Multi-Question Learning for Visual Question Answering

Chenyi Lei,[1,2*] Lei Wu,[3*] Dong Liu,[1†] Zhao Li,[2] Guoxin Wang,[2] Haihong Tang,[2] Houqiang Li[1]
[1]University of Science and Technology of China
[2]Alibaba Group, [3]Zhejiang University
{chenyi.lcy, lizhao.lz}@alibaba-inc.com, shenhai1895@zju.edu.cn,
{dongeliu, lihq}@ustc.edu.cn, {xiaogong.wgx, piaoxue}@taobao.com

Abstract

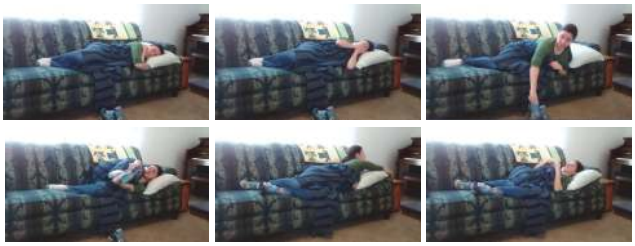
Visual Question Answering (VQA) raises a great challenge for computer vision and natural language processing communities. Most of the existing approaches consider video-question pairs individually during training. However, we observe that there are usually multiple (either sequentially generated or not) questions for the target video in a VQA task, and the questions themselves have abundant semantic relations. To explore these relations, we propose a new paradigm for VQA termed Multi-Question Learning (MQL). Inspired by the multi-task learning, MQL learns from multiple questions jointly together with their corresponding answers for a target video sequence. The learned representations of video-question pairs are then more general to be transferred for new questions. We further propose an effective VQA framework and design a training procedure for MQL, where the specifically designed attention network models the relation between input video and corresponding questions, enabling multiple video-question pairs to be co-trained. Experimental results on public datasets show the favorable performance of the proposed MQL-VQA framework compared to state-of-the-arts.

Introduction

Visual Question Answering (VQA), which is considered as one of the highest goals of artificial intelligence, has raised a great challenge for both computer vision and natural language processing communities. Specifically, VQA is a task about automatically returning the relevant answer to the given question with respect to the reference visual content. Currently, for ease of evaluation, most of the researches (including this paper) focus on selecting answers from given multiple choices. Producing answers in natural language from scratch is deemed a much more difficult task.

Most of the existing works considered the problem of VQA as a multimodal understanding task, that is to combine the visual cues based on image/video understanding with the question based on natural language understanding and reasoning. With the development of attention mechanism, many studies explored the attention-based fusion of image/video representations learned from deep convolutional neural network and question representations learned via the sequential model (Anderson et al. 2018; Nguyen and Okatani 2018; Yang et al. 2016; Yu et al. 2019; 2017b; Li et al. 2019; Gao et al. 2019b), which made significant progress in the last few years. For example, (Anderson et al. 2018) proposed to combine bottom-up and top-down attention mechanism that enables attention to be estimated at the level of objects and other salient image/video regions. Meanwhile, some recent works also considered the modeling of answer representations to enrich the multimodal understanding (Hu, Chao, and Sha 2018; Bai et al. 2018). For instance, (Bai et al. 2018) proposed a deep attention neural tensor network to fuse the information of image-question-answer triplets. Although the existing works have achieved promising performance in VQA, it is still not clear whether they do actually reason, or they reason in a human-comprehensible way (Bai et al. 2018; Cadene et al. 2019).

One issue that was superficially ignored in prior arts is about the multiple questions for one video sequence. Indeed, we observe that there are more and more VQA datasets (Jang et al. 2017; Zhu et al. 2016; Ranjay et al. 2016) containing



Q1: What is in the living room?
Q2: What is the color of the wall?
Q3: What is the person in the video doing?
Q4: Where is the pillow?
Q5: Is the person in the video sitting or lying?

Figure 1: An example of a video sequence with multiple questions. There are abundant relations between the multiple questions themselves, either explicitly *e.g.* between Q3 and Q5, or more implicitly *e.g.* between Q1 and Q4.

*Equal contribution. This work was done when Lei Wu was an intern at Alibaba.

†Dong Liu is the corresponding author.

multiple questions for a video sequence. This is also a common phenomenon in applications such as education for children (Calhoun 1999) and multi-turn video question answering (Zhao et al. 2018). It is worth noting that the semantic relational information among multiple video-question pairs plays an important role for human performing actually reasoning in VQA tasks. Actually, multiple questions complement information and provide clues for each other. Taking the question Q3 in Fig.1 as an example, existing VQA methods are subject to bias in Q3 itself to recognize the action of the person in the video sequence, with the underlying models probably not truly understanding the relevance of textual and visual concepts. The visual and textual concepts in Q1 and Q4 (*e.g*. "pillow", "living room" and corresponding potential answers) are strongly related to Q3 (*e.g*. "sleeping"). Furthermore, being able to better exploit the textual and visual concepts from the heavy tails of the question and answer distributions (*e.g*. the concept "sleeping" in Q3) would enable more accurate question answering. It is therefore of great importance to learn these heavy tails answers with related questions and visual information jointly. Nevertheless, almost all of the previous works considered and learned from each video-question pair independently.

