Hierarchical Recurrent Contextual Attention Network for Video Question Answering

Fei Zhou[1,2] and Yahong Han[1(✉)]

[1] College of Intelligence and Computing, Tianjin University, Tianjin, China
{beiyang_flychou,yahong}@tju.edu.cn
[2] Tianjin International Engineering Institute, Tianjin University, Tianjin, China

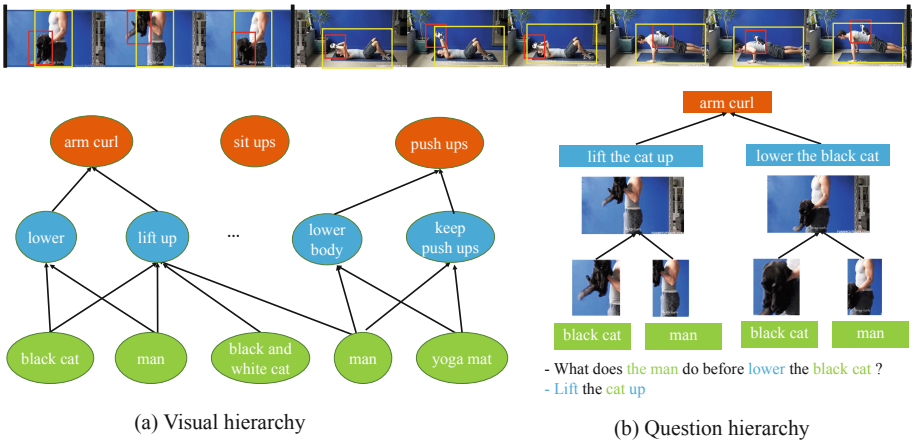
Abstract. Video question answering (VideoQA) is a task of answering a natural language question related to the content of a video. Existing methods that utilize the fine-grained object information have achieved significant improvements, however, they rely on costly external object detectors or fail to explore the rich structure of videos. In this work, we propose to understand video from two dimensions: temporal and semantic. In semantic space, videos are organized in a hierarchical structure (pixels, objects, activities, events). In temporal space, video can be viewed as a sequence of events, which contain multiple objects and activities. Based on this insight, we propose a reusable neural unit called recurrent contextual attention (RCA). RCA receives a 2D grid feature and conditional features as input, and computes multiple high-order compositional semantic representations. We then stack these units to build our hierarchy and utilize recurrent attention to generate diverse representations for different views of each subsequence. Without the bells and whistles, our model achieves excellent performance on three VideoQA datasets: TGIF-QA, MSVD-QA, and MSRVTT-QA using only grid features. Visualization results further validate the effectiveness of our method.

Keywords: Video question answering · Video understanding · Multi-modal fusion and inference

1 Introduction

Research on video-language tasks has flourished in the past few years. Video question answering (VideoQA) is one of the most prominent as it can develop agent to communicate with the dynamic visual world through natural language. From the vision perspective, fully extracting and utilizing the information contained in the video and filtering clues according to the linguistic context is the key to video question answering. Recent advancements [1,4,7,14] of VideoQA can also be mainly attributed to the exploration of finer-grained spatial information within the video frames. As a representative example, L-GCN [7] first uses an additional object detector to detect the spatial bounding boxes of important

L. Fang et al. (Eds.): CICAI 2022, LNAI 13605, pp. 280–290, 2022.
https://doi.org/10.1007/978-3-031-20500-2_23



(a) Visual hierarchy

(b) Question hierarchy

Fig. 1. Hierarchy of video and question. The entire video can be divided into three events in time: curling, sit-ups, and prone; from the semantic dimension, the low-level grid forms a series of objects: cats, people; different objects interactions form different activities; all activities are organized in sequence to construct events.

objects, and then uses a graph neural network to model the relation among all objects. Objects and their relations are undoubtedly crucial for video question answering, since interactions between objects can be explicitly captured to better understand the complex content of videos.

