RESEARCH ARTICLE

Modality attention fusion model with hybrid multi-head self-attention for video understanding

Xuqiang Zhuang¹, Fang’ai Liu¹*, Jian Hou¹, Jianhua Hao¹, Xiaohong Cai²

¹ School of Information Science & Engineering, Shandong Normal University, Jinan, China, ² College of Intelligence and Information Engineering, Shandong University of Traditional Chinese Medicine, Jinan, China

* lfa@sdu.edu.cn

Abstract

Video question answering (Video-QA) is a subject undergoing intense study in Artificial Intelligence, which is one of the tasks which can evaluate such AI abilities. In this paper, we propose a Modality Attention Fusion framework with Hybrid Multi-head Self-attention (MAF-HMS). MAF-HMS focuses on the task of answering multiple-choice questions regarding a video-subtitle-QA representation by fusion of attention and self-attention between each modality. We use BERT to extract text features, and use Faster R-CNN to extract visual features to provide a useful input representation for our model to answer questions. In addition, we have constructed a Modality Attention Fusion (MAF) framework for the attention fusion matrix from different modalities (video, subtitles, QA), and use a Hybrid Multi-headed Self-attention (HMS) to further determine the correct answer. Experiments on three separate scene datasets show our overall model outperforms the baseline methods by a large margin. Finally, we conducted extensive ablation studies to verify the various components of the network and demonstrate the effectiveness and advantages of our method over existing methods through question type and required modality experimental results.

Introduction

In recent years, studies on video question answering (Video-QA) based on both vision and natural language have successfully benefited from deep neural networks [1–5]. This task aims to select the reasoning process of the correct answer from the answer candidates in the video [6]. Machine’s understanding of images and videos is transitional from labeling the image with a few words to learn to generate a complete sentence.

Most existing methods [7–11] attracted a lot of attention and Video-QA has experienced outstanding advancement. Weinzaepfel et al. [12] proposed a spatio-temporal action localization approach, they applied a tracking-by-detection model which scored video with a combination of static and motion CNN features. To capture more detail in videos, Krishna et al. [13] utilized contextual information from past and future events to jointly describe all events. Lu et al. [14] proposed a multi-step semantic attention network, which learning visual relation
facts as semantic knowledge to help infer the correct answer. However, Video-QA tasks based on vision and natural language require the visual representation of the video combined with subtitles to infer the correct answer, so the Video-QA task is more difficult than image captioning tasks.

The Video-QA task is essentially the fusion of multiple modal data to generate accurate answers to questions related to the video story. Most models for Video-QA, e.g., [15–17] usually use multimodal data to calculate the picture features through the deep convolutional neural network, and the question text features through the recurrent neural network, and then map the input picture and the question features to a common representation space. Finally, the common feature map vector is input to an answer classifier to determine the final answer. However, in real life, the questions people ask about pictures are often related to the target entity in the picture. Therefore, in order to further understand the characteristics of the picture, the image information representation space can be constructed by combining the objects in the picture with the understanding of visual information, and inference the stage focuses on the target entity area of the image and the adjacent caption information. In addition, the characteristics of the questions can be used to focus on different target entity instances according to different questions, thereby improving the accuracy of the answers selected by the model.

In this paper, we propose a novel modal attention fusion model with hybrid multi-head self-attention for Video-QA task based on BERT [18]. We experimented with three Video-QA datasets. First, on the TVQA dataset, we showed that our system outperforms their baseline parser. Second, although the MSVD-QA and MSRVTT-QA datasets are much more challenging and open-ended nature, we are still able to achieve 36.8% and 35.26% accuracy, an absolute improvement of 2.67 points and 0.22 points over baselines, respectively. Additionally, we perform an ablation study and analysis to clarify the strengths of our model, demonstrate the effectiveness of our model in Video-QA. Finally, we demonstrate the effectiveness and advantages of our method over existing methods through question type and required modality experimental results.

Related work

Video question answering

A recent direction in Video-QA leverage text modality such as subtitle in addition to video modality for video understanding. It has been gathering a rising attention in recent years, with the release of various Video-QA benchmarks such as STAGE (Lei et al. [19]), it has proposed a Spatio-Temporal Video Question Answering task of requiring intelligent systems to simultaneously retrieve visual concepts of relevant moments to answer spatio-temporal video question, such as a dual-LSTM based approach (Jang et al. [20]) with both spatial and temporal attention, generates spatial and temporal attention to localize which regions in the video need to attend, such as a video question answering framework (Kim et al., [21]) that requires to simultaneously retrieve the relevant moments and referenced visual concepts, such as the progressive attention memory network (PAMN) proposed by Kim et al. [22], the method extracted features through progressive attention mechanism that utilizes cues from both question and answer to progressively prune out irrelevant temporal parts in memory, such as Two-stream method (Lei et al. [23]) based on a bi-directional LSTM to encode both textual and visual sequences. Different from previous studies, we use BERT in our work to model the information captured in the video clips.

