Journal of Advanced Engineering Research
ISSN: 2393-8447
Volume 10, Issue 2, 2023, pp.214-218

Review of Reasoning Methods for Video Question Answering

Archana K S

Department of Computer Science And Engineering, Govt. Engineering College, Thrissur Kerala-680009, India

email: archanasathya121@gmail.com

ABSTRACT

Now a days in the field of Computer Vision there is an emerging challenging and complex task named as Video Question Answering (VQA). VQA retrieve spatial and temporal information from the video and then it will be interpreted. The methods that are used to implement VQA have been extended from the methods of Image-QA. The main difference between Image QA and VQA is that VQA have to process both the motion information as well as the appearance information. Therefore multi-step reasoning is used for VQA. There are mainly four category of reasoning methods are there. The first category of reasoning implements spatial and temporal attention mechanism. This will iteratively select useful information for answering questions. But the problem is that they do not consider relationship between different objects. The second category of reasoning methods are memory based networks that are previously used in Text QA and Image QA. But when they perform multistep reasoning neglect the visual relation information. The third category perform relational reasoning through relational network. But they can only model a limited number of objects. The fourth group of reasoning methods are based on GNN. GNN is more powerful and flexible than relational reasoning. So more researches are coming in this field. This survey reviews a number of methods that belongs to different reasoning methods. These methods shows recent advancement in VQA.

Keywords - Video Question Answering (VQA), Graph Neural Network (GNN)

1. INTRODUCTION

In the field of computer vision Video Question Answering (VQA) is a challenging task. There exist several visual information retrieval techniques. Video captioning/Description and video guided machine translation are some of the other related works that preceded the task of VQA. Out of them VQA is more challenging task. Because it requires an understanding of visual data based on the question asked in natural language.

During the task of VQA the agent have to understand videos first and then have to perform relational reasoning. The relational reasoning is performed according to question's visual, textual as well as spatio-temporal contents. The visual information we face in the real world is either in the form of image or videos. So the task of VQA will help to extract important information for building real life applications. VQA retrieve temporal and spatial information from the input video scene and then interpret it.

There exist many categories of Question answers. The task of VQA is classified in to 2 subtasks: multiple-choice QA and open-ended QA. Open-ended QA is more challenging, because it does not have fixed answer set. Another problem is that it is necessary to have an in-depth understanding of the video scene to produce correct answer from a pool of possible answers. Since it helps to understand complex scene and generate relevant information without the constraints choice, Open-ended QA has real world applicability.

Most of the models for VQA have been created by extending the methods of Image QA. But VQA is more challenging compared to Image QA, because it expects spatial as well as temporal mapping and requires complex object interactions in the video scenes. Earlier approaches for constructing models of VQA only considered the spatial information from the video. Later introduced approach for extracting both spatial and temporal information from video.

To solve the above challenges now we use multi-step reasoning. Previous reasoning based methods are classified in to 4 categories. The first one implements spatial and temporal attention mechanism to iteratively select useful information to answer the given question. The second group focuses on memory based network.

The third group aims to perform relational reasoning through simple modules like the relation network. The fourth category is based on the graph neural network.

This survey reviews a number of methods that belongs to different categories of reasoning. It includes the main working of different methods for solving the task of VQA, their advantages, disadvantages and contribution of each work is discussed here.

2. DIFFERENT REASONING METHODOLOGIES

Video QA is challenging and high-level multimedia task. It requires the agents to understand videos. And then perform relational reasoning according to questions also based on visual, textual as well as spatial-temporal contents. Most of the VQA methods are developed from the Image QA methods. But there is difference between Image QA and Video QA is that, VQA have motion information in addition with appearance information. Also VQA needs to have spatial-temporal reasoning over object. Since VQA method different from image QA it have to perform multi-step reasoning over objects. We can categories the reasoning methods in to 4 groups:

- Spatio-temporal Reasoning
- Memory Network based Reasoning
- Relational Network Reasoning
- Graph Neural Network based Reasoning

2.1 Spatio-Temporal Reasoning

This method provides a joint reasoning of spatial and temporal structures of video for solving the task of VQA. The special structure will give the information about in frame actions. Temporal structure will analyses the sequence of actions taking place in the given video.

2.1.1 Spatio-temporal Relation Network (STRN)

First we have to understand about the Relational reasoning, that means the ability to reason about relationships among entities. During the task of VQA it is necessary to answer Complex questions. Spatio-Temporal Relation Network provides joint reasoning. This joint reasoning is performed over both spatial and temporal domains. The input consist of ordered temporal sequence of spatial frame descriptors. STRN models the interactions among objects and how they evolve over time.

2.1.2 Multi-step Reasoning

When we combine multiple logical operations for answering a given question during the task of VQA. This process is called as multi-step reasoning [1]. It is more challenging and difficult in comparison with single step reasoning. So there are several reasoning based method for solving the complex task of VQA. Few of them are discussed in this paper.

