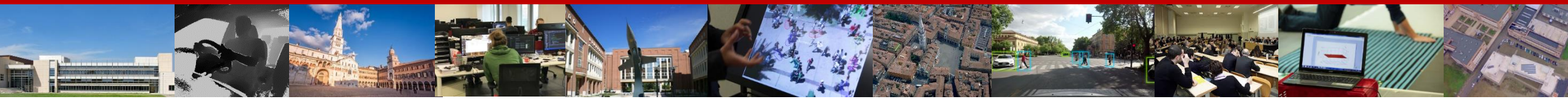
UNIMORE
UNIVERSITÀ DEGLI STUDI DI MODENA E REGGIO EMILIA

AImage Lab

Artificial Intelligence
Research and Innovation Center (AIRI)

Vision, Language and Action:

from Captioning to Embodied AI



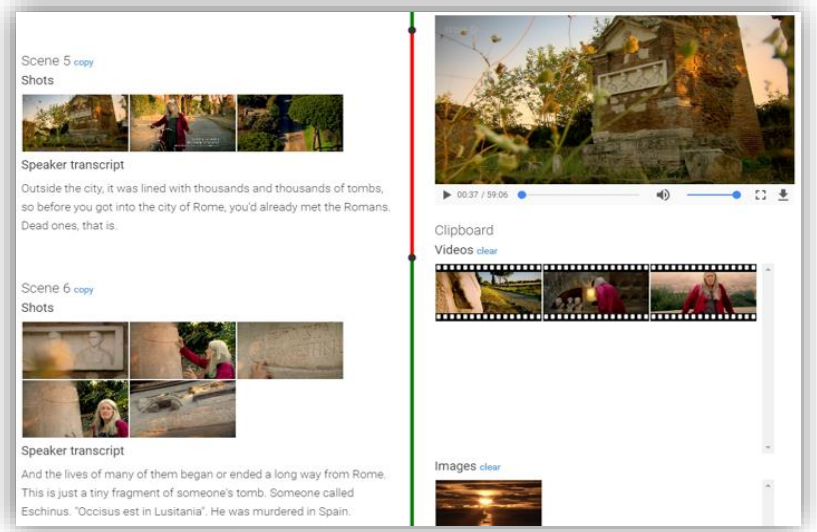
Lorenzo Baraldi, Marcella Cornia, Federico Landi

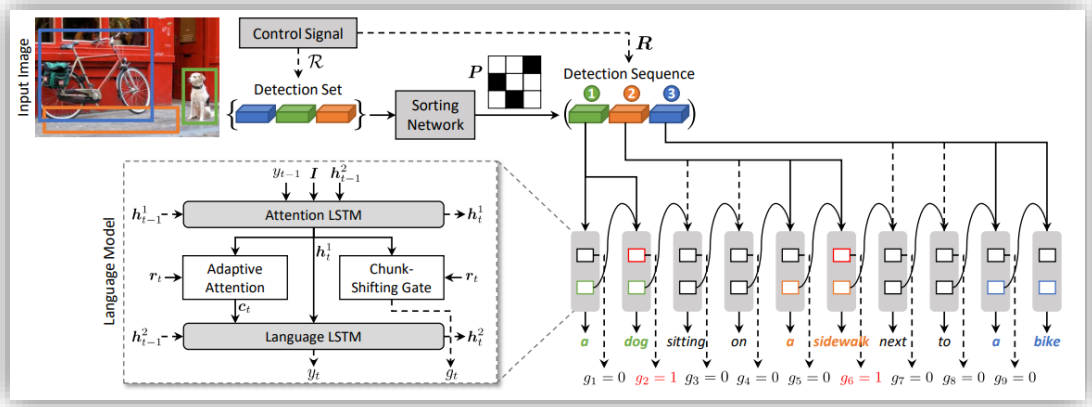
{name.surname}@unimore.it

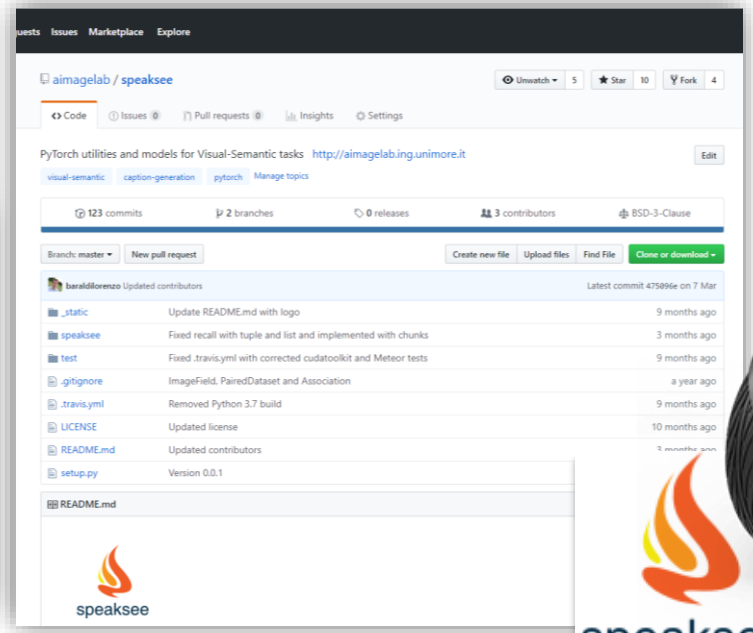
University of Modena and Reggio Emilia, Italy

RESEARCH ACTIVITIES

- Expertise on Sequence modelling
  - LSTM-based, Transformer-based
- Integrating Vision and Language
- Embodied AI: Vision, Language and Navigation
- Video modelling
- soon to come: **Explainable AI**







CONTROLLABLE LSTMS

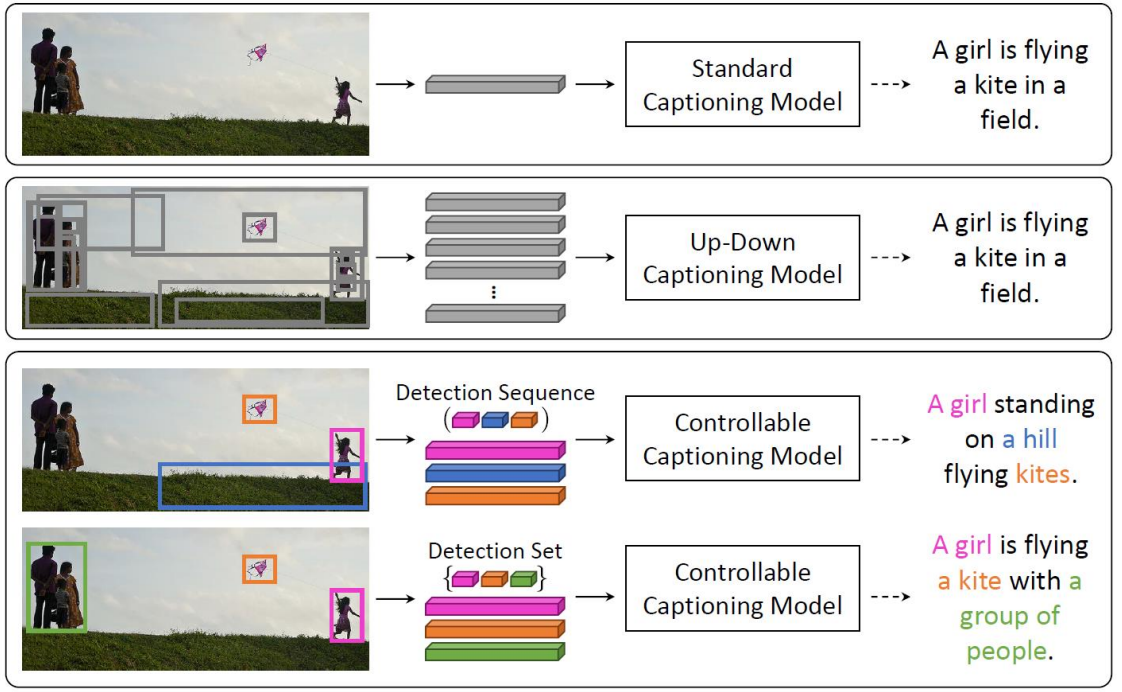
Image Captioning:

- Describe an image through a conditional language model
- Usually: (a set of) visual feature vectors + LSTM-based language model with attention
- One image, one caption

