EgoVQA - An Egocentric Video Question Answering Benchmark Dataset

Chenyou Fan
Google
fanchenyou@gmail.com

Abstract

Recently, much effort and attention has been devoted to Visual Question Answering (VQA) on static images and Video Question Answering (VideoQA) on third-person videos. In the meantime, first-person question answering has more natural use cases while this topic remains seldom studied. A typical meaningful scenario is an intelligent agent provides assistance to handicapped people to perceive the environment by the queries, localize objects and persons based on descriptions, and identify intentions of surrounding people to guide their reactions (e.g., shake hands or avoid punches). However, due to the lack of first-person video datasets, seldom study had been carried on first-person VideoQA task. To address this issue, we collected a novel egocentric VideoQA dataset called EgoVQA with 600 question-answer pairs with visual contents across 5,000 frames from 16 first-person videos. Various types of queries such as "Who", "What", "How many" are provided to form a semantically rich corpus. We use this database to evaluate the performance of four mainstream third-person VideoQA methods to illustrate their performance gap between first-person related questions and third-person related questions. We believe that EgoVQA dataset will facilitate future research on the imperative task of first-person VideoQA.

1. Introduction

Egocentric videos are taken from first-person perspective which record the environmental scenes with wearable cameras. Egocentric videos have several noticeable and unique features that make them different from third-person videos. First of all, the camera wearers themselves generally don't appear in their own videos except for their hands and arms. In addition, first-person videos usually record the camera wearers' interactions with other people who may appear in the videos. Last but not least, egocentric videos always record the ego-motions which could make the videos blurry and ambiguous.

Video Question Answering (VideoQA) is a task to infer the corresponding answer given a video and the visual content related question. Existing VideoQA research [5,6,21,35,37,42,59] focused mostly on third-person videos and performed experiments on public third-person video datasets such as TGIF-QA [35], MSVD-QA [54], MSRVTT-QA [54] and Youtube2Text-QA [60].

Video question answering on egocentric video (EgoVQA) is an interesting but less studied topic. However, due to the uniqueness of egocentric video, EgoVQA is more challenging and ambiguous than QA on normal third-person videos. Let us consider the question of "what am I doing ?" – a deceptively straightforward question which is actually difficult to fulfill. If we remember that "I" am not appearing in my own videos, a normal action recognition technique such as C3D [51] may not be directly useful. To make things harder, egocentric videos contain both ego-motions and third-person motions during the interactions between camera wearers and other people. That said, action recognition techniques that rely on motion information (e.g., optical flows in [48] and [15]) could also perform poorly. However, current state-of-the-art third-person VideoQA methods heavily depend on both the third-person appearance and motion information to perform learning and inference. Hereby, direct reuse existing third-person VideoQA methods on first-person videos is not an ideal solution.

The level of difficulty does not drop when we query on objects instead of actions. Let us consider two first-person QA examples: "Q: what is placed on the desk ? A: monitor" and "Q: what am I holding in my hands ? A: a bottle". Both questions require understanding the keywords (e.g., desk and hand) in the question and then localizing the objects according to the positional information (e.g., on desk and in hands). In a static and well composed third-person video, the world coordinate does not shift and the localization is relative easy. The subtleties immediately come when dealing with first-person videos with ego-motions. The positions of my hands in previous frames could be away from the positions in current frame. Therefore, a well-designed video encoder of EgoVQA should take into account of ego-motions properly.

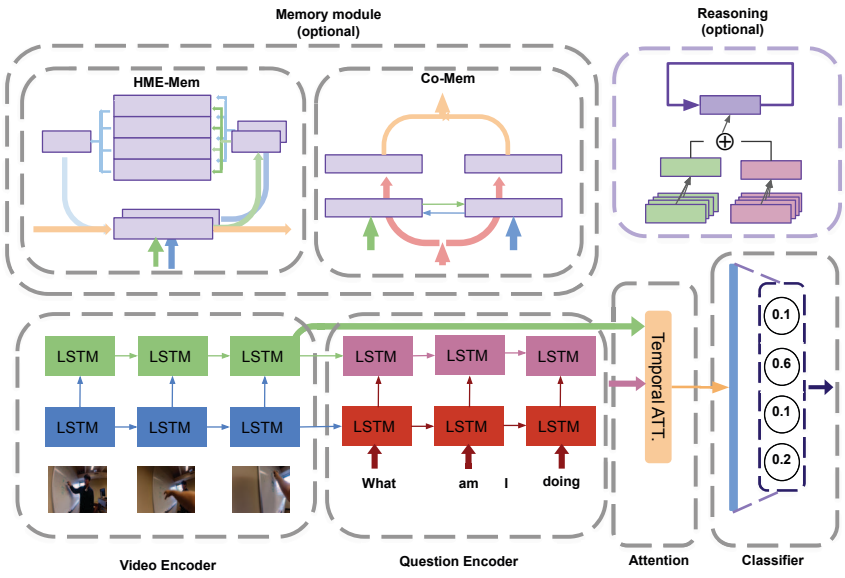

Figure 1. A typical VideoQA neural network architecture with video encoder, question encoder, attention module, classifier, as well as different memory modules and reasoning modules in different architecture designs.

