IJCSE International Journal of Computer Sciences and Engineering Open Access
Survey Paper Volume-5, Issue-8 E-ISSN: 2347-2693

Image Caption Generation: A Survey

Khushboo Khurana[1*], Shyamal Mundada[2]

[1*] Computer Science and Engineering Department, Shri Ramdeobaba College of Engineering and Management, Nagpur, India
[2] Computer Science and Engineering Department, Shri Ramdeobaba College of Engineering and Management, Nagpur, India

[*]*Corresponding Author: khuranak1@rknec.edu*

Available online at: www.ijcseonline.org

Received: 19/Feb//2018, Revised: 25/Feb2018, Accepted: 17/Mar/2018, Published: 30/Mar/2018

***Abstract*—** In recent years, the amount of images are increasing due to advancement in the technologies. This proliferation in Image data demands analysis and generation of image descriptions. The generated image captions must describe the image in a precise manner, covering all aspects of the image. Automatic image caption generation has emerged as an important task of research in the new integrated community of language-vision. Image captioning techniques can be broadly divided into data-driven and feature-driven methods. Data-driven techniques involve extraction of a similar image, whose caption is as it is copied or extraction of multiple images and then combining their captions to form appropriate caption for the input image. Feature based methods involve analyzing the visual content of the image and then generating natural language sentences. In this paper, we have reviewed both the methods along with the most efficient feature based technique that uses Convolution Neural Networks (CNNs) and Recurrent Neural Networks (RNNs). Various CNN-RNN based techniques are proposed by researchers to solve the Image captioning task and have achieved remarkable results. First, image objects and their relations are analyzed using CNN and then RNN are used for sentence generation. We have also elaborated the concept of CNN.

***Keywords*—** Image captioning, image understanding, CNN, RNN, natural language generation

I. INTRODUCTION

The task of automatically generating image description has attracted a lot of attention in the field of Artificial Intelligence. Automatic image description generation requires full understanding of image and sophisticated techniques of natural language generation. Thus the area of research in Image Captioning requires the techniques of both Natural Language Processing (NLP) and Computer Vision (CV).

A number of other problems such as Image Captioning, Visual Question Answering, Visual Storytelling, Visually Grounded Dialog, Image synthesis from text descriptions, etc. can also be addressed using the integration of the two technologies. In this paper we focus only on Image Captioning.

Major software industries and social media giants have realized the importance of Image captioning and have taken efforts in the direction. Facebook released an automatic image captioning tool that uses deep neural networks. The system can identify particular objects in a photo. It can pick out particular characteristics of the people in the photo, including smile, beard, glasses, etc. It can also describe the scene characteristics. Google and Microsoft also have their Image Captioning Systems that they presented in Microsoft Common Objects in Context (COCO) Captioning Challenge for automatically coming up with captions for images in papers [16], [27-28].

Humans can perceive the data around them. Humans have sensory perception generating signals from the environment. The tools for perception are sound, vision, smell, taste and touch. The association between perception and physical reality is predominantly strong for the visual sense and sound for humans. For perception of data by machine we need visual understanding. Computer vision algorithms are applied for understanding of visual data. Let us consider for example, we have an image of an outdoor scene. Scene understanding involves the extraction of objects in the image, action performed (if any), location of the scene, etc. and attaching this information as image annotations. Much of the research work in computer vision focuses on image annotation, by generation of labels for an image. The labeling is based on the content of the image. These labels are however not arranged in a meaningful sentence. Such a task of arranging the image labels to form meaningful sentence requires use of sophisticated natural language processing techniques.

In the next section we discuss image captioning and the methods used for image caption generation. In Section III, we have discussed the process for neural network based image understanding and sentence Generation and then

presented the review of CNN and RNN based Techniques in section IV. Application domains of image captioning are presented in section V followed by conclusion in section VI.

