JOURNAL OF CRITICAL REVIEWS

ISSN- 2394-5125 VOL 6, ISSUE 04, 2019

Object Detection Model using YOLO-based Deep Learningin Images and Videos

Sukanth Behera[1], Saradiya Kishore Parida[1]

[1]Assistant Professor,Department of Electronics and Communication Engineering

[1]Gandhi Institute for Technology(GIFT), Bhubaneswar, India

ABSTRACT

Deep learning has gained a tremendous influence on how the world is adapting to Artificial Intelligence since past few years.The image classification is a classical problem of image processing, computer vision and machine learning fields.We present YOLO, a new approach to object detection. Prior work on object detection repurposes classifiers to perform detection. Instead, we frame object detection as a regression problem to spatially separated bounding boxes and associated class probabilities. A single neural network predicts bounding boxes and class probabilities directly from full images in one evaluation. Since the whole detection pipeline is a single network, it can be optimized end-to-end directly on detection performance. Our unified architecture is extremely fast.

Keywords: image classification, object detection, convolutional neural networks, deep learning, YOLO algorithm.

1. INTRODUCTION

Classification is a systematic arrangement in groups and categories based on its features. Image classification came into existence for decreasing the gap between the computer vision and human vision by training the computer with the data. The image classification is achieved by differentiating the image into the prescribed category based on the content of the vision. Motivation by [1], in this paper, we explore the study of image classification using deep learning. The conventional methods used for image classifying is part and piece of the field of artificial intelligence (AI) formally called as machine learning. The machine learning consists of feature extraction module that extracts the important features such as edges, textures etc and a classification module that classify based on the features extracted. The main limitation of machine learning is, while separating, it can only extract certain set of features on images and unable to extract differentiating features from the training set of data. This disadvantage is rectified by using the deep learning [2].

Deep learning (DL) is a sub field to the machine learning, capable of learning through its own method of computing. A deep learning model is introduced to persistently break down information with a homogeneous structure like how a human would make determinations. To accomplish this, deep learning utilizes a layered structure of several algorithms expressed as an artificial neural system (ANN). The architecture of an ANN is simulated with the help of the biological neural network of the human brain. This makes the deep learning most capable than the standard machine learning models [3, 4]. In deep learning, we consider the neural networks that identify the image based on its features. This is accomplished for the building of a complete feature extraction model which can solve the difficulties faced due to the conventional methods. The extractor of the integrated model should be able to learn extracting the differentiating features from the training set of images accurately. Many methods like GIST, histogram of gradient oriented and Local Binary Patterns, SIFT are used to classify the feature descriptors from the image.

Pierre et al. [5] bridged between the lower layer's output and the classifier to take the global shape and local details into account. This use of multi-stage features improved the accuracy over systems that use single stage features on several tasks, such as in pedestrian detection and certain sorts of classification. Motivated by many advantages of the multi-layers features, we propose an alternative multistage strategy that can be applied to a standard one track CNN whose weight parameter is fixed after the training has been finished without the multi-stage strategy in mind.

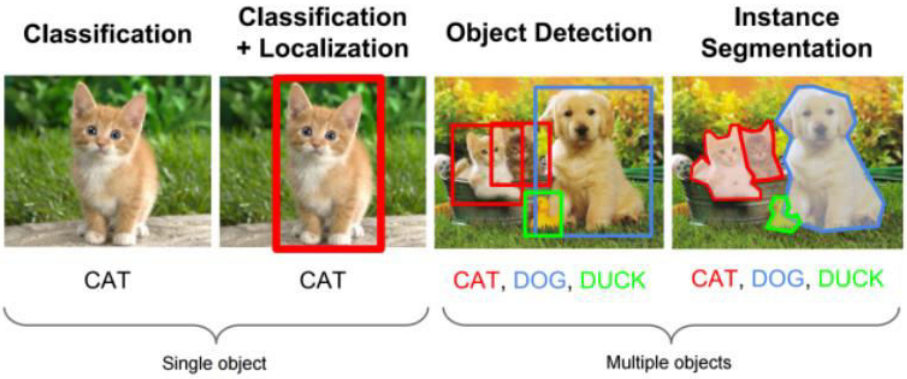

Fig. 1: Example of image classification using object detection.