Self-attention for Video-QA

In the past few years, many studies have thus focused on self-attention models, which aim to identify various complex relations to answering questions. For example, Li et al. [24] proposed
position self-attention to simultaneously attend both visual and textual information for improving answer prediction. Kim et al. [25] apply the multi-head attention and self-attention networks to learn the latent concepts in scene frames and captions. Zhang et al. [26] proposed a hierarchical convolutional self-attention encoder-decoder network to efficiently model video contents from videos to answer questions. Jin et al. [27] proposed a multi-interaction network to learn the potential relations between videos and questions. Kim et al. [28] proposed the modality shifting attention network for localizes the temporal moment of interest, and predict the answer using self-attention mechanism on both video and text modality. DFAF framework [29] learn the relationship between multimodalities by adopting the self-attention mechanism.

**Preliminaries**

Our work is designed on Video-QA tasks with BERT and Fast R-CNN. In this section, we formally describe the BERT and Faster R-CNN models.

**BERT**

We use BERT (Devlin et al., [18]) as the backbone of our model architecture to represent all text data. BERT is a language representation model that used bi-directional Transformers to pre-train on a large dataset, and then used the model parameters of the pre-trained model to fine-tune other NLP tasks. In summary, the BERT model further increased the generalization ability of the word vector model, fully describing the character level, word level and sentence level.

**Faster R-CNN**

We use the Faster R-CNN [30] model to extract visual characteristics. Faster R-CNN is mainly composed of RPN network and VGG16 network. RPN network is used to generate regional candidate frames, and VGG16 network is used to extract feature maps of candidate images. Faster R-CNN detects and recognizes the target in the candidate area based on the candidate frame extracted by RPN. The overall process of Faster R-CNN has 4 units marked with different colors and shown in Fig 1.

1. Feature extraction: Faster R-CNN utilizes the VGG16 network to extract the feature map of the candidate image, which is shared for the subsequent RPN layer and fully connected layer.

2. RPN network: RPN network is used to generate candidate area frames. This layer determines that the anchor point belongs to the foreground or the background, and then uses the bounding box regression to correct the anchor frame to obtain an accurate candidate frame.

3. ROI pooling layer: This layer collects the input feature maps and candidate target regions to extract the feature maps of the target region, and then transmits them to the subsequent fully connected layer.

4. Target classification and regression: Faster R-CNN utilizes the feature map of the target area to calculate the category of the target area, and utilizes the bounding box regression to obtain the final precise position of the detection frame.

**Methodology**

The structure of our proposed framework aims to select the correct answer in the Video-QA tasks (shown in Fig 2). We process video and subtitle features with two independent BERT
layers, and use BERT to combine visual conceptual features and subtitles and questions with each candidate answer for embedding. These input features use modal attention fusion (MAF) and a hybrid multi-head self-attention (HMS) mechanism to obtain the final answer prediction.

Fig 1. Faster R-CNN network.
https://doi.org/10.1371/journal.pone.0275156.g001

Fig 2. The network architecture of MAF-HMS.
https://doi.org/10.1371/journal.pone.0275156.g002
Input representation

Text representation. We create 5 hypotheses by concatenating a question representation $q \in \mathbb{R}^{768}$ with 5 candidate answer representations $\{a_i\}_{i=1}^5 \in \mathbb{R}^{768}$. The question and each of the answer candidates were concatenated to form 5 hypotheses $qa \in \mathbb{R}^{n_{qa} \times 768}$, where $n_{qa}$ represents the maximum number of tokens per hypotheses. For each hypothesis, MAF-HMS learns to predict its correctness score and to maximize the score of the correct answer.

Similarly, we create subtitle representations as $s \in \mathbb{R}^{n_s \times 768}$.

