ISSN: 2321-7782 (Online)
e-ISJN: A4372-3114
Impact Factor: 7.327

Volume 6, Issue 12, December 2018

International Journal of Advance Research in Computer Science and Management Studies

Research Article / Survey Paper / Case Study

Available online at: www.ijarcsms.com

Object Detection and Classification in Globally Inclusive Images Using Yolo

Swetha M S[1]
Assistant Professor, Department of IS&E,
BMS Institute of Technology & Management,
Yelahanka, Bangalore -560064,
Karnataka – India.

Vineetha M[2]
Student, Department of IS&E,
BMS Institute of Technology & Management,
Yelahanka, Bangalore -560064,
Karnataka – India.

Abstract: Object detection is the process of object recognition and classification. There are several Training sets available online for training an object detection model. But the models aren't trained to detect the same object from different geographical regions. This paper proposes a model that can detect globally inclusive images. The given model is created using YOLO Neural networks and it is initially trained using the Inclusive Images dataset. Furthermore, it is trained using a newly curated, more representative dataset. A machine learning model should be efficient and accurate even when it learns from imperfect data sources. This model shows an improvement in the accuracy when detecting geographically diverse images.

Keywords: CNN, R-CNN, YOLO, SSD, Neural Networks, Object Detection, Bounding boxes.

I. Introduction

In recent years, with the rapid development of Machine learning and Artificial Intelligence a number of research areas with respect to these concepts have increased immensely. Along with this there is a continuous improvement of convolution neural networks and with this Computer vision algorithms and architecture has developed immensely. Vision starts with the eyes but all the computation happens in the brain. Thus, the principal target of computer vision and AI is detecting the object and their relation, action and intention.

The biggest challenge, however, is that it needs training big data. In general, Artificial Intelligence technologies have have to learn thoroughly from the training set. That is, in order to recognize or detect an object from a video stream, image data should be used as training data. Furthermore, it requires very well-refined and well-represented dataset, for good training.

The dataset it's compliant to be COCO, Pascal VOC and Image Net and also recently Google Images V4. The ultimate goal is that, through experimentation to strengthen the understanding of the model, and through the analysis of the results, learn the importance of targeted and inclusive datasets for deep learning. In addition, to this optimize the model for efficient utilization when integrated with the necessary system or application. Therefore, this paper proposes a system that efficiently detects objects belonging to different geographical regions.

II. Related Work

Convolution Neural Networks [1] uses the concepts of deep learning and becomes the golden standard for image classification. But CNN can recognize an object in a single image. This implies that CNN can't detect different objects in a single image. Object detection not only includes recognizing whether specific object is present or not, but also determines the exact position of the target region. Furthermore, the objects should be separated by bounding boxes.

On the other hand, R-CNN uses selective search method to extract 2000 regions (region proposals) from the image. The problem with this algorithm is it cannot be implemented in real time operations as it is extremely slow. The R-CNN algorithm is typically selective search to get objects of interest in the scene. The Fast R-CNN [2] approach varies from the R-CNN as it applies this method of selective search on the entire image instead of selecting one region at a time. Faster R-CNN [3] method came up was object detection algorithm that eliminates the selective search algorithm to perform the operation. Single Shot MultiBox Detector detects objects by applying feature maps based on CNN. The feature points are formed in a grid on the image and the bounding box which centred on each of the feature grid cell determines the presence of an objects. SSD [4] is faster than Faster R-CNN, but its accuracy is lower because limited number of the anchors.

You Only Look Once (YOLO) [5] outperformed all the customary methodologies. YOLO engineering contains a single neural network which predicts bounding boxes and class probabilities particularly from full pictures in a single assessment. Since the whole detection pipeline is a single network and it will in general be improved and it can be used in real-time object detection. Just after this the second version of YOLO was released called YOLO9000 [6]. Although this trained the model for 9000 object categories it still was not sufficient to detect globally inclusive images.

Therefore in order to get good performance and accurate results (especially related to globally inclusive images) with YOLO, it should be trained with the necessary targeted datasets. This paper proposes training the YOLO model with the Google's Inclusive Image dataset and newly curated more representative dataset.