Therefore, we expect that egocentric VideoQA requires special feature and attention designs to suit for its special needs. Motivated by the importance and uniqueness of this topic, we collected a compact egocentric VideoQA dataset named EgoVQA which includes 600 question-answer pairs and covers queries of actions, positions, objects and counting. We describe the details of this dataset and experiment results to show that directly applying existing methods on EgoVQA produces sub-optimal results without considering the characteristics of egocentric videos. We have released our dataset and benchmark code for boosting further research[1].

2. Related Work

Egocentric video study has become an important computer vision task since the prevalence of wearable cameras such as Sensecam [32], GoPro [1], Xiaoyi [2] and Narrative Clips [3]. More and more people are using wearable cameras for various applications such as lifelogging [13,22,30], children behavior study [8,9], sports video analysis [11,12], event detection [50], etc. Recent computer vision research topics on studying first-person videos have included object detection [19,24,25,39], hands detection [10], scene understanding [26], activity recognition [23,47], person identification and segmentation [20,40,57,58], gaze detection and tracking [34,62], etc. Ego-motion, as the signature of first-person videos, has also been especially considered in person identification [20,58] and motion estimation [58,61]. The combination of natural language processing and egocentric video is also an interesting topic. Due to the redundancy and ambiguity of egocentric videos, several research paper [14, 22] proposed to perform video captioning and summarization to abstract and categorize egocentric videos.

Visual Question Answering (VQA) and Video Question Answering (VideoQA) have become emerging research topics [4,5,21,27,42,43,59,60] in the past few years. VQA and VideoQA study how to infer the correct answers of given questions related to the visual contents of images and videos respectively. Existing studies have been carried on images or videos from third-person perspective. However, there is a lack of egocentric QA benchmark dataset to facilitate QA study on egocentric videos.

Perhaps the most related work are EmbodiedQA [18,53] which proposed to use robot agent to navigate the environment and fulfill the query intents. Though their designed QA module and existing VideoQA methods share similar designs such as encoder -decoder and attention mechanisms, their work focused on learning navigation policies in simulated environment, while we considered real-world scenarios that how QA methods could better understand the footage of wearable cameras to perceive the world even with considerable ego-motions.

To evaluate the capacities of existing VideoQA methods on egocentric videos, we propose a benchmark egocentric VideoQA dataset and establish the baselines by applying existing third-person VideoQA methods on egocentric videos directly. This paper will be organized as follows. In Section 3, we will give a brief survey of current VideoQA techniques that were designed for QA on third-person videos. In Section 4, we will review existing third-person VideoQA datasets. In Section 5, we will show the details and samples of our EgoVQA dataset. In Section 6,

[1]https://github.com/fanchenyou/EgoVQA.git

we will show the performance of existing methods when directly applied to our new EgoVQA dataset as baselines. Finally we will discuss potential future efforts that could improve this novel task of egocentric VideoQA.

3. VideoQA Approaches

In this section, we will briefly review existing VideoQA frameworks and main modules. We first introduce the mainstream VideoQA architecture ST-VQA [35] with its encoder-decoder and attention design, then we introduce its variants that either add new components such as neural memory modules [21] or modify existing attention modules [27, 54]. We will investigate the shortages of current VideoQA methods when applied to egocentric VideoQA task, and briefly discuss potential directions of improving egocentric VideoQA.

3.1. Video Encoder

As shown in *Video Encoder* module in Figure 1, previous work [21, 27, 35, 54, 60] extracted video features from sampled video frames with pre-trained neural networks such as ResNet [31], VGG [49] and C3D [51] to obtain appearance and motion features. Some work [35, 54] early fused features and fed into LSTM or GRU based video encoders to obtain encoded video features. Fan *et al*. [21] and Gao *et al*. [27] applied late fusion on different types of video features and integrated them with memory modules. The illustrations of their different memory designs HME-Mem [21] and Co-Mem [27] are shown by *Memory* module in Figure 1. However, these techniques were designed to understand third-person videos without considering the entangled ego-motions and third-person motions concurred in first-person videos.