The success of multi-task learning (Caruana 1997) implies that inductive transfer may improve generalization by using the domain information contained in the training signal of related tasks as an inductive bias. Analogous to multi-task learning, we propose a new modeling paradigm for VQA tasks termed Multi-Question Learning (MQL) mechanism to address the challenges mentioned above. Our main idea is to co-train multiple video-question pairs jointly to generate a generic video-question representation, which can produce better answers for the questions and can be transferred and utilized in new questions. The representation can extract general and effective features from complex visual and textual information source, which is learned across several related semantic concepts of different questions.

It is worth noting that our MQL is different from the VQA in a visual dialog scenario (Teney et al. 2018; Zhao et al. 2018), where for a new question, the previous questions and answers can be utilized as context. Instead, in our setting, we have multiple questions only without any answer, and we consider the questions in no order. In addition, even when there is only one question for testing, our trained MQL still works and performs well, which indicates that the improvements of trained MQL are not only from the information gain comparing to Single Question Learning (SQL) but also from the better understanding and reasoning capability studied by MQL mechanisms.

To fulfill the MQL, we propose an effective multi-head attention-based deep neural network for video question answering task. As noted video QA is a more difficult problem than image QA (Zhao et al. 2017). Our framework integrates the techniques of convolutional neural network (CNN), gated recurrent unit (GRU), attention mechanism, and multi-question supervision. Specifically, video representations are modeled by CNN; GRU as recurrent neural network (RNN) is used as the building block to learning desired representations from multiple questions. A multi-head attention network is designed on top of the GRU and CNN, which learns attention-weighed representations within questions, video frames and related video regions. Then, all of these attention-weighed representations are integrated as a general representation, which is made more robust and reliable by sharing representations and learning within the multi-question setting. Finally, the general representation is co-trained with different question-answer pairs jointly.

To summary, the main contributions of this paper are as follows:

- Different from the previous studies, we propose the Multi-Question Learning (MQL) mechanism as a new formulation of the VQA problem, which learns visual content with corresponding questions jointly to increase the capacity of modeling latent relation among textual and visual information. Our proposed MQL mechanism provides a new paradigm for VQA tasks.
- We propose effective multi-head attention-based deep learning framework for the video question answering task, in which the video and corresponding questions are co-trained jointly to capture the complex semantic relations among visual and textual information source.
- We validate the proposed MQL mechanism and the Video QA framework on two public datasets. Experimental results demonstrate that the proposed MQL mechanism and the Video QA framework achieve superior performance than the state-of-the-art Video QA methods significantly. Furthermore, we also perform several in-depth analyses to give insights into our proposed approach.

Related Work

In this section, we briefly review some related works about visual question answering and multi-task learning.

Visual Question Answering. The visual question answering task is to provide an accurate answer for the question given in natural language about the given visual content (Antol et al. 2015). Most of the previous works formulated the visual question answering task as a classification problem and solved it with a deep neural network that implements the joint representation of image and question (Gao et al. 2019a; Anderson et al. 2018; Nguyen and Okatani 2018; Yang et al. 2016; Lu et al. 2016; Yu et al. 2019; 2017b; Shih, Singh, and Hoiem 2016; Liang et al. 2018). With the advancement of visual attention, (Shih, Singh, and Hoiem 2016) introduced the spatial attention that selects the relevant image regions to the text-based questions. (Yang et al. 2016) developed the stacked attention network, and (Liang et al. 2018) proposed the focal visual-text attention network for image question answering, where both visual and text sequence information is presented.

To better exploit the reasoning for visual question answering, some works tried to introduce the answer representation into consideration (Hu, Chao, and Sha 2018; Bai et al. 2018; Teney et al. 2018). The researches in (Hu, Chao, and Sha 2018; Teney et al. 2018) tried to project the joint image-question representation into the answer embedding space learned from a text corpus. Meanwhile, multimodal feature

fusion is crucial for visual question answering, which has been studied in many works (Fukui et al. 2016a; Gao et al. 2016). (Fukui et al. 2016a) showed that a more complicated fusion method does improve the performance. However, the reasoning of visual question answering may still be ineffective due to the complexity of visual and textual information (Bai et al. 2018; Jabri, Joulin, and VanderMaaten 2016; Nguyen and Okatani 2018).

Recently, with the development of datasets (Jang et al. 2017; Zhu et al. 2016; Ranjay et al. 2016) and applications (Zhao et al. 2018; Teney et al. 2018) with multiple questions for a piece of visual content, a few works tried to introduce the information of multiple questions into the reasoning for visual question answering. (Zhao et al. 2018) proposed the hierarchical attention context network for context-aware question understanding, which takes a series of questions and answers as sequential input to train and predict for the target question. Different from these works, in this paper we propose to co-train multiple video-question pairs and to answer them simultaneously. Our considered scenario is that multiple questions are to be answered for a video sequence, instead of visual dialog or multi-turn visual question answering.

Multi-Task Learning. The success of deep learning in general depends on data representation, as different representations can entangle and hid more or less different properties of variation behind the data (Ni et al. 2018; Caruana 1997). Deep network learned representations for a single task can only capture information required by the task and other information of the data may be discarded. In contrast, multi-task learning allows statistical strength sharing and knowledge transfer, thus the representations can capture more underlying factors. This hypothesis has been confirmed by a number of previous works (Misra et al. 2016; Collobert and Weston 2008; Hashimoto et al. 2016), demonstrating the effectiveness of representation learning in scenarios of multi-task learning and transfer learning. However, most of traditional methods of visual question answering treat each video-question pair independently, lack of considering the information of and relations to other questions. Therefore, we are keen to re-frame visual question answering as a multi-question learning problem.

