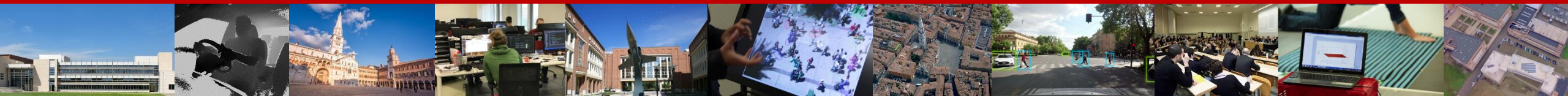
Part III:

Controllable Image Captioning



Marcella Cornia

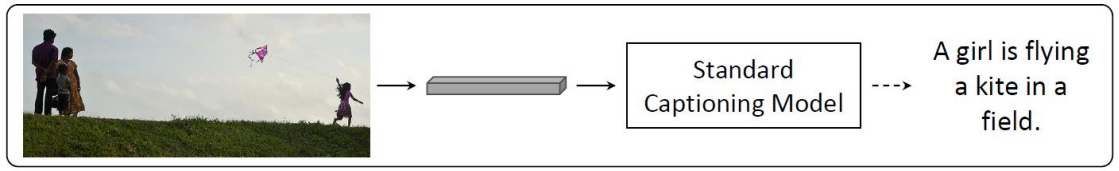
marcella.cornia@unimore.it

University of Modena and Reggio Emilia, Italy

Introduction

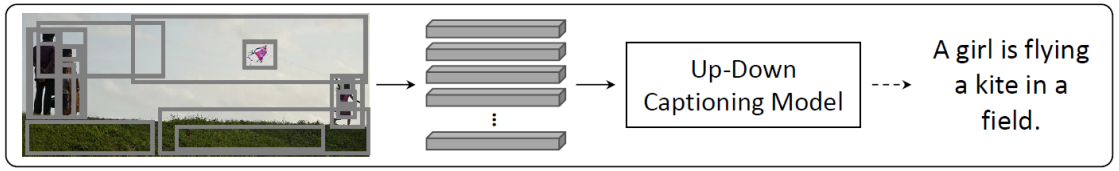
Early captioning approaches:

- Global image feature vector



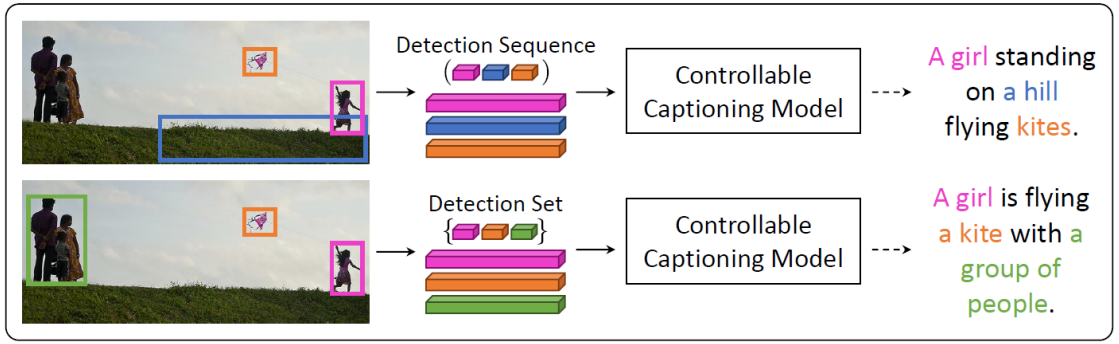
Attention-based approaches:

- Weakly interpretable (through attention)
- Not controllable.
  - We can't decide which regions get processed
  - No control over the generation process.



Show, Control and Tell

- Controllable via regions
  - A sequence (ordered)
  - A set (unordered)

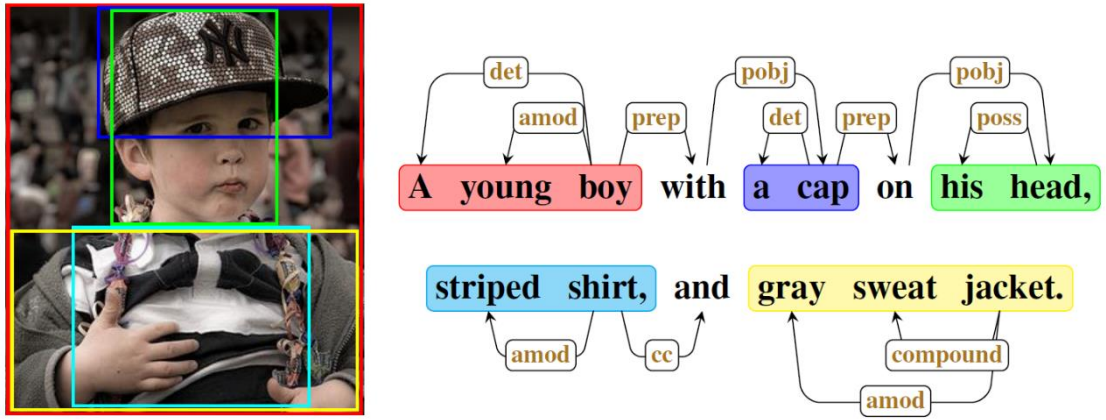


Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. "Show, Control and Tell: A Framework for Generating Grounded and Controllable Captions." CVPR 2019.

Structure and Diversity of Captions

A step back: what is the right abstraction level for generating sentences?

- So far: word level
- *Ours: noun chunks*



Noun chunk: a noun grouped together with its modifiers.

