Computers, Materials & Continua CMC, vol.65, no.3, pp.2529-2541, 2020

Stacked Attention Networks for Referring Expressions Comprehension

Yugang Li[1, *], Haibo Sun[1], Zhe Chen[1], Yudan Ding[1] and Siqi Zhou[2]

Abstract: Referring expressions comprehension is the task of locating the image region described by a natural language expression, which refer to the properties of the region or the relationships with other regions. Most previous work handles this problem by selecting the most relevant regions from a set of candidate regions, when there are many candidate regions in the set these methods are inefficient. Inspired by recent success of image captioning by using deep learning methods, in this paper we proposed a framework to understand the referring expressions by multiple steps of reasoning. We present a model for referring expressions comprehension by selecting the most relevant region directly from the image. The core of our model is a recurrent attention network which can be seen as an extension of Memory Network. The proposed model capable of improving the results by multiple computational hops. We evaluate the proposed model on two referring expression datasets: Visual Genome and Flickr30k Entities. The experimental results demonstrate that the proposed model outperform previous state-of-the-art methods both in accuracy and efficiency. We also conduct an ablation experiment to show that the performance of the model is not getting better with the increase of the attention layers.

Keywords: Stacked attention networks, referring expressions, visual relationship, deep learning.

1 Introduction

Great progresses have been made on computer vision, natural language processing, knowledge embedding and reasoning, even quantum computation [Liu, Gao, Yu et al. (2018); Liu, Chen, Liu et al. (2018); Liu, Gao, Wang et al. (2019)]. But the task of locating the image region described by a natural language expression is far from being solved. This task, also known as referring expression comprehension or grounding referring expression [Hu, Rohrbach, Andreas et al. (2016); Rohrbach, Rohrbach, Hu et al. (2016)]. It is challenging to address this problem, because it involves both natural language processing and computer vision. Referring expressions comprehension can be

[1] Academy of Broadcasting Science, Beijing, 100866, China.

[2] School of Electrical and Electronic Engineering, Nanyang Technological University, 639798, Singapore.

[*] Corresponding Author: Yugang Li. Email: liyugang@abs.ac.cn.

Received: 03 June 2020; Accepted: 03 July 2020.

widely used in many applications which need natural language interfaces to communicate with people, e.g., human-robot interaction or image retrieval, in which objects are queried by their attributes, location or relationships of the objects with other objects.

Referring expressions often refer to several objects and the relationships between them, for example, in Fig. 1, the expression *the man wearing yellow clothes* describes a *man* entity which participates in a *wearing* relationship with a *yellow clothes* entity. While most previous work [Nagaraja, Morariu and Davis (2016); Rohrbach, Rohrbach, Hu et al. (2016); Yu, Poirson, Yang et al. (2016)] locates the region by selecting the relevant entity described by the referring expression from a set of candidate proposals. As a result, the precision of these approaches depends on the performance of the object detection module, and when the number of objects is huge, it is time-consuming to select the relevant entity corresponding to the referring expression. We find that localizing the referred entity from an image often requires multi-step reasoning. Take Fig. 1 as an example, there are many entities in the image, to find *the man wearing yellow clothes*, we first locate all those entities (e.g., *man*, *yellow clothes*) and concepts (e.g., *wearing*), then rule out irrelevant entities gradually, finally localize the most relevant entity.

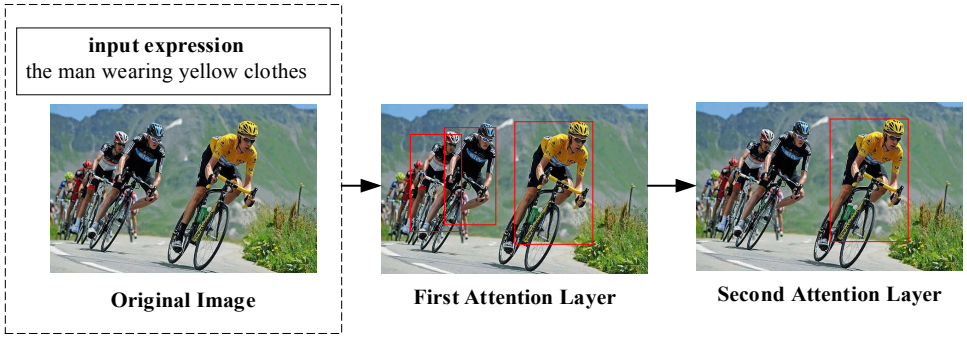

Figure 1: An illustration of our framework for referring expression comprehension