2. MOTIVATION

Since Matthew et al. [6] invented a probe to look inside a feature map, if one carefully observes the visualized features at each layer, one can obtain intuition as to why multi-stage features could enable further improvement of image classification. When comparing the visualization of features and the corresponding image patches, the latter has the greater variation since CNN mainly focuses on a discriminant structure. For other discoveries in [6], lower layer features are usually simpler than those of higher layers. The meaning of this discovery is that simple images are well activated at lower layers and complex images have high activation value at higher layers. Also, the lower layer features are focused on a smaller area in an image and the higher layer features are focused on a larger area in an image. For these reasons, the deep neural network model that uses the last layer features only finds it hard to classify the dataset which contains both simple and complex objects. This forces us to bind features from multi-stages in an effective way.

2.1. Artificial Neural Networks

A neural network is a combination of hardware bonded or separated by the software system which operates on the small part in the human brain called as neuron. A multi layered neural network can be proposed as an alternative of the above case. The training image samples should be more than nine times the number of parameters essential for tuning the classical classification under particularly good resolution. The multi-layered neural network is so complicated task with respect to its architecture in the real-world implementations. The multi-layered neural network is at present expressed as the Deep Learning. In deep neural networks every node decides its basic inputs by itself and sends it to the next tier on behalf of the previous tier.We train the data in the networks by giving an input image and conveying the network about its output. Neural networks are expressed in terms of number of layers involved for producing the inputs and outputs and the depth of the neural network. Neural networks are involved in many principles like fuzzy logic, genetic algorithms, and Bayesian methods. These layers are generally referred to as hidden layers. They are expressed in terms of number of hidden nodes and number of inputs and outputs every node consists. The Convolutional Neural Network (ConvNet) is most popular algorithm used for implementing the deep learning technique. The ConvNet consists of Feature detection layers and classification. A ConvNet is composed of several layers, and they are convolutional layers, maxpooling or average-pooling layers, and fully connected layers.

2.2. Alexnet

The ConvNet is categorized into two types named LeNet and AlexNet. The LeNet is expressed as the Shallow Convolutional Neural Networks which is designed to classify the hand-written digits. The LeNet comprises of 2 convolutional layers, 2 subsampling layers, 2 hidden layers and 1 output layer [5]. The AlexNet is expressed as the deep convolutional neural networks which are used for classifying the input image to one of the thousand classes. AlexNet is used to solve many problems like indoor sense classification which is highly seen in artificial neural intelligence. It is a powerful method of knowing the features of the image with more differential vision in the computer field for the recognition of patterns. This paper discusses about the classification of a size of image of required choice. It can very effectively classify the training sample of images present in the AlexNet for better vision. The AlexNet comprises of 5 convolutional layers, 3 sub sampling layers and 3 fully connected layers. The main difference between the LeNet and AlexNet are the type of Feature Extractor. We use the non-linearity in the Feature Extractor module in AlexNet whereas Log sinusoid is used in LeNet. AlexNet uses dropout which is not observed in any other data sets of networking.

3. YOLO ALGORITHM

There are a few different algorithms for object detection, and they can be split into two groups:

3.1. Algorithms based on classification

They work in two stages. In the first step, we are selecting from the image interesting regions. Then we are classifying those regions using convolutional neural networks. This solution could be terribly slow because we must run prediction for every selected region. Most known example of this type of algorithms is Region-based convolutional neural network (RCNN), Fast-RCNN and Faster-RCNN.

3.2. Algorithms based on regression

Instead of selecting interesting parts of an image, we are predicting classes and bounding boxes for the whole image in one run of the algorithm. Most known example of this type of algorithms is YOLO (You only look once) commonly used for real-time object detection.Our task is to predict a class of an object and the bounding box specifying object location.

