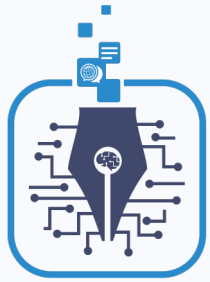
Generating Natural-Language Navigation Instructions from Panoramic Images

RWBC-OIL, Tokyo Tech, AIST

Erick Mendieta M2

Naoaki Okazaki, Hiroya Takamura

AIST



OKAZAKILAB

Outline

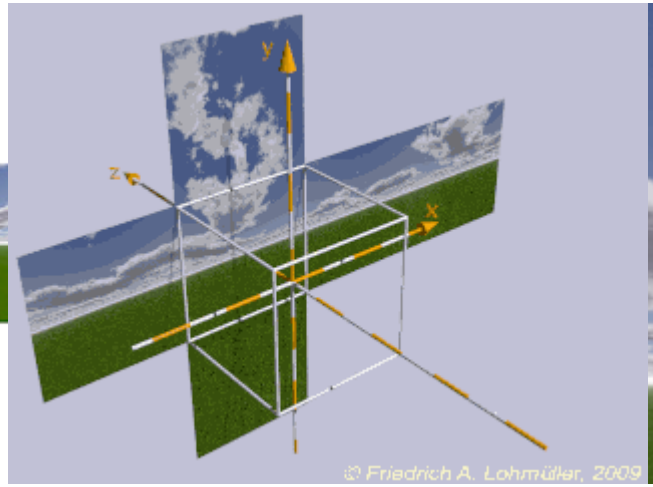
- Overview
- Problem
- Goal
- Dataset
- Method
- Results
- Conclusion

Overview 1/3

The idea of an agent that can follow instructions base on natural language is a long-held goal in AI.

With the rise of AR/VR we can create realistic environments from panoramic images and train an agent to follow instructions in the generated environment.*

Start with a 360º panorama



Wrap the panorama to a box

Obtain a simulated environment

* image from Friedrich A. lohmuller: http://www.f-lohmueller.de/

Overview 2/3

Moreover, we can connect multiple real-life panoramic images to create complex environments.

A complex environment is generated from hundreds of panoramic images arranged according to physical order.

Video from: https://matterport.com/

Overview 3/3

In such complex environments, we can provide natural language instructions for an agent to follow.

The agent needs to generate a route to get as close as possible to the place described in the instruction.

Leave the bedroom, and enter the kitchen. Walk forward, and take a left at the couch. Stop in front of the window.

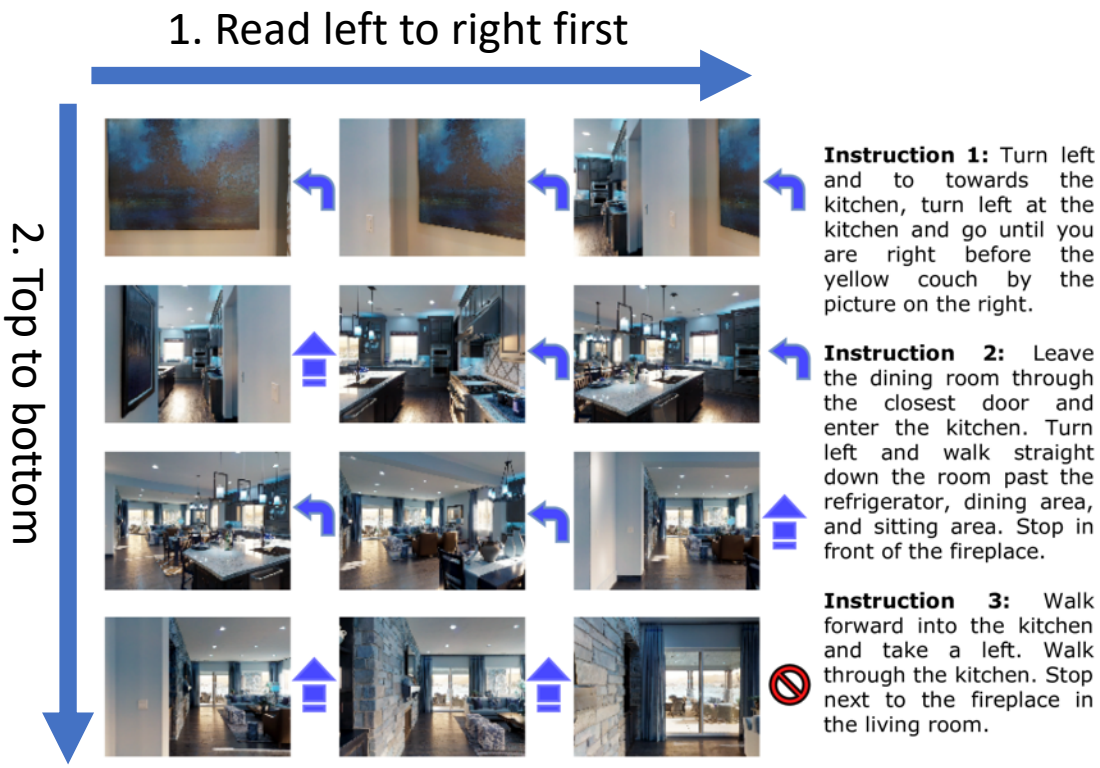
The video shows a route travelled by an agent. the actions taken are represented by the arrow [left, right, up, down, stop, forward].