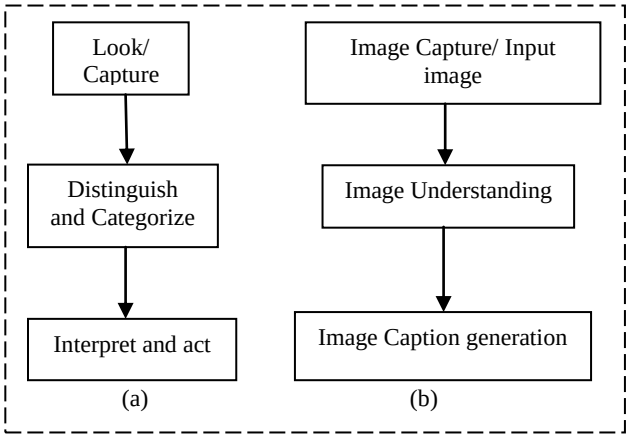

Figure 1. Co-relation between Human Perception and Image Captioning xample of a figure caption. (a) Human Perception of scene (b) Image captioning system

II. IMAGE CAPTIONING

In the new integrated community of language-vision, automatic image caption generation has emerged as an important task of research. Image captioning involves taking an image, analyzing the visual content and then generating natural language description that describes the most aspects of the image. The natural language generated is mostly in the form of natural language sentences; but not necessarily.

Image captioning is a challenging task from the CV as well as NLP point of view. Image understanding requires identification of objects, their attributes, high-level features of the image like indoor image or outdoor image, action performed, etc. But challenges exist, such as, implied objects where the objects or people that are not present in the image may be referred. The captions generated thus require understanding of all aspects of the image and may also require some knowledge database. The captions generated must be easy to understand and describe the image precisely. The image captions generated can be broadly classified as:

- Content Based
- Context Based

The techniques for image caption generation either generate captions that are based on the content by performing image content analysis or can be based on context that require strategies based on image as well as natural language.

The task of image captioning can be co-related to the way humans perceive the information and process it. Perception of information by humans is followed by information processing by performing interpretation and realization of object or environment. As shown in figure 1, human perception and understanding of a scene involves getting acquainted with the context (look), distinguish, categorize and then interpretation of sensory information. For example, the person first looks around to capture the information, then distinguishes the different objects in the scene to categorize the objects after their recognition and act accordingly. Image captioning thus has two main components- Image understanding and Image Caption Generation. Image understanding can be accomplished using image processing techniques and Caption generation requires techniques of Natural Language Generation (NLG).

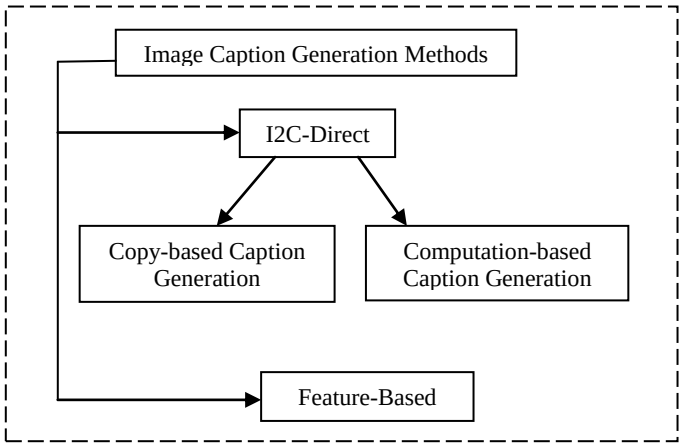

Figure 2. Image Caption Generation Methods

Image Caption Generation Methods can be broadly divided into I2C-Direct and feature-based as shown in figure 2.

- I2CDirect methods are data-driven methods that directly convert the images into captions. These methods are data-driven that model the image caption generation as a retrieval problem; where the dataset is assumed to contain images along with associated captions. Data-driven methods can be further divided into copy-based and computation-based. In copy-based caption generation the closest image to the image in consideration is found and the caption associated with the extracted image is copies as the caption for the image in consideration. In computation-based caption generation, a number of relevant images are extracted for the test image and captions associated with the extracted images are used by the model to generate new caption for the test image.
- In Feature-Driven methods image is first analyzed by image processing algorithms and then sentence generation strategies are applied. These methods can be Graph based/ Tree based or based on machine learning algorithms.

