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

Peter Anderson[1†], Xiaodong He[2‡], Chris Buehler[2], Damien Teney[3], Mark Johnson[4], Stephen Gould[1], Lei Zhang[2]

[1]Australian National University, [2]Microsoft Research, [3]University of Adelaide, [4]Macquarie University, [†]Moving to Georgia Tech, [‡]Now at JD AI Research

2018 CVPR
SALT LAKE CITY • JUNE 18-22

Visual attention

- Vision and language tasks often require fine-grained visual processing, e.g. VQA:

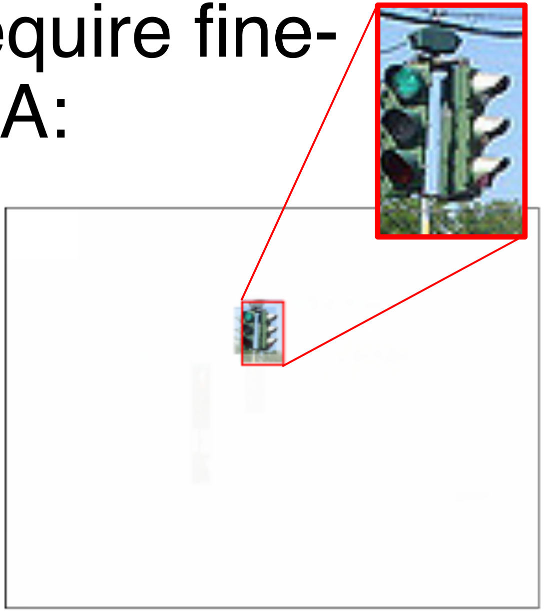
```
Q: What color
is illuminated
on the traffic
light?
```

Visual attention

- Vision and language tasks often require fine-grained visual processing, e.g. VQA:

```
Q: What color
is illuminated
on the traffic
light?

A: green
```



Visual attention

- Visual attention mechanisms learn to focus on image regions that are relevant to the task

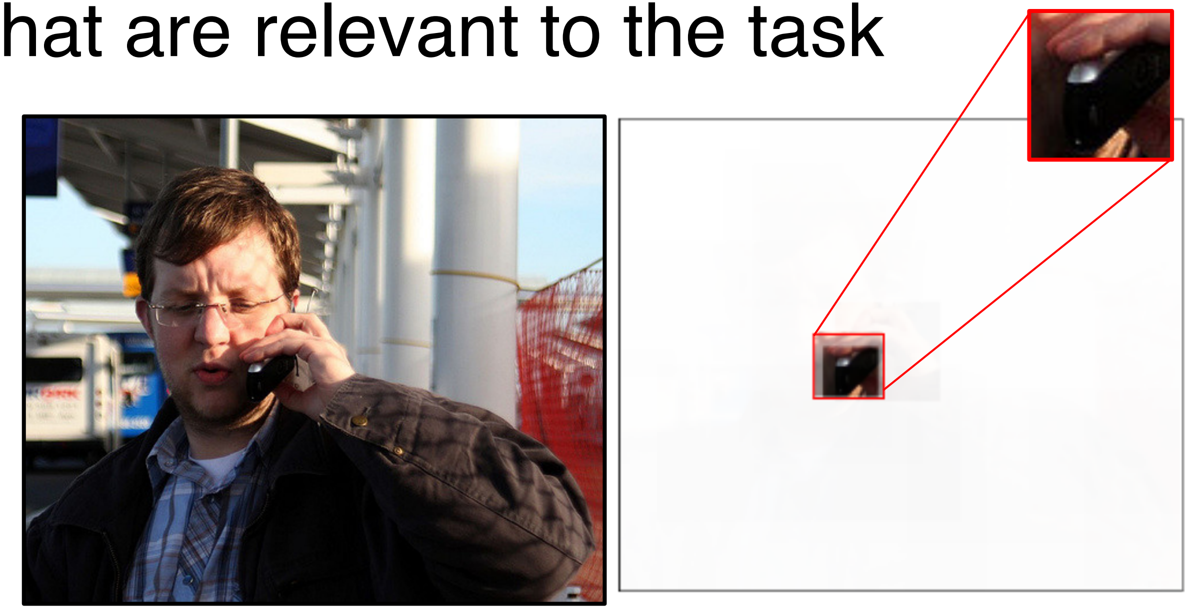
Q: What is the man holding?

Visual attention

- Visual attention mechanisms learn to focus on image regions that are relevant to the task

`Q: What is the man holding?`

`A:` **`phone`**



Components of visual attention

- Visual attention mechanisms learn to focus on image regions that are relevant to the task

attended feature $\longrightarrow$ $\hat{\boldsymbol{v}} = f(\boldsymbol{h}, V)$

Components of visual attention

- Visual attention mechanisms learn to focus on image regions that are relevant to the task

attended feature $\longrightarrow$ $\hat{\boldsymbol{v}} = f(\boldsymbol{h}, V)$

1. set of attention candidates, V

Components of visual attention

- Visual attention mechanisms learn to focus on image regions that are relevant to the task

2. task context representation

attended feature $\longrightarrow$ $\hat{\boldsymbol{v}} = f(\boldsymbol{h}, V)$

1. set of attention candidates, V

Components of visual attention

- Visual attention mechanisms learn to focus on image regions that are relevant to the task

