RNIC-A Retrospect Network for image captioning

XIU LONG YI
Shandong University of Science and Technology
YOU FU
Shandong University of Science and Technology
DU LEI ZHENG
Shandong University of Science and Technology
XIAO PENG LIU
Shandong University of Science and Technology
RONG HUA (huarong@sdust.edu.cn)
Shandong University of Science and Technology

Research Article

Keywords: LSTM, image caption, visual attention, textual attention, Retrospect

Posted Date: October 20th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-985124/v1

Version of Record: A version of this preprint was published at Soft Computing on January 20th, 2022. See the published version at https://doi.org/10.1007/s00500-021-06622-3.

Noname manuscript No.
(will be inserted by the editor)

RNIC-A Retrospect Network for image captioning

Xiu-Long Yi[1] · RONG HUA[1,*] · You Fu[1] · Du-Lei Zheng[1] · Zhi-Yu Wang[2]

Received: date / Accepted: date

Abstract As cross-domain research combining computer vision and natural language processing, the current image captioning research mainly considers how to improve the visual features, less attention has been paid to utilizing the inherent properties of language to boost captioning performance. Facing this challenge, we proposed a textual attention mechanism, which can obtain semantic relevance between words by scanning all generated words. The Retrospect Network for image captioning(RNIC) proposed in this paper aims to improve input and prediction process by using textual attention. Concretely, the textual attention mechanism is applied to the model simultaneously with the visual attention mechanism to provide the input of the model with the maximum information required for generating captions. In this way, our model can learn to collaboratively attend on both visual and textual features. Moreover, the semantic relevance between words obtained by retrospect is used as the basis for prediction, so that the decoder can simulate the human language system and better make predictions based on the already generated contents. We evaluate the effectiveness of our model on the COCO image captioning datasets and achieve superior performance overthe previous methods.

Keywords LSTM · image caption · visual attention · textual attention · Retrospect

✉ Xiulong Yi
E-mail: yxl1620159241@163.com
✉ Rong Hua
E-mail: huarong@sdust.edu.cn

[1] College of Computer Science and Engineering, Shandong University of Science and Technology, Qingdao 266590 Shandong, China
[2] Huangdao District Central Hospital, Qingdao 266555 Shandong, China

1 Introduction

Image captioning is a fundamental study in computer vision, which aims to identify objects within an image, understand the relationships between objects, and represent them in a natural language that humans can understand. The difficulty of the image caption research is to make the computer 'see' the visible objects and 'understand' the invisible object relationships, which is much more difficult than image classification and object detection [1–3]. Because of its remarkable role in image/video retrieval and assisting visually impaired groups to perceive their environment, image captioning has attracted wide interest from academia [4, 5] and industry [6, 7].

In recent years, attention mechanism is widely used on various tasks [8–10], which only focuses on selective parts of the whole visual space when and where as needed. However, as cross-domain research combining computer vision and natural language processing, relying on visual features alone is still not sufficient to generate high-quality captions, textual information is also crucial for improving model performance. State-of-the-art image caption models with Long short-term memory (LSTM[11]) as the decoder is too simple in its utilization of textual information. There are two manifestations in the model. Firstly, the decoder only uses adjacent textual information as the input, and more textual information is passed through the memory unit of the LSTM, which is not effective in dealing with long-term dependency problems. As shown in 1, when 'paddle' is to be predicted, information about "surfing" is not well transferred to that moment because the interval is too long. The second is that the semantic correlation between words is ignored in the final prediction process, and the inherent properties of language cannot be exploited to improve the performance of the model.

In this paper, following the conventional encoder-decoder framework, we propose the RNIC model, which can improve the input and prediction process. Different from previous methods which boost captioning performance by improving the visual attention mechnism, our RNIC applied attention in both visual and textual domain.

Fig. 1: The attention weight distribution over the past generated words is shown when predicting the word'paddle'.The thicker line indicates arelatively larger weight.

The main contributions of this paper are as follows:

1. In response to the problem that the current mainstream models overemphasize how to improve visual features, we propose textual attention mechanism for image captioning. The textual attention mechanism allows the model to trace back to the text information that is most relevant to the current moment prediction.

2. We explore the role of textual attention mechanism - it can effectively improve model input and prediction.

3. RNIC model applies both textual attention and visual attention to the model, so that the model is able to make predictions based on what has been generated, and the model's ability to handle long-term dependencies is significantly enhanced.

2 Related work

Visual Attention. Xu et al.[12]first introduced attention mechanism into image captioning,which generates a matrix to weight each receptive field in the encoded feature map. Instead of only attending to the receptive field in the encoded feature map, Chen et al.[13] added a feature channel attention module.Lu et al.[14]proposed adaptive attention,which adaptively decides when and where to rely on the visual information.In order to solve the problem that above models lack accurate positioning of informative regions in the original image,Anderson et al.[15]proposed bottom-up and top-down attention mechanism,where bottom-up attention first uses object detection models to detect multiple informative regions in the image, then top-down attention attends to the most relevant detected regions when generating a word.Yao et al.[16]injected a graph convolutional neural network to relate detected informative regions, and therefore refine their features before feeding into the decoder.

Textual Attention. Though no prior work has explored textual attention in image caption, there are some related works in natural language processing. In [38], the author propose RNNSearch to learn an alignment over the input sentences. Tim et al.[39] propose a more fine-grained attention mechanism to reason about the entailment in two sentences. Yin et al. [40] propose an attention-based bigram CNN for jointly performing attention between two CNN hierarchies.

Visual Attention with Textual Attention in Visual Question Answering (VQA). In the VQA task, Hyeonseob Nam et al.[17] combined visual attention with textual attention to capture the fine-grained interactions between vision and text, and by focusing on specific regions in images and text to gather the necessary information. Lu J, Yang et al.[18] proposed Co-Attention to make the model focus on different regions of the image as well as different segments of the text (questions), and model the text at three levels to capture different granularity of information.

Unlike the VQA task, the image caption research is a language-generating process and only the generated textual information is known at the time of prediction. Lei et al.[19] proposed Reflective Attention, which combines textual attention with visual attention for the first time and applies it to the image caption research. The Reflective Attention calculates the attention of the hidden units for all moments and uses the results as a basis for prediction. In this paper, we propose a more direct textual attention mechanism by calculating the similarity between hidden units and generated words to obtain the required textual information at that time step, and use it to improve the input and prediction of the model.

3 Method

We adopt the popular encoder-decoder framework for image caption generation. Our model (see Figure 2 for the model structure) takes a single raw image and generates a caption S encoded as a sequence of 1-of-k encoded words:

$$S = S_1, ..., S_n, S_i \in \mathbb{R}^k \tag{1}$$

where k is the size of the dictionary and n is the length of the caption.

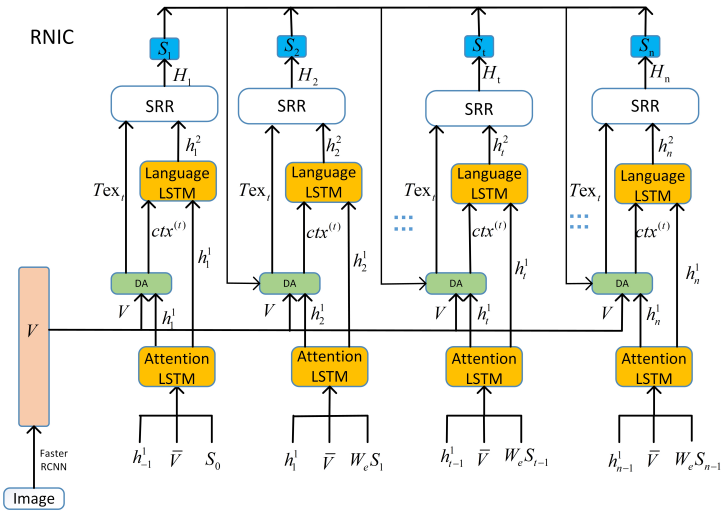

Fig. 2: The overall architecture of RNIC

As shown in Figure2, the RNIC model proposed in this paper is implemented based on a two-layer LSTM, with the first LSTM as the attention model and the second LSTM as the language model. Dual Attention Module is used to generate a joint context vector to maximize the visual and textual information needed to generate caption. Semantic Relevance Retrospect

Module allows the model to better predict based on the generated content, which better simulates the human language system. The two modules make the model significantly more capable of handling long-term dependencies.

3.1 Object-Level Encoder

To generate captions, the first step is to extract the visual features from images. In this paper, the visual features V of an image are extracted by a pre-trained Faster RCNN. The extractor generates L region vectors V_i, each region vector is a D-dimensional representation corresponding to a part of the image:

$$V = V_1, ..., V_L, V_i \in \mathbb{R}^D \tag{2}$$

Compared to the convensional uniform meshing method on CNN features, the object-level encoder focuses more on objects in an image that is closely related to the perception mechnism in human visual system.

3.2 Retrospect Decoder

Given a set of region image features V proposed by encoder, the goal for the retrospect decoder is to generate the caption S. The generated caption should not only capture the content information from the image but also be meaningful and coherent. Similar to[15], retrospect decoder contains two layer LSTM, with the first LSTM as the attention model and the second LSTM as the language model.

The input vector to the Attention LSTM at each time step consists of the previous output of the Language LSTM, concatenated with the mean-pooled image feature $\bar{V} = \frac{1}{L}\sum_{i=1}^{L} V_i$ and encoding of the previously generated word, given by:

$$x_t^1 = [h_{t-1}^2, \bar{V}, W_e S_{t-1}] \tag{3}$$

where $W_e \in \mathbb{R}^{E\times|\mathbb{Z}|}$ represents the word embedding matrix, and S_{t-1} is the output vocabulary at time step t-1, represented by the one-hot vector.

The input vector to the Language LSTM at each time step consists of the output of the Dual Attention module concatenated with the output of Attention LSTM, given by:

$$x_t^2 = [ctx^{(t)}, h_t^1] \tag{4}$$

where the joint context vector $ctx^{(t)}$ is generated by the proposed Dual Attention Module.

At each time step t the conditional distribution over possible output words is given by:

$$p(S_t|S_1, ..., S_{t-1}) = softmax(W_p H_t + b_p) \tag{5}$$

where, $W_p \in \mathbb{R}^{|Z|\times M}$,$b_p \in \mathbb{R}^{|Z|}$ are parameters that need to be learned. H_t is generated by the Semantic Relevance Retrospect Module.

3.3 Dual Attention Module

Previous works have shown that visual attention alone can perform fairly well for localizing objects and aiding caption generation. However, as cross-domain research combining computer vision and natural language processing, relying on visual features alone is still not sufficient to generate high-quality captions, textual information is also crucial for improving model performance. To this end, we propose Dual Attention module which can simultaneously attend to visual and textual modalities. The structure is illustrated in Figure3.

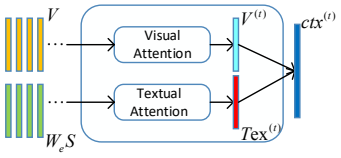

Fig. 3: Dual Attention Module Structure Diagram

As shown in Figure3, The application of the Dual Attention Module allows the model to learn to collaborate on visual and textual features by generating a joint context vector $ctx^{(t)}$:

$$ctx^{(t)} = [V^{(t)}; Tex^{(t)}] \tag{6}$$

where $V^{(t)}$, $Tex^{(t)}$ represent the results of visual attention and textual attention respectively.

Visual Attention. Visual attention aims to generate a vector by attending to certain parts of the input image. At time step t, given the output of the Attention LSTM h_t^1, the visual context vector $V^{(t)}$ is generated by:

$$a_{t,i} = W_\alpha^T \tanh(W_{v\alpha} V_i + W_{h\alpha} h_t^1) \tag{7}$$

$$\alpha_t = softmax(a_t) \tag{8}$$

$$V^{(t)} = \sum_{i=1}^{L} V_i \alpha_{t,i} \tag{9}$$

where, $W_\alpha \in \mathbb{R}^H$,$W_{V\alpha} \in \mathbb{R}^{H\times V}$,$W_{h\alpha} \in \mathbb{R}^{H\times M}$ are the parameters to be learned.

Textual Attention. To make better use of the inherent properties of language, we propose the textual attention mechanism. To our knowledge, this is the first work exploring textual attention in image captioning. The textual attention mechanism can review all the generated words at each time step and extract important information to guide the prediction process. At time step t,

given the output of the Attention LSTM h_t^1, the text context vector $Tex^{(t)}$ is generated by:

$$B_{t,i} = W_\beta^T \tanh(W_{S\beta} S_i + W_{h\beta} h_t^1) \tag{10}$$

$$\beta_t = softmax(B_t) \tag{11}$$

$$Tex^{(t)} = \sum_{i=1}^{t-1} S_i \beta_{t,i} \tag{12}$$

3.4 Semantic Relevance Retrospect Module

State-of-the-art image captioning methods mostly uses the hidden state alone to generate captions. In this way, the historical sequence information cant be used well. Our Semantic Relevance Retrospect Module models the dependencies between pairs of words at different time steps. The structure of the Semantic Relevance Retrospect Module is illustrated in Figure4.

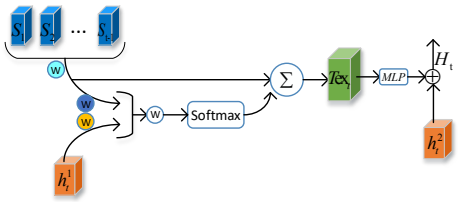

Fig. 4: SRR Module structure Diagram

As shown in Figure 4, the SRR module puts the results of the text attention mechanism through the multilayer perceptron before summing with the hidden state as the basis for the model prediction. The role of the multilayer perceptron is twofold: one is to match the results of the text attention mechanism with the hidden state dimension, and the other is to enable the model to further explore the semantic relatedness between words.

In this way, the probability of the output words at timestep t is calculated as follows:

$$Tex_t^{'} = MLP(Tex_t) \tag{13}$$

$$H_t = h_t^2 + Tex_t^{'} \tag{14}$$

$$p(S_t|S_1, ..., S_{t-1}) = softmax(W_p H_t + b_p) \tag{15}$$

where, $W_p \in \mathbb{R}^{|Z| \times M}$,$b_p \in \mathbb{R}^{|Z|}$ are parameters that need to be learned.

The SRR module uses the semantic correlation joint hidden state as the basis for prediction, which enables the decoder to better reason and predict based on the already generated content.

4 Experiments

4.1 Datasets and Evaluation Metrics

We evaluate our model on the MS-COCO dataset[20]. MS-COCO dataset contains 123,287 images labeled with 5 captions for each. We follow the splits provided by Karpathy et al.[21], where 5000 images are used for validation, 5000 for testing and the rest for training. Following most image caption research, we use different metrics, including BLEU[22], METEOR[23], ROUGE-L[24] and CIDEr[25] to evaluate the proposed method and compared with other methods. For simplicity, B-n is used to denote the n-gram BLEU score and M, R, C are used to represent METEOR, ROUGE-L,CIDEr, respectively. All of the above evaluation metrics evaluate the performance of the model by measuring the similarity between the generated and labeled sentences.

4.2 Implementation Details

To represent image regions, we employ a pre-trained Faster-RCNN[26] model on ImageNet[27] and Visual Genome[28]. The dimension of the original vectors is 2048 and we project them to a new space with the dimension of 1024, which is also the hidden size of the LSTM in the decoder. To represent words, we drop the words that occur less than 5 times and end up with a vocabulary of 9945 words. We use one-hot vectors and linearly project them to dimension of 1024. As for training process, we first train RNIC model under XE loss for 25 epochs with the learning rate set to 5e-4, then we optimize the CIDEr-D score with SCST[29] for another 15 epochs with the learning rate set to 5e-5.

4.3 Experiment Results

We report the performance on the MS-COCO Karpathy test split of our model as well as the compared models in Table1. The models include: Stack-VS Attention[30], which proposes a visual-semantic attention based multi-stage framework; GCN-LSTM[16], which explores visual relationship for boosting image captioning; LBPF[31], which can embed previous visual information and look into future; SGAE[32], which introduce auto-encoding scene graphs into its model; ORT[33], which takes into account geometric information in the encoder phase; MAD+SAP[34], which demonstrate that selecting appropriate-subsequent attributes to attend to is beneficial for imagecaptioning models; AoANet[35], which extends the conventional attention mechanisms to determine the relevance between attention resultsand queries; ETA[36], which extends the Transformer model to exploit complementary information of visual

Table 1: Performance of our model and other models on MS-COCO Karpathy test split,All results are reported after the reinforce optimization stage.

Model	B-1	B-2	B-3	B-4	M	R	C
Baseline[15]	79.8	-	-	36.3	27.7	56.9	120.1
Stack-VS Attention[30]	79.4	63.6	49.0	37.2	27.9	57.7	122.6
GCN-LSTM[16]	80.5	-	-	38.2	28.5	58.3	127.6
LBPF[31]	80.5	-	-	38.3	28.5	58.4	127.6
SGAE[32]	80.8	-	-	38.4	28.4	58.6	127.8
ORT[33]	80.5	-	-	38.6	28.7	58.4	128.3
MAD+SAP[34]	-	-	-	38.6	28.7	58.5	128.8
AoANet[35]	80.2	-	-	38.9	29.2	58.8	129.8
ETA[36]	81.5	-	-	39.3	28.8	58.9	126.6
X-Transformer[37]	80.9	65.8	51.5	**39.7**	29.5	59.1	132.8
RNIC(single)	80.3	64.8	50.4	38.5	29.2	58.8	130.1
RNIC(Ensemble)	**81.5**	**66.0**	**51.7**	39.4	**29.7**	**59.4**	**133.2**

regions and semantic attributes simultaneously; X-Transformer[37], which employs bilinear pooling to selectively capitalize on visual information or perform multi-modal reasoning.

As can be seen from Table 1, the RNIC model has a significant improvement over the baseline, which is because the beseline mainly considers how to improve visual features and do not make enough use of textual information. Visual attention-based models rely only on the memory units of LSTM to utilize the generated textual information, which is not satisfactory when the sentence is too long. Our model uses the textual attention mechanism to improve the input and prediction of the model so that the model can learn to collaborate on visual features and textual features, and uses the semantic correlation between words for prediction. The use of the textual attention mechanism enables our model to make better predictions based on the generated text information.

4.4 Ablation Study and Analysis

To verify the effects of the two modules DA and SRR of our model, the ablation experiments are designed as follows: (1) Baseline: represents the model without both DA and SRR modules; (2) DA represents the removal of the SRR module and keeping only the DA module; (3) SRR represents the removal of the DA module; (4) RNIC represents the removal of the DA module and the SRR modules are applied to the model at the same time. The experimental results are shown in Table 2.

As can be seen from Table2, both the DA module and the SRR module are important for the improvement of the model performance, and both modules improve all the metrics compared to Baseline. This proves that the textual information is crucial for the prediction of the model and indeed solves to some extent the problem that relying only on the memory units of the LSTM for using textual information is not enough to generate high-quality captions.

Table 2: Ablation study.All results are reported after the reinforce optimization stage

Model	B-1	B-2	B-3	B-4	M	R	C
Baseline[7]	79.8	-	-	36.3	27.7	56.9	120.1
DA	80.1	64.5	49.9	38.3	28.9	58.6	129.9
SRR	80.0	64.2	49.8	38.0	28.8	58.5	129.2
RNIC(single)	80.3	64.8	50.4	38.5	29.2	58.8	130.1

To better show the specific difference in the image captioning prediction results of our model with the ablation study section, we visualized some of generated captions on the COCO dataset in Figure 5.

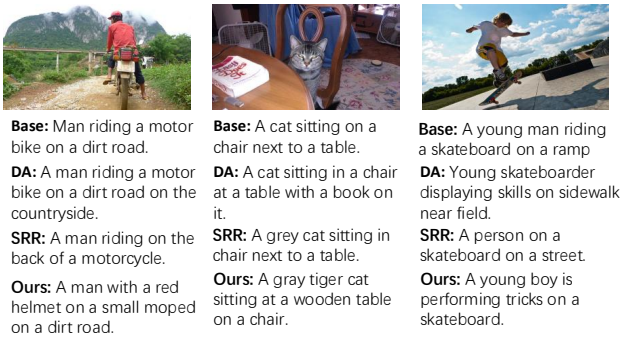

Fig. 5: Examples of captions generated by our approach and the original model, as well as the ablation study section.

From the Figure 5 we can see, on average, our model is able to generate more accurate and descriptive captions.

4.5 Textual Attention weight visualization and Analysis

To better understand and illustrate our model, we visualize how the RNIC model makes inferences and predictions based on the words that have been generated, as shown in Figure 6. Take the first case in Figure 6 as an example, when it comes to predicting 'water', 'boat' can play a very important role in the prediction. The textual attention mechanism allows our model to trace back to the textual information most relevant to the prediction and act on both the input and prediction aspects of the model, thus improving the performance of the model.

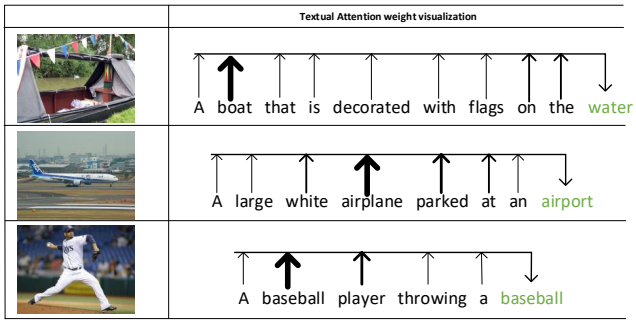

Fig. 6: Examples of captions and textual attention weight visualization generated by RNIC. The thicker line indicates a relatively larger weight,and the word to be predicted is highlighted in green.

4.6 complexity and efficiency Analysis

Compared to visual attention-based models, RNIC Added calculation of text attention mechanism. Because there is no dependency between visual attention and textual attention, it can be carried out in parallel. This allows the RNIC model to achieve better performance without reducing computational efficiency.

5 Conclusion

In this paper, we devise RNIC model for image captioning. By introducing the textual attention, the original visual attention based model is extened to learn on both visual and textual information to maximize the information needed for generating captions. Our model can better mimic the human language system - making predictions based on what has been generated.Moreover, comprehensive comparisions with state-of-the art methods and adequate ablation studies demonstrate the effectiveness of our framework. In future work, we intend to apply textual attention in RNIC to video captioning. We also explore how to incorporate textual attention mechanism with Transformer framework.

Author Contributions Xiulong Yi proposed the method and conducted the experiments, analysed the data and wrote the manuscript. Rong Hua supervised the project and participated in manuscript revisions. You Fu, Dulei Zheng, Zhiyu Wang provided critical reviews that helped improve the manuscript.

Funding This study was funded by National key research and development project(2017YFB0202002)

Ethical approval This article does not contain any studies with human participants or animals performed by any of the authors.

Informed consent Informed consent was obtained from all individual participants included in the study.

6 Declarations

Conflict of Interest Author XIU-LONG YI, RONG HUA ,YOU FU, DU-LEI ZHENG, Zhi-Yu Wang declare that they have no conflict of interest.

References

1. Liu T, Liu H, Li Y F, et al. Flexible FTIR spectral imaging enhancement for industrial robot infrared vision sensing[J]. IEEE Transactions on Industrial Informatics, 2019, 16(1): 544-554.
2. Liu T, Liu H, Li Y, et al. Efficient blind signal reconstruction with wavelet transforms regularization for educational robot infrared vision sensing[J]. IEEE/ASME Transactions on Mechatronics, 2018, 24(1): 384-394.
3. Liu T, Liu H, Chen Z, et al. Fast blind instrument function estimation method for industrial infrared spectrometers[J]. IEEE Transactions on Industrial Informatics, 2018, 14(12): 5268-5277.
4. Liu H, Fang S, Zhang Z, et al. MFDNet: Collaborative Poses Perception and Matrix Fisher Distribution for Head Pose Estimation[J]. IEEE Transactions on Multimedia, 2021.
5. Liu H, Nie H, Zhang Z, et al. Anisotropic angle distribution learning for head pose estimation and attention understanding in human-computer interaction[J]. Neurocomputing, 2021, 433: 310-322.
6. Cornia M, Stefanini M, Baraldi L, et al. Meshed-memory transformer for image captioning[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 10578-10587.
7. Ji J, Luo Y, Sun X, et al. Improving image captioning by leveraging intra-and inter-layer global representation in transformer network[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2021, 35(2): 1655-1663.
8. Huang Q, Zhang Y, Peng H, et al. Deep subspace clustering to achieve jointly latent feature extraction and discriminative learning[J]. Neurocomputing, 2020, 404: 340-350.
9. Fan Z, Dan T, Yu H, et al. Single Fundus Image Super-Resolution Via Cascaded Channel-Wise Attention Network[C]//2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC). IEEE, 2020: 1984-1987.
10. Huang J, Zhuo E, Li H, et al. Achieving accurate segmentation of nasopharyngeal carcinoma in mr images through recurrent attention[C]//International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, Cham, 2019: 494-502.
11. Hochreiter S, Schmidhuber J. Long short-term memory[J]. Neural computation, 1997, 9(8): 1735-1780.

12. Xu K, Ba J, Kiros R, et al. Show, attend and tell: Neural image caption generation with visual attention[C]//International conference on machine learning. PMLR, 2015: 2048-2057.
13. Chen L, Zhang H, Xiao J, et al. Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 5659-5667.
14. Lu J, Xiong C, Parikh D, et al. Knowing when to look: Adaptive attention via a visual sentinel for image captioning[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 375-383.
15. Anderson P, He X, Buehler C, et al. Bottom-up and top-down attention for image captioning and visual question answering[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 6077-6086.
16. Yao T, Pan Y, Li Y, et al. Exploring visual relationship for image captioning[C]//Proceedings of the European conference on computer vision (ECCV). 2018: 684-699.
17. Nam H, Ha J W, Kim J. Dual attention networks for multimodal reasoning and matching[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 299-307.
18. Lu J, Yang J, Batra D, et al. Hierarchical question-image co-attention for visual question answering[J]. Advances in neural information processing systems, 2016, 29: 289-297.
19. Ke L, Pei W, Li R, et al. Reflective decoding network for image captioning[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019: 8888-8897.
20. Chen X, Fang H, Lin T Y, et al. Microsoft coco captions: Data collection and evaluation server[J]. arXiv preprint arXiv:1504.00325, 2015.
21. Karpathy A, Fei-Fei L. Deep visual-semantic alignments for generating image descriptions[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2015: 3128-3137.
22. Papineni K, Roukos S, Ward T, et al. Bleu: a method for automatic evaluation of machine translation[C]//Proceedings of the 40th annual meeting of the Association for Computational Linguistics. 2002: 311-318.
23. Banerjee S, Lavie A. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments[C]//Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization. 2005: 65-72.
24. Lin C Y. Rouge: A package for automatic evaluation of summaries[C]//Text summarization branches out. 2004: 74-81.
25. Vedantam R, Lawrence Zitnick C, Parikh D. Cider: Consensus-based image description evaluation[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2015: 4566-4575.
26. Ren S, He K, Girshick R, et al. Faster r-cnn: Towards real-time object detection with region proposal networks[J]. Advances in neural information processing systems, 2015, 28: 91-99.

27. Deng J, Dong W, Socher R, et al. Imagenet: A large-scale hierarchical image database[C]//2009 IEEE conference on computer vision and pattern recognition. Ieee, 2009: 248-255.
28. Krishna R, Zhu Y, Groth O, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations[J]. arXiv preprint arXiv:1602.07332, 2016.
29. Rennie S J, Marcheret E, Mroueh Y, et al. Self-critical sequence training for image captioning[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 7008-7024.
30. Cheng L, Wei W, Mao X, et al. Stack-VS: Stacked Visual-Semantic Attention for Image Caption Generation[J]. IEEE Access, 2020, 8: 154953-154965.
31. Qin Y, Du J, Zhang Y, et al. Look back and predict forward in image captioning[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 8367-8375.
32. Yang X, Tang K, Zhang H, et al. Auto-encoding scene graphs for image captioning[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 10685-10694.
33. Herdade S, Kappeler A, Boakye K, et al. Image captioning: Transforming objects into words[J]. arXiv preprint arXiv:1906.05963, 2019.
34. Huang Y, Chen J, Ouyang W, et al. Image captioning with end-to-end attribute detection and subsequent attributes prediction[J]. IEEE Transactions on Image Processing, 2020, 29: 4013-4026.
35. Huang L, Wang W, Chen J, et al. Attention on attention for image captioning[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019: 4634-4643.
36. Li G, Zhu L, Liu P, et al. Entangled transformer for image captioning[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019: 8928-8937.
37. Pan Y, Yao T, Li Y, et al. X-linear attention networks for image captioning[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 10971-10980.
38. Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate[J]. arXiv preprint arXiv:1409.0473, 2014.
39. Rocktäschel T, Grefenstette E, Hermann K M, et al. Reasoning about entailment with neural attention[J]. arXiv preprint arXiv:1509.06664, 2015.
40. Yin W, Schütze H, Xiang B, et al. Abcnn: Attention-based convolutional neural network for modeling sentence pairs[J]. Transactions of the Association for Computational Linguistics, 2016, 4: 259-272.