Learning to Reason with Relational Video Representation for 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

How does machine learn to reason about the content of a video in answering a question? A Video QA system must simultaneously understand language, represent visual content over space-time, and iteratively transform these representations in response to lingual content in the query, and finally arriving at a sensible answer. While recent advances in textual and visual question answering have come up with sophisticated visual representation and neural reasoning mechanisms, major challenges in Video QA remain on dynamic grounding of concepts, relations and actions to support the reasoning process. We present a new end-to-end layered architecture for Video QA, which is composed of a question-guided video representation layer and a generic reasoning layer to produce answer. The video is represented using a hierarchical model that encodes visual information about objects, actions and relations in space-time given the textual cues from the question. The encoded representation is then passed to a reasoning module, which in this paper, is implemented as a MAC net. The system is evaluated on the SVQA (synthetic) and TGIF-QA datasets (real), demonstrating state-of-the-art results, with a large margin in the case of multi-step reasoning.

1. Introduction

How can machine learn to reason about a dynamic scene as human does? A powerful demonstration of such a capability is answering unseen natural questions about a video. Recall that reasoning is the mental faculty to produce new knowledge from the previously acquired knowledge base in response to a query [2]. Thus the task of video question answering (Video QA) boils down to learning to acquire and manipulate visual knowledge distributed through space and time conditioned on the compositional linguistic cues. Although it is tempting to extrapolate from the recent successes on visual QA over static images [1, 10, 33, 35], Video QA is relatively new and great challenges remain [31, 22].

While acquiring visual knowledge of objects and relations from static images has advanced hugely in recent years [7], deep video understanding remains elusive. Compared to static images, video poses new challenges, primarily due to the inherent dynamic nature of visual content over time [6, 34]. At the lowest level, we have correlated motion and appearance [6]. At a higher level, we have objects that are persistent over time, actions that are local in time, and the relations that can span over an extended length. Thus searching for an answer from a video facilitates solving simultaneous sub-tasks in both the visual and lingual spaces, probably in an iterative and compositional fashion. In the visual space, the sub-tasks at each step involve extracting and attending to objects, actions, and relations in time and space. In the textual space, the tasks involve extracting and attending to concepts in the context of sentence semantics.

A plausible approach to Video QA is to prepare video content to accommodate the retrieval of information specified in the question [12, 17, 37]. But this has not yet offered a more complex reasoning capability that involves multi-step inference and handling of compositionality. More recent works have attempted to add limited reasoning capability into the system through memory and attention mechanisms [6, 22]. Little attention has been paid for objects, actions, and relations over space-time.

Our approach to Video QA is to separate the processes of *visual representation* and *reasoning,* conditioned on the textual cues. This division of labor realizes a *dual-process* cognitive view that the two processes are qualitatively different, in that visual recognition can be reactive but reasoning is usually deliberative [3, 14]. In our system, visual representation precedes and makes its output accessible to the reasoning process, which is largely domain independent. We have observed this division in visual QA before, e.g., [35] for QA in static images. However, different from [35], we do not seek a neural-symbolic approach, but relying on implicit reasoning capability in a differentiable neural system [2, 11]. More specifically, at the visual understanding level, we derive a hierarchical model over time, dubbed Clip-based Relational Network (CRN), that can accommodate objects,

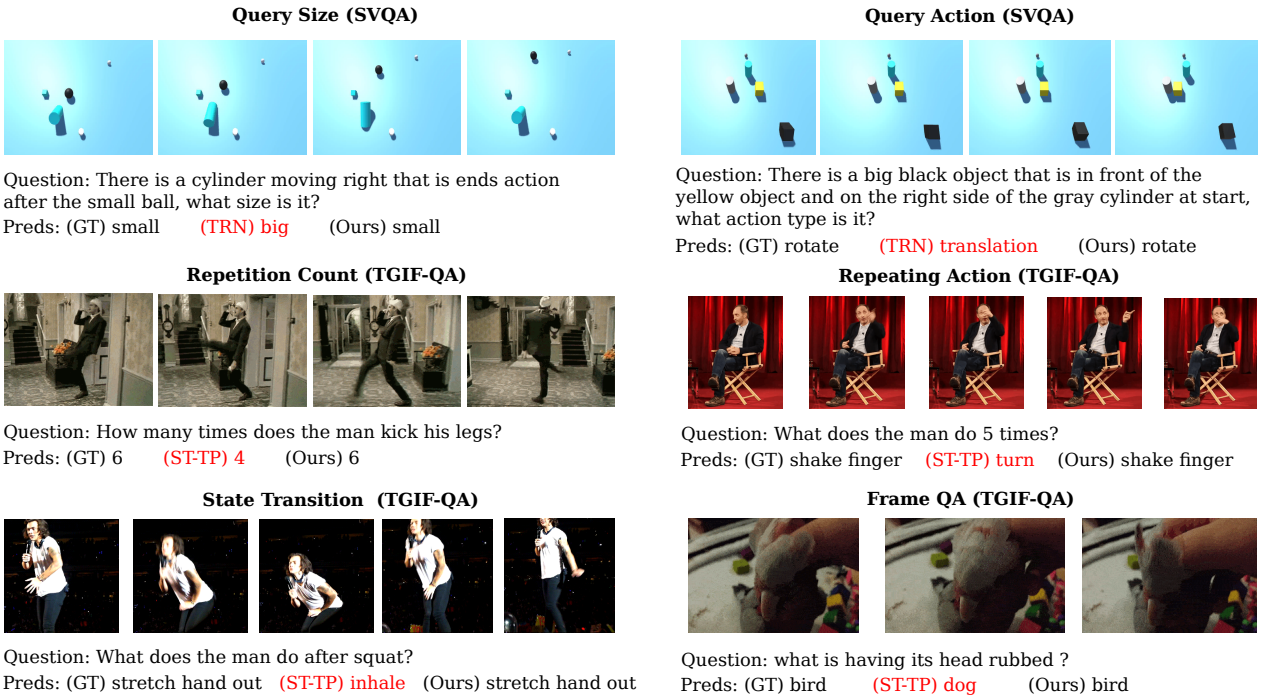

Figure 1. Examples of SVQA and TGIF-QA dataset. GT: groundtruth; TRN: our baseline utilizing TRN of [38]; ST-TP: method introduced in [12]. Best viewed in color.

actions, and relations in space-time. This is followed by a generic reasoning module, known as Memory-Attention-Composition (MAC) network, which takes prepared visual content as a knowledge base, and iteratively co-attends to the textual concepts and the visual concepts/relations to extract the answer [11].

