Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/patcog

Long video question answering: A Matching-guided Attention Model

Weining Wang[a,b], Yan Huang[a,b], Liang Wang[a,b,c,*]

[a] Center for Research on Intelligent Perception and Computing National Laboratory of Pattern Recognition Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China
[b] School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing 100049, China
[c] Center for Excellence in Brain Science and Intelligence Technology Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China

ARTICLE INFO

Article history:
Received 6 June 2019
Revised 29 December 2019
Accepted 26 January 2020
Available online 30 January 2020

Keywords:
Long video QA
Matching-guided attention

ABSTRACT

Existing video question answering methods answer given questions based on short video snippets. The underlying assumption is that the visual content indicating the ground truth answer ubiquitously exists in the snippet. It might be problematic for long video applications, since involving large numbers of answer-irrelevant snippets will dramatically degenerate the performance. To deal with this issue, we focus on a rarely investigated but practically important problem, namely long video QA, by predicting answers directly from long videos rather than manually pre-extracted short video snippets. We accordingly propose a Matching-guided Attention Model (MAM) which jointly extracts question-related video snippets and predicts answers in a unified framework. To localize questions accurately and efficiently, we calculate corresponding matching scores and boundary regression results for candidate video snippet proposals generated by sliding windows of limited granularity. Guided by the matching scores, the model pays different attention to the extracted video snippet proposals for each question. Finally, we use the attended visual features along with the question to predict the answer in a classification manner. A key obstacle to training our model is that publicly available video QA datasets only contain short videos especially designed for short video QA. Thus, we generate two new datasets for this task on the top of *TACoS Multi-level* dataset and *MSR-VTT* dataset by generating QA pairs from the video captions, called *TACoS-QA* and *MSR-VTT-QA*. Experimental results show the effectiveness of our proposed method on both datasets by comparing with two short video QA methods and a baseline method.

© 2020 Elsevier Ltd. All rights reserved.

1. Introduction

Visual question answering (VQA) [1] has drawn much attention recently, which is a difficult problem due to the joint modeling and interaction between vision and language. In this direction, image-based QA has been extensively studied and great progress has been achieved [2–13]. Video-based QA [14–20] is less studied even though it has various applications, *e.g.*, video-based chatting robot and language-driven video understanding.

Most of the existing video QA methods deal with just short video snippets rather than original long videos. They usually make an underlying assumption that the specific snippets related to desired answers are precisely pre-detected which makes them less applicable to many real-world tasks, *e.g.*, video surveillance and egocentric video analysis, in which long videos commonly appear and the answer-related video snippets cannot be easily pre-extracted. When roughly applying them to these tasks, the performance would largely degenerate since their used global visual features include large numbers of answer-irrelevant contents.

Different from them, we are interested in a more challenging and general problem in terms of long video QA, *i.e.*, answering questions directly from long videos without pre-knowing the locations of answer-related short snippets. The key of this problem lies in how to accurately detect target snippets based on given questions from long videos with varying lengths. It combines multiple vision sub-problems such as language understanding and action detection, which can also be regarded as a question-driven video detection problem.

To deal with this problem, we explore a Matching-guided Attention Model (MAM) to jointly extract question-driven answer-related video snippets and predict the final answers. We generate video snippet proposals for each long video and then find the most relevant video snippet for a given question in a matching-based manner. Although we could densely sample sliding windows at multiple scales, it is computationally infeasible for long videos and makes the matching task more difficult due to the extremely

[*] Corresponding author.
E-mail addresses: weining.wang@cripac.ia.ac.cn (W. Wang), yhuang@nlpr.ia.ac.cn (Y. Huang), wangliang@nlpr.ia.ac.cn (L. Wang).

https://doi.org/10.1016/j.patcog.2020.107248
0031-3203/© 2020 Elsevier Ltd. All rights reserved.

large variations of match space. Thus, we narrow down the match space by generating rough video snippet proposals with sliding windows in limited granularity. To refine them for more accurate answer prediction, we add location regression module on the top of the generated video proposals. After obtaining matching scores and location regression results of a given question, we firstly extract refined video snippets according to the regression results. Then, we use the matching scores as the attention weights to calculate a weighted visual feature vector for each question. Finally, the attended visual feature vector is combined with the sentence embedding to predict the final answer.

To the best of our knowledge, we are probably the first to study this problem and currently there is no such dataset for model training. Accordingly, we build two new datasets on the top of two video caption datasets, namely *TACoS Multi-Level* dataset and *MSR-VTT* dataset. *TACoS Multi-level* is designed for the cooking behavior, while *MSR-VTT* focuses on general videos in our life. We use a classic question generation method [3] to automatically convert the video captions into pairwise questions and answers. In this way, we can obtain a large number of QA pairs, as well as their associated long videos, as our experimental datasets. We perform extensive experiments on the extended datasets and demonstrate the effectiveness of the proposed method.

The main contributions of this work can be summarized as follows:

- We study a rarely investigated but practically important problem, namely long video QA, which can be suitably applied to many long video tasks.
- We propose a Matching-guided Attention Model (MAM) to deal with the long video QA problem, which jointly extracts question-related video snippets and predicts the answers based on attended visual features.
- We generate two new datasets (a simple one and a complex one) including long videos, QA pairs and corresponding temporal boundaries of QA pairs, which can be used for evaluating the study of the long video QA problem. Experimental results demonstrate the effectiveness of the proposed method.

2. Related work

Visual question answering (VQA). Deep learning has achieved a big breakthrough in computer vision [21–23]. With the successful applications of deep learning, image based VQA has been extensively studied in recent years. Some of the early works solve the VQA problem with Bayesian approaches. Malinowski et al. [24] set up the problem of VQA as a visual Turing Test. Kafle and Kanan [25] use a Bayesian framework to predict the answer type and achieve good performance by combining a discriminative work.

Following the successful application of soft attention [26–31], many VQA methods also utilize attention mechanism to selectively attend to parts of images or questions. Xu and Saenko [32] design an attention architecture which uses each word embedding to capture fine-grained alignment between the image and question. Yang et al. [4] exploit stacked attention networks which iteratively search for answer-related image regions within multiple steps of reasoning. Lu et al. [33] propose a bidirectional co-attention mechanism that simultaneously utilizes the question guided visual attention and a visual guided attention over the input question. Fang et al. [34] propose coherent dropout and siamese dropout mechanism to improve the performance of a visual spacial attention model. Liang et al. [11] introduce an end-to-end approach that uses a hierarchical process to dynamically determine what media and what time to focus on to answer the question. Lioutas et al. [35] use two separate word embedding models to increase the expressive power of the attention model. Yang et al. [36] utilize self-attention to find the most informative components of the question and use new question representation to guide visual attention of images. Yu et al. [13] present a Modular Co-Attention Network (MCAN) which consists of self-attention and guided-attention units to model the intra- and inter-modal interactions simultaneously.

Memory-augmented neural networks have also been developed. Xiong et al. [37] propose new modules on the top of the DMN (Dynamic Memory Networks) framework to tackle the VQA problem. Ma et al. [9] exploit memory-augmented neural networks to maintain the relatively long-term memory of scarce training exemplars, which can even predict accurate answers to visual questions occurring rarely in the training set.

Moreover, recent researches begin to incorporate external knowledge into VQA. Wu et al. [38] first introduce the Fact-based VQA (FVQA) task and build a new dataset, where questions require deeper reasoning with external knowledge. Narasimhan et al. [39,40] further study this problem and investigate fact retrieval based methods on FVQA. Marino et al. [41] introduce OK-VQA dataset with more diverse unstructured knowledge and propose a set of baselines that exploit unstructured knowledge.

Although image based VQA methods have achieved impressive progress, they are inadequate for the video QA due to the lack of modeling the temporal dynamics of video contents. In this work, we leverage the spatio-temporal information from videos by employing C3D [22] feature extraction to better understand the video data.

