



A survey on automatic image caption generation

Shuang Bai^{a,*}, Shan An^b

^a School of Electronic and Information Engineering, Beijing Jiaotong University, No.3 Shang Yuan Cun, Hai Dian District, Beijing, China

^b Beijing Jingdong Shangke Information Technology Co., Ltd, Beijing, China

ARTICLE INFO

Article history:

Received 5 May 2017

Revised 13 April 2018

Accepted 19 May 2018

Available online 26 May 2018

Communicated by Dr. Min Xu

Keywords:

Image captioning

Sentence template

Deep neural networks

Multimodal embedding

Encoder–decoder framework

Attention mechanism

ABSTRACT

Image captioning means automatically generating a caption for an image. As a recently emerged research area, it is attracting more and more attention. To achieve the goal of image captioning, semantic information of images needs to be captured and expressed in natural languages. Connecting both research communities of computer vision and natural language processing, image captioning is a quite challenging task. Various approaches have been proposed to solve this problem. In this paper, we present a survey on advances in image captioning research. Based on the technique adopted, we classify image captioning approaches into different categories. Representative methods in each category are summarized, and their strengths and limitations are talked about. In this paper, we first discuss methods used in early work which are mainly retrieval and template based. Then, we focus our main attention on neural network based methods, which give state of the art results. Neural network based methods are further divided into subcategories based on the specific framework they use. Each subcategory of neural network based methods are discussed in detail. After that, state of the art methods are compared on benchmark datasets. Following that, discussions on future research directions are presented.

© 2018 Elsevier B.V. All rights reserved.

1. Introduction

Humans are able to relatively easily describe the environments they are in. Given an image, it is natural for a human to describe an immense amount of details about this image with a quick glance [1]. This is one of humans' basic abilities. Making computers imitate humans' ability to interpret the visual world has been a long standing goal of researchers in the field of artificial intelligence.

Although great progress has been made in various computer vision tasks, such as object recognition [2], [3], attribute classification [4], [5], action classification [6], [7], image classification [8] and scene recognition [9], [10], it is a relatively new task to let a computer use a human-like sentence to automatically describe an image that is forwarded to it.

Using a computer to automatically generate a natural language description for an image, which is defined as image captioning, is challenging. Because connecting both research communities of computer vision and natural language processing, image captioning not only requires a high level understanding of the semantic contents of an image, but also needs to express the information in a human-like sentence. Determination of presences, attributes and

relationships of objects in an image is not an easy task itself. Organizing a sentence to describe such information makes this task even more difficult.

Since much of human communication depends on natural languages, whether written or spoken, enabling computers to describe the visual world will lead to a great number of possible applications, such as producing natural human robot interactions, early childhood education, information retrieval, and visually impaired assistance, and so on.

As a challenging and meaningful research field in artificial intelligence, image captioning is attracting more and more attention and is becoming increasingly important.

Given an image, the goal of image captioning is to generate a sentence that is linguistically plausible and semantically truthful to the content of this image. So there are two basic questions involved in image captioning, i.e. visual understanding and linguistic processing. To ensure generated sentences are grammatically and semantically correct, techniques of computer vision and natural language processing are supposed to be adopted to deal with problems arising from the corresponding modality and integrated appropriately. To this end, various approaches have been proposed.

Originally, automatic image captioning is only attempted to yield simple descriptions for images taken under extremely constrained conditions. For example, Kojima et al. [11] used concept hierarchies of actions, case structures and verb patterns to generate natural languages to describe human activities in a fixed office

* Corresponding author.

E-mail address: shuangb@bjtu.edu.cn (S. Bai).

Table 1
Summary of image captioning methods.

Method	Representative methods
Early work	Retrieval based Farhadi et al. [13], Ordonez et al. [15], Gupta et al. [16], Hodosh et al. [32], Mason and Charniak [49], Kuznetsova et al. [50].
Neural networks based	Template based Yang et al. [14], Kulkarni et al. [51], Li et al. [52], Mitchell et al. [53], Ushiku et al. [54]. Augmenting early work by deep models Socher et al. [55], Karpathy et al. [37], Ma et al. [56], Yan and Mikolajczyk [57], Lebrete et al. [58]. Multimodal learning Kiros et al. [59], Mao et al. [60], Karpathy and Li [61], Chen and Zitnick [62]. Encoder–decoder framework Kiros et al. [63], Vinyals et al. [64], Donahue et al. [34], Jia et al. [65], Wu et al. [66], Pu et al. [67]. Attention guided Xu et al. [68], You et al. [69], Yang et al. [70]. Compositional architectures Fang et al. [33], Tran et al. [71], Fu et al. [72], Ma and Han [73], Oruganti et al. [74], Wang et al. [75]. Describing novel objects Mao et al. [76], Hendricks and Venugopalan, [36].

environment. Hede et al. used a dictionary of objects and language templates to describe images of objects in backgrounds without clutters [12]. Apparently, such methods are far from applications to describing images that we encounter in our everyday life.

It is not until recently that work aiming to generate descriptions for generic real life images is proposed [13]–[16]. Early work on image captioning mainly follows two lines of research, i.e. retrieval based and template based. Because these methods accomplish the image captioning task either by making use of existing captions in the training set or relying on hard-coded language structures, the disadvantage of methods adopted in early work is that they are not flexible enough. As a result, expressiveness of generated descriptions by these methods is, to a large extent, limited.

Despite the difficult nature of the image captioning task, thanks to recent advances in deep neural networks [17–22], which are widely applied to the fields of computer vision [23–26] and natural language processing [27–31], image captioning systems based on deep neural networks are proposed. Powerful deep neural networks provide efficient solutions to visual and language modelling. Consequently, they are used to augment existing systems and design countless new approaches. Employing deep neural networks to tackle the image captioning problem has demonstrated state of the art results [32]–[37].

With the recent surge of research interest in image captioning, a large number of approaches have been proposed. To facilitate readers to have a quick overview of the advances of image captioning, we present this survey to review past work and envision future research directions. Although there exist several research topics that also involve both computer vision and natural language processing, such as visual question answering [38–42], text summarization [43], [44] and video description [45–48], because each of them has its own focus, in this survey we mainly focus on work that aims to automatically generate descriptions for generic real life images.

Based on the technique adopted in each method, we classify image captioning approaches into different categories, which are summarized in Table 1. Representative methods in each category are listed. Methods in early work are mainly retrieval and template based, in which hard coded rules and hand engineered features are utilized. Outputs of such methods have obvious limitations. We review early work relatively briefly in this survey. With the great progress made in research of deep neural networks, approaches that employ neural networks for image captioning are proposed and demonstrate state of the art results. Based on the framework used in each deep neural network based method, we further classify these methods into subcategories. In this survey, we will focus our main attention on neural network based methods. The framework used in each subcategory will be introduced, and the corresponding representative methods will be discussed in more detail.

This paper is organized as follows. In Sections 2 and 3, we first review retrieval based and template based image captioning methods, respectively. Section 4 is about neural network based methods,

in this section we divide neural network based image captioning methods into subcategories, and discuss representative methods in each subcategory, respectively. State of art methods will be compared on benchmark datasets in Section 5. After that, we will envision future research directions of image captioning in Section 6. The conclusion will be given in Section 7.

2. Retrieval based image captioning

One type of image captioning methods that are common in early work is retrieval based. Given a query image, retrieval based methods produce a caption for it through retrieving one or a set of sentences from a pre-specified sentence pool. The generated caption can either be a sentence that has already existed or a sentence composed from the retrieved ones. First, let us investigate the line of research that directly uses retrieved sentences as captions of images.

Farhadi et al. establish a $\langle \text{object}, \text{action}, \text{scene} \rangle$ meaning space to link images and sentences. Given a query image, they map it into the meaning space by solving a Markov Random Field, and use Lin similarity measure [77] to determine the semantic distance between this image and each existing sentence parsed by Curran et al. parser [78]. The sentence closest to the query image is taken as its caption [13].

In [15], to caption an image Ordonez et al. first employ global image descriptors to retrieve a set of images from a web-scale collection of captioned photographs. Then, they utilize semantic contents of the retrieved images to perform re-ranking and use the caption of the top image as the description of the query.

Hodosh et al. frame image captioning as a ranking task [32]. The authors employ the Kernel Canonical Correlation Analysis technique [79], [80] to project image and text items into a common space, where training images and their corresponding captions are maximally correlated. In the new common space, cosine similarities between images and sentences are calculated to select top ranked sentences to act as descriptions of query images.

To alleviate impacts of noisy visual estimation in methods that depend on image retrieval for image captioning, Mason and Charniak first use visual similarity to retrieve a set of captioned images for a query image [49]. Then, from the captions of the retrieved images, they estimate a word probability density conditioned on the query image. The word probability density is used to score the existing captions to select the one with the largest score as the caption of the query.

The above methods have implicitly assumed that given a query image there always exists a sentence that is pertinent to it. This assumption is hardly true in practice. Therefore, instead of using retrieved sentences as descriptions of query images directly, in the other line of retrieval based research, retrieved sentences are utilized to compose a new description for a query image.

Provided with a dataset of paired images and sentences, Gupta et al. use Stanford CoreNLP toolkit¹ to process sentences in the dataset to derive a list of phrases for each image. In order to generate a description for a query image, image retrieval is first performed based on global image features to retrieve a set of images for the query. Then, a model trained to predicate phrase relevance is used to select phrases from the ones associated with retrieved images. Finally a description sentence is generated based on the selected relevant phrases [16].

With a similar idea, Kuznetsova et al. propose a tree based method to compose image descriptions by making use of captioned web images [50]. After performing image retrieval and phrase extraction, the authors take extracted phrases as tree fragments and model description composition as a constraint optimization problem, which is encoded by using Integer Linear Programming [81], [82] and solved by using the CPLEX solver². Before this paper, the same authors have reported a similar method in [83].

Disadvantages of retrieval based image captioning methods are obvious. Such methods transfer well-formed human-written sentences or phrases for generating descriptions for query images. Although the yielded outputs are usually grammatically correct and fluent, constraining image descriptions to sentences that have already existed can not adapt to new combinations of objects or novel scenes. Under certain conditions, generated descriptions may even be irrelevant to image contents. Retrieval based methods have large limitations to their capability to describe images.

3. Template based image captioning

In early image captioning work, another type of methods that are commonly used is template based. In template based methods, image captions are generated through a syntactically and semantically constrained process. Typically, in order to use a template based method to generate a description for an image, a specified set of visual concepts need to be detected first. Then, the detected visual concepts are connected through sentence templates or specific language grammar rules or combinatorial optimization algorithms [84], [53] to compose a sentence.

A method to use a sentence template for generating image descriptions is presented in [14] by Yang et al., where a quadruplet (Nouns-Verbs-Scenes-Prepositions) is utilized as a sentence template. To describe an image, the authors first use detection algorithms [2], [85] to estimate objects and scenes in this image. Then, they employ a language model [86] trained over the Gigaword corpus³ to predicate verbs, scenes and prepositions that may be used to compose the sentence. With probabilities of all elements computed, the best quadruplet is obtained by using Hidden Markov Model inference. Finally, the image description is generated by filling the sentence structure given by the quadruplet.

Kulkarni et al. employ Conditional Random Field to determine image contents to be rendered in the image caption [87], [51]. In their method, nodes of the graph correspond to objects, object attributes and spatial relationships between objects, respectively. In the graph model, unary potential functions of nodes are obtained by using corresponding visual models, while pairwise potential functions are obtained by making statistics on a collection of existing descriptions. Image contents to be described are determined by performing Conditional Random Field inference. Outputs of the inference is used to generate a description based on a sentence template.

