Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 4352–4358
Marseille, 11–16 May 2020

LifeQA: A Real-life Dataset for Video Question Answering

Santiago Castro, Mahmoud Azab, Jonathan C. Stroud, Cristina Noujaim, Ruoyao Wang, Jia Deng, Rada Mihalcea
University of Michigan
{sacastro, mazab, stroud, cnoujaim, ruoyaow, jiadeng, mihalcea}@umich.edu

Abstract

We introduce LifeQA, a benchmark dataset for video question answering that focuses on day-to-day real-life situations. Current video question answering datasets consist of movies and TV shows. However, it is well-known that these visual domains are not representative of our day-to-day lives. Movies and TV shows, for example, benefit from professional camera movements, clean editing, crisp audio recordings, and scripted dialog between professional actors. While these domains provide a large amount of data for training models, their properties make them unsuitable for testing real-life question answering systems. Our dataset, by contrast, consists of video clips that represent only real-life scenarios. We collect 275 such video clips and over 2.3k multiple-choice questions. In this paper, we analyze the challenging but realistic aspects of LifeQA, and we apply several state-of-the-art video question answering models to provide benchmarks for future research. The full dataset is publicly available at https://lit.eecs.umich.edu/lifeqa/.

Keywords: natural language processing, question answering, video question answering, computer vision

What is the name of the younger girl?		How many people are playing?	
A. Caitlin	B. Lucy	A. 2	B. 4
C. Jane	D. Cindy	**C. 3**	D. 1

Figure 1: An instance from LifeQA. The image shows a frame from the video, part of the transcriptions, two questions along with the candidate answers, and the correct answers in bold.

1. Introduction

Video Question Answering (Video QA) is one of the most challenging and crucial problems for artificial intelligence. In this task, we are given a video and must answer natural language questions about its content, such as "What game is the little girl playing?". Answering these questions requires a rich understanding of the visual and auditory content in the video, as well as the ability to relate this content to natural language concepts. Like many challenging tasks, much of the recent progress on Video QA is due to the introduction of several large-scale datasets, which consist primarily of movies and TV shows (Tapaswi et al., 2016; Rohrbach et al., 2017; Lei et al., 2018). Movies and TV shows provide for countless hours of clean, crisply-edited video and accurately-captioned audio, and are therefore easily adapted into datasets. However, these same features mean that movies and TV are not representative of day-to-day life. Therefore, these datasets cannot be used to evaluate how well models perform when applied to realistic videos of day-to-day life.

To address this issue, we introduce **Life Question Answering** (**LifeQA**), a Video QA benchmark dataset that consists of videos and questions about day-to-day life. LifeQA is drawn from hand-picked YouTube videos, which depict scenarios such as children playing, a family having a meal together, or a snapshot from a daycare. These videos are not professionally shot, edited, or scripted, making them much more representative of daily life than prior datasets. They also benefit from increased diversity in terms of the number of people and scenes that appear, since they are not drawn for a small set of shows or films. In addition, the questions include few proper names or references to known locations, which are commonly referenced in TV datasets that feature well-known characters (such as "Sheldon", or "Monica's apartment"), and therefore the questions have to be answered without prior knowledge about the scene. Moreover, the questions are challenging as they cover visual grounding ("what color is the blanket?"), intent ("what does the father want to do with the box?"), and commonsense reasoning ("what is in the bottle?"), all hallmarks of a comprehensive QA dataset.

LifeQA consists of 275 videos and 2,326 multiple-choice questions, making it a suitable complement for existing datasets and a challenging benchmark for existing Video QA systems. To enable future research, we are making LifeQA publicly available, along with automatically and manually generated transcriptions (from the speech in the audio channel) and pre-computed features for every video. In this paper, we describe the LifeQA dataset, present several analyses, and evaluate the performance of several baselines that highlight the difficulty of the task.

