

Multimodal Emotion Recognition Using Transfer Learning on Audio and Text Data

James J. Deng[1(✉)], Clement H. C. Leung[2(✉)], and Yuanxi Li[3]

[1] MindSense Technologies, Hong Kong, People's Republic of China
james@mindsense.ai
[2] School of Science and Engineering, The Chinese University of Hong Kong, Shenzhen, People's Republic of China
clementleung@cuhk.edu.cn
[3] Hong Kong Baptist University, Hong Kong, People's Republic of China
csyxli@comp.hkbu.edu.hk

Abstract. Emotion recognition has been extensively studied in a single modality in the last decade. However, humans express their emotions usually through multiple modalities like voice, facial expressions, or text. In this paper, we propose a new method to find a unified emotion representation for multimodal emotion recognition through speech audio, and text. Emotion-based feature representation from speech audio is learned by an unsupervised triplet-loss objective, and a text-to-text transformer network is constructed to extract latent emotional meaning. As deep neural network models trained by huge datasets exhaust a lot of unaffordable resources, transfer learning provides a powerful and reusable technique to help fine-tune emotion recognition models trained on mega audio and text datasets respectively. Automatic multimodal fusion of emotion-based features from speech audio and text is conducted by a new transformer. Both the accuracy and robustness of proposed method are evaluated, and we show that our method for multimodal fusion using transfer learning in emotion recognition achieves good results.

Keywords: Multimodal emotion recognition · Multimodal fusion · Transformer network

1 Introduction

In the last decade, deep learning like convolution neural network (CNN), recurrent neural network (RNN), and other deep models have proven useful in many domains, including computer vision, speech and audio processing, and natural language processing. Many applications like face recognition, speech recognition, and machine translation have achieved great success. Research on emotion recognition makes quick improvements as well. However, as emotion is complex and determined by a joint function of pharmacological, cognitive, and environmental variables, emotion recognized by different people may be ambiguous or even

© Springer Nature Switzerland AG 2021
O. Gervasi et al. (Eds.): ICCSA 2021, LNCS 12951, pp. 552–563, 2021.
https://doi.org/10.1007/978-3-030-86970-0_39

opposite. The single modality for emotion recognition is insufficient and incomplete. For example, in a conversation, people's voices, text of speech content, facial expressions, body language, or gestures all convey emotional meanings. Thus, it is inappropriate to recognize people's emotions through a single modality. Multimodal modeling is a natural and reasonable process for emotion recognition. Although multimodal analysis has been extensively studied and some have achieved remarkable results in specific constraints, it cannot simply apply these methods in different environments. Therefore, building a universal representation for emotional information from multiple modalities is necessary and useful in the task of multimodal emotion recognition. In this paper, to simplify, we propose a method to find a good emotional information representation using both speech audio and text information for emotion recognition.

Nowadays training models using mega dataset consumes huge resource and have become more and more difficult and less affordable for most of researchers, institutes or companies. For example, training GPT-3 would cost at least $4.6 million, because training deep learning models is not a clean, one-shot process. There are a lot of trial and error and hyperparameter tuning that would probably increase the cost sharply. Transfer learning provides a powerful and reusable technique to help us solve resource shortages and save model training costs. This paper adopts this strategy and employs the excellent fruits of pre-trained models by mega datasets of audio and text as the basis of our work. The pre-trained model of speech audio is trained on a Youtube-8M dataset named AudioSet by Google. This is an extremely large-scale dataset of manually annotated audio clips. A model named VGGish is an audio feature embedding produced by training a modified VGGNet model to predict video-level tags from this dataset, which is widely used for audio classification. In addition, another model TRIpLet Loss network (TRILL) is trained from AudioSet again and achieves good results for several audio tasks like speaker identification, and emotion recognition. We use the fine-tuned TRILL model to extract the speech audio features as the representation for the modality of speech audio. As for text representation, we adopt fine-tuned Text-To-Text Transfer Transformer (T5) [23] model trained by a common crawl (C4) dataset to extract text embeddings. Operations of transfer learning have greatly accelerated model training and applied in specific domains. To reuse the fruits of transfer learning especially for learning embeddings from speech audio and text, we adopts the strategy of transfer learning to construct emotional information representation.

