Object Detection using Supervised Learning: Survey & Discussions

Prof. Neeraj Sharma[1], Prof. Priyanka Sharma[2], Prof. Saurabh Mandloi[3], Prof. Kretika Tiwari[4]

[1]Assistant Professor, Department of CSE, CIST, Bhopal (India)

[2]Assistant Professor, Department of CSE, CIRT, Bhopal (India)

[3]Assistant Professor, Department of CSE, PGOI, Bhopal (India)

[4]Assistant Professor, Department of CSE, BSSS, Bhopal (India)

ABSTRACT

For decades, object recognition and detection have been important problems in real-life applications. There are many applications for these utilities, including human detection, intrusion detection, motion detection, face detection etc. The detection and recognition of an object or pedestrian present growing and challenging problems in the field of computer vision. In this paper presents the literature survey for the object detection and the classification of any object or an image using the machine learning approach.

Keywords: Object detection, Red Green Blue, Supervised learning, Convolution neural network Feature extraction.

INTRODUCTION

Many computer vision tasks can be viewed in the context of detection and grouping: detecting smaller visual units and grouping them into larger structures. For example, in multi-person pose estimation we detect body joints and group them into individual people; in instance segmentation we detect pixels belonging to a semantic class and group them into object instances; in multi-object tracking we detect objects across video frames and group them into tracks. In all of these cases, the output is a variable number of visual units and their assignment into a variable number of visual groups [9]. Such tasks are often approached with two-stage pipelines that perform detection first and grouping second. But such approaches may be suboptimal because detection and grouping are tightly coupled.

A safe and robust autonomous driving system relies on accurate perception of the environment. To be more specific, an autonomous vehicle needs to accurately detect cars, pedestrians, cyclists, road signs, and other objects in real time in order to make the right control decisions that ensure safety [3]. A CNN is an effective solution in classification and recognition problems for large datasets, such as ImageNet. In contrast with other learning algorithms, a CNN has characteristics such as a local receptive fields and shared weights. A receptive field exploits the sparse connectivity of neurons to only a local region of an adjacent layer. In shared weights, replicated units create a feature map using shared parameters and increase robustness, which is efficient in lane detection problems that include different road environments [6]. Unmanned aerial vehicles (UAV) are now increasingly used as a cost effective and timely method of capturing remote sensing (RS) images. The advantages of UAV technology include low cost, small size, safety, ecological operation, and most of all, the fast and on-demand acquisition of images. The advance of UAV technology has reached the stage of being able to provide extremely high resolution remote sensing images encompassing abundant spatial and contextual information. This has enabled studies proposing many novel applications for UAV image analysis, including vegetation monitoring, urban site analysis, disaster management, oil and gas pipeline monitoring, detection and mapping of archaeological sites, and object detection [8].

A CNN can be viewed as a simplified version of the Neocognitron model [2], which was proposed to simulate the human visual system in 1980. CNNs initially appeared in the early 1990s, but they did not enjoy much popularity at the time due to limited computational resources. However, with the advent of powerful graphics processing unit (GPU) computing and abundance of labeled training data, CNNs have once again emerged as a powerful feature extraction and classification tool, yielding record-breaking results in major computer vision challenges.

The success of CNNs in computer vision has widely inspired investigators in the medical

imaging community, resulting in a number of publications in a short period of time, which collectively demonstrates the effectiveness of CNNs for a variety of medical imaging tasks. CNNs are so-named due to the convolutional layers in their architectures. Convolutional layers are responsible for detecting certain local features in all locations of their input images. To detect local structures, each node in a convolutional layer is connected to only a small subset of spatially connected neurons in the input image channels. To enable the search for the same local feature all over the input channels, the connection weights are shared between the nodes in the convolutional layers. Each set of shared weights is called a kernel or a convolution kernel. Thus, a convolutional layer with n kernels learns to detect n local features whose strength across the input images is visible in the resulting n feature maps. To reduce computational complexity and achieve a hierarchical set of image features, each sequence of convolution layers is followed by a pooling layer.

