Knowledge-Constrained Answer Generation for Open-Ended Video Question Answering

Yao Jin[1]*, Guocheng Niu[2], Xinyan Xiao[2], Jian Zhang[3], Xi Peng[4], Jun Yu[1]†

[1]Hangzhou Dianzi University
[2]Baidu Inc.
[3]Zhejiang International Studies University
[4]College of Computer Science, Sichuan Univerisity
jy.struggling.cu@gmail.com, niuguocheng@baidu.com, xiaoxinyan@baidu.com, jeyzhang@outlook.com, pengx.gm@gmail.com, yujun@hdu.edu.cn

Abstract

Open-ended Video question answering (open-ended VideoQA) aims to understand video content and question semantics to generate the correct answers. Most of the best performing models define the problem as a discriminative task of multi-label classification. In real-world scenarios, however, it is difficult to define a candidate set that includes all possible answers. In this paper, we propose a Knowledge-constrained Generative VideoQA Algorithm (KcGA) with an encoder-decoder pipeline, which enables out-of-domain answer generation through an adaptive external knowledge module and a multi-stream information control mechanism. We use ClipBERT to extract the video-question features, extract framewise object-level external knowledge from a commonsense knowledge base and compute the contextual-aware episode memory units via an attention based GRU to form the external knowledge features, and exploit multi-stream information control mechanism to fuse video-question and external knowledge features such that the semantic complementation and alignment are well achieved. We evaluate our model on two open-ended benchmark datasets to demonstrate that we can effectively and robustly generate high-quality answers without restrictions of training data.

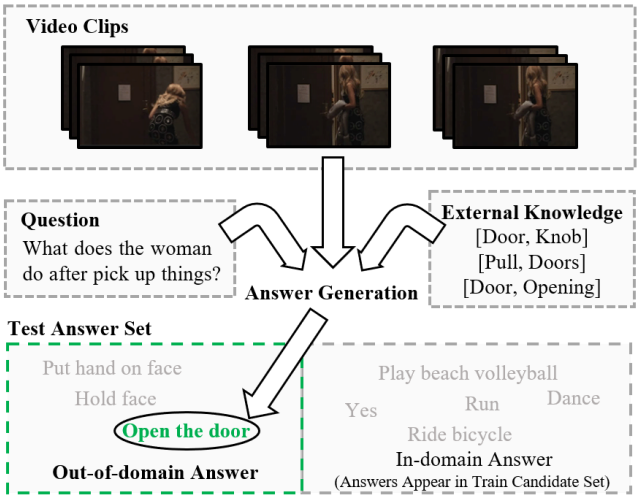

Figure 1: An illustration of out-of-domain answer generation. Our model generates out-of-domain answers (answers that do not appear in the training set) by combining video clips, question features and external knowledge features.

.

Introduction

Open-ended Video Question Answering (open-ended VideoQA) (Fan et al. 2019) means generating the answer according to a given video and question from scratch without having to choose from several pre-supplied answers or fill the missing part of an incomplete answer. To achieve the freeform answer generation, open-ended VideoQA should not only understand the video and question well, but also conduct comprehensive cross-modal reasoning. Therefore, open-ended VideoQA is more challenging compared with other types of VideoQA. Many well-performed works treat the open-ended VideoQA as a multi-label classification task and solve it using attention mechanisms (Gao et al. 2018; Li et al. 2019b; Gao et al. 2019), graph networks (Cherian et al. 2022; Park, Lee, and Sohn 2021; Jiang and Han 2020; Huang et al. 2020) or causal analysis (Li et al. 2022). This requires to form a candidate set using answers with the top-k occurrence frequency, and then infer the answer from the candidate set. Since the candidate set consists of only answers appearing in the training set, these methods can only predict in-domain answers, but cannot predict out-of-domain answers (answers that do not appear in the training set) at all. Furthermore, due to the structural characteristics of the answers in the training set, the candidate set basically contains very shot answers with length one or exceptionally two. Hence, the predicted answers usually lack rich semantics and details. In contrast, freely generated out-of-domain answers often contain long and semantically complete answers with more details. Motivated by this, some works (Xue, Zhao, and Cai 2017; Zhao et al. 2018b; Lee et al. 2021) tackle the open-ended VideoQA in a

*Work is done during an internship at Baidu Inc.
†Corresponding author

generative manner. They rely on not the candidate set but the entire training set, which nevertheless may still unable to provide sufficient semantic for generating high quality out-of-domain answers.

