Chapter 1

Related Work

This chapter discusses a number of works related to this project. Firstly, a selection of VideoQA datasets are presented in Section 1.1. Then we discuss some of the existing VideoQA implementations in Section 1.2. Since most existing work done in VideoQA involves end-to-end neural networks, Section 1.3 outlines a number of 'hybrid' approaches used in VQA. Many of the authors of the VideoQA datasets discussed in Section 1.1 also present original VideoQA models. These models are therefore discussed in Section 1.2.

1.1 Datasets for VideoQA

A number of datasets are available for the VideoQA problem, in this section we discuss each of the available datasets. A comparison of all the datasets discussed in this section is shown in Table 1.1.

Firstly, the CLEVRER [28] dataset contains 20,000 synthetic videos generated in a controlled environment. The CLEVRER dataset is similar to the CLEVR [10] dataset for VQA, but extends the task to videos and adds a number of question types which are specific to videos. Each video includes a number of objects, where an object can be identified by a combination of three properties: shape, material and colour. Each video also contains a number of events, of which there are three types: enter, exit and collision. In total, CLEVRER contains 305,280 questions, of four different types:

- **Descriptive.** Require the model to reason about a video's content.
- **Explanatory.** Ask about whether an object is responsible for an event.
- **Predictive.** Evaluate a model's ability to predict future events (events that may occur after the video ends).
- **Counterfactual.** Ask about the outcome of a video under hypothetical conditions.

Figure 1.1 shows frames from an example video, along with a number of QA pairs and their associated types.

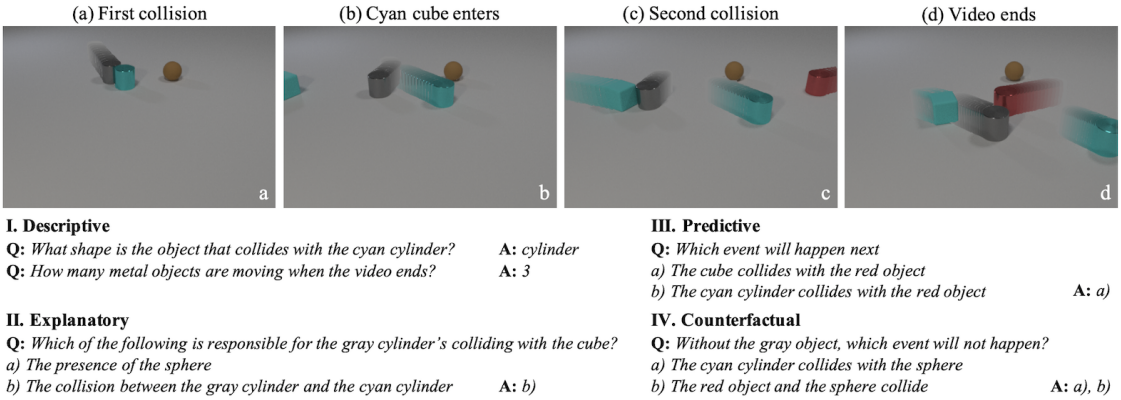

Figure 1.1: An example video and a number of QA pairs from the CLEVRER dataset. Image from [28].

The MovieQA dataset [21] is a VideoQA dataset consisting of 14,944 multiple-choice questions about parts of movies. The clips come from a collection of 408 movies and the Question-Answer (QA) pairs were generated by humans. The questions and each of the possible answers are written in natural language, and there are five possible answers for each question.

Zeng et al. [30] create a much larger VideoQA dataset by automatically generating QA pairs from videos and their associated descriptions collected online. Their dataset consists of 18,100 videos as well 151,263 and 21,352 automatically generated QA pairs in the training and validation sets, respectively. The dataset also contains 2,461 human-generated QA pairs to be used for testing. Their questions and answers are free-form natural language, however, a large number of their answers are yes and no (32.5% and 32.5%, respectively).

