M2FNet: Multi-modal Fusion Network for Emotion Recognition in Conversation

Supplementary Material

Vishal Chudasama, Purbayan Kar, Ashish Gudmalwar, Nirmesh Shah, Pankaj Wasnik*, Naoyuki Onoe
Media Analysis Group, Sony Research India, Bangalore, India
{vishal.chudasama1, purbayan.kar, ashish.gudmalwar, nirmesh.shah,
pankaj.wasnik, naoyuki.onoe}@sony.com

The supplementary material covers additional experiments further to validate the efficacies of the proposed M2FNet model. In Section 1, a detailed description of the datasets is presented. Section 2 describes the training performance of the proposed model on MELD and IEMOCAP datasets. The performance of the proposed M2FNet model is analyzed in Section 3. Finally, Section 4 presents the visual analysis of the proposed model.

1. Details of used Datasets

We have evaluated our proposed model on two well-known benchmark datasets: IEMOCAP [1] and MELD [3]. The details of these datasets are given below:

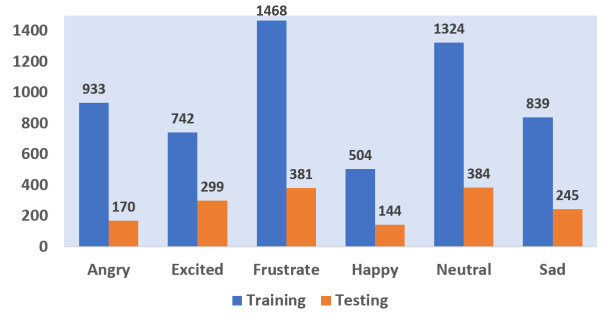
1) Interactive Emotional Dyadic Motion Capture (IEMOCAP): IEMOCAP is a multimodal, multi-speaker database that contains videos of dyadic sessions with around 12 hours of audiovisual data and text transcriptions. It was collected by the Speech Analysis and Interpretation Laboratory (SAIL) at the University of Southern California (USC). This database was created using markers placed on the face, head, and hands of 10 actors in dyadic sessions, which offer precise information on their facial expressions and hand movements throughout scripted and natural spoken communication scenarios. Although there are 11 distinct emotions in the original dataset, only six emotions (i.e., Neutral, Frustrate, Sad, Angry, Excited, and Happy) are utilized in the evaluation. The distributions of these emotions for the training and testing process are illustrated in Figure 1(a).

2) Multimodal EmotionLines Dataset (MELD): MELD is a multimodal dataset that consists of raw videos, acoustic segments and text transcripts for the task of emotion recognition in conversation. It has been generated by extending the EmotionLines dataset [2] where it contains the same dialogue texts as the EmotionLines dataset; however, it includes audio and visual modalities in addition to text. There are about 1400 dialogues and 13000 utterances created from the Friends TV series. Multiple speakers have participated in each dialogue, while each utterance in the di-

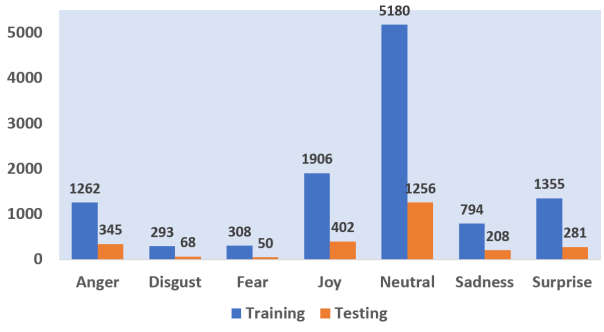
*Pankaj Wasnik is the corresponding author.

Table 1. Statistics of the IEMOCAP dataset

Statistics	Train	Dev	Test
Number of dialogues	100	20	31
Number of utterances	4778	980	1622
Numer of speakers	10 (Male: 5 & Female: 5)		
Maximum conversation length	110		
Minimum conversation length	8		
Average conversation length	50		
Number of classes	6		



(a) IEMOCAP



(b) MELD

Figure 1. Emotion-wise distribution of benchmark datasets.

