Research on Visual Reasoning Model Based on Active Learning

Hao Ma[1,a], Xuefeng Zhu[1,b] and Yifeng Zheng[1,c]

[1]Beijing Key Lab of Petroleum Data Mining, China University of Petroleum, Beijing
a. 547860313@qq.com, b. xuefeng.zhu@cup.edu.cn, c. zyf@mnnu.edu.cn

Keywords: Visual reasoning, active learning, overfitting, generalization ability, labeling effort.

Abstract: Nowadays, the visual reasoning model has been able to answer complex questions that cannot be answered by the visual question answering model on the CLEVR dataset. However, most visual reasoning models require a large amount of data for strongly supervised learning, which is easy to increase the cost of data labeling. In addition, these approaches will lead to over-fitting, thereby reducing the generalization performance of the model. To solve the above problem, in this paper, we propose a novel model combined with active learning. It utilizes active learning to select the most informative and representative sample as the training data efficiently and accurately. Therefore, fewer samples can be employed to train a visual inference model to obtain higher accuracy and better generalization ability. The experimental results from three aspects show the effectiveness of the proposed approach with active learning for the visual reasoning model.

1. Introduction

With the development of single-modal deep learning, it has been widely employed in various fields, such as computer vision and natural language processing. Therefore, researchers increasingly hope to apply it to more complex scenes, for example, inference on daily visual input. However, the reasoning is a manifestation of human intelligence, which is a very complex task for artificial intelligence. In order to solve the above problems, a multi-modal visual question answering model was derived.

Early Visual Question Answering (VQA) models [1] tended to utilize bias in the data to answer questions, and many effective general deep learning models have been derived by using existing data sets. However, because the model cannot capture the complex underlying structure behind the problem, and thus it achieves poorly results on the visual inference data set (CLEVR data set). The illustrate of the problem is showing in Figure 1.

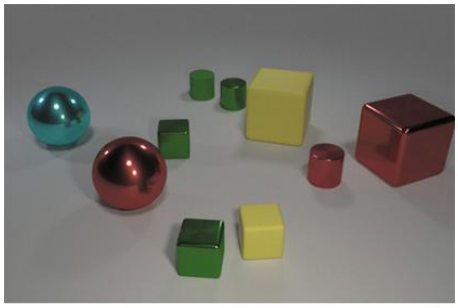

Figure 1: CELVR data set example.

To solve the above problem, researchers achieve the underlying reasoning process based on existing models to implement the visual reasoning model [2-6]. It should be noted that this process requires a large number of training samples for strongly supervised learning, which is expensive for data labeling, and leads to over-fitting of the model. For the CLEVR-Human dataset with more linguistic diversity, it cannot achieve excellent performance.

For this problem, in this paper, we propose a novel visual reasoning model with active learning and then evaluate it on the CLEVR and CLEVR-Human data sets to validate the effectiveness of models. The following tasks of visual reasoning are demonstrated to illustrate that active learning can help to enhance model accuracy and generalization performance:

- The active learning can help to enhance the accuracy of the model on the CLEVR data set.
- Through the fusion of active learning, it can effectively reduce the dependence on data and the cost of data labeling, so that fewer training samples are utilized to obtain the same accuracy as the baseline.
- Combined with active learning, while reducing the over-fitting of the model, it helps to improve the generalization performance of the model.

2. Model Introduction

To executing complex inference tasks successfully, it is necessary to introduce combinatorial inference into the model explicitly. In this paper, the visual inference model with active learning consists of two parts, a program generator, and an execution engine. The overall structure of the model is shown in Figure 2.

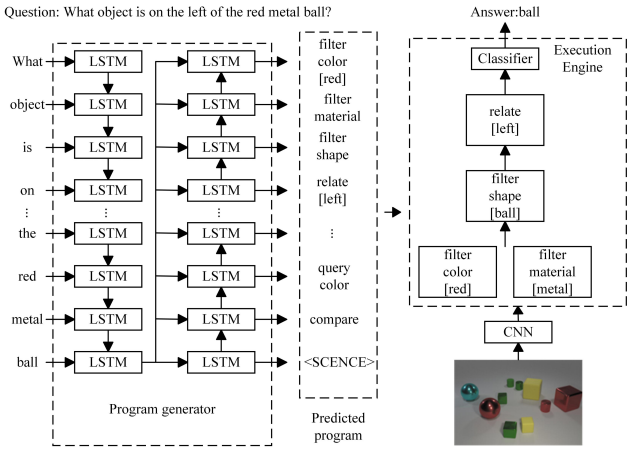

Figure 2: Model structure diagram.

