Available online at: http://www.ijmtst.com/vol6issue04.html

International Journal for Modern Trends in Science and Technology

ISSN: 2455-3778 :: Volume: 06, Issue No: 04, April 2020

A Survey on Object Detection Based Automatic Image Captioning using Deep Learning

Trushna Patel[1] | Dr. Darshak G Thakore[1] | Dr. Narendra M Patel[1]

[1]Department of Computer Engineering, Birla Vishvakarma Mahavidyalaya, Vallabh Vidyanagar, Gujarat, India

To Cite this Article

Trushna Patel, Dr. Darshak G Thakore and Dr. Narendra M Patel, "A Survey on Object Detection Based Automatic Image Captioning using Deep Learning", *International Journal for Modern Trends in Science and Technology*, Vol. 06, Issue 04, April 2020, pp.:274-280.

Article Info

Received on 25-March-2020, Revised on 15-April-2020, Accepted on 19-April-2020, Published on 23-April-2020.

ABSTRACT

The aim of an Image Caption Generation system is to generate captions for an image. It provides a descriptive sentence for an image, that helps people to better understand the semantic meaning of an image. Image Captioning is an application for both Natural Language Processing and Deep Learning. Manually writing captions for an image is a tedious task. Instead, this will automatically generate captions for the input image. This enables the necessity for detecting objects and establishing relationships among objects. Feature extraction and Scene Classification are used for extracting some useful features and understanding the semantics behind the scene from the image. Both these techniques enable the use of the Convolutional Neural Network(CNN). For generating captions, it is necessary for a system to establish a relation between phrases and objects. Recurrent Neural Network(RNN) is such a method that enables connecting words and phrases and providing a descriptive sentence. The paper provides a brief survey of certain requirements for Image Captioning such as Image Captioning methods, datasets, and evaluation metrics. This paper also discusses the categories of image caption generation.

KEYWORDS: *Image Captioning, Object Detection, Sentence Generation, Image to Text Conversion.*

I. INTRODUCTION

Nowadays, every day, one can come across many images from various sources such as news articles, internet, documents or advertisements[1]. A human can interpret the image content easily while a machine may face difficulty in understanding the semantics of an image. Image Captioning refers to such a system, which generates captions for an image automatically. It refers to analyzing an input image, extracting objects, establishing relationships among the objects and then generating a description for the image. It also refers to understanding the scene of an image, the object attributes, and their interactions. Image Captioning provides effective use of Natural Language Processing and Image Processing, that together enables user automation in life. The difficult task in this system may be to detect image objects and create a description. For description generation, the words and their relationship must be established. The sentence may be generated by understanding the semantic meaning of the image and the predicted words. In order to understand the semantics behind the image, the relationship between the objects that are extracted from an input image. Image Captioning can be used in many areas such as image web searching, social media, military, education, and video caption generation.

The process requires establishing the relation between image features and word sequences. This enables the system to create a sentence using words extracted fromthe image. Image Captioning provides the textual form of an image by visualizing the image content.

There are two broad categories for image feature extraction namely, Traditional Machine Learning-Based Technique and Deep Learning-Based Technique. Feature Extraction refers to a reduction in the number of features in the dataset by creating new features from existing features[2]. Machine Learning provides some handcrafted features such as Local Binary Patterns (LBP), Scale-Invariant Feature Transform (SIFT), the Histogram of Oriented Gradients (HOG), and Speeded Up Robust Features(SURF)[1]. These techniques perform feature extraction on an image while Support Vector Machine(SVM) performs the classification of an object[1]. Since these features are not feasible for large datasets, Deep Learning features can be used. It helps to extract features from a large and diverse dataset of images and videos using the Convolutional Neural Network(CNN)[1]. CNN uses a Softmax activation function for the classification of an object[1].

Most of the existing Image Captioning articles rely on the three broad categories of Image Captioning methods: Template-Based Image Captioning, Retrieval-Based Image Captioning, and Novel Image Captioning[1]. Certain Image Captioning methods will be described in a further section.

Fig. 1 and Fig. 2 provide users to better understand an image captioning system. With the help of these sample images, the user can also understand the input image(in Fig. 1) and output caption(in Fig. 2) of an image captioning system.

