iJRASET
International Journal For Research in Applied Science and Engineering Technology

# INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

**Volume:** 8 **Issue:** I **Month of publication:** January 2020

**DOI:** http://doi.org/10.22214/ijraset.2020.1059

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
*ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.177*
*Volume 8 Issue I, Jan 2020- Available at www.ijraset.com*

# A Survey on Deep Visual Semantic Alignments for Generating Image Descriptions

Ms. Malge Shraddha V.[1], Prof. Shah H. P.[2]
[1, 2]*Computer Science and Engineering, M. S. Bidve Engineering College, Latur, Dr. Babasaheb Ambedkar Technological University, Lonere.*

***Abstract:** In a generalized way, describing an image is called as image captioning. As a human being it is easy for us to look at a scene/image and describe it in depth with all the details needed to be mentioned and spatial connectivity between the objects and entities. For a machine to deal with images is itself a huge task and to deal with all the fine details within an image needs lots of techniques to be dealt with. In this paper we are going to have the overview of multiple methods which were invented and growing day by day so as to make the task of image captioning easier. Since it deals with images and sentences we are going to view the techniques needed to handle computer vision and natural language processing. To describe images we are going to study the deep learning methods.*
***Keywords:** Image Captioning, Deep Learning, CNN, RNN and Encoder-Decoder Framework*

## I. INTRODUCTION

The main purpose of this research is to introduce all the new algorithms and concepts related with field of CVPR (Computer Vision and Pattern Recognition). So as to make all the relevant information available at one place. Image captioning is one of the interesting area in Artificial Intelligence . This deals with understanding and recognizing the objects and their correlation. To identify the objects within the images, need to obtain image features. Before getting introduced with Deep machine learning methods, image features were retrieved using Traditional machine learning methods.

In traditional machine learning methods, hand crafted features Local Binary Patterns (LBP), the Histogram of Oriented Gradients (HOG), the Scale Invariant Feature Transform(SIFT) and the combination of such features are widely used. In these techniques features are extracted from input data and then passed to the Support Vector Machine (SVM) classifier to classify the objects.

In Deep machine learning techniques, basically CNN and RNN are used for image captioning purpose. In which CNN i.e. Convolutional Neural Networks are used to deal with images for image recognition and object classification. And RNN (Recurrent Neural Networks) are used to deal with the caption part i.e. text to be generated for the corresponding image. CNNs are used for feature learning followed by softmax as classifier and then RNN used for natural language processing. CNN and RNN are the two models dealing with different entities images and text respectively and combining them is necessary so as to relate the entities within images and words within description, this is carried out through multimodal embedding.

In further sections we will overview the overall procedure for image captioning, the datasets on which models could be trained, the neural network and other approaches followed as the core of image captioning task and the metrics on which accuracy is measured.

## II. PROCESS OF IMAGE CAPTIONING

The process generally followed to generate the textual description of images is a follows:

1) *Dataset:* As the initial thing we need to collect a dataset containing number of images with their descriptions which will be provided as input data to train the model. We will see different datasets which are freely available for educational purposes.
2) *Preprocessing:* To make the dataset available for training the model, dataset must be in a defined input format like <image_id, caption> or any other desired form. As image data and text data are not compatible so they are converted in vector format for training purpose and other operations.
3) *Training:* As the model is generated, the model should be trained against the dataset to understand the regional and syntactical correspondences between images and captions. So whichever dataset is used more than half of entries are used for training purposes. Training can be repeated to make the model learn to generate the desired description of
4) *Testing:* Once the model is trained against dataset, the model must be evaluated with the dataset which has not seen by the model before. Accuracy of model is confirmed if for test images, it generates the captions which were mentioned in the dataset.
5) *Generate New Captions:* Now as we have the completely ready model which is also evaluated on test dataset, is ready to generate the captions for entirely new images.

