Automatic Caption Generation from Image: A Comprehensive Survey

Ms. Yugandhara A. Thakare[1] , Prof. K. H. Walse[2]

[1] P.G Dept. of Computer Science, SGBAU, Amravati, MH., India.
[2] Dept. of Computer Science & Engineering, Anuradha Engineering College, Chikhli, MH., India

Abstract

In recent years, image captioning has progressively received attention by various researchers due to the speedy progress of AI, and become a remarkable task. Image caption automatically generates textual description in line with the contents ascertained in a picture, which is the association of knowledge of computer vision (CV) and natural language processing (NLP). This paper gives a precise view of different architectures, benefits, and limitations of these architectures. It also provides different datasets and performance assessment criteria used in this field. Finally, this review paper discusses several unsolved issues in the image captioning task.

Keywords
Image captioning, Computer Vision, CNN, RNN, LSTM

1. Introduction

Deep learning has received the greatest attention over the last decade as a result of its capacity to expand and solve problems that were not solved previously. Explaining images with captions has impacted many applications and it has become an important area of research for many people, who connect via media as a language. This would lead to creating a need for variation in architectures that can be converted to sentences. From a CV & AI perspective, there is a need to bridge the semantic gap among low-level visual and high-level abstract data. Image captions are a common technique for filling semantic gaps in many real-world applications.

Image captioning has recently become an important area of computer vision and attracted the interest of researchers. Image captioning automatically generates textual description of contents from an image in a syntactically, semantically and meaningful or expressive way, indirectly it tells us what the picture is all about. The job of image captioning is straightforward – a single sentence should be generated as a output which describes what is in fact presented in the image – the things existing, the activities being performed, the correlation among the things and their properties etc. This survey purposes to present a broad summary of image caption generation models and recent developments in these architectures.

The remainder of the paper is arranged as follows. Section 2 discusses image captioning architecture and models as well as recent developments in it. Section 3 discusses the challenges of image captioning. Section 4 introduces benchmark datasets and compares the results of various models. Different evaluation methods are discussed in Section 5. Section 6 outlines the review of existing work and suggestions for future direction.

2.   Classical Models

We conducted a wide review of the literature on image captioning and classified the models according to their approaches for language generation.

2.1 Retrieval based caption generation model

Earlier Retrieval based image captioning task was common work for captioning purpose. In this method one or set of sentences are retrieved for the given query image from the pre-specified pool of sentences. Image caption generated is either a combination of sentence from the retrieval ones or it can be a description which is already presented.

[1] The author of this paper offered a novel strategy which does not rely on any classifiers, object detectors or any handwritten rules for generating image description like human. The Stanford CoreNLP toolkit is used to process sentences in the dataset, from which a list of phrases for each image is produced. This method generates a set of query images and generates a description for the query image by doing image retrieval based on global image features. Here the model is being trained to predict the relevance of sentences. Then this model is used for selecting phrases from ones which are associated with the retrieved images. At last, based on selected relevant phrases, description sentence is generated.

[2] Image captioning was exhibited as a ranking task by Hodosh. To integrate images and text into a common region, the author have used Kernel Canonical Correlation technique, in which the maximum training images and captions are correlated.Similarities between images and descriptions are calculated in the novel common area to generate the top-ranked query image description.

[3] The author presented a model that uses deep, multimodal embedding of visual and natural language data to retrieve images and texts in the both directions. Their model embeds sentence fragments and objects present in image into a common area.

2.2 Template based caption generation model

The template-based captioning paradigm has a long history, where each piece of sentence is aligned with the words received from the image content and the description is constructed using the pre-defined language templates.

[4] Author introduced novel method for generating a short description sentence from an image. The proposed architecture computes score from image, which is used to link to a sentence description to given image. This score is calculated by making the comparison of the assessment of meaning obtained from both the image and the phrase. In this approach, for the given image, based on the calculated score one can search for the best caption with respect to image from the large set and vice versa. As a result, both image description and image illustration are generated using the same approach. This space of meanings is one of the essential factors in their model, as it is found among the space of sentences and images. This approach provides a simplified sentence model for sentence generation

[5] Author proposed an architecture which is proficient for generating natural description with simple and true to the image as possible. Proposed architecture generates caption based on syntactic trees instead of using fixed template which generates one kind of word. Such method creates data driven model which is able to automatically parse and train on unrestricted amount of text and generate description for detected objects in proper manner. This "Midge" system is capable of deciding what will be objects and subjects in the description and false detection can be filter out from this based on statistics from co-occurrence of words and generate description as concise as possible. The drawback of this system is that, it often detects incorrect objects and missed the silent and un-likely objects from the image.

