

# BERT Representations for Video Question Answering

Zekun Yang<sup>1</sup> Noa Garcia<sup>1</sup> Chenhui Chu<sup>1</sup> Mayu Otani<sup>2</sup> Yuta Nakashima<sup>1</sup> Haruo Takemura<sup>1</sup>  
<sup>1</sup>Osaka University, Japan <sup>2</sup>CyberAgent, Inc., Japan

yang.zekun@lab.ime.cmc.osaka-u.ac.jp, {noagarcia, chu, n-yuta}@ids.osaka-u.ac.jp,  
otani\_mayu@cyberagent.co.jp, takemura@ime.cmc.osaka-u.ac.jp

## Abstract

*Visual question answering (VQA) aims at answering questions about the visual content of an image or a video. Currently, most work on VQA is focused on image-based question answering, and less attention has been paid into answering questions about videos. However, VQA in video presents some unique challenges that are worth studying: it not only requires to model a sequence of visual features over time, but often it also needs to reason about associated subtitles. In this work, we propose to use BERT, a sequential modelling technique based on Transformers, to encode the complex semantics from video clips. Our proposed model jointly captures the visual and language information of a video scene by encoding not only the subtitles but also a sequence of visual concepts with a pre-trained language-based Transformer. In our experiments, we exhaustively study the performance of our model by taking different input arrangements, showing outstanding improvements when compared against previous work on two well-known video VQA datasets: TVQA and Pororo.*

## 1. Introduction

Answering questions automatically is considered as one of the highest goals for an intelligent system. To achieve such a goal, visual question answering (VQA) aims to answer questions about images by extracting the semantic information contained in both the language content (*i.e.* the question) and the visual content (*i.e.* the image). A typical VQA system [42, 1] takes an image and a question pair as input, encodes their visual and language features into high-dimensional vectors, and processes them using attention mechanisms [43] to predict the correct answer.

In the last few years, VQA has attracted a lot of attention and the field has experienced outstanding advancements [11, 1, 16, 2, 34]. However, current frameworks still present some limitations. For example, whereas VQA has been mainly focused on modelling static image-related questions, less attention has been paid to answer questions

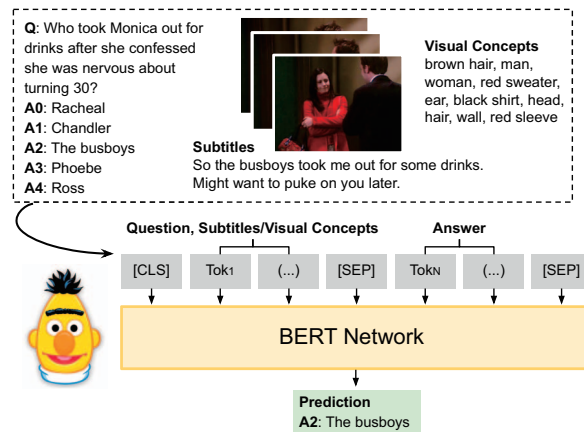


Figure 1. Our proposed model for video-QA based on BERT.

about videos, which requires reasoning over temporal sequences of images. Moreover, most of the efforts in VQA systems are mainly focused on extracting better visual representation from images [23, 6, 26], leaving the modelling of the semantic language contribution to standard recurrent neural networks (RNN).

To address VQA in temporal sequences, video question answering (video-QA) where models need to correctly answer questions about video clips, has been recently studied [36]. Due to the temporal nature of videos, video-QA presents some unique challenges with respect to standard VQA: 1) it requires to understand the temporal coherence in a set of frames [49, 28], and 2) it often needs to consider plot-related questions based on the associated subtitles [21, 18, 8]. This means that video-QA models need to process considerably more input data than standard VQA systems, and hence they need specific methods to extract and represent such amount of visual and language content.

Most models for video-QA introduced so far [36, 21, 18] encode the language information from the questions and the subtitles using RNNs, especially long short-term memory (LSTM) networks [13]. However, LSTM representations may be failing at capturing semantic relationships in long

text sequences, such as the ones that appear in the subtitles of long video clips (*e.g.* about 30 seconds).

In this work, we propose to improve video-QA by capturing the visual and language semantic information from the video clips using BERT representations [7]. BERT is a powerful bidirectional network based on language Transformers [39] and it has been shown to outperform LSTMs in several natural language processing tasks [7, 9, 24]. However, BERT has been barely explored in computer vision applications. For video-QA, Lei *et al.* [22] proposed to use BERT off-the-shelf to extract pre-trained representations from the language information (*i.e.* questions, answers, and subtitles). In this work, we go a step further and perform a deep study about BERT representations for video understanding. We not only fine-tune the network for the task of interest, but also rely on BERT for encoding both the language and the visual information, as shown in Fig. 1.

We address video-QA as a multiple choice task [21, 18]. In our proposed model, we first extract the visual semantic information from each video frame as visual concepts using Faster-RCNN [32] fine-tuned on the Visual Genome dataset [20]. Then, the subtitles and the extracted visual concepts are processed in two independent flows along with the question and candidate answers. In each flow, a fine-tuned BERT network is used to predict the correct answer. The outputs of the two flows are jointly processed to obtain the final prediction. We extensively evaluate our model on two well-known video-QA datasets: TVQA [21] and Pororo [18]. In our experiments, we conduct several ablation studies and comparisons against previous work, showing that our proposed framework improves the accuracy of video-QA by at least 3.34% on the TVQA dataset and 4.89% on the Pororo dataset compare with TVQA model [21] and MDAM [17], respectively.

