

# Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering

Peter Anderson<sup>1†</sup>, Xiaodong He<sup>2‡</sup>, Chris Buehler<sup>2</sup>, Damien Teney<sup>3</sup>, Mark Johnson<sup>4</sup>, Stephen Gould<sup>1</sup>, Lei Zhang<sup>2</sup>

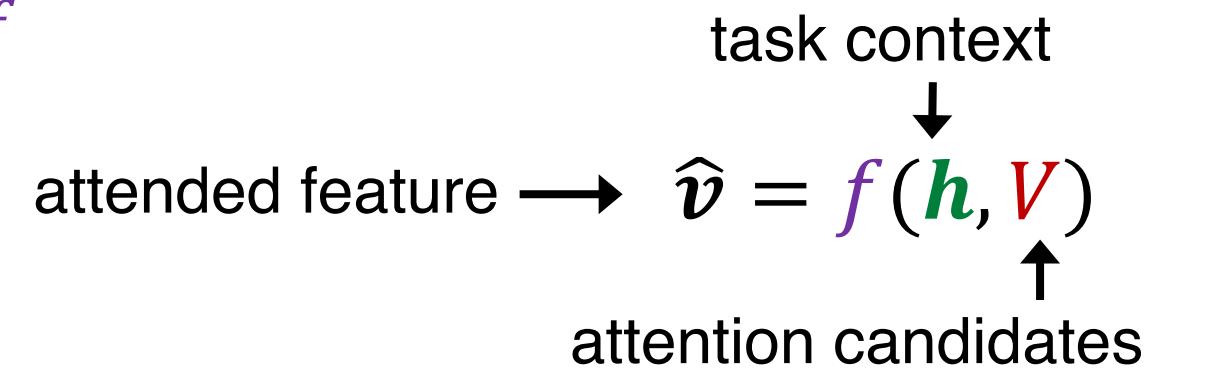
<sup>1</sup>Australian National University, <sup>2</sup>Microsoft Research, <sup>3</sup>University of Adelaide, <sup>4</sup>Macquarie University, <sup>†</sup>Transitioning to Georgia Tech, <sup>‡</sup>Now at JD AI Research



## 1. Visual attention

Visual attention mechanisms learn to focus on image regions that are relevant to the task, requiring:

1. Learned attention function (network),  $f$
2. A set of attention candidates,  $V$
3. Task context representation,  $h$



## 4. Pre-training Faster R-CNN

We pre-train Faster R-CNN on Visual Genome<sup>6</sup> data, using:

- 1600 object classes
- 400 attribute classes

To select  $k$  attention candidates, a detection confidence threshold is used

Example training data:



<sup>6</sup>Krishna et al. arXiv 1602.07332, 2016

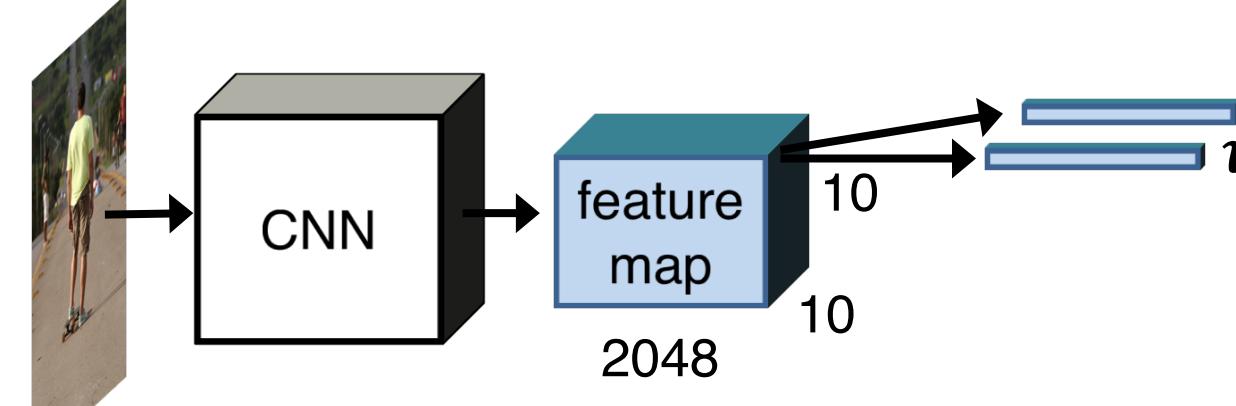


Code, models and pre-trained features available:  
<http://www.panderson.me/up-down-attention>

