Hierarchical Conditional Relation Networks for Video Question Answering

Thao Minh Le, Vuong Le, Svetha Venkatesh, and Truyen Tran
Applied Artificial Intelligence Institute, Deakin University, Australia
{lethao,vuong.le,svetha.venkatesh,truyen.tran}@deakin.edu.au

Abstract

Video question answering (VideoQA) is challenging as it requires modeling capacity to distill dynamic visual artifacts and distant relations and to associate them with linguistic concepts. We introduce a general-purpose reusable neural unit called Conditional Relation Network (CRN) that serves as a building block to construct more sophisticated structures for representation and reasoning over video. CRN takes as input an array of tensorial objects and a conditioning feature, and computes an array of encoded output objects. Model building becomes a simple exercise of replication, rearrangement and stacking of these reusable units for diverse modalities and contextual information. This design thus supports high-order relational and multi-step reasoning. The resulting architecture for VideoQA is a CRN hierarchy whose branches represent sub-videos or clips, all sharing the same question as the contextual condition. Our evaluations on well-known datasets achieved new SoTA results, demonstrating the impact of building a general-purpose reasoning unit on complex domains such as VideoQA.

1. Introduction

Answering natural questions about a video is a powerful demonstration of cognitive capability. The task involves acquisition and manipulation of spatio-temporal visual representations guided by the compositional semantics of the linguistic cues [7, 17, 20, 30, 33, 36]. As questions are potentially unconstrained, VideoQA requires deep modeling capacity to encode and represent crucial video properties such as object permanence, motion profiles, prolonged actions, and varying-length temporal relations in a hierarchical manner. For VideoQA, the visual representations should ideally be question-specific and answer-ready.

The current approach toward modeling videos for QA is to build neural architectures in which each sub-system is either designed for a specific tailor-made purpose or for a particular data modality. Because of this specificity, such hand crafted architectures tend to be non-optimal for changes in data modality [17], varying video length [24] or question types (such as frame QA [20] versus action count [6]). This has resulted in proliferation of heterogeneous networks.

In this work we propose a general-purpose *reusable neural unit* called Conditional Relation Network (CRN) that encapsulates and transforms an array of objects into a new array conditioned on a contextual feature. The unit computes sparse high-order relations between the input objects, and then modulates the encoding through a specified context (See Fig. 2). The flexibility of CRN and its encapsulating design allow it to be replicated and layered to form deep hierarchical conditional relation networks (HCRN) in a straightforward manner. The stacked units thus provide contextualized refinement of relational knowledge from video objects – in a stage-wise manner it combines appearance features with clip activity flow and linguistic context, and follows it by integrating in context from the whole video motion and linguistic features. The resulting HCRN is homogeneous, agreeing with the design philosophy of networks such as InceptionNet [31], ResNet [9] and FiLM [27].

The hierarchy of the CRNs are as follows – at the lowest level, the CRNs encode the relations *between* frame appearance in a clip and integrate the *clip motion as context*; this output is processed at the next stage by CRNs that now integrate in the *linguistic context*; in the following stage, the CRNs capture the relation *between* the clip encodings, and integrate in *video motion as context*; in the final stage the CRN integrates the video encoding with the linguistic feature as context (See Fig. 3). By allowing the CRNs to be stacked hierarchically, the model naturally supports modeling hierarchical structures in video and relational reasoning; by allowing appropriate context to be introduced in stages, the model handles multimodal fusion and multi-step reasoning. For long videos further levels of hierarchy can be added enabling encoding of relations between distant frames.

We demonstrate the capability of HCRN in answering questions in major VideoQA datasets. The hierarchical architecture with four-layers of CRN units achieves favorable answer accuracy across all VideoQA tasks. Notably, it performs consistently well on questions involving either appearance, motion, state transition, temporal relations, or action repetition demonstrating that the model can analyze and

(a) Question: What does the girl do 9 times?
Baseline: walk
HCRN: blocks a person's punch
Ground truth: blocks a person's punch

(b) Question: What does the man do before turning body to left?
Baseline: pick up the man's hand
HCRN: breath
Ground truth: breath

Figure 1. Example questions for which frame relations are key toward correct answers. (a) *Near-term frame relations* are required for counting of fast actions. (b) *Far-term frame relations* connect the actions in long transition. HCRN with the ability to model hierarchical conditional relations handles successfully, while baseline struggles. See more examples in supplemental materials.

combine information in all of these channels. Furthermore HCRN scales well on longer length videos simply with the addition of an extra layer. Fig. 1 demonstrates several representative cases those were difficult for the baseline of flat visual-question interaction but can be handled by our model.

Our model and results demonstrate the impact of building general-purpose neural reasoning units that support native multimodality interaction in improving robustness and generalization capacities of VideoQA models.

2. Related Work

Our proposed HCRN model advances the development of VideoQA by addressing two key challenges: (1) Efficiently representing videos as amalgam of complementing factors including appearance, motion and relations, and (2) Effectively allows the interaction of such visual features with the linguistic query.

Spatio-temporal video representation is traditionally done by variations of recurrent networks (RNNs) among which many were used for VideoQA such as recurrent encoder-decoder [49, 47], bidirectional LSTM [15] and two-staged LSTM [44]. To increase the memorizing ability, external memory can be added to these networks [7, 44]. This technique is more useful for videos that are longer [40] and with more complex structures such as movies [33] and TV programs [17] with extra accompanying channels such as speech or subtitles. On these cases, memory networks [15, 24, 35] were used to store multimodal features [36] for later retrieval. Memory augmented RNNs can also compress video into heterogenous sets [6] of dual appearance/motion features. While in RNNs, appearance and motion are modeled separately, 3D and 2D/3D hybrid convolutional operators [34, 28] intrinsically integrates spatio-temporal visual information and are also used for VideoQA [10, 20]. Multiscale temporal structure can be modeled by either mixing short and long term convolutional filters [37] or combining pre-extracted frame features non-local operators [32, 18]. Within the second approach, the TRN network [48] demonstrates the role of temporal frame relations as an another important visual feature for video reasoning and VideoQA [16]. Relations of predetected objects were also considered in a separate processing stream [11] and combined with other modalities in late-fusion [29]. Our HCRN model emerges on top of these trends by allowing all three channels of video information namely appearance, motion and relations to iteratively interact and complement each other in every step of a hierarchical multi-scale framework.