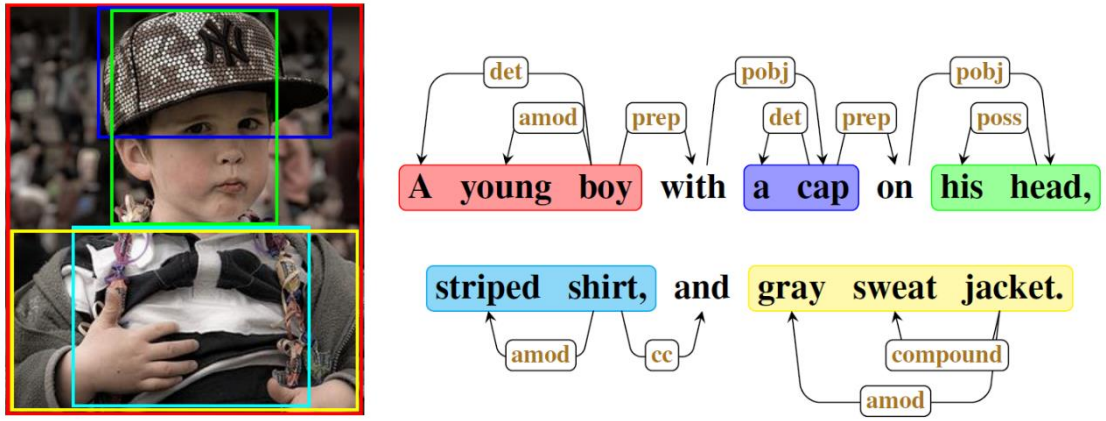
- Easy to obtain via NLP tools (dependency tree)

Noun chunks can be mapped to image regions

Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. "Show, Control and Tell: A Framework for Generating Grounded and Controllable Captions." CVPR 2019.

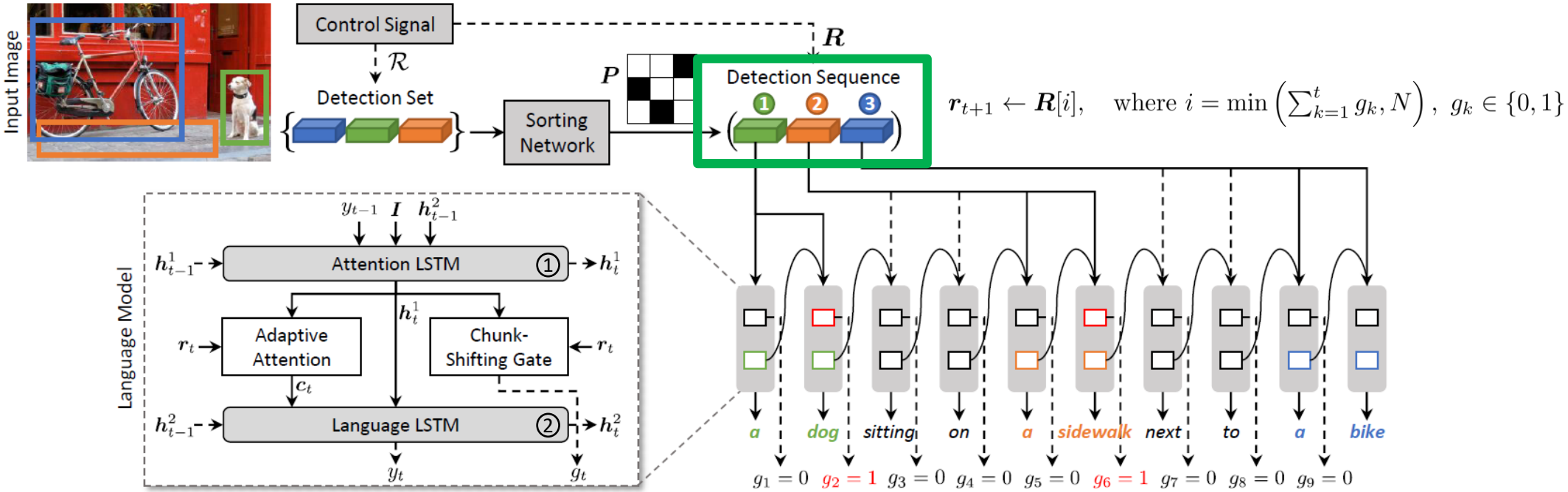
Structure and Diversity of Captions

- A sentence is a sequence of noun chunks, eventually associated with an image region
- Different captions → different chunks → different regions
- **Captions differ in terms of**:
  - The set of selected regions
  - The order in which they are described
  - Their mapping to noun chunks (linguistic abilities of the annotator) → Learned!



Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. "Show, Control and Tell: A Framework for Generating Grounded and Controllable Captions." CVPR 2019.

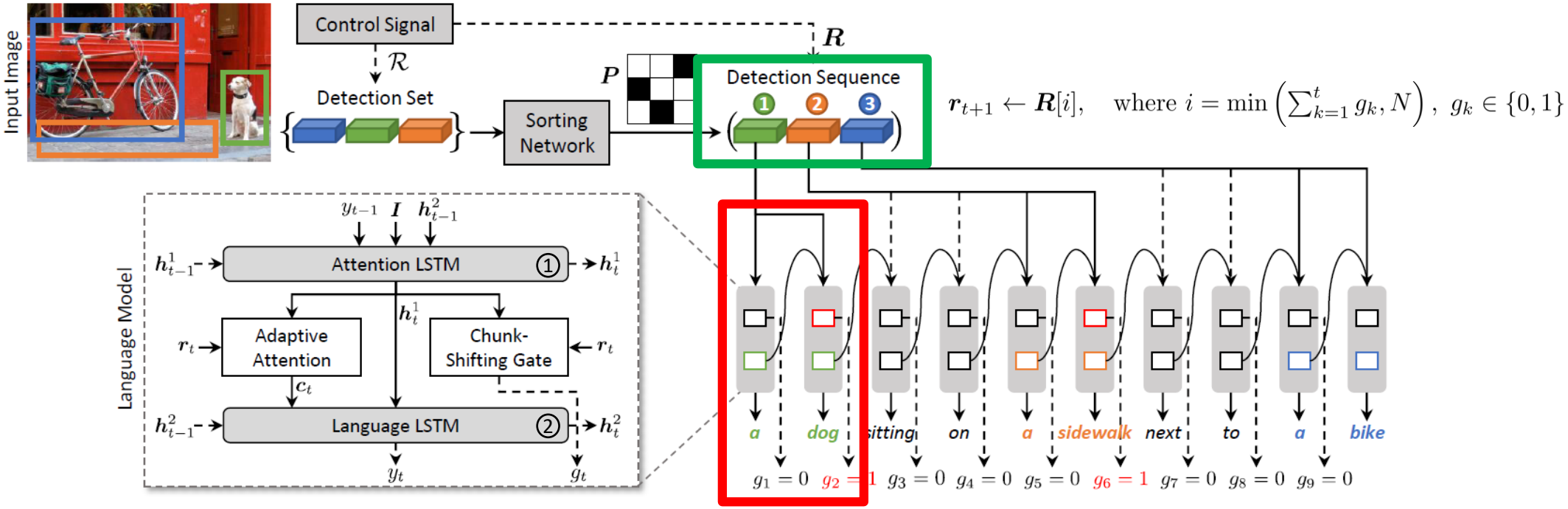
Generating Controllable Captions



- Language model takes as input **a sequence of regions**
- Switches between one region and the next one via a learned chunk-shifting gate
  - When it's done with the generation of chunk, it moves to the next region in the sequence

Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. "Show, Control and Tell: A Framework for Generating Grounded and Controllable Captions." CVPR 2019.

