Multimodal Emotion Prediction in Group Conversation

TOMA ALLARY, THIERRY BÉDARD-CORTEY, PHILIPPE BERGERONI, YUAN LI

Abstract

Emotion prediction in group conversations is a challenging task due to the complexity of human interactions and the multimodal nature of emotional expressions.This project aims to develop a multimodal model that integrates textual and audio features for emotion prediction in group conversations. It combines RoBERTa for text embeddings, WavLM for audio embeddings, and a transformer-based fusion layer to enhance emotion recognition. The project is based on the study ”Emotional Cues Extraction and Fusion for Multimodal Emotion Prediction and Recognition in Conversation”[1].

1. Introduction

Human emotions are important for the communication. In recent years, deep learning has been achieving tremendous success on numerous machine learning applications. In natural language and speech processing areas, conversational AI had a huge growth in recent years — from virtual assistants to customer support chatbots. But there's still a big gap: most systems don't understand how we feel. That's why emotion-aware systems are so valuable — they enable more natural, empathetic, and human-like interactions. In this project, Our goal is to build a multimodal machine learning model that could predict the next emotion in a group conversation using traces of previous audio and text. More specifically, we explore simplified architectures to improve on current existing models that are complex and heavy to train and implement. Light and re-usable architecture are generally more suited for real-life tasks where cost and time are important. This work could serve as the basis for applications such as social robots, therapeutic robots and intelligent tutoring systems.

2. Relate Work

There are three major approaches commonly used in emotion detection models. First is RNN-based models, which are good for temporal sequences but suffer with long-term dependencies and slow processing. The second is Transformer-based models, which excel at capturing long-range dependencies and can process data in parallel — though they are resource-intensive. The third is Graph-based models, which represent conversations as interactions between speakers, allowing them to infer emotions based on relationships. These models are powerful but often complex to implement. Our project mainly leveraged transformer-based models for their strong performance on language and speech tasks.

This study (**?**) explores the extraction and fusion of emotional cues from multiple modalities, including text and audio, to enhance emotion prediction in conversations. The authors propose a transformer-based fusion layer to combine embeddings from different modalities, achieving state-of-the-art results in emotion recognition tasks.

2.1. RoBERTa

RoBERTa (Robustly Optimized BERT Pretraining Approach) (**?**)is a transformer-based language model developed by Facebook AI. This model significantly enhances the original BERT architecture through several key modifications in its pre-training methodology. Unlike BERT, RoBERTa was trained on a substantially larger dataset, totaling 160 GB of diverse text and code. This extensive training corpus allowed the model to develop a more robust and generalized understanding of language. One of the critical improvements in RoBERTa is the removal of the Next Sentence Prediction (NSP) objective used in BERT, which was found to not consistently improve downstream task performance. Additionally, RoBERTa employs dynamic masking during pre-training, where the masked tokens change in each epoch, forcing the model to learn more effectively from different parts of the input. The model also benefits from being trained with much larger batch sizes, contributing to more stable and efficient learning. RoBERTa's architecture, being based on the transformer, inherently excels at capturing long-range dependencies within text, making it particularly adept at understanding context. In dialogue, emotions are often

[*] Equal contribution . **AUTHORERR: Missing \icmlcorrespondingauthor.**

conveyed subtly and are deeply intertwined with the surrounding linguistic context. RoBERTa's enhanced semantic capabilities, stemming from its extensive training and architectural refinements, enable it to discern these intricate relationships and make more accurate emotion predictions by considering the nuances of the conversational flow and the interplay of words.

2.2. WavLM

WavLM (**?**), short for Waveform Language Model, represents a significant advancement in self-supervised learning for comprehensive speech processing, developed by Microsoft. This large-scale model is meticulously engineered to learn rich, universal representations directly from raw audio waveforms, eliminating the need for extensive manual feature engineering. Building upon the success of the wav2vec 2.0 framework, WavLM introduces innovative techniques to further enhance the quality of learned representations. A key contribution of WavLM is its emphasis on capturing prosodic features like tone, rhythm, and pitch – elements that carry crucial information about a speaker's emotional state. By explicitly modeling these acoustic characteristics, WavLM gains a distinct advantage in tasks such as emotion analysis, where these subtle vocal cues are paramount. Beyond emotional understanding, WavLM demonstrates its versatility by supporting a wide array of speech-related tasks, including speaker identification, automatic speech recognition, and emotion recognition. This broad applicability makes WavLM an exceptionally valuable component for multimodal fusion approaches, where information from different modalities, such as audio and text, can be combined to achieve a more holistic and accurate understanding of communicative signals. Its ability to extract meaningful features from raw audio across various tasks underscores its potential to significantly improve the performance of speech-centric artificial intelligence systems.

3. Task Definition

Emotion Prediction in Conversation (**EPC**) In a multi-modal multi-party (or dyadic) dialogue containing text and audio D $= (u_1, s_1), (u_2, s_2), ..., (u_N, s_N)$, where (u_i, s_i) represents the ith utterance-speaker pair in the conversation, N is the number of utterances in the dialogue. EPC aims to predict the emotion category label $emotion_{n+1}$ of the future utterance speaker pair (u_{n+1}, s_{n+1}) by given the historical dialogues $(u_1, s_1), (u_2, s_2), ..., (u_n, s_n)$.

3.1. Dataset

MELD is a popular dataset for tasks involving the analysis of emotions expressed by multiple speakers. It encompasses a collection of over 1400 dialogues and 13,000 speech instances extracted from the television show Friends. Emotion annotation in the dataset includes: neutral, happiness, surprise, sadness, anger, disgust, and fear.

