Towards Learning a Joint Representation from Transformer in Multimodal Emotion Recognition

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

[1] MindSense Technologies, Pok Fu Lam, Hong Kong, PRC
james@mindsense.ai
[2] The Chinese University of Hong Kong, Shenzhen, People's Republic of China
clementleung@cuhk.edu.cn

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. This paper proposes a new method to learn a joint emotion representation for multimodal emotion recognition. Emotion-based feature for speech audio is learned by an unsupervised triplet-loss objective, and a text-to-text transformer network is used to extract text embedding for latent emotional meaning. Transfer learning provides a powerful and reusable technique to help fine-tune emotion recognition models trained on mega audio and text datasets respectively. The extracted emotional information from speech audio and text embedding are processed by dedicated transformer networks. The alternating co-attention mechanism is used to construct a deep transformer network. Multimodal fusion is implemented by a deep co-attention transformer network. Experimental results show the proposed method for learning a joint emotion representation achieves good performance in multimodal emotion recognition.

Keywords: Multimodal emotion recognition · Multimodal fusion · Transformer network

1 Introduction

Deep learning like convolution neural network (CNN), recurrent neural network (RNN), and other deep models have proven extremely 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 also yields significant importance. 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 opposite. The single modality for emotion recognition is insufficient and incomplete. For example, in a conversation, people's voices, text of speech content,

M. Mahmud et al. (Eds.): BI 2021, LNAI 12960, pp. 179–188, 2021.
https://doi.org/10.1007/978-3-030-86993-9_17

facial expressions, body language, or gestures all convey emotional meaning. Thus, it is inappropriate to recognize people's emotions only 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, learning a joint representation for emotional information from multiple modalities is necessary and useful in downstream tasks like multimodal emotion recognition. In this paper, we propose a learning method to find a joint emotion representation using both speech audio and text information from a transformer netowrk with co-attention.

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. 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) [18] model trained by a common crawl (C4) dataset to extract text embeddings. Operations of transfer learning have greatly accelerated model training and that being 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 obtain speech features and text embedding.

We expect the fused feature of speech and text are more informative and synthetic. To retain more hidden emotional information for multimodal fusion, we use the transformer network architecture to process modalities, with each modality passed to a dedicated transformer. Considering that speech audio and speech content obviously have some extent of correlation, and both exert an effect on emotion recognition, the co-attention mechanism is adopted here. This operation can well reflect internal influence of each modality. A quantitative evaluation shows that learned fused features can outperform the existing methods in emotion recognition. The main contribution of this paper is summarized as follows: (1) we propose an effective method to learn a joint emotion representation; (2) we evaluate the performance of the proposed method and validate the co-attention mechanism. Section 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.

2 Literature Review

Many research works of emotion recognition have been done in a single modality setting in past decades. A survey [4] 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 [16]. 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 [9], Deep Belief Network (DBN) [8], Long Short-Term Memory (LSTM), Autoencoder [2,11] have been used to construct various deep neural networks, achieving good performance in speech emotion recognition. Sentiment analysis [14] 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 [20] is rather popular in dealing with natural language processing like General Language Understanding Evaluation (GLUE). The model of Bidirectional Encoder Representations from Transformers (BERT) [3] 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 provides valuable insight for performing semantic classification like recognizing emotion from the text. This paper inherits this text-to-text transformer architecture.

Transfer learning and domain adaptation have been extensively practiced in machine learning. In specific domains, there is often only a few dataset 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 [5] or deep models like VGGish or BERT is transferred to be used on another learning task, usually achieving an exciting performance. [12] uses DBN of transfer learning to build sparse autoencoder for speech emotion recognition. [10] introduces a sent2affect framework, a tailored form of transfer learning to recognize emotions from the text. Another Universal Language Model Fine-tuning (ULMFiT) [7] 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 [15], joint feature representation

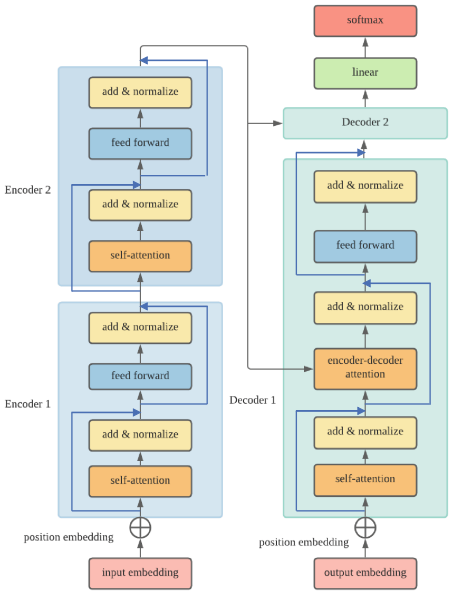

Fig. 1. An encoder-decoder transformer architecture with residual skip connections.

