Carnegie Mellon University

Contextual Priming and Feedback for Faster R-CNN

Abhinav Shrivastava and Abhinav Gupta
Carnegie Mellon University

THE ROBOTICS INSTITUTE

Goal

Incorporate top-down information, feedback and contextual information in Faster R-CNN

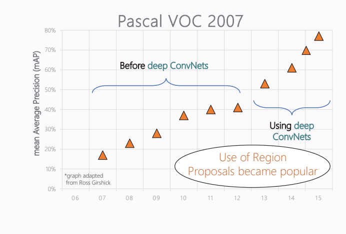
Contribution

Using Semantic segmentation for contextually priming region proposal & object detection modules, and providing iterative feedback to the entire network

Results

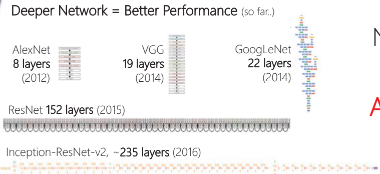
Improvement across all three tasks: object detection, semantic segmentation and region proposals.

Key Ingredients of a Region-based ConvNet Object Detector [most state-of-the-art in Object Detection systems]



1 Deeper, Feedforward ConvNets

Deeper Network = Better Performance (so far..)



Networks getting deeper, but remain feedforward

Are we on the right path?

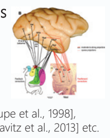
Human Visual Pathway

Strong evidence of Feedback connections

- Outnumber feedforward
- Feedback even to V1

Support that Object Detection uses:

- Top-down information
- Contextual Priming



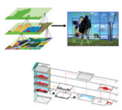
[Hupe et al., 1998], [Kravitz et al., 2013] etc.

2 Recognition using Regions

E.g., Selective Search, Randomized Prim's, CPMC, Bing, EdgeBoxes, Rigor, Geodesic, MCG, DeepMask, SharpMask, AttractioNet, etc.

Reduces Search Space
Allows use of richer features

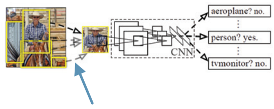
Focuses 'attention' in right areas
Reduces false positives



Generally, bottom-up, segmentation driven

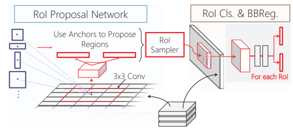
From Fast R-CNN to Faster R-CNN

Fast R-CNN



- Contextual Priming (attention) using bottom-up regions
- But no top-down feedback

Faster R-CNN



- Learns ConvNet to propose regions
- No Segmentation driven or bottom-up regions

Can we bridge this gap between empirical results and theory?

Incorporate top-down information, feedback and/or contextual reasoning in object detection

Contextual Priming and Feedback: Incorporating top-down information Faster R-CNN

Main Contributions:

Semantic segmentation as a top-down signal for:

- Contextual Priming
  For region proposals & object detection
- Iterative Feedback
  Top-down feedback to the entire network

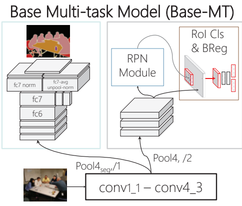
0 Faster R-CNN + Segmentation

Ideal Segmentation Network:

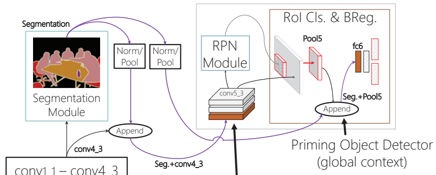
- Should be Fast
- Closely follow Faster R-CNN network (e.g., VGG16)
- No post-processing (e.g., CRFs)
  - Helps with end-to-end training

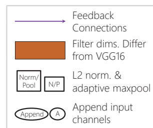
We use ParseNet [Liu 2015].

