CTAL: Pre-training Cross-modal Transformer for Audio-and-Language Representations

Hang Li, Wenbiao Ding, Yu Kang, Tianqiao Liu, Zhongqin Wu, Zitao Liu*
TAL Education Group, Beijing, China
{lihang4,dingwenbiao,kangyu,liutianqiao}@tal.com
{wuzhongqin,liuzitao}@tal.com

Abstract

Existing approaches for audio-language task-specific prediction focus on building complicated late-fusion mechanisms. However, these models face challenges of overfitting with limited labels and poor generalization. In this paper, we present a **C**ross-modal **T**ransformer for **A**udio-and-**L**anguage, i.e., CTAL, which aims to learn the intra- and inter- modalities connections between audio and language through two proxy tasks from a large number of audio-and-language pairs: masked language modeling and masked cross-modal acoustic modeling. After fine-tuning our CTAL model on multiple downstream audio-and-language tasks, we observe significant improvements on different tasks, including emotion classification, sentiment analysis, and speaker verification. Furthermore, we design a fusion mechanism in the fine-tuning phase, which allows CTAL to achieve better performance. Lastly, we conduct detailed ablation studies to demonstrate that both our novel cross-modality fusion component and audio-language pre-training methods contribute to the promising results. The code and pre-trained models are available at https://github.com/tal-ai/CTAL_EMNLP2021.

1 Introduction

Speech processing (SP) requires the understanding of a set of acoustic and language content, including phonemes, tones, words and semantic meanings. Different from human, who benefit from self-learning through real-world experiences, current SP methods are more like narrow experts relying heavily on a large number of task-specific human annotations and a small change may cause the learning process to start all over again. In recent years, pre-training for single modality, such as language and audio signals, is widely explored due to its ease-of-use and competent generalization to various downstream tasks.

In the field of NLP, pre-trained models, such as BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), XLNet (Yang et al., 2019) and GPT2 (Radford et al., 2019), share the same idea of first leveraging large-scale unlabeled corpus to perform contextualized language model pre-training then fine-tuning the model to adapt to downstream tasks, such as machine reading comprehension (Lai et al., 2017), question answering (Rajpurkar et al., 2016) and natural language inference (Bowman et al., 2015), etc. Following the success of pre-trained models in NLP, BERT-like models are also applied to SP community (Schneider et al., 2019; Baevski et al., 2020a,b), where robust audio representations are learned through an audio-style self-supervised context prediction task.

Despite these influential unimodal methods, for tasks at the intersection of audio and language, such as speech emotion classification (Livingstone and Russo, 2018; Busso et al., 2008), speaker verification (Panayotov et al., 2015) and sentiment analysis (Zadeh et al., 2018), large-scale pre-training for the modality-pair of audio and language is barely explored. The previous attempt is to train task-specific experts upon the concatenation of language representations and audio representations in a late fusion manner (Ramirez et al., 2011; Glodek et al., 2011; Zadeh et al., 2017; Yoon et al., 2019, 2018; Xu et al., 2019; Li et al., 2020b, 2021), without any generic audio-and-language pre-training. These task-specific experts will suffer from the overfitting problem when trained with limited data. Meanwhile, due to the heterogeneity across language and audio modalities, late fusion of high-level representations lacks surface-level cross-modal alignment and complementation during the pre-training phase.

Therefore, we propose CTAL, a pre-training cross-modal Transformer for audio-and-language representations, and has shown its strong performance on three established audio-and-language

* The corresponding author: Zitao Liu.

Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 3966–3977
November 7–11, 2021.

tasks: emotion classification (Busso et al., 2008), sentiment analysis (Zadeh et al., 2018) and speaker verification (Panayotov et al., 2015). We propose multimodal Transformer as our backbone model, which consists of two modules, a language stream encoding module which adapts word as input element, and a text-referred audio stream encoding module which accepts both frame-level Mel-spectrograms and token-level output embeddings from the language stream encoding module as input elements. In order to learn both intra- and inter- modalities connections, we pre-train our model with two tasks: (1) masked language modeling (MLM); and (2) masked cross-modal acoustic modeling (MCAM). Different from unimodal pre-training, e.g., masked acoustic modeling in MOCKINGJAY (Liu et al., 2020), our cross-modal pre-training is able to reconstruct masked audio features from both intra- and inter-modalities information. On the basis of our pre-trained model, a regularization term based on feature orthogonality is introduced during the model fine-tuning stage, which is designed to ensure that features of different modalities provide information from different perspectives. Please notice that this orthogonal regularization mechanism is general and not limited to audio-language tasks.

The main contributions of our paper are listed as follows:

- We present CTAL, a pre-training framework for learning audio-and-language representations with Transformer, which considers both intra- and inter- modalities connections. To the best of our knowledge, we are the first to introduce the pre-training cross audio-and-language modalities.
- We propose a novel cross-modality fusion mechanism at the fine-tuning stage, which forces our pre-trained model learn composite features from different views.
- Comprehensive empirical results demonstrate that our CTAL achieves the state-of-the-art results on various downstream SP tasks, such as emotion classification, sentiment analysis, and speaker verification. We conduct detailed ablation studies and analysis to show the effectiveness of our model components and our pre-training strategies. To encourage reproducible results, we put our code publicly available at https://github.com/tal-ai/CTAL_EMNLP2021.

2 Related Work

2.1 Unimodal Pre-training

There has been a long interest around self-supervised representation learning. Previous works have explored alternative approaches to improve word embedding (Mikolov et al., 2013; Le and Mikolov, 2014; Pennington et al., 2014), which is a low-level linguistic representation. After that, pre-trained NLP models based on multi-layer Transformers, such as BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), XLNet (Yang et al., 2019) and GPT2 (Radford et al., 2019), benefit from context-sensitive representation learning on large-scale corpus, showing significant improvements in various downstream language understanding tasks. Self-supervised learning in speech processing has also shown increasing promise. Following BERT, many approaches (Jiang et al., 2019; Liu et al., 2021, 2020; Chi et al., 2021) are proposed to learn high-level acoustic representations rather than surface features such as log Mel-spectrograms or waveform, which can reveal the abundant information within audio signals.