Knowledge is defined as high-level awareness and understanding of the input information and its surroundings by (Yu et al. 2022). It contains commonly recognized facts and regulations that are known as common sense, which benefits human as well as neural networks in learning, communicating or reasoning. We deem knowledge can provide key information to out-of-domain answer generation with comprehensive details. Consequently, we propose a Knowledge-constrained Generative VideoQA Algorithm (KcGA) to introduce implicit semantic constraints and richer semantic support for out-of-domain answer generation, as shown in Figure 1. Specifically, this model adopts the traditional encoder-decoder structure (Sutskever, Vinyals, and Le 2014), in which the encoder maps the input features to a group of fixed-sized vectors and the decoder subsequently maps the vectors to answers. The encoder comprises three modules. The representation module combines the question embeddings and video clips to generate video-question features. In the meanwhile, the adaptive external knowledge module extracts entity objects as the visual knowledge from video frames, which are then fed to an external commonsense knowledge base (ConceptNet (Speer, Chin, and Havasi 2017)) to obtain the external knowledge. To suppress the potential noise during knowledge introduction, both the visual and external knowledge are denoised by CLIP (Radford et al. 2021), a pre-trained model for image-text matching, which can be used to measure the correlation between text and images. Then, the multi-stream information control module forges the external knowledge features into the video-question features by computing the local attentions of external knowledge features and video-question features based on the global attention between the two modes of features. Finally, the knowledge video-question features are fed to a decoder, GPT2 (Radford et al. 2019) in this paper, to generate the freeform answers. By these means, the KcGA significantly improves the quality of out-of-domain answer generation, which has been justified by extensive experiments.

Our contributions are summarized in three aspects. (1) We first propose an open-ended framework that can efficiently generate out-of-domain answers for VideoQA. (2) We propose an adaptive external knowledge module and a multi-stream information control mechanism to introduce the commonsense knowledge into the generation of the out-of-domain answer with rich semantics. (3) On two open-ended benchmark datasets (i.e. NExT-QA (Xiao et al. 2021), TGIF-QA (Jang et al. 2017)), we conduct extensive experiments and obtain the state-of-the-art results.

Related Work

Video Question Answering

VideoQA requires understanding the question and video content to predict the answer. In the past few years, many works have been explored based on the attention mechanism (Xu et al. 2017; Kim et al. 2018; Zhao et al. 2017). Zhou et al. (Zhao et al. 2018a) constructed a multi-stream spatio-temporal attention network for learning joint representations and context-aware question embeddings for dynamic video content. In recent years, the in-depth development of graph networks (Huang et al. 2020; Jiang and Han 2020; Park, Lee, and Sohn 2021), neural modules (Le et al. 2020), and memory networks (Kim et al. 2019; Fan et al. 2019) have also provided more methods for video question answering. Li et al. (Li et al. 2019a) proposed learnable aggregating net with diversity learning, which is based on a multi-path pyramidal attention structure and a diversity learning mechanism to achieve the diversity of attention. All of the above methods define tasks as multi-label classification task and infer answers by setting candidate answer set, which only contain the answers in the train set and cannot predict out-of-domain answers.

In the generative question answering, there had been some attempts (Li et al. 2021; Zhao et al. 2018b; Lee et al. 2021), but these works had problems such as poor model performance and simple question format. Xue et al. (Xue, Zhao, and Cai 2017) proposed sequential video attention and temporal question attention models. This method includes a module for automatically generating answers, but the overall focus is on how to use the timing information of videos, and does not reflect the ability to generate answers outside the domain. Recently, Yang et al. (Yang et al. 2021) also tried to use generative methods to generate answers and made some breakthroughs, but it still needed to construct a set of positive and negative samples, and fundamentally, it can only choose from a limited set. The reason these methods do not generate good out-of-domain answers is that their models do not provide sufficient semantic support for the generation process.

Methods

In this section, we propose a knowledge-constrained open-ended framework to efficiently generate out-of-domain answers. We implement the framework of KcGA, whose architecture is depicted in Figure 2 where the Video-Text Encoder (Figure 2 (a)) extracts the video-question features based on the context information of the video and question, the Adaptive External Knowledge Module (Figure 2 (b)) fits the detected video objects to a knowledge base to obtain the External Knowledge related to video content that provides implicit constraints for answer generation, the Multi-stream Information Control Mechanism (MsICM) (Figure 2 (c)) fuses the video-question features and external knowledge through local and global attention to obtain knowledge enhanced video-question features, which are finally fed to the language decoder (Figure 2 (d)) for free form answer generation.

Feature Representation

We randomly divide the video V of L frames into N clips $\mathbf{C} = \{\mathbf{c}_1, ..., \mathbf{c}_N\}$ of equal length, and embed the question into $\mathbf{Q}$, a group of 100-dimensional vectors corresponding to the words, using WordPiece (Wu et al. 2016) that improves

