow



22 layers

GoogLeNet, 2014

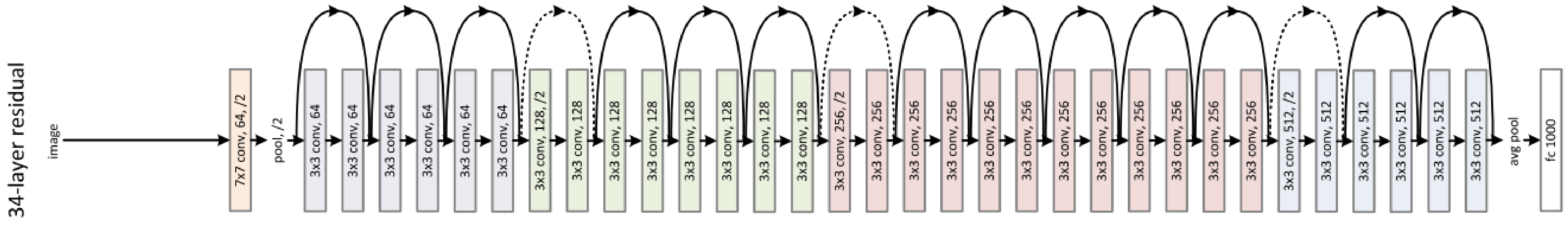
Improve information flow



100+ layers

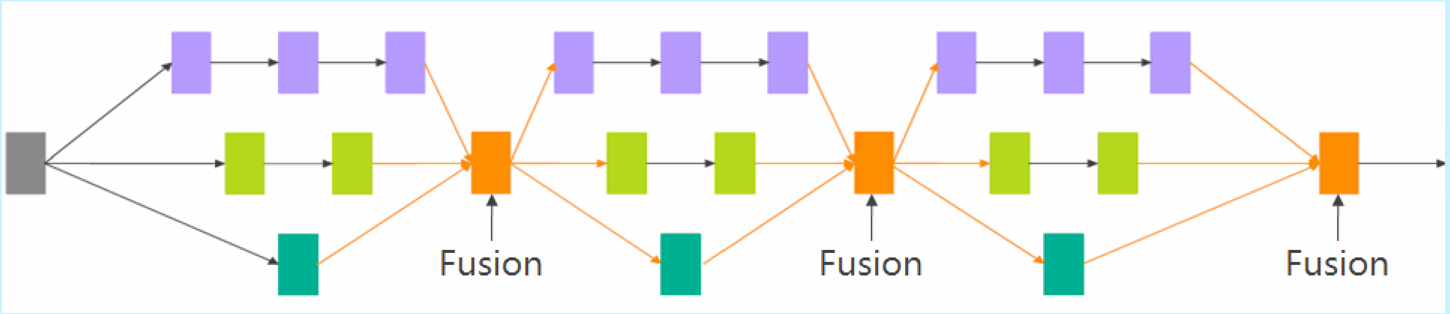
Highway, 2015

Improve information flow

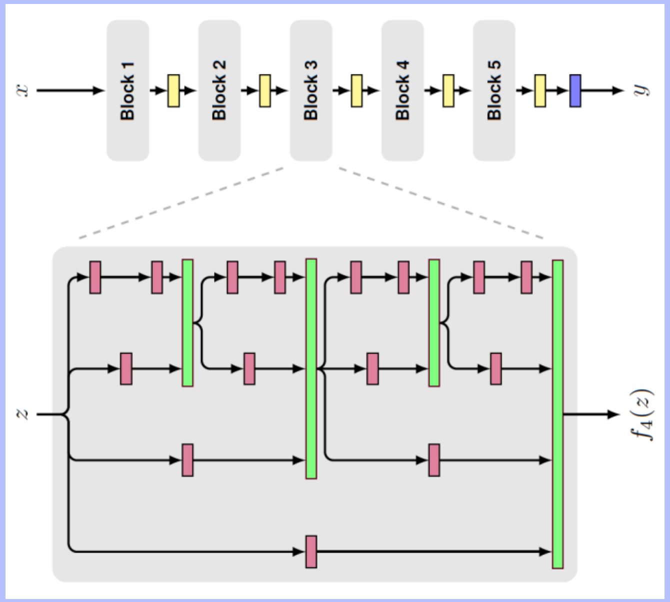


152 layers
ResNet, 2015

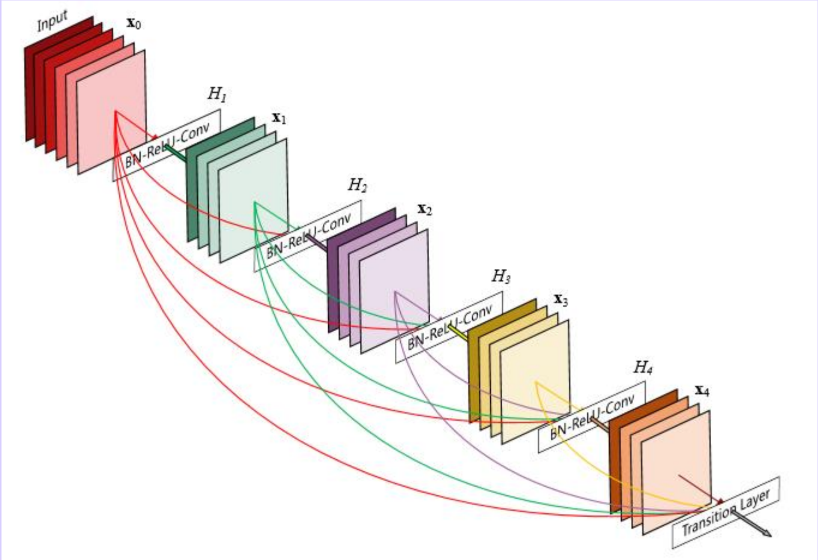
Improve information flow



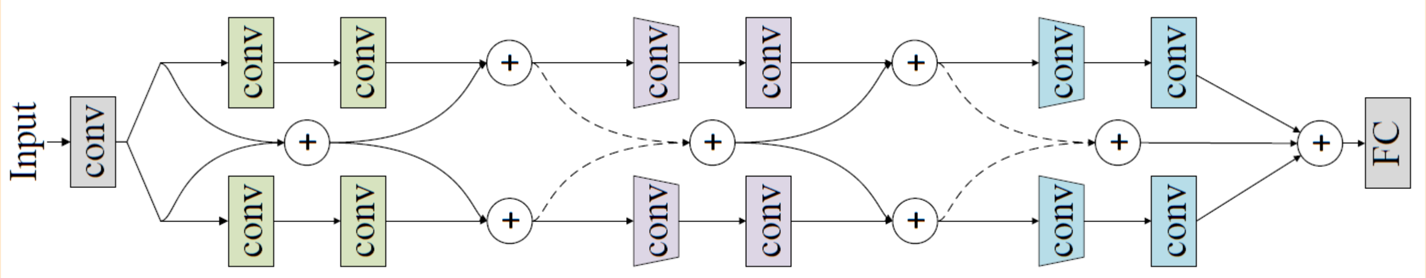
Deeply-fused nets, 2016



FractalNets, 2016



DenseNets, 2016



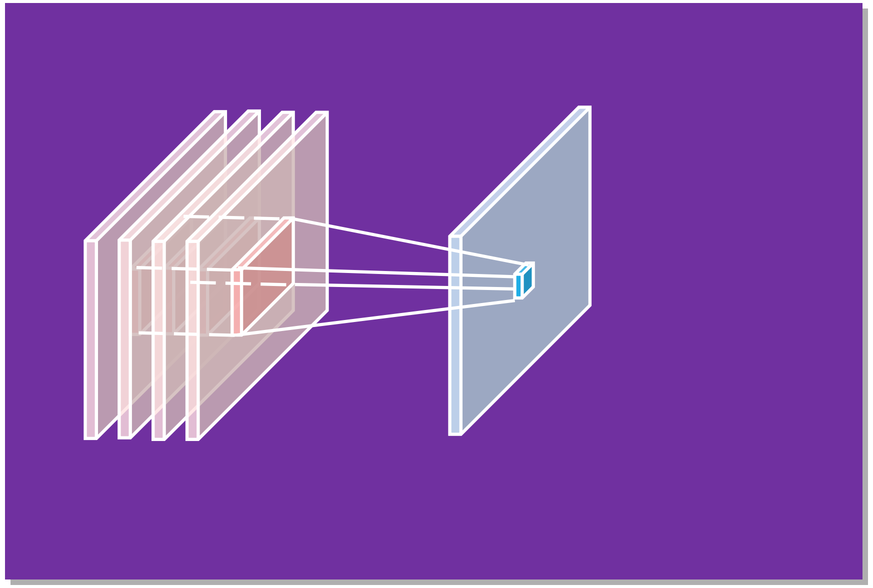
Merge-and-Run Networks (DMRNets), 2016

Two architecture design paths

- Going deeper
  - Stack multiple blocks: VGG
  - Improve information flow by skip connections
    - GoogleNet, Highway, ResNet, Deeply-Fused Nets, FractalNets, DenseNets, Merge-and-run
- Eliminate the redundancy
  - Convolution operations

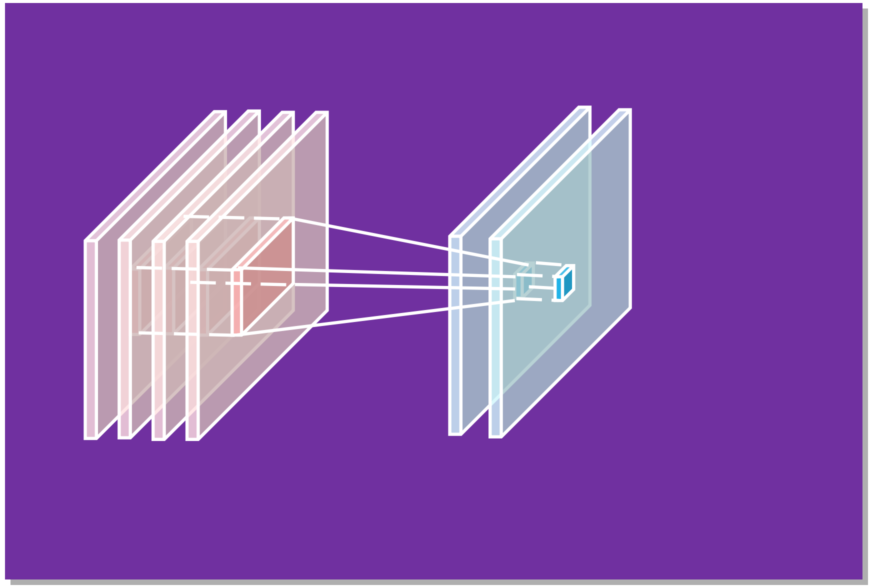
Convolution

$$\begin{pmatrix} 7.39 \\ \\ \\ \end{pmatrix} = \begin{pmatrix} 2.91 & -8.93 & 3.06 & -9.22 & \cdots & 9.01 \\ \\ \\ \end{pmatrix} \times \begin{pmatrix} 0.38 \\ -0.94 \\ 0.47 \\ -0.65 \\ \vdots \\ -0.29 \end{pmatrix}$$



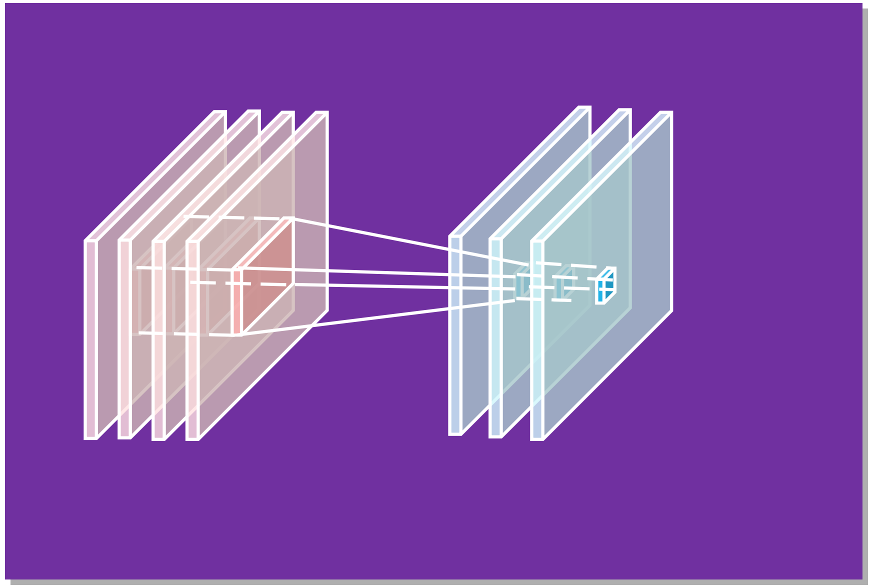
Convolution

$$\begin{pmatrix} 7.39 \\ -4.82 \\ \\ \\ \end{pmatrix} = \begin{pmatrix} 2.91 & -8.93 & 3.06 & -9.22 & \cdots & 9.01 \\ -3.12 & 6.01 & -5.94 & -5.83 & \cdots & -2.75 \\ \\ \\ \end{pmatrix} \times \begin{pmatrix} 0.38 \\ -0.94 \\ 0.47 \\ -0.65 \\ \vdots \\ -0.29 \end{pmatrix}$$



