A novel framework for image description generation

Qiang Cai[1], Ziyu Xue[1], Xiaoyu Zhang[2], Xiaobin Zhu[1], Wei Shao[3], and Lei Wang[4]

1.Beijing Key Laboratory of Big Data Technology for Food Safety, School of Computer and Information Engineering, Beijing Technology and Business University, No. 11, Fucheng Road, Haidian District, Beijing, China
2.Institute of Information Engineering, Chinese Academy of Sciences, A, No. 89, Minzhuang Road, Haidian District, Beijing, China.
3.CCTV High-Tech Television Development CO.,Ltd, 12B, Huabao building, Lianhua bridge, Haidian District, Beijing, China.
4.Information Technology InstituteAcademy of Broadcasting Science, SAPPRFT, NO.2 Fuxingmenwai Street, Xicheng District, Beijing, China.
{caiq@th.btbu.edu.cn,xueziyucs@gmail.com, zhangxiaoyu@iie.ac.cn,brucezhucas@gmail.com, shaowei@cctvht.cn,wanglei@abs.ac.cn}
http://www.springer.com/lncs

Abstract. The existing *image description* generation algorithms always fail to cover rich semantics information in natural images with single sentence or dense object annotations. In this paper, we propose a novel semi-supervised generative visual sentence generation framework by jointly modeling *Regions Convolutional Neural Network* (RCNN) and improved *Wasserstein Generative Adversarial Network* (WGAN), for generating diverse and semantically coherent sentence description of images. In our algorithm, the features of candidate regions are extracted with RCNN and the enriched words are polished by their context with an improved WGAN. The improved WGAN consists of a structured sentence generator and a multi-level sentence discriminators. The generator produces sentences recur-rently by incorporating region-based visual and language attention mechanisms, while the discriminator assesses the quality of generated sentences. The experimental results on publicly available dataset show the promising performance of our work against other related works.

Keywords: WGAN, Image Description, RCNN

1 Introduction

Image description can be used in many real application scenarios, e.g., video retrieval and automatic video subtitling. It is a challenging task due to the intersection of computer vision, natural language processing and other disciplines. In recent years, image description generation has attracted lots of focus in research

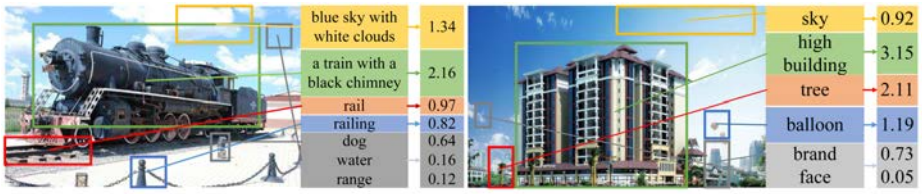

Fig. 1. Image description of dense regions.

domain. And great progress has been made in labeling images with a pre-defined close set of visual categories, as shown in Fig. 1.

Image description generation is thoroughly studied by the computer vision communities, where some well established models have been developed. A majority of algorithms proposed to describe image using individual semantic sentence. Everingham et al.[1] focused on labeling images with a fixed set of visual categories, while Kulkarni et al.[2] relied on hard-coded visual concepts and sen-tence templates to generate sentence descriptions. These two methods all ig-nored the location of dense objects in images. Recent works[3],[4],[5] intended to use Recurrent Neural Network to generate the dense image descriptions and corre-sponding location information, which were spliced according to the trained sen-tence model. However, the simplified sentence models always generate single smoothly sentences using fixed set of visual categories without consideration of rich vocabulary, which will fail to cover rich underlying semantics of images.

To address above-mentioned issue, we propose a novel approach for image description generation, by cooperation between modeling Convolutional Neural Network (CNN) and improved Wasserstein Generative Adversarial Network (WGAN). When using CNN in visual feature acquisition, alignment objective method will be used in order to avoid the deviation of regions from words. We have described the transformations that map every image and sentence into a set of vectors in a common h-dimensional space. After matching the dense ob-jects and words, the higher score words are more likely to be used to generate sentence. The improved WGAN model is an adversarial training mechanism, which jointly by a structured sentence generator and multi-level sentence dis-criminators. The discriminator learns to distinguish between real and synthesized sentences which from generator. The visual features with higher scores are used as the input in improved WGAN to generate sentences.