We validate our model on two large public datasets, the TGIF-QA and the SVQA. The TGIF-QA is a real dataset, and is relatively well-studied [6, 12, 22]. See Fig. 1, last two rows for example frames and question types. The SVQA is a new synthetic dataset designed to mitigate the inherent biases in the real datasets and to promote multi-step reasoning [31]. Several cases of complex, multi-part questions are shown in Fig. 1, first row. On both datasets, the proposed model (CRN+MAC) achieves new records, and the margin on the SVQA is qualitatively different from the best known results. Some example responses are displayed in Fig. 1, demonstrating how our proposed method works in different scenarios.

Our contributions are 1) Introducing a new neural architecture for learning to reason in video question answering. The system separates low-level visual representation from high-level reasoning, both conditioned on textual cues. 2) Proposing a hierarchical model for preparing video representation taking into account of query-driven frame selectivity within a clip and temporal relations between clips.

2. Related Work

Video representation in Video QA Similarly to other video understanding tasks, most of the available methods for Video QA relied on recurrent networks or 3D convolutions to extract video features. Variations of LSTM were used in [17] with a bidirectional LSTM, [37] in the form of a two-staged LSTM. Likewise, [6, 22] used two levels of GRU, the first one for extracting "facts" and the second one in each iteration of the memory based reasoning. In another direction, convolutional networks were used to integrate visual information with either 2D or 3D kernels[12, 22].

Different to these two traditional trends, in this work we propose CRN, a query-driven hierarchical relational feature extraction strategy, which supports strong modeling for both near-term and far-term spatio-temporal relations. The model subsumes a recent framework known as Temporal Relation Network (TRN), which has succeeded in action recognition [38] – a special case of Video QA with only one question about action. In our CRN, we generalize the relational extraction framework to support multiple levels of granularity in the temporal scale. This development is necessary to address the nondeterministic queries in video QA tasks.

Reasoning for Video QA Early approaches toward video question answering formulated the problem as a visual information retrieval task. The work in [17, 37] both utilized memory network as a platform to retrieve the information

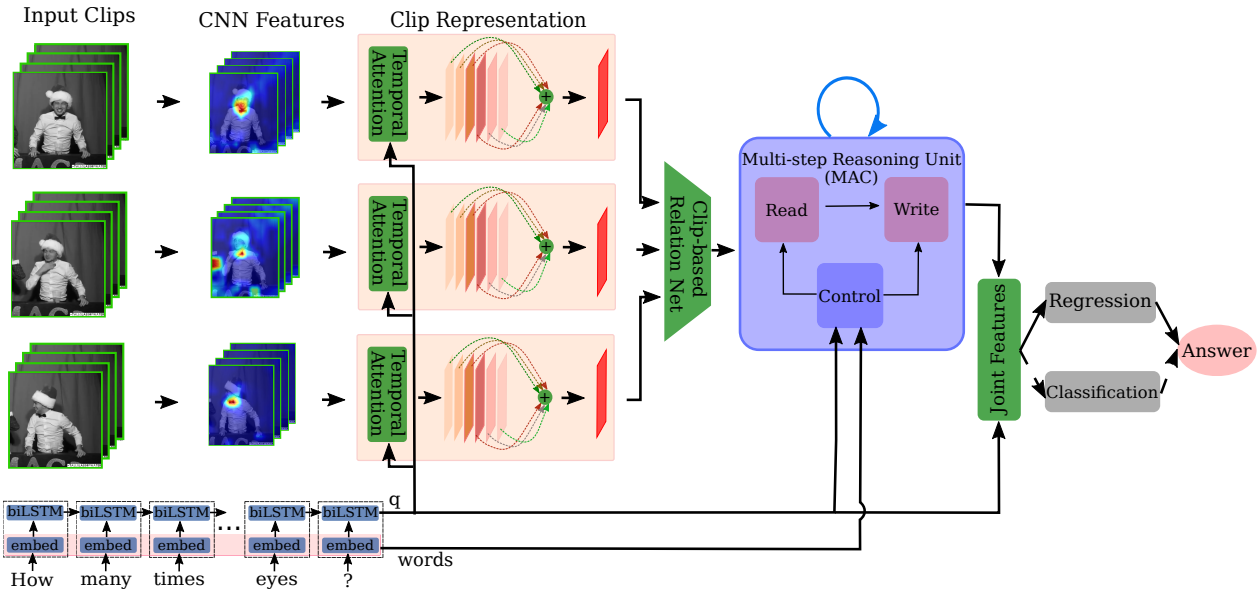

Figure 2. Overview of Network Architecture for Video QA. The model is viewed as a dual-process system of hierarchical video representation with Clip-based Relation Network (CRN) and high-level multi-step reasoning with MAC cells, in which both two processes are guided by textual cues. Inputs of CRN are the aggregated feature representations of equal-size clips obtained by a temporal attention mechanism. The high-level reasoning module takes contextual words of question as well as the vector representation of the question to retrieve relevant information from the output of CRN which acts as a knowledge base. Finally, an output classifier computes the final prediction based on the retrieved information at the final reasoning step and the question vector.

in the video features related to the question embedding to regress for the most relevant output. Beyond information retrieval, answering questions on videos is by nature a spatio-temporal reasoning problem. To achieve this ability, recent Video QA methods started interleaving reasoning mechanisms into the pattern matching network operations. In [12], Jang *et al*. calculated the attention weights on the video LSTM features queried by the question. This attention-based reasoning mechanism aims at identifying the spatio-temporal regions that are relevant to the question but does not support deducing new information based on the data provided. In an effort toward deeper reasoning, Gao *et al*. [6] proposed to parse the two-stream video features through a dynamic co-memory module which iteratively refines the episodic memory. In the most recent development, [22] used self-attention mechanism to internally contemplate about video and question first, then put them through a co-attention block to match the information contained in the two sources of data. The current trend set by these works pushes the sophistication of the reasoning processes on finding the correlation between data pattern and the query without having a real controlling mechanism.

