INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Multi-Scale Attention for Audio Question Answering

Guangyao Li[1], *Yixin Xu*[1], *Di Hu*[1,2,*]

[1]Gaoling School of Artificial Intelligence, Renmin Uniiversity of China, Beijing
[2]Beijing Key Laboratory of Big Data Management and Analysis Methods, Beijing
{guangyaoli, xu_yixin, dihu}@ruc.edu.cn

Abstract

Audio question answering (AQA), acting as a widely used proxy task to explore scene understanding, has got more attention. The AQA is challenging for it requires comprehensive temporal reasoning from different scales' events of an audio scene. However, existing methods mostly extend the structures of visual question answering task to audio ones in a simple pattern but may not perform well when perceiving a fine-grained audio scene. To this end, we present a **M**ulti-scale **W**indow **A**ttention **F**usion **M**odel (**MWAFM**) consisting of an *asynchronous hybrid attention module* and a *multi-scale window attention module*. The former is designed to aggregate unimodal and cross-modal temporal contexts, while the latter captures sound events of varying lengths and their temporal dependencies for a more comprehensive understanding. Extensive experiments are conducted to demonstrate that the proposed MWAFM can effectively explore temporal information to facilitate AQA in the fine-grained scene. [1]

Index Terms: Audio Question Answering, Multi-scale Attention, Temporal reasoning.

1. Introduction

We are surrounded by a complex mixture of audio signals in daily life, and our auditory perception system unconsciously focuses on sound sources of interest [1]. Imagine sitting on your sofa at home with your eyes closed, and you can hear and recognize a succession of sounds: baby crying, family members speaking, their footsteps, etc. Understanding all these sounds and interpreting the perceived scene as a domestic scene comes naturally to humans but is still challenging for machines. Hence, making machines leverage audio information, especially authentic sounds in natural scenes, to achieve considerable audio scene perception and understanding ability as humans is an interesting and valuable topic.

Recently, we have seen significant progress in sound event detection [2, 3], speech recognition [4, 5] and enhancement [6], music transcription [7], audio source separation [8], and audio question answering [9, 10] toward audio scene understanding. Among them, question answering has been widely used as a proxy task to explore scene understanding along with getting more and more attention. Inspired by CLEVR [11], some researchers proposed a program-generated dataset containing fixed-length audio sequences of different notes to explore the AQA task [12]. Following this trajectory, the DAQA [9] and NAAQA [13] are proposed to extend the questions to more acoustically realistic situations and have achieved outstanding achievements. However, the generated data usually lack diversity and challenges in the natural scene. To this end, some others tend to focus on answering questions with spoken [14] and voice [15] information and answering questions about videos might require joint reasoning about visual and audio cues. For the audio scenes above main consist of human speech, to explore more natural sound scene understanding, Clotho-AQA [10] is proposed that contains audio files of day-to-day sounds occurring in the environment, such as birds, etc., while avoiding human speech. Meanwhile, MUSIC-AVQA [16] also includes many natural sounds and includes the AQA sub-task. These datasets provide a good testbed for the AQA task and attract researchers' attention. To our knowledge, current research on AQA is mostly presented in the form of new datasets and baseline networks (e.g., DAQA, Clotho-AQA), but ignores the characteristics of sound scenarios such as event duration and long-term dependence. Additionly, reasoning exploration in VQA tasks focuses on spatial or spatio-temporal (e.g., Audio-MUSIC-AQA) problems, and its spatial information plays an indispensable role in model, which is not suitable for directly transferring reasoning network in VQA to AQA tasks. Hence, how to explore temporal reasoning problems in AQA tasks effectively, especially capturing sound events at different scales and their correlation, is challenging but significant for fine-grained sound scene understanding.

