### Spatio-temporal Relational Reasoning for Video Question Answering

by

Gursimran Singh

### A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

#### **Master of Science**

in

# THE FACULTY OF GRADUATE AND POSTDOCTORAL STUDIES

(Computer Science)

The University of British Columbia (Vancouver)

(vaneouver)

October 2019

© Gursimran Singh, 2019

The following individuals certify that they have read, and recommend to the Faculty of Graduate and Postdoctoral Studies for acceptance, the thesis entitled:

Spatio-temporal Relational Reasoning for Video Question Answering

submitted by	Gursimran Singh	in partial fulfillment of the requirements for
the degree of	Master of Science	
in	Computer Science	

#### **Examining Committee:**

James J. Little, Computer Science Supervisor

Helge Rhodin, Computer Science Second Reader

## Abstract

Video question answering is the task of automatically answering questions about videos. Apart from direct practical interest, it provides a good way to benchmark our progress on various tasks in video understanding. A successful algorithm must ground objects of interest and model relationships among them in both the spatial and temporal domains jointly. We show that the existing state-of-the-art approaches, which are based on Convolutional Neural Networks or Recurrent Neural Networks, are not effective at joint reasoning in both spatial and temporal domains. Moreover, they are short-sighted and struggle with long-range dependencies in videos. To address these challenges, we present a novel spatio-temporal reasoning neural module that models complex multi-entity relationships in space and longterm dependencies in time. Our model captures both time-changing object interactions and action dynamics of individual objects in an effective way. We evaluate our module on two benchmark datasets which require spatio-temporal reasoning: TGIF-QA and SVQA. We achieve state-of-the-art performance on both datasets. More significantly, we achieve substantial improvements in some of the most challenging question types, like counting, which demonstrate the effectiveness of our proposed spatio-temporal relational module.

## Lay Summary

Video question answering is an important task in computer vision, which aims to automatically answer questions about videos. Apart from direct practical interest, it provides a good way to benchmark our progress towards different tasks in video understanding. A major challenge for this task is that many challenging queries require long-range reasoning in both spatial and temporal domains. Existing techniques struggle to model long-range sequences and do not incorporate effective prior knowledge for joint spatio-temporal reasoning. In this thesis, we propose an end-to-end approach that uses relational networks to jointly perform long-range spatial and temporal reasoning in videos: spatio-temporal relational reasoning. In doing so, we capture the time-evolving object-interactions between multiple objects and the time-evolving action-dynamics of individual objects. We show the effectiveness of our approach on two benchmark datasets and achieve state-of-theart performance.

## Preface

The entire work presented here is original work done by the author, Gursimran Singh, performed under the supervision of Prof. James J. Little. The author would like to thank Prof. Leonid Sigal for invaluable technical advice in the successful completion of this work. A version of this work, authored by Gursimran Singh, James J. Little, and Leonid Sigal has been an oral presentation and published as:

• G. Singh, L.Sigal, and J.J.Little. *Spatio-temporal Relational Reasoning for Video Question Answering*. In British Machine Vision Conference (BMVC), September 2019

## **Table of Contents**

Ał	ostrac	ti	ii
La	ıy Suı	<b>nmary</b>	v
Pr	eface		v
Ta	ble of	f Contents	vi
Li	st of [	Fables	ii
Li	st of l	Figures	X
Ac	know	ledgments	ĸi
1	Intr	oduction	1
	1.1	Problem Definition	4
		1.1.1 Scope	6
		1.1.2 Data	7
	1.2	Method Outline	8
	1.3	Thesis Organization	0
2	Rela	ited Work	1
	2.1	Visual Question Answering	1
		2.1.1 Image Question Answering	1
		2.1.2 Video Question Answering	3
	2.2	Video Understanding	5

	2.3	Relational Reasoning	17
3	Арр	roach	19
	3.1	Text and Visual Representation	20
	3.2	Spatio-temporal Relational Network	21
	3.3	Answer Decoder	24
	3.4	Training	25
4	Exp	eriments	27
	4.1	Datasets	27
		4.1.1 TGIF-QA	27
		4.1.2 SVQA	29
	4.2	Implementation Details	30
		4.2.1 Setup	30
		4.2.2 Metrics	31
	4.3	Comparison with State-of-the-art Methods	32
		4.3.1 TGIF-QA Dataset	32
		4.3.2 SVQA Dataset	33
	4.4	Ablations	35
	4.5	Qualitative Results	36
5	Con	clusion	39
Bi	bliogi	raphy	42

## **List of Tables**

Table 4.1	Statistics of the TGIF-QA dataset	28
Table 4.2	Statistics of the SVQA dataset.	30
Table 4.3	Comparison with state-of-the-art on TGIF-QA dataset. ↑ means	
	higher numbers correspond to better performance (ACC) and $\downarrow$	
	means lower numbers correspond to better performance (MSE).	33
Table 4.4	Comparison with the state-of-the-art on different categories of	
	the SVQA dataset. We split the table into two parts (Top - Exist,	
	Count, Integer and Attribute Comparison; Bottom - Query and	
	aggregate of All) for better readability. Higher numbers corre-	
	spond to better performance.	34
Table 4.5	Effectiveness of joint spatio-temporal reasoning as opposed to	
	individual spatial or temporal relational reasoning on TGIF-QA	
	dataset	35

## **List of Figures**

Figure 1.1	The figure shows a question and selected frames of a video	
	in SVQA dataset [41]. We show individual spatial-relations	
	among relevant objects, which are far, close and closer.	
	The change of spatial-relations over time corresponds to the	
	temporal-relation which is getting close. Having ob-	
	served these temporal-relations among object-pairs, the algo-	
	rithm can infer the answer – Sphere	3
Figure 1.2	Figure shows different question-types used in this work (illus-	
	trations taken from TGIF-QA dataset). Note that all questions	
	in the SVQA dataset are open-ended word type	5
Figure 1.3	The figure shows the block diagram of our proposed model.	
	First, we extract features using pre-trained CNNs (shown in	
	green). The video features are passed to Spatial Relation Mod-	
	ule (SRM) and Global Context Encoder LSTM (GCE) which	
	models object interactions and action dynamics. Finally, the	
	output of SRM and GCE are passed to Temporal Relation Mod-	
	ule (TRM) to model relationships in time. Finally, these are	
	passed to Answer Encoder to obtain the answer. Separately,	
	we encode the question and answer-options using the Ques-	
	tion/ Answer Encoder LSTM which is passed to SRM, GCE	
	and TRM modules to condition on the question	9

Figure 3.1	Spatio-temporal Relational Network architecture. The Spa-	
	tial Relations Module (top) models arbitrary spatial-relations	
	among all possible groups of objects for each frame individu-	
	ally. The Global Context Encoder LSTM (bottom left) models	
	the action-dynamics with global context at time $t$ . The con-	
	catenated output of these modules is then fed to the Tempo-	
	ral Relations Module (bottom right) which computes temporal	
	relations among a temporally-ordered group of frames. No-	
	tice that, for simplicity, we have shown object-groups as pairs,	
	however, in general they can be more than two	21
Figure 4.1	[Best viewed in color] A comparison of the qualitative results	
	of ST-TP [13] and STRN (Ours). Green and Red refers to cor-	
	rect and incorrect predictions, respectively.	36
Figure 4.2	[Best viewed in color] Qualitative results for different cate-	
	gories of the SVQA dataset where our approach predicted the	
	correct answers.	37
Figure 4.3	[Best viewed in color] Qualitative results for different cate-	
	gories of the SVQA and TGIF-QA datasets where our approach	
	predicted the <b>incorrect</b> answers	38

## Acknowledgments

I would like to express my heartfelt gratitude to my supervisor, James Little. It was his vital advice, support, and encouragement which motivated me to take up and successfully conclude this research. I think his critical sense of dissecting problems had a lasting impact on me being a successful research student.

I would also like to thank Prof. Leonid Sigal who is a co-author of a version of this work presented at British Machine Vision Conference, 2019. His invaluable technical advice was crucial in conducting and designing experiments needed for this work. Also, I would like to thank Prof. Helge Rhodin for agreeing to be the second reader of my thesis.

Apart from that, I'm very grateful to Prof. Leonid Sigal, Prof. Frank Wood, and Prof. Mark Schmidt for teaching amazing courses at UBC which helped me build a solid foundation in machine learning. Also, thanks to my course project collaborators, Mohit, Siddhesh, Setareh, and Saeid for invaluable discussions, ideas and help in coding know-how.

I would also like to thank my fellow members at the Computer Vision group, especially Rayat, Soheil, Tanzila, and Polina for interesting discussions, talks and being so nice, kind, and helpful. Also, thanks to the Department of Computer Science at the University of British Columbia for accepting me as a student in this prestigious institution.

Lastly, I would always be grateful and indebted to my family for their unconditional love and support. My parents, most of all, have the greatest contribution in shaping me for success in all my endeavors. Special thanks to my girlfriend, Amrit, who helped me sail smoothly with her love and support.

### **Chapter 1**

## Introduction

Perceiving and understanding video data is vital to human intelligence. It is crucial to every day-to-day activity ranging from low-level tasks like eating to high-level tasks like driving. Similarly, the machines which aim to emulate these activities need to perceive and make sense of time-evolving visual data. It must single out objects of interest and understand their relationships at various levels of abstraction. Moreover, it must do so in a hierarchical manner building relationships of relationships in both spatial and temporal domains. Towards this goal, the task of video understanding has been proposed and is an active area of research in computer vision literature.

In contrast to individual images, which require reasoning in the spatial domain, videos require joint reasoning over both spatial and temporal domains. Due to the richness of the problem, it is being studied as different sub-tasks, each differing in terms of simplifying assumptions used to tame the problem. For instance, Activity Recognition [10, 43] aims to classify the short and trimmed video among a vocabulary of activities. Temporal Action Localisation [40, 47] aims to temporally localize activities in a possibly longer video. More recently, Video Question Answering [8, 13, 17, 51] aims to answer various types of natural-language questions about a video.

Video Question Answering (aka VideoQA) is arguably the most challenging among video tasks since it may contain a myriad of queries, including those encompassing other video understanding tasks. Each of the many possible naturallanguage questions, encapsulated by a smaller set of queries, can be seen as belonging to one of the sub-tasks in video understanding. Hence, it is arguably a more general and challenging among other specific video understanding tasks. For instance, simpler questions, similar to image question answering (ImageQA), involve attribute identification in a single frame of a video [32, 34]. More complex questions, similar to activity recognition and localization, require looking at multiple frames in a local temporal region [13, 34]. The most complex questions require recognizing activities across time, counting them or reasoning about their temporal order [13, 41].