2.2 Multimodal Pre-training

While pre-training for audio-and-language representations has rarely been studied, several attempts have been made to pre-train models for vision-and-language tasks on visual question answering (Antol et al., 2015) and visual commonsense reasoning (Zellers et al., 2019) datasets. In general, these vision-and-language pre-training methods can be divided into two categories, according to their different encoder architectures as follows: (a) prior works like ViLBERT (Lu et al., 2019) and LXMERT (Tan and Bansal, 2019), apply two unimodal networks to encode input text and images respectively and adapt cross-modal interactions in a symmetric fusion manner; (b) the other category of pre-training frameworks like VisualBert (Li et al., 2019), Unicoder-VL (Li et al., 2020a) and UNITER (Chen et al., 2020), concatenate vision and language features as a unified single-stream input and utilize a universal encoder to learn joint multimodal representations.

However, transfer above algorithms directly from vision-and-language to audio-and-language field faces challenges, including: (1) unified architecture is not suitable for audio-language modalities, since both text and audio signals are generally long sequences, and cross-modal aggregation at the very beginning phase with Transformer self-attention mechanism will lead to higher computational complexity; (2) audio signals are more informative than language texts, which contain both semantic information of text content and personal feelings. Thus, it is not suitable to apply the symmetric cross-modal fusion module proposed in prior vision-and-language pre-training researches. Based on these facts, we design our backbone model with a language stream encoding module and a text-referred audio stream encoding module, which allow necessary intra- and inter-modality connections during pre-training with less computational cost.

The closest work to our approach is from Haque et al. (2019) and our approach differs from it in two aspects. First, we use a more explicit, multi-component design for the cross-modality connections (i.e., with a text-referred audio stream encoding module and a novel cross-modality fusion component). Second, we employ different pre-training tasks which accept both text and audio frames as input to conduct contextualized masked language modeling and masked cross-modal acoustic modeling tasks.

3 Our Approach

In this section, we first present our cross-modal pre-training framework CTAL, including details of text and audio pre-processing and encoding modules for separate modalities. Then we present our pre-training tasks. In the end, we propose a novel fusion mechanism which can be utilized in the fine-tuning stage. Following conventions, we use bold upper case letters to represent matrices and bold lower case letters to represent vectors.

3.1 The CTAL Framework

We build our cross-modal Transformer by extending the original Transformer (Vaswani et al., 2017) into the multimodal paradigm. As shown in Figure 1, CTAL takes audio sequences and their corresponding text sequences as the input. Each audio sequence is represented as a sequence of frames, and each text sequence is represented as a sequence of tokens. Then we encode the input to the linguistic embedding and audio embedding, and feed them into a text encoding module and a text-referred audio encoding module respectively to generate final language representations and text referred audio representations. Following the notations used by Vaswani et al. (2017), we adapt $\mathbf{Q}$, $\mathbf{K}$ and $\mathbf{V}$ as queries, keys and values for attention mechanism, MultiHead($\mathbf{Q}$, $\mathbf{K}$, $\mathbf{V}$) as multi-head attention, FFN($\mathbf{X}$) as position-wise feed forward networks and LayerNorm($\mathbf{X}$) as layer normalization.

3.1.1 Input Embeddings

Linguistic Embedding. To encode any input text with a modest size (30K units) of subword vocabulary, we follow the text pre-processing of RoBERTa, which tokenizes each input text $w = \{w_0, w_1, ..., w_T\}$ with byte-level byte-pair encoding (BBPE) (Radford et al., 2019). Besides, we also add the special tokens <s> and </s> to represent start and end tokens. Then we sum up each token embedding and its corresponding position embedding to get the final input token embeddings $\{\mathbf{e}_{w_0}, \mathbf{e}_{w_1}, ..., \mathbf{e}_{w_T}\}$ for language modality. T is the total length of input tokens.

Audio Embedding. The input audio signal is first transformed into frames of width 50ms and step 12.5ms. Then the 80 dimension Mel-spectrograms are extracted from each frame and concatenated with their first order derivatives, making the feature dimension to 160. In this way, the raw signal is converted into sequence of frame-level acoustic surface features $\{a_0, a_1, ..., a_\mathcal{T}\}$, where $\mathcal{T}$ is the total number of frames. For simplicity, we denote this audio feature sequence as input acoustic features after this section. Then, we feed these surface features to a projection layer and add them with the position embeddings to obtain the input audio embeddings $\{\mathbf{e}_{a_0}, \mathbf{e}_{a_1}, ..., \mathbf{e}_{a_\mathcal{T}}\}$ for audio modality.

3.1.2 Text Encoding Module

As shown in Figure 1, we apply the original Transformer encoder to language stream inputs, each language stream encoding layer consists of one multi-head self-attention sublayer and one position-wise feed forward sublayer. We stack N such language encoding layer and use the output of k-th layer as the input to the $(k+1)$-th layer. We initialize $\mathbf{H}_w^0$ with $\{\mathbf{e}_{w_0}, \mathbf{e}_{w_1}, ..., \mathbf{e}_{w_T}\}$ and obtain the language representations for the k-th layer with the followings:

$$\begin{aligned}
\hat{\mathbf{H}}_w^{k+1} &= \mathrm{MultiHead}(\mathbf{Q} = \mathbf{H}_w^k, \mathbf{K} = \mathbf{H}_w^k, \mathbf{V} = \mathbf{H}_w^k) \\
\tilde{\mathbf{H}}_w^{k+1} &= \mathrm{LayerNorm}(\hat{\mathbf{H}}_w^{k+1} + \mathbf{H}_w^k) \\
\mathbf{H}_w^{k+1} &= \mathrm{LayerNorm}(\mathrm{FFN}(\tilde{\mathbf{H}}_w^{k+1}) + \tilde{\mathbf{H}}_w^{k+1})
\end{aligned}$$

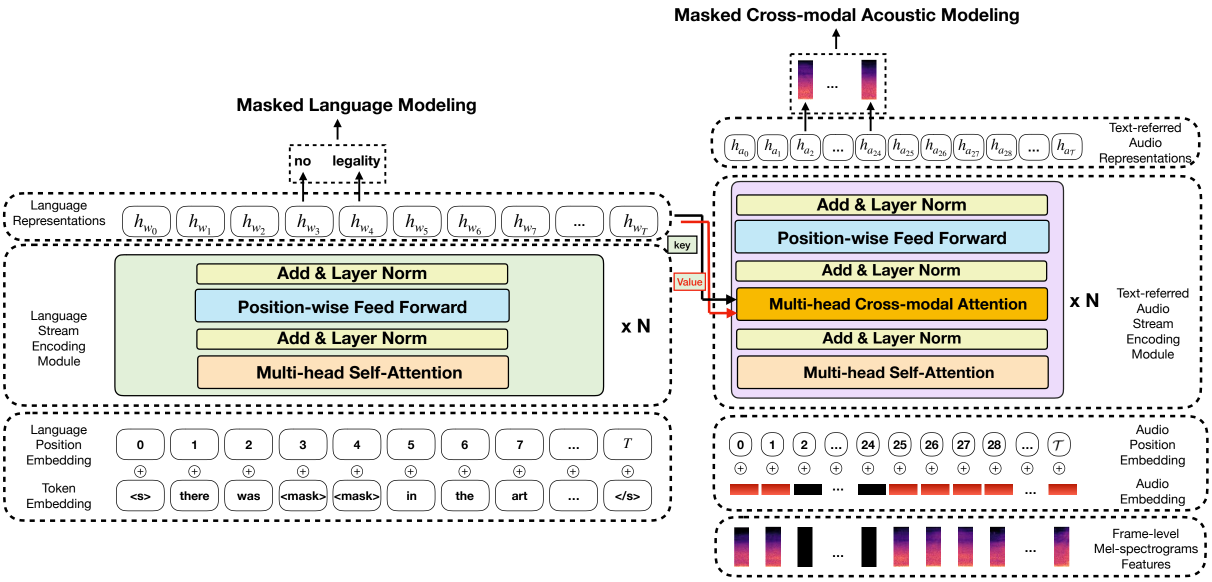


Figure 1: The proposed CTAL pre-training framework.

We get the final output $\mathbf{H}_w^N \in \mathbb{R}^{T \times d_w}$ from our language stream encoding module, where d_w denotes the hidden size of the language stream representations. The first token of every text sequence is always a special start token (<s>), and the final hidden state corresponding to this token is always used as the aggregated text sequence representation for classification tasks.

3.1.3 Text-Referred Audio Encoding Module

For text-referred audio encoding module, we first initialize hidden representations $\mathbf{H}_a^0$ with $\{\mathbf{e}_{a_0}, \mathbf{e}_{a_1}, ..., \mathbf{e}_{a_\mathcal{T}}\}$, and pass them to a stack of N text-referred audio encoding layers to acquire the final audio stream representations $\mathbf{H}_a^N$.

Our text-referred audio encoding module is different from the original Transformer decoder by modifying two kinds of multi-head attention mechanism. Firstly, in order to learn the bi-directional intra-modality representation for audio, we get rid of the future mask in the masked multi-head self-attention. Specifically for the $(l+1)$-th layer:

$$\hat{\mathbf{H}}_a^{l+1} = \mathrm{MultiHead}(\mathbf{Q} = \mathbf{H}_a^l, \mathbf{K} = \mathbf{H}_a^l, \mathbf{V} = \mathbf{H}_a^l)$$
$$\tilde{\mathbf{H}}_a^{l+1} = \mathrm{LayerNorm}(\hat{\mathbf{H}}_a^{l+1} + \mathbf{H}_a^l)$$

Secondly, we apply multi-head cross-modal attention which accepts the final language stream representations as keys and values in each layer to apply the inter-modality interactions:

$$\bar{\mathbf{H}}_a^{l+1} = \mathrm{MultiHead}(\mathbf{Q} = \tilde{\mathbf{H}}_a^{l+1}, \mathbf{K} = \mathbf{H}_w^N, \mathbf{V} = \mathbf{H}_w^N)$$
$$\ddot{\mathbf{H}}_a^{l+1} = \mathrm{LayerNorm}(\bar{\mathbf{H}}_a^{l+1} + \tilde{\mathbf{H}}_a^{l+1})$$
$$\mathbf{H}_a^{l+1} = \mathrm{LayerNorm}(\mathrm{FFN}(\ddot{\mathbf{H}}_a^{l+1}) + \ddot{\mathbf{H}}_a^{l+1})$$

Finally, we obtain the text-referred audio representation of N-th layer $\mathbf{H}_a^N \in \mathbb{R}^{\mathcal{T} \times d_a}$, where d_a denotes the hidden size of the audio stream representations.

3.2 Pre-training Tasks

3.2.1 Masked Language Modeling

For language stream, we take the MLM task for language intra-modality learning. As shown in Figure 1, the MLM task setup is almost the same as RoBERTa (Liu et al., 2019), we dynamically mask out the input tokens with a probability of 15%. Masked tokens are replaced with a special <mask> token 80% of the time, a random token 10%, and unaltered 10%. The goal of MLM is to predict these masked tokens based on the observed tokens. Here, we do not introduce acoustic information for masked token prediction, since semantic information of text can be well enough captured through language input. It is redundant to introduce cross-modal inputs here and it is demonstrate through the ablation study discussed in Section 5.1.

