Visual Question Answering

I went through the video and the demo link sent by you. I found it pretty interesting. The video explanation was quite basic and it served the purpose of showing the current state in the domain of VQA. The main topics attached with the VQA are tokenization, text generation, word embeddings and context modelling. The processing of questions is the most important task in these types of scenarios. The analysis of questions is important to target the spot in the image for answering the questions provided by the users. For example: There is an image shown in the demo showing a white bird in the garden. On performing convolution on the image, the spot where the bird is sitting is located. Thus, on the basis of the location of the bird, questions like “Where is the bird sitting?” , “Where is the image captured?” can be answered easily.

VQA has seen recent advances by applying the knowledge base with normal question answering techniques. Recently VQA 360 has been introduced. It is a novel technique of visual question answering on images from a 360 degree view. Unlike the normal field-view image, it captures the entire visual content around the optical center of a camera demanding more reasoning and understanding.To handle these kinds of tasks, knowledge based learning is required. For example, if an image consists of a pizza box consisting of a pizza with 4 slices left, the model will not be able to count the slices in the image because of the lack of ability to calculate. In this case(VQA 360), an external knowledge base can help in improving the results.

While playing with the cloud CV model, I witnessed a few things. The model is capable of answering the questions based on colors and figures. It also answers satisfactorily on the questions based on expressions. Whereas, questions based on counting are often not answered correctly. One more surprising thing which I observed was that it was capable of detecting the origin of a person. For example, when I uploaded my image with 2 friends of mine, it was capable of answering that we belong to india. Another spectacular result which I witnessed was that it could identify the background of that image very precisely.

Explored sources:

1. Visual question answering on 360 degree images https://arxiv.org/abs/2001.03339
2. Visual Question Answering: A Survey of Methods and Datasets https://arxiv.org/abs/1511.03416
3. Visual7W: Grounded Question Answering in Images https://arxiv.org/abs/1607.05910