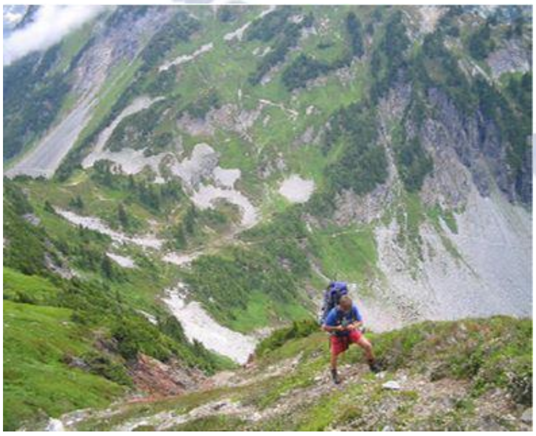
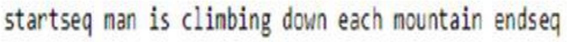


Fig. 1 Input Image [3]



startseq man is climbing down each mountain endseq

Fig. 2 Associated Caption[3]

II. Image Captioning System

As discussed, Image Captioning methods have three variations, Template-Based, Retrieval-Based and Novel Image Captioning. Template-Based Image Captioning uses fixed templates with a number of blank slots to generate a sentence. In this approach, blank slots are filled with the detected objects, attributes, and actions from an image[1]. The example of the Template-Based approach is provided byD. Hutchison et al.[4] in their paper, which uses triplet <*object, action, scene*>of scene elements to fill the blank slots. Another example of a template-based approach is provided by S. Li et al.[5], which extracts the phrases from the detected objects, attributes and their relationships for generating captions. Though this approach provides grammatically correct captions, it does not generate variable length captions. In the Retrieval-Based Approach, captions are retrieved from the existing captions[1]. In this approach, a visually similar image and caption are extracted from the training dataset. The corresponding caption is retrieved from a pool of captions[1]. The example of this approach is determined by M. Hodosh[6] and P. Kuznetsova[7]. Though this approach generates syntactically correct captions, it does not generate semantically correct captions. In the Novel Image Caption Generation Approach, it analyzes the image content and then generates image captions with the help of language models[1]. This approach generates new captions for each image that are accurate and semantically correct. Most novel caption generation systems use deep learning techniques, thus it is the main focus of this survey. The deep learning approach is used by many authors([3], [8]–[12]) in order to achieve Image Captioning. All of these papers use CNN as a deep learning technique for feature extraction and object detection.

M. Z. Hossain et al.[1] in their paper, discussed various types of Deep Learning methods such as Visual Space vs. Multimodal Space, Dense Captioning vs. Captions for the whole scene, Encoder-Decoder Architecture vs. Compositional Architecture and LSTM vs. Others.

According to M. Z. Hossain et al.[1], in Visual Space-Based methods, image features and their captions are separately passed to the language decoder while in Multimodal-Based methods, a shared multimodal is learned from an image and their captions and passed to the language.

Dense Captioning is well described by M. Z. Hossain et al.[1], in which, only a single caption is generated for the entire image. It uses different image regions to extract information of image objects and provide region-wise captions whereas the latter method provides a caption for the whole image, irrespective of the image regions[1]. This method generates a sentence for each object of an image and combines them to create a full image description.

In Encoder-Decoder[1] based methods, it provides an end to end manner of generating captions using encoder-decoder architecture. This method contains an encoder neural network that converts an image into an intermediate form and a decoder recurrent neural network that generates a description[13]. It uses CNN for image feature extraction and then fed to RNN for the generation of words related to an image[1]. In Compositional architecture[1] based methods, image captioning is composed of several independent methods. This method integrates independent building blocks into a pipeline to generate captions[13]. It uses CNN for image understanding and then a set of candidate caption is generated. The final caption to be generated as output by re-ranking these candidate captions[1].