However, it is costly to extract fine-grained object features through object detectors. Moreover, the annotation of object detection is expensive, and pre-trained object detectors can not generalize well to datasets with large domain gap. A recent study [9] compared grid-based features and object-based features in image question answering, and showed that incorporating object detector can significantly slow down the model by 4.6 to 23.8 times, but does not bring significant performance improvements compared to plain CNN features (grid features). They also concluded that the semantic content that the feature represent is more critical than the format of features. Inspired by them, we revisit grid features for VideoQA and propose a reusable unit called Recurrent Contextual Attention Network (RCAN) that encapsulates and transforms a 2D sequence into a new higher-order 1D sequence conditioned on contextual features, where the 2D sequence can represent both the temporal and semantic dimension. The flexibility of RCAN allows it to be replicated and layered to form deep hierarchical recurrent contextual attention network (HRCAN), which can temporally divide and conquer long video clips and semantically form different levels of video concepts from the bottom up.

The hierarchy of the RCANs are as follows - at the lowest level, the RCANs encode the relations between raw *grid* features (considered as the input 2D sequence) in a frame and then aggregates multiple *regions* conditioned on frame-level context and linguistic context; at the next level, we combine multiple regions

of a clip to form a 2D sequence, then use RCANs to perform inter-region message passing between adjacent frames guided by motion context, and finally generate high-order *events* representation conditioned on clip-level context and linguistic context; in the final stage, the attention mechanism is used to aggregate the representations of multiple sub-events to form a compact global video representation for answer inference. As shown in Fig. 1, at the lowest level, we need to first aggregate the main objects from the original pixels: *black cat, people, yoga mat*, etc.; and then form a series of actions based on motion information: *lowering, raising, sit-ups, push-ups*, etc.; and finally arrange the different actions to form the overall event: *exercise*. Specifically, the aggregation of video elements (*objects, activities, events*) is achieved through the recurrent attention of RCANs, and the interaction of visual elements is achieved through message passing in RCANs. Therefore, our method incorporates the advantages of current grid-based [10] and object-based methods [7].

The contributions are summarized as follows: (1) We propose a Recurrent Attention (RA) to extract semantic content in grid features to further explore its potentials in VideoQA; (2) Further, we extend the operating objects of RA and propose a general neural unit RCAN, which receives a sequence of low-level video elements and outputs a compositional deep-semantic video element; (3) Finally, we construct a hierarchical network based on RCAN to divide and conquer videos in time sequence, and construct different levels of video semantics from bottom to top in semantic space. State-of-the-art results on three datasets validate the effectiveness of our method.

2 Related Work

VideoQA on Grid Features. Earlier work [8] used simple spatial and temporal attention mechanism, but did not bring much improvement. Some subsequent works [2,3,13] focuses on multimodal fusion and temporal modeling with using simple pooling in space. QueST [10] is a further attempt at the spatio-temporal attention of grids features, which decomposed the question semantics in space and time to guide the attention generation in space and time, respectively. Although obtaining the performance improvement, it only attend once per frame, resulting in the lack of richness of the features obtained from each frame and the inability to model the interactions between the attended regions. Therefore, we propose a recurrent attention mechanism to alleviate this shortcoming.

VideoQA on Object Features. Most recent methods for video question answering methods employ pretrained object detectors to extract object features and model interactions between them. L-GCN [7] proposed a location-aware graph neural network to model interactions between objects. HOSTR [1] used object detection and tracking to establish object trajectories, and then applied object-based spatio-temporal attention mechanisms to model object interactions. HAIR [14] utilized both object detection features and attribute features for visual and semantic relational reasoning. These methods generally achieve better performance than grid feature-based methods, the most notable difference being

that they can extract distinct object regions and explicitly model pairwise interactions between objects without the influence of irrelevant backgrounds. Thus in this paper, in addition to using the recurrent attention mechanism to extract diverse regions, we also use the message passing mechanism to model the interactions between them. A key difference is that our region extraction method is lightweight and does not require additional object detectors.

Hierarchical Architectures. Compared to images, videos contain more complex structures. Thus, HCRN [12] designed and stacked conditional relation blocks to represent videos as amalgam of complementing factors including appearance, motion and relations. However, it mainly focused on reasoning about temporal relations and used simple mean-pooling to model relations. Lack of fine-grained spatial information makes it not generalizable well to scenes involving multiple objects. Follow a similar design philosophy, HOSTR [1] introduced nested graphs for spatio-temporal reasoning over object trajectories to learn hierarchical video representations, and achieved better performance. However, the good performance of HOSTR relies on accurate object trajectories, which is difficult to achieve in practice. In this work, we also follow the hierarchical design principle, but do not rely on any additional object detectors and object trackers.

