UC San Diego

Vision-and-Language Navigation: Interpreting visually-grounded navigation instructions in real environments

Presenter: Quan

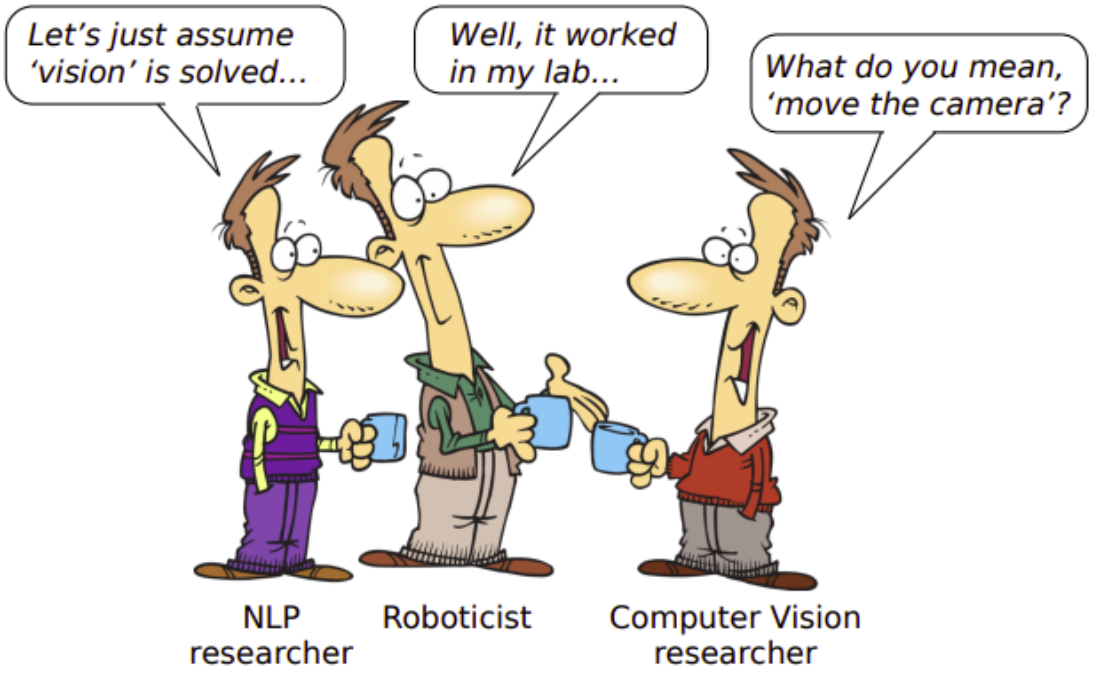
June 02

Outline

- Introduction
- Related work
- Method
- Experiments
- Conclusion

Introduction

- Connect language and vision to actions.
- Recent availability of 3D reconstructions at large scale is an enabler for research on embodied agents.
- Timely to refocus on the intersection of computer vision, NLP and robotics.



Introduction

Example task of vision-and-language navigation

Given a natural language navigation instruction, navigate through a real environment to find the goal location.

Instruction: Head upstairs and walk past the piano through an archway directly in front. Turn right when the hallway ends at pictures and table. Wait by the moose antlers hanging on the wall.

Related works

Image Captioning (Chen et. al. 2015)

A man in a pink bow tie and a pink shirt is being hugged by a man in a blue shirt.

