VIETNAM GENERAL CONFEDERATION OF LABOR

TON DUC THANG UNIVERSITY

INFORMATION TECHNOLOGY FACULTY

INTRODUCTION TO DEEP LEARNING

MIDTERM REPORT

IMAGE CAPTIONING

Instructor:	**Prof. Lê Anh Cường**
Student 1:	**Đỗ Phạm Quang Hưng - 520K0127**
Student 2:	**Lê Phước Thịnh - 520K0343**
Student 3:	**Chibuike Timothy Benedict - 5190K0078**

HO CHI MINH CITY, 2023

Table of Contents

Abstract

Image captioning means generating 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.

I. Introduction

It is not difficult to quickly recognize and understand an image by capturing 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 color, 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 the image. 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.

[1] 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)

[2] 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.

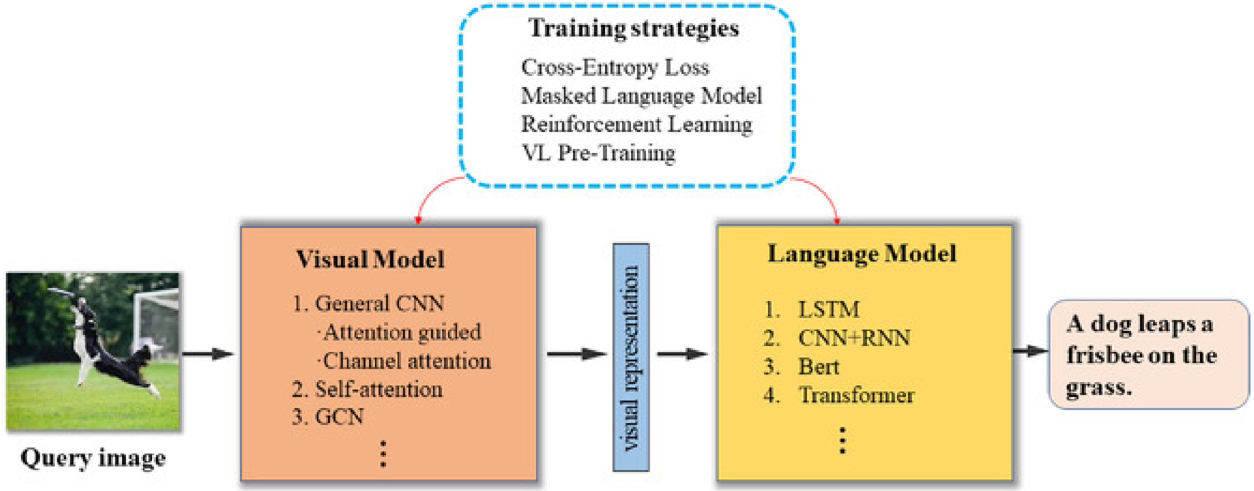


FIGURE 1. Overview of automatic image captioning tasks and a simple example of the most relevant approaches

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 to invest more energy to make it prosperous. At first, researchers adopted search-based methods and 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[3,4] 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. There is still a lack

[3] 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.
[4] Bai, S., An, S.: A survey on automatic image caption generation. Neurocomputing (2018), 311, 291–304.

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 modeling 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.