Earlier attempts for generic multimodal fusion for visual reasoning includes bilinear operators, either applied directly [13] or through attention [13, 43]. While these approaches treat the input tensors equally in a costly joint multiplicative operation, HCRN separates conditioning factors from refined information, hence it is more efficient and also more flexible on adapting operators to conditioning types.

Temporal hierarchy has been explored for video analysis [22], most recently with recurrent networks [25, 1] and graph networks [23]. However, we believe we are the first to consider hierarchical interaction of multi-modalities including linguistic cues for VideoQA.

Linguistic query–visual feature interaction in VideoQA has traditionally been formed as a visual information retrieval task in a common representation space of independently transformed question and referred video [44]. The retrieval is more convenient with heterogeneous memory slots [6]. On top of information retrieval, co-attention between the two modalities provides a more interactive combination [10]. Developments along this direction include attribute-based attention [42], hierarchical attention [21, 45, 46], multi-head attention [14, 19], multi-step progressive attention memory [12] or combining self-attention with co-attention [20]. For higher order reasoning, question can interact iteratively with video features via episodic memory or through switching mechanism [41]. Multi-step reasoning for VideoQA is also approached by [39] and [30] with refined attention.

Unlike these techniques, our HCRN model supports conditioning video features with linguistic clues as a context factor in every stage of the multi-level refinement process. This allows linguistic cue to involve earlier and deeper into video presentation construction than any available methods.

Neural building blocks - Beyond the VideoQA domain, CRN unit shares the idealism of uniformity in neural architecture with other general purpose neural building blocks

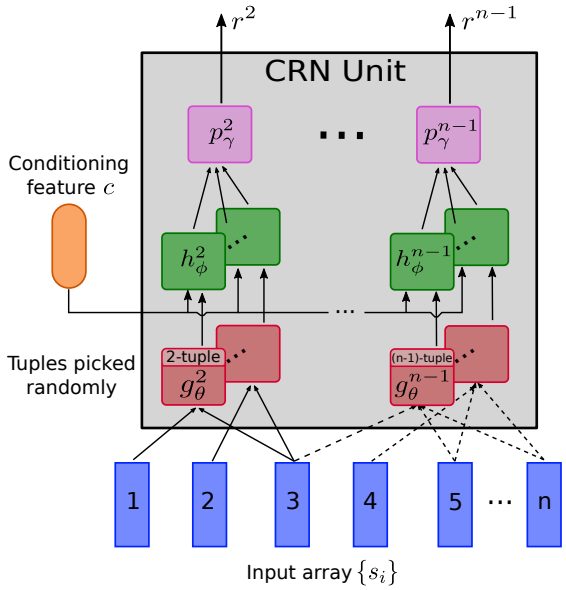

Figure 2. Conditional Relation Network. a) Input array $\mathcal{S}$ of n objects are first processed to model k-tuple relations from t sub-sampled size-k subsets by sub-network $g^k(.)$. The outputs are further conditioned with the context c via sub-network $h^k(.,.)$ and finally aggregated by $p^k(.)$ to obtain a result vector r^k which represents k-tuple conditional relations. Tuple sizes can range from 2 to $(n-1)$, which outputs an $(n-2)$-dimensional output array.

such as the block in InceptionNet [31], Residual Block in ResNet [9], Recurrent Block in RNN, conditional linear layer in FiLM [27], and matrix-matrix-block in neural matrix net [5]. Our CRN departs significantly from these designs by assuming an array-to-array block that supports conditional relational reasoning and can be reused to build networks of other purposes in vision and language processing.

3. Method

The goal of VideoQA is to deduce an answer $\tilde{a}$ from a video $\mathcal{V}$ in response to a natural question q. The answer $\tilde{a}$ can be found in an answer space $\mathcal{A}$ which is a pre-defined set of possible answers for open-ended questions or a list of answer candidates in case of multi-choice questions. Formally, VideoQA can be formulated as follows:

$$\tilde{a} = \underset{a \in \mathcal{A}}{\text{argmax}} \mathcal{F}_\theta \left(a \mid q, \mathcal{V}\right), \tag{1}$$

where θ is the model parameters of scoring function $\mathcal{F}$.

Visual representation We begin by dividing the video $\mathcal{V}$ of L frames into N equal length clips $C = (C_1, ..., C_N)$. Each clip C_i of length $T = \lfloor L/N \rfloor$ is represented by two sources of information: frame-wise appearance feature vectors $V_i = \left\{v_{i,j} | v_{i,j} \in \mathbb{R}^{2048}\right\}_{j=1}^{T}$, and the motion feature vector at clip level $f_i \in \mathbb{R}^{2048}$. In our experiments, $v_{i,j}$ are the *pool5* output of ResNet [9] features and f_i are derived by ResNeXt-101 [38, 8].

Notation	Role
$\mathcal{S}$	Input array of n objects (e.g. frames, clips)
c	Conditioning feature (e.g. query, motion feat.)
k_{max}	Maximum subset (also tuple) size considered
k	Each subset size from 2 to k_{max}
Q^k	Set of all size-k subsets of $\mathcal{S}$
t	Number of subsets randomly selected from Q^k
Q^k_{selected}	Set of t selected subsets from Q^k
$g^k(.)$	Sub-network processing each size-k subset
$h^k(.,.)$	Conditioning sub-network
$p^k(.)$	Aggregating sub-network
R	Result array of CRN unit on $\mathcal{S}$ given c
r^k	Member result vector of k-tuple relations

Table 1. Notations of CRN unit operations

Algorithm 1: CRN Unit

Input : Array $\mathcal{S} = \{s_i\}_{i=1}^n$, conditioning feature c
Output : Array R
Metaparams : $\{k_{\text{max}}, t \mid k_{\text{max}} < n\}$
Build all sets of subsets $\{Q^k \mid k = 2, 3, ..., k_{\text{max}}\}$ where Q^k is set of all size-k subsets of $\mathcal{S}$
Initialize $R \leftarrow \{\}$
for $k \leftarrow 2$ **to** k_{max} **do**
  Q^k_{selected} = randomly select t subsets from Q^k
  for each *subset* $q_i \in Q^k_{\text{selected}}$ **do**
    $g_i = g^k(q_i)$
    $h_i = h^k(g_i, c)$
  end
  $r^k = p^k(\{h_i\})$
  add r^k to R
end

Subsequently, linear feature transformations are applied to project $\{v_{ij}\}$ and f_i into a standard d-dimensions feature space to obtain $\{\hat{v}_{ij} | \hat{v}_{ij} \in \mathbb{R}^{\text{d}}\}$ and $\hat{f}_i \in \mathbb{R}^{\text{d}}$, respectively.

