Multimodal Emotion Modelling with Deep Learning using Shared Feature Space

Vivek Tiwari
University of Southern California
Los Angeles, United States
vivekprt@usc.edu

ABSTRACT

This purpose of this paper is to introduce a study of deep learning techniques for emotion prediction, multimodal emotion prediction. Deep learning techniques have been explored for modeling human emotions across different modalities. There are model agnostic and model based approaches for emotion modeling. In this paper, we are going to explore model based approaches paraticularly neural networks. Multimodal emotion modeling with fusion techniques is also an explored domain that tend to combine the features from each modality in an early, later or hybrid way.

1 MOTIVATION

The idea is to employ novel deep neural networks (DNN) for multimodal fusion of audio, video and text modalities for emotion prediction. The proposed works walks through a previous approaches with DNN architectures for emotion prediction with independent DNNs and shared layers which aim to learn the representation for each modality, as well as a combined representation to achieve prediction. The motivation is to employ different modular architectures from the input stage of feeding data to the output stage of emotion prediction. This implies that this work will seek out to test different approaches in feature extraction like CNN architectures, RNN architectures and myriad ways to combine independent features into a shared feature space. In the end, the shared feature space is processed by neural network architectures to extract emotion prediction for the input. The work is relevant to the coursework in human emotion recognition using multi-modular approaches for emotion prediction in an attempt to study the fusion layers and provide an review of the experiments.

2 BACKGROUND

Over the past decade, facial expression recognition (ER)has been a topic of significant interest. Many ER techniques have been proposed to automatically detect the seven universally recognizable types of emotions like joy, surprise, anger,fear, disgust, sadness and neutral from a single still facial image [4], [5], [6].

Multimodal machine learning enables a wide range of applications: from audio-visual speech recognition to image captioning. With the development of artificial intelligence, there is an explosion of interest in realizing more natural human-machine interaction (HMI) systems. The emotion, as an important aspect of HMI, is also attracting more and more attention. Due to the complexity of emotion recognition and the diversity of application scenarios, the single modality is difficult to meet the demand. Multimodal recognition methods, which take into account the audio, video, text and biological information, can improve the recognition performance [1].

It is important to extract more discriminative features in the emotion classification. Before the popularity of deep neural networks (DNNs), frame-level handcraft features are wildly studied and utilized [2] including Histogram of Oriented Gradient (HOG), Local Binary Patterns (LBP), Local Phase Quantization (LPQ) and Scale Invariant Feature Transform (SIFT). Three Orthogonal Planes (TOP), summarizing functionals (FUN), Fisher Vector encoding (FV), Spatial Pyramid Matching (SPM) and Bag of Words (BOW) are also utilized to capture temporal information [15], [16]. Now the DNNs based approach generates the state-of-the-art performance in many fields. However, due to limited training samples in the AFEW database, complex DNNs are difficult to train [22]. To deal with that problem, transfer learning is adopted. Then bottleneck features are extracted from fine-tuned models [7], [8], [9].

To gain better performance, fusion methods that merge different modalities are essential. Fusion methods can be classified into feature level fusion (or called early fusion), decision level fusion (or called late fusion) and model level fusion. Most teams chose late fusion in pervious challenges [10], [11], [12], [13]. Vielzeuf et al. [13] discussed five fusion methods: Majority Vote, Mean, ModDrop, Score Tree and Weighted Mean. They found that Weighted Mean was the most effective fusion method, which had less risk of overfitting. Ouyang et al. [14] utilized reinforcement learning strategy to find the best fusion weight.

A novel architecture for the fusion of different modalities called conditional attention fusion was proposed in [3]. We use Long-short term memory recurrent neural networks (LSTMs) as thebasic model for each unimodality since LSTMs are able to capturelong time dependencies. For each

time step, the fusion model learn show much of attentions it should put on each modality conditioning on its current input multi-modal features and recent history information. The approach is similar to human perceptions since humans can dynamically focus on more obvious and trustful modalities to understand emotions. Unlike early fusion, we dynamically combine predictions of different modalities, which avoids the curse of dimensionality and synchronization between different features. And unlike late fusion, the input features are interacted in a higher level to learn the current attention instead of being isolated without any interactions among different modalities.

