GCNet: Non-local Networks Meet Squeeze-Excitation Networks and Beyond

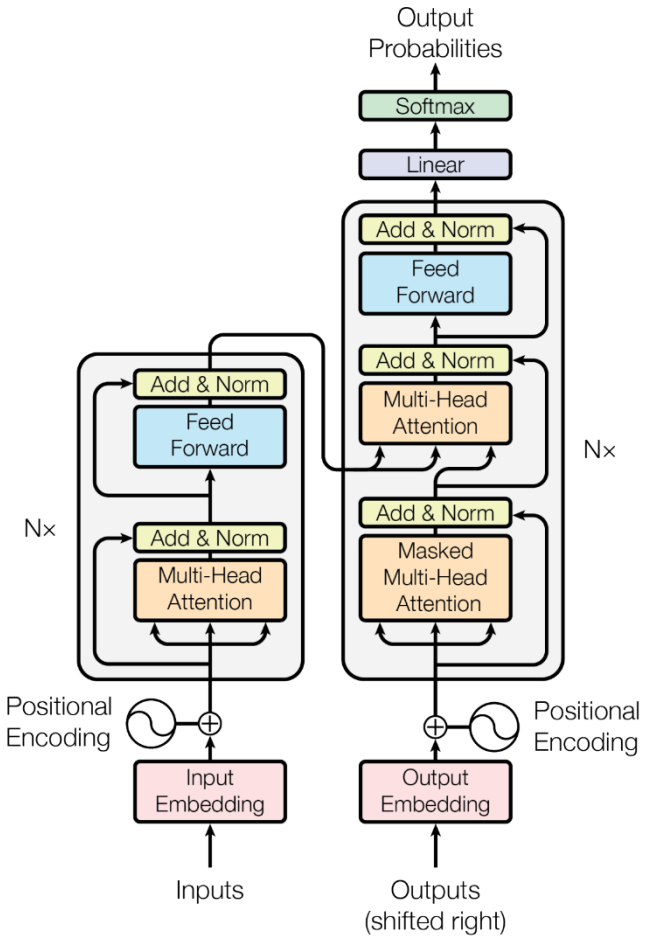
Yue Cao*, Jiarui Xu*, Stephen Lin, Fangyun Wei, Han Hu

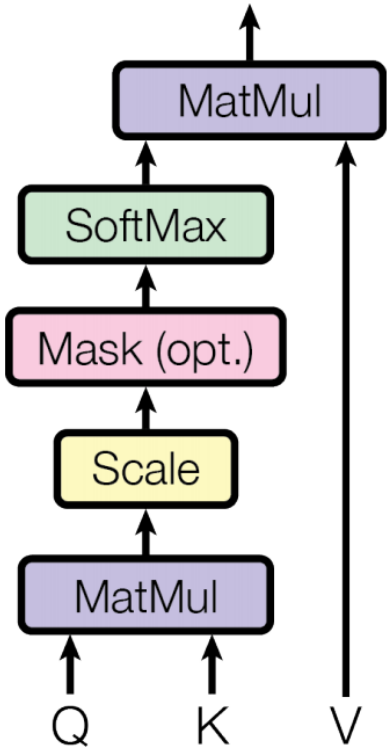
MSRA & HKUST

Code available at:
https://github.com/xvjiarui/GCNet

Related Works: Self Attention Mechanism

- Transformer is a milestone for machine translation, which applies a self-attention mechanism to model long-range dependencies.