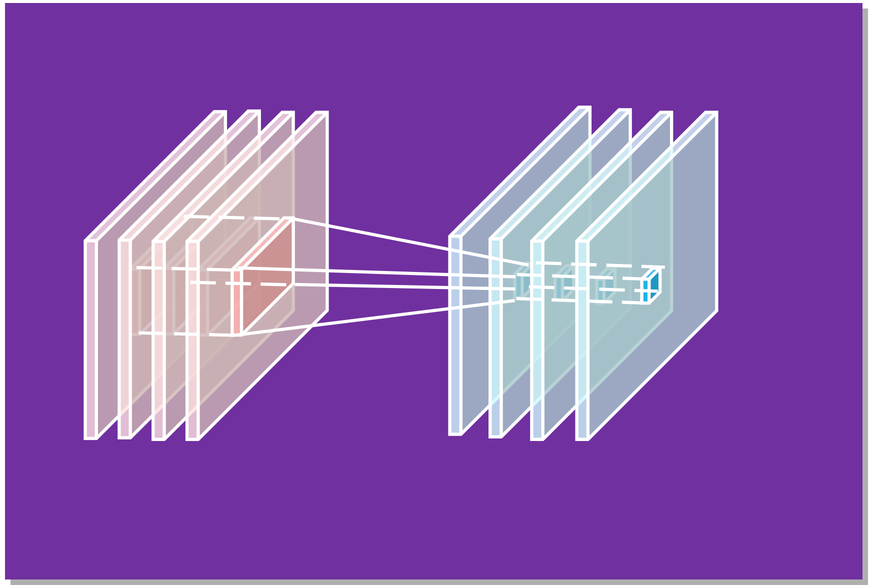
Convolution

$$\begin{pmatrix} 7.39 \\ -4.82 \\ 8.14 \\ \\ \end{pmatrix} = \begin{pmatrix} 2.91 & -8.93 & 3.06 & -9.22 & \cdots & 9.01 \\ -3.12 & 6.01 & -5.94 & -5.83 & \cdots & -2.75 \\ -9.13 & -5.82 & 8.78 & 6.23 & \cdots & -8.82 \\ \\ \end{pmatrix} \times \begin{pmatrix} 0.38 \\ -0.94 \\ 0.47 \\ -0.65 \\ \vdots \\ -0.29 \end{pmatrix}$$



Convolution

$$\begin{pmatrix} 7.39 \\ -4.82 \\ 8.14 \\ -5.27 \end{pmatrix} = \begin{pmatrix} 2.91 & -8.93 & 3.06 & -9.22 & \cdots & 9.01 \\ -3.12 & 6.01 & -5.94 & -5.83 & \cdots & -2.75 \\ -9.13 & -5.82 & 8.78 & 6.23 & \cdots & -8.82 \\ 9.04 & 3.21 & -6.15 & -2.94 & \cdots & 5.94 \end{pmatrix} \times \begin{pmatrix} 0.38 \\ -0.94 \\ 0.47 \\ -0.65 \\ \vdots \\ -0.29 \end{pmatrix}$$



Two architecture design paths

- Going deeper
  - Stack multiple blocks: VGG
  - Improve information flow by skip connections
    - GoogleNet, Highway, ResNet, Deeply-Fused Nets, FractalNets, DenseNets, Merge-and-run
- Eliminate the redundancy
  - Convolution operations
  - Low-precision kernels

Low-precision kernels

- Binarization
- Integer
- Quantization

$$\begin{pmatrix} 2.91 & -8.93 & 3.06 & -9.22 & -1.56 & 9.01 \\ -3.12 & 6.01 & -5.94 & -5.83 & 6.90 & -2.75 \\ -9.13 & -5.82 & 8.78 & 6.23 & -8.38 & -8.82 \\ 9.04 & 3.21 & -6.15 & -2.94 & 5.63 & 5.94 \end{pmatrix} \times \begin{pmatrix} 0.38 \\ -0.94 \\ 0.47 \\ -0.65 \\ 0.46 \\ -0.29 \end{pmatrix}$$

$$\Downarrow$$

$$\begin{pmatrix} 1 & -1 & 1 & -1 & -1 & 1 \\ -1 & 1 & -1 & -1 & 1 & -1 \\ -1 & -1 & 1 & 1 & -1 & -1 \\ 1 & 1 & -1 & -1 & 1 & 1 \end{pmatrix} \times \begin{pmatrix} 0.38 \\ -0.94 \\ 0.47 \\ -0.65 \\ 0.46 \\ -0.29 \end{pmatrix}$$

Low-precision kernels

- Binarization
- Integer
- Quantization

$$\begin{pmatrix} 2.91 & -8.93 & 3.06 & -9.22 & -1.56 & 9.01 \\ -3.12 & 6.01 & -5.94 & -5.83 & 6.90 & -2.75 \\ -9.13 & -5.82 & 8.78 & 6.23 & -8.38 & -8.82 \\ 9.04 & 3.21 & -6.15 & -2.94 & 5.63 & 5.94 \end{pmatrix} \times \begin{pmatrix} 0.38 \\ -0.94 \\ 0.47 \\ -0.65 \\ 0.46 \\ -0.29 \end{pmatrix}$$

$$\downarrow$$

$$\begin{pmatrix} 2 & -8 & 3 & -9 & -1 & 9 \\ -3 & 6 & -5 & -5 & 6 & -2 \\ -9 & -5 & 8 & 6 & -8 & -8 \\ 9 & 3 & -6 & -2 & 5 & 5 \end{pmatrix} \times \begin{pmatrix} 0.38 \\ -0.94 \\ 0.47 \\ -0.65 \\ 0.46 \\ -0.29 \end{pmatrix}$$

Low-precision kernels

- Binarization
- Integer
- Quantization

$$\begin{pmatrix} 2.91 & -8.93 & 3.06 & -9.22 & -1.56 & 9.01 \\ -3.12 & 6.01 & -5.94 & -5.83 & 6.90 & -2.75 \\ -9.13 & -5.82 & 8.78 & 6.23 & -8.38 & -8.82 \\ 9.04 & 3.21 & -6.15 & -2.94 & 5.63 & 5.94 \end{pmatrix} \times \begin{pmatrix} 0.38 \\ -0.94 \\ 0.47 \\ -0.65 \\ 0.46 \\ -0.29 \end{pmatrix}$$

$$\downarrow$$

$$\begin{pmatrix} 3 & -9 & 3 & -9 & -3 & 9 \\ -3 & 6 & -6 & -6 & 6 & -3 \\ -9 & -6 & 9 & 6 & -9 & -9 \\ 9 & 3 & -6 & -3 & 6 & 6 \end{pmatrix} \times \begin{pmatrix} 0.38 \\ -0.94 \\ 0.47 \\ -0.65 \\ 0.46 \\ -0.29 \end{pmatrix}$$

Two architecture design paths

- Going deeper
  - Stack multiple blocks: VGG
  - Improve information flow by skip connections
    - GoogleNet, Highway, ResNet, Deeply-Fused Nets, FractalNets, DenseNets, Merge-and-run
- Eliminate the redundancy
  - Convolution operations
  - Low-precision kernels
  - Low-rank kernels

Low-rank kernels

- Filter pruning
- Channel pruning

$$\begin{pmatrix} 7.39 \\ -4.82 \\ 8.14 \\ -5.27 \end{pmatrix} = \begin{pmatrix} 2.91 & -8.93 & 3.06 & -9.22 & \cdots & 9.01 \\ -3.12 & 6.01 & -5.94 & -5.83 & \cdots & -2.75 \\ -9.13 & -5.82 & 8.78 & 6.23 & \cdots & -8.82 \\ 9.04 & 3.21 & -6.15 & -2.94 & \cdots & 5.94 \end{pmatrix} \times \begin{pmatrix} 0.38 \\ -0.94 \\ 0.47 \\ -0.65 \\ \vdots \\ -0.29 \end{pmatrix}$$

Low-rank kernels

- Filter pruning
- Channel pruning

$$\begin{pmatrix} 7.39 \\ -4.82 \\ \cancel{8.14} \\ \cancel{-5.27} \end{pmatrix} = \begin{pmatrix} 2.91 & -8.93 & 3.06 & -9.22 & \cdots & 9.01 \\ -3.12 & 6.01 & -5.94 & -5.83 & \cdots & -2.75 \\ \cancel{-9.13} & \cancel{-5.82} & \cancel{8.78} & \cancel{6.23} & \cancel{\cdots} & \cancel{-8.82} \\ \cancel{9.04} & \cancel{3.21} & \cancel{-6.15} & \cancel{-2.94} & \cancel{\cdots} & \cancel{5.94} \end{pmatrix} \times \begin{pmatrix} 0.38 \\ -0.94 \\ 0.47 \\ -0.65 \\ \vdots \\ -0.29 \end{pmatrix}$$

Low-rank kernels

- Filter pruning
- Channel pruning

$$\begin{pmatrix} 7.39 \\ -4.82 \\ 8.14 \\ -5.27 \end{pmatrix} = \begin{pmatrix} 2.91 & -8.93 & 3.06 & -9.22 & \cdots & 9.01 \\ -3.12 & 6.01 & -5.94 & -5.83 & \cdots & -2.75 \\ -9.13 & -5.82 & 8.78 & 6.23 & \cdots & -8.82 \\ 9.04 & 3.21 & -6.15 & -2.94 & \cdots & 5.94 \end{pmatrix} \times \begin{pmatrix} 0.38 \\ -0.94 \\ 0.47 \\ -0.65 \\ \vdots \\ -0.29 \end{pmatrix}$$

Low-rank kernels

- Filter pruning

- Channel pruning

$$
\begin{pmatrix}
2.91 & -8.93 & \cancel{3.06} & \cancel{-9.22} & \cdots & 9.01 \\
-3.12 & 6.01 & \cancel{-5.94} & \cancel{-5.83} & \cdots & -2.75 \\
-9.13 & -5.82 & \cancel{8.78} & \cancel{6.23} & \cdots & -8.82 \\
9.04 & 3.21 & \cancel{-6.15} & \cancel{-2.94} & \cdots & 5.94
\end{pmatrix}
\times
\begin{pmatrix}
0.38 \\
-0.94 \\
\cancel{0.47} \\
\cancel{0.65} \\
\vdots \\
-0.29
\end{pmatrix}
$$

Two architecture design paths

- Going deeper
  - Stack multiple blocks: VGG
  - Improve information flow by skip connections
    - GoogleNet, Highway, ResNet, Deeply-Fused Nets, FractalNets, DenseNets, Merge-and-run
- Eliminate the redundancy
  - Convolution operations
  - Low-precision kernels
  - Low-rank kernels
  - Composition from low-rank kernels

Composition from low-rank kernels

$$\begin{pmatrix} 2.91 & -8.93 & 3.06 & -9.22 & -1.56 & 9.01 \\ -3.12 & 6.01 & -5.94 & -5.83 & 6.90 & -2.75 \\ -9.13 & -5.82 & 8.78 & 6.23 & -8.38 & -8.82 \\ 9.04 & 3.21 & -6.15 & -2.94 & 5.63 & 5.94 \end{pmatrix} \times \begin{pmatrix} 0.38 \\ -0.94 \\ 0.47 \\ -0.65 \\ 0.46 \\ -0.29 \end{pmatrix}$$

$$\begin{pmatrix} 1.57 & 2.34 \\ 5.76 & 1.51 \\ -3.78 & 9.03 \\ -7.48 & 5.46 \end{pmatrix} \times \begin{pmatrix} -0.53 & 6.70 & 2.09 & 5.31 & 1.53 & -7.87 \\ 6.22 & 9.16 & 8.12 & -2.69 & -3.48 & 1.55 \end{pmatrix} \times \begin{pmatrix} 0.38 \\ -0.94 \\ 0.47 \\ -0.65 \\ 0.46 \\ -0.29 \end{pmatrix}$$

Two architecture design paths

- Going deeper
  - Stack multiple blocks: VGG
  - Improve information flow by skip connections
    - GoogleNet, Highway, ResNet, Deeply-Fused Nets, FractalNets, DenseNets, Merge-and-run
- Eliminate the redundancy
  - Convolution operations
  - Low-precision kernels
  - Low-rank kernels
  - Composition from low-rank kernels
  - Sparse kernels

Sparse kernels

- Non-structured sparse

- Structured sparse

$$\begin{pmatrix} 2.91 & -8.93 & 3.06 & -9.22 & -1.56 & 9.01 \\ -3.12 & 6.01 & -5.94 & -5.83 & 6.90 & -2.75 \\ -9.13 & -5.82 & 8.78 & 6.23 & -8.38 & -8.82 \\ 9.04 & 3.21 & -6.15 & -2.94 & 5.63 & 5.94 \end{pmatrix} \times \begin{pmatrix} 0.38 \\ -0.94 \\ 0.47 \\ -0.65 \\ 0.46 \\ -0.29 \end{pmatrix}$$

$$\begin{pmatrix} 0 & -8.93 & 0 & -9.22 & 0 & 9.01 \\ 0 & 6.01 & 0 & -5.83 & 6.90 & 0 \\ -9.13 & 0 & 8.78 & 6.23 & -8.38 & -8.82 \\ 9.04 & 0 & -6.15 & 0 & 0 & 0 \end{pmatrix} \times \begin{pmatrix} 0.38 \\ -0.94 \\ 0.47 \\ -0.65 \\ 0.46 \\ -0.29 \end{pmatrix}$$

Sparse kernels

- Non-structured sparse
- Structured sparse

$$\begin{pmatrix} 2.91 & -8.93 & 3.06 & -9.22 & -1.56 & 9.01 \\ -3.12 & 6.01 & -5.94 & -5.83 & 6.90 & -2.75 \\ -9.13 & -5.82 & 8.78 & 6.23 & -8.38 & -8.82 \\ 9.04 & 3.21 & -6.15 & -2.94 & 5.63 & 5.94 \end{pmatrix} \times \begin{pmatrix} 0.38 \\ -0.94 \\ 0.47 \\ -0.65 \\ 0.46 \\ -0.29 \end{pmatrix}$$

$$\Downarrow$$

$$\begin{pmatrix} 2.91 & -8.93 & 3.06 & 0 & 0 & 0 \\ -3.12 & 6.01 & -5.94 & 0 & 0 & 0 \\ 0 & 0 & 0 & 6.23 & -8.38 & -8.82 \\ 0 & 0 & 0 & -2.94 & 5.63 & 5.94 \end{pmatrix} \times \begin{pmatrix} 0.38 \\ -0.94 \\ 0.47 \\ -0.65 \\ 0.46 \\ -0.29 \end{pmatrix}$$