3 SEWA DATASET

The main aim of the SEWA DB is to provide enough suitable data of labelled examples to facilitate the development of robust tools for automatic machine understanding of human behaviour. To create the SEWA dataset, a data collection experiment has been conducted. To promote natural interactions, participants within each pair were required to know each other personally in advance of the experiment. The entire watching of adverts and the subsequent conversation between the volunteers is recorded using web-cameras and microphones integrated into the laptops/PCs of the volunteers.

The SEWA database includes annotations of the recordings in terms of facial landmarks, facial action unit (FAU) intensities, various vocalisations, verbal cues, mirroring, and rapport, continuously valued valence, arousal, liking, and prototypic examples (templates) of (dis)liking and sentiment. The data has been annotated in an iterative fashion, starting with a sufficient amount of examples to be annotated in a semi-automated manner and used to train various feature extraction algorithms developed in SEWA, and ending with a large DB of annotated facial behaviour recorded in the wild.

In my work, I've focused on working with the data where each volunteer is asked to watch adverts (each person watches 4 adverts, each being about 60 seconds long). These adverts have been chosen to elicit mental states including amusement, empathy, liking and boredom. After watching the advert, the volunteer is also asked to fill-in a questionnaire to self-report his/her emotional state and sentiment toward the advert. It also includes conversation with another volunteer usually known to the first volunteer by means of a video-chat software (on average, 3 minutes long conversations). This provides video, audio and text transcripts of the subject's conversation with another. The work is based on multi-modal modeling of dimensional emotions on this data. The data is annotated by 5 annotators who annotate on video data alone, audio data alone and video audio data for valence and arousal using a joystick.

4 FEATURE EXTRACTION

FEATURE SAMPLING

The text transcripts of the data provide start and end time slots between which words were spoken along with what was spoken. A sample of what it looks like can be seen from Figure 1. Based, on the start and end time for each text transcript, audio and video features are extracted from that particular interval. Each row therefore, corresponds to one sample of the dataset used in this work which has all three modalities - image, audio and text.

start_time	end_time	subject	text
-0.25	0.4	246	<laughter>
0.96	5.17	245	and like... But its good, cuz it like, apply to all different situations I guess, like
5.2	6.14	246	Yeah, a kitchen??
5.84	6.56	245	Children
6.76	7.67	245	as well as
7.07	7.55	246	Yeah
8.28	11.12	245	adults? See, yeah, that was pretty good, but
9.07	9.61	246	...
11.44	12.33	245	It was alright
12.72	13.85	245	wasnt particularly

Figure 1: Sample Text Transcript Data

DATA MODELING

Since, the SEWA dataset was annotated using a joystick by annotators as they viewed the data, there was no specific time given to them to annotate a piece of data. This means that as the data was displayed to them, they were annotating it with the movement of the joystick. This much likely reduces the temporal dependency for data annotation. The work was therefore focused to understand emotion predictions using video, audio and text data at any instance. For this purpose, the data was modeled to include instances sampled from intervals defined by the text data. For example, the text data provided transcript for interval (T0,T1) as shown in Figure 1. This interval was substantially small and did not depict much difference in video data over frames. Hence, corresponding to the interval video frame and audio data was extracted to form a sample of video,audio and text data. The data was Z-score normalized before training.

AUDIO FEATURE MODELING

The audio data is available in .wav files in the SEWA database. For this work, Mel Frequency Cepstrum Coefficients (MFCC) features were extracted from the .wav files using the librosa library at the original sampling rate. A total of n features were experimented with the audio data, where n ranges between 10 and 40, stacked to create a mxn dimensional data. Here, m is the total number of samples. Experiments performed on the MFCC features didn't show facilitate audio based model training. Further insepection revealed that the MFCC features posed no significant correlation with the target values of valence, with pearonr correlation coefficient in the order of 1e-3.

