# Video Story Question Answering with Character-Centric Scene Parsing and Question-Aware Temporal Attention

Anonymous ICCV submission

Paper ID ****

## Abstract

*With the exploding growth of videos, there is an increasing interests for automatic video understanding. Video Story Question Answering (VSQA) proves to be an effective way for benchmarking the comprehension ability of a model. Recent VSQA approaches merely extract visual features from the whole scene or detected objects in each frame. However, it is hard to claim a method really understands a video without considering the characters inside. Additionally, relations and actions acquired by scene parsing are indispensable in the comprehension of video stories. In this work, we incorporate character-centric scene parsing to assist the VSQA task. Our reasoning framework consists of two parts: the first part utilizes question-aware temporal attention to locate the corresponding frame intervals; the second part involves a cross-attention transformer for multiple stream fusion. We train and test our VSQA model on the recently released TVQA dataset, which is the largest VSQA dataset until now. The experiments show that all modules in our framework work collaboratively and significantly improve the overall performance. Ablation studies demonstrate that our scene parsing based framework is efficacious for deeper understanding of video semantics.*

## 1. Introduction

The explosive growth of videos calls for effective techniques to understand the rich visual and language contents within them. A convincing way to measure how well a model understands a video is to correctly answer relevant questions about it. The task of Video Story Question Answering usually takes three steps: 1) extract key features for multimodal contents; 2) fuse multimodal features after extracting them; 3) utilize the fused features to make right predictions. For the first step, current state-of-the-art methods [7, 11, 2, 5] mainly focus on global visual features on image level. Namely, they treat one or several frames as input and extract features that represent a holistic understanding of these frames. Consequently, a general comprehension of what happens in the frames is achieved, while small details are often missed due to the large granularity of those global features. Such details include individual objects, their relationships and attributes, and perhaps more importantly, the identities of people inside the video if human beings exist. These components are often critical for answering semantic questions such as "What instrument is Raj playing when Raj and Howard have their show?" (Ji: Maybe change the teaser figure to this example), where the instrument (object) that Raj (human identity) is playing (relationship) is the decisive factor of this video, and global features usually have very limited power to capture them. Such limitation motivates the need for a framework that focuses on visual clues at a finer level and provide richer knowledge of the images' scenes for the following steps. This is the main goal of our work. Furthermore, localizing the frame interval corresponding to the moment described by the question is another important ability for VSQA models. In fact, a input video clip usually contains lots of moments while only a small proportion of its frames are closely relevant to answer the question. Taken the same question as example, if the frames related to "when Raj and Howard

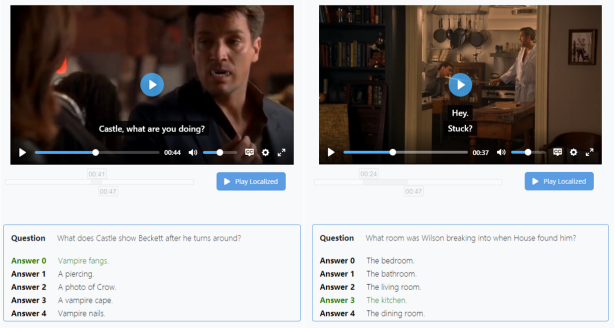


Figure 1. A simple illustration of **Video Story Question Answering** task: given a video clip with subtitles and a multiple choice question about the contextual information, the goal is to correctly predict the right answer.

have their show" are retrieved, the VSQA model can focus on these frames and get rid of the unrelated noisy frames.

(Ji: Modify this paragraph once the model is determined.) In this work, we build a VSQA framework with character-centric scene parsing and Question-Aware Temporal Attention

character-centric relationships and multi-modal attention flow. We train and test our model on the recently released large-scale video QA dataset - TVQA [1]. As Fig. 1 shows, the dataset offers a large number (21.8K) of video clips. In each video clip, there are corresponding subtitles and several multiple choice questions about the contextual information, the goal of our framework is to correctly predict the right answers of these questions.

The key contributions of the paper can be summarized as follows:

- We propose to conduct character-centric scene parsing for video story question answering. To the best of our knowledge, this is the first attempt to leverage for video story understanding researches.

