1. Title

Multimodal Signal Processing for Communicative Robots

2. Background

When humans engage in a conversation, we rely on multimodal signal data (i.e. audio, visual, haptic, etc.), from which we can detect and classify classes such as emotion, sentiment, age, gender, etc. All of these are very important information for us to proceed with our conversation.

For example, when you talk to someone who's emotionally angry, you'd try to calm him/her or try to understand the reason behind it. As another example, in some cultures (e.g. South Korea), people address each other differently based on their age. This kind of things are deeply embedded to us that we don't even consciously think about it, but it's something that robots don't have a clue about.

When a human and a robot engage in a conversation, we assume that this information is also important from the robot side. Only when the robots can detect and classify such things, it is possible for the robot and the human to have a proper conversation.

For example, let's say that you are talking to a robot and it misclassifies your gender. In some languages, English for example, pronouns can change their forms based on the gender. Now that the robot has misclassified your gender, many things that robot say might sound very awkward to you.

We first visit the existing literature.

- Poria *et al.*, 2018 [1] has annotated emotion and sentiment from TV-series Friends multimodal data. The data are public, with which we can not only save our time but also compare the results with other researchers.
- Levi et al., 2015 [2] uses a simple CNN to classify one's gender and age from their faces. Their data is also publicly available. There have been more sophisticated models since then, but nonetheless this is a good starting point.
- Ghosal *et al.*, 2020 [3] uses mental states, events, and causal relations to recognize emotion/sentiment in the utterance level. They test their model on datasets such as MELD [1]. Reproducing their work will give us insight how an utterance level emotion/sentiment recognition can be done.

3. Research questions

Given the time constraint, we can't answer all of the below research questions. We should prioritize and be selective of what we can do for the next two months. Here are some of my thoughts.

1. Emotion/sentiment recognition and age/gender detection are currently two separate lines of work. Can one complement the other?
2. How useful is the information (i.e. emotion/sentiment and age/gender) for communicative robots?

3. How is this project connected to the other two projects (“From dialogue to Kinship relations and back” by Jaap Kruijt and “In need of an explanation?” by Lea Krause)? Can they benefit from our work and/or can we benefit from their work?

4. Some of the classes such as age at a given time are not context dependent, whereas other classes such as emotion/sentiment are highly dependent on the context. Is the context well modeled in the existing works? If they are, how are they modeled? If they are not well modeled, are there any ways to do it better?

5. (highly technical) Can we do better than the state-of-the art age/gender detection and emotion/sentiment recognition? The simplest way to achieve this is to fine-tune the hyper-parameters used in their training. The most complicated way would be to come up with a more suitable neural network architecture (e.g. enhancing the used RNNs, CNNs, transformers, and attention modules).

4. Research methodology

1. Reproduce the existing state of the art will be the first step. Unix or Unix-like machines with a python environment are highly desirable. The TA of this project (Taewoon Kim) will try to do it on google colab so that anyone can follow easily.

2. Sync with the other two PhDs (Jaap Kruijt and Lea Krause) for the big picture. For this test data needs to be created with the other projects for integrated system tests.

5. Bibliography

[1] Poria, Soujanya, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. "Meld: A multimodal multi-party dataset for emotion recognition in conversations." arXiv preprint arXiv:1810.02508 (2018)

[2] G. Levi and T. Hassncer, "Age and gender classification using convolutional neural networks," 2015 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Boston, MA, 2015, pp. 34-42, doi: 10.1109/CVPRW.2015.7301352.

[3] Ghosal, D., Majumder, N., Gelbukh, A., Mihalcea, R., & Poria, S. (2020). COSMIC: COmmonSense knowledge for eMotion Identification in Conversations. ArXiv, abs/2010.02795.