Investigation into audio features for emotion prediction tasks proposed an candidate for audio feature extraction called Deep Spectrum features. Deep Spectrum [18] employs deep neural networks trained on images for audio feature extraction. For this work, fc2 features from VGG16 network are extracted providing a 4096 dimensional audio feature representation. It was observed that out of 4096, 2500 features had a pearsonr correlation greater than 0.1, 981 had greater than 0.2 and 40 features had correlation greater than 0.5 with the target variable.

The audio model trained on these features has 11 Dense layers. The activation to the Dense layer that serves as the feature input for the fusion layer in the multimodal fusion model has no activation. The activation after each other layer is Relu. There is no activation after the final Dense layer. The loss function used is MSE Loss. There were two experiments carried out. In the first experiment, all of the 11 layers were used for emotion prediction. In the second experiment, the first 5 layers were used as features extractors whose weights came from the 11 layers trained on the training data. The first 5 layers were then used as feature extractors for the test data, which was then passed to a Scikit-learn regressor. The results for the input features in the two experiments in displayed in Table 1

Table 1: Audio Model Results

DL/DL + ML Regressors	Best CCC
DL	0.7955666
DL + ML	0.9564198

TEXT FEATURE MODELING

The text data is available in .csv files providing transcriptions in defined start and end time slots. For this work, Glove embeddings trained over news dataset with embedding dimension of 50,100 and 300 were employed to train a emotion prediction model. The model didn't train well requiring further insepection of the embedding features. To investigate the features, a sentence based embedding of the transcripts revealed no significant pearsonr correlation with the target variable. It was hypothesized that this could be the result of difference in the transcriptions and the data on which the embeddings were trained on. To validate this, sentence based embedding of twitter data trained glove embeddings was obtained which revealed pearsonr correlation of the order of 1e-1, better than previous embeddings. Other text feature extractors were including BERT [17], Universal Sentence Encoder [20] and LiWC [21] were explored to model text data for multimodal prediction in a series of individual and fusion experiments.

Investigation into text feature extraction techniques led to Universal Sentence Enocder (USE). The USE embeddings were employed to obtain a 512 dimensional embedding for a sentence.

The text model trained on these features has 11 Dense layers. The activation to the Dense layer that serves as the feature input for the fusion layer in the multimodal fusion model has no activation. The activation after each other layer is Relu. There is no activation after the final Dense layer. The loss function used is MSE Loss. There were two experiments carried out. In the first experiment, all of the 11 layers were used for emotion prediction. In the second experiment, the first 5 layers were used as features extractors whose weights came from the 11 layers trained on the training data. The first 5 layers were then used as feature extractors for the test data, which was then passed to a Scikit-learn regressor. The results for the input features in the two experiments in displayed in Table 2

Table 2: Video Model Results

DL/DL + ML Regressors	Best CCC
DL	0.009
DL + ML	0.14

VIDEO FEATURE MODELING

The video data is available in .wav files. OpenFace model for celebrity face recognition was employed to begin with. The model provides a 128 dimensional embedding of the face. The features demonstrated pearsonr correlation of the order of 1e-3 with the target variable, basically no significant correlation. The video features extracted didn't show any training improvements.

Investigating into other feature extraction techniques for video data, OpenFace [19] was employed for landmark feature extraction over frames. These features included action units, gaze and other landmark positions providing a 465 dimensional feature vector for each frame.

The video model trained on these features has 11 Dense layers. The activation to the Dense layer that serves as the feature input for the fusion layer in the multimodal fusion model has no activation. The activation after each other layer is Relu. There is no activation after the final Dense layer. The loss function used is MSE Loss. There were two experiments carried out similar to the audio model.The results for the input features in the two experiments in displayed in Table 3

Table 3: Video Model Results

DL/DL + ML Regressors	Best CCC
DL	0.706959
DL + ML	0.8883883

EARLY FUSION RESULTS

In order to determine the performance with features when they were combined before passing them to a machine learning model, a set of experiments were carried out. These experiments were carried out with Random Forest and Support Vector regressors with the inputs being different sets of video, audio and text modalities. It utlized OpenFace features for video, DPFS features for audio, BERT features extracted from BERT Base Uncased model from the last four layer encodings and concatenated to form sentence embeddings as well as Universal Sentence Encoder (USE) embeddings for text. The results for the experiments are described in Table 4.

