


REVIEW

A thorough review of models, evaluation metrics, and datasets on image captioning

Gaifang Luo  | Lijun Cheng | Chao Jing | Can Zhao | Guozhu Song

School of Software, Shanxi Agricultural University,
Jinzhou, China

Correspondence

Guozhu Song, School of Software, Shanxi Agricultural University, Jinzhou 030801, China.
Email: songgz@sxau.edu.cn

Funding information

Higher Agricultural College Branch of the China Educational Technology Association, Grant/Award Number: C21ZD05

Abstract

Image captioning means generate descriptive sentences from a query image automatically. It has recently received widespread attention from the computer vision and natural language processing communities as an emerging visual task. Currently, both components have evolved considerably by exploiting object regions, attributes, attention mechanism methods, entity recognition with novelties, and training strategies. However, despite the impressive results, the research has not yet come to a conclusive answer. This survey aims to provide a comprehensive overview of image captioning methods, from technical architectures to benchmark datasets, evaluation metrics, and comparison of state-of-the-art methods. In particular, image captioning methods are divided into different categories based on the technique adopted. Representative methods in each class are summarized, and their advantages and limitations are discussed. Moreover, many related state-of-the-art studies were quantitatively compared to determine the recent trends and future directions in image captioning. The ultimate goal of this work is to serve as a tool for understanding the existing literature and highlighting future directions in the area of image captioning for Computer Vision and Natural Language Processing communities may benefit from.

1 | INTRODUCTION

It is not difficult to quickly recognize and understand an image by captured visual content. However, letting the computer dig out the helpful information from images and playing its tremendous value for us is still a problem that needs to be solved urgently. For a long time, researchers have tried to perceive and understand the high-level semantic information of images, such as scenes, objects, and relationships, through low-level visual features such as colour, texture, and shape. Unfortunately, computers cannot generate high-level semantic features through low-level visual features as humans do, making a “semantic gap” between image content and image understanding [1,2].

Image captioning means that given an image, the machine perceives its content and generates descriptions automatically. In the early days of the development of computer vision, researchers tried to use computers to simulate the human visual system and let the computer tell people what it saw. After that, researchers put forward higher requirements: let the computer recognize the objects in the image, determine the target

attributes, and even determine the relationship between the recognized entities in the form of natural language to describe image [3]. So far, there have been many related methods of captioning, and still a continuous improvement.

Figure 1 gives an overview of automatic image captioning tasks and a simple example of the most relevant approaches. The purpose of these studies is to find an effective pipeline to process the query image, represent its content, and transform it into a sequence of words by generating connections between visual and textual elements while maintaining the fluency of the language. In its standard configuration, image captioning is an image-to-sequence issue. These images are coded into one or more feature vectors in the visual coding step, and the input is prepared for the second decoder generation step, called a language model. A sequence of words or sub-words decoded from a given vocabulary through a decoder.

Visual understanding is an essential part of artificial intelligence. As one of the tasks of visual understanding, image caption generation has received extensive attention. Still need researchers to invest more energy to make it prosperous. At first, researchers adopted search-based methods and

This is an open access article under the terms of the [Creative Commons Attribution-NonCommercial-NoDerivs](https://creativecommons.org/licenses/by-nc-nd/4.0/) License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.

© 2021 The Authors. *IET Image Processing* published by John Wiley & Sons Ltd on behalf of The Institution of Engineering and Technology

language template-based methods to let computers generate image description sentences but could not achieve satisfactory performance. With the continuous development of deep learning technology, researchers use neural networks to directly learn the mapping of images to describe sentences from a large amount of data. Its performance is far better than previous methods. Moreover, several domain-specific scenarios and variants of this task have been studied. Besides, the computer vision and natural language processing (NLP) communities have solved the challenge of constructing appropriate evaluation metrics to compare the results with human-generated ground truths. Nevertheless, the achieved results are still far from our goal.

With the recent surge of research interest in image captioning, a large number of studies on image captioning review have been proposed. It is noticed that they prefer to focus on specific aspects of this emerging vision to language tasks, such as the technical framework, evaluation indicators, training strategies, or publicly available datasets. However, the existing studies on the review of image captioning have been considered slightly out of vogue or fail to provide a comprehensive overview of the current research, including technologies, benchmark datasets, and evaluation metrics [3,4,120–122]. There is still a lack of literature that comprehensively reviews the research status, innovative technologies, and development prospects. Intending to give a testament to the journey that captioning has taken so far and to encourage novel ideas, in this paper, we provide a holistic overview of the models developed in the last years.

Following the two inherent stages of the captioning model, we developed a taxonomy of visual encoding and language modelling methods, focusing on their key aspects and limitations. We focus on the technical frameworks adopted in the literature over the past few years, from search-based approaches to language model-based approaches and the latest developments obtained through neural networks. Furthermore, we review the primary datasets used to explore image captions, from domain-specific benchmarks to domain-specific datasets collected to investigate specific aspects of the problem. Also, we analyzed the standard metrics employed for model performance evaluation.

Another contribution of this study is to quantitatively compare the main image captioning methods considering standard metrics, and discuss the strengths and weaknesses of various techniques, thereby clarifying the performance, differences and characteristics of the most critical models. Finally, we outlined the recent research trends of image captioning and discussed some open challenges and future directions.

This paper is organized as follows. We first review search-based and Language template-based image captioning methods in Section 2. Then we review the recent deep learning methods in Section 3. In this section, we divide them into sub-categories and discuss representative strategies in each sub-category, respectively. State-of-art methods are compared on benchmark datasets in Section 4. After that, we summarize the latest research trends and envision future research directions of image captioning in Section 5. The conclusion is given in Section 6.

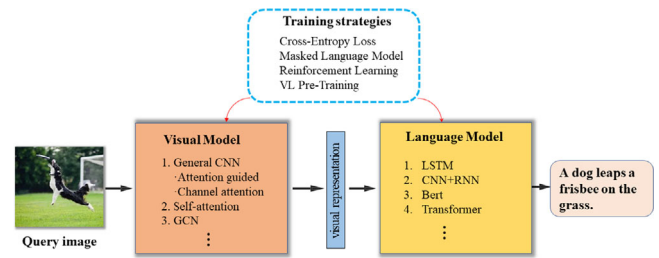


FIGURE 1 Overview of automatic image captioning tasks and a simple example of the most relevant approaches

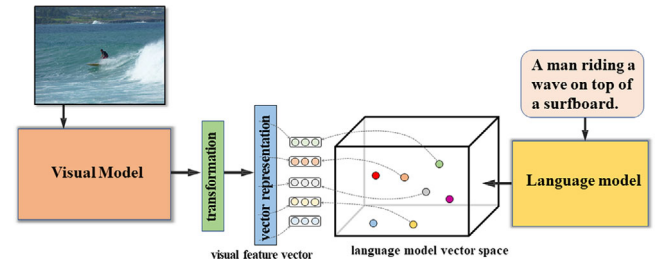


FIGURE 2 Search-based approach and language template-based approach research framework

2 | THE EARLIER IMAGE CAPTION METHODS

This section introduces the two earlier methods: the search-based method and the language template-based method. The research framework of the two types of methods is shown in Figure 2. Both of them first extract image visual features, such as objects, actions, relationships, scenes, and so forth, across visual models, and then transform them into vector representations. The difference is their subsequent steps. The search-based method is to search for similar images through key information of the image and use the corresponding description as the final description result. While the language template-based method maps the image description and the vectorized image content to the same metric space through the language template and selects the result with the highest similarity as its final description.

2.1 | Search-based approaches

Search-based image captioning approaches usually construct an image and the corresponding caption into an “image-caption” dataset firstly. When performing an image caption task, compare the query image with the image in the training set to search for similar images in the training set. Then, the captions of similar images in the dataset are marked as candidate captions set, which are usually sentences or phrases. Finally, the final image description is determined by re-ranking all candidate descriptions.

Ordóñez et al. [5] grabbed a large number of images from the Internet and manually labelled the title and caption of images. When performing the description task, they calculated

the global similarity (scene) between the image to be described and the image in the network image library. The most similar images are found, and their corresponding captions are used as the final result. Hodosh et al. [6] also regard the image caption task as a sorting task. Still, the difference is that the nuclear category correlation analysis technology [7,8] is used to project the image and the attribute items extracted from the letters into a public space. Through training, the image and its corresponding caption have the most remarkable correlation. Then put the undescribed image in this public space, and select the caption with the highest ranking as the final description by calculating their cosine similarity between all captions. Mason and Charniak [9] took the lead in considering the effect of noise on the method. They use visual similarity to search for images from the dataset similar to undescribed images and then obtain captions corresponding to these images. The image captions are ranked by calculating the probability density of words, and finally, the image description with the highest ranking is selected as a result.

The above methods are all trying to find the description sentence that best matches the query image from the existing image description in the data set. The rationality of this type of method must follow a premise: there must be a caption in the data set that matches the query image. However, in practical applications, this is impossible. Therefore, instead of directly finding the best matching description sentence, some methods try to refine the phrases in the best matching description sentence and synthesize them into a new caption sentence.

Gupta et al. [10] used the Stanford CoreNLP toolkit to split the caption sentences in the dataset into descriptive phrases. When given a query image, search through image features to obtain similar images as candidate image sets, then use the trained model to select relevant descriptive phrases from image descriptions corresponding to the candidate image sets, and finally pass these related Phrases to generate a new caption sentence. There are many similar types of research works. For example, Kuznetsova et al. [11] proposed a method based on a tree structure. They obtaining relevant description phrases from existing image captions as the leaves of the tree, and then selectively combine some phrases to form new sentences as a caption; Socher [12] uses a tree structure to embed the image caption into vector space, learns to extract the subject and action in the caption from the word order and syntactic structure. And finally completes the image caption by subject and action.

Such methods rely heavily on existing data sets. Because in the specific image captioning, it reorders the caption of similar images in the data set, and finally uses the description sentence as the description of the query image. Therefore, this method cannot generate new sentences very well, which is also the obvious defect of this method: First, if there are only a few particularly good image description sentences in the data set, the final caption will be difficult to produce satisfactory results; Secondly, if the query image differs greatly from the image in the data set, for example, the difference in content or style is obvious, it is often difficult to find a similar image in the data set for the query image and it is also difficult to obtain better results.

2.2 | Language template-based approaches

In the earlier work of image caption, another commonly used approach is the language template-based approach. This type of method usually first makes a basic understanding of visual features of the picture and finds out some visual features, such as objects, relationships, attributes and so forth, and then generates caption based on these visual features obtained through a language model. Generally, multiple sentences are generated to form a candidate set, and then the description sentences in the candidate set are sorted, and the sentence with a higher ranking is selected as the final result. In this type of method, the most important idea is still to extract handicraft visual features and then generate captions through a language model.

Yang et al. [13] proposed an image captioning method that uses the nouns-verbs-scenes-prepositions quadruplet as a sentence template. In order to describe an image, the detection algorithm [14, 15] is first used to analyze the objects and scenes in the image, and then the language model [16] trained on the Gigaword corpus is used to determine the verbs, scenes, and prepositions that can be used to form sentences. Then combine the calculated probabilities of all elements and use hidden Markov model inference to obtain the best quadruplet. Finally, the image caption is generated according to the information of the determined quadruplet and a language template.

Kulkarni et al. [17] used conditional random fields(CRF) to extract statistical data from a large number of visual descriptive text pools to smooth the output of vision detection and recognition algorithms to determine the words' content in the image caption. Their method can generate more realistic descriptive sentences of the content. Specifically, they built a graph structure. The nodes of the graph represent objects and attributes, and the spatial positions between nodes represent the relationships between objects. Obtain the unary potential functions of nodes to learn the representation vector of them by using corresponding visualization model, obtain the pairwise potential functions through statistics of existing descriptions' set to learn the relationship between two nodes, and then predict the image content through the conditional random field according to the unary potential functions and pairwise potential functions. Finally, the predicted image content generates captions through the template-based method.

Li et al. [18] used the visual model to extract the object, attribute, and spatial relationship information of the image, and defined it in the triplet of "(attribute-object), preposition, (attribute-object)." When given a query image, employ web-scale n-gram data for phrase selection to collect candidate phrases that may form a triplet. Then, the dynamic programming method is used to realize phrase fusion to find the optimal compatible set of phrases as the caption of the query image.

Mitchell et al. [19] also used handicraft visual algorithms to process images. The process of extracting objects, actions, and scenes in the image to represent the image according to the visual algorithm. After that, the entire image caption task is formulated as the generation of a decision tree. They cluster and sort object nouns to determine the content to be described, and

finally generate the content seen by the computer vision system in detailed descriptive sentences through the Trigram language model [20].

Some studies have made new attempts on this basis. Compared with words, phrases can often express more content better, so some researchers have proposed methods for phrase generation. They believe that it is not good enough to use visual models to obtain visual content, such as objects, attributes, actions, scenes, and prepositions, from images and express them as words, and then generate descriptive sentences from these words.

Fang et al. [21] proposed a method of visual, language models and multi-modal similarity detectors that directly learn from image caption data sets. The author uses multiple instance learning directly from the images and their related descriptions, inputs the words returned by these detectors into the language model to generate descriptions, and then reorders them to select the most similar description as the generation result.