Show, control and tell

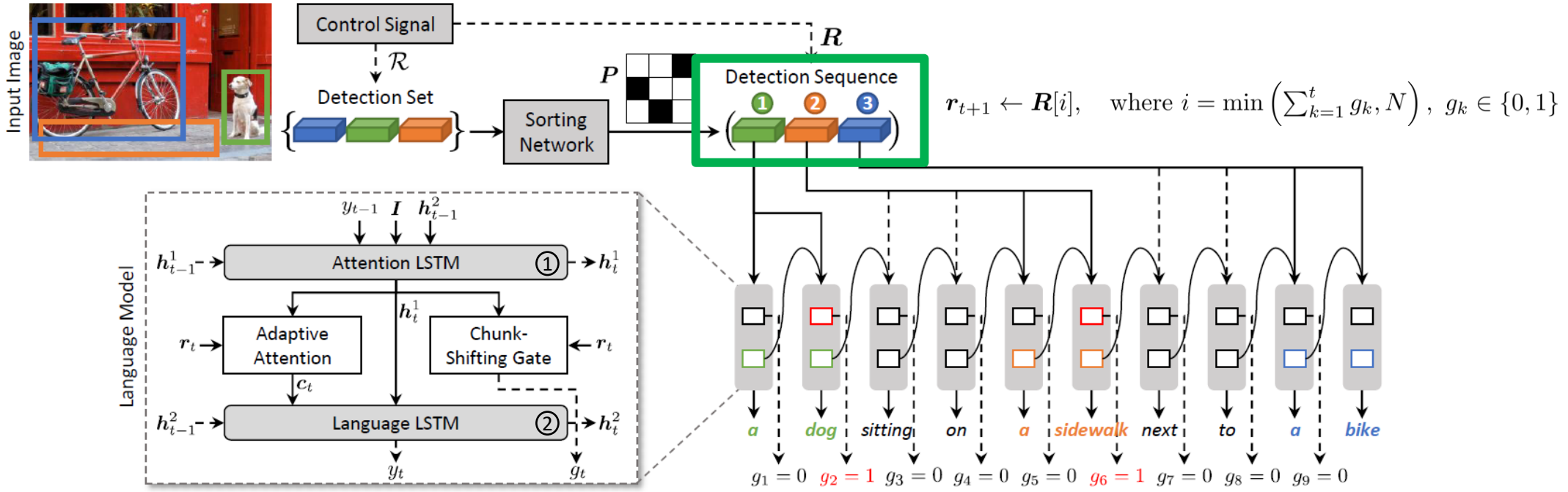
- Make an LSTM-based language model controllable via:
  - A sequence of regions (ordered)
  - A set (unordered)

Different captions for the same image!



