Multimodal Emotion Recognition and Sentiment Analysis Using Masked Attention and Multimodal Interaction

Tatiana Voloshina, Olesia Makhnytkina
ITMO University
Saint Petersburg, Russian Federation
tatyana.shimohina23@gmail.com, makhnytkina@itmo.ru

***Abstract*—People express emotions verbally (with the linguistic part) and non-verbally (with facial expressions and speech tone). For better emotion recognition it is expedient to use both types of expressions. In this paper, a multimodal approach for emotion recognition based on fusion of textual, audio, and video data with masked multimodal attention and multimodal interaction are suggested. Our models is built on top of the language representation model Bidirectional Encoder Representations from Transformers (BERT). BERT is mostly used to work with text data, while the approaches with the interaction of text and audio modalities with fine-tuning a pre-trained BERT model are less common. In this work, a new 3-Modal Cross-BERT model that utilizes BERT fine-tuning based on textual, audio, and video data using masked multimodal attention and the model with multimodal interaction are proposed.**

Our algorithms were evaluated on publicly available multimodal sentiment and emotion analysis datasets CMU-MOSI, CMU-MOSEI, IEMOCAP and MELD. Experimental results show significant improvements in the performance across all metrics compared to the previous state-of-the-art methods for chosen datasets.

I. Introduction

Automatic emotion recognition and sentiment analysis are very close tasks. They are utilized in a range of applications such as video game development, education, patient care, car security, recruiting, smart home, etc(https://www.gartner.com/smarterwithgartner/13-surprising-uses-for-emotion-ai-technology).

Scientists and psychologists have been studying ways to express and show human emotions. Now there are three approaches to the study of the foundation of emotions: a dimensional (regression) model, a categorical (discrete) model of emotions, and mixed model.

In the discrete model (with a limited list of available classes), the emotional sphere consists of a certain number of primary, basic or fundamental emotions. Although each person expresses emotions differently. It has been established that a number of emotions are universal and can be understood regardless of human characteristics. However, different authors suggest a different number of basic emotions. P. Ekman [1] identifies six emotions (happiness, sadness, anger, fear, surprise, disgust), R. Plutchik [2] identifies eight basic emotions, which are the basis for all others and can be grouped into polar opposites: joy and sadness, acceptance and disgust, fear and anger, surprise and anticipation.

Limited number of variables (axes in space) are specified in the dimensional model. This model is focused on studying the similarities and differences between emotions and provides ways to express a wide range of emotional states. In this model, emotion is described using two or three fundamental characteristics, and affective states are expressed in a multidimensional space. Russell's model represents the affective state as a circle in a two-dimensional bipolar space [3]. Suggested dimensions are valence and arousal. Valence (pleasure) reflects positive or negative emotional states, and a value close to zero means a neutral emotion. Arousal expresses the active or passive component of emotion.

Usually researchers are only interested in measuring valency. In this case, observations are classified in a one-dimensional emotive space "positive-negative". This direction is called Sentiment analysis. As a rule, such a classification has two, three or five dimensions. In the first case, a classification is made into "positive" or "negative". In the second case, "neutrality" is added to these two dimensions. In the third case, each of the classes ("positive" or "negative") is expanded to two according to the degree of expression of this emotion.

The approaches proposed in the article will be applied both for recognizing emotions represented by a discrete model and for estimating valence (sentiment analysis) with emotional degree expressed in 5 classes (from -3 to 3).

According to various evaluations, the accuracy of manual emotion recognition ranges from 50% to 70%; based on audio data, the accuracy averages 70%. In the presence of high environmental noise (with a signal-to-noise ratio = +16 dB), the linguistic information is deformed to the point where listeners cannot recognize words but the perception of emotions is still possible with a probability of more than 50% [4]. The compilers of the IEMOCAP dataset estimate the accuracy of manual emotion recognition from video data to be 72% [5]. For most datasets, the accuracy of automatic emotion recognition appears inferior to human capabilities.

According to the leaderboards crowdsourcing efforts PapersWithCode, for the CMU-MOSEI dataset, the best accuracy of recognition of 7 sentiment classes reaches 52.0% [6]; for the MOSI dataset, the best result is 44.9% [7]. For the MELD

dataset, the best accuracy reaches 66.7% [8] for the emotion recognition in dialogues and 42.3% for emotion recognition in separated messages regardless of the context [9]. In our research we classify single utterance without context and previous dialogues. For the IEMOCAP dataset, the best result for emotion recognition is 68.2% [10].

Multiclass emotion or sentiment classification represents a more challenging task, so no experiments for binary classification were made. The article focuses only on multiclass classification. The purpose of our study is not only to represent fusion techniques, but also to pay attention to the different features representation and combinations. The most common approaches to joining modalities are at the feature level and at the decision level. The best performance is achieved by intermediate modality fusion and neural networks based on Graph-MFN (Memory Fusion Network) [11]. This paper proposes a new methods for multimodal emotion and sentiment recognition. First method was inspired by [7] and extends the proposed approach by including video modality and replacing audio features. The code is publicly available at https://github.com/T-Sh/3-Modal-Cross-Bert. The second model utilizes multimodal interactions on different levels and refines the previous approach.

II. Related works

The improvement of automatic emotion recognition remains a key issue. Generally, models with a fusion of different modalities (audio, video, and text) show better results. There is a fairly large number of papers about the automatic recognition of emotions, both in individual modalities and multimodal. Methods and feature's extractions contributed to better results.