## 2. Related Work

**Video Story Question Answering.** The Read Write Memory Networks (RWMN) [7] utilizes Compact Bilinear Pooling to fuse individual captions with corresponding frames and store them in memory slots. Multi-layered CNNs are used to represent adjacent slots in time. The Layered Memory Network (LMN) [11] learns a layered representation of movie content, which not only encodes the correspondence between words and visual content inside frames but also encodes the temporal alignment between sentences and frames inside movie clips. The Multimodal Dual Attention Memory (MDAM) [2] provides the dual attention structure that captures a high-level abstraction of the full video content by learning the latent variables of the video input (frames and captions). Late multimodal fusion is applied to get a joint representation.

**Scene Graph Parsing.** Scene graph parsing has recently emerged as a task that goes one step further from object detection towards holistic image semantic understanding[VRD, VG, Danfei, OpenImages]. The task is to first detect any visually related pair of objects and recognize the predicate that describes the relation, then build the scene graph by taking objects as nodes and their relations as edges. Most recent approaches achieve this goal by learning classifiers that predict relations based on different types of features of the object pairs [blablabla]. It is also demonstrated in recent works that scene graphs can provide rich knowledge of image semantics and help boost high-level tasks such as Image Captioning and Visual Question Answering[blablabla]. We are interested in how scene graphs can be utilized in videos, which to our best knowledge has not been explored.

**Character Naming.** Previous methods usually train their face assignment model with the whole dataset. However, it is impossible for people to recognize a character after watching all episodes. On the contrary, people are able to recognize the characters just with a short video clip.

**Sentence Localization in Videos.** TBA

## 3. The Proposed Method

In this section, we will introduce the data preparation for subtitles, questions, answers and video frames. Moreover, we will show the details of our multi-modal cross-attention flow model.

### 3.1. Method Overview

TBA

### 3.2. Data Preprocessing

**Tokenization and Vocabulary Building.** We first tokenize the subtitle sentences, option sentences and correct answers using Spacy. We then implements a vocabulary module to store the word ids, word tokens and pad tokens, with their corresponding embeddings. Besides, we have also included many methods, such as ids look-up, filter the less-frequency words etc so as we can interact with vocabularies with flexibility.

**Text Embedding.** Empirically, 300 dimension word vectors trained on 840-billion tokens gave us the best performances on most question answering tasks. Therefore, We use pre-trained 300 dimension word vectors from GloVe [8] to embed the words after tokenization. Here we transfer the word ids to their corresponding embedding vector after padding to fix size as input for neural network models. To alleviate out of words(OOV) issues, we have tried to evaluate different strategies. Currently, we used averaged character vectors of the people's name to represent the name appeared in the context.

**Video Feature Extraction.** For objects and actor-centric relations detection, we adopt the state-of-the-art relationship detection approach from [15]. The original model is pre-trained on Visual Genome [3] dataset. The task of visual relationship detection consists of object detection and relationship classifier. It means that we can run an object detector on the input image to obtain labels, boxes and visual features for subject and object, then use these as input features to the relationship classifier which only needs to output a label. Figure 3 gives a overview of the adopted relationship detection approach. However, some objects and relations detection results of original model are not related to the plots of the TV shows. So we can see that actor-centric relations is more important for the task of TV

[1] http://tvqa.cs.unc.edu/

show understanding. Hence we only consider the action classes included in the recently released actor-centric AVA dataset [1]. We retrain our relation detection model on the basis of AVA and then use it to extract visual concept features which serves as one stream of our final model. Figure 2 shows some examples of actor-centric relations and actor-unrelated relations.


Figure 2. In this work, we focus on actor-centric relations. The three images with green borderlines are actor-centric relations, while the right-down image with red borderline is an example of actor-unrelated relations.

