

Topic-specific Image Caption Generation

Chang Zhou, Yuzhao Mao, and Xiaojie Wang

School of Computer,
Beijing University of Posts and Telecommunications,
Beijing, China
{elani, maoyuzhao, xjwang}@bupt.edu.cn

Abstract. Recently, image caption which aims to generate a textual description for an image automatically has attracted researchers from various fields. Encouraging performance has been achieved by applying deep neural networks. Most of these works aim at generating a single caption which may be incomprehensive, especially for complex images. This paper proposes a topic-specific multi-caption generator, which infer topics from image first and then generate a variety of topic-specific captions, each of which depicts the image from a particular topic. We perform experiments on flickr8k, flickr30k and MSCOCO. The results show that the proposed model performs better than single-caption generator when generating topic-specific captions. The proposed model effectively generates diversity of captions under reasonable topics and they differ from each other in topic level.

Keywords: Image caption · Topic model · Encoder-decoder

1 Introduction

Image caption is a cross-modal task which links the visual and the natural language modality. It aims at generating textual descriptions for an image automatically, and have received attentions worldwide. Inspired by successful advances in neural machine translation [1,2], most image caption generators are based on the encoder-decoder framework and trained in an end-to-end fashion. In machine translation, an LSTM is used to encode the sentence in source language, and another LSTM is employed to decode the intermedia into target language. By replacing the encoder with a CNN, encouraging performance has been achieved in image caption [6,7,8,9,10,11].

Most image caption generators generate a single caption for an image. However, an image is rich of information and an individual caption may be insufficient to depict it, especially for complex images. On the other hand, when facing an image, human may focus on different aspects and describe it from various of angles, resulting in multiple captions for the same image. There are also demands for captioning image under particular topics in real life. Sometimes one is interested just in a particular aspect of the image. Textual information related to the target topic should be extracted and information beyond the topic should be omitted in this circumstance.

To simulate the multi-caption results of human beings, as well as generate captions specific to particular topics, we propose a topic-specific multi-caption generator for image. It takes latent topic attributes of captions into account. We employ an unsupervised topic model to infer latent topics from captions first. Each topic is represented as an embedding and then integrated into the decoder, guiding the generating of topic-specific captions. We also propose a method of inferring topics from images to limit the range of topics to perform captioning. The results show that the topic-specific image caption generator effectively generates diverse captions under reasonable topics and each of them depicts the image in a particular aspect.

The remainder of the paper is organized as follows. Section 2 introduces some previous works in image caption generation. Detailed formulation and model structure are given in Section 3. Experimental settings, evaluation metrics and experimental results are shown in Section 4. Conclusion and discussion of future works are included in Section 5.

2 Related Work

As for image caption generating, there are various studies, from early pipeline methods [14,17] to the end-to-end models commonly used now [3,4,5,9,10,11]. Inspired by the success of the encoder-decoder framework in machine translation, Karpathy et al. [4] employed CNN as encoder to extract features from image and RNN is used as decoder to generate sentences. For the advantage of LSTM dealing with long distance dependence, Vinyals et al.[7] replaced the RNN generator with LSTM and made continuous improvement.

In order to pay additional attention to visual or semantic information, the attention mechanism is employed and improve the performance of image caption to a great extent. Xu et al. [9] use visual attention method to pay attention to particular parts of the image with different ratios at each time step. You et al. [10] and Zhou et al.[11] propose different semantic attention approaches separately. You et al. extract several key words as semantic attributes for each image and then integrate the semantic information into input and output. While Zhou et al. use image feature filtered by text feature (text-conditional image feature) as semantic guidance for the gLSTM decoder. Both of the two methods take current generated words into account and use them to impact input and output states, or filter image features.

Recently, a number of studies resulting in multi-captions generating arise. Johnson et al. [12] propose a dense image caption model to generate an individual caption for multiple objects separately. They add a dense localization layer between the CNN encoder and the RNN decoder to handle the localization and description task jointly. Captions in various granularities, namely words, phrases and sentences, are generated as bounding-box moving throw the image. Mao et al. [13] aim at generating unambiguous captions for similar objects in an image. They link descriptions to corresponding bounding-boxes and train the model by minimizing the max-margin loss between positive samples and negative samples.

