Involution: Inverting the Inherence of Convolution for Visual Recognition

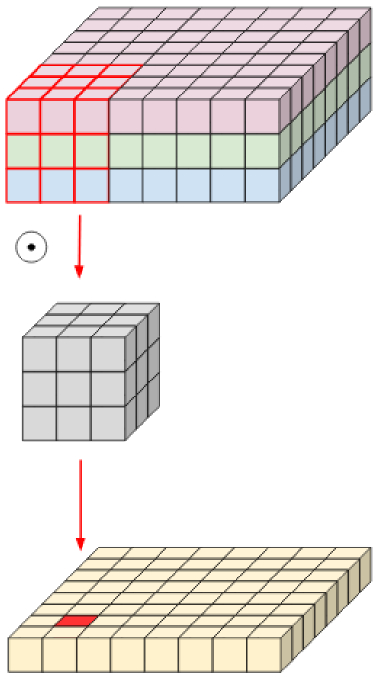
Duo Li, Jie Hu, Changhu Wang, Xiangtai Li, Qi She, Lei Zhu, Tong Zhang, Qifeng Che

CVPR 2021

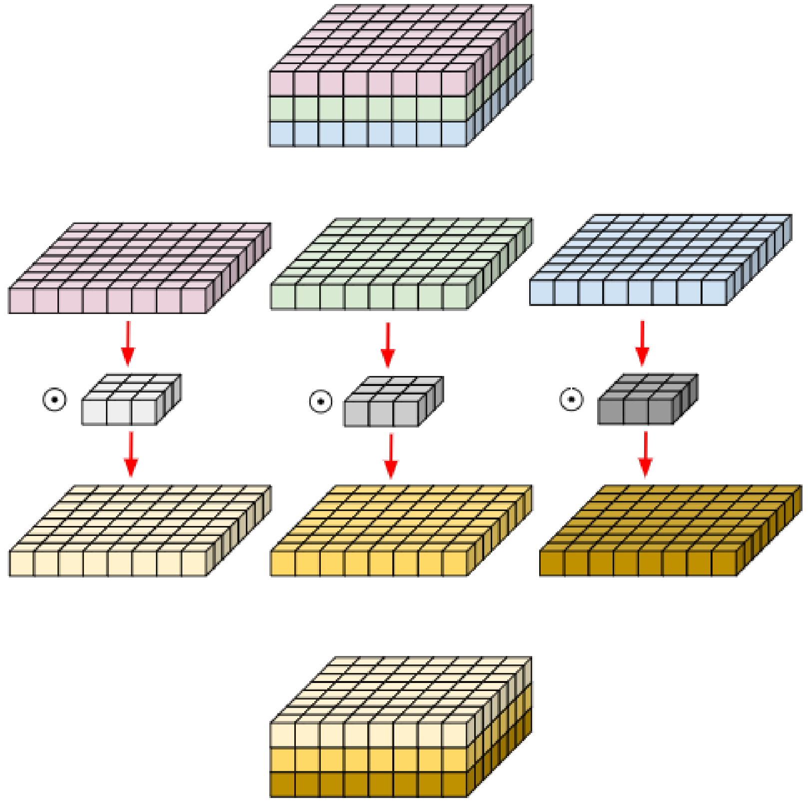
2021.05.31 윤주열

Channel-specific, Spatial-agnostic

- Inherent principle of Conv. layers
- Easily achieves translation equivariance



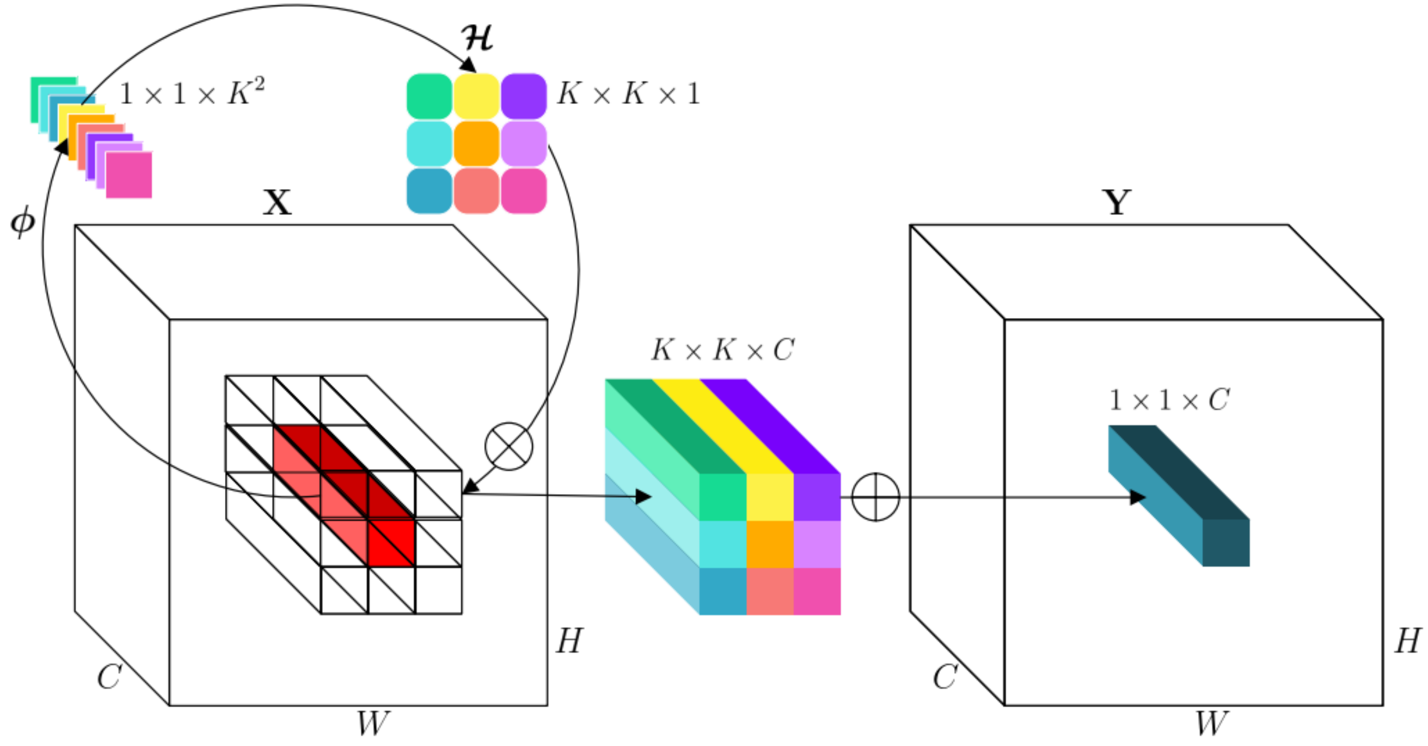
Standard Conv.



Depth-wise Conv.

Involutions

- Channel-agnostic, Spatial-specific kernels
- How do we achieve Spatial-specificness?
  → Generate kernels
- Proposes RedNet
  - Channel-spatial
  - Spatial alone
  - Channel alone
  - All in one



Involutions

- Architecture profile

Architecture	GPU time (ms)	CPU time (ms)	Top-1 Acc. (%)
ResNet-50 [18]	11.4	895.4	76.8
ResNet-101 [18]	18.9	967.4	78.5
SAN19 [64]	33.2	N/A	77.4
Axial ResNet-S [50]	35.9	377.0	78.1
RedNet-38	11.4	156.3	77.6
RedNet-50	14.3	211.2	78.4