For complex structured videos with multimodal features such as in movies [32] and TV programs [20], recent method leveraged Memory network [17] and its variations [26] to store multimodal features into episodic memory for later retrieval of related information to the question. More sophisticated reasoning mechanisms are also developed with hierarchical attention[23], multi-head attention[16] or multi-step progressive attention memory[15] to jointly reason on video/audio/text concurrent signals.

Compared to the previous work, our framework steps toward specialized but generic powerful reasoning with a centralized reasoning engine controlling attention and memory on rich relational representation. This intensive reasoning module has the distinctive strength of end-to-end differentiability and multi-step self-error-correcting mechanism which are key for robust deep information deducing.

3. Method

In this section, we present our main contribution to addressing the challenges posed in Video QA. In particular, we propose a modular end-to-end neural architecture, as illustrated in Fig. 2.

3.1. Dual-Process System View

Our architecture is partly inspired by the *dual-process theory* dictating that there are two cognitive processes serving separate purposes in reasoning: the lower pattern recognition that tends to be associative, and the higher-order reasoning faculty that tends to be deliberative [3, 14]. Translated into our Video QA scenarios, we have the pattern recognition process for extracting visual features, representing objects and relations, and making the representation accessible to the higher reasoning process. The interesting and challenging

aspects come from two sources. First, video spans over both space and time, and hence calling for methods to deal with object persistence, action span and repetition, and long-range relations. Second, Video QA aims to respond to the textual query, hence the two processes should be conditional, that is, the textual cues will guide both the video representation and reasoning. Although the language coding process is critical to extracting textual cues, it is not our primary concern, and in our system, we make use of the standard bi-LSTM with GloVe word embedding.

The visual representation process has been studied intensively in recent years, in particular for action recognition [38], trajectory modeling [25] and video prediction [24]. Video QA represents a challenging class compared to action recognition because a video can contain more than one trimmed action. In addition, questions may address more complicated relations between entities and actions in the video over time. Therefore, we treat a video as a composition of a number of video clips in this work, each clip can be viewed as an action. While previous studies have explored the importance of hierarchical representation of video [39], it is vital to model the relationships between clips. Inspired by [30] and the recent work [38] on action recognition, known as Temporal Relation Network (TRN), we propose a Clip-based Relation Network (CRN) for video representation, where clip features are selectively query-dependent. It is expected that CRN is more effective in terms of modeling a temporal sequence than the the simplistic TRN which comes with a certain number of sampled frames. The CRN represents the video as a tensor for the use in the reasoning stage.

The reasoning process, due to its deliberative nature, involves multiple steps in an iterative fashion. We utilize Memory-Attention-Composition (MAC) cells [11] for the task due to its generality and modularity. More specially, the MAC cells are called repeatedly conditioned on the textual cues to manipulate information from given video representations as a knowledge base. Finally, information prepared by the MAC, combined with the textual cues is presented to a decoder to produce an answer.

In short, our system consists of three components interacting with each other: (1) hierarchical video representation with CRN, (2) visual multi-step reasoning with MAC cells and (3) answer decoders. We detail these components in what follows.

3.2. Clip-based Relation Network

Given a video of sequential frames, let $C = (C_1, ..., C_K)$ presents K segmented clips with $C_i = (F_{i,1}, ..., F_{i,T})$ being the i-th clip, where F_{ij} is extracted CNN feature of frame j-th in clip C_i, T is the number of frames in a clip. As some frames can be redundant or irrelevant to the question, it is important to selectively attend to frames. We thus utilize

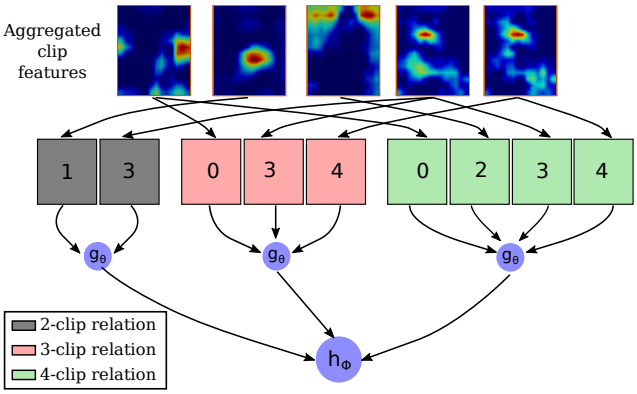

Figure 3. Illustration of Clip-based Relation Network (CRN). Aggregated features of equal size clips are fed into k-clip relation modules. Inputs to relation modules are selected on a random basis whilst keeping their temporal order unchanged. In this figure, our CRN represents a video as aggregated features of five video clips using 2-clip relation, 3-clip relation, and 4-clip relation modules. This results in the final feature of the same shape as one clip feature.

a temporal attention mechanism conditioned on question vector q to compute the aggregated clip feature $\bar{C}_i$ of the corresponding clip C_i:

$$s_{i,j} = w^\top \left(\left(W^{(q)} q + b^{(q)} \right) \odot \left(W^{(r)} F_{i,j} + b^{(r)} \right) \right) \quad (1)$$

$$\bar{C}_i = \sum_{j=1}^{T} F_{i,j} \cdot \text{softmax}(s_{i,j}) \quad (2)$$

where $w, W^{(q)}, W^{(r)}, b^{(q)}$ and $b^{(r)}$ are learnable weights, and $\odot$ is element-wise multiplication.