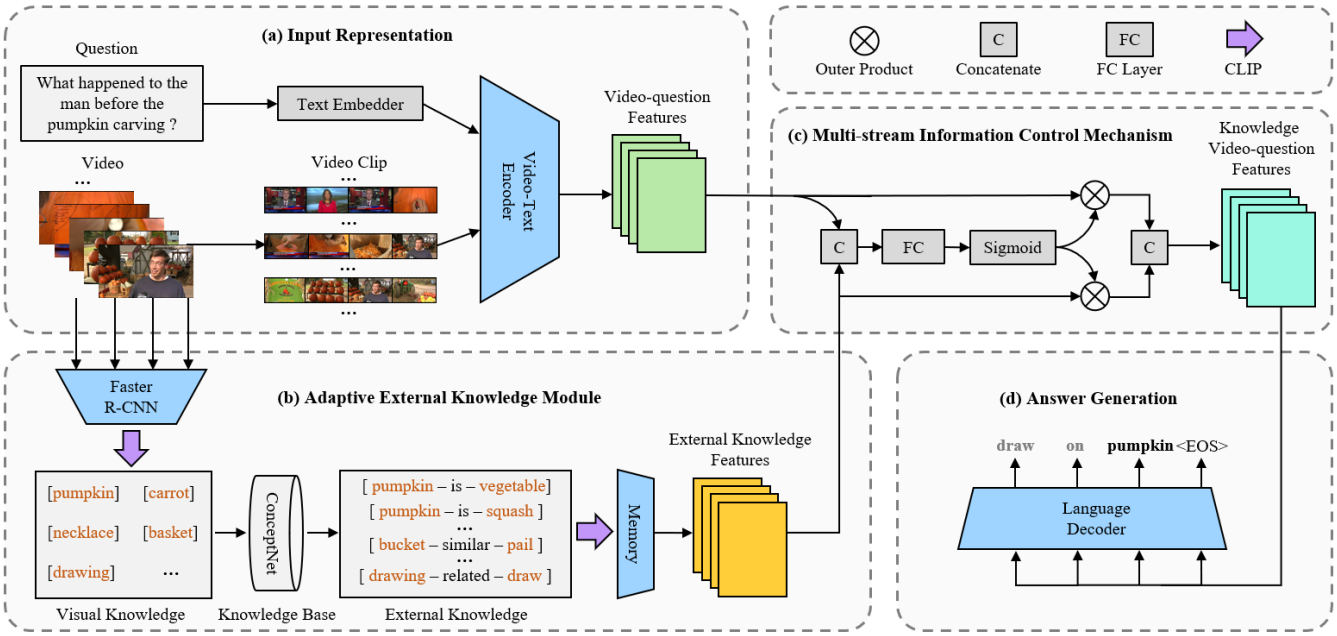

Figure 2: Overall architecture of our model: (a) The Video-question features are extracted by the Video-Text Encoder, which contains the context information of the video and the question. (b) We use the adaptive external knowledge module to obtain External Knowledge, which provides implicit constraints for answer generation. (c) We propose a multi-stream information control mechanism to fully integrate and complement multi-stream information. (d) The language decoder turned out to generate the correct answer in the form of free text.

the semantic distinction between different words. Then, we adopt the baseline model ClipBERT (Lei et al. 2021) as the video-text encoder to extract the well-performed common context information of the video and the question. ClipBERT are efficient in not only visual and linguistic feature extraction, but also memory and computation power consumption. The video-question features are represented as:

$$\mathbf{F}^V = \mathrm{E}_{\mathrm{ClipBERT}}(\mathbf{C}, \mathbf{Q}) \tag{1}$$

where $\mathrm{E}_{\mathrm{ClipBERT}}(\cdot)$ denotes the encoder model.

Adaptive External Knowledge Module

To represent the external knowledge about the videos, we propose to fit the video frames $\mathbf{f}_i (i = 1, ..., L)$ to an external knowledge base to query the external knowledge $\mathbf{k}^E$ and then construct the knowledge features $\mathbf{F}^K$ by iteratively memorizing the focus of attentions existing in $\mathbf{k}^E$.

To this end, we first use Faster R-CNN with ResNet-101 (Anderson et al. 2018) to extract video objects from each frame $\mathbf{f}_i$ and denote the text descriptions of the objects as o_i. However, irrelevant objects to the video could also be extracted by the pre-trained model, therefore we further use CLIP (Radford et al. 2021) to denoise o_i by calculating the similarity between $\mathbf{f}_i$ and o_i to filter off the irrelevant objects. This process can be formulated as:

$$\mathrm{k}^V = \{\mathrm{o}_i | \mathrm{CLIP}(\mathbf{f}_i, \mathrm{o}_i) > \alpha\}_{i=1}^{L} \tag{2}$$

where we name the outputted object label set k^V as the visual knowledge, $\mathrm{CLIP}(\cdot)$ stands for model CLIP for image-text matching, and α is a pred-defined threshold, here we set $\alpha = 0.2$.

k^V is still the knowledge of the entities contained in the video. To obtain external knowledge, we retrieve a set of triples $< \mathrm{o}, \mathrm{r}, \mathrm{s} >$ (both o and s are also text labels of objects, r is the relation between objects) from the commonsense knowledge base ConceptNet (Speer, Chin, and Havasi 2017) by comparing each object of o_i in k^V with the o and s in all the triples and selecting the triples with similar o or s to o_i. We then reorganize the o or s in these triples into o_i and s_i, and denoise them through CLIP:

$$\mathrm{k}^E = \{\mathrm{o}_i \cup \mathrm{s}_i | \mathrm{CLIP}(\mathbf{f}_i, \mathrm{o}_i \cup \mathrm{s}_i) > \beta\}_{i=1}^{L}, \tag{3}$$

where we reuse o_i to represent the objects depicted by the external knowledge, k^E is the denoised external knowledge, $\beta = 0.19$ is a threshold. For further computation, we embed k^E with WordPiece (Wu et al. 2016) to obtain the corresponding vectorized representation:

$$\mathbf{k}^E = \{\mathrm{o}_i \cup \mathrm{s}_i\}_{i=1}^{L} = \mathrm{WordPiece}(\mathrm{k}^E). \tag{4}$$