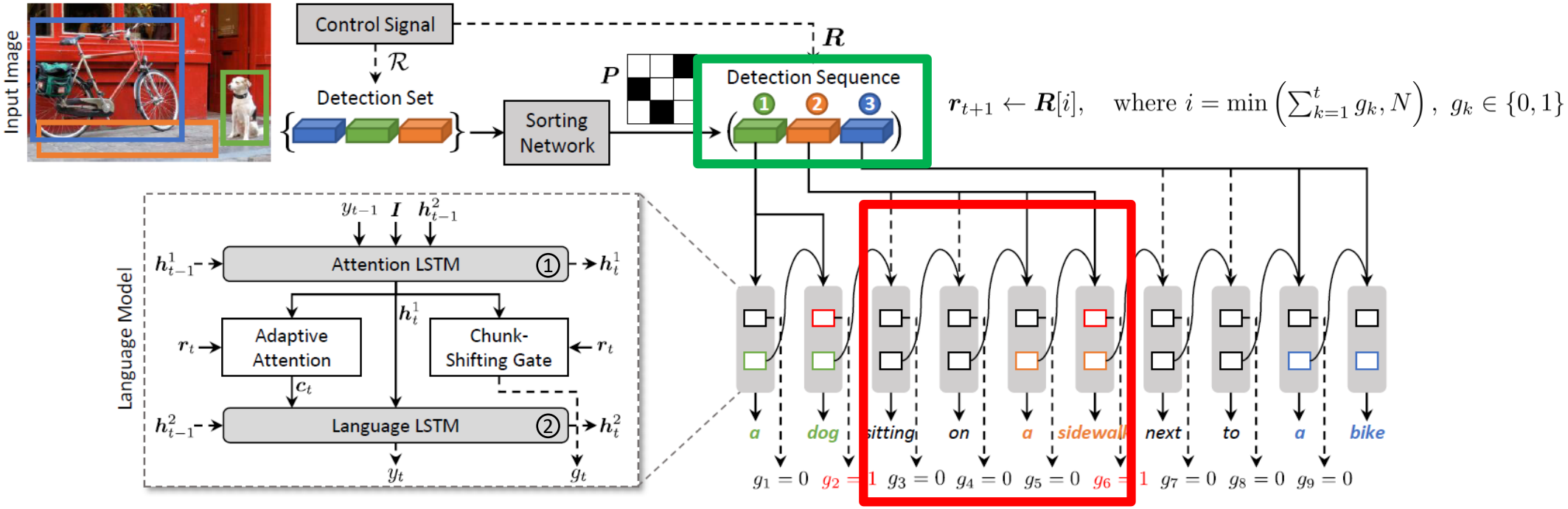
Generating Controllable Captions



- Language model takes as input a sequence of regions
- **Switches between one region and the next one via a learned chunk-shifting gate**
  - When it's done with the generation of chunk, it moves to the next region in the sequence

Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. "Show, Control and Tell: A Framework for Generating Grounded and Controllable Captions." CVPR 2019.

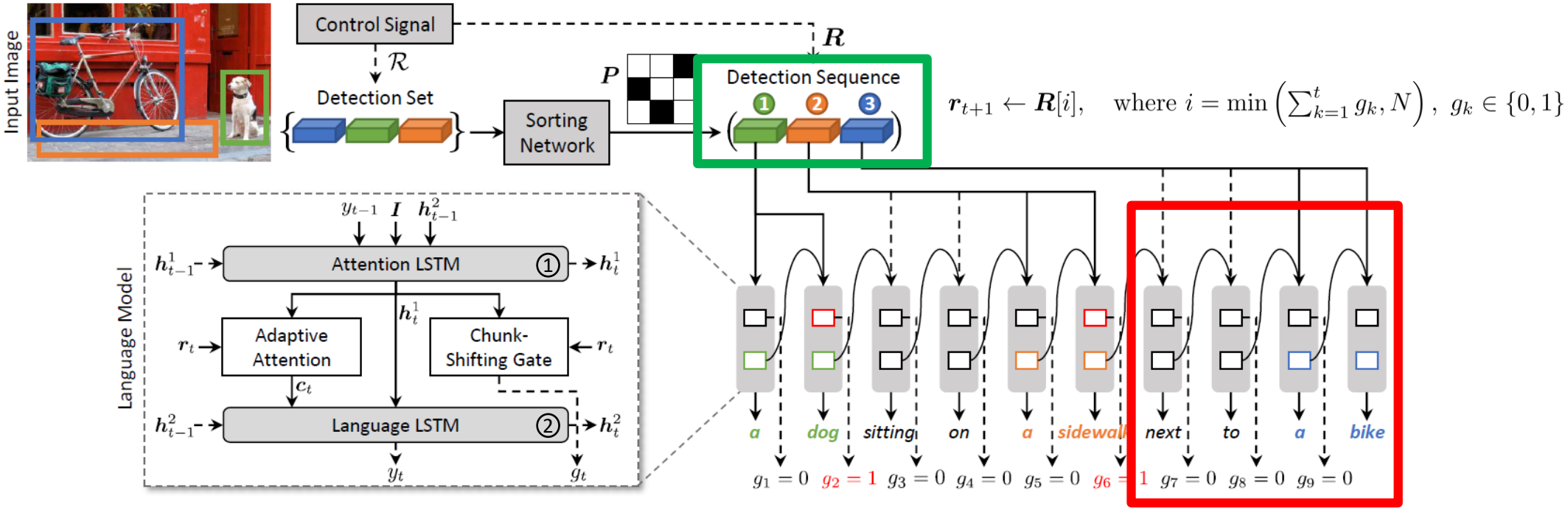
Generating Controllable Captions



- Language model takes as input a sequence of regions
- **Switches between one region and the next one via a learned chunk-shifting gate**
  - When it's done with the generation of chunk, it moves to the next region in the sequence

Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. "Show, Control and Tell: A Framework for Generating Grounded and Controllable Captions." CVPR 2019.