The program generator adopts a sequence-to-sequence (Seq2Seq) structure to map the input questions to corresponding ground-truth programs. Meanwhile, the ground-truth programs are defined by the grammar, and the semantics defines their behavior. That is, the CLEVR data set stores semantic rules by employing pre-specifying the function set Φ composed of functions ϕ. For each function f, it has a fixed number of parameters $v_\phi \in \{1,2\}$. The valid program z, that is, the ground-truth program is stored in the form of a syntax tree, where each node contains a function $\phi(\phi \in \Phi)$, and the number of children of each node is the same as the number of parameters of ϕ.

The Seq2Seq model is divide into an encoder and a decoder, which composed of several long-short term memory (LSTM) units. This structure is more suitable for the above mapping method. The structure of the program generator is shown in Figure 3.

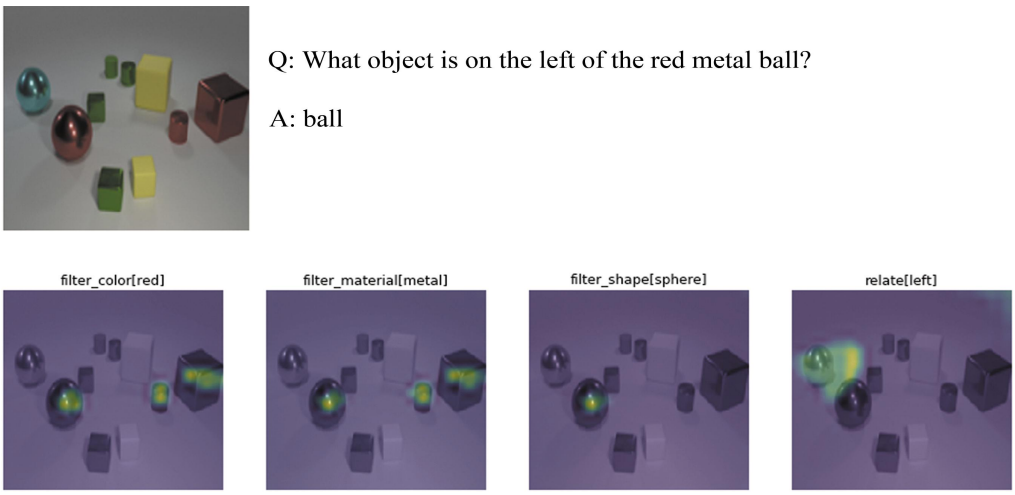

Figure 3: Model execution process visualization.

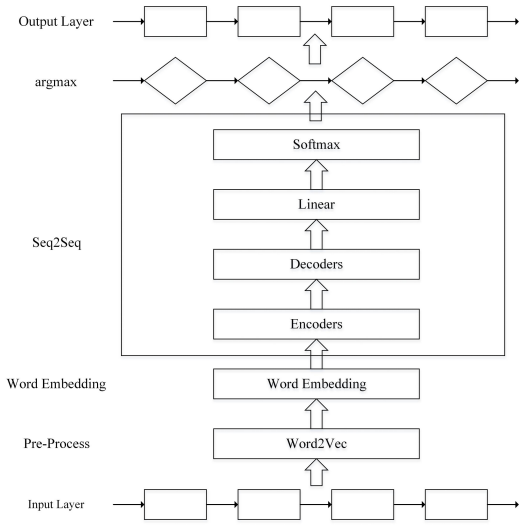

Figure 4: Program generator structure diagram.

In the training process of the program generator, it needs to convert the syntax tree with a non-sequential discretization structure into a function sequence by adopting the way of preorder traversal, and then map the problem to the function sequence.