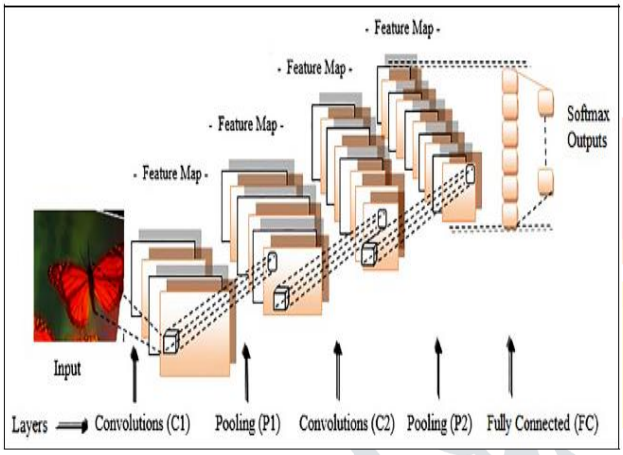

Fig 1: A Convolution Neural Network architecture [11].

Fully-convolutional networks (FCN) were popularized by Long et al., who applied them to the semantic segmentation domain. FCN defines a broad class of CNNs, where the output of the final parameterized layer is a grid rather than a vector. This is useful in semantic segmentation, where each location in the grid corresponds to the predicted class of a pixel. FCN models have been applied in other areas as well. To address the image classification problem, a CNN needs to output a 1-dimensional vector of class probabilities. One common approach is to have one or more fully connected layers, which by definition output a 1D vector – 1×1×Channels [3].

In recent years, deep learning methods have emerged as powerful machine learning methods for object recognition and detection. Deep learning methods are different from traditional approaches in that they automatically and quickly learn the features directly from the raw pixels of the input images without using approaches such as SIFT, HOG, and SURF. In deep learning methods, local receptive fields grow in a layer-by-layer manner. The low-level layers extract fine features, such as lines, borders, and corners, while high-level layers exhibit higher features, such as object portions, like pedestrian parts, or the whole object, like cars and traffic signs [10].

While recent research has been primarily focused on improving accuracy, for actual deployment in an autonomous vehicle, there are other issues of image object detection that are equally critical. For autonomous driving some basic requirements for image object detectors include the following: a) Accuracy. More specifically, the detector ideally should achieve 100% recall with high precision on objects of interest. b) Speed. The detector should have real-time or faster inference speed to reduce the latency of the vehicle control loop. c) Small model size [3].

The rest of this paper is organized as follows in the first section we describe an introduction of about the object detection and machine learning approach. In section II we discuss about the object detection, In section III we discuss about the weakly supervised learning, In section IV we presents the literature review for the object detection in number of applications. Finally in section V we conclude and discuss the future scope.

II OBJECT DETECTION

Object detection is the process of detecting instances of semantic objects of a certain class (such as humans, airplanes, or birds) in digital images and video. A common approach for object detection frameworks includes the creation of a large set of candidate windows that are in the sequel classified using CNN features. For example, the method described employs selective search to derive object proposals, extracts CNN features for each proposal, and then feeds the features to an SVM classifier to decide whether the windows include the object or not. A large number of works is based on the concept of
Regions with CNN features. Approaches following the Regions with CNN paradigm usually have good detection accuracies; however, there is a significant number of methods trying to further improve the performance of Regions with CNN approaches, some of which succeed in finding approximate object positions but often cannot precisely determine the exact position of the object [15]. The goal of human pose estimation is to determine the position of human joints from images, image sequences, depth images, or skeleton data as provided by motion capturing hardware. Face recognition is one of the hottest computer vision applications with great commercial interest as well.