[6] The author's method is divided into two steps.The clatter output generated by computer vision systems is smoothed in content planning using statistics gathered from visually descriptive natural

language. After choosing the contents by conditional random field to be used in generation succeeding surface realization step is used for searching words to describe the content chosen by previous step to generate description based on sentence template. Their approach is based on a graph, with nodes representing objects, attributes and the relationships among them.

2.3 Deep neural network based caption generation model

In the early work, retrieval and template-based image captioning systems were widely accepted. Deep neural networks were recently used to produce image descriptions, which was a major breakthrough in deep learning.

- **Encoder-Decoder image captioning Framework**

Recent advancements in multimodal learning and machine translation for image caption generation have sparked interest in the encoder-decoder technique. The encoder-decoder framework's general approach is to encode an image using an encoder neural network to an intermediate level, which is then sent as input to a RNN decoder, which will subsequently generate output in terms of sentences word by word. The basic model of the encoder-decoder image captioning structure is shown in the diagram below.
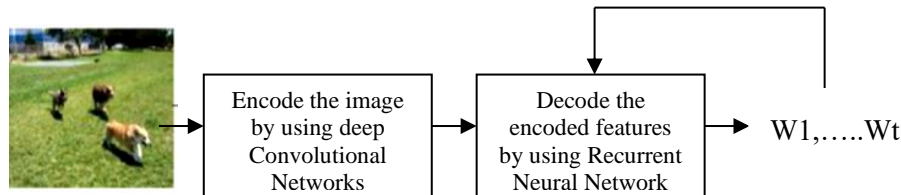
Figure 1: General Structure of encoder–decoder image captioning framework [35]

[7] The author described a neural image caption (NIC) system that encodes images into an intermediate representation using a deep CNN as an encoder. After that, a decoder known as a RNN produces equivalent descriptions. The author used a more sophisticated RNN model in which visual input is directly delivered to the RNN, making it easier for the RNN to maintain track of things described by text. As a result, the system's output outperformed traditional benchmarks by a large margin.

[8] Author presented an approach called weakly-supervised image captioning i.e. WICA. This approach is able to generate image caption with rich contextual information with incomplete dataset or incomplete annotation i.e. weakly supervised on contextual level. Using a sequence to sequence approach, they first apply an encoder decoder neural network to obtain the essential features that characterize the picture. To enhance the captioning task with contextual information they use object detection model Faster-RCCN detects features of objects in an image.

[9] The author presented an encoder-decoder framework that can combine joint picture text embedding models with multimodal neural language models. As a result, a word-by-word output sentence can be formed by providing an image as an input. For textual data, they utilized a LSTM RNN, while for visual data, they used a deep CNN.

- **Attention based image captioning framework**

As the image hold within a large amount of information, it is not required to describe the image's entire contents. For this reason, an attention-based image captioning framework is used. The attention model, which helps to handle this problem by extracting the essential image regions with respect to image context, solves the limitation of the encoder-decoder approach. Attention is the ability to choose one's own interests. Image captioning quality has been enhanced significantly with the help of attention mechanism.

