Project-Related Paper Report

Zeyuan Hu, iamzeyuanhu@utexas.edu

EID:zh4378 Spring 2018

Abstract

In this writeup, we first summarize a CVPR 2018 paper "Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering" [1]. We then discuss the limitation of the paper and extension of the paper's model in our project.

1 Summarization

For visual question answering (VQA) and image captioning, top-down visual attention mechanisms have been widely used. Usually, the attention is trained to predict the weight for each spatial location in the Convolutional Neural Network (CNN) output and then the model incoporates the weights and the representation of VQA questions into a recurrent neural network (RNN) to generate the answers for the questions. However, the authors think that this top-down approach fails to consider how the image regions that are subject to attention are determined. Specifically, the authors argue that attention models should not be agnostic to the content of the image: attention models should operate on objects in the image instead of on CNN features that correspond to a uniform grid of equally-sized image regions. The authors think that the top-down attention mechanism cannot attain both coarse and fine levels of details due to the fixed number of image regions. In addition, they find it is hard for top-down attention model to detect the objects that are poorly aligned to the equally-sized image regions and bind visual concepts with those objects.

The authors propse a bottom-up attention mechanism to fix those issues. Specifically, the bottom-up mechanism, implemented using an object detection model Faster R-CNN [2], proposes a set of salient image regions, with each region represented by a pooled convolutional feature vector. The authors use the combination of top-down and bottom-up attention mechanism in both captioning model and VQA model. For captioning model, the authors use two layer LSTMs with both the partially-generated captions and the mean-pooled image features proposed by Faster R-CNN as input. For VQA model, the authors implement a deep neural network with joint embedding of the question and the image features, which is followed by a prediction of regression of scores over a set of candidate answers.

Authors' Up-Down model provides a significant improvement over the best ResNet baseline across all question types in 2017 VQA challenge. In addition, the model also achieves a state-of-the-art results on MSCOCO test server.

2 Discussion

The paper uses two attention mechanisms in VQA task. Bottom-up attention weights the importance of salient image regions (i.e., objects in image) without considering the task-specific context (e.g., questions in VQA). Top-down attention takes the task-specific context into account and re-weights the set of image regions proposed by bottom-up attention model before producing the answers. In our project, we build our model based on the bottom-up attention mechanism proposed in the paper to have all the benefits mentioned in the previous section.

The paper's main contribution is the integration of bottom-up and top-down attention models. But, the authors consider image captioning and VQA as two separate tasks. However, we think that two tasks are closely related to each other: captions provide hint for the questions and questions provide more context for captioning. Thus, we propose to solve two tasks simultaneously with one model. There are several limitations in the paper when they consider two tasks separately. For VQA task, for example, answers can be easily learned from training set questions without looking at image features at all (e.g., answers to most question which starts with 'is there? are 'yes'). The paper does not have a detailed error analysis of the model each question category in the VQA v2.0 dataset (i.e., Yes/No, Number, Other). Thus, in our project, we want to make sure that both language features and image features are considered equally before answering the questions. In addition, how the system reaches the answer is unaddressed in the paper. In our project, we use the question specific captions that are generated from both question and image features as explanations to the VQA answers.

For the image captioning task, since the VQA model and captioning model are separate in the paper, authors' captioning model cannot integrate both the questions and the answers from VQA, which may be helpful for the captioning task. Furthermore, the impact of bottom-up attention mechanism on the captioning is not fully explored in the paper. For example, captioning model may easily focus on the most significant visual content and ignores others due to the bottom-up attention model, which may lead to a simple and short caption that reflects one object.

References

[1] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang, "Bottom-up and top-down attention for image captioning and visual question answering," *arXiv preprint arXiv:1707.07998*, 2017.

[2] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'15, (Cambridge, MA, USA), pp. 91–99, MIT Press, 2015.