M. Z. Hossain et al.[1], in their paper, discussed variations of RNNs for sentence generation. LSTM is a type of RNN with a memory cell that maintains the information of an image over a long period of time. LSTM is used in sequence to the sequence learning task. LSTM contains various gates such as the input gate, the output gate, and the forget gate[14]. These gates that learns what information is relevant to keep or forget[14]. A Gated Recurrent Unit(GRU) is similar to LSTM but does not use a separate memory cell and uses less number of gates to flow the information[13]. It contains two types of gates namely, update gate and reset gate. The Update gate takes care of what information to keep and what to throw away while the Reset gate decides how much past information to forget[14].

N. K. Kumar et al.[3], in their research,provided various tasks for generating descriptions such as understanding visual representations of objects, the relationship between objects and generating a sentence. This research work includes multiple methods. It uses Region Proposal Approach for Object Detection, CNN(Convolutional Neural Network) for Feature Extraction and Scene Classification, RNN(Recurrent Neural Network) for human and object attributes and for generating a description of the image. C. Amritkar et al.[8] proposed a different way of generating captions. It uses a pre-trained VGG16 CNN model, RNN model, and LSTM model. LSTM is a storing unit in RNN that stands for Long Short Term Memory. It observed categories in types of results like a description with errors, without errors, related and unrelated captions. These categories are due to considerations of the neighborhood of words. P. Shah et al.[9] provides the use of Show and Tell method for image captioning. Show and Tell method is an advancement in image recognition and neural machine translation in image captioning. It is a combination of the Inception-v3 model and LSTM. Inception-v3 model is used for object recognition and LSTM is used for storing intermediate words during sentence generation. An input image is given to the Inception-v3 model. At the end of this model, a single fully connected layer is added, which transforms the output of the Inception-v3 model into a word embedding vector. With the help of this vector, LSTM generates sentences. F. Fang et al.[10] proposed a Word Level Attention model for Image Captioning. A word-level attention layer is designed to process image features with two models for accurate word prediction. The first mode is the Bidirectional spatial embedding module to handle feature maps. The second model is Attention to extract word-level attention. Word Level Attention Layer has Line Level Bidirectional Embedding used to process features using bidirectional LSTM networks. It has also Word Level Attention Extraction used to extract visual information to predict the next word using a softmax activation function. D.-J. Kim et al.[11] focuses on better sentence learning using Deep Convolutional Network. Image feature extraction is implemented using deep fisher kernel. An author has used fine-tuning CNN on sentences and datasets for aggregating the activations from CNN into the fisher vector. All the activations are aggregated to form a Fisher Vector. Instead of LSTM, this approach uses gLSTM. Since the information of the input image is only provided in the first step, the input image information gets diluted as it proceeds. gLSTM provides a piece of additional information called 'guide', that pertains to the input image information throughout the process. A. Poghosyan et al.[12] proposed a new approach that uses LSTM with Read-Only LSTM. LSTM cell provides the facility to store while the modified LSTM cell will provide image features. As the input image content is only provided in the first step, LSTM generated word may not be associated with the input image description. To predict the word related to both previous word and input image content, an additional unit is used, which is LSTM known as LSTM with Read-Only Unit. The Read-Only Unit along with LSTM provides better accuracy for caption generation.

Fig. 3 illustrates a general workflow from literature for an Image Captioning System proposed by X. Yin et al.[15]. According to them, a caption generation system using object detection should comprise of feature extraction, object detection, and sentence generation. The image from a dataset is input to the system. Object Detection is performed on the input image in order to get an object and its location. Features are extracted from the image using the CNN pre-trained model. The output of Object Detection and Feature Extraction are encoded in the same size to input it in the RNN model. The RNN language model is used to generate a description of the image.