Linguistic representation All words in the question and answer candidates in case of multi-choice questions are first embedded into vectors of 300 dimensions, which are initialized with pre-trained GloVe word embeddings [26]. We further pass these context-independent embedding vectors through a biLSTM. Output hidden states of the forward and backward LSTM passes are finally concatenated to form the question representation $q \in \mathbb{R}^d$.

With these representations, we now describe our new hierarchical architecture for VideoQA (see Fig. 3). We first present the core compositional computation unit that serves as building blocks for the architecture in Section 3.1. In the following sub-section, we propose to design $\mathcal{F}$ as a layer-by-layer network architecture that can be built by simply stacking the core units in a particular manner.

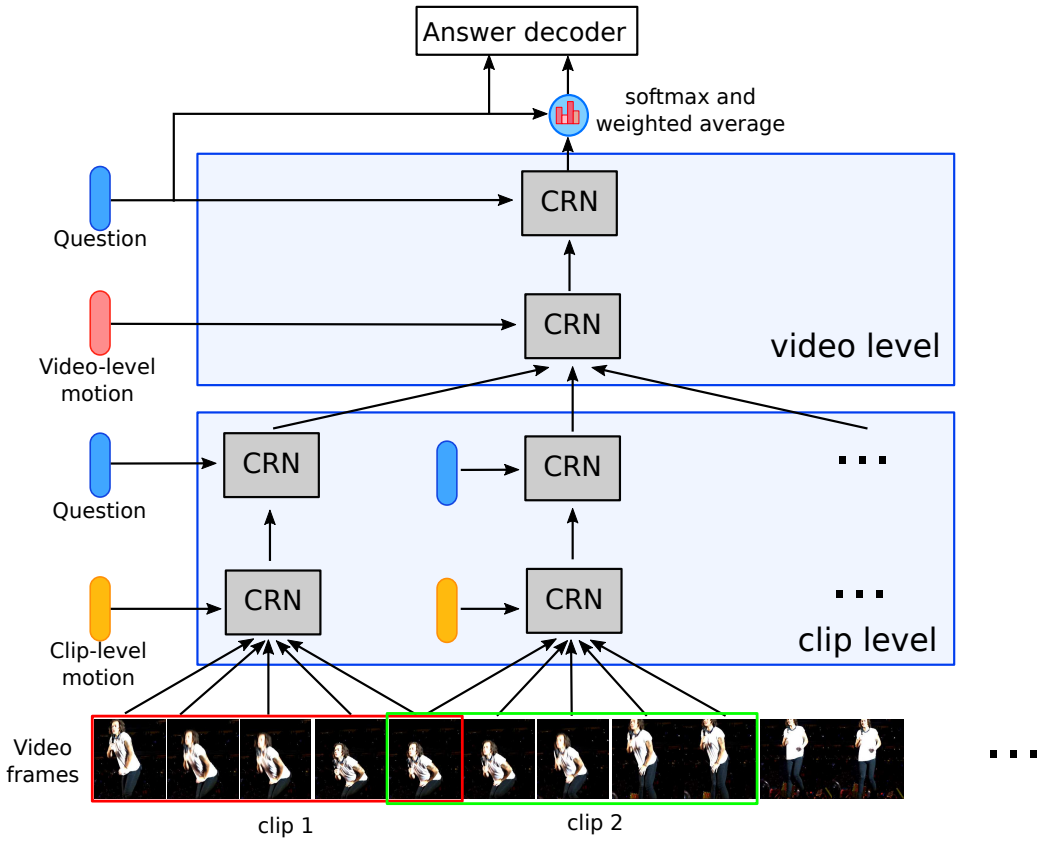

Figure 3. Hierarchical Conditional Relation Networks (HCRN) Architecture for VideoQA. The CRNs are stacked in a hierarchy, embedding the video input at different granularities including frame, short clip and entire video levels. The video feature embedding is conditioned on the linguistic cue at each level of granularity. The visual-question joint representation is fed into an output classifier for prediction.

3.1. Conditional Relation Network Unit

We introduce a reusable computation unit, termed Conditional Relation Network (CRN), which takes as input an array of n objects $\mathcal{S} = \{s_i\}_{i=1}^{n}$ and a conditioning feature c - both in the same vector space $\mathbb{R}^d$ or tensor space $\mathbb{R}^{W\times H\times d}$. CRN generates an output array of objects of the same dimensions containing high-order object relations of input features given the global context. The operation of CRN unit is presented algorithmically in Alg. 1 and visually in Fig. 2. Table 1 summarizes the notations used across these presentations.

When in use for VideoQA, CRN's input array is composed of features at either frame or short-clip levels. The objects $\{s_i\}_{i=1}^{n}$ greatly share mutual information and it is redundant to consider all possible combinations of given objects. Therefore, applying a sampling scheme on the set of subsets (line 4 of Alg. 1) is crucial for redundancy reduction and computational efficiency. We borrow the sampling trick in [48] to build sets of t selected subsets Q^k_{selected}. Regarding the choice of k_{max}, we choose $k_{\text{max}} = n - 1$ in later experiments, resulting in the output array of size $n - 2$ if $n > 2$ and array of size 1 if $n = 2$.

As a choice in implementation, the functions $g^k(.), p^k(.)$ are simple average-pooling. In generic form, they can be any aggregation sub-networks that join a random set into a single representation. Meanwhile, $h^k(.,.)$ is a MLP running on top of feature concatenation that models the non-linear relationships between multiple input modalities. We tie parameters of the conditioning sub-network $h^k(.,.)$ across the subsets of the same size k. In our implementation, $h^k(.,.)$ consists of a single linear transformation followed by an ELU [3] activation.