Image From: https://bringmeaspoon.org/

Problem

The problem with the previous set up is that is very time-consuming for a human create an instruction for a route.

Different people generate different instructions for the same route. So it is difficult to automate the generation of human-like instructions.



The figure shows 3 different instructions collected from humans referring to the same route.

RWBC-OIL

OKAZAKILAB

Goal

The **goal** of the research will be to ***use simulator's panoramic image routes*** ***to generate*** ***Human-like Natural Language instructions***.

入力

?
モデル

出力

Generated Instruction: Walk forward past the dining table and past the white sofas. Wait near the glass doorway.

Principles

- Humans tend to select different reference objects to create their instructions.
- Humans tend to produce different instructions even when given the same route.

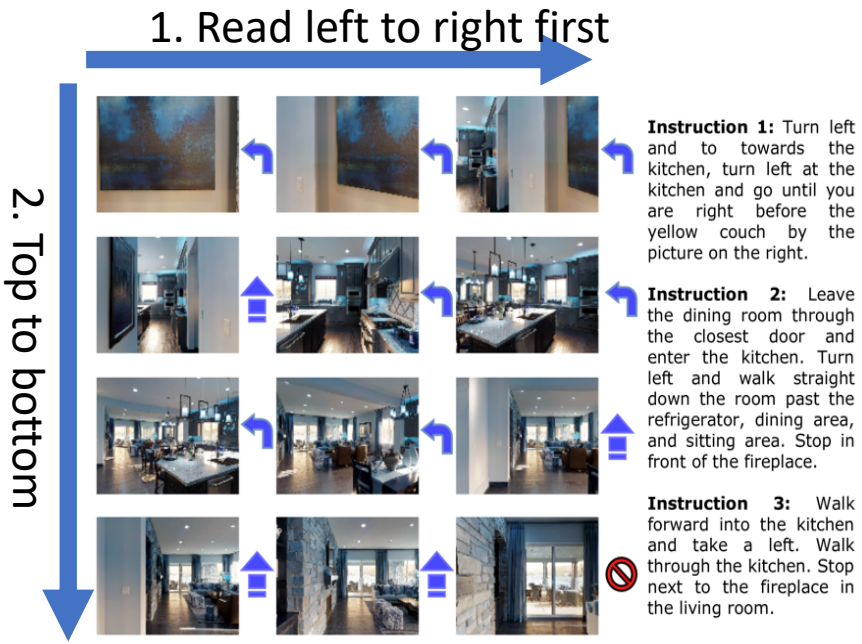
RWBC-OIL

OKAZAKILAB

Dataset

Matterport dataset*

- 10.800 panoramic images of 90 houses.
- 7,189 Routes
- 3 instructions per route for a total of 21,567 instruction.
- Vocabulary size 3.1k words with 1.2k with more than 5 mentions.
- Average instruction length is 29 words.
- Splits: training, validation seen, validation unseen and test.



