IJCSE International Journal of Computer Sciences and Engineering Open Access
Survey Paper Volume-6, Issue-3 E-ISSN: 2347-2693

Image Caption Generation: A Comprehensive Survey

Sailee P. Pawaskar[1*], J. A. Laxminarayana[2]

[1*]Computer Engineering Department, Goa College Of Engineering, Goa University, Farmagudi-Ponda, Goa, India
[2] Computer Engineering Department, Goa College Of Engineering, Goa University, Farmagudi-Ponda, Goa, India

[*]*Corresponding Author: sailee.spawaskar516@gmail.com*

Available online at: www.ijcseonline.org

Received: 23/Feb//2018, Revised: 28/Feb2018, Accepted: 21/Mar/2018, Published: 30/Mar/2018

Abstract— From the viewpoint of humans and computers, images could be interpreted in different ways. In case of humans, an image could be simply some description or scene of an action or environment etc.; while with respect to computers, it is just some combination of pixels or digital numbers. The process of Image Captioning deals with assigning internal data in the form of captions or keywords to a digital image. This paper is a comprehensive survey of different methodologies to generate appropriate image captions. Here, we have compared various approaches available for implementation of image captioning. We have also described the evaluation metrics that could be used by such systems. Appropriate captions will assist the users to search images with long queries. Automatic image captioning could also be useful for visually impaired people in understanding pictures.

Keywords—Automatic image captioning, Deep CNN, Hidden Markov Model, LSTM, Neural Network, RNN,

I. INTRODUCTION

Image Caption Generation has emerged as a challenging and important research area following advances in statistical language modeling and image recognition[1]. Automatic caption generation of images provides us with various benefits. It helps the visually impaired and also enables automatic labeling of the millions of images uploaded over the Internet every day.

Generation of image captions includes identifying and detecting objects, people etc. It also requires to determine the properties of objects and then combining several sources of information into a sentence. Therefore it is a very difficult task to define an image; which is also an important problem in the field of computer vision. Despite being a difficult problem, the research community has recently made headway into this area, thanks to large labeled datasets, and progresses in learning expressive neural network models[2]. The image description must capture not only the objects present in an image, but it must also express how these objects are related to each other. It should also include their attributes and the activities they are involved in. The field also brings together state-of-the-art models in Natural Language Processing and Computer Vision, two of the major fields in Artificial Intelligence. Also, it could provide more accurate information of images or videos in scenarios such as image sharing in social network or video surveillance systems [3]. The main purpose of this paper is to address the various techniques available to generate accurate captions for given images. We have also described the evaluation metrics that are used by such systems.

In this work, we study the contributions of researchers in the field of Automatic Image Captioning. Different captioning models are discussed in Section II, whereas the different available approaches for the implementation of an image captioning system are discussed in Section III. In Section IV, different evaluation metrics that are used to evaluate image captioning techniques are discussed. Section V compares different image captioning techniques and Section VI concludes our work.

II. IMAGE CAPTIONING MODELS

There are various caption generation models.

Model 1: Generate the Whole Sequence—The whole textual description for a given an image can be generated using this model. Figure 1 shows the important steps in this model.

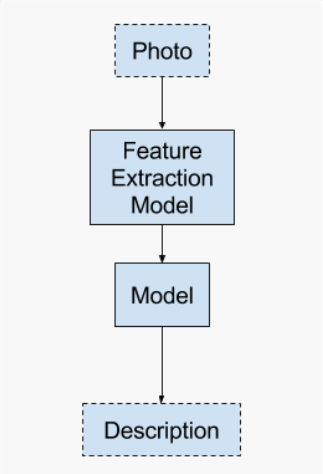

Figure 1. Model 1-Generate the whole sequence

This model generates the entire output in a one-shot manner. Here the image passes through a feature extraction model. All sequences are padded to the same length.

Model 2: Generate Word from Word—Given an image and one word as input the LSTM generates a prediction of one word. Figure 2 explains the proposed model.

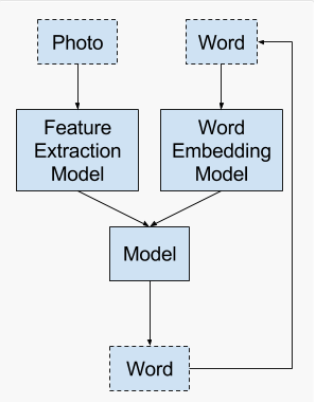

Figure 2. Model 2-Generate word from word

Here the textual description is generated via recursive calls to the model. The one word input is either a token to indicate the start of the sequence in the case of the first time the model is called or is the word generated from the previous time the model was called. The image passes through a feature extraction model. The input word is integer encoded and passes through a word embedding. This process is repeated until an end of sequence token is generated.

