Joint Visual Context for Pedestrian Captioning

Quan Liu[1,2,4], Yingying Chen[3,4], Jinqiao Wang[3,4], and Sijiong zhang[1,2,4]

[1] National Astronomical Observatories/Nanjing Institute of Astronomical Optics&Technology, Chinese Academy of Sciences, Nanjing 210042, China
[2] Key Laboratory of Astronomical Optics&Technology, Nanjing Institute of Astronomical Optics&Technology, Chinese Academy of Sciences, Nanjing 210042, China
[3] National Lab of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China
[4] University of Chinese Academy of Sciences, Beijing 100190, China
{quanliu, sjzhang}@niaot.ac.cn;{yingying.chen, jqwang}@nlpr.ia.ac.cn

Abstract. Image captioning is a fundamental task connecting computer vision and natural language processing. Recent researches usually concentrate on generic image captioning or video captioning among thousands of classes. However, they can not effectively deal with a specific class of objects, such as pedestrian. Pedestrian captioning is critical for analysis, identification and retrieval in massive collections of data. Therefore, in this paper, we propose a novel approach for pedestrian captioning with joint visual context. Firstly, a deep convolutional neural network (CNN) is employed to obtain the global attributes of a pedestrian (e.g., gender, age, and actions), and a Faster R-CNN is utilized to detect the local parts of interest for identification of the local attributes of a pedestrian (e.g., cloth type, color type, and the belongings). Then, we splice the global and local attributes into a fixed length vector and input it into a Long-Short Term Memory network (LSTM) to generate descriptions. Finally, a dataset of 5000 pedestrian images is collected to evaluate the performance of pedestrian captioning. Experimental results show the superiority of the proposed approach.

Keywords: Image captioning, pedestrian description

1 Introduction

Effectively describing the content of an image relies on rich semantic knowledge of a visual scene including the location, objects, attributes and actions etc. This is a particular challenging task in computer vision, but simultaneously it could help people understand the content of an image more directly compared to the fixed-categories image classification or object detection tasks. The majority of recent researches[1–3] have attempted to describe generic images with fixed vocabularies of visual concepts. Some object detection and region description methods[4, 5] have been added to expand the label space.

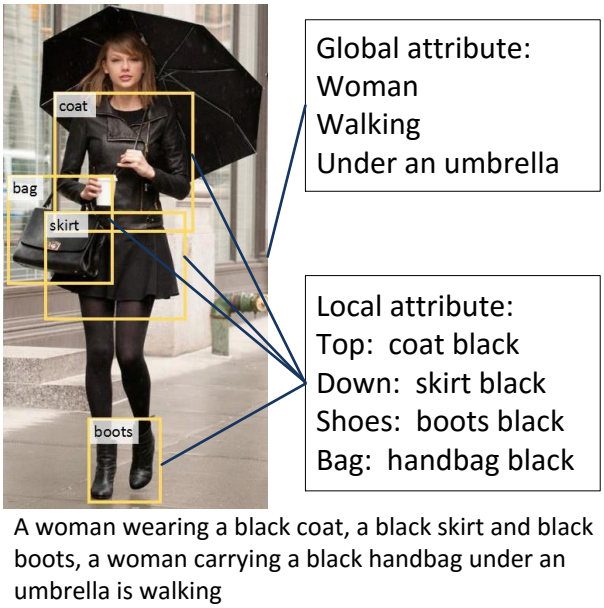

Fig. 1. An example of pedestrian captioning.

