HIGH SCHOOL EDITION
Journal of Student Research
Volume 10 Issue 4 (2021)

Evaluating State-of-the-Art Visual Question Answering Models Ability to Answer Complex Counting Questions

Krish Gangaraju[1] and Khaled Jedoui[#]

[1]The International School Bangalore, Bangalore, Karnataka, India
[#]Advisor

ABSTRACT

Visual Question Answering (VQA) is a relatively newer area of computer science involving computer vision, natural language processing, and deep learning. It has the ability to answer questions (currently in English) related to particular images that it is shown. Since the original VQA dataset was made publicly available in 2014, we've seen datasets such as the OK-VQA, Visual7W, and CLEVR that have all explored new concepts, various algorithms exceeding previous benchmarks, and methods to evaluate these models. However, to the best of my research, I have not seen any math or word problems being integrated into any of the VQA datasets. In this paper, I incorporate the four basic mathematical operations into the 'counting' questions of the CLEVR dataset and compare how different models fair against this modified dataset of 100,00 images and 2.4 million questions. The models we used achieved circa 50% validation accuracy within 4 epochs showing room for improvement. If VQA models can assimilate mathematics into its question understanding ability, then this can open new pathways for the future.

Introduction

Visual Question Answering has become a topic of interest for many computer vision and deep learning researchers as it poses a multidisciplinary problem. It is the amalgamation of both language and visual tasks where complex systems attempt to answer natural language questions by extracting relevant information, either from the finer details or the general scene of an image. Before the novel DAQUAR[15] dataset in 2014, due to insufficient training data and computational power to tackle this problem, this was perceived to be an impractical challenge. But in the years that followed, immense progress has been made with several new datasets and improvements to neural network architectures.

As humans, we can effortlessly classify, count, compare, make inferences, and recognize spatial relationships between objects in images using our common sense and previous experience to guide us. On the contrary, even the most intricate AI systems are far from reaching this level of nuanced visual comprehension. To achieve this understanding, models must be able to reason and apply external information to answer the question at hand, similar to what humans do.[1] Furthermore, datasets must reduce biases, as they give a false impression of a system's increasing ability to use and understand the details of the visual information to answer the questions. In the past, we've seen systems answer correctly based on a prior, following patterns and regularities, exploiting question semantics instead of focusing on the image itself.[1] Another instance of bias is the visual priming bias which addresses that people ask questions about objects only when they are present in the picture.[2] An example would be asking, "Do you see a ball in the image?" only on images that contain a ball. Many advancements have been made to combat these biases, including the CLEVR (Compositional Language and Elementary Visual Reasoning) diagnostic dataset with 100k synthetic images and 853k unique questions that test different visual reasoning abilities.[3]

The images in this dataset are visually simple and only consist of basic 3D shapes reorganized with different spatial relations and different physical properties. Each image contains several 3D-rendered objects that have one of each of the following attributes, seen in Figure 1:

(i) **Size:** Large, Small
(ii) **Shape:** Cube, Cylinder, Sphere
(iii) **Color:** Blue, Brown, Cyan, Gray, Green, Purple, Red, Yellow
(iv) **Material:** Metal, Rubber

However, the questions are more complex than the standard VQA question. By including the reduction of "question-conditional bias via rejection sampling", avoiding convoluted questions that in actuality have simple shortcuts, and containing ground-truth detailed image annotations, it ensures that a variety of reasoning skills are necessary to answer these complex questions. While the CLEVR dataset requires relatively complex reasoning skills from these machine learning models, it seems the current models are able to solve this type of complexity with ease and achieve high accuracy. With the modified dataset we will compare and evaluate the best state-of-the-art models for CLEVR to see if they can achieve the same accuracies after incorporating four fundamental operations into the CLEVR counting questions: addition, subtraction, multiplication, and division.

