Image Scene Understanding - Object Detection in Aerial Images using Convolutional Neural Networks

Nisha Patil[1] Niketan Bothe[2] Shivam Gulve[3]
[1]Assistant Professor [2,3]UG Scholar
[1,2]Sandip University, Nashik, Maharashtra, India [3]Mukesh Patel School of Technology Management and Engineering, NMIMS, Mumbai, India

Abstract— This paper reviews and analyses the approaches concerning information updating from colored aerial photographs with the aim to transmit a detailed database of symbolic information about the objects detected in the aerial images fed to the system. Detecting objects in aerial images is obstructed/challenged by multiple problems such as variance of objects, undetermined obstructions and cluttered background. This paper, analyses the use of Convolutional Neural Networks and its features from multiple layers to perform robust aerial object detection. An image classification-based approach is used to localize the region-of-interest (ROIs) of multiple aspect ratios and further classify them into positive or negative SVM classifiers.

Keywords: Object detection in Aerial Images, Convolutional Neural Networks, Region-of-Interest pool, SVM

I. INTRODUCTION

In the past few years, there has been a sudden surge of the use of unmanned aerial vehicles (UAVs) for management of civil and public infrastructure assets. A few common instances include – a routine bridge inspection, power and cable line inspection, cross-border surveillances, traffic surveying and so on. As the application of UAV devices becomes widespread, higher levels of autonomy and increased independent- decision-making abilities are of importance for increasing the efficiency, accuracy and reliability of the devices. With improvements in remote-image-sensing technology, the spatial and overall resolution of remote sensing images has been continuously improved. Using image blocks or scenes as the basic unit for conducting image segmentation can make effective use of spatial contextual information to remove the ambiguity of interpretation.

A considerable number of methods have been proposed for object detection in aerial images, but the orientation robustness and cluttered background is an unsolved issue. In aerial images, objects in varying orientation and with multiple obstructions have large appearance variation, which challenges the existing object detection methods and approaches. In addition, the aspect ratio is an additional problem as it varies with varying orientations and multiple tiny obstructions in the image which introduces difficulty in localizing the object block.

Also, methods for image classification and segmentation make use of statistical models that are derived from labeled training datasets. In the usual setting, a learning algorithm passively accepts randomly selected images from the training set. Therefore, providing labeled datasets is costly in terms of human time and efforts. Furthermore, providing small training datasets can eventually lead to a poor performance in terms of image classification.

This paper aims to review the efforts behind image scene understanding in aerial images and image classification by using Deep Convolutional Neural Networks. Multiple problems have been tackled by various authors such as:

1) Firstly, we need to extract the present/active objects from the aerial images provided in the dataset
2) Secondly, we need to take care of the cluttered background, undetermined obstructions and orientation variation problem
3) Thirdly, assign every present object a label
4) Send symbolic information and train the model to assign vectors to each type/class of image in the data.

We see, that the above method, though produces highly accurate results, is very expensive in terms of human efforts and time. Hence, in this paper we will also suggest an alternative to this method which aims to drastically reduce the human efforts and time while maintaining the same level of accuracy in multi-class image classification and object detection.

II. APPROACH

A. Convolutional Neural Network

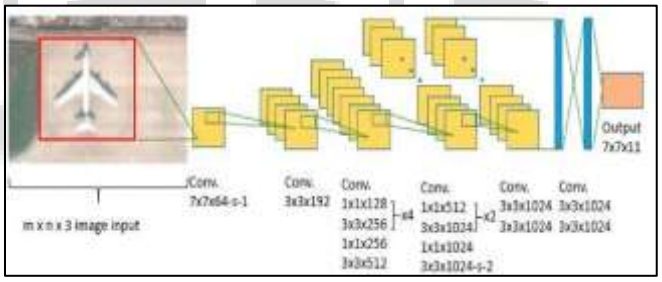

Fig. 1: Graphical representation of multilayered convolutional neural network (CNN)

Fig [1] above shows the approach followed to train the machine model to recognize airplanes by feeding it into a multi-layered convolutional neural network layer. The figure is input with three parameters [width, height, channel number] which is the predefined format according for the Conv2D () function of the Keras library used to train the machine model.