3.2.2 Masked Cross-modal Acoustic Modeling

For audio stream, we propose MCAM to train the text-referred audio representations. Prior research by Baevski et al. (2020b) indicates that the performance of acoustic pre-trained models on downstream tasks is improved with the increment in size of continuous masked frames during pre-training phase. However, due to the complexity of audio signals, the long-term dependencies in audio sequences is hard to be captured with acoustic features alone. To mitigate that problem, we propose MCAM to capture effective information of audio through learning both intra- and inter- modalities connections between audio and language.

To implement MCAM, we first split the audio in separate segments according to C consecutive frames per segment, where C is uniformly sampled from 20 to 50. Then we randomly select 15% of these segments and for each of them, we mask it all to zero 80% of the time, replace it with the other C randomly selected frames within the audio 10% of the time, and keep it unchanged for the remaining cases. In this manner, we prevent the model exploiting local smoothness of acoustic frames and the model is required to make inference based on global information rather than local messages. Finally, the goal is to reconstruct these masked acoustic features based on the remaining acoustic features and the language stream prompt, by minimizing the ℓ_1 loss between the original masked acoustic features and the predicted ones.

Overall, our final objective is to minimize the sum of the loss functions above.

3.3 Fine-Tuning CTAL

CTAL is designed to be a generic pre-training model for various audio-language tasks. It is relatively simple to fine-tune CTAL for various downstream tasks with just one additional output layer. To further combine information from different modalities, we propose a novel and flexible fusion mechanism at the fine-tuning stage. We denote $\mathbf{H}_w^N \in \mathbb{R}^{T \times d}$ and $\mathbf{H}_a^N \in \mathbb{R}^{T \times d}$ as the final representation from text encoding module and text-referred audio encoding module. We assume that both modules have the same hidden size d.

In SP tasks, we use the compressed hidden vectors to represent both the language and audio input sequences. Following the idea from Wang (2018), which proves that max pooling mechanism tends to make false negatives while attention pooling mechanism prefers making false positives, we come up with both attention-pooling layer and max-pooling layer to let them complement each other. After applying attention-pooling and max-pooling to audio stream final representations $\mathbf{H}_a^N$, we obtain $\mathbf{h}_a^{attn} \in \mathbb{R}^d$ and $\mathbf{h}_a^{max} \in \mathbb{R}^d$ respectively.

$$\mathbf{h}_a^{attn} = \mathrm{Softmax}(\mathbf{v}_a^{attn} \cdot \tanh(\mathbf{W}_a^{attn} \cdot \mathbf{H}_a^N)) \cdot \mathbf{H}_a^N$$
$$\mathbf{h}_a^{max} = \mathrm{MaxPool}(\mathbf{H}_a^N)$$

where $\mathbf{v}_a^{attn}$ and $\mathbf{W}_a^{attn}$ are parameters in the audio attention-pooling layer.

As discussed in Section 3.1.2, for language stream, we adapt the final hidden state of the start token $\mathbf{h}_{w0} \in \mathbb{R}^d$ as the aggregated text sequence representation $\mathbf{h}_w^{attn}$ for attention-pooling, and we conduct additional max-pooling for text stream output $\mathbf{H}_w^N$ to obtain $\mathbf{h}_w^{max}$.Then we fuse the aggregated sequence representations from two modalities as follows:

$$\mathbf{h}^{fuse} = (\mathbf{h}_a^{attn} + \mathbf{h}_w^{attn}) \oplus (\mathbf{h}_a^{max} + \mathbf{h}_w^{max})$$

where $\oplus$ denotes the vector concatenation, and the final hidden state $\mathbf{h}^{fuse}$ is always used as the audio-and-language representation for classification tasks.

3.3.1 Orthogonal Regularization

One key characteristic of multimodal learning is the generated representations of different modalities are supposed to depict a sample from different point of views. In order to encourage the two modules to get representations from different perspectives, we introduce a regularization term which aims at achieving the representation orthogonality during the fine-tuning stage:

$$\mathcal{L}_{\text{Orth}} = \frac{|\mathbf{h}_a^{attn\,T} \mathbf{h}_w^{attn}|}{\|\mathbf{h}_a^{attn}\| \, \|\mathbf{h}_w^{attn}\|} + \frac{|\mathbf{h}_a^{max\,T} \mathbf{h}_w^{max}|}{\|\mathbf{h}_a^{max}\| \, \|\mathbf{h}_w^{max}\|}$$

4 Experimental Setup and Results

4.1 Pre-training Details

We pre-train our CTAL on the public dataset LibriSpeech (Panayotov et al., 2015), which is a dataset of reading English speech, including both audio recordings and corresponding authorized transcripts. It has 7 subsets in total (train-clean-100, train-clean-360, train-other-500, dev-clean, dev-other, test-clean, test-other). The subsets with "clean" in their names contain audios with higher recording quality, while the other subsets have low-quality recordings. We use all three training subsets for pre-training, including approximately 960 hours of speech and 280K utterances.

Following Radford et al. (2019), we consider training a BBPE tokenizer on the LibriSpeech corpus with additional special tokens (<s>, </s>, <mask>, <pad>) as our language stream tokenizers. We tokenize the input text into token sequence as described in Section 3.1.1. For audio stream, we use Librosa (McFee et al., 2015), which is a well-established audio analysis Python package, to extract the 160-dimension input acoustic feature for each frame as described in Section 3.1.1. We denote the number of layers (i.e., language stream encoding layer and text-referred audio stream encoding layer) as N, the number of self-attention heads as A, and the number of hidden size as H. We primarily report results on two model sizes: $\mathbf{CTAL_{BASE}}$ (N=3, A=12, H=768) and $\mathbf{CTAL_{LARGE}}$ (N=6, A=12, H=768). The total number of parameters for $\mathbf{CTAL_{BASE}}$ is 60M and 110M for $\mathbf{CTAL_{LARGE}}$. We take the Adam (Kingma and Ba, 2015) as the optimizer with initial learning rate of 5e-5 and a linear-decayed learning rate schedule with warm up (Devlin et al., 2019). We pre-train our model using 4 16G-V100 GPUs with a batch size of 16 for 1,000,000 steps, and the whole pre-training process takes roughly 48 hours.