Table 2: Runtime analysis for representative networks. The speed benchmark is on a single NVIDIA TITAN Xp GPU and Intel® Xeon® CPU E5-2660 v4@2.00GHz.

Architecture	#Params (M)	FLOPs (G)	Top-1 Acc. (%)
ResNet-26 [18]	13.7	2.4	73.6
LR-Net-26 [20]	14.7	2.6	75.7
Stand-Alone ResNet-26 [39]	10.3	2.4	74.8
SAN10† [64]	10.5	2.2	75.5
RedNet-26	**9.2**	**1.7**	**75.9**
ResNet-38 [18]	19.6	3.2	76.0
Stand-Alone ResNet-38 [39]	14.1	3.0	76.9
SAN15† [64]	14.1	3.0	77.1
RedNet-38	**12.4**	**2.2**	**77.6**
ResNet-50 [18]	25.6	4.1	76.8
LR-Net-50 [20]	23.3	4.3	77.3
AA-ResNet-50 [2]	25.8	4.2	77.7
Stand-Alone ResNet-50 [39]	18.0	3.6	77.6
SAN19† [64]	17.6	3.8	77.4
Axial ResNet-S‡ [50]	**12.5**	3.3	78.1
RedNet-50	15.5	**2.7**	**78.4**
ResNet-101 [18]	44.6	7.9	78.5
LR-Net-101 [20]	42.0	8.0	78.5
AA-ResNet-101 [2]	45.4	8.1	78.7
RedNet-101	**25.6**	**4.7**	**79.1**
ResNet-152 [18]	60.2	11.6	**79.3**
AA-ResNet-152 [2]	61.6	11.9	79.1
Axial ResNet-M‡ [50]	**26.5**	**6.8**	79.2
Axial ResNet-L‡ [50]	45.8	11.6	**79.3**
RedNet-152	34.0	**6.8**	**79.3**

