



Joining Cross-Modality Retrieval to Reweight Attributes for Image Caption Generation

Yuxuan Ding¹, Wei Wang¹, Mengmeng Jiang¹, Heng Liu¹, Donghu Deng², Wei Wei³, and Chunna Tian¹(✉)

¹ VIPS Lab, School of Electronic Engineering, Xidian University, Xi'an, China
chnatian@xidian.edu.cn

² Troops 95841 of PLA, Jiuquan, China

³ School of Computer Science, Northwestern Polytechnical University, Xi'an, China

Abstract. Automatic natural language description for images is one of the key issues towards image understanding. In this paper, we propose an image caption framework, which explores specific semantics jointing with general semantics. For specific semantics, we propose to retrieve captions of the given image in a visual-semantic embedding space. To explore the general semantics, we first extract the common attributes of the image by Multiple Instance Learning (MIL) detectors. Then, we use the specific semantics to re-rank the semantic attributes extracted by MIL, which are mapped into visual feature layer of CNN to extract the jointing visual feature. Finally, we feed the visual feature to LSTM and generate the caption of image under the guidance of BLEU₄ similarity, incorporating the sentence-making priors of reference captions. We evaluate our algorithm on standard metrics: BLEU, CIDEr, ROUGE_L and METEOR. Experimental results show our approach outperforms the state-of-the-art methods.

Keywords: Cross-modality retrieval · Image captioning · Semantic attribute

1 Introduction

Image captioning aims to automatically describe an image with natural language captions. It first grabs information of main objects, relationships among objects and the scene context as well in the images, and then describes the information with natural languages. Thus, it involves the techniques of both computer vision and natural language processing. However, how to well represent the visual information of images and describe them reasonably are still challenging. Thus, image captioning is still a hot research topic. A mass of methods are proposed to address these issues in recent years.

The first author Yuxuan Ding is a Ph.D. candidate.

Retrieval Based Caption: In retrieval base captioner, caption was retrieved from captions of similar images in the training set. As we can see, retrieval based methods need a large amount of annotated sentences for searching valid similar descriptions. Ordonez et al. [20] utilize global image representations to retrieve related captions from a large dataset and then transfer to the query image. Devlin et al. [7] find the visually similar k -Nearest Neighbor (k -NN) of the testing images in the training set, and then select best captions from the captions of k -NN images based on highest average lexical similarity. Kiros et al. [14] proposed an end-to-end method to train embedding mapping with triplet loss. Faghri et al. [9] uses rank loss to optimize the embedding, which has achieved the state-of-the-art performance in image-caption retrieval. However, these approaches cannot generate novel descriptions.

Encoder-Decoder Based Caption: Image captioning has big progress in recent years, because of deep learning based feature representation and sequential machine translation. Inspired by the end-to-end machine translation [2, 6], Encoder-Decoder captioner extract the visual feature of images with Convolutional Neural Network (CNN) and then use Recurrent Neural Networks (RNN) to translate visual representation into natural language descriptions [13, 19, 26]. Mao et al. [19] propose a multimodal Recurrent Neural Network (m-RNN) model for generating captions, it consists of a CNN for image representation, a RNN for text embedding. Vinyals et al. [26] adopt GoogLeNet as an image encoder and apply LSTM [11] as the decoder. Karpathy et al. [13] attempt to align sentence fragments to image regions, and then aim at generating descriptions of visual regions using RNN. In the end-to-end translation framework, some approaches introduce an attention mechanism to improve the performance of image captioning [18, 28]. Recently, Reinforcement Learning (RL) [24] has been applied to optimize the image captioning model by using the test metrics as rewards, such as BLUE in [1] or CIDEr in [22], which improve the results distinctively.