3.2. Question Encoder

Similar to image captioning [36] and language translation [52], a question is tokenized as a word sequence and each word is represented as a fixed-length word embedding in previous work [21, 27, 35, 54, 60]. These word embeddings are commonly initialized with the pre-trained GloVe 300-D [46] features. Similar to video encoder, a separate LSTM or GRU based encoder is adopted to encode word sequence, as shown by *Question Encoder* module in Figure 1.

3.3. Attention Mechanism

Attention mechanism has been widely applied in sequence-to-sequence modeling such as machine translation [7, 33, 52] and image captioning [56]. Applying attention mechanism to associate the question with most relevant frames has become an essential part of existing VideoQA techniques. As shown by *Attention* block in Figure 1, Jang *et al*. [35] applied temporal attention mechanism by interacting the encoded question feature with encoded video features and generated soft attended video features by measuring the importance of each frame. Xu *et al*. [54] further applied attention on each step of question encoding stage to gradually refine the attention on each word. Niu *et al*. [44] applied recursive attention to question answering dialog to refine the attention with the progress of the conversation.

However, current attention mechanisms, if directly shared across and learned on both first-person and third-person questions, could generate conflicted attentions on first-person video sets.

3.4. Reasoning

Instead of simply applying temporal attention once to build the final answer abstraction, several work [21, 54, 60] adopted multi-step reasoning to better relate question words with encoded video features. make complex reasoning in multiple cycles with refined attention such as [27, 60]. As shown by *Reasoning* block in Figure 1, a typical multi-step reasoning module utilizes a controller such as AMU [54] or LSTM [21] which iteratively refines the attention over frames given the question feature or the co-attention between frames and question words. However, if generated attentions were incorrect, such that attentions were paid to third-person actions for first-person queries, multi-step reasoning could make attentions even more unreliable.

4. Existing Dataset

In this section, we briefly review four existing third-person VideoQA datasets which are publicly available and commonly used by previous work. In Table 1, we adopted the statistics of existing datasets from [21] with minor modifications. Also we will compare them with our proposed EgoVQA dataset in next section.

As shown in Table 1 *3rd-Person* part, all existing VideoQA datasets are from third-person perspective. **TGIF-QA** [35] is a dataset of over 165,000 questions on 71741 animated pictures originated from TGIF dataset [41]. Multiple tasks are formulated upon this dataset including counting repetitions of the queried action, detecting transitions of two actions and image-based QA. **MSVD-QA** and **MSRVTT-QA** [54] are two datasets with third-person videos originated from MSVD [16] and MSVTT [55] respectively. The VideoQA tasks formulated in both of these two datasets are open-ended questions of types *what, who, how, when and where*, and their answer sets are of size 1000. **YouTube2Text-QA** [60] is a dataset with both open-ended and multiple-choice tasks of three major question types (*what, who and other*. The video source is MSVD [16] while the questions are derived from YouTube2Text [28] video description corpus.

Perspective	Dataset	Vocab size	Video num	Question num			Answer set	Choice num
				Train	Val	Test		
3rd-Person	TGIF-QA [35]	8,000	71,741	125,473	13,941	25,751	1746	5
	MSVD-QA [54]	4,000	1,970	30,933	6,415	13,157	1000	NA
	MSRVTT-QA [54]	8,000	10,000	158,581	12,278	72,821	1000	NA
	Youtube2Text-QA [60]	6,500	1,970	88,350	6,489	4,590	1000	4
1st-Person	EgoVQA (ours)	4000	520	250	150	120	101	5

Table 1. Dataset statistics of four existing third-person VideoQA datasets and our EgoVQA. The columns from left to right indicate the video perspective, dataset name, vocabulary size, sampled video length, number of videos, size of QA splits, pre-defined answer set size for open-ended questions and number of options for multiple-choice questions.

