Comparison and Analysis of algorithms used for detecting slums in satellite images

Mrs. Pallavi Saindane  
Asst. Professor  
Vivekanand Education Society's Institute of Technology  
Mumbai, Maharashtra, India  
pallavi.saindane@ves.ac.in

Gayatri Ganapathy  
Student  
Vivekanand Education Society's Institute of Technology  
Mumbai, Maharashtra, India  
2015gayatri.ganapathy@ves.ac.in

Nilesh Bhatia  
Student  
Vivekanand Education Society's Institute of Technology  
Mumbai, Maharashtra, India  
2015nilesh.bhatia@ves.ac.in

Aishwarya Vaidya  
Student  
Vivekanand Education Society's Institute of Technology  
Mumbai, Maharashtra, India  
2015aishwarya.vaidya@ves.ac.in

Neha Prabhavalkar  
Student  
Vivekanand Education Society's Institute of Technology  
Mumbai, Maharashtra, India  
2015neha.prabhavalkar@ves.ac.in

Abstract—This paper aims to redefine one of the applications of satellite images, slum identification. Slums also formally termed as informal settlements which can be identified from the satellite images using various image segmentation as well as object detection techniques. Policy makers spend substantial time and resources to discover certain regions. This paper presents a brief study of an algorithm that could be helpful for classifying residential areas as developed and underdeveloped using satellite images as input. This can further help to improve the approach required to identify the slums from satellite images. Such a process would be easily scalable and reduce time and resources. This will help developing and under-developed countries in investing appropriate resources and lead to economic progress.

Keywords—Slum identification, Satellite images, Convolutional Neural Networks.

I. INTRODUCTION

Underdeveloped areas in any city are a matter of concern for most policymakers and government officials. Detecting such areas and gaining knowledge of the economic status of such areas often involves work in three different dimensions that is legal, social and spatial. Tax/Revenue Department would help in legal dimension, Census Survey and local group/communities are approached for better understanding. Even then there still would exist areas which are unsurveyed due to unawareness of their existence. So the last plausible dimension is the spatial dimension, using satellite images to identify infrastructural facilities/connectivities. A need for an automated system which can accurately identify the underdeveloped areas of concern in an ongoing manner would prove to be of great influence in such cases. It is also necessary to understand that the features of a slum in a particular area or region cannot be applied to other regions. This paper explores various approaches to detect slums using satellite imagery and the accuracy, efficiency associated with these approaches.

II. METHODS

A. CNN

Convolutional Neural Networks are types of Neural Networks whose basic constituents are neurons which have weights and bias. A typical Convolutional Neural Network has an input layer, an output layer and many hidden layers which are used for the intention of Pooling, Rectifying, etc.

B. RCNN

RCNN stands for Region Convolutional Neural Network. An algorithm named Selective Search is used in order to extract region proposals from the input image which are then fed to a CNN whose purpose is to serve as a feature extractor.

C. Faster RCNN

Faster RCNN is considered to be an extended version of RCNN. In older versions, Selective Search was used which is a time consuming algorithm which can hamper the performance of a CNN. Hence, to identify region proposals, a new CNN is used which is called as Region Proposal Network (RPN).

D. Mask RCNN

Mask RCNN is the most efficient algorithm of all the other existing algorithms. It includes two basic modules viz, a CNN which is used for extraction of features and a RPN which is used in order to locate region proposals.
We stack the layers given below in CNN such that output of one layer becomes input for other layer. Layers can be repeated several times. So the image becomes more and more filtered as it goes through convolution layer and becomes smaller as it goes through pooling layer. ReLU layer is to keep the math from breaking out.

