Self-supervised Cross-modal Pretraining for Speech Emotion Recognition and Sentiment Analysis

Iek-Heng Chu[1]*, Ziyi Chen[1]*, Xinlu Yu[1], Mei Han[1], Jing Xiao[2] and Peng Chang[1]

[1]PAII Inc., Palo Alto, USA
[2]Ping An Technology, Shenzhen, China
{zhuyixing276,chenziyi253,yuxinlu698,hanmei613,changpeng805}@paii-labs.com
xiaojing661@pingan.com.cn

Abstract

Multimodal speech emotion recognition (SER) and sentiment analysis (SA) are important techniques for human-computer interaction. Most existing multimodal approaches utilize either *shallow* cross-modal fusion of pretrained features, or *deep* cross-modal fusion with raw features. Recently, attempts have been made to fuse pretrained feature representations in a deep fusion manner during fine-tuning stage. However, those approaches have not led to improved results, partially due to their relatively simple fusion mechanisms and lack of proper cross-modal pretraining. In this work, leveraging single-modal pretrained models (RoBERTa and HuBERT), we propose a novel deeply-fused audio-text bi-modal transformer with carefully designed cross-modal fusion mechanism and a stage-wise cross-modal pretraining scheme to fully facilitate the cross-modal learning. Our experiment results show that the proposed method achieves state-of-the-art results on the public IEMOCAP emotion and CMU-MOSEI sentiment datasets, exceeding the previous benchmarks by a large margin.

1 Introduction

Speech emotion recognition and sentiment analysis are tasks of analyzing people's attitude and opinions from their speeches. Emotion can be categorized into different classes. The most well-known emotion categorization is given by Erman who proposes 6 basic emotion classes (Gu et al., 2019): *fear*, *anger*, *joy*, *sadness*, *disgust*, and *surprise*. Sentiment reflects human attitude towards an event or object, and it is often labeled as positive, neutral or negative. Among the emotion classes, *joy* can be considered as a positive sentiment whereas *anger, sadness, fear* and *disgust* are negative sentiments.

Emotion understanding helps people communicate with each other more effectively. For human-computer interaction (HCI), speech emotion recognition (SER) plays a vital role in assisting computers to understand people's opinions. With the high-demand applications in HCI, such as voice assistants and callbots, it is essential to recognize the users' emotions and respond accordingly. With the rapid progress in deep learning (DL), many DL methods have been applied to SER (Ng et al., 2015; Sun et al., 2021b; Chen and Rudnicky, 2021) and sentiment analysis (SA) (Zhang et al., 2018; Devlin et al., 2018). There are two key challenges in these tasks at the current stage.

The first challenge is the limited availability of annotated data. Emotion annotation is known to be difficult because the pre-established annotation schemes are based on human psychology and are not conducive to reliable emotion annotation (Öhman, 2020). Therefore, compared to other tasks such as automatic speech recognition (ASR), existing SER and SA annotated datasets are fairly small in size (Busso et al., 2008; Zadeh and Pu, 2018). Direct training using such datasets may be prone to overfitting and poor model generalization ability. One way to address the issue regarding the scarcity of labeled data is via self-supervised learning (SSL). SSL has gained great success in areas of natural language processing (NLP) and speech. It has become the standard approach to build general-purpose pretrained models by utilizing large amount of unlabeled data. The pretrained models have achieved state-of-the-art (SOTA) performance on both NLP and speech related tasks. For SER and SA tasks, fine-tuning single-modal pretrained model also leads to impressive results (Chen and Rudnicky, 2021; Pepino et al., 2021; Wang et al., 2021).

The second challenge is the learning of a multimodal feature space that can well distinguish among different emotions or sentiments, especially in the case of multimodal modeling. Finding the most effective mechanism to fuse features from different modalities remains an open problem. Pre-

* Equal contribution

Findings of the Association for Computational Linguistics: EMNLP 2022, pages 5105–5114
December 7-11, 2022 ©2022 Association for Computational Linguistics

vious studies have explored different approaches to fuse multimodal information for SER and SA tasks. However, most existing approaches focus mainly on shallow fusion mechanisms, which are unlikely to capture the deep latent relationships among different modalities.

Motivated by the aforementioned challenges, we here propose a novel speech-text bi-modal transformer built on top of single-modal pretrained models RoBERTa and HuBERT. The proposed model has a carefully designed deep cross-modal fusion mechanism, which is trained with a novel cross-modal SSL based pretraining. Our experiments on SER task (Busso et al., 2008) and SA task (Zadeh et al., 2018) showcase that our model architecture leads to effective cross-modal fusion and improves the performance on both tasks. We also propose a novel stage-wise training scheme from initial SSL pretraining to final fine-tuning to fully leverage the large amount of unlabeled data. Our pretrained cross-modal transformer achieves the SOTA performance on both the IEMOCAP emotion dataset and CMU-MOSEI sentiment dataset. It is worth noting that our proposed training scheme is not limited to SER and SA tasks.