Multi-Question Learning

Setup, Notations, and Main Idea

In this paper, we focus on the choice-based visual question answering task, which is to select an answer a from an answer set $\mathcal{A}$ for the given visual content v and the target question q. The answer set $\mathcal{A}$ is usually defined as all possible answers in the related task (Anderson et al. 2018; Benyounes et al. 2017), *i.e.* the same for all (v, q) pairs. We also define the set $\mathcal{C}$ that contains all the correct answers for (v, q). $\mathcal{C}$ is usually a singleton, but may also contain multiple correct answers depending on the task. Classically, the training dataset is denoted by a set of distinctive triplets $\mathcal{D} = \{(v_n, q_n, \mathcal{C}_n \subseteq \mathcal{A})\}$, which is treated independently in the training procedure. However, this traditional paradigm cannot capture all the information encoded in the visual and textual information to derive better models. This is mainly attributed to two challenges.

First, by separating two different questions q_i and q_j for the same visual content v, with the ordinal numbers i and j, we lose the semantic kinship between the two. Specifically, if there are two triplets $(v, q_i, a_i \in \mathcal{C}_i)$ and $(v, q_j, a_j \in \mathcal{C}_j)$ having semantic relations between q_i and q_j, we would expect the reasoning of a_i and a_j to have some degrees of semantic relations, too. In the traditional setup, two triplets are considered independently such that the learning procedure focuses on the part of the visual and textual information corresponding to the specific triplet, which does not preserve and takes full advantage of related semantic relevance information in the questions. This semantic relevance information from different questions is beneficial and adds robustness for the more general multimodal representation, which can be transferred for answering and reasoning other questions.

Second, traditional methods that treat triplets $\mathcal{D}$ as for an independent classification problem, are severely influenced by the bias of answers of the training data. Class imbalance has been a longstanding challenge for classification, thus it is difficult to learn the visual and textual information for low-frequency questions and answers (Chao, Hu, and Sha 2018). In contrast, learning multiple questions jointly provides a new possibility to narrow the bias of answers of the training data by the transferred information learned from other questions.

Therefore, we propose a new paradigm for studying visual question answering called Multi-Question Learning (MQL) in this work. Specifically, we further consider the question set $\mathcal{Q}$, that is all questions for the specific visual content v. For the specific visual content v_n, we assume there is a set of questions $\mathcal{Q}_n$, and the question number is Q_n, then the corresponding training dataset is re-defined as $\{(v_n, q_k \in \mathcal{Q}_n, \mathcal{Q}_n, \mathcal{C}_k \subseteq \mathcal{A})\}$, which will be co-trained.

Problem Formulation of Multi-Question Learning

For the specific visual content v_n with Q_n questions $q_k \in \mathcal{Q}_n$, we first define a joint embedding function $f_\theta(v_n, \mathcal{Q}_n)$ to generate the general embedding representation r_n. Then, we learn multiple question-answer pairs simultaneously. For each question, the other questions are viewed as regularization. By sharing representations with MQL, we can enable joint visual and textual representations to be general and reliable. We also define a function $g_\phi(r_n, q_k)$ for the target question to emphasize its uniqueness. The related functions are parameterized by θ and ϕ, respectively. In the MQL mechanism, any multimodal representation network can be used, such as multi-layer perceptron (MLP) and Stacked Attention Network (SAN) (Yang et al. 2016). In our work, we also design effective multi-head attention-based network to implement the function f_θ and g_ϕ.

Formally, given the data $(v_n, q_k \in \mathcal{Q}_n, \mathcal{Q}_n, a \in \mathcal{C}_k)$ corresponding to visual content v_n and target question $q_k \in \mathcal{Q}_n$, we define the following probabilistic model,

$$p(a|v_n, q_k, \mathcal{Q}_n) = \frac{\exp(g_\phi(r_n, q_k)^\top W_a)}{\sum_{a' \in \mathcal{A}} \exp(g_\phi(r_n, q_k)^\top W_{a'})} \quad (1)$$