Yatskar et al. [22] used a deep CRF model to deal with the situation-driven prediction of objects and activities and collected a large-scale data set containing more than 500 activities, 1700 characters, 11,000 objects, 125,000 images, and 200,000 unique situations. Ushiku et al. [23] proposed a “model and similarity general subspace” method, which is used to directly learn a phrase classifier to describe an image.

This type of method first obtains visual content information from the image, such as objects, attributes, actions, scenes and so forth, and then generates descriptive sentences through the language template. Therefore, this kind of method can usually represent the image visual content elements better, and at the same time, it can also generate grammatically correct caption through the language model. However, sentences generated by language models are often simple in form and lack diversity. They are not natural and fluent in many situations. Moreover, due to the limitations of language models, their generalization ability needs to be strengthened.

3 | THE RECENT DEEP LEARNING METHODS

The relatively early researches used the search-based method and the language template-based method for an image caption, and these two methods have their significant defects. With the remarkable progress of deep learning in many fields, more and more researchers have begun to pay attention to neural networks. The same is true in image captioning, especially when the Encoder-Decoder model [24] has made significant progress in machine translation tasks. Affected by this idea, image captioning has also begun to try this mapping to learn visual features to describe sentences directly from data and outperforms the above two methods. Since this method of image description added to the neural network model mainly uses sequential network models such as VGG, ResNet and so forth, convolutional neural network (CNN) in the encoder or RNN and LSTM in the decoder part [25–29], it is known as

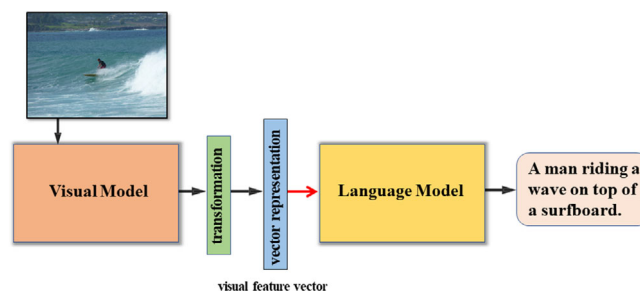


FIGURE 3 Sequence-based approach research framework

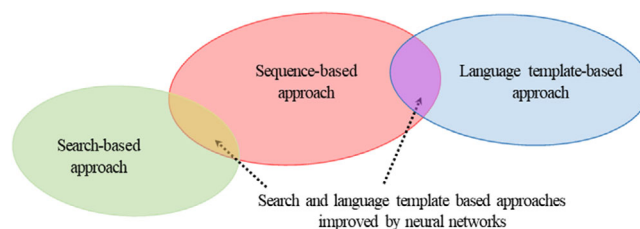


FIGURE 4 Schematic diagram of the relationship between the three methods

the Sequence-based approach. The overall framework of the sequence-based approach is illustrated in Figure 3.

The relationship between the search-based approach, the language template-based approach, and the sequence-based approach is illustrated in Figure 4. Even though deep neural networks are widely adopted for tackling image captioning task, different methods may be based on different technical frameworks. Therefore, we classify sequence-based methods into subcategories according to the main technical framework and discuss each subcategory, respectively.

3.1 | Search and language template based approaches improved by neural networks

Different from search-based and language template-based methods, inspired by advances in the field of deep neural networks, deep neural networks are employed to perform image captioning tasks as an encoder or decoder part of the visual-language task. When the neural network is adopted as an optical encoder, it mainly learns the expression of images to visual features. In contrast, when adopted as a linguistic decoder, it needs to learn to map transformed visual features from the query image to the corresponding description sentences. The framework of search and language template-based approaches improved by neural networks is shown in Figure 5.

Socher et al. [30] employed the CNN model proposed in [31] to extract visual features from a query image, analyzed the phrases order and sentence syntax of captions in the data set, and expressed them as vectors by relying on trees. Then mapped these features to a common vector space through the maximum margin objective function. Finally, search for the corresponding

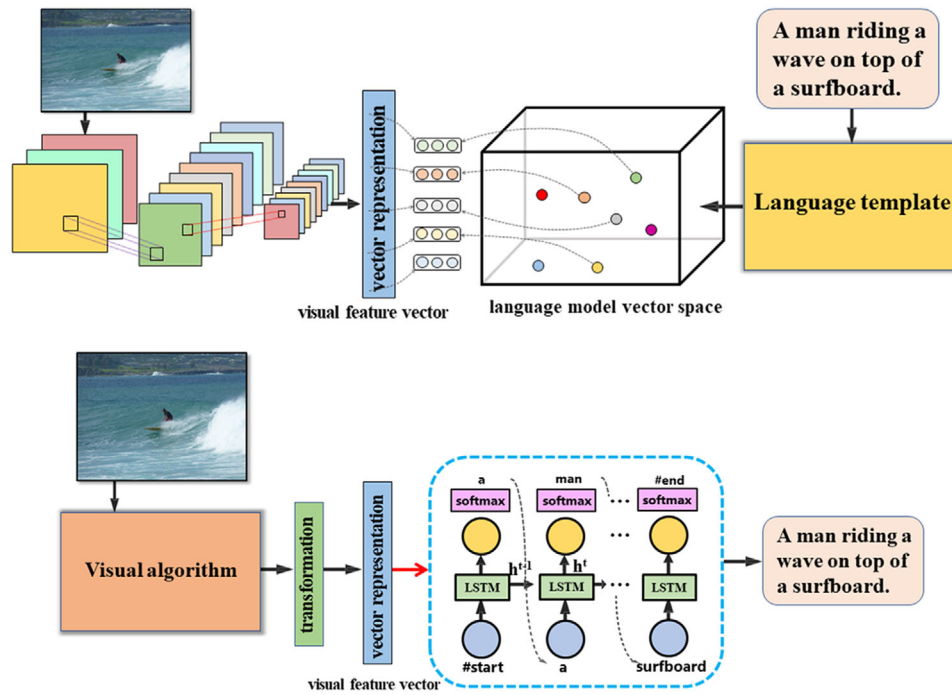


FIGURE 5 The framework of search and language template based approaches improved by neural networks

image description by calculating the inner product of the image visual features and the description vector.

At the same time, Karpathy et al. [32] proposed a method to map image fragments and sentence fragments to the same common space, and then calculate the similarity between the caption and the query image. Rather than directly mapping entire images and captions into a common embedding space, the difference is that the author uses more fine-grained units. The author employed the RCNN model [33] to represent the image as an image fragment, uses the dependency tree relationship to process the description sentence to obtain the sentence segment, and finally designs a maximizing margin target to align the features. The similarity between image fragment features and description sentence fragments is calculated by the inner product to select the description sentence. However, it should be noted that this method cannot generate descriptive sentences, and its follow-up work [46] makes up for this shortcoming, which will be introduced later in this section.

Lebert et al. [34] proposed a phrase-based image caption method. The author uses a pre-trained CNN model [35] to generate image representations, uses SENNA software to extract phrases from description sentences, and then expresses them as high-dimensional vectors through some representation methods of word vectors [36–38]. Finally, a bilinear model is trained to measure the generated image representation and phrase vector to search for the description corresponding to the image.

Kiro et al. [39] learned an image-caption vector space and a language model to decode this space. The author uses CNN and LSTM models to encode image and caption sentences

respectively. The encoded image and sentence representations are mapped to the same computing space through two fully connected networks, and then the CNN model and the LSTM model are trained separately. Besides, the author proposes Log-bilinear neural language models and Multiplicative neural language models to decode the representation vector and form a new description. Similar work also [40].

In the above studies, although manual visual algorithms, language models, or measurement systems are still used in the entire framework, the performance has been improved due to adopted neural network models. However, the framework formed by the manual algorithms and the neural networks trained by separated often does not achieve optimal performance. There are roughly three reasons for this: (1) multiple modules in the entire framework cannot learn from each other during the design or training process; (2) the objective training function deviates from the overall performance index of the system; (3) The algorithm itself cannot completely exclude the limitations of artificial design methods, and it imposes restrictions on the generated descriptions.

Therefore, researchers try to generate descriptions through end-to-end systems. They hope that by reducing manual pre-processing and subsequent processing, as much as possible to make the model from the original input to the final output, using a pipelined model, to avoid the inherent shortcomings of the multi-module mentioned above. Moreover, the end-to-end approach reduces the project's complexity and gives the model more room for free play. We comprehensively review them in the next several sections.

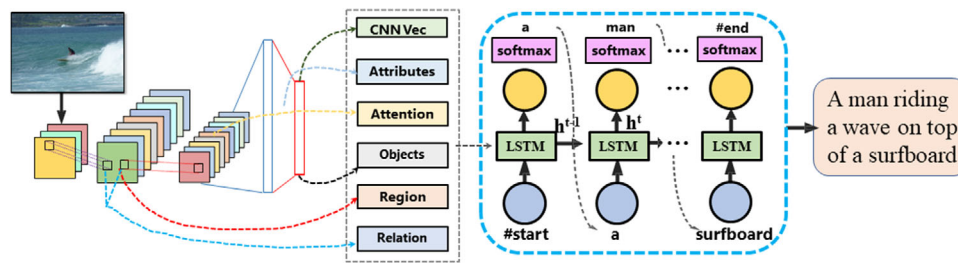


FIGURE 6 Image captioning with high-level representations methods

3.2 | Image captioning with high-level representations

A glance at the image is enough for humans to point out and describe many details about the visual scene. However, it turns out that this extraordinary ability is an elusive task for our visual recognition model. Some studies expect to design a sufficiently rich model to simultaneously reason about the high-level semantic contents of the query image and its representation in the field of natural language. Image captioning with high-level representations methods is given in Figure 6.

Wu et al. [41] provided a new idea. The author believes that the feature map after the CNN model should not be directly connected to the image caption problem, but should have high-level semantic features. They extracted the 256 words that appear most frequently (at least 5 times) from the descriptive sentences in the training set as the most representative attributes. The VGG [35] network model pre-trained on ImageNet [42] is used to modify its final output into a 256-dimensional attribute vector, corresponding to the extracted 256 attributes, and this CNN structure is used as the encoder. Since a picture may correspond to multiple attributes, through the training of multi-label task classification, the features extracted by the CNN structure contain semantic information. Furthermore, to ensure that the model can effectively extract semantic information, the author also added a detector to further improve the accuracy of the model. The decoder continues to use the LSTM structure, and finally generates an image caption rich in semantic information.

Karpathy et al. [43] adopted a similar structure. Their previous work [32] continued in-depth, replacing the previous language model with RNN to extract the features of the description sentence. Since the RNN contains contextual information, it is considered to be related to the semantics of the entire sentence, and the trained RNN model can better generate description sentences. Yao et al. [44] enhanced attribute learning by integrating the correlation between attributes into multi-instance learning (MIL). To incorporate the attributes into the image description, the author constructs different structures by inputting the image representation and attributes into the RNN in different ways to explore the relationship between them.

Cognitive evidence shows that vision-based language is not directly output from end-to-end, but is related to high-level abstract symbols. Therefore, in addition to describing the properties of objects, visual relationships are used to help the generation of image descriptions, so that the final captions are more in line with the artistic conception of expression.

Yao et al. [45] proposed a GCN-LSTM structure to describe the relationship between objects under the framework of the attention mechanism. They propose the structure of the combination of graph convolutional network (GCN) and LSTM to integrate the semantics and the relationship of objects in space into the image encoder. The relational graph is constructed according to the spatial and semantic connections of the objects detected in the image. Then, the relationship graph refines the representation of each region through the GCN graphic structure to obtain the regional-level relationship perception features, and finally inject it into the LSTM to generate a description sentence. Then, the representation of each region in the relationship graph is refined through the GCN graph, to transform the regional-level relationship perception feature, and finally injected into the LSTM to generate a caption.

Fan et al. [46] propose a Theme Concepts extended Image Captioning (TCIC) framework that incorporates theme concepts to represent high-level cross-modality semantics. They model theme concepts as memory vectors and present a Transformer with Theme Nodes (TTN) to incorporate those vectors for image captioning. On the vision side, TTN is configured to take both scene graph-based features and theme concepts as input for visual representation learning. On the language side, TTN is configured to take both captions and theme concepts as input for text representation re-construction. Both settings aim to generate target captions with the same transformer-based decoder.

The general method based on semantic relation graph is also followed by Song et al. [47]. Here, the main innovation is to explore a comprehensive understanding of contextual interactions reflected on various visual relationships between objects. The region-based bidirectional encoder from the transformers (regional BERT) represents the drawing of global interactions between detected objects without extra relational annotations. Liu et al. [48] achieve unpaired image captioning by bridging the vision and the language domains with high-level semantic information.

