Modern Convolution Network Design

杨健程 YANG Jiancheng

Oct 25, 2017

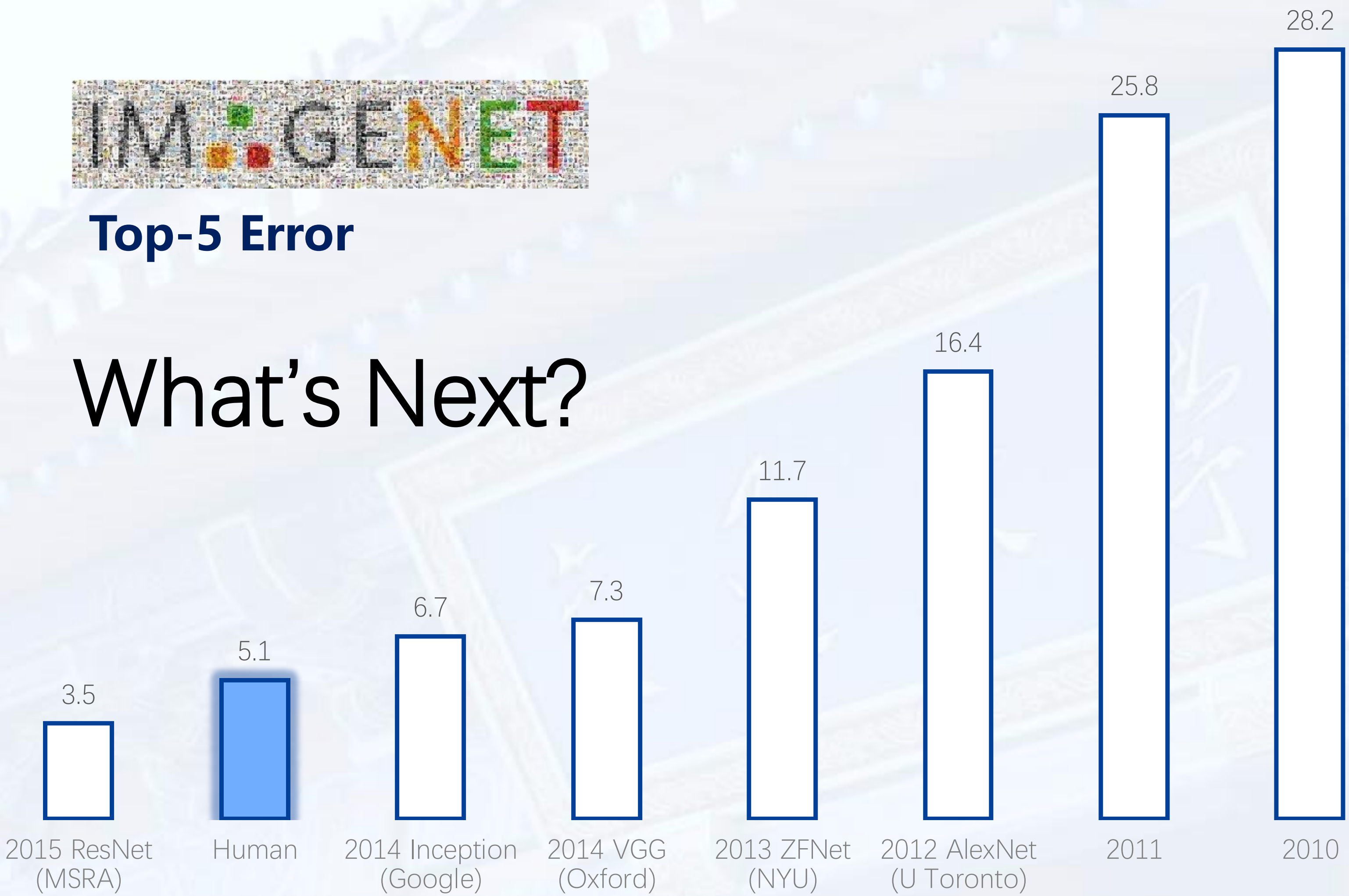
上海交通大學
SHANGHAI JIAO TONG UNIVERSITY

Milestone

IMAGENET

Top-5 Error

What's Next?



Agenda

*Times are last modified version on arXiv

- Deeper and easy to train (Res / Dense)
  - Pre-Activation (2016.07)
  - DenseNet (2016.12)
  - Dual Path Networks (2017.07)
- Wider and light-weight (Group Conv)
  - ResNeXt (2017.04)
  - Xception (2017.04)
  - ShuffleNet (2017.07)
  - Merge-and-Run (2017.07)
  - Interleaved Group Conv (2017.07)
- Global Context
  - Dilated Conv: Dilated-8 (2016.04) and Dilated Residual Network (2017.05)
  - Squeeze-and-Excitation (2017.09)