A. I2C-Direct

Data-driven methods have proven to be highly effective for image caption generation [24]. The authors have proposed to integrate an object-based semantic image representation into a deep features-based retrieval framework to select the relevant images. Given an image, the system retrieves a number of relevant images and their associated captions from a large set of images from which descriptions can be generated. CNN features are used to retrieve the images of similar scenes. Then, those candidate images are re-ranked using a novel object-based semantic representation to extract most relevant images. Then most important image regions are retrieved from each retrieved image, and only those images that share common important regions are considered further. All descriptions that belong to each cluster to the same bag, and use textual relations with these descriptions to generate a novel description of the query image.

Yansong Feng, et al. [2] have proposed a system for video and image retrieval along with tool that aid visually impaired individuals to access pictorial information. The caption generation is for news domain images. The process is a two step process- content selection and surface realization. Content selection identifies what the image is about and surface realization, verbalizes the chosen content. Probabilistic image annotation model is used for keyword suggestion that uses machine learning approach. The results are comparable to the handwritten captions. Rather than enumerating the objects in the picture and how they relate to each other, the captions created are news-worthy text that draws the reader into the accompanying article.

Another paper that works on news dataset is [15]. The dataset contains images along with other information such as news article, keywords, category labels, etc. Semantics between such heterogeneous information sources is often implemented in a common representation space. Images can be divided into image classes. The context representation from each source reduces to estimating the probability distribution of the context categories conditioned on the source of context for the training and the test items. These probability distributions are then combined in a generative model for image annotation, thus incorporating the estimated context. Annotations of the images are generated by the first step of the framework. In the second step articles related to the image are found that helps in the caption generation. The caption generation is inspired by extractive summarization and headline generation methods proposed in natural language processing. Context-based captions are generated.

B. Feature-Based

In [1], authors have presented a framework that generates text descriptions from image and video content. As in natural language processing, the text is first parsed into tokens, in this paper the authors have parsed the image by decomposing it into constituent visual patterns. The scene of the image or video frame is decomposed into objects and their relationships. These objects can then be further decomposed. Figure 3, referred from the research paper well explains the process. The scene is decomposed into different visual objects and their relations are indicated. Person object is further decomposed into its parts. These visual patterns are represented using an AND-OR graph (AoG) that represent syntactic and semantic relations of primitives, parts, objects and scenes. Interactive image parser (IIP) is used, that first applies edge based segmentation to find different objects. The framework is capable to recognize visible objects, occluded objects and surface. Visual patterns are then converted into semantically meaningful text by a text generation engine. The natural language text is generated using the relations between the objects.

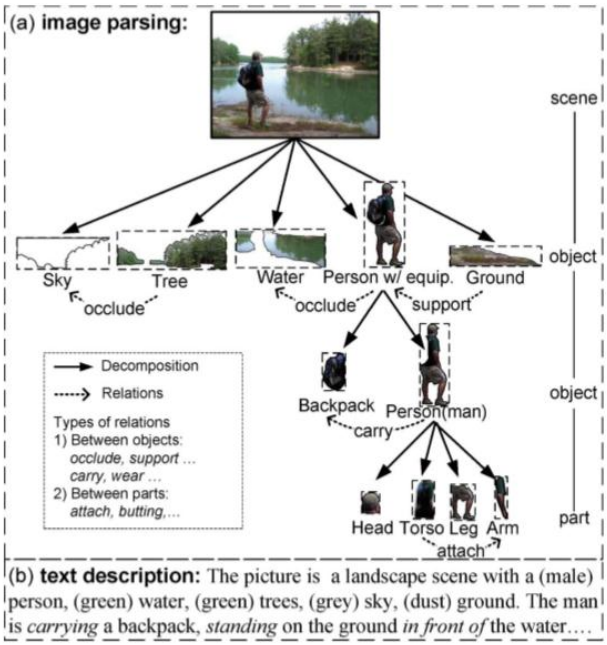

Figure 1. (a) image parsing and (b) text description. (Referred from [1])

Description for images of objects shot in uniform background is determined in [13]. Manually created database of objects indexed by an image signature (e.g., color and texture) and two keywords (the object's name and category) are used. Images are first segmented into objects and their signatures are generated. Using the signatures description is generated using template based sentence generation technique.

