Learning Question-Guided Video Representation for Multi-Turn Video Question Answering

Guan-Lin Chao
Carnegie Mellon University
guanlinchao@cmu.edu

Abhinav Rastogi
Google AI
abhirast@google.com

Semih Yavuz
University of California, Santa Barbara
syavuz@cs.ucsb.edu

Dilek Hakkani-Tür
Amazon Alexa AI
dilek@ieee.org

Jindong Chen
Google AI
jdchen@google.com

Ian Lane
Carnegie Mellon University
lane@cmu.edu

Abstract

Understanding and conversing about dynamic scenes is one of the key capabilities of AI agents that navigate the environment and convey useful information to humans. Video question answering is a specific scenario of such AI-human interaction where an agent generates a natural language response to a question regarding the video of a dynamic scene. Incorporating features from multiple modalities, which often provide supplementary information, is one of the challenging aspects of video question answering. Furthermore, a question often concerns only a small segment of the video, hence encoding the entire video sequence using a recurrent neural network is not computationally efficient. Our proposed question-guided video representation module efficiently generates the token-level video summary guided by each word in the question. The learned representations are then fused with the question to generate the answer. Through empirical evaluation on the Audio Visual Scene-aware Dialog (AVSD) dataset (Alamri et al., 2019a), our proposed models in single-turn and multi-turn question answering achieve state-of-the-art performance on several automatic natural language generation evaluation metrics.

1 Introduction

Nowadays dialogue systems are becoming more and more ubiquitous in our lives. It is essential for such systems to perceive the environment, gather data and convey useful information to humans in an accessible fashion. Video question answering (VideoQA) systems provide a convenient way for humans to acquire visual information about the environment. If a user wants to obtain information about a dynamic scene, one can simply ask the VideoQA system a question in natural language, and the system generates a natural-language answer. The task of a VideoQA dialogue system in

Figure 1: An example from the AVSD dataset. Each example contains a video and its associated question answering dialogue regarding the video scene.

this paper is described as follows. Given a video as grounding evidence, in each dialogue turn, the system is presented a question and is required to generate an answer in natural language. Figure 1 shows an example of multi-turn VideoQA. It is composed of a video clip and a dialogue, where the dialogue contains open-ended question answer pairs regarding the scene in the video. In order to answer the questions correctly, the system needs to be effective at understanding the question, the video and the dialogue context altogether.

Recent work on VideoQA has shown promising performance using multi-modal attention fusion for combination of features from different modalities (Xu et al., 2017; Zeng et al., 2017; Zhao et al., 2018; Gao et al., 2018). However, one of the challenges is that the length of the video sequence can be very long and the question may concern only

a small segment in the video. Therefore, it may be time inefficient to encode the entire video sequence using a recurrent neural network.

In this work, we present the question-guided video representation module which learns 1) to summarize the video frame features efficiently using an attention mechanism and 2) to perform feature selection through a gating mechanism. The learned question-guided video representation is a compact video summary for each token in the question. The video summary and question information are then fused to create multi-modal representations. The multi-modal representations and the dialogue context are then passed as input to a sequence-to-sequence model with attention to generate the answer (Section 3). We empirically demonstrate the effectiveness of the proposed methods using the AVSD dataset (Alamri et al., 2019a) for evaluation (Section 4). The experiments show that our model for single-turn VideoQA achieves state-of-the-art performance, and our multi-turn VideoQA model shows competitive performance, in comparison with existing approaches (Section 5).

2 Related Work

In the recent years, research on visual question answering has accelerated following the release of multiple publicly available datasets. These datasets include COCO-QA (Ren et al., 2015a), VQA (Agrawal et al., 2017), and Visual Madlibs (Yu et al., 2015) for image question answering and MovieQA (Tapaswi et al., 2016), TGIF-QA (Jang et al., 2017), and TVQA (Lei et al., 2018) for video question answering.

2.1 Image Question Answering

The goal of image question answering is to infer the correct answer, given a natural language question related to the visual content of an image. It assesses the system's capability of multi-modal understanding and reasoning regarding multiple aspects of humans and objects, such as their appearance, counting, relationships and interactions (Lei et al., 2018). State-of-the-art image question answering models make use of spatial attention to obtain a fixed length question-dependent embedded representation of the image, which is then combined with the question feature to predict the answer (Yang et al., 2016; Xu and Saenko, 2016; Kazemi and Elqursh, 2017; Anderson et al., 2018).

Dynamic memory (Kumar et al., 2016; Xiong et al., 2016) and co-attention mechanism (Lu et al., 2016; Ma et al., 2018) are also adopted to model sophisticated cross-modality interactions.

2.2 Video Question Answering

VideoQA is a more complex task. As a video is a sequence of images, it contains not only appearance information but also motion and transitions. Therefore, VideoQA requires spatial and temporal aggregation of image features to encode the video into a question-relevant representation. Hence, temporal frame-level attention is utilized to model the temporal dynamics, where frame-level attribute detection and unified video representation are learned jointly (Ye et al., 2017; Xu et al., 2017; Mun et al., 2017). Similarly, Lei et al. (2018) use Faster R-CNN (Ren et al., 2015b) trained with the Visual Genome (Krishna et al., 2017) dataset to detect object and attribute regions in each frame, which are used as input features to the question answering model. Previous works also adopt various forms of external memory (Sukhbaatar et al., 2015; Kumar et al., 2016; Graves et al., 2016) to store question information, which allows multiple iterations of question-conditioned inference on the video features (Na et al., 2017; Kim et al., 2017; Zeng et al., 2017; Gao et al., 2018; Chenyou Fan, 2019).

