Intra-Image Region Context for Image Captioning

Shihao Wang, Hong Mo, Yue Xu, Wei Wu, and Zhong Zhou(✉)

State Key Laboratory of Virtual Reality Technology and Systems,
Beihang University, Beijing 100191, China
zz@buaa.edu.cn

Abstract. Image captioning is a challenging task involving computer vision and natural language processing. In recent works, visual attention mechanisms have been extensively used. However, they consider little about the correlations among different regions and the attention on regions. This paper is try to make up for the deficiencies in existing approaches and propose a novel captioning model, which extracts the salient region correlations from the image feature, synthesizes intra-image regions' context, and automatically distributes an appropriate attention over regions. The Intra-Image Region Context (IIRC) model proposed in this paper jointly learns regions' semantic correlations in one image. It consists of two main parts. The first is to extract feature vectors of image through convolutional neural work (CNN) and get correlations among regions from feature vectors by recurrent neural network (RNN). The second is to generate the caption according to the synthesis of region contexts from the first network with attention on different region contexts. The model and baseline are evaluated on MSCOCO test server. The experimental results have illustrated that the model is superior over many outstanding models on the metrics of BLEU, METEOR, ROUGE-L and CIDEr. Moreover, the model excels in describing details, especially those related to position and action.

Keywords: Image captioning · Intra-image region · Regions correlations

1 Introduction

Image captioning is a fundamental research issue which aims at automatically generating a natural description of an image. It has received a significant amount of attention in both computer vision and natural language processing research communities [1, 2]. The image captioning's task is to generate semantically and syntactically appropriate target sentence with consecutive words to represent the image content, which can be quite challenging in two ways. First of all, the model needs to learn and capture the semantic information of image with great precision. Secondly, the generation of the target sentence must take into account both the correctness of the syntax and the correlation between the semantics and the image content, which thus requires complex interactions among them.

In recent years, many approaches which achieve impressive results on image captioning [10, 11, 22] have been raised with the availability of larger datasets [3, 4, 9].

R. Hong et al. (Eds.): PCM 2018, LNCS 11166, pp. 212–222, 2018.
https://doi.org/10.1007/978-3-030-00764-5_20

Particularly, a strong and effective approach was proposed to generate captions in high quality [11]. The image features are encoded by the input image with a deep convolutional neural work (CNN), then the encoded feature is used to generate the output caption by the Long Short Term Memory (LSTM) recurrent neural network (RNN) decoder. This encoder-decoder model becomes the baseline of recent research methods.

To improve the quality of the output captions and help the decoder focus on the key image information, the model needs to perform some fine-grained visual processing. Therefore, visual attention mechanisms have been widely applied in image captioning tasks [12, 17, 22]. Most traditional visual attention mechanisms used in image captioning are the top-down variety. These mechanisms are generally trained to selectively attend to the output of one or more layers of a CNN [16, 22]. However, they give little consideration to how the image regions which are subject to attention are selected, and how those different image regions are related with each other.

In this paper we propose a model based on encoder-decoder architecture, which allows the network to generate captions by the correlation of context among image regions. Our mechanism extracts several major regions of the image feature as region features, with each region feature represented by a pooled feature vector. Then we form a sequence including these features in order, and use RNN encoder to read each region feature sequentially. That is encoder maps these image regions sequence into a continuous feature vectors. We call this intra-image region context modelling, which considers the correlations among image regions. After that, the decoder transforms the continuous feature vectors from the encoder to a sequence as the output sentence. Both the encoder and the decoder adopt the LSTM as recurrent neuron. This process with sequence-to-sequence encoding and decoding enables correlation learning of region features in the model. Allowing the decoder units to determine which region features is more helpful and important for each time step, we introduce the non-visual attention mechanism into this framework. In this way, the model can use the context to predict the attention distribution over regions.

In order to evaluate the performance of our method, our model is trained and tested on MSCOCO caption dataset [9] and MSCOCO test server. MSCOCO is a large and popular dataset containing more than 120,000 images. Our results on the test server not only achieve remarkable performance at CIDEr, METEOR, ROUGE-L and BLEU scores, but also outperform current baseline. The scores of evaluation metrics thoroughly reflect the effectiveness of our model.

