Evaluation of Systematic Errors in Visual Question Answering

Aya Nuseir[1], Moritz Vannahme, and Marc Ebner[1]

University of Greifswald, Institute of Mathematics and Computer Science, Germany,
{s-aynuse,marc.ebner}@uni-greifswald.de,moritzav98@gmail.com

Abstract. Counting questions are considered to be a subfield of the Visual Question Answering (VQA) research area. To evaluate VQA systems properly, a VQA dataset is needed in which all possible answers for all possible counting questions occur equally often. For this purpose, a generator program is developed to create a balanced dataset automatically to help in analyzing the VQA general network architecture and the VQAv2 dataset. The results show that the achieved accuracy of VQAv2 is mostly due to the structure of the questions and answers. On the other hand, when using the generated dataset, the VQA network is not able to achieve an accuracy of more than 12.12%, which is far below the 35.18% in the evaluation of the VQAv2 dataset. We found that two types of information can be exploited by a VQA network in the image to achieve better results: a characteristic object colour and a fixed association of image positions with certain numbers. Our work is a starting point for further work on the analysis of systemic errors in VQA, especially in the area of counting.

Keywords: Counting, Visual Question Answering, dataset, generator, balanced.

1 Introduction

Over the last few decades, Deep Learning has played an important role in improving computer vision, natural language processing, and knowledge reasoning [7, 3]. The progress in these disciplines enabled researchers to extend this work by combining two or more of these disciplines to produce multimodal disciplines like Image Captioning, and Visual Question Answering (VQA). Image captioning takes an image as input and generates a textual description for this image as output [8].

Regardless of the tremendous work on image captioning, there are some obstacles; as the image caption focuses on a certain part or object of an image plus the absence of interaction between the computer and the user Gao et al.[9]. Visual question answering combines computer vision, natural language processing and knowledge reasoning. Visual Question Answering uses two components as input: an image and a textual question about this image. An answer is generated as output [7]. VQA is more complex than Image Captioning.

VQA employs Knowledge reasoning in comprehending and inferring information about a specific component of the image[10]. Two categories exist for the question of a Visual Question Answering task: closed-ended and open-ended. Examples of closed-ended questions are multiple-choice, drop-down, checkboxes, and ranking questions, these questions can often be answered with yes or no. An open-ended answer is a free-form sentence. The open-ended question could be a question about fine-grained recognition, object detection, activity recognition, knowledge base reasoning, and commonsense reasoning [3]. Open-ended counting questions are a sub-task of VQA, where the goal is to count the number of instances in an image given a question formulated in a natural language [11].

The counting questions require several tasks to fulfil [13]: understanding the nature of the goal object to count, recognizing them in the image and finally adding them up. While the Counting problem is considered a slight task for Humans, the available VQA models fight to answer any counting questions successfully outside of dataset biases. According to [14] the existence of biases in the dataset affects the performance of the models; where the inferences obtained from a biased dataset are not correct since its lack of reproducibility and generalizability. Also, most of the models tend to determine against subjects of under-represented categories.

For this, we aim to show that the performance of the VQA network by [3], with the VQAv2 dataset for the counting questions tends to the hidden information in the questions rather than involving the visual information. To demonstrate this, a VQA dataset was needed in which all possible Answers for all possible numbers of questions occur the same number of times. Therefore we generate a synthesised dataset based on OpenSceneGraph.

This paper is organized as follows. Section 2 shows some of the available counting problem studies. Section 3 describes the general structure of our proposed framework. Section 4 provides a brief analysis and descriptions of the bias that is frequently present in counting questions. Section 5 describes the dataset characteristics. The results and discussion of our analysis are discussed in Sections 6 and 7. Finally, the paper is concluded in Section 8.

2 Related Work

According to Chattopadhya et al.[12] their counting models help in improving object detection performance. An important question with respect to visual question answering is whether a VQA network is actually processing the visual information properly to answer the question. However, it is also possible that the VQA network is using information that is hidden inside the question to answer it. We have generated a VQA dataset in which all possible answers for all possible counting questions occur equally often to solve the bias problem.

