Self-Supervised Audio-and-Text Pre-training with Extremely Low-Resource Parallel Data

Yu Kang[1], **Tianqiao Liu**[1], **Hang Li**[1], **Yang Hao**[1], **Wenbiao Ding**[1, 2 *]
[1] TAL Education Group, Beijing, China
[2] Tencent, Beijing, China
{kangyu, liutianqiao, lihang4, haoyang2}@tal.com, darwinding@tencent.com

Abstract

Multimodal pre-training for audio-and-text has recently been proved to be effective and has significantly improved the performance of many downstream speech understanding tasks. However, these state-of-the-art pre-training audio-text models work well only when provided with large amount of parallel audio-and-text data, which brings challenges on many languages that are rich in unimodal corpora but scarce of parallel cross-modal corpus. In this paper, we investigate whether it is possible to pre-train an audio-text multimodal model with extremely low-resource parallel data and extra non-parallel unimodal data. Our pre-training framework consists of the following components: (1) Intra-modal Denoising Auto-Encoding (IDAE), which is able to reconstruct input text (audio) representations from a noisy version of itself. (2) Cross-modal Denoising Auto-Encoding (CDAE), which is pre-trained to reconstruct the input text (audio), given both a noisy version of the input text (audio) and the corresponding translated noisy audio features (text embeddings). (3) Iterative Denoising Process (IDP), which iteratively translates raw audio (text) and the corresponding text embeddings (audio features) translated from previous iteration into the new less-noisy text embeddings (audio features). We adapt a dual cross-modal Transformer as our backbone model which consists of two unimodal encoders for IDAE and two cross-modal encoders for CDAE and IDP. Our method achieves comparable performance on multiple downstream speech understanding tasks compared with the model pre-trained on fully parallel data, demonstrating the great potential of the proposed method.

Introduction

The recent rise of large-scale unsupervised pre-training models, i.e. BERT (Devlin et al. 2019), RoBERTa (Liu et al. 2019), Wav2Vec (Schneider et al. 2019), Mockingjay (Liu et al. 2020), etc., in artificial intelligence communities like natural language processing (NLP) and speech signal processing (SSP) has demonstrated the effectiveness of pretraining-finetuning framework. Subsequent study like CTAL (Li et al. 2021) successfully extends this framework to the multimodal field through designing cross-modal self-supervised tasks on the audio-and-text parallel corpus, and has significantly improved the performance of many downstream tasks at the intersection of audio and text which we denote as speech understanding tasks in the following.

However, different from collecting non-parallel unimodal corpus, the collection of parallel corpus like LibriSpeech always inquires additional manual filtering or annotating works. Due to this fact, the scale of available parallel corpus is limited comparing to the existing unimodal corpus, which restricts the benefits of large scale pre-training process.

In this paper, in order to get rid of the dependence of audio-and-text pre-training on parallelism between text and audio data, we present a novel multimodal pre-training framework mainly based on non-parallel unimodal corpora, which incorporates three components: (1) Intra-modal Denoising Auto-Encoding (IDAE), (2) Cross-modal Denoising Auto-Encoding (CDAE) and (3) Iterative Denoising Process (IDP). IDAE (Artetxe et al. 2018) is an effective algorithm for self-supervised training, through restoring the corrupted inputs in a unimodal manner, the model becomes capable to extract useful intra-modality information. CDAE is an extension of IDAE, which reconstructs the input text (audio), given both a noisy version of the input text (audio) and the corresponding translated noisy audio features (text embeddings). By learning to exploit the complementary information between noisy cross-modal inputs, this reconstruction process allows the model to learn inter-modal connections efficiently.

To construct the pseudo-parallel cross-modal inputs for CDAE, we propose a novel training procedure named IDP, which is inspired by the back-translation (Sennrich, Haddow, and Birch 2016; He et al. 2016) in neural machine translation. Back-translation is one of the most effective ways to leverage monolingual data for translation. Unlike back-translation, which regenerates pseudo-parallel data in each training round. IDP only performs one generation at the beginning of training, and applies the current model to eliminate the noise in the generated data in each subsequent training round to get pseudo-parallel inputs with less noise. IDP actually applies the reconstruction capabilities learned during CDAE, then the new pseudo-parallel data from IDP is used to further train CDAE at the next iteration, until convergence of the algorithm.

With the help of these components, our pre-training model is capable to learn both intra-modality and inter-modality connections with large-scale non-parallel unimodal corpus. To demonstrate its effectiveness, we apply our pre-training

[*] The corresponding author: Wenbiao Ding. Work done at TAL.

model to three established speech understanding tasks: emotion classification (Busso et al. 2008), sentiment analysis (Bagher Zadeh et al. 2018) and speaker verification (Panayotov et al. 2015). The empirical results show that our method outperforms all baselines. The main contributions of our paper are listed as follows:

- We present a novel multimodal pre-training method with large-scale non-parallel unimodal corpus for strong audio-and-text representations including both intra-modality and inter-modality connections. We are the first to introduce multimodal pre-training in the low-resource scenarios which scarce of parallel cross-modal data.
- Comprehensive empirical evidence demonstrates that our pre-training model outperforms all baselines on various downstream speech understanding tasks, such as emotion classification, sentiment analysis, and speaker verification. We conduct detailed ablation studies and analysis to prove the effectiveness of our pre-training strategies.
- Further experiments show that our pre-training method can be effective in semi-supervised manner as well. That is, in scenarios where sufficient parallel cross-modal data is available, adding extra non-parallel unimodal data can further improve the performance.