Although the two models above can generate multi-captions for an image, they are object-driven essentially while ours are topic-driven. We aim at generating topic-specific multiple captions. Compared with generating results of previous models, our generation is more tendentious and diverse. The generated captions depict the image from various points of view which is more human-like.

3 Model

In this paper, we propose a model which can infer latent topics from image and generate multiple topic-specific captions based on the encoder-decoder framework. Remarkable results can be achieved by maximizing the probability of the captions conditioned on the given image together with the topic and minimizing the divergence between the predicted topic distribution and the real one for the image simultaneously.

We first formulize the topic-specific image generation task in Section 3.1. A overview of the architecture of the proposed model is given in Section 3.2. Detailed introduction of each sub module including the unsupervised topic model training, image-topic distribution predicting and topic-specific caption generating is presented in Section 3.3, 3.4 and 3.5 respectively. Finally, we address the loss function and the details of training in Section 3.6.

3.1 Problem Formulization

For single-caption generators, the target of the image caption is to maximize the probability of the description given an image. Suppose an image is presented as I and a caption is presented as S , the target is to maximize:

$$\log P(S|I) = \sum_{t=1}^{N_S} \log P(w_t|w_0, \dots, w_{t-1}, I). \quad (1)$$

where $S = \{w_0, \dots, w_{N_S}\}$ is the caption of image I with N_S words.

Different from the traditional single-caption generator, we aim at inferring latent topics from an image first and then generating multiple topic-specific captions each of which depicts the image from a particular topic point of view. Based on the assumption that topic-caption pairs are independent, our target is to maximize the probability of all topic-caption pairs given an image as follows:

$$\log P(S, Z|I) = \sum_{k=1}^{N_K} \log P(s_k, z_k|I). \quad (2)$$

Where Z , S denote the topic set and the corresponding caption set for image I , z_k and s_k are the k -th topic and caption specific to it. N_K denotes the amount of the topic-caption pairs (s_k, z_k) which can be inferred from the image.

Note that each item in the summation can be decomposes as follows:

$$P(s_k, z_k|I) = P(s_k|z_k, I)P(z_k|I). \quad (3)$$

We decompose the target into two parts, the second part $P(z_k|I)$ can be seen as an image-topic classifier, indicating whether the topic z_i can be inferred from the image. The first part $P(s_k|z_k, I)$ can be seen as a sentence generator similar to the traditional image caption generator, while with a topic restriction in addition. So the target can be represented as follows,

$$\log P(S, Z|I) = \sum_{k=1}^{N_K} \log P(z_k|I) + \sum_{k=1}^{N_K} \log P(s_k|z_k, I). \quad (4)$$

As shown above, the target of finding topic-caption pairs for a given image is divided into two parts, predicting topic distribution from the given image first and then predicting descriptions conditioned on the image and a particular topic.