Model 3: Generate Word from Sequence—Given an image and a sequence of words that are already generated for the image, it predicts the next word in the description. Figure 3 demonstrates the working of this model.

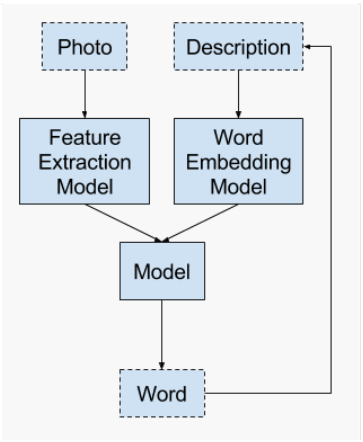

Figure 3. Model 3-Generate word from sequence

This model generates a textual description via recursive calls to the model. It is a generalization of the Model 2. The sequence of words that are given as input provides the model a context for generating the next word in the sequence. The image passes through a feature extraction. The image may be provided each time step with the sequence, or once at the beginning, which may be the preferred approach. The input sequence is padded to a fixed-length and integer encoded to pass through a word embedding. The recursive word generation process is repeated until an end of sequence token is generated.

III. APPROACHES TO IMAGE CAPTIONING

There are various approaches to generate image captions.

A. Neural Network Approaches

Xinlei Chen et al presented bi-directional mapping between images and their sentence-based descriptions[4]. This model is capable of both generating novel captions given an image, and reconstructing visual features given an image description. The evaluation was done on several tasks like sentence generation, sentence retrieval, and image retrieval. When compared to human-generated captions, captions generated by this model were preferred by humans 21.0% of the time.

Jeff Donahue et al developed a novel recurrent convolutional architecture suitable for large-scale visual learning which is end-to-end trainable and demonstrated the value of these models on benchmark video recognition tasks, image description, and retrieval problems, and video narration challenges[5]. Recurrent convolutional models are "doubly deep" in that they can be compositional in spatial and

temporal "layers". Such models may have advantages when target concepts are complex and/or training data are limited. These recurrent long-term models are directly connected to modern visual convnet models and can be jointly trained to simultaneously learn temporal dynamics and convolutional perceptual representations.

Hao Fang et al proposed a novel approach for automatically generating image descriptions: visual detectors, language models, and multimodal similarity models learned directly from a dataset of image captions[6]. Then used multiple instance learning to train visual detectors for words that commonly occur in captions, including many different parts of speech such as nouns, verbs, and adjectives. The word detector outputs serve as conditional inputs to a maximum-entropy language model. The language model learns from a set of over 400,000 image descriptions to capture the statistics of word usage. This model then captured global semantics by re-ranking caption candidates using sentence-level features and a deep multimodal similarity model. Used Microsoft COCO benchmark, producing a BLEU-4 score of 29.1%. When compared with human-generated captions this model produced better quality captions.

Andrej Karpathy et al presented a model that generates natural language descriptions of images and their regions[7]. This model is based on a novel combination of Convolutional Neural Networks over image regions, bidirectional Recurrent Neural Networks over sentences, and a structured objective that aligns the two modalities through a multimodal embedding. Then used a Multimodal Recurrent Neural Network (MRNN) architecture that uses the inferred alignments to learn to generate novel descriptions of image regions. Used the Flickr8K, Flickr30K and MSCOCO datasets for the experiment. The Multimodal RNN model is subject to multiple limitations. First, the model can only generate a description of one input array of pixels at a fixed resolution.

Vinayshekhar Bannihatti Kumar et al improvised existing technologies used in popular social networking sites like Twitter, to include some of the new states of the art technologies in machine learning to build features[8]. The model developed is similar to twitter along with additional features that include Image Captioning, Auto-Tagging of Tweets, Sentiment Detection, Spam Filtering and an Innovative News Feed Generator.

A generative model based on a deep recurrent architecture that combines recent advances in computer vision and machine translation and that can be used to generate natural sentences describing an image is proposed by Oriol Vinyals et al [9]. The model is trained to maximize the likelihood of the target description sentence given the training image. Experiments on several datasets show the accuracy of the model and the fluency of the language it learns solely from image descriptions. This model is often quite accurate when verified both qualitatively and quantitatively. For instance, while the current state-of-the-art BLEU-1 scores on the Pascal dataset is 25, this approach yields 59, to be compared to human performance around 69. Also showed BLEU-1 score improvements on Flickr30k, from 56 to 66, and on SBU, from 19 to 28. Lastly, on the newly released COCO dataset, it achieved a BLEU-4 of 27.7, which is the current state-of-the-art.