However, due to the limitation of the collected visual concepts and sentence templates of existing models, it has difficulties in generating detailed description for a specific target, such as pedestrian. In this paper, we strive to take a step for generating dense captions for a pedestrian image. Pedestrian captioning is to help a machine "understanding" a pedestrian image by several sentences. The rich description includes global attributes (e.g., gender, age, and action) and local attributes (e.g., cloth type, color type, and the belongings). Pedestrian captioning is critical for many applications. Taking video surveillance as an example, detailed description of the pedestrian is conducive to quickly retrieve the target of interest. Moreover, it's a basic part of our task to describe the dressing of a person, as a result, this technology can also provide intelligent recommendation for online shopping such as buying clothes, bags etc. For image captioning, many previous methods have attempted to go from an image to caption by combining a deep Convolutional Neural Network (CNN) with a Recurrent Neural Network (RNN). Mao et al.[1] proposed a multi-modal Recurrent Neural Network model (m-RNN) for generating novel sentence descriptions of images, which was the first work incorporating the RNN in a deep multimodal architecture. A. Karpathy et al.[4] also put forward an m-RNN model. However, their RNN inputs the image information only at the first time step in comparison to[1]. Since Long Short-Term Memory network (LSTM), as a special variation of RNN, has achieved remarkable results on nature language processing tasks such as speech recognition[6] and machine translation[7, 8], recent approaches introduced LSTM to image caption and achieved a significant improvement, such

as NIC model[2] and LRCN model[3]. Wu et al.[9] incorporated high-level attribute vector instead of high-level image features into a CNN-RNN framework and made further progress in image caption task.

In this paper, to generate a rich description for a pedestrian image, we pay attention to the description of a specific object class, then generate individual and particular description for a pedestrian image, including global and local attribute information as well as the pedestrian activities. Figure 1 gives an example of pedestrian caption, where we can see that the two sentences give a more detail description including the global and local characteristic for this girl. To obtain these personalized attributes, we first train a global attribute classifier based on CNN to predict the global characteristics of pedestrians, and use a part detector based on Faster R-CNN[10] with a deep CNN to locate each part of the pedestrian for the local attributes. Then, we splice the global and local predictive probability distributions into a fixed length attribute vector and input it into a LSTM-based language model to generate a structured and detailed description of the pedestrian. To evaluate the proposed approach, a dataset of 5000 pedestrian images is collected and published online and this is the first dataset for pedestrian captioning[1].

2 METHODOLOGY

The architecture of our approach is illustrated in Figure 2. The model consists of two parts: attribute analysis and sentence generation. The attribute analysis is decomposed into global and local attribute classification, which eventually generates the prediction probability for each attribute. To generate sentences, a recurrent neural network(RNN) is trained to predict each word of the sentencewe adopt the LSTM architecture for the RNN. Recent researches[1–3] usually input the high level image features into the LSTM model, however this approach cannot work well in describing local features. Inspired by[9], we input a high-level attribute vector that contains the prediction probabilities for particular attributes generated by the image analysis part into the LSTM model to generate the sentence. $S_{0:N}$ is the words that the LSTM model generated at every time step, more details about the prediction procedure will be illustrated in section 2.2.

2.1 Global and Local Attributes

The salient context extracted from an entire image is often treated as the global attribute. The global attribute recognition can be seen as a traditional classification problem, so we design a deep CNN to predict the probability of a particular attribute, which is a part of the attribute vector. We use a powerful VGGNet[11] pretrained on ImageNet[12] to extract the global attribute. This

[1] Dataset can be downloaded at:
`www.nlpr.ia.ac.cn/iva/homepage/jqwang/pedestrian_caption_dataset.zip`

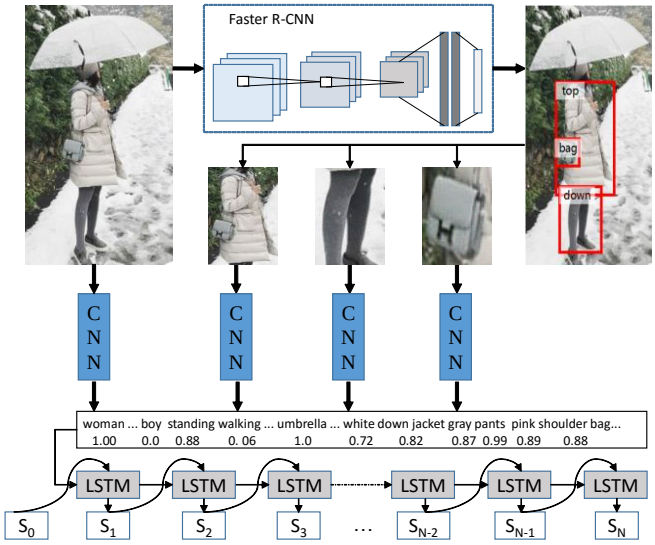

Fig. 2. Overview of pedestrian captioning.The model includes attribute analysis and sentence generation.

model is finetuned on the pedestrian dataset that we created and labeled and will be illustrated in section 3.