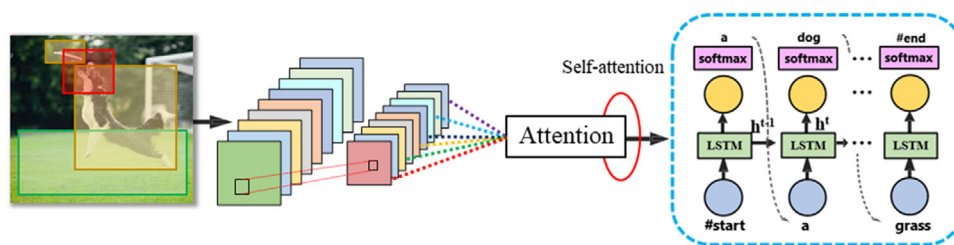


FIGURE 7 General framework of enhanced image captioning with attention correction

3.3 | Enhanced image captioning with attention correction

Images can convey rich semantics due to rich visual information. However, only the most prominent content needs to be paid attention to in image captioning tasks. Moreover, while the neural network exhibits powerful representation capabilities, it also brings redundant and disturbing information due to its complex structure. Inspired by the human visual attention mechanism, methods of using attention to guide image caption generation are proposed. In such methods, the attention mechanism based on various kinds of feature representation of the input image is integrated into the language generation framework to make the generation process focus on the interest regions of the visual features at each step to generate a description of the input image. The most relevant visual coding strategy for image captions with attention mechanism is given in Figure 7.

Vinyals et al. [49] proposed a model based on CNN+LSTM. This structure provides a general idea for image description tasks. The framework first encodes the image with GoogleNet [50] as feature vectors, and then decodes them with LSTM, which outputs the probability of all words in the word list, selects the highest word as the output, and finally forms the image description. The model achieved state-of-the-art performance at the time.

Xu et al. [51] added an attention mechanism on this basis to further improve the performance of the model. The author also uses the image as input, CNN as the encoder to extract the image features to form a feature map, and then through the attention mechanism to enhance or suppress the feature map. As the input data into the LSTM model, the data after the attention mechanism at different moments will be adjusted by the output data of the LSTM model at the previous moment, and finally, the image description is generated through the LSTM model.

Yang et al. [52] believe that the attention mechanism only pays attention to the part each time, and does not consider the impact of global factors on the prediction. Therefore, they proposed two models: CNN encoder + RNN decoder and RNN encoder + RNN decoder. The CNN model feature map that captures the global features of the image is input to the LSTM decoder unit to obtain a more compact and abstract vector representation, the thought vector. And the thought vector is used as the input of the attention mechanism in the decoder to ensure

the global information while not omitting the local information. Finally, the vector is passed through the RNN model to generate a caption. Besides, the author also designed a recognizable supervisory training mechanism concerning the research in [21].

Chen et al. [53] proposed an attention mechanism combining space and channel for CNN+RNN structure. The author believes that the previous attention mechanism only considers the spatial relationship, so it introduces the channel attention mechanism in the multi-layer feature map. Since the feature mapping of the channel direction is essentially the detector response mapping of the corresponding filter, the maintenance of the channel direction can be regarded as a process of selecting semantic attributes according to the needs of the sentence context. The network uses an encoding-decoding framework to generate image descriptions. Through multi-level channels and spatial attention mechanisms, the feature maps in each level of CNN are given the ability to adapt to sentence context. There are similar related works through the attention mechanism [54–58].

Another interesting observation is that the ability of captions to distinguish target images from other similar images has not been fully explored. This causes the relationship between objects in the similar image group to be ignored. Wang et al. [59] improved the distinctiveness of image captions using a group-based distinctive captioning model, and proposed a new evaluation metric DisWordRate to measure the distinctiveness of captions.

Recently, transformer has shown good performance when dealing with serialized information. More importantly, the transformer has been recognized as the latest technology for sequence modelling tasks such as language understanding and machine translation. Some studies have adopted the newest transformer architecture for image captioning. The general transformer-based methods framework is given in Figure 8. Transformer-like architectures could be applied directly on visual context patches, thus excluding or limiting the usage of the convolutional operator [60,61]. On this line, Liu et al. [62] designed the first non-convolutional architecture for image captioning. Specifically, a pre-trained Vision Transformer in [60] is employed as an encoder, and then a general Transformer is accepted as decoder to generate captions.

Dong et al. [63] proposed dual graph convolutional networks (Dual-GCN) with transformer and curriculum learning to explore the contextual relevance between contextual images for image captioning, see Figure 9. Two independent GCNs

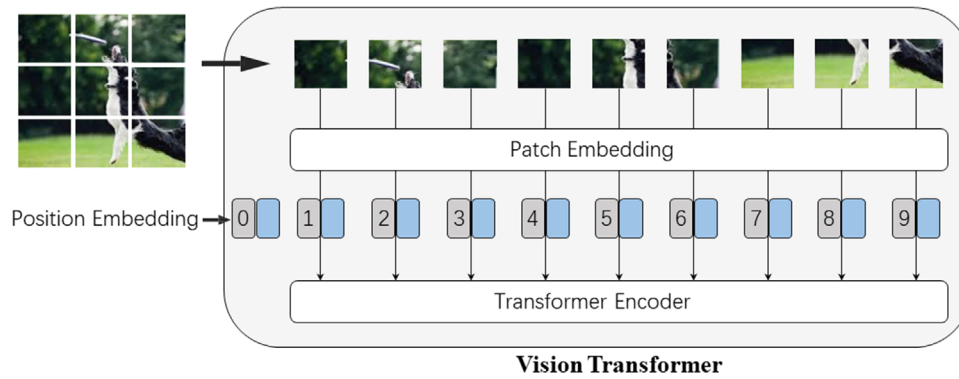


FIGURE 8 General transformer-based methods framework for automatic image captioning

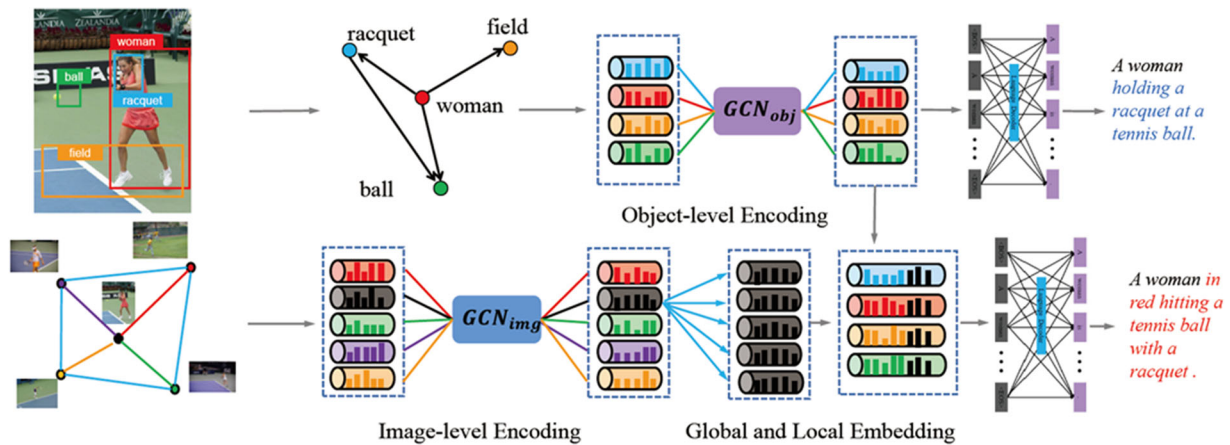


FIGURE 9 An illustration example of dual-GCN proposed by Dong et al. [63]

encode the entire image and the objects from the image, and then the captions are generated by a Transformer linguistic decoder. Ji et al. [64] introduce a Global Enhanced Transformer to extract a more comprehensive global representation and then adaptively guide the decoder to generate high-quality captions.

Sariyildiz et al. [65] introduced transformer-based image conditional masking language modelling (ICMLM) for learning the visual representation of image-caption pairs. Lee et al. [66] proposed a new metric UMIC, an unreferenced metric for image captioning which does not require reference captions to evaluate image captions, and adopted a pre-trained transformer to generate captions. Yang et al. proposed [67] a novel transformer, ReFormer, adapted to generate features embedded in relational information and clearly express the paired relations between objects in images.

3.4 | Image captioning based on entity recognition with novelties

Since current image caption tasks are usually based on image-caption pairs, the generated description sentences can only capture the goals learned during the training process, and cannot be well extended to many novel scenes and objects, in

which images outside of the datasets. In practical applications, some pictures contain rich information while cannot be fully expressed by existing models. Therefore, how to promote the vocabulary expansion of the description sentence and integrate the recognized target into the description is also a problem for researchers.

Yao et al. [123] present a new architecture long short-term memory with copying mechanism (LSTM-C) for describing novel objects in captions. Specifically, freely available object recognition datasets are leveraged to develop novel objects classifiers. Then LSTM-C combines the standard verbatim sentence generation of the decoder RNN with a copy mechanism, and selects words from novel objects at the appropriate position in the output captions. Lu et al. [124] adopted a similar framework. The difference is that they first generate a sentence ‘template’ with slot locations explicitly tied to specific image regions. These slots are then filled in by visual concepts identified in the regions by object detectors instead of novel objects classifiers.

Yang et al. [68] employed a scene graph model to integrate each object and the relationship between its attributes and other objects through graph convolution and unified as the input of the network. As shown in Figure 10, their studies detects objects through Faster-RCNN [69], detects object relationships through MOTIFS [70], and detects attributes through a fully

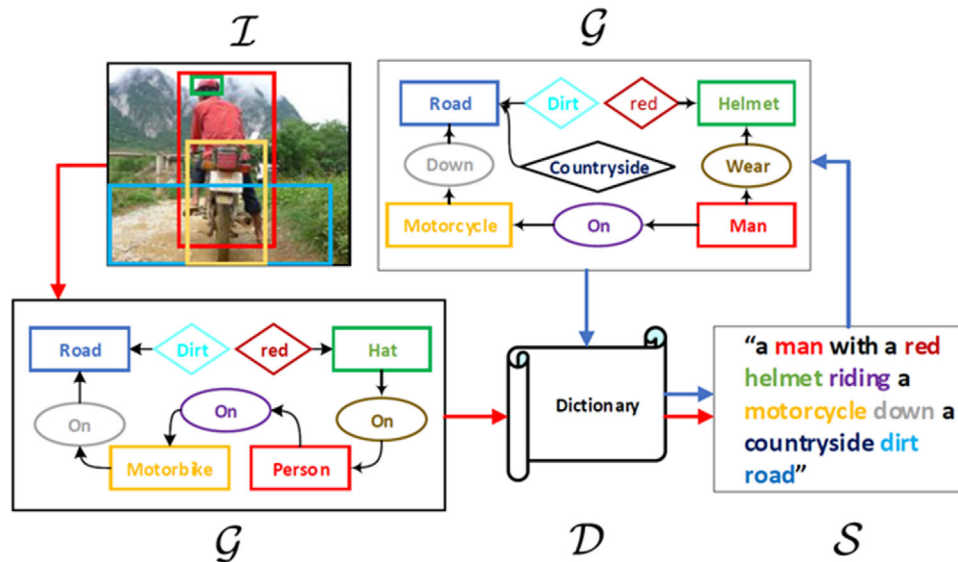


FIGURE 10 Auto-encoding scene graphs for image captioning proposed by Yang et al. [68]

connected network. To introduce the prior knowledge in the corpus, they constructed the concept of a dictionary, which pre-trained on the text corpus, and then used it as an initialization directly in the description model. Finally, the attributes, objects, and relationships after the dictionary mapping are used as the input of RNN to describe and generate. Anderson et al. [71] adopted constrained beam search to force the inclusion of selected tag words in the output, and fixed, pretrained word embeddings to facilitate vocabulary expansion to previously unseen tag words. Yao et al. [72] proposed Hierarchy Parsing (HIP) to analyze the hierarchical structure of images. The author regards HIP as a feature optimizer, integrating features from the overall image level, the regional level, and the instance level. Chen et al. [73] proposed a structure ASG to represent user intention in fine-grained level and control what and how detailed the generated description should be.

Note that the above studies could not fully consider the connection of contextual descriptions in the dataset. Mahajan et al. [74] leverage the contextual descriptions in the dataset to explain similar contexts in different visual scenes. For this reason, a new type of latent space decomposition is proposed, named contextual-object split, to model the diversity of contextual descriptions of images and texts in the dataset, see Figure 11.

Li et al. [75] proposed a pointing mechanism for automatic captioning. They first used the target detection model pre-trained on other data sets to do a detection classification on the image, and then the obtained prediction results were fused with LSTM. Based on the original caption vocabulary, the label vocabulary in the target detection data set is added to form an extended vocabulary as training data set. Since the image captioning model usually learns too many priors in the training set, which leads to hallucinations, the pointing mechanism is used to balance the decoder-generated words or copy directly from

the recognized words. The methods of [76–78] are also similar studies.

Another interesting observation is that all of the above models require a training set of fully annotated image-sentence pairs. However, the cost of obtaining large amounts of such data is prohibitive. Demirel et al. [79] proposed and studied a practically important variant of this problem where test images may contain visual objects with no corresponding visual or textual training examples. For this problem, they proposed a detection-driven method based on a generalized zero-shot detection model and a template-based sentence generation model.

