Audio-Visual Agreement for Audio-Visual Question Answering

Kyung Myung Ko[1]

[1]Department of Computer Science, Purdue University

Abstract

Recently, transformer architectures has brought simplicity to contextualize multiple modalities to maximize joint information. This is especially beneficial in understanding complex scenes such as musical performances. However, the challenges arise from the natural ambiguity in each modalities. In this work, we propose agreement and alignment between the modalities to maximize the joint information. We attend the text queries to each modalities to measure the agreement while maximizing their similarities. We believe this work will narrow down to target the necessary information for understanding such complex scenes. The code is available at https://github.com/andyko208/AVAG-AVQA.

1 Introduction

Humans understand the world through multiple senses, where each of them contribute uniquely to the understanding the complex environments. However, not all sensory information can be captured and digitized for machines to learn in the same way. This presents a significant challenge in developing systems capable of understanding the world, particularly when constrained by limited modalities. Multimodal learning aims to integrate information from different sources, such as vision and sound, to form a more holistic understanding. Still, abundant information does not always lead to better understanding of the world, as the modalities can interfere with each other, leading to confusion rather than clarity.

In the context of musical performances, the challenge becomes even more evident. For instance, if someone is unfamiliar with the sound of a specific instrument, they have to rely on visual cues like the motion of the performer to reason about the scene. However, when multiple similar instruments are played simultaneously, distinguishing between them based on visual or auditory information alone becomes difficult. This task is further complicated by the abstract nature of audio, which often lacks objective details that can be easily inferred, requiring subjective interpretation. Machines face a unique challenge in this context, as they require not just multi-modal inputs but also sophisticated reasoning abilities to make sense of incomplete or ambiguous information.

In this work, we propose a novel approach to improve the understanding of complex music performance scene by minimizing the information gap in the modalities. We utilize a dataset constructed with a collection of musical performances in diverse scenes, of which the questions are paired with videos along with 42 different answers as labels to each audio, visual, and audio-visual types of question. We approach the problem from the idea of disagreement between the modalities when each are individually utilized to answer such questions. Therefore, we propose two different modules to maximize the information from each modalities when answering questions that require both modalities. The primary goal for this project is to to improve performance on the existential questions on the audio-visual question category. Such example would be, "How many instruments are sounding in the video?". By integrating each modules that captures agreement between the modalities and the similarity between the modalities, we seek to advance the machine's ability to perceive and such dynamic scenes, moving closer to human-like understanding.

2 Related Work

Audio-visual multimodal learning explores complex relationship between the separate modalities. Various work in this domain had been published, such as audio-visual source separation [1, 2, 3] and audio-visual event localization [8, 9, 11]. As an extension of tradition VQA, audio-visual question answering task has gained attention as to model the temporal relationship of diverse videos rather than from static images or videos with the existence of human speech. Unlike to traditional videos, audio-visual question answering put emphasis on the audio itself rather than the transcribed

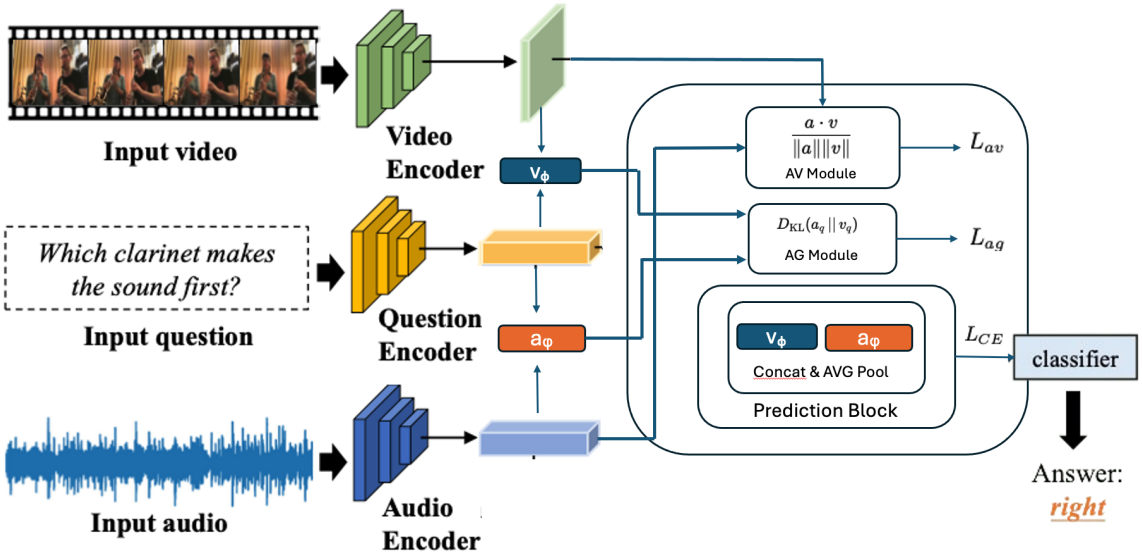

Figure 1: The proposed AV-AG model. Encoded visual and audio embeddings to compute similarity and each question-attened embeddings for maximizing agreement with KL-divergence.

text. This poses challenges, as audio itself may be interpreted in diverse ways depending on the situation and the environment. Past work has to introduced this challenge with novel QA benchmarks. Pano-AVQA [12] first set up a QA benchmark that focused on 360 degree videos. Later on, MUSIC-AVQA [6] proposed a new benchmark that further incorporates temporal dynamics that was not captured in previous benchmarks.

