2018 24th International Conference on Pattern Recognition (ICPR)
Beijing, China, August 20-24, 2018

Towards Unconstrained Pointing Problem of Visual Question Answering: A Retrieval-based Method

Wenlong Cheng, Yan Huang, Liang Wang
Center for Research on Intelligent Perception and Computing (CRIPAC)
National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences
University of Chinese Academy of Sciences, Beijing 100190, China
Email: {wenlong.cheng, yhuang, wangliang}@nlpr.ia.ac.cn

Abstract—**The pointing problem of visual question answering (VQA) is that given an image and a question which asks for the location of the interested object, find a region that answers the question. It is an important research problem in VQA tasks and has many potential applications in our daily life. Most of the existing work on this task can only solve it in the form of multiple choices, i.e., given candidate answers in advance, and then selecting a correct one. In this paper, we propose a retrieval model, which can not only deal with the multiple-choices task, but also provide a feasible solution for the no-candidate-answer task. The principle of our method is to pull the question and correct answer close, and push the question and incorrect answer away in a common feature space. To our best knowledge, we are the first to use retrieval method to solve the unconstrained (no-candidate-answer) pointing problem of VQA. Furthermore, our proposed method outperforms the state-of-the-art methods on the Visual7W [1] dataset in terms of the pointing problem of VQA.**

I. Introduction

With the development of computer vision (CV) and natural language processing (NLP), as well as the increment of computing ability and the availability of relevant large-scale datasets, there emerge some new tasks, combining the knowledge of CV and NLP, such as image captioning [2]–[5] and VQA [6]–[8]. Compared with image captioning, VQA is a more complex problem, because it needs deeper reasoning between visual representations and textual semantics. What is VQA? In short, VQA is that given an image and a question, output a natural language answer according to the given question-answer pair [6]. The types of VQA can be divided into four categories, namely joint learning approaches, attention mechanisms, compositional models, and using external knowledge bases models, from a perspective of used methods. Besides, they can be classified into two categories from a perspective of how to generate answers, one is classification, and the other is generation [7].

Compared with image captioning, VQA needs more detailed correlation between the words of the given question and regions of the corresponding image. Compared with textual question answering (QA), VQA also needs additional visual information of the corresponding image besides the textual question information. Dealing with visual information brings great difficulties, because the image contains low-level visual information, which contains relatively less high-level semantics than textual information. What's more, the image lacks the structural and grammatical rules of language, so VQA itself is already a more complex problem than QA [7].

Zhu et al. [1] propose the pointing problem of VQA, and they publish a Viusal7W dataset, which consists of 7W questions, i.e., what, where, when, who, why, how and which. The Visual7W dataset is a subset of Visual Genome [9], the largest VQA dataset, and it adds the visual questions, compared with most of the exsiting VQA datasets. In this paper, we will mainly focus on the pointing (which) problem of VQA. Research on the pointing problem of VQA can help us save a lot of time to focus on the relevant regions in images and aid us in understanding the natural language answer. For this problem, Zhu et al. [1] propose a LSTM + attention mechanism model, which adds a spatial attention mechanism to the LSTM architecture, to improve the precision of question encoding. But there is a notable gap between human performance and the LSTM + attention mechanism model. In addition, the LSTM + attention mechanism model [1] only deals with the multiple-choices pointing problem. Therefore, the scope of its application is limited.

To address the above problems, we propose a retrieval model, which has the ability to solve the pointing problem of VQA without candidate answers. First, it can obtain better performance than the LSTM + attention mechanism model. Second, it can not only deal with the multiple-choices pointing problem of VQA, but also deal with the no-candidate-answer pointing problem. Considering that object detection has gradually become a mature technique, and accurate object detection is the key factor to generate qualified candidate regions [10], we use the object proposals methods, such as Edge Boxes [11], to generate candidate regions in the given image, and then obtain a relatively good result by using the generated candidate answers.

