# Review

## Introduction to ML for Natural Language Processing

Daniele Passabì [221229], Data Science

*Making the V in VQA Matter:*
*Elevating the Role of Image Understanding in Visual Question Answering*

# 1 Summary

The paper [4] is concerned with *Visual Question Answering* (VQA), a task in computer vision in which a system receives a textual question about an image and has to deduce the answer [5]. According to *Goyal et al.*, the inherent structure of the world we live in and the biases inevitably present in our language are signals that are most easily grasped by machine learning models, which mostly ignore the visual aspect of the data. Using a dataset not constructed keeping these considerations in mind leads to a distorted conception of the real capabilities of the models, which are overestimated.

The authors do something that in our opinion research should always do: question itself in order to improve. They expand the dataset proposed by *Antol et al.* [1], seeking to improve its balance. They then employ the new dataset to test several state-of-the-art VQA models. The results obtained on this balanced dataset are significantly worse than those obtained on the original one. This empirical evidence shows that the models were exploiting language priors, instead of using the combined information of text and image.

Furthermore, the authors develop a new interpretable architecture: given an input (image, question) the model provides a counter-explanation based on examples. This means that in addition to the answer to the question asked, an image is also returned as output which the model considers visually similar to the one taken as input, but that has a different answer to the question asked.

# 2 Strength and weaknesses

One of the most striking and convincing points of this article is the way the authors justify every choice behind their actions: this gives the impression that nothing is left to chance. Furthermore, with the exception of a few minor repetitions, the text is written in an extremely clear and concise manner.

Although it would have been more than sufficient, the authors did not stop at augmenting the dataset and make use of existing architectures. In fact, they even created their own model. An implementation of this kind demonstrates in our opinion a real and deep knowledge of models and architectures, increasing the reader's confidence in the researchers.

It is not obvious to find weaknesses in this paper, which we consider to be very precise and of high quality. The methodology followed by the authors is very robust and will be discussed in detail in section 4. However, one choice we could not help but question is the use of Amazon Mechanical Turk (AMT), employed during the balancing phase of the dataset by the authors, who outsourced the task. Platforms such as AMT are often in the spotlight for ethical and moral issues [2]. The so-called *workers* perform extremely repetitive tasks for very little money and do not benefit from any of the conventional workers' rights. In fact, the platform itself is often labelled a *digital sweatshop*. The authors of the paper do not specify what salary was provided to each worker, which might suggest an unethical use of the already ambiguous platform.

# 3 Potential Impact

The authors of the article started from an intuition and, thanks to their extensive research, were able to demonstrate for the first time how the models considered to be state-of-the-art in the field of VQA were exploiting the biases present in language. This could potentially be a major milestone from which other researchers can build on and extend the work already accomplished.

An indication of how much the work of *Goyal et al.* was appreciated and acknowledged may be inferred by using Google Scholar [3], a free search engine that indexes the full text and metadata of scientific literature.

By searching within the platform for the keyword *"VQA"*, the article we are discussing is the second result (out of about 20 thousands possible papers). Furthermore, at the time of the writing of this review, the paper has more than 1300 citations.

## 4 Soundness of the Work

We find the methodology followed by the authors of the paper to be nearly flawless: they perceived a problem, questioned the current results obtained in the literature, asked interesting and complex questions that needed to be answered, and finally proceeded in the best possible way to address these questions.

Firstly, the motivation behind the enlargement of the initial dataset is very valid. The latter, in fact, presented strong imbalances. For example, as is mentioned in the paper, the answer to questions starting with *"Do you see a"* was yes in 87% of the cases. The authors' data collection work was commendable: they managed to almost double the size of the original dataset, which was already considerable, obtaining a total of 1.1 million pairs (image, question). To achieve this, they used Amazon Mechanical Turk, a crowdsourcing platform that allowed them to outsource the image selection process. Despite the misgivings we expressed in section 2, AMT is a tool that enables valid results, thanks to the possibility of pre-selection of workers.

Testing pre-existing state of the art models in the literature represents another very logical step in the authors' methodology. This choice, along with the one of using a baseline model that does not employ images to make predictions allowed them to empirically demonstrate the presence of language priors and bias in the models, and the extent to which they were not making full use of the information contained in the images.

