Special Paper
방송공학회논문지 제28권 제7호, 2023년 12월 (JBE Vol. 28, No. 7, December 2023)
https://doi.org/10.5909/JBE.2023.28.7.849
ISSN 2287-9137 (Online) ISSN 1226-7953 (Print)

Video Question Answering with Overcoming Spatial and Temporal Redundancy in Feature Extraction

Ju-Hee Lee[a)], Seong Jong Ha[b)], and Je-Won Kang[a)‡]

Abstract

The current video question answering (video QA) models tend to produce inaccurate results, when key video features are obscured by a large amount of redundant video data. In this paper, we newly define the problems as “spatial redundancy” and “temporal redundancy” in video QA and develop an effective appearance and motion features to resolve the drawbacks. We generate a motion feature from adjacent appearance features to distinguish meaningful events in adjacent frames. Further, question-to-video attention is applied to consider the inter-modal correlation during feature extraction to focus on more relevant features. For benchmark tests, we create MSVD dynamic QA dataset to include various motions and scene changes, by sampling video clips from the MSVD QA dataset. The performance of video QA methods can be evaluated when the test videos have different temporal dynamics. The proposed method is adaptable to state-of-the-art networks, and it is model-agnostic and end-to-end trainable. Experimental results demonstrate that the proposed method provides a superior performance to a baseline model both in the MSVD dynamic QA dataset and the original MSVD dataset.

Keyword : Video question answering, attention model, multi modal data learning

Ⅰ. Introduction

Video question answering (video QA) is a computer vision task to answer a question through semantic reasoning with an input video[1,2]. In video QA, handling large amounts of both visual and linguist information plays an important role in improving performance. Recent studies use convolutional neural network (CNN) such as VGG[3] and C3D[4] networks to extract an appearance feature and a motion feature, respectively. The features are fused with a language feature to answer a question, using a long short-term memory (LSTM)[5]. Based on the CNN and LSTM models, researchers have introduced attention models to focus on more relevant local regions and video intervals[6,7,8]. Spatial attention and temporal attention have been proposed to find a region of interest and temporally local-

a) Department of Electronic and Electrical Engineering and Graduate Program in Smart Factory, Ewha W. University, Seoul Korea
b) CJ AI
‡ Corresponding Author : Je-Won Kang
E-mail: jewonk@ewha.ac.kr
Tel: +82-2-3277-2347
ORCID: https://orcid.org/0000-0002-1637-9479
※ This research was supported by the NRF grant funded by MSIT (No. NRF-2022R1A2C4002052)
· Manuscript September 6, 2023; Revised October 31, 2023; Accepted October 31, 2023

ize the subject in each frame[9] - [12]. However, they had limitations to understand global contexts in a video[13,14]. Later, memory modules were used to keep attended spatial and temporal features[7,8]. In this manner, a question could be answered after a model examines a whole video. However, it was difficult to focus on a region of interests among many input video frames. Further, inter-domain correlation has been largely ignored without carefully examining key information.

We present two problematic scenarios caused by the inherent nature of a video as shown in Figure 1. "Temporal redundancy" refers to the problem that a model is more likely to learn the frequent motion and reflectively answer the question, ignoring a key action. For example, in Figure 1, the model answers "Cooking" instead of the ground-truth "Standing" when the question "What is a woman who wearing a blue skirt doing?" is given. This is because the model has linked "Woman" and "Cooking" as the most occurring motion in the part of the video. Second, "Spatial redundancy" happens when the model chooses an answer which is primary within all scenes but wrong, where the model has learned dominant appearance unrelated to the question. When the question "What is a woman wearing a blue skirt grabbing?" is given, the video presents the "Onion" in most scenes that are less relevant than the answer "Hands" in each scene. Therefore, video data can be learned with some inherent biases caused by a superficial inter-domain correlation. As a result, the model predicts a wrong answer without carefully looking at any key information.

