LLQA - Lifelog Question Answering Dataset

Ly-Duyen Tran[1(✉)], Thanh Cong Ho[2,3], Lan Anh Pham[2,3], Binh Nguyen[2,3], Cathal Gurrin[1], and Liting Zhou[1]

[1] Dublin City University, Dublin, Ireland
ly.tran2@mail.dcu.ie
[2] Vietnam National University,
Ho Chi Minh University of Science, Ho Chi Minh City, Vietnam
[3] AISIA Research Lab, Ho Chi Minh City, Vietnam

Abstract. Recollecting details from lifelog data involves a higher level of granularity and reasoning than a conventional lifelog retrieval task. Investigating the task of Question Answering (QA) in lifelog data could help in human memory recollection, as well as improve traditional lifelog retrieval systems. However, there has not yet been a standardised benchmark dataset for the lifelog-based QA. In order to provide a first dataset and baseline benchmark for QA on lifelog data, we present a novel dataset, *LLQA*, which is an augmented 85-day lifelog collection and includes over 15,000 multiple-choice questions. We also provide different baselines for the evaluation of future works. The results showed that lifelog QA is a challenging task that requires more exploration. The dataset is publicly available at https://github.com/allie-tran/LLQA.

Keywords: Lifelogging · Question answering

1 Introduction

Lifelogging has gained popularity within the research community in recent years with the main focus on lifelog retrieval. The term lifelogging refers to the process of capturing a personal digital diary by technologies such as body cameras and various other wearable sensors. The most extensive published lifelog data, used in the Lifelog Search Challenge workshop 2020 [12], features a collection of first-person images captured throughout the day, as well as the corresponding metadata such as time, GPS coordinates, and biometrics data. Such lifelog data can be processed in lifelog systems, which can serve as a form of 'prosthetic' memory. Lifelogs can support users in memory-related activities such as recollecting, reminiscing, retrieving, reflecting, and remembering intentions, as defined by Sellen and Whittaker's five R's [24]. Out of the five R's, retrieving lifelog data, typically lifelog photos, has been the subject of the majority of lifelog research, as seen in various workshops [11,12,21]. Recollecting details in past lifelog data, on the other hand, involves a higher level of granularity and reasoning; for example, it might involve answering

B. Þór Jónsson et al. (Eds.): MMM 2022, LNCS 13141, pp. 217–228, 2022.
https://doi.org/10.1007/978-3-030-98358-1_18

memory questions such as '*What did I do, where did I go, and who did I see on [Tuesday][afternoon], [July 14, 2018]*?'. Thus, it becomes clear that Question Answering (QA) is an important related topic for research and this paper introduces the first QA dataset for lifelogs.

QA systems are designed to automatically answer questions posed in natural language and are considered to be one of the ultimate goals for retrieval systems [27]. For instance, users may prefer getting concise answers to specific questions instead of browsing an entire document. The same argument could be made for other types of media such as photos and videos; Visual QA systems can save the user from extraneous effort by automatically inferring a user's question regarding an image/video and producing a short and accurate answer. To produce the correct answer, the model needs to be able to interpret the question and focus on the relevant part of the image/video. Due to advances in the field of computer vision, visual QA has been a fast-growing area with various techniques for images [1,8,14] and videos [7,15,17]. Applying such visual QA techniques to lifelogs suggests that lifelog QA can be a valuable and impactful research area, since lifelog data is heavily visual-based. Having the ability to understand the whole context of a real-world event, Lifelog QA systems ultimately could provide help in human memory recollection, as well as improve traditional lifelog retrieval systems.

Despite the similarities to visual QA, the data used in Lifelog QA has several distinct aspects that render the direct application of Visual QA techniques less effective. Image QA techniques do not exploit the temporal nature of lifelog data. In the case of Video QA, standard action recognition techniques such as C3D [15] may not be useful as lifelog data are discontinuous (with an average frequency of 1 snapshot every 30 s) in the current generation of lifelog datasets. Moreover, current state-of-the-art video QA methods learn inference by relying on the appearance and motion data from a third-person point of view, which is different from the first-person photos in lifelog data. The most related work to Lifelog QA is EgoVQA [7], an egocentric video question answering dataset containing first-person perspective videos similarly to lifelog photos. However, videos still hold different characteristics compared to lifelog photos. For this reason, a novel benchmark dataset for Lifelog QA is a prerequisite to evaluate a model's ability to 'recollect' details in lifelog data.

