Embodied Vision-and-Language Navigation with Dynamic Convolutional Filters

Federico Landi, Lorenzo Baraldi, Massimiliano Corsini, Rita Cucchiara
University of Modena and Reggio Emilia

1. Motivations

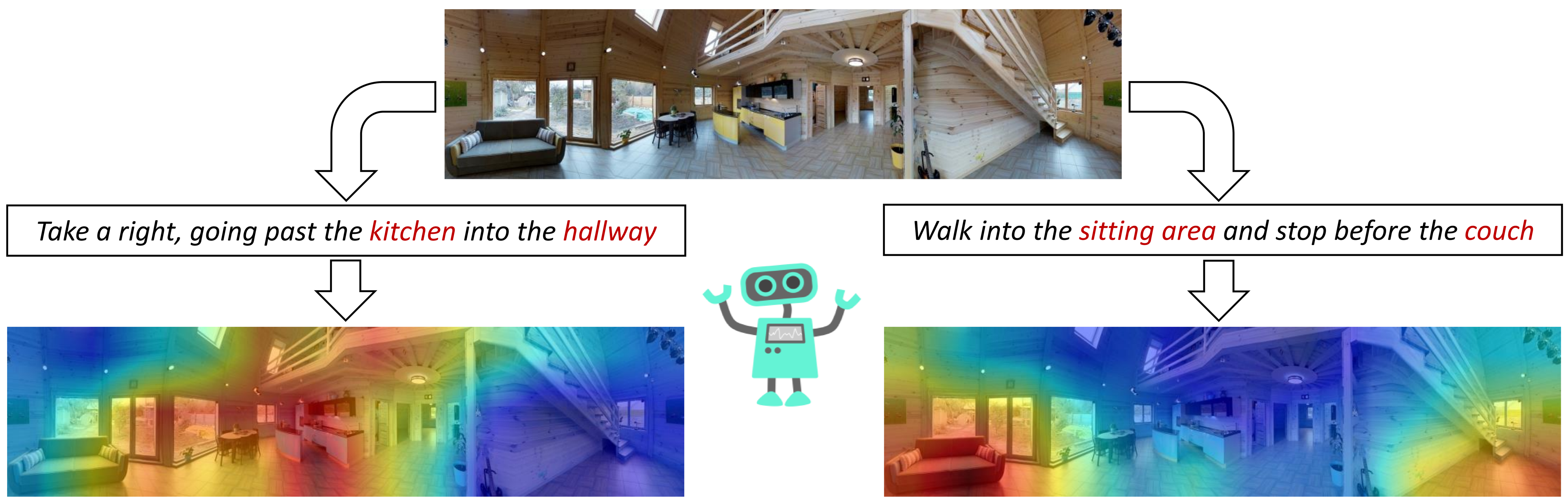
In Vision-and-Language Navigation (VLN), an embodied agent needs to reach a target destination with the only guidance of a natural language instruction. We exploit dynamic convolutional filters to ground the lingual description into the visual observation in an elegant and efficient way.

Our contributions:

- New encoder-decoder architecture which employs dynamic convolutional filters;
- Novel categorization for VLN: we distinguish between *low-level* and *high-level* actions methods;
- State-of-Art results for low-level VLN.

2. Dynamic Convolutional Filters

Rather than learning a fixed set of convolutional filters, we learn to generate them depending on the natural language specification.



Dynamic convolutional filters act as specialized and flexible feature extractors.

3. Architecture

For each navigation episode, the agent receives a natural language instruction that is encoded into X. At each time step t, the agent observes I_t from the surroundings. The goal is to decode the atomic action for the current step.

Encoder-Decoder Attention:

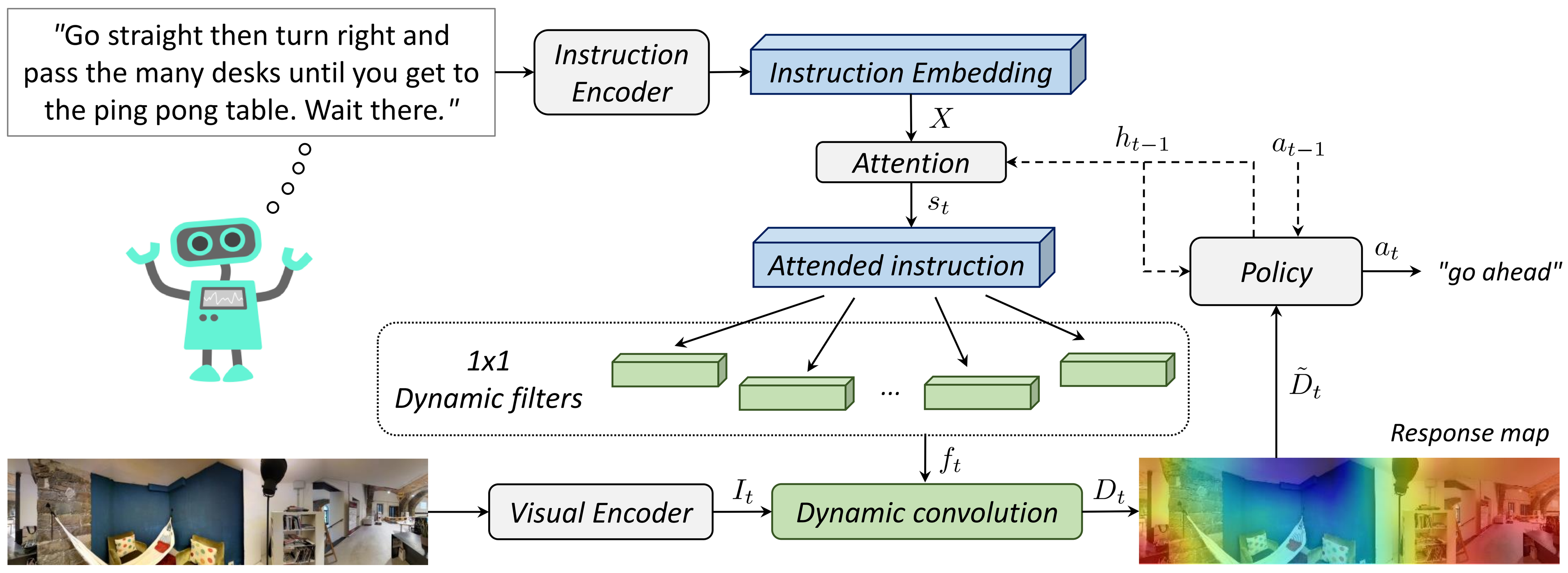
$$q_t = W_q h_{t-1} + b_q \qquad K = W_k X + b_k$$
$$\alpha_t = q_t K^T / \sqrt{d_{att}} \qquad s_t = \mathrm{softmax}(\alpha_t) X$$

Dynamic Convolution:

$$f_t = \ell_2[\tanh(W_f s_t + b_f)]$$
$$D_t = f_t * I_t$$

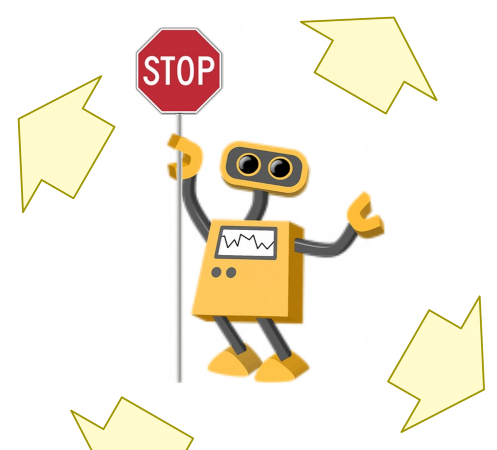
Action Decoding:

$$h_t = \mathrm{LSTM}([\tilde{D}_t, a_{t-1}], h_{t-1})$$
$$p_t = \mathrm{softmax}(W_a h_t + b_a)$$



