Audio-Oriented Multimodal Machine Comprehension via Dynamic Inter- and Intra-modality Attention

Zhiqi Huang[1*], Fenglin Liu[1*], Xian Wu[2], Shen Ge[2], Helin Wang[1], Wei Fan[2], Yuexian Zou[1,3†]

[1] ADSPLAB, School of ECE, Peking University, China
[2] Tencent, China
[3] Peng Cheng Laboratory, China
{zhiqihuang, fenglinliu98, wanghl15, zouyx}@pku.edu.cn, {kevinxwu, shenge, davidwfan}@tencent.com

Abstract

While Machine Comprehension (MC) has attracted extensive research interests in recent years, existing approaches mainly belong to the category of Machine Reading Comprehension task which mines textual inputs (paragraphs and questions) to predict the answers (choices or text spans). However, there are a lot of MC tasks that accept audio input in addition to the textual input, e.g. English listening comprehension test. In this paper, we target the problem of Audio-Oriented Multimodal Machine Comprehension, and its goal is to answer questions based on the given audio and textual information. To solve this problem, we propose a Dynamic Inter- and Intra-modality Attention (DIIA) model to effectively fuse the two modalities (audio and textual). DIIA can work as an independent component and thus be easily integrated into existing MC models. Moreover, we further develop a Multimodal Knowledge Distillation (MKD) module to enable our multimodal MC model to accurately predict the answers based only on either the text or the audio. As a result, the proposed approach can handle various tasks including: Audio-Oriented Multimodal Machine Comprehension, Machine Reading Comprehension and Machine Listening Comprehension, in a single model, making fair comparisons possible between our model and the existing unimodal MC models. Experimental results and analysis prove the effectiveness of the proposed approaches. First, the proposed DIIA boosts the baseline models by up to 21.08% in terms of accuracy; Second, under the unimodal scenarios, the MKD module allows our multimodal MC model to significantly outperform the unimodal models by up to 18.87%, which are trained and tested with only audio or textual data.

Introduction

Recently, there is a surge of research interests in Machine Comprehension (MC), which aims to teach the machine to answer questions after giving comprehension materials (Nguyen et al. 2016; Rajpurkar et al. 2016; Lai et al. 2017). As shown in Figure 1, conventional MC system accepts unimodal textual inputs, and predict the corresponding answers to the given multiple-choice questions. By adopting various deep learning techniques, many models have been developed for the MC problem and are proven to be effective (Liu et al. 2019c; Qiu et al. 2019).

However, conventional MC only focus on accepting single modal textual inputs, while in real life many multimodal (audio and textual modalities) scenarios exist, such as playing music with lyrics, and taking a listening examination in TOEFL, etc. Moreover, multimodal inputs often convey more information than single modality inputs, and it is easy to make wrong judgments under single modal scenarios. For example, people could express opposite intentions by using different tones to say the sentence "that's interesting". If the emphasis is put on the word "interesting!", he/she may really be interested and wants to know more; on the other hand, if the whole sentence is expressed intermittently, e.g., "that's... um... interesting.", he/she may not be interested at all. Another example in English exams like TOEFL, i.e., "He hasn't seen his parent four years!" and "He hasn't seen his parent for years!", may only be distinguished by their different sound emphasis, showing a clear trap in the audio. As a result, different tones of voice could lead to different meanings, even when the textual sentences are nearly identical. Inspired by these real-world applications and observations, we propose the novel problem of *Audio-Oriented Multimodal Machine Comprehension*. As shown in Figure 1, the novel problem requires the system to consider both the audio and textual modality inputs in selecting the correct answers. Compared to the conventional setting of unimodal based MC, our new problem poses two fundamental challenges. First, in the learning of multimodal tasks, due to the great disparities between the textual and the audio domains, plus the distinct features of the modalities, one core challenge is to effectively bridge the gap between textual and the audio domains and learn an effective fusion of multimodality features (Lu et al. 2016; Gao et al. 2019; Liu et al. 2019a, 2020b). Second, due to the abundance of unimodal (textual or audio) real-world application scenarios, e.g., the conventional machine reading comprehension (textual) (Lai et al. 2017) and the end-to-end spoken language understanding (audio) (Serdyuk et al. 2018), it is necessary to enable the multimodal MC model to work in the unimodal scenarios. In other words, we should empower the proposed multimodal MC model with the capability to accurately predict the answers based only on the textual or only on audio input.

To tackle the first challenge, we propose a novel Dy-

[*] Equal Contribution.
[†] Corresponding Author.

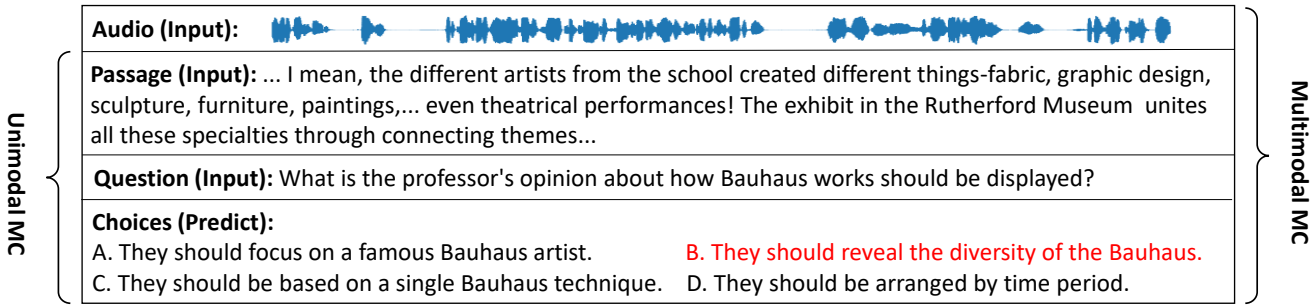

Figure 1: Comparison between the conventional unimodal machine comprehension and the proposed audio-oriented multimodal machine comprehension. The red colored text indicate the ground-truth answer.

namic Inter- and Intra-modality Attention (DIIA) model to better capture the high-level interactions between audio and textual features, resulting in an efficient multimodality feature fusion to answer the questions accurately. As shown in Figure 2, the DIIA integrates the self-attention and co-attention to learn the inter- and intra-relationships between audio and textual modalities in an effective manner (Liu et al. 2019a,b, 2020a). The core intuition behind our motivation is that, each textual word should obtain information not only from its associated audio information but also from related words/phrases to infer the answer to the question, and so do audio. To tackle the second challenge, based on our proposed DIIA model, we further develop a Multimodal Knowledge Distillation (MKD) module, which transfers representative knowledge from multimodal to either textual or audio modality. In implementation, as shown in Figure 2, the learned multimodal representations from our pre-trained multimodal MC model, i.e., DIIA, are used to guide the learning of unimodal representations. As a result, the MKD module associates the multimodal knowledge hidden behind the fused multimodal features to facilitate the understanding of unimodal information. The design of the proposed MKD allows our approach to be applied to scenarios where only single modal data is available. In other words, our approach can accurately answer the questions based only on input audio or input text, so that our approach can be used for fair comparisons with existing textual based MC models.