### 3.3. Character-Centric Scene Parsing

#### 3.3.1 Character Naming with Partially Supervision

TBA

#### 3.3.2 Hierarchical Scene Parsing

TBA

### 3.4. Question-Aware Temporal Attention

TBA

### 3.5. Multimodal Temporal Attentional Transformer Network

Originated from the research of human visual system, attention mechanism has been widely adopted in many tasks, such as image caption generation and machine translation [10]. The advantage of attention processing is that it enables selectively focus on salient parts rather than the whole scene. In textual question answering task, many models introduce the attention mechanism to better align the passage and question to get better knowledge of their relationship between the representations of the question and document [9, 14, 6, 12].

For our model, after the step of data processing and video feature extraction, we first encode each input contextual stream $C$ (dialogue, objects, or actor-centric relation) and the corresponding question-answers pair $Q$, $A_0$, $A_1$, $A_2$, $A_3$, $A_4$ using depthwise separable convolution layers. Afterwards, our model enables cross-attention flow to generate context-aware-question representation $M$ between question-answers pair and the multi-modal input streams:

$$M_{\mathrm{Q}} = f(C, Q), \;\; M_{\mathrm{A_i}} = f(C, A_i) \tag{1}$$

Finally, the intermediate embeddings for different input streams are added together and be fed into a softmax layer to make the final selection.

$$\begin{aligned} H_{\mathrm{A_i}} &= [C; M_{\mathrm{Q}}; M_{\mathrm{Ai}}; C \odot M_{\mathrm{Q}}; C \odot M_{\mathrm{Ai}}] \\ p &= \mathrm{softmax}(H_{\mathrm{A_0}} \oplus H_{\mathrm{A_1}} \oplus H_{\mathrm{A_2}} \oplus H_{\mathrm{A_3}} \oplus H_{\mathrm{A_4}}) \end{aligned} \tag{2}$$

Figure 4 shows the whole framework of our multi-modal cross-attention flow model.

## 4. Experiments

**TVQA dataset.** The recently released TVQA dataset [4] is a large-scale video question answering dataset based on 6 popular TV shows. It consists of 152.5K QA pairs from 21.8K video clips, spanning over 460 hours of video. To encourage questions requiring both visual and language comprehension to answer, the questions are designed to be compositional in the format [What/How/ Where/Why/...] _____ [when/before/after] _____.

**More facts about TVQA.** (a) 152.5K questions: 84.8K What, 17.7K Who, 17.8K Where, 15.8K Why, 13.6K How. The type distribution of question-answer pairs is shown in Figure **??**. (b) 925 episodes from 6 TV shows: 3 situation comedies (Friends, The Big Bang Theory, How I Met Your Mother), 2 medical comedies (Grey's Anatomy, House M.D.), 1 crime comedy (Castle). (c) Each video clip is associated with 7 questions and a dialogue (consists of character names and subtitles).

### 4.1. Compared Methods

TBA

### 4.2. Evaluation and Discussion

In the training process, we set batch size to 64. The learning rate is set to 0.0002 and apply a step learning rate scheduler with step size 5. To best valuate our models, we employ answer retrieval accuracy (%) as metrics. Similar to the evaluation in [13] and other machine reading comprehension and dialogue systems, the accuracy of the in answer retrieval analysis is one of the dominant and straightforward evaluation strategy. Each question followed with five answer candidates has one of that considering as the

