CPSC 490 Project Proposal: Learning to Orient Towards the Focus of Attention of a Group Conversation Using Variational Auto-encoders

Sally Ma - Advised by Marynel Vázquez

January 2020

1 Introduction

Social robots often need to operate with and around groups of people, including partaking in group conversations. Past work on controlling robots' spatial behaviors in human conversations includes relying on tele-operations [6], [4] and generating autonomous robot motion through rule-based approaches [7] or generative models of proxemic behavior [3]. Rule-based approaches tend to generalize poorly to new situations, and generative models have not been tested with adapting a robot's position with respect to more than one person. Recognizing the importance of a robot to turn towards the speaker in a group conversation as a subtle and effective strategy of conveying attentiveness to the conversation and redirecting the focus of a conversation, Vázquez et al. evaluate using reinforcement learning (RL) to control robot orientation in a simulated environment that they built based on models from social psychology explaining spatial behavior during group conversations. They find that model-based RL agents are capable of learning good policies within a few minutes of interaction time [5].

Inspired by the potential of Variational Autoencoders (VAEs) in generating complex data [1], we at the Interactive Machines Group intend to explore using VAEs to allow the robot to learn motion policies in orienting towards the speaker of a conversation. Built on top of standard function approximators (neural networks) and able to be trained with stochastic gradient descent, VAEs emerge as an appealing approach to unsupervised learning of complicated distributions. In comparison with standard autoencoders, which are discriminative, VAEs are generative models that learn a low-dimensional latent representation of the input. This latent representation is parameterized by Gaussian means and standard deviations, which correspond to a set of Gaussian distributions that can be sampled for generating output. Because of this architecture, VAEs are effective models for encoding sequential data, allowing them to perform tasks such as disentangling dynamic features from static features in video recordings [2]. VAEs thus demonstrate potential in disentangling complex social contexts for a robot, allowing the robot to understand and synthesize sequential context clues from its surroundings and better respond to changing social dynamics.

2 Project

This project investigates the effectiveness of using VAEs to compress high-dimensional observations, from which a robot could learn motion policies in orienting towards the speaker of a conversation. The main goals of the project are:

1. Understand how to make the VAEs work in our specific task;
2. Analyze the tradeoffs for hyperparameters;
3. Compare VAEs with more direct, traditional methods such as mapping high-dimensional observations directly to actions.

This project is the senior project in partial fulfillment for the Combined Bachelor of Science and Master of Science program in Computer Science.

2.1 Experiments

For the experiments I plan to conduct in order to evaluate model performance, I will use a simulator that Professor Vázquez implemented for simulating group conversations [5]. Figure 1 presents an example of the context of the simulation, where people maintain distinct spatial organizations known as F-formations (short for face-formations) while standing freely in open, public spaces. Face formations begin when the members of a group position themselves such that their transnational segments (segments that extend in front of each person) intersect, the intersection of which is known as the o-space of the F-formation. In the simulation, the robot and the people in its conversation form a circular F-formation.

We concentrate the evaluation on situations where the users are the active speakers and the focus of attention. These situations are more interesting to study than their counterparts because the robot does not have control of the interaction dynamics and must adapt to the flow of the conversation. Thus in our simulation, the robot will not be allowed to speak.

2.2 Models

I intend to start with the simplest, most generic VAE model, and make it function in our use case. Once it works for our task, I might explore more complex VAE models such as BetaVAE and the one proposed by Li et al. [2], which learns a latent representation of data split into a static and dynamic part and thus enforces the disentanglement of dynamic and static features.

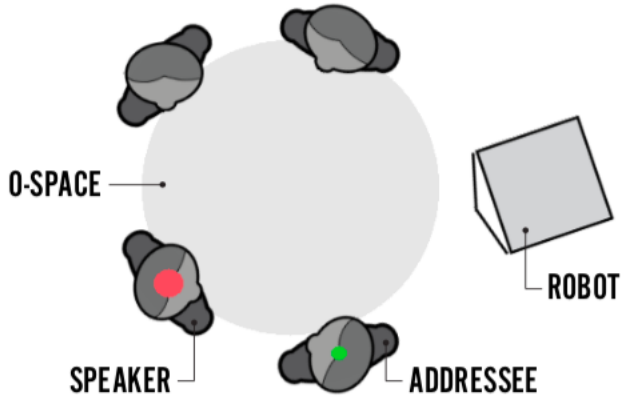

Figure 1: A simulated group conversation between a robot and four people. The red and green circles on top of the agents identify the speaker and addresses, respectively. The big gray circle represents the o-space of the group's face-formation

3 Deliverables

At the end of this project, I will provide an abstract, the source code, and a final report summarizing my work and findings, which will include a comparative analysis of how effectively VAEs and more direct, traditional methods work in our use case.

References

[1] Carl Doersch. Tutorial on variational autoencoders. *ArXiv*, abs/1606.05908, 2016.

[2] Yingzhen Li and Stephan Mandt. Disentangled sequential autoencoder. In *ICML*, 2018.

[3] Ross Mead and Maja Mataric. *Perceptual Models of Human-Robot Proxemics*, volume 109, pages 261–276. 01 2016.

[4] J. Vroon, M. Joosse, M. Lohse, J. Kolkmeier, J. Kim, K. Truong, G. Englebienne, D. Heylen, and V. Evers. Dynamics of social positioning patterns in group-robot interactions. In *2015 24th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, pages 394–399, Aug 2015.

[5] Marynel Vázquez, Aaron Steinfeld, and Scott Hudson. Maintaining awareness of the focus of attention of a conversation: A robot-centric reinforcement learning approach. pages 36–43, 08 2016.

[6] Marynel Vázquez, Aaron Steinfeld, Scott E. Hudson, and Jodi Forlizzi. Spatial and other social engagement cues in a child-robot interaction: Effects of a sidekick. In *Proceedings of the 2014 ACM/IEEE International Conference on Human-Robot Interaction*, HRI '14, page 391–398, New York, NY, USA, 2014. Association for Computing Machinery.

[7] Mohammad Abu Yousuf, Yoshinori Kobayashi, Yoshinori Kuno, Akiko Yamazaki, and Keiichi Yamazaki. Development of a mobile museum guide robot that can configure spatial formation with visitors. In De-Shuang Huang, Changjun Jiang, Vitoantonio Bevilacqua, and Juan Carlos Figueroa, editors, *Intelligent Computing Technology*, pages 423–432, Berlin, Heidelberg, 2012. Springer Berlin Heidelberg.