3.2 Model Architecture

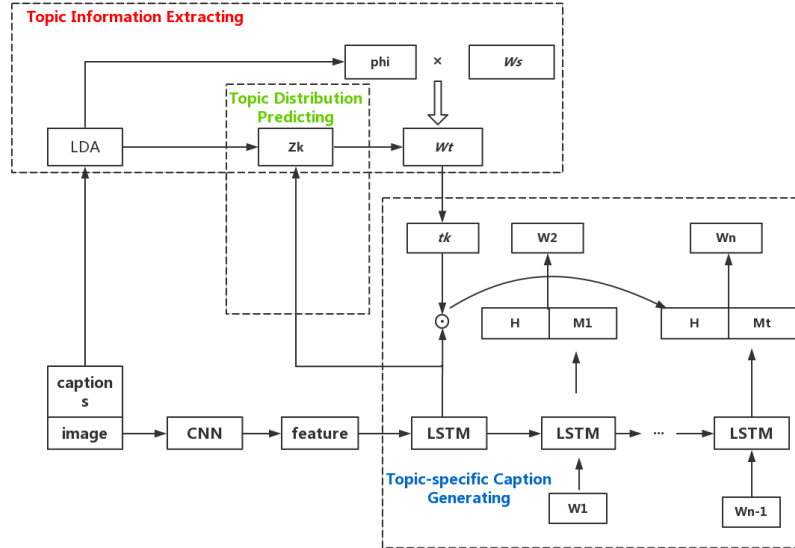


Fig. 1. The framework of the proposed topic-specific image caption generator. After applying LDA on textual captions, the topic distribution for each document (z_k) and the word distribution for each topic (ϕ) can be gained. Then train the predictor to infer topic distributions and the generator for image caption. W_s denotes the word embedding matrix, W_t denotes the topic embedding matrix. A particular topic embedding t_k supervises the caption procedure along with the image.

Following discussions above, we now depict the topic-specific image generating model. We employ the encoder-decoder framework, with CNN such as

VggNet to encode the image and LSTM to decode image representation into a textual sentence. The decoding side mainly consists of three modules: topic extracting, topic distribution predicting and topic-specific caption generating. The topic extracting module employs an unsupervised topic model to extract topic information from the captions, such as topic distribution features among different captions and word distribution features among different topics. The topic distribution prediction module is an image-topic classifier essentially. It approximates the topic distribution in the image to the inferred results in the topic extracting module. The topic-specific caption generating module takes the image and a particular topic as input and generates caption with topic restriction. The model architecture is shown in Fig. 1. When generating captions for new images, topic predictor infers topic sets first. Each of the inferred topics is utilized separately, leading generating of topic-specific captions.

3.3 Topic Information Extracting

Topics are latent features in captions. As there is no image caption dataset offering captions with ground-truth topic label at present, we have to obtain the latent topic information in an unsupervised manner. LDA [15] is an unsupervised topic model which has been widely used in tasks such as document classification. It introduces latent topics into document-word distribution and can infer topic distribution of unlabeled corpus. Representing each caption as a bag of words, topic distribution of each caption and word distribution of each topic can be learnt by applying LDA. Each caption with the inferred topic label composes a topic-caption pair which is used in the subsequent training procedure. It's worth to notice that we take the average of the inferred topic distributions of all captions for the same image as the target topic distribution in topic-distribution prediction training.

3.4 Topic Distribution Predicting

To generate multiple topic-specific captions for an image, it is necessary to infer topics which are occurred in the image first. It can be solved as a multi-label classification task, while we tackle it in a more elaborated way. An image may contain more than one topics simultaneously with different probabilities. We train a probability distribution predictor which approximates the topic distribution of the image to the inferred one in topic information extracting period. If the probability of a topic is higher than a threshold, we believe that the topic exists in the image and a caption should be generated.

It is worth to notice that we take the output of the first LSTM unit in decoder as input to train the topic distribution predictor, which is represented as I afterwards. Commonly, the output of the encoder can be seen as the image representation and then fed to the first LSTM unit to initialize the decoder. While the output of the first LSTM unit carries visual feature as well. It can be seen as a more abstract representation of the original image and utilizing it

results in better performance than the encoder output when training the topic distribution predictor.

As mentioned before, the target is to minimizing the divergence of the topic distributions between the predicted one and the inferred one by applying LDA. We use sigmoid-entropy as loss function, and the loss function is shown below,

$$z' = \text{sigmoid}(WI + b) . \quad (5)$$

$$L_{\text{topic_dist}} = - \sum_{k=1}^K (z_k \log z'_k + (1 - z_k)(1 - z'_k)) . \quad (6)$$

Where I denotes the abstract representation of the image mentioned above, z_k denotes the probability of topic k inferred by LDA, and z'_k denotes the one predicted by topic-distribution predictor. W, b are training parameters.

