Reinforced Cross-Modal Matching and Self-Supervised Imitation Learning for Vision-Language Navigation

Xin Wang, Qiuyuan Huang, Asli Celikyilmaz, Jianfeng Gao, Dinghan Shen, Yuan-Fang Wang, William Yang Wang and Lei Zhang

CVPR 2019

Best Student Paper Award: 1/5160=0.02%

于家硕
19210240064

Vision-and-Language Navigation

- Natural language instruction
- First person camera view
- Multimodal machine learning

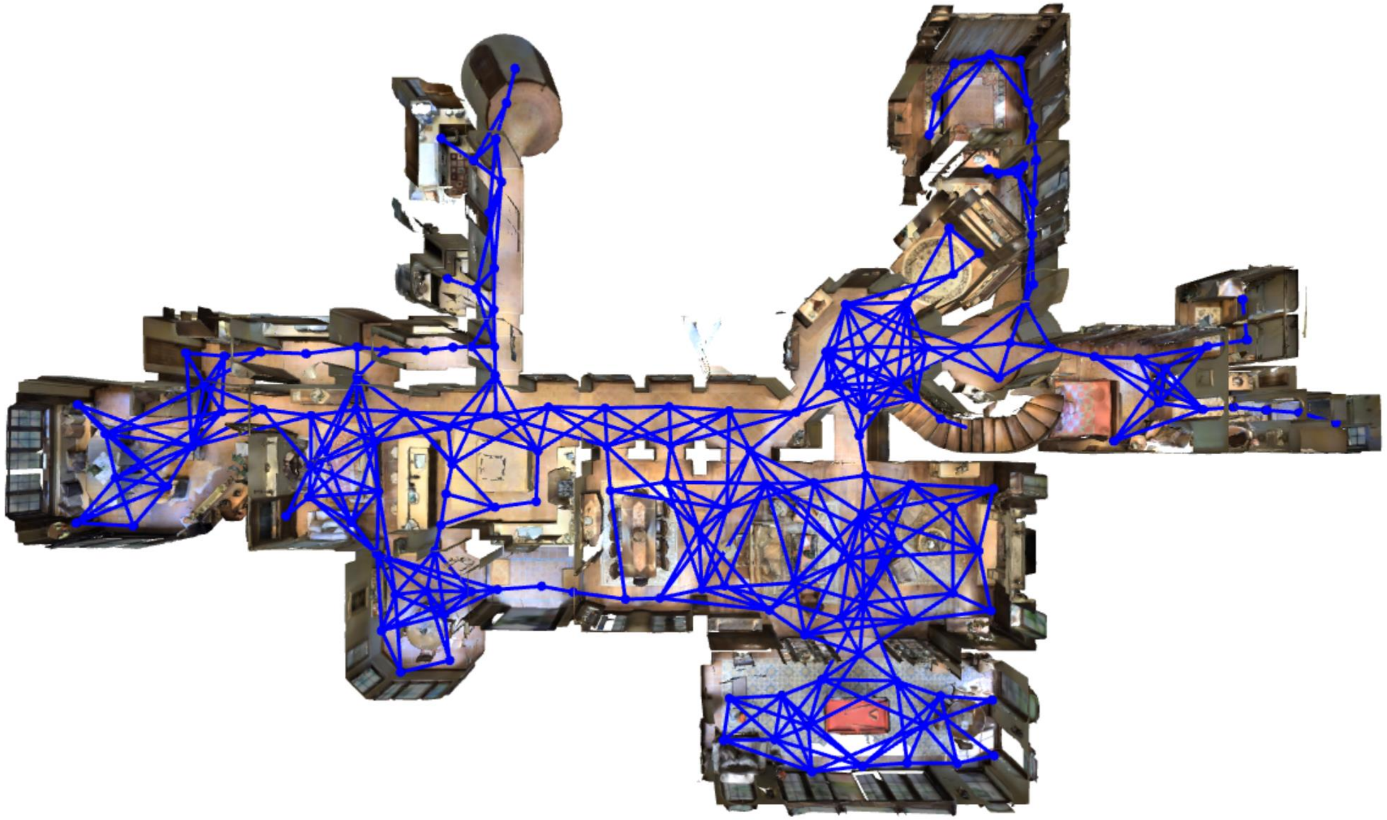
Figure 3. X. Wang et al. Look before you leap: Bridging model-free and model-based reinforcement learning for planned-ahead vision-and-language navigation. ECCV2018

Matterport3D Simulator

Figure 1. P. Anderson et al. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments.CVPR2018

- A new large-scale visual reinforcement learning simulation environment
- Based on Matterport3D Dataset
- 10800 panoramic views
- 194400 RGB-D images
- 90 building-scale scenes
- Construct a simulator by assign virtual 3D position, heading and camera elevation
- Discretized motions

Matterport3D Simulator



Example navigation graph for a partial floor of one building-scale scene in the Matterport3D Simulator

Navigable paths between panoramic viewpoints are illustrated in blue. Stairs can also be navigated to move between floors.

Figure 3. P. Anderson et al. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. CVPR2018

Room-to-Room(R2R) Navigation

Pass the pool and go indoors using the double glass doors. Pass the large table with chairs and turn left and wait by the wine bottles that have grapes by them.

Walk straight through the room and exit out the door on the left. Keep going past the large table and turn left. Walk down the hallway and stop when you reach the 2 entry ways. One in front of you and one to your right. The bar area is to your left.

Enter house through double doors, continue straight across dining room, turn left into bar and stop on the circle on the ground.