Execution Engine is composed of neural module network [13] according to the corresponding program execution sequence ζ Specifically, given ζ the execution engine maps the function $\phi(\phi \in$

Φ) in ζ to the corresponding neural module network $\mu(\zeta)$ to generate a neural network μ_ϕ that can answer visual inference questions. Then $\mu(\zeta)$ accepts ζ and picture ξ as inputs, and outputs the prediction of the answer α, $\alpha = \phi(\xi,\zeta)$. It can be seen that, in the process of problem-solving, each intermediate output is interpretable; thus, the entire model has the ability to combine reasoning.

Since the intermediate output of each neural module network during the execution of the model is retained and visualized, the change of the model's attention (that is, the model's inference process) can be known, as shown in Figure 4. In the execution engine, each neural module network is a standard residual block, which includes two 3x3 convolutional layers. For unary operations, such as filter_color, filter_shape, employ a single residual block directly. For binary operations, such as query_size, equal_color, it needs to connect the two residual blocks in the channel dimension direction of the picture feature matrix. Before the execution engine processes the pictures, it is necessary to input the pictures to pre-trained ResNet-101 [16] on the ImageNet [17] website for feature extraction.

In the training phase, the two parts of the model adopt their training samples, respectively. And then, jointly trained (called fine-tune) is conducted. In other words, we utilize the output of the program generator as the input of the execution engine. In the process of fine-tune, reinforcement learning [15] is employed to update the parameters of both models simultaneously.

In recent years, visual reasoning has become a research hotspot. Johnson et al. [2] employed neural module networks to model the underlying reasoning process explicitly. Runtao Liu et al. [3] constructed a synthetic diagnostic data set based on expression understanding CLEVR-Ref+ and proposed a network modular, call IEP-Ref, which can capture combinability clearly. Sjoerd van Steenkiste et al. [4] studied whether the entangled representation is suitable for abstract inference tasks. Jiaxin Shi et al. [5] utilized the display neural module to solve complex visual reasoning tasks. Perez E et al. [6] presented a novel method with linear feature modulation to influence neural network calculation.

The above method performs well on both the training set and the test set, but only Johnson et al. [2] conducted generalization performance experiments on the CLEVR-Human data set. Therefore, we adopt it as a baseline to prove that the proposed approach with the active learning can help to reduce the over-fitting of the model and the labeling cost, and can enhance the generalization performance of the model effectively.

3. Active Learning Methods

3.1. Active Learning

Although, researchers have proposed various visual reasoning models, which have been verified on the CLEVR dataset. However, there is no research on the visual reasoning model of active learning. Most work on active learning mainly involves image classification, and only a few involve natural language processing. Active learning can provide a training sample selection method for the model that can obtain the same or even exceed the baseline performance using fewer labeled samples.

Wang et al. proposed an active learning method based on uncertainty sampling, using CNN for active learning of image classification [7]. Gal et al. proved that active learning based on Bayesian uncertainty sampling could achieve better results [8-9]. Samarth Sinha et al. proposed a pool-based semi-supervised active learning algorithm to learn the sampling mechanism implicitly in an adversarial manner [10]. Zhang et al. utilized CNN for active learning of sentence classification [11]. Akshay Krishnamurthy et al. designed an active learning algorithm for cost-sensitive multiclass classification [12]. The above-mentioned active learning methods mainly focus on image classification, and do not involve natural language processing and the Seq2Seq model too much.

Researches show that the accuracy of the program execution sequence is the key to whether the entire model can output the correct answer. It shows that the program generator is significant for the model. Therefore, in this paper, we merge the program generator with active learning. From [14], it can be seen that there are a variety of active learning methods to utilize. Therefore, we mainly consider uncertainty sampling. Three uncertainty sampling approaches are employed in the program generator to validate the advantages of the proposed model in this paper.