We expect the fused data is more informative and synthetic. To retain more hidden emotional information for multimodal fusion, we concatenate speech audio features and text embeddings extracted from fine-tuned pertained models. The concatenated feature representation is passed to a transformer-based encoder-decoder network with an attention mechanism. This means that the final predictor module can balance the influence of information transmission. A quantitative evaluation shows that our models outperform the existing methods in emotion recognition. The main contribution of this paper is summarized as follows: (1) we propose an effective method to fuse multiple modalities on

the basis of transfer learning; (2) we evaluate the performance of the proposed method and make an attempt at a general semantic framework. The rest of the paper is structured as follows: Sect. 2 describes the literature review, Sect. 3 discusses the proposed methodologies and overall architecture, Sect. 4 explains the experimental setup and results, and Sect. 5 summarizes our work and discusses the future work.

2 Literature Review

In psychology and philosophy, emotion is a highly complex subjective, and conscious experience. There is no clear definition of what an emotion is. Consequently, different emotion theories or models have been proposed by different researchers in the last century. These models are usually rooted in two emotion theories: discrete emotion theory and dimensional emotion theory. Discrete emotion theory employs a finite number of emotional descriptors or adjectives [11] to express basic human emotions (e.g., joy, sadness, anger, contempt, happiness). For example, Ortony et al. [19] proposed an emotion cognition model commonly known as the OCC model to hierarchically describe 22 emotion descriptors. For the sake of simplicity, some researchers [21] used coarser-grained partition such as happy, neutral, and sad. Dimensional emotion theory states that emotion should be depicted in a psychological dimensional space, which can resolve the shortcomings of discrete emotion theory. Two or three dimensional emotion models such as arousal-valence [24], resonance-arousal-valence [3], and arousal-valence-dominance [16] are widely used in different application domains for example music emotion recognition. Dimensional emotion theory is more likely to be used in computational emotion systems or emotion synthesis. The advantage of discrete emotion theory is that it is easy to understand and use in practice, especially in emotion recognition, In this paper, for simplicity, we shall adopt discrete emotion theory to represent emotion.

In recent years many research works of emotion recognition have been done in a single modality setting. A survey [5] of methods is summarized to address three important aspects of the design of a speech emotion recognition system. The first one is the choice of suitable features for speech representation. The second issue is the design of an appropriate classification scheme and the third issue is the proper preparation of an emotional speech database for evaluating system performance. Domain expert knowledge makes significant for manually constructing high-quality features [18]. Recent research has mostly focused on deep representation learning methods, either supervised, semi-supervised, or unsupervised. Successful representations improve the sample efficiency of ML algorithms by extracting most information out of the raw signal from the new data before any task-specific learning takes place. This strategy has been used successfully in many application domains. Many deep learning methods like CNN [10], Deep Belief Network (DBN) [9], Long Short-Term Memory (LSTM) [28], Autoencoder [2,13] have been used to construct various deep neural networks, achieving a good performance in speech emotion recognition. Sentiment analysis

[15] is the task of automatically determining from the text the attitude, classified by positive, negative, and neutral attitude. Knowledge-based and statistical methods are usually used to extract text features like word embedding (e.g., Word2Vec), pair-wise correlation of words, and parts-of-speech (POS) tag of the sentence. Recently, transformer architecture [26] is rather popular in dealing with natural language processing like General Language Understanding Evaluation (GLUE) [27]. The model of Bidirectional Encoder Representations from Transformers (BERT) [4] is constructed by 12 Encoders with bidirectional self-attention mechanism. A unified text-to-text format, in contrast to BERT-style models that can only output either a class label or a span of the input, can flexibly be used on the same model, loss function, and hyperparameters on any NLP tasks. This unified framework provides valuable insight for performing semantic classification like recognizing emotion from the text. This paper inherits this text-to-text transformer architecture.