Video representation. At first, we extract the sequence of image frames at 3 fps for the video. Then, extract a high-level semantic representation for each image frame. CNN has been recognized as a powerful deep learning model that can capture the visual concept of the image, so our paper uses the Faster R-CNN model to extract visual characteristics $v \in \mathbb{R}^{n_v \times 768}$ from top-20 object proposals. As visual characteristics are in the text domain, they are embedded in the manner as the subtitle.

We extracted video representations $V \in \mathbb{R}^{n_v \times 768}$, text representations $S \in \mathbb{R}^{n_s \times 768}$ in the subtitle, and QA pairs $\{QA_i\}_{i=1}^5 \in \mathbb{R}^{n_{qa} \times 768}$ from the second-to-last layer of two BERT layers.

BERT. Two separate BERT layers are first applied on the subtitle-QA (s and qa) and Video-QA (v and qa) representations. The pre-trained BERT model can be automatically fine-tuned to achieve the most advanced performance in various NLP tasks. The first token in each input by BERT is [CLS], which is used to obtain the output in the classification task. The [SEP] token is added to indicate the separation between the two inputs. In this article, we consider the tokens of input for two BERT layers as follows:

$$[CLS] + s + [SEP] + qa$$

$$[CLS] + v + [SEP] + qa$$

(1)

The output of BERT layers consists of a set of text and video features. These text and video features are flattened and expressed as $V \in \mathbb{R}^{n_v \times 768}$, $S \in \mathbb{R}^{n_s \times 768}$, $\{QA_i\}_{i=1}^5 \in \mathbb{R}^{n_{qa} \times 768}$.

Modality attention fusion framework

Modality attention fusion framework is designed to fuse video features, subtitle features and QA features into the final MAF features, which are then fed into the hybrid multi-head self-attention layer to obtain the final prediction result.

The QA features $QA_i \in \mathbb{R}^{n_{qa} \times 768}$ and video features $V \in \mathbb{R}^{n_v \times 768}$ are fused together as $V$-QA features $S^{V-QA}_i \in \mathbb{R}^{n_s \times 768}$. Similarly, the QA features $QA_j \in \mathbb{R}^{n_{qa} \times 768}$ and subtitle features $S \in \mathbb{R}^{n_s \times 768}$ are fused together as $S$-QA features $S^{S-QA}_i \in \mathbb{R}^{n_v \times 768}$. Then, we use max pooling operation to reduce the size of the fused features of the two different modalities. The S-QA features $S^{S-QA}_i$ as follows:

$$S^{S-QA}_i = QA^T S_i$$

(2)

$$QA^{S-QA}_i = S^T QA_i$$

(3)
\[ S_{i}^{att} = \text{softmax}(S_{i}^{QA}) \cdot S_{i} \]  
\[ QA^{att} = \text{softmax}(QA_{i}^{QA}) \cdot S_{i} \]  
\[ S_{m}^{m} = \text{maxpool}(\text{fc}(\{S_{i}; QA^{att}; S_{i} \odot QA^{att}\})) \]  
\[ QA_{m}^{m} = \text{maxpool}(\text{fc}(\{QA; S_{i}^{att}; QA \odot S_{i}^{att}\})) \]

Where \( \text{fc} \) is a fully-connected layer. The fused subtitle features \( \hat{s}_{i} \) from different directions are integrated by concatenating as follows:
\[ \hat{s}_{i} = \text{fc}(\{QA_{m}^{m}; S_{m}^{m} \odot QA_{m}^{m} + S_{m}^{m}\}) \]

Similarly, we can define the fused video features as follows:
\[ \hat{v}_{i} = \text{fc}(\{QA_{v}^{m}; V_{v}^{m} \odot V_{v}^{m} \odot QA_{v}^{m} + V_{v}^{m}\}) \]

We add fused subtitle features and fused video features to get the final fused features, also called MAF features \( \hat{u} \) as follows:
\[ \hat{u} = \hat{s}_{i} + \hat{v}_{i} \]

**Hybrid multi-head self-attention**

More recently, studies on Video-QA task [24–29] show that learning both visual and textual forms of multi-head self-attention leads to more accurate predictions. Multi-head self-attention mechanism is to map the query matrix (Q), key matrix (K) and value matrix (V) to multiple different subspaces. The subspaces are calculated without interference with each other, and finally the output is stitched together. In this paper, \( \hat{u} \) Containing visual and subtitle semantic features are used as input to the multi-head self-attention layer.