2 Proposed Model

Our Intra-Image Region Context model consists of two major subnetworks components: intra-image region context (Fig. 1(a)) and language decoder (Fig. 1(b)). The attention module is used between subnetworks. Different from CNN-RNN encoder-decoder architecture like show-and-tell model [11], we use the RNN-RNN encoder-decoder just like sequence-to-sequence model [2]. When our model gets the target image, it will first extract an image feature by the deep CNN model. The center region, top left region, top right region, bottom left region, bottom right region and entire region of the image feature map are serialized, and then they are put into the LSTM encoder to extract the correlations of different region features. After that, the region context feature which is produced at the last time step of LSTM encoder will be put into the first time step to LSTM decoder to generate the caption for the image. To overcome the loss of regional semantic information without fine-grained localisation, we introduce an attention mechanism to our model. Attention mechanism can focus on the most relevant sections of the input region feature vectors sequence and guide our decoder to those sections for feature extraction. An illustration of our complete model for image captioning is provided in Fig. 1.

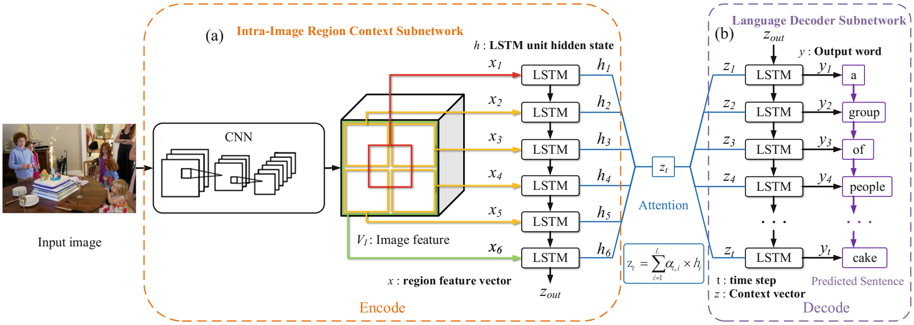


Fig. 1. An overview of the proposed model for image captioning by intra-image region context.

2.1 Intra-Image Region Context Subnetwork

The Intra-Image Region Context subnetwork includes a deep CNN (e.g. Inception-v4 [7]) and a RNN with LSTM recurrent neuron, as illustrated in Fig. 1(a). The CNN has already been well pre-trained on ImageNet [5] which is a large dataset for image classification mission. The pre-trained network which has a well generalization capability, has already learnt the ability of how to get some useful features. The transfer learning is widely used in lots of computer vision tasks. In our method, we use the well pre-trained CNN to extract the semantic feature of the full image. We get the feature from the last layer before pooling layer and full-connected layer. The feature map V_i will be with the shape like (V_h, V_w, V_c), where V_h, V_w, V_c represents height, width, and channel of the feature separately.

Then we divide the feature map into 5 pieces, which have same shapes of (V_h/2, V_w/2, V_c). We get 5 different important regions of feature map as illustrated in Fig. 2. Their directions are: center, top left, top right, bottom left, bottom left and bottom right. These parts map the semantic information of the corresponding areas of the source image. The attention in the human visual system is able to be focused automatically by top-down and bottom-up signals [18, 19]. When a person wants to observe what the picture is talking about, he usually focuses his attention on the center area of the image first to get the main semantics information of the image. Then he will look at the remaining area of the image to obtain some scene of other semantic information, so that he can provide the caption for the image. Drew on the experience of the method of human observation, we simplify the surrounding area as four corners of image feature map, and center area as the center of image feature map. For keeping the full image semantic information joining the generation of caption, we pool the whole feature map to the size of 5 splits of image feature map. For each feature vector of 6 regions which are mentioned above, we flatten and full-connect them to the m-dimensional feature vector with the length as the number of hidden units. Then, we form the region feature vectors x_n sequence $X = (x_1, \ldots, x_6)$ in order: center, top left, top right, bottom left, bottom right, full.

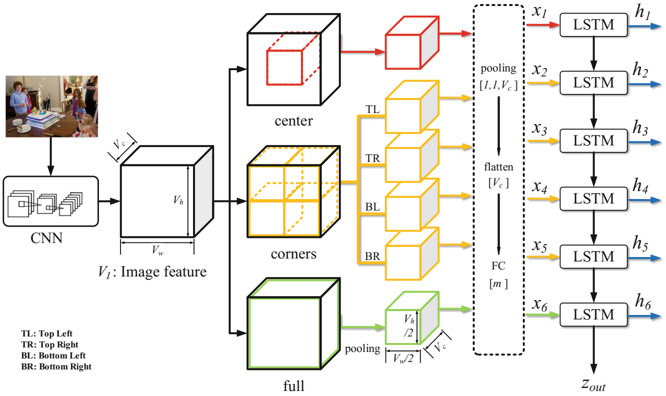


Fig. 2. Intra-Image Region Context Subnetwork. It utilizes the regions of image feature to extract the contexts and connection between them, then puts the region feature into LSTM to get hidden status and the output of our encoder.