Related Work

Multimodal Pre-training

Inspired by the success of language pre-training like BERT (Devlin et al. 2019) , the research community has started to pay more attention to pre-training in multimodal scenarios and has achieved remarkable results. There are quite a few attempts that have been made to pre-train models for vision-and-language tasks. In general, these pre-training methods can be divided into two categories, according to their different encoder architectures. (a) Previous 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 late fusion manner. (b) The other category of pre-training frameworks like VisualBert (Li et al. 2019) and Unicoder-VL (Li et al. 2020), concatenate vision and language features as a unified single-stream input and utilize a universal encoder to learn joint multimodal representations. It is worth noting that, the authors of VL-BERT (Su et al. 2020) claim that their unified architecture outperforms the two-stream designs. However, the unified architecture may not be suitable for our pre-training method, since the universal encoder lacks of the ability to perform the translation between two modalities.

The architecture of our dual Transformer is similar to that of LXMERT and ViLBERT. However, the model structure on which our proposed pre-training method relies is not fixed. Any model structure with the capability of translating one modality to the other can work well with our approach.

While pre-training for vision-and-language has made some progresses in recent years, audio-and-text pre-training has also started to evolved recently. SpeechBERT (Chuang et al. 2019) proposes a pre-training audio-and-text model for spoken question answering task. However, the pre-training task in SpeechBERT relies on parallel cross-modal data with forced alignment information between words and audio signals, and the downstream task to which the pre-training model applies are too limited. CTAL (Li et al. 2021) proposes a pre-training model to learn audio-and-text representations for multiple downstream speech understanding tasks. However, the self-supervised pre-training task introduced in CTAL still needs large-scale of parallel audio-and-text data, which is unavailable in many low-resource languages. So, in this paper, we design several truly self-supervised tasks for audio-and-text pre-training mainly based on non-parallel unimodal data.

Back-Translation

Back-translation (Sennrich, Haddow, and Birch 2016; He et al. 2016) in neural machine translation is a very effective data-augmentation scheme under the semi-supervised setting. A translation model is first trained on the available parallel data, then the model is used to produce translations from the extra monolingual corpus. The pairs composed of these translations with their corresponding monolingual data are then used as additional training data for the original translation system to further boost the performance. The same approach is also applied by (Tjandra, Sakti, and Nakamura 2017) to enhance the performance of automatic speech recognition (ASR) and text to speech (TTS) models simultaneously with extra unpaired audio and text data. However, it still relies on a number of audio-text pairs to do supervised training and get valid ASR and TTS models firstly.

Some similar works to ours are the unsupervised machine translation (Lample et al. 2018a,b; Artetxe et al. 2018) with back-translation mechanism. The authors have carefully designed some initialization methods to obtain a rough translation model to replace the supervised training process in the semi-supervised setting. Then the denoising effect of language models and automatic generation of parallel data by iterative back-translation are leveraged round by round until convergence of the algorithm. (Ren et al. 2019; Xu et al. 2020) also proposed the unsupervised ASR and TTS models in the similar way.

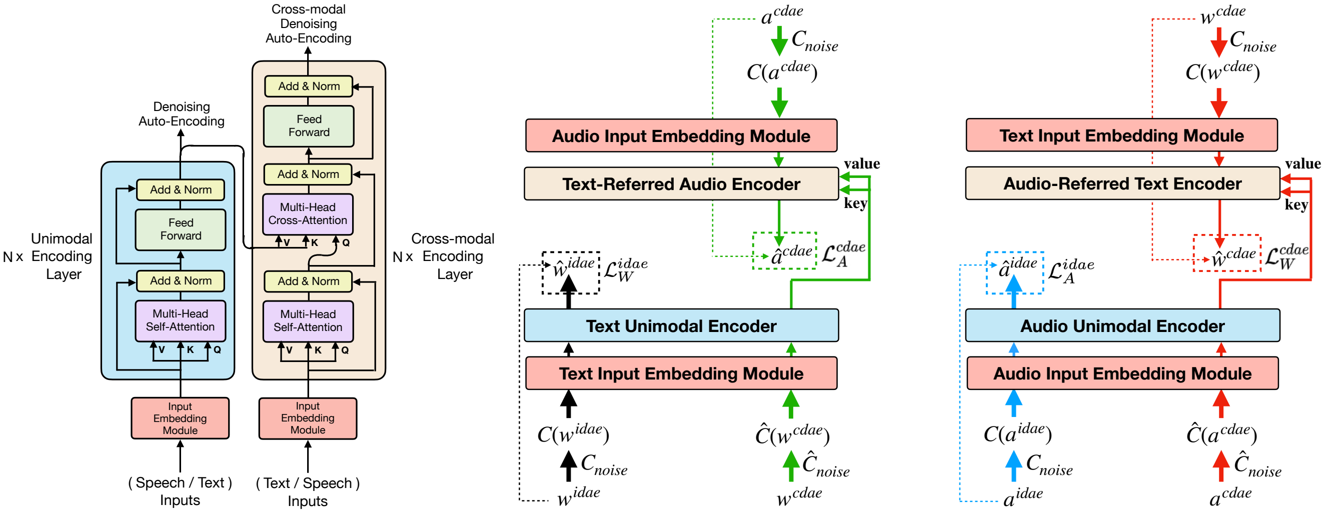
In our work, (1) the IDP leverages the denoising capability of model learned during the training process to iteratively eliminate the noise in the translated text embeddings (audio features), instead of regenerating new translation results each round like back-translation. And (2) our cross-modal encoder is able to learn bidirectional audio-and-text representation which is important for finetuning on downstream tasks, while the autoregressive decoder in previous works is only able to learn representation with unidirectional information. (3) During back-translation, the autoregressive decoder must generate translations step by step, which is very slow for long sequences like audio signals. Therefore, this is unacceptable in large-scale pre-training scenarios. In contrast, IDP can take full advantage of the GPU to eliminate noise in parallel.

Approach

In this section, we first describe the model architecture of the dual Transformer, and then we detail our pre-training strategies.

