Language-Vision Guided 3D Indoor Navigation with Reinforcement Learning

Jiali Duan (jialidua@usc.edu)

Project Page: http://mcl-lab.usc.edu:3000/trajectory.html

Department of Electrical Engineering, University of Southern California, USA

Motivation

Combining language instruction and visual observation as guidance for 3D indoor navigation

Action Decision is Ambiguous

Misalignment between language instruction and vision information

Hard to interpret decision logic

Proposed Solution

(a) Vision and Language Co-Attention

(b) Maximum Likelihood Estimation

Better generalization to unseen scenarios

Framework Pipeline

Our visual navigation is mapless and only uses language instruction $X = (x_1, x_2 ... x_L)$ & visual observation $V_t = (v_{t,1}, v_{t,2}, ..., v_{t,K})$ as input, where L is the number of words and K the number of navigable direction. The visual-language navigator framework which we abbreviate as VLN involve four major steps.

Step 1: Language Encoding

– Encode language instruction into word embeddings.

Step 3: Synthesize

– Train Synthesizer for data augmentation and pragmatic inference.

Step 2: Attention

– Extract ResNet-152 visual features with attention.

Step 4: Joint Training

– Joint train synthesizer and navigator with MLE.

Vision and language Co-Attention is performed in steps 1–2 and maximum likelihood estimation is performed in step 3–4.

Step 1: Language Encoding

The agent is expected to understand the context of instruction given current panoramic visual observations. The attention weight over L words of the instruction is computed as:

$$z_{t,l}^{textual} = (W_x h_{t-1})^T x_l \tag{1}$$

$$\alpha_t = softmax(z_t^{textual}) \tag{2}$$

where W_x are parameters to be learnt. $z_{t,l}^{textual}$ denote the correlation between word l and previous hidden state h_{t-1} and α_t is the weight over textual features X at time t.

Based on the attention distribution, the textual feature $\hat{x}_t$ is the weighted combination of textual representation $\hat{x}_t = \alpha_t^T X$.

Step 2: Attention

For each decision making, the agent needs to identify the most salient visual regions from current visual observations. We perform visual attention over image features from current views:

$$z_{t,k}^{visual} = (W_{v_1} h_{t-1})^T W_{v_2} v_{t,k}, \ \beta_t = softmax(z_t^{visual}) \tag{3}$$

where W_{v_1} and W_{v_2} are parameters to be learnt. Similar to Eqn 1. The grounded visual feature $\hat{v}_t$ is the weighted combination of visual features $\hat{v}_t = \beta_t^T V$.

Based on textual grounding, visual grounding above, action chosen at time step t is a bilinear dot product involving past history h_t and navigable action embedding at current step as: $y_t = (W_{o_1} h_t)^T W_{o_2} a_t$ and $p_t = softmax(y_t)$.

Step 3-4: Joint Training and MLE

Training process involves two steps:

– Pretrain synthesizer for data augmentation.

– Joint train synthesizer with navigator.

Specifically, the synthesizer is pretrained using Eqn. 4

$$\hat{d}_k = argmax_d P_S(d \mid \hat{r}_k) \tag{4}$$

We augment navigation instruction and route pairs $\mathcal{D} = (d_1, r_1) \dots (d_N, r_N)$ by greedily generating synthetic instructions on sampled new routes in the environment. Then, the synthesizer model $P_S(d \mid r)$ is joint-trained with the navigator model $P_N(r \mid d)$ by approxmating Eqn. 5

$$argmax_{r \in R(d)} P_S(d \mid r)^{\lambda} \cdot P_N(r \mid d)^{(1-\lambda)} \tag{5}$$

where $lambda$ is a hyper-parameter in the range $[0, 1]$. When λ is close to 1, it means that we rely mostly on the score of synthesizer to select routes. We observe the best performance with $\lambda = 0.1$.

Results and Conclusions

we submit our result to Vison and Language Navigation challenge online test server. We achieved 55.67% (corresponds to Table 1, † means with data augmentation) success rate on test-split, better than CVPR2018, ECCV2018 and NIPS2018 results.

Method	Validation-Seen NE ↓	Validation-Seen SR ↑	Validation-Seen OSR ↑	Validation-Unseen NE ↓	Validation-Unseen SR ↑	Validation-Unseen OSR ↑	Test (unseen) NE ↓	Test (unseen) SR ↑	Test (unseen) OSR ↑
Random	9.45	15.9	21.4	9.23	16.3	22.0	9.77	13.2	18.3
Student-forcing [1]	6.01	38.6	52.9	7.81	21.8	28.4	7.85	20.4	26.6
RPA [2]	5.56	42.9	52.6	7.65	24.6	31.8	7.53	25.3	32.5
Speaker-follower[3]	3.88	63.0	71.0	5.24	50.0	63.0	-	-	-
Speaker-follower(†)	3.08	70.1	78.3	4.83	54.6	65.2	4.87	53.5	96.0
Ours	3.26	67.58	74.93	4.91	53.26	64.96	-	-	-
Ours (†)	**2.88**	**71.79**	**80.80**	**4.76**	**54.79**	**67.65**	**4.57**	**55.67**	95.81

Qualitative results are available at Project Page above or https://sites.google.com/view/submission-2019.

Bibliography

[1] Anderson et.al. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. *CVPR*. 2018.

[2] Wang et.al. Look Before You Leap: Bridging Model-Free and Model-Based Reinforcement Learning for Planned-Ahead Vision-and-Language Navigation. *ECCV*. 2018.

[3] Fried et. al. Speaker-Follower Models for Vision-and-Language Navigation. *NIPS*. 2018.

[4] Anderson et.al. On Evaluation of Embodied Navigation Agents. *arXiv preprint arXiv: 1807.06757*. 2018.