Natural Language Object Retrieval

Ronghang Hu[1] Huazhe Xu[2] Marcus Rohrbach[1,3]
Jiashi Feng[4] Kate Saenko[5] Trevor Darrell[1]

[1]UC Berkeley [2]Tsinghua University [3]ICSI, Berkeley
[4]NUS [5]Boston University

Natural Language Object Retrieval

query='*window upper right*'

query='*bottom left window*'

Example: ground-truth in ***yellow***, prediction in ***green***

Given an input **scene** and a natural language **query**, we want to *localize the target object* in the scene.

Related Tasks

image captioning

a room filled with a wooden table and a green chair.

text-based image retrieval

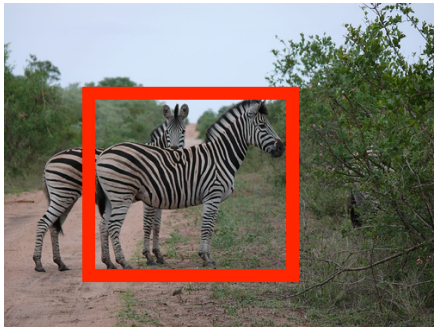




query="a man wearing sunglasses sits on a bench"

this task: natural language object retrieval

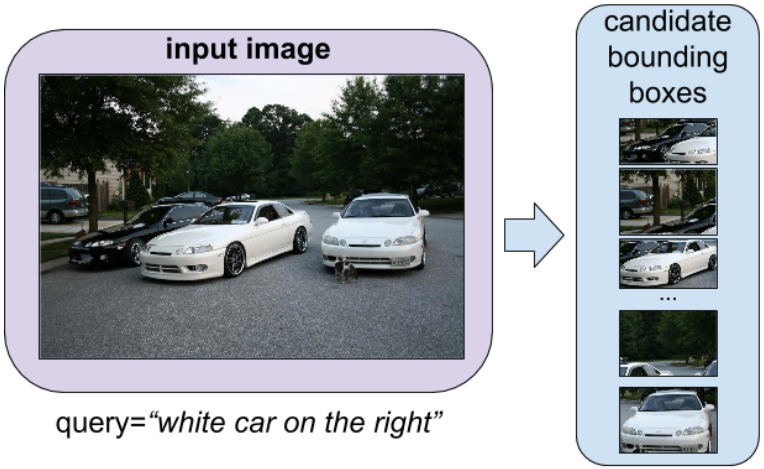
query="zebra on the right"



Our Method

Basic pipeline:

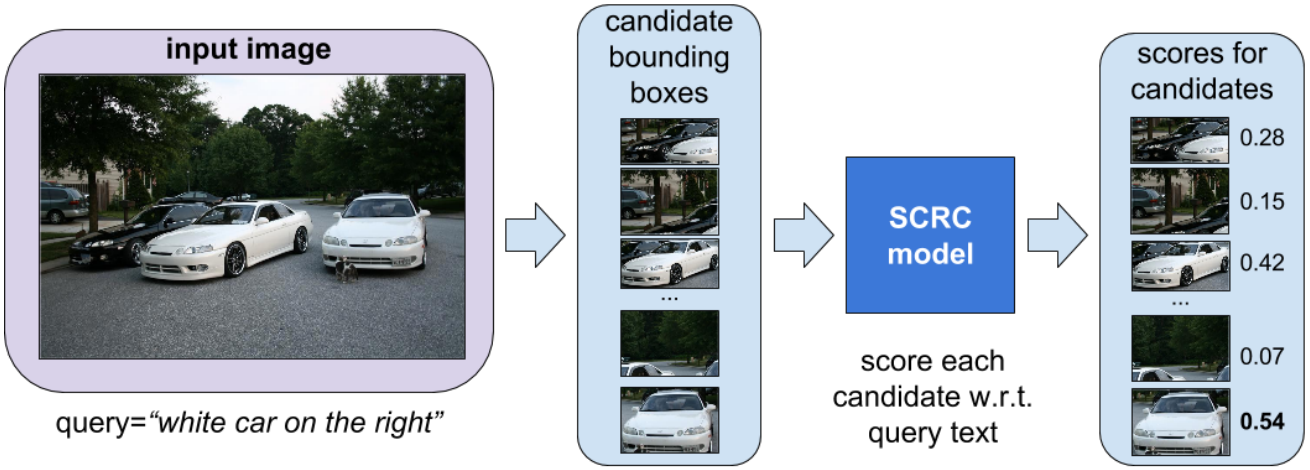
1. Extract candidate bounding box locations



Our Method

Basic pipeline:

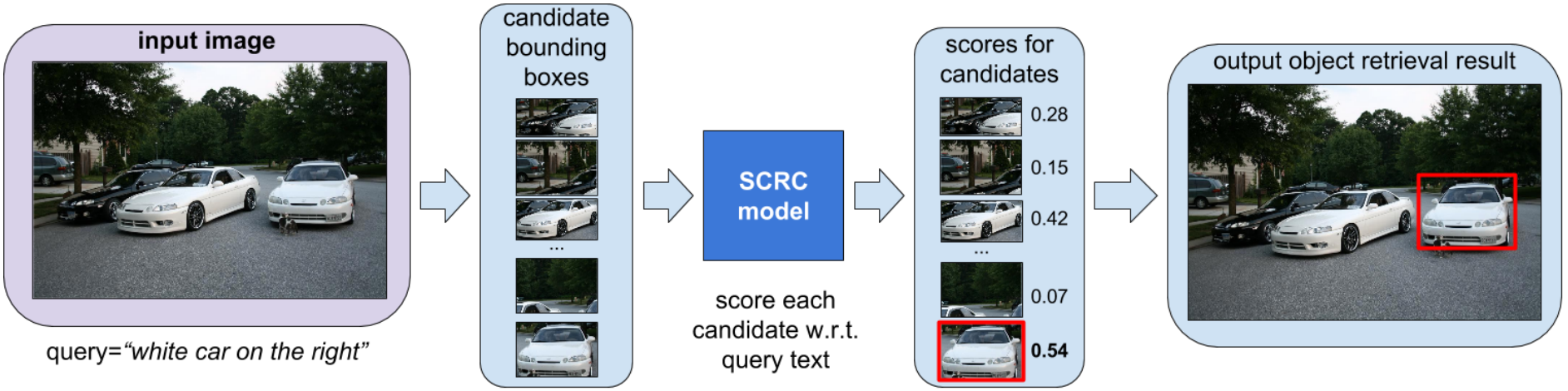
1. Extract candidate bounding box locations
2. Score each candidate w.r.t query text