The main contributions of our paper include:

- We propose an audio-text pretrained cross-modal transformer model built on top of RoBERTa and HuBERT single-modal pretrained models.
- We propose a novel stage-wise training scheme for our cross-modal transformer model that includes initial pretraining, task adaptive pretraining, and fine-tuning on downstream tasks. To our best knowledge, we are the first to introduce task adaptive pretraining step for cross-modal representation learning.
- On top of the stage-wise training scheme, we also perform detailed investigation of the impact of different factors on the model performance, including orthogonality regularization and layer pooling etc. Our best pre-trained cross-modal model achieves SOTA performance on both CMU-MOSEI sentiment and IEMOCAP emotion datasets.

2 Related Work

2.1 Single-modal pretrained models

Transformer-based pretrained models have shown great success in various downstream tasks including SER and SA. In recent years, audio pretrained models such as Wav2Vec 2.0 (Baevski et al., 2020) and HuBERT (Hsu et al., 2021) are widely explored in the SER task (Chen and Rudnicky, 2021; Wang et al., 2021; Morais et al., 2022; Pepino et al., 2021). These trained models can achieve results competitive to SOTA results on public datasets upon various training schemes. For example, Chen et al. (Chen and Rudnicky, 2021) illustrate that task adaptive pretraining on Wav2Vec 2.0 model is beneficial in further improving model performance. Morais et al. (Morais et al., 2022) show that layer pooling and weight averaging over best model checkpoints also help boost the performance.

Sentiment analysis has been an important research topic in NLP area as it is closely related to text semantic information (Kearney and Liu, 2014; Mehta and Pandya, 2020). Recently, text-based pretrained models such as BERT (Devlin et al., 2018) and RoBERTa (Liu et al., 2019) are proven to achieve significant improvement on public sentiment datasets as compared to those without pretraining. Moreover, Gururangan et al.(Gururangan et al., 2020) demonstrate that adaptive pretraining on task-domain datasets can effectively mitigate domain mismatch issue that often occurs in NLP tasks.

2.2 Multimodal Fusion and pretraining

For multimodal tasks such as SER and SA, most previous studies focus on building models from raw or low-level input features in a supervised (Tsai et al., 2019; Sun et al., 2021a; Yoon et al., 2019; Krishna and Patil, 2020) or self-supervised manner (Li et al., 2021; Yang et al., 2022).

Regarding fusion among single-modal features, it remains a popular research topic. There are two major fusion schemes at the current stage: i) Shallow modality fusion, where high-level single-modality features concatenate and serve as the multimodal representation prior to passing to the output head for prediction (Siriwardhana et al., 2020; Makiuchi et al., 2021). ii) Deep modality fusion where different raw or low-level single-modality features are fused via methods such as cross attention mechanism (Tsai et al., 2019).

Most existing approaches focus mainly on shallow fusion strategies for SER and SA tasks. Siriwardhana et al. (Siriwardhana et al., 2020) explore different types of shallow fusion of audio and text high-level features from single-modal pretrained models. They find that simple concatenation of the two modality features yields better performance than that using the co-attention fusion. Makiuchi et al. (Makiuchi et al., 2021) propose simple score fusion wherein the final prediction score is simply the weighted average over single-modality prediction scores from text- and audio-based trained models.

Some previous studies also explore deep cross-modality fusion. Tsai et al. (Tsai et al., 2019) propose a cross-modal transformer model to perform deep fusion via the cross-modality attention mechanism. In that work, raw or low-level features from vision, audio and text are adopted. More recently, Li et al. (Li et al., 2021) propose a transformer-based cross-modal pretrained model containing a text encoding module and a text-referred audio encoding module. They adopt raw or low-level features during pretraining.

3 Method

We first briefly review HuBERT (Hsu et al., 2021) and RoBERTa (Liu et al., 2019), the two single-modal pretrained models on top of which our bi-modal pretrained model is built. Then we introduce the proposed model architecture, which is followed by the training strategy during the pretraining and fine-tuning phases.

3.1 Single-modal pretrained model

HuBERT (Hsu et al., 2021) is a transformer-based self-supervised audio model which can be used to extract speech representations. Using the offline k-means clustering to generate the labels, HuBERT is trained with the masked language modelling (MLM) task to predict the clustering assignment of the continuous masked speech. The HuBERT model takes as input the raw audio sequence, and outputs the corresponding sequence of audio representations.

RoBERTa (Liu et al., 2019) is an extended version of the transformer-based BERT model (Devlin et al., 2018) with an optimized pretraining approach. Unlike BERT, RoBERTa is pretrained with the MLM task but the next sentence prediction task is excluded. The pretrained model takes as input the word sequence that are tokenized using GPT-2 tokenizer (Radford et al., 2019), and it outputs the corresponding sequence of word representations.