4.2 Fine-tuning on Downstream Tasks

We transfer our pre-trained CTAL model to three established SP tasks, with simple and necessary modifications on the output layers, loss function and training strategy.

4.2.1 Emotion Classification

In emotion classification task, given a speech clip, the model is asked to predict which emotion category the speech belongs to. Here, we conduct experiments on the widely-used dataset IEMOCAP (Busso et al., 2008). The dataset was recorded from ten actors, divided into five sessions, and each session has dialogues between two speakers with different genders. The dataset contains audio, transcriptions, and video recordings, and we only use audio and transcriptions in our study. The recorded dialogues have been sliced into utterances and labeled in 10 categories by three annotators and utterances without any text content are filtered out in our experiment. For consistent comparison with previous works, we follow the settings with Xu et al. (2019), which use four emotions (angry, happy, neutral and sad) for classification and perform 5-fold cross-validation over sessions, where each session is used as the test set in turn and remaining as training dataset. We adopt two widely used metrics for evaluation: weighted accuracy (WA) that is the overall classification accuracy and unweighted accuracy (UA) that is the average recall over all four classes. We report the averaged WA and UA over the 5-fold cross-validation experiments, and higher WA and UA results represent better model performances.

Methods	WA ↑	UA ↑
LSTM_alignment (Xu et al., 2019)	0.6900	0.7014
MRDE (Yoon et al., 2018)	0.6702	0.6764
MHA (Yoon et al., 2019)	0.6780	0.6880
$\mathrm{CTAL_{BASE}}$	0.7286	0.7370
$\mathrm{CTAL_{LARGE}}$	**0.7395**	**0.7463**

Table 1: Comparison to the SOTA methods on the IEMOCAP dataset.

To fine-tune on IEMOCAP, we represent the input sequence (for a pair of audio and text) as described in Section 4.1, and use the final hidden vector $\mathbf{h}^{fuse}$ as the audio-and-language representation. The only new parameters introduced during fine-tuning are classification layer weights $\mathbf{W} \in \mathbb{R}^{4\times d}$ and CTAL fine-tuning is driven by the cross-entropy loss between the predicted class and the gold label. We use a batch size of 4 and fine-tune for 20 epochs over each fold with one 16G-V100 GPU. We take AdamW (Loshchilov and Hutter, 2018) as the optimizer in fine-tuning stage, the learning rate is initialized as 1e-5 and we apply a cosine annealing learning rate schedule (Loshchilov and Hutter, 2017) to reach the optimum.

We select multiple models that claim to achieve the SOTA results on IEMOCAP dataset as our baselines. Please notice that previous methods are specifically designed for the task with no pre-training stage. Xu et al. (2019) aims to produce more strong multimodal representations by learning the alignment between speech frames and text words using an attention mechanism, i.e., "LSTM_alignment". Yoon et al. (2018) uses a dual-RNNs to encode the information from audio and text separately, then combines them by simple representations concatenation to predict emotion classes, i.e., "MDRE". Yoon et al. (2019) proposes a multi-hop attention mechanism to infer the correlation between audio and language modalities based on the output hidden representations of two bi-directional long short-term memory encoders, and output the final classification result from the concatenation of audio and language representations, i.e., "MHA".

Table 1 presents our experimental results on IEMOCAP dataset. Since some prior works experiment with different train/test split, we re-implement baseline models with their published

Methods	$\text{Acc}_2 \uparrow$	F1 $\uparrow$	MAE $\downarrow$	Corr $\uparrow$
MulT	0.7966	0.8008	0.6367	0.6292
$\text{CTAL}_{\text{BASE}}$	0.8036	0.8055	0.6061	**0.6828**
$\text{CTAL}_{\text{LARGE}}$	**0.8077**	**0.8101**	**0.6027**	0.6809

Table 2: Comparison to the SOTA methods on the CMU-MOSEI dataset.

codes[1,2]. Both $\text{CTAL}_{\text{BASE}}$ and $\text{CTAL}_{\text{LARGE}}$ outperform all three baselines by a substantial margin, obtaining 3.86% and 4.95% respective absolute WA improvement, and 3.56% and 4.49% respective absolute UA improvement over the prior state of the art.

4.2.2 Sentiment Analysis

The goal of the sentiment analysis task is to predict the degree of positive and negative sentiment. Compared to the emotion classification task, sentiment analysis is a regression task rather than a classification task. We adopt CMU-MOSEI (Zadeh et al., 2018) dataset for evaluation, which contains 23,454 movie review video clips from YouTube. We use only audio and corresponding transcriptions as input in our experiments. Each sample in the dataset is labeled with a sentiment score from -3 (strongly negative) to 3 (strongly positive) by human annotators. We follow the same experimental protocol as MulT (Tsai et al., 2019), with the same train/test data split and the same evaluation metrics, which includes two classification metrics: (1) binary accuracy (i.e., Acc_2: accuracy over positive/negative sentiments classification), and F1 score; (2) two regression metrics: mean absolute error (MAE), and the Pearson correlation coefficient (Corr) between model's predictions and human annotations. Since the prior top results reported on the CMU-MOSEI dataset are all achieved using all three modalities, so does MulT[3], we prune the vision-related components in MulT and re-train the model using only audio and text information.

During fine-tuning on sentiment analysis, we introduce additional parameters $\mathbf{w} \in \mathbb{R}^d$ to project the final hidden representation $\mathbf{h}^{fuse}$ to the sentiment score, and the model is trained to minimize the ℓ_1 loss between the predicted scores and the gold annotations. The other fine-tuning settings over CMU-MOSEI are almost the same as IEMOCAP. As show in Table 2, we observe improvements across all 4 metrics for CTAL over MulT baseline under both base and large settings.