Attention based image captioning task comes with spatial attention, Semantic attention, Self-adaptive attention etc. Figure 2 and 3 shows the pictorial representation of model with attention and no-attention mechanism.
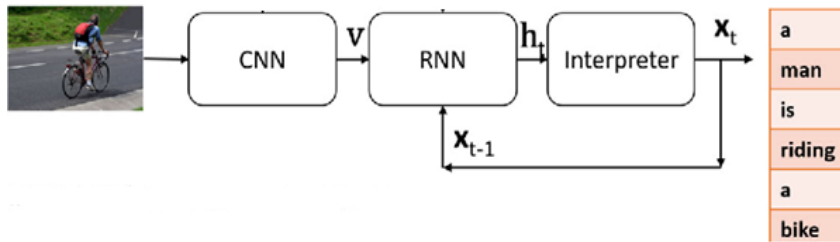
Figure 2: Image Captioning Model without Attention Mechanism [10]
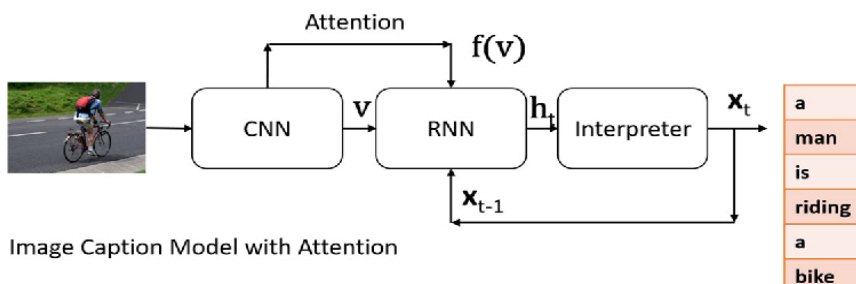
Figure 3: Image Captioning Model with Attention Mechanism [10]

[11] Author described image captioning system in which at first stage image is examined and signified by means of several visual regions to extract visual features/ global visual context. These feature vectors are provided as an input to LSTM network, where hidden states of it is used for predicting where the next sequence of visual focus on different regions should be and also for predicting sequence of generating upcoming word in the caption through scene specific language model.

[12] GLA: Global-local attention approach for generating image captions was introduced by the author, and it generates more relevant image captions. The proposed method focuses on key regions which are semantic in nature with maintaining global context information via attention mechanism on integrated local and global features. For extraction of image features they used VGG16, for object detection purpose used Faster R-CNN and for language model used stacked two layer LSTM.

[13] Author initiated novel model for captioning purpose. They proposed saliency prediction model that make decision on two attentive parts one is silent region and another is contextual region of the image for caption generation as saliency can enrich image description. These two parts cooperate with each other during the generation of caption and in turn can generate enhanced image description. For extracting high-level image features and LSTM on features and saliency map, the model uses a fully CNN as an encoder.

- **Dense captioning based framework**

Dense captioning based framework is the new approach for understanding of the image. Dense captioning deals with locating a silent region or region of interest in the image and generating regional language caption simultaneously.

[14] The proposed work aims to construct a deep neural network model which will reason about image content and representation in the natural language domain. Author initiated a ranking model that aligns language modalities to visual regions through a multimodal embedding. Multimodal RNN is used for generating description from visual data or image. This architecture performance is evaluated on datasets like Flicker 8k, Flicker 30k and COCO. However, the model's shortcoming is that it can only provide

descriptions for one-pixel array input at a fixed resolution. Multiple saccades throughout the image can be used to determine all items, their interactions, and the larger context for caption generation.

[15] Author designed an architecture called Fully Convolutional Localization Network (FCLN) which is able to localize region of interest from image and defines each region with natural language. This model jointly solves the localization and sentence generation task. This architecture processes image with the help of effective forward pass which doesn't requires external proposal of regions with end to end training and having single optimization round. Architecture includes Convolutional network, RNN as language model for description generation and novel dense localization layer which can be included in any neural network for image processing task with region level training and prediction like region of interest. Architecture is evaluated on benchmark dataset like Visual Genome.

[16] In this work, author proposed a unique architecture entitled Context and Attribute Grounded Dense Captioning (CAG-Net). This is an end-to-end architecture which uses target i.e. global and contextual i.e. neighboring hints for dense captioning. Contextual feature extractor and attribute grounded caption generator are the two components that make up CAG-Net.

Figure 4. shows some illustration of the effects of image captions on the various methods proposed by Karpathy and Fei-fei [14], Vinyals et al. [7], Xu et al. [37] and Fang et al. [38]. The results show that the approach followed by Xu et al. uses additional information into encoder-decoder framework i.e. attention mechanism with dynamically attend salient image regions throughout the procedure of generating image description which gives superior performance in terms of generating more accurate caption for the given images.

[14]	A pan filled with broccoli and meat	A street sign on the side of the road	A group of people standing on top of a snow covered slope	A baseball player pitching a ball on top of a field
[7]	A pan filled with broccoli and meat cooking	A stop sign on the side of the road	A group of people standing on top of a snow covered slope	A baseball player pitching a ball.
[37]	A pan filled with broccoli and meat on a stove	A stop sign on a road with trees	A group of people sitting on a ski on a snow covered slope	A baseball player throwing a ball in green field.
[32]	A pot of broccoli on a stove	A yellow sign on a dirt road	A group of people posing for a picture on a ski lift	A baseball player throwing a ball.