Transfer learning [20] and domain adaptation have been extensively practiced in machine learning. In specific domains, there is often only a few data available, and it is difficult to train an accurate model by using these small datasets. However, many modalities (e.g., speech, text) have the same essence and low-level elements. Thus, transfer learning can well overcome the small dataset limitation. A sharing learned latent representation [6] or deep models like VGGish or BERT is transferred to be used on another learning task, usually achieving an exciting performance. [14] uses DBN of transfer learning to build sparse autoencoder for speech emotion recognition. [12] introduces a sent2affect framework, a tailored form of transfer learning to recognize emotions from the text. Another Universal Language Model Fine-tuning (ULMFiT) [8] is carried out to evaluate several text classification and outperforms state-of-the-art models. Therefore, to make full use of pre-trained models by mega dataset, we adopt transfer learning of multiple modalities. Though there exists some work on multimodal emotion recognition, for example, canonical correlational analysis [17], joint feature representation [22,25], or generative adversarial network (GAN) for multimodal modeling, there is less work on the aspect of transfer learning on multimodal. This inspires us to use transfer learning and multimodal fusion to build a unified emotional representation of speech audio and text.

3 Methodology

This paper concentrates on dual modalities: speech and text. The aim is to find a unified and reusable representation of emotional information from these dual modalities. This section will first introduce the emotional features of speech audio, and text, respectively. Then the proposed method for fusing speech and text for multimodal emotion recognition will be described.

3.1 Emotion-Based Feature Representation for Audio Data

As speech recognition has been extensively researched and achieved great success in the last decade, many speech emotion recognition tasks also adopt acoustic

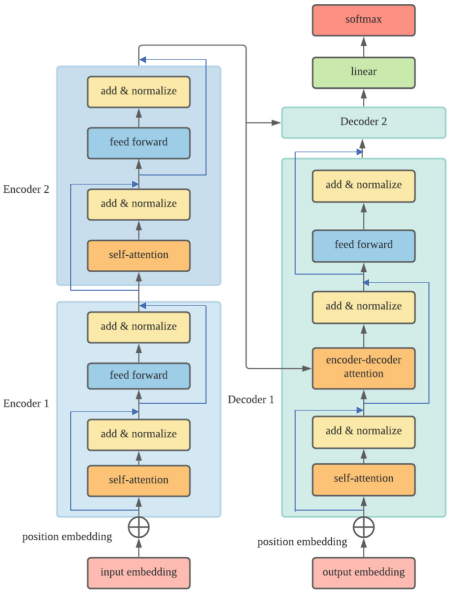

Fig. 1. An encoder-decoder Transformer architecture with residual skip connections.

features like Mel-frequency Cepstral Coefficient (MFCC), deep auto-encoders in Deep Neural Network (DNN), or end-to-end with attention-based strategy. However, the same content of speech usually expresses the different emotions corresponding to different voice attributes like pitch, rhythm, timbre, or context. Emotional Aspects of the speech signal generally change more slowly than the phonetic and lexical aspects used to explicitly convey meaning. Therefore, we need to find a good representation for emotion-related tasks to be considerably more stable in time than what is usually adopted in speech recognition.

To consider the temporal characteristic of speech audio, we represent a large and unlabeled speech collection as a sequence of spectrogram context windows $X = x_1, x_2, \ldots, x_N$, where each $x_i \in^{F \times T}$. We aim to learn a embedding $g :^{F \times T} \rightarrow^{d}$ from spectrogram context windows to a $d-$dimensional embedding space such that $\|g(x_i) - g(x_j)\| \leq \|g(x_i) - g(x_k)\|$ when $i - j \leq i - k$. We can express this embedding formulated by learning a triplet loss function. Suppose a large collection of example triplets is represented by $z = (x_i, x_j, x_k)$, where $i - j \leq \tau$ and $i - k \rangle \tau$ for some suitably chosen time scale τ. The τ represent the specific duration of each given audio clip. The whole loss function $\Theta(z)$ is expressed as follows:

$$\Theta(z) = \sum_{i=1}^{N} [\|g(x_i) - g(x_j)\|_2^2 + \|g(x_i) - g(x_k)\|_2^2 + \delta] \tag{1}$$

where $\|\bullet\|$ is the L_2 norm, $[\bullet]$ represents standard hinge loss and δ is non-negative margin hyperparameter.

