Object Detection With Voice Feedback

Abstract

Computer vision technologies, particularly the Deep Convolutional Neural Network, were developed rapidly in recent years. Use of state-of-the-art computers is providing promising vision techniques to help vision loss sufferers. In this project we try to visualize an object kept at a particular spatial location using a camera and then output the result as an audio speech so that the visually impaired person could hear.

1. Introduction

In recent years computer vision technologies have been developed which are very accurate and give promising results presented real-time object detection system using a CNN in order to recognize objects. In this project, we explored the possibility of using the hearing sense to understand visual objects. We aim to build a real-time object detection with a voice feedback system with the goal of telling the user what all is in the surroundings with its spatial position. This will help the visually impaired navigate in their surroundings safely. The detection of objects will be done from real-time video taken from their mobile phone camera. YOLO is a much improved version in terms of speed when compared with its preceding models like R-CNN, Fast R-CNN. We are trying to implement the "You Only Look Once: Unified, Real-Time Object Detection" YOLOv3 algorithm to identify the object present before the person. Then the label of the object and its location in the frame is identified and then converted into audio by using Text to Speech conversion. The person can use the output of our system in order to make him aware of his surroundings.

2. Literature Review

This paper [1] introduces a live object recognition system that serves as a blind aid for visually impaired people. The project used a Convolutional Neural Network for recognition of pre-trained objects on the ImageNet dataset. A camera aligned with the systems pre-determined orientation, serves as input to a computer system, which has the object recognition Neural Network deployed to carry out real-time object detection. Output from the network was presented to the visually impaired person in a suitable format.

This paper [2] discusses various techniques for object recognition and a method for multiple object detection in an image. The paper proposes the idea of detecting multiple objects in an image by using different object detectors simultaneously.

The paper [3] gives us insights about the open source computer vision library, also known as OpenCV. To start with motivation the authors point out to the the need to advance vision research and disseminate vision knowledge. The paper also deals with the latest functions available in opencv2 library. Matrix is the basic data structure for opencv. It provides the functions to initialize an image and allocate memory (which is not required for current version) and methods to access the pixel values at a particular location and channel. OpenCV has implemented several algorithms for image processing including some of the basic computations like histogram and equalize histogram which are helpful in analysing the image. Besides, the algorithm which is used for key point feature detection is SURF which is of our interest in our goal of object detection in this project. Finally it demonstrates the experiments with stereo images and it produced worst result. Nevertheless, it still does better than other global methods as it automatically takes the parameters by default unlike the global ones which take assigned parameters and find the optimal ones. The basic takeaway from the paper is about the library and the illustration of algorithms which detect features of image.

The