Table 4: Early Fusion Results

Video OpenFace	DPFS	BERT	USE	Best CCC
1	1	1	0	0.7010
1	1	0	0	0.7325
1	1	0	1	0.69

LATE FUSION RESULTS

In order to determine the performance with features when they were combined after passing them to a machine learning model, a set of experiments were carried out. These experiments were carried out with Random Forest and Support Vector regressors with the inputs being different sets of video, audio and text modalities. It utlized OpenFace features for video, DPFS features for audio, BERT features extracted from BERT Base Uncased model from the last four layer encodings and concatenated to form sentence embeddings for text. For each feature, there was a machine learning model that predicted valence. These predictions from different modalities were aggregated to get the late fusion results The results for the experiments are described in Table 5.

Table 5: Late Fusion Results

Video OpenFace	DPFS	BERT	Best CCC
1	1	1	0.39588
1	1	0	0.4781

HYBRID FUSION RESULTS

The hybrid fusion model incorporates video, audio and text modalities. The video features are extracted from OpenFace, audio features are extracted from DPFS and text features come from three different sources. The text features are extracted from BERT using the BERT Base Uncased model by concatenating embeddings from last four layers to get embedding for a sentence, and from Universal Sentence Encoder (USE) which provides an alternate to BERT for sentence embeddings, LiWC features retrieved from VADER. These features are Z-score normalized before splitting them to train and test data.

The model consists of 5 Dense layers for each of the modalities with activation Relu except to the fifth layer which has no activation. The output activations are then added to create a single activation embedding which is passed through 6 Dense layers with activation Relu except the final Dense layer. The loss function used is MSE.

There are multiple experiments carried out to investigate the features with the performance of the fusion model. The experiments were centered to also find out which set of features work out best with the fusion model and data. There were two sets of experiments carried out with the fusion model. The first set of experiments utilized all the layers of the model The second set of experiments were carried out using the fusion model as a feature extractor by training the model on train data. The model was used to extract features from the fusion layer activations which were then used to make valence prediction using a Random Forest or Support Vector regressors. The best results were found when the deep learning model was used as a feature extractor. The results for the experiments are shown in Table 6.

Table 6: Hybrid Fusion Results

OpenFace	DPFS	BERT	LiWC	USE	Best CCC
1	1	1	1	0	0.938
1	1	1	0	0	0.940
1	1	0	0	0	0.967
1	0	0	0	0	0.888
0	1	0	0	0	0.956
1	1	0	0	1	0.879

BEST RESULTS

The baseline for this work is described in [1] for instance based valence predictions on the SEWA dataset. It doesn't take into account the affect on predcitions over time series. Due to change in feature extraction techniques, feature engineering using correlation analysis and fusion methods, the

model outperforms the baseline with a significant margin from the baseline. The best model incorporates features from OpenFace, Deep Spectrum for video and audio data. Incorporating text data features reduces the model performance, primarily because the text data is not utilized over time series. The results are described in Table 7.

Table 7: Best Model Results

Model	Best CCC
BEST Model	0.967
Baseline	0.534

5 EVALUATION METRIC

The evaluation metric used for this work is Lin's Concordance Correlation Coefficient (CCC). Lin's concordance correlation coefficient (CCC) is the concordance between a new test or measurement (Y) and a gold standard test or measurement (X). This statistic quantifies the agreement between these two measures of the same variable. This metric is used to evaluate the work as the baseline provides results with this metric. The Lin's Concordance Correlation Coefficient (CCC) is defined in Eqn 1.

$$\rho(y, y') = \frac{2 * \sigma(y, y')}{\sigma^2(y) + \sigma^2(y') + (\bar{y'} - \bar{y})^2} \quad (1)$$

where y' and y are the sets for which the correlation is calculated, $\sigma^2(y)$ and $\sigma^2(y')$ are the variances calculated on sets y' and y respectively and $\bar{y'}$ and $\bar{y}$ are the means of y' and y, respectively.