The convolution neural network is different from the usual neural network owing to the fact that it has multiple convolution layers. The CNN is trained and the correct parameters to be passed are determined. The batch size, iteration number, detection threshold and decay are all task specific parameters or arguments that have to be passed by the user into the platform. The total number of epochs that the network needs to be trained upon is to be determined empirically. Here, 'epoch' refers to a single presentation of the entire data on the CNN. For batch-wise training, each and every training sample must be passed learning algorithm at once in a single epoch before their overall weight is updated for the next epoch. A few index terms here are:

1) Batch size – total number of training samples in a single backward/forward pass
2) Learning rate – user-defined constant to control the learning rate of the neuron
3) Decay – The ratio between learning rate and epoch

B. Multi-Class Classification for Image Scene Understanding

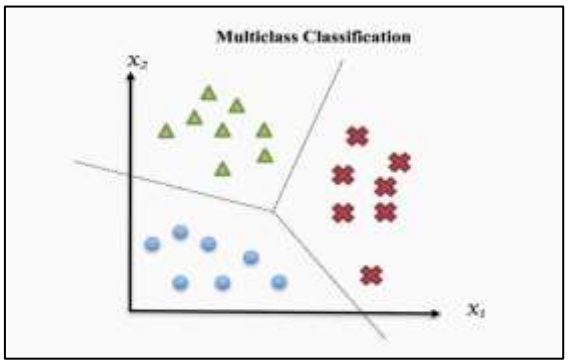

Fig. 2: Multiclass Image Classification Graphical Representation

Another important part of this review paper is the multiclass image classification. Considering there are two types of image classification possible – binary classification and non-binary/multi-class classification. Binary consider a single classifier in contention at a time i.e. the class in consideration is positive and all the rest are altogether classified as negative. Binary classifiers generate classes than may overlap

i.e. instances can be both A and B and C or neither. However, in case of multiclass classification every class has its own classifier that separates it from the others i.e. no grouping takes place. Multiclass classifiers generate mutually exclusive classes i.e. instance is either A, B or C.

Support Vector Machines (SVMs) are binary classifiers which require full-labelling of the data and is directly applied to the two classes available in the dataset but real-life application like object detection in aerial images require multi-class classification, which causes a problem. Hence, Multiclass Support Vector Machines are used as it addresses this very issue. It forms multiples of two classifiers based on the feature vector derived from the class of data.

Multiclass SVMs are basically made of learning modules and classification modules of which the classification model is applied to new data. It can be implemented by converting every single class vector machine into multiples of the binary segmentations which can be done by distinguishing the classifiers on the basis of the particular label vs the rest or every pair of classes.

III. LITERATURE SURVEY

A detailed and summarized review of the various papers that were referred to are as follow:

In [1], the authors have proposed an object detection algorithm which comprises of two parts – a rotation invariant Deep CNN feature extraction procedure and an object detection pipeline. The Deep CNN is deployed via Alex Net Architecture which consist of POOL5, FC6 & FC7 layer features each of which has a dimension of 9216, 4096 and 4096. It consists of five convolutional layers, of which, three are pooling layers and the rest two are fully-connected layers that are packed together at the end. The feature selection procedure takes place by a specific disentangling learning algorithm which shows that it is proper to use separate groups of features to model several distinct factors.

The object detection process is the next step that follows the rotation invariant Deep CNN features that were obtained in the previous step. To reduce the number of windows to be accessed, a graph-cut image segmentation method is used which produces color consistent areas. Then, the most similar groups are grouped together and this method is called selective search.

The next step that follows is called Fine Classification – this is where the Region-of- Interest (ROIs) generated in the Coarse Utilization step is used to train SVM classifiers. Once all the candidate windows have been classified, Non-Max Suppress (NMS) is applied.

When the said approach was applied, multiple combinations of the convolutional layers have been used and the probability of it giving the closest predictions were recorded. The plane recall was set at 0.9 and vehicle recall was set at 0.8. It was observed that the DCNN significantly outperforms the ACF (Aggregate Channel Feature). For a single channel feature, it was observed that the FC7(fully-connected layer 7) only significantly outperforms both POOL5 and FC6 with a precision of 0.861 when the vehicle recall has been set to 0.8. Whereas for plane recall, the POOL5 outperforms the FC6&7, y producing a precision of 0.891 when the plane recall has been set to 0.9.

In Vehicle detection the precision reached a maximum of 0.945 when a POOL5+FC6 was used. Whereas, in Plane detection, the max precision reached was 0.972 when a POOL5+FC7 was used.

In [2], the authors have used the YOLO approach of 'You Only Look Once' to train and test the machine models. The YOLO approach has many advantages over the traditionally used CNN software. Most of the CNNs use regional proposal approach to detect and suggest potential bounding boxes in an image. This is followed by multiclass classification and refinement while simultaneously eliminating duplicates.

The authors have approached the object detection problem as a tensor-regression problem. A bounding box is assigned to every probable object in the test image. The bounding box contains information related to the x and y coordinates of the image, its width(b) and height(h) and the probability of it being the required object to be detected.

Non-maximal suppression (NMS) is used to remove duplicates. To implement it, the authors have used a loss function which is expressed as:

$$\lambda_{coor} \sum_{i=0}^{S^2} \sum_{j=0}^{B} 1_{ij}^{obj} \left(\sqrt{w_i} - \sqrt{\hat{w}_i} \right)^2 + \left(\sqrt{h_i} - \sqrt{\hat{h}_i} \right)^2$$

The authors propose a detection network that has only 24 convolutional and two fully- connected fc6 & fc7 layers. The 26 layers application as shown in Figure 1. According to the authors, alternating and shifting (1 X 1) convolutions reduce the feature space. The final and fully connected layer makes object classification depending whether it is probable to belong to the bounding box or not.