Table 1: The architecture profiles on ImageNet val set. Single-crop testing with 224 × 224 crop size is adopted. We compare with improved re-implementations if available and extract the other results from their original publications.

Results

- Fundamental vision tasks

Detector	Backbone	Neck	Head	#Params (M)	FLOPs (G)	AP^{bbox}	AP^{bbox}_{50}	AP^{bbox}_{75}	AP^{bbox}_{S}	AP^{bbox}_{M}	AP^{bbox}_{L}
Faster R-CNN [40]	ResNet-50	convolution	convolution	41.5	207.1	37.7	58.7	40.8	21.7	41.6	48.4
	RedNet-50	convolution	convolution	31.6	177.9	39.5 (+1.8)	60.9 (+2.2)	42.8 (+2.0)	23.3 (+1.6)	42.9 (+1.3)	52.2 (+3.8)
	RedNet-50	involution	convolution	29.5	135.0	40.2 (+2.5)	62.1 (+3.4)	43.4 (+2.6)	24.2 (+2.5)	43.3 (+1.7)	52.7 (+4.3)
	RedNet-50	involution	involution	29.0	91.5	39.2 (+1.5)	61.0 (+2.3)	42.4 (+1.6)	23.1 (+1.4)	43.0 (+1.4)	50.7 (+2.3)

