ActivityNet-QA: A Dataset for Understanding Complex Web Videos via Question Answering

Zhou Yu[1], Dejing Xu[2], Jun Yu[1]*, Ting Yu[1], Zhou Zhao[2], Yueting Zhuang[2], Dacheng Tao[3]
[1] Key Laboratory of Complex Systems Modeling and Simulation,
School of Computer Science and Technology, Hangzhou Dianzi University, Hangzhou, China.
[2] College of Computer Science, Zhejiang University, Hangzhou, China
[3] UBTECH Sydney AI Centre, SIT, FEIT, University of Sydney, Australia
{yuz, yujun , yuting}@hdu.edu.cn, {xudejing, zhaozhou, yzhuang}@zju.edu.cn, dacheng.tao@sydney.edu.au

Abstract

Recent developments in modeling language and vision have been successfully applied to image question answering. It is both crucial and natural to extend this research direction to the video domain for video question answering (VideoQA). Compared to the image domain where large scale and fully annotated benchmark datasets exists, VideoQA datasets are limited to small scale and are automatically generated, etc. These limitations restrict their applicability in practice. Here we introduce ActivityNet-QA, a fully annotated and large scale VideoQA dataset. The dataset consists of 58,000 QA pairs on 5,800 complex web videos derived from the popular ActivityNet dataset. We present a statistical analysis of our ActivityNet-QA dataset and conduct extensive experiments on it by comparing existing VideoQA baselines. Moreover, we explore various video representation strategies to improve VideoQA performance, especially for long videos. The dataset is available at `https://github.com/MILVLG/activitynet-qa`

Introduction

Recent developments in deep neural networks have significantly accelerated the performance of many computer vision and natural language processing tasks. These advances stimulated research into bridging the semantic connections between vision and language, such as in visual captioning (Donahue et al. 2015; Xu et al. 2015), visual grounding (Rohrbach et al. 2016; Chen, Kovvuri, and Nevatia 2017; Yu et al. 2018b) and visual question answering (Malinowski, Rohrbach, and Fritz 2015; Fukui et al. 2016).

Visual question answering (VQA) aims to generate natural language answers to free-form questions about a visual object (*e.g.*, an image or a video). Compared to visual captioning, VQA is *interactive* and provides fine-grained visual understanding. Image question answering (ImageQA) in particular has shown recent success, with many approaches proposed to investigate the key components of this task, *e.g.*, discriminative feature representation (Anderson et al. 2018), multi-modal fusion (Kim et al. 2017a; Yu et al. 2017; Yu et al. 2018a) and visual reasoning (Nam, Ha, and Kim 2016; Lu et al. 2016; Johnson et al. 2017b). This success has been facilitated by large scale and well annotated training datasets, such as Visual Genome (Krishna et al. 2016) and VQA (Antol et al. 2015; Goyal et al. 2017).

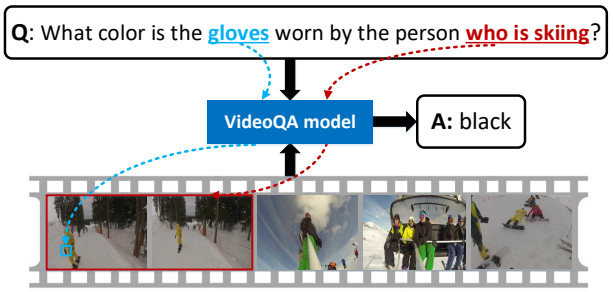


Figure 1: A VideoQA example. To answer the question correctly, one should fully understand the fine-grained semantics of the questions (*i.e.*, the underlined keywords) and perform spatio-temporal reasoning on the visual contents of the video (*i.e.*, frames in red border and objects in blue box).

Video question answering (VideoQA) can be seen as a natural but more challenging extension of ImageQA, due to the additional complexity of understanding of image sequences and more diverse types of questions asked. Figure 1 shows an example of VideoQA. To accurately answer the question, a VideoQA model requires simultaneous fine-grained video content understanding and spatio-temporal reasoning. Existing approaches mainly focus on the temporal attention mechanism (Jang et al. 2017; Xu et al. 2017) or memory mechanism (Na et al. 2017; Kim et al. 2017b; Zhao et al. 2018a). Na *et al.* introduced a read-write memory network to fuse multi-modal features and store temporal information using a multi-stage convolutional neural networks model (Na et al. 2017). Xu *et al.* represented a video as appearance and motion stream features and introduced a gradually refined attention model to fuse the two-stream features together. (Xu et al. 2017). Gao *et al.* proposed a co-memory network to jointly model and interact with the motion and appearance information (Gao et al. 2018). Zhao *et al.* introduced an adaptive hierarchical encoder to learn the segment-level video representation with adaptive video

*Jun Yu is the corresponding author

Table 1: Comparison of existing VideoQA datasets with ours (OE: open-ended, and MC: multiple-choice).