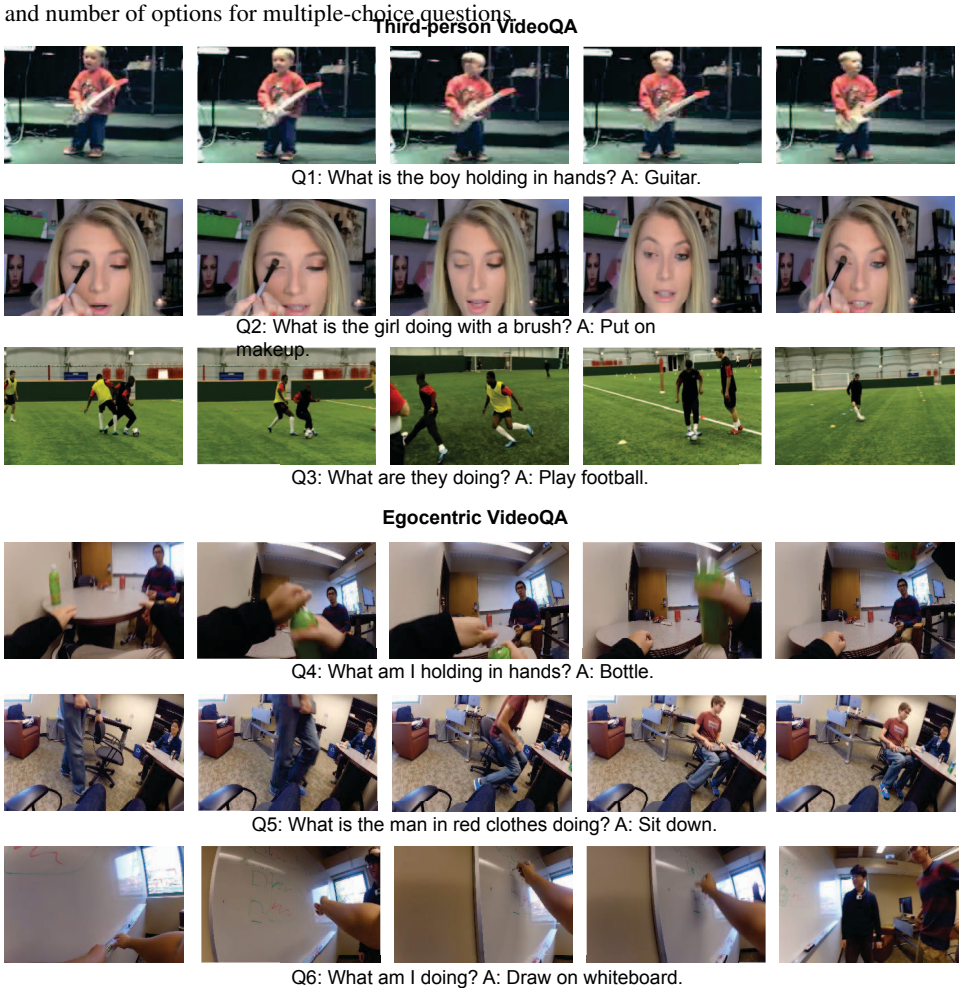

Figure 2. Examples of VideoQA on third-person videos and egocentric videos.

5. EgoVQA

In this paper, we propose an Egocentric VideoQA dataset named EgoVQA. Built upon an existing egocentric video dataset, we manually annotated more than 600 question-answer pairs with eight designed question types.

5.1. Egocentric Videos

The video source of EgoVQA is the public IU Multi-view dataset[2] which were collected for multi-view egocentric video studies [20, 57]. In original IU Multiview dataset, there are 8 sets of indoor scenarios, and each has two synchronized first-person videos taken by two different participants in that scenario with wearable cameras. Each first-person video lasts 5-10 minutes, and each scenario has 3-4 participants performing actions such as having snacks, shaking hands, drawing on whiteboard, etc. We extend IU Multiview dataset with 580 QA pairs covering the same amount of video clips selected from 16 long first-person videos. Each video clip lasts from 20 to 100 seconds. We manually select those video clips that satisfy one or more of

[2]http://vision.soic.indiana.edu/firstthird-eccv2018

the following conditions: when the camera wearer was performing personal actions such as drinking, having snacks, typing, etc; when the camera wearer was interacting with other persons such as shaking hands; when other participants were performing intended actions; and there were recognizable objects in camera wearers' views. Based on the video contents, we annotated eight types of questions which will be discussed in next section.

In Figure 2, we illustrate the difference of existing third-person VideoQA annotations and our EgoVQA annotations. The top three video samples (Q1-Q3) from MSVD [16] show that typical third-person videos are stable and well composed; the main characters are appearing in the center of the frames; and the actions are clearly recognizable and well defined. In contrast, the bottom three video samples from our EgoVQA dataset illustrate that egocentric videos are blurry and ambiguous; the camera wearers are not in their own videos except for their hands and arms; and the actions are generally more difficult to recognize due to ego-motions and camera angles.

