Generating Descriptions for Sequential Images with Local-Object Attention and Global Semantic Context Modelling

Jing Su[1], Chenghua Lin[2], Mian Zhou[3], Qingyun Dai[4], Haoyu Lv[4]
[1]Guangdong Ocean University, [2]University of Aberdeen
[3] Tianjin University of Technology, [4] Guangdong University of Technology
jingsuw@163.com,chenghua.lin@abdn.ac.uk
zhoumian@tjut.edu.cn,1144295091@qq.com,lvhaoyuchn@163.com

Abstract

In this paper, we propose an end-to-end CNN-LSTM model for generating descriptions for sequential images with a local-object attention mechanism. To generate coherent descriptions, we capture global semantic context using a multi-layer perceptron, which learns the dependencies between sequential images. A paralleled LSTM network is exploited for decoding the sequence descriptions. Experimental results show that our model outperforms the baseline across three different evaluation metrics on the datasets published by Microsoft.

1 Introduction

Recently, automatically generating image descriptions has attracted considerable interest in the fields of computer vision and nature language processing. Such a task is easy to humans but highly non-trivial for machines as it requires not only capturing the semantic information from images (e.g., objects and actions) but also needs to generate human-like natural language descriptions.

Existing approaches to generating image description are dominated by neural network-based methods, which mostly focus on generating description for a single image (Karpathy and Li, 2015; Xu et al., 2015; Jia et al., 2015; You et al., 2016). Generating descriptions for sequential images, in contrast, is much more challenging, i.e., the information of both individual images as well as the dependencies between images in a sequence needs to be captured.

Huang et al. (2016) introduce the first sequential vision-to-language dataset and exploit Gated Recurrent Units (GRUs) (Cho et al., 2014) based encoder and decoder for the task of visual storytelling. However, their approach only considers image information of a sequence at the first time step of the decoder, where the local attention mechanism is ignored which is important for capturing the correlation between the features of an individual image and the corresponding words in a description sentence. Yu et al. (2017) propose a hierarchically-attentive Recurrent Neural Nets (RNNs) for album summarisation and storytelling. To generate descriptions for an image album, their hierarchical framework selects representative images from several image sequences of the album, where the selected images might not necessary have correlation to each other.

In this paper, we propose an end-to-end CNN-LSTM model with a local-object attention mechanism for generating story-like descriptions for multiple images of a sequence. To improve the coherence of the generated descriptions, we exploit a paralleled long short-terms memory (LSTM) network and learns global semantic context by embedding the global features of sequential images as an initial input to the hidden layer of the LSTM model. We evaluate the performance of our model on the task of generating story-like descriptions for an image sequence on the sequence-in-sequence (SIS) dataset published by Microsoft. We hypothesise that by taking into account global context, our model can also generate better descriptions for individual images. Therefore, in another set of experiments, we further test our model on the Descriptions of Images-in-Isolation (DII) dataset for generating descriptions for each individual image of a sequence. Experimental results show that our model outperforms a baseline developed based on the state-of-the-art image captioning model (Xu et al., 2015) in terms of BLEU, METEOR and ROUGE, and can generate sequential descriptions which preserve the dependencies between sentences.

Proceedings of the Workshop on Intelligent Interactive Systems and Language Generation (2IS&NLG), pages 3–8,
Tilburg, The Netherlands, November 5 2018.