Li et al. use visual models to perform detections in images for extracting semantic information including objects, attributes and spatial relationships [52]. Then, they define a triplet of the format $\langle \langle \text{adj1}, \text{obj1} \rangle, \text{prep}, \langle \text{adj2}, \text{obj2} \rangle \rangle$ for encoding recognition results. To generate a description with the triplet, web-scale n -gram data, which is able to provide frequency counts of possible n -gram sequences, is resorted to for performing phrase selection, so that candidate phrases that may compose the triplet can be collected. After that, phrase fusion is implemented to use dynamic programming to find the optimal compatible set of phrases to act as the description of the query image.

Mitchell et al. employ computer vision algorithms to process an image and represent this image by using $\langle \text{objects}, \text{actions}, \text{spatial relationships} \rangle$ triplets [53]. After that, they formulate image description as a tree-generating process based on the visual recognition results. Through object nouns clustering and ordering, the authors determine image contents to describe. Then sub-trees are created for object nouns, which are further used for creating full trees. Finally, a trigram language model [88] is used to select a string from the generated full trees as the description of the corresponding image.

Methods mentioned above use visual models to predicate individual words from a query image in a piece-wise manner. Then, predicted words such as objects, attributes, verbs and prepositions are connected in later stages to generate human-like descriptions. Since phrases are combinations of words, compared to individual words, phrases carry bigger chunks of information [89]. Sentences yielded based on phrases tend to be more descriptive. Therefore, methods utilizing phrases under the template based image captioning framework are proposed.

Ushiku et al. present a method called Common Subspace for Model and Similarity to learn phrase classifiers directly for captioning images [54]. Specifically, the authors extract continuous words [84] from training captions as phrases. Then, they map image features and phrase features into the same subspace, where similarity based and model based classification are integrated to learn a classifier for each phrase. In the testing stage, phrases estimated from a query image are connected by using multi-stack beam search [84] to generate a description.

Template based image captioning can generate syntactically correct sentences, and descriptions yielded by such methods are usually more relevant to image contents than retrieval based ones. However, there are also disadvantages for template based methods. Because description generation under the template based framework is strictly constrained to image contents recognized by visual models, with the typically small number of visual models available, there are usually limitations to coverage, creativity, and complexity of generated sentences. Moreover, compared to human-written captions, using rigid templates as main structures of sentences will make generated descriptions less natural.

4. Deep neural network based image captioning

Retrieval based and template based image captioning methods are adopted mainly in early work. Due to great progress made in the field of deep learning [18], [90], recent work begins to rely on deep neural networks for automatic image captioning. In this section, we will review such methods. Even though deep neural networks are now widely adopted for tackling the image captioning task, different methods may be based on different frameworks. Therefore, we classify deep neural network based methods into subcategories on the basis of the main framework they use and discuss each subcategory, respectively.

¹ <http://nlp.stanford.edu/software/corenlp.shtml>.

² ILOG CPLEX: High-performance software for mathematical programming and optimization. <http://www.ilog.com/products/cplex/>.

³ <https://catalog.ldc.upenn.edu/LDC2003T05>.

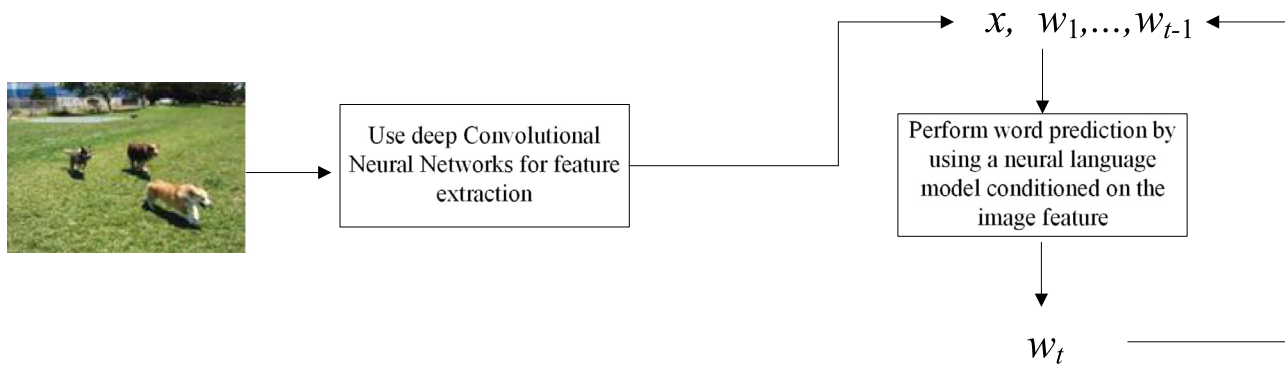


Fig. 1. General structure of multimodal learning based image captioning methods.

4.1. Retrieval and template based methods augmented by neural networks

Encouraged by advances in the field of deep neural networks, instead of utilizing hand-engineered features and shallow models like in early work, deep neural networks are employed to perform image captioning. With inspiration from retrieval based methods, researchers propose to utilize deep models to formulate image captioning as a multi-modality embedding [91] and ranking problem.

To retrieve a description sentence for a query image, Socher et al. propose to use dependency-tree recursive neural networks to represent phrases and sentences as compositional vectors. They use another deep neural network [92] as visual model to extract features from images [55]. Obtained multimodal features are mapped into a common space by using a max-margin objective function. After training, correct image and sentence pairs in the common space will have larger inner products and vice versa. At last, sentence retrieval is performed based on similarities between representations of images and sentences in the common space.

Karpathy et al. propose to embed sentence fragments and image fragments into a common space for ranking sentences for a query image [37]. They use dependency tree relations [93] of a sentence as sentence fragments and use detection results of the Region Convolutional Neural Network method [3] in an image as image fragments. Representing both image fragments and sentence fragments as feature vectors, the authors design a structured max-margin objective, which includes a global ranking term and a fragment alignment term, to map visual and textual data into a common space. In the common space, similarities between images and sentences are computed based on fragment similarities, as a result sentence ranking can be conducted at a finer level.

In order to measure similarities between images and sentences with different levels of interactions between them taken into consideration, Ma et al. propose a multimodal Convolutional Neural Network [56]. Ma's framework includes three kinds of components, i.e. image CNNs to encode visual data [94], [95], matching CNNs to jointly represent visual and textual data [96], [97] and multi-layer perceptions to score compatibility of visual and textual data. The authors use different variants of matching CNNs to account for joint representations of images and words, phrases and sentences. The final matching score between an image and a sentence is determined based on an ensemble of multimodal Convolutional Neural Networks.

Yan and Mikolajczyk propose to use deep Canonical Correlation Analysis [98] to match images and sentences [57]. They use a deep Convolutional Neural Network [8] to extract visual features from images and use a stacked network to extract textual features from Frequency-Inverse Document Frequency represented sentences. The Canonical Correlation Analysis objective is employed to map visual and textual features to a joint latent space

with correlation between paired features maximized. In the joint latent space, similarities between an image feature and a sentence feature can be computed directly for sentence retrieval.

Besides using deep models to augment retrieval based image captioning methods, utilizing deep models under the template based framework is also attempted. Lebre et al. leverage a kind of soft-template to generate image captions with deep models [58]. In this method, the authors use the SENNA software⁴ to extract phrases from training sentences and make statistics on the extracted phrases. Phrases are represented as high-dimensional vectors by using a word vector representation approach [31], [99], [100], and images are represented by using a deep Convolutional Neural Network [94]. A bilinear model is trained as a metric between image features and phrase features, so that given a query image, phrases can be inferred from it. Phrases inferred from an image are used to generate a sentence under the guidance of statistics made in the early stage.

With the utilization of deep neural networks, performances of image captioning methods are improved significantly. However, introducing deep neural networks into retrieval based and template based methods does not overcome their disadvantages. Limitations of sentences generated by these methods are not removed.

4.2. Image captioning based on multimodal learning

Retrieval based and template based image captioning methods impose limitations on generated sentences. Thanks to powerful deep neural networks, image captioning approaches are proposed that do not rely on existing captions or assumptions about sentence structures in the caption generation process. Such methods can yield more expressive and flexible sentences with richer structures. Using multimodal neural networks is one of the attempts that rely on pure learning to generate image captions.

General structure of multimodal learning based image captioning methods is shown in Fig. 1. In such kind of methods, image features are first extracted by using a feature extractor, such as deep convolutional neural networks. Then, the obtained image feature is forwarded to a neural language model, which maps the image feature into the common space with the word features and perform word predication conditioned on the image feature and previously generated context words.

Kiros et al. propose to use a neural language model which is conditioned on image inputs to generate captions for images [59]. In their method, log-bilinear language model [30] is adapted to multimodal cases. In a natural language processing problem, a language model is used to predicate the probability of generating a

⁴ Available at <http://ml.nec-labs.com/senna/>.

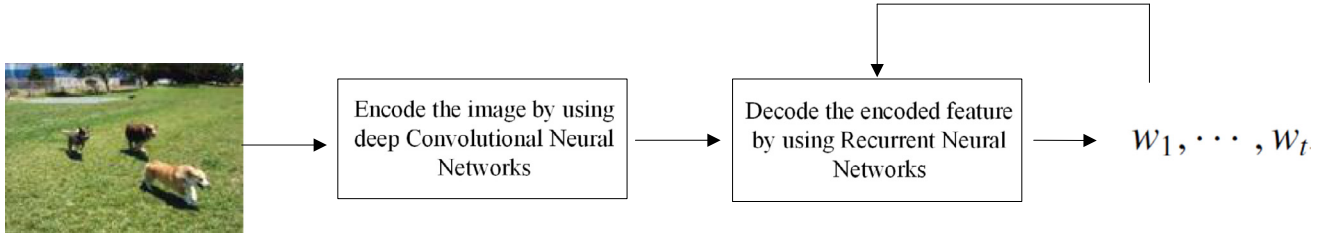


Fig. 2. General structure of encoder-decoder based image captioning methods.

word w_t conditioned on previously generated words w_1, \dots, w_{t-1} , which is shown below:

$$P(w_t | w_1, \dots, w_{t-1}). \quad (1)$$

The authors make the language model become dependent on images through two different ways, i.e. adding an image feature as an additive bias to the representation of the next predicted word and gating the word representation matrix by using the image feature. Consequently, in the multimodal case the probability of generating a word w_t is as follows:

$$P(w_t | w_1, \dots, w_{t-1}, I). \quad (2)$$

where I is an image feature. In their method, images are represented by a deep Convolutional Neural Network, and joint image-text feature learning is implemented by back propagating gradients from the loss function through the multimodal neural network model. By using this model, an image caption can be generated word by word, with the generation of each word conditioned on previously generated words and the image feature.