5.2. Question Annotation

To generate QA pairs from video clips, we spent about 30-hour human time manually annotating questions and corresponding correct answers in English sentences. Each question-answer pair is querying and answering visually related content of the video clip. In Table 2, we list QA examples of eight major types of questions in EgoVQA dataset. **1st action** questions query the egocentric actions of camera wearers, e.g., "what am I doing"; **3rd action** questions query the third-person actions from views of camera wearers, e.g., "what is the man in red clothes doing"; **1st who** questions query the persons who were interacting with the camera wearers, e.g., "who am i talking with"; **3rd who** questions query actions of other persons, e.g., "who is standing beside the door"; **Count** questions ask the number of persons or objects in the scenes; and **Color** questions ask colors of major objects or clothes of participants in the scenes.

5.3. Answer Generation

We now describe our approach to generate the candidate answers for multiple-choice questions, and the answer set for open-ended questions. **Multiple-choice** task of VideoQA requires to choose the only correct answer out of K candidates. In EgoVQA, each multiple-choice question comes with five candidate answers in which exactly one is correct. To generate other confusing options, we categorize questions by types of "what action", "what object", "who", "count" and "color", and then for each category we aggregate all the correct answers as a candidate pool for that category. We randomly sample four candidates from the pool without replacement as negative candidates. We apply this strategy to generate candidates for "what action", "what object", "who" and "color" questions as well. For "count" questions, we fix the candidate answers to be zero to four. This sampling strategy avoids generating simple candidates which can be inferred simply by the grammar or context of the question without understanding the visual contents. **Open-ended** task is to choose one correct word or short phrase as the answer from a pre-defined answer set. We generate an answer set of 101 words or phrases by aggregating unique correct answers of the entire dataset. We reserve this task for future research.

5.4. Splits

For robust evaluation purposes, we generate three distinct training, evaluation and testing splits from the database. For the eight scenarios (numbered from 1 to 8) in IU Multiview dataset, we build three train/validation/test sets with scenario numbers of $4 : 2 : 2$. Each set reserves four scenarios (eight videos as there are two first-person cameras) for training, two scenarios (four videos) for validation and testing respectively. As we extract multiple video clips (each of 25-100 seconds) from every first-person video, this way of splitting ensures that video clips from the same videos are not in both training and testing sets. The total numbers of video clips for training, validation and testing are shown in Table 3.

6. Experiments and Discussions

We evaluate four existing third-person VideoQA models with our EgoVQA dataset and report their performance in percent accuracy on test set of all three splits. Specifically, we compare ST-VQA [35] (w/ and w/o attention mechanism), Co-Mem [27] and HME-VQA [21] with respect to Multiple-Choice task (five-way classification) on all three splits of EgoVQA dataset. Note that all models were originally designed for third-person VideoQA. Since our EgoVQA dataset is relatively small in terms of training samples and vocabulary size, we pre-trained all four different models on YouTube2Text-QA [60] dataset (please refer to Table 1 for details) and then fine-tuned on our EgoVQA dataset. Without pre-training, we found the models tend to overfit quickly.

6.1. Overall result

In Table 4, we report the percent accuracy of each model's performance on each of three splits (Col. *1,2,3*), the average accuracy of all splits (Col. *Avg*), as well as the performance boost by pre-training (Col. *Init* $\uparrow$). In overall, HME-VQA [21] outperformed the other three methods, which is consistent to their reported results on third-person video datasets. Surprisingly, ST-VQA [35] with temporal attention does not outperform ST-VQA without attention in accuracy. Co-Mem [27], with dual attention mechanisms,

Type	Question	Answer
1st action	what am i doing	have snack
1st action	what am i doing	point on whiteboard
3rd action	what is the man in black clothes doing	draw on white board
3rd action	what is the man on my right side doing	open snack bag
1st who	who am i talking with	the man standing in front of me
1st who	who am i looking at	a man in blue
3rd who	who is standing beside the door	the man in blue shirt
3rd who	who is sitting on the chair	the man on my left side
1st obj	what am i holding in hands	phone
1st obj	what am i holding in hands	laptop
3rd obj	what is the man in front of me holding	pen
3rd obj	what is the man in gray clothes holding	bottle
count	how many people can i see in the scene	three
color	what is the color of the toy in my hands	blue

Table 2. Sample questions of different types in EgoVQA.

Split No.	Video No.	Train Num	Val Num	Test Num
1	{1,2,3,4}{5,6}{7,8}	276	173	132
2	{1,3,5,7}{2,8}{4,6}	241	167	173
3	{1,5,6,8}{4,7}{2,3}	283	111	187

Table 3. Three train/val/test splits of EgoVQA datasets.