3.1.1. Least Confidence

The least confidence (LC) formula cannot be applied in the Seq2Seq model directly. Therefore, we sum the minimum confidence of each word in the sequence. Then, the average value is calculated as the minimum confidence of the current sequence and sorted in descending order. Finally, filter samples according to the pre-set threshold. The formulation is defined as follows.

$$\frac{1}{\mathrm{n}}\sum_{i=1}^{n}\left(1-\max_{y_1,\ldots,y_n} \mathrm{P}[y_1,\ldots,y_n|\{x_{ij}\}]\right) \tag{1}$$

where y_{n} represents the number of the n-th category, x_{ij} represents the j-th word in the i-th sequence.

In the program generator, to obtain the optimal global solution needs to consider all possible values, which is an NP-hard problem. Therefore, we adopt the greedy algorithm of sequence decoding to approximate the optimal solution.

3.1.2. Maximum Normalized Log-Probability

Researches have shown that LC methods tend to choose long sequence samples since long sequences have more labeling information. To address this problem, in this paper, we employ the LC-based Maximum Normalized Log-Probability (MNLP), defined as follows.

$$\max_{y_1,\ldots,y_n} \frac{1}{n}\sum_{i=1}^{n}\lg \mathrm{P}[y_i|y_1,\ldots,y_n,\{x_{ij}\}] \tag{2}$$

It takes the logarithm of the output classification probability and normalizes it, which can reduce the error caused by the LC tendency effectively.

3.1.3. Sequence Entropy

Sequence Entropy (SE) is utilized to filter samples. The entropy value is calculated from the output classification probability, thereby obtaining the amount of information contained in the sample. And then sort the results in descending order. The higher the entropy of the sequence information, the greater the degree of sample confusion, and the more difficult it is to classify. Finally, filter the samples according to the preset threshold. The formulation is defined as follows.

$$-\sum_{\hat{y}} \mathrm{P}(\hat{y}|x;\theta)\lg \mathrm{P}(\hat{y}|x;\theta) \tag{3}$$

In the input active learning algorithm, the matrix vector is defined as $\mathbf{A}$, and its specification is [64,27,44], where the first dimension represents the data batch size, the second dimension

represents the length of the semantic vector, the third dimension represents the number of classification categories. The above three algorithms all adopt double nested loops, the time complexity of all three algorithms is $O(n^2)$.

3.2. Training Method

In this paper, the active learning model is defined as follows:

$$\mathrm{M}=（\mathrm{L}, \mathrm{U}, \phi(\cdot), \mathrm{B}） \tag{4}$$

where L represents the labeled data set, U is the unlabeled data set, B denotes the query batch size, and $\phi(\cdot)$ is the query function which contains the active learning algorithm introduced in Section 3.1.

The following steps are adopted for active learning.

$$x_b^* = \arg\max_{x \in U} \phi(x) \tag{5}$$

$$L = L \cup \{x_b^*, label(x_b^*)\} \tag{6}$$

$$U = U - x_b^* \tag{7}$$

where x_b^* denotes the most worthwhile sample obtained by the query function by filtering the corresponding sample information.

The algorithm process is as follows: First, we randomly select 2% samples in U as training samples and train a new model as the warm-start model for the active learning process. Then, in the new round of training, input all the samples in U into the warm-start model, and enter the obtained sample classification probability into $\phi(\cdot)$ to obtain the amount of information that can measure whether the sample has a labeled value. The top 5% of samples is obtained by screening the amount of information. Finally, take this part of the sample from U (do not put it back), mark it, and put it in L as the active learning training sample. Each subsequent round of screening will take 5% of the sample from U, mark it, and put it in L, and test the accuracy of the model. If the accuracy meets the pre-set threshold, the loop ends, otherwise, repeat the above steps. It is worth noting that after completing the sample selection step, the entire model will be retained on L to alleviate the over-fitting.

4. Experiment

In order to verify the validity of the model, this paper mainly uses two data sets: CLEVR and CLEVR-Human, both of which are proposed by Johnson et al. [14]. CLEVR dataset is stored in the form of tuples (pictures, questions, answers, program execution sequences). The CLEVR-Human dataset does not have a corresponding ground-truth program, so the Form storage. The detail of datasets is shown in Table 1.

Table 1: Basic data set information.