2.3 Video Question Answering Dialogue

Recently in DSTC7, Alamri et al. (2019a) introduce the Audio-Visual Scene-aware Dialog (AVSD) dataset for multi-turn VideoQA. In addition to the challenge of integrating the questions and the dynamic scene information, the dialogue system also needs to effectively incorporate the dialogue context for coreference resolution to fully understand the user's questions across turns. To this end, Alamri et al. (2019b) use two-stream inflated 3D ConvNet (I3D) model (Carreira and Zisserman, 2017) to extract spatiotemporal visual frame features (I3D-RGB features for RGB input and I3D-flow features for optical flow input), and propose the Naïve Fusion method to combine multi-modal inputs based on the hierarchical recurrent encoder (HRE) architecture (Das et al., 2017). Hori et al. (2018) extend the Naïve Fusion approach and propose the Attentional Fusion method which learns multi-modal attention weights to fuse features from different modalities. Zhuang et al. (2019) modify the Attentional Fu-
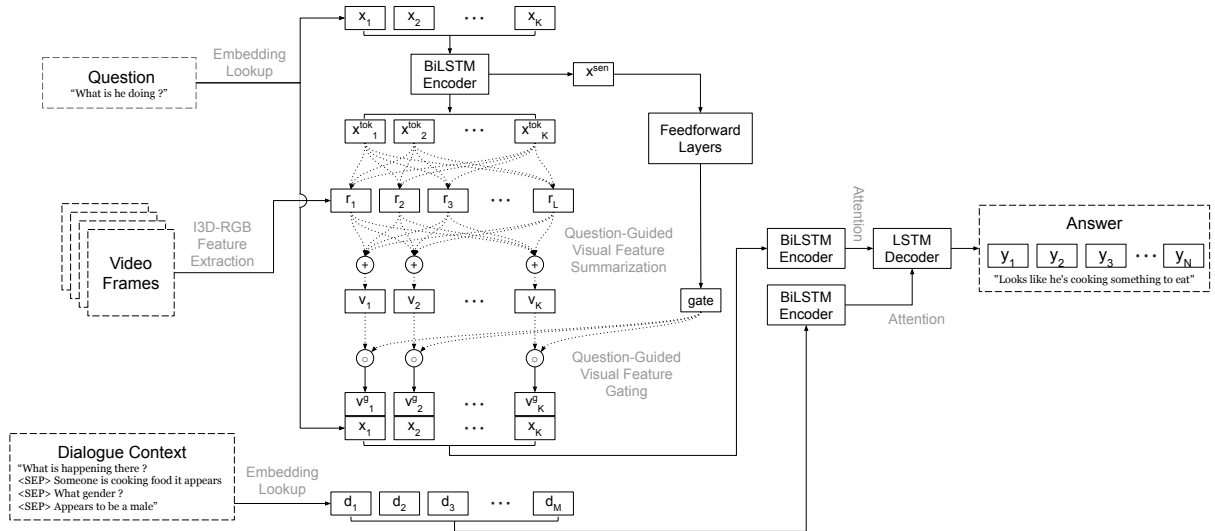
Figure 2: Overview of the proposed approach. First the I3D-RGB frame features are extracted. The question-guided video representation module takes as input the question sentence and the I3D-RGB features, generates a video representation for each token and applies gating using question as guidance. Then the question tokens are augmented by the per-token video representations and encoded by a bidirectional LSTM encoder. Similarly, the dialogue context is encoded by a bidirectional LSTM encoder. Finally, the LSTM answer decoder predicts the answer sequence.

sion method and propose to use Maximum Mutual Information (MMI) (Bahl et al., 1986) as the training objective. Besides the HRE architecture, the multi-source sequence-to-sequence (Multi-Source Seq2Seq) architecture with attention (Zoph and Knight, 2016; Firat et al., 2016) is also commonly applied (Pasunuru and Bansal, 2019; Kumar et al., 2019; Yeh et al., 2019). Previous works (Sanabria et al., 2019; Le et al., 2019; Pasunuru and Bansal, 2019) also explore various attention mechanisms to incorporate the different modal inputs, such as hierarchical attention (Libovický and Helcl, 2017) and cross attention (Seo et al., 2017). For modeling visual features, Lin et al. (2019) propose to use Dynamic memory networks (Kumar et al., 2016) and Nguyen et al. (2019) propose to use feature-wise linear modulation layers (Perez et al., 2018).

3 Approach

We formulate the multi-turn VideoQA task as follows. Given a sequence of raw video frames $\mathbf{f}$, the embedded question sentence $\mathbf{x} = \{x_1, \ldots, x_K\}$ and the single concatenated embedded sentence of the dialogue context $\mathbf{d} = \{d_1, \ldots, d_M\}$, the output is an answer sentence $\mathbf{y} = \{y_1, \ldots, y_N\}$.

The architecture of our proposed approach is illustrated in Figure 2. First the Video Frame Feature Extraction Module extracts the I3D-

RGB frame features from the video frames (Section 3.1). The Question-Guided Video Representation Module takes as input the embedded question sentence and the I3D-RGB features, and generates a compact video representation for each token in the question sentence (Section 3.2). In the Video-Augmented Question Encoder, the question tokens are first augmented by their corresponding per-token video representations and then encoded by a bidirectional LSTM (Section 3.3). Similarly, in the Dialogue Context Encoder, the dialogue context is encoded by a bidirectional LSTM (Section 3.4). Finally, in the Answer Decoder, the outputs from the Video-Augmented Question Encoder and the Dialogue Context Encoder are used as attention memory for the LSTM decoder to predict the answer sentence (Section 3.5). Our encoders and decoder work in the same way as the multi-source sequence-to-sequence models with attention (Zoph and Knight, 2016; Firat et al., 2016).