In this work, we propose a **M**ulti-scale **W**indow **A**ttention **F**usion **M**odel (**MWAFM**) on Clotho-AQA and Audio-MUSIC-AVQA dataset. Concretely, we apply a new asynchronous hybrid attention module (AHAM) to leverage unimodal and cross-modal temporal contexts. Then, given that multi-length sound event, whose temporal context is crucial for understanding complex scenes, a develop multi-scale window attention module (MWAM) to capture sound events of different scales. Furthermore, since the audio event changes over time dynamically, the model uses question features as queries to attend crucial temporal segments for encoding question-aware audio embedding effectively. As an open-ended problem, the correct answers to questions can be predicted by choosing words from a predefined answer vocabulary. To validate the proposed model, we conducted a large number of experiments, including comparison with mainstream QA methods, various ablation study module et al,. Sufficient experimental results demonstrate the effectiveness and generalization of the proposed MWAFM.

The main contributions include: **1)** A novel asynchronous hybrid attention module to leverage unimodal and cross-modal temporal contexts. **2)** An effective attentive multi-scale window attention module captures sound events of various lengths and their temporal contextual dependencies. **3)** The proposed MWAFM provides a powerful ability for AQA tasks, which achieves superior performance on two public benchmarks and demonstrates the model's effectiveness and generalization.

[1] Code:https://github.com/GeWu-Lab/MWAFM, *Corresponding author.

10.21437/Interspeech.2023-1606

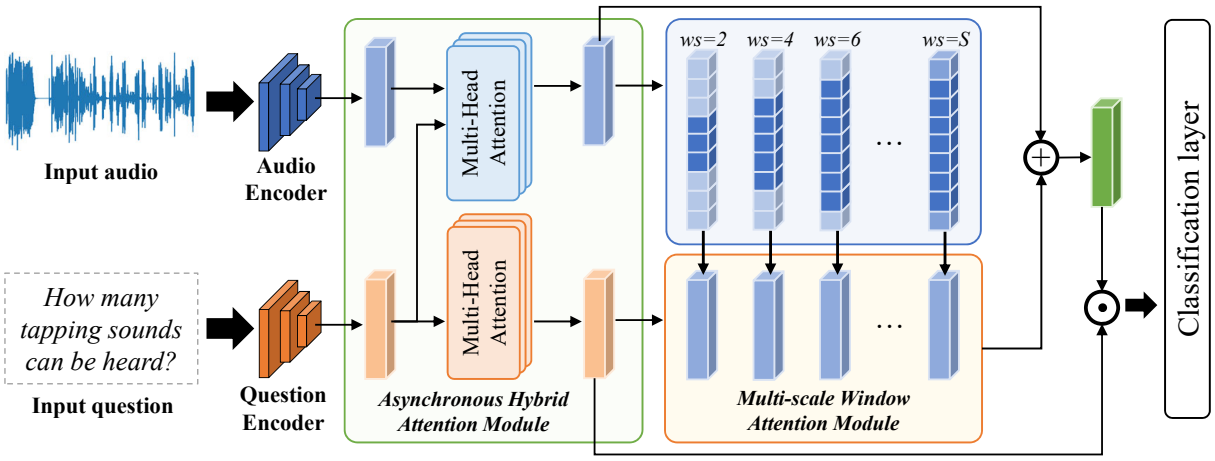

Figure 1: *The pipeline of our proposed **M**ulti-scale **W**indow **A**ttention **F**usion **M**odel (**MWAFM**). The model takes pre-trained CNNs to extract audio features and uses a Fasttext pre-trained word vectors to obtain question features. Firstly, we aggregate unimodal and cross-modal temporal contexts. Then we highlight audio features of key timestamps for temporal multi-scales information through question queries. Finally, multimodal fusion is exploited to integrate audio and question information for predicting the answer to the input question. (**ws** indicate **w**indow **s**ize).*

2. Method

To facilitate audio question answering task effectively, we proposed **M**ulti-scale **W**indow **A**ttention **F**usion **M**odel (**MWAFM**), which is composed of an asynchronous hybrid attention module (AHAM), and a multi-scale window attention module (MWAM). The architecture of **MWAFM** is illustrated in Figure. 1.

2.1. Representations for Different Modalities

Audio Representation. Given an input audio sequence, we first divide it into T non-overlapping audio segments $A = \{a_1, a_2, ..., a_T\}$, where each segment is $1s$ long, and T is the timestamp. We encode each audio segment A_t into a feature vector f_a^t using a pre-trained VGGish model [17], which is a VGG-like 2D CNN network, employing over transformed audio spectrograms. The audio representation is extracted offline and the model is not fine-tuned.

