university of groningen / faculty of mathematics and natural sciences / artificial intelligence

Explainability in Visual Question Answering

Graduation Project Proposal
(Computational Intelligence and Robotics)

Nino Jansen (s296590)

November 11, 2020

Internal Supervisor(s):
Prof. Dr. Lambert Schomaker (Artificial Intelligence, University of Groningen)
MA Zhenxing Zhang (Artificial Intelligence, University of Groningen)

Artificial Intelligence / Human-Machine Communication
University of Groningen, The Netherlands

1 Introduction

Artificial Intelligence (AI) has seen massive advancements over recent years. Deep neural networks have allowed AI to achieve very high performance on single-domain tasks, such as object recognition and machine translation. A common trend in recent AI research is to move away from the relatively simple single-domain tasks to more complex multi-domain ones. Visual question answering (VQA), originally proposed in (Antol et al., 2015). is such a multi-domain task. In VQA, computer vision (CV) and natural language processing (NLP) are combined to answer free-form open-ended questions on an image. These questions can be of different types, requiring different skills. They can require reasoning ability like the question: ”Will the dog catch the frisbee”?. Others require object recognition like: ”How many dogs are there?”. Some require spatial understanding like: ”What object is to the left of the table?”. The diverse nature of the questions and the skills required to answer them makes this a complicated task for AI to handle.

The models utilized to address a multi-domain problem like VQA are generally very complex and deep. The complexity makes it hard to understand the reasoning a model follows to come to its output. This is a common problem faced in the use of deep neural networks. The network can be seen as a black box, that maps input to output, without the insight of what goes on in between. Understanding the reasoning of a model can aid in improving the mistakes it makes. Being able to follow the reasoning would also help in improving people’s trust in AI models, as the output can be explained.

This thesis will concern itself with this problem, through exploring explainability in VQA. The general theme of current approaches towards explainability utilize attention heatmaps. Attention mechanisms are used to focus on the most important regions of the image for answering the question. Therefore, an attention heatmap shows the locations the model deemed most important. In Ghosh et al. (2019) the attention heatmap is utilized in combination with the scene graph to generate a natural language explanation. A scene graph is a graphical representation of the image, where nodes depict objects and edges represent relations between objects. The important entities of the image’s scene graph are determined through the heat map and used to build the explanation.

In this thesis, a novel approach towards explainability based on image generation will be explored. Instead of a natural language explanation, an image will be generated to serve as an explanation. The image is generated based on the question and the answer from the VQA model. The generated image then depicts how an image should look to produce that answer according to the model. For example, for the question and answer: ”Q: How many dogs are in the picture? A: Three”, the generated image would show a depiction of three dogs. This approach also has some other benefits. The generated image can be reused as input for the VQA model, as a way to force consistency of the answer. As the generated image will never be completely equal to the original, it can also be utilized to augment the image data. For these reasons, image generation is an interesting approach to explore.

2 Theoretical Framework

VQA can formally be seen as the task of predicting/generating a natural language answer A for an image and question pair $< I, Q >$. Image generation, in this case, reverses the task to generating an I for a $< Q, A >$ pair. Forcing consistency between the generated image and the original image could improve the VQA model performance and allow us to use the generated image as an explanation. The idea of forcing consistency has been used in image generation from a textual description. In (Qiao et al., 2019) they proposed a model they call MirrorGAN. It is an image generation model that incorporates re-description from the generated image. The idea of using cycle-consistency has also been applied in VQA to get a more robust model in (Shah et al., 2019). They use the answer of the VQA system to generate a question that is used in place of the original question on the VQA system again. Consistency between the original question and the rephrased question, in addition to consistency between the answers, is used to make the model more robust. Their architecture idea can be used as a basis for this thesis.

2.1 Visual Question Answering

Many different approaches to VQA have been proposed. The most common approach to VQA involves a joint embedding of the question and image, that is used to predict an answer using a multi-layer perceptron (MLP), as first proposed in (Antol et al., 2015). They used the last hidden layer of the pre-trained VGGNet CNN (Simonyan and Zisserman, 2014) to embed the image. The question was embedded through an LSTM and combined with the image embedding through element-wise multiplication. In terms of VQA models, this is a quite simple approach and is often seen as a baseline model.