In the field of lifelog QA, the novel dataset proposed in this paper supports the following research contributions:

1. Describing a new semi-automatic process of constructing a Lifelog QA dataset, based on an existing lifelog collection;
2. Providing 15,065 lifelog QA pairs, comprising of both multiple-choice questions and yes/no questions;
3. Presenting results of a pilot experiment to identify the gap between the human gold standard and existing QA models.

2 Related Work

2.1 Lifelogs and Personal Data Analytics

The inspiration for lifelogging dates back to Vannevar Bush's 1945 article *As We May Think* [3], which describes a blueprint personal information system which he called *Memex*. Bush considered *Memex* as 'a device in which an individual stores all his books, records, and communications, which is mechanised to be consulted with exceeding speed and flexibility. It is an enlarged intimate supplement to his memory'. However, it was not until a research project of Microsoft Research, called MyLifeBits [10] was started by Gordon Bell in 2001, that lifelogging began to gain attention from the research community. The MyLifeBits system attempted to capture every possible aspect of the daily life of Bell, including every web page visited, all Instant Message (IM) chat sessions, all telephone conversations, meetings, radio, television programs, as well as all mouse and keyboard activities and media files in his personal computers. All digitised data were stored in a SQL database to support a simple interface for different functionalities such as organising, associating metadata, assessing, and reporting information. Since then, due to advances in sensor technology and the availability of low-cost data storage, lifelogging has become an achievable activity for many. However the primarily passive nature of lifelogging means that the amount of data generated can be massive (over 1 TB of multimodal data per individual per year), and therefore effectively searching through such extensive archives of lifelog data remains an important yet challenging task.

Different lifelog benchmarking workshops/challenges have been established with distinctive evaluation metrics to assess lifelog systems, with the common objective being to facilitate the effective retrieval of specific lifelog images in an interactive or automatic manner. The standard approach taken by existing lifelog retrieval systems, such as MyScéal [26] and LifeSeeker [20], is assigning semantic context, e.g., visual concepts, to lifelog photos and applying traditional information retrieval techniques to produce a ranked list of relevant images. This approach treats each lifelog photo individually, which does not exploit the temporal and continuous nature of lifelog data. This is important because an individual snapshot of lifelog data is likely to not fully convey the whole context of an event [13].

There have been a number of lifelog datasets released since 2015 and at the most recent lifelog retrieval workshop, LSC'20 [12], the organisers published a large collection including six months of anonymised lifelog data, consisting of 50 GB of fully redacted wearable camera images at 1024×768 resolution, captured using OMG Autographer and Narrative Clip devices. These images were collected during periods in 2015, 2016, and 2018; some private information (for example, faces and identifiable materials) appearing in these images are anonymised in a manual or semi-manual process. The metadata for the collection consists of textual metadata representing time, physical activities, biometrics, locations, as well as visual concepts extracted from the non-redacted version of the visual dataset using a CAFFE CNN-based object detector [16]. This dataset forms the basis of the dataset augmented and released in this paper.

2.2 Video Question Answering (Video QA)

Video QA, an application of QA, is a task requiring the generation of correct answers to given questions related to a video or video archive. The questions are either in the form of fill-in-the-blank, multiple-choice, or open-ended types.

All existing Video QA datasets except for EgoVQA [7] are from third-person perspective. TGIF-QA [15] is a dataset of over 165,000 questions on 71,741 animated pictures. Multiple tasks are formulated upon this dataset, including counting the repetitions of the queried action, detecting the transitions of two actions, and image-based QA. MSVD-QA and MSRVTT-QA [28] are two datasets with third-person videos. The Video QA tasks formulated in both of these two datasets are open-ended questions of the types what, who, how, when, and where, and their answer sets are of size 1000. YouTube2Text-QA [29] is a dataset with both open-ended and multiple-choice tasks of three major question types (what, who, and other). TVQA and TVQA+ [17,18] are built on 21,793 video clips of 6 popular TV shows with 152.5K human-written QA pairs. EgoVQA [7] was proposed due to the lack of first-person point-of-view videos in these datasets; however, the size of the dataset is small, with just over 600 QA pairs.