3.2 Bi-modal pretrained model

An overview of the bi-modal pretrained model architecture is shown in Figure 1. The model takes as inputs the text sequence of word-piece tokens and the raw audio signals. The text and audio inputs simultaneously pass through the pretrained RoBERTa and HuBERT models respectively. The outputs of the last encoder layers of the two models are noted as $e_w \in \mathbb{R}^{T_w \times d_w}$ and $e_a \in \mathbb{R}^{T_a \times d_a}$, and used as the text and audio embeddings, respectively.

3.2.1 Cross-modal encoding module

Inspired by the cross-modal attention mechanism proposed by Tsai et al. (Tsai et al., 2019), we here introduce a cross-modal encoding module that has a symmetric architecture consisting of a text referred cross-modal transformer and an audio referred cross-modal transformer. Both cross-modal transformers are extended from the original transformer into a bi-modal scenario.

Taking the cross-modal audio transformer (right hand side of Figure 1) as an example, we first apply multi-head self-attention on the audio sequence for each block of the cross-modal audio encoder.

$$\acute{h}_a^{[l+1]} = Attn(Q = h_a^{[l]}, K = h_a^{[l]}, V = h_a^{[l]}), \tag{1}$$

where $h_a^{[l]}$ is the output of lth cross-modal audio encoder block and $h_a^{[0]} = e_a$. Q, K and V represent query, key and value in multi-head attention. Similar to the previous work on transformer, we add a residual connection and layer normalization (LN) (Ba et al., 2016) to the self-attention output.

$$\dot{h}_a^{[l+1]} = LN(h_a^{[l]} + \acute{h}_a^{[l+1]}) \tag{2}$$

After layer normalization, cross-modal attention is employed to learn the interactions between the audio and text modalities. We pass $\dot{h}_a^{[l+1]}$ as query and RoBERTa text representation e_w as key and value to get the representation of $(l+1)$th block in the following way:

$$\ddot{h}_a^{[l+1]} = Attn(Q = \dot{h}_a^{[l+1]}, K = e_w, V = e_w) \tag{3}$$

$$\tilde{h}_a^{[l+1]} = LN(\ddot{h}_a^{[l+1]} + \dot{h}_a^{[l+1]}) \tag{4}$$

$$h_a^{[l+1]} = LN(FFN(\tilde{h}_a^{[l+1]}) + \tilde{h}_a^{[l+1]}) \tag{5}$$

Eventually, we obtain the final cross-modal audio representation $h_a \in \mathbb{R}^{T_a \times d_a}$ from the last block of

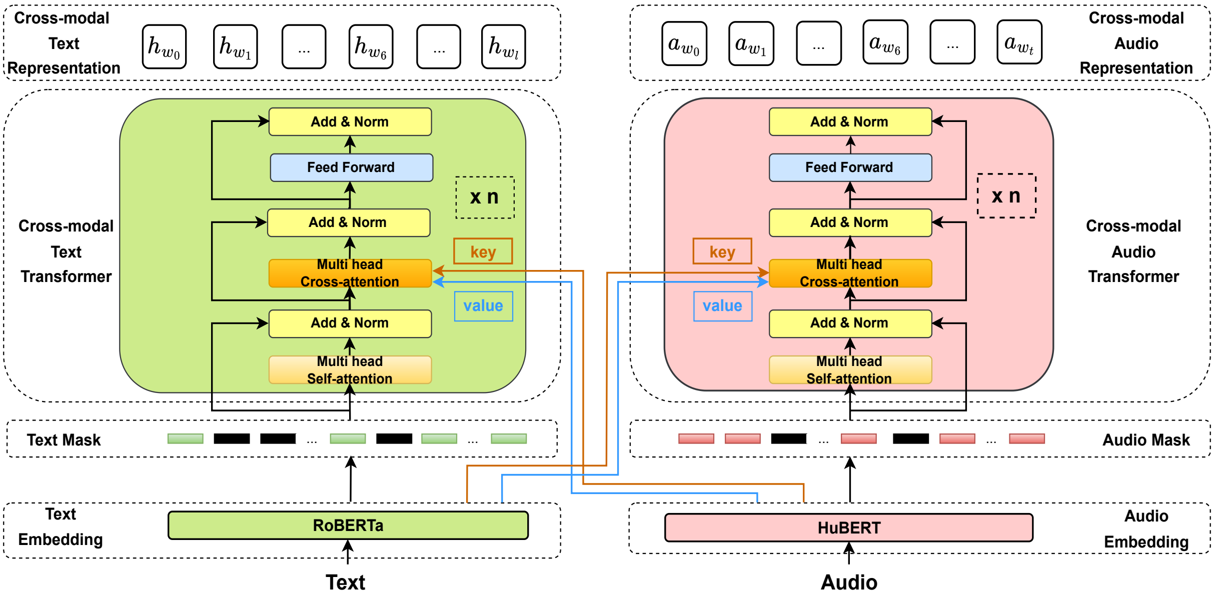

Figure 1: The architecture of the proposed cross-modal pretrained transformer model.

the cross-modal encoder. Due to the symmetric architecture of the cross-modal pretrained model, the final text representation $h_w \in \mathbb{R}^{T_w \times d_w}$ is computed in a mirroring manner.