To fully exploit the contextual information among adjacent frames from $\mathbf{k}^E$, we adopt the Attention based GRU (Kumar et al. 2016) to compute the context aware knowledge representations. In the meanwhile, since the video contains abundant content, we believe the model may focus on distinct but pivotal content of the video each time it spectates the video, then form the final perception about the video after multiple playbacks. The external knowledge could be perceived in the similar way. To this end, we use a group

of episodic memory units $\{\mathbf{m}^1, \mathbf{m}^2, ..., \mathbf{m}^T\}$ to record the focus of attentions of $\mathbf{k}^E$ at each time t. Specifically, we conduct following computations:

$$\mathbf{m}^0 = \tanh(\mathbf{W}_p \mathbf{k}^E + \mathbf{b}_p), \tag{5}$$

$$\mathbf{z}^t = [\mathbf{k}^E \odot \mathbf{m}^0, \mathbf{k}^E \odot \mathbf{m}^{t-1}, |\mathbf{k}^E - \mathbf{m}^0|, |\mathbf{k}^E - \mathbf{m}^{t-1}|], \tag{6}$$

$$\mathbf{g}^t = \mathrm{softmax}(\mathbf{W}_{\mathbf{g}_1} \tanh(\mathbf{W}_{\mathbf{g}_2} \mathbf{z}^t + \mathbf{b}_{\mathbf{g}_2}) + \mathbf{b}_{\mathbf{g}_1}), \tag{7}$$

$$e_r^t = \mathrm{g}_r^t \mathrm{GRU}(k_r^E, e_{r-1}^t) + (1 - \mathrm{g}_r^t) e_{r-1}^t, \tag{8}$$

$$\mathbf{m}^t = \mathrm{ReLU}(\mathbf{W}_m [\mathbf{m}^{t-1}, e_R^t, \mathbf{m}^0] + \mathbf{b}_m) \tag{9}$$

where $\mathbf{z}^t$ captures the similarity between the external knowledge and the memory, $\mathbf{g}^t$ defines the focus of attention each time the model checks on the knowledge, $|\cdot|$ is the absolute value, $\odot$ represents the element-wise product, and $\mathbf{e}^t$ is the state vector of GRU at time t with $\mathbf{e}^0 = 0$. By appropriately setting the number of parameters and GRU states, $\mathbf{g}^t$ and $\mathbf{e}^t$ have the same size as $\mathbf{k}^E$, such that we can compute $\mathbf{e}^t$ element by element through (7) where e_r^t, k_r^E, and g_r^t are the r_{th} element corresponding to the r_{th} state in the GRU. Subsequently, we update the memory using formula (8) where e_R^t is the final state of the GRU. After completing the iteration, the external knowledge features $\mathbf{F}^K$ are given by

$$\mathbf{F}^K = \mathrm{ReLU}(\mathbf{W}_f [\mathbf{m}^1, ..., \mathbf{m}^t] + \mathbf{b}_f) \tag{10}$$

where $\mathbf{m}^t$ is the final episode memory.

Multi-stream Information Control Mechanism

In order to fuse the video-question features and the knowledge features by discovering the mutual complementarity between them, we design a multi-stream information control mechanism. First, we compute the global attention $\mathbf{F}^C$ between video-question features and external knowledge features:

$$\mathbf{F}^C = \mathrm{Sigmoid}(\mathrm{BN}(\mathbf{W}_c [\mathbf{F}^V, \mathbf{F}^K] + \mathbf{b}_c)) \tag{11}$$

where BN indicates the batch normalization. Then, based on the global attention $\mathbf{F}^C$, we compute the local attention for different features separately, and integrate the local attentions to obtain the knowledge enhanced video-question features $\mathbf{F}^M$:

$$\mathbf{F}^M = \mathrm{ReLU}(\mathbf{W}_M [(\mathbf{F}^C)^T \mathbf{F}^V, (\mathbf{F}^C)^T \mathbf{F}^K] + \mathbf{b}_M). \tag{12}$$

Owing to the attention mechanism, the important contents of videos, questions and external knowledge can be well alignment, such that the answer generation module can obtain semantically correct feature representations.

To verify the superiority of MsICM we used, we also design four different multimodal interaction methods based on some existing multimodal fusion methods (Diao et al. 2021; Lee et al. 2018), as shown in Figure 3. We will compare these different multimodal fusion schemes in the experimental section.