Visual Attribute: Feeding high-level semantic concepts to RNN usually results in better captioning results. Therefore, visual attributes are incorporated into image captioning in many ways [27, 29, 30], among which attribute extraction is one of the most successful method. Wu et al. [27] demonstrate that the high level visual concepts play an important role in image captioning. They feed the detected region-based attributes rather than CNN feature into the Decoder. Yao et al. [29] confirm that feeding image feature to LSTM at the first time step and feeding attributes at every step is the best choice. You et al. [30] use a Fully Convolutional Network (FCN) [23] selectively focus on visual semantic words while extracting image feature.

However, there are two issues need to be solved: (1) The detected attributes emphasize on the general attributes in the training set, which may not the most related to testing images; (2) Using attributes instead of visual features ignores the spatial layout and context of the attributes. Once the information is lost, it is hard to be recovered during decoding. To address these issues, we use cross-

modal retrieval to find related captions for image. Considering the embedding space should align feature of both objects and scene, we concatenate the scene feature and object feature to build the multi-feature visual-semantic embedding (MVSE++) based cross-modal retrieval. Retrieved captions are used to re-rank the detected attributes, which pick the specific semantics. Then, we map attributes into the CNN to extract their visual feature. This feature contains objects, layout and context of general and specific semantic attributes. We adopt the $Bleu_4$ similarity in the decoding to further use the specific semantics, improving the performance of sentence generation. The framework of our method is illustrated as Fig. 1. The experimental results on MS-COCO show our method achieves the best performance on almost all evaluation metrics compared with the state-of-the-art methods, especially, the BLEU_4 reaches 0.342 and CIDEr reaches 1.058.

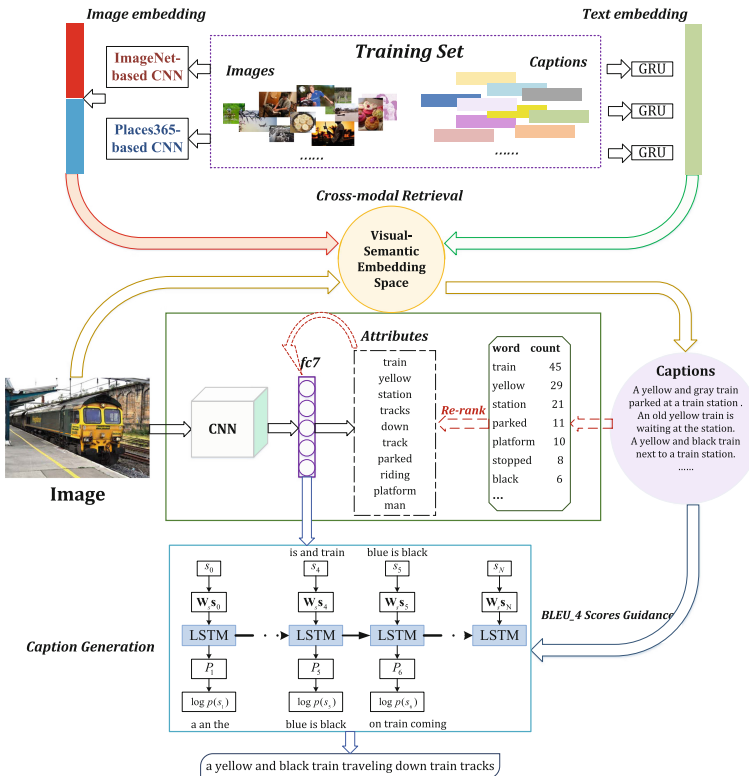


Fig. 1. The main framework of the proposed method.

2 Our Model

In this paper, we propose a retrieval based attribute model for captioning task. Our model consists of three parts: Attribute extraction by MIL detectors [10] to obtain the general semantics existing in the testing image for captioning, attributes re-ranking based on the specific semantics provided by MVSE++ retrieval and caption generation guided by BLEU.4 similarity. As the detectors were obtained from the whole dataset, we think they can represent general semantics of captions. We think specific semantics of an image can be provided by other images which correspond to it, so we called retrieval semantics as specific semantics.

