Watch, Listen, and Answer: Open-ended VideoQA with Modulated Multi-stream 3D ConvNets

1st Taiki Miyanishi
ATR, RIKEN AIP
Kyoto, Japan
miyanishi@atr.jp

2nd Motoaki Kawanabe
ATR, RIKEN AIP
Kyoto, Japan
kawanabe@atr.jp

***Abstract*—We propose an open-ended multimodal video question answering (VideoQA) method that predicts textual answers by referring to multimodal information derived from videos. Most current open-ended VideoQA methods focus on motion and appearance features from videos and ignore the audio features that are useful for understanding video content in more detail. A few prior works that use motion, appearance, and audio features showed poor results on public benchmarks since they failed to (e.g., region or grid-level) multimodal features effectively fuse the features with details for video reasoning. We overcame these limitations with multi-stream 3-dimensional convolutional networks (3D ConvNets) and a transformer-based modulator for VideoQA. Our network represents detailed motion and appearance features as well as an audio feature on multiple 3D ConvNets and modulates each intermediate representation with question information to extract their relevant spatiotemporal features over the frames. Based on the question content, our network fuses the multimodal information of 3D ConvNets and predicts the final answers. Our VideoQA method, which effectively combined multimodal data yields, outperformed both a previous multimodal VideoQA method and a state-of-the-art method on standard benchmarks. Visualization suggests that our method can predict the correct answers by listening to the audio information, even when the motion and appearance features are inadequate for understanding the video content.**

***Index Terms*—Video Question Answering**

I. Introduction

Video question answering (VideoQA) is an emerging visual question answering (VQA) task [1] for video contents. Its goal is to return appropriate answers about videos in response to textual questions posed by humans. Recently, various VideoQA methods for both short- and long-form video datasets have been proposed. Most existing methods use clip- and frame-level motion and appearance features and show competitive performance on public VideoQA datasets [2]–[5]. However, these works often inadequately model the detail motion and appearance information, such as the region or grid-level features that represent objects and human motions in the video. Moreover, even though video often contains audio information associated with the motion of objects and people [6], [7], existing works generally ignore it when answering questions. Fig. 1 shows an open-ended VideoQA example that can be answered correctly by simultaneously using motion, appearance, and audio information. The VideoQA system must first recognize the detailed motion and appearance information embedded in the video (e.g., region-level or grid-level features) that represents "a boy" and "playing instruments," posed in the given question. Then the system predicts the answer "flute"

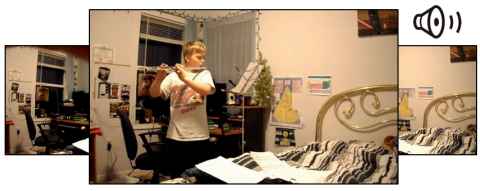

Fig. 1. Example of VideoQA: Each **M**, **V**, and **A** marks indicate using motion, appearance, and audio information for VideoQA. This example could be answered correctly by additionally using audio information, although it failed to answer by only using motion and appearance one.

more correctly by listening to the associated audio information. To generate correct answers with more understanding, integrating multimodal information, including motion, appearance and audio, is essential. Multimodal integration played an important role in other video recognition tasks [8], [9]. However, since few works use audio information for open-ended VideoQA, their performance is relatively low [10] because they fail to use detailed multimodal information and effectively fuse them.

Motivated by these issues, we propose a novel VideoQA architecture that effectively integrates grid-level motion, appearance, and audio features for video reasoning. Our method uses multi-stream three-dimensional Convolutional Networks (3D ConvNets) based on ResNet bottlenecks [11] to represent the grid-level spatiotemporal features of motion, appearance, and audio information. For answering questions about the video, the method individually modulates three-stream bottlenecks using feature-wise affine transformation [12] integrated with transformer-based self-attention blocks [13] that aggregate the temporal information between question and multiple frame- or clip-level features. We also use a controller to perform different modulations to each bottleneck's block guided by the question information. Finally, our method combines multimodal representations based on the question and generates textual answers. We demonstrate our method's effectiveness on short- and long-form open-ended VideoQA datasets, including motion, appearance, and audio information. Our experimental results show that our method significantly outperformed a recent multimodal VideoQA method [10] and a state-of-the-art method [14] that uses detailed motion-appearance features on major question types. The visualization shows the effectiveness of using audio information.