Video question answering (Video QA). Compared with the studies on image QA mentioned above, video QA is less studied. Zhu et al. [42] introduce the problem of video QA and present an RNN-based encoder-decoder approach to answer multiple-choice questions. Tapaswi et al. [15] introduce the problem of Movie QA where given questions can be answered by using multiple sources of information including full-length movies, subtitles, scripts and plots. Zeng et al. [14] extend several image QA approaches to video QA and introduce a new dataset. Maharaj et al. [43] present a fill-in-the-bank QA dataset and evaluate five different models on the dataset. Mun et al. [44] construct a dataset collected from Super Mario video gameplay and propose spatio-temporal attention models to conduct temporal event reasoning. Xue et al. [45] introduce the task of free-form open-ended video QA and propose an attention model to generate the answers. Jang et al. [16] establish a dataset for video QA named TGIF-QA containing short video snippets and propose a dual-LSTM based approach. Ye et al. [17] propose to use the frame-level attributes for video QA. Gao et al. [18] introduce the memory network to video QA, which utilizes both motion and appearance information. Xue et al. [20] propose a heterogeneous tree-structured memory network for video QA. Yu et al. [19] exploit a joint sequence fusion model to measure hierarchical semantic similarity between two multimodal sequence data. Fan et al. [46] propose a heterogeneous memory network with a new multimodal fusion layer which can better understand complex questions and attend to salient visual hints. Zhao et al. [47] study the multi-turn video QA task with a multi-stream hierarchical attention context reinforced network. Li et al. [48] utilize self-attention and co-attention mechanism replacing RNNs in the propose model which can better exploit the global dependencies of question and temporal information in the video.

However, these methods are generally designed to answer questions for short videos. Involving large numbers of answer-irrelevant snippets will dramatically degenerate their performance. Different from them, we focus on long video QA and propose a novel Matching-guided Attention Model to selectively attend to parts of

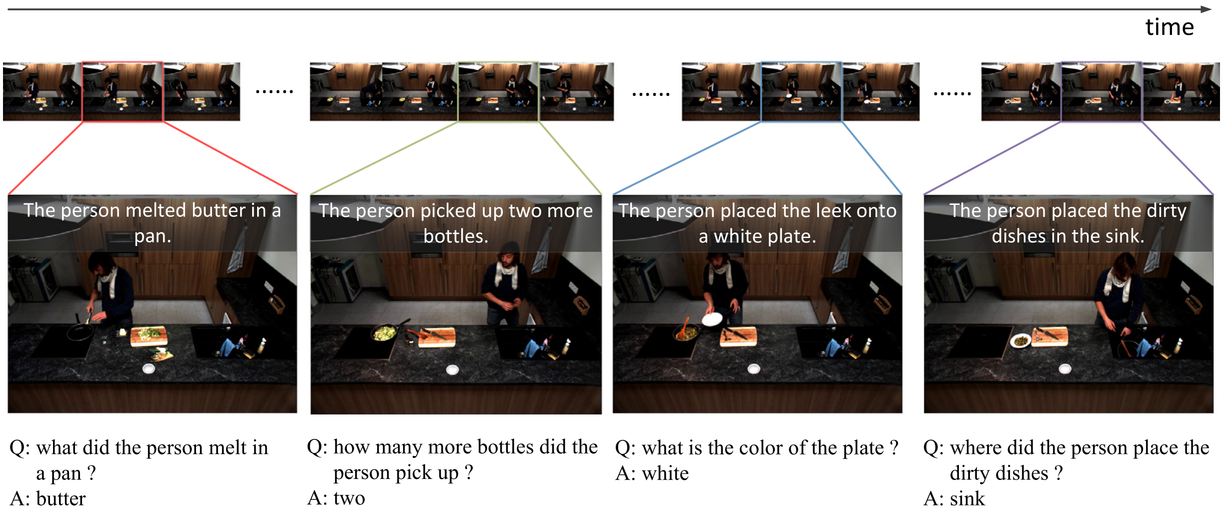

time
......
The person melted butter in a pan.
The person picked up two more bottles.
The person placed the leek onto a white plate.
The person placed the dirty dishes in the sink.
Q: what did the person melt in a pan ?
A: butter
Q: how many more bottles did the person pick up ?
A: two
Q: what is the color of the plate ?
A: white
Q: where did the person place the dirty dishes ?
A: sink

Fig. 1. The problem definition of long video QA. The top row illustrates a long video with example frames, the middle row shows four selected snippets and their associated textual descriptions and the bottom row presents the generated pairs of questions and answers.

the long videos and predict the answers based on the attended visual feature.

3. Problem description

The problem of long video QA is defined as follows: given a long video containing varying visual contents along with the time axis, as well as a question referring to certain content of a target video snippet, the goal is to answer the question without pre-knowing the location of the target snippet. As shown in Fig. 1, we take a long video in which a person is cooking food for example. The video consists of different snippets with various visual contents, *e.g.*, "*the person melted butter in a pan*" and "*the person placed the leek onto a white plate*". Two example questions can be raised according to these two snippets such as "*what did the person melt in a pan ?*" and "*what is the color of the plate ?*". The goal of long video QA is to generate the answers, respectively, *i.e.*, "*butter*" and "*white*".

Note that there is a major difference between this problem and the existing video QA works [14–18]. They deal with short video snippets that are manually pre-extracted from raw long videos, which exactly contain the contents corresponding to the desired answers. However, such an assumption does not always hold in real-world scenarios, because most videos in real applications are untrimmed and contain various video contents over large time span, especially in video surveillance. To address this problem, we propose long video QA in this work which directly handles raw long videos. In addition to the target answer-related snippet, there will be many answer-irreverent snippets, which makes the problem much more challenging.

4. Long video QA dataset

We cannot directly exploit the publicly available video QA datasets [14–16] for long video QA, since they only contain short videos especially designed for short video QA. To study the long video QA problem, we have to create new datasets. We do not create the datasets from scratch because collecting human generated QA pairs is very time-consuming. Inspired by [3], we extend two existing video caption datasets for long video QA, through generating pairwise questions and answers from the given video descriptions.

4.1. Pairwise QA generation

Although there are quite a few available video caption datasets, most of them cannot be used because they only provide short video snippets. Rohrbach et al. [49] selected a subset from the *MPII Cooking 2* dataset [50], totally 185 long videos, and collected a corpus named *TACoS Multi-Level* with about 20 triples of descriptions for each video, totally 52,593 descriptions. The lengths of videos range from several minutes to tens of minutes. The corpus provides the start and end frames of each description, which allows us to learn the relationship between video snippet and language description, as well as serving as the ground truth for performance evaluation. Therefore, we exploit *TACoS Multi-Level* dataset and *MPII Cooking 2* dataset to produce the desired *TACoS-QA* dataset.

Another video caption dataset *MSR-VTT* [51] provides 10K web video clips with 41.2 hours and 200K clip-sentence pairs in total. It covers many activity categories and contains diverse visual contents, which is richer than *TACoS Multi-Level* in terms of sentence and vocabulary. The corpus of this dataset provides the start and end time of each clip in the raw long video, as well as the link of the raw long video (the dataset does not provide the raw long video directly). Because many links of the raw long videos are dead, we finally download 3,852 raw long videos in total.

Another reason that we choose the above two datasets is that we would like to provide long video QA datasets with varying difficulty levels. *TACoS Multi-Level* dataset is relatively simple that the videos are designed for the cooking behavior and the background of videos is the same in a single scene. *MSR-VTT* dataset is relatively complex that the videos are collected from general videos in our life. The videos are collected in 20 representative categories (including cooking, sports, gaming, *etc.*).

We then use a commonly used question generation method [3] to generate pairs of questions and answers from those descriptions. This method is able to generate 4 kinds of questions: object question (what), number question (how many), color question (what color) and location question (where). It rejects the answers that appear too rarely or too often in the generated datasets. After the generation, we obtain a number of pairs of questions and answers. However, we find that among the video descriptions there are some repeated sentences, which causes replicated pairs of questions and answers for the same video snippet. Accordingly,

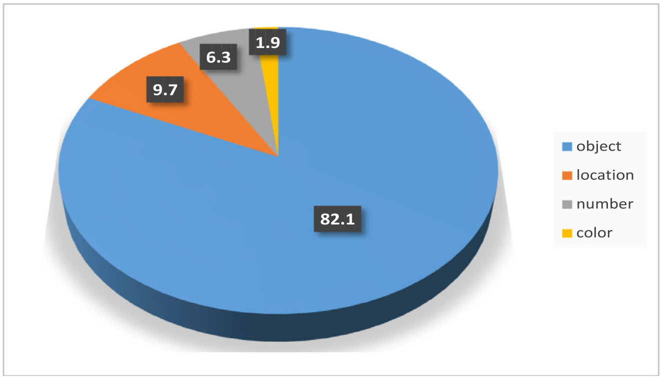


Fig. 2. Proportions for four types of questions in the *TACoS-QA* dataset.