2.1 Image-Caption Retrieval

We propose to retrieve the captions for the input image in a multi-feature visual-semantic embedding (MVSE++) space, which avoids the visual semantic misalignment on both aspects of objects and scene context. In the original VSE++ based cross-modal retrieval method, the visual feature is extracted by the CNN based object classifier trained on ImageNet dataset, which mainly focus on the objects in images and lacks the scene information. Thus, we joint the object-scene feature based visual-semantic embedding to retrieve the image specific captions. The feature of image and semantic features of captions, which are encoded by GRU as same as basic VSE++ method, which are mapped into the same space. Thus, we can get the candidates in another modality in this common space by finding near neighbors, which compose our specific semantics.

2.2 The General and Specific Semantics Jointed Visual Feature Extraction

First, we need figure out what are the generally happened semantics in image captions. Follow most attributes detection method, we analyze the distribution of word frequency in the training captions, and collect 1000 most frequently appearing words to build an attribute set $A = \{Att_1, Att_2, \dots, Att_N\}$, $N = 1000$, as the common semantics. This set covers 92% words in all captions and acts as the initial semantic categories for 1000 attribute detectors, which is trained by a CNN based MIL model.

MIL views each training image as a bag of labels. An image \mathbf{I} is a bag of semantic features. For one attribute Att_i , image is a positive sample if its caption contains Att_i , regions in the image build a positive bag, otherwise, it is a negative sample and we think it is a negative bag. The MIL detectors are alternatively optimized. Attributes usually describe complex and some of them cannot be demarcated boundaries clearly, such as “red”, “holding” or “beautiful” etc. So we follow the work of Lebet et al. [15], detecting attributes with a noisy-OR version of MIL. We resize an image sample to 567×567 and feed into CNN, which is based on a modified VGG.16 network. Five convolution layers in front are kept, in order to maintain regional information for visual words extracting, we replace

fc layers by three convolutions then obtain a fully convolutional network. So, the penultimate convolution layer fc_7conv represents image feature reserving location information of original input image. After above steps, it generates a 12×12 coarse response map corresponding to slide the original CNN over image with stride of 32 and get fc_8conv 's output on each location. For each image \mathbf{I} , $p_j^{Att_i}$ is the probability of sub-region j corresponding an attribute Att_i , then we calculate an integrated probability combine all regions probabilities in this image as follows

$$p^{Att_i} = 1 - \prod_{j \in \mathbf{I}} (1 - p_j^{Att_i}) \quad (1)$$

We train the network with a multi-label classification task. The class is top 1000 frequent words and labels are built from the ground truth captions. This is our MIL detector.

Given a testing image \mathbf{I} , we detect a set of attributes $A = \{Att_1, Att_2, \dots, Att_N\}$ by MIL detectors obtain the general semantics existing in the image for captioning. We obtain the specific semantics by retrieving top sentences in MVSE++ space. We count the frequency of attribute Att_i in A as c_{Att_i} , which is used to reweight the original attributes probabilities $\{p_{att}^{Att_1}, p_{att}^{Att_2}, \dots, p_{att}^{Att_N}\}$ as follows.

$$p_{re-att}^{Att_i} = p_{att}^{Att_i} + \alpha * c_{Att_i} \quad (2)$$

In Eq. (2), α represents the weighting coefficient, means the proportion of retrieved words in the overall attributes. According to Eq. (2), we re-rank the attributes in A and selected top T attributes $\{Att_1, Att_2, \dots, Att_T\}$, whose probabilities $\{p_{re-att}^{Att_1}, p_{re-att}^{Att_2}, \dots, p_{re-att}^{Att_T}\}$ is defined as ρ , to maintain the testing-specific attributes and reduce the influence of uncorrelated attributes. Finally, we map the re-ranked attributes to fc_7conv layer of CNN to obtain the general and specific semantics collaborated visual feature. Corresponding visual features of the re-ranked attributes are extracted as follows.

$$\rho' = \sum_{i=1}^T \rho_i \mathbf{fc}_8 \mathbf{w}_i \quad (3)$$

$$\mathbf{z}_{re-att} = \rho' \odot \text{GAP}(fc_7conv) \quad (4)$$

where $\mathbf{fc}_8 \mathbf{w}_i$ is weight from fc_7conv to fc_8conv for Att_i , \odot represents dot multiplication. GAP is global average pooling operation, which is a merging of region feature so it can provide more details in image. We consider the \mathbf{z}_{re-att} as an importance-weighted visual feature with the most caption-relevant information. This feature is used as the input of LSTM to improve the accuracy of image captioning. These re-ranked attributes contain a variety types of word such as noun, adjective, verb and so on.