The rest of the paper is organized as follows. Section 2 overviews the related works. In Section 3, we elaborate our method. The experimental evaluation is given in Section 4, and we draw conclusion in Section 5.

2 Related work

Image description generation is an active area of research on its own [5]. Traditional methods concentrated on features extraction. Deep model is developing

rapidly because of its accuracy and applicability, which generally divide into dense object recognition, visual sentence generation and visual paragraph generation. Generating high score sentence plays an important role in paragraph generation. In this section, we will briefly review sentence generation related works.

Kiros R et al. [4] proposed a model that CNN is used to feature extraction and object detection. Borrowing the above method of feature extraction, Karpathy A et al. [3] modeled Recurrent Neural Networks framework learns a joint image-label embedding to characterize the semantic label dependency as well as the image-label relevance. However, RNN failed to adapt to generate sentences, because it has no time sequence. RNN with Long Short-Term Memory (LSTM) can effectively model the long-term temporal dependency in a sequence [6]. To rea-son about long-term linguistic structures with multiple sentences, hierarchical recurrent network has been widely used to directly simulate the hierarchy of language. In [8],[9], a framework based on CNN-LSTM was developed to image description that the generated descriptions significantly outperform baselines. However, LSTM uses fixed set of visual categories without consideration of rich vocabulary, which will fail to cover rich underlying semantics of images.

Dai b et al. [15] proposed Conditional GAN (CGAN) to generate sentence, which proposed a cooperation model between native CNN and CGAN. CGAN overcomes the difficulty by Policy Gradient, a strategy stemming from Reinforcement Learning, which allows the generator to receive early feedback along the way. Liang X et al. [10] proposed Recurrent Topic-Transition GAN (RTT-GAN) to generate paragraph, which composed of LSTM model and generated diverse and semantically coherent sentences by reasoning over both local semantic re-gions and global sentence context. The generator selectively incorporates visual and language cues of semantic regions to produce each sentence. Nevertheless, to the best of our knowledge, the previous methods failed to consider the align-ment between dense objects and words, which had a great impact on the accu-racy of the words.

3 Our method

The framework of our algorithm is shown as Fig.2. The red box part illustrates the flowchart of RCNN model. The green box part details the procedure of sentence generation. In our algorithm, we first generate descriptions for candidate image regions with RCNN. Sequentially, the sentence generator will generate meaningful sentences by incorporating the fine-grained visual and textual cues in a selective way. To ensure quality of sentences, we apply a sentence discriminator on each generated sentence to measure the plausibility and smoothness of semantic transition with preceding sentences. The generator and discriminator are learned jointly within an adversarial framework.

3.1 Learning to align visual and language data

Image-Feature Extraction. Following prior works, we observe that sentence de-scriptions make frequent references to objects and their attributes. Therefore,

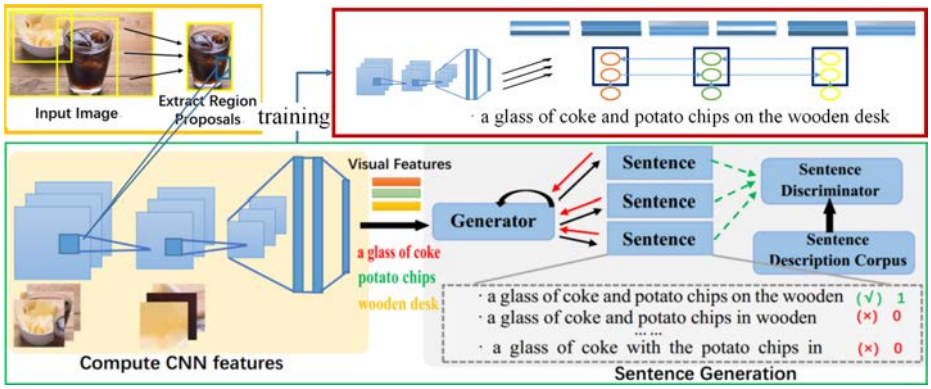

Fig. 2. The framework of our algorithm.