In this paper, we propose a framework for referring expressions comprehension based on stacked attention networks. We focus on the referring expressions which can be decomposed as a (subject, predicate, object) triplet. As shown in Fig. 2, the proposed framework consists of three components: (1) the image model, which employs a convolution neural network (CNN) to extract image features, here we use VGGNet [Simonyan and Zisserman (2014)]; (2) referring expressions parsing model, which employs soft attention to parse the referring expression into a subject, a relationship and an object and (3) stacked attention networks model, which localizes the most relevant region relate to the referring expression by multi-step reasoning. As illustrated in Fig. 2, our model first parses the referring expression into subject, predicate and object, and concatenates them into an expression vector. Then we use the expression vector to search the represent of the image in the first attention layer, then refine the expression vector by combining with the retrieved image vectors. By implementing this process iteratively, we get a more relevant region to the input referring expression. Finally, by mapping the vectors from the last attention layer to the input image, we locate the region described by the input expression with a bounding box.

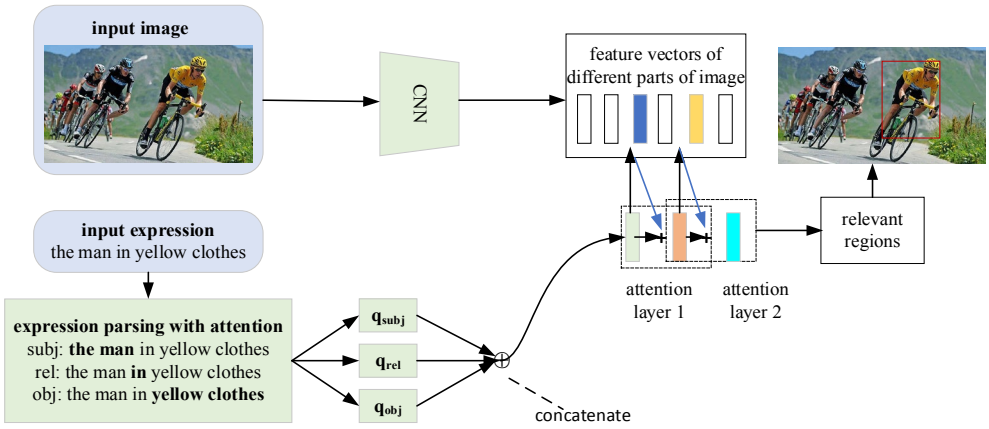

Figure 2: Our proposed model architecture for ground referential expressions

To sum up, there are three main contributions in our work. First, we propose a novel framework for referring expressions comprehension which uses stacked attention networks to reason the result by multi-step. Second, we perform comprehensive evaluation on two datasets and results demonstrate that the proposed model outperforms previous state-of-the-art approaches by a substantial margin. Third, we verify that the accuracy of the results is not proportional to the number of attention layers.

2 Related work

Referring expressions comprehension can be formulated as the problem of retrieving the most relevant region from a set of candidate regions of an image *I*, given a natural language expression *q*:

$$\hat{a} = \arg\max_{a \in A} p(a \mid I, q, \theta) \tag{1}$$

where θ is the parameters to be learned and *a* is the set of candidate regions.