Zhang et al. [13] proposed a counting module that is able to learn to count through differentiable bounding box deduplication; where their module is based on the interaction between object proposals and attention maps. What makes their module differ from the other approaches is they created counting features

using information present in the attention map over the object proposals rather than depending on counting features being present In the input

Acharya et al. (2019) [15] propose a model based on region proposals that conclude relationships between objects and background image regions. Also, they collected new complex questions using Amazon Mechanical Turk (AMT) and imported both simple and complex questions from other datasets (VQA2 and Visual Genome)

According to [14] there a two standard types of bias: Class imbalance and Covariate shift. To prevent bias in developing our datasets, we adopted the equality of the number of images and questions and answers for both training and testing sets. Also, the distribution of answers is equal. In our work, we used the binary question(yes/no answer) plus the "How many?" and developed our own images artificially. The reasons for adopting binary question type are: relevant semantic information is available in the question and this type of question is easier to evaluate [2]

3 Visual Question Answering Network

We have used the VQA-Network by Antol et al. [3]. The network is split into three distinct components: (a) Image processing: The VQA Network uses a pre-trained VGG_ILSVRC_19_layers neural network [4] for image processing. The weights for this network are frozen during the training process on the VQA dataset and are not adjusted. The result of the last fully connected layer of the network is extracted as a 2048-element feature vector. The image embedding is first transformed to 1024-dim by a fully connected layer + tanh non-linearity.

(b) Question processing: The questions are converted into a one-hot-encoding prior to processing by the network. The encoded questions are fed into the network one word at a time into a series of chained LSTM blocks. The output of a previous block together with the encoded next word in the question form the input of the next block. The output of each block is a 2048-element vector. The final vector is the output of the entire component. To cope with variable question length, before each update step, the gradients are normalized over all LSTM elements. The output of each block element vector is followed by a fully connected layer + tanh non-linearity to transform 2048-dim embedding to 1024-dim.

(c)The Final component: combines the image vector with question embedding, The resulting vector is linearly transformed into a 1000-element vector by passing the resulting vector to an MLP – a fully connected neural network classifier with two hidden layers and 1000 hidden units (dropout 0.5) in each layer with tanh non-linearity. The index of the maximum entry of this vector (1000 element vector) corresponds to the given answer using a softmax layer

This network was initially trained and evaluated on the VQA dataset created by Antol et al. [3]. This dataset is based on the Microsoft Common Objects in Context (COCO) dataset[5]. The VQA dataset was later improved upon by Goyal et al. [1]. There are several different types of questions, but for this paper,

we have reduced all types of questions to just three: 'How many...', 'Are there...', 'Is there...'.

4 Analysis

The original work by Antol et al. defined a scoring metric that awarded partial points based on how many individual answers agreed with the answer given by the network [3, 1]. For our work, we required a more strict definition of what the correct answer is. By deciding on one correct answer per question we were able to tally how many times each answer, i.e. number, showed up in the ground truth, how many times it was actually answered and how many times both coincided and the given answer was the correct one. Our accuracy scores are computed using the following formula:

$$\text{Accuracy} = \frac{\sum \text{Correct Answers}}{\sum \text{Given Answers}}$$

On the evaluation of the original VQAv2 dataset, this yields a lower accuracy score than that given by the metric used in the work of Goyal et al. [1].

As discussed by Zhang et al. [2] and applied by Goyal et al. in their improvement of the VQA dataset[1], an imbalance between different answers is a significant problem in creating and evaluating VQA datasets. It allows the network to improve its accuracy without furthering its understanding of the task by making use of biases that are inherent in the training data. While the work of Goyal et al. focused especially on binary yes/no questions, i.e. questions with only one dominating answer. They discarded other question types; where each question has a small pool of frequent answers.

Regardless, the dataset by Goyal et al. [1] is considered to be an improvement over the VQA dataset. The count questions are still highly biased, the non-uniform distribution of answers in this dataset is shown in the table below. Table 1, describes the rate for all number answers, and also shows the rate of answers for questions of the type 'How many'. For each one of them, The first column named "Answers" gives the answers whose percentage was measured. The "Expected" column shows the percentage of all expected 'number' responses (all in the validation dataset). The "Given by Neural Network (NN)" column shows the percentage out of all actually given answers.

Table 1 shows the percentage of all number answers and integer answers less than 98. Although the possible answers contain a wide range of possible numbers, the expected answers are dominated by the numbers 0 to 10 and in particular 0 to 4. This is reflected even more strongly in the distribution of the answers given. Also, Table 1 shows the percentage of answers for questions of the type 'How many' (counting questions). Here, too, the answers 0 to 10, and especially 0 to 4, dominate.

Table 1: Summary of the distribution of the answers

Answer	**Answer**[1]		**Answer**[2]		**Answer**[3]		**Answer**[4]	
	Expected	Given by NN	Expected	Given by NN	Expected	Given by NN	Expected	Given by NN
0 to 4	54%	83%	73%	91%	63%	93%	76%	94%
0 to 10	68%	88 %	92%	96%	79%	97%	94%	98%

From the previous table, the entire dataset, and thus in particular for certain questions, the answers that occur are limited to a small collection of realistic answers. While the nominal pool of all possible answers is significantly larger.