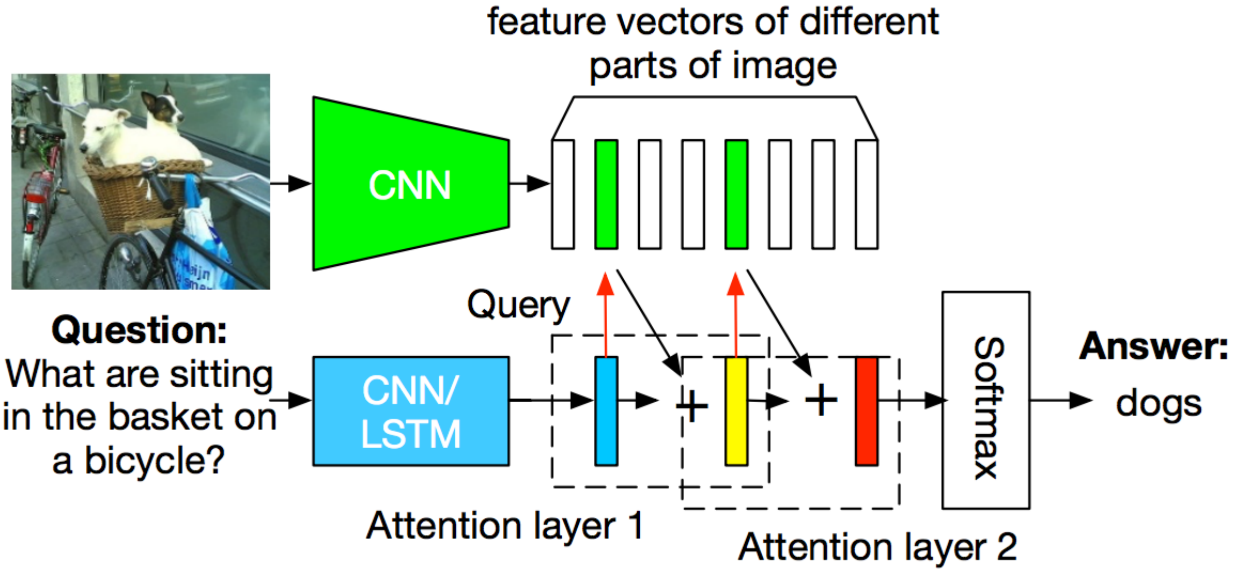
2. task context representation

attended feature $\longrightarrow$ $\hat{\boldsymbol{v}} = f(\boldsymbol{h}, V)$

3. learned attention function

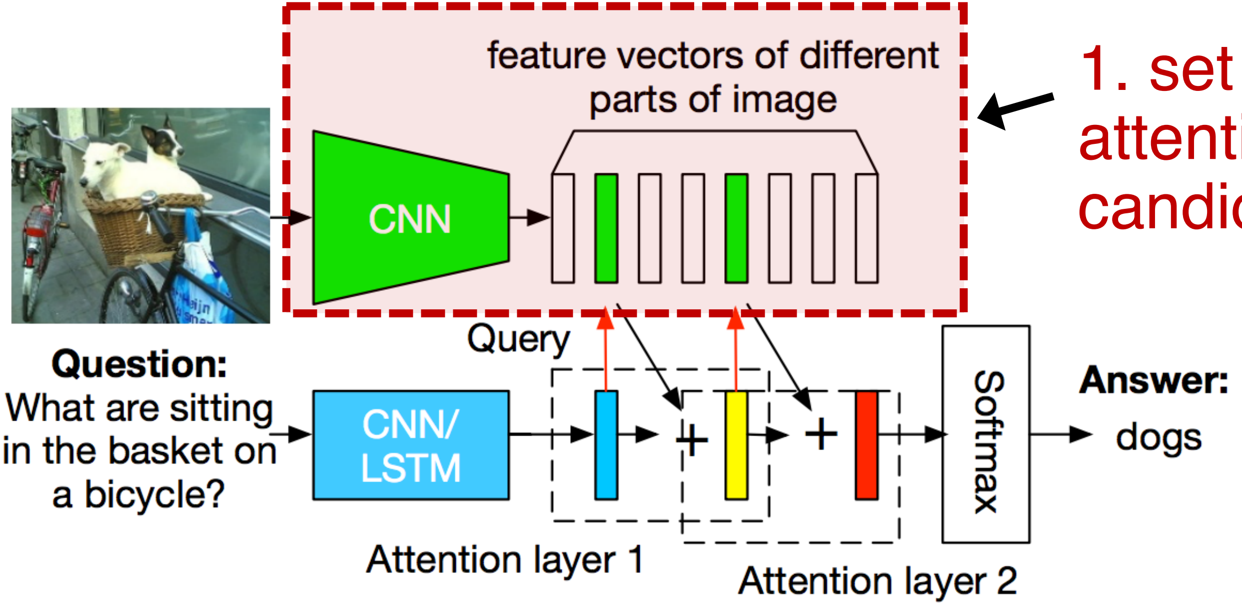
1. set of attention candidates, V

Example: Stacked attention networks[1]