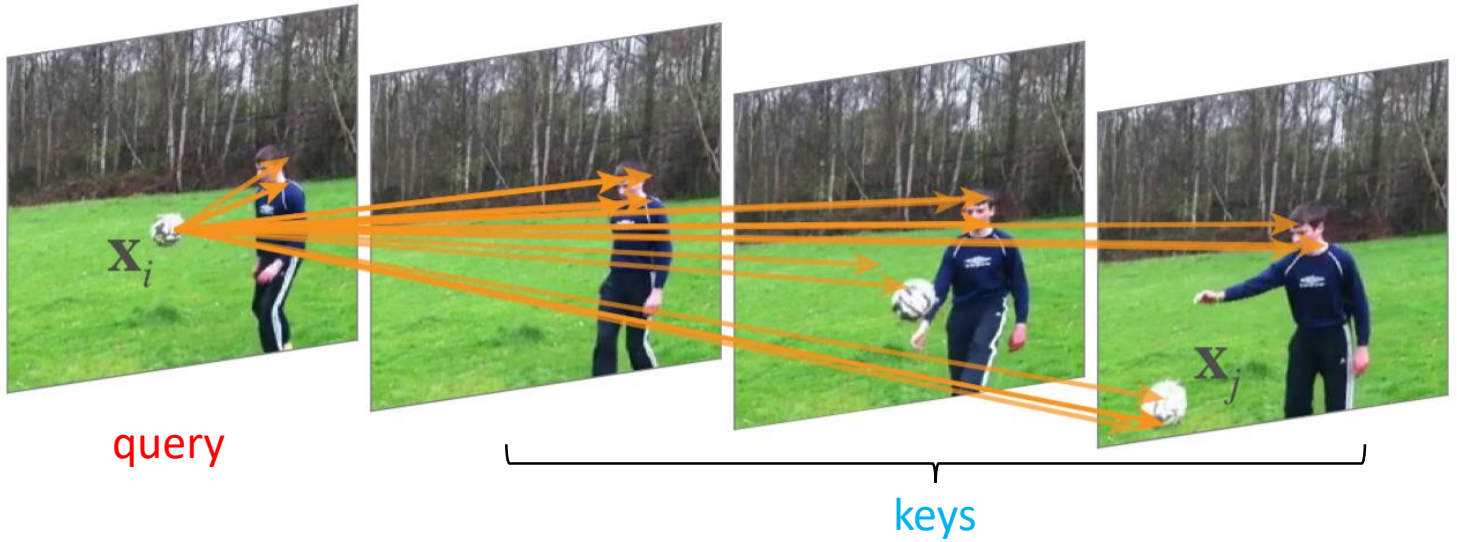




A. Vaswni et al. *Attention is all your need*. NIPS'2017

Related Works: Non-local Neural Networks

- Each query pixel (x_i) will aggregate values from each key pixel (x_j) by attention weight averaging.



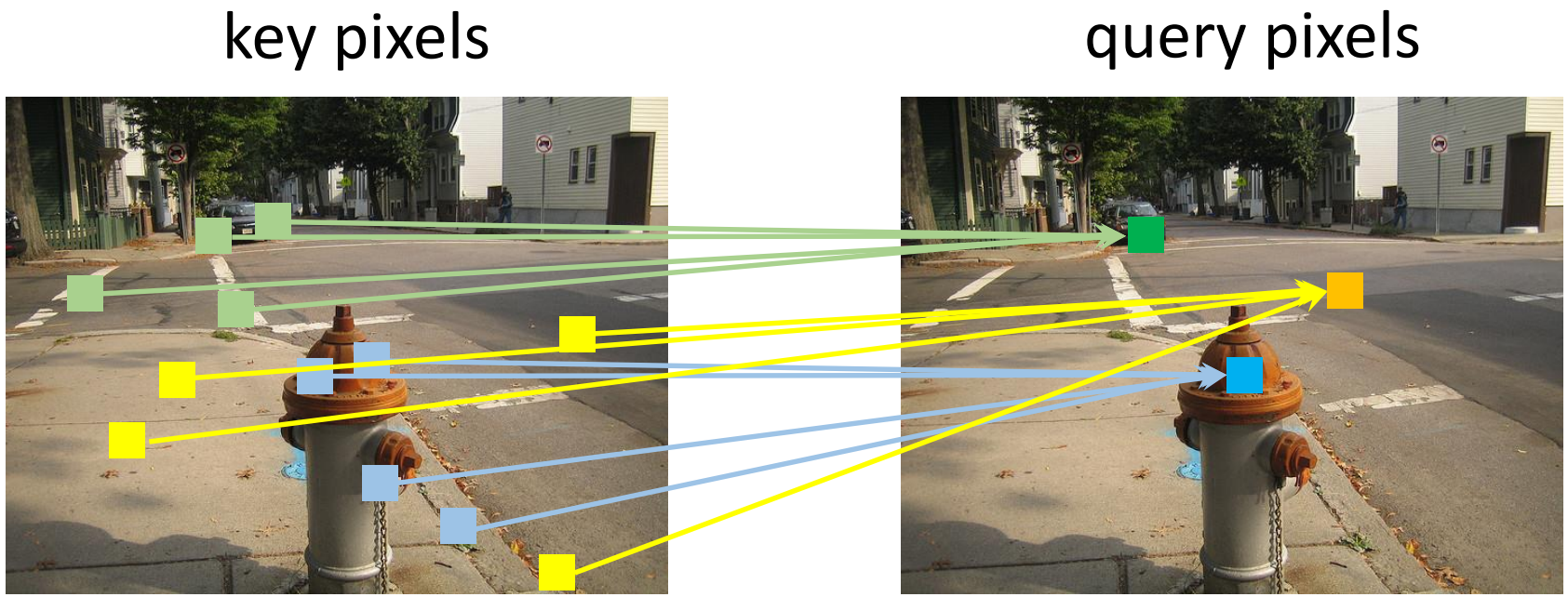
$$\mathrm{y_i} = \frac{1}{C(x)} \sum_{\forall j} f(x_i, x_j)\, g(x_j)$$

- Model dependency between distant pixels (long range dependency)
- Complementary to convolution, which prove to work well on many visual understanding tasks.

