Multi-Stage Selective Re-Decoding Module for Image Paragraph Captioning

Guozhang Nie[1], Xian Zhong[2,a,*], Chengming Zou[2], Qi Cu[3]and Luo Zhong[2]

[1]School of Computer Science and Technology, Wuhan University of Technology, China
[2]Hubei Province Key Laboratory of Transportation Internet of Things,
Wuhan University of Technology, China
[3]School of Computer Science and Technology, Wuhan University of Technology, China
a. zhongx@whut.edu.cn.
**Xian Zhong*

Keywords: Image Paragraph Captioning, Encoder-Decoder, Multi-Stage Re-Decoding.

Abstract: Image paragraph captioning describes an image with a paragraph. Existing methods typically train hierarchical networks with a one-stage strategy, where one-stage means those models directly generate a description without multi-stage modification. Due to the exposure bias, we have observed that there may be errors and omissions in the description generation process, such as one object in the image is wrongly expressed or one subregion in the image is neglected. To solve this problem,we present a novel approach for image paragraph captioning, called the multi-stage selective re-decoding (MSSRD) module,which extends the conventional one-stage methods to generate richer captions. After gaining a preliminary caption, our module dynamically selects appropriate words and un-decoded visual features that are in the previous stage. These selected features are re-decoded into a new caption in the next stage. The new caption is more diverse and finer than previous one. We conduct extensive experiments to demonstrate the significance of our work.

1. Introduction

Image captioning [1] focuses on describing an image with a single sentence. However, it is often insufficient to provide the entire visual content of the image. Recent work proposed image paragraph captioning [2] to describe images with a paragraph of detailed and fine-grained stories. With the rapid progress of deep learning, the achievements in image captioning [1], [3]–[6] are remarkable. Nevertheless, when these strong single-sentence captioning models are trained on paragraph captioning dataset [2], they produce repetitive paragraphs that are unable to concisely describe image.

Prior works [2], [7]–[13] tried to address this repetition with architectural changes and training policy, which separate the generation of sentence topics and words. Krause et al. [2] first initiated the idea of image paragraph captioning and proposed a hierarchical RNN for image paragraph generation.To solve the issue of repetitive paragraphs, Melas-Kyriazi etal. [7] introduced an integrated penalty on trigram repetition into the image captioning model, which produces much more diverse paragraphs. Wang et al. [8] presented convolutional auto-encoding plus LSTM, which explored the modeling of sentence topics to boost image paragraph generation. Wu etal. [11] proposed an approach that formulated hierarchical rewards and values at both word and sentence levels, which provided dense supervision cues for learning effective para graph generator. To the best of our knowledge, existing ap proaches all generate paragraph caption by one-stage. They all generate the description text directly from the decoder in the model, without modifying the generated caption. However, the captions generated by them may contain errors and omissions due to exposure bias [14]. Thus, we deem multi-stage process,including eliminating errors and supplementing omissions,is necessary. In this paper, we demonstrate how multi-stage decoding, which allows modifying words/sentences during the sentence-making process, can greatly improve the performance of paragraph captioning.

Our idea is inspired by humans' daily writing with images.The most universal method is to write a rough and coarse paragraph in the first stage. In other stages, most approaches are to learn from the sentence obtained in the previous stage and observe the distinction between that and the content of the image. If there is a blemish in the paragraph, it will be altered. If one region of the image is not expressed in words,we will append a new sentence as a supplement to describe the missing part. To address the aforementioned problems,we introduce a multi-stage selective re-decoding (MSSRD) module. Concretely, after the decoding of the first stage, we can gain the initial paragraph descriptions. In other stages, the caption generated by the previous stage and visual features are fed into the proposed module. In our MSSRD, the previous caption is selected based on visual features attention, and the words that closely resembled image semantics will be selected.These words will be fed into the decoder in the next stage and become a part of the new caption. The visual features are selected based on text attention. Those visual features which have not been decoded in the previous stage will be re-selected, which will be passed to the decoder with selected words features in the next stage to generate a new caption. After many steps of such dynamic selections, the sentences generated by MSSRD are more consistent with ground-truth in terms of diversity and correctness.