III WEAKLY SUPERVISED LEARNING

Deep neural networks give rise to many breakthroughs in computer vision by using huge amounts of labeled training data. Supervised object detection and semantic segmentation require object or even pixel level annotations, which are much more labor-intensive to obtain than image level labels. On the other hand, when there exist image level labels only, due to incomplete annotations, it is very challenging to predict accurate object locations, pixel-wise labels, or even image level labels in multi-label image classification [10]. Given image level supervision only, researchers have proposed many weakly supervised algorithms for detecting objects and labeling pixels. These algorithms employ different mechanisms, including bottom-up, top-down and hybrid approaches , to dig out useful information. In bottom-up algorithms, pixels are usually grouped into many object proposals, which are further classified, and the classification results are merged to match ground truth image labels. In top-down algorithms, images first go through a forward pass of a deep neural network, and the result is then propagated backward to discover which pixels actually contribute to the final result. There are also hybrid algorithms that consider both bottom-up and top-down cues in their pipeline. Although there exist many weakly supervised algorithms, the accuracy achieved by top weakly supervised algorithms is still significantly lower than their fully supervised counterparts. This is reflected in both the precision and recall of their results. In terms of precision, results from weakly supervised algorithms contain much more noise and outliers due to indirect and incomplete supervision. Likewise, such algorithms also achieve much lower recall because there is insufficient labeled information for them to learn comprehensive feature representations of target object categories. However, different types of weakly supervised algorithms may return different but complementary subsets of the ground truth.

IV RELATED WORK

[1] The paper surveys some recent progress in deep learning based fine-grained image classification and semantic segmentation. They can be directly adapted to find-grained image classification. Since the subtle differences of visually similar fine-grained objects usually exist in some common parts, many approaches resort to deep learning technology to boost the performance of part localization, while some approaches integrate the part localization into the deep learning framework and can be trained end-to-end. In this paper they review four types of deep learning based fine rained image classification approaches, including the general convolutional neural networks (CNNs), part detection based, ensemble of networks based and visual attention based fine-grained image classification approaches. Besides, the deep learning based semantic segmentation approaches are also covered in this paper. The region proposal based and fully convolutional networks based approaches for semantic segmentation are introduced respectively. [2] In this study they have compared these two successful learning machines both experimentally and theoretically. For that purpose, they considered two well-studied topics in the field of medical image analysis: detection of lung nodules and distinction between benign and malignant lung nodules in computed tomography (CT). For a thorough analysis, they used 2 optimized MTANN architectures and 4 distinct CNN architectures that have different depths. Their experiments demonstrated that the performance of MTANNs was substantially higher than that of CNN when using only limited training data. With a larger training dataset, the performance gap became less evident even though the margin was still significant. Specifically, for nodule detection, MTANNs generated 2.7 false positives per patient at 100% sensitivity, which was significantly ($p < 0.05$) lower than the best performing CNN model with 22.7 false positives per patient at the same level of sensitivity. [3] In this work, they propose SqueezeDet, a fully convolutional neural network for object detection that aims to simultaneously satisfy all of the above constraints. In their network they use convolutional layers not only to extract feature maps, but also as the output layer to compute bounding boxes and class probabilities. The detection pipeline of our model only contains a single forward pass of a neural network, thus it is extremely fast. Their model is fully convolutional, which leads to small model size and better energy efficiency. [4] In this paper, they extend adversarial examples to semantic segmentation and object detection which are much more difficult. Our observation is that both segmentation and detection are based on classifying multiple targets on an image (e.g., the target is a pixel or a receptive field in segmentation, and an object proposal in detection). This inspires us to optimize a loss function over a set of targets for generating adversarial perturbations. Based on this, they propose a novel algorithm named Dense Adversary Generation (DAG), which applies to the state-of-the-art networks for segmentation and detection. [5] They propose a real-time RGB-based pipeline for object detection and 6D pose estimation. Our novel 3D orientation estimation is based on a variant of the Denoising Auto encoder that is trained on simulated views of a 3D model using Domain Randomization. This so-called Augmented Auto encoder has several advantages over existing methods: It does not require real, pose-annotated training data, generalizes to various test sensors and inherently handles object and view symmetries. Instead of learning an