3.2.2 Pretraining task

We pretrain our cross-modal model with the MLM task to learn the audio-text cross-modal representations. During pretraining, we simultaneously apply masking to both audio and text embeddings that serve as the inputs to the cross-modal text and audio transformers. An output linear layer is added to each transformer so as to make MLM predictions. The training details for each transformer are given below.

Cross-modal text MLM For the masking of the text embedding sequence, we follow the setup as RoBERTa (Liu et al., 2019) where we dynamically mask out each token embedding with a probability of 15%. Masked tokens are replaced with the <masked> special token, a random token, and unchanged token with a probability of 80%, 10%, and 10%, respectively. Note that when the masked sequence is passed to the text transformer, we use the unmasked key and value pair from the audio embedding for the cross-attention. This allows the model to predict the masked tokens using the information from the other modality. The text representation sequence from the transformer are then passed to the linear head for the prediction of the masked tokens. The corresponding loss L_w is computed with cross entropy loss.

Cross-modal audio MLM For the masking of the audio representation, we first perform similar setup used in wav2vec 2.0 (Baevski et al., 2020) and HuBERT (Hsu et al., 2021), where we randomly sample two audio frames with probability $p = 0.075$ as the starting frame of the mask, and the mask span is set to 10. Masked audio frames are subsequently replaced with a learnable vector. This setup forces the model to learn the masked audio frames based on global audio information rather than the local information from the neighboring audio frames.

To apply MLM task with the continuous speech representation, we follow the same prediction steps used for HuBERT. Specifically, we utilize the set of cluster codewords from HuBERT model as the labels for the frame-level audio representation. We predict the codeword for each masked audio frame based on cosine similarity betwteen the predicted and codeword embeddings. The corresponding loss L_a is computed as the cross entropy loss. The goal of our pretraining objective is to minimize the total loss $L_{\text{tot}} = L_a + L_w$.

3.3 Task adaptive pretraining

The task of adaptive pretraining (TAPT) adds an additional phase of pretraining on the task-specific unlabeled data. TAPT reduces the mismatch between pretraining domain and the task-specific domain so that the representation of the pretrained model better reflects the task distribution (Chen and Rudnicky, 2021; Gururangan et al., 2020). The previous study shows that TAPT can effectively improve the performance of single-modal pretrained model on the SER task (Chen and Rudnicky, 2021). In this paper, we use TAPT as the second phase of the cross-modal pretraining.

3.4 Fine-tuning

We fine-tune the pretrained bi-modal model for the downstream SER task and SA task. As both tasks require sequence-level prediction, we add a pooling layer on top of the cross-modal encoding module, which is followed by a linear head for predictions. For the cross-modal text representation, we use the first token embedding, i.e. *CLS* ($h_{w_0} \in \mathbb{R}^{d_w}$) as the sentence representation. For the cross-modal audio representation, we simply average over all audio frame embeddings to yield the utterance-level audio representation, denoted as $\bar{h}_a \in \mathbb{R}^{d_a}$. Then we fuse these two embeddings to form a final bi-modal representation $h_{\text{fuse}} \in \mathbb{R}^{d_a+d_w}$

$$h_{\text{fuse}} = \bar{h}_a \oplus h_{w_0}, \tag{6}$$

where $\oplus$ denotes vector concatenation. Finally, we pass it to the linear head for predictions. The task-specific loss function is denoted as L_{task}. It is the cross-entropy loss function for the SER task and mean squared error loss function for the SA task.

To encourage the text and audio transformer modules to learn from different perspectives, we introduce an orthogonal regularization term L_{ortho} (Li et al., 2021) to the loss function that is defined below,

$$L_{\text{ortho}} = \frac{|\bar{h}_a^{\text{T}} \cdot \bar{h}_w|}{||\bar{h}_a|| \cdot ||\bar{h}_w||}. \tag{7}$$

The total loss function during fine-tuning stage (L_{FT}) reads

$$L_{\text{FT}} = L_{\text{task}} + \alpha \cdot L_{\text{ortho}}, \tag{8}$$

where α is a hyper-parameter to adjust the effect of the orthogonality regularization.

4 Experiments

4.1 Dataset and evaluation metrics

We adopt the 960 hours of LibriSpeech corpus (Panayotov et al., 2015) for our bi-modal self-supervised pretraining. This dataset provides audio and the corresponding transcripts of English audio books without any emotion- or sentiment-related annotation. To evaluate our proposed algorithms, we use two public multimodal datasets for SER and SA tasks, IEMOCAP (Busso et al., 2008) and CMU-MOSEI (Zadeh and Pu, 2018). Both datasets have been widely adopted for comparison of model performance in these tasks. In this work, we only utilize the audio and text transcriptions in our experiments. For CMU-MOSEI, we extract all sample annotations using the CMU-MultimodalSDK (Zadeh et al., 2018).