The problem of referring expression comprehension has got more and more attentions recent years [Fukui, Huk Park, Yang et al. (2016); Hu, Rohrbach, Andreas et al. (2016); Hu, Xu, Rohrbach et al. (2015); Mao, Huang, Toshev et al. (2016); Nagaraja, Morariu and Davis (2016); Rohrbach, Rohrbach, Hu et al. (2016); Yu, Poirson, Yang et al. (2016)]. A commonly used architecture was to extract a set of candidate regions by employing object detection methods [Girshick (2015); Redmon, Divvala, Girshick et al. (2016); Ren, He, Girshick et al. (2015)]. Next, score each candidate region by extracting the region feature and looking at whether it matches with the input referring expression. Finally, we get the localization result by returning the bounding box of a region with the highest score. In order to establish an explicit correspondences between the regions in the image and the components in the referring expression, some work [Hu, Rohrbach, Andreas et al. (2016); Yu, Lin, Shen et al. (2018)] proposed modular network which composed of different modules, one used to decide whether a region matches the subject or object textual component, another used to decide whether a pair of regions matches the relationship described in the referring expression. Although these methods have got

impressive results, it is time-consuming when there are many candidate proposals in the image. Suppose there are n regions in the candidate proposal set A, and the referring expression refers to m objects, we need to compute O ($n \times m$) scores. In this work, we employ stacked attention network to locate the region described by the expression directly.

Stacked attention network is mainly based on complex recurrent neural networks which include an encoder and a decoder [Sutskever, Vinyals and Le (2014)]. It has been firmly applied as a state of the art approach in machine translation [Luong, Pham and Manning (2015)], image captioning [Lu, Xiong, Parikh et al. (2017); Xu, Ba, Kiros et al. (2015)], visual question answering (VQA) [Ilievski, Yan and Feng (2016)] and image classification [Li and Wang (2018)]. Stacked attention networks (SANs) can be seen as a variation of the attention mechanism which allow multi-step reasoning to get the result, and has been widely used in machine translation [Sukhbaatar, Weston and Fergus (2015)] and VQA [Yang, He, Gao et al. (2016)]. In this work, we use SANs to pinpoint the image region described by a referring expression. To the best of our knowledge, we propose the first SANs for the task of referring expressions comprehension.

Visual relationships are a pair of localized objects connected via a predicate. We represent visual relationship by a triplet in the form of (subject, predicate, object). There have been numerous efforts in visual relationship detection [Li, Wang and Chen (2019); Zhang, Kyaw, Chang et al. (2017)]. As visual relationships provide large semantic capacities, being able to endow a system with auxiliary information far beyond what individual object detectors provide. Some studies utilize visual relationships as semantic knowledge complementary [Lu, Ji, Zhang et al. (2018); Teney, Liu and Hengel (2017)], proposed a framework to learn visual relation for VQA [Xu, Zhu, Choy et al. (2017)], or utilized visual relationships to generate scene graphs via massage passing. In order to extract rich semantic knowledge from the input referring expression, we proposed an expression parsing method to decompose the expression into the triplet format (subject, predicate, object).

3 Our model

In this paper, we propose a framework to localize a specific region described by a referring expression of an image. The motivation in our model is that grounding referring expression in an image often requires multi-step reasoning. For referring expressions like *the cat on top of the table*, we need to first locate the entity *cat* and *table*, and comprehend the concept *on top of*, then gradually rule out irrelevant entities, finally pinpoint the relevant entity *cat*. To achieve this goal, we employ stacked attention networks to localize the region by multiple steps of reasoning.

The overall architecture of the proposed model is shown in Fig. 2. As illustrated, the architecture is composed of three components: the image model, the referring expression parsing model and the stacked attention networks model. Given a set of candidate regions and a referring expression, our model first parses the expression into three components: subject, predicate and object. Then we concatenate them together into a vector which is input to the stacked attention network to pinpoint the relevant region. In this section, we will describe the three models in detail.

3.1 Image model

The image model uses a convolution neural network (CNN) [Krizhevsky, Sutskever and Hinton (2012)] to get the representation of input images. In this work, we use VGGNet [Simonyan and Zisserman (2014)] to extract the image feature map f_I. We rescale the input images to be 448×448 pixels, and in order to retain spatial information of the original images, we get rid of fully connected layer of the VGGNet and only use features from the last pooling layer:

$$f_I = CNN_{vgg}(I) \tag{2}$$

The dimension of f_I is 512×14×14, where 14×14 denotes the number of split regions of the original image and 512 is the dimension of the feature vector of each region. As shown in Fig. 3, each feature vector in f_I corresponds to a 32×32 pixel region of the original image.

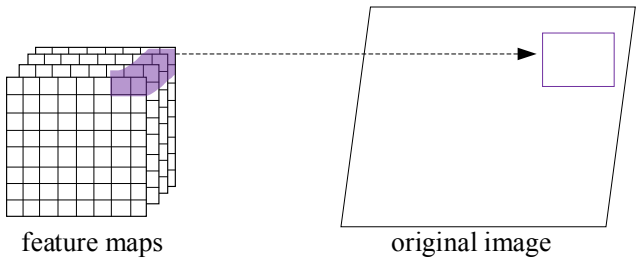

Figure 3: The relation between feature maps and original image

In this work, we transform each feature vector to a new vector which has the same dimension with the subject entity embedding (or object entity embedding):

$$v_I = \tanh(W_I f_I + b_I) \tag{3}$$

At test time, after multiple steps of searching we get a vector set v_Q ($v_Q \in v_I$), then by mapping v_Q to the original image, we pinpoint the relevant region.

3.2 Referring expression parsing model

Unlike previous studies using long short-term memory (LSTM) to represent the referential expression holistically, which fall short of determining whether a region matches an expression. Similar to Hu et al. [Hu, Rohrbach, Andreas et al. (2016)], we employ soft attention mechanism to decompose the input referring expression Q into three components: subject, predicate and object, and then compute their vector representations:

$$q_{subj} = \sum_{t=1}^{T} a_{t,subj} e_t \tag{4}$$

$$q_{rel} = \sum_{t=1}^{T} a_{t,rel} e_t \tag{5}$$

$$q_{obj} = \sum_{t=1}^{T} a_{t,obj} e_t \tag{6}$$

where e_t is the vector computed by embedding each word w_t of Q in a continuous space, $a_{t,subj}$, $a_{t,rel}$ and $a_{t,obj}$ are attention weights of subject, predicate, object respectively:

$$a_{t,subj} = \frac{\exp(\beta_{subj}^T h_t)}{\sum_{\tau=1}^{T} \exp(\beta_{subj}^T h_\tau)} \tag{7}$$

$$a_{t,rel} = \frac{\exp(\beta_{rel}^T h_t)}{\sum_{\tau=1}^{T} \exp(\beta_{rel}^T h_\tau)} \tag{8}$$

$$a_{t,obj} = \frac{\exp(\beta_{obj}^T h_t)}{\sum_{\tau=1}^{T} \exp(\beta_{obj}^T h_\tau)} \tag{9}$$

where h_t is concatenation of the hidden state of a 2-layer bidirectional LSTM network:

$$h_t = [h_t^{(1,fw)} h_t^{(1,bw)} h_t^{(2,fw)} h_t^{(2,bw)}] \tag{10}$$

where $h_t^{(1,fw)}$ and $h_t^{(1,bw)}$ is the forward and backward hidden state of first layer respectively, and they are concatenated into h_t^1 which is feed into the second layer and output $h_t^{(2,fw)}$ and $h_t^{(2,bw)}$. As a result, h_t incorporates information from context and the word w_t itself. We concatenate q_{subj}, q_{rel}, q_{obj} into a vector v_Q=[q_{subj} q_{rel} q_{obj}] as a representation of Q. By concatenating the three components into one vector, we explore semantic knowledge of the referring expression sufficiently.

3.3 Stacked attention networks model

Usually, an image consists of a lot of objects, many of which are irrelevant to the referring expression, so using the one global image feature vector to predict the answer could produce a suboptimal result. In this work, we use stacked attention networks (SANs) to locate the relevant region gradually via multi-step reasoning. SANs are able to rule out irrelevant regions and pinpoint the regions which are relevant to the referring expression.

Stacked attention network is based on memory network [Sukhbaatar, Weston and Fergus (2015); Weston, Chopra and Bordes (2014)] which is a recurrent neural network with an explicit attention mechanism that can select certain parts of the image vectors and stored them in an external memory. Given the image I and a natural language referring expression Q, the model uses Q to choose relevant regions which are stored in the external memory, this process named a "hop". Our attention networks model aligns Q with I in the first hop, and obtain improved results by adding a second attention hop which chooses visual evidence based on the results of the first hop.

Given the representation of referring expression v_Q and the image feature matrix v_I. We first compute the correlation between v_I and v_Q:

$$C = v_Q \cdot (v_I \cdot W_A + b_A)^T \tag{11}$$

where W_A corresponds to the attention embedding weights of visual features v_I, then compute the attention probability of each image region through a softmax function:

$$p_I = soft\max(W_p \cdot C + b_p) \tag{12}$$

We compute the weighted sum of the image vectors $\tilde{v}_I$ based on p_I, then combine $\tilde{v}_i$ with v_Q to form a refined vector u:

$$\tilde{v}_I = \sum_i p_i v_i \tag{13}$$

$$u = \tilde{v}_I + v_Q \tag{14}$$

where u is considered as a refined vector because it combines both referring expression and the visual information which is related to the final answer. As mentioned above, in complicated case, a single attention layer is usually not enough to find the most relevant region. As a result, we use multiple attention layers to iterate the above process, each layer extracts more fine-grained visual information for prediction. Formally, for the k-th attention layer, we calculate:

$$C^k = v_Q^k \cdot (v_I \cdot W_A^k + b_A^k)^T \tag{15}$$

$$p_I^k = soft\max(W_p^k \cdot C^k + b_p^k) \tag{16}$$

We repeat this process k times, and at the top of the network we infer the result by aggregating the image feature vector:

$$\tilde{v}_I^k = \sum_i p_I^k v_i \tag{17}$$

After selecting the relevant regions, we find the corresponding regions in the original image, and denote them by bounding boxes.

4 Experiments

In this section, we evaluate our model by conducting a series of experiments on two datasets: Visual Genome (VG) dataset [Krishna, Zhu, Groth et al. (2017)] and Flickr30k Entities [Plummer, Wang, Cervantes et al. (2015)]. We compare our model with some approaches proposed recently [Fukui, Huk Park, Yang et al. (2016); Hu, Rohrbach, Andreas et al. (2016); Rohrbach, Rohrbach, Hu et al. (2016)] on referring expression comprehension. In the experiment, we also increase the attention layers gradually to find out the relation between the accuracy and the amount of attention layers in the model.

4.1 Model training and setup

For the image model, a 16 layers VGG network [Simonyan and Zisserman (2014)] is employed to extract image features, and the network is pre-trained on ImageNet 2012 classification challenge dataset [Deng, Dong, Socher et al. (2009)]. We implement our model using Tensorflow deep learning framework [Abadi, Barham, Chen et al. (2016)], which is widely used in computer vision tasks. We use stochastic gradients descent with learning rate ϵ=0.0001 to train the model, dropout is adopted in fully connected layers and after the LSTM layers. The batch size is fixed to be 200.

4.2 Experiments on Visual Genome dataset

VG dataset is proposed for cognitive scene reasoning and understanding tasks, it contains interactions and relationship expressions describe pairs of objects, such as *girl feeding large elephant* and *a man wearing a hat*. Similar to Hu et al. [Hu, Rohrbach, Andreas et al. (2016)], we evaluate our model in two tasks: retrieving the subject-object pair and retrieving the subject region. Given an image and a referring expression, the former task is to locate both the subject and the object and denote it with a bounding box, the latter is to locate the region only corresponding to the subject of the expression.

Table 1: Comparison of our model with previous methods on Visual Genome dataset

Methods	AP-subj	AP-pair
CMNs	44.24%	28.52%
MMI-SoftMax	42.60%	23.70%
Ours:		
One hop	44.50%	25.23%
Two hops	46.43%	27.56%
Three hops	46.32%	27.60%
Four hops	46.44%	27.58%

The predication is correct if the predicted bounding box overlaps with the ground-truth bounding box by more than 50% intersection over union (IoU). We use average precision (AP) and accuracy (AC) to evaluate our proposed model. AP-subj denotes subject regions matching subject grounding-truth, AP-pair denotes both the subject and object regions match the grounding-truth. We compare the performance of our model with CMNs [Hu, Rohrbach, Andreas et al. (2016)] and MMI-SoftMax [Mao, Huang, Toshev et al. (2016)] in Tab. 1. From the results we can see that our method outperforms the previous two methods even using only one attention network layer, showing the advantage of attention network. The reason could be that attention network capable of putting higher weights on the regions that are more relevant to the referring expression. Another observation is that the performance of the model is not getting better with the increase of the attention layers, when it has more than two attention layers the precision hardly improved. This may be due to that two attention layers can get the whole information of the image.

Table 2: Results without using referring expression parsing model

Methods	AP-subj	AP-pair
CMNs	44.24%	28.52%
MMI-SoftMax	42.60%	23.70%
Ours	45.36%	26.74%

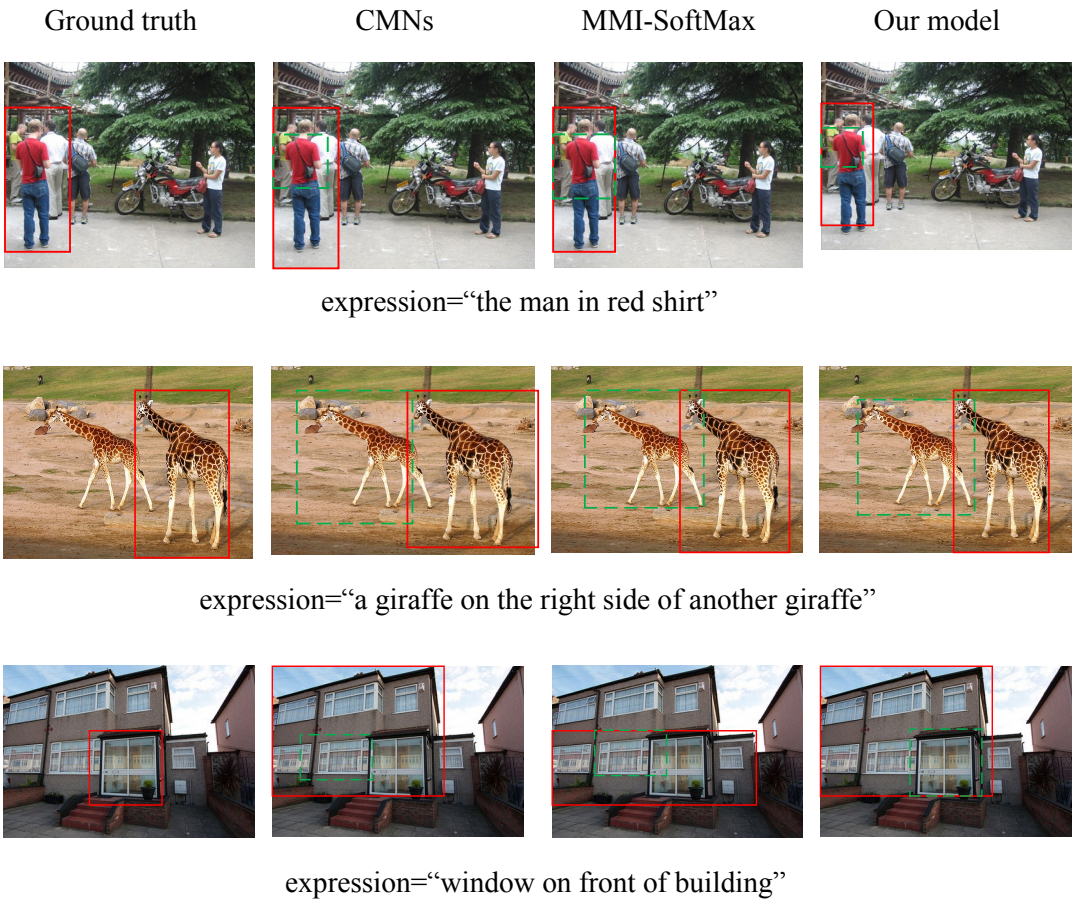

Figure 4: Examples of referring expressions comprehension in the VG dataset

The visualization results of the examples demonstrate that by using multiple attention layers to perform multi-step locating can produce more precise results. For example, consider the expression *window on front of building* in the bottom row of Fig. 4, only our model can locate the subject and object correctly.

4.3 Experiments on Flickr30k Entities

We also apply our model to Flickr30k Entities dataset. Different with Visual Genome in which region-level annotations were produced independently from its captions, and phrases in captions are not linked to these regions. Flickr30k Entities is an extension of Flickr30k dataset [Young, Lai, Hodosh et al. (2014)]. Fig. 5 illustrates two examples of Flickr30k Entities.

Figure 5: Examples of Flickr30k Entities

We randomly chose 6,000 images as the validation set, 6,000 images as the testing set and the remaining images as the training set. Similar to Section 4.2 we compare the performance of our model with CMNs and MMI-SoftMax, and evaluate on this dataset using AP-subj and AC which is measured as percentage of query expressions which have been localized correctly:

Table 3: Comparison of our model with previous methods on Flickr30k Entities

Methods	AC	AP-subj
CMNs	56.18%	43.36%
MMI-SoftMax	59.10%	45.52%
Our model	68.64%	51.48%

From the results we can see that our model outperforms the other two methods both in AC and AP-subj by a large margin. Our experimental results demonstrate the superior performance of the proposed model and the effectiveness of the stacked attention architecture for referring expression comprehension.

5 Conclusions

In this paper, we propose a novel model for referring expression comprehension. Our model uses an attention model to parse the input referring expression into three components: subject, relationship and object, then concatenate them to represent the expression. By using a multiple-layer memory network which queries an image multiple times to locate the relevant visual region. Experimental results on Visual Genome and Flickr30k Entities demonstrate that the proposed model outperforms previous state-of-the-art approaches by a large margin and show the effectiveness of the stacked attention architecture for referring expression comprehension. We also conduct an

ablation experiment to prove that the attention layer is not the more the better, the model performs best when it has two attention layers.

In the future, we will focus on improving the result of referring expression comprehension by using method. To move a forward step, we plan to set foot in the field of exploiting the deeper information of images, e.g., relational reasoning.

Funding Statement: We would like to thank Yongbin Wang of Communication University of China for his suggestions and constructive feedback. This work was supported in part by audio-visual new media laboratory operation and maintenance of Academy of Broadcasting Science, Grant No. 200304, in part by the National Key Research and Development Program of China (Grant No. 2019YFB1406201).

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

Abadi, M.; Barham, P.; Chen, J.; Chen, Z.; Davis, A. et al. (2016): Tensorflow: a system for large-scale machine learning. *12th Symposium on Operating Systems Design and Implementation*, pp. 265-283.

Deng, J.; Dong, W.; Socher, R.; Li, L. J.; Li, K. et al. (2009): Imagenet: a large-scale hierarchical image database. *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248-255.

Fukui, A.; Huk Park, D.; Yang, D.; Rohrbach, A.; Darrell, T. et al. (2016): Multimodal compact bilinear pooling for visual question answering and visual grounding. https://arxiv.org/pdf/1606.01847.pdf.

Girshick, R. (2015): Fast R-CNN. https://arxiv.org/pdf/1504.08083.pdf.

Hu, R.; Rohrbach, M.; Andreas, J.; Darrell, T.; Saenko, K. (2017): Modeling relationships in referential expressions with compositional modular networks. *The IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1115-1124.

Hu, R.; Xu, H.; Rohrbach, M.; Feng, J.; Saenko, K. et al. (2015): Natural language object retrieval. *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4555-4564.

Ilievski, I.; Yan, S.; Feng, J. (2016): A focused dynamic attention model for visual question answering. https://arxiv.org/pdf/1604.01485.pdf.

Krishna, R.; Zhu, Y.; Groth, O.; Johnson, J.; Hata, K. et al. (2017): Visual Genome: connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, vol. 123, no. 1, pp. 32-73.

Krizhevsky, A.; Sutskever, I.; Hinton, G. E. (2012): Imagenet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*, pp. 1097-1105.

Li, Y.; Wang, Y. (2018): A multi-label image classification algorithm based on attention model. *IEEE/ACIS 17th International Conference on Computer and Information Science*, pp. 728-731.

Li, Y.; Wang, Y.; Chen, Z. (2019): Image classification with visual relationship. *IEEE/ACIS 18th International Conference on Computer and Information Science*, pp. 214-219.

Liu, W. J.; Gao, P. P.; Yu, W. B.; Qu, Z. G.; Yang, C. N. (2018): Quantum relief algorithm. *Quantum Information Processing*, vol.17, no. 10, pp. 280.

