# R-FCN: Object Detection via Region-based Fully Convolutional Networks

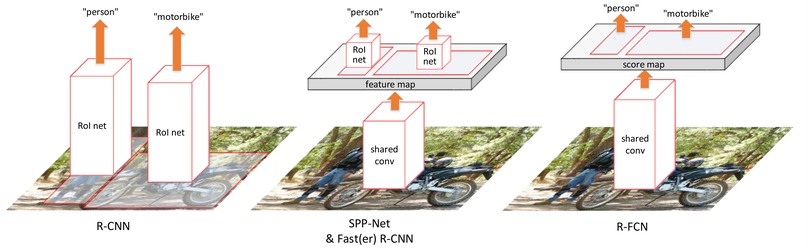
Jifeng Dai, Yi Li, Kaiming He, Jian Sun
Microsoft Research Asia

Microsoft Research 微软亚洲研究院

## Highlights

- Region-based, fully-convolutional networks for object detection
- 1-st in VOC 2012 detection leaderboard (comp4)
- 2.5-20× faster than Faster-RCNN (0.17 sec/image on K40 using ResNet-101)
- Code is available at `https://github.com/daijifeng001/R-FCN`

## Region-based Object Detectors




**Methodologies of *region-based* detectors using ResNet-101**

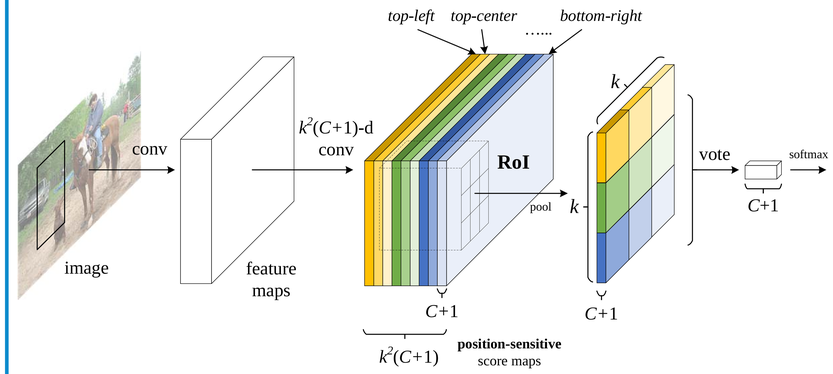
| | R-CNN | Faster R-CNN | R-FCN [ours] |
|---|---|---|---|
| depth of shared conv subnetwork | 0 | 91 | 101 |
| depth of RoI-wise subnetwork | 101 | 10 | **0** |

## Respecting Translation Variance for Detection

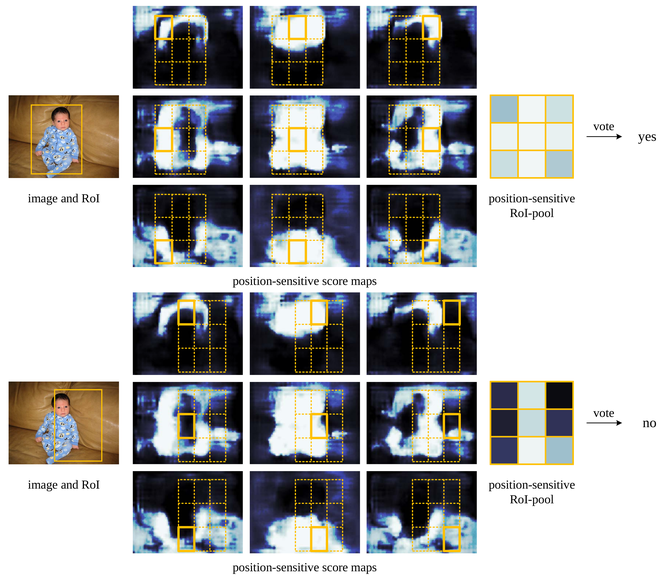
- Increasing translation invariance for image classification
  - Shift of an object inside an image should be indiscriminative
  - Leading deep (fully) convolutional architectures are translation-invariant
- Respecting translation variance for object detection
  - Responses should reflect how candidate boxes overlap with objects
  - A considerable deep per-ROI subnet in Faster-RCNN using ResNet-101


## R-FCN

- Key idea of **R-FCN** for object detection
  - Position-sensitive score maps ($k \times k$, *e.g.*, $k = 3$)
  - Position-sensitive RoI pooling




- Spatial information is encoded by position-sensitive score maps




- Key properties of **R-FCN**
  - Negligible per-RoI computational cost (in both training/inference)
  - The whole architecture is end-to-end trainable

## Experiments

**Comparisons among "almost" fully convolutional strategies using ResNet-101**

| method | RoI output size ($k \times k$) | mAP on VOC 07 (%) |
|---|---|---|
| naïve Faster R-CNN | $1 \times 1$ | 61.7 |
| | $7 \times 7$ | 68.9 |
| class-specific RPN | - | 67.6 |
| R-FCN (w/o position-sensitivity) | $1 \times 1$ | *fail* |
| R-FCN | $3 \times 3$ | 75.5 |
| | $7 \times 7$ | **76.6** |

**Comparisons between Faster R-CNN and R-FCN using ResNet-101**

| | depth of per-RoI subnetwork | training w/ OHEM? | train time (sec/img) | test time (sec/img) | mAP (%) on VOC07 |
|---|---|---|---|---|---|
| Faster R-CNN | 10 | | 1.2 | 0.42 | 76.4 |
| **R-FCN** | 0 | | 0.45 | 0.17 | 76.6 |
| Faster R-CNN | 10 | ✓ (300 RoIs) | 1.5 | 0.42 | 79.3 |
| **R-FCN** | 0 | ✓ (300 RoIs) | 0.45 | 0.17 | **79.5** |
| Faster R-CNN | 10 | ✓ (2000 RoIs) | 2.9 | 0.42 | *N/A* |
| **R-FCN** | 0 | ✓ (2000 RoIs) | 0.46 | 0.17 | 79.3 |

**On the Impact of Depth**

| | training data | test data | VGG-16 | ResNet-50 | ResNet-101 | ResNet-152 |
|---|---|---|---|---|---|---|
| R-FCN | 07+12 | 07 | 75.6 | 77.0 | 79.5 | 79.6 |
| R-FCN multi-sc train | 07+12 | 07 | 76.5 | 78.7 | 80.5 | 80.4 |

**Experiments on COCO using ResNet-101**

| | training data | test data | AP@0.5 | AP | test time (sec/img) |
|---|---|---|---|---|---|
| Faster R-CNN | train | val | 48.4 | 27.2 | 0.42 |
| **R-FCN** | train | val | **50.2** | **29.0** | 0.17 |