To generate novel captions for images, Mao et al. adapt a Recurrent Neural Network language model to multimodal cases for directly modelling the probability of generating a word conditioned on a given image and previously generated words [60], [35]. Under their framework, a deep Convolutional Neural Network [8] is used to extract visual features from images, and a Recurrent Neural Network [101] with a multimodal part is used to model word distributions conditioned on image features and context words. For the Recurrent Neural Network language model, each unit consists of an input word layer w , an recurrent layer r and an output layer y . At the t_{th} unit of the Recurrent Neural Network language model, the calculation performed by these three layers is shown as follows:

$$x(t) = [w(t) \ r(t-1)], \quad (3)$$

$$r(t) = f(U \cdot x(t)), \quad (4)$$

$$y(t) = g(V \cdot r(t)), \quad (5)$$

where $f(\cdot)$ and $g(\cdot)$ are element-wise non-linear functions, and U and V are matrices of weights to be learned. The multimodal part calculates its layer activation vector $m(t)$ by using the equation below:

$$m(t) = g_m(V_w \cdot w(t) + V_r \cdot r(t) + V_I \cdot I), \quad (6)$$

where g_m is a non-linear function. I is the image feature. V_w , V_r and V_I are matrices of weights to be learned. The multimodal part fuses image features and distributed word representations by mapping and adding them. To train the model, a perplexity based cost function is minimized based on back propagation.

Schuster and Paliwal present an approach to align image regions represented by a Convolutional Neural Network and sentence segments represented by a Bidirectional Recurrent Neural Network [102] to learn a multimodal Recurrent Neural Network model to generate descriptions for image regions [61]. In their method, after

representing image regions and sentence segments by using corresponding neural networks, a structured objective is used to map visual and textual data into a common space and associate each region feature to the textual feature that describes the region. The aligned two modalities are then employed to train a multimodal Recurrent Neural Network model, that can be used to predicate the probability of generating the next word given an image feature and context words.

Recurrent Neural Networks are known to have difficulties in learning long term dependencies [103], [104]. To alleviate this weakness in image captioning, Chen and Zitnick propose to dynamically build a visual representation of an image as a caption is being generated for it, so that long term visual concepts can be remembered during this process [62]. To this end, a set of latent variables U_{t-1} are introduced to encode visual interpretation of words W_{t-1} that have already been generated. With these latent variables, the probability of generating a word w_t is given below:

$$P(w_t, V | W_{t-1}, U_{t-1}) = P(w_t | V, W_{t-1}, U_{t-1})P(V | W_{t-1}, U_{t-1}), \quad (7)$$

where V denotes observed visual features, and W_{t-1} denotes generated words (w_1, \dots, w_{t-1}) . The authors realize the above idea through adding recurrent visual hidden layer u into the Recurrent Neural Networks. The recurrent layer u is helpful for both reconstructing the visual features V from previous words W_{t-1} and predicting the next word w_t .

4.3. Image captioning based on the encoder-decoder framework

Inspired by recent advances in neural machine translation [28], [105], [106], the encoder-decoder framework is adopted to generate captions for images. General structure of encoder-decoder based image captioning methods is shown in Fig. 2. This framework is originally designed to translate sentences from one language into another language. Motivated by the neural machine translation idea, it is argued that image captioning can be formulated as a translation problem, where the input is an image, while the output is a sentence [63]. In image captioning methods under this framework, an encoder neural network first encodes an image into an intermediate representation, then a decoder recurrent neural network takes the intermediate representation as input and generate a sentence word by word.

Kiros et al. introduce the encoder-decoder framework into image captioning research to unify joint image-text embedding models and multimodal neural language models, so that given an image input, a sentence output can be generated word by word [63] like language translation. They use Long Short-Term Memory (LSTM) Recurrent Neural Networks to encode textual data [107] and a deep Convolutional Neural Network to encode visual data. Then, through optimizing a pairwise ranking loss, encoded visual data is projected into an embedding space spanned by LSTM hidden states that encode textual data. In the embedding space, a structure-content neural language model is used to decode visual features conditioned on context word feature vectors, allowing for sentence generation word by word.

With the same inspiration from neural machine translation, Vinyals et al. use a deep Convolutional Neural Network as an encoder to encode images and use Long Short-Term Memory (LSTM) Recurrent Neural Networks to decode obtained image features into sentences [64] [108]. With the above framework, the authors formulate image captioning as predicating the probability of a sentence conditioned on an input image:

$$S^* = \arg \max_S P(S | I; \theta) \quad (8)$$

where I is an input image and θ is the model parameter. Since a sentence S equals to a sequence of words (S_0, \dots, S_{T+1}) , with chain rule Eq. (8) is reformulated below:

$$S^* = \arg \max_S \prod P(S_t | I, S_0, \dots, S_{t-1}; \theta). \quad (9)$$

Vinyals et al. use a Long Short-Term Memory neural network to model $P(S_t | I, S_0, \dots, S_{t-1}; \theta)$ as hidden state h_t , which can be updated by a update function below:

$$h_{t+1} = f(h_t, x_t), \quad (10)$$

where x_t is the input to the Long Short-Term Memory neural network. In the first unit, x_t is an image feature, while in other units x_t is a feature of previously predicated context words. The model parameter θ is obtained by maximizing the likelihood of sentence image pairs in the training set. With the trained model, possible output word sequences can be predicted by either sampling or beam search.

Similar to Vinyals's work [64], [108], Donahue et al. also adopt a deep Convolutional Neural Network for encoding and Long Short-Term Memory Recurrent Networks for decoding to generate a sentence description for an input image [34]. The difference is that instead of inputting image features to the system only at the initial stage, Donahue et al. provide both image feature and context word feature to the sequential model at each time step.

It has demonstrated promising results to use the encoder-decoder framework to tackle the image captioning problem. Encouraged by the success, approaches aiming to augment this framework for obtaining better performances are proposed.

Aiming to generate image descriptions that are closely related to image contents, Jia et al. extract semantic information from images and add the information to each unit of the Long Short-Term Memory Recurrent Neural Networks during the process of sentence generation [65]. The original forms of the memory cell and gates of a LSTM unit [109] are defined as follows:

$$i_l = \sigma(W_{ix}x_l + W_{im}m_{l-1}), \quad (11)$$

$$f_l = \sigma(W_{fx}x_l + W_{fm}m_{l-1}), \quad (12)$$

$$o_l = \sigma(W_{ox}x_l + W_{om}m_{l-1}), \quad (13)$$

$$c_l = f_l \odot c_{l-1} + i_l \odot h(W_{cx}x_l + W_{cm}m_{l-1}), \quad (14)$$

$$m_l = o_l \odot c_l, \quad (15)$$

where $\sigma(\cdot)$ and $h(\cdot)$ are non-linear functions, variables i_l , f_l and o_l stand for input gate, forget gate, output gate of a LSTM cell, respectively, c_l and m_l stand for the state and hidden state of the memory cell unit, x_l is the input, $W_{[\cdot][\cdot]}$ are model parameters, and \odot denotes an element-wise multiplication operation. With the addition of semantic information to an LSTM unit, the forms of the memory cell and gates are changed to be as follows:

$$i'_l = \sigma(W_{ix}x_l + W_{im}m'_{l-1} + W_{ig}g), \quad (16)$$

$$f'_l = \sigma(W_{fx}x_l + W_{fm}m'_{l-1} + W_{fg}g), \quad (17)$$

$$o'_l = \sigma(W_{ox}x_l + W_{om}m'_{l-1} + W_{og}g), \quad (18)$$

$$c'_l = f'_l \odot c'_{l-1} + i'_l \odot h(W_{cx}x_l + W_{cm}m'_{l-1} + W_{cg}g), \quad (19)$$

$$m'_l = o'_l \odot c'_l, \quad (20)$$

where g is the representation of semantic information, which can be from any sources as long as it can provide guidance for image captioning.

Given an image, approaches introduced above seek to directly derive a description from its visual features. In order to utilize high-level semantic information for image captioning, Wu et al. incorporate visual concepts into the encoder-decoder framework [66]. To this end, the authors first mined a set of semantic attributes from the training sentences. Under the region-based multi-label classification framework [110], a Convolutional Neural Network based classifier is trained for each attribute. With trained semantic attribute classifiers, an image can be represented as a prediction vector $V_{att}(I)$ giving the probability of each attribute to be present in the image. After encoding an image I as $V_{att}(I)$, a Long Short-Term Memory network [107] is employed as a decoder to generate a sentence describing the contents of the image based on the representation. Under this condition, the image captioning problem can be reformulated below:

$$S^* = \arg \max_S P(S | V_{att}(I); \theta) \quad (21)$$

where I is the input image. S is a sentence. θ is the model parameter.

Because in practical applications, there may be far less captioned images than uncaptioned ones, semi-supervised learning of image captioning models is of significant practical values. To obtain an image captioning system by leveraging the vast quantity of uncaptioned images available, Pu et al. propose a semi-supervised learning method under the encoder-decoder framework to use a deep Convolutional Neural Network to encode images and a Deep Generative Deconvolutional Network to decode latent image features for image captioning [67]. The system uses the deep Convolutional Neural Network to provide an approximation to the distribution of the latent features of the Deep Generative Deconvolutional Network and link the latent features to generative models for captions. After training, given an image, the caption can be generated by averaging across the distribution of latent features of Deep Generative Deconvolutional Network.

4.4. Attention guided image captioning

It is well-known that images are rich in information they contain, while in image captioning it is unnecessary to describe all details of a given image. Only the most salient contents are supposed to be mentioned in the description. Motivated by the visual attention mechanism of primates and humans [111], [112], approaches that utilize attention to guide image description generation are proposed. By incorporating attention to the encoder-decoder image captioning framework, sentence generation will be conditioned on hidden states that are computed based on attention mechanism. General structure of attention guided image captioning methods is given Fig. 3. In such methods, attention mechanism based on various kinds of cues from the input image is incorporated into the encoder-decoder framework to make the decoding process focus on certain aspects of the input image at each time step to generate a description for the input image.

Encouraged by successes of other tasks that employ attention mechanism [113–115], Xu et al. propose an attentive encoder-decoder model to be able to dynamically attend salient image regions during the process of image description generation [68]. Forwarding an image to a deep Convolutional Neural Network and extracting features from a lower

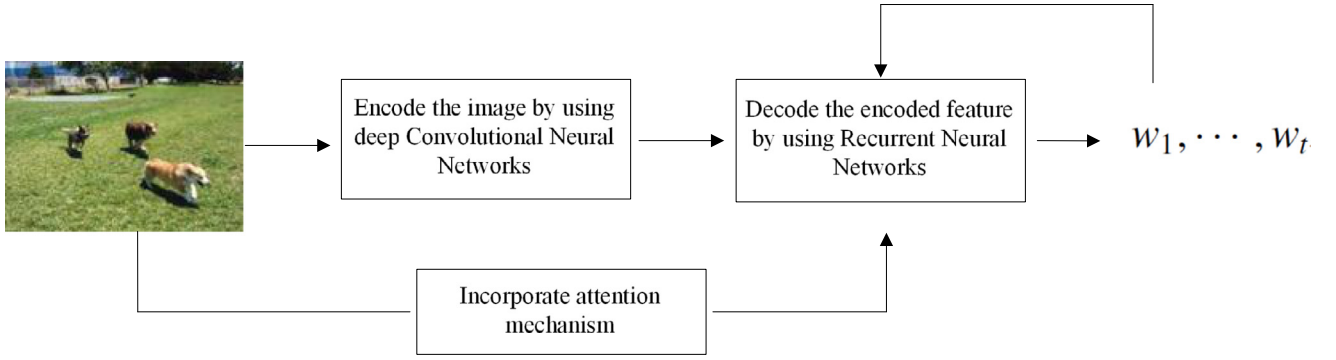


Fig. 3. General structure of attention guided image captioning methods.

convolutional layer of the network, the authors encode an image as a set of feature vectors which is shown as follows:

$$a = (a_1, \dots, a_N), \quad a_i \in \mathbb{R}^D, \quad (22)$$

where a_i is a D -dimensional feature vector that represents one part of the image. As a result, an image is represented by N vectors. In the decoding stage, a Long Short-Term Memory network is used as the decoder. Different from previous LSTM versions, a context vector z_l is utilized to dynamically represent image parts that are relevant for caption generation at time l . Consequently, the memory cell and gates of a LSTM unit become the forms given below:

$$i_l = \sigma(W_{ix}x_l + W_{im}m_{l-1} + W_{iz}z_l), \quad (23)$$

$$f_l = \sigma(W_{fx}x_l + W_{fm}m_{l-1} + W_{fz}z_l), \quad (24)$$

$$o_l = \sigma(W_{ox}x_l + W_{om}m_{l-1} + W_{oz}z_l), \quad (25)$$

$$c_l = f_l \odot c_{l-1} + i_l \odot h(W_{cx}x_l + W_{cm}m_{l-1} + W_{cz}z_l), \quad (26)$$

$$m_l = o_l \odot c_l. \quad (27)$$

Attention is imposed to the decoding process by using the context vector z_l , which is a function of image region vectors (a_1, \dots, a_N) and weights associated with them $(\alpha_1, \dots, \alpha_N)$:

$$z_l = \phi(\{a_i\}, \{\alpha_i\}). \quad (28)$$

With different function forms, different attention mechanisms can be applied. In [68], Xu et al. proposed a stochastic hard attention and a deterministic soft attention for image captioning. In each time step, the stochastic hard attention mechanism selects a visual feature from one of the N locations as the context vector to generate a word, while the deterministic soft attention mechanism combines visual features from all N locations to obtain the context vector to generate a word.