Datasets	Video source	QA pairs generation	QA tasks	# Videos	# QA pairs	Average video length
MSVD-QA (Xu et al. 2017)	MSVD	Automatic	OE	1,970	50,505	10s
MSRVTT-QA (Xu et al. 2017)	MSRVTT	Automatic	OE	10,000	243,680	15s
TGIF-QA (Jang et al. 2017)	TGIF	Automatic & Human	OE & MC	56,720	103,919	3s
MovieQA (Tapaswi et al. 2016)	Movies	Human	MC	6,771	6,462	200s
Video-QA (Zeng et al. 2017)	Jukinmedia	Automatic	OE	18,100	174,775	45s
ActivityNet-QA (Ours)	ActivityNet	Human	OE	5,800	58,000	180s

segmentation, and devised a reinforced decoder to generate the answer for long videos (Zhao et al. 2018b).

As noted above, high-quality datasets are of considerable value for VQA research. Several VideoQA datasets have been compiled for different scenarios, such as MovieQA (Tapaswi et al. 2016), TGIF-QA (Jang et al. 2017), MSVD-QA, MSRVTT-QA (Xu et al. 2017), and Video-QA (Zeng et al. 2017). Most of these VideoQA datasets exploit video source data from other datasets and then add question-answer pairs to them. The detailed statistics of these datasets are listed in Table 1. We can see that these existing datasets are imperfect and have at least one of the following limitations:

- The datasets are small scale. Without sufficient training samples, the obtained model suffers from under-fitting. Without sufficient testing samples, the evaluated results are unreliable.
- The questions and answers are automatically generated by algorithms (*e.g.*, obtained from the captioning results or narrative descriptions using off-the-shelf algorithms) rather than human annotation. Automatically generated question-answer pairs lack diversity, making the learned model easy to over-fit.
- The videos are short. The length of a video is closely related to the complexity of video content. Questions on short videos (*e.g.*, less than 10 seconds) are usually too easy to answer making it difficult distinguish the performance of different VideoQA approaches on the dataset.
- The videos represent a small number of activities. This severely restricts the generalizability of the VideoQA models trained on these datasets and poorly reflects model performance in real-world use.

In this paper, we construct a new benchmark dataset *ActivityNet-QA* for evaluating VideoQA performance. Our dataset exploits 5,800 videos from the ActivityNet dataset, which contains about 20,000 untrimmed web videos representing 200 action classes (Fabian Caba Heilbron and Niebles 2015). We annotate each video with ten question-answer pairs using crowdsourcing to finally obtain 58,000 question-answer pairs. Compared with other VideoQA datasets, ActivityNet-QA is of large scale, fully annotated by humans, and with very long videos. To better understand the properties of ActivityNet-QA, we present statistical and visualization analyses. We further conduct experiments on ActivityNet-QA and compare results produced by existing VideoQA baselines.

ActivityNet-QA Dataset

We first introduce the ActivityNet-QA dataset from three perspectives, namely *video collection*, *QA generation*, and *statistical analysis*.

Video Collection

We first collect videos for the dataset. Due to time limitation, we are unable to annotate every video in ActivityNet. Instead, we sample 5,800 videos from the 20,000 videos in ActivityNet dataset. Specifically, we sample 3,200/1,800/800 videos from the original train/val/test splits of ActivityNet respectively. Moreover, we take class diversity and balance into consideration. Since the videos in the train and val splits of ActivityNet possess class labels, we use this information as a prior to guide sampling and force a uniform distribution of class labels in the sampled videos. The class information is not available for the test split, so we adopt a simple random sampling strategy instead.

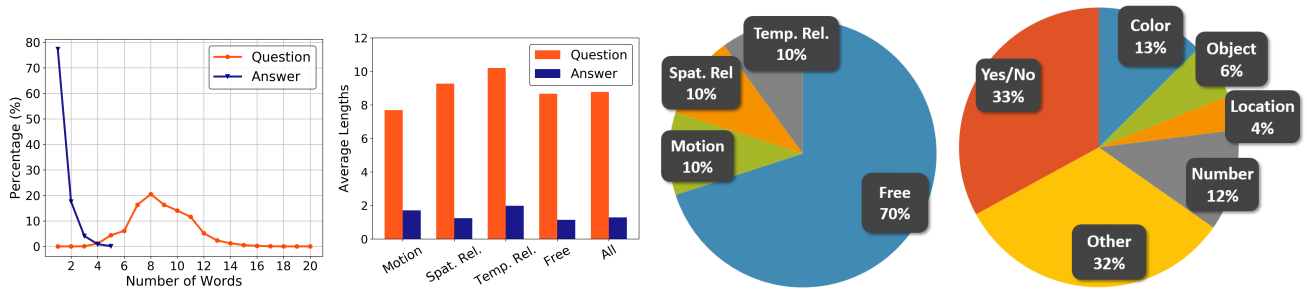
QA Generation

As the videos are collected, we generate QA pairs for each video. To reduce the labor costs, some VideoQA datasets exploit the narrative descriptions or captions of videos (Jang et al. 2017; Xu et al. 2017), to automatically generate QA pairs using off-the-shelf algorithms (Ren, Kiros, and Zemel 2015). However, since the textual descriptions contains relatively little information about the videos, these generated QA pairs lack diversity and are often redundant. Therefore, we generate QA pairs by human crowdsourcing.

