ShowNTell: Visual Attention Prediction for Robot Spatial Arrangement Tasks

ANNOTATIONS

We thought of two different ways to formulate the problem:

Paragraphs for Problem Statement version A

Paragraphs for Problem Statement version B

Paragraphs to write

I. INTRODUCTION

Motivations For robots to live and work with us, we want them to act out what they hear according to what they see. However, many instructions involving spatial references are imprecise. For example, when asked to "put the coffee cup on the right side of the plate," how much right to the plate should the robot move to? How about "on the right side of the laptop?" Empirically, the areas that are considered appropriate for the two reference objects should be different. The sets should depend on the task and domain specified by the instruction, as well as the physical configuration and limitation indicated in the visual observation. We consider this language-directed visual prediction fundamental to robot spatial arrangement tasks.

Objective and scope [A] In this paper, we argue that both natural language instructions and demonstration image sequences are essential in developing the robot's visual prediction ability for following instructions. Instructions are not only the human way of communicating intentions but also provide context and constraints for robots to predict visual attention more accurately. Visual observations are not only the background for target objects, but they also provide other objects and space as factors to affect the attention prediction.

Objective and scope [B] In this paper, we argue that infusing prior knowledge about spatial relationships makes it more data-efficient to learn visual prediction ability. If we assume spatial relationships are learned, learning to predict visual attention based on spatial instructions becomes fine-tuning the prediction to specific tasks and domains.

Proposed model We present ShowNTell, a neural network model that predicts visual attention given natural language instructions in spatial arrangement tasks. Much like Neural Module Networks [1], ShowNTell dynamically lay out a deep network according to linguistic triples that are translated from natural language instructions. The network takes in the image before the action indicated by the natural language instruction is performed, and outputs an attention map that marks the visual goal for the action. The regions with the highest probability on the attention map and the linguistic triples are used by the robot to execute the instructions. The modules of the network are reusable and are trained together end-to-end.

Technical novelty In contrast to the sequence-to-sequence models used in instruction grounding and unlike [1] that uses parse tree to represent semantics, our model uses linguistic triples as structured representations of sentence semantics. Triples with the form `<subject relation object>` are obtained using the START parser, enabling important spatial relations and object properties be singled out for encoding prior knowledge about spatial relationships and references shared between instructions.

Implementation and evaluation [A] We train ShowNTell in the task of setting dinner tables. The training images show the top-down view of the table. The instructions are provided by humans who have little prior knowledge in table-setting and are complex with sequences and clauses. ShowNTell can successfully apply the visual prediction ability to unseen objects such as in setting up a tea ceremony. It can also transfer the ability to other domains such as setting up an office table with little training data.

Implementation and evaluation [B] We train ShowNTell in the task of table-setting. The training images show a top-down view of the table. First, we train the model with images and instructions for learning spatial relationships such as "left" and "between." Then we use table-setting image sequences and instructions provided by humans with little prior knowledge in table-setting. Compared to the model that is directly trained from scratch, our ShowNTell model can achieve the same performance with XX% fewer data. Moreover, ShowNTell model has better accuracy in applying the visual prediction ability to unseen objects such as in setting up a tea ceremony. It can also transfer the ability to other domains such as

setting up an office table with less training data.

Contributions The contributions of this project include:

- On the task level, we designed a model that predicts visual attention on visual observations given natural language instructions. The model is object general and easy to transfer to other domains.
- On the technical level, we designed a neural network model that combines the structured symbolic representation of semantics and data-driven approach for learning a robust system.
- We collected a data set of instructions with corresponding visual demonstrations for the table-setting task.
- Implemented a simulated robot arm that can properly set up tables according to different sequences of user instructions, as judged by humans.

II. Related Work

Past works and gap Mapping instructions and visual observations to actions have been explored in the past two years [2]–[10]. However, most of projects focus on blocks world task in 2D [5], [8]–[10] or 3D [4], [6] that involve grids and squares. Among those tasks that involve simulated real objects [2], [3], [7], none have focused on spatial arrangement tasks but navigation tasks, with [2] and [7] not concerning about spatial references.