Our Method

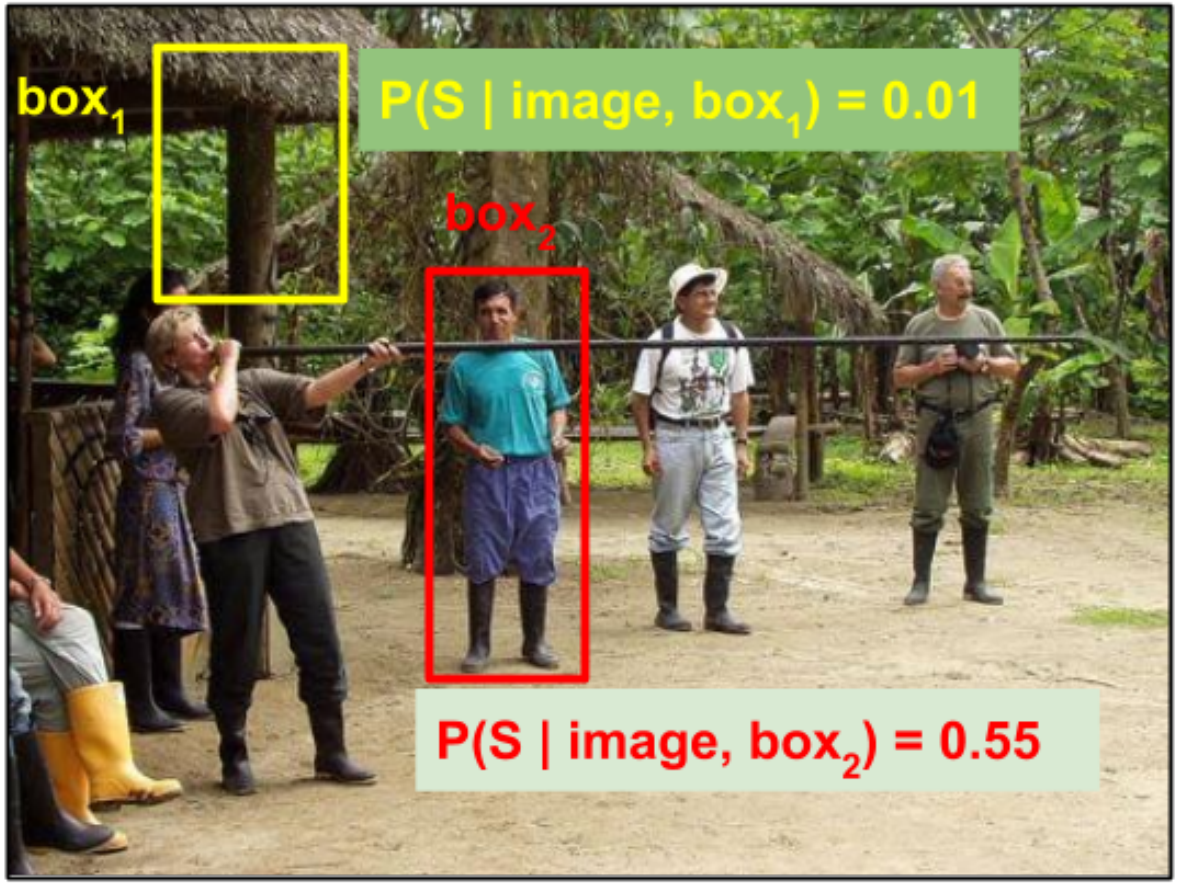
Basic pipeline:

1. Extract candidate bounding box locations
2. Score each candidate w.r.t query text
3. Return the highest scoring candidate box

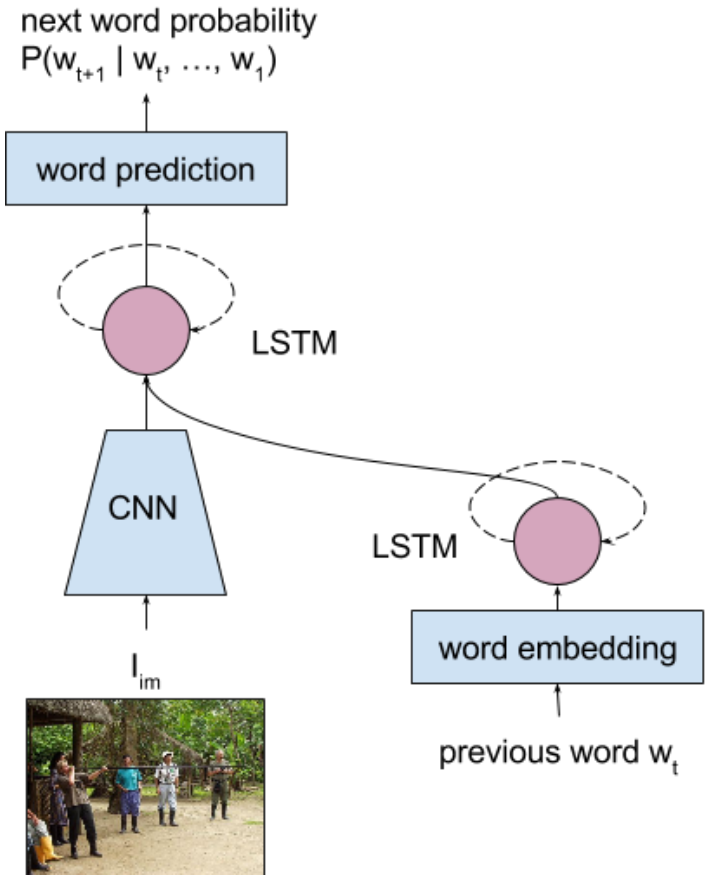


Use the **probability of query text** on a candidate box as its score.

query text S='*man in middle with blue shirt and blue shorts*'



Score the Candidate Boxes



LRCN model for image captioning
[Donahue et al. 2015]

LSTM-based image captioning models can compute p(sentence | image)

- Run an image captioning model to score the candidate regions?