The following vector representations are commonly used as text representations in recent years:

- FastText [12] is a vector representation of words with the simultaneous use of skipgram and C-BOW (Continuous Bag-of-Words) models. The dimension of the representation is 300, a lot of multi languages pre-trained models are publicity available, especially for English.
- Bert Embeddings [13] represents and connects the token itself (pre-trained), the number of its offer, and the position of the token within its. The input data is received and processed by the network in parallel, not sequentially, but the information about the mutual arrangement of words in the original sentence is stored, being included in the positional part of the embedding of the corresponding token. The Bidirectional Encoder Representations from Transformers (BERT) [14], Robustly Optimized BERT Pretraining Approach (RoBERTa) [15], Generative Pre-trained Transformer (GPT) [16] are commonly used for text data and shows SOTA results in a lot of tasks based on text data.

As characteristics and features of audio, 3 presentation options can be distinguished:

- Spectrogram represents the image that shows signal power spectral density versus time. It is required to analyze each segment of the signal. Although the original spectrum contains many components that are not essential for emotional recognition, spectrograms can find their application [17], [18].
- The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) [19] - two versions of the acoustic parameter set representation are proposed: a minimalist parameter set that implements the prosodic, excited, vocal, and spectral descriptors that have been found to be the most important, and an extension of the minimalist set (eGeMAPS) that contains additional descriptors that are claimed to improve accuracy automatic affect recognition compared to a set of pure prosodic and spectral parameters. The application for emotional recognition is presented in [20].
- Mel-Frequency Cepstral Coefficients (MFCC) [21] are another characteristics of speech signals. The chalk scale relates the perceived frequency or pitch of a pure tone to its actual measured frequency. Humans are much better at distinguishing small pitch changes at low frequencies than they are at high frequencies. This approach makes it possible to bring the machine perception of speech closer to the human one. The application is presented in [22].
- It is also possible to use the MFCC and eGeMAPS features at the same time, as shown in paper [23].

As shown in [24] and [25], various versions of convolutional and recurrent neural networks are typically used to extract features from video data. This models are also used as feature extractors and classifiers. The following architectures are the most common:

- RNN (recurrent neural network) [26] is recurrent neural network. Designed for modeling serial data, they are widely used in text and video [27] processing. The paper [28] shows the use of various modifications of RNN in the problem of multimodal sentiment recognition, however, all of them show low results (Concordance Correlation Coefficient ¡ 0.5) compared to other approaches.
- LSTM [29] is long short-term memory network. Compared to RNNs, are capable of storing long-term dependencies. It is similarly applied in various tasks. Along with RNN, LSTM configurations are used, but in the problem of multimodal analysis they show close results.

For video modality some of the possible neural architectures were evaluated, for example CNN and LSTM. Also, combining the received frames into sequences shows much better results [30]. In addition, other experiments were carried out to select the best number of frames, the best selection of n-th frames, the size of overlapping windows and etc.

Combining the selected features is possible at different stages, depending on the chosen approach:

- Earle fusion or data level fusion. Early fusion applies to raw data or pre-processed data from sensors. Features of the data must be extracted from the data before merging, otherwise the process will be complicated, especially when the data sources have different sampling rates

between modalities. Synchronization of data sources is also more complicated when one data source is discrete while others are continuous. Therefore, converting data sources into a single feature vector is a major challenge in the early stage of data fusion. There are two drawbacks to using data merge early on. One of the main disadvantages of this method is that a large amount of data will be subtracted from the modalities to arrive at a consensus before the merge. Once the data has common matrices, it is analyzed using a machine learning algorithm. Another disadvantage of this method is the synchronization of timestamps of different modalities. A common way to overcome this shortcoming is to collect data or signals at a common sample rate.

- Late fusion or decision level fusion. A late merge uses data sources independently, followed by a merge at the decision stage. This method is much simpler than the early merge method, especially when the data sources differ significantly from each other in terms of sample rate, data dimensions, and units. A late merge often gives better performance because problems from multiple models are handled independently so errors don't correlate. However, a number of researchers use late merging or decision-level merging to analyze problems with multimodal data [31], [32], [33].
- Intermediate level fusion. The architecture of the intermediate fusion is built on the basis of a deep neural network. This method is more flexible, allowing to combine data at different stages of model training. For example, in paper [34], unimodal features of each modality are combined using Attention Networks. Afterwards, the merged modalities go through the various RNN variants again to get the final sentiment polarity.
- Another fusion methods. The authors of [22] suggest using not only data from each modality, but also intramodal, intermodal, and interbimodal interactions as features. A bimodal information-oriented architecture based on multilevel attention has been developed to extract independent and consistent information from different modalities for efficient fusion. Also graph representations [35] can be used to efficiently represent modalities and combine them.

The Table I presents the results for our implementation of each modality individually and in various combinations. Experiments are provided for MOSI dataset due to it's small size and a fairly large variety of actors. As you can see, the main part of the information on sentiment is extracted from the textual modality. The video and audio modalities are rather complementary. In this work, the text modality is also taken as the basis, and the audio and video modalities are taken as additional ones.

