# ION

# Inside-Outside Net: Detecting Objects in Context with Skip Pooling and Recurrent Neural Networks

Sean Bell[1], C. Lawrence Zitnick[2], Kavita Bala[1], Ross Girshick[2].

*[1]Cornell University. [2]Microsoft Research (now both at Facebook AI Research).*

Microsoft Research

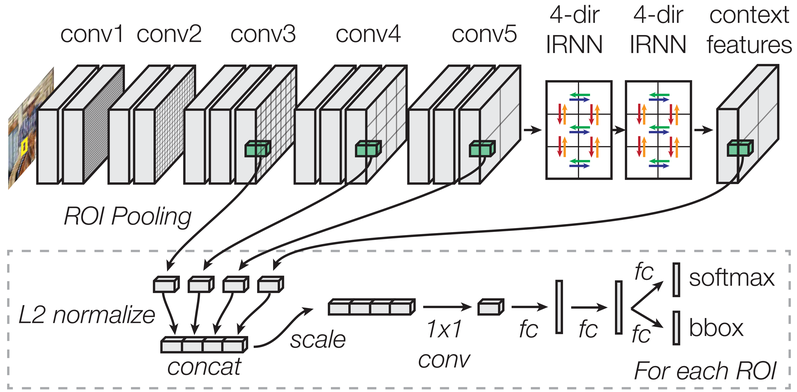
## Highlights

- New object detection architecture: "ION"
- Extensive experiments to validate design decisions
- MS COCO 2015: Won Best Student Entry (3rd overall)

## ION Architecture

- Based on Fast R-CNN [Girshick 2015]
- **Skip pooling:** pool from multiple layers (+3.8 mAP*)
- **Context features:** lateral stacked RNNs (+1.9 mAP*)
- **Normalization:** L2 normalize each pooled blob and re-scale [Liu 2015] (doesn't work without it)

**metric: object detection on PASCAL VOC 2007 test*

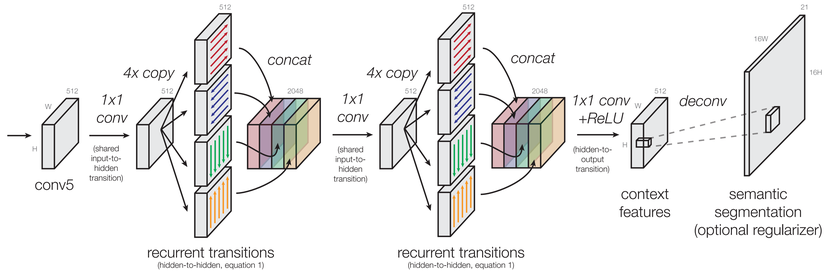



## Context Features

### 2x Stacked 4-Direction IRNN

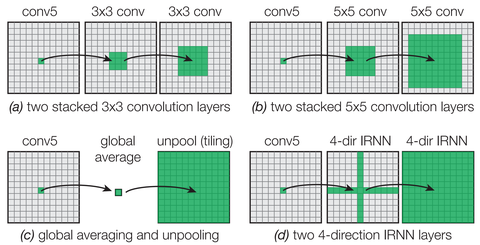
- 4 RNNs move in different directions, laterally across the feature map
- Stack multiple RNNs together (input/output have same shape)
- We use IRNNs: ReLU RNNs initialized with the identity
- Simplify the update rule using 1x1 convolutions:

$$h_{i,j}^{\text{right}} \leftarrow \max\left(\mathbf{W}_{hh}^{\text{right}} h_{i,j-1}^{\text{right}} + h_{i,j}^{\text{right}}, 0\right)$$




### Comparing approaches to adding context

2x stacked IRNN output is both "global" and "local": every cell in the output depends on every cell in the input, and it is spatially varying




| Context method | Seg. | mAP |
|---|---|---|
| Without context | | 74.6 |
| (a) 2x stacked 512x3x3 conv | | 74.8 |
| (b) 2x stacked 256x5x5 conv | | 74.6 |
| (c) Global average pooling | | 74.9 |
| (d) 2x stacked 4-dir IRNN | | **75.6** |
| (a) 2x stacked 512x3x3 conv | ✓ | 75.2 |
| (d) 2x stacked 4-dir IRNN | ✓ | **76.5** |

Table 7. **Comparing approaches to adding context.** All rows also pool out of conv3, conv4, and conv5. Metric: detection mAP on VOC07 test. **Seg:** if checked, the top layer received extra supervision from semantic segmentation labels.

