Semantic representation for visual reasoning

Xubin Ni[1], *Lirong* Yin[2], *Xiaobing* Chen[1], *Shan* Liu[1,3], *Bo* Yang[1], and *Wenfeng* Zheng[1,*]

[1]School of Automation, University of Electronic Science and Technology of China, Chengdu, 610054, China
[2]Geographical & Sustainability Sciences Department, University of Iowa, Iowa City, IA, 52242, USA
[3]Department of Modelling, Simulation, and Visualization Engineering, Old Dominion University, Norfolk, VA 23529, USA

Abstract. In the field of visual reasoning, image features are widely used as the input of neural networks to get answers. However, image features are too redundant to learn accurate characterizations for regular networks. While in human reasoning, abstract description is usually constructed to avoid irrelevant details. Inspired by this, a higher-level representation named semantic representation is introduced in this paper to make visual reasoning more efficient. The idea of the Gram matrix used in the neural style transfer research is transferred here to build a relation matrix which enables the related information between objects to be better represented. The model using semantic representation as input outperforms the same model using image features as input which verifies that more accurate results can be obtained through the introduction of high-level semantic representation in the field of visual reasoning.

1 Introduction

Visual Question Answering (VQA) is a technology that combines natural language processing with digital image processing. The general process for solving a VQA problem is to take the image and the corresponding question in natural language as input and finally get the answer. If the question involves reasoning, it is called visual reasoning. The issues studied by visual reasoning are similar to VQA but may require more interdependent inference steps to solve the problem.

The models used in VQA can be divided into classical machine learning models and deep learning models. Most classical machine learning models are based on Bayesian theory. [1] proposed a Bayesian framework for VQA, predicting the type of answer to a question and using it to generate an answer. [2] combined the semantic tree obtained from the semantic analyser and image to construct the SWQA model to predict the corresponding answer. After CLEVR dataset was proposed [3], those classical machine learning models performed poorly on this dataset compared with some deep learning models [4-7]. The workflow of the existing deep learning models can be roughly divided into three steps: extracting features from the question sentence, extracting features from the image, and combining the question embedding and image features to generate answers. For question

* Corresponding author: wenfeng.zheng.cn@gmail.com

embedding, techniques such as Long Short-Term Memory (LSTM) can be used [8, 9]. Regarding image features, it is generally considered to extract image features using a convolutional neural network (CNN). The iBOWIMG model used a pre-trained GoogleNet image classification model to extract image features, using the word embedding of each word in the question as a text feature. After stitching the image features and text features, the answer was obtained by Softmax regression [4]. [5, 6] continuously generated a neural network for each image and problem selected from various problem-based sub-modules and combined them to generate a neural network. The relational network made it possible to grasp the key of relational reasoning by constraining the structure of the function and obtained the state-of-art result [7].

Although deep learning models have made significant progress over classical machine learning models, it still has a large gap with the level of human reasoning on complex issues. In this research it is planned to use the semantic representation of the image as input in the visual reasoning task, instead of directly using the pixel or the image features extracted by CNN as input, to explore whether the result of introducing high-level semantic representation can be better. If better results can be gained, this idea can be transferred to other areas of computer vision, even other areas of deep learning.

2 Method

2.1 Dataset：Enhanced Sort-of-CLEVR(ESOC)

Most current visual reasoning works are based on the CLEVR dataset [3]. To simplify the image processing part and the natural language understanding part, the ESOC dataset based on Sort-of-Clevr dataset [7] was built in this research. ESOC dataset is similar to Sort-of-Clevr dataset but the scenes of ESOC dataset are more detailed and complex.

ESOC contains images, several questions for the image and corresponding answers. Each image contains several 2D geometric objects. The property of each object is randomly chosen from Figure 1. The questions are divided into relational questions (RQ) and non-relational questions (NRQ). The non-relational question involves only one object in the image, querying its shape, position or size. The relational question involves the positional relationship between two objects. The entire dataset is divided into three groups which are 2-shapes dataset, 4-shapes dataset, and 6-shapes dataset to test the performance of the model on different complexity scenes. N-shapes means that there are n objects in each scene.