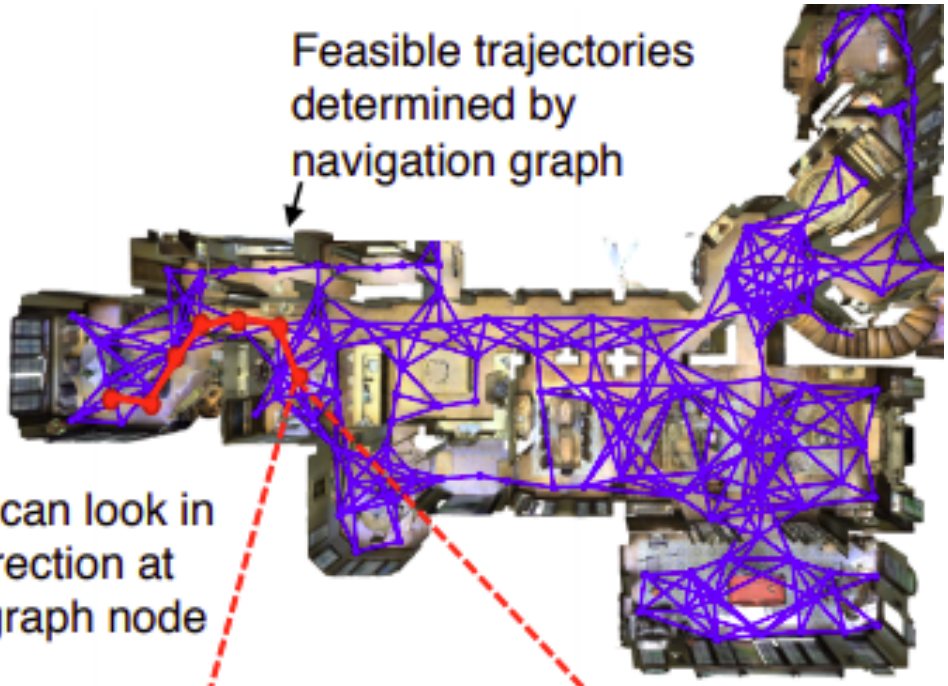
Related works

- Many language and vision grounding tasks.
- But they are passive tasks.
- The agent is not allowed to move the camera

Matterport3D Simulator

- Simulator for embodied visual agents, based on the Matterport3D dataset[5] containing:
  - 10,800 panoramas
  - 90 diverse buildings
- Discrete motion but with continuous camera control and real images.

[5]Chang *et al.* 3DV, 2017



Room-to-Room (R2R) navigation dataset

Data Collection[6]:

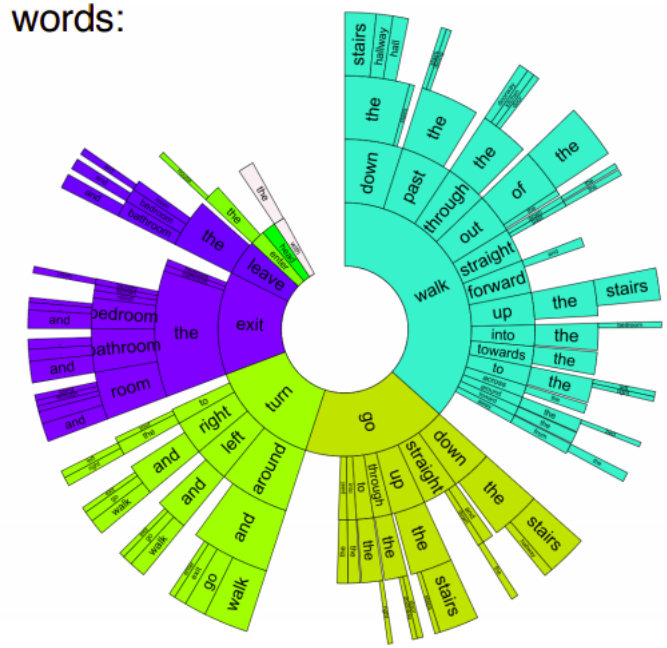
- Sampled 7,189 shortest paths between locations (mostly) in different rooms.
- Collected 21,567 navigation instructions (3 per path) using crowd workers and a WebGL interface (1,600 hours).

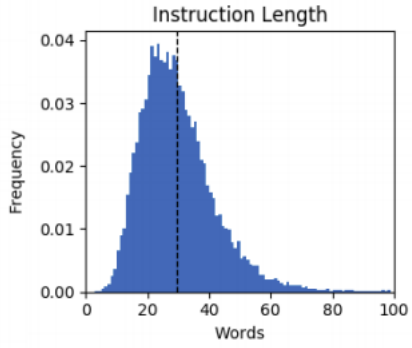
Environment splits:

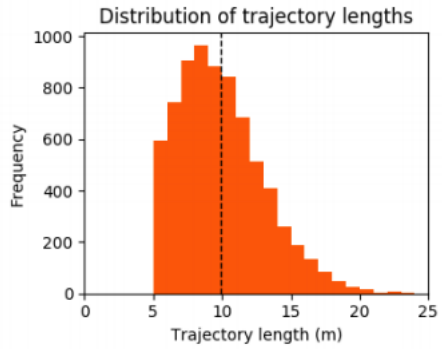
- 61 training / val-seen, 11 val-unseen, 18 test (unseen).