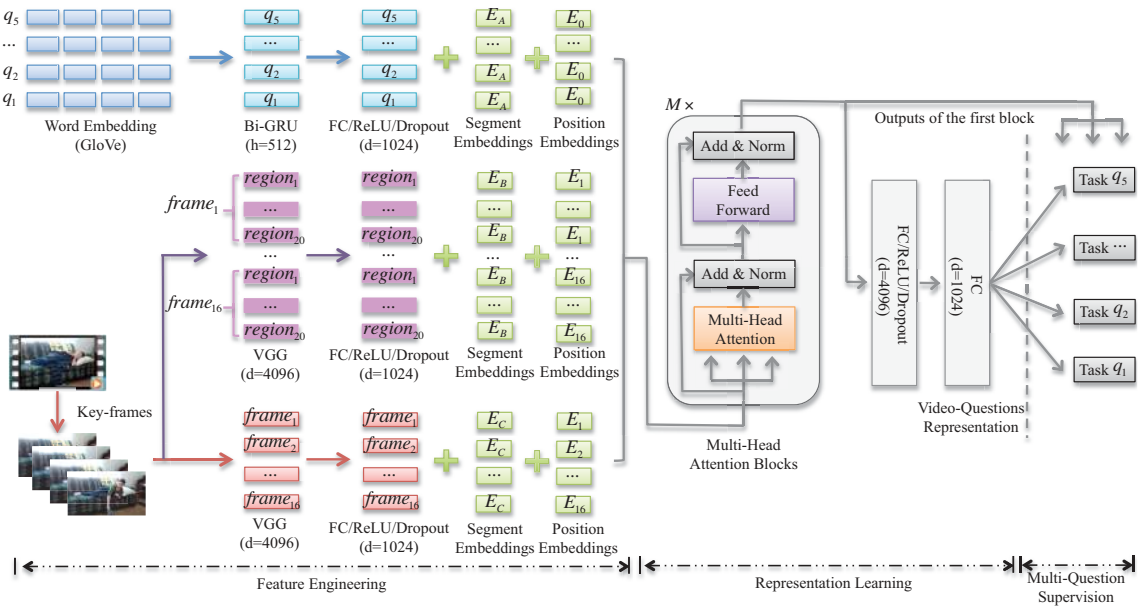


Figure 2: The proposed video question answering framework with the Multi-Question Learning (MQL) mechanism. The framework has three parts including *Feature Engineering*, *Representation Learning*, and *Multi-Question Supervision*.

where $r_n = f_\theta(v_n, \mathcal{Q}_n)$ and W_a is the learned representation for the answer $a \in \mathcal{A}$.

In practice, we need to predefine a constant denoted as M_Q to represent the number of the questions used for a single visual content. There are three different cases regarding the constant for each question. First, if $Q_n = M_Q$, there is nothing to do. Second, if $Q_n < M_Q$, we augment $\mathcal{Q}_n$ with either all-zero vector or duplication. Third, if $Q_n > M_Q$, then for each question q_k, we construct a set $\mathcal{Q}_n^k \subset \mathcal{Q}_n$ such that the subset $\mathcal{Q}_n^k$ has exactly M_Q questions; here we select the top M_Q relevant questions to q_k into the subset, according to some pre-trained models such as BERT (Devlin et al. 2018). In our work, we set $M_Q = 5$ by default. Note that if we set $M_Q = 1$, then the MQL degenerates to single-question learning (SQL), *i.e.* the traditional visual question answering setting.

Given Eq. (1), it is natural to learn the parameters to maximize the likelihood of the probabilistic model. As there might be different numbers of questions for different visual content items, we introduce the weighted likelihood in our work, which has also been exploited in several previous works (Goyal et al. 2017; Hu, Chao, and Sha 2018). Specifically, we have

$$\mathcal{L} = -\sum_{n}^{N} \sum_{k}^{|\mathcal{Q}_n^k|} \frac{|\mathcal{Q}_n^k|}{M_Q} \log p(a|v_n, q_k, \mathcal{Q}_n^k) \tag{2}$$

whose intuition is the hypothesis that the v_n with more related questions contributes more to the MQL mechanism.

During prediction, given the learned parameters mentioned above, we can apply the following decision rule

$$a^* = \arg\max_{a \in \mathcal{A}} g_\phi(r_n, q_k)^\top W_a \tag{3}$$

to identify the answer to the question q_k of visual content v_n, where $r_n = f_\theta(v_n, \mathcal{Q}_n^k)$.

Video Question Answering Framework

The proposed MQL mechanism is generic for different visual question answering tasks as long as there are available multi-question-per-content data for training. In this paper, we focus on the Video QA task as it is considered the most challenging problem among peers (Zhao et al. 2017). Motivated by recent attention mechanisms in NLP (Devlin et al. 2018) and CV (Gao et al. 2019b), as illustrated in Fig. 2, we design a multi-head attention-based neural network to better modeling the correlations and interactions between the multi-questions and videos, which is divided into three parts for a better description. We first extract visual and textual features in the *Feature Engineering* part. Then, the multi-head attention-based question and video representations are generated in the *Representation Learning* part. Finally, there is a *Multi-Question Supervision* part for co-training different questions.

Feature Engineering. From an input video sequence, we first select 16 key-frames and resize each frame to the resolution of 224×224. Then, there are two streams for visual information. One is global information for each frame. In this stream, the initial visual representation of each frame is a 4096-dimensional vector that is extracted by the pre-trained VGG model (Simonyan and Zisserman 2014). The other is region information for each region inside a frame, which is motivated by the claim that the global representation of the frame may fail to capture all necessary information for answering region-specific questions (Li and Jia 2016). We generate 20 candidate regions per frame using the method in (Li and Jia 2016). For each region, we also em-

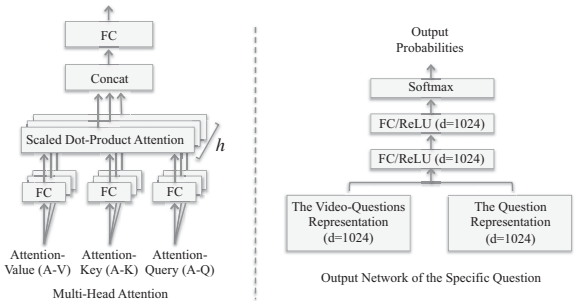


Figure 3: More details of the proposed Video QA framework shown in Fig. 2. Left: the multi-head attention mechanism; Right: the *Multi-Question Supervision* part.