Refer also to our related work: *Tips and Tricks for Visual Question Answering: Learnings From the 2017 Challenge*,  
Poster J21, Wednesday June 20, 10:10-12:30 Poster Session P2-1

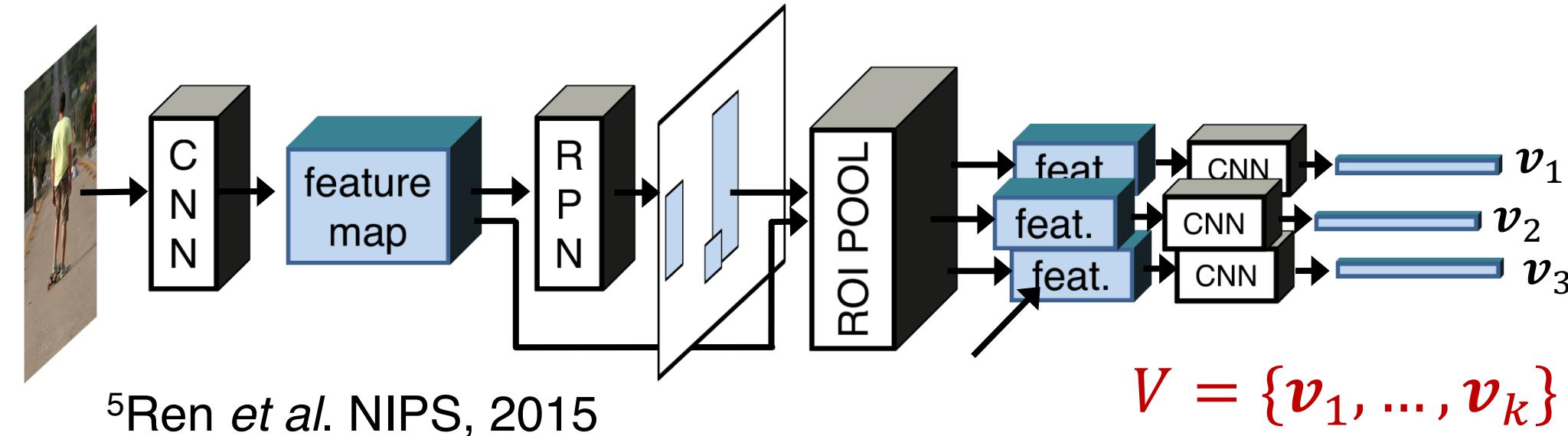
## 2. Generation of attention candidates, $V$

