Image Captioning with Relational Knowledge

Huan Yang, Dandan Song(✉), and Lejian Liao

Beijing Engineering Research Center of High Volume Language Information Processing and Cloud Computing Applications, Beijing Key Laboratory of Intelligent Information Technology, School of Computer Science and Technology, Beijing Institute of Technology, Beijing 100081, China
{yh,sdd,liaolj}@bit.edu.cn

Abstract. People have learned extensive relational knowledge from daily life. This is one of the facts that enables human to describe the information from images easily. In this paper, we propose a novel framework called Image Captioning with Relational Knowledge ($ICRK$) that combines relational knowledge with image captioning model and utilizes relational knowledge to strengthen the learning process of representing words. As more precise syntactic and semantic word relationships were learned, the image captioning model acquires more semantic features that help to generate more accurate image descriptions. Experiments on several benchmark datasets, using automatic evaluation metrics, have all demonstrated that our model can significantly improve the quality of image captioning.

Keywords: Image captioning · Relational knowledge
Word embedding

1 Introduction

Image captioning, a challenging task which combines Natural Language Processing with Computer Vision, has attracted more and more attention recently. Generating the descriptions of images automatically not only is of benefit for applications like image retrieval but also helps visually impaired people to see the world. It is so important that has been treated as a core problem in Computer Vision.

Recognizing and describing details in images is natural and easy for people. However, it can be a challenging task for image captioning models. One of the important reasons is that when looking at an image, people are not just recognizing a large number of objects in it, but also able to detect the relationship between them. For example, when we see "*girl*" and "*bed*" in an image, we will naturally describe it as "*A girl is sitting on the bed*", and when given an object "*meal*" and the relationship "*is presented in*", we can also easily come up with "*tray*" as the object where the meal is presented in. Because of the relational

X. Geng and B.-H. Kang (Eds.): PRICAI 2018, LNAI 11013, pp. 378–386, 2018.
https://doi.org/10.1007/978-3-319-97310-4_43

knowledge people have, we can recognize the objects in the image more accurate and make a description more fluent. In contrast, the image captioning model cannot do it without learning this relational knowledge.

Continuous skip-gram model and continuous bag-of-word model (CBOW) and [12] have been proposed for computing continuous vector representations of words from very large data sets, and it has been proven that these models can learn high-quality word vectors from huge data. However, these models learned word representations from the continuously distributed representation of the context, so if there are little context information about two syntactically or semantically similar words, they cannot learn the relationship between them. In that case, when we put these word representations into image captioning model, the model that has learned relational knowledge will perform better than the model that hasn't learned. Furthermore, learning from the amount of context could be noisy or biased, and these word representations cannot reflect the inherent relationship between words.

In order to combine relational knowledge with image captioning model and get better word representations, we propose a novel model that incorporates the relational knowledge of words from knowledge graph into the learning process and treats relational knowledge as regularization function. Concretely, the main contribution of this paper is proposing a new image captioning algorithm which combines relational knowledge, and we define a new learning objective to strengthen the learning of word representation in image captions. We validate the effectiveness of this approach on several datasets in which we outperform competing methods and achieve state-of-the-art consistently across different evaluation metrics.

2 Related Work

Image captioning model can be divided into two categories generally: bottom-up and top-down. Bottom-up approaches use the visual concepts, objects detected from the image and pretrained neural network to get the words corresponding to these visual features, and then combine these words into sentences using language models. Representative works include [5,8], and these methods rely on the effectiveness of the visual detectors and the ability of language model to generate sentences. However, unlike bottom-up approaches need to detect visual concept, words and put them together, top-down approaches can be trained from end to end. These approaches [6,11,16,20] use a Convolutional Neural Network (CNN) to extract image features and combine these features with Recurrent Neural Network (RNN) to accomplish image captioning. The main difference between these approaches is that different methods use different CNN and RNN.