The TGIF-QA dataset [9] is commonly used for assessing the performance of VideoQA models. The dataset contains 165,165 human-generated QA pairs collected from 71,741 GIFs, sourced from the TGIF dataset [13], which contains a number of GIFs and associated descriptions. There are four possible types of questions in the TGIF-QA dataset, three of which are specific to VideoQA; requiring temporal knowledge to answer. The question types are as follows:

- **Repetition Count**. Counting the number of repetitions of an action. There are 11 possible answers (0, ..., 9, 10+).
- **Repeating Action**. A multiple-choice question about identifying an action that has been repeated in the video.
- **State Transition**. A multiple-choice question about identifying the state before or after another state.
- **FrameQA**. Open-ended questions related to a single frame.

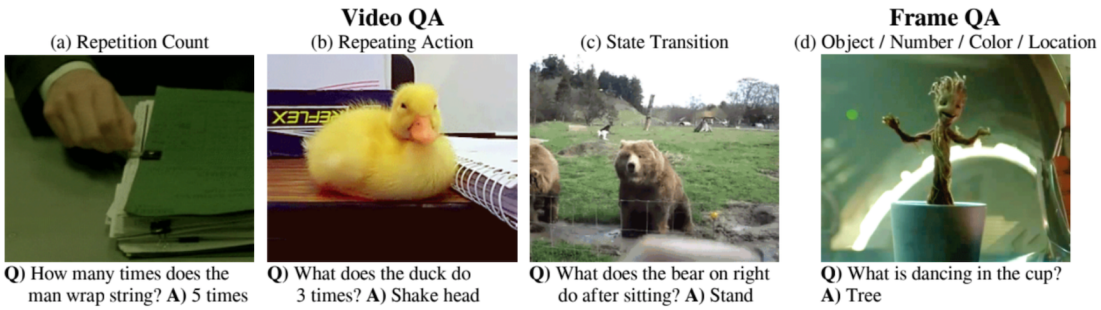

Figure 1.2: Example videos and QA pairs included in the TGIF-QA dataset, split by question type. Figure from [9].

For the VideoQA questions the authors created templates for questions and used a large number of human annotators to speed up the generation process. The FrameQA questions are generated using the descriptions from the TGIF dataset. A number of quality control checks were also included. A few example questions and videos from the TGIF-QA dataset are shown in Figure **??**.

Zhu et al. [31] have proposed a VideoQA dataset containing fill-in-the-blank (FIB) style questions, with multiple-choice answers. The dataset contains over 100,000 real-world video clips and 400,000 questions. The dataset is generated from three different annotated video sources. On top of questions which ask the model to describe the present (describe the current video), for two of three video sources the authors also introduce two additional question types: infer the past and predict the future. For these two types of questions the model is asked a question on a part of the video which it is not explicitly given; these questions require the model to use some form of commonsense reasoning to generate a correct answer. One of the advantages of using a multiple-choice dataset, such as this, is that it is more amenable to quantitative evaluation than datasets with long free-form answers, since answers are either right or wrong.

The EgoVQA dataset [4] attempts to address the lack of first-person VideoQA datasets. The dataset contains 581 QA pairs with both multiple-choice questions (with 5 possible answers per question) and open-ended questions. The dataset was created by manually generating QA pairs from a pre-existing set of 16 first-person videos. The authors also show that existing VideoQA models only marginally outperformed random choice on some types of questions in their dataset. They conjecture that existing models struggle to separate attentions on camera wearers from attentions on third persons.

Xu et al. [25] generate two VideoQA datasets by converting video captions into QA pairs. The first dataset, known as MSVD-QA, is generated from the Microsoft Research Video Description Corpus [3] which is used in many video captioning experiments. MSVD-QA contains 1,970 video clips and 50,505 QA pairs. Similarly, the second dataset, known as MSRVTT-QA, is generated from the MSR-VTT dataset [26]. The MSRVTT-QA dataset contains 10,000 video clips and 243,680 QA pairs.

The YouTube2Text-QA dataset [27] is another large dataset for VideoQA generated from a pre-existing video description dataset, in this case the YouTube2Text [7] dataset is used. The YouTube2Text-QA dataset consists of 1,970 videos and 99,421 QA pairs.

The TVQA dataset [12] contains 21,793 video clips and 152,545 QA pairs based on 6 popular TV shows. The QA pairs were annotated manually using *Amazon Mechanical Turk*. Workers were asked to generate questions in the format: [What/How/Where/Why/Who/Other] _____ [when/before/after] _____. The second part of the question localises the relevant video moment within the clip, while the first part contains the question about that moment. The answers to the questions are given in multiple-choice format, with five candidate answers for each question.