In [3], the authors have tried four different CNN models to find out which model suits the best for object detection. The four models used are – AlexNet , VGGNet – 19, GoogleNet, ResNet50. Each of the model forms different layer of the neural network. Namely, AlexNet forms layer 23, VGGNet- 19 is layer 45, Google Net is layer 142 and ResNet is layer 175.

The features were extracted from multiple combination layers and full connection layer. It was found after testing with datasets and pretrained models that the ResNet performs much better than when compared with other models.

[4] Human species have an ability to recognize and acquire strange patterns. This principle applies to unseen images as well. A neural network can be programmed to show similar ability. This can be implemented using A Twin Network setup, this network accepts distinct inputs and is tied using an energy function.

It ensures the consistency of its predictions. Weight tying guarantees that two extremely similar images could not possibly be mapped by their respective networks to very different locations in feature space because each network computes the same function.

The network is symmetric: if we present two distinct images to the twin networks, the top conjoining layer will compute the same metric as if we were to we present the same two images but to the opposite twins.

[5] Facial expressions play an important role in recognition of emotions and are used in the process of non-verbal communication, as well as to identify people. They are very important in daily emotional communication, just next to the tone of voice. They are also an indicator of feelings, allowing a man to express an emotional state. People, can immediately recognize an emotional state of a person. As a consequence, information on the facial expressions are often used in automatic systems of motion recognition.

- Each subject participated in two sessions.
- Each session consisted of three trials in which a participant mimicked, in sequence, all seven examined emotional states.
- As a result, forty-two (2×3×7=42) 5- second sessions were registered for each user.
- The entire database contained a total of 252 (6×42) facial expressions.

[6] Image classification is the process of assigning land cover classes to pixels. For example, these 9 global land cover data sets classify images into forest, urban, agriculture and other classes.

There are three main technique in image classification: -

- Unsupervised classification
- Supervised classification

1) Unsupervised Classification: -

In unsupervised classification it creates groups of "pixels" into "cluster" which are totally based on their properties. In order to create cluster, the unsupervised uses KMEANS and ISODATA.

Unsupervised Classification steps: -

- Generate Cluster
- Assign Classes

2) Supervised Classification: -

In supervised classification we select representative sample for each land cover class. Then the software uses the 'training sites' and applies it to an entire image.

Supervised Classification steps: -

- Select training areas
- Generate signature file
- Classify

In [7], authors have used 3 classification models Decision Tree (DT) and Convolutional Neural Network (CNN).

A decision tree is a decision support tool that uses a tree-like graph or model of decisions and their possible consequences, including chance event outcomes, resource costs, and utility. It is one way to display an algorithm that only contains conditional control statements. A CNN & DT have nothing in common. Decision Trees cannot be trained using Gradient Descent process and cannot represent linear relations between features

A. Inference

The main observation made from the reviewed papers are:

- Each paper analyses a different feature of the CNN
- Different models have different efficiency
- Labelled training sets are preferred over unlabeled training sets but they have their own cons.
- The highest training accuracy was obtained for the ResNet50 Model.
- Different POOL and FC layer combination give different level of accuracy for different classes.

B. Suggestion for Improvements

Since most of the methods mentioned above make use of the traditional Convolutional Neural Network, the training dataset size is considerably large. The part where the convolutional neural networks have a disadvantage is this. Unlike human, the traditional neural networks can't learn from a single image. It is to be trained with thousands of images so that it can understand and recognize a similar pattern and then, when tested, try to classify the testing image into one of such classes. This method however fails when the testing data is less

i.e. when there is insufficient images training set. In such a case, it gives a poor accuracy in terms of object recognition.

ACKNOWLEDGEMENT

The authors would, with utmost respect, like to thank Prof. Nisha Patil for providing support and help with the research methodology and paper writing technique.

REFERENCES

[1] Haigang Zhu, Xiaogang Chen, Weiqun Dai, Kun Fu, Qixiang Fe, Jainbin Jaio, *Orientation Robust Object Detect in Aerial Images Using Deep CNN*
[2] Matija Radovic, Offei Adarkwa and Qiaosong Wang, * Object Recognition in Aerial Images using CNN*
[3] Mohammed Abbas Kadhim and Mohammed Hamzah Abed, * Convolution Neural Network for Satellite Image Classification *

[4] Gregory Koch, Richard Zemel, Ruslan Salakhutdinov Department of Computer Science, University of Toronto. Toronto, Ontario, Canada. * Siamese Neural Networks for One-shot Image Recognition
[5] Tarik A Rashid * Convolutional Neural Networks based Method for Improving Facial Expression Recognition