[1]Yang *et al.* CVPR 2016

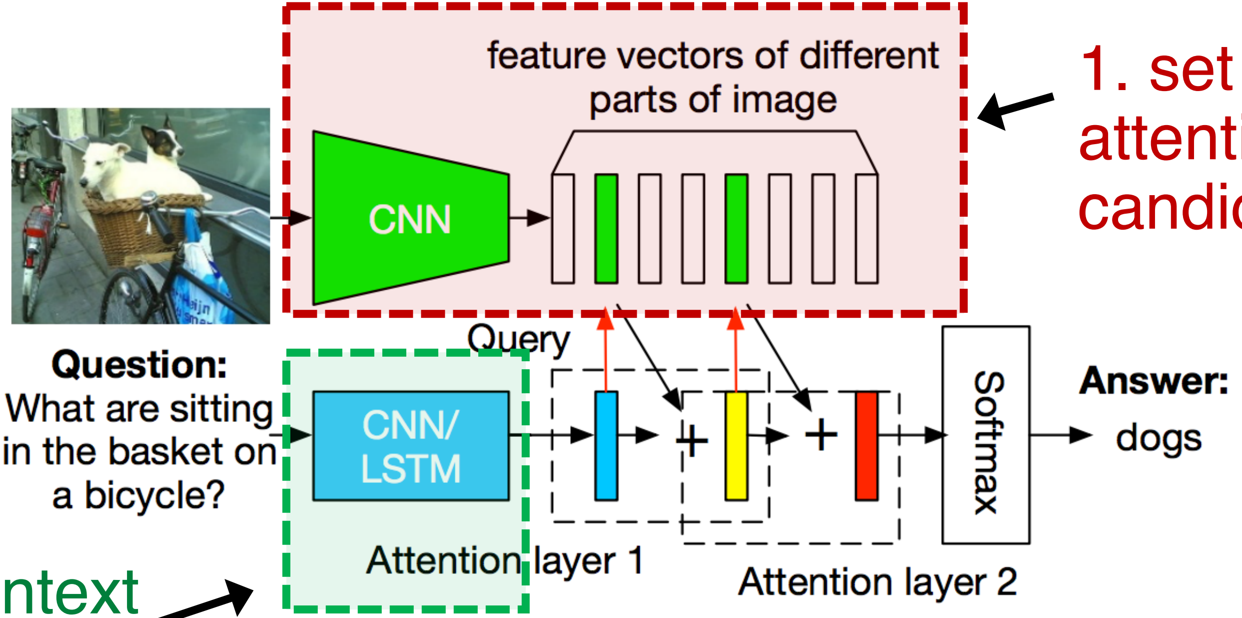
Example: Stacked attention networks[1]



1. set of attention candidates, V

[1]Yang *et al.* CVPR 2016

Example: Stacked attention networks[1]