This paper is the first attempt to define and address the problems due to the large amount of redundant video data. The performance of a video QA model can be significantly degraded by the nature of video and drawbacks of conventional approaches. Although the previous studies applied attention mechanisms[6] - [8],[12], the features had disadvantages because of insufficient representation capability to express diverse scenes in a video. Therefore, we develop

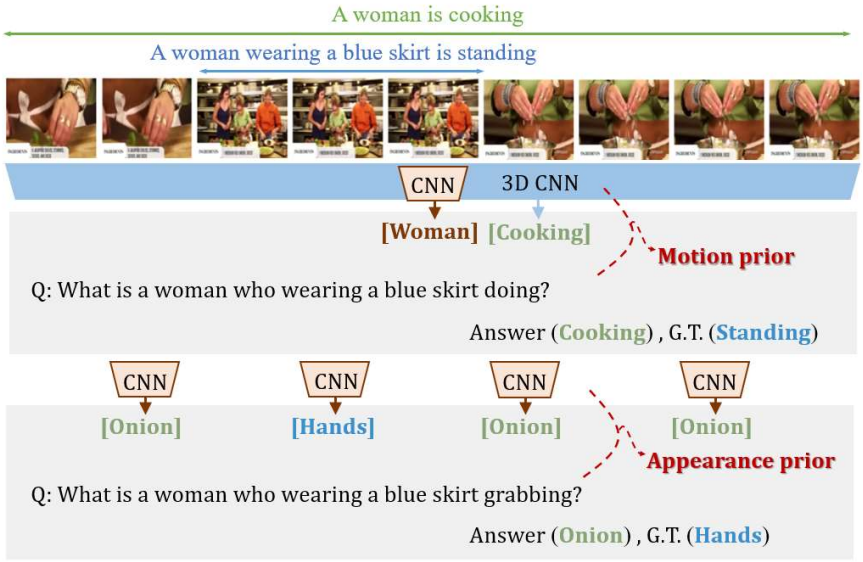


Fig. 1. Our motivation is presented. Video QA models can suffer two problems in video. A large amount of redundant information of motion and appearance can degrade the performance of a model due to the nature of video and drawbacks of conventional approaches.

a novel deep feature extraction to cope with the problem. Our work has several primary contributions as follows:

We propose a deep video QA model to provide more reliable video features although video data presents various characteristics. Motion features are generated from surrounding appearance features to distinguish meaningful temporal changes among other frequent motions in frames.

We develop a MSVD dynamic video QA test dataset. The proposed dataset is used for benchmark tests to measure the different characteristics of videos with various temporal dynamics. The dataset is more difficult to respond by including detailed questions with several scene changes in the video.

Ⅱ. Related Works

1. Visual question answering

Most previous works exploited high-level correlations between visual and textual information. To capture multi-modal interaction, Fukui et al.[15] proposed a pooling method to combine the visual and textual representations. In [16], a combined bottom-up and top-down attention mechanism was used to generate features. Video QA is a more challenging task due to motion or causality in the temporal dimension. Several temporal features have been developed to find relevant events to queries[2],[18].

Some researchers exploited multi-modal features. Object aware temporal attention was presented to learn appearance-question interactions[19]. A spatio-temporal relational network was developed to treat temporal changes among different objects[20]. Object features were used with a graph structure to enhance co-attention between the appearance and question[21]. However, the video feature was delivered from the hidden layer of the LSTM, making it difficult to effectively utilize the temporal associations among the previous frames.

2. Attention and memory module

Temporal attention was proposed to exploit temporal correlation in a video, and it was extended to use spatial and temporal attention[12],[18]. In [18], a hierarchical dual-level attention network was proposed to obtain the question aware video representation. In addition, language-guided attention and video-guided methods have been developed to exploit multi-modal information[6, 22]. In [1], an end-to-end learning model was proposed to uses question for gradually refining its temporal attention. Self-attention methods have been introduced to to identify the importance of questions and videos in video QA [25],[26]. Memory modules have been used for attending features[7],[8],[17]. In [7], a memory module was used for combining motion and appearance features in co-memory attention. In [8], a shared memory module was used to learn global video features. However, the existing methods still have not explicitly addressed the forgotten features by spatial and temporal redundancy in natural videos.