The effect this distortion has on the actual learning of the neural network is especially apparent when considering the scores achieved by a network trained only on questions and answers without associated image data. Such an approach already achieves a score (under the VQA metric) of 31.55% on 'number'-answers. A network trained and evaluated on Question+Answer+Image only improves this score to 35.18%, by 3.6%[1].This showcases the drastically small influence that image information has on the network's 'thought process'. Scores also drop significantly under the metric we chose for this paper, leading to 21.49% accuracy for 'number'-answers and 23.57% for 'How many...'-type questions.

5 Evaluation

The imbalance of the existing dataset led us to create a new balanced dataset. In order to properly analyze neural network performance for counting questions, we require a dataset where all numbers show up uniformly as answers to all questions. While sub-sampling the VQAv2-dataset was considered, we abandoned this approach as it would severely limit our ability to expand the resulting dataset. Instead, inspired by the work of Johnson et al. and their CLEVR dataset[6], we wrote a generator tool which is able to automatically create a dataset from 3D models according to specifications. We will discuss the different datasets and their properties below.

In order to perform our evaluations we trained the VQA neural network on a training set generated by our generator and then evaluated the answers it gave on a validation set generated to the same specifications. It should be noted that we did not retrain the image processing component for the neural network as it had not been trained in the original work by Antol et al. either. The hyperparameters for training were the standard ones built into the VQA-Network, except for the number of training cycles which was cut to 50000.

The images created by our generator show a 7x7 chessboard with 49 3D models placed on top of it. Each model is placed on a separate tile. Lighting is randomly positioned above the chessboard. The questions come in the form of one of the following three types:

[1] All 'number' answers.
[2] All 'number' answers< 98.
[3] All 'How many' answers.
[4] 'How many' answers < 98.

- Counting questions of the form 'How many [objects] are in the image?', subsequently abbreviated as count questions.
- Binary amount questions of the shape 'Are there [X objects] in the image?', subsequently abbreviated as amount questions.
- Binary presence questions of the shape 'Is there an [object] in the image?', subsequently abbreviated as presence questions.

As the answer is always known with perfect accuracy, all ten individual answers will be identical (we have ten answers for one question in our dataset since the VQAv2 dataset has ten answers for each question). The options that were changed between different evaluated datasets were: the distribution of questions, the question of whether or not to fix the position of counted objects and camera position, and the mechanism used to colour objects and the chessboard. The first parameter is the change between the two distributions of these questions.

There are two possible distributions of these questions. In the standard version, the counting question, as the main topic of this paper, is generated for each image and the two other questions are generated for every third image (once per constellation with a moving camera, once per three constellations with a fixed camera). Since the yes/no questions are generated in pairs, this leads to 3 counting questions to 2 number questions to 2 precedence questions (3:2:2) split and this is distribution B. For distribution A, we generate yes/no questions for each image, leading to a ratio of 1:2:2 between the three question types.

The second parameter is the location of the camera; where we have two versions of datasets static and non-static. In the static version, the camera position is locked to the board. While in the non-static version, the camera position is freely rotated on a ring around the board and object positions are completely random. The third parameter is the colouring model; three colouring models were tested. The first one was leaving all objects naively in the colouring dictated by their model, the second model was to randomly colour the objects. However, we abandoned this due to the poor interaction with texturing. It also gave the entire image a much more unrealistic feel. Instead, we switched to randomly colouring the floor tiles using the previously extracted characteristic colours of each object. The third colouring model simply does not address the issue.

6 Results

For our analysis, we generated the following five configurations for distributions A and B.

- CO-NF: coloured object and black/white floor
- NO-CF: non-modified object and coloured floor
- NO-NF: non-modified object and black/white floor
- NO-CF-S: non-modified object and coloured floor with fixed camera and fixed object positions
- NO-NF-S: non-modified object and black/white floor with fixed camera and fixed object positions

The CO-NF approach was included to ensure that our NO-CF approach is able to cope with the impact of the characteristic colour of objects.