2. Related Work

2.1. Text-based Question Answering

Question answering based on text has been extensively explored (Richardson et al., 2013; Hermann et al., 2015; Weston et al., 2015). Early question answering systems were developed for restricted domains, relied on manually crafted features, and had limited capabilities (Katz et al., 2002; Soricut and Brill, 2004; Benamara, 2004). Recently, the rise of deep learning methods motivated the need for large question answering datasets to leverage the capabilities of such models. With that goal in mind, several large-scale reading comprehension datasets were introduced (Rajpurkar et al., 2016; Richardson et al., 2013; Bajgar et al., 2016;

Nguyen et al., 2016). (Rajpurkar et al., 2016) introduced the SQuAD dataset, which is composed of Wikipedia articles and the answers are specified as spans from a text passage. Similarly, (Richardson et al., 2013) collected the MCTest dataset, a multiple-choice open-domain reading comprehension dataset. Given a paragraph, a question, and a set of multiple answers, the task of a QA system is to select the correct answer.

2.2. Multimodal Question Answering

Recently, question answering systems have been constructed to answer questions about other modalities, such as images (Visual QA) and video (Video QA). For the former, several datasets have been proposed such as VQA (Agrawal et al., 2017), Visual7W (Zhu et al., 2016), VisDial (Das et al., 2017), GQA (Hudson and Manning, 2019) and DREAM (Sun et al., 2019). These benchmarks aim to help building visual understanding systems that can reason about the contents of a given image. Given an image and a question, the system would either select a correct answer from multiple choices or generate a free-form textual answer.

Video QA is more challenging, in that it allows for a broader range of question types, and requires the use of temporal information. Many datasets have been proposed for Video QA, such as LSMDC 16 (Rohrbach et al., 2017), TGIF-QA (Jang et al., 2017), MovieQA (Tapaswi et al., 2016), PororoQA (Kim et al., 2017), MarioQA (Mun et al., 2017), VCQA (Zhu et al., 2017), TVQA (Lei et al., 2018), and ActivityNet-QA (Yu et al., 2019). LSMDC, TGIF-QA, PororoQA, and MarioQA consist of short video clips (just a few seconds), which is difficult to understand what is going on in a scene beyond several actions that can be identified. Additionally, they are completely dependent the on visual cues, with no presence of speech and other audio cues.

MovieQA and TVQA consist of movies and TV series. The questions and answers were generated based on the dialog and visual information presented in short video clips from TV shows. However, these acted and well-directed video clips are hard to find in the real world. As in them, we constructed our questions and answers based on both textual and visual cues from short video clips. However, unlike them, our proposed dataset relies on video clips that were recorded naturally by people, without predefined scripts. Therefore, understanding videos requires overcoming challenges such as environmental noise, camera movements, lighting conditions, and naturally occurring dialogues. In addition, scenes are less defined, with undefined characters, lack of subject permanence, and sometimes incoherent conversations. That makes our dataset more challenging for Visual QA tasks.

2.3. In-the-Wild Datasets

Recent work in computer vision has focused on evaluating models "in the wild" — that is, on realistic datasets that depict real-life situations. This is evident in recent video datasets, such as Charades (Sigurdsson et al., 2016) and VLOG (Fouhey et al., 2018), both of which include indoor scenes of human activities. These datasets include rich annotations about human actions, objects, and scenes, but do not include questions and answers as in LifeQA. To the best of our knowledge, our LifeQA dataset is the first real-life dataset for Video QA.

ActivityNet QA consists of short YouTube clips originally selected for an activity recognition dataset (Heilbron et al., 2015). Unlike our dataset, these datasets do not explicitly include videos of real-life settings.

VCQA (Zhu et al., 2017) consists of cooking and in-the-wild YouTube videos (about half of the dataset), and clips from movies (the other half). Questions in VCQA are automatically generated from templates and are not written by humans. Additionally, these automatically generated questions only focus on nouns and verbs, as well as short-term temporal reasoning questions, while in LifeQA we have a more challenging question set about reasons, emotions, and locations. Moreover, VCQA does not consider dialogues, texts, and audio information, which are equally important to understand real-life scenes.

