NCL Team at SemEval-2024 Task 3: Fusing Multimodal Pre-training Embeddings for Emotion Cause Prediction in Conversations

Shu Li[1] and **Zicen Liao**[2] and **Huizhi Liang**[3]
[1] Beijing Accent Advertising Co., Ltd.
[2] School of Computing, Newcastle University, Newcastle upon Tyne, UK
[3] School of Computing, Newcastle University, Newcastle upon Tyne, UK
15510366636@163.com
and liaozicen55@gmail.com
and huizhi.liang@newcastle.ac.uk

Abstract

In this study, we introduce an MLP approach for extracting multimodal cause utterances in conversations, utilizing the multimodal conversational emotion causes from the ECF dataset. Our research focuses on evaluating a bi-modal framework that integrates video and audio embeddings to analyze emotional expressions within dialogues. The core of our methodology involves the extraction of embeddings from pre-trained models for each modality, followed by their concatenation and subsequent classification via an MLP network. We compared the accuracy performances across different modality combinations including text-audio-video, video-audio, and audio only.

1 Introduction

In recent times, multimodal sentiment analysis has become a critical research frontier in the realm of natural language processing, moving beyond the confines of traditional text analysis to embrace a richer blend of audio, visual, and text data. This comprehensive approach aims to deepen our understanding of sentiments and emotions.

Previous research has highlighted the effectiveness of hierarchical fusion techniques and context modelling in improving the precision of multimodal sentiment analysis by adeptly merging features from varied modalities (Wang et al., 2023). Additionally, initiatives such as the Unified Multimodal Sentiment Analysis and Emotion Recognition UniMSE have proven the benefits of applying contrastive learning techniques to enhance performance in both sentiment analysis and emotion recognition, underscoring the significance of integrated frameworks within this field (Hu et al., 2022). CubeMLP delves into the realm of feature mixing for multimodal data processing (Sun et al., 2022). Meanwhile, the MMLatch model sheds light on the critical roles of bottom-up and top-down fusion mechanisms (Paraskevopoulos et al., 2022), offering insights into the impact of high-level representations on the synthesis of sensory information.

This study proposes the development of a Multilayer Perceptron network, specifically designed to extract causal utterances from the Multimodal Emotion-Cause Pair Extraction in Conversations (ECF) dataset (Wang et al., 2024).

2 Data Description

For this research, the ECF dataset has been selected as the primary source of data for training and testing our model. The ECF dataset contains several key elements that are integral to our study:

- Video Clips: Each sample in the dataset includes a video clip from the show Friends, capturing the visual expressions, body language, and interactions between characters.
- Audio Tracks: Audio tracks in the video clips, which include the spoken dialogues, tone of voice, laughter, and other paralinguistic features.
- Transcribed Text: For each clip, the spoken dialogues are transcribed to provide textual context to the interactions.
- Emotion and Sentiment Annotations: The dataset provides detailed annotations for each dialogue segment, including the emotion category, the emotion utterance, and the cause utterance.

Our research leverages the video and audio components of the ECF dataset. By analyzing the video and audio modalities, our goal is to uncover the underlying patterns and triggers of emotional expressions, without the direct influence of textual information.

To adapt the ECF dataset for our specific research objectives, a meticulous data preparation

Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024), pages 285–290

process is undertaken. This involves: - Annotation Mapping: Aligning the cause utterance annotations with the corresponding audio and video segments for supervised learning. - Dataset Split: The dataset is divided into training validating subsets as 8:2, the testing subset is provided by the task provider.

Our research endeavours to architect a model that harnesses the strengths of each modality to provide a comprehensive understanding of sentiment. At the core of our methodology is a model architecture designed to seamlessly integrate these diverse data types, leveraging the power of pre-trained models to extract embeddings from text, video, and audio streams for sentiment extraction and classification.

The model lies in the process of concatenating the embeddings generated by these modularity extractors. This approach not only preserves the richness of each modality data but also facilitates the creation of a unified representation that embodies the composite sentiment conveyed across text, video, and audio. The concatenated embeddings serve as input to a Multilayer Perceptron (MLP) classifier, which is designed to discern the integrated sentiment.

3 Methodology

We propose an MLP network architecture designed to synergize the embeddings extracted from video and audio. This network aims to process and integrate these multimodal inputs, facilitating the classification of cause utterances within the framework of sentiment analysis without relying on textual information. The decision to exclude textual data from our analysis stems from a desire to investigate the intrinsic value of audio-visual cues in sentiment analysis.