After a comprehensive review of research on video QA, we observe that there are three unique characteristics of Lifelog QA compared with Video QA: (1) lifelog QA deals with more channels of information because of the inherent multimodality of lifelog data; (2) the collected activities in lifelog are captured in snapshots instead of being continuous, rendering the motion features ineffective; (3) unlike most video QA datasets, the point of view in lifelog visual data is first-person instead of third-person. Therefore, it is clear that the existing approaches and datasets for visual QA are not representative of the challenge posed by lifelog QA, hence it becomes necessary to investigate Lifelog QA in more detail, which is the primary motivation for this research.

3 LLQA - A Lifelog Question Answering Dataset

We define **Lifelog Question Answering (Lifelog QA)** as a task to produce correct answers to given textual representations of an individual's information needs concerning a past moment or experience from a lifelogger's daily life. In the scope of this initial research, we will consider only multiple-choice questions and yes/no questions due to the straightforward means of evaluation. It is anticipated that other types of answers will be explored at a later point.

In this section, a detailed explanation about how to build the first lifelog QA dataset is covered. This process is part of our contribution to the field of lifelog QA. To save time and effort, we applied automated steps where possible. The pipeline of the entire process is summarised in Fig. 1 and the description of each component is as follows:

3.1 Data Collection

The lifelog QA dataset for this work is based on the LSC'20 collection [12] mentioned in Sect. 2.1. Specifically, 26 d of data were selected from the year 2015

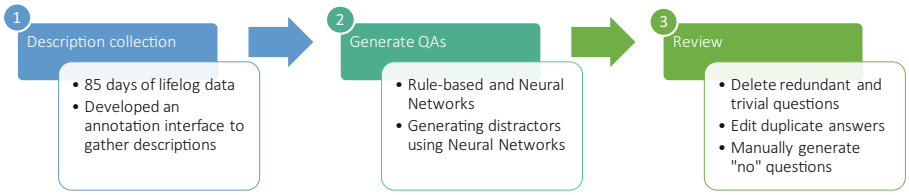

Fig. 1. The process of dataset construction.

and 59 d were selected in 2016. Each day is segmented into short events of the date based on the locations and activities of the lifelogger, which is based on the event segmentation approach of Doherty and Smeaton [6]. This encourages the annotators to focus on individual events. From the provided metadata throughout the day, whenever the location (work, home, etc.) or the activity (walking, driving, etc.) is changed, a new segment will be created. The process results in a total of 2,412 segments.

An annotation system was developed that presents annotators with all images in each segment along with the metadata such as time, GPS location, and the relative position of the segment in the whole day.

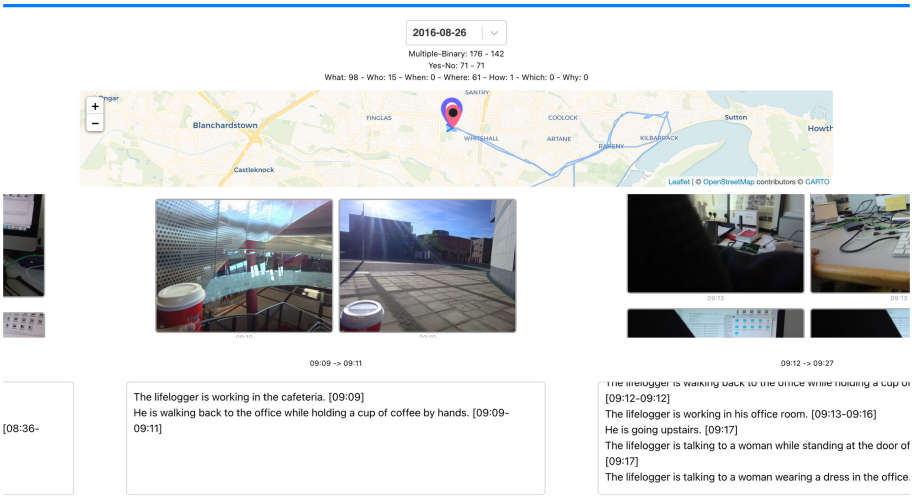

Fig. 2. Annotation Interface.

Annotators, who are volunteers from undergraduate Computer Science programmes, were asked to describe the events happening in each segment as seen in Fig. 2. Every description is annotated along with its starting and ending times.