We notice that word representations are vital for image captioning no matter what model we use as the description of an image is organized by single words. Some recent effort, such as continuous skip-gram model and CBOW model [12], have attempted to learn word representations that can capture both the syntactic and the semantic information among words. However, in prior work [6], little change has been found in final performance of image captioning when adding these trained word vectors. In contrast, inspired by a popular study on the multi-relation model [2] that builds relationships between entities, we observe that there are also relationships between the objects in the image and this feature can be used in image captions. Instead of putting the word vectors which trained by word2vec [12] model into the image captioning model directly, we extract relational knowledge from the descriptions and extend the objective function of word2vec model by combining the relational knowledge as regularization function. What's more, instead of using the popular knowledge graphs, such as Freebase [1] and WordNet [14], to train our model, we build a knowledge base by our own which is tailored to this task without much noisy.

3 Proposed Model

3.1 Overall Framework

Following several previous works [6,11,16,20], we use a CNN to extract image features and RNN to connect images features with sentences features. In this work, our particular design is combining relational knowledge with image captioning model. We extract relational triplets from the descriptions of images and use this relational knowledge to construct semantic features, and then combine these semantic features with visual features of images to generate the descriptions of images automatically. Our overall image captioning model is illustrated in Fig. 1. We describe our method to construct semantic features based on relational knowledge in Sect. 3.2. In Sect. 3.3 we outline the architecture of our image captioning model.

3.2 Relational Knowledge with Word Representations

We adopt the continuous skip-gram model as the basis of the proposed relational knowledge embedding framework[1]. It is a word embedding model using a neural network architecture and has been proved efficient for learning high-quality distributed vector representations. The continuous skip-gram model focuses on finding word representations that are meaningful for predicting the surrounding words in a sentence.

[1] Note that although we use the continuous skip-gram model as an example to illustrate our framework, the similar framework can be developed on the basis of any other word embedding models.

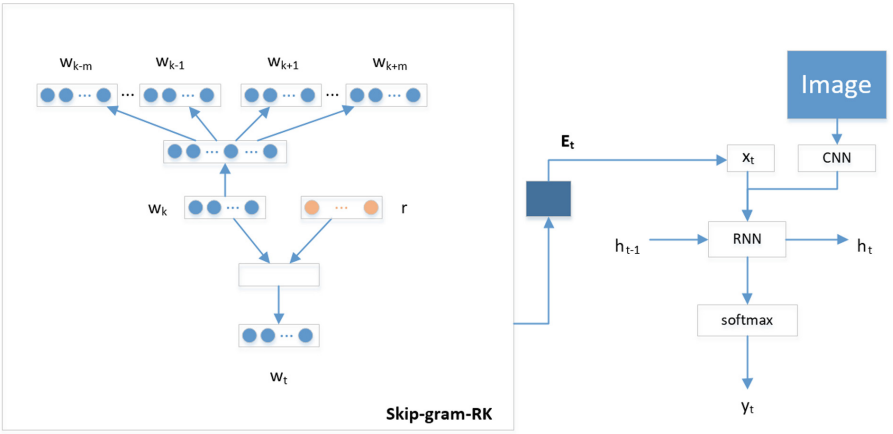

Fig. 1. An illustration of our image captioning model ICRK. It is comprised of a CNN, a RNN and our Skip-gram-RK model.

Given a sequence of training words w_1, w_2, w_3,...,w_K, the objective of the continuous skip-gram model is to maximize the log probability:

$$\xi = \sum_{w_k \in C} \log p\left(Context\left(w_k\right)|w_k\right) \tag{1}$$

where C are words in the vocabulary, $Context(w_k)$ is the training context $\{w_k - m, ..., w_k - 1, w_k + 1, ..., w_k + m\}$, and m indicates the context window size to be $2m + 1$.

Since the relational knowledge in knowledge graph is usually represented in the triplet $(head, relation, tail)$ (denoted (h, r, t)), each of which often can be extracted from text. The principle of previously developed translation-based model [2] is that $h + r \approx t$, if (h, r, t) holds, the embedding of the head entity h plus the embedding of the relationship r should be close to the tail entity t, otherwise $h + r$ should be far away from t.

