Neural Reasoning, Fast and Slow, for Video Question Answering

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

***Abstract*—What does it take to design a machine that learns to answer natural questions about a video? 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 lingual and visual question answering have enabled sophisticated representations and neural reasoning mechanisms, major challenges in Video QA remain on dynamic grounding of concepts, relations and actions to support the reasoning process. Inspired by the dual-process account of human reasoning, we design a *dual process neural architecture*, which is composed of a question-guided video processing module (System 1, fast and reactive) followed by a generic reasoning module (System 2, slow and deliberative). System 1 is a hierarchical model that encodes visual patterns about objects, actions and relations in space-time given the textual cues from the question. The encoded representation is a set of high-level visual features, which are then passed to System 2. Here multi-step inference follows to iteratively chain visual elements as instructed by the textual elements. The system is evaluated on the SVQA (synthetic) and TGIF-QA datasets (real), demonstrating competitive results, with a large margin in the case of multi-step reasoning.**

I. Introduction

A long standing goal in AI is to design a learning machine capable of reasoning about dynamic scenes it sees. A powerful demonstration of such a capability is answering unseen natural questions about a video. Video QA systems must be able to learn, acquire and manipulate visual knowledge distributed through space-time conditioned on the compositional linguistic cues. Recent successes in image QA [1], [11], [12], [30] pave possible roads, but Video QA is largely under-explored [27], [21]. Compared to static images, video poses new challenges, primarily due to the inherent dynamic nature of visual content over time [7], [29]. At the lowest level, we have correlated motion and appearance [7]. At higher levels, we have objects that are persistent over time, actions that are local in time, and the relations that can span over an extended length.

Searching for an answer from a video facilitates solving interwoven 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 [14], [19], [32]. 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 that are tightly entangled with visual representation [7], [21]. These systems are thus non-modular, and less comprehensible as a result.

Our approach to Video QA is to disentangle the processes of visual pattern recognition and compositional reasoning [30]. This division of labor realizes a *dual process* cognitive view that the two processes are qualitatively different: visual cognition can be reactive and domain-specific but reasoning is usually deliberative and domain-general [3], [16]. In our system, pattern recognition precedes and makes its output accessible to the reasoning process. More specifically, at the visual understanding level, we derive a hierarchical model over time, dubbed Clip-based Relational Network (CRN), that can accommodate objects, actions, and relations in space-time. This is followed by a generic differentiable reasoning module, known as Memory-Attention-Composition (MAC) network [13], which iteratively manipulates a set of objects in the knowledge base given a set of cues in the query, one step at a time. In our setting, MAC 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. The overall dual-process system is modular and fully differentiable, making it easy to compose modules and train.

We validate our dual process model on two large public datasets, the TGIF-QA and the SVQA. The TGIF-QA is a real dataset, and is relatively well-studied [7], [14], [21]. 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 [27]. 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 modular neural architecture for learning to reason in Video QA. The system implements dual process theory by disentangling reactive visual cognition and language understanding from deliberative reasoning. 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.