To control the generated questions, we define three template question types and ask the annotators to cover all the three question types for every video. Beyond these three questions, annotators are free to ask arbitrary questions about the videos. The three question types are as follows:

Motion. This type of question interrogates coarse temporal action understanding. Compared with traditional action recognition, this task is more challenging in this setting. To correctly answer the question with respect to a long video, a VideoQA model needs to correctly localize the action referred to by the question.

Spatial Relationship. This type of question tests spatial reasoning on one static frame. In contrast to the spatial reasoning in ImageQA (Johnson et al. 2017a), this task



(a) QA length distributions (b) Average QA lengths (c) Question distribution (d) Answer distribution for *Free* type

Figure 2: The statistics of our ActivityNet-QA dataset.

additionally examine the temporal attention ability to find to the frame from the whole video first.

Temporal Relationship. As a companion to spatial relationship, this type of questions examines the ability of reasoning temporal relationships of objects from a sequence of frames. As a prerequisite, one should find the related frames from the whole video first.

For the free type questions, it is hard to classify them into non-overlapped types even for humans. Referring to the taxonomy in existing VQA datasets (Ren, Kiros, and Zemel 2015; Antol et al. 2015), we manually categorize the samples into the following six classes by their answer types: *Yes/No*, *Number*, *Color*, *Object*, *Location* and *Other*.

To control the quality of generated questions, the following practical principles are applied:

- Questions and answers that are too long are probably caused by improper representation. Therefore, we empirically restricted the maximum question length to 20 words and maximum answer length to 5 words.
- For each QA pair, the question annotator and answer annotator are separate. If the answer annotator regards the generated question *unanswerable*, this question is double-checked and may have been further regenerated. Employing this strategy effectively improve question objectivity, which is important for obtaining high-quality annotations.
- A portion of questions are randomly selected and sent to multiple annotators. The multiple answers to one question are merged by majority voting. Employing this strategy reduced the probability of erroneous answers and evaluated annotator reliability.

The initial QA pairs are in Chinese, since our crowdsourcing platform is located in China. As the lengths of questions and answers are well controlled, the state-of-the-art machine translation algorithms can easily translate them into English. To further improve the quality of the translated results, we use a novel strategy to automatically detect potential mistakes in the translated results. For each sentence in Chinese, we use the APIs from the four commercial translation engines of Google, Baidu, Sogou and Youdao to obtain four translated versions in English. We evaluate the average similarities of each two natural language sentences using CIDEr score (Vedantam, Lawrence Zitnick, and Parikh 2015), setting an empirical threshold to the average CIDEr score, and manually checking samples that did not reach the threshold. For the remaining samples that did reach the threshold, we use the results obtained from the Baidu translation engine.

Table 2: Examples of questions in different types.

Types	Questions
Motion	What are the person wearing earphones doing? What are people doing at the beginning of the video? What is person wearing red t-shirt doing?
Spat. Rel.	What is on the left of the lawn? What is on the left side of the man on his knees? What is behind of the person sitting in the video?
Temp. Rel.	What happened to the person in black before falling down? What happened to the woman before drying her hair? What happened to the person before playing violin?
Free	How many people are there in the video? [Number] Is the athlete in the room? [Yes/No] What are the animals that appear in the video? [Object] What is the color of the person's pants? [Color] Where is the person in a black coat? [Location] What is the gender of the athlete? [Other]

To better understand the different question types and the quality of translations, some examples from our dataset are shown in Table 2.

Statistical Analysis

Here we present the detailed statistics of our ActivityNet-QA dataset. The distributions of question and answer lengths are shown in Figure 2(a). As noted above, the maximum question length is 20 and the maximum answer length is 5 respectively (in English). The average QA lengths for all the question types are reported in Figure 2(b). Regardless of question type, the average question length is 8.67 and average answer length is 1.85. Similar to (Antol et al. 2015), the answer lengths in our dataset are relatively short. Short answers are easier to process, and one can simply treat the answering problem as multi-class classification.