Even though the captions generated by the above methods can accurately describe the images, they are universal to similar images and lack distinctiveness, cannot accurately convey the uniqueness of each image. Wang et al. [80] address this problem through training with sets of similar images. A distinctiveness metric CIDErBtw is proposed to evaluate the distinctiveness of a caption with respect to those of similar images. Then, the distinctiveness of the caption generated by each image is encouraged based on the training strategy using CIDErBtw in the weighted loss function or as a reinforcement learning reward.

Although the above-mentioned methods have achieved good results, there are still some problems:

1. Since the process of RNN generating sentences is a word-by-word process, during training, the last word input to the RNN is from the label in the training set, but at the time of prediction, the words input to the RNN depend on the words already generated by the RNN itself. Moreover, the sequence structure causes the accumulation of errors, in other words, once one word not generated well, leading to the following words are also not generated well;
2. The cross-entropy loss function is used when the model is trained, but the evaluation index in the machine translation

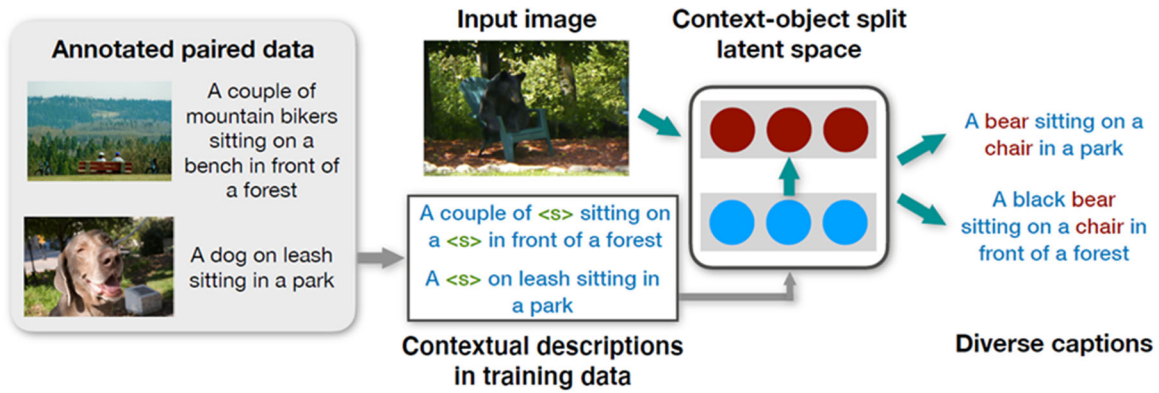


FIGURE 11 The context-object split conditional variational autoencoder model proposed by Mahajan et al. [74]

is used when it is verified, and there is a problem that it does not correspond.

3.5 | Improved image captioning via reinforcement learning

Image captioning, as a visual-text task, most methods to solve this task require training a statistical model on a data set of (image-caption) pairs. The deep models are usually trained to maximize the log-likelihood of the training set. However, the log-likelihood score of captions does not have a reasonable correlation with the human evaluation of quality. Standard syntactic evaluation metrics are also not very relevant or traditionally tricky to optimize. Therefore, some studies adopt reinforcement learning to optimize image descriptions.

Rennie et al. [81] proposed an SCST method to train image description models. However, considering that the variance of reinforcement learning in calculating gradients will be large, which will make the training unstable, the author used the sentences generated by the model during testing as a benchmark instead of profiling feedback to reduce the variance. In the sampling process, sentences that are better than the benchmark will get a positive weight, and vice versa, bad sentences will be suppressed.

Liu et al. [82] proposed an image annotation method based on reinforcement learning. The author adopts the Encoder-Decoder model in [49], adopts SPIDER as the reward function, and uses policy gradient optimization based on Monte Carlo rollouts. The evaluation metric SPIDER is used to ensure the fluency of caption sentences and the semantical faithfulness of the image. Wang et al. [83] combined reinforcement learning and imitation learning and proposed a method of Reinforced Cross-Modal Matching (RCM).

Wu et al. [84] designed a novel global-local discrimination target to facilitate the generation of fine-grained descriptions, illustrated in Figure 12. Their model consists of a global discriminative objective and a local discriminative objective. The global discriminant constraint is developed from a global perspective, which pulls the generated sentence to better distinguish the

corresponding image from all other images in the entire data set. From a local perspective, a local discriminant constraint is proposed to increase attention and make it emphasize less frequent but more specific words or phrases, thereby facilitating the generation of titles that better describe the visual details of a given image.

There are also some researches starting from the definition of the image caption task and providing some new ideas for image caption generation. Jia et al. [85] proposed LSTM based method to guide the model towards solutions more tightly coupled to the image content. A picture is worth a thousand words. A picture is worth a thousand words. Justin and Xu et al. [86, 87] divided the image into regions. They then described them separately and finally generated a series of simple sentence sets for the image, taking into account the two aspects of object positioning and description. The examples of Dense captioning task is illustrated in Figure 13. Venugopalan et al. [88] realized an end-to-end video description model for video frame sequence input and text sequence output. Deng et al. [89] propose using a simple length level embedding to give the ability to control the title, e.g., choosing to describe the image either roughly or in detail. Chen et al. [90] make a brave attempt towards an annotation-free evaluation of cross-lingual image captioning.

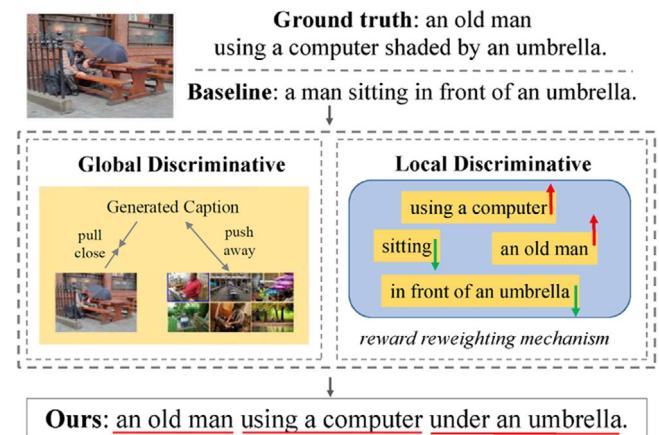


FIGURE 12 Fine-grained image captioning with global-local discriminative objective proposed by Wu et al. [84]

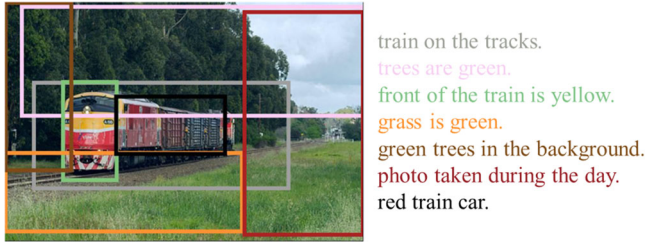


FIGURE 13 Dense captioning task examples display as proposed Justin et al. [86]; Xu et al. [87]

As shown in Table 1, we summarize the image caption methods discussed in this survey according to method categories. The table lists the specific methods, datasets, and evaluation metrics for every reference mentioned in this survey to compare. It can be seen that with the adoption of neural networks, newer frameworks are no longer satisfied with the expression described in the data set, and hope to generate detailed or novel description sentences. Additionally, the evaluation methods have also become more consistent. More methods have adopted evaluation measures (CIDEr, SPICE) related to human judgment and perform well. Moreover, large data sets such as COCO and Visual Genome are used more and more frequently.

4 | COMPARISON OF STATE-OF-THE-ART METHODS

4.1 | Evaluation metrics

In this section, we compare image captioning methods that give state-of-the-art results. The image captioning task is comprehensive of computer vision and NLP. It can be simply understood that this task requires the model to recognize objects, actions, scenes, and relationships between objects in the image, and then map these contained visual contents into descriptive sentences. In general, this vision-language task requires two basic requirements:

1. The correctness of the grammar—the language grammar needs to be followed during the mapping process to make the result readable;
2. The richness of the description sentence—the generated caption needs to be able to accurately describe the details of the corresponding image, and produce a sufficiently complex description.

Due to the complexity of the output of the image description task, how to evaluate the description is very difficult. There are currently many evaluation metrics to assess the image caption in terms of language quality and semantic correctness [91–97]. The more commonly used evaluation metrics mainly include BLEU, ROUGE, METEOR, CIDEr, SPICE. Among them, BLEU, ROUGE-L, and METEOR originated from machine translation, used to judge the language quality of machine translation,

and have been widely used in image caption tasks, while CIDEr and SPICE are more inclined to the evaluation of semantic information. In fact, the most intuitive evaluation index is through direct human judgment. However, because manual evaluation requires a large amount of non-reusable workforce, it is not easy to scale up. Therefore, in this article, we report a comparison of methods based on automatic image caption evaluation metrics.

BLEU The Bilingual Evaluation Understudy (BLEU) method [91] is adopted to evaluate the quality of translated sentences in machine translation. It compares each translation segment with a set of reference translations with good translation quality and calculates each segment score then estimates the overall quality of the translation. In the field of image description, as a similarity measurement method, BLEU adopts an n -gram matching rule. The BLEU evaluation metric can be evaluated by analyzing the co-occurrence frequency of n -gram in the predicted caption and the label. Let Candidates and Reference be the predicted caption and the label respectively. For an n -gram, the precision of the sentence can be expressed as follows:

$$P_n = \frac{\sum_{C \in \{Candidates\}} \sum_{n\text{-gram} \in C} \text{Conut}_{clp}(n\text{-gram})}{\sum_{C' \in \{Candidates\}} \sum_{n\text{-gram}' \in C'} \text{Conut}(n\text{-gram}')} \quad (1)$$

where $\text{Conut}_{clp}(n\text{-gram})$ represents the minimum number of occurrences of $n\text{-gram}$ in the Candidates and the reference. $\text{Conut}(n\text{-gram}')$ indicates the number of times $n\text{-gram}$ appears in the Candidates.

As the number of tuples increases, the P_n of the $n\text{-gram}$ statistics will decrease. To prevent the training results from tending to short sentences, multiply by the brevity penalty factor to get the final evaluation formula:

$$BLEU = BP \times \exp\left(\sum_{n=1}^N w_n \log P_n\right) \quad (2)$$

where w_n is the weight of the precision of $n\text{-gram}$. Let c be the length of the Candidates and r be the Reference corpus length. The value of BP is as follows:

$$BP = \begin{cases} 1 & \text{if } c < r \\ e^{1-r/c} & \text{if } c > r \end{cases} \quad (3)$$

Specifically, BLEU-1 divides the description sentence and the label into words, counts the number of times the words in the description sentence appear in the label one by one, and records the minimum number of times that tuple appears in the description sentence and label, and Carry out the ratio with the description sentence, and finally, to avoid the bias problem of the generated description sentence being too short, multiply it by a penalty factor to get the final result. BLEU-2 divides the description sentence and label into 2-tuples containing two words for statistics and calculations. Under normal circumstances, up to 4-tuples are calculated.

TABLE 1 An overview of the approaches in this survey and organized in method categories order

Year	Reference	Method	Evaluation metrics	Datasets
2011	Ordonez [5]	Search-based	BLEU	SBU1M
2012	Gupta [10]		Human, BLEU, ROUGE	Pascal1K, IAPR
2013	Hodosh [6]		Human, BLEU, ROUGE	Pascal1K, Flickr8K
2014	Mason [9]		SBU1M	Human, BLEU
2014	Kuznetsova [11]		Human, BLEU, Meteor	SBU1M
2011	Yang [13]	Language template-based	BLEU, ROUGE, Meteor, CIDEr	IAPR, Flickr8K/30K, MSCOCO
2011	Li [18]		Human, BLEU	Pascal1K
2012	Mitchell [19]		Human	Pascal1K
2013	Kulkarni [17]		Human, BLEU	Pascal1K
2014	Ushiku [23]		BLEU	Pascal1K, IAPR, SBU1M, MS COCO
2015	Fang [21]	CNN + search-based	Human, BLEU, ROUGE, Meteor, CIDEr	MS COCO
2014	Socher [30]		ROUGE	Pascal1K
2014	Karpathy [32]		BLEU, Meteor, CIDEr	Flickr8K/30K, MSCOCO
2014	Kiros [40]		ROUGE	Flickr8K/30K
2015	Lebret [34]		BLEU, ROUGE	Flickr30K, MSCOCO
2015	Wu [41]	Attributes	BLEU, Meteor, CIDEr, PPL	Flickr8K/30K, MSCOCO
2015	Karpathy [43]		BLEU, Meteor, CIDEr, mRank, ROUGE	Flickr8K/30K, MSCOCO
2015	Ji [61]		BLEU, Meteor	Flickr8K/30K, MSCOCO
2015	Jia [85]		BLEU, Meteor, CIDEr	Flickr8K/30K, MSCOCO
2016	Justin [86]		Meteor, ROUGE	MS COCO, Visual Genome
2018	Yao [45]	Attention + relationship	BLEU, Meteor, ROUGE, CIDEr, SPICE	MS COCO
2021	Fan [46]		BLEU, Meteor, ROUGE, CIDEr, SPICE	MS COCO
2015	Vinyals [49]		BLEU, Meteor, CIDEr, ROUGE	Pascal1K, SBU1M, Flickr8K/30K
2015	Xu [51]		BLEU, Meteor	Flickr8K/30K, MSCOCO
2016	Yang [50]		CS-k	HabeasCorpus
2016	Chen [53]	CaPtion transformeR	BLEU, Meteor, ROUGE, CIDEr	Flickr8K/30K, MSCOCO
2016	You [57]		BLEU, Meteor	Flickr30K, MSCOCO
2017	Lu [54]		BLEU, Meteor, ROUGE, CIDEr	Flickr30K, MSCOCO
2018	Jiang [54]		BLEU, Meteor, ROUGE, CIDEr, SPICE	MS COCO
2020	Pan [58]		BLEU, Meteor, ROUGE, CIDEr	MS COCO
2021	Liu [61]	GCN+transformer+ curriculum learning	BLEU, Meteor, ROUGE, CIDEr	MS COCO
2021	Dong [63]		BLEU, Meteor, ROUGE, CIDEr	MS COCO, Visual Genome
2021	Ji [64]		BLEU, Meteor, ROUGE, CIDEr, SPICE	MS COCO
2021	Lee [66]		BLEU, Meteor, ROUGE, CIDEr, SPICE, UMIC	Composite, Flickr8k, PASCAL50s
2018	Lu [124]		BLEU, Meteor, CIDEr, SPICE	Flickr30K, MSCOCO
2019	Yang [68]	Relationship + novel object	BLEU, Meteor, ROUGE, CIDEr, SPICE	MS COCO, Visual Genome
2019	Yao [72]		BLEU, Meteor, ROUGE, CIDEr, SPICE	MS COCO, Visual Genome
2019	Li [75]		Meteor, ROUGE, SPICE	held-out MSCOCO
2019	Yue [78]		Meteor, CIDEr, SPICE	MS COCO
2020	Chen [73]		BLEU, Meteor, ROUGE, CIDEr, SPICE	MS COCO, Visual Genome
2021	Mahajan [74]	Contextobject split	BLEU, Meteor, ROUGE, CIDEr, SPICE	MS COCO