Geetika et al developed a model based on a deep recurrent neural network that generates brief statement to describe an image. CNN was used for extracting features[10]. They had also used ranking objective to subtle difference between similar images to generate discriminatory captions. Their model was evaluated using BLEU, METEOR and CIDEr scores.

The existing neural network approach to generate image caption has the following architecture:

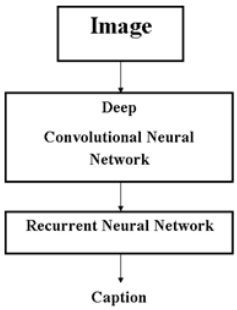

Figure 5. Existing architecture for image caption generation

B. Hidden Markov Approaches

Arnab Ghoshal et al presented a novel method for automatic annotation of images with keywords from a generic vocabulary of concepts or objects for the purpose of content-based image retrieval[11]. An image, represented as a sequence of feature vectors characterizing low-level visual features such as color, texture or oriented-edges, is modeled as having been stochastically generated by a hidden Markov model, whose states represent concepts. The parameters of the model are estimated from a set of manually annotated (training) images. Each image in a large test collection is then automatically annotated with the a posteriori probability of concepts present in it.

David Zajic & Bonnie Dorr proposed a novel application of Hidden Markov Models to automatic generation of informative headlines for English texts[12]. This model described four decoding parameters to make the headlines appear more like Headlines, the language of informative newspaper headlines. It also allowed morphological variation in words between headline and story English. Informal and

formal evaluations indicate that this approach produces informative headlines, mimicking a Headlines style generated by humans.

PHILO SUMI et al presented automatic caption generation for news images in association with the related news article[13]. Here it will input one image and news article to the system. The system will generate most important keywords which are associated with the image in association with the image. To find the image related keywords, first, we will find out the input image's features using SIFT (Scale Invariant Feature Translation) method. And using these features it will compare the image with the images which are stored in the database. After finding the best-matched image we will extract the keywords associated with that image. After applying grammatical rules to the keywords an appropriate caption is generated. This approach combines the textual modalities with the visual one.

C. Other Methods

Krishnan Ramnath et al developed a system that helps a smartphone user generate a caption for their photos[14]. It operates by uploading the photo to a cloud service where a number of parallel modules are applied to recognize a variety of entities and relations. The outputs of the modules are combined to generate a large set of candidate captions, which are returned to the phone. The phone client includes a convenient user interface that allows users to select their favorite caption, reorder, add, or delete words to obtain the grammatical style they prefer. The user can also select from multiple candidates returned by the recognition modules.

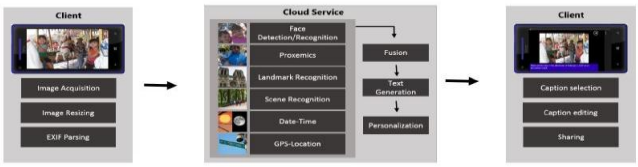

Figure 6. The architecture of the system proposed by Ramnath et al [14]

Kaustubh Shivdikar et al presented a hybrid engine that utilizes a combination of feature detection algorithms coupled with context-free grammar to create a model that serves to semantically and logically describe an image in its entirety[15]. This hybrid engine model has an F1 score of 94.33% and a unigram score of 75% when evaluated on a novel dataset trained on human-annotated images.

IV. EVALUATION METRICS

Evaluation metrics are used to assess the quality of automatically generated texts. Some evaluation metrics used to calculate score are as follows:

BLEU(Bilingual evaluation understudy)-
It is an algorithm for evaluating the quality of text which has been machine-translated from one natural language to another. The central idea behind BLEU is "The closer a machine translation is to human translation, the better it is". BLEU was one of the first metrics to achieve a high correlation with the human judgment of quality. It is one of the most popular automated and inexpensive metrics. Scores are calculated for individual translated segments generally sentences by comparing them with a set of good quality reference translations. Those scores are then averaged over the whole corpus to reach an estimate of the translation's overall quality.

CIDEr-
Automatically describing an image with a sentence is a challenge in computer vision and natural language processing. Due to recent progress in object detection, attribute classification, action recognition, etc there is renewed interest in this area.

METEOR (Metric for Evaluation of Translation with Explicit ORdering)-
It is a metric for the evaluation of machine translation output. The metric is based on the harmonic mean of unigram precision and recall, with recall weighted higher than precision. It also has several features that are not found in other metrics, such as stemming and synonymy matching, along with exact word matching. It produces a good correlation with human judgment at the sentence or segment level.