Question Representation. For an asked question $Q = \{q_n\}_{n=1}^N$ input, the Fasttext pre-trained model [18] is used to process projected word vectors. If the input question Q has N words, the word embedding shape using Fasttext is $N \times 300$. And the word embeddings are passed through a linear layers into a feature vector f_q.

2.2. Asynchronous Hybrid Attention Module

Natural audios tend to contain continuous and repetitive rather than isolated sound event. Especially, sound events usually redundantly recur many times. To capture multimodal temporal contexts, we design a new temporal mechanism: Asynchronous Hybrid Attention module (AHAM), which uses two unimodal encoders and a multimodal encoder to learn which snippets for each audio adaptively.

Unimodal Encoder. For each of the question and audio input representations $\{f_q^n\}_{n=1}^N, \{f_a^t\}_{t=1}^T$, we first apply the layer normalization and feed them into the corresponding unimodal encoder, in which we use self-attention module to update f_q^n and f_a^t, respectively. An asynchronous hybrid attention function H in AHAM will be learned from audio and question features: $\{q_i\}_{i=1}^N, \{a_i\}_{i=1}^T$ to update f_q^n and f_a^t, respectively. The updated audio and question feature $\hat{f}_a^t$, $\hat{f}_q^n$ can be obtained with same computation:

$$\hat{f}_m^l = f_m^l + \sum_{l=1}^{\intercal} w_l f_m^l = f_m^l + \sigma(\frac{f_m^l \boldsymbol{f}_m^\intercal}{\sqrt{d}}) \boldsymbol{f}_m \ , \quad (1)$$

where m is the question or audio modality, and l is their corresponding length; $f_a^t = [f_a^1; ...; f_a^T]$ and $f_q^n = [f_q^1; ...; f_q^N]$; skip-connections can help preserve the identity information from the input sequences. The σ is softmax function; the d is a scaling factor with the same size as the feature dimension and $(\cdot)^\intercal$ denotes the transpose operator.

Multimodal Encoder. To highlight the audio key timestamps closely associated with the question, we utilize the asynchronous cross-modal attention, designed to attend critical temporal segments among the changing audio scenes and capture question-aware audio embeddings. Concretely, given $\{f_q^n\}_{n=1}^N$ and audio features $\{f_a^t\}_{t=1}^T$ from input embedding, the multimodal encoder will learn to aggregate question-aware audio features. The module will produce a more robust audio feature representation F_a^t:

$$F_a^t = f_a^t + \sum_{t=1}^{\intercal} w_t^a f_a^t = f_a^t + \sigma(\frac{f_q \boldsymbol{f}_a^\intercal}{\sqrt{d}}) \boldsymbol{f}_a \ . \quad (2)$$

The AHAM will assign large weights to audio snippets which are more relevant to the question. Then the obtained audio and question features $F_a, F_q = \hat{f}_q$ are input to the MWAM.

2.3. Multi-scale Window Attention Module

To capture the intrinsic property that different events have various duration, the multi-scale window attention with different window size is designed, such a stacked shifted window transformer with window size increasing with layers becoming deeper. Given the importance of local context, our attention pattern employs multi-size window attention surrounding each token. Given a fixed window size S, each token attends to $S/2$ on each side (The dark part in the green area in Figure. 2). For

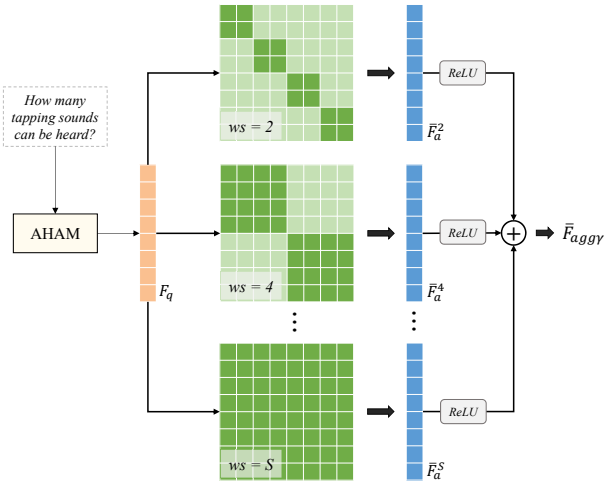

Figure 2: *Multi-scale Window Attention Module (**MWAM**). The Module will aggregate multi-scale question-aware audio features ( ), which is obtained through question queries on the output from sliding window of different size.*