Method	Acc(%) on Split				
	1	2	3	Avg	Init ↑
ST-VQA w/o att	31.82	35.26	30.48	32.52	+4.71
ST-VQA [35]	31.82	**37.57**	27.27	32.33	+8.16
Co-Mem [27]	**32.58**	32.37	25.13	30.03	+1.15
HME-VQA [21]	**32.58**	36.42	**31.02**	**33.34**	+2.86

Table 4. Experiment results on EgoVQA dataset. The individual and average classification accuracy on three splits are shown. The last column (Init ↑) shows the accuracy

under-performed all other methods. We conjecture that the strong ego-motions of egocentric videos bring incorrect attentions due to the lack of capacity to separate attentions on camera wearers from attentions on third persons. The memory slots inside the memory module of HME-VQA alleviated such problem (see Figure 1) by implicitly learning separate attentions of different factors, yet its performance advantage is less than 1%. Future work could focus on disentangling first-person and third-person attentions to better find out relevant visual contents. We also find out that pretraining each model on a large third-person dataset and then fine-tuning on EgoVQA could increase accuracy up to 8%, as the word vectors and video encoders are better trained.

6.2. Per-category result

In Table 5, we further compare the performance of different models on different question types. From column 1 and 2, we observe that he first-person action query accuracy is about 9% less than third-person action query (32.84% v.s. 41.67%) on the best-performing HME-VQA, and 2% and 5% for ST-VQA and Co-Mem. This suggests that classifying actions of camera wearers is generally harder than actions of third persons in the scene. This is understandable as camera wearers don't appear in own videos and only ego-motion evidence is usable for classification. Also pretraining on third-person videos could benefit third-person queries more on first-person videos. From column 3 and 4, we observe that 1st-person object query has generally higher accuracy than 3rd-person object query. This is likely that the objects that camera-wearers interact with are usually near to the cameras and centered in the frames. Similar trend also applies on who queries, as first-person who queries have generally better accuracy than third-person who queries (Column 5 and 6). For color queries, we found most of the models fail to outperform the random guessing (20%) by more than 5%. It is likely that localizing the queried object is already a hard task while assigning color to noisily localized pixels becomes even harder. Future work could better localize queried objects using object detection or segmentation techniques to better refine attended areas in image planes.

6.3. Future Work

To mitigate the gaps of action query accuracy on first- and third-person videos, we conjecture that disentangling first- and third-person attentions could be the first step. As a direct approach, it's worth trying to **explicitly** estimating ego-motions and third-person motions simultaneously and learning first-person attentions on frames of significant ego-motions for first-person queries. While jointly learning first- and third-person attentions by forming a dual problem (first-person question in camera A would become third-person question in camera B, vice versa) could also **implicitly** refine both first- and third-person attentions. The

Method	Question type							
	Action 1st(67)	Action 3rd(108)	Object 1st(54)	Object 3rd(86)	Who 1st(13)	Who 3rd(63)	Count(64)	Color(31)
ST-VQA w/o att	29.85	29.63	**37.04**	30.23	38.46	**38.10**	**39.06**	22.58
ST-VQA [35]	28.36	30.56	31.48	**31.40**	**46.15**	34.92	35.94	**32.26**
Co-Mem [27]	25.37	30.56	**37.04**	29.07	38.46	26.98	31.25	22.58
HME-VQA [21]	**32.84**	**41.67**	27.78	30.23	23.08	28.57	42.19	22.58

Table 5. Experiment results on different types of questions.

EgoVQA, with both first- and third-person queries on synchronized multi-camera videos, supports both directions of research as a benchmark dataset.

6.4. Implementation details

We implemented the benchmark neural networks in PyTorch [45] and updated the parameters with Adam optimizer [38]. Across the experiments, we use a fixed batch size of 32 and learning rate of 10^{-3}. For the video and text encoders inside each model, we use a same two-layer LSTMs with hidden size 256. For the HME-VQA method, we set the video and question memory sizes to the same as LSTM hidden size. We will release the EgoVQA dataset, pre-trained models and evaluated methods to encourage further research on this topic.

7. Conclusion

In this paper, we proposed an egocentric VideoQA dataset named EgoVQA which is, to our knowledge, the first VideoQA dataset dedicated for egocentric video studies. We also established the baselines by evaluating existing state-of-the-art third-person VideoQA methods. Experimental results suggested that existing methods generally lack the ability to handle ego-motions or separate attentions for first-person activities and third-person activities. We will make public of the dataset and code to boost further studies on this novel and important research topic. The author also encourages the community to annotate question-answer pairs on larger datasets such as ImageCLEF Lifelog [17] and NTCIR Lifelog [29] datasets to explore larger neural network designs.

References

[1] http://www.gopro.com.

[2] http://www.xiaoyi.com.