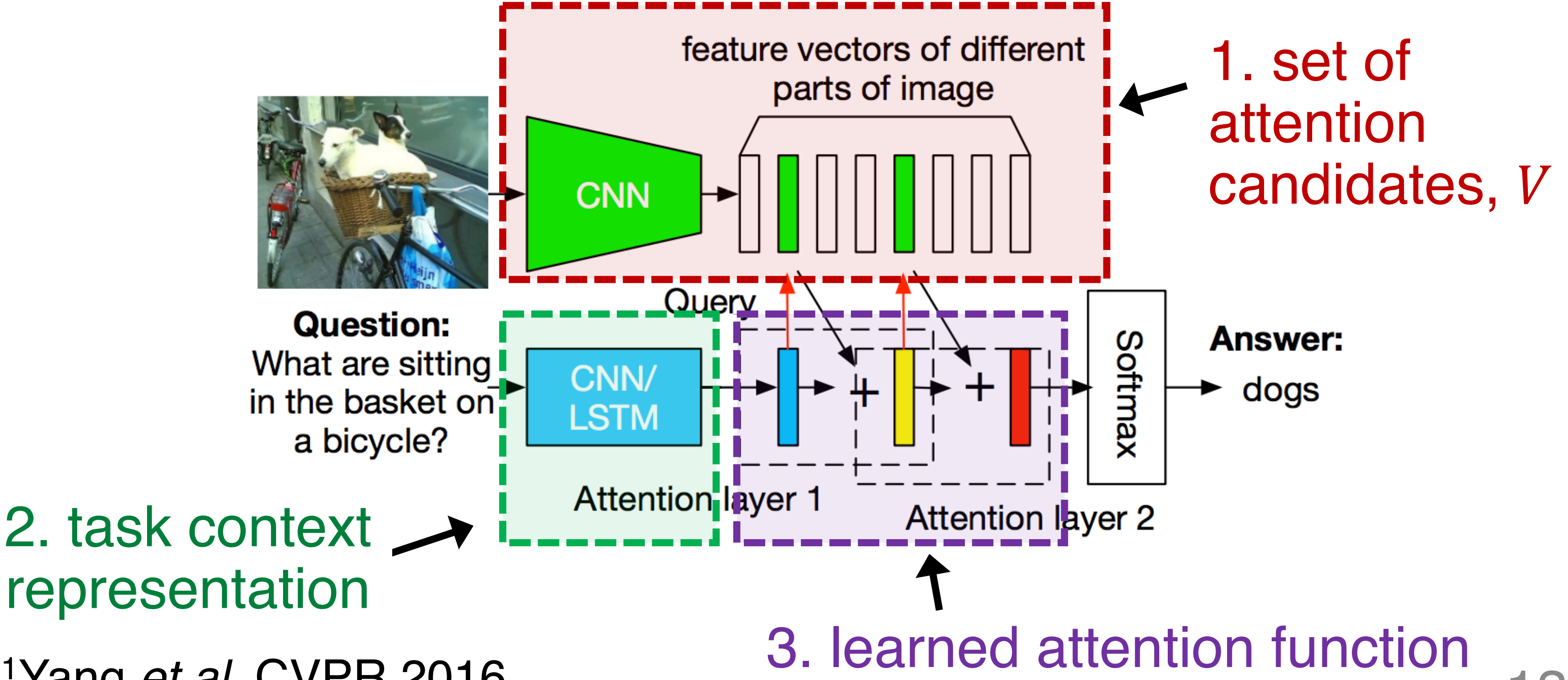


1. set of attention candidates, V

2. task context representation

[1]Yang *et al.* CVPR 2016

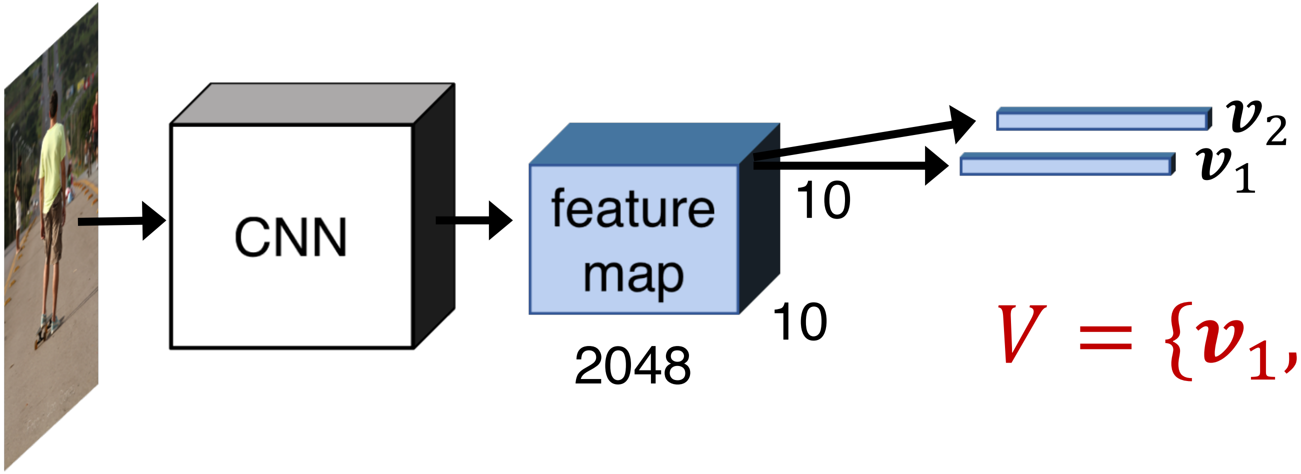
Example: Stacked attention networks[1]



[1]Yang *et al.* CVPR 2016

Attention candidates, V

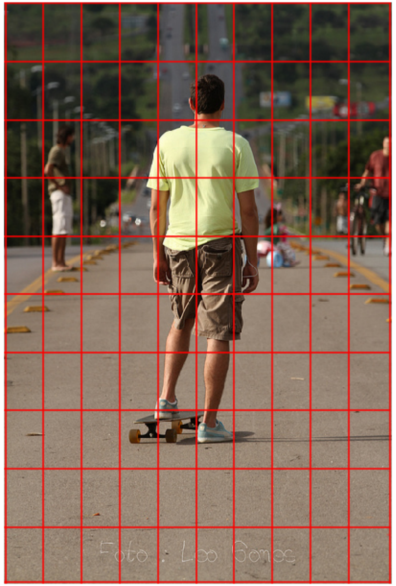
Standard approach: use the spatial output of a CNN to extract vectors for each position in a grid



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

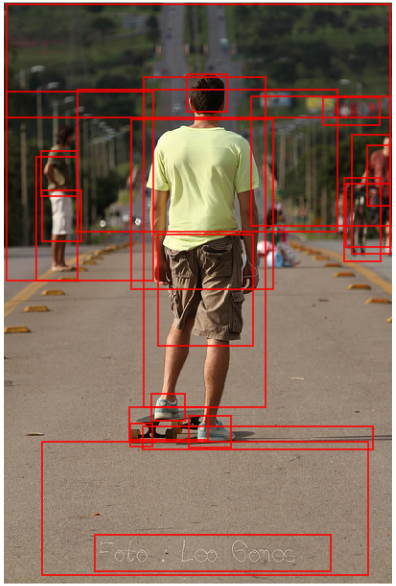
Attention candidates, V

10



10

Standard approach:
spatial output of a CNN



k regions

Our approach:
object-based attention

Objects are a natural basis for attention

- Human visual attention can select discrete objects, not just spatial regions[1]

[1]Egly et al. 1994, Scholl 2001

Objects are a natural basis for attention

- Human visual attention can select discrete objects, not just spatial regions[1]
- Image captioning and VQA are concerned with objects

`A young man on a skateboard looking down street with people watching.`

Q: `Is the boy in the yellow shirt wearing head protective gear?` **A:** `No`

[1]Egly et al. 1994, Scholl 2001

Objects are a natural basis for attention

- Human visual attention can select discrete objects, not just spatial regions[1]
- Image captioning and VQA are concerned with objects

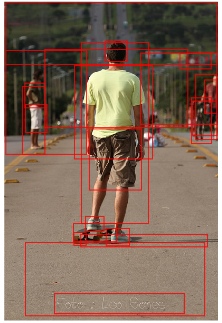
A young man on a skateboard looking down street with people watching.

Q: Is the boy in the yellow shirt wearing head protective gear? **A:** No

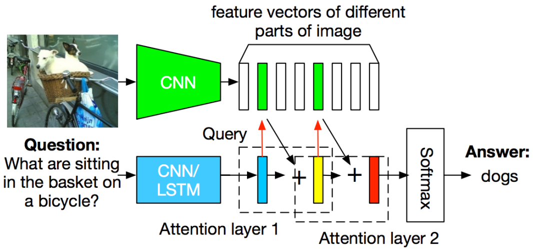
[1] Egly et al. 1994, Scholl 2001

Bottom-up and top-down attention

V



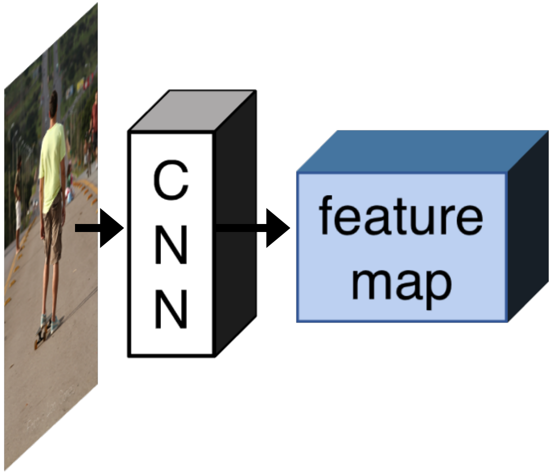
Bottom-up process: Extract all objects and other salient regions from the image (independent of the question / partially-completed caption)