## 2. Related Work

**Image-Based Question Answering** Image-based question answering, or standard VQA, takes an image and a related question as input, extracts features from the image and the question, and fuse them to predict the correct answer. Lots of methods have been proposed in the last few years. For example, in Spatial Memory Network [42], neuron activations from different regions of the image are stored in a memory; in [29], an entity graph based question answering is proposed, where graph convolutional network is used to simulate the reasoning about the correct answer; in interactive question answering [10], a task of answering questions that require an autonomous agent to interact with a dynamic visual environment is introduced; in [14], a graph network for bridging the gap between the neural and symbolic artificial intelligence is proposed. Besides these improvements, many other related work, such as [33, 25, 30, 12, 3, 46, 11, 1, 16, 2, 34], has contributed to

the advancement of the field.

**Video-Based Question Answering** In contrast to image-based VQA, video-based question answering, or video-QA, needs a joint understanding of the question and candidate answers, a temporal sequence of video frames, and also the associated subtitles. In the last few years, some work related to video-QA has been proposed. For example, [21] proposes a method to answer video-based questions where the visual and language features are embedded by a LSTM, whereas the same authors improve the results in [22] with a network that grounds evidence in both the spatial and temporal domains; in [8] a video understanding task by fusing external knowledge and video-QA is introduced; in [15] a video question answering framework that requires to simultaneously retrieve the relevant moments and referenced visual concepts is proposed; in [49] video-QA is studied in the temporal domain to do inference and prediction; and in [47] a scalable approach to automatically harvest videos and descriptions online and generate candidate QA pairs is proposed. Other related work can be found in [41, 40, 36, 48]. Different from previous studies, we use BERT to model the information captured in the video clips in our work.

**Language Representations** Language representations associate each word in a sentence with a real-valued vector. They are widely used for processing the language information in question answering models and different methods have been proposed in the last few years. For example, GloVe [31] is proposed to leverage statistical information by training only on the non-zero elements in a word-word co-occurrence matrix; Skip-Thoughts [19] learns a generic, distributed sentence encoder in an unsupervised way; in [35], sentiment tree banks and recursive neural tensor networks are used to represent language features; in [27], a LSTM encoder from attentional sequence-to-sequence model is used to contextualize word vectors; in [44], XLNet, a generalized auto-regressive pretraining method that enables learning bidirectional contexts by maximizing the expected likelihood over all permutations of the factorization order is proposed. Besides these methods, some other language representations [5, 38] have also been proposed.

## 3. Proposed Framework

The basic structure of our proposed framework, which aims at answering multiple choice video-QA questions, is shown in Fig. 2. We process the visual and language information in two independent flows, which are lately fused to obtain the jointly answer prediction. In the visual flow, we represent the visual semantic information from each video frame as the set of objects and attributes that appear in the scene, named visual concepts. In the language flow, the

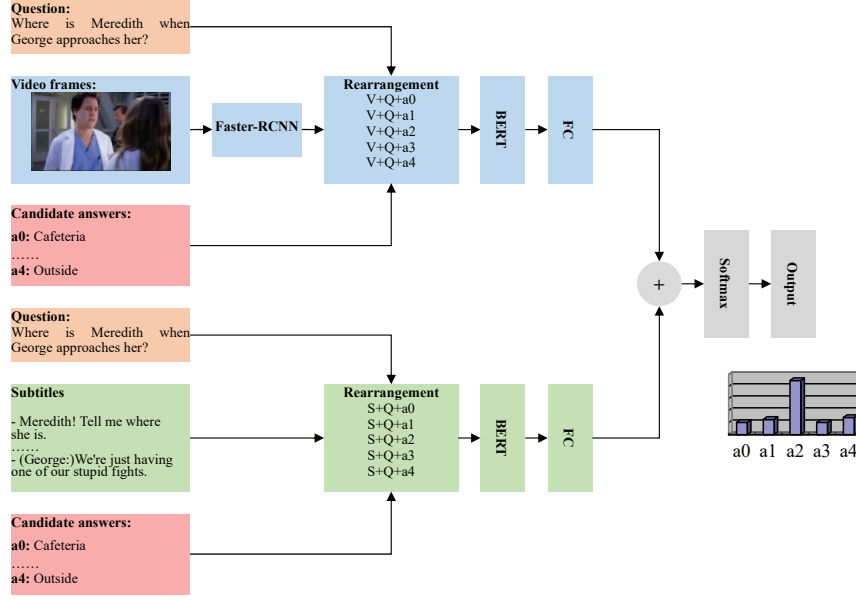


Figure 2. The framework of our proposed model.

language semantic information is extracted from the subtitles. For each flow, visual concept features and subtitles, respectively, are processed along with the question and each candidate answer with a BERT network.

### 3.1. Introduction to BERT

BERT is a language representation model designed to extract pre-trained deep bidirectional representations [7]. It uses bi-directional Transformers [39], meaning every word attends to the context to both sides in every layer of the network. Pre-trained BERT representations can be fine-tuned automatically, achieving state-of-the-art performance in a wide range of tasks [7, 9, 24].

For a given input sequence of word tokens, the input representation of each token is the combination of the corresponding token embeddings, segment embeddings and position embeddings. An example of input representation is shown in Fig. 3, where segment embeddings denote the sentence of each token (A: the former sentence; B: the latter sentence) and position embeddings denote the position of each token within the input sequence. The first token in every sentence is [CLS], which is used to obtain the output in classification tasks. The [SEP] token is added to indicate the separation between two sentences.

### 3.2. Feature Representations and Predictions

We use two independent BERT networks to predict the correct answer of each question based on the information obtained from the visual concept features and subtitles.