3.1 Video Frame Feature Extraction Module

In this work, we make use of the I3D-RGB frame features as the visual modality input, which are pre-extracted and provided in the AVSD dataset (Alamri et al., 2019a). Here we briefly describe the I3D-RGB feature extraction process, and we refer the readers to (Carreira and Zisser-

man, 2017) for more details of the I3D model. Two-stream Inflated 3D ConvNet (I3D) is a state-of-the-art action recognition model which operates on video inputs. The I3D model takes as input two streams of video frames: RGB frames and optical flow frames. The two streams are separately passed to a respective 3D ConvNet, which is inflated from 2D ConvNets to incorporate the temporal dimension. Two sequences of spatiotemporal features are produced by the respective 3D ConvNet, which are jointly used to predict the action class. The I3D-RGB features provided in the AVSD dataset are intermediate spatiotemporal representations from the "Mixed_5c" layer of the RGB stream's 3D ConvNet. The AVSD dataset uses the I3D model parameters pre-trained on the Kinetics dataset (Kay et al., 2017). To reduce the number of parameters in our model, we use a trainable linear projection layer to reduce the dimensionality of I3D-RGB features from 2048 to 256. Extracted from the video frames $\mathbf{f}$ and projected to a lower dimension, the sequence of dimension-reduced I3D-RGB frame features are denoted by $\mathbf{r} = \{r_1, \ldots, r_L\}$, where $r_i \in \mathrm{R}^{256}, \forall i$.

3.2 Question-Guided Video Representation Module

We use a bidirectional LSTM network to encode the sequence of question token embedding $\mathbf{x} = \{x_1, \ldots, x_K\}$. The token-level intermediate representations are denoted by $\mathbf{x}^{\mathrm{tok}} = \{x_1^{\mathrm{tok}}, \ldots, x_K^{\mathrm{tok}}\}$, and the embedded representation of the entire question is denoted by x^{sen}. These outputs will be used to guide the video representation.

$$\vec{h}_0 = \overleftarrow{h}_{K+1} = \mathbf{0} \tag{1}$$

$$\vec{h}_k = \mathrm{LSTM}_{\mathrm{guide}}^{\mathrm{forw}}(x_k, \vec{h}_{k-1}) \tag{2}$$

$$\overleftarrow{h}_k = \mathrm{LSTM}_{\mathrm{guide}}^{\mathrm{back}}(x_k, \overleftarrow{h}_{k+1}) \tag{3}$$

$$x_k^{\mathrm{tok}} = \vec{h}_k \oplus \overleftarrow{h}_k \tag{4}$$

$$\forall k \in \{1, \ldots, K\}$$

$$x^{\mathrm{sen}} = \vec{h}_K \oplus \overleftarrow{h}_1 \tag{5}$$

where $\oplus$ denotes vector concatenation; $\vec{\mathbf{h}}$ and $\overleftarrow{\mathbf{h}}$ represent the local forward and backward LSTM hidden states.

3.2.1 Per-Token Visual Feature Summarization

Generally the sequence length of the video frame features is quite large, as shown in Table 1. There-

fore it is not computationally efficient to encode the video features using a recurrent neural network. We propose to use the attention mechanism to generate a context vector to efficiently summarize the I3D-RGB features. We use the trilinear function (Seo et al., 2017) as a similarity measure to identify the frames most similar to the question tokens. For each question token x_k, we compute the similarity scores of its encoded representation x_k^{tok} with each of the I3D-RGB features $\mathbf{r}$. The similarity scores $\mathbf{s}_k$ are converted to an attention distribution $\mathbf{w}_k^{\mathrm{att}}$ over the I3D-RGB features by the softmax function. And the video summary v_k corresponding to the question token x_k is defined as the attention weighted linear combination of the I3D-RGB features. We also explored using dot product for computing similarity and empirically found out it yields suboptimal results.

$$s_{k,l} = \mathrm{trilinear}(x_k^{\mathrm{tok}}, r_l) \tag{6}$$

$$= W_{\mathrm{sim}}[x_k^{\mathrm{tok}} \oplus r_l \oplus (x_k^{\mathrm{tok}} \odot r_l)] \tag{7}$$

$$\forall l \in \{1, \ldots, L\}$$

$$\mathbf{w}_k^{\mathrm{att}} = \mathrm{softmax}(\mathbf{s}_k) \tag{8}$$

$$v_k = \sum_{l=1}^{L} w_{k,l}^{\mathrm{att}} \, r_l \tag{9}$$

$$\forall k \in \{1, \ldots, K\}$$

where $\odot$ denotes element-wise multiplication, and W_{sim} is a trainable variable.

3.2.2 Visual Feature Gating

Not all details in the video are important for answering a question. Attention helps in discarding the unimportant frames in the time dimension. We propose a gating mechanism which enables us to perform feature selection within each frame. We project the sentence-level question representation x^{sen} through fully-connected layers with ReLU nonlinearity to generate a gate vector g. For each question token x_k, its corresponding video summary v_k is then multiplied element-wise with the gate vector g to generate a gated visual summary v_k^{g}. We also experimented applying gating on the dimension-reduced I3D-RGB features $\mathbf{r}$, prior to the per-token visual feature summarization step, but it resulted in an inferior performance.

$$g = \mathrm{sigmoid}(W_{\mathrm{g},1}(\mathrm{ReLU}(W_{\mathrm{g},2}x^{\mathrm{sen}} + b_{\mathrm{g},2}) + b_{\mathrm{g},1}) \tag{10}$$

$$v_k^g = v_k \odot g \tag{11}$$

$$\forall k \in \{1, \ldots, K\}$$

where $W_{g,1}$, $b_{g,1}$, $W_{g,2}$, $b_{g,2}$ are trainable variables.

3.3 Video-Augmented Question Encoder

Given the sequence of per-token gated visual summary $\mathbf{v}^{g} = \{v_1^{g}, \ldots, v_K^{g}\}$, we augment the question features by concatenating the embedded question tokens $\mathbf{x} = \{x_1, \ldots, x_K\}$ with their associated per-token video summary. The augmented question features are then encoded using a bidirectional LSTM. The token-level video-augmented question features are denoted by $\mathbf{q}^{tok} = \{q_1^{tok}, \ldots, q_K^{tok}\}$, and the sentence-level feature is denoted by q^{sen}.

$$\vec{h}_0 = \overleftarrow{h}_{K+1} = \mathbf{0} \tag{12}$$

$$\vec{h}_k = \text{LSTM}_{\text{ques}}^{\text{forw}}(x_k \oplus v_k^{g}, \vec{h}_{k-1}) \tag{13}$$

$$\overleftarrow{h}_k = \text{LSTM}_{\text{ques}}^{\text{back}}(x_k \oplus v_k^{g}, \overleftarrow{h}_{k+1}) \tag{14}$$

$$q_k^{tok} = \vec{h}_k \oplus \overleftarrow{h}_k \tag{15}$$

$$\forall k \in \{1, \ldots, K\}$$

$$q^{sen} = \vec{h}_K \oplus \overleftarrow{h}_1 \tag{16}$$

where $\vec{h}$ and $\overleftarrow{h}$ represent the local forward and backward LSTM hidden states.