Moreover, to better handle this problem, we also collect two audio-oriented multimodal machine comprehension datasets, i.e., L-TOEFL and CET, from English listening tests, which contain questions and answers in the form of text, as well as the comprehension passages in both textual and audio modalities. The extensive experiments and analysis on the proposed L-TOEFL and CET datasets validate our arguments and prove the effectiveness of our approach.

Overall, our main contributions are as follows:

- We propose the audio-oriented multimodal machine comprehension task, which requires the system to understand both input audio and textual information together, rather than only use textual information in previous works. We also assemble two audio-oriented multimodal machine comprehension datasets (L-TOEFL and CET) for the task.
- A novel Dynamic Inter- and Intra-modality Attention (DIIA) model is proposed to obtain multimodality fusion by interleaving inter- and intra-modality feature-level fusion. Such a framework provides a solid bias for the audio-oriented multimodal machine comprehension task.
- We further develop the Multimodal Knowledge Distillation (MKD) module, which can transfer representative knowledge from multimodal to either textual modal or audio modal, enabling a single multimodal MC model to accurately predict the answers based only on the text or audio. As a result, our model can handle various tasks at the same time with a single model.
- The extensive experiments show consistent performance gains achieved by the proposed novel DIIA over baseline systems. The experiments also show that the proposed MKD module enables the multimodal MC model to be applied in the unimodal scenarios and outperform the conventional unimodal models significantly.

Related Work

The related works are introduced from two aspects: 1) Multi-Choices Machine Comprehension and 2) Machine Comprehension of Spoken Content.

Multi-Choices Machine Comprehension

There are four types of Machine Comprehension (MC) (Chen 2018), including cloze-style (Hermann et al. 2015), multi-choices (Lai et al. 2017), span extraction (Rajpurkar et al. 2016; Rajpurkar, Jia, and Liang 2018), and generative format (Kociský et al. 2018). In this paper, we focus on multi-choices machine comprehension: the goal is to find the only correct answer in the multiple (usually 4) choices based on the given inputs, i.e., passage, question, and multiple choices.

Machine Comprehension of Spoken Content

Tseng et al. (2016) proposed to deal with the MC task of spoken content by first employ an ASR model (Yu and Li 2017) to recognize speech into text, then, an MC model is designed to process the ASR transcriptions for selecting the correct answer out of 4 choices. Although such method take the spoken content into consideration, the system is still a text-based MC system. Chuang et al. (2020); Kuo, Luo, and

Chen (2020) studied the end-to-end spoken question answering problem by introducing a pre-trained modality fusion model learned from audio and text, however, in the inference phase, they need input with both modalities while ours can predict with only audio or only text input. And there are also some related work studied different tasks (Yang et al. 2003) or introduced auxiliary tool (Zhang et al. 2018).

Approach

In this section, we first formulate the conventional machine comprehension (MC) and the proposed multimodal machine comprehension problems; then we describe the proposed approach in detail. Specifically, to better capture the high-level interactions between audio and textual features, and generate efficient multimodality feature fusion to accurately answer questions, we propose the novel Dynamic Inter- and Intra-modality Attention (DIIA) model. Based on our DIIA model, we further propose the Multimodal Knowledge Distillation (MKD) to enable our multimodal MC model to accurately predict the answers based only on the text or audio. Figure 2 illustrates the proposed approach.

Problem Formulation

In this section, we formulate the problems of conventional multi-choices machine comprehension and the proposed multi-choices multimodal machine comprehension.

Problem Formulation of Conventional Multi-Choices Machine Comprehension Task Taking a passage $\mathbf{P}_\mathrm{I}$ as input, the goal of conventional multi-choices machine comprehension is to predict the correct choice $\mathbf{C}_\text{ans}$ based on the given questions $\mathbf{Q}$ and candidate choices $\mathbf{C}_\text{candidate}$. Some well-performing frameworks (Wang et al. 2018; Dhingra et al. 2017; Devlin et al. 2019) normally include a text encoder and an answer predictor, which can be formulated as:

$$\text{Text Encoder} : \mathbf{P}_\mathrm{I} \rightarrow \mathbf{P}$$
$$\text{Answer Predictor} : \mathbf{P}, \mathbf{Q}, \mathbf{C}_\text{candidate} \rightarrow \mathbf{C}_\text{ans}$$

The text encoder aims to generate the textual features $\mathbf{P}$ of the input passage $\mathbf{P}_\mathrm{I}$. In implementation, given an input sequence with length N, the text features are usually generated by 300-dimensional word embeddings with GloVe vectors (Pennington, Socher, and Manning 2014), and represented as $\mathbf{P} \in \mathbb{R}^{N \times d}$ ($d = 300$). The answer predictor, e.g., co-matching (Wang et al. 2018), is used to predict the correct choice $\mathbf{C}_\text{ans}$ from the given questions $\mathbf{Q}$, the candidate choices $\mathbf{C}_\text{candidate}$ and $\mathbf{P}$. Given the ground truth choice, we can simply train the framework by minimizing training loss, e.g., cross-entropy loss.

