VQA dataset generation

Nabih Nebbache Zeryab Moussaoui Leila Hamdad

1 Introduction

The Visual Question Answering (VQA) [1] is an AI domain that has known considerable attraction lately due to the potential benefits that can result from the development of such a field. The recency of the domain implies that this latter is still a research domain, and one of the reasons the migration to the industry is not done yet is that VQA systems are very general. This can be explained by the research goal which is to create human-like systems that can answer the widest possible range of answers, whereas the industry goal is to have performant domain specialised systems. In this work, we contribute by creating a technique of VQA datasets specialisation. Indeed, as most VQA systems are Deep Learning systems, their specialisation in some field requires them to be trained on a specialised dataset in that field.

2 Approach

We designed two approaches to reach our goal. The first one consists of filtering existent VQA datasets to get one related to the chosen theme. The second one consists of creating questions/answers from an image-only dataset that is related to the chosen theme.

2.1 Filtering VQA datasets

Each element of a VQA dataset is an image, a question and one or many answers. Applying filters on the dataset means filtering on one of the data composing an element of the dataset. We chose to apply 2 kinds of filters in our work.

2.1.1 Filtering on images

The filtering on images consist of extracting the objects present on the image and then comparing them with a set of keywords containing the chosen theme related objects' names. This technique fits with VQA datasets that have information about the objects present in the images.

2.2 Filtering on questions

The filtering on questions consist of extracting the words from the questions and then comparing them with a set of keywords containing the chosen theme related words. The advantage of this technique is that it works with every VQA dataset.

After applying these two techniques, we found that there are some false positive elements in the filtered dataset. We applied a negative filtering, that is: We apply a second filter on the filtered dataset, we take elements whose images doesn't contain an object present in a keyword set in the case of filtering on images, or we take elements whose questions doesn't contain words present in the keyword set in the case of filtering on questions. The keywords set contains keywords that belong to a theme that opposes the desired theme (for example: outdoor scenes opposes indoor scenes).

2.3 Generating questions and answers from image-only datasets

Image-only datasets like MSCOCO [2] are more important in term of data volume than VQA datasets, hence exploiting these datasets by generating questions answers from images will increase the amount of resulting VQA data and their variance. In order to generate questions/answers from image, we went through these following steps:

2.3.1 Object detection

The first step is to get from each image a higher representation especially contained objects with their position. The resulting data must have a scheme as shown in the figure 1.

Dataset	Number of filtered Elements	Original number of elements	Ratio of filtered elements
Visual Genome	439,425	1,773,258	24.8%
GQA dataset	6,645,451	16,317,209	40.7%
Train VQA V2	71041	443,757	16.0%
Val VQA V2	34732	214,354	16.2%
Total	7190649	18748578	38.4%

Table 1: This table summerizes the results obtained on some VQA datasets after applying the VQA dataset filtering method

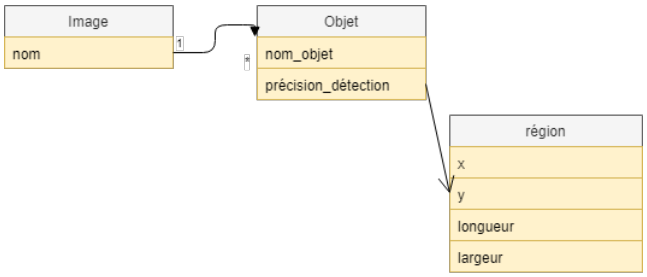

Figure 1: UML diagram representing the new image representation data format

Once the images transformed, we can apply the next step.

2.3.2 Generate questions and answers

After retrieving in each image the present objects and their positions, we applied these data to a bunch of question templates we designed. There are 5 type of generated questions:

Counting questions: this type questions the number of objects in the image. One of the templates used is: How many *{object}*s are there ?

Detection questions: This type questions the existence of objects in the image. One of the templates used is: Is there a *{object}* *[here]* ?

Relative detection questions: This type questions the existence of objects relatively to other objects' positions. One of the templates used is: Is there a *{object1}* on the *{position}* of the *{object2}* ?

Relative counting questions: This type questions the number of objects relatively to other object's position. One of the templates used is: How many *{object1}*s are on the *{position}* of the *{object2}* ?