Our main contributions can be summarized as follows. First, we propose a retrieval model to solve the multiple-choices pointing problem of VQA. Second, our proposed model can obtain a better performance than the LSTM + attention mechanism [1] model on the pointing task of multiple choices. Third, we attempt to solve a new problem, the unconstrained pointing problem of VQA. To our best knowledge, it is the first attempt to use the retrieval method to solve the unconstrained pointing task.

978-1-5386-3788-3/18/$31.00 ©2018 IEEE

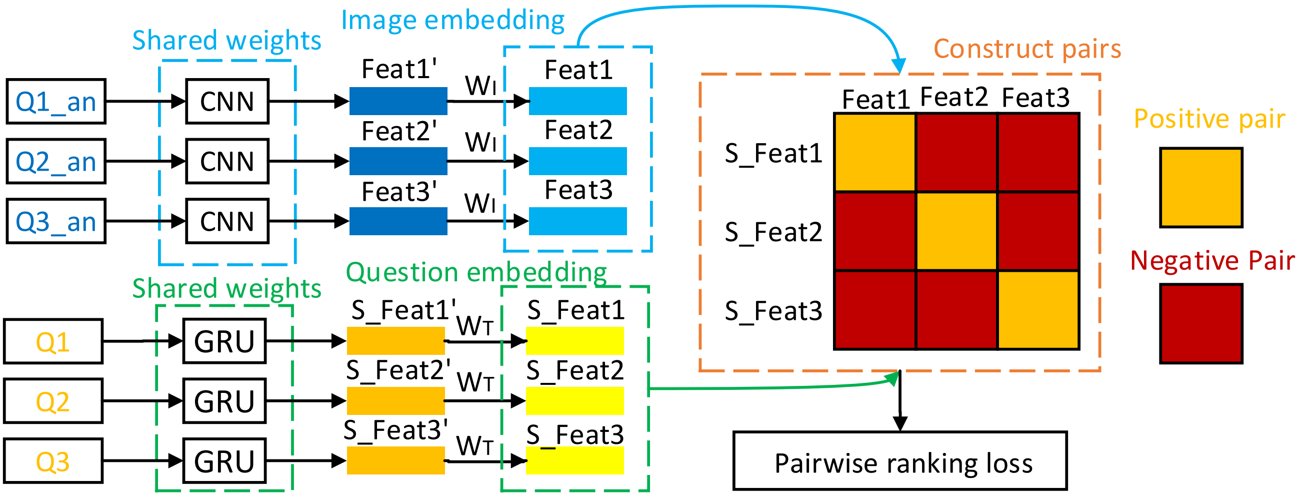

Fig. 1. The training process of our proposed retrieval model. The model is trained in a batch. Suppose the batch size is n, and $n = 3$ in this example. It contains three modules, namely the module of image embedding, the module of question embedding and the module of constructing pairs. Suppose the given questions are Q1, Q2 and Q3, and the corresponding correct answers (image regions) are Q1_an, Q2_an and Q3_an, respectively. Here, 'Feat' denotes the feature of images, and 'S_Feat' denotes the feature of questions. W_I and W_T are the image mapping matrix and textual mapping matirx, respectively. In the phase of constructing pairs, the positive pairs are (S_Feat1, Feat1), (S_Feat2, Feat2), (S_Feat3, Feat3), while the negative pairs we choose are (S_Feat1, Feat2), (S_Feat1, Feat3), (S_Feat2, Feat1), (S_Feat2, Feat3), (S_Feat3, Feat1), (S_Feat3, Feat2). Then we use pairwise ranking loss to update the weights of our model. (Best viewed in color.)

II. Related Work

A. Image-Sentence Embedding

The representation of image and sentence is the crucial step in solving multi-modal tasks, such as image-sentence matching, image captioning and VQA. Socher et al. [12] propose an approach to learn a joint image-sentence embedding. Karpathy et al. [13] propose a fine-grained model, which can embed fragments of image and fragments of sentence into a common feature space. Besides, Gong et al. [14] propose an improved approach, by using the weakly annotated images to improve the joint image-sentence embedding. Kapathy and Li [4] propose a visual-semantic alignment model to further improve the precision of image and sentence embeddings, and build a finer level relationship between the given image and correspnding sentence. Huang et al. [15] propose a semantic-enhanced model to improve the precision of image embedding. Different from the above methods, our model use the pairwise ranking loss to learn a mapping so that the features of questions and correct answers are near, while the features of questions and incorret answers are far, in the embedding space.