The PororoQA dataset [11] is created from video clips and subtitles of the children's cartoon series, *Pororo*. The dataset contains 8,913 multiple-choice QA pairs and 16,066 video clips.

Finally, the MovieFIB dataset [14] is a large-scale fill-in-the-blank style dataset generated from movie descriptions. The dataset contains 128,085 video clips and 348,998 QA pairs. The questions concern entities, actions and objects; answering these questions therefore implies that a model has some level of visual understanding of the scene, rather than being able to answer based purely on the given partial sentence. Answers are open-ended (not multiple choice) but each answer is only a single word.

Table 1.1: Comparison of discussed VideoQA datasets. Each row contains data on: the number of videos/clips, the number of QA pairs, whether the uses multiple-choice questions, whether the dataset uses fill-in-the-blank questions and the video source. Where some questions in a dataset are multiple-choice and others are not, the table shows which is true for the largest number of questions, but includes a footnote for other types of questions.

Dataset	#Videos	#QA pairs	MC	FIB	Source
CLEVRER	20,000	305,280	N[1]	N	Generated
MovieQA	408[2]	14,944	Y	N	Movies
Zeng et al.	18,100	175,076	N	N	Online videos
TGIF-QA	71,741	165,165	Y[3]	N	Online videos
Zhu et al.	>100,000	400,000	Y	Y	Various
EgoVQA	16	581	Y	N	First-person videos
MSVD-QA	1,970	50,505	N	N	Video desc. corpus
MSRVTT-QA	10,000	243,680	N	N	Video desc. corpus
Youtube2Text-QA	1,970	99,421	Y	N	YouTube videos
TVQA	21,793	152,545	Y	N	TV shows
PororoQA	16,066	8,913	Y	N	Cartoon series
MovieFIB	128,085	348,998	N	Y	Movie description

[1] Non-desciptive questions (85,362 questions) are MC.
[2] Full length movies. Some of the QAs come with timestamps, allowing more video clips.
[3] FrameQA questions (53,083 QA pairs) are not MC.

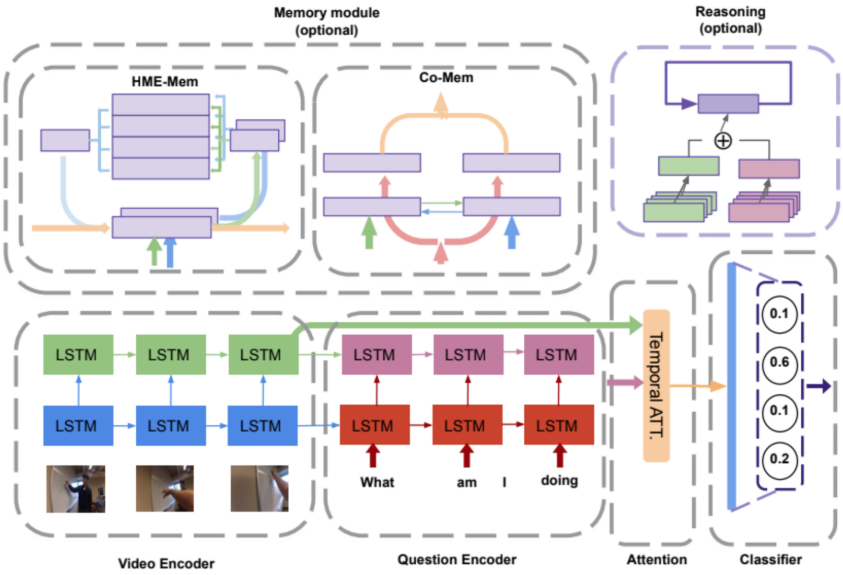

Figure 1.3: Typical VideoQA architecture. Figure from [4].

1.2 VideoQA Implementations

All of the datasets described in the section above (except [4]) were presented along with original neural network models for solving the VideoQA task. This section attempts to summarise some of these approaches, along with a few others [6, 29, 5, 18]. However, since the focus of our approach will not be an end-to-end neural architecture, we do not give detailed descriptions. An example of a typical VideoQA neural architecture is given in Figure 1.3.