3. LifeQA Dataset

3.1. Dataset Collection

To collect this dataset, we begin by searching for videos on YouTube, using manually-chosen keywords that lead to videos of people living out their daily lives in varied settings (e.g., "my morning routine," "dialogue," "kids playing," "class in elementary school" and "watching TV"). We then hand-pick 59 such videos, based on the condition that they must contain recordings of natural interactions in natural settings. We explicitly exclude videos that do not contain language interactions.

The identification of such videos turns out to be a challenging task, requiring significant manual effort. This is primarily because most of the recordings available online are in the form of vlogs, which include video recordings with voice layovers, and are therefore not typical of natural interactions.

We manually split the source videos into 275 video clips such that each clip includes coherent scenes and lasts for 1–2 minutes. We obtain transcriptions for the video clips using the Google Cloud Speech-to-Text platform. We also collect manual transcriptions for each video.

Next, two annotators write five questions per video. For each question, we ask the annotators to write the correct answer to the question as well as three distractors (which we define as incorrect but semantically related answers). The annotators are instructed to formulate a diverse set of questions, which require an understanding of both the visual and linguistic content of the videos. We then instruct a third annotator to merge the two sets of questions from the original annotators, manually eliminate any duplicate questions, and correct typographical errors. In total, we collect 2,326 questions using this procedure.

We present a summary of the dataset in Table 1. Figure 1 shows an example from the LifeQA dataset, showing two sample questions that require either linguistic or visual clues to be answered. Additional questions are illustrated in Figure 5.

3.2. Dataset Analysis

We examine LifeQA's common question types in Figure 2. A majority of the questions are "what" questions, which were previously acknowledged to be among the most frequent and

Source videos	59
Clips	275
Clips per source video	4.7 ± 3.6
Clip duration	1m 14s ± 16s
Modalities	video, audio, text
Questions	2326
Questions per clip	8.5 ± 2.0
Candidate answers	4
Tokens per question	6.7 ± 2.1
Tokens per correct answer	1.5 ± 1.1
Tokens per incorrect answer	1.4 ± 0.9

Table 1: Statistics of the LifeQA dataset. Here we report totals and averages along with standard deviation.

Figure 2: Distribution of the LifeQA questions' tokens.

also most ambiguous types of questions. We find that "what" questions most frequently reference "color", "number", and "kind", each of which require visual clues from the video. Not pictured in Figure 2: we find that nouns referring to people, such as "girl", "woman", "man", and "boy" are the first noun in more than 21% of the questions, and we find very few proper names.

We then analyze the type of data required to answer the questions, as shown in Figure 3. To obtain these results, we manually inspect each question and answer to determine whether the question requires the visual (video) or speech (audio or transcription) modalities to answer. We find that 61% of questions need the video to be answered, 29% require the speech or audio information, and 10% need both modalities.

In addition, we analyze the questions based on the expected answer types, as shown in Figure 4. This analysis is inspired by (Tapaswi et al., 2016) and (Lei et al., 2018) as a way of more deeply understanding the type of information needed to answer each question. The graph shows that many questions reference basic visual features, such as count (how many),

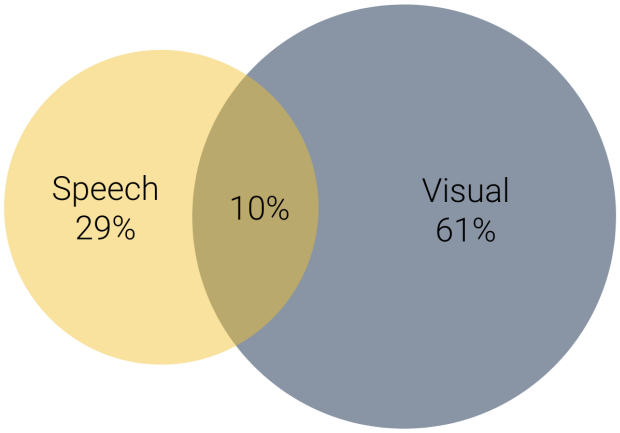

Figure 3: Venn diagram at scale showing the amount of questions by answer type.

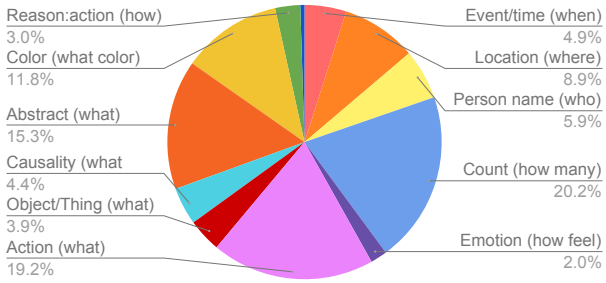

Figure 4: Distribution of the LifeQA questions by type.

color (what color), and location (where) answers. However, there are also many questions that require both language and visual features. For example, abstract ("what") questions ("What is the job of the woman?") can require more than one mode of information to answer.

Dataset Comparison. In Table 2 we compare our dataset with other Video QA datasets. We highlight the presence of multiple modalities and its real-life nature, which differentiates it from prior work. Specifically, LifeQA is the only existing Video QA dataset that focuses on real-life understanding and is carefully constructed from hand-picked in-the-wild videos. In addition, it spans all typical audio and visual modalities, and contains videos that are much longer than those in many other datasets. These qualities lead to a diverse, high-quality video dataset that is suitable for benchmarking current video QA systems and serves as a complement to existing QA datasets. Please refer to Section 2. for more details on the comparison.

More Examples. In Figure 5, we present additional examples of instances in LifeQA. These examples demonstrate the wide variety of scenes and question types present in LifeQA.

4. Experiments

To show the difficulty of the task and explore biases, we implement several models and compare their performance by measuring the question answering accuracy.

Dataset	Task	Source	Answer	Questions	Instances	Avg dur. (s)	Real life?	D	T	A	I	V
DREAM	Reading Comprehension	Exams	MC	10,197	6,444	-		✓	✓			
VisDial	Dialog QA	Images	MC	1,261,510	133,351	-		✓	✓		✓	
LSMDC 16	Video Description	Movies	Text	128,118	128,085	4.1				✓	✓	✓
TGIF-QA	Temporal Reasoning	Tumblr GIFs	MC/Txt	165,165	71,741	≈3.6					✓	✓
MovieQA	Story Understanding	Movies	MC	14,944	6,771	202.7		✓	✓	✓	✓	✓
PororoQA	Story Understanding	Cartoons	MC	8,913	16,066	4.6		✓	✓	✓	✓	✓
MarioQA	Temporal Reasoning	Video games	MC	187,757	187,757	4.5					✓	✓
TVQA	Story Understanding	TV Series	MC	152,545	21,793	76.2		✓	✓	✓	✓	✓
VCQA	Temporal Reasoning	Movies/Web	FB/MC	390,744	109,895	≈30.0					✓	✓
LifeQA	**Real-life Understanding**	**YouTube**	**MC**	**2,326**	**275**	**74.0**	✓	✓	✓	✓	✓	✓

Table 2: Video and Dialog QA datasets comparison. Answer = answer type, h = hours of video, s = seconds per video clip, D = dialog, T = text, A = audio, I = image, V = video, MC = multiple choice, FB = fill in the blanks.

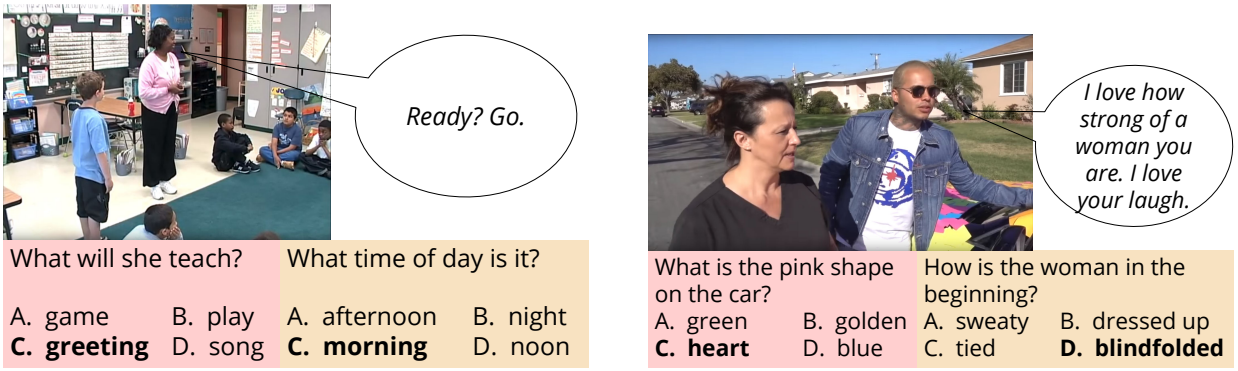

Figure 5: Additional instances from LifeQA. The videos capture a broad range of indoor and outdoor scenes, and the questions refer to both visual and auditory concepts.

4.1. Baselines

We implement and evaluate several baselines, including simple heuristics as well as neural methods. We categorize these baselines according to what inputs they use (the question, the transcriptions, or the visual content) and whether they are trained from scratch or pretrained. By analyzing these baselines, we demonstrate the differences between evaluations on our data versus other non-real-life datasets.

Human baseline. We provide a human baseline, in which two workers were asked to answer a random sample of 101 questions. One of them first listended to the audio in the video without looking at the visual content, then answered the questions, and then repeated the same task by using both modalities (i.e., listen to the audio and watch the video). The other worker did the same but using the visual content — i.e., they first watched the video without listening to the audio, then they answered the questions, and then repeated the same task with both modalities. Note that this differs from the previous analysis in Figure 3 as workers answer the questions here by using one modality at a time and without knowing the correct answer a-priori.

Question-only. We implement several baselines that use only the questions and their candidate answers. Three of these baselines use only the answers, without the question; *Random* chooses one out of the four options uniformly at random, and *Longest answer* and *shortest answer* choose the answer with most or fewest number of tokens, respectively. The first two baselines that also use the question are based on computing some measure of similarity between the question and candidate answers. The first is *Word matching*, as defined by (Yih et al., 2013), which finds the answer with the most overlapping words with the question. The second is *Most similar answer*, which looks at word-level similarity, which we compute by using the average GloVe embedding (Pennington et al., 2014) of the question and each answer, and selecting the answer with the highest cosine similarity with the question. We use GloVe embeddings (Pennington et al., 2014) with size 300 pretrained on 6B tokens from Wikipedia 2014 (Rajpurkar et al., 2016) and Gigaword5 (Parker et al., 2011).

Finally, we implement *ST-VQA-Text*, a variant of Spatio-Temporal VQA (ST-VQA) (Jang et al., 2017) which uses no visual information. It encodes the question with a 2-layer LSTM, then encodes the candidate answers and assigns a score to each one. The text is tokenized and represented using GloVe embeddings (Pennington et al., 2014) of size 300 pretrained on the Common Crawl dataset.

Question + Transcriptions. We present several neural baselines that use the questions, answers, and transcriptions, but omit the videos and audio.

Text-only LSTM and *text-only CNN* both use neural models to separately encode the transcript, question, and answers. The former is a one-layer BiLSTM of hidden size 100. The latter is a 1D CNN with 100 filters of size two tokens and 100 filters of size 3 tokens. We then concatenate the transcript and question encodings, and embed them with a two-layer fully-connected network. We compute the dot product similarity between the question+transcript encoding with each of the candidate answers, and select the one with the highest

score.

Second, we use a variant of BiDAF (Seo et al., 2017) in which we remove the component that predicts the likelihood of each token being the start and end of the span that is needed for SQuAD (Rajpurkar et al., 2016), because in LifeQA there are no such spans. We then compute the dot product between the final hidden state of the Modeling Layer and the representation of each answer choice, which serves as a score. This same process is repeated for both the question and the transcript.

Finally, we use a modified version of the end-to-end Memory Network (MemN2N) proposed by (Tapaswi et al., 2016) based on (Sukhbaatar et al., 2015) to handle multiple-choice question answering. The input to the model are the transcriptions, questions, and candidate answers. The transcription segments are obtained by mean-pooling the GloVe representation of the words for each segment. Our network has an attention layer over the transcriptions to pick the segments that are most relevant to the given question and trained in an end-to-end fashion to select the correct answer.

Question + Vision. We use two variants of ST-VQA (Jang et al., 2017). Both encode the video using a CNN followed by an LSTM, whose final hidden state is then used as in *ST-VQA-Text*. *ST-VQA-Tp.* uses the concatenation of the output of an ImageNet (Deng et al., 2009) pretrained ResNet152 (He et al., 2016) `pool5` layer and of a Sports1M (Karpathy et al., 2014) pretrained C3D (Tran et al., 2015) `fc6` layer as the video encoder. *ST-VQA-Sp.Tp.* computes a spatial attention map to decide what parts of the image are most useful, and uses the `res5c` and `conv5b` of the two CNN encoders. Both use temporal attention maps to pool important information across video frames. We also tried a variant that uses RGB-I3D (Carreira and Zisserman, 2017) (with `avg_pool` and `mixed_5c` layers respectively) instead of C3D, pretrained on ImageNet and Kinetics but do not report it because we obtained similar results.

Question + Transcriptions + Vision. We implement two neural models that use all modalities, TVQA (Lei et al., 2018) and MovieQA (Tapaswi et al., 2016). Both models use object detection networks to identify visual concepts in the corresponding video frames, allowing them to make use of the visual modality. For both we use as visual inputs the output predictions of a Faster R-CNN (Ren et al., 2015) object detection model pretrained on Visual Genome (Krishna et al., 2017).

Pretrained Baselines. Finally, we utilize the TVQA model pretrained on the TVQA dataset. We evaluate it in two versions, with and without fine-tuning on LifeQA.

4.2. Results

In Table 3 we evaluate each model with a five-fold cross-validation, grouping by source video.[1] Similar to (Lei et al., 2018), the baselines trained from scratch do not generally benefit from using the visual information. In fact, most models do not surpass ST-VQA-Text, a baseline which uses only the question and the candidate answers as input. This shows the presence of biases in the dataset, including the multiple-choice setup as opposed to free answer, which allows models to overfit to obtain better-than-random performance. It also demonstrates that leveraging real-life video data is a challenge for existing systems.

Inputs	Model	Accuracy
	Random	25.0
A	Longest answer	30.6
	Shortest answer	21.5
Q+A	Word matching	24.8
	Most similar answer	35.2
	ST-VQA-Text	45.4
T+Q+A	BiDAF	43.3
	Text-only CNN	43.5
	Text-only LSTM	44.0
	Text-only Memory Network	37.9
	Human	63.4
V+Q+A	ST-VQA-Tp.	45.0
	ST-VQA-Sp.Tp.	44.6
	Human	48.5
V+T+Q+A	Multimodal Memory Network	38.2
	TVQA from scratch	41.1
	Pretr. TVQA w/o fine-tuning	51.8
	Pretr. TVQA w/ fine-tuning	51.6
	Human	90.6

Table 3: Baselines on the LifeQA dataset. In the first column, "A" stands for *answer*, "Q" for *question*, "T" for *transcripts* and "V" for *visual modality*. When the transcripts are part of the input, the human performance is measured by using the audio instead.

The TVQA model shows a significant gain in performance when is pretrained on the TVQA dataset, possibly due to the significant larger training size. However, there is still a big gap with respect to human accuracy, providing evidence that this is a challenging benchmark. The same model is able to obtain 66.5% accuracy on the TVQA dataset with five answer choices instead of four. Moreover, the model is not able to perform better even when fine-tuning, showing that the task is still hard when given in-domain training data and giving hints that more robust models should be considered to close the gap as opposed to labeling a larger amount of data to train on.

5. Conclusion

In this work, we introduced LifeQA, a real-life dataset for evaluating Video QA systems on real-life scenarios. Through several analyses and experimental evaluations, we showed that LifeQA presents a challenging task for existing models, with a significant gap in accuracy compared to human performance, thus suggesting that future research is necessary to leverage the multimodal features in this domain. The dataset is publicly available at `https://lit.eecs.umich.edu/lifeqa/`.[2]

[1] Note: we used 221 out of the 275 video clips (50 out of 59 source videos) that were available when running the experiments.

[2] Given the relatively small amount of video data we share and the fact that it is drawn from public sources, the sharing of this data falls under "fair use."

Acknowledgments

We are grateful to Aurelia Bunescu, Daniel D'Souza, Penghao He, Shubham Dash, and Yu-Wei Chao for their help with the collection and annotation of the dataset. This material is based in part upon work supported by the Toyota Research Institute ("TRI"). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of TRI or any other Toyota entity.

References

Agrawal, A., Lu, J., Antol, S., Mitchell, M., Zitnick, C. L., Parikh, D., and Batra, D. (2017). Vqa: Visual question answering. *International Journal of Computer Vision*, 123(1):4–31.

Bajgar, O., Kadlec, R., and Kleindienst, J. (2016). Embracing data abundance: Booktest dataset for reading comprehension. In *ICLR 2017 — Workshop Track*.

Benamara, F. (2004). Cooperative question answering in restricted domains: the WEBCOOP experiment. In *ACL 2004: Question Answering in Restricted Domains*.

Carreira, J. and Zisserman, A. (2017). Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308.

Das, A., Kottur, S., Gupta, K., Singh, A., Yadav, D., Moura, J. M. F., Parikh, D., and Batra, D. (2017). Visual dialog. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July.

Deng, J., Dong, W., Socher, R., Li, L., Kai Li, and Li Fei-Fei. (2009). Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, June.

Fouhey, D. F., Kuo, W.-c., Efros, A. A., and Malik, J. (2018). From lifestyle vlogs to everyday interactions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4991–5000.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Heilbron, F. C., Escorcia, V., Ghanem, B., and Niebles, J. C. (2015). Activitynet: A large-scale video benchmark for human activity understanding. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 961–970.

Hermann, K. M., Kocisky, T., Grefenstette, E., Espeholt, L., Kay, W., Suleyman, M., and Blunsom, P. (2015). Teaching machines to read and comprehend. In *Advances in neural information processing systems*, pages 1693–1701.

Hudson, D. A. and Manning, C. D. (2019). Gqa: a new dataset for compositional question answering over real-world images. *arXiv preprint arXiv:1902.09506*.

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*, pages 2758–2766.

Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., and Fei-Fei, L. (2014). Large-scale video classification with convolutional neural networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June.

Katz, B., Felshin, S., Yuret, D., Ibrahim, A., Lin, J., Marton, G., McFarland, A. J., and Temelkuran, B. (2002). Omnibase: Uniform access to heterogeneous data for question answering. In *International Conference on Application of Natural Language to Information Systems*, pages 230–234. Springer.

Kim, K.-M., Heo, M.-O., Choi, S.-H., and Zhang, B.-T. (2017). Deepstory: Video story qa by deep embedded memory networks. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, IJCAI'17, pages 2016–2022. AAAI Press.

Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.-J., Shamma, D. A., et al. (2017). Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123(1):32–73.

Lei, J., Yu, L., Bansal, M., and Berg, T. L. (2018). TVQA: Localized, compositional video question answering. In *2018 Conference on Empirical Methods in Natural Language Processing*.

Mun, J., Hongsuck Seo, P., Jung, I., and Han, B. (2017). Marioqa: Answering questions by watching gameplay videos. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2867–2875.

Nguyen, T., Rosenberg, M., Song, X., Gao, J., Tiwary, S., Majumder, R., and Deng, L. (2016). Ms marco: A human generated machine reading comprehension dataset. *arXiv preprint arXiv:1611.09268*.

Parker, R., Graff, D., Kong, J., Chen, K., and Maeda, K. (2011). English gigaword fifth edition. *Linguistic Data Consortium*.

Pennington, J., Socher, R., and Manning, C. D. (2014). Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Rajpurkar, P., Zhang, J., Lopyrev, K., and Liang, P. (2016). Squad: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392.

Ren, S., He, K., Girshick, R., and Sun, J. (2015). Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99.

Richardson, M., Burges, C. J., and Renshaw, E. (2013). Mctest: A challenge dataset for the open-domain machine comprehension of text. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 193–203.

Rohrbach, A., Torabi, A., Rohrbach, M., Tandon, N., Pal, C., Larochelle, H., Courville, A., and Schiele, B. (2017). Movie description. *International Journal of Computer Vision*, 123(1):94–120, May.

Seo, M., Kembhavi, A., Farhadi, A., and Hajishirzi, H. (2017). Bidirectional attention flow for machine comprehension. In *ICLR*.

Sigurdsson, G. A., Varol, G., Wang, X., Farhadi, A., Laptev, I., and Gupta, A. (2016). Hollywood in homes: Crowdsourcing data collection for activity understanding. In *European Conference on Computer Vision*, pages 510–526. Springer.

Soricut, R. and Brill, E. (2004). Automatic question answering: Beyond the factoid. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004*.

Sukhbaatar, S., szlam, a., Weston, J., and Fergus, R. (2015). End-to-end memory networks. In C. Cortes, et al., editors, *Advances in Neural Information Processing Systems 28*, pages 2440–2448. Curran Associates, Inc.

Sun, K., Yu, D., Chen, J., Yu, D., Choi, Y., and Cardie, C. (2019). DREAM: A challenge data set and models for dialogue-based reading comprehension. *Transactions of the Association for Computational Linguistics*.

Tapaswi, M., Zhu, Y., Stiefelhagen, R., Torralba, A., Urtasun, R., and Fidler, S. (2016). Movieqa: Understanding stories in movies through question-answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4631–4640.

Tran, D., Bourdev, L., Fergus, R., Torresani, L., and Paluri, M. (2015). Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 4489–4497.

Weston, J., Bordes, A., Chopra, S., Rush, A. M., van Merriënboer, B., Joulin, A., and Mikolov, T. (2015). Towards ai-complete question answering: A set of prerequisite toy tasks. *arXiv preprint arXiv:1502.05698*.

Yih, W.-t., Chang, M.-W., Meek, C., and Pastusiak, A. (2013). Question answering using enhanced lexical semantic models. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1744–1753.

Yu, Z., Xu, D., Yu, J., Yu, T., Zhao, Z., Zhuang, Y., and Tao, D. (2019). Activitynet-qa: A dataset for understanding complex web videos via question answering. *arXiv preprint arXiv:1906.02467*.

Zhu, Y., Groth, O., Bernstein, M., and Fei-Fei, L. (2016). Visual7w: Grounded question answering in images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4995–5004.

Zhu, L., Xu, Z., Yang, Y., and Hauptmann, A. G. (2017). Uncovering the temporal context for video question answering. *International Journal of Computer Vision*, 124(3), Sep.