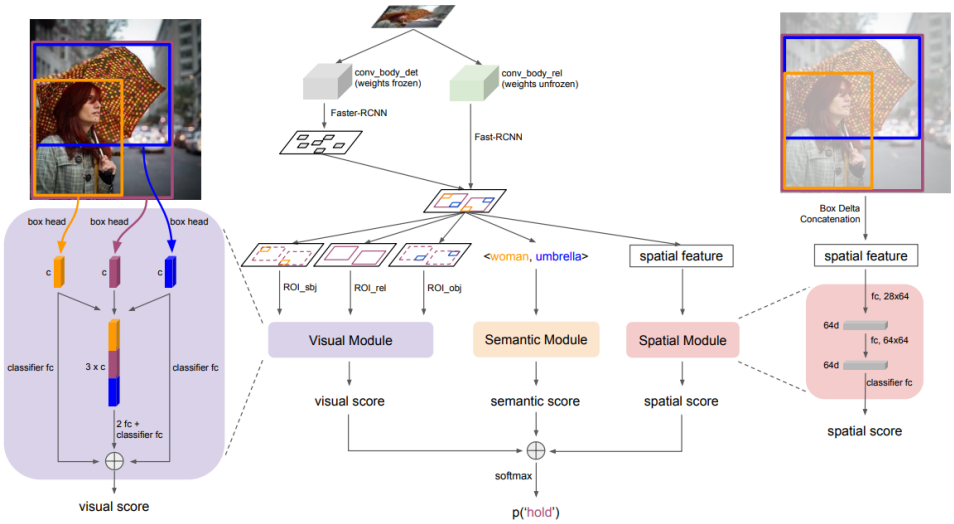


Figure 3. Overview of the adopted object and relation detection approach [15].

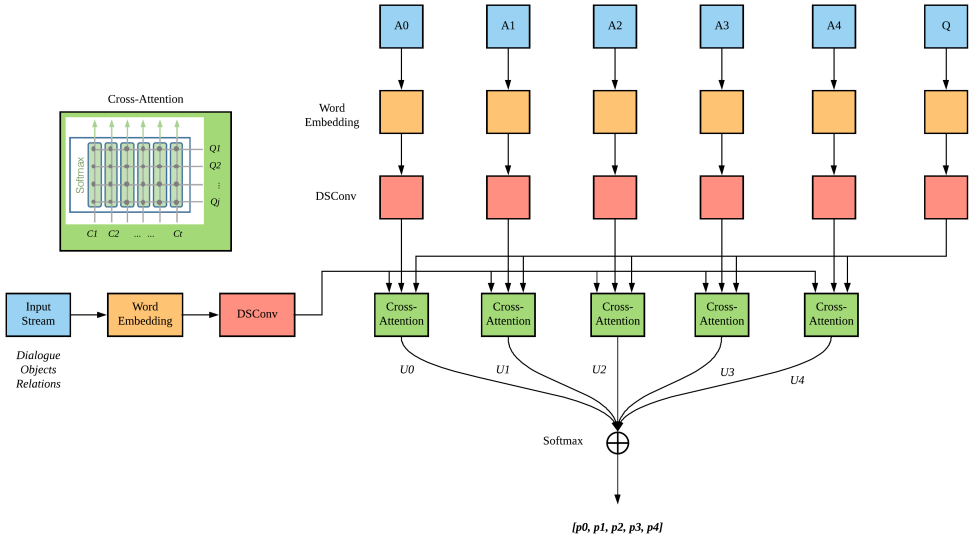


Figure 4. Our multi-modal cross-attention flow model.

only ground truth candidates. We will retrieve the most relevant one through our model as our choice.

As shown in Table **??**, dial w/ actor names outperforms dial w/o actor name largely. This means that the actor names in the dialogue play a key role in contextual inference and should be maintained. Another vital part would be relation caption. We follow the method from [15] with their object and relation detection results. By comparing Method 3 (D + O) and Method 4 (D + R), we can conclude that actor-centric relations embody more information of the plot than purely object concepts. With relation captions added to Method 3 (D + O), the accuracy of Method 5 (D + O + R) outperforms other baselines. It is easy to notice the necessity of adding actor-centric relations as one input stream, because it carries more contextual information.

### 4.3. Ablation Study

TBA

### 4.4. Case Study

TBA

## 5. Conclusion

In this project, we explore to apply actor-centric relations and object concepts to assist the video QA task. Our model

| Dataset | Test-Public w/ ts | | | | | | |
|---|---|---|---|---|---|---|---|
| Show | BBT | Friends | HIMYM | Grey | House | Castle | All |
| Multi-Stream[4] | 70.1892 | 65.6183 | 64.8148 | 68.2093 | 69.7010 | 69.7928 | 68.4770 |

Table 1. Results of TVQA test-public evaluation for models that used time-stamp annotation ('ts'). The performance of our method is also compared to other baselines on six TVQA sub-datasets: BBT, Friends, HIMYM, Grey, House, and Castle. All numbers in the table are percentage precision numbers (%).