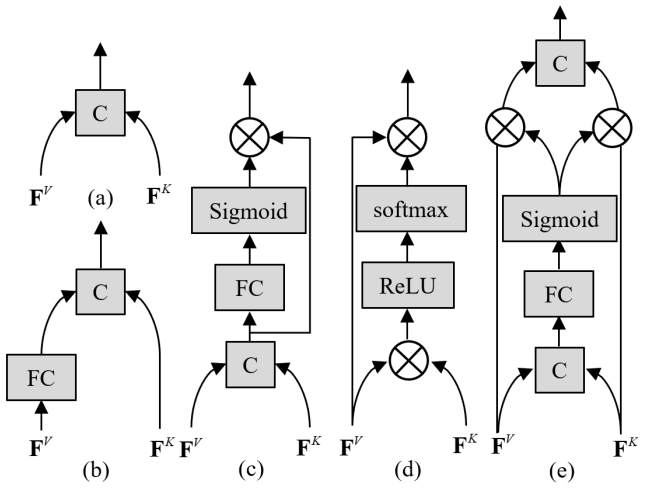

Figure 3: Five different multimodal fusion methods. (a) and (b) concatenate multimodal features, (c) and (d) achieve multimodal fusion using attention, (e) is the proposed multi-stream information control mechanism.

.

Answer Generation and Evaluation

We employ the transformer (Vaswani et al. 2017) as the answer generation module, and use the pretrained model GPT2 (Radford et al. 2019) for parameter initialization, which allows the model to selectively focus on the most relevant parts of the input features to the answer. This can be depicted as:

$$\mathbf{y} = \mathrm{D}_{\mathrm{GPT2}}(\mathbf{F}^M) \tag{13}$$

where $\mathbf{y}$ is our predicted answer and $\mathrm{D}_{\mathrm{GPT2}}(\cdot)$ is the decoder model initialized with GPT2.

Experiments

In this section, we firstly introduce the open-ended benchmark datasets (i.e. NExT-QA (Xiao et al. 2021), TGIF-QA (Jang et al. 2017)) and implementation details, then compare with the state-of-the-art methods and conduct ablation experiments of the model. Particularly, we evaluate the result of out-of-domain answer generation.

Datasets

Open-QA Dataset is an originally constructed dataset from TGIF-QA (Jang et al. 2017) for open-ended Video QA. TGIF-QA dataset is initially built for evaluating the multi-label classification-based VideoQA. It contains high proportion of single-word answer but the multi-word answers are insufficient. Single-word answers cannot reflect well the superiority and generalization ability of generative models owing to the deficient semantic. To this end, we take the text of correct answer to the muti-label classification as the answer to the question, and choose 35862, 7317 and 8506 questions with valid answers (whose frequency of occurrence is more than 10 times) from TGIF-QA to build the train, validation

	Acc			
Methods	Overall	1-L(37.73%)	2-L(53.72%)	3-L(6.70%)
STVQA(Jang et al. 2017)	13.43	20.01	10.94	0
HME(Fan et al. 2019)	16.89	28.54	11.40	0
HCRN(Le et al. 2020)	15.45	25.02	11.18	0
*CoMVT(Seo, Nagrani, and Schmid 2021)	19.55	30.88	14.71	0
*HERO(Li et al. 2020)	16.09	24.74	12.58	0
*ClipBERT(Lei et al. 2021)	17.43	27.30	13.29	0
ClipBERT+Decoder(G)	20.36	26.83	15.85	25.79
KcGA (ours)	27.04	31.01	25.32	25.96

Table 1: Comparisons on Open-QA with the state of the art. 1/2/3-L means the question with answer length 1/2/3 words, and the following percentage indicates the proportion of question with this length of answer. (G) means the decoder is designed to generate freeform answers. "*" indicates pretrained methods. "Acc" represents the accuracy (%)

Methods	Acc
STVQA(Jang et al. 2017)	23.04
HCRN(Le et al. 2020)	23.92
HME(Fan et al. 2019)	24.06
UATT(Xue, Zhao, and Cai 2017)	24.25
HGA(Jiang and Han 2020)	25.18
ClipBERT(Lei et al. 2021)	24.17
KcGA (C)	26.94
KcGA (ours)	28.21

Table 2: Comparisons on NExT-QA with the state of the art. "Acc" represents the accuracy (%). (C) means the decoder is replaced by a multi-label classifier.

and test sets of Open-QA Dataset, within which the questions with answer length of two or more account for 63% of the total questions.

NExT-QA Dataset is a dataset that focuses on video exploitability (Xiao et al. 2021), whose fundamental goal is to evaluate the model's performance in causal behavioral inference, temporal behavioral inference, and common-scenario inference. This requires the model to have higher level of abstraction and logical reasoning ability about videos and questions. The dataset contains 3,870 train, 570 validation, and 1,000 test videos with 37523, 5343 and 9178 open-ended questions respectively.

Implementation Details

In input representation, we refer to the settings of ClipBERT (Lei et al. 2021) on VideoQA task to extract video-question features. In addition to denoising the acquired visual knowledge and external knowledge by CLIP, we exclude some uninformative knowledge, such as "people", "hair" and "sky". In addition, in retrieving the knowledge base with video objects, we reserve only three triples with top-3 leading weights indicating the knowledge confidence, and the retrieved objects from knowledge base adds up to 30 for each video. We extract the external knowledge for videos rather than for video clips. The dimension of an episodic memory unit is 768. We use Aadm (Loshchilov and Hutter 2019) to train our model end-to-end. During the training phase, we set an initial learning rate of 5e-5 to warm up in the first 10% of training steps, then let it decay linearly to 0. The batch size is set to 256 and the dropout rate is set to 0.3. For each task, we train the model for 50 epochs.