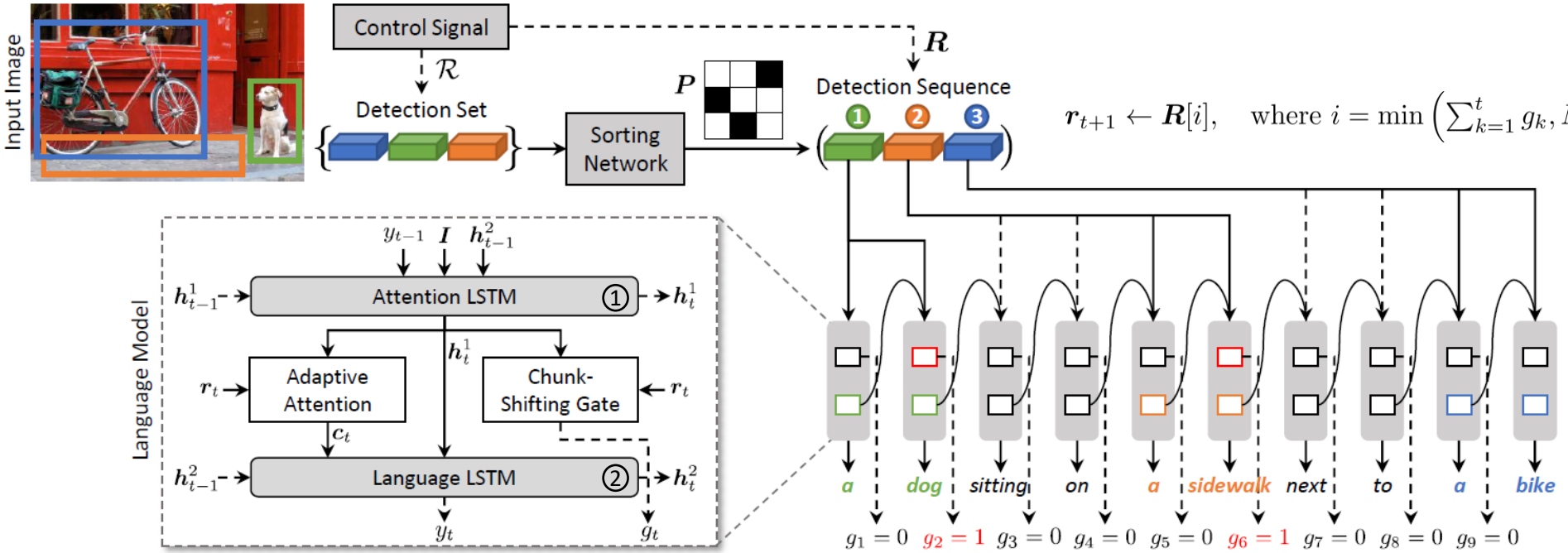
Generating Controllable Captions



- Language model takes as input a sequence of regions
- **Switches between one region and the next one via a learned chunk-shifting gate**
  - When it's done with the generation of chunk, it moves to the next region in the sequence

Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. "Show, Control and Tell: A Framework for Generating Grounded and Controllable Captions." CVPR 2019.

Training

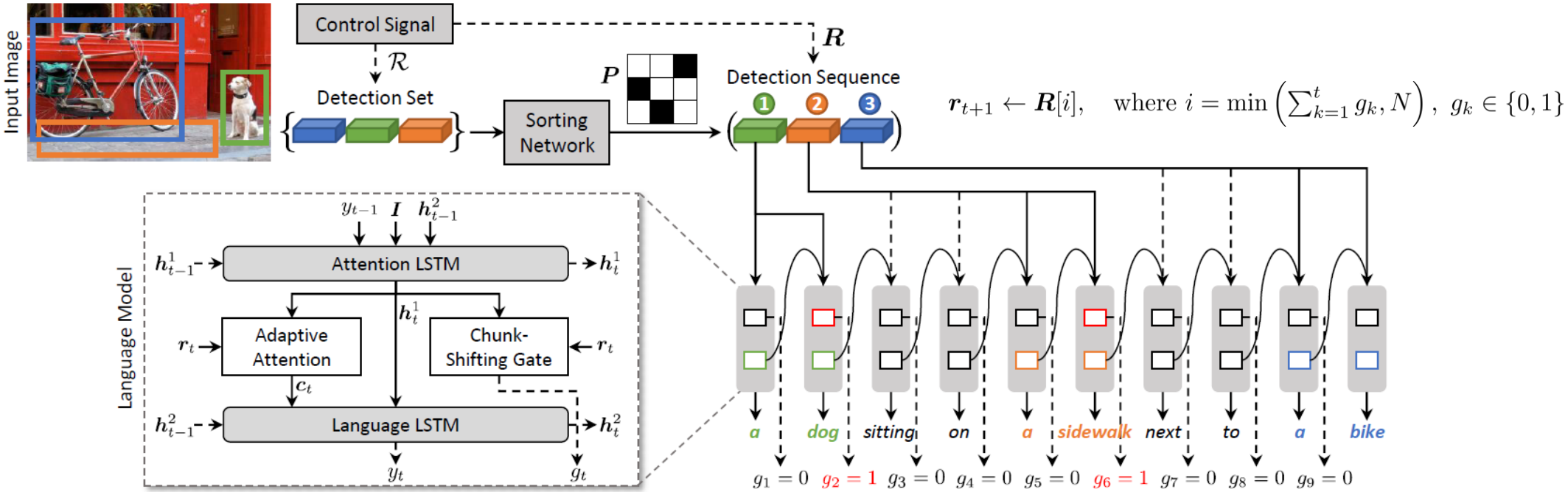


- Train on GT words and shifting gate values (obtained via NLP)

$$L(\theta) = -\sum_{t=1}^{T}\Big(\log \overbrace{p(y_t^*|\boldsymbol{r}_{1:t}^*, \boldsymbol{y}_{1:t-1}^*)}^{\text{Word-level probability}} + g_t^* \log p(g_t = 1|\boldsymbol{r}_{1:t}^*, \boldsymbol{y}_{1:t-1}^*) + (1 - g_t^*)\log(1 - \underbrace{p(g_t = 1|\boldsymbol{r}_{1:t}^*, \boldsymbol{y}_{1:t-1}^*)}_{\text{Chunk-level probability}})\Big)$$

Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. "Show, Control and Tell: A Framework for Generating Grounded and Controllable Captions." CVPR 2019.

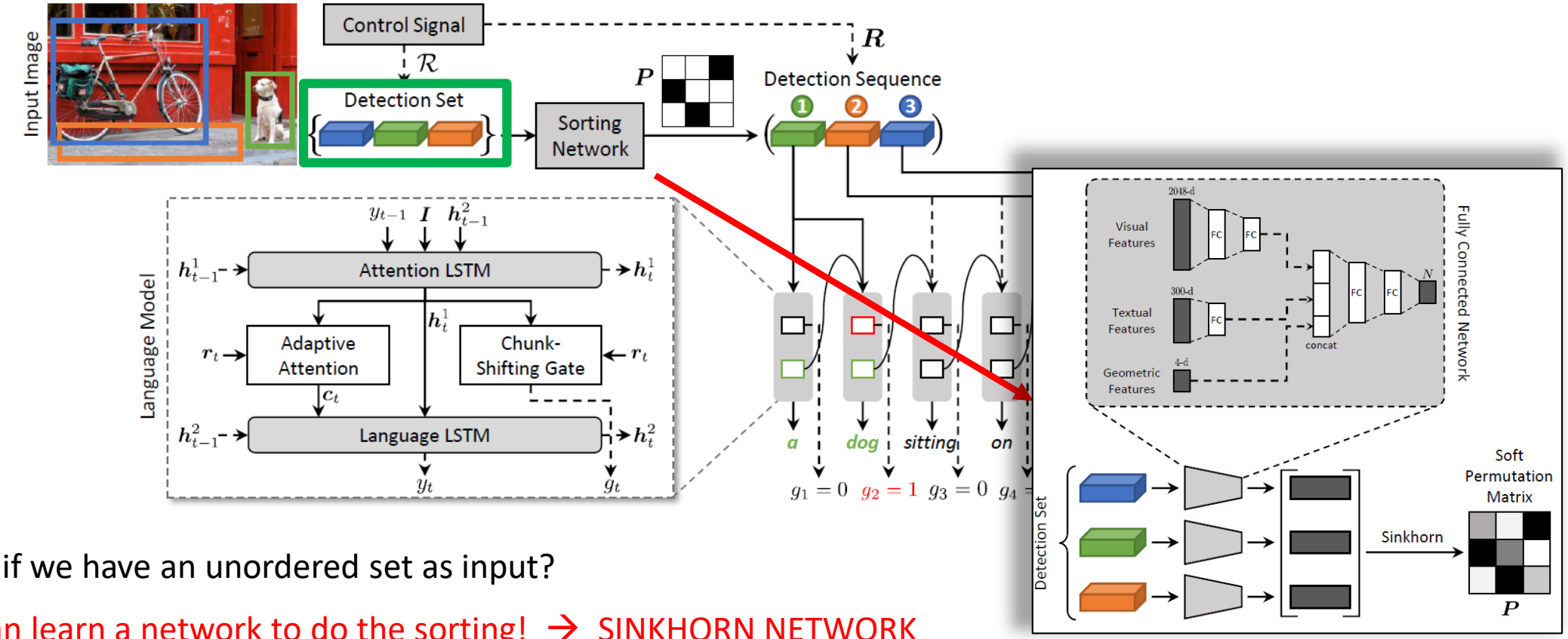
Rewarding Correct Captions



- Train on GT words and shifting gate values (obtained via NLP)
- then, finetune using Reinforcement Learning
  - CIDEr wrt GT caption (caption quality)
  - Plus, use the alignment between the predicted and GT chunks as reward (Needleman-Wunsch algorithm)

Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. "Show, Control and Tell: A Framework for Generating Grounded and Controllable Captions." CVPR 2019.

Sorting Network



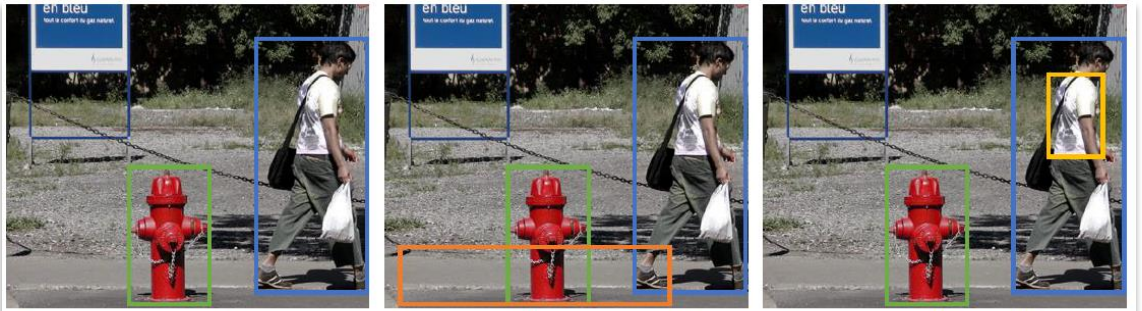
- What if we have an unordered set as input?
- We can learn a network to do the sorting! → SINKHORN NETWORK
  - Approximates a derivable permutation matrix
  - Train on real data, then use the Hungarian to get the true permutation matrix.

Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. "Show, Control and Tell: A Framework for Generating Grounded and Controllable Captions." CVPR 2019.