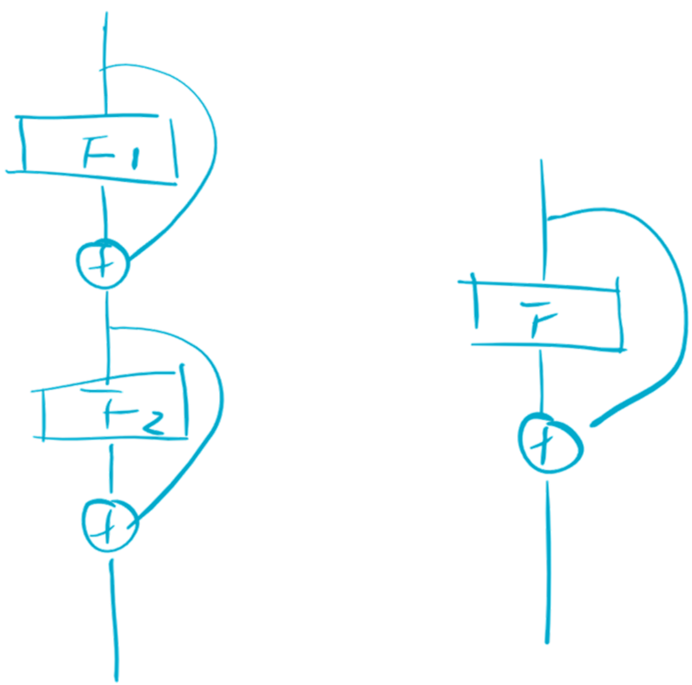
Agenda

*Times are last modified version on arXiv

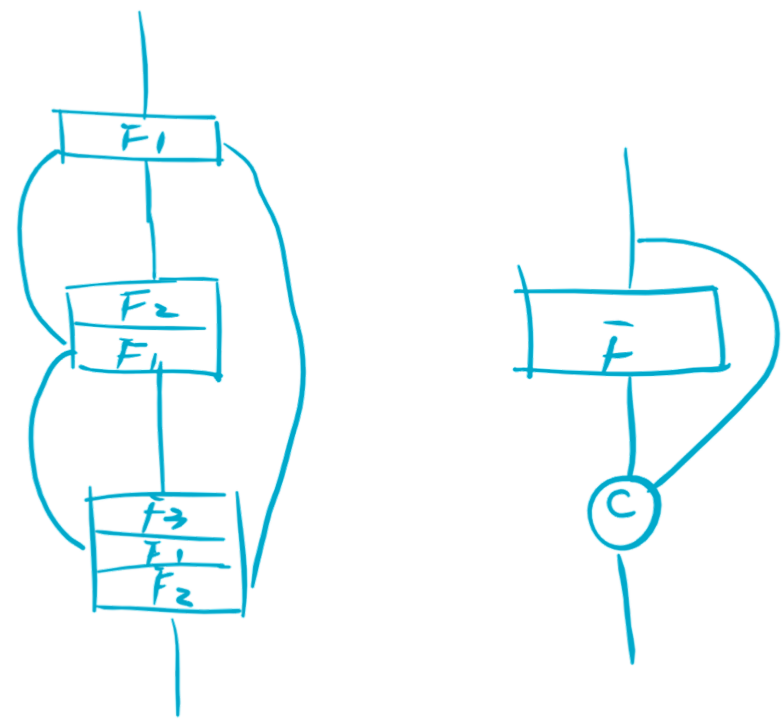
- **Deeper** and easy to train (Res / Dense)
  - Pre-Activation (2016.07)
  - DenseNet (2016.12)
  - Dual Path Networks (2017.07)
- **Wider** and light-weight (Group Conv)
  - ResNeXt (2017.04)
  - Xception (2017.04)
  - ShuffleNet (2017.07)
  - Merge-and-Run (2017.07)
  - Interleaved Group Conv (2017.07)
- Global Context
  - Dilated Conv: Dilated-8 (2016.04) and Dilated Residual Network (2017.05)
  - Squeeze-and-Excitation (2017.09)

Deeper and easy to train

Res / Dense



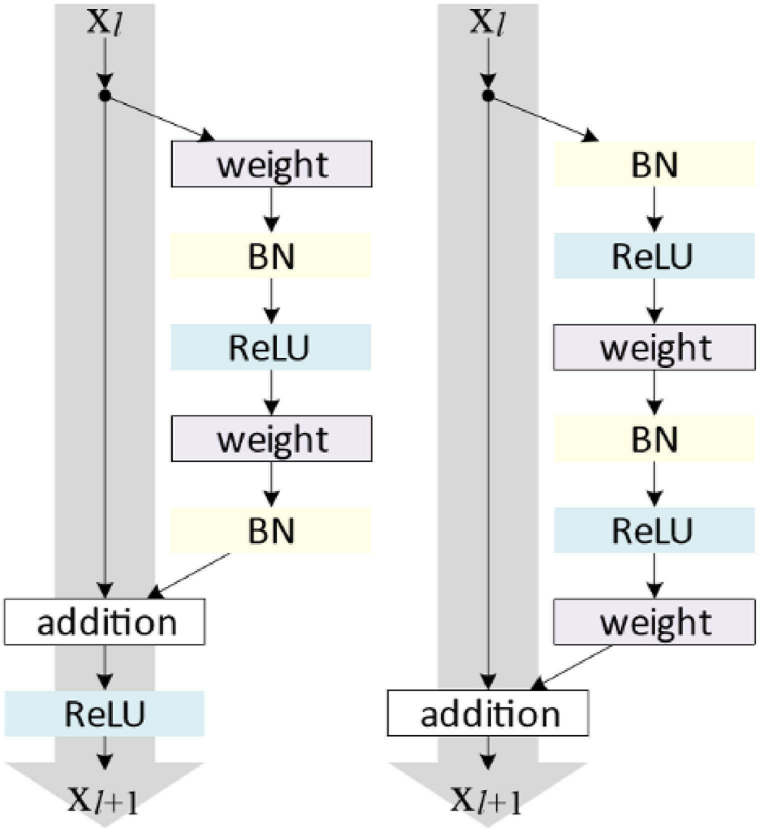
ResNet and Residual Block



DenseNet and Dense Block

Deeper and easy to train

Pre-Activation



(a) original (b) proposed

- 1. Ease of Optimization

$$\mathbf{x}_L = \mathbf{x}_l + \sum_{i=l}^{L-1} \mathcal{F}(\mathbf{x}_i, \mathcal{W}_i), \tag{4}$$

$$\frac{\partial \mathcal{E}}{\partial \mathbf{x}_l} = \frac{\partial \mathcal{E}}{\partial \mathbf{x}_L}\frac{\partial \mathbf{x}_L}{\partial \mathbf{x}_l} = \frac{\partial \mathcal{E}}{\partial \mathbf{x}_L}\left(1 + \frac{\partial}{\partial \mathbf{x}_l}\sum_{i=l}^{L-1} \mathcal{F}(\mathbf{x}_i, \mathcal{W}_i)\right). \tag{5}$$

We also find that the impact of f = ReLU is not severe when the ResNet has fewer layers (*e.g.*, 164 in Fig. 6(right)). The training curve seems to suffer

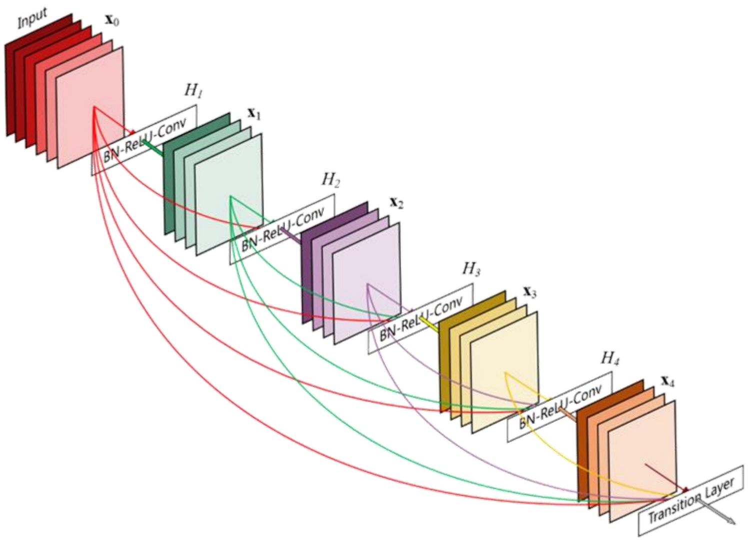
- 2. Reducing Overfitting

the next weight layer. On the contrary, in our pre-activation version, the inputs to all weight layers have been normalized.