A generic VideoQA algorithm must learn to ground objects of interest in video and reason about their interactions in both the spatial and temporal domains. Conceptually, a VideoQA algorithm can be broken into three different sub-tasks. First, the algorithm should understand the intent of the query from the natural language description of the question. Second, it should compute relationships which are relevant to the query in the spatial domain. Finally, it should reason how these spatial relations evolve in the temporal domain. For instance, consider the question and the sequence of frames in Figure 1.1. To answer, the algorithm considers spatial relations among all possible object-pairs in each frame individually. In the case of the blue cylinder and sphere, these are calculated as far in the first frame, close and very close in the subsequent frames. Having done that, it needs to reason how these spatial relationships change in the temporal domain. This leads to identifying the correct interpretation that cylinder and sphere are getting close. Having observed these spatio-temporal relationships among all possible object-pairs, the algorithm can figure out the correct answer which is a sphere in this case.

Traditional approaches use 3D Convolutional Neural Network (CNN) [30, 47], Long Short Term Memory Unit (LSTM) [26, 51], or attention [8, 13, 48] to model such relationships. Although successful, they are limited in capacity. For instance, 3D CNNs are useful for identifying local spatio-temporal action events, demonstrated by success in Activity Recognition datasets, however, they struggle in modeling long-range temporal relationships [46, 54]. Similarly, LSTMbased approaches, although known to do well in long-range text sequences, struggle to model videos [8, 23]. This is because, unlike text, videos contain longer and information-richer sequences of spatial data, which LSTMs, or their spatial-



Figure 1.1: The figure shows a question and selected frames of a video in SVQA dataset [41]. We show individual spatial-relations among relevant objects, which are far, close and closer. The change of spatial-relations over time corresponds to the temporal-relation which is getting close. Having observed these temporal-relations among object-pairs, the algorithm can infer the answer – Sphere.

attention variants, cannot model naturally and effectively. More importantly, these network architectures do not provide an effective way to model videos and need to learn relational reasoning from scratch which is inefficient and data-hungry [38].

In this work, we leverage and extend relational networks [31] to model spatio-

temporal relationships in videos. Relational networks are designed and provide effective prior knowledge to effectively model relationships among its inputs. The ability to reason is embedded right in the formulation of relational networks, just like the ability to reason about sequences is encoded in LSTMs. Previously, relational reasoning has been used effectively in image question answering [38] and activity recognition [54]. However, it was limited to either the spatial or the temporal domain individually. Inspired by these two works, we present spatio-temporal relational networks which can perform joint relational reasoning in both spatial and temporal domains.

**Contributions:** We make several contributions in this work. First, we present a novel general-purpose neural network module which acts as an effective prior for spatio-temporal relational reasoning in videos. This allows us to capture objectinteractions and how they change over time. Second, we present Global Context Encoder LSTM to model action-dynamics of individual objects with global context at time t. Hence, our approach captures both spatio-temporal relations (capturing object-interactions) and action-dynamics (capturing how individual objects change over time) in videos. To our knowledge, this is the first attempt to perform joint spatio-temporal reasoning using relational networks. Third, we show the effectiveness of our proposed Spatio-Temporal Relational Network on a variety of VideoQA tasks, which include both real-world (TGIF-QA [13]) and synthetic (SVQA [41]) datasets. Our approach achieves state-of-the-art results on both these datasets. Finally, we show substantial improvement in the challenging counting task that requires capturing spatio-temporal dynamics in different parts of a video. Also, to the best of our knowledge, this is the first attempt to approach VideoQA using relational networks.

#### **1.1 Problem Definition**

We are addressing the problem of video question answering (VQA) in this work. More formally, the input consists of a video, consisting of a sequence of frames; a question, consisting of a sequence of words; and optionally five different answer options. The task is to answer the question based on the information present in the video. The exact type of output depends on the specific task in the video question

#### (a) Multiple-Choice



Q) What does the woman do after

1. Open lid with spill sauce

hold cat?

Options:

2. Shake cat 3 In mouth 4. On the around 5. Show things in hand Answer: Shake Cat



lighting up? Answer: Cigarette (c) Open-ended number



Q) What is the man with his guitar Q) What is the color of the hair? Answer: Brown

Figure 1.2: Figure shows different question-types used in this work (illustrations taken from TGIF-QA dataset). Note that all questions in the SVQA dataset are open-ended word type.

answering dataset. Here are the possible options:

- 1. Multiple Choice: The output consists of picking the correct option among a set of five possibilities.
- 2. Open-ended word: The output consists of picking the correct word in the entire vocabulary.
- 3. Open-ended number: The output consists of predicting a number like the count of events in a video.

We model the first two tasks as a classification problem and the last task as a regression problem. Irrespective of the type of output, the model needs to understand what is being asked in the question and reason over relevant parts of the video to arrive at the answer. Here are some of the desired properties of the model:

- Robust to various ways (natural language descriptions) of describing the same query in a video.
- Able to deal with different length of videos and identify objects of interest to reason over them.

- It should be able to identify the correct answer irrespective of the type of question-answering task.
- Able to deal with different viewpoints of camera and zoom levels in videos.

The task of video question answering is inherently difficult due to several reasons. First, video question answering may contain questions about many different sub-tasks like classification, counting, or even grounding. To succeed, the same VQA algorithm must do well on all these tasks. Second, unlike images which require reasoning in the only spatial dimension, videos require joint reasoning in both spatial and temporal dimensions. Hence, traditional architectures like CNNs and LSTMs, which are designed to explicitly reason only in one domain, are not the best choices when used off the shelf in the case of videos. The 3D alternative of the 2D CNNs are narrow-sighted and are only able to focus on a short part of a video at a time. Similarly, LSTMs are known to struggle for long sequences, especially in case of videos which contain long information-rich frames instead of sequences of words. Third, training a neural network architecture for videos is resource-intensive in terms of memory, time and computation power. For instance, in comparison to images, typically a video dataset size is 100 times (order of terabytes) and the corresponding neural network model has 100X more parameters (and takes 3-4 days to converge). Fourth, videos can be very diverse with different lengths, event-speed, number of objects, frame-rate, resolution, camera angle, etc. Amongst this diversity, the algorithm needs to focus on the relevant parts and capture both the local and the global detail. Consider as an example a video with a human clapping. While the local information may signify hands moving closer or farther, only when the algorithm sees the larger picture (global information), it can understand that it is a clapping video. All these factors make the general problem of video question answering very challenging.

#### 1.1.1 Scope

We limit the scope of our problem by making several simplifying assumptions. First, we tame the computational complexity by uniformly sampling 35-40 frames, effectively reducing the frame rate. Second, in both of our datasets, we only have trimmed videos consisting of a few seconds. This simplifies the problem of irrelevant areas in the start and the end of a video. With these assumptions, we reduce the size of our dataset to one terabytes. Third, the set of possible answers is limited by either the number of available options or the size of the vocabulary or numbers from 0-9. Fourth, we only consider a limited number of tasks out of all possibilities in the general problem of video question answering problem. Specifically, we only consider tasks in the TFIF-QA and the SVQA dataset which are limited to counting, transition, comparison, exists, query, etc. They do not contain tasks like grounding of events or verifying the order of events. Fifth, we assume the availability of a standard pre-trained 2D and 3D CNNs (ResNet[12] and C3D[43]) to extract features vectors.

#### 1.1.2 Data

We validate the importance of joint spatio-temporal reasoning using two video question answering datasets, namely TGIF-QA [13] and SVQA [41]. These datasets have been carefully chosen as they contain complex questions requiring spatio-temporal reasoning in videos. This is in contrast to many other VideoQA datasets like LSMDC-QA [35] which are image-centric, merely querying about visual concepts like color, objects, and locations [13]. Also, we would like to clarify the rationale behind choosing Video Question Answering as opposed to other video understanding tasks like Activity Recognition (used in a closely related work [54]). While activity recognition requires spatio-temporal reasoning, it is limited to "recognizing actions" which are usually localized to a short sequence of frames. On the other hand, miscellaneous VQA tasks require more elaborate and long-term reasoning over both spatial and temporal domains. For instance, the Count task requires the network to recognize all possible actions of interest and then compute the sum to arrive at the answer. Hence, the proposed video question answering datasets are a better judge for evaluating spatio-temporal reasoning in Videos.

**TGIF-QA** [13] is a large-scale dataset containing 165K QA pairs collected from 71K real-world animated Tumblr GIFs. The questions are categorized into four separate tasks. 1) Repeating Action (Action) aims to name the event that happened a specific number of times in the video. This is a multiple-choice task where the

correct answer is one of the five available options (Fig 4.1a). 2) State Transition (Trans), similarly, is a multiple-choice task with five options. Questions ask about state transitions like facial expressions (from happy to sad), among others (Fig 4.1b). 3) FrameQA is an open-ended task which, similar to image-QA, can be answered by looking at one of the "appropriate" frames in the video. However, the range of possible answers spans the entire vocabulary (Fig 4.1c). 4) Repetition Count (Count) aims to count the number of times a given event happens in the video. This is an open-ended task and answers lie in a range of integers: 0 to 10 (Fig 4.1d).

**SVQA** [41] is synthetically generated dataset designed to control and minimize language biases in existing videoQA datasets. It contains 120K questions asked on 12K videos with moving objects like sphere, cylinder or cube (Fig 4.2). Similar to FrameQA, answers span the entire vocabulary. Questions are compositional and require a series of reasoning steps (like comparison and arithmetic) in both the spatial and the temporal domains. Questions are further categorized into exist, count, integer comparison, attribute comparison and query subtypes. Since the exact train-val-test subsets of the SVQA dataset are not readily available, we randomly sample a new split similar to Table 1 of Song [41]. In comparison to TGIF-QA, it contains more complex questions requiring more elaborate spatio-temporal reasoning. However, unlike real-world GIFs in TGIF-QA, it contains perceptually-simpler scenes consisting of a few synthetic objects. These two datasets are well suited for our task because they contain well-formed questions that require complex spatio-temporal reasoning.

#### **1.2 Method Outline**

The input to our model consists of a sequence of frames  $(\{v^t\}_{t=1}^T)$  in the form of a *video*, and a sequence of words  $(\{w^i\}_{i=1}^L)$  in the form of a *question*; where T is the length of the video and L is the length of the question. Optionally, in case of multiple-choice questions, it also consists of five separate sequences of words  $(\{c_k^j\}_{j=1}^{S_k})$  representing *answer-options*; where  $S_k$  is the corresponding sequence length and  $k \in [0, 4]$ . The block diagram of our approach is shown in Figure 1.3.



**Figure 1.3:** The figure shows the block diagram of our proposed model. First, we extract features using pre-trained CNNs (shown in green). The video features are passed to Spatial Relation Module (SRM) and Global Context Encoder LSTM (GCE) which models object interactions and action dynamics. Finally, the output of SRM and GCE are passed to Temporal Relation Module (TRM) to model relationships in time. Finally, these are passed to Answer Encoder to obtain the answer. Separately, we encode the question and answer-options using the Question/Answer Encoder LSTM which is passed to SRM, GCE and TRM modules to condition on the question.