The description should include actions or activities; objects that the lifelogger interacted with along with their properties such as size, shape, or colour; the location where the lifelogger was in, heading towards to or away from; and people (with a general identity description to preserve privacy). One example could be 'The lifelogger is reading a book in a cafe with a person in a black t-shirt.'

3.2 Generation of Question and Answers

The descriptions were converted to a list of questions by an automatic system which is summarised in Fig. 3. Entity extraction and syntax transformation were completed using hand-crafted rules based on POS tags and semantic role labels. To generate question words (who, what, where, etc.), a Seq2Seq neural network was trained on the questions and answers in the CoQA [23] dataset. False answers (distractors), are generated using RACE [9] with the gathered knowledge from ConceptNet [25] facts as context.

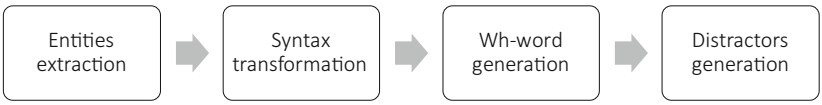

Fig. 3. The procedure of question-answer generation.

Given the description 'The lifelogger was reading a book in a cafe.', the generation process would be as follows:

1. Entities extraction
   The lifelogger, *reading a book*, and *in a cafe* are examples of entities in the sentence. We will choose *reading a book* in this example to illustrate further. Thus, the correct answer to this generated question-answer pair would be *reading a book*;
2. Syntax transformation - yes/no
   By moving *was* to the beginning of the sentence, we get 'Was the lifelogger reading a book in the cafe?' - 'Yes' as a yes/no question-answer pair;
3. Syntax transformation - multiple
   First, based on the POS tags, an automated process decides the entity is a *phrasal verb*, thus by replacing it with *doing* in the sentence and by applying a rule-based syntax transformation, we get '[...] was the lifelogger doing in the cafe?'
4. Wh-word generation
   Since questions in this dataset start with a Wh word, a pretrained S2S model chooses appropriate question word for this question. In this case, a sensible one would be *What*.

5. Distractor generations
   So far, we get the question-answer pair as 'What was the lifelogger doing in the cafe?' - 'Reading a book'. To make this a multiple-choice question, we use RACE [9], a distractor generator for reading comprehension questions, and get the other wrong answers as 'Using his phone', 'Drinking coffee', and 'Playing football'.

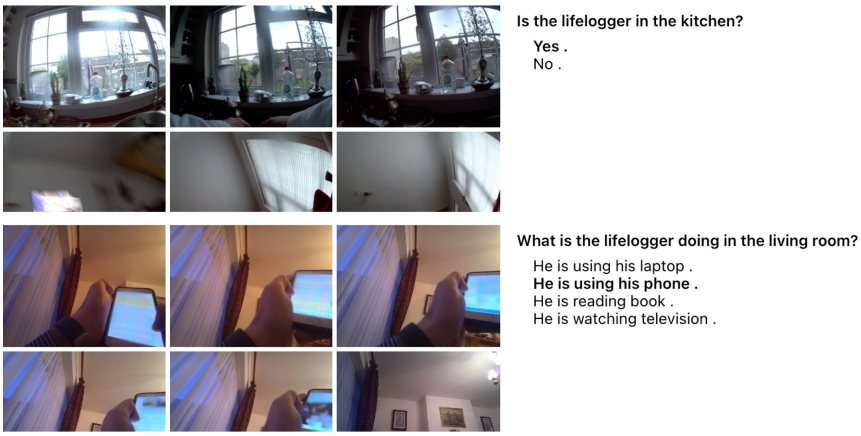

Fig. 4. Two example question-answer pairs in the dataset. The dataset contains both yes/no questions and multiple-choice questions.

3.3 Review

The generated questions and answers are reviewed by the annotators to correct semantics and delete duplicates, as well as ensuring constraints such as:

1. There are no duplicate answers for the same question,
2. The ratios between yes and no questions are balanced. As the automatic syntax transformation could only generate positive yes/no questions, the annotators are asked to create negative ones manually.