4.2.3 Speaker Verification

Speaker verification focuses on verifying the speaker identity of an utterance through comparing it with the pre-recorded voiceprint information. In this experiment, we adopt LibriSpeech (Panayotov et al., 2015) for evaluation, which includes 292K utterances collected from more than 2,438 speakers. Following the same experiment setting with prior works (Wan et al., 2018; Jung et al., 2019), we fine-tune our pre-trained model with all training splits (train-clean-100, train-clean-360 and train-other-500), and evaluate it with test-clean part, which contains 40 brand new speakers to the training part. Please note that, although the train set for our speaker verification task is identical with the one we used for pre-training, the speaker identity information and test-clean data are not released during the pre-training process. Thus, it is fair to perform comparisons between our models with other prior works. We add a classifier over the head of fused embeddings $\mathbf{h}^{fuse}$ and adopt cross-entropy loss to fine-tune it. The output size of the classifier is same to the number of unique speakers in train set.

Methods	EER $\downarrow$
GE2E (Wan et al., 2018)	0.0379
RawNet (Jung et al., 2019)	0.0253
$\text{CTAL}_{\text{BASE}}$	0.0194
$\text{CTAL}_{\text{LARGE}}$	**0.0155**

Table 3: Comparison to the SOTA methods on the LibriSpeech dataset.

For evaluation, we utilize the representation before classifier as the input audio's identity embedding. Cosine distance of each paired audio embeddings is used as the indicator for the final decision. Similar to prior studies, we report the equal error rate (EER) as the evaluation metric, and lower EER represents better model performance. We choose two SOTA models as our baselines (Wan et al., 2018; Jung et al., 2019) where GE2E (Wan et al., 2018) designs a general loss function that emphasizes examples that are difficult to verify at each step of the training process, and RawNet (Jung et al., 2019) proposes an end-to-end network

[1]MDRE:https://github.com/david-yoon/multimodal-speech-emotion.git

[2]LSTM_alignment:https://github.com/didi/delta

[3]MulT:https://github.com/yaohungt/Multimodal-Transformer

that input raw audio waveforms to extract speaker embeddings. The comparison results are shown in Table. 3. From the table, we observe that our $\text{CTAL}_{\text{BASE}}$ outperforms GE2E and RawNet by 1.85% and 0.59% respectively, and $\text{CTAL}_{\text{LARGE}}$ outperforms two baselines by 2.24% and 0.98% respectively.

5 Analysis

5.1 Ablation Studies

We present the ablation result of different key components in CTAL in Table 4. For experimental efficiency, all of the ablation experiments are conducted with $\text{CTAL}_{\text{LARGE}}$.

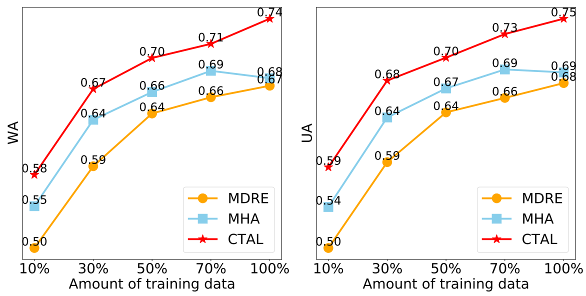
Overall, the pre-training of CTAL improves the performance across all the three downstream tasks (by comparing settings "w/o Pre-training" and $\text{CTAL}_{\text{LARGE}}$), and we find that $\text{CTAL}_{\text{LARGE}}$ significantly outperforms $\text{CTAL}_{\text{BASE}}$ across all tasks. Besides, with the increment in the size of pre-training data, CTAL achieves better performances on all evaluation metrics except Acc_2 and F1 in sentiment analysis task (by comparing settings (a) "pre-train with train-clean-360" and $\text{CTAL}_{\text{LARGE}}$). The effectiveness of the asymmetry encoder design for audio-and-language representations is demonstrated by comparing $\text{CTAL}_{\text{LARGE}}$ to LXMERT and VisualBERT, where all models are designed to have similar size of parameters.

By comparing (b) "w/o MLM" to "w/o Pre-training" and (c) "w/o MCAM" to "w/o Pre-training", we see the benefits of pre-training on MCAM and MLM respectively. However, by comparing (b) and (c) with $\text{CTAL}_{\text{LARGE}}$, both of them suffer dramatically performance decrease over all downstream tasks. This indicates the importance of joint-training with MLM and MCAM tasks during the pre-training stage. So far, the effectiveness of pre-training and different tasks are demonstrated.

Setting (d) "w/o Orthogonal Fusion" removes our proposed cross-modality orthogonal-fusion component and by comparing it with $\text{CTAL}_{\text{LARGE}}$, we observe that the model's performance decreases on all three downstream tasks, which proves its effectiveness. Setting (e) "w/o audio Outputs" and (f) "w/o language Outputs" only use the output embeddings from either audio or language encoding module for downstream fine-tuning. Through comparing them to (d), we find each kind of embeddings contributes to the Audio-and-Language tasks and the best performance is achieved through the appropriate fusion of both parts. At last, setting (g) "w/o Cross-modal Pre-training" utilizes unimodal pre-training models, RoBERTa and Mockingjay pre-trained with LibriSpeech dataset, and fuses their output embeddings for the downstream tasks. To be noticed, "w/o Cross-modal Pre-training" is chosen to have the same model size as $\text{CTAL}_{\text{LARGE}}$ for the comparison purpose. Additionally, we present the performance of each single modality pre-trained model, Mockinjay and RoBERTa, to demonstrate the advantages of multimodal pre-training. From the results, we find our approach still holds better performance across all three tasks, which proves the importance of introducing inter-modality learning during pre-training phase.