Deeper and easy to train

DenseNet

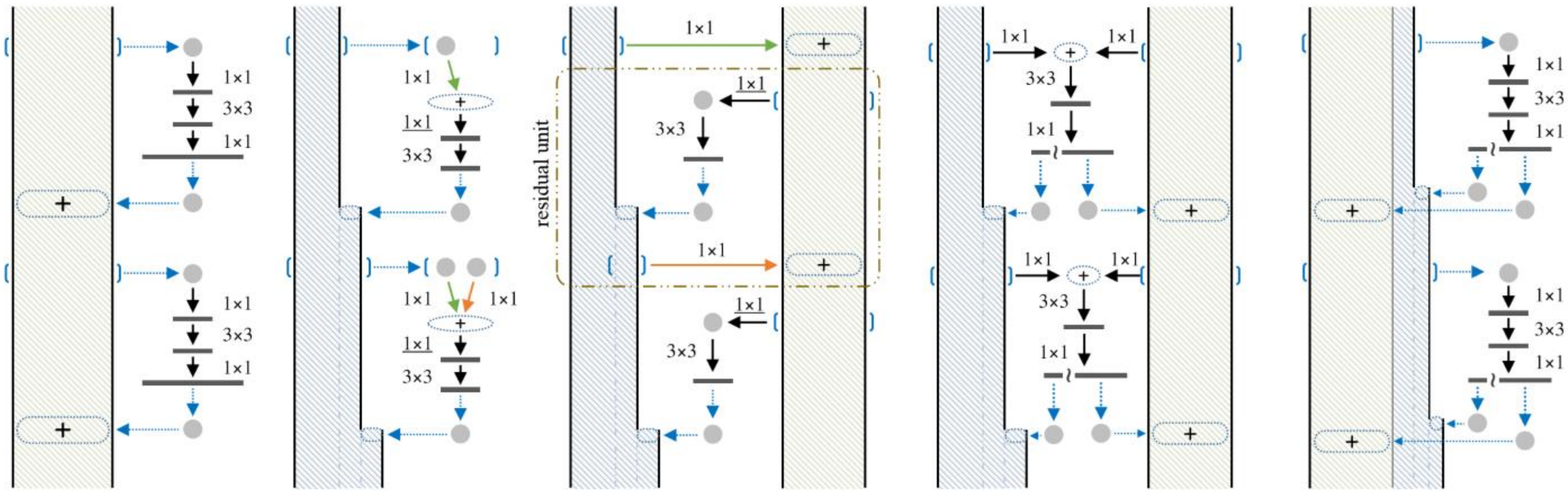
- DenseNet-**BC**
  - Bottleneck
  - Compression
- 1 / 3 parameters of ResNet counterpart
- Less prone to overfitting
- Able to train from scratch
- Large memory consuming



Deeper and easy to train

Dual Path Networks

- Res + Dense



(a) Residual Network (b) Densely Connected Network (c) Densely Connected Network (with shared connections) (d) Dual Path Architecture (e) DPN

- split => one Res path + one Dense path => merge

Agenda

*Times are last modified version on arXiv

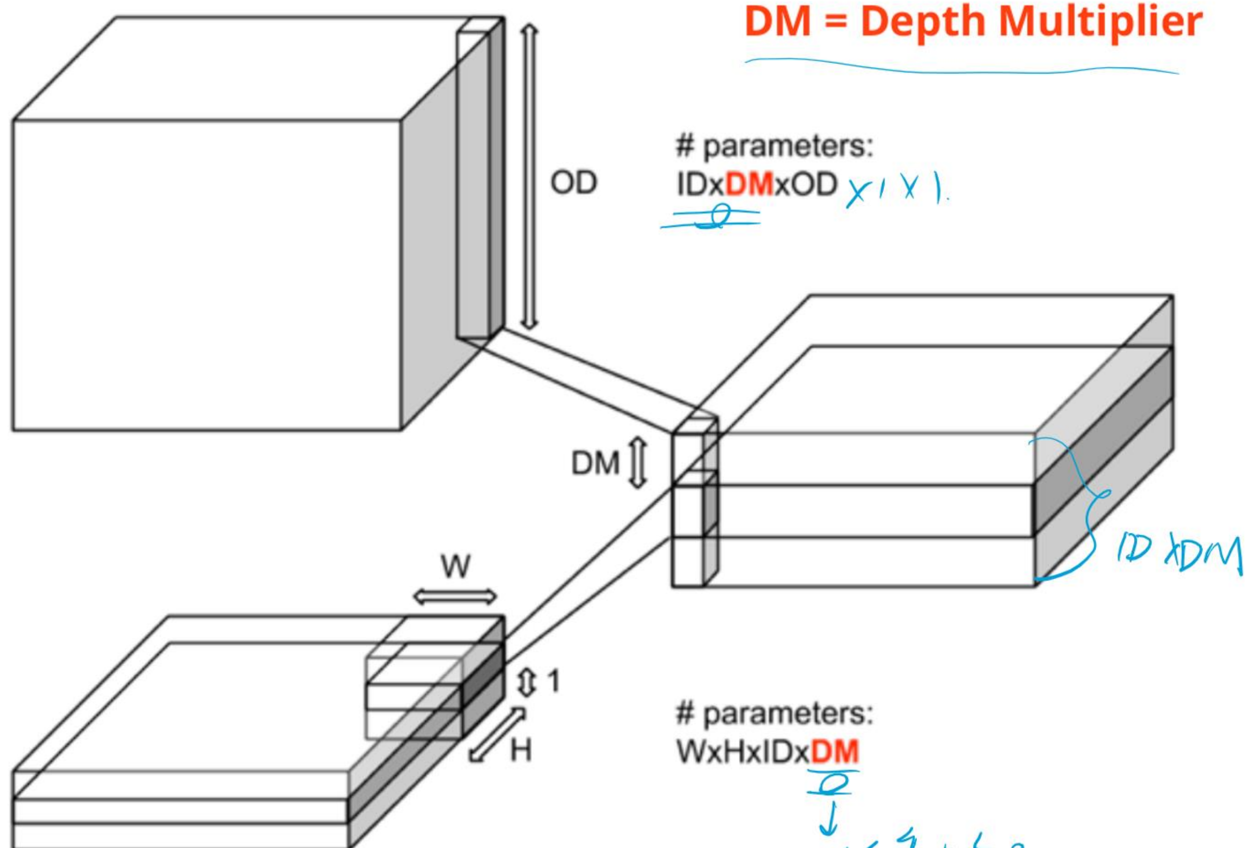
- Deeper and easy to train (Res / Dense)
  - Pre-Activation (2016.07)
  - DenseNet (2016.12)
  - Dual Path Networks (2017.07)
- Wider and light-weight (Group Conv)
  - ResNeXt (2017.04)
  - Xception (2017.04)
  - ShuffleNet (2017.07)
  - Merge-and-Run (2017.07)
  - Interleaved Group Conv (2017.07)
- Global Context
  - Dilated Conv: Dilated-8 (2016.04) and Dilated Residual Network (2017.05)
  - Squeeze-and-Excitation (2017.09)