Dataset and Category		QA Pairs			Video Sequences		
		Train	Test	Total	Train	Test	Total
TGIF-QA	Object	16755	5586	22341	15584	3209	18793
	Number	8096	3148	11244	8033	1903	9936
	Color	11939	3904	15843	10872	3190	14062
	Location	2602	1053	3655	2600	917	3517
Total		84124	19795	103919	48198	8522	56720
CPT-QA	Object	7431	769	8200	7288	751	8039
	Number	1435	160	1595	1159	159	1593
	Color	7938	875	8813	1434	868	8749
	Location	7276	795	8071	7242	788	8030
	Action	7108	762	7870	7105	760	7865
	Either-OR	6150	681	6831	5834	644	6478
	Other	2542	358	2900	2115	279	2394
Total		39880	4400	44280	7976	880	8856

Table 1: Statistics of the two datasets, in which we divide questions into several different categories, such as "Object" and "Number."

ploy the pre-trained VGG to generate a 4096-dimensional vector. For textual information, the GloVe model (Pennington, Socher, and Manning 2014) is employed to extract the semantic word embeddings for each question, and a single-layer bidirectional GRU (Bi-GRU) with the hidden dimension of 512 is used on top of each question. All of these Bi-GRUs are sharing parameters. After the extraction of initial representations, we employ three full-connection layers to change dimension to 1024 for each kind of features, respectively. To distinguish the different types of features, three kinds of segment embeddings are added to corresponding features, known as E_A, E_B and E_C, respectively. Since the temporal information is believed more important for video, we further design a position embedding layer like in recent works (Devlin et al. 2018). Note that we allow questions to be out of order, so all questions for a video sequence have the same position embedding E_0, and the frame and its related regions are assigned position embedding E_i for the i-th key-frame. Finally, we have a 1024-dimensional feature for each question, region, and key-frame.

Representation Learning. Motivated by the success of multi-head attention mechanism in recent works (Devlin et al. 2018; Vaswani et al. 2017), we design a multi-head attention-based representation learning approach to capture the complex semantic relations among questions, regions, and key-frames. There are M multi-head attention blocks where M is a hyper-parameter and equals 6 in our work. Each block refers to the structures in (Vaswani et al. 2017), contains two sub-layers. The first is a multi-head self-attention, and the second is a simple input-wise fully connected feed-forward. In this way, the model could exploit a series of important semantic information for representing and reasoning the questions and the video, including the relations among questions, the temporal information among key-frames, and the spatial information and the overall semantic relations among questions and frames.

The left part of Fig. 3 illustrates the multi-head attention mechanism. We first linearly project an input vector h times (h-heads) with different, learned linear projections to d_q, d_k and d_v dimensions, respectively, known as attention-query (A-Q), attention-key (A-K) and attention-value (A-V). In our work, A-Q, A-K, and A-V are the same, and particularly, they are the extracted feature vectors from the Feature Engineering part or the previous multi-head attention block. Next, we perform scaled dot-product attention for each head, and we concatenate all scaled dot-product attention output vectors along each head. A full-connection layer is then employed to get the output. In our work, we set $h = 8$ and $d_q = d_k = d_v = 1024/h = 128$. The feed-forward network includes two full-connection layers, which is applied to each input vector separately and identically.

After multi-head attention blocks, we concatenate all output vectors and utilize two full-connection layers to generate the final question/video representations. Note that the output vectors of the first block corresponding to the input question vector are employed to represent each question, respectively, which will be used in the following *Multi-Question Supervision* part. Our motivation is that the output vectors of the first multi-head attention block preserve and distinguish the information of each question to the largest extent. At the same time, the relations within questions, regions, and frames can be integrated to a certain extent.

Multi-Question Supervision. As mentioned above, we believe that jointly train the multiple questions could make the question/video representations more general and reliable. Thus we define a network for co-training, which is to share parameters for each question and is illustrated in the right part of Fig. 3. We concatenate the representations of questions/video and the target question, in order to emphasize the uniqueness of the target question on the basis of the shared semantic expression of multiple questions and the video. Both of them are generated from the previous *Representation Learning* part. Then, there are two full-connection layers to generate the final feature vector for prediction.

Experiments

Datasets and Metrics

TGIF-QA dataset[1]. It is a large-scale public dataset introduced by (Jang et al. 2017). Since we focus on the problem of selecting appropriate answers from a pool of candidate answers in this paper, we perform our experiments on the "Frame QA" task described in TGIF-QA. In this dataset,

[1]https://github.com/YunseokJANG/tgif-qa

Method	Accuracy(%)
VIS+LSTM(avg) (Ren, Kiros, and Zemel 2015)	34.97
Yu *et al*. (Fukui et al. 2016b)	39.64
ST-TP(R+F) (Jang et al. 2017)	49.50
Co-memory (full) (Gao et al. 2018)	51.50
PSAC(R) (Li et al. 2019)	55.7
Ours (SQL)	57.03
Ours (M-S)	58.20
Ours (MQL)	**59.83**

Table 2: Results on the Frame QA task of TGIF-QA dataset. "Ours (M-S)" stands for using MQL for training but testing with single question.

Method	Accuracy(%)
SAN (SQL)	53.14
SAN (MQL)	58.08
r-STAN (SQL)	55.62
r-STAN (MQL)	59.82
MLAN (SQL)	56.11
MLAN (MQL)	61.54
Ours (SQL)	58.26
Ours (M-S)	59.00
Ours (MQL)	**62.85**

Table 3: Results on the CPT-QA dataset.

each video sequence is accompanied with one or more questions, and each question is associated with only one answer. We follow the training and test dataset of TGIF-QA.