Ⅲ. Proposed Method

Figure 2 presents the proposed neural network architecture based on the conventional modules[2,12,18] colored with gray. The proposed network consists of a motion feature extraction (blue) and a question-to-video attention during feature extraction (red). The attention module takes inputs from the question attention (green) and produces more relevant motion and appearance features denoted by $\bar{a}$ and $\overline{m}$ before encoding by LSTM. The process is designed to effectively remove the spatial and temporal redundancy during feature extraction.

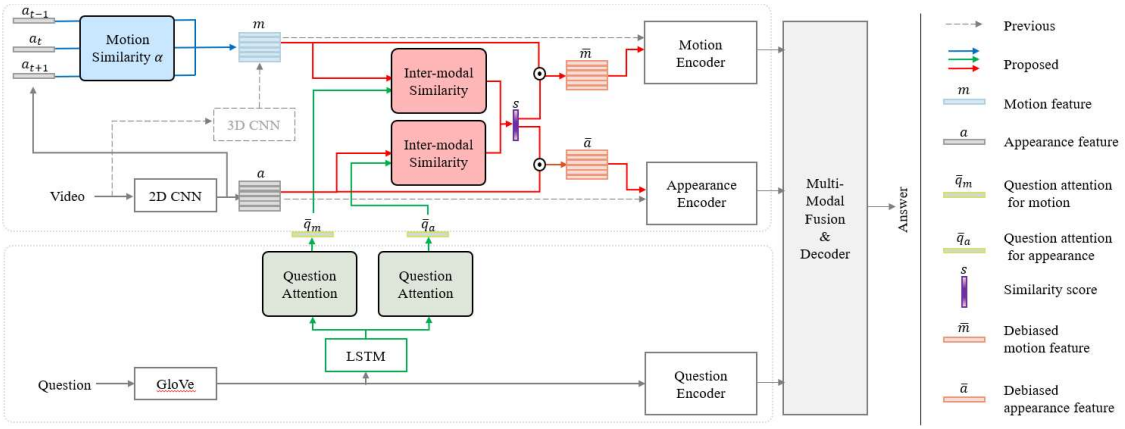


Fig. 2. A blockdiagram of the proposed video QA network architecture on top of the conventional modules colored with gray. The motion feature extraction replacing the 3D CNN is colored with blue. Question-to-video attention is colored with red. Question attention is colored with green.

1. Appearance and motion feature extraction

In the proposed method, the conventional 3D CNN has been removed, because it was used for homogeneous motion video. It hardly captures local motion with some scene changes. Instead, a motion feature is generated with adaptive weights of input frames by computing the similarity to adjacent appearance features. If there are substantial temporal changes among adjacent appearance features, the motion feature is extracted from few key frames near the current appearance feature. Otherwise, it is extracted by examining more frames. We explain more details as follows.

An appearance feature a_t is extracted using a VGG-16 network, and N consecutive frames appearance features around the current time t are used for input. Once the appearance features are ready, we calculate the similarity c_t as $c_t = F_s(a_t, a_{t+r})$, where r is a time index, and we use a cosine similarity function for F. For normalization we use a softmax function. Then, a temporal feature ϵ_t is generated as follows:

$$\epsilon_t = \sum_{-N/2}^{N/2} \alpha_t a_{t+r} \quad \cdots \quad (1)$$

where α_t is a normalized one from c_t. It is ensured that the features are more synchronized by the current time as follows:

$$m_t = (a_t + \epsilon_t)/2 \quad \cdots \quad (2)$$

where m_t is calculated in all the sampled appearance frame, but the complexity increases only slightly because it uses a pertained CNN.

2. Question-to-video attended feature

In the proposed method, a_t and m_t are enhanced by measuring the correlation with a given query. To determine the correlation, we propose to use $\overline{q_a}$ and $\overline{q_m}$ that are question attentions for appearance and motion and use question-to-video attention deep module to output attended feature $\overline{a_t}$ and $\overline{m_t}$ given as

$$\overline{a_t},\ \overline{m_t} = A(a,\ m,\ \overline{q_a},\ \overline{q_m}) \quad \cdots \quad (3)$$

where A is the deep model to generate the question-to-video attended features. The attention is applied

during feature extraction to efficiently capture useful motion and appearance features among spatial and temporal abundant information. We first explain question attention and then the generation process of the proposed features from question-to-video attention. We used a two-layers LSTM[5] to create motion and appearance query.