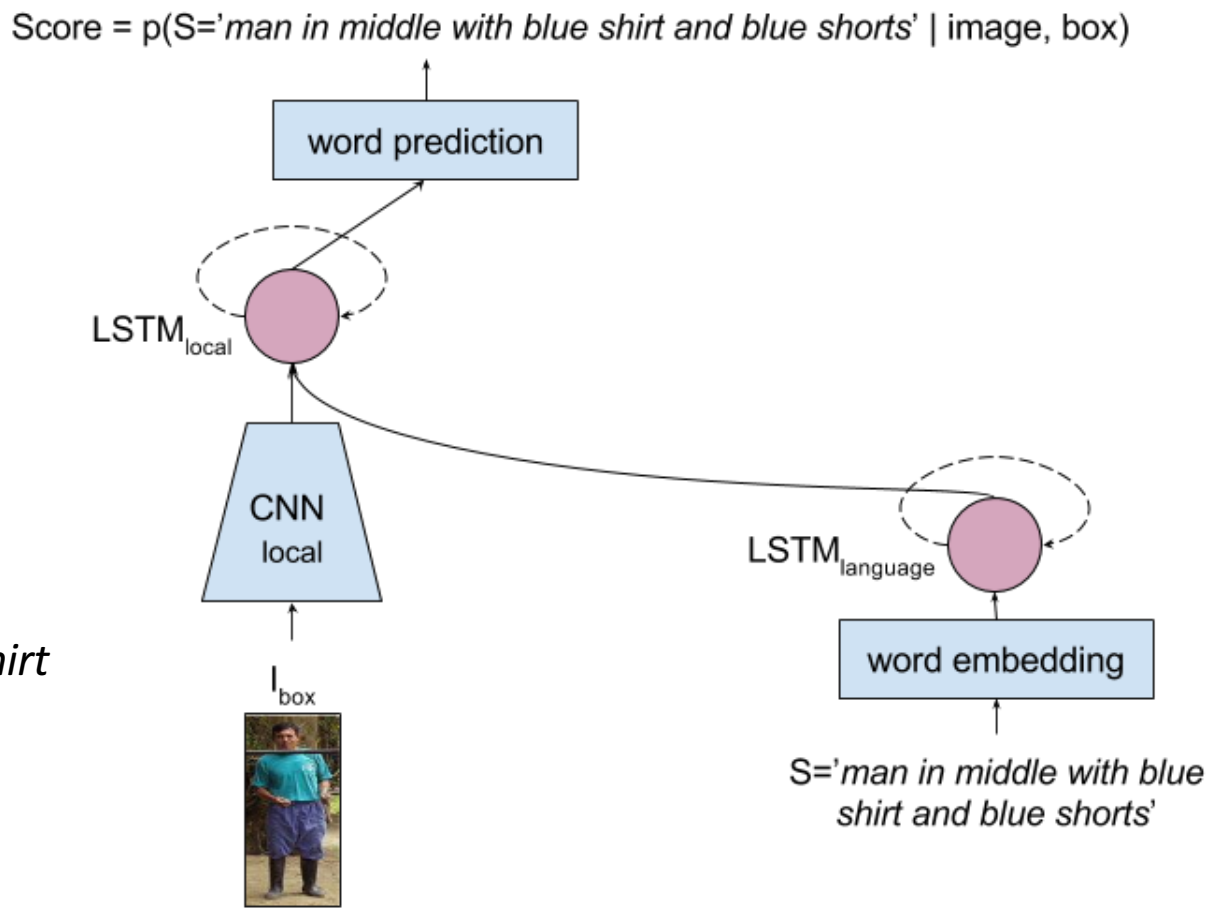
Score the Candidate Boxes

Compute p(query | box) using LRCN image captioning model

[Donahue et al. 2015].

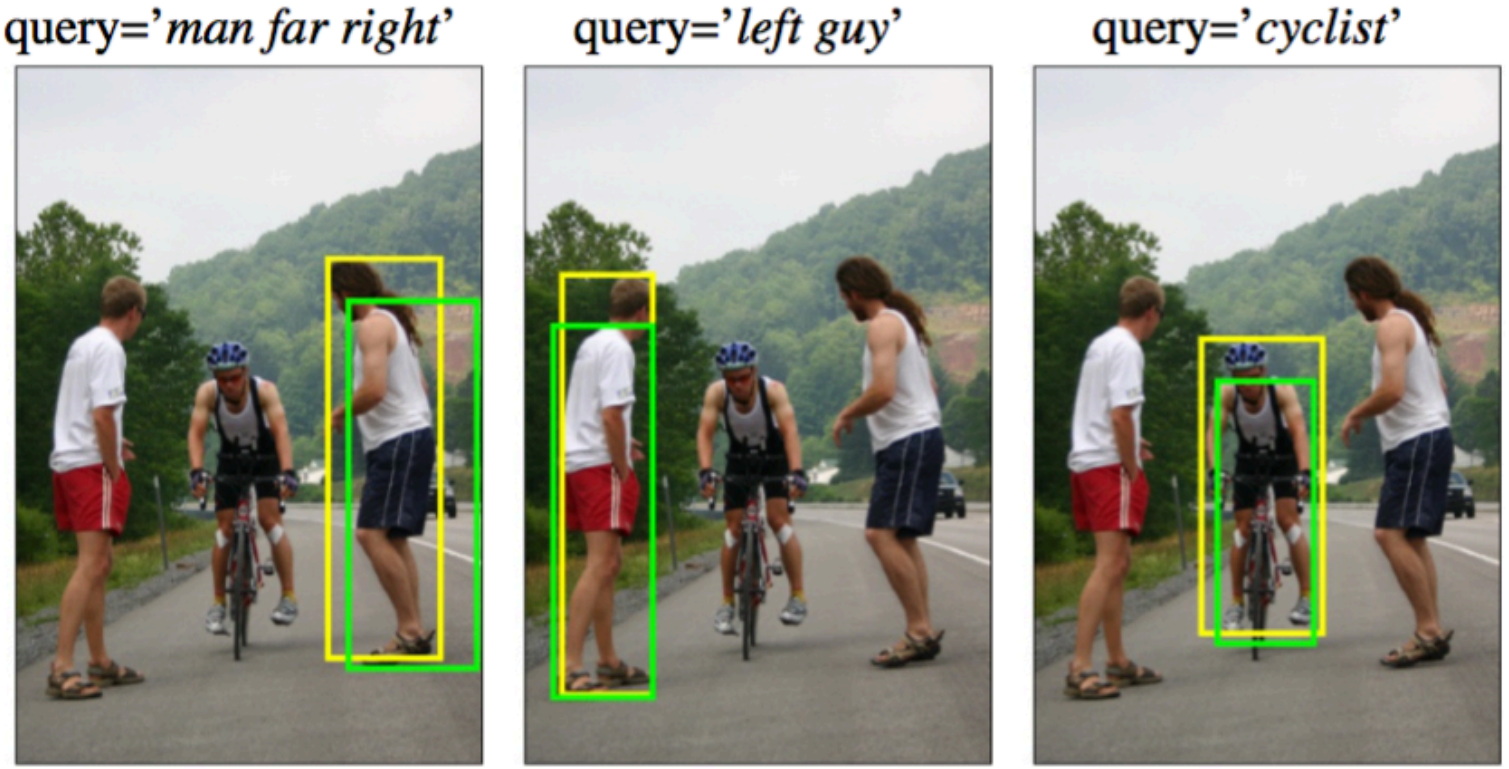
input image

query='*man in middle with blue shirt and blue shoes*'



Spatial Cues and Global Context

Spatial locations and **scene context** needed in this task.