Specifically, in the stochastic hard attention mechanism, at time step l , for each location i , the positive weight $\alpha_{l,i}$ associated with it is taken as the probability for this location to be focused on for generating the corresponding word. The context vector z_l is calculated as follows:

$$z_l = \sum_i^N s_{l,i} a_i. \quad (29)$$

where $s_{l,i}$ is an indicator variable, which is set to 1, if the visual feature a_i from the i_{th} location out of N is attended at time step l , otherwise 0. The distribution of the variable $s_{l,i}$ is treated as a multinoulli distribution parametrized by $\{\alpha_{l,i}\}$, and its value is determined based on sampling.

Contrarily, in the deterministic soft attention mechanism, the positive weight $\alpha_{l,i}$ associated with location i at time step l is used to represent the relative importance of the corresponding location in blending visual features from all N locations to calculate the context vector z_l , which is formulated below:

$$z_l = \sum_i^N \alpha_{l,i} a_i. \quad (30)$$

Finding that both bottom-up [13], [87], [116] and top-down [34], [61], [62] image captioning approaches have certain limitations, You et al. propose a semantic attention model to take advantages of the complimentary properties of both types of approaches [69]. To achieve this goal, the authors use a deep Convolutional Neural Network and a set of visual attribute detectors to extract a global feature v and a list of visual attributes $\{A_i\}$ from an input image, respectively. With each attribute corresponding to one entry of the used vocabulary, words to generate and attributes to detect share the same vocabulary. Under the encoder–decoder framework, the global visual feature v is only forwarded to the encoder at the initial step. In the decoding stage, using an input attention function $\phi(\cdot)$, certain cognitive visual cues in the attribute list $\{A_i\}$ will be attended with a probability distribution:

$$\{\alpha_t^i\} = \phi(y_{t-1}, \{A_i\}), \quad (31)$$

where α_t^i is the weight assigned to an attribute in the list, and y_{t-1} is the previously generated word. These weights are used to calculate input vector x_t to the t_{th} unit of a Long Short-Term Memory neural network. With an output attention function $\varphi(\cdot)$, the attention on all the attributes will be modulated by using the weights given below:

$$\{\beta_t^i\} = \varphi(m_t, \{A_i\}), \quad (32)$$

where β_t^i is the weight assigned to an attribute. m_t is the hidden state of t_{th} unit of the Long Short-Term Memory neural network. The obtained weights are further used to predicate probability distribution of the next word to be generated.

Arguing that attentive encoder–decoder models lack global modelling abilities due to their sequential information processing manner, Yang et al. propose a review network to enhance the encoder–decoder framework [70]. To overcome the above-mentioned problem, a reviewer module is introduced to perform review steps on the hidden states of the encoder and give a thought vector at each step. During this process, attention mechanism is applied to determine weights assigned to hidden states. Through this manner, information encoded by the encoder can be reviewed and learned by the thought vectors which can capture global properties of the input. Obtained thought vectors are used by the decoder for word predication. Specifically, the authors use the VGGNet [94], which is a commonly used deep Convolutional

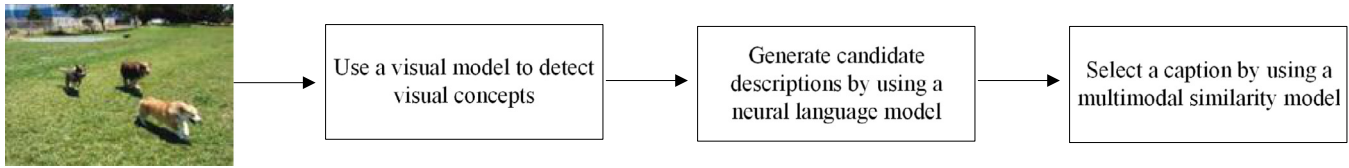


Fig. 4. General structure of compositional image captioning methods.

Neural Network to encode an image as a context vector c and a set of hidden states $H = \{h_t\}$. A Long-Short Term Memory neural network is used as reviewer to produce thought vectors. A thought vector f_t at the t_{th} LSTM unit is calculated as follows:

$$f_t = g_t(H, f_{t-1}), \quad (33)$$

where g_t is a function performed by a reviewer with attention mechanism applied. After obtaining thought vectors $F = \{f_t\}$, a Long-Short Term Memory neural network decoder can predicate word probability distribution based on them as given below:

$$y_t = g'_t(F, s_{t-1}, y_{t-1}), \quad (34)$$

where s_t is the hidden state of the t_{th} LSTM unit in the decoder. y_t is the t_{th} word.

4.5. Compositional architectures for image captioning

In Section 4, we focus on image captioning methods that are based on deep neural networks. Most of the approaches in previous subsections are based on end-to-end frameworks, whose parameters can be trained jointly. Such methods are neat and efficient. However, believing that each type of approaches have their own advantages and disadvantages, architectures composed of independent building blocks are proposed for image captioning. In this subsection, we will talk about compositional image captioning architectures that are consisted of independent functional building blocks that may be used in different types of methods.

General structure of compositional image captioning methods is given Fig. 4. In contrast to end-to-end image captioning framework, compositional image captioning methods integrate independent building blocks into a pipeline to generate captions for input images. Generally, compositional image captioning methods use a visual model to detect visual concepts appearing in the input image. Then, detected visual concepts are forwarded to a language model to generate candidate descriptions, which are then post-processed to select one of them as the caption of the input image.

Fang et al. propose a system that is consisted of visual detectors, language models and multimodal similarity models for automatic image captioning [33]. The authors first detect a vocabulary of words that are most common in the training captions. Then, corresponding to each word, a visual detector is trained by using a Multiple Instance Learning approach [117]. Visual features used by these detectors are extracted by a deep Convolutional Neural Network [8]. Given an image, conditioned on the words detected from it, a maximum entropy language model [118] is adopted to generate candidate captions. During this process, left-to-right beam search [119] with a stack of pre-specified length of l is performed. Consequently, l candidate captions are obtained for this image. Finally, a deep multimodal similarity model, which maps images and text fragments into a common space for similarity measurement is used to re-rank the candidate descriptions.

Based on Fang's et al. work [33], Tran et al. presented a system for captioning open domain images [71]. Similar to [33], the authors use a deep residual network based vision model to detect a broad range of visual concepts [120], a maximum entropy language model for candidate description generation, and a deep multimodal similarity model for caption ranking. What's more, the

authors added detection for landmarks and celebrities and a confidence model for dealing with images that are difficult to describe.

To exploit parallel structures between images and sentences for image captioning, Fu et al. propose to align the word generation process to visual perception of image regions [72]. Furthermore, the authors introduce scene-specific contexts to capture high-level semantic information in images for adapting word generation to specific scene types. Given an image, Fu et al. first use the selective search method [121] to extract a large number of image regions. Then, based on the criterion of being semantically meaningful, non-compositional and contextually rich, a small number of them are selected for further processing. Each selected region is represented as a visual feature by using the ResNet network [120]. These features are dynamically attended by an attention-based decoder, which is a Long-Short Term Memory neural network [107]. Finally, to exploit semantic-contexts in images for better captioning, Latent Dirichlet Allocation [122] and a multilayer perceptron are used to predicate a context vector for an image to bias the word generation in the Long-Short Term Memory neural network.

To be able to produce detailed descriptions about image contents, Ma and Han propose to use structural words for image captioning [73]. Their method consists of two-stages, i.e. structural word recognition and sentence translation. The authors first employ a multi-layer optimization method to generate a hierarchical concepts to represent an image as a tetrad $\langle \text{objects}, \text{attributes}, \text{activities}, \text{scenes} \rangle$. The tetrad plays the role of structural words. Then, they utilize an encoder-decoder machine translation model, which is based on the Long-Short Term Memory neural network, to translate the structural words into sentences.

Oruganti et al. present a fusion based model which consists of an image processing stage, a language processing stage and a fusion stage [74]. In their method, images and languages are independently processed in their corresponding stages based on a Convolutional Neural Network and a Long-Short Term Memory network, respectively. After that, the outputs of these two stages are mapped into a common vector space, where the fusion stage associate these two modalities and make predications. Such a method is argued to be able to make the system more flexible and mitigate the shortcomings of previous approaches on their inability to accommodate disparate inputs.

A parallel-fusion RNN-LSTM architecture is presented in [75] by Wang et al. to take advantages of the complementary properties of simple Recurrent Neural Networks and Long-Short Term Memory networks for improving the performance of image captioning systems. In their method, inputs are mapped to hidden states by Recurrent Neural Network units and Long-Short Term Memory units in parallel. Then, the hidden states in these two networks are merged with certain ratios for word predication.

4.6. Generating descriptions for images with novelties

So far, all of the introduced image captioning methods are limited to pre-specified and fixed word dictionaries and are not enabled to generate descriptions for concepts that are not trained with paired image-sentence training data. Humans have the ability to recognize, learn and use novel concepts in various visual understanding tasks. And in practical image description

applications, it is quite possible to come across situations where there are novel objects which are not in the pre-specified vocabulary or have not been trained with paired image-sentence data. It is undesirable to retrain the whole system every time when a few images with novel concepts appear. Therefore, it is a useful ability for image captioning systems to adapt to novelties appearing in images for generating image descriptions efficiently. In this subsection, we talk about approaches that can deal with novelties in images during image captioning.

In order to learn novel visual concepts without retraining the whole system, Mao et al. propose to use linguistic context and visual features to hypothesize semantic meanings of new words and use these words to describe images with novelties [76]. To accomplish the novelty learning task, the authors build their system by making two modifications to the model proposed in [35]. First, they use a transposed weight sharing strategy to reduce the number of parameters in the model, so that the over fitting problem can be prevented. Second, they use a Long-Short Term Memory (LSTM) layer [107] to replace the recurrent layer to avoid the gradient explosion and vanishing problem.