Dataset	Number of pictures	Number of questions	Number of answers	Number of ground-truth programs
CLEVR	70,000	700,000	700,000	700,000
CLEVR-Human	17817	7202	7202	--

In addition, the model of Johnson et al. [2] is selected as a baseline to verify the performance of the proposed model from three aspects. The experimental of this paper is composed of 36 experiments, and each experiment repeats three times, and the mean is taken as the final experimental result to reduce the experimental error.

4.1.Model Performance Under the Same Amount of Data

In order to verify the performance of the visual inference model with active learning, we test the performance of our proposed model and the baseline on the number of training samples of 9k and 18k, respectively. Figure 5 shows the accuracy comparison of the program generator at 9k and 18k data, and Figure 6 shows the accuracy comparison of the model after fine-tune at 9k and 18k data.

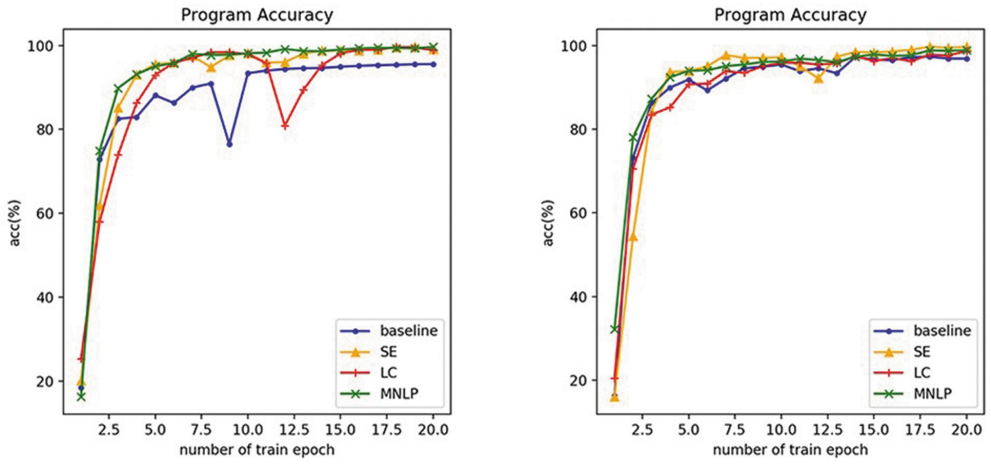

Figure 5: The accuracy comparison chart of the program generator of the model under the data volume of 9k (left) and 18k (right).

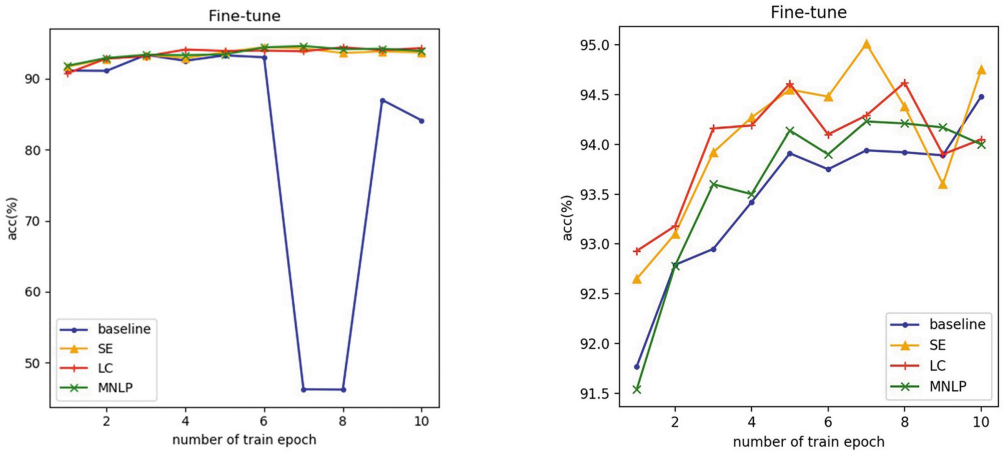

Figure 6: The accuracy comparison chart of the model after fine-tune under the data volume of 9k (left) and 18k (right).