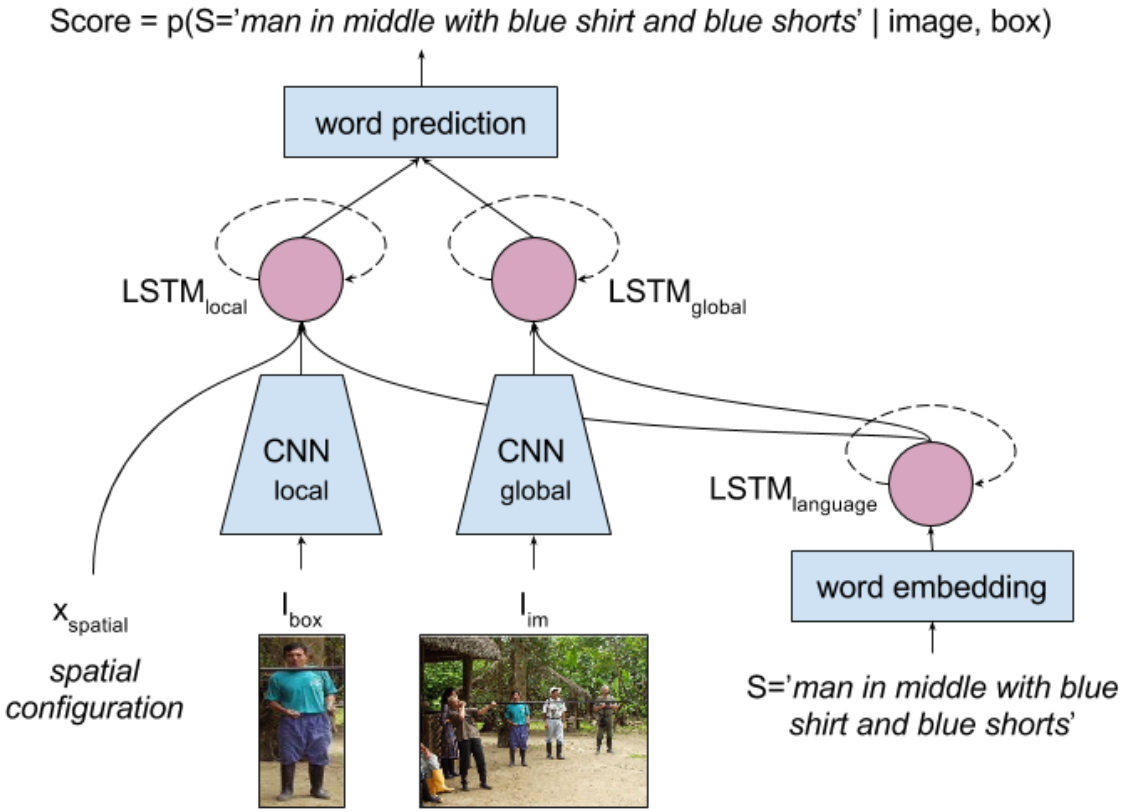
(***yellow***: ground-truth, ***green***: our prediction)

Spatial Context Recurrent ConvNet (SCRC)

- Spatial cue: bounding box coordinates
- Scene context: whole-image feature

input image

query='*man in middle with blue shirt and blue shoes*'

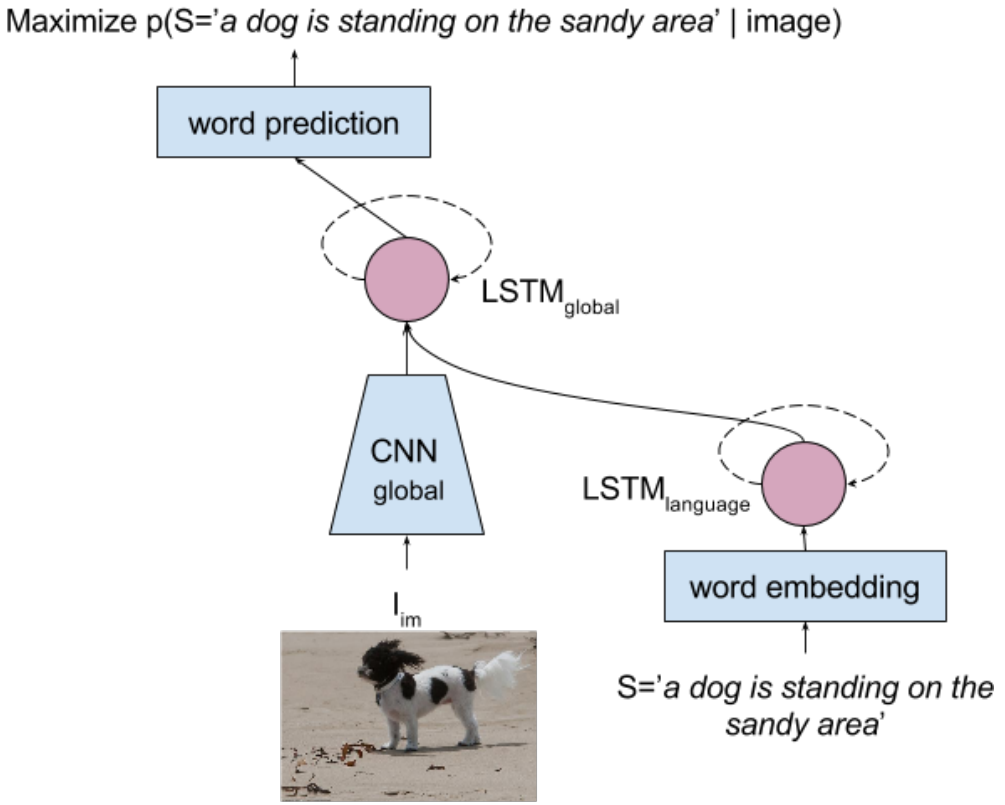


Knowledge Transfer from Image Captioning

- Transferring visual-linguistic knowledge by pretraining on image captioning data.

image

caption='*a dog is standing on the sandy area*'



Training on Object Retrieval Data

- Training: maximize the probabilities of description text on annotated regions.
- End-to-end trainable with back propagation.

bounding box with text annotation

text=*'man in middle with blue shirt and blue shoes'*

maximize p(text | image, box) during training

Experiments

- The ReferIt Dataset [Kazemzadeh et al. 2014]
  - Referring expressions on image regions
  - 20000 images (10000 trainval + 10000 test)
  - both object and stuff regions

- Training procedure
  - pretrain on MS COCO image captions
  - fine-tune on the ReferIt dataset

ReferIt Experiments

Candidate bounding boxes at test time:
top-100 EdgeBox proposals [Zitnick and Dollár, 2014].