With the aim of describing novel objects that are not present in the training image-sentence pairs, Hendricks et al. propose the Deep Compositional Captioner method [36]. In this method, large object recognition datasets and external text corpora are leveraged, and novel object description is realised based on knowledges transferred between semantically similar concepts. To achieve this goal, Hendricks et al. first train a lexical classifier and a language model over image datasets and text corpora, respectively. Then, they trained a deep multimodal caption model to integrate the lexical classifier and the language model. Particularly, as a linear combination of affine transformation of image and language features, the caption model enables easy transfer of semantic knowledge between these two modalities, which allows predication of novel objects.

5. State of the art method comparison

5.1. Image captioning evaluation metrics

In this section, we will compare image captioning methods that give state of the art results. Being plagued by the complexity of the outputs, image captioning methods are difficult to evaluate. In order to compare image captioning systems as for their capability to generate human-like sentences with respect to linguistic quality and semantic correctness, various evaluation metrics have been designed. For state of the art method comparison, we need to introduce the commonly used evaluation metrics first.

In fact, the most intuitive way to determine how well a generated sentence describes the content of an image is by direct human judgements. However, because human evaluation requires large amounts of un-reusable human efforts, it is difficult to scale up. Furthermore, human evaluation is inherently subjective making it suffer from user variances. Therefore, in this paper we report method comparison based on automatic image captioning evaluation metrics. The used automatic evaluation metrics include BLEU [123], ROUGE-L [124], METEOR [125] and CIDEr [126]. BLEU, ROUGE-L and METEOR are originally designed to judge the quality of machine translation. Because the evaluation process of image captioning is exactly the same as machine translation, in which generated sentences are compared against ground truth sentences, these metrics are widely used for image captioning evaluation.

BLEU [123] is to use variable lengths of phrases of a candidate sentence to match against reference sentences written by human to measure their closeness. In other words, BLEU metrics are determined by comparing a candidate sentence with reference sentences in n-grams. Specifically, to determine BLEU-1, the candidate

sentence is compared with reference sentences in unigram, while for calculating BLEU-2, bigram will be used for matching. A maximum order of four is empirically determined to obtain the best correlation with human judgements. For BLEU metrics, the unigram scores account for the adequacy, while higher n-gram scores account for the fluency.

ROUGE-L [124] is designed to evaluate the adequacy and fluency of machine translation. This metric employs the longest common subsequence between a candidate sentence and a set of reference sentences to measure their similarity at sentence-level. The longest common subsequence between two sentences only requires in-sequence word matches, and the matched words are not necessarily consecutive. Determination of the longest common subsequence is achieved by using dynamic programming technique. Because this metric automatically includes longest in-sequence common n-grams, sentence level structure can be naturally captured.

METEOR [125] is an automatic machine translation evaluation metric. It first performs generalized unigram matches between a candidate sentence and a human-written reference sentence, then computes a score based on the matching results. The computation involves precision, recall and alignments of the matched words. In the case of multiple reference sentences, the best score among all independently computed ones is taken as the final evaluation result of the candidate. Introduction of this metric is for addressing weakness of the BLEU metric, which is derived only based on the precision of matched n-grams.

CIDEr [126] is a paradigm that uses human consensus to evaluate the quality of image captioning. This metric measures the similarity of a sentence generated by the image captioning method to the majority of ground truth sentences written by human. It achieves this by encoding the frequency of the n-grams in the candidate sentence to appear in the reference sentences, where a Term Frequency Inverse Document Frequency weighting for each n-gram is used. This metric is designed to evaluate generated sentences in aspects of grammaticality, saliency, importance and accuracy.

5.2. Comparison on benchmark datasets

Three benchmark datasets that are widely used to evaluate image captioning methods are employed as the testbed for method comparison. The datasets are Flickr8K [32], Flickr30k [127] and Microsoft COCO Caption dataset [128].

Flickr8K [32] contains 8,000 images extracted from Flickr. The images in this dataset mainly contain human and animals. Each image is annotated by five sentences based on crowdsourcing service from Amazon Mechanical Turk. During image annotation, the Amazon Mechanical Turk workers are instructed to focus on the images and describe their contents without considering the context in which the pictures are taken.

Flickr30k [127] is a dataset that is extended from the Flickr8k dataset. There are 31,783 annotated images in Flickr30k. Each image is associated to five sentences purposely written for it. The images in this dataset are mainly about humans involved in everyday activities and events.

Microsoft COCO Caption dataset [128] is created by gathering images of complex everyday scenes with common objects in their natural context. Currently, there are 123,287 images in total, of which 82,783 and 40,504 are used for training and validation, respectively. For each image in the training and validation set, five human written captions are provided. Captions of test images are unavailable publicly. This dataset poses great challenges to the image captioning task.

The comparison is based on an experiment protocol that is commonly adopted in previous work. For datasets Flickr8k and Flickr30k, 1,000 images are used for validation and testing

Table 2

Method comparison on datasets Flickr8k and Flickr30k. In this table, B-n, MT, RG, CD stand for BLEU-n, METEOR, ROUGE-L and CIDEr, respectively.

Category	Method	Flickr8k							Flickr30k						
		B-1	B-2	B-3	B-4	MT	RG	CD	B-1	B-2	B-3	B-4	MT	RG	CD
Multimodal learning	Karpathy and Fei-Fei [61]	0.579	0.383	0.245	0.160	–	–	–	0.573	0.369	0.240	0.157	–	–	–
	Mao et al. [35]	0.565	0.386	0.256	0.170	–	–	–	0.600	0.410	0.280	0.190	–	–	–
encoder–decoder framework	Kiros et al. [59]	0.656	0.424	0.277	0.177	0.173	–	–	0.600	0.380	0.254	0.171	0.169	–	–
	Donahue et al. [34]	–	–	–	–	–	–	–	0.587	0.391	0.251	0.165	–	–	–
Attention guided	Vinyals et al. [64]	0.630	0.410	0.270	–	–	–	–	0.670	0.450	0.300	–	–	–	–
	Jia et al. [65]	0.647	0.459	0.318	0.216	0.202	–	–	0.646	0.446	0.305	0.206	0.179	–	–
Compositional architectures	You et al. [69]	–	–	–	–	–	–	–	0.647	0.460	0.324	0.230	0.189	–	–
	Xu et al. [68]	0.670	0.457	0.314	0.213	0.203	–	–	0.669	0.439	0.296	0.199	0.185	–	–
	Fu et al. [72]	0.639	0.459	0.319	0.217	0.204	0.470	0.538	0.649	0.462	0.324	0.224	0.194	0.451	0.472

Table 3

Method comparison on Microsoft COCO Caption dataset under the commonly used protocol. In this table, B-n, MT, RG, CD stand for BLEU-n, METEOR, ROUGE-L and CIDEr, respectively.

Category	Method	MSCOCO						
		B-1	B-2	B-3	B-4	MT	RG	CD
Multimodal learning	Karpathy and Fei-Fei [61]	0.625	0.450	0.321	0.230	0.195	–	0.660
	Mao et al. [35]	0.670	0.490	0.350	0.250	–	–	–
encoder–decoder framework	Donahue et al. [34]	0.669	0.489	0.349	0.249	–	–	–
	Jia et al. [65]	0.670	0.491	0.358	0.264	0.227	–	0.813
Attention guided	Vinyals et al. [64]	–	–	–	0.277	0.237	–	0.855
	Wu et al. [66]	0.74	0.56	0.42	0.31	0.26	–	0.94
Compositional architectures	Xu et al. [68]	0.718	0.504	0.357	0.250	0.230	–	–
	You et al. [69]	0.709	0.537	0.402	0.304	0.243	–	–
	Fang et al. [33]	–	–	–	0.257	0.236	–	–
	Fu et al. [72]	0.724	0.555	0.418	0.313	0.248	0.532	0.955

respectively, while all the other images are used for training. For the Microsoft COCO Caption dataset, since the captions of the test set are unavailable, only training and validation sets are used. All images in the training set are used for training, while 5,000 validation images are used for validation, and another 5,000 images from the validation set are used for testing. Under the experiment setting described above, image captioning comparison on datasets Flickr8k and Flickr30k is shown in Table 2, and comparison results on the Microsoft COCO Caption dataset are shown in Table 3.

In the method Karpathy and Li [61], a multimodal Recurrent Neural Network is trained to align image regions and sentence fragments for image captioning. The authors report their results on the benchmark datasets Flickr8k, Flickr30k and Microsoft COCO Caption dataset in Tables 2 and 3, respectively. On Flickr8k, the achieved BLEU-1, BLEU-2, BLEU-3 and BLEU-4 scores are 0.579, 0.383, 0.245 and 0.160, respectively. Similar results are achieved on the Flickr30k dataset, which are 0.573, 0.369, 0.240 and 0.157, respectively. Higher scores are achieved by their method on the Microsoft COCO Caption dataset for all the BLEU-n evaluation metrics. Furthermore, on this dataset, METEOR and CIDEr scores are reported, which are 0.195 and 0.660, respectively.

Another multimodal learning based image captioning method is Mao et al. [35], where a deep Convolutional Neural Network is used to extract visual features from images, and a Recurrent Neural Network with a multimodal part is used to model word distributions conditioned on image features and context words. In their method, words are generated one by one for captioning images. They evaluate their method on all three benchmark datasets, with respect to BLEU-n metrics. Their method outperforms Karpathy and Li [61] on all three benchmarks. The results show that multimodal learning based image captioning method that generates image descriptions word by word can outperform the one using language fragments due to its flexibility.

After the encoder–decoder framework is introduced to solve the image captioning problem, it becomes a popular paradigm, and promising performances are demonstrated. Donahue et al. adopt

a deep Convolutional Neural Network for encoding and a Long Short-Term Memory Recurrent Network for decoding to generate sentence descriptions for input images [34]. In Donahue's method, both image feature and context word feature are provided to the sequential model at each time step. On the Flickr30k dataset, the achieved BLEU-n scores are 0.587, 0.391, 0.251 and 0.165, respectively. On the Microsoft COCO Caption dataset, the achieved BLEU-n scores are 0.669, 0.489, 0.349 and 0.249, respectively. The results are superior to Karpathy and Li [61], but a little bit inferior to Mao et al. [35].

With the same encoder–decoder framework, Vinyals et al. [64] outperform Donahue et al. [34] by feeding image features to the decoder network at only the initial time step. In Vinyals' method, inputs to the decoder at the following time steps are features of previously predicated context words. They report BLUE-1, BLUE-2 and BLUE-3 scores on the Flickr8k and Flickr30k datasets and report BLUE-4, METEOR and CIDEr scores on the MSCOCO dataset. As for the reported results, they outperform multimodal learning based image captioning methods [35], [61] and the other encoder–decoder based method [34]. The results show that compared to multimodal learning based image captioning framework, the encoder–decoder framework is more effective for image captioning.

Following the encoder–decoder paradigm, Jia et al. [65] propose to extract semantic information from images and add the information to each unit of the Long Short-Term Memory Recurrent Neural Network during the process of sentence generation for generating image descriptions that are closely related to image contents. Through this manner, the BLEU-n scores on the Flickr8k dataset are improved to 0.647, 0.459, 0.318 and 0.216, respectively. And the BLEU-n scores on the Flickr30k dataset are improved to 0.646, 0.446, 0.305 and 0.206, respectively. The METEOR scores on the Flickr8k and Flickr30k are 0.202 and 0.179, respectively. Compared to the basic encoder–decoder framework, results achieved by their method are much higher. And scores reported by the authors on the MSCOCO dataset are also competitive with other methods.