A three-step task of image description generation is presented in [18]. The first step is to identify objects in an image, the second step detects spatial relations between object pairs on the basis of language and visual features; and in the third step, spatial relations are mapped to natural

language (NL) descriptions. Results for various natural language generation techniques are presented.

In [3], content planning and surface realization are the main steps for generation of natural language sentences. First the output of computer vision-based detection and recognition algorithms are smoothened and then best content words are selected for image description. Object detectors are used to detect the objects present in the image. Attribute classifiers then determine the amount attribute present in each object for a predefined set of attributes. After this the prepositions are found find the relationship between the objects. Relations such as near, against, beside, etc. are used by the authors. Labeling of graph is predicted after the unary image potentials are incorporated into a conditional random field (CRF), which leads to sentence generation. Image recognition system extracts visual information as a set of triples describing the depicted objects, their attributes and spatial relationships, such as <<white;cloud>; in; <blue;sky>>. Models like n-gram, ILP-based optimization and template-based approach are used and compared. Authors found that the template method is simplest and produces somewhat robotic sounding sentences, it does generate consistent sentences by construction. Both the simple decoding method and ILP-based decoding sometimes produce grammatically incorrect sentences.

III. NEURAL NETWORK BASED IMAGE UNDERSTANDING AND SENTENCE GENERATION

Convolutional Neural Networks (CNNs) are analogous to traditional Artificial Neural Networks, except that in CNN the neurons self-optimise through learning. Each neuron will still receive an input and perform an operation. The only notable difference between CNNs and traditional ANNs is that CNNs are mostly used in the field of pattern recognition within images. This allows encoding image-specific features into the architecture, making the network.

CNNs can comprise different types of layers- convolutional, ReLu, pooling and fully-connected layers. These layers are stacked to build the CNN architecture. The input layer will hold the pixel values of the image. The convolutional layer will determine the output of neurons of which are connected to local regions of the input through the calculation of the scalar product between their weights and the region connected to the input volume. The rectified linear unit (ReLu) is used to find the activation function to produce the output of the previous layer. The pooling layer will then simply perform down sampling to reduce the number of parameters within that activation. The fully-connected layers perform the same duties as in standard ANNs to produce class scores from the activations, to be used for classification [33]. The convolution, ReLu and pooling layers can be repeated as a part of hidden layer after which fully connected layer is used. CNNs can have any number of hidden layers 10s to 100s. Each hidden layer learns to detect a feature in the image.

Development of Convolution Neural Networks (CNNs) has made multi object recognition in images very efficient.

Let us consider a simple example as shown in figure 3 to understand the process of CNN. Let first understand what an image is? An image can be considered as a 2x2 array of pixels, where a single pixel can have value between 0 to 255. But colored images require an additional field representing 3 color channels- R, G and B. So, colored images are 3-dimensional containing width, height and depth. Width and height are width and height of the image and depth represents the color channel. Hence colored image will have 3 matrices associated with each image, one for each of the color channels.

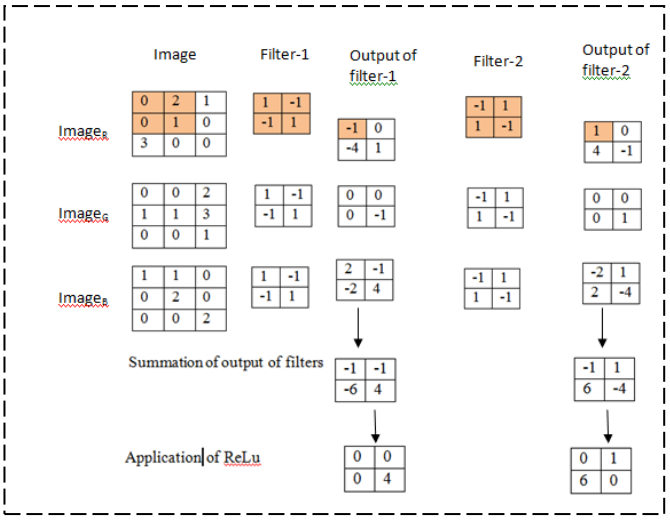

Figure 2. CNN Process

Let us consider an image of the size 3x3x3 and two simple filters of the size 2x2. We consider that the first filter detects the back-slash("\") and second filter detects the forward-slash("/").