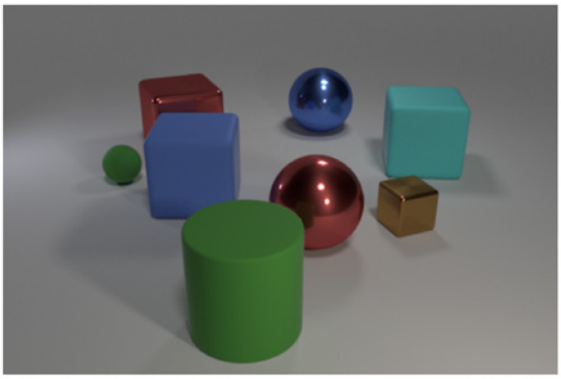

Figure 1. Attributes of CLEVR Questions. The attributes (Size, Shape, Color, Material) of each object are adjacently grouped together. The questions have empty spaces for the attributes, which are either filled or kept empty depending on the ground truth representations for the image.

In this paper we propose a modification to the synthetic CLEVR dataset to incorporate mathematical reasoning. We alter each counting type question and add another layer to the basic VQA structure, so it now combines language, visual, and simple arithmetic reasoning. To achieve the stated research goal, the paper is structured as follows. We will review previous advancements in this field to provide the reader a baseline for understanding our research.

Then, we discuss the method of constructing the CLEVR-Ops dataset, and assess whether the current VQA models can attain similar results as they did on the CLEVR.

Related Works

As previously mentioned, VQA is a challenging task due to models exploiting biases in the datasets and therefore only being able to superficially comprehend the images and questions.[4] It runs into many issues because of its inherent complexity, being at the intersection NLP and Computer Vision. Challenges arise from both subdomains such as in object detection, object counting, entity recognition, and language generation.

In many cases models and networks tend to fail on novel test instances, and lack complete understanding of both the question being asked and the image perceived. On the original VQA dataset, popular architectures such as the CNN+LSTM, CNN+LSTM with-attention, and MCB (multimodal compact bilinear) return answers after 'listening' to only half of the question.[4,5] Furthermore, there is no significant change in performance when the visual entity is removed defeating the purpose of visual question answering. CLEVR is able to resolve this issue due to its 'highly compositional questions'. Models needs to be adjusted to account for these questions that require a variety of visual reasoning abilities such as attribute identification, counting, comparing, spatial relationships, and logical operations.

In recent years a variety of VQA benchmarks have been proposed.[6-12] These benchmarks add a new component to the underlying simplicity of VQA using images from renowned and expansive datasets such as MS COCO with 120,000 images.[14] OK-VQA augments the idea that to answer challenging questions, external knowledge, apart from the image itself, is required to fulfill this task.[6] Visual Genome is one of the largest datasets allowing more diverse answers due to the free-form and region-based questioning method for each QA pair.[7] This method reduced biases and with the lack of binary questions, driving more complex questions to exist. Visual7w a subset of Visual genome, provides object-level grounding annotations that correlate the object in the question to their respective 'bounding box', for the seven different 'W' questions (what, where, when, who, why, how and which).[8] Visual Madlibs proposes the idea that instead of general image description, algorithms can answer for focused and detail-oriented questions by collecting automated fill-in-the-blank templates and multiple-choice questions for evaluation.[9,13] COCO-QA uses its own question generation algorithm which helps construct a dataset by converting image descriptions to questions.[10] For example, the image caption could be "There is a red balloon floating in the sky." Using that a natural language question like "What color is the balloon" can be formed. FM-IQA is a multilingual dataset that contains Chinese QA pairs which produces full sentence answers.[11] The evaluation was done quite distinctly, by human judges. They conduct a Visual Turing test to determine whether the answer was given by a human or their mQA model, and proceed to rate the answer on a scale from 0-2. MovieQA uses subtitled movies, scripts written by screenwriters, DVS narrations, and video clips to benchmark 'automatic story comprehension' from both video and textual information aiming to evaluate semantic understanding over extended temporal data.[12]

There have also been datasets that utilize synthetic images and questions for visual reasoning. A few advantages that synthetic datasets have over real-world data include: **(i)** ability to control the data; **(ii)** no data collection cost; **(iii)** lower and easier to reduce biases; **(iv)** explores high-level reason instead of purely vision tasks. The first major VQA dataset, DAQUAR, was able to use these advantages. Though relatively small, it contains text templates to generate questions.[15] A subset of the popular VQA dataset - SYNTH-VQA consists of 50,000 cartoon images of various scenarios with human models (different races, ages, genders, expressions), 31 animals and over 100 objects all in different positions. SHAPES is a very small unbiased dataset which is a collection of 2D geometric shapes with binary questions about attributes, positions, layout and their properties.[16] The concept is similar to that of CLEVR instead the latter is vastly bigger and utilises 3D objects.