3.2 Text-to-Text Transformer Representation

Since the transformer architecture has achieved high performance in a number of NLP tasks like The General Language Understanding Evaluation (GLUE) benchmark, Sentiment analysis (SST-2), SQuAD question answering, we adopt the transformer architecture as well. Given a sequence of text obtained from speech, we first map the tokens of the initial input sequence to an embedding space, and then input the embedded sequence to the encoder layer. The encoder layer is composed of a stack of blocks, and each block consists of a self-attention layer followed by a small feed-forward network. Layer normalization in the self-attention layer and feed-forward network is considered, where the activations are only re-scaled and no additive bias is applied. In addition, a residual skip is connected from input to output of the self-attention layer and feed-forward network, respectively. After that, the dropout is calculated within the feed-forward network, on the skip connection, on the attention weights, and at the input and output of the entire stack. As for the decoder layer, except for the similar block in the encoding layer, it contains a standard attention operation after each self-attention layer. The self-attention mechanism in the decoder also uses a form of auto-regressive or causal self-attention, which only allows the model to attend to past outputs. The output of the final decoder block is feed into a dense layer with a softmax output, whose weights are shared with the input embedding matrix. All attention mechanisms in the transformer are split up into independent "heads" whose outputs are concatenated before being further processed. The whole text-to-text transformer architecture is illustrated in Fig. 2. In a transformer, instead of using a fixed embedding for each position, relative position embeddings produce a different learned embedding according to the offset between the "key" and "query" being compared in the self-attention mechanism. We use a simplified form of position embeddings where each "embedding" is simply a scalar that is added to the corresponding logit used for computing the attention weights.

3.3 Multimodal Fusion of Speech Audio and Text

Considering the success of transformer architecture, we also adopt this attention-enhanced architecture in the fusion of speech audio features and text embeddings extracted along the operations described in previous subsections. As the speech consists of both audio and corresponding text, to retain more emotional information used for multimodal fusion, we directly concatenate the speech audio features and text embeddings. The combined features are processed by a transformer model, and the output of the model is regarded as the fusion of multimodal. The overall fusion of speech audio and text is illustrated in Fig. 2. In the process of fusion, we want to minimize the loss of both global and regional segments that convey emotional information. Speech audio features and texts are treated as sequences to be encoded into the learning procedure, while on the decoder side the target fused emotional features are generated. Suppose concatenated features of emotional representation of speech audio and text are denoted

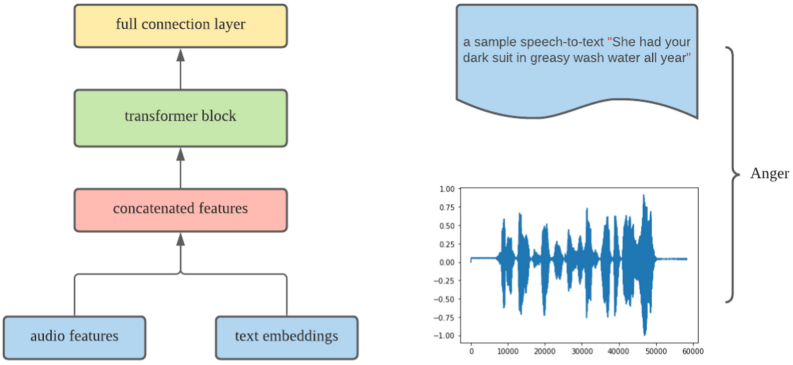

Fig. 2. Left part shows multimodal fusion of speech audio and text on embeddings extracted from transfer learning; right part shows an sample speech with anger emotion.

by x, the reconstruction x' of the same shape as x is achieved by decoder, the simplest loss function is calculated by

$$\Phi = ||x - x'||^2 \tag{2}$$

We add an explicit regularizer in the above loss function that forces the model to learn an encoding robust to slight variations of input features extracted from transfer learning. The regularization loss term is regarded as the squared Frobenius norm $\| A \|_F$ of the Jacobian matrix J for the hidden layer activations with respect to the input features. The final loss function is represented by

$$\Phi = ||x - x'||^2 + \lambda \sum_i \| \bigtriangledown_x h_i \|^2 \tag{3}$$

where $\bigtriangledown_x h_i$ refers to the gradient field of hidden layer activations with respect to the input x, summed over all i training samples. The Frobenius norm and Jacobian matrix J are given as follows:

$$\| A \|_F = \sqrt{\sum_{i=1}^{m} \sum_{j=1}^{k} \| a_{ij} \|^2} \tag{4}$$

$$J = \begin{pmatrix} \frac{\delta a_1^{(h)}(x)}{\delta x_1} & \cdots & \frac{\delta a_1^{(h)}(x)}{\delta x_m} \\ \frac{\delta a_2^{(h)}(x)}{\delta x_1} & \cdots & \frac{\delta a_2^{(h)}(x)}{\delta x_m} \\ \vdots & \ddots & \vdots \\ \frac{\delta a_n^{(h)}(x)}{\delta x_1} & \cdots & \frac{\delta a_n^{(h)}(x)}{\delta x_m} \end{pmatrix} \tag{5}$$

Finally, we consider the emotional information representation as multimodal fusion that satisfies the minimization of the loss function in Eq. 3.