The image has 3 matrices for each color channel. Figure 3 shows the process of CNN over an image. The 3 matrices for each color channel are as on the left-side. Since the size of filter is 2x2, we consider the first 2x2 matrix (Stride-1) of the ImageR and compute the dot product with filter-1 and filter 2 respectively to get the output for each filter, as highlighted in the figure 3. For example, convolution of filter-1 with stride-1 will be, 0x1+2x-1+0x-1+1x1= -1. Similarly filters are applied to all the strides of the image until complete image is convoluted. This process is repeated for the other 2 image matrices- ImageG and ImageB.

After convolution, summation of outputs for both filters is computed. Then ReLu is applied to make negative numbers zero and non-negative numbers are considered as it is. The last step is to apply pooling, after which flattening is performed to generate the final output predication.
There are different types of functions that can be used for activation by ReLu and then for down sampling by pooling. Some of the pooling methods are max-pooling, overlapping-pooling, spatial pyramid pooling [33], kernel pooling [34], wavelet pooling, etc.

A CNN is designed to recognize patterns across images (components) and then learn to combine these components to recognize larger structures (e.g., objects). Patterns are found on all the different subfields of the image. Recurrent Neural Networks (RNNs) are used to recognize patterns across time, that is output of previous step is also considered as the input to next step along with other inputs, by building memory in this process. RNNs are ideal for text analysis. Approaches based on CNN-RNN combination is used for generation of image captions. CNN is used for image understanding and RNN for Caption generation.

Caption generation can be extractive or abstractive. We generally need to generate a single sentence as caption which can be done by ranking the list of keywords extracted by the image annotation. Extractive methods yield grammatical captions and require relatively little linguistic analysis. But it is possible that the image content cannot be described using a single. Abstractive caption generation is used in [2]. Sentence generation/ Natural language generation can be broadly divided into grammar based and template based techniques. Template based generation of sentences is used by many researchers. A fixed template is filled based on the image concepts extracted [1, 9-11, 19, 20]. Farhadi et al. [5] use detections to infer a triplet of scene elements which is converted to text using templates. Grammar-based approach are used in [21], [22] to assemble descriptions.

Phrase combination method utilises a phrase-structure parser in order to isolate phrases from the collected descriptions. The phrase-structure trees are then traversed from root to the leaves and each node corresponding to an NP (noun phrase), VP (verb phrase) or PP (prepositional phrase) is extracted. Phrase-based language models use phrases extracted with shallow parsing (chunking) to build the language model. For each query image, language models extracted from the retrieved descriptions of the visually similar images from the same dataset are used to generate candidate sentences and the best sentences are selected after reordering them according to their probabilities and their lengths. A shallow parser is trained in order to chunk the sentences [24].

IV. REVIEW OF RNN AND CNN BASED TECHNIQUES

In [6] a feed-forward neural net is used to predict the next word given the image and previous words. In [7] a recurrent neural network is used for the prediction task.

CNN-RNN-based image captioning framework is presented in [29]. A generative model based on a deep recurrent architecture that combines recent advances in computer vision and machine translation that can be used to generate natural sentences describing an image is proposed in [4]. The model is trained to maximize the likelihood of the target description sentence given the training image. A single joint model that takes an image I as input, and is trained to maximize the likelihood P(S|I) of producing a target sequence of words S= {S1;S2;...}where each word Si comes from a given dictionary, that describes the image adequately. Deep convolution neural network (CNN) is used by the authors. They combine state-of-art sub-networks for vision and language models. These can be pre-trained on larger corpora and thus can take advantage of additional data.

In [35], authors have used CNN for image feature extraction, LSTM for caption generation along with a ranking objective.