2.3 Image Caption Generator

We use the LSTM model as caption generator. At t time step, LSTM is formulated as below.

$$\mathbf{i}_t = \sigma(\mathbf{W}_{ix}\mathbf{x}_t + \mathbf{W}_{ih}\mathbf{h}_{t-1}) \quad (5)$$

$$\mathbf{f}_t = \sigma(\mathbf{W}_{fx}\mathbf{x}_t + \mathbf{W}_{fh}\mathbf{h}_{t-1}) \quad (6)$$

$$\mathbf{o}_t = \sigma(\mathbf{W}_{ox}\mathbf{x}_t + \mathbf{W}_{oh}\mathbf{h}_{t-1}) \quad (7)$$

$$\mathbf{c}_t = \mathbf{f}_t \odot \mathbf{c}_{t-1} + \mathbf{i}_t \odot \tanh(\mathbf{W}_{cx}\mathbf{x}_t + \mathbf{W}_{ch}\mathbf{h}_{t-1}) \quad (8)$$

$$\mathbf{h}_t = \mathbf{o}_t \odot \tanh(\mathbf{c}_t) \quad (9)$$

where \mathbf{x}_t , \mathbf{h}_t and \mathbf{c}_t are input vector, hidden state and cell state of LSTM. \mathbf{W} represents the embedding matrix. σ is sigmoid function and \tanh is hyperbolic tangent. \odot represents dot multiplication of two vectors. \mathbf{x}_t , \mathbf{h}_{t-1} and \mathbf{c}_{t-1} are given at each time step. \mathbf{i}_t , \mathbf{f}_t , \mathbf{o}_t are input gate, forget gate and output gate respectively.

Instead of feeding the simple image feature directly, we input the attributes re-ranked image visual feature \mathbf{z}_{re-att} in Sect. 3.2 to LSTM as the ‘‘source language’’. Therefore, we establish a dependence relationship between words and sentences in the training dataset. The decoder maximizing the probability of the correct description is formulated by Eq. (10).

$$\theta^* = \arg \max_{\theta} \sum_{(I,S)} \log p(S|\mathbf{z}_{re-att}; \theta) \quad (10)$$

we define $S = \{S_1, S_2, \dots, S_L\}$ as a sequence of words. It usually uses chain rule to model the joint probability of previously generated words as

$$\log p(S|\mathbf{z}_{re-att}) = \sum_{t=0}^L \log p(S_t|\mathbf{z}_{re-att}, S_0, \dots, S_{t-1}), \quad (11)$$

where N is caption length and $\log p(S_t|\mathbf{z}_{re-att}, S_0, \dots, S_{t-1})$ means the probability of generating the current word S_t conditioned on attribute based vector \mathbf{z}_{re-att} and previously generated words.

During training, suppose we have image visual feature \mathbf{z}_{re-att} and its description sequence $\{S_0, S_1, S_2, \dots, S_N, S_{N+1}\}$, where S_0 is the start symbol and S_{N+1} is the end symbol. Each element in the sequence is a one-hot word vector, whose size is 11518. In the proposed approach, feature vector is mapped into H -dimensional space by the embedding matrix \mathbf{W}_f . LSTM was initialized as follows.

$$\mathbf{h}_0 = \text{LSTM}(\mathbf{W}_f\mathbf{z}_{re-att}, \mathbf{0}) \quad (12)$$