All models from work previously presented here contain a video encoder for extracting features from frames of the video. These usually include both appearance and motion features, which are extracted from pre-trained networks (ResNet [8], VGG [17] or GoogLeNet [20], for appearance, and C3D [22], for motion). The features extracted from each frame are then usually passed into LSTM or GRU networks to obtain encodings for the whole video.

Questions are often encoded by generating word embeddings for each word of the sentences and then passing the list of words to a sentence encoder, such as the LSTM or GRU architectures.

Visual attention [16] is used to help neural networks focus on the most relevent areas of an image or video. Applying attention mechanisms to associate a question with its most relevant frames (or areas within frames) has become a key part of more recent VideoQA models [9, 25, 27, 12, 6, 29, 5, 11]. Temporal attention is commonly applied to help the model focus on the most salient frames of the video, however, [9] applies both temporal and spatial attention, which also allows the model to attend to the most relevant regions of a frame.

1.3 External Knowledge for VQA

While VideoQA research has focused on end-to-end neural network architectures, some recent research in VQA (single frame setting) has experimented with using explicit reasoning layers and integrating external knowledge. A number of VQA datasets which require some form of external 'commonsense' reasoning have been proposed recently [15, 24, 23]. It has been shown that end-to-end neural networks which do not attempt to make use of knowledge which is external to the training data perform poorly on some of these datasets [15]. This section discusses some of the attempts that have been made to integrate external knowledge into VQA systems.

The authors of the three datasets outlined above propose models which make use of external knowledge, ususally stored in some structured knowledge base (KB). Examples of KBs include DBpedia [2], which stores structured information extracted from Wikipedia, and ConceptNet [19], which contains automatically generated 'commonsense' relations between objects.

As opposed to using a structured KB, the authors of [15] provide a neural network model which is trained to find the answer to an image-question pair from Wikipedia articles. They also propose a number of methods for combining this network with state-of-the-art VQA models and show that this provides an improvement in performance on their dataset.

The authors of [1] propose a model which first extracts properties from an image using a pre-trained neural network and represents these properties explicitly using logic. They also extract relations between nouns, adjectives and the question word from the question and represent these in logic. Finally, they reason over the extracted relations using a probabilistic reasoning engine to find the most likely answer. This method of reasoning not only allows the model to make use of external knowledge, but also helps improve the transparency and explainability of the model.

Chapter 2

Conclusion

TODO discuss:

1. trade-offs (providing a env spec and requiring that everything be discrete) and advantages (learning the model of an envionment explicitly to allow explanations, error correction) of using hybrid models generally (as well our specific models). (See listed trade-offs in H-PERL chapter, advantages may come from observations made in the evaluation section)

2. Future work (adding NLP to the model, other types of questions that can be added to give further advantages of using hybrid, neural models (graph networks) for event detection, extending the ILP to be faster, more accurate and allow larger search spaces, multiple moving objects in videos, non-determinism in videos, removing the need for question type, dymanically updating the attention (object detection) of the model based on the question, learning the QA system rules).

3. How this model can be applied to related tasks (eg. VQA, captioning, maybe RL).

4. Compare to objecttives to see if completed

Bibliography

[1] Somak Aditya, Yezhou Yang, and Chitta Baral. "Explicit reasoning over end-to-end neural architectures for visual question answering". In: *Thirty-Second AAAI Conference on Artificial Intelligence*. 2018.

[2] Sören Auer et al. "Dbpedia: A nucleus for a web of open data". In: *The semantic web*. Springer, 2007, pp. 722–735.

[3] David L Chen and William B Dolan. "Collecting highly parallel data for paraphrase evaluation". In: *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*. Association for Computational Linguistics. 2011, pp. 190–200.

[4] Chenyou Fan. "EgoVQA - An Egocentric Video Question Answering Benchmark Dataset". In: *The IEEE International Conference on Computer Vision (ICCV) Workshops*. 2019.

[5] Chenyou Fan et al. "Heterogeneous Memory Enhanced Multimodal Attention Model for Video Question Answering". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2019, pp. 1999–2007.