we adopt RCNN to detect object, which followed by [3]. The feature vectors for de-tected regions are computed as below:

$$v = W_M[CNN_{\theta c}(I_p)] \ . \tag{1}$$

where θ_c denotes CNN model parameter sets; the dimension of matrix W_M is $h * 4096$ (h is the size of the multimodal embedding space). Every image is transformed as a sequence of N words, which encoded in a representation vector. Every sequence will be transformed into a h-dimensional vector $\{V_i|i = 120\}$.

Objects Alignment. We formulate an image-sentence score as a function of the individual region-word scores, which used a Bidirectional Recurrent Neural Network (BRNN) to compute the word representations. Like the model of Karpathy et al.[3] interprets the dot product vi T-st between the i-th region and t-th word as a measure of similarity and use it to define the score between image k and sentence l as Eq.2:

$$S_{\mathrm{kl}} = \sum_{t \in g_l} \sum_{i \in gi} \max(0, v_i^T s_t) \ . \tag{2}$$

where, g_i denotes the $i-th$ image fragments. l denote the images and sentences in the training set respectively. s_t is a function of all words in the entire sentence.

Every word s_t aligns to the single best image region. As we shown in the experiments, this simplified model also leads to improvements in the final ranking performance. Assuming that $i = l$ denotes a corresponding image and sentence pair, the final max-margin, structured loss remains:

$$\partial(\theta) = \sum_i [\sum_{l(images)} \max(0, S_{il} - S_{ii} + 1) + \sum_{l(sentence)} \max(0, S_{li} - S_{ii} + 1)] \ . \tag{3}$$

This objective encourages aligned image-sentences pairs to have a higher score than misaligned pairs, by a margin. We get the high scores to form word vectors.

3.2 Sentence Generator

The architecture of the generator G shown in Fig 3, which recurrently retains different levels of context states with a hierarchy constructed by sentence LSTM, word LSTM, and two attention modules. Firstly, the visual attention module selectively focuses on semantic regions, generating the visual representation of the sentence from the visual vectors. The sentence LSTM can be able to encode a topic vector for a new sentence. Secondly, the language attention module embedded in local phrases of focused semantic regions to facilitate word generation from the word LSTM, meanwhile learnt to incorporate linguistic knowledge.

LSTM for Sentence Description. Sentence LSTM is a single layer model. When we training the model, it takes the image pixels I and a sequence of input vectors from the region description $(v_1, ..., v_T)$. And computes a sequence of hidden states $(h_1, ..., h_t)$ and a sequence of outputs $(y_1, ..., y_t)$ by iterating the following recurrence relation from 1 to T. An attention mechanism is applied in the visual features V of all semantic regions, resulting in a visual context vector $f_v t$ which represents the next sentence at $t - th$ step:

$$\mathrm{y}_t = soft\max\{W_{oh} \times f(W_{hx}x_t + W_{hh}h_{t-1} + b_h + W_{hi}[CNN_{\theta_c}(i)]) + b_o\} \quad (4)$$

here $W_h i$, $W_h x$, $W_h h$, $W_o h$, x_i and b_h, bo are learnable parameters. $CNN_\theta c(i)$ is the last layer in all net. y_t is the output vector, which holds the log probabilities of words in the dictionary, according to the highest score.

Word LSTM with Language Attention. To get plausibility and smoothness sentences, the proposed model should recognize and describe substantial details such as objects, attributes, and relationships, word replacement model is appro-priate for context. We selectively incorporate the embedding of local phrases based on the topic vector and use Word LSTM to generate the better word repre-sentation. Considering that each local phrase relates to the respective visual fea-ture, we reuse the visual attentive weights to enhance the language attention. By computing the contribution of a word to the whole sentence in hidden layer, word LSTM embeds the words which have the high contribution into sentence.