## Results

### PASCAL VOC 2007

| Method | Boxes | R | W | D | Train | Time | mAP |
|---|---|---|---|---|---|---|---|
| FRCN [9] | SS | | | | 07+12 | 0.3s | 70.0 |
| Faster [32] | RPN | | | | 07+12 | 0.2s | 73.2 |
| MR-CNN [8] | SS+EB | ✓ | | | 07+12 | 30s | 78.2 |
| ION [ours] | SS | | | | 07+12 | 0.8s | 74.6 |
| ION [ours] | SS | ✓ | | | 07+12 | 0.8s | 75.6 |
| ION [ours] | SS | ✓ | ✓ | ✓ | 07+12 | 1.2s | 77.6 |
| ION [ours] | SS+EB | ✓ | ✓ | ✓ | 07+12 | 2.0s | 79.4 |
| ION [ours] | SS | ✓ | | | 07++12+S | 0.8s | 76.5 |
| ION [ours] | SS | ✓ | ✓ | | 07++12+S | 1.2s | 78.5 |
| ION [ours] | SS | ✓ | ✓ | ✓ | 07++12+S | 1.2s | 79.2 |
| ION [ours] | SS+EB | ✓ | ✓ | ✓ | 07++12+S | 2.0s | **80.1** |

### PASCAL VOC 2012

| Method | Boxes | R | W | D | Train | mAP |
|---|---|---|---|---|---|---|
| FRCN [9] | SS | | | | 07++12 | 68.4 |
| Faster [32] | RPN | | | | 07++12 | 70.4 |
| FRCN+YOLO [31] | SS | | | | 07++12 | 70.4 |
| HyperNet [21] | RPN | | | | 07++12 | 71.4 |
| MR-CNN [8] | SS+EB | ✓ | | | 07+12 | 73.9 |
| ION [ours] | SS | ✓ | ✓ | ✓ | 07+12 | 74.7 |
| ION [ours] | SS+EB | ✓ | ✓ | ✓ | 07+12 | 76.4 |
| ION [ours] | SS | ✓ | ✓ | ✓ | 07++12+S | 76.4 |
| ION [ours] | SS+EB | ✓ | ✓ | ✓ | 07++12+S | **77.9** |

**Legend**
**07+12:** 07 trainval + 12 trainval,
**07++12:** 07+12 + 07 test
**07++12+S:** 07+12 plus SBD segmentation labels [Hariharan 2011],
**R:** Use IRNN context features,
**W:** 2 rounds of bbox regression [Gidaris 2015],
**D:** no dropout,
**SS:** Selective Search [Uijlings 2013],
**EB:** Edge Boxes [Zitnick 2014],
**RPN:** Region Proposal Network [Ren 2015].
See paper for details.

### MS COCO 2015 Competition

| | test-comp. | test-dev | runtime | test scales | ensemble |
|---|---|---|---|---|---|
| Ours (competition) | 31.0 mAP | 31.2 mAP | 2.7s | 1 | 1 net |
| Ours (post-competition) | | 33.1 mAP | 5.5s | 1 | 1 net |
| ResNet [He 2015] | | 32.2 mAP | | 1 | 1 net |
| ResNet [He 2015] | | 34.9 mAP | | 2 | 1 net |
| ResNet [He 2015] (winner) | 37.1 mAP | 37.4 mAP | | 2 | 3 nets |

### Accuracy Breakdown (MS COCO 2015 test-dev, post-competition)

| | |
|---|---|
| +5.1 mAP: | New ION detection architecture |
| +3.9 mAP: | Better box proposals (RPN + MCG), more data (train+val) |
| +1.3 mAP: | Semantic segmentation (regularizer, not used for test) |
| +1.3 mAP: | Iterative bbox regression [Gidaris 2015] with new thresholds |
| +0.8 mAP: | Larger mini-batches (4 images/batch) and longer training |
| +0.8 mAP: | Left/right flips during inference (average results) |
| +0.6 mAP: | Remove dropout |
| +0.2 mAP: | Add two 3x3 convolution layers after conv5 |

### References

R. Girshick. "Fast R-CNN." ICCV 2015.
S. Gidaris, N. Komodakis. "Object detection via a multi-region & semantic segmentation-aware CNN model." ICCV 2015.
S. Ren, K. He, R. Girshick, J. Sun. "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks." NIPS 2015.
P. Arbelaez, J. Pont-Tuset, J. T. Barron, F. Marques, J. Malik. "Multiscale Combinatorial Grouping." CVPR 2014.
J. Uijlings, K. van de Sande, T. Gevers, A. Smeulders. "Selective search for object recognition". IJCV 2013.
B. Hariharan, P. Arbelaez, L. Bourdev, S.Maji, J. Malik. "Semantic contours from inverse detectors." ICCV 2011.
C. L. Zitnick and P. Dollar. "Edge boxes: Locating object proposals from edges." ECCV 2014.
W. Liu, A. Rabinovich, A. C. Berg. "ParseNet: Looking Wider to See Better". arXiv 2015.
K. He, X. Zhang, S. Ren, J. Sun, "Deep Residual Learning for Image Recognition". CVPR 2016.