X. Wang et al. *Non-local Neural Networks*. CVPR'2018

What Is Expected To Be Learnt

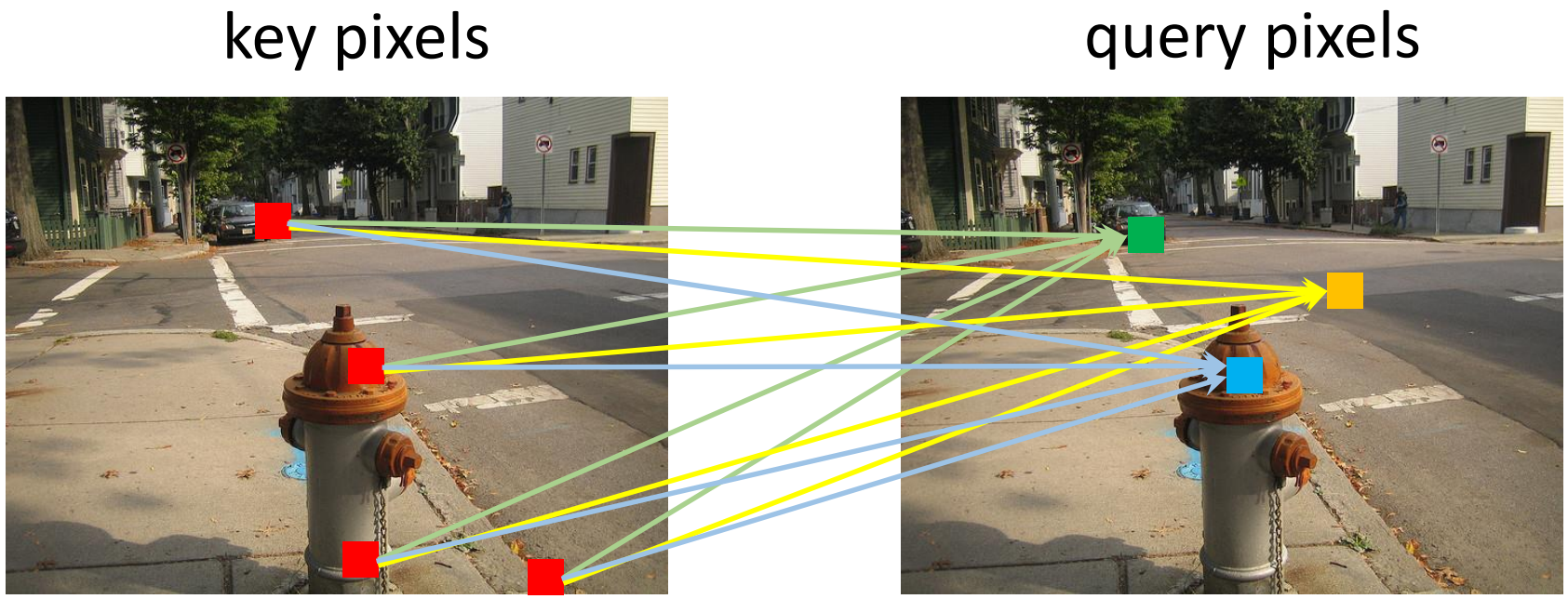
- Different query pixels impacted **by different sets** of key pixels