B. Image Captioning

Image captioning is a very hot topic, and it has achieved great progress in recent years. Kiros et al. [16] propose an encoder-decoder pipeline for the image captioning task. Oriol et al. [17] propose an end-to-end model to solve the problem of image captioning. Xu et al. [5] propose a two-attention-based image caption generator, namely a "soft" deterministic attention and a "hard" stochastic attention mechanism to guide "where" and "what" the attention should focuse on. Yang et al. [18] propose an object detection and location model to solve the image captioning problem. Dai and Lin [19] propose a contrastive learning method to solve the problem of image captioning. Recently, inspired by the successful application of machine translation and conditional image generation, Aneja et al. [20] propose a convolutional image captioning technique. The VQA task bears some resemblance to the image captioning task. Our proposed method is inspired by work of Kiros et al. [16].

C. Visual Question Answering

With the development of deep learning, visual question answering has received more and more attention. Malinowski et al. [21] propose a "Neural-Image-QA" model, which uses the Long Short-Term Memory cells (LSTM) to encode the sentence with the holistic image feature, and the decode process is also implemented by LSTM. Ren et al. [22] propose two models, the first is "VIS+LSTM", which adds the visual information at the start, and the second is "2-VIS+BLSTM", which adds the visual information at the start and the end. Zhu et al. [1] summarize 7W questions and propose a new task "which" in VQA, as well as propose a LSTM + attention mechanism model. Yang et al. [23] propose stacked attention networks, which use semantic representation of a question to search the relevant regions of a given question. Shih et al. [8] propose a method, which generates proposals by the object proposals method and learns to answer visual questions by selecting the proposals relevant to the given question. Anderson et al. [24] propose a combined bottom-up and top-down visual attention mechanism for both image captioning and visual question answering. In addition, Trott et al. [25] propose an iterpretable approach for solving the problem of counting in VQA. Compared with the above methods, our method uses the retrieval method to train the model without the features of the whole image and attempts to deal with a new task, the unconstrained pointing problem of VQA.

III. Our Model

We propose a retrieval model to solve the pointing problem of VQA. The training process of our proposed model is shown in Fig. 1. Given a question-answer pair, we use convolutional neural network (CNN) to extract visual features of candidate answers with shared weights, and use GRU to extract the textual features of the question. Then we map the visual features and textual features into a common feature space with the weight matrices W_I and W_T, respectively. We use the visual image features and textual question features in the common feature space to construct positive pairs and negative pairs. In addition, we use the constructed pairs to update the model with a pairwise ranking loss. In the following sections, we will introduce how to train our retrieval model and how to test our model's performance.

A. Sentence Representation

Compared with the LSTM [26], the GRU [27] model has fewer parameters, but can obtain a comparable result. Besides, the GRU can solve the problem of vanishing gradient and exploding gradient, so the GRU can learn good textual features to express the sentence. For efficiency and simplicity, we choose GRU to encode the question in the given question-answer pair. The GRU unit includes update gate and reset gate. Suppose the length of the sentence is N, then the hidden state of GRU at time step N is the representation of the whole sentence.

B. Image Region Representation

We use a CNN to encode image region information. In our experiments, we use VGG16 [28] with shared weights to extract visual information of candidate answers based on the given image. Image region features are extracted from 'fc7' in the VGG16 neural network, and the dimensionality of the image feature is 4096. Then, we map the image region features into the common feature space.

C. Pairwise Ranking Loss

In our experiments, we use pairwise ranking loss to update our model. Let the dimensionality of image features be D_I, the dimensionality of sentence features be D_T, the dimensionality of the common feature space be D, the vocabulary size be V. In addition, let $W_I \in \mathbb{R}^{D\times D_I}$ and $W_T \in \mathbb{R}^{D\times D_T}$ be the image mapping matrix and textual mapping matrix, respectively. Given the question $Q = \{w_1, w_2, \ldots, w_N\}$, where $w_1, w_2, \ldots, w_N$ are the words in the question sentence. Suppose their corresponding word embeddings are $\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_N$, and $\mathbf{w}_i \in \mathbb{R}^D$, $i = 1, 2, \ldots, N$. The question sentence embedding v can be represented by the hidden state of GRU at time step N. Let X_I denote the image feature extracted by VGG16, which is mapped into the common feature space, where $x = W_I X_I$, $X_I \in \mathbb{R}^{D_I}$, $x \in \mathbb{R}^D$.

Similar to the principle of contrastive loss and triplet loss, the pairwise ranking loss [4], [13], [16] is an effective method to maintain the similarity of the paired samples. Define the similarity function $s(x, v) = x^T v$, where x and v have been normalized. Let θ be all the learned parameters, the pairwise ranking loss can be formulated as following.

TABLE I
Comparisions on existing models of multiple-choices pointing QA task. (in accuracy)

Method	Pointing Task
Human (Question+Image) [1]	0.973
Logistic Regression (Question+Image) [1]	0.307
LSTM (Question+Image) [1]	0.521
LSTM-Att (Question+Image) [1]	0.561
LSTM-Retrieval-PRL (Question+Image)	**0.630**

$$\begin{aligned}\min_{\theta}\{\sum_{x}\sum_{k} max\{0, \alpha - s(x, v) + s(x, v_k)\}+ \\ \sum_{v}\sum_{k}\max\{0, \alpha - s(v, x) + s(v, x_k)\}\}\end{aligned} \tag{1}$$

where v_k is a contrastive question for visual candidate answer x, vice versa, x_k is a contrastive candidate answer for question v. The contrastive terms are sampled from the same batch of training samples, and Fig. 1 shows how to construct pairs in the given mini-batch question-answer pairs.

The principle of the pairwise ranking loss is that pulling the question and its corresponding answer close, as well as push the question and incorrect answers away in the common feature space. Considering both efficiency and complexity, we do not use the incorrect candidate answers in the same question-answer pair to construct the negative pairs here. Because the candidate answers from different question-answer pairs can give the training process harder samples, the model can have the better discrimination ability.

D. Training

We use pairwise ranking loss to train our model by the strategy of mini-batch samples. The whole architecture of training process can be seen in Fig. 1. We use the Adam optimizer to optimize our model. In our experiments, the question is encoded by the GRU. The details of our experiments will be discussed in the following experiments.

E. Testing

We use the trained model to obtain the features of the questions and the features of their corresponding candidate answers (image regions) in the common feature space, respectively. We use the feature of one question to retrieve the features of its corresponding candidate answers. Then we choose the most similar one as the predicted answer.

IV. Experiments

We conduct our experiments on the Visual7W[1] [1] dataset. In our experiments, we conduct two tasks to verify the pro-

[1] To our best knowledge, the Visual7W dataset is currently the only one dataset for the pointing problem of VQA, so we can only conduct our experiments on this dataset.

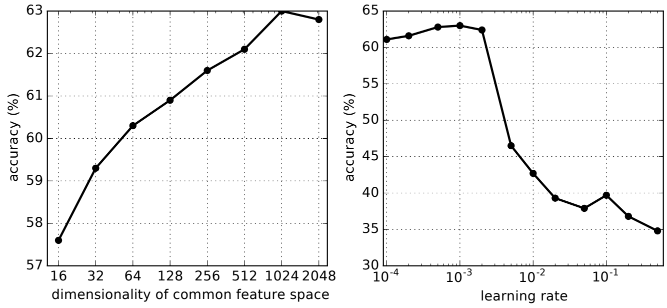

Fig. 2. The sensitivity analysis of some hyperparameters.

posed model. One is the pointing problem of multiple-choices VQA, and the other is the pointing problem of no-candidate-answer VQA. The experiment settings are as follows. The batch size is 128. The dimensionality of word embedding is 300. The dimensionality of the common feature space is 1024. The learning rate is 0.001. The margin is 0.20.

A. Visual7W Dataset

The Visual7W [1] dataset contains 7W question types, including what, where, when, who, why, how and which. It contains a part of Visual Genome [9] dataset, which contains 1.7 million QA pairs of the 7W question types. Compared with the Visual Genome dataset, the Visual7W dataset provides more annotations, such as object groundings, multiple choices and human experiments. The Visual7W dataset contains the train dataset and test dataset. In our experiments, we randomly choose 5000 question-answer pairs from the train dataset as the validation dataset, and choose the remained train dataset as the train dataset. The test dataset remains unchanged.

B. Multiple-choices VQA

The task of multiple-choices VQA can be defined to choose one possible answer from four candidate answers, which include one correct answer. We compare our retrieval model with a current state-of-the-art method [1]. The results are shown in TABLE I. In the table, the last row shows our proposed method **LSTM-Retrieval-PRL**, where '**PRL**' indicates that our model uses pairwise ranking loss to update the learned parameters. Compared with Zhu et al. [1] in the pointing problem of multiple-choices VQA, we can see that our proposed method achieves significant improvement. We also test the sensitivity of hyperparameters, such as the dimensionality of the common feature space and the learning rate. The experiments of the sensitivity of some hyperparameters are shown in Fig. 2. From the figure, we can see that with the increment of the dimensionality of the common feature space, the accuracy slowly increases, and then decreases if the dimensionality of the common feature space is too high. We choose 1024 as the dimensionality of the common feature space. We can also see that with the increment of the learning rate, the accuracy slowly increases, and then decreases dramatically. The learning rate is finally equal to 0.001. From the figure we can draw a conclusion that compared with the dimensionality of common feature space, the learning rate has more significant effect on the model.

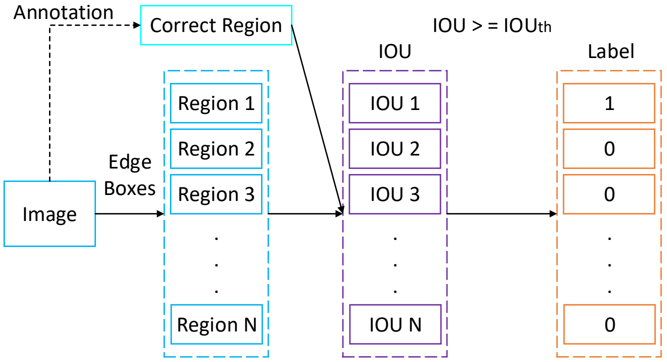

Fig. 3. The flowchart of how to generate candidate answers. If $IOU >= IOU_{th}$, then the label of the generated region is 1, otherwise, the label is 0. Here, 1 denotes the correct candidate answer, while 0 indicates the incorrect candidate answer.

C. No-candidate-answer VQA

The task of no-candidate-answer VQA can be defined to find one possible answer from the corresponding image based on the given question. Because there is no candidate answer, we should generate a lot of candidate answers. Here we use Edge Boxes [11] to generate region proposals (candidate answers). Fig. 3 is the flowchart of how to label the generated proposals, and how to judge the quality of generated proposals. IOU_{th} is an IOU threshold to discriminate the generated correct proposals from the generated proposals. The process of how to label the generated candidate answers is described as followings. First, we use Edge Boxes to generate region proposals, then we compute the intersection-over-union (IOU) between the region of the correct answer and generated region proposals. If IOU is greater than or equal to IOU_{th}, we think the generated region proposal is correct, and the label of the region proposal is 1. Otherwise, we think the generated region proposal is incorrect, and the label of the region proposal is 0.

TABLE II shows our results on no-candidate-answer VQA. The Fig. 4 is the corresponding figure. In the experiment, we change the number of generated proposals from 150 to 1200 with an interval of 150, and change the IOU_{th} from 0.3 to 0.5 with an interval of 0.1. From TABLE II and Fig. 4, we can observe that the recall and accuracy both improve with the decrement of IOU_{th}. Because the standard of correct answers decreases, the total number of correct candidate answers increases while the total number of candidate answers remains unchanged in a given question-answer pair. We can also see that the recall substantially increases with the increment of the number of generated proposals. But the accuracy increases slowly at first, and then slowly decreases with the further increment of the number of generated proposals. This is because the increased choices not only bring the increment of recall, but also bring the difficulty to choose the correct answer from generated candidate proposals. How to balance

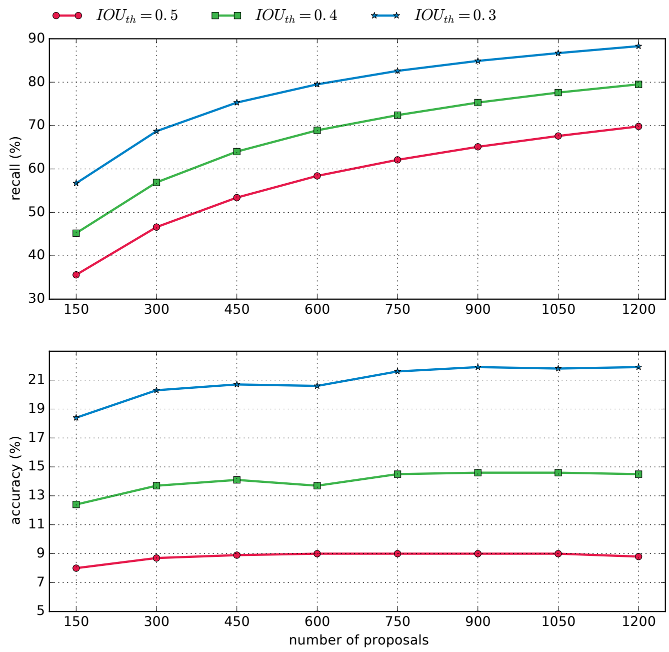

Fig. 4. The recall and accuracy curves of the pointing problem of VQA without candidate answers.

the recall and the number of generated proposals is a very important problem. In other words, how to generate qualified proposals is crucial to solve this task.

Compared with the multiple-choices task, the no-candidate-answer task obtains a much lower performance. There are two reasons for the poor performance. One is that the generated answers may not include the correct answer. The other is that the number of generated candidate answers is much larger than the number of multiple choices. We introduce the ratio between the total correct answers and total candidate answers to eveluate the difficulty of selecting the correct answer. If the ratio is higher, the difficulty of selection is smaller. There are always 4 candidate answers in the given question-answer pair in multiple-choices task. So its ratio is 0.25. The ratios of the no-candidate-answer task are shown in Fig. 5. The ratios in no-candidate-answer task are much lower than the multiple-choices task, which can explain why the no-candidate-answer task obtains a much lower performance than the multiple-choices task.

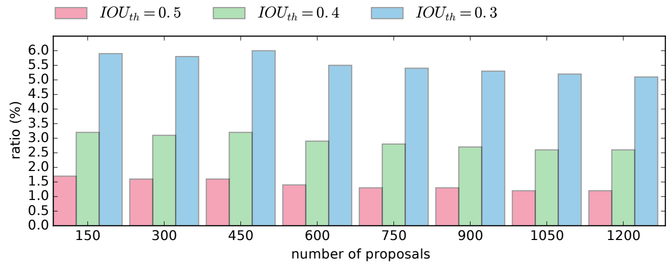

Fig. 5. The ratios of different numbers of generated proposals and different IOU_{th} values. (Best viewed in color.)

TABLE II
THE PERFORMANCE OF OUR PROPOSED MODEL IN NO-CANDIDATE-ANSWER VQA TASK. (IN ACCURACY)

proposals number	IOU_{th}	recall	accuracy
150	0.5	0.356	0.076
150	0.4	0.452	0.117
150	0.3	0.567	0.176
300	0.5	0.466	0.087
300	0.4	0.569	0.137
300	0.3	0.687	0.203
450	0.5	0.534	0.084
450	0.4	0.640	0.135
450	0.3	0.753	0.202
600	0.5	0.584	0.090
600	0.4	0.689	0.137
600	0.3	0.795	0.206
750	0.5	0.621	0.090
750	0.4	0.724	0.145
750	0.3	0.826	0.216
900	0.5	0.651	0.090
900	0.4	0.753	0.146
900	0.3	0.849	0.219
1050	0.5	0.676	0.090
1050	0.4	0.776	0.146
1050	0.3	0.867	0.218
1200	0.5	0.698	0.088
1200	0.4	0.795	0.145
1200	0.3	0.883	0.219

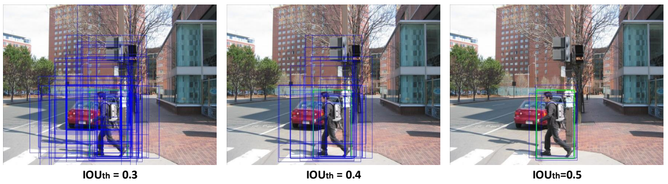

Fig. 6. The visualization of generated correct answers. From left to right, the values of IOU_{th} are 0.3, 0.4, 0.5, respectively. In the images, the green bounding boxes are the groundtruth correct answers, while the blue bounding boxes are the generated correct answers according to different IOU_{th} values. (Best viewed in color.)

There are several key problems in no-candidate-answer VQA task. First, how to choose the IOU_{th} is an important problem, because the IOU_{th} determines the quality of generated proposals. If we set the IOU_{th} low, the generated correct answer may contain many irrelevant objects or large surroudings. If we set the IOU_{th} high, the generated region proposals can not include the correct answer. Fig. 6 shows the visualization of generated correct answers with different IOU_{th}. From the figure, we can see that when the IOU_{th} equals to 0.3, many generated correct proposals cover some irrelevant objects except for the correct object, and the range of the bounding box is too large or small. If we increase the IOU_{th} to 0.4, the range of bounding box becomes more accurate. Furthermore, when IOU_{th} equals to 0.5, the generated proposals have great overlap with the correct answer, so we can think the generated proposals are the correct answers. Second, if we use this way to generate candidate answers, more than one correct answer or no correct answer will be generated. Third, we need to get region proposals firstly, so the computational efficiency is also a problem. In the following work, we will attempt to solve the above problems.

V. Conclusion

In this paper, we have proposed a retrieval model to solve the pointing problem of VQA. To our best knowledge, we are the first to use the retrieval method to solve the unconstrained pointing problem of VQA. In addition, our model outperforms the existing work on the pointing problem of VQA on the Visual7W dataset. Furthermore, our model can not only deal with the pointing problem of multiple-choices VQA, but also provide a feasible solution to the no-candidate-answer problem. In the future work, we will add the attention mechanism and the whole image visual features to improve the precision of the question encoding. For the no-candidate-answer task, we will try other region proposals methods and deep reinforcement learning to improve the quality of the generated proposals.

Acknowledges

This work is jointly supported by National Key Research and Development Program of China (2016YFB1001000), National Natural Science Foundation of China (61525306, 61633021, 61721004, 61420106015), Beijing Natural Science Foundation (4162058), and Capital Science and Technology Leading Talent Training Project (Z181100006318030).

References

[1] Y. Zhu, O. Groth, M. Bernstein, and L. Fei-Fei, "Visual7w: Grounded question answering in images," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 4995–5004. 1, 2, 3, 4

[2] J. Mao, W. Xu, Y. Yang, J. Wang, Z. Huang, and A. Yuille, "Deep captioning with multimodal recurrent neural networks (m-rnn)," *arXiv preprint arXiv:1412.6632*, 2014. 1

[3] X. Chen and C. Lawrence Zitnick, "Mind's eye: A recurrent visual representation for image caption generation," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 2422–2431. 1

[4] A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3128–3137. 1, 2, 3

[5] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *International Conference on Machine Learning*, 2015, pp. 2048–2057. 1, 2

[6] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. Lawrence Zitnick, and D. Parikh, "Vqa: Visual question answering," in *IEEE International Conference on Computer Vision*, 2015, pp. 2425–2433. 1

[7] Q. Wu, D. Teney, P. Wang, C. Shen, A. Dick, and A. van den Hengel, "Visual question answering: A survey of methods and datasets," *Computer Vision and Image Understanding*, 2017. 1

[8] K. J. Shih, S. Singh, and D. Hoiem, "Where to look: Focus regions for visual question answering," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 4613–4621. 1, 2

[9] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma *et al.*, "Visual genome: Connecting language and vision using crowdsourced dense image annotations," *International Journal of Computer Vision*, vol. 123, no. 1, pp. 32–73, 2017. 1, 4

[10] R. Hu, H. Xu, M. Rohrbach, J. Feng, K. Saenko, and T. Darrell, "Natural language object retrieval," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 4555–4564. 1

[11] C. L. Zitnick and P. Dollár, "Edge boxes: Locating object proposals from edges," in *European Conference on Computer Vision*. Springer, 2014, pp. 391–405. 1, 4

[12] R. Socher, A. Karpathy, Q. V. Le, C. D. Manning, and A. Y. Ng, "Grounded compositional semantics for finding and describing images with sentences," *Transactions of the Association for Computational Linguistics*, vol. 2, pp. 207–218, 2014. 2

[13] A. Karpathy, A. Joulin, and L. F. Fei-Fei, "Deep fragment embeddings for bidirectional image sentence mapping," in *Advances in Neural Information Processing Systems*, 2014, pp. 1889–1897. 2, 3

[14] Y. Gong, L. Wang, M. Hodosh, J. Hockenmaier, and S. Lazebnik, "Improving image-sentence embeddings using large weakly annotated photo collections," in *European Conference on Computer Vision*. Springer, 2014, pp. 529–545. 2

[15] Y. Huang, Q. Wu, and L. Wang, "Learning semantic concepts and order for image and sentence matching," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 2

[16] R. Kiros, R. Salakhutdinov, and R. S. Zemel, "Unifying visual-semantic embeddings with multimodal neural language models," *arXiv preprint arXiv:1411.2539*, 2014. 2, 3

[17] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3156–3164. 2

[18] Z. Yang, Y.-J. Zhang, S. ur Rehman, and Y. Huang, "Image captioning with object detection and localization," in *International Conference on Image and Graphics*. Springer, 2017, pp. 109–118. 2

[19] B. Dai and D. Lin, "Contrastive learning for image captioning," in *Advances in Neural Information Processing Systems*, 2017, pp. 898–907. 2

[20] J. Aneja, A. Deshpande, and A. Schwing, "Convolutional image captioning," in *IEEE International Conference on Computer Vision*, 2018. 2

[21] M. Malinowski, M. Rohrbach, and M. Fritz, "Ask your neurons: A neural-based approach to answering questions about images," in *IEEE International Conference on Computer Vision*, 2015, pp. 1–9. 2

[22] M. Ren, R. Kiros, and R. Zemel, "Exploring models and data for image question answering," in *Advances in Neural Information Processing Systems*, 2015, pp. 2953–2961. 2

[23] Z. Yang, X. He, J. Gao, L. Deng, and A. Smola, "Stacked attention networks for image question answering," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 21–29. 2

[24] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang, "Bottom-up and top-down attention for image captioning and vqa," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 2

[25] A. Trott, C. Xiong, and R. Socher, "Interpretable counting for visual question answering," in *International Conference on Learning Representations*, 2018. 2

[26] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997. 3

[27] K. Cho, B. van Merriënboer, Ç. Gülçehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using rnn encoder–decoder for statistical machine translation," in *Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2014, pp. 1724–1734. 3

[28] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *International Conference on Learning Representations*, 2015. 3