State of the Art Comparison

Open-QA: Table 1 shows the comparison between our method (KcGA) and some recent state-of-the-arts including both non-pretrained and pretrained methods on Open-QA dataset. Our method has a clear improvement over previous methods on this complex dataset, with a 6.68% improvement in Overall Acc. We believe this is because we introduce external knowledge from knowledge base to provide much semantic, and our multi-stream information control mechanism captures both the global and local attention for multimodal fusion. For questions with different answer lengths, our method has significant improvement, which amounts to 9.47% when the answer length is 2. The primary reason is that the classification method needs to predefine the candidate answer set for training, and long answers with low occurrence rate often do not appear in the candidate set.

NExT-QA: Table 2 shows the comparison on NExT-QA dataset between KcGA with the spatial-temporal reasoning (Jang et al. 2017), hierarchical conditional relation network (Le et al. 2020), heterogeneous memory (Fan et al. 2019), heterogeneous graph (Jiang and Han 2020), and ClipBERT(baseline) (Lei et al. 2021) methods. Our method achieves significant 4.04% performance improvement against the baseline. Actually, this is predictable based on our previous discussion about the experimental results in Table 1. In addition, the performance of the generative version of our method (KcGA) is also improved by 1.27% compared with the classification version (KcGA(C)). The discrimination between KcGA and KcGA(C) lies in only the decoder.

Ablation Study

In Table 3, we provide an ablation experiment about the performance of multimodal fusion methods on Open-QA. In addition to our multi-stream information control mechanism, we also provide four other methods. Cat1 and Cat2 are simple multimodal concatenation corresponding to Figure 3

Multimodal Interaction Methods	Acc
- Cat1	25.26
- Cat2	25.01
- SRL (Diao et al. 2021)	26.11
- SCA (Lee et al. 2018)	23.04
KcGA (ours)	27.04

Table 3: Ablation study of multimodal fusion methods on Open-QA. Cat1 and Cat2 are simple multimodal concatenation, SRL and SCA are attention-based multimodal fusions, KcGA is the proposed method.

Methods	Type	Acc
ClipBERT (Lei et al. 2021)	C	17.43
ClipBERT+Decoder	G	20.36
KcGA	C	22.97
KcGA (ours)	G	27.04

Table 4: Ablation study on Open-QA using different answer prediction tasks. "Type" means the property of task, "C" refers to multi-label classification and "G" refers to freeform generation.

(a) and Figure 3 (b). SRL means similarity representation learning method proposed by (Diao et al. 2021) (Figure3 (c)), SCA refers to stacked cross attention proposed by (Lee et al. 2018) (Figure3 (d)), and both methods are based on attention. KcGA corresponds to Figure3 (e). It can be seen that Cat1 performs similarly to Cat2, with 0.25% difference. SRL tends to extract global video-question features from multimodal information, while SCA focuses more on video information. This explains why SRL surpassed SCA in this experiment. As comparison, the proposed multi-stream information control mechanism yields 0.93% performance improvement against SRL and 4% improvement against SCA, and the reason lies in the joint usage of global and local attention existing in the multimodal features.

Table 4 explores the accuracies of different answer generation manners on Open-QA, where ClipBERT adopts the multi-label classification task (C) to achieve answer generation, ClipBERT+Decoder replaces the classifier with a text generator (Decoder) to generate the freeform answer (G) and this scheme is similar to the proposed KcGA. We can also replace the decoder of KcGA with a classifier for answer generation. In general, the generative method has higher flexibility and stronger generalization ability in multi-word answer generation. ClipBERT+Decocer outperforms ClipBERT by 2.93% and KcGA outperforms KcGA with classifier-based answer generator by 4.07%. This can similarly be attributed to the independence of the generative decoder on the pre-defined candidate label set. We also find that the improvement by applying generative decoder to KcGA exceeds the improvement by applying generative decoder to ClipBERT. Therefore, we believe that the abundant semantic implied in the external knowledge benefits the freeform answer generation more than the multi-label classification. It is noteworthy that the KcGA with multi-label classifier exceeds the ClipBERT+Decocer by 2.61, which

Methods	Acc
ClipBERT+Decoder (G)	20.36
- No E_CLIP	24.91
- No V_CLIP	26.55
- No Memory	26.46
KcGA (ours)	27.04

Table 5: Ablation study about CLIP and memory operation on Open-QA. (G) represents that we use the generative approach for VideoQA. "No E_CLIP" means that we did not denoise the external knowledge with CLIP, "No V_CLIP" means that we did not denoise the visual knowledge, and "No Memory" means we remove memory computation from KcGA.