Methodology

Constructing CLEVR-Ops

We created the CLEVR-Ops dataset which was built upon the original CLEVR diagnostic dataset. A diagnostic dataset is one that it is used to test the complex reasoning abilities of models and understand the capabilities of VQA systems. To do this, biases must be kept at a minimum and therefore synthetic images, which help create a controlled and balanced dataset are beneficial. These images are created with scene graph annotations (containing the attributes of the objects, their positions, and spatial relations). With this information, the images are rendered using the 3D computer graphics software, Blender. Furthermore, as CLEVR tests various reasoning abilities it contains several question templates that each belong to specific 'question families', that help create natural language questions. For example,

one template could be "How many <C> <M> are there?", where the <C> and <M> are parameters that need to be filled with a color and a material. In each question there are several parameters in the template but not all of them need an attribute to be filled, it can be left empty or nil. Examples of these templates can be seen in Figure 1. With 90 families, each having an average of 4 templates along with up to 19 parameters for each template (including synonyms for the attributes), CLEVR has more than 853,000 unique questions, managing to avoid ill-posed and degenerate questions resulting in reduced question-conditional bias. Out of the various question types in CLEVR, the counting questions seemed to be the most challenging as it had one of the lowest accuracies while testing – only reaching 52.2% at best.[3]

Over the last few years models such as MDETR, OCCAM, MAC, and NS-VQA have been able increase this accuracy tremendously to achieve over 97% accuracy on the counting questions.[17-20] These new models have successfully overcome the challenge that CLEVR posed. In our dataset CLEVR-Ops, we take this a step further, focusing solely on counting questions and adding a new layer of complexity, where the model is not only expected to connect visual reasoning to counting, but also identify the correct arithmetic operation and perform it within the given context of the question. For this task, we design a dataset with 2.4 million questions that follow similar question structures as the original CLEVR with modifications to incorporate math.

To incorporate math into the dataset, we only use questions that have a numerical answer i.e. the counting questions, which account for 23.6% of the original CLEVR dataset. Their answers range from zero to ten, but because there are only three to ten objects in an image, having a uniform answer distribution becomes challenging. An unequal answer distribution could cause the model to pre-emptively decide on an output/answer simply because it is very common. For example, if 50% of the answers were 1, the model may take a shortcut, choose 1, and get the answer correct despite not understanding the question.

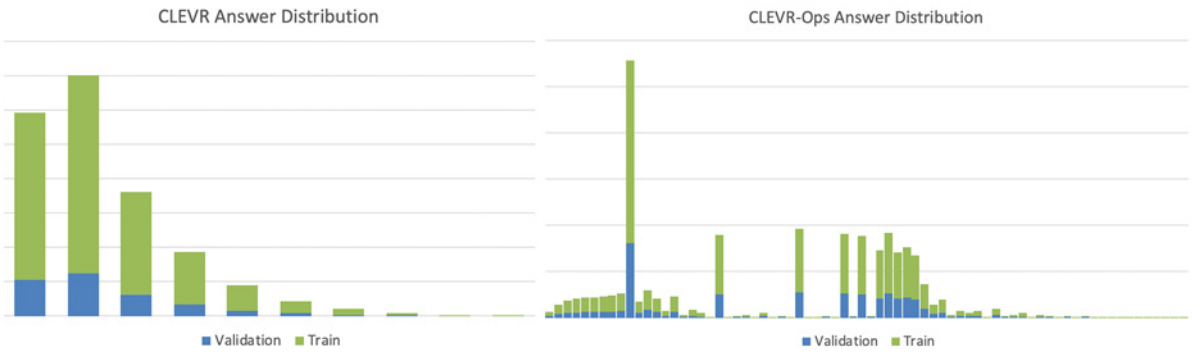

Figure 2. CLEVR vs CLEVR-Ops Answer Distribution. Compared to CLEVR, our dataset has a more even distribution. The most common answer in both datasets is 0. However, in Ops only 18% of answers in 0 whereas in the original 35% of the answers are 0.