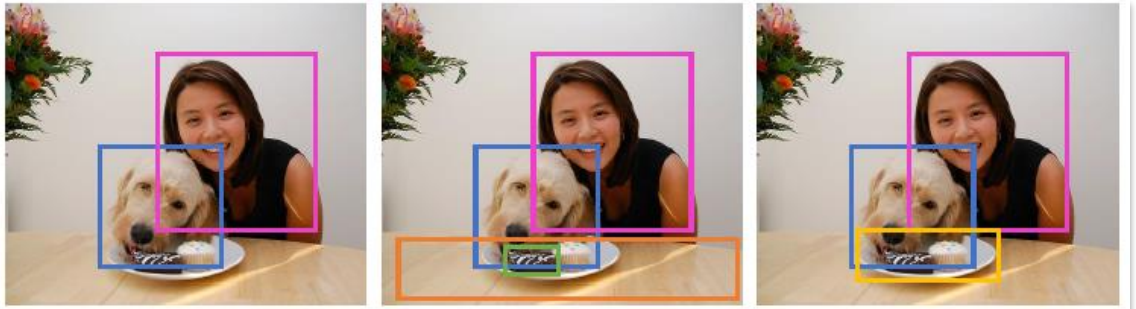
COCO Entities Dataset

For training and evaluation, we collect COCO-Entities
→ more than 120,000 images

- COCO with noun chunks associated to regions
- Semi-automatically annotated



A young man walking past a red fire hydrant. | A man walks past a red fire hydrant on the sidewalk. | A man in a white t-shirt walking past a fire hydrant.



A young girl is sitting down with her dog. | A woman sitting at a table with a dog eating cake. | A woman and a dog that is eating from a plate.

Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. "Show, Control and Tell: A Framework for Generating Grounded and Controllable Captions." CVPR 2019.

Controllable Image Captioning: Results

- Controllability via a set of regions
  - 75.5% intersection-over-union with GT chunks!
  - Adds more diversity than methods tailored for diversity ☺

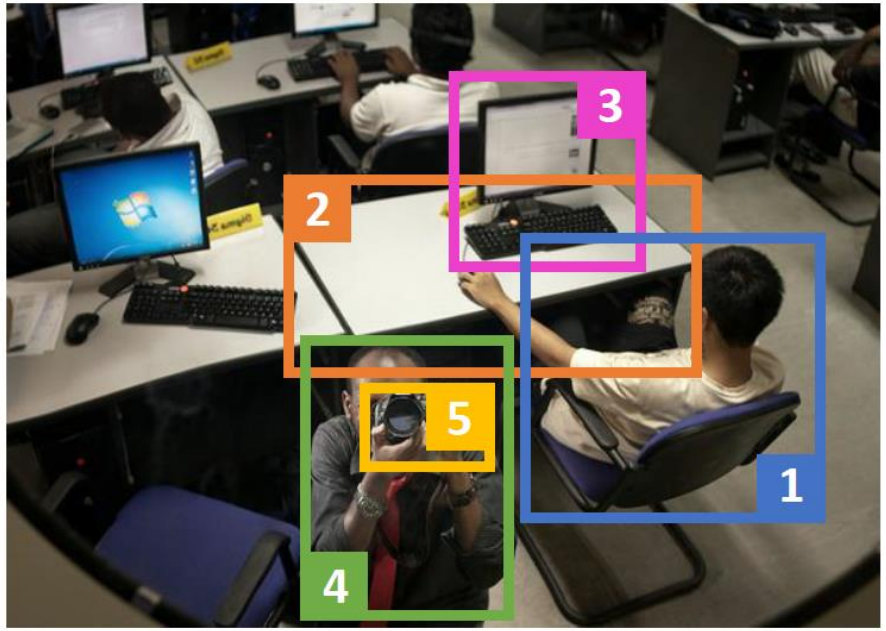
Method	B-4	M	R	C	S	IoU
FC-2K† [36]	12.5	18.5	39.6	116.5	26.6	61.0
Up-Down† [3]	14.4	20.0	42.2	132.8	29.7	63.2
Neural Baby Talk† [24]	13.1	19.2	40.5	119.1	29.2	62.6
Controllable LSTM	12.9	19.3	41.3	123.4	28.7	64.2
Controllable Up-Down	**18.1**	23.6	48.4	170.5	40.4	71.6
Ours *w/* single sentinel	17.4	23.6	48.4	168.4	43.7	75.4
Ours *w/o* visual sentinel	17.6	23.4	48.5	168.9	43.6	75.3
Ours	18.0	**23.8**	**48.9**	**173.3**	**44.1**	**75.5**

Method	Samples	B-4	M	R	C	S
AG-CVAE [44]	20	**47.1**	30.9	63.8	130.8	24.4
POS [8]	20	44.9	36.5	67.8	146.8	27.7
Ours	20	44.8	**36.6**	**68.9**	**156.5**	**30.9**

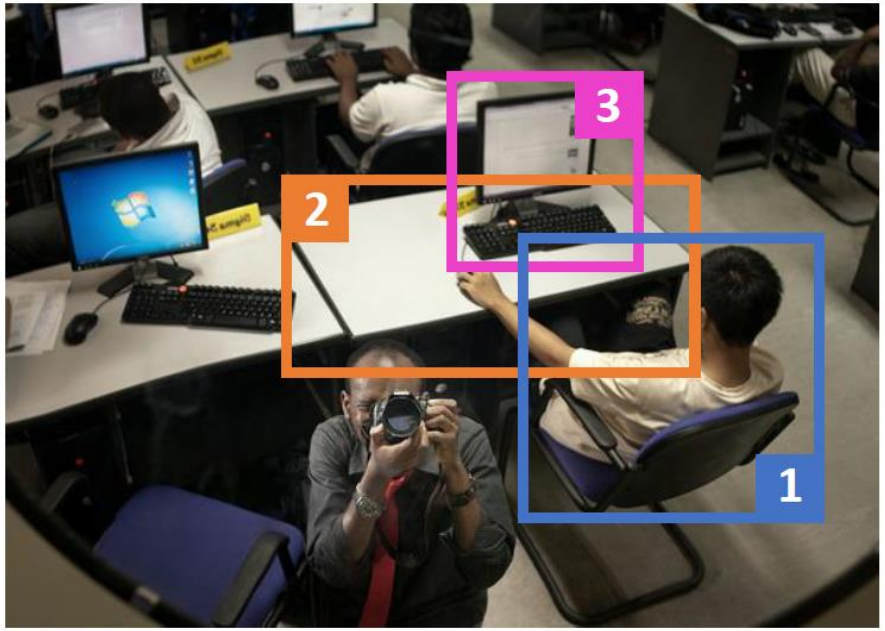
Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. "Show, Control and Tell: A Framework for Generating Grounded and Controllable Captions." CVPR 2019.