3.3 Sentence Discriminator

The sentence discriminator D_s aim to distinguish between real sentences and synthesized ones based on the linguistic characteristics of a natural sentence description. In our algorithm, the discriminator D_s is an LSTM that recurrently takes the input word embedding within a sentence, meanwhile produces a real value plausibility score of the synthesized sentence, which evaluates the plausibility of individual sentences.

The discrete nature of text samples disappeared gradient back-propagation from the discriminators to the generator. To overcome this problem, we use deterministic approximation. At each word generation step, we select the most probable word according to the soft-max emission distribution, and propagate

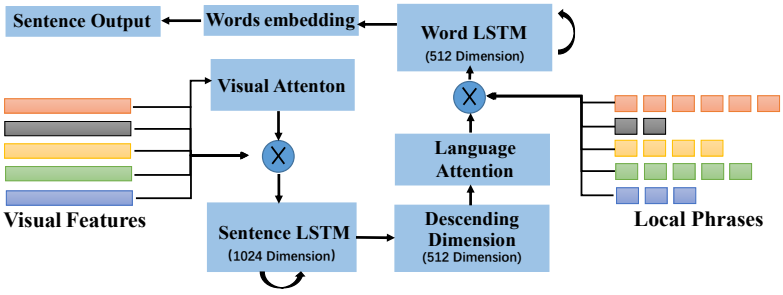

Fig. 3. The diagram of generator.

gradient only through the words. More simply, we apply max-pooling operation over the soft-max to overcome the gradient disappearance in GAN model. Fig.4 shows the discriminator of the framework.

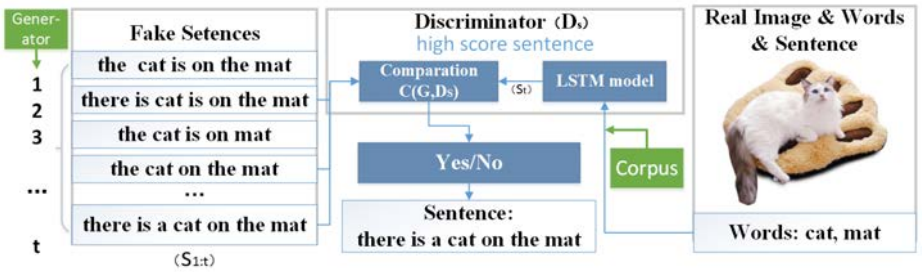

Fig. 4. The diagram of discriminator.

The Sentence Input (as shown in Fig.4 with Fake Sentences) is the output of generator in Section 3.2. The objective of the adversarial framework, which followed WGAN[10] method by minimizing an approximated Wasserstein distance, is thus written as:

$$\min_{G} \max_{D^s} C(G, D^s) = E_{\hat{s} \sim S}[D^s(\hat{s})] - E_{\hat{s} \sim S_{1:t}}[D^s(\hat{s})] \tag{5}$$

where $\hat{s}$ is the true sentence, S and $S_(1:t)$ denote the true data distributions of generation sentences and true sentences, which are constructed from a sentence description corpus. sentence discriminator D^s that optimizes a critic between

real/fake sentences. $\hat{s} \sim S(1 : t)$ distribution of generated sentences by the generator G. The objective for the generator G is:

$$G^{'} = \arg\min \max_{G,D^s} \gamma C(G, D^s) \tag{6}$$

where γ is a balancing parameter fixed at 0:001 in implementation. The optimization is performed in an alternating min-max manner [10].

4 Experiments

To validate the effectiveness of our proposed algorithm, we conduct experiments on MSCOCO datasets: the dataset contains 123,000 images respectively and each is annotated with five sentences using Amazon Mechanical Turk. Through-out all the experiments, 5,000 images are used in both validation and testing.

To evaluate the performatnce of the LSTM model in generating sentences, we adopt VGGNet [11] in the full image experiments to calculate BLEU [12], METEOR [13] and CIDEr [14] scores. We evaluates a candidate sentence by matching five reference sentences written by humans. Word generator can optimize individual vocabulary. This step does not improve the accuracy of the sentence, because the above steps have generated vocabulary and phrases already. However, after the vocabulary of the replacement, we can get a smoother sentence.