Model Architecture

We adapt dual Transformer as our backbone model, which has been verified to be an effective cross-modal architecture



(a) Model Architecture

(b) Pre-training Procedure

Figure 1: (a): The structure of each encoder in our unimodal encoders and cross-modal encoders. (b): The overall pre-training flow of our method, which consists of the intra-modal denoising auto-encoding (IDAE) and the cross-modal denoising auto-encoding (CDAE), it is worth noting that, we do not present the iterative denoising process (IDP) in the figure. The solid lines denote the forward propagation and the dash lines denote the self-supervised signals.

by (Tan and Bansal 2019) and (Lu et al. 2019), and we extend the original implementation to suit our training procedure as shown in Figure 1a. It consists of three subcomponents, the input embedding module, the unimodal encoder and the cross-modal encoder, each corresponding to two different modalities, and all modules have separate parameters. We will dive into details in the following words.

Input Embedding Module The input embedding module for text consists of a token embedding layer and a position embedding layer. We follow the text preprocessing of RoBERTa (Liu et al. 2019) to encode the input text to token embeddings for text modality.

The input module for audio is composed of a position embedding layer and a dense layer which projects the audio features to hidden size. Following (Chi et al. 2021), each 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. Finally, we feed these features to the dense layer and add them with the position embeddings to obtain the audio embeddings for audio modality.

Unimodal Encoders We apply the original implementation of the Transformer encoder to encode the input embeddings to the unimodal representations for text and audio separately, named Text Unimodal Encoder and Audio Unimodal Encoder.

Cross-modal Encoders Then, we apply two cross-modal encoders, named Audio-Referred Text Encoder and Text-Referred Audio Encoder respectively, each encoder includes N identical layers. We input the embeddings of one modality to the first layer, and by stacking N these layers, we can get the final cross-modal representations (i.e. text-referred audio representations or audio-referred text representations). Inside each layer, the self-attention sub-layer is first applied to learn the intra-modality representation, then we apply cross-modal attention sub-layer which accepts the final representations from the unimodal encoder of the other modality as key and value to learn the inter-modality interactions. Finally, we obtain the audio-referred text representations $\boldsymbol{H}_w^N \in \mathbb{R}^{\mathcal{T}_w \times d}$ and text-referred audio representations $\boldsymbol{H}_a^N \in \mathbb{R}^{\mathcal{T}_a \times d}$, where d denotes the hidden size of the representations. $\mathcal{T}_w$ and $\mathcal{T}_a$ are the output lengths for each modality.

Pre-training Problem Statement

We consider a dataset which consists of two large-scale non-parallel unimodal corpora and a limited parallel cross-modal corpus. We denote the non-parallel unimodal corpora for text and audio as $\mathcal{W}_{unpaired}$ and $\mathcal{A}_{unpaired}$ respectively, and these two corpora do not correspond to each other. Meanwhile, the parallel corpus includes a small set of paired data $(\mathcal{W}_{paired}, \mathcal{A}_{paired})$. Then, our objective is to conduct a well performing audio-and-text pre-training model with the above dataset.

Intra-modal Denoising Auto-Encoding

Given the large amount of non-parallel audio and text data, building capabilities of learning context-sensitive unimodal representations and reconstructing corrupted inputs is of great importance. To this end, we leverage denoising auto-encoder (Vincent et al. 2008) to reconstruct the audio and text inputs from the corrupted version of itself, which is widely used in self-supervised learning. The objective function $\mathcal{L}^{idae}$ of the intra-modal denoising auto-encoding (IDAE) is formulated as:

$$\begin{aligned}\mathcal{L}^{idae} &= \mathcal{L}_W(w|C(w);\theta_{uni}^W; w \in \{\mathcal{W}_{unpaired}, \mathcal{W}_{paired}\}) \\ &+ \mathcal{L}_A(a|C(a);\theta_{uni}^A; a \in \{\mathcal{A}_{unpaired}, \mathcal{A}_{paired}\})\end{aligned}$$

where θ_{uni}^W and θ_{uni}^A denote the model parameters of the text unimodal encoder and the audio unimodal encoder respectively. C is a corrupt function, where for text modality, we dynamically select some input tokens with a probability of 15% and the selected tokens are replaced with a special <mask> token 80% of the time, a random token 10%, and unaltered 10%. For audio modality, the corrupt function is a little bit different, at first, we split the audio features in separate segments according to S_{num} successive frames per segment, where S_{num} 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 S_{num} 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 from exploiting local smoothness of acoustic frames and the model is required to reconstruct inputs based on global information rather than local messages. $\mathcal{L}_W$ and $\mathcal{L}_A$ denote the loss for text and audio respectively, and we implement them as:

$$\begin{aligned}\mathcal{L}_W(y|x;\theta^W_{uni}) &= -\log P(y|x;\theta^W_{uni})\\ \mathcal{L}_A(y|x;\theta^A_{uni}) &= L1(y, f(x;\theta^A_{uni}))\end{aligned}$$

where $L1$ denotes the L1 loss function. It is worth noting that although the $\mathcal{L}_W$ and $\mathcal{L}_A$ have different scales of loss, our attempts to balance them do not have any effect on the final performance, so we choose to leave them unchanged for all experiments.