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

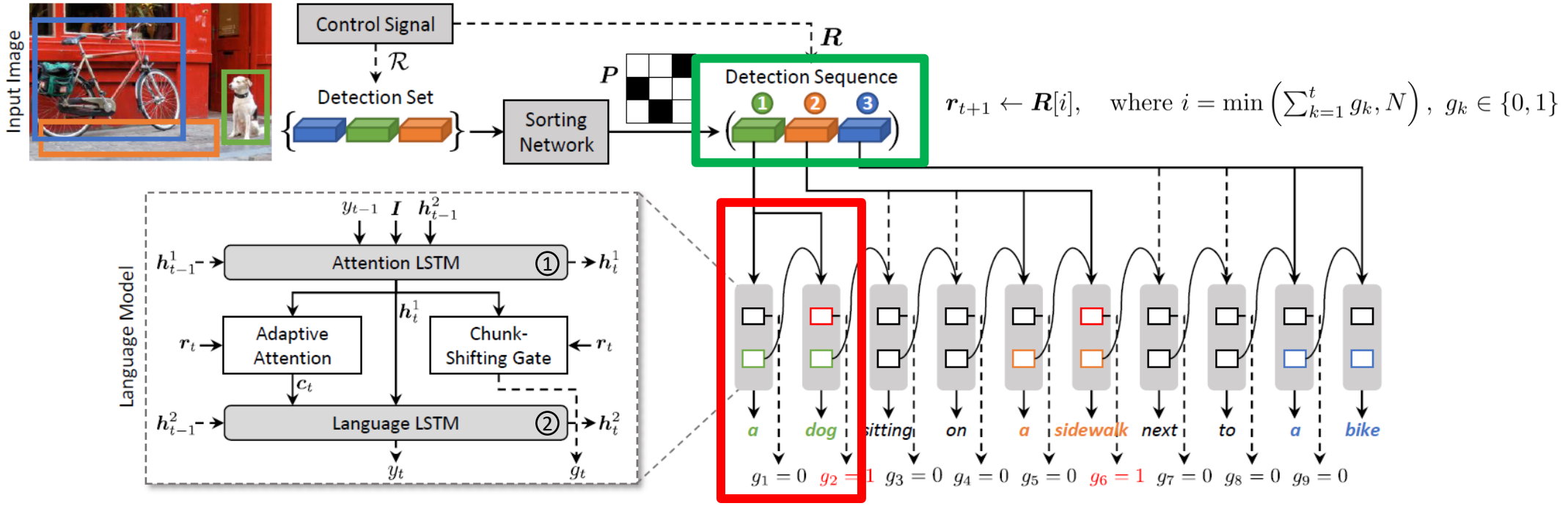
CONTROLLABLE IMAGE CAPTIONING



- Language models 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

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

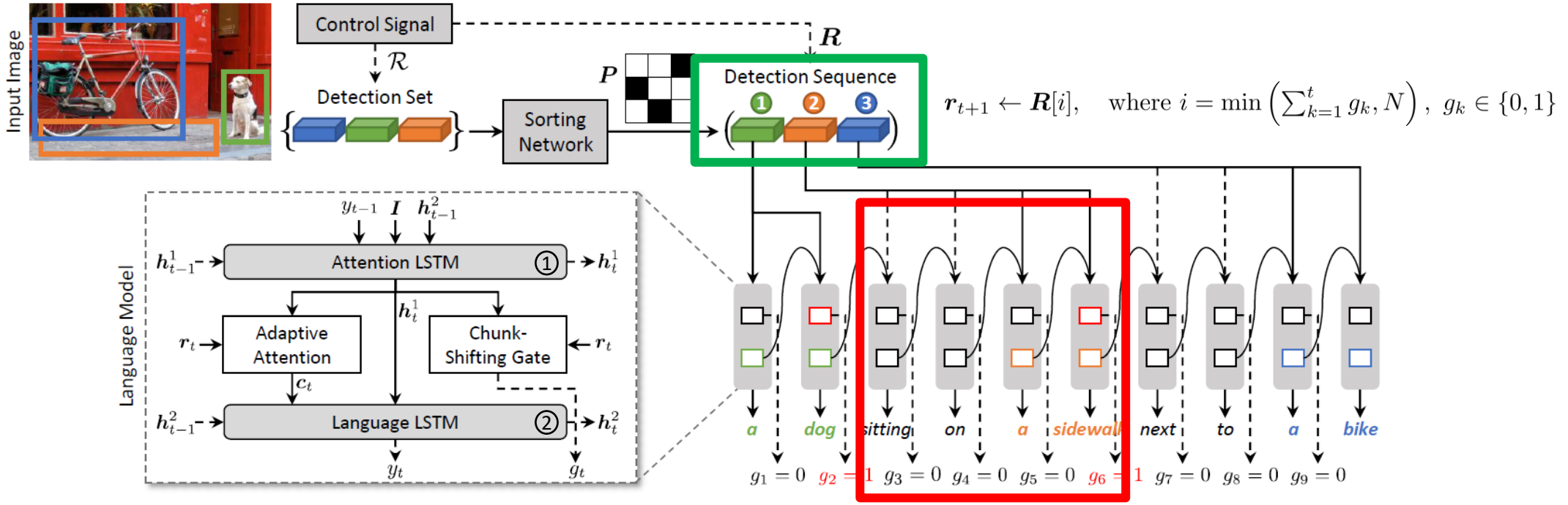
CONTROLLABLE IMAGE CAPTIONING



- Language models 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

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

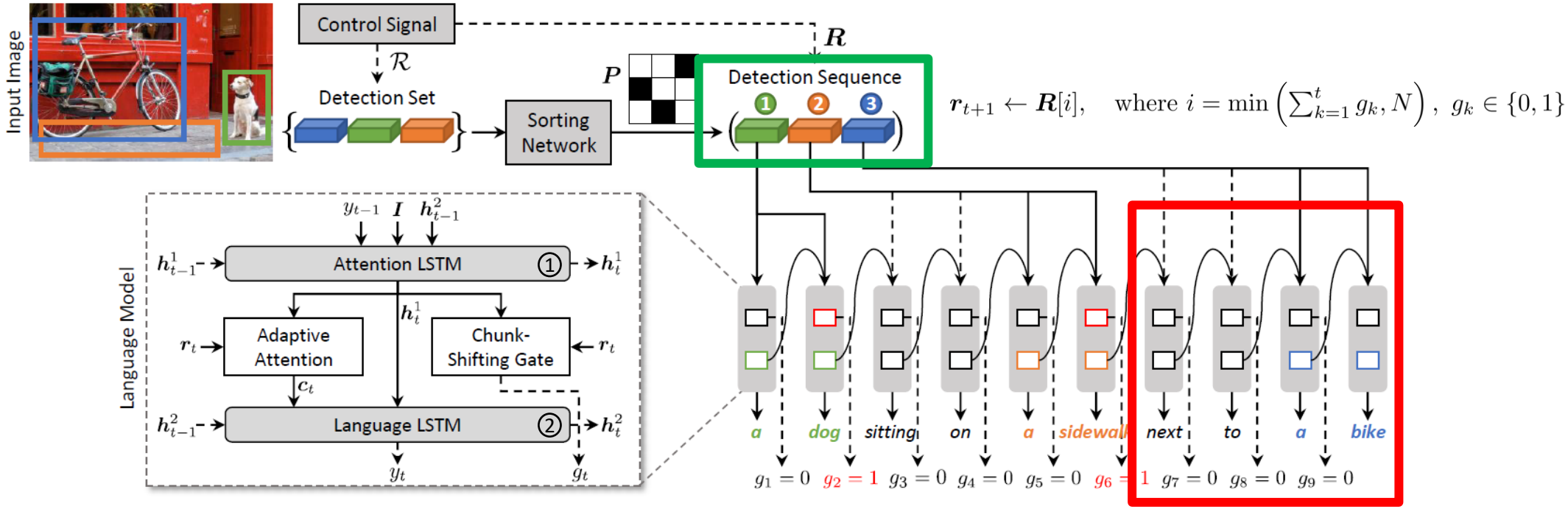
CONTROLLABLE IMAGE CAPTIONING



- Language models 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

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

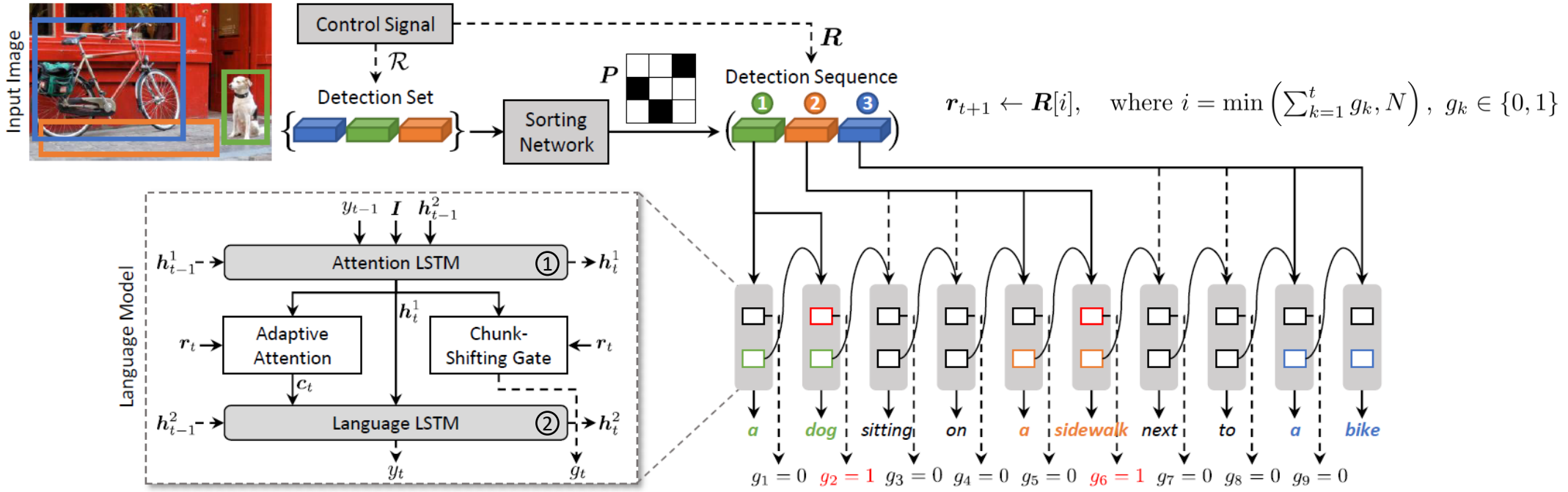
CONTROLLABLE IMAGE CAPTIONING



- Language models 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

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

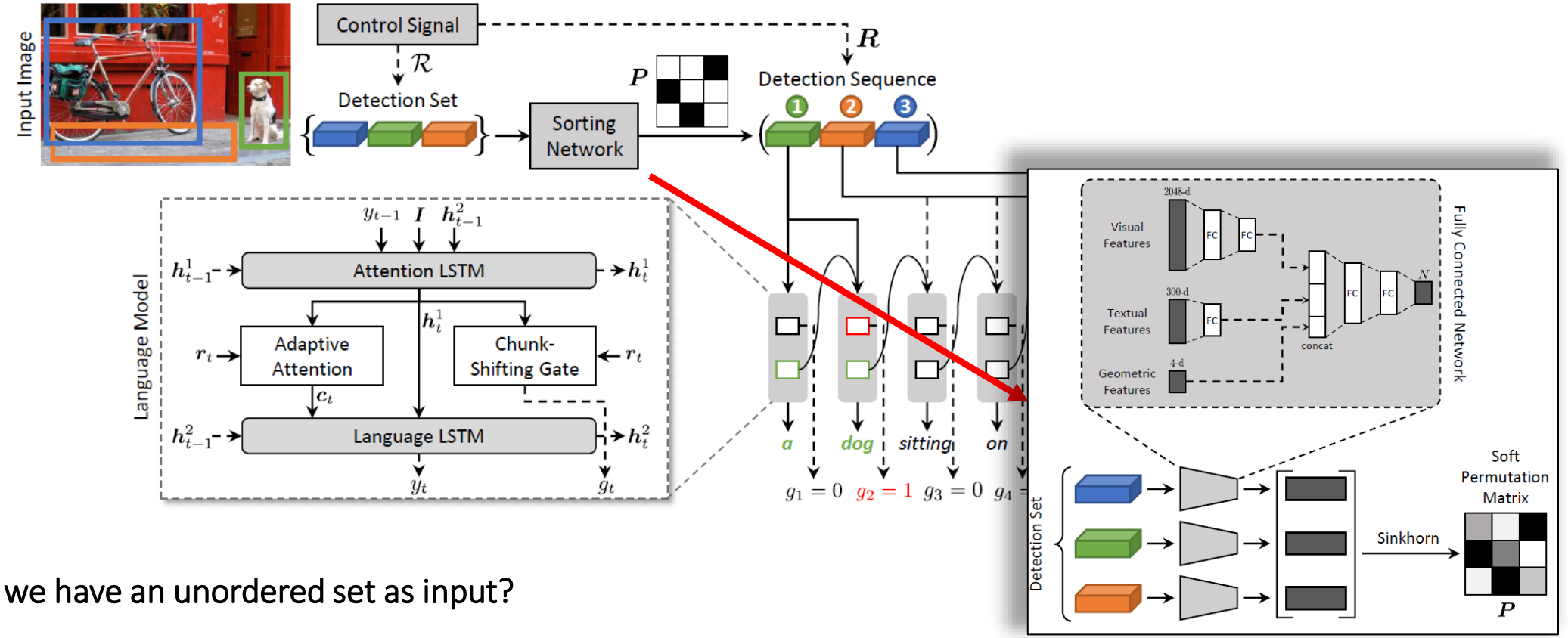
CONTROLLABLE IMAGE CAPTIONING



- Train on GT words and shifting gate values (obtained via NLP)
- … and since we are at it, finetune using Reinforcement Learning
  - CIDEr wrt GT caption (as usual)
  - Plus, use the alignment between the predicted and GT chunks as reward (Needleman-Wunsh algorithm)