\[ Q = \hat{u} \ast W_{Q}^{i}, i = 1, 2, \cdots, h \]  
\[ K = \hat{u} \ast W_{K}^{i}, i = 1, 2, \cdots, h \]  
\[ V = \hat{u} \ast W_{V}^{i}, i = 1, 2, \cdots, h \]  
\[ \theta_{i, \text{att}}(Q, K, V) = \text{Softmax}(\frac{QK^{T}}{\sqrt{d}})V \]  
\[ H_{i} = \theta_{i, \text{att}}(QW_{Q}^{i}, KW_{K}^{i}, VW_{V}^{i}), i = 1, 2, \cdots, h \]  
\[ \text{MulHead} = \text{Concat}([H_{1}; \cdots; H_{h}])W^{m} \]

Where \( W_{Q}^{i}, W_{K}^{i}, W_{V}^{i} \) are the linear mapping matrices of the query matrix (Q), key matrix (K), and value matrix (V) in the multi-head attention layer.

However, [31] showed that there is inevitably a limitation called the low-rank bottleneck in the self-attention mechanism. Specifically, increasing the number of heads under the premise of fixed model parameters will lead to a reduction in the size of each head, and a smaller head size will introduce rank constraints on the projection matrices of each head, resulting in a
decrease in their expressiveness. Therefore, inspired by [32], we design a hybrid multi-headed self-attentive (HMS) mechanism with the aim of alleviating the low-rank bottleneck issue in multi-headed self-attentive mechanisms for connecting heads to each other to improve the expressive ability of our model.

In the hybrid multi-head self-attention, we add multiple intrinsic inter-vector similarities $\gamma$ to express complex intrinsic relationships of heads on Eq 11. The hybrid multi-head self-attention model is shown in Fig 3, as illustrated, the model requires only very few additional parameters to allow the model to capture complex multimodal information.

The hybrid multi-head self-attention mechanism is defined as:

$$\gamma_{ij} = \begin{pmatrix} \gamma_{11} & \cdots & \gamma_{ih} \\ \vdots & \ddots & \vdots \\ \gamma_{hi} & \cdots & \gamma_{hh} \end{pmatrix}$$  \hspace{1cm} (14)

$$\theta_{s-att}(Q, K, V) = \text{Softmax}\left(\sum_{i=1}^{h} \gamma_{ij} \frac{Q_i K_i^T}{\sqrt{d}} \right) V, i = 1, 2, \ldots, h$$  \hspace{1cm} (15)

$$H_{hyb}^i = \theta_{s-att}(QW_i^Q, KW_i^K, VW_i^V), i = 1, 2, \ldots, h$$  \hspace{1cm} (16)

$$\text{Hyb}_\text{MulHead} = \text{Concat}(H_{hyb}^1, \ldots, H_{hyb}^h) W_{hyb}$$  \hspace{1cm} (17)

where $h$ is the number of heads, and we introduce a new matrix $\gamma_{ij} \in \mathbb{R}^{h \times h}$. $\gamma_{ij}$ denotes the learnable parameter that can automatically measure the correlation between each head during training.

After obtaining the feature vector through hybrid multi-head self-attention, the probability $y$ of each answer is the correct answer is predicted by the fully-connect and softmax layer:

$$y = \text{Softmax}(w_{HMS} \ast \text{Hyb}_\text{MulHead}(\hat{u}) + b_{HMS})$$  \hspace{1cm} (18)

**Experiments**

Dataset

We evaluate our method on three video QA datasets. More details are given below.
TVQA [22]. TVQA dataset is a benchmark for Video-QA, containing 152545 human annotated multiple-choice question-answer pairs (84768 what, 13644 how, 17777 where, 15798 why, 17654 who questions), 21.8K video clips from 6 TV shows (The Big Bang Theory, Castle, How I Met Your Mother, Grey’s Anatomy, House M.D., Friends). The questions in TVQA dataset have five answer candidates and only one of them is ground-truth answer. The format of the questions is designed as follows: “[What/How/Where/Why/who]___[when/before/after/...]?”, and the two part of the question requires visual and linguistic understanding. There are total 122,039 QAs for train set, 15,253 QAs for validation set and 7,623 QAs for test set, respectively.

MSVD-QA [33]. MSVD-QA is based on MSVD [34] video dataset. This is a small open-ended dataset of 50,505 question answer pairs annotated from 1,970 short clips. It consists of five types of questions, including what, how, where, when and who, of which 61% of questions are used for training whilst 13% and 26% are used as validation set and test set, respectively.