We also investigate the distribution of the questions (Fig-

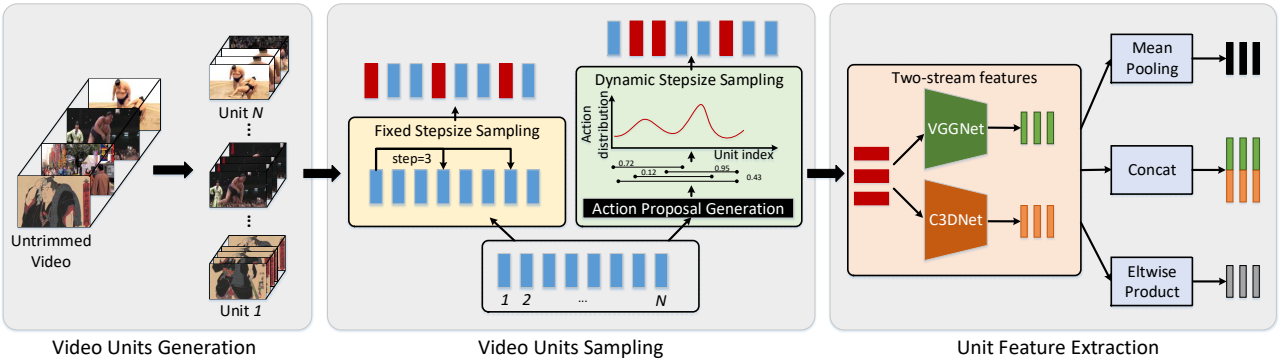


Figure 3: The flowchart of video feature representation procedures, including video units generation (left), video units sampling (middle) and unit feature extraction (right).

ure 2(c)). For each video, we generate exactly ten QA pairs, including one *motion* type question, one *spatial relationship* type question, and one *temporal relationship* type question, respectively. The remaining seven questions are classified as *free* type as they are generated without constraints. To eliminate the effect of the answer prior and improve the role of video understanding for *Yes/No* type questions, we balance the ratio of *yes* and *no* samples to make it close to one. To better understand the organization of the *free* type questions, the answer distribution is shown in Figure 2(d).

Methods

In this section, we explore the difficulty of the ActivityNet-QA dataset using several baseline models based on different types of video features.

Video Feature Representation

Existing VideoQA approaches usually extract *two-stream* features from the motion channel and the appearance channel of video, respectively (Xu et al. 2017).

Since the untrimmed videos are very long, we split the videos into small units, each unit containing 16 consecutive frames without overlap between any two units. The average number of units counted on all the videos is 270, which is still too large for existing VideoQA models. Besides, the number of units varies for different videos, further complicating model training. Therefore, we propose two alternative sampling strategies to sample a fixed number of units T for all videos:

Fixed Stepsize (FS). For a video, assume it contains N units and we expect to output T units. This strategy evenly sample T units from the N units with a fixed stepsize $\lceil \frac{N}{T} \rceil$.

Dynamic Stepsize (DS). An untrimmed video may contain many worthless frames with little information. To make the sampled video units more discriminative, we propose a sampling strategy with dynamic stepsize such that the selected units have a high probability of containing meaningful actions. To achieve this, we first introduce an external temporal action proposal model (Fabian Caba Heilbron and Ghanem 2016) to generate a set of action proposals:

$$P = \{(t_i^{\text{start}}, t_i^{\text{end}}, c_i)\} \quad (1)$$

where each proposal $p_i \in P$ contains a start index $t_i^{\text{start}} \in \mathbb{R}$, an end index $t_i^{\text{end}} \in \mathbb{R}$ and a confidence score $c_i \in \mathbb{R}$. For each video unit, we regard it as a bin in a histogram. If the duration of a proposal p_i covers the video unit, its confidence score c_i is added to this unit. After traversing the candidate set P, we obtain a histogram w.r.t the video units. By normalizing the histogram using the softmax function, we obtain an action score distribution over the video units indicating the probability that one unit has valid actions. Finally, we sample the units w.r.t the score distribution to obtain T units with dynamic stepsize.

As we have obtained the sampled video unit set, for each unit u_i, we used the VGG-16 network pre-trained on the ImageNet dataset (Simonyan and Zisserman 2014) to extract the appearance feature (the `fc7` feature $x_i \in \mathbb{R}^{4096}$) given the central frame of the unit, and the C3D network pre-trained on the Sport-1M dataset (Tran et al. 2015) to extract the motion features (the `fc7` feature $y_i \in \mathbb{R}^{4096}$) given the whole 16 consecutive frames of the unit.

To fuse the two-stream features, we use three fusion strategies: *Mean Pooling*, *Concat* and *Eltwise Product* to obtain the 4096-D, 8192-D, and 4096-D fused visual features for each video unit, respectively. We then perform L_2 normalization on each fused visual feature.

The overall flowchart for video feature representation is illustrated in Figure 3.

VideoQA Baselines

Based on the extracted visual features, we implement the following VideoQA baselines. Note that the focus of this paper is the constructed ActivityNet-QA dataset and a discussion of what influences the performance on the dataset. Therefore, we do not perform comparison with complex VideoQA models, such as (Gao et al. 2018; Xu et al. 2017).

