Generating Natural Language Descriptions of Trajectories Using Long Short Term Memory Neural Networks

Rodolfo Corona and Rolando Fernandez

I. Problem Description

Given a point-cloud $p \in P$ and a manipulation trajectory $t \in T$, our goal is to output a free-form Natural Language (NL) description $l \in L$ that describes the trajectory t:

$$f : T \times P \mapsto L \tag{1}$$

II. Motivation

Currently there is not much research in the area of Explainable Artificial Intelligence (XAI), an area of AI that aims at creating systems that allow for an agent's actions to be understood by a human user. Lomas et al., discusses how giving an agent the ability to explain it's actions would help human users gain trust for the actions taken by an agent [1].

Our goal is to create a system that allows an agent to explain the actions it will take or that need to be performed to complete a given task, something that would allow for better cooperation between the agents and human users, while at the same time allowing the human users to better understand the intentions of the agent.

III. Hypothesis

Given $(t, p) \in T \times P$, a Long Short Term Memory (LSTM) Neural Network Architecture may be trained to generate NL descriptions that accurately describe the actions the agent performs under a trajectory $t \in T$.

IV. Methods

A. Dataset

We propose to use the Robobarista data set, which contains 116 point clouds of objects, and 250 natural language descriptions of 1225 trajectories. The set will also contain embeddings which were trained to map all three modalities (T, P, and L) into a common space. Specifically, these embeddings were generated for singular trajectories as well as for point cloud and description pairs [2].

B. Baseline

For a baseline generative model, we propose to take the inputted pair $(t, p) \in T \times P$ and find the k nearest neighbor pairs $(p', l') \in P \times L$ of t in the training set within the shared embedding space. These k-nearest neighbors will then be re-ranked based on how similar their corresponding point cloud p' is to p. This similarity will be measured by comparing bag-of-keypoint vectors generated for p and p' using NARF [3] descriptors with a method analogous to [4].

C. Contribution

Long Short Term Memory networks (LSTMs) have been shown to be able serve as generative models for text [5]. Therefore, we would like to train an LSTM to generate a description $l \in L$ given a pair $(t, p) \in T \times P$ through methods inspired by [6].

V. Evaluation

We propose quantitative and qualitative criteria under which to evaluate the correctness of a generated phrase against a ground truth description.

A. Quantitative

For automatic evaluation, we will explore metrics such as Levenshtein distance and the cosine similarity of the description's skip-thought vectors [7].

B. *Qualitative*

To qualitatively test our proposed system we plan on using a human rating system, where human participants will compare the NL description generated by the system against that of the NL description in the ground truth. The participants will then rate the NL description on a scale by judging how semantically similar the generated description is to the ground truth.

REFERENCES

[1] M. Lomas, R. Chevalier, E. V. Cross II, R. C. Garrett, J. Hoare, and M. Kopack, "Explaining robot actions," in *Proceedings of the seventh annual ACM/IEEE international conference on Human-Robot Interaction*. ACM, 2012, pp. 187–188.

[2] J. Sung, S. H. Jin, I. Lenz, and A. Saxena, "Robobarista: Learning to manipulate novel objects via deep multimodal embedding," *arXiv preprint arXiv:1601.02705*, 2016.

[3] B. Steder, R. B. Rusu, K. Konolige, and W. Burgard, "Narf: 3d range image features for object recognition," in *Workshop on Defining and Solving Realistic Perception Problems in Personal Robotics at the IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS)*, vol. 44, 2010.

[4] G. Csurka, C. Dance, L. Fan, J. Willamowski, and C. Bray, "Visual categorization with bags of keypoints," in *Workshop on statistical learning in computer vision, ECCV*, vol. 1, no. 1-22. Prague, 2004, pp. 1–2.

[5] A. Graves, "Generating sequences with recurrent neural networks," *arXiv preprint arXiv:1308.0850*, 2013.

[6] S. Venugopalan, M. Rohrbach, J. Donahue, R. Mooney, T. Darrell, and K. Saenko, "Sequence to sequence - video to text," in *The IEEE International Conference on Computer Vision (ICCV)*, December 2015.

[7] R. Kiros, Y. Zhu, R. R. Salakhutdinov, R. Zemel, R. Urtasun, A. Torralba, and S. Fidler, "Skip-thought vectors," in *Advances in Neural Information Processing Systems 28*, C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, Eds. Curran Associates, Inc., 2015, pp. 3294–3302. [Online]. Available: http://papers.nips.cc/paper/5950-skip-thought-vectors.pdf