Next, an inter-modal similarity score is calculated to consider the other modality when training an unimodal feature. We first calculate the appearance and motion similarity scores s_a and s_m. The scores are obtained from the video features a and m and the attention $\overline{q_a}$ and $\overline{q_m}$ as

$$s_a = softmax(\overline{q_a} W_s a) \quad \ldots \quad (4)$$

and

$$s_m = softmax(\overline{q_m} W_s m) \quad \ldots \quad (5)$$

where W_s is a weight similarity matrix[27] with learnable parameters, and we use the SoftMax function to compute the probability. The scores are end-to-end learnable without ground-truth. We note that the calculation of the similarity can give a similar effect to an intermediate fusion[28]. Next, an inter-modal similarity score s to have more intermodal synchronization is calculated as

$$s = \frac{s_a + s_m}{2} \quad \ldots \quad (6)$$

Because s plays a role in highlighting the spatial and temporal regions associated with a question, it can activate different areas for each question even for the same video. For example, for the first question in Figure 3, the frames around t = 1 and N are more attended due to more relevant objects such as “man”. Then, we multiply the similarity scores to the video features a_t and m_t to extract the enhanced feature $\bar{a}$ and $\overline{m}$ as the dot products with s.

3. Loss function

We set the loss function L to train whole network as follows:

$$L = \delta_1 L_1 + \delta_2 L_2, \quad \ldots \quad (7)$$

where L_1 is a cross entropy loss for an open-end question and δ_1=0.9 and δ_2=0.1. In addition, we propose to use L_2 to impose restrictions to the larger variance of the inter-modal scores by

$$L_2 = 1 - \frac{1}{N}\sum_{j=1}^{N}(s_j - \mu), \quad \ldots \quad (8)$$

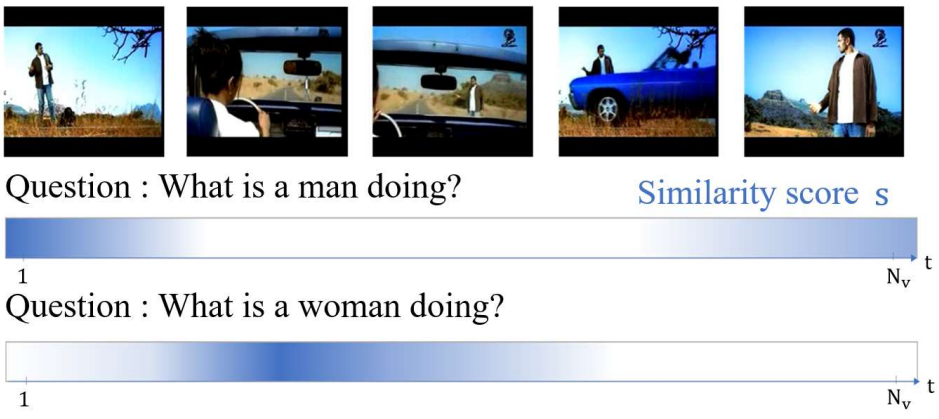


Fig. 3. Visualization of inter-modal similarity score s. The region colored with blue stands for the strong attention

where μ is the average of s_j.

IV. Experimental Results

1. MSVD and MSVD Dynamic QA Dataset

The original MSVD QA dataset[23] includes 520 video clips, in which the majority of the videos display homogeneous motion. There are 71.8 % of videos with noticeable temporal changes less than four. That is, original MSVD QA dataset is not suitable for evaluating video QA models, when the models are supposed to manage video scenes with complicated temporal dynamics and diverse questions. So, we carefully sampled 30 videos that contain various motion and content dynamics from original set. The minimum number of scene changes in MSVD dynamic QA is 3, and the average scene change number is around 10.4. We also added 570 open-ended questions to the MSVD dynamic QA dataset. The questions ask more concrete answers about video scenes. Some previous methods may fail to provide good performance if they have not carefully examined both motion and appearance with the question or focused only the narrow parts of the video. Some of the questions are shown in Figure 4. The original question asks only the part of the video related with "A woman is rowing in a boat". However, the modified question sets have more verities. For example, We add questions about two men who appear in the middle of the video scene. To accurately answer the question, a model needs to capture both the intermediate motion and appearance. Further, a model should not be confused by the dominant motion, i.e. rowing. In addition, the last question asks "What are two girls doing?" where the two girls appear in the last frames. To answer the question, a model needs to address spatial redundancy.