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

CONTROLLABLE IMAGE CAPTIONING

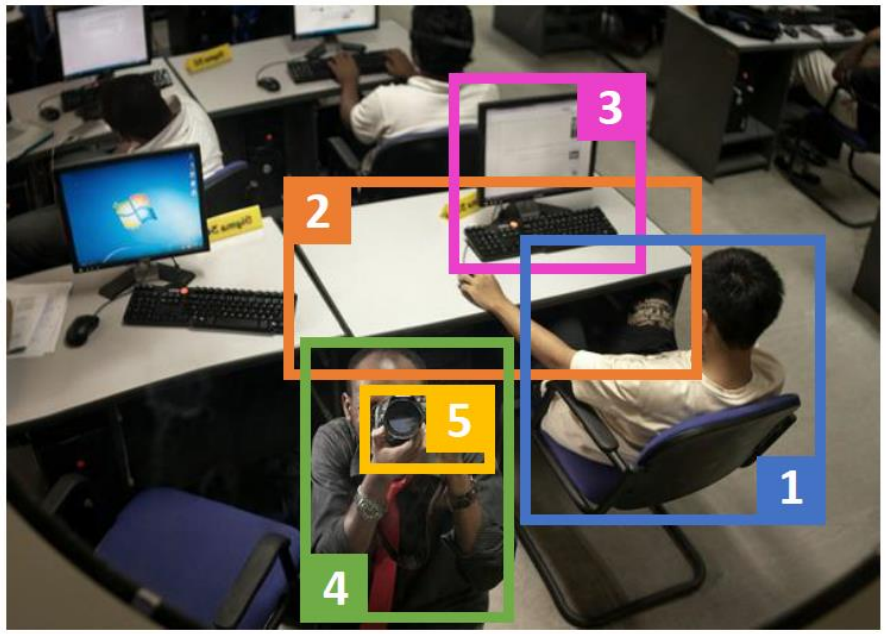


- What if we have an unordered set as input?
- We can learn a network to do the sorting! → SINKHORN NETWORK
  - Approximates a derivable permutation matrix
  - Train on real data, then use the Hungarian to get the true permutation matrix.

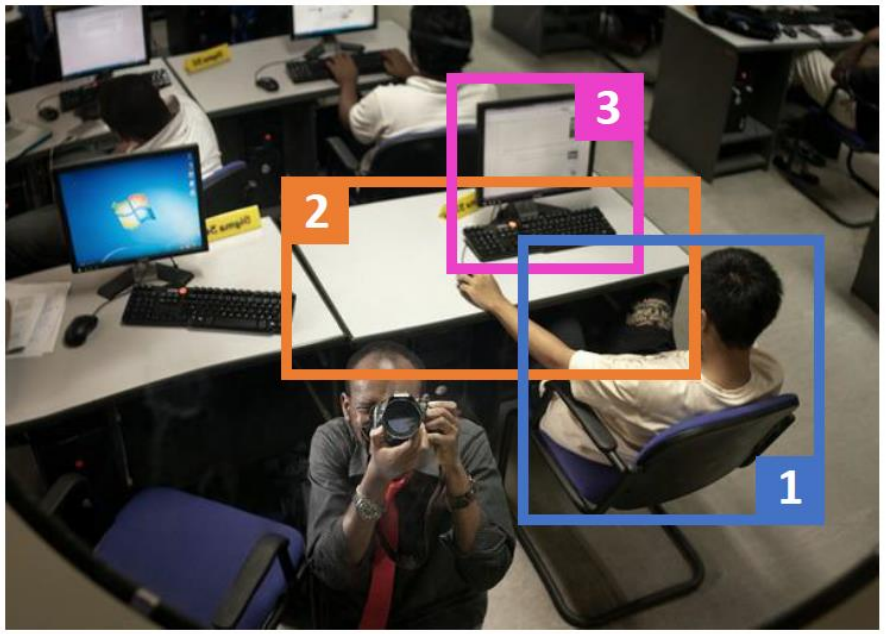
Cornia Marcella, Lorenzo Baraldi, and Rita Cucchiara. "Show, Control and Tell: A Framework for Generating Grounded and Controllable Captions." ***CVPR 2019.***

CONTROLLABLE IMAGE CAPTIONING

Results when Controlling with a sequence of regions



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

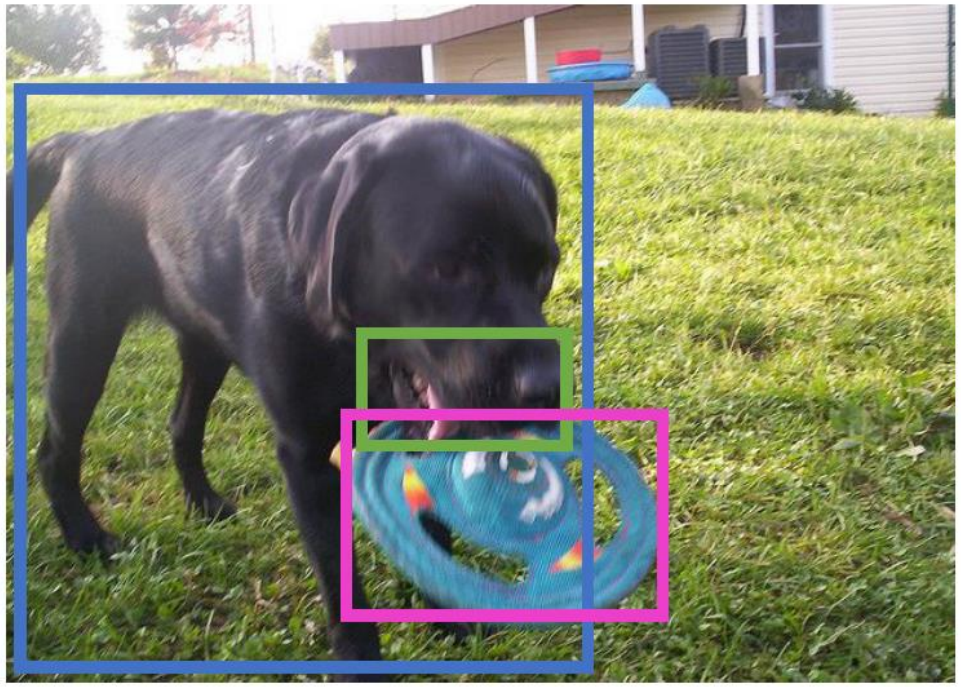


A man sitting at a desk with a computer.

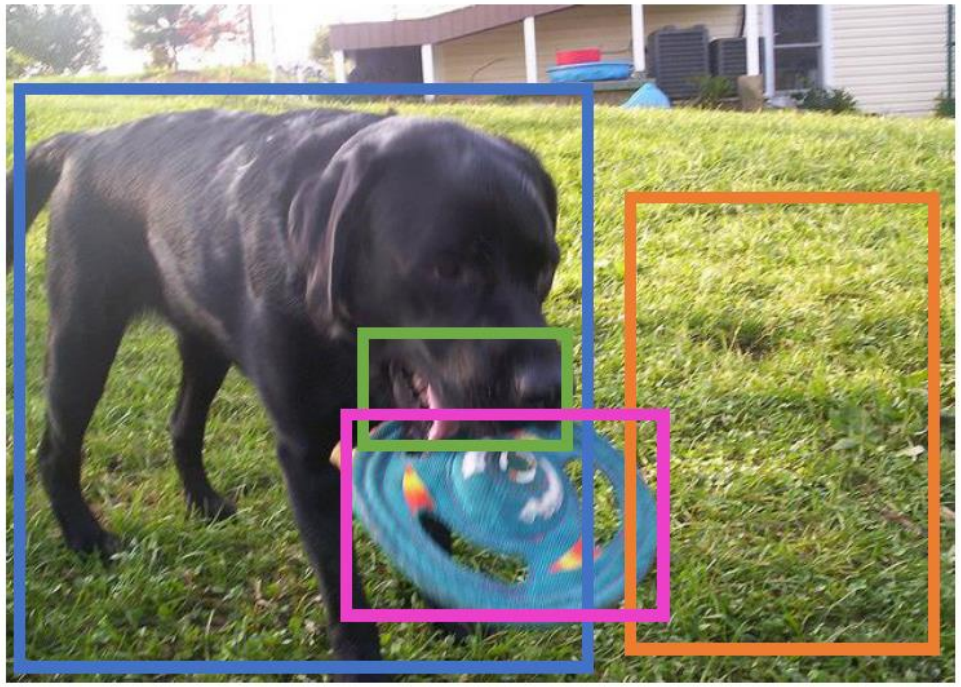
Cornia Marcella, Lorenzo Baraldi, and Rita Cucchiara. "Show, Control and Tell: A Framework for Generating Grounded and Controllable Captions." ***CVPR 2019.***

CONTROLLABLE IMAGE CAPTIONING

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.