Methods	Out-of-domain
ClipBERT (Lei et al. 2021)	-
ClipBERT+Decoder (G)	1.34
KcGA (C)	-
- No Memory	4.16
KcGA (ours)	4.55

Table 6: Study about the ability of out-of-domain answer generation (the answers were not included in the training set). We use accuracy (%) to evaluate this experiment. "-" indicates the method tackles a multi-label classification task, which does not support out-of-domain answer generation.

justifies the effect of the external knowledge.

Table 5 shows the ablation study about CLIP and memory operation on Open-QA, where "No E_CLIP" mean removing the external knowledge denoising from KcGA while keeping other part unchanged, "No V_CLIP" means removing the visual knowledge denoising from KcGA while keeping other part unchanged, and "No Memory" means we remove memory computation from KcGA and feed the external knowledge directly to the next module. If we denoise only the visual knowledge, the performance drops by 2.13% compared with KcGA. If we only denoise the external knowledge, the performance decrease is 0.49%. External knowledge denoising tends to play more important role because this operation filters off some objects involved in the knowledge base but irrelevant to the video. "No Memory" means that the external knowledge is directly passed to the next module without being encoded in memory. By further optimizing the features with memory units, the accuracy of our method is improved by 0.58Table 6 demonstrates the performance of different methods on out-of-domain answer generation. ClipBERT cannot predict out-of-domain answers because it addresses a multi-label classification task. We substitute the classifier with the decoder module to enable it to generate out-of-domain answers (ClipBERT+Decoder), and the performance on out-of-domain answer generation is 1.34%. Our method (KcGA) achieves 4.55% accuracy on out-of-domain answer generation, thanks to the rich semantics provided by the proposed adaptive external knowledge module. Furthermore, we can see that optimizing the external knowledge through

Question: What does the man do after look downward?

Answer: Look upward

KcGA (Out-of-domain): **Look upward**

Question: What does the man do after look down at hand?

Answer: Put hand to face

KcGA (Out-of-domain): **Put hand to face**

Question: What does the man do after smile?

Answer: Nod

KcGA: **Nod head**

Question: What does the girl do after look surprise?

Answer: Smile

KcGA: **Laugh**

Figure 4: Typical results generated by our method (KcGA) on Open-QA dataset. The green texts denote the same out-of-domain answers as the ground truth ones, and the blue texts denote the semantically similar answers to the ground truth ones.

the memory scheme benefits the out-of-domain answer generation.

Qualitative Results: In Figure 4, we enumerate some typical results generated by our method. The upper part shows the out-of-domain answers (rendered in green) generated by KcGA that are the same as the ground truth ones. The lower part shows the answers generated by KcGA that are semantically similar to the ground truth ones. In order to comply with the current video question answering work, we have to use the Acc for evaluation, which has obvious limitations. For example, it can be observed that the generated answer "Nod head" is semantically identical to the correct answer "Nod", but the semantic consistency cannot be reflected by the Acc. This explains why our results in table 6 are particularly low. Even though we obtain semantically dissimilar result to the ground truth answer e.g. "laugh" versus "smile", it still cannot be regarded as a wrong answer either. These properties are not possessed by the multi-label classification tasks.

Conclusion

Open-ended VideoQA task needs to generate answer according to the video content and target question. We provide a framework that can efficiently generate out-of-domain answers. The framework follows a traditional encoder-decoder structure. In the decoding phase, we propose an adaptive external knowledge module and a novel multi-stream information control mechanism to use knowledge constraints to optimize the feature encoding, and subsequently improve the quality of answer generation. Our model achieves state-of-the-art performance on two challenging VideoQA datasets. Especially, our method is effective to multi-word answer generation task.

Acknowledgements

This work was supported by NSFC NO. 62125201 and NSFC No. 62020106007.

References

Anderson, P.; He, X.; Buehler, C.; Teney, D.; Johnson, M.; Gould, S.; and Zhang, L. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 6077–6086.

Cherian, A.; Hori, C.; Marks, T. K.; and Le Roux, J. 2022. (2.5+ 1) D Spatio-Temporal Scene Graphs for Video Question Answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 444–453.

Diao, H.; Zhang, Y.; Ma, L.; and Lu, H. 2021. Similarity reasoning and filtration for image-text matching. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 1218–1226.

Fan, C.; Zhang, X.; Zhang, S.; Wang, W.; Zhang, C.; and Huang, H. 2019. Heterogeneous memory enhanced multimodal attention model for video question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 1999–2007.

Gao, J.; Ge, R.; Chen, K.; and Nevatia, R. 2018. Motion-appearance co-memory networks for video question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 6576–6585.

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

the AAAI Conference on Artificial Intelligence, volume 33, 6391–6398.

Huang, D.; Chen, P.; Zeng, R.; Du, Q.; Tan, M.; and Gan, C. 2020. Location-aware graph convolutional networks for video question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 11021–11028.

Jang, Y.; Song, Y.; Yu, Y.; Kim, Y.; and Kim, G. 2017. Tgif-qa: Toward spatio-temporal reasoning in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2758–2766.

Jiang, P.; and Han, Y. 2020. Reasoning with heterogeneous graph alignment for video question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 11109–11116.