(Continues)

TABLE 1 (Continued)

Year	Reference	Method	Evaluation metrics	Datasets
2021	Wang [80]	Pretrained model + self-critical sequence training	BLEU, Meteor, ROUGE, CIDEr, SPICE, CIDErBtw	MS COCO
2017	Rennie [81]	Reinforcement learning	BLEU, Meteor, ROUGE, CIDEr	MS COCO
2017	Liu [82]	Policy gradient	BLEU, Meteor, ROUGE, CIDEr,	MS COCO
2019	Wang [83]	Reinforcement learning	—	R2R dataset, VLN Challenge
2020	Wu [84]	Global-local discriminative objective	BLEU, Meteor, ROUGE, CIDEr, SPICE	MS COCO
2020	Deng [89]	LaBERT	BLEU, Meteor, ROUGE, CIDEr, SPICE	MS COCO

ROUGE The automatic abstract evaluation (recall-oriented understudy for gisting evaluation, ROUGE) method [92] evaluates abstracts based on the co-occurrence information of the N-tuples in the evaluation abstracts. It is an evaluation method oriented to the recall rate of N-tuples and is used to evaluate the machine's fluency of translation. In the evaluation, ROUGE uses dynamic programming to determine the longest common subsequence between the caption and the label and then calculates the recall of them based on the calculated common subsequence to determine the similarity between the caption and the label.

Take the most commonly used ROUGE-L as an example. Given the Candidate C and Reference R , let $LCS(C, R)$ be the length of the longest common subsequence, it can be an expression as follows:

$$ROUGE - L = \frac{(1 + \beta^2) R_{LCS} P_{LCS}}{R_{LCS} + \beta^2 P_{LCS}} \quad (4)$$

where $R_{LCS} = \frac{LCS(C, R)}{c}$, $P_{LCS} = \frac{LCS(C, R)}{r}$, $\beta = \frac{P_{LCS}}{R_{LCS}}$. c and r represent the length of the candidate and reference.

Like BLEU, the higher the value of the ROUGE indicator, the higher its quality, but neither considers the grammatical accuracy nor the semantic level of description. Since the longest common subsequence does not require words to be continuous, ROUGE can capture the structure of the sentence. Additionally, ROUGE is composed of a series of evaluation methods, including ROUGE-N, ROUGE-L, ROUGE-S, ROUGE-W, ROUGE-SU and so forth. The appropriate ROUGE method can be selected according to the specific research content.

METEOR Metric for evaluation of translation with explicit ordering (METEOR) [93] is also an evaluation index in the field of machine translation. For a query image caption, the METEOR metric calculates the precision and recall and then performs a harmonic average.

Let w_c be the number of words in the candidate, let w_r be the number of words in the reference, and let m be the number of common words in the candidate and the reference. The precision and recall can be expressed as $P = \frac{m}{w_c}$ and $R = \frac{m}{w_r}$. Thus, the harmonic mean can be expressed as $F_{mean} = \frac{PR}{\alpha P + (1-\alpha)R}$. We naturally think that the longer the continuous length of the longest common subsequence matched the better. However, the

evaluation metric considers the case of only a single word, so a penalty factor pen is introduced. Finally, the METEOR metric is calculated as follows:

$$METERO = (1 - pen) \times F_{mean} \quad (5)$$

where penalty factor $pen = \gamma(\frac{ch}{m})^\theta$, ch is the number of chunks, which means contiguous ordered block. α , θ , and γ are hyper-parameters determined according to different datasets.

METEOR overcomes the problem that BLEU does not measure recall and accurately match words. For high-order tuples, BLEU evaluates indirectly by measuring high-order n-grams, while METEOR measures accurate word-to-word matching. However, some researchers have not adopted this evaluation method due to the relatively cumbersome calculation steps and a large number of method parameters.

CIDEr Consensus-based Image Description Evaluation (CIDEr) [94] regards each sentence as a document, and then calculates the cosine angle of the word frequency-inverse document frequency (TF-IDF) vector, and then obtains the similarity between the description sentence and the label. Finally, the final result is obtained by averaging the similarity of tuples of different lengths. The specific formula is as follows:

$$CIDEr_n(c, S) = \frac{1}{M} \sum_{i=1}^M \frac{g^n(c) \cdot g^n(S_i)}{\|g^n(c)\| \times \|g^n(S_i)\|} \quad (6)$$

where c represents a candidate caption, S represents a set of reference captions, n represents an n-gram to be evaluated, M represents the number of reference captions, and $g^n(\cdot)$ represents an n-gram-based TF-IDF vector.

Since tuples appearing more frequently in the entire corpus often contain a smaller amount of information, this method can allow different tuples to have different weights with different TF-IDFs. Therefore, CIDEr can be grammatically and significantly Evaluate descriptive sentences for sex and accuracy.

SPICE Semantic Propositional Image Caption Evaluation (SPICE) proposed by Anderson et al. [95] to use graph-based semantic representation to encode the objects, attributes, and relationships in the description sentence, and to evaluate the description sentence at the semantic level. SPICE parses the candidate and references captions into syntactic dependencies

trees through a dependency parser [97]. After the dependency tree is generated, a rule-based method is used to map the dependency tree into a scene graph. Specifically, the syntactic dependencies tree is generated through three post-processing steps for simplifying quantitative modifiers, analyzing pronouns, and processing plural nouns. After that, the generated tree structure is parsed according to nine simple language rules to extract the objects, relationships, and attributes that make up the scene graph.

Suppose c is a candidate and S represents a set of reference captions. The scene graph generated by the candidate is denoted as $G(c) = \langle O(c), E(c), K(c) \rangle$ and the scene graph generated by S is denoted as $G(s)$. $T(\cdot)$ means to convert a scene graph into a collection of tuples in the form of $T(G(c)) \Leftrightarrow O(c) \cup E(c) \cup K(c)$. Then the precision and recall can be expressed as:

$$P(c, S) = \frac{|T(G(c)) \otimes T(G(S))|}{|T(G(c))|} \quad (7)$$

$$R(c, S) = \frac{|T(G(c)) \otimes T(G(S))|}{|T(G(S))|} \quad (8)$$

Consequently, SPICE metric calculation can be expressed as the following form:

$$SPICE(c, S) = F_1(c, S) = \frac{2 \cdot P(c, S) \cdot R(c, S)}{P(c, S) + R(c, S)} \quad (9)$$

where \otimes the binary matching operator, which returns matching tuples.

Although SPICE can better judge semantic information, it ignores the grammatical requirements and therefore cannot judge the natural fluency of sentences. Besides, since the evaluation mainly checks the similarity of nouns, it is not suitable for tasks such as machine translation.

4.2 | Image caption datasets

So far, there have been a large number of data sets used for image captioning. These data sets are different to a certain extent in terms of data collection and sorting, presentation of data labels, as well as the volume and specifications of the datasets, which lays the data foundation for the task of image description generation. **MS COCO** dataset [98] is a large-scale dataset launched by Microsoft in 2014 that can be used for tasks such as image recognition, object detection, semantic segmentation, and image caption. The images in the dataset consist of nearly 100 object categories from images of daily complex scenes containing ordinary objects under natural backgrounds, and each image is artificially annotated using Amazon Mechanical Turk (AMT).

The dataset contains 82,783 training image samples and 40,504 verification image samples. Besides, there are 40,775 test images whose labels are not open to the public. Each image contains five caption sentences. Due to the complex scenes and diverse annotations of the MS COCO data set, this poses a

greater challenge to the researcher's model, and it is also one of the factors that have attracted more and more attention from researchers.

Flickr8K dataset [6] was released for public use by researchers in 2013. The images in the dataset are all from the photo and image sharing website Flickr and contain 8000 images. Compared with MS MSCOCO, the data scale is small, and the image content is mainly human and animal. The label caption is also through crowdsourcing services by Amazon's manual labelling platform. Each image has five sentences as description.

Flickr30K dataset [99] is an extension of Flickr8K, contains 31783 image data, each image has five sentences corresponding description.

In addition to the above three data sets, some data sets can be used for image caption tasks. For example, the Visual Genome dataset [100] contains 108,077 pictures. On average, each picture has 35 objects, 26 attributes, and more than 5.4 million descriptions of image regions. The Pascal1K dataset [101] includes 1,000 pictures selected from the Pascal object recognition dataset [102], and each picture is annotated with 5 description sentences through the Amazon manual labelling platform. SBU1M [1] contains about one million pictures obtained from Flickr. The image data can be obtained by downloading a CSV file containing the image URL and the corresponding caption. There are also many data sets [103–110] for the majority of researchers to learn and use. Some examples of images and associate description labels are given in Figure 15.

This section mainly introduces three datasets that are currently widely used in image captioning: MS MSCOCO dataset, Flickr8K, and Flickr30K, and organizes datasets in other image description fields or data sets that can be used for image description generation tasks. The datasets are summarized in Figure 14. To understand the scale of the dataset more intuitively, ImageNet, Pascal, and some other datasets are added as a reference in the figure. Please note that the number of specific categories of Flickr-8k, Flickr-30k, and SBU1M is not counted.

4.3 | Comparison on benchmark datasets and discussion

In this section, we compared representative studies for each type of method. The comparison is based on an experiment protocol that is commonly adopted in previous work. For datasets Flickr8k and Flickr30k, 1000 images are used for validation and testing 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 5000 validation images are used for validation, and another 5000 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 dataset are shown in Table 3. The optimal performance has been marked in bold for easy observation.



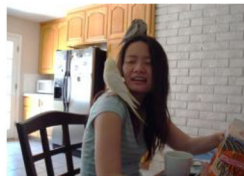
1. A pizza on a pan sitting on a table.
2. A close up of a pizza in a pan on a table.
3. A pizza sits on a plate on a dark surface.
4. A person sitting at a table where a pizza is sitting.
5. A pizza topped with different toppings is brought to a table.

MS COCO



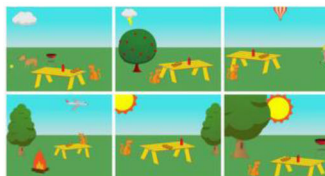
1. A girl going into a wooden building.
2. A little girl climbing into a wooden playhouse.
3. A little girl climbing the stairs to her playhouse.
4. A little girl in a pink dress going into a wooden cabin.
5. A child in a pink dress is climbing up a set of stairs in an entry way.

Flickr8K



1. A woman has a bird on her shoulder, and another bird on her head
2. A woman with a bird on her head and a bird on her shoulder.
3. A women sitting at a dining table with two small birds sitting on her.
4. A young Asian woman sitting at a kitchen table with a bird on her head and another on her shoulder.

Pascal1K



cat anxiously sits in the park and stares at an unattended hot dog that someone has left on a yellow bench.

Abstract Scenes



1. a yellow building with white columns in the background;
2. two palm trees in front of the house;
3. cars are parking in front of the house;
4. a woman and a child are walking over the square.

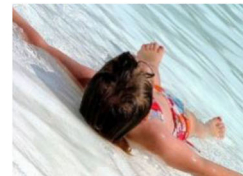
IAPR TC-12



Alt-text: Musician Justin Timberlake performs at the 2017 Pilgrimage Music & Cultural Festival on September 23, 2017 in Franklin, Tennessee.

Conceptual: pop artist performs at the festival in a city.