It may be of concern that the relation formed by a particular subset may be unnecessary to model k-tuple relations, we optionally design a self-gating mechanism similar to [4] to regulate the feature flow to go through each CRN module. Formally, the conditioning function $h^k(.,.)$ in that case is given by:

$$h^k\left(x, y\right) = \text{ELU}(W_{h_1}\left[x, y\right]) * \sigma\left(W_{h_2}\left[x, y\right]\right), \quad (2)$$

where $[.,.]$ denotes the tensor concatenation, σ is sigmoid function, and W_{h_1}, W_{h_2} are linear weights.

3.2. Hierarchical Conditional Relation Networks

We use CRN blocks to build a deep network architecture to exploit inherent characteristics of a video sequence namely temporal relations, motion, and the hierarchy of video structure, and to support reasoning guided by linguistic questions. We term the proposed network architecture

Hierarchical Conditional Relation Networks (HCRN) (see Fig. 3). The design of the HCRN by stacking reusable core units is partly inspired by modern CNN network architectures, of which InceptionNet [31] and ResNet [9] are the most well-known examples.

A model for VideoQA should distill the visual content in the context of the question, given the fact that much of the visual information is usually not relevant to the question. Drawing inspiration from the hierarchy of video structure, we boil down the problem of VideoQA into a process of video representation in which a given video is encoded progressively at different granularities, including short clip (consecutive frames) and entire video levels. It is crucial that the whole process conditions on linguistic cue. In particular, at each hierarchy level, we use two stacked CRN units, one conditioned on motion features followed by one conditioned on linguistic cues. Intuitively, the motion feature serves as a dynamic context shaping the temporal relations found among frames (at the clip level) or clips (at the video level). As the shaping effect is applied to all relations, self-gating is not needed, and thus a simple MLP suffices. On the other hand, the linguistic cues are by nature selective, that is, not all relations are equally relevant to the question. Thus we utilize the self-gating mechanism in Eq. (2) for the CRN units which condition on question representation.

With this particular design of network architecture, the input array at clip level consists of frame-wise appearance feature vectors $\{\hat{v}_{ij}\}$, while that at a video level is the output at the clip level. Meanwhile, the motion conditioning feature at clip level CRNs are corresponding clip motion feature vector $\hat{f}_i$. They are further passed to an LSTM, whose final state is used as video-level motion features. Note that this particular implementation is not the only option. We believe we are the first to progressively incorporate multiple modalities of input in such a hierarchical manner in contrast to the typical approach of treating appearance features and motion features as a two-stream network.

To handle a long video of thousand frames, which is equivalent to dozens of short-term clips, there are two options to reduce the computational cost of CRN in handling large sets of subsets $\{Q^k | k = 2, 3, ..., k_{\text{max}}\}$ given an input array S: limit the maximum subset size k_{max} or extend the HCRN to deeper hierarchy. For the former option, this choice of sparse sampling may have potential to lose critical relation information of specific subsets. The latter, on the other hand, is able to densely sample subsets for relation modeling. Specifically, we can group N short-term clips into $N_1 \times N_2$ hyper-clips, of which N_1 is the number of the hyper-clips and N_2 is the number of short-term clips in one hyper-clip. By doing this, our HCRN now becomes a 3-level of hierarchical network architecture.

At the end of the HCRN, we compute the average visual feature based on conditioning to the question representation q. Assume outputs of the last CRN unit at video level are an array $O = \left\{o_i \mid o_i \in \mathbb{R}^{H \times d}\right\}_{i=1}^{N-4}$, we first stack them together, resulting in an output tensor $o \in \mathbb{R}^{(N-4) \times H \times d}$, and further vectorize this output tensor to obtain the final output $o' \in \mathbb{R}^{H' \times d}, H' = (N-4) \times H$. The weighted average information is given by:

$$I = [W_{o'} o', W_{o'} o' \odot W_q q], \tag{3}$$

$$I' = \text{ELU}\left(W_I I + b\right), \tag{4}$$

$$\gamma = \text{softmax}\left(W_{I'} I' + b\right), \tag{5}$$

$$\tilde{o} = \sum_{h=1}^{H'} \gamma_h o'_h;\ \tilde{o} \in \mathbb{R}^d, \tag{6}$$

where, $[.,.]$ denotes concatenation operation, and $\odot$ is the Hadamard product.

3.3. Answer Decoders and Loss Functions

Following [10, 30, 6], we adopt different answer decoders depending on the task. Open-ended questions are treated as multi-label classification problems. For these, we employ a classifier which takes as input the combination of the retrieved information from visual cue $\tilde{o}$ and the question representation q, and computes label probabilities $p \in \mathbb{R}^{|\mathcal{A}|}$:

$$y = \text{ELU}\left(W_o\left[\tilde{o}, W_q q + b\right] + b\right), \tag{7}$$

$$y' = \text{ELU}\left(W_y y + b\right), \tag{8}$$

$$p = \text{softmax}\left(W_{y'} y' + b\right). \tag{9}$$

The cross-entropy is used as the loss function.

For repetition count task, we use a linear regression function taking y' in Eq. (8) as input, followed by a rounding function for integer count results. The loss for this task is Mean Squared Error (MSE).

For multi-choice question types (such as repeating action and state transition in TGIF-QA), each answer candidate is processed in the same way with the question. In detail, we use the shared parameter HCRNs with either question or each answer candidate as language cues. As a result, we have a set of HCRN outputs, one conditioned on question ($\tilde{o}_q$), and the others conditioned on answer candidates ($\tilde{o}_a$). Subsequently, $\tilde{o}_q$, $\{\tilde{o}_a\}$, question representation q and answer candidates a are fed into a final classifier with a linear regression to output an answer index, as follows:

$$y = [\tilde{o}_q, \tilde{o}_a, W_q q + b, W_a a + b], \tag{10}$$

$$y' = \text{ELU}\left(W_y y + b\right), \tag{11}$$

$$s = W_{y'} y' + b. \tag{12}$$

We use the popular hinge loss [10] of pairwise comparisons, $\max\left(0, 1 + s^n - s^p\right)$, between scores for incorrect s^n and correct answers s^p to train the network.