Wider and light-weight

Group Convolution

Separable Convolution

DM = Depth Multiplier

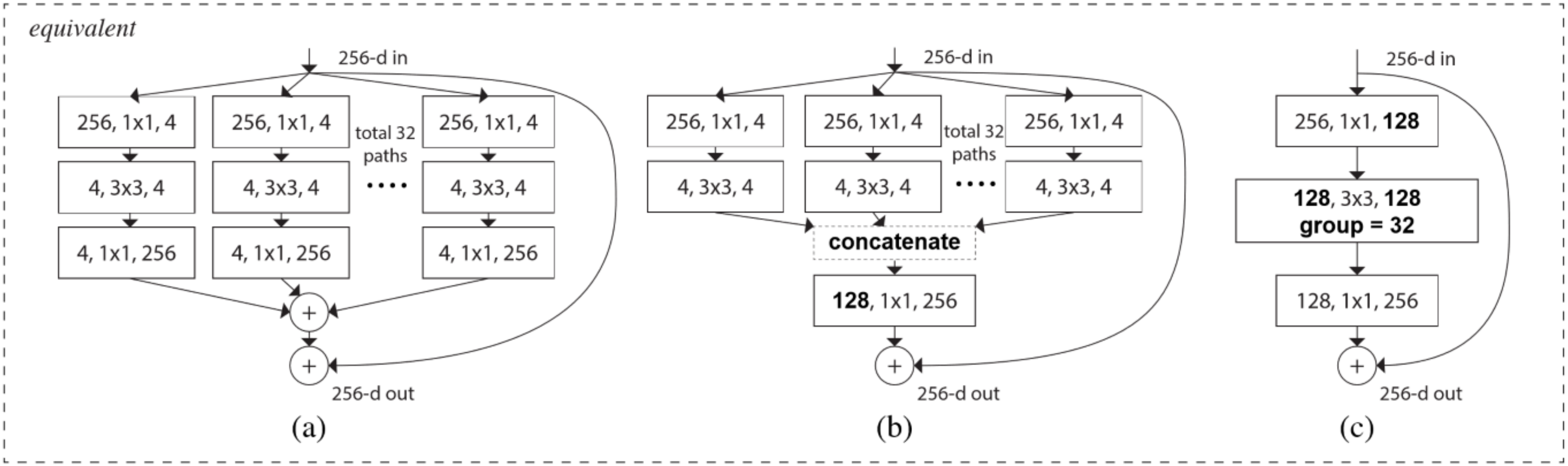


\# parameters:
IDxDMxOD x1x1.