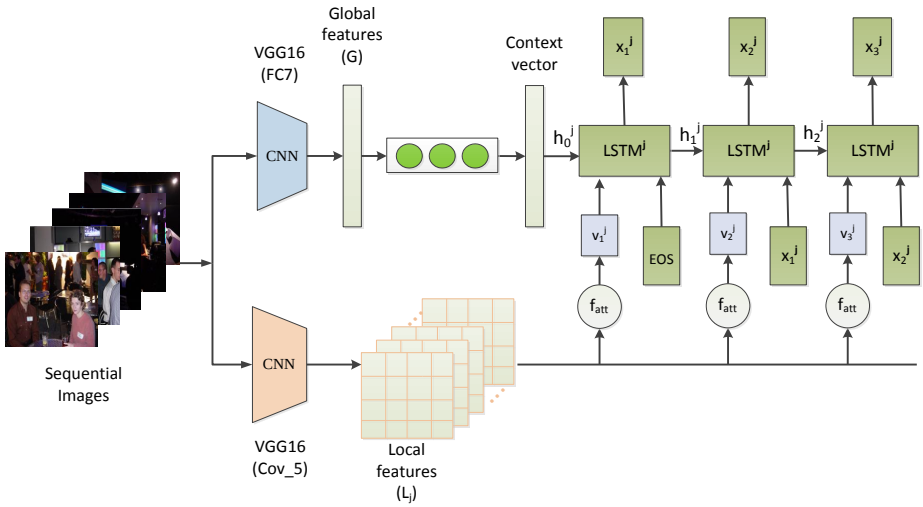


Figure 1: The architecture of our CNN-LSTM model with global semantic context.

2 Related Work

Recent successes in machine translation using Recurrent Neural Network (RNN) (Bahdanau et al., 2014; Cho et al., 2014) catalyse the adoption of neural networks in the task of image caption generation. Early works of image caption generation based on CNN-RNN networks have been made great progress.Vinyals et al. (2014) propose an encoder-decoder model which utilises a Convolutional Neural Network (CNN) for encoding the input image into a vector representation and a Recurrent Neural Network (RNN) for decoding the corresponding text description. Similarly, Karpathy and Li (2015) present an alignment model based on a CNN and a bidirectional RNN which can align segment regions of an image to the corresponding words of a text description. Donahue et al. (2014) propose a Long-term Recurrent Convolutional Network (LRCN) which integrates convolutional layers and long-range temporal recursion for generating image descriptions.

Recently, the attention mechanism (Xu et al., 2015; You et al., 2016; Lu et al., 2016; Zhou et al., 2016) has been widely used and proved to be effective in the task of image description generation. For instance, Xu et al. (2015) explore two kinds of attention mechanism for generating image descriptions, i.e., soft-attention and hard-attention, whereas You et al. (2016) exploits a selective semantic attention mechanism for the same task.

There is also a surge of research interest in visual storytelling (Kim and Xing, 2014; Sigurdsson et al., 2016; Huang et al., 2016; Yu et al., 2017). Huang et al. (2016) collect stories using Mechanical Turk and translate a sequence of images into story-like descriptions by extending a GRU-GRU framework. Yu et al. (2017) utilise a hierarchically-attentive structures with combined RNNs for photo selection and story generation. However, the above mentioned approaches for generating descriptions of sequential images do not explicitly capture the dependencies between each individual images of a sequence, which is the gap that we try to address in this paper.

3 Methodology

In this section, we describe the proposed CNN-LSTM model with local-object attention. In order to generate coherent descriptions for an image sequence, we introduce global semantic context and a paralleled LSTM in our framework as shown in Figure. 1. Our model works by first extracting the global features of sequential images using a CNN network (VGG16) (Simonyan and Zisserman, 2014), which has been extensively used in image recognition. Here a VGG16 model contains 13 convolutional layers, 5 pooling layers and 3 fully connected layers. The extracted global features are then embedded into a global semantic vector with a multi-layer perceptron as the initial input to the hidden layer of a paralleled LSTM model. Our model then applies the last convolutional-layer operation from the VGG16 model to generate the local features of each image in sequence. Finally, we introduce a paralleled LSTM model and a local-object attention mecha-

nism to decode sentence descriptions.

3.1 Features Extraction and Embedding

Sequential image descriptions are different from single image description due to the spatial correlation between images. Therefore, in the encoder, we exploit both global and local features for describing the content of sequential images. We extract global features of the sequential images with the second fully connected layer (FC7) from VGG16 model. The global features are denoted by G which are a set of 4096-dimension vectors. Then, we select the features of the final convolutional layer (Cov_5) from the VGG16 model to represent local features for each image in the sequence. The local features are denoted as L_j ($j = 1,...,N$), where N is the number of images in the sequence. In our experiment, we follow Huang et al. (2016) and set 5 as the number of images in a sequence. Finally, we embed the global features G into a 512-dimension context vector via a multilayer perceptron which is then used as the initial input of the hidden layer in LSTM model.