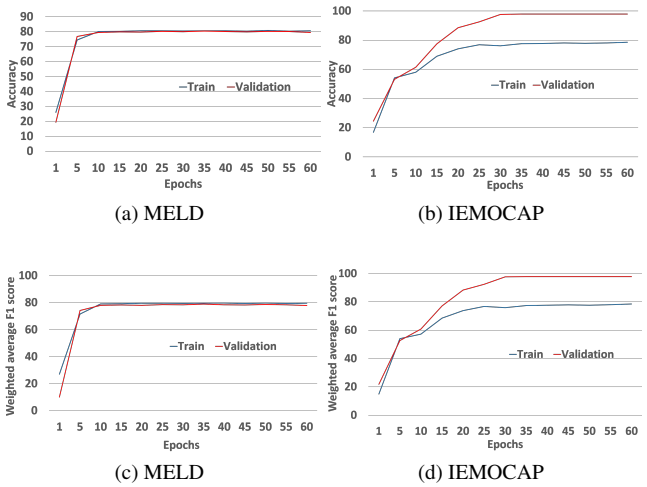
Table 2. Statistics of the MELD datasets

Statistics	Train	Dev	Test
Number of dialogues	1039	114	280
Number of utterances	9989	1109	2610
Number of speakers	260	47	100
Number of unique words	10,643	2,384	4,361
Avergae utterance length	8.03	7.99	8.28
Maximum utterance length	69	37	45
Number of emotion shifts	4003	427	1003
Average number of emotions per dialogue	3.30	3.35	3.24
Average duration of utterances	3.59s	3.59s	3.58s

alogue has been assigned with seven emotions, i.e., Anger, Disgust, Sadness, Joy, Neutral, Surprise and Fear. It also includes a sentiment annotation like positive, negative, or neutral. The statistics of the MELD dataset are described in Table 2 while the training/testing distribution of each emotion is demonstrated in Figure 1(b). Here, it is visualized that there is an imbalanced distribution of samples for emotions. Furthermore, the MELD dataset consists of almost twice the number of samples as the IEMOCAP dataset, making it more challenging to analyze than the IEMOCAP dataset.

2. Training Performance

We train the proposed M2FNet network using the categorical cross-entropy loss function on each utterance for the M number of dialogues. 10% of the training data is used as the validation set. During training process, we measure the corresponding accuracy and weighted average F1 score on training and validation datasets. The corresponding graphs are illustrated in Figure 2 (a-d) for MELD and IEMOCAP datasets, respectively.



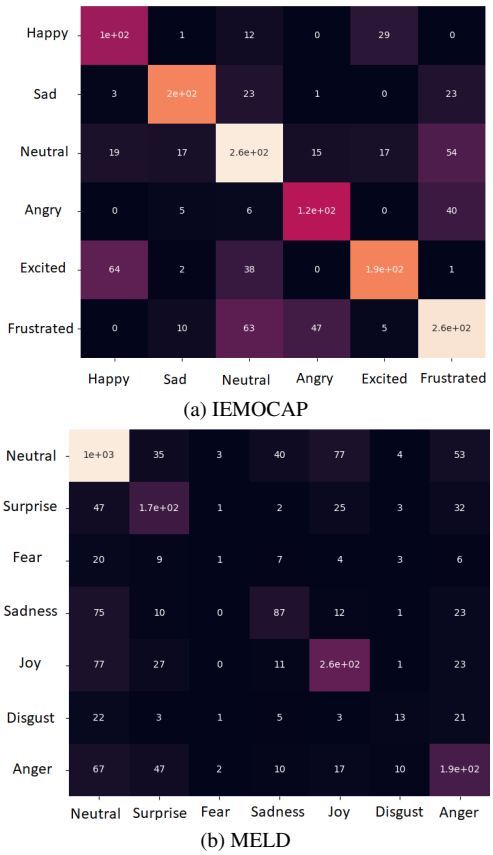
(a) MELD (b) IEMOCAP

(c) MELD (d) IEMOCAP

Figure 2. Learning curves of the proposed network for MELD and IEMOCAP training dataset.