3.4 Dialogue Context Encoder

Similar to the video-augmented question encoder, we encode the embedded dialogue context tokens $\mathbf{d} = \{d_1, \ldots, d_M\}$ using a bidirectional LSTM. The embedded token-level representations are denoted by $\mathbf{d}^{tok} = \{d_1^{tok}, \ldots, d_M^{tok}\}$.

$$\vec{h}_0 = \overleftarrow{h}_{M+1} = \mathbf{0} \tag{17}$$

$$\vec{h}_m = \text{LSTM}_{\text{dial}}^{\text{forw}}(d_m, \vec{h}_{m-1}) \tag{18}$$

$$\overleftarrow{h}_m = \text{LSTM}_{\text{dial}}^{\text{back}}(d_m, \overleftarrow{h}_{m+1}) \tag{19}$$

$$d_m^{tok} = \vec{h}_m \oplus \overleftarrow{h}_m \tag{20}$$

$$\forall m \in \{1, \ldots, M\}$$

where $\vec{h}$ and $\overleftarrow{h}$ represent the local forward and backward LSTM hidden states.

3.5 Answer Decoder

The final states of the forward and backward LSTM units of the question encoder are used to initialize the state of answer decoder. Let y_n be the output of the decoder at step n, where $1 \leq n \leq N$, y_0 be the special start of sentence token and y_n^{emb} be the embedded representation of y_n. At a decoder step n, the previous decoder hidden state

h_{n-1} is used to attend over $\mathbf{q}^{tok}$ and $\mathbf{d}^{tok}$ to get the attention vectors $h_n^{att,q}$ and $h_n^{att,d}$ respectively. These two vectors retrieve the relevant features from the intermediate representations of the video-augmented question encoder and the dialogue context encoder, both of which are useful for generating the next token of the answer. At each decoder step, the decoder hidden state h_n is used to generate a distribution over the vocabulary. The decoder output y_n^* is defined to be $\text{argmax}_{y_n} p(y_n|y_{\leq n-1})$.

$$h_0 = q^{sen} \tag{21}$$

$$s_{n,k}^{q} = v_{\text{ans,q}}^{\top} \tanh(W_{\text{ans,q}}[h_{n-1} \oplus q_k^{tok}]) \tag{22}$$

$$\forall k \in \{1, \ldots, K\}$$

$$\mathbf{w}_n^{q} = \text{softmax}(\mathbf{s}_n^{q}) \tag{23}$$

$$h_n^{att,q} = \sum_{k=1}^{K} w_{n,k}^{q} q_k^{tok} \tag{24}$$

$$s_{n,m}^{d} = v_{\text{ans,d}}^{\top} \tanh(W_{\text{ans,d}}[h_{n-1} \oplus d_m^{tok}]) \tag{25}$$

$$\forall m \in \{1, \ldots, M\}$$

$$\mathbf{w}_n^{d} = \text{softmax}(\mathbf{s}_n^{d}) \tag{26}$$

$$h_n^{att,d} = \sum_{m=1}^{M} w_{n,m}^{d} d_m^{tok} \tag{27}$$

$$h_n = \text{LSTM}_{\text{ans}}(y_{n-1}^{emb}, [h_n^{att,q} \oplus h_n^{att,d} \oplus h_{n-1}]) \tag{28}$$

$$p(y_n|y_{\leq n-1}) = \text{softmax}(W_{\text{ans}}h_n + b_{\text{ans}}) \tag{29}$$

$$\forall n \in \{1, \ldots, N\}$$

where $\mathbf{h}$ represents the local LSTM hidden states, and $W_{\text{ans,q}}$, $W_{\text{ans,d}}$, W_{ans}, b_{ans} are trainable variables.

4 Experiments

4.1 Dataset

We consider the Audio-Visual Scene-aware Dialog (AVSD) dataset (Alamri et al., 2019a) for evaluating our proposed model in single-turn and multi-turn VideoQA. We use the official release of train set for training, and the public (*i.e.,* prototype) validation and test sets for inference. The AVSD dataset is a collection of text-based human-human question answering dialogues based on the video clips from the CHARADES dataset (Sigurdsson et al., 2016). The CHARADES dataset contains video clips of daily indoor human activities, originally purposed for research in video activity classification and localization. Along with

	Training	Validation	Test
# of dialogues	7659	732	733
# of turns	153,180	14,680	14,660
# of words	1,450,754	138,314	138,790
Avg. length of question (K)	8.5	8.4	8.5
Avg. length of I3D-RGB (L)	179.2	173.0	171.3

Table 1: Data statistics of the AVSD dataset. We use the official training set, and the public (*i.e.,* prototype) validation and test sets. We also present the average length of the question token sequences and the I3D-RGB frame feature sequences to highlight the importance of time efficient video encoding without using a recurrent neural network. The sequence lengths of the questions and I3D-RGB frame features are denoted by K and L respectively in the model description (Section 3).

the video clips and associated question answering dialogues, the AVSD dataset also provides the pre-extracted I3D-RGB visual frame features using a pre-trained two-stream inflated 3D ConvNet (I3D) model (Carreira and Zisserman, 2017). The pre-trained I3D model was trained on the Kinetics dataset (Kay et al., 2017) for human action recognition.