LSTM encoder reads region features sequentially, and gives final output z_{out} which represents the context of 6 region semantic features. Recurrent neural network is effective for modeling sequence data. In theory [6], RNN could handle long-term dependencies, but actually it can only remember the limited contents of time steps due to problems of gradient vanishing and explosion. To address this problem, a special RNN neuron called LSTM is proposed and it establishes the state-of-the-art for the sequence task. Therefore, we feed region features sequence X into LSTM. Particularly, at each time step t, the LSTM updates states using the input x_t, previous status $h_{t\text{-}1}$ and $c_{t\text{-}1}$, as follows:

$$f_t = \sigma\left(W_{fh}h_{t-1} + W_{fx}x_t + b_f\right) \tag{1}$$

$$i_t = \sigma(W_{ih}h_{t-1} + W_{ix}x_t + b_i) \tag{2}$$

$$o_t = \sigma(W_{oh}h_{t-1} + W_{ox}x_t + b_o) \tag{3}$$

$$g_t = tanh\left(W_{gh}h_{t-1} + W_{gx}x_t + b_g\right) \tag{4}$$

$$c_t = f_t \odot c_{t-1} + i_t \odot g_t \tag{5}$$

$$h_t = o_t \odot \tanh(c_t) \tag{6}$$

where all of $\sigma(\cdot)$ refer to the sigmoid function, $tanh(\cdot)$ the hyperbolic tangent function, $\odot$ the operator of element-wise product. The LSTM has five components, four gates and one memory cell: forget gate f, input gate i, output gate o, input modulation gate g, memory cell c_t, with the learned parameters Ws, bs. The cell c_t depends on $c_{t\text{-}1}$ which is the previous memory cell, adjusted by forget gate f_t, and g_t adjusted by input gate i_t. Therefore, LSTM not only can solve the problems of gradient vanishing and explosion, but also is able to capture complex and long-term dynamics or dependency in sequence data. Importantly, this allows the model to selectively extract and encode the spatial and semantical dependency among different regions of the image feature. As Fig. 1(a) shown, the LSTM take sequentially an element x_t of X at each time step t. Then, it updates its single hidden state h_t of step t as:

$$h_t = f_\lambda(h_{t-1}, x_t) \tag{7}$$

where f_λ represents the non-linear activation function of parameter λ. After six time steps we will have $h_t(t = 1, \ldots, 6)$, the hidden states, and z_{out}, the comprehensive of region contexts of the image feature.

2.2 Language Decoder Subnetwork

To model the potential high-level region semantic correlation subject to learning a caption sequence generator, we construct a LSTM decoder. Specifically, the LSTM decoder aims at modeling sequential recurrent regions correlation within both intra-image region context z and the comprehensive of region contexts z_{out} and generation dynamic length output as predicted sequence of words y_t over time step t. This is our purpose because of varying co-occurring semantic attributes among regions of the feature. The appropriate caption of the image will be generated from pretreatment list of words. The Language Decoder subnetwork is shown in Fig. 1(b), which consists of one LSTM decoder and attention module. In order to obtain the initial hidden state h_1^2 of decoder, we use the comprehensive of region contexts vector z_{out} to initialize it. This step is for the purpose of incorporating the intra-image region context correlation into the decoding procedure. Different from the encoder, when we infer a caption, the output word and hidden state of time step t, $y_{t\text{-}1}$ and h_t^2 rely on the previous $h_{t\text{-}1}^2$ and $z_{t\text{-}1}$, which is initialized by the start token of words (e.g. "<S>"). In fundamental, our model is able to mine the potential high-level region semantic correlation of dynamic sequence precisely because of this recurrent feedback connection in sequence. Different from Eq. (7), h_t^2 is update as follows:

$$h_t^2 = f_\lambda\left(h_{t-1}^2, z_{t-1}\right) \tag{8}$$

Similar to Eq. (1)–Eq. (6), the gates and cells of the decoder LSTM update as following:

$$f_t = \sigma\left(W_{fh}h_{t-1}^2 + W_{fz}z_{t-1} + b_f\right) \tag{9}$$

$$i_t = \sigma\left(W_{ih}h_{t-1}^2 + W_{iz}z_{t-1} + b_i\right) \tag{10}$$

$$o_t = \sigma\left(W_{oh}h_{t-1}^2 + W_{oz}z_{t-1} + b_o\right) \tag{11}$$

$$g_t = tanh\left(W_{gh}h_{t-1}^2 + W_{gz}z_{t-1} + b_g\right) \tag{12}$$

$$c_t = f_t \odot c_{t-1} + i_t \odot g_t \tag{13}$$

$$h_t^2 = o_t \odot \tanh(c_t) \tag{14}$$

The decoder LSTM is also updated by previous states and some parameters as the encoder LSTM did before. Notation $y_{1:T}$ refers to a sequence of words $(y_1, \ldots, y_T)$. The conditional distribution over possible result at each time step t, given by:

$$p(y_t|y_{1:t-1}) = \text{softmax}\left(W_p h_t^2 + b_p\right) \tag{15}$$

where W_p and b_p are learned matrixes. The complete sequence is calculated as:

$$p(y_{1:T}) = \prod_{t=1}^{T} p(y_t|y_{1:t-1}) \tag{16}$$

Recurrent Region Attention. Regional correlation patterns in images of real world can have many significant and complex changes. A considerable amount of image semantic information might not be well encoded, because each region context vector could only hold its limited information of semantics. In order to overcome this limitation, we introduce the attention mechanism into our model to improve its performance. So that it will automatically locate at the most relevant sections of the input region feature vector sequence and focus on these sections, when the model is predicting the current words. This is actually a standard sequence-to-sequence alignment mechanism which is different from the attention mechanism in [22]. We implement the mechanism by importing a special structure between the encoder output and reformulated decoder inputs.

Given the output h of the encoder LSTM, at each time step t we generate an attention weight $\alpha_{i,t}$ for each encoder hidden state h_i as:

$$u_{i,t} = h_t^{2^\top} W_u h_i \tag{17}$$

$$\alpha_t = \mathrm{softmax}(u_t) \tag{18}$$

$$z_t = \sum\nolimits_{i=1}^{L} \alpha_{i,t} h_i \tag{19}$$

where W_u is learned parameters, $u_{i,t}$ is the score at i-th hidden state of time t, $i = 1, \ldots, L(L = 6)$, and n is the splits number of image features we discussed before. Similar to [8], our approach of attention gets the decoder hidden state at time step t. Then we calculate attention scores, and from the calculated scores, we get the context vector z_t which will be concatenated with hidden state h_t^2 of the decoder. After that, we can predict a word of the caption sequence by Eqs. (15) and (16).

At last, the objective of our method is to minimize the cross entropy loss L_{CE} by given target ground truth sequence $y_{1:T}^*$ and captioning model with parameters θ, as follows:

$$L_{CE}(\theta) = -\sum\nolimits_{t=1}^{T} \log p_\theta\left(y_t^* | y_{1:T}^*\right) \tag{20}$$

We use the stochastic gradient descent (SGD) with gradient decay to optimize the goal function, which is efficient for optimizing our model, and the comprehensive of region contexts just feed at the beginning of the decoder LSTM only once at training time.

3 Experiments

3.1 Dataset

To evaluate our proposed model, a large and high-quality dataset is necessary. In view of this, we use the Microsoft COCO (MSCOCO) 2014 caption dataset [9]. For validation of model parameters and offline evaluation, we use the data splits from the method of 'Karpathy' [10]. These splits have been widely used to demonstrate results of models in the previous woks. The training split contains 113,287 images with five captions each, 5 K images for validation, and 5 K images for testing as well. We also submit our results to MSCOCO test server to get how effective our model is. Following other practicing standard, we slightly filter the model vocabulary. We keep words that appear above five times, convert all captions to lower case and tokenize on space. We end up with a vocabulary of length 10,116. We report results with seven extensively used evaluation metrics: BLEU (1, 2, 3, 4) [23], METEOR [25], ROUGE-L [24], and CIDEr [21].

3.2 Implementation Details

Our proposed IIRC subnetwork consists of two components, CNN and LSTM encoder. Particularly, in this work, we use Inception-v4 [7] CNN model which is well pre-trained on ImageNet [5] to extract the semantic feature of the image for image embedding. We elicit the feature from the layer after Inception-C blocks as our image feature V_I which has the shape of 8 × 8 × 1536 i.e. $V_h \times V_w \times V_w$. Cutting out from V_I, our each region feature vector has the shape of 4 × 4 × 1536. Then region feature vectors are pooled, flattened and full-connected to the 512-dimensional feature vector i.e. m = 512. Determined by experience of others works, both the encoder and the decoder LSTM of our model has 512 hidden state units (neurons). Similarly, word and attention embedding dimension are fixed to 512. Empirically, we set the initial learning rate as 0.5 with learning rate decay factor of 0.5 per 8 epochs for our SGD optimizer, and we find that it is a suitable way for our model optimizing. We initialize our model by the fixed pre-trained parameters of the CNN with given hyperparameters. After the model converges (i.e. we have a nice set of parameters), we unfix parameters of Inception-v4 and fine-tune the model to get the better performance on MSCOCO dataset. The learning rate is fixed to value of 5×10^{-4}.

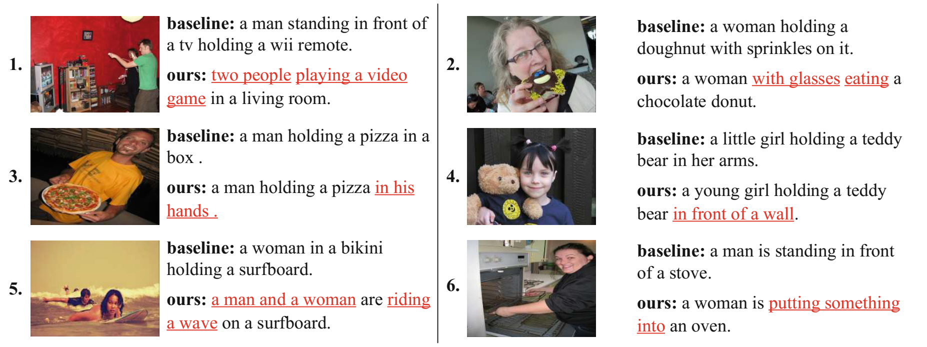
To quantify the effectiveness of our approach, similar to model in [11], our baseline model uses CNN as encoder and LSTM as decoder in encoder-decoder architecture. The difference is that we upgrade its CNN encoder from Inception-v3 to Inception-v4. The shape of CNN net's last layer output as image feature is 8 × 8 × 1536. This is equivalent to the original net's last layer output in [11] which has the shape of 8 × 8 × 2048. The number of LSTM hidden state units is similarly set to 512. Moreover, we set another model called All Regions Context (ARC) for comparative experiment. The ARC model is similar to our proposed IIRC model. However, it uses 64 regions of size 1 × 1 × 1536 as input region features in Sect. 2.1. To be fair, we trained both the baseline model and the ARC model in the same way as our IIRC model.

Table 1. Results on the online MSCOCO test server.

	BLEU-1		BLEU-2		BLEU-3		BLEU-4		METEOR		ROUGE-L		CIDEr	
	c5	c40	c5	c40	c5	c40	c5	c40	c5	c40	c5	c40	c5	c40
SCA-CNN [16]	71.2	89.4	54.2	80.2	40.4	69.1	30.2	57.9	24.4	33.1	52.4	67.4	91.2	92.1
NIC [11]	71.3	89.5	54.2	80.2	40.7	69.4	30.9	58.7	25.4	34.6	53.0	68.2	94.3	94.6
Review Net [12]	72.0	90.0	55.0	81.2	41.4	70.5	31.3	59.7	25.6	34.7	53.3	68.6	96.5	96.9
ATT_VC [13]	73.1	90.0	56.5	81.5	42.4	70.9	31.6	59.9	25.0	33.5	53.5	68.2	94.3	95.8
MSM [14]	73.9	91.9	57.5	84.2	43.6	74.0	33.0	63.2	25.6	35.0	54.2	70.0	98.4	100.3
PG-BCMR [15]	**75.4**	91.8	59.1	84.1	44.5	73.8	33.2	62.4	25.7	34.0	55.0	69.5	101.3	103.2
Ours: baseline	71.7	88.9	54.5	79.2	40.1	67.5	29.2	55.9	25.3	33.5	52.9	67.1	94.4	96.9
Ours: ARC	71.8	89.2	54.7	79.8	40.3	68.4	29.4	56.7	25.5	33.9	53.1	67.3	95.4	98.4
Ours: IIRC	74.9	**92.0**	**58.5**	**84.4**	**44.8**	**74.3**	**34.2**	**63.5**	**27.0**	**36.3**	**55.5**	**70.8**	**105.7**	**105.5**

3.3 Results and Discussion

To evaluate the effectiveness of our approach, we evaluate our model against prior works as well as our comparative models including baseline and ARC model. The evaluation results of comparison are illustrated in Table 1, where row IIRC is the results of our model. The results in the table show that, our IIRC model has achieved better performance at BLEU (1, 2, 3, 4), METEOR, ROUGE-L as well as CIDEr metrics, and exceeded our baseline and ARC in all metrics. Obviously, the scores of our IIRC model is higher than other models in Table 1 on all metrics only except c5 of BLEU-1 metric. The gap between ARC and IIRC shows the superiority of our approach in choosing salient regions. The approach in [14] utilizes both attributes information and image feature encoding to decode captions, but we only use the image feature from the picture. The approach in [15] uses reinforcement learning in optimizing metrics of its model, and it gives a higher weight on ROUGE metric. The CIDEr metric is different from other evaluation metrics, because it is proposed to aim at image abstract issues and have high matching rate of artificial consensus [21]. Specially, SCST model [17] has an optimizing target of CIDEr score with reinforcement learning, and it has established a state-of-the-art on the caption task. Therefore, the scores of our model on the CIDEr metric are more sufficient to prove the effectiveness of our approach. Simultaneously, the METEOR and ROUGE-L scores can also demonstrate that than BLEUs [20]. The score gap between our model and other models in Table 1 is sufficient to illustrate the validity of our model.



1. **baseline:** a man standing in front of a tv holding a wii remote.
ours: two people playing a video game in a living room.

2. **baseline:** a woman holding a doughnut with sprinkles on it.
ours: a woman with glasses eating a chocolate donut.

3. **baseline:** a man holding a pizza in a box .
ours: a man holding a pizza in his hands .

4. **baseline:** a little girl holding a teddy bear in her arms.
ours: a young girl holding a teddy bear in front of a wall.

5. **baseline:** a woman in a bikini holding a surfboard.
ours: a man and a woman are riding a wave on a surfboard.

6. **baseline:** a man is standing in front of a stove.
ours: a woman is putting something into an oven.

Fig. 3. Qualitative analysis on impact of our IIRC model. In the examples above our method can give more precise details of the picture such as positions, number of objects and the color detail, which baseline fail to do. The red and underlined words are caption details given by our model.

We also conduct a qualitative analysis on the role of intra-image region context in caption generation. We compare our model with baseline, which has similar architecture as [11]. Some samples of the caption generated by our approach method and baseline method are shown in Fig. 3. Our model can get details of position and action from perception of intra-image region context. Moreover, it achieves a well performance relative to our baseline. As examples 1, 2, 5 and 6 in Fig. 3, our model generates captions from the intra-image region context, which are more accurate at action details (e.g. in example 1, baseline just gives ‘standing in front a tv holding a wii remote’ but ours gives ‘playing a video game’ which shows intra-image region context information between TV and remote.). As example 3 and 4, our model shows more position details in caption results (e.g. in example 3, from the perception of region context, we get ‘in his hands’ rather than ‘in a box’, and the location of pizza is more appropriate). In addition, we can also give a more accurate number of objects in the caption (e.g. example 1 and 5).

4 Conclusion

In this paper, we present an approach that generates captions of images from intra-image region context. Our approach enables the salient region context to be effectively extracted from the image semantic feature, and it is able to automatically perceive the correlation among regions. Applying this approach, we can generate the description of an image based on the fusion of intra-image region contexts. The method is tested on MSCOCO test server. The experiment results demonstrate its superiority on all general caption metrics over other models and its effectiveness of perceiving the intra-image region context. Meanwhile, the IIRC model is able to generate captions with more details on position and action.

Acknowledgements. This work is supported by the Natural Science Foundation of China under Grant No. 61472020, 61572061.