[3] http://www.getnarrative.com.

[4] A. Agrawal, J. Lu, S. Antol, M. Mitchell, C. L. Zitnick, D. Batra, and D. Parikh. VQA: Visual question answering. In *ICCV*, 2015.

[5] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *CVPR*, 2018.

[6] J. Andreas, M. Rohrbach, T. Darrell, and D. Klein. Neural module networks. In *CVPR*, 2016.

[7] D. Bahdanau, K. Cho, and Y. Bengio. Neural machine translation by jointly learning to align and translate. In *ICLR*, 2015.

[8] S. Bambach, D. Crandall, L. Smith, and C. Yu. Toddler-inspired visual object learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

[9] S. Bambach, D. J. Crandall, L. B. Smith, and C. Yu. An egocentric perspective on active vision and visual object learning in toddlers. In *ICDL)*, 2017.

[10] S. Bambach, S. Lee, D. Crandall, and C. Yu. Lending a hand: Detecting hands and recognizing activities in complex egocentric interactions. In *ICCV*, 2015.

[11] G. Bertasius and J. Shi. Using cross-model egosupervision to learn cooperative basketball intention. In *ICCV*, 2017.

[12] G. Bertasius, H. Soo Park, S. X. Yu, and J. Shi. Am i a baller? basketball performance assessment from first-person videos. In *CVPR*, 2017.

[13] M. Bolanos, M. Dimiccoli, and P. Radeva. Toward storytelling from visual lifelogging: An overview. *IEEE Transactions on Human-Machine Systems*, 2016.

[14] M. Bolaños, Á. Peris, F. Casacuberta, S. Soler, and P. Radeva. Egocentric video description based on temporally-linked sequences. *Journal of Visual Communication and Image Representation*, 2018.

[15] J. Carreira and A. Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *CVPR*, 2017.

[16] D. L. Chen and W. B. Dolan. Collecting highly parallel data for paraphrase evaluation. In *ACL*, 2011.

[17] D.-T. Dang-Nguyen, L. Piras, M. Riegler, M.-T. Tran, L. Zhou, M. Lux, T.-K. Le, V.-T. Ninh, and C. Gurrin. Overview of ImageCLEFlifelog 2019: Solve my life puzzle and Lifelog Moment Retrieval. In *CLEF2019 Working Notes*.

[18] A. Das, S. Datta, G. Gkioxari, S. Lee, D. Parikh, and D. Batra. Embodied Question Answering. In *CVPR*, 2018.

[19] C. Fan, J. Lee, and M. S. Ryoo. Forecasting hands and objects in future frames. In *AHB@ECCVW*, 2018.

[20] C. Fan, J. Lee, M. Xu, K. Kumar Singh, Y. Jae Lee, D. J. Crandall, and M. S. Ryoo. Identifying first-person camera wearers in third-person videos. In *CVPR*, 2017.

[21] C. Fan, X. Zhang, S. Zhang, W. Wang, C. Zhang, and H. Huang. Heterogeneous memory enhanced multimodal attention model for video question answering. In *CVPR*, 2019.

[22] C. Fan, Z. Zhang, and D. J. Crandall. Deepdiary: Lifelogging image captioning and summarization. *Journal of Visual Communication and Image Representation*, 2018.

[23] A. Fathi, Y. Li, and J. M. Rehg. Learning to recognize daily actions using gaze. In *ECCV*. 2012.

[24] A. Fathi, X. Ren, and J. M. Rehg. Learning to recognize objects in egocentric activities. In *CVPR*, 2011.

[25] A. Furnari, S. Battiato, K. Grauman, and G. M. Farinella. Next-active-object prediction from egocentric videos. *Journal of Visual Communication and Image Representation*, 2017.

[26] A. Furnari, G. M. Farinella, and S. Battiato. Recognizing personal contexts from egocentric images. In *EPIC@ICCV*, 2015.

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

[28] S. Guadarrama, N. Krishnamoorthy, G. Malkarnenkar, S. Venugopalan, R. Mooney, T. Darrell, and K. Saenko. Youtube2text: Recognizing and describing arbitrary activities using semantic hierarchies and zero-shot recognition. In *ICCV*, 2013.

[29] C. Gurrin, H. Joho, F. Hopfgartner, L. Zhou, and R. Albatal. Ntcir lifelog: The first test collection for lifelog research. In *SIGIR*, 2016.

[30] C. Gurrin, A. F. Smeaton, D. Byrne, N. Hare, G. J. Jones, and N. Connor. An examination of a large visual lifelog. In *Information Retrieval Technology*. 2008.

[31] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

[32] S. Hodges, L. Williams, E. Berry, S. Izadi, J. Srinivasan, A. Butler, G. Smyth, N. Kapur, and K. Wood. Sensecam: A retrospective memory aid. In *International Conference on Ubiquitous Computing*, 2006.

[33] C. Hori, T. Hori, T.-Y. Lee, Z. Zhang, B. Harsham, J. R. Hershey, T. K. Marks, and K. Sumi. Attention-based multimodal fusion for video description. In *ICCV*, 2017.

[34] Y. Huang, M. Cai, Z. Li, and Y. Sato. Predicting gaze in egocentric video by learning task-dependent attention transition. In *ECCV*, 2018.

[35] Y. Jang, Y. Song, Y. Yu, Y. Kim, and G. Kim. TGIF-QA: Toward spatio-temporal reasoning in visual question answering. In *CVPR*, 2017.

[36] A. Karpathy and L. Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *CVPR*, 2015.

[37] A. Kembhavi, M. Seo, D. Schwenk, J. Choi, A. Farhadi, and H. Hajishirzi. Textbook question answering for multimodal machine comprehension. In *CVPR*, 2017.

[38] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015.

[39] M. Korayem, R. Templeman, D. Chen, D. Crandall, and A. Kapadia. Enhancing lifelogging privacy by detecting screens. In *CHI*, 2016.

[40] Y. J. Lee, J. Ghosh, and K. Grauman. Discovering important people and objects for egocentric video summarization. In *CVPR*, 2012.

[41] Y. Li, Y. Song, L. Cao, J. Tetreault, L. Goldberg, A. Jaimes, and J. Luo. Tgif: A new dataset and benchmark on animated gif description. In *CVPR*, 2016.

[42] C. Ma, C. Shen, A. Dick, Q. Wu, P. Wang, A. van den Hengel, and I. Reid. Visual question answering with memory-augmented networks. In *CVPR*, 2018.

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

[44] Y. Niu, H. Zhang, M. Zhang, J. Zhang, Z. Lu, and J.-R. Wen. Recursive visual attention in visual dialog. In *CVPR*, 2019.

[45] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer. Automatic differentiation in pytorch. 2017.

[46] J. Pennington, R. Socher, and C. Manning. GloVe: Global vectors for word representation. In *EMNLP*, 2014.

[47] M. Ryoo and L. Matthies. First-person activity recognition: What are they doing to me? In *CVPR*, 2013.

[48] K. Simonyan and A. Zisserman. Two-stream convolutional networks for action recognition in videos. In *NIPS*, 2014.

[49] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. 2014.

[50] Y.-C. Su and K. Grauman. Detecting engagement in egocentric video. In *ECCV*, 2016.

[51] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri. Learning spatio-temporal features with 3d convolutional networks. In *ICCV*, 2015.

[52] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. In *NIPS*, 2017.

[53] E. Wijmans, S. Datta, O. Maksymets, A. Das, G. Gkioxari, S. Lee, I. Essa, D. Parikh, and D. Batra. Embodied Question Answering in Photorealistic Environments with Point Cloud Perception. In *CVPR*, 2019.

[54] D. Xu, Z. Zhao, J. Xiao, F. Wu, H. Zhang, X. He, and Y. Zhuang. Video question answering via gradually refined attention over appearance and motion. In *ACMMM*, 2017.

[55] J. Xu, T. Mei, T. Yao, and Y. Rui. MSR-VTT: A large video description dataset for bridging video and language. In *CVPR*, 2016.

[56] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *ICML*, 2015.

[57] M. Xu, C. Fan, Y. Wang, M. S. Ryoo, and D. J. Crandall. Joint person segmentation and identification in synchronized first- and third-person videos. In *ECCV*, 2018.

[58] T. Yagi, K. Mangalam, R. Yonetani, and Y. Sato. Future person localization in first-person videos. In *CVPR*, 2018.

[59] Z. Yang, X. He, J. Gao, L. Deng, and A. Smola. Stacked attention networks for image question answering. In *CVPR*, 2016.

[60] Y. Ye, Z. Zhao, Y. Li, L. Chen, J. Xiao, and Y. Zhuang. Video question answering via attribute-augmented attention network learning. In *SIGIR*, 2017.

[61] R. Yonetani, K. M. Kitani, and Y. Sato. Ego-surfing first-person videos. In *CVPR*, 2015.

[62] M. Zhang, K. Teck Ma, J. Hwee Lim, Q. Zhao, and J. Feng. Deep future gaze: Gaze anticipation on egocentric videos using adversarial networks. In *CVPR*, 2017.