6 FUSION LAYER ANALYSIS

The hybrid model with a fusion layer outperformed other models explored in this work. In order to determine what the fusion layer learns and which features from each modality are impactful for the fusion layer output, further investigation was carried out. Since, it is not possible to draw a one-to-one mapping between the units in the fusion layer and the input features, correlation analysis was carried out. Specifically, pearson correlation coefficent was calculated between the input features and mean of the fusion layer activations.

In case of video modality, Action Units features demonstrated most correlations with the fusion layer. Action Units AU23-r (lip tightner), AU26-c (Jaw drop) had correlations greater than 0.9. Action Units AU04-r, AU25-r, AU05-c, AU15-c (Brow Lowerer, Lips Part, Upper Lid Raiser, Lip corner depressor) had correlations greater than 0.7 and Action Units AU01-r, AU02-r, AU05-r, AU01-c, AU04-c, AU07-c, AU17-c (Inner Brow Raiser, Outer Brow Raiser, and lastly AU17 Chin Raiser) had correlations greater than 0.5 with the fusion layer.

Deep Spectrum features extracted for audio data demonstrated the most correlation with the fusion layer. Out of 4096 dimensional feature space, 122 features had correlation greater than 0.9, 245 features had correlation greater than 0.7 and 245 features had correlation greater than 0.5. BERT embeddings did not depict as significant correlations as the Deep Spectrum features. Both sets of features from Deep Spectrum and BERT are encodings, making it difficult to correlate back to original inputs, which is the audio spectrum for Deep Spectrum encodings and text for BERT encodings.

7 FUSION LAYER PATTERNS

In order to investigate OpenFace features and their interactions with fusion layer, three experiments were carried out. In the first set of experiments, all of the OpenFace features were used to train a model. In the second set of experiments, selected OpenFace features including Action Units AUxy-r, where xy denotes the action unit number, and head pose location and rotation features were used to train a model. In the third set of experiments, all features except the ones used in the second experiment were used to train a model. The OpenFace features were scaled and translated randomly and their effects on the activation layer was observed. The audio and text features were not tampered with and were consistent across the two experiments. A pattern emerges in the fusion layer activations extracted from the two models. A subtle pattern emerges from the two experiments that remains consistent for data transformations mentioned earlier performed for that experiment. The selected model shows less activated units than the activations of the model trained with all features.