TABLE I. COMPARISON OF THE ACCURACY OF ALGORITHMS FOR DIFFERENT APPLIED MODALITIES ON THE MOSI DATASET

Features	Model	Accuracy	F1	Precision	Recall
T	BERT	41.5	41.2	50.2	41.3
A	LSTM + Linear	25.59	25.63	32.76	20.1
V	2*3D-CNN + Linear	29.2	29.2	32.8	29.2
T+A	CM-BERT	44.7	44.6	44.3	43.9
T+A+V	3-Modal Cross-BERT	47.5	47.3	48.6	47.6

III. DATASETS & METHODS

A. Datasets

Datasets were preprocessed before models training. During preprocessing broken audio and video sources were removed, all data were represented in unificated form. Modalities were preprocesse separately. The information about resulted datasets are presented in Table II.

TABLE II. DATASETS INFORMATION

Dataset	№ of actors	№ of classes	№ of samples
MELD	260	7 emotions	13019
MOSEI	1000	6 emotions 7 sentiments	8206
MOSI	89	7 sentiments	2185
IEMOCAP	10	6 emotions	7368

CMU - MOSEI (CMU Multimodal Opinion Sentiment and Emotion Intensity) [11] is the largest dataset for emotionality evaluation. It contains over 65 hours of open source videos, 23,453 tagged videos from 1,000 speakers on 250 different topics. Videos from Youtube were selected according to certain conditions: the monologue format, the presence of only one person during the recording, shooting mainly from the front. The most common 3 topics are reviews (16.2%), debates (2.9%) and consultations (1.8%). For markup, the authors used the Ekman emotion system [36] with 6 emotions happiness, sadness, anger, fear, disgust, surprise and 7 sentiment classes. The dataset is unbalanced, the emotion of happiness and positive classes prevail.

CMU - MOSI (Multimodal Corpus of Sentiment Intensity and Subjectivity Analysis in Online Opinion Videos) [37] was created based on short (two to five minutes) video blogs from Youtube. A total of 93 videos were selected with 89 speakers, including 41 women and 48 men. Each video was then split into segments for a total of 3702 segments. Commonly, English is used, but not all speakers are native. The data are labeled according to 7 classes of sentiment. The dataset is unbalanced for different sentiment classes.

Dataset MELD [38] contains dialogues from the popular series Friends. The authors used the Ekman emotion system [36] with 6 emotions and the additional neutral class. The dataset contains over 1,400 dialogues and 13,000 utterances

from the series. Both already cut segments of the video with transcription are presented, as well as an indication of the season and episode. While other datasets consist of monologues or dialogues of only two people, this database aims to increase the number of participants in a conversation. It presents a more difficult task for researchers. Also, unlike other datasets, it often has a noisy soundtrack and off-screen laughter. Several people can be present in one frame as well, which makes it difficult to determine the speaker. The dataset is unbalanced.

IEMOCAP (Interactive Emotional Dyadic Motion Capture Database) [5]. Audio and video streams, text description of scenarios and face capture were used. Invited professional actors (5 men and 5 women) were recorded for 5 episodes of a couple of people each. As a result, the total size of the dataset is 302 video dialogues, about 12 hours. The markup was carried out in two dimensions. Discrete emotions were assessed, such as anger, excitement, fear, sadness, surprise, frustration, joy, annoyance, disappointment and neutral class, as well as continuous ratings on scales of valence (valence) (1 - negative, 5 - positive), activation (activation) (1 - calm, 5 - excited) and dominance (1 - weak, 5 - strong). There are multiple labels for one entry. Since weakly expressed emotions are poorly represented in the dataset (for example, fear, surprise, disgust), they were merged with the nearest strongly expressed emotion. As a result, the following set of classes was obtained: sadness, neutral, joy, delight, anger and disappointment.

B. Methods

In this paper, the 3-Modal Cross-BERT model, which combines textual, audio, and video data for emotion and sentiment recognition, and Multimodal Interaction Model, that carries modalities interactions on different levels of features processing, are suggested. The base algorithm is presented in Fig. 1. Both algorithms implement same feature extractions stage, but use different approaches in the fusion stage.

Algorithms receive a sequence of the first 50 tokens, a dedicated audio track, and a sequence of frames extracted from the video stream as input. The first 25 data from the video sequence were used. The following sections describe approaches for handling each modality and modalities fusions.

1) Text: The sequence of tokens is fed to the algorithm. The pre-trained Bert model extracts features from the sequence. Bert Embeddings connect pre-trained representations of the token itself, its sentence's number, and the position of the token within its sentence. The input data is received and processed by the network in parallel, not sequentially, but the information about the mutual arrangement of words in the original sentence is stored, being included in the positional part of the embedding of the corresponding token. The dimension of the representation is 768. A pre-trained model for the English language is used. For research, the bert-base-uncased (12-layer, 768-hidden, 12-heads, 110M parameters, trained on on BookCorpus, a dataset consisting of 11,038 unpublished books and English Wikipedia (excluding lists, tables and headers)) version from the transformers library is applied. Finally we get one-dimensional embedding for text utterance. Since the dimensionality of the data is different, a convolutional layer is used to control the dimensionality of the output.