There are five main components: (a) Question/ Answer Encoder LSTM (b) Spatial Relation Module (SRM), (c) Global Context Encoder LSTM (GCE), and (d) Temporal Relations Module (TRM) (e) Answer Decoder Module. As a preprocessing step, we extract appearance  $({A^t}_{t=1}^T \in \mathbb{R}^{7 \times 7 \times 2048})$  and motion  $({M^t}_{t=1}^T \in \mathbb{R}^{4096})$  features using pretrained ResNet-152 [12] (*res5c*) and C3D [43] (*fc6*), respectively. Further, we use a Downscale CNN to reduce the size of feature vector from  ${A^t}_{t=1}^T \in \mathbb{R}^{7 \times 7 \times 2048}$  to  ${O^t}_{t=1}^T \in \mathbb{R}^{3 \times 3 \times 256}$ .

We use the Question Encoder LSTM and the Answer Encoder LSTM to encode the question and answer options (if available) respectively. The output of the Question/ Answer Encoder LSTM is used as the question encoding  $\gamma$  to condition the output of other modules. The Spatial Relation Module takes appearance features ( $\{A^t\}_{t=1}^T$ ) as input and computes spatial relations among various objects. This can be seen as modeling object interactions in each frame individually. The Global Context Encoder LSTM takes motion features ( $\{M^t\}_{t=1}^T$ ) as input and captures action-dynamics with global context at time t. Finally, the Temporal Relation Module takes the concatenated SRM-encoding ( $f_t$ ) and the GCE-encoding ( $\rho_t$ ) as input and computes how the spatial-relations and action-dynamics change over time. This corresponds to modeling temporal changes in both the interactions among different objects and the motion-dynamics of individual objects. Apart from the video features, the SRM, TRM and GCE modules also take the question/ answer encoding ( $\gamma$ ) as input to conditions their output on the query. The output (Y) of the TRM is passed to the Answer Decoder Module which generates the answer. In the case of multiple-choice and open-ended word questions, we use classification while in the case of open-ended number questions regression is used (details in section 3.3).

#### **1.3** Thesis Organization

We have organized the thesis as follows: We present the background and related work in Chapter 2. In this chapter, we start with presenting background on existing question answering systems and their limitations. We then discuss the state-of-theart methods to solve the problem of video question answering. We also review the background material on relational networks and their applications. In Chapter 3, we describe our Spatio-Temporal Relation Network (STRN) and define each component in detail. Additionally, we discuss the necessary details for implementation, training, and inference. In Chapter 4, we present the experimental results. We present a comparison of our approach with the state-of-the-art methods on the two videoQA datasets. Also, we present an ablation study to highlight the importance of joint spatio-temporal relational reasoning. Finally, in Chapter 5, we highlight our main contributions and discuss future directions to conclude the thesis.

### **Chapter 2**

## **Related Work**

Our work on video question answering is related to the rich literature of visual question answering, video analysis, and relational networks. First, we review recent successes on image question answering and discuss the challenges in extending it to video question answering. In particular, we discuss what makes videos more challenging. Then, we review existing literature on video analysis, especially on capturing spatio-temporal relationships. Finally, we describe relational networks in detail as part of the background material.

#### 2.1 Visual Question Answering

#### 2.1.1 Image Question Answering

Image question answering [1], also known as visual question answering, is the task of answering queries about images. It has been a popular area for research for computer vision researchers over the past few years [6, 11, 24]. Typical questions focus on identifying attributes, counting objects or reasoning about their spatialrelations. Overall, question answering is a fairly complex task involving a combination of visual perception, language understanding, and deductive reasoning. Earliest methods use end-to-end neural networks to learn a mapping from input images and questions, directly to answers. In particular, a typical approach extracts image features from the last layer of the convolutional neural networks pre-trained on object recognition datasets. Questions are fed word-by-word into recurrent neural networks to obtain fixed-length feature vectors as question representation. The question and the image features are then jointly embedded, followed by a multiclass classification to arrive at the answer [7, 25, 32]. However, since the image and question are separately encoded, their representations tend to encode irrelevant information which hurts performance.

As a remedy, the visual attention mechanism [49] was proposed to allow selective focusing on parts of the image relevant to answer the question. Chen [3] proposed a configurable convolutional kernel, derived from the question semantics, which can be used to identify relevant image features. Yang [42] proposed multiple steps of attention in a stacked manner for improved reasoning. Lu [22] proposed co-attention between the question and image, which also attends to relevant words in the question in addition to relevant areas in the image. Another successful approach, Fukri [6] proposed multimodal compact bi-linear pooling (MCB), which efficiently combined the image and question features to predict attention.

Despite the fact that black-box neural network approaches can deliver impressive VQA performance, they have been argued to exploit dataset-specific biases rather than doing structured reasoning. This has been corroborated by poor generalization on the CLEVR dataset, a diagnostic dataset containing queries ranging from simple perception tasks to complex logical and spatial reasoning tasks[11]. As a remedy, various approaches were proposed which used neural networks to obtain discrete representations, allowing incorporation of symbolic-reasoning priors. For instance, Johnson et al. [15] proposed to infer a specific chain of reasoning (called program) from a given question and use it to build a custom neural network architecture, by assembling modules. Yi [50], apart from the question, also inferred the structured scene representation of the image. This allowed them to obtain the answer by executing the program directly on the scene representation using a deterministic symbolic executor, a very effective prior. However, these approaches depend on discrete representations (programs or structured scene representation) which are hard to obtain in the case of real-world data.

Alternatively, Santoro [38] proposed relational networks and achieved state-ofthe-art and superhuman performance on the CLEVR dataset. Relational networks inherently provide effective prior knowledge to learn relations among spatial elements of an image. This allows better reasoning, however, it was limited in the spatial domain. We extend this work to videos and propose a model which does joint relational reasoning in both the spatial and the temporal domain.

#### 2.1.2 Video Question Answering

Video question answering, a relatively newer problem, is the task of answering natural language questions on videos. A video consists of a sequence of frames representing complex interactions and manipulations of objects that evolve over time. Unlike image question answering, which is confined to spatial reasoning, it focuses on queries which require joint reasoning in both spatial and temporal domain. These may include identifying a single activity spanning a few contiguous frames, or more generally multiple such activities and inferring relationships between them.

Despite significant successes, existing methods in image question answering do not naturally extend to videos as they fail to capture the temporal aspect. For instance, simple baselines which average the CNN feature representation of individual frames do not give satisfactory performance. By averaging, they loses both crucial details and temporal order among different frames. Similarly, the use of attention to attend to the most relevant frame or predicting an answer based on each frame separately makes sense only for shallow queries which requires looking at only one frame to answer the question. The earliest approaches [26, 48, 51] used LSTMs to capture temporal order among frames of a video. First, they extracted feature vectors for each frame independently from the final layer of a pre-trained convolutional neural network. Then they passed these feature vectors through an LSTM to model temporal dependency, and optionally, leveraged temporal attention to selectively attend to important frames in a video. These approaches modeled temporal reasoning through a combination of LSTM and attention but lacked any spatial reasoning.

Jang [13] proposed an architecture which utilized both spatial and temporal attention. Apart from the final-layer CNN features, they also extracted convolution features maps. A spatial-attention mechanism was used to selectively attend to important regions in CNN feature maps. The attended convolutional vectors, along with the final-layer CNN feature vectors are passed into an LSTM, followed by a temporal attention mechanism. Another approach [52] proposed a hierarchical spatio-temporal attention network which jointly learns the representation of sequentially critical frame with the targeted objects. It allows multi-step reasoning for refining the joint representation of the spatio-temporal attentional video and the textual question. They employ an object generator to produce a set of candidate regions and used spatial attention to automatically localize targeted regions in each frame according to the question. The attended representations are further attended by an attentional GRU, which learns order sensitive representation of the relevant frames. Similarly, Zhou [53] employs hierarchical attention context network to model hierarchically sequential questions and multi-step reasoning. Again, they use a combination of long short-term memory (LSTM) unit and a multi-stream spatio-temporal attention network to learn a joint representation of the video and the question. Although these approaches allowed spatio-temporal reasoning, however, it is limited in terms of modeling long-range temporal dependencies. They are attributed to recurrent neural networks (RNN) and its variants like long short-term memory (LSTM) units [8, 23]. First, LSTMs are ineffective with long sequences of data, and videos are usually composed of long sequences of frames. Second, LSTMs struggled to model information-rich frames vectors, which contained more detailed information in comparison to text-data vectors.

To address this problem, Song [41] proposed a more granular spatial-attention and a modified-GRU incorporating a temporally-attended hidden state transfer. The spatial attention mechanism is designed to address crucial and multiple logical sub-tasks embedded in the questions like 'filter shape', 'query color', etc. The GRU-based temporal attention captures long-term temporal dependencies to gather all relevant visual cues. Another set of approaches used memory networks [8, 17, 27] to handle long-term dependencies. Memory networks provide explicit memory representation for each token in the sequence, which can be read and replaced. Providing memory representations aids long short-term memory (LSTM) units to recover information in long sequences. Moreover, multi-step attention allows iterative reasoning over memory representations, useful for question answering tasks. Gao [8] used a co-memory attention. First, they extract motion and appearance features by using a two-stream network. Then, they used a 1D convolution-deconvolution network to build multi-level contextual facts with the same temporal resolution but a different contextual range. Finally, they use two memory networks, each for appearance and motion. Their method incorporates a co-memory attention mechanism which takes motion cues for appearance attention generation and appearance cues for motion attention generation. Li [20] used a self-attention based technique to exploit global dependencies among words of a question and frames of a video. They used a positional self-attention and calculated attention at each position by attending to all positions within the same sequence. This allows capturing global dependencies of question and temporal information in the video. Further, they utilized co-attention to attend to both the question and the video simultaneously. An important aspect of this approach is that it avoids the use of recurrent neural networks (RNN), and hence requires less computation time and achieves better performance.

Although modified-GRU [41], co-attention [8], and self-attention [20] performed better than pure LSTM-based approaches, however, they still had to learn relational reasoning from scratch. The ability to compute arbitrary relations is essential to question answering tasks. LSTMs, memory networks, and self-attention based techniques do not model the ability to compute relations explicitly. Instead, they learn it implicitly using supervised data, which is inefficient and data-hungry. Our approach takes a different route and uses relational networks which provide effective prior knowledge for performing relational reasoning. Moreover, our approach allows joint relational reasoning in both spatial and temporal domains. This allows capturing complex spatio-temporal relationships and hence, it is dataefficient and outperforms the above techniques.

