Yifei Zhu

Dr. Osama Alshaykh

EC601 A1

September 15th, 2020

Project 1: Visual Question Answering

The project I choose as my first project is the topic on Visual Question Answering (VQA), which is one specific area related with Data Science and Intelligent Systems and machine learning. I think this project can help me learn more on what I will study through graduate school and help me better understand how related applications apply in real world.

As VQA is one area under data Science and intelligent Systems based on deep learning, Visual Question Answering is one area involves with Computer Vision (CV) and Natural Language Processing (NLP). In modern society, after the internet develops for years. People put effort in Artificial intelligent and try to let computers do things normally only human can do. They expect through study that computers can finish specific missions for human in a faster and accurate way. Before the deeper research in VQA and AI, the most common thing we feel in our real life related the importance of AI could be speech recognition, such as Siri people use in their iPhone, and the voice help we call banks or service companies for questions. The speech recognition helps human deals huge amount of work used only can be handled by human. Also, we see autonomous cars get more and more popular and this technology is based on AI. Data

Science and Intelligent Systems become to play a significant role towards our society. The way we live will constantly be changed by AI.

Under this background, VQA also plays an important role in changing our lives. Explanation of VQA could be: ‘A VQA system takes as input an image and a free-form, open-ended, natural-language question about the image and produces a natural-language answer as the output.’ [1] For example, if we give a picture to computer, can the computer recognize how many people are there on the picture? What are their races? Or do they look happy or sad? VQA is the technology that makes computer to understand the information of a given input in human’s way and makes computer to output the answer correctly. Here VQA is not only the simple question and answer technology that describe the information, but require the logical reasoning like human being. For example, if given a picture of a street, the VQA not only have to identify how many cars are there parking, but also to push forward to identify where can park for a new car and are those cars violate the traffic rules. Because of this requirement, VQA is one of the hardest areas in AI and still need deep study to push it develops. However, it is clearly that through the accurate rate raises, VQA could have many direct applications to our life. Such as VQA can apply to blind disable people by helping them better get information from outside. Also, VQA can apply to imagine search system, which will be an important change to social media and E-commerce. Imagine VQA plays a role to replace part of work for human in such areas and help gain huge amount of data and information, which will push these areas to improve faster.

Currently, to achieve VQA and have an accurate model, there are different ways to approach. The most common model of VQA is to use the idea of CNN (Convolutional neural network) to build. CNN is a way to process imagines and identify objects based on the shared-weight architecture of the convolution kernels or filters that slide along input features and provide translation equivariant responses known as feature maps.[2] Since computers only get inputs of imagines as pixels and it is a way to help computers identify the objects contain in an imagine. Based on CNN, one model called R-CNN (Region Proposal-CNN). From that we also have Fast R-CNN, Faster R-CNN, YOLO, and SSD to achieve object detection. Besides this model, there are also Deeper LSTM Q + norm I model [1], VIS+LSTM model [3], Neural-Image-QA model [4], etc.

One open source provided by this class is the example of using Faster R-CNN with ResNet-101 on GitHub [5]. A bottom-up-attention model. The open source has a license under the MIT License and totally free with the authority for public to use and download on GitHub. Through the information on GitHub, it provides a 70.3% overall accuracy on 2017 VQA Challenge, which seems a high rate. The open source also contains imagines that the model used for test and training. People can easily download and try the model or use their own model to test how this open source works with different features. Another open source is based on Bilinear Attention Networks, which is also famous for apply to VQA. It uses ‘attention map’ as the strategy to solve VQA questions, and unlike usual attention map, Bilinear Attention Networks uses attention to two features instead of one. Regularly: output=feature * attention map Bilinear Attention Networks: output=feature1 * bilinear attention map * feature2 [6]. It also has a MIT License and free for public to use. Similarly, as the R-CNN one, this model also achieves tests about 70% accuracy.

In conclusion, the topic on VQA has a large graph for me to research on, and by doing this first research, I gain a deeper opinion on what VQA is and how could we apply it. Moreover, since VQA is one area under data since and intelligent system, I have a more clearly and detailed picture of this large background. The solution towards VQA through CNN and R-CNN methods is a great part could be researching more with, and I have a preference to study this method with models and hope to finish a complete test through the semester.

Work Cited

[1] Antol S, Agrawal A, Lu J, et al. VQA: Visual question answering[C] //Proceedings of the IEEE International Conference on Computer Vision. 2015: 2425-2433.

[2] “Convolutional Neural Network.” Wikipedia, Wikimedia Foundation, 17 Sept. 2021, en.wikipedia.org/wiki/Convolutional_neural_network.

[3] Ren M, Kiros R, Zemel R. Exploring models and data for image question answering[C]//Advances in Neural Information Processing Systems. 2015: 2953-2961.

[4] Malinowski M, Rohrbach M, Fritz M. Ask your neurons: A neural-based approach to answering questions about images[C]//Proceedings of the IEEE International Conference on Computer Vision. 2015: 1-9.

[5] peteanderson80. “Peteanderson80/Bottom-up-Attention: Bottom-up Attention Model for Image Captioning and VQA, Based on Faster R-CNN and Visual Genome.” GitHub, github.com/peteanderson80/bottom-up-attention.

[6] .css-1cd9gw4{margin-left:.3em;}phd . “【Vqa】Bilinear Attention Networks, zhuanlan.zhihu.com/p/361318213.