3 Method

To tackle the challenges in AVQA, we propose a mechanism that not only focuses on the alignment of different modalities using contextualized embedding but also measures the agreement between the modalities with respect to the question. In this task, there are three modalities: images, audio, and text.

Audio Encoder We encode each audio segment A_t into a feature vector f_t a using a pre-trained VGGish model [5]. The audio representation is extracted offline and the model is not fine-tuned.

Visual Representation We sample a fixed number of frames for all video segments. We then apply pre-trained ResNet-18 [4] on video frames to extract visual feature map $f^t_{v,m}$ for each video segment V_t . The used pre-trained ResNet-18 model is not fine-tuned.

Question Representation For an asked question $Q = \{q_n\}^N_{n=1}$, a RoBERTa [7] is used to process projected word embeddings $\{f_q\}^N_{n=1}$and encode the question into a feature vector f_q using the last hidden state. The question encoder is not fine-tuned.

3.1 AV Alignment Module

One of the main problems of this task is the misalignment of the audio and visual information. Thus, maximizing their joint information is crucial. We randomly sample the batches and obtain the corresponding audio and visual feature pairs. We compute the cosine similarity between the modalities with a threshold and maximize their similarity through backpropagation.

3.2 AV Agreement Module

Our main contribution of this work is to maximize the disagreement of individual modalities with respect to the provided question. First, each modality is performed cross-attention with the text modality. Then, we maximize the attended modalities agreement with Kullback–Leibler Divergence. We empirically determined that KL-Divergence works the best over cosine-similarity and soft-dtw.

3.3 Answer Prediction

We concatenate the embedded representation from each module and perform adaptive average pooling to aggregate their information to pass through the prediction block and perform multi-class classification for the 42 answer labels with cross-entropy loss. We formulate the multimodal fusion step as $\mathbf{a}_{qv} = \text{FC}(\text{MLP}(\text{AdaptiveMaxPool1d}(\text{Concat}[\mathbf{a}_q, \mathbf{v}_q])))$ For each modules we empirically determine the scale values. In the end, our loss function looks as follows:

$$L = \lambda_1 L_{av} + \lambda_2 L_{ag} + L_{CE} \quad (1)$$

We utilize a softmax function to output a probabilities $p \in R^C$ for candidate answers. During testing, we

Audio			Visual			Audio-Visual						
Count	Comp	Avg	Count	Location	Avg	Existential	Location	Count	Comp	Temporal	Avg	Total Avg
78.18	**67.05**	**74.06**	**71.56**	**76.38**	**74.00**	81.81	**64.51**	**70.80**	**66.01**	**63.23**	**69.54**	**71.52**
64.44	60.17	62.31	55.51	57.75	56.63	**82.20**	40.72	44.00	62.98	40.34	54.05	57.67

Figure 2: AVQA result comparison by question types, baseline on the top.

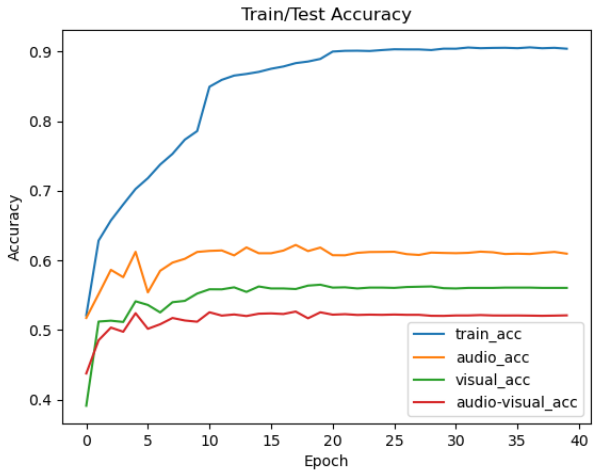

Figure 3: With AG & AV module

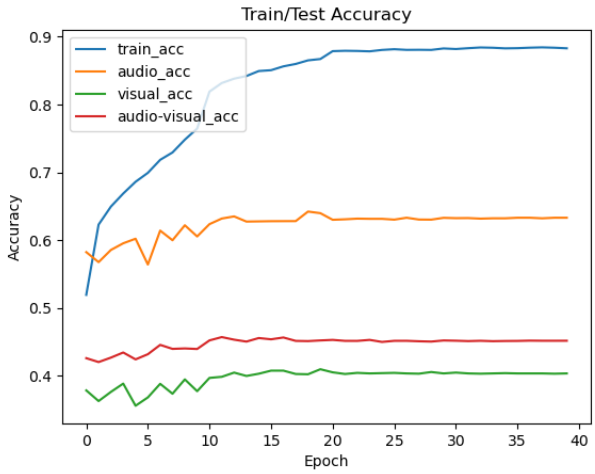

Figure 4: Without AG & AV module

select the predicted answer by $c^* = argmax_c(p)$. A general description of the model is shown in Figure 1.

