VQuAD: Video Question Answering Diagnostic Dataset

Vivek Gupta *
IIT Kanpur
vivg126@gmail.com

Badri N. Patro †
IIT Kanpur
badri@iitk.ac.in

Hemant Parihar
IIT Kanpur
hemantp@iitk.ac.in

Vinay P. Namboodiri
University of Bath
vpn22@bath.ac.uk

Abstract

In this paper, we investigate the task of Video based Question Answering. We provide a diagnostic dataset that can be used to evaluate the extent of reasoning abilities of various methods for solving this task. Previous datasets proposed for this task do not have this ability. Our dataset is large scale (around 1.3 million questions jointly for train and test) and evaluates both the spatial and temporal properties and the relationship between various objects for these properties. We evaluate state of the art language model (BERT) as a baseline to understand the extent of correlation based on language features alone. Other existing networks are then used to combine video features along with language features for solving this task. Unfortunately, we observe that the currently prevalent systems do not perform significantly better than the language baseline. We hypothesise that this is due to our efforts in ensuring that no obvious biases exist in this dataset and the dataset is balanced. To make progress, the learning techniques needs to obtain an ability to reason, going beyond basic correlation of biases. This is an interesting and significant challenge provided through our work. We release our dataset and source code for our baseline modules in the following webpage `https://delta-lab-iitk.github.io/vquad/`.

1. Introduction

Significant progress has been achieved by the community in terms of answering queries over images [1, 24]. This progress has further lead to interest in answering queries over videos [30, 17, 35]. However, it is crucial to ensure that the progress achieved is in terms of actual visual reasoning abilities. This assurance was possible in images through the use of a diagnostic dataset CLEVR [14]. Introducing this dataset allowed evaluation of true reasoning capabilities in terms of answering questions as compared to fatuous correlation based capabilities. However, such an image based diagnostic dataset alone does not suffice to evaluate video based question answering systems. To address this requirement, we make available a video-based question answering diagnostic dataset (VQuAD). Specifically, through this dataset, we ensure that a variety of object properties, the relationship between object properties and temporal properties can be analyzed using the proposed dataset.

The dataset is motivated by the fact that while reasoning about a video, one needs the ability to analyse various factors. In addition to the spatial reasoning capabilities required for answering queries related to images, one needs to obtain reasoning abilities concerning temporal queries. Further one also requires intuitive physics based reasoning abilities [7, 3, 9] to answer queries related to collision based queries. We ensure by using an automatically function based question generation approach that there is minimal bias and therefore correlation based approaches to answer queries would not succeed. We further analyze popular baselines based on visual question answering (VQA) systems and video based question answering systems. Our analysis suggests that currently, prevalent systems are deficient in answering reasoning based queries for videos. Importantly, even methods such as FILM [29] that perform well on CLEVR [14] dataset do not do well in the context of videos. The emphasis in this dataset is on the spatio-temporal reasoning abilities. Hence, the videos themselves do not pose complex visual analytics challenges; rather, the challenge lies in the reasoning abilities. Using the currently prevalent video question answering datasets [30, 35] in addition to the proposed dataset would allow thorough understanding and diagnostics being available for the various methods. We also show that analysis of the results can be obtained by using visual explanation techniques such as GradCam [32] to understand whether the visual attention regions the methods focus on are correct or not.

The motivation for investigating video based question answering systems are for enabling a variety of applications related to obtaining systems that aid visually impaired users to visual surveillance systems. To develop reliable systems that have accurate reasoning abilities, it is crucial that we analyze the various reasoning abilities of such systems. The proposed work is a step in that direction to obtain

*Currently working at Oracle
†Currently working at KU Leuven

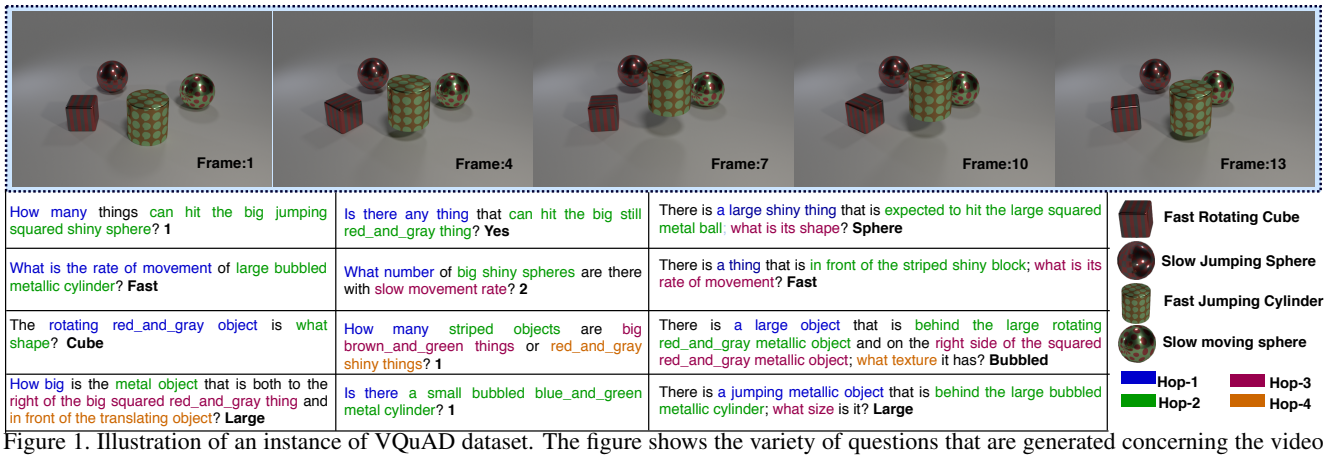


Figure 1. Illustration of an instance of VQuAD dataset. The figure shows the variety of questions that are generated concerning the video created and the difference in complexity in terms of hops for the questions.

comprehensive diagnostic analysis for video based question answering systems.