Table 4

Automatic metric scores on the MSCOCO test server. In this table, B-n, MT, RG, CD stand for BLEU-n, METEOR, ROUGE-L and CIDEr, respectively.

Category	Method	MSCOCO c5							MSCOCO c40						
		B-1	B-2	B-3	B-4	MT	RG	CD	B-1	B-2	B-3	B-4	MT	RG	CD
Multimodal learning encoder–decoder framework	Mao et al. [35]	0.680	0.506	0.369	0.272	0.225	0.499	0.791	0.865	0.760	0.641	0.529	0.304	0.640	0.789
	Donahue et al. [34]	0.700	0.530	0.380	0.280	0.240	0.520	0.870	0.870	0.770	0.650	0.530	0.320	0.660	0.890
	Vinyals et al. [64]	0.713	0.542	0.407	0.309	0.254	0.530	0.943	0.895	0.802	0.694	0.587	0.346	0.682	0.946
Attention guided	Wu et al. [66]	0.730	0.560	0.410	0.310	0.250	0.530	0.920	0.890	0.800	0.690	0.580	0.330	0.670	0.930
	Xu et al. [68]	0.705	0.528	0.383	0.277	0.241	0.516	0.865	0.881	0.779	0.658	0.537	0.322	0.654	0.893
	You et al. [69]	0.731	0.565	0.424	0.316	0.250	0.535	0.943	0.9	0.815	0.709	0.599	0.335	0.682	0.958
Compositional architectures	Yang et al. [70]	–	–	–	–	–	–	–	–	–	–	0.597	0.347	0.686	0.969
	Fang et al. [33]	0.695	–	–	0.291	0.247	0.519	0.912	0.880	–	–	0.567	0.331	0.662	0.925
	Fu et al. [72]	0.722	0.556	0.418	0.314	0.248	0.530	0.939	0.902	0.817	0.711	0.601	0.336	0.680	0.946

With the encoder–decoder framework, Xu et al. [68] propose to add the attention mechanism to the model, so that the attentive encoder–decoder model is able to dynamically attend salient image regions during the process of image description generation. Xu et al. reported their BLEU-n and METEOR scores on all three benchmark datasets. Their results are comparable to Jia et al. [65].

To take advantages of the complimentary properties of bottom-up and top-down image captioning approaches, You et al. [69] propose a semantic attention model to incorporate cognitive visual cues into the decoder as attention guidance for image captioning. Their method is evaluated on the Flickr30k and MSCOCO dataset, with BLEU-n and METEOR scores reported. The experiment results show that their method can improve the scores further compared to Xu et al. [68] and Jia et al. [65]. The results show that appropriate modifications to the basic encoder–decoder framework by introducing attention mechanism can improve the image captioning performances effectively.

A compositional architecture is used by Fu et al. [72] to integrate independent building blocks for generating captions for input images. In their method, the word generation process is aligned to visual perception of image regions, and scene-specific contexts are introduced to capture high-level semantic information in images for adapting word generation to specific scene types. The authors report their experiment results on all three benchmark datasets with respect to evaluation metrics BLEU-n, METEOR and CIDEr. Most of the reported results can outperform other methods. However, although methods based on compositional architectures can utilize information from different sources and take advantages of strengths of various methods to give better results than most of the other methods, they are usually much more complex and relatively hard to implement.

To ensure consistency in evaluation of image captioning methods, a test server is hosted by the MSCOCO team [128]. For method evaluation, this server allows researchers to forward captions generated by their own models to it for computing several popular metric scores. The computed metric scores include BLEU, METEOR, ROUGE and CIDEr. The evaluation on the server is on the “test 2014” test set of the Microsoft COCO Caption dataset, whose ground truth captions are unavailable publicly. With each image in the test set accompanied by 40 human-written captions, two types of metrics can be computed for caption evaluation, i.e. c5 and c40, which means to compare one caption against 5 reference captions and 40 reference captions for metric score computation, respectively. Evaluation results of previous methods on the test server are summarized in Table 4.

From Table 4, it can be seen that image captioning evaluation metric scores computed based on c40 are higher than the ones computed based on c5. This is because the evaluation metrics are computed based on the consistency between the generated description and the reference descriptions. Therefore, more refer-

ences can usually lead to higher probability of matching, resulting higher metric scores.

From Tables 3 and 4, it can be seen that although image captioning evaluation metric scores computed on the MSCOCO test server are different from the ones computed under the commonly used protocol, the tendency of the performances of the methods are similar. The method Mao et al. [35], which is multimodal learning based, is outperformed by encoder–decoder based image captioning methods Donahue et al. [34] and Vinyals et al. [64]. Although both methods Donahue et al. [34] and Vinyals et al. [64] are based on the encoder–decoder framework, with different decoding mechanisms, like in Tables 2 and 3, Vinyals et al. [64] achieve higher scores than Donahue et al. [34], with respect to all used evaluation metrics.

Incorporating additional information into the encoder–decoder framework can improve the image captioning performance further. For example, by using the attention mechanism, Xu et al. [68] give superior performances to Donahue et al. [34]. By incorporating visual concepts into the encoder–decoder framework, Wu et al. [66] outperform Xu et al. [68]. By using a semantic attention model, You et al. [69] achieve superior performances to nearly all the other methods.

These results show that various kinds of cues from the images can be utilized to improve image captioning performances of the encoder–decoder framework. And effectiveness of different information may be different for improving the image captioning performance. And even with the same structure, when information are fed to the framework in different ways, quite different results may be achieved.

On MSCOCO test server image captioning methods based on compositional architectures can usually give relatively good results. Fu et al. [72], which is a compositional architecture, achieve image captioning scores comparable to You et al. [69], and another compositional method Fang et al. [33] can also outperform multimodal based method Mao et al. [35] and encoder–decoder based method Donahue et al. [34] and Xu et al. [68].

In summary, from Table 4, it can be observed that when using the MSCOCO test server for image captioning method evaluation, image captioning methods based on the encoder–decoder framework [34], [64] outperform the multimodal learning image captioning method [35], noticeably. When semantic information or attention mechanisms are used [66], [69], the performance can be improved further. Currently, the best results on the MSCOCO test server are achieved by image captioning methods that utilize attention mechanisms to augment the encoder–decoder framework [69] [70], which outperform the compositional method [72] slightly (Accessed in March, 2017).

Finally, in Fig. 5 we show examples of image captioning results obtained based on different approaches to give readers a straightforward impression for different kinds of image caption methods.






					
Karpathy and Fei-fei [61]:	A pan filled with broccoli and meat	A street sign on the side of the road	A group of people standing on top of a snow covered slope	A baseball player pitching a ball on top of a field	A group of cows standing in a field
Vinyals et al.[64]:	A pan filled with broccoli and meat cooking	A stop sign on the side of the road	A group of people standing on top of a snow covered slope	A stop sign on the side of a road	A herd of cattle standing on top of a lush green field
Xu et al. [68]:	A pan filled with broccoli and meat on a stove	A stop sign on a road with trees	A group of people sitting on a ski lift on a snow covered slope	A baseball player throwing a ball in a green field	A herd of cattle standing in a grass covered field
Fang et al. [33]:	A pot of broccoli on a stove	A yellow sign on a dirt road	A group of people posing for a picture on a ski lift	A baseball player throwing a ball	A herd of cattle standing on top of a grass covered field
Human caption:	A wok with a cooked broccoli meal in it	A floodway sign sitting on the side of a road in a field	Three people wearing ski gear sitting on a ski lift	A pitcher holds his arm far behind him during a pitch	Several cows are gathered together in a grassy field

Fig. 5. Examples of image captioning results obtained based on different methods.

6. Future research directions

Automatic image captioning is a relatively new task, thanks to the efforts made by researchers in this field, great progress has been made. In our opinion there is still much room to improve the performance of image captioning. First, with the fast development of deep neural networks, employing more powerful network structures as language models and/or visual models will undoubtedly improve the performance of image description generation. Second, because images are consisted of objects distributed in space, while image captions are sequences of words, investigation on presence and order of visual concepts in image captions are important for image captioning. Furthermore, since this problem fits well with the attention mechanism and attention mechanism is suggested to run the range of AI-related tasks [129], how to utilize attention mechanism to generate image captions effectively will continue to be an important research topic. Third, due to the lack of paired image-sentence training set, research on utilizing unsupervised data, either from images alone or text alone, to improve image captioning will be promising. Fourth, current approaches mainly focus on generating captions that are general about image contents. However, as pointed by Johnson et al. [130], to describe images at a human level and to be applicable in real-life environments, image description should be well grounded by the elements of the images. Therefore, image captioning grounded by image regions will be one of the future research directions. Fifth, so far, most of previous methods are designed to image captioning for generic cases, while task-specific image captioning is needed in certain cases. Research on solving image captioning problems in various special cases will also be interesting.

7. Conclusion

In this paper, we present a survey on image captioning. Based on the technique adopted in each method, we classify image captioning approaches into different categories. Representative methods in each category are summarized, and strengths and limitations of each type of work are talked about. We first discuss early image captioning work which are mainly retrieval based and template based. Then, our main attention is focused on neural network based methods, which give state of the art results. Because different frameworks are used in neural network based methods, we further divided them into subcategories and discussed each subcategory, respectively. After that, state of the art methods are compared on benchmark datasets. Finally, we present a discussion on future research directions of automatic image captioning.

Acknowledgments

This work was supported by National Natural Science Foundation of China (61602027).

References

- [1] L. Fei-Fei, A. Iyer, C. Koch, P. Perona., What do we perceive in a glance of a real-world scene? *J. Vis.* 7 (1) (2007) 1–29.
- [2] P.F. Felzenszwalb, R.B. Girshick, D. McAllester, D. Ramanan, Object detection with discriminatively trained part based models, *IEEE Trans. Pattern Anal. Mach. Intell.* 32 (9) (2010) 1627–1645.
- [3] R. Girshick, J. Donahue, T. Darrell, J. Malik, Rich feature hierarchies for accurate object detection and semantic segmentation, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Columbus, OH, USA, 2014, pp. 580–587.
- [4] C.H. Lampert, H. Nickisch, S. Harmeling, Learning to detect unseen object classes by between class attribute transfer, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Miami, FL, USA, 2009, pp. 951–958.
- [5] C. Gan, T. Yang, B. Gong, Learning attributes equals multi-source domain generalization, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Miami, FL, USA, 2016, pp. 87–97.
- [6] L. Bourdev, J. Malik, S. Maji, Action recognition from a distributed representation of pose and appearance, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Providence, RI, 2011, pp. 3177–3184.
- [7] Y.-W. Chao, Z. Wang, R. Mihalcea, J. Deng, Mining semantic affordances of visual object categories, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Boston, MA, USA, 2015, pp. 4259–4267.
- [8] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, in: *Proceedings of the Twenty Fifth International Conference on Neural Information Processing Systems*, 2012, pp. 1097–1105.
- [9] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, A. Oliva, Learning deep features for scene recognition using places database, in: *Proceedings of the Advances in Neural Information Processing Systems (NIPS)*, 2014, pp. 487–495.
- [10] Y. Gong, L. Wang, R. Guo, S. Lazebnik, Multi-scale orderless pooling of deep convolutional activation features, in: *Proceedings of the European Conference on Computer Vision*, 2014, pp. 392–407.
- [11] A. Kojima, T. Tamura, K. Fukunaga, Natural language description of human activities from video images based on concept hierarchy of actions, *Int. Comput. Vis.* 50 (2002) 171–184.
- [12] P. Hede, P. Moellic, J. Bourgeois, M. Joint, C. Thomas, Automatic generation of natural language descriptions for images, in: *Proceedings of the Recherche Dinformatique Assistée Par Ordinateur*, 2004.
- [13] A. Farhadi, M. Hejrati, M.A. Sadeghi, P. Young, C. Rashtchian, J. Hockenmaier, D. Forsyth, Every picture tells a story: Generating sentences from images, in: *Proceedings of the European Conference on Computer Vision*, 2010, pp. 15–29.
- [14] Y. Yang, C.L. Teo, H. Daume, Y. Aloimono, Corpus-guided sentence generation of natural images, in: *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2011, pp. 444–454.
- [15] V. Ordonez, G. Kulkarni, T.L. Berg., Im2Text: describing images using 1 million captioned photographs, in: *Proceedings of the Advances in Neural Information Processing Systems*, 2011, pp. 1143–1151.
- [16] A. Gupta, Y. Verma, C.V. Jawahar., Choosing linguistics over vision to describe images, in: *Proceedings of the AAAI Conference on Artificial Intelligence*, 5, 2012.