978-1-7281-6926-2/20/$31.00 ©2020 IEEE

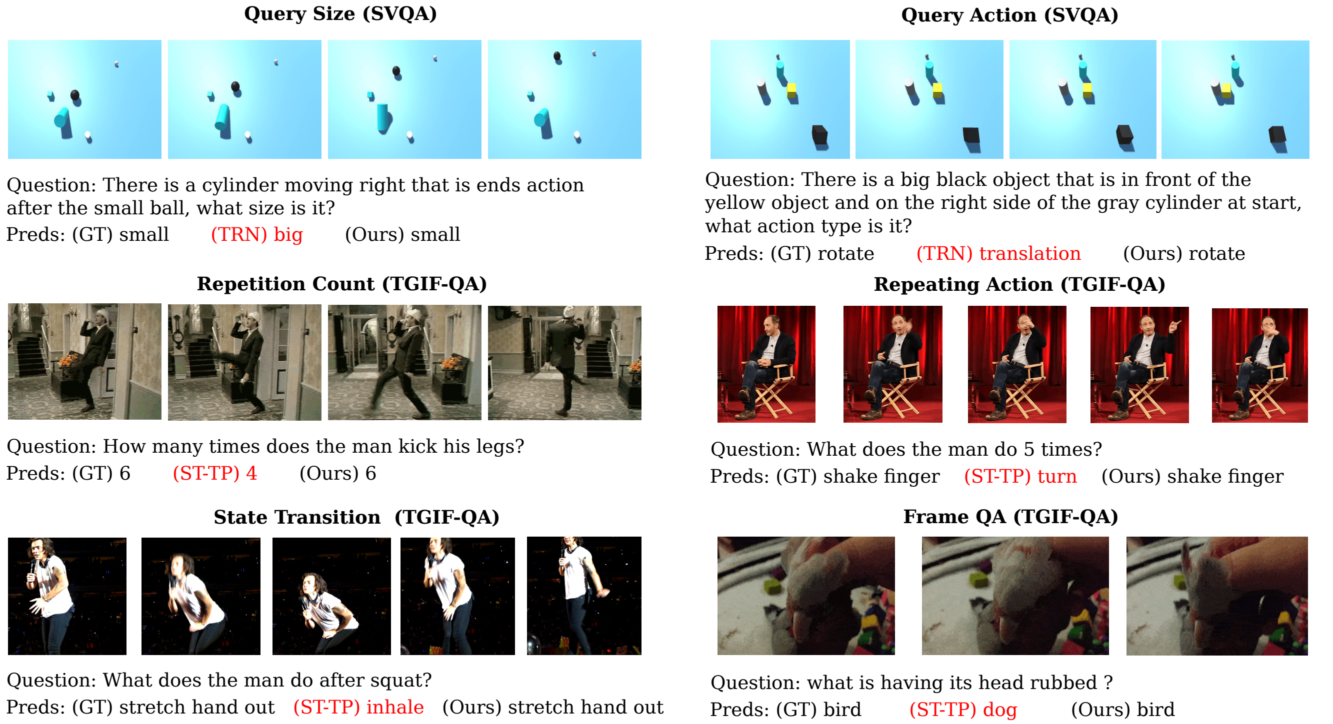

Figure 1. Examples of SVQA and TGIF-QA dataset. GT: ground truth; TRN: our baseline utilizing TRN [33]; ST-TP: method introduced in [14]. Best viewed in color.

II. Related Work

Video representation in Video QA: Available methods for Video QA typically relied on recurrent networks or 3D convolutions to extract video features. Variations of LSTM were used in [19] with a bidirectional LSTM, [32] in the form of a two-staged LSTM. Likewise, [7], [21] 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 [14], [21].

Different from 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. CRN supports multiple levels of granularity in the temporal scale. This development is necessary to address the nondeterministic queries in Video QA tasks.

Neural reasoning for Video QA: The work in [19], [32] both utilized memory network as a platform to retrieve the information in the video features related to the question embedding. More recent Video QA methods started interleaving simple reasoning mechanisms into the pattern matching network operations. In [14], Jang *et al.* calculated the attention weights on the video LSTM features queried by the question. In an effort toward deeper reasoning, Gao *et al.* [7] proposed to parse the two-stream video features through a dynamic co-memory module which iteratively refines the episodic memory. Lately [21] used self-attention mechanisms 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. For complex structured videos with multimodal features, recent method leveraged memories [4], [19], [23] 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 [22], multi-head attention [18] or multi-step progressive attention memory [17] to jointly reason on video/audio/text concurrent signals.

The current trend set by these works pushes the sophistication of the reasoning processes on finding the correlation between data pattern and the query. Pattern recognition and reasoning are tightly entangled, and reasoning tends to be specific to visual/textual patterns as a result. Compared to the previous works, our framework steps toward disentangling these two processes.

Dual process systems: Reasoning systems that exhibit behaviors consistent with dual process theories are typically neural-symbolic hybrids (e.g., see [8] for an overview). In [30], visual pattern recognition modules form elements of a symbolic program whose execution would find answers for image question answering. Different from [30], we rely on implicit reasoning capability in a fully differentiable neural system [2], [13].