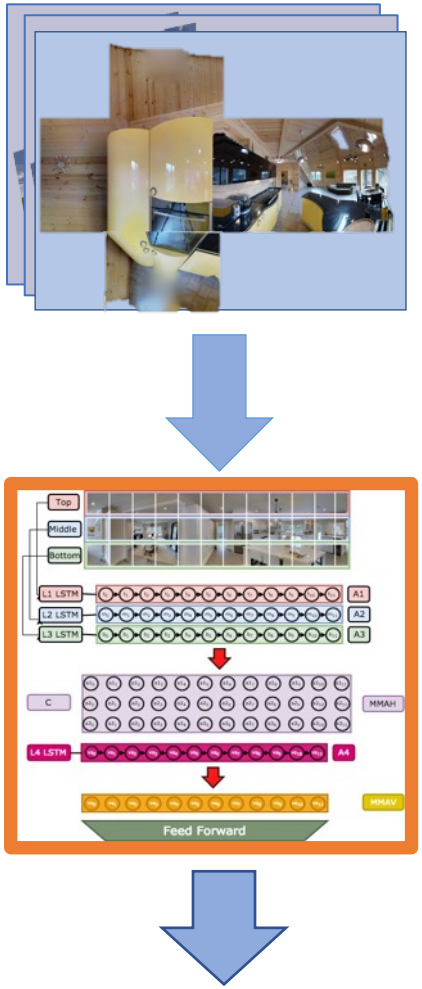
*Peter Anderson et al. “Vision-and-Language Navigation: Interpreting visually-grounded navigation instructions in real environments.

Approach

We create a new architecture inspired by video processing techniques.*

The architecture implicitly captures the positional and temporal relations of parts of panoramic images and routes.

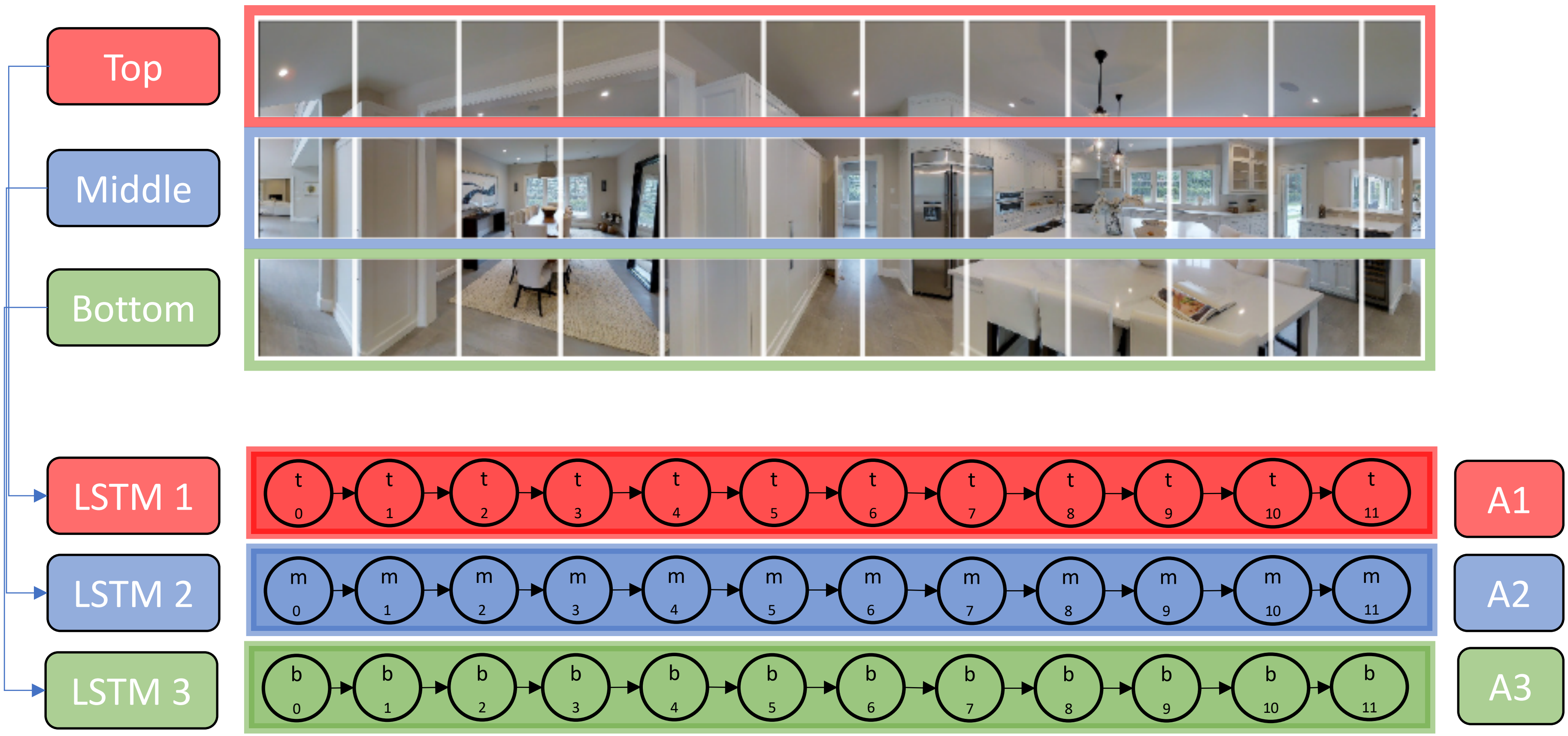
By doing that we want to focus on different reference objects along the way in the panoramic image and produce different human-like instructions.