In Table 1, we present the statistics of the AVSD dataset. Given the fact that the lengths of the I3D-RGB frame feature sequences are more than 20 times longer than the questions, using a recurrent neural network to encode the visual feature sequences will be very time consuming, as the visual frames are processed sequentially. Our proposed question-guided video representation module summarizes the video sequence efficiently - aggregating the visual features by question-guided attention and weighted summation and performing gating with a question-guided gate vector, both of which can be done in parallel across all frames.

4.2 Experimental Setup

We implement our models using the Tensor2Tensor framework (Vaswani et al., 2018). The question and dialogue context tokens are both embedded with the same randomly-initialized word embedding matrix, which is also shared with the answer decoder's output embedding. The dimension of the word embedding is 256, the same dimension to which the I3D-RGB features are transformed. All of our LSTM encoders and decoder

have 1 hidden layer. Bahdanau attention mechanism (Bahdanau et al., 2015) is used in the answer decoder. During training, we apply dropout rate 0.2 in the encoder and decoder cells. We use the ADAM optimizer (Kingma and Ba, 2015) with $\alpha = 2 \times 10^{-4}, \beta_1 = 0.85, \beta_2 = 0.997, \epsilon = 10^{-6}$, and clip the gradient with L2 norm threshold 2.0 (Pascanu et al., 2013). The models are trained up to 100K steps with early stopping on the validation BLEU-4 score using batch size 1024 on a single GPU. During inference, we use beam search decoding with beam width 3. We experimented with word embedding dimension {256, 512}, dropout rate {0, 0.2}, Luong and Bahdanau attention mechanisms, {1, 2} hidden layer(s) for both encoders and the decoder. We found the aforementioned setting worked best for most models.

5 Results

5.1 Comparison with Existing Methods

We evaluate our proposed approach using the same natural language generation evaluation toolkit NLGEval (Sharma et al., 2017) as the previous approaches. The corpus-wide scores of the following unsupervised automated metrics are reported, including BLEU-1 through BLEU-4 (Papineni et al., 2002), METEOR (Banerjee and Lavie, 2005), ROUGE-L (Lin and Och, 2004) and CIDEr (Vedantam et al., 2015). The results of our models in comparison with the previous approaches are shown in Table 2. We report the mean and standard deviation scores of 5 runs using random initialization and early stopping on the public (prototype) validation set. We apply our model in two scenarios: single-turn and multi-turn VideoQA. The only difference is that in single-turn VideoQA, the dialogue context encoder is excluded from the model.

First we observe that our proposed multi-turn VideoQA model significantly outperforms the single-turn VideoQA model. This suggests that the additional dialogue context input can provide supplementary information from the question and visual features, and thus is helpful for generating the correct answer. Secondly, comparing the single-turn VideoQA models, our approach outperforms the existing approaches across all automatic evaluation metrics. This suggests the effectiveness of our proposed question-guided video representations for VideoQA. When comparing

Single-Turn VideoQA Models	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDEr
Naïve Fusion	27.7	17.5	11.8	8.3	11.7	28.8	74.0
Multi-source Seq2Seq	-	-	-	8.83	12.43	34.23	95.54
Ours	**29.56**±0.75	**18.60**±0.49	**13.16**±0.33	**9.77**±0.21	**13.19**±0.20	**34.29**±0.19	**101.75**±1.03

Multi-Turn VideoQA Models	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDEr
Naïve Fusion	27.7	17.6	12.0	8.5	11.8	29.0	76.5
Attentional Fusion	27.6	17.7	12.2	8.7	11.7	29.3	78.7
Modified Attn. Fusion	27.7	17.6	12.0	8.5	11.8	29.0	76.5
+MMI objective	28.3	18.1	12.4	8.9	12.1	29.6	80.5
Hierarchical Attention	29.1	18.6	12.6	9.0	12.7	30.1	82.4
+pre-trained embedding	**30.7**	**20.4**	14.4	10.6	13.6	32.0	99.5
Multi-Source Seq2Seq	-	-	-	10.58	**14.13**	36.54	105.39
Ours	30.52±0.34	20.00±0.20	**14.46**±0.14	**10.93**±0.11	13.87±0.10	**36.62**±0.23	**113.28**±1.37

Table 2: Comparison with existing approaches: Naïve Fusion (Alamri et al., 2019b; Zhuang et al., 2019), Attentional Fusion (Hori et al., 2018; Zhuang et al., 2019), Multi-Source Sequence-to-Sequence model (Pasunuru and Bansal, 2019), Modified Attentional Fusion with Maximum Mutual Information objective (Zhuang et al., 2019) and Hierarchical Attention with pre-trained embedding (Le et al., 2019), on the AVSD public test set. For each approach, we report its corpus-wide scores on BLEU-1 through BLEU-4, METEOR, ROUGE-L and CIDEr. We report the mean and standard deviation scores of 5 runs using random initialization and early stopping on the public (prototype) validation set.

Model	BLEU-4	METEOR	ROUGE-L	CIDEr
Ours	10.94	13.73	36.30	111.12
-TokSumm	10.46	13.49	35.81	110.08
-Gating	10.59	13.64	36.11	108.51
-TokSumm-Gating	10.06	13.20	35.35	104.01

Table 3: Ablation study on the AVSD validation set. We observe that the performance degrades when either of both of the question-guided per-token visual feature summarization (TokSumm) and feature gating (Gating) techniques are removed.

with previous multi-turn VideoQA models, our approach that uses the dialogue context (questions and answers in previous turns) yields state-of-the-art performance on the BLEU-3, BLEU-4, ROUGE-L and CIDEr metrics and competitive results on BLEU-1, BLEU-2 and METEOR. It is worth mentioning that our model does not use pre-trained word embedding or audio features as in the previous hierarchical attention approach (Le et al., 2019).

5.2 Ablation Study and Weights Visualization

We perform ablation experiments on the validation set in the multi-turn VideoQA scenario to analyze the effectiveness of the two techniques in the question-guided video representation module. The results are shown in Table 3.