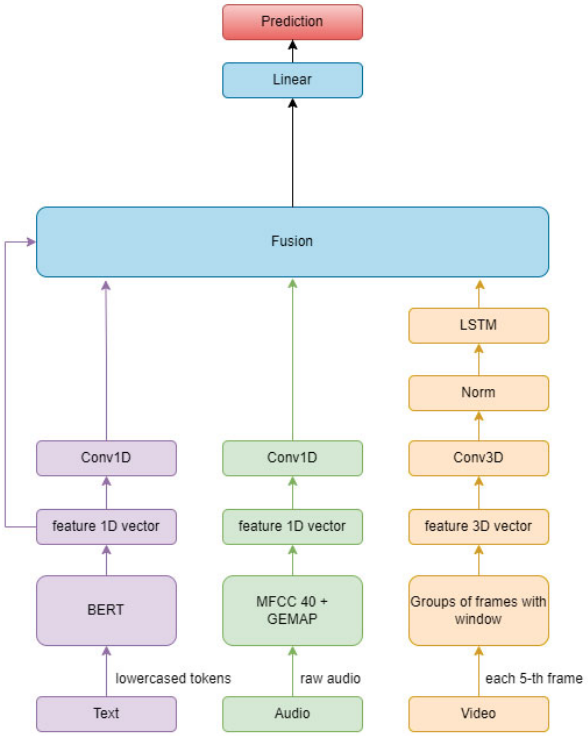

Fig. 1. Pipeline for the multimodal multiclass classification

2) Audio: 40-dimensional MFCC features are extracted using the Librosa library and 88-feature GEMAPS are extracted using the OPENSMILE library for audio tracks. Then the resulting vectors are combined into a single one-dimensional feature vector. The vector is then fed to the input of the convolutional layer to extract features with the desired dimension.

3) Video: First, every 5th frame was extracted from a video stream. Then faces were searched, cropped, and aligned with the help of OpenCV2 and DLib within the selected frames. After the faces were converted to grayscale, the size of the resulting images was reduced to 64 by 64 pixels. The sequence of two dimensional frames is formed into groups of 10 frames with an overlapping window between groups of 5 frames. Finally a three dimensional vector with a convolutional 3D layer is used for feature extraction and dimensionality control. Further, the results are normed and fed to the LSTM layer.

4) Fusion of Modalities with mask: The masked multimodal attention is designed to merge audio, video, and text modalities. A close-up of the fusion architecture is shown in Fig. 2.

Features $X_{modality}$ and Keys $K_{modality}$ for each modality are defined as $K_{modality} = X^{\top}_{modality}$. Then the attention matrix $W_{modality}$ is evaluated through matrix multiplication.

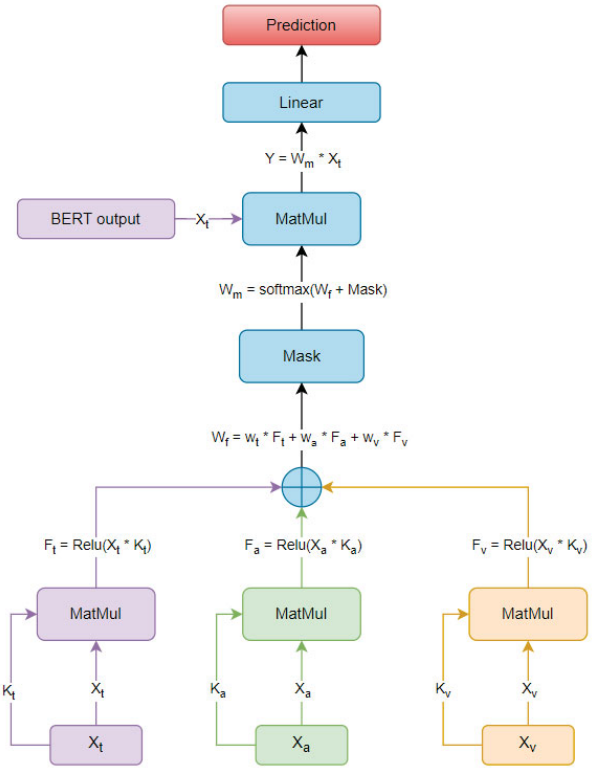

Fig. 2. The architecture of multimodal masked attention fusion

With the attention matrix and weights $w_{modality}$ the fusion attention matrix W_m is calculated. To reduce the influence of the padding sequence, a mask matrix is introduced. This matrix uses 0 to represent the token position and infinite—to represent the padding position (after the softmax function the attention score of the padding position will be 0) [7].

After obtaining the multimodal attention matrix, W_m multiplied by X_t to add text modality features and get the output of the attention Y. Then the final prediction of the class is calculated by a linear layer.

5) Fusion of Modalities with multimodal interaction: The previous method was then extended using inter-modal interaction and named Multimodal interaction model. Combinations are created with each modality in pairs, in triplets and in the union of pairs. The resulting sets are then merged into a result set, applying a per-set bias. For original features weights for each modality were applied. The method in details is presented in Fig. 3. Then the final prediction of the class is calculated by a linear layer.

IV. Experimental methodology

The datasets were pre-processed to extract features. Then each dataset was divided into training (80%) and testing (20%) subsets. Classes are evenly distributed in subsets for MOSEI, IEMOCAP and MELD datasets. Additionally, for the MOSI dataset, the partition from [7] is recreated.

Following training parameters were used: the AdamW optimizer with a learning rate = $2e-6$ and the BCELoss loss and Cross Entropy functions for masked fusion and interaction fusion respectively. The pre-trained BERT model is also fine-tined. For training the 3-Modal Cross-BERT model, the batch size is setted to 24 and the max sequences length is setted to 50. For each dataset, the model was trained separately for 80 epochs. The best results were selected based on the test sample from all epochs.