A. Visual Goal Prediction

B. Grounding Spatial Relations

C. Neural Modular Networks

III. Problem Formulation

Technical problem The ability to predict visual goals can be formulated as a mapping from a natural language instruction to a region in visual observation. Formally, each training datum can be thought of as a 3-tuple (x, W, y):

- x: the input natural language instruction
- W: the input visual observation of the world before x is followed
- y: the output visual goal, the most likely region referred in the instruction x.

A. Table-Setting Task

Task The goal of a table-setting task is to evaluate how well an agent can execute sequences of spatial arrangement instructions to assimilate how humans might execute them. The task is evaluated by a subjective judgement score of how the arrangement is proper.

Task decomposition Setting up a table may involve 8-15 objects, thus a similar number of steps. Each step is likely to involve at least one reference object and a spatial relationship. Figure 1 shows two sample final arrangements and the sequence of objects added according to two online video tutorials.

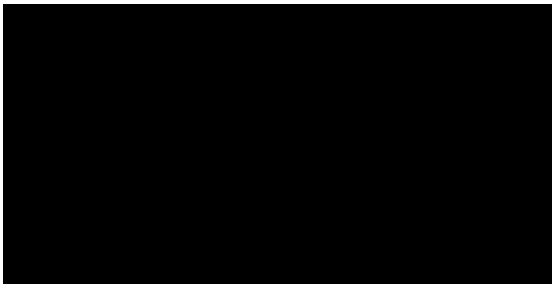

Fig. 1: Two different arrangements of the table-setting tasks.

B. Table-Setting Dataset

The Table-Setting Dataset we collected for this paper contains both instruction and visual observation data in a simulated table environment powered by Unity. Assumptions The table is observed by top-down orthogonal camera, thus 3D spatial relations like *behind* are not considered and *above* is interpreted in 2D plane.

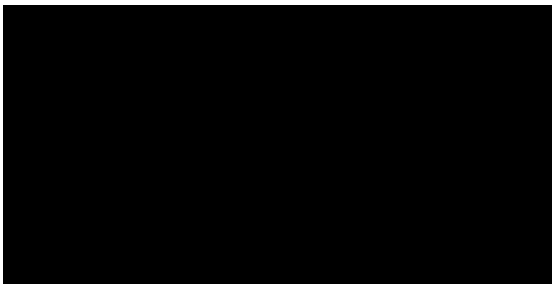

Fig. 2: Crowdsourced instructions for one table-setting step.

Collect x data To train the model with real instructions that humans naturally use to describe the table-setting steps, we first crowdsource sequences of instructions through Amazon Mechanical Turk. The instruction workers were given pairs of screenshots, which are the before and after scenes of an action, and were asked to describe the actions in terms of natural language imperatives. The screenshots came from five different online table-setting tutorial videos. The objects in the vicinity of the reference region are labeled. This collection results in 100 caption-like sequences of instructions for each of the five table-setting videos, including approximately 5000 instructions x. Figure 2 shows a list of sample instructions collected from one pair of before and after scenes.

Collect W data Then, to increase the variation of demonstration table-setting visual observations, each sequence of instructions was used to instruct about 100 human workers to set up the table by dragging and rotating objects on a webpage. To improve the quality of data, human workers are shown example table-setting photos as a training process for them to be table-setting experts.

After these two rounds of data collection, we have 500,000 3-tuples for training. Table I shows the dataset statistics and Table II presents the qualitative analysis of the instructions.

TABLE I: Summary statistics of the instruction data

Dataset Statistic	Count
Number of instructions	
Average number of instructions for each video	
Average length of instructions	
Average number of actions per instruction	

TABLE II: Qualitative analysis of Table-Setting instructions. (The count indicates the occurrences in a sample of 200 instructions)

Category	Count	Example
Spatial relations with one object		
Spatial relations with two objects		
Conjunctions of two spatial relations		
Co-reference		
Comparatives		

IV. ShowNTell for Visual Goal Prediction

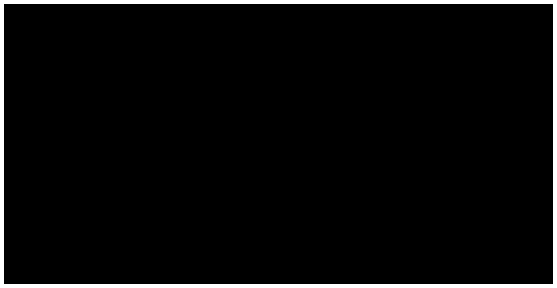

Fig. 3: A schematic representation of our proposed model

We propose the neural modular networks for visual goal prediction (Figure 3). In this section, we will use instruction *put the fork on the right of the plate* to explain the architecture.