## III. RELATED WORK

### A. *Datasets Used*

Machine learning is a method of analyzing data to build a neural network model. We use Machine Learning Algorithms to iteratively learn from data [1]. There are many datasets used for training purpose. The pre-trained models again go under training for better accuracy. Following are few datasets which provides wide amount image-sentence pairs among which few are dedicated for training purpose and rest are for testing/ evaluating purpose. Following are the few datasets used.

1) *MSCOCO Dataset:* MSCOCO is a large dataset used for object detection, caption generation. The Microsoft COCO dataset contain 82,783 training images and 40,504 validation images, each with 5 descriptions.
2) *Flickr8k Dataset:* Flickr8k dataset is a dataset used for training the model which will be used as image to sentence descriptor. This contains a total of 8000 images with 5 descriptions for each image. Among which dedicated 6000 are used for training, 1000 are used for development and the rest 1000 are used for testing purpose.
3) *Flickr30k Dataset:* It has become a benchmark for image description. Dataset presents Flickr30k entities, which have captions from flickr30k with 244k chains, linking existence of same entities across different captions for the same image, and associating them with 276k annotated bounding boxes. Annotations of this type are necessary for continuous progress in automatic image description and grounded language understanding.
4) *Conceptual Captions:* This is the latest dataset and could be the challenge for image captioning. As this Conceptual Captions, a new dataset consist of about 3.3 million image-caption pairs that are formed with the use of billions of web pages on which image features' extraction and filtering of those features is performed.
5) *ImageNet Dataset:* ImageNet is a dataset which is organized according to wordNet hierarchy. WordNet contain near about 100,000 phrases and for these many phrases ImageNet provided around 1000 images on an average to illustrate a single phrase.

### B. *Methodologies Used*

Image Captioning is the task of describing a given image in terms of entities present within the images, their correlation, and the semantically accurate textual description of that image. And dealing with two completely different entities i.e images and text, we need to convert them into some comparable form in which their ranks will be compared based on same data type. Both are needed to convert into vectors. So one of the mechanisms is Encoder-Decoder. Another one is Attention Mechanism which mainly focuses on replicating natural human behavior. Once image is seen, people think before summarizing an image paying attention to specific object/region within an image and form the sentence keeping focus on that object. The same approach is used in attention model. It can be carried out in top-down and bottom-up approach but top-down approach is used most of the time since it gives more accuracy, in such a way that results produced with machine and human are similar.[8] The two keywords are trending now-a-days with reference to image captioning are Novel and Semantics. These keywords are important in solving the challenge which is: to generate the textual descriptions which are not distinguishable from human written ones. Implementing Semantics [2] in image captioning system means to impose sentiments into the system. Novel objects must be included for the expansion of scenarios. As the core part of image captioning using deep learning methods RCNN (Region based CNN) and BRNN (Bidirectional RNN) play a vital role in aligning visual data to descriptions to be semantically correct.

1) *Convolutional Neural Network:* Convolutional Neural Networks (ConvNets) is one of the main categories of neural networks which deals with the image classification, object detection, image recognition, etc. CNN acts as an encoder part in encoder-decoder mechanism. Suppose we are using CNN for image classification there are multiple layers which forms CNN model which are convolution layer with filters for feature extraction (convolution + ReLU), Pooling layer, Fully Connected layer and the softmax function to classify the object. The RCNN stands for region based CNN is the CNN focusing on regions of objects within the image.

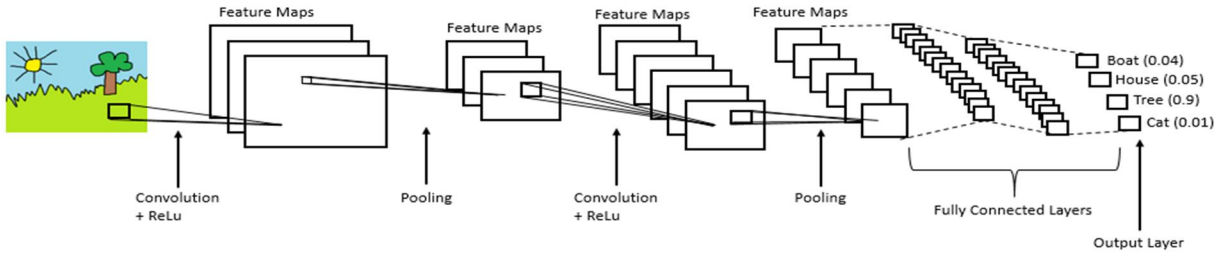



Figure 1 Complete CNN Architecture[11]