2. Training and testing details

We use Adam optimizer and NVIDIA-RTX 6000 GPU in training. The batch size is set to 64, the ratio of drop-out is set to 0.2, and the size of hidden states in LSTM is set to 256. The original MSVD QA dataset is used for training, and the MSVD dynamic QA dataset is used for testing.

Who is rowing in a green canoe down a flooded street?	Woman
Who paddles a canoe through very shallow water?	Woman
Who is sailing the boat in the river?	Woman
Who is rowing a boat in a water flooded locality?	Woman
Who is paddling in a boat?	Newscaster

(a) MSVD QA dataset

Who is rowing a canoe down a flooded street?	Woman
Who is rowing a boat in the flooded city?	Woman
Who is passing by the boat on which a woman is paddling?	Man
What are two men doing?	Walk
What are two girls doing?	Stand

(b) MSVD dynamic QA dataset

Fig. 4. Samples in (a) the original MSVD QA dataset and (b) MSVD dynamic QA dataset

3. Performance analysis

We verified the performance of proposed algorithms in the MSVD dynamic QA dataset. AMU[1], CoMem[7], and HME[8] are used for the baseline models, because the models use attention mechanism. Because the dataset has open-ended questions, we use the top-1 accuracy (Acc@1), top-10 accuracy (Acc@10), and the WuPalmer Similarity (WUPS) score, which measures how the predicted answer is semantically close to the actual answer. WUPS@0.0 and WUPS@0.9 are used with the different thresholds in the measurement[24].

As shown in Table 1, the proposed method provides consistently improved performance over the baseline models. In Acc@1, we observe the performance difference by the proposed method around 3.6%, 5.4%, and 5.5% on AMU, CoMem, and HME, respectively. In Acc@10, the gaps become larger on the average. In particular, for HME, the proposed method improves the accuracy around 16.6% over HME only. We also report WUPS scores. As WUPS@0.9 requires more precise answer than WUPS@0.0, the similar behaviors are observed. The average difference in WUPS@0.0 and in WUPS@0.9 are around 2.6% and 6.7%, respectively. We also conduct the quantitative performance analysis on the original MSVD QA dataset to see how the performance changes with different video characteristics. HME is used as the baseline model to measure Acc@1. Table 2 shows experimental results in the original MSVD dataset. HME+Ours provides the accuracy around 34.8%, which improves around 1.1% over the original HME. It also outperforms LGCN[30] in the accuracy. It is noted that the difference was larger around 5.5% in the dynamic set. This result implies that the proposed network provides more accurate outputs although input videos contain more scene changes.

Table 1. Experimental results on the MSVD dynamic QA dataset

Model	ACC@1	ACC@10	WUPS@0.9	WUPS@0.0
AMU	19.8	35.2	30.4	60.1
CoMem	21.0	36.1	33.0	62.1
HME	23.3	38.4	34.1	70.2
AMU + PROP	23.4	38.6	34.2	70.0
CoMem + PROP	26.4	45.2	33.9	70.4
HME + PROP	28.8	55.1	37.3	72.0
Avg. △	△4.8	△9.7	△2.6	△6.7

Table 2. Experimental results on the MSVD QA dataset

Model	AMU	CoMem	HGA	L-GCN	HME	HME+ PROP
ACC@1	31.7	32.0	34.7	34.3	33.7	34.8

4. Ablation tests

4.1 Motion feature

Our network has replaced the conventional motion feature obtained from 3D CNN. In the ablation test, we compare the performance when using the proposed motion feature and the C3D feature in the MSVD dynamic QA dataset and the original dataset. As shown in Table 3, the proposed motion feature provides some improvements around 0.9% in the original set. The performance difference is significant in the dynamic set as 5.3%. We observe the proposed motion feature is more effective to the dynamic videos.