\# parameters:
WxHxIDxDM

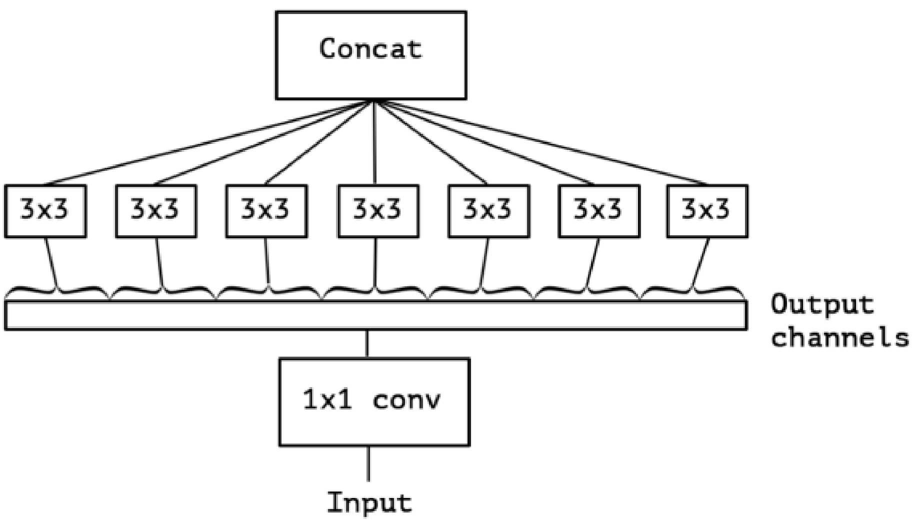
Wider and light-weight

ResNeXt

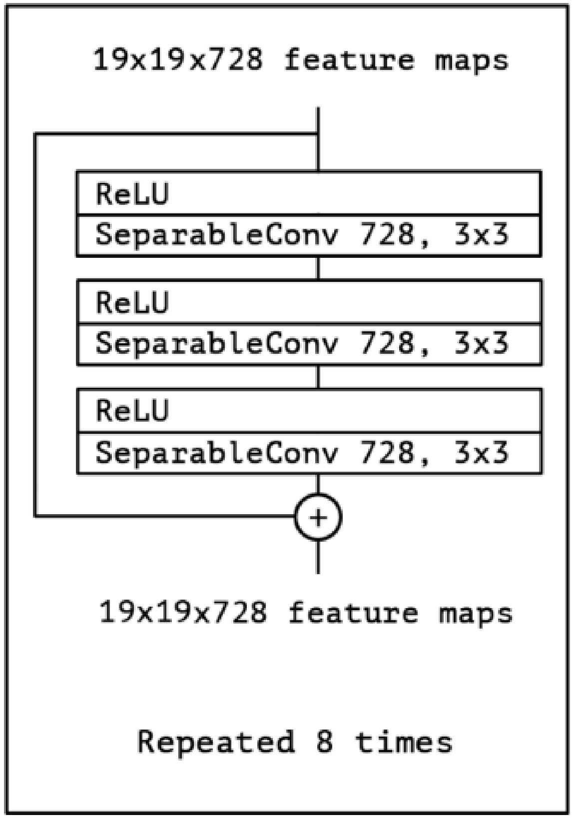


Wider and light-weight

Xception

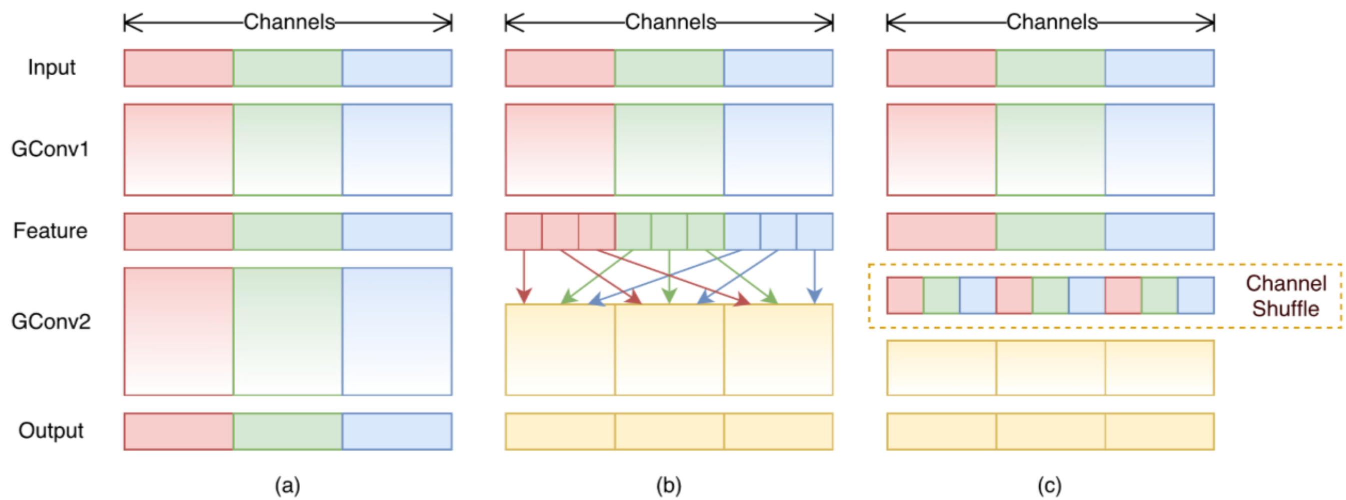


SeparableConv



Wider and light-weight

ShuffleNet



- Based on Xception, also focus on 1x1 conv (pointwise gconv)
- Use “channel shuffle” to encourage interl-channel communication
- Very efficient on small models (never mention on large models)

Wider and light-weight

Merge-and-Run

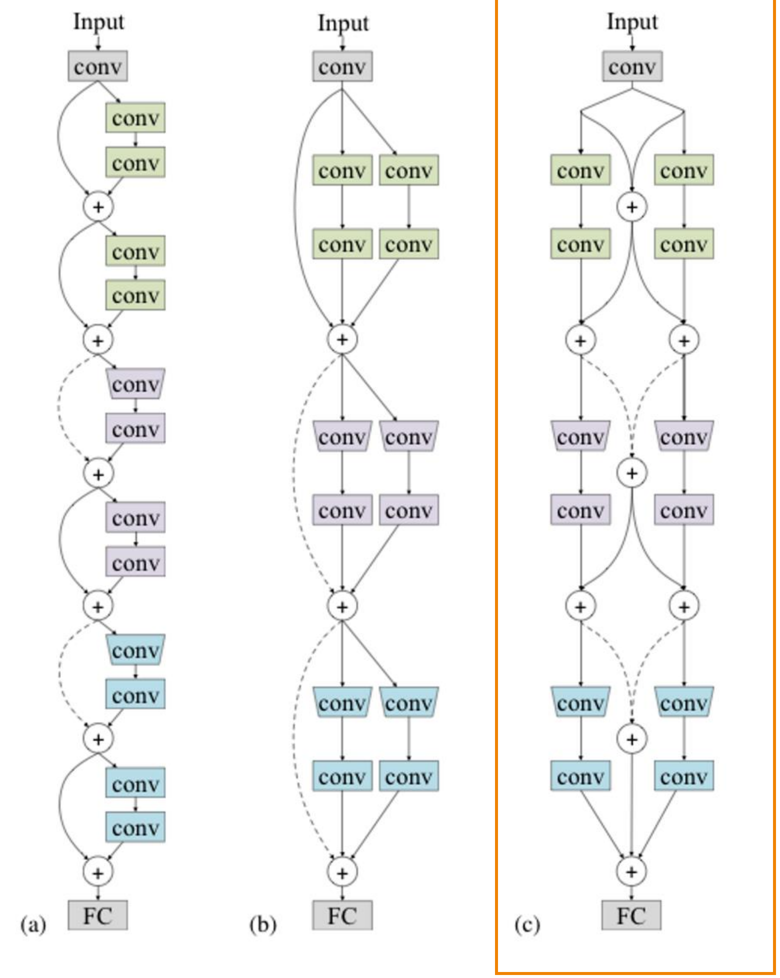
- Like ResNeXt
- Improved information flow

Information flow improvement. We transform Equation 3 into the matrix form,

$$\begin{bmatrix} \mathbf{x}_{2(t+1)} \\ \mathbf{x}_{2(t+1)+1} \end{bmatrix} = \begin{bmatrix} H_{2t}(\mathbf{x}_{2t}) \\ H_{2t+1}(\mathbf{x}_{2t+1}) \end{bmatrix} + \frac{1}{2} \begin{bmatrix} \mathbf{I} & \mathbf{I} \\ \mathbf{I} & \mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{x}_{2t} \\ \mathbf{x}_{2t+1} \end{bmatrix}, \tag{4}$$

where $\mathbf{I}$ is an $d \times d$ identity matrix and d is the dimension of $\mathbf{x}_{2t}$ (and $\mathbf{x}_{2t+1}$). $\mathbf{M} = \frac{1}{2} \begin{bmatrix} \mathbf{I} & \mathbf{I} \\ \mathbf{I} & \mathbf{I} \end{bmatrix}$ is the transformation matrix of the merge-and-run mapping.

- Shorter paths
- Increased width



Wider and light-weight

Interleaved Group Convolution

- Highly related to ShuffleNet (Channel Shuffle operation)

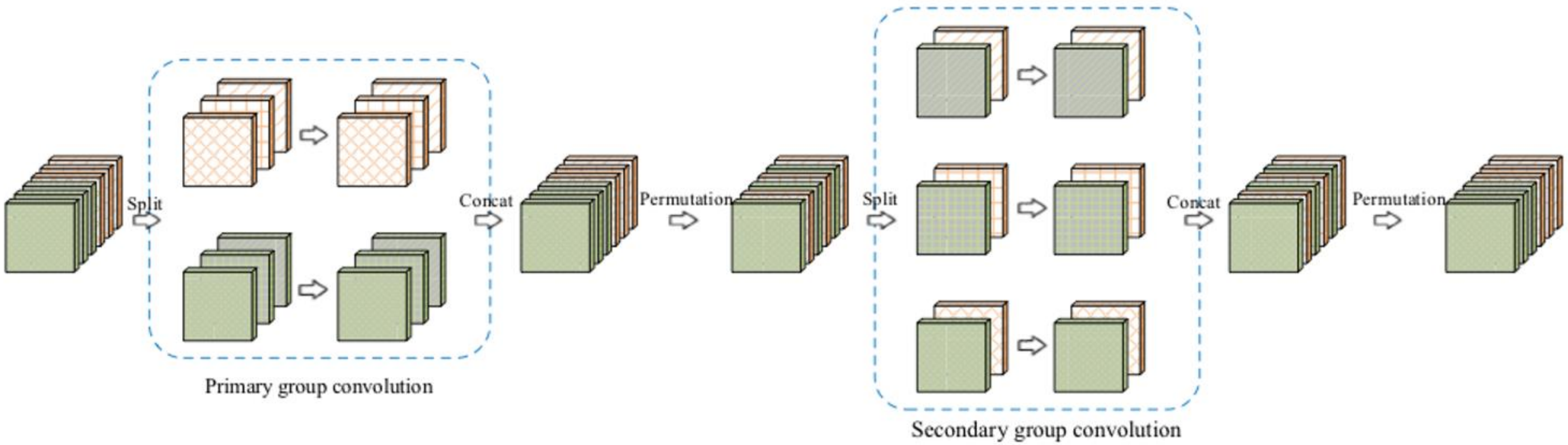


Figure 1. Illustrating the interleaved group convolution, with $L = 2$ primary partitions and $M = 3$ secondary partitions. The convolution for each primary partition in primary group convolution is spatial. The convolution for each secondary partition in secondary group convolution is point-wise (1×1). Details are given in Section 3.1.

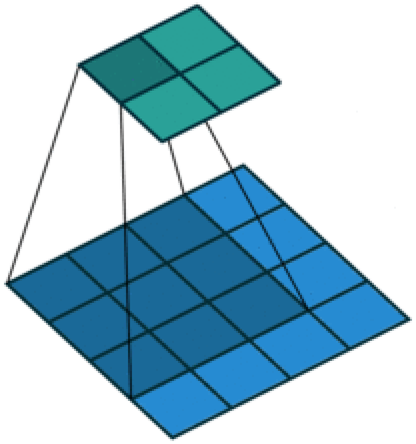
Agenda

*Times are last modified version on arXiv

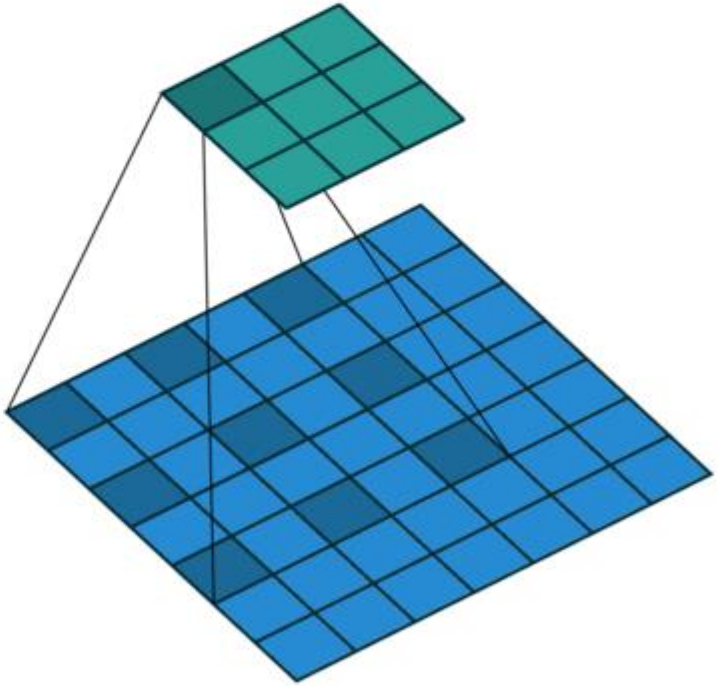
- **Deeper** and easy to train (Res / Dense)
  - Pre-Activation (2016.07)
  - DenseNet (2016.12)
  - Dual Path Networks (2017.07)
- **Wider** and light-weight (Group Conv)
  - ResNeXt (2017.04)
  - Xception (2017.04)
  - ShuffleNet (2017.07)
  - Merge-and-Run (2017.07)
  - Interleaved Group Conv (2017.07)
- Global Context
  - Dilated Conv: Dilated-8 (2016.04) and Dilated Residual Network (2017.05)
  - Squeeze-and-Excitation (2017.09)