2. Related Work

Our work relates to the broader domain of solving problems in the space of vision and language based multimodal reasoning domain. There has been extensive work done in the vision and language domain for solving image captioning [40, 16, 15, 45], Visual Question Answering (VQA) [25, 1, 22, 49, 10, 20, 33] and Visual Dialog [5, 2]. In this context various datasets [26, 1, 12, 14, 41, 19, 49] have been proposed for solving image based question answering. While significant progress has been achieved in these tasks, our work relates mainly to video based question answering. The video based question answering task has shown increasing interest. There have been a number of methods and datasets devoted to understanding video content. Several datasets have been proposed to do reasoning over video content, e.g., MovieFIB [23], VideoQA [48], LSMDC [31, 36], MovieQA [35], PororoQA [17], MarioQA [28], and TVQA [18]. All these datasets focus on video content supported by natural language in the form of dialogs. However, it is possible that the methods that solve these could be based on cues in dialogs [4, 2] or visual content and could be correctly answering these without actually arriving at the 'correct' visual reasoning [34]. For instance, queries based on actions could be answered based on correlations between the image and text features without actually succeeding in reasoning. We present a diagnostic dataset that quantifies the understanding of video contents without any natural language support. In particular, we intend to evaluate the capability of statistical learning systems to capture spatiotemporal knowledge of data and perform reasoning over them. It is clear that there is a sufficient gap in the abilities for the various systems to actually reason based on the observed accuracies for various baselines in the proposed dataset.

3. The VQuAD Dataset

The VQuAD dataset provides complex and challenging questions to answer over video contents. It contains synthetic videos and questions with a balanced distribution of object properties. A *video scene* contains all the ground truth information about the object properties in the video. This ground truth assists us in the generation of questions based on various aspects of the videos.

3.1. Objects Properties

The VQuAD dataset offers two classes of properties, *spatial* and *temporal*. Spatial properties can be analyzed and reasoned over by looking at a single frame of video. To reason over the temporal property, one has to look across the video frames of the video. A brief overview of VQuAD objects is presented in Figure 2 (Left).

Spatial Properties: VQuAD family contains three object **shapes** *(cube, cylinder, and sphere)* of two **sizes** *(small and large)* and two **materials** *(metal and rubber)*. It contains three bichromatic **textures** *(squared, bubbled and striped)* on each object surface. Six **color combos** *red_and_green, red_and_gray, brown_and_green, brown_and_gray, blue_and_green, blue_and_gray* are used on objects across the dataset.

Temporal Properties: Each object in the VQuAD dataset has a **movement** *(still, rotating, jumping and translating)* associated with it. A *still* type object does not contain any actions. An object with *translating* movement has an initial position (first frame) and a final position (last frame). Spatial property of *texture* aids in recognition of *rotation* movement of some object types. A *rotating* object can rotate in either clockwise or anti-clockwise direction. Objects with actions (rotate, jump, translate) have **speed** *(fast, slow)* associated with it. Since we are maintaining a fixed video length, so time is constant, and *speed* $\propto$ *distance*.

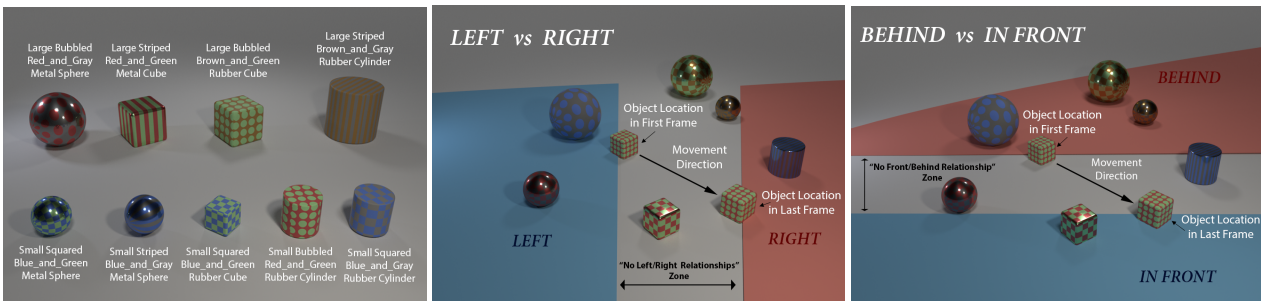


Figure 2. Object properties (Left), Relationship definitions for translating objects (Middle, Right)

For the rotating objects, we determine its pace with the rotation values (in degrees) for a single frame transition. For jumping objects, speed corresponds to jump height and for translating object it refers to the Euclidean distance. We maintain a difference in pace of fast and slow objects by a multiplicative factor of *three*, to make them distinguishable by humans. Humans also have a visual sense of predicting **possible collisions**. We have incorporated the idea of collision detection in our dataset by calculating the possibility of crashes after the last frame. For simplicity, we have avoided collisions in the ongoing video scene. We have verified that humans can answer these questions with high accuracy.

3.2. Object Relationships

This dataset contains two types of object relationships such as spatial and temporal.

3.2.1 Spatial Relationships

The VQuAD dataset includes 'left,' 'right,' 'behind' and 'in front' direction relationships among objects. We calculate them by projecting the directions concerning the camera perspective onto the ground plane. The viewpoint vector direction of the camera becomes the 'behind' vector, and the vector in its opposite direction becomes 'in front' vector. We calculate the left, and right directions similarly.

