Learning Multiparty & Multimodal Social Interactions

Yan Pan
Advisors: Paul Liang, Louis-Philippe Morency
Carnegie Mellon University

1 Abstract

As intelligent systems become widespread in the real world, the area of **artificial social intelligence** has become a prominent research area. In order for artificial intelligence (AI) to understand social interactions, the ability to understand multimodal interactions is required. The purpose of this research is to study the learning of **multiparty & multimodal social interactions**, with a particular focus on multimodal dialogs. We propose a new standard of evaluation benchmarks to evaluate the realism of interactive social AI in three different stages: **imitation**, **self-play**, and **interaction**. With our benchmark for interactive social AI, we will build methods based on multimodal multitask learning to build interactive social AI.

2 Objectives and Contribution

The fundamental goal of **artificial social intelligence** is to build socially intelligent AI that can comprehend human social cues, intents, and affective states, engage in social conversation, and understand social norms and commonsense in order to maintain a rich level of interpersonal interaction with humans. Progress in real-world social AI would bring about real-world advances in human sensing and robot design, with the end goal of engaging people through social and physical interactions [3], monitoring human behavior to understand and predict the types of help people need [1, 2], and offering assistance in schools, hospitals, and the workplace [4].

This research aims to study how AI can learn **multiparty & multimodal social interactions**, specifically in the setting of multi-person multimodal dialogs. In multimodal dialogs, various speakers communicate with each other and adjust their behaviors based on multiple sensory modalities, including speech, gestures, and facial expressions. The goal for the AI is to perceive and process the multimodal information demonstrated by humans and interact with them. Multimodal dialogs can be viewed as the combination of multiple microtasks that describe a subset of social interactions, including dialog generation, question answering, sentiment analysis, etc. Current work has focused on text-based dialog without building benchmarks and models that study the integration and generation of social cues important to social interactions. Existing benchmarks lack the component of interactivity and focus primarily on supervised tasks. To address this problem, we first propose a new standard of evaluation benchmarks for interactive social intelligence through 3 steps:

1. imitating human response in a supervised learning manner;
2. interact with itself in a simulated environment to strengthen the learned knowledge in imitation;
3. interacts with human in an online setting to collect online data and feedback.

With this evaluation benchmark, we hope to devise learning algorithms that can learn in interactive multi-person multimodal settings and take a major step forward in constructing socially intelligent AI capable of interacting with humans.

3 Methodology

3.1 Project Design and Feasibility

Any AI model can only be as good as the metrics used to evaluate it. Therefore, we plan to **collect realistic interactive benchmarks to better evaluate real-world social AI**. Existing benchmarks lack the component of interactivity and focus primarily on supervised tasks. Instead, we envision a new series

Stage	Agents	Data collection	Learning algorithm	Evaluation metric
1. Imitation	AI	Human demonstration	Supervised learning	Generation likelihood
2. Self-play	AI with AI	Simulated environment	Reinforcement learning	Cumulative reward & Generation likelihood
3. Interaction	AI with Human	Simulated environment & Human-in-the-loop labeling	Reinforcement learning & Active learning	Human judgement & Generation likelihood

Table 1: We envision a new standard of evaluation benchmarks that increasingly assess the realism of interactive social AI across imitation, self-play, and interaction stages, each building on top of the previous stage. We also plan to collect more realistic interactive benchmarks that better represent real-world social AI in dialogue, robotics, and healthcare.

of evaluation **microtasks** that each specializes in a subset of social interactions. Each of these microtasks increasingly assesses the realism of social intelligence, categorized according to the agents involved, data collection process, learning algorithms, and evaluation metrics as illustrated in Table 1. Stage 1, the **imitation** stage, tests whether AI is able to imitate humans in social settings. Most of the existing work in supervised affect recognition and dialog modeling falls under this category. Stage 2, the **self-play** stage, tests whether AI is able to interactively engage with itself. Stage 3, the **interactive** stage, tests whether AI is able to engage in interactive social communication with a real human. We aim to leverage human-in-the-loop learning and active learning to provide useful human labels in an interactive multimodal setting. The culmination of these 3 stages will more accurately benchmark the interactive capabilities of social AI and uncover the shortcomings of existing models.

3.1.1 Stage 1: Imitation learning from offline data

In stage 1, we plan to perform our baseline experiment on a multi-person multimodal egocentric communication dataset with 38.5 hours of conversation videos recorded from a first-person perspective [6]. We plan to train several independent modules on the existing annotations of the dataset to learn specific downstream tasks. The tasks include dialog response prediction, predicting and generating informative responses with conversation history; speaker identification, identifying and predicting the current and next speaker given the multimodal information of participants; and laugh prediction, predicting whether the participants will laugh in a short period of time based on current conversation content. Besides, we also plan to collect more tasks from the dataset through further annotation, including two types of question answering, social-related and physical-related questions, in order to probe AI's understanding of human interactions. Specifically, the physical questions are based on summarizing AI's multimodal observations of the world, including object properties and spatial-temporal relationships, while the social questions aim to facilitate the perception of emotions, feelings, and personalities of the participants.