Typical: spatial output of a CNN

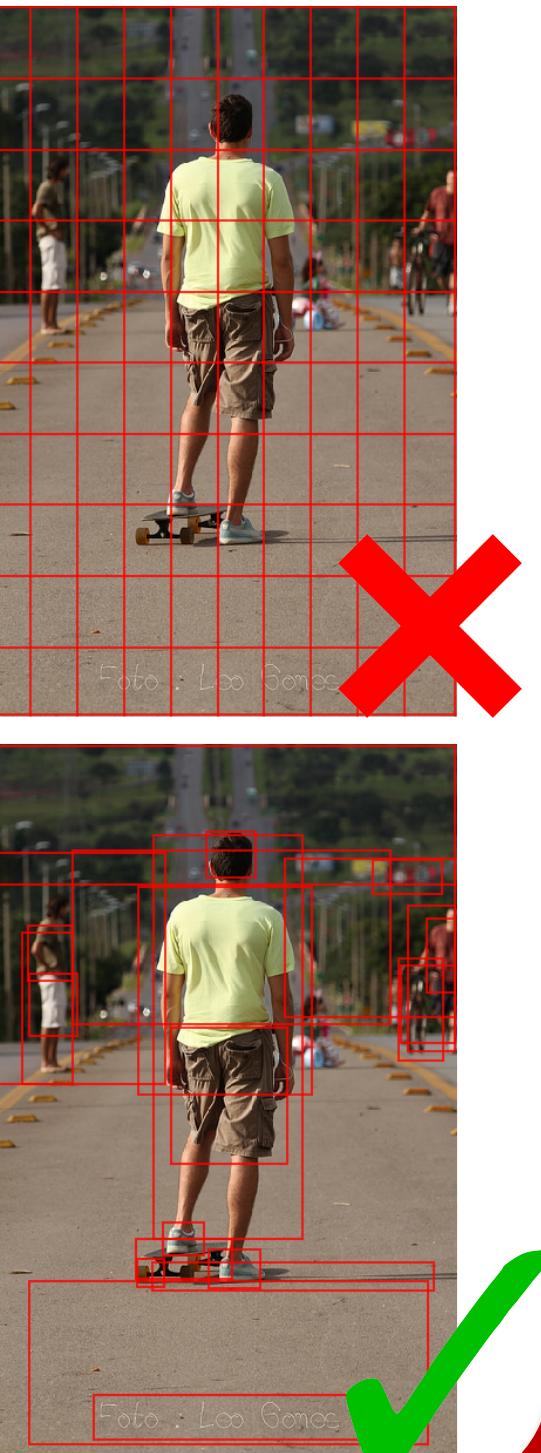


$$V = \{v_1, \dots, v_{100}\}$$