Base Multi-task Model (Base-MT)



1 Contextual Priming via Segmentation

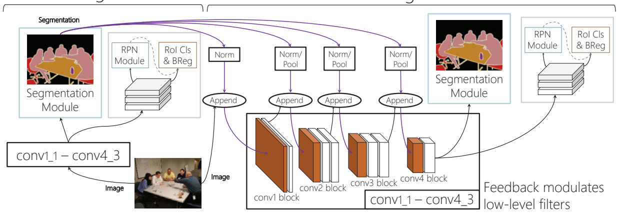




Priming Object Detector (global context)

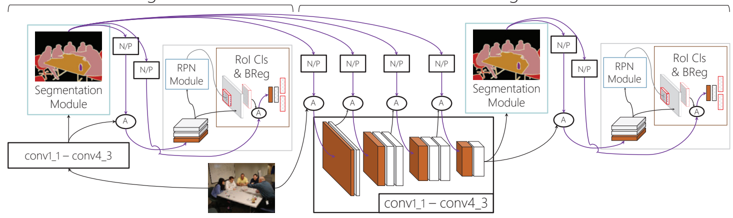
- Priming Region Proposal Generation
- Priming Object Detector (local context)

2 Iterative Feedback via Segmentation



Feedback modulates low-level filters

3 Joint Model: Contextual Priming and Feedback



Experiments to study the impact of Priming & Feedback

Ablation Analysis: Contextual Priming

	mAP	mIOU
Base-MT	75.6	65.8
Priming to conv5_1	77.0	65.8
Priming to conv5_1, each fc6	77.8	65.3

+ Priming to each RoI (which adds global context) helps detection.
- Gradients from each RoI overpower segmentation network.

Ablation Analysis: Iterative Feedback

	Stage-2 Init.	mAP	mIOU
Base-MT	-	75.6	65.8
Feedback to conv1_1	ImageNet	76.5	69.3
	Stage-1	76.3	69.3
Feedback to conv{1,2,3,4}_1	ImageNet	76.3	69.1
	Stage-1	77.3	69.5

- More feedback helps when initializing with Stage-1 network (cf., unrolled self-feedback)

Test set: VOC12 Segmentation val. set

Detection results

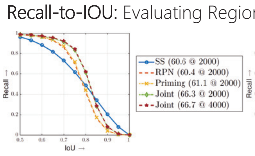
	S	P	F	mAP
Fast R-CNN				71.6
Faster R-CNN				75.3
Base-MT	✓			75.6
Ours [priming]	✓	✓		77.0
Ours [feedback]	✓		✓	77.3
Ours [joint]	✓	✓	✓	**77.8**

Segmentation results

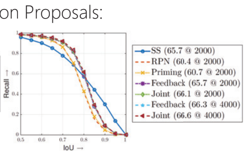
	S	P	F	mIOU
ParseNet	✓			68.2
ParseNet*	✓			66.0
Base-MT	✓			65.8
Ours [priming]	✓	✓		65.3
Ours [feedback]	✓		✓	69.5
Ours [joint]	✓	✓	✓	**69.6**

*with detection hyperparams (see paper)

Recall-to-IOU: Evaluating Region Proposals:



VOC 2007 detection test set



VOC 2012 segmentation val set

This top-down information improves **all** three tasks: object detection, semantic segmentation and region proposals.

Main Results on standard dataset splits

Detection results on VOC07 detection test set. All methods are trained on VOC07 trainval and VOC12 trainval

	S	mAP	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	persn	plant	sheep	sofa	train	tv
Fast R-CNN		70.0	77.0	78.1	69.3	59.4	38.3	81.6	78.6	86.7	42.8	78.8	68.9	84.7	82	76.6	69.9	31.8	70.1	74.8	80.4	70.4
Faster R-CNN		73.2	76.5	79.0	70.9	65.5	52.1	83.1	84.7	86.4	52	81.9	65.7	84.8	84.6	77.5	76.7	38.8	73.6	73.9	83.0	72.6
Base-MT	✓	74.7	78.4	79.3	75.9	63.2	56.8	85.9	85.4	88.4	54.9	83.9	68.6	84.6	85.6	78.5	78.1	41.3	74.6	74.8	84.0	72.4
Ours [joint]	✓	**76.4**	79.3	80.5	76.8	72.0	58.2	85.1	86.5	89.3	60.6	82.2	69.2	87.0	87.2	81.6	78.2	44.6	77.9	76.7	82.4	71.9
						+8.8					+5.7							+3.3	+3.3			