We tried to correct this in our dataset shown in Figure 2. Once we segregated the counting questions, we modified them, as exemplified in Figure 3. The questions now incorporate a total of 80 unique text-templates (20 for each operation: addition, subtraction, multiplication, and division). For each CLEVR counting question, we added 12 variations to CLEVR-Ops (3 for each operation) while keeping key phrases from the original. Each of these three only differ by a single number. This means that each of these 3 questions follow the exact same template with different numerical information. For example, "How many objects are the same size as the blue ball plus 2?" and "How many objects are the same size as the blue ball plus 8?" are questions built from "How many objects are the same size as the blue ball?" This is to test the pure mathematical capability of the model and give it the data required to understand and learn basic arithmetic.

Applying MAC to CLEVR-Ops

To evaluate the current state of the art models for VQA on CLEVR-Ops, we implement and assess the performance of the Memory Attention and Composition [MAC] model. It performed notably well on the original CLEVR - including the counting questions - and its neural network architecture is theoretically compatible with learning of rudimentary arithmetic.

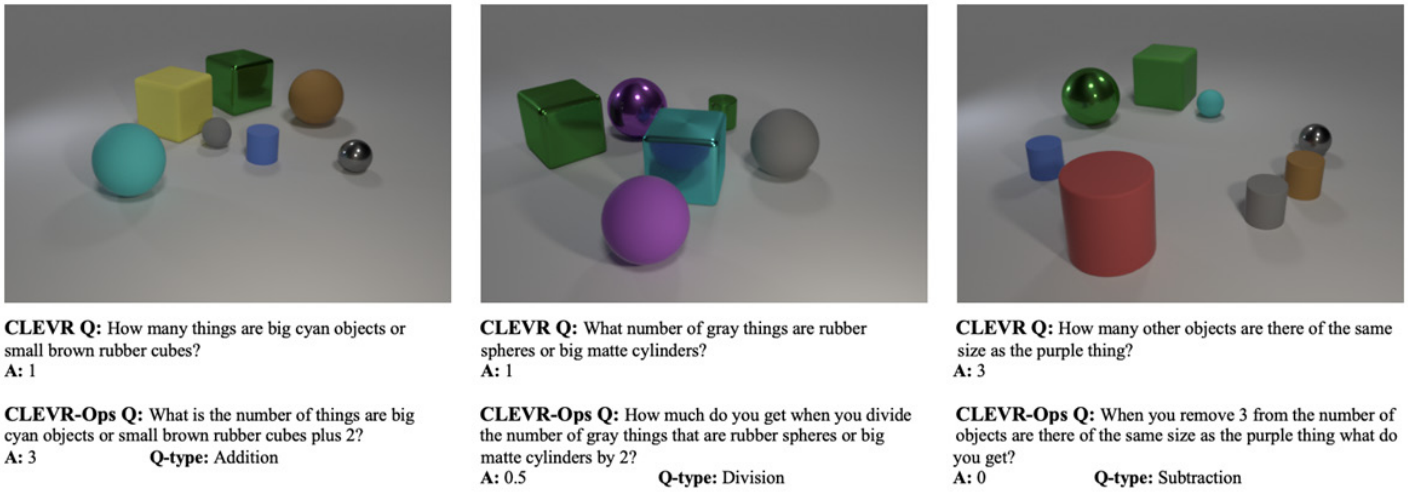

Figure 3. Modified CLEVR-Ops Questions. The original CLEVR questions are transformed. The modification adds a new layer of complexity and due to this, the structure of the question itself becomes more elaborate. The question changes from simple counting to counting and arithmetic.