different scales of S, we compute the self-attention score of the S window size on audio, where $Q\text{-}value, K\text{-}value, V\text{-}value$ are the corresponding values in the sliding window. The audio feature expression $\hat{F}_a^i$ at different scales by:

$$\hat{F}_a^i = MHAttn(F_a^i, F_a^i, F_a^i), i = 2, 4, 6, ..., S, \quad (3)$$

where S is the size of the sliding window. Then, the model will learn to aggregate question-aware audio features on different scales $\hat{F}_a^i$ by *Equation (2)*. And the output feature $\bar{F}_a^i$ can be computed as:

$$\bar{F}_a^i = MHAttn(F_q, F_a^i, F_a^i), i = 2, 4, 6, ..., S. \quad (4)$$

We aggregate $\bar{F}_a^i$ and F_a to better capture multi-scale sound events under the guidance of question by:

$$\bar{F}_{aggr} = Norm(\sum_i^S (ReLU(\bar{F}_a^i) + ReLU(F_a))) \ , \quad (5)$$

where $\bar{F}_{aggr}$ is audio feature through MWAM and AHAM. The illustration of *Equation (5)* can be seen in the blue part at the bottom of Figure 2. Among them, a linear projection and a dropout layer are operated before *ReLU*. Hence, the quesiton queried audio contextual embeddings are more capable for predicting correct answers.

2.4. Feature Fusion and Answer Prediction

Different modalities can contribute to correctly answer questions. To combine the features: $\bar{F}_{aggr}$ and F_q, we introduce a simple multimodal fusion network. Specifically, we integrate audio and question features with employing an element-wise multiplication operation by $e = \bar{F}_{aggr} \circ F_q$.

To achieve AQA task, we predict the answer for a given question based on the joint embedding e. It can be formulated as an open-ended task, which aims to choose one correct word as the answer from a pre-defined answer vocabulary. We utilize a linear layer and softmax function to output a probability $p \in \mathcal{R}^C$ for candidate answers. With the predicted probability vector and the corresponding ground-truth label y, we can optimize our network using a cross-entropy loss:

$$\mathcal{L}_{qa} = -\sum_{c=1}^{C} y_c log(p_c). \quad (6)$$

During testing, we can select the predicted answer by $\hat{c} = \arg\max_c(p)$.

Table 1: *AQA results of different methods on the test set of Chloto-AQA. The top-1 results are highlighted.*

Method	Top-1 Acc	Top-5 Acc	Top-10 Acc
GRU [20]	09.21	26.02	36.87
BiLSTM [21]	14.39	35.90	47.00
HME [22]	12.79	32.85	42.54
PSAC [23]	14.29	35.80	46.32
LongFormer [24]	15.98	36.72	47.67
FCNLSTM [9]	20.16	44.43	55.57
CONVLSTM [9]	18.99	40.70	49.27
ST-AVQA [16]	17.20	40.16	50.73
AquaNet [10]	14.78	36.19	46.71
MWAFM	**22.24**	**46.56**	**57.75**

3. Experiments and Analysis

3.1. Datasets

To benchmark the performance of **MWFAM**, we conduct experiments on two large-scale datasets, including Clotho-AQA and Audio-MUSIC-AVQA. **Clotho-AQA**, an audio question answering dataset consisting of 1,991 audio files selected from the Clotho [19] each between 15 to 30 seconds in duration. The training, validation and test splits of Clotho-AQA contain 1,174, 344, and 473 audio files and 828, 512 and 801 unique answers without *yes/no*, respectively. **Audio-MUSIC-AVQA** is a subset of MUSIC-AVQA [16] dataset that only contains audio-related question-answer pairs and the corresponding audio. We split the dataset into training, validation, and testing sets for training and evaluation with 5,633, 806, and 1,611 QA pairs, respectively.