A. Modules

B. From images to embeddings

C. From instruction to triples

D. From triples to networks

E. Predicting visual goals

F. From predictions to actions

V. Experiment Setup

A. Implementation details

Our model was trained end-to-end using PyTorch with a batch size of 32. The visual observation W has dimension 224 $\times$ 224 $\times$ 3 (height, width, channel). The first coevolutionary layer uses a filter of size 5 $\times$ 5 and the second of size 3 $\times$ 5, each followed by a `tanh`. The final prediction layer is a 56 $\times$ 56 filter that ...

B. Baseline Approaches

We compare our model with the following models:

1) Without linguistic triples: By replacing linguistic triples with LSTM encodings, ...

2) Without modules: Using Compositional Attention Networks [11]

VI. Evaluation

A. Interpretability

1) Visualizing processes:

2) Comparing linguistic paraphrases:

B. Example Pheonomena

In the following example, three instructions for the same step are used to test the model.

VII. Expected Partial/Final Outcomes

1) **Table-Setting instruction dataset**: 1,000 sequences of natural language instructions that describe how to set a dinner table.
2) **Table-Setting demonstration dataset**: 1,000,000 sequences of pictures showing the steps in setting a dinner table.
3) **Instruction parser**: an LSTM-based language module that takes in natural language instructions and generates linguistic triples.
4) **Visual identifier**: a CNN-based vision module that takes in the scene, masks of the objects, and linguistic triples, then output a map of likelihood

References

[1] J. Andreas, M. Rohrbach, T. Darrell, and D. Klein, "Neural module networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 39–48.

[2] D. Misra, A. Bennett, V. Blukis, E. Niklasson, M. Shatkhin, and Y. Artzi, "Mapping instructions to actions in 3d environments with visual goal prediction," *arXiv preprint arXiv:1809.00786*, 2018.

[3] V. Blukis, N. Brukhim, A. Bennett, R. A. Knepper, and Y. Artzi, "Following high-level navigation instructions on a simulated quadcopter with imitation learning," *arXiv preprint arXiv:1806.00047*, 2018.

[4] N. Kitaev and D. Klein, "Where is misty? interpreting spatial descriptors by modeling regions in space," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 2017, pp. 157–166.

[5] D. Misra, J. Langford, and Y. Artzi, "Mapping instructions and visual observations to actions with reinforcement learning," *arXiv preprint arXiv:1704.08795*, 2017.

[6] Y. Bisk, K. J. Shih, Y. Choi, and D. Marcu, "Learning interpretable spatial operations in a rich 3d blocks world," in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[7] D. S. Chaplot, K. M. Sathyendra, R. K. Pasumarthi, D. Rajagopal, and R. Salakhutdinov, "Gated-attention architectures for task-oriented language grounding," in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[8] M. Janner, K. Narasimhan, and R. Barzilay, "Representation learning for grounded spatial reasoning," *Transactions of the Association of Computational Linguistics*, vol. 6, pp. 49–61, 2018.

[9] W. Xiong, X. Guo, M. Yu, S. Chang, B. Zhou, and W. Y. Wang, "Scheduled policy optimization for natural language communication with intelligent agents," *arXiv preprint arXiv:1806.06187*, 2018.

[10] A. Sinha, M. Sarkar, B. Krishnamurthy *et al.*, "Attention based natural language grounding by navigating virtual environment," *arXiv preprint arXiv:1804.08454*, 2018.

[11] D. A. Hudson and C. D. Manning, "Compositional attention networks for machine reasoning," *arXiv preprint arXiv:1803.03067*, 2018.