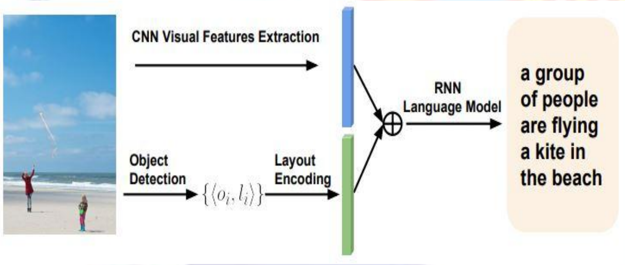

CNN Visual Features Extraction
Object Detection
$\{(o_i, l_i)\}$
Layout Encoding
RNN Language Model
a group of people are flying a kite in the beach

Fig. 3General Work Flow[15]

III. Survey Of Existing Methods

There are certain methods that can be carried out in order to implement image captioning. These methods include Object Detection, Feature Extraction, Scene Classification, and Sentence or Caption Generation. Object Detection method combines two methods viz. Image Classification and Object Localization. Image Classification is used for classifying the detected object while Object Localization is used for locating the specific object from an image. R-CNN family is a popular technique for this purpose. R-CNN stands for Region-Based Convolutional Neural Network. Certain variations of Region-Based techniques are R-CNN, Fast R-CNN, Faster R-CNN, and Mask R-CNN.. All these region-based object detections are slow due to the high computations of the regions. Thus an object detection method called YOLO(You Only Look Once)[16] was introduced.

Region convolutional neural network(RNN)[17]:Itis used to detect the region from the image which contains the object. The R-CNN uses the Selective Search method for extracting regions from an image in order to detect objects.

RCNN model contains the four steps. It takes an image as the input and extracts the region proposal that is enclosed into a square. The extracted Region Proposal is provided into aConvolutional Neural Network that produces a feature vector. The last extracted feature is fed into a Support Vector Machine to calculateand classify the objects into class labels.

You Only Look Once(YOLO)[16]: YOLO is an Object Detection method that divides the input image into the S*S grid. Each grid is responsible for detecting one object. Together with object detection, each grid is responsible for predicting a fixed number of bounding boxes. Each grid cell predicts B bounding boxes and a bounding box score. It also provides C conditional class probabilities and one object per grid irrespective of a number of bounding boxes. The bounding box provided by this method contains various parameters such as X-Coordinate(x), Y-Coordinate(y), Height of box(h), Width of the box(w) and confidence. If the confidence is high, the probability of an object in that box is high. The class probability provides the category to which the extracted object belongs to. There may be many bounding boxes, but the highest confidence and area intersected by several bounding boxes contain the object. Mask R-CNN is similar to YOLO method, but it provides a precise masked object along with the bounding box

Single Shot Detector(SSD)[18]: SSD is a Feed-Forward Convolutional Neural Network that generates a fixed-size collection of bounding boxes and the scores of the object for its presence in the box. Along with these, in the final step, a non-max supression is used for final detections. A Non-Max Supression[19] is a post-processing method that is used to smooth the response. SSD[18] uses a VGG16 model for the early layers in its network. SSD is a simple object detection method relative to other methods that require region proposals. Instead, it encapsulates all the computations required for object detection in a single network.

Convolutional Neural Network(CNN)is used for an effective caption generation by understanding the features and semantics of the scene. Feature Extraction is used to extract important features from an image to understand the image content. It reduces the dimensions of raw data to process the image content. Scene Classification is a technique for understanding semantics behind the detected objects and meaningful features. It helps to provide the correct meaning of an object and feature according to the scene attributes. Both these techniques are implemented using the Convolutional Neural Network(CNN)[20], with the help of VGG model.

CNN[20]: It is widely used in areas like image classification, object detection, and recognition. In image classification, CNN processes an input image and categorizes it into classes. An input image passes through a number of convolutional layers. There are three types of layers in a CNN - Convolutional Layer, Pooling Layer, and Fully Connected Layer. Convolutional Layer is responsible for extracting features from the input image. It maintains the relation between pixels of an input image and performs convolution using an input image and a kernel. Based on the type of kernel, Feature Map is generated as an output. The activation function that is used by CNN is ReLU, which stands for the Rectified Linear Unit for a non-linear operation. It finds maximum from an input, which implies that, the pooling layer reduces the number of parameters from an image. Down-sampling or subsampling in CNN refers to which reduces the dimensions of a feature map but pertains to the important features of an image. Max Pooling and Min Pooling are used to find maximum and minimum among the kernel size respectively. The Fully Connected layer which uses Softmax or Sigmoid Activation Function to classify the outputs into classes.