Standing in front of the family picture, turn left and walk straight through the bathroom past the tub and mirrors. Go through the doorway and stop when the door to the bathroom is on your right and the door to the closet is to your left.

Walk with the family photo on your right. Continue straight into the bathroom. Walk past the bathtub. Stop in the hall between the bathroom and toilet doorways.

Walk straight passed bathtub and stop with closet on the left and toilet on the right.

- Scene chose from Matterport3D Simulator

- Navigation instructions collected from Amazon Mechanical Turk
- US-based AMT workers use 3D WebGL environment writing directions
- 400 workers annotate about 1600 hours

- Three instructions provided for each scene

Figure 4 a),b). P. Anderson et al. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. CVPR2018

Motivation

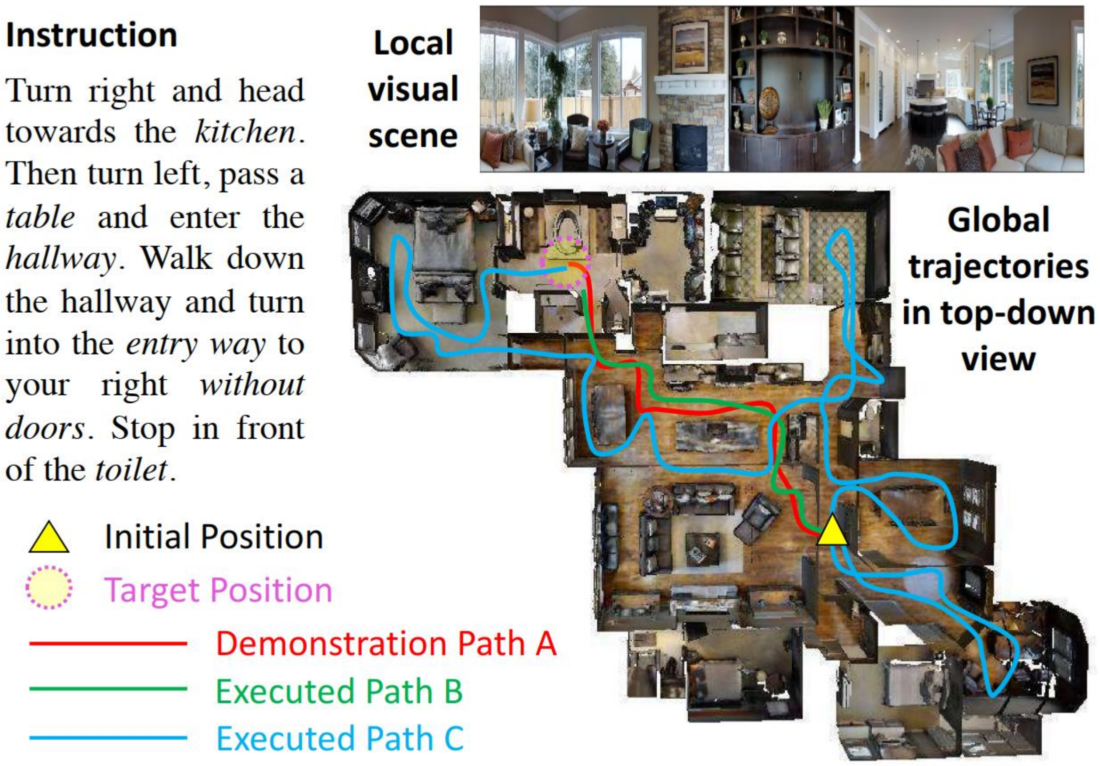
- Reasoning over visual images and natural language instructions can be difficult

Reasoning

Figure 3. X. Wang et al. Look before you leap: Bridging model-free and model-based reinforcement learning for planned-ahead vision-and-language navigation. ECCV2018

Motivation

- Feedback is rather coarse



Motivation

- Existing work suffer from the generalization problem

	Trajectory Length (m)	Navigation Error (m)	Success (%)	Oracle Success (%)
Val Seen:				
SHORTEST	10.19	0.00	100	100
RANDOM	9.58	9.45	15.9	21.4
Teacher-forcing	10.95	8.01	27.1	36.7
Student-forcing	11.33	6.01	38.6	52.9
Val Unseen:				
SHORTEST	9.48	0.00	100	100
RANDOM	9.77	9.23	16.3	22.0
Teacher-forcing	10.67	8.61	19.6	29.1
Student-forcing	8.39	7.81	21.8	28.4
Test (unseen):				
SHORTEST	9.93	0.00	100	100
RANDOM	9.93	9.77	13.2	18.3
Human	11.90	1.61	86.4	90.2
Student-forcing	8.13	7.85	20.4	26.6

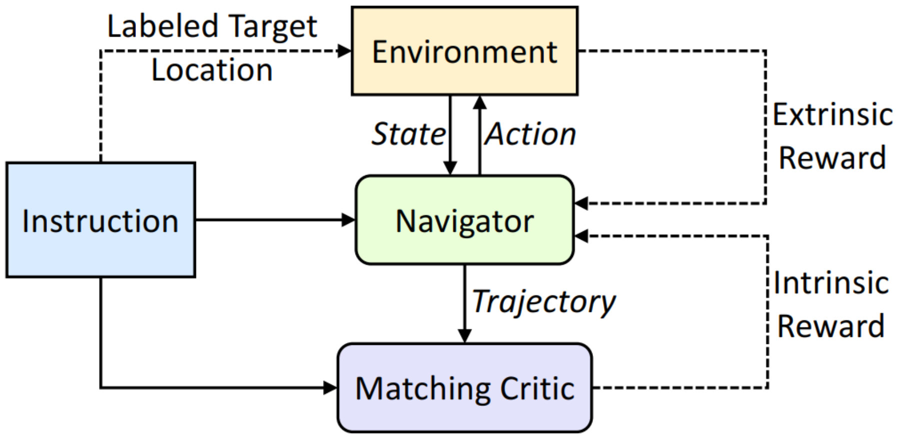
Table 1. P. Anderson et al. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. CVPR2018