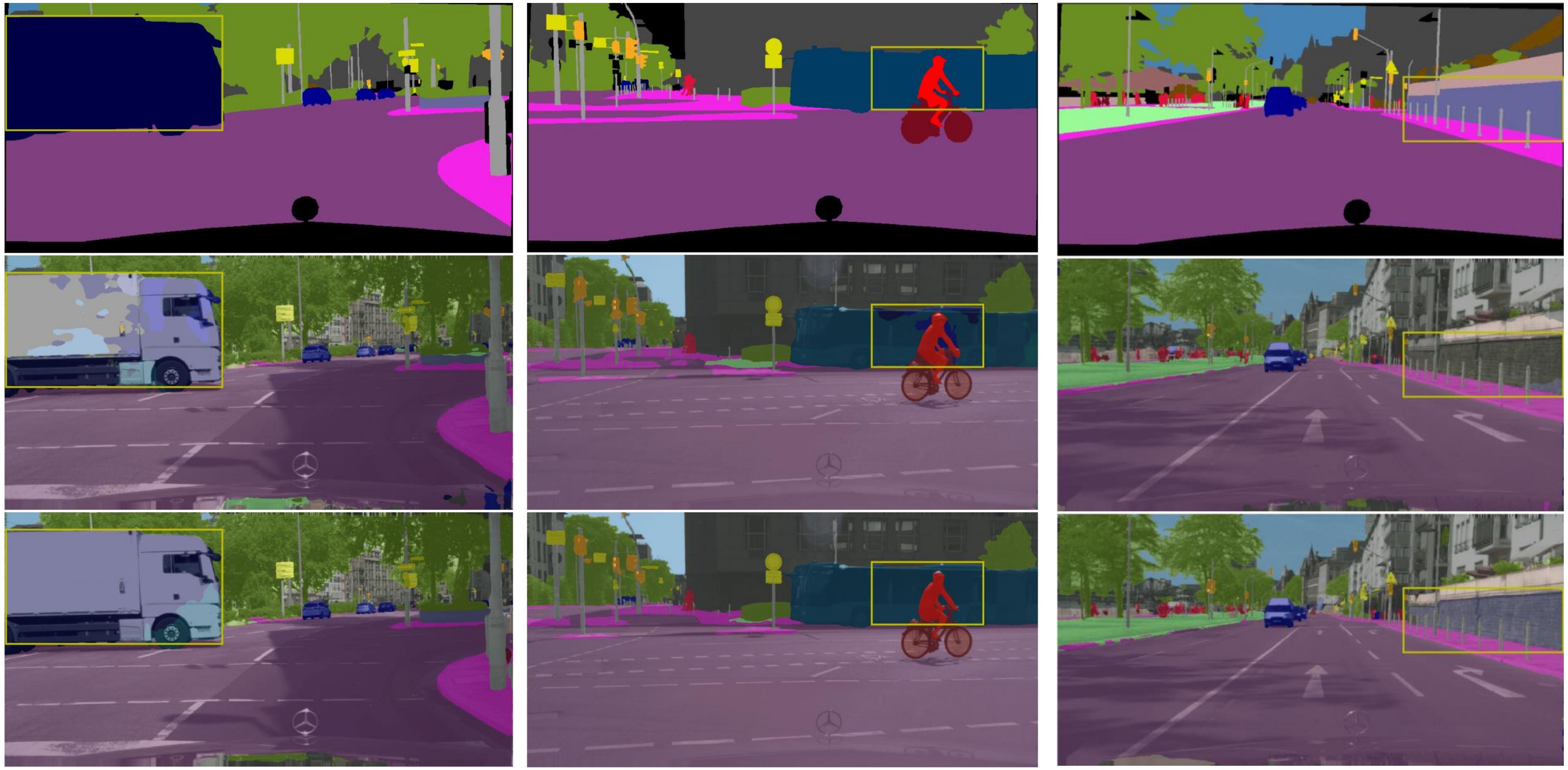
Detector	Backbone	Neck	Head	#Params (M)	FLOPs (G)	AP	AP_{50}	AP_{75}	AP_{S}	AP_{M}	AP_{L}
Mask R-CNN [17]	ResNet-50	convolution	convolution	44.2	253.4	38.4 35.1	59.2 56.3	41.9 37.3	21.9 18.5	42.3 38.6	49.7 46.9
	RedNet-50	convolution	convolution	34.2	224.2	40.2 (+1.8) 36.1 (+1.0)	61.4 (+2.2) 58.1 (+1.8)	43.7 (+1.8) 38.2 (+0.9)	24.2 (+2.3) 19.9 (+1.4)	43.4 (+1.1) 39.3 (+0.7)	52.5 (+2.8) 48.9 (+2.0)
	RedNet-50	involution	convolution	32.2	181.3	40.8 (+2.4) 36.4 (+1.3)	62.3 (+3.1) 59.0 (+2.7)	44.3 (+2.4) 38.5 (+1.2)	24.2 (+2.3) 19.9 (+1.4)	44.0 (+1.7) 39.4 (+0.8)	53.0 (+3.3) 49.1 (+2.2)
	RedNet-50	involution	involution	29.5	104.6	39.6 (+1.2) 35.1 (+0.0)	60.7 (+1.5) 57.1 (+0.8)	42.7 (+0.8) 37.3 (+0.0)	23.5 (+1.6) 19.2 (+0.7)	43.1 (+0.8) 38.5 (−0.1)	51.1 (+1.4) 47.3 (+0.4)

Table 3: Performance comparison on COCO detection and segmentation. The bounding box AP is reported for the object detection track in the upper table. The bounding box and mask AP are simultaneously reported for the instance segmentation track in the lower table, listed in the two separate lines following a single detector. In the parentheses are the gaps to the fully convolution-based counterparts. Highlighted in green are the gaps of at least +2.0 points, the same in Table 4 and 5.

Results

- Qualitative results/ behavior

Segmentor	Backbone	Neck	#Params (M)	FLOPs (G)	mean IoU (%)	wall	truck	bus
	ResNet-50	convolution	28.5	362.8	74.5	39.4	58.6	72.2
Semantic FPN [26]	RedNet-50	convolution	18.5	293.9	78.3 (+3.8)	52.7 (+13.3)	77.3 (+18.7)	87.6 (+15.4)
	RedNet-50	involution	16.4	205.2	79.2 (+4.7)	56.9 (+17.5)	82.1 (+23.5)	88.5 (+16.3)



Results

- Ablation Analysis

Kernel size

Kernel Size	#Params (M)	FLOPs (G)	Top-1 Acc. (%)
3×3	14.7	2.4	76.9
5×5	15.1	2.5	77.4
7×7	15.5	2.6	77.7
9×9	16.2	2.7	77.8

(a) Accuracy saturates with **kernel size** increasing.

Group Channel

#Group Channel	#Params (M)	FLOPs (G)	Top-1 Acc. (%)
1	30.2	5.0	77.9
4	18.5	3.0	77.7
16	15.5	2.6	77.7
C	14.6	2.4	76.5

(b) Appropriate **grouping channels** improves efficiency.