**Visual Representations** Recent work [45, 21] has found that using detected object labels as input has comparable or better performance to using CNN features directly in image captioning and video-QA tasks. Thus, to represent the semantic content of the video scene we use detected object labels, named as visual concept features. The visual concept features contain both objects and attributes, such as *grey pants*, *woman*, *blonde hair*, etc. We extract visual concept features from each video frame using Faster R-CNN [32] fine-tuned on the Visual Genome [20] as in [1]. Frames are extracted at 3 fps. In every extracted frame, the visual concept features are represented by corresponding words or noun phrases. The unique visual concept features from a whole scene  $v$  are then obtained by aggregating the visual concepts from all the frames and removing duplicates. Then, the question  $q$ , the unique visual concept features and each candidate answer  $a_i$  ( $i=0,1,2,3,4$ ) are concatenated and rearranged into a single string  $c_i$ . Each rearranged string is tokenized to obtain the sequence  $T_{c_i}$ .

$$c_i = [v, q, a_i] \quad (1)$$

$$T_{c_i} = \text{tokenize}(c_i) \quad (2)$$

Here, the concatenation of  $v$  and  $q$ ,  $[v, q]$ , is set as the former sentence and  $[a_i]$  is set as the latter sentence. The last token(s) in the former and latter sentences are truncated until the number of words in  $T_{c_i}$  is no more than a maximum number of words  $L$ .

Next,  $T_{c_i}$  is fed into the BERT network, which outputs  $V_{c_i}$ , a matrix containing the vector representation of each

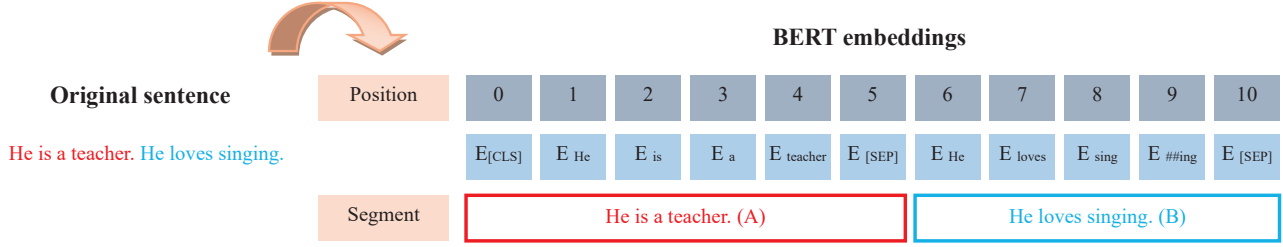


Figure 3. An example of input representation using BERT.

word in the input sentence. The output vector corresponding to the [CLS] token  $V_{c_i}^0$  is fed to a fully connected layer to obtain the visual flow prediction  $R_{c_i}$  for answer  $i$ , where  $F_c$  is a trainable parameter.

$$V_{c_i} = \text{BERT}_c(T_{c_i}) \quad (3)$$

$$R_{c_i} = F_c V_{c_i}^0 \quad (4)$$

**Language Representations** Similarly, in the language flow, we concatenate the subtitles  $s$ , the question  $q$  and the candidate answer items  $a_i$  ( $i=0,1,2,3,4$ ) to form the string  $w_i$ . The rearranged string is tokenized to form the sequence of tokens  $T_{w_i}$ , which is fed into the BERT network to obtain  $V_{w_i}$ . The language flow prediction  $R_{w_i}$  for answer  $i$ , then, is obtained as:

$$w_i = [s, q, a_i] \quad (5)$$

$$T_{w_i} = \text{tokenize}(w_i) \quad (6)$$

$$V_{w_i} = \text{BERT}_w(T_{w_i}) \quad (7)$$

$$R_{w_i} = F_w V_{w_i}^0 \quad (8)$$

**Prediction** Finally, the predictions of visual and language flows for each candidate answer are summed to obtain  $R_{p_i}$ , and softmax is used to convert the summed vector into the answer scores  $R_f$ :

$$R_{p_i} = R_{c_i} + R_{w_i} \quad (9)$$

$$R_p = [R_{p_0}, R_{p_1}, R_{p_2}, R_{p_3}, R_{p_4}] \quad (10)$$

$$R_f = \text{softmax}(R_p) \quad (11)$$

The answer with the maximum score is selected as the final predicted answer  $a_p$ , with:

$$p = \text{argmax}(R_f) \quad (12)$$

## 4. Experiments

**Experimental Settings** Our evaluation is performed on a computer with Core i7 8700K CPU (3.70GHz), 32G RAM and Nvidia TITAN RTX GPU. We use  $BERT_{BASE}$  uncased model, which has 12 layers, 768 hidden sizes, 12 self-attention sizes, 110 million parameters and makes no distinction between upper case and lower case tokens. The learning rate is set to  $2e-5$ , the number of epochs is set to 10, the batch size is set to 8 and  $L$ , the maximum number of tokens per sequence, is set to 128.

**Datasets** We use two video-QA datasets: TVQA [21] and Pororo [18]. TVQA is based on six TV shows with of 152,500 question-answer pairs (Q/A pairs) from 21,800 clips, while Pororo dataset is based on a children's cartoon video series called *Pororo* with 8,834 Q/A pairs from 171 episodes. In both datasets, the subtitles corresponding to each video scene are provided and questions are formulated as a multiple-choice task with one correct answer out of five candidates. The questions in both datasets require a joint understanding about visual and language features to find the correct answer. In TVQA, the corresponding video and language elements are annotated with time stamps in each Q/A pair to denote the related segment of the question. Since the number of total submissions to the test server is limited, we split 15,253 Q/A pairs from the training set to form a test\* set, while the validation set is kept the same. We also report some results on the official test set. In Pororo dataset, besides the videos and subtitles, descriptions about the scenes are given. For comparison, we report results on TVQA [21], STAGE [22] and MDAM [17]. We also wanted to compare against [15] and perform experiments on MovieQA [37], but the dataset was not available when we wrote this paper.

**Input Sequence** We consider three ways of rearranging the input sequences of tokens,  $c_i$  and  $w_i$ :

- 1) [CLS] +V/S+Q+ [SEP] +A
  - 2) [CLS] +V/S+ . +Q+ [SEP] +A
  - 3) [CLS] +V/S+ [SEP] +Q+ [SEP] +A
- (13)



Table 1. Accuracy (in %) of proposed method on TVQA dataset with time stamp annotations. Note that one only has limited chances to submit their results to the test server for evaluation, thus we only show representative results of our method.