Problem Formulation of Multimodal Multi-Choices Machine Comprehension Task The main difference between the audio-oriented multimodal machine comprehension and conventional machine comprehension is the different available information that can be used to predict the correct answer. Specifically, for audio-oriented multimodal machine comprehension task, it requires the system to further consider the audio information $\mathbf{A}_\mathrm{I}$ when selecting the correct answer, which can be formulated as:

$$\text{Text Encoder} : \mathbf{P}_\mathrm{I} \rightarrow \mathbf{P}$$
$$\text{Audio Encoder} : \mathbf{A}_\mathrm{I} \rightarrow \mathbf{A}$$
$$\text{Answer Predictor} : \mathbf{P}, \mathbf{A}, \mathbf{Q}, \mathbf{C}_\text{candidate} \rightarrow \mathbf{C}_\text{ans}$$

where $\mathbf{A}$ denotes the extracted audio features from the input audio $\mathbf{A}_\mathrm{I}$. In implementation, we adopt VGGish (Hershey et al. 2017) pre-trained on AudioSet (Gemmeke et al. 2017) to extract the audio features, represented as $\mathbf{A} \in \mathbb{R}^{M \times d}$.

Dynamic Inter- and Intra-modality Attention

Our Audio-Oriented Multimodal Machine Comprehension task requires the MC system to understand both input audio and textual information, thus learning fine-grained joint representations of audio and text are of paramount importance. In other words, it is vital to learn the alignments and relationships between audio and textual modalities. To this end, as shown in Figure 2, we propose the Dynamic Inter- and Intra-modality Attention (DIIA) to effectively fuse the audio and textual features before predicting answers. In particular, inspired by the success of Multi-Head Attention (MHA) (Vaswani et al. 2017), we refer to the MHA mechanism and propose the DIIA model, which consists of an Inter-modality Attention module and an Intra-modality Attention module, to learn the inter- and intra-relationships of audio and textual modalities in an effective manner.

Multi-Head Attention In order to extract the relationship between the intra-modality and inter-modality of audio features and textual features, we adopt the Multi-Head Attention (MHA) (Vaswani et al. 2017), which compute the association weights between different features. The attention mechanism allows probabilistic many-to-many relations instead of monotonic relations, as in Xu et al. (2015); Liu et al. (2019a). The following MHA consists of n parallel heads and each head is represented as scaled dot-product attention.

$$\text{Att}(\mathbf{Q},\mathbf{K},\mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathrm{W}_\mathrm{Q}(\mathbf{K}\mathrm{W}_\mathrm{K})^T}{\sqrt{d_k}}\right)\mathbf{V}\mathrm{W}_\mathrm{V}$$

where $\mathbf{Q} \in \mathbb{R}^{l \times d}$, $\mathbf{K} \in \mathbb{R}^{k \times d}$ and $\mathbf{V} \in \mathbb{R}^{k \times d}$ represent respectively the query matrix, the key matrix and the value matrix; The l and k denote the length of the query and key/value, respectively; $\mathrm{W}_\mathrm{Q}, \mathrm{W}_\mathrm{K}, \mathrm{W}_\mathrm{V} \in \mathbb{R}^{d \times d_k}$ are the learnable parameters of linear transformations and $d_k = d/n$ is the scaling factor, where n is the number of heads.

Following the multi-head attention is a fully-connected Feed-Forward Network (FFN), which is defined as follows:

$$\text{FFN}(\mathbf{x}) = \text{ReLU}(\mathbf{x}\mathrm{W}_\mathrm{f} + b_\mathrm{f})\mathrm{W}_\mathrm{ff} + b_\mathrm{ff}$$

where W_f and W_ff denote matrices for linear transformation; b_f and b_ff represent the bias terms. Each sub-layer, i.e., MHA and FFN, is followed by an operation sequence[1] of dropout (Srivastava et al. 2014), shortcut connection (He et al. 2016), and layer normalization (Ba, Kiros, and Hinton 2016).

We take advantage of the MHA to implement the idea of learning the inter- and intra-relationships of audio and textual modalities.

[1]For conciseness, the operation sequence in this paper is omitted.