3.2. Evaluation

We follow the standard evaluation protocol of each dataset. We use answer prediction accuracy as the metric and evaluate model performance on answering different types of questions. It should be noted that we use *Top*-1, *Top*-5, and *Top*-10 accuracy metrics to evaluate the performance of the Clotho-AQA dataset because the number of unique answer classes is high (828). But we only use *Top*-1 accuracy as a metric on the Audio-MUSIC-AVQA dataset because its answer vocabulary consists of 42 possible answers to different questions. For training, we use one single model to handle all questions without training separated models for each type.

3.3. Implementation Details

The audio stream is divided into non-overlapping 1s segments, and the sampling rate of sounds is 16 kHz. We use a linear layer for each 1s-long audio segment to process the extracted 128-D VGGish feature into a 512-D feature vector. The audio is fixed to 24 seconds through the *interpolate* operation, and the question length is fixed to 20. Batch-size and number of epochs are 64 and 50, respectively. The initial learning rate is $1e$-4, and our network is trained with the Adam optimizer. All the experiments are conducted on NVIDIA-V100 GPUs.

Table 2: *Ablation study on different window-size and the proposed modules. We observe that leveraging UHAM and MWAM can boost AQA task. (**ws** indicate **w**indow **s**ize).*

Method	**ws**	*Top*-1	*Top*-5	*Top*-10
w/ ws-attn	2	21.07	45.10	55.64
w/ ws-attn	4	21.27	44.57	55.45
w/ ws-attn	6	21.42	44.90	56.07
w/ ws-attn	12	21.32	45.17	56.06
w/o un-queried	2,4,6,12	20.30	43.51	54.07
w/ bi-queried	2,4,6,12	21.17	43.22	55.57
w/o AHAM	2,4,6,12	13.32	36.14	47.24
w/o MWAM	-	21.75	44.48	55.23
MWAFM′	12,12,12,12	20.35	43.17	54.89
MWAFM*	——	21.22	46.08	56.44
MWAFM	2,4,6,12	**22.24**	**46.56**	**57.75**

3.4. Results and analysis

In this subsection, we evaluate the proposed MWAFM on Clotho-AQA and Audio-MUSIC-AVQA to present the model performance for the AQA task.

Comparison with QA Models. To validate our MWAFM on the Clotho-AQA dataset, we compare it with recent QA methods: GRU [20], BiLSTM [21], AquaNet [10], HME [22], LongFromer [24], FCNLSTM [9], CONVLSTM [9], ST-AVQA [16]. Specifically, FCNLSTM is applied to temporal reasoning of sound events, but it is difficult to capture multi-lengths events because of its only modeling of global information. AquaNet is designed for AQA tasks, but it simply migrates from VQA methods without reasoning capabilities. Although the ST-AVQA method is dedicated to real-world scene understanding and reasoning, it combines audio and visual modalities to improve the perception ability of the model. For the scene containing only audio modality, it is difficult for the temporal module of the ST-AVQA to carry out fine-grained understanding. Tabel. 1 shows results of those comparable methods, where we only use the *Temporal Grounding Module* in the ST-AVQA method. The results demonstrate that our model achieves considerable improvement on most QA methods. It is worth noting that our method achieves excellent performance on which metrics are *Top*-1, *Top*-5, and *Top*-10, indicating the effectiveness of the proposed MWAFM.