Input	Name	Visual	Language	Val	Test*	Test
Q+A	TVQA [21]	-	GloVe + LSTM	42.77	-	<b>43.50</b>
	Ours [CLS] +Q+ [SEP] +A	-	BERT	<b>46.88</b>	47.54	-
V+Q+A	TVQA [21]	GloVe + LSTM	-	45.03	-	<b>45.44</b>
	Ours [CLS] +V+Q+ [SEP] +A	BERT	-	48.91	49.45	-
	Ours [CLS] +V+ . +Q+ [SEP] +A	BERT	-	<b>48.95</b>	49.23	-
	Ours [CLS] +V+ [SEP] +Q+ [SEP] +A	BERT	-	48.74	<b>49.53</b>	-
S+Q+A	TVQA [21]	-	GloVe + LSTM	65.15	-	<b>66.36</b>
	Ours [CLS] +S+Q+ [SEP] +A	-	BERT	70.08	69.42	-
	Ours [CLS] +S+ . +Q+ [SEP] +A	-	BERT	70.09	70.13	-
	Ours [CLS] +S+ [SEP] +Q+ [SEP] +A	-	BERT	<b>70.65</b>	<b>70.22</b>	-
V+S+Q+A	TVQA [21]	GloVe + LSTM	GloVe + LSTM	67.70	-	68.48
	STAGE [22]	GloVe + LSTM	BERT	70.50	-	70.23
	Ours [CLS] +V/S+Q+ [SEP] +A	BERT	BERT	72.06	<b>72.54</b>	<b>73.57</b>
	Ours [CLS] +V/S+ . +Q+ [SEP] +A	BERT	BERT	<b>72.41</b>	72.23	72.71
	Ours [CLS] +V/S+ [SEP] +Q+ [SEP] +A	BERT	BERT	72.35	72.50	73.06

where V means visual concepts, S means subtitles, Q means question and A means answer. V/S means visual concept and subtitles are taken in the visual and subtitle flows, respectively. Ablation studies are made by removing both visual concepts and subtitles (Q+A), only visual concepts (S+Q+A) or only subtitles (V+Q+A).

**Results on TVQA Dataset** Results on TVQA dataset are in Table 1. We report results using the provided time stamp annotations, meaning that we only input the visual concepts and subtitles corresponding to each time stamp. For comparison, we report results of the TVQA model [21], using LSTM for both visual and language representations, and STAGE [22], using LSTM for visual and BERT for language representations. Results show that when we use both visual and subtitle representations, our method obtains an accuracy up to a 5.09% higher than the one obtained with GloVe + LSTM and up to 3.34% higher than STAGE.

**Results on Pororo Dataset** Results on Pororo dataset are in Table 2. As input, we only use the video scenes and subtitles and we do not use the provided video scene descriptions. We compare our method against MDAM [17] and TVQA [21] models. We obtain an accuracy up to 4.89% higher than MDAM and up to 11.26% higher than TVQA. Note that we do not report DEMN model [18] results as they use the video scene descriptions.

**Ablation Study** Table 1 shows that when visual concepts along with questions and answers are used on the TVQA dataset, the accuracy improves by about 2% compared to those using questions and answers only. On the Pororo dataset, the improvement is by more than 13%. We also find that the use of subtitles makes a big leap in the accuracy

Table 2. Accuracy (in %) of proposed method on Pororo dataset.

Input	Name	Model	Val	Test
S+Q+A	MDAM [17]	LSTM	-	42.50
	TVQA [21]	LSTM	37.60	33.90
	Ours [CLS] +S+Q+A	BERT	55.57	52.54
	Ours [CLS] +S+ . +Q+A	BERT	48.93	50.04
	Ours [CLS] +S+ [SEP] +Q+A	BERT	<b>56.49</b>	<b>55.41</b>
V+S+Q+A	MDAM [17]	LSTM	-	48.90
	TVQA [21]	LSTM	37.78	42.53
	Ours [CLS] +V/S+Q+A	BERT	48.93	48.42
	Ours [CLS] +V/S+ . +Q+A	BERT	<b>54.14</b>	<b>53.79</b>
	Ours [CLS] +V/S+ [SEP] +Q+A	BERT	52.45	52.18

of video-QA tasks. When subtitles along with questions and answers are contained in the input on the TVQA dataset, the accuracy increase by more than 20% compared to those using questions and answers only. These results indicate the importance of a strong visual and language representation model in video-QA tasks.

In the TVQA dataset, all of the three proposed rearrangement methods from Eq. (13) give a better prediction than the TVQA model and STAGE. However, there are small differences between each method, implying that BERT deals with the three rearrangements (especially with the separation marks such as "." and [SEP]) in different ways. When the inputs are visual concepts, subtitles, questions, and answers, [CLS] +V/S+ . +Q+ [SEP] +A performs the best in validation set, while [CLS] +V/S+Q+ [SEP] +A performs the best in both test\* and test sets.

## 5. Discussion

**Training Time** On the TVQA dataset, it takes about 3 hours to train an epoch for either the visual or the language flow, while it takes about 5.5 hours to train an epoch for both flows. As reference, the TVQA model [21] on the same ma-

Table 3. Accuracy results (in %) on the TVQA dataset using full length subtitles (without timestamps annotations). Note that one only has limited chances to submit their results to the test server for evaluation, thus we only show representative results of our method.