Once given an image (knowledge base) and a question (task description), both of these inputs are converted to vector representations and encoded into the input unit. With a core recurrent network seen in Figure 4, the question is broken down into multiple series of operations. In the MAC cell the control state is updated after each iteration, and using this, information is extracted from the image and entered into the memory state. This process repeats until an answer is computed using the question representation and final memory (containing the final intermediate result from the relevant extracted information from the image).

The primary challenge CLEVR-Ops presents is, after being able to achieve this level of reasoning, it must also be able to solve the simple arithmetic that we have augmented into the CLEVR-Ops questions. Natural Language Processing techniques have been able to do this in the past by extracting the relevant information from simple word problems.[21]

Despite the MAC model following a similar concept of segmenting and extracting information from the larger question, the question remains of whether it will be able to overcome the new degree of difficulty.

Model Performance Evaluation on CLEVR-Ops

We evaluate the MAC model's performance on the CLEVR-Ops to establish a baseline for our dataset. We split the data into 80k images and 2.4 million question pairs into 82% training and 18% validation. After encoding the input images feature vectors using faster-RCNN architecture, we train three different variants of the model on our dataset:

(i) Vanilla MAC
(ii) MAC with Non-recurrent Control Unit

(iii) Vanilla MAC with Self-Attention Write Unit

Each variant has the same model architecture but with a different recurrent network or cell.

Vanilla MAC: In the recurrent cell there are three components (control, read and write unit), and two states: control and memory. The control unit segments the question and updates the control state as to which reasoning operation/segment it will perform. Based on the reasoning operation, the control state "guides" the read unit in extracting relevant information from the image. Lastly, the write unit integrates this information and stores an intermediate result within the memory.

Non-recurrent Control MAC: In this variant, instead of passing through each reasoning operation one by one, it handles the question as a whole. There is a slight difference in this control unit when compared to the original and due to that, it converges faster.

Figure 4. MAC Recurrent Network. The three units: Control, Read, and Write work in conjunction to perform iterative reasoning steps for individual smaller sections of the question with information from the knowledge base (image).

Self-Attention Write MAC: When incorporating self-attention into the write unit, instead of each cell only considering the preceding intermediate result, all of the previous results are taken into account. An attention distribution over the prior steps is computed, determining the relevance i.e. how much attention should be paid to the reasoning step. Based on this, a weighted average of the memory states combines with the current intermediate result and this process repeats.

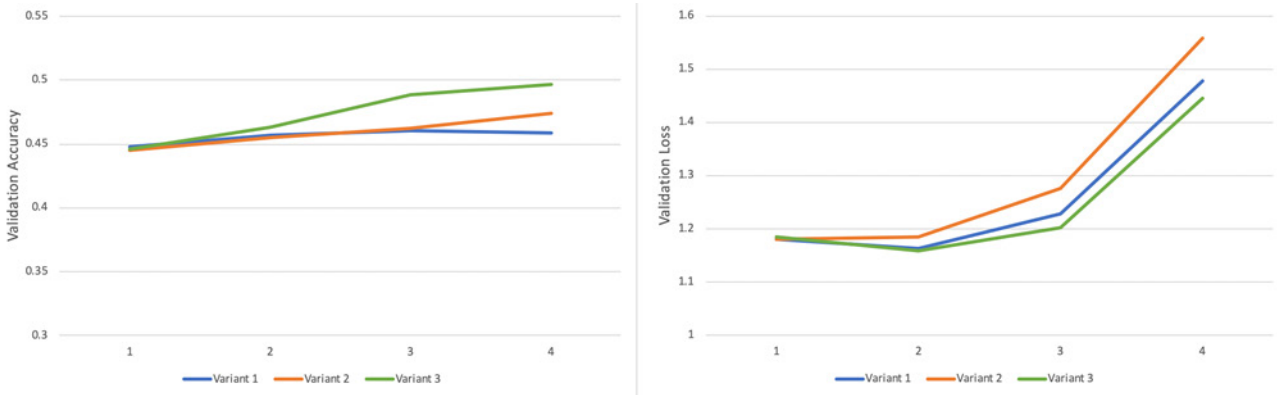

Figure 5. Model Training over 4 epochs. From left to right: (i) shows that all 3 variants are performing almost identically and using the equation of the graphs, after 20 epochs the accuracy can reach up to roughly 80%, (ii) shows us an increase in validation loss which suggests that the model is either currently overfitting or due to a diverse range of probable values or answers.