2) *Convolution Layer:* This layer contains one more layer with it which is ReLU i.e Rectified Linear Unit. Convolution layer and ReLU layer are used for extracting features like color, edges. This layer helps in identifying different features (such as performing blurring, sharpening operations) of images using different filters. And ReLU is a non-linear operation, intending to introduce non-linearity in our ConvNet, Since the real world data would want ConvNet to learn only about positive values. Feature map is generated at the end and is input to the next layer.
3) *Pooling Layer:* Pooling layer reduces the number of parameters when dealing with large images. Spatial Pooling is also called sub-sampling or down-sampling. Max Pooling, average pooling, sum pooling are the types of spatial pooling. In max pooling, the maximum of all the parameter values is passed to the next layer. Taking average of all elements of feature map is average pooling. Adding all the elements result in sum pooling.
4) *Fully Connected Layer:* The output of previous layer is flattened and a vector is generated. Then softmax or sigmoid functions [3] which are activation functions are applied on vectors for further classification of images. There are many CNNs developed with very few differences in each but have all the layers above in common. Following are the few CNNs used for feature extraction:

*a)* VGG-16 Net[9];
*b)* ResNet;
*c)* GoogleNet;
*d)* AlexNet;
*e)* DenseNet.

5) *Recurrent Neural Network:* RNNs are the neural networks developed to make use of sequential information. As RNNs deal with sentence formation in which next word is predicted based on the previous word. This task is repeated for all elements until the complete sentence is formed that is the reason this neural network is recurrent. Multimodal RNN used to generate novel image captions.[7]

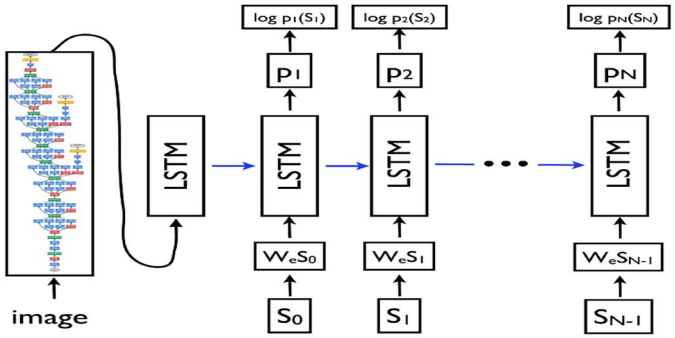



Figure 2 LSTM Model combined with a CNN image embedder [12]

6) *LSTM (Long Short Term Memory):* It is a special kind of RNN, capable of learning Long-term dependencies. LSTMs have a nature of remembering information for a longer period of time. A L.S.T.M network follows the pre-trained VGG16. The L.S.T.M network is used for language generation.[13]

### C. Evaluation Metrics

How could one measure the success of any machine learning model? Depending only on accuracy of model may lead to poor predictions when model is deployed on unseen data. So, evaluation metrics are bind to machine learning tasks. There are different metrics for different machine learning tasks like classification, regression, clustering, etc. The basic task behind using all the metrics is to measure the similarity between machine generated and human generated sentences. Some evaluation metrics used to calculate the score as follows:

1) *BLEU [4]: (*Bilingual evaluation Understudy) is an algorithm to evaluate the quality of the generated text. Quality in terms of comparing machine generated text and human generated text. The closer the two texts are the more will be the BLEU score. It is one of the first metric attaining high correlations. This is the precision based.
2) *Meteor:* (Metric for Evaluation of Translation with Explicit ORdering) it is a metric to evaluate machine translation output. This metric is based on the phenomena of harmonic mean of unigram precision and recall, where recall has higher weight than precision.