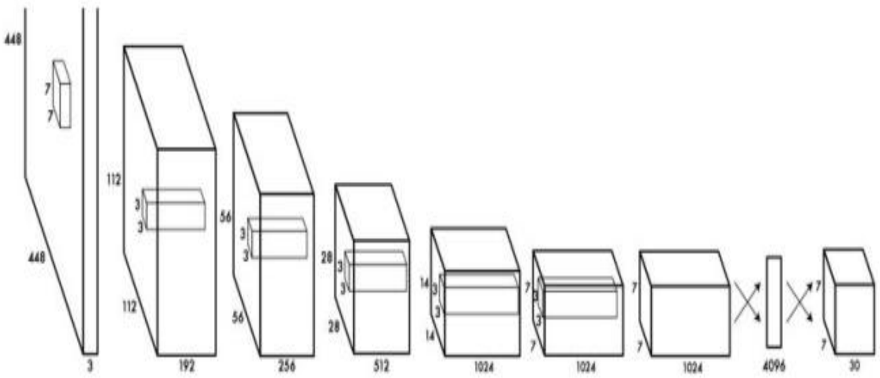

Fig 1 Architecture of YOLO

III. METHODOLOGY

The model is trained using the datasets Inclusive Images by Google and the newly curated inclusive images. The model itself is built using YOLO neural networks and it is trained using the Nvidia GTX 1080ti GPU.

Rudimentary Algorithm for Object detection:

- STEP 1: A model or algorithm is utilized to produce regions of interest or region recommendations. These region recommendations are a substantial set of bounding boxes traversing the full picture (that is, an object localization segment).
- STEP 2: Secondly, visual features are separated for every one of the bounding boxes, they are assessed and it is resolved whether and which objects are available in the proposition dependent on visual features (i.e. an object order segment).
- STEP 3: In the last post-processing step, covering boxes are joined into a single bounding
- Box and the image are detected and it looks as given in fig 2.

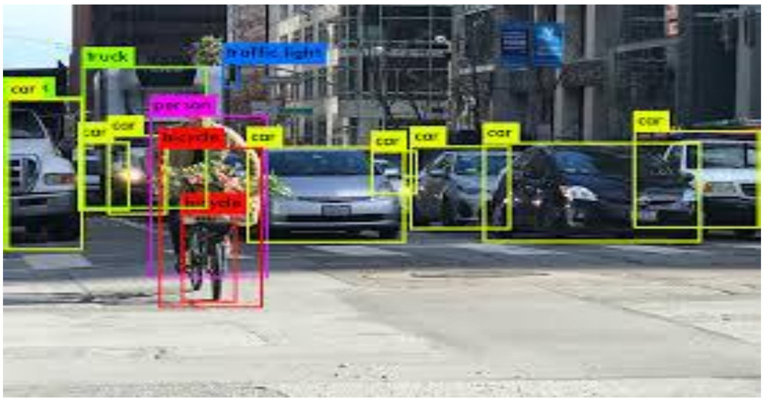

Fig. 2 Objects detected

A. Collection of data

The data to input in this model are images. As mentioned previously the datasets being used are:

- **Inclusive Images** (subset of Google v4): This subset contains 1,743,042 training images with annotated bounding box around every object in the image. I is divided into the training and tuning set. Since there will be a great difference between the train and the test set, tuning plays a very important role. (Fig. 3)
- **Newly curated Inclusive images**: This dataset was developed by parsing the web for related images. After this was done the images were annotated. In the first step the dataset is annotated manually using Labelbox. And in the second step the annotations for the remaining samples is derived using a model which was trained with the first stage annotations. This is done after the following steps.