3.2.2 Temporal Relationships

We ensure consistent temporal relationships. This consists of the direction relationships involving the translating objects. We define the relationship of an object with translating object if and only if it holds this relationship with both the initial and final position of the translating object. Refer to Fig. 2(Middle, Right) for clear understanding. For example, a rotating object x is 'left' related to the translating object y if and only if it has 'left' relationship with both the initial and final position of the object y. The zone between the initial and final position of the translating object is called a 'no relationship zone.' The objects present in this zone will not have the said relationship with this translating object. The VQuAD dataset also contains *same* attribute relationship in spatial and temporal context. The spatial context involves *same* relationship over $texture$, $color$, $shape$, $size$, and $material$, whereas temporal context includes the relationship over $movement$ and $speed$. We define *'same_speed'* relationship to understand the reasoning ability involving intuitive physics [7].

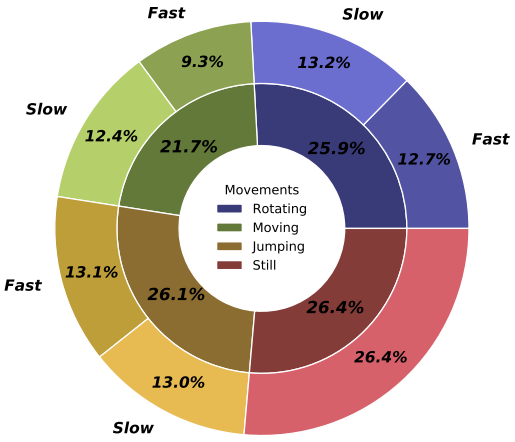
3.3. Video Scene Representation

The dataset stores the ground truth information of every video sequence in a JSON structure. It contains all the scene information about object properties such as size, color, shape, material, movement, texture, and speed. It also includes knowledge about the temporal and spatial relationships among the objects. This scene representation assists in generating questions & ground truth answers for the videos.

3.4. Video Generation

Video, in its elemental sense, is a sequence of images representing an event. We need to process videos and convert them to image sequences for their analysis. To escape this standard conversion, we directly render the frames of the videos. First, we sample a scene graph by randomly selecting objects' spatial properties like shapes, sizes, materials, colors, and texture. We place the objects such that no objects intersect, all the objects are at least partially visible, and there is some margin between different objects; this helps in avoiding ambiguity in establishing spatial relationships. Considering this as our first frame, we then sample for temporal properties like movement and speed. If the movement type is *'jump'* or *'rotate'* then movement sequence is created for all the frames with corresponding speed values. For the *'translate'* movement type, we sample a destination point for the object based on its speed. We sample the location from the circular annulus of radius quantified by each speed type. This point must have some margin with all other objects in the scene, and there must be no intersection with any objects in the path. This requirement is mandatory to avoid conflicts in temporal relationships. The movement sequence is then produced by breaking the pathway into frame length. We recursively do sampling for the destination point and if we do not find a suitable destination point after a fixed number of retries, we convert the object movement type to non-translating,

(a) Movement distribution



(b) Feature distribution

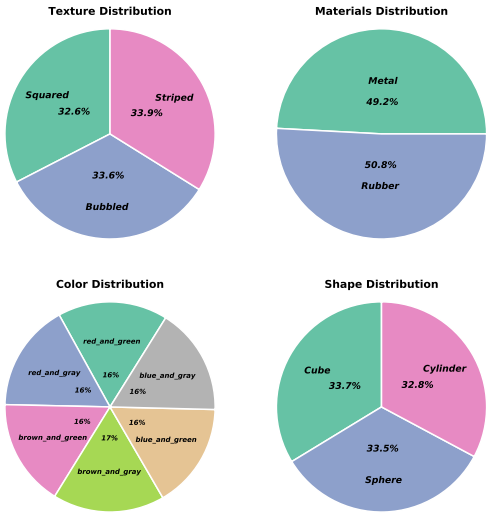


Figure 3. (a) Movement Distribution: **Inner circle:** Movement values, **Outer circle:** *speed* value per movement. (b) Feature distribution: Distribution of object attributes across all the objects in VQuAD dataset.

Question Type	Train sample	Test sample
What	579615	199627
How	77134	31987
Are	97757	36800
Is	141524	52936
Does	38425	13563
Do	21635	8047

Table 1. Distribution of Question statement type

Split	Questions	Unique questions	Videos
Train	750000	673322	3000
Val	249999	241856	3000
Test	360000	344223	1000

Table 2. VQuAD statisics

i.e., rotating, jumping, and static. This conversion does not have much effect on the distribution of the object movement types, as shown in Fig. 3(a). Finally, the video frames are rendered using Blender by iteratively applying spatial and temporal properties on objects for each frame.

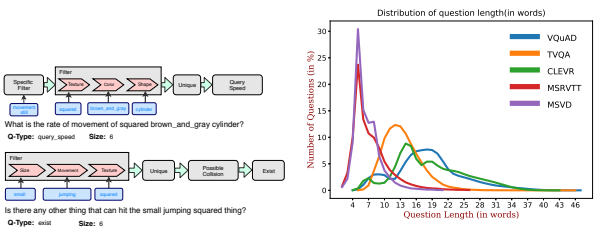


Figure 4. Explanation of question generation (Above) and Question length comparison with other VQA datasets (Below)