Top-down process: Given task context, weight the attention candidates (i.e., use existing VQA / captioning models)

Attention candidates, V

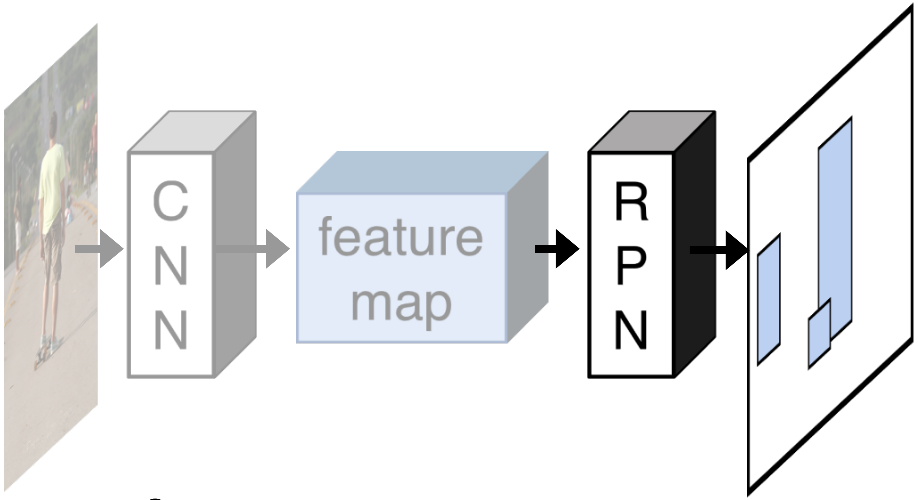
Our approach: bottom-up attention (using Faster R-CNN[2])



[2]Ren *et al*. NIPS, 2015

Attention candidates, V

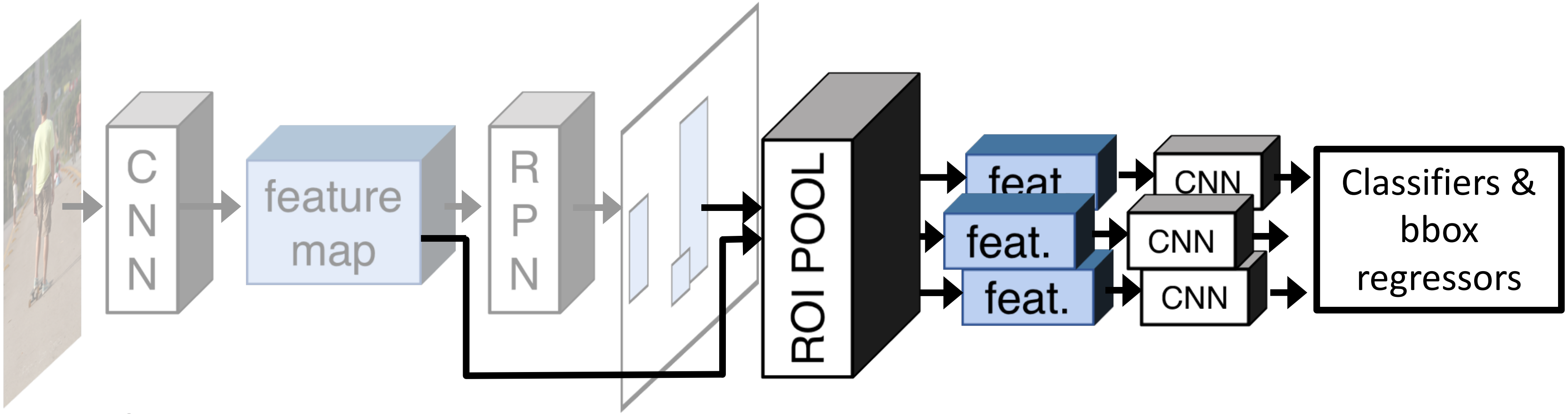
Our approach: bottom-up attention (using Faster R-CNN[2])



[2]Ren *et al*. NIPS, 2015

Attention candidates, V

Our approach: bottom-up attention (using Faster R-CNN[2])

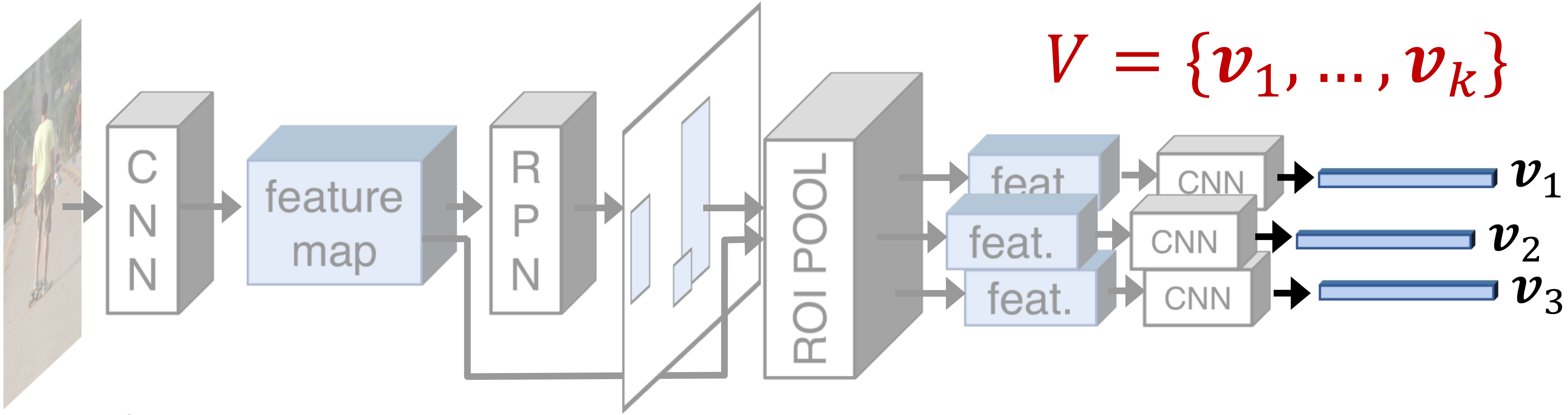


[2]Ren *et al*. NIPS, 2015

Attention candidates, V

Our approach: bottom-up attention (using Faster R-CNN[2])