The decoding procedure is given in Eqs. (13)–(15). After initializing LSTM, one-hot word vector is embedded by \mathbf{W}_S as input vector. Hidden state \mathbf{h}_t at each step is computed by LSTM and mapped into 11518-dimensional word space by \mathbf{W}_h . The generator is formulated to minimize the loss, which is the negative log likelihood in Eq. (17).

$$\text{init}(\mathbf{h}) = \mathbf{h}_0 \quad (13)$$

$$\mathbf{h}_t = \text{LSTM}(\mathbf{W}_S S_t, \mathbf{h}_{t-1}), t \geq 0 \quad (14)$$

$$p(S_t | \mathbf{z}_{re-att}, S_0, S_1, \dots, S_{t-1}) = \text{Softmax}(\mathbf{W}_h \mathbf{h}_t) \quad (15)$$

Similar to Chen et al. [4], LSTM network infers image descriptions in the testing phase. We use retrieved sentences as references to guide the description generation by comparing the similarities between current generated sentence and top k retrieved captions. So during the generation of each sentence, it can correct the deviation of focus, making descriptions fit the evaluation metrics. Inspired by Devlin et al. [7], we introduce the consensus score concept that calculated by the descriptions of similar images from the training set. This consensus scoring function between image \mathbf{I} and generated sentence S as

$$r(S, \mathbf{I}) = \frac{1}{k} \sum_{\omega=1}^k \varphi(S, \omega), \quad (16)$$

where S is current generated sentence and \mathbf{I} is the given image, the k sentences $\{\omega_1, \omega_2, \dots, \omega_k\}$ are retrieved by image \mathbf{I} using cross-modal embedding, and $\varphi(S, \omega)$ is the similarity score between two captions: (S, ω) . We choose BLEU-4 similarity function which measures 4-gram overlap. At each inference time step of LSTM, the probability of generating is decided by log likelihood Eq. (11) and consensus score Eq. (16) together. We use λ to balance these two terms, the final predict probability as follow.

$$l(S, \mathbf{z}_{re-att}) = \lambda \log p(S | \mathbf{z}_{re-att}) + (1 - \lambda)r(S, \mathbf{I}) \quad (17)$$

3 Experiments and Results

3.1 Datasets and Experimental Setup

MSCOCO contains 82,783 training images, 40,504 validation images and 40,775 testing images [17]. As MSCOCO is the most common dataset of image captioning task and many related works only evaluate on it, we also explore evaluation result of our model. Each image has five captions annotated by Amazon Mechanical Turk (AMT). Since the original testing set of is not completely available, we follow standard testing way of previous methods. For comparison with other approaches fairly, we split the training set and validation set together into three parts: training, validation and testing as Vinyals et al. [26] did. This split reserves 10% unused 5000 images of MSCOCO validation randomly for testing.

Network Architectures: As for feature extracting, we modify VGG_16 by keeping the convolutional layers $conv_1$ to $conv_5$ and replacing the fully connected layers fc_7 fc_8 with three fully convolutional layers. Finally, a MIL layer is followed for visual attributes prediction. We select all attributes with probabilities higher than 0.3 as candidate terms. For MVSE++, we implemented ResNet_152 CNN trained on ImageNet and Places365 datasets to obtain two

2048D feature vectors, which are concatenated as one 4096D visual feature. The top 20 retrieved sentences are used as the specific semantic prior to re-rank candidate attributes to weight layer and output a 4096D feature vector. We feed this feature to LSTM with a 512D state vector from Google NIC network for captioning. All these models are trained on NVIDIA Titan Xp.

Evaluation Metrics: The methods are evaluated on the standard metrics: BLEU_n [21], CIDEr [25], ROUGE_L [16] and METEOR [3] following coco-caption [5]. BLEU measures the similarity between two sentences in machine translation task, which is defined as the geometric mean of n-gram (up to 4) precision scores multiplied by a brevity penalty on short sentences. CIDEr measures the consensus between generated descriptions and the reference sentences, which is a specific evaluation metric designed for image captioning recently. METEOR is defined as the harmonic mean of precision and recall of unigram matches between sentences. For all the metrics, the higher is the better.