II. 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.

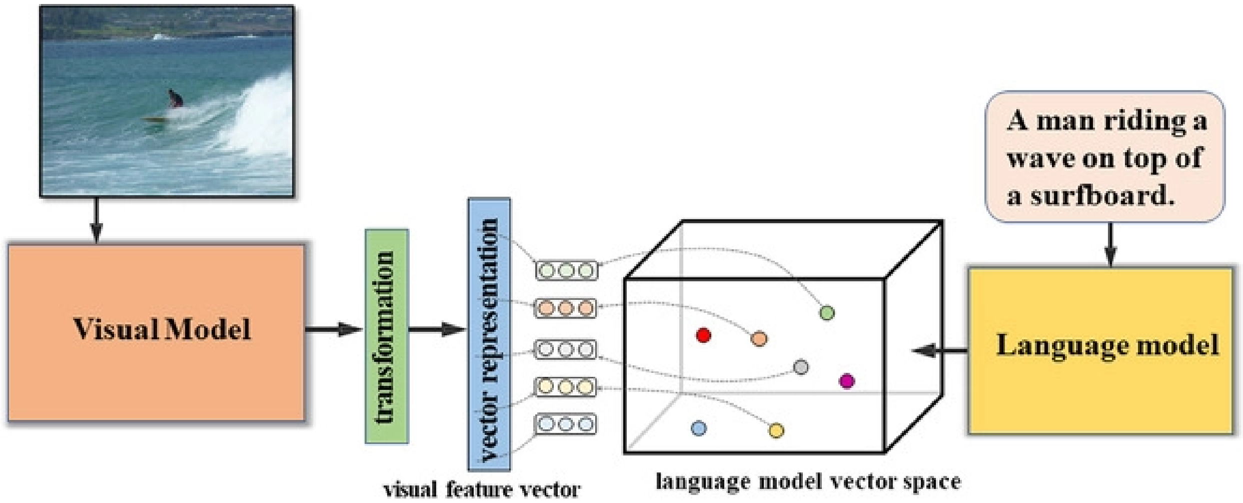


FIGURE 2. Search-based approach and language template-based approach research framework

2.1. Search-based Method

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.

- *Ordonez et al.*[5] grabbed a large number of images from the Internet and manually labeled 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

[5] Ordonez, V., Kulkarni G., Berg, T.: Im2text: Describing images using 1 million captioned photographs. Adv. Neural Inf. Process. Syst. 24, 1143–1151 (2011)

[6] 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)

[7] Bach, F.R., Jordan, I.: Kernel independent component analysis. J. Mach. Learn. Res. 3(Jul), 1–48 (2002)

[8] 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)

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.

[9] 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)

[10] 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)

[11] Kuznetsova, P., et al.: Treetalk: Composition and compression of trees for image descriptions. Trans. Assoc. Comput. Linguist. 2, 351–362 (2014)

[12] 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)

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 captions 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.

[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. Cosmput. Linguist. 19(1), 61–74 (1993)

- *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 CRF 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

[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)

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 captions* 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.

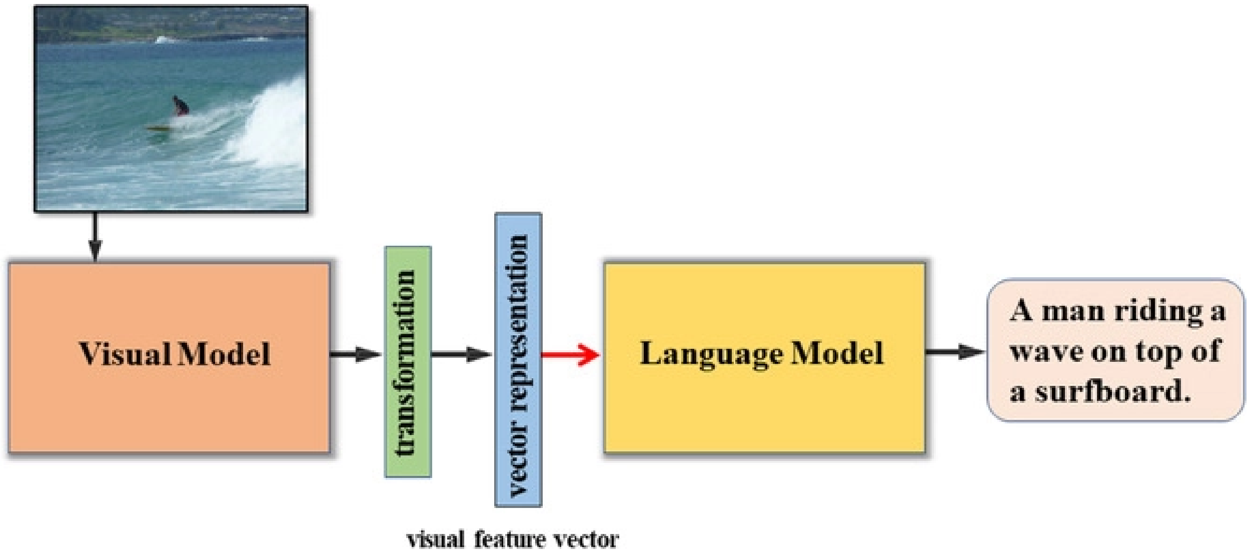
[21] Fang, H., et al.: From captions to visual concepts and back. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2015)