The extracted object attributes and scene attributes are used for sentence generation. In the Image Captioning system, the most important part is to generate an effective sentence with the help of those attributes. There are several methods to generate captions such as Recurrent Neural Network(RNN)[21], Long Short Term Memory(LSTM) and Gated Recurrent Unit(GRU).According to the author, RNN[21]can be described as an application in sentiment classification, image captioning, Language translation.

RNN[21]: It's basic task is to what next word will be in a sentence. In RNN firstly a neuron is supplied to the network it is a one-time step then calculation of its current state using a combination of the current input and the previous state. As the problem demands, compute and combine the information from all the previous states. After completion of all steps, the final current state is used to calculate the output. The output is afterward compared to the original output and the error is computed. The error is then backpropagated to the network to update the weights. LSTM is used to store the intermediate words during caption generation until the end of the sentence. Using LSTM, one can ensure the relevance of the current word from all the previous words. LSTM and GRU aim in solving Vanishing Gradient Problem of RNN.

Long Short Term Memory[22]: LSTM networks layer consists of a set of recurrently connected blocks known as memory. It contains a cell and three units: input gate, output gate and forget gate. The LSTM cell keeps values over an arbitrary time period and LSTM gates regulate the flow of information into and out of the cell. GRU is a variation of LSTM with fewer gates.

GRU(Gated Recurrent Unit)[23]: It contains two gates viz. update gate and reset gate. An Update gate helps to decide how much information of past needs to be passed to the future while the Reset gate decides how much past information to forget.

IV. Datasets And Evaluation Metrics

There are three datasets available for Image Captioning – Flickr8K[24], Flickr30K[25] and MSCOCO[26]. Flickr8K Dataset[24] is used for caption generation. It has 8000 images with 6000 training images, 1000 validation images, and 1000 testing images. Flickr30K Dataset[25] is a dataset for image description and grounded language understanding[1]. It has 30000 images with 28000 training images, 1000 validation images, and 1000 testing images. MSCOCO Dataset[26]is used for an Image Recognition, Object detection and Image Captioning dataset provided by Microsoft. MSCOCO contains 328K images with 82783 training images, 40504 validation images, and 40775 testing images. Table 1. provides the sample images and their associates captions for each of the mentioned datasets.

Image Captioning technique provides various evaluation metrics to evaluate the generated captions. The most popular metric is BLEU(Bilingual Evaluation Understudy)[27], which is used to evaluate a machine-generated text. The generated text segments are compared with the set of reference text and scores for each text segment[1]. The overall evaluation can be determined by averaging that individual text. Though the BLEU score is popular for machine translation, it is only suitable for short captions[27].Another metric is METEOR(Metric for Evaluation of Translation with explicit ORdering)[28], which is used to evaluate a machine-generated text. It compares the word segment with reference text. It also provides matching of synonyms of words, thus making a better correlation of sentence[1]. CIDEr(Consensus-Based Image Description Evaluation)[29] is a metric for evaluation image descriptions. It also provides a consensus between generated and human suggested descriptions[1].

V. Summary

Image Captioning system is a technique for generating the descriptions for an image. This paper concludes the variety of methods to implement a caption generation system. Certain methods include Template-Based Image Captioning, Retrieval-Based Image Captioning, and a Novel Image Caption Generation system. The advantages and limitations of these methods are also discussed in this paper. It also provides a comprehensive survey for different techniques used in an effective description generation. The survey also provides several datasets for image captioning such as Flickr8K[24], Flickr30K[25] and MSCOCO[26] Dataset. Captions generated using deep learning techniques can be evaluated using certain evaluation metrics such as BLEU[27], METEOR[28], and CIDEr[29].