3 Proposed Method

3.1 Visual and Linguistic Representation

Video Representations. Given a video, we uniformly sample T frames and divide it into K clips with L frames, where $T = K \times L$. For each frame, we use pre-trained 2D ResNet [6] to extract appearance features F^a, then take this frame as the center frame and combine 8 frames before and after this frame to form a segment and use pre-trained 3D ResNeXt [5] to extract motion features F^m. Specifically, we use the output of the last convolutional layer as the grid feature representation, so $F^a \in \mathbb{R}^{K\times L\times 7\times 7\times 2048}$, $F^m \in \mathbb{R}^{K\times L\times 4\times 4\times 2048}$. Next, linear feature transformations are used to transform them into the standard d-dimensional feature space. Following [7], We add spatial position encoding and frame position encoding on grid features.

Linguistic Representation. We apply GloVe converts each question word into a 300-dimensional word representations, then feed the word representations into a bidirectional LSTM to model contextual dependencies. Then We obtain word-level question representations $F^q \in \mathbb{R}^{M\times d}$ by stacking the hidden states at each time step, and then use the output of the last hidden unit as the global question representation $q_L \in \mathbb{R}^d$, where M is the length of the question.

3.2 Recurrent Contextual Attention Network Unit

As illustrated in Fig. 3(a), RCAN consists of three operations: cross-modal attention (CMAT), intra-modal graph attention (GAT), and recurrent context

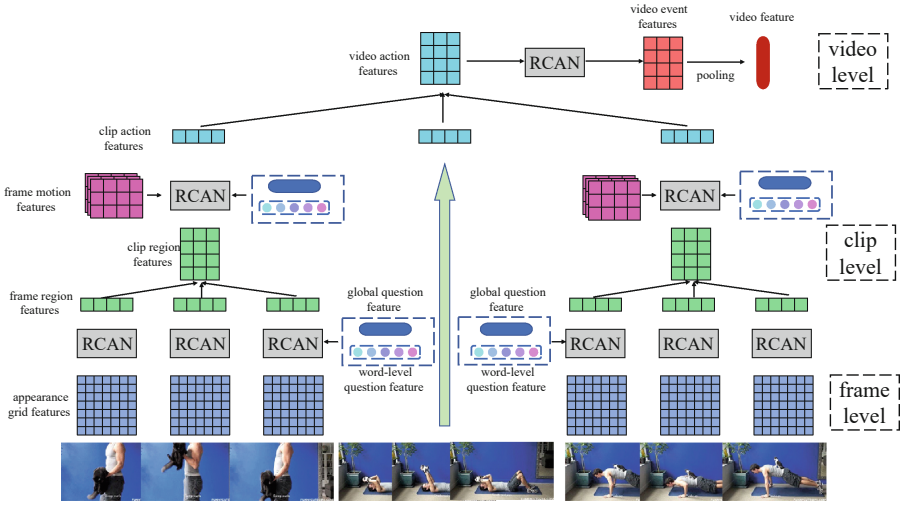


Fig. 2. Hierarchical recurrent contextual attention network (HRCAN) for VideoQA. At the frame-level, the RCAN receives grid features of a single frame as input and generates multiple frame-wise regional features in the question and frame-wise visual context. At the clip-level, we concatenate the regional features of multiple frames of a clip to generate clip-wise action features in the context of motion. Finally, we use a graph convolutional network to model the relation between video action features and generate the final video representation.

attention (RCA). The CMAT and the GAT are based on the self-attention mechanism $SAT(X, Y) = U$ [16]:

$$A = (W_1 X)((W_2 Y^T)/\sqrt{d} \tag{1}$$

$$H = softmax(A)(W_3 Y)) \tag{2}$$

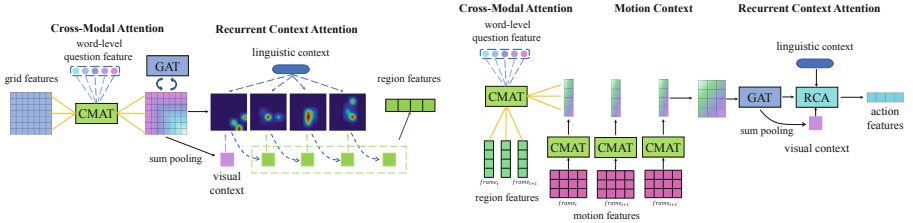
$$U = LN(X + H) \tag{3}$$

where $X \in \mathbb{R}^{N \times d}$ and $Y \in \mathbb{R}^{M \times d}$ are the inputs, $U \in \mathbb{R}^{N \times d}$ is the output, and $W_1 \sim W_3 \in \mathbb{R}^{d \times d}$ are the learned weight matrices.