The dataset contains 15,065 QA pairs in total. Examples of the QA pairs can be seen in Fig. 4. On average, our questions contain 7.66 words. Correct answers tend to contain 3.57 words compared to 4.34 words in the generated wrong answers. Figure 5 and Table 1 present the breakdown of questions generated. The dataset is split into two sets: training and testing sets consisting of 10,668 (70.81%) and 4,397(29.19%) question-answer pairs, respectively. The splitting was done in a manner that ensures there are no overlapping days between the subsets, or in other words, the lifelog data in the testing set are unseen.

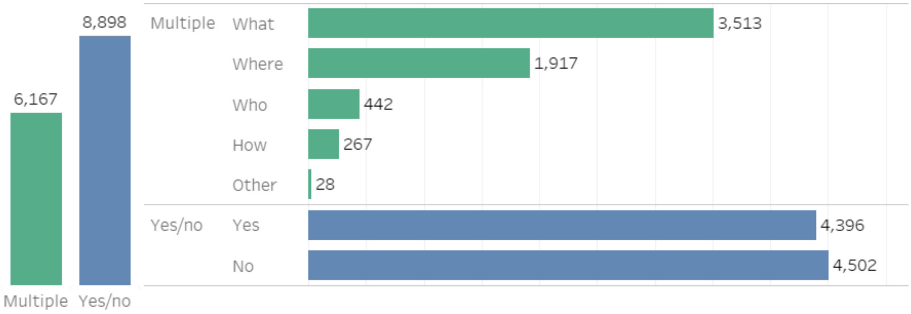

Fig. 5. Numbers of each question type in Lifelog QA dataset.

Table 1. Numbers of questions in each month in LSC'20 lifelog data collection.

Month	#Days	Days	#Images	#Questions
Feb, 2015	06	Feb 24–28	8549	941
Mar, 2015	20	Mar 01–20	28563	2745
Aug, 2016	24	Aug 08–31	32026	4871
Sep, 2016	30	Sep 01–30	51195	5595
Oct, 2016	05	Oct 01–05	7375	913
Total	**85**	–	**127708**	**15065**

4 Pilot Experiment

In order to evaluate the dataset and provide accompanying baselines for subsequent comparison, a pilot experiment has been carried out on several baselines, which are described below.

4.1 Human Gold-Standard Baseline

To determine the targeted performance (in terms of accuracy) on our dataset, we performed a user study, asking different groups of 10 volunteer students to complete the question-answering task. Each volunteer was asked to answer 20 yes/no questions and 20 multiple-choice questions chosen randomly from the testing set. Each question was accompanied by the relevant images. To avoid bias, there was no overlap between the annotators that have worked on the questions and the students participating in this study. The gold standard accuracy was found to be 0.8417 for yes/no questions and 0.8625 for multiple-choice questions. The reason that the scores are less than 1.0 is because the volunteers were presented with the relevant section for the question, rather than the lifelog data for the whole day, so in some cases, they did not fully understand the context of the event mentioned in the question. Another interesting feedback from the participants, as well as the annotators, concerns the volume of lifelog data causing issues in

understanding. This is a common problem in lifelog analytics when the decisions regarding lifelog data are often made by a third party and not the original data gathering lifelogger, for example, as seen in the studies carried out by Byrne et al. [4].

4.2 Question-Only

We implement several heuristic baselines that use only the questions and their candidate answers in a similar approach to Castro et al. [5]. Specifically, *Longest answer* and *Shortest answer* choose one out of the four options with the most or the fewest number of tokens, respectively. *Word matching* chooses the answer based on the number of tokens they have in common with the question. Because yes/no answers have no difference either in length or the number of common words with the questions, we omit these models for this experiment.

Moreover, we implement *Sequence-to-sequence (S2S)* model based on the architecture of UniLMv2 [2], the state-of-the-art model in natural language understanding and generation tasks. We trained S2S on the CoQA [23] question-answer pairs. It encodes the question with a 2-layer LSTM, then encodes the candidate answers and assigns a score to each one. The text is tokenised and represented using Glove 300-D embeddings [22].

4.3 Question and Vision

Because of the similarity to Video QA task, we implemented *TVQA*, the original TVQA [17] model, trained on TVQA dataset. This is the state-of-the-art system in Video QA. To evaluate the application to lifelog data, we consider each day to be a one fps video with each image (along with the attached metadata) as one single frame in that video. We converted the annotated starting and ending times into the ordinal index of the frames in the video. Moreover, we replaced the subtitles intended for videos with a concatenation of metadata associated with the frames. While it may seem strange to treat visual lifelog data as motion video, it is temporal in nature and many of the participants in the LSC challenge [12] have modified existing Video Search systems from the VBS challenge [19] to treat lifelog data as 1 fps video.