4.1 Objective Alignment Evaluation

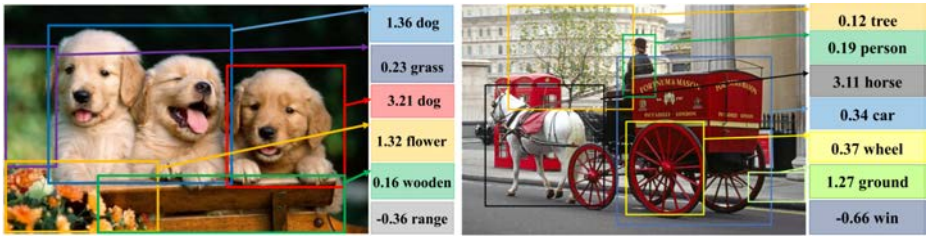

Fig. 5. Example alignments predicted by our model.

RCNN is pre-trained on ImageNet and fine-tuned on the 200 classes of the ImageNet Detection Challenge. In our algorithm, the top 19 ranked regions are selected. Image-Sentence alignment evaluation is used to build the Image-Sentence rank in our work, which retrieve the most compatible test sentence and visualize the highest-scoring region for each word and the associated scores, as shown in Fig 5.

Inputting all the words into the generator in order, until the sentence generated in Section 3.2. The word LSTM is recurrently forwarded to optimize the

	R@1	R@5	R@10	Med r
Our model: 1K test images	38.4	69.9	80.5	1.0
Our model: 5K test images	16.5	39.2	52.0	9.0

Table 1. Image-Sentence ranking experiment results.

Model	B-1	B-2	B-3	B-4	METEOR	CIDEr
Vinyals[9]	48.0	28.1	16.6	10.0	15.7	38.3
Google NIC	66.6	46.1	32.9	-	-	-
LRCN[5]	62.8	44.1	30.4	-	-	-
MS Research[7]	-	-	-	21.1	20.7	-
Chen and Zitnick[8]	-	-	-	19.0	20.4	-
Andrej Karpathy and Li Fei-Fei[3]	62.5	45.0	32.1	23.0	19.5	66.0
Our model	**62.7**	**45.3**	**32.5**	**23.5**	**19.7**	**66.7**

Table 2. Evaluation of full image predictions on 1,000 test images.

words iteratively. We report the median rank of the closest ground truth result and Recall@K, which can get a correct sentence in top K results. The result of these experiments can be found in Table 1[3].

4.2 The Process of Adversarial Training

The generator in WGAN can be regarded as an image captioning model due to the lack of adversarial loss, similar as Word-Concat. The discriminator identify with the closest ground truth result, comparing with corpus. It justifies that the sentence plausibility is very critical for generating long, convincing sentences.

Table 2 shows the model evaluation of full image predictions on 1,000 test images. Where B-n is BLEU-score that uses up to n-grams, its output is always a number between 0 and 1. This value indicates how similar the candidate text is to the reference texts, with values closer to 1 representing more similar texts.

We compare our method with others. Vinyals et al.[9] use LSTM, which get the first word from through a bias term on the first step. Karpathy et al.[3] use Mul-timodel-RNN, the assembly result become better. Donahue et al.[5] use a 2-layer factored LSTM and GoogLeNet, which is a different CNN, and report results of a model ensemble. Others[8], [12] appear to work worse than ours, but this is likely in large part due to their use of the less powerful AlexNet[7] features. Compared to these approaches, our model generates more accurate descriptions.

4.3 Personalized Sentence Generation

Different with prior works, the proposed model supports the personalized sentence generation, which will produce diverse descriptions by first words. The generator can sequentially output diverse and topic-coherent sentences for an image. If we give two different first words, two models will produce two personalized sentences for the same image, respectively. We present qualitative results of our model in Fig.6.

4.4 Objective Alignment Evaluation

Fig. 6. Example sentences generated by the proposed model for test images.