3.2 Sequential Descriptions Generation

In the decoding stage, our goal is to obtain the most likely text descriptions of a given sequence of images. This can be generated by training a model to maximize the log likelihood of a sequence of sentences S, given the corresponding sequential images I and the model parameters θ, as shown in Eq. 1.

$$\theta^* = \arg\max_{\theta} \sum_{j=1}^{N} \sum_{(I,s_j)} log\, p(s_j|I,\theta) \quad (1)$$

Here s_j denotes a sentence in S, and N is the total number of sentences in S.

Assuming a generative model of each sentence s_j produces each word in the sentence in order, the log probability of s_j is given by the sum of the log probabilities over the words:

$$\log p(s_j|I) = \sum_{t=1}^{C} log\, p(s_{j,t}|I, s_{j,1}, s_{j,2}...s_{j,t-1}) \quad (2)$$

where $s_{j,t}$ represents the t^{th} word in the j^{th} sentence and C is the total number of words of s_j.

We utilize a LSTM network (Hochreiter and Schmidhuber, 1997) to produce a sequence descriptions conditioned on the local feature vectors, the previous generated words, as well as the hidden state with a global semantic context. Formally, our LSTM model is formulated as follows:

$$\begin{aligned}
i_t^j &= \sigma(W_{xi}x_{t-1}^j + W_{hi}h_{t-1}^j + W_{vi}v_t^j + b_i) \\
f_t^j &= \sigma(W_{xf}x_{t-1}^j + W_{hf}h_{t-1}^j + W_{vf}v_t^j + b_f) \\
o_t^j &= \sigma(W_{xo}x_{t-1}^j + W_{ho}h_{t-1}^j + W_{vo}v_t^j + b_o) \\
q_t^j &= \varphi(W_{xq}x_{t-1}^j + W_{hq}h_{t-1}^j + W_{vq}v_t^j + b_q) \\
c_t^j &= f_t^j \odot c_{t-1}^j + i_t^j \odot q_t^j \\
h_t^j &= o_t^j \odot \varphi(c_t^j)
\end{aligned} \quad (3)$$

where i_t^j, f_t^j, o_t^j and c_t^j represents input gates, forget gates, output gates and memory, respectively. q_t^j represents the updating information in the memory c_t^j. σ denotes the sigmoid activation function, $\odot$ represents the element-wise multiplication, and φ indicates the hyperbolic tangent function. $W_\bullet$ and $b_\bullet$ are the parameters to be estimated during training. Also h_t^j is the hidden state at time step t which will be used as an input to the LSTM unit at the next time step.

Here, we utilize a multilayer perceptron to model the global semantic context which can be viewed as the initial input of the hidden state h_0^j, where every initial value h_0^j in the LSTM model is equal and is defined as:

$$h_0^j = W_0\, \varphi(W_g G + b_g) \quad (4)$$

When modelling local context, the local context vector v_t^j is a dynamic representation of the relevant part of the j^{th} image in a sequence at time t. In Eq. 6, we use the attention mechanism f_{att} proposed by (Bahdanau et al., 2014) to compute the local attention vector v_t^j, where the corresponding weight k_t^j of each local features L_j is computed by a softmax function with input from a multilayer perceptron which considers both the current local vector L_j and the hidden state h_{t-1}^j at time $t-1$.

$$k_t^j = softmax(W_k\, tanh(W_{lv}L^j + W_{hv}h_{t-1}^j + b_v)) \quad (5)$$

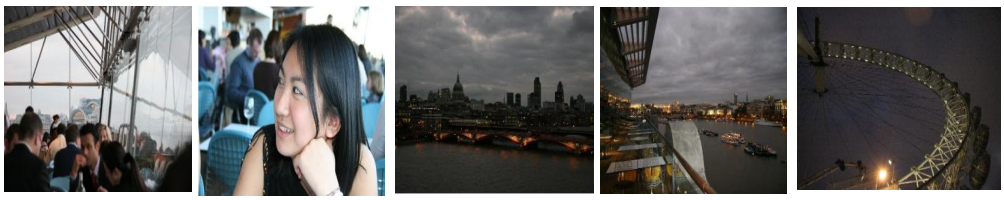
$$v_t^j = \sum_{i=1}^{M} k_{it}^j L_i^j \quad (6)$$

4 Experiments

Dataset.