MSRVTT-QA [6]. It is an open-ended dataset that contains 10K videos and 243K question answer pairs. It consists of five types of questions, including what, how, where, when and who. Compared to the other two datasets, videos in MSRVTT-QA contain more complex scenes. They are also much longer, ranging from 10 to 30 seconds long, equivalent to 300 to 900 frames per video. Splits for train, validation and test are with the proportions are 65%, 5%, and 30%, respectively.

Baselines

We compared our model with several methods:

- **Two-stream (Lei et al., [23])**. Combines information from different modalities with LSTMs and cross-attention.
- **PAMN (Kim et al., [22])**. Utilize progressive attention memory to update the belief for each answer.
- **Multi-task (Kim et al., [21])**. Using Word2Vec and Bi-LSTM for visual and language representations.
- **STAGE (Lei et al., [19])**. Using RCNN for visual and BERT for language representations.

Implementation

In all the experiments, the recommended train/validation/test split was strictly followed, we independently repeated each experiment 10 times and reported average results of ACC. (accuracy).

We use BERT-Base uncased model, it has 12 layers and 768 hidden sizes. In our experiments, the maximum number of tokens per sequence is set to 128, the batch size is 64, the learning rate is set to 0.0001, and the epochs are set to 10. Our evaluation is performed on a machine with Intel(R) Xeon(R) Gold 6132 CPU (2.60GHz), 256G RAM and Nvidia GeForce RTX 2080 Ti.

Experiment result

Results

Results on TVQA. From Table 1, we noted that our approach outperforms the best previous method by 1.53/1.74 accuracy points on the test/val. As compared to the STAGE, we noted the performance is substantially improved. Considering that there are 15,253 and 7,623 validation and test questions, this establishes the strength of our task. We can see that, our model outperforms the baseline models by a large margin. In particular, the scores of our model
across all the TV shows are more balanced than the scores from other models, which mean our model is more consistent and robust.

We believe the reasons behind the performance boost are the following:

1. Compared with LSTM-based models (Two-stream, PAMN, Multi-task), BERT-based models (STAGE and MAF-HMS) enable capturing longer dependencies between and within different modalities, especially when there is a long subtitle.

2. Our approach can properly integrate input features from different modalities to help answer questions.

3. Hybrid multi-head self-attention can more fully consider the contribution of each modality, and fusion of multi-head results can make the model extract more important features more accurately and improve the performance of our model.

Results on MSVD-QA and MSRVTT-QA. As is shown in Fig 4, we compare our MAF-HMS with Two-stream, PAMN, Multi-task and STAGE on MSVD-QA dataset. Our MAF-HMS achieves the most promising performance in overall accuracy, which demonstrates the superiority of the proposed method on the non-trivial scenarios.

The MSVD-QA and MSRVTT-QA datasets represent highly challenging benchmarks for machine compared to the TVQA, thanks to their open-ended nature. Our MAF-HMS model outperforms existing methods on both datasets, achieving 36.82% and 35.26% accuracy which are 2.67 points and 0.22 points improvement on MSVD-QA and MSRVTT-QA, respectively. This suggests that the model can handle both small and large datasets better than existing methods.
Ablation study

For further analysis, our model shows substantial improvement over the baseline, we evaluated our models with different variants. In these experiments, we choose the same train-and-test setting. From Table 2, we have the following observations.

Effect of model. We design a simpler model with GloVe and LSTM for text representation. As is shown in Table 2, we summarize the ablation analysis of MAF-HMS on TVQA dataset in order to measure the validity of the key components of MAF-HMS. Using BERT for contextual word embeddings significantly improves performance comparing to GloVe embedding and LSTM counterpart.

Effect of variants. To measure to effectiveness of all components, we evaluate the following settings:

MAF-HMS w/o HMS. Remove the hybrid multi-head self-attention unit in the MAF-HMS.
MAF-MS. To verify the validity of hybrid multi-head attention unit, we employ the standard multi-head attention as a substitute.
MAF-HMS w/o MAF. To measure to effectiveness of MAF component, MAF-HMS w/o MAF underperforms MAF-HMS, which shows that the modality attention fusion framework is important in understanding video questions.
MAF-HMS w/ VTF. We use VTF (Video-Text Fusion, from STAGE) instead of MAF.
MAF-HMS w/ DMF. DMF (Dynamic Modality Fusion) is the modal fusion mechanism of PAMN.
MAF-HMS w/ JMCQ. JMCQ (Video-Text Fusion) is the modal fusion mechanism of Two-stream.