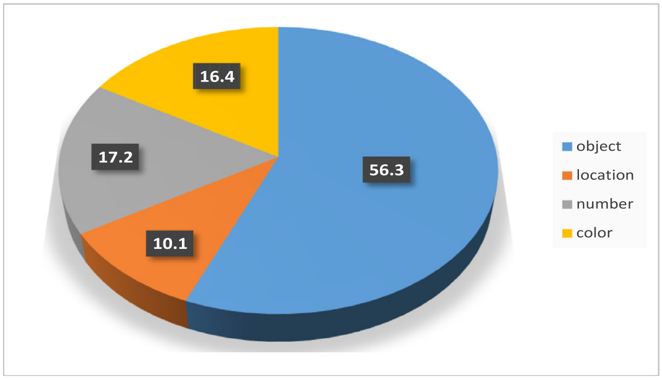


Fig. 3. Proportions for four types of questions in *MSR-VTT-QA* dataset.

we further refine the generated QA pairs based on the rule that the questions must be different for a single video snippet. After the above generation and refinement, we finally obtain our experimental datasets and denote them as *TACoS-QA* dataset and *MSR-VTT-QA* dataset, respectively. Examples of generated questions and answers are illustrated in Fig. 1. It should be noted that in *TACoS-QA* dataset there are several video snippets annotated with QA pairs in one long video, while in *MSR-VTT-QA* dataset there is only one video snippet annotated with QA pairs in one long video.

4.2. Dataset statistics

TACoS-QA dataset contains 185 long videos and 23,431 pairs of questions and answers in total. We split the dataset into a training set, a validation set and a testing set, which consists of 13,666 pairs of questions and answers with 120 videos, 1773 pairs of questions and answers with other 17 videos and 5871 pairs of questions and answers with the rest 48 videos, respectively. The answer set contains the most frequent 93 answers in the training dataset. At a rate of 29 frames per second (FPS), the maximum and minimum lengths of the videos are 40,658 and 1,392 frames, respectively. The mean and median lengths of the videos are 8,961 and 6,467 frames, respectively. For the generated questions, the maximum and minimum lengths are 8 and 4, respectively. Both the mean and median lengths are 8. The distribution for the 4 types of generated questions is shown in Fig. 2. We can see that the proportions of object question (what), location question (where), number question (how many) and color question (what color) are 82.1%, 9.7%, 6.3% and 1.9%, respectively. The reason for such a large proportion of object question is that the video contents are about cooking behavior so that there are many kinds of actions, tools, places, foods and ingredients in the descriptions.

MSR-VTT-QA dataset contains 3852 long videos and 19,748 pairs of questions and answers in total. We split the dataset into a training set, a validation set and a testing set, which consists of 11,830 pairs of questions and answers with 2430 videos, 1970 pairs of questions and answers with other 379 videos and 5948 pairs of questions and answers with the rest 1043 videos, respectively. The answer set contains the most frequent 100 answers in the training dataset. At a rate of 29 frames per second (FPS), the maximum and minimum lengths of the videos are 108,575 and 254 frames, respectively. The mean and median lengths of the videos are 10,441 and 7635 frames, respectively. The maximum and minimum lengths of questions are 21 and 4, respectively. The mean and median lengths are 8 and 7, respectively. The distribution for the 4 types of generated questions is shown in Fig. 3. We can see that the proportions of object question (what), location question (where), number question (how many) and color question (what color) are 56.3%, 10.1%, 17.2% and 16.4%, respectively.

Table 1
Statistics of *TACoS-QA* dataset.

	Train	Val	Test	Total
Num of videos	120	17	48	185
Num of QA pairs	13,666	1773	5871	21,310
Num of snippets	5433	751	2118	8302
Mean video length	95,559	583	3708	8983
Mean snippet length	484	172	432	443

Table 2
Statistics of *MSR-VTT-QA* dataset.

	Train	Val	Test	Total
Num of videos	2430	379	1043	3852
Num of QA pairs	11,830	1970	5948	19,748
Num of snippets	2430	379	1043	3852
Mean video length	10,148	10,017	11,166	10,441
Mean snippet length	464	485	472	469

Other statistics of the two generated datasets are summarized in Tables 1 and 2, respectively. From the tables we can find that the lengths of most videos are very long (*i.e.*, thousands of frames) containing very complex visual contents, while those ground truth snippets are very short (*i.e.*, several hundreds of frames). Thus, accurately localizing the answer-related video snippet is very challenging.

5. Matching-guided Attention Model

The problem of long video QA can be formulated as follows: given a question q and a long video $v = \{v_1, v_2, ..., v_N\}$, where $v_i (i = 1, 2, ..., N)$ is the ith video snippet proposal which can be obtained using a sliding window, we aim to predict the desired answer a. As shown in Fig. 4, we propose a Matching-guided Attention Model (MAM) which contains three modules: a) video and question embedding, b) question-related video content localization and c) answer prediction. In this section, we will describe the three modules one by one.

5.1. Video and question embedding

To discriminately represent the video snippets, we use a 3D convolutional neural network (C3D) [22] focusing on both appearance and motion information to extract features from the last fully-connected layer. For a video snippet proposal v_i, we first extract features for its every 16 consecutive frames with 8-frame overlap and then use mean pooling over all the features as its embedding. A fully connected layer with ReLU further embeds the visual feature to f_v. Note that we also compare and analyze other sampling frequencies during feature extraction in Section 6.5.

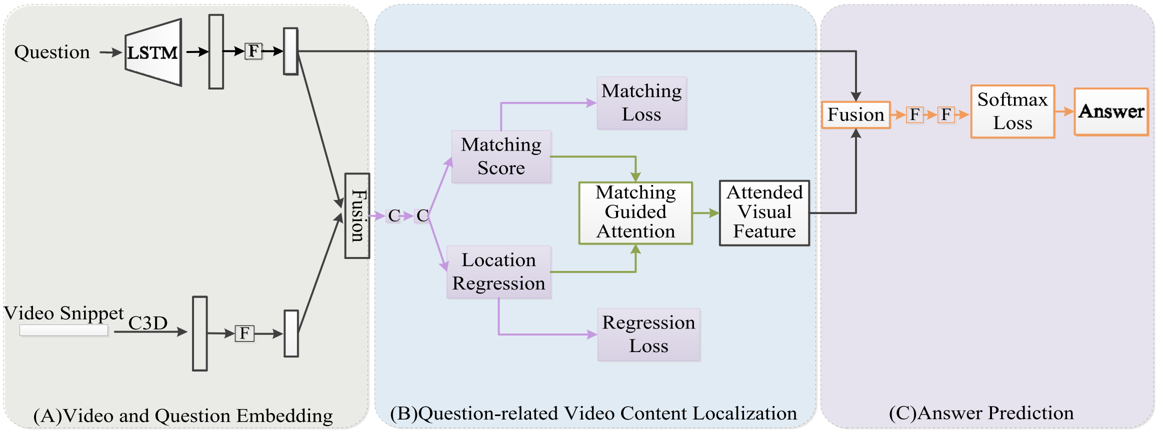


Fig. 4. An overview of Matching-guided Attention Model (MAM). The model contains three modules: a) video and question embedding, b) question-related video content localization and c) answer prediction. A matching loss is used to match related question and video snippet. A regression loss is used to regress the temporal boundary of the question-related video content in the video snippet proposal. A softmax loss is used to categorize the fused question and attended visual feature into an answer class.