Iterative Denoising Process

Constructing pseudo-parallel data without introducing too much noise is the key component in leveraging the large-scale non-parallel corpora for cross-modal pre-training. For this purpose, we design an iterative denoising process (IDP) to perform modality translation with step-by-step noise reduction. We can regard the dual Transformer as a text-to-audio Transformer and a audio-to-text Transformer. In order to translate text to audio, we input the unimodal text corpus to the text unimodal encoder and input the translated audio features from last iteration to the text-referred audio encoder. These two encoders make up the text-to-audio Transformer. Similarly, for audio-to-text translation, we switch to use the audio-to-text Transformer which composed of the audio unimodal encoder and the audio-referred text encoder, accepting unimodal audio corpus and translated text embeddings from last iteration as inputs respectively. Before the start of each iteration in the pre-training phase, we first translate the non-parallel unimodal text (audio) data w (a) into less-noisy $\tilde{a}$ ($\tilde{w}$). Specifically, we obtain the translated results for the k-th training iteration with the followings:

$$\begin{aligned}\tilde{w}_k &= f(\tilde{w}_{k-1}, f(a;\theta^A_{uni});\theta^W_{cross})\\ \tilde{a}_k &= f(\tilde{a}_{k-1}, f(w;\theta^W_{uni});\theta^A_{cross})\end{aligned}$$

where θ^W_{cross} and θ^A_{cross} denote the model parameters of the audio-referred text encoder and the text-referred audio encoder respectively. $f(;\theta)$ denotes one forward propagation of an encoder in the dual Transformer, and θ determines exactly which encoder it is.

Instead of getting raw text or raw audio signal after modality translation, $\tilde{w}_k$ and $\tilde{a}_k$ should retain information of a and w as much as possible. To this end, for text modality we output embeddings rather than tokens, and for audio modality we output the acoustic features mentioned in input embedding module section.

Warm-Up Stage

To jump start the process, we suggest a warm-up stage at the very beginning of training to obtain a rough but reasonable initial translation results, and in the subsequent training process, we use the IDP to gradually eliminate the noise in it.

Firstly, we train the text-to-audio Transformer and audio-to-text Transformer with the low-resource parallel data $(\mathcal{W}_{paired}, \mathcal{A}_{paired})$ as follows:

$$\begin{aligned}\mathcal{L}^{warm} &= \mathcal{L}_W(w|a, w_{masked};\theta^A_{uni},\theta^W_{cross})\\ &\quad + \mathcal{L}_A(a|w, a_{masked};\theta^W_{uni},\theta^A_{cross})\end{aligned}$$

where $(w,a) \in (\mathcal{W}_{paired}, \mathcal{A}_{paired})$, w_{masked} (a_{masked}) is the sequence of special mask tokens (segments) with the equal length to w (a). For training the text-to-audio Transformer, we input w to the text unimodal encoder and input a_{masked} to the text-referred audio encoder to reconstruct a. Training the audio-to-text Transformer is similar. In this way, the model is constrained to perform modality translation based solely on the source modality. After the initial training, we initialize $\tilde{w}_0$ and $\tilde{a}_0$ for unimodal non-parallel corpora as follows:

$$\begin{aligned}\tilde{w}_0 &= f(w_{masked}, f(a;\theta^A_{uni});\theta^W_{cross})\\ \tilde{a}_0 &= f(a_{masked}, f(w;\theta^W_{uni});\theta^A_{cross})\end{aligned}$$

where $w \in \mathcal{W}_{unpaired}$ and $a \in \mathcal{A}_{unpaired}$.

It's worth noting that since we do not know the exact length of the ground truth, we set the lengths of w_{masked} and a_{masked} as two hyper-parameters which actually limit the maximum amount of information that can be included in the translation results. And we set 256 for text and 1000 for audio according to experiments. If the exact length of the ground truth is much longer than the length hyper-parameter, the model should learn to retain as much important information as possible during the translation process, and similarly, if the exact length is much shorter than the length hyper-parameter, then the model should learn to use something like padding information to fill the information space.

Cross-modal Denoising Auto-Encoding

We extend the intra-modal denoising auto-encoding (IDAE) to cross-modal denoising auto-encoding (CDAE) for learning inter-modality connections and cross-modal denoising capability. We train the model by reconstructing the input text (audio), given both a noisy version of the same input text (audio) and the noisy translated audio features (text embeddings) from IDP.

Non-parallel Corpus For training the text-to-audio Transformer, we input $\tilde{w}$ to the text unimodal encoder and input corrupted a to the text-referred audio encoder to reconstruct a. As for training the audio-to-text Transformer, we switch to use $(\tilde{a}, w)$ as the pseudo-parallel inputs. The objective function is formulated as follows:

$$\begin{aligned}\mathcal{L}^{cdae}_{unpaired} &= \mathcal{L}_W(w|\tilde{a}, C(w);\theta^A_{uni},\theta^W_{cross})\\ &\quad + \mathcal{L}_A(a|\tilde{w}, C(a);\theta^W_{uni},\theta^A_{cross})\end{aligned}$$

We adapt almost the same corrupt function C as described in IDAE section. However, in order to prevent the model from reconstructing inputs based only on itself and ignoring the cross-modal information, we increase the probability of corruption in C from 15% to 30%.

The intuition behind CDAE process is that as long as the initial translation results $\tilde{w}_0$ and $\tilde{a}_0$ retain at least some information of the source modalities, the unimodal encoders will map such translation into a representation in feature space that also corresponds to a cleaner version of the input, since the unimodal encoders are trained to denoise in IDAE. At the same time, the cross-modal encoders are trained to predict noiseless outputs, conditioned on noisy cross-modal inputs in CDAE. Putting these two pieces together will eliminate noise in the translations, which will enable better IDP outputs at the next iteration, and so on so forth.