3.5 Topic-specific Caption Generator

Topic Embedding Construction The topic label is represented as embedding before feeding into the decoder. For each topic in the topic sets, we represent it as a weighted sum of all word embeddings as follows:

$$\text{topic}_k = \sum_{i=1}^K \phi_{k,i} w_i . \quad (7)$$

Where $\phi_{k,i}$ denotes the probability of the i th word in topic k , which can be learned by LDA, and w_i denotes the embedding form of the i th word which can be learnt from LSTM networks during training procedure.

Topic-specific Caption Generator As shown in Fig. 1, to generate a topic-specific caption for a given image, a simple but effective way is performing softmax function on a mixture of topic feature and the image feature when predicting words. To compute the probability of each candidate word in each time step with restriction of topic_k , we use following formulations,

$$H = W_I I \odot W_{\text{topic}} \text{topic}_k . \quad (8)$$

$$i_t = \sigma(W_{ix} x_t + W_{im} m_{t-1}) . \quad (9)$$

$$f_t = \sigma(W_{fx} x_t + W_{fm} m_{t-1}) . \quad (10)$$

$$o_t = \sigma(W_{ox} x_t + W_{om} m_{t-1}) . \quad (11)$$

$$c_t = f_t \odot c_{t-1} + i_t \odot h(W_{cx} x_t + W_{cm} m_{t-1}) . \quad (12)$$

$$m_t = o_t \odot c_t . \quad (13)$$

$$p_{t+1} = \text{softmax}(m_t, H) . \quad (14)$$

where I is the image representation mentioned above, H is the representation mixed with image and topic features, x_t is the input word embedding, m_t is the

output of the LSTM unit, i_t , f_t , o_t are input, forget, output gates and c_t is the memory in LSTM unit. W matrices are trained parameters.

The basic idea behind the formulation is simply but reasonable. We take the first LSTM output as a more abstract representation of the image. Then construct a latent variable H which extracting intersection features of image and topic and concatenate it with the output of the LSTM unit at each time step. It results in a mixture which contains not only the predictions made by the language model, but also the intersection restriction of the image and the topic. Apply softmax function on the mixed output and choose the word with the maximum probability in vocabulary as the predicted one.

3.6 Loss Function

The deviation of the topic-specific caption generating model comes from two parts, namely the deviation of the language model (the LSTM generator), and the deviation of the topic distribution predictor. The loss function is shown below:

$$\begin{aligned} \log L(S, Z|I) = & - \sum_{k=1}^{N_K} (z_k \log z'_k + (1 - z_k)(1 - z'_k)) \\ & - \sum_{k=1}^{N_K} \sum_{t=1}^{N_{S_k}} \log P(w_{k,t} | w_{k,0}, \dots, w_{k,t-1}, I, z_k). \end{aligned} \quad (15)$$

Where N_{S_k} denotes the length of the caption s_k specific to topic z_k and $w_{k,t}$ denotes the t -th word in caption s_k . Our model is trained by minimizing the loss function above with stochastic gradient descent. Batch size is set to 100 and learning rate is set to 0.0005. We also set dropout ratio to 0.5 for both the input embedding layer and output layer of the LSTM to combat overfitting.

4 Experiments

4.1 Datasets and Experimental Settings

Datasets We test our model on 3 public datasets, which is flickr8k [16], flickr30k[19] and MS-COCO[20] with total images of 8000, 31000 and 123000 respectively. Five references are provided by human annotators for each image in both Flickr8k and Flickr30k dataset. As for MS-COCO dataset, some images have more than 5 references which for dataset consistency we retain only 5 of them. We use public splits [4] to perform train, validation and test.

Topic Model Training In order to extract topic features from captions, LDA is applied in training set of the three dataset separately. We implement LDA with publicly available code, plda [21]. The parameters are set as $\alpha = 0.1$, $\beta = 0.01$. We train LDA with various topic numbers from 20 to 100 with varying step.

Image Features For image representations, we adopt the pre-trained VggNet as encoder and take the penultimate layer with 4096 dimensions as encoder output. We don't fine-tune the encoder CNN during the experiment to be consistent with the compared models.