To account for relations between clips, we borrow the strategy of TRN described in [38] which adapts and generalizes the proposal in [30] to temporal domain. Different from [38], our relational network operates at the clip level instead of frame level. More specifically, the k-order relational representation of video is given as:

$$R^{(k)}(C) = h_\Phi \left(\sum_{i_1 < i_2 ... < i_k} g_\theta \left(\bar{C}_{i_1}, \bar{C}_{i_2}, ..., \bar{C}_{i_k} \right) \right) \quad (3)$$

for $k = 2, 3, .., K$, where h_ϕ and g_θ are any function with parameters ϕ and θ, respectively, for feature fusion. We term this resulted model as *Clip-based Relation Network* (CRN). Fig. 3 illustrates our procedure for our CRN.

Remark The CRN subsumes TRN as a special case when $T \to 1$. However, by computing the relations at the clip level, we can better model the hierarchical structure of video and avoid computational complexity inherent in TRN. For example, we neither need to apply sparse sampling of frames

nor use the multi-resolution trick as in TRN. Consider a lengthy video sequence, in TRN, the chance of having pairs of distantly related frames is high; hence, their relations are less important than those of near-term frames. In the worst case, those pairs can become noise to the feature representation and damage the reasoning process in later stages. In contrast, not only our clips representation can preserve such near-term relations but also the far-term relations between short snippets of the video can be guaranteed with the CRN.

3.3. Multi-step Centralized Reasoning

Higher-order reasoning on the rich relational information from video features is the key for reliably answering questions. The popular approaches for Video QA reasoning conform with static image QA by injecting reasoning mechanisms such as attention and memory into end-to-end visual data processing networks [12, 6]. However, compared to those in images, the video low level features are much more sophisticated and redundant. This complexity and distracting repetitions in feature tensors pose challenges to the saliency locating and record keeping abilities of attention and memory modules inside the mixed up end-to-end networks. We approach this limitation by disentangling the slow, deliberative reasoning steps out of fast, automatic feature extraction and temporal relation modeling. This "slow-thinking" reasoning is done with a dedicated module that iteratively distills and purifies the key relational information contained in the CRN features.

In our experiments, we use the MAC network [11] as the option for the reasoning module. At the core of MAC network are the recurrent cells called *control units*, collaborating with *read units* and *write units* to iteratively make reasoning operations on a knowledge base using a sequence of clues extracted from the query. Compared to mixed-up feature extraction/reasoning mechanisms, the control units give the MAC network distinctive features of a centralized reasoning module that can make a well-informed decision on attention and memory read/writes. MAC is also powered by the flexible retrieving/processing/reference mechanism while processing the query and looking up in the knowledge base. These characteristics are well suited to explore the rich, condensed relational information in CRN features. The iterative reasoning process of MAC supports a level of error self-correcting ability that also helps to deal with the possible remaining redundancy and distraction.

In our setup, the knowledge base B used in MAC network is gathered from the CRN features from all available orders:

$$B = \sum_{k=2}^{K} R^{(k)}(C) \tag{4}$$

where $R^{(k)}(C)$ are the k-order CRN representations calculated as in Eq. (3).

Concurrently, the question text of length N is processed through a bi-LSTM into two sets of clues: contextual words $\{w_1, ..., w_N\}$ which are the output states of the LSTM at each step, and the position-aware question vector q, the joint representation of the final hidden states from forward and backward LSTM passes:

$$q = [\overleftarrow{w_1}, \overrightarrow{w_N}] \tag{5}$$

where $[.]$ denotes a vector concatenation operator.

For each reasoning step i, the relevant aspects of question to this step is estimated from q:

$$q_i = W_i q + b_i \tag{6}$$

where W_i and b_i are learned linear transformation.

Based on the pair of clues a pair of clues (w_n, q_i), the control unit calculates a soft self-attention weight α_n^c:

$$\alpha_n^c = \text{softmax}(W^\top([W^{(c)\top} c_{i-1}, W^{(q)\top} q_i] \odot w_n))$$

and infers the control state at this step:

$$c_i = \sum_{n=1}^{N} \alpha_n^c w_n \tag{7}$$

The read unit uses this control signal and the prior memory m_{i-1} to calculate the read attention weights $\alpha_{i,x,y}^r$ for each location x, y in the knowledge base B and retrieves the related information:

$$r_i = \sum_{x,y} \alpha_{i,x,y}^r B_{x,y} \tag{8}$$

where $\alpha_{i,x,y}^r = \text{softmax}(W^T(c_i \odot [m_{i-1} \odot B_{x,y}, B_{x,y}]))$.

To finish each reasoning iteration, the write unit calculates the intermediate reasoning result m_i by updating previous record m_{i-1} with the new information derived from the retrieved knowledge r_i , say $m_i = f(m_{i-1}, r_i)$. In our experments, the function f is a linear transformation on top of a mere vector concatenation operator.

At the end of the process of P steps, the final memory state m_P emerges as the output of the reasoning module to be used in decoding for answers in the next step.

3.4. Answer Decoders

Following [12, 31], we adopt different answer decoders depending on the task. These include open-ended words, open-ended number, and multi-choice.

For open-ended words (for example, presented in Frame QA in TGIF-QA dataset and all subtasks in SVQA dataset – see Sec. 4.1), we treat this as a multi-label classification problem of $\mathbb{V}$ labels from answer vocabularies. Hence, we employ a linear function that takes the combination of the final memory output m_P of the last reasoning iteration and

output question vector q after bi-LSTM as the input, resulting in the output $o^h = \text{concat}(m_P, q)$; $o^h \in \mathbb{R}^{1,024}$ followed by a softmax layer:

$$p = \textit{softmax}\left(W^{(\omega)\top} o^h + b^{(\omega)}\right) \tag{9}$$

where $p \in \mathbb{R}^{|\mathbb{V}|}$ is a confidence vector of probabilities of labels; and $W^{(\omega)\top} \in \mathbb{R}^{|\mathbb{V}| \times 1,024}$ and $b^{(\omega)} \in \mathbb{R}^{|\mathbb{V}|}$ are weights. The cross-entropy is used as the loss function of the network in this case.