Figure 4: Example of image captioning results on different approaches.

- **Scene Graph image captioner framework**

Although deep neural networks have recently showed promise in the image captioning challenge, they don't utilize the structural visual and textual knowledge inherent inside an image explicitly. So, the concept of scene graph comes into a picture. Scene Graph comprises of structured semantic information of an image.

[17] Author introduced unique Graph Convolutional Networks - Long Short-Term Memory (GCN-LSTM) design that fuses the modelling visual relationship for captioning task with attention-based encoder-decoder framework. As modelling relationships associated with objects in an image ultimately plays supportive role for describing image. In this, semantic relationship and spatial relationship of objects are integrated into image encoder. Precisely they construct spatial/semantic graph with directed edges based on detected objects spatial and sematic relationship and representation of objects region detected by faster R-CNN are then refined by graph structure using Graph Convolutional network. Here vertex defines each region and edges define relationship between them. By learning those regions features, proposed architecture take advantage of LSTM with attention mechanism as a decoder for generation of sentence. The architecture is evaluated on MSCOCO dataset.

[18] The author proposed the Scene Graph Auto-Encoder (SGAE), a revolutionary unsupervised learning approach that incorporates inductive bias into a dictionary.This language inductive bias is included in to fundamental encoder-decoder architecture to generate more human-like caption, which in turn work as re-encoder for generation of language. This ultimately results in improvement in the performance of encoder-decoder architecture. The performance of SGAE designs is tested using the MSCOCO benchmark dataset.

[19] Author presented a novel framework named as Scene Graph Captioner (SGC) for captioning task. This framework is capable of capturing the structural semantic visual scene through objects, its attributes and relationship between those objects. First Author developed methodology to generate scene graph based on different parameters of objects. Second, they suggested Scene graph captioner incorporates high-level graph and visual attention information into a deep captioning framework. The author presented a system that can capture semantic notion and graph topology by inserting scene graph into structural representation. They create a scene graph driven method for constructing graphs with attention in advance. Finally, an LSTM-based architecture turns the information into a description. By using high-level concepts and the attention clustering region, SGC is able to build descriptions based on graph-based construction. The MSCOCO dataset was used to test the proposed methodology.

[20] The author developed a unique model for high-level image understanding called Context-based Captioning and Scene Graph Generation Network (C2SGNet). The model at the same time creates scene graphs as well as natural language descriptions from images. The performance of the C2SGNet model was assessed using the Visual Genome data set as a benchmark dataset. However, the C2SGNet model has the limit that the context information for each layer is only available from the lower layer, not the upper layer.

3.   Challenges on Image Captioning

Human is capable of easily classifying contents in the scene and describing the same in natural language description. But this is quite difficult for computer system to perform the same task. Computer system can identify activities of human in a video to a certain level [21]. But the task of automatically generating visual scene description has remained unsolved. Furthermore, despite the fact that human action identification is a well-studied topic in CV, interpreting complex and long-term human activities automatically is a difficult challenge. [22].

Other major challenges include:

- Identifying the reasonable details of visual contents of image and interaction of the detected objects is a challenging task. Occasionally some refine actions are tough to detect for vision technique or it is non-visible. For example, it may create difficulty in interpreting the human activity in an image/video due to unclear boundaries and occlusions of interactive objects.
- The current architecture is primarily concerned with the problem of visual description. Designing a visual understanding system, like visual reasoning and visual question responding, to think one step

ahead would be more engaging. Such high-level visual understanding systems are estimated to function well in afterward. Generation of inaccurate natural language description of image due to reasons like failure to recognize the unpredicted objects or views, singular v/s plural errors in the textual description, though some word are not present in the image their presence are usually associated with each other, absence of visual temporal informal leads to improper action detection in image/video.

4.   Datasets

Data are the base of AI. For assessing the performance of classification approach, number of benchmark datasets has been proposed. Many datasets have been built mainly for image/video captioning task. The number of photos in each dataset is shown in Table 1.