	Val Seen				Val Unseen				Test (unseen)			
Model	TL (m)	NE (m)	SR (%)	OSR (%)	TL (m)	NE (m)	SR (%)	OSR (%)	TL (m)	NE (m)	SR (%)	OSR (%)
Shortest	10.19	0.00	100	100	9.48	0.00	100	100	9.93	0.00	100	100
Random	9.58	9.45	15.9	21.4	9.77	9.23	16.3	22.0	9.93	9.77	13.2	18.3
Teacher-forcing	10.95	8.01	27.1	36.7	10.67	8.61	19.6	29.1	-	-	-	-
Student-forcing	11.33	6.01	38.6	52.9	8.39	7.81	21.8	28.4	8.13	7.85	20.4	26.6
Ours												
XE	11.51	5.79	40.2	**54.1**	8.94	7.97	21.3	28.7	9.37	7.82	22.1	30.1
Model-free RL	10.88	5.82	41.9	53.5	8.75	7.88	21.5	28.9	8.83	7.76	23.1	30.2
RPA	8.46	**5.56**	**42.9**	52.6	7.22	**7.65**	**24.6**	**31.8**	9.15	**7.53**	**25.3**	**32.5**

Table 1. X. Wang et al. Look before you leap: Bridging model-free and model-based reinforcement learning for planned-ahead vision-and-language navigation. ECCV2018

Reinforced Cross-Modal Matching

Overview



Annotation

- Instruction $\chi = x_1, x_2, \ldots, x_n$
- Matching critic V_β
- Reasoning navigator π_θ
- Actions $a_1, a_2, \ldots, a_T \in \mathcal{A}$
- trajectory τ generated from $\mathcal{A}$
- Target location s_{target}

Reinforced Cross-Modal Matching

Reasoning Navigator

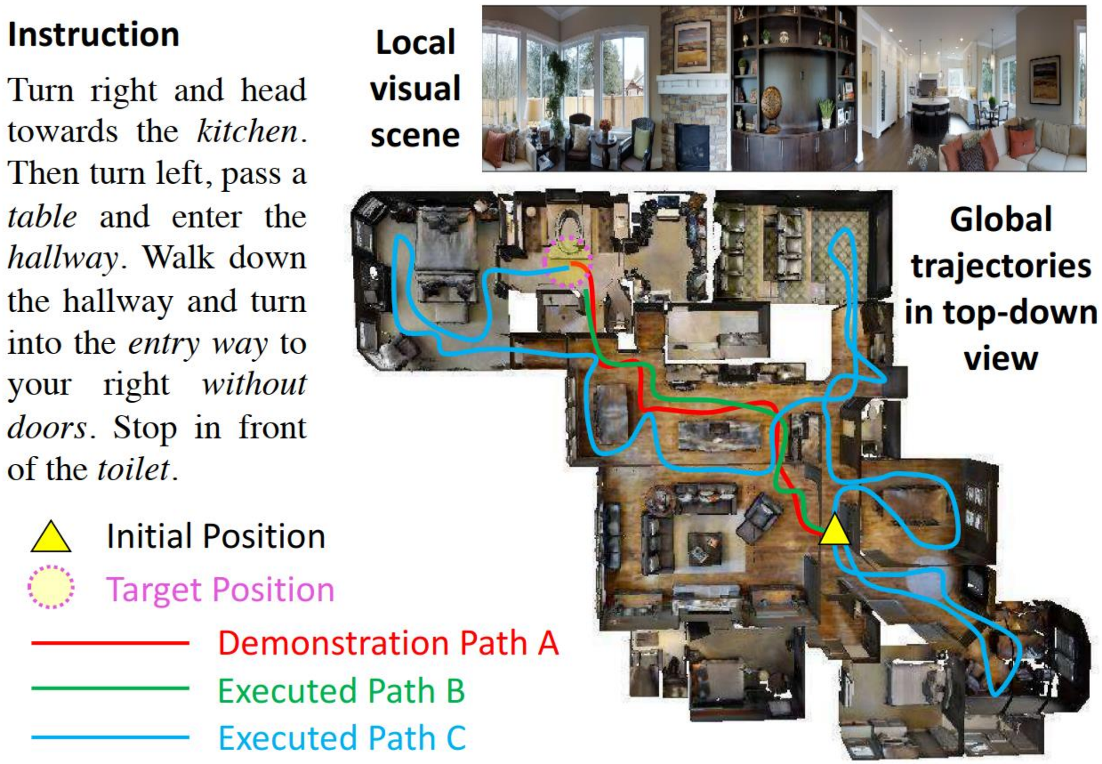
Navigator goals:

Mapping instruction to a sequence of actions:

$$\pi_\theta : \chi \rightarrow \mathcal{A} = \{a_1, \ldots, a_T\}$$

Navigator learns:

- Trajectory history
- Focus of textual instruction
- Local visual attention

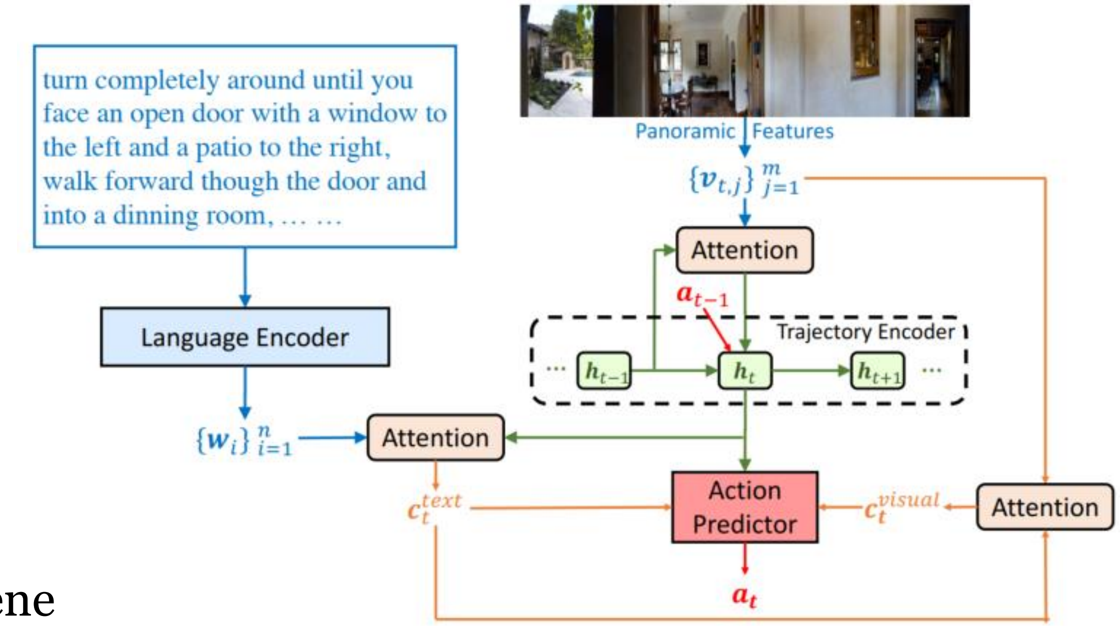


Reinforced Cross-Modal Matching

Reasoning Navigator

At time step t:

- Receive a state s_t from the environment
- Split panoramic view into m different viewpoints
- Extract image patch features at step t as $\{v_{t,j}\}_{j=1}^{m}$
- Ground the textual instruction χ in the local visual scene



Cross-modal reasoning navigator at step t

Reinforced Cross-Modal Matching

Reasoning Navigator

History Content:

Attention-based*(dot-product)* trajectory LSTM encoder:

$$h_t = LSTM([v_t, a_{t-1}], h_{t-1})$$

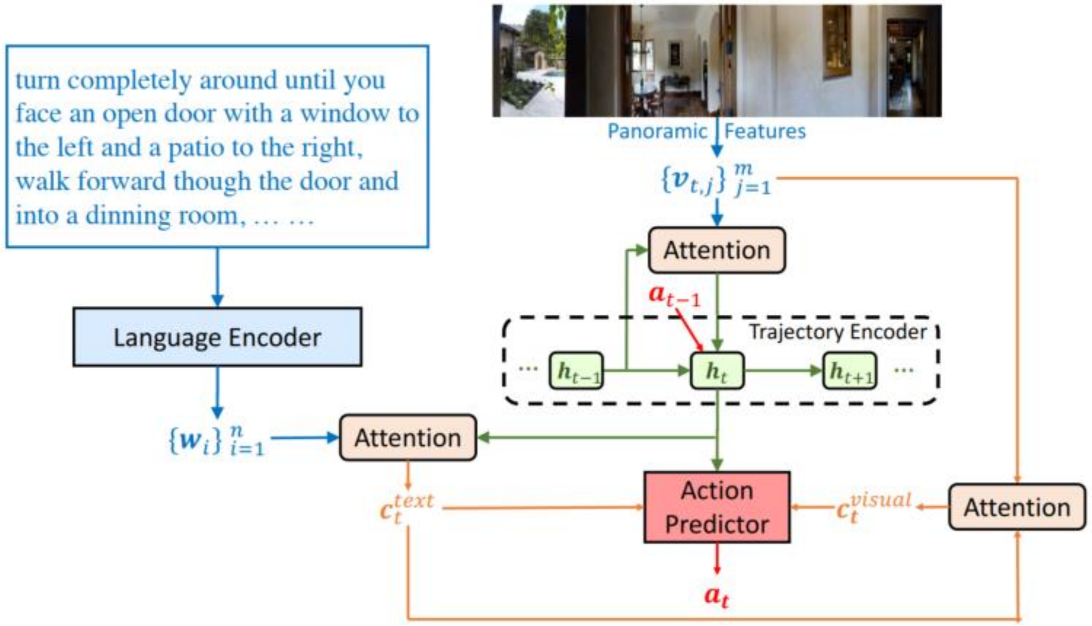
$$v_t = attention\left(h_{t-1}, \{v_{t,j}\}_{j=1}^{m}\right)$$