Ablation study. Table. 2 shows the effects of AHAM and MWAM and their components on our model performance. Obviously, temporal contexts can improve the performance of the model. In AHAM, the model only uses the entire question feature as a query *(MWAFM)* to calculate the attention score on audio feature sequence, which outperforms their cross-modal attention *(w/ bi-queried)*. At the same time, the performance is further degraded if both are not used *(w/o un-queried)*. It shows that it is tending to associate the semantics of the sentence with a particular segment of the audio. In MWAM, we conducted a large number of experiments with multiple random seeds to take the average value, and calculated the value of the standard deviation under each evaluation index to be about 0.5. It can be seen that the proposed model aggregating audio features at multiple scales is much better than using only features of a single window size. Furthermore, the experiment demonstrates that the proposed MWAFM can leverage unimodal and cross-modal to capture the machine dependencies of sound events at different scales. Additionally, we provide convincing results by using multiple large window sizes, like MWAFM′ in Table. 2. The result shows the combined effect of attention module networks with multi-scale window sizes is the best. The fact is that short-term sound events tend to be more related to adjacent sound segments and less to distant. Meanwhile, considering the long duration of some acoustic events, the proposed model retains the large window attention mechanism. Therefore, an effective attentive multi-scale window attention module captures sound events of various lengths and their temporal contextual dependencies well. Apart from this, we replace the MWAM with a conventional transformer that stacks asynchronous hybrid attention modules (MWAFM*), and set the stacked layers to 4 for fair comparison with 4 different window sizes in MWAM. We conduct experiments on Cloto-AQA and find that this way achieves *Top*-1, *Top*-5, and *Top*-10 scores of 21.22%, 46.08%, and 56.44%, respectively. Compared to using MWAM, the results are lower by 1.02%, 0.48%, and 1.31%, primarily because of the variable length of events in the audio that make it challenging for fixed-size window attention stacking to capture them. Hence, it is necessary to use a multi-scale window attention module, which can enhance the model's performance.

Table 3: *AQA results of different methods on the test set of Audio-MUSIC-AVQA. The top-1 results are highlighted.*

Method	Counting	Comparative	Average
GRU [20]	65.00	64.31	64.74
BiLSTM [21]	65.59	45.62	58.22
HCAttn [25]	66.08	56.23	62.45
HME [22]	67.65	63.97	66.29
PSAC [23]	69.03	60.94	66.05
FCNLSTM [9]	66.96	62.96	65.49
CONVLSTM [9]	68.04	62.63	66.05
ST-AVQA [16]	67.75	**64.65**	66.60
AquaNet [10]	65.59	52.86	60.89
MWAFM	**69.42**	64.31	**67.54**

Generalization of the model. We expect the proposed method to be effective in other similar real-world sound scene understanding. Audio-MUSIC-AVQA proposed by Li et al. [16] contains rich dynamic and complex audio scenes, which is very suitable for verifying the performance of MWAFM. Therefore, we conduct extensive experiments on this dataset. As shown in Tabel. 3, the overall results of the proposed MWAFM method achieve the best. Although the performance on some sub-problems is not the best, it also achieves satisfactory results. They are sufficient to illustrate the generalization performance of the proposed MWAFM method.

4. Conclusion

In this paper, we present a novetly **M**ulti-scale **W**indow **A**ttention **F**usion **M**odel (**MWAFM**) for audio question answering task. The proposed MWAFM provides a powerful ability for AQA tasks, which can capture sound events of various lengths and their temporal contextual dependencies well and assign large weights to key audio snippets that are more relevant to the question. We evaluate the MWAFM on two benchmark datasets, which illustrates the effectiveness and generalization of the proposed model. We believe our research will facilitate the development of fine-grained audio scene understanding, especially in terms of temporal reasoning.

Acknowledgement. This research was supported by National Natural Science Foundation of China (NO.62106272), the Young Elite Scientists Sponsorship Program by CAST (2021QNRC001), and Public Computing Cloud, Renmin University of China.

5. References

[1] A. Mesaros, T. Heittola, T. Virtanen, and M. D. Plumbley, "Sound event detection: A tutorial," *IEEE Signal Processing Magazine*, vol. 38, no. 5, pp. 67–83, 2021.

[2] E. Cakır, G. Parascandolo, T. Heittola, H. Huttunen, and T. Virtanen, "Convolutional recurrent neural networks for polyphonic sound event detection," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 6, pp. 1291–1303, 2017.

[3] Z. Mnasri, S. Rovetta, and F. Masulli, "Anomalous sound event detection: A survey of machine learning based methods and applications," *Multimedia Tools and Applications*, vol. 81, no. 4, pp. 5537–5586, 2022.