CPT-QA dataset[2]. CPT-QA dataset is a recent Video QA dataset released by Zhejiang Lab. For this dataset, the providers pre-define several types of questions and invite volunteers to ask and answer questions for video sequences. For each video sequence there are up to five questions, and for each question there are three persons providing "ground truth" answers, thus, there can be up to three correct answers for one question. Compared with the TGIF-QA dataset, CPT-QA is smaller but has more variety in the category of questions, as can be observed in Table 1. As this dataset is prepared for a competition, we do not have the ground truth of its test set. In our experiments, we divide its training set into two parts: randomly selecting 90% video sequences for training and using the rest for the test.

Metrics. We use the classification accuracy (ACC) as the metric to evaluate the performance of different methods. For the CPT-QA dataset, it is worth noting that the true positive is defined as the predicted answer hits any one of the correct answers for each question.

Compared Approaches

Our full approach is denoted by Ours (MQL). To identify the general effect of MQL mechanism, we also test a single question learning (SQL) version of our approach, *i.e*. we set $M_Q = 1$ in Eq. (2). Furthermore, we test the performance of our model that is trained based on MQL mechanism but deals with a single question in the test, which is denoted by Ours (M-S). This is to cope with scenarios where only one question is given for a video sequence in practice.

Since there are several state-of-the-art works reported their results on the TGIF-QA dataset, we directly compare our result with these results and mark the related work to each method in Table 2. However, for the CPT-QA dataset, there is no published result to our best knowledge. So we design several baselines following (Zhao et al. 2017):

SAN (SQL) is the incremental algorithm based on stacked attention networks (Yang et al. 2016). Following (Zhao et al. 2017), we add the GRU network to fuse the sequential representations of spatially attended frames for Video QA.

[2]https://tianchi.aliyun.com/competition/entrance/231676/information

MLAN (SQL) is modified from the MLAN algorithm (Yu et al. 2017a), where we add the Bi-GRU network for obtaining the representations of the video and the target question.

r-STAN (SQL) is a state-of-the-art Video QA method (Zhao et al. 2017), which proposed a hierarchal spatio-temporal attentional encoder-decoder learning method with multi-step reasoning process for Video QA.

SAN (MQL), MLAN (MQL) and r-STAN (MQL) are the MQL versions of SAN, MLAN, and r-STAN, respectively, which are compared with Ours (MQL) to reveal the effects of different architectures for modeling the relationship between the questions and videos in MQL. They perform a concatenation and add a full-connection layer on top of the corresponding SQL version.

In summary, we design the comparisons between SQL and MQL to demonstrate the advantages of MQL. Considering that the MQL setting provides more information than the SQL setting during prediction, and it is intuitive that MQL performs better than SQL. For fairer comparisons, we also design and implement several MQL baselines, and compare Ours (MQL) with the other MQL methods. Besides, we also compare all SQL-based methods with Ours (M-S), which only involves a single question during prediction.

All approaches employ common initial textual and visual features. Configurations of our method can be found in Fig. 2 and related sections. All approaches are implemented on the TensorFlow, utilizing 100 parameter servers and 2000 workers, each of with runs with 15 CPU cores. The batch size of Ours (MQL) is set to be 16. All the batches are trained for 30 epochs. We use the Adam optimizer for all the approaches. The initial learning rate is set to 1e-4.

Overall Results

Table 2 and 3 give a comprehensive evaluation for baselines on TGIF-QA and CPT-QA datasets, respectively. Specifically, in all cases in Table 2 and 3, Ours (MQL) achieves the best performance, which shows that the MQL mechanism with an effective multimodal features fusion model can further improve the performance of visual question answering. Meanwhile, these experimental results also reveal a number of interesting points:

- In Table 3, all the methods based on MQL mechanism outperform the corresponding SQL version by a noticeable margin, which suggests the MQL mechanism is critical for the performance of VQA tasks. We attribute the reasons for two points. First, multiple questions contribute to the better description and supplement for the target

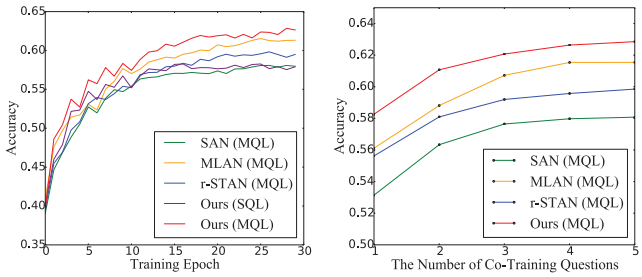


Figure 4: Left: the learning curves of different approaches. Right: the performance with respect to different numbers of co-trained questions (*i.e*. M_Q in our MQL).

question. Second, co-training multiple questions jointly makes the model better understanding the semantic relationships among questions and the visual content for reasoning, and leads to a more general multimodal representation. Furthermore, it indicates that, by better modeling the correlations and interactions between the questions and videos, Ours (MQL) is in favor of outperforming all other MQL-based methods. We make more in-depth discussions of MQL in the next sections.

- Ours (SQL) outperforms the state-of-the-art approaches in Table 2, and Ours (SQL) surpasses the other single question learning baselines in Table 3. It indicates the effectiveness of our proposed video question answering framework, especially the network design. Most probably, the multi-head attention mechanism has a strong ability to correlate and integrate three different types of features, *i.e*. question, region, and frame.
- The accuracy of Ours (M-S) drops a lot compared to that of Ours (MQL). It emphasizes the importance to identify the relations between multiple questions for answer prediction. However, it still outperforms Ours (SQL), which indicates that the model trained with MQL setting is good enough to cope with different test settings: it is applicable when there is only one question for the video sequence during prediction.