For a given question $q = \{x_1, x_2, ..., x_T\}$ where x_t is the tth word, we adopt LSTM to model the sequential relations of words in the question:

$$\begin{aligned}
g^t &= \phi(W_g(Ex^t) + R_g h^{t-1} + b_g) \\
i^t &= \sigma(W_i(Ex^t) + R_i h^{t-1} + b_i) \\
f^t &= \sigma(W_f(Ex^t) + R_f h^{t-1} + b_f) \\
c^t &= g^t \odot i^t + c^{t-1} \odot f^t \\
o^t &= \sigma(W_o(Ex^t) + R_o h^{t-1} + b_o) \\
g^t &= \phi(c^t) \odot o^t
\end{aligned} \tag{1}$$

where c^t, x^t and h^t are the memory cell, input and hidden state at the tth timestep, respectively. E denotes a matrix of learned representations of words in the questions, which can be indexed by the one-hot vector x^t. W, R and b denote the input weights, recurrent weights and biases, respectively. g, i, f and o are subscripts indicating the input passway, input gate, forget gate and output gate, respectively. σ and ϕ refer to the sigmoid and hyperbolic tangent functions, respectively. $\odot$ means the dot product operation. At the last time step T, we regard the hidden state h_T followed by a fully connected layer with ReLU as the embedding of the input question, denoted as f_q.

The inputs of the first fusion module are the video representation f_v and the question representation f_q, which have the same dimension. Following [52] which achieves good performance on language-driven temporal activity localization, we use element-wise addition, element-wise multiplication and vector concatenation followed by a fully connected layer to fuse information from both modalities.

$$f_{vq} = (f_q \times f_v) \| (f_q + f_v) \| FC(f_q \| f_v) \tag{2}$$

The cross-modal representation f_{vq} is used as the input to the next module, namely question-related video content localization.

5.2. Question-related video content localization

Directly answering questions from original long videos is very difficult, since long videos contain various question-irrelevant contents. We propose to firstly extract question-related video snippet proposals for each question. To avoid generating video snippet proposals with a large number of scales, we utilize a matching score to evaluate the relation between each video snippet proposal and the question and then further adjust the temporal boundary of the video snippet proposal.

Taking the cross-modal feature f_{vq} as input, we use two 1 × 1 convolutional layers to generate the matching score denoted as s, which indicates the similarity between a question and a video snippet. The formulation is denoted as $s = \theta_s(f_{vq})$. To encourage matched snippet-question pairs having positive match scores and mismatched pairs having negative scores, we minimize the following objective:

$$\begin{aligned}
L_{match} = \frac{1}{N}\sum_{i=0}&[w_1 \log(1 + \exp(-s_{i,i})) \\
&+ \sum_{j=0, j\neq i}^{N} w_2 \log(1 + \exp(s_{i,j}))]
\end{aligned} \tag{3}$$

where N is the batch size, $s_{i,j}$ is the matching score between question q_j and video snippet v_i, w_1 and w_2 are the hyper parameters that control the balance between positive (matched) and negative (mismatched) snippet-question pairs, respectively.

It is straightforward to generate a large number of video snippet proposals with different scales and regard the most matching video snippet proposal as the question-related video snippet. However, generating video snippet proposals with a large number of scales is computationally expensive. Instead, we control the number of video snippet proposals by using sliding windows in limited granularity. Then, we perform location regression on the top of the generated video proposals to further adjust the temporal boundary.

Using the same network as the matching score network, the start time and end time are also generated after two 1 × 1 convolution layers, which is formulated as $(t_s, t_e) = \theta_{se}(f_{vq})$. We exploit an L_2 regression loss to optimize the temporal boundary as follows:

$$L_{reg} = \frac{1}{N}\sum_{i=0}^{N} \|(t_s^i, t_e^i) - (g_s^i, g_e^i)\| \tag{4}$$

where t_s^i and t_e^i denote the start time and end time of the i_{th} video snippet proposal respectively, g_s^i and g_e^i denote the ground truth start time and end time of question-related video snippet respectively and N is the batch size.

Due to the difficulty of question-driven video content localization, the precision of the top-1 detected video snippet is relatively low such that it is problematic to directly regard it as the final extracted video snippet. Thus, we propose a matching-guided attention mechanism, where the top-n video snippet proposals are selected and fused together weighted by the matching scores.

For each question, we select top-n video snippet proposals $\{v_1, v_2...v_n\}$ with top-n matching scores. We then extract question-related video snippets precisely from the top-n video snippet proposals according to the regression results $\{(t_s^1, t_e^1), (t_s^2, t_e^2)...(t_s^n, t_e^n)\}$. Consequently, we obtain n extracted video snippets $\{v_1^*, v_2^* ... v_n^*\}$ with matching scores $\{s_1, s_2 ... s_n\}$. The final visual representation corresponding to a question is calculated as:

$$v_{attend} = s_1 \cdot \theta_v(v_1^*) + s_2 \cdot \theta_v(v_2^*) + \ldots + s_N \cdot \theta_v(v_n^*) \quad (5)$$

where θ_v denotes the function of video embedding as illustrated in Section 5.1 and v_{attend} is the final attended visual feature vector.

5.3. Answer prediction

After obtaining the attended visual feature vector, we concatenate it with question embedding as the final cross-modal feature vector. We then adopt two fully connected layers with ReLU and a standard softmax layer to generate the answers in a classification manner.

$$\mathbf{a} = \mathrm{softmax}\left[W_a^\top (f_q \parallel v_{attend}) + \mathbf{b}_a\right] \quad (6)$$

where f_q is the embedding of question and v_{attend} is the attended visual feature. We use a cross-entropy loss for answer prediction as follows:

$$L_{ans} = -\sum_{i=1}^{N} a_i \log \hat{a}_i \quad (7)$$

where a_i is the predicted answer of the i_{th} question, $\hat{a}_i$ is the ground truth answer of the i_{th} question and N is the batch size.

5.4. Model learning

During model learning, to jointly perform question-related video content localization and answer prediction, we need to minimize the following objective:

$$L = \lambda_1 L_{match} + \lambda_2 L_{reg} + \lambda_3 L_{ans} \quad (8)$$

where L_{match}, L_{reg} and L_{ans} denote the matching loss, the regression loss and the cross-entropy loss, respectively. λ_1, λ_2, λ_3 are tuning parameters.

If we simply minimize the matching loss, regression loss and cross-entropy loss simultaneously, the model converges to some suboptimal solutions where the matching module and regression module perform badly. Here, we propose to train our model in two stages: 1) training the model with matching loss and regression loss together to produce accurate estimations for matching scores and regression results, and then, 2) combining the obtained attended visual feature vector and question embedding to predict the final answers with the cross-entropy loss. In the first stage, the parameters of the answer prediction module remain unchanged. In the second stage, we keep the parameters related to the first stage unchanged. We experimentally find such a two-stage training process can well address the multi-task problem which maintains the convergence of our model and reduces the computational time greatly.

As mentioned above, we first consider to use LSTM to encode the questions. However, the model performs not well on question-related video content localization and answer prediction. We find that for a given long video, there are a lot of identical predicted answers of different questions. The possible reason is that the scales of *TACoS-QA* and *MSR-VTT-QA* are not large enough to sufficiently train the LSTM encoder and the sequential information of sentences is not well captured. Hence, we use an off-the-shelf skip-thoughts [53] sentence embedding extractor to encode the questions, which is pre-trained over a large-scale book corpus producing more generic sentence representations.

6. Experiments

In this section, we will present the experimental configurations and results of the proposed method.

6.1. Implementation details

Data preparation. Before training, we use multi-scale temporal sliding windows with [64, 128, 256, 512] frames and 80% overlap to generate video snippet proposals. At testing stage, we only use sliding windows with [128, 256] frames for efficiency. We then use these video snippet proposals to collect training samples which are used as the input to our framework. For a video snippet proposal, we align it as a positive training sample if it satisfies two constraints as follows: 1) the IoU (intersection over union) of the sliding window snippet and the ground truth temporal interval is larger than 0.5. 2) the nIoL (non intersection over length) of the sliding window snippet and ground truth temporal interval is smaller than 0.2. We also collect negative samples which have no intersection with any question annotation.

