Marcel Aubry  
EC601  
Professor Osama  
Sept 18th, 2022

Visual Question Answering  
Data Science and Intelligent Systems

Visual Question Answering (VQA) is an emerging technology that can answer natural language questions about an image. This task was previously thought to be purely science fictional but through the advancement of deep learning and natural language processing, systems are now capable of answering questions that are considered “human”. VQA uses Natural Language Processing as one of the inputs (a natural language question), and the other input is the content of an image. Therefore, the system needs advances object recognition, object detection, and scene classification among many others.

Datasets are at the core of VQA, at the bare minimum, each element of the dataset contains an image, a question, and an answer to that question. Recent datasets have been generated through crowdsourcing to have an actual human give the answer. Often these questions could be simple enough for a four-year-old to answer. For a dataset to be acceptable, it must be large enough to capture the range of possibilities among the questions and the contents of each image in a real-world scenario. The most used dataset is the Microsoft Common Objects in Contexts (COCO). It consists of 3.28 \* 105 elements and a total of 2.5 \* 106 labeled instances.

In general, a VQA system has the following three procedures: image featurization, question featurization, and joint comprehension of text and image. The first step is to describe the image in terms of vectors to be able to apply different mathematical operations to it, for example, an RBG vector. A pre-trained convolutional neural network is often chosen to help in accentuating the features of an image easily. Next comes the processing of the natural question posed. Each word in the query is converted into an input vector which is looked-up on a table. Finally, the general idea of joint comprehension of the image and text is to use mathematical operations such as addition, multiplication, dot products and concatenation of similar regions.

An application for VQA could be to help blind and visually impaired individuals. With the help of these systems, these individuals could be provided with information on any image. VQA can be used for educational purposes as well. As the technology improves, and systems can answer even more complex questions, VQAs could potentially answer questions that an average human is not able to.

There is great societal significance for this area of research. For example, ultra-specialized surgeons are in huge demand, but most importantly, they are in demand across the world, so one patient may be in Turkey but another patient in critical condition is in Canada and needs to be operated the same day. To help both within the limited time frame, the doctor may decide to execute the surgery remotely. The surgery would be done over a 5G network to limit the amount of delay, but the surgeon does not have time to train a whole crew of assistant doctors/surgeons in the new location. This is where VQA can have an enormous impact.

A well-trained system would be able to assist the surgeon by responding to queries such as “zoom in on the tumor on the left of the limbic lobe” or “locate the pineal gland”. This way, the surgeon would be able to assist patients all over the world with the help of a VQA system. Using this new technology on a global scale to improve a higher quality in humanity is a key mission of creating the societal engineer.

Table

Description automatically generatedUsing the mentioned example, we will explore the status of VQA in the surgical scenes. Questions are always being asked in the medical field; however, they are not always being answered. Whether it is a shy student in class too scared of asking their highly claimed professors, or expert surgeons who are overworked and do not have time to answer questions from junior residents, the knowledge is sometimes not being shared as well as it could be. Having access to a surgical VQA system could be of great benefit as a backup option for a learner to not have to bother a more advanced surgeon. Unfortunately, there is no database containing massive amounts of annotated medical data which has been a barrier for creating such a system. The National University of Singapore recognized this lack of technology and designed a surgical-VQA task to “generate answers for questions related to surgical tools, their interaction with tissue and surgical procedures.” Using the surgical scene segmentation dataset from the MICCAI (Medical Image Computing and Computer Assisted Interventions) endoscopic vision challenge, a workflow recognition challenge dataset (Cholec80), and extending it further to introduce two new datasets for Surgical-VQA task, they employed two vision-text attention-based transformer models to perform classification-based and sentence-based answering for the surgical-VQA. This proposed encoder (VisualBERT ResMLP) outperformed the current medical VQA state-of-the-art MedFuse model in almost all datasets.

While this system answers less-complex questions, from the application’s standpoint, it unfolds the possibility of incorporating open-ended questions where the model could be used to answer complex questions related to the surgery. As we can see, VQA is still in its development stage and much more work will have to be done for systems to be able to answer more complex and less precise questions. From the research I have done, the limiting factor seems to be the lack of annotated data. Thus, form a personal standpoint, I would be interested in either developing systems that can be trained in an unsupervised way or approaching crowdsourcing from a different angle. This could be some sort of trade off for users to develop more complex questions and answers for a large dataset of images following a certain formula.

# Bibliography

* Damien Teney, Qi Wu, and Anton van den Hengel, Visual Question Answering: a tutorial, EEE: Signal Processing Magazine, 2017; 34(6):63- 75, <https://hdl.handle.net/2440/116146>
* Manmadhan, S., Kovoor, B.C. Visual question answering: a state-of-the-art review. Artif Intell Rev 53, 5705–5745 (2020). <https://doi.org/10.1007/s10462-020-09832-7>
* “Visual Question Answering.” *Visual Question Answering*, https://visualqa.org/.
* Antol, Stanislaw, et al. "Vqa: Visual question answering." Proceedings of the IEEE international conference on computer vision. 2015.
* *PDF for 2206.11053 - Arxiv.org*. https://arxiv.org/pdf/2206.11053.