III. 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.

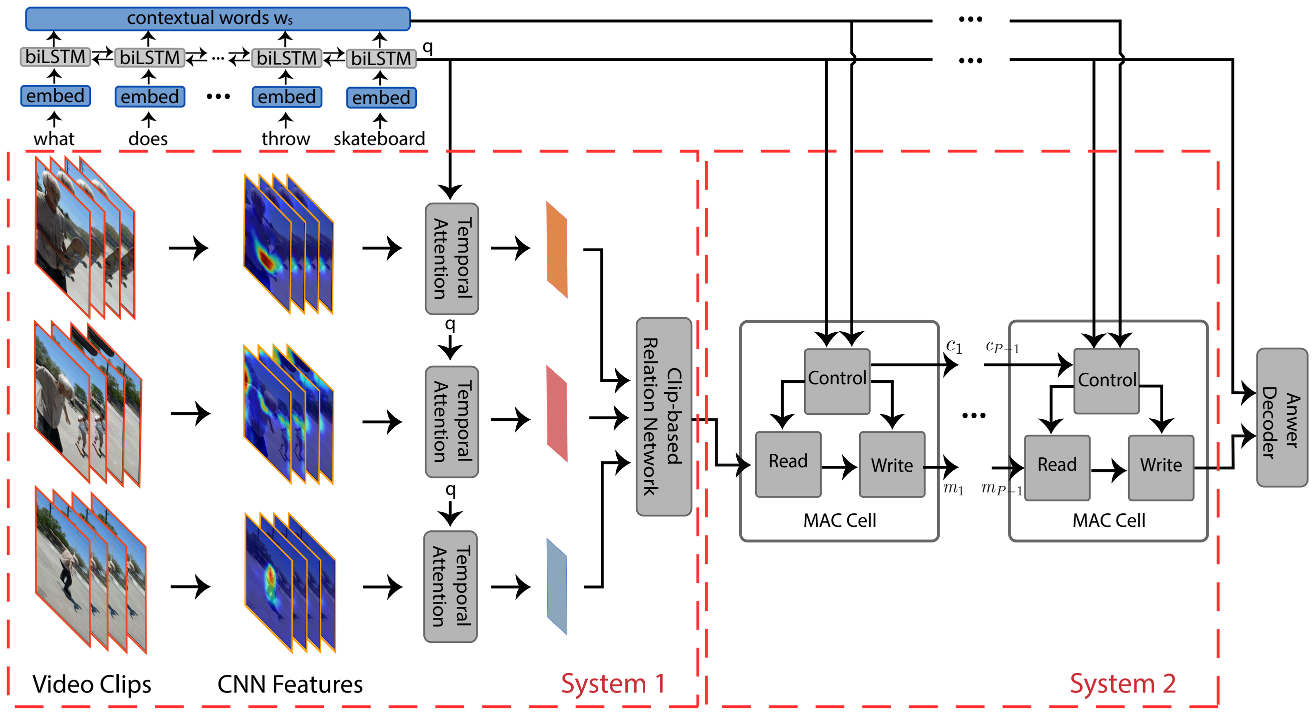

Figure 2. Overview of Network Architecture for Video QA. The model is viewed as a dual process system of hierarchical video representation with Clip-based Relation Network (CRN) and high-level multi-step reasoning with MAC cells, in which the two processes are guided by textual cues. Inputs of CRN are the aggregated features of equal-size clips obtained by a temporal attention mechanism. The high-level reasoning module iteratively co-attends to the contextual words of a given question and the visual concepts/relations prepared by the CRN unit to extract relevant visual information to the answer. At the end of the network, an answer decoder taking as input the joint representation of the question feature and the retrieved visual information is used for prediction.

A. *Dual Process System View*

Our architecture is partly inspired by the *dual process theory* dictating that there are two loosely coupled 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], [16]. 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.