The three “w/” variants have little performance improvement compared to MAF-HMS w/o MAF, and our MAF-HMS has increased by 3.96 points. Thus, MAF component contribution is huge. Finally, we compare our method with two simple variants that drop the HMS and replace the HMS with a standard multi-headed self-attentive unit to show the contribution of our hybrid multi-head self-attention unit. Without HMS, we find out that the performance of our MAF-HMS model decreases by 1.08 points on TVQA dataset. We adopt the multi-headed self-attentive to replace the HMS, the result demonstrates the positive role of HMS.

Qualitative results

In this section, we provide analysis the qualitative examples of TVQA benchmark solved by MAF–HMS. We selected a few successful and unsuccessful prediction samples from TVQA dataset. Following, we list some of these observations.

Table 2. Ablation study on model variants of MAF-HMS on the validation set of TVQA.

| Methods                    | Model          | Val. (%) | Δ (%) |
|----------------------------|----------------|----------|-------|
| MAF-HMS w/o BERT          | GloVe + LSTM   | 69.57    | 2.64  |
| MAF-HMS w/o MAF           | BERT           | 68.25    | 3.96  |
| MAF-HMS w/ VTF (STAGE)    | BERT           | 69.13    | 3.08  |
| MAF-HMS w/ DMF (PAMN)     | BERT           | 68.52    | 3.69  |
| MAF-HMS w/ JMCQ (Two-stream) | BERT         | 68.58    | 3.63  |
| MAF-HMS w/o HMS           | BERT           | 71.13    | 1.08  |
| MAF-MS (multi-head self-attention) | BERT       | 71.76    | 0.45  |
| MAF-HMS                   | BERT           | 72.21    | -     |

https://doi.org/10.1371/journal.pone.0275156.t002
For example, the utterance “How did you come to work today” is very important to answer the question, information from neighboring utterances, e.g., “By bus” is the key message to answer the question. All this information is contained in a single modality and is easy to capture, whether it is baselines or our model. Such contextual relationships are prevalent throughout the TVQA dataset. In addition, these are difficult to find the answer information from the subtitles. If there is a lack of key information in the subtitles, such as the question, “what did he hand her”, if the context dependence is not obvious in the subtitles, we can find the answer from the video features, e.g., “A man’s hand holding a hamburger”. There are strongly related entities in the visual features, our method will give correct predictions based on these fusion features.

We further investigate the performance of MAF-HMS by comparing with STAGE on TVQA dataset. For example, the question “Where did she go after she confessed to him”, there are strongly related entities in the visual features, and the answer can be found from the visual modalities, both MAF-HMS and STAGE obtain accurate prediction for such questions. However, in another question “what did he hand her”, there is some disturbing information with confusion. We can find the wrong answer “diaper bag” from the subtitle “well, this is his diaper bag”, and find the correct answer “hamburger” in the video representation, there are strongly related entities in the question features. Other answers can be found in subtitles and video representations, and our model can more accurately eliminate interference options and find the answer directly from the subtitles with understanding the semantics of the question better.

Finally, we show two failure cases. For example, in the question “Who tells Anton that he has quite the temper when interrogating him in the interrogation room?” and the answer is the names of 5 characters, character’s face which is very challenging to capture using the visual concepts extracted by Faster R-CNN. MAF-HMS fails to predict the correct answer as the visual concept features are insufficient in capturing textual cues in the video. Furthermore, in another question “Who is sitting at the computer when the group is talking”, our model is difficult to find the action concept features associated with the question and answer from the subtitles and visual features, so we predict failure. When the answer is not explicit in the video frames or the subtitles, our method gives an incorrect prediction, however our MAF-HMS still achieves outstanding results.

From these examples, we can see that our model is able to solve both visual and language modalities multiple-choice questions.

**Performance by question type**

In this section, we evaluate the proposed model with different question types on TVQA dataset.

This section describes the analysis of MAF-HMS by comparing the accuracy with respect to question type. The Fig 5 exhibits the performance comparison by 5W1H (Who, Where, When, What, Why and How) question type between Two-stream, PAMN, Multi-task, STAGE and MAF-HMS on the validation set of TVQA benchmark.

For fair comparison with existing methods, we tried to reproduce the results on Two-stream, PAMN, Multi-task and STAGE. Notwithstanding the effect achieved on “who” question is not satisfactory, MAF-HMS obtains an average of 72.19% accuracy for the majority of question types, significantly higher than other baselines. Especially, reached 81.93% accuracy performance on “when” question. The reason is that “when” type sentences contain more interactions and inferences between video representation and text representation, and our model can further integrate patterns with BERT and Faster R-CNN to answer multiple choice questions more effectively. The results confirm that our approach shows additional benefit in
mining the modal relations between text and video representation, which implies the superiority of MAF-HMS to help infer the correct answer.