Two architecture design paths

- Going deeper
  - Stack multiple blocks: VGG
  - Improve information flow by skip connections
    - GoogleNet, Highway, ResNet, Deeply-Fused Nets, FractalNets, DenseNets, Merge-and-run
- Eliminate the redundancy
  - Convolution operations
  - Low-precision kernels
  - Low-rank kernels
  - Composition from low-rank kernels
  - Sparse kernels
  - Composition from sparse kernels

Composition from sparse kernels

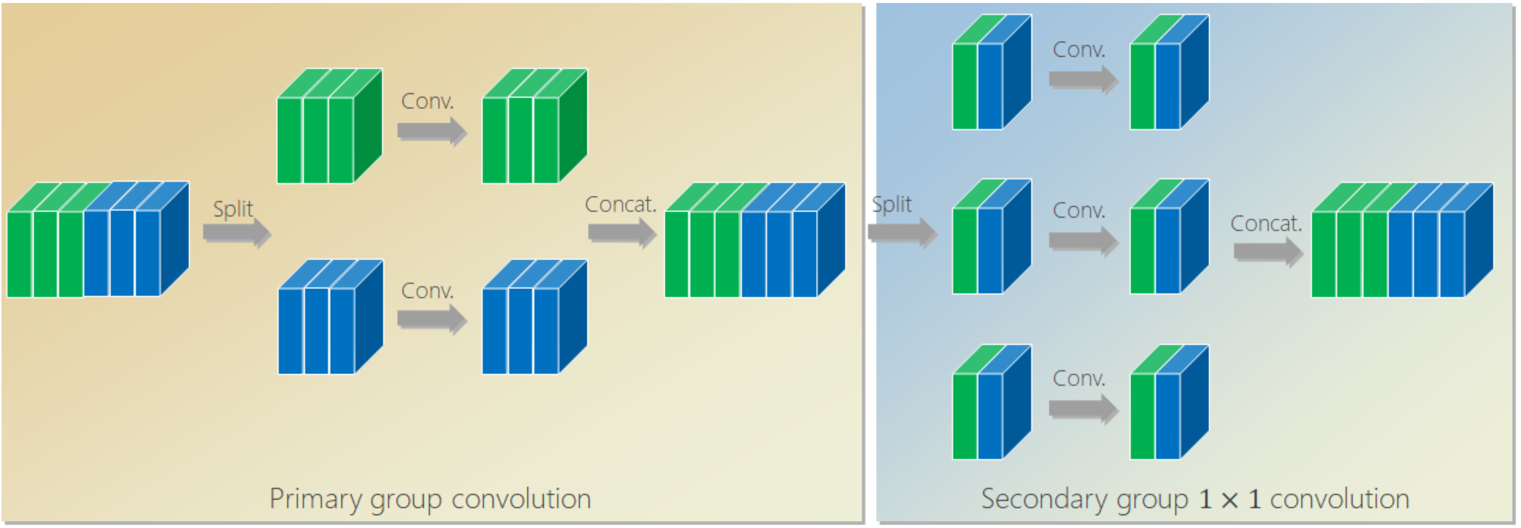
$$\begin{pmatrix} 2.91 & -8.93 & 3.06 & -9.22 & -1.56 & 9.01 \\ -3.12 & 6.01 & -5.94 & -5.83 & 6.90 & -2.75 \\ -9.13 & -5.82 & 8.78 & 6.23 & -8.38 & -8.82 \\ 9.04 & 3.21 & -6.15 & -2.94 & 5.63 & 5.94 \end{pmatrix} \times \begin{pmatrix} 0.38 \\ -0.94 \\ 0.47 \\ -0.65 \\ 0.46 \\ -0.29 \end{pmatrix}$$

$$\begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} -1.03 & -5.25 & 0 & 0 \\ 7.45 & -6.93 & 0 & 0 \\ 0 & 0 & 9.02 & -3.58 \\ 0 & 0 & 6.53 & -1.97 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \times \begin{pmatrix} 4.73 & 6.47 & -5.02 & 0 & 0 & 0 \\ 4.08 & 2.59 & 8.54 & 0 & 0 & 0 \\ 0 & 0 & 0 & -5.85 & 3.87 & 9.01 \\ 0 & 0 & 0 & -4.67 & -3.31 & 5.05 \end{pmatrix} \times \begin{pmatrix} 0.38 \\ -0.94 \\ 0.47 \\ -0.65 \\ 0.46 \\ -0.29 \end{pmatrix}$$

Interleaved group convolutions

[1] Ting Zhang, Guo-Jun Qi, Bin Xiao, Jingdong Wang: Interleaved Group Convolutions. ICCV 2017: 4383-4392
[2] Guotian Xie, Jingdong Wang, Ting Zhang, Jianhuang Lai, Richang Hong, and Guo-JunQi. IGCV2: Interleaved Structured Sparse Convolutional Neural Networks. CVPR 2018.
[3] Ke Sun, Mingjie Li, Dong Liu, and Jingdong Wang. IGCV3: Interleaved Low-Rank Group Convolutions for Efficient Deep Neural Networks. BMVC 2018.

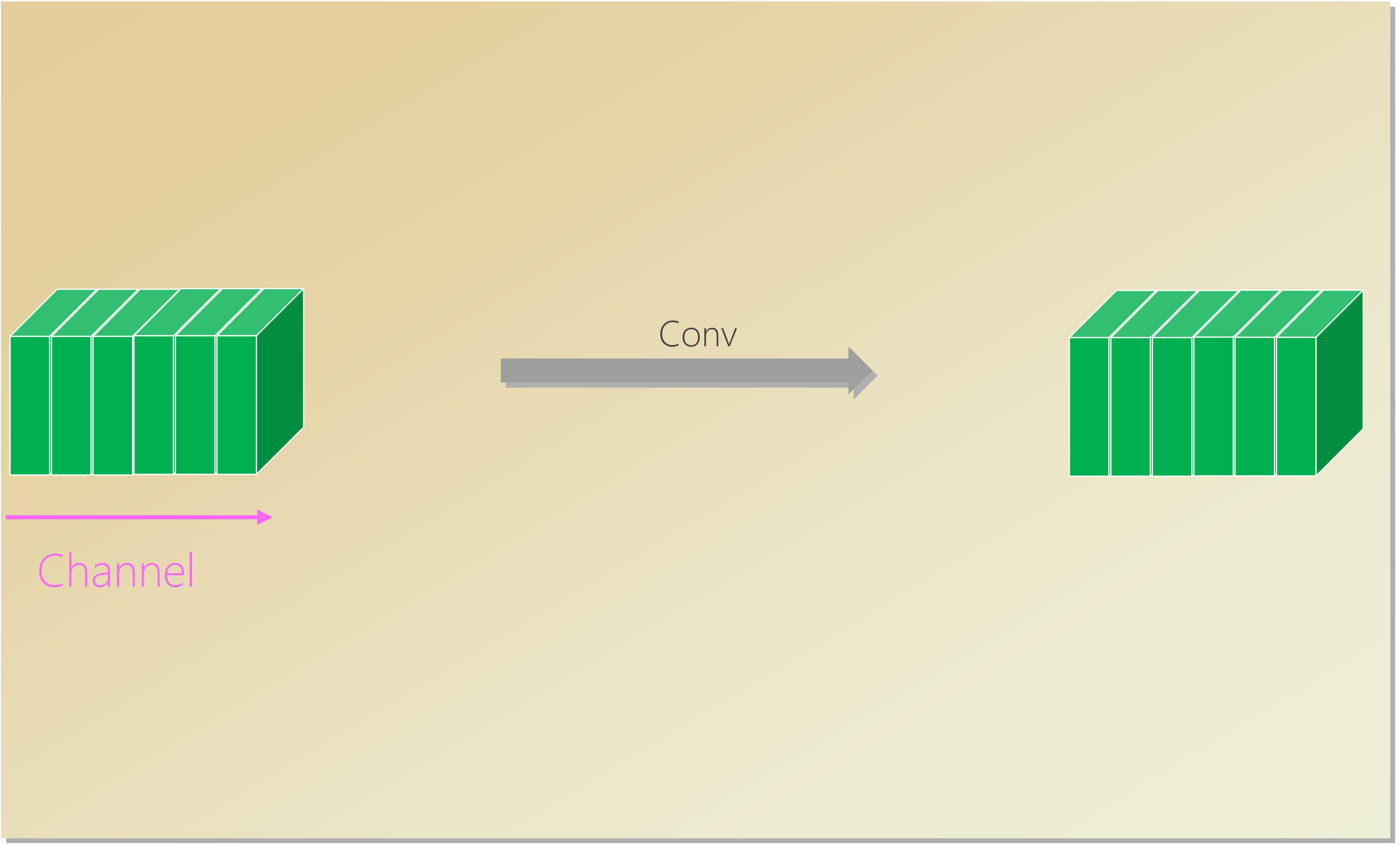
IGCV1: Interleaved group convolutions for large models



Ting Zhang, Guo-Jun Qi, Bin Xiao, Jingdong Wang: Interleaved Group Convolutions, ICCV 2017.

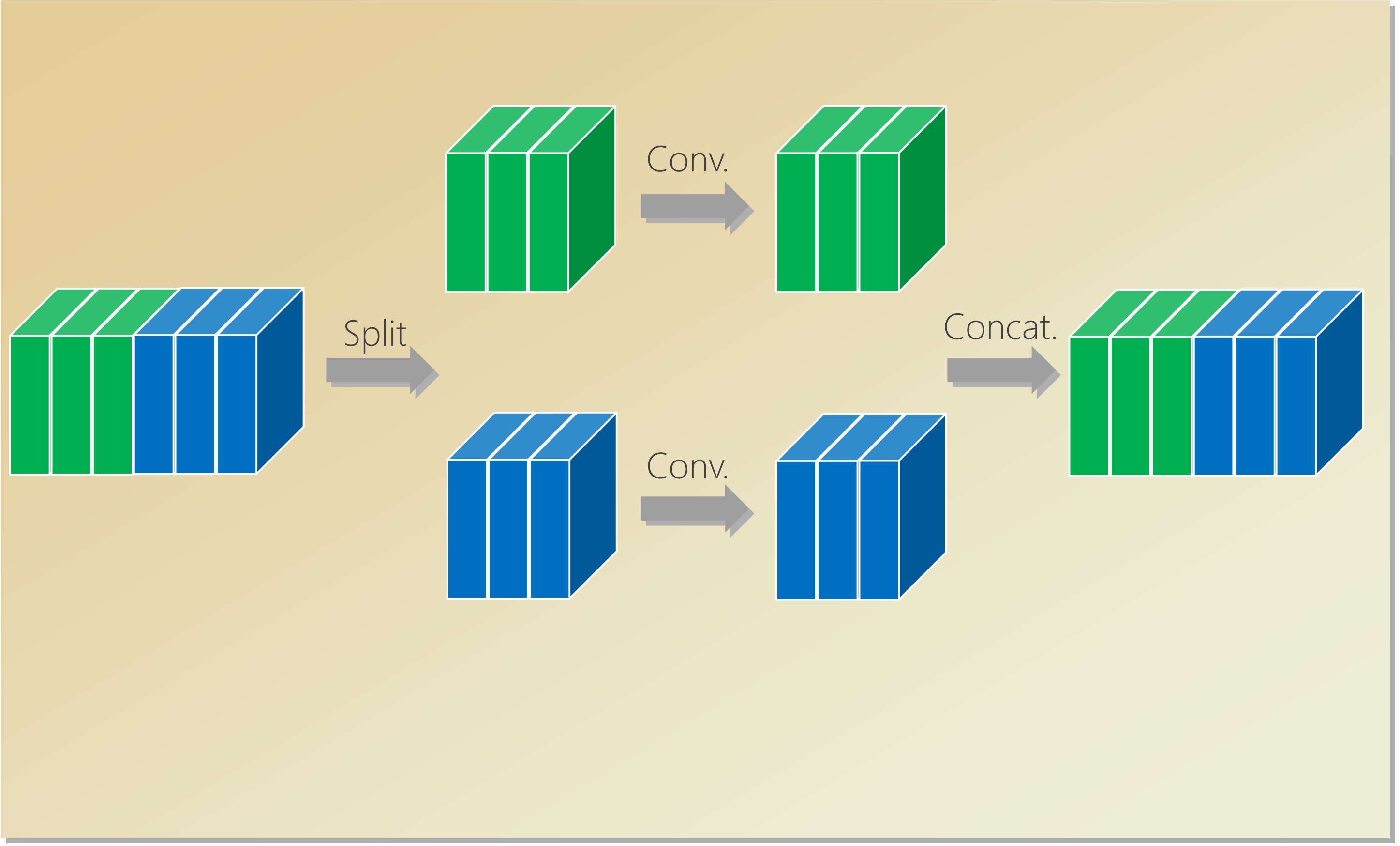
Blog: https://mp.weixin.qq.com/s/PiQB2AvhtDceMJxYN8O8jA