Parallel Corpus We treat the cross-modal parallelism provided by parallel corpus as prior knowledge and use it to guide the non-parallel training process. The CDAE for parallel corpus is straightforward and the objective function is defined as follows:

$$\begin{aligned}\mathcal{L}^{cdae}_{paired} &= \mathcal{L}_W(w|a, C(w); \theta^A_{uni}, \theta^W_{cross}) \\ &\quad + \mathcal{L}_A(a|w, C(a); \theta^W_{uni}, \theta^A_{cross})\end{aligned}$$

where $(w, a) \in (\mathcal{W}_{paired}, \mathcal{A}_{paired})$, and the noise function C is exactly the same as that of non-parallel data. However, during non-parallel training, the input of unimodal encoder $\tilde{w}$ and $\tilde{a}$ are noisy, which shows inconsistency from the parallel corpus scenario, as the input w and a are noiseless. Thus we apply a new corrupt function $\hat{C}$ on input a and w in $\mathcal{L}_W$ and $\mathcal{L}_A$ respectively, to imitate the noise from modality translation. In detail, we first apply IDP to parallel corpus following the same procedures aforementioned, which translates w (a) into $\tilde{a}$ $(\tilde{w})$. Then we select 30% tokens (segments) in w (a), and replace them with the tokens (segments) of $\tilde{w}$ $(\tilde{a})$ in the same position. We re-formulate the CDAE for paired corpus as follows:

$$\begin{aligned}\mathcal{L}^{cdae}_{paired} &= \mathcal{L}_W(w|\hat{C}(a), C(w); \theta^A_{uni}, \theta^W_{cross}) \\ &\quad + \mathcal{L}_A(a|\hat{C}(w), C(a); \theta^W_{uni}, \theta^A_{cross})\end{aligned}$$

At last, we denote $\mathcal{L}^{cdae} = \mathcal{L}^{cdae}_{unpaired} + \mathcal{L}^{cdae}_{paired}$.

Final Objective Function

The general procedure of our pre-training method is shown in Figure 1b, and the final objective function of our learning algorithm is as follows:

$$\begin{aligned}\mathcal{L}(\theta^W_{uni}, \theta^A_{uni}, \theta^W_{cross}, \theta^A_{cross}) &= \mathcal{L}^{idae}(\theta^W_{uni}, \theta^A_{uni}) \\ &\quad + \mathcal{L}^{cdae}(\theta^W_{uni}, \theta^A_{uni}, \theta^W_{cross}, \theta^A_{cross})\end{aligned}$$

Overall Training Flow

The overall algorithm is described in Algorithm 1. A potential direction of improvement for our algorithm is that there is discrepancy of the type of noise in the inputs of cross-modal encoders between the CDAE and IDP. Specifically, the inputs of cross-modal encoders contain <mask> token for text and masked all-zero segment for audio in CDAE phase, while these special mask tokens and segments do not appear in IDP. To mitigate this discrepancy, we decrease the probability of replacing the selected time-steps with masked ones from 80% to 60% during CDAE pre-trainig, as a lower probability will cause the training to become unstable or even difficult to converge according to our experiment results.

Algorithm 1: Self-supervised Multimodal Pre-training with Low-Resource Parallel Data

```
Input: ${\mathcal{W}_{unpaired}, \mathcal{A}_{unpaired}}$: large-scale non-parallel corpora;
    $(\mathcal{W}_{paired}, \mathcal{A}_{paired})$: low-resource parallel corpus; $K$: the total
    number of training epochs; $T$: the total number of warm-up
    epochs
Output: $\theta_{all}$: $\theta^W_{uni}$; $\theta^A_{uni}$; $\theta^W_{cross}$; $\theta^A_{cross}$
Warm Up
for $t = 1, T$ do
    $\theta_{all} \leftarrow arg \min \mathcal{L}^{warm}$
end for
$\tilde{w}_0 = f(w_{masked}, f(a; \theta^A_{uni}); \theta^W_{cross})$
$\tilde{a}_0 = f(a_{masked}, f(w; \theta^W_{uni}); \theta^A_{cross})$
Iterative Denoising
for $k = 1, K$ do
    $\theta^W_{uni}, \theta^A_{uni} \leftarrow arg \min \mathcal{L}^{idae}$
    $\theta_{all} \leftarrow arg \min \mathcal{L}^{cdae}$
    with no gradient:
        $\tilde{w}_k = f(\tilde{w}_{k-1}, f(a; \theta^A_{uni}); \theta^W_{cross})$
        $\tilde{a}_k = f(\tilde{a}_{k-1}, f(w; \theta^W_{uni}); \theta^A_{cross})$
end for
return $\theta_{all}$: $\theta^W_{uni}$; $\theta^A_{uni}$; $\theta^W_{cross}$; $\theta^A_{cross}$
```

Experimental Setup and Result

Pre-training Details

We pre-train our model on the LibriSpeech (Panayotov et al. 2015) dataset, which includes both audio recordings and corresponding authorized transcripts of English reading speech. We consider three subsets of LibriSpeech for pre-training: *train-clean-100*, *train-clean-360*, *train-other-500*. In our experiments, we sample the low-resource parallel corpus from *train-clean-100* subset, and in order to build the non-parallel corpus, we first combine the remaining two subsets, then split it in half and take the text from one half and the audio from the other half. In the end, we form a parallel corpus including 1 hour of speech and 500 utterances, and a non-parallel corpus including 430 hours of speech and 126k utterances.

For the dual Transformer, each encoder in both unimodal encoders and cross-modal encoders has 3 layers, the number of multi-head attention heads is 12 and the hidden size is 768. The total number of parameters of the whole dual Transformer is 110M.

We take Adam (Kingma and Ba 2015) as our optimizer with initial learning rate of 2e-5 and a linear-decayed learning rate schedule with warm-up (Devlin et al. 2019). We pre-train our model using 4 32G-V100 GPUs with a batch size of 8 for 500,000 steps, and the whole pre-training process takes roughly 72 hours.