Similarly to this approach, we extract the triple (w_h, r, w_t) from training data, and it consists of two words w_h, w_t and the relationship r contacting them. To combine the relational knowledge with word embedding model, we assume that relationships between words can be interpreted as translation operations and they can be represented by vectors. The basic idea of our model is that $w_h + r \approx w_t$. However, instead of learning vectors embedding by minimizing a margin-based ranking criterion over the training set which results in complex combined optimization problem [19], we adopt an objective to maximize the probability as below:

$$J = \sum_{r \in R_{w_h}} \log p(w_t|w_h + r) \tag{2}$$

To incorporate relational knowledge into word representations learning system, we get the following combined objective D:

$$D = \xi + \alpha J \tag{3}$$

where α is the combination coefficient. R_{w_h} contains all the relationships related to w_h. Our goal is to maximize the combined objective D.

Traditional neural networks often define the conditional probability $p\,(y|x)$ in $softmax$ function, which is impractical in this task due to the high cost of computing $\nabla \log p\,(y|x)$ in the case of having hundreds of words in the vocabulary ($10^5 - 10^7$ terms). In training process, we use negative sampling (NEG) [13] to solve this problem.

3.3 Join Relational Knowledge with Captioning Model

In general image captioning model, we often use CNN to extract image features and use RNN to combine the image feature with the corresponding caption. In this work, we adopt the Multimodal RNN mentioned in [6] as the captioning model, and we use a pretrained VGGNet [17] to extract spatial image features. Furthermore, by combining relational knowledge with the captioning model, our method attains the state-of-the-art performance.

We get the output word vectors E from the relational knowledge embedding model described in Sect. 3.2, and then use the E_t to represent the input vector x_t of the multimodal RNN, where E_t is the word encoding of the input word at timestep t. Besides the x_t, the multimodal RNN also takes the image pixels during training. It computes a sequence of outputs (y_1, ..., y_t) by iterating the following recurrence relation:

$$b_v = W_{hi}[CNN(I)] \tag{4}$$

$$h_1 = f(W_{hx}x_1 + b_h + b_v) \tag{5}$$

$$h_t = f(W_{hx}x_t + W_{hh}h_{t-1} + b_h) \quad t > 1 \tag{6}$$

$$y_t = softmax(W_{oh}h_t + b_o) \tag{7}$$

where W_{hi}, W_{hx}, W_{hh}, W_{oh}, b_h and b_o are learnable parameters, and $CNN(I)$ is the last layer of a CNN. We provide the image context vector b_v to the RNN only at the first iteration, which has been proven work better than at each time step in [6].

4 Experimental Results and Discussion

4.1 Datasets

To evaluate our proposed image captioning model, we experiment with MSCOCO [10] datasets. It contains 123,287 images, and we use the publicly available Karpathy splits [6] that have been used extensively in prior work to report our results. We get 113,287 images for training, 5,000 images respectively for validation and testing. Each image is annotated with 5 sentences.

We convert all sentences to lower case, discard non-alphanumeric characters and filter words whose frequency less than 5 in the training set, resulting in 9,488 words for training. We report our results using the standard automatic evaluation metrics, BLEU [15], METEOR [3], ROUGE-L [9] and CIDEr [18].

4.2 Evaluation

To verify the effectiveness of relational knowledge, we evaluate our full model ($ICRK$) against DeepVS model as well as other state-of-the-art models on image captioning.

In training, we encodes the full-size input image with VGGNet [17] and set the size of hidden layer of RNN and the size of the input word embedding to 512, and we use Adam [7] algorithm to do model updating with an initial learning rate of $4 \times e^{-4}$.

Table 1 reports the performance of our ICRK which adds Skip-gram-RK to DeepVS relative to DeepVS baseline on the MSCOCO Karpathy test split. We also illustrate some qualitative captioning results of our model and the baseline in Fig. 2.

Table 1. Performance of our method on MSCOCO dataset

	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDEr
LRCN [4]	62.8	44.2	30.4	21.0	-	-	-
DeepVS [6]	62.5	45.0	32.1	23.0	19.5	-	66.0
Our baseline: DeepVS	64.3	45.3	31.7	22.9	20.5	46.7	69.7
Our model: ICRK	**65.9**	**47.2**	**33.1**	**23.7**	**21.1**	**47.9**	**73.0**

DeepVS: a kitchen with a refrigerator and a stove
Our ICRK: a refrigerator filled with lots of food and drinks
Human: a refrigerator filled with lots of soft drinks