Table 1. Comparison of multimodal fusion performance with single modality through the different embeddings on different emotion datasets.

Models	IEMOCAP	SAVEE	Mean
VGGis FC1	65.3%	57.7%	61.5%
VGGish Finetuned	61.4%	59.3%	60.4%
YAMNet layer 10	63.2%	62.3%	62.8%
YAMNet Finetuned	67.6%	62.7%	65.2%
TRILL distilled	70.5%	67.8%	69.2%
TRILL Finetuned	73.8%	68.6%	71.2%
Text-to-Text Transformer (T5)	75.7%	72.3%	75.5%
TRILL-T5 Multimodal Fusion	81.3%	77.4%	79.4%

4 Experiments and Results

Before presenting the results from our empirical study, we briefly introduce the necessary background topics required to understand the experiments and results. Youtube-8M dataset is a benchmark dataset that expands ontology of 632 audio event classes and a collection of 2,084,320 human-labeled 10-second sound clips drawn from YouTube videos. This dataset have been used in a number of audio tasks like speaker identification, emotion recognition. Several famous models like VGGish, Yamnet, and TRILL are trained by this benchmark dataset. VGGish is audio embedding generated by training a modified VGGNet model, and Yamnet employs Mobilenet_v1 depthwise-separable convolution architecture to predicts 521 audio event classes. TRILL is trained to find a good non-semantic representation of speech and exceeds state-of-the-art performance on a number of transfer learning tasks. Another new open-source pre-training dataset, called the Colossal Clean Crawled Corpus (C4) is used in many NLP tasks. The text-to-text transfer transformer, named T5 model, also achieves state-of-the-art results on many NLP benchmarks. Therefore, in our experiments, we reuse these pre-trained models in single modality, respectively.

To evaluate the results of multimodal fusion through transfer learning, we choose two emotion datasets. Emotional Dyadic Motion Capture (IEMOCAP) [1] is a multimodal and multispeaker database, containing approximately 12 h of audiovisual data, including video, speech, motion capture of face, text transcriptions. Another dataset is SAVEE [7] database recorded from four native English male speakers, supporting 7 emotion categories: anger, disgust, fear, happiness, neutral, sadness, and surprise.

In the experiment of transfer learning, we select both the final output and intermediate representations of the given pre-trained models. As for speech audio, in the TRILL model, we use the final 512-dimensional embedding layer and the pre-ReLU output of the first 19-depth convolutional layer. For Vggish,

we use the final layer and the first fully connected layer. For YAMNet, we use the final pre-logit layer and the 5 depth-separable convolutional layer outputs. We use the emotion dataset of IEMOCAP and SAVEE for fine-tune training. As for text processing, we use the Text-to-Text transformer model with a maximum sequence length of 512 and a batch size of 128 sequences. During fine-tuning, we continue using batches with 128 length-512 sequences, and a constant learning rate of 0.001. The number of embedding dimensionality is set to 512.

We used different pre-trained models like VGGish, YAMNet, TRILL, and T5 to fine-tune for emotion recognition in a single modality. The obtained emotion representation of speech audio and text embedding are concatenated to pass through a transformer network. Multimodal fusion of emotional information through a transformer network generates a unified representation for emotion recognition. Table 1 shows the comparison of several single modalities and multimodal fusion results for emotion recognition. We can see that fine-tuning the final embedding of pre-trained models gives a clear boost to emotion recognition. In addition, TRILL shows better performance than VGGish and YAMNet. Text-to-Text Transformer (T5) model shows better performance in emotion recognition from the text. The average of Text-to-Text Transformer accuracy achieves up to 75.5% in the emotion dataset. Multimodal fusion shows better results than single modality for emotion recognition. This explains that multiple modalities like speech and text convey more emotional information than a single modality. The multimodal fusion can well employ complementary information.

Table 2. Average performance of the different embeddings on four selected emotions.

Models + Dataset	Happy	Anger	Sad	Natural
TRILL (IEMOCAP_Audio)	77.2%	81.9%	72.3%	66.8%
TRILL (SAVEE_Audio)	73.3%	84.3%	73.3%	66.7%
T5 (IEMOCAP_Text)	79.6%	82.7%	72.2%	64.9%
T5 (SAVEE_Text)	75.0%	81.7%	71.7%	66.7%
Multimodal Fusion (IEMOCAP)	83.3%	85.2%	77.5%	71.8%
Multimodal Fusion (SAVEE)	85.0%	86.7%	78.3%	71.7%
Mean	78.9%	83.8%	74.2%	68.1%

As speech audio samples of IEMOCAP datasets in some emotion categories are rather small, we only used four emotion categories (e.g., happy, sad, anger, and natural) for analysis of differences in both two emotion datasets. Table 2 shows the comparison of single and multiple modalities on different emotion categories. We can see that the emotion category happy and anger achieves the best recognition results than that of sad and neutral. The mean accuracy of anger emotion recognition is up to 83.8%, while the mean accuracy of neutral emotion recognition is only 68.1%. Through studies of false-positive results, we find that in the above-mentioned emotion datasets, it is not obvious and easy for a human

to recognize samples that convey neutral or sad emotional meaning. Table 2 also shows the same conclusion that the performance of multimodal fusion through a transformer network exceeds single modality in Table 1. We also investigated the robustness of multimodal features' representation. The RAVDESS dataset is a multimodal validated English dataset that contains speech, song, audio, and video files that represent 8 emotions. As the speech audio and text are rather short, we used them to test the robustness of emotion representation through multimodal fusion. The results of the selected four emotion categories are shown in Table 3. The robustness is evaluated by mean squared error (MSE) criteria. The MSE of prediction on RAVDESS dataset is calculated and the values are limited to a small range, and we think transfer learning achieves good results on a different dataset, the proposed method on multimodal emotion feature fusion indeed contains rich emotional information and can be used in other zero-shot or one-shot emotion learning problems.

Table 3. Evaluation of robustness of our multimodal emotion recognition method on different datasets.

Dataset	Happy	Anger	Sad	Natural
IEMOCAP & SAVEE	84.2%	86.0%	77.9%	71.8%
RAVDESS	82.2%	87.8%	77.8%	76.1%

5 Conclusion

In this paper, we proposed a new method for multimodal emotion recognition. As deep neural network models trained by huge datasets exhaust a lot of unaffordable resources, we made full use of transfer learning. Large excellent pre-trained models like TRILL and Text-to-Text Transformer from a single modality are used and fine-tuned on the emotion dataset. The extracted emotional information from speech audio and text embedding are concatenated and passed to a transformer network to fuse. The experiments compared the performance of single modality and multiple modalities (speech and text) for emotion recognition which showed noticeable advantages of the latter over the former. Furthermore, the robustness of transfer learning and multimodal fusion is also tested by an extra dataset. We showed that our proposed method for multimodal emotion recognition achieves good results and can be used in other zero-shot or one-shot emotion learning tasks.

References

1. Busso, C., et al.: IEMOCAP: interactive emotional dyadic motion capture database. Lang. Resour. Eval. **42**(4), 335–359 (2008)