Video representation. For each video snippet proposal in the training set, we extract features for every 16 consecutive frames with 8-frame overlap and perform mean pooling over them. At the end of video feature extraction, we add an *L*2-normalization. Due to the high FPS rate, we sample the frames with frequencies of 8, 16, 32 and 64, respectively. For example, under the sampling frequency of 8, we extract C3D features at the 8_{th}, 16_{th}, 24_{th}, ..., $(8n)_{th}$ frames of the video, respectively. We compare the impact of different sampling frequencies in Section 6.5.

Question embedding. The question embedding is performed using LSTM with a maximum length of 100, where we initialize the word embeddings using a continuous bag-of-words model [54] to compute 300 dimensional vectors. In the deep video-question embedding model, we encode both the questions and video snippets into 1024 dimensional vectors. Besides LSTM, we also use skip-thoughts [53] to encode the sentences, because skip-thoughts is trained on a large corpus of documents, which can produce generic sentence representations that are robust and perform well in cross-modal tasks.

Optimization. We use Adam [55] to train our network with a learning rate of 0.001. The batch size is set as 24. In Eq. (3) and (8), w_1, w_2, λ_1, λ_2 and λ_3 are all set as 1. During training, the temporal boundaries of the question-related video content within the video are given in the positive examples. In testing stage, the temporal information is only used for performance evaluation. We experimentally find that a fused representation of three video snippets is sufficient for answer prediction.

Evaluation metric. We compute Recall@*n* as the evaluation metric, which means that at least one of the top-*n* results is the right answer. The metric itself is on question level, so the overall performance is the average among all the questions. $Recall@n = \frac{1}{N}\sum_{i=1}^{N} r(n, q_i)$, where $r(n, q_i)$ is the recall for question q_i, N is the number of questions and Recall@*n* is the averaged performance.

6.2. A baseline method

Because there is no previous work focusing on long video QA in the literature, we further propose a baseline method for long video QA which is used to demonstrate the effectiveness of our method. We develop a two-step baseline method. 1) We first learn a deep video-question embedding model to map the question q and video snippets $\{v_1, v_2, ..., v_N\}$ into a common space, where the similarity score between the matched question-snippet is higher than mismatched ones. 2) After selecting the most related snippet of question through the deep video-question embedding model, we train a classifier on the concatenated representations of selected snippet and question to finally predict the desired answer.

We define the product of a snippet and a question $s(\tilde{v}, \tilde{q})$ as the similarity score $s(\tilde{v}, \tilde{q})$, in which $\tilde{v}$ and $\tilde{q}$ are scaled to have unit norm. In this way, the similarity function is equivalent to cosine similarity. Let θ denotes all the trainable parameters, the model

Table 3
Detection results of different methods on *TACoS-QA*.

Methods	IoU=0.1	IoU=0.2	IoU=0.3	IoU=0.4	IoU=0.5
Baseline method	6.38	5.33	4.37	3.93	3.17
MAM (LSTM)	6.78	5.42	3.66	1.64	1.29
MAM (skip-thoughts)	**14.48**	**10.99**	**7.95**	**5.91**	**3.95**

Table 4
Detection results of different methods on *MSR-VTT-QA*.

Methods	IoU=0.1	IoU=0.2	IoU=0.3	IoU=0.4	IoU=0.5
Baseline method	4.48	5.26	4.96	3.52	2.68
MAM (LSTM)	5.82	5.01	3.39	1.53	2.10
MAM (skip-thoughts)	**9.23**	**8.41**	**6.35**	**5.85**	**3.64**

can be learnt by optimizing a ranking loss as follows:

$$\min_{\theta} \sum_{\tilde{v}} \sum_{k} max\{0, \alpha - s(\tilde{v}, \tilde{q}) + s(\tilde{v}, \tilde{q}_k))\} + \sum_{\tilde{q}} \sum_{k} max\{0, \alpha - s(\tilde{q}, \tilde{v}) + s(\tilde{q}, \tilde{v}_k)\} \tag{9}$$

where $\tilde{v}$ and $\tilde{q}$ are matched video snippet and question, $\tilde{v}_k$ is the k-th negative video snippet to question $\tilde{q}$ and $\tilde{q}_k$ is the k-th negative question to video snippet $\tilde{v}$. The negative pairs are chosen randomly from the training set and re-sampled at each epoch during the training phase. α is a margin parameter.

After obtaining the deep video-question embedding model, we embed each question and corresponding video snippet proposals of the given video into a common space, *i.e.*, $\tilde{q}$, $\tilde{v}_1$, $\tilde{v}_2$,..., $\tilde{v}_N$. Based on these embeddings, we can compute the similarity scores of the question with all the proposals. We select one snippet $\tilde{v}^*$ with the highest score as the question-related video snippet. We then concatenate the embeddings of the selected snippet and the question together:

$$v_{JR} = \tilde{v}^* \parallel \tilde{q} \tag{10}$$

where $\parallel$ is the concatenation operation. By treating each answer in the training set as a class, we train a linear SVM classifier to classify the v_{JR} to the desired answer.

6.3. Question-driven video detection

In the testing phase, we extract the most relevant video snippet for each question, which can be regarded as the task of question-driven video detection. We adopt a similar metric used in [52] to compute "IoU=m", which means the percentage of the top-1 results whose IoUs are larger than m. The metric is also evaluated on sentence level. Thus, the overall performance is the average among all the questions.

We denote the Matching-guided Attention Model as MAM. LSTM means the questions are embedded with LSTM, while skip-thoughts means the questions are embedded with skip-thoughts. The detection results on two datasets are presented in Tables 3 and 4, respectively.

As shown in Tables 3 and 4, the MAM (skip-thoughts) model consistently outperforms the baseline method in terms of all the evaluation metrics by a large margin. The MAM (LSTM) model outperforms the baseline method at IoU =0.1 and 0.2, but it is inferior to baseline model at high IoUs. The possible reason is that the scales of the two generated datasets are not large enough to train the LSTM. Since skip-thoughts is trained on a large corpus of documents, it can produce more generic sentence representations which leads to better detection results.

Table 5
Comparison of different methods on *TACoS-QA*.

Methods	Recall@1	Recall@5	Recall@10
Random	1.08	5.38	10.75
HME-VideoQA	10.6	30.8	38.1
TGIF-QA	13.53	20.89	29.24
Baseline method	16.35	27.43	35.81
MAM (LSTM)	13.35	23.06	35.70
MAM (skip-thoughts)	**22.84**	**33.38**	**46.77**

Table 6
Comparison of different methods on *MSR-VTT-QA*.

Methods	Recall@1	Recall@5	Recall@10
Random	0.01	0.05	0.10
Baseline method	11.45	18.23	22.56
MAM (LSTM)	9.57	15.90	20.25
MAM (skip-thoughts)	**12.59**	**20.23**	**25.18**

6.4. Long video QA

We test the proposed Matching-guided Attention Model (MAM) and comparison methods on *TACoS-QA* and *MSR-VTT-QA* datasets and report Recall@{1, 5, 10}. The results are shown in Tables 5 and 6, where all methods use the same C3D features. 'Random' means that we randomly select n answers from the test set and evaluate Recall@n. MAM (LSTM) uses LSTM as the question encoder, while MAM (skip-thoughts) uses pre-trained skip-thoughts as the question encoder. We can see that our proposed matching-guided attention model consistently outperforms the baseline method by a large margin under all evaluation metrics on both two datasets, which demonstrates the effectiveness of our model on long video QA. The performance of the MAM (skip-thoughts) model outperforms the MAM (LSTM) model, because the scales of *TACoS-QA* and *MSR-VTT-QA* are not large enough to train the LSTM.

Besides, we can observe that the answer prediction results are consistent with the detection results shown in Tables 3 and 4, where better detection results lead to better answer prediction demonstrating the importance of question-related video content localization.

