1 Abstract

The ability to detect empathy is an important capability for socially compliant robots. We will build an end-to-end multimodal learning system that predicts the empathy of a participant in a conversation. The novelty of our project will be in the way we are fusing multiple inputs into our system. We will be using an attention mechanism to fuse audio and video inputs to predict empathy. Using attention will improve the interpretability of our system to allow us to better understand what it means to be empathetic.

2 Background

Empathy is defined as the ability to understand and share feelings of one another. This ability to understand other people’s emotions is critical for humans to interact with each other and understand one another. Any automated machine or robot would benefit from having an empathy module to detect this expressed emotion from the person they are conversing with. Once able to understand another’s emotional state, they can react accordingly. This can ultimately improve the likeability of robots by allowing them to carry more thoughtful and natural conversations.

Even for humans, detecting empathy is a difficult task because it requires picking up on subtle voice inflections or facial expressions that the other person expresses. Because the problem of detecting emotion and empathy from humans is highly nuanced and complex, we are using deep learning to try to learn and
understand what features are most important to the expression of empathy in audio and gesture.

Our project will incorporate the use of attention to fuse the audio and video inputs. Attention is currently used for tasks such as speech recognition, summarization, and image captioning. It assigns importance to different characteristics of the input and learns which features are highly correlated to the correct outcome. It has gained popularity in recent years because it improves interpretability of what the neural network is doing and which features it is paying attention to.[1] This is also one of the main reasons we are interested in using attention for this project—to be able to better interpret what features make a person more or less empathetic.

3 Dataset

We will be using the OMG-Empathy challenge dataset.[2] This dataset has seven hours of recordings of subjects conversing (mp4 files). In these videos, one participant tells a semi-scripted story, and the other is told to react emotionally. The videos are labeled after the conversation (by the listening participant) with how empathetic they felt towards the other person at each second. They labeled empathy on a scale of -1 (no empathy) to 1 (empathetic). Because our system is multi-modal, we will be using the audio from the video files and the video itself as inputs to our system.

4 Proposed Project

We will be implementing a neural network for empathy recognition using audio and video. The input to our system will be audio and video clips, and the output is an empathy prediction. Here are all the relevant intermediate steps within our system:

1. Preformatting data: Before training, we need to give some thought to the inputs to our system. The dataset provides mp4 videos with labeled empathy label at every frame.

   - Audio: We will need to first extract the audio from the videos. We will first try using all the audio (including both the storyteller and the listener) as inputs to our system. If time permits, we can test segmented out just the audio of the listener’s response.
• Video: The video in this dataset contains both speakers and their full bodies. We will first attempt to use only a cropped image of the listeners face as input to our system. If time permits, we will consider adding facial landmarks or full body crops into our model to improve performance as this was done in previous work. [5]

2. Feature Extraction:

• Audio: Previous work for feature extraction of audio has used pre-built toolkits such as openSmile.[3] We will experiment with different methods that have worked to see what can achieve the best results for our system.

• Video: Feature extraction for video is most commonly done by using convolutional neural networks. There are many pre-trained CNNs we can use (such as the VGG Face CNN descriptor) for feature extraction. [3]

3. Fusing feature vectors for multi-modal input: After computing the two feature vectors, we will fuse them together by taking a weighted sum, where the weights are the attention coefficients. Attention coefficients can be computed by following the work of our advisor’s paper on multimodal robot navigation. [4]

4. Model to predict empathy: Once we have a fused feature vector, we will pass it through a feedforward network that will output an empathy prediction.

We will be using the Concordance Correlation Coefficient (CCC) evaluation metric to test our model’s prediction. The CCC is measure of correlation that penalizes divergences in either mean or variance between two data sequences. [5]

5 Deliverables

The deliverables for our project will include individual final reports detailing our work and a Github repository containing our code and weights for the trained model.
6 Division of Work

This is a two person project, so we will be splitting up the project in a way that is fair and challenging for both members. We will both be in charge of doing preliminary research of previous models and methods that have been successful in solving this problem and similar problems. In addition, we will both be writing final reports. Julia will focus on formatting and feature extraction of video, and Roland will work on formatting and feature extraction of audio. After we get features, we will work together on implementing the attention mechanism and model for predicting empathy from the fused feature vector.

References