Room detection: This type of questions is specific to the indoor scenes dataset, It aims to query on the location provided by the image. One of the templates used is: Is this a *{room}*?

To generate questions and answers, we replace the arguments inside the questions templates either by data present in the image or by random fake data. Concretely, we apply the following substitutions:

- *{object}* or *{objectN}* (with N being a Natural number) by the name of an object present or not in the current image.
- *{position}* by either *left* or *right*.
- *{room}* by the name of a room.
- *[x]* either by x with x being any text or nothing (what inside the brackets is optional), choosing between the two is done with Bernoulli's law with p = 0.5.

3 Application

We applied our technique to generate a VQA indoor scenes dataset. We started by filtering existing VQA datasets. We first generated theme related keywords set for filtering and outdoor scenes keywords set for negative filtering. Then we applied filtering by images on Visual Genome dataset and GQA dataset because these last contain information about objects in each image. We applied the filtering on questions on the VQA V2 dataset [3]. In order to enrich the resulted dataset, we applied the second technique which is the generation of questions and answers from image only dataset on the NYU Depth V2 dataset [4] which is an indoor scenes images dataset containing 400 000 element. The object detection was done with YOLO 9000 [5]

4 Results

The table 1 summarizes the results obtained after applying the above described techniques for generating an indoor scenes dataset. The figure 2 summarizes the results obtained after applying the technique of generating questions answers from image-only datasets on 60,000 images from NYU Depth V2 dataset. We remark that no object is detected in most images in the dataset, this is due to many factors amoung them: (i) YOLO is not the most performant object detection algorithm. (ii) We use a general pretrained model, not specifically trained on indoor scenes. (iii) We only tested on bathroom scenes, the algorithm may not be efficient on detecting objects present on that kind of scenes.

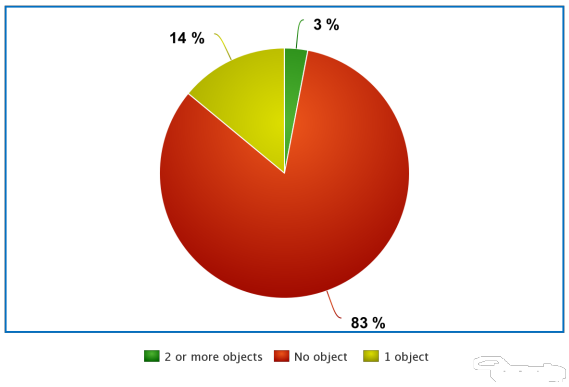

Figure 2: Chart reprsenting the distribution of the number of detected objects in 60,000 images from NYU Depth V2 dataset. In 83% of cases in the dataset, no object was detected. In 14% of cases in the dataset, only one object was detected. In 3% of cases in the dataset, 2 or more objects were detected

5 Perspectives and improvements

This project my help to create specialized VQA dataset in any real-life domain, hence we expect more specialized dataset created by this technique. We also expect improving these techniques. We present in the following points some paths to follow in order to bring some improvements to the project:

- Automatically generating filtering-VQA-datasets keywords-files from the chosen theme by using NLP-related methods.
- Using a better objects detection algorithm to generate questions answers from image-only datasets
- Automatically training the object detection algorithm on the chosen theme by training this latter on the filtered Visual Genome images since this latter contain images whose objects are surrounded with boxes.

References

[1] Stanislaw Antol et al. “Vqa: Visual question answering”. In: *Proceedings of the IEEE international conference on computer vision*. 2015, pp. 2425–2433.

[2] Tsung-Yi Lin et al. “Microsoft coco: Common objects in context”. In: *European conference on computer vision*. Springer. 2014, pp. 740–755.

[3] Yash Goyal et al. “Making the V in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering”. In: *Conference on Computer Vision and Pattern Recognition (CVPR)*. 2017.

[4] Pushmeet Kohli Nathan Silberman Derek Hoiem and Rob Fergus. “Indoor Segmentation and Support Inference from RGBD Images”. In: *ECCV*. 2012.

[5] Joseph Redmon and Ali Farhadi. “YOLO9000: better, faster, stronger”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 7263–7271.