Following the use of grid-like features from a CNN, many researchers started incorporating visual attention mechanisms into their models. In (Anderson et al., 2018) bottom-up and top-down attention mechanisms were utilized. Bottom-up attention involves the recognition of the most salient regions of the image. The Faster R-CNN network (Ren et al., 2015) is used to determine these regions. Top-down attention uses the question embedding to weigh the importance of the regions determined by the bottom-up attention. Their approach performed well, while not overcomplicating the VQA model.

2.2 Image generation

Image generation has commonly been studied as the task of generating an image for a text description or class label. Adaptations of the generative adversarial network (GAN) (Goodfellow et al., 2014) have been a common approach to image generation. In the network, two neural networks, a generator and a discriminator are trained together in the form of a zero-sum game. The discriminator is trained to determine if an image is generated or real. the generator is trained to generate images from a noise input that are able to fool the discriminator. As both networks are constantly trying to improve, eventually the generator is able to generate real looking images.

As the original GAN framework generates images only from a noise input, conditional GAN (cGAN) frameworks were introduced that add a condition (Mirza and Osindero, 2014), which allows for generation from class labels or text descriptions. In (Zhang et al., 2017) the StackGAN was proposed utilizing cGANs. This network can generate high-quality 256x256 images from a text description. It uses two stacked GANs conditioned on the text description. The first GAN sketches a 64x64 image, which is downsampled and used as input for the second GAN that fills in the details to generate the final 256x256 image. At the time of publishing, this was the state-of-the-art for image generation from sentences.

Another more recent interesting approach to image generation is the use of scene graphs. In (Johnson et al., 2018), scene graphs were used as input for an image generation network. They used graph convolutions on the scene graph to predict the scene layout. This predicted scene layout is then used in a cascaded refinement network (Chen and Koltun, 2017) to form the image. They compared their method with the StackGAN by generating a scene graph from a text description. This shows that their network produces better images than StackGAN. In this study the $< Q,A >$ pair is used for generation, which might not be informative enough to form a descriptive enough scene graph, so it is unsure how viable this method is for this case.

2.3 Datasets

Building a dataset for VQA is a resource and time-intensive process, as generating questions and answers for images is a manual process. Much effort has already been put into building well-balanced representative datasets. One of those datasets is the Visual Genome dataset (Krishna et al., 2017). It is a collection of images with corresponding annotation on objects, attributes and relationships portrayed in the image. It also contains one or more question-answer pairs for each image. In total it has over 100k images and 1700k question-answer pairs.

Another commonly used VQA dataset is the aptly named VQA v1.0/2.0 dataset. This dataset is based on another dataset called MS COCO (Lin et al., 2014), which contains images depicting complex and diverse scenes. In the first iteration (Antol et al., 2015) it contains 204k images with 760K questions and 10M answers. The second iteration (Goyal et al., 2017) aims to counter language priors by balancing the dataset. Complementary images are collected for questions to make sure each question has a pair of similar images that lead to different answers. This dataset contains 1100K questions and 11M answers. In contrary to the visual genome dataset, these datasets are specifically designed for VQA.

3 Research Question

The main goal of this thesis is to achieve explainability for VQA. The main research question therefore is:

- How can image generation be used as an explainability method for VQA.

In addition, the following sub-question is considered:

- Can forced cycle-consistency between a VQA model and image generation model improve the VQA model.

4 Methods

Ideally, the full architecture would be trained end to end, but this will not be viable as the training would require an immense amount of resources to finish in a reasonable time. Therefore, the VQA and image generation models will be pre-trained on the same dataset. The pre-trained models will be used in the full architecture to further train the VQA model keeping the others static. The performance of the VQA model will be measured before and after being trained in the full architecture on a hold-out test set. It will be measured through an accuracy score in general and on different question types. The performance of the image generation model will be evaluated quantitatively through the inception score (Salimans et al., 2016). This is a measure of the visual quality and is also used in (Zhang et al., 2017). The final experiments will be done on the Visual Genome dataset (Krishna et al., 2017). Intermediate experiments might be done with smaller datasets, to speed up the training and testing.