For language coding, we make use of the standard bi-LSTM with GloVe word embeddings. Let S be a given question's length, we subsequently obtain two sets of linguistic clues: contextual words $\left\{w_s | w_s \in \mathbb{R}^d\right\}_{s=1}^{S}$ which are the output states of the LSTM at each step, and the question vector $q = [\overleftarrow{w_1}; \overrightarrow{w_S}], q \in \mathbb{R}^{\mathrm{d}}$ which is the joint representation of the final hidden states from forward and backward LSTM passes.

We treat a video as a composition of video clips, in which each clip can be viewed as an activity. While previous studies have explored the importance of hierarchical representation of video [34], we hypothesize that it is also vital to model the relationships between clips. Inspired by [26] and a recent work [33] 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 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 [13] 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 where the outputs of one component are the inputs to another: (1) temporal relational pattern recognition, (2) multi-step reasoning with MAC cells and (3) answer decoders. We detail these components in what follows.

B. *Temporal Relational Pattern Recognition*

Given a video of continuous frames, we begin with dividing the video into L equal-length clips $C = (C_1, ..., C_l, ..., C_L)$.

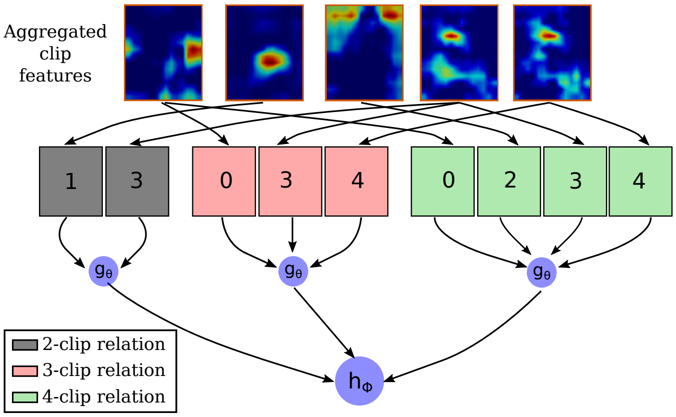

Figure 3. Illustration of Clip-based Relation Network (CRN). Aggregated features of equal size clips are fed into k-clip relation modules. Inputs to relation modules are selected on a random basis whilst keeping their temporal order unchanged. In this figure, our CRN represents a video as aggregated 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.

Each clip C_l of T frames is represented by $C_l = \{V_{l,t} \mid V_{l,t} \in \mathbb{R}^{W \times H \times D}\}_{t=1}^{T}$, where $V_{l,t}$ is frame features extracted by ResNet-101 [10] of the t-th frame in clip C_l; W, H, D are dimensions of the extracted features. Frame-level features are subsequently projected to a d dimensional space via linear transformations, resulting $C_l = \{V'_{l,t} \mid V'_{l,t} \in \mathbb{R}^{W \times H \times d}\}_{t=1}^{T}$.

As consecutive frames are redundant or irrelevant to the question, it is crucial to selectively attend to frames. In addition, this would greatly reduce the computational cost. We thus utilize a temporal attention mechanism conditioned on the question vector q to compute the aggregated clip feature $\hat{C}_l$ of the corresponding clip C_l:

$$V_{l,t}^{\text{pool}} = \frac{1}{W \times H} \sum_{w=1}^{W} \sum_{h=1}^{H} V'_{l,t,w,h}; V_{l,t}^{\text{pool}} \in \mathbb{R}^{d}, \tag{1}$$

$$s_{l,t} = W\left((W^q q + b^q) \odot \left(W^v V_{l,t}^{\text{pool}} + b^v\right)\right), \tag{2}$$

$$\hat{C}_l = \sum_{t=1}^{T} V'_{l,t} \cdot \text{softmax}(s_{l,t}), \tag{3}$$

where, W, W^q, W^v, b^q and b^v are learnable weights, and $\odot$ is element-wise multiplication.