RCANs takes as input a grid feature $V \in \mathbb{R}^{L_v \times d}$ ($L_v = H_v \times H_v$), which can be CNN feature map of a frame or region features of a Clip, and linguistic context $F_q \in \mathbb{R}^{M \times d}$, then produce compositional features: $R = RCAN(V, F_q) \in \mathbb{R}^{N \times d}$.

Cross-modal Attention. Through the cross-attention of word-level linguistic representation and visual features, the video elements mentioned in the question will have a stronger response.

$$\tilde{V} = CMAT(V, F^q) = SAT(V, F^q) \tag{4}$$



(a) Frame-level RCAN. (b) Clip-level RCAN.

Fig. 3. Illustration of recurrent context attention unit (RCAN). Cross-modal attention (CMAT) is first used to model inter-modal interactions, then graph attention (GAT) further model intra-modal interactions, and finally recurrent contextual attention (RCA) extracts multiple higher-order representations. Additional motion information is introduced at the clip-level.

After obtaining enhanced visual features, we use graph attention (GAT) to model interactions within modalities:

$$\hat{V} = GAT(\tilde{V}, \tilde{V}) = SAT(\tilde{V}, \tilde{V}) \tag{5}$$

Graph attention enables message passing between related visual elements, such as different parts of the same object, different objects of an action, which can facilitate subsequent RCA to extract more semantic-related region features.

The detailed computation of recurrent context attention (RCA) is as follows. We first compute a guidance vector $g_t \in \mathbb{R}^d$ based on the globally average-pooled visual feature or the last attended region feature $r_{t-1} \in \mathbb{R}^d$, the l-th grid feature $\hat{v}_l$ and the linguistic context $q_L \in \mathbb{R}^d$.

$$g_{t,l} = W_{g,t}(W_{q,t}q_L + W_{r,t}r_{t-1} + W_{l,t}\hat{v}_l) + b_{g,t} \tag{6}$$

$$\alpha_{t,l} = W_{att,t}LeakyReLU(g_{t,l}) + b_{att,t} \tag{7}$$

$$\alpha_t = Softmax([\alpha_{t,1}, ..., \alpha_{t,L_v}]) \tag{8}$$

$$r_t = \sum_{l=1}^{L_v} \alpha_{t,l}\hat{v}_l \tag{9}$$

where $W_{g,t}$, $W_{q,t}$, $W_{r,t}$, $W_{l,t}$, $W_{att,t} \in \mathbb{R}^{d\times d}$ are learnable weights, and $b_{g,t}$ and $b_{att,t}$ are the biases, $a_{t,l}$ is a scale which is the weight on the l-th grid feature used to generate the t-th region. r_t is the visual context at step t, $r_0 = sumpooling(\hat{V})$. Finally, we iteratively generate N regions $R = \{r_1, ..., r_n\} \in \mathbb{R}^{N\times d}$.

3.3 Hierarchical Recurrent Contextual Attention Network

As shown in Fig. 2, at the lowest level, RCAN receives frame-level grid features F^a as input, then captures inter-modal and intra-modal interactions and finally generates a series of regional features $F^r \in R^{K\times L\times N_f\times d}$:

$$F^r_{k,l} = RCAN(F^a_{k,l}, F^q, q_L) \tag{10}$$

where, $F^a_{k,l}$ is the grid features of the l-th frame of the k-th clip, $F^r_{k,l}$ is the regional feature of the corresponding frame.