Controllability via a Sequence of Regions

Results when controlling with a sequence of regions



A man sitting at a desk with a computer and a man holding a camera.

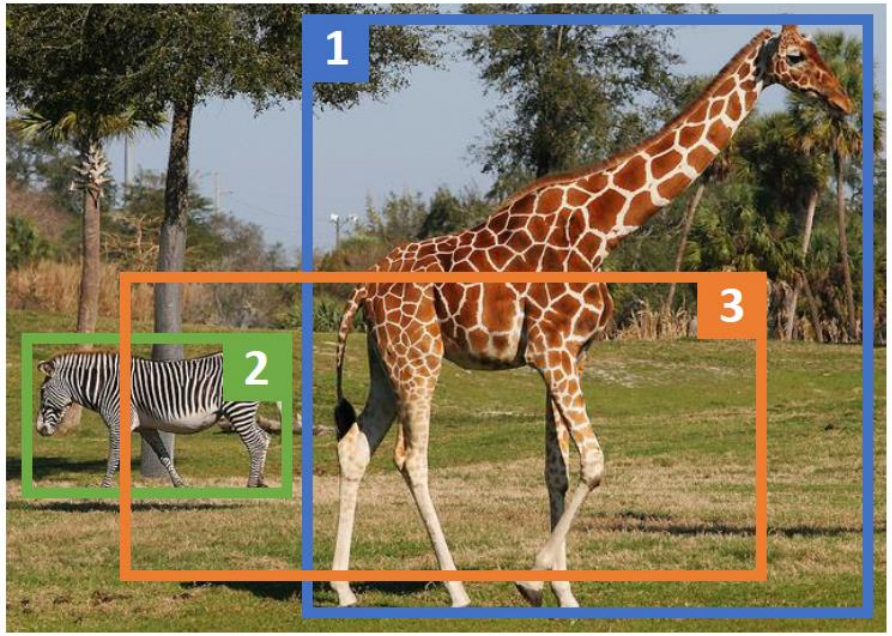


A man sitting at a desk with a computer.

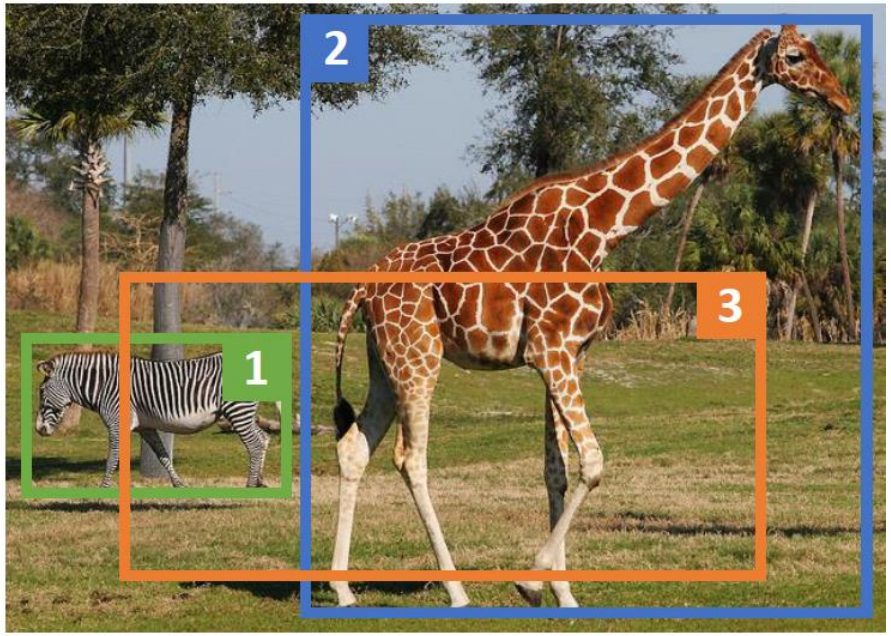
Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. "Show, Control and Tell: A Framework for Generating Grounded and Controllable Captions." CVPR 2019.

Controllability via a Sequence of Regions

Results when controlling with a sequence of regions



A giraffe standing in front of a zebra in a field.

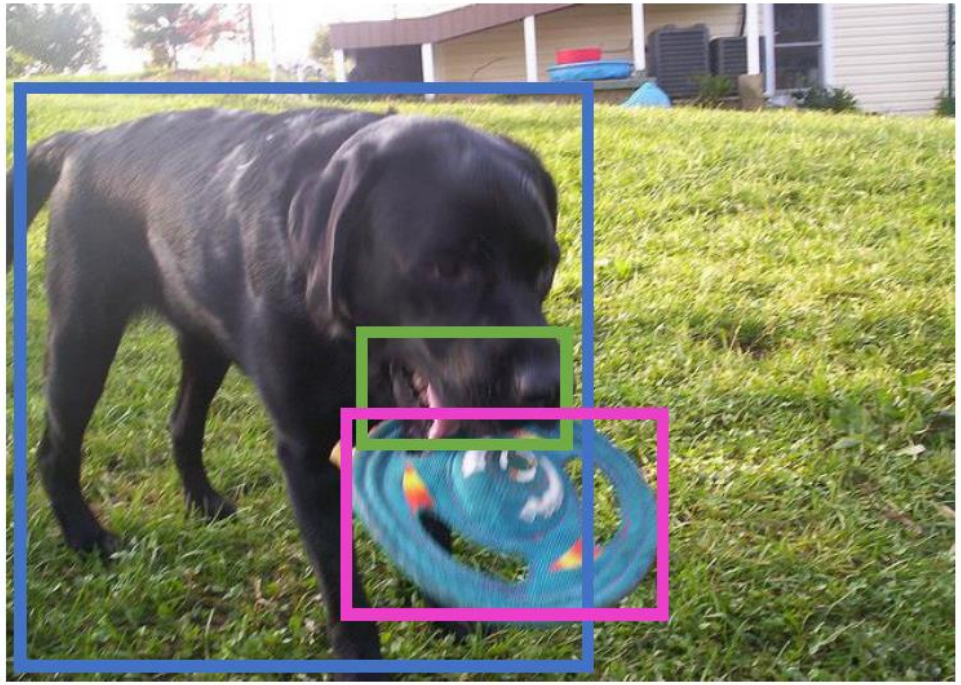


A zebra standing next to a giraffe in a field.