2. Cibau, N.E., Albornoz, E.M., Rufiner, H.L.: Speech emotion recognition using a deep autoencoder. Anales de la XV Reunion de Procesamiento de la Informacion y Control **16**, 934–939 (2013)
3. Deng, J.J., Leung, C.H., Milani, A., Chen, L.: Emotional states associated with music: classification, prediction of changes, and consideration in recommendation. ACM Trans. Interact. Intell. Syst. (TiiS) **5**(1), 4 (2015)
4. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
5. El Ayadi, M., Kamel, M.S., Karray, F.: Survey on speech emotion recognition: features, classification schemes, and databases. Pattern Recogn. **44**(3), 572–587 (2011)
6. Hamel, P., Davies, M.E., Yoshii, K., Goto, M.: Transfer learning in mir: sharing learned latent representations for music audio classification and similarity (2013)
7. Haq, S., Jackson, P.J., Edge, J.: Speaker-dependent audio-visual emotion recognition. In: AVSP, pp. 53–58 (2009)
8. Howard, J., Ruder, S.: Universal language model fine-tuning for text classification. arXiv preprint arXiv:1801.06146 (2018)
9. Huang, C., Gong, W., Fu, W., Feng, D.: A research of speech emotion recognition based on deep belief network and SVM. Mathematical Problems in Engineering, 2014 (2014)
10. Huang, Z., Dong, M., Mao, Q., Zhan, Y.: Speech emotion recognition using CNN. In: Proceedings of the 22nd ACM international conference on Multimedia, pp. 801–804 (2014)
11. Izard, C.E., Malatesta, C.Z.: Perspectives on emotional development i: differential emotions theory of early emotional development. In: The First Draft of this Paper was Based on an Invited Address to the Eastern Psychological Association, 1 April 1983. John Wiley & Sons (1987)
12. Kratzwald, B., Ilić, S., Kraus, M., Feuerriegel, S., Prendinger, H.: Deep learning for affective computing: text-based emotion recognition in decision support. Decis. Support Syst. **115**, 24–35 (2018)
13. Latif, S., Rana, R., Khalifa, S., Jurdak, R., Epps, J., Schuller, B.W.: Multi-task semi-supervised adversarial autoencoding for speech emotion recognition. IEEE Trans. Affect. Comput. **abs/1907.06078** (2019)
14. Latif, S., Rana, R., Younis, S., Qadir, J., Epps, J.: Transfer learning for improving speech emotion classification accuracy. arXiv preprint arXiv:1801.06353 (2018)
15. Medhat, W., Hassan, A., Korashy, H.: Sentiment analysis algorithms and applications: a survey. Ain Shams Eng. J. **5**(4), 1093–1113 (2014)
16. Mehrabian, A.: Basic dimensions for a general psychological theory implications for personality, social, environmental, and developmental studies (1980)
17. Mittal, T., Bhattacharya, U., Chandra, R., Bera, A., Manocha, D.: M3er: multiplicative multimodal emotion recognition using facial, textual, and speech cues. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, pp. 1359–1367 (2020)
18. Nwe, T.L., Foo, S.W., De Silva, L.C.: Speech emotion recognition using hidden markov models. Speech Commun. **41**(4), 603–623 (2003)
19. Ortony, A., Clore, G.L., Collins, A.: The Cognitive Structure of Emotions. Cambridge University Press, Cambridge (1990)
20. Pan, S.J., Yang, Q.: A survey on transfer learning. IEEE Trans. Knowl. Data Eng. **22**(10), 1345–1359 (2009)

21. Pohle, T., Pampalk, E., Widmer, G.: Evaluation of frequently used audio features for classification of music into perceptual categories. In: Proceedings of the Fourth International Workshop on Content-Based Multimedia Indexing, vol. 162 (2005)
22. Poria, S., Chaturvedi, I., Cambria, E., Hussain, A.: Convolutional MKL based multimodal emotion recognition and sentiment analysis. In: 2016 IEEE 16th international conference on data mining (ICDM), pp. 439–448. IEEE (2016)
23. Roberts, A., Raffel, C.: Exploring transfer learning with t5: the text-to-text transfer transformer. Accessed on, pp. 23–07 (2020)
24. Thayer, R.E.: The Biopsychology of Mood and Arousal. Oxford University Press, Oxford (1990)
25. Tzirakis, P., Trigeorgis, G., Nicolaou, M.A., Schuller, B.W., Zafeiriou, S.: End-to-end multimodal emotion recognition using deep neural networks. IEEE J. Selected Topics Signal Process. **11**(8), 1301–1309 (2017)
26. Vaswani, A., et al.: Attention is all you need. arXiv preprint arXiv:1706.03762 (2017)
27. Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., Bowman, S.R.: GLUE: a multi-task benchmark and analysis platform for natural language understanding. arXiv preprint arXiv:1804.07461 (2018)
28. Zhao, J., Mao, X., Chen, L.: Speech emotion recognition using deep 1d & 2d cnn lstm networks. Biomed. Signal Process. Control **47**, 312–323 (2019)