5.2.1 Question-Guided Per-Token Visual Feature Summarization (TokSumm)

Instead of using token-level question representations $\mathbf{x}^{\text{tok}} = \{x_1^{\text{tok}}, \ldots, x_K^{\text{tok}}\}$ to generate per-token video summary $\mathbf{v} = \{v_1, \ldots, v_K\}$, we experiment with using the sentence-level representation of the question x^{sen} as the query vector to attend over the I3D-RGB visual features to create a visual summary v, and use v to augment each of the question tokens in the video-augmented question encoder.

$$s_l = \text{trilinear}(x^{\text{sen}}, r_l) \qquad (30)$$
$$\forall l \in \{1, \ldots, L\}$$
$$\mathbf{w}^{\text{att}} = \text{softmax}(\mathbf{s}) \qquad (31)$$
$$v = \sum_{l=1}^{L} w_l^{\text{att}} r_l \qquad (32)$$

We observe the performance degrades when the sentence-level video summary is used instead of the token-level video summary.

Figure 3 shows an example of the attention weights in the question-guided per-token visual feature summarization. We can see that for different question tokens, the attention weights are shifted to focus on the different segment in the sequence of the video frame features.

5.2.2 Question-Guided Visual Feature Gating (Gating)

We also experiment with using the non-gated token-level video summary $\mathbf{v} = \{v_1, \ldots, v_K\}$ to
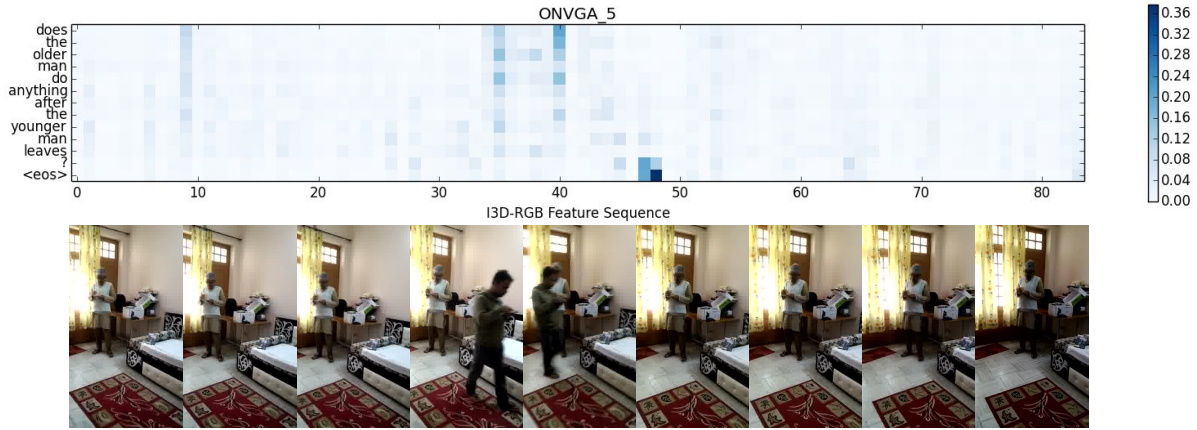
Figure 3: Question-guided per-token visual feature summary weights on a question. Each row represents the attention weights $\mathbf{w}_k^{\mathrm{att}}$ of the corresponding encoded question token x_k^{tok} over the I3D-RGB visual features. We can observe that the attention weights are shifted to focus on the relevant segment of the visual frame features for the question tokens "after the younger man leaves <eos>?"
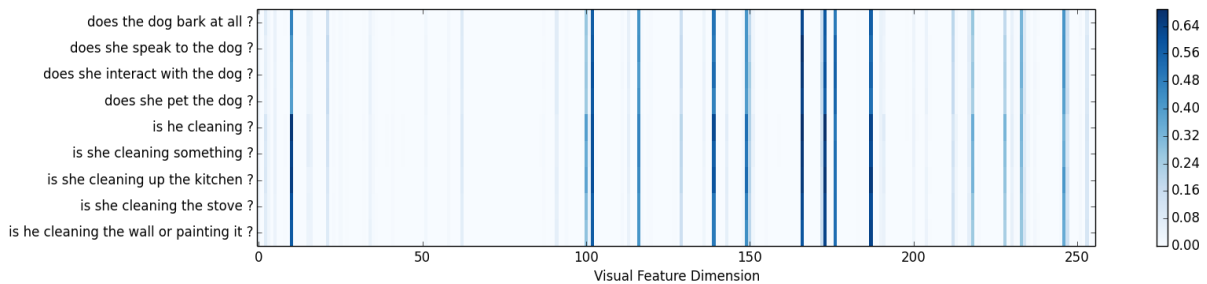

Figure 4: Question-guided gate weights g for some example questions. Across the questions about similar subjects, we observe a similar trend of weight distribution over visual feature dimensions. Conversely, questions about different topics show different gate weights patterns.

augment the question information in the video-augmented question encoder. We observe the model's performance declines when the question-guided gating is not applied on the video summary feature. Removing both the per-token visual feature summarization and the gating mechanism results in further degradation in the model performance.

Figure 4 illustrates the question-guided gate weights g of several example questions. We observe that the gate vectors corresponding to the questions regarding similar subjects assign weights on similar dimensions of the visual feature. Although many of the visual feature dimensions have low weights across different questions, the feature dimensions of higher gate weights still exhibit certain topic-specific patterns.

6 Conclusion and Future Work

In this paper, we present an end-to-end trainable model for single-turn and multi-turn VideoQA.