To obtain local attribute, we detect the important parts of a pedestrian and then recognize the categories and color attributes. In this paper, we detect the local attributes with a Faster Region Convolutional Neural Network (Faster R-CNN). Following the common training method of object detection, our detection model is finetuned on the pedestrian dataset with annotated groundings to regions of different parts of the pedestrian. Then, we design a deep CNN to analyze the category and color of each part in a pedestrian image. Thereby, we achieve local information like what kind of clothes the man is wearing and what color it is. Both global and local attributes are shown in Table 1. At last, we splice the global and local predictive probability distributions into a fixed length attribute vector whose length is the size of the attribute set. Then, we will be able to take the vector as an input and apply a LSTM-based language model to generate pedestrian captions.

2.2 Sentence Generator

Recent advances in machine translation have shown that a powerful sequence model is the guarantee to achieve state-of-the-art results. Thus, we use a Long-Short Term Memory (LSTM) network for sequence generation. LSTM is a particular variant of RNN and has achieved great success in translation[7, 8] and sequence generation[13] tasks. The choice of LSTM is up to its ability to deal with gradient vanishing and exploding, the most common challenge in designing

Table 1. Global and local attributes samples for pedestrian

attribute		words
global	gender	girl,boy,man,woman,old man/woman etc
	actions	walking,running,holding a baby/umbrella etc
local	top	T-shirt,shirt,coat,down jacket,suit etc
	down	pants,shorts,skirt,dress etc
	bag	handbag,shoulder bag,backpack,suitcase etc
	color	black,white,red,blue,yellow,gray,green etc

and training RNNs. Similar to[9], given the image and annotation sentence, we train the LSTM model by maximizing the probability of the correct captions. However, unlike typical approaches, we utilize the attribute vector described in the previous section rather than high level image features as the input. The formulation can be written as follows:

$$\log p(S|V_{attr}) = \sum_{t=1}^{N} \log p(S_t|S_{0:t-1}, V_{attr}) \tag{1}$$

where $S_0, ..., S_L$ is a sequence of words, each word S_t is represented in one-hot vector whose dimension is equal to the size of words dictionary. Note that S_0 and S_N respectively represent the special start and end token which donates the start and end of the sentence. $p(S_t|S_{0:t-1}, V_{attr})$ is the probability of the next word S_t given the previous words $S_{0:t-1}$ and the attribute vector V_{attr}.

We train the LSTM language generation model to predict each word of the final caption sentence in an unrolled form. Given the attribute vector V_{attr} and previous sentence $S_{1:t-1}$, the probability distribution of the next word is predicted in a sequence by iterating the following recurrence relation:

$$b_v = W_{hv} V_{attr} \tag{2}$$

$$h_t = LSTM(W_{hx} x_t, W_{hh} h_{t-1}, b_v(t=1), b_h) \tag{3}$$

$$p_t = softmax(W_{ho} h_t + b_o) \tag{4}$$

where $W_{hv}, W_{hx}, W_{hh}, W_{ho}$ and b_h, b_o are the learnable parameters and we provide the embedded atrributes vector b_v to the RNN only at the first iteration, because we found it works better than at each time step. At each time step the hidden state(h_t) of the LSTM layer is used to predict a distribution(p_t) over the words in the vocabulary. Particularly, if the stop word is predicted, a complete sentence has been generated.

All the parameters in LSTM are leant by minimizing the following loss function which is the sum of negative log likelihood of the correct word at each

step:

$$L(S, V_{attr}) = -\sum_{t=1}^{N} logp_t(S_t) \tag{5}$$

Stochastic Gradient Descent (SGD) is used to solve Eq.5 with mini-batches of 10 image-sentence pairs. The embedding size of attributes is set to 65, while the sizes of the embedding word and hidden state are set to 1000 in all the experiments. At time step $t = -1$, we set $x_{-1} = 0$ and $h_{-1} = 0$. The LSTM memory state is initialized to the range (-0.1, 0.1) with a uniform distribution. In inference process, the first sentence uses beam search (beam size = 3) and the second is sampled from the output distribution and choose the most probable caption which is different to the first sentence.

