Contents lists available at ScienceDirect

Signal Processing: Image Communication

journal homepage: www.elsevier.com/locate/image

Repeated review based image captioning for image evidence review

Jinning Guan, Eric Wang *

Shenzhen Key Laboratory of Internet Information Collaboration, Shenzhen Graduate School, Harbin Institute of Technology, Shenzhen, 518055, China

ARTICLE INFO

Keywords:
Repeated review
Image captioning
Encoder–decoder
Multimodal layer

ABSTRACT

We propose a repeated review deep learning model for image captioning in image evidence review process. It consists of two subnetworks. One is the convolutional neural network which is employed to extract the image features and the other is the recurrent neural network which is used to decode the image features into captions. Our model combines the advantages of the two subnetworks by recalling visual information different from the traditional model of encoder–decoder, and then introduces multimodal layer to fuse the image and caption effectively. The proposed model has been validated on benchmark datasets (MSCOCO, Flickr). It shows that the proposed model performs well on bleu-3 and bleu-4, even to some extent, beyond the best models available today (such as NIC, m-RNN, etc.).

1. Introduction

Evidence review is one of key steps in digital image forensic, which includes reviewing on authenticity, legacy and relativity. Among them, authenticity and legacy review can be fulfilled by hardware or software tools, while, content relativity review has been done by hand for a long time. Particularly when there are a great many of the images reviewed as evidence, the work is a time consuming process since reviewers need to check the content of the images one by one.

Automatic review of image contents is of practical value, and it needs understanding image contents, annotating them and presenting relations among subjects in images precisely by machine. Although describing the content of the given images is a very challenging task, it is still a fascinating area to be explored.

Image captioning is the key process for automatic image review. It has an image as the input, and the annotation of the image content as the output. The task requires that it can recognize objects, understand their relations and present it in natural language. Rapid development of computer vision and natural language processing technologies inspire many ideas on it. And most of them, they consider it retrieving mapping problem on both sentences and images to a same semantic space. It relies on datasets, which means, it is hard to describe a new unseen image if there are no such new objects or scenes in datasets. While, a new unseen image to be recognized is a common scene.

In this work, we proposed a Repeated-Review Neural Image Captioning (RRNIC) to generate image captioning, and realized an end-to-end neural network system which is able to automatically review the content of an image and generate a reasonable description in plain English. RRNIC is a combination of the convolution neural network and the recurrent neural network. The convolutional neural network (CNN) is employed to generate a vector expression of an image, which will feed into a recurrent neural network (RNN). In order not to forget the raw image along the transfer of RNN, a multimodal layer is designed to review the raw image at each round. Experiment on common datasets shows RRNIC is robust on the task of image captioning and in some metrics it is better than the current methods.

Our system can be served for automatic image evidence review in the future. All the image contents can be reviewed by machine and descriptions are generated automatically. Image pointers and content descriptions are stored in database for future retrieval. The evidence reviewer only need to set up key words related to current case to retrieve those sensitive images.

2. Related work

Image captioning involves two popular research areas (computer vision and natural language), which recently both have many achievements by deep learning. Some researchers keep working on improving the accuracy of caption generation for images and some results show encouraging effects.

Deep learning in computer vision and natural language. In computer vision field, Krizhevsky et al. [1] proposed a deep Convolutional Neural Networks (CNN) with 8 layers (denoted as AlexNet) and its performance in the image classification task of the ImageNet competition outperforms previous methods. Recently, the designer of

* Corresponding author.
E-mail addresses: guanjinning@stu.hit.edu.cn (J. Guan), wk_hit@hit.edu.cn (E. Wang).

https://doi.org/10.1016/j.image.2018.02.005
Received 2 December 2017; Received in revised form 31 January 2018; Accepted 5 February 2018
Available online xxxx

Please cite this article in press as: J. Guan, E. Wang, Repeated review based image captioning for image evidence review, Signal Processing: Image Communication (2018), https://doi.org/10.1016/j.image.2018.02.005.

GoogleLeNet [2], which has 22 layers, became the winner of the 2014 ImageNet competition. Later on, a better approach called "Rethinking the Inception Architecture for Computer Vision" [3] (Inception-v3) was proposed, which achieves significant improvement on the ImageNet task with 3.5% of top-5 error rate on the validation dataset (3.6% error rate on the test dataset) and 17.3% of top-1 error rate on the validation dataset. Inception structure has a good performance. Residual connection is different from the traditional network structure, and the corresponding network Resnet proposed by Kaiming He [4] who won the championship in 2015, ILSVRC. Later, Christian Szegedy & Sergey Ioffeto [5] combined Inception structure with Residual connection and proposed a new type of networks, Resnet-Inception, of which the top-5 error rate decreased to be 3.08%.

In natural language, Recurrent Neural Network (RNN) has the top performance in many tasks, such as speech recognition and word embedding learning [6–8]. And most of caption generations are built on recurrent neural networks. All of them mainly prove the effectiveness of storing context information in a recurrent layer.

