MENDOZA MICCEL
2022320099

Modeling Multimodal Social Interactions: New Challenges and Baselines with Densely Aligned Representations

Sangmin Lee[1] Bolin Lai[2] Fiona Ryan[2] Bikram Boote[1] James M. Rehg[1]
[1]University of Illinois Urbana-Champaign [2]Georgia Institute of Technology
{sangminl,boote,jrehg}@illinois.edu {bolin.lai,fkryan}@gatech.edu

Summary:
This paper addresses the complexity of modeling fine-grained social interactions in multi-party environments by introducing three novel tasks:

1. Speaking Target Identification: Determining the intended listener of an utterance.
2. Pronoun Coreference Resolution: Resolving third-person pronouns to specific individuals.
3. Mentioned Player Prediction: Identifying players explicitly referred to by name.

These tasks are curated and annotated within the domain of social deduction games (e.g., Werewolf, Avalon) using two datasets: YouTube and Ego4D. The authors highlight the limitations of previous work that rely on holistic visual features or focus on single-person behaviors, and propose a novel multimodal baseline that uses densely aligned language-visual representations. This alignment enables synchronized modeling of verbal and non-verbal cues, including gesture, gaze, and spatial positioning.

Extensive experiments demonstrate the effectiveness of their model over unimodal baselines and existing multimodal methods. Ablation studies further show the contribution of gesture, gaze, conversational context, and player permutation learning to model generalization.

Strengths and Weaknesses:

Originality

Strengths

- Introduces three novel tasks in multi-party interaction: speaking target identification, pronoun coreference resolution, and mentioned player prediction.

Weaknesses

- Tasks are limited to social deduction games, raising concerns about generalizability.
- Pronoun coreference is already a well-studied NLP task; the novelty mainly lies in its multimodal framing.

Quality

Strengths

- Evaluation spans two datasets, multiple language models, and several baselines, ensuring methodological rigor.
- High-quality annotations with strong inter-annotator agreement improve dataset reliability.

Weakness

- The method depends heavily on AlphaPose tracking, which may not be robust in real-world, occluded settings.

Clarity

Strengths

- Figures effectively visualize the dense alignment paradigm and model pipeline.
- Qualitative examples demonstrate how multimodal cues help disambiguate referents.

Weakness

- None major, though occasional jargon could still be simplified for broader accessibility.

Significance

Strengths

- Provides benchmarks and a baseline that the community can build on.

Weaknesses

- Reported performance gains are modest in some tasks, limiting practical impact.
- No evaluation of efficiency or real-time feasibility, which is key for interactive applications.

Possible Enhancements:

- Broaden Application Domains: Evaluate the model on general multi-person datasets (e.g., AMI meeting corpus, classroom interactions) to test generalization beyond games.
- Integrate Advanced Multimodal Models: Compare or integrate with newer V+L architectures such as Flamingo or OmniNet for stronger baselines.
- Improve Robustness: Introduce noise (e.g., dropped frames, tracking loss) to test model robustness in more realistic settings.
- Augment Error Analysis: Provide categorized error cases (e.g., visual ambiguity, linguistic vagueness) to guide future model improvements.
- Expand Temporal Modeling: Incorporate long-term interaction modeling across multiple rounds or game phases to test memory and context retention.
- User Study or Human Baseline: Include human performance benchmarks to contextualize the model's capability in referent prediction.