Discussion and Analysis

For better understanding our proposed MQL mechanism and video question answering framework, we perform further in-depth discussion and analysis in this section.

We observed the convergence of training with different approaches on the CPT-QA dataset, as shown in the left part of Fig. 4. The learning curves of Ours (SQL) and SAN (MQL) converge quickly after about 15 epochs, and the accuracy on validation set of Ours (SQL) even begins to decline after 15 epochs. In contrast, the learning curves of other MQL based methods grow continually. This phenomenon indicates that multi-question learning mechanism with an effective multimodal fusion model could learn more semantical information from the training data continually, which exploits the training data better.

In our approach, the maximal number of questions for each video sequence during training–M_Q in Eq. (2) is an essential parameter. Since most of the video sequences in the two datasets have up to 5 questions, we test different M_Q values from 1 to 5. Note that when $M_Q = 1$ it is indeed single question learning. The related accuracy results are displayed in the right part of Fig. 4, in which we can observe that the performance increases continually as the number of co-training questions increases. This result suggests the advantage of MQL. Note that the gain becomes marginal when M_Q is greater than 3. One reason is that most of the video sequences in the training data contain no more than 3 questions. The other reason is due to the diversity of questions, *i.e*. introducing more irrelevant questions is not helpful.

Method	Accuracy						
	Object	Color	Number	Location	Action	Either-OR	Other
SAN (SQL)	0.4859	0.5859	0.7379	0.6849	0.2029	0.7667	0.1487
SAN (MQL)	0.5051	0.6332	0.7711	0.7321	0.2414	0.8212	0.1789
r-STAN (SQL)	0.5098	0.6020	0.7524	0.6973	0.2156	0.8001	0.1712
r-STAN (MQL)	0.5386	0.6640	0.8006	0.7468	0.2564	0.8611	0.1953
MLAN (SQL)	0.5244	0.6151	0.8087	0.7149	0.2196	0.8096	0.1856
MLAN (MQL)	0.5537	0.6873	0.8422	0.7644	0.2604	0.8756	0.2093
Ours (SQL)	0.5329	0.6332	0.8310	0.7273	0.2223	0.8240	0.1937
Ours (MQL)	**0.5722**	**0.7033**	**0.8865**	**0.7768**	**0.2770**	**0.8928**	**0.2150**

Table 4: Results of accuracy per each category on the CPT-QA dataset.

Table 4 summarizes the evaluation results of accuracy for different types of questions on the CPT-QA dataset. All MQL based models outperform their corresponding single question learning versions on all kinds of questions. Furthermore, comparing Ours (MQL) with Ours (SQL), we observe that for questions about "Actions," such as "What is the person in the video doing?" the relative improvement is the highest among different types of questions, reaches as much as 24.60%. This can be attributed to the characteristics of these questions, to answer which we need the global information of the video and multiple questions can help here. It also implies that multi-task learning can be helpful for video action recognition.

Conclusion

In this paper, we propose the multi-question learning mechanism as a new paradigm for the visual question answering tasks, analogous to multi-task learning. Multi-question learning mechanism helps generate more robust and generic multimodal representations for visual and textual information, which can not only benefit for the target question, but also be transferred for answering and reasoning other questions. We also propose a multi-head attention-based video question answering framework to implement the multi-question learning mechanism, which captures the complex semantic relations among questions, spatially variant visual information, and temporal information. The empirical evaluations on two large-scale public datasets demonstrate that our proposed approach builds a new state-of-the-art for video question answering.

Acknowledgements

This work was supported in part to Dr. Houqiang Li by NSFC under contract No. 61836011, and was partly supported by Alibaba-Zhejiang University Joint Research Institute of Frontier Technologies.

References

Anderson, P.; He, X.; Buehler, C.; Teney, D.; Johnson, M.; Gould, S.; and Zhang, L. 2018. Bottom-up and top-down attention for image captioning and visual question answering. *In CVPR*.

Antol, S.; Agrawal, A.; Lu, J.; Mitchell, M.; Batra, D.; C.Zitnick; and Parikh, D. 2015. Vqa: Visual question answering. *In ICCV*.

Bai, Y.; Fu, J.; Zhao, T.; and Mei, T. 2018. Deep attention neural tensor network for visual question answering. *In ECCV*.

Benyounes, H.; Cadene, R.; Cord, M.; and Thome, N. 2017. Mutan: Multimodal tucker fusion for visual question answering. *In ICCV*.

Cadene, R.; Ben-younes, H.; Cord, M.; and Thome, N. 2019. Murel: Multimodal relational reasoning for visual question answering. *In CVPR*.

Calhoun, E. 1999. Teaching beginning reading and writing with the picture word inductive model. *Association for Supervision Curriculum Development* 138.

Caruana, R. 1997. Multitask learning. *Machine Learning* 28(1):41–75.

Chao, W.; Hu, H.; and Sha, F. 2018. Cross-dataset adaptation for visual question answering. *In CVPR*.

Collobert, R., and Weston, J. 2008. A unified architecture for natural language processing: Deep neural networks with multitask learning. *In ICML*.