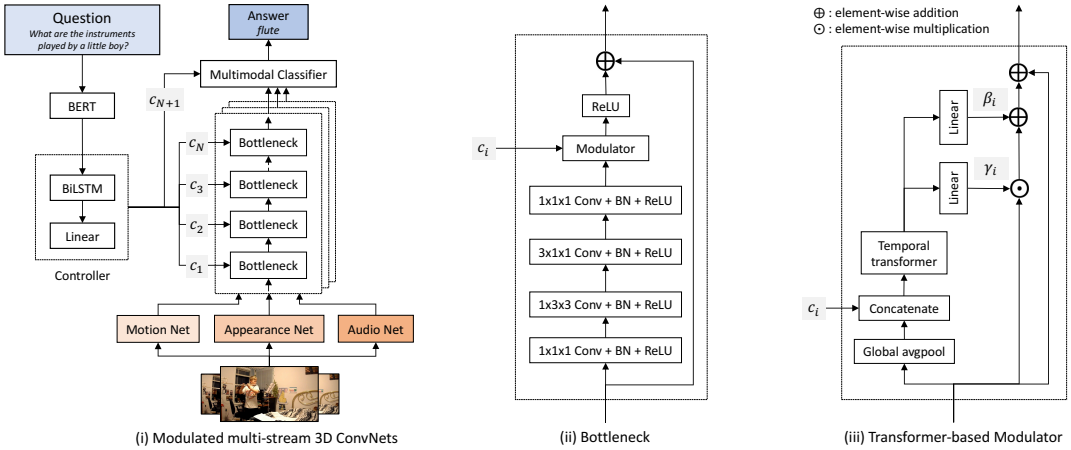

Fig. 2. Bottlenecks take as input multimodal video features (i.e., motion, appearance, audio) and generate intermediate representations. Each bottleneck has a modulation layer that manipulates features in bottleneck based on a control state updated by a controller. Multimodal classifier computes most probable answers based on pooled features of multi-stream bottlenecks.

II. Related Work

Video question answering. VideoQA is a natural extension of image-based visual question answering (VQA) [1] to the video domain. This challenging task requires both language and video understanding. Many existing works use the attention mechanism [15] to find important frames (or clips) relevant to a given question [16], [17]. Recently some works use a self-attention mechanism [13] for capturing temporal relationships over frames (or clips), which helps find relevant events across time [3], [18]. Our transformer-based modulation layer also uses a self-attention mechanism to capture temporal relationships over frames used to modulate spatiotemporal features. Unlike the standard VQA that targets static images, VideoQA targets a video that contains dynamics information, e.g., motions of objects and humans. To capture the dynamics information of videos, existing VideoQA approaches use frame- or clip-level motion-appearance information [5], [19]. Some works simultaneously use detailed appearance and motion features for video reasoning [14]. In contrast, our method uses audio features for VideoQA in addition to the detailed motion and appearance features. We assume that audio information has important clues that cannot be captured by motion and appearance information.

Audio-visual video recognition. For understanding video contents, many attempts have explored the integration of the audio and visual information of videos, such as video retrieval [20], recognizing activities [9], captioning events [8], [21], and visual dialogue [22], [23]. Even though numerous video understanding tasks have used both visual and audio information, some prior works on the open-ended VideoQA task [10] use textual descriptions and audio features as well as visual and motion features. Unfortunately, that study shows that the audio did not improve the VideoQA performance because it failed to effectively fuse the multimodal data. In contrast, our method can further improve the VideoQA performance more than methods that use uni-modal data by modeling the detailed spatiotemporal information of the multimodal data.