A. Convolution layer

CNN can be roughly classified into CNN for objects and CNN for grasping type. CNN takes binary image as input (light(1) or dark(-1) pixel) An image is considered as two dimensional array of pixel. The object or pattern inside an image can be straight, translated, scaled, rotated or weighted. As a human, we can see such differences clearly. But it is hard for a computer to differentiate or identify. Hence convolutional network analyses image by breaking it into smaller parts. Using these parts matching is done. For matching the parts of an image with image, filtering process is used. Filtering includes following:
1. Line up the image and the feature patch.
2. Multiply each image pixel by the corresponding feature pixel.
3. Add them up.
4. Divide by the total number of pixel in the feature.
Using this feature convolution tries to match each feature with the entire image. Hence after applying bunch of features, we get a stack of filtered image for each image. This was a convolution layer.

B. Pooling layer

In Pooling, we shrink the image stack. Shrinking the size of images makes the calculations much more manageable and easier. Process of max pooling is as follows:
1. Pick a window size (usually 2 or 3).
2. Pick a stride (usually 2).
3. Walk your image across your filtered images.
4. From each window, take the maximum value.
In pooling we don’t care about the position of maximum value in that window. This makes it little less sensitive to position.

C. ReLU layer

ReLU stands for rectified linear units. It is a computational unit which performs normalisation. In this process all the negative values in the filtered image are replaced by zeros. So, we perform this process with all other images in the stack.
At last, in fully connected layer, we take our stack of images and form a single list so it can become easier for visualization. Certain values in the list has more weightage when we feed certain image (as values represent the presence of features). Feature values in the list act as a list of votes. Initially, we may not get the correct features and correct value (votes) for the features. Hence we have to use the method of backpropagation. In backpropagation, error is calculated (error = right answer - actual answer) and is propagated back to the previous layers. Thus, the feature values will get modified according to the error. So designer have to decide about: - number of features and size of features in convolution, window size and window stride in pooling, number of neurons in fully connected layers. In architecture, one has to make decisions about the number of each type of layer and the order of the layers.
IV. STUDY OF RCNN

The process of object detection in RCNN algorithm is classified into the following two sub-processes:-

A. Selective Search

In order to locate region proposals, selective search is used for object detection. The fundamental principle behind the working of selective search is that it performs clustering of similar regions based on their size, texture, color and shape. The initial step in this technique is to over-segment the image considering an important aspect which is intensity of the pixels. The result expected from this step is segmented regions. Taking into consideration these segmented regions, the algorithm works as follows:-
1. Create bounding boxes corresponding to segmented parts to the list of proposed regions.
2. Group neighboring segments on the basis of similarity.
3. Perform Step 1.

So, the output of each iteration is segments which are smaller in size that add up into larger segments which get added to the region proposals’ list.

B. SVM and Classification

After obtaining region proposals using selective search algorithm, they are combined into a single unit and fed to the convolutional neural network. Here, CNN is used as a feature extractor and the result generated is a 4096 dimension feature vector. The output is given to an SVM classifier which classifies the presence of objects contained in the proposed regions. An SVM classifier classifies the data points by separating them using a hyperplane. Hence, bounding boxes are created around these areas that can have a potential object. The precision of the bounding boxes drawn can be increased by predicting offset values. These offset values assist in manipulating the bounding boxes of region proposals.

V. STUDY OF FASTER RCNN

Faster RCNN uses Fast RCNN and region proposal network as base.

A. Fast RCNN

In RCNN, we have to send all the region of interest to the CNN which will result in excessive computational time. So, in Fast RCNN we overcome this problem by sending whole image through the CNN and get output as convolutional feature map. This feature maps will be used to generate region of interest. Selective search algorithm will be used for this purpose. Size of the bounding box will be scaled with the size of feature map. Fixed size small regions will be created, using ROI pooling as input to fully connected layers as they accept only fixed size inputs. Class of the proposed regions will be determined using softmax layer. Softmax classifier gives us the probabilities for each class label. Bounding box regressors will be used to calculate offset values for bounding box. In faster RCNN we use different network for finding regions instead of selective search algorithm.

B. Region proposal network

After we have found our feature map, instead of using selective search, we will use different network for region proposal. Here, we slide a window over convolutional feature. From each window, we generate k anchor boxes. Anchor boxes can be of different scales and aspect ratio (width of image / height of image). Anchor boxes are the boxes which may contain objects. Here, at this time, we are not concerned about the class of the object, but only the presence of object inside anchor box. From this anchor boxes we try to find the correct coordinate for bounding boxes. Hence, RPN has a classifier and regression layer. Classifier determines the probability(scores) of proposal having the target object and regression regresses the coordinates of proposal. After RPN, we get proposed reasons of different sizes. So to match this with CNN feature map, ROI pooling is done. After that, we feed them to fully connected layers.
VI. STUDY OF MASK RCNN

Mask RCNN is a new type of neural network architecture which is an extension of Faster RCNN which adds a branch of segmentation mask to the latter one. It works towards the problem of instance segmentation which in turn consists of two sub problems viz object detection and semantic segmentation. Mask RCNN consists of two major stages:-

A. Region Proposal Network

Region Proposal Networks are used for the purpose of object detection. The main goal is to draw bounding boxes around the regions which contain an object. Another important task that takes place is ROI Align. ROI Align is the betterment of ROI pooling. Sometimes, data is lost when ROI pooling is used. The accuracy of the model is improved because of ROI Align.

B. Fully Convolutional Network

Fully Convolutional Networks are used for the purpose of semantic segmentation. The network is trained such that it perfectly maps each pixel of the input image to the particular class it belongs to. One thing about FCN is that it does not mark individual instances of a class separately. For example, if an image contains four cats then, the network would mark the region which contains the cats rather than separately specifying the four cats.

Hence, the Mask RCNN model performs two tasks:-

1. Performing object detection in order to obtain instances of objects.
2. Performing semantic segmentation in order to obtain area in which the object is present.
CNN is the base for object detection and image classification. CNN now outperforms humans on ImageNet challenge. RCNN was the first application of CNN for object detection. A typical CNN can tell us about the class of the object in an image. But in RCNN, CNN was forced to focus on a particular region as only a single object will dominate in that region. Thus after detecting regions by selective search algorithm, this resized regions was fed to CNN for further process. Fast RCNN was invented to speed up and simplify RCNN. In CNN, we were having three different models for image feature extraction, classification and regression. But in Fast RCNN we used a single network to compute all three. Fast RCNN also used selective search algorithm for region detection which was further improved by Faster RCNN. It used different network for region proposal. Using same feature map generated by CNN, it was able to generate region proposal. So for this, only one CNN was needed to be trained. Mask R CNN further extended Faster RCNN for pixel level segmentation. It added a branch to Faster R CNN which gave binary mask (says whether or not a given pixel is a part of an object) as output. In slum identification, to locate and identify slum area in an image, we need to go for pixel level segmentation. Thus Mask R CNN is the best way to identify the slums.

VIII. REFERENCES

[1] M. Kuffer, K. Pfeffer, and R. Sluzas, “Slums from space-15 years of slum mapping using remote sensing,” Remote Sens., vol. 8, no. 6, 2016.
[2] Kaiming He, Georgia Gkioxari, Piotr Dollar, Ross Girshick, “Mask RCNN”.
[3] Risbud, N. Policies for tenure security in Delhi. In Holding Their Ground-Secure Land Tenure for the Urban Poor in Developing Countries; Durand-Lasserve, A., Royston, R., Eds.; Earthscan: London, UK, 2002; pp. 59–74.
[4] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun, “Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks”.
[5] Bowen Cheng, Yunchao Wei, Honghui Shi, Rogerio Feris, Jinjun Xiong, and Thomas Huang1 “Revisiting RCNN: On Awakening the Classification Power of Faster RCNN”.
[6] Zhaozhuo Xu, Zhihan Jiang and Yicheng Li Department of Electrical Engineering, Stanford University