E-VQA is the extension of an ImageQA baseline (Antol et al. 2015), where one long-short term memory (LSTM) network (Hochreiter and Schmidhuber 1997) is used to

Table 3: The accuracies of the methods in different question types. Q-type prior denotes a simple baseline using the most popular answer per question type as the prediction.

Methods	Accuracy (%)					WUPS (%)	
	Motion	Spat. Rel.	Temp. Rel.	Free	All	WUPS@0.9	WUPS@0.0
Q-type prior	2.9	5.8	1.4	19.7	14.8	16.4	35.1
E-VQA	2.5	6.6	1.4	34.4	25.1	29.3	53.5
E-MN	3.0	8.1	1.6	36.9	27.1	31.5	55.9
E-SA	**12.5**	**14.4**	**2.5**	**41.2**	**31.8**	**34.9**	**56.4**

encode all words in the question and another different LSTM network is used to encode the frames in the video. The features of the question and videos are then fused into the joint feature representation with element-wise multiplication for answer prediction.

E-MN is the extension of the end-to-end memory networks model (Sukhbaatar et al. 2015) for ImageQA, where the bidirectional LSTM networks are used to update the frame representations of the video. The updated representations are mapped into the memory and the question representation is used to perform multiple inference steps to predict the answer.

E-SA is the extension of the soft attention model (Yao et al. 2015) for ImageQA, where the question are fist encoded using a LSTM network. The encoded question feature is used to attend on features of video features. Finally, the question feature and weighted video feature are fused to predict the answer.

All the above baselines are trained in an end-to-end manner. Since they are decoupled from the video features, they can be flexibly combined with the video features obtained by different strategies.

Experiments

We evaluate the aforementioned VideoQA models on our ActivityNet-QA dataset. We use 3,200 videos and 32,000 corresponding QA pairs in the train split to train the models, and 1,800 videos and 18,000 corresponding QA pairs in the val split to tune hyper-parameters. We report the predicted results on 800 videos and 8,000 QA pairs in the test split.

Experimental Setup

We formulate the VideoQA problem as a multi-class classification problem with each class corresponding to an answer. To generate the answer vocabulary, we choose the top 1,000 most frequent answers in the train split as the answer vocabulary, which covers 84.7% / 86.2% / 85.6% of the train/val/test answers, respectively. To generate the question vocabulary, we select the top 8,000 most frequent words from the questions in the train split. We take the token *unk* for out-of-vocabulary words. Since there are several video feature representation strategies, we use the FS sampling with $T = 20$ and the mean-pooling fusion strategies as the default options in the experiments unless otherwise stated.

We implement all the methods and train the models using TensorFlow. For all models, we use the Adam solver with a base learning rate $\alpha = 0.001$, $\beta_1 = 0.9$, and $\beta_2 = 0.99$ and train the models to up to 100 epochs with a batch size of 100. The early stopping strategy is used if the accuracy on the validation set does not improve for 10 epochs. All models use the pre-trained 300-dimensional GloVe embedding (Pennington, Socher, and Manning 2014) to initialize the question embedding layer. For the models using LSTM networks, the number of LSTM hidden units is set to 300, and the common space dimension is set to 256 as suggested by (Xu et al. 2017). The number of memory units for E-MN is set to 500 as suggested by (Zeng et al. 2017).

Evaluation Criteria

We evaluate the performance using two common evaluation criteria for VideoQA, *i.e.*, accuracy (Xu et al. 2017) and WUPS (Malinowski and Fritz 2014). For the QA pairs in the test set with size N, given any testing question $\mathbf{q}_i \in Q$ and its corresponding ground-truth answer $\mathbf{y}_i \in Y$, we denote the predicted answer from the VideoQA model by $\mathbf{a}_i$. Note that $\mathbf{a}_i$ or $\mathbf{y}_i$ corresponds to a sentence which can be seen as a set of words. Based on the definition above, the two evaluation criteria are:

Accuracy is a criterion that is used to commonly used to measure the performance of classification tasks.

$$\text{Accuracy} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{1}[\mathbf{a_i} = \mathbf{y_i}] \tag{2}$$

where $\mathbf{1}[\cdot]$ is an indicator function that accuracy of the sample is 1 only if $\mathbf{a_i}$ and $\mathbf{y_i}$ are identical, and 0 otherwise.