4.1.1 IEMOCAP

IEMOCAP (Busso et al., 2008) is a widely-used dataset for evaluating SER models. This 12-hour dataset was recorded by ten actors and it is split into five sessions, each with a male and a female speakers. Each recording is annotated with one of the 9 emotion classes. To have a direct comparison with the previous works, we follow the same setting and consider only four of the emotion classes: angry, happy, neutral, and sad, wherein we merge the class “excited" into the class “happy". We perform a leave-one-session-out 5-fold cross validation on the dataset. We evaluate our model performance using three metrics: i) binary accuracy of each emotion class, ii) unweighted accuracy (UA) that is the average of the recall of each emotion class, and iii) weighted accuracy (WA) that is the overall accuracy of the 4-class classification model.

4.1.2 CMU-MOSEI

CMU-MOSEI (Zadeh and Pu, 2018) is an emotion and sentiment analysis dataset that contains 23,454 movie review video clips extracted from YouTube. Each sample is labeled by human annotators with a score varying from -3 (strongly negative) to +3 (strongly positive). We follow the same evaluation protocol as MulT (Tsai et al., 2019) and we evaluate the following five metrics: i) binary accuracy (Acc_2) of positive/negative sentiment classification with score in [-3, 0) are considered negative sentiment whereas score in (0, 3] as positive sentiment; ii) F1 score, iii) 7-class accuracy (Acc_7) for the classification of integer sentiment score $\in$ [-3, 3], and iv) mean absolute error (MAE) of the score.

4.2 Training configuration

We implement the proposed cross-modal model in Figure 1 within PyTorch framework (Paszke et al., 2019). We obtain the checkpoints of HuBERT[1] and RoBERTa[2] pretrained models via Huggingface

[1] https://huggingface.co/facebook/hubert-large-ll60k

[2] https://huggingface.co/roberta-large

Methods	Angry ↑	Happy ↑	Neutral ↑	Sad ↑	WA ↑	UA ↑
MulT (Tsai et al., 2019)	0.739	0.848	0.625	0.777	-	-
JBLS (Siriwardhana et al., 2020)	**0.920**	0.870	0.809	0.908	-	0.734
CTAL (Li et al., 2021)	-	-	-	-	0.740	0.746
HuBERT	0.908	0.825	0.785	0.885	0.703	0.711
RoBERTa	0.902	0.850	0.782	0.869	0.702	0.709
Shallow-Fusion	0.901	0.849	0.789	0.895	0.717	0.728
CMT BASE	0.907	0.869	0.815	0.912	**0.751**	**0.763**
CMT LARGE	0.898	**0.872**	**0.817**	**0.913**	0.750	0.761

Table 1: Main experimental results on IEMOCAP emotion dataset, where emotion-wise (angry/happy/neutral/sad) binary accuracy, weighted accuracy (WA) and unweighted accuracy (UA) are presented.

Methods	Acc_7 ↑	Acc_2 ↑	F1-score ↑	MAE ↓
MulT (Tsai et al., 2019)	0.507	0.816	0.816	0.591
JBLS (Siriwardhana et al., 2020)	0.521	0.878	-	0.518
CTAL (Li et al., 2021)	-	0.808	0.810	0.603
HuBERT	0.486	0.796	0.799	0.634
RoBERTa	0.521	0.876	0.877	0.523
Shallow-Fusion	0.538	0.861	0.860	0.518
CMT BASE	**0.546**	0.880	0.878	0.501
CMT LARGE	0.545	**0.885**	**0.885**	**0.500**

Table 2: Main experimental results on CMU-MOSEI sentiment dataset, where 7-class accuracy (Acc_7), 2-class accuracy (Acc_2), F1 score, and mean absolute error (MAE) are presented.

interface. Both models have 24 transformer layers with an output embedding dimension of 1024. In our proposed model, we consider two different configurations: CMT BASE and CMT LARGE. CMT BASE has two cross-modal transformer layers and 4 attention heads for each modality. CMT LARGE has four cross-modal transformer layers and 8 attention heads. The numbers of model parameters for CMT BASE and CMT LARGE are 64M and 128M. The total number of parameters for the entire model are 703M and 767M, respectively. Both CMT BASE and CMT LARGE models are firstly pretrained, then adaptively pretrained with downstream-task specific unlabeled data, and finally fine-tuned with orthogonality regularization.

4.2.1 Pretraining

We pretrain our model using 960h Librispeech corpus (Panayotov et al., 2015). We take AdamW (Loshchilov and Hutter, 2017) as the optimizer with an initial learning rate of 5e-5 and linear-decayed learning rate schedule. We use an effective batch size of 256 for 100,000 updates.