Input	Visual and Language Representation	Val	Test*	Test
	TVQA [21]	64.42	-	66.46
	PAMN [15]	-	-	66.77
	STAGE [22]	<b>68.56</b>	-	<b>69.67</b>
V+S+Q+A	Ours [CLS] +V/S+Q+ [SEP] +A	60.17	60.48	-
	Ours [CLS] +V/S+. +Q+ [SEP] +A	60.42	60.47	-
	Ours [CLS] +V/S+ [SEP] +Q+ [SEP] +A	61.97	61.67	-
	Ours [CLS] +V/S+Q+ [SEP] +A with pruning	63.07	<b>62.77</b>	62.72
	Ours [CLS] +V/S+. +Q+ [SEP] +A with pruning	62.05	61.94	-
	Ours [CLS] +V/S+ [SEP] +Q+ [SEP] +A with pruning	62.87	62.54	-

chine and the same batch size takes about 2.25 hours to train an epoch for either the visual or the language flow, while it takes about 4.25 hours to train an epoch for both flows. This difference is because BERT encodes the answers along with the question and visual concepts or subtitles, respectively, making it time-consuming and memory-consuming in the tokenization and training. Usually, the best validation accuracy can be obtained within 2 epochs in both flows.

**Evaluation with Full-Length Elements** To test the performance of our model under non-annotation situations, we use the full-length elements (*i.e.* visual concepts and subtitles without time stamp annotations) instead of the time stamp annotated elements. The maximum number of tokens per input,  $L$ , is 512. The results are in Table 3. The best validation accuracy of our proposed method is 61.97% being 2.45% lower than the TVQA model without time stamp annotations and 6.59% lower than STAGE [22]. The drop in performance with respect using the time stamp annotations may be because the input sequence is limited in size by the maximum number of tokens  $L$ , being many full-length subtitles longer than 512.

To overcome this limitation, we follow [4] and prune the irrelevant part of the subtitles using similarities between their TF-IDF sentence representations. First, we generate a vocabulary of about 44,000 words with all the tokens that appear in the TVQA train set at least 5 times. Then, we segment the tokenized subtitles into sections of 400 tokens and compute the cosine similarity between the TF-IDF representations of each section and the question. Finally, we select the section with the highest cosine similarity as the input subtitle. The best validation accuracy of our proposed method with pruning is a 1.1% higher than without pruning, however, on the test server, the result is still 4.05% lower than PAMN [15] and 6.95% lower than STAGE.

By analysing the results in detail, there may be two main reasons for this phenomena. First, full-length elements contain too many words to be covered in the embeddings. Our method embeds the visual concepts/subtitles, the question, and candidate answers altogether, and truncating the

Table 4. Statistics of the input sequences on the TVQA test\* set.

	Max	Min	Avg	>128	>256	>512
Visual flow	527	20	135.73	45.11%	3.94%	0.01%
Subtitle flow	684	18	95.58	15.30%	3.30%	0.75%

remaining words when the input sequence is longer than  $L$ . However, in [21, 15, 22], the questions, the answers, the subtitles and the visual features are embedded independently, enabling the tokenized input to contain more words and convey more information. Second, our work does not use attention mechanisms to find the corresponding part of visual/subtitle elements that is related to the question. Even if we are using TF-IDF for pruning, there are about 20% of the the pruned tokens are more than 512 words. In TVQA model, context matching modules are used to build context-aware vectors, which is helpful for prediction. In PAMN, dual memory embedding is adopted to enable pinpointing different temporal part for each module. In STAGE, guided attention is applied to match the words in questions/answers to the visual concepts and subtitles. In our future work we will explore attention mechanisms to improve the prediction in the full-length elements.

**Evaluation with Different Sequence Lengths** Next, we explore whether different  $L$  values have influence on the video-QA accuracy. The maximum number of words, the minimum number of words, the average number of words and the percentage of having more than 128, 256, and 512 words with time stamp annotations are calculated in Table 4 for the visual and language sequence of the TVQA dataset test\* set. There are more than 45% of the sequences having more than 128 words in visual flow, and more than 15% of the sequences with more than 128 words in subtitle flow. We can also see that the percentage of sequence with more than 512 words in both flows is less than 1%.

We evaluate our model with three different values of  $L$  (128, 256, and 512) on the TVQA dataset test\* set with time stamp annotations. The results are shown in Table 5. When  $L$  increases, the test\* accuracies of our model also rise up. The reason is that  $L$  implies the amount of information con-

Table 5. Accuracy results (in %) on te TVQA test\* set using different  $L$  values (128, 256, and 512).

Input Sequence	128	256	512
[CLS] +V/S+Q+ [SEP] +A	72.54	72.82	73.05
[CLS] +V/S+ . +Q+ [SEP] +A	72.23	72.68	72.79
[CLS] +V/S+ [SEP] +Q+ [SEP] +A	72.35	72.96	<b>73.15</b>

Table 6. Accuracy results (in %) on TVQA test\* set without  $\langle \text{eos} \rangle$  and ".".

Input Sequence	Original	EOS	EOS, "."
[CLS] +V/S+Q+ [SEP] +A	72.54	73.15	72.67
[CLS] +V/S+ . +Q+ [SEP] +A	72.23	72.71	72.79
[CLS] +V/S+ [SEP] +Q+ [SEP] +A	72.50	72.94	72.79

veyed into the BERT network.

**Evaluation with Minor Embedding Changes** To examine how the prediction is affected by minor changes in the embeddings, we remove the  $\langle \text{eos} \rangle$  mark and both  $\langle \text{eos} \rangle$  and full stop (".") in the TVQA dataset subtitles before tokenization on test\* set with time stamp annotations.  $L$  is set as 128. The results are shown in Table 6. By applying these changes, the accuracy increases. Intuitively, the input sequence has more content words when  $\langle \text{eos} \rangle$  and full stop are removed. Results show that when [CLS] +V/S+ . +Q+ [SEP] +A is adopted, the accuracy of removing  $\langle \text{eos} \rangle$  and "." is higher than removing merely  $\langle \text{eos} \rangle$ . However, when [CLS] +V/S+Q+ [SEP] +A and [CLS] +V/S+ [SEP] +Q+ [SEP] +A are adopted, the accuracy of removing  $\langle \text{eos} \rangle$  and "." is lower than removing merely  $\langle \text{eos} \rangle$ . This implies that there is some influence of these token marks in the sentence segmentation. For example, when a full stop is input, the network may identify a sentence before it, and when both  $\langle \text{eos} \rangle$  and full stops are removed, the network may be find it hard to segment the input sequence. From our experiments, "." mark has a stronger influence than  $\langle \text{eos} \rangle$  in making correct predictions.