Frame-level RCAN have extracted diverse regional features from each frame and modeled intra-frame relation via GAT and RCA. The goal of clip-level RCAN is combine motion context and regional information to form different actions, such as "lower the cat, lift up the cat". As shown in Fig. 3 (b), we add a CMAT to introduce motion context into region features. Specifically, the clip-level RCAN receives the output of the frame-level RCAN R_f and the frame-level motion grid features $F^m \in R^{K\times L\times L_m\times d}$ as input, and cyclically generates a series of action representations $F^c \in \mathbb{R}^{K\times N_c\times d}$:

$$F^c_k = RCAN(F^r_k, F^m_k, F^q, q_L) \tag{11}$$

where $F^r_k \in R^{L\times N_f\times d}$ is the feature of all regions of the k-th clip, $F^m_k \in R^{L\times L_m\times d}$ is the motion feature of all frames of the k-th clip, and $F^c_k \in \mathbb{R}^{N_c\times d}$ is the action features generated by the k-th clip. Finally, we use a RCAN without RCA to model the dependencies between the clip features of the video: $F^v = RCAN(F^c, F^q) \in \mathbb{R}^{(K\times N_c)\times d}$ and use attention pooling to generate the final visual feature: $z = Attn(F^v, q_L) \in \mathbb{R}^d$.

3.4 Answer Decoder

Following the previous work [10,12], for multiple-choice QA, we concatenate each candidate and question to form a holistic query. The global query feature q_L is fused with the final video feature z and a multilayer perceptron (MLP) is used to predict scores:

$$s = MLP([z; q_L]) \tag{12}$$

For multi-choice QA, we maximize the margin between positive and negative QA-pairs: $max(0, 1 + s^n - s^p)$. For opened QA, We treat it as a classification task on a pre-defined set of answers, then use the decoder to predict a class probability and train it using cross-entropy.

4 Experiments

4.1 Experiment Setup

Datasets. **TGIF-QA** [8] contains 165K QA pairs collected from 72K animated GIFs. We use action repetition (Action), state transition (Trans.), frame-level question (FrameQA) tasks for evaluation. FrameQA is an opened QA, the others are multi-choice QA. **MSVD-QA** and **MSRVTT-QA** [18] contain 50K and 243K Q&A pairs respectively, and consist of five different types of questions, including *what,who,how,when* and *where*. The task is open-ended. For all datasets, we report accuracy (percentage of correctly answered questions) as an evaluation metric according to the standard.

Table 1. Comparison with state-of-the-art methods.

Methods	TGIF-QA			MSRVTT-QA	MSVD-QA
	Action	Trans.	FrameQA		
ST-VQA [8]	62.9	69.4	49.5	30.9	31.3
PSAC [13]	70.4	76.9	55.7	–	–
QueST [10]	75.9	81.0	59.7	34.6	36.1
Co-mem [3]	68.2	74.3	51.5	31.9	31.7
HME [2]	73.9	77.8	53.8	33.0	33.7
L-GCN [7]	74.3	81.1	56.3	33.7	34.3
HGA [11]	75.4	81.0	55.1	35.5	34.7
GMIN [4]	73.0	81.7	57.5	36.1	35.4
BTA [15]	75.9	82.6	57.5	36.9	37.2
HCRN [12]	75.0	81.4	55.9	35.6	36.1
HOSTR [1]	75.0	83.0	58.0	35.9	39.4
HQGA [17]	76.9	85.6	61.3	38.6	41.2
HRCAN	**81.8**	83.6	**63.7**	**38.8**	**41.8**

Implementation Details. For each video, we uniformly sample $T = 16$ frames, then divide them into $K = 4$ clips with $L = 4$ frames per clip. We use pretrained ResNet152 from [9] to extract appearance features, and use ResNeXt101 pretrained on Kinetics to extract motion features. We set the dimension of the hidden units d to 512, and $N_f = N_c = 4$. For training details, we train our model for 50 epochs with a batch size of 32. The learning rate is set to 10^{-4}, warms up for 5 epochs, and then cosine anneals.

4.2 Comparison with Prior Work

In Table 1, we compare our method with methods involving 4 main categories: cross-attention, memory-based methods, graph-structed methods and hierarchical models. The results show that our HRCAN model consistently outperforms other models on all experimental datasets.