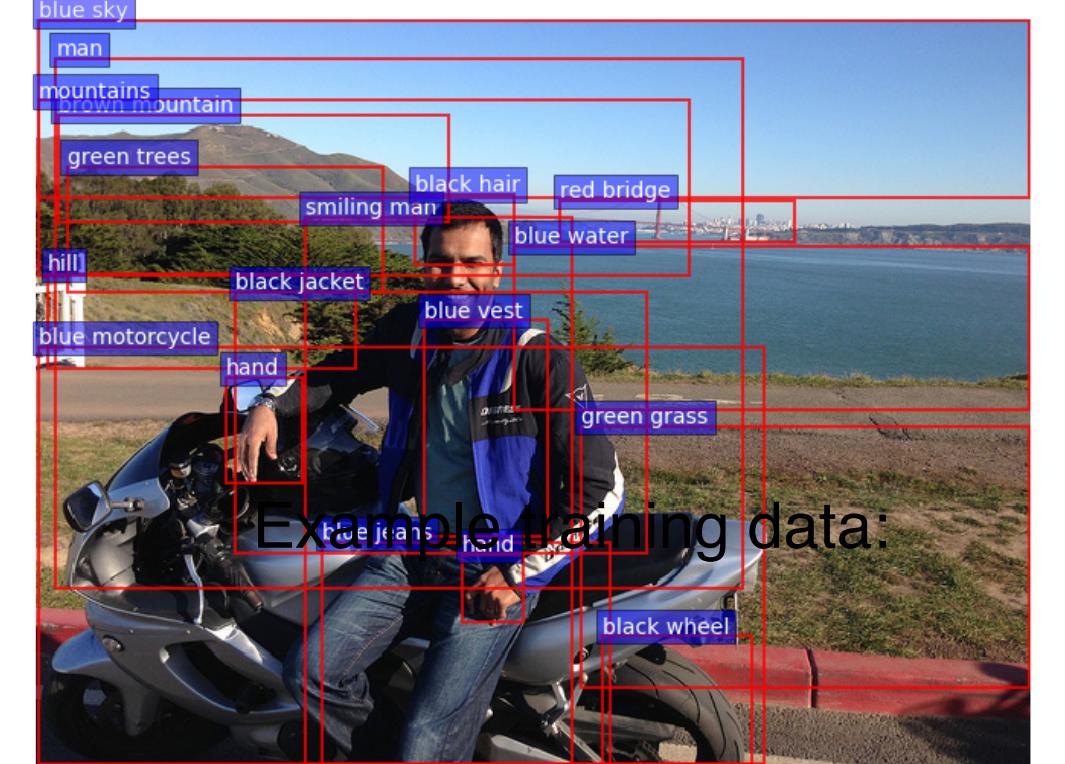
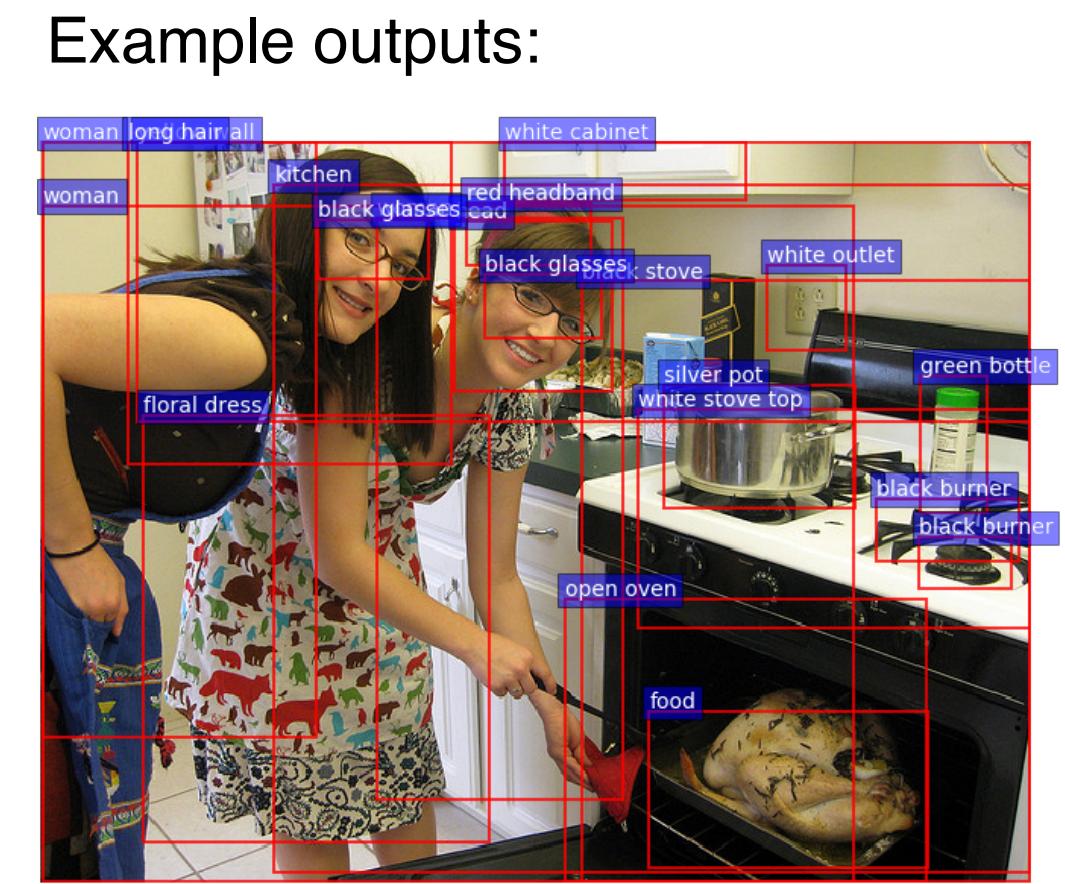
Ours: bottom-up attention (using Faster R-CNN<sup>5</sup>)