Both the SIS and DII datasets are published by Microsoft[1], which have a similar data structure,

[1] http://visionandlanguage.net/VIST/



DII (our model)	(1) a group of people that are on the beach. (2) a man and a woman pose for a picture together. (3) a city at night with many buildings in the backgroud. (4) a bridge that is next to the water. (5) a large ship is being enjoyed by the crowd.
DII (cnn-att-lstm)	(1) a group of people that are next to each other. (2) a man and a woman sitting at a table. (3) a group of friends pose for a picture. (4) the man is blowing out into the camera. (5) a woman is smilling.
DII (ground truth)	(1) a variety of people sitting in a window filled restanrant. (2) closeup of a woman looking to her right in a restaurant setting. (3) many buildings by the beach. (4) a waterfront scence from an outside restaurant at night. (5) people on the ferris wheel.
SIS (our model)	(1) the family went to restaurant. (2) the family was very excited to have a party. (3) the sun was going down to the beach. (4) the family decide to go to restaurant. (5) i was so excited to have a great time.
SIS (cnn-att-lstm)	(1) the city is a small windows. (2) the girls are ready to go to the day. (3) the beautiful fireworks. (4) the city has a great view. (5) we drove up.
SIS (ground truth)	(1) me and my lover went on a vacation to see some sights. here we are getting something to eat. (2) we liked the food but the place was rather crowded for our tastes. here is a view of the city from our hotel. (3) it was so lovely to look out every night as the sun went down. another shot from high up. (4) it was breath taking to watch the city light up as the sun went down. (5) we where in line for a ferris wheel. i thought that this would make a good pic, and i think it came out well.

Figure 2: Example of sequential descriptions generated by our model, the baseline, and the ground truth.

Positive example

(1) the kids had a lot of fun. (2) the people were very happy to celebrate. (3) the people brought their favorite. (4) the people were enjoying themselves. (5) the people were very happy.

Failure Example

(1) there was a great time. (2) i had a great time. (3) we took a great time. (4) this is a picture. (5) we had a great time.

Figure 3: Error analysis of our model. First row: our model generates correct captions. Second row: failure cases due to severe overfitting.

Dataset	Train	Test	Vocab. Size
DII	23,415	1,665	10,000
SIS	110,905	10,370	18,000

Table 1: Dataset statistics.

Dataset	Method	BLEU	METEOR	ROUGE
DII	cnn-att-lstm	36.1	9.2	26.9
	Our model	**40.1**	**11.2**	**29.1**
SIS	cnn-att-lstm	15.2	4.6	13.6
	Our model	**17.2**	**5.5**	**15.2**

Table 2: Evaluation of the quality of descriptions generated for sequential images.

i.e., each image sequence consists of five images and their corresponding descriptions. The key difference is that descriptions of SIS consider the dependencies between images, whereas the descriptions of DII are generated for each individual image, i.e., no dependencies are considered. As the full DII and SIS datasets are quite large, we only used part of both datasets for our initial experiments, where the dataset statistics are shown in Table 1.

Evaluation. We compare our model with the sequence-to-sequence baseline (cnn-att-lstm) with attention mechanism (Xu et al., 2015). The cnn-att-lstm baseline only utilises the local attention mechanism which combines visual concepts of an image with the corresponding words in a sentence. Our model, apart from adopting a local-object attention, can further model global semantic context for capturing the correlation between sequential images.