To account for relations between clips, we borrow the strategy of TRN described in [33] which adapts and generalizes the proposal in [26] to temporal domain. Different from [33], our relational network operates at the clip level instead of frame level. More specifically, the k-order relational representation of video is given as

$$R^{(k)}(C) = h_\Phi\left(\sum_{l_1 < l_2 \ldots < l_k} g_\theta\left(\hat{C}_{l_1}, \hat{C}_{l_2}, \ldots, \hat{C}_{l_k}\right)\right), \tag{4}$$

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

a) 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 scenario, 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 a video can be guaranteed with the CRN.

C. Multi-step Centralized Reasoning

Higher-order reasoning on the rich relational temporal patterns is the key for reliably answering questions. Our approach is to disentangle 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 [13] 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 reads/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:

$$B = \sum_{k=2}^{K} R^{(k)}(C), \tag{5}$$

where, $R^{(k)}(C)$ are the k-order CRN representations calculated as in Eq. (4).

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

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

where, W_i^q and b_i^q are network weights.

Let $[;]$ denote the concatenation operator of two tensors. Based on the pair of clues contextual words and step-aware question vector $(\{w_s\}_{s=1}^{S}, q_i)$, recall that S is a given question's length, and the control state of the previous reasoning step

c_{i-1}, the control unit calculates a soft self-attention weight $\alpha_{i,s}^{\text{control}}$ over words in the question:

$$F_i = W_i^{c1}[W_i^{c0}c_{i-1}; q_i], \tag{7}$$

$$\alpha_{i,s}^{\text{control}} = \text{softmax}(W_i^{\alpha}(F_i \odot w_s) + b^{\alpha}), \tag{8}$$

and infers the control state at this step:

$$c_i = \sum_{s=1}^{S} \alpha_{i,s}^{\text{control}} w_s. \tag{9}$$

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

$$r_i = \sum_{x,y} \alpha_{i,x,y}^{\text{read}} B_{x,y}, \tag{10}$$

where,

$$I_{i,x,y} = [m_{i-1} \odot B_{x,y}; B_{x,y}], \tag{11}$$

$$I'_{i,x,y} = W_i^I I_{i,x,y}, \tag{12}$$

$$\alpha_{i,x,y}^{\text{read}} = \text{softmax}(W_i^{\alpha}(c_i \odot I'_{i,x,y}) + b^{\alpha}). \tag{13}$$

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 experiments, the function f is simply a linear transformation on top of a vector concatenation operator.

At the end of the process of P steps (P MAC cells), the final memory state m_P emerges as the output of the reasoning module used by the answer decoders in the next stage.

D. Answer Decoders

Following [14], [27], we adopt different answer decoders depending on the tasks. These include open-ended words, open-ended number, and multi-choice question.

For open-ended words (e.g. those in Frame QA in TGIF-QA dataset and all QA pairs in SVQA dataset – see Sec. IV-A), we treat them as multi-label classification problems of $\mathcal{V}$ labels defined in an answer space $\mathcal{A}$. Hence, we employ a classifier which composes 2-fully connected layers, and the following softmax, and takes as of input the combination of the memory state m_P and the question representation q:

$$p = \text{softmax}(W^{o2}(W^{o1}[m_p; W^q q + b^q] + b^{o1}) + b^{o2}), \tag{14}$$

where, $p \in \mathbb{R}^{\mathcal{V}}$ is a confidence vector of probabilities of labels. 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 representation of m_P and q. We further pass the regression output through a rounding function for prediction:

$$s = \lfloor W^{o2}(W^{o1}[m_p; W^q q + b^q] + b^{o1}) + b^{o2} \rfloor. \tag{15}$$

Mean Squared Error (MSE) is used as the loss function during the training process in this case.

Regarding the multi-choice question type, which includes repeating action and state transition in TGIF-QA dataset in later experiments, we treat each answer candidate of a short sentence in the same way as with questions. In particular, we reuse one MAC network for both questions and answer candidates in which network parameters are shared. As a result, there are two types of memory output, one derived by question $m_{q,P}$, and the other one conditions on answer candidates $m_{a,P}$. Inputs to a classifier are from four sources, including $m_{q,P}$, $m_{a,P}$, question representation q and answer candidates a:

$$y = [m_{q,P}; m_{a,P}; W^q q + b^q; W^a a + b^a], \tag{16}$$

$$y' = \sigma(W^y y + b^y); \sigma = \text{ELU}(.), \tag{17}$$

Finally, a linear regression is used to output an answer index:

$$s = W^{y'} y' + b^{y'}. \tag{18}$$

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.

IV. Experiments

A. Datasets

We evaluate our proposed method on two recent public datasets: Synthetic Video Question Answering (SVQA) [27] and TGIF-QA [14].

a) SVQA: This dataset is a benchmark for multi-step reasoning. Resembling the CLEVR dataset [15] 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.

b) TGIF-QA: This is currently the largest dataset for Video QA, containing 165K QA pairs collected from 72K animated GIFs. This dataset covers four sub-tasks 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 require strong spatio-temporal reasoning abilities. *Repetition Count*: This is one of the most challenging tasks in Video QA where machines are asked 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

Table I
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 II
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**

choice task asking machines to perceive the transition between two states/events. 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?" aim 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 found in one of the frames in a video. None of temporal relations is necessary to answer questions.

B. 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 [10] pretrained on ImageNet 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 pre-trained GloVe embeddings [24]. Unless otherwise stated, we use $P = 12$ MAC cells for multi-step reasoning in our network, similar to what described in [13], while all hidden state sizes are set to 512 for both CRN and MAC cells.

Our network is trained using Adam, 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 with proportions of 70-10-20% for training, cross-validation, and testing set, accordingly. As for TGIF-QA dataset, we take 10% of training videos in each sub-task as the validation sets. Reported results are at the epochs giving best of performance on the validation sets.

a) Evaluation Metrics: For the TGIF-QA, to be consistent with prior works [14], [7], [21], we use accuracy as the evaluation metric for all tasks except the repetition count task whose evaluation metric is Mean Square Error (MSE). For the SVQA, we report accuracy for all sub-tasks, which are considered as multi-label classification problems.

C. Results

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. The ablation results are presented in Table I, II 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. From Table I, 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 taken 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 MAC. As the SVQA dataset contains questions with compositional sequences, it greatly benefits from performing reasoning process in a multi-step manner.

Avg.Pool+MAC: A baseline to assess the significance of temporal information in the simplest form of average temporal pooling comparing to ones using a single frame. We follow [33] 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 comparing to the previous baselines on both datasets. Due to the linguistic bias, the contribution of visual information to the overall performance on the TGIF-QA is much modest than that on the 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 [33]. The model mitigates the limit of feature engineering process for video representation of a single frame as well as simply temporal average pooling. Apparently, using a single frame loses crucial temporal information of the video and is likely to fail when strong temporal reasoning capability plays a crucial role, particularly in state transition and counting. We use visual features processed in the Avg.Pool+MAC experiment to feed into a TRN module for fair comparisons. TRN improves by more than 12% of overall performance from the one using a single video frame on the SVQA, while the increase for state transition task of the 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 the SVQA comparing to the experiment Avg.Pool+MAC, the improvement on the TGIF-QA is minimal.

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 the outputs of CRN for the knowledge base of MAC. We witness significant improvements on all sub-tasks in the SVQA over the simplistic TRN whilst results on the TGIF-QA dataset are less noticeable. The results reveal the strong CRN's capability as well as a better suit of video representation for reasoning over the TRN, especially in case of compositional reasoning. The results also prove our earlier argument that sparsely sampled frames from the video are insufficient to embrace fast-pace actions/events such as repeating action and repetition count.

2) Benchmarking against SOTAs: We also compare our proposed model with other state-of-the-art methods on both datasets, as shown in Table III (SVQA) and Table IV (TGIF-QA). 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 I and Table III reveal that the contributions of visual information to the overall performance of the best known results are very little. This means their system is largely suffered from the linguistic bias of the dataset for the decision making process. In contrast, our proposed methods do not seem to suffer from this problem. We establish new qualitatively different SOTAs on all sub-tasks and a massive jump from 44.9% accuracy to 75.8% accuracy overall.