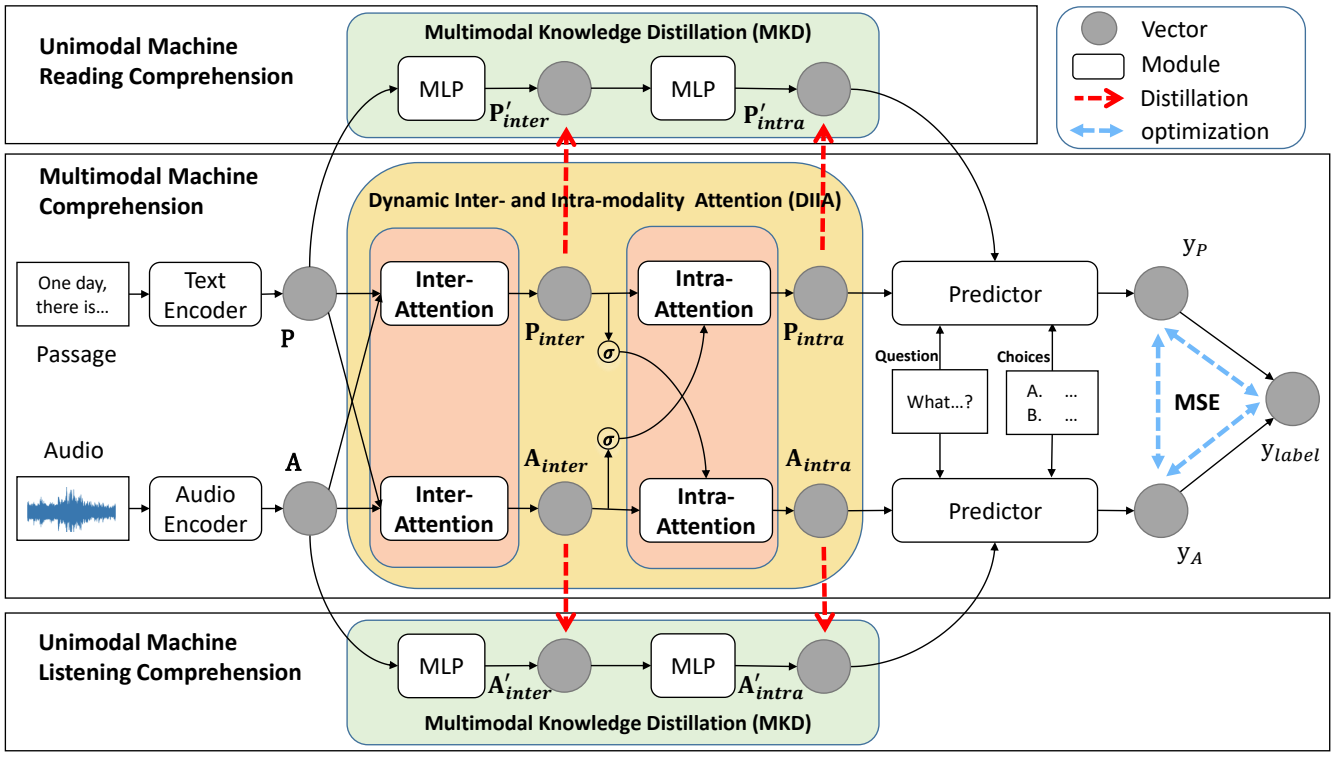

Figure 2: The architecture of our proposed approach. The DIIA model consists of the inter-modality attention module and the intra-modality attention module, aiming to capture the correlation and build the relationship between audio and textual modalities. The blue dashed lines represent the MSE loss between $\mathbf{y}_A$, $\mathbf{y}_P$ (the training label) and $\mathbf{y}_{label}$ (the ground truth label). The red dashed lines represent the hidden states based distillation. The predictor is adapted from the existing machine comprehension models, such as Co-Matching (Wang et al. 2018).

Inter-modality Attention 1) To represent textual features $\mathbf{P}$ with high quality, we need to find the most relevant audio descriptions $\mathbf{A}$ to identify the direct relations between audio and text. 2) Similarly, we need to find the most relevant textual features $\mathbf{P}$ to summarize the properties of the audio features $\mathbf{A}$.

According to the attention theorem (Vaswani et al. 2017), taking the first situation as example, the textual features $\mathbf{P} \in \mathbb{R}^{N\times d}$ serve as query, and the audio features $\mathbf{A} \in \mathbb{R}^{M\times d}$ server as key and value. Consequently, the result $\mathbf{A}_{\text{inter}} \in \mathbb{R}^{M\times d}$ turns out to be a set of attended audio features for textual features:

$$\mathbf{A}_{\text{inter}} = \text{FFN}(\text{MHA}(\mathbf{A}, \mathbf{P}, \mathbf{P}))$$

Similarly, the $\mathbf{P}_{\text{inter}} \in \mathbb{R}^{N\times d}$ can be computed as follow:

$$\mathbf{P}_{\text{inter}} = \text{FFN}(\text{MHA}(\mathbf{P}, \mathbf{A}, \mathbf{A}))$$

Now we can assume that the relation between the audio and textual features are built and represent the two updated features as $\mathbf{A}_{\text{inter}} \in \mathbb{R}^{M\times d}$ and $\mathbf{P}_{\text{inter}} \in \mathbb{R}^{N\times d}$.

Intra-modality Attention After the inter-modality attention module, the cross-modal relations between audio and text have been modeled. However, information from different modalities may have varying predictive power and noise. We argue that modeling relationships in a single modality can make up for this deficiency. The intra-modality attention module explores how the knowledge learned from two modalities can be fused in an appropriate way to help the training of the multimodal MC model. We adapt the following formula to learn the intra-relationships of audio and textual features:

$$\mathbf{A}_{\text{intra}} = \text{FFN}(\text{MHA}(\hat{\mathbf{A}}_{\text{inter}}, \hat{\mathbf{A}}_{\text{inter}}, \mathbf{A}_{\text{inter}}))$$
$$\mathbf{P}_{\text{intra}} = \text{FFN}(\text{MHA}(\hat{\mathbf{P}}_{\text{inter}}, \hat{\mathbf{P}}_{\text{inter}}, \mathbf{P}_{\text{inter}}))$$

It is worth noticing that to better promote the learning of intra-relationships (Hu et al. 2020; Liu et al. 2018), we further design a conditional gate operation $\mathbf{G}$ to update the queries and keys. The process is defined as follows:

$$\hat{\mathbf{A}}_{\text{inter}} = (1 + \mathbf{G}_{\mathbf{P}}) \odot \mathbf{A}_{\text{inter}}$$
$$\hat{\mathbf{P}}_{\text{inter}} = (1 + \mathbf{G}_{\mathbf{A}}) \odot \mathbf{P}_{\text{inter}}$$

where $\odot$ represents the element-wise multiplication. The conditional gate operation $\mathbf{G}_{\mathbf{P}}$ and $\mathbf{G}_{\mathbf{A}}$ are defined as:

$$\mathbf{G}_{\mathbf{A}} = \sigma\left(\text{Avg\_pool}(\mathbf{A})\mathbf{W}_{\mathbf{A}}\right)$$
$$\mathbf{G}_{\mathbf{P}} = \sigma\left(\text{Avg\_pool}(\mathbf{P})\mathbf{W}_{\mathbf{P}}\right)$$

where the σ and Avg_pool denote the sigmoid function and average pooling, respectively.

Through the formula, in the audio domain, the intra-modality attention learns salient audio groupings and integrates naturally related audio information. In the textual domain, it learns text collocations and has the ability to consider associations and collocations of sentences in the passage during answer predicting. The learned intra-relationships of audio and textual features are super beneficial for multimodal machine comprehension task.

Multimodal Knowledge Distillation The intuition of enabling a model to accurately predict the answer based only on either the text or the audio is that conventionally the model is only allowed to accept a single modality as input. However, due to the fact that the proposed DIIA have learned the fine-grained multimodal representations, inspired by the knowledge distillation technique (Romero et al. 2015), we further introduce the Multimodal Knowledge Distillation Module (MKD) to distill the representative knowledge from multimodal learnt by DIIA to either textual modal or audio modal to enhance the input features with single modality only. In particular, the MKD consists of two Multi-Layer Perceptrons (MLPs). We represent the output of the MLPs as $\mathbf{A}^{'}_{\text{inter}}$, $\mathbf{A}^{'}_{\text{intra}}$ for audio distillation block and $\mathbf{P}^{'}_{\text{inter}}$, $\mathbf{P}^{'}_{\text{intra}}$ for passage distillation block, which are defined as:

$$\mathbf{A}^{'}_{\text{inter}} = \text{MLP}(\mathbf{A}); \quad \mathbf{A}^{'}_{\text{intra}} = \text{MLP}(\mathbf{A}^{'}_{\text{inter}})$$
$$\mathbf{P}^{'}_{\text{inter}} = \text{MLP}(\mathbf{P}); \quad \mathbf{P}^{'}_{\text{intra}} = \text{MLP}(\mathbf{P}^{'}_{\text{inter}})$$