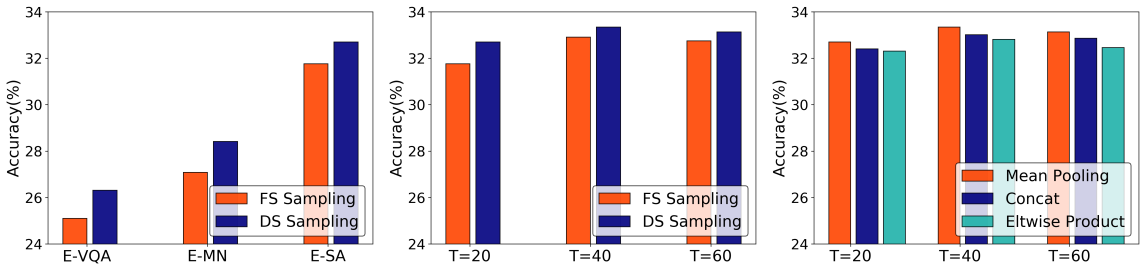
WUPS is a generalization of the accuracy measure that accounts for word-level ambiguities in the answer words. The WUPS score with the threshold γ is given by

$$\text{WUPS} = \frac{1}{N}\sum_{i=1}^{N}\min\{\prod_{w\in\mathbf{a}_i}\max_{v\in\mathbf{y}_i}\mu_\gamma(w,v), \prod_{v\in\mathbf{y}_i}\max_{w\in\mathbf{a}_i}\mu_\gamma(w,v)\} \tag{3}$$

where w and v are the words in the each predicted answer and ground-truth answer respectively. $\mu_\gamma(w,v)$ is given by

$$\mu_\gamma(w,v) = \begin{cases} \text{WUP}(w,v) & \text{if WUP}(w,v) \geq \gamma \\ 0.1 \cdot \text{WUP}(w,v) & \text{otherwise} \end{cases} \tag{4}$$

Following the setting in (Malinowski, Rohrbach, and Fritz 2015), we choose two thresholds $\gamma = 0.0$ and $\gamma = 0.9$ for calculating the WUPS score and denote them by WUPS@0.0 and WUPS@0.9, respectively.



(a) Sampling strategies (b) Sampling frequencies (c) Fusion strategies

Figure 4: Overall accuracies of different strategies in video feature representations. (a) sampling strategies w.r.t. different VideoQA methods; (b) sampling frequencies w.r.t. different sampling strategies for E-SA; (c) sampling frequencies w.r.t. different fusion strategies for E-SA.

Figure 5: Visualizations of three video examples with two sampling strategies (DS sampling on the left and FS sampling on the right). Each row shows the sampled video units (represented by their central frames) for a video with sampling frequency T=5.

Results and Discussion

Table 3 shows the performance of the baselines on all question types based on the two evaluation criteria. From these results, we can make the following observations: 1) all baselines significantly outperform the *Q-type prior* baseline with respect to the overall accuracy and WUPS scores, indicating that without understanding the visual content of videos, one cannot achieve good performance on our dataset; 2) the accuracy of the *temporal relationship* type is lower than the others. This can be explained by the fact that temporal reasoning over long videos is still not well solved by baselines, and there remains significant room for further improvement; 3) E-SA slightly outperforms E-VQA and E-MN both in terms of accuracy and WUPS, respectively. However, the overall performance is still far from satisfactory, reflecting the difficulty of the dataset.

Table 4 provides the detailed accuracies for the free type questions. The accuracy for the *Object* class is lower than the others due to the diversity of possible answers. E-SA still outperforms the other two methods steadily, indicating its effectiveness in modeling temporal attention.

We next investigate the effect of using different video feature representation strategies.

Sampling strategies. The effect of using different sampling strategies for different methods is shown in Figure 4(a). The results show that the performance of all baselines improved by at least 1% when the FS sampling is replaced with the DS sampling. This verifies that the action distribution is an important prior when extracting video features, especially when the videos are long. The sampled frames by the DS sampling can be seen as the *key-frames*, which better reflect fine-grained video semantics. To better understand the differences between the two sampling strategies, we visualize the video units (represented by their central frames) in Figure 5. It can be seen that the DS sampling obtains more representative and diverse video units compared to the FS sampling.

Table 4: The detailed accuracies of the Free type questions.

	Y/N	Color	Obj.	Loc.	Num.	Other
E-VQA	52.7	27.3	7.9	8.8	44.2	20.6
E-MN	55.1	28.0	12.0	12.2	44.4	24.2
E-SA	**59.4**	**29.8**	**14.2**	**25.9**	**44.6**	**28.4**

Sampling frequencies. Figure 4(b) shows the effect of T=$\{20, 40, 60\}$ for E-SA. As the sampling strategy is correlated with the sampling frequency, we report the accuracies with respect to different sampling strategies. The results show that as T increases, the performance gap between the fixed sampling and dynamic sampling narrows. This can be interpreted as denser sampling better preserve the detailed information in videos. Moreover, using the video features generated with dense sampling frequency (*e.g.*, T=60) greatly increase the complexity of VideoQA models, leading to degraded performance.

Figure 6: Visualizations of the results obtained by different methods. For each video example, we show the questions (Q), ground-truth answers (GT) and the predictions of different methods, respectively. The left column shows the examples that at least one method give correct predictions, while the right column shows the examples that all methods give wrong predictions.