A deep neural network model is presented in [8] that generate natural language descriptions of images and their regions. Datasets of images and their sentence descriptions are used to learn about the inter-modal correspondences between language and visual data. It infers latent alignment between segments of sentences and the region of the image that they describe. A novel combination of convolutional Neural Networks over image regions, bidirectional Recurrent Neural Networks (RNN) over sentences is proposed. Multimodal Recurrent Neural Network architecture is used that takes as input an image and generates its description in text. A model is then trained on the inferred correspondences and its performance is evaluated on a new dataset of region-level annotations.

In [12], Long-term Recurrent Convolutional Networks (LRCNs), a class of architectures for visual recognition and description which combines convolutional layers and long-range temporal recursion is presented. Convolutional Neural Network (CNN) is used for extraction of visual features, followed by LSTM. They have performed the task of activity recognition and description of video by providing input to CNN then LSTM. The task of image captioning is also performed. Long Short-Term Memory (LSTM) units are recurrent modules which enable long-range learning. LSTM units have hidden state augmented with nonlinear mechanisms to allow state to propagate without modification, be updated, or be reset, using simple learned gating functions [14]. LSTM decoders can be driven directly from conventional computer vision methods which predict higher-level discriminative labels.

Another paper that uses CNN and LSTM is [16], where a scene-specific language model for generating text is proposed. The process of generating the next word, given the previously generated ones, is aligned with the visual perception experience where the attention shifts among the visual regions, imposing a thread of ordering in visual perception. An image is first decomposed and represented with multiple visual regions from which visual features are extracted. The visual feature vectors are then fed into a Long Short Term Memory network which predicts both the sequence of focusing on different regions and the sequence of generating words based on the transition of visual attention. The neural network model is also governed by a scene vector, a global visual context extracted from the whole image. Region-based attention or scene-specific contexts improve the system.
LSTM model can be trained on video-sentence pairs and machine learning can be applied to associate a video to a sentence [23].

A new cascade recurrent neural network (CRNN) for image caption generation rather than using classical multimodal recurrent neural network, which only uses a single network for extracting unidirectional syntactic features can be used [30]. CRNN adopts a cascade network for learning visual-language interactions from forward and backward directions, which can exploit the deep semantic contexts contained in the image. In the proposed framework, two embedding layers for dense word expression are constructed. A new stacked Gated Recurrent Unit is designed for learning image-word mappings.

The major problems faced problem of missing objects while generating the image description and misprediction, when one object may be recognized in a wrong category. In [25], a novel method called as global–local attention (GLA) for generating image description is presented. The proposed GLA model utilizes an attention mechanism to integrate local representation at object-level with global representation at image-level. This leads to processing of all objects and context information concurrently.

CNN for extracting image feature and RNN to predict next produced word, make the obtained features unadaptable to the word generated at current time. A new time-varying parallel RNN (TVPRNN) to deal with this task is proposed in [26]. This system uses two classical CNNs namely VggNet and inception v3 for extracting global image feature, jointly with RNN to obtain time-varying feature at each time step, which are used for representing current word. Visual and textual representation in a multimodal space is fused. Moreover, visual attention mechanism is introduced to guild the proposed network.

V. APPLICATION DOMAINS

News articles contain heading, text, image and caption of the image. The most attracting feature in a news article apart from the heading is the image and its caption which determines whether the reader is interested in reading the news text. Due to their prominence, journalists try to write good captions. These captions must be such that they identify persons, specify their task and describe the image in a precise manner. This task of creating captions of the news images, have a strong relation to headline generation [32], where the aim is to create a very short summary for a document. Here the intension is the generation of captions for images.

Caption generation for remote sensing images have various applications like Military intelligence generation. At war time, images of battle field are captured by spy drones or satellites. If these images are transformed into text or voice messages its can save human anlaysis of the images only important images can further be used. The messages can be further sent to the frontline combat soldiers or the command center [17].

In Robotics, enabling effective human–robot interaction is crucial. In this context, a fundamental aspect is the development of a user-friendly human–robot interface, such as a natural language interface. The robot must be able to generate complete sentences describing the scene, dealing with the hierarchical nature of the temporal information contained in image sequences. The robot side of the interface, in particular the ability to generate natural language descriptions for the scene using deep recurrent neural network architecture completely based on the gated recurrent unit paradigm is proposed in [31].