Test on 100 EdgeBox proposals	SCRC Ablation Study			R@1	R@10
	transfer	spatial	context		
CAFFE-7K [Guadarrama et al. 2014]				10.38%	26.20%
LRCN [Donahue et al. 2015]				8.59%	31.86%
SCRC	×	×	×	14.53%	40.72%

ReferIt Experiments

Candidate bounding boxes at test time:
top-100 EdgeBox proposals [Zitnick and Dollár, 2014].

Test on 100 EdgeBox proposals	SCRC Ablation Study			R@1	R@10
	transfer	spatial	context		
CAFFE-7K [Guadarrama et al. 2014]				10.38%	26.20%
LRCN [Donahue et al. 2015]				8.59%	31.86%
SCRC	×	×	×	14.53%	40.72%
SCRC	✓	×	×	15.78%	42.54%

ReferIt Experiments

Candidate bounding boxes at test time:
top-100 EdgeBox proposals [Zitnick and Dollár, 2014].

Test on 100 EdgeBox proposals	SCRC Ablation Study			R@1	R@10
	transfer	spatial	context		
CAFFE-7K [Guadarrama et al. 2014]				10.38%	26.20%
LRCN [Donahue et al. 2015]				8.59%	31.86%
SCRC	×	×	×	14.53%	40.72%
SCRC	✓	×	×	15.78%	42.54%
SCRC	✓	✓	×	17.68%	44.77%

ReferIt Experiments

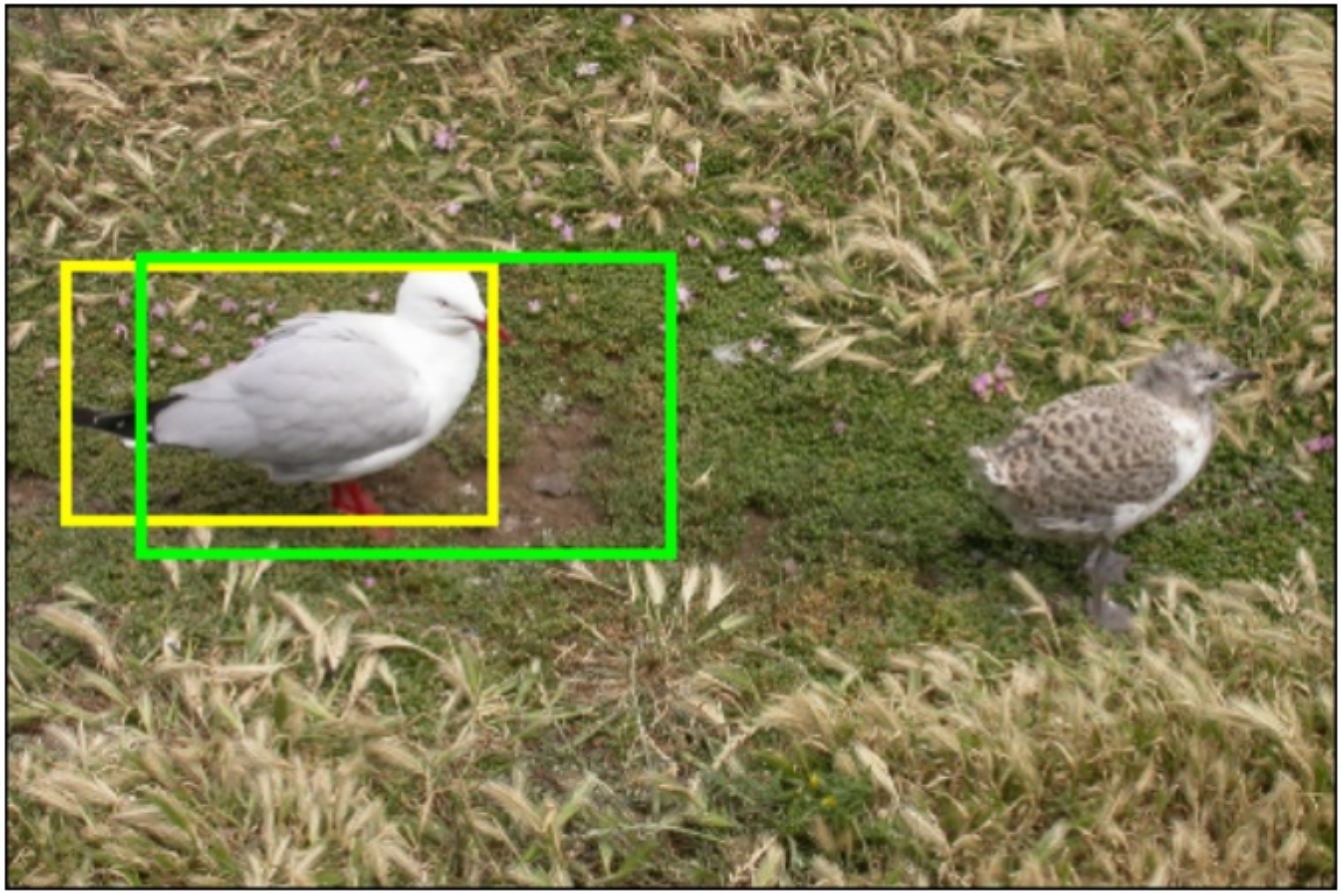
Candidate bounding boxes at test time:
top-100 EdgeBox proposals [Zitnick and Dollár, 2014].