Room-to-Room (R2R) navigation dataset

Distribution of navigation instructions based on their first words:







Examples of new vocabulary encountered in unseen environments:

hieroglyphs

Squiggle painting

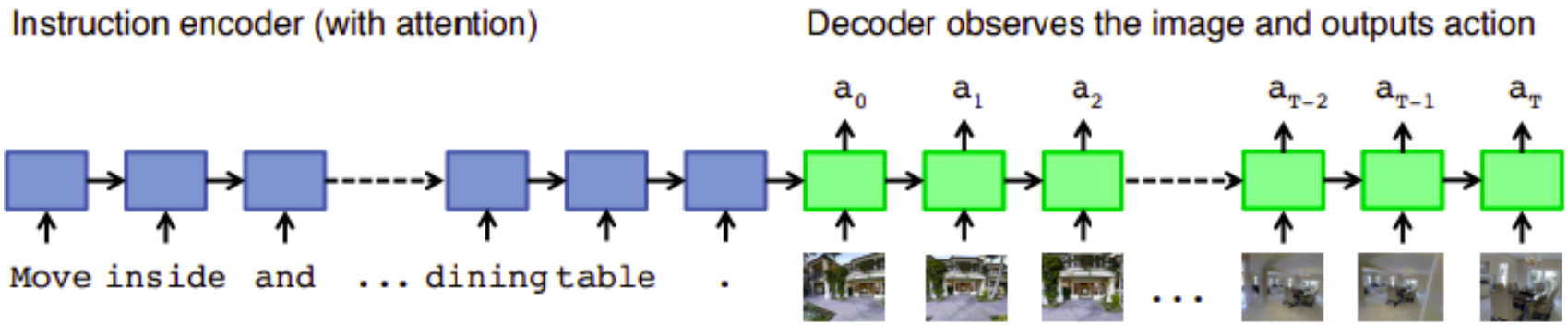
mannequins

teapot

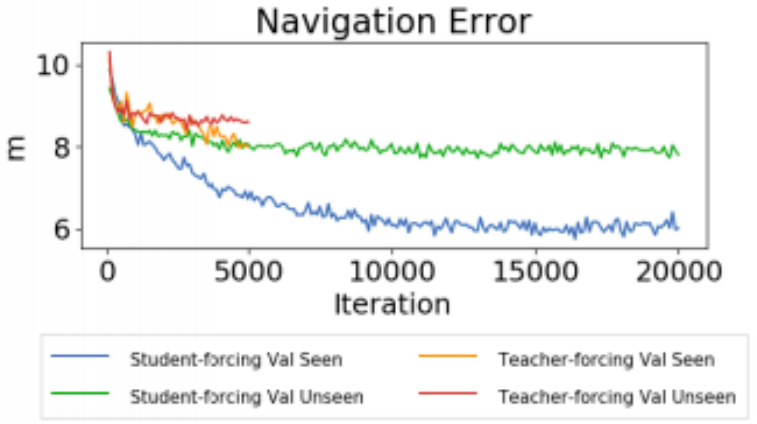
Example of room to room navigation

Go indoors. Go past the wall with holes in it. Go past the large table with chairs. Turn right and wait there.

Baseline Seq2Seq Model



- LSTM-based Seq2Seq baseline model outputting a distribution over 6 actions: `left`, `right`, `up`, `down`, `forward` & `stop`.
- Image features from ResNet-152.
- Training with 'student-forcing' (sampling the next action) outperforms 'teacher-forcing' (selecting the ground-truth action).



The model is trained with Cross Entropy to mimick an oracle agent

Evaluation

Clear Evaluation Protocol:

- Report navigation error (distance from goal) for each instruction in the unseen test environments.
- 'Success' when navigation error < 3m.
- Agent must choose to stop (also report success rate with oracle stopping).