Table 3. Ablation study for video features

Dataset	Motion feature	All (13,157)	What (8,149)	Who (4,552)	Other (456)
MSVD QA	C3D	33.7	22.4	50.1	70.8
	PROP	34.8	23.5	51.0	74.3
MSVD dynamic QA	C3D	23.5	26.6	18.7	21.3
	PROP	28.8	28.7	29.7	26.2

4.2 Question-to-Attention Module

Table 4 shows the performance comparisons between the original video features a_t and m_t and question-attended video features $\overline{a}_t$ and $\overline{m}_t$. The attended features provides better performance both in the MSVD QA and MSVD dy-

namic QA datasets. In particular, the proposed video features significantly improve the performance around 6.8% in MSVD dynamic QA dataset.

Table 4. Ablation study for a question-to-attention module

Dataset	Video Feature	All (13,157)	What (8,149)	Who (4,552)	Other (456)
MSVD QA	$a_t,\ m_t$	33.9	22.7	49.9	71.2
	$\overline{a_t},\ \overline{m_t}$	34.8	23.5	51.0	74.3
MSVD dynamic QA	$a_t,\ m_t$	22.0	25.4	16.5	18.0
	$\overline{a_t},\ \overline{m_t}$	28.8	28.7	29.7	26.2

4.3 Loss function

Since we used two cost functions in Eq. (7), we turn on or off the L_2 cost function to see the performance difference. As shown in Table 5, we observe some improvements around 2.7% in the dynamic set while there has been only slight performance improvement in MSVD QA. This experimental results imply that the regularization enables the model to have more restrictions to select key video frames in the dynamic set.

Table 5. Ablation study for a loss function

Loss function	MSVD QA	MSVD Dynamic QA
L1	34.5	26.1
L1+L2	34.8	28.8

V. Conclusion

In this paper, we first defined and addressed two major intrinsic problems in video data, which are spatial redundancy and temporal redundancy, and attempted to solve the problems. The proposed network consisted of a novel motion feature extraction replacing 3D CNN and question-to-video attention during to use more reliable features. The motion feature was generated by adjacent appearance features. Unlike 3D CNN, the motion feature could represent the local motion synchronized more to the current appearance. The question-to-video attention modules generated attended motion and appearance features, and they were efficiently used for highlighting more relevant regions about the question. We also created a new setting named MSVD dynamic QA dataset to have more dynamic scenes. Experimental results showed that the proposed method outperformed the conventional methods. In the future works, we will create a larger number of QA samples so that the video QA models can be trained and tested with the samples.

Reference

[1] D. Xu, Z. Zhao, J. Xiao, F. Wu, H. Zhang, X. He, and Y. Zhuang, "Video question answering via gradually refined attention over appearance and motion," P*roceedings of the 25th ACM international conference on Multimedia*,*international,* Paris, France, pp. 1645-1653, 2017.
doi: https://dl.acm.org/doi/abs/10.1145/3123266.3123427

[2] Y. Jang, Y. Song, Y. Yu, Y. Kim, and G. Kim, "Tgif-qa: Toward spatio-temporal reasoning in visual question answering," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,* Honolulu, Hawaii, pp. 2758 - 2766, 2017.
doi: https://doi.org/10.48550/arXiv.1704.04497

[3] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:*1409.1556, 2014.
doi: https://arxiv.org/pdf/1409.1556.pdf%E3%80%82

[4] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutiontional networks," *Proceedings of the IEEE international conference on computer vision*, Santiago, Chile, pp. 4489 - 4497, 2015.
doi: https://arxiv.org/abs/1412.0767

[5] A. G. Salman, Y. Heryadi, E. Abdurahman, and W. Suparta, "Single layer & multi-layer long short-term memory (lstm) model with intermediate variables for weather forecasting," *Procedia Computer Science*, Vol.135, No. pp. 89 - 98, 2018
doi: https://doi.org/10.1016/j.procs.2018.08.153