4.2 Evaluation Strategy

We carry out the evaluation in two steps: evaluating performance of the topic-distribution predictor first and then the topic-specific caption generator.

Image-topic Classification Evaluation For each caption, we take the topic with the highest probability inferred by LDA as its topic label. For each image, the set of topic labels of all its captions can be regard as the latent topics for the image. When predicting topics from image, we filter topics with low probability and get predicted topics. We adopt two metrics, namely $F1$ score and one-recall to evaluate topic predicting performance. $F1$ score is calculated as below:

$$F = \frac{2PR}{P + R}. \quad (16)$$

where P denotes precision and R denotes recall of the classifier. One-recall measures the percentage of topic classification results which shoot at least one of the topics in the inferred ones by LDA. The higher the one-recall score is, the better the predictor performs.

Caption Generating Evaluation We adopt four metrics to evaluate the generated captions, namely BLEU@N [22], METEOR [23], ROUGE-L [24], and CIDEr-D [25]. To check whether or not the multi-caption generator is able to generate captions inclining to a particular topic, captions and reference are compared grouping by topic in evaluation. Topics of references can be inferred by applying the pre-trained LDA model. Topic-specific captions can be generated guiding by the same LDA inferred topics.

It's worth to notice that our model can generate multiple captions each of which is specific to a particular topic. However, previous models only generate one single caption in the matter of "topic". In order to show the difference in topic point of view among captions, we are forced to fake the traditional image caption generator as a multiple one. We copy the best generation of the single-caption generator to all topics, based on the assumption that generations in all topics are always the same.

4.3 Results

Classification Results and Analysis We evaluate the classification performance under different topic number settings. The result is shown in Fig. 2.

From the results we can see, the $F1$ score is unsatisfactory among all datasets. While the one-recall score is pretty high. It indicates that the predictor makes at least one correct prediction for most images though failing to infer all topics precisely. As the latent topics are obtained by applying LDA model, the

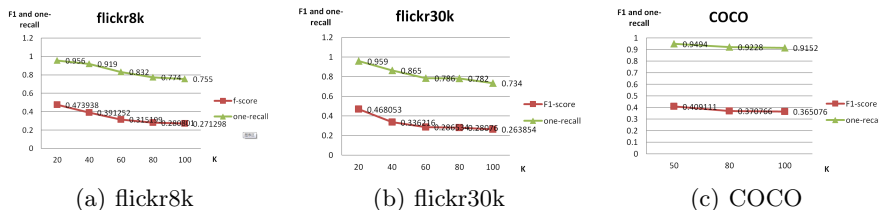


Fig. 2. image-topic classification performance with varying numbers of topic for three datasets. Classification performance declines with topic number increasing.

performance of the topic model affects the classification performance to a certain extent. Note that most of the captions are short sentences with length no more than 10 words. Each caption is handled as a short document and topics underlying in short documents may not be well learnt by topic models like LDA.

Image Caption Results and Analysis We compare our model with another two models without CNN fine-tuning: the Google NIC model [7] as baseline model and the glstm model [3]. Comparison results are shown in Table 1.

Table 1. topic-specific caption generating results for flickr8k, flickr30k and COCO dataset. The proposed topic-specific model(TS) outperforms the baseline model and the glstm model in topic-specific image caption generating. Best performance is obtained when the topic number k is set to 100.

Datasets	models	Bleu-1	Bleu-2	Bleu-3	Bleu-4	METEOR	ROUGE_L	CIDEr
flickr8k	baseline	0.348	0.204	0.125	0.078	0.128	0.311	0.403
	glstm	0.371	0.227	0.145	0.092	0.141	0.335	0.506
	TS	0.379	0.241	0.156	0.101	0.154	0.366	0.662
flickr30k	baseline	0.337	0.192	0.112	0.069	0.114	0.287	0.241
	glstm	0.34	0.204	0.128	0.082	0.122	0.311	0.385
	TS	0.348	0.22	0.142	0.091	0.136	0.336	0.517
COCO	baseline	0.426	0.265	0.168	0.109	0.145	0.356	0.54
	TS	0.476	0.321	0.218	0.148	0.181	0.403	0.876

As the results shown, our model surpasses both of the compared models in all metrics. With restriction of a particular topic, generations of the proposed model are more close to the human ones with the same topic. It indicates that the topic embedding captures latent topic features indeed. It guides the decoder to generate captions inclining to the target topic, depicting things related to the target topic and omitting the irrelevant ones.

4.4 Topic Analysis

Go deeper into the generated results, we confirm the topicality within topic-specific captions first, and then compare captions under different topics as well as captions under the same topic for different images.

Table 2. Topic consistency in flickr8k, flickr30k and COCO. The inferred topic is same as the supervised one most of the times.

Datasets	accuracy
flickr8k	0.799068767908
flickr30k	0.845175766642
coco	0.78639553716

Topic Consistency We take experiments to verify topic consistency between topic inferred by LDA and the supervised one used during generating. The topic label of each supervised topic in generating stage is recorded. After acquiring all topic-specific generations, we apply the pre-trained LDA model to them and infer topic distributions. Comparing the inferred topic with the supervised one and a good consistence occurs, as shown in Table 2.

The result shows that the inferred topic label is the same with the supervised one for nearly 80% of the generated descriptions, which indicates a good consistency. We can also confirm the rationality and effectiveness of the topic embedding.

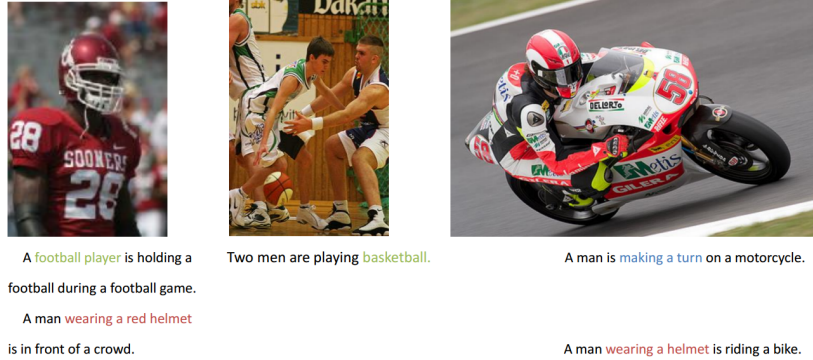


Fig. 3. Similarity of captions among different images with the same topic and diversity of captions for the same image. Topics are represented as different colors.

Generating Diversity For image with rich topics, multiple captions can be obtained. Some examples are shown in Fig. 3. The topic-specific captions of the first image depict it from "sports" and "clothing" topics separately, and captions related with "action" and "clothing" are generated for the third one.

We also observe generation similarities among different images which are specific to the same topic. Both of the captions colored in green in Fig. 3 are generated under the same topic, which is obviously a topic related with sports.

5 Conclusion

In this paper, we proposed a topic-specific caption generator which can generate multiple captions each of which depict the image from a particular topic. Topic information is encoded in form of embedding and utilized to supervise generating topic-specific captions. Compared with results of single-caption generator, the generating results are more diverse and topic-specific.

As the experimental results show, the topic distribution predictor has great potential for improvement. In the future, we plan to train a multi-modal topic model which can capture topic features from image and texts jointly.

Acknowledgments. This paper is supported by 111 Project(No.B08004), NSFC(No.61273365), Beijing Advanced Innovation Center for Imaging Technology, Engineering Research Center of Information Networks of MOE, and ZTE.

References

1. Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., Bengio, Y.: Learning phrase representations using RNN encoder-decoder for statistical machine translation. arXiv preprint arXiv:1406.1078 (2014)
2. Sutskever, I., Vinyals, O., Le, Q. V.: Sequence to sequence learning with neural networks. In Advances in neural information processing systems ,pp. 3104-3112 (2014)
3. Jia, X., Gavves, E., Fernando, B., Tuytelaars, T.: Guiding the long-short term memory model for image caption generation. In Proceedings of the IEEE International Conference on Computer Vision, pp. 2407-2415. (2015)
4. Karpathy, A., Fei-Fei, L.: Deep visual-semantic alignments for generating image descriptions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3128-3137 (2015)
5. Kiros, R., Salakhutdinov, R., Zemel, R.: Multimodal neural language models. In Proceedings of the 31st International Conference on Machine Learning (ICML-14), pp. 595-603 (2014)
6. Mao, J., Xu, W., Yang, Y., Wang, J., Huang, Z., Yuille, A.: Deep captioning with multimodal recurrent neural networks (m-rnn). arXiv preprint arXiv:1412.6632.(2014)
7. Vinyals, O., Toshev, A., Bengio, S., Erhan, D.: Show and tell: A neural image caption generator. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3156-3164 (2015)
8. Wu, Q., Shen, C., Liu, L., Dick, A., van den Hengel, A.: What value do explicit high level concepts have in vision to language problems?. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 203-212 (2016)
9. Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., ... , Bengio, Y.: Show, attend and tell: Neural image caption generation with visual attention. In International Conference on Machine Learning, pp. 2048-2057 (2015, June)

10. You, Q., Jin, H., Wang, Z., Fang, C., Luo, J.: Image captioning with semantic attention. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4651-4659 (2016)
11. Zhou, L., Xu, C., Koch, P., Corso, J. J.: Image Caption Generation with Text-Conditional Semantic Attention. arXiv preprint arXiv:1606.04621 (2016)
12. Johnson, J., Karpathy, A., Fei-Fei, L.: Densecap: Fully convolutional localization networks for dense captioning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4565-4574 (2016)
13. Mao, J., Huang, J., Toshev, A., Camburu, O., Yuille, A. L., Murphy, K.: Generation and comprehension of unambiguous object descriptions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 11-20 (2016)
14. Farhadi, A., Hejrati, M., Sadeghi, M., Young, P., Rashtchian, C., Hockenmaier, J., Forsyth, D.: Every picture tells a story: Generating sentences from images. Computer vision CECCV 2010, 15-29 (2010)
15. Blei, D. M., Ng, A. Y., Jordan, M. I.: Latent dirichlet allocation. Journal of machine Learning research, 3(Jan), 993-1022 (2003)
16. Hodosh, M., Young, P., Hockenmaier, J.: Framing image description as a ranking task: Data, models and evaluation metrics. Journal of Artificial Intelligence Research, 47, 853-899 (2013)
17. Kulkarni, G., Premraj, V., Ordonez, V., Dhar, S., Li, S., Choi, Y., ..., Berg, T. L.: Babytalk: Understanding and generating simple image descriptions. IEEE Transactions on Pattern Analysis and Machine Intelligence, 35(12), 2891-2903 (2013)
18. Yang, Y., Teo, C. L., Daum III, H., Aloimonos, Y.: Corpus-guided sentence generation of natural images. In Proceedings of the Conference on Empirical Methods in Natural Language Processing, pp. 444-454. Association for Computational Linguistics (2011, July)
19. Young, P., Lai, A., Hodosh, M., Hockenmaier, J.: 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)
20. Lin, T. Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., ..., Zitnick, C. L.: Microsoft coco: Common objects in context. In European Conference on Computer Vision, pp. 740-755. Springer International Publishing (2014, September)
21. Wang, Y., Bai, H., Stanton, M., Chen, W. Y., Chang, E. Y.: Plda: Parallel latent dirichlet allocation for large-scale applications. In International Conference on Algorithmic Applications in Management, pp. 301-314. Springer Berlin Heidelberg (2009, June)
22. 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, July)
23. Banerjee, S., Lavie, A.: METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization, Vol. 29, pp. 65-72 (2005, June).
24. Lin, C. Y.: Rouge: A package for automatic evaluation of summaries. In Text summarization branches out: Proceedings of the ACL-04 workshop, Vol. 8 (2004, July)
25. Vedantam, R., Lawrence Zitnick, C., Parikh, D.: Cider: Consensus-based image description evaluation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4566-4575 (2015)