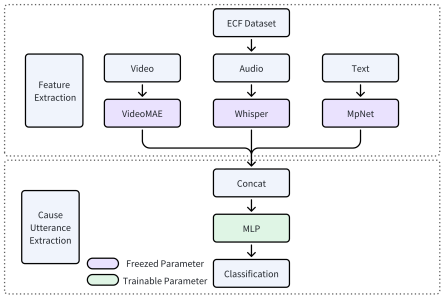

Figure 1: Overview of the cause utterance classification model.

As shown in Figure 1, this model contains two parts, feature extraction and classification. First, we extract audio, video, and text embedding from the pre-trained model. The text embedding extraction is for the comparison of experiments. Then the embeddings are concatenated and put into the MLP network which acts as the classifier for extracting cause utterances. To facilitate the extraction of emotion category and cause utterance, these tasks are regarded as classification tasks. The labels associated with the emotion category and cause utterance are regarded as the classes for these two tasks. This approach enables the MLP to execute the classification.

3.1 Embedding Extraction

VideoMAE is utilized for extracting video embeddings. This model, based on the Masked Autoencoder principle, selectively masks portions of the input video frames and reconstructs the missing parts, thereby learning robust video representations. (Tong et al., 2022) Given an input video V, the model produces an embedding E_V as follows:

$$E_V = \text{VideoMAE}(V) \tag{1}$$

To obtain the video embedding, frames are initially extracted from the video at their native resolution and compiled into a list. Temporal subsampling is applied to this collection of frames, a measure aimed at reducing computational time. Each subsampled set of frames applied normalization and resizing as data augmentation before being inputted into the pre-trained VideoMAE model to acquire the corresponding embeddings.

Whisper is used to extract audio embeddings from the corresponding audio tracks. Whisper processes the raw audio signals, focusing on capturing the nuances of speech, tone, and other auditory features relevant to sentiment analysis (Radford et al., 2022). For an audio input A, the Whisper model outputs an embedding E_A as:

$$E_A = \text{Whisper}(A) \tag{2}$$

Audio information was segregated from the video content and resampled to a 16000Hz sample rate to align with the Whisper model. In leveraging the pre-trained Whisper model for embedding extraction, the classification head was removed to get the pooler output.

MPNet is used to extract text embeddings. MPNet integrates the strengths of both masked language modelling (MLM) and permuted language modelling (PLM) to effectively capture the context

of words in a sentence, both from left-to-right and right-to-left, making it effective for understanding the full context of textual data.For an text input T, the Whisper model outputs an embedding E_T as:

$$E_T = \text{MPNet}(T) \tag{3}$$

3.2 Integration of Embeddings

The embeddings E_V and E_A are concatenated to form a unified representation E_{VA} of the video and audio modalities:

$$E_{VA} = \text{Concat}(E_V, E_A) \tag{4}$$

The embeddings E_V, E_A and E_T are concatenated for the ablation test:

$$E_{VAT} = \text{Concat}(E_V, E_A, E_T) \tag{5}$$

This concatenated embedding serves as the input to the MLP network. The decision to concatenate these embeddings is based on the hypothesis that doing so preserves the distinctiveness of each modality while allowing the network to learn from the intermodal dynamics, essential for identifying cause utterances.

3.3 Network Design

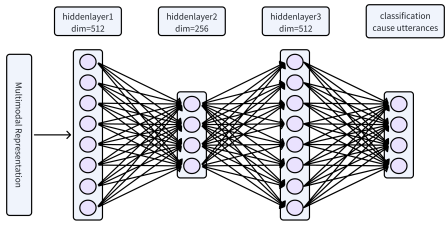

Figure 2: An MLP model aims to classify cause utterance from multimodal embeddings.

Our model employs a MLP architecture, crafted to process and classify concatenated video and audio embeddings. The simplicity and interpretability were significant considerations in choosing the MLP network as the classifier.

The MLP consists of four fully connected layers. The first layer expands the input to 512 hidden units, followed by a reduction to 256 units in the second layer, and an expansion back to 512 units in the third layer, before concluding with the output layer that matches the number of cause utterance classes.

Each hidden layer is equipped with a ReLU activation function to introduce non-linearity, allowing the model to learn complex patterns in the data. To combat overfitting, a dropout rate of 0.5 is applied after each ReLU activation, regularizing the network by randomly omitting a subset of features at each iteration of the training process.