- [17] H. Goh, N. Thome, M. Cord, J. Lim, Learning deep hierarchical visual feature coding, *IEEE Trans. Neural Netw. Learn. Syst.* 25 (12) (2014) 2212–2225.
- [18] Y. Bengio, A. Courville, P. Vincent, Representation learning: a review and new perspectives, *IEEE Trans. Pattern Anal. Mach. Intell.* 35 (8) (2013) 1798–1828.
- [19] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, T. Darrell, DeCAF: a deep convolutional activation feature for generic visual recognition, in: *Proceedings of The Thirty First International Conference on Machine Learning*, 2014, pp. 647–655.
- [20] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, T. Darrell, Caffe: convolutional architecture for fast feature embedding, *arXiv:1408.5093v1* (2014).
- [21] N. Zhang, S. Ding, J. Zhang, Y. Xue, Research on point-wise gated deep networks, *Appl. Soft Comput.* 52 (2017) 1210–1221.
- [22] J.P. Papa, W. Scheirer, D.D. Cox, Fine-tuning deep belief networks using harmony search, *Appl. Soft Comput.* 46 (2016) 875–885.
- [23] C. Farabet, C. Couprie, L. Najman, Y. LeCun, Learning hierarchical features for scene labeling, *IEEE Trans. Pattern Anal. Mach. Intell.* 35(8).
- [24] E.P. Iijina, C.K. Mohan, Hybrid deep neural network model for human action recognition, *Appl. Soft Comput.* 46 (2016) 936–952.
- [25] S. Wang, Y. Jiang, F.-L. Chung, P. Qian, Feedforward kernel neural networks, generalized least learning machine, and its deep learning with application to image classification, *Appl. Soft Comput.* 37 (2015) 125–141.
- [26] S. Bai, Growing random forest on deep convolutional neural networks for scene categorization, *Expert Syst. Appl.* 71 (2017) 279–287.
- [27] D. Bahdanau, K. Cho, Y. Bengio, Neural machine translation by jointly learning to align and translate, *arXiv:1409.0473v7* (2016).
- [28] K. Cho, B.V. Merriboer, C. Gulcehre, Learning phrase representations using RNN encoder–decoder for statistical machine translation, *arXiv:1406.1078v3* (2014).
- [29] R. Collobert, J. Weston, A unified architecture for natural language processing: deep neural networks with multitask learning, in: *Proceedings of the Twenty Fifth International Conference on Machine Learning*, 2008, pp. 160–167.
- [30] A. Mnih, G. Hinton, Three new graphical models for statistical language modelling, in: *Proceedings of the Twenty Fourth International Conference on Machine Learning*, 2007, pp. 641–648.
- [31] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, J. Dean, Distributed representations of words and phrases and their compositionality, in: *Proceedings of the Advances in Neural Information Processing Systems*, 2013.
- [32] M. Hodosh, P. Young, J. Hockenmaier, Framing image description as a ranking task: data, models and evaluation metrics, *J. Artif. Intell. Res.* 47 (2013) 853–899.
- [33] H. Fang, S. Gupta, F. Iandola, R. Srivastava, From captions to visual concepts and back, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1473–1482.
- [34] J. Donahue, L. Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, Long-term recurrent convolutional networks for visual recognition and description, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 2625–2634.
- [35] J. Mao, W. Xu, Y. Yang, J. Wang, Z. Huang, A. Yuille, Deep captioning with multimodal recurrent neural networks, in: *Proceedings of the International Conference on Learning Representation*, 2015.
- [36] M.R.R.M.S. L A Hendricks, S. Venugopalan, Deep compositional captioning: describing novel object categories without paired training data, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 1–10.
- [37] A. Karpathy, A. Joulin, F. Li, Deep fragment embeddings for bidirectional image sentence mapping, in: *Proceedings of the Twenty Seventh Advances in Neural Information Processing Systems (NIPS)*, 3, 2014, pp. 1889–1897.
- [38] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C.L. Zitnick, D. Parikh, VQA: Visual question answering, *arXiv:1505.00468v7* (2016).
- [39] M. Malinowski, M. Fritz, A multi-world approach to question answering about real-world scenes based on uncertain input, in: *Proceedings of the Advances in Neural Information Processing Systems*, pp. 1682–1690.
- [40] M. Malinowski, M. Rohrbach, M. Fritz, Ask your neurons: a neural-based approach to answering questions about images, in: *Proceedings of the International Conference on Computer Vision*, 2015.
- [41] H. Gao, J. Mao, J. Zhou, Z. Huang, L. Wang, W. Xu., Are you talking to a machine? Dataset and methods for multilingual image question answering, in: *Proceedings of the Advances in Neural Information Processing Systems*, pp. 2296–2304.
- [42] D. Geman, S. Geman, N. Hallonquist, L. Younes, Visual Turing test for computer vision systems, in: *Proceedings of the National Academy of Sciences of the United States of America*, vol. 112, pp. 3618–3623.
- [43] Y. Feng, M. Lapata, Automatic caption generation for news images, *IEEE Trans. Pattern Anal. Mach. Intell.* 35(4).
- [44] A. Tariq, H. Foroosh, A context-driven extractive framework for generating realistic image descriptions, *IEEE Trans. Image Process.* 26(2).
- [45] S. Guadarrama, N. Krishnamoorthy, G. Malkarnkar, S. Venugopalan, R. Mooney, T. Darrell, K. Saenko, YouTube2text: recognizing and describing arbitrary activities using semantic hierarchies and zero-shot recognition, in: *Proceedings of the International Conference on Computer Vision*, pp. 2712–2719.
- [46] J. Thomason, S. Venugopalan, S. Guadarrama, K. Saenko, R. Mooney, Integrating language and vision to generate natural language descriptions of videos in the wild, in: *Proceedings of the International Conference on Computational Linguistics*, 2014.
- [47] S. Venugopalan, M. Rohrbach, J. Donahue, R. Mooney, T. Darrell, K. Saenko, Sequence to sequence – video to text, in: *Proceedings of the International Conference on Computer Vision*, 2015.
- [48] S. Venugopalan, L. Hendricks, R. Mooney, K. Saenko, Improving LSTM-based video description with linguistic knowledge mined from text, in: *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2016.
- [49] R. Mason, E. Charniak, Nonparametric method for data driven image captioning, in: *Proceedings of the Fifty Second Annual Meeting of the Association for Computational Linguistics*, 2014.
- [50] P. Kuznetsova, V. Ordonez, T. Berg, Y. Choi, TREETALK: composition and compression of trees for image descriptions, *Trans. Assoc. Comput. Linguist.* 2 (10) (2014) 351–362.
- [51] G. Kulkarni, V. Premraj, V. Ordonez, S. Dhar, S. Li, Y. Choi, A.C. Berg, T.L. Berg, BabyTalk: understanding and generating simple image descriptions, *IEEE Trans. Pattern Anal. Mach. Intell.* 35 (12) (2013) 2891–2903.
- [52] S. Li, G. Kulkarni, T.L. Berg, A.C. Berg, Y. Choi, Composing simple image descriptions using web-scale n-grams, in: *Proceedings of the Fifteenth Conference on Computational Natural Language Learning*, 2011.
- [53] M. Mitchell, J. Dodge, A. Goyal, K. Yamaguchi, K. Stratos, X. Han, A. Mensch, A. Berg, T. Berg, H. Daume, Midge: Generating image descriptions from computer vision detections, in: *Proceedings of the Thirteenth Conference of the European Chapter of the Association for Computational Linguistics*, 2012.
- [54] Y. Ushiku, M. Yamaguchi, Y. Mukuta, T. Harada, Common subspace for model and similarity: phrase learning for caption generation from images, in: *IEEE International Conference on Computer Vision*, 2015, pp. 2668–2676.
- [55] R. Socher, A. Karpathy, Q.V. Le, C.D. Manning, A.Y. Ng, Grounded compositional semantics for finding and describing images with sentences, *TACL* 2 (2014) 207–218.
- [56] L. Ma, Z. Lu, Lifeng, S.H. Li, Multimodal convolutional neural networks for matching image and sentences, in: *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 2623–2631.
- [57] F. Yan, K. Mikolajczyk, Deep correlation for matching images and text, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3441–3450.
- [58] R. Lebre, P.O. Pinheiro, R. Collobert, Phrase-based image captioning, in: *Proceedings of the International Conference on Machine Learning*, 2015.
- [59] R. Kiros, R. Zemel, R. Salakhutdinov, Multimodal neural language models, in: *Proceedings of the International Conference on Machine Learning*, 2014.
- [60] J. Mao, W. Xu, Y. Yang, J. Wang, A.L. Yuille, Explain images with multimodal recurrent neural networks, *arXiv:1410.1090v1* (2014).
- [61] A. Karpathy, F. Li, Deep visual-semantic alignments for generating image descriptions, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3128–3137.
- [62] X. Chen, C. Zitnick, Mind's eye: a recurrent visual representation for image caption generation, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 2422–2431.
- [63] R. Kiros, R. Salakhutdinov, R. Zemel, Unifying visual-semantic embeddings with multimodal neural language models, *arXiv:1411.2539*(2018).
- [64] O. Vinyals, A. Toshev, S. Bengio, D. Erhan, Show and tell: a neural image caption generator, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3156–3164.
- [65] X. Jia, E. Gavves, B. Fernando, T. Tuytelaars, Guiding the long-short term memory model for image caption generation, in: *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 2407–2415.
- [66] Q. Wu, C. Shen, L. Liu, A. Dick, A. van den Hengel, 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*, 2016, pp. 203–212.
- [67] Y. Pu, Z. Gan, R. Henao, X. Yuan, C. Li, A. Stevens, L. Carin, Variational autoencoder for deep learning of images, labels and captions, in: *Proceedings of the Advances in Neural Information Processing Systems*, 2016.
- [68] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhutdinov, R. Zemel, Y. Bengio, Show, attend and tell: neural image caption generation with visual attention, *arXiv:1502.03044v3* (2016).
- [69] Q. You, H. Jin, Z. Wang, C. Fang, J. Luo, Image captioning with semantic attention, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 4651–4659.
- [70] Z. Yang, Y. Yuan, Y. Wu, R. Salakhutdinov, W.W. Cohen, Review networks for caption generation, in: *Proceedings of the Advances in Neural Information Processing Systems*, 2016, pp. 2361–2369.
- [71] K. Tran, X. He, L. Zhang, J. Sun, Rich image captioning in the wild, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 434–441.
- [72] K. Fu, J. Jin, R. Cui, F. Sha, C. Zhang, Aligning where to see and what to tell: image captioning with region-based attention and scene-specific contexts, *IEEE Trans. Pattern Anal. Mach. Intell.* (2016).
- [73] S. Ma, Y. Han, Describing images by feeding LSTM with structural words, in: *Proceedings of the IEEE International Conference on Multimedia and Expo*, 2016, pp. 1–6.
- [74] R. Oruganti, S. Sah, S. Pillai, R. Ptucha, Image description through fusion based recurrent multi-modal learning, in: *Proceedings of the IEEE International Conference on Image Processing*, 2016, pp. 3613–3617.
- [75] M. Wang, L. Song, X. Yang, C. Luo, A parallel-fusion RNN-LSTM architecture for image caption generation, in: *Proceedings of the IEEE International Conference on Image Processing*, 2016.