Our proposed framework takes the question, I3D-RGB video frame features and dialogue context as input. Using the question information as guidance, the video features are summarized as compact representations to augment the question information, which are jointly used with the dialogue context to generate a natural language answer to the question. Specifically, our proposed question-guided video representation module is able to summarize the video features efficiently for each question token using an attention mechanism and perform feature selection through a gating mechanism. In empirical evaluation, our proposed models for single-turn and multi-turn VideoQA outperform existing approaches on several automatic natural language generation evaluation metrics. Detailed analyses are performed, and it is shown that our model effectively attends to relevant frames in the video feature sequence for summarization, and the gating mechanism shows topic-specific patterns in the feature dimension selection within a frame. In future work, we plan

to extend the models to incorporate audio features and experiment with more advanced techniques to incorporate the dialogue context with the question and video information, such as hierarchical attention and co-attention mechanisms. We also plan to employ our model on TVQA, a larger scale VideoQA dataset.

References

Aishwarya Agrawal, Jiasen Lu, Stanislaw Antol, Margaret Mitchell, C Lawrence Zitnick, Devi Parikh, and Dhruv Batra. 2017. VQA: Visual question answering. *International Journal of Computer Vision (IJCV)*.

Huda Alamri, Vincent Cartillier, Abhishek Das, Jue Wang, Stefan Lee, Peter Anderson, Irfan Essa, Devi Parikh, Dhruv Batra, Anoop Cherian, Tim K. Marks, and Chiori Hori. 2019a. Audio visual scene-aware dialog. In *Computer Vision and Pattern Recognition (CVPR)*.

Huda Alamri, Chiori Hori, Tim K. Marks, Dhruv Batra, and Devi Parikh. 2019b. Audio visual scene-aware dialog (avsd) track for natural language generation in dstc7. In *DSTC7 at AAAI 2019 Workshop*.

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *Computer Vision and Pattern Recognition (CVPR)*.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *International Conference on Learning Representations (ICLR)*.

Lalit R Bahl, Peter F Brown, Peter V De Souza, and Robert L Mercer. 1986. Maximum mutual information estimation of hidden markov model parameters for speech recognition. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *ACL workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*.

Joao Carreira and Andrew Zisserman. 2017. Quo vadis, action recognition? a new model and the Kinetics dataset. In *Computer Vision and Pattern Recognition (CVPR)*.

Shu Zhang Wensheng Wang Chi Zhang Heng Huang Chenyou Fan, Xiaofan Zhang. 2019. Heterogeneous memory enhanced multimodal attention model for video question answering. In *Computer Vision and Pattern Recognition (CVPR)*.

Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José MF Moura, Devi Parikh, and Dhruv Batra. 2017. Visual dialog. In *Computer Vision and Pattern Recognition (CVPR)*.

Orhan Firat, Kyunghyun Cho, and Yoshua Bengio. 2016. Multi-way, multilingual neural machine translation with a shared attention mechanism. In *Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*.

Jiyang Gao, Runzhou Ge, Kan Chen, and Ram Nevatia. 2018. Motion-appearance co-memory networks for video question answering. In *Computer Vision and Pattern Recognition (CVPR)*.

Alex Graves, Greg Wayne, Malcolm Reynolds, Tim Harley, Ivo Danihelka, Agnieszka Grabska-Barwińska, Sergio Gómez Colmenarejo, Edward Grefenstette, Tiago Ramalho, John Agapiou, et al. 2016. Hybrid computing using a neural network with dynamic external memory. *Nature*.

Chiori Hori, Huda Alamri, Jue Wang, Gordon Winchern, Takaaki Hori, Anoop Cherian, Tim K Marks, Vincent Cartillier, Raphael Gontijo Lopes, Abhishek Das, et al. 2018. End-to-end audio visual scene-aware dialog using multimodal attention-based video features. *Computing Research Repository*, arXiv:1806.08409.

Yunseok Jang, Yale Song, Youngjae Yu, Youngjin Kim, and Gunhee Kim. 2017. TGIF-QA: Toward spatio-temporal reasoning in visual question answering. In *Computer Vision and Pattern Recognition (CVPR)*.

Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. 2017. The kinetics human action video dataset. *Computing Research Repository*, arXiv:1705.06950.

Vahid Kazemi and Ali Elqursh. 2017. Show, ask, attend, and answer: A strong baseline for visual question answering. *Computing Research Repository*, arXiv:1704.03162.

Kyung-Min Kim, Min-Oh Heo, Seong-Ho Choi, and Byoung-Tak Zhang. 2017. DeepStory: video story QA by deep embedded memory networks. In *International Joint Conferences on Artificial Intelligence (IJCAI)*.

D Kingma and Jimmy Ba. 2015. A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. 2017. Visual Genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision (IJCV)*.

Ankit Kumar, Ozan Irsoy, Peter Ondruska, Mohit Iyyer, James Bradbury, Ishaan Gulrajani, Victor Zhong, Romain Paulus, and Richard Socher. 2016. Ask me anything: Dynamic memory networks for natural language processing. In *International Conference on Machine Learning (ICML)*.

Shachi H Kumar, Eda Okur, Saurav Sahay, Juan Jose Alvarado Leanos, Jonathan Huang, and Lama Nachman. 2019. Context, attention and audio feature explorations for audio visual scene-aware dialog. In *DSTC7 at AAAI 2019 workshop*.

Hung Le, S Hoi, Doyen Sahoo, and N Chen. 2019. End-to-end multimodal dialog systems with hierarchical multimodal attention on video features. In *DSTC7 at AAAI 2019 workshop*.

Jie Lei, Licheng Yu, Mohit Bansal, and Tamara Berg. 2018. TVQA: Localized, compositional video question answering. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Jindřich Libovickỳ and Jindřich Helcl. 2017. Attention strategies for multi-source sequence-to-sequence learning. In *Annual Meeting of the Association for Computational Linguistics (ACL)*.

Chin-Yew Lin and Franz Josef Och. 2004. Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics. In *Annual Meeting of the Association for Computational Linguistics (ACL)*.

Kuan-Yen Lin, Chao-Chun Hsu, Yun-Nung Chen, and Lun-Wei Ku. 2019. Entropy-enhanced multimodal attention model for scene-aware dialogue generation. In *DSTC7 at AAAI 2019 workshop*.

Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. 2016. Hierarchical question-image co-attention for visual question answering. In *Advances in Neural Information Processing Systems*.