<table>
<tr><th>Dataset Name</th><th>Sample</th><th>Associated Captions</th></tr>
<tr><td>Flickr8K [24]</td><td> </td><td>• 102455176_5f8ead62d 5.jpg#0 A man uses ice picks and crampons to scale ice.
• 102455176_5f8ead62d 5.jpg#1 an iceclimber in a blue jacket and black pants is scaling a frozen ice wall.
• 102455176_5f8ead62d 5.jpg#2An ice climber scaling a frozen waterfall.
• 102455176_5f8ead62d 5.jpg#3 A person in blue and red ice climbing with two picks.
• 102455176_5f8ead62d 5.jpg#4 Climber climbing an ice wall</td></tr>
<tr><td>Flickr30 K[25]</td><td> </td><td>• 1000919630.jpg| 0| A man sits in a chair while holding a large stuffed animal of a lion.
• 1000919630.jpg| 1| A man is sitting on a chair holding a large stuffed animal.
• 1000919630.jpg| 2| A man completes the finishing touches on a stuffed lion.
• 1000919630.jpg| 3| A man holds a large stuffed lion toy.
• 1000919630.jpg| 4| A man is smiling at a stuffed lion</td></tr>
<tr><td>MSCOC O[26]</td><td> </td><td>• {"image_id": 100563,"id": 533194,"caption": "An airplane parked on the runway next to some trucks."}
• {"image_id": 100563,"id": 535342,"caption": "An airliner parked next to a jetway at the airport"}
• {"image_id": 100563,"id": 538138,"caption": "A white and grey airplane sits at a gate at an airport."}
• {image_id": 100563,"id": 539827,"caption": "A plane parked on a tarmac of an airport"}
• {"image_id": 100563,"id": 540349,"caption": "An airplane at the gate about to be pushed back"}</td></tr>
</table>

Table 1. Dataset Image and Associated Caption[24],[25],[26]

References

[1] M. Z. Hossain, F. Sohel, M. F. Shiratuddin, and H. Laga, "A Comprehensive Survey of Deep Learning for Image Captioning," *ArXiv181004020 Cs Stat*, Oct. 2018.

[2] P. P. Ippolito, "Feature Extraction Techniques," *Medium*, 11-Oct-2019. [Online]. Available: https://towardsdatascience.com/feature-extraction-techniques-d619b56e31be. [Accessed: 24-Dec-2019].

[3] N. K. Kumar, D. Vigneswari, A. Mohan, K. Laxman, and J. Yuvaraj, "Detection and Recognition of Objects in Image Caption Generator System: A Deep Learning Approach," in *2019 5th International Conference on Advanced Computing Communication Systems (ICACCS)*, 2019, pp. 107–109, doi: 10.1109/ICACCS.2019.8728516.

[4] D. Hutchison *et al.*, "Every Picture Tells a Story: Generating Sentences from Images," in *Computer Vision – ECCV 2010*, vol. 6314, K. Daniilidis, P. Maragos, and N. Paragios, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2010, pp. 15–29.

[5] S. Li, G. Kulkarni, T. L. Berg, A. C. Berg, and Y. Choi, "Composing Simple Image Descriptions using Web-scale N-grams," p. 9

[6] M. Hodosh, P. Young, and J. Hockenmaier, "Framing Image Description as a Ranking Task Data, Models and Evaluation Metrics Extended Abstract," p. 5.

[7] P. Kuznetsova, V. Ordonez, A. Berg, T. Berg, and Y. Choi, "Collective Generation of Natural Image Descriptions," in *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Jeju Island, Korea, 2012, pp. 359–368.

[8] C. Amritkar and V. Jabade, "Image Caption Generation Using Deep Learning Technique," in *2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA)*, 2018, pp. 1–4, doi: 10.1109/ICCUBEA.2018.8697360.

[9] P. Shah, V. Bakrola, and S. Pati, "Image captioning using deep neural architectures," in *2017 International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS)*, 2017, pp. 1–4, doi: 10.1109/ICIIECS.2017.8276124.

[10] F. Fang, H. Wang, and P. Tang, "Image Captioning with Word Level Attention," in *2018 25th IEEE International Conference on Image Processing (ICIP)*, 2018, pp. 1278–1282, doi: 10.1109/ICIP.2018.8451558.

[11] D.-J. Kim, D. Yoo, B. Sim, and I. S. Kweon, "Sentence learning on deep convolutional networks for image Caption Generation," in *2016 13th International Conference on Ubiquitous Robots and Ambient Intelligence (URAI)*, 2016, pp. 246–247, doi: 10.1109/URAI.2016.7625747.