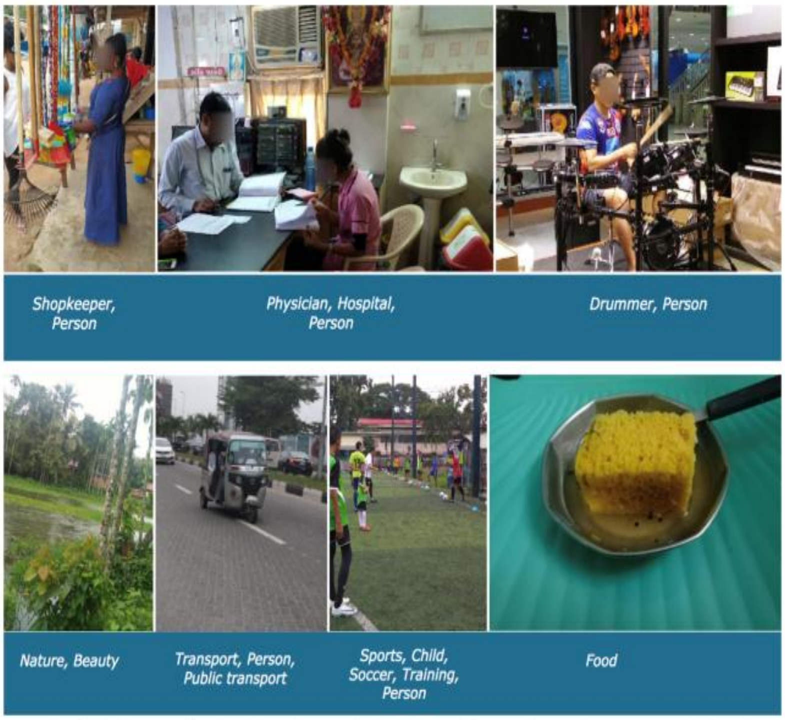
Shopkeeper, Person
Physician, Hospital, Person
Drummer, Person
Nature, Beauty
Transport, Person, Public transport
Sports, Child, Soccer, Training, Person
Food

Fig. 3 Examples of the Google Inclusive Image dataset

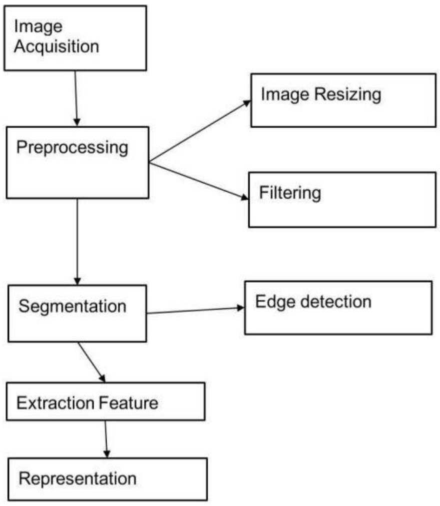

Fig. 4: Image Processing

B. Image Processing [6] (Fig. 4) YOLO has 24 convolutional layers followed by 2 fully connected layers.

- **Image acquisition:** Two ways for creating a digital picture are with a digital camera or a Flatbed scanner. Most of these images need to go through the following stages.
- **Pre-processing**: This means to specifically evacuate the repetition present in the image without affecting the interest region of the image. This includes the accompanying two phases-Image Resizing and Filtering. Vulnerabilities are brought into the picture in several ways. This results in data loss and even a noisy image. Filtration the process of removing this.
- **Segmentation:** This is the process of isolating the various interest regions in an image using techniques such as edge detection.
- **Feature Extraction:** Feature is the important portion of the image; hence it is essential to recognize them dependably. This is done in the first few convolutions as shown in the fig.
- **Representation:** This step involves changing the information to a frame appropriate for computer processing.
- **Training the model:** The model is trained with the datasets as mentioned above. YOLO predicts several bounding boxes for each grid in the image. During training we require only one bounding box predictor for each of the object. Therefore, one predictor will be assigned to be “responsible” for predicting an object based on the prediction that has the highest current IOU with the ground truth. Each predictor will be better at predicting certain sizes, aspect ratios, and basically the overall recall. During training of the model the objective is to optimize the loss function:

$$\begin{aligned}
&\lambda_{\text{coord}} \sum_{i=0}^{S^2} \sum_{j=0}^{B} \mathbb{1}_{ij}^{\text{obj}} \left[(x_i - \hat{x}_i)^2 + (y_i - \hat{y}_i)^2 \right] \\
&+ \lambda_{\text{coord}} \sum_{i=0}^{S^2} \sum_{j=0}^{B} \mathbb{1}_{ij}^{\text{obj}} \left[\left(\sqrt{w_i} - \sqrt{\hat{w}_i} \right)^2 + \left(\sqrt{h_i} - \sqrt{\hat{h}_i} \right)^2 \right] \\
&+ \sum_{i=0}^{S^2} \sum_{j=0}^{B} \mathbb{1}_{ij}^{\text{obj}} \left(C_i - \hat{C}_i \right)^2 \\
&+ \lambda_{\text{noobj}} \sum_{i=0}^{S^2} \sum_{j=0}^{B} \mathbb{1}_{ij}^{\text{noobj}} \left(C_i - \hat{C}_i \right)^2 \\
&+ \sum_{i=0}^{S^2} \mathbb{1}_{i}^{\text{obj}} \sum_{c \in \text{classes}} \left(p_i(c) - \hat{p}_i(c) \right)^2
\end{aligned}$$

Fig. 5: The loss function

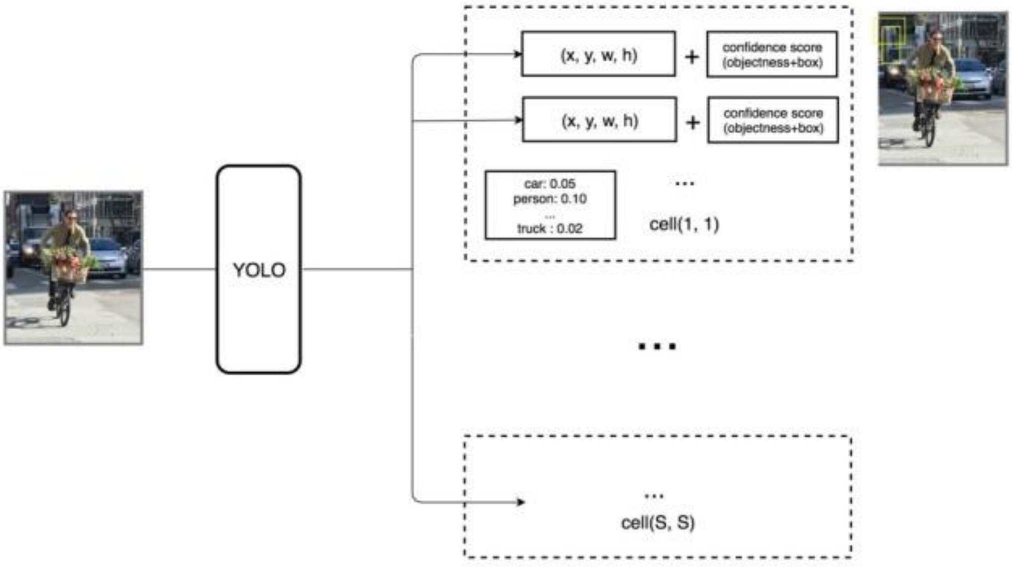

Fig.6: The bounding box prediction stages

IV. RESULTS

The developed is way more accurate in 20% more number of trials compared to the YOLO model which is not trained using the Inclusive Images when the objects to be detected are from different geographical as regions(belonging to the regions shown in Fig.7).

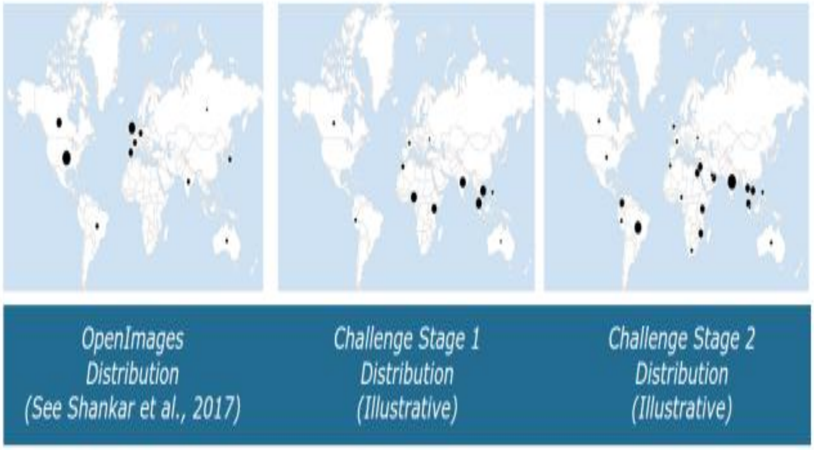
OpenImages Distribution (See Shankar et al., 2017)
Challenge Stage 1 Distribution (Illustrative)
Challenge Stage 2 Distribution (Illustrative)