3.4. Complexity Analysis

We provide a brief analysis here, leaving detailed derivations in Supplement. For a fixed sampling resolution t, a single forward pass of CRN would take quadratic time in $k_{\max}$. For an input array of length n, feature size F, the unit produces an output array of size $k_{\max}-1$ of the same feature dimensions. The overall complexity of HCRN depends on design choice for each CRN unit and specific arrangement of CRN units. For clarity, let $t=2$ and $k_{\max}=n-1$, which are found to work well in later experiments. Suppose there are N clips of length T, making a video of length $L=NT$. A 2-level architecture of Fig. 3 needs $2TLF$ time to compute the CRNs at the lowest level, and $2NLF$ time to compute the second level, totaling $2(T+N)LF$ time.

Let us now analyze a 3-level architecture that generalizes the one in Fig. 3. The N clips are organized into M sub-videos, each has Q clips, i.e., $N=MQ$. The clip-level CRNs remain the same. At the next level, each sub-video CRN takes as input an array of length Q, whose elements have size $(T-4)F$. Using the same logic as before, the set of sub-video-level CRNs cost $2\frac{N}{M}LF$ time. A stack of two sub-video CRNs now produces an output array of size $(Q-4)(T-4)F$, serving as an input object in an array of length M for the video-level CRNs. Thus the video-level CRNs cost $2MLF$. Thus the total cost for 3-level HCRN is in the order of $2(T+\frac{N}{M}+M)LF$.

Compared to the 2-level HCRN, the a 3-level HCRN reduces computation time by $2(N-\frac{N}{M}-M)LF \approx 2NLF$ assuming $N \gg \max\{M, \frac{N}{M}\}$. As $N=\frac{L}{T}$, this reduces to $2NLF = 2\frac{L^2}{T}F$. In practice T is often fixed, thus the saving scales quadratically with video length L, suggesting that hierarchy is computational efficient for long videos.

4. Experiments

4.1. Datasets

TGIF-QA [10] This is currently the most prominent dataset for VideoQA, containing 165K QA pairs and 72K animated GIFs. The dataset covers four tasks addressing unique properties of video. Of which, the first three require strong spatio-temporal reasoning abilities*: Repetition Count* - to retrieve number of occurrences of an action, *Repeating Action*- multi-choice task to identify the action repeated for a given number of times, *State Transition* - multi-choice tasks regarding temporal order of events. The last task - *Frame QA* - is akin to image QA where a particular frame in a video is sufficient to answer the questions.

MSVD-QA [39] This is a small dataset of 50,505 question answer pairs annotated from 1,970 short clips. Questions are of five types, including what, who, how, when and where.

MSRVTT-QA [40] The dataset contains 10K videos and 243K question answer pairs. Similar to MSVD-QA, questions are of five types. Compared to the other two datasets, videos in MSRVTT-QA contain more complex scenes. They are also much longer, ranging from 10 to 30 seconds long, equivalent to 300 to 900 frames per video.

Model	Action	Trans.	Frame	Count
ST-TP [10]	62.9	69.4	49.5	4.32
Co-mem [7]	68.2	74.3	51.5	4.10
PSAC [20]	70.4	76.9	55.7	4.27
HME [6]	73.9	77.8	53.8	4.02
HCRN	**75.0**	**81.4**	**55.9**	**3.82**

Table 2. Comparison with the state-of-the-art methods on TGIF-QA dataset. For count, the lower the better.

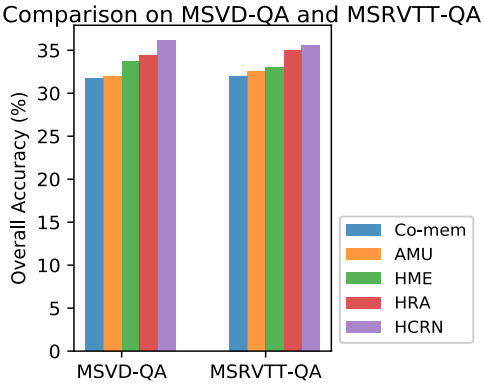

Figure 4. Performance comparison on MSVD-QA and MSRVTT-QA dataset with state-of-the-art methods: Co-mem [7], HME [6], HRA [2], and AMU [39].

We use accuracy to be the evaluation metric for all experiments, except those for repetition count on TGIF-QA dataset where Mean Square Error (MSE) is applied.

4.2. Implementation Details

Videos are segmented into 8 clips, each clip contains 16 frames by default. Long videos in MSRVTT-QA are additionally segmented into 24 clips for evaluating the ability of handling very long sequences. Unless otherwise stated, the default setting is with a 2-level HCRN as depicted in Fig. 3, and $d=512$, $t=1$. We train the model initially at learning rate of 10^{-4} and decay by half after every 10 epochs. All experiments are terminated after 25 epochs and reported results are at the epoch giving the best validation accuracy. Pytorch implementation of the model is available online[1].

4.3. Results

4.3.1 Benchmarking against SoTAs

We compare our proposed model with state-of-the-art methods (SoTAs) on aforementioned datasets. For TGIF-QA, we compare with most recent SoTAs, including [6, 7, 10, 20],

[1]https://github.com/thaolmk54/hcrn-videoqa

Model	Appear.	Motion	Hiera.	Relation
ST-TP [10]	✓	✓		
Co-mem [7]	✓	✓		
PSAC [20]	✓			
HME [6]	✓	✓		
HCRN	✓	✓	✓	✓

Table 3. Model design choices and input modalities in comparison. See Table 2 for corresponding performance on TGIF-QA dataset.

over four tasks. These works, except for [20], make use of motion features extracted from optical flow or 3D CNNs.

The results are summarized in Table 2 for TGIF-QA, and in Fig. 4 for MSVD-QA and MSRVTT-QA. Reported numbers of the competitors are taken from the original papers and [6]. It is clear that our model consistently outperforms or is competitive with SoTA models on all tasks across all datasets. The improvements are particularly noticeable when strong temporal reasoning is required, i.e., for the questions involving actions and transitions in TGIF-QA. These results confirm the significance of considering both near-term and far-term temporal relations toward finding correct answers.

The MSVD-QA and MSRVTT-QA datasets represent highly challenging benchmarks for machine compared to the TGIF-QA, thanks to their open-ended nature. Our model HCRN outperforms existing methods on both datasets, achieving 36.1% and 35.6% accuracy which are 1.7 points and 0.6 points improvement on MSVD-QA and MSRVTT-QA, respectively. This suggests that the model can handle both small and large datasets better than existing methods.