Table 2: Accuracy Comparison - distribution A

Dataset	Total	Count	Amount	Presence
A-CO-NF	45.09	9.86	53.07	54.72
A-NO-CF	45.18	9.86	52.96	55.07
A-NO-NF	46.73	11.62	**55.45**	**55.57**
A-NO-CF-S	45.29	**11.85**	55.17	52.14
A-NO-NF-S	46.52	11.5	55.3	55.26

Table 3: Accuracy comparison - distribution B

Dataset	Total	Count	Amount	Presence
B-CO-NF	33.4	9.63	51.38	51.07
B-NO-CF	33.49	9.39	50.80	52.32
B-NO-NF	33.89	9.57	51.56	**52.72**
B-NO-CF-S	35.94	**12.12**	55.75	51.87
B-NO-NF-S	35.83	11.76	**57.49**	50.27

The accuracy scores obtained by the neural network is shown in the tables 2 and 3. The dataset configuration is shown in the first column, A/B indicates the distribution of the questions. In the following, we will refer to these datasets using their abbreviation. If the distribution is omitted, then the analysis refers to both A- and B- results.

The comparison between the CO-NF and NO-CF datasets validates our choice of using coloured floor tiles to defeat the network's ability to make use of characteristic object colours. The accuracy between the two remains the same or drops.

In both datasets, worse or similar accuracy for count and number questions is seen when switching from CO-NF to NO-CF. There is an improvement for presence questions since in NO-CF data sets it is possible to distinguish one object from another by color.

The necessity of having something counteracting the use of characteristic colour is further demonstrated in the comparison of A-NO-CF and A-NO-NF where the lack of colouring leads to a notable increase in performance for count and amount questions. The near-equal distribution of answers across all 11 possible answers clearly shows the connection between a skew in the ground truth and a skew in the given answers.

It especially shows that as the skew increases for given answers that were visible in our analysis of the original VQAv2 dataset, it is absent on equally

distributed ground truths. The accuracy of the NO-CF dataset, which can be considered the one closest to reality, is only 9.86% for count questions (9.39% on the distribution B), a little better than the accuracy of 9.09% achieved by guessing randomly. This supports our hypothesis that in its training and evaluation on the VQAv2 dataset, the network draws most of its information in correctly answering questions from the question itself rather than the image. It is especially notable that even with restrictions on the complexity of the dataset the network achieves a maximal accuracy of 12.12% (on B-NO-CF-S). Combined with the accuracy of NO-CF datasets and the performance of the language-only model on the VQAv2 dataset, This effectively demonstrates that most of the 35.18% accuracy on the VQAv2 dataset stems from patterns in the distribution of answers on questions.

For count and amount questions there is a notable improvement in the change from A-NO-CF to A-NO-NF or A-NO-CF-S and in the change from B-NO-CF to B-NO-CF-S. This demonstrates effectively that The network is able to make use of the characteristic colour to improve its performance.

However the lack of change between B-NO-CF and B-NO-NF showcases that a significant presence of amount and presence questions in the dataset are necessary to enable this use of the characteristic colour. However the lack of improvement from A-NO-CF-S and A-NO-NF to A-NO-NF-S and from B-NO-CF-S to B-NO-NF-S suggest that the network is not able to make effective use of both the artificial pattern of characteristic colour and locked object position at the same time. All three question categories lose accuracy when we change the dataset from A-CO-NF and A-NO-NF to B-CO-NF and B-NO-CF. This shows that in datasets without added patterns around which the network can orient itself (and thus implicitly real-world datasets), a high percentage of amount and presence questions can improve even the performance of face value unrelated question types. One possible explanation is that count and amount questions do not provide feedback that is useful to solving the initial challenge of correctly classifying objects. Especially for counting questions where random guesses are unlikely to hit the correct answer, negative answers provide very little information.

For presence questions accuracy drops in the change from A-NO-CF to A-NO-CF-S and it rises again with the change from A-NO-CF-S to A-NO-NF-S. This tentatively suggests that in distribution A the use of fixed object positions and the characteristic colour is not just incompatible, but that the use of fixed object positions comes with the cost of a general drop in attention to the correct classification of objects or a lower attention to the entire image. Especially the change in performance happening only in presence questions in the change from A-NO-CF-S to A-NO-NF-S and the strong similarity in accuracies between A-NO-NF and A-NO-NF-S might imply that fixed image positions are disregarded entirely by the network on the A-NO-NF-S dataset. This might be investigated in future work which evaluates a network trained on an A-NO-NF dataset on an A-NO-NF-S dataset and vice versa.

7 Discussion

This paper has demonstrated that the VQA network has difficulties answering counting questions. Why this may be the case is still unknown. We propose two different theories on why the VQA fails to answer counting questions.