Image captions generator. A few new approaches have been introduced in recent years. Mao et al. [9] employs a recurrent neural network to predict the text to be generated. Vinyals et al. [10] constructs a neural image caption generator. They encoded an image and decoded it into sentences. In addition, some approaches try to introduce attention mechanism into image region. Xu et al. [11] proposed a model to generate words on the image regions that systems select. The alignment for generating image caption is employed by Karpathy et al. [12] during training. Finally, Chen et al. [13] build a visual representation for sentences while generating descriptions. And they proposed a model based on the discussion of the role of visual attention model in space and channel. Snapchat [14] and google cooperatively research on employing reinforcement learning to train image description and generate neural networks, adopting the Actor-critic framework. The application of adaptive attention mechanism in image description generation is attracted by Lu et al. [15]

3. Our model

Our model combines the advantages of NIC [10] and mrnn [9], employs a encoder–decoder approach, and solves the problem of forgetting the original image information in the decoding process of NIC. At the same time, we proposed a multimodal layer which can merge the original images with text information.

It mainly involves following parts, (a) vector representations on images, (b) vector representations on words, (c) Bi-RNN model on feature capture and decoding, (d) a mix layer which can repeatedly review, (e) a word generation model. The details are explained as follows:

3.1. Flow

Fig. 1 shows the general flow of our approach. First, the original image is extracted by a convolutional neural network for feature extraction at different levels. After continuous convolution and pooling operations, all feature planes are compressed to get a fixed length feature vector. Then the feature vectors of the image are decoded as the initial state of the Bi-RNN, and meanwhile the first word of the image captioning is produced. Then the vector representation of the previous word is regarded as the input of the next Bi-RNN neuron to predict the output of the next word, so it keeps going until the whole sentence is generated completely. At the same time, the image features are repeatedly reviewed in the decoding process of each step to prevent the information of the original image from being forgotten during the process of recurrent neural network transmission.

On the step of generating a sentence for a given image, we employ beam search method [16], beam search is a heuristic search algorithm that explores a graph by expanding the most promising node in a limited set, and it iteratively considers the set of the k best sentences up to time t as candidates to generate sentences of size $t + 1$, and keeps only the resulting best k of them.

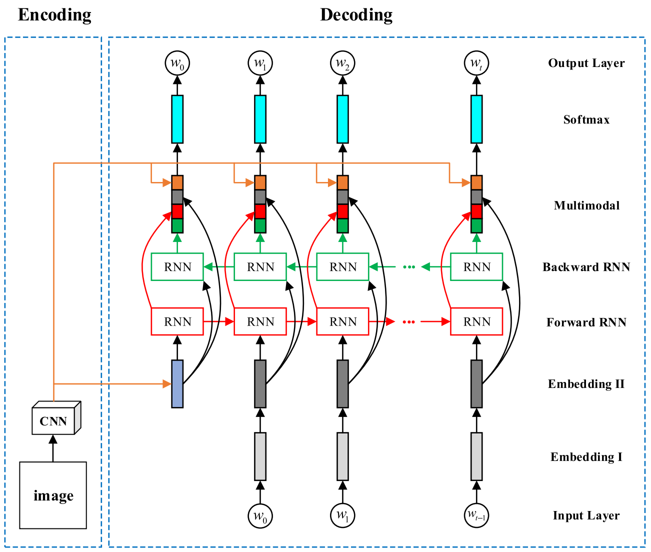

Fig. 1. The flow of our approach.

3.2. Images encoding

As described above, we employed a deep convolutional neural network (CNN) to perform encoding for images. The visual representation comes from a convolutional neural network which is pre-trained by image classification on a large-scale dataset such as ImageNet.

The variables are initialized in pre-trained models (inception V3 and Resnet-Inception) and after that, the vector representation of the given image is generated, which is denoted as $g(I)$. Then two fully connected networks are trained: one is to map $g(I)$ into the first input expected by LSTM network, and the other is to map $g(I)$ into the multimodal layer.

3.3. Caption decoding

In the stage of caption decoding, we employed bidirectional recurrent neural network (Bi-RNN) [17]. The basic idea of Bi-RNN is that each training sequence consist of two RNNs, which represent forward and backward, and both are connected to an output layer. The structure provides the complete past and future context information for each point in the input layer. We believe that the information is not only related to the previous information, but also the subsequence, therefore we choose Bi-RNN as the main framework. Besides, the performance of Bi-RNN should be better than original RNN, and Bi-RNN can employ LSTM [18] or GRU [19] as its basic neuron cell, which is called Bi-LSTM or Bi-GRU. The structures of Bi-LSTM and Bi-GRU are showed in Figs. 2 and 3. It can be seen from the diagram that the output of the Bi-RNN is determined by the output of the forward recurrent neural network and the output of the reverse recurrent neural network.

Since Bi-RNN can consider the influence of the previous and the latter moment on the current moment, the surrounding information that can be captured at the current moment will be more abundant. Also, our model compared the effects of Bi-LSTM and Bi-GRU on sentence generation.

Suppose the output of forward recurrent neural network is h_{ft}, the output of reverse recurrent neural network is h_{bt}, then the output of Bi-RNN is $h_t = [h_{ft}^T, h_{bt}^T]^T$.

Long Short Term Memory Network (LSTM) [18] and Gated Recurrent Unit (GRU) [19] are two kinds of recurrent neural etworks commonly used in natural language processing (see Figs. 4 and 7).

LSTM has solved the problem of "Long-Term Dependencies" by introducing forget gate, input gate and output gate. The forget gate formula is defined as follows:

$$f_t = \sigma(W_{xf} X_t + W_{hf} h_{t-1} + b_f) \tag{3.1}$$

Please cite this article in press as: J. Guan, E. Wang, Repeated review based image captioning for image evidence review, Signal Processing: Image Communication (2018), https://doi.org/10.1016/j.image.2018.02.005.