Regular convolution



Complexity: $6 \times 5 \times 5 \times 6$

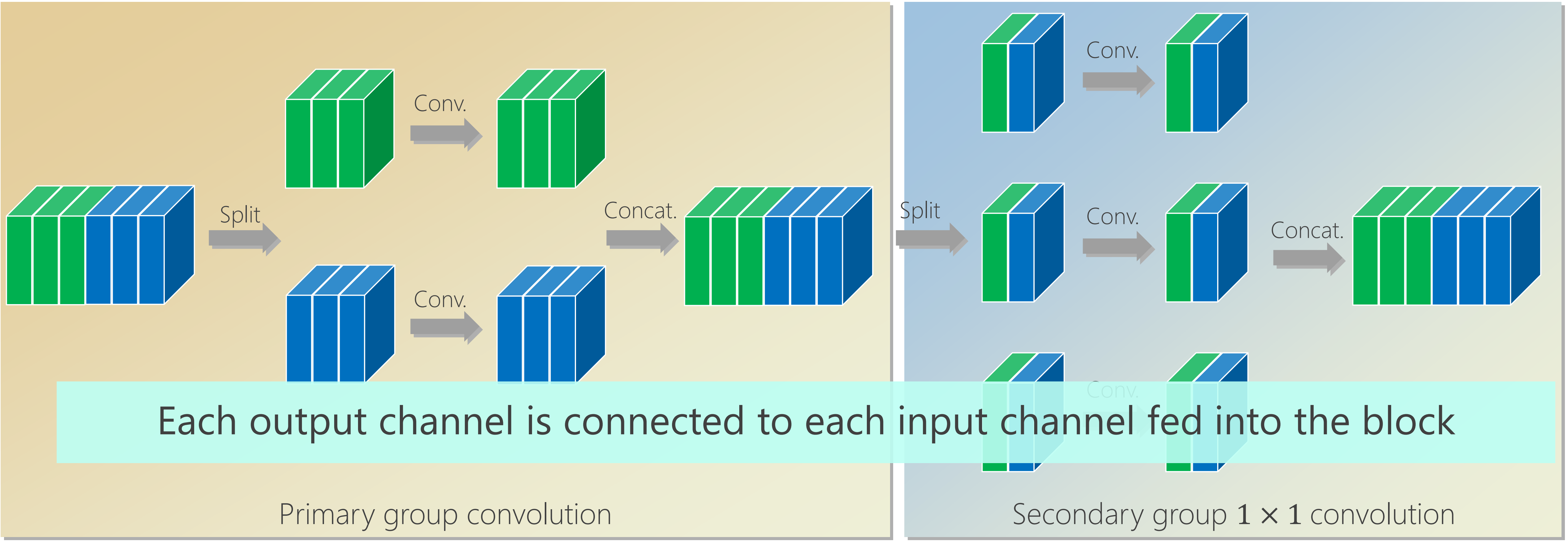
Group convolution



Complexity: $2 \times (3 \times 5 \times 5 \times 3)$

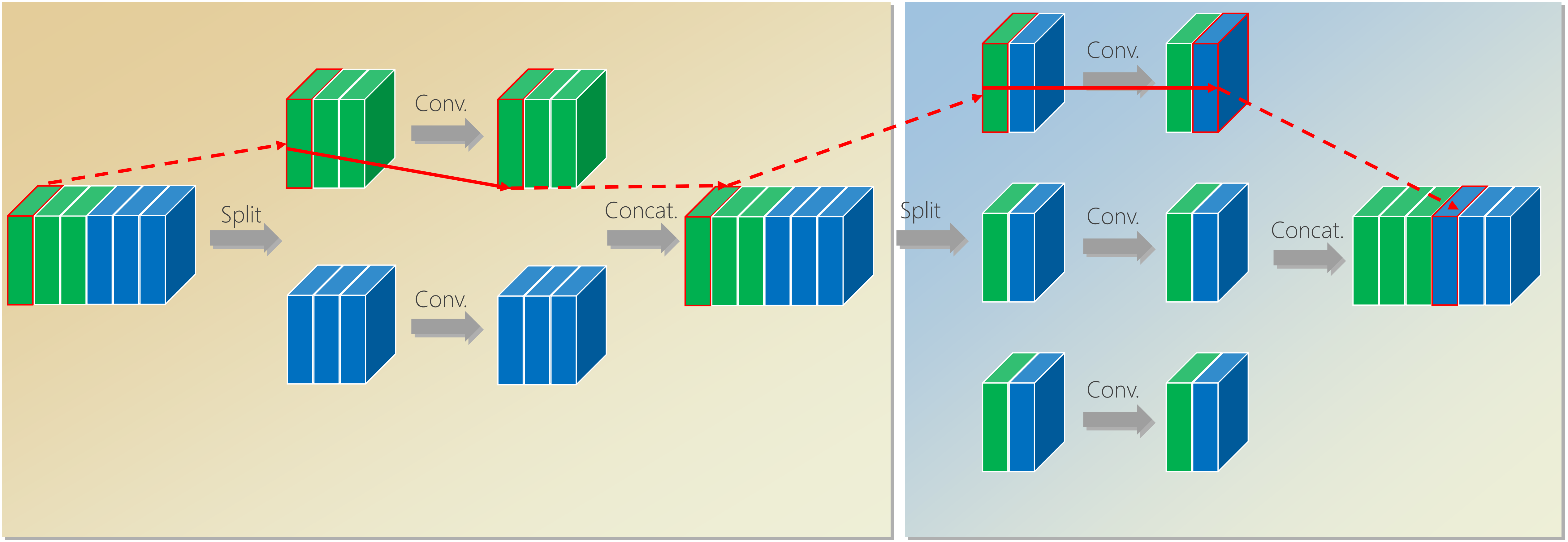
Conduct convolutions *separately* over the partitions
Computation cost is lower than regular convolutions

Interleaved group convolution (IGC)



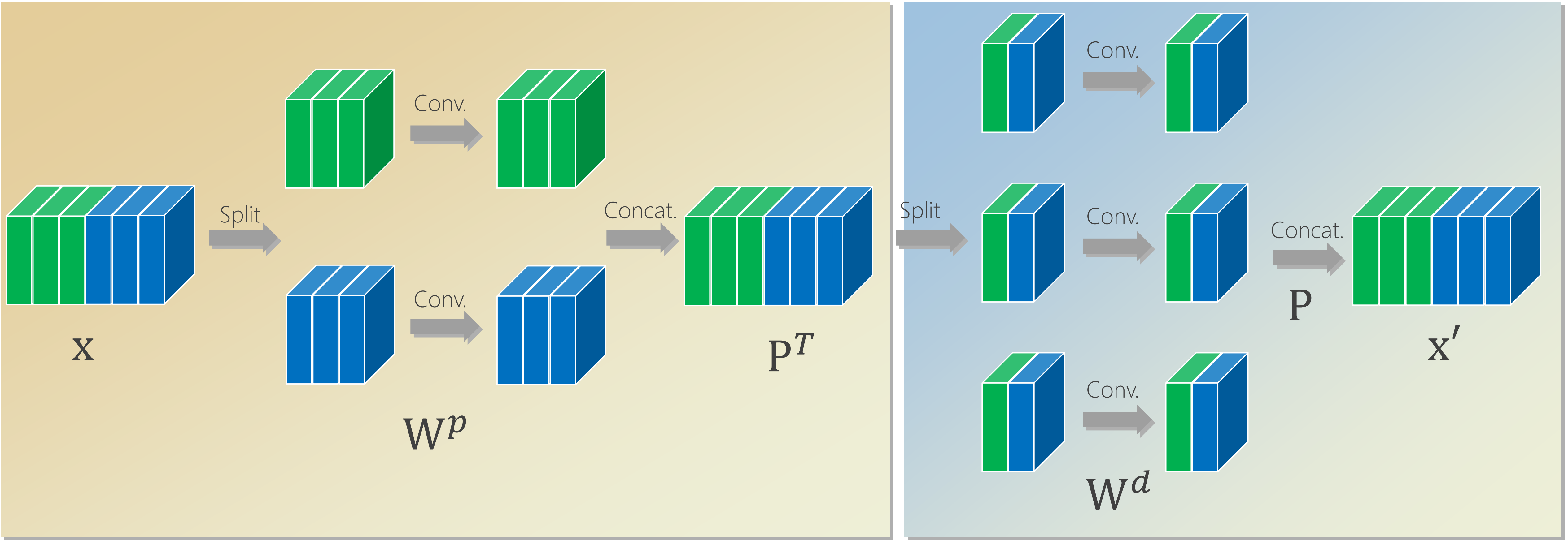
L=2 primary partitions
3 channels in each partition

M=3 secondary partitions
2 channels in each partition



The path connecting each input channel with each output channel

Matrix form



$$\mathrm{x}' = \mathrm{P}\mathrm{W}^d\mathrm{P}^T\mathrm{W}^p\mathrm{x}$$

Criterion: Strict complementary condition

Strict complementary condition:

The channels lying in the *same* branch in one group convolution lie in *different* branches and come from *all* the branches in the other group convolution.

We have: the resulting convolution kernel matrix is *dense*

$$\mathrm{x}' = \boxed{\mathrm{P}\mathrm{W}^{d}\mathrm{P}^{T}\mathrm{W}^{p}}\mathrm{x}$$

Advantages: Wider than regular convolutions

Condition: $$\frac{L}{L-1} < MS$$

#(Primary partitions) #(Secondary partitions) Spatial kernel size

Thus, our IGC is *wider* except $\boldsymbol{L = 1}$ under the same #parameters

Comparison to regular convolutions

CIFAR-10 classification accuracy

depth	RegConv-18	IGC	
20	92.55 ± 0.14	92.84 ± 0.26	
38	91.57 ± 0.09	92.24 ± 0.62	
62	88.60 ± 0.49	90.03 ± 0.85	$+1.43$

Model size: #params ($\times 10^6$)

depth	RegConv-18	IGC
20	0.34	0.15
38	0.71	0.31
62	1.20	0.52

Computation complexity: FLOPS ($\times 10^8$)

depth	RegConv-18	IGC
20	0.51	0.29
38	1.1	0.57
62	1.7	0.95

Comparison to regular convolutions

CIFAR-100 classification accuracy

depth	RegConv-18	IGC	
20	68.71 ± 0.32	70.54 ± 0.26	
38	65.00 ± 0.57	69.56 ± 0.76	
62	58.52 ± 2.31	65.84 ± 0.75	$+7.32$

Model size: #params ($\times 10^6$)

depth	RegConv-18	IGC
20	0.34	0.15
38	0.71	0.31
62	1.20	0.52

Computation complexity: FLOPS ($\times 10^8$)

depth	RegConv-18	IGC
20	0.51	0.29
38	1.1	0.57
62	1.7	0.95

Comparison to ResNets on ImageNet classification

	#params ($\times 10^7$)	**FLOPS** ($\times 10^9$)	**Training error**		**Validation error**	
			Top-1	**Top-5**	**Top-1**	**Top-5**
ResNet (Reg. Conv.)	1.133	2.1	21.43	5.96	30.58	10.77
Our approach	0.861	1.3	13.93	2.75	26.95	8.92
					+3.63	+1.85

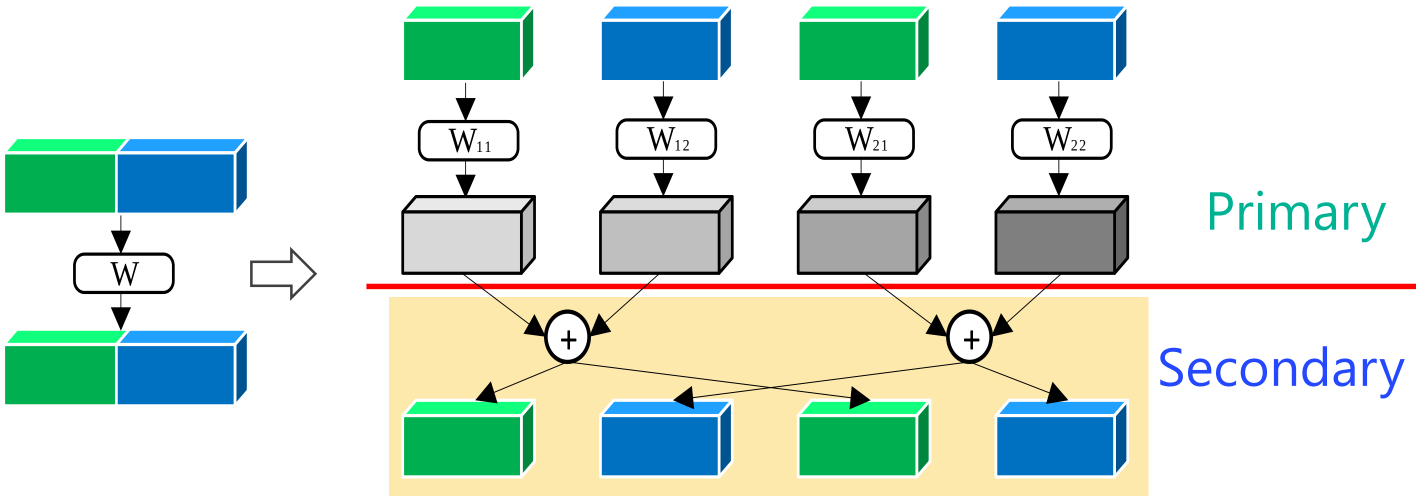
Our approach: replace regular convolutions with our interleaved group convolutions

Regular convolutions are interleaved group convolutions

- Four-branch representation
- Primary group convolution

$$\mathbf{W} = \begin{bmatrix} \mathbf{W}_{11} & \mathbf{W}_{12} \\ \mathbf{W}_{21} & \mathbf{W}_{22} \end{bmatrix}$$