explicit mapping from input images to object poses, it provides an implicit representation of object orientations defined by samples in a latent space. [6] In this paper they use a CNN for image enhancement and the detection of driving lanes on motorways. In general, the process of lane detection consists of edge extraction and line detection. A CNN can be used to enhance the input images before lane detection by excluding noise and obstacles that are irrelevant to the edge detection result. However, training conventional CNNs requires considerable computation and a big dataset. Therefore, they suggest a new learning algorithm for CNNs using an extreme learning machine (ELM). The ELM is a fast learning method used to calculate network weights between output and hidden layers in a single iteration and thus, can dramatically reduce learning time while producing accurate results with minimal training data. A conventional ELM can be applied to networks with a single hidden layer; as such, they propose a stacked ELM architecture in the CNN framework. [7] They apply for the first time an object detection model previously used on natural images to identify cells and recognize their stages in bright field microscopy images of malaria-infected blood. Many micro-organisms like malaria parasites are still studied by expert manual inspection and hand counting. This type of object detection task is challenging due to factors like variations in cell shape, density, and color, and uncertainty of some cell classes. In addition, annotated data useful for training is scarce, and the class distribution is inherently highly imbalanced due to the dominance of uninfected red blood cells. [8] Their proposed method begins by segmenting the input image into small homogeneous regions, which can be used as candidate locations for car detection. Next, a window is extracted around each region, and deep learning is used to mine highly descriptive features from these windows. They use a deep convolutional neural network (CNN) system that is already pre-trained on huge auxiliary data as a feature extraction tool, combined with a linear support vector machine (SVM) classifier to classify regions into "car" and "no-car" classes. The final step is devoted to a fine-tuning procedure which performs morphological dilation to smooth the detected regions and fill any holes. [10] In this paper, they propose a novel weakly supervised curriculum learning pipeline for multi-label object recognition, detection and semantic segmentation. In this pipeline, we first obtain intermediate object localization and pixel labeling results for the training images, and then use such results to train task-specific deep networks in a fully supervised manner. The entire process consists of four stages, including object localization in the training images, filtering and fusing object instances, pixel labeling for the training images, and task-specific network training. To obtain clean object instances in the training images, they propose a novel algorithm for filtering, fusing and classifying object instances collected from multiple solution mechanisms. [13] They present a method to address this issue, and learn object detectors incrementally, when neither the original training data nor annotations for the original classes in the new training set are available. The core of our proposed solution is a loss function to balance the interplay between predictions on the new classes and a new distillation loss which minimizes the discrepancy between responses for old classes from the original and the updated networks. This incremental learning can be performed multiple times, for a new set of classes in each step, with a moderate drop in performance compared to the baseline network trained on the ensemble of data.

V CONCLUSION AND FUTURE SCOPE

Over the last years supervised learning and deep learning methods have been shown to outperform previous state-of-the-art machine learning techniques in several fields, with computer vision being one of the most prominent cases. Accurate detection of objects and image classification in is a central problem in many applications, such as autonomous navigation, housekeeping robots, and augmented/virtual reality. learning allows computational models of multiple processing layers to learn and represent data with multiple levels of abstraction mimicking how the brain perceives and understands multimodal information, thus implicitly capturing intricate structures of large-scale data. In this paper focus on the literature survey for the various machine learning techniques, future work includes the implementation model to the extend existing model whose suffer from the accuracy and precision.

REFERENCES:-

[1] Bo Zhao, Jiashi Feng, Xiao Wu, Shuicheng Yan, "A Survey on Deep Learning-based Fine-grained Object Classification and Semantic Segmentation", International Journal of Automation and Computing, April 2017, pp 119-135.

[2] Nima Tajbakhsh, Kenji Suzuki, "Comparing two classes of end-to-end machine-learning models in lung nodule detection and classification: MTANNs vs. CNNs", Elsevier ltd. Pattern Recognition, pp 476–486.

[3] Bichen Wu, Forrest Iandola, Peter H. Jin, Kurt Keutzer, "SqueezeDet: Unified, Small, Low Power Fully Convolutional Neural Networks for Real-Time Object Detection for Autonomous Driving", IEEE 2017, pp 129-137.

[4] Cihang Xie, Jianyu Wang, Zhishuai Zhang, Yuyin Zhou, Lingxi Xie, Alan Yuille,

"Adversarial Examples for Semantic Segmentation and Object Detection", IEEE 2017, pp 1369-1378.