4.2.2 Fine-tuning

We take Adam (Kingma and Ba, 2014) as the optimizer with an initial learning rate of 1e-5 during fine-tuning. We use an effective batch size of 16 with the number of epoch as 25 for both IEMOCAP and CMU-MOSEI datasets. We introduce an orthogonal regularization term to the loss function, as shown in Eq.(8). For the regularization ratio α, we set it to 10 since it leads to the best tuning results, as illustrated in Figure 4.

4.3 Results and discussion

4.3.1 Main results

The results of IEMOCAP and CMU-MOSEI are illustrated in Table 1. We compare the following 8 fine-tuned models: 1) Speech-only HuBERT model. 2) Text-only RoBERTa model. 3) A model that fuses RoBERTa (text) and HuBERT (audio) representations by shallow fusion. 4) Our pretrained and task-adapted cross-modal transformer (CMT) model with two transformer layers, denoted as CMT BASE. 5) Our pretrained and task-adapted model with four transformer layers, denoted as CMT LARGE. 6) A multimodal transformer model using low-level features (Tsai et al., 2019). 7) A shallow fusion model with Speech-BERT and RoBERTa as the single-modal pretrained models (Siriwardhana et al., 2020). We reproduce the results using their public repository[3]. 8) A pretrained cross-modal transformer model without any large single-modal pretrained models.

[3] https://github.com/shamanez/BERT-like-is-All-You-Need

IEMOCAP emotion First, we find that both single-modal models have similar performance with UA and WA about 0.7. Upon bi-modal shallow fusion, we achieve an absolute 1.5% improvement on both metrics, suggesting that fusion from both modalities is critical to SER task. This is also consistent with previous multimodal works. Moreover, we find that the results are significantly improved upon the addition of the pretrained cross-modal transformer layers in CMT BASE and CMT LARGE. The absolute gain in both UA and WA is over 4%. This highlights the importance of the more complex cross-modal attention for learning the latent correlations between the two modalities. Both our CMT BASE and CMT LARGE models also outperform previous bi-modal works such as JBLS (Siriwardhana et al., 2020) and CTAL (Li et al., 2021). It should be noted that the CMT BASE and CMT LARGE models have very close UA and WA, which is within 0.3%.

CMU-MOSEI sentiment First, we notice that speech-only HuBERT model always under performs compared to other models whereas text-only RoBERTa model achieves competitive results compared to shallow-fusion-based model. This is attributed to the fact that sentiment analysis is closely related to text semantics. With shallow fusion, all metrics except for Acc$_2$ and F1-score are slightly improved. These metrics are further improved with our CMT BASE model, and it achieves the best Acc$_7$ (0.546) among all multimodal studies. With more cross-modal transformer layers, CMT LARGE model performs slightly better than CMT BASE in terms of all metrics except for Acc$_7$, which only differs by about 0.1%. It should be noted that CMT LARGE model achieves the SOTA performance in Acc$_2$, F1-score and MAE. We should point out that even with the simple shallow fusion, JBLS and our shallow-fusion approach still achieve significant better performance than other multimodal methods. This suggests the large contribution of large single modal SSL models provide better representation for the SA task.

Layer index	1	2	3	4
Weight	0.175	0.438	0.277	0.109

Table 3: Learned weights of the layer pooler associated with different audio cross-modal transformer layers in the pretrained CMT-4 model.

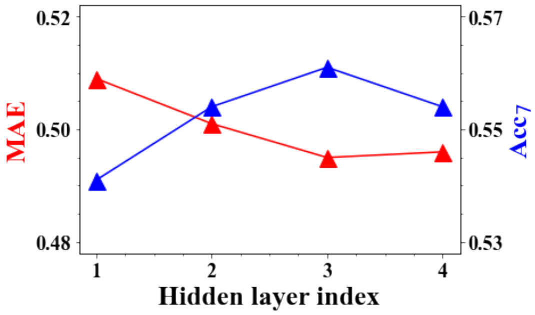

Figure 2: Impact of using different hidden states as the CMT-4 audio transformer representation on CMU-MOSEI metrics MAE (red) and Acc$_7$ (blue).

4.3.2 Effect of the cross-modal audio transformer layers

Previous study (Chen et al., 2021) shows that different transformer layers of audio SSL model may contain different types of audio information. The last layer representation of the cross-modal audio transformer may not be the optimal to SER and SA tasks. To analyze the effect of the cross-modal audio transformer layers, we first add a pooling layer on pretrained CMT with four cross-modal blocks (CMT-4 PT). During fine-tuning stage, this pooling layer performs weighted averaging over all hidden states from the cross-modal audio transformer, and it outputs the averaged one as the final cross-modal audio representation. Table 3 lists the weights of all 4 cross-modal transformer layers. Layer 2 and 3 have the highest weights, and the bottom and top layers have lower weights. This indicates the second and third layer have the most audio sentiment-related information. In addition, we fine-tune the CMT-4 PT model on CMU-MOSEI using only one of the cross-modal audio layer hidden states as the final cross-modal audio transformer representation. As shown in Figure 2, the third hidden layer achieves the best results, suggesting that the middle layers may store the most sentiment-related information. It should be noted that adding weight averaging layer does not bring any performance gain in our study. More details will be discussed in Section 4.3.4.