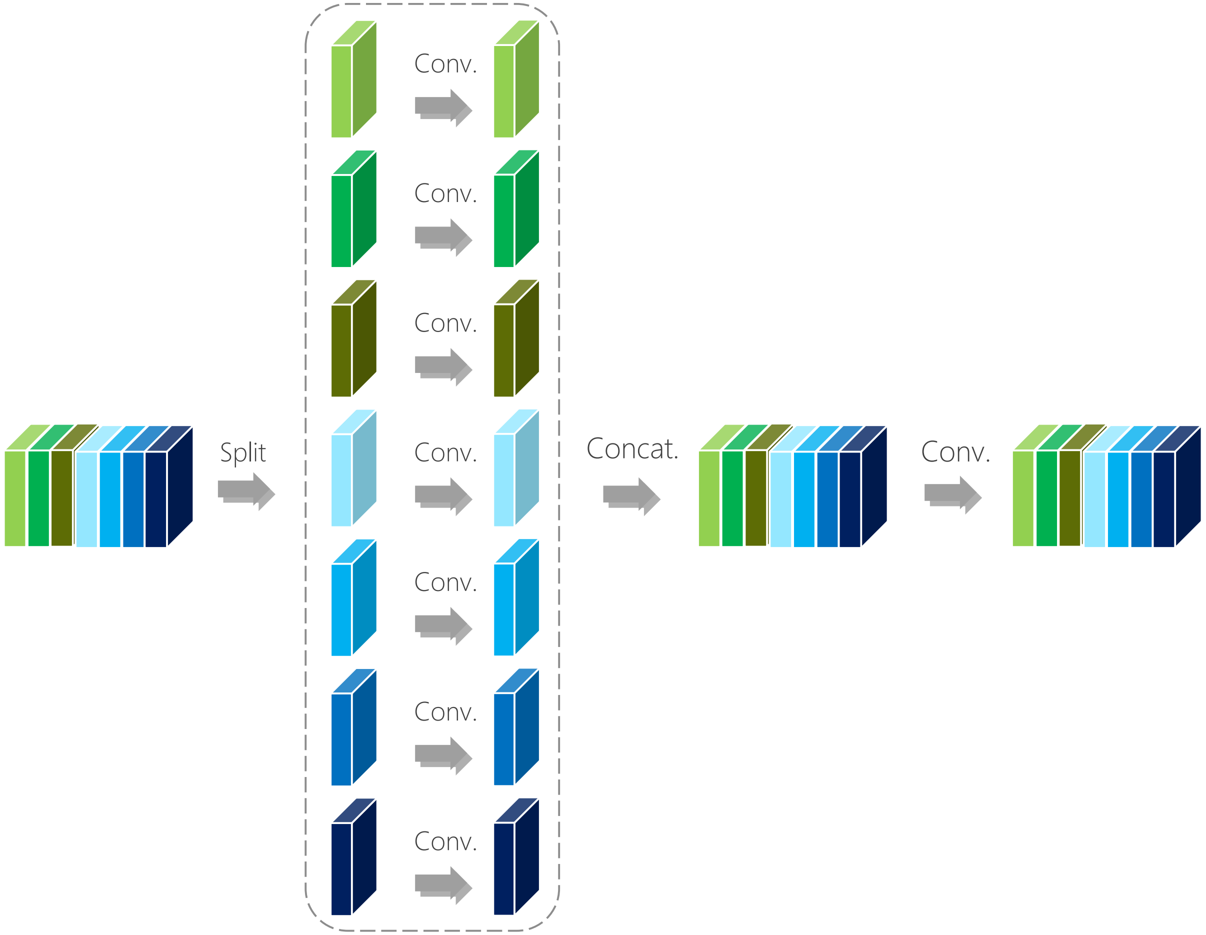
$$\mathbf{W}^p = \mathrm{diag}(\mathbf{W}_{11}, \mathbf{W}_{12}, \mathbf{W}_{21}, \mathbf{W}_{22})$$

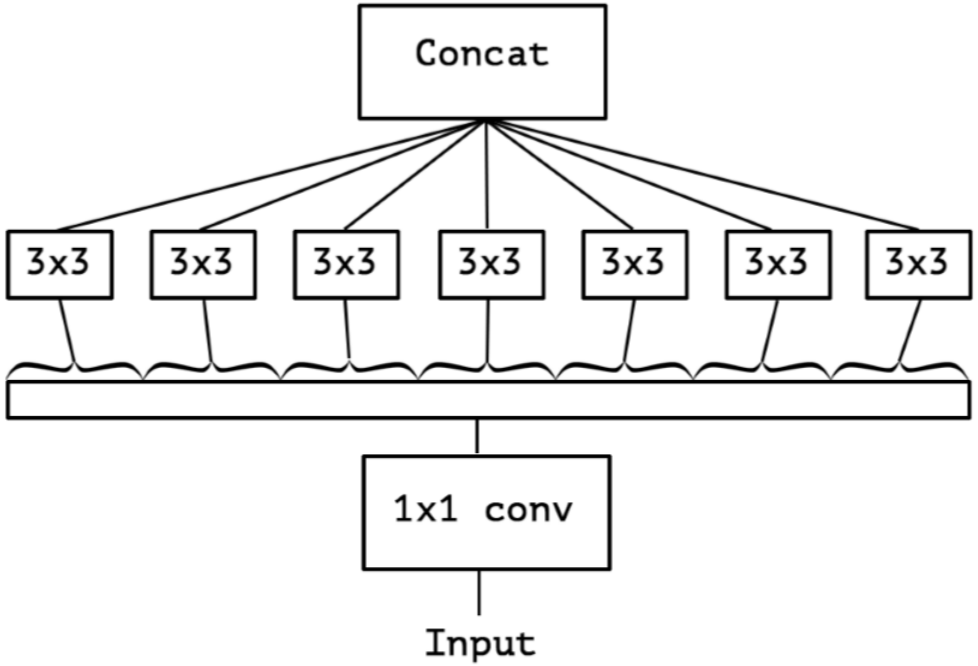


- Secondary group convolution

$$\mathbf{W}^d_{11} = \mathbf{W}^d_{22} = \cdots = \mathbf{W}^d_{MM} = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \end{bmatrix}$$

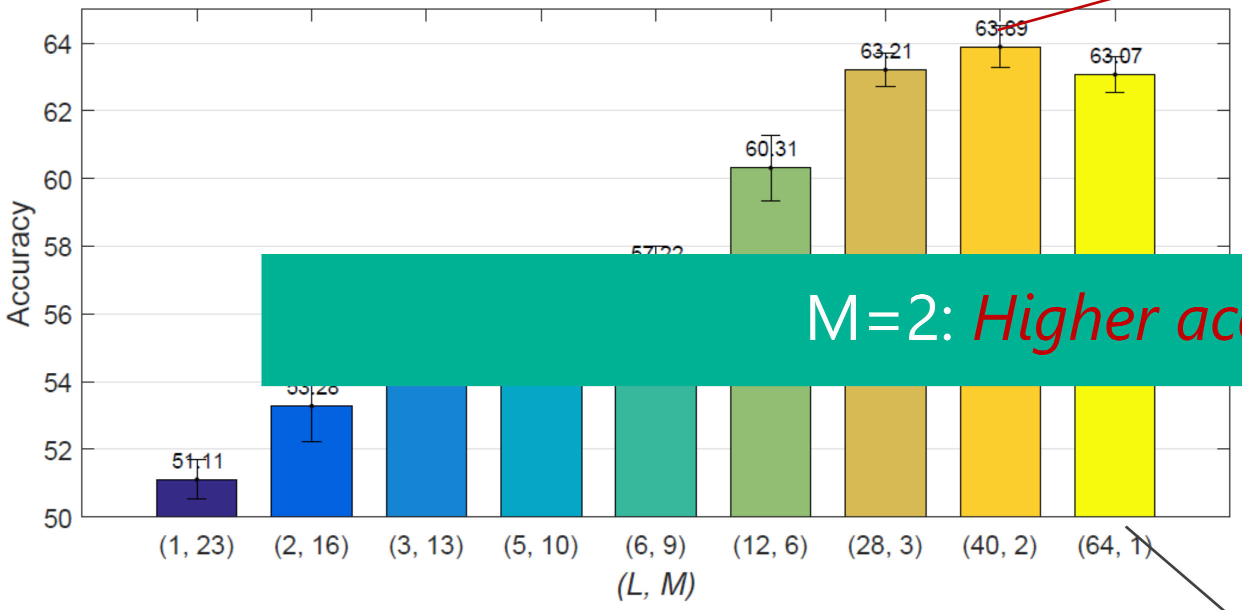
Xception is a special case of IGC





Accuracy under same #params and FLOPS

Best accuracy



M=2: *Higher accur*

width	23	32	39	50	54	72	**84**	80	64

Xception

Comparison to Xception

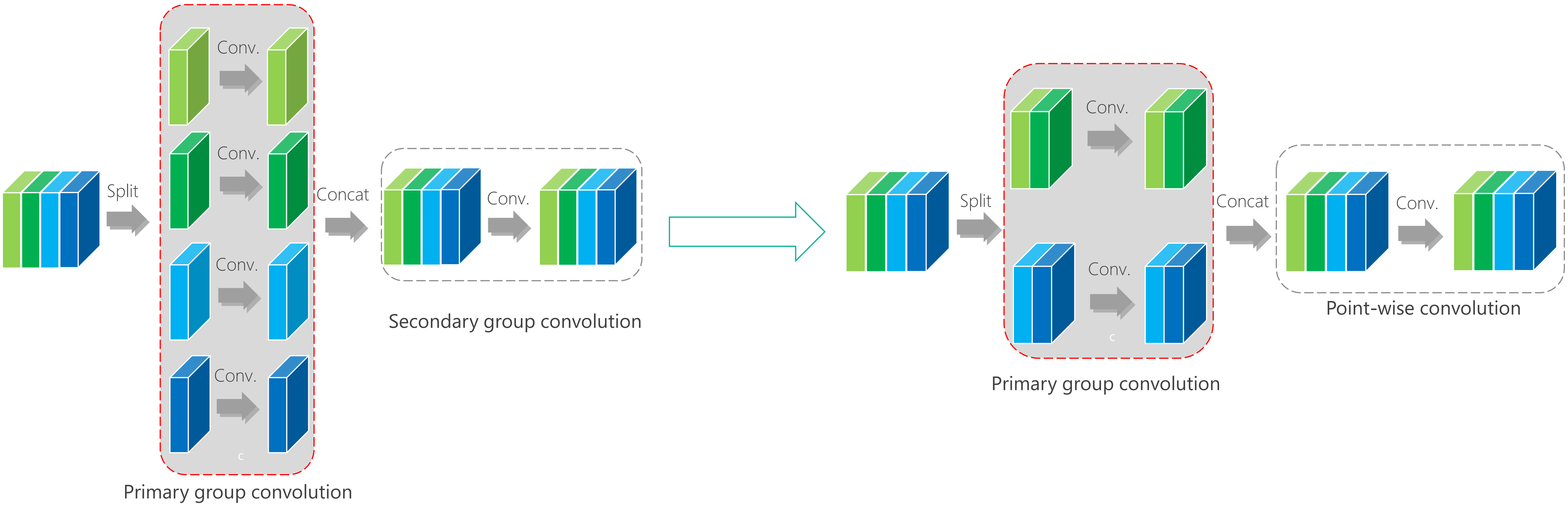
CIFAR-100, Small network

	Xception	Our approach	
testing error	36.93 ± 0.54	36.11 ± 0.62	-0.82
#params	3.62×10^4	3.80×10^4	
FLOPS	3.05×10^7	3.07×10^7	

CIFAR-100, Large network

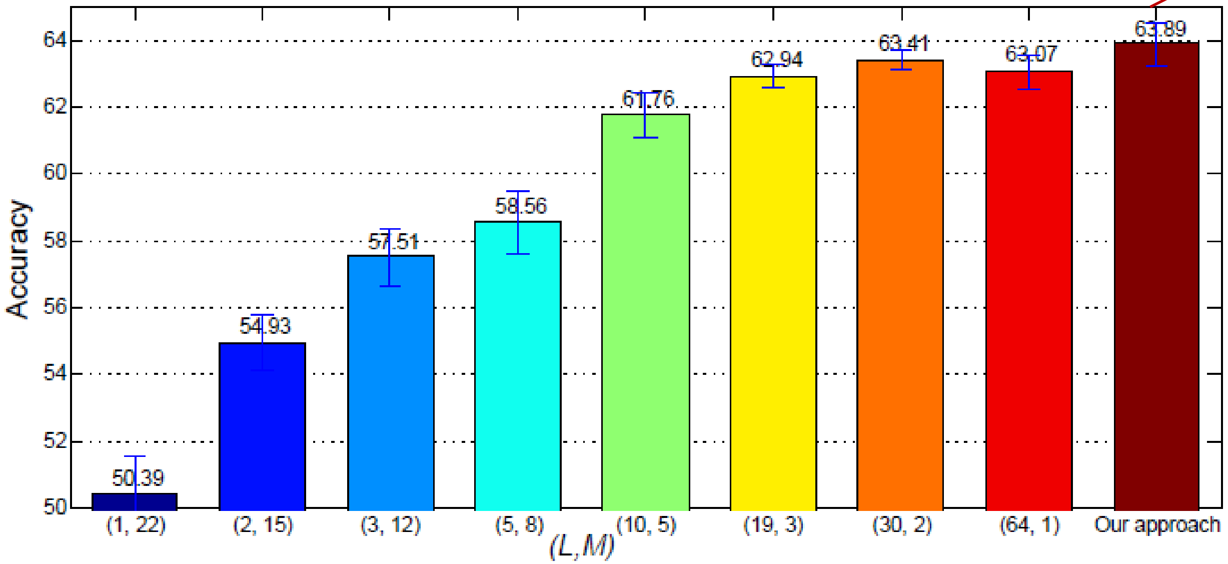
	Xception	Our approach	
testing error	32.87 ± 0.67	31.87 ± 0.58	-1.00
#params	1.21×10^5	1.26×10^5	
FLOPS	1.11×10^8	1.12×10^8	