Finally, we provide a justification for the competitive performance of our HCRN against existing rivals by comparing model features in Table 3. Whilst it is not straightforward to compare head-to-head on internal model designs, it is evident that effective video modeling necessitates handling of motion, temporal relation and hierarchy at the same time. We will back this hypothesis by further detailed studies in Section 4.3.2 (for motion, temporal relations, shallow hierarchy) and Section 4.3.3 (deep hierarchy).

4.3.2 Ablation Studies

To provide more insight about our model, we conduct extensive ablation studies on TGIF-QA with a wide range of configurations. The results are reported in Table 4. *Full 2-level HCRN* denotes the full model of Fig. 3 with $k_{max} = n-1, t = 2$. Overall we find that ablating any of design components or CRN units would degrade the performance for temporal reasoning tasks (actions, transition and action counting). The effects are detailed as follows.

Effect of relation order k_{max} and resolution t Without relations ($k_{max} = 1$) the performance drops significantly on actions and events reasoning. This is expected since those questions often require putting actions and events in relation

Model	Act.	Trans.	F.QA	Count
Relations (k_{max}, t)				
$k_{max} = 1, t = 1$	65.2	75.5	54.9	3.97
$k_{max} = 1, t = 3$	66.2	76.2	55.7	3.95
$k_{max} = 1, t = 5$	65.4	76.7	56.0	3.91
$k_{max} = 1, t = 9$	65.6	75.6	56.3	3.92
$k_{max} = 1, t = 11$	65.4	75.1	56.3	3.91
$k_{max} = 2, t = 2$	67.2	76.6	56.7	3.94
$k_{max} = 2, t = 9$	66.3	76.7	56.5	3.92
$k_{max} = 4, t = 2$	64.0	75.9	56.2	3.87
$k_{max} = 4, t = 9$	66.3	75.6	55.8	4.00
$k_{max} = \lfloor n/2 \rfloor, t = 2$	73.3	81.7	56.1	3.89
$k_{max} = \lfloor n/2 \rfloor, t = 9$	72.5	81.1	56.6	3.82
$k_{max} = n-1, t = 1$	75.0	81.4	55.9	3.82
$k_{max} = n-1, t = 3$	75.1	81.5	55.5	3.91
$k_{max} = n-1, t = 5$	73.6	82.0	54.7	3.84
$k_{max} = n-1, t = 7$	75.4	81.4	55.6	3.86
$k_{max} = n-1, t = 9$	74.1	81.9	54.7	3.87
Hierarchy				
1-level, video CRN only	66.2	78.4	56.6	3.94
1.5-level, clips→pool	70.4	80.5	56.6	3.94
Motion conditioning				
w/o motion	70.8	79.8	56.4	4.38
w/o short-term motion	74.9	82.1	56.5	4.03
w/o long-term motion	75.1	81.3	56.7	3.92
Linguistic conditioning				
w/o linguistic condition	66.5	75.7	56.2	3.97
w/o quest.@clip level	74.3	81.1	55.8	3.95
w/o quest.@video level	74.0	80.5	55.9	3.92
Gating				
w/o gate	74.1	82.0	55.8	3.93
w/ gate quest. & motion	73.3	80.9	55.3	3.90
Full 2-level HCRN	75.1	81.2	55.7	3.88

Table 4. Ablation studies on TGIF-QA dataset. For count, the lower the better. Act.: Action; Trans.: Transition; F.QA: Frame QA. When not explicitly specified, we use $k_{max} = n-1, t = 2$ for relation order and sampling resolution.

with a larger context (e.g., what happens before something else). In this case, the frame QA benefits more from increasing sampling resolution t because of better chance to find a relevant frame. However, when taking relations into account ($k_{max} > 1$), we find that HCRN is robust against sampling resolution t but depends critically on the maximium relation order k_{max}. The relative independence w.r.t. t can be due to visual redundancy between frames, so that resampling may capture almost the same information. On the other hand, when considering only low-order object relations, the performance is significantly dropped in all tasks, except frame QA. These results confirm that high-order relations are required for temporal reasoning. As the frame QA task requires only reasoning on a single frame, incorporating temporal information might confuse the model.

Depth of hierarchy	Overall Acc.
2-level, 24 clips $\rightarrow$ 1 vid	35.6
3-level, 24 clips $\rightarrow$ 4 sub-vids $\rightarrow$ 1 vid	35.6

Table 5. Results for going deeper hierarchy on MSRVTT-QA dataset. Run time is reduced by factor of 4 for going from 2-level to 3-level hierarchy.

Effect of hierarchy We design two simpler models with only one CRN layer: ▶ 1-*level,* 1 *CRN video on key frames only*: Using only one CRN at the video-level whose input array consists of key frames of the clips. Note that video-level motion features are still maintained. ▶ 1.5-*level, clip CRNs* $\rightarrow$ *pooling*: Only the clip-level CRNs are used, and their outputs are mean-pooled to represent video. The pooling operation represents a simplistic relational operation across clips. The results confirm that a hierarchy is needed for high performance on temporal reasoning tasks.

Effect of motion conditioning We evaluate the following settings: ▶ *w/o short-term motions*: Remove all CRN units that condition on the short-term motion features (clip level) in the HCRN. ▶ *w/o long-term motions*: Remove the CRN unit that conditions on the long-term motion features (video level) in the HCRN. ▶ *w/o motions*: Remove motion feature from being used by HCRN. We find that motion, in agreeing with prior arts, is critical to detect actions, hence computing action count. Long-term motion is particularly significant for counting task, as this task requires maintaining global temporal context during the entire process. For other tasks, short-term motion is usually sufficient. E.g. in action task, wherein one action is repeatedly performed during the entire video, long-term context contributes little. Not surprisingly, motion does not play the positive role in answering questions on single frames as only appearance information needed.

Effect of linguistic conditioning and gating Linguistic cues represent a crucial context for selecting relevant visual artifacts. For that we test the following ablations: ▶ *w/o quest.@clip level*: Remove the CRN unit that conditions on question representation at clip level. ▶ *w/o quest.@video level*: Remove the CRN unit that conditions on question representation at video level. ▶ *w/o linguistic condition:* Exclude all CRN units conditioning on linguistic cue while the linguistic cue is still in the answer decoder. Likewise, gating offers a selection mechanism. Thus we study its effect as follows: ▶ *wo/ gate*: Turn off the self-gating mechanism in all CRN units. ▶ *w/ gate quest. & motion*: Turn on the self-gating mechanism in all CRN units.