3.5. Question representation and families

We use an approach similar to that used by CLEVR [14] for representation and generation of questions. In this approach, questions are the representation of functional programs build using basic building blocks, which, on execution over scene information yields ground truth answers. These programs are made up of basic building blocks which handle the elementary operations like *querying*, *counting*, *comparing*, etc. Unlike other video datasets, the functional representation enables VQuAD to contain multi-hop question architecture (Fig. 1) which requires multiple iterations over questions to answer them correctly. Each iteration indulges different aspects of reasoning. For example, `What number of big shiny spheres are there with slow movement rate?` This questions needs two hops, first hop to find *big shiny spheres with slow movements* and second hop to *count them*. Some questions induce prior knowledge about some features incorporated in answering the question. For example, `What is the rate of movement of square cylinder?` imposes an *understanding* that the object in question must have a *non-still* movement type. We introduce an elementary operation of *specific_filter* to handle such priors; it *filters* out the objects with specific features enforcing the priors. We have also added a basic block of *possible_collision* which returns the set of objects expected to collide with the input object after the last video frame. A question family comprises of a template which contains functional program nodes, parameters, and constraints. It also contains multiple text templates to represent the program in natural language. For example, `What is the rate of movement of squared cylinder?` can be generated from text template *What is the rate of movement of* $< T >< M >< S >$ by mapping $< T >< M >< S >$ to *texture, material, and shape* and assigning values *squared, Nil, cylinder* respectively. The answer to this question can

(a) Question distribution (b) Answer distribution

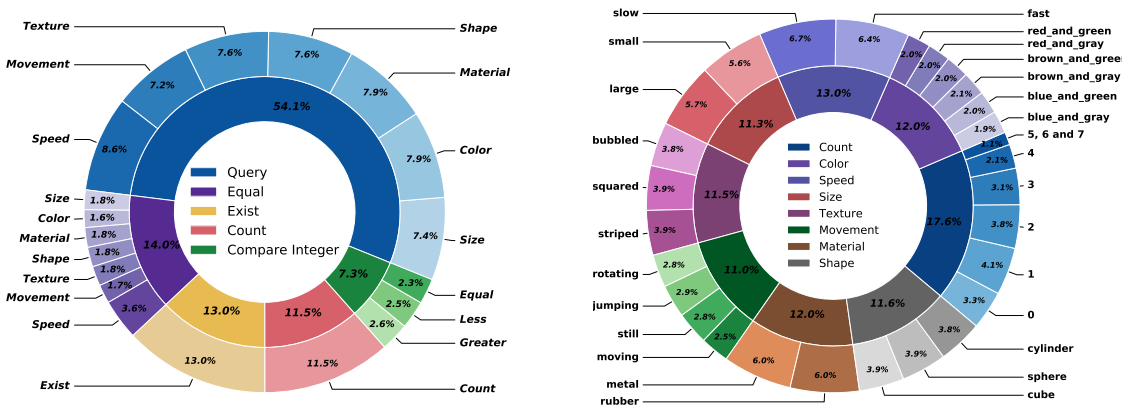


Figure 5. (a) Question distribution (left) : **Inner Circle:** Elementary Reasoning Block, **Outer Circle:** Attribute types per reasoning block. (b) Answer distribution (right) : **Inner Circle:** Answer Classes, **Outer Circle:** Values per class

be obtained by instantiating the program with *query_speed(unique(filter_shape(cylinder, filter_texture(squared, specific_filter(movement_still, scene()))))).*

3.6. Question Generation

To generate questions, we select a family among 158 question families, fill-in the template parameters, and execute them over the video scene to produce ground truth answers. The parameter values are then substituted to the text templates to generate the natural language question. Each VQuAD template can take up to 27 parameters as input, which gives it the capability to produce numerous unique questions. We use depth-first search to search for the various valid parameter instantiations for question templates. In this way, we traverse through the depth-first tree looking for parameters and pruning undesired branches using the ground truth scene information. We also use rejection sampling to maintain a uniform answer distribution, which assists in minimizing question-conditional bias. We have chosen the question-type as the type of last functional block used in its representation. The number of functional blocks in questions determines the question size.

For a model it is not easy to learn the template. There is lot of variety in the templates. The templates are designed to offer a multistep reasoning over video contents, avoiding any superficial clues to answer the questions. We have verified this by giving only question features (LSTM, Bert) to the model, no video information is provided, there is not much question conditional bias. Training data of VQuAD contains 999999 questions generated over 3000 videos. To ensure the proper validation, we shuffle the question-video pairs and split it into 750000 training and 349999 validation samples. The test data includes 360000 questions created over 1000 videos, rendered separately.

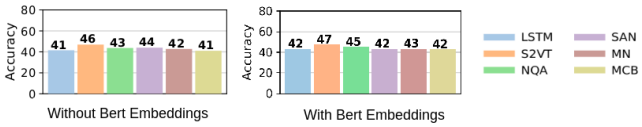


Figure 6. Overall Accuracy Comparison with or without Bert

4. Baseline Models

Video QA methods commonly use encoded frame features as video representations. C3D features [37] are also common in video domain. Some methods use recurrent neural networks [44] with iterations over video frames and question to answer them. Other methods include memory components in network architecture [43, 21, 11] for spatio-temporal attention. To answer a question based on the video, attention would play a major role. This stems by considering that methods such as S2VT [39], NQA [1], LSTM [13] are able to solve for VQA and videoQA. We compare this proposed model with the other existing models such as LSTM [13], S2VT [39], Neural QA[1], SAN [46], Memory Network [42], MCB [10] and provide results in Figure 7.

5. Experiments

We evaluate different aspects of VQuAD dataset in following ways: First, we assess the distribution of spatiotemporal properties & relationships in Fig 3 along with question and answer type assessments Fig.- 5. Second, we perform a statistical analysis of dataset presented in Table 1 & Table 2 with question length based comparison in Fig 4. Third, we evaluate various state of the art methods on our dataset, shown in Fig.- 7. Finally, we present some insights on Attention visualization Fig.- 8 comparison of each method over discrete reasoning categories. We provide video samples and more analysis results on supplementary material.

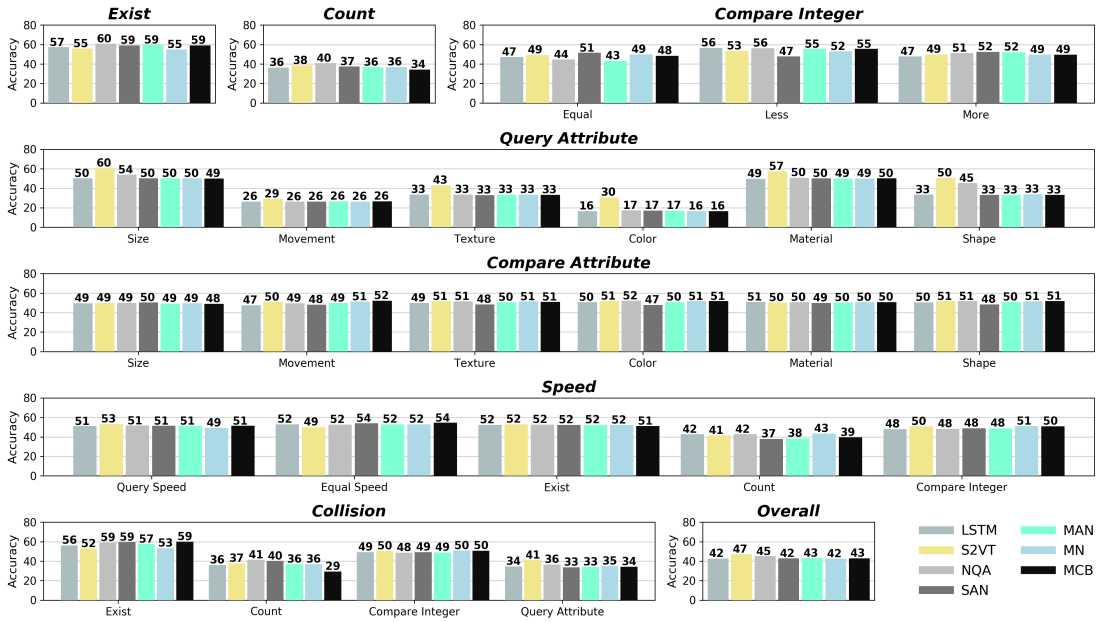


Figure 7. Question type accuracy of seven VQA methods with bert embedding in VQuAD.

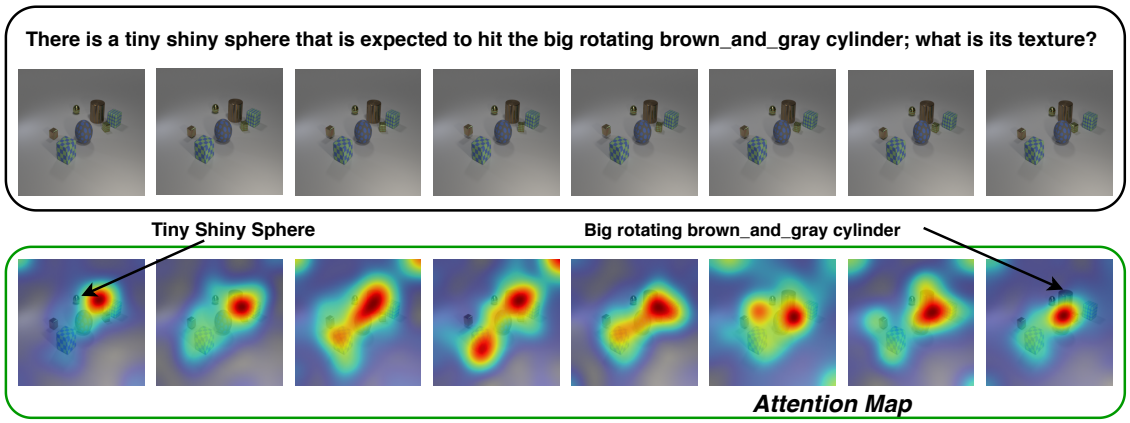


Figure 8. Attention map of SAN (Stacked Attention Network) across Visual Frames

5.1. Analysis On Question length

People usually feel that longer questions are challenging to answer since they involve multiple steps of reasoning. Here we try to investigate the importance of question size. We take the length as the number of words present in the question. For each of our models, we plot their accuracy against question size. Surprisingly, we found no correlation between question length and accuracy. Figure-4 shows the analysis of question length with accuracy.

5.2. Model Configurations

We use VGG-19 to extract 7x7x512 dimensional frame features for S2VT & NQA, and 14x14x512 dimensional features for Memory Networks, MCB, and SAN. The VGG-19 models were pretrained on Imagenet [6] and no finetuning is done. The frame was resized to 224x224 before feature extraction. We used Word2Vec[27] with embedding dimension of 200 to train the word vectors on the question representation of our dataset. LSTM is used for encoding question features and frame features with the dimension of 512 for the hidden layer. Multi-Layer Perceptron is used during the early stages of the classification with ReLU activation. All the hyperparameters, including learning rate, word embedding size, dropout, hidden size of LSTMs, and MLP layers are tuned according to the accuracy over validation dataset. Further we extend our models by extracting the features for questions using pre-trained BERT model [8], experiment with them to see the improvement in existing models. In order to get question encoding features we use LSTM over BERT embedding features. Though it gives better result compared to not using BERT, the improvement is around 1-2% which is not quite significant. Thus we use state of the art language model (BERT) for question features encoding, for videos features encoding and to

Q-Type	BERT	Best	Acc.	Human
Exist	57.0	NQA	60.0	85
Count	36	SAN	40.2	87
Compare Int	50.8	MAN	51.5	89
Query Attr.	34.9	S2VT	45.2	91
Compare Attr.	49.8	S2VT	51.0	80
Speed	44.8	MCB	49.5	81
Collision	45.2	S2VT	50.2	92
Overall	42.5	S2VT	47.1	86

Table 3. Accuracy comparison of best performing method with conditional-bias model(BERT) and human-evaluation.

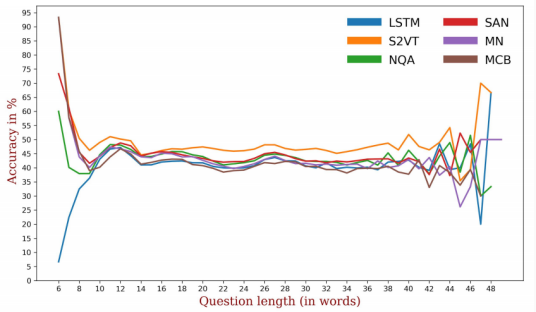


Table 4. This figure shows accuracy vs number of the word in the question in VQuAD dataset.

obtain joint embedding in order to improve reasoning between question and video. However, in our dataset, none of above approaches perform as expected. The results are presented in table 3 and summarized in Figure 7. It seems the task defined in our dataset based upon the attributes such as speed, direction, jumping, rotation and colliding appear to be much harder as compared to other videos/movies QA dataset. Hence this raises a need for pursuing research for new reasoning based methods.

5.3. Attention visualization

In order to accurately answer questions based on image sequences, a shift in attention might be required across the frames. Here we will evaluate implementation of some popular attention based methods like SAN over video QA tasks in VQuAD dataset. In Fig 8, the question asks about existence of *any small shiny sphere* which is about to hit *rotating brown and gray cylinder*. This question compels us to look at distinctive portions of images across the frames to arrive at correct answer. We can observe that SAN goes around all the objects missing the attention on the relevant *small sphere*. This shows the inability of attention based methods to visit suitable areas essential for temporal reasoning.

5.4. Model analysis of Reasoning Types

Exist: Existence type questions ask about the presence of the objects based on particular conditions. For example, *Is there anything else that has the same size as the jumping squared red_and_gray metal thing?*. The answers to these type of questions are *yes* or *no*. All the baseline models give accuracy above 50%. NQA performs best among all with an accuracy of 60% see Figure 7.

Count: Count type questions reason over the ability to count objects satisfying some conditions. For example, *How many things are there to the right of the jumping cube?*. There are eight possible answers for count question (zero to seven), so maintaining a uniform answer distribution is very difficult. We observe the LSTM accuracy of 36.0% suggesting the presence of some question-conditional bias. Relatively less occurrence of 5, 6, and 7 in the answers might be a contributing factor [Figure 5 (b)]. Counting the objects is a challenging problem, as shown in [47, 38]. NQA is the best performer in this section with an accuracy of 40% see Figure 7. We have also incorporated FiLM [29] technique in temporal context on our dataset.

Compare Integer: This type of reasoning involves comparing the integer counts in spatiotemporal zone. For Example, *Is the number of red_and_gray metallic things less than the number of translating bubbled spheres?*. Answer to this type of questions contain either *yes* or *no*. This question type requires memory, counting, and comparison to answer them correctly.

Query Attribute: In this reasoning type, the questions query about some attribute of the object. For example: *What is the texture of the jumping red_and_gray object?*. In VQuAD dataset, we have two sizes, two materials, three textures, four movements, six color combos, and three shapes. Interestingly we see that four of our models NQA, SAN, MN, MAN and MCB, produces equally likely results for attributes of movement, texture, color, and material. S2VT outperforms all the models with a significant margin for all the query attribute types.

Compare Attribute: In this category, we ask questions involving the comparison of two attributes. For example: *Does the rotating block have the same size as the jumping squared shiny thing?*. The answers under this category are *yes* or *no* type. All the models failed to capture the similarity and dissimilarity between the attribute values of all the six attributes.

Speed: This type involves speed of objects with non-still movement type, in questions. It includes the spatiotemporal reasoning type of query speed and equal speed. In query speed, we ask about the speed attribute of the object. For example: *What is the rate of movement of the large bubbled metallic cylinder?*. Answers to these questions are either *fast* or *slow*.In equal speed, we reason about the equality of the speed attribute. For example, *Does the large rotating striped red_and_gray thing have same rate as the big shiny cube?*. The answers to this type of questions are either *yes* or *no*. The collision type contains *Exist*, *Count* and *Compare Integer* as their sub-types. For example, *What number of large metallic things are there with fast movement rate?*. We can see that none of the methods can capture the *speed*