#### 2.2 Video Understanding

Video understanding is one of the fundamental problems in computer vision. In the recent years, many related sub-tasks including Activity Recognition [10, 43], Temporal Action Localisation [40, 47], Dense Video Captioning [19, 55], and Video Question Answering [8, 13, 17, 51] have been active area of research. A significant part of research deals with developing robust spatio-temporal features, which represent how spatial elements interact and change over time. Earliest attempts were

centered around hand-designed representations based on histograms and pyramids [4, 5, 18, 37, 45]. Encouraged by the success of deep learning techniques for still images, many attempts focused on extracting features for each frame of the video separately using pre-trained CNNs and then integrating temporal information using mean-pooling [9] or recurrent neural networks (RNNs) [26, 48, 51]. However, mean-pooling is prone to losing temporal order and detailed information in frames and RNNs are known to struggle with long sequences, especially when they contain information-rich frames [8, 23]. Alternately, 3D convolutional networks [2] were proposed which allowed reasoning in both spatial and temporal domains. The 4D kernel consumes a set of RGB frames in a small temporal window and learns spatio-temporal patterns over a local region in the video. 3D convolutional networks (3D CNNs) and many of its variants [30, 44] achieved strong performance on many video understanding tasks like action recognition [43], action detection [40], and video captioning [28]. Despite the success, 3D CNNs struggle to achieve good performance on challenging datasets, where long-range spatio-temporal reasoning is required. This is due to the limited size of the temporal window in convolutional kernels.

As a remedy, a recent work by Wang [46] models videos as space-time graphs where nodes represent regions of interest in the video. The nodes, which are generated by region proposal networks, are connected by either appearance similarity or spatio-temporal proximity. Finally, given the graph representation, they perform reasoning and inference on the graph using graph neural networks (GNNs) [39]. This technique facilitates modeling temporal shape dynamics and functional relationships between humans and objects. The utility of this approach was shown on two benchmark Activity Recognition datasets. This approach is similar to ours as it explicitly models spatio-temporal relationships. However, they use Graph Neural Networks (GNNs) which depend on structured data like bounding boxes extracted using Region Proposal Networks (RPNs) [33]. In comparison, our approach use relation networks (RNs) which are flexible and can work with a broad range of unstructured inputs like raw RGB values, CNN and LSTM outputs. Moreover, RNs are simpler, more exclusively focused on relational reasoning and easily integrable within broader architectures.

#### 2.3 Relational Reasoning

Relational reasoning is the ability to reason about relationships among entities. It is central to general intelligent behavior and is essential to answer complex questions in VQA tasks. Relation networks (RN), first introduced by Roposo [31], are designed by constraining the functional form of neural networks such that the ability to reason about relations is baked right in its architecture. Hence, relational reasoning is inherent to relation networks (RN) without needing to be learned from data. Santoro [38] defined the relation network as a composite function below:

$$\operatorname{RN}(O) = f_{\phi}\left(\sum_{i,j} g_{\theta}(o_i, o_j)\right)$$
(2.1)

where the input consists of a set of objects,  $O = o_1, o_2, ..., o_i, ..., o_j, ..., o_n$ ; and functions  $g_{\theta}$  and  $f_{\phi}$  are general purpose functions (e.g. neural networks) with learnable parameters  $\theta$  and  $\phi$ . The function  $g_{\theta}$  computes arbitrary relations between any two input objects  $o_i, o_j$ . The function  $f_{\phi}$  reasons over the computed relations and combines them into desired output. As stated in Santoro [38], relation networks have following strengths.

- 1. **Infer relations**: The RNs consider potential relations among all possible object pairs  $(o_i, o_j)$ , and they learn these relations by learning the parameters of the relation function  $g_{\theta}$  from data.
- 2. **Data efficiency**: The RNs learns relations for all possible object pairs  $(o_i, o_j)$  using the same relation function  $g_{\theta}$ . This ensures greater generalization and less overfitting.
- 3. **Invariant**: The summation in Eq. 2.1 makes the computed relations order invariant. Hence the output captures relations which are generally representative of the set of object pairs.

Roposo [31] and Santoro [38] showed the effectiveness of relation networks on scene description data and image question answering, respectively. They demonstrated that even the powerful CNNs or MLPs struggle to solve questions which

require relational reasoning. However, when augmented with relational networks (RNs), they achieve superhuman performance even in complex datasets like CLEVR [14]. Zhou [54] extended relational networks to the temporal domain and introduced Temporal Relational Networks (TRNs) for videos. They define the temporal relation network (TRN) as a composite function below:

$$T_2(V) = h_\phi \left( \sum_{i < j} g_\theta(f_i, f_j) \right)$$
(2.2)

where the input is the video, represented by a set of *n* frames,  $V = f_1, f_2, ..., f_n$ . The  $g_\theta$  and  $f_\phi$  functions are the usual relation and reasoning functions, respectively. However, unlike the relation network (RN), in a temporal relation network, we only consider pairs which are temporally ordered in time. This is enforced by the i < j condition in the summation operation in Eq. 2.2. TRN computes temporal relationships among video frames and achieved the state-of-the-art result in Activity Recognition datasets.

Our work is different from the above networks in two main ways. First, we show the effectiveness of relational networks in a more challenging setting: Video Question Answering, where interesting events may occur at different parts of the video. Second, and more importantly, we model joint spatio-temporal relationships, unlike [38] and [54] which work either in spatial or temporal domains individually.

### **Chapter 3**

## Approach

Video Question Answering is a very challenging task which requires joint reasoning in both spatial and temporal domains. It requires looking at multiple spatial elements in a frame, finding relationships among them, and observing how these relationships evolve over time. Moreover, these relationships may span long sequences of frames which are hard to model using an LSTM. Our proposed Spatial-Temporal Relational Network (STRN) uses relational networks in both spatial and temporal domain to model multi-entity and long-term relationships in videos.

The overall architecture of our Spatio-Temporal Relational Network (STRN) is shown in Figure 3.1. The input consists of a video, a question and optionally answer-options (only in the case of multiple-choice questions). We extract appearance  $({A^t}_{t=1}^T \in \mathbb{R}^{7 \times 7 \times 2048})$  and motion  $({M^t}_{t=1}^T \in \mathbb{R}^{4096})$  features from the video using a pre-trained ResNet-152 [12] (*res5c*) and a pre-trained C3D [43] (*fc*6), respectively, where T is the sequence length. The question and answer options (when available) are encoded using two separate two-layer LSTMs, named as the Question Encoder LSTM and the Answer Encoder LSTM, respectively.

Following this, we pass the video features and the question/ answer encoding to the proposed spatio-temporal relational network, which consists of three main components: (a) Spatial Relation Module (SRM), (b) Global Context Encoder LSTM (GCE), and (c) Temporal Relations Module (TRM). The Spatial Relation Module takes appearance features ( $\{A^t\}_{t=1}^T$ ) as input and computes spatial relations among various objects. This can be seen as modeling object interactions in each frame individually. The Global Context Encoder LSTM takes motion features ( $\{M^t\}_{t=1}^T$ ) as input and captures action-dynamics of individual objects with global context at time t. Finally, the Temporal Relation Module takes the concatenated SRM-encoding ( $f_t$ ) and the GCE-encoding ( $\rho_t$ ) as input and computes how the spatial-relations and action-dynamics change over time. This corresponds to modeling temporal changes in both the interactions among different objects and the motion-dynamics of individual objects. The output encoding (Y) of the STRM module is passed to the Answer Decoder Module which produces an answer. The exact form depends on the specific question-answering task (Section 3.3). In the case of multiple-choice questions, we classify into one of five available options. Similarly, in the case of open-ended word questions, classification is used to predict one of the words in the vocabulary. On the other hand, regression is used to predict the actual number in case of open-ended number questions. The details are explained in the sections below.

#### **3.1** Text and Visual Representation

**Question and Answer representation**: Both the question and answer consist of a sequence of words in a vocabulary. We represent each word as a 300D vector using the GloVe word embedding [29] pretrained using the Common Crawl dataset. We denote the question and the answer by  $qi_{i=1}^N$  and  $ai_{i=1}^M$ , where M, N are respective sequence lengths. Similar to previous work [13], we encode the question and each of the answer options (when available) using the Question Encoder LSTM and the Answer Encoder LSTM, respectively. The final encoding is denoted as  $\gamma$ .

**Visual representation**: We extract appearance and motion information from a video using ResNet-152 [12] and C3D [43], pre-trained on ImageNet 2012 classification dataset [36] and Sport1M dataset [16], respectively. The appearance features, denoted as  $\{A^t\}_{t=1}^T \in \mathbb{R}^{7 \times 7 \times 2048}$ , are extracted from the *res5c* layer of the pre-trained ResNet-152 [12]. For computational reasons, we downscale these CNN feature maps using a *Downscale CNN* and denote the resultant feature descriptor as  $\{O^t\}_{t=1}^T \in \mathbb{R}^{3 \times 3 \times 256}$ . It can be seen as capturing possible objects representations in respective areas of the original image corresponding to  $3 \times 3$  locations in the



**Figure 3.1:** Spatio-temporal Relational Network architecture. The Spatial Relations Module (top) models arbitrary spatial-relations among all possible groups of objects for each frame individually. The Global Context Encoder LSTM (bottom left) models the action-dynamics with global context at time t. The concatenated output of these modules is then fed to the Temporal Relations Module (bottom right) which computes temporal relations among a temporally-ordered group of frames. Notice that, for simplicity, we have shown object-groups as pairs, however, in general they can be more than two.

CNN feature maps. Similarly, we extract motion features from the *fc*6 layer of a pre-trained C3D [43], denoted by  $\{M^t\}_{t=1}^T \in \mathbb{R}^{4096}$  where T is the sequence length.

#### **3.2** Spatio-temporal Relational Network

Inspired by relational networks [38, 54], we encode the ability to model spatiotemporal relationships right in the formulation of STRN. Hence, it acts as an effective prior for situations which require joint reasoning over both spatial and temporal domains. The input consists of an ordered temporal sequence of spatial frame-descriptors  $\{O^t\}_{t=1}^T$ , where each  $O^t$  contains *L* spatial object-representations  $\{o_x\}_{x=1}^L$ . In general, the spatial frame-descriptors can be any representation of interest. It can be structured in the case of bounding boxes or unstructured in the case of CNN feature maps. In this work, we use CNN feature maps  $(\{O^t\}_{t=1}^T \in \mathbb{R}^{3\times3\times256})$  which we obtain from ResNet-152 features  $(\{A^t\}_{t=1}^T \in \mathbb{R}^{7\times7\times2048})$  using a Downscale-CNN layer (see Figure 3.1). Given  $\{O^t\}_{t=1}^T$ , we define the basic-STRN as a composite function below:

$$\text{STRN}_{B}(O) = h_{\beta}^{T} \left( \sum_{a < b} g_{\alpha}^{T}(f_{a}, f_{b}) \right)$$
(3.1)

$$f_t = h_{\phi}^S \left( \sum_{a,b} g_{\theta}^S(o_a, o_b) \right)$$
(3.2)

Equation (3.2) corresponds to SRM and is responsible for computing spatial relations ( $f_t$ ) for each  $O^t = \{o_x\}_{x=1}^L$ . In particular, spatial-relation function  $g_{\theta}^S$  infers whether and how the two inputs are related to each other. The relations are computed for all possible input combinations  $o_a, o_b \in \{O^t\}$ . The individual objectobject relations are then agglomerated and reasoned over by the function  $h_{\phi}^S$ . In a similar way, Equation (3.1) corresponds to TRM and computes the temporal relations among a sequence of ordered inputs  $\{f_t\}_{t=1}^T$  obtained from  $\{O^t\}_{t=1}^T$  using Equation (3.2). The temporal-relation function  $g_{\alpha}^T$  computes the individual frame-frame relations, which are agglomerated, and reasoned over by the function  $h_{\beta}^T$ . Hence, the combination of Equations (3.1) and (3.2) models the temporal relations among the spatial relations, achieving spatio-temporal relational reasoning. In other words, STRN\_B models the interactions among objects and how they evolve over time. Following previous work [38, 54], we use fully-connected layers to represent the functions  $g_{\alpha}^T, h_{\beta}^T, g_{\theta}^S$ , and  $h_{\phi}^S$ , which are parameterized by  $\alpha, \beta, \theta$ , and  $\phi$ .