The following libraries for the Python 3 programming language were used for experiments:

- pytorch - creation of basic models, usage of structures and methods for experiments.
- scikit-learn - use of metrics for evaluation.
- transformers - implementation of Transformer class models, like BERT.
- librosa - audio MFCC features extraction.
- openSMILE - audio eGEMAPS features extraction.
- Dlib and OpenCV - determination of the face in the image, cropping and straightening of the image;

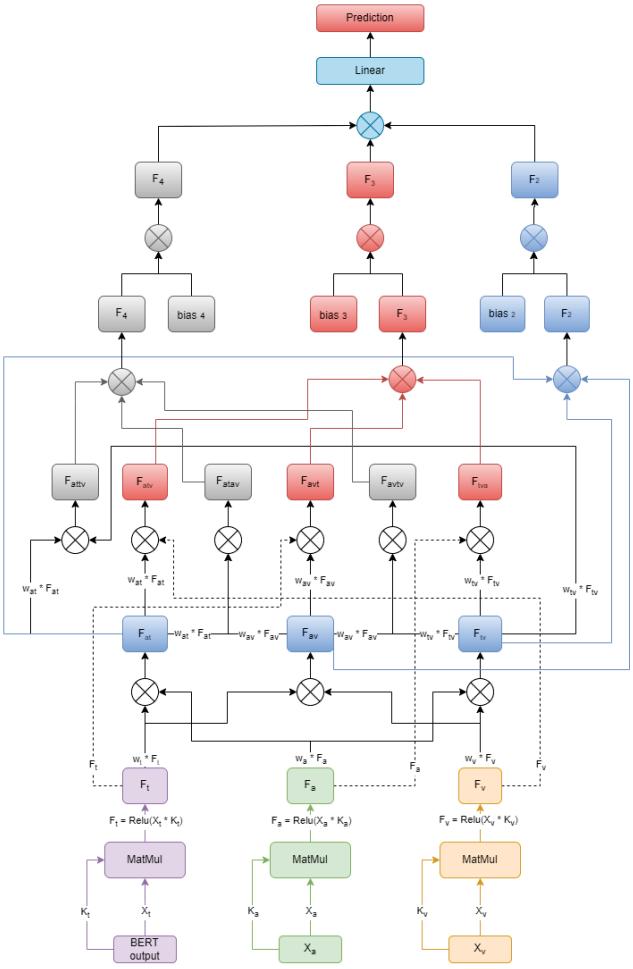

Fig. 3. The architecture of multimodal interaction fusion

TABLE III. RESULTS AND COMPARISON WITH SOTA MODELS

Dataset	Task	Classes	Model	Weighted Accuracy	Weighted F1	Weighted Precision	Weighted Recall
MELD	emotion recognition	7	HGFM [9]	42.3	-	-	-
			3-Modal Cross-BERT	44.0	45.0	52.7	41.3
			Multimodal interaction	**62.2**	60.7	60.8	62.2
MOSEI	sentiment analysis	7	MMLatch [6]	52.1	-	-	-
			3-Modal Cross-BERT	58.4	59.0	62.2	58.4
			Multimodal interaction	**63.3**	61.0	61.2	63.4
MOSEI	emotion recognition	6	3-Modal Cross-BERT	34.7	33.2	49.6	34.7
			Multimodal interaction	52.4	46.0	43.5	52.4
			MMIM [39]	**54.2**	-	-	-
MOSI	sentiment analysis	7	CM-BERT [7]	44.9	-	-	-
			3-Modal Cross-BERT	48.5	48.2	49.8	48.5
			Multimodal interaction	**51.0**	50.2	50.5	50.1
IEMOCAP	emotion recognition	6	3-Modal Cross-BERT	34.0	34.0	37.0	33.0
			Multimodal interaction	57.0	57.0	57.0	57.0
			COGMEN [10]	**68.2**	67.6	-	-

V. RESULTS

Evaluations of the following results were presented: the emotion recognition results for the MELD, MOSEI and IEMOCAP and the sentiment analysis results for MOSI and MOSEI. In our experiments, consistent with the previous work [7], the same metrics were used to evaluate the performance of the baselines and our model. Multiclass weighted accuracy (WA) and F1 score are selected.

The performance of 3-Modal Cross-BERT and Interaction model was compared with previous models on the multimodal sentiment analysis and emotion recognition tasks. Results are presented in Table III. The chosen models for comparison are:

- Hierarchical grained and feature model (HGFM), where the frame-level and utterance-level structures of acoustic samples were processed by the recurrent neural network. The model included a frame-level representation module with before and after information, an utterance-level representation module with context information, and a different level acoustic feature fusion module [9].
- MMLatch, a neural network module that used representations from higher levels of the architecture to create top-down masks for the low-level input features. Mechanism extracted high-level representations for each modality and used these representations to mask the sensory inputs, allowing the model to perform top-down feature masking [6].
- Cross-Modal BERT (CM-BERT), which relied on the interaction of text and audio modality to fine-tune the pre-trained BERT model. As the core unit of the CM-BERT, masked multimodal attention was designed to dynamically adjust the weight of words by combining the information of text and audio modality [7].
- MMIM, MultiModal InfoMax, which hierarchically maximizes the Mutual Information (MI) in unimodal input pairs (inter-modality) and between multimodal fusion result and unimodal input in order to maintain task-related information through multimodal fusion. The framework is jointly trained with the main task (MSA) to improve the performance of the downstream MSA task. To address the intractable issue of MI bounds, a set of computationally simple parametric and non-parametric methods were formulated to approximate their truth value [39].
- COGMEN, COntextualized Graph Neural Network based Multimodal Emotion recognitioN (COGMEN) system that leverages local information (i.e., inter/intra dependency between speakers) and global information (context). The model uses Graph Neural Network (GNN) based architecture to model the complex dependencies, local and global information, in a conversation [10].