Test on 100 EdgeBox proposals	SCRC Ablation Study			R@1	R@10
	transfer	spatial	context		
CAFFE-7K [Guadarrama et al. 2014]				10.38%	26.20%
LRCN [Donahue et al. 2015]				8.59%	31.86%
SCRC	×	×	×	14.53%	40.72%
SCRC	✓	×	×	15.78%	42.54%
SCRC	✓	✓	×	17.68%	44.77%
SCRC	✓	✓	✓	**17.93%**	**45.27%**
Upper bound (coverage of candidate boxes)				59.38%	59.38%

Visualized Results

***green*: highest scoring prediction**
***yellow*: ground-truth region**

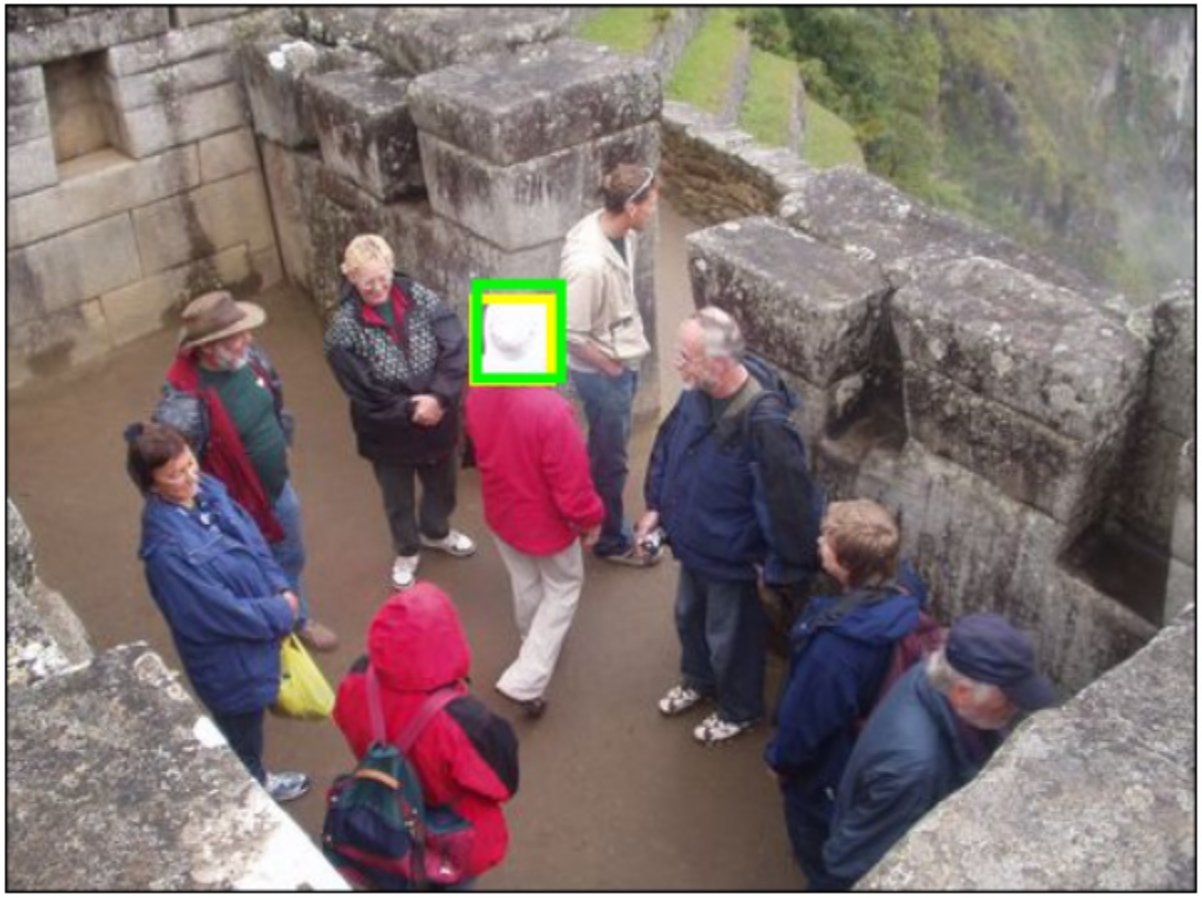
query='*bird on the left*'



Visualized Results

***green*: highest scoring prediction**
***yellow*: ground-truth region**

query='*white hat*'



Visualized Results

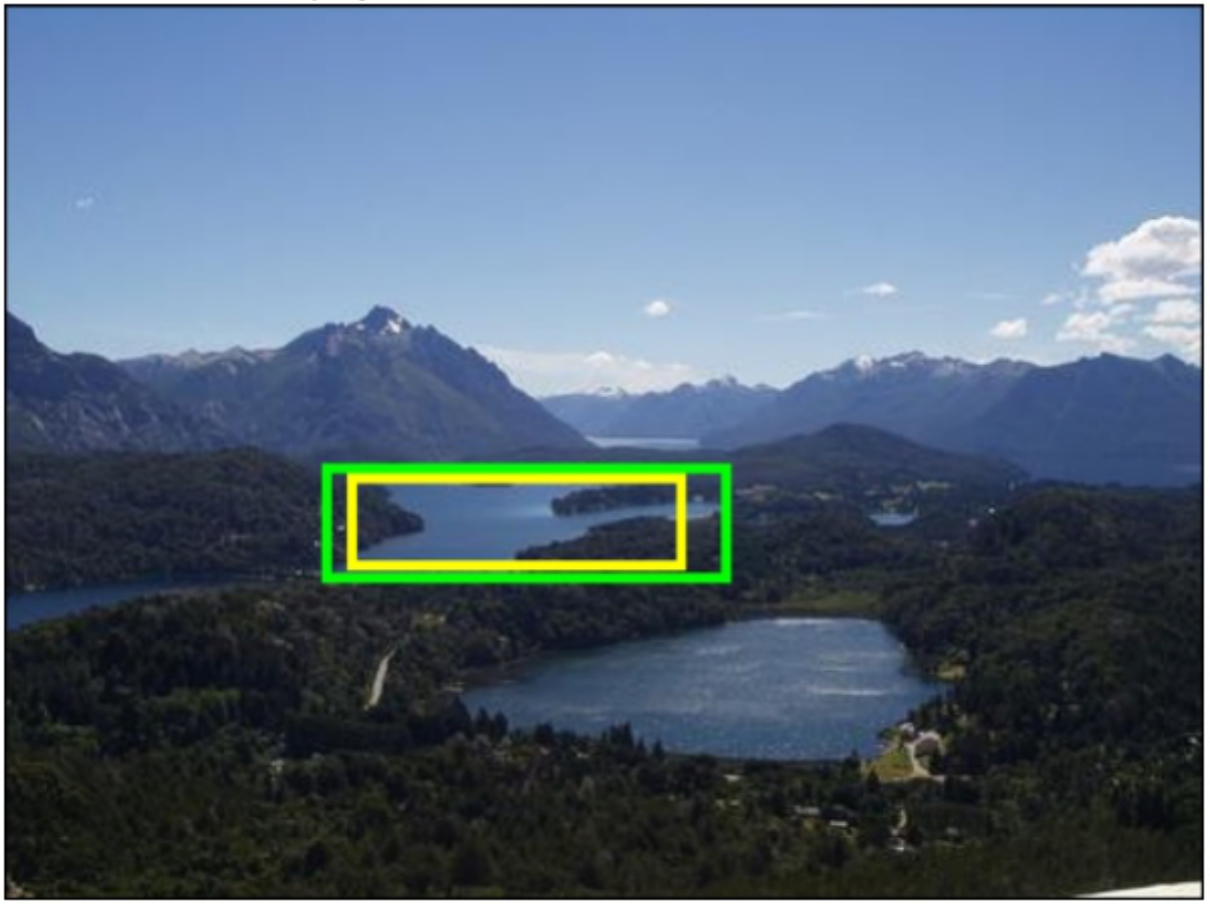
***green*: highest scoring prediction**
***yellow*: ground-truth region**