3) *Rouge:* (Recall-Oriented Understudy for Gisting Evaluation) it is a set of metrics to evaluate automatic text summarization. This is the recall based metric.
4) *CIDEr:* (Consensus based Image Description Evaluation) this measures the closeness of machine generated sentence against the one by human. Using sentence harmony, this metric identifies the notions like saliency, precision and recall.[5]

## IV. CONCLUSION

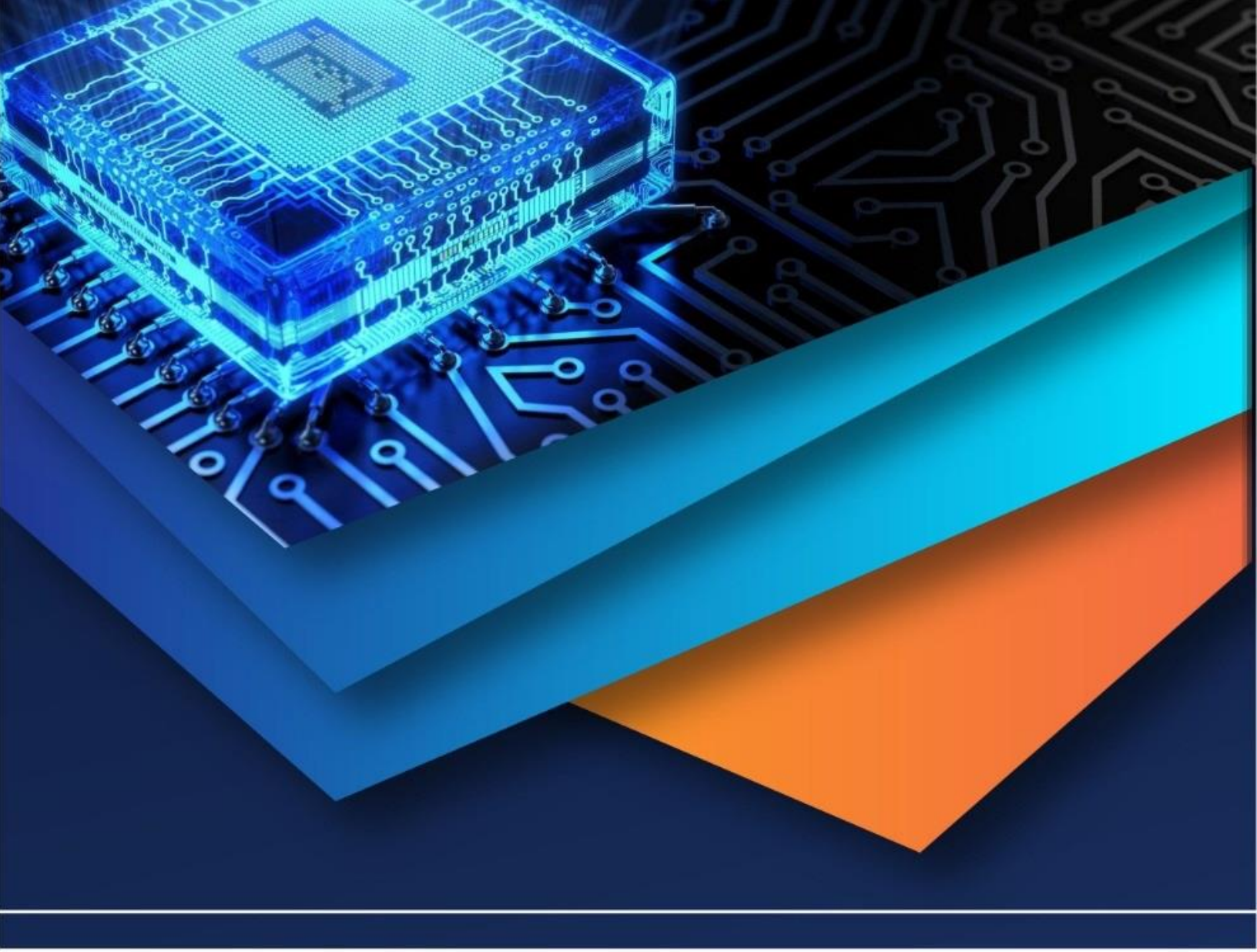
Image Captioning is a kind of field of study which is like never ending process as it gets improved with better algorithms and vast amount of data used for training the model. There are many articles which provide lot information. This paper tries to summarize all of those papers or all of the things which come in discussion when talking about image captioning. And there is always scope for improvement. We hope this paper could help other aspirants with the newest information to image captioning and introduce all the aspects of it.