[17,19], or generative adversarial network (GAN) for multimodal modeling, there is less work on the aspect of transfer learning on multimodal. The co-attention mechanism [13] has showed good performance on visual question answering. This inspires us to use transfer learning, transformer network, and multimodal fusion to learn a joint emotional representation of modalities.

3 Methodology

This paper concentrates on dual modalities: speech and text. The aim is to learn a joint representation of emotional information from these dual modalities. This section will first introduce the emotional features extracted from speech audio, and text, respectively. Then the proposed method for joint emotion representation from transformer will be interpreted.

3.1 Feature Extraction from Modalities

Speech recognition has been extensively researched and achieved great success in the last decade, and many speech emotion recognition tasks adopt acoustic features like Mel-frequency Cepstral Coefficient (MFCC), deep auto-encoders in Deep Neural Network (DNN), or end-to-end with attention-based strategy that usually are used in speech recognition. 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 suitable 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 $x_i \in^{F \times T}$, F and T denotes for the dimensionality of spectrogram. 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 $|i - k| > \tau$ 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.

Considering that 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 fed 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. 1. 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.

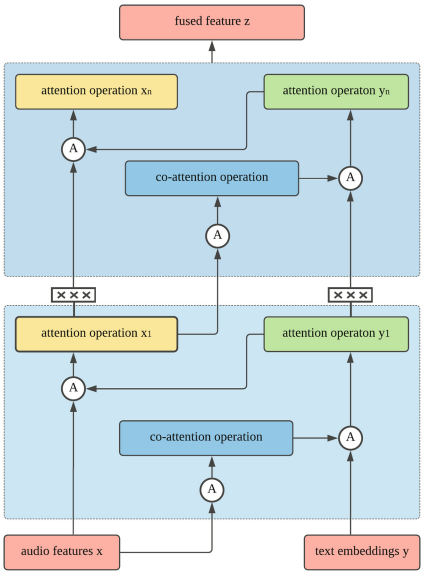

Fig. 2. Multimodal fusion by learning a joint emotion representation through a deep network with alternating co-attention operations.

3.2 Joint Emotion Representation from Transformer Network

After we obtain emotional representation of speech and text, it is natural to concatenate speech audio features and text embeddings, and feed them to a network. Here, we make some changes. Each modality is processed by a dedicated transformer. That's to say, speech audio features are passed to a transformer, and text embeddings are sent to another. Given a piece of speech, people recognize emotion both from speech audio and speech content. Speech audio and speech content obviously have some extent of correlation, and both exert an effect on emotion recognition. Thus, the co-attention mechanism is adopted. Taking the aforementioned Speech audio features X and text embedding Y as input, we pass the input features to a transformer network with a deep co-attention model. The network architecture of this deep co-attention is illustrated in Fig. 2. We sequentially alternate between speech and text attention. We first attend to the speech based on the text embedding vector. Then we attend to the text based on the attended speech feature. The co-attention operations are represented by

$$\begin{aligned} \hat{x} &= \Lambda(X; g_x) \\ \hat{y} &= \Lambda(Y; g_y) \end{aligned} \tag{2}$$

where attention guidance g_x derived from speech, and attention guidance g_y derived from text. The detailed computation is given as follows:

$$\begin{aligned} H &= tanh(W_x X + (W_{g_x} g_x)V^T) \\ \hat{a} &= softmax(w_{hx}^T H) \\ \hat{x} &= \sum a_i^x x_i \end{aligned} \tag{3}$$

where V is a vector with all elements equalling one. W_x and W_{g_x} denotes for $k \times d$ matrix parameter, and w_{hx} refers to k dimensional vector parameter. a^x is the attention weight of speech feature X. The computations of $\hat{y}$ follows the same process in Eq. 3. At the first step of alternating co-attention, g_x is 0. At the second step, g_y is intermediate attended text embedding from the first step. At last, we use the speech feature $\hat{x}$ as the guidance to attend the text again. We use a linear function to fuse attended features $\hat{x}$ and $\hat{y}$. The fused feature z is represented by

$$z = LayerNorm(W_x^T \hat{x} + W_y^T \hat{y}) \tag{4}$$

Finally, the binary cross-entropy is used as loss function to train a classifier.