4.3.3 Effect of pretraining

To further analyze the effect of our pretrained CMT model, we fine-tune the pretrained and non-pretrained CMT-4 models with different proportion of CMU-MOSEI dataset. Figure 3 shows the MAE and 7 class accuracy (Acc$_7$) of CMU-MOSEI test set with 20%, 50%, 80%, and full training

Methods	Acc$_7$ ↑	Acc$_2$ ↑	F1-score ↑	MAE ↓
CMT-4 w/o PT	0.545	0.874	0.875	0.508
CMT-4 PT	0.554	0.870	0.871	0.496
CMT-4 PT + TAPT	**0.559**	0.866	0.869	0.502
CMT-4 PT + Layer pooler	0.545	0.864	0.863	0.509
CMT-4 PT + Ortho	0.554	0.879	0.878	**0.493**
CMT BASE	0.546	0.880	0.878	0.501
CMT LARGE	0.545	**0.885**	**0.885**	0.500

Table 4: The ablation analysis of our proposed CMT-4 model using CMU-MOSEI. The terms PT, TAPT, Layer pooler and Ortho refer to pretrained, task adaptive pretraining, weighted average layer of cross-modal audio transformer hidden states, and the orthogonality regularization term.

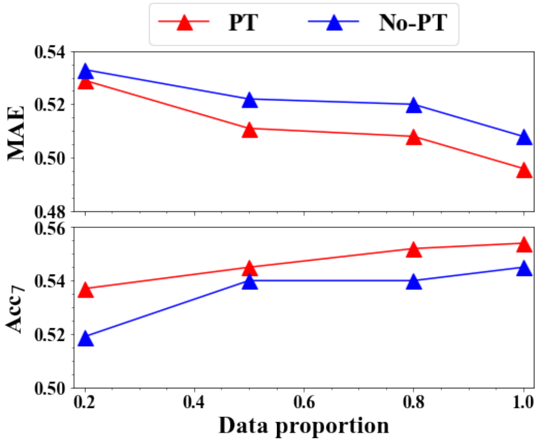

Figure 3: The performance of pretrained (PT) and non-pretrained (No-PT) CMT-4 models with different proportions of CMU-MOSEI training set.

data. The pretrained CMT-4 model consistently outperform the non-pretrained counterpart across different training sizes. Benefiting from the SSL of paired audio-text data, the pretrained CMT-4 can achieve the similar result to the non-pretrained CMT-4 with only 50% of the training data.

4.3.4 Ablation study

We conduct an ablation analysis for our proposed CMT model using the CMU-MOSEI dataset. The results are detailed in Table 4. In general, our proposed CMT BASE and CMT LARGE models perform better than the CMT model without pretraining. We also observe that task adaptive pretraining and orthogonality regularization boost the final performance of the CMT-4 pretrained model. However, adding the layer pooler does not bring any gain in the results. Our analysis in Section 4.3.2 suggests that some lower hidden layers have significantly less sentiment-related information. Since the CMT-4 model only has four cross-modal transformer layers, and the weighted average pooling still assigns some weights to the lower layers. This tends to adversely affect the final performance.

5 Conclusion

In this work, we propose a novel bi-modal model with a symmetric cross-modal attention mechanism that efficiently fuses the representations from single-modal pretrained models. We show that upon pretraining followed by task adaptative pretraining and fine-tuing with additional modality orthogonal regularization, the proposed bi-modal model can achieve SOTA performance on both IEMOCAP and CMU-MOSEI datasets.

Limitations

In this paper we mainly focus on how to improve the overall SER/SA performance through the self-supervised cross-modal pretraining scheme. We believe the underlying cross-modal attention mechanism can also be further improved, to better explore the complementary information across the modalities. The current pretraining utilizes the standard MLM pretext tasks, which can be further improved with pretext tasks better accommodating SER/SA analysis. These limitations also constitute our future work.

References

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. 2016. Layer normalization. *arXiv preprint arXiv:1607.06450*.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in Neural Information Processing Systems*, 33:12449–12460.

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. 2008. Iemocap: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42(4):335–359.

Li-Wei Chen and Alexander Rudnicky. 2021. Exploring wav2vec 2.0 fine-tuning for improved speech emotion recognition. *arXiv preprint arXiv:2110.06309*.

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, et al. 2021. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *arXiv preprint arXiv:2110.13900*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Simeng Gu, Fushun Wang, Nitesh P. Patel, James A. Bourgeois, and Jason H. Huang. 2019. A model for basic emotions using observations of behavior in drosophila. *Frontiers in Psychology*, 10.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A Smith. 2020. Don't stop pretraining: adapt language models to domains and tasks. *arXiv preprint arXiv:2004.10964*.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3451–3460.

