Image Question Answering using Convolutional Neural Network with Dynamic Parameter Prediction II

Liangjie Cao

Jun 24, 2018

1. Related Work

Deep learning based approaches demonstrate competitive performances in ImageQA [1, 8, 9, 4, 5]. Most approaches based on deep learning commonly use CNNs to extract features from image while they use different strategies to handle question sentences. Some algorithms employ embedding of joint features based on image and question [1, 5, 8]. However, learning a softmax classifier on the simple joint features-concatenation of CNN-based image features and continuous bag-of-words representation of a questionłperforms better than LSTM-based embedding on COCO-QA [9] dataset. Another line of research is to utilize CNNs for feature extraction from both image and question and combine the two features [6]; this approach demonstrates impressive performance on DAQUAR [7] dataset by allowing to fine-tune the whole parameters.

The prediction of the weight parameters in deep neural networks has been explored in [2] in the context of zeroshot learning. To perform classification of unseen classes, it trains a multi-layer perceptron to predict a binary classifier for class-specific description in text. However, this method is not directly applicable to ImageQA since finding solutions based on the combination of question and answer is a more complex problem than the one discussed in [2], and ImageQA involves a significantly larger set of candidate answers, which requires much more parameters than the binary classification case. Recently, a parameter reduction technique based on a hashing trick is proposed by Chen *et al.* [3] to fit a large neural network in a limited memory budget. However, applying this technique to the dynamic prediction of parameters in deep neural networks is not attempted yet to our knowledge.

2. Network Architecture

Figure 1 illustrates the overall architecture of the proposed algorithm. The network is composed of two subnetworks: classification network and parameter prediction network. The classification network is a CNN. One of the fully-connected layers in the CNN is the dynamic parameter layer, and the weights in the layer are determined adaptively by the parameter prediction network. The parameter prediction network has GRU cells and a fully-connected layer. It takes a question as its input, and generates a real-valued vector, which corresponds to candidate weights for the dynamic parameter layer in the classification network. Given an image and a question, their algorithm estimates the weights in the dynamic parameter layer through hashing with the candidate weights btained from the parameter prediction network. Then, it feeds the input image to the classification network to obtain the final answer. More details of the proposed network are discussed in the following subsections.

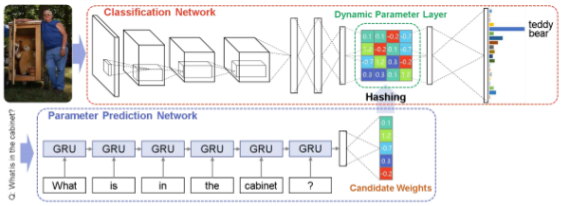

Figure 1. Overall architecture of the proposed Dynamic Parameter Prediction network (DPPnet), which is composed of the classification network and the parameter prediction network. The weights in the dynamic parameter layer are mapped by a hashing trick from the candidate weights obtained from the parameter prediction network

3. Classification Network

They put the dynamic parameter layer in the second last fully-connected layer instead of the classification layer because it involves the smallest number of parameters. As the number of parameters in the classification layer increases in proportion to the number of possible answers, predicting the weights for the classification layer may not be a good option to general ImageQA problems in terms of scalability. Their choice for the dynamic parameter layer can be interpreted as follows. By fixing the classification layer while adapting the immediately preceding layer, we

obtain the task-independent semantic embedding of all possible answers and use the representation of an input embedded in the answer space to solve an ImageQA problem. Therefore, the relationships of the answers globally learned from all recognition tasks can help solve new ones involving unseen classes, especially in multiple choice questions. For example, when not the exact ground-truth word (*e.g.*, kitten) but similar words (*e.g.*, cat and kitty) are shown at training time, the network can still predict the close answers (*e.g.*, kitten) based on the globally learned answer embedding. Even though they could also exploit the benefit of answer embedding based on the relations among answers to de?ne a loss function, they leave it as their future work.

4. Parameter Prediction Networ

Let ω_1,...,ω_T be the words in a question q, where T is the number of words in the question. In each time step t, given the embedded vector x_t for a word ω_t, the GRU encoder updates its hidden state at time t, denoted by h_t, using the following equations:

$$r_t = \sigma(W_\tau x_t + U_r h_{t-1}) \tag{1}$$

$$z_t = \sigma(W_z x_t + U_z h_{t-1}) \tag{2}$$

$$\bar{h}_t = tanh(W_h x_t + U_h(r_t \odot h_{t-1})) \tag{3}$$

$$h_t = (1 - z_t) \odot h_{t-1} + z_t \odot \bar{h}_t \tag{4}$$

where r_t and z_t respectively denote the reset and update gates at time t, and $\bar{h}_t$ is candidate activation at time t. In addition, $\odot$ indicates element-wise multiplication operator and $\sigma()$ is a sigmoid function. Note that the coefficient matrices related to GRU such as W_r, W_z, W_h, U_r, U_z, and U_h are learned by our training algorithm.

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] J. L. Ba, K. Swersky, S. Fidler, and R. Salakhutdinov. Predicting deep zero-shot convolutional neural networks using textual descriptions. In *ICCV*, 2016. 1

[3] W. Chen, S. Tyree, S. Tyree, K. Q. Weinberger, and Y. Chen. Compressing neural networks with the hashing trick. In *ICCM*, 2015. 1

[4] J. Deng, W. Dong, R. Socher, L. J. Li, K. Li, and F. F. Li. ImageNet: a large-scale hierarchical image database. In *CVPR*, 2009. 1

[5] H. Gao, J. Mao, J. Zhou, Z. Huang, L. Wang, and W. Xu. Are you talking to a machine? dataset and methods for multilingual image question answerin. In *NIPS*, 2015. 1

[6] L. Ma, Z. Lu, and H. Li. Learning to answer questions from image using convolutional neural network. In *AAAI*, 2016. 1

[7] M. Malinowski and M. Fritz. A multi-world approach to question answering about real-world scenes based on uncertain input. In *NIPS*, 2014. 1

[8] M. Malinowski, M. Rohrbach, and M. Fritz. Ask your neurons: a neural-based approach to answering questions about images. In *ICCV*, 2015. 1

[9] M. Ren, R. Kiros, and R. S. Zemel. Exploring models and data for image question answering. In *NIPS*, 2015. 1