Based on Table III, it is easy to see that the Multimodal Interaction model produces new state-of-the-art results on chosen datasets and improves the performance on weighted accuracy. For the task of emotion recognition without context, an improvement of 1.7% was achieved for the MELD dataset. For the sentiment analysis task, the classification accuracy for the MOSI and MOSEI datasets increased by 3.6% and 6.3%, respectively.

Confusion matrices for each dataset with the best results are presented in Fig. 4- 8. For the sentiment analysis problem, the most mistakes are confusions of close classes. For example, for the MOSEI dataset, an erroneous assignment of neutral messages to weakly positive ones and vice versa. This is

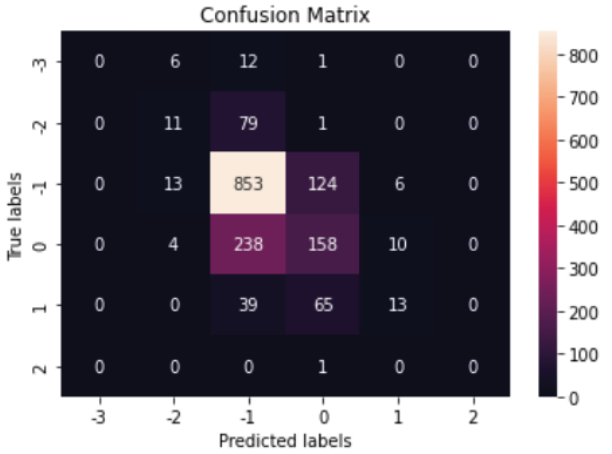

Fig. 4. MOSEI confusion matrix (sentiment)

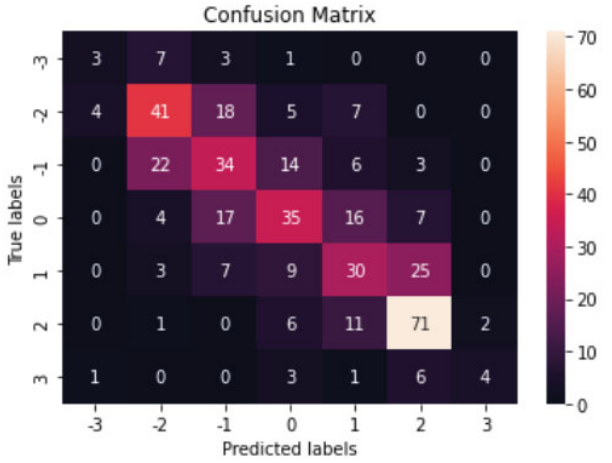

Fig. 5. MOSI confusion matrix (sentiment)

because people show a weak positive attitude almost as neutral.

For the task of emotion recognition, the largest part of the errors is associated with the confusion of a neutral emotion with the rest, which is explained by a possibly weak expression of emotions. This is especially evident in the results for the MELD dataset. For the MOSEI dataset algorithm usually misplaced emotions with happiness. This may be due to the stronger expression of happiness compared to others emotions. The model, trained on IEMOCAP dataset, commonly classifies emotions correctly, but like the other versions, confuses neutral class and close emotions (like anger and frustration).

VI. Conclusions

In this paper, new 3-Modal Cross-BERT model and Multimodal Interaction model for multiclass sentiment analysis and emotion recognition were proposed. An extension of the pretrained BERT model with audio and video modalities were suggested. Audio and video data are used to fine-tune the textual BERT model through the use of masked multimodal attention. Research has been conducted with various methods

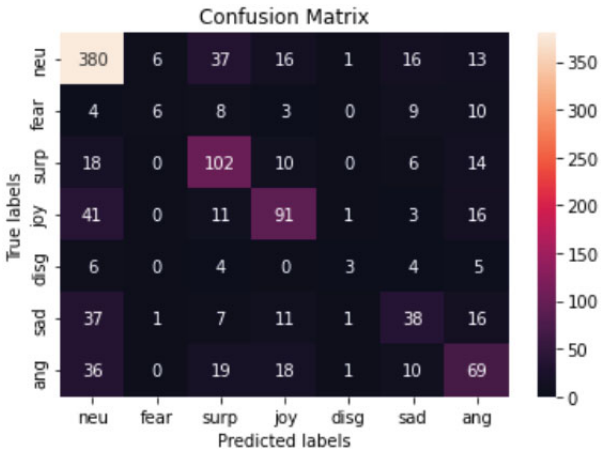

Fig. 6. MELD confusion matrix (emotion)