X. Wang et al. *Non-local Neural Networks*. CVPR'2018

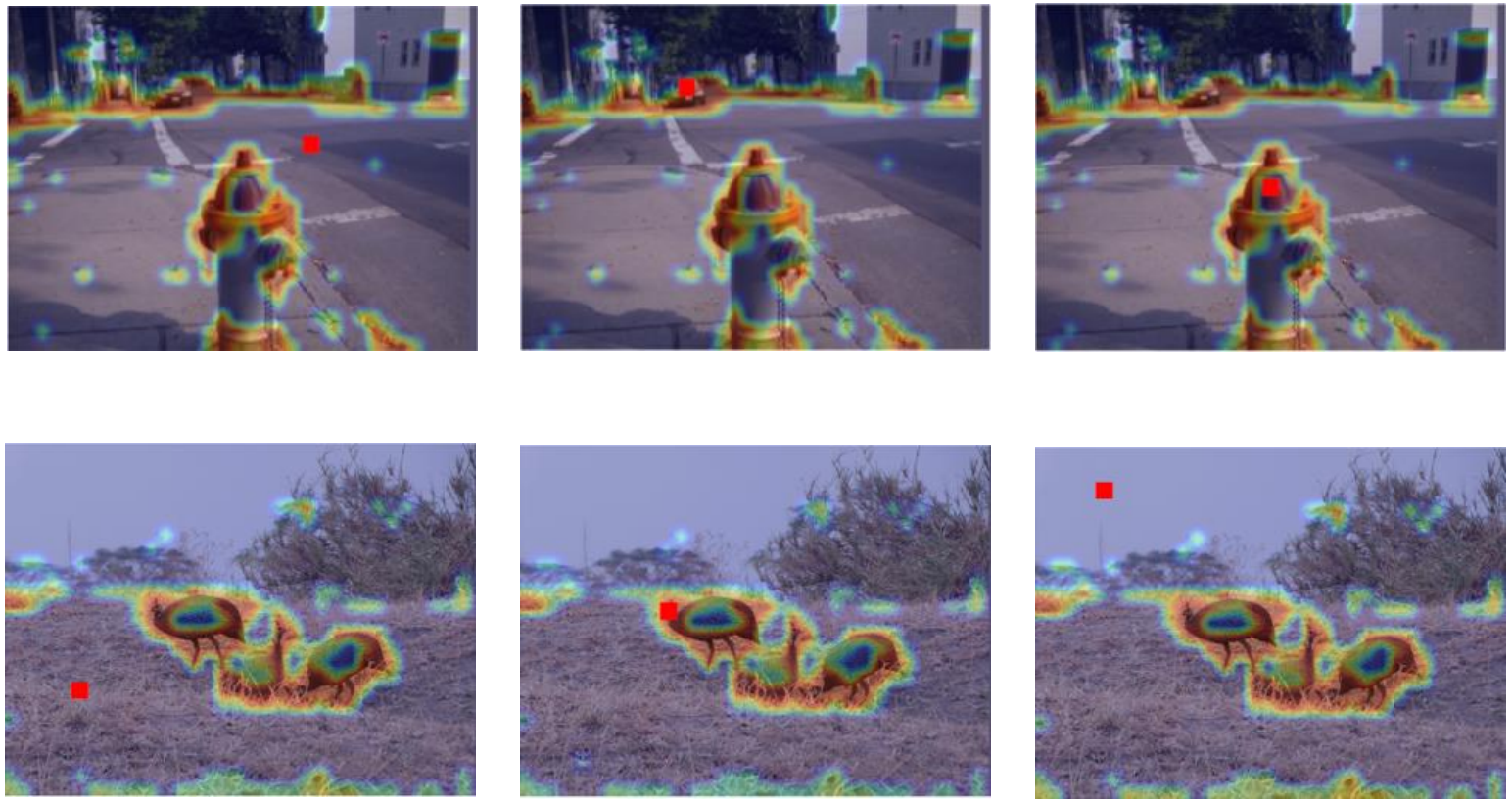
What Is Actually Learnt

- Different query pixels impacted **by the same set** of key pixels



X. Wang et al. *Non-local Neural Networks*. CVPR'2018

Attention Maps For Different Query Pixels



■ query pixels

The effectiveness of non-local neural networks ***do not come from the modeling of dependency between distant pixels, but from the global context modeling****.*

Statistical Analysis On COCO, ImageNet, Kinetics

- We computed the **cosine distance** of different parts of Non-Local Network to verify the dependency.
- It turns out that what Non-Local modeling is query independent, namely global context from statistical perspective.

			Cosine distance		
Dataset	AP(bbox)	AP(mask)	Input	Attention Map	ouput
COCO	38.0	34.7	0.401	0.020	0.012
Dataset	Top-1	Top-5			
ImageNet	77.2	91.9	0.358	0.004	0.003
Dataset	Top-1	Top-5			
Kinetics	75.9	92.2	0.301	0.115	0.074