Specifically, both L-GCN and GMIN use graph-based methods to model object-level interactions for question answering. However, they fail to construct the hierarchical nature of the video and interactions are constructed without the guidance of language query. Through the inter-modal and inter-modal interactions in RCAN, and the semantic hierarchy of HRCAN, our model shows clear superiority on the experimental dataset. HCRN, HOSTR and HQGA are similar to us in designing hierarchical conditional architectures. However, HCRN is limited to hierarchical temporal relations between frames, which are only modeled by simple average pooling. The lack of spatial fine-grained information makes it insufficient to understand complex object interactions in space-time, which limits not only its performance in single-frame question answering, but also its tempo-

ral reasoning ability. HOSTR advances HCRN by building hierarchies on object trajectories and employing graph operations for relational reasoning. However, it lacks object-word level fine-grained matching which results in its sub-optimal results, and relies on costly object detection and trajectory tracking. Finally, the HQGA, like us, tries to aligns words in the linguistic queries with visual elements of the hierarchy in the video. However, it relies on multiple visual encoder to build different visual hierarchies, *e.g.*, 2D & 3D CNN and Faster-RCNN, which makes it limited in practical applications.

4.3 Ablation Studies

Hierarchy. In the top section of Table 2, we layer-wisely replace the RCAN with average pooling to study the effect of the hierarchy. It can be seen that the lack of any level will cause performance degradation. And the frame-level has the greatest impact on the results, we argue that the lack of spatially fine-grained information extraction will introduce noise to subsequent levels. This show that just modeling the relation between clips is suboptimal. We also study the effect of the number of iterations in RCAN and find that increasing both N_f and N_c can lead to better overall performance.

Linguistic Conditioning. From the middle section of Table 2, the lack of language conditions jeopardize the overall performance, indicating the necessity of injecting query cues when encoding video features. Specially, we replace word-level representations F^q in cross-modal attention with global one q_L, and the performance drop shows the importance of fine-grained cross-modal matching.

Motion Conditioning. Removing the motion context at the clip level or putting the motion information before the frame level will cause performance

Table 2. Ablation studies on TGIF-QA dataset, for action repetition and frameqa tasks.

Model	Action	Frame
Hierarchy		
w/o frame-level	77.90	58.60
w/o clip-level	78.40	63.50
w/o video-level	81.70	64.01
Linguistic conditioning		
w/o linguistic cond.	78.20	62.80
w/ global linguistic cond.	80.10	63.40
Motion conditioning		
w/o motion cond. in clip-level	80.50	63.30
w motion cond. in frame-level	80.20	63.40
Full model	81.80	63.70

Model	Action	Frame
RCA(N_c and N_f)		
$N_c = 1$, $N_f = 1$	80.80	63.0
$N_c = 1$, $N_f = 2$	80.60	63.7
$N_c = 1$, $N_f = 4$	80.60	63.6
$N_c = 1$, $N_f = 1$	80.80	63.0
$N_c = 2$, $N_f = 1$	81.60	62.7
$N_c = 4$, $N_f = 1$	80.70	63.7

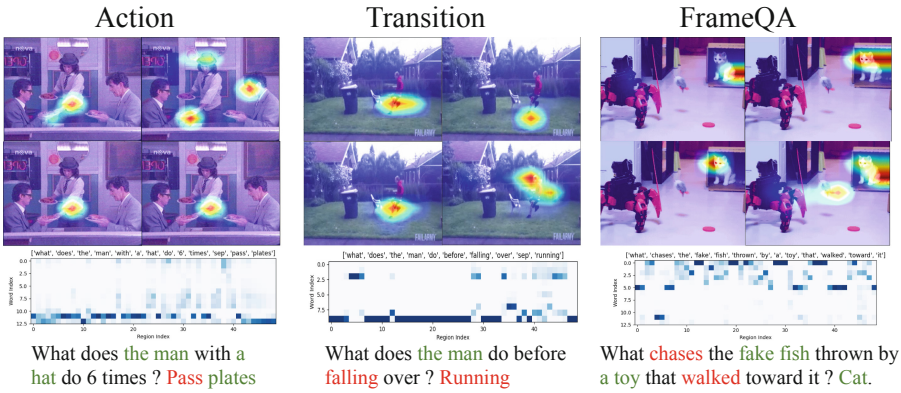


Fig. 4. Visualization of our cross-modal attention and recurrent attention in frame level. The above is the attended region heat map for recurrent attention, the below is the attention weights between grid and words in cross-modal attention.

degradation. One is that the introduction of motion context can make up for the loss of information caused by sparse sampling, and the other is that adding motion information directly to the appearance grid features may bring noise without frame-level RCAN enhancement and extraction of question-specific regions.