We find that the conditioning question provides an important context for encoding video. Conditioning features (motion and language), through the gating mechanism in Eq. (2), offers further performance gain in action and counting tasks, possibly by selectively passing question-relevant information up the inference chain.

4.3.3 Deepening model hierarchy

We test the scalability of the HCRN on long videos in the MSRVTT-QA dataset, which are organized into 24 clips (3 times longer than other two datasets). We consider two settings: ▶ 2-*level hierarchy,* 24 *clips*$\rightarrow$1 *vid*: The model is as illustrated in Fig. 3, where 24 clip-level CRNs are followed by a video-level CRN. ▶ 3-*level hierarchy*, 24 *clips*$\rightarrow$4 *sub-vids*$\rightarrow$1 *vid*: Starting from the 24 clips as in the 2-level hierarchy, we group 24 clips into 4 sub-videos, each is a group of 6 consecutive clips, resulting in a 3-level hierarchy. These two models are designed to have similar number of parameters, approx. 50M.

The results are reported in Table 5. Unlike existing methods which usually struggle with handling long videos, our method is scalable for them by offering deeper hierarchy, as analyzed theoretically in Section 3.4. Using a deeper hierarchy is expected to significantly reduce the training time and inference time for HCRN, especially when the video is long. In our experiments, we achieve *4 times reduction in training and inference time* by going from 2-level HCRN to 3-level counterpart whilst maintaining the same performance.

5. Discussion

We introduced a general-purpose neural unit called Conditional Relational Networks (CRNs) and a method to construct hierarchical networks for VideoQA using CRNs as building blocks. A CRN is a relational transformer that encapsulates and maps an array of tensorial objects into a new array of the same kind, conditioned on a contextual feature. In the process, high-order relations among input objects are encoded and modulated by the conditioning feature. This design allows flexible construction of sophisticated structure such as stack and hierarchy, and supports iterative reasoning, making it suitable for QA over multimodal and structured domains like video. The HCRN was evaluated on multiple VideoQA datasets (TGIF-QA, MSVD-QA, MSRVTT-QA) demonstrating competitive reasoning capability.

Different to temporal attention based approaches which put effort into selecting objects, HCRN concentrates on modeling relations and hierarchy in video. This difference in methodology and design choices leads to distinctive benefits. CRN units can be further augmented with attention mechanisms to cover better object selection ability, so that related tasks such as frame QA can be further improved.

The examination of CRN in VideoQA highlights the importance of building generic neural reasoning unit that supports native multimodal interaction in improving robustness of visual reasoning. We wish to emphasize that the unit is general-purpose, and hence is applicable for other reasoning tasks, which we will explore. These includes an extension to consider the accompanying linguistic channels which are crucial for TVQA [17] and MovieQA [33] tasks.

References

[1] Lorenzo Baraldi, Costantino Grana, and Rita Cucchiara. Hierarchical boundary-aware neural encoder for video captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1657–1666, 2017. 2

[2] Muhammad Iqbal Hasan Chowdhury, Kien Nguyen, Sridha Sridharan, and Clinton Fookes. Hierarchical relational attention for video question answering. In *2018 25th IEEE International Conference on Image Processing (ICIP)*, pages 599–603. IEEE, 2018. 4

[3] Djork-Arné Clevert, Thomas Unterthiner, and Sepp Hochreiter. Fast and Accurate Deep Network Learning by Exponential Linear Units (ELUs). *arXiv preprint arXiv:1511.07289*, 2015. 3.1

[4] Yann N Dauphin, Angela Fan, Michael Auli, and David Grangier. Language modeling with gated convolutional networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 933–941. JMLR. org, 2017. 3.1

[5] Kien Do, Truyen Tran, and Svetha Venkatesh. Learning deep matrix representations. *arXiv preprint arXiv:1703.01454*, 2018. 2

[6] 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*, pages 1999–2007, 2019. 1, 2, 3.3, 4.2, 4, 4.3.1

[7] Jiyang Gao, Runzhou Ge, Kan Chen, and Ram Nevatia. Motion-appearance co-memory networks for video question answering. *CVPR*, 2018. 1, 2, 4.2, 4, 4.3.1

[8] Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 6546–6555, 2018. 3

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CVPR*, 2016. 1, 2, 3, 3.2

[10] Yunseok Jang, Yale Song, Youngjae Yu, Youngjin Kim, and Gunhee Kim. Tgif-qa: Toward spatio-temporal reasoning in visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2758–2766, 2017. 2, 3.3, 3.3, 4.1, 4.2, 4.3.1

[11] Weike Jin, Zhou Zhao, Mao Gu, Jun Yu, Jun Xiao, and Yueting Zhuang. Multi-interaction network with object relation for video question answering. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 1193–1201. ACM, 2019. 2

[12] Junyeong Kim, Minuk Ma, Kyungsu Kim, Sungjin Kim, and Chang D Yoo. Progressive attention memory network for movie story question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8337–8346, 2019. 2

[13] Jin-Hwa Kim, Jaehyun Jun, and Byoung-Tak Zhang. Bilinear attention networks. In *Advances in Neural Information Processing Systems*, pages 1564–1574, 2018. 2

[14] Kyung-Min Kim, Seong-Ho Choi, Jin-Hwa Kim, and Byoung-Tak Zhang. Multimodal dual attention memory for video story question answering. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 673–688, 2018. 2

[15] Kyung-Min Kim, Min-Oh Heo, Seong-Ho Choi, and Byoung-Tak Zhang. DeepStory: video story QA by deep embedded memory networks. In *IJCAI*, pages 2016–2022. AAAI Press, 2017. 2

[16] Thao Minh Le, Vuong Le, Svetha Venkatesh, and Truyen Tran. Learning to reason with relational video representation for question answering. *arXiv preprint arXiv:1907.04553*, 2019. 2

[17] Jie Lei, Licheng Yu, Mohit Bansal, and Tamara L Berg. Tvqa: Localized, compositional video question answering. *Conference on Empirical Methods in Natural Language Processing*, 2018. 1, 2, 5