**Capturing Action Dynamics:** In STRN\_B, we used the  $f_i$ 's in Equation (3.1) to be spatial relations, which helped us model evolving object interactions. However, apart from interactions, some queries may also inquire about changes in motion (or appearance) of individual objects. In order to capture action-dynamics, we

leverage motion features  $({M^t}_{t=1}^T \in \mathbb{R}^{4096})$ , which represent course motion information corresponding to each object in a video [43]. However, both C3D and Flow features are known to encode only short-term temporal information [46]. Hence, we additionally make use of a Global Context Encoder LSTM to capture longterm global context. Instead of using only the SRM-encoding, we consider the concatenation of the SRM-encoding ( $f_t$ ) and the GCE\_LSTM-encoding ( $\rho_t$ ) while computing temporal relations. The resultant model, STRN\_GC, captures both the interactions among object-groups and the action-dynamics of individual objects with global-context.

$$\mathrm{STRN}_{\mathrm{GC}}(O) = h_{\beta}^{T} \left( \sum_{a < b} g_{\alpha}^{T}(\Omega_{a}, \Omega_{b}) \right)$$
(3.3)

$$\rho_t = \text{LSTM}(M^t, \rho_{t-1}) \tag{3.4}$$

where  $\Omega_t$  is obtained by concatenating  $f_t$  (Eq. 3.2) and  $\rho_t$  (Eq. 3.4),  $\rho_t$  is the hidden state of the Global Context Encoder LSTM at time *t* and  $M^t$  is the  $t^{th}$  motion embedding.

**Conditioning and Multi-scale Relations**: For tasks like video question answering, different questions may require different types of relations. Hence, we model dependence on questions by conditioning the relation-functions  $g_{\alpha}^{T}$ , and  $g_{\theta}^{S}$ to obtain query-specific variants. For instance, functions  $g_{\alpha}^{T}(f_{a}, f_{b})$ , and  $g_{\theta}^{S}(o_{a}, o_{b})$ transform to  $g_{\alpha}^{T}(f_{a}, f_{b}, \gamma)$ , and  $g_{\theta}^{S}(o_{a}, o_{b}, \gamma)$ , where  $\gamma$  is the question encoding obtained through a text-encoder LSTM, similar to one used in Jang [13].

Additionally, inspired by multi-scale temporal relational networks [54], instead of computing relations among only two possible frames/objects at a time, we generalize the relation-functions  $g_{\alpha}^{T}(f_{a}, f_{b}, \gamma)$ , and  $g_{\theta}^{S}(o_{a}, o_{b}, \gamma)$  to consider multiple frames/objects:  $g_{\alpha}^{T}(f_{a}, f_{b}, ...f_{m}, \gamma)$ , and  $g_{\theta}^{S}(o_{a}, o_{b}, ...o_{n}, \gamma)$ , for *m* frames and *n* objects, respectively. Then, we consider multiple relation-functions each specializing to capture relationships for a given value of (m, n) frames/ objects at a time. This allows modelling relationships at multiple scales. We define the **M** multi-scale, **N** multi-object Spatio-Temporal Relational Network (STRN) as:

$$\mathrm{STRN}_{S}(O, m, n) = h_{\beta}^{T} \left( \sum_{a < b.. < m} g_{\alpha}^{T}(\Omega_{a}, \Omega_{b}, ..\Omega_{m}, \gamma) \right)$$
(3.5)

$$f_t = h_{\phi}^S \left( \sum_{a,b..n} g_{\theta}^S(o_a, o_b, ...o_n, \gamma) \right)$$
(3.6)

$$\operatorname{STRN}(O, M, N) = \sum_{m=2, n=2}^{M, N} \left( \operatorname{STRN}_{S}(O, m, n) \right)$$
(3.7)

Each STRN\_S(O, m, n) in Equation (3.5) computes relationships among a given value of *m* temporal-objects and *n* spatial-objects and has its own *h* and *g* functions. Additionally, we consider the temporal-relation function,  $g_{\alpha}^{T}$  (from Eq. 3.3 and 3.4), which captures both object-interactions and action-dynamics. STRN(O, M, N) in Equation (3.7) accumulates relationships from multiple STRN\_S(O, m, n) for all values of (m, n), ranging from (2,2) to (M, N). Hence, we obtain the M-multi-scale and N multi-object Spatio-Temporal Relation Network (STRN) which we use as our final model.

#### **3.3** Answer Decoder

The final encoding from the spatio-temporal relation network (Y) is passed to the Answer Decoder Module which generates the actual answer. Depending of the question type, we have three different types of modules:

**Multiple-choice:** It is modeled as a classification task to choose among five different options. We define a linear regression function which takes the TRM-encoding (Y) as input and outputs a real-valued score for each multiple-choice answer-candidate:

$$s = W_{MC}^T Y \tag{3.8}$$

where  $W_{MC}$  are trainable model parameters. To optimize, we use hinge-loss:  $max(0, 1 + s_n - s_p)$ , where  $s_p$  and  $s_n$  are scores of the correct and incorrect answer, respectively. We use this decoder for repeating action and state transition tasks in the TGIF-QA dataset.

**Open-ended numbers:** We model it has a regression problem to predict the correct number. Similar to the above, we define a linear regression function:

$$s = [W_N^T Y + b] \tag{3.9}$$

where [.] denotes rounding, Y is the TRM-encoding,  $W_N$  are model parameters and b is the bias. We optimize the network using  $l_2$  loss between the ground truth and the predicted value. This decoder is used for the repetition count task in the TGIF-QA dataset.

**Open-ended word:** This is modeled as a classification task among a vocabulary of words. We define a linear classifier which selects an answer from a vocabulary V:

$$o = softmax(W_w^T Y + b) \tag{3.10}$$

where  $W_w$  are model parameters and *b* is the bias. We use cross-entropy loss and the final answer is obtained using:  $y = argmax_{y \in V}(o)$ . We use this decoder for the SVQA dataset and also for the FrameQA task in the TGIF-QA dataset.

#### 3.4 Training

Following previous work [8, 13, 20, 41], we train separate models for each task of the TGIF-QA dataset and one model for the entire SVQA dataset. We use pre-trained ResNet-152, C3D, and GloVe word embeddings for video and text representations. We train the rest of the network in an end-to-end manner using backpropagation. Specifically, the parameters of Downscale CNN, Spatial Temporal Module, Temporal Relation Module, Global Context Encoder LSTM, Question Encoder LSTM, Answer Encoder LSTM, and Answer Decoder Module are trained together in an end-to-end manner.

Similar to previous work [54], we choose the functions  $g_{\alpha}^{T}$ ,  $h_{\beta}^{T}$ ,  $g_{\theta}^{S}$ ,  $h_{\phi}^{S}$  in Equation (3.5) and (3.6) as fully-connected neural network layers. This is quite simple in comparison to previous work on video question answering which uses LSTMs or complex memory networks [8, 13]. Moreover, feed-forward networks, unlike LSTMs, can be parallelized efficiently which leads to notably fast training time

and real-time inference.

One drawback of relational networks is their quadratic  $O(n^2)$  computational complexity with respect to its inputs (*n*). This is because they consider all possible combinations while computing relationships among its inputs. In case of the SRM, we make it tractable by downscaling the CNN feature maps from  $7 \times 7$  to  $3 \times 3$  using a Downscale CNN. In the case of TRM, where we have long sequences of frames, we solve this problem using subsampling S = 3, as done in Zhou [54]. In Equation (3.7), we choose M=10 different scales, which means we consider 2-10 frames at a time while computing temporal relations. Similarly, we choose N=3, which means we consider 2-3 spatial-objects at a time while computing spatial relations. For more implementation details, refer to Section 4.2.1.

### **Chapter 4**

## **Experiments**

This chapter outlines the details of our experiments. First, we explain the datasets and details used for evaluation. Then, we compare our method with the state-ofthe-art baselines on both the TGIF-QA and the SVQA dataset. Finally, we show an ablation study which demonstrates the effectiveness of joint spatio-temporal relational reasoning.

#### 4.1 Datasets

We validate our model on two large-scale video question answering datasets: TGIF-QA, based on real world animated GIFs and SVQA, based on synthetically generated videos. Both the datasets are explained in detail below.

#### 4.1.1 TGIF-QA

TGIF-QA [13] is a large-scale dataset containing 165K QA pairs collected from 71K real-world animated Tumblr GIFs. TGIF-QA dataset is based on the e Tumblr GIF (TGIF) dataset [21]. The dataset contains two types of questions: videoQA and FrameQA. VideoQA requires looking at multiple frames and understanding the spatio-temporal relationships in the video. On the other hand, FrameQA, similar to imageQA, requires looking at only one appropriate frame in the video. A total of 112,082 VideoQA pairs from 53,247 GIFs were generated using a combination of

crowdsourcing and template-based approach. For FrameQA, a total of 53,083 QA pairs from 39,479 GIFs were automatically generated from captions in the TGIF dataset [21] using the same setup of Ren [32]. The questions are categorized into four separate tasks, which are described below. For detailed statistics, refer to Table 4.1

- 1. **Repeating Action (Action)** aims to name the event that happened a specific number of times in the video. This is a multiple-choice task where the correct answer is one of the five available options (Fig 4.1a). For example, "what does the woman do 3 times?".
- 2. **State Transition (Trans)**, similarly, is a multiple-choice task with five options. Questions ask about state transitions like facial expressions (from happy to sad), among others (Fig 4.1b).
- 3. **FrameQA** is an open-ended task which, similar to image-QA, can be answered by looking at one of the "appropriate" frames in the video. However, the range of possible answers span the entire vocabulary (Fig 4.1c). For example, "what is dancing in the cup?".
- 4. **Repetition Count (Count)** aims to count the number of times a given event happens in the video. This is an open-ended task and answers lie in a range of integers: 0 to 10 (Fig 4.1d). For example, "how many times does the cat lick?"

Toolz	#	# QA pair	s	# GIFs			
Task	Train	Test	Total	Train	Test	Total	
Rep. Action	20,475	2,274	22,749	20,475	2,274	22,749	
Transition	52,704	6,232	58,936	26,352	3,116	29,468	
FrameQA	39392	13691	53083	37089	9219	40308	
Rep. Count	26,843	3,554	30,397	26,843	3,554	30,397	
Total	139,414	25,751	165,165	62,846	9,575	71,741	

Table 4.1: Statistics of the TGIF-QA dataset.

#### 4.1.2 SVQA

SVQA [41] is a synthetically generated dataset designed to control and minimize language biases in existing videoQA datasets. It contains 120K questions asked on 12K videos with moving objects like sphere, cylinder or cube (Fig 4.2). The questions are compositional and require a series of reasoning steps (like comparison and arithmetic) in both spatial and temporal domains. All QA pairs are openended word type and are categorized into multiple categories and subcategories as described below. The questions are picked verbatim from the SVQA dataset.

- 1. **Count** Questions about counts of objects, events, and actions in the video. For example, "what number of cubes are there?"
- Exist Yes/ No questions about the existence of actions, events, and objects.
   For example, "are there any yellow balls rotating present?"
- 3. **Query** Questions querying about Color, Size, Action, Direction (of motion) and Shape of objects. For example, "what action type is the small object that is to the right of the green object at start?"
- 4. **Integer Comparison** Comparison questions (More, Less, and Equal) about counts of objects, actions, events, etc. For example, "what action type is the red object that has the same size as the white cylinder?"
- 5. Attribute Comparison Comparison questions about attributes (Color, Size, Action, Direction, and Shape) of objects. For example, "do the small ball that is to the left of the big object at start and the small cube have the same color?"

Since the exact train-val-test subsets of the SVQA dataset are not readily available, we randomly sample a new split as shown in Table 4.2, which is consistent with previous work [41]. In comparison to TGIF-QA, it contains more complex questions requiring more elaborate spatio-temporal reasoning. However, unlike real-world GIFs in TGIF-QA, it contains perceptually-simpler scenes consisting of a few synthetic objects. These two datasets are well suited for our task because they contain well formed questions that require complex spatio-temporal reasoning.

Question Category	Sub Category	Train	Val	Test
Count		19320	2760	5520
Exist		6720	960	1920
	Color	7560	1056	2160
	Size	7560	1056	2160
Query	Action Type	6720	936	1920
	Direction	7560	1056	2160
	Shape	7560	1056	2160
	More	2520	600	720
Integer Comparison	Equal	2520	600	720
	Less	2520	600	720
	Color	2520	216	720
	Size	2520	216	720
Attribute Comparison	Action Type	2520	216	720
	Direction	2520	216	720
	Shape	2520	216	720
Total QA pairs	83160	11880	23760	
Total Videos	8400	1200	2400	

 Table 4.2: Statistics of the SVQA dataset.

#### 4.2 Implementation Details

#### 4.2.1 Setup

We implement our model and design our experiments using PyTorch. Following previous work [8, 13, 20, 41], we train separate models for each task of the TGIF-QA dataset and one model for the entire SVQA dataset. Similarly, we set the maximum number of uniformly-sampled frames in a video to 35. To encode text, we use the 300D Glove [29] word embeddings and take the output of the final layer of a text-encoder LSTM as the question-encoding (taken as answer-encoding in case of multiple-choice questions). Both the text-encoder and global-context-encoder are two layer LSTMs with 512 hidden units. In all our experiments, we use a batch size of 64. We train our networks in an end-to-end fashion using Adam optimizer with an initial learning rate of 0.001. Wherever applicable, we use a dropout of 0.2. The functions  $g^T_{\alpha}, h^T_{\beta}, g^S_{\theta}, h^S_{\phi}$  in Equation (3.5) and (3.6) are fully-connected networks

with 2, 1, 2, 2 layers and and 256, 256, 256, 256 hidden units, respectively. In Equation (3.7), we choose M=10 different scales, which means we consider 2-10 frames at a time while computing temporal relations. Since the number of possible combinations of frames can be large, we follow Zhou [54] and randomly sample S = 3 possible frame-sequences for each separate scale. Similarly, we choose N=3, which means we consider 2-3 spatial-objects at a time while computing spatial relations. We do not subsample spatial-relations but we downscale the appearance-features from  $\{A^t\}_{t=1}^T \in \mathbb{R}^{7 \times 7 \times 2048}$  to  $\{O^t\}_{t=1}^T \in \mathbb{R}^{3x3x256}$  using a Downscale-CNN having (384, 192, 256) filters and (2,3,2) kernels.

#### 4.2.2 Metrics

Following previous work [8, 13, 20, 41], we use classification accuracy (ACC) as an evaluation metric for all tasks of the SVQA dataset and also the Trans, Action and FrameQA tasks of the TGIF-QA dataset, which is defined as,

$$ACC = \frac{\text{Questions Answered Correctly}}{\text{Total Questions}} \times 100.$$
 (4.1)

For the Count task of the TGIF-QA dataset, we use Mean Squared Error (MSE) between the predicted value and the ground truth value as an evaluation metric, which is defined as,

$$MSE = \frac{\sum_{i}^{n} (\text{Prediction}_{i} - \text{Ground Truth}_{i})^{2}}{\text{Total Questions (n)}} .$$
(4.2)

For the TGIF-QA dataset, we train separate models corresponding to each task. On the other hand, for the SVQA dataset, we train only one model model and accuracy is computed by considering appropriate subsets of questions. This is consistent with previous work on video question answering [8, 13, 20, 41].

#### **4.3** Comparison with State-of-the-art Methods

#### 4.3.1 TGIF-QA Dataset

The results of our model and existing baselines on the TGIF-QA dataset are shown in Table 4.3. At the very top, *Random* and *Text* correspond to selecting an answer randomly and learning a model without any visual input, respectively. In the next four lines, we show results obtained using the image-VQA based baselines, which either mean-pool the video features (aggr) or average the results (avg). VIS+LSTM [32] combines image-representation with textual features encoded by an LSTM, and VQA-MCB [6], uses multimodal compact bilinear pooling to handle imagetext fusion and spatial attention.

The next six lines show the results correspond to videoQA methods (refer to Section 2.1 for a detailed comparison). The letters inside the brackets correspond to the features used to train the model: R means ResNet, C means C3D and F means Flow. The first set of models, ST and its variants [13], are shown in lines 7-10 in Table 4.3. Suffixes SP and TP corresponds to whether Spatial and Temporal attention mechanism has been used. We encourage the reader to refer to Fig 3, 4 of Jang [13] for more details. Co-memory [8] model is based on Dynamic Memory Networks (DNMs) which uses a motion-appearance co-memory attention memory. PSAC (Positional Self-Attention with Co-attention) [20] uses a Self-Attention based approach as opposed to LSTMs which models global dependencies by attending to all positions in the sequence.

The last three rows show the result of our models. Our STRN model outperforms all other approaches on all tasks which require spatio-temporal reasoning: Action (2.74%), Trans (2.4%) and Count (4.63%)  $(\frac{4.10-3.91}{4.10})$  by a significant margin. On the other task, FrameQA, which can be answered using a single frame, we outperform all but one approach (PSAC). We gain this increase in performance despite not taking advantage of Flow features and complex memorynetworks (used in Co-memory), or co-attention mechanisms (used in PSAC). In the STRN-GC variants, we do not use the Global Context Encoder LSTM. As shown, we get good results even without using action-dynamics with global-context.

Model	Action ↑	Trans $\uparrow$	FrameQA $\uparrow$	$Count \downarrow$
Random [13]	20.00	20.00	0.06	6.92
Text [13]	47.91	56.93	39.26	5.01
VIS+LSTM(aggr) [13]	46.80	56.90	34.60	5.09
VIS+LSTM(avg) [13]	48.80	34.80	35.00	4.80
VQA-MCB(aggr) [13]	58.90	24.30	25.70	5.17
VQA-MCB(avg) [13]	29.10	33.00	15.50	5.54
ST(R+C) [13]	60.10	65.70	48.20	4.38
ST-SP(R+C) [13]	57.30	63.70	45.50	4.28
ST-TP(R+C) [13]	60.80	67.10	49.30	4.40
ST-SP-TP(R+C) [13]	57.00	59.60	47.80	4.56
Co-memory (R+F) [8]	68.20	74.30	51.50	4.10
PSAC (R) [20]	70.40	76.90	55.70	4.27
STRN-GC (R) [ours]	72.16	79.18	52.90	4.42
STRN-GC (C) [ours]	71.42	78.85	50.04	4.10
STRN (R+C) [ours]	73.14	79.30	52.96	3.91

**Table 4.3:** Comparison with state-of-the-art on TGIF-QA dataset.  $\uparrow$  means higher numbers correspond to better performance (ACC) and  $\downarrow$  means lower numbers correspond to better performance (MSE).

#### 4.3.2 SVQA Dataset

Results of our model and existing baselines on the SVQA dataset are summarized in Table 4.4. Similar to Table 4.3, the two lines at the top correspond to random and text-only models. *GRU+AVG* is an image-VQA based approach which averages the sequential video representation and concatenates it with the question encodng generated by the GRU. *2GRU* uses two GRUs to encode questions and videos separately and concatenates them to generate the answer. SP-TP is the same as Jang [13] with temporal attention (TP). *SVQA* is the model proposed in Song [41]. It is comprised of a refined GRU (temporal-attention GRU, abbreviated as TA-GRU), to capture long-term temporal dependencies for multi-step reasoning.