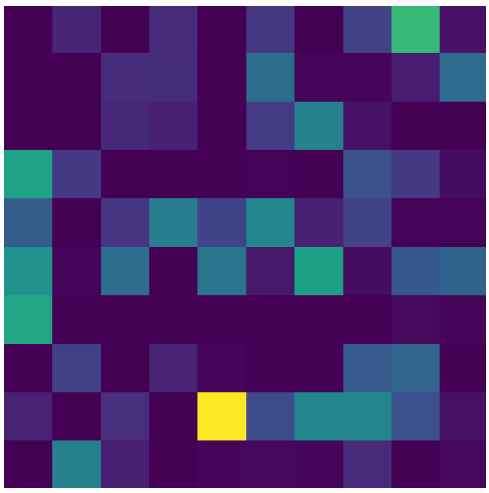

Figure 2: All Features Model Fusion Layer Activations

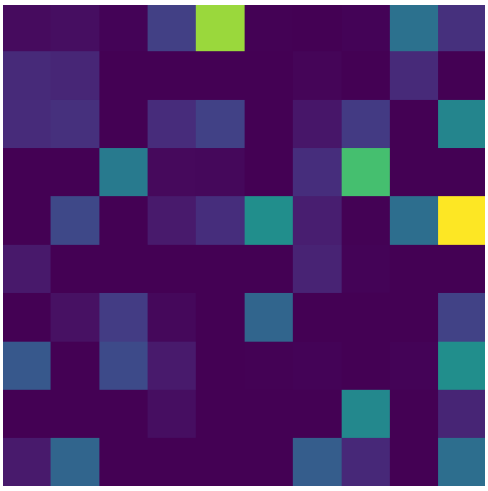

Figure 3: Selected Features Model Fusion Layer Activations

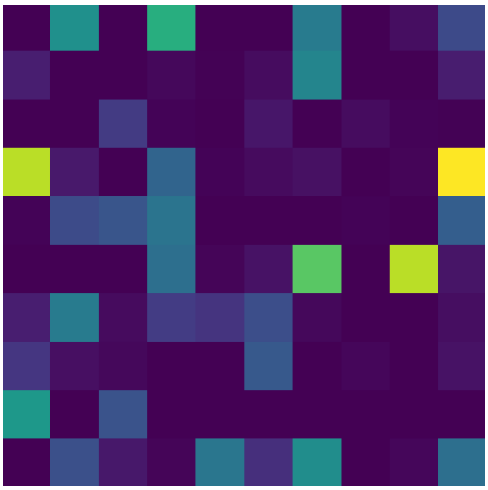

Figure 4: Remaining Features Model Fusion Layer Activations

The activation layer retains the activation of central units across the two experiments. Since, audio and text features were same across experiments, the activations show the interaction of features in the All Features model and the Selected Features model is concentrated with the central units in the activation space. Particularly, what can be drawn from these experiments is that the selected Action Units that also demonstrated high correlations with the fusion layer activations interact with the central units in the activation layer when combined with features from other modalities. This is established from the inactivation of central units in the fusion layer when the selected features from the second experiment were not utilized.

8 OBSERVATIONS

Early and Late fusion techniques were tried to establish the effectiveness of hybrid fusion for the dataset in order to compare it to hypbrid fusion explored in this work. Early fusion results were better than Late fusion results, providing directions for further inspection. BERT encoded embeddings worked better than Universal Sentence Encoder (USE) embeddings for the fusion model.âĂŃ Action Units correlate significantly to the fusion layer providing increase in net CCC score than when tested without action units. âĂŃLIWC features specifically didn't perform well by themselves. In the fusion model, their contribution was close to nothing or decrease in scores when used with other modalities. This could primarily be due to not modeling the features in time series. The experiments provide substantial proof that eventhough the annotations were made on the fly using a joytstick as the media was put forward to the annotators, the measures themeselves were not atomic and followed from time series. As a result, text features didn't contribute positively to the valence predictions. There is a subtle pattern that emerges in the fusion layer depending on the video input features considered that is invariant of the random changes to the inputs themselves. Action units among the video features display strong interactions to selected units in the fusion layer.

9 FUTURE WORK

Since BERT and Deep Spectrum features couldn't be correlated back to their original inputs, one work will be to focus on exploring dependencies of the fusion layer on the Deep Spectrum and BERT features, and correlating them back to the original inputs. This will help to analyze what in the original text or audio spectrum correlates with the fusion layer for any instance without the temporal dependancy. Since, it is established that the data instances are not to be treated as atomic as text data features didn't contribute positively to predictions, exploring temporal dependencies after creating a clean conversation text datasetâĂŃ becomes the area of interest in multimodal modeling and fusion layer analysis.

REFERENCES

[1] Juan D. S. Ortega, Mohammed Senoussaoui, Eric Granger, Marco PedersoliPatrick Cardinal and Alessandro L. Koerich. Multimodal Fusion with Deep Neural Networks for Audio-VideoEmotion Recognition.

[2] Zheng Lian, Ya Li, Jianhua Tao , Jian Huang. Investigation of Multimodal Features, Classifiers and Fusion Methods for Emotion Recognition.

[3] Shizhe Chen, Qin Jin. Multi-modal Conditional AÂĆention Fusion for DimensionalEmotion Prediction.

[4] M. J. Cossetin, J. C. Nievola, and A. L. Koerich. Facial expression recognition using a pairwise feature selection and classificationapproach.InInternational Joint Conference on Neural Networks(IJCNN'2016), pages 5149-5155. IEEE, 2016.