[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)

III. THE RECENT DEEP LEARNING METHODS

The relatively early researchers 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, 26, 27, 28, 29], it is known as the Sequence-based approach. The overall framework of the sequence-based approach is illustrated in Figure 3.



[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).

[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).

FIGURE 3. Sequence-based approach research framework

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 tasks, 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.

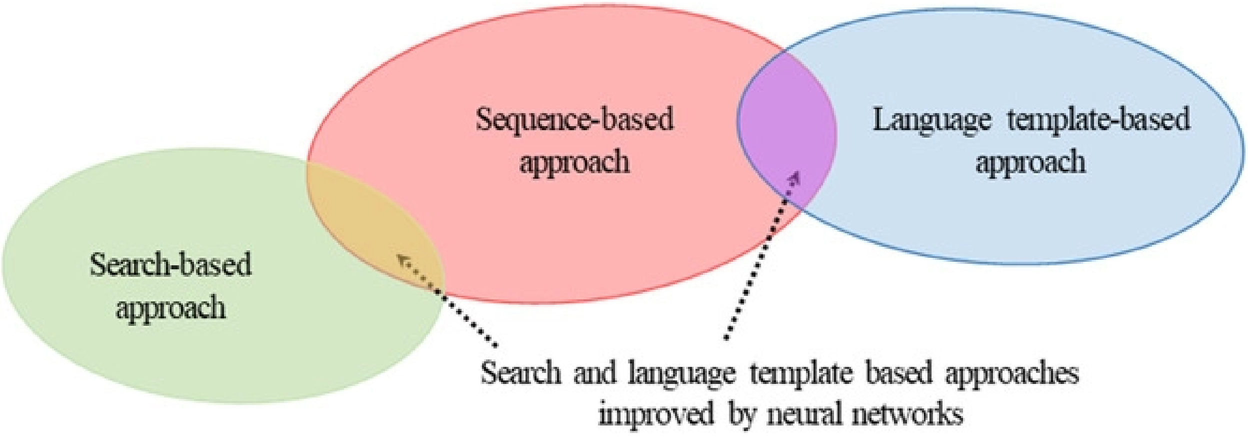

Sequence-based approach
Language template-based approach
Search-based approach
Search and language template based approaches improved by neural networks

FIGURE 4. Schematic diagram of the relationship between the three methods

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.