Fusion strategies. In Figure 4(c), we explore the effect of using different fusion strategies with (T=$\{20, 40, 60\}$) and FS sampling for E-SA. It can be seen that *Mean Pooling* achieves the best performance compared to other two fusion strategies in terms of accuracy and robustness.

For qualitative analysis, we present some successful and failed cases obtained by different methods in Figure 6. These methods show a greater probability of correctly answering questions that focused on static frame, but fail to answer the questions involving temporal reasoning. These observations are useful for guiding further improvements for VideoQA models in the future.

Conclusion and Future Work

In this paper, we present a new large scale dataset *ActivityNet-QA* for understanding complex web videos by question answering. Compared with existing VideoQA datasets, our dataset is unique in that: 1) the videos originate from ActivityNet, a large-scale video understanding dataset with long web videos; 2) the QA pairs are fully annotated by crowdsourcing. To guarantee the quality of our dataset, we conduct significant pre- and post-processing by both algorithmic and human efforts. Based on the constructed dataset, we apply several baselines to analyze the difficulty of our dataset and also investigate the strategies to learn better video feature representation; and 3) the QA pairs of our dataset are bilingual with alignment. This property may inspire multi-lingual VideoQA studies.

Since the models studied here represent the baseline, there remains significant room for improvement. For example, by introducing a more advanced video feature representation model that can learn better discriminative visual features or introducing a more powerful VideoQA model that can perform accurate spatio-temporal reasoning. Furthermore, auxiliary information on ActivityNet, *e.g.*, dense captions (Krishna et al. 2017) can be utilized to help better understanding the fine-grained semantics of videos.

Acknowledgments

This work was supported in part by National Natural Science Foundation of China under Grant 61702143, Grant 61836002, Grant 61622205 and Grant 61602405, and in part by the China Knowledge Centre for Engineering Sciences and Technology, and in part by the Australian Research Council Projects under Grant FL-170100117, Grant DP-180103424 and Grant IH-180100002.

References

[Anderson et al. 2018] 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. *CVPR*.

[Antol et al. 2015] Antol, S.; Agrawal, A.; Lu, J.; Mitchell, M.; Batra, D.; Lawrence Zitnick, C.; and Parikh, D. 2015. Vqa: Visual question answering. In *ICCV*, 2425–2433.

[Chen, Kovvuri, and Nevatia 2017] Chen, K.; Kovvuri, R.; and Nevatia, R. 2017. Query-guided regression network with context policy for phrase grounding. *ICCV*.

[Donahue et al. 2015] Donahue, J.; Anne Hendricks, L.; Guadarrama, S.; Rohrbach, M.; Venugopalan, S.; Saenko, K.; and Darrell, T. 2015. Long-term recurrent convolutional networks for visual recognition and description. In *CVPR*, 2625–2634.

[Fabian Caba Heilbron and Ghanem 2016] Fabian Caba Heilbron, J. C. N., and Ghanem, B. 2016. Fast temporal activity proposals for efficient detection of human actions in untrimmed videos. In *CVPR*.

[Fabian Caba Heilbron and Niebles 2015] Fabian Caba Heilbron, Victor Escorcia, B. G., and Niebles, J. C. 2015.

Activitynet: A large-scale video benchmark for human activity understanding. In *CVPR*, 961–970.

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

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

[Goyal et al. 2017] Goyal, Y.; Khot, T.; Summers-Stay, D.; Batra, D.; and Parikh, D. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *CVPR*, 6904–6913.

[Hochreiter and Schmidhuber 1997] Hochreiter, S., and Schmidhuber, J. 1997. Long short-term memory. *Neural computation* 9(8):1735–1780.

[Jang et al. 2017] Jang, Y.; Song, Y.; Yu, Y.; Kim, Y.; and Kim, G. 2017. Tgif-qa: Toward spatio-temporal reasoning in visual question answering. In *CVPR*, 2758–2766.

[Johnson et al. 2017a] Johnson, J.; Hariharan, B.; van der Maaten, L.; Fei-Fei, L.; Zitnick, C. L.; and Girshick, R. 2017a. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *CVPR*, 1988–1997.

[Johnson et al. 2017b] Johnson, J.; Hariharan, B.; van der Maaten, L.; Hoffman, J.; Fei-Fei, L.; Zitnick, C. L.; and Girshick, R. B. 2017b. Inferring and executing programs for visual reasoning. In *ICCV*, 3008–3017.

[Kim et al. 2017a] Kim, J.-H.; On, K. W.; Lim, W.; Kim, J.; Ha, J.-W.; and Zhang, B.-T. 2017a. Hadamard Product for Low-rank Bilinear Pooling. In *ICLR*.

[Kim et al. 2017b] Kim, K.-M.; Heo, M.-O.; Choi, S.-H.; and Zhang, B.-T. 2017b. Deepstory: Video story qa by deep embedded memory networks. *IJCAI*.