3 EXPERIMENTS

Since we are stepping towards generating a well-structured description of a specific pedestrian, there is no public dataset available for pedestrian captioning. Therefore, we establish a new dataset of pedestrian caption to evaluate our approach. For fair comparison, we implement some representative methods trained on our dataset for comparison. The generated captions are evaluated on multiple widely used evaluation metrics.

3.1 Dataset

Most previous works in image captioning[1–4] are evaluated on Flickr8k[14], Flickr30k[15] and Microsoft COCO[16] dataset. However, these datasets describe the diversity of the multiple objectives, and there is no detailed descriptions for each pedestrian, thus it is not suitable for evaluating our proposed generation method. Therefore, we initially select 5,000 pedestrian images under different scenes, manually adding descriptions as the ground truth. Of which, 4500 images are randomly selected to train our LSTM-based language model and 500 images for testing. These pictures cover a variety of seasons, time, and back-ground locations. Each image has two descriptions, the format can be described with reference to Figure 1.

Table 2. Results of our approach with only global attributes, local attributes input and global+local attributes input on BLEU-1,2,3,4, METEOR, ROUGE and CIDEr metrics.

Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE	CIDEr
`global attributes`	0.743	0.616	0.525	0.439	0.325	0.677	1.001
`local attributes`	0.802	0.711	0.646	0.586	0.455	0.817	2.620
`global+local attributes`	0.866	0.806	0.755	0.702	0.516	0.871	3.084

Table 3. BLEU-1,2,3,4, METEOR, ROUGE and CIDEr metrics compared with the typical methods and our approach on our own dataset.High is good in all columns.

Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE	CIDEr
LRCN[3]	0.774	0.677	0.600	0.531	0.405	0.773	2.207
Google-NIC[2]	0.803	0.735	0.679	0.633	0.436	0.804	2.880
Ours-all attributes	**0.866**	**0.806**	**0.755**	**0.702**	**0.516**	**0.871**	**3.084**

Fig. 3. Examples of pedestrian captioning with our approach.

3.2 Main results

To evaluate the quality of sentences generated by the language model for the given images and their reference sentences, we adopt the following four widely used evaluation metrics in image caption: BLEU[17], METEOR[18], ROUGE[19], and CIDEr[20]. All the scores are calculated using the coco-evaluation code, so the format of our ground truth file is similar to MS COCO dataset's test file format. To demonstrate the effectiveness of our approach, we first evaluate our approach by using different attributes as the input to the LSTM model, "global attributes", "local attributes" and "global+local attribute" respectively denote model with global attributes input, local attributes input and all the attributes input. Results are shown in Table 2. Obviously, local attributes plays a crucial role in the caption task. With all the attributes inputted into the LSTM model, we achieve a better performance. We also implement some baseline models that are representative approaches in the field of image captioning on our dataset. We show the results of baseline models and our approach on our own datasets in Tables 3. As shown in the table, our method has achieved remarkable results in the mainstream of several evaluation metrics and the scores are much higher than the LRCN[3] and NIC[2] model.We also show some example pedestrian captions generated by our approach in Fig 3. We can see that our method can generate descriptions with high quality for both global or local information, which mainly

benefits from utilizing the Faster R-CNN model to detect local attributes as well as inputting the attributes vector to the LSTM language model directly.

4 Conclusions

In this paper, we propose a novel caption approach to generate descriptions for pedestrian on joint visual context. We pay more attention to personalized descriptions for pedestrians, thus we use a Faster R-CNN model along with deep CNNs to classify the different parts of the pedestrian body to get the local attributes. Along with the global attributes, we create an attribute vector which contains the probabilities of various attributes and input it into a LSTM-based language model to generate sentences. Experiments and comparison with multiple popular evaluation metrics on our own dataset show the promise of the proposed approach.