Detection results on VOC12 detection test set. All methods are trained on VOC07 trainval+test and VOC12 trainval

	S	mAP	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	persn	plant	sheep	sofa	train	tv
Fast R-CNN		68.4	82.3	78.4	70.8	52.3	38.7	77.8	71.6	89.3	44.2	73.0	55.0	87.5	80.5	80.8	72	35.1	68.3	65.7	80.4	64.2
Faster R-CNN		70.4	84.9	79.8	74.3	53.9	49.8	77.5	75.9	88.5	45.6	77.1	55.3	86.9	81.7	80.9	79.6	40.1	72.6	60.9	81.2	61.5
Base-MT	✓	71.1	84.2	80.9	73.1	55.1	50.6	78.2	75.6	89.0	48.6	76.7	54.8	87.6	82.5	83.0	80.0	41.7	74.2	60.7	81.4	63.1
Ours [joint]	✓	**72.6**	84.0	81.2	75.9	60.4	51.8	81.2	77.4	90.9	50.2	77.6	58.7	88.4	83.6	82.0	80.4	41.5	75.0	64.2	82.9	65.1
						+5.3		+3.0					+3.9							+3.5		

Segmentation results on VOC12 segmentation test set. All methods are trained on 07 trainval+test and 12 trainval

	S	mIOU	bg	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	persn	plant	sheep	sofa	train	tv
Base-MT	✓	66.4	91.3	82.0	37.7	77.6	58.8	58.8	84.0	75.6	83.1	25.1	70.9	57.8	74.0	74.6	76.4	75.0	48.8	73.7	45.6	72.3	52.0
Ours [joint]	✓	**71.4**	93.0	89.3	41.4	84.1	63.8	65.2	88.1	80.9	88.6	28.4	75.4	60.6	80.3	80.9	83.1	79.7	55.4	77.9	48.2	75.8	58.8
				+7.3	+3.7	+6.5	+5.0	+6.4	+4.1	+5.3	+5.5	+3.3	+4.5		+6.3	+6.3	+6.7	+4.7	+6.6	+4.2		+3.5	+6.8

Detection results on COCO 2015 test-dev set. All methods are trained COCO 2014 trainval35k

	S	P	F	AP, IoU: 0.5:0.95	0.5	0.75	AP, Area: Small	Med.	Large	AR, #Dets: car	10	100	AR, Area: Small	Med.	Large
Faster R-CNN				24.5	46.0	23.7	8.2	26.4	36.9	24.0	34.8	35.5	13.4	39.2	54.3
Base-MT	✓			25.0	47.0	24.2	8.1	27.1	38.1	24.3	35.1	35.8	13.2	39.8	55.0
Ours [priming]	✓	✓		25.8	48.2	25.3	8.3	27.8	38.6	24.5	35.7	36.5	13.6	40.6	54.7
Ours [joint]	✓	✓	✓	**27.5**	**49.2**	**27.8**	**8.9**	**29.5**	**41.5**	**25.5**	**37.4**	**38.3**	**14.6**	**42.5**	**57.4**
COCO Detection 2016 Challenge Entry: 1. **Training** with more smaller proposals 2. **Testing**: a) multi-scale, b) average across LR-flip, c) add AttractioNet proposals, d) box refinement and weighted voting				32.4	52.9	34.3	15.0	35.4	45.7	29.5	46.3	47.2	25.5	52.1	65.3

Ranked 4th in 2016 COCO detection challenge with a single VGG16 model!