[4] A. Baevski, W.-N. Hsu, A. Conneau, and M. Auli, "Unsupervised speech recognition," *Advances in Neural Information Processing Systems*, vol. 34, pp. 27 826–27 839, 2021.

[5] J. Li *et al.*, "Recent advances in end-to-end automatic speech recognition," *APSIPA Transactions on Signal and Information Processing*, vol. 11, no. 1, 2022.

[6] D. Michelsanti, Z.-H. Tan, S.-X. Zhang, Y. Xu, M. Yu, D. Yu, and J. Jensen, "An overview of deep-learning-based audio-visual speech enhancement and separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 1368–1396, 2021.

[7] E. Benetos, S. Dixon, Z. Duan, and S. Ewert, "Automatic music transcription: An overview," *IEEE Signal Processing Magazine*, vol. 36, no. 1, pp. 20–30, 2018.

[8] S. Makino, *Audio source separation*. Springer, 2018, vol. 433.

[9] H. M. Fayek and J. Johnson, "Temporal reasoning via audio question answering," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2283–2294, 2020.

[10] S. Lipping, P. Sudarsanam, K. Drossos, and T. Virtanen, "Clotho-aqa: A crowdsourced dataset for audio question answering," in *2022 30th European Signal Processing Conference (EUSIPCO)*. IEEE, 2022, pp. 1140–1144.

[11] J. Johnson, B. Hariharan, L. Van Der Maaten, L. Fei-Fei, C. Lawrence Zitnick, and R. Girshick, "Clevr: A diagnostic dataset for compositional language and elementary visual reasoning," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2901–2910.

[12] J. Abdelnour, G. Salvi, and J. Rouat, "Clear: A dataset for compositional language and elementary acoustic reasoning," *arXiv preprint arXiv:1811.10561*, 2018.

[13] J. Abdelnour, J. Rouat, and G. Salvi, "Naaqa: A neural architecture for acoustic question answering," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[14] T. Zhang, D. Dai, T. Tuytelaars, M.-F. Moens, and L. Van Gool, "Speech-based visual question answering," *arXiv preprint arXiv:1705.00464*, 2017.

[15] A. Chuklin, A. Severyn, J. R. Trippas, E. Alfonseca, H. Silen, and D. Spina, "Using audio transformations to improve comprehension in voice question answering," in *International Conference of the Cross-Language Evaluation Forum for European Languages*. Springer, 2019, pp. 164–170.

[16] G. Li, Y. Wei, Y. Tian, C. Xu, J.-R. Wen, and D. Hu, "Learning to answer questions in dynamic audio-visual scenarios," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022, pp. 19 108–19 118.

[17] J. F. Gemmeke, D. P. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2017, pp. 776–780.

[18] T. Mikolov, É. Grave, P. Bojanowski, C. Puhrsch, and A. Joulin, "Advances in pre-training distributed word representations," in *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, 2018.

[19] K. Drossos, S. Lipping, and T. Virtanen, "Clotho: An audio captioning dataset," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 736–740.

[20] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh, "Vqa: Visual question answering," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 2425–2433.

[21] C. Hori, T. Hori, T.-Y. Lee, Z. Zhang, B. Harsham, J. R. Hershey, T. K. Marks, and K. Sumi, "Attention-based multimodal fusion for video description," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 4193–4202.

[22] C. Fan, X. Zhang, S. Zhang, W. Wang, C. Zhang, and H. Huang, "Heterogeneous memory enhanced multimodal attention model for video question answering," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 1999–2007.

[23] X. Li, J. Song, L. Gao, X. Liu, W. Huang, X. He, and C. Gan, "Beyond rnns: Positional self-attention with co-attention for video question answering," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, 2019, pp. 8658–8665.

[24] I. Beltagy, M. E. Peters, and A. Cohan, "Longformer: The long-document transformer," *arXiv preprint arXiv:2004.05150*, 2020.

[25] J. Lu, J. Yang, D. Batra, and D. Parikh, "Hierarchical question-image co-attention for visual question answering," *arXiv preprint arXiv:1606.00061*, 2016.