V. COMPARISON OF VARIOUS APPROACHES

In this section, we will compare the various approaches to generate image captions that are discussed in Section IV.

In case of Neural Networks, the image is first fed to the deep convolutional network. Then the image passes through the various layers of deep CNN. The convolutional network is used to generate image vector for the particular image that is fed. Using the image vector as an input caption generator will then generate appropriate captions. Here in this case caption generator that is used is called Recurrent Neural Network(RNN). These RNNs are used to handle long-term dependencies. That is unlike feedforward networks these RNNs are able to produce independent output at every timestamp. The RNNs are replaced by Long Short-Term Memory(LSTM) and Gated Recurrent Units(GRUs) as a solution to vanishing gradient problem. D. J. Kim et al have

proposed gLSTM wherein the input image is fed to the network at every timestamp for better results[16].

In case of Hidden Markov(HM)Approach for generation of news caption generation, the image and news article together is given as an input. This is then fed to the keyword extraction algorithm which will find out all possible keywords. These keywords are then given to hidden Markov model to generate most likely keywords. Then passed on to the sentence generator to produce the captions. So we find that deep neural network approach provides us with accurate information as compared to other approaches. Features can be more accurately extracted using deep neural network approaches. In some other cases, context-free grammars are also used in order to generate semantically correct captions.

From the evaluation point of view, METEOR score mentioned in Section IV has a higher correlation with human judgment (0.964) than BLEU score (0.817).

VI. CONCLUSION

The focus of this paper is to address various techniques for automatically generating caption for images, which is important for many image-related applications. Here we found that deep neural network approach provides us with accurate information as compared to other approaches. Features can be more accurately extracted using deep neural network approaches.

REFERENCES

[1] Moses Soh, "*Learning CNN-LSTM Architectures for Image Caption Generation* ", 2016.

[2] Mathews, Alexander & Xie, Lexing & He, Xuming, " *SentiCap: Generating Image Descriptions with Sentiments*", 2015.

[3] Jianhui Chen, Wenqiang Dong, Minchen Li, "*Image Caption Generator Based On Deep Neural Networks*".

[4] X. Chen and C. L. Zitnick, "*Mind's eye: A recurrent visual representation for image caption generation*," 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, 2015, pp. 2422-2431.

[5] 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, Issue 4, pp. 677-691, April 1 2017.

[6] 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.

[7] 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.

[8] V. B. Kumar, T. R. Baadkar, and V. Joshi, "*CRYPTANITE: A New Look to the World of Social Networks Using Deep Learning*," 2016 12th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS), Naples, 2016, pp. 358-364.

[9] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "*Show and tell: A neural image caption generator*," 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, 2015, pp. 3156-3164.

[10] Geetika, Tulsi Jain, "*Discriminatory Image Caption Generation Based on Recurrent Neural Networks and Ranking Objective*", International Journal of Computer Sciences and Engineering, Vol. 5, Issue.10, pp.260-265, 2017.

[11] Arnab Ghoshal, Pavel Ircing, Sanjeev Khudanpur "*Hidden Markov Models for Automatic Annotation and Content-Based Retrieval of Images and Video*".

[12] Zajic R. Schwartz, D & Door, B & Schwartz, Richard "*Automatic Headline Generation for Newspaper Stories*", 2018.

[13] PHILO SUMI , ANU.T.P " *A Systematic Approach for News Caption Generation*", International Journal of Advanced Research in Computer Science & Technology (IJARCST 2014), Vol. 2, Issue 2, Ver. 1 (April - June 2014)

[14] K. Ramnath *et al*., "*AutoCaption: Automatic caption generation for personal photos*," *IEEE Winter Conference on Applications of Computer Vision*, Steamboat Springs, CO, 2014, pp. 1050-1057.

[15] K. Shivdikar, A. Kak, and K. Marwah, "*Automatic image annotation using a hybrid engine*," 2015 Annual IEEE India Conference (INDICON), New Delhi, 2015, pp. 1-6.

[16] D. J. Kim, D. Yoo, B. Sim and I. S. Kweon, "*Sentence learning on deep convolutional networks for image Caption Generation*," 2016 13th International Conference on Ubiquitous Robots and Ambient Intelligence (URAI), Xi'an, 2016, pp. 246-247.

Authors Profile

Miss Sailee P. Pawaskar pursued Bachelor of Engineering in Information Technology from Goa College Of Engineering, Goa University in 2016. At present, she is pursuing Master Of Engineering in Computer Science & Engineering from Goa College Of Engineering, Goa University.