The models were trained and tested several times, on different combinations of datasets:

- $UU$ trained on Unbalanced dataset, tested on Unbalanced dataset
- $UB$ trained on Unbalanced dataset, tested on Balanced dataset
- $B_{half}B$ trained on half Balanced dataset, tested on Balanced dataset
- $BB$ trained on Balanced dataset, tested on Balanced dataset

This train/test combination chosen by the authors provides a comprehensive answer to the research question proposed in the article. From the results, one can observe not only how VQA models trained on the unbalanced set decrease in accuracy when tested on the balanced dataset, but also how training performed on a balanced dataset (whether $B_{half}$ or $B$) benefits the performance of the architectures. Out of sheer curiosity, we would be intrigued to have the performance of models trained on $B_{half}/B$ and tested on $U$.

Focusing further on the performances of the models, we find the distinction in terms of accuracy based on the type of response the model was intended to give quite logical. In this manner, the researchers highlight how the binary case is yet more prone to language bias. The authors' approach hints at how separating the types of possible model responses can simplify the distinction of truly performing models from those that are merely exploiting language priors.

We conclude this section by discussing the novel model capable of creating counter-examples, which is very difficult to evaluate given the inherent nature of the problem. The authors could have decided the *"rules of the game"* in their favour but, although their results are the highest, they seemed to us to be as unbiased and objective as possible.

## 5 Replicability

Thanks to the very detailed explanation on the creation of the balanced dataset using AMT, it would be technically possible for an interested researcher to recreate the data augmentation procedure. Naturally, obtaining exactly the same results as the authors is statistically unlikely, which is why we find it very positive that the final balanced dataset has been made public. This not only allows everyone to assess its quality, but also enables it to be employed for further experiments which could lead to advances in the field of VQA.

The tests performed on the models should also be technically straightforward to reproduce, as the architectures used by the authors are public.

Although it can by no means be said that the explanations concerning the model implemented by the researchers are lacking in clarity, when discussing such complex architectures, it is in our opinion preferable to have the code available. Without it, trying to reproduce exactly the same architecture, even if only to verify the same thesis demonstrated by the authors, becomes very complex.

We firmly believe that the world of research should be more balanced between *competition* and *cooperation*. After all, it would not have been possible to write this article without the state-of-the-art VQA models made public, or without the open-source dataset from which the authors started to create their balanced dataset.

As a final note, we do understand that the paper aims to be very brief and concise, but we would have welcomed some more technical information. For example, what are the hyperparameters chosen for the convolutional networks or LSTMs in the model implemented by the authors? What framework was used to develop the architecture? What means are needed to be able to train models on such a large number of images?

## 6 Substance

Notwithstanding the fact that the paper was the result of a very extensive work by the authors, there are always means by which one can broaden the work undertaken. Some proposals follow in this section.

In section *4. Benchmarking Existing VQA Models*, the authors argue that more pairs (image, question) could benefit the performance of the models, that could be data starved. This is supported by showing that models trained on $B$ were more accurate than those trained on $B_{half}$. To further confirm this hypothesis, it would be interesting to train the models on progressively fewer observations. Maintaining the authors' logic and notation, the datasets could be $B_{\frac{1}{4}}$, $B_{\frac{1}{8}}$, etc. It would then become possible to show empirically how abundant data benefits VQA architectures.

Concurrently, we believe that it would be of interest to train the models on a fully balanced dataset, even if this means fewer pairs (image, question). This dataset, which we will call $B_{perfectly\_balanced}$, should have the same probability of answering every possible question. For example: if the question is *"Which is the sport being played in the picture?"* the possible answers should be tennis (25%), football (25%), swimming (25%), baseball (25%).

These approaches would enable to study the trade-off between *balance* and *data abundance*, understanding which of the two aspects is more predominant in determining the architectures' performance.

## References

[1] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433, 2015.

[2] Karën Fort, Gilles Adda, and Kevin Bretonnel Cohen. Amazon mechanical turk: Gold mine or coal mine? *Computational Linguistics*, pages 413–420, 2011.

[3] Google. Google Scholar.
https://scholar.google.com/.
Accessed: 05-06-2022.

[4] 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 6904–6913, 2017.

[5] Kushal Kafle and Christopher Kanan. Visual question answering: Datasets, algorithms, and future challenges. *Computer Vision and Image Understanding*, 163:3–20, 2017.