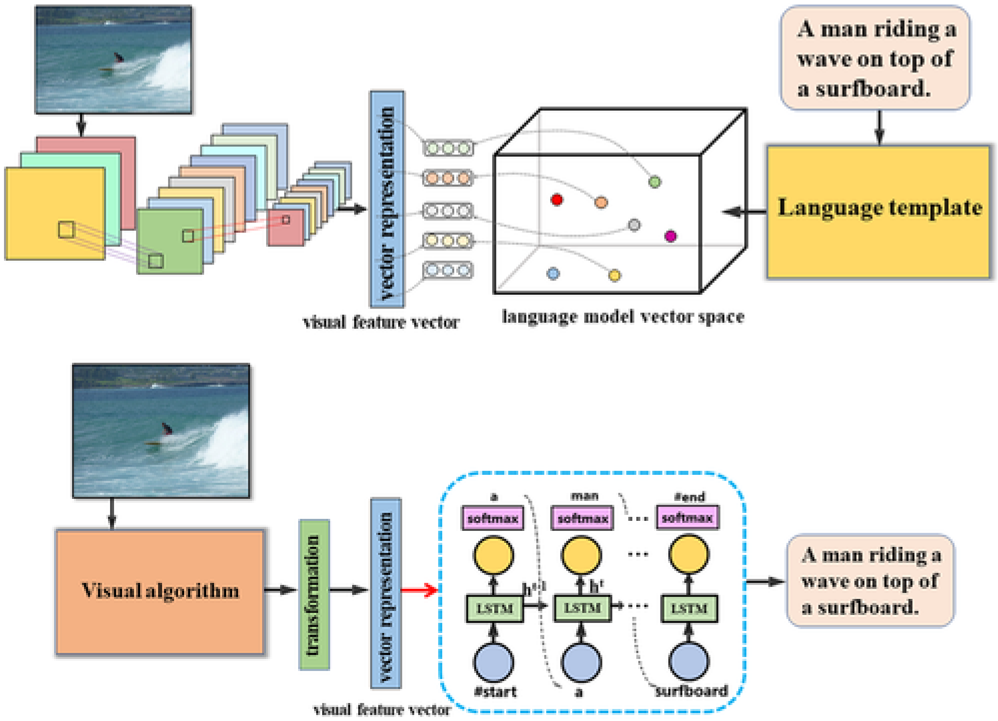


FIGURE 5. The framework of search and language template based approaches improved by neural networks

- *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 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

[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]] 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] [37] [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].

[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] Lebret, 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).

[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).

[38] Mikolov, T., Chen, K., Corrado, G., et al.: Efficient Estimation of Word Representations in Vector Space. Comput. Sci. (2013).

[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 MachineLearningSociety (2014)

[40] Kiros, R., Salakhutdinov, R., Zemel, R.: Multimodal neural language models. In: International Conference on Machine Learning, pp. 595–603 (2014)

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 separation 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.

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.

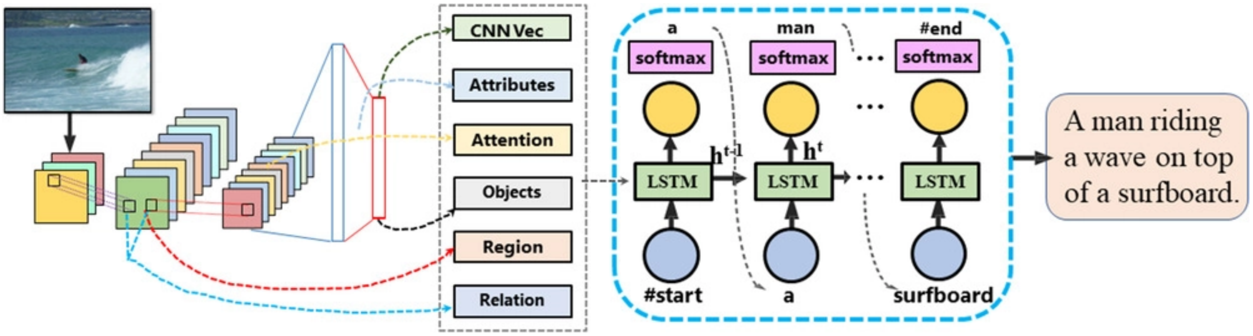


FIGURE 6. Image captioning with high-level representations methods

- *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

[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)

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.

[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)

The general method based on semantic relation graphs 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.

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.