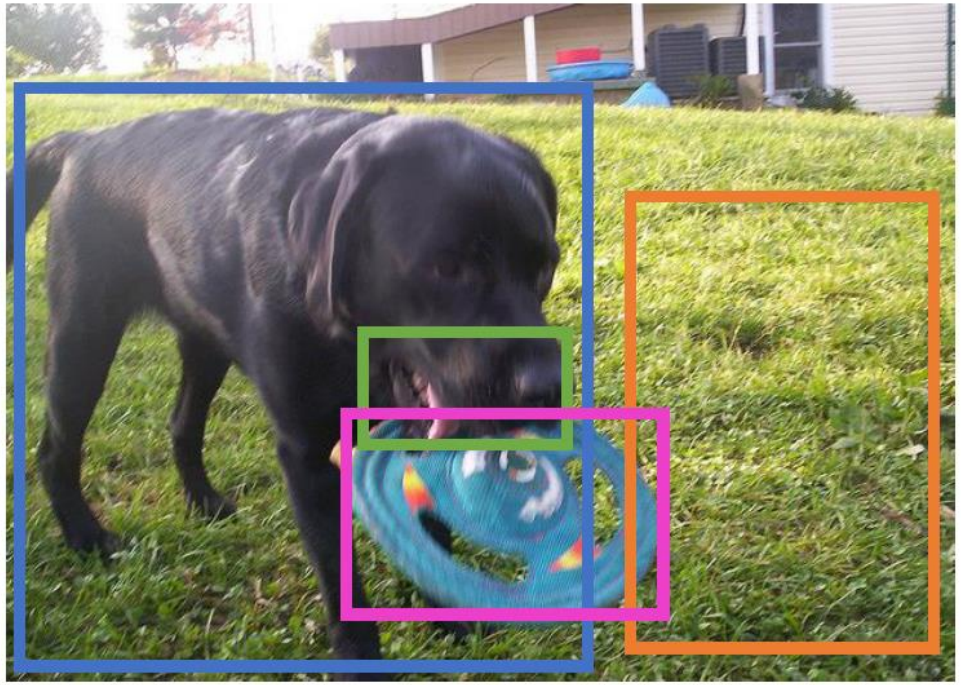
Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. "Show, Control and Tell: A Framework for Generating Grounded and Controllable Captions." CVPR 2019.

Controllability via a Set of Regions

Results when controlling with a set of regions



A dog holding a frisbee in its mouth.



A dog standing in the grass with a frisbee in its mouth.

Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. "Show, Control and Tell: A Framework for Generating Grounded and Controllable Captions." CVPR 2019.

Controllability via a Set of Regions

Results when controlling with a set of regions

A man in a black jacket skiing down a hill.

A man on skis down a snow covered slope.

Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. "Show, Control and Tell: A Framework for Generating Grounded and Controllable Captions." CVPR 2019.

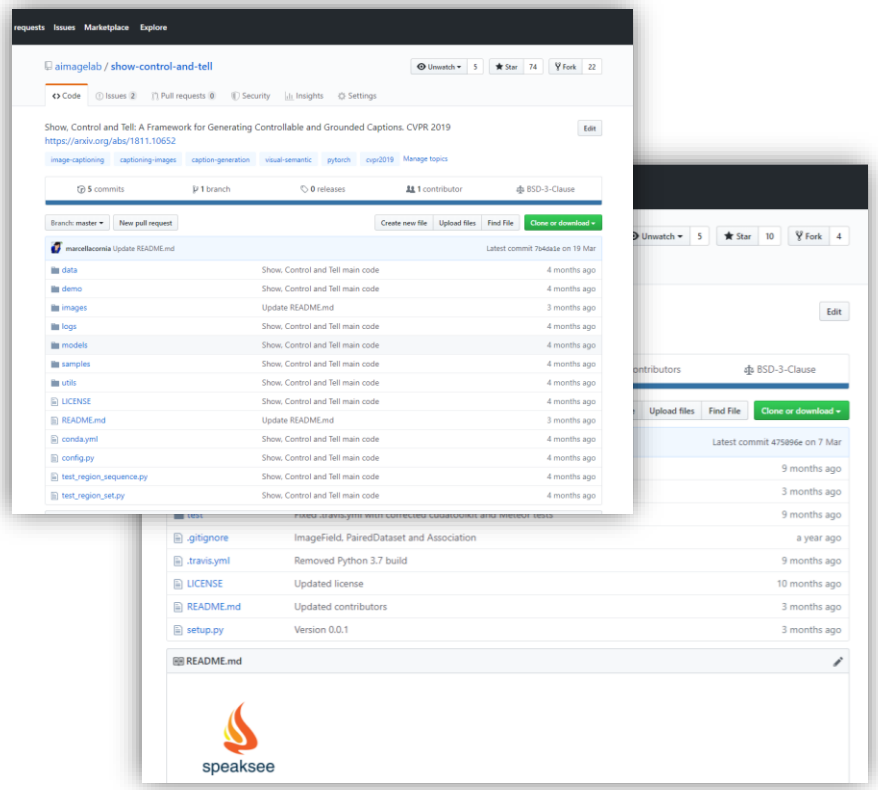
Show, Control and Tell: Source Code

↓ Get your hands dirty

Dataset, code, pre-trained models are available at

- https://github.com/aimagelab/show-control-and-tell

  → based on **speaksee**



Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. "Show, Control and Tell: A Framework for Generating Grounded and Controllable Captions." CVPR 2019.