query='*Window with closed curtains*'

Visualized Results

***green*: highest scoring prediction**
***yellow*: ground-truth region**

query='*lake underneath the mountain on the left*'



Visualized Results

***green*: highest scoring prediction**
***yellow*: ground-truth region**

query='*leaves of left tree*'

Visualized Results

***green*: highest scoring prediction**
***yellow*: ground-truth region**

query='*far right person*'

query='*lady in black shirt*'

Visualized Results

***green*: highest scoring prediction**
***yellow*: ground-truth region**

query='*2 people on left*'

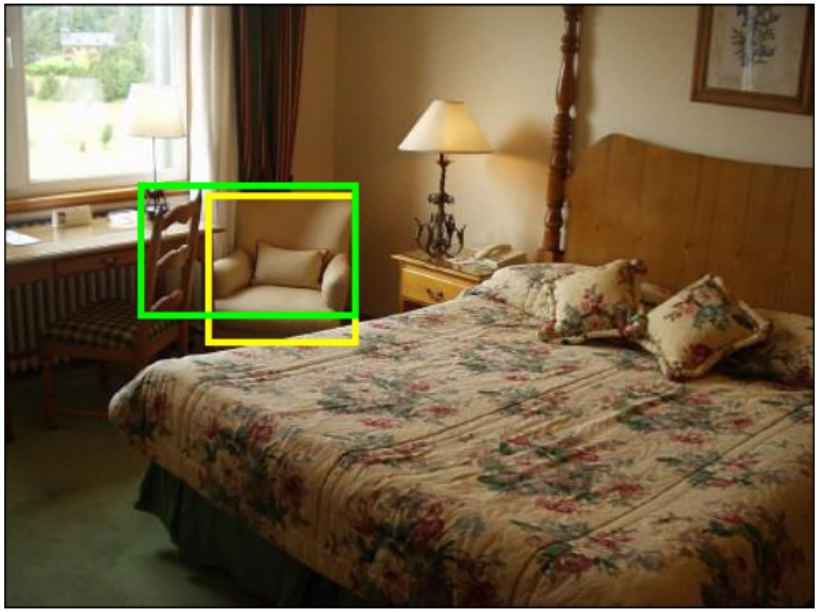
query='*dude center with backpack blue*'

Visualized Results

***green*: highest scoring prediction**
***yellow*: ground-truth region**

query='*chair left*'

query='*nice plush chair*'



Failure Cases

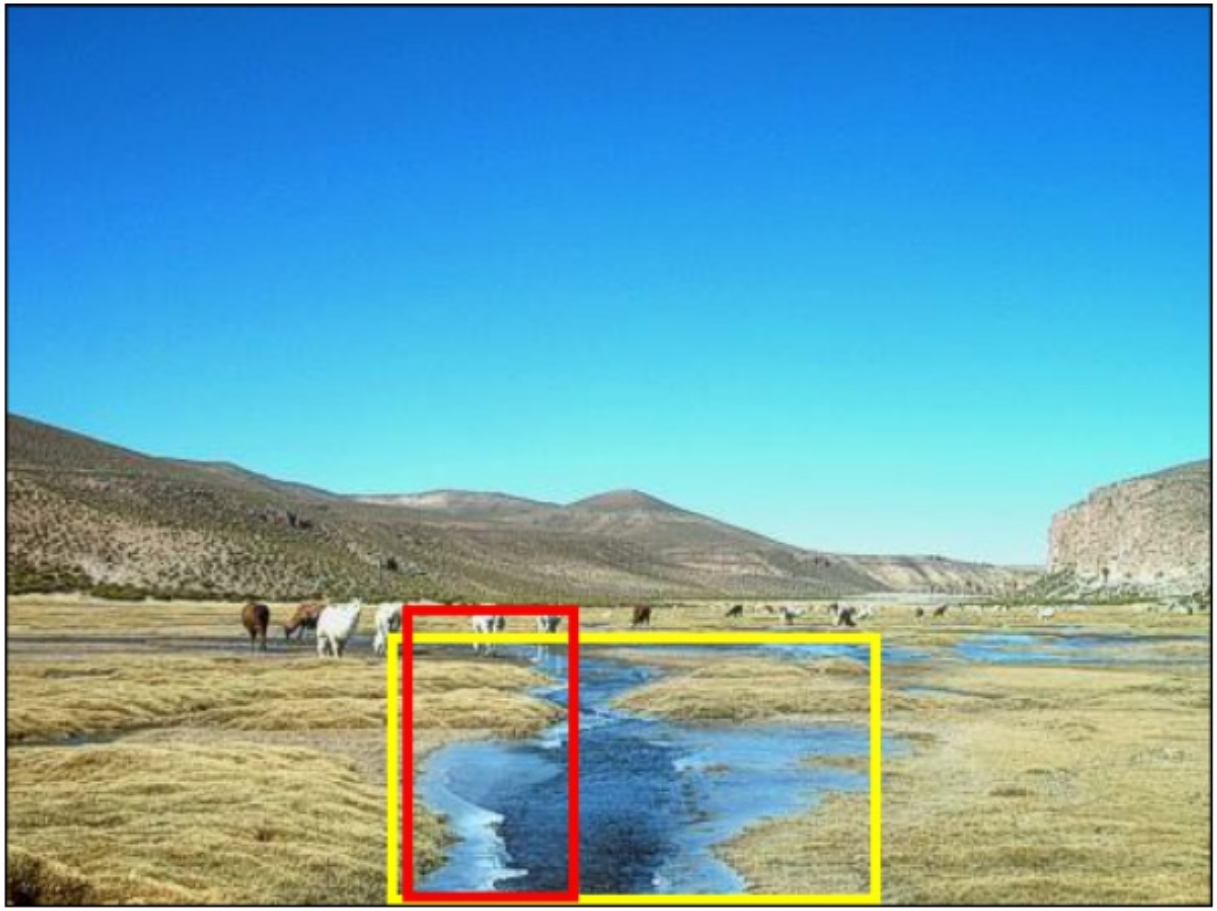
***red*: highest scoring prediction**
***yellow*: ground-truth region**

query='*picture 2nd from left*'

Failure Cases

***red*: highest scoring prediction**
***yellow*: ground-truth region**

query='*water in the center bottom*'



Flickr30K Entities Experiment

- On the Flickr30K Entities dataset [Plummer et al. 2015], our SCRC method outperforms the previous Canonical Correlation Analysis (CCA) results.