Liu, W.; Chen, Z. Y.; Liu, J. S.; Su, Z. F.; Chi, L. H. (2018): Full-blind delegating private quantum computation. *Computers, Materials & Continua*, vol. 56, no. 2, pp. 211-224.

Liu, W.; Gao, P.; Wang, Y.; Yu, W.; Zhang, M. (2019): A unitary weights based one-iteration quantum perceptron algorithm for non-ideal training sets. *IEEE Access*, vol. 7, pp. 36854-36865.

Lu, J.; Xiong, C.; Parikh, D.; Socher, R. (2017): Knowing when to look: adaptive attention via a visual sentinel for image captioning. *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 375-383.

Lu, P.; Ji, L.; Zhang, W.; Duan, N.; Zhou, M. et al. (2018): R-VQA: learning visual relation facts with semantic attention for visual question answering. *International Conference on Knowledge Discovery & Data Mining*, pp. 1880-1889.

Luong, M. T.; Pham, H.; Manning, C. D. (2015): Effective approaches to attention-based neural machine translation. https://arxiv.org/pdf/1508.04025.pdf.

Mao, J.; Huang, J.; Toshev, A.; Camburu, O.; Yuille, A. L. et al. (2016): Generation and comprehension of unambiguous object descriptions. *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 11-20.

Nagaraja, V. K.; Morariu, V. I.; Davis, L. S. (2016): Modeling context between objects for referring expression understanding. *European Conference on Computer Vision*, pp. 792-807.

Plummer, B. A.; Wang, L.; Cervantes, C. M.; Caicedo, J. C.; Hockenmaier, J. et al. (2015): Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2641-2649.

Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. (2016): You only look once: Unified, real-time object detection. *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 779-788.

Ren, S.; He, K.; Girshick, R.; Sun, J. (2015): Faster R-CNN: Towards real-time object detection with region proposal networks. *Advances in Neural Information Processing Systems*, pp. 91-99.

Rohrbach, A.; Rohrbach, M.; Hu, R.; Darrell, T.; Schiele, B. (2016): Grounding of textual phrases in images by reconstruction. *European Conference on Computer Vision*, pp. 817-834.

Simonyan, K.; Zisserman, A. (2014): Very deep convolutional networks for large-scale image recognition. https://arxiv.org/pdf/1409.1556.pdf.

Sukhbaatar, S.; Weston, J.; Fergus, R. (2015): End-to-end memory networks. *Advances in Neural Information Processing Systems*, pp. 2440-2448.

Sutskever, I.; Vinyals, O.; Le, Q. V. (2014): Sequence to sequence learning with neural networks. *Advances in Neural Information Processing Systems*, pp. 3104-3112.

Teney, D.; Liu, L.; van den Hengel, A. (2017): Graph-structured representations for visual question answering. *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1-9.

Weston, J.; Chopra, S.; Bordes, A. (2014): Memory networks. https://arxiv.org/pdf/1410.3916.pdf.

Xu, D.; Zhu, Y.; Choy, C. B.; Li, F. (2017): Scene graph generation by iterative message passing. *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5410-5419.

Xu, K.; Ba, J.; Kiros, R.; Cho, K.; Courville, A. et al. (2015): Show, attend and tell: Neural image caption generation with visual attention. *International Conference on Machine Learning*, pp. 2048-2057.

Yang, Z.; He, X.; Gao, J.; Deng, L.; Smola, A. (2016): Stacked attention networks for image question answering. *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 21-29.

Young, P.; Lai, A.; Hodosh, M.; Hockenmaier, J. (2014): From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, vol. 2, pp. 67-78.

Yu, L.; Lin, Z.; Shen, X.; Yang, J.; Lu, X. et al. (2018): Mattnet: Modular attention network for referring expression comprehension. *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1307-1315.

Yu, L.; Poirson, P.; Yang, S.; Berg, A. C.; Berg, T. L. (2016): Modeling context in referring expressions. *European Conference on Computer Vision*, pp. 69-85.

Zhang, H.; Kyaw, Z.; Chang, S. F.; Chua, T. S. (2017): Visual translation embedding network for visual relation detection. *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5532-5540.