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

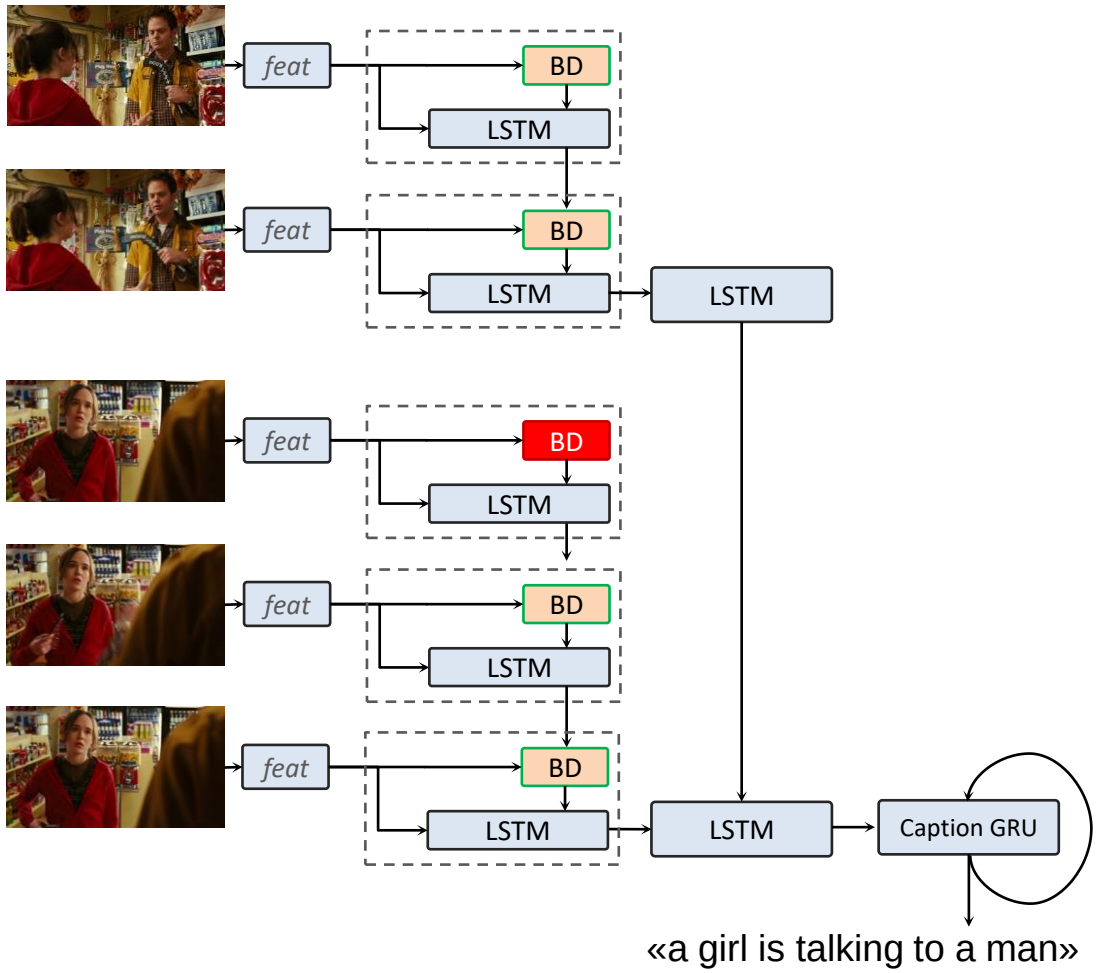
HIERARCHICAL LSTMS

Describing videos

- LSTMs do not show good learning capabilities on long sequences
- Plus, they do not deal with the **layered structure of videos**.

Our proposal

- A layered LSTM module which can **adaptively modify its structure** to input data.
- The result is a **variable length and adaptive encoding of the video**, whose length and granularity depends on the input video itself.

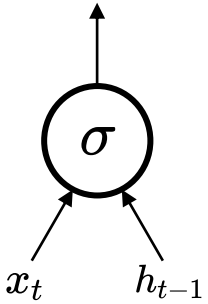


L. Baraldi, C. Grana, R. Cucchiara. "Hierarchical Boundary-Aware Neural Encoder for Video Captioning." ***CVPR 2017.***

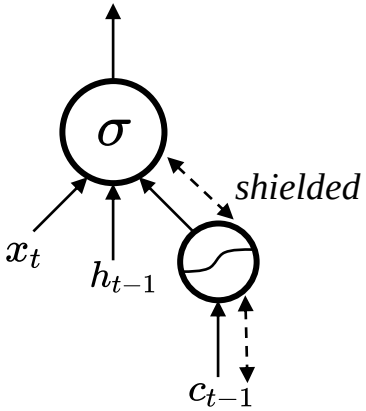
LSTMS WITH WORKING MEMORY CONNECTION

Latest work

- We redefine the LSTM cell, endowing it with a better connection between long and short memory.
- Solves learning issues of the peephole connection.
- Improves learning, especially on long sequences, on a variety of sequence modelling tasks.



$$\mathbf{g}_t = \tanh(\mathbf{W}_{gx}\mathbf{x}_t + \mathbf{W}_{gh}\mathbf{h}_{t-1} + \mathbf{b}_g)$$
$$\mathbf{i}_t = \sigma(\mathbf{W}_{ix}\mathbf{x}_t + \mathbf{W}_{ih}\mathbf{h}_{t-1} + \mathbf{b}_i)$$
$$\mathbf{f}_t = \sigma(\mathbf{W}_{fx}\mathbf{x}_t + \mathbf{W}_{fh}\mathbf{h}_{t-1} + \mathbf{b}_f)$$
$$\mathbf{c}_t = \mathbf{f}_t \odot \mathbf{c}_{t-1} + \mathbf{i}_t \odot \mathbf{g}_t$$
$$\mathbf{o}_t = \sigma(\mathbf{W}_{ox}\mathbf{x}_t + \mathbf{W}_{oh}\mathbf{h}_{t-1} + \mathbf{b}_o)$$
$$\mathbf{h}_t = \mathbf{o}_t \odot \tanh(\mathbf{c}_t)$$



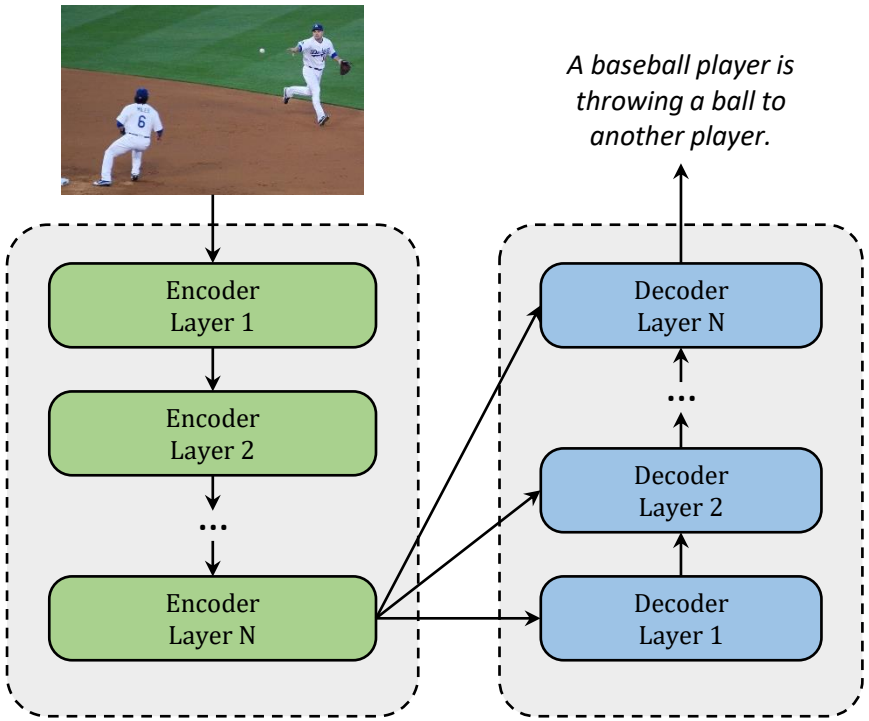
$$\mathbf{i}_t = \sigma(\mathbf{W}_{ix}\mathbf{x}_t + \mathbf{W}_{ih}\mathbf{h}_{t-1} + \tanh(\mathbf{W}_{ic}\mathbf{c}_{t-1}) + \mathbf{b}_i)$$
$$\mathbf{f}_t = \sigma(\mathbf{W}_{fx}\mathbf{x}_t + \mathbf{W}_{fh}\mathbf{h}_{t-1} + \tanh(\mathbf{W}_{fc}\mathbf{c}_{t-1}) + \mathbf{b}_f)$$
$$\mathbf{o}_t = \sigma(\mathbf{W}_{ox}\mathbf{x}_t + \mathbf{W}_{oh}\mathbf{h}_{t-1} + \tanh(\mathbf{W}_{oc}\mathbf{c}_t) + \mathbf{b}_o)$$