Baseline: In order to completely verify the effectiveness of our method, we use an original VGG₁₆ to extract image features as baseline, however, without attribute involves. We just make a common VGG Network to extract f_{c7} feature as one 4096D vector which is fed into the LSTM directly.

3.2 Results and Analysis

For an intuitive presentation of our joint language retrieval attribute-conditional approach, we design the following experiment. Table 1 shows how the language retrieval results improve captioning accuracy. **ATTR** means an attribute based feature only mapping visual concepts on f_{c7} fed into LSTM, its Bleu₄ score just reaches 0.256. **Re-ATTR** means the model combined attribute with the retrieval results, as we can see, Bleu₄ score rapidly increases to 0.32 while CIDEr increases from 0.765 to 1.001, nearly one-third. Other metrics have an excellent performance, too. **B4-Re-ATTR** model consists of visual attribute, MVSE++ retrieval based attribute distribution re-rank, and BLEU₄ similarity guidance caption generation, that achieves the best performance on all metrics obviously.

Table 1. Results comparison on variety of parameters, testing on MSCOCO dataset of 5000 images

Model	Bleu ₁	Bleu ₂	Bleu ₃	Bleu ₄	CIDEr	ROUGE _L	METEOR
Baseline	0.658	0.478	0.347	0.255	0.776	0.491	0.223
ATTR	0.663	0.481	0.347	0.256	0.765	0.491	0.226
Re-ATTR	0.730	0.561	0.423	0.320	1.001	0.540	0.260
B4-Re-ATTR	0.749	0.586	0.446	0.337	1.051	0.548	0.260

We report performance of our method and other state-of-the-art methods on MSCOCO in Table 2. The state-of-the-art algorithms are three main types: (1) The simple encode-decode based model Google NIC [26], LRCN [8] and m-RNN [19]. (2) Attention based methods such as Guiding LSTM [12] and Soft/Hard Attention [28], and 3. High level attributes based model ATT_FCNC [30], ATT_CNN_LSTM [27], (3) LSTM-A [29]. Experiment demonstrates that our joint retrieval attribute-conditional approach achieves almost the excellent performance on metrics, BLEU_4 is a more convincing evaluation metrics that measures the matching degree between phrases. Our model has an outstanding performance in BLUE_4, it reaches to 0.342, outperforms all the compared state-of-the-art approaches. As for the specialized evaluation metric CIDEr, our 1.058 better than all the comparison methods, too. Soft/Hard Attention model performances better than other models because of the “attention” mechanism. However, our attribute model still has best results under most metrics. As the same type, our approach performs better than ATT_FCNC, ATT_CNN_LSTM and LSTM-A, it is not difficult to judge that our cross-modal retrieval method provides effective scene context and spatial layout similarity of attributes for image caption task. And BLUE_4 similarity is a key supplement in generation stage.

Table 2. Performance of our proposed method and other state-of-the-art methods on MSCOCO

Model	Bleu_1	Bleu_2	Bleu_3	Bleu_4	CIDEr	ROUGE_L	METEOR
NIC [26]	0.666	0.451	0.304	0.203	0.855	0.491	0.237
LRCN [8]	0.628	0.442	0.304	0.210	-	-	-
m-RNN [19]	0.670	0.490	0.350	0.250	-	-	-
Soft Attention [28]	0.707	0.492	0.344	0.243	-	-	0.239
Hard Attention [28]	0.718	0.504	0.357	0.250	-	-	0.230
Guiding LSTM [12]	0.670	0.490	0.360	0.260	-	-	0.230
ATT_FCNC [30]	0.709	0.537	0.402	0.304	-	-	0.243
ATT_CNN_LSTM [30]	0.740	0.560	0.420	0.310	0.940	-	0.260
LSTM-A [29]	0.730	0.565	0.429	0.325	0.986	0.538	0.261
Our model	0.730	0.561	0.423	0.320	1.001	0.540	0.260
Our B4 Model	0.749	0.586	0.446	0.337	1.051	0.548	0.260

