GENERATING SENTENCES BASED ON IMAGE USING DEEP LEARNING

Mr. S.M. Shinde, Priya Landage, Kajal Landage & Aishwarya Gurav

[1]*Assistant Professor, Computer Science and Engineering, SVERI's College of Engineering, Punyashlok Ahilyadevi Holkar Solapur University, Solapur, Maharashtra, India.*
[2,3,4]*Computer Science and Engineering, SVERI's College of Engineering, Punyashlok Ahilyadevi Holkar Solapur University, Solapur, Maharashtra, India*

Abstract: *Describing the contents of an Image is an easy task for Human Beings. To achieve the same using computers is a rather difficult task. To correctly describe an image, not only accurate recognition of objects is required, but also their attributes, relationships and scene information is required. This idea of describing contents of an image can be of a lot of help for elderly and blind people, but it is also challenging. To achieve this, we utilize the regional image details. Each region goes through the process of Object Detection, and combine multiple techniques to learn their attributes. The result with this is high level labels that can be used to generate the description of Image. Functions like scene Classification, attribute learning, sentence generation, and relationship detection, object detection and classification are also included in the system.*

Keywords: **Machine Learning, Deep Learning, Computer Vision, Image Classification, Image Localization, Object Detection.**

1. INTRODUCTION

Human beings are able to classify and describe most of the Images quite accurately. These descriptions are usually rich in details such as the object of interest, its neighboring objects and their interactions. Because they are in sentence form, much is omitted by Humans which they find less significant. As pictures can be represented digitally, it covers all the details that are not visible to the Human eye. However, achieving the same using computers is a rather difficult task. Recent advancements have shown that not only the label, but the surrounding often plays a vital role in determining the action in the Image [1].

In this paper, we try to achieve short descriptions of the Image in English. To reduce the number of comparisons, we limit the number of objects to two per image. Research contributions of this paper are summarized below:

- This research utilizes local image regions to initialize image descriptions rather than focuses on single holistic feature of an input image [2].
- It generates more descriptive labels and attributes for both objects and people for subsequent sentence generation.
- The system possesses the capability to successfully describe images from outside the domain of training. Most state-of-the-art previous methods were only applicable on the data-sets that they have been trained on. By combining multiple data sets, this system can be used on a much wider variety of images for testing.

Applications of this system range from Hazard Detection, Autonomous Driving, Image search, Anomaly Detection among others.

2. Literature Survey

The current state-of-art techniques of Object Classification include Support Vector Machine [SVM] and Convolutional Neural Network [CNN]. A brief description of these methods is given below:

SVM [Support Vector Machine] - Support Vector Machines are based on the concept of decision planes that define decision boundaries. A decision plane can be defined as the plane that can divide different data points [having different member class] on the basis of features. SVM technique is able to outperform almost every image classification techniques [3]. The only limitation of SVM is that it is suitable when data-set is small, as it requires the entire data-set to classify the image. Thus, speed and size are the only limitations of SVM.

CNN [Convolutional Neural Network]: Convolutional networks are inspired by biological processes in the brain [4]. Convolution Networks usually consist of Convolution layers along with Average/Max Pooling Layers. Convolution layers help with learning small kernels that can extract features out of an Image, while Average or Max Pooling Layers are responsible for reducing the Input size in the Deeper layers. Together, they are able to build complex networks that can learn high scaled features without adding unnecessary connections as seen in the Fully Connected Artificial Neural network. This architecture enables it to scale the network on a very deep level, while also allowing it to reduce the number of parameters, and eventually computation to a very low level.

Traditional machine learning relies on shallow nets, which consists of a single Input layer [Neurons same as that of Image], single Output layer [Single Neuron for Regression, multiple neurons for Classification], and a single Hidden layer connecting the two. “Deep” in Deep Learning refers to the use of multiple Hidden layers stacked on top of each other, to learn complex functions that can approximate the given dataset. So Deep Learning simply means that the network contains more than three layers [including Input and Output] to approximate the function. The further you advance into the neural net, the more complex the features your nodes can recognize, since they aggregate and recombine features from the previous layer.

3. Methodology

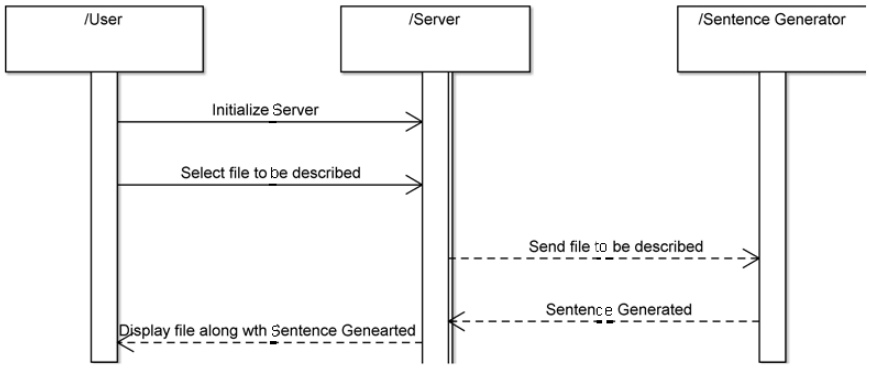
Artificial neural networks are structurally and conceptually inspired by the biological nervous system of Animals, mostly Human Beings. The earliest Neural Network to be constructed was the Perceptron. It consisted of input layer, an output layer, along with one Hidden layer that fully connected the above two. It was good to classify linearly separable patterns.

To solve more complex problems, neural networks were introduced that had a layered architecture i.e., input layer, output layer and more hidden layers stacked on top of each other. Neural network consists of interconnected neurons that takes the input, performs some processing on the input data, and finally forward the current layer output to the next layer. The general architecture of neural network is such that the neurons perform weighted sum of the neurons in the previous layer, pass it through the activation function, and pass along the activated neurons to the next layer. Thus adding more hidden layer allows to deal with more complex patterns as hidden layer capture non-linear relationship. These neural networks are known as Deep Neural network. Deep learning provides new cost effective to train. Extra layers in DNN enable composition of features from lower Layers to the upper layer by giving the potential of modeling complex data.

Deep learning is the growing trend to develop automated application. It is improvement of artificial neural network that consist of more hidden layer that permits higher level of abstraction and improved image analysis. It becomes extensively applied method due to its recent unparalleled result for several applications i.e. object detection, speech recognition, face recognition and medical imaging.

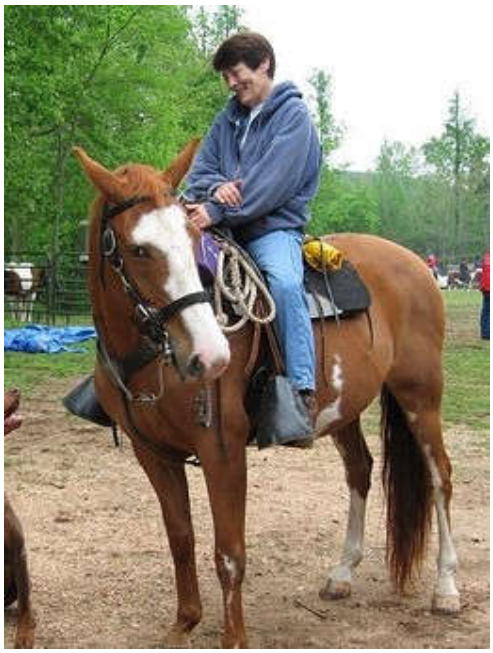
In our project, we have used Convolutional Neural Network [CNN] to classify images. CNN outperforms traditional image classification methods like SVM and Fully connected Neural network due to reduction in computations.

The following Image displays our proposed System Architecture:



4. Conclusion

Inspired by the recent work in machine learning and object detection, we introduce a model that can generate short descriptive sentences using the relative location of Objects [5]. We describe how we can train the Neural Network to learn attributes of objects and use their relative position to generate its short description. Applications of this system includes Autonomous Driving, Image search, Hazard Detection among others. Though this system does a good job, this fails under the criteria where objects are distant from each other, or are overlapped to a great extent. With the exception of these, the model performs remarkably well on photos of wide range of resolution. Following is an example Image followed by its description as tested by our system:



Sentence: Man is on Horse

Automatically generating captions for an image is a huge task that is very close to the entire domain of Computer Vision. Not only must caption generation models be able to solve the computer vision challenges of determining what objects are in an image, but they must also be powerful enough to capture and express their relationships in natural language.

5. References

[1] *Farhadi A. et al. (2010) Every Picture Tells a Story: Generating Sentences from Images. In: Daniilidis K., Maragos P., Paragios N. (eds) Computer Vision – ECCV 2010. ECCV 2010. Lecture Notes in Computer Science, vol 6314. Springer, Berlin, Heidelberg*

[2] *Zhongliang Yang, Yu-Jin Zhang, Sadaqat ur Rehman, Yongfeng Huang, "Image Captioning with Object Detection and Localization", Department of Electronic Engineering, Tsinghua University, Beijing*

[3] *Matsumoto M. (2012) SVM-Based Object Detection Using Self-quotient ε-Filter and Histograms of Oriented Gradients. In: Madani K., Dourado Correia A., Rosa A., Filipe J. (eds) Computational Intelligence. Studies in Computational Intelligence, vol 399. Springer, Berlin, Heidelberg*

[4] *Dhillon, A., Verma, G.K. Convolutional neural network: a review of models, methodologies and applications to object detection. Prog Artif Intell (2019)*

[5] *Jadczyk M., Tomczyk A. (2015) Object Localization Using Active Partitions and Structural Description. In: Rutkowski L., Korytkowski M., Scherer R., Tadeusiewicz R., Zadeh L., Zurada J. (eds) Artificial Intelligence and Soft Computing. ICAISC 2015. Lecture Notes in Computer Science, vol 9119. Springer, Cham*