[5] Martin Sundermeyer, Zoltan-Csaba Marton, Maximilian Durner, Manuel Brucker, Rudolph Triebel, "Implicit 3D Orientation Learning for 6D Object Detection from RGB Images", IEEE 2017, pp 1-17.

[6] Jihun Kim, Jonghong Kim, Gil-Jin Jang, Minho Lee, "Fast learning method for convolutional neural networks using extreme learning machine and its application to lane detection", Elsevier ltd. Neural Networks, 2017, pp 109–121.

[7] Jane Hung, Anne Carpenter, "Applying Faster R-CNN for Object Detection on Malaria Images", IEEE Explore 2017, pp 56-61.

[8] Nassim Ammour, Haikel Alhichri, Yakoub Bazi, Bilel Benjdira, Naif Alajlan, Mansour Zuair, "Deep Learning Approach for Car Detection in UAV Imagery", Remote sensing 2017, pp 1-15.

[9] Alejandro Newell, Zhiao Huang, Jia Deng, "Associative Embedding: End-to-End Learning for Joint Detection and Grouping", 31st Conference on Neural Information Processing Systems,2017, pp 1-11.

[10] Weifeng Ge, Sibei Yang, Yizhou Yu, "Multi-Evidence Filtering and Fusion for Multi-Label Classification, Object Detection and Semantic Segmentation Based on Weakly Supervised Learning", IEEE Explore 2017, pp 1277-1286.

[11] Aysegul Ucxar, Yakup Demir, Cuneyt Guzelis, "Object recognition and detection with deep learning for autonomous driving applications", Transactions of the Society for Modeling and Simulation International 2017, Vol. 93, PP 759–769.

[12] Thijs Kooi, Geert Litjens, Bram van Ginneken, Albert Gubern-Mérida, Clara I. Sánche , Ritse Mann, Ard den Heeten , Nico Karssemeijer, "Large scale deep learning for computer aided detection of mammographic lesions", Elsevier ltd. Medical Image Analysis, 2017, pp 303–312.

[13] Konstantin Shmelkov, Cordelia Schmid, Karteek Alahari, "Incremental Learning of Object Detectors without Catastrophic Forgetting", IEEE 2017, pp 3400-3409.

[14] Yin Zhou, Oncel Tuzel, "VoxelNet: End-to-End Learning for Point Cloud Based 3D Object Detection", IEEE 2017, pp 4490-4499.

[15] Athanasios Voulodimos, Nikolaos Doulamis, Anastasios Doulamis, Eftychios Protopapadakis, "Deep Learning for Computer Vision: A Brief Review", Hindawi Computational Intelligence and Neuroscience, 2018, pp 1-14.

[16] H. Su, S. Maji, E. Kalogerakis, and E. G. Learned-Miller, "Multi-view convolutional neural networks for 3d shape recognition", In Proc. ICCV, to appear, 2015.

[17] J. Sun, M. Ovsjanikov, and L. Guibas, "A concise and provably informative multi-scale signature based on heat diffusion", In Computer graphics forum, volume 28, pages 1383–1392. Wiley Online Library, 2009.

[18] O. Vinyals, S. Bengio, and M. Kudlur, "Order matters: Sequence to sequence for sets", arXiv preprint arXiv:1511.06391, 2015.

[19] D. Z. Wang and I. Posner, "Voting for voting in online point cloud object detection", Proceedings of the Robotics: Science and Systems, Rome, Italy, 1317, 2015.

[20] Z. Wu, R. Shou, Y. Wang, and X. Liu, " Interactive shape cosegmentation via label propagation", Computers Graphics, 248-254, 2014.

[21] Z. Wu, S. Song, A. Khosla, F. Yu, L. Zhang, X. Tang, and J. Xiao, "3d shapenets: A deep representation for volumetric shapes", In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 1912–1920, 2015.

[22] L. Yi, V. G. Kim, D. Ceylan, I.-C. Shen, M. Yan, H. Su, C. Lu, Q. Huang, A. Sheffer, and L. Guibas, "A scalable active framework for region annotation in 3d shape collections", SIGGRAPH Asia, 2016.