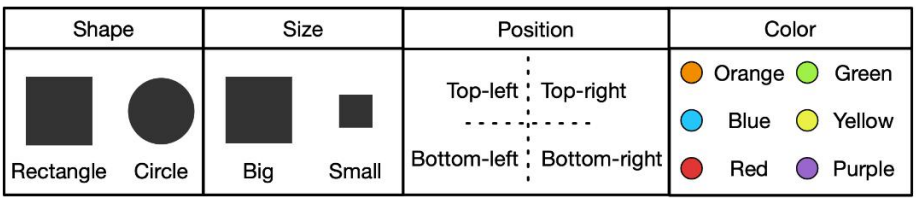

Fig. 1. The elemental composition of the ESOC dataset.

2.2 Process of extracting semantic representation

Assuming that the input image is x, the question for the image is q, the answer is a. The purpose of visual reasoning task is to find a system H to get the answer $a = H(x,q)$. In this research, a semantic network capable of characterizing enough information of the image is

used as input rather than image features extracted by CNN. The process is shown in Figure 2. Suppose O is the set of objects contained in the image and $O = \{O_1,O_2,\cdots,O_n\}$. Image segmentation technology is needed for extracting the collection O. For simple images, it can be solved according to edge segmentation. If the content of the dataset is a complex image, such as CLEVR dataset, which contains 3D objects that may overlap each other, instance segmentation techniques such as Mask R-CNN [10] and SSD [11] may be employed. For each object, the set of its attributes is $P = \{P_1,P_2,\cdots,P_m\}$. In this research, these attributes may include shape, size, color, and location. Modules M to extract these properties can be designed separately, where $M = \{M_1,M_2,\cdots,M_m\}$. The module $M_i \in M$ for extracting image semantic representation has no fixed requirements and is selected according to specific tasks. For example, the shape of the target object can be obtained by CNN. Assuming that the value of the attribute P_i is V_i. The value V_i of the attribute P_i of the object O_k can be obtained by $V_i = M_i(O_k)$. Thus, a set of triples describing the entire image can be obtained which is $S = \{S_{O_1},S_{O_2},\cdots,S_{O_n}\}$, where

$$S_{O_i} = \{(O_i,P_1,V_1),(O_i,P_2,V_2),\cdots,(O_i,P_m,V_m)\} \quad (1)$$

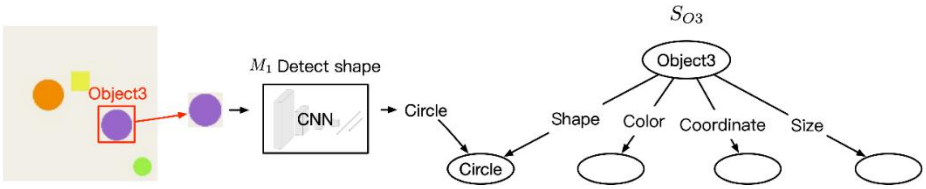

Fig. 2. The baseline of extracting a semantic net form the image.

Natural language cannot be understood by computers directly, so questions need to be pre-processed first. It can be done using LSTM, like the method in machine translation called Seq2Seq [12], to map the question to a vector space. Suppose the question is q, the encoder is L, and the question embedding is characterized by $Q = L(q)$. Finally, $a = H(S,Q)$.

Each sample in ESOC dataset contains (x,S,Q,a), where x is the original image; S is the semantic representation of the original image; Q is the question embedding; a is the ground-truth answer. (x,Q,a) and (S,Q,a) were trained in the neural network of the same architecture to compare the results of the two. Adam optimizer was used for gradient descent training and dropout mechanism was used to prevent overfitting [13]. The termination of training was judged by the early stopping method [14].

2.3 Relation matrix

At the beginning of the research, the way to organize semantic representation was to directly add the semantic representation of each object to a list to form a multidimensional vector. However, in practical experiments, such effects were found to be limited. The probable reason is that the network only captures the individual information of each object, but does not learn a way to characterize the relative relationship between different objects. A similar situation was also found in the study of style transfer. The visual style is an overall intuitive perception. When comparing the style differences between the two images, it is inefficient to take the feature layers of the two images directly because each feature map can only reflect some of its own rather than the overall feature. Similarly, the semantic vectors here were independent of each other, and the model did not have a good understanding of the relationship between them and the relationship between the individual and the whole scene. In [15], a Gram matrix of feature layers was constructed to describe

the style of an image. The reason why the Gram matrix is valid in that research is that it can be used to measure the characteristics of each dimension and the relationship between different dimensions. In the multi-scale matrix obtained after the inner product, the diagonal elements provide information about the different feature maps themselves, and the remaining elements provide relevant information between the different feature maps. Such a matrix can not only reflect the characteristics of the corresponding image but also reflect the closeness of different features.