2.1.3 Dynamic Hierarchical Reinforced Network

Open-ended long-form video question answering is a difficult task. Here according to the given input question and long-form video content, an automatically generated natural language answer is predicted. The previous works mainly focus on short-form video question answering. This is because lack of modelling semantic representations from long-form video contents. The dynamic hierarchical reinforced network can be used for open-ended long-form video question answering. It has an encoder–decoder architecture. That is it consist of a dynamic hierarchical encoder and a reinforced decoder.

2.1.4 Long-form VQA Via Dynamic Hierarchical Reinforced Network

Here discussing the problem of open-ended long-form video question answering. And considering from the viewpoint of dynamic hierarchical reinforced encoder-decoder network learning [2]. First implement the dynamic hierarchical encoder which is used to segment long-form video contents. And then as per the given question, learn the jointly video semantic representations. After that develop the reinforced decoder network with a hierarchical attention mechanism. This will generate the natural language answer for open-ended long-form video question answer. There are several previous methods existing for short-form video question answering. It has input video of length few seconds. Since the videos in the internet become lengthier it become challenging task. But applying this methods inappropriately to long-form VQA, it lacks modelling semantic representations. This work deals with a spatio-temporal reasoning mechanism for that it introduces an encoder decoder architecture. It contribute a lot in the long-form VQA.

2.2 Memory Network based Reasoning

This is another type of reasoning method that focuses on memory based network. This kind of networks are quite popular in Text QA and Image QA. After the success of Dynamic Memory Network in Image QA, some work use memory-based network on Video QA tasks. The traditional models like RNN and LSTM are powerful sequence predictors. But the memory used by these methods are limited. And they perform badly on long-term dependency tasks. The main use for memory in VQA is that, it models the need to keep track of the past and future frames. The answers may require reference from multiple frames in time.

2.2.1 Motion-Appearance Co-memory Network

In order to answer a question sometimes we need to keep track of some things in our mind. Similarly some VQA questions also need to keep track of past and future frames to answer question. The answer may require reference from different frames. This [3] paper deals with a motion-appearance co-memory network for VQA which is based on a Dynamic Memory Network (DMN). This method is developed from the observation that, considering about processing the input question, it mainly deals with temporal reasoning. So the model needs to have more temporal reasoning. And another observation is that unlike Images video contains lots of frames that consist rich amount of information. It makes reasoning process more complicated. So here in this paper they model a co-memory mechanism to joint and model a motion-appearance information.

2.2.2 Motion-Appearance Co-memory Network

Design In this [3] work they first design a co-memory attention mechanism. It will jointly model motion and appearance information. After that build multi-level contextual facts by using a temporal convolutional and de-convolutional, that is used for solving VQA. Then created a method called Dynamic Fact Ensemble. This is used to dynamically produce temporal facts in each cycle of fact encoding.

So in this model first we will have an input video. After that the video will be converted to sequence of motion and appearance features by using a two stream models. Then these features are fed in to a temporal convolutional and de-convolutional neural network. Because this will be used to build multi-level contextual facts that have the same temporal resolution, but represent different contextual information. These contextual facts are further used as input to the memory network. The co-memory network hold 2 separate memory states. One for motion and other for the appearance information. Then jointly model them as a co-memory attention mechanism. It will take motion cues for appearance attention generation. Similarly appearance cues for motion attention generation. Based on the above, design Dynamic fact ensemble method that produce temporal facts dynamically at each cycle of encoding.

Most of the memory based methods are used in short-term video, and they perform well. But on complicated videos, it lacks depth hierarchical decomposition video semantics. It leads failure to deliver excellent performance on untrimmed long-term videos. Also long term video contains large amount of noise and small quality of relevant information. So it is important to remove redundancy and unwanted information from the video.

2.3 Relational Network

A Relation Network is a component of an artificial neural network. It has a structure that can be used to reason about relations among objects. Relation network can infer relations. RN are data efficient. RN operates on set of objects. They do not consider the order of the object.

2.3.1Hierarchical Conditional Relation Network

This [4] paper deals with a reasoning method that is based on relation network. This paper introduces a general purpose and reusable neural unit named as Conditional Relation Network (CRN). It acts as a building block to construct structures for reasoning over video. Using this CRN they create an architecture for VQA. This design supports m and multi-step reasoning high order relational.

By using CRN in VQA it provides robustness in visual reasoning. It supports iterative reasoning. HCRN works on longer length videos that has addition of extra layers. Flexibility of CRN allow it to be replicated and layered to form deep hierarchical conditional relation network (HCRN).Since CRN is a general purpose neural unit it can be used for reasoning task in other applications like TVQA, Movie QA and so on.

2.4 Graph based Reasoning

The interaction between objects and their properties, movement all the details are represented in the form of graph. So it become more easy to understand and represent the complicated relation and interaction between different objects.

2.4.1 Location-Aware Graph Convolutional Network

This [5] paper deals with a graph based reasoning method. The paper introduces a network called Location-Aware Graph Convolutional Network (L-GCN). L-GCN is used to model the interaction between objects related to given question. The content of the video is represented as a graph. It is important to consider the interactions between the objects in the given videos that makes answering the given question related to the movement of object will become easy. Graph based method like this contains irrelevant information too. That will cause uninformative noise in to relational reasoning.