Each boundary box contains 5 elements: (x, y, w, h) and a box confidence score. The confidence score reflects how likely the box contains an object (objectness) and how accurate is the boundary box. We normalize the bounding box width w and height h by the image width and height. x and y are offsets to the corresponding cell. Hence, x, y, w, and h are all between 0 and 1. Each cell has 20 conditional class probabilities. The conditional class probability is the probability that the detected object belongs to a class (one probability per category for each cell). So, YOLO's prediction has a shape of (S, S, B×5 + C) = (7, 7, 2×5 + 20) = (7, 7, 30).

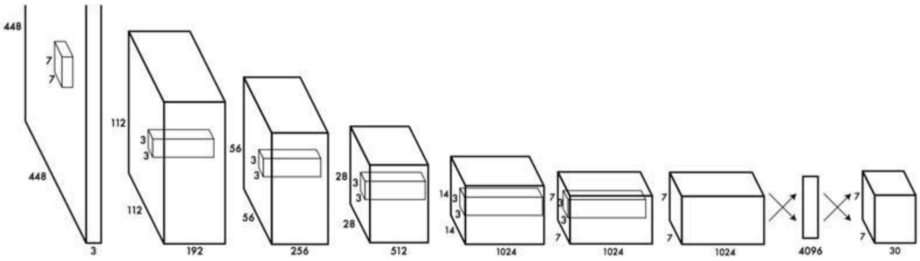

Fig. 2: CNN model.

The major concept of YOLO is to build a CNN network to predict a (7, 7, 30) tensor. It uses a CNN network to reduce the spatial dimension to 7×7 with 1024 output channels at each location. YOLO performs a linear regression using two fully connected layers to make 7×7×2 boundary box predictions (the middle picture below). To make a final prediction, we keep those with high box confidence scores (greater than 0.25) as our final predictions (the right picture).

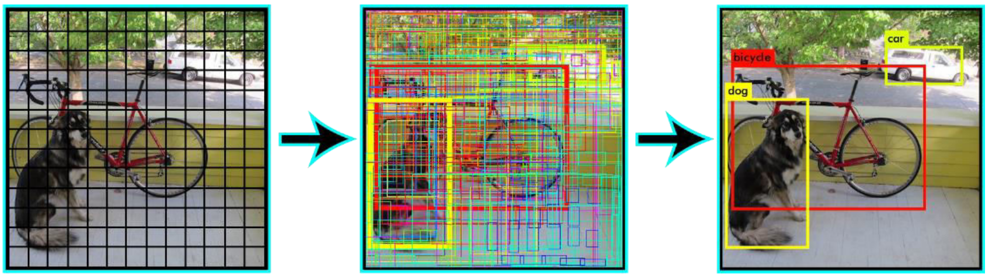

Fig. 3: Output of image classification using YOLO algorithm.

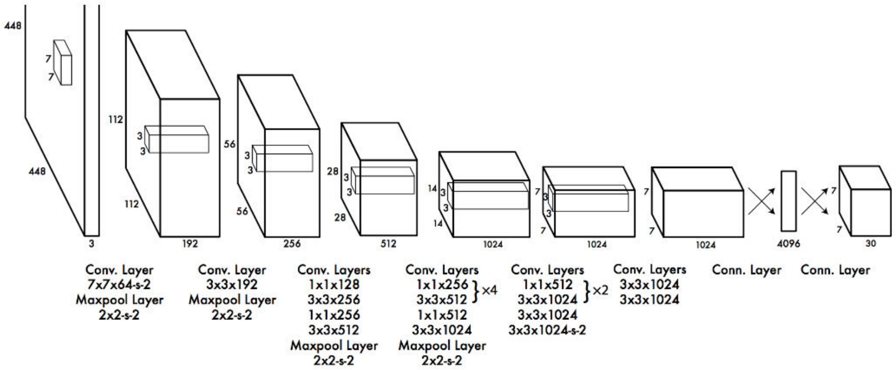

Fig. 4: YOLO network model.