4 Experiments and Results

We briefly introduce the necessary background topics required to understand the experiments and results before presenting the results from our empirical study. 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-s 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 multi-speaker database, containing approximately 12 h of audiovisual data, including video, speech, motion capture of face, text transcriptions. Another dataset is SAVEE [6] 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

Table 1. Comparison of multimodal fusion by learning a joint emotion representation 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	81.7%	75.9%	78.8%

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 set 4-layer deep co-attention network for learning a joint emotion representation.

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 by learning a joint emotion representation 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. As the length of each speech in dataset of SAVEE is short, corresponding text numbers are rather small. Thus, the performance is lower than that of IEMOCAP dataset.

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,

Table 2. Average performance of the different emotion representations on four selected emotion categories.

Models + Dataset	Happy	Anger	Sad	Natural	Mean
TRILL (IEMOCAP_Audio)	77.2%	81.9%	72.3%	66.8%	74.6%
TRILL (SAVEE_Audio)	73.3%	84.3%	73.3%	66.7%	74.4%
T5 (IEMOCAP_Text)	79.6%	82.7%	72.2%	64.9%	74.9%
T5 (SAVEE_Text)	75.0%	81.7%	71.7%	66.7%	73.8%
Multimodal fusion (IEMOCAP)	83.1%	84.6%	76.3%	71.1%	78.9%
Multimodal fusion (SAVEE)	84.7%	85.2%	74.5%	70.3%	78.7%

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. 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.

5 Conclusion

This paper proposed a new method for learning a joint emotion representation for multimodal emotion recognition. Considering that deep neural network models trained by huge datasets exhaust a lot of unaffordable resources, large excellent pre-trained models like TRILL and Text-to-Text Transformer from a single modality are used and fine-tuned on the used emotion datasets. The extracted emotional information from speech audio and text embedding are processed by dedicated transformer networks. The alternating co-attention mechanism is constructed in a deep transformer network. The fused features of each modality for emotional information are used in a classifier. 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. We showed that our proposed method for learning a joint emotion representation 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. 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)
4. 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)
5. Hamel, P., Davies, M.E., Yoshii, K., Goto, M.: Transfer learning in MIR: sharing learned latent representations for music audio classification and similarity (2013)
6. Haq, S., Jackson, P.J., Edge, J.: Speaker-dependent audio-visual emotion recognition. In: AVSP, pp. 53–58 (2009)
7. Howard, J., Ruder, S.: Universal language model fine-tuning for text classification. arXiv preprint arXiv:1801.06146 (2018)
8. Huang, C., Gong, W., Fu, W., Feng, D.: A research of speech emotion recognition based on deep belief network and SVM. Math. Probl. Eng. **2014** (2014)
9. 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)
10. 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)
11. 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. (2020)
12. Latif, S., Rana, R., Younis, S., Qadir, J., Epps, J.: Transfer learning for improving speech emotion classification accuracy. arXiv preprint arXiv:1801.06353 (2018)
13. Lu, J., Yang, J., Batra, D., Parikh, D.: Hierarchical question-image co-attention for visual question answering. Adv. Neural Inf. Process. Syst. **29**, 289–297 (2016)
14. Medhat, W., Hassan, A., Korashy, H.: Sentiment analysis algorithms and applications: a survey. Ain Shams Eng. J. **5**(4), 1093–1113 (2014)
15. 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)
16. Nwe, T.L., Foo, S.W., De Silva, L.C.: Speech emotion recognition using hidden Markov models. Speech Commun. **41**(4), 603–623 (2003)
17. 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)
18. Roberts, A., Raffel, C.: Exploring transfer learning with T5: the text-to-text transfer transformer. Accessed 23 July 2020
19. Tzirakis, P., Trigeorgis, G., Nicolaou, M.A., Schuller, B.W., Zafeiriou, S.: End-to-end multimodal emotion recognition using deep neural networks. IEEE J. Sel. Top. Signal Process. **11**(8), 1301–1309 (2017)
20. Vaswani, A., et al.: Attention is all you need. arXiv preprint arXiv:1706.03762 (2017)