Similarly, we also use a linear regression function to predict real-value numbers (repetition count) directly from the joint output features o^h. We further pass the regression output through a rounding function for prediction:

$$s = W^{(r)\top} o^h \tag{10}$$

where $W^{(r)}$ is network parameters. Mean Squared Error (MSE) is used as the loss function during the training process.

Regarding multi-choice question type, which includes repeating action and state transition in TGIF-QA dataset, we first adopt a linear regression function to project feature space to the answer space of answer candidate indices. The regression function takes the fused features $o^m = \text{concat}(m_P^q, m_P^a, q, a)$; $o^m \in \mathbb{R}^{2,048}$, which is the combination of memory outputs from the final compositional reasoning unit for both question m_P^q and answer candidates m_P^a with the question vector q as well as the corresponding answer option a, as the input. It is import to note that each answer candidate a of multi-choice question goes through the multi-step centralized reasoning module in the same way as the query:

$$s = W^{(m)\top} o^m \tag{11}$$

Accordingly, the model in this case is trained with hinge loss of pairwise comparisons, $\max(0, 1 + s^n - s^p)$, between scores for incorrect s^n and correct answers s^p.

4. Experiments

4.1. Datasets

We evaluate our proposed method on two recent public datasets: Synthetic Video Question Answering (SVQA) [31] and TGIF-QA [12].

SVQA: This dataset is a benchmark for multi-step reasoning introduced by Song *et al.* [31]. Resembling the CLEVR dataset [13] for traditional visual question answering task, SVQA provides long questions with logical structures along with spatial and temporal interactions between objects. SVQA was designed to mitigate several drawbacks of current Video QA datasets including language bias and the shortage of compositional logical structure in questions. It contains 120K QA pairs generated from 12K videos that cover a number of question types such as count, exist, object attributes comparison and query.

TGIF-QA: This is currently the largest dataset for Video QA, containing 165K QA pairs collected from 72K animated GIFs. This dataset covers four task types mostly to address the unique properties of video including repetition count, repeating action, state transition and frame QA. Of the four tasks, the first three demand a strong spatio-temporal reasoning ability. *Repetition Count*: This is one of the most challenging tasks of Video QA where we count the repetitions of an action. For example, one has to answer questions like "How many times does the woman shake hips?". This is defined as an open-ended task with 11 possible answers in total ranging from 0 to 10+. *Repeating Action*: This is a multiple choice task of five answer candidates corresponding to one question. The task is to identify the action that is repeated for a given number of times in the video (e.g. "what does the dog do 4 times?"). *State Transition*: This is also a multiple choice task addressing a special property of video that is to understand the transition between two states. There are certain states characterized in the dataset including facial expressions, actions, places and object properties. Questions like "What does the woman do before turn to the right side?" and "What does the woman do after look left side?" are aimed at identifying previous state and next state, respectively. *Frame QA*: This task is akin to the traditional visual QA where the answer to a question can be distilled from one of the frames in a video. The key frame may locate in any time steps of the video.

4.2. Implementation Details

Each video is segmented into five equal clips, each of which has eight consecutive frames. The middle frame of each clip is determined based on the length of the video. We take *conv4* output features from ResNet-101 [9] pretrained on ImageNet [29] as the visual features of each video frame. Each frame feature has dimensions of $14 \times 14 \times 1024$. Each word in questions and answer candidates in case of multiple choice question is embedded into a vector of dimension 300 and initialized by a pre-trained GloVe [27]. In order to calculate temporal attention in each clip, we first apply mean pooling over entire spatial dimensions of a frame feature and further pass the filtered output through a linear transformation before multiplying it with question vector q. Unless otherwise stated, we use $P = 12$ MAC cells for multi-step reasoning in our network as in [11], while all hidden state sizes are set to 512 for both CRN and MAC units.

Our network is trained using Adam [18], with a learning rate of 5×10^{-5} for repetition count and 10^{-4} for other tasks, and a batch size of 16. The SVQA is split into three parts

Table 1. Ablation studies. (*) For count, the lower the better.

Model	SVQA	TGIF-QA (*)			
		Action	Trans.	Frame	Count
Linguistic only	42.6	51.5	52.8	46.0	4.77
Ling.+S.Frame	44.6	51.3	53.4	50.4	4.63
S.Frame+MAC	58.1	67.8	76.1	57.1	4.41
Avg.Pool+MAC	67.4	70.1	77.7	58.0	4.31
TRN+MAC	70.8	69.0	78.4	58.7	4.33
CRN+MLP	49.3	51.5	53.0	53.5	4.53
CRN+MAC	**75.8**	**71.3**	**78.7**	**59.2**	**4.23**

Table 2. Ablation studies with different reasoning iterations. (*) For count, the lower the better.

Reasoning iterations	TGIF-QA (*)			
	Action	Trans.	Frame	Count
4	69.9	77.6	58.5	4.30
8	70.8	78.8	58.6	4.29
12	**71.3**	**78.7**	**59.2**	**4.23**

with proportions of 70-10-20% for training, cross-validation, and testing set, accordingly. We also take 10% of training videos in each subtask in TGIF-QA as validation set. Similar to the work in [11], we also experience better performance when employing ELU as non-linearity instead of standard ReLU. In addition, we observe a favorable performance when replacing the summation operator in Eq. (3) with a mere concatenation. All experiments are conducted with one single NVIDIA Tesla V100 GPU.