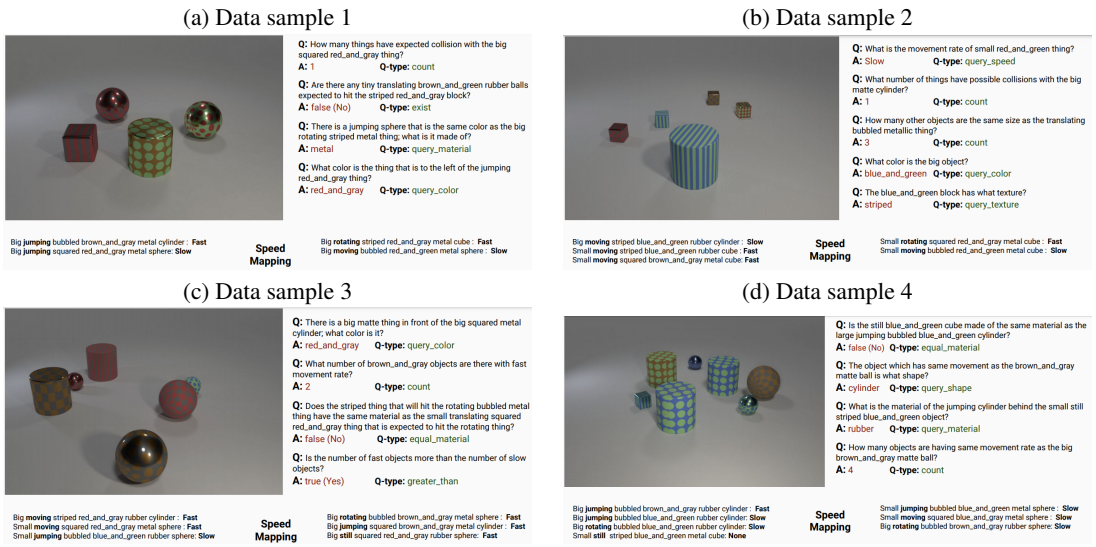


Figure 9. This figure shows data sample of our VQuAD dataset based on various object attributes.

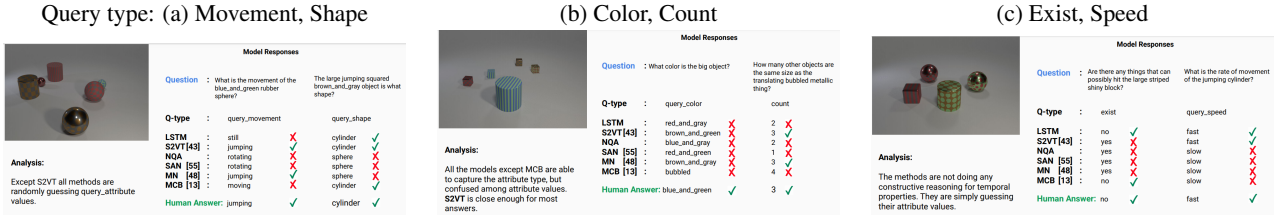


Figure 10. We show responses of various start of the art models in VQuAD dataset. We observed responses for various query types such as : Color, Count, Exist, Speed, Movement and Shape.

information of the objects. Each shows the attribute values as equally likely. This analysis indicates that identifying the speed of an object is a difficult task.

Collision: This type of reasoning asks questions about the temporal property of the expected collision of the object. For example: *Are there any other things that can hit the large striped shiny block?*. It includes *Exist*, *Count*, *Compare Integer* and *Query Attribute* as its sub-types. The accuracy of around 50% for compare integer and exist type questions shows reasoning inability of models in these areas. This result shows that all the models failed to capture the collision property, and capturing this property is a significantly tough task.

5.5. Data Samples and Model Responses

In figure-9, we provide few sample example of our dataset. In each example, example contains a given video and its corresponding questions. The question based upon temporal and spatial relation of the video. We also provide query types such as shape, movement, speed, color, direction, count, exist, texture, compare and material. We also show variant of objects based on the materials such metal or rubber in supplementary material..

We obtained responses for various baseline models as shown in figure-10. For a given video and its corresponding questions, we obtain answer based on the query types. We observe that almost all the model fails to answer the question of a video for all query-types(count, movement, shape, color etc.) in the dataset. In moment and shape query type, S2VT performs better than others. Similarly, exist and speed base query, LSTM based model preforms better than others.

6. Discussion and Future Work

In this paper we have introduced 'a VQuAD' dataset, which provides rich analysis for video based question answering. It contains a wide variety of questions over videos with multiple reasoning steps. This also contains less question conditional bias and a uniform distribution of question types and answer classes. This type of variability and analysis flexibility is not present in other datasets. In our experiments, we showed that the conventional VQA methods for Image question answering and video question answering significantly fail in spatio-temporal reasoning. We have also proposed a model (MAN) in this paper that performs comparably on this dataset with other models. To conclude, with this dataset, we pose a challenge to the community to better understand the video content and perform reasoning over them.

References

[1] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. VQA: Visual Question Answering. In *International Conference on Computer Vision (ICCV)*, 2015.

[2] Yusuf Aytar, Lluis Castrejon, Carl Vondrick, Hamed Pirsiavash, and Antonio Torralba. Cross-modal scene networks. *IEEE transactions on pattern analysis and machine intelligence*, 2017.

[3] Michael Chang, Tomer Ullman, Antonio Torralba, and Joshua B. Tenenbaum. A compositional object-based approach to learning physical dynamics. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2017.

[4] Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José MF Moura, Devi Parikh, and Dhruv Batra. Visual dialog. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[5] Abhishek Das, Satwik Kottur, José M.F. Moura, Stefan Lee, and Dhruv Batra. Learning cooperative visual dialog agents with deep reinforcement learning. In *IEEE International Conference on Computer Vision (ICCV)*, 2017.

[6] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[7] Misha Denil, Pulkit Agrawal, Tejas D Kulkarni, Tom Erez, Peter Battaglia, and Nando de Freitas. Learning to perform physics experiments via deep reinforcement learning. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2017.

[8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, 2019.

[9] Panna Felsen, Pulkit Agrawal, and Jitendra Malik. What will happen next? forecasting player moves in sports videos. In *International Conference on Computer Vision (ICCV)*, 2017.

[10] Akira Fukui, Dong Huk Park, Daylen Yang, Anna Rohrbach, Trevor Darrell, and Marcus Rohrbach. Multimodal compact bilinear pooling for visual question answering and visual grounding. *arXiv preprint arXiv:1606.01847*, 2016.

[11] Jiyang Gao, Runzhou Ge, Kan Chen, and Ram Nevatia. Motion-appearance co-memory networks for video question answering. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Jun 2018.

[12] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–9, 2017.

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

[14] Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2901–2910, 2017.