The proposed model generates sensible descriptions of images as shown in Fig.6. The images shown in Fig.6 with blue background translated more accurate, the first prediction a large building with a clock tower on top of it does not appear in the training set. However, there are a large building, a clock tower, on top of are occurrence, which the model may have composed to describe the first image. However, picture of the running car, the model cannot accurately determine the state. The last two pictures, when judge the objects (especially color judgment), there have been some problems. They cannot accurately determine the color of the horse and water. Further, we will output the phrase more accurately by adjusting the recognition within the bounding boxes.

In general, we find a relatively large portion of generated sentences can be found in the training data. The proposed method can be repeated many times without any syntactic specification. Therefore, the generated sentence is more smooth and complete, and the generated words are relatively rich and appropriate.

5 Conclusion

In this paper, we propose a novel approach for cooperation between modeling CNN and RTT-GAN. When we use CNN to extract features, Image-Sentence alignment evaluation will be used to build the Image-Sentence rank, higher scores words will be selected to generate sentences. RTT-GAN uses heuristic rules to generate smoothly sentence with rich vocabularies. We evaluated the performance on full frame experiments and showed that our model outperforms retrieval baselines. As our future work, we will attempt to update RTT-GAN, improve levels and types on discriminators to enhance sentences recognition effect, and apply it to some other applications, such as paragraph generation.

References

1. M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. IJCV, 88(2):303C338, June 2010.
2. G. Kulkarni, V. Premraj, S. Dhar, S. Li, Y. Choi, A. C. Berg, and T. L. Berg. Baby talk: Under-standing and generating simple image descriptions. In CVPR, pp.1601-1608, 2011.
3. A. Karpathy and L. Fei-Fei. Deep visual-semantic alignments for generating image descrip-tions. In CVPR, pp.3128C3137, 2015.
4. R. Kiros, R. Salakhutdinov and R. S. Zemel. Unifying visual-semantic embeddings with mul-timodal neural language models, Computer Science, pp.3412C3415, 2014.
5. C. Zhang, Z. Xue, X. Zhu, Q. Huang, and Q. Tian, Boosted random contextual semantic space based representation for visual recognition, Information Sciences, 369:160-170, 10 Nov, 2016.
6. J. Wang, J. Yang, K. Yu, et al. Locality-constrained linear coding for image classification. In CVPR, pp.3360-3367, 2010.
7. H. Fang, S. Gupta, F. Iandola, R. Srivastava, L. Deng, P. Dollar, J. Gao, X. He, M. Mitchell, J. Platt, et al. From captions to visual concepts and back. arXiv:1411.4952, 2014.
8. X. Chen and C. L. Zitnick. Learning a recurrent visual representation for image caption gen-eration. CoRR, abs/1411.5654, 2014.
9. O. Vinyals, A. Toshev, S. Bengio, et al. Show and tell: A neural image caption generator, Computer Science, pp.3156-3164, 2015.
10. X. Liang, Z. Hu and H. Zhang. Recurrent topic-transition GAN for visual paragraph genera-tion. arXiv preprint arXiv: 1703.07022, 2017.
11. K. Simonyan and A. Zisserman, Very Deep Convolutional Networks for Large-Scale Image Recognition, Computer Science, pp.1543-1544, 2014.
12. Xiao-Yu Zhang, Shupeng Wang, Xiaobin Zhu, Xiaochun Yun, Guangjun Wu: Update vs. up-grade: modeling with indeterminate multi-class active learning. Neurocomputing (NEUCOM), 162, pp. 163-170, 2015.
13. M. Denkowski and A. Lavie, Meteor universal: Language specific translation evaluation for any target language, The Workshop on Statistical Machine Translation, pp 376-380, 2014.
14. R. Vedantam, C. L. Zitnick, and D. Parikh. Cider: Consensus-based image description evalua-tion. CoRR, abs/1411.5726, 2014.

15. Xiao-Yu Zhang, Shupeng Wang, Xiaochun Yun: Bidirectional active learning: a two-way ex-ploration into unlabeled and labeled dataset. IEEE Transactions on Neural Networks and Learning Systems (TNNLS), 26(12), pp. 3034-3044, 2015.