[6] Jiyang Gao et al. "Motion-appearance co-memory networks for video question answering". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018, pp. 6576–6585.

[7] Sergio Guadarrama et al. "Youtube2text: Recognizing and describing arbitrary activities using semantic hierarchies and zero-shot recognition". In: *Proceedings of the IEEE international conference on computer vision*. 2013, pp. 2712–2719.

[8] Kaiming He et al. "Deep residual learning for image recognition". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 770–778.

[9] Yunseok Jang et al. "Tgif-qa: Toward spatio-temporal reasoning in visual question answering". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2017, pp. 2758–2766.

[10] Justin Johnson et al. "Clevr: A diagnostic dataset for compositional language and elementary visual reasoning". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2017, pp. 2901–2910.

[11] Kyung-Min Kim et al. "DeepStory: Video Story QA by Deep Embedded Memory Networks". In: IJCAI17. AAAI Press, 2017, 20162022.

[12] Jie Lei et al. “Tvqa: Localized, compositional video question answering”. In: *arXiv preprint arXiv:1809.01696* (2018).

[13] Yuncheng Li et al. “TGIF: A new dataset and benchmark on animated GIF description”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2016, pp. 4641–4650.

[14] Tegan Maharaj et al. “A Dataset and Exploration of Models for Understanding Video Data through Fill-in-the-Blank Question-Answering”. In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2016), pp. 7359–7368.

[15] Kenneth Marino et al. “Ok-vqa: A visual question answering benchmark requiring external knowledge”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2019, pp. 3195–3204.

[16] Volodymyr Mnih, Nicolas Heess, Alex Graves, et al. “Recurrent models of visual attention”. In: *Advances in neural information processing systems*. 2014, pp. 2204–2212.

[17] Karen Simonyan and Andrew Zisserman. “Very deep convolutional networks for large-scale image recognition”. In: *arXiv preprint arXiv:1409.1556* (2014).

[18] Gursimran Singh. “Spatio-temporal relational reasoning for video question answering”. PhD thesis. University of British Columbia, 2019.

[19] Robyn Speer, Joshua Chin, and Catherine Havasi. “Conceptnet 5.5: An open multilingual graph of general knowledge”. In: *Thirty-First AAAI Conference on Artificial Intelligence*. 2017.

[20] Christian Szegedy et al. “Going deeper with convolutions”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015, pp. 1–9.

[21] Makarand Tapaswi et al. “MovieQA: Understanding Stories in Movies through Question-Answering”. In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2015), pp. 4631–4640.

[22] Du Tran et al. “Learning spatiotemporal features with 3d convolutional networks”. In: *Proceedings of the IEEE international conference on computer vision*. 2015, pp. 4489–4497.

[23] Peng Wang et al. “Explicit Knowledge-based Reasoning for Visual Question Answering”. In: *IJCAI-17*. 2017, pp. 1290–1296.

[24] Peng Wang et al. “Fvqa: Fact-based visual question answering”. In: *IEEE transactions on pattern analysis and machine intelligence* 40.10 (2018), pp. 2413–2427.

[25] Dejing Xu et al. “Video Question Answering via Gradually Refined Attention over Appearance and Motion”. In: *MM ’17*. 2017.

[26] Jun Xu et al. “MSR-VTT: A Large Video Description Dataset for Bridging Video and Language”. In: IEEE International Conference on Computer Vision and Pattern Recognition (CVPR), 2016.

[27] Yunan Ye et al. “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*. ACM. 2017, pp. 829–832.

[28] Kexin Yi et al. “Clevrer: Collision events for video representation and reasoning”. In: *arXiv preprint arXiv:1910.01442* (2019).

[29] Youngjae Yu et al. “End-to-End Concept Word Detection for Video Captioning, Retrieval, and Question Answering”. In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2016), pp. 3261–3269.

[30] Kuo-Hao Zeng et al. “Leveraging Video Descriptions to Learn Video Question Answering”. In: *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*. AAAI Press, 2017, 4334–4340.

[31] Linchao Zhu et al. “Uncovering the temporal context for video question answering”. In: *International Journal of Computer Vision* 124.3 (2017), pp. 409–421.