[5] J. Kumari, R. Rajesh, and K. Pooja. Facial expression recognition:A survey.Procedia Computer Science, 58:486 - 491, 2015. 2nd IntlSymposium on Computer Vision and the Internet.

[6] L. E. S. Oliveira, M. Mansano, A. L. Koerich, and A. S. Britto Jr. 2d principal component analysis for face and facial expression recognition. Computing in Science and Engineering, 13(3):9-13, 2011.

[7] J. F. Cohn, T. S. Kruez, I. Matthews, Y. Yang, M. H. Nguyen, M. T.Padilla, F. Zhou, and F. D. la Torre. Detecting depression from facialactions and vocal prosody. In3rd Intl Conf. on Affective Computingand Intelligent Interaction and Workshops, pages 1-7, Sept 2009.

[8] D. L. Tannugi, A. S. Britto Jr., and A. L. Koerich. Memory integrityof cnns for cross-dataset facial expression recognition.InIEEEInternational Conference on Systems, Man, and Cybernetics (SMC),pages 1-6, 2019.

[9] G. Trigeorgis, F. Ringeval, R. Brueckner, E. Marchi, M. A. Nicolaou,B. Schuller, and S. Zafeiriou.Adieu features? end-to-end speechemotion recognition using a deep convolutional recurrent network. InIEEE Intl Conf. on Acoustics, Speech and Signal Processing, pages5200-5204, March 2016.

[10] Fan, X. Lu, D. Li, and Y. Liu, "Video-based emotion recognition using CNN-RNN and C3D hybrid networks," in Proceedings of the 18th ACM International Conference on Multimodal Interaction, 2016, pp. 445-450: ACM.

[11] J. Yan et al., "Multi-clue fusion for emotion recognition in the wild," in ACM International Conference on Multimodal Interaction, 2016, pp. 458-463.

[12] A. Yao, D. Cai, P. Hu, S. Wang, L. Sha, and Y. Chen, "HoloNet: towards robust emotion recognition in the wild," in ACM International Conference on Multimodal Interaction, 2016, pp. 472-478.

[13] V. Vielzeuf, S. Pateux, and F. Jurie, "Temporal multimodal fusion for video emotion classification in the wild," pp. 569-576, 2017

[14] X. Ouyang et al., "Audio-visual emotion recognition using deep transfer learning and multiple temporal models," in The ACM International Conference, 2017, pp. 577-582.

[15] Richard Socher, Alex Perelygin, Jean Y Wu, Jason Chuang, Christopher D Man-ning, Andrew Y Ng, and Christopher PoÂŁs. Recursive deep models for semanticcompositionality over a sentiment treebank. InProceedings of the conference onempirical methods in natural language processing (EMNLP), volume 1631, page1642. Citeseer, 2013.

[16] JunKai Chen, Zenghai Chen, Zheru Chi, and Hong Fu. Emotion recognitionin the wild with feature fusion and multiple kernel learning. InProceedings ofthe 16th International Conference on Multimodal Interaction, ICMI 2014, Istanbul,Turkey, November 12-16, 2014, pages 508-513, 2014.

[17] Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, 2018.

[18] Z. Zhao et al., "Exploring Deep Spectrum Representations via Attention-Based Recurrent and Convolutional Neural Networks for Speech Emotion Recognition," in IEEE Access, vol. 7, pp. 97515-97525, 2019, doi: 10.1109/ACCESS.2019.2928625.

[19] B. Amos, B. Ludwiczuk, M. Satyanarayanan, "Openface: A general-purpose face recognition library with mobile applications," CMU-CS-16-118, CMU School of Computer Science, Tech. Rep., 2016.

[20] Universal Sentence Encoder Daniel Cera, Yinfei Yanga, Sheng-yi Konga, Nan Huaa, Nicole Limtiacob,Rhomni St. Johna, Noah Constanta, Mario Guajardo-C ÌĄespedesa, Steve Yuanc,Chris Tara, Yun-Hsuan Sunga, Brian Stropea, Ray Kurzweila.

[21] VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text, C.J. Hutto, Eric Gilbert, Georgia Institute of Technology, Atlanta, GA 30032.