Fine-tuning on Downstream Tasks

We apply our pre-training model to three different kinds of widely-studied speech understanding tasks, only with simple modifications. During fine-tuning, we directly input the parallel data in downstream dataset to our model and obtain

Methods	WA ↑	UA ↑
LSTM_alignment (Xu et al. 2019)	0.6900	0.7014
MRDE (Yoon, Byun, and Jung 2018)	0.6702	0.6764
MHA (Yoon et al. 2019)	0.6780	0.6880
Our Method	**0.7254**	**0.7339**

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

the audio-referred text representations $\boldsymbol{H}_w^N \in \mathbb{R}^{\mathcal{T}_w \times d}$ and text-referred audio representations $\boldsymbol{H}_a^N \in \mathbb{R}^{\mathcal{T}_a \times d}$ from the cross-modal encoders. Our final representation is $h^{fuse} = [\mathcal{MP}(\boldsymbol{H}_w^N); \mathcal{MP}(\boldsymbol{H}_a^N)] \in \mathbb{R}^{2\times d}$, where $\mathcal{MP}$ is mean pooling over the output length.

Emotion Classification In emotion classification task, given a speech clip, the model is asked to predict which emotion category the speech belongings to. Here, we use the widely-used dataset IEMOCAP (Busso et al. 2008). IEMOCAP is recorded across 5 sessions with 5 pairs of speakers. We follow the settings with (Xu et al. 2019) for consistent comparisons with previous works, which perform 5-fold cross-validation over sessions. We adopt the same two metrics as previous works for evaluation: weighted accuracy (WA) that is the overall classification accuracy and unweighted accuracy (UA) that is the average recall over all classes. The reported WA and UA are averaged over the 5-fold cross-validation experiments, and higher WA and UA results represent better model performances.

To fine-tune on IEMOCAP, the only new parameters are weights of classification layer $\mathbf{W} \in \mathbb{R}^{4\times(2\times d)}$, which is applied to h^{fuse}. The training is driven by the cross-entropy loss between the predicted class and the gold label.

We select multiple models that claim to achieve SOTA results on IEMOCAP dataset as our baselines. Table 1 presents our experimental results on IEMOCAP dataset. Since some prior works experiment with different train/test split, we reimplement baseline models with their published codes and unify the split as aforementioned. Our proposed method outperforms all three baselines, obtaining 3.54% and 3.25% absolute improvement on WA and UA respectively over the prior state of the art.

Sentiment Analysis We adopt CMU-MOSEI (Zadeh et al. 2018) dataset to evaluate the sentiment analysis task, which aims to predict the degree of positive and negative sentiment. We follow the same experimental protocol as MuIT (Tsai et al. 2019), with the same train/test data split. Unlike the emotion classification task, sentiment analysis is a regression task, so we adopt two widely-used regression metrics for evaluation: 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. So we reimplement it, prune the vision-related components, and re-train the model using only audio and text information.

During fine-tuning on MOSEI, we introduce additional parameters $\mathbf{W} \in \mathbb{R}^{1\times(2\times d)}$ to project the final representations h^{fuse} to the sentiment score, and the model is trained to minimize the L1 loss between the predicted scores and the labels. As show in Table 2, we observe that our method improves the MAE and Corr by 2.03% and 5.12% over MulT.

Methods	MAE ↓	Corr ↑
MulT (Tsai et al. 2019)	0.6367	0.6292
Our Method	**0.6164**	**0.6804**

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

Methods	EER ↓
GE2E (Wan et al. 2018)	0.0379
RawNet (Jung et al. 2019)	0.0253
Our Method	**0.0210**

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

Speaker Verification The goal of the speaker verification task is to verify the speaker identity of an utterance through comparing it with the pre-recoded voiceprints. In the text-independent speaker verification task, text can provide much more information than just text content, as a strong audio-and-text representation will include fine-grained cross-modal information such as the way and speed of pronunciation of each word, which also contains strong speaker features. In this experiment, we adopt LibriSpeech (Panayotov et al. 2015) for evaluation, which consists of 7 subsets in total. Following the same experiment settings with prior works (Wan et al. 2018; Jung et al. 2019), we fine-tune our pre-trained model with all training subsets, and evaluate it with *test-clean* part, which contains 40 brand new speakers to the training part. Please note here, 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 model with other prior works. We add a two-layer dense layer and a classifier over the head of final representations h^{fuse} and adopt cross-entropy loss to fine-tune it. The output size of the classifier is the same as the number of unique speakers in train set.

For evaluation, we utilize the representations before classifier as the speaker embedding of input speech recording. And the cosine distance of each paired speaker embeddings are 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). The comparison results are shown in Table. 3. From the table, we observe that the proposed method outperforms GE2E and RawNet by 1.69% and 0.43% respectively.

Analysis

Ablation Studies

In order to study the effectiveness of different key components in our method, we present the ablation results in Table. 4.

Effect of Key Components in Pre-training Strategies and Model Overall, our pre-training method improves the per-

Settings	Emotion Classification		Sentiment Analysis		Speaker Verification
	WA ↑	UA ↑	MAE ↓	Corr ↑	EER ↓
w/o Pre-training	0.7083	0.7119	0.6586	0.6238	0.0354
w/o IDAE	0.7112	0.7201	0.6405	0.6506	0.0296
w/o IDP	0.7241	0.7297	0.6199	0.6772	0.0232
w/o Audio Outputs	0.6998	0.7079	0.6353	0.6333	_
w/o Text Outputs	0.7174	0.7217	0.6211	0.6655	0.0273
w/o Paired Data	0.7108	0.7220	0.6424	0.6342	0.0307
w/ Less Unpaired Data	0.7129	0.7154	0.6469	0.6318	0.0335
RoBERTa	0.6858	0.7003	0.6560	0.6152	_
Mockingjay	0.6701	0.6823	0.6676	0.6011	0.0320
Late Fusion	0.7103	0.7257	0.6409	0.6434	0.0313
Our Method	0.7254	0.7339	0.6164	**0.6804**	0.0210
w/ Fully Paired Data	**0.7336**	**0.7406**	**0.6045**	0.6794	**0.0181**

Table 4: The results for performing ablation studies with proposed method. The EER is not reported for settings "w/o Audio Outputs" and "RoBERTa" because it does not make sense to perform speaker verification with only text features.

formance across all three downstream tasks (by comparing settings "w/o Pre-training" and "Our Method"). By comparing settings "w/o IDAE" with "Our Method", we see the benefits of IDAE's denoising capability and the importance of intra-modal connections. Setting "w/o IDP" removes our proposed iterative denoising process (IDP), and in each training iteration, we regenerate the translation results only conditioned on the data of source modality, like back-translation. By comparing it with "Our Method", we observe that the model's performances decrease on all three tasks, which proves the effectiveness of IDP. Each setting of "w/o Audio Outputs" and "w/o Text Outputs" only uses the output representations from either audio-referred text encoder or text-referred audio encoder for fine-tuning, and through comparing them to "Our Method", we find each of the outputs contributes to the downstream speech understanding tasks.