[Krishna et al. 2016] Krishna, R.; Zhu, Y.; Groth, O.; Johnson, J.; Hata, K.; Kravitz, J.; Chen, S.; Kalantidis, Y.; Li, L.-J.; Shamma, D. A.; et al. 2016. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *arXiv preprint arXiv:1602.07332*.

[Krishna et al. 2017] Krishna, R.; Hata, K.; Ren, F.; Fei-Fei, L.; and Niebles, J. C. 2017. Dense-captioning events in videos. In *ICCV*, 706–715.

[Lu et al. 2016] Lu, J.; Yang, J.; Batra, D.; and Parikh, D. 2016. Hierarchical question-image co-attention for visual question answering. In *NIPS*, 289–297.

[Malinowski and Fritz 2014] Malinowski, M., and Fritz, M. 2014. A multi-world approach to question answering about real-world scenes based on uncertain input. In *NIPS*, 1682–1690.

[Malinowski, Rohrbach, and Fritz 2015] Malinowski, M.; Rohrbach, M.; and Fritz, M. 2015. Ask your neurons: A neural-based approach to answering questions about images. In *ICCV*, 1–9.

[Na et al. 2017] Na, S.; Lee, S.; Kim, J.; and Kim, G. 2017. A read-write memory network for movie story understanding. In *ICCV*.

[Nam, Ha, and Kim 2016] Nam, H.; Ha, J.-W.; and Kim, J. 2016. Dual attention networks for multimodal reasoning and matching. *arXiv preprint arXiv:1611.00471*.

[Pennington, Socher, and Manning 2014] Pennington, J.; Socher, R.; and Manning, C. D. 2014. Glove: Global vectors for word representation. In *EMNLP*, volume 14, 1532–1543.

[Ren, Kiros, and Zemel 2015] Ren, M.; Kiros, R.; and Zemel, R. 2015. Exploring models and data for image question answering. In *NIPS*, 2953–2961.

[Rohrbach et al. 2016] Rohrbach, A.; Rohrbach, M.; Hu, R.; Darrell, T.; and Schiele, B. 2016. Grounding of textual phrases in images by reconstruction. In *ECCV*, 817–834.

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

[Sukhbaatar et al. 2015] Sukhbaatar, S.; Weston, J.; Fergus, R.; et al. 2015. End-to-end memory networks. In *NIPS*, 2440–2448.

[Tapaswi et al. 2016] Tapaswi, M.; Zhu, Y.; Stiefelhagen, R.; Torralba, A.; Urtasun, R.; and Fidler, S. 2016. Movieqa: Understanding stories in movies through question-answering. In *CVPR*, 4631–4640.

[Tran et al. 2015] Tran, D.; Bourdev, L.; Fergus, R.; Torresani, L.; and Paluri, M. 2015. Learning spatiotemporal features with 3d convolutional networks. In *ICCV*, 4489–4497.

[Vedantam, Lawrence Zitnick, and Parikh 2015] Vedantam, R.; Lawrence Zitnick, C.; and Parikh, D. 2015. Cider: Consensus-based image description evaluation. In *CVPR*, 4566–4575.

[Xu et al. 2015] Xu, K.; Ba, J.; Kiros, R.; Cho, K.; Courville, A. C.; Salakhutdinov, R.; Zemel, R. S.; and Bengio, Y. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *ICML*, volume 14, 77–81.

[Xu et al. 2017] 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 *ACM Multimedia*, 1645–1653.

[Yao et al. 2015] Yao, L.; Torabi, A.; Cho, K.; Ballas, N.; Pal, C.; Larochelle, H.; and Courville, A. 2015. Describing videos by exploiting temporal structure. In *ICCV*, 4507–4515.

[Yu et al. 2017] Yu, Z.; Yu, J.; Fan, J.; and Tao, D. 2017. Multi-modal factorized bilinear pooling with co-attention learning for visual question answering. *ICCV* 1821–1830.

[Yu et al. 2018a] Yu, Z.; Yu, J.; Xiang, C.; Fan, J.; and Tao, D. 2018a. Beyond bilinear: Generalized multimodal factorized high-order pooling for visual question answering. *IEEE Transactions on Neural Networks and Learning Systems*. doi:10.1109/TNNLS.2018.2817340.

[Yu et al. 2018b] Yu, Z.; Yu, J.; Xiang, C.; Zhao, Z.; Tian, Q.; and Tao, D. 2018b. Rethinking diversified and

discriminative proposal generation for visual grounding. *IJCAI* 1114–1120.

[Zeng et al. 2017] Zeng, K.-H.; Chen, T.-H.; Chuang, C.-Y.; Liao, Y.-H.; Niebles, J. C.; and Sun, M. 2017. Leveraging video descriptions to learn video question answering. In *AAAI*, 4334–4340.

[Zhao et al. 2018a] 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*, 3690–3696.

[Zhao et al. 2018b] 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*, 3683–3689.