In addition, we compare our method with two classic models for short video QA including TGIF-QA [16] and HME-VideoQA [46] on *TACoS-QA* dataset. TGIF-QA [16] is a commonly used short video QA approach. HME-VideoQA [16] is the current state-of-the-art method for short video QA. The comparison results are shown in Table 5. As shown in Table 5, we can see that our MAM model outperforms both TGIF-QA and HME-VideoQA on all evaluation metrics by a large margin, which demonstrate the effectiveness of our method on long video QA. The reason that short QA methods perform not well is that they cannot attend to question-related

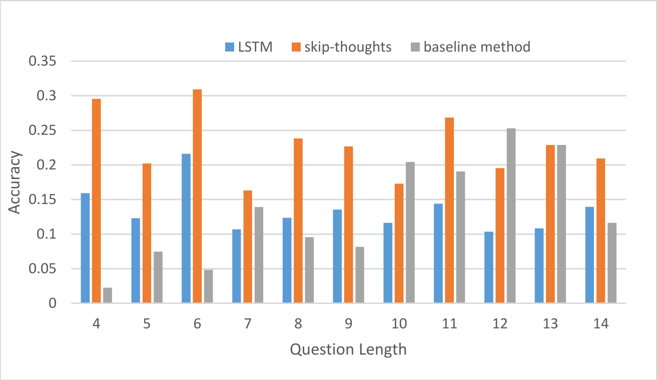


Fig. 5. The accuracy of questions with different lengths.

Table 7
Results with different model configurations.

Configurations		Accuracy		
Video snippet	Question	Recall@1	Recall@5	Recall@10
√	×	16.03	24.78	30.51
×	√	18.97	29.45	41.44
√	√	22.84	33.38	46.77

Table 8
Results corresponding to different types of questions under different sampling frequencies. *q*: question type, and *f*: sampling frequency.

f/*q*	Object	Number	Color	Location	Total
8	**14.85**%	**41.53**%	45.88%	**37.89**%	**24.82**%
16	14.50%	40.21%	**47.06**%	37.46%	22.13%
32	14.55%	30.69%	47.05%	34.79%	22.17%
64	14.50%	35.71%	45.88%	35.07%	23.88%

Table 9
Mean IoU (mIoU) of question-driven video content detection under different sample frequencies. *f*: sampling frequency.

mIoU/*f*	8	16	32	64
Baseline method	3.64%	6.58%	3.86%	**7.71**%
MAM (skip-thoughts)	5.57%	**8.04**%	5.88%	6.77%

video snippet in long videos. Although they use attention mechanism, it is not effective in long videos as the attention weights are not constrained explicitly. In our model, we deal with these challenges by using the location regression to precisely localize the question-related video snippets and use the proposed matching-guided attention mechanism to the localized video snippets. As illustrated in Tables 3 and 4, our method can improve the detection results significantly compared with the baseline model.

We report the accuracies of the questions with different lengths on *TACoS-QA* in Fig. 5. The lengths of questions vary from 4 to 14 words. We observe several points as follows: 1) the baseline method performs well when the lengths of questions are longer than 10, where the performance is even better than MAM (skip-thoughts) when the lengths are equal to 10 and 12. 2) the MAM (skip-thoughts) performs better than the MAM (LSTM) for both long questions and short questions.

To estimate the importance of video and question data in our model, we conduct experiments with two variational MAM(skip-thoughts) models. In the first model, we delete the question input and train the model with the rest architecture. Contrarily, we delete the visual input in the second model. The comparison results of different model configurations on *TACoS-QA* dataset are presented in Table 7.

As shown in Table 7, when the video snippet is deleted, the accuracy drops 4%. When the question input is removed, the accuracy drops 6%. Then, the importance of video and question input can be summarized from the accuracies, namely the question might play a more important role in the attention mechanism. The possible reason is that the model with only video snippet input is not able to attend to question-related video snippet, due to the various contents and long range of long videos. Therefore, it cannot output reasonable answers. However, questions are more distinct in semantic meanings so that they can lead to more reasonable answers.

6.5. *Impact of sampling frequencies*

We train the MAM (skip-thoughts) model with different sampling frequencies on *TACoS-QA* dataset and calculate Recall@1 of four types of questions as shown in Table 8. As we can see, the accuracies of color questions are the best, although the number of color questions is the least. In particular, there are totally 8 kinds of answers for color questions and an accuracy of more than 45% is achieved. Answering object questions is the hardest, due to the large potential answer space including various actions, tools, places, *etc*. Note that the accuracy of number questions drops a lot as the sampling frequency increases. It might result from the fact that using a high sampling frequency will lose much detailed and counting-related information. Similar observations can also be obtained in location questions.

The mean IoUs of question-driven video content detection under different sampling frequencies are shown in Table 9. From the table we can see that, except for the case of 32 and 16, as the sampling frequency becomes higher, the IoU becomes better. However, the overall mean IoUs are relatively low. It is mainly attributed to the fact that, the lengths of most videos are very long (*i.e.*, thousands of frames) while the ground truth snippets are very short (*i.e.*, several hundreds of frames), so precisely localizing these snippets is very difficult.

6.6. *Impact of the number of attended video snippets*

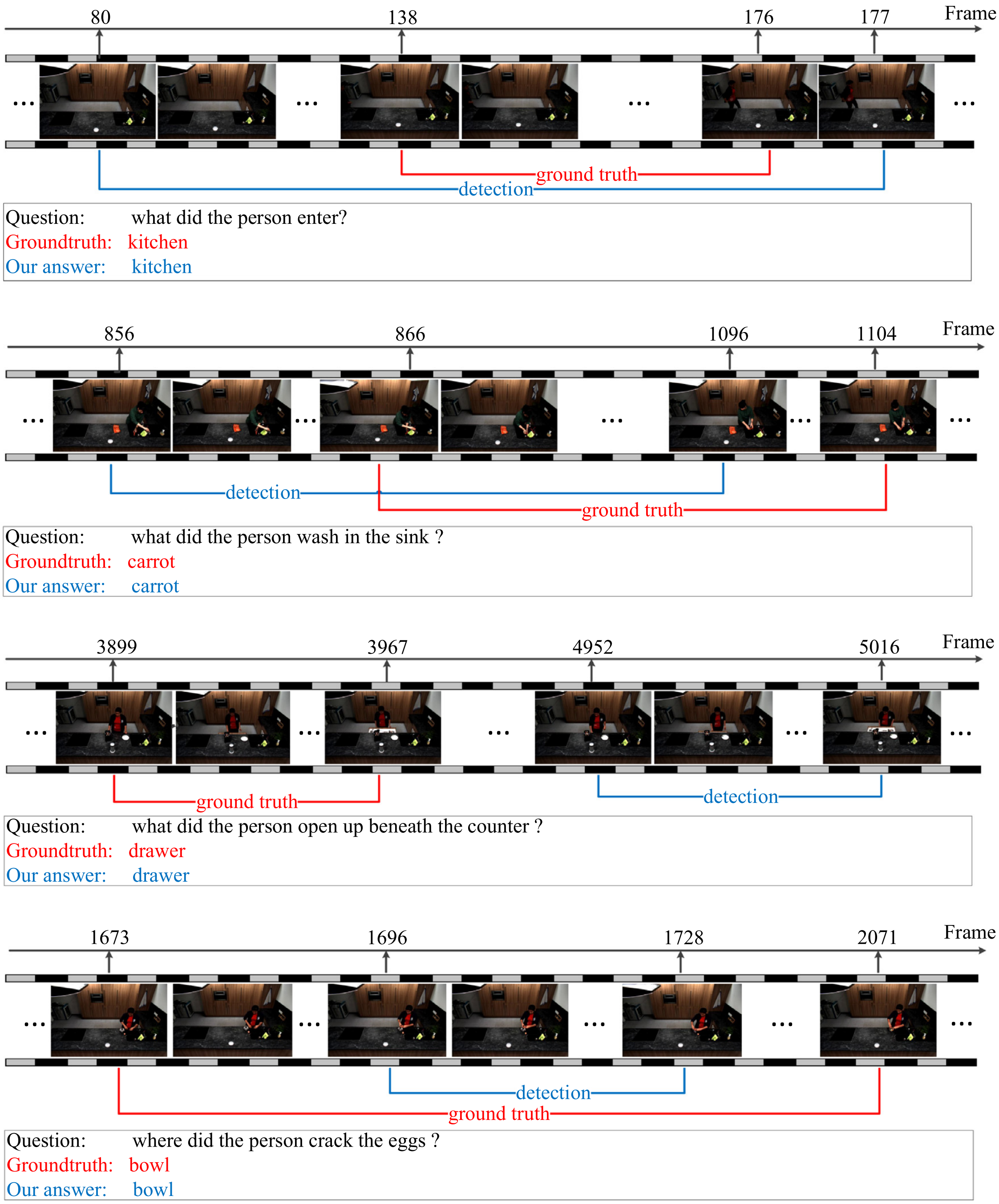
When training the proposed Matching-guided Attention Model with skip-thoughts on *TACoS-QA* dataset, we vary the number of attended snippets in the attention module from 1 to 5. For all the cases, we keep the sampling frequency as 8 and other hyperparameters unchanged. The corresponding results in Table 10 show that, when the number of attended snippets is 3, the model achieves the best performance. The possible reason is that when the model attends to more snippets the model obtains more data noise. When the number of attended snippets is less than 3, due to the low precision of detection, the attended visual feature vector may not contain the question-related video content.