**Qualitative Results** We show some examples of successful and unsuccessful predictions of our framework and TVQA model in Fig. 4. [CLS] +V/S+ [SEP] +Q+ [SEP] +A embedding is used. The words or the boxes in blue are hints to the correct prediction. Questions 1, 3 and 6 are related to the visual features, while questions 2, 4 and 5 are related to the subtitles.

Our framework correctly predicts question 1, 2 and 5, but incorrectly predicts 3, 4 and 6. In question 1, the scene looks like an office. Even if the time annotated subtitles are long, our framework captures the visual features related to an office in the video scene, so it gives a correct prediction. In question 2, the answer can be found from the subtitles, where Sheldon explains the difference between comics and comic books. This question spans about 12 seconds

and is quite challenging. Our model gives the correct prediction according to the subtitles. In question 3, we can see that Cameron is staring at the computer while talking in the video frame, but our framework finds it difficult to capture this feature, as there are many people in the scene. In question 4, we can read from the subtitles Stuart says that "your mother already gave me the money", then Howard says "What?". Our model also failed to predict the correct answer to this question. This may be because our framework find "Stuart" and "mother" in the same line, but cannot tell who is the person that gives the money. In question 5, the answer can be directly found from the subtitles, saying "But the bad memories crowded out the good and she ran". In question 6, Bernadette's face shows a sad expression, which is very challenging to capture using the visual concepts extracted by Faster R-CNN.

From these examples, we can see that our model is able to solve both visual and language related questions that cannot be solved by LSTM. However, when the answer is not explicit in the video frames or the subtitles and needs outside knowledge sources, our method gives a bad prediction. The reason may be that BERT uses a self-attention bi-directional structure, making every word to attend its context on both sides, while in LSTM, the follow-up words in the long sentence may have a weak attendance to its long previous words. However, the attention mechanism in TVQA model helps to pinpoint relevant words, which might be the reason for correctly predicting questions 3 and 4.

## 6. Conclusion

In this work, we proposed an improved framework for video-QA tasks. We used language representations for visual concept features and subtitles based on BERT to capture the semantics in both the video scenes and subtitles more accurately. Experiments were conducted to test the performance of our model by taking different input arrangements, subtitles with/without time stamp annotations, different maximum length and some minor changes of the input sentence into BERT models. Results show that our model gave correct predictions from the language and visual representations on TVQA and Pororo datasets. Our model improved the performance by 5.09% compared to the previous work based on LSTM, and 3.34% compared with the STAGE network. However, our model present some limitations when using full-length subtitles. As a future work, we will explore the use of attention mechanism to identify the relevant part of the long full-length subtitles.

## Acknowledgement

This work was supported by JSPS KAKENHI No. 18H03264 and ACT-I, JST.







<p>(1) </p> <p><b>Subtitles:</b> 00:00:21,578 --&gt; 00:00:23,346 I told him he was gonna get us both killed. 00:00:23,413 --&gt; 00:00:26,115 (Castle:)But he only got himself killed. Trying to save you. 00:00:27,685 --&gt; 00:00:30,586 (Beckett:)We have an audio recording of Boothe from that night.</p> <p><b>Film:</b> <i>Castle</i> <b>Time stamp:</b> 21.81-29.82</p>	<p><b>Q:</b> Where were Castle and Beckett when they were talking to Trey about Lance's death? <b>a0:</b> In a jail cell <b>a1:</b> In an office <b>a2:</b> In a squad car <b>a3:</b> At the crime scene <b>a4:</b> In the parking lot</p> <p><b>Our prediction:</b> 1 (✓) <b>TVQA prediction:</b> 0 (✗)</p>
<p>(2) </p> <p><b>Subtitles:</b> 00:00:23,334 --&gt; 00:00:26,464 Oh, hey, could you pick me up a few comics for my nephew's birthday? 00:00:26,629 --&gt; 00:00:28,629 No, I think you mean comic books. 00:00:28,798 --&gt; 00:00:30,678 Comics are feeble attempts at humor... 00:00:30,841 --&gt; 00:00:33,631 (Sheldon:)...featuring talking babies and anthropomorphized pets... 00:00:33,803 --&gt; 00:00:38,053 (Sheldon:)...found traditionally in the optimistically named "funny pages."</p> <p><b>Film:</b> <i>The Big Bang Theory</i> <b>Time stamp:</b> 25.84-37.81</p>	<p><b>Q:</b> What does Sheldon explain the difference between after Penny asks for a favor? <b>a0:</b> Nuclear fusion and nuclear fission. <b>a1:</b> Sausage and sausage patties. <b>a2:</b> Comics and Comic books. <b>a3:</b> A yard and a meter. <b>a4:</b> Organic chemistry and inorganic chemistry.</p> <p><b>Our prediction:</b> 2 (✓) <b>TVQA prediction:</b> 3 (✗)</p>
<p>(3) </p> <p><b>Subtitles:</b> 00:00:08,067 --&gt; 00:00:12,231 (Cameron:)House would let you out of it in a heartbeat. Or he wouldn't, just to jerk me around.</p> <p><b>Film:</b> <i>House</i> <b>Time stamp:</b> 9.83-11.17</p>	<p><b>Q:</b> Who is sitting at the computer when the group is talking? <b>a0:</b> Cameron <b>a1:</b> Chase <b>a2:</b> Foreman <b>a3:</b> House <b>a4:</b> Cuddy</p> <p><b>Our prediction:</b> 2 (✗) <b>TVQA prediction:</b> 0 (✓)</p>
<p>(4) </p> <p><b>Subtitles:</b> 00:00:27,494 --&gt; 00:00:28,924 (Stuart:)I appreciate the offer, 00:00:28,928 --&gt; 00:00:31,898 (Stuart:)but actually your mother already gave me the money. 00:00:33,700 --&gt; 00:00:36,000 (Howard:)What? 00:00:36,002 --&gt; 00:00:38,202 (Stuart:)Yeah. I told her it was too much, 00:00:38,204 --&gt; 00:00:41,304 (Stuart:)but she said she was happy to help out her bubala.</p> <p><b>Film:</b> <i>The Big Bang Theory</i> <b>Time stamp:</b> 28.07-41.19</p>	<p><b>Q:</b> Who does Stuart say gave him money to reopen his comic book store after Howard offers him some? <b>a0:</b> He inherited some money from a relative <b>a1:</b> Stuart's mom <b>a2:</b> Sheldon <b>a3:</b> Leonard <b>a4:</b> Howard's mom</p> <p><b>Our prediction:</b> 1 (✗) <b>TVQA prediction:</b> 4 (✓)</p>
<p>(5) </p> <p><b>Subtitles:</b> 00:00:26,373 --&gt; 00:00:28,568 (Masters:)If it's hit her brain, that could mean she doesn't have long. 00:00:28,676 --&gt; 00:00:32,203 So then the question becomes, "Will the sister show up at the funeral?" 00:00:32,279 --&gt; 00:00:33,541 (Chase:)She tried to reconcile. 00:00:33,614 --&gt; 00:00:36,242 (Chase:)But the bad memories crowded out the good and she ran. 00:00:36,350 --&gt; 00:00:37,749 (House:)Nobody's perfect.</p> <p><b>Film:</b> <i>House</i> <b>Time stamp:</b> 27.6-36.5</p>	<p><b>Q:</b> What reason did Chase give for the patient's sister not talking to her when talking to House on the phone? <b>a0:</b> The sister is actually her mother <b>a1:</b> She didn't remember her <b>a2:</b> They hate each other <b>a3:</b> She doesn't have a sister <b>a4:</b> Bad memories</p> <p><b>Our prediction:</b> 4 (✓) <b>TVQA prediction:</b> 4 (✓)</p>
<p>(6) </p> <p><b>Subtitles:</b> 00:00:00,222 --&gt; 00:00:03,862 (Howard:)...before my dad left me and my mom... 00:00:03,859 --&gt; 00:00:07,399 (Howard:)he used to... take me to the comic book store. 00:00:08,463 --&gt; 00:00:11,963 (Howard:)It was one of the few things we did together. 00:00:11,967 --&gt; 00:00:14,837 (Bernadette:)Oh. Howie, I had no idea.</p> <p><b>Film:</b> <i>The Big Bang Theory</i> <b>Time stamp:</b> 0-14.18</p>	<p><b>Q:</b> How did Bernadette feel when Howard told about his dad? <b>a0:</b> Happy <b>a1:</b> Sad <b>a2:</b> Nervous <b>a3:</b> Anxious <b>a4:</b> Angry</p> <p><b>Our prediction:</b> 3 (✗) <b>TVQA prediction:</b> 3 (✗)</p>

Figure 4. Successful and unsuccessful predictions of our framework and TVQA model.



## References

- [1] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6077–6086, 2018.
- [2] R. Cadene, H. Ben-Younes, M. Cord, and N. Thome. Murel: Multimodal relational reasoning for visual question answering. In *Proc. CVPR*, pages 1989–1998, 2019.
- [3] L. Chen, Q. Li, H. Wang, and Y. Long. Static correlative filter based convolutional neural network for visual question answering. In *2018 IEEE International Conference on Big Data and Smart Computing (BigComp)*, pages 526–529. IEEE, 2018.
- [4] C. Clark and M. Gardner. Simple and effective multi-paragraph reading comprehension. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 845–855, 2018.
- [5] A. Conneau, D. Kiela, H. Schwenk, L. Barrault, and A. Bordes. Supervised learning of universal sentence representations from natural language inference data. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 670–680, 2017.
- [6] M. T. Desta, L. Chen, and T. Kornuta. Object-based reasoning in vqa. In *2018 IEEE Winter Conference on Applications of Computer Vision*, pages 1814–1823. IEEE, 2018.
- [7] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4171–4186, 2019.
- [8] N. Garcia, M. Otani, C. Chu, and Y. Nakashima. KnowIT VQA: Answering knowledge-based questions about videos. In *Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence*, 2020.
- [9] L. Gong, D. He, Z. Li, T. Qin, L. Wang, and T. Liu. Efficient training of BERT by progressively stacking. In *International Conference on Machine Learning*, pages 2337–2346, 2019.
- [10] D. Gordon, A. Kembhavi, M. Rastegari, J. Redmon, D. Fox, and A. Farhadi. Iqa: Visual question answering in interactive environments. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4089–4098, 2018.
- [11] Y. Goyal, T. Khot, D. Summers-Stay, D. Batra, and D. Parikh. Making the V in VQA matter: Elevating the role of image understanding in Visual Question Answering. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.
- [12] D. Gurari, Q. Li, A. J. Stangl, A. Guo, C. Lin, K. Grauman, J. Luo, and J. P. Bigham. Vizwiz grand challenge: Answering visual questions from blind people. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3608–3617, 2018.
- [13] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.
- [14] D. A. Hudson and C. D. Manning. Learning by abstraction: The neural state machine. *arXiv preprint arXiv:1907.03950*, 2019.
- [15] 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*, pages 8337–8346, 2019.
- [16] J.-H. Kim, J. Jun, and B.-T. Zhang. Bilinear attention networks. In *Proc. NeurIPS*, pages 1564–1574, 2018.
- [17] K.-M. Kim, S.-H. Choi, J.-H. Kim, and B.-T. Zhang. Multimodal dual attention memory for video story question answering. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 673–688, 2018.
- [18] K.-M. Kim, M.-O. Heo, S.-H. Choi, and B.-T. Zhang. Deep-story: video story qa by deep embedded memory networks. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, pages 2016–2022. AAAI Press, 2017.
- [19] R. Kiros, Y. Zhu, R. R. Salakhutdinov, R. Zemel, R. Urtasun, A. Torralba, and S. Fidler. Skip-thought vectors. In *Advances in Neural Information Processing Systems*, pages 3294–3302, 2015.
- [20] R. Krishna, K. Hata, F. Ren, L. Fei-Fei, and J. Carlos Niebles. Dense-captioning events in videos. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 706–715, 2017.
- [21] J. Lei, L. Yu, M. Bansal, and T. Berg. Tvqa: Localized, compositional video question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1369–1379, 2018.
- [22] J. Lei, L. Yu, T. L. Berg, and M. Bansal. Tvqa+: Spatio-temporal grounding for video question answering. *arXiv preprint arXiv:1904.11574*, 2019.
- [23] Q. Li, J. Fu, D. Yu, T. Mei, and J. Luo. Tell-and-answer: Towards explainable visual question answering using attributes and captions. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1338–1346, 2018.
- [24] Z. Li, X. Ding, and T. Liu. Story ending prediction by transferable BERT. *arXiv Preprint arXiv:1905.07504*, 2019.
- [25] H. Liu, S. Gong, Y. Ji, J. Yang, T. Xing, and C. Liu. Multimodal cross-guided attention networks for visual question answering. In *2018 International Conference on Computer Modeling, Simulation and Algorithm (CMSA 2018)*. Atlantis Press, 2018.
- [26] P. Lu, L. Ji, W. Zhang, N. Duan, M. Zhou, and J. Wang. 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*, pages 1880–1889. ACM, 2018.
- [27] B. McCann, J. Bradbury, C. Xiong, and R. Socher. Learned in translation: Contextualized word vectors. In *Advances in Neural Information Processing Systems*, pages 6294–6305, 2017.
- [28] J. Mun, P. Hongsuck Seo, I. Jung, and B. Han. MarioQA: Answering questions by watching gameplay videos. In *Proc. ICCV*, pages 2867–2875, 2017.