Conceptual



1. A girl is stretched out in shallow water.
2. A girl wearing a red and multicolored bikini is laying on her back in shallow water.
3. Girl wearing a bikini lying on her back in a shallow pool of clear blue water.
4. A young girl is lying in the sand, while ocean water is surrounding her.
5. A little girl in a red swimsuit is laying on her back in shallow water.

Flickr30K



En: A double decker bus driving down a street.

Ja: 二階建てのバスが道路を走っている。

STAIR Captions

FIGURE 14 Datasets for the image caption task models. Note that some widely used datasets in computer vision have been added to facilitate an intuitive observation and understanding

The current research in this field is mainly focused on deep learning-based methods, among which the attention mechanism and Transformer-based methods seem to be at the forefront of this research topic. In the method of Lu et al. [54], the attention mechanism is trained to filter image regions and guide sentence fragments for image captioning. The authors report their

results on the benchmark dataset Flickr8k and Microsoft COCO Caption dataset in Tables 2 and 3, respectively. On Flickr8k, the BLEU-1, BLEU-2, BLEU-3, and BLEU-4 scores obtained are 0.677, 0.494, 0.354, and 0.251, respectively, which are the best performance. Furthermore, on this dataset, ROUGE scores are reported, which is 0.531. As shown in Table 3,

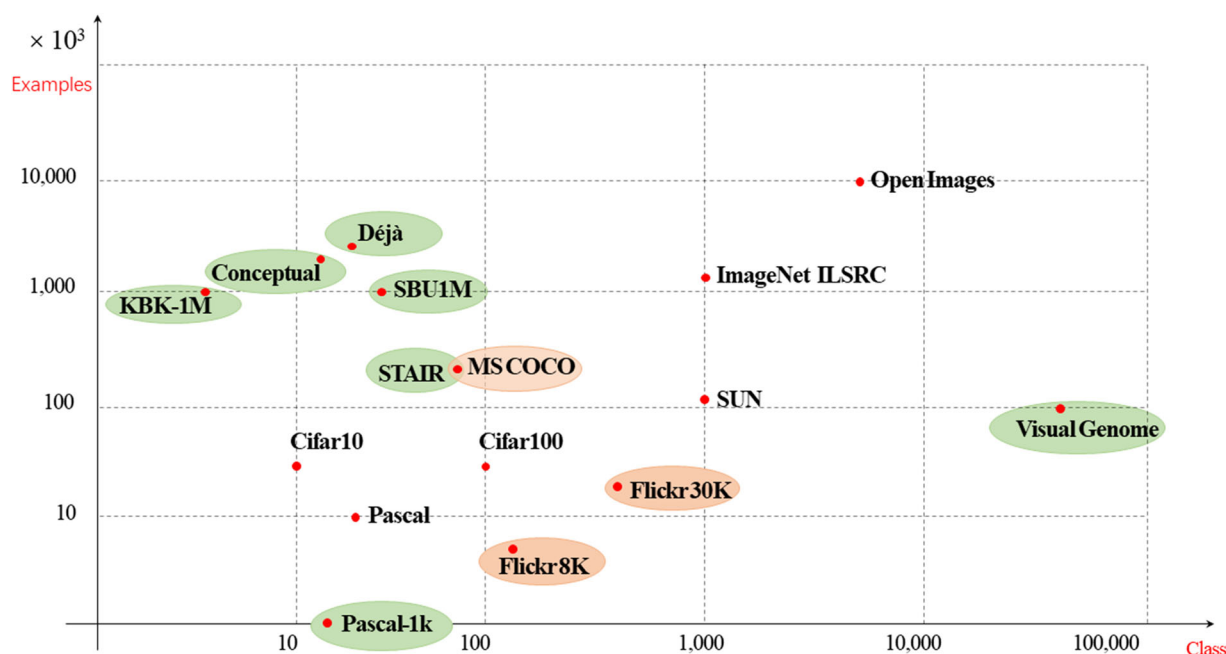


FIGURE 15 Example images and associate description sentences from the benchmark image captioning datasets

they achieved the best performance in most evaluation metrics. The method of Pan et al. [58], another attention mechanism-based image captioning method, achieved higher scores on the Microsoft COCO Caption dataset. The outstanding performance of the attention mechanism may be because current captioning encoders usually adopt graph convolutional networks (GCN) to represent the relationship between objects in the query image. Moreover, deep CNNs are employed to represent the objects and scenes in the image. The attention mechanism can work well on encoders, including CNN and GCN, and sequence-based decoders, forming a perfect framework. By integrating attention into the encoder-decoder image captioning framework, sentence generation will be conditioned on the hidden state calculated based on the attention mechanism, and captions can perform better.

Similarly, the emerging direction of novelty entities is also worthy of attention. We have observed that this type of method has better performance on the smaller Flickr dataset. Fu et al. [77] and Lu et al. [124] have achieved better performance on the Flickr30k dataset. Specifically, the obtained BLEU-1, BLEU-2, BLEU-3 and BLEU-4 scores are 0.720, 0.459, 0.319 and 0.285, respectively. In particular, the CIDEr score is 0.648, which outperforms other techniques. It indicated that the captions generated by other solutions could only capture the goals learned during the training process and thus cannot be extended to many novel scenes or objects, resulting in poor performance. Therefore, identifying novel targets and promoting vocabulary expansion of descriptive sentences is an exciting topic.

However, as a rising star, the Transformer-based method has achieved the best scores on BLEU-1, BLEU-2, and ROUGE, 0.822, 0.676, and 0.597 on the MS COCO dataset, respectively. This may be due to the RNN's inability to deal with the long-distance dependency problem, resulting in the gradients tend-

ing to disappear or explode after multiple propagation stages. Not only that, deep learning is representation learning, and numerous layers have significant advantages, while RNN is not easy to expand to multiple layers in engineering. These factors limit the further development of this type of framework. Even so, attention-guided image captioning combined with powerful visual encoders can still compete with subsequent Transformer-based methods in terms of performance. These methods are slower to train but generally smaller than Transformer-based methods. The Transformer-based method solves the problem of long-distance dependence of RNN. And because of its structural characteristics, it is easy to expand to deeper layers according to actual layout needs. We look forward to more researchers developing novel Transformer-based architectures.

Image captioning with high-level representations has advantages over other methods since that they can easily define the high-level representations of the image to limit the image content. However, their success largely depends on how accurately they estimate visual content. In particular, they explicitly use different convolutional neural networks to predict the most likely meaning of a given image. These methods have limited accuracy in practice, so if they fail to identify the most critical objects and their attributes, they cannot generate compelling descriptions.

Generating descriptions for images with novelties is also an exciting topic. These studies have the advantage of developing human-like descriptions because they can retrieve the most similar but independent novel descriptions from the extensive pre-trained datasets. However, generating these captions requires a similarity metric space to compare images or sentences. Compared with direct measures in visual or language space, these measures are difficult to define in the neural network embedding space. In addition, the metric space of training images or

TABLE 2 Method comparison on datasets Flickr8k and Flickr30k. In this table, B-n, ROUGE stand for BLEU-n and ROUGE-L, respectively

Category	Method	Flickr30k						Flickr8k					
		B-1	B-2	B-3	B-4	Meteor	ROUGE	CIDEr	B-1	B-2	B-3	B-4	Meteor
High-level representations	Karpathy et al. [43]	0.573	0.369	0.240	0.157	—	—	—	0.579	0.383	0.245	0.160	—
	Jia et al. [85]	0.647	0.459	0.318	0.216	—	0.202	—	0.646	0.446	0.305	0.206	—
Attention correction	Lee et al. [66]	0.274	—	—	0.286	0.403	0.300	0.419	—	—	—	—	—
	Xu et al. [51]	0.670	0.457	0.314	0.213	0.203	—	—	0.669	0.439	0.296	0.199	0.185
Novelties entity	Lu et al. [54]	—	—	—	—	—	—	—	0.677	0.494	0.354	0.251	0.204
	Fu et al. [77]	0.639	0.459	0.319	0.217	0.204	0.470	0.538	0.649	0.462	0.324	0.224	0.194
	Lu et al. [124]	0.720	—	—	0.285	0.231	—	0.648	—	—	—	—	—
													0.472

sentences requires a large amount of training data. On the bright side, such an embedding space can also be used for the reverse problem, searching the most description image for the query captions. This is not possible with methods based on generation or visual search.

As for the training strategy, sentence-level fine-tuning with reinforcement learning leads to significant performance improvement. Common training strategies are based on cross-entropy loss, aiming to operate at the word level and optimize the probability of each word in the natural sequence without considering the longer-range dependence between the generated terms. Therefore, there is a problem of violence bias [111] caused by the difference between the distribution of training data and the distribution of its own prediction words. Given the limitations of word-level training strategies, a significant improvement has been achieved by applying the reinforcement learning paradigm to the training image caption model. In this framework, the image captioning model is regarded as the agent of parameter decision strategy. At each time step, the agent executes a strategy to select an action, which is to predict the next word in the generated sentence. Once the end of the sequence is reached, the agent will be rewarded based on the generated sentence. The purpose of training is to optimize the agent parameters to maximize the expected reward. Many studies adopt this paradigm and explore different sequence-level indicators as rewards. Since the random strategy is difficult to improve in an acceptable time, the general programs need first to use cross-entropy or mask language model for pre-training, and then use reinforcement learning to fine-tune the stage sequence level metric as a reward. This ensures that the initial reinforcement learning strategy is more appropriate than the random strategy.

5 | RECENT TRENDS AND FUTURE RESEARCH DIRECTIONS

5.1 | Recent trends

Automatic image captioning is a relatively new task and has been made significant progress thanks to researchers in this field. The discussion in the previous subsections (Sections 2, 3, and 4) clarifies that each image description approach has its particular strengths and weaknesses. We believe that the current research mainly revolves around the following points:

First of all, more researchers are currently focusing on image captioning through the attention mechanism. Since this problem is very consistent with the attention mechanism, how to use the attention mechanism to generate image captions effectively will continue to be an essential research topic. Secondly, the transformer-based method has made preliminary attempts in this task and achieved outstanding performance. The use of transformers to improve image captioning will be promising. Furthermore, current image captioning studies are usually based on image-caption pairs. Still, even the MS COCO data set only contains a small part of the objects we encounter in real life. Since the generated description sentences can only capture the goals learned during the training process, they cannot be

TABLE 3 Method comparison on MS COCO dataset under the commonly used protocol

Category	Method	MS COCO							
		B-1	B-2	B-3	B-4	Meteor	ROUGE	CIDEr	SPICE
High-level representations	Karpathy et al. [43]	0.625	0.450	0.321	0.230	—	0.195	0.660	—
	Jia et al. [85]	0.670	0.491	0.358	0.264	0.227	—	0.813	—
	Yao et al. [45]	0.809	—	—	0.383	0.286	0.585	1.287	0.221
	Fan et al. [46]	0.818	—	—	0.40.8	0.295	0.592	1.353	0.225
Attention correction	Xu et al. [51]	0.718	0.504	0.357	0.250	0.230	—	—	—
	Lu et al. [54]	0.742	0.580	0.439	0.332	0.266	—	1.085	—
	Pan et al. [58]	0.817	0.668	0.526	0.407	0.299	0.597	1.353	0.238
Transformer-based	Dong et al. [63]	0.822	0.676	0.524	0.398	0.298	0.597	1.294	—
	Ji et al. [64]	0.816	0.665	0.519	0.397	0.294	0.591	1.303	—
	Liu et al. [62]	0.818	0.665	0.518	0.395	0.291	0.592	1.254	—
Novelties entity	Lu et al. [124]	0.759	—	—	0.349	0.274	—	1.089	0.201
	Yang et al. [68]	0.810	0.656	0.507	0.385	0.282	0.586	1.238	0.222
	Yao et al. [72]	0.816	0.662	0.515	0.393	0.288	0.590	1.279	0.223
	Chen et al. [73]	—	—	—	0.230	0.245	0.501	2.042	0.421
	Mahajan et al. [74]	0.777	0.620	0.473	0.354	0.270	0.563	1.144	0.204
Reinforcement learning	Liu et al. [82]	0.754	0.591	0.445	0.332	0.257	0.550	1.013	—
	Wu et al. [84]	0.790	0.630	0.482	0.363	0.277	0.571	1.179	0.216
	Deng et al. [89]	0.776	0.613	0.469	0.353	0.286	0.574	1.182	0.223

In this table, B-n, ROUGE stand for BLEU-n and ROUGE-L, respectively

extended to many new scenes and objects. Therefore, promoting the vocabulary expansion of descriptive sentences, identifying new targets, and integrating them into the description is also a problem researchers face. The fourth is an emerging direction of vision-language pre-training that significantly boosts image or video captioning. Despite having impressive vision-language pertaining with various deep models, the pertaining of a universal technical framework for different vision-language tasks remains challenging. Therefore, redevelopment and treatment of pre-trained models obtained from large datasets could extend the image captioning to other similar functions by transfer learning, such as vision-language understanding, short video captioning, and Visual Question Answering [125–127]. Finally is about training strategy. Since several commonly used evaluation metrics are more from machine translation, the optimization direction is more inclined to cross-entropy loss. Therefore, researchers are also putting more effort into making more muscular reward functions [66,90,112–114].