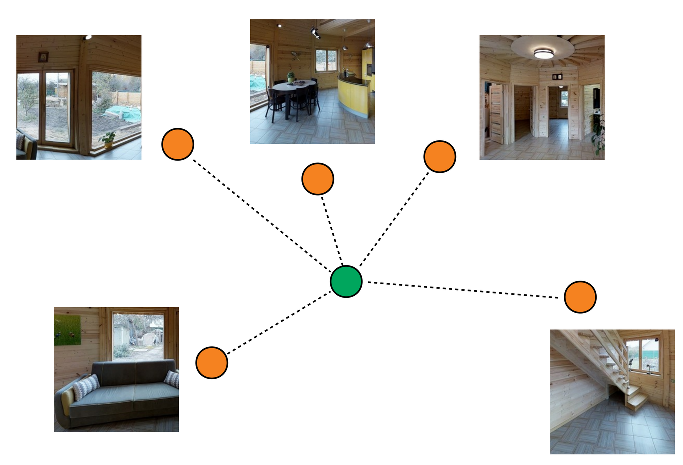
4. Low-level and High-level Methods

Low-level Action Space
Simulates continuous control



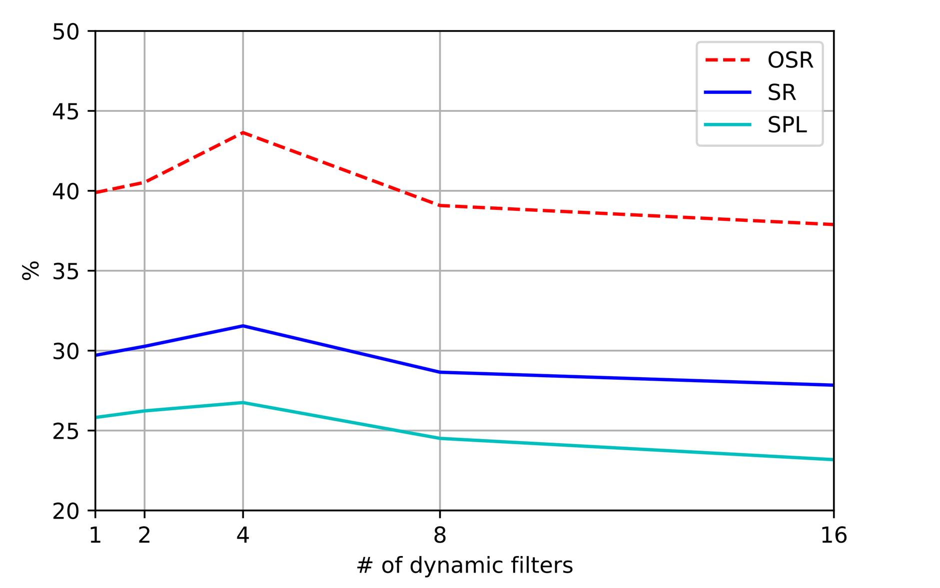
This work!

High-level Action Space
Path selection on a discrete graph



6. Number of Dynamic Filters

# of filters	Val-Unseen NE ↓	Val-Unseen SR ↑
1	6.79	29.7
2	6.77	30.3
4	**6.65**	**31.6**
8	7.19	28.7
16	7.03	27.8



One filter is enough, best setup with four response maps.

5. Qualitative Results

Next action: L *left* R *right* F *forward* E *end episode*

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.*

7. Comparison with State-of-the-Art

Low-level (ours): possible actions are *move forward*, *turn left 30°*, *turn right 30°*, *raise elevation*, *lower elevation*, and *end episode*.

	Validation-Seen				Validation-Unseen				Test (Unseen)			
Low-level Actions Methods	NE ↓	SR ↑	OSR ↑	SPL ↑	NE ↓	SR ↑	OSR ↑	SPL ↑	NE ↓	SR ↑	OSR ↑	SPL ↑
Random	9.45	0.16	0.21	-	9.23	0.16	0.22	-	9.77	0.13	0.18	0.12
Student-forcing [1]	6.01	0.39	0.53	-	7.81	0.22	0.28	-	7.85	0.20	0.27	0.18
RPA [2]	5.56	0.43	0.53	-	7.65	0.25	0.32	-	7.53	0.25	0.33	0.23
Ours	4.68	0.53	0.66	0.46	6.65	0.32	0.44	0.27	7.14	0.31	0.42	0.27
Ours w/ data augmentation	**3.96**	**0.58**	**0.73**	**0.51**	**6.52**	**0.34**	**0.43**	**0.29**	**6.55**	**0.35**	**0.45**	**0.31**

High-level: the agent selects the destination from a set of adjacent nodes. Moves are made thanks to global information from the simulator.

	Validation-Seen				Validation-Unseen				Test (Unseen)			
High-level Actions Methods	NE ↓	SR ↑	OSR ↑	SPL ↑	NE ↓	SR ↑	OSR ↑	SPL ↑	NE ↓	SR ↑	OSR ↑	SPL ↑
Speaker-Follower [3]	3.36	0.66	0.74	-	6.62	0.36	0.45	-	6.62	0.35	0.44	0.28
Self-Monitoring [4]	**3.22**	0.67	**0.78**	0.58	5.52	0.45	0.56	0.32	5.99	0.43	0.55	0.32
Regretful [5]	3.23	**0.69**	0.77	**0.63**	**5.32**	**0.50**	**0.59**	**0.41**	**5.69**	**0.48**	**0.56**	**0.40**

NE: Navigation Error (**m**); **OSR**: Oracle Success Rate;
SR: Success Rate; **SPL**: Success rate normalized on Path Length.

8. References

[1] Anderson *et al.* Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In *CVPR*, 2018.
[2] Wang *et al.* Look before you leap: Bridging model-free and model-based reinforcement learning for planned-ahead VLN. In *ECCV*, 2018.
[3] Fried *et al.* Speaker-follower models for vision-and-language navigation. In *NeurIPS*, 2018.
[4] Ma *et al.* Self-Monitoring Navigation Agent via Auxiliary Progress Estimation. In *ICLR*, 2019.
[5] Ma *et al.* The Regretful Agent: Heuristic-Aided Navigation through Progress Estimation. In *CVPR*, 2019.

BMVC CARDIFF 2019

Correspondence to
federico.landi@unimore.it