- Each salient object / image region is detected and represented by its mean-pooled feature vector

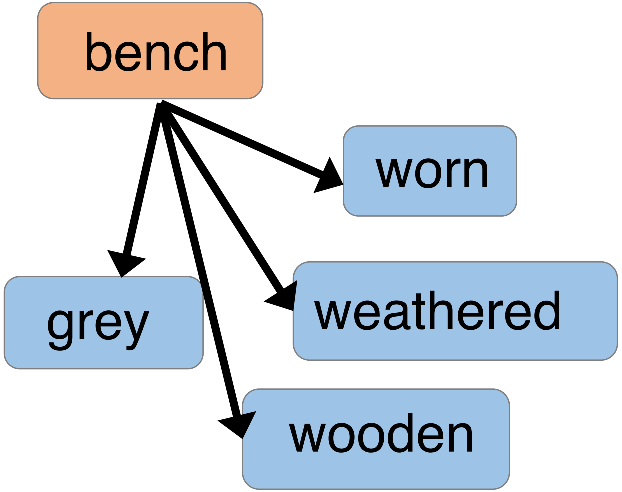


[2]Ren *et al*. NIPS, 2015

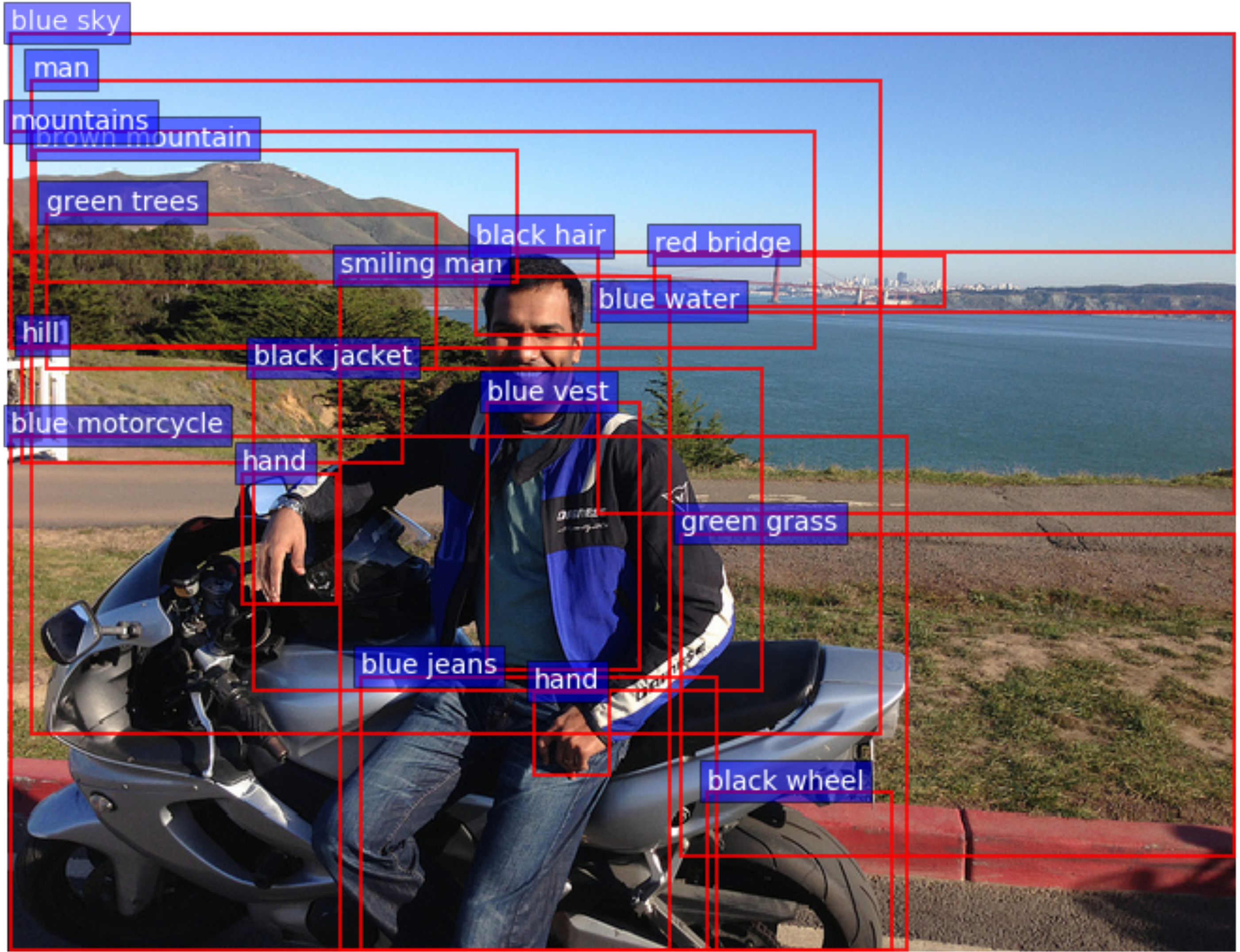
Faster R-CNN pre-training

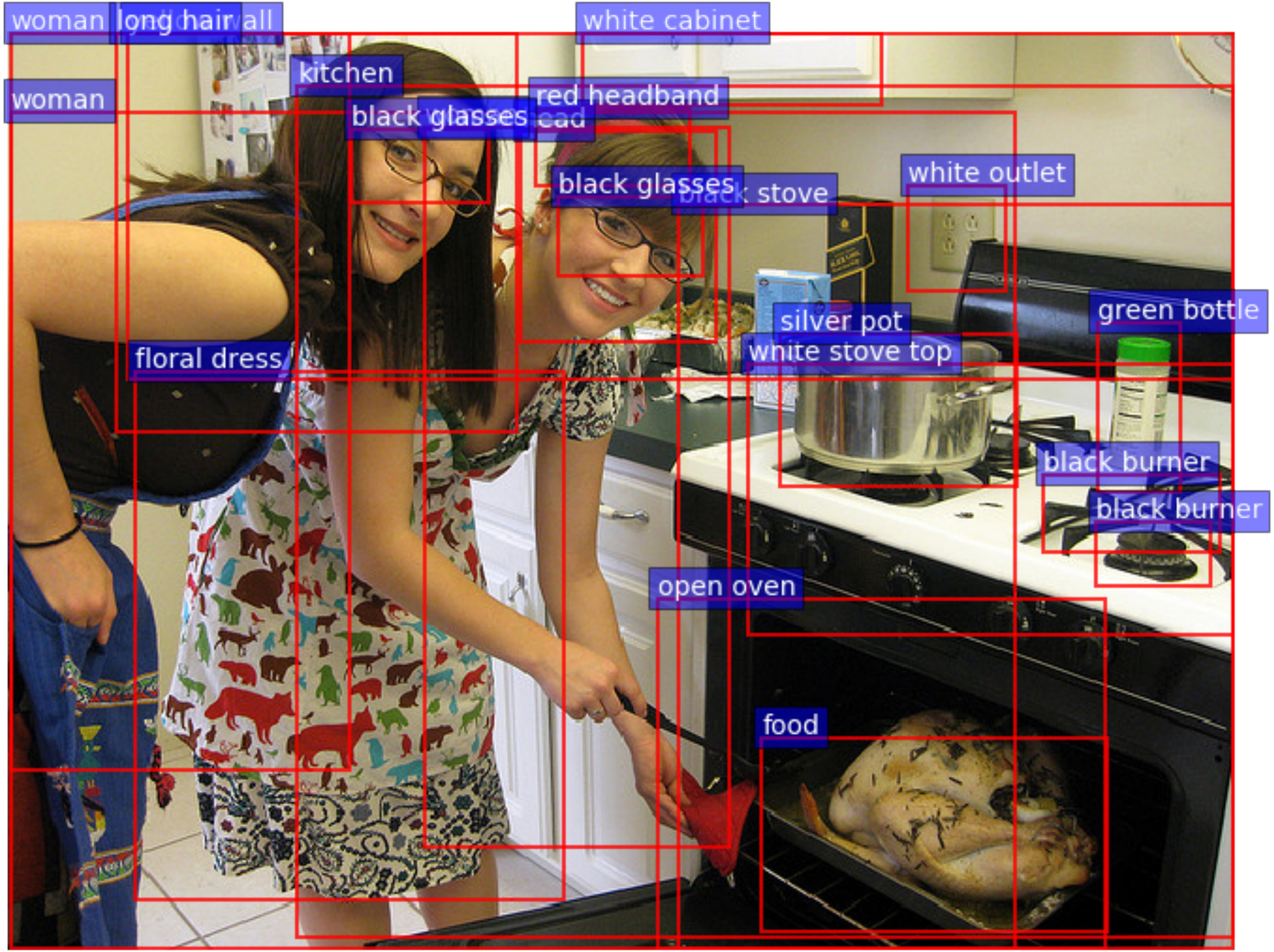
Using Visual Genome[3] with:

- 1600 filtered object classes
- 400 filtered attribute classes



[3]Krishna *et al.* arXiv 1602.07332, 2016

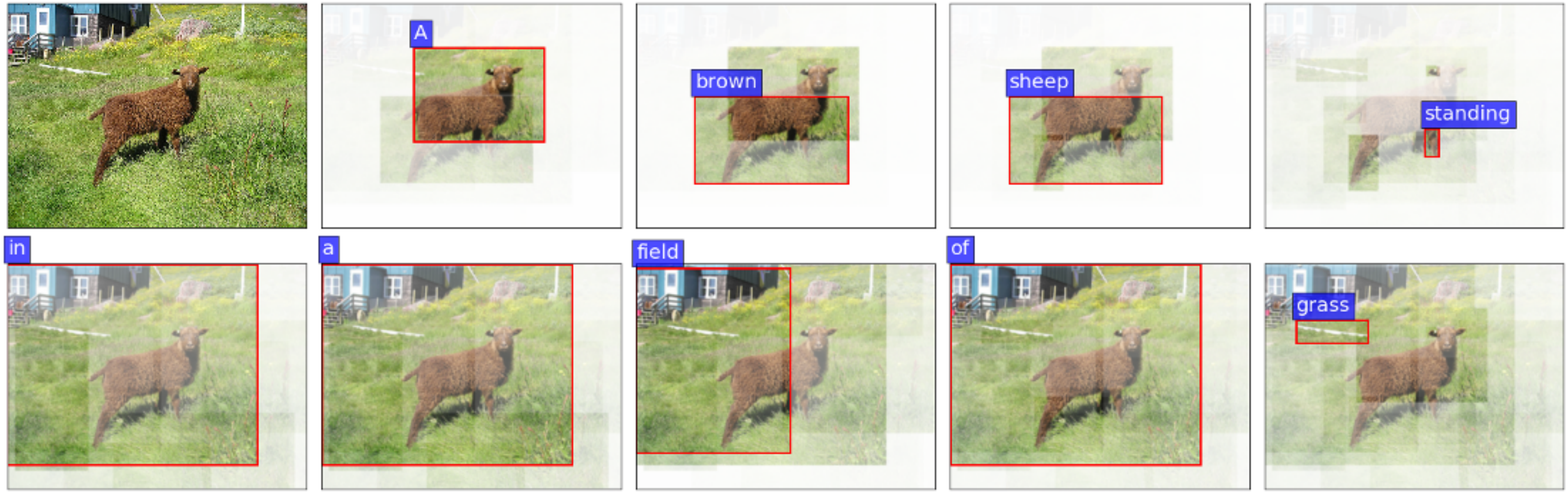

blue sky
man
mountains
brown mountain
green trees
black hair
red bridge
smiling man
blue water
hill
black jacket
blue vest
blue motorcycle
hand
green grass
blue jeans
hand
black wheel