[12] A. Poghosyan and H. Sarukhanyan, "Short-term memory with read-only unit in neural image caption generator," in *2017 Computer Science and Information Technologies (CSIT)*, 2017, pp. 162–167, doi: 10.1109/CSITechnol.2017.8312163.

[13] S. Bai and S. An, "A Survey on Automatic Image Caption Generation," *Neurocomputing*, vol. 311, May 2018, doi: 10.1016/j.neucom.2018.05.080.

[14] M. Nguyen, "Illustrated Guide to LSTM's and GRU's: A step by step explanation," *Medium*, 10-Jul-2019. [Online]. Available: https://towardsdatascience.com/illustrated-guide-to-lstms-and-gru-s-a-step-by-step-explanation-44e9eb85bf21. [Accessed: 01-Jan-2020].

[15] X. Yin and V. Ordonez, "OBJ2TEXT: Generating Visually Descriptive Language from Object Layouts," *ArXiv170707102 Cs*, Jul. 2017.

[16] J. Hui, "Real-time Object Detection with YOLO, YOLOv2 and now YOLOv3," *Medium*, 27-Aug-2019. [Online]. Available: https://medium.com/@jonathan_hui/real-time-object-detection-with-yolo-yolov2-28b1b93e2088. [Accessed: 25-Nov-2019].

[17] R. Gandhi, "R-CNN, Fast R-CNN, Faster R-CNN, YOLO — Object Detection Algorithms," *Medium*, 09-Jul-2018. [Online]. Available: https://towardsdatascience.com/r-cnn-fast-r-cnn-faster-r-cnn-yolo-object-detection-algorithms-36d53571365e. [Accessed: 07-Jan-2020].

[18] W. Liu *et al.*, "SSD: Single Shot MultiBox Detector," *ArXiv151202325 Cs*, vol. 9905, pp. 21–37, 2016, doi: 10.1007/978-3-319-46448-0_2

[19] R. Rothe, M. Guillaumin, and L. Van Gool, "Non-maximum Suppression for Object Detection by Passing Messages Between Windows," in *Computer Vision -- ACCV 2014*, vol. 9003, D. Cremers, I. Reid, H. Saito, and M.-H. Yang, Eds. Cham: Springer International Publishing, 2015, pp. 290–306.

[20] Prabhu, "Understanding of Convolutional Neural Network (CNN) — Deep Learning," *Medium*, 21-Nov-2019. [Online]. Available: https://medium.com/@RaghavPrabhu/understanding-of-convolutional-neural-network-cnn-deep-learning-99760835f148. [Accessed: 25-Nov-2019].

[21] "Fundamentals of Deep Learning – Introduction to Recurrent Neural Networks." [Online]. Available: https://www.analyticsvidhya.com/blog/2017/12/introduction-to-recurrent-neural-networks/. [Accessed: 25-Nov-2019].

[22] "Long short-term memory," *Wikipedia*. 28-Dec-2019.

[23] S. Kostadinov, "Understanding GRU Networks," *Medium*, 10-Nov-2019. [Online]. Available: https://towardsdatascience.com/understanding-gru-networks-2ef37df6c9be. [Accessed: 07-Jan-2020].

[24] "Flickr8K." [Online]. Available: https://kaggle.com/shadabhussain/flickr8k. [Accessed: 25-Nov-2019]

[25] "Image captioning." [Online]. Available: https://kaggle.com/hsankesara/image-captioning. [Accessed: 25-Nov-2019].

[26] "COCO - Common Objects in Context." [Online]. Available: http://cocodataset.org/#home. [Accessed: 25-Nov-2019].

[27] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a Method for Automatic Evaluation of Machine Translation," in *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, Philadelphia, Pennsylvania, USA, 2002, pp. 311–318, doi: 10.3115/1073083.1073135.

[28] S. Banerjee and A. Lavie, "METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments," p. 8.

[29] "[1411.5726] CIDEr: Consensus-based Image Description Evaluation." [Online]. Available: https://arxiv.org/abs/1411.5726. [Accessed: 24-Dec-2019].