6.8620 Spring 2024

Project Proposal

Do not exceed two pages (excluding references). Do not modify the margins or the font size.

1 Team Members

Member #1: Vincent Lin (linv)

2 Problem

Sarcasm is a fundamentally multimodal form of communication. Many dimensions of human communication contribute to conveying the meaning or indication of sarcasm: exaggerated facial expressions, amplified pitch modulations, text-based conversational context, etc. However, without a multimodal architecture, it's difficult for deep learning methods to capture all modes of human input. However, given the intuitive adjacency of sarcasm detection and general emotion detection, is there some implicit correlation between the two that can be learned by a deep learning model without having to be trained on both? Can we learn something about the internal representation of one by training solely on the other? This is the question I intend to investigate via modern transformer-based architectures and benchmark datasets.

3 Methods

As a basis for audio encoding, I intend to use HuBERT [1], a BERT-based audio encoder that has published results in high performance for downstream audio tasks. To research the cross-modal relationship between sarcasm and emotion classification, as well as HuBERT's learned representation of each, I intend to investigate the performance of a sarcasm fine-tuned instance of HuBERT on the emotion dataset benchmark, and vice versa with a emotion fine-tuned model on the sarcasm benchmark (datasets described below). The intention of this experimental framework is to discover whether fine-tuning a transformer-based encoder on one task can implicitly improve the performance of the encoder on the other, without having seen explicit training data on the other. Ideally, identifying some correlation between the two "modalities" of speech will lead to a deeper analysis of their relationship, e.g., in the comparing the HuBERTs' latent space embeddings.

4 Experiments

4.1 Datasets

CREMA-D: The Crowd-Sourced Emotional Multimodal Actors Dataset (CREMA-D) [1] is a dataset of audio and video clips from several actors, each classified as one of six different emotions (Anger, Disgust, Fear, Happy, Neutral, and Sad) at four different levels of expression (Low, Medium, High, and Unspecified). This dataset provides diversified data across gender, age, race/ethnicity, etc. and will be the training foundation for emotion classification fine-tuning.

MUStARD: The Multimodal Sarcasm Dataset (MUStARD) [2] is a multimodal video corpus of TV show clips compiled for automated sarcasm recognition. Each clip contains video, audio, and text data of each utterance annotated with binary sarcasm labels. I will focus primarily on the audio data provided in this dataset and use it as the training foundation for sarcasm classification fine-tuning.

4.2 Experimental Procedure

Experimentation will include isolated fine-tuning of two separate instances of HuBERT on their respective datasets. Then, each instance will be evaluated on both the emotion and sarcasm benchmarks (ideally, performing better on the benchmark corresponding to its respective training data). Comparing these results to a baseline non-fine-tuned HuBERT instance, I am curious to discover whether training on general emotion classification data boosts the performance of HuBERT on classifying sarcasm (without encountering explicit sarcasm training), and vice versa.

Since I will be started with HuBERT's pretrained weights and fine-tuning from that starting point, the required computational resources should be manageable, given that I also perform the necessary dataset pre-processing & data truncation.

There may also be opportunity to perform a joint training of both emotion and sarcasm classification combined in a multi-task training paradigm to compare against the performance of the individually-trained HuBERT instances.

5 Timeline

Proposed timeline:

April 30: Initial work with initializing datasets and data processing pipeline

May 4: Fine-tuning HuBERT pipeline finalized and producing consist results on benchmark datasets

May 7: Compilation of results and deeper analysis of why correlations may arise

[1] https://github.com/CheyneyComputerScience/CREMA-D

[1] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, *Hubert: Self-supervised speech representation learning by masked prediction of hidden units*, 2021. arXiv: 2106.07447 [cs.CL].

[2] S. Castro, D. Hazarika, V. Pérez-Rosas, R. Zimmermann, R. Mihalcea, and S. Poria, "Towards multimodal sarcasm detection (an _Obviously_ perfect paper)," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, A. Korhonen, D. Traum, and L. Màrquez, Eds., Florence, Italy: Association for Computational Linguistics, Jul. 2019, pp. 4619–4629.

[3] .