References

1. Mao, Junhua and Xu, Wei and Yang, Yi and Wang, Jiang and Huang, Zhiheng and Yuille, Alan: Deep captioning with multimodal recurrent neural networks (m-rnn). arXiv preprint arXiv:1412.6632 (2014)
2. Vinyals, Oriol and Toshev, Alexander and Bengio, Samy and Erhan, Dumitru: Show and tell: A neural image caption generator. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 3156–3164 (2015)
3. Donahue, Jeffrey and Anne Hendricks, Lisa and Guadarrama, Sergio and Rohrbach, Marcus and Venugopalan, Subhashini and Saenko, Kate and Darrell, Trevor: Long-term recurrent convolutional networks for visual recognition and description. Proceedings of the IEEE conference on computer vision and pattern recognition, 2625–2634 (2015)
4. Karpathy, Andrej and Fei-Fei, Li: Deep visual-semantic alignments for generating image descriptions. Proceedings of the IEEE conference on computer vision and pattern recognition, 3128–3137 (2015)
5. Fang, Hao and Gupta, Saurabh and Iandola, Forrest and Srivastava, Rupesh K and Deng, Li and Dollár, Piotr and Gao, Jianfeng and He, Xiaodong and Mitchell, Margaret and Platt, John C and others: From captions to visual concepts and back. Proceedings of the IEEE conference on computer vision and pattern recognition, 1473–1482 (2015)
6. Graves, Alex and Jaitly, Navdeep: Towards End-To-End Speech Recognition with Recurrent Neural Networks. ICML.14, 1764–1772 (2014)
7. Sutskever, Ilya and Vinyals, Oriol and Le, Quoc V: Sequence to sequence learning with neural networks. Advances in neural information processing systems, 3104–3112 (2014)
8. Cho, Kyunghyun and Van Merriënboer, Bart and Bahdanau, Dzmitry and Bengio, Yoshua: On the properties of neural machine translation: Encoder-decoder approaches. arXiv preprint arXiv:1409.1259 (2014)
9. Wu, Qi and Shen, Chunhua and Liu, Lingqiao and Dick, Anthony and van den Hengel, Anton: What value do explicit high level concepts have in vision to language problems?. Proceedings of the IEEE conference on computer vision and pattern recognition, 203–212 (2016)

10. Ren, Shaoqing and He, Kaiming and Girshick, Ross and Sun, Jian: Faster r-cnn: Towards real-time object detection with region proposal networks. Advances in neural information processing systems, 91–99 (2015)
11. Simonyan, Karen and Zisserman, Andrew: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
12. Deng, Jia and Dong, Wei and Socher, Richard and Li, Li-Jia and Li, Kai and Fei-Fei, Li: Imagenet: A large-scale hierarchical image database. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 248–255 (2009)
13. Graves, Alex: Generating sequences with recurrent neural networks. arXiv preprint arXiv:1308.0850 (2013)
14. Hodosh, Micah and Young, Peter and Hockenmaier, Julia: Framing image description as a ranking task: Data, models and evaluation metrics. Journal of Artificial Intelligence Research.47, 853–899 (2013)
15. Young, Peter and Lai, Alice and Hodosh, Micah and Hockenmaier, Julia: From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. Transactions of the Association for Computational Linguistics.2, 67–78 (2014)
16. Lin, Tsung-Yi and Maire, Michael and Belongie, Serge and Hays, James and Perona, Pietro and Ramanan, Deva and Dollár, Piotr and Zitnick, C Lawrence: Microsoft coco: Common objects in context. European Conference on Computer Vision, 740–755 (2014)
17. Papineni, Kishore and Roukos, Salim and Ward, Todd and Zhu, Wei-Jing: BLEU: a method for automatic evaluation of machine translation. Proceedings of the 40th annual meeting on association for computational linguistics, 311–318 (2002)
18. Lavie, Michael Denkowski Alon: Meteor universal: Language specific translation evaluation for any target language. ACL 2014. 376 (2014)
19. Lin, Chin-Yew: Rouge: A package for automatic evaluation of summaries. Text summarization branches out: Proceedings of the ACL-04 workshop. 8 (2004)
20. Vedantam, Ramakrishna and Lawrence Zitnick, C and Parikh, Devi: Cider: Consensus-based image description evaluation. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 4566–4575 (2015)