Structured Data - Visual Question Answering Systems

Raphael Attali

May 2019

1 Introduction

Detecting elements in a picture is a common task in Deep Learning. As well, the current stage in Natural Language Processing allows an artifical intelligence to answer some questions. In 2015, the Visual Question Answering brought these two fields together. The idea is simple, ask a question to the algorithm about an image, and getting the correct answer.

Mixing many fields of study, NLP and image processing, has been already done before, namely with the image captioning. The image captioning, well-named because it depicts an image, requires an excellent understanding of the objects and their interactions one to another in the first hand, and on the other hand it is necessary to have a great understanding of the language to be able to generate a correct sentence.
In the field of Question Answering, we do have these constraints, and we have even more complexity. Indeed, the algorithm need to understand correctly the question, and to exploit the image - more precisely, the interactions implied by the question itself that are present in the image. Finally, it has to find the correct answer and also the correct format of answer. For us, "how many chairs are on the picture" means naturally that the answer is a number, but the network has to learn this kind of things.

All these difficulties make the Question Answering a real challenge, but with great stakes. Many studies were lead until now, only a few were published. The authors of the article [1] we are studying explain they found some improvements in the field and exceed the result of the current state-of-art, and evaluate the enhancements in the network as it will be 1.8 times better than any published result. To be able to understand these improvements, we will first give an overview about the state of the art in this field of study.

2 General Results - Baselines

As we have seen before, the field of Question Answering is the result of many different problems : we understand the question with NLP network, then we generally add a CNN to process the image, and the result of these two complex parts will lead to an output, the answer. We propose to discuss about each of these fields separately, to have a better overall understanding, and we will discuss how to regroup everything at the end of the section. Also, a few of baselines only focus on one side of the problem (often leading to mixed performance) then that is quite important to separate the problems.

All the techniques that will be depicted will be compared on the DAQUAR - *DAtaset for QUestion Answering on Real-world images* - dataset. The choice of that dataset is quite simple, as it is the only publicly available one for image-based Question Answering. We will study this dataset and all the stakes around it below.

2.1 NLP : Question understanding

The first part of the Question Answering problem is straight-forward : that is the question. Basically, the network will have to understand the question, else it will necessarily be wrong when answering. The understanding of the language is something extremely frequent in AI, and well-named Natural Language Processing. We will try to find what is the best way to achieve our goal among all the methods of NLP.

The NLP techniques referenced in the article are the following, the LSTM, the BLSTM, or the BOW.
The BOW - Bag of Words - is a commonly used model that allows you to count all words in a piece of text. Basically it creates an occurrence matrix for the questions set, disregarding grammar and word order, then the word frequencies are used as features for training a classifier. This approach seems "too easy" to work, and it is in many situations. However, in our case it may be quite a good way to manage our questions. First, the BOW will understand the type of the question easily : "how many","what", ... all these interrogative words would be spotted and the answer will at least have the good format, and we will reduce the field of answers. It may also learn about the other words, but it would not be very elaborate.
The other option is then the LSTM, or its bidimensionnal version the BLSTM. These are RNN, Recursive Neural Network, and they are extremely used in all NLP tasks. Indeed, the LSTM has been proved to be the most efficient RNN in a large set of problems, then it is quite natural that the authors chose it.

Without image, we can try to guess the answer of the question. A random guess among the most common answers (depending on the dataset) has a score of 0.3 on DAQUAR dataset, according to the article. The BOW and LSTM

alone got respectively a score of 0.4319 and 0.4350. As planned, the LSTM is the best, however the BOW is a good competitor despite its apparent simplicity for this kind of problem. We will see in the following what would behave the best when paired to the rest of the network.

2.2 CNN : exploiting the image

It is extremely common to use CNNs *Convolutional Neural Network* to work on images. CNNs can be thought of automatic feature extractors from the image. If we use another algorithm with pixel vector, the spatial information would be lost, a CNN effectively uses adjacent pixel information to effectively downsample the image by convolution, then it uses a prediction layer at the end.

The choice of the CNN is then pretty obvious. However, the choice of the architecture is very important. Basically, the article seems to focus on two main decisions about the architecture : with, or without dimensionality reduction in the middle.
Dimensionality Reduction plays a really important role in machine learning, especially in our case where we are working with thousands of features. Principal Components Analysis are one of the top dimensionality reduction algorithm, however the article does not mention it. It is not mentionned either which one is the best on his own as "deaf" prediction, that is to say which one has the best score without being given the question (we just give the type though, else it would lead to awful performance).

If we don't have the "deaf" performance, we may just assume that the two can be linked to different NLP methods. More precisely, it would be difficult to link a CNN without dimensionality reduction to a LSTM, because it would remain extremely long to train.

2.3 Mixing NLP and CNN

The paragraph above explained that we can't use a CNN without dimensionality reduction (that we will call IMG in the rest of the report, like in the article) with complex NLP network. We have to find a balance : using IMG, but not being able to use the LSTM, or using LSTM and help the training by adding a dimensionnality reduction layer. That are exactly the solutions proposed in the article.
Then, we have the following networks that are proposed : VIS (CNN with dimensionality reduction) + LSTM, or IMG + BOW, as BOW can handle high dimensionality easily. We also add another one, a bidimensional VIS + LSTM, well-named 2VIS+BLSTM. Finally, we also propose something interesting, a "FULL" mode that is an average of the 3 models above.