Effect of Pre-training Data Size Setting "w/o Paired Data" removes the parallel corpus and pre-trains the model only using non-parallel corpus, by comparing it with "Our Method", we observe that the model's performances drop on all downstream tasks, which proves the importance of parallel data and its guidance to the pre-training procedure. Besides, with the increment in the size of non-parallel data, our model achieves better performances on all evaluation metrics (by comparing setting "w/ Less Unpaired Data" and "Our Method"). At last, setting "w/ Fully Paired Data" is the model pre-trained with fully parallel data (431 hours in total). We use the text data in the non-parallel corpus with their corresponding audio and combine with the parallel corpus to pre-train this model. From the results, we find "Our Method" can achieve comparable performance across all three tasks and even better on correlation coefficient in sentiment analysis.

Comparisons with Unimodal Pre-training Models Then, setting "RoBERTa" (Liu et al. 2019) and setting "Mockingjay" (Liu et al. 2020) are unimodal pre-training models pre-trained only with text data and audio data in our non-parallel corpus respectively. Setting "Late Fusion" concatenates their output representations for downstream fine-tuning. It is worth noting that, the total number of these two unimodal models' parameters equals to that of "Our Method" for comparison purposes. By comparing setting "Late Fusion" with "Our Method", we find our approach still outperforms all three tasks, which proves the importance of introducing inter-modality learning during pre-training phase (the effect of cooperation between IDP and CDAE). Furthermore, by comparing settings "RoBERTa" and "Mockingjay" with settings "w/o Audio Outputs" and "w/o Text Outputs", we can find that using either text outputs or audio outputs from our cross-modal encoders in downstream tasks can achieve better performances than using that from unimodal models, which indicates that each of our cross-modal encoders' outputs already includes information from both modalities, and the best performance is achieved through the fusion of both parts.

Effectiveness Under Semi-supervised Setting

Settings	Emotion Classification		Sentiment Analysis		Speaker Verification
	WA ↑	UA ↑	MAE ↓	Corr ↑	EER ↓
w/ Sufficient Paired Data	0.7315	0.7362	0.6093	0.6801	0.0192
w/ External Unpaired Data	**0.7352**	**0.7420**	**0.6047**	**0.6807**	**0.0167**

Table 5: The results for demonstrating the effectiveness of our method under semi-supervised setting.

At last, we further study the effectiveness of our method under semi-supervised setting. We use *train-clean-360* subset in LibriSpeech as our parallel corpus, and form an external non-parallel corpus including 300 hours of speech from remaining training subsets in the same way as mentioned in pre-training details section, the results are presented in Table. 5. From the results, we can see that even if we have sufficient parallel data for pre-training, leveraging external non-parallel data by our method can further boost the performance.

Conclusion

In this work, we proposed an audio-and-text pre-training framework which leverages only low-resource parallel corpus and extra non-parallel corpus. Our pre-training model achieves comparable performance on multiple downstream tasks compared with the model pre-trained on fully parallel data and outperforms all baselines, demonstrating the great potential of our method.

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

Artetxe, M.; Labaka, G.; Agirre, E.; and Cho, K. 2018. Unsupervised Neural Machine Translation. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.

Bagher Zadeh, A.; Liang, P. P.; Poria, S.; Cambria, E.; and Morency, L.-P. 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 (Volume 1: Long Papers)*, 2236–2246. Melbourne, Australia: Association for Computational Linguistics.

Busso, C.; Bulut, M.; Lee, C.-C.; Kazemzadeh, A.; Mower, E.; Kim, S.; Chang, J. N.; Lee, S.; and Narayanan, S. S. 2008. IEMOCAP: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42(4): 335–359.

Chi, P.-H.; Chung, P.-H.; Wu, T.-H.; Hsieh, C.-C.; Chen, Y.-H.; Li, S.-W.; and Lee, H.-y. 2021. Audio albert: A lite bert for self-supervised learning of audio representation. In *2021 IEEE Spoken Language Technology Workshop (SLT)*, 344–350. IEEE.

Chuang, Y.-S.; Liu, C.-L.; Lee, H.-Y.; and Lee, L.-s. 2019. Speechbert: An audio-and-text jointly learned language model for end-to-end spoken question answering. *ArXiv preprint*, abs/1910.11559.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 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, Volume 1 (Long and Short Papers)*, 4171–4186. Minneapolis, Minnesota: Association for Computational Linguistics.

He, D.; Xia, Y.; Qin, T.; Wang, L.; Yu, N.; Liu, T.; and Ma, W. 2016. Dual Learning for Machine Translation. In Lee, D. D.; Sugiyama, M.; von Luxburg, U.; Guyon, I.; and Garnett, R., eds., *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, 820–828.

Jung, J.-w.; Heo, H.-S.; Kim, J.-h.; Shim, H.-j.; and Yu, H.-J. 2019. Rawnet: Advanced end-to-end deep neural network using raw waveforms for text-independent speaker verification. *ArXiv preprint*, abs/1904.08104.