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

Kim, K.-M.; Choi, S.-H.; Kim, J.-H.; and Zhang, B.-T. 2018. Multimodal dual attention memory for video story question answering. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 673–688.

Kumar, A.; Irsoy, O.; Ondruska, P.; Iyyer, M.; Bradbury, J.; Gulrajani, I.; Zhong, V.; Paulus, R.; and Socher, R. 2016. Ask me anything: Dynamic memory networks for natural language processing. In *International conference on machine learning*, 1378–1387. PMLR.

Le, T. M.; Le, V.; Venkatesh, S.; and Tran, T. 2020. Hierarchical conditional relation networks for video question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9972–9981.

Lee, D.; Choi, S.; Jang, Y.; and Zhang, B.-T. 2021. Mounting Video Metadata on Transformer-based Language Model for Open-ended Video Question Answering. *arXiv preprint arXiv:2108.05158*.

Lee, K.-H.; Chen, X.; Hua, G.; Hu, H.; and He, X. 2018. Stacked cross attention for image-text matching. In *Proceedings of the European conference on computer vision (ECCV)*, 201–216.

Lei, J.; Li, L.; Zhou, L.; Gan, Z.; Berg, T. L.; Bansal, M.; and Liu, J. 2021. Less is more: Clipbert for video-and-language learning via sparse sampling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7331–7341.

Li, L.; Chen, Y.-C.; Cheng, Y.; Gan, Z.; Yu, L.; and Liu, J. 2020. Hero: Hierarchical encoder for video+ language omni-representation pre-training. *arXiv preprint arXiv:2005.00200*.

Li, X.; Gao, L.; Wang, X.; Liu, W.; Xu, X.; Shen, H. T.; and Song, J. 2019a. Learnable aggregating net with diversity learning for video question answering. In *Proceedings of the 27th ACM international conference on multimedia*, 1166–1174.

Li, X.; Song, J.; Gao, L.; Liu, X.; Huang, W.; He, X.; and Gan, C. 2019b. Beyond rnns: Positional self-attention with co-attention for video question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 8658–8665.

Li, Y.; Wang, X.; Xiao, J.; Ji, W.; and Chua, T.-S. 2022. Invariant grounding for video question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2928–2937.

Li, Z.; Li, Z.; Zhang, J.; Feng, Y.; and Zhou, J. 2021. Bridging text and video: A universal multimodal transformer for audio-visual scene-aware dialog. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29: 2476–2483.

Loshchilov, I.; and Hutter, F. 2019. Decoupled weight decay regularization (2017). *arXiv preprint arXiv:1711.05101*.

Park, J.; Lee, J.; and Sohn, K. 2021. Bridge to answer: Structure-aware graph interaction network for video question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15526–15535.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 8748–8763. PMLR.

Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; Sutskever, I.; et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1: 9.

Seo, P. H.; Nagrani, A.; and Schmid, C. 2021. Look before you speak: Visually contextualized utterances. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16877–16887.

Speer, R.; Chin, J.; and Havasi, C. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Thirty-first AAAI conference on artificial intelligence*.

Sutskever, I.; Vinyals, O.; and Le, Q. V. 2014. Sequence to sequence learning with neural networks. *Advances in neural information processing systems*, 27.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Wu, Y.; Schuster, M.; Chen, Z.; Le, Q. V.; Norouzi, M.; Macherey, W.; Krikun, M.; Cao, Y.; Gao, Q.; Macherey, K.; et al. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.

Xiao, J.; Shang, X.; Yao, A.; and Chua, T.-S. 2021. Next-qa: Next phase of question-answering to explaining temporal actions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9777–9786.

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

Xue, H.; Zhao, Z.; and Cai, D. 2017. Unifying the video and question attentions for open-ended video question answering. *IEEE Transactions on Image Processing*, 26: 5656–5666.

Yang, A.; Miech, A.; Sivic, J.; Laptev, I.; and Schmid, C. 2021. Just ask: Learning to answer questions from millions of narrated videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 1686–1697.

Yu, W.; Zhu, C.; Li, Z.; Hu, Z.; Wang, Q.; Ji, H.; and Jiang, M. 2022. A survey of knowledge-enhanced text generation. *ACM Computing Surveys (CSUR)*.

Zhao, Z.; Jiang, X.; Cai, D.; Xiao, J.; He, X.; and Pu, S. 2018a. Multi-turn video question answering via multi-stream hierarchical attention context network. In *IJCAI*, volume 2018, 27th.

Zhao, Z.; Yang, Q.; Cai, D.; He, X.; Zhuang, Y.; Zhao, Z.; Yang, Q.; Cai, D.; He, X.; and Zhuang, Y. 2017. Video Question Answering via Hierarchical Spatio-Temporal Attention Networks. In *IJCAI*, volume 2, 8.

Zhao, Z.; Zhang, Z.; Xiao, S.; Yu, Z.; Yu, J.; Cai, D.; Wu, F.; and Zhuang, Y. 2018b. Open-ended long-form video question answering via adaptive hierarchical reinforced networks. In *IJCAI*, volume 2, 8.