6.7. *Qualitative evaluation*

We illustrate 4 examples of detected video snippets and predicted answers by our proposed method in Fig. 6. In the top row, the given question is "*what did the person enter*". The prediction is longer than the ground truth, as the prediction contains some frames before the person walking into the kitchen. In the second row, the key words in the question are "wash" and "sink". The model can find the relevant content but the start time and end time are not accurately equal to the ground truth. In the third row, the key information is "open up beneath the counter". The prediction of our model has no intersection with the ground truth. The prediction is reasonable due to the fact that this long video contains multiple snippets with the same content. In the bottom row,

Table 10
Results of the MAM (skip-thoughts) model using different numbers of attention snippets.

Number of attention snippets	1	2	3	4	5
Accuracy	23.15%	23.86%	**24.82%**	19.14%	17.15%



Question: what did the person enter?
Groundtruth: kitchen
Our answer: kitchen

Question: what did the person wash in the sink ?
Groundtruth: carrot
Our answer: carrot

Question: what did the person open up beneath the counter ?
Groundtruth: drawer
Our answer: drawer

Question: where did the person crack the eggs ?
Groundtruth: bowl
Our answer: bowl

Fig. 6. Examples of detected snippets and predicted answers by our proposed Matching-guided Attention Model on *TACoS-QA* dataset.

the question is about cracking eggs. The prediction is shorter than the ground truth.

By carefully analyzing these examples, we can obtain the following conclusions. 1) Different from the IoU measurement which is calculated in a strict way, the answer prediction does not require a very accurate video location detection but can still get the correct answer. It can also partially explain the inconsistent performance trends between the IoU in Table 9 and the answer accuracy in Table 8. 2) Each long video contains very diverse contents, so that our question-driven video detection method may reasonably detect the “wrong” snippet having the same or similar content with the ground truth snippet. For example, given a question “*what did the person open up beneath the counter ?*” and a long video containing several snippets with the same content of a person opening up the drawer, our proposed method may detect any snippet of them and output the correct answer "*drawer*".

7. Conclusions and future work

This paper has investigated a challenging but hardly studied problem, namely long video QA. Given a long video and a question, the goal of long video QA is to answer the question without pre-knowing the location of the target snippet. Two long-video QA datasets are built on the top of two video caption datasets, namely *TACoS-QA* dataset in cooking scene and *MSR-VTT-QA* dataset in diverse life scenes. Compared with existing video QA datasets, the proposed datasets have the following strengths. 1) Videos in the datasets are untrimmed long videos containing various visual contents, which enables the evaluation of long video QA. 2) The datasets provide the temporal boundaries of QA pairs in the long videos, such that the question-related video content can be localized in an accurate manner. 3) There are several different QA pairs in a given video which increases the difficulty of long video QA.

To deal with this problem, we have proposed a Matching-guided Attention Model (MAM). Extensive experimental results show the effectiveness of the proposed method by comparing with two short video QA methods and a baseline method. The proposed model localizes the question-related video content and answers the question simultaneously, which is efficient and end-to-end trainable. The proposed matching-guided attention module is much more effective than traditional attention mechanism, since a matching loss is used to supervise the attention weights (matching scores). The weakness of the proposed model may be that the linguistic knowledge of questions is not well employed. In the future, we will study how to use compositional linguistic structure in questions or use external knowledge-based methods for better answer prediction. Moreover, we will improve our model by incorporating more advanced components, *e.g.*, augmented memory modeling [32].

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work is jointly supported by National Key Research and Development Program of China (2016YFB1001000, 2018AAA0100402), National Natural Science Foundation of China (61525306, 61633021, 61721004, 61420106015, 61806194, U1803261, 61976132), Capital Science and Technology Leading Talent Training Project (Z181100006318030), HW2019SOW01, and CAS-AIR.

References

[1] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. Lawrence Zitnick, D. Parikh, Vqa: visual question answering, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 2425–2433.

[2] H. Gao, J. Mao, J. Zhou, Z. Huang, L. Wang, W. Xu, Are you talking to a machine? dataset and methods for multilingual image question, in: Advances in Neural Information Processing Systems, 2015, pp. 2296–2304.

[3] M. Ren, R. Kiros, R. Zemel, Exploring models and data for image question answering, in: Advances in Neural Information Processing Systems, 2015, pp. 2953–2961.

[4] Z. Yang, X. He, J. Gao, L. Deng, A. Smola, Stacked attention networks for image question answering, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 21–29.

[5] Y. Zhu, O. Groth, M. Bernstein, L. Fei-Fei, Visual7w: grounded question answering in images, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 4995–5004.

[6] Y. Li, N. Duan, B. Zhou, X. Chu, W. Ouyang, X. Wang, M. Zhou, Visual question generation as dual task of visual question answering, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 6116–6124.

[7] G. Peng, H. Li, H. You, Z. Jiang, P. Lu, S. Hoi, X. Wang, Dynamic fusion with intra-and inter-modality attention flow for visual question answering, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019.

[8] A. Agrawal, D. Batra, D. Parikh, A. Kembhavi, Don't just assume; look and answer: overcoming priors for visual question answering, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 4971–4980.

[9] C. Ma, C. Shen, A. Dick, Q. Wu, P. Wang, A. van den Hengel, I. Reid, Visual question answering with memory-augmented networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 6975–6984.

[10] Y. Goyal, T. Khot, D. Summers-Stay, D. Batra, 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, 2017, pp. 6904–6913.

[11] J. Liang, L. Jiang, L. Cao, L.-J. Li, A.G. Hauptmann, Focal visual-text attention for visual question answering, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 6135–6143.

[12] Q. Li, Q. Tao, S. Joty, J. Cai, J. Luo, Vqa-e: Explaining, elaborating, and enhancing your answers for visual questions, in: Proceedings of the European Conference on Computer Vision, 2018, pp. 552–567.

[13] Z. Yu, J. Yu, Y. Cui, D. Tao, Q. Tian, Deep modular co-attention networks for visual question answering, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 6281–6290.

[14] K.-H. Zeng, T.-H. Chen, C.-Y. Chuang, Y.-H. Liao, J.C. Niebles, M. Sun, Leveraging video descriptions to learn video question answering, in: Proceedings of the AAAI Conference on Artificial Intelligence, 2017, pp. 4334–4340.

[15] M. Tapaswi, Y. Zhu, R. Stiefelhagen, A. Torralba, R. Urtasun, S. Fidler, MovieQA: understanding stories in movies through question-answering, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 4631–4640.

[16] Y. Jang, Y. Song, Y. Yu, Y. Kim, G. Kim, Tgif-qa: Toward spatio-temporal reasoning in visual question answering, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 2758–2766.

[17] Y. Ye, Z. Zhao, Y. Li, L. Chen, J. Xiao, Y. Zhuang, Video question answering via attribute-augmented attention network learning, in: Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2017, pp. 829–832.

[18] J. Gao, R. Ge, K. Chen, R. Nevatia, Motion-appearance co-memory networks for video question answering, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 6576–6585.

[19] Y. Yu, J. Kim, G. Kim, A joint sequence fusion model for video question answering and retrieval, in: Proceedings of the European Conference on Computer Vision, 2018, pp. 471–487.