Kingma, D. P.; and Ba, J. 2015. Adam: A Method for Stochastic Optimization. In Bengio, Y.; and LeCun, Y., eds., *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Lample, G.; Conneau, A.; Denoyer, L.; and Ranzato, M. 2018a. Unsupervised Machine Translation Using Monolingual Corpora Only. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.

Lample, G.; Ott, M.; Conneau, A.; Denoyer, L.; and Ranzato, M. 2018b. Phrase-Based & Neural Unsupervised Machine Translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 5039–5049. Brussels, Belgium: Association for Computational Linguistics.

Li, G.; Duan, N.; Fang, Y.; Gong, M.; and Jiang, D. 2020. Unicoder-VL: A Universal Encoder for Vision and Language by Cross-Modal Pre-Training. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, 11336–11344. AAAI Press.

Li, H.; Kang, Y.; Liu, T.; Ding, W.; and Liu, Z. 2021. CTAL: Pre-training Cross-modal Transformer for Audio-and-Language Representations. arXiv:2109.00181.

Li, L. H.; Yatskar, M.; Yin, D.; Hsieh, C.-J.; and Chang, K.-W. 2019. Visualbert: A simple and performant baseline for vision and language. *ArXiv preprint*, abs/1908.03557.

Liu, A. T.; Yang, S.; Chi, P.; Hsu, P.; and Lee, H. 2020. Mockingjay: Unsupervised Speech Representation Learning with Deep Bidirectional Transformer Encoders. In *2020 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2020, Barcelona, Spain, May 4-8, 2020*, 6419–6423. IEEE.

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. Roberta: A robustly optimized bert pretraining approach. *ArXiv preprint*, abs/1907.11692.

Lu, J.; Batra, D.; Parikh, D.; and Lee, S. 2019. ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks. In Wallach, H. M.; Larochelle, H.; Beygelzimer, A.; d'Alché-Buc, F.; Fox, E. B.; and Garnett, R., eds., *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, 13–23.

Panayotov, V.; Chen, G.; Povey, D.; and Khudanpur, S. 2015. Librispeech: An ASR corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2015, South Brisbane, Queensland, Australia, April 19-24, 2015*, 5206–5210. IEEE.

Ren, Y.; Tan, X.; Qin, T.; Zhao, S.; Zhao, Z.; and Liu, T. 2019. Almost Unsupervised Text to Speech and Automatic Speech Recognition. In Chaudhuri, K.; and Salakhutdinov, R., eds., *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, 5410–5419. PMLR.

Schneider, S.; Baevski, A.; Collobert, R.; and Auli, M. 2019. wav2vec: Unsupervised Pre-Training for Speech Recognition. In *Interspeech 2019, 20th Annual Conference of the International Speech Communication Association*, 3465–3469.

Sennrich, R.; Haddow, B.; and Birch, A. 2016. Improving Neural Machine Translation Models with Monolingual Data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 86–96. Berlin, Germany: Association for Computational Linguistics.

Su, W.; Zhu, X.; Cao, Y.; Li, B.; Lu, L.; Wei, F.; and Dai, J. 2020. VL-BERT: Pre-training of Generic Visual-Linguistic Representations. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Tan, H.; and Bansal, M. 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 (EMNLP-IJCNLP)*, 5100–5111. Hong Kong, China: Association for Computational Linguistics.

Tjandra, A.; Sakti, S.; and Nakamura, S. 2017. Listening while speaking: Speech chain by deep learning. In *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 301–308. IEEE.

Tsai, Y.-H. H.; Bai, S.; Liang, P. P.; Kolter, J. Z.; Morency, L.-P.; and Salakhutdinov, R. 2019. Multimodal Transformer for Unaligned Multimodal Language Sequences. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 6558–6569. Florence, Italy: Association for Computational Linguistics.

Vincent, P.; Larochelle, H.; Bengio, Y.; and Manzagol, P. 2008. Extracting and composing robust features with denoising autoencoders. In Cohen, W. W.; McCallum, A.; and Roweis, S. T., eds., *Machine Learning, Proceedings of the Twenty-Fifth International Conference (ICML 2008), Helsinki, Finland, June 5-9, 2008*, volume 307 of *ACM International Conference Proceeding Series*, 1096–1103. ACM.

Wan, L.; Wang, Q.; Papir, A.; and Lopez-Moreno, I. 2018. Generalized End-to-End Loss for Speaker Verification. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2018, Calgary, AB, Canada, April 15-20, 2018*, 4879–4883. IEEE.

Xu, H.; Zhang, H.; Han, K.; Wang, Y.; Peng, Y.; and Li, X. 2019. Learning alignment for multimodal emotion recognition from speech. *ArXiv preprint*, abs/1909.05645.

Xu, J.; Tan, X.; Ren, Y.; Qin, T.; Li, J.; Zhao, S.; and Liu, T. 2020. LRSpeech: Extremely Low-Resource Speech Synthesis and Recognition. In Gupta, R.; Liu, Y.; Tang, J.; and Prakash, B. A., eds., *KDD '20: The 26th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Virtual Event, CA, USA, August 23-27, 2020*, 2802–2812. ACM.

Yoon, S.; Byun, S.; Dey, S.; and Jung, K. 2019. Speech Emotion Recognition Using Multi-hop Attention Mechanism. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2019, Brighton, United Kingdom, May 12-17, 2019*, 2822–2826. IEEE.

Yoon, S.; Byun, S.; and Jung, K. 2018. Multimodal speech emotion recognition using audio and text. In *2018 IEEE Spoken Language Technology Workshop (SLT)*, 112–118. IEEE.

Zadeh, A.; Liang, P. P.; Poria, S.; Vij, P.; Cambria, E.; and Morency, L. 2018. Multi-attention Recurrent Network for Human Communication Comprehension. In McIlraith, S. A.; and Weinberger, K. Q., eds., *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, 5642–5649. AAAI Press.