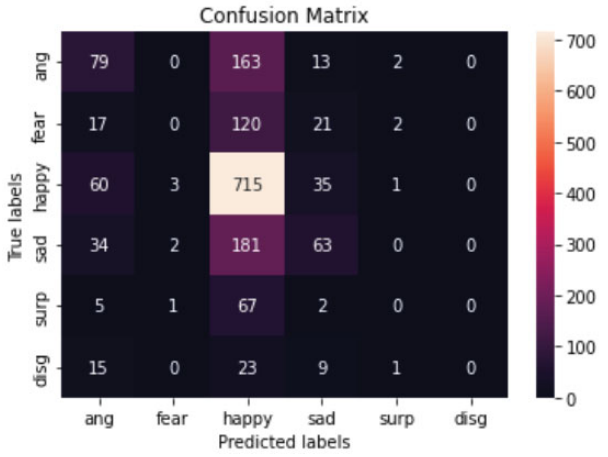

Fig. 7. MOSEI confusion matrix (emotion)

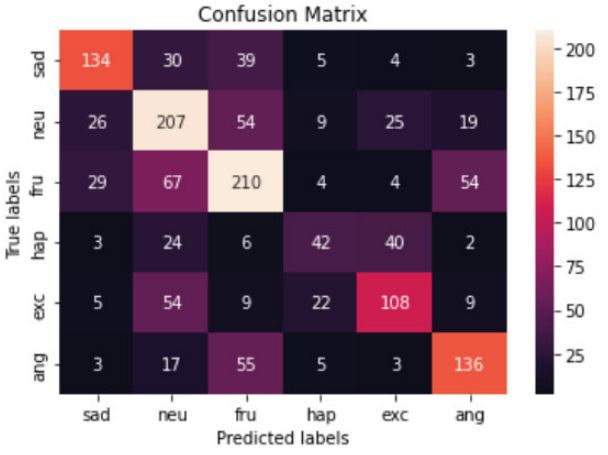

Fig. 8. IEMOCAP confusion matrix (emotion)

for extracting features from audio and video data. The best results for audio data were obtained using 40-dimensional MFCC and 88-feature eGEMAPS, a combination of convolutional 3D layer and LSTM for video data.

Experimental results show that 3-Modal Cross-BERT and Multimodal Interaction model significantly improved accuracy on MELD, MOSI, and MOSEI data compared to the previous state-of-the-art models. The value of including video modality is being proven experimentally. For example in the MOSI dataset, adding modality improved accuracy by 3.4%. It also managed to outperform the state-of-the-art models for other datasets, which indicates the validity of adding a video modality. The work presents only the best parameters and features sets.

ACKNOWLEDGMENT

The research was financially supported by the Russian Science Foundations (project 22-11-00128).

REFERENCES

[1] P. Ekman, E. R. Sorenson, and W. V. Friesen, "Pan-cultural elements in facial displays of emotion," *Science*, vol. 164, pp. 86 – 88, 1969.

[2] R. Plutchik, "Psychophysiology of individual differences with special reference to emotions*," *Annals of the New York Academy of Sciences*, vol. 134, pp. 776 – 781, 12 2006.

[3] J. Russell, "A circumplex model of affect," *Journal of Personality and Social Psychology*, vol. 39, pp. 1161–1178, 12 1980.

[4] V. Morozov, *Emotional language and emotional ear. Selected works*. Institute of Psychology of the Russian Academy of Sciences, 2017. [Online]. Available: https://books.google.ru/books?id=aaNUDwAAQBAJ

[5] C. Busso, M. Bulut, C.-C. Lee, E. A. Kazemzadeh, E. M. Provost, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "Iemocap: interactive emotional dyadic motion capture database," *Language Resources and Evaluation*, vol. 42, pp. 335–359, 2008.

[6] G. Paraskevopoulos, E. Georgiou, and A. Potamianos, "Mmlatch: Bottom-up top-down fusion for multimodal sentiment analysis," 2022. [Online]. Available: https://arxiv.org/abs/2201.09828

[7] K. Yang, H. Xu, and K. Gao, "Cm-bert: Cross-modal bert for text-audio sentiment analysis," in *Proceedings of the 28th ACM International Conference on Multimedia*, ser. MM '20. New York, NY, USA: Association for Computing Machinery, 2020, p. 521–528. [Online]. Available: https://doi.org/10.1145/3394171.3413690

[8] V. Chudasama, P. Kar, A. Gudmalwar, N. Shah, P. Wasnik, and N. Onoe, "M2fnet: Multi-modal fusion network for emotion recognition in conversation," 06 2022.

[9] Y. xu, H. Xu, and J. Zou, "Hgfm : A hierarchical grained and feature model for acoustic emotion recognition," 05 2020, pp. 6499–6503.

[10] A. Joshi, A. Bhat, A. Jain, A. Singh, and A. Modi, "COGMEN: COntextualized GNN based multimodal emotion recognitioN," in *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Seattle, United States: Association for Computational Linguistics, Jul. 2022, pp. 4148–4164. [Online]. Available: https://aclanthology.org/2022.naacl-main.306

[11] A. Zadeh, P. P. Liang, S. Poria, E. Cambria, and L.-P. Morency, "Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph," in *ACL*, 2018.

[12] A. Joulin, E. Grave, P. Bojanowski, and T. Mikolov, "Bag of tricks for efficient text classification," 07 2016.

[13] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," *CoRR*, vol. abs/1810.04805, 2018. [Online]. Available: http://arxiv.org/abs/1810.04805

[14] S. Padi, S. Sadjadi, D. Manocha, and R. Sriram, "Multimodal emotion recognition using transfer learning from speaker recognition and bert-based models," 02 2022.

[15] T. Kim and P. Vossen, "Emoberta: Speaker-aware emotion recognition in conversation with roberta," 08 2021.

[16] J. Liu, D. Shen, Y. Zhang, B. Dolan, L. Carin, and W. Chen, "What makes good in-context examples for GPT-3?" in *Proceedings of Deep Learning Inside Out (DeeLIO 2022): The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*. Dublin, Ireland and Online: Association for Computational Linguistics, May 2022, pp. 100–114. [Online]. Available: https://aclanthology.org/2022.deelio-1.10

[17] H. Kaya, A. A. Salah, A. Karpov, O. Frolova, A. Grigorev, and E. Lyakso, "Emotion, age, and gender classification in children's speech by humans and machines," *Computer Speech Language*, vol. 46, pp. 268–283, 2017. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0885230816301346

[18] C. Luna Jiménez, D. Griol, Z. Callejas, R. Kleinlein, J. Montero, and F. Fernández-Martínez, "Multimodal emotion recognition on ravdess dataset using transfer learning," *Sensors*, vol. 21, p. 7665, 11 2021.

[19] F. Eyben, K. R. Scherer, B. Schuller, J. Sundberg, E. André, C. Busso, L. Devillers, J. Epps, P. Laukka, S. S. Narayanan, and K. P. Truong, "The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing," *IEEE Transactions on Affective Computing*, vol. 7, pp. 190–202, 2016.

[20] Z. Peng, J. Dang, M. Unoki, and M. Akagi, "Multi-resolution modulation-filtered cochleagram feature for lstm-based dimensional emotion recognition from speech," *Neural Networks*, vol. 140, pp. 261–273, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0893608021001155

[21] M. Padmanabhan and M. Picheny, "Comparison of parametric representation for monosyllabic word recognition in continuously spoken sentences," 2017.

[22] T. Wu, J. Peng, W. Zhang, H. Zhang, S. Tan, F. Yi, C. Ma, and Y. Huang, "Video sentiment analysis with bimodal information-augmented multi-head attention," *Knowledge-Based Systems*, vol. 235, p. 107676, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0950705121009382

[23] E. Ryumina, O. Verkholyak, and A. Karpov, "Annotation confidence vs. training sample size: Trade-off solution for partially-continuous categorical emotion recognition," 08 2021, pp. 3690–3694.

[24] M. Singh and Y. Fang, "Emotion recognition in audio and video using deep neural networks," 06 2020.

[25] ——, "Emotion recognition in audio and video using deep neural networks," 06 2020.

[26] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *Nature*, vol. 323, pp. 533–536, 1986.

[27] E. Cambria, D. Hazarika, S. Poria, A. Hussain, and R. Subramanyam, "Benchmarking multimodal sentiment analysis," 07 2017.

[28] D. Kollias and S. Zafeiriou, "Exploiting multi-cnn features in cnn-rnn based dimensional emotion recognition on the omg in-the-wild dataset," *IEEE Transactions on Affective Computing*, vol. 12, pp. 595–606, 2019.

[29] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, pp. 1735–80, 12 1997.

[30] D. Dresvyanskiy, E. Ryumina, H. Kaya, M. Markitantov, A. Karpov, and W. Minker, "End-to-end modeling and transfer learning for audiovisual emotion recognition in-the-wild," *Multimodal Technologies and Interaction*, vol. 6, no. 2, 2022. [Online]. Available: https://www.mdpi.com/2414-4088/6/2/11

[31] D. Kollias and S. Zafeiriou, "Exploiting multi-cnn features in cnn-rnn based dimensional emotion recognition on the omg in-the-wild dataset," 10 2019.

[32] S. Amiriparian, A. Sokolov, I. Aslan, L. Christ, M. Gerczuk, T. Hübner, D. Lamanov, M. Milling, S. Ottl, I. Poduremennykh, E. Shuranov, and B. Schuller, "On the impact of word error rate on acoustic-linguistic speech emotion recognition: An update for the deep learning era," 04 2021.

[33] L. Stappen, A. Baird, L. Christ, L. Schumann, B. Sertolli, E. Messner, E. Cambria, G. Zhao, and B. Schuller, "The muse 2021 multimodal sentiment analysis challenge: Sentiment, emotion, physiological-emotion, and stress," 04 2021.

[34] A. Agarwal, A. Yadav, and D. Vishwakarma, "Multimodal sentiment analysis via rnn variants," 05 2019, pp. 19–23.

[35] S. Mai, H. Hu, and S. Xing, "Modality to modality translation: An adversarial representation learning and graph fusion network for

multimodal fusion," 2019. [Online]. Available: https://arxiv.org/abs/1911.07848

[36] P. Ekman, W. V. Freisen, and S. Ancoli, "Facial signs of emotional experience." *Journal of Personality and Social Psychology*, vol. 39, pp. 1125–1134, 1980.

[37] A. Zadeh, R. Zellers, E. Pincus, and L.-P. Morency, "Mosi: Multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos," *ArXiv*, vol. abs/1606.06259, 2016.

[38] S. Poria, D. Hazarika, N. Majumder, G. Naik, E. Cambria, and R. Mihalcea, "Meld: A multimodal multi-party dataset for emotion recognition in conversations," *ArXiv*, vol. abs/1810.02508, 2019.

[39] W. Han, H. Chen, and S. Poria, "Improving multimodal fusion with hierarchical mutual information maximization for multimodal sentiment analysis," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 9180–9192. [Online]. Available: https://aclanthology.org/2021.emnlp-main.723