From Figure 5, it can be seen that the accuracy of the proposed model is higher than the baseline. Meanwhile, we can observe the influence degree of different active learning algorithms in model accuracy. At 9k data, MNLP has the most significant improvement; at 18k data volume, SE has the best effect.

As can be seen from Figure 6, the accuracy of the model combined with active learning after fine-tune, is higher than the baseline. Meanwhile, we can observe the influence degree of different active learning algorithms in model accuracy. At 9k data, MNLP is slightly better than other active learning algorithms; at 18k data, SE is slightly better than other active learning algorithms. The reason is that the MNLP algorithm tends to employ the sum of the logarithm of the sample classification probability as the screening basis, while the SE algorithm tends to utilize the sample's confusion as to the screening basis. It can be seen that the samples selected by the MNLP algorithm are more representative, while the samples selected by the SE algorithm have more information.

In summary, the proposed model in this paper has better performance under the same number of training samples. The reason is that the training samples selected by active learning have higher training value.

4.2. The Amount of Data Required for the Model to Reach the Same Accuracy

To verify the effectiveness of active learning, the method introduced in section 3.2 is employed to train the model. Besides, compared our model with the baseline program generator and model fine-tune, respectively, the amount of data required to achieve the same accuracy.

Table 2: The amount of data required by the program generator to achieve the same accuracy as the baseline model.

Method	The amount of data		
	9,000	18,000	700,000
Baseline	100%	100%	100%
Ours-LC	60%	55%	60%
Ours-SE	60%	55%	55%
Ours-MNLP	60%	50%	50%

Table 2 shows the number of training samples required in the program generator to achieve baseline accuracy. From Table 2, we can see that that the mode with active learning requires 60% of the amount of data to achieve the same accuracy compared with baseline model on the 9k data; on the 18k data, only need 53.33% to achieve the same accuracy; on the 700k data, compared with baseline model, only 55% of the data used to achieve the same accuracy.

Table 3 shows the number of training samples required to achieve baseline accuracy after the model is fine-tuned. From Table 3, we can see that the model with active learning only needs 61.17% of the amount of data to achieve the same accuracy compared with the baseline model on 9k data volume; on the 18k data, only need 53% of the amount of data to achieve the same accuracy; on 700k data, compared with baseline model, only 56.67% of data required to achieve the same accuracy.

Table 3: The amount of data required to achieve the same accuracy as the baseline model after fine-tune.

Method	The amount of data		
	9,000	18,000	700,000
Baseline	100%	100%	100%
Ours-LC	65%	60%	65%
Ours-SE	60%	55%	55%
Ours-MNLP	60%	50%	50%

In summary, the model based on active learning has a lower dependence on data and lower cost of data labeling and can use fewer training samples to achieve the same accuracy as the baseline model.

4.3.Model Generalization Performance Verification

In this subsection, we test our proposed model on the CLEVR-Human dataset to verifies the generalization performance of the model.

We first utilize the CLEVR dataset to train our model and test the model on the CLEVR-Human dataset directly. Then, we fine-tune the model on CLEVR-Human and test the performance of the model on CLEVR-Human to verify the effectiveness of the model with active learning when facing more linguistic diversity problems. Table 4 shows the difference in overall accuracy between the model and the baseline. It can be seen that the accuracy of the visual reasoning model based on active learning on the CLEVR-Human data set is 11.33% higher than the baseline, especially MNLP method. The accuracy of the model after fine-tune on CLEVR-Human is 10.83% higher than the baseline, of which the MNLP method has the highest accuracy.

Table 4: The accuracy of the model trained on the CLEVR dataset is tested directly on CLEVR-Human and after fine-tune on CLEVR-Human.

Method	Before fine-tune	After fine-tune
Baseline	54.0%	66.6%
Ours-LC	61.3%	72.2%
Ours-SE	66.1%	79.6%
Ours-MNLP	68.6%	80.5%

Through the comparison of the above three aspects, we found that active learning can be used to determine whether the sample has higher training value by calculating the relevant feature information of the sample. Due to LC cannot measure the uncertainty of the sample well, it performs the worst in the three comparison verifications. MNLP can performs best when the sample demand is small. The reason is that, after performing the logarithmic normalization on LC, MNLP can better measure the uncertainty of samples and select more representative sample. SE performs better when the sample demand is large. The reason is that, the sample selected by the sample information entropy is more informative. Therefore, MNLP is more suitable for a smaller number of samples, while SE is more suitable for a large sample size.