Devlin, J.; Chang, M.; Lee, K.; and Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv:1810.04805*.

Fukui, A.; D.Park; Yang, D.; Rohrbach, A.; Darrell, T.; and Rohrbach, M. 2016a. Multimodal compact bilinear pooling for visual question answering and visual grounding. *In ECCV*.

Fukui, A.; Park, D. H.; Yang, D.; Rohrbach, A.; Darrell, T.; and Rohrbach, M. 2016b. Multimodal compact bilinear pooling for visual question answering and visual grounding. *arXiv preprint arXiv:1606.01847*.

Gao, Y.; Beijbom, O.; Zhang, N.; and Darrell, T. 2016. Compact bilinear pooling. *In CVPR*.

Gao, J.; Ge, R.; Chen, K.; and Nevatia, R. 2018. Motion-appearance co-memory networks for video question answering. *In CVPR*.

Gao, L.; Zeng, P.; Song, J.; Li, Y.; Liu, W.; Mei, T.; and Shen, H. 2019a. Structured two-stream attention network for video question answering. *In AAAI*.

Gao, P.; Jiang, Z.; You, H.; Lu, P.; Hoi, S.; Wang, X.; and Li, H. 2019b. Dynamic fusion with intra- and inter-modality attention flow for visual question answering. *In CVPR*.

Goyal, Y.; Khot, T.; Summers-Stay, D.; Batra, D.; and Parikh, D. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. *In CVPR*.

Hashimoto, K.; Xiong, C.; Tsuruoka, Y.; and Socher, R. 2016. A joint many-task model: Growing a neural network for multiple nlp tasks. *arXiv preprint arXiv:1611.01587*.

Hu, H.; Chao, W.; and Sha, F. 2018. Learning answer embeddings for visual question answering. *In CVPR*.

Jabri, A.; Joulin, A.; and VanderMaaten, L. 2016. Revisiting visual question answering baselines. *In ECCV*.

Jang, Y.; Song, Y.; Yu, Y.; Kim, Y.; and Kim, G. 2017. Tgif-qa: To- ward spatio-temporal reasoning in visual question answering. *In CVPR*.

Li, R., and Jia, J. 2016. Visual question answering with question representation update (qru). *In NIPS*.

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

Liang, J.; Jiang, L.; Cao, L.; Li, L.; and Hauptmann, A. 2018. Focal visual-text attention for visual question answering. *In CVPR*.

Lu, J.; Yang, J.; Batra, D.; and Parikh, D. 2016. Hierarchical question-image co-attention for visual question answering. *In Neural Information Processing Systems (NIPS)*.

Misra, I.; Shrivastava, A.; Gupta, A.; and Hebert, M. 2016. Cross-stitch networks for multi-task learning. *In CVPR*.

Nguyen, D., and Okatani, T. 2018. Improved fusion of visual and language representations by dense symmetric co-attention for visual question answering. *In CVPR*.

Ni, Y.; Ou, D.; Liu, S.; Li, X.; Ou, W.; Zeng, A.; and Si, L. 2018. Perceive your users in depth: Learning universal user representations from multiple e-commerce tasks. *In KDD*.

Pennington, J.; Socher, R.; and Manning, C. 2014. Glove: Global vectors for word representation. *https://nlp.stanford.edu/projects/glove/*.

Ranjay, K.; Yuke, Z.; Oliver, G.; Justin, J.; Kenji, H.; Joshua, K.; Stephanie, C.; Yannis, K.; Jia, L.; David, S.; Michael, B.; and Li, F. 2016. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *https://arxiv.org/abs/1602.07332*.

Ren, M.; Kiros, R.; and Zemel, R. 2015. Exploring models and data for image question answering. *Advances in Neural Information Processing Systems* 2953–2961.

Shih, K.; Singh, S.; and Hoiem, D. 2016. Where to look: Focus regions for visual question answering. *In CVPR*.

Simonyan, K., and Zisserman, A. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Teney, D.; Anderson, P.; He, X.; and Hengel, A. 2018. Tips and tricks for visual question answering: Learnings from the 2017 challenge. *In CVPR*.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.; Kaiser, L.; and Polosukhin, I. 2017. Attention is all you need. *arXiv:1706.03762*.

Yang, Z.; He, X.; Gao, J.; Deng, L.; and Smola, A. 2016. Stacked attention networks for image question answering. *In CVPR*.

Yu, D.; Fu, J.; Mei, T.; and Rui, Y. 2017a. Multi-level attention networks for visual question answering. *In CVPR*.

Yu, Y.; Ko, H.; Choi, J.; and Kim, G. 2017b. End-to-end concept word detection for video captioning, retrieval, and question answering. *In CVPR*.

Yu, Z.; Yu, J.; Cui, Y.; Tao, D.; and Tian, Q. 2019. Deep modular co-attention networks for visual question answering. *In CVPR*.

Zhao, Z.; Yang, Q.; Cai, D.; He, X.; and Zhuang, Y. 2017. Video question answering via hierarchical spatio-temporal attention networks. *In IJCAI*.

Zhao, Z.; Jiang, X.; Cai, D.; Xiao, J.; He, X.; and Pu, S. 2018. Multi-turn video question answering via multi-stream hierarchical attention context network. *In IJCAI*.

Zhu, Y.; Groth, O.; Bernstein, M.; and Li, F. 2016. Visual7w: Grounded question answering in images. *In CVPR*.