It might be possible that the failure of the network is due to single image processing. In the VQA network, there are only two layers in which the network can alter the encoding of its image information: The layer that accepts the feature vector and prepares it for the combination of the question vector and the layer that converts the combined vector. The second, however, is reserved for extracting the answer and is thus severely limited in its ability to process image information. This means that the network has only a single processing step dedicated to processing image information, and the reason behind the network applying only a single mode of analysis is the lack of association between the use of characteristic colour and fixed image positions.

Besides, this prevents the network from deciding which areas of the image to consider based on the image content. At this point of processing, the question has not yet been included in the processing, since the combination of the vectors takes place later. This leads to the fact that the network modifies its analysis of the image according to the entire dataset rather than adjusting it to fit the question.

A second reason for the failure might be due to the ratio of correct to false answers. Counting questions are complex questions because they combine multiple different sources of failure, which are not effectively addressed the binary feedback of correct/incorrect. They combine classification ('Which object am I counting?'), detection ('Have I found all instances of the specified object?') and the counting itself ('How many objects have I found?'). The network can fail in all three of these steps without receiving feedback on why it failed. Especially in the important early steps of training, this means that the network will not be able to effectively adjust its internal weights to improve its accuracy on the question.

When the detection issue is eliminated the performance improves, either via the use of the characteristic colour or through fixed object positions. Also, the measured improvement is increased when the proportion of presence questions that provide binary feedback on classification (correct or incorrect classification) is increased.

8 Conclusion

This paper deals with a sub-field of VQA. It focuses on questions that require the network to count objects in the image. We have used the VQA-LSTM-CNN network to evaluate the performance of VQA on counting questions. Due to the distribution of questions and answers in VQAv2, the accuracy is better than when evaluated on our dataset. The accuracy on our dataset failed to exceed 12.12%, while the VQAv2 accuracy is 35.18%. In addition to that, we found

that two features can be used by a VQA network in the image to improve its performance. For future work, we're looking to examine new models with our datasets. The model architectures that we intend to use are based on attention mechanisms, the standard VQA model without attention and the Transformers multimodal

References

1. Goyal, Y., Khot, T., Summers-Stay, D., Batra, D. and Parikh, D., 2017. 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 (pp. 6904-6913).
2. Zhang, P., Goyal, Y., Summers-Stay, D., Batra, D. and Parikh, D., 2016. Yin and yang: Balancing and answering binary visual questions. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 5014-5022).
3. Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C.L. and Parikh, D., 2015. Vqa: Visual question answering. In Proceedings of the IEEE international conference on computer vision (pp. 2425-2433).
4. Simonyan, K. and Zisserman, A., 2014. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.
5. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P. and Zitnick, C.L., 2014. Microsoft coco: Common objects in context. In Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13 (pp. 740-755). Springer International Publishing.
6. Johnson, J., Hariharan, B., Van Der Maaten, L., Fei-Fei, L., Lawrence Zitnick, C. and Girshick, R., 2017. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 2901-2910).
7. Teney, D., Wu, Q. and van den Hengel, A., 2017. Visual question answering: A tutorial. IEEE Signal Processing Magazine, 34(6), pp.63-75.
8. Sharma, P., Ding, N., Goodman, S. and Soricut, R., 2018, July. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 2556-2565).
9. Gao, H., Mao, J., Zhou, J., Huang, Z., Wang, L. and Xu, W., 2015. Are you talking to a machine? dataset and methods for multilingual image question. Advances in neural information processing systems, 28.
10. Wu, Q., Teney, D., Wang, P., Shen, C., Dick, A. and Van Den Hengel, A., 2017. Visual question answering: A survey of methods and datasets. Computer Vision and Image Understanding, 163, pp.21-40.
11. Dancette, C., Cadene, R., Chen, X. and Cord, M., 2020. Overcoming statistical shortcuts for open-ended visual counting. arXiv preprint arXiv:2006.10079. @Commentjabref-meta: databaseType:bibtex;
12. Chattopadhyay, P., Vedantam, R., Selvaraju, R.R., Batra, D. and Parikh, D., 2017. Counting everyday objects in everyday scenes. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 1135-1144).
13. Zhang, Yan and Hare, Jonathon and Prügel-Bennett, Adam, 2018. Learning to count objects in natural images for visual question answering.arXiv preprint arXiv:1802.05766

14. Glauner, Patrick and Valtchev, Petko and State, Radu, 2018. Impact of biases in big data. arXiv preprint arXiv:1803.00897
15. Acharya, Manoj and Kafle, Kushal and Kanan, Christopher, 2019.Proceedings of the AAAI conference on artificial intelligence (pp. 8076-8084)