The MLP network is designed with four fully connected layers, integrating nonlinear activation functions and dropout for regularization. Given the concatenated embedding E_{VA}, the forward pass through the MLP can be described by the following set of equations:

- **First Layer Transformation:** The input is passed through the first fully connected layer, transforming it to a higher-dimensional space.

  $$H_1 = \text{ReLU}(W_1 E_{VA} + b_1) \tag{6}$$

  where W_1 and b_1 are the weights and biases of the first linear layer, respectively, and E_{VA} represents the concatenated embeddings. ReLU activation follows to introduce non-linearity.

- **Applying Dropout:** To prevent overfitting, dropout is applied to the output of the ReLU activation,

  $$D_1 = \text{Dropout}(H_1) \tag{7}$$

- **Second and Third Layer Transformations:** The second and third layers further process the data through linear transformations and ReLU activations:

  $$H_2 = \text{ReLU}(W_2 D_1 + b_2) \tag{8}$$

  and

  $$H_3 = \text{ReLU}(W_3 D_2 + b_3) \tag{9}$$

  where W_2, W_3, b_2, and b_3 correspond to the weights and biases of these layers. Each transformation is followed by dropout to enhance model generalization.

- **Final Layer Transformation:** The last step in the network involves passing the output through a final fully connected layer without subsequent ReLU activation, resulting in the output logits,

  $$O = W_4 D_3 + b_4 \tag{10}$$

  This layer maps the processed features to the target output space.

Where W_i and b_i represent the weights and biases of the i^{th} layer, respectively, and ReLU is the Rectified Linear Unit activation function. The dropout is applied after each activation except the final layer to mitigate overfitting.

The output O represents the logits corresponding to each class, which in this case are the possible cause utterances. The model employs a cross-entropy loss function to compute the difference between the predicted probabilities and the actual class labels. This loss guides the training process through backpropagation, adjusting the weights W_i and biases b_i to minimize prediction errors. The network is optimized using the Adam optimizer, with a learning rate of 0.0002.

4 Experiments

In the subtask2 dataset, 1,374 conversations have been annotated by human evaluators. The dataset comprises 13,619 video clips, each tagged with a cause utterance label, delineating the specific cause associated with the clip. These cause utterances are distributed across 29 distinct categories. 66.49 percentage of the cause utterances can be attributed to the context provided by the current video clip itself.

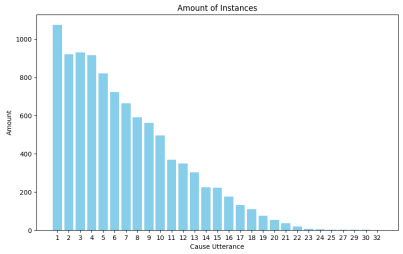

Figure 3: Overview of the cause utterance instances distribution. The cause utterance tends to be more related to earlier situations.

The histogram of Fig.3 illustrates the frequency of instances for each cause utterance category, with a descending order of occurrence. Categories are indexed from 1 to 32 on the x-axis, reflecting a diverse range of causes utterances. The y-axis quantifies the amount of instances, highlighting the prevalence of lower-indexed categories.

4.1 Training Process

The training of our MLP model follows a systematic approach. We utilize the cross-entropy loss, which combines a softmax activation and a log loss in one function. This choice is particularly suited for multi-class classification problems, as it measures the performance of a classification model whose output is a probability value between 0 and 1. The Adam optimizer is chosen for its effectiveness in handling sparse gradients and adapting the learning rate for each parameter, which is crucial given the complexity of our model and the diverse nature of our data. The learning rate is set to 0.0002, offering a balance between fast convergence and the risk of overshooting minimal loss. Our model undergoes training for 2000 epochs. This training period ensures that the model has the opportunity to learn from the entire dataset, optimizing its parameters to identify cause utterances. The alignment of these choices with our research objectives and dataset characteristics ensures a rigorous yet efficient training process, tailored to maximize performance while mitigating the risk of overfitting.

4.2 Metrics

To evaluate the model's performance, we employ the F1 score and weighted F1 score as our primary metrics. These metrics are particularly chosen for their relevance in classification tasks.

F1 Score is calculated as the harmonic mean of precision (P) and recall (R), providing a comprehensive measure of the model's accuracy across all classes. It is given by the equation:

$$F1 = 2 \times \frac{P \times R}{P + R} \tag{11}$$

This metric effectively balances the precision and recall, offering a singular view of model performance.