Using a similar idea, a relation matrix was constructed to describe each object's own information and the relationship between different objects. Suppose that the semantic representations of object O_1 and object O_2 are S_{O_1} and S_{O_2}, and the question embedding is Q. The network F consisting of multiple fully connected layers is used to extract the relationship between O_1 and O_2 which is $F(S_{O_1},S_{O_2},Q)$. Assuming that there are n objects in the graph, the size of the relationship matrix is $n \times n$, where the elements of the i_{th} row and the j_{th} column is $R(i,j) = F(S_{O_i}, S_{O_j}, Q)$. The relation matrix is:

$$R = \begin{bmatrix} R(1,1) & R(1,2) & \cdots & R(1,n) \\ R(2,1) & R(2,2) & \cdots & R(2,n) \\ \vdots & \vdots & \ddots & \vdots \\ R(n,1) & R(n,2) & \cdots & R(n,n) \end{bmatrix} \tag{2}$$

2.4 Model

The model of this research consists of two parts which are feature extraction layer and inference layer, as shown in Figure 3. Feature extraction mainly involves image features extraction, semantic representation extraction and question features extraction. The structure of the inference layer of the image-feature-based model and semantic-representation-based model is the same which is multi-layer fully connected neural networks. The convolutional neural network for extracting image features is composed of four layers of convolutional layers.

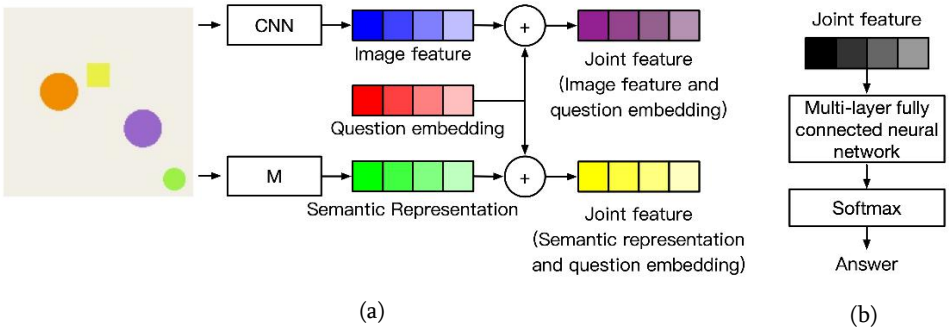

Fig. 3. Two parts of the model in this research. (a) Feature extraction layer. (b) Inference layer.

For image-feature-based model, the features of the image are first extracted by the convolutional layer. Then the image features and question embedding are jointly embedded as the input to inference layer, and the answer is finally obtained. For semantic-representation-based model, the feature extraction method is as described in section 2.2 and 2.3. After that, the semantic representation vector and the question embedding can be jointly embedded as the input to inference layer.

3 Result

The experiment was based on the ESOC dataset and the model described in section 2.2. The main task was to compare the difference in results with semantic representation as input and image features as input. At first, semantic vector was directly arranged into a large-size vector as input. The result is shown in Table 1. The accuracy promotion after replacing image features with semantic representation as input is shown in Figure 4.

Table 1. The accuracy (%) of image-feature-based (IF) models, semantic-representation-based (SR) models and semantic-representation-based models with relation matrix.

Model Type	**IF**			**SR**			**SR + relation matrix**		
Object numbers	**2**	**4**	**6**	**2**	**4**	**6**	**2**	**4**	**6**
Accuracy of NRQ	69.64	59.39	60.14	99.03	97.81	95.77	99.75	99.72	99.61
Accuracy of RQ	69.33	64.10	55.61	99.06	66.79	57.90	99.76	94.91	93.28

It can be seen that the accuracy is improved on three datasets after the semantic representation is used as the input. Among them, on the non-relational question, semantic-representation-based model has a significant effect on the improvement of accuracy, but on the relational question, the accuracy promotion of the 4-shapes dataset and 6-shapes dataset is not apparent, only about 4%. The reason why the accuracy of the 2-shapes dataset on the relational question is as high as that on the non-relational question is that when there are only two objects in the scene, the relational question and the non-relational question are equivalent.

In order to further improve the results, the relation matrix was used in the data pre-processing part, and other parts did not change. The results are shown in Table 1 and Figure 5. It can be found that after using the relation matrix, the accuracy of the relational question is also significantly improved compared to the image-feature-based model. For 4-shapes dataset, the accuracy promotion increases from 4.20% to 48.07%, and for the 6-shapes dataset it is from 4.12% to 67.74%. The improvement is undeniable. It can be observed that a simple model can complete complex reasoning tasks as long as the semantic representation is processed slightly.

Fig. 4. The promotion of accuracy after using semantic representation as input directly.

Fig. 5. The promotion of accuracy after using semantic representation with relation matrix.

4 Conclusion

This paper demonstrates how the semantic representation of an image can be used as the input to a visual reasoning system and verifies that changing the representation of the image can further improve system performance. After replacing the visual features with unprocessed semantic representation vectors, the accuracy of the model on non-relational questions was significantly improved, but the accuracy on the relational questions was only slightly improved. Then the semantic vector was pre-processed by constructing a relation matrix. This simple operation made the model have a 43% to 67% improvement on relational and non-relational questions. It can be seen that the effect of semantic

representation itself is not worse than image features. And semantic representation is simple and easy to carry out other processing. After this, experiment results can be further improved. In this experiment, the baseline model was fundamental, with only a convolutional neural network and a multi-layer fully connected layer. If the semantic representation is more complex and more specific in a particular field, it could be expected that the introduction of higher-level representation will make results more accurate, which is one of the possible future research directions.

The problem with the current work is that although the final accuracy is improved, the work of extracting semantic information requires manual intervention to select appropriate semantic features, which undoubtedly increases system complexity and additional time overhead compared to the end-to-end system. So, the next work will focus on how to make system extract useful semantic information automatically and reduce manual intervention.

Acknowledgments

Funds for International S&T Cooperation and Exchange R&D Project of Sichuan Province (Grant No. 2017HH0054).

References

1. Kafle, K. and C. Kanan. *Answer-Type Prediction for Visual Question Answering*. in *Computer Vision and Pattern Recognition*. (2016).
2. Malinowski, M. and M. Fritz. *A multi-world approach to question answering about real-world scenes based on uncertain input*. in *Advances in Neural Information Processing Systems*. (2014).
3. Johnson, J., et al., *CLEVR: A Diagnostic Dataset for Compositional Language and Elementary Visual Reasoning.* (2017).
4. Zhou, B., et al., *Simple Baseline for Visual Question Answering.* Computer Science, (2015).
5. Andreas, J., et al. *Neural Module Networks*. in *IEEE Conference on Computer Vision and Pattern Recognition*. (2016).
6. Johnson, J., et al., *Inferring and Executing Programs for Visual Reasoning.* (2017).
7. Santoro, A., et al., *A simple neural network module for relational reasoning.* (2017).
8. Sak, H., A. Senior, and F. Beaufays, *Long Short-Term Memory Based Recurrent Neural Network Architectures for Large Vocabulary Speech Recognition.* Computer Science, (2014): p. 338-342.
9. Andreas, J., et al., *Deep Compositional Question Answering with Neural Module Networks.* Computer Science, (2015). **27**: p. 55-56.
10. He, K., et al., *Mask R-CNN.* IEEE Transactions on Pattern Analysis & Machine Intelligence, (2017). **PP**(99): p. 1-1.
11. Liu, W., et al. *SSD: Single Shot MultiBox Detector*. in *European Conference on Computer Vision*. (2016).
12. Sutskever, I., O. Vinyals, and Q.V. Le, *Sequence to Sequence Learning with Neural Networks.* (2014). **4**: p. 3104-3112.
13. Srivastava, N., et al., *Dropout: a simple way to prevent neural networks from overfitting.* Journal of Machine Learning Research, (2014). **15**(1): p. 1929-1958.
14. Yao, Y., L. Rosasco, and A. Caponnetto, *On Early Stopping in Gradient Descent Learning.* Constructive Approximation, (2007). **26**(2): p. 289-315.
15. Gatys, L.A., A.S. Ecker, and M. Bethge, *A Neural Algorithm of Artistic Style.* Computer Science, (2015).