$$= \sum_j softmax(h_{t-1} W_h (v_{t,j} W_v)^T) v_{t,j}$$

v_t represents weighted sum of panoramic features

h_t represents trajectory $\tau_{1:t}$ till step t

W_h and W_v are learnable projection matrices



Cross-modal reasoning navigator at step t

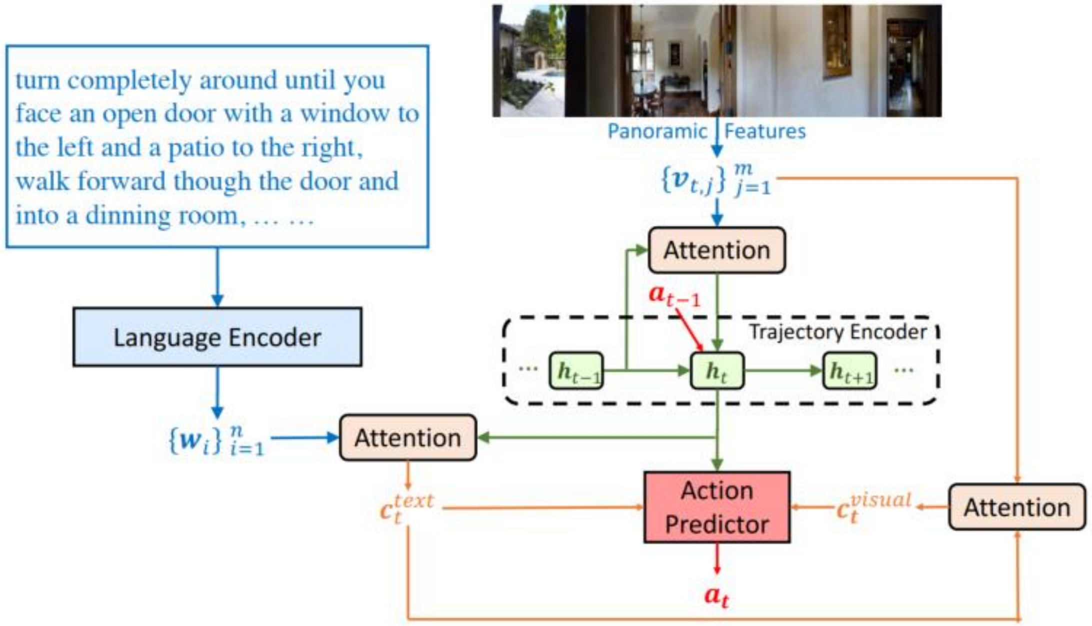
Reinforced Cross-Modal Matching

Reasoning Navigator

Visually Conditioned Textual Context:

- Use a language encoder LSTM to encode instruction χ into a set of textual features $\{w_i\}_{i=1}^{n}$
- Compute textual context:

$$c_t^{text} = attention(h_t, \{w_i\}_{i=1}^{n})$$



Cross-modal reasoning navigator at step t

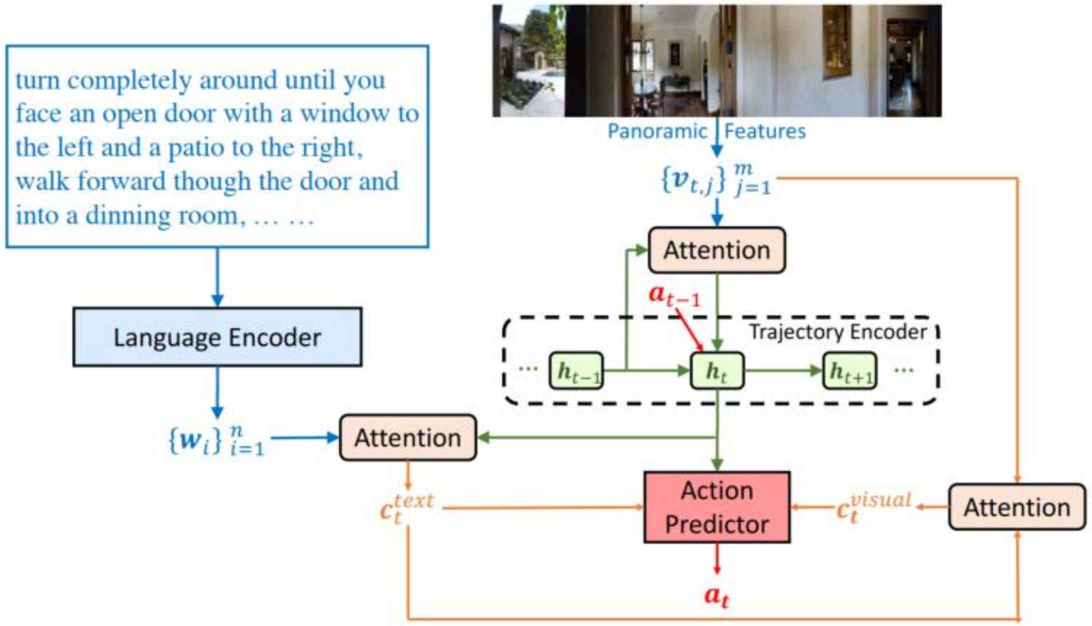
Reinforced Cross-Modal Matching

Reasoning Navigator

Textually Conditioned Visual Context:

- Compute textual context:

$$c_t^{visual} = attention(c_t^{text}, \{v_j\}_{j=1}^{m})$$



Cross-modal reasoning navigator at step t

Reinforced Cross-Modal Matching

Reasoning Navigator

Action Prediction:

$$p^k = softmax([h_t, c_t^{text}, c_t^{visual}]W_c(u_k W_u)^T)$$

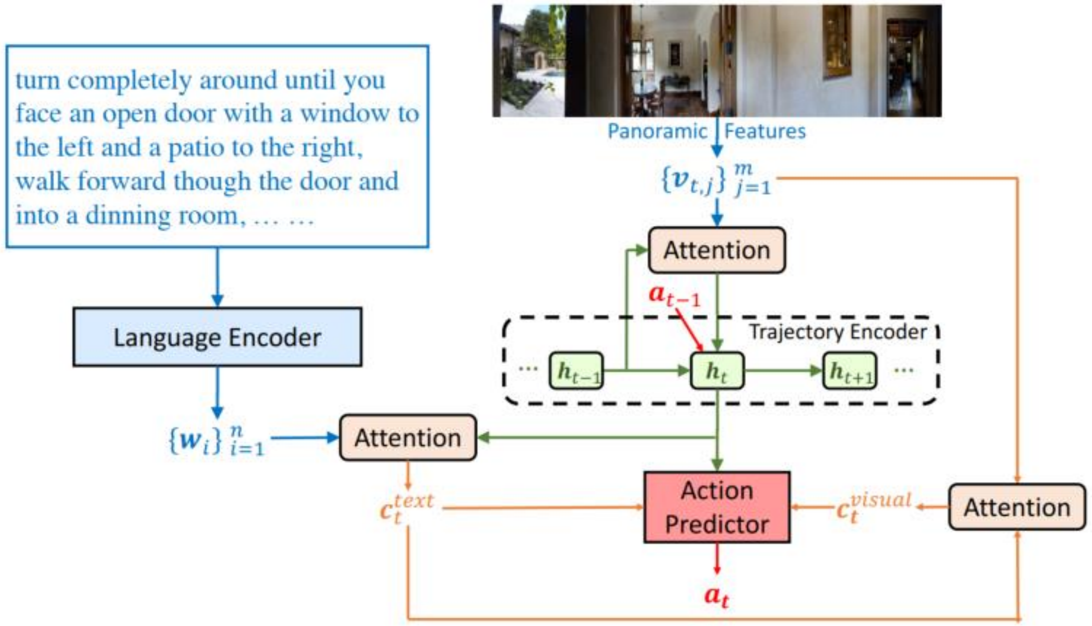
u_k is the action embedding that represents the k-th navigable direction

$$u_k = concat(vec_{appear}, vec_{orien})$$

vec_{appear} is a CNN feature vector

vec_{orien} is a 4-d feature vector $[sin\psi; cos\psi; sin\omega; cos\omega]$

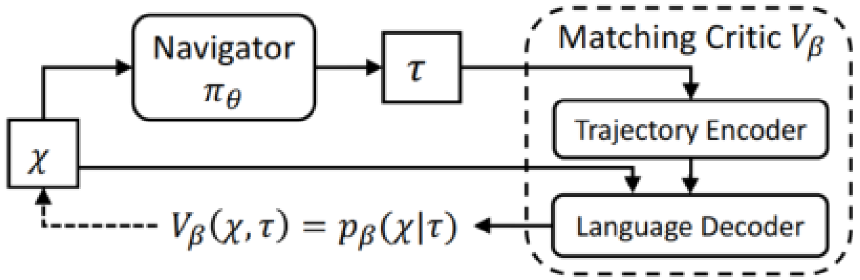
ψ and ω are heading and elevation angles



Cross-modal reasoning navigator at step t

Reinforced Cross-Modal Matching

Cross-Modal Matching Critic



- Matching Critic provided intrinsic reward:

$$R_{intr} = V_\beta(\chi,\tau) = V_\beta(\chi,\pi_\theta(\chi))$$

- Measure cycle-reconstruction reward as intrinsic reward:

$$R_{intr} = p_\beta(\chi|\pi_\theta(\chi)) = p_\beta(\chi|\tau)$$

- Matching Critic is pre-trained with human-demonstrations via supervised learning

Reinforced Cross-Modal Matching

Learning: warmup

- Use demonstration actions to conduct supervised learning with MLE

$$L_{sl} = -\mathbb{E}[\log(\pi_\theta(a_t^*|s_t))]$$

- To quickly approximate a relatively good policy
- Bad generalizability, which need further RL

Reinforced Cross-Modal Matching

Learning: extrinsic reward

$D_{target}(s_t)$ is the distance between s_t and s_{target}

$$r(s_t, a_t) = D_{target}(s_t) - D_{target}(s_{t+1}), \qquad t < T$$

$$r(s_T, a_T) = \mathbb{1}(D_{target}(s_T) \leq d) \quad t = T$$

Use discounted cumulative reward to incorporate the influence of action in the future and account for local greedy search

$$R_{extr}(s_t, a_t) = r(s_t, a_t) + \sum_{t'=t+1}^{T} \gamma^{t'-t} r(s_{t'}, a_{t'})$$

Reinforced Cross-Modal Matching

Learning: intrinsic reward

$$R_{intr} = p_{\beta}(\chi | \pi_{\theta}(\chi)) = p_{\beta}(\chi | \tau)$$

Reinforced Cross-Modal Matching

Learning: training

- Loss Function:

$$A_t = R_{intr} + \delta R_{extr}$$

$$L_{lr} = -\mathbb{E}_{a_t \sim \pi_\theta}[A_t]$$

- Training function:

$$\nabla_\theta L_{lr} = - A_t \nabla_\theta \log \pi_\theta(a_t|s_t)$$

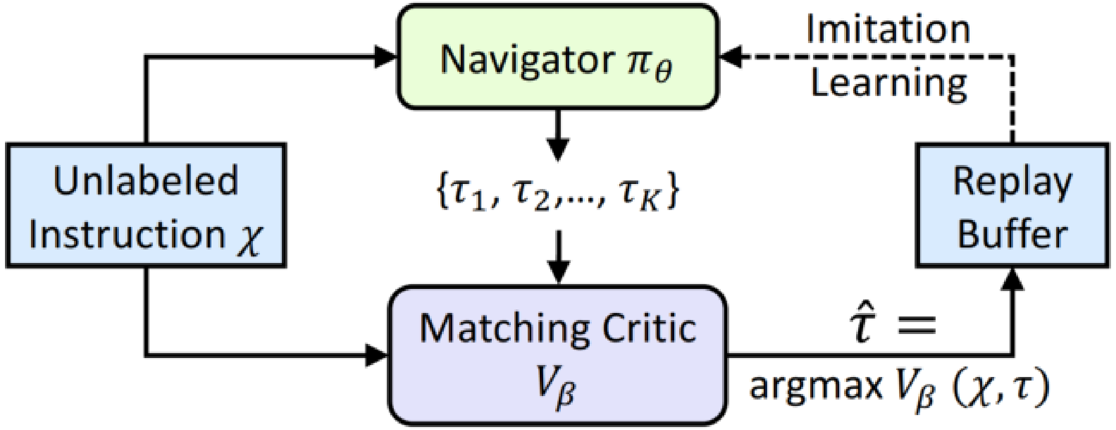
Self-Supervised Imitation Learning

Exploitation and Exploration:

- Explore in new environment without ground-truth demonstration
- Exploit useful trajectory in a replay buffer

Facilitates lifelong learning and adaption to new environments

Self-Supervised Imitation Learning



SIL for exploration on unlabeled data

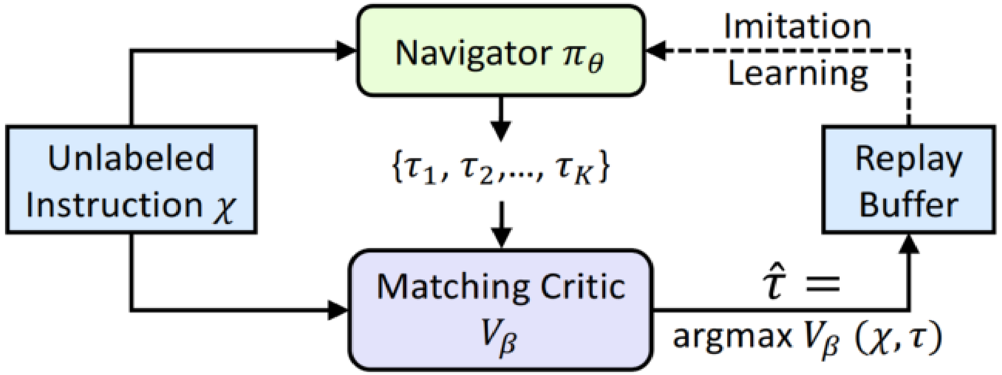
Given an instruction χ without target location

Using matching critic to choose the best trajectory

$$\hat{\tau} = argmax V_\beta(\chi, \tau)$$

Target location is unknown - No Supervision!

Self-Supervised Imitation Learning



SIL for exploration on unlabeled data

Use loss for policy gradient

use off-policy Monte-Carlo return instead of on-policy return

$$L_{sil} = -R_{intr} log \pi_\theta(a_t|s_t)$$

Loss function can also be interpreted as supervised learning loss

Regard $\hat{\tau}$ as ground-truth and $\hat{a_t}$ is action stored in replay buffer

$$L_{sil} = -\mathbb{E}[log(\pi_\theta(\hat{a_t}|s_t))]$$

Experiments

- R2R dataset:

  7129 paths and 21567 human-annotated instructions

  Testing scenario of VLN task is standard and fully-unseen

- Standard evaluation metric:

  Path Length(PL), Navigation Error(NE),Oracle Success Rate(OSR),

  Success Rate(SR), Success rate weighted by inverse Path Length(SPL)

- Implementation Details:

  ResNet-152 CNN features and pretrained GloVe word embeddings

Experiments

Comparison with SOTA

Test Set (VLN Challenge Leaderboard)					
Model	PL ↓	NE ↓	OSR ↑	SR ↑	**SPL** ↑
Random	9.89	9.79	18.3	13.2	12
seq2seq [3]	**8.13**	7.85	26.6	20.4	18
RPA [50]	9.15	7.53	32.5	25.3	23
Speaker-Follower [13]	14.82	6.62	44.0	35.0	28
+ beam search	*1257.38*	4.87	96.0	53.5	*1*
Ours					
RCM	15.22	**6.01**	**50.8**	**43.1**	35
RCM + SIL (train)	**11.97**	6.12	49.5	43.0	**38**
RCM + SIL (unseen)[3]	9.48	4.21	66.8	60.5	59

- Outperform significantly on SPL
- SIL shortens Path Length

Experiments

Ablation study

#	Model	Seen Validation PL ↓	NE ↓	OSR ↑	SR ↑	Unseen Validation PL ↓	NE ↓	OSR ↑	SR ↑
0	Speaker-Follower (no beam search) [13]	-	3.36	73.8	66.4	-	6.62	45.0	35.5
1	RCM + SIL (train)	**10.65**	3.53	75.0	66.7	**11.46**	6.09	50.1	**42.8**
2	RCM	11.92	3.37	76.6	67.4	14.84	**5.88**	**51.9**	42.5
3	− intrinsic reward	12.08	3.25	**77.2**	**67.6**	15.00	6.02	50.5	40.6
4	− extrinsic reward = pure SL	11.99	3.22	76.7	66.9	14.83	6.29	46.5	37.7
5	− cross-modal reasoning	11.88	**3.18**	73.9	66.4	14.51	6.47	44.8	35.7
6	RCM + SIL (unseen)	**10.13**	**2.78**	**79.7**	**73.0**	**9.12**	**4.17**	**69.31**	**61.3**