Results and Discussion

The results of our experiments are shown in Table 1 and the performance of each model is plotted against the number of epochs in Figure 5. Due to time and resource constraints, the testing was limited to 4 epochs. However as shown in Figure 5 the models were improving significantly after each consecutive cycle. This leads us to believe that with more time and training the model will become much closer to understanding the complex counting questions and accurately answer them.

MAC Model Variants	Train Accuracy	Validation Accuracy
Vanilla	0.6121	0.4588
Non-Recurrent Control	0.6359	0.4739
Self-Attention Write	0.6397	0.4964

Table 1. Final Results. After 4 epochs, shown above are the final train and validation accuracies for each of the 3 models to 4 decimal points.

Methods

The final accuracies of the models differ only slightly, meaning that the addition of these variants have very little added benefit. With this we can conclude that even the Vanilla MAC is able to answer the CLEVR-Ops questions relatively accurately.

We can see that the performance of the MAC models and its variants fall drastically on the CLEVR-Ops when compared to the original CLEVR. Our evaluation has revealed that despite the continuous progression in this field, seemingly simple tasks are demanding for these systems. Having said that, considering the vast number of possible answers available in the CLEVR-Ops dataset, achieving a 64% train accuracy and 50% validation accuracy demonstrate that the model is learning how to solve these questions to some extent and therefore shows a good proof of concept. These models are able to evaluate and perform the required steps to answer the CLEVR-Ops complex counting questions to some extent. The observations in Figure 5 suggest that with further training and execution of several more epochs, there will be improvements in accuracy and we can see to what extent these models truly understand the modified questions.

Conclusion

This paper introduces the CLEVR-Ops dataset which is designed to augment a new layer of complexity that can aid in a diagnostic analysis of state-of-the-art visual question answering models. We've successfully incorporated simple arithmetic into the original CLEVR and provide a dataset with over 2.4M questions and 80k images. We describe the usage of text templates in creating the dataset and the importance of bias minimization in our dataset.

It is important in this field of research to test this theory and discern whether the limitation lies in the available resources or in the architecture of the current models itself. Our study indicates that VQA systems still need to be improved greatly before the attempts and the practical applications of scene understanding can be tried and tested. We plan to use CLEVR-Ops as a diagnostic dataset focused solely on challenging the counting ability of the present and future VQA models in hopes of detecting question biases and evaluating new significant areas of research.

Acknowledgments

I would like to thank my advisor Khaled Jedoui for helping me with this project.

References

1. Bhattacharya, N.; Li, Q.; Gurari, D. Why Does a Visual Question Have Different Answers? In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*; IEEE, 2019.

2. Zhang, P.; Goyal, Y.; Summers-Stay, D.; Batra, D.; Parikh, D. Yin and Yang: Balancing and Answering Binary Visual Questions. *arXiv [cs.CL]*, 2015.

3. Goyal, Y.; Khot, T.; Summers-Stay, D.; Batra, D.; Parikh, D. Making the V in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE, 2017.

4. Johnson, J.; Hariharan, B.; van der Maaten, L.; Fei-Fei, L.; Zitnick, C. L.; Girshick, R. CLEVR: A Diagnostic Dataset for Compositional Language and Elementary Visual Reasoning. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE, 2017.

5. Agrawal, A.; Batra, D.; Parikh, D. Analyzing the Behavior of Visual Question Answering Models. *arXiv [cs.CL]*, 2016.

6. Agrawal, A.; Lu, J.; Antol, S.; Mitchell, M.; Zitnick, C. L.; Batra, D.; Parikh, D. VQA: Visual Question Answering. *arXiv [cs.CL]*, 2015.

