Graph-Structured Representations for Visual Question Answering

Chaonan Song

May 13, 2018

1 Related work

Since Antol *et al.* [2] seminal paper, the task of visual quizzes has received more attention from people. The most recent method is to combine images with deep neural networks. The image is preprocessed by a Convolutional Neural Network (CN-N) for image classification, and intermediate features are extracted to describe the image. This problem is usually generated by a Recurrent Neural Network (RNN), such as LSTM, which produces a vector that represents a sequence of words. These two representations are mapped to the joint space through one or more nonlinear layers. Then input them into the classifier of the output vocabulary and predict the final answer. Recent papers on VQA have proposed improvements and changes to this basic idea. See [9]. A great progress of basic method is using attention mechanism [1, 3, 6, 11–13]. It models the interaction between the actual content of the input image. Visual input is usually represented as a spatial feature map, not as an overall image width feature. Their method uses a weighted operation, which is the same as their chart representation, and they match it as a sub map. Graphic nodes that represent question words are associated with graphic nodes that represent scene objects. Similarly, Lu*et al.* [8] common concern model determines the attention weights for image regions and problem words. Their best method is to proceed in order. First is the problem-guided visual attention, then the image-guided problem attention.

Recall rate at 100 FP on FDDB		
They describe how to use graph repre sentations of scene and question for VQA.	They showed how to use off-the-shelf language parsing tools to generate graphical representations of grammatical relationships.	They trained the proposed model on the VQA abstract scene benchmark [2] and improved accuracy and accuracy.

Table 1: Contribution of this article

In their case, they found that a single version of the joint version performed better. Their methods have some contribution in Table. 1. A major contribution of their model is the use of input scenarios and structured representations of the problem. This contrasts with typical CNN and RNN models. The Dynamic Memory Network (DMN) applied to VQA in [10] also maintains a set representation of the input. Similar with their model, DMN modeling inputs the interaction between different parts. Their method can also take as input the features of any relationship between input parts (edge features in the graph). This specifically allows the use of syntactic dependencies between words after the pre-parsing problem. The neural network in the figure has recently received much attention [4, 5, 7]. The method most similar to them is the gating sequence neural network [15], which associates a gating cycle unit (GRU [**?**] with each node and updates each node by iteratively passing messages between neighbors. Their formula similarly merges information from neighbors into each node feature through multiple iterations, but using the attention mechanism within the recurring unit does not find any advantage.

References

[1] Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Dan Klein. Neural module networks. *In Proc. IEEE Conf. Comp. Vis. Patt. Recogn*, 2016.

[2] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. *In Proc. IEEE Int. Conf. Comp. Vis*, 2015.

[3] Kan Chen, Jiang Wang, Liang Chieh Chen, Haoyuan Gao, Wei Xu, and Ram Nevatia. Abc-cnn: An attention based convolutional neural network for visual question answering. *Computer Science*, 2015.

[4] David Duvenaud, Dougal Maclaurin, Jorge Aguilera-Iparraguirre, Rafael Gmez-Bombarelli, Timothy Hirzel, Aln Aspuru-Guzik, and Ryan P Adams. Convolutional networks on graphs for learning molecular fingerprints. pages 2224–2232, 2015.

[5] Ashesh Jain, Amir R. Zamir, Silvio Savarese, and Ashutosh Saxena. Structural-rnn: Deep learning on spatio-temporal graphs. pages 5308–5317, 2015.

[6] Aiwen Jiang, Fang Wang, Fatih Porikli, and Y-i Li. Compositional memory for visual question answering. *Computer Science*, 2015.

[7] Yujia Li, Daniel Tarlow, Marc Brockschmidt, and Richard Zemel. Gated graph sequence neural networks. *Computer Science*, 2015.

[8] Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. Hierarchical question-image co-attention for visual question answering. 2016.

[9] Qi Wu, Damien Teney, Peng Wang, Chunhua Shen, Anthony Dick, and Anton Van Den Hengel. Visual question answering: A survey of methods and datasets. *Computer Vision Image Understanding*, 2017.

[10] Caiming Xiong, Stephen Merity, and Richard Socher. Dynamic memory networks for visual and textual question answering. 2016.

[11] Huijuan Xu and Kate Saenko. Ask, attend and answer: Exploring question-guided spatial attention for visual question answering. In *European Conference on Computer Vision*, pages 451–466, 2016.

[12] Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, and Alex Smola. Stacked attention networks for image question answering. pages 21–29, 2015.

[13] Yuke Zhu, Oliver Groth, Michael Bernstein, and Fei Fei Li. Visual7w: Grounded question answering in images. In *Computer Vision and Pattern Recognition*, pages 4995–5004, 2016.