woman
long hair
white cabinet
kitchen
woman
red headband
black glasses
black glasses
white outlet
green bottle
silver pot
white stove top
floral dress
black burner
black burner
open oven
food

ResNet (10×10): A man sitting on a ~~toilet~~ in a bathroom.

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

Up-Down (Ours): `A brown sheep standing in a field of grass.`



COCO Captions results

1st COCO Captions leaderboard (July 2017)

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**

COCO Captions results

1st COCO Captions leaderboard (July 2017)

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**

+8%

COCO Captions results

1st COCO Captions leaderboard (July 2017)

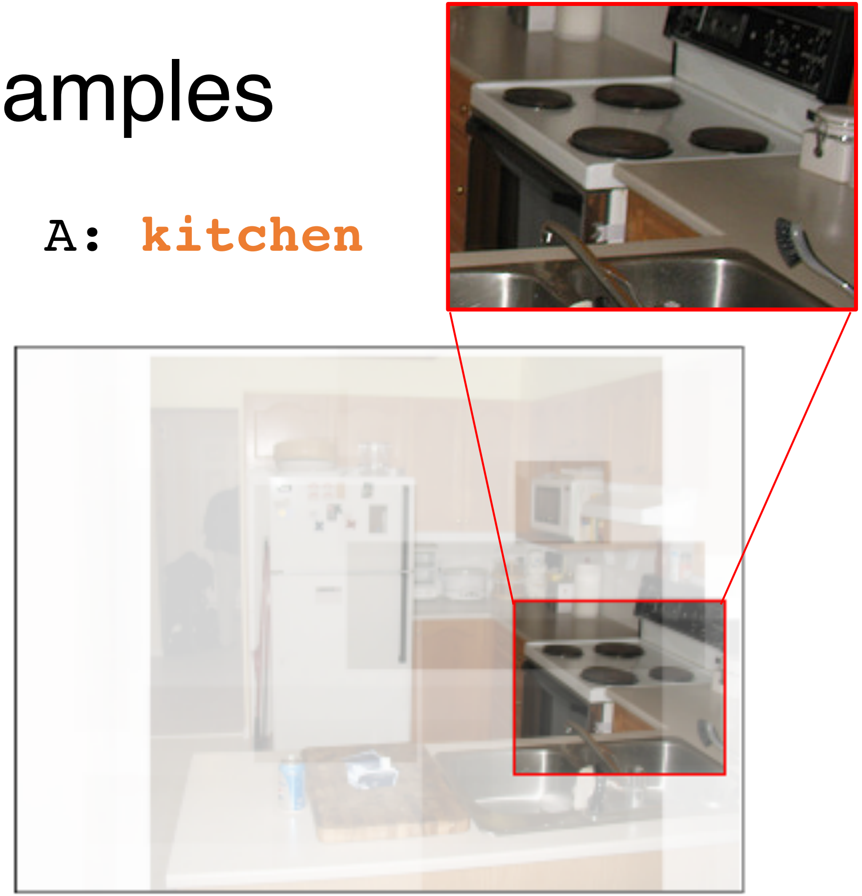
COCO Captions "Karpathy" test set (single-model):

	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**
			+8%	+6%

VQA examples

Q: What room are they in?

A: kitchen



VQA examples - counting

Q: How many oranges are on pedestals?

A: ~~two~~

VQA examples - reading

Q: What is the name of the realty company?

A: ~~none~~

VQA results

- **1st 2017 VQA Challenge** (June 2017)
- Top three 2018 Challenge entries used our approach

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**

VQA results

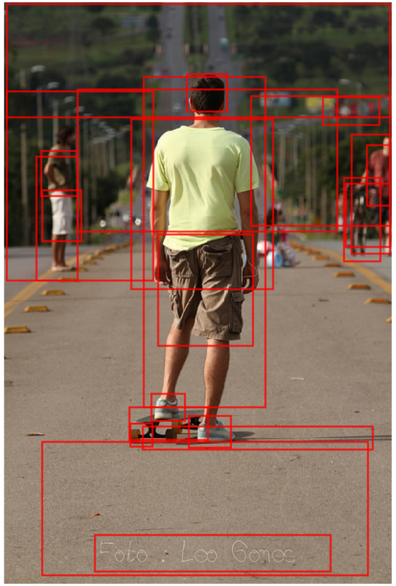
- **1st 2017 VQA Challenge** (June 2017)
- Top three 2018 Challenge entries used our approach

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**

+4%

Benefits of 'Up-Down' attention



- Natural approach
- Unifies vision & language tasks with object detection models
- Transfer learning by pre-training on object detection datasets
- Complementary to other models (just swap attention candidates)
- Can be fine-tuned
- More interpretable attention weights
- Significant improvements on multiple tasks

Poster C12

Code, models and drop-in pre-trained COCO image features available at:

http://www.panderson.me/up-down-attention

Related Work: ‘Tips and Tricks for Visual Question Answering: Learnings From the 2017 Challenge’, also at CVPR 2018, Poster J21