4 Experiments

Implementation Details The sampling rates of sounds and video frames are sampled at 16 kHz and 1 fps. For each video, we divide it into non-overlapping segments of the same length with 1 frame and generate a 512-D feature vector for each visual segment. For each 1s-long audio segment, we use a linear layer to process the extracted 128-D VGGish feature into a 512-D feature vector. The dimension of the word embedding is set to 512. With the limitation of computing resources, we sampled the videos by taking 1s every 6s. Batch size and number of epochs are set to 16 and 40. The initial learning rate is 1e-3 and will drop by multiplying 0.1 every 10 epochs. Our networks is trained with the Adam optimizer.

Training Strategy We utilize Pre-LN Transformer [10] and stabilize the learning process by applying layer norm before applying the multi-head attention to the query, key, and value matrix. We empirically set λ_1 to 0.5 and λ_2 to 0.2 in our experiment.

Evaluation We use answer prediction accuracy as the main metric to evaluate model performance on the provided QA benchmark. The answer classes consists of 42 possible answers (22 objects, 12 counting choices, 6 location types, and yes/no). For training, we use one single model to handle all questions without training separated models for each type. So the accuracy with random choice is 1/42≈2.4%. Additionally, all models are trained on our AVQA dataset using the same features for a fair comparison. We perform comparison of the performance with and without the proposed AV & AG modules in Figure 3 & 4. The table of comparison of the evaluation by the question types is shown in Figure 2.

5 Discussion

In this work, we investigate the audio-visual question answering problem, which aims to answer questions regarding videos by fully exploiting multisensory content. We propose audio-visual agreement module to explore how the effect of each modalities relating to the same question results in answering the question in a bigger scope. Our results show a strong performance in answering simple question such as the existence of musical instrument, suggesting the potential of this work to fit in a single-objective task. We believe this work answers the ambiguity with regards to providing sufficient information would not always lead to a better result.

Limitation Although the proposed agreement module improved the performance on the evaluation sector specific to existential questions, the rest of the

questions presented severe decrease in performance. This may signify that not one general objective function may not work for in all scenarios. For example, temporal understanding obtained the lowest mark, suggesting the lack of information to model the temporal consistency across the frames. More advanced model that could bridge the temporal association across modalities is expected to boost performance further.

Future Work Understanding the limitation from this work, we acknowledge the importance of proper alignment of modalities that crucially affects temporal understanding of the dynamic scene. Future work must therefore address the enhancement of the alignment between modalities to ensure a more balanced and comprehensive understanding across these dimensions. By refining how audio, visual, and other inputs interact and are represented within the model, we hope to improve its ability to handle more complex queries that require reasoning about the both the temporal spatial relations.

References

[1] Moitreya Chatterjee, Jonathan Le Roux, Narendra Ahuja, and Anoop Cherian. Visual scene graphs for audio source separation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1204–1213, 2021.

[2] Chuang Gan, Deng Huang, Hang Zhao, Joshua B Tenenbaum, and Antonio Torralba. Music gesture for visual sound separation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10478–10487, 2020.

[3] Ruohan Gao and Kristen Grauman. Co-separating sounds of visual objects. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3879–3888, 2019.

[4] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[5] Shawn Hershey, Sourish Chaudhuri, Daniel PW Ellis, Jort F Gemmeke, Aren Jansen, R Channing Moore, Manoj Plakal, Devin Platt, Rif A Saurous, Bryan Seybold, et al. Cnn architectures for large-scale audio classification. In *2017 ieee international conference on acoustics, speech and signal processing (icassp)*, pages 131–135. IEEE, 2017.

[6] Guangyao Li, Yake Wei, Yapeng Tian, Chenliang Xu, Ji-Rong Wen, and Di Hu. Learning to answer questions in dynamic audio-visual scenarios. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19108–19118, 2022.

[7] Yinhan Liu. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[8] Yu Wu, Linchao Zhu, Yan Yan, and Yi Yang. Dual attention matching for audio-visual event localization. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6292–6300, 2019.

[9] Yan Xia and Zhou Zhao. Cross-modal background suppression for audio-visual event localization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 19989–19998, 2022.

[10] Ruibin Xiong, Yunchang Yang, Di He, Kai Zheng, Shuxin Zheng, Chen Xing, Huishuai Zhang, Yanyan Lan, Liwei Wang, and Tieyan Liu. On layer normalization in the transformer architecture. In *International Conference on Machine Learning*, pages 10524–10533. PMLR, 2020.

[11] Hanyu Xuan, Zhenyu Zhang, Shuo Chen, Jian Yang, and Yan Yan. Cross-modal attention network for temporal inconsistent audio-visual event localization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 279–286, 2020.

[12] Heeseung Yun, Youngjae Yu, Wonsuk Yang, Kangil Lee, and Gunhee Kim. Pano-avqa: Grounded audio-visual question answering on 360deg videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2031–2041, 2021.