Generated Instruction: Walk forward past the dining table and past the white sofas. Wait near the glass doorway.

* Huanyu Yu et al. “Fine-grained Video Captioning for Sports Narrative”

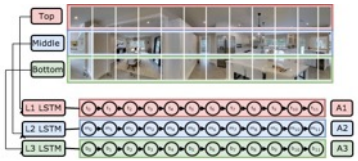
Method 1/4

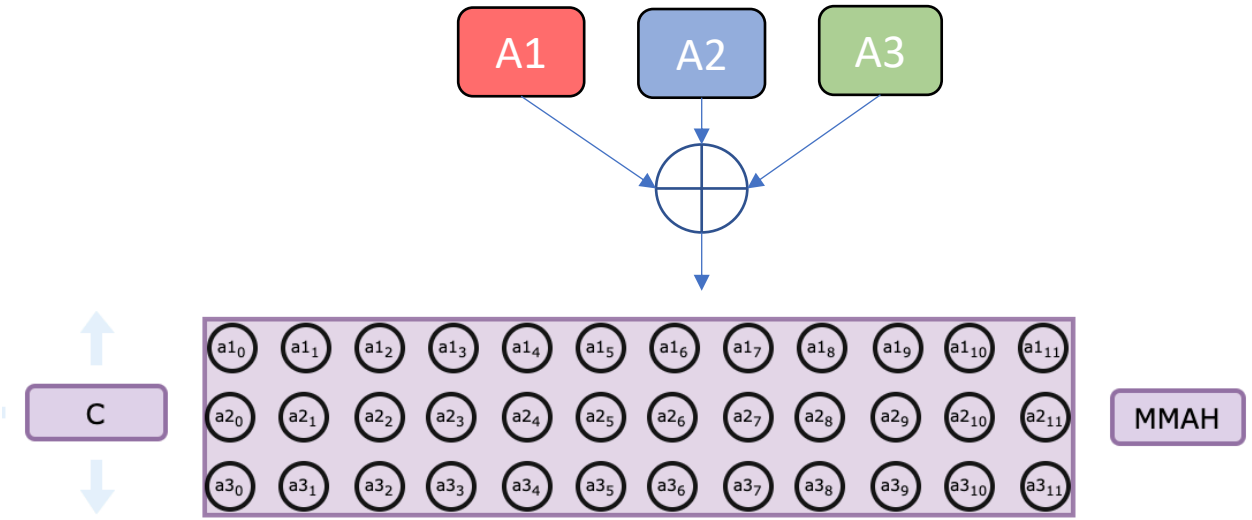


… the roof, main vision line, and the floor, respectively.