Then we apply the following formulas to distill the knowledge from the output of DIIA to the MKD, which can be computed with mean squared error (MSE) loss and be represented as:

$$\mathcal{L}_{\text{MKD}_A} = \text{MSE}(\mathbf{A}_{\text{inter}}, \mathbf{A}^{'}_{\text{inter}}) + \text{MSE}(\mathbf{A}_{\text{intra}}, \mathbf{A}^{'}_{\text{intra}})$$
$$\mathcal{L}_{\text{MKD}_P} = \text{MSE}(\mathbf{P}_{\text{inter}}, \mathbf{P}^{'}_{\text{inter}}) + \text{MSE}(\mathbf{P}_{\text{intra}}, \mathbf{P}^{'}_{\text{intra}})$$

In this way, with the help of MKD, we can use only one modal data input in the answer prediction process, while implicitly use the interaction information of the two modalities to enhance the unimodal representations.

Implementation

From the above process, the proposed DIIA model focuses on learning the relationships between audio and textual features to obtain the multimodal representations, and the MKD module can distill the multimodal knowledge learned from the DIIA. In this section, we describe the training process detail of our approach by introducing three applied problems, i.e., Multimodal Machine Comprehension, Unimodal Machine Reading Comprehension and Unimodal Machine Listening Comprehension.

Multimodal Machine Comprehension As shown in Figure 2, the outputs of the predictors of multimodal machine comprehension problem are $\mathbf{y}_A$ and $\mathbf{y}_P$ for the audio and textual features, respectively. Given the ground truth $\mathbf{y}_{label}$, we adopt the Cross-Entropy (CE) loss function to optimize our multimodal MC model, including the DIIA. The optimization is defined as:

$$\ell_{pred}(\mathbf{y}_A, \mathbf{y}_{label}) = \text{CE}(\mathbf{y}_A, \mathbf{y}_{label})$$
$$\ell_{pred}(\mathbf{y}_P, \mathbf{y}_{label}) = \text{CE}(\mathbf{y}_P, \mathbf{y}_{label})$$

Specifically, for better optimization, we further distill the knowledge loss on the logits to narrow the distance between the audio logits and the textual logits through MSE loss:

$$\ell_{pred}(\mathbf{y}_A, \mathbf{y}_P) = \text{MSE}(\mathbf{y}_A, \mathbf{y}_P)$$

Overall, the final objective loss function is computed as:

$$\mathcal{L}_1 = \ell_{pred}(\mathbf{y}_A, \mathbf{y}_{label}) + \ell_{pred}(\mathbf{y}_P, \mathbf{y}_{label}) + \ell_{pred}(\mathbf{y}_A, \mathbf{y}_P)$$

At the testing stage, the input $\mathbf{P}$ and $\mathbf{A}$ are sent to the DIIA to obtain the $\mathbf{P}_{\text{intra}}$ and $\mathbf{A}_{\text{intra}}$, then sent to the predictor to obtain the $\mathbf{y}_A$ and $\mathbf{y}_P$. Finally, we add the two predicted logits as the final predicted logits to predict the answer.

Unimodal Machine Reading Comprehension For practical use, we further propose to enable our multimodal MC model to accurately predict the answers based only on the text, which means that we only use unimodal textual information in inference. Because the unimodal scenario requires no modal interaction, we remove the DIIA, instead, we introduce an MKD module to transfer the representative multimodal knowledge learned from our DIIA to the textual modality (see Figure 2). Specifically, at the training stage, we first directly adopt the pre-trained multimodal MC model. Next, we freeze the parameters of the text encoder, DIIA and predictor, and use the proposed $\mathcal{L}_{\text{MKD}_P}$ to train the MKD. At the testing stage, we obtain the input passage features of the predictor $\hat{\mathbf{P}}_{\text{inter}}$ through the MKD module, and then output the predicted logits $\mathbf{y}_P$.

In this way, our model can work on unimodal input in practice with multimodal information being introduced during the distillation training process. So that our model can be compared fairly with conventional MC models.

Unimodal Machine Listening Comprehension Similar to the Unimodal Machine Reading Comprehension scenario, we train the MKD using $\mathcal{L}_{\text{MKD}_A}$ with the pre-trained multimodal MC model froze. And generate the predicted logits $\mathbf{y}_A$ with the trained MKD at the test stage.

Experiments

We describe the collected datasets and the training details, followed by the evaluation of the proposed approach.

Datasets

In this paper, to address the audio-oriented multimodal MC problem, we propose the L-TOEFL and CET datasets, where L-TOEFL is collected from the TOEFL Educational Testing Service (TOEFL ETS)[2], which is an English ability test designed to measure the ability to listen for basic comprehension, pragmatic understanding and synthesizing information, and CET is collected from the College English Test (CET)[3], a national English as a foreign language test in the People's Republic of China. Designed by educational experts, L-TOEFL and CET datasets aim to test non-native English speakers using various types of complicated questions, and the passages are divided into types of conversation and lecture. Specifically, L-TOEFL and CET are collected from a total of 106 official examinations, with each

[2]https://www.ets.org/toefl/

[3]http://cet.neea.edu.cn/

Modalities		Datasets	L-TOEFL						CET					
		Methods	GAReader		Co-Matching		DCMN		GAReader		Co-Matching		DCMN	
		Data Partitions	Dev	Test	Dev	Test	Dev	Test	Dev	Test	Dev	Test	Dev	Test
Unimodal	Audio	Conventional	31.29	30.63	34.55	33.01	43.42	41.61	34.11	33.82	39.96	39.44	52.86	51.61
		w/ MKD	43.76	43.31	48.34	48.09	62.40	60.51	47.11	46.44	52.66	52.50	65.39	63.99
	Text	Conventional	36.35	36.19	38.26	38.75	46.85	46.77	41.73	41.22	47.73	47.90	60.07	58.68
		w/ MKD	44.69	43.60	49.98	49.20	64.33	62.68	49.04	47.98	55.31	55.00	67.63	66.18
Multimodal		Shallow-fusion	46.28	45.19	50.68	50.75	63.22	62.69	51.33	48.62	57.81	56.11	66.52	64.39
		w/ Inter-	47.15	46.11	52.40	51.34	64.01	63.77	53.13	50.74	58.03	56.50	68.94	67.80
		w/ Intra-	45.98	45.60	51.93	51.14	63.79	62.92	52.05	50.22	57.85	56.49	67.35	66.01
		w/ DIIA	**48.36**	**47.78**	**53.22**	**52.03**	**65.94**	**63.68**	**54.20**	**51.12**	**59.01**	**57.83**	**69.13**	**68.09**