5.2 | Future research directions

As this survey demonstrated, the CV and NLP communities have witnessed a surge in interest in image description systems. With the latest developments in deep models, the quality of image description has been significantly improved. Due to the comprehensive application scenarios of image description tasks,

which can be applied to multimedia search, seeing chatbot, incident report for surveillance, and life assistance for the visually impaired, the value of researching image caption is obvious. Nevertheless, this task still faces a series of challenges. In the following, we discuss future directions that this line of research is likely to benefit.

First of all, with the rapid development of neural networks, a more robust network structure will undoubtedly improve the performance of image caption generation. However, most current research uses supervised training, which requires a large amount of labelled data and is limited by image description sentences [115]. Due to this limitation, how to adapt the model to out-of-domain data is an important research topic. Therefore, in the future, using weakly supervised or unsupervised data to improve image captioning research will be promising.

Secondly, more stylized image captions are also one of the future development directions. Stylized image captioning aims to generate captions not only semantically related to a given image but also consistent with given style descriptions. It includes emotional image captioning, personality image captioning. For emotional image captioning, people may recall similar emotions when they are in similar scenes and often use identical style phrases to express similar sentiments [116]. Therefore, emotional image captioning is necessary. Personalized image captioning aims to describe images with natural language caption with given personality characteristics [117]. The current image captioning studies has mainly considered

the factual setting for image captioning where the generated captions should faithfully present the visual content of images. A major limitation for this factual setting concern its failure to incorporate human factors, like personalities or traits, into the caption generation process. Therefore, in the future, people might prefer to produce engaging captions in which their personality characteristics are clearly expressed, and the theme concepts in the image may not be included in all its details.

Third, current models usually rely on direct representations of the descriptions they see during training, making the descriptions generated during testing very similar. It leads to many repetitions and limits the diversity of generated descriptions, making it challenging to reach human performance. The system that generates diversified captions repeats what has been seen and infers underlying semantics, which remains an open challenge. For example, Yang et al. [118] generated accurate descriptions for online fashion items, which is not only important for enhancing customers' shopping experience, but also for increasing online sales. Wang et al. [119] proposed a caption generator with instance-awareness and cross-hierarchy attention for remote sensing image.

Fourth, it was found from this survey that the earliest image description work used relatively small datasets, such as Pascal1K, Flickr30K, Flickr8K and so forth. Recently, the introduction of large datasets such as MS COCO and Visual Genome has made it possible to train more complex models. Nonetheless, the area is likely to benefit from more extensive and diversified datasets that share a common, unified, comprehensive vocabulary. Therefore, collecting larger and more comprehensive datasets and developing more general methods to generate natural descriptions across domains will continue to be an important research topic.

Finally, although more and more researchers have begun to work on image caption, experimental results show that the performance has not entirely surpassed that of humans. The evaluation methods currently used are not entirely consistent with human judgments about caption sentences. Therefore, the visual-text-based evaluation method is also one of the issues worth considering.

6 | CONCLUSION

In this survey, we review and analyze most of the studies on the image caption. According to each research method's passing characteristics and differences, we divide the image caption methods into different categories. This survey divides the image caption methods into three categories: the search-based approach, the language template-based approach, and the sequence-based approach using neural networks, and focuses on the refinement of the sequence-based approach. In terms of related evaluation measures, we introduced several evaluation measures that are most commonly used at present. Through the analysis and discussion of each type of evaluation's restrictive factors, we summarize the advantages and disadvantages. In addition, we provide a brief review of the widely used datasets and visually displayed them in the form of figures. After that,

state-of-the-art methods are compared on benchmark datasets. Finally, we put forward the discussion of image captioning for future research directions.

ACKNOWLEDGEMENTS

This work was supported by the 2021 Smart Campus Special Project of the Higher Agricultural College Branch of the China Educational Technology Association (No. C21ZD05).

AUTHOR CONTRIBUTIONS

Gaifang Luo: conceptualization; investigation; methodology; writing – original draft. **Lijun Cheng:** conceptualization; formal analysis; investigation. **Chao Jing:** investigation; writing – review and editing. **Can Zhao:** project administration; supervision. **Guozhu Song:** funding acquisition; resources; supervision; writing – review and editing

CONFLICT OF INTEREST

The authors declare no conflict of interest.

DATA AVAILABILITY STATEMENT

All data, models, and code generated or used during the study appear in the submitted article. Data sharing is not applicable to this article as no new data were created or analyzed in this study.

ORCID

Gaifang Luo  <https://orcid.org/0000-0002-6792-8853>

REFERENCES

- Farhadi, A., Hejrati, M., Sadeghi, M.A., et al.: Every picture tells a story: Generating sentences from images. In: European Conference on Computer Vision, pp. 15–29. (2010)
- Yang, J., Sun, Y., Liang, J., et al.: Image captioning by incorporating affective concepts learned from both visual and textual components. *Neurocomputing* (2019), 328, 56–68.
- Bernardi, R., Cakici, R., Elliott, D., et al.: Automatic description generation from images: A survey of models, datasets, and evaluation measures. *J. Artif. Intell. Research* (2016), 55, 409–442.
- Bai, S., An, S.: A survey on automatic image caption generation. *Neurocomputing* (2018), 311, 291–304.
- Ordonez, V., Kulkarni G., Berg, T.: Im2text: Describing images using 1 million captioned photographs. *Adv. Neural Inf. Process. Syst.* 24, 1143–1151 (2011)
- Hodosh, M., Young P., Hockenmaier, J.: Framing image description as a ranking task: Data, models and evaluation metrics. *J. Artif. Intell. Research* 47, 853–899 (2013)
- Bach, F.R., Jordan, I.: Kernel independent component analysis. *J. Mach. Learn. Res.* 3(Jul), 1–48 (2002)
- Hardoon, D.R., Szedmak S., Shawe-Taylor, J.: Canonical correlation analysis: An overview with application to learning methods. *Neural Comput.* 16(12), 2639–2664 (2004)
- Mason, R., Charniak, E.: Nonparametric method for data-driven image captioning. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, vol. 2 (Short Papers) (2014)
- Gupta, A., Verma, Y., Jawahar, C.: Choosing linguistics over vision to describe images. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 26 (2012)
- Kuznetsova, P., et al.: Treetalk: Composition and compression of trees for image descriptions. *Trans. Assoc. Comput. Linguist.* 2, 351–362 (2014)
- Socher, R., Fei-Fei, L.: Connecting modalities: Semi-supervised segmentation and annotation of images using unaligned text corpora. In: 2010

- IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 966–973 (2010)
13. Yang, Y., et al.: Corpus-guided sentence generation of natural images. In: Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing. (2011)
14. Felzenszwalb, P.F., et al.: Object detection with discriminatively trained part-based models. *IEEE Trans. Pattern Anal. Mach. Intell.* 32(9), 1627–1645 (2009)
15. Oliva, A., Torralba, A.: Modeling the shape of the scene: A holistic representation of the spatial envelope. *Int. J. Comput. Vision* 42(3), 145–175 (2001)
16. Dunning, T.E.: Accurate methods for the statistics of surprise and coincidence. *Comput. Linguist.* 19(1), 61–74 (1993)
17. Kulkarni, G., et al.: Babytalk: Understanding and generating simple image descriptions. *IEEE Trans. Pattern Anal. Mach. Intell.* 35(12), 2891–2903 (2013)
18. Li, S., et al.: Composing simple image descriptions using web-scale n-grams. In: Proceedings of the Fifteenth Conference on Computational Natural Language Learning (2011)
19. Mitchell, M., et al.: Midge: Generating image descriptions from computer vision detections. In: Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics (2012)
20. Koehn, P.: Europarl: A parallel corpus for statistical machine translation. In: MT Summit, vol. 5 (2005)
21. Fang, H., et al.: From captions to visual concepts and back. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2015)
22. Yatskar, M., Zettlemoyer L., Farhadi, A.: Situation recognition: Visual semantic role labeling for image understanding. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2016)
23. Ushiku, Y., et al.: Common subspace for model and similarity: Phrase learning for caption generation from images. In: Proceedings of the IEEE International Conference on Computer Vision (2015)
24. Cho, K., et al.: Learning phrase representations using RNN encoder-decoder for statistical machine translation. *arXiv:1406.1078* (2014)
25. Lipton, Z.C., Berkowitz, J., Elkan, C.: A critical review of recurrent neural networks for sequence learning. *Comput. Sci. abs/1506.00019*, (2015). <http://arxiv.org/abs/1506.00019>
26. Zaremba, W., Sutskever I., Vinyals, O.: Recurrent neural network regularization. *arXiv:1409.2329* (2014)
27. Sutskever, I., Vinyals, O., Le, Q.V.: Sequence to sequence learning with neural networks. *Adv. Neural Inf. Process. Syst.* 2, 3104–3112 (2014)
28. Gers, F.A., Eck D., Schmidhuber, J.: Applying LSTM to time series predictable through time-window approaches. In: *Neural Nets WIRN Vietri-01*, pp. 193–200. Springer, London (2002)
29. Sak, H., Senior A., Beaufays, F.: Long short-term memory based recurrent neural network architectures for large vocabulary speech recognition. *Comput. Sci. abs/1402.1128*, 338–342 (2014). <http://arxiv.org/abs/1402.1128>
30. Socher, R., et al.: Grounded compositional semantics for finding and describing images with sentences. *Trans. Assoc. Comput. Linguist.* 2, 207–218 (2014)
31. Le, Q.V.: Building high-level features using large scale unsupervised learning. In: 2013 IEEE International Conference on Acoustics, Speech and Signal Processing (2013)
32. Karpathy, A., Joulin, A., Fei-Fei, L.: Deep fragment embeddings for bidirectional image sentence mapping. In: Proceedings of the Twenty Seventh Advances in Neural Information Processing Systems (NIPS), vol. 3, pp. 1889–1897 (2014)
33. Girshick R., Donahue J., Darrell T., Malik J.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 580–587. Columbus, OH (2014)
34. Lebrecht, R., Pinheiro P., Collobert R.: Phrase-based image captioning. In: International Conference on Machine Learning (2015)
35. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. *Comput. Sci.* (2014). <http://arxiv.org/abs/1409.1556v6>
36. Mikolov, T., Chen, K., Corrado, G., et al.: Distributed representations of words and phrases and their compositionality. *Adv. Neural Inf. Process. Syst.* 26, 3111–3119 (2013)
37. Mnih, A., Kavukcuoglu, K.: Learning word embeddings efficiently with noise-contrastive estimation. *Adv. Neural Inf. Process. Syst.* 26, 2265–2273 (2013). <https://proceedings.neurips.cc/paper/2013/file/db2b4182156b2f1f817860ac9f409ad7-Paper.pdf>
38. Mikolov, T., Chen, K., Corrado, G., et al.: Efficient Estimation of Word Representations in Vector Space. *Comput. Sci.* (2013). <http://arxiv.org/abs/1301.3781v3>
39. Kiros R., Salakhutdinov R., Zemel R.S.: Unifying Visual-Semantic Embeddings with Multimodal Neural Language Models. In: Proceedings of the NIPS Workshop on International Machine Learning Society (2014)
40. Kiros, R., Salakhutdinov, R., Zemel, R.: Multimodal neural language models. In: International Conference on Machine Learning, pp. 595–603 (2014)
41. Wu, Q., Shen, C. & Liu, L.: What value do explicit high level concepts have in vision to language problems? In: IEEE conference on computer vision and pattern recognition, pp. 203–212 (2016)
42. Russakovsky O., Deng J., Su H., Krause J., Satheesh S., Ma S., Huang Z., Karpathy A., Khosla A., Bernstein M., Berg A.C., Fei-Fei, L.: ImageNet large scale visual recognition challenge. *Int. J. Comput. Vision (IJCV)* 115(3), 211–252 (2015)
43. Karpathy, A. & Fei-Fei L.: Deep visual-semantic alignments for generating image descriptions. In: Proceedings of the IEEE Conference On Computer Vision And Pattern Recognition (2015)
44. Yao, T., Pan, Y., Li, Y., et al.: Boosting image captioning with attributes. In: IEEE International Conference on Computer Vision, pp. 4904–4912 (2017)
45. Yao, T., et al.: Exploring visual relationship for image captioning. In: Proceedings of the European Conference On Computer Vision (ECCV) (2018)
46. Fan, Z., Wei, Z., Wang, S., et al.: TCIC: Theme concepts learning cross language and vision for image captioning. *arXiv:2106.10936* (2021)
47. Song, Z. & Zhou, X.: Exploring explicit and implicit visual relationships for image captioning. In: 2021 IEEE International Conference on Multimedia and Expo (ICME) (2021)
48. Liu, F., Gao, M., Zhang, T., et al.: Exploring semantic relationships for unpaired image captioning. *arXiv:2106.10658* (2021)
49. Vinyals, O., et al.: Show and tell: A neural image caption generator. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2015)
50. Szegedy, C., Liu, W., Jia, Y., et al.: Going deeper with convolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2015): 1–9
51. Xu, K., et al.: Show, attend and tell: Neural image caption generation with visual attention. In: International Conference on Machine Learning (2015)
52. Yang Z., Yuan Y., Wu Y., Cohen W.W., Salakhutdinov R.: Review networks for caption generation. In: Proceedings of the Twenty Seventh Advances in Neural Information Processing Systems (NIPS) (2016)
53. Chen, L., Zhang, H., Xiao, J., et al.: SCA-CNN: Spatial and channel-wise attention in convolutional networks for image captioning. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016)
54. Lu, J., Xiong, C., Parikh, D., et al.: Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017)
55. Anderson, P., et al.: Bottom-up and top-down attention for image captioning and visual question answering. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2018)
56. Jiang W., Ma L., Jiang Y.-G., Liu W., Zhang T.: Recurrent fusion network for image captioning. In: Proceedings of the European Conference on Computer Vision (ECCV) (2018)