Performance by required modality

To investigate whether MAF-HMS is sensitive to the modality entities, we provide the analysis of MAF-HMS by the required modality for all datasets. For comparison purpose, we also designed three types of labels according to which modality is required for answering prediction.

V. There are only strongly related entities in the visual features.
S. There are only strongly related entities in the subtitle features.
V&S. There are strongly related entities in the visual features and subtitle features.

As is shown in Fig 6, we have manually labeled 3000 (1000 per each type) examples for each dataset.

Fig 5. Performance of all the tasks on TVQA dataset by question type. M1-M5 represent Two-stream, PAMN, Multi-task, STAGE, and MAF-HMS, respectively.

https://doi.org/10.1371/journal.pone.0275156.g005

Fig 6. Analysis by required modality of MAF-HMS.

https://doi.org/10.1371/journal.pone.0275156.g006
It can be seen that require subtitle and video modalities for answer prediction (i.e., label V&S) achieves the highest accuracy score among all datasets, label V has the lowest accuracy score. It means that effectively fusing multi-modal information can strengthen the model’s ability to understand video, and achieving the impressive performance on multi-modal tasks.

Conclusion

In this work, we presented the MAF-HMS model for Video-QA tasks. We use a modality attention fusion framework that combines visual and subtitles representation features to capture the semantics more accurately. We process video and subtitle features with two independent BERT layers, and use BERT to combine visual conceptual features and subtitles and questions with each candidate answer for embedding. These input features use modal attention fusion and a hybrid multi-head self-attention mechanism to obtain the final answer prediction.

Experiments were conducted to test the performance of our model. Results show that our model gave correct predictions from the language and visual representations on TVQA dataset. In our experiments, the MAF-HMS was evaluated on multiple Video QA datasets, especially, has achieved a test accuracy rate of 72.21% on TVQA dataset. Although the MSVD-QA and MSRVTT-QA datasets are much more challenging and open-ended nature, we are still able to achieve 36.82% and 35.26% accuracy which are 2.67 points and 0.22 points improvement on MSVD-QA and MSRVTT-QA. In addition, we perform an ablation study and analysis to clarify the strengths of our model, demonstrate the effectiveness of our model in Video-QA. Finally, experimental results demonstrate the effectiveness of our method in question type and required modality tasks.

Supporting information

S1 Dataset.
(ZIP)

Author Contributions

Conceptualization: Xuqiang Zhuang.
Data curation: Xuqiang Zhuang.
Methodology: Xuqiang Zhuang.
Project administration: Fang’ai Liu.
Resources: Jianhua Hao.
Writing – original draft: Xuqiang Zhuang.
Writing – review & editing: Xuqiang Zhuang, Fang’ai Liu, Jian Hou, Xiaohong Cai.

References

1. T. Le, V. Le, et al. Hierarchical Conditional Relation Networks for Multimodal Video Question Answering. arXiv preprint arXiv:2010.10019 (2020).
2. Wang A., Luu A. T., Foo C.S., Zhu H., Tay Y., and Chandrasekhar V.. Holistic multi-modal memory network for movie question answering. IEEE Transactions on Image Processing, vol. 29, pp. 489–499 (2019).
3. H. Xu, K. Saenko. Ask, attend and answer: Exploring question-guided spatial attention for visual question answering. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 451–466 (2016).
4. Y. Ye, Z. Zhao, Y. Li, L. Chen, J. Xiao, and 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. ACM, pp. 829–832 (2017).

5. A.U. Khan, A. Mazaheri, N.V. Lobo, et al. MMFT-BERT: Multimodal Fusion Transformer with BERT Encodings for Visual Question Answering. arXiv preprint arXiv:2010.14095 (2020).

6. J. Xu, T. Mei, T. Yao, and 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, pp. 5288–5296 (2016).

7. K.M. Kim, M.O. Heo, S.H. Choi, and B.T. Zhang. Deepstory: Video story QA by deep embedded memory networks. International Joint Conferences on Artificial Intelligence, pp. 2016–2022 (2017).