| Dataset | Test-Public w/o ts | | | | | | |
|---|---|---|---|---|---|---|---|
| Show | BBT | Friends | HIMYM | Grey | House | Castle | All |
| Multi-Stream[4] | 70.2544 | 65.7783 | 64.0212 | 67.2032 | 66.8439 | 63.9586 | 66.4568 |
| JunyeongKim | 69.6021 | 65.9382 | 64.5503 | 68.2093 | 66.5116 | 66.6848 | 67.0471 |
| PAMN | 67.6451 | 63.5928 | 62.1693 | 67.6056 | 64.1860 | 63.1407 | 64.6071 |

Table 2. Results of TVQA test-public evaluation for models that did not use time-stamp annotation ('ts'). The performance of our method is also compared to other baselines on six TVQA sub-datasets: BBT, Friends, HIMYM, Grey, House, and Castle. All numbers in the table are percentage precision numbers (%).

is based on multi-modal cross-attention flow and is implemented with purely depth-wise convolution networks. In the experiments on TVQA dataset, our Full Dial + Objects + Relations model achieves the best 65.41% accuracy among all baselines, which proves the effectiveness of actor-centric as contextual information. In the future, we plan to increase explainability of the model by applying neural module network to do inference. To enrich available contextual information for question answering, we will extract plot graph both from video clips and dialogues. For building a plot graph with multimodal information, we plan to extend our work with dialogue-based actor naming and temporal causality arrangement.

## References

[1] C. Gu, C. Sun, D. A. Ross, C. Vondrick, C. Pantofaru, Y. Li, S. Vijayanarasimhan, G. Toderici, S. Ricco, R. Sukthankar, et al. Ava: A video dataset of spatio-temporally localized atomic visual actions. *CVPR*, 2018. 3

[2] K.-M. Kim, S.-H. Choi, J.-H. Kim, and B.-T. Zhang. Multimodal dual attention memory for video story question answering. *ECCV*, 2018. 1, 2

[3] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *IJCV*, 2017. 2

[4] J. Lei, L. Yu, M. Bansal, and T. L. Berg. Tvqa: Localized, compositional video question answering. In *EMNLP*, 2018. 3, 5

[5] X. Li, J. Song, L. Gao, X. Liu, W. Huang, X. He, and C. Gan. Beyond rnns: Positional self-attention with co-attention for video question answering. *AAAI*, 2019. 1

[6] R. Liu, W. Wei, W. Mao, and M. Chikina. Phase conductor on multi-layered attentions for machine comprehension. *arXiv*, 2017. 3

[7] S. Na, S. Lee, J. Kim, and G. Kim. A read-write memory network for movie story understanding. In *ICCV*, 2017. 1, 2

[8] J. Pennington, R. Socher, and C. Manning. Glove: Global vectors for word representation. In *EMNLP*, 2014. 2

[9] M. Seo, A. Kembhavi, A. Farhadi, and H. Hajishirzi. Bidirectional attention flow for machine comprehension. *ICLR*, 2017. 3

[10] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. In *NIPS*, 2017. 3

[11] B. Wang, Y. Xu, Y. Han, and R. Hong. Movie question answering: Remembering the textual cues for layered visual contents. *AAAI*, 2018. 1, 2

[12] C. Xiong, V. Zhong, and R. Socher. Dynamic coattention networks for question answering. *arXiv*, 2016. 3

[13] Z. Yan, N. Duan, J. Bao, P. Chen, M. Zhou, Z. Li, and J. Zhou. Docchat: An information retrieval approach for chatbot engines using unstructured documents. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 516–525, 2016. 3

[14] A. W. Yu, D. Dohan, M.-T. Luong, R. Zhao, K. Chen, M. Norouzi, and Q. V. Le. Qanet: Combining local convolution with global self-attention for reading comprehension. *ICLR*, 2018. 3

[15] J. Zhang, Y. Kalantidis, M. Rohrbach, M. Paluri, A. Elgammal, and M. Elhoseiny. Large-scale visual relationship understanding. *AAAI*, 2019. 2, 4