[6] Z.-J. Zha, J. Liu, T. Yang, and Y. Zhang, "Spatiotemporal-textual coattention network for video question answering," *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, Vol. 15, no. 2s, pp. 1 - 18, 2019.
doi: https://doi.org/10.1145/3320061

[7] J. Gao, R. Ge, K. Chen, and R. Nevatia, "Motion-appearance co-memory networks for video question answering," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,*

Salt Lake City, Utah, pp. 6576 – 6585, 2018.
doi: https://doi.org/10.48550/arXiv.1803.10906

[8] C. Fan, X. Zhang, S. Zhang, W. Wang, C. Zhang, and H. Huang, "Heterogeneous memory enhanced multimodal attention model for video question answering," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,* Long Beach, CA, USA, pp. 1999 – 2007, 2019
doi: https://doi.org/10.48550/arXiv.1904.04357

[9] Z. Yang, X. He, J. Gao, L. Deng, and A. Smola, "Stacked attention networks for image question answering," *Proceedings of the IEEE conference on computer vision and pattern recognition,* Las Vegas, NV, USA, pp. 21 – 29, 2016 .
doi: https://doi.org/10.48550/arXiv.1511.02274

[10] Y. Bai, J. Fu, T. Zhao, and T. Mei, "Deep attention neural tensor network for visual question answering," *Proceedings of the European Conference on Computer Vision,* Munich, Germany, pp. 20-35, 2018.
doi: https://link.springer.com/chapter/10.1007/978-3-030-01258-8_2

[11] H. Ben-Younes, R. Cadene, M. Cord, and N. Thome, "Mutan: Multimodal tucker fusion for visual question answering," *Proceedings of the IEEE international conference on computer vision,* Venice, Italy, pp. 2612 – 2620, 2017.
doi: https://doi.org/10.48550/arXiv.1705.06676

[12] M. Seo, A. Kembhavi, A. Farhadi, and H. Hajishirzi, "Bidirectional attention flow for machine comprehension," *arXiv preprint arXiv*:1611.01603, 2016
doi: https://doi.org/10.48550/arXiv.1611.01603

[13] C. Xiong, S. Merity, and R. Socher, "Dynamic memory networks for visual and textual question answering," *International conference on machine learning. PMLR,* New York City, NY, USA, pp. 2397 – 2406, 2016.
doi: https://doi.org/10.48550/arXiv.1603.01417

[14] C. Ma, C. Shen, A. Dick, Q. Wu, P. Wang, A. van den Hengel, and I. Reid, "Visual question answering with memory-augmented networks,", *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,* Salt Lake City, Utah, pp. 6975 – 6984, 2018.
doi: https://doi.org/10.48550/arXiv.1707.04968

[15] A. Fukui, D. H. Park, D. Yang, A. Rohrbach, T. Darrell, and M. Rohrbach, "Multimodal compact bilinear pooling for visual question answering and visual grounding," *arXiv preprint arXiv*:1606.01847
doi: https://doi.org/10.48550/arXiv.1606.01847

[16] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang, "Bottom-up and top-down attention for image captioning and visual question answering," *Proceedings of the IEEE conference on computer vision and pattern recognition*, Salt Lake City, Utah, pp. 6077 – 6086, 2018.
doi: https://doi.org/10.48550/arXiv.1707.07998

[17] I. Ilievski, S. Yan, and J. Feng, "A focused dynamic attention model for visual question answering," *arXiv preprint arXiv*:1604.01485, 2016.
doi: https://doi.org/10.48550/arXiv.1604.01485

[18] Z. Zhao, J. Lin, X. Jiang, D. Cai, X. He, and Y. Zhuang, "Video question answering via hierarchical dual-level attention network learning," *Proceedings of the 25th ACM international conference on Multimedia,* Mountain View, CA , USA, pp. 1050 – 1058, 2017.
doi: https://doi.org/10.1145/3123266.3123364