5 Scientific Relevance for Artificial Intelligence

VQA is one of the recently proposed multi-domain tasks that is a big challenge for current AI research. A model that could answer any question about an image correctly would make the perception of AI being truly intelligent to get ever closer. Unfortunately, while achieving promising results, VQA research is still far from achieving this. Current models can make mistakes that seem completely silly to a human. Due to the black-box nature of deep learning, it is hard for researchers to follow the decision-making process of the model to solve these mistakes. explainability of the model could help with finding this out and consequently solving the problem. This study aims to improve VQA's explainability through the novel approach of image generation.

Once VQA models start getting used in applications outside of research, people must have trust in the models. Being able to explain the thought process of the model would help in generating this trust. Even if the model makes a mistake, being able to explain the mistake instead of having to rely on the black box explanation, is much more trustworthy. VQA models could for example be used in visual aid systems, where the user can ask the system questions about a real-life scene. Especially in such a case, trust in the model is very important. These reasons make exploring explainability in VQA scientifically relevant.

6 Planning

The preliminary research and writing of the proposal happened in between mid-August and mid-October. The main research will adhere to the following planning:

October & November

- Familiarize with peregrine and PyTorch
- Implement image generation model

December

- Test and train image generation model on a simple dataset
- Implement a baseline VQA model

January

- Test and train image generation on a VQA dataset
- Test and train VQA model

February

- Implement attention mechanism into VQA model
- Implement the full architecture

March

- Run experiments on the full architecture

April

- Finalize experiments
- Write thesis

May

- Write result section

June

- Hand in first version of thesis

July

- Hand in final version of thesis
- Buffer

7 Resources and Support

The peregrine supercomputer environment provided by the RUG will be used to perform most of the training and experiments. The author owns a PC with a GPU that can also be used for these. Due to the current corona situation, all work will be done from home and meetings through a video-calling platform.

References

Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., and Zhang, L. (2018). Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering. Technical report.

Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C. L., and Parikh, D. (2015). VQA: Visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision*, volume 2015 Inter, pages 2425–2433.

Chen, Q. and Koltun, V. (2017). Photographic image synthesis with cascaded refinement networks. In *Proceedings of the IEEE international conference on computer vision*, pages 1511–1520.

Ghosh, S., Burachas, G., Ray, A., and Ziskind, A. (2019). Generating Natural Language Explanations for Visual Question Answering using Scene Graphs and Visual Attention. Technical report.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680.

Goyal, Y., Khot, T., Summers-Stay, D., Batra, D., and Parikh, D. (2017). Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6904–6913.

Johnson, J., Gupta, A., and Fei-Fei, L. (2018). Image Generation from Scene Graphs. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 1219–1228.

Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.-J., Shamma, D. A., et al. (2017). Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123(1):32–73.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. (2014). Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer.

Mirza, M. and Osindero, S. (2014). Conditional Generative Adversarial Nets. Technical report.

Qiao, T., Zhang, J., Xu, D., and Tao, D. (2019). Mirrorgan: Learning text-to-image generation by redescription. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2019-June, pages 1505–1514.

Ren, S., He, K., Girshick, R., and Sun, J. (2015). Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99.

Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., and Chen, X. (2016). Improved techniques for training gans. In *Advances in neural information processing systems*, pages 2234–2242.

Shah, M., Chen, X., Rohrbach, M., and Parikh, D. (2019). Cycle-consistency for robust visual question answering. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2019-June, pages 6642–6651.

Simonyan, K. and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Zhang, H., Xu, T., Li, H., Zhang, S., Wang, X., Huang, X., and Metaxas, D. N. (2017). Stack-gan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 5907–5915.