The last line presents the results of our proposed model. As shown in the last column (All) of Table 4.4, we outperform all methods by a margin of **2.68%**. We perform better in Exist, Count and five out of thirteen sub-categories of Integer Comparison, Attribute Comparison and Query. We do competitively in three and

	Exist	Count	Intege	er Com	parisor	1 <i>1</i>	Attribu	te Con	nparise	n
			More	Equal	Less	Color	Size	Туре	Dir	Shape
Random [41]	50.00	22.00	50.00	50.00	50.00	50.00	50.00	50.00	50.00	50.00
Text [41]	52.92	32.41	75.14	56.39	57.81	47.73	52.56	53.12	53.55	51.56
GRU+AVG [41]	51.77	33.18	59.66	54.12	59.38	52.27	50.00	51.13	53.27	47.58
2GRU [41]	53.54	35.02	68.18	53.70	56.10	54.12	51.28	51.70	52.70	47.86
ST-TP [41]	51.46	32.54	58.46	50.39	53.52	49.74	54.56	53.12	51.95	50.39
SVQA [41]	52.03	38.20	74.28	57.67	61.60	55.96	55.90	53.40	57.50	52.98
STRN [ours]	54.01	44.67	72.22	57.78	62.92	56.39	55.28	50.69	50.14	50.00
			Query	,		All				
	Color	Size	Type	Dir	Shape					
Random [41]	12.50	50.00	50.00	25.00	33.33	33.33				
Text [41]	12.27	51.07	48.65	25.23	32.70	39.95				
GRU+AVG [41]	19.78	51.91	53.33	28.26	38.29	41.43				
2GRU [41]	19.59	53.50	58.38	34.79	38.34	41.85				
ST-TP [41]	21.23	53.81	55.70	36.08	40.60	40.47				
SVQA [41]	23.39	63.30	62.90	43.20	41.69	44.90				
STRN [ours]	24.31	59.68	59.32	28.24	44.49	47.58				

**Table 4.4:** Comparison with the state-of-the-art on different categories of the SVQA dataset. We split the table into two parts (Top - Exist, Count, Integer and Attribute Comparison; Bottom - Query and aggregate of All) for better readability. Higher numbers correspond to better performance.

worse in five sub-categories. However, we would also like to highlight a substantial improvement of **6.47% in the Count category**, which unlike sub-categories, forms a significant portion (23%) of the total dataset (see Fig 3 of [41]). This result is in consonance with TGIF-QA dataset (Table 4.3), where we gain a substantial improvement of **4.63% in the Count** task. Since counting is considered a complex task requiring elaborate spatio-temporal reasoning, we believe this improvement conclusively demonstrates the effectiveness of our approach.

Our approach falls short by 1-4% in eight out of thirteen subcategories of Integer Comparison, Attribute Comparison, and Query. Interestingly, for two subcategories, Dir of Query and Attribute Comparison, we notice a substantial drop in performance by 7.36% and 14.96%, respectively. We posit the reason to be the spatial-location agnostic nature of how we compute spatial relations. As mentioned in Section 3.2, we agglomerate the spatial relations, computed by the spatialrelation function  $g^{S}_{\theta}$ , using a summation operator. The summation operator loses any order in terms of the direction in the spatial coordinates, hence, drop the in performance in Dir category. However, this hypothesis needs to be tested with more experiments as part of future work.

#### 4.4 Ablations

We conduct Ablations on only the TGIF-QA dataset. In this experiment, we show the effectiveness of joint spatio-temporal relational reasoning, as opposed to individual spatial or temporal relational reasoning. To show that our experiment generalizes over different modalities, we consider separate models trained individually on both ResNet (ResNet-res5c) and C3D (C3D-conv5b). In order to avoid interference, we do not use Global Context Encoder LSTM and we call the resultant model as STRN-GC. We summarize the results in Table 4.5. We consider two baselines. In STRN-GC-TRM, we replace the Temporal Relations Module (TRM) with a two-layer LSTM as a baseline to model temporal relations. We initialize the hidden state of the LSTM using the last hidden state of the text-encoder LSTM, following the ST models of Jang [13]. In STRN-GC-SRM, we replace the Spatial Relations Module (SRM) using an expressive CNN. As shown in the table, STRN-GC significantly outperforms both baselines in all four tasks, which shows the effectiveness of joint spatio-temporal relational reasoning.

ResNet-resSc						
Model	Action	Trans	Frame	Count		
STRN-GC-TRM	64.95	71.25	44.86	4.50		
STRN-GC-SRM	66.09	77.36	49.57	4.54		
STRN-GC	72.16	79.18	52.90	4.42		
	C3D-conv5b					
Model	Action	Trans	Frame	Count		
STRN-GC-TRM	63.10	71.26	44.63	4.18		
STRN-GC-SRM	67.72	77.70	47.53	4.40		

DacNat

Table 4.5: Effectiveness of joint spatio-temporal reasoning as opposed to individual spatial or temporal relational reasoning on TGIF-QA dataset.

### 4.5 Qualitative Results

#### (a) Repeating Action



Q) What does the guy do four times?
 (Ours): Throw money
 ST-TP: Spin

(b) State Transition



Q) What does the man do after lick lips? (Ours): Bite donut ST-TP: Running

(c) FrameQA



Q) What is the color of the hair? (Ours): Brown ST-TP: Pleasant





Q) How many times does the man shake shoulders? (Ours): 4 ST-TP: 5

**Figure 4.1:** [Best viewed in color] A comparison of the qualitative results of ST-TP [13] and STRN (Ours). Green and Red refers to correct and incorrect predictions, respectively.

#### (a) Exist



**Q)** Are there any balls moving left that have the same size as the red ball? **Prediction:** Yes

#### (c) Integer Comparison (more)



**Prediction:** Yes

#### (e) Query (color)



**Q)** What color is the small object rotating that has the same shape as the big object? **Prediction:** Cyan

#### (g) Attribute Comparison (direction)



**Q)** Do the small cyan object and the blue ball have the same action direction? **Prediction:** No

#### (b) Count



**Q)** How many objects are either big objects that have the same color as the cylinder moving forward or small blue cubes rotating? **Prediction:** Two

#### (d) Integer Comparison (less)



**Q)** Are there fewer big cylinders than big red object? **Prediction:** No

#### (f) Query (shape)



Q) What shape is the big white object that is starts moving after the green object? Prediction: Cube

#### (h) Attribute Comparison (size)



**Q)** Does the yellow cylinder have the same size as the grey cube? **Prediction:** No

Figure 4.2: [Best viewed in color] Qualitative results for different categories of the SVQA dataset where our approach predicted the correct answers.

#### (a) SVQA (Count)



Q) What number of things are either small objects that are behind the small blue cube rotating at start or big gray cylinders? Prediction: One Ground Truth: Three

#### (c) SVQA (Integer Comparison)



Q) Is the number of cyan objects the same as the number of blue cubes? Prediction: No Ground Truth: Yes

(e) TGIF-QA (Trans)



Q) What does the man do before bow head? Prediction: Tip over Ground Truth: Sing a song

#### (g) TGIF-QA (FrameQA)



Q) what is the color of the hair? Prediction: Black Ground Truth: Brown

#### (b) SVQA (Attribute Comparison)



Q) Do the small cube rotating and the small object moving backward have the same color? Prediction: No Ground Truth: Yes

#### (d) SVQA (Exist)



Q) Are there any balls that have the same color as the cylinder? Prediction: No Ground Truth: Yes

#### (f) TGIF-QA (Count)



Q) How many times does the cat step ? Prediction: 9 Ground Truth: 10

#### (h) TGIF-QA (Action)



Q) What does the man on left do 3 times? Prediction: Clap Ground Truth: Shake fists

**Figure 4.3:** [Best viewed in color] Qualitative results for different categories of the SVQA and TGIF-QA datasets where our approach predicted the **incorrect** answers.

### **Chapter 5**

## Conclusion

We propose a general-purpose neural network module named Spatio-Temporal Relational Network (STRN). Our module uses relational networks (RN) which are inherently designed with the prior knowledge to model relationships. It computes spatial relationships using a Spatial Relation Module (SRM) and models how they change in the temporal domain using a Temporal Relation Module (TRM). This captures object interactions and how they change over time. Separately, it models how the dynamics of individual objects changes over time using a combination of a pre-trained 3D Conv (C3D) and a Temporal Relation Module (TRM), which can be seen as capturing time-changing action-dynamics of individual objects. We demonstrate the effectiveness of our module on two benchmark video question answering datasets which require spatio-temporal reasoning: a real-world (TGIF-QA) and a synthetic (SVQA) datasets. We achieve state-of-the-art results on both the datasets. Additionally, we achieve substantial improvement (4-7%) in the challenging counting task.

One line of future work is to explore different types of agglomeration of spatialrelations in the Spatial Relation Module. As discussed in Section 4.3.2, the summation agglomeration seems to lose the spatial order of relations. Hence, it is hypothesized to be the reason for the substantial drop in performance (7-15%) in the Dir subcategory of Attribute comparison and Query types. This hypothesis needs to be tested by conducting experiments on different types of agglomerations. As a natural next step, these experiments can be conducted in both the spatial and the temporal domains to infer their advantages and disadvantages.

Another line of research is to explore the use of a query-dependent agglomeration in both spatial and temporal relation modules. Instead of an agnostic agglomeration (sum or concatenation) of spatial and temporal relations, we can condition these using the question-encoding  $\gamma$ . As an example, consider the spatial relation module as follows.

$$f_t = h_{\phi}^S \left( \sum_{a,b} g_{\theta}^S(o_a, o_b) . \alpha_{a,b} \right)$$
(5.1)

$$\alpha_{a,b} = \operatorname{align}(g^{S}_{\theta}(o_{a}, o_{b}), \gamma)$$
(5.2)

where 
$$\operatorname{align}(a,b) = \frac{\exp(\operatorname{score}(a,b))}{\sum_{a'=1,b'=1}^{n} \exp(\operatorname{score}(a',b'))}$$
 (5.3)

In Eq. 5.2,  $\alpha_{a,b}$  represents the relative strength of a relation between two particular objects (a,b) with respect to the particular query represented by question encoding  $\gamma$ . Each relation in Eq. 5.1 is weighted by their relative strength to compute the attended relation.

We would also like to expand the scope of our work by avoiding one or more assumptions mentioned in Section 1.1.1. For instance, our model can also be enhanced by considering more frames per video. Currently, our model only considers 35-40 frames per video, which is a fixed number depending on the dataset. This limits the ability of our algorithm to detect fleeting events in the video and it also caps the number of detectable events. This limitation is similar to earlier object detection algorithms which were dependent on a fixed number of regions proposed by off-the-shelf Region Proposal Networks (RPE). As a first step, we can explore an off-the-shelf Event Proposal Networks (counterpart of RPEs for videos), however, an end-to-end system to train videos which can propose events and compute features at the same time should be the ultimate goal (similar to the image-based object detection system Faster-RCNN [33]).

Our model can also be enhanced by considering structured representations of spatial objects. Currently, we consider CNN feature-maps as implicitly capturing unstructured object representations at different locations of the image. However, it has limitations in the representation of multiple and a variable number of objects. For instance, CNN feature maps in the current implementation only allows representing one object per region and caps the number of objects by the size of feature maps. As a remedy, an off-the-shelf pre-trained object detector can be used to extract explicit object representations which can be used in spatial relation module to capture relationships. Since this is also the preferred input type for Graph Neural Networks (GNNs), another interesting future research can compare the spatio-temporal reasoning abilities of relational networks and graph neural networks.

Lastly, there is a need for a more comprehensive real-world video question answering dataset. TGIF-QA, although a good step in this direction, lacks enough diversity of tasks and is limited in the number of training examples per task. Also, we need a formal framework detailing a taxonomy of different tasks in video understanding. It should detail the steps required to collect data for each of this task and the specific skills which each of this task aims to benchmark. Also, there is a need for a deeper understanding of what neural network architecture works better for what task. A comparative study to benchmark different neural network architectures on different tasks will give concrete direction to future researchers in this field.