Qualitative Analysis: In addition to the above exact results, we also draw a qualitative analysis chart as Fig. 2 to show the superiority of our method. It compares our model with the initial attributes model **ATTR**, two decoding networks use the same LSTM structure so the gap of results only depends on the differences of image features. We show some captioning examples from the validation set. As we can see, the visual words often corresponds to salient objects



Fig. 2. Qualitative analysis of attributes with caption retrieval result. The top line shows simple visual attributes feature captions. The bottom line shows retrieval reweighted descriptions.

or relationships of images. Since the retrieval based attributes provide both main objects and surroundings, the captions of our final network have more fine details, such as the type and number of objects, the color information and the spatial relationship between goals. All the above results illustrate that our attribute-conditional model guided by caption retrieval leads to an overall increase in caption generation performance.

4 Conclusions

In this paper, we propose a novel caption generation approach based on reweighted semantic attributes. We use cross-modality retrieval results to re-rank key visual attributes in image and obtain an attribute-conditional feature, on the other hand, retrieval results also provide BLEU_4 similarity information to guide caption generating for testing image. For attribute extraction, a MIL based VGG_16 network detects preliminary key attributes from sets of image regions as candidates, these attributes always pay more attention on the regions with richer semantic information in given image. For cross-modality retrieval, a MVSE++ model searches similar captions in joint visual-semantic embedding space. Then, we reweight the candidate attributes distribution according to the retrieved similar image captions from the training set, moreover, the retrieved captions also participate in sentence generating on the LSTM decoding stage. Experiments verify the accuracy of our method. It outperforms several state-of-the-art methods on MSCOCO 2014 dataset.

Acknowledgement. This work was supported in part by the National Natural Science Foundation of China under Grants 61571354 and 61671385. In part by China Post doctoral Science Foundation under Grant 158201.

References

1. Bahdanau, D., et al.: An actor-critic algorithm for sequence prediction. In: 5th International Conference on Learning Representations, ICLR 2017, Toulon, France, 24–26 April 2017, Conference Track Proceedings (2017)
2. Bahdanau, D., Cho, K., Bengio, Y.: Neural machine translation by jointly learning to align and translate. In: 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, 7–9 May 2015, Conference Track Proceedings (2015)
3. Banerjee, S., Lavie, A.: METEOR: an automatic metric for MT evaluation with improved correlation with human judgments. In: Proceedings of the Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization@ACL 2005, Ann Arbor, Michigan, USA, 29 June 2005, pp. 65–72 (2005)
4. Chen, M., Ding, G., Zhao, S., Chen, H., Liu, Q., Han, J.: Reference based LSTM for image captioning. In: Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, San Francisco, California, USA, 4–9 February 2017, pp. 3981–3987 (2017)
5. Chen, X., et al.: Microsoft COCO captions: data collection and evaluation server. CoRR abs/1504.00325 (2015). <http://arxiv.org/abs/1504.00325>
6. Cho, K., et al.: Learning phrase representations using RNN encoder-decoder for statistical machine translation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, Doha, Qatar, 25–29 October 2014. A meeting of SIGDAT, a Special Interest Group of the ACL, pp. 1724–1734 (2014)
7. Devlin, J., Gupta, S., Girshick, R.B., Mitchell, M., Zitnick, C.L.: Exploring nearest neighbor approaches for image captioning. CoRR abs/1505.04467 (2015). <http://arxiv.org/abs/1505.04467>
8. Donahue, J., et al.: Long-term recurrent convolutional networks for visual recognition and description. *IEEE Trans. Pattern Anal. Mach. Intell.* **39**(4), 677–691 (2017)
9. Faghri, F., Fleet, D.J., Kiros, J., Fidler, S.: VSE++: improving visual-semantic embeddings with hard negatives. In: British Machine Vision Conference 2018, BMVC 2018, 3–6 September 2018, p. 12. Northumbria University, Newcastle, UK (2018)
10. Fang, H., et al.: From captions to visual concepts and back. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, 7–12 June 2015, pp. 1473–1482 (2015)
11. Hochreiter, S., Schmidhuber, J.: Long short-term memory. *Neural Comput.* **9**(8), 1735–1780 (1997)
12. Jia, X., Gavves, E., Fernando, B., Tuytelaars, T.: Guiding the long-short term memory model for image caption generation. In: 2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, 7–13 December 2015, pp. 2407–2415 (2015)
13. Karpathy, A., Fei-Fei, L.: Deep visual-semantic alignments for generating image descriptions. *IEEE Trans. Pattern Anal. Mach. Intell.* **39**(4), 664–676 (2017)

14. Kiros, R., Salakhutdinov, R., Zemel, R.S.: Unifying visual-semantic embeddings with multimodal neural language models. CoRR abs/1411.2539 (2014). <http://arxiv.org/abs/1411.2539>
15. Lebet, R., Pinheiro, P.H.O., Collobert, R.: Simple image description generator via a linear phrase-based approach. In: 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, 7–9 May 2015, Workshop Track Proceedings (2015)
16. Lin, C.Y.: ROUGE: a package for automatic evaluation of summaries. Text Summarization Branches Out (2004)
17. Lin, T.-Y., et al.: Microsoft COCO: common objects in context. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8693, pp. 740–755. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10602-1_48
18. Lu, J., Xiong, C., Parikh, D., Socher, R.: Knowing when to look: adaptive attention via a visual sentinel for image captioning. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, 21–26 July 2017, pp. 3242–3250 (2017)
19. Mao, J., Xu, W., Yang, Y., Wang, J., Yuille, A.L.: Deep captioning with multimodal recurrent neural networks (m-RNN). In: 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, 7–9 May 2015, Conference Track Proceedings (2015)
20. Ordonez, V., Kulkarni, G., Berg, T.L.: Im2text: describing images using 1 million captioned photographs. In: Advances in Neural Information Processing Systems 24: 25th Annual Conference on Neural Information Processing Systems 2011. Proceedings of a Meeting Held 12–14 December 2011, Granada, Spain, pp. 1143–1151 (2011)
21. Papineni, K., Roukos, S., Ward, T., Zhu, W.: BLEU: a method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, 6–12 July 2002, Philadelphia, PA, USA, pp. 311–318 (2002)
22. Rennie, S.J., Marcheret, E., Mroueh, Y., Ross, J., Goel, V.: Self-critical sequence training for image captioning. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, 21–26 July 2017, pp. 1179–1195 (2017)
23. Shelhamer, E., Long, J., Darrell, T.: Fully convolutional networks for semantic segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **39**(4), 640–651 (2017)
24. Sutton, R.S., Barto, A.G.: Reinforcement Learning - An Introduction. Adaptive Computation and Machine Learning. MIT Press, Cambridge (1998)
25. Vedantam, R., Zitnick, C.L., Parikh, D.: Cider: consensus-based image description evaluation. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, 7–12 June 2015, pp. 4566–4575 (2015)
26. Vinyals, O., Toshev, A., Bengio, S., Erhan, D.: Show and tell: a neural image caption generator. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, 7–12 June 2015, pp. 3156–3164 (2015)
27. Wu, Q., Shen, C., Liu, L., Dick, A.R., van den Hengel, A.: What value do explicit high level concepts have in vision to language problems? In: 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, 27–30 June 2016, pp. 203–212 (2016)
28. Xu, K., et al.: Show, attend and tell: Neural image caption generation with visual attention. In: Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6–11 July 2015, pp. 2048–2057 (2015)

29. Yao, T., Pan, Y., Li, Y., Qiu, Z., Mei, T.: Boosting image captioning with attributes. In: IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, 22–29 October 2017, pp. 4904–4912 (2017)
30. You, Q., Jin, H., Wang, Z., Fang, C., Luo, J.: Image captioning with semantic attention. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, 27–30 June 2016, pp. 4651–4659 (2016)