2. The Proposed Method

2.1. Problem Formulation

We formulate the task of image paragraph captioning into a sequential decision-making process. In our method, the paragraph is considered a long sentence $X' = \{x'_1, x'_2, \ldots, x'_T\}$ with length of T words. Each words is mapped to a unique vec tor. For sequence-level training, similar to previous works [4],[7], we leverage two learning stages: (1) standard supervised learning with cross-entropy loss; (2) reinforcement learning (RL) by policy gradient method using a self-critical relative base reward [4].

Training with Cross Entropy Loss: In the first training stage, the objective function of the framework is to minimize the cross-entropy loss (XE):

$$\mathcal{L}_{XE}(\theta) = -\sum_{t=1}^{T} \log(p(x_t|x_1, \ldots, x_{t-1})) \tag{1}$$

where $x_{1:T}$ denotes the target ground-truth sequence, and θ is the parameters to the decoder.

CIDEr-D Score Optimization: Then we directly optimize the non-differentiable metrics with self-critical sequence train ing (SCST) [4]:

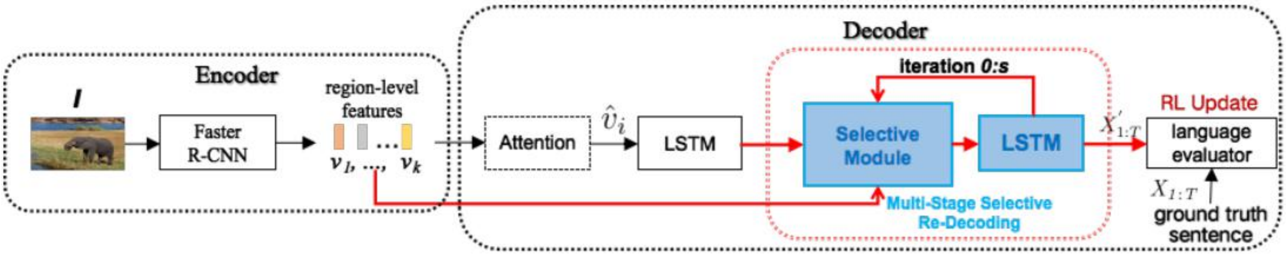
$$\mathcal{L}_{RL}(\theta) = -\mathbf{E}_{\mathbf{y}_{1:T} \sim p_\theta}[r(\mathbf{y}_{1:T})] \tag{2}$$

where the reward r(·) uses the score of metric CIDEr-D [15].The gradients can be approximated:

$$\nabla_\theta \mathcal{L}_{RL}(\theta) \approx -(r(\mathbf{y}^s_{1:T}) - r(\hat{\mathbf{y}}_{1:T}))\, \nabla_\theta \log p_\theta(\mathbf{y}^s_{1:T}) \quad (3)$$

where $\mathbf{y}^s$ ys means it's a result sampled from a probability distribution, while $\hat{\mathbf{y}}$ indicated a result of greedy decoding.

2.2. Framework



Figureure 1: An overview of our MSSRD framework for image paragraph generation. There are six interconnected components: Faster R-CNN, Attention, LSTM, Selective Module, LSTM, and Language Evaluator.

As in many previous studies, we devise our MSSRD based on traditional encoder-decoder structure. An overview of our method is depicted in Figureure 1. Firstly, image I is encoded by Faster R-CNN [16] model to extract region-level visual features. Next, these visual features can be processed by optional Attention modules. Then, these visual features can be sent to the first LSTM model that can be as decoder to get the generated sentences $X'_{1:T}$. These sentences go through a Language Evaluator to train the entire model through reinforcement learning. All the above processes are the traditional one-stage method. It can be found that our proposed MSSRD can be employed as an extension of any captioning models.It accepts the sentences generated in the previous stage and visual features, makes the dynamic selection, and feeds the selected features to the decoder LSTM for re-decoding. This process can be iterated s times and trained by RL.

2.3. One-Stage Encoder-Decoder

Encoder: We first encode I to obtain image features using a deep CNN. In this paper, we use the region-level features to be consistent with precious works [7], thus we extract a set of feature vectors $V = \{v_1, v_2, ..., v_K\}$ using a Faster R-CNN pre-trained by [5] on Visual Genome [17], where $v_i \in \mathbb{R}^D$, K is the number of vectors in V , and D is the dimension of each vector, and visual features $V \in \mathbb{R}^{K \times D}$.

Decoder: Our decoder implements a top-down attention [5] hierarchical LSTM composed of 2 LSTM layers, a top-down attention layer, and a multilayer perceptron layer followed by softmax function. The structure of our first-stage paragraph decoder is absolutely consistent with the previous work [7].

2.4. Multi-Stage Selective Re-Decoding Module

After the decoding of the first stage, supposing we have a coarse paragraph caption X' of I, some sentences are associated with a certain region of the image but lack diversity with other sentences. Moreover, these sentences may be inaccurate due to exposure bias [14]. We deem that multi-stage re-decoding would facilitate the decoder to find out inner relations between image and sentences.

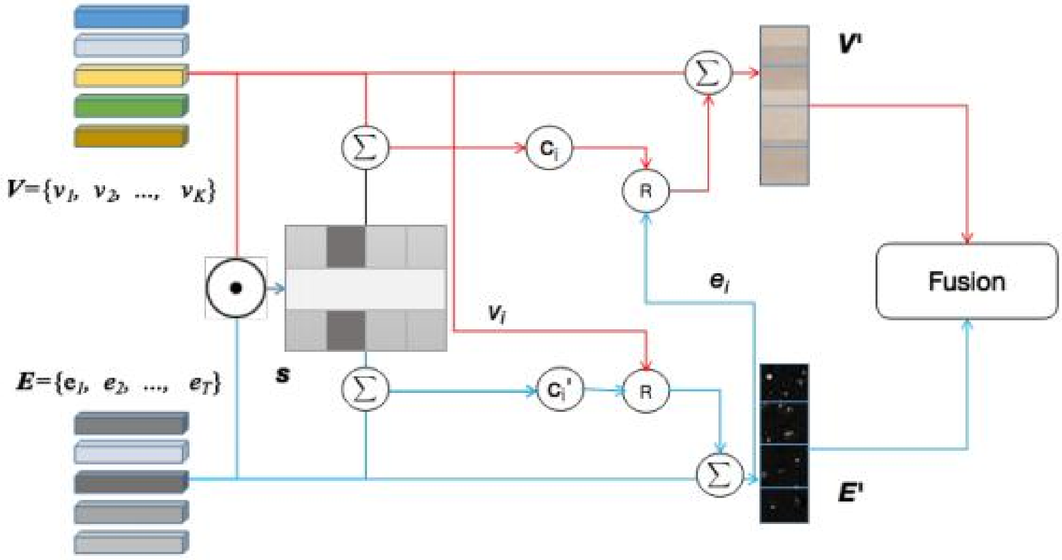


Figure 2: An illustration of the proposed Selective Module selecting the proper words and unused visual features from the previous stage.

In our proposed module (see Figure 2), the selection and fusion require visual features and text features. We need to encode X0 generated in the previous stage. Inspired by [18], here we employ a bi-directional gated recurrent unit (Bi-GRU) as the text encoder to obtain the feature vectors for each word. Specifically, we use a learned word embedding as the inputs of the Bi-GRU to summarize information from both forward and backward directions in the word x'_i. Then the final feature vector ei for the word x'_i is computed by averaging both hidden states from the forward and the backward GRU. Finally, we obtain feature vector for each X0 generated by previous stage,which is denoted as $E = \{e_i | i = 1, \ldots, n, e_i \in \mathbb{R}^M\}$, E ∈ R175×M, where each e_i encodes a word information, and M is the dimension of e_i, and all sentences in X' are padded and truncated to the same length *T*.

Image-based Word Selection: We first calculate the sim ilarity matrix for all possible pairs of words in the sentences and sub-regions in the image by:

$$s = E(W_v V)^T \tag{4}$$

where $W_v \in \mathbb{R}^{D\times M}$ is learned parameters. $(\cdot)^T$ denotes the transpose of visual features matrix, $s \in \mathbb{R}^{175\times K}$ and $s_{i,j}$ is the dot-product similarity between the *i*-th word of the sentences and the *j*-th sub-region of the image. Then, we normalize the similarity matrix as follows:

$$\bar{s}_{i,j} = \frac{\exp(s_{i,j})}{\sum_{k=0}^{T-1}\exp(s_{k,j})} \tag{5}$$

Then, similar to [19], we build an attention model to computer a region-context vector for each word (query). The region context vector c_i is a dynamic representation of the image's sub-regions related to the i-th word of the sentence. It is computed as the weighted sum over all visual vectors:

$$c_i = \sum_{j=0}^{K-1} \alpha_j v_j, \text{where } \alpha_j = \frac{\exp(\gamma \bar{s}_{i,j})}{\sum_{t=0}^{K-1}\exp(\gamma \bar{s}_{i,t})} \tag{6}$$

where γ is a factor that determines how much attention is paid to features of its relevant sub-regions when computing the region-context vector for a word. After that, we define the relevance between the i-th word and the image using the cosine similarity between c_i and e_i, $R(c_i, e_i) = \frac{c_i^T e_i}{\|c_i\|\|e_i\|}$. The $R(c_i, e_i)$ can simply represent how well the i-th word and image match. It is a scalar in the range [0, 1]. A value of 0 implies that the *i*-th word is not consistent with the image and will be not considered in the next stage; otherwise 1. Finally,we can obtain all words that are consistent with the image. Itis computed as follows:

$$E' = \sum_{i=0}^{T-1} R(c_i, e_i) e_i \tag{7}$$

where $E' \in \mathbb{R}^D$.

Sentences-based Visual Selection: Similarly, to gain the visual features that haven't been decoded in previous stages,we re-normalize the similarity matrix as follows:

$$\bar{s}'_{i,j} = \frac{\exp(s_{i,j})}{\sum_{t=0}^{K-1} \exp(s_{i,t})} \tag{8}$$

The sentence-context vector c'_i is a dynamic representation of the sentences' words related to the i-th region-level feature of the image. Since the sentence generated in the previous stage may have errors, we use the correct word that has been selected as a reference. We redefine (6) by:

$$c'_i = \sum_{j=0}^{T-1} \beta_j R(c_j, e_j) e_j, \text{where } \beta_j = \frac{\exp(\gamma \bar{s}'_{i,j})}{\sum_{t=0}^{T-1} \exp(\gamma \bar{s}'_{t,j})} \tag{9}$$

where $R(c_j, e_j)e_j$ is the same as (7). Then, we redefine $R(c'_i, v_i) = \frac{c_i'^T v_i}{\|c'_i\|\|v_i\|}$ and substitute it to (7), obtaining the features that were not decoded in the previous stage. The $R(c'_i, e_i)$ has the same meaning as $R(c_i, e_i)$. However, its usage is discrepant. Since if a sentence is consistent with the sub-region in the image, it indicates that the visual features of this sub-region have been decoded in the previous stage and have been retained in the sentence. To generate other diverse sentences, we should choose those region-level visual features with low sentence similarity, which will be retained to the next stage for re-decoding. Eventually, we can obtain whole visual features that were not decoded in previous stages. It is computed as follows:

$$V' = \sum_{i=0}^{K-1} [1 - R(c'_i, v_i)] \; v_i \tag{10}$$

where $V' \in \mathbb{R}^D$.

Multi-Stage Re-Decoding: After obtaining the final visual region and word features for paragraph captioning, such fea tures could then be fused. Then, these fused features will be transmitted to a one-layer LSTM module for re-decoding. This process can be iterated s times (see Figureure 1). We experiment with the three fusion approaches by 2 stages in which the features concatenation between visual and language represen tations achieves the best performance with a trivial margin.

3. Exrerimental Results

3.1. Datasets and Experimental Setting

Dataset and Metrics: We conduct experiments on Stan ford image-paragraph dataset released in [2], which contains 14,575, 2,489, and 2,487 images for training, validation, and testing, respectively. Each image was annotated with a single paragraph, which contains multiple (8–10) sentences. Six widely used evaluation metrics BLEU-1~ 4 [20], ME TEOR [21], and CIDEr [15] are adopt in our experiments for quantitative evaluation.

Implementation details: We adopt the same hyperparam eters settings in [7]. In our proposed module, T = 175 and word features are encoded into features of dimension M = 1024 by GRU. The factor γ in (6) and (9) is set to 5. We use a single layer LSTM with hidden size of 1024 in the MSSRD module. For other detailed parameters, please refer to [7].

3.2. Comparison with State-of-the-Art

To verify the effectiveness of the proposed method, we conduct experiments and compare our model with the state of-the-arts for image paragraph captioning. Table I reports the quantitative performance comparison. We can observe that the proposed MSSRD achieves the best performance in almost all metrics. For example, our MSSRD improves CIDEr, a specifically designed metric for evaluation captions, by 10.44% over the second-best method baseline [7] w/o MSSRD. Note that the only difference between our method and the baseline [7] is whether the MSSRD module is used or not. The results show our method can solve the problems of errors and omissions in some degree.

3.3. Ablation Study of MSSRD

To further verify the effectiveness of our MSSRD, we perform extensive ablation studies, and the results are shown in Table II. In fact, the factor that affects the performance of our model is the number of stages in multiple stages. To verify it, we fix all other parameters and observed the model effect by setting different stage sizes in Figure 1. The results obtained are shown in Table II that appropriately increasing the stage size *s* can increase the performance of the model. Overall, the best number of multi-stage selection for re-decoding is 2. When it continues to increase, the performance of the model will be hindered. The reason may be over-fitting and it is difficult to train again. Therefore, we fix the stage-size s = 2 and try different feature fusion type via either features concatenation, feature element-wise product, and feature addition to obtain fused features. Finally, we get the best result when feature fusion type is features concatenation. All indicators have exceeded the other two fusion methods. However, the gap between the three is almost negligible.

Table 1: Performance (%) with the state-of-the-art methods on stanford, where m, c, and b-1~4 are short for meteor, CIDEr, and bleu-1~4. bold numbers are the best results.

Method	M	C	B-1	B-2	B-3	B-4
Regions-Hierarchical [2]	15.95	13.52	41.9	24.11	14.23	8.69
RP-GAN [12]	17.40	14.71	41.94	24.99	15.01	9.38
DAM [10]	13.90	17.30	35.00	20.20	11.70	6.60
Dual-CNN [13]	15.60	17.40	41.60	24.40	14.30	8.60
CAPG-VAE [9]	18.62	20.93	42.38	25.52	15.15	9.43
DHPV [11]	17.02	22.47	43.35	26.73	16.92	10.99
CAE-LSTM [8]	18.82	25.15	-	-	-	9.67
Baseline [7] w/o MSSRD	17.86	30.63	43.54	27.44	17.33	10.58
Ours w/ MSSRD	18.40	33.83	45.72	27.98	17.85	11.53

Table 2: Performance (%) of the selection of stage size s in MSSRD module, bold numbers are the best results.

s	METEOR	CIDEr	BLEU-1	BLEU-2	BLEU-3	BLEU-4
1	18.08	33.76	44.69	27.76	17.78	11.55
2	18.40	33.83	45.72	27.98	17.85	11.53
3	18.35	34.02	45.14	27.90	17.81	11.12
4	17.65	30.34	42.43	27.23	17.04	10.11

Conclusions

In this paper, we proposed the MSSRD module, an ex tension of any traditional caption models, to solve errors and omissions for image paragraph captioning. The proposed MSSRD module, which can dynamically select words that match the semantics of images and un-decoded visual features in the previous stage, can generate caption based on diversity and correctness. More remarkably, we significantly improved performance with MSSRD by a large margin on Standford image paragraph captioning dataset.

Acknowledgments

This work was supported in part by Fundamental Research Funds for the Central Universities under Grant 191010001 and Hubei Key Laboratory of Transportation Internet of Things under Grant 2018IOT003, 2020III026GX.

References

[1] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," in CVPR, 2015, pp. 3156–3164.

[2] J. Krause, J. Johnson, R. Krishna, and F. Li, "A hierarchical approach for generating descriptive image paragraphs," in CVPR, 2017, pp. 3337– 3345.

[3] K. Xu, J. Ba, R. Kiros, K. Cho, A. C. Courville, R. Salakhutdinov, R. S. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in ICML, 2015, pp. 2048–2057.

[4] S. J. Rennie, E. Marcheret, Y. Mroueh, J. Ross, and V. Goel, "Self critical sequence training for image captioning," in CVPR, 2017, pp. 1179–1195.

[5] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang, "Bottom-up and top-down attention for image captioning and visual question answering," in CVPR, 2018, pp. 6077–6086.

[6] C. Chen, S. Mu, W. Xiao, Z. Ye, L. Wu, and Q. Ju, "Improving image captioning with conditional generative adversarial nets," in AAAI, 2019, pp. 8142–8150.

[7] L. Melas-Kyriazi, A. M. Rush, and G. Han, "Training for diversity in image paragraph captioning," in EMNLP, 2018, pp. 757–761.

[8] J. Wang, Y. Pan, T. Yao, J. Tang, and T. Mei, "Convolutional auto encoding of sentence topics for image paragraph generation," in IJCAI, 2019, pp. 940–946.

[9] M. Chatterjee and A. G. Schwing, "Diverse and coherent paragraph generation from images," in ECCV, 2018, pp. 747–763.

[10] Z. Wang, Y. Luo, Y. Li, Z. Huang, and H. Yin, "Look deeper see richer: Depth-aware image paragraph captioning," in ACM MM, 2018, pp. 672– 680.

[11] S. Wu, Z. Zha, Z. Wang, H. Li, and F. Wu, "Densely supervised hierarchical policy-value network for image paragraph generation," in IJCAI, 2019, pp. 975–981.

[12] W. Che, X. Fan, R. Xiong, and D. Zhao, "Visual relationship embedding network for image paragraph generation," IEEE Trans. Multim., vol. 22, no. 9, pp. 2307–2320, 2020.

[13] R. Li, H. Liang, Y. Shi, F. Feng, and X. Wang, "Dual-CNN: A convolutional language decoder for paragraph image captioning," Neu rocomputing, vol. 396, pp. 92–101, 2020.

[14] J. Gu, K. Cho, and V. O. K. Li, "Trainable greedy decoding for neural machine translation," in NeurIPS, 2017, pp. 1968–1978.

[15] R. Vedantam, Z. C. Lawrence, and D. Parikh, "CIDEr: Consensus-based image description evaluation," in CVPR, 2015, pp. 4566–4575.

[16] S. Ren, K. He, R. B. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in NeurIPS, 2015, pp. 91–99.

[17] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L. Li, D. A. Shamma, M. S. Bernstein, and F. Li, "Visual genome: Connecting language and vision using crowdsourced dense image annotations," Int. J. Comput. Vis., vol. 123, no. 1, pp. 32–73, 2017.

[18] H. Chen, G. Ding, Z. Lin, S. Zhao, and J. Han, "Cross-modal image-text retrieval with semantic consistency," in ACM MM, 2019, pp. 1749–1757.

[19] T. Xu, P. Zhang, Q. Huang, H. Zhang, Z. Gan, X. Huang, and X. He, "AttnGAN: Fine-grained text to image generation with attentional generative adversarial networks," in CVPR, 2018, pp. 1316–1324.

[20] K. Papineni, S. Roukos, T. Ward, and W. Zhu, "Bleu: a method for automatic evaluation of machine translation," in ACL, 2002, pp. 311– 318.

[21] S. Banerjee and A. Lavie, "METEOR: An automatic metric for MT evaluation with improved correlation with human judgments," in ACLW, 2005, pp. 65–72.