◇ **MSCOCO**

MSCOCO [23] is the most often used dataset for the captioning of images. There are 82,783 training images, 40,504 validation images, and 5 human-annotated descriptions per image in this dataset. Furthermore, all descriptions in the training set are transformed to lowercase, and some unusual words that appear below 5 times are surplus, which results in a total dictionary of 10,201 different words in the dataset.

◇ **Flicker 8K**

 Flicker 8k [24] image derives from the Flicker site which is the Yahoo's photo album site. Flicker 8K has an image volume of 8,000 images, with 6000 images for training, 1000 images for verification, and 1000 images for testing. The image captioning result by Jia et al. [36] on Flicker8k dataset are presented in Figure 5.

Ground Truth Caption:
A little boy runs away from the approaching waves of the ocean.

Generated Caption:
A young boy is running on the beach.

Figure 5: Image Caption result by Jia et al. [36] on a sample image of Flickr8k dataset.

◇ **Flicker 30K**

Flicker 30k [25] contains 31,783 images, all of which were gathered from the Flicker website. It has a total of 28000 images intended for training, 1000 for verification, and 1000 for testing. These images are typically representing people participating in an event. For each image, the equivalent human annotation is still five sentences.

◇ **Conceptual Captions Dataset**

The Conceptual Captions dataset [26] has around 3.3 million images for training, validation, and test set 22530. The image captioning dataset has roughly 3.3 million examples, which is far larger than MSCOCO. It has a broad variety of images, including nature images, professional photos, cartoons, and drawings. Its captions are based on descriptions extracted from original Alt-text properties, which have been automatically converted to make a balance of cleanliness, in formativeness, and learnability.

◇ **Visual Gnome**

Visual Genome is a database, a knowledge base, and an ongoing challenge that aims to link structured image concepts to language. The Visual Gnome dataset contains 108,077 images, with an average of 35 objects, 26 attributes, and 21 pairwise associations between objects in each image.

Table 1 Statistics of images count in every dataset.

Dataset Name	Total Volume	Overall Images		
		Train	Valid	Test
MSCOCO	330K	82783	40504	40775
Filckr8k	8091	6000	1000	1000
Filckr30k	31783	28000	1000	1000
Conceptual Caption	3.3 M	3.3 M	28355	22530
Visual Gnome	108077	-	-	-

5. Evaluation Metrics

The quantitative findings of certain representative methodologies are presented in this part, which highlights many types of widely used metrics for evaluation. Evaluation of image captioning methods is a not easy task. Capability of image captioning system can be compared in terms of how generated sentence is close to the human generated sentence and in semantic correctness also. Widely adopted evaluation metrics are BLEU [30], ROUGE [33], METEOR [34], etc. BLEU@N [30], METEOR [28], ROUGE-L [29], CIDEr-D [31], and SPICE [27] are five usually used metrics for quantitatively analysing the outcome of image or video captioning.

◇ **BLEU (Bilingual Evaluation Understudy)**

The evaluation metric used to assess the quality of generated text is BLUE. For this, bleu employs a measure in which each text is compared to a set of reference texts written by humans. On the other hand, there is no need to pay attention to syntactical accuracy while determining the proximity of a system's generated description to ground truth and a score assigned to each of them. Finally, the quality of the created text is assessed using the computed average score.

◇ **ROGUE (Recall-Oriented Understudy for Gisting Evaluation)**

Rouge metric match generated sentences words pair, words sequences and n-gram with the human annotated reference sentence. ROGUE is also available in several tasks specific terms like ROUGE-W, ROUGESU, ROUGE-1, 2. For small description ROUGE-SU and ROUGH-2 provides better performance. ROUGH-1and ROUGE-W is good for single document evaluation. Limitation of ROGUE is to compute on multi-document text summarization.

◇ **METEOR (Metric for Evaluation of Translation with Explicit ORdering)**

METEOR metric is used to compute machine-generated language. The concept of a generalized unigram match is used by METEOR metrics. This is done by comparing the machine-generated text to human-annotated sentences. If there are several references, the similarity score for each is examined, and the best score among the separately calculated ones is chosen.

◇ **CIDEr**

Aside from the measures stated above, CIDEr is an essential metric for image and video captioning. For each N-gram, the CIDEr metric applies a Term Frequency Inverse Document Frequency (TF-IDF) weighting to determine the consensus in image or video captioning. The assessment measures BLEU@N, METEOR, ROUGE-L, and CIDEr-D are mostly subtle to N-gram overlap. For two sentences to express the same meaning, this is neither essential nor sufficient. To solve this problem, a new evaluation metric known as SPICE is introduced.

✧ **SPICE**

The SPICE metric was recently designed to assess how well captions recover objects, properties, and relationships between objects in scene graphs that more closely resemble human judgment.

Table 2 gives a brief of the performance of image caption models on the MS-COCO, Flicker 8K, Flicker 30K, PASCAL and SALICON image dataset in terms of BLUE, METEOR and CIDEr, respectively. Models shown in the given table mainly adopts the architecture of CNN-RNN and CNN-LSTM architecture. From Table 2, conclusions could be made: GCN-LSTM [17] evaluations on COCO and [13] on the PASCAL database achieves high performance compared to other attention and non attention-based approach. As expected, in [17] CIDER points increased to 128.7% when improved with CIDER-D score. It is a unique design of Graph Convolutional Networks and Short-Term Memory (GCN-LSTM) that integrates the modelling visual relationship for captioning task with attention-based encoder-decoder framework. As modelling relationships associated with objects in an image ultimately plays supportive role for describing image.

Table 2 The results of several models on various benchmark datasets. BLUE-1, BLUE-2, BLUE-3, BLUE-4, METEOR, ROUGE-L, and CIDEr are represented by the metrics B@1, B@2, B@3, B@4, M, R, C.

Reference	Dataset	B@1	B@2	B@3	B@4	M	R	C
[7]	COCO	-	-	-	27.7	23.7	-	85.5
[8]	COCO	30.9	17.1	10.6	7.1	-	-	-
[11]	COCO	72.4	55.5	41.8	31.3	24.8	53.2	95.5
[12]	MSCOCO	72.5	55.6	41.7	31.2	24.9	53.3	96.4
	FLICKER 8K	57.2	37.9	23.9	14.8	16.6	41.9	36.2
	FLICKER 30K	56.8	37.2	23.2	14.6	16.6	41.9	36.2
[13]	SALICON	69.2	51.4	37.2	26.9	22.9	50.4	73.3
	COCO	70.8	53.6	39.1	28.4	24.8	52.1	89.8
	FLICKER 8K (VALIDATION)	62.8	44.5	30.2	19.9	20.3	46.5	50.1
	FLICKER 8K(TEST)	63.5	45.6	31.5	21.2	21.1	47.5	54.1
	FLICKER 30K (VALIDATION)	61.3	43.3	30.1	20.9	20.2	45.0	44.5
	FLICKER 30K(TEST)	61.5	43.8	30.5	21.3	20.0	45.2	46.4
	PASCAL-50S	82.4	70.2	57.5	45.7	32.9	66.3	70.7
[14]	MSCOCO-2014	62.5	45.0	32.1	23.0	19.5	-	66.0
	FLICKER 8K	57.9	38.3	24.5	16.0	-	-	-
	FLICKER 30K	57.3	36.9	24.0	15.7	-	-	-
[17]	COCO	80.9	65.5	50.8	38.3	28.6	58.5	128.7
[18]	MSCOCO	80.8	-	-	38.4	28.4	58.6	127.8
[19]	MSCOCO	67.9	49.3	34.7	24.3	22.2	48.8	75.4

6. Conclusion and Future Perspective

Reviewed several image captioning models and their limitations in this paper. Different benchmark datasets and evaluation measures were also presented and discussed. The results of numerous approaches applied to various datasets are illustrated, and several issues in image captioning are explored. The major flaw in recent work is that it is unsuccessful in constructing context combinations, and it has a major constraint in generating relationships among the many components in the image. The main reason behind this is that the context is not defined effectively and recurrent units are not able to generalize and recognize them. Though there is huge success achieved in recent years in image

captioning still there is a big room for enhancement. Future work must be progress in the direction of building context and generalization, with more accurate textual description generation. Another serious issue is the time it takes to train, test, and generate textual descriptions for the model in an efficient manner to increase performance. Another future direction will be to design a system in such a way that it is capable of describing an image by summarizing object relationships even if some objects are not precisely recognized or absent.

7. References

[1] Gupta A, Verma Y, Jawahar C. Choosing Linguistics over Vision to Describe Images. AAAI. 2021Sep. 26(1):606-12. https://ojs.aaai.org /index.php/AAAI/article/view/8205

[2] Hodosh, M., Young, P., Hockenmaier, J.: Framing Image Description as a Ranking Task: Data, Models and Evaluation Metrics. Journal of Artificial Intelligence Research 47, 853– 899 (2013), 10.1613/jair.3994;https://dx.doi.org/10.1613/jair.3994

[3] Karpathy A. Nips2014-1. Adv Neural Inf Process Syst. 2014;1–9.

[4] Farhadi A, Hejrati M, Sadeghi MA, Young P, Rashtchian C, Hockenmaier J, et al. Every picture tells a story: Generating sentences from images. Lect Notes Comput Sci (including Subser Lect Notes Artif Intell Lect Notes Bioinformatics). 2010;6314 LNCS(PART 4):15–29.

[5] Mitchell M, Han X, Dodge J, Mensch A, Goyal A, Berg A et al. Midge: Generating image descriptions from computer vision detections. In EACL 2012 - 13th Conference of the European Chapter of the Association for Computational Linguistics, Proceedings. Association for Computational Linguistics (ACL). 2012. p. 747-756

[6] Kulkarni G, Premraj V, Ordonez V, Dhar S, Li S, Choi Y, et al. Baby talk: Understanding and generating simple image descriptions. IEEE Trans Pattern Anal Mach Intell. 2013;35(12):2891–903.

[7] Vinyals O, Toshev A, Bengio S, Erhan D. Show and tell: A neural image caption generator. Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit. 2015;07-12-June-2015:3156–64

[8] Zheng HT, Wang Z, Ma N, Chen J, Xiao X, Sangaiah AK. Weakly-supervised image captioning based on rich contextual information. Multimed Tools Appl. 2018;77(14):18583–99.

[9] Kiros R, Salakhutdinov R, Zemel RS. Unifying Visual-Semantic Embeddings with Multimodal Neural Language Models. 2014;1–13. Available from: http://arxiv.org/abs/1411.2539

[10] Sur, C. Survey of deep learning and architectures for visual captioning—transitioning between media and natural languages. Multimed Tools Appl 78, 32187–32237 (2019). https://doi.org/10.1007/s11042-019-08021-1

[11] K. Fu, J. Jin, R. Cui, F. Sha and C. Zhang, Aligning Where to See and What to Tell: Image Captioning with Region-Based Attention and Scene-Specific Contexts, in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 39, no. 12, pp. 2321-2334, 1 Dec. 2017, doi: 10.1109/TPAMI.2016.2642953.

[12] L. Li, S. Tang, Y. Zhang, L. Deng and Q. Tian, GLA: Global–Local Attention for Image Description, in IEEE Transactions on Multimedia, vol. 20, no. 3, pp. 726-737, March 2018, doi: 10.1109/TMM.2017.2751140.

[13] Marcella Cornia, Lorenzo Baraldi, Giuseppe Serra, and Rita Cucchiara. 2018. Paying More Attention to Saliency: Image Captioning with Saliency and Context Attention. ACM Trans. Multimedia Comput. Commun. Appl. 14, 2, Article 48 (May 2018), 21 pages. DOI:https://doi.org/10.1145/3177745

[14] Karpathy A, Fei-Fei L. Deep Visual-Semantic Alignments for Generating Image Descriptions. IEEE Trans Pattern Anal Mach Intell. 2017;39(4):664–76

[15] Johnson, J., Karpathy, A., Fei-Fei, L.: DenseCap: Fully Convolutional Localization Networks for Dense Captioning. IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016)

[16] Yin, G., Sheng, & Lu, Liu, Bin, Yu, Nenghai, Wang, Shao, J.:Context and Attribute Grounded Dense Captioning. 6234-6243. IEEE/CVF Conference on Computer Vision and Pattern Recognition (2019)

[17] Yao, T., Pan, Y., Li, Y., Mei, T. (2018). Exploring Visual Relationship for Image Captioning. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds) Computer Vision – ECCV 2018. ECCV 2018. Lecture Notes in Computer Science(), vol 11218. Springer, Cham. https://doi.org/10.1007/978-3-030-01264-9_42

[18] Yang, X., Tang, K., Zhang, H., Cai, J.: Auto-Encoding Scene Graphs for Image Captioning.In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 10677–10686 (2019)

[19] Xu, N., Liu, A.A., Liu, J., Nie, W., Su, Y.: Scene graph captioner: Image captioning based on structural visual representation. Journal of Visual Communication and Image Representation 58, 477–485 (2019), 10.1016/j.jvcir.2018.12.027; https://dx.doi.org /10.1016 /j.jvcir.2018.12.027

[20] Shin, Kim, I.: Deep Image Understanding Using Multilayered Contexts. Mathematical Problems in Engineering. 2018. 1-11. Doi: 10.1155/2018/5847460. (2018)

[21] Torralba, A., Murphy, K., Freeman, W., Rubin, M.: wContext-based vision system for place and object recognition. Proc. IEEE Int. Conf. Comput. Vis pp. 273–280 (2003)

[22] Wallraven, C., Schultze, M., Mohler, B., Vatakis, A., Pastra, K.: The Poeticon enacted scenario corpus-A tool for human and computational experiments on action understanding. Proc. 9th IEEE Conf. Autom. Face Gesture Recognit pp. 484–491 (2011)

[23] Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick,C.L.: Microsoft COCO: Common objects in context. Proceedings of the IEEE International Conference on Computer Vision pp. 740–755 (2014)

[24] K. Anitha Kumari, C. Mouneeshwari, R. B. Udhaya, R. Jasmitha Automated Image Captioning for Flickr8K Dataset,Proceedings of International Conference on Artificial Intelligence, Smart Grid and Smart City Applications, 2020 ,ISBN : 978-3-030-24050-9

[25] Plummer, B.A., Wang, L., Cervantes, C.M., Caicedo, J.C., Hockenmaier, J., Lazebnik, S.: Flickr30k Entities: Collecting Region-to-Phrase Correspondences for Richer Image-to- Sentence Models. In: International Journal of Computer Vision. vol. 123, pp. 74–93. Springer Science and Business Media LLC (2017), 10.1007/s11263-016-0965-7;https://dx.doi.org/ 10.1007/s11263-016-0965-7

[26] Sharma, P., Ding, N., Goodman, S.: 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 (2018)

[27] Anderson, P., Fernando, B., Johnson, M., Gould, S.: SPICE: Semantic propositional image caption evaluation. Proceedings of the European Conference on Computer Vision pp. 382– 398 (2016)

[28] Banerjee, S., Lavie, A.: METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization pp. 65–72 (2005)

[29] Lin, C.Y.: Rouge: A package for automatic evaluation of summaries. Proceedings of the ACL Workshop on Text Summarization Branches Out. 10 pages (2004)

[30] Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: BLEU: A method for automatic evaluation of machine translation. Proceedings of the Annual Meeting on Association for Computational Linguistics pp. 311–318 (2002)

[31] Vedantam, C.L.R., Zitnick, D., Parikh: Cider: Consensus-based image description evaluation. Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition pp. 4566–4575 (2015)

[32] Fang, H., Gupta, S., Iandola, F., Srivastava, R. K., Deng, L., Dollár, P., ... & Zweig, G. (2015). From captions to visual concepts and back. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 1473-1482).

[33] Lin, C.Y., Och, F.J.: Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics. Meeting on Association for Computational Lin- guistics (2004)

[34] Lavie, A., Agarwal, A.: Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. The Second Workshop on Statistical Machine Translation pp. 228–231 (2007)

[35] Bai, S., An, S.: A survey on automatic image caption generation. Neurocomputing 311, 291–304 (2018), 10.1016/j.neucom.2018.05.080; https://dx.doi.org/10.1016/ j.neucom. 2018. 05.080

[36] Xu Jia, Efstratios Gavves, Basura Fernando, and Tinne Tuytelaars. 2015. Guiding the long-short term memory model for image caption generation. In Proceedings of the IEEE International Conference on Computer Vision. 2407–2415.

[37] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhutdinov, R. Zemel, Y. Bengio, Show, attend and tell: neural image caption generation with visual attention, arXiv:1502.03044v3 (2016)

[38] R Core Team, R: A language and environment for statistical computing, 2019. URL: https://www.R-project.org/.

[39] S. Anzaroot, A. McCallum, UMass citation field extraction dataset, 2013. URL: http://www.iesl.cs.umass.edu/data/data-umasscitationfield.