Kernel Generation

Function Form	#params (M)	FLOPs (G)	Top-1 Acc. (%)
$\mathbf{W}$	18.1	3.0	77.8
$\mathbf{W}_1\sigma\mathbf{W}_0, r=1$	19.4	3.2	77.8
$\mathbf{W}_1\sigma\mathbf{W}_0, r=4$	15.5	2.6	77.7
$\mathbf{W}_1\sigma\mathbf{W}_0, r=16$	14.6	2.4	77.4

(c) Introducing the **bottleneck structure** reduces complexity.

- Not Sensitive to hyper-parameters

Results

- Visualization

Sum of K×K values from each kernel (easy to visualize)

Different groups highlight different semantics

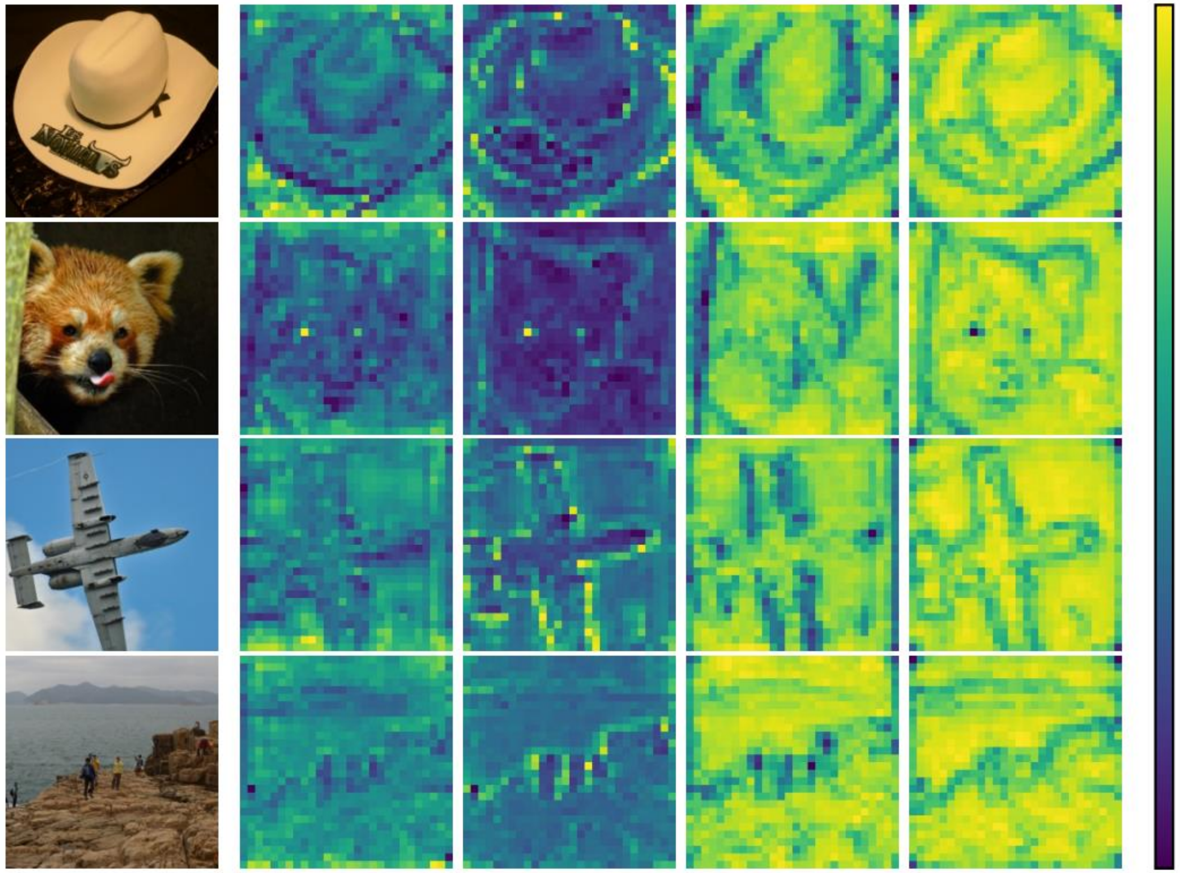


Figure 3: The heat maps in each row interpret the generated kernels for an image instance from the ImageNet validation set, drawn from four different classes, including cowboy hat, lesser panda, warplane, and cliff (from top to bottom).