7. Marino, K.; Rastegari, M.; Farhadi, A.; Mottaghi, R. OK-VQA: A Visual Question Answering Benchmark Requiring External Knowledge. *arXiv [cs.CV]*, 2019.

8. Krishna, R.; Zhu, Y.; Groth, O.; Johnson, J.; Hata, K.; Kravitz, J.; Chen, S.; Kalantidis, Y.; Li, L.-J.; Shamma, D. A.; Bernstein, M. S.; Fei-Fei, L. Connecting language and vision using crowdsourced dense image annotations https://visualgenome.org/static/paper/Visual_Genome.pdf (accessed Oct 5, 2021).

9. Zhu, Y.; Groth, O.; Bernstein, M.; Fei-Fei, L. Visual7W: Grounded Question Answering in Images. *arXiv [cs.CV]*, 2015.

10. Yu, L.; Park, E.; Berg, A. C.; Berg, T. L. Visual Madlibs: Fill in the Blank Image Generation and Question Answering. *arXiv [cs.CV]*, 2015.

11. Du, T.; Cao, J.; Wu, Q.; Li, W.; Shen, B.; Chen, Y. CocoQa: Question Answering for Coding Conventions over Knowledge Graphs. In *2019 34th IEEE/ACM International Conference on Automated Software Engineering (ASE)*; IEEE, 2019; pp 1086–1089.

12. Gao, H.; Mao, J.; Zhou, J.; Huang, Z.; Wang, L.; Xu, W. Are You Talking to a Machine? Dataset and Methods for Multilingual Image Question Answering. *arXiv [cs.CV]*, 2015.

13. Tapaswi, M.; Zhu, Y.; Stiefelhagen, R.; Torralba, A.; Urtasun, R.; Fidler, S. MovieQA: Understanding Stories in Movies through Question-Answering. *arXiv [cs.CV]*, 2015.

14. Vinyals, O.; Toshev, A.; Bengio, S.; Erhan, D. Show and Tell: A Neural Image Caption Generator. *arXiv [cs.CV]*, 2014.

15. Pietro Perona Deva Ramanan C. Lawrence Zitnick Piotr Dollar, T.-Y. L. M. M. S. B. L. B. R. G. J. H. Microsoft COCO: Common objects in context http://arxiv.org/abs/1405.0312v3 (accessed Oct 5, 2021).

16. Ren, M.; Kiros, R.; Zemel, R. Exploring Models and Data for Image Question Answering. *arXiv [cs.LG]*, 2015.

17. Korchi, A. E.; Ghanou, Y. 2D Geometric Shapes Dataset - for Machine Learning and Pattern Recognition. *Data Brief* 2020, *32* (106090), 106090.

18. Kamath, A.; Singh, M.; LeCun, Y.; Misra, I.; Synnaeve, G.; Carion, N. MDETR -- Modulated Detection for End-to-End Multi-Modal Understanding. *arXiv [cs.CV]*, 2021.

19. Wang, Z.; Wang, K.; Yu, M.; Xiong, J.; Hwu, W.-M.; Hasegawa-Johnson, M.; Shi, H. Interpretable Visual Reasoning via Induced Symbolic Space. *arXiv [cs.CV]*, 2020.

20. Hudson, D. A.; Manning, C. D. Compositional Attention Networks for Machine Reasoning. *arXiv [cs.AI]*, 2018.

21. Yi, K.; Wu, J.; Gan, C.; Torralba, A.; Kohli, P.; Tenenbaum, J. B. Neural-Symbolic VQA: Disentangling Reasoning from Vision and Language Understanding. *arXiv [cs.AI]*, 2018.

22. Goyal, A. P. S. B. Are NLP Models really able to Solve Simple Math Word Problems? http://arxiv.org/abs/2103.07191v2 (accessed Oct 6, 2021).

23. Mac-Network: Implementation for the Paper “Compositional Attention Networks for Machine Reasoning” (Hudson and Manning, ICLR 2018).