Deep roots: IGC's variant



Comparison to deep roots

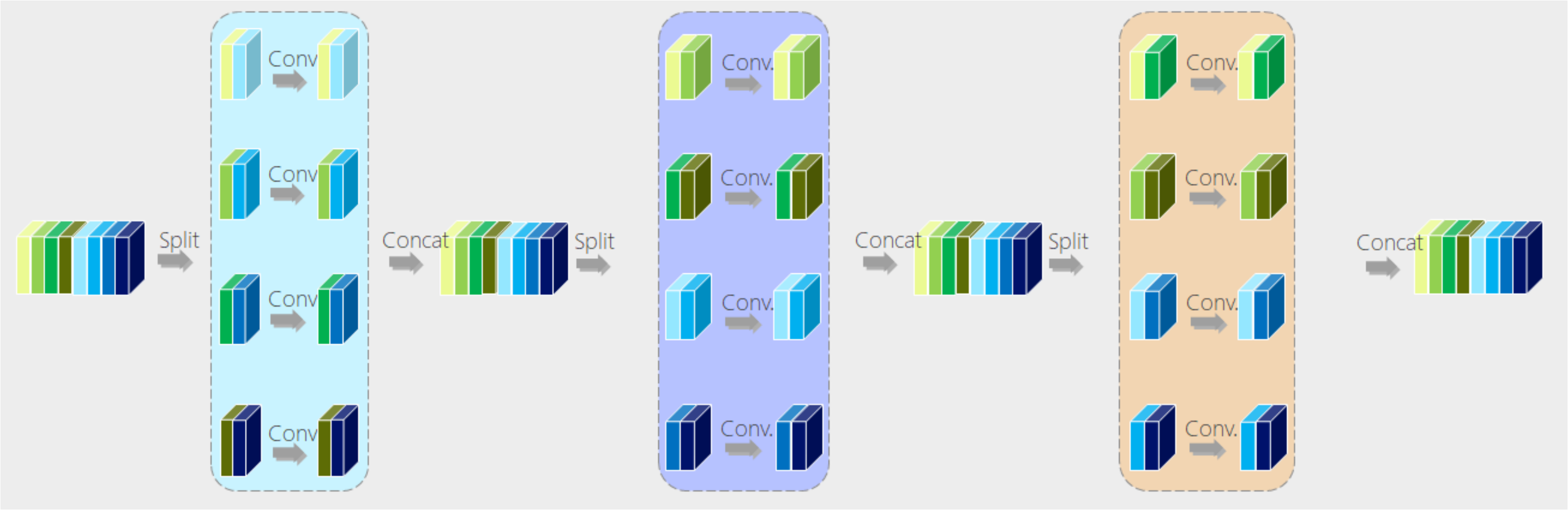
Our approach



Comparison with state-of-the-arts

Method	Depth	#Params.	CIFAR-10	CIFAR-100	SVHN
FractalNet with DO/DP	21	38.6M	5.22	23.30	2.01
	21	38.6M	4.60	23.73	1.87
ResNet	110	1.7M	6.41	27.22	2.01
Multi ResNet	200	10.2M	4.35	20.42	-
Wide ResNet	16	11.0M	4.81	22.07	-
	28	36.5M	4.17	20.50	-
DenseNet	40	1.0M	5.24	24.42	1.79
	100	27.2M	3.74	19.25	1.59
DMRNet	56	1.7M	4.94	24.46	1.66
DMRNet-Wide	32	14.9M	3.94	19.25	1.51
DMRNet-Wide	50	24.8M	3.57	19.00	1.55
IGC-L16M32	20	17.7M	3.37	19.31	1.63
IGC-L450M2	20	19.3M	3.30	19.00	-
IGC-L32M26	20	24.1M	3.31	18.75	1.56

IGCV2: Interleaved Structured Sparse Convolutional Neural Networks



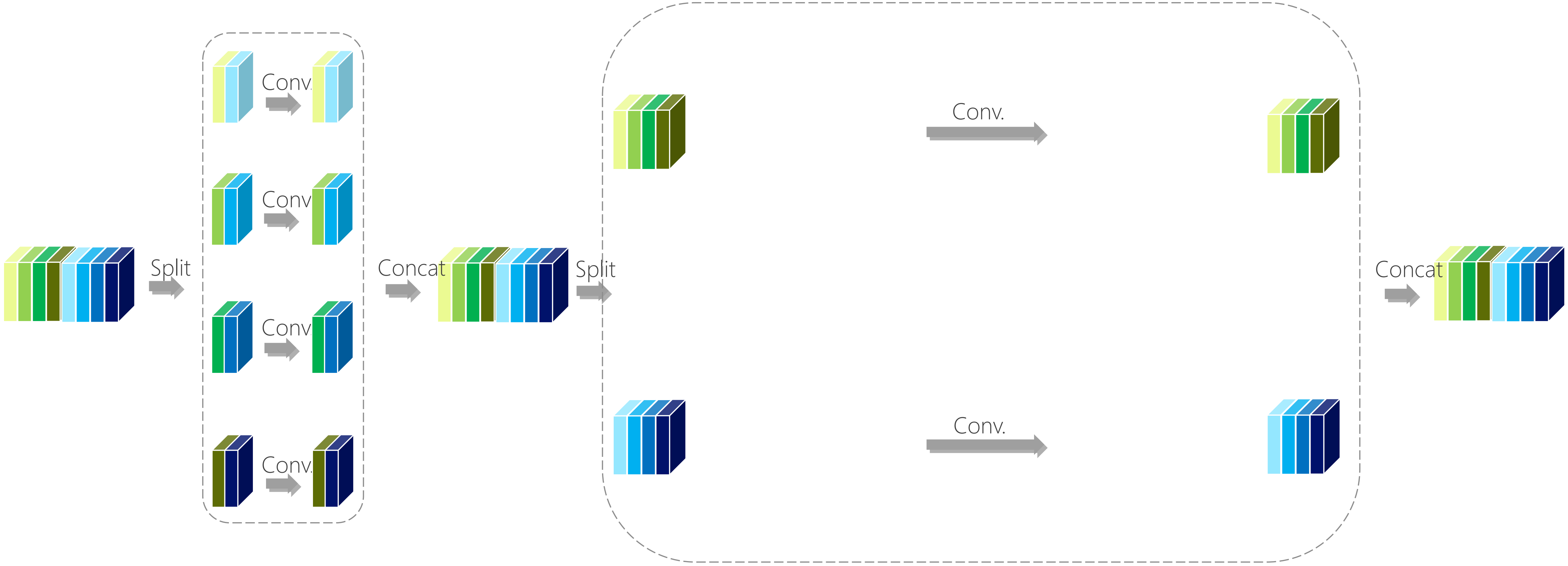
Guotian Xie, Jingdong Wang, Ting Zhang, Jianhuang Lai, Richang Hong, and Guo-JunQi. IGCV2: Interleaved Structured Sparse Convolutional Neural Networks. CVPR 2018.

Interleaved group convolutions for small models

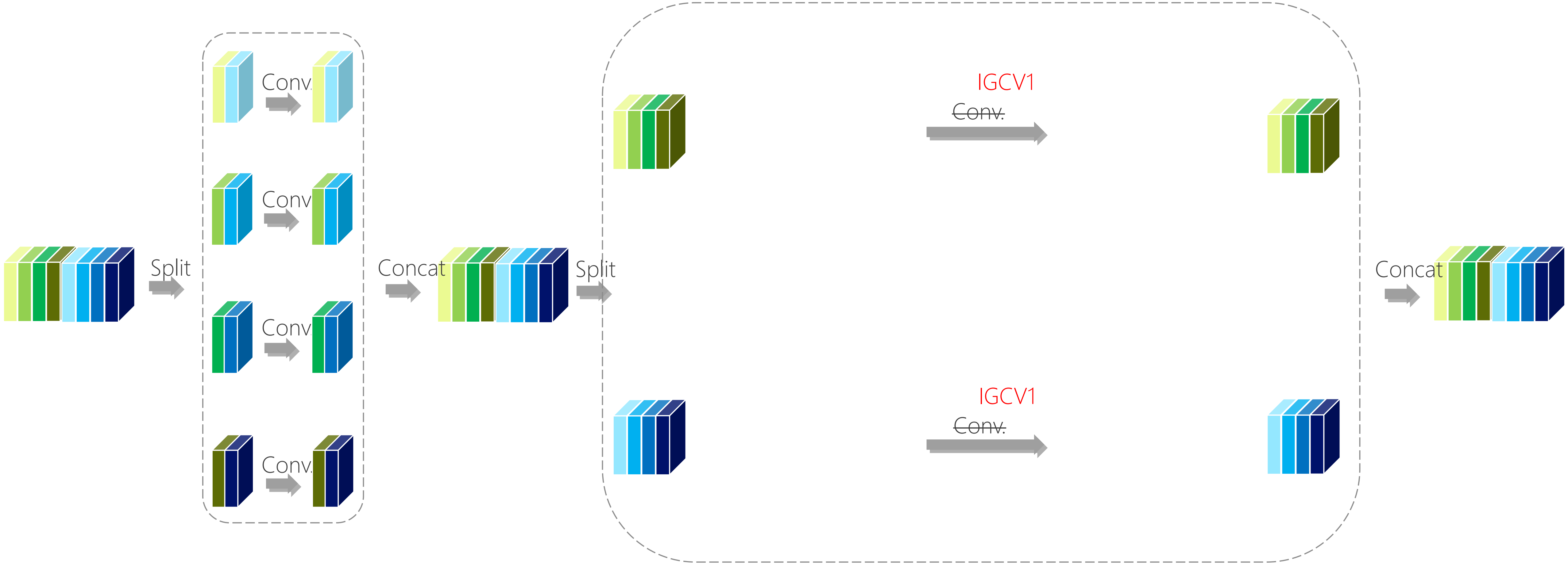
- Structured sparse matrix composition
  - 2 → Multiple structured sparse matrices
  - Benefit: further redundancy reduction (much sparser)

Guotian Xie, Jingdong Wang et. al. IGCV2: Interleaved Structured Sparse Convolutional Neural Networks. CVPR 2018.