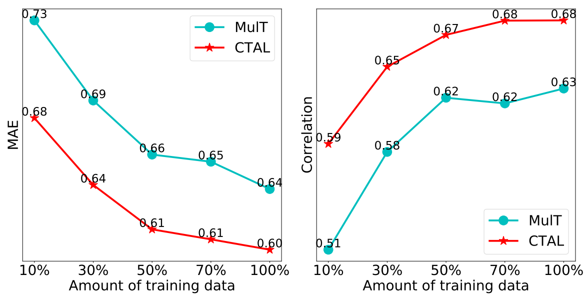
5.2 Effect of Pre-training

We analyze the effect of pre-trained CTAL by visualizing its performance on downstream tasks versus different proportions of training data being used.



(a) WA vs proportions (b) UA vs proportions

Figure 2: Results of models on different proportions of training data on IEMOCAP in terms of WA and UA.



(a) MAE vs proportions (b) Correlation vs proportions

Figure 3: Results of models on different proportions of training data on CMU-MOSEI in terms of MAE and Correlation.

In Figure 2a and Figure 2b, we show the performance on IEMOCAP dataset. First of all, on both metrics, CTAL outperforms all baselines across different proportions of training data. Secondly, the

Settings	MLM	MCAM	Orthognal Fusion	Cross-modal Pre-train	Text Outputs	Audio Outputs	Pre-train 960 Hours	Pre-train 360 Hours	Emotion Classification (IEMOCAP)		Sentiment Analysis (MOSEI)				Speaker Verification (LibriSpeech)
									WA ↑	UA ↑	Acc_2 ↑	F1 ↑	MAE ↓	Corr ↑	EER ↓
w/o pre-training			√		√	√			0.7004	0.7110	0.7804	0.7809	0.6654	0.6086	0.0366
(a)	√	√	√	√	√	√		√	0.7262	0.7386	**0.8127**	**0.8150**	0.6050	0.6804	0.0204
(b)		√	√	√	√	√	√		0.7077	0.7185	0.7834	0.7842	0.6629	0.6096	0.0244
(c)	√		√		√	√	√		0.7080	0.7171	0.7812	0.7809	0.6442	0.6440	0.0327
(d)	√	√		√	√	√	√		0.7338	0.7444	0.7948	0.7939	0.6035	**0.6832**	0.0176
(e)	√	√		√	√		√		0.6497	0.6586	0.7804	0.7795	0.6235	0.6639	-
(f)	√	√		√		√	√		0.7315	0.7412	0.7989	0.7915	0.6065	0.6750	0.0190
(g)	√	(MAM)	√		√	√	√		0.7116	0.7270	0.7820	0.7834	0.6323	0.6527	0.0306
Mockingjay		(MAM)				√	√		0.5505	0.5672	0.6887	0.7199	0.8056	0.3556	0.0551
RoBERTa	√				√		√		0.6377	0.6411	0.7451	0.7412	0.6598	0.5760	-
LXMERT	(LXMERT)		√	√	√	√	√		0.7145	0.7222	0.7749	0.7740	0.6405	0.6430	0.0320
VisualBERT	(VisualBERT)		√	√	√	√	√		0.6778	0.6848	0.7769	0.7722	0.6621	0.6243	0.0375
$\mathrm{CTAL}_{\mathrm{BASE}}$	√	√	√	√	√	√	√		0.7286	0.7370	0.8036	0.8055	0.6061	0.6828	0.0194
$\mathrm{CTAL}_{\mathrm{LARGE}}$	√	√	√	√	√	√	√		**0.7395**	**0.7463**	0.8077	0.8101	**0.6027**	0.6809	**0.0155**

Table 4: The results for performing ablation study with $\mathrm{CTAL}_{\mathrm{LARGE}}$. Notation "(MAM)" represents the acoustic stream encoding module is pre-trained with mask audio modeling (MAM) task. The EER is not reported for setting (d) and RoBERTa, because it does not make sense to perform speaker verification with only semantic embeddings.

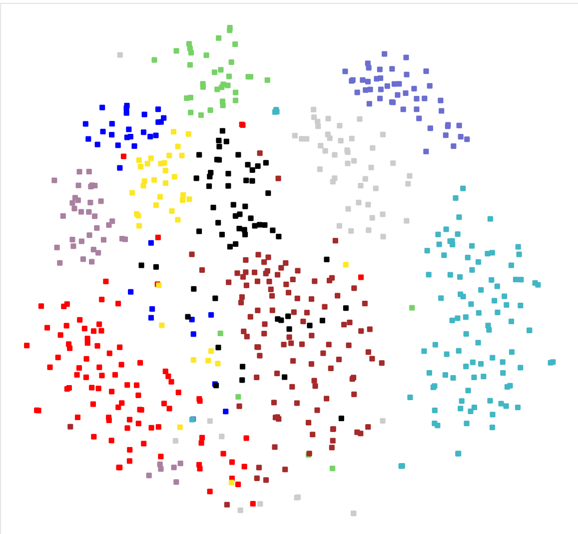


Figure 4: Visualization of 10 speakers embeddings via t-SNE. Different colors represent different speakers.

figures show that CTAL only needs half the amount of training data to achieve a better performance than baselines. The results on MOSEI dataset are shown in Figure 3a and Figure 3b, and the same conclusion can also be drawn.

In Figure 4, we use t-SNE (Van der Maaten and Hinton, 2008) to visualize the speaker embeddings in test set extracted from pre-trained CTAL without training on downstream tasks. Here, each point represents an utterance and different speakers have different colors. We can observe that the model can have some capability to distinguish utterances of different speakers with only pre-training.

6 Conclusion

In this work, we proposed CTAL, a novel pre-training cross-modal Transformer to learn effective representations for audio-and-language tasks. It is pre-trained with two pre-training tasks on a large-scale dataset of audio-and-language pairs. Extensive empirical analysis demonstrates that our pre-trained model improves various speech understanding performance significantly and achieves new SOTA results. Besides, we show the effectiveness of different model components and the competent generalization capability via detailed ablation studies and analysis.

Acknowledgements

This work was supported in part by National Key R&D Program of China, under Grant No. 2020AAA0104500 and in part by Beijing Nova Program (Z201100006820068) from Beijing Municipal Science & Technology Commission.