2.4.2 Graph Neural Network based Reasoning

The fourth group of reasoning method uses graph neural networks. This will integrate the relation information into the framework. It considers GNN's powerful representation ability on relation modelling. GNN is a relation encoder. It captures the relationship between the objects. So it enables reasoning with rich relational information. In relational reasoning Graph neural network is more flexible and powerful.

The graph based reasoning methods give very promising results and they are less explored. Recently researchers doing in this field. The model which works on a graph based mechanism will contain a graph based representation of the video that is used to answer questions related to the spatial-temporal and contextual aspects of the whole video.

This is a 3 stage process. Initially the video will be converted to image frames. Then the mean absolute difference between pixel values of adjacent frames is computed. This value or it's change is used to detect key frames for further processing. The second phase consist of the detection of objects and dense captioning by application of the Regional-Convolutional Neural Network(R-CNN). Finally, each caption is converted to a scene graph using Natural Language Toolkit.

2.4.3Heterogeneous Graph Alignment

This [6] paper also consist of a graph based reasoning methodology. They build a deep Heterogeneous Graph Alignment Network (HGA). It is made on the question words and video shots. In video question answering usually the processing of video and question will be done separately and finally these two modalities are combined together using some fusion network. All these methods will use information of one modalities to boost the other. But most of the cases they neglect the co relations of both inter and intra modalities. Within the HGA network inter and intra-modality information can be aligned and interacted parallel over the heterogeneous graph.

This graph based method uses an undirected heterogeneous graph. The question word is considered as node. This is used to integrate correlations of both inter- and intra-modality. These graph based methods also contains irrelevant information to the input question.

3. EXPERIMENT RESULTS ON TGIF-QA

For the comparison of various methods that we discussed, here I will show the result of performance of each method corresponding to TGIF-QA data set. TGIF-QA is widely used benchmark dataset. TGIF-QA dataset consisting of animated GIF and question answer pairs around 72K and 165K respectively. For the evaluation considering following facts such as, Repetition count (Count), Repeating action (Action), State transition (Trans.) and Frame QA (Frame QA).

Repetition count (Count) in a video is the count of number of repetitions of an action. It is an Open-ended task. The next measure is repeating action (Action), which is the task to identify a repetitive action from 5 candidate answer. State transition (Trans.) is also a multi-choice question in which has 5 choices. It measures the transition of 2 states. The last one is Frame QA (Frame QA) is task of answer that can make from one single frame. And this task is categorized as a multi-classification problem, in which a correct answer will found from a dictionary.

TABLE 1: Table of Comparison: TGIF-QA Dataset

Model	Action	Trans.	Frame	Count
Co-Memory	68.2	74.3	51.5	4.10
HCRN	75.0	81.4	55.9	3.82
LGCN	74.3	81.1	56.3	4.95
HGA	75.4	81.0	55.1	4.09

5. CONCLUSION

In the field of Computer Vision VQA is a complex task. There are plenty of methods for implementing the task of

VQA. In this survey I included some of the recent methods that belongs to different categories of reasoning. Most of these methods are inspired from Image QA .The main difference is that VQA methods contains motion features in addition with appearance features. There are lot of improvements can be seen in different categories of reasoning methods for the task of VQA. So there is a lot of scope for future researches in the field of video question answering.

From studying about spatial-temporal reasoning based method, it can be observed that they consider less about the object relationship. It mainly focuses on the motion information in the video. For the case of memory based reasoning method it can be observed that they neglect visual relation information during multi-step reasoning. They mainly focuses on the memory element. From the case of relation network based reasoning, we can find that, at a time they can only process a limited number of objects. Finally the fourth category of reasoning came. The graph based reasoning methods are more powerful and flexible than previous reasoning methods. GNN captures the inter-object relations and enables reasoning with rich relational information.

REFERENCE

[1] Xiaomeng Song, Yucheng Shi, Xin Chen, and Yahong Han. *Explore multi-step reasoning in video question answering*. In Proc. ACM MM, pages 239–247, 2018.

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

[3] Jiyang Gao, Runzhou Ge, Kan Chen, and Ram Nevatia. *Motionappearance co-memory networks for video question answering*. In Proc. CVPR, pages 6576–6585, 2018.

[4] Thao Minh Le, Vuong Le, Svetha Venkatesh, and Truyen Tran. *Hierarchical conditional relation networks for video question answering*. In Proc. CVPR, pages 9972–9981, 2020.

[5] Deng Huang, Peihao Chen, Runhao Zeng, Qing Du, Mingkui Tan, and Chuang Gan. *Location aware graph convolutional networks for video question answering.* In Proc. AAAI, pages 11021–11028, 2020.

[6] Pin Jiang and Yahong Han. *Reasoning with heterogeneous graph alignment for video question answering*. In Proc. AAAI, pages 11109–11116, 2020.

[7] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: *A large video description dataset for bridging video and language*. In Proc. CVPR, pages 5288–5296, 2016.

[8] Dejing Xu, Zhou Zhao, Jun Xiao, Fei Wu, Hanwang Zhang, Xiangnan He, and Yueting Zhuang. *Video question answering via gradually refined attention over appearance and motion*. In Proc. ACM MM, pages 1645–1653, 2017.