VI. CONCLUSION

Image Caption Generation techniques based on data and features are reviewed in the paper. The direct techniques convert the image directly to sentences either by copying the caption of related image or by computing the caption using the captions extracted from relevant images. These methods are highly dependent on the extracted images, correctness and preciseness of their associated captions. Initial research in feature based methods use image parsing. Then the parts of images are recognized using visual primitive recognizers, combined with graphs or logic systems, which are further converted to natural language via rule-based systems. Such systems are mostly hand-designed and have been demonstrated only on limited domains. Recent techniques that use combination of CNN and RNN for generation of image captions prove to give better results for caption generation.

REFERENCES

[1] B. Z. Yao, X. Yang, L. Lin, M. W. Lee, and S.-C. Zhu, "I2T: Image parsing to text description," in Proc. IEEE, vol. 98, no. 8, pp. 1485–1508, Aug. 2010.

[2] Y. Feng and M. Lapata, "Automatic Caption Generation for News Images," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 35, no. 4, pp. 797-812, April 2013.

[3] G. Kulkarni et al., "BabyTalk: Understanding and Generating Simple Image Descriptions," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 35, no. 12, pp. 2891-2903, Dec. 2013.

[4] G. Kulkarni, V. Premraj, S. Dhar, S. Li, Y. Choi, A.C. Berg, and T.L. Berg, "Baby Talk: Understanding and Generating Image Descriptions," Proc. IEEE Conf. Computer Vision and Pattern Recognition, pp. 1601-1608, 2011.

[5] A. Farhadi "Every picture tells a story: Generating sentences from images," in Proc. 11th Eur. Conf. Comput. Vis.: Part IV, 2010, pp. 15–29.

[6] R. Kiros and R. Z. R. Salakhutdinov, "Multimodal neural language models," in Proc. Neural Inf. Process. Syst. Deep Learn. Workshop, 2013.

[7] J. Mao, W. Xu, Y. Yang, J. Wang, and A. Yuille, "Explain images with multimodal recurrent neural networks," in arXiv:1410.1090, 2014.

[8] A. Karpathy and L. Fei-Fei, "Deep Visual-Semantic Alignments for Generating Image Descriptions," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 39, no. 4, pp. 664-676, April 1 2017.

[9] Y. Yang, C. L. Teo, H. Daume III, and Y. Aloimonos, "Corpusguided sentence generation of natural images," in Proc. Conf. Empirical Methods Natural Language Process., 2011, pp. 444–454.

[10] D. Elliott and F. Keller, "Image description using visual dependency representations," in Proc. Empirical Methods Natural Language Process., 2013, pp. 1292–1302.

[11] A. Gupta and P. Mannem, "From image annotation to image description," in Neural Information Processing. Berlin, Germany: Springer, 2012

[12] J. Donahue et al., "Long-Term Recurrent Convolutional Networks for Visual Recognition and Description," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 39, no. 4, pp. 677-691, April 1 2017.

[13] P. He´de, P.A. Moellic, J. Bourgeoys, M. Joint, and C. Thomas, "Automatic Generation of Natural Language Descriptions for Images," Proc. Recherche d 'Information Assistee par Ordinateur, 2004.

[14] S. Hochreiter and J. Schmidhuber, "Long short-term memory," in Neural Computation. Cambridge, MA, USA: MIT Press, 1997.

[15] A. Tariq and H. Foroosh, "A Context-Driven Extractive Framework for Generating Realistic Image Descriptions," in IEEE Transactions on Image Processing, vol. 26, no. 2, pp. 619-632, Feb. 2017.

[16] 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, Dec. 1 2017.

[17] Z. Shi and Z. Zou, "Can a Machine Generate Humanlike Language Descriptions for a Remote Sensing Image?," in IEEE Transactions on Geoscience and Remote Sensing, vol. 55, no. 6, pp. 3623-3634, June 2017.

[18] A. Muscat and A. Belz, "Learning to Generate Descriptions of Visual Data Anchored in Spatial Relations," in IEEE Computational Intelligence Magazine, vol. 12, no. 3, pp. 29-42, Aug. 2017.