YOLO has 24 convolutional layers followed by 2 fully connected layers (FC). Some convolution layers use 1 × 1 reduction layers alternatively to reduce the depth of the feature's maps. For the last convolution layer, it outputs a tensor with shape (7, 7, 1024). The tensor is then flattened. Using 2 fully connected layers as a form of linear regression, its outputs 7×7×30 parameters and then reshapes to (7, 7, 30), i.e. 2 boundary box predictions per location.

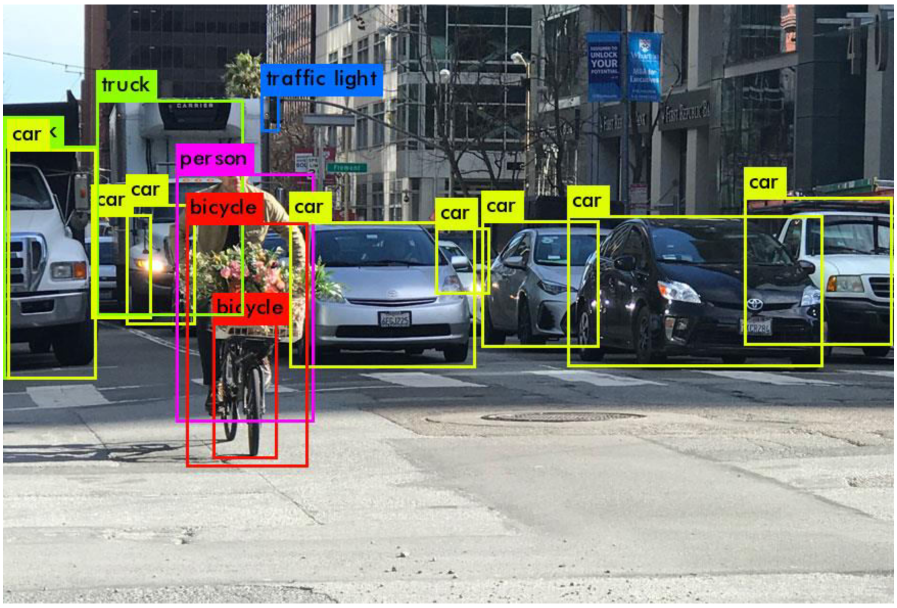

Fig. 5: Output of object detection and classification in real-time videos.

4. CONCLUSIONS

This article presented YOLO, a new approach to object detection, where the object detection is framed as a regression problem to spatially separated bounding boxes and associated class probabilities.Our unified architecture is extremely fast. This model showed excellent detection and tracking results on the object trained and can further utilized in specific scenarios to detect, track, and respond to the targeted objects in the video surveillance. This real time analysis of the ecosystem can yield great results by enabling security, order, and utility for any enterprise. Further extending the work to detect ammunition and guns to trigger alarm in case of terrorist attacks. The model can be deployed in CCTVs, drones, and other surveillance devices to detect attacks on many places like schools, government offices and hospitals where arms are completely restricted

REFERENCES

[1] https://in.mathworks.com/matlabcentral/fileexchange/59133neural-network-toolbox-tm--model-for-alexnet-network.

[2] H. Lee, R. Grosse, R. Ranganath, and A.Y. Ng. Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations. In Proceedings of the 26th Annual International Conference on Machine Learning, pages 609–616. ACM, 2009.

[3] Deep Learning with MATLAB – matlab expo2018.

[4] Introducing Deep Learning with the MATLAB – Deep Learning E-Book provided by the mathworks.

[5] Sermanet, P., Kavukcuoglu, K., Chintala, S., LeCun, Y.: Pedestrian detection with unsupervised multi-stage feature learning. In: 2013 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3626–3633. IEEE (2013)

[6] Zeiler, M.D., Fergus, R.: Visualizing and understanding convolutional neural networks. arXiv preprint arXiv:1311.2901 (2013)