Test on 100 EdgeBox proposals	R@1	R@10
Canonical Correlation Analysis (CCA) [Plummer et al. 2015]	25.3%	59.7%
SCRC	**27.8%**	**62.9%**
Upper bound (coverage of candidate boxes)	76.9%	76.9%

Kitchen Experiment

- Test scenario: retrieving cropped objects regions
  - 11 candidate regions: 1 true object + 10 distractors (sampled from either Kitchen itself or ImageNET).

query='*whisk with red tipped handle*'

query='*mobile phone the pink color*'

	SCRC transfer	10 distractors from Kitchen	10 distractors from ImageNET
CAFFE-7K [Guadarrama et al. 2014]		51.34%	57.50%
LRCN [Donahue et al. 2015]		40.35%	63.22%
SCRC	✗	54.02%	74.08%
SCRC	✓	**61.62%**	**81.15%**

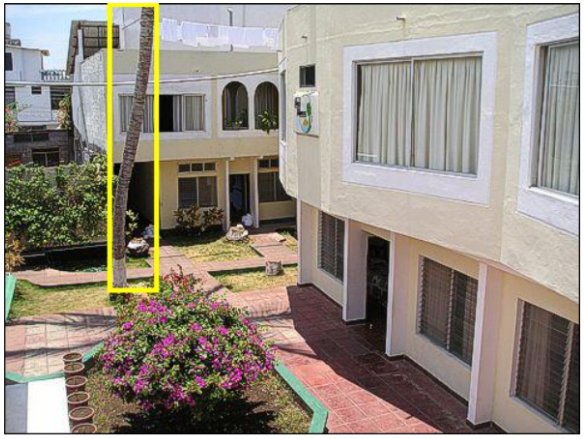
Generating Region Descriptions

- Our model can also be applied to generate text descriptions for given regions in an image using p(text | image, box).

***yellow*: given image region to describe**

generated description=
'yellow car'

generated description=
'desk in front of kid with red shirt'



generated description=
'tree trunk left'

Conclusion

Natural language object retrieval with Spatial Context Recurrent ConvNet (SCRC) model

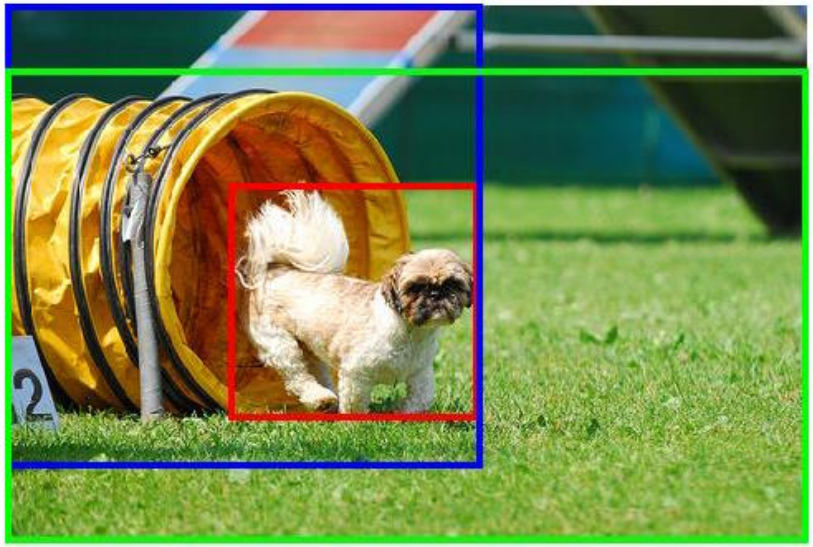
- local descriptors, spatial cues and scene context
- transfers knowledge from image captioning

More details:

- Project page: ronghanghu.com/text_obj_retrieval
- Extended paper: arxiv.org/pdf/1511.04164

Follow-up work: GroundeR

- Discriminatively trained
- Significant improvement over SCRC



A little brown and white dog emerges from a yellow collapsible toy tunnel onto the lawn.

- a little brown and white dog
- a yellow collapsible toy tunnel
- the lawn

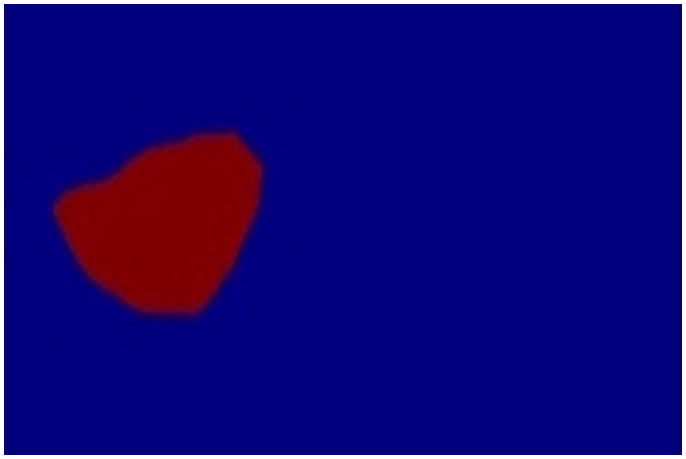
Anna Rohrbach, Marcus Rohrbach, Ronghang Hu, Trevor Darrell, Bernt Schiele.
Grounding of textual phrases in images by reconstruction.

Details: arxiv.org/pdf/1511.03745

Segmentation from Referential Expressions

query='*the bird on the left*'

input image



output segmentation results

Ronghang Hu, Marcus Rohrbach, and Trevor Darrell.
Segmentation from Natural Language Expressions.

Details: arxiv.org/pdf/1603.06180