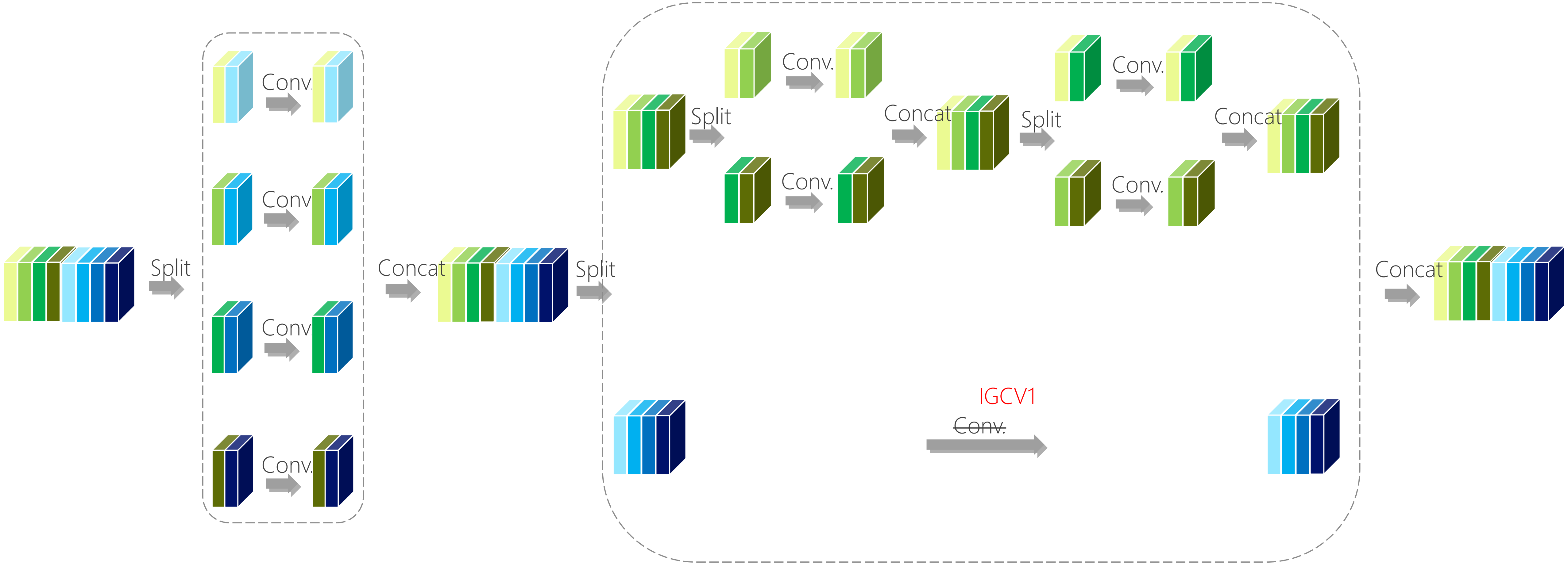
IGCV1



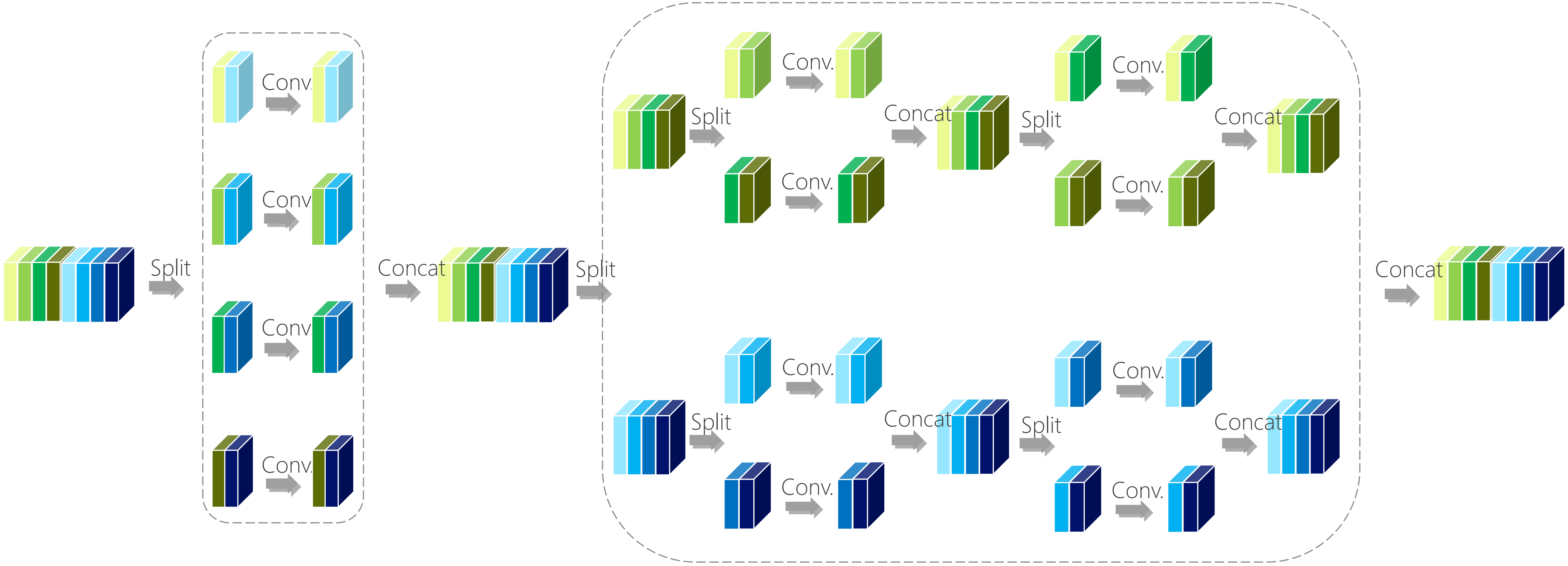
Multiple structured sparse matrix composition



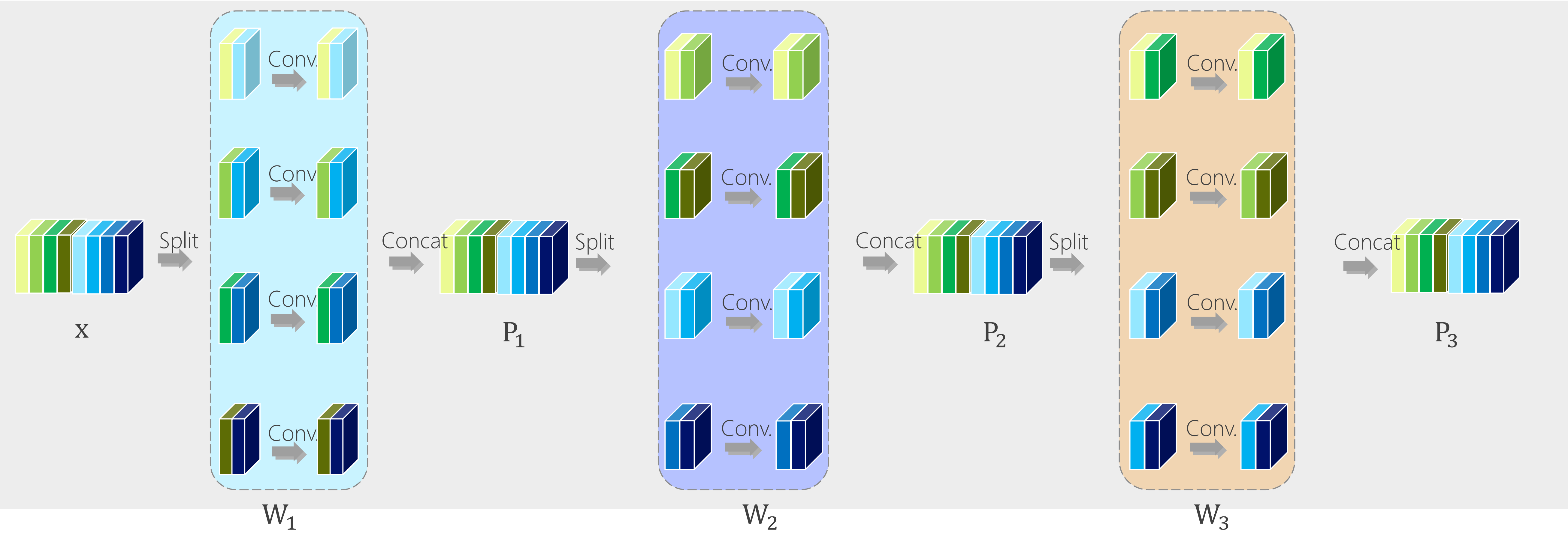
Multiple structured sparse matrix composition



Multiple structured sparse matrix composition



Multiple structured sparse matrix composition



$$x' = \boxed{P_3 W_3 P_2 W_2 P_1 W_1} x$$

Dense due to complementary condition

L (e.g., 3 or >3) group convolutions

Criterion: Strict complementary condition

Divide the *L* group convolutions into two convolutions:

$$\mathbf{x}' = \mathbf{P}_L \boxed{(\mathbf{W}_L \prod_{l=L-1}^{m} \mathbf{P}_l \mathbf{W}_l)} \mathbf{P}_m \boxed{(\mathbf{W}_{m-1} \prod_{l=m-2}^{1} \mathbf{P}_l \mathbf{W}_l)} \mathbf{x}$$

Strict complementary condition:

1. The multiple group convolutions can be merged to two *group* convolutions.
2. The channels lying in the *same* branch in one group convolution lie in *different* branches and come from *all* the branches in the other group convolution.

The resulting convolution kernel matrix is *dense*

How to design L group convolutions

IGC block: 1 channel-wise 3×3 convolution + $(L-1)$ group 1×1 convolutions

Balance condition: To have *minimum total #parameters*, three conditions hold:

1) #parameters for $(L-1)$ group convolutions are the *same*
2) #branches for $(L-1)$ group convolutions are the *same*
3) #channels in each branch are the *same*

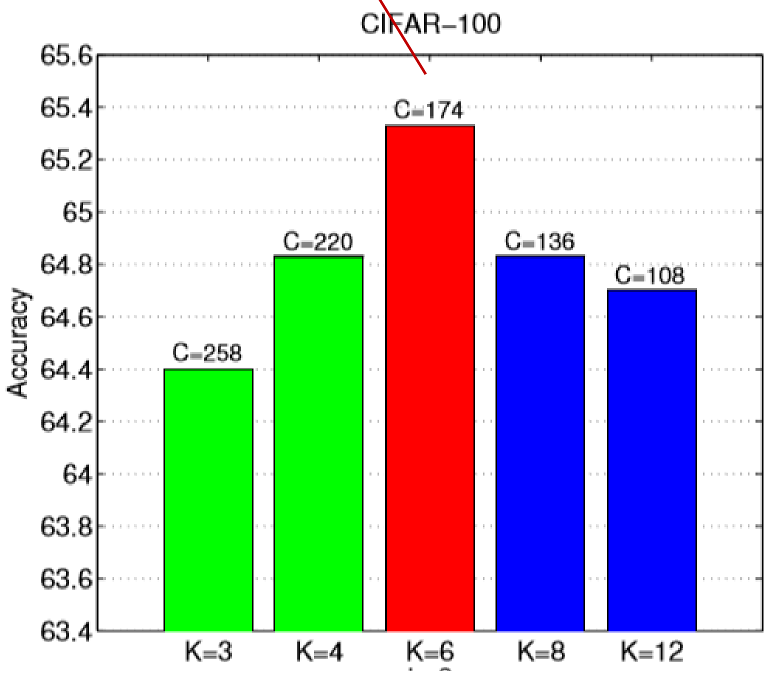
Design: #channels in each branch: $K_2 = \cdots = K_L = C^{\frac{1}{L-1}} = K$

$$\#\text{branches} = \frac{C}{K}$$

width

Empirical justification

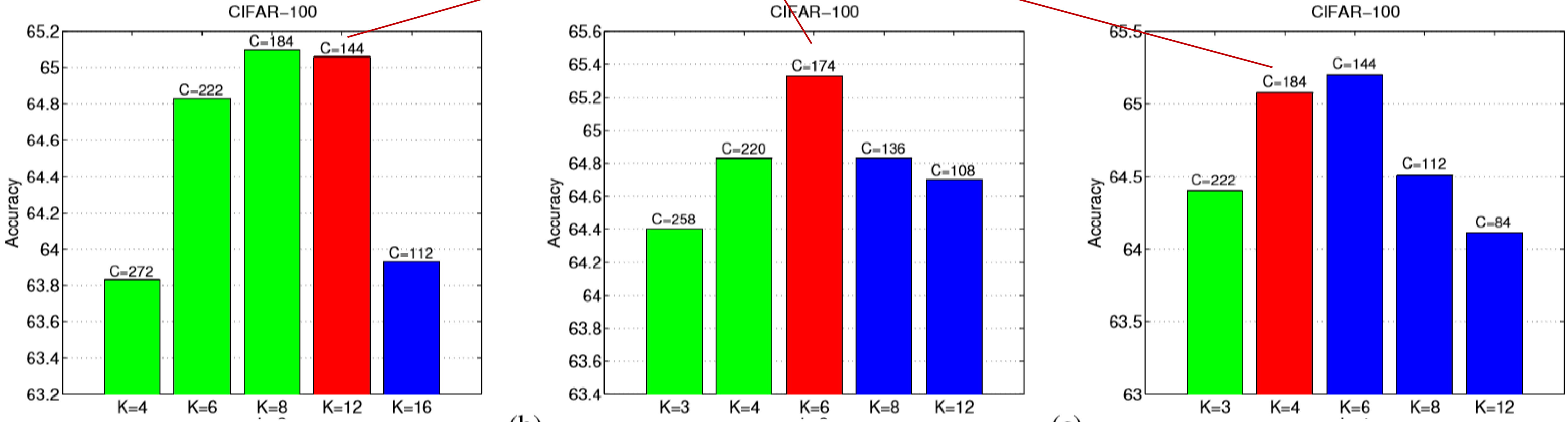
Meet complementary and balance conditions



$L = 4$

Empirical justification

Meet complementary and balance conditions



(a) $L = 3$

(b) $L = 4$

(c) $L = 5$

How many group convolutions?

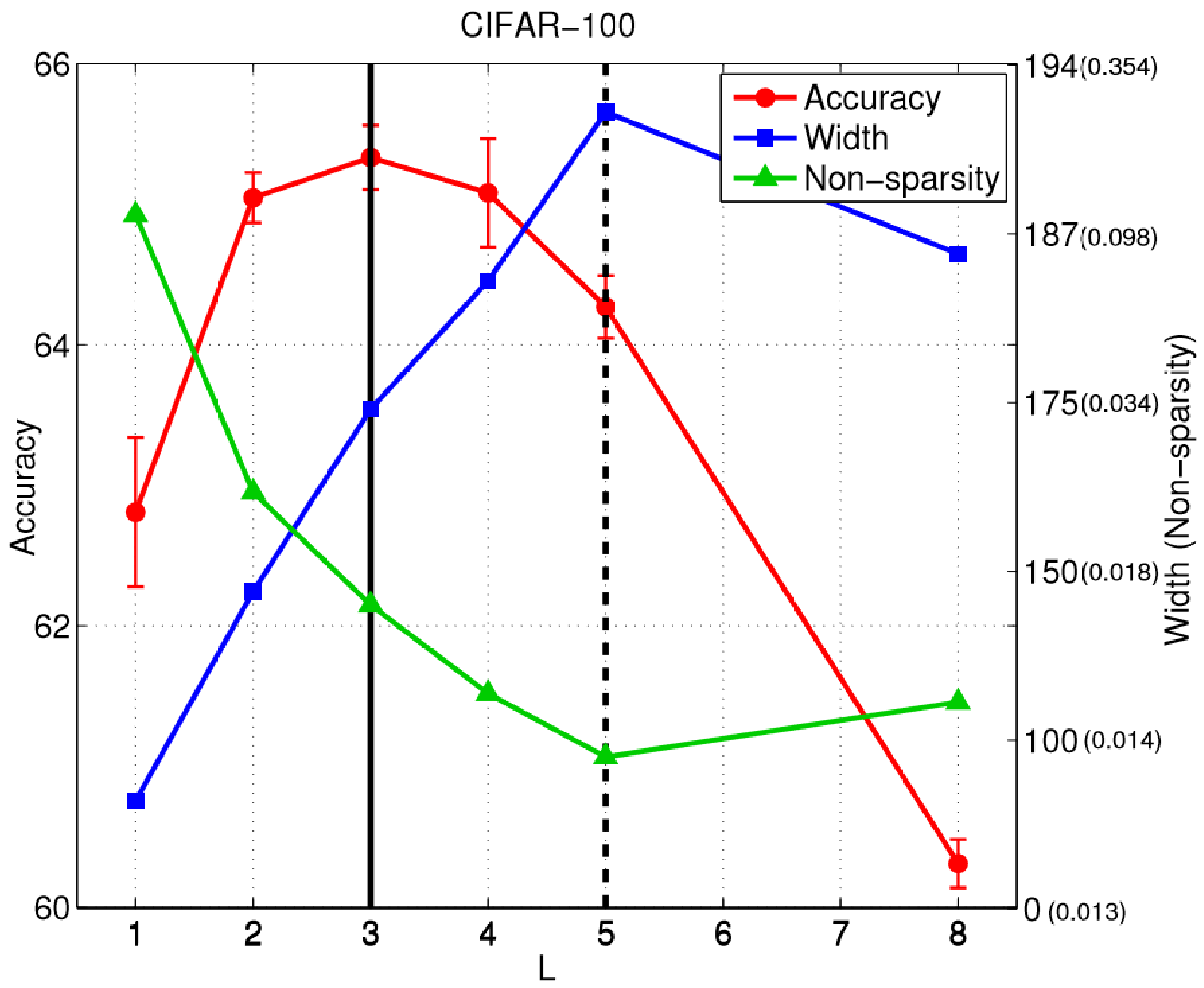
#parameters: When balance condition holds, the amount of parameters for 1 channel-wise 3×3 convolution and $(L-1)$ group 1×1 convolutions:

$$Q(L) = C(L-1)(C)^{\frac{1}{L-1}} + 9C$$

We have $Q(L=3) < Q(L=2) \text{ if } C > 4$ and $L = \log(C) + 1 \text{ so that } Q \text{ is minimum}$