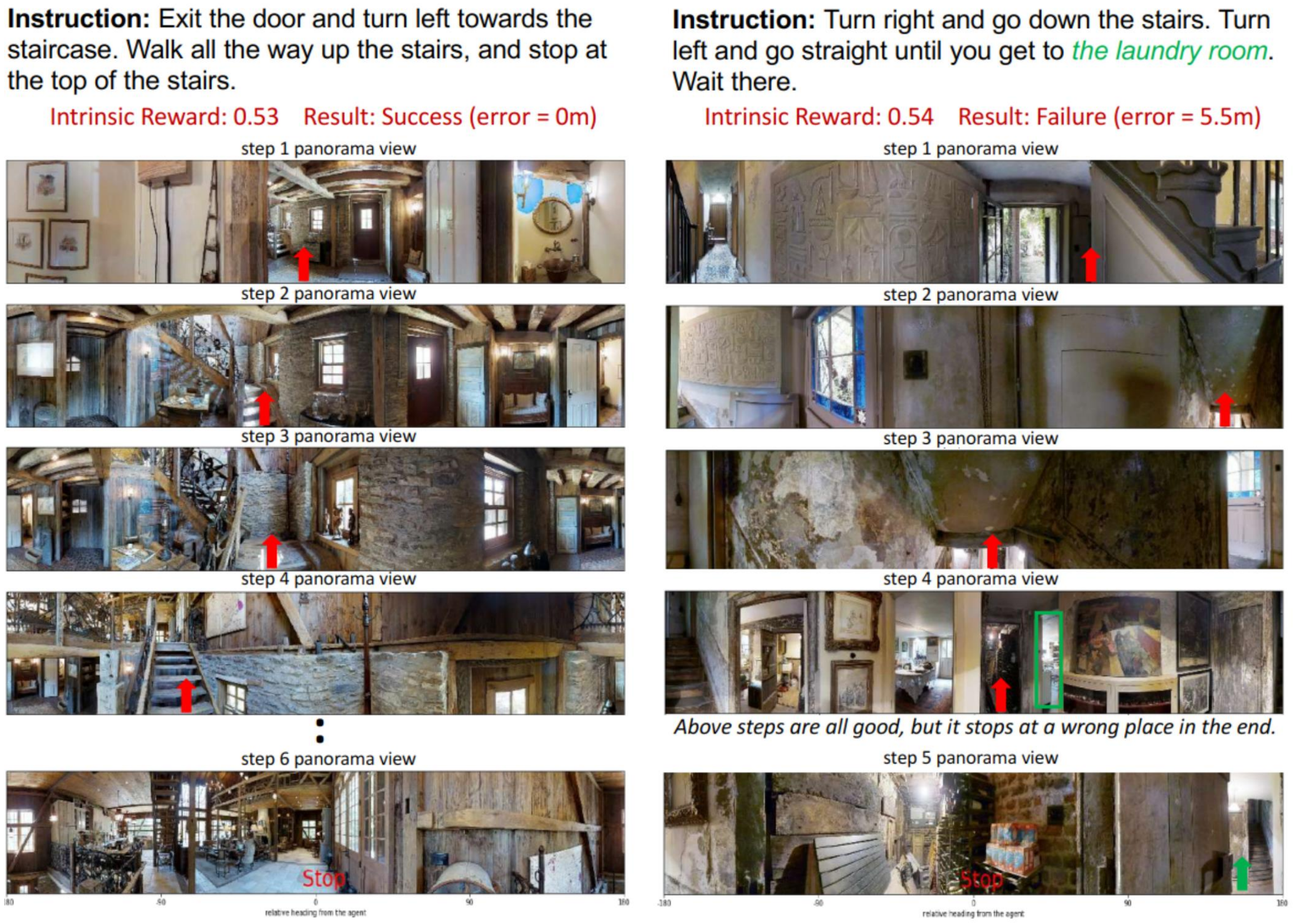
Extrinsic reward and intrinsic reward are complementary
Intrinsic reward can improve exploration on unseen environments
Extrinsic reward can guarantee the stability of RL
Cross-modal reasoning and SIL are also useful

Experiments

Qualitative Analysis

Unable to understand
the laundry room

More precise visual grounding



(a) A successful case

(b) A failure case

More about VLN…

VLN leaderboard

bringmeaspoon.org

B - Baseline submission

Rank	Participant team	length	error	oracle success	success	spl	Last submission at
1	human	11.85	1.61	0.90	0.86	0.76	1 year ago
2	Self-Supervised Auxiliary Reasoning Tasks (Beam Search)	40.85	3.24	0.81	0.71	0.21	11 days ago
3	Back Translation with Environmental Dropout (with Beam Search) (null)	686.82	3.26	0.99	0.69	0.01	11 months ago
4	Self-Supervised Auxiliary Reasoning Tasks (Pre-explore)	10.43	3.69	0.75	0.68	0.65	16 days ago
5	vBot (Greedy)	10.24	3.76	0.71	0.65	0.62	4 months ago
6	Back Translation with Environmental Dropout (exploring unseen environments before testing)	9.79	3.97	0.70	0.64	0.61	11 months ago
7	Reinforced Cross-Modal Matching (optimized for SR; with beam search)	357.62	4.03	0.96	0.63	0.02	1 year ago
8	Self-Monitoring Navigation Agent (with beam search) (Self-Aware Co-Grounded Model)	373.09	4.48	0.97	0.61	0.02	1 year ago
9	Tactical Rewind - long	196.53	4.29	0.90	0.61	0.03	11 months ago


THANKS!