*The images are preprocessed with Restnet151 before going to the LSTM

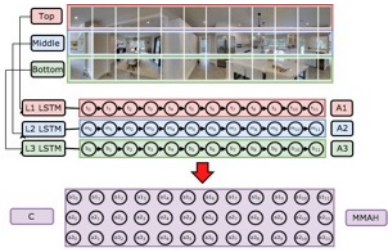
Method 2/4

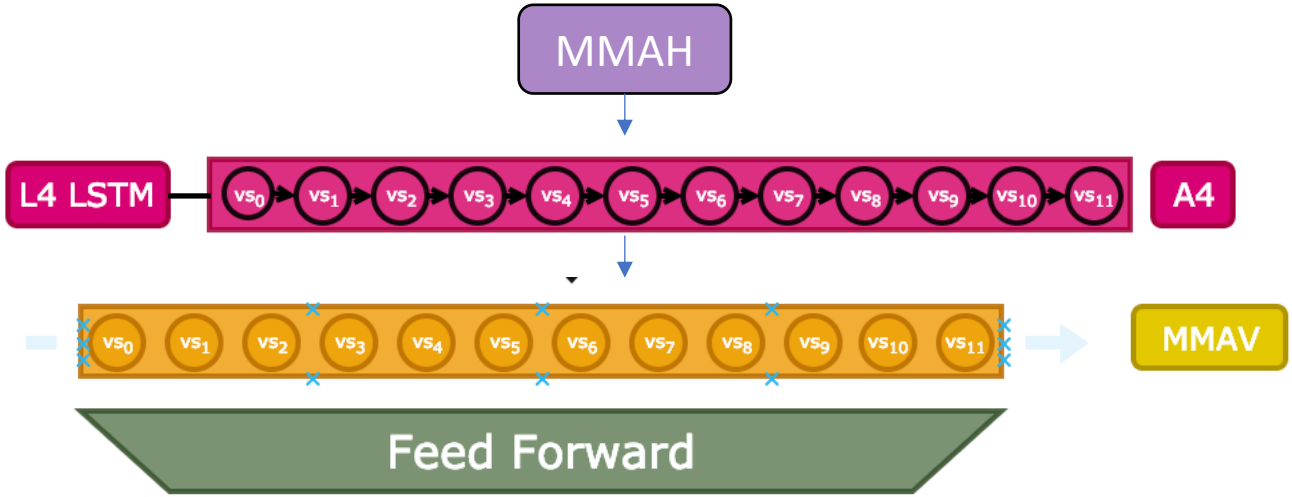




4. We concatenate the unrolled results from the LSTM layers A1,A2, A3 and perform attention to them to see the different object Horizontally. (From left to right)