<sup>5</sup>Ren et al. NIPS, 2015



Example training data:



## 5. Quantitative results

- 1<sup>st</sup> 2017 VQA Challenge (June 2017)
- 1<sup>st</sup> COCO Captions leaderboard (July 2017)
- Up-Down approach now incorporated into many other models (including many 2018 VQA Challenge entries)

VQA v2 val set (single-model):

	Yes/No	Number	Other	Overall
ResNet (1×1)	76.0	36.5	46.8	56.3
ResNet (14×14)	76.6	36.2	49.5	57.9
ResNet (7×7)	77.6	37.7	51.5	59.4
Up-Down (Ours)	80.3	42.8	55.8	63.2

## 6. Qualitative results

### Image captioning:

ResNet: A man sitting on a **toilet** in a bathroom.

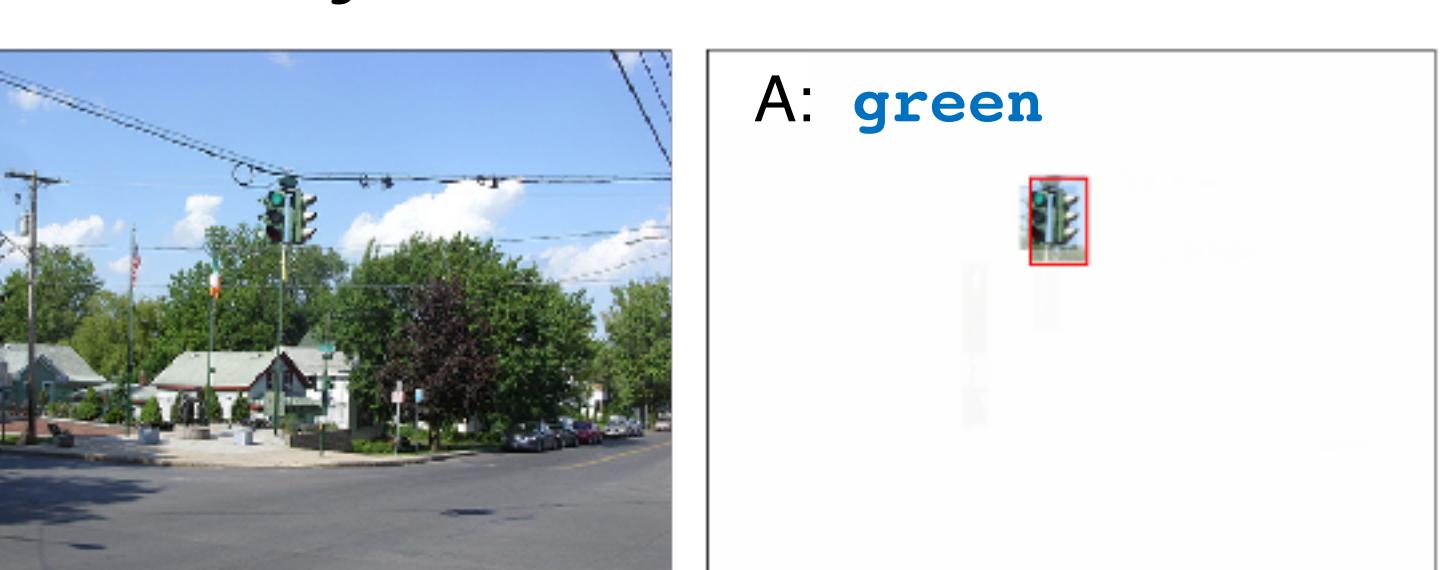


Up-Down: A man sitting on a **couch** in a bathroom.



### VQA:

Q: What color is illuminated on the traffic light?



COCO Captions "Karpathy" test set (single-model):

	BLEU-4	METEOR	CIDEr	SPICE
ResNet (10×10)	34.0	26.5	111.1	20.2
Up-Down (Ours)	36.3	27.7	120.1	21.4