**Brief research for Visual Question Answering**

The topic I have chosen is visual question answering, and the topic is about using a picture as input, analyze it using machine learning. After the process, we can achieve the function of a user asking a free-form, open-ended question using plain text, and we use natural language processing tools to analyze what the user is asking and feedback the answer.

I chose this topic because if someday, we can improve its efficiency to the point that quick analysis and feedback are possible. It can help visual-impaired people by providing things like hazard warnings or helping them locate buildings, which can be very beneficial for society. Right now, there aren’t many real-life applications I can’t find, and this might be because the processing time is taking very long, limiting its usage. In most situations, we are expecting real-time feedbacks, and in most cases, if this feedback is not in real-time, it is not helpful for us.

After doing some research, I found an open-source framework for visual question answering called MMF. It is from Facebook AI Research and is powered by PyTorch using the Python language. Its last GitHub release was back in Aug 2019, but it is still maintained by more than 20 people. The latest update was within this month, so I think it’s safe to say it is still an ongoing project. The license it uses is a 3-cause BSD license, which means that we can use it to redistribute and modify it any way we want as long as we keep the 3-cause BSD license and cite the original author.

The most common way to achieve visual question answering is by separating it into two parts and then combining it with joint embedding. So, one factor is to extract information from the given picture using methods like CNN. Then, the question is processed separately using language processing methods, things as natural language parser. Then these two data go through a classifier and give the feedback from a set of predefined answers. Another way of doing this is called compositional models, which again processes the question using language processing methods. Still, this time the picture is not processed until the question part is done, the program will pick the related models based on the critical components of the question, then use the models to compose an output. However, sometimes we do have to have more knowledge on the question that we are going to answer, models using an external knowledge base is the method that addresses this problem, it is similar to the first method, but there is an extra step in image processing, which classifies the image and get the image attribute, search that in the knowledge base, and encode the caption after searching.

I did some tests using the Pythia VQA model on CloudCV, I have noticed that it will take about 5 to 10 seconds to get the feedback, and the prediction accuracy was about 80% on simple questions like where is something and what color is something. Still, the problem with classification is that it can come with feedbacks that don't make sense if you ask a question that is not related to the picture. An example is that I asked the question "who am I,” and the feedback was "nobody" with a 60% confidence, even higher than most of the predictions that were made by this model, which is true, but it has nothing to do with the image given.

Overall, I think visual question answering topic still has a long way to go before using it in our daily lives. Still, I do think when that day comes. It can help many people in many ways, including education, picture classification/search in social media, and assisting visually impaired people.