Evaluation Metrics: For the TGIF-QA, to be consistent with prior work [12, 6, 22], we use accuracy as the evaluation metric for all tasks of other than repetition count, which is considered as a regression problem and assessed using Mean Square Error. For the SVQA, we report accuracy for all sub-tasks, which are considered as multi-label classification problems.

4.3. Results

4.3.1 Ablation Studies:

To demonstrate the effectiveness of each component on the overall performance of the proposed network, we first conduct a series of ablation studies on both the SVQA and TGIF-QA datasets. In all ablation studies, we ensure to encode questions and answers (sequences) if needed in the same way as explained above for fair comparisons. The ablation results are presented in Table 1, 2 showing progressive improvements, which justify the added complexity. We explain below the baselines.

Linguistic only: With this baseline, we aim to assess how much linguistic information affects overall performance. This also suggests the linguistic bias of the dataset. From Table 1, it is clear that TGIF-QA is highly linguistically biased while the problem is mitigated with SVQA dataset to some extent.

Ling.+S.Frame: This is a very basic model of VQA that combines the encoded question vector with CNN features of a random frame from a given video. As expected, this baseline gives modest improvements over the model using only linguistic features.

S.Frame+MAC: To demonstrate the significance of multi-step reasoning in Video QA, we randomly select one video frame and then use its CNN feature maps as the knowledge base of the multi-step centralized reasoning module MAC. In other words, this experiment reflects the importance of multi-step reasoning in a question answering problem. As SVQA dataset contains questions with compositional sequences, it greatly benefits from the multi-step centralized reasoning module.

Avg.Pool+MAC: A baseline to assess the significance of temporal information in the simplest form of average temporal pooling comparing to ones with a single frame. We follow [38] to sparely sample 8 frames which are the middle frames of the equal size segments from a given video. As can be seen, this model is able to achieve significant improvements in both SVQA and TGIF-QA. Due to the linguistic bias, the contribution of visual information to the overall performance on TGIF-QA is much less than that of SVQA.

TRN+MAC: This baseline is a special case of ours where we flatten the hierarchy, and the relation network is applied at the frame level, similar to what proposed in [38]. The model mitigates the limit of feature engineering process of using only one single frame for video representation as well as simply temporal pooling over the whole sequence of frames. Apparently, using a single frame loses crucial temporal information of the video and is likely to fail when strong temporal reasoning is needed, particularly in state transition and counting. We use visual features processed in the Avg.Pool+MAC experiment to input into a TRN module for fair comparisons. TRN improves by more than 12% of overall performance from one with a single video frame, while that of state transition task on TGIF-QA is more than 2%, around 1.5% for both repeating action and frame QA and 0.08 MSE in case of repetition count. Although this baseline produces great increments on SVQA comparing to the experiment Avg.Pool+MAC, it barely outperforms those on TGIF-QA.

CRN+MLP: In order to evaluate how the reasoning module affects the overall performance, we conduct this experiment by using a feed-forward network as the reasoning module with the proposed visual representation CRN.

CRN+MAC: This is our proposed method in which we opt CRN as the knowledge base of the compositional attention module. We experience significant improvements on all tasks over the simplistic TRN on SVQA whilst that of

TGIF-QA dataset is more moderate. The results reveal the strong CRN's capability as well as a better suit of video representation for reasoning over the simplistic TRN, especially in case of compositional reasoning. The results are also consistent with our earlier arguments that sparsely sampled frames from the video are insufficient to embrace fast-pace actions/events such as repeating action and repetition count.

4.3.2 Benchmarking against SOTAs:

We also compare our proposed model with other state-of-the-art methods on both two datasets, as shown in Table 3 (SVQA, synthetic) and Table 4 (TGIF-QA, real). As the TGIF-QA is older, much effort has been spent on benchmarking it and significant progress has been made in recent years. The SVQA is new, and hence published results are not very indicative of the latest advance in modeling.

For the SVQA, Table 1 and Table 3 reveal that the contribution of visual information to the overall performance of the best known results is very little. This means their system is mostly based on the linguistic bias of the dataset to make the decision. In contrast, our proposed methods do not seem to suffer from this problem. We establish new qualitatively different SOTAs, jumping massively from 44.9% accuracy to 75.8% accuracy overall.

For the TGIF-QA dataset, Jang *et al*. [12] extended winner models of the VQA 2016 challenge to evaluate on Video QA task, namely VIS+LSTM [28] and VQA-MCB [5]. Early fusion and late fusion are applied to both two approaches. We also list some other methods provided by [12] including those proposed by [5] and [36]. Interestingly, none of the previous work reported ablation study of utilizing only textual cues as the input of the system to assess the linguistic bias of the dataset, and the fact that some of the reported methods produced worse performance than this baseline. We suspect that improper integrating of visual information caused confusion to the system giving such low performance. In [12] , Jang *et al*. also proposed their own methods of leveraging spatial and temporal attention. In Table 4, SP indicates spatial attention, ST presents temporal attention while "R" and "C" mean ResNet features and C3D features, respectively. Later, Gao *et al*. [6] greatly advanced the performance on this dataset with a co-memory mechanism on two video feature streams. Li *et al*. [22] recently set the state of the art performance on TGIF-QA with only ResNet features by using a novel self-attention mechanism. Our method, similarly in the sense that we only use ResNet features, is able to achieve significant improvements on three tasks including repeating action, state transition and frame QA while we are slightly advanced on repetition count task compared to our competitor [22] using the same video features. Specifically, we gain better performance than [22] of around 1% for repeating action, approximately 2% for state transition, 3.5% for frame QA, and 0.04 MSE for repetition count.

For repetition counting in TGIF-QA, although we have not outperformed [6], it is not directly comparable since they utilized motion in addition to appearance, whilst in our case motion is not explicitly used and thus the action boundaries are not clearly detected. We hypothesize that counting task needs a specific network, as evident in recent work [21, 33].