Table 1: Performance (Accuracy (%)) on the proposed L-TOEFL and CET datasets. Multimodal and Unimodal represent the input modalities we use for the models, i.e., audio and text, audio only, or text only. Conventional means the conventional models that only employ the predictor and the feature encoder. "w/ MKD" means that we employ our proposed MKD module as well as our training method in the unimodal setting. Shallow fusion means directly add the two unimodal representations for prediction and bypass the DIIA module.

official examination contains 1 to 6 passage(s) with corresponding audio, each passage contains 1 to 6 question(s), and each question is accompanied with 4 choices. Thus, our datasets consist of 4-attributes pair: {audio, passage, question, 4 answer choices with the correct one}. After deleting some improper pairs, such as multiple answers (more than one correct answer), etc., we get a total of 2,200 such A-P-Q-C pairs. We randomly divide the L-TOEFL and CET datasets into 1000/162/162 and 657/110/109 examples as for train/dev/test data partitioning, respectively, following ratios of 0.75/0.125/0.125. The amount of collected English listening test data set is larger than the one used in Tseng et al. (2016) (963 examples).

Experimental Settings

The encoder takes the audio features and text features as input. In implementation, we adopt the VGGish (Hershey et al. 2017) pre-trained on AudioSet (Gemmeke et al. 2017) to extract the audio features denoted as $\mathbf{A} \in \mathbb{R}^{M \times 300}$. First, we resample the raw audio file to the rate assumed by VGGish, then generate a 128-dimensional embedding of each AudioSet segment. After that, we employ a linear transformation to map the dimension from 128 to 300. Thus, the feature dimension of each audio is $frame \times 300$ where the $frame$ range from 109 to 469. Given an input sequence with length N, the text features are initialized by 300-dimensional word embeddings with GloVe vectors (Pennington, Socher, and Manning 2014) denoted as $\mathbf{W} \in \mathbb{R}^{N \times 300}$. Considering the actual length of the datasets, we set the maximum length M and N as 384 empirically. We adopt the Adam optimizer for optimizing the parameters, with a mini-batch size of 12 and initial learning rate of 0.001. After training 100 epochs, we select the model which works the best on the dev set, and then evaluate it on the test set in terms of accuracy (%).

Experimental Results

In this section, we will present our evaluation of the unimodal MC models, i.e., conventional Machine Reading Comprehension model, Machine Listening Comprehension model, as well as the models with our proposed MKD module, and the multimodal MC model with our proposed DIIA model. We conduct experiments on three representative baseline MC systems, i.e., GA Reader (Dhingra et al. 2017), Co-Matching (Wang et al. 2018), and DCMN (Zhang et al. 2019). As can be seen in Table 1, under the unimodal setting, the models using text features outperform the models using audio features, which is due to the text is cleaner and easier to be processed by machines. Besides, our proposed MKD help the model distill multimodal knowledge in the unimodal scenarios, and can consequently achieve better performance than the conventional MC models, which verifies the effectiveness of our approach. From the last four rows of Table 1, we find that the fusion of multimodal features can significantly improve the performance of the model, which proves our motivation and effectiveness in proposing the audio-oriented MC task. In addition, compared to the shallow-fusion setting that simply add audio and textual features together, both the proposed Inter-modality Attention and Intra-modality Attention can achieve better performance by learning the inter- and intra-relationships between the two modalities, respectively, so as to obtain better multimodal representations. We also implement two unimodal models (audio and text) via averaging with Co-Matching as the predictor. The ensemble model acquires accuracy of 44.64 and 52.74 on the L-TOEFL and CET datasets, while the proposed model achieves 52.03 and 57.83 accuracy, therefore the proposed model consistently outperforms the ensemble of unimodal models.

Please note that our approach is not a replacement of existing approaches, but can be easily integrated into them to boost the performance, such as Zhang et al. (2020); Zhu, Zhao, and Li (2020).

Analysis

In this section, we analyze the effectiveness of the proposed DIIA and visualize the attention weights to show the advantage of the proposed approach in an intuitive manner. We also analyze the effectiveness of different passage types.

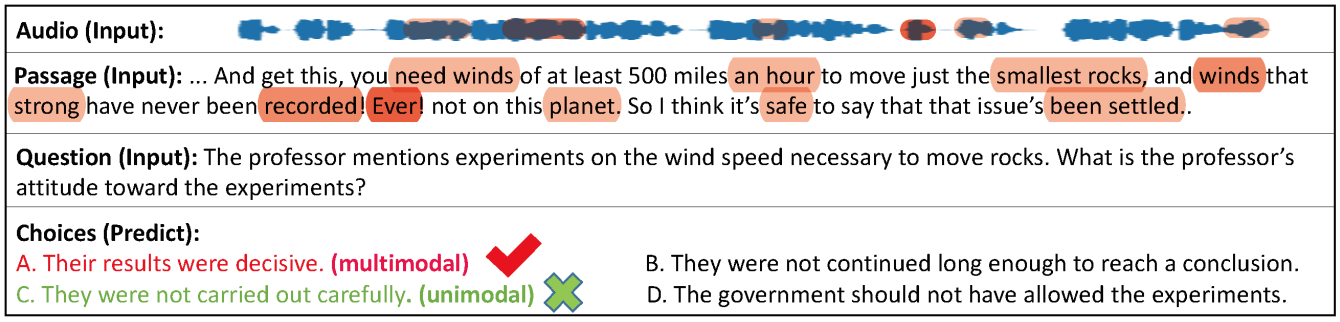

Figure 3: Visualization of the attention weights on the audio and passage in the inter-modality module of our proposed Dynamic Inter- and Intra-modality Attention. Each token's importance score is calculated by summing up the attention weight between this token and the tokens of other input modality. Darker color means higher weights.