III. Network Architecture

Given a video and a textual question, VideoQA method's goal is to predict an answer that matches the correct one. In this this section, we introduce multi-stream 3D ConvNets (M3DC) for open-ended VideoQA. Fig. 2 (left) shows the overall architecture of our proposed network, which mainly consists of following bottlenecks, a controller, a modulator, and a multimodal classifier.

Multi-stream bottlenecks. For learning detailed spatiotemporal visual representation in videos, we use the bottleneck of 3D ResNets [24] as a template. Fig. 2 (center) shows our bottleneck, which contains four Conv layers and one modulation layer with a residual connection. Each convolutional layer is followed by batch normalization (BN) [25] for normalizing the layer inputs and a rectified linear unit (ReLU) for nonlinearity. We use the P3D style bottleneck [26] to reduce the computational cost. First, the $1 \times 1 \times 1$ Conv layer reduces the D channel dimensions to $D/2$. Second, the $1 \times 3 \times 3$ Conv layer aggregates the spatial information in a video frame. Third, the $3 \times 1 \times 1$ Conv layer aggregates the temporal information over the video frames. Fourth, the $1 \times 1 \times 1$ Conv layer recovers the channel dimensions followed by the modulation layer that regulates the aggregated spatiotemporal information with conditioning information.

Controller. The representation of bottlenecks should be refined with each layer of stacks. To this end, we used a lightweight, practical controller proposed in a visual reasoning task [27] and performed different modulations on each bottleneck. First, we encoded a given question using a pre-trained, 12-layer BERT model [28] and extracted word feature vector $w \in \mathbb{R}^{768}$ from the last layer, which is fine-tuned during the VideoQA training. The controller encodes the extracted word features using a one-layer bi-directional LSTM (BiLSTM) and uses a series of its output states $\{\boldsymbol{cw}_i\}_{i=1}^{M} \in \mathbb{R}^{D}$ as question word embeddings, where M is the number of words in a question and D is the input channels of the bottleneck.

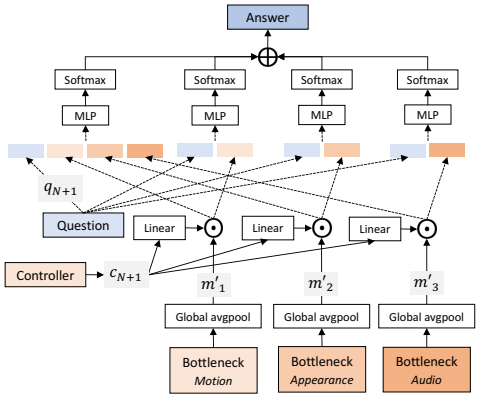

Fig. 3. Overview of multimodal classifier: Classifier gates bottleneck's outputs pooled in both spatiotemporal dimensions using control state and selects most probable answer from candidates.

For a question sentence embedding, the controller uses the final hidden states from the backward and forward LSTMs as $\boldsymbol{q} = [\boldsymbol{h}_b; \boldsymbol{h}_f] \in \mathbb{R}^{2D}$, where ; denotes horizontal vector concatenation. For layer-wise modulation, the controller applies a linear transformation to $\boldsymbol{q}$ and generates $\{\boldsymbol{q}_i\}_{i=1}^{N} \in \mathbb{R}^D$, where N is the number of bottlenecks in each stream. We also make an additional question embedding feature, $\boldsymbol{q}_{N+1} \in \mathbb{R}^D$, for gating the multimodal representation from each stream using question information. The controller produces control state $\boldsymbol{c}_i \in \mathbb{R}^D$ for modulating the i^{th} bottleneck as follows:

$$\begin{aligned} \boldsymbol{cq}_i &= \text{Linear}([\boldsymbol{c}_{i-1}; \boldsymbol{q}_i]), \quad ca_{i,m} = \text{Linear}(\boldsymbol{cq}_i \odot \boldsymbol{cw}_m), \\ cv_{i,m} &= \text{Softmax}(ca_{i,m}), \quad \boldsymbol{c}_i = \sum_{m=1}^{M} cv_{i,m} \cdot \boldsymbol{cw}_m, \end{aligned} \tag{1}$$

where $\text{Linear}(\cdot)$ is the linear projection layer, $\odot$ denotes element-wise multiplication, and $\boldsymbol{cq}_i \in \mathbb{R}^D$ and $ca_{i,m} \in \mathbb{R}$ denote the intermediate representations to calculate the attention value of m^{th} word $cv_{i,m} \in \mathbb{R}$.

Modulator. For modulating bottlenecks conditioned on a question, we extend the feature-wise linear modulation (FiLM) [12] specialized in VideoQA. Fig 2 (right) shows our extended FiLM modulation layer. In general, a FiLM layer is used to influence the intermediate representation of a bottleneck based on the conditioning information by scaling and shifting:

$$\text{FiLM}(\boldsymbol{X}_{i,j}|\boldsymbol{\gamma}_{i,j}, \boldsymbol{\beta}_{i,j}) = \boldsymbol{\gamma}_{i,j}\boldsymbol{X}_{i,j} + \boldsymbol{\beta}_{i,j}, \tag{2}$$

where $\boldsymbol{X}_i \in \mathbb{R}^{h \times w \times D}$ ($\boldsymbol{X}_{i,j} \in \mathbb{R}^{h \times w}$) denotes the activation outputs of a previous layer that has h height, w width, and D channels, $\boldsymbol{\gamma}_i = \text{Linear}(\boldsymbol{x}_i) \in \mathbb{R}^D$ and $\boldsymbol{\beta}_i = \text{Linear}(\boldsymbol{x}_i) \in \mathbb{R}^D$ are modulation parameters, and $\boldsymbol{x}_i$ is the conditioning information for the i^{th} block modulation. The vanilla FiLM layer, which is mainly designed for static images, cannot consider the interaction between spatiotemporal features across frames even though the videos consist of a sequence of frames. To leverage the spatiotemporal information over video frames, we used a multi-layer transformer [13] with a self-attention mechanism that can relate various frame sequence positions. The transformer block is defined as follows:

$$\begin{aligned} \text{TransformerBlock}(\boldsymbol{V}) &= \text{LayerNorm}(\boldsymbol{V}' + \text{FFN}(\boldsymbol{V}')), \\ \boldsymbol{V}' &= \text{LayerNorm}(\boldsymbol{V} + \text{MultiHeadAtten}(\boldsymbol{V})), \end{aligned}$$

where $\boldsymbol{V} \in \mathbb{R}^{K \times 2D}$ is a sequence of $\boldsymbol{V}_k \in \mathbb{R}^{2D}$, which is the concatenation of control state $\boldsymbol{c}$ and the k^{th} frame-level feature that applies global average pooling in the spatial dimension to the previous 3D Conv layer's outputs, K is the number of frames, FFN is row-wise feed-forward networks, LayerNorm is a layer normalization [29], and MultiHeadAtten is a multi-head attention block [13] that performs self-attention to the input sequence of the features. We refer to this temporal transformer-based FiLM as T-FiLM. We use the k^{th} frame-level output of the T-FiLM ($\boldsymbol{v}_k \in \mathbb{R}^{2D}$) as the conditioning information for modulating the spatiotemporal features of each frame.

Multimodal classifier. Our VideoQA model predicts answers using a multimodal classifier. Fig. 3 shows its overview. First, the classifier gates output $\boldsymbol{m'}_j \in \mathbb{R}^D$ from the j^{th} modality bottleneck pooled in both the spatial and temporal dimensions using last step control state $\boldsymbol{c}_{N+1}$:

$$\boldsymbol{m}_j = \text{Sigmoid}(\text{Linear}(\boldsymbol{c}_{N+1})) \odot \boldsymbol{m'}_j, \tag{3}$$

where gated output $\boldsymbol{m}_j \in \mathbb{R}^D$. The multimodal classifier predicts answers using multimodal joint training [30] to avoid overfitting the multimodal VideoQA model. To compute the final answer, we use a simple classifier with a multi-layer perceptron (MLP) that takes as input the question and the gated output of the multi-stream bottleneck:

$$\boldsymbol{o'}_s j = \text{Linear}([\boldsymbol{q}_{N+1}; \boldsymbol{m}_j]), \quad \boldsymbol{o}_j = \text{Softmax}(\text{GELU}(\text{Linear}(\boldsymbol{o'}_j))),$$

where GELU is a Gaussian error linear unit [31], $\boldsymbol{o'}_j \in \mathbb{R}^D$, and $\boldsymbol{o}_j \in \mathbb{R}^{|A|}$ denotes the softmax scores of answer candidates A. Finally, we select the answer with the highest values based on the scores' ensemble from all the modalities.

IV. Experimental Settings

Datasets. We evaluated our method using two public open-ended VideoQA datasets, MSRVTT-QA [16] and ActivityNet-QA [2], which both contain audio and visual contents. MSRVTT-QA is a short-form VideoQA dataset (avg. video length is 15 sec.) that has five question types: what, who, how, when, and where. ActivityNet-QA is a long-form VideoQA dataset (avg. video length is 116 sec.) is relatively balanced. In contrast to the superficial question categorization of MSRVTT, ActivityNet-QA's questions are semantically categorized into four main question types: motion, spatial relationship, temporal relationship, and free. Free questions have six sub-question types: yes/no, number, color, object, location, and other.

Features. So that MotionNet can extract grid-level motion features ($\boldsymbol{f}^m \in \mathbb{R}^{h_m \times w_m \times 2304}$) in Fig. 2, we used the SlowFast networks [32], which are commonly employed for video recognition tasks, where $h_m \times w_m \times 2304$ is the size of the SlowFast networks' feature map. So that AppearanceNet can extract grid-level visual appearance features

TABLE I
MSRVTT-QA RESULTS

	What 49,869	Who 20,385	How 1,640	When 677	Where 250	All 72,821
ESA [2], [16]	0.220	0.416	0.796	0.731	0.332	0.293
ST-VQA [36]	0.245	0.412	0.780	0.765	0.349	0.309
HME [5]	0.265	0.436	0.824	0.760	0.286	0.330
CAN [4]	0.267	0.434	**0.837**	0.753	0.352	0.332
HCRN [3]	0.295	0.451	0.821	0.783	0.344	0.355
TS-STMAC [14]	0.336	0.488	0.831	**0.786**	0.336	0.394
Ours: M3DC	**0.346**	**0.516**	0.800	0.780	**0.376**	**0.408**

TABLE II
ACTIVITYNET-QA RESULTS

	Motion 800	Spatial 800	Temporal 800	Free 5600	All 8,000
ESA [2], [16]	0.125	0.144	0.025	0.412	0.318
HME [5]	0.174	0.159	0.023	0.423	0.331
CAN [4]	0.211	0.173	0.036	0.445	0.354
MAR [10]	0.158	0.159	0.026	0.443	0.344
HCRN [3]	0.215	0.171	0.031	0.457	0.362
TS-STMAC [14]	0.355	0.183	0.039	0.492	0.402
Ours: M3DC	**0.374**	**0.209**	**0.050**	**0.497**	**0.411**

($\boldsymbol{f}^v \in \mathbb{R}^{h_v \times w_v \times 2048}$), we used a pre-trained model for the grid features [33], where $h_v \times w_v \times 2048$ is the size of this network's feature map. To reduce storage space and computational costs, we used principal component analysis to lower the motion and appearance features' dimensions to 128. So that AudioNet can extract audio features, we used a pre-trained model of PANNs [34] and extracted audio features $\boldsymbol{f}^a \in \mathbb{R}^{2048}$ from 2.56-sec. video clips. We used $K = 20$ frames at even intervals to extract these features. For answer candidates A, we used the top 1,000 most frequent answers in a training split.

Training details. We trained our method using RADAM [35] for optimization with a learning rate of $\alpha = 0.0003$ and a batch size of 32. We stopped the training to mitigate overfitting when the validation accuracy did not increase for ten epochs. We converted the words in the questions and answers to lower cases and set 3D ConvNets $D = 256$. We set the number of network's blocks to $N = 4$. Learning rate α and dimension D were selected from candidates $\{0.0001, 0.0002, 0.0003\}$ and $\{128, 256, 512\}$ based on the validation set.

V. EXPERIMENTS

We evaluated our proposed method by comparing it to the current VideoQA methods. Since the number of questions in some question types is relatively small, we report the accuracy for all the questions and each question type with the number of instances. We used the reported accuracy of original papers or additional experiments using the existing methods' public codes.

Comparison to state-of-the-art. We show the VideoQA performance on MSRVTT-QA in Table I. Our method, M3DC, achieved an overall accuracy of 0.408 and outperformed all the existing methods. In particular, it outperformed the latest method, TS-STMAC [14], on the question types with many instances (i.e., what and who). Tables II and III show the performance of the main and sub-question types of ActivityNet-QA. Our proposed method outperformed the others for all the main question types and achieved the best accuracy of 0.411. In particular, across all question types, our method significantly outperformed the existing multimodal open-ended VideoQA method MAR [10], which simultaneously uses motion, appearance, and audio features. Also, our method highly outperformed the existing SoTA method on the spatial and temporal question types and is 14% and 28% better than TS-STMAC [14]. Our method is competitive or outperformed the other methods for the sub-question types of free-form questions (Table III). Because TS-STMAC uses the extracted features of Faster-RCNN trained on the Visual Genome dataset, which contains the object's bounding boxes and color attribute labels, these features did well for object and color question types.

TABLE III
RESULTS ON FREE TYPE QUESTIONS OF ACTIVITYNET-QA

	Yes/No 2,094	Color 697	Object 318	Location 386	Number 606	Other 1,499
ESA [2], [16]	0.594	0.298	0.142	0.259	0.446	0.284
HME [5]	0.607	0.304	0.132	0.277	0.475	0.297
CAN [4]	0.626	0.311	0.201	0.306	0.480	0.333
MAR [10]	0.645	0.311	0.195	0.295	0.446	0.310
HCRN [3]	0.657	0.316	0.220	0.298	0.454	0.336
TS-STMAC [14]	0.683	**0.364**	**0.258**	0.316	0.500	0.376
Ours: M3DC	**0.685**	0.356	0.255	**0.365**	**0.507**	**0.381**

TABLE IV
ABLATION STUDIES OF OUR METHOD M3DC

	MSRVTT-QA	ActivityNet-QA
Number of layers		
$N = 2$	0.407	0.394
$N = 3$	0.405	0.401
$N = 5$	0.406	0.398
Layer-size		
$D = 64$	0.380	0.379
$D = 128$	0.396	0.394
Text Embedding		
Glove	0.394	0.402
Modality		
audio only	0.337	0.327
motion only	0.378	0.386
visual only	0.380	0.383
motion + visual	0.399	0.403
Classifier		
w/o modality ensemble	0.399	0.402
w/o modality gate	0.399	0.399
Modulation		
w/o temporal transformer	0.409	0.395
w/o controller	0.400	0.402
Full	0.408	0.411

Ablation study. To validate the effectiveness of each component of our method, we conducted ablation studies on both VideoQA datasets with various settings. Table IV shows the ablation results. Our multimodal method outperformed the uni-modal methods. The results indicate that modality gating and an ensemble of classifiers' scores are effective. Our proposed T-FiLM is also useful for long-form VideoQA. We found the controller plays a vital role in improving VideoQA performance.

Qualitative results. Figure 4 shows the representative VideoQA results. Our method effectively fused the motion,

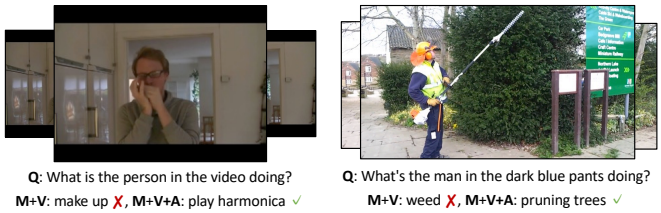

Fig. 4. Qualitative results: **M**, **V**, and **A** marks indicate using motion, appearance, and audio information for VideoQA.

appearance, and audio information for VideoQA. The case on the left shows that motion and appearance information is insufficient to answer the question because the instrument being held is hard to see, and the man does not move much. However, our model answered correctly by listening to the audio from a harmonica. The right case shows that our model predicted the correct answer, "pruning trees," by listening to the pruning machine's audio information, even when motion and appearance were insufficient to understand the video content.

VI. Conclusion

We developed a multimodal video question answering method that predicts answers using motion, appearance, and audio features. We evaluated our modulated multi-stream 3D ConvNets and described improvements on two standard VideoQA datasets and all the modalities' complementary effects.

Acknowledgment

This work was supported by JSPS KAKENHI Grant Number JP18KK0284 and JST CREST Grant Number JPMJCR15E2.

References

[1] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh, "VQA: Visual Question Answering," in *ICCV*, 2015, pp. 2425–2433.

[2] Zhou Yu, Dejing Xu, Jun Yu, Ting Yu, Zhou Zhao, Yueting Zhuang, and Dacheng Tao, "ActivityNet-QA: A dataset for understanding complex web videos via question answering," in *AAAI*, 2019, pp. 9127–9134.

[3] Thao Minh Le, Vuong Le, Svetha Venkatesh, and Truyen Tran, "Hierarchical conditional relation networks for video question answering," in *CVPR*, 2020, pp. 9972–9981.

[4] Ting Yu, Jun Yu, Zhou Yu, and Dacheng Tao, "Compositional attention networks with two-stream fusion for video question answering," *IEEE Transactions on Image Processing*, vol. 29, pp. 1204–1218, 2020.

[5] Chenyou Fan, Xiaofan Zhang, Shu Zhang, Wensheng Wang, Chi Zhang, and Heng Huang, "Heterogeneous memory enhanced multimodal attention model for video question answering," in *CVPR*, 2019, pp. 1999–2007.

[6] Hang Zhao, Chuang Gan, Andrew Rouditchenko, Carl Vondrick, Josh McDermott, and Antonio Torralba, "The sound of pixels," in *ECCV*, 2018.

[7] Hang Zhao, Chuang Gan, Wei-Chiu Ma, and Antonio Torralba, "The sound of motions," in *ICCV*, 2019, pp. 11735–1744.

[8] Vladimir Iashin and Esa Rahtu, "A better use of audio-visual cues: Dense video captioning with bi-modal transformer," in *BMVC*, 2020.

[9] Fanyi Xiao, Yong Jae Lee, Kristen Grauman, Jitendra Malik, and Christoph Feichtenhofer, "Audiovisual slowfast networks for video recognition," *arXiv:2001.08740*, 2020.

[10] Yueting Zhuang, Dejing Xu, Xin Yan, Wenzhuo Cheng, Zhou Zhao, Shiliang Pu, and Jun Xiao, "Multichannel attention refinement for video question answering," *ACM Transactions on Multimedia Computing, Communications, and Applications*, vol. 16, no. 1s, pp. 1–23, Mar. 2020.

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *CVPR*, 2016, pp. 770–778.

[12] Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron C. Courville, "Film: Visual reasoning with a general conditioning layer," in *AAAI*, 2018, pp. 3942–3951.

[13] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin, "Attention is all you need," in *NeurIPS*, 2017, pp. 5998–6008.

[14] Taiki Miyanishi, Takuya Maekawa, and Motoaki Kawanabe, "Two-stream spatiotemporal compositional attention network for videoqa," in *BMVC*, 2020.

[15] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio, "Neural machine translation by jointly learning to align and translate," in *ICLR*, 2015.

[16] Dejing Xu, Zhou Zhao, Jun Xiao, Fei Wu, Hanwang Zhang, Xiangnan He, and Yueting Zhuang, "Video question answering via gradually refined attention over appearance and motion," in *ACMMM*, 2017, pp. 1645–1653.

[17] Zhu Zhang, Zhou Zhao, Zhijie Lin, Jingkuan Song, and Xiaofei He, "Open-ended long-form video question answering via hierarchical convolutional self-attention networks," in *IJCAI*, 2019, pp. 4383–4389.

[18] Pin Jiang and Yahong Han, "Reasoning with heterogeneous graph alignment for video question answering," in *AAAI*, 2020, pp. 11109–11116.

[19] Jiyang Gao, Runzhou Ge, Kan Chen, and Ram Nevatia, "Motion-appearance co-memory networks for video question answering," in *CVPR*, 2018, pp. 6576–6585.

[20] Valentin Gabeur, Chen Sun, Karteek Alahari, and Cordelia Schmid, "Multi-modal Transformer for Video Retrieval," in *ECCV*, 2020.

[21] Tanzila Rahman, Bicheng Xu, and Leonid Sigal, "Watch, listen and tell: Multi-modal weakly supervised dense event captioning," in *ICCV*, 2019, pp. 8908–8917.

[22] C. Hori, H. Alamri, J. Wang, G. Wichern, T. Hori, A. Cherian, T. K. Marks, V. Cartillier, R. G. Lopes, A. Das, I. Essa, D. Batra, and D. Parikh, "End-to-end audio visual scene-aware dialog using multimodal attention-based video features," in *ICASSP*, 2019, pp. 2352–2356.

[23] Idan Schwartz, Alexander G. Schwing, and Tamir Hazan, "A simple baseline for audio-visual scene-aware dialog," in *CVPR*, 2019, pp. 12548–12558.

[24] Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh, "Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet?," in *CVPR*, 2018, pp. 6546–6555.

[25] Sergey Ioffe and Christian Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *ICML*, 2015, pp. 448–456.

[26] Zhaofan Qiu, Ting Yao, and Tao Mei, "Learning spatio-temporal representation with pseudo-3d residual networks," in *ICCV*, 2017, pp. 5533–5541.

[27] Drew A Hudson and Christopher D Manning, "Compositional attention networks for machine reasoning," in *ICLR*, 2018.

[28] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *NAACL-HLT*, 2019, pp. 4171–4186.

[29] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton, "Layer normalization," *arXiv:1607.06450*, 2016.

[30] Weiyao Wang, Du Tran, and Matt Feiszli, "What makes training multi-modal classification networks hard?," in *CVPR*, 2020, pp. 12695–12705.

[31] Dan Hendrycks and Kevin Gimpel, "Gaussian Error Linear Units (GELUs)," *arXiv:1606.08415*, 2016.

[32] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He, "Slowfast networks for video recognition," in *ICCV*, 2019, pp. 6202–6211.

[33] Huaizu Jiang, Ishan Misra, Marcus Rohrbach, Erik Learned-Miller, and Xinlei Chen, "In defense of grid features for visual question answering," in *CVPR*, 2020, pp. 10267–10276.

[34] Qiuqiang Kong, Yin Cao, Turab Iqbal, Yuxuan Wang, Wenwu Wang, and Mark D. Plumbley, "Panns: Large-scale pretrained audio neural networks for audio pattern recognition," *arXiv:1912.10211*, 2020.

[35] Liyuan Liu, Haoming Jiang, Pengcheng He, Weizhu Chen, Xiaodong Liu, Jianfeng Gao, and Jiawei Han, "On the variance of the adaptive learning rate and beyond," in *ICLR*, April 2020.

[36] Yunseok Jang, Yale Song, Youngjae Yu, Youngjin Kim, and Gunhee Kim, "TGIF-QA: Toward spatio-temporal reasoning in visual question answering," in *CVPR*, 2017, pp. 2758–2766.