Qualitative Results: Fig. 1 shows example frames and associated question types in the TGIF-QA and the SVQA datasets. The figure also presents corresponding responses by our CRN+MAC, and those by ST-TP method [12] (on TGIF-QA) and TRN+MAC (our own special case of flat video representation, on SVQA) for reference. The questions clearly demonstrate challenges that video QA systems must face such as visual ambiguity, subtlety, compositional language understanding as well as concepts grounding. The questions in SVQA were designed for multi-step reasoning, and the dual-process system of CRN+MAC Net proves to be effective in these cases.

5. Discussion

The proposed layered neural architecture is in line the hypothesis that the high-level reasoning capability (known as System 2 in the literature) came fairly late in the evolution history, probably after reactive perception systems (known as System 1) had been developed (e.g., see [3, 4, 14]). Hence it is plausible that video representation precedes and is accessible to reasoning, e.g., along the line of proposals in [4, 8]. In addition, we observed that the generic reasoning scheme of MAC net is surprisingly powerful for the domain of Video QA, especially for the problems that demand multi-step inference (e.g., on the SVQA dataset)[1]. This suggests that it is worthy to spend effort to advance reasoning functionalities for both general cases and in spatio-temporal settings. We also observed that counting over space-time remains a challenging problem. We conjecture that it might benefit from accurate explicit region proposals for objects and duration proposals for action, rather than the implicit detection as currently implemented.

The dual-process view opens a wide room for future study. As System 1 usually involves short-term sensory memory and System 2 working memory, we expect that creating and linking these two memory components will enhance the capacity of the system (e.g., see [19]). Second, although we have presented a seamless feedforward integration of System 1 and System 2, it still opens on how the two systems interact. For example, it might be beneficial to have a full feed-back loop where the high-level reasoning guides the

[1] We observe in passing that our model loosely resembles the modularity hypothesis pushed forward by Fodor [4]. He suggests that the mind is composed of relatively independent domain-specific modules (e.g., CRNs, bi-LSTM) and a central processing part (e.g., MAC), which is domain-general.

Table 3. Comparison with the state-of-the-art method on SVQA.

	Exist	Count	Integer Comparison			Attribute Comparison					Query					All
			More	Equal	Less	Color	Size	Type	Dir	Shape	Color	Size	Type	Dir	Shape	
SA(S) [31]	51.7	36.3	72.7	54.8	58.6	52.2	53.6	52.7	53.0	52.3	29.0	54.0	55.7	38.1	46.3	43.1
TA-GRU(T) [31]	54.6	36.6	73.0	57.3	57.7	53.8	53.4	54.8	55.1	52.4	22.0	54.8	55.5	41.7	42.9	44.2
SA+TA-GRU [31]	52.0	38.2	74.3	57.7	61.6	56.0	55.9	53.4	57.5	53.0	23.4	63.3	62.9	43.2	41.7	44.9
CRN+MAC	**72.8**	**56.7**	**84.5**	**71.7**	**75.9**	**70.5**	**76.2**	**90.7**	**75.9**	**57.2**	**76.1**	**92.8**	**91.0**	**87.4**	**85.4**	**75.8**

Table 4. Comparison with the state-of-the-art method on TGIF-QA dataset. For count, the lower the better. R: ResNet, C: C3D features, F: flow features.

Model	Action	Trans.	Frame	Count
VIS+LSTM (aggr)[28]	46.8	56.9	34.6	5.09
VIS+LSTM (avg)[28]	48.8	34.8	35.0	4.80
VQA-MCB (aggr)[5]	58.9	24.3	25.7	5.17
VQA-MCB (avg)[5]	29.1	33.0	15.5	5.54
Yu et al.[36]	56.1	64.0	39.6	5.13
ST(R+C)[12]	60.1	65.7	48.2	4.38
ST-SP(R+C)[12]	57.3	63.7	45.5	4.28
ST-SP-TP(R+C)[12]	57.0	59.6	47.8	4.56
ST-TP(R+C)[12]	60.8	67.1	49.3	4.40
ST-TP(R+F)[12]	62.9	69.4	49.5	4.32
Co-memory(R+F)[6]	68.2	74.3	51.5	**4.10**
PSAC(R)[22]	70.4	76.9	55.7	4.27
CRN+MAC(R)	**71.3**	**78.7**	**59.2**	4.23

video representation steps, e.g., through a top-down, iterative attentional scheme. These added capabilities may allow more complex queries over movie-length videos.

5.1. Conclusion

In summary, we have proposed a new layered neural architecture for learning to reason in video question answering. The architecture is founded on the premise that Video QA tasks necessitate a conditional dual-process of associative video representation and deliberative multi-step reasoning, given textual cues. The two processes are ordered in that the former process prepares query-specific representation of video to support the latter reasoning process. With that in mind, we designed a hierarchical relational model for query-guided video representation named Clip-based Relational Network (CRN) and integrated it with a generic neural reasoning module (MAC Net) to infer an answer. The system is fully differentiable and hence amenable to end-to-end training. Compared to existing state-of-the-arts in Video QA, the new system is more open to accommodate a wide range of low-level visual processing and high-level reasoning capabilities. Tested on SVQA (synthetic) and TGIF-QA (real) datasets, the proposed system demonstrates a new state-of-the-art performance in a majority of cases. The gained margin is strongly evident in the case where the system is defined for – multi-step reasoning.