- [76] J. Mao, X. Wei, Y. Yang, J. Wang, Learning like a child: fast novel visual concept learning from sentence descriptions of images, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 2533–2541.
- [77] D. Lin, An information-theoretic definition of similarity, in: Proceedings of the Fifteenth International Conference on Machine Learning, pp. 296–304.
- [78] J. Curran, S. Clark, J. Bos, Linguistically motivated large-scale NLP with CC and boxer, in: Proceedings of the Forty Fifth Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions, pp. 33–36.
- [79] F.R. Bach, M.I. Jordan, Kernel independent component analysis, *J. Mach. Learn. Res.* 3 (2002) 1–48.
- [80] D.R. Hardoon, S.R. Szedmak, J.R. Shawe-Taylor, Canonical correlation analysis: an overview with application to learning methods, *Neural Comput.* 16 (2004) 2639–2664.
- [81] D. Roth, W. tau Yih, A linear programming formulation for global inference in natural language tasks, in: Proceedings of the Annual Conference on Computational Natural Language Learning, 2004.
- [82] J. Clarke, M. Lapata, Global inference for sentence compression an integer linear programming approach, *J. Artif. Intell. Res.* 31 (2008) 339–429.
- [83] P. Kuznetsova, V. Ordonez, A.C. Berg, T.L. Berg, Y. Choi, Collective generation of natural image descriptions, in: Proceedings of the Meeting of the Association for Computational Linguistics, 2012.
- [84] Y. Ushiku, T. Harada, Y. Kuniyoshi, Efficient image annotation for automatic sentence generation, in: Proceedings of the Twentieth ACM International Conference on Multimedia, 2012.
- [85] A. Oliva, A. Torralba, Modeling the shape of the scene: a holistic representation of the spatial envelope, *Int. J. Comput. Vis.* 42 (3) (2001) 145–175.
- [86] T. Dunning, Accurate methods for the statistics of surprise and coincidence, *Comput. Linguist.* 19 (1) (1993) 61–74.
- [87] G. Kulkarni, V. Premraj, S. Dhar, S. Li, Y. Choi, A.C. Berg, T.L. Berg, Baby talk: understanding and generating simple image descriptions, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2011.
- [88] P. Koehn, Europarl: a parallel corpus for statistical machine translation, in: MT Summit, 2005.
- [89] A. Farhadi, M.A. Sadeghi, Phrasal recognition, *IEEE Trans. Pattern Anal. Mach. Intell.* 35 (12) (2013) 2854–2865.
- [90] Y. LeCun, Y. Bengio, G. Hinton, Deep learning, *Nature* 521 (7553) (2015) 436–444.
- [91] A. Frome, G.S. Corrado, J. Shlens, S. Bengio, J. Dean, T. Mikolov, Devise: a deep visual-semantic embedding model, in: Proceedings of the Twenty Sixth International Conference on Neural Information Processing Systems, 2013, pp. 2121–2129.
- [92] Q.V. Le, M. Ranzato, R. Monga, M. Devin, K. Chen, G. Corrado, J. Dean, A.Y. Ng, Building high-level features using large scale unsupervised learning, in: Proceedings of the International Conference on Machine Learning, 2012.
- [93] M. Marneffe, B. MacCartney, C. Manning, Generating typed dependency parses from phrase structure parses, in: Proceedings of the LREC, 2006, pp. 449–454.
- [94] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, arXiv:1409.1556v6 (2015).
- [95] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, Going deeper with convolutions, arXiv:1409.4842 (2018).
- [96] B. Hu, Z. Lu, H. Li, Q. Chen, Convolutional neural network architectures for matching natural language sentences, in: Proceedings of the Twenty Seventh International Conference on Neural Information Processing Systems, 2014, pp. 2042–2050.
- [97] N. Kalchbrenner, E. Grefenstette, P. Blunsom, A convolutional neural network for modelling sentences, arXiv:1404.2188v1 (2014).
- [98] G. Andrew, R. Arora, J. Bilmes, K. Livescu, Deep canonical correlation analysis, in: Proceedings of the International Conference on Machine Learning, 2013, pp. 1247–1255.
- [99] A. Mnih, K. Kavukcuoglu, Learning word embeddings efficiently with noise-contrastive estimation, in: Proceedings of the Advances in Neural Information Processing Systems, 2013, pp. 2265–2273.
- [100] T. Mikolov, K. Chen, G. Corrado, J. Dean, Efficient estimation of word representations in vector space, arXiv:1301.3781v3 (2013).
- [101] J.L. Elman, Finding structure in time, *Cognit. Sci.* 14 (2) (1990) 179–211.
- [102] M. Schuster, K. Paliwal, Bidirectional recurrent neural networks, *IEEE Trans. Signal Process.* 45 (11) (1997) 2673–2681.
- [103] Y. Bengio, P. Simard, P. Frasconi, Learning long-term dependencies with gradient descent is difficult, *IEEE Trans. Neural Netw.* 5(5).
- [104] T. Mikolov, M. Karafiat, L. Burget, J. Cernocky, S. Khudanpur, Recurrent neural network based language model, in: Proceedings of the Conference of the International Speech Communication Association, 2010, pp. 1045–1048.
- [105] N. Kalchbrenner, P. Blunsom, Recurrent continuous translation models, in: Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2013.
- [106] I. Sutskever, O. Vinyals, Q.V. Le, Sequence to sequence learning with neural networks, in: Proceedings of the Advances in Neural Information Processing Systems, 2014.
- [107] S. Hochreiter, J. Schmidhuber, Long short-term memory, *Neural Comput.* 9 (8) (1997) 1735–1780.
- [108] O. Vinyals, A. Toshev, S. Bengio, D. Erhan, Show and tell: lessons learned from the 2015 MSCOCO image captioning challenge, *IEEE Trans. Pattern Anal. Mach. Intell.* 39(4).
- [109] K. Greff, R.K. Srivastava, J. Koutník, B.R. Steunebrink, J. Schmidhuber, LSTM: a search space odyssey, arXiv:1503.04069v2 (2017).
- [110] Y. Wei, W. Xia, J. Huang, B. Ni, J. Dong, Y. Zhao, S. Yan, CNN: single-label to multi-label, arXiv:1406.5726v3 (2014) 1–14.
- [111] R. A. Rensink, The dynamic representation of scenes, *Vis. Cognit.* 7 (1) (2000) 17–42.
- [112] M. Spratlting, M.H. Johnson, A feedback model of visual attention, *J. Cognit. Neurosci.* 16 (2) (2004) 219–237.
- [113] D. Bahdanau, K. Cho, Y. Bengio, Neural machine translation by jointly learning to align and translate, arXiv:1409.0473v7 (2017).
- [114] J. Ba, V. Mnih, K. Kavukcuoglu, Multiple object recognition with visual attention, in: Proceedings of the International Conference on Learning Representation, 2015.
- [115] V. Mnih, N. Hees, A. Graves, K. Kavukcuoglu, Recurrent models of visual attention, in: Proceedings of the Advances in Neural Information Processing Systems, 2014.
- [116] D. Elliott, F. Keller, Image description using visual dependency representations, in: Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2013, pp. 1292–1302.
- [117] C. Zhang, J.C. Platt, P.A. Viola, Multiple instance boosting for object detection, in: Proceedings of the Advances in Neural Information Processing Systems, 2005, pp. 1419–1426.
- [118] A.L. Berger, S.A.D. Pietra, V.J.D. Pietra, A maximum entropy approach to natural language processing, *Comput. Linguist.* 22 (1) (1996) 39–71.
- [119] A. Ratnaparkhi, Trainable methods for surface natural language generation, in: Proceedings of the North American chapter of the Association for Computational Linguistics conference, 2000, pp. 194–201.
- [120] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.
- [121] J.R. Uijlings, K.E. van de Sande, T. Gevers, A.W. Smeulders, Selective search for object recognition, *Int. J. Comput. Vis.* 104 (2) (2013) 154–171.
- [122] D.M. Blei, A.Y. Ng, M.I. Jordan, Latent Dirichlet allocation, *J. Mach. Learn. Res.* 3 (2003) 993–1022.
- [123] K. Papineni, S. Roukos, T. Ward, W. Zhu, BLEU: a method for automatic evaluation of machine translation, in: Proceedings of the Meeting on Association for Computational Linguistics, vol. 4 (2002).
- [124] C.-Y. Lin, F.J. Och, Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics, in: Proceedings of the Meeting on Association for Computational Linguistics, 2004.
- [125] A. Lavie, A. Agarwal, METEOR: an automatic metric for MT evaluation with improved correlation with human judgments, in: Proceedings of the Second Workshop on Statistical Machine Translation, 2007, pp. 228–231.
- [126] R. Vedantam, C.L. Zitnick, D. Parikh, CIDEr: consensus-based image description evaluation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 4566–4575.
- [127] P. Young, A. Lai, M. Hodosh, J. Hockenmaier, From image descriptions to visual denotations: new similarity metrics for semantic inference over event descriptions, in: Proceedings of the Meeting on Association for Computational Linguistics, 2014, pp. 67–78.
- [128] X. Chen, H. Fang, T. Lin, R. Vedantam, S. Gupta, P. Dollar, C. Zitnick, Microsoft COCO captions: data collection and evaluation server, arXiv:1504.00325v2 (2015).
- [129] K. Cho, A. Courville, Y. Bengio, Describing multimedia content using attention-based encoder-decoder networks, *IEEE Trans. Multimed.* 17 (11) (2015) 1875–1886.
- [130] J. Johnson, A. Karpathy, L. Fei-Fei, DenseCap: fully convolutional localization networks for dense captioning, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 4565–4574.



Shuang Bai received the degrees of B.Eng. and M.Eng. from the School of Electrical Engineering and Automation of Tianjin University, Tianjin, China in 2007 and 2009, respectively. In 2013, he received the degree of D.Eng. in the Graduate School of Information Science of Nagoya University. Currently, he is an associate professor in the School of Electronic and Information Engineering of Beijing Jiaotong University, Beijing, China. His research interests include machine learning and computer vision.



Shan An received the degree of B.Eng. from the school of Electrical Engineering and Automation of Tianjin University, China in 2007 and received the degree of M.Eng. from the school of Control Science and Engineering of Shandong University, China, in 2010. Currently, he is a senior algorithm engineer in JD.COM. Before joining JD.COM, he worked for China Academy of Space Technology and Alibaba.com. His research interests include machine learning and computer vision.