[19] Y. Yang, C. L. Teo, H. Daumé III, and Y. Aloimonos, "Corpus-guided sentence generation of natural images," in Proc. 16th Conf. on Empirical Methods in Natural Language Processing, Edinburg, Scotland, July 27-31, 2011, pp. 444–454.

[20] D. Elliott and F. Keller, "Image description using visual dependency representations," in Proc. 18th Conf. on Empirical Methods in Natural Language Processing, Seattle, Oct. 18-21, 2013, pp. 1292–1302.

[21] M. Mitchell, X. Han, J. Dodge, A. Mensch, A. Goyal, A. Berg, K. Yamaguchi , T. Berg, K. Stratos, and H. Daumé, III, "Midge: Generating image descriptions from computer vision detections," in Proc. 13th Conf. of the European Chapter of the Association for Computational Linguistics, Avignon, France, Apr. 23-27, 2012, pp. 747–756.

[22] P. Kuznetsova, V. Ordonez, T. L. Berg, and Y. Choi, "Treetalk: Composition and compression of trees for image descriptions," Trans. Assoc. Comput. Linguist., vol. 2, no.10, pp. 351–362, June 2014.

[23] L. Gao, Z. Guo, H. Zhang, X. Xu and H. T. Shen, "Video Captioning With Attention-Based LSTM and Semantic Consistency," in IEEE Transactions on Multimedia, vol. 19, no. 9, pp. 2045-2055, Sept. 2017.

[24] M. Kilickaya, B. K. Akkus, R. Cakici, A. Erdem, E. Erdem and N. Ikizler-Cinbis, "Data-driven image captioning via salient region discovery," in IET Comp. Vision, vol. 11, no. 6, pp. 398-406, 9 2017.

[25] 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.

[26] L. Yang and H. Hu, "TVPRNN for image caption generation," in Electronics Letters, vol. 53, no. 22, pp. 1471-1473, 10 26 2017.

[27] H. Fang et al., "From captions to visual concepts and back," 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, 2015, pp. 1473-1482.

[28] J. Devlin, H. Cheng, H. Fang, S Gupta, L Deng, X. He, G. Zweig, M. Mitchell, "Language Models for Image Captioning: The Quirks and What Works" Cornell University Library, Oct.-2015.

[29] X. He and L. Deng, "Deep Learning for Image-to-Text Generation: A Technical Overview," in IEEE Signal Processing Magazine, vol. 34, no. 6, pp. 109-116, Nov. 2017.

[30] J. Wu and H. Hu, "Cascade recurrent neural network for image caption generation," in Electronics Letters, vol. 53, no. 25, pp. 1642-1643, 12 7 2017. doi: 10.1049/el.2017.3159

[31] S. Cascianelli, G. Costante, T. A. Ciarfuglia, P. Valigi and M. L. Fravolini, "Full-GRU Natural Language Video Description for Service Robotics Applications," in IEEE Robotics and Automation Letters, vol. 3, no. 2, pp. 841-848, April 2018.

[32] M. Banko, V. Mittal, and M. Witbrock, "Headline Generation Based on Statistical Translation," Proc. 38th Ann. Meeting Assoc. for Computational Linguistics, pp. 318-325, 2000.

[33] K. He, X. Zhang, S. Ren and J. Sun, "Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 37, no. 9, pp. 1904-1916, Sept. 1 2015.

[34] Y. Cui, F. Zhou, J. Wang, X. Liu, Y. Lin and S. Belongie, "Kernel Pooling for Convolutional Neural Networks," 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, 2017, pp. 3049-3058.

[35] G, T. Jain "Discriminatory Image Caption Generation Based on Recurrent Neural Networks and Ranking Objective," International Journal of Computer Science and Engineering, Vol 5 (10), pp.260-265, Oct 2017.

Authors Profile

Mrs. Khushboo Khurana is working as an Assistanst Professor in Shri Ramdeobaba College of Engineering and Management, Nagpur. Her area of intrests include Image, Video Processing and Big Data Analytics.

Mrs. Shyamal Mundada is working as an Assistanst Professor in Shri Ramdeobaba College of Engineering and Management, Nagpur. Her area of intrests include Image Processing and Real Time Scheduling.