References

[1] 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. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6077–6086, 2018. 1

[2] L. Bottou. From machine learning to machine reasoning. *Machine Learning*, 94(2):133–149, 2014. 1

[3] J. S. B. Evans. Dual-processing accounts of reasoning, judgment, and social cognition. *Annu. Rev. Psychol.*, 59:255–278, 2008. 1, 3.1, 5

[4] J. A. Fodor. *The modularity of mind*. MIT press, 1983. 5, 1

[5] 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. *EMNLP*, 2016. 4.3.2, 4

[6] J. Gao, R. Ge, K. Chen, and R. Nevatia. Motion-appearance co-memory networks for video question answering. *CVPR*, 2018. 1, 2, 2, 3.3, 4.2, 4.3.2, 4

[7] M. Garnelo and M. Shanahan. Reconciling deep learning with symbolic artificial intelligence: representing objects and relations. *Current Opinion in Behavioral Sciences*, 29:17–23, 2019. 1

[8] S. Harnad. The symbol grounding problem. *Physica D: Nonlinear Phenomena*, 42(1-3):335–346, 1990. 5

[9] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. *CVPR*, 2016. 4.2

[10] R. Hu, J. Andreas, M. Rohrbach, T. Darrell, and K. Saenko. Learning to reason: End-to-end module networks for visual question answering. In *Computer Vision (ICCV), 2017 IEEE International Conference on*, pages 804–813. IEEE, 2017. 1

[11] D. A. Hudson and C. D. Manning. Compositional attention networks for machine reasoning. *ICLR*, 2018. 1, 3.1, 3.3, 4.2

[12] Y. Jang, Y. Song, Y. Yu, Y. Kim, and G. 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. 1, 2, 2, 3.3, 3.4, 4.1, 4.2, 4.3.2, 4.3.2, 4

[13] J. Johnson, B. Hariharan, L. van der Maaten, L. Fei-Fei, C. Lawrence Zitnick, and R. Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2901–2910, 2017. 4.1

[14] D. Kahneman. *Thinking, fast and slow*. Farrar, Straus and Giroux New York, 2011. 1, 3.1, 5

[15] J. Kim, M. Ma, K. Kim, S. Kim, and C. D. Yoo. Progressive attention memory network for movie story question answer-

ing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8337–8346, 2019. 2

[16] K.-M. Kim, S.-H. Choi, J.-H. Kim, and B.-T. 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

[17] K.-M. Kim, M.-O. Heo, S.-H. Choi, and B.-T. Zhang. DeepStory: video story QA by deep embedded memory networks. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, pages 2016–2022. AAAI Press, 2017. 1, 2, 2

[18] D. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 4.2

[19] H. Le, T. Tran, T. Nguyen, and S. Venkatesh. Variational memory encoder-decoder. In *NeurIPS*, 2018. 5

[20] J. Lei, L. Yu, M. Bansal, and T. L. Berg. Tvqa: Localized, compositional video question answering. *Conference on Empirical Methods in Natural Language Processing*, 2018. 2

[21] O. Levy and L. Wolf. Live repetition counting. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3020–3028, 2015. 4.3.2

[22] 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. In *AAAI*, 2019. 1, 2, 2, 4.2, 4.3.2, 4

[23] J. Liang, L. Jiang, L. Cao, L.-J. Li, and A. 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

[24] M. Mathieu, C. Couprie, and Y. LeCun. Deep multi-scale video prediction beyond mean square error. *arXiv preprint arXiv:1511.05440*, 2015. 3.1

[25] R. Morais, V. Le, T. Tran, B. Saha, M. Mansour, and S. Venkatesh. Learning regularity in skeleton trajectories for anomaly detection in videos. In *CVPR'19*, 2019. 3.1

[26] S. Na, S. Lee, J. Kim, and G. Kim. A read-write memory network for movie story understanding. In *International Conference on Computer Vision (ICCV 2017). Venice, Italy*, 2017. 2

[27] J. Pennington, R. Socher, and C. D. Manning. Glove: Global vectors for word representation. In *EMNLP*, volume 14, pages 1532–1543, 2014. 4.2

[28] M. Ren, R. Kiros, and R. Zemel. Exploring models and data for image question answering. In *Advances in neural information processing systems*, pages 2953–2961, 2015. 4.3.2, 4

[29] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015. 4.2

[30] A. Santoro, D. Raposo, D. G. Barrett, M. Malinowski, R. Pascanu, P. Battaglia, and T. Lillicrap. A simple neural network module for relational reasoning. In *Advances in neural information processing systems*, pages 4974–4983, 2017. 3.1, 3.2

[31] X. Song, Y. Shi, X. Chen, and Y. Han. Explore multi-step reasoning in video question answering. In *2018 ACM Multimedia Conference on Multimedia Conference*, pages 239–247. ACM, 2018. 1, 3.4, 4.1, 4.1, 3

[32] M. Tapaswi, Y. Zhu, R. Stiefelhagen, A. Torralba, R. Urtasun, and S. 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. 2

[33] A. Trott, C. Xiong, and R. Socher. Interpretable counting for visual question answering. *ICLR*, 2018. 1, 4.3.2

[34] B. Wang, Y. Xu, Y. Han, and R. Hong. Movie question answering: Remembering the textual cues for layered visual contents. *AAAI'18*, 2018. 1

[35] K. Yi, J. Wu, C. Gan, A. Torralba, P. Kohli, and J. Tenenbaum. Neural-symbolic vqa: Disentangling reasoning from vision and language understanding. In *Advances in Neural Information Processing Systems*, pages 1039–1050, 2018. 1

[36] Y. Yu, H. Ko, J. Choi, and G. Kim. End-to-end concept word detection for video captioning, retrieval, and question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3165–3173, 2017. 4.3.2, 4

[37] K.-H. Zeng, T.-H. Chen, C.-Y. Chuang, Y.-H. Liao, J. C. Niebles, and M. Sun. Leveraging video descriptions to learn video question answering. In *Thirty-First AAAI Conference on Artificial Intelligence*, 2017. 1, 2, 2

[38] B. Zhou, A. Andonian, A. Oliva, and A. Torralba. Temporal relational reasoning in videos. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 803–818, 2018. 1, 2, 3.1, 3.2, 4.3.1

[39] J. Zhu, Z. Zhu, and W. Zou. End-to-end video-level representation learning for action recognition. In *2018 24th International Conference on Pattern Recognition (ICPR)*, pages 645–650. IEEE, 2018. 3.1