## V. ACKNOWLEDGMENT

I want to express my sincere gratitude to the guide of the project Prof. Shah H. P. for her valuable collaboration and guidance throughout my research. I would also like to thank to our respected head of the department Prof. Tandle S. R. and all the faculty members and friends for their cooperation.

## REFERENCES

[1] J. Donahue, L. A. Hendricks, S. Guadarrama, M. Rohrbach,S. Venugopalan, K. Saenko, and T. Darrell. “Long-term recurrent convolutional networks for visual recognition and description” CoRR,vol. Abs/1411.4389, 2014.
[2] You, Q.; Jin, H.; Luo, J. “Image captioning at will: A versatile scheme for effectively injecting sentiments into image descriptions”. arXiv 2018, arXiv:1801.10121.
[3] Vijayaraju, Nivetha. "Image Retrieval Using Image Captioning" (2019).Master's Projects. 687.
[4] K.Papineni,S.Roukos,T.Ward,andW.-J.Zhu. “Bleu:a method for automatic evaluation of machine translation”. In Proceedings of the 40th annual meeting on association for computational linguistics, pages 311–318. Association for Computational Linguistics, 2002.
[5] Ramakrishna Vedantam, C. Lawrence Zitnick, Devi Parikh.” CIDEr: Consensus-based Image Description Evaluation”. arXiv:1411.5726v2 [cs.CV] 3 Jun 2015.
[6] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. “Show and Tell: Lessons learned from the 2015 MSCOCO Image Captioning Challenge” arXiv:1609.06647v1 [cs.CV] 21 Sep 2016.
[7] J. Mao, W. Xu, Y. Yang, J. Wang, Z. Huang, and A. Yuille.”Deep captioning with multimodal recurrent neural networks (m-rnn)”. ICLR, 2015.
[8] He, S. Tavakoli, H.R.; Borji, A.; Pugeault, N. “A synchronized multi-modal attention-caption dataset and analysis”. arXiv 2019, arXiv:1903.02499..
[9] Simonyan, Karen, and Andrew Zisserman. "Very deep convolutional networks for large-scale image recognition." arXiv preprint arXiv:1409.1556 (2014).
[10] Yin Cui1, Guandao Yang, Andreas Veit, Xun Huang, Serge Belongie. “Learning to Evaluate Image Captioning” arXiv:1806.06422v1 [cs.CV] 17 Jun 2018.
[11] R. Prabhu. “Understanding of Convolutional Neural Network (CNN)— Deep Learning”. [Online].Available: https://medium.com/@RaghavPrabhu/understanding-of-convolutional-neuralnetwork-cnn-deep-learning-99760835f148.
[12] S. Ioffe and C. Szegedy, “Batch normalization: Accelerating deep network training by reducing internal covariate shift,” in ICML, 2015.
[13] Amey Arvind Bhile, Varsha Hole. “Real-Time Environment Description Application for Visually Challenged People”. ICCNCT 2019.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

# INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY