Journal on Big Data
DOI:10.32604/jbd.2021.016674
Article

Encoder-Decoder Based Multi-Feature Fusion Model for Image Caption Generation

Mingyang Duan, Jin Liu* and Shiqi Lv

Shanghai Maritime University, Shanghai, 201306, China
*Corresponding Author: Jin Liu. Email: jinliu@shmtu.edu.cn
Received: 08 January 2021; Accepted: 07 April 2021

Abstract: Image caption generation is an essential task in computer vision and image understanding. Contemporary image caption generation models usually use the encoder-decoder model as the underlying network structure. However, in the traditional Encoder-Decoder architectures, only the global features of the images are extracted, while the local information of the images is not well utilized. This paper proposed an Encoder-Decoder model based on fused features and a novel mechanism for correcting the generated caption text. We use VGG16 and Faster R-CNN to extract global and local features in the encoder first. Then, we train the bidirectional LSTM network with the fused features in the decoder. Finally, the local features extracted is used to correct the caption text. The experiment results prove that the effectiveness of the proposed method.

Keywords: Image understanding; image captioning; deep learning; fused features

1 Introduction

Nowadays, with the rapid technology development, self-portraits and hand-to-hand shooting have gradually become a mainstream social way, which makes the number of images grows at an exponential rate. Conventional methods of image retrieval which usually involves manual annotation of images and brief image captions, can no longer deal with images of this magnitude. Thus, enabling computer the vision processing ability, especially the image understanding ability with neural networks, has attracted researchers' attentions.

Image captioning is an essential task in image understanding, which involves image semantic segmentation, image recognition, NLP, etc. The early methods are based on template extraction and matching or based on machine learning and statistical models. Most of these methods depend on the formulation of rules, and the training efficiency is not satisfactory. However, recent development in neural networks and deep learning have made them popular methods in image captioning. And Encoder-Decoder model has been used as the basis of image captioning models. However, traditional methods only use global image features in the part of Encoder while local features are ignored. Besides, to improve accuracy of image captioning, it is nature to utilized the finding of biologists' work on the visual attention mechanism of humans to build the deep learning model.

This paper proposed a new Encoder-Decoder model-based method to fuse multiple features from an image and conduct description text correction. We utilize the global image features and the local features in the encoder and highlights the representation of the local targets in the image caption. At the same time, the fused feature is trained by the Bi-LSTM network with attention mechanism to increase its representational ability. By comparing the subject in the generated image caption with the extracted local features, the final description text will be rectified.

This work is licensed under a Creative Commons Attribution 4.0 International License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

2 Related Work

Early methods of image captioning relied on rules. Those methods match description text with labeled image features in advance. Researchers label the image features without image caption in this way.

From 1999 to 2003, Mori et al. proposed the methods that divided the image into regions and then described the content in each region separately [1–2]. Jeon proposed a cross-media correlation model in 2003 [3]. In 2010, Farhadi et al. created a triple of (object, motion, background) by object extraction and feature extraction on the original image [4]. In 2009, Liu et al. proposed an image annotation method based on graph learning [5]. Besides grapy models, Zhou et al. proposed a method, transforming the semantic partitioning problem into machine learning problem through Multi-Instance Multi-Label Learning (MIML) [6].

In recent years, deep learning has become widespread and it provides a way to generate image caption completely different from the above methods [7–8]. And what is more important, the deep learning-based methods had achieved higher accuracy for image caption generation. Vinyals et al. proposed an encoder-decoder model in 2015 [9].

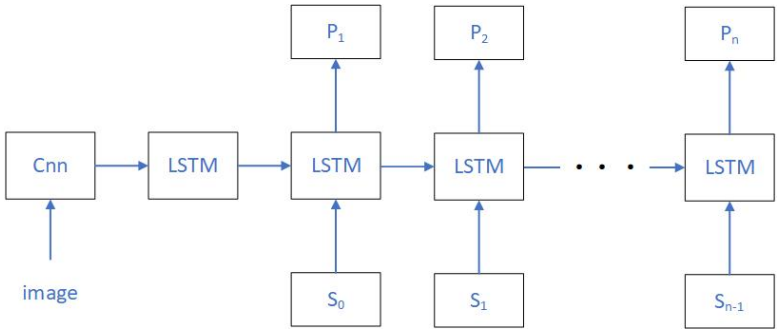

Figure 1: The network structure of the encoder-decoder model

After 2014, various deep learning-based image caption methods were proposed. In 2014, Chen et al. changed the structure of the RNN network in decoder, so that the RNN network could not only translate image features into text, but also got image features from text [10]. Xu et al. proposed a method that adding a layer of attention mechanism between the original encoder and decoder layers [11]. You et al. proposed a novel attention algorithm [12].

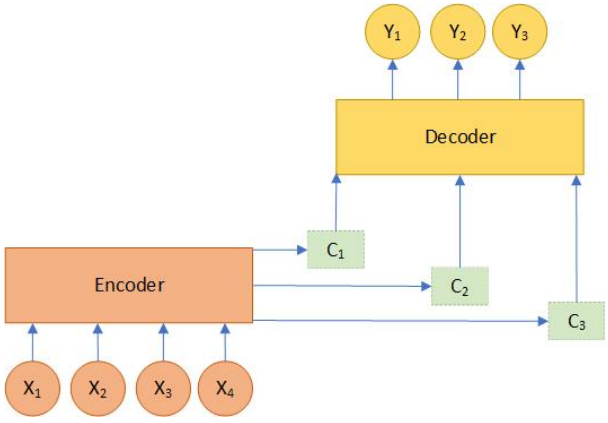

Figure 2: The structure of the encoder-decoder model with attention mechanism

Deep learning-based methods are becoming the de-facto methods for image captioning [13–15]. However, in most of these traditional deep learning-based methods, the image caption models only use the global feature alone [16–18]. The relationships between objects in image is not well described, and some obvious errors appears in the final caption text. In this work, to overcome the disadvantages of

existing methods, we propose a novel image caption model that utilizes a fused features extraction algorithm and a text correction mechanism (FF-TC).

3 Method

The proposed FF-TC method consists of three major components, fused image features extraction, Bi-LSTM model with attention mechanism, and text correction for image caption. The above three methods are embedded in an encoder-decoder model.

3.1 Overview of Features Extraction

In this section, a features fusion algorithm is presented. A VGG16 [19] is used to extract global features. With multiple convolution layers, it can extract highly abstract image features. Then, a Faster-RCNN [20] is used for local image feature extraction. The mixture feature obtained by the fusion algorithm is used as the output of the encoder, reducing the impact of useless information and enhancing the expression of the key information [21–23]. We use the attention mechanism and the bidirectional LSTM network in the decoder [24–26]. And the Bi-LSTM is used to generate image caption. For the wrong objects description in the caption text, we identify the real category of objects in the image and make corrections [27–29].

3.1.1 Global Features Extraction

We extract local and global image features respectively. We select the VGG16 network for the global feature extraction. The formal expression of the convolution process is shown in Eq. (1).

$$x_j^l = f(\sum_i^{M_l} a_i^{l-1} \times w_{ij}^l + b_j^l) \tag{1}$$

The specific network model is shown in Fig. 3. Since there is no need to identify the category of the image, the fully connected layers used by the original VGG16 is removed in our FF-TC.

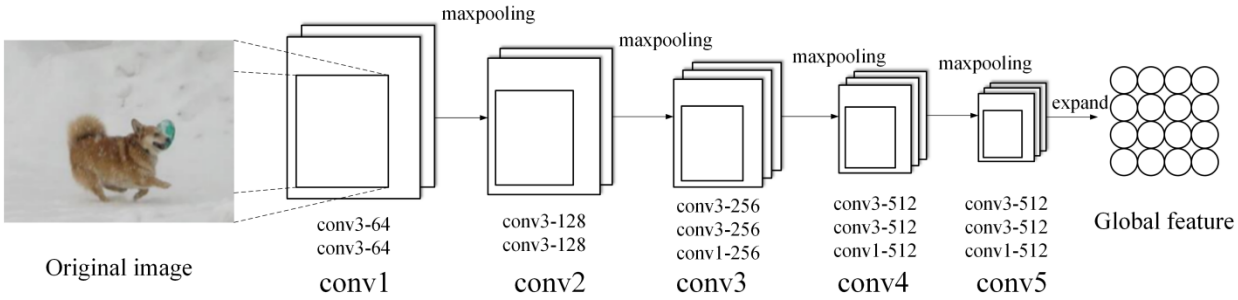

Figure 3: The structure of our VGG 16 for the global feature extraction of images

3.1.2 Local Features Extraction

We extract the information of objects by the Faster-RCNN. The process of object detection is shown in Fig. 4. Then, a VGG16 extracts local image features. The local features need to maintain the same dimension as the global features, so the set of local features is converted into a set of N*N matrices. The local feature is defined in Eq. (2).

$$L_f = \begin{pmatrix} b_{00} & \cdots & \cdots & b_{0j} \\ \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots \\ b_{i0} & \cdots & \cdots & b_{ij} \end{pmatrix}_{i*j} , i = j = N \tag{2}$$

Since people usually pay more attention to the object occupying a large proportion of the image, we use the proportion of each object to the whole image to evaluate the importance of them.

$$P = \frac{Area_{object}}{Area_{image}} \tag{3}$$

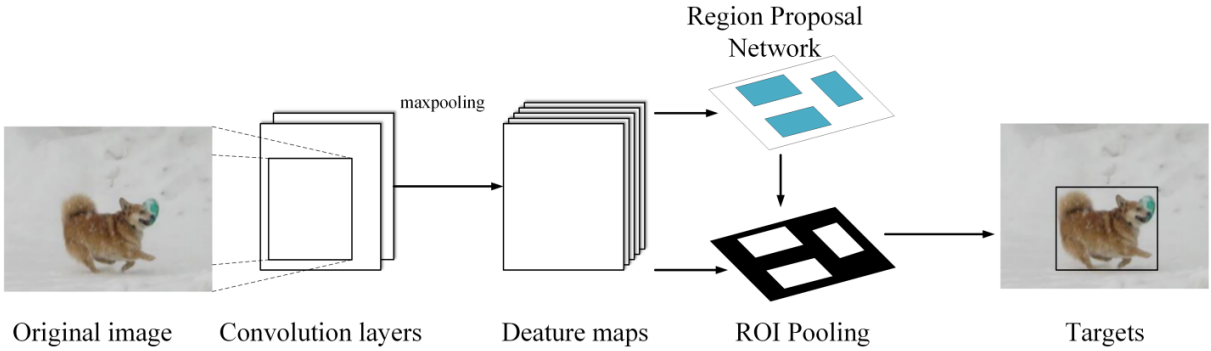

Figure 4: The specific steps of object detection by a Faster-RCNN

3.2 Bidirectional LSTM with Attention Mechanism

We use the encoder-decoder structure with attention mechanism to build our image caption model. The calculation of attention is as follows:

$$C_i = \sum_{j=1}^{T} a_{ij} h_i \tag{4}$$

$$\delta_i = softamx(f_{att}(h_i, s_j)) \tag{5}$$

$$f_{att}(h_i, s_j) = tanh(W_1 h_i + W_2 s_j) \tag{6}$$

where C_i refers to the environment vector. h_i refers to the hidden state in t step. s_j refers to the hidden state before t step. a_{ij} refers to the attention coefficient. The environment vector and the current hidden state h_i can be calculated at the same time. δ_i refers to the weight of the attention. $f_{att}(h_i, s_j)$ refers to the assigned value between h_i and s_j.

The encoder extracts global feature G_f and local feature L_f, and outputs the mixture feature M_f. The attention mechanism assigns the ratio between stronger associations and weaker associations. The decoder part consists of a Bi-LSTM network. We process the mixture feature through an attention layer and then send it into the Bi-LSTM. The representation of each word can be used to calculate the contextual relationship of the word through the Bi-LSTM unit. Using t = 1…N to represent the index of words in different time steps, the formal expression of the Bi-LSTM unit is as follows:

$$x_t = W_\omega \vartheta_t \tag{7}$$

$$e_t = f(W_e x_t + b_e) \tag{8}$$

$$h_t^f = f(e_t + W_f f_{t-1}^f + b_f) \tag{9}$$

$$h_t^b = f(e_t + W_b f_{t-1}^b + b_b) \tag{10}$$

$$S_t = f(W_d(h_t^f + h_t^b) + b_d) \tag{11}$$

where ϑ_t refers to a vector of the index of the words. The Bi-LSTM has two separate workflows. h_t^f refers to the work flow from left to right, while h_t^b refers to the work flow from right to left.

4 Experiment

Our image caption model is based on the fused features of images and text correction mechanism; thus FF-TC consists of multiple neural networks. We train these different networks with different datasets.

We train a VGG 16 model with the ImageNet dataset in the experiment. We scale the images to a

size of 224 * 224 * 3, and then use the mean filtering on the images to reduce the impact of noise. There are 15 layers, including 5 convolutional layers, 5 max pooling layers, 1 flattening layer, 3 fully connection layers and 1 softmax layer.

The Pascal VOC dataset is used to train the Faster-RCNN. During training, we set the numbers of anchors of RPN network in the Faster-RCNN as 256 and maintain a 1:1 positive and negative sample ratio.

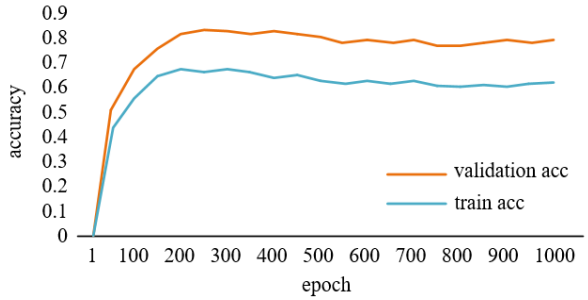

Figure 5: The loss and accuracy curves of training

Figure 6: Part of the experiment results

The MSCOCO dataset is used for the training of the bidirectional LSTM. The word vector obtained by the embedding and the mixture features obtained by the feature fusion algorithm will be sent to the attention layer and the bidirectional LSTM layers. The loss and accuracy curves are shown in Fig. 5. The test accuracy in the training set can reach 78.2%, and the test accuracy in the verification set can reach 66.5%. Part of the experiment results are shown in Fig. 6.

Table 1: The Bleu values

Method	Bleu1	Bleu2	Bleu3	Bleu4
BRNN	64.2	45.1	30.4	20.3
NIC	66.6	46.1	32.9	24.6
M-RNN	67	49	35	25
LRCN	62.79	44.19	30.41	21
Emb-gLSTM	67.0	49.1	35.8	26.4
FF (no TC)	68.3	50.3	33.7	22
FF-TC	**68.9**	**51.1**	**33.9**	**22.5**

As shown in Tab. 1, our method has improved the performance of image caption for most images. The mixture features combined with global and local features can better guide image caption generation. In addition, with the text correction algorithm, parts of the classification errors of objects can be corrected. The text correction algorithm greatly improves the accuracy of the image caption.

5 Conclusion

We proposed a novel Encoder-Decoder model-based image captioning model FF-TC. The algorithm for fusing image features and the algorithm for correcting description text with local features is our major contribution, where the fused features capture both the details and relationships of objects. For the possible errors in the description text, our method rectifies them by utilizing the similarity got from the objects' information in the image and the nouns in the text.

The proposed model takes advantages of various neural networks. The image features are extracted by a VGG16 and a Faster R-CNN respectively. The bidirectional LSTM with attention mechanism is used to generate image caption text. The encoder-decoder model connects image features extraction and caption generation. According to experiments results, the fused features have proven to be superior to either global features or local features alone. The text correction method is also proved to be feasible.

Funding Statement: This work is supported by the National Natural Science Foundation of China (6187223).

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

References:

[1] J. Liu and C. K. Gu, "Multi-scale multi-class conditional generative adversarial network for handwritten character generation," *The Journal of Supercomputing*, vol. 75, no. 4, pp. 1922–1940, 2019.

[2] Y. Mori and H. Takahashi, "Image-to-Word transformation based on dividing and vector quantizing images with words," in *Int. Workshop on Multimedia Intelligent Storage & Retrieval Management*, pp. 1–9, 1999.

[3] J. Jeon and V. Lavrenko, "Automatic image annotation and retrieval using cross-media relevance models," in *Int. ACM SIGIR Conf. on Research and Development in Information Retrieval*, pp. 119–126, 2003.

[4] A. Farhadi and M. Hejrati, "Every picture tells a story: generating sentences from images," in *European Conf. on Computer Vision*, pp. 15–29, 2010.

[5] J. Liu and M. Li., "Image annotation via graph learning," *Pattern Recognition*, vol. 42, no. 2, pp. 218–228, 2009.

[6] Z. H. Zhou and M. L. Zhang, "Multi-instance multi-label learning with application to scene classification," in *Int. Conf. on Neural Information Processing Systems*, pp. 1609–1616, 2006.

[7] J. Liu, M. J. Zhou, L. Lin, H. J. Kim and J. Wang, "Rank web documents based on multi-domain ontology," *Journal of Ambient Intelligence and Humanized Computing*, pp. 1–10, 2018.

[8] J. Liu, J. J. Song, L. Kong, J. U. Kim and J. Wang "A novel parallel method for denoising and deduplicating mass web documents," *Journal of Internet Technology*, vol. 17, no. 5, pp. 889–896, 2016.

[9] O. Vinyals, A. Toshev and S. Bengio, "Show and tell: A neural image caption generator," *Computer Vision and Pattern Recognition. IEEE*, pp. 3156–3164, 2015.

[10] X. Chen and Z. C. Lawrence, "Mind's eye: a recurrent visual representation for image caption generation," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 2422–2431, 2015.

[11] K. Xu and J. Ba, "Show, attend and tell: neural image caption generation with visual attention," *Computer Science*, pp. 2048–2057, 2015.

[12] H. Fang, S. Gupta, F. Iandola, R. Srivastava, L. Deng *et al.,* "From captions to visual concepts and back," *Computer Vision and Pattern Recognition*, pp. 1473–1482, 2015.

[13] J. Liu, L. Lin, Z. H. Cai, J. Wang and H. J. Kim, "Deep web data extraction based on visual information processing," *Journal of Ambient Intelligence and Humanized Computing*, pp. 1–11, 2017.

[14] J. Liu, L. N. Wang, M. J. Zhou, J. Wang, S. Lee *et al.,* "Fine-grained entity type classification with adaptive context," *Soft Computing*, vol. 22, no. 13, pp. 4307–4318, 2018.

[15] J. Liu, H. L. Ren, M. L. Wu, J. Wang, H. J. Kim *et al.,* "Multiple relations extraction among multiple entities in unstructured text," *Soft Computing*, vol. 22, pp. 4295–4305, 2018.

[16] J. Liu, Y. H. Yang, S. Q. Lv, J. A. Wang, H. Chen *et al.,* "Attention-based BiGRU-CNN for Chinese question classification," *Journal of Ambient Intelligence and Humanized Computing*, pp. 1–12, 2019.

[17] J. Liu, Y. Li, X. Tian, A. Sangaiah and J. Wang, "Towards semantic sensor data: an ontology approach," *Sensors (Basel, Switzerland)*, vol. 19, no. 5, 2019.

[18] J. Liu, L. Lin, H. L. Ren and M. H. Gu, "Building neural network language model with pos-based negative sampling and stochastic conjugate gradient descent," *Soft Computing*, vol. 22, no. 20, pp, 6705–6717, 2018.

[19] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," arXiv preprint arXiv:1409.1556, 2014.

[20] S. Ren, K. He, R. Girshick and J. Sun, "Faster R-CNN: towards real-time object detection with region proposal networks," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. 39, no. 6, pp. 1137–1149,2017.

[21] A. Krizhevsky, I. Sutskever and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Int. Conf. on Neural Information Processing Systems*, Curran Associates Inc., pp. 1097–1105, 2012.

[22] Z. Yaqing and B. Liuzhong, "Image feature extraction by fusing global and local features," *Journal of Huaqiao University (Natural Science)*, vol. 36, no. 4, pp. 406–411, 2015.

[23] V. Mnih, N. Heess, A. Graves and K. Kavukcuoglu, "Recurrent models of visual attention," NIPS, vol. 2, pp. 2204–2212, 2014.

[24] M. T. Luong, H. Pham and C. D. Manning, "Effective approaches to attention-based neural machine translation," arXiv preprint arXiv:1508.04025, 2015.

[25] C. Fellbaum, "WordNet, theory and applications of ontology," *Computer Applications*, pp. 231–243, 2010.

[26] X. Rong, "Word2vec parameter learning explained," *Computer Science*, 2014.

[27] F. Zhang, H. Zhao, W. Ying, Q. Liu, A. Noel *et al., "*Human face sketch to RGB image with edge optimization and generative adversarial networks," *Intelligent Automation & Soft Computing*, vol. 26, no. 6, pp. 1391–1401, 2020.

[28] H. Wu, Q. Liu and X. Liu, "A review on deep learning approaches to image classification and object segmentation," *Computers, Materials & Continua*, vol. 60, no. 2, pp. 575–597, 2019.

[29] J. Zhang, Y. Li, S. Niu, Z. Cao and X. Wang, "Improved fully convolutional network for digital image region forgery detection," *Computers, Materials & Continua*, vol. 60, no. 1, pp. 287–303, 2019.