Method 3/4



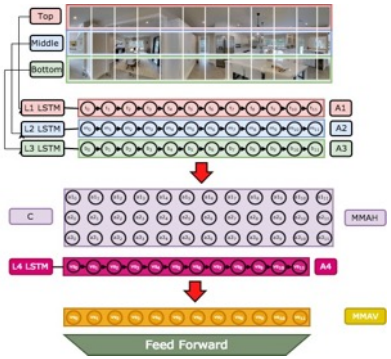


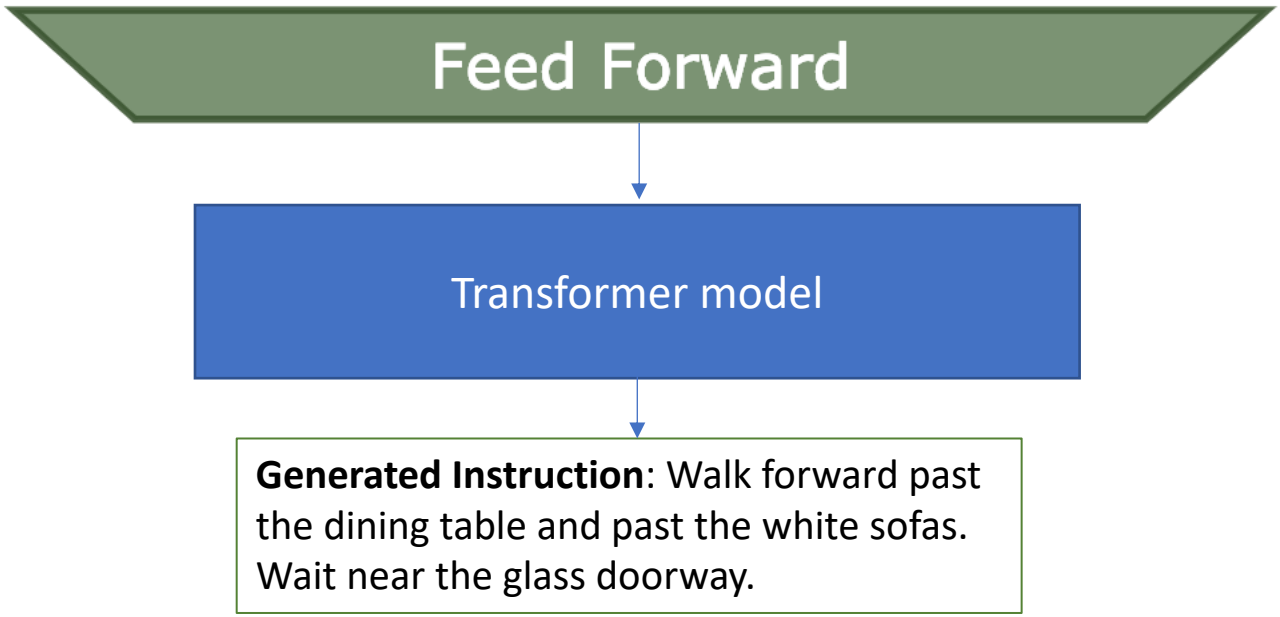
5. We then input the results of the attention to another LSTM layer with the purpose of capturing the relationship of the multiple panoramic images in the route vertically.

6. With the unrolled results of that layer A4 we perform attention once again to capture the relationship of the panoramic images in the route and then a Feed Forward layer to create the panoramic embeddings.