[18] Fu Li, Chuang Gan, Xiao Liu, Yunlong Bian, Xiang Long, Yandong Li, Zhichao Li, Jie Zhou, and Shilei Wen. Temporal modeling approaches for large-scale youtube-8m video understanding. *CVPR workshop*, 2017. 2

[19] Xiangpeng Li, Lianli Gao, Xuanhan Wang, Wu Liu, Xing Xu, Heng Tao Shen, and Jingkuan Song. Learnable aggregating net with diversity learning for video question answering. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 1166–1174. ACM, 2019. 2

[20] Xiangpeng Li, Jingkuan Song, Lianli Gao, Xianglong Liu, Wenbing Huang, Xiangnan He, and Chuang Gan. Beyond RNNs: Positional Self-Attention with Co-Attention for Video Question Answering. In *AAAI*, 2019. 1, 2, 4.2, 4.3.1

[21] Junwei Liang, Lu Jiang, Liangliang Cao, Li-Jia Li, and Alexander G Hauptmann. Focal visual-text attention for visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6135–6143, 2018. 2

[22] Rainer Lienhart. Abstracting home video automatically. In *Proceedings of the seventh ACM international conference on Multimedia (Part 2)*, pages 37–40. ACM, 1999. 2

[23] Feng Mao, Xiang Wu, Hui Xue, and Rong Zhang. Hierarchical video frame sequence representation with deep convolutional graph network. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 0–0, 2018. 2

[24] Seil Na, Sangho Lee, Jisung Kim, and Gunhee Kim. A read-write memory network for movie story understanding. In *International Conference on Computer Vision (ICCV 2017). Venice, Italy*, 2017. 1, 2

[25] Pingbo Pan, Zhongwen Xu, Yi Yang, Fei Wu, and Yueting Zhuang. Hierarchical recurrent neural encoder for video representation with application to captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1029–1038, 2016. 2

[26] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *EMNLP*, volume 14, pages 1532–1543, 2014. 11

[27] Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. Film: Visual reasoning with a general conditioning layer. In *AAAI*, 2018. 1, 2

[28] Zhaofan Qiu, Ting Yao, and Tao Mei. Learning spatio-temporal representation with pseudo-3d residual networks. In

Proceedings of the IEEE International Conference on Computer Vision, pages 5533–5541, 2017. 2

[29] Gursimran Singh, Leonid Sigal, and James J Little. Spatio-temporal relational reasoning for video question answering. In *BMVC*, 2019. 2

[30] Xiaomeng Song, Yucheng Shi, Xin Chen, and Yahong Han. Explore multi-step reasoning in video question answering. In *2018 ACM Multimedia Conference on Multimedia Conference*, pages 239–247. ACM, 2018. 1, 2, 3.3

[31] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015. 1, 2, 3.2

[32] Yongyi Tang, Xing Zhang, Lin Ma, Jingwen Wang, Shaoxiang Chen, and Yu-Gang Jiang. Non-local netvlad encoding for video classification. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 0–0, 2018. 2

[33] Makarand Tapaswi, Yukun Zhu, Rainer Stiefelhagen, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. Movieqa: Understanding stories in movies through question-answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4631–4640, 2016. 1, 2, 5

[34] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 6450–6459, 2018. 2

[35] Anran Wang, Anh Tuan Luu, Chuan-Sheng Foo, Hongyuan Zhu, Yi Tay, and Vijay Chandrasekhar. Holistic multi-modal memory network for movie question answering. *IEEE Transactions on Image Processing*, 29:489–499, 2019. 2

[36] Bo Wang, Youjiang Xu, Yahong Han, and Richang Hong. Movie question answering: Remembering the textual cues for layered visual contents. *AAAI'18*, 2018. 1, 2

[37] Chao-Yuan Wu, Christoph Feichtenhofer, Haoqi Fan, Kaiming He, Philipp Krahenbuhl, and Ross Girshick. Long-term feature banks for detailed video understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 284–293, 2019. 2

[38] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1492–1500, 2017. 3

[39] 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 *Proceedings of the 25th ACM international conference on Multimedia*, pages 1645–1653. ACM, 2017. 2, 4.1, 4

[40] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5288–5296, 2016. 2, 4.1

[41] Tianhao Yang, Zheng-Jun Zha, Hongtao Xie, Meng Wang, and Hanwang Zhang. Question-aware tube-switch network for video question answering. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 1184–1192. ACM, 2019. 2

[42] Yunan Ye, Zhou Zhao, Yimeng Li, Long Chen, Jun Xiao, and Yueting Zhuang. Video question answering via attribute-augmented attention network learning. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 829–832. ACM, 2017. 2

[43] Zhou Yu, Jun Yu, Jianping Fan, and Dacheng Tao. Multi-modal factorized bilinear pooling with co-attention learning for visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 1821–1830, 2017. 2

[44] Kuo-Hao Zeng, Tseng-Hung Chen, Ching-Yao Chuang, Yuan-Hong Liao, Juan Carlos Niebles, and Min Sun. Leveraging video descriptions to learn video question answering. In *Thirty-First AAAI Conference on Artificial Intelligence*, 2017. 2

[45] Zhou Zhao, Xinghua Jiang, Deng Cai, Jun Xiao, Xiaofei He, and Shiliang Pu. Multi-turn video question answering via multi-stream hierarchical attention context network. In *IJCAI*, pages 3690–3696, 2018. 2

[46] Zhou Zhao, Qifan Yang, Deng Cai, Xiaofei He, and Yueting Zhuang. Video question answering via hierarchical spatio-temporal attention networks. In *IJCAI*, pages 3518–3524, 2017. 2

[47] Zhou Zhao, Zhu Zhang, Shuwen Xiao, Zhenxin Xiao, Xiaohui Yan, Jun Yu, Deng Cai, and Fei Wu. Long-form video question answering via dynamic hierarchical reinforced networks. *IEEE Transactions on Image Processing*, 28(12):5939–5952, 2019. 2

[48] Bolei Zhou, Alex Andonian, Aude Oliva, and Antonio Torralba. Temporal relational reasoning in videos. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 803–818, 2018. 2, 3.1

[49] Linchao Zhu, Zhongwen Xu, Yi Yang, and Alexander G Hauptmann. Uncovering the temporal context for video question answering. *International Journal of Computer Vision*, 124(3):409–421, 2017. 2