For the TGIF-QA dataset, Jang *et al.* [14] extended winner models of the VQA 2016 challenge to evaluate on Video QA task, namely VIS+LSTM [25] and VQA-MCB [6]. Early fusion and late fusion are applied to both two approaches. We also list here some other methods provided by [14] including those proposed in [6] and [31]. Interestingly, none of the previous works reported ablation studies of utilizing only textual cues as the input 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 the improper integrating of visual information caused confusion to the systems giving such low performance. In Table IV, SP indicates spatial attention, ST presents temporal attention while "R", "C" and "F" indicate ResNet, C3D and optical-flow features, respectively. Later, Gao *et al.* [7] greatly advanced the performance on this dataset with a co-memory mechanism on two video feature streams. Li *et al.* [21] recently achieved respected performance on TGIF-QA with only ResNet features by using a novel self-attention mechanism. Our method, which is also relied on ResNet features only, could achieve new state-of-the-art performance on the state transition task and the frame QA task with a big gap comparing to prior works on the frame QA task. It appears that methods using both appearance features (RestNet features) and motion features (C3D or optical-flow features) perform bad on the frame QA task, suggesting the need for an adaptive feature selection mechanism. For action and counting tasks, although we have not outperformed [7], [4], it is not directly comparable since they utilized motion in addition to appearance features. Our method, on the other hand, models the temporal relationships without explicitly using motion features and thus the action boundaries are not clearly detected. We hypothesize that counting task needs a specific network, as evident in recent work [20], [28].

3) Qualitative Results: Fig. 1 shows example frames and associated question types in the TGIF-QA and SVQA datasets. The figure also presents corresponding responses by our proposed method, and those by ST-TP [14] (on the TGIF-QA) and TRN+MAC (our own special case of flat video representation, on the 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 the SVQA were designed for multi-step reasoning, and the dual process system of CRN+MAC Net proves to be effective in these cases.

V. Discussion

We have proposed a new differentiable 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 cognition 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 modular, and thus 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.

The proposed layered neural architecture is in line with proposals in [5], [9], where reactive perception (System 1) precedes and is accessible to deliberative reasoning (System 2). Better perception capabilities will definitely make it easier for visual reasoning. For example, action counting might benefit from accurate explicit region proposals for objects and duration proposals for action, rather than the implicit detection as currently implemented. We also 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). This suggests that it is worthy to spend effort to advance reasoning functionalities for both general cases and in spatio-temporal settings. Finally, although we have presented a seamless feedforward integration of System 1 and System 2, it is still open on how the two systems interact.

Table III
COMPARISON WITH THE SOTA METHODS 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) [27]	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) [27]	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 [27]	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 IV
COMPARISON WITH THE SOTA METHODS ON TGIF-QA. FOR COUNT, THE LOWER THE BETTER. R: RESNET, C: C3D FEATURES, F: FLOW FEATURES.

Model	Action	Trans.	Frame	Count
VIS+LSTM (aggr)[25]	46.8	56.9	34.6	5.09
VIS+LSTM (avg)[25]	48.8	34.8	35.0	4.80
VQA-MCB (aggr)[6]	58.9	24.3	25.7	5.17
VQA-MCB (avg)[6]	29.1	33.0	15.5	5.54
Yu et al.[31]	56.1	64.0	39.6	5.13
ST(R+C)[14]	60.1	65.7	48.2	4.38
ST-SP(R+C)[14]	57.3	63.7	45.5	4.28
ST-SP-TP(R+C)[14]	57.0	59.6	47.8	4.56
ST-TP(R+C)[14]	60.8	67.1	49.3	4.40
ST-TP(R+F)[14]	62.9	69.4	49.5	4.32
Co-memory(R+F)[7]	68.2	74.3	51.5	4.10
PSAC(R)[21]	70.4	76.9	55.7	4.27
HME(R+C)[4]	**73.9**	77.8	53.8	**4.02**
CRN+MAC(R)	71.3	**78.7**	**59.2**	4.23

REFERENCES

[1] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *CVPR*, pages 6077–6086, 2018.