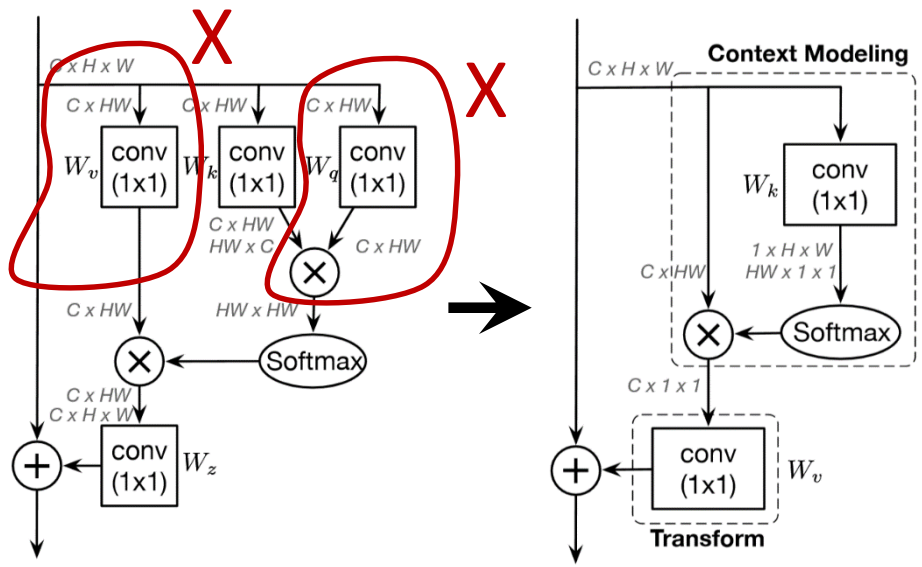
Statistical Analysis On Cityscapes (Exception)

- However, compared with aforementioned 3 datasets, Cityscapes seems to be an exception.

			Cosine distance		
Dataset	AP(bbox)	AP(mask)	Input	Attention Map	ouput
COCO	38.0	34.7	0.401	0.020	0.012
Dataset	Top-1	Top-5			
ImageNet	77.2	91.9	0.358	0.004	0.003
Dataset	Top-1	Top-5			
Kinetics	75.9	92.2	0.301	0.115	0.074
Dataset	mIoU				
Cityscapes	77.59		0.315	0.383	0.354