RWBC-OIL
OKAZAKILAB

Method 4/4



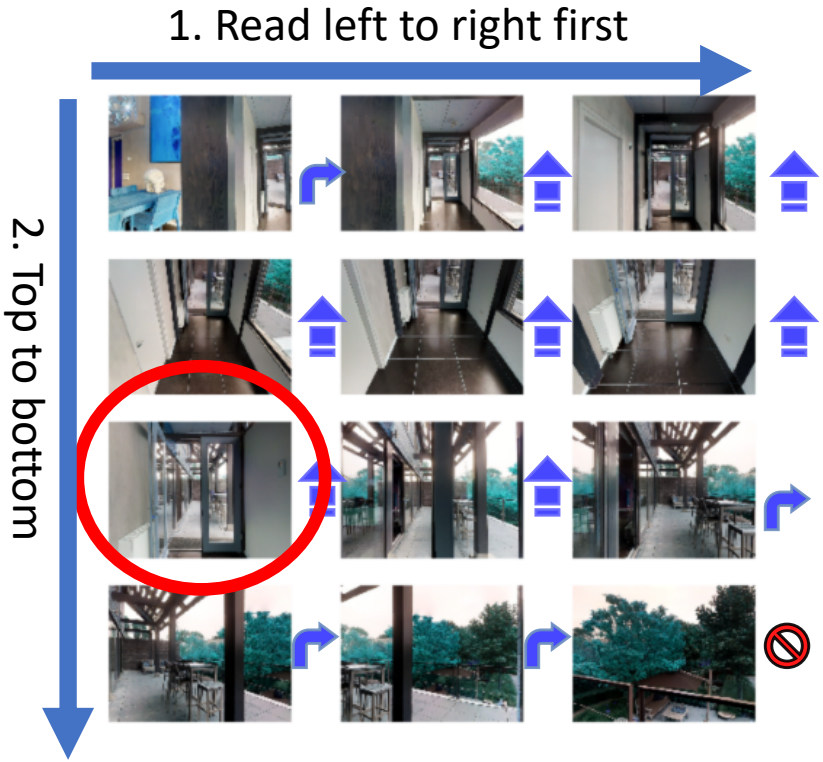


7. As Last step we use the transformer model to generate the instruction from our Panoramic embeddings.

Results 1/3

As a result, our method can recognize objects by their specific names making the descriptions more like human.

For example “glass door” is predicted rather than a generic term “doorway”.



Previous work: walk down the hallway and turn right . walk down the hall and stop in the doorway .

Our work: turn around and go out the door . go out the glass door and go outside . wait on the porch .

Results 2/3

Golden: Exit the kitchen heading towards the small dining nook and turn right. Continue forward and take the entrance ahead and to the right. Wait at the corner next to the end of both couches.

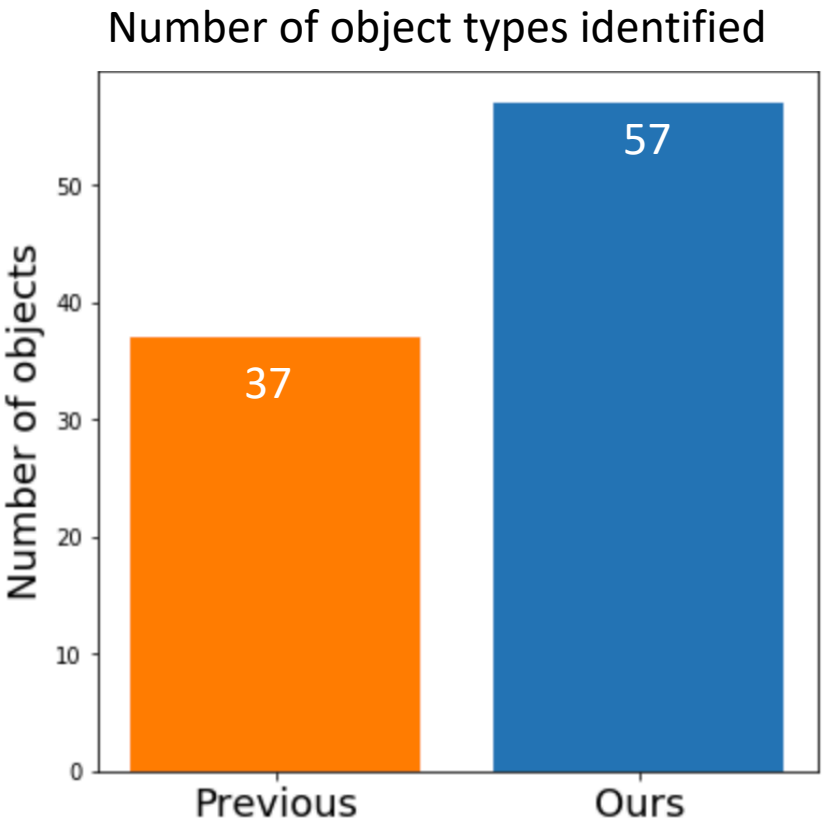
Previous work: turn around and walk through the kitchen . turn right and walk past the dining table . stop in front of the dining table .