Qualitative Analysis. In Fig. 4, we visualize frame-level cross-modal attention and recurrent attention. It can be seen that in CMA, important objects or actions in the query are emphasized, such as "hat","passing plates", "running", "fake fish", "cat". Subsequently, the corresponding visual regions are extracted by recurrent attention. Moreover, we can observe: 1) CMA is sparse, which may be the reason that word-level supervision is better than sentence-level one. Namely, only a few query subjects in sentences need to be emphasized; 2) the attention of grid features is more flexible than that of object. Obejct features are often limited to a rectangular instance-related bounded by box, while our grid-based recurrent attention can attend flexible regions, which may only be part of the instance, *e.g.*, "plate", "legs".

5 Conclusion

In this paper, we propose a new VideoQA model termed as HRCAN, that uses a reusable attention unit RCAN to perform hierarchical reasoning on visual elements. Specifically, multimodal interactions are modeled in RCAN through inter-modal and intra-modal attention, and then low-level visual elements are aggregated into diverse high-level visual elements through recurrent attention. Our extensive experimental analysis have validated the effectiveness of the propose method. Additional visual analysis can also further validate the insights.

References

1. Dang, L.H., Le, T.M., Le, V., Tran, T.: Hierarchical object-oriented spatio-temporal reasoning for video question answering. arXiv preprint. arXiv:2106.13432 (2021)
2. Fan, C., Zhang, X., Zhang, S.: Heterogeneous memory enhanced multimodal attention model for video question answering. In: CVPR, pp. 1999–2007 (2019)
3. Gao, J., Ge, R., Chen, K.: Motion-appearance co-memory networks for video question answering. In: CVPR, pp. 6576–6585 (2018)
4. Gu, M., Zhao, Z., Jin, W., Hong, R., Wu, F.: Graph-based multi-interaction network for video question answering. IEEE Trans. Image Process. **30**, 2758–2770 (2021)
5. Hara, K., Kataoka, H., Satoh, Y.: Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? In: CVPR, pp. 6546–6555 (2018)
6. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
7. Huang, D., Chen, P., Zeng, R.: Location-aware graph convolutional networks for video question answering. In: AAAI, pp. 11021–11028 (2020)
8. Jang, Y., Song, Y., Yu, Y.: Tgif-qa: toward spatio-temporal reasoning in visual question answering. In: CVPR, pp. 2758–2766 (2017)
9. Jiang, H., Misra, I., Rohrbach, M., Learned-Miller, E., Chen, X.: In defense of grid features for visual question answering. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10267–10276 (2020)
10. Jiang, J., Chen, Z.: Divide and conquer: question-guided spatio-temporal contextual attention for video question answering. In: AAAI, pp. 11101–11108 (2020)
11. Jiang, P., Han, Y.: Reasoning with heterogeneous graph alignment for video question answering. In: AAAI, pp. 11109–11116 (2020)
12. Le, T.M., Le, V., Venkatesh, S.: Hierarchical conditional relation networks for video question answering. In: CVPR, pp. 9972–9981 (2020)
13. Li, X., Song, J., Gao, L.: Beyond rnns: positional self-attention with co-attention for video question answering. In: AAAI, pp. 8658–8665 (2019)
14. Liu, F., Liu, J., Wang, W., Lu, H.: Hair: hierarchical visual-semantic relational reasoning for video question answering. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 1698–1707 (2021)
15. Park, J., Lee, J., Sohn, K.: Bridge to answer: Structure-aware graph interaction network for video question answering. In: CVPR, pp. 15526–15535 (2021)
16. Vaswani, A., Shazeer, N., Parmar, N.: Attention is all you need. In: NeurIPS, pp. 5998–6008 (2017)
17. Xiao, J., Yao, A., Liu, Z., Li, Y., Ji, W., Chua, T.S.: Video as conditional graph hierarchy for multi-granular question answering. AAAI (2022)
18. Xu, D., Zhao, Z., Xiao, J.: Video question answering via gradually refined attention over appearance and motion. In: ACM MM, pp. 1645–1653 (2017)