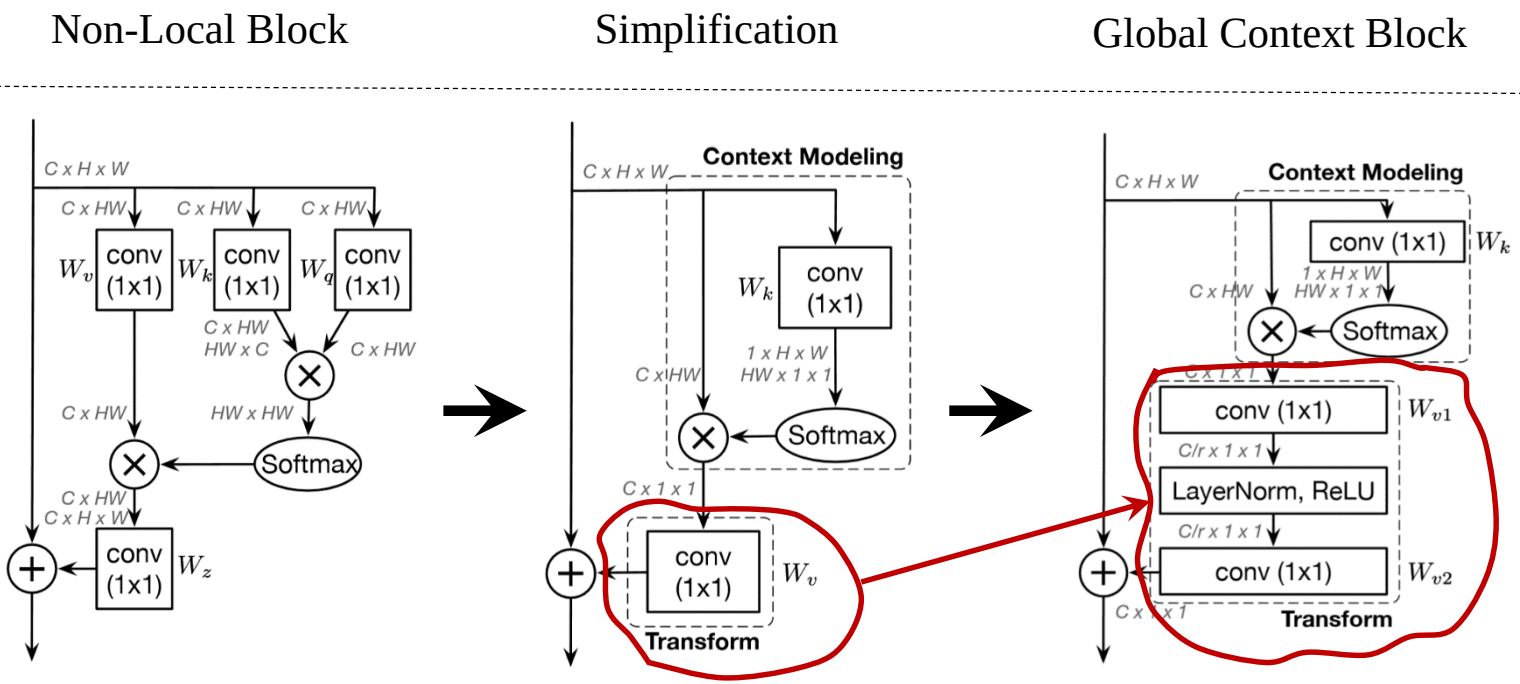
Explicitly Use The Same Attention Map

Non-Local Block Simplification



FLOPs	9.3G	→	**5.0M**
model size	2.1M		1.0M
accuracy (mAP)	38.0		38.1

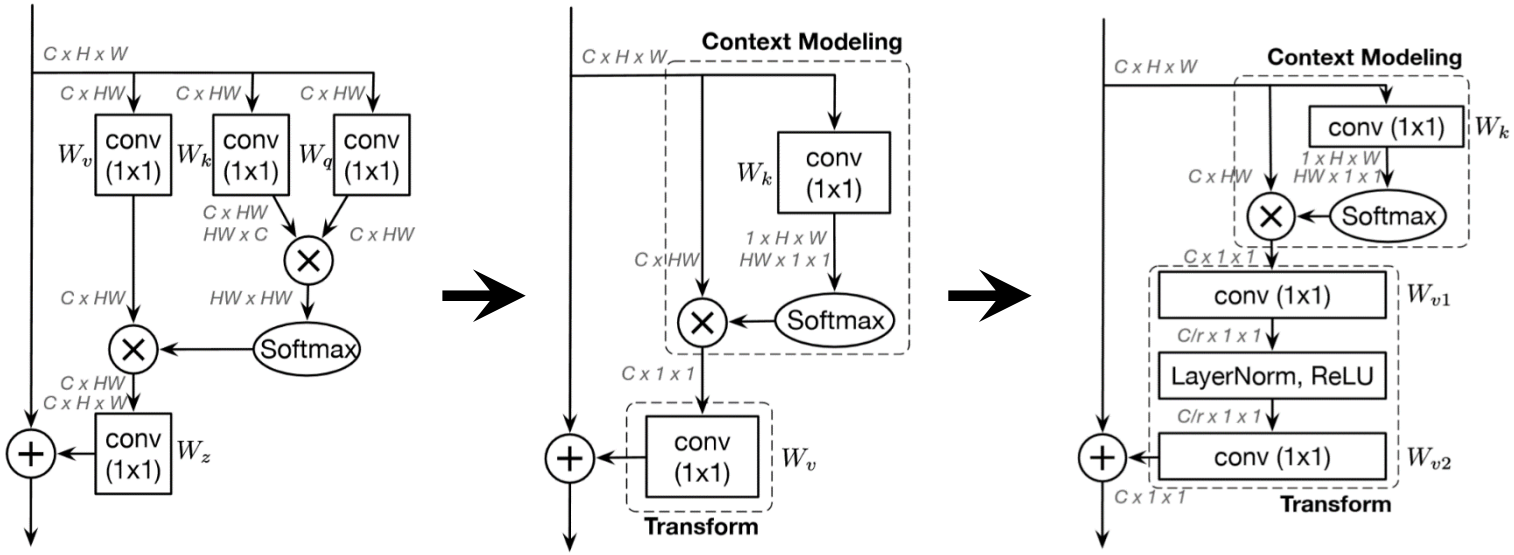
Explicitly Use The Same Attention Map



FLOPs	9.3G	5.0M	4.0M
model size	2.1M	1.0M	**0.1M**
accuracy (mAP)	38.0	38.1	**38.1**

borrowed from SE-Net (champion of 2017 ImageNet Challenge)

Explicitly Use The Same Attention Map

	Non-Local Block	Simplification	Global Context Block	Reduction
				
FLOPs	9.3G	5.0M	**4.0M**	**2,300x**
model size	2.1M	1.0M	**0.1M**	**20x**
accuracy (mAP)	38.0	38.1	**38.1**	**unchanged**

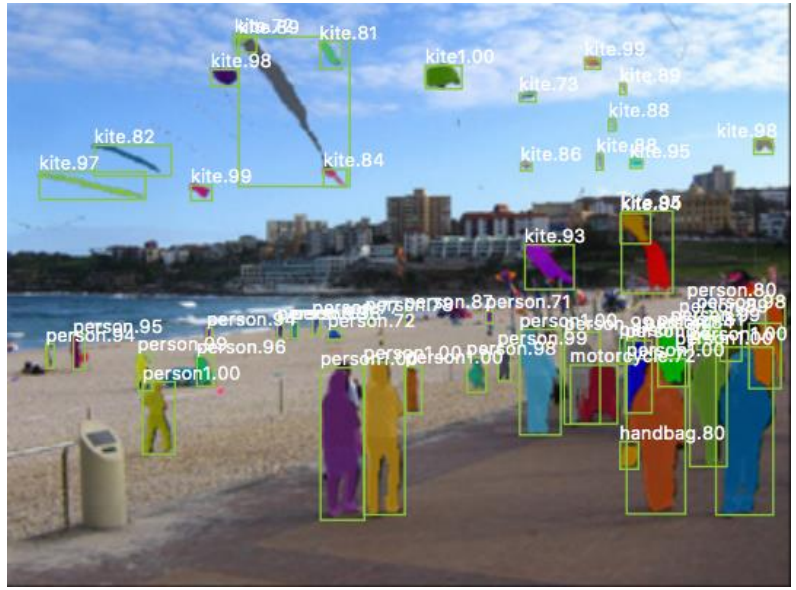
Ablation study of Global Context Network

Tasks	Backbone	Dataset	Evaluation
Image Classification	ResNet-50	ImageNet	Top Acc
Object Detection	Faster R-CNN+FPN+ResNet-50	COCO	Mean AP
Action Recognition	ResNet-50 Slow only	Kinetics 500	Top Acc
Semantic Segmentation	Dilated ResNet-101	Cityscapes	Mean IoU

COCO Object Detection Results

- Baseline: Mask R-CNN + ResNet50 + FPN

method	AP (bbox)	AP (mask)	#param	FLOPs
baseline	37.2	33.8	44.4M	279.4G
NL-Net	38.0	34.7	46.5M	288.7G
SNL-Net	38.1	35.0	45.4M	279.4G
GC-Net (1 block)	38.1	34.9	44.5M	279.4G
GC-Net (all layers)	39.4	35.7	46.9M	279.6G



+2.2 mAP +1.9 mAP

with little computation and model size overhead!

ImageNet Image Classification Results

- Baseline: ResNet-50

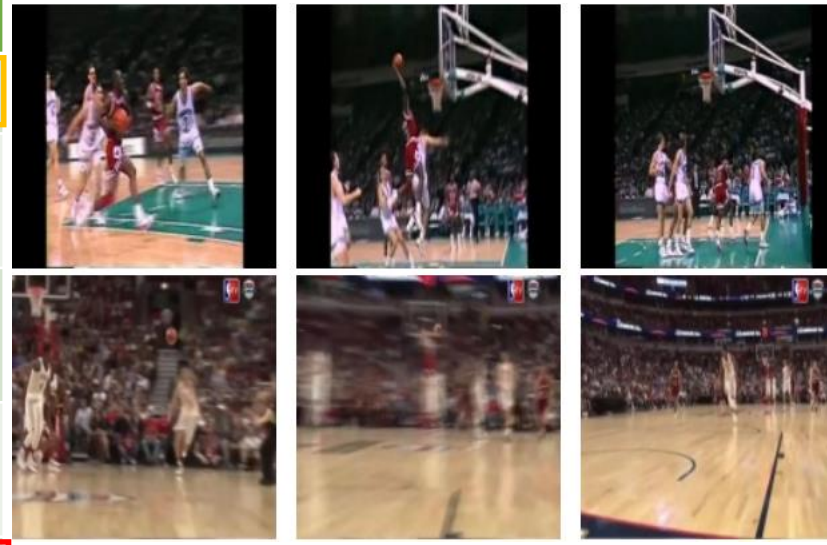
method	Top-1 Acc	Top-5 Acc	#param	FLOPs
baseline	76.51	93.35	25.56M	3.86G
NL-Net	77.21	93.64	27.66M	4.11G
SNL-Net	77.10	93.56	26.61M	3.86G
GC-Net (1 layer)	77.20	93.47	25.69M	3.86G
GC-Net (all layers)	77.49	93.67	28.08M	3.87G




Kinetics Action Recognition Results

- Baseline: ResNet-50 Slow-only

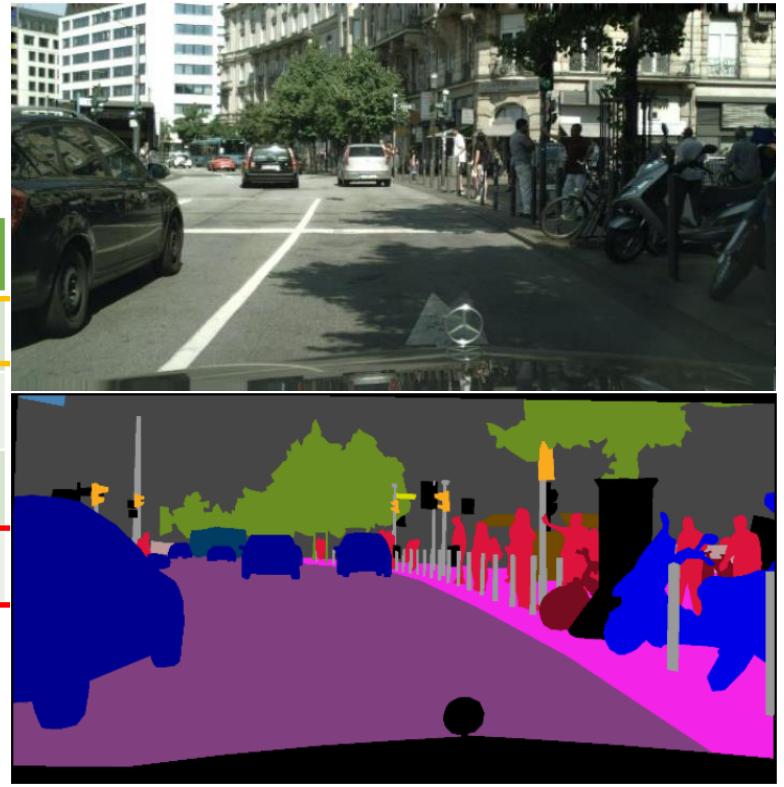
method	Top-1 Acc	Top-5 Acc	#param	FLOPs
baseline	74.94	91.90	32.45M	39.29G
NL-Net(5 blocks)	75.95	92.29	39.81M	59.60G
SNL-Net(5 blocks	75.76	92.44	36.13M	39.32G
GC-Net (5 blocks)	75.85	92.25	34.30M	39.31G
GC-Net (all layers)	76.00	92.34	42.45M	39.35G



Cityscapes Semantics Segmentation Results

- Baseline: ResNet101 Dilated

method	mIoU	#param	FLOPs
baseline	75.42%	70.96M	646.88G
NL-Head	77.59%	71.22M	649.36G
SNL-Head	77.22%	71.22M	646.86G
GC-Head	78.55%	71.09M	646.89G



COCO Object Detection Results

- Stronger backbone

backbone	method	AP (bbox)	AP (mask)	#param	FLOPs
ResNet-50	Baseline	37.2	33.8	44.4M	279.4G
	+GC r16	39.4	35.7	46.9M	279.5G
	+GC r4	**39.9**	**36.2**	54.4M	279.6G
ResNet-101	Baseline	39.8	36.0	63.4M	354.1G
	+GC r16	41.1	37.4	68.1M	354.2G
	+GC r4	**41.7**	**37.6**	82.4M	354.3G
ResNeXt-101	Baseline	41.2	37.3	63.0M	357.8G
	+GC r16	42.4	38.0	67.8M	358.1G
	+GC r4	**42.9**	**38.5**	81.9M	358.2G

COCO Object Detection Results

- Stronger method

backbone	method	AP (bbox)	AP (mask)	#param	FLOPs
ResNeXt-101	Baseline	41.2	37.3	63.0M	357.8G
	+GC r16	42.4	38.0	67.8M	358.1G
	+GC r4	**42.9**	**38.5**	81.9M	358.2G
ResNeXt-101 +Cascade	Baseline	44.7	38.3	95.9M	536.9G
	+GC r16	45.9	39.3	100.7M	537.2G
	+GC r4	**46.5**	**39.7**	114.9M	537.3G
ResNeXt-101 +DCN +Cascade	Baseline	47.1	40.4	98.5M	547.5G
	+GC r16	47.9	40.9	103.3M	547.7G
	+GC r4	**47.9**	**40.8**	117.5M	547.8G

Conclusion

- We have found empirically that **non-local network only models query-independent context** on several important visual recognition tasks.
- We simplify non-local networks while preserve the long-range dependency modeling capability and performance.
- We proposed a novel Global Context Network which can effectively model long-range dependency with light computation, which shows consistent improvements on four fundamental benchmarks.