Global Context

Dilated Convolution



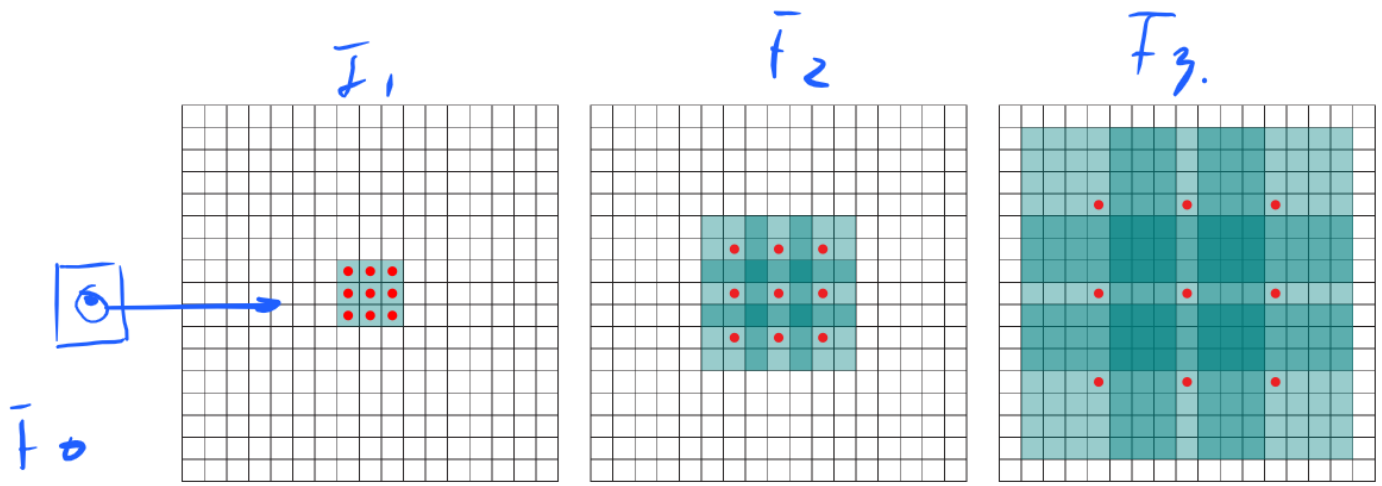
Regular Conv
No padding, no strides



Dilated Conv
No padding, no stride

Global Context

Dilated-8



(a) (b) (c)

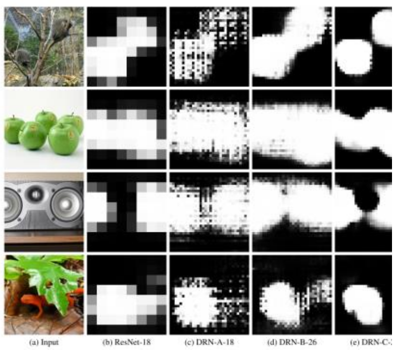
dilated-8

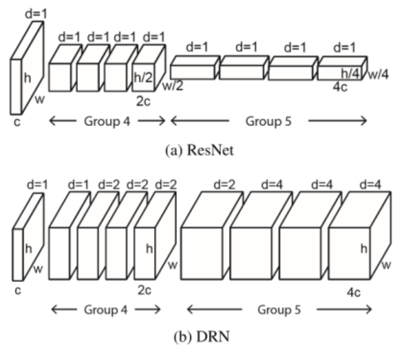
Layer	1	2	3	4	5	6	7	8
Convolution	3×3	3×3	3×3	3×3	3×3	3×3	3×3	1×1
Dilation	1	1	2	4	8	16	1	1
Truncation	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No
Receptive field	3×3	5×5	9×9	17×17	33×33	65×65	67×67	67×67
Output channels								
Basic	C	C	C	C	C	C	C	C
Large	$2C$	$2C$	$4C$	$8C$	$16C$	$32C$	$32C$	C

relu.