Modalities	Methods	Con.	Δ	Lec.	Δ
Unimodal (Text)	Conventional	37.94	-	37.72	-
	w/ MKD	47.63	+9.69	46.37	+8.65
Multimodal	w/ DIIA	**52.63**	+14.69	**52.17**	+14.45

Table 2: Comparison between different type of passage and audio in the L-TOEFL dataset. Con. and Lec. stand for the Conversation passages and Lecture passages, respectively. Δ denotes the improvement over the model under the conventional setting. The models are explored with the Co-Matching (Wang et al. 2018) as the predictor.

Effect of the DIIA

To explore the effectiveness of the multimodal features and the correlations between audio and textual features learned by DIIA, we visualize the attention weights in the inter-modality attention module. As shown in Figure 3, we find that the unimodal MC system predicts the wrong answer choice 'C' while the multimodal MC system predicts the correct answer choice 'A'. To explore the reason for this difference, we visualize the attention weights in the inter-modality attention module, finding that by introducing the audio features and the DIIA model, the passage can put more attention on the key information including transitional, time and noun words, while the audio can extract more useful information, e.g., the tone information. The visualization of the attention weights on audio and textual information also verify our hypothesis and demonstrate the effectiveness of our approach. Since our model can capture the alignments and relationships between audio and textual modalities, the distribution of attention weights between audio and textual features is similar, which indicates that the textual features are properly enriched by the aligned audio features.

Types of Passage

L-TOEFL is composed of different types of passages, namely conversation and lecture, and they also correspond to different types of audio. Such difference may be exhibited in audio length, the number of characters or the overall audio emotions, etc. To better understand the improvement brought by audio information, we further explore the impacts of the passage and audio types on the MC system. Specifically, we conduct experiments for conversation and lecture passages on different model settings, i.e., the conventional unimodal model, the "w/ MKD" model, and the proposed multimodal model. We use the co-matching model (Wang et al. 2018) as the predictor. Table 2 shows that our approach brings greater improvements in passages of conversation type than passages of lecture type. We attribute this to the fact that the human mood and tone change more evidently in human conversation and are relatively smoother in lecture, which means the conversation audio can provide more useful information to the model and the multimodal MC system can predict the answer more precisely.

Conclusions

In this paper, we introduce the Audio-Oriented Multimodal Machine Comprehension task. To achieve this goal, we collect two datasets named L-TOEFL and CET, which consist of 1,324 and 876 audio-passage-question-choices pairs, respectively, and propose a Dynamic Inter- and Intra-modality Attention (DIIA) model, which consists of an inter-modality attention model and an intra-modality attention model. The DIIA model can learn the inter- and intra- relationships between the audio and the textual modalities. DIIA can work as an independent component and thus can be easily integrated into existing machine comprehension models. Furthermore, considering the abundance of unimodal (textual or audio) real-world application scenarios, we further develop a Multimodal Knowledge Distillation (MKD) module to enable our multimodal MC model to accurately predict the answers based only on either the text or the audio. The experimental results and the analysis demonstrates that audio input can improve the accuracy of machine comprehension models that solely relies on textual input. Specifically, the multimodal MC models achieve better results than the unimodal MC models, and the proposed DIIA model could fuse the audio and textual modalities effectively, thus boosts the baseline models by up to 21.08% in terms of accuracy; Furthermore, the MKD module allows our multimodal MC model (pretrain with both audio and textual input, predict with either audio or textual input) to outperform the unimodal models (train and predict with only audio or textual input) by up to 18.87% in terms of accuracy.

Acknowledgments

Special acknowledgments are given to AOTO-PKUSZ Joint Research Center for Artificial Intelligence on Scene Cognition Technology Innovation for its support. We thank all the anonymous reviewers for their constructive comments and suggestions.

References

Ba, L. J.; Kiros, R.; and Hinton, G. E. 2016. Layer Normalization. *arXiv preprint arXiv:1607.06450* .

Chen, D. 2018. Neural Reading Comprehension and Beyond. In *PhD thesis, Stanford University*.

Chuang, Y.; Liu, C.; Lee, H.; and Lee, L. 2020. SpeechBERT: An Audio-and-Text Jointly Learned Language Model for End-to-End Spoken Question Answering. In *INTERSPEECH*, 4168–4172.

Devlin, J.; Chang, M.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL-HLT*, 4171–4186.

Dhingra, B.; Liu, H.; Yang, Z.; Cohen, W. W.; and Salakhutdinov, R. 2017. Gated-Attention Readers for Text Comprehension. In *ACL*, 1832–1846.

Gao, P.; Jiang, Z.; You, H.; Lu, P.; Hoi, S. C. H.; Wang, X.; and Li, H. 2019. Dynamic Fusion With Intra- and Inter-Modality Attention Flow for Visual Question Answering. In *CVPR*, 6639–6648.

Gemmeke, J. F.; Ellis, D. P. W.; Freedman, D.; Jansen, A.; Lawrence, W.; Moore, R. C.; Plakal, M.; and Ritter, M. 2017. Audio Set: An ontology and human-labeled dataset for audio events. In *ICASSP*, 776–780.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep Residual Learning for Image Recognition. In *CVPR*, 770–778.

Hermann, K. M.; Kociský, T.; Grefenstette, E.; Espeholt, L.; Kay, W.; Suleyman, M.; and Blunsom, P. 2015. Teaching Machines to Read and Comprehend. In *NIPS*, 1693–1701.

Hershey, S.; Chaudhuri, S.; Ellis, D. P. W.; Gemmeke, J. F.; Jansen, A.; Moore, R. C.; Plakal, M.; Platt, D.; Saurous, R. A.; Seybold, B.; Slaney, M.; Weiss, R. J.; and Wilson, K. W. 2017. CNN architectures for large-scale audio classification. In *ICASSP*, 131–135.

Hu, J.; Shen, L.; Albanie, S.; Sun, G.; and Wu, E. 2020. Squeeze-and-Excitation Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* 42(8): 2011–2023.