- [29] M. Narasimhan, S. Lazebnik, and A. Schwing. Out of the box: Reasoning with graph convolution nets for factual visual question answering. In *Advances in Neural Information Processing Systems*, pages 2654–2665, 2018.
- [30] M. Narasimhan and 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 (ECCV)*, pages 451–468, 2018.
- [31] J. Pennington, R. Socher, and C. Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pages 1532–1543, 2014.
- [32] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems*, pages 91–99, 2015.
- [33] N. Ruwa, Q. Mao, L. Wang, J. Gou, and M. Dong. Mood-aware visual question answering. *Neurocomputing*, 330:305–316, 2019.
- [34] A. Singh, V. Goswami, V. Natarajan, Y. Jiang, X. Chen, M. Shah, M. Rohrbach, D. Batra, and D. Parikh. Pythia-a platform for vision & language research. In *SysML Workshop, NeurIPS*, volume 2018, 2018.
- [35] R. Socher, A. Perelygin, J. Wu, J. Chuang, C. D. Manning, A. Ng, and C. Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, 2013.
- [36] M. Tapaswi, Y. Zhu, R. Stiefelhagen, A. Torralba, R. Urtasun, and S. Fidler. Movieqa: Understanding stories in movies through question-answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4631–4640, 2016.
- [37] M. Tapaswi, Y. Zhu, R. Stiefelhagen, A. Torralba, R. Urtasun, and S. Fidler. Movieqa: Understanding stories in movies through question-answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4631–4640, 2016.
- [38] J. Turian, L. Ratinov, and Y. Bengio. Word representations: a simple and general method for semi-supervised learning. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 384–394. Association for Computational Linguistics, 2010.
- [39] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. In *Proc. NIPS*, pages 5998–6008, 2017.
- [40] Y.-C. Wu, C.-H. Chang, and Y.-S. Lee. Clvq: Cross-language video question/answering system. In *IEEE Sixth International Symposium on Multimedia Software Engineering*, pages 294–301. IEEE, 2004.
- [41] Y.-C. Wu and J.-C. Yang. A robust passage retrieval algorithm for video question answering. *IEEE Transactions on Circuits and Systems for Video Technology*, 18(10):1411–1421, 2008.
- [42] H. Xu and K. Saenko. Ask, attend and answer: Exploring question-guided spatial attention for visual question answering. In *European Conference on Computer Vision*, pages 451–466. Springer, 2016.
- [43] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *Proc. ICML*, pages 2048–2057, 2015.
- [44] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. Salakhutdinov, and Q. V. Le. Xlnet: Generalized autoregressive pretraining for language understanding. *arXiv preprint arXiv:1906.08237*, 2019.
- [45] X. Yin and V. Ordonez. Obj2text: Generating visually descriptive language from object layouts. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 177–187, 2017.
- [46] D. Yu, X. Gao, and H. Xiong. Structured semantic representation for visual question answering. In *2018 25th IEEE International Conference on Image Processing (ICIP)*, pages 2286–2290. IEEE, 2018.
- [47] K.-H. Zeng, T.-H. Chen, C.-Y. Chuang, Y.-H. Liao, J. C. Niebles, and M. Sun. Leveraging video descriptions to learn video question answering. In *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.
- [48] Z. Zhao, X. Jiang, D. Cai, J. Xiao, X. He, and S. Pu. Multi-turn video question answering via multi-stream hierarchical attention context network. In *IJCAI*, pages 3690–3696, 2018.
- [49] L. Zhu, Z. Xu, Y. Yang, and A. G. Hauptmann. Uncovering the temporal context for video question answering. *International Journal of Computer Vision*, 124(3):409–421, 2017.