3. Performance Analysis

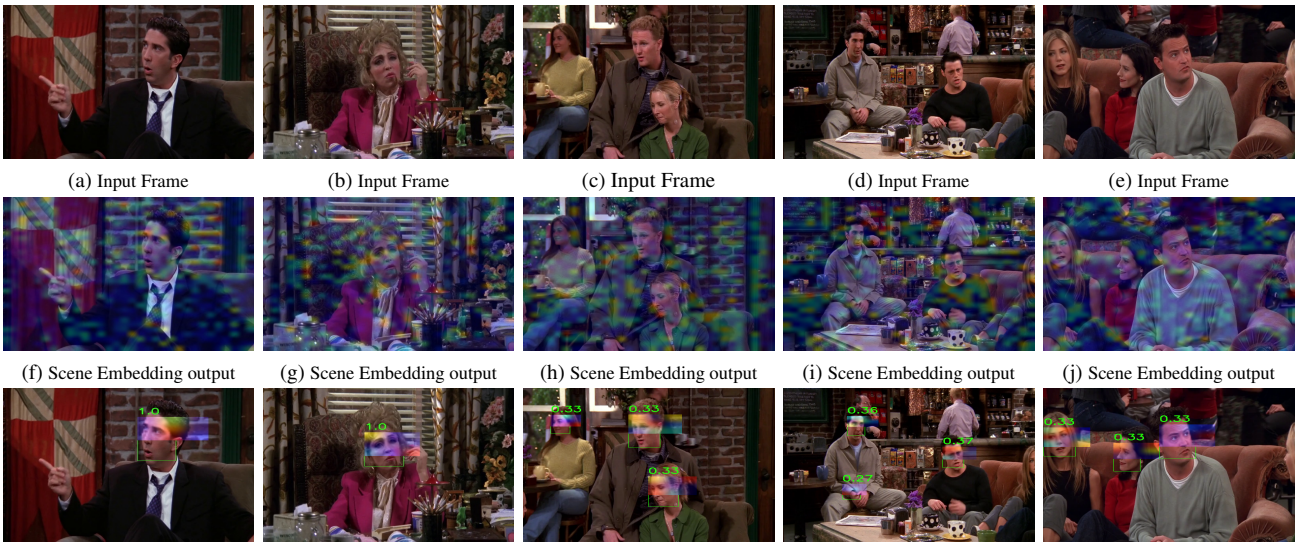
To observe how the proposed model works well to distinguish different emotions, we show a confusion matrix-based analysis in Figure 3 for the test sets of IEMOCAP and MELD datasets, respectively. From this Figure, we can see that the proposed model can separate different emotions very well. In case of the IEMOCAP dataset (i.e., from Figure 3(a)), it is observed that the proposed model tends to get confused between Frustration and Anger as well as between Happy and Excited emotions. This happens due to these emotions having similar visual and acoustic cues. The misclassification is more prominent on the MELD dataset due to the imbalanced distribution of emotions, with a large percent (i.e., approximately 45%) of the labels being neutral. This can be visualized from Figure 3(b) where the proposed model misclassifies some instances of other classes as neutral. Here, one can also observe the confusion between Surprise-Anger and Disgust-Anger that might happen due to the low number of Disgust and Surprise instances and their similarity with Anger emotion.



(a) IEMOCAP

(b) MELD

Figure 3. Heatmap predictions generated by the proposed network on the IEMOCAP and MELD test sets.



(a) Input Frame (b) Input Frame (c) Input Frame (d) Input Frame (e) Input Frame

(f) Scene Embedding output (g) Scene Embedding output (h) Scene Embedding output (i) Scene Embedding output (j) Scene Embedding output

(k) Weighted Face model output (l) Weighted Face model output (m) Weighted Face model output (n) Weighted Face model output (o) Weighted Face model output

Figure 4. Visual analysis

4. Visual Analysis

To observe what the proposed model focus on, we adopt the GradCAM [4] technique and generate the images from the prediction of the proposed model. In Figure 4, we show such GradCAM images. Here, from individual faces' GradCAM outputs, it is observed that the proposed weighted face model identifies the most prominent facial features from the frame and tends to improve the score.

References

[1] Carlos Busso, Murtaza Bulut, et al. Iemocap: interactive emotional dyadic motion capture database. *Language Resources and Evaluation*, 42:335–359, 2008.

[2] Sheng-Yeh Chen, Chao-Chun Hsu, Chuan-Chun Kuo, Lun-Wei Ku, et al. Emotionlines: An emotion corpus of multi-party conversations. *arXiv preprint arXiv:1802.08379*, 2018.

[3] Soujanya Poria, 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.

[4] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. *International Journal of Computer Vision*, 128(2):336–359, Oct 2019.