Model	Test Bit per Character (BPC)			
	Fixed # Params ($\sim 2.2M$)		Fixed # Hidden Units (512)	
	$T_{PTB} = 150$	$T_{PTB} = 300$	$T_{PTB} = 150$	$T_{PTB} = 300$
LSTM	1.334 ± 0.0006	1.343 ± 0.0004	1.386 ± 0.0005	1.395 ± 0.0005
LSTM-PH	1.339 ± 0.0048	1.343 ± 0.0009	1.383 ± 0.0004	1.394 ± 0.0005
LSTM-WM	$\mathbf{1.299 \pm 0.0005}$	$\mathbf{1.302 \pm 0.0008}$	$\mathbf{1.299 \pm 0.0005}$	$\mathbf{1.302 \pm 0.0008}$

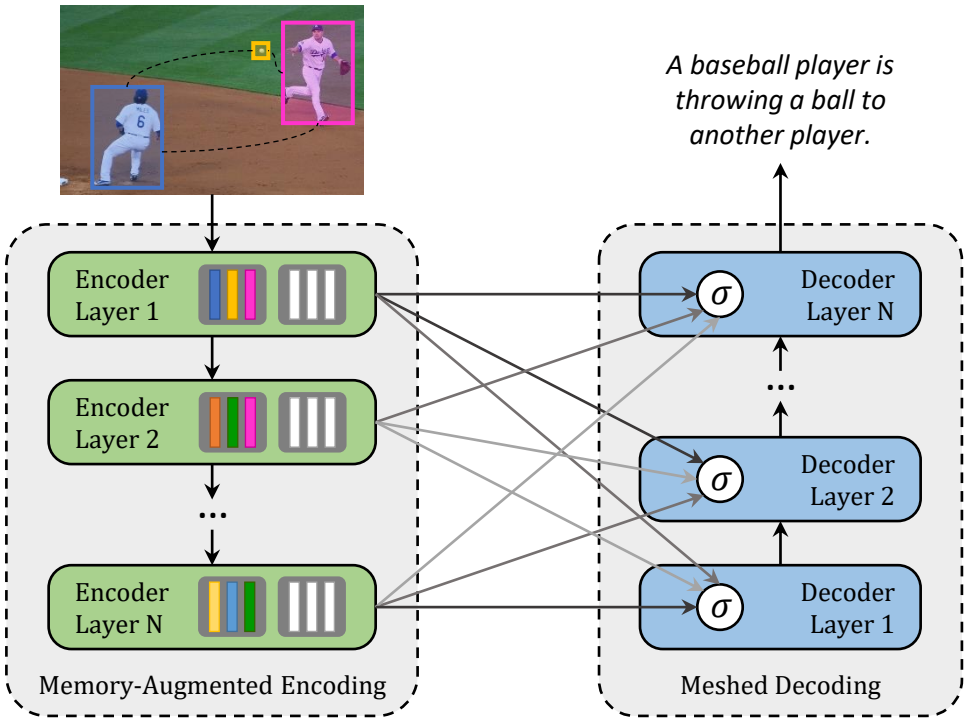
F. Landi, L. Baraldi, M. Cornia, R. Cucchiara. “Working Memory Connection for LSTM." ***To appear.***

Meshed-Memory Transformer

Original Transformer



M^2 Transformer



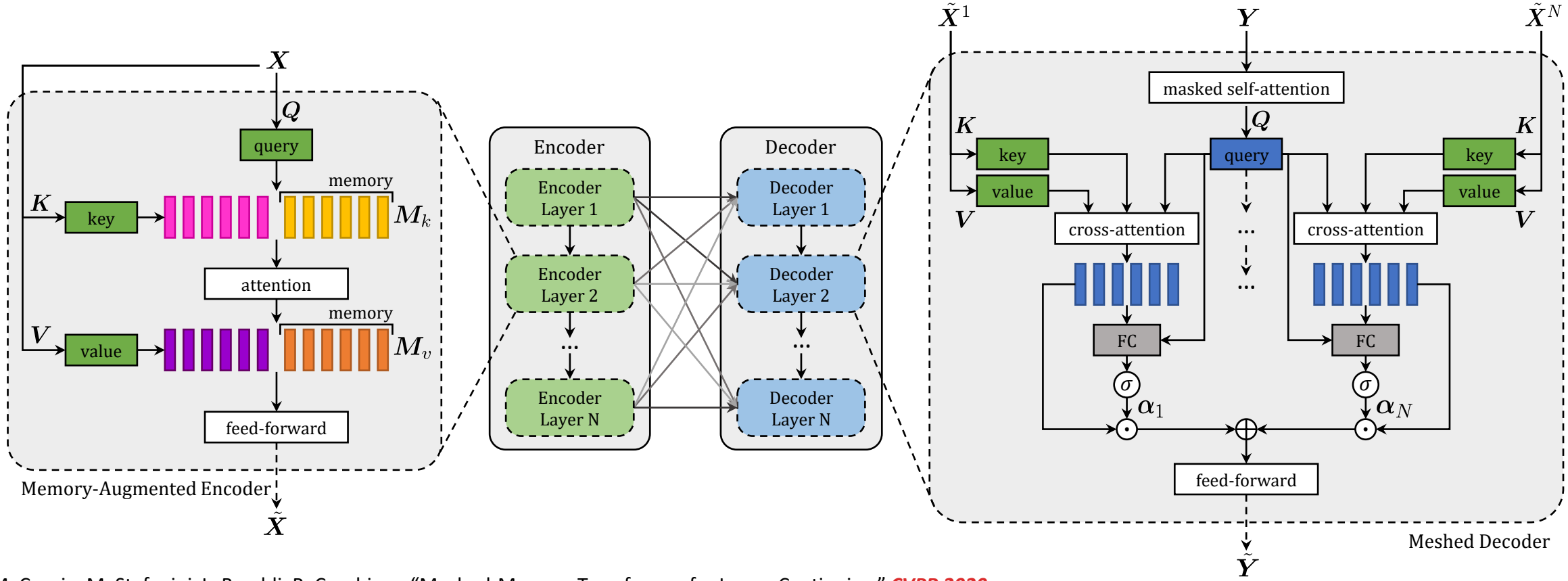
Relationships between image regions are modeled via **persistent memory vectors**.

Encoder and decoder layers are connected in **a mesh-like structure**.

M. Cornia, M. Stefanini, L. Baraldi, R. Cucchiara. "Meshed-Memory Transformer for Image Captioning." ***CVPR 2020***

Meshed-Memory Transformer

- In our encoder, the set of keys and values is extended with learnable vectors that can encode **a priori information**.
- A mesh connectivity is operated through **a learnable gating mechanism** which modulates the contribution of each encoder layer during cross attention.



M. Cornia, M. Stefanini, L. Baraldi, R. Cucchiara. "Meshed-Memory Transformer for Image Captioning." ***CVPR 2020***

Research direction – Captioning Novel Objects

- Final goal: describing *everything* even if it is not in the training set.
- **Joint collaboration with NVAITC**:
  - Extension to multi-GPU and multi-node training
  - Optimization of data loading and processing, mixed-precision training
  - (soon to come) migration to Marconi-100
- Next research steps: extension to the description of novel objects, via self-supervised training.

"A cat looking at its reflection in a mirror."

"A little girl sitting in a shopping cart eating a hot dog."

"A plate of breakfast food with eggs and toast."

"A green truck parked next to a pile of hay."

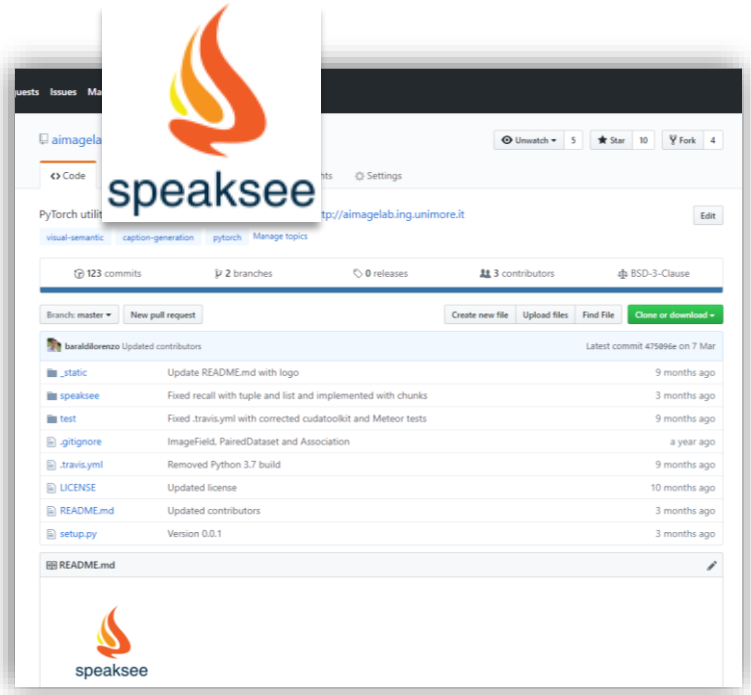
M. Cornia, M. Stefanini, L. Baraldi, R. Cucchiara. "M^2: Meshed-Memory Transformer for Image Captioning." ***CVPR 2020***

↓ Get your hands dirty

Speaksee: **our library for visual-semantic tasks** (captioning, retrieval, …) built on top of PyTorch

Data-loading, Models, Evaluation

https://github.com/aimagelab/speaksee



CADL WORKSHOP AT ICPR 2020

"**International Workshop on Computational Aspects of Deep Learning**"

Workshop organized with NVAITC, at ICPR 2020 (January 2021)

- Discussing innovations in the field of computational optimization
- Also, a place for the NVAITC community to meet and enlarge itself!

Topics:

- Design of innovative architectures and operators for data-intensive scenarios
- Video understanding and spatio-temporal feature extraction
- Distributed reinforcement learning algorithms
- Applications of large-scale pre-training techniques
- Distributed training approaches and architectures
- HPC and massively parallel architectures in Deep Learning

See the webpage for CfP, dates and organization:
www.cadl.it

VISION AND LANGUAGE NAVIGATION

Dynamic Response Map

Agent position (and next action)

Instruction:

Walk up the stairs.
Turn right at the top of the stairs and walk along the red ropes.
Walk through the open doorway straight ahead along the red carpet.
Walk through that hallway into the room with couches and a marble coffee table.

VISION AND LANGUAGE NAVIGATION

Dynamic Response Map

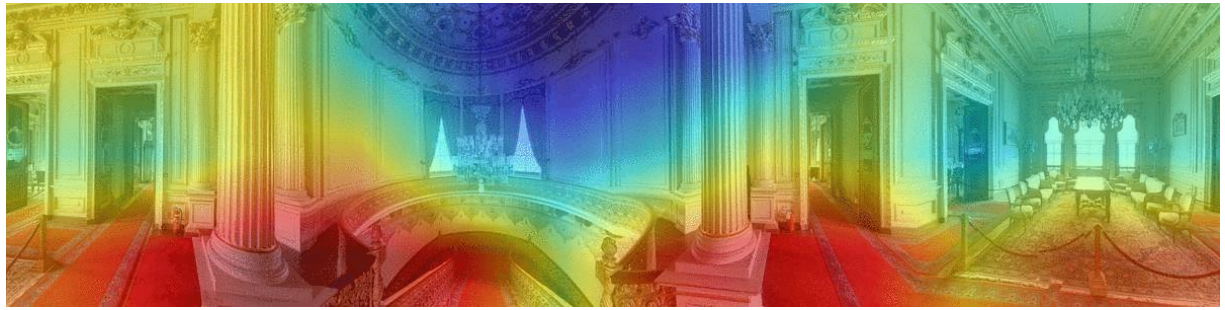
Agent position (and next action)

Instruction:

Walk up the stairs.
Turn right at the top of the stairs and walk along the red ropes.
Walk through the open doorway straight ahead along the red carpet.
Walk through that hallway into the room with couches and a marble coffee table.

VISION AND LANGUAGE NAVIGATION

Dynamic Response Map



Agent position (and next action)

Instruction:

Walk up the stairs.
Turn right at the top of the stairs and walk along the red ropes.
Walk through the open doorway straight ahead along the red carpet.
Walk through that hallway into the room with couches and a marble coffee table.

VISION AND LANGUAGE NAVIGATION

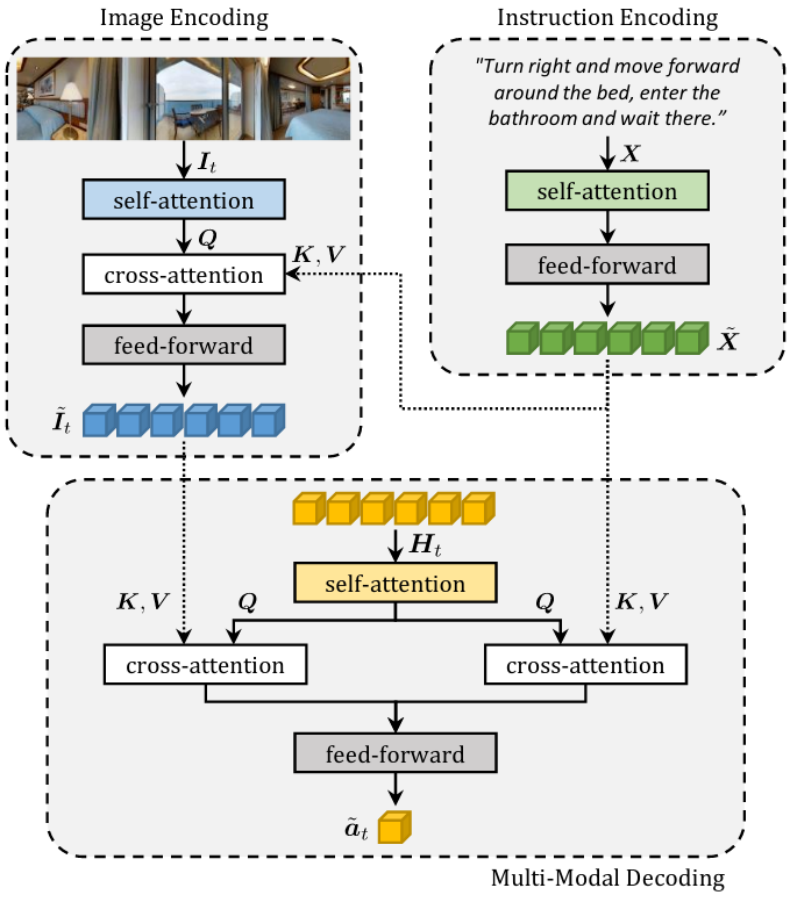
Dynamic Response Map

Agent position (and next action)

Instruction:

Walk up the stairs.
Turn right at the top of the stairs and walk along the red ropes.
Walk through the open doorway straight ahead along the red carpet.
Walk through that hallway into the room with couches and a marble coffee table.

PTA



Locomotion is guided by visual perception.
Not only does it depend on perception, but perception depends on locomotion.

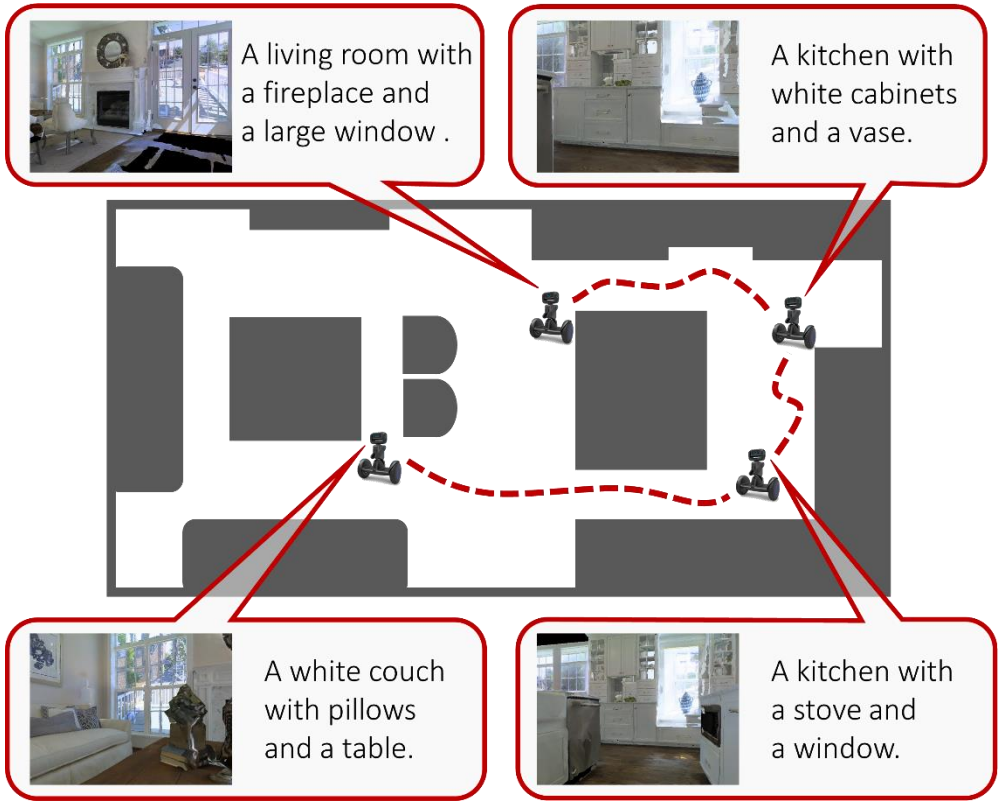
The Ecological Approach to Visual Perception
James J. Gibson

PTA is the first Transformer-like architecture merging three different modalities: text, vision, and action.

With a focus on low-level motion: a more realistic setup; closer to real-world applications

F. Landi, L. Baraldi, M. Corsini, R. Cucchiara. "Embodied Vision-and-Language Navigation with Dynamic Convolutional Filters." ***BMVC 2019***
F. Landi, M. Cornia, L. Baraldi, M. Corsini, R. Cucchiara. "Perceive, Transform, and Act: Multi-Modal Attention Networks for Vision-and-Language Navigation." ***To appear***

SELF-SUPERVISED NAVIGATION AND RECOUNTING



Explore and Explain – a new setting for embodied AI

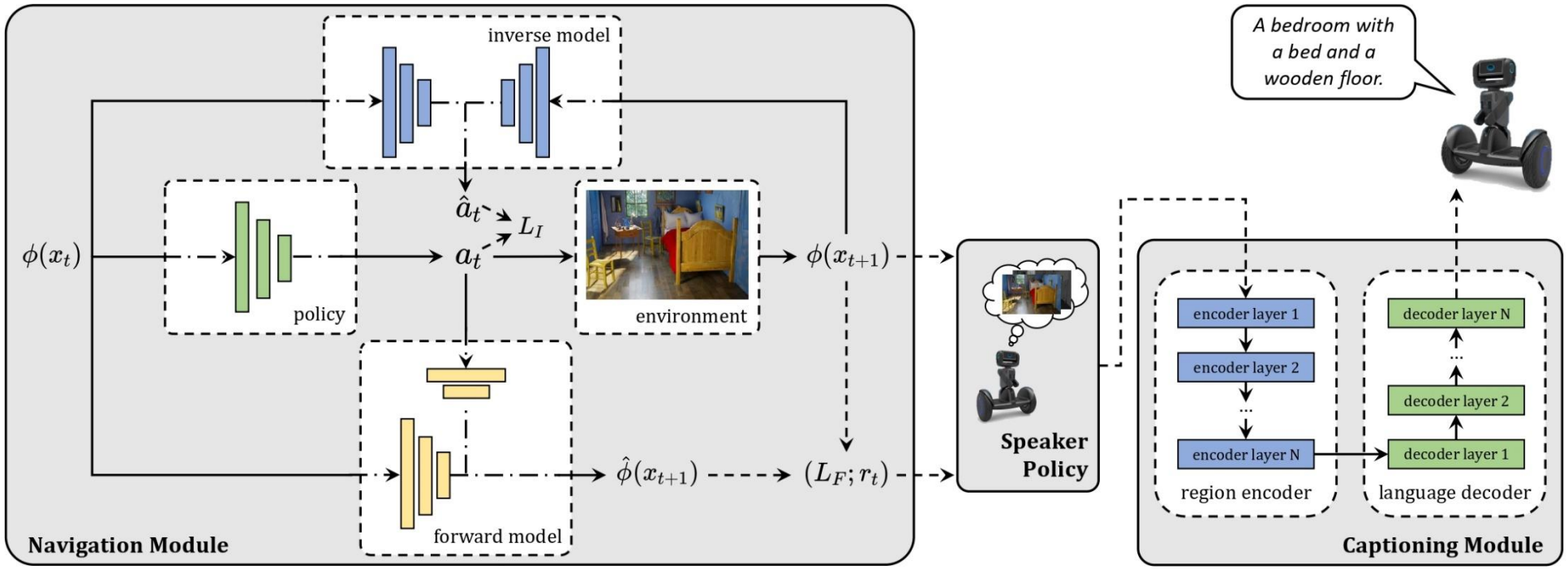
The agent must jointly deal with two different tasks:

- Exploration of unseen environments
- Description of the most relevant visual features

R. Bigazzi, F. Landi, M. Cornia, S. Cascianelli, L. Baraldi, R. Cucchiara. “Explore and Explain: Self-supervised Navigation and Recounting.” ***ICPR 2020***

OUR APPROACH – EX^2

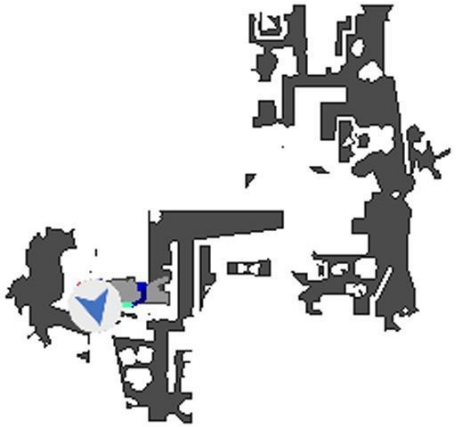
Curiosity-driven exploration module + **Transformer-based** captioner



R. Bigazzi, F. Landi, M. Cornia, S. Cascianelli, L. Baraldi, R. Cucchiara. "Explore and Explain: Self-supervised Navigation and Recounting." ***ICPR 2020***

QUALITATIVE RESULTS - EXPLORATION

Random exploration



Vanilla curiosity

EX^2

Our navigation module enriches traditional curiosity-based navigation with a **penalty** on repeated actions

R. Bigazzi, F. Landi, M. Cornia, S. Cascianelli, L. Baraldi, R. Cucchiara. “Explore and Explain: Self-supervised Navigation and Recounting.” ***ICPR 2020***

QUALITATIVE RESULTS - EXPLAINABILITY

A living room with a couch and a television.

A bathroom with a bath tub and a window.

A bedroom with a bed and a painting on the wall.

A living room with a fireplace and a table.

A kitchen with white cabinets and a glass door.

EX^2 can **identify** and **describe** relevant object during exploration

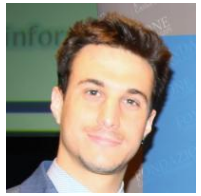
- An important means to provide some **explainability** for the actions performed by the agent

R. Bigazzi, F. Landi, M. Cornia, S. Cascianelli, L. Baraldi, R. Cucchiara. "Explore and Explain: Self-supervised Navigation and Recounting." ***ICPR 2020***

Thanks!
Questions?


Matteo Stefanini


Matteo Tomei


Federico Landi


Roberto Bigazzi


Marcella Cornia


Silvia Cascianelli


Lorenzo Baraldi


Rita Cucchiara

UNIMORE
UNIVERSITÀ DEGLI STUDI DI MODENA E REGGIO EMILIA

AImage Lab

Artificial Intelligence
Research and Innovation Center (AIRI)