Benefit: Further redundancy reduction by composing more $(L > 2)$ sparse matrices

Empirical analysis of the number of group convolutions L



Comparison to IGCV1

CIFAR-100 classification accuracy

Depth	IGCV1	IGCV2	
8	66.52 ± 0.12	67.65 ± 0.29	
20	70.87 ± 0.39	72.63 ± 0.07	
26	71.82 ± 0.40	73.49 ± 0.34	$+1.67$

Model size: #params ($\times 10^6$)

Depth	IGCV1	IGCV2
8	0.046	0.046
20	0.151	0.144
26	0.203	0.193

Computation complexity: FLOPS ($\times 10^7$)

Depth	IGCV1	IGCV2
8	1.00	1.18
20	2.89	3.20
26	3.83	4.21

Comparison to IGCV1

Tiny ImageNet classification accuracy

Depth	IGCV1	IGCV2	
8	48.57 ± 0.53	51.49 ± 0.33	
20	54.83 ± 0.27	56.40 ± 0.17	
26	56.64 ± 0.15	57.12 ± 0.09	$+0.48$

Model size: #params ($\times 10^6$)

Depth	IGCV1	IGCV2
8	0.046	0.046
20	0.151	0.144
26	0.203	0.193

Computation complexity: FLOPS ($\times 10^7$)

Depth	IGCV1	IGCV2
8	1.00	1.18
20	2.89	3.20
26	3.83	4.21

Comparison with small models

Method	Depth	#Params.	CIFAR-10	CIFAR-100	Tiny ImageNet
FractalNet	21	38.6M	5.22	23.30	-
ResNet	110	1.7M	5.52	28.02	46.5
DFN-MR1	56	1.7M	4.94	24.46	-
RiR	18	10.3M	5.01	22.90	-
ResNet34	34	21.4M	-	-	46.9
ResNet18-2x	18	25.7M	-	-	44.6
WRN-32-4	32	7.4M	5.43	23.55	39.63
WRN-40-4	40	8.9M	4.53	21.18	-
DenseNet (k=12)	40	1.0M	-	-	39.09
DenseNet (k=12)	40	1.0M	5.24	24.42	-
DenseNet-BC (k=12)	100	0.8M	4.51	22.27	-
IGC-V2*-C416	20	0.65M	5.49	22.95	38.81

Interleaved group convolutions for small models

- Structured sparse matrix composition
  - 2 → Multiple structured sparse matrices
  - Benefit: further redundancy reduction (much sparser)

- Complementary condition
  - Strict → loose

Guotian Xie, Jingdong Wang et. al. IGCV2: Interleaved Structured Sparse Convolutional Neural Networks. CVPR 2018.

Design criterion: Strict complementary condition

Goal: There is one and only one path between each pair of input and output channels such that each output channel gets information from each input channel

Design criterion: ~~Strict~~ complementary condition
Loose

are more paths

Goal: There ~~is one and only one path~~ between each pair of input and output channels such that each output channel gets ^ information from each input channel
rich

Design criterion: Loose complementary condition

Loose

~~Strict~~ complementary condition:

1. The multiple group convolutions can be merged to two *group* convolutions.
2. The ~~channels~~ lying in the *same* branch in one group convolution lie in *different* branches and come from *all* the branches in the other group convolution.

super-channels

A super-channel is composed of 2 or more channels

Interleaved group convolutions for small models

- Structured sparse matrix composition
  - 2 → Multiple structured sparse matrices
  - Benefit: further redundancy reduction (much sparser)

IGCV2

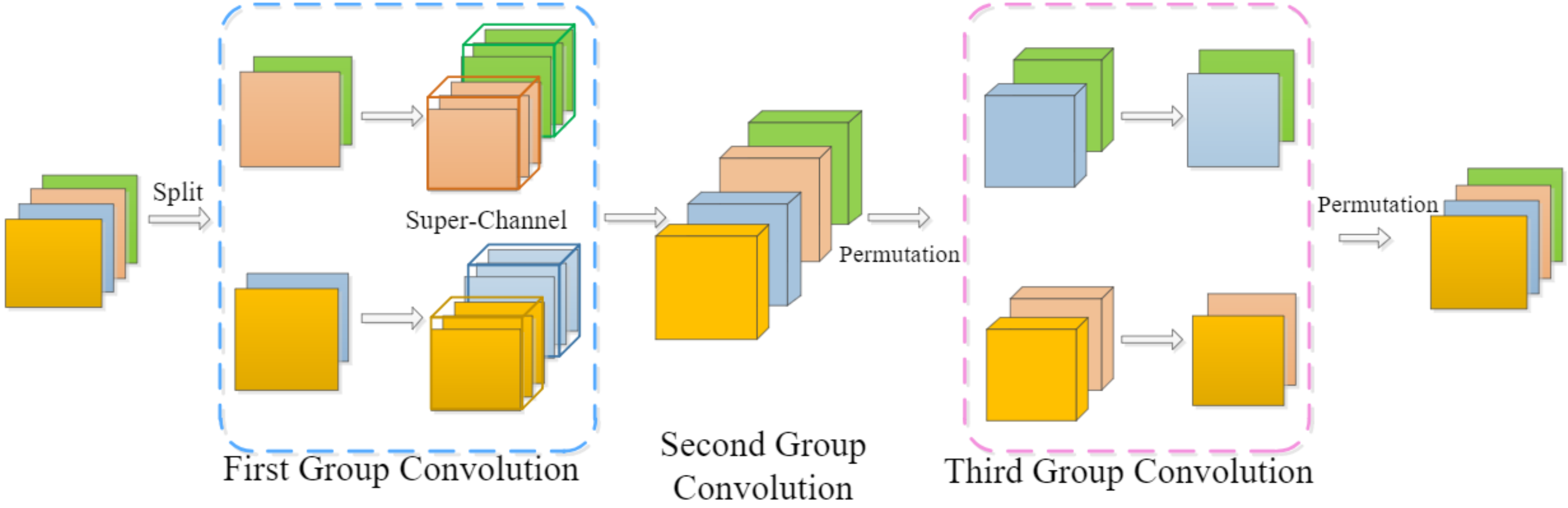
- Complementary condition
  - Strict → loose

Guotian Xie, Jingdong Wang et. al. IGCV2: Interleaved Structured Sparse Convolutional Neural Networks. CVPR 2018.

Comparison to MobileNetV1 on ImageNet

Model	#Params. (M)	FLOPS (M)	Accuracy (%)
MobileNetV1-1.0	4.2	569	70.6
IGCV2-1.0	4.1	564	**70.7**
MobileNetV1-0.5	1.3	149	63.7
IGCV2-0.5	1.3	156	**65.5**
MobileNetV1-0.25	0.5	41	50.6
IGCV2-0.25	0.5	46	**54.9**

IGCV3: Interleaved Low-Rank Group Convolutions



Ke Sun, Mingjie Li, Dong Liu, and Jingdong Wang. IGCV3: Interleaved Low-Rank Group Convolutions for Efficient Deep Neural Networks. BMVC 2018.

Interleaved group convolutions for small models

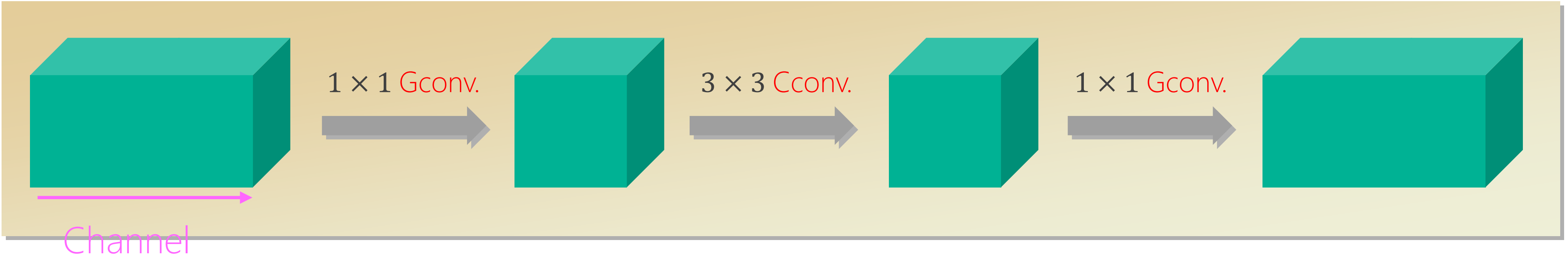
- Structured sparse matrix composition
  - 2 → Multiple structured sparse matrices
  - Benefit: further redundancy reduction (much sparser)
- Complementary condition
  - Strict → loose
- Low-rank structured sparse matrix composition
  - Structured sparse + low-rank

Ke Sun, Mingjie Li, Dong Liu, and Jingdong Wang. IGCV3: Interleaved Low-Rank Group Convolutions for Efficient Deep Neural Networks. BMVC 2018.

Structured sparse + low-rank

Structured sparse + low-rank:

1. Group 1×1 convolution (Gconv.): dimension reduction
2. Channel-wise 3×3 convolution (Cconv.)
3. Group 1×1 convolution (Gcon.): dimension increasing



Interleaved group convolutions for small models

- Structured sparse matrix composition
  - 2 → Multiple structured sparse matrices
  - Benefit: further redundancy reduction (much sparser)
- Complementary condition
  - Strict → loose

IGCV3

- Low-rank structured sparse matrix composition
  - Structured sparse + low-rank

Ke Sun, Mingjie Li, Dong Liu, and Jingdong Wang. IGCV3: Interleaved Low-Rank Group Convolutions for Efficient Deep Neural Networks. BMVC 2018.

Comparison to MobileNetV2 on ImageNet

Network	#Params. (M)	FLOPS (M)	Accuracy (%)
MobileNetV1-1.0	4.2	569	70.6
IGCV2-1.0	4.1	564	70.7
MobileNetV2-1.0 (paper)	3.4	300	72.0
MobileNetV2-1.0 (our impl.)	3.4	300	71.01
IGCV3-1.0	3.5	320	72.2

Comparison to MobileNetV2 on ImageNet

Network	#Params. (M)	FLOPS (M)	Accuracy (%)
MobileNetV1-1.0	4.2	569	70.6
IGCV2-1.0	4.1	564	70.7
MobileNetV2-1.0 (paper)	3.4	300	72.0
MobileNetV2-1.0 (our impl.)	3.4	300	71.01
IGCV3-1.0	3.5	320	72.2
MobileNetV2-0.7 (our impl.)	1.9	160	66.57
IGCV3-0.7	2.0	170	68.46

COCO object detection

Network	#Params. (M)	mAP
SSD	36.1M	23.2
YOLOV2	50.7M	21.6
MobileNetV1 SSDLite	5.1M	22.2
MobileNetV2 SSDLite	4.3M	22.1
IGCV3+ SSDLite	4.0M	22.2

Summary

- Advantages
  - Small model
  - Fast computation
  - High accuracy
- Drop-in replacement of regular convolutions
  - Interleaving
  - Group 1×1 convolutions
  - Low rank
- Superior to Google's MobileNets

Publications

[1] Guotian Xie, Jingdong Wang et. al. IGCV2: Interleaved Structured Sparse Convolutional Neural Networks. CVPR 2018.

[2] Ting Zhang, Guo-Jun Qi, Bin Xiao, and Jingdong Wang: Interleaved Group Convolutions for Deep Neural Networks. ICCV (2017)

[3] Liming Zhao, Mingjie Li, Depu Meng, Xi Li, Zhuowen Tu, and Jingdong Wang: Deep Convolutional Neural Networks with Merge-and-Run Mappings. IJCAI 2018.

[4] Guotian Xie, Ting Zhang, Kuiyuan Yang, Jianhuang Lai, and Jingdong Wang: Decoupled Convolutions for CNNs. AAAI 2018

[5] Ke Sun, Mingjie Li, Dong Liu, and Jingdong Wang: IGCV3: Interleaved Low-Rank Group Convolutions for Efficient Deep Neural Networks. BMVC 2018.

[6] Jingdong Wang, Zhen Wei, and Ting Zhang: Deeply-fused nets. 2016.

Collaborators

- Liming Zhao
- Ting Zhang
- Zhen Wei
- Guotian Xie
- Ke Sun
- Depu Meng
- Mingjie Li

- Bin Xiao
- Guojun Qi

Thanks!
Q&A

Deep fusion code:
https://github.com/zlmzju/fusenet

IGC Code:
https://github.com/homles11/IGCV3

Homepage:
https://jingdongwang2017.github.io/