In summary, first, active learning can ensure that the training cost is reduced without reducing the accuracy of the model; second, it can reduce the dependence of the model on the data and the

cost of data labeling; finally, it can alleviate the over-fitting to enhance the generalization performance.

5. Conclusions

This paper proposes a visual reasoning model based on active learning and finally proves that the visual reasoning model with active learning is not only higher in accuracy than the baseline model, but also better than the baseline model in generalization performance. In addition, active learning can help to reduce the cost of data labeling and the dependence of the model on the data.

However, our proposed model has certain limitations. In the process of active learning, the model needs to filter samples and retrain the model each time, which reduces training efficiency. In future work, we will continue to explore possible directions to optimize existing models. Meanwhile, we will also conduct research and exploration on the other data sets.

References

[1] S. Antol, A. Agrawal, J. Lu, et al. VQA: Visual question answering, The IEEE International Conference on Computer Vision（ICCV）. Santiago, Chile, 2015:2425-2433.

[2] Johnson J, Hariharan B, van der Maaten L, et al. Inferring and executing programs for visual reasoning, Proceedings of the IEEE International Conference on Computer Vision. Venice, Italy, 2017: 2989-2998.

[3] Runtao Liu, Chenxi Liu, Yutong Bai, et al. CLEVR-Ref+: Diagnosing Visual Reasoning With Referring Expressions，The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 4185-4194.

[4] Van Steenkiste, Sjoerd and Locatello, et al. Are Disentangled Representations Helpful for Abstract Visual Reasoning?, Advances in Neural Information Processing Systems 32(NIPS), 2019, pp.14245--14258.

[5] Jiaxin Shi, Hanwang Zhang, Juanzi Li. Explainable and Ex-plicit Visual Reasoning Over Scene Graphs，The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 8376-8384.

[6] Perez E, Strub F, De Vries H, et al. Film: Visual reasoning with a general conditioning layer, Thirty-Second AAAI Conference on Artificial Intelligence. Louisiana: AAAI Press, 2018.

[7] Keze Wang, Dongyu Zhang, Ya Li, et al. Cost-effective active learning for deep image classification， IEEE Transac-tions on Circuits and Systems for Video Technology, vol.27, no. 12, pp. 2591-2600, Dec. 2017.

[8] Yarin Gal, Riashat Islam, and Zoubin Ghahramani. Deep bayesian active learning with image data, ICML'17: Proceedings of the 34th International Conference on Ma-chine Learning-Volume 70, August, 2017, Pages 1183–1192.

[9] Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision?, NIPS: arXiv preprint arXiv:1703.04977, 2017.

[10] Samarth Sinha, Sayna Ebrahimi, Trevor Darrell. Variational Adversarial Active Learning, The IEEE International Conference on Computer Vision (ICCV),2019, pp. 5972-5981.

[11] Ye Zhang, Matthew Lease, and Byron C Wallace. Active discriminative text representation learning, AAAI, pp. 3386–3392, 2017.

[12] Krishnamurthy, Akshay, et al. Active Learning for Cost-Sensitive Classification. Journal of Machine Learning Research, 2019, 20.65: 1-50.

[13] David Mascharka, Philip Tran, Ryan Soklaski, et al. Transparency by Design: Closing the Gap Between Performance and Interpretability in Visual Reasoning, The IEEE Con-ference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 4942-4950.

[14] Culotta A, McCallum A. Reducing labeling effort for structured prediction tasks, AAAI. 2005, 5: 746-751.

[15] Johnson J, Hariharan B, van der Maaten L, et al. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Venice, Italy, 2017: 2901-2910.

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, et al. Deep Residual Learning for Image Recognition， Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, Nevada, 2016, pp. 770-778.

[17] O. Russakovsky, J. Deng, H. Su, et al. ImageNet large scale visual recognition challenge, International Journal of Computer Vision volume 115, pages211–252(2015).