Image Question Answering using Convolutional Neural Network with Dynamic Parameter Prediction

Liangjie Cao

Jun 22, 2018

Abstract

The authors tackle image question answering (ImageQA) problem by learning a convolutional neural network (CNN) with a dynamic parameter layer whose weights are determined adaptively based on questions. For the adaptive parameter prediction, they employ a separate parameter prediction network, which consists of gated recurrent unit (GRU) taking a question as its input and a fully-connected layer generating a set of candidate weights as its output. However, it is challenging to construct a parameter prediction network for a large number of parameters in the fully-connected dynamic parameter layer of the CNN. We reduce the complexity of this problem by incorporating a hashing technique, where the candidate weights given by the parameter prediction network are selected using a predefined hash function to determine individual weights in the dynamic parameter layer. The proposed networkłjoint network with the CNN for ImageQA and the parameter prediction networkł is trained end-to-end through back-propagation, where its weights are initialized using a pre-trained CNN and GRU. The proposed algorithm illustrates the state-of-the-art performance on all available public ImageQA benchmarks.

1. Introduction

One of the ultimate goals in computer vision is holistic scene understanding [10], which requires a system to capture various kinds of information such as objects, actions, events, scene, atmosphere, and their relations in many different levels of semantics. Although significant progress on various recognition tasks [2, 3, 5, 7, 8, 9, 11] has been made in recent years, these works focus only on solving relatively simple recognition problems in controlled settings, where each dataset consists of concepts with similar level of understanding (e.g. object, scene, bird species, face identity, action, texture etc.). There have been less efforts made on solving various recognition problems simultaneously, which is more complex and realistic, even though this is a crucial step toward holistic scene understanding.

Figure 1. Sample images and questions in VQA dataset![1]. Each question requires different type and/or level of understanding of the corresponding input image to find correct answers

Image question answering (ImageQA) [1] aims to solve the holistic scene understanding problem by proposing a task unifying various recognition problems. ImageQA is a task automatically answering the questions about an input image as illustrated in Figure 1. The critical challenge of this problem is that different questions require different types and levels of understanding of an image to find correct answers. For example, to answer the question like "how is the weather?" they need to perform classification on multiple choices related to weather, while they should decide between yes and no for the question like "is this picture taken during the day?" For this reason, not only the performance on a single recognition task but also the capability to select a proper task is important to solve ImageQA problem.

Contrary to the existing approaches, the authors define a different recognition task depending on a question. To realize this idea, they propose a deep CNN with a dynamic parameter layer whose weights are determined adaptively based on questions. They claim that a single deep CNN architecture can take care of various tasks by allowing adaptive weight assignment in the dynamic parameter layer. For the adaptive parameter prediction, they employ a parameter

prediction network, which consists of gated recurrent units (GRU) taking a question as its input and a fully-connected layer generating a set of candidate weights for the dynamic parameter layer. The entire network including the CNN for ImageQA and the parameter prediction network is trained end-to-end through back-propagation, where its weights are initialized using pre-trained CNN and GRU.

2. Problem Formulation

ImageQA systems predict the best answer $\hat{a}$ given an image I and a question q. Conventional approaches [4, 6] typically construct a joint feature vector based on two inputs I and q and solve a classification problem for ImageQA using the following equation:

$$\hat{a} = \arg\min_{a \in \Omega} p(a \mid I, q; \theta) \tag{1}$$

where Ω is a set of all possible answers and θ is a vector for the parameters in the network. On the contrary, they use the question to predict weights in the classifier and solve the problem. They find the solution by

$$\hat{a} = \arg\max_{a \in \Omega} p(a \mid I; \theta_s, \theta_d(q)) \tag{2}$$

where θ_s and $\theta_d(q)$ denote static and dynamic parameters, respectively. Note that the values of $\theta_d(q)$ are determined by the question q.

References

[1] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh. Vqa: Visual question answering. In ICCV*, 2015. 1*

[2] M. Cimpoi, S. Maji, I. Kokkinos, S. Mohamed, and A. Vedaldi. Describing textures in the wild. In CVPR*, 2014. 1*

[3] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell. DeCAF: a deep convolutional activation feature for generic visual recognition. In ICMI*, 2014. 1*

[4] L. Ma, Z. Lu, and H. Li. Learning to answer questions from image using convolutional neural network. In AAAI*, 2016. 2*

[5] M. Oquab, L. Bottou, I. Laptev, and J. Sivic. Learning and transferring mid-level image representations using convolutional neural networks. In CVPR*, 2014. 1*

[6] M. Ren, R. Kiros, and R. S. Zemel. Exploring models and data for image question answering. In NIPS*, 2015. 2*

[7] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In ICLR*, 2015. 1*

[8] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and b. y. Rabinovich, Andrew. Going deeper with convolutions. 1

[9] Y. Taigman, M. Yang, Marc, and L. Wolf. DeepFace: Closing the gap to human-level performance in face verification. In CVPR*, 2014. 1*

[10] J. Yao, S. Fidler, and R. Urtasun. Describing the scene as a whole: Joint object detection, scene classification and semantic segmentation. In CVPR*, 2012. 1*

[11] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva. Learning deep features for scene recognition using places database. In NIPS*, 2014. 1*