DeepVS: a man standing next to a fire hydrant
Our ICRK: a parking meter sitting on the side of a road
Human: series of parking meters and cars are located next to each other

DeepVS: two zebras are standing in a field of grass
Our ICRK: two zebras standing next to each other in a zoo
Human: zebras standing behind the fence in a zoo

DeepVS: a tennis player in action on the court
Our ICRK: a tennis player is getting ready to serve the ball
Human: a woman in a skirt gets ready to hit a tennis ball

DeepVS: a plate of food with a sandwich and salad
Our ICRK: a white plate topped with meat and vegetables
Human: a white plate with a variety of meat and vegetables

DeepVS: an elephant is standing in the middle of a field
Our ICRK: a group of elephants standing next to each other
Human: a group of elephants walking in muddy water.

Fig. 2. Qualitative captioning results of our method and DeepVS baseline. The descriptions generated by our model are more accurate than the descriptions generated by DeepVS, and our model combined with relational knowledge can recognize more reliable objects in image and make a better description.

5 Conclusion

In this paper, we present a novel image captioning model which combines relational knowledge with captioning model. Qualitative evaluation suggest that using relational knowledge as regularization function to learning word representations effectively improves the performance of image captioning model. Compared this method with two captioning baseline models and other works, our method achieves state-of-the-art performance.

Acknowledgments. This work was supported by National Key Research and Development Program of China (Grant No. 2016YFB1000902), National Program on Key Basic Research Project (973 Program, Grant No. 2013CB329600), and National Natural Science Foundation of China (Grant No. 61472040).

References

1. Bollacker, K., Evans, C., Paritosh, P., Sturge, T., Taylor, J.: Freebase: a collaboratively created graph database for structuring human knowledge. In: Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data, SIGMOD 2008, pp. 1247–1250 (2008)
2. Bordes, A., Usunier, N., Garcia-Duran, A., Weston, J., Yakhnenko, O.: Translating embeddings for modeling multi-relational data. In: Advances in Neural Information Processing Systems, pp. 2787–2795 (2013)
3. Denkowski, M., Lavie, A.: Meteor universal: language specific translation evaluation for any target language. In: Proceedings of the Ninth Workshop on Statistical Machine Translation, pp. 376–380 (2014)
4. Donahue, J., et al.: Long-term recurrent convolutional networks for visual recognition and description. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2625–2634 (2015)
5. Fang, H., et al.: From captions to visual concepts and back. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2015
6. Karpathy, A., Fei-Fei, L.: Deep visual-semantic alignments for generating image descriptions. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2015
7. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
8. Lebret, R., Pinheiro, P.O., Collobert, R.: Simple image description generator via a linear phrase-based approach. In: ICLR (2015)
9. Lin, C.Y.: Rouge: a package for automatic evaluation of summaries. In: Text Summarization Branches Out (2004)
10. Lin, T.-Y., et al.: Microsoft COCO: common objects in context. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014, Part V. LNCS, vol. 8693, pp. 740–755. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10602-1_48
11. Lu, J., Xiong, C., Parikh, D., Socher, R.: Knowing when to look: adaptive attention via a visual sentinel for image captioning. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017)
12. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. In: ICLR (2013)

13. Mikolov, T., Sutskever, I., Chen, K., Corrado, G., Dean, J.: Distributed representations of words and phrases and their compositionality. Adv. Neural Inf. Process. Syst. **26**, 3111–3119 (2013)
14. Miller, G.A.: Wordnet: a lexical database for the english language. Commun. ACM **38**(11), 39–41 (2002)
15. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: BLEU: a method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, pp. 311–318. Association for Computational Linguistics (2002)
16. Rennie, S.J., Marcheret, E., Mroueh, Y., Ross, J., Goel, V.: Self-critical sequence training for image captioning. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017)
17. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
18. Vedantam, R., Lawrence Zitnick, C., Parikh, D.: CIDER: consensus-based image description evaluation. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4566–4575 (2015)
19. Xu, C., et al.: RC-NET: a general framework for incorporating knowledge into word representations. In: Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management, pp. 1219–1228. ACM (2014)
20. Xu, K., et al.: Show, attend and tell: Neural image caption generation with visual attention. ICML (2015)