4.4 Results

Both S2S and TVQA models have been retrained on the training set of the lifelog QA dataset and achieved a small improvement in accuracy compared to the untrained versions, as seen in Table 2. Furthermore, there is no considerable difference between the question-only models. Although the average length of the correct answers are shorter than the wrong ones, *Shortest answer* did not perform well at the lowest accuracy of 0.1717 for multiple-choice questions. Amongst the models, the retrained TVQA achieved the best performance with the accuracy of 0.6338 and 0.6136 for yes/no questions and multiple-choice questions, respectively. However, humans still significantly outperformed the models.

The results highlighted that the existing approaches are still far from the human gold standard for the lifelog QA task, so they should be optimised to improve performance. This will be a potential and promising topic for future research in lifelog domain in general, and especially in lifelog QA.

Table 2. Accuracy of different models in the pilot experiment.

Model	Yes/no	Multiple-choice
Longest answer	–	0.3202
Shortest answer	–	0.1717
Word matching	–	0.3041
S2S	0.5206	0.3148
S2S (retrained)	0.5066	0.3626
TVQA	0.4956	0.4085
TVQA (retrained)	**0.6338**	**0.6136**
Human baseline	0.8417	0.8625

5 Conclusion

In this work, we introduced Lifelog QA, a question answering dataset for lifelog data. The dataset consists of over 15,000 yes/no questions and multiple-choice questions. Through several baseline experiments, we assessed the suitability of the dataset for the task of lifelog QA. We note that there is still a significant gap between the proposed baselines and human performance on the QA accuracy, meaning that there is a significant research challenge to be addressed. Our findings suggest that a large proportion of the dataset involves the lifelogger's actions or interactions with other objects, therefore it is crucial to improve the standard action recognition mechanism. One possible approach is to sample video frames with a lower rate similarly to lifelog data and develop models based on this. Furthermore, we could develop respective sequences of features for other metadata instead of using the existing textual subtitle stream as in the TVQA model. Additionally, temporal reasoning is also essential to this task, especially for questions containing *before* or *after* actions. These three points can be integrated in future works to improve the semantic understanding of lifelog data.

The dataset is published at https://github.com/allie-tran/LLQA. We also include the annotated description with timestamps, which can be used to develop models for lifelog captioning tasks. By creating this dataset, we hope it can encourage more researchers to participate in and explore this research area further.

Acknowledgements. This work was conducted with the financial support of the Science Foundation Ireland under grant agreement 13/RC/2106_P2 and the Centre for Research Training in Digitally-Enhanced Reality (d-real) under Grant No.

18/CRT/6224. For the purpose of Open Access, the author has applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission.