## **Bibliography**

- S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. Lawrence Zitnick, and D. Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2425–2433, 2015. → page 11
- [2] M. Baccouche, F. Mamalet, C. Wolf, C. Garcia, and A. Baskurt. Sequential deep learning for human action recognition. In *International workshop on Human Behavior Understanding*, pages 29–39. Springer, 2011. → page 16
- [3] K. Chen, J. Wang, L.-C. Chen, H. Gao, W. Xu, and R. Nevatia. Abc-cnn: An attention based convolutional neural network for visual question answering. arXiv preprint arXiv:1511.05960, 2015. → page 12
- [4] N. Dalal, B. Triggs, and C. Schmid. Human detection using oriented histograms of flow and appearance. In *European Conference on Computer Vision*, pages 428–441. Springer, 2006. → page 16
- [5] P. Dollár, V. Rabaud, G. Cottrell, and S. Belongie. Behavior recognition via sparse spatio-temporal features. VS-PETS Beijing, China, 2005. → page 16
- [6] A. Fukui, D. H. Park, D. Yang, A. Rohrbach, T. Darrell, and M. Rohrbach. Multimodal compact bilinear pooling for visual question answering and visual grounding. *arXiv preprint arXiv:1606.01847*, 2016. → pages 11, 12, 32
- [7] H. Gao, J. Mao, J. Zhou, Z. Huang, L. Wang, and W. Xu. Are you talking to a machine? dataset and methods for multilingual image question. In *Advances in Neural Information Processing Systems*, pages 2296–2304, 2015. → page 12
- [8] J. Gao, R. Ge, K. Chen, and R. Nevatia. Motion-appearance co-memory networks for video question answering. In *Proceedings of the IEEE*

Conference on Computer Vision and Pattern Recognition, pages 6576–6585, 2018.  $\rightarrow$  pages 1, 2, 14, 15, 16, 25, 30, 31, 32, 33

- [9] R. Girdhar, D. Ramanan, A. Gupta, J. Sivic, and B. Russell. Actionvlad: Learning spatio-temporal aggregation for action classification. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 971–980, 2017. → page 16
- [10] R. Goyal, S. E. Kahou, V. Michalski, J. Materzynska, S. Westphal, H. Kim, V. Haenel, I. Fruend, P. Yianilos, M. Mueller-Freitag, et al. The" something something" video database for learning and evaluating visual common sense. In *Proceedings of the IEEE International Conference on Computer Vision*, volume 1, page 3, 2017. → pages 1, 15
- [11] Y. Goyal, T. Khot, D. Summers-Stay, D. Batra, and D. Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6904–6913, 2017. → pages 11, 12
- [12] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. → pages 7, 9, 19, 20
- [13] 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. → pages x, 1, 2, 4, 7, 13, 15, 20, 23, 25, 27, 30, 31, 32, 33, 35, 36
- [14] 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. → page 18
- [15] J. Johnson, B. Hariharan, L. van der Maaten, J. Hoffman, L. Fei-Fei, C. L. Zitnick, and R. B. Girshick. Inferring and executing programs for visual reasoning. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3008–3017, 2017. → page 12
- [16] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei. Large-scale video classification with convolutional neural networks. In

Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, pages 1725–1732, 2014.  $\rightarrow$  page 20

- [17] K.-M. Kim, M.-O. Heo, S.-H. Choi, and B.-T. Zhang. Deepstory: video story qa by deep embedded memory networks. arXiv preprint arXiv:1707.00836, 2017. → pages 1, 14, 15
- [18] A. Klaser, M. Marszałek, and C. Schmid. A spatio-temporal descriptor based on 3d-gradients. 2008. → page 16
- [19] R. Krishna, K. Hata, F. Ren, L. Fei-Fei, and J. Carlos Niebles.
   Dense-captioning events in videos. In *Proceedings of the IEEE International* Conference on Computer Vision, pages 706–715, 2017. → page 15
- [20] X. Li, J. Song, L. Gao, X. Liu, W. bing Huang, X. He, and C. Gan. Beyond rnns: Positional self-attention with co-attention for video question answering. In AAAI 2019, 2019. → pages 15, 25, 30, 31, 32, 33
- [21] Y. Li, Y. Song, L. Cao, J. Tetreault, L. Goldberg, A. Jaimes, and J. Luo. Tgif: A new dataset and benchmark on animated gif description. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4641–4650, 2016. → pages 27, 28
- [22] J. Lu, J. Yang, D. Batra, and D. Parikh. Hierarchical question-image co-attention for visual question answering. In *Advances In Neural Information Processing Systems*, pages 289–297, 2016. → page 12
- [23] C.-Y. Ma, A. Kadav, I. Melvin, Z. Kira, G. AlRegib, and H. Peter Graf. Attend and interact: Higher-order object interactions for video understanding. In *Proceedings of the IEEE Conference on Computer Vision* and Pattern Recognition, pages 6790–6800, 2018. → pages 2, 14, 16
- [24] M. Malinowski and M. Fritz. A multi-world approach to question answering about real-world scenes based on uncertain input. In Advances in Neural Information Processing Systems, pages 1682–1690, 2014. → page 11
- [25] M. Malinowski, M. Rohrbach, and M. Fritz. Ask your neurons: A neural-based approach to answering questions about images. In *Proceedings* of the IEEE International Conference on Computer Vision, pages 1–9, 2015. → page 12
- [26] J. Mun, P. Hongsuck Seo, I. Jung, and B. Han. Marioqa: Answering questions by watching gameplay videos. In *Proceedings of the IEEE*

International Conference on Computer Vision, pages 2867–2875, 2017.  $\rightarrow$  pages 2, 13, 16

- [27] S. Na, S. Lee, J. Kim, and G. Kim. A read-write memory network for movie story understanding. In *Proceedings of the IEEE International Conference* on Computer Vision, pages 677–685, 2017. → page 14
- [28] Y. Pan, T. Mei, T. Yao, H. Li, and Y. Rui. Jointly modeling embedding and translation to bridge video and language. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 4594–4602, 2016. → page 16
- [29] J. Pennington, R. Socher, and C. Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on Empirical Methods in Natural Language Processing*, pages 1532–1543, 2014. → pages 20, 30
- [30] Z. Qiu, T. Yao, and T. Mei. Learning spatio-temporal representation with pseudo-3d residual networks. In *proceedings of the IEEE International Conference on Computer Vision*, pages 5533–5541, 2017. → pages 2, 16
- [31] D. Raposo, A. Santoro, D. Barrett, R. Pascanu, T. Lillicrap, and P. Battaglia. Discovering objects and their relations from entangled scene representations. arXiv preprint arXiv:1702.05068, 2017. → pages 3, 17
- [32] 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. → pages 2, 12, 28, 32
- [33] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In Advances in Neural Information Processing Systems, pages 91–99, 2015. → pages 16, 40
- [34] A. Rohrbach, M. Rohrbach, N. Tandon, and B. Schiele. A dataset for movie description. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3202–3212, 2015. → page 2
- [35] A. Rohrbach, A. Torabi, M. Rohrbach, N. Tandon, C. Pal, H. Larochelle,
   A. Courville, and B. Schiele. Movie description. *International Journal of Computer Vision*, 123(1):94–120, 2017. → page 7
- [36] 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. → page 20

- [37] S. Sadanand and J. J. Corso. Action bank: A high-level representation of activity in video. In 2012 IEEE Conference on Computer Vision and Pattern Recognition, pages 1234–1241. IEEE, 2012. → page 16
- [38] 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 4967–4976, 2017. → pages 3, 4, 12, 17, 18, 21, 22
- [39] F. Scarselli, M. Gori, A. C. Tsoi, M. Hagenbuchner, and G. Monfardini. The graph neural network model. *IEEE Transactions on Neural Networks*, 20(1): 61–80, 2009. → page 16
- [40] Z. Shou, D. Wang, and S.-F. Chang. Temporal action localization in untrimmed videos via multi-stage cnns. In *Proceedings of the IEEE* conference on Computer Vision and Pattern Recognition, pages 1049–1058, 2016. → pages 1, 15, 16
- [41] 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. → pages ix, 2, 3, 4, 7, 8, 14, 15, 25, 29, 30, 31, 33, 34
- [42] Q. Sun and Y. Fu. Stacked self-attention networks for visual question answering. In *Proceedings of the 2019 on International Conference on Multimedia Retrieval*, pages 207–211. ACM, 2019. → page 12
- [43] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4489–4497, 2015. → pages 1, 7, 9, 15, 16, 19, 20, 21, 23
- [44] D. Tran, H. Wang, L. Torresani, J. Ray, Y. LeCun, and M. Paluri. A closer look at spatiotemporal convolutions for action recognition. In *Proceedings* of the IEEE conference on Computer Vision and Pattern Recognition, pages 6450–6459, 2018. → page 16
- [45] H. Wang and C. Schmid. Action recognition with improved trajectories. In Proceedings of the IEEE International Conference on Computer Vision, pages 3551–3558, 2013. → page 16
- [46] X. Wang and A. Gupta. Videos as space-time region graphs. In Proceedings of the European Conference on Computer Vision, pages 399–417, 2018. → pages 2, 16, 23

- [47] P. Weinzaepfel, Z. Harchaoui, and C. Schmid. Learning to track for spatio-temporal action localization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3164–3172, 2015. → pages 1, 2, 15
- [48] D. Xu, Z. Zhao, J. Xiao, F. Wu, H. Zhang, X. He, and Y. Zhuang. Video question answering via gradually refined attention over appearance and motion. In *Proceedings of the 25th ACM International Conference on Multimedia*, pages 1645–1653. ACM, 2017. → pages 2, 13, 16
- [49] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International Conference on Machine Learning*, pages 2048–2057, 2015. → page 12
- [50] 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 1031–1042, 2018. → page 12
- [51] 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. → pages 1, 2, 13, 15, 16
- [52] Z. Zhao, Q. Yang, D. Cai, X. He, and Y. Zhuang. Video question answering via hierarchical spatio-temporal attention networks. In *International Joint Conference on Artificial Intelligence*, pages 3518–3524, 2017. → page 14
- [53] Z. Zhao, X. Jiang, D. Cai, J. Xiao, X. He, and S. Pu. Multi-turn video question answering via multi-stream hierarchical attention context network. In *IJCAI*, pages 3690–3696, 2018. → page 14
- [54] B. Zhou, A. Andonian, A. Oliva, and A. Torralba. Temporal relational reasoning in videos. In *Proceedings of the European Conference on Computer Vision*, pages 803–818, 2018. → pages 2, 4, 7, 18, 21, 22, 23, 25, 26, 31
- [55] L. Zhou, Y. Zhou, J. J. Corso, R. Socher, and C. Xiong. End-to-end dense video captioning with masked transformer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8739–8748, 2018. → page 15