[15] Justin Johnson, Andrej Karpathy, and Li Fei-Fei. Densecap: Fully convolutional localization networks for dense captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4565–4574, 2016.

[16] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3128–3137, 2015.

[17] Kyung-Min Kim, Min-Oh Heo, Seong-Ho Choi, and Byoung-Tak Zhang. Deepstory: video story qa by deep embedded memory networks. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, pages 2016–2022. AAAI Press, 2017.

[18] Jie Lei, Licheng Yu, Mohit Bansal, and Tamara Berg. Tvqa: Localized, compositional video question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1369–1379, 2018.

[19] Junwei Liang, Liangliang Cao, Yannis Kalantidis, Li-Jia Li, Alexander G Hauptmann, et al. Focal visual-text attention for memex question answering. *IEEE transactions on pattern analysis and machine intelligence*, 2019.

[20] Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. Hierarchical question-image co-attention for visual question answering. In *Advances In Neural Information Processing Systems*, pages 289–297, 2016.

[21] Chao Ma, Chunhua Shen, Anthony Dick, Qi Wu, Peng Wang, Anton van den Hengel, and Ian Reid. Visual question answering with memory-augmented networks. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Jun 2018.

[22] Lin Ma, Zhengdong Lu, and Hang Li. Learning to answer questions from image using convolutional neural network. In *Thirtieth AAAI Conference on Artificial Intelligence*, 2016.

[23] Tegan Maharaj, Nicolas Ballas, Anna Rohrbach, Aaron Courville, and Christopher Pal. A dataset and exploration of models for understanding video data through fill-in-the-blank question-answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6884–6893, 2017.

[24] Mateusz Malinowski and Mario Fritz. A multi-world approach to question answering about real-world scenes based on uncertain input. In *Advances in neural information processing systems*, pages 1682–1690, 2014.

[25] Mateusz Malinowski and Mario Fritz. A multi-world approach to question answering about real-world scenes based on uncertain input. In *Advances in Neural Information Processing Systems (NIPS)*, 2014.

[26] M. Malinowski, M. Rohrbach, and M. Fritz. Ask your neurons: A neural-based approach to answering questions about images bibtex. In *International Conference on Computer Vision (ICCV)*, 2015.

[27] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[28] Jonghwan Mun, Paul Hongsuck Seo, Ilchae Jung, and Bohyung Han. Marioqa: Answering questions by watching gameplay videos. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2867–2875, 2017.

[29] Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. Film: Visual reasoning with a general conditioning layer. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[30] Anna Rohrbach, Marcus Rohrbach, Niket Tandon, and Bernt Schiele. A dataset for movie description. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3202–3212, 2015.

[31] Anna Rohrbach, Atousa Torabi, Marcus Rohrbach, Niket Tandon, Christopher Pal, Hugo Larochelle, Aaron Courville, and Bernt Schiele. Movie description. *International Journal of Computer Vision*, 123(1):94–120, 2017.

[32] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017.

[33] Kevin J Shih, Saurabh Singh, and Derek Hoiem. Where to look: Focus regions for visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4613–4621, 2016.

[34] Shane Storks, Qiaozi Gao, and Joyce Y Chai. Commonsense reasoning for natural language understanding: A survey of benchmarks, resources, and approaches. *arXiv preprint arXiv:1904.01172*, 2019.

[35] Makarand Tapaswi, Yukun Zhu, Rainer Stiefelhagen, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. Movieqa: Understanding stories in movies through question-answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4631–4640, 2016.

[36] Atousa Torabi, Niket Tandon, and Leonid Sigal. Learning language-visual embedding for movie understanding with natural-language. *arXiv preprint arXiv:1609.08124*, 2016.

[37] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 4489–4497, 2015.

[38] Alexander Trott, Caiming Xiong, and Richard Socher. Interpretable counting for visual question answering. 2018.

[39] Subhashini Venugopalan, Marcus Rohrbach, Jeff Donahue, Raymond Mooney, Trevor Darrell, and Kate Saenko. Sequence to sequence – video to text. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2015.

[40] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3156–3164, 2015.

[41] Peng Wang, Qi Wu, Chunhua Shen, Anthony Dick, and Anton van den Hengel. Fvqa: Fact-based visual question answering. *IEEE transactions on pattern analysis and machine intelligence*, 40(10):2413–2427, 2018.

[42] Jason Weston, Antoine Bordes, Sumit Chopra, Alexander M Rush, Bart van Merriënboer, Armand Joulin, and Tomas Mikolov. Towards ai-complete question answering: A set of prerequisite toy tasks. *arXiv preprint arXiv:1502.05698*, 2015.

[43] Jason Weston, Sumit Chopra, and Antoine Bordes. Memory networks. *arXiv preprint arXiv:1410.3916*, 2014.

[44] Ronald J. Williams and David Zipser. A learning algorithm for continually running fully recurrent neural networks. *Neural Computation*, 1(2):270280, Jun 1989.

[45] Xinchen Yan, Jimei Yang, Kihyuk Sohn, and Honglak Lee. Attribute2image: Conditional image generation from visual attributes. In *European Conference on Computer Vision*, pages 776–791. Springer, 2016.

[46] Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, and Alex Smola. Stacked attention networks for image question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 21–29, 2016.

[47] Yan Zhang, Jonathon Hare, and Adam Prügel-Bennett. Learning to count objects in natural images for visual question answering. 2018.

[48] Linchao Zhu, Zhongwen Xu, Yi Yang, and Alexander G Hauptmann. Uncovering the temporal context for video question answering. *International Journal of Computer Vision*, 124(3):409–421, 2017.

[49] Yuke Zhu, Oliver Groth, Michael Bernstein, and Li Fei-Fei. Visual7w: Grounded question answering in images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4995–5004, 2016.