Huang, Z.; Liu, F.; Zhou, P.; and Zou, Y. 2021. Sentiment Injected Iteratively Co-iteractive Network for Spoken Language Understanding. In *ICASSP*.

Huang, Z.; Liu, F.; and Zou, Y. 2020. Federated Learning for Spoken Language Understanding. In *COLING*, 3467–3478.

Kociský, T.; Schwarz, J.; Blunsom, P.; Dyer, C.; Hermann, K. M.; Melis, G.; and Grefenstette, E. 2018. The NarrativeQA Reading Comprehension Challenge. *TACL* 6: 317–328.

Kuo, C.; Luo, S.; and Chen, K. 2020. An Audio-Enriched BERT-Based Framework for Spoken Multiple-Choice Question Answering. In *INTERSPEECH*, 4173–4177.

Lai, G.; Xie, Q.; Liu, H.; Yang, Y.; and Hovy, E. H. 2017. RACE: Large-scale ReAding Comprehension Dataset From Examinations. In *EMNLP*, 785–794.

Liu, F.; Liu, Y.; Ren, X.; He, X.; and Sun, X. 2019a. Aligning Visual Regions and Textual Concepts for Semantic-Grounded Image Representations. In *NeurIPS*, 6847–6857.

Liu, F.; Ren, X.; Liu, Y.; Lei, K.; and Sun, X. 2019b. Exploring and Distilling Cross-Modal Information for Image Captioning. In *IJCAI*, 5095–5101.

Liu, F.; Ren, X.; Liu, Y.; Wang, H.; and Sun, X. 2018. simNet: Stepwise Image-Topic Merging Network for Generating Detailed and Comprehensive Image Captions. In *EMNLP*, 137–149.

Liu, F.; Wu, X.; Ge, S.; Fan, W.; and Zou, Y. 2020a. Federated Learning for Vision-and-Language Grounding Problems. In *AAAI*, 11572–11579.

Liu, F.; Wu, X.; Ge, S.; Zhang, X.; Fan, W.; and Zou, Y. 2020b. Bridging the Gap between Vision and Language Domains for Improved Image Captioning. In *ACM Multimedia*, 4153–4161.

Liu, S.; Zhang, X.; Zhang, S.; Wang, H.; and Zhang, W. 2019c. Neural Machine Reading Comprehension: Methods and Trends. *CoRR* abs/1907.01118.

Lu, J.; Yang, J.; Batra, D.; and Parikh, D. 2016. Hierarchical Question-Image Co-Attention for Visual Question Answering. In *NIPS*, 289–297.

Nguyen, T.; Rosenberg, M.; Song, X.; Gao, J.; Tiwary, S.; Majumder, R.; and Deng, L. 2016. MS MARCO: A Human Generated MAchine Reading COmprehension Dataset. In *NIPS*.

Pennington, J.; Socher, R.; and Manning, C. D. 2014. Glove: Global Vectors for Word Representation. In *EMNLP*, 1532–1543.

Qiu, B.; Chen, X.; Xu, J.; and Sun, Y. 2019. A Survey on Neural Machine Reading Comprehension. *CoRR* abs/1906.03824.

Rajpurkar, P.; Jia, R.; and Liang, P. 2018. Know What You Don't Know: Unanswerable Questions for SQuAD. In *ACL*, 784–789.

Rajpurkar, P.; Zhang, J.; Lopyrev, K.; and Liang, P. 2016. SQuAD: 100, 000+ Questions for Machine Comprehension of Text. In *EMNLP*, 2383–2392.

Romero, A.; Ballas, N.; Kahou, S. E.; Chassang, A.; Gatta, C.; and Bengio, Y. 2015. FitNets: Hints for Thin Deep Nets. In *ICLR*.

Serdyuk, D.; Wang, Y.; Fuegen, C.; Kumar, A.; Liu, B.; and Bengio, Y. 2018. Towards End-to-end Spoken Language Understanding. In *ICASSP*, 5754–5758.

Srivastava, N.; Hinton, G. E.; Krizhevsky, A.; Sutskever, I.; and Salakhutdinov, R. 2014. Dropout: a simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research* 15(1): 1929–1958.

Tseng, B.-H.; Shen, S.-S.; Lee, H.-Y.; and Lee, L.-S. 2016. Towards machine comprehension of spoken content: Initial TOEFL listening comprehension test by machine. In *INTERSPEECH*, 2731–2735.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; and Polosukhin, I. 2017. Attention is All you Need. In *NIPS*, 5998–6008.

Wang, S.; Yu, M.; Jiang, J.; and Chang, S. 2018. A Co-Matching Model for Multi-choice Reading Comprehension. In *ACL*, 746–751.

Xu, K.; Ba, J.; Kiros, R.; Cho, K.; Courville, A.; Salakhudinov, R.; Zemel, R.; and Bengio, Y. 2015. Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. In *ICML*, 2048–2057.

Yang, H.; Chaisorn, L.; Zhao, Y.; Neo, S.; and Chua, T. 2003. VideoQA: question answering on news video. In *ACM Multimedia*, 632–641.

Yu, D.; and Li, J. 2017. Recent progresses in deep learning based acoustic models. *IEEE CAA J. Autom. Sinica* 4(3): 396–409.

Zhang, S.; Zhao, H.; Wu, Y.; Zhang, Z.; Zhou, X.; and Zhou, X. 2019. Dual Co-Matching Network for Multi-choice Reading Comprehension. *CoRR* abs/1901.09381.

Zhang, Y.; Dai, H.; Kozareva, Z.; Smola, A. J.; and Song, L. 2018. Variational Reasoning for Question Answering With Knowledge Graph. In *AAAI*, 6069–6076.

Zhang, Z.; Wu, Y.; Zhou, J.; Duan, S.; Zhao, H.; and Wang, R. 2020. SG-Net: Syntax-Guided Machine Reading Comprehension. In *AAAI*, 9636–9643.

Zhou, P.; Huang, Z.; Liu, F.; and Zou, Y. 2020. PIN: A Novel Parallel Interactive Network for Spoken Language Understanding. In *ICPR*.

Zhu, P.; Zhao, H.; and Li, X. 2020. Dual Multi-head Co-attention for Multi-choice Reading Comprehension. *CoRR* abs/2001.09415.