Speaker-Follower Models for Vision-and-Language Navigation

Daniel Fried*[1] Ronghang Hu*[1] Volkan Cirik*[2] Anna Rohrbach[1] Jacob Andreas[1]
Louis-Philippe Morency[2] Taylor Berg-Kirkpatrick[2] Kate Saenko[3] Dan Klein**[1] Trevor Darrell**[1]

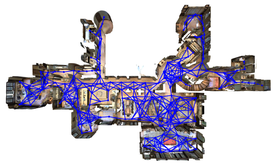
[1]University of California, Berkeley [2]Carnegie Mellon University [3]Boston University

Follower and Speaker Models for VLN

The Vision-and-Language Navigation (VLN) task [1]

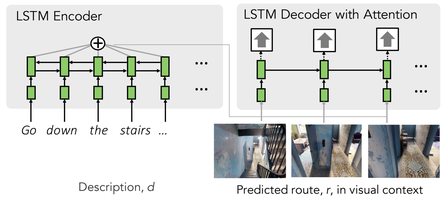
- Given a *textual instruction*, take actions to navigate to the described target location.

> *"Walk between the columns and make a sharp turn right. Walk down the steps and stop on the landing."*



➢ Follower model for instruction following

$P_F(r|d)$



The VLN task requires learning a *follower model* to predict routes from instructions.

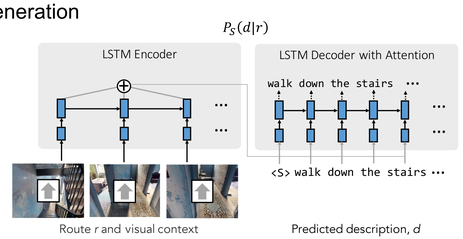
- We implement our follower model as an attentional sequence-to-sequence model that translates the instruction text into a navigation route.
- At each step, the decoder attends to parts of the instruction text and visual evidences, and predicts an action.

➢ Speaker model for instruction generation

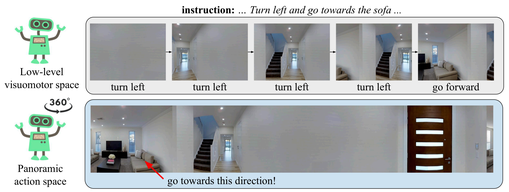
$P_S(d|r)$

In this work, we introduce a new *speaker model* to also learn the instruction generation process.

- We implement our speaker model in a symmetric way, translating visual inputs and actions into instructions.



➢ Panoramic action space

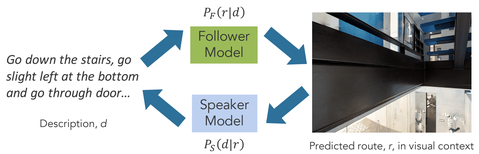


Our agent acts directly on a panoramic representation that matches the granularity of the VLN task.

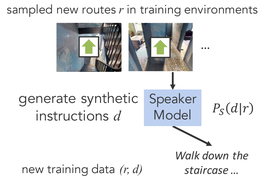
- Our panoramic representation allows high-level actions, converted from low-level visuomotor control.

Speaker-Driven Data Augmentation and Pragmatic Inference

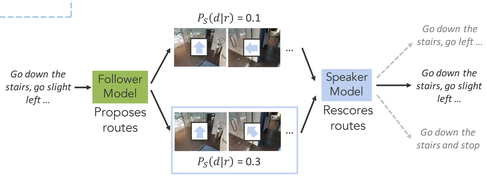
➢ Our speaker model helps the follower at both training time for *data augmentation* and test time for *pragmatic inference*



Data augmentation (during training): generate synthetic instructions on new routes as additional training data for the follower



Pragmatic inference (at test time): rank possible routes by measuring how likely the instruction can be generated from each route



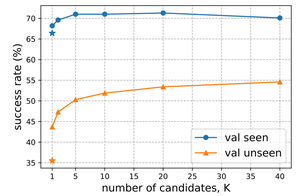
Ablation Study

➢ Dataset and evaluation metric

Dataset: Room-to-Room (R2R) dataset [1] with real scenes and human-written instructions.
Evaluation metric: NE is navigation error. SR and OSR are success rate and oracle success rate (%) respectively. Success is defined as stopping within 3 meters of the goal location.

➢ Ablations: all components in our model are helpful and complementary.

#	Data Augmentation	Pragmatic Inference	Panoramic Space	Validation-Seen			Validation-Unseen		
				NE↓	SR↑	OSR↑	NE↓	SR↑	OSR↑
1				6.08	40.3	51.6	7.90	19.9	26.1
2	✓			5.05	46.8	59.9	7.30	24.6	33.2
3		✓		5.23	51.5	60.8	6.62	34.5	43.1
4			✓	4.86	52.1	63.3	7.07	31.2	41.3
5	✓	✓		4.28	57.2	63.9	5.75	39.3	47.0
6	✓		✓	3.36	66.4	73.8	6.62	35.5	45.0
7		✓	✓	3.88	63.3	71.0	5.24	49.5	63.4
8	✓	✓	✓	**3.08**	**70.1**	**78.3**	**4.83**	**54.6**	**65.2**



Improvements obtained both from
- the beam search procedure (stars vs. the circle/triangle points at K = 1)
- more candidates in pragmatic inference (larger values of K)

Comparison to Previous Work and Qualitative Results

➢ Comparison: our model shows large improvements over previous work on both seen and unseen environments.

Method	Validation-Seen			Validation-Unseen			Test (unseen)			
	NE↓	SR↑	OSR↑	NE↓	SR↑	OSR↑	NE↓	SR↑	OSR↑	TL↓
Random	9.45	15.9	21.4	9.23	16.3	22	9.77	13.2	18.3	9.89
Student-forcing [1]	6.01	38.6	52.9	7.81	21.8	28.4	7.85	20.4	26.6	8.13
RPA [2]	5.56	42.9	52.6	7.65	24.6	31.8	7.53	25.3	32.5	9.15
ours	**3.08**	**70.1**	**78.3**	**4.83**	**54.6**	**65.2**	**4.87**	**53.5**	63.9	11.63
ours (challenge participation)*	–	–	–	–	–	–	**4.87**	**53.5**	**96.0**	1257.38
Human	–	–	–	–	–	–	1.61	86.4	90.2	11.9

Trajectory length (TL) on the test set is reported for completeness.
*: When submitting to the VLN Challenge, we modified our beam search to maintain physical plausibility and to comply with the challenge guidelines. The resulting trajectory has higher oracle success rate while being very long (details in supplemental).

Attention visualization

instruction: *"Walk up stairs. Turn left and walk to the double doors by the living room."*

Comparison between greedy and pragmatic inference

instruction: *"Walk past hall table. Walk into bedroom. Make left at table clock. Wait at bathroom door threshold."*

greedy inference without pragmatics: confused about "walk into bedroom", and entered a wrong room without a table clock

pragmatic inference: entered the correct bedroom on the right, where it could see a "table clock"

Navigation steps of the panoramic agent. The red arrow shows the direction chosen by the agent to go next.

References

[1] Anderson, P., Wu, Q., Teney, D., Bruce, J., Johnson, M., Su¨nderhauf, N., Reid, I., Gould, S., Hengel, A.v.d.: Vision-and-language navigation: Interpreting visually- grounded navigation instructions in real environments. In CVPR (2018)
[2] Wang, X., Xiong, W., Wang, H., Wang, W.Y.: Look before you leap: Bridging model-free and model-based reinforcement learning for planned-ahead vision-and- language navigation. In ECCV (2018)