Weighted F1 Score extends the F1 score by weighting each class's score according to its presence in the dataset. This adjustment makes the metric more representative of the model's performance across classes of varying sizes. The weighted F1 score can be expressed as:

$$\text{Weighted F1} = \sum_{i=1}^{n} w_i \times F1_i \tag{12}$$

where w_i is the weight or relative frequency of class i in the dataset, and $F1_i$ is the F1 score for class i. This calculation ensures the final score reflects the proportional significance of each class, making it invaluable for datasets with class imbalances. These metrics provide an assessment of the model's performance, reflecting its effectiveness in classifying the embeddings in alignment with cause utterance extraction.

5 Ablation Studies

In the context of investigating the ECF dataset, our research undertook a series of ablation studies. These studies were aimed at elucidating the impact of various combinations of modalities on the efficacy of cause utterance classification. These studies are crucial for understanding how combining video, audio, and text data can enhance performance. These studies also help people assess the individual impact of each modality of data on the task cause utterance classification. Ablation Study Design The ablation studies were designed to compare the following configurations:
- Utilization of video, audio, and text embeddings. We utilized video, audio, and text embeddings to assess the maximum potential of multimodal data fusion. This configuration represents the most comprehensive approach.
- Utilization of video and audio embeddings without text. By employing video and audio embeddings while excluding text, our objective is to test whether the information conveyed by the audio modality is equivalent to that of the text modality. This comparison helps us understand the extent to which visual and auditory information alone can drive the classification process.
- Utilization of either video or audio embeddings exclusively. This test helps determine the standalone capabilities of visual and auditory data in identifying cause utterances.

5.1 Ablation Study Results

The network design does not incorporate any combination of modalities.

Configuration	F1	wF1
Video + Audio + Text	0.0253	0.0552
Video + Audio	0.0237	0.0694
Video Only	0.0144	0.0255

Table 1: Ablation Study Results on the ECF Dataset development set.

Configuration	F1	wF1
Video + Audio + Text	-	-
Video + Audio	0.0152	0.0146
Video Only	0.0222	0.0119

Table 2: Ablation Study Results on the ECF Dataset test set.

The combination of video and audio embeddings emerged as a configuration, showcasing its ineffectiveness in the absence of textual data.

6 Conclusion

We proposed a bimodal framework incorporating visual and acoustic modalities for emotion extraction from the "Friends" series, with the addition of a text modality to discern its performance enhancement. The results demonstrate that as the number of modalities increases, the accuracy of emotion extraction gradually improves. Particularly, the Visual-Acoustic model exhibits relatively good accuracy, with a significant improvement upon the addition of the textual modality. The experiment highlights:

- The crucial role of the text modality in sentiment analysis.
- In scenarios lacking textual data, the application of bi-modal models incorporating visual and acoustic modalities can effectively accomplish recognition tasks.

However, the experiment has several limitations concerning the target task. For instance, it did not utilize state-of-the-art pre-trained models, resulting in intra-modality comparisons without specifying the most suitable model for the task. To overcome this limitation, we will develop an evaluation system in our future work to further investigate the effects of embedding extraction using different modalities with pre-trained models. Due to time and resource constraints, the experiment did not extensively tune the models, thereby might not show their optimal performance. Future research could explore using Multi-modal LLMs and task-specific pre-trained models to predict emotion cause in conversations.

References

Guimin Hu, Ting-En Lin, Yi Zhao, Guangming Lu, Yuchuan Wu, and Yongbin Li. 2022. Unimse: Towards unified multimodal sentiment analysis and emotion recognition.

Georgios Paraskevopoulos, Efthymios Georgiou, and Alexandros Potamianos. 2022. Mmlatch: Bottom-up top-down fusion for multimodal sentiment analysis.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision.

Hao Sun, Hongyi Wang, Jiaqing Liu, Yen-Wei Chen, and Lanfen Lin. 2022. Cubemlp: An mlp-based

model for multimodal sentiment analysis and depression estimation. In *Proceedings of the 30th ACM International Conference on Multimedia*, MM '22. ACM.

Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. 2022. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training.

Fanfan Wang, Zixiang Ding, Rui Xia, Zhaoyu Li, and Jianfei Yu. 2023. Multimodal emotion-cause pair extraction in conversations. *IEEE Transactions on Affective Computing*, 14(3):1832–1844.

Fanfan Wang, Heqing Ma, Rui Xia, Jianfei Yu, and Erik Cambria. 2024. Semeval-2024 task 3: Multimodal emotion cause analysis in conversations. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 2022–2033, Mexico City, Mexico. Association for Computational Linguistics.