Colm Kearney and Sha Liu. 2014. Textual sentiment in finance: A survey of methods and models. *International Review of Financial Analysis*, 33:171–185.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXivf:1412.6980*.

DN Krishna and Ankita Patil. 2020. Multimodal emotion recognition using cross-modal attention and 1d convolutional neural networks. In *Interspeech*, pages 4243–4247.

Hang Li, Yu Kang, Tianqiao Liu, Wenbiao Ding, and Zitao Liu. 2021. Ctal: Pre-training cross-modal transformer for audio-and-language representations. *arXiv preprint arXiv:2109.00181*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Mariana Rodrigues Makiuchi, Kuniaki Uto, and Koichi Shinoda. 2021. Multimodal emotion recognition with high-level speech and text features. *arXiv preprint arXiv:2111.10202*.

Pooja Mehta and Sharnil Pandya. 2020. A review on sentiment analysis methodologies, practices and applications. *International Journal of Scientific and Technology Research*, 9(2):601–609.

Edmilson Morais, Ron Hoory, Weizhong Zhu, Itai Gat, Matheus Damasceno, and Hagai Aronowitz. 2022. Speech emotion recognition using self-supervised features. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6922–6926. IEEE.

Hong-Wei Ng, Viet Dung Nguyen, Vassilios Vonikakis, and Stefan Winkler. 2015. Deep learning for emotion recognition on small datasets using transfer learning. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, ICMI '15, page 443–449, New York, NY, USA. Association for Computing Machinery.

Emily Öhman. 2020. Emotion annotation: Rethinking emotion categorization. In *DHN Post-Proceedings*, pages 134–144.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: An asr corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5206–5210.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32.

Leonardo Pepino, Pablo Ernesto Riera, and Luciana Ferrer. 2021. Emotion recognition from speech using wav2vec 2.0 embeddings. In *Interspeech*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Shamane Siriwardhana, Andrew Reis, Rivindu Weerasekera, and Suranga Nanayakkara. 2020. Jointly Fine-Tuning "BERT-Like" Self Supervised Models to Improve Multimodal Speech Emotion Recognition. In *Proc. Interspeech 2020*, pages 3755–3759.

Licai Sun, Bin Liu, Jianhua Tao, and Zheng Lian. 2021a. Multimodal cross- and self-attention network for speech emotion recognition. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4275–4279.

Licai Sun, Bin Liu, Jianhua Tao, and Zheng Lian. 2021b. Multimodal cross-and self-attention network for speech emotion recognition. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4275–4279. IEEE.

Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2019. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the conference. Association for Computational Linguistics. Meeting*, volume 2019, page 6558. NIH Public Access.

Yingzhi Wang, Abdelmoumene Boumadane, and Abdelwahab Heba. 2021. A fine-tuned wav2vec 2.0/hubert benchmark for speech emotion recognition, speaker verification and spoken language understanding. *arXiv preprint arXiv:2111.02735*.

Ziyi Yang, Yuwei Fang, Chenguang Zhu, Reid Pryzant, Dongdong Chen, Yu Shi, Yichong Xu, Yao Qian, Mei Gao, Yi-Ling Chen, et al. 2022. i-code: An integrative and composable multimodal learning framework. *arXiv preprint arXiv:2205.01818*.

Seunghyun Yoon, Seokhyun Byun, Subhadeep Dey, and Kyomin Jung. 2019. Speech emotion recognition using multi-hop attention mechanism. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2822–2826. IEEE.

Amir Zadeh, Paul Pu Liang, Soujanya Poria, Prateek Vij, Erik Cambria, and Louis-Philippe Morency. 2018. Multi-attention recurrent network for human communication comprehension. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Amir Zadeh and Paul Pu. 2018. Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Long Papers)*.

Lei Zhang, Shuai Wang, and Bing Liu. 2018. Deep learning for sentiment analysis: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 8(4):e1253.

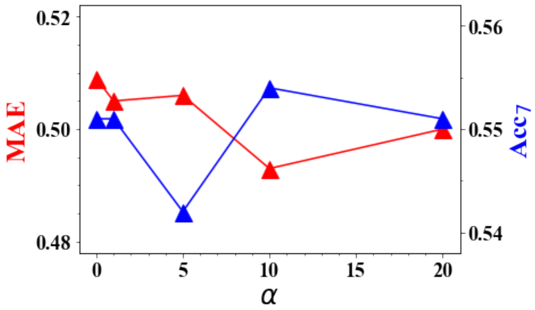

Figure 4: Tuning of orthogonality regularization on CMU-MOSEI using MAE and 7-class accuracy (Acc_7). Here, the ratio parameter α is defined in Eq.(8).

A Appendix

A.1 Effect of orthogonality regularization

We adopt the pretrained CMT-4 model to evaluate the effect of orthogonality regularization on the final results. Using CMU-MOSEI dataset, we perform fine-tuning using various values of the regularization term ratio α, as defined in Eq.(8). The results are demonstrated in Figure 4. It shows that $\alpha = 10$ achieves the best performance.