[19] W. Jin, Z. Zhao, M. Gu, J. Yu, J. Xiao, and Y. Zhuang, "Multi-interaction network with object relation for video question answering," *Proceedings of the 27th ACM international conference on multimedia*, Nice, France, pp. 1193 – 1201, 2019.
doi: https://doi.org/10.1145/3343031.3351065

[20] G. Singh, *Spatio-temporal relational reasoning for video question answering,* Ph.D. dissertation of University of British Columbiam Vancouver, Canada, 2019
doi: http://hdl.handle.net/2429/72033

[21] D. Huang, P. Chen, R. Zeng, Q. Du, M. Tan, and C. Gan, "Location-aware graph convolutional networks for video question answering." *Proceedings of the AAAI Conference on Artificial Intelligence,* New York, USA, Vol. 34, no. 07, pp. 11021 – 11028, 2020.
doi: https://doi.org/10.1609/aaai.v34i07.6737

[22] P. Gao, H. Li, S. Li, P. Lu, Y. Li, S. C. Hoi, and X. Wang, "Question-guided hybrid convolution for visual question answering," *Proceedings of the European Conference on Computer Vision,* Munich, Germany, pp. 469 – 485, 2018.
doi: https://doi.org/10.1007/978-3-030-01246-5_29

[23] D. L. Chen and W. B. Dolan. "Collecting highly parallel data for paraphrase evaluation," *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 190 – 200, 2011.
doi: https://aclanthology.org/P11-1020

[24] M. Malinowski and M. Fritz, "A multi-world approach to question ans wering about real-world scenes based on uncertain input," *Advances in neural information processing systems*, Montreal, Quebec, Canada, pp. 1682 – 1690, 2014.
doi: https://doi.org/10.48550/arXiv.1410.0210

[25] Z. Yu, J. Yu, Y. Cui, D. Tao, and Q. Tian, "Deep modular co-attention networks for visual question answering," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition,* Las Vegas, NV, USA, pp: 6281 – 6290, 2019.
doi: https://doi.org/10.48550/arXiv.1906.10770

[26] 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," *Proceedings of the AAAI Conference on Artificial Intelligence,* Hawaii, USA, Vol. 33, no. 01, pp. 8658 – 8665, 2019.
doi: https://doi.org/10.1609/aaai.v33i01.33018658

[27] C. Matthew and J. Foote, "Summarizing video using non-negative similarity matrix factorization," *2002 IEEE Workshop on Multimedia Signal Processing. Elsevier,* pp. 25 – 28, 2022.
doi: https://doi.org/10.1109/MMSP.2002.1203239

[28] H. R. V. Joze, A. Shaban, M. L. Iuzzolino, and K. Koishida, "Mmtm: Multimodal transfer module for cnn fusion," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13 289 – 13 29, 2020.
doi: https://doi.ieeecomputersociety.org/10.1109/CVPR42600.2020.01330

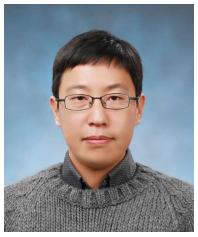
Introduction Authors

Ju-Hee Lee

- 2020.2 ~ Present : Ewha Womans University, Department of Electronic and Electrical Engineering
- ORCID : https://orcid.org/0000-0002-9245-4688
- Research interests : Video-Language pre-training, Video understanding

Seong Jong Ha

- 2009.3 ~ 2012.8 : Ph.D., Seoul National University
- 2013.1 ~ 2018.7 : Senior Engineer, Samsung SDS
- 2018.7 ~ 2022.9 : Principal Vision AI Researcher, NCSOFT
- 2022.9 ~ Present : Principal Vision AI Researcher, CJ Corporation
- ORCID : https://orcid.org/0009-0006-7231-5206
- Research interests : Video Understanding, Object Tracking, 3D Pose Estimation, Video Object Segmentation, Video Inpainting



Je-Won Kang

- 2008.8 ~ 2012.7 : Ph.D, University of Southern California
- 2012.8 ~ 2014.2 : Senior engineer, Qualcomm Inc.
- 2014.3 ~ Present : Professor, Ewha Womans University
- ORCID : https://orcid.org/0000-0002-1637-9479
- Research interests : Video coding, Video understanding, Machine Learning