Chao Ma, Chunhua Shen, Anthony Dick, Qi Wu, Peng Wang, Anton van den Hengel, and Ian Reid. 2018. Visual question answering with memory-augmented networks. In *Computer Vision and Pattern Recognition (CVPR)*.

Jonghwan Mun, Paul Hongsuck Seo, Ilchae Jung, and Bohyung Han. 2017. MarioQA: Answering questions by watching gameplay videos. In *International Conference on Computer Vision (ICCV)*.

Seil Na, Sangho Lee, Jisung Kim, and Gunhee Kim. 2017. A read-write memory network for movie story understanding. In *International Conference on Computer Vision (ICCV)*.

Dat Tien Nguyen, Shikhar Sharma, Hannes Schulz, and Layla El Asri. 2019. From film to video: Multi-turn question answering with multi-modal context. In *DSTC7 at AAAI 2019 workshop*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A method for automatic evaluation of machine translation. In *Annual Meeting of the Association for Computational Linguistics (ACL)*.

Razvan Pascanu, Tomas Mikolov, and Yoshua Bengio. 2013. On the difficulty of training recurrent neural networks. In *International Conference on Machine Learning (ICML)*.

Ramakanth Pasunuru and Mohit Bansal. 2019. Dstc7-avsd: Scene-aware video-dialogue systems with dual attention. In *DSTC7 at AAAI 2019 workshop*.

Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. 2018. FiLM: Visual reasoning with a general conditioning layer. In *AAAI Conference on Artificial Intelligence (AAAI)*.

Mengye Ren, Ryan Kiros, and Richard Zemel. 2015a. Exploring models and data for image question answering. In *Advances in Neural Information Processing Systems*.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015b. Faster R-CNN: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems*.

Ramon Sanabria, Shruti Palaskar, and Florian Metze. 2019. Cmu sinbads submission for the dstc7 avsd challenge. In *DSTC7 at AAAI 2019 workshop*.

Minjoon Seo, Aniruddha Kembhavi, Ali Farhadi, and Hannaneh Hajishirzi. 2017. Bidirectional attention flow for machine comprehension. In *International Conference on Learning Representations (ICLR)*.

Shikhar Sharma, Layla El Asri, Hannes Schulz, and Jeremie Zumer. 2017. Relevance of unsupervised metrics in task-oriented dialogue for evaluating natural language generation. *Computing Research Repository*, arXiv:1706.09799.

Gunnar A Sigurdsson, Gül Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. 2016. Hollywood in homes: Crowdsourcing data collection for activity understanding. In *European Conference on Computer Vision (ECCV)*.

Sainbayar Sukhbaatar, Jason Weston, Rob Fergus, et al. 2015. End-to-end memory networks. In *Advances in Neural Information Processing Systems*.

Makarand Tapaswi, Yukun Zhu, Rainer Stiefelhagen, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. 2016. MovieQA: Understanding stories in movies through question-answering. In *Computer Vision and Pattern Recognition (CVPR)*.

Ashish Vaswani, Samy Bengio, Eugene Brevdo, Francois Chollet, Aidan Gomez, Stephan Gouws, Llion Jones, Łukasz Kaiser, Nal Kalchbrenner, Niki Parmar, et al. 2018. Tensor2Tensor for neural machine translation. In *Conference of the Association for Machine Translation in the Americas*.

Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2015. CIDEr: Consensus-based image description evaluation. In *Computer Vision and Pattern Recognition (CVPR)*.

Caiming Xiong, Stephen Merity, and Richard Socher. 2016. Dynamic memory networks for visual and textual question answering. In *International Conference on Machine Learning (ICML)*.

Dejing Xu, Zhou Zhao, Jun Xiao, Fei Wu, Hanwang Zhang, Xiangnan He, and Yueting Zhuang. 2017. Video question answering via gradually refined attention over appearance and motion. In *International Conference on Multimedia*.

Huijuan Xu and Kate Saenko. 2016. Ask, attend and answer: Exploring question-guided spatial attention for visual question answering. In *European Conference on Computer Vision (ECCV)*.

Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, and Alex Smola. 2016. Stacked attention networks for image question answering. In *Computer Vision and Pattern Recognition (CVPR)*.

Yunan Ye, Zhou Zhao, Yimeng Li, Long Chen, Jun Xiao, and Yueting Zhuang. 2017. Video question answering via attribute-augmented attention network learning. In *SIGIR Conference on Research and Development in Information Retrieval*.

Yi-Ting Yeh, Tzu-Chuan Lin, Hsiao-Hua Cheng, Yi-Hsuan Deng, Shang-Yu Su, and Yun-Nung Chen. 2019. Reactive multi-stage feature fusion for multimodal dialogue modeling. In *DSTC7 at AAAI 2019 Workshop*.

Licheng Yu, Eunbyung Park, Alexander C Berg, and Tamara L Berg. 2015. Visual Madlibs: Fill in the blank description generation and question answering. In *International Conference on Computer Vision (ICCV)*.

Kuo-Hao Zeng, Tseng-Hung Chen, Ching-Yao Chuang, Yuan-Hong Liao, Juan Carlos Niebles, and Min Sun. 2017. Leveraging video descriptions to learn video question answering. In *AAAI Conference on Artificial Intelligence (AAAI)*.

Zhou Zhao, Zhu Zhang, Shuwen Xiao, Zhou Yu, Jun Yu, Deng Cai, Fei Wu, and Yueting Zhuang. 2018. Open-ended long-form video question answering via adaptive hierarchical reinforced networks. In *International Joint Conferences on Artificial Intelligence (IJCAI)*.

Bairong Zhuang, Wenbo Wang, and Takahiro Shinozaki. 2019. Investigation of attention-based multimodal fusion and maximum mutual information objective for dstc7 track3. In *DSTC7 at AAAI 2019 workshop*.

Barret Zoph and Kevin Knight. 2016. Multi-source neural translation. In *Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*.