Fig 7: The geographical regions covered

$$Precision = \frac{TP}{TP + FN}.$$

$$Recall = \frac{TP}{TP + FP}.$$

TP, FP, and FN respectively represent the number of samples that efficiently recognized the objects, the number of samples that identified another object as interest object, and the number of samples that identified the interest object is another object.

	Pre-Trained YOLO	YOLO trained with inclusive datasets
Precision	96.82%	97.63%
Recall	87.83%	89.28%
IOU	74.37%	76.67%

Fig. 9: Comparison of the two models

V. Conclusion

A dataset can greatly affect the efficiency of the training and the performance of a deep learning mode. The represented in this paper adopts improved image processing methods and effective targeted datasets in order to effectively detect globally inclusive objects i.e., the same object but with few differences in appearance because it's from different geographical region. This model can further be improved by creating a model that can detect inclusive images even when sufficient data is not available to train it.

References

1. Jawadul H. Bappy and Amit K. Roy-Chowdhury, "CNN based region proposals for efficient object detection", 2016 IEEE International Conference on Image Processing (ICIP), IEEE, 2016, pp.3658-3662.
2. Ross Girshick, "Fast R-CNN", IEEE International Conference on Computer Vision, 2015, pp. 1440-1448.
3. Shaoqing Ren, Kaiming He, Ross Girshick, Jian Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks" IEEE Transactions on Pattern Analysis and Machine Intelligence (Volume: 39, Issue: 6),2017, pp. 1137-1149.
4. Chengcheng Ning, Huajun Zhou, Yan Song , linhui Tang, "INCEPTION SINGLE SHOT MULTIBOX DETECTOR FOR OBJECT DETECTION", Proceedings of the IEEE International Conference on Multimedia and Expo Workshops (ICMEW), 2017, pp. 549-554.
5. Joseph Redmon, Ali Farhadi, "YOLO9000: Better, Faster, Stronger",IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 6517-6525
6. Hyeok-June Jeong, Kyeong-Sik Park, Young-Guk Ha "Image Preprocessing for Efficient Training of YOLO Deep Learning Networks", IEEE International Conference on Big Data and Smart Computing, 2018, pp.635-637
7. Chanran Kim, Younkyoung Lee, Jong-II park, " Diminishing unwanted Objects Based on Object Detection using Deep Learning and Image Inpainting", IEEE, 2018.
8. Swetha M S and Vallae Haritha , "Object Detection using Single Shot Detector for a Self-Driving Car", in International Journal for Innovative Research in Science & Technology (IJIRST) VOLUME 5, ISSUE 7 December 2018 ISSN (online): 2349-6010 pp.
9. Swetha M S and Veena M Shellikeri2, " Survey of Object Detection using Deep Neural Networks", in International Journal of Advanced Research in Computer and Communication Engineering Vol. 7, Issue 11, November 2018 ISSN (Online) 2278-1021 ISSN (Print) 2319-5940 pp.19-24
10. Swetha M S and Srishti Suman, "Survey on Faster Region Convolution Neural Network for Object Detection", in International Journal on Future Revolution in Computer Science & Communication Engineering ISSN: 2454-4248 Volume: 4 Issue: 11 November 2018 pp. 79 – 85
11. Swetha M Dr.Thungamani M and,Ankita Mishra, Enhancement of Performance Analysis in Anonymity MANET through Trust-Aware Routing Protocol. In the proceedings of International Journal of Advance Research in Computer Science and Management Studies, Volume 5, Issue 5, May 2017.