4. Method

4.1. Reference baseline method

Our baseline architecture is from study (**?**). As seen in Figure 1, they propose an architecture using both text and audio modalities. They also extract a mel-spectrogram as a third modality from the audio. Their first steps are to extract features from modalities using pre-trained models (i.e. RoBERTa, WaveLM and a Spectrum Extraction Tool)

TODO: explain KWRT

TODO: explain Prosody enhencement

TODO: explain Transformer+fusion

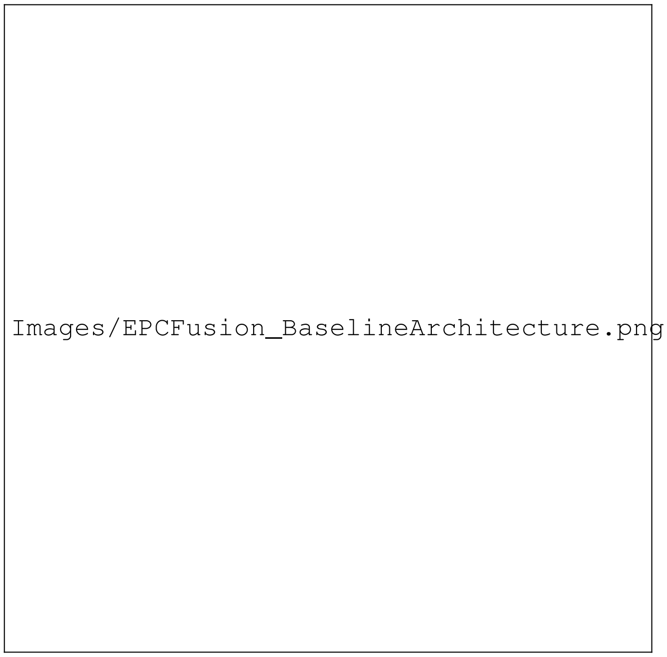

Figure 1. Schema of the baseline architecture. The left side is the main framework of the model. (a) multi-modal fusion module, the green part represents the pre-trained language transformer layer, the yellow part represents the pre-trained vision transformer layer, the pink part is trainable, and other parts are frozen.

4.2. Convolution Neural Network

Our first approach TODO

4.3. Decoder Transformer & MLP

TODO

4.4. Encoder Transformer & MLP

TODO

5. Experiments

For our experiments, we utilize pre-trained wavLM and RoBERTa models to extract 768-dimensional features, then use serval different architectures to predict next emotion. architecture 1 architecture 2 architecture 3
TODO:
Experiments: Comparisons to baselines using metrics (explain what the metrics are), with visualizations in terms of graphs, tables etc. (if you can show more, e.g., qualitative results with images - the better).

6. Optional sections if needed

6.1. Figures

You may want to include figures in the paper to illustrate your approach and results. Such artwork should be centered, legible, and separated from the text. Lines should be dark and at least 0.5 points thick for purposes of reproduction, and text should not appear on a gray background.

Label all distinct components of each figure. If the figure takes the form of a graph, then give a name for each axis and include a legend that briefly describes each curve. Do not include a title inside the figure; instead, the caption should serve this function.

Number figures sequentially, placing the figure number and caption *after* the graphics, with at least 0.1 inches of space before the caption and 0.1 inches after it, as in Figure **??**. The figure caption should be set in 9 point type and centered unless it runs two or more lines, in which case it should be flush left. You may float figures to the top or bottom of a column, and you may set wide figures across both columns (use the environment `figure*` in LaTeX). Always place two-column figures at the top or bottom of the page.

6.2. Algorithms

If you are using LaTeX, please use the "algorithm" and "algorithmic" environments to format pseudocode. These require the corresponding stylefiles, algorithm.sty and algorithmic.sty, which are supplied with this package. Algorithm 1 shows an example.

Algorithm 1 Bubble Sort

```
Input: data $x_i$, size $m$
repeat
  Initialize $noChange = true$.
  for $i = 1$ to $m - 1$ do
    if $x_i > x_{i+1}$ then
      Swap $x_i$ and $x_{i+1}$
      $noChange = false$
    end if
  end for
until $noChange$ is $true$
```

Table 1. Classification accuracies for naive Bayes and flexible Bayes on various data sets.

DATA SET	NAIVE	FLEXIBLE	BETTER?
BREAST	95.9± 0.2	96.7± 0.2	√
CLEVELAND	83.3± 0.6	80.0± 0.6	×
GLASS2	61.9± 1.4	83.8± 0.7	√
CREDIT	74.8± 0.5	78.3± 0.6	
HORSE	73.3± 0.9	69.7± 1.0	×
META	67.1± 0.6	76.5± 0.5	√
PIMA	75.1± 0.6	73.9± 0.5	
VEHICLE	44.9± 0.6	61.5± 0.4	√

6.3. Tables

You may also want to include tables that summarize material. Like figures, these should be centered, legible, and numbered consecutively. However, place the title *above* the table with at least 0.1 inches of space before the title and the same after it, as in Table 1. The table title should be set in 9 point type and centered unless it runs two or more lines, in which case it should be flush left.

Tables contain textual material, whereas figures contain graphical material. Specify the contents of each row and column in the table's topmost row. Again, you may float tables to a column's top or bottom, and set wide tables across both columns. Place two-column tables at the top or bottom of the page.

References

Langley, P. Crafting papers on machine learning. In Langley, P. (ed.), *Proceedings of the 17th International Conference on Machine Learning (ICML 2000)*, pp. 1207–1216, Stanford, CA, 2000. Morgan Kaufmann.