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

[3] Jonathan St BT Evans. Dual-processing accounts of reasoning, judgment, and social cognition. *Annu. Rev. Psychol.*, 59:255–278, 2008.

[4] 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.

[5] Jerry A Fodor. *The modularity of mind*. MIT press, 1983.

[6] Akira Fukui, Dong Huk Park, Daylen Yang, Anna Rohrbach, Trevor Darrell, and Marcus Rohrbach. Multimodal compact bilinear pooling for visual question answering and visual grounding. *EMNLP*, 2016.

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

[8] Artur d'Avila Garcez, Marco Gori, Luis C Lamb, Luciano Serafini, Michael Spranger, and Son N Tran. Neural-symbolic computing: An effective methodology for principled integration of machine learning and reasoning. *arXiv preprint arXiv:1905.06088*, 2019.

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

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CVPR*, 2016.

[11] Ronghang Hu, Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Kate Saenko. Learning to reason: End-to-end module networks for visual question answering. In *ICCV*, pages 804–813. IEEE, 2017.

[12] Drew Hudson and Christopher D Manning. Learning by abstraction: The neural state machine. In *NeurIPS*, pages 5901–5914, 2019.

[14] Yunseok Jang, Yale Song, Youngjae Yu, Youngjin Kim, and Gunhee Kim. Tgif-qa: Toward spatio-temporal reasoning in visual question answering. In *CVPR*, pages 2758–2766, 2017.

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

[15] Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *CVPR*, pages 2901–2910, 2017.

[16] Daniel Kahneman. *Thinking, fast and slow*. Farrar, Straus and Giroux New York, 2011.

[17] Junyeong Kim, Minuk Ma, Kyungsu Kim, Sungjin Kim, and Chang D Yoo. Progressive attention memory network for movie story question answering. In *CVPR*, pages 8337–8346, 2019.

[18] Kyung-Min Kim, Seong-Ho Choi, Jin-Hwa Kim, and Byoung-Tak Zhang. Multimodal dual attention memory for video story question answering. In *ECCV*, pages 673–688, 2018.

[19] 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.

[20] Ofir Levy and Lior Wolf. Live repetition counting. In *ICCV*, pages 3020–3028, 2015.

[21] 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.

[22] Junwei Liang, Lu Jiang, Liangliang Cao, Li-Jia Li, and Alexander G Hauptmann. Focal visual-text attention for visual question answering. In *CVPR*, pages 6135–6143, 2018.

[23] 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.

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

[25] Mengye Ren, Ryan Kiros, and Richard Zemel. Exploring models and data for image question answering. In *Advances in neural information processing systems*, pages 2953–2961, 2015.

[26] Adam Santoro, David Raposo, David G Barrett, Mateusz Malinowski, Razvan Pascanu, Peter Battaglia, and Tim Lillicrap. A simple neural network module for relational reasoning. In *NIPS*, pages 4974–4983, 2017.

[27] 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.

[28] Alexander Trott, Caiming Xiong, and Richard Socher. Interpretable counting for visual question answering. *ICLR*, 2018.

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

[30] Kexin Yi, Jiajun Wu, Chuang Gan, Antonio Torralba, Pushmeet Kohli, and Josh Tenenbaum. Neural-symbolic VQA: Disentangling reasoning from vision and language understanding. In *NeurIPS*, pages 1039–1050, 2018.

[31] Youngjae Yu, Hyungjin Ko, Jongwook Choi, and Gunhee Kim. End-to-end concept word detection for video captioning, retrieval, and question answering. In *CVPR*, pages 3165–3173, 2017.

[32] 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 *AAAI*, 2017.

[33] Bolei Zhou, Alex Andonian, Aude Oliva, and Antonio Torralba. Temporal relational reasoning in videos. In *ECCV*, pages 803–818, 2018.

[34] Jiagang Zhu, Zheng Zhu, and Wei Zou. End-to-end video-level representation learning for action recognition. In *ICPR*. IEEE, 2018.