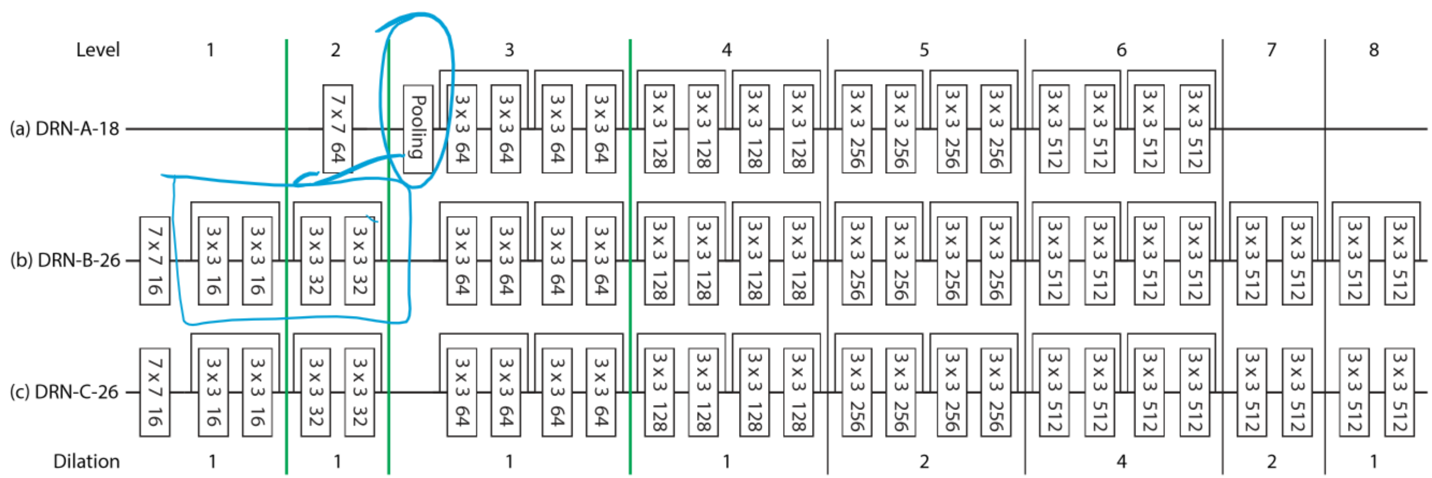
Global Context

Dilated Residual Network



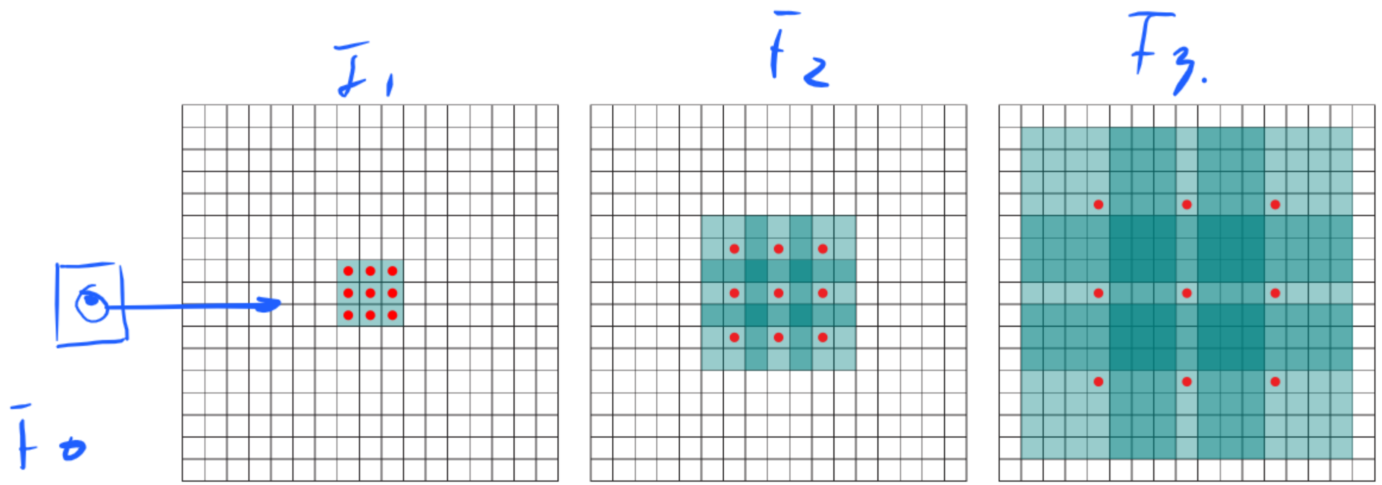


- Low resolution problem
- Degridding for dilated conv
- High res means high memory cost



Global Context

Dilated-8



(a) (b) (c)

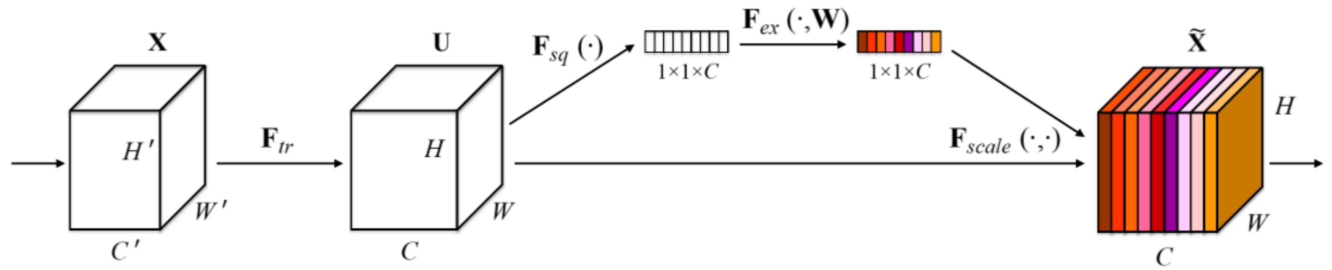
dilated-8

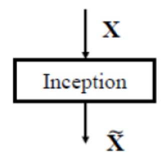
Layer	1	2	3	4	5	6	7	8
Convolution	3×3	3×3	3×3	3×3	3×3	3×3	3×3	1×1
Dilation	1	1	2	4	8	16	1	1
Truncation	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No
Receptive field	3×3	5×5	9×9	17×17	33×33	65×65	67×67	67×67
Output channels								
Basic	C	C	C	C	C	C	C	C
Large	$2C$	$2C$	$4C$	$8C$	$16C$	$32C$	$32C$	C

relu.

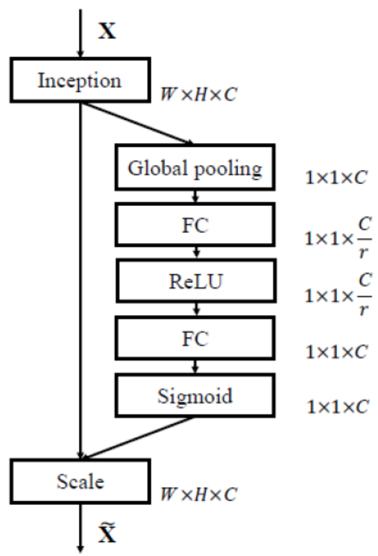
Global Context

Squeeze-and-Excitation

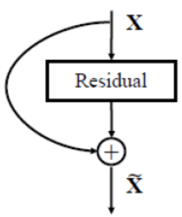




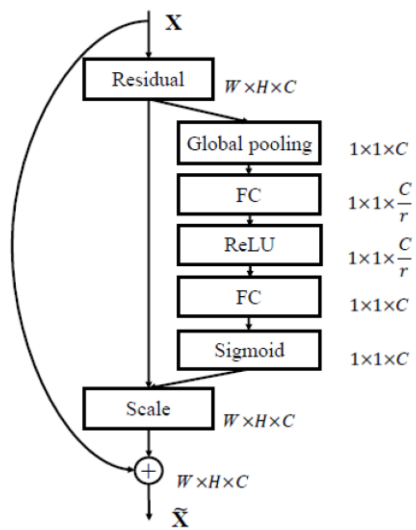
Inception Module



SE-Inception Module



ResNet Module



SE-ResNet Module

Global Context
Squeeze-and-Excitation

- Squeeze: Global Information Embedding
  - Larger than local receptive field of regular conv
- Excitation: Adaptive Recalibration
  - Capture channel-wise dependencies
- Details of ImageNet 2017 best entry (SENet)
  a) Based on ResNeXt-152
  b) Halved bottleneck
  c) 7x7 conv=> 2 stack of 3x3 conv
  d) Down-sampling 1x1 stride-2 conv => 3x3 stride-2 conv
  e) Dropout before fc
  f) Label smoothing
  g) BN parameter frozen for last few epochs
  h) 2048 batch size with initial learning rate of 0.1 SGD on 64 GPUs

Thank You