[20] H. Xue, W. Chu, Z. Zhao, D. Cai, A better way to attend: Attention with trees for video question answering, IEEE Trans. Image Process. 27 (11) (2018) 5563–5574.

[21] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, in: Advances in Neural Information Processing Systems, 2012, pp. 1097–1105.

[22] D. Tran, L. Bourdev, R. Fergus, L. Torresani, M. Paluri, Learning spatiotemporal features with 3d convolutional networks, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 4489–4497.

[23] Q. Li, Z. Sun, R. He, T. Tan, Deep supervised discrete hashing, in: Advances in Neural Information Processing Systems, 2017, pp. 2482–2491.

[24] M. Malinowski, M. Fritz, A multi-world approach to question answering about real-world scenes based on uncertain input, in: Advances in Neural Information Processing Systems, 2014, pp. 1682–1690.

[25] K. Kafle, C. Kanan, Answer-type prediction for visual question answering, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 4976–4984.

[26] J. Wang, W. Wang, L. Wang, Z. Wang, D.D. Feng, T. Tan, Learning visual relationship and context-aware attention for image captioning, Pattern Recognit. 98 (2020) 107075.

[27] X. Liu, J. Geng, H. Ling, Y. Cheung, Attention guided deep audio-face fusion for efficient speaker naming, Pattern Recognit. 88 (2019) 557–568.

[28] C. Luo, L. Jin, Z. Sun, Moran: A multi-object rectified attention network for scene text recognition, Pattern Recognit. 90 (2019) 109–118.

[29] J. Gu, G. Meng, S. Xiang, C. Pan, Blind image quality assessment via learnable attention-based pooling, Pattern Recognit. 91 (2019) 332–344.

[30] A.K. Bhunia, A. Konwer, A.K. Bhunia, A. Bhowmick, P.P. Roy, U. Pal, Script identification in natural scene image and video frames using an attention based convolutional-LSTM network, Pattern Recognit. 85 (2019) 172–184.

[31] S. Xie, H. Hu, Y. Wu, Deep multi-path convolutional neural network joint with salient region attention for facial expression recognition, Pattern Recognit. 92 (2019) 177–191.

[32] H. Xu, K. Saenko, Ask, attend and answer: Exploring question-guided spatial attention for visual question answering, in: European Conference on Computer Vision, 2016, pp. 451–466.

[33] J. Lu, J. Yang, D. Batra, D. Parikh, Hierarchical question-image co-attention for visual question answering, in: Advances In Neural Information Processing Systems, 2016, pp. 289–297.

[34] Z. Fang, J. Liu, Y. Li, Y. Qiao, H. Lu, Improving visual question answering using dropout and enhanced question encoder, Pattern Recognit. 90 (2019) 404–414.
[35] V. Lioutas, N. Passalis, A. Tefas, Explicit ensemble attention learning for improving visual question answering, Pattern Recognit. Lett. 111 (2018) 51–57.
[36] C. Yang, M. Jiang, B. Jiang, W. Zhou, K. Li, Co-attention network with question type for visual question answering, IEEE Access 7 (2019) 40771–40781.
[37] C. Xiong, S. Merity, R. Socher, Dynamic memory networks for visual and textual question answering, in: International Conference on Machine Learning, 2016, pp. 2397–2406.
[38] P. Wang, Q. Wu, C. Shen, A. Dick, A. van den Hengel, Fvqa: fact-based visual question answering, IEEE Trans. Pattern Anal. Mach.Intell. 40 (10) (2018) 2413–2427.
[39] M. Narasimhan, A.G. Schwing, Straight to the facts: learning knowledge base retrieval for factual visual question answering, in: Proceedings of the European Conference on Computer Vision, 2018, pp. 451–468.
[40] M. Narasimhan, S. Lazebnik, A. Schwing, Out of the box: reasoning with graph convolution nets for factual visual question answering, in: Advances in Neural Information Processing Systems, 2018, pp. 2654–2665.
[41] K. Marino, M. Rastegari, A. Farhadi, R. Mottaghi, Ok-vqa: a visual question answering benchmark requiring external knowledge, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 3195–3204.
[42] L. Zhu, Z. Xu, Y. Yang, A.G. Hauptmann, Uncovering the temporal context for video question answering, Int. J. Comput. Vis. 124 (3) (2017) 409–421.
[43] T. Maharaj, N. Ballas, A. Rohrbach, A. Courville, C. Pal, A dataset and exploration of models for understanding video data through fill-in-the-blank question-answering, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 6884–6893.
[44] J. Mun, P. Hongsuck Seo, I. Jung, B. Han, Marioqa: Answering questions by watching gameplay videos, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 2867–2875.
[45] H. Xue, Z. Zhao, D. Cai, Unifying the video and question attentions for open-ended video question answering, IEEE Trans. Image Process. 26 (12) (2017) 5656–5666.
[46] C. Fan, X. Zhang, S. Zhang, W. Wang, C. Zhang, H. Huang, Heterogeneous memory enhanced multimodal attention model for video question answering, in: Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 2000–2007.
[47] Z. Zhao, Z. Zhang, X. Jiang, D. Cai, Multi-turn video question answering via hierarchical attention context reinforced networks, IEEE Trans. Image Process. (2019).
[48] X. Li, J. Song, L. Gao, X. Liu, W. Huang, X. He, C. Gan, Beyond rnns: positional self-attention with co-attention for video question answering, in: Proceedings of the AAAI Conference on Artificial Intelligence, 33, 2019, pp. 8658–8665.
[49] A. Rohrbach, M. Rohrbach, W. Qiu, A. Friedrich, M. Pinkal, B. Schiele, Coherent multi-sentence video description with variable level of detail, in: German Conference on Pattern Recognition, 2014, pp. 184–195.
[50] M. Rohrbach, A. Rohrbach, M. Regneri, S. Amin, M. Andriluka, M. Pinkal, B. Schiele, Recognizing fine-grained and composite activities using hand-centric features and script data, Int. J. Comput. Vis. (2015) 1–28.
[51] J. Xu, T. Mei, T. Yao, Y. Rui, Msr-vtt: a large video description dataset for bridging video and language, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 5288–5296.
[52] J. Gao, C. Sun, Z. Yang, R. Nevatia, Tall: temporal activity localization via language query, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 5277–5285.
[53] R. Kiros, Y. Zhu, R.R. Salakhutdinov, R. Zemel, R. Urtasun, A. Torralba, S. Fidler, Skip-thought vectors, in: Advances in Neural Information Processing Systems, 2015, pp. 3294–3302.
[54] T. Mikolov, K. Chen, G. Corrado, J. Dean, Efficient estimation of word representations in vector space, in: International Conference on Learning Representations Workshop, 2013.
[55] D. Kingma, J. Ba, Adam: a method for stochastic optimization, arXiv:1412.6980 (2014).

Weining Wang received the B.E. degree in automation from North China Electric Power University, China, in 2015. She is now a PhD candidate working at the National Laboratory of Pattern Recognition (NLPR), Institute of Automation, Chinese Academy of Sciences (CASIA), Beijing, China. Her research interests include video question answering and video action detection.

Yan Huang received the BSc degree from the University of Electronic Science and Technology of China (UESTC), in 2012 and the PhD degree from the University of Chinese Academy of Sciences (UCAS), in 2017. Since July 2017, he has joined the National Laboratory of Pattern Recognition (NLPR), Institute of Automation, Chinese Academy of Sciences (CASIA) and he is an associate professor now. His research interests include machine learning and pattern recognition.

Liang Wang received both the BEng and MEng degrees from Anhui University, in 1997 and 2000, respectively, and the PhD degree from the Institute of Automation, Chinese Academy of Sciences (CASIA), in 2004. From 2004 to 2010, he was a research assistant at Imperial College London, United Kingdom, and Monash University, Australia, a research fellow with the University of Melbourne, Australia, and a lecturer with the University of Bath, United Kingdom, respectively. Currently, he is a full professor of the Hundred Talents Program at the National Lab of Pattern Recognition, CASIA. His major research interests include machine learning, pattern recognition, and computer vision. He is currently an IEEE Fellow and IAPR Fellow.