In order to choose which one is the best, we need to know the issues and difficulties that would be encountered, thus we have to study the dataset. Then we would be able to find the correct preprocessing scheme and the best model to solve our Question Answering issue.

3 The Dataset

3.1 DAQUAR dataset

Understanding the dataset is crucial. The dataset that would be used is DAQUAR, a dataset with 12 468 questions for image-based Question Answering. There are mainly three types of questions in this dataset: object type, object color, and number of objects. Some questions may be hard, so hard that a human may fail.

We have several issues that may make the answering difficult showing both NLP and CNN parts must be very robust. We propose to enumerate and explain some of them. [2]

Some of the questions refer to spatial relation like 'behind'. Answering that question is dependent on the reference frame, then the network should be aware of such a difficulty ! We usually tend to think that we uses observer-centric view, however that's not always the truth.
Also, some annotators use variations on spatial relations that are similar, for example 'beneath' is closely related to 'below'. Moreover the annotators are using different names to call the same things, such as 'nightstand', 'stool', and 'cabinet'. That means the NLP model should be well-enough trained to be able to handle synonyms.
Some questions are ambiguous, for example : "how many doors in the picture ?". A locker may count, or not, in the doors count : more than languages understanding, it is now refering to common-sense.

Another tricky questions may refer to references like 'corner' that are difficult to resolve given current computer vision models. Such scene features are frequently used by humans. Another example where the network would fail contrarily to a human is the case where we have to find objects with a particularity - for example, "how many lights are on in the image ?" may be a problem for the model, because it may detect only the lamps, without understanding which one is on or off.

With all these problem, we understand even better the difficulties and the stakes about the efficiency of our models to being able to overcome such difficulties.

3.2 To another dataset

The DAQUAR dataset contains 12 468 questions, which may be not enough, regarding all the complex situations we have to face. The article propose to use another dataset to train the model and to generate questions. We can take any dataset of Image Captioning, and define some rules to generate the questions and their answer properly from the captions. We will briefly explain some of this rules.
There will be some problems about synonyms as written above. To avoid the problem, we may apply some transformations upon synonyms, a simple one is to replace every indefinite determiners by definite ones (a/an $\rightarrow$ the).
Also we will apply some transformations on the questions in order to make it more readable and less ambiguous for our model. We have seen that there may be mistakes the model can do that are linked to the complexity of the sentence. The author of the article have then decided to split all composed sentences into multiple sentences, whose structures are more easy to understand.// The article also mention a operation on the interrogative words, but we can imagine plenty transformations that will lead to proper questions/answers. Being able to generate questions will make the training way better, and the model would be able to handle cases way more complex than with DAQUAR only.

4 The model

The authors has proposed three models we already talked about, plus the "FULL" one that averages them. The authors have tested them, we will explain their results.
The authors have split the results into the 4 categories of main questions : object, number (how many objects), location (where is an object) and color (color of one object).

We are quite surprised to realize the best performances are obtained with the Bag of Words method, except in number category. The best NLP method is known to be the LSTM, however it has been here overtaken by the BOW. We can guess that the BOW is also better because the CNN linked to it has no dimensionality reduction ; this that may be the factor that makes the BOW wins. We could think the best model would have been a CNN without dimensionality reduction linked to a LSTM, but this model is really difficult to train because of this too high dimension ; moreover we don't have any evidence of that statements, maybe we can just explain that the BOW is better in efficiency than LSTM in the field of Question Answering.
The model that should be chosen is the FULL, because it obtains the best score for every category.

5 Conclusion

The field of Question Answering is just walking its first steps. With some limitations, such as the restricted number of questions type and the output that is not structured, we have built a model that can predict quite well the answer to a question.
There is still much to do in this field of research, considering all the limitations on the questions. However, this article have made a great step ahead in Question Answering. Indeed, the model performs 1.8 times better than every existing QA-model, and we have made great progress in the choice of the problem. Namely, we showed that a simple BOW can perform equally well - or better - compared to a recurrent network, the LSTM, the one we use in Image Captioning field.
There are many potential applications for QA. Probably the most direct application is to help blind and visually-impaired users. A QA system could provide information about an image on the Web or any social media. Another obvious application is to integrate QA into image retrieval systems. This could have a huge impact on social media or e-commerce. QA can also be used with educational or recreational purposes [3]. As described in this article, we can really say that the systems capable of answering image-based questions are emerging with promising results, and that all these applications may become soon a reality.

References

[1] Mengye Ren1, Ryan Kiros1, Richard S. Zemel
Exploring Models and Data for Image Question Answering.
http://papers.nips.cc/paper/5640-exploring-models-and-data-for-image-question-answering.pdf

[2] Mateusz Malinowski and Mario Fritz
Visual Turing Challenge
https://www.mpi-inf.mpg.de/departments/computer-vision-and-multimodal-computing/research/vision-and-language/visual-turing-challenge/c7669

[3] Introduction to Visual Question Answering: Datasets, Approaches and Evaluation https://tryolabs.com/blog/2018/03/01/introduction-to-visual-question-answering/