Situated Task-Driven Multimodal Intent

Adriana Kovashka, Malihe Alikhani, Rebecca Hwa, Diane Liman (SCI Computer Science)

Motivation

- In AI, the communicative role of images and text is underexplored, and it is not clear how the goals of a communicative act affect visual-linguistic composition.
- How can a robot help a team of humans perform a task, if it captures the complementary roles of different modalities?
- What cues can a machine perception system use to gauge the role of each communicative modality?
- How does the physical environment in which the interaction takes place correlate with the choice of actions humans take?

Project Description

- We will capture diverse communicative intents in multimodal interactions where two humans, or a human and a web interface, aim to accomplish a predefined task.
- We will collect each communicative act (e.g., speaking, showing), and explanations of the intent of each act, to develop a taxonomy of the relations visual and linguistic acts have to each other and the interaction goal.
- We will use sensing to capture interactions: (a) wearable cameras, (b) sound and speech, (c) proximity, motion and touch sensors.
- We will train models to detect each type of communicative intent for the visual/linguistic modalities, conditioned on the task. We will model intent as an operator and learn separate functions to map data to meaning.

Context

- The visual (e.g., showing, pointing) and language modality (e.g., comparing, giving instructions) have complementary roles. Yet prior work often models these through simple fusion mechanisms such as addition. We argue that in collaborative task completion, the visual and textual modalities may only serve an auxiliary or modification role. We will innovate techniques that learn image representations as a projective transformation of the text (and vice versa).

University of Pittsburgh

How should AI systems support **multimodal task-driven interactions**? How to gauge the role of each **communicative modality**, i.e., naming, pointing, demonstration? How does the **meaning** of each action change with the actor's **goal** and **physical context**?

Project Deliverables

- The output of this project will be:
- (1) A dataset of collaborative human interactions with sensing data (30 sessions of 30 minutes each), and extensive analyses of the relationships between actions that humans take in different modalities, the intent of each individual action, overall goal of the interaction, and the physical properties of the environment.
- (2) Multiple models to infer intent, and compute visual and text representations conditioned on intent.
- (3) A predictive model to infer the human participants' needs in completing their tasks, which can be used by an autonomous assistive agent.

- For (1) an undergraduate student will help set up the environment, using one of the PI's labs as a makeshift "store", "kitchen", "office", Sennott Square, and public outside spaces.
- For (2), a graduate student will develop five techniques to capture intent: (a) An approach that learns separate image and text representation models based on different intents used; (b) An approach like (a) but using sensing data beyond image and text as additional features; (c) An approach to capture the complementary roles of images and text.
- For (3), the graduate and undergraduate students will collaborate to develop an approach that anticipates the human participants' next action. They will then test this approach in two of the application settings we used for data collection, e.g., pretend-doctor play, and information campaign design.

- Following this project, we will apply for external grants (NSF, DARPA, ONR, NIH) to enable more extensive data collection, more innovative intent modeling (based on insights from this project), and more in-depth applications.

- Timeline: (1) in Fall 2022, (2) in Spring 2023, (3) in Summer 2023.

Potential Impact

- This work contextualizes why people speak vs why they show/point, in the environment in which they operate. Our data collection will capture multimodal communication precisely, and train more accurate models for tasks such as image-text retrieval, visual dialog for autonomous assistants, virtual doctor-patient visits, and information campaign analysis and design.