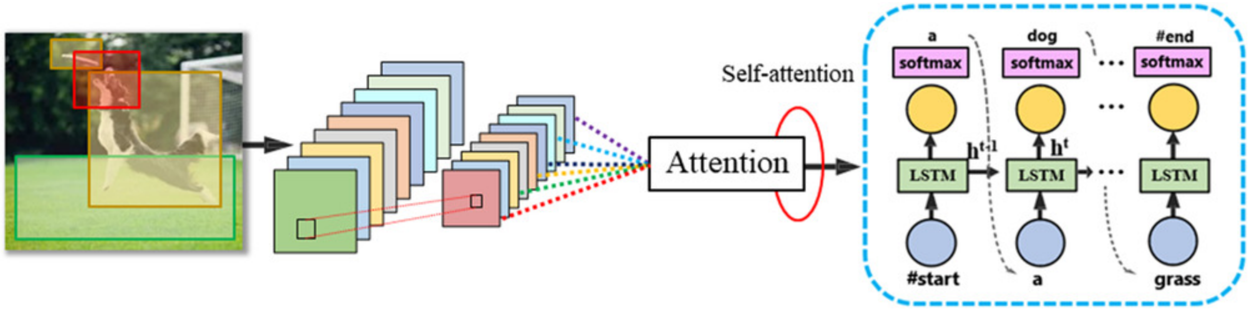


FIGURE 7. General framework of enhanced image captioning with attention correction

- *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

[47] Fan, Z., Wei, Z., Wang, S., et al.: TCIC: Theme concepts learning cross language and vision for image captioning. arXiv:2106.10936 (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)

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

[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

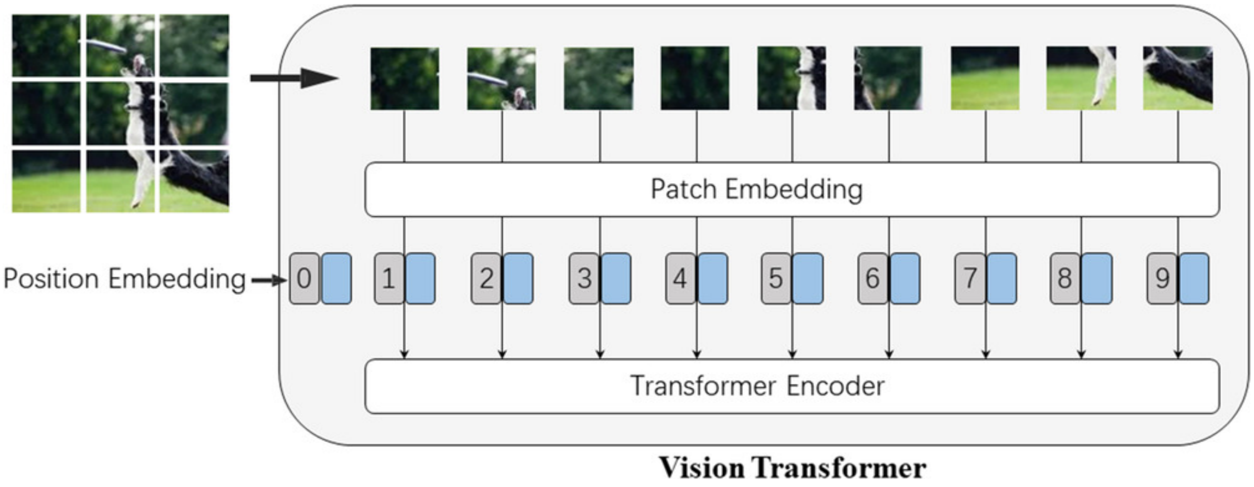
[51] Xu, K., et al.: Show, attend and tell: Neural image caption generation with visual attention. In: International Conference on Machine Learning (2015)Xu, K., et al.: Show, attend and tell: Neural image caption generation with visual attention. In: International Conference on Machine Learning (2015)

[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)

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 55 56]. 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, transformers have shown good performance when dealing with serialized information. More importantly, the transformer has been recognized as the latest technology for sequence modeling 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 [57 58]. 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.



[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] Dosovitskiy, A., Beyer, L., Kolesnikov, A., et al.: An Image is Worth 16×16 Words: Transformers for Image Recognition at Scale (2020)

[58] Liu, W., Chen, S., Guo, L., et al.: CPTR: Full transformer network for image captioning. arXiv:2101.10804 (2021).

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

- *Dong et al.*[59] 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 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.

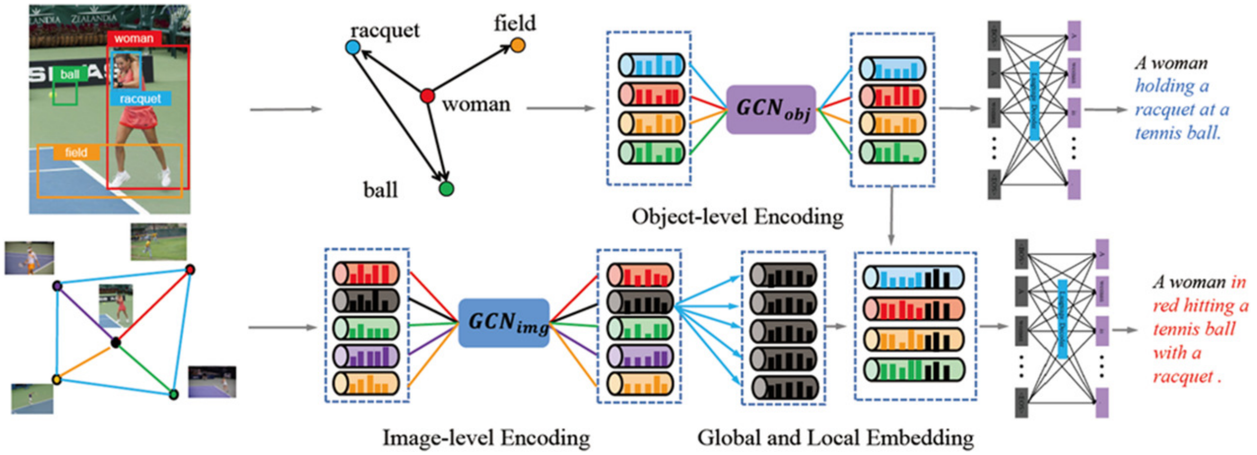

woman
ball
racquet
field
racquet
field
woman
ball
GCN_{obj}
Object-level Encoding
A woman holding a racquet at a tennis ball.
GCN_{img}
A woman in red hitting a tennis ball with a racquet .
Image-level Encoding
Global and Local Embedding

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

- *Sariyildiz et al.*[60] introduced transformer-based image **conditional masking language modeling** (ICMLM) for learning the visual representation of image-caption pairs. *Lee et al.*[61] 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[62] a novel transformer, **ReFormer**, adapted to generate features embedded in relational information and clearly express the paired relations between objects in images.

[59] Dong, X., Long, C., Xu, W., et al.: Dual graph convolutional networks with transformer and curriculum learning for image captioning. arXiv:2108.02366 (2021)

[60] 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

[61] Lee, H., Yoon, S., Dernoncourt, F., et al.: UMIC: An unreferenced metric for image captioning via contrastive learning. arXiv:2106.14019 (2021)

[62] Yang, X., Liu, Y., Wang, X: ReFormer: The relational transformer for image captioning. arXiv:2107.14178 (2021)

IV. 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.

4.1.1. 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:

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.

4.1.2. 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.

4.1.3. 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.

4.1.4. 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.

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 dataset, 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 labeling platform. Each image has five sentences as description.

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

V. RECENT TRENDS AND FUTURE RESEARCH DIRECTIONS

Automatic image captioning is a relatively new task and has made significant progress thanks to researchers in this field. The discussion in the previous subsections (Sections 2, 3, 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 dataset 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 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. *Finally*, it 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

References

1. A thorough review of models, evaluation metrics, and datasets on image captioning - Luo - 2022 - IET Image Processing - Wiley Online Library