References

1. Anderson, P., et al.: Bottom-up and top-down attention for image captioning and visual question answering, pp. 6077–6086 (2018)
2. Bao, H., et al.: Unilmv2: pseudo-masked language models for unified language model pre-training. In: International Conference on Machine Learning, pp. 642–652. PMLR (2020)
3. Bush, V., et al.: As we may think. The atlantic monthly **176**(1), 101–108 (1945)
4. Byrne, D., Kelliher, A., Jones, G.J.: Life editing: third-party perspectives on lifelog content. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 1501–1510 (2011)
5. Castro, S., Azab, M., Stroud, J., Noujaim, C., Wang, R., Deng, J., Mihalcea, R.: Lifeqa: a real-life dataset for video question answering. In: Proceedings of the 12th Language Resources and Evaluation Conference, pp. 4352–4358 (2020)
6. Doherty, A., Smeaton, A.: Automatically segmenting LifeLog data into events
7. Fan, C.: EgoVQA - an egocentric video question answering benchmark dataset. In: 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW), pp. 4359–4366 (Oct 2019), iSSN: 2473–9944
8. Fukui, A., Park, D.H., Yang, D., Rohrbach, A., Darrell, T., Rohrbach, M.: Multimodal Compact Bilinear Pooling for Visual Question Answering and Visual Grounding. arXiv:1606.01847 [cs], September 2016
9. Gao, Y., Bing, L., Li, P., King, I., Lyu, M.R.: Generating distractors for reading comprehension questions from real examinations. In: AAAI-19 AAAI Conference on Artificial Intelligence (2019)
10. Gemmell, J., Bell, C., Lueder, R.: Mylifebits: a personal database for everything. Commun. ACM **49**, 89–95 (2006)
11. Gurrin, C., et al.: Overview of the NTCIR-14 lifelog-3 task. In: Proceedings of the 14th NTCIR Conference, p. 13. NII (2019)
12. Gurrin, C., et al.: Introduction to the third annual lifelog search challenge (LSC'20). In: Proceedings of the 2020 International Conference on Multimedia Retrieval, ICMR 2020, pp. 584–585. Association for Computing Machinery
13. Gurrin, C., Smeaton, A.F., Doherty, A.R., et al.: Lifelogging: personal big data. Found. Trends Inform. Retrieval **8**(1), 1–125 (2014)
14. Hu, R., Andreas, J., Rohrbach, M., Darrell, T., Saenko, K.: Learning to reason: end-to-end module networks for visual question answering. arXiv:1704.05526 [cs], Septrmber 2017. arXiv: 1704.05526 version: 3
15. Jang, Y., Song, Y., Yu, Y., Kim, Y., Kim, G.: TGIF-QA: toward spatio-temporal reasoning in visual question answering
16. Jia, Y., et al.: Caffe: convolutional architecture for fast feature embedding
17. Lei, J., Yu, L., Bansal, M., Berg, T.L.: TVQA: localized, compositional video question answering. arXiv:1809.01696 [cs] (May 2019), arXiv: 1809.01696
18. Lei, J., Yu, L., Berg, T.L., Bansal, M.: TVQA+: spatio-temporal grounding for video question answering. arXiv:1904.11574 [cs], May 2020. arXiv: 1904.11574
19. Lokoč, J., et al.: Is the reign of interactive search eternal? findings from the video browser showdown 2020. ACM Trans. Multimedia Comput. Commun. Appl. **17**(3), July 2021

20. Nguyen, T.N., et al.: Lifeseeker 3.0: An interactive lifelog search engine for lsc'21. In: Proceedings of the 4th Annual on Lifelog Search Challenge, pp. 41–46 (2021)
21. Ninh, V.T., Le, T.K., Zhou, L., Piras, L., Riegler, M.: Overview of ImageCLEFlifelog 2020: Lifelog moment retrieval and sport performance lifelog. In: CLEF (Working Notes), p. 17 (2020)
22. Pennington, J., Socher, R., Manning, C.D.: GloVe: global vectors for word representation. In: Empirical Methods in Natural Language Processing (EMNLP), pp. 1532–1543
23. Reddy, S., Chen, D., Manning, C.D.: CoQA: a conversational question answering challenge. Trans. Assoc. Comput. Linguist. **7**, 249–266 (2019)
24. Sellen, A.J., Whittaker, S.: Beyond total capture: a constructive critique of lifelogging 53(5), 70–77
25. Speer, R., Chin, J., Havasi, C.: Conceptnet 5.5: an open multilingual graph of general knowledge. In: Thirty-First AAAI Conference on Artificial Intelligence (2017)
26. Tran, L.D., Nguyen, M.D., Thanh Binh, N., Lee, H., Gurrin, C.: Myscéal 2.0: a revised experimental interactive lifelog retrieval system for lsc'21. In: Proceedings of the 4th Annual on Lifelog Search Challenge, pp. 11–16 (2021)
27. Trotman, A., Geva, S., Kamps, J.: Report on the sigir 2007 workshop on focused retrieval. In: ACM SIGIR Forum, vol. 41, pp. 97–103. ACM, New York (2007)
28. Xu, D., et al.: Video question answering via gradually refined attention over appearance and motion. In: Proceedings of the 25th ACM International Conference on Multimedia, MM 2017, pp. 1645–1653. Association for Computing Machinery, event-place: Mountain View, California, USA
29. Ye, Y., Zhao, Z., Li, Y., Chen, L., Xiao, J., Zhuang, Y.: Video question answering via attribute-augmented attention network learning. In: Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 829–832 (2017)