References

1. Russakovsky, O., et al.: Imagenet large scale visual recognition challenge. Int. J. Comput. Vis. **115**(3), 211–252 (2015)
2. Farhadi, A., Endres, I., Hoiem, D., Forsyth, D.: Describing objects by their attributes. In: 2009 Computer Vision and Pattern Recognition, pp. 1778–1785. IEEE (2009)
3. Hodosh, M., Young, P., Hockenmaier, J.: Framing image description as a ranking task: data, models and evaluation metrics. In: International Conference on Artificial Intelligence, pp. 4188–4192 (2015)
4. Young, P., Lai, A., Hodosh, M., Hockenmaier, J.: From image descriptions to visual denotations: new similarity metrics for semantic inference over event descriptions. Nlp.cs. illinois.edu (2014)
5. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: a large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2009, pp. 248–255. IEEE (2009)
6. Bengio, Y., Simard, P., Frasconi, P.: Learning long-term dependencies with gradient descent is difficult. IEEE Trans. Neural Netw. **5**(2), 157–166 (1994)
7. Szegedy, C., Ioffe, S., Vanhoucke, V., Alemi, A.A.: Inception-v4, inception-resnet and the impact of residual connections on learning. In: AAAI, vol. 4, p. 12 (2017)
8. Luong, T., Pham, H., Manning, C.D.: Effective approaches to attention-based neural machine translation. In: Empirical Methods in Natural Language Processing, pp. 1412–1421 (2015)
9. Lin, T.-Y., et al.: Microsoft COCO: common objects in context. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8693, pp. 740–755. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10602-1_48

10. Karpathy, A., Fei-Fei, L.: Deep visual-semantic alignments for generating image descriptions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3128–3137 (2015)
11. Vinyals, O., Toshev, A., Bengio, S., Erhan, D.: Show and tell: a neural image caption generator. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3156–3164. IEEE (2015)
12. Yang, Z., Yuan, Y., Wu, Y., Cohen, W.W., Salakhutdinov, R.R.: Review networks for caption generation. In: Advances in Neural Information Processing Systems, pp. 2361–2369 (2016)
13. You, Q., Jin, H., Wang, Z., Fang, C., Luo, J.: Image captioning with semantic attention. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition, pp. 4651–4659 (2016)
14. Yao, T., Pan, Y., Li, Y., Qiu, Z., Mei, T.: Boosting image captioning with attributes. In: Computer Vision and Pattern Recognition, pp. 4894–4902 (2017)
15. Liu, S., Zhu, Z., Ye, N., Guadarrama, S., Murphy, K.: Improved image captioning via policy gradient optimization of spider. In: Proceedings of IEEE Conference on Computer Vision and Pattern, vol. 3 (2017)
16. Chen, L., et al.: SCA-CNN: spatial and channel-wise attention in convolutional networks for image captioning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5659–5667 (2017)
17. Rennie, S.J., Marcheret, E., Mroueh, Y., Ross, J., Goel, V.: Self-critical sequence training for image captioning. In: CVPR, vol. 1, p. 3 (2017)
18. Corbetta, M., Shulman, G.L.: Control of goal-directed and stimulus-driven attention in the brain. Nat. Rev. Neurosci. **3**(3), 201 (2002)
19. Buschman, T.J., Miller, E.K.: Top-down versus bottom-up control of attention in the prefrontal and posterior parietal cortices. Science **318**(5847), 1860–1862 (2007)
20. Elliott, D., Keller, F.: Comparing automatic evaluation measures for image description. In: Meeting of the Association for Computational Linguistics, pp. 452–457 (2014)
21. Vedantam, R., Zitnick, C.L., Parikh, D.: Cider: consensus-based image description evaluation. In: Computer Vision and Pattern Recognition, pp. 4566–4575 (2015)
22. Xu, K., et al.: Show, attend and tell: neural image caption generation with visual attention. In: International Conference on Machine Learning, pp. 2048–2057 (2015)
23. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting on As-sociation for Computational Linguistics, pp. 311–318. Association for Computational Linguistics (2002)
24. Lin, C.Y.: Rouge: a package for automatic evaluation of summaries. In: Proceedings of Workshop on Text Summarization Branches Out, Post Conference Workshop of ACL 2004 (2004)
25. Denkowski, M., Lavie, A.: Meteor universal: language specific translation evaluation for any target language. In: Proceedings of the Ninth Workshop on Statistical Machine Translation, pp. 376–380 (2014)