3.1.2 Stage 2 and 3: Learning from online interactions

In stages 2 and 3, we focus on building models that can learn to participate in multimodal interactions based on the downstream tasks learned in the previous stage through self-playing and eventually become able to interact with real humans and learn from the process. In these stages, our goal is to build and train a model based on the microtasks that is able to engage in interactive social communications. Our proposed solution is **multimodal multitask learning**, in which representation is shared across multiple tasks. We pretrain a series of modules on task-specific data and then incorporates shared representation across the modules, so they are able to share information between them. The modules are constructed in a hierarchical manner, where lower levels process more general information and higher levels process more task-specific information. The model can be trained with reinforcement learning in a simulated environment of the multimodal interaction and using active learning in interactions with real humans. We plan to clearly define the reinforcement learning problem mathematically and design the details of the model that incorporates the pretrained modules and capable of engaging in multimodal interactions with humans.

We plan to evaluate our results in both stages quantitatively. For single components of the interaction,

the result can be evaluated using quantitative automatic metrics, such as BLEU [7] and perplexity [5]. In the multimodal interaction setting, we can perform human evaluations in an interactive environment to determine the relevance and informativeness of the AI's responses and behaviors.

3.2 Preliminary Results

Our preliminary result includes analysis and preprocessing of the dataset and training independent models to learn the specific downstream tasks. We trained a DialoGPT [8] model on the transcripts of the dataset to predict the next sentence in the conversation. With the conversation history as input, the model is able to generate responses to the current conversation. We also analyzed the application of several other downstream tasks on the dataset, including speaker identification, laugh prediction, and question answering that we previously mentioned. We conclude that it is possible to perform these microtasks on the existing annotations of the dataset effectively with independent modules.

3.3 Project Time Table

May 15 - June 7: Stage 1: Imitation, which includes:

1. Conduct experiments of the downstream tasks on existing annotations (1 week);
2. Annotate the question answering tasks on the dataset (1 week);
3. Conduct experiments of question answering on the annotated dataset (1 week).

June 1 - August 14: Stage 2 & 3: Self-play and interaction, which includes:

1. Build the simulation environment of multimodal interaction (1 week);
2. Design the model (3 weeks);
3. Conduct experiments and evaluations of the model (2 weeks).

4 Background

I am currently a sophomore student at Carnegie Mellon University majoring in computer science. I assert that I satisfy the necessary requirements to work on the research project. I have completed coursework in Ph.D. level machine learning and optimization (10-701, 10-725), computer vision (16-385), probability and statistics (21-325 and 36-226), mathematics (21-355). I have relevant experience in applying machine learning methods and writing research papers during my research experience at Peking University during high school. I have worked on several machine learning and data science projects since high school, including my project for the Ph.D. level machine learning class 10-701 which received top grade in the course. I am familiar with the essential tools for machine learning, including Python and PyTorch, and the major machine learning models and techniques. I am also familiar with algorithms and mathematics and have relevant competitive programming and competitive math backgrounds. I am currently working with the advisors at MultiComp Lab and familiar with the research topics.

5 Feedback and Evaluation

I will work closely with Dr. Louis-Philippe Morency and PhD student Paul Liang for this project. I will check in twice a week with Paul where we will discuss technical details of the project, experiments, and results, and provide feedback and evaluations. I will also have monthly one-on-one meetings with Dr. Morency to discuss high-level progress and future directions. I will also attend our research group's weekly group meetings and share my progress with other students in the group.

6 Dissemination of Knowledge

I plan to communicate our ideas and results with the research community through publications in open-access journals and conferences. I also plan to package the source code and pretrained models for release to the community. At Carnegie Mellon University, I will present our result in the Meetings of the Minds and communicate with other faculties and students.

References

[1] Paul Pu Liang, Ziyin Liu, Amir Zadeh, and Louis-Philippe Morency. Multimodal language analysis with recurrent multistage fusion, 2018.

[2] Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. Multimodal local-global ranking fusion for emotion recognition. In *Proceedings of the 20th ACM International Conference on Multimodal Interaction*, ICMI '18, page 472–476, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450356923. doi: 10.1145/3242969.3243019. URL https://doi.org/10.1145/3242969.3243019.

[3] Paul Pu Liang, Jeffrey Chen, Ruslan Salakhutdinov, Louis-Philippe Morency, and Satwik Kottur. On emergent communication in competitive multi-agent teams, 2020.

[4] Paul Pu Liang, Terrance Liu, Liu Ziyin, Nicholas B. Allen, Randy P. Auerbach, David Brent, Ruslan Salakhutdinov, and Louis-Philippe Morency. Think locally, act globally: Federated learning with local and global representations, 2020.

[5] Remi Mir, Bjarke Felbo, Nick Obradovich, and Iyad Rahwan. Evaluating style transfer for text, 2019.

[6] Curtis Northcutt, Shengxin Zha, Steven Lovegrove, and Richard Newcombe. Egocom: A multi-person multi-modal egocentric communications dataset. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[7] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002.

[8] Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and William B Dolan. Dialogpt: Large-scale generative pre-training for conversational response generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 270–278, 2020.