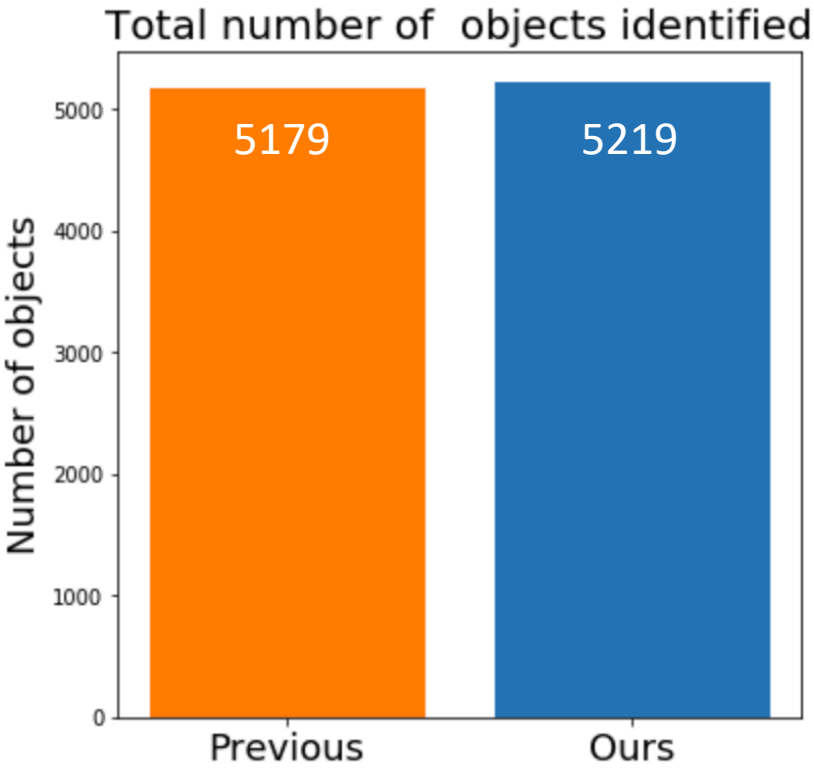
Our work: walk into the kitchen and turn right . walk past the table and into the living room . stop in front of the fireplace .

RWBC-OIL

OKAZAKILAB

Results 3/3





Even when the total number of objects detected is similar in both experiments our approach identifies almost 54% more object types.

Results 4/4

- The previous result was calculated as follow.

1. Identify the words that represent objects in the environment and vocabulary.
2. Extract the sentences from both experiments that contain those words.
3. The total number of identifier objects was calculated here.
4. Remove the sentences that doesn't make sense
5. Count the number of types of objects in the remaining sentences.

Conclusions

- The proposed method seems to be as good as the previous model in terms of BLEU score.

- Qualitative analysis indicates that the generated sentences make more use of image features to select the reference object. *

Model	Val_seen	Val_unseen
Speaker	28.3	**27.5**
Ours	**28.5**	27.0

BLEU score comparison, on Validation seen and Validation unseen datasets. Higher is better.

* The comparison was performed using the Validation unseen set.

References

[1] Peter Anderson et al. “Vision-and-Language Navigation: Interpreting visually-grounded navigation instructions in real environments”. In: CVPR. 2018.

[2] Angel Chang et al. “Matterport3D: Learning from RGB-D Data in Indoor Environments”. In:3DV. 2017.

[3] Jia Deng et al. “ImageNet: A Large-Scale Hierarchical Image Database”. In: CVPR. 2009.

[4] Fried et al. “Speaker-Follower Models for Vision-and-Language Navigation”. In: NeurIPS. 2018.

[5] Kaiming He et al. “Deep Residual Learning for Image Recognition”. In: CVPR. 2015.

[6] Kishore Papineni et al. “BLEU: a Method for Automatic Evaluation of Machine Translation”. In: ACL. 2002.

References

[7] Jeffrey Pennington, Richard Socher, and Christopher D.Manning. “Glove: Global vectors for word representation”. In: EMNLP. 2014.

[8] anonymous - under review. “Self-Monitoring Navigation Agent via Auxiliary Progress Estimation”. In: ICLR. 2019.

[9] Ashish Vaswani et al. “Attention Is All You Need”. In: NeurIPS. 2017.

[10] Xin Wang et al. “Look Before You Leap: Bridging Model-Free and Model-Based Reinforcement Learning for Planned-Ahead Vision-and-Language Navigation”. In: ECCV. 2018.

[11] Huanyu Yu et al. “Fine-grained Video Captioning for Sports Narrative”. In: CVPR. 2018.

[12] Kiela et al. “Learning Visually Grounded Sentence Representations”. In: NAACL. 2018.