57. You, Q., Jin H., Wang Z., Fang C., Luo J.: Image captioning with semantic attention. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4651–4659 (2016)
58. Pan, Y., Yao T., Li Y., Mei T.: X-linear attention networks for image captioning. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10971–10980 (2020)
59. Wang, J., Xu, W., Wang, Q., et al.: Group-based distinctive image captioning with memory attention. *arXiv:2108.09151* (2021)
60. Dosovitskiy, A., Beyer, L., Kolesnikov, A., et al.: An Image is Worth 16X 16 Words: Transformers for Image Recognition at Scale (2020)
61. Liu, Z., Lin, Y., Cao, Y., et al.: Swin transformer: Hierarchical vision transformer using shifted windows. *arXiv:2103.14030* (2021)
62. Liu, W., Chen, S., Guo, L., et al.: CPTR: Full transformer network for image captioning. *arXiv:2101.10804* (2021)
63. Dong, X., Long, C., Xu, W., et al.: Dual graph convolutional networks with transformer and curriculum learning for image captioning. *arXiv:2108.02366* (2021)
64. Ji, J., Luo, Y., Sun, X., et al.: Improving image captioning by leveraging intra-and inter-layer global representation in transformer network. In: *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, pp. 1655–1663 (2021)
65. Sariyildiz, M.B., Perez, J., Larlus, D.: Learning visual representations with caption annotations. In: *Computer Vision—ECCV 2020: 16th European Conference*, pp. 153–170. Glasgow, UK, 23–28 August 2020
66. Lee, H., Yoon, S., Dernoncourt, F., et al.: UMIC: An unreferenced metric for image captioning via contrastive learning. *arXiv:2106.14019* (2021)
67. Yang, X., Liu, Y., Wang, X.: ReFormer: The relational transformer for image captioning. *arXiv:2107.14178* (2021)
68. Yang, X., et al.: Auto-encoding scene graphs for image captioning. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2019)
69. Ren, S., He K., Girshick R., Sun J.: Faster R-CNN: Towards real-time object detection with region proposal networks. *IEEE Trans. Pattern Anal. Mach. Intell.* 39(6), 1137–1149 (2016)
70. Milo, R., Shen-Orr S., Itzkovitz S., Kashtan N., Chklovskii D., Alon U.: Network motifs: Simple building blocks of complex networks. *Science* 298(5594), 824–827 (2002)
71. Anderson, P., Fernando, B., Johnson, M., et al.: Guided open vocabulary image captioning with constrained beam search. In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 936–945 (2017)
72. Yao, T., et al.: Hierarchy parsing for image captioning. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2019)
73. Chen, S., et al.: Say as you wish: Fine-grained control of image caption generation with abstract scene graphs. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2020)
74. Mahajan, S., Roth, S.: Diverse image captioning with context-object split latent spaces. *arXiv:2011.00966* (2020)
75. Li, Y., et al.: Pointing novel objects in image captioning. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2019)
76. Hendricks, L.A., Venugopalan S., Rohrbach M., Mooney R., Saenko K., Darrell T.: Deep compositional captioning: Describing novel object categories without paired training data. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–10 (2016)
77. Fu, K., Jin, J., Cui, R., et al.: Aligning where to see and what to tell: Image captioning with region-based attention and scene-specific contexts. *IEEE Trans. Pattern Anal. Mach. Intell.* 39(12), 2321–2334 (2016)
78. Zheng, Y., Li Y., Wang S.: Intention oriented image captions with guiding objects. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8395–8404 (2019)
79. Demirel, B., Cinbis, R.G.: Detection and captioning with unseen object classes. *arXiv:2108.06165* (2021)
80. Wang, J., Xu, W., Wang, Q., et al.: Compare and reweight: Distinctive image captioning using similar images sets. In: *Proceedings of the European Conference on Computer Vision*, pp. 370–386 (2020)
81. Rennie, S.J., et al.: Self-critical sequence training for image captioning. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2017)
82. Liu, S., Zhu Z., Ye N., Guadarrama S., Murphy K.: Improved image captioning via policy gradient optimization of spider. In: *Proceedings of the IEEE International Conference on Computer Vision*, pp. 873–881 (2017)
83. Wang, X., et al.: Reinforced cross-modal matching and self-supervised imitation learning for vision-language navigation. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2019)
84. Wu, J., Chen, T., Wu, H., et al.: Fine-grained image captioning with global-local discriminative objective. *IEEE Trans. Multimedia* 23, 2413–2427 (2021)
85. Jia, X., Gavves E., Fernando B. & Tuytelaars T.: Guiding the long-short term memory model for image caption generation. In: *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2407–2415 (2015)
86. Johnson, J., Karpathy A., Fei-Fei L.: Denscap: Fully convolutional localization networks for dense captioning. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4565–4574 (2016)
87. Xu, G., Niu, S., Tan, M., et al.: Towards Accurate Text-based Image Captioning with Content Diversity Exploration. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12637–12646 (2021)
88. Venugopalan, S., Rohrbach M., Donahue J., Mooney R., Darrell T., Saenko K.: Sequence to sequence-video to text. In: *Proceedings of the IEEE International Conference on Computer Vision*, pp. 4534–4542 (2015)
89. Deng, C., Ding, N., Tan, M., et al.: Length-controllable image captioning. In: *Computer Vision—ECCV 2020: 16th European Conference*, pp. 712–729. Glasgow, UK, 23–28 August 2020
90. Chen, A., Huang, X., Lin, H., et al.: Towards annotation-free evaluation of cross-lingual image captioning. In: *Proceedings of the 2nd ACM International Conference on Multimedia in Asia*, pp. 1–7 (2021)
91. Papineni, K., Roukos S., Ward T., Zhu W.J.: Bleu: A method for automatic evaluation of machine translation. In: *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pp. 311–318 (2002)
92. Lin, C.-Y., Hovy E.: Automatic evaluation of summaries using n-gram co-occurrence statistics. In: *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 150–157 (2003)
93. Banerjee, S., Lavie A.: METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In: *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pp. 65–72 (2005)
94. Vedantam, R., Lawrence Zitnick C., Parikh A.: Cider: Consensus-based image description evaluation. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4566–4575 (2015)
95. Anderson, P., Fernando B., Johnson M., Gould S.: Spice: Semantic propositional image caption evaluation. In: *Proceedings of the European Conference on Computer Vision*, pp. 382–398 (2016)
96. Cui, Y., Yang G., Veit A., Huang X., Belongie S.: Learning to evaluate image captioning. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5804–5812 (2018)
97. Klein, D., Manning D.: Accurate unlexicalized parsing. In: *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pp. 423–430 (2003)
98. Lin, T.-Y., Maire M., Belongie S., Hays J., Perona P., Ramanan D., Dollár P., Zitnick C.L.: Microsoft coco: Common objects in context. In: *Proceedings of the European Conference on Computer Vision*, pp. 740–755 (2014)
99. Young, P., Lai A., Hodosh M., Hockenmaier J.: From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Trans. Assoc. Comput. Linguist.* 2, 67–78 (2014)
100. Krishna, R., Zhu Y., Groth O., Johnson J., Hata K., Kravitz J., Chen S., et al.: Visual genome: Connecting language and vision using

- crowdsourced dense image annotations. *Int. J. Comput. Vision* 123(1), 32–73 (2017)
101. Rashtchian, C., Young P., Hodosh M., Hockenmaier J.: Collecting image annotations using amazon's mechanical turk. In: *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pp. 139–147 (2010)
 102. Everingham, M., Gool L.V., Williams C.K.I., Winn J., Zisserman A.: The pascal visual object classes (VOC) challenge. *Int. J. Comput. Vision* 88(2), 303–338 (2010)
 103. Kleppe, M., Elliott, D., Doing Visual Big Data—Creating the KBK-1M dataset containing 1, 6 million newspaper images available for researchers
 104. Elliott, D., Keller F.: Image description using visual dependency representations. In: *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pp. 1292–1302 (2013)
 105. Zitnick, C.L., Parikh D.: Bringing semantics into focus using visual abstraction. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3009–3016 (2013)
 106. Zitnick, C.L., Parikh D., Vanderwende L.: Learning the visual interpretation of sentences. In: *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1681–1688 (2013)
 107. Chen, J., Kuznetsova P., Warren D., Choi Y.: Déjà image-captions: A corpus of expressive descriptions in repetition. In: *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 504–514 (2015)
 108. Sharma, P., Ding N., Goodman S., Soricut R.: Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, vol 1 (Long Papers), pp. 2556–2565 (2018)
 109. Yoshikawa Y., Shigeto Y., Takeuchi A.: STAIR captions: Constructing a large-scale japanese image caption dataset. In: *Annual Meeting of the Association for Computational Linguistics (ACL) (Short Paper)*, (2017)
 110. Chen, F., Ji R., Sun X., Wu Y., Su J.: Groupcap: Group-based image captioning with structured relevance and diversity constraints. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1345–1353 (2018)
 111. Ranzato, M.A., Chopra, S., Auli, M., et al.: Sequence level training with recurrent neural networks. *arXiv:1511.06732* (2015)
 112. Wang, Z., Feng, B., Narasimhan, K., et al.: Towards unique and informative captioning of images. In: *Computer Vision—ECCV 2020: 16th European Conference*, pp. 629–644. Glasgow, UK, 23–28 August 2020
 113. Hessel, J., Holtzman, A., Forbes, M., et al.: CLIPScore: A reference-free evaluation metric for image captioning. *arXiv:2104.08718* (2021)
 114. Lee, H., Scialom, T., Yoon, S., et al. QACE: Asking questions to evaluate an image caption. *arXiv:2108.12560* (2021)
 115. Chen, N., Pan, X., Chen, R., et al.: Distributed attention for grounded image captioning. *arXiv:2108.01056* (2021)
 116. Li, G., Zhai, Y., Lin, Z., et al.: Similar scenes arouse similar emotions: Parallel data augmentation for stylized image captioning. *arXiv:2108.11912* (2021)
 117. Nguyen, M.T., Phung, D., Hoai, M., et al.: Structural and functional decomposition for personality image captioning in a communication game. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pp. 4587–4593 (2020)
 118. Yang, X., Zhang, H., Jin, D., et al.: Fashion captioning: Towards generating accurate descriptions with semantic rewards. In: *Computer Vision—ECCV 2020: 16th European Conference*, pp. 1–17. Glasgow, UK, 23–28 August 2020
 119. Wang, C., Jiang, Z., Yuan, Y.: Instance-aware remote sensing image captioning with cross-hierarchy attention. In: *IGARSS 2020-2020 IEEE International Geoscience and Remote Sensing Symposium*, pp. 980–983 (2020)
 120. Elhagry, A., Kadaoui, K.: A thorough review on recent deep learning methodologies for image captioning. *arXiv:2107.13114* (2021)
 121. Al Sulaimi, M., Ahmad, I., Jeragh, M.: Deep image captioning survey: A resource availability perspective. In: *2021 29th Conference of Open Innovations Association (FRUCT)*, pp. 3–13 (2021)
 122. Sharma, H., Agrahari, M., Singh, S.K., et al.: Image captioning: A comprehensive survey. In: *2020 International Conference on Power Electronics & IoT Applications in Renewable Energy and its Control (PARC)*, pp. 325–328 (2020)
 123. Yao, T., Pan, Y., Li, Y., et al.: Incorporating copying mechanism in image captioning for learning novel objects. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6580–6588 (2017)
 124. Lu, J., Yang, J., Batra, D., et al.: Neural baby talk. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 7219–7228 (2018)
 125. Zhou, L., Palangi, H., Zhang, L., et al.: Unified vision-language pre-training for image captioning and VQA. In: *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, pp. 13041–13049 (2020)
 126. Pan, Y., Li, Y., Luo, J., et al.: Auto-captions on GIF: A large-scale video-sentence dataset for vision-language pre-training. *arXiv:2007.02375* (2020)
 127. Li, Y., Pan, Y., Yao, T., et al.: Scheduled sampling in vision-language pre-training with decoupled encoder-decoder network. In: *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, pp. 8518–8526 (2021)

How to cite this article: Luo, G., Cheng, L., Jing, C., Zhao, C., Song, G.: A thorough review of models, evaluation metrics, and datasets on image captioning. *IET Image Process.* 16, 311–332 (2022).
<https://doi.org/10.1049/ipr2.12367>