Table 2 shows the experimental results of our model on the task of generating descriptions for sequential images with three popular evaluation metrics, i.e. BLEU, Meteor and ROUGE. It can be observed from Table 2 that our model outperforms the baseline on both SIS and DII datasets for all evaluation metrics. It is also observed that the scores of the evaluation metric are generally higher for the DII dataset than the SIS dataset. The main reason is that the SIS dataset contains more sentences descriptions in a sequence and more abstract content descriptions such as "breathtaking" and "excited" which are difficult to understand and prone to overfitting.

Figure 2 shows an example sequence of five images as well as their corresponding descriptions generated by our model, the baseline (cnn-att-lstm), and the ground truth. For the SIS dataset, it can observed that our model can capture more coherent story-like descriptions. For instance, our model can learn the social word "family" to connect the whole story and learn the emotional words "great time" to summarise the description. However, the baseline model failed to capture such important information. Our model can learn dependencies of visual scenes between images even on the DII dataset. For example, compared to the descriptions generated by cnn-att-lstm, our model can learn the visual word "beach" in image 1 by reasoning from the visual word "water" in image 4.

Our model can generally achieve good results by capturing the global semantics of an image sequence such as the example in the first row of Figure 3. However, our model also has difficulties in generating meaningful descriptions in a number of cases. For instance, our model generates fairly abstractive descriptions such as "a great time" due to severe overfitting, as shown in the second row of Figure 3. We suppose the issue of overfitting is likely to be alleviated by adding more training data or using more effective algorithm for image feature extraction.

5 Conclusion

In this paper, we present a local-object attention model with global semantic context for sequential image descriptions. Unlike other CNN-LSTM models that only employ a single image as input for image caption, our proposed method can generate descriptions of sequential images by exploiting the global semantic context to learn the dependencies between sequential images. Extensive experiments on two image datasets (DII and SIS) show promising results of our model.

Acknowledgement

We thank Ehud Reiter and Kees van Deemter for their helpful discussion. This paper was also supported partly by Science and Technology Project of Guangdong Province (No.503314759024).

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *Computer Science*.

Kyunghyun Cho, Bart Van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger

Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *Computer Science*.

Jeff Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. 2014. Long-term recurrent convolutional networks for visual recognition and description. *CoRR*.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation*, 9(8):1735–1780.

Ting-Hao (Kenneth) Huang, Francis Ferraro, Nasrin Mostafazadeh, Ishan Misra, Aishwarya Agrawal, Jacob Devlin, Ross Girshick, Xiaodong He, Pushmeet Kohli, Dhruv Batra, C. Lawrence Zitnick, Devi Zitnick Zitnick, Lucy Vanderwende, Michel Galley, and Margaret Mitchell. 2016. Visual storytelling. In *the Conference of the North American Chapter of the Association for Computational Linguistics*, pages 1233–1239.

Xu Jia, Efstratios Gavves, Basura Fernando, and Tinne Tuytelaars. 2015. Guiding the long-short term memory model for image caption generation. In *IEEE International Conference on Computer Vision*, pages 2407–2415.

Andrej Karpathy and Fei Fei Li. 2015. Deep visual-semantic alignments for generating image descriptions. In *Computer Vision and Pattern Recognition*, pages 3128–3137.

G. Kim and E. P. Xing. 2014. Reconstructing storyline graphs for image recommendation from web community photos. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3882–3889.

Jiasen Lu, Caiming Xiong, Devi Parikh, and Richard Socher. 2016. Knowing when to look: Adaptive attention via A visual sentinel for image captioning. *CoRR*.

Gunnar A. Sigurdsson, Xinlei Chen, and Abhinav Gupta. 2016. Learning visual storylines with skipping recurrent neural networks. *CoRR*, abs/1604.04279.

Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *Computer Science*.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2014. Show and tell: A neural image caption generator. pages 3156–3164.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. *Computer Science*, pages 2048–2057.

Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo. 2016. Image captioning with semantic attention. pages 4651–4659.

Licheng Yu, Mohit Bansal, and Tamara L. Berg. 2017. Hierarchically-attentive RNN for album summarization and storytelling. *CoRR*.

Luowei Zhou, Chenliang Xu, Parker A. Koch, and Jason J. Corso. 2016. Image caption generation with text-conditional semantic attention. *CoRR*.