8. L. Zhu, Z. Xu, Y. Yang, and A.G. Hauptmann. Uncovering the temporal context for video question answering. International Journal of Computer Vision, vol. 124, no. 3, pp. 409–421 (2017).

9. Zhao Z., Zhang Z., Xiao S., Xiao Z., Yan X., Yu J., et al. Long-form video question answering via dynamic hierarchical reinforced networks. IEEE Transactions on Image Processing, vol. 28, no. 12, pp. 5939–5952 (2019). https://doi.org/10.1109/TIP.2019.2922062 PMID: 31217111

10. X. Li, L. Gao, X. Wang, W. Liu, X. Xu, H.T. Shen, et al. Learnable aggregating net with diversity learning for video question answering. In Proceedings of the 27th ACM International Conference on Multimedia. ACM, pp. 1166–1174 (2019).

11. Z. Yuan, S. Sun, L. Duan, et al. Adversarial Multimodal Network for Movie Question Answering. arXiv preprint arXiv:1906.09844 (2019).

12. P. Weinzaepfel, Z. Harchaoui, C. Schmid. Learning to track for spatio-temporal action localization. In: Proceedings of the IEEE international conference on computer vision, pp. 3164–3172 (2015).

13. P. Lu, J. Li, W. Zhang, et al. R-VQA: Learning visual relation facts with semantic attention for visual question answering. In: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp. 1880–1889. ACM (2018).

14. R. Krishna, K. Hata, et al. Dense-captioning events in videos. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 706–715 (2017).

15. J. Devlin, M Chang, K. Lee, et al. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018).

16. J. Lei, L. Xu, M. Bansal, et al. TVQA+: Spatio-temporal grounding for video question answering. In arXiv:1904.11574 (2019).

17. Y. Jang, Y. Song, Y. Yu, Y. Kim, and G. Kim. TGIF-QA: Toward spatio-temporal reasoning in visual question answering. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, 21–26 July 2017, pp. 2758–2766 (2017).

18. J. Kim, M. Ma, et al. Gaining extra supervision via multi-task learning for multi-modal video question answering. In: Proceedings of the IEEE Conference on International Joint Conference on Neural Networks, pp. 1–8. IEEE (2019).

19. J. Kim, M. Ma, K. Kim, S. Kim, and C.D. Yoo. Progressive attention memory network for movie story question answering. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 8337–8346 (2019).

20. Lei J., Yu L., Bansal M., et al. TVQA: Localized, compositional video question answering. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, EMNLP, pp. 1532–1543 (2018).

21. Li X., Song J., et al. Beyond rnns: Positional self-attention with co-attention for video question answering. In Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 33, No. 01, pp. 8658–8665 (2019).

22. Zhang, Z., Zhao, Z., et al. Open-ended long-form video question answering via hierarchical convolutional self-attention networks. arXiv preprint arXiv:1906.12158 (2019).
27. Jin, W., Zhao, Z., et al. Multi-interaction network with object relation for video question answering. In Proceedings of the 27th ACM international conference on multimedia, pp. 1193–1201 (2019).

28. J. Kim, M. Ma, T. Pham, et al. Modality Shifting Attention Network for Multi-Modal Video Question Answering. In: 2020 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, 13–19 June 2020, pp. 10106–10115 (2020).

29. Gao, P., Jiang, Z., et al. Dynamic fusion with intra-and inter-modality attention flow for visual question answering. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 6639–6648 (2019).

30. Ren S., He K., et al. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. IEEE Trans. Pattern Anal. Mach. Intell. 39(6), 1137–1149 (2017). https://doi.org/10.1109/TPAMI.2016.2577031 PMID: 27295650

31. S. Bhojanapalli, C. Yun, A.S. Rawat, S.J. Reddi, S. Kumar. Low-rank bottleneck in multi-head attention models. In International Conference on Machine Learning, pp. 864–873 (2020).

32. N. Shazeer, Z. Lan, et al. Talking-Heads Attention. arXiv preprint arXiv:2003.02436, (2020).

33. D. Xu, Z. Zhao, J. Xiao, F. Wu, H. Zhang, X. He, et al. Video question answering via gradually refined attention over appearance and motion. In Proceedings of the 25th ACM international conference on Multimedia. ACM, pp. 1645–1653. (2017).

34. S. Guadarrama, N. Krishnamoorthy, G. Malkamenkar, et al. Youtube2text: Recognizing and describing arbitrary activities using semantic hierarchies and zero-shot recognition. Proceedings of the IEEE international conference on computer vision. pp. 2712–2719 (2013).