References

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2425–2433.

Alexei Baevski, Steffen Schneider, and Michael Auli. 2020a. vq-wav2vec: Self-supervised learning of discrete speech representations. In *8th International Conference on Learning Representations*.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020b. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642.

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.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. UNITER: Universal image-text representation learning. In *European Conference on Computer Vision*, pages 104–120. Springer.

Po-Han Chi, Pei-Hung Chung, Tsung-Han Wu, Chun-Cheng Hsieh, Yen-Hao Chen, Shang-Wen Li, and Hung-yi Lee. 2021. Audio albert: A lite bert for self-supervised learning of audio representation. In *IEEE Spoken Language Technology Workshop, SLT 2021, Shenzhen, China, January 19-22, 2021*, pages 344–350.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Michael Glodek, Stephan Tschechne, Georg Layher, Martin Schels, Tobias Brosch, Stefan Scherer, Markus Kächele, Miriam Schmidt, Heiko Neumann, Günther Palm, et al. 2011. Multiple classifier systems for the classification of audio-visual emotional states. In *International Conference on Affective Computing and Intelligent Interaction*, pages 359–368. Springer.

Albert Haque, Michelle Guo, Prateek Verma, and Li Fei-Fei. 2019. Audio-linguistic embeddings for spoken sentences. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 7355–7359. IEEE.

Dongwei Jiang, Xiaoning Lei, Wubo Li, Ne Luo, Yuxuan Hu, Wei Zou, and Xiangang Li. 2019. Improving transformer-based speech recognition using unsupervised pre-training. *arXiv preprint arXiv:1910.09932*.

Jee-weon Jung, Hee-Soo Heo, Ju-ho Kim, Hye-jin Shim, and Ha-Jin Yu. 2019. RawNet: Advanced end-to-end deep neural network using raw waveforms for text-independent speaker verification. In *The 20th Annual Conference of the International Speech Communication Association*, pages 1268–1272.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *The 3rd International Conference on Learning Representations*.

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard H. Hovy. 2017. RACE: large-scale reading comprehension dataset from examinations. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*, pages 785–794.

Quoc Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *International Conference on Machine Learning*, pages 1188–1196. PMLR.

Gen Li, Nan Duan, Yuejian Fang, Ming Gong, and Daxin Jiang. 2020a. Unicoder-VL: A universal encoder for vision and language by cross-modal pre-training. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 11336–11344.

Hang Li, Wenbiao Ding, Zhongqin Wu, and Zitao Liu. 2021. Learning fine-grained cross modality excitement for speech emotion recognition. In *Interspeech 2021, 22th Annual Conference of the International Speech Communication Association*, pages 3375–3379.

Hang Li, Yu Kang, Wenbiao Ding, Song Yang, Songfan Yang, Gale Yan Huang, and Zitao Liu. 2020b. Multimodal learning for classroom activity detection. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 9234–9238. IEEE.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019. VisualBERT: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*.

Andy T. Liu, Shang-Wen Li, and Hung-yi Lee. 2021. TERA: self-supervised learning of transformer encoder representation for speech. *IEEE ACM Trans. Audio Speech Lang. Process.*, 29:2351–2366.

Andy T Liu, Shu-wen Yang, Po-Han Chi, Po-chun Hsu, and Hung-yi Lee. 2020. Mockingjay: Unsupervised speech representation learning with deep bidirectional transformer encoders. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 6419–6423. IEEE.

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*.

Steven R Livingstone and Frank A Russo. 2018. The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english. *PloS one*, 13(5):e0196391.

Ilya Loshchilov and Frank Hutter. 2017. SGDR: stochastic gradient descent with warm restarts. In *The 5th International Conference on Learning Representations*.

Ilya Loshchilov and Frank Hutter. 2018. Fixing weight decay regularization in adam.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. ViLBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 13–23.

Brian McFee, Colin Raffel, Dawen Liang, Daniel PW Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto. 2015. librosa: Audio and music signal analysis in python. In *Proceedings of the 14th Python in Science Conference*, volume 8, pages 18–25. Citeseer.

Tomás Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*.

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*, pages 5206–5210. IEEE.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pages 1532–1543.

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

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100, 000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392.

Geovany A Ramirez, Tadas Baltrušaitis, and Louis-Philippe Morency. 2011. Modeling latent discriminative dynamic of multi-dimensional affective signals. In *International Conference on Affective Computing and Intelligent Interaction*, pages 396–406. Springer.

Steffen Schneider, Alexei Baevski, Ronan Collobert, and Michael Auli. 2019. wav2vec: Unsupervised pre-training for speech recognition. In *Interspeech 2019, 20th Annual Conference of the International Speech Communication Association*, pages 3465–3469.

Hao Tan and Mohit Bansal. 2019. LXMERT: Learning cross-modality encoder representations from transformers. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 5099–5110.

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 Annual Meeting of the Association for Computational Linguistics*.

Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9(11).

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Li Wan, Quan Wang, Alan Papir, and Ignacio Lopez Moreno. 2018. Generalized end-to-end loss for speaker verification. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 4879–4883. IEEE.

Yun Wang. 2018. Polyphonic sound event detection with weak labeling. *PhD thesis*.

Haiyang Xu, Hui Zhang, Kun Han, Yun Wang, Yiping Peng, and Xiangang Li. 2019. Learning alignment for multimodal emotion recognition from speech. In *Interspeech 2019, 20th Annual Conference of the International Speech Communication Association*, pages 3569–3573.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime G. Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 5754–5764.

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

Seunghyun Yoon, Seokhyun Byun, and Kyomin Jung. 2018. Multimodal speech emotion recognition using audio and text. In *2018 IEEE Spoken Language Technology Workshop*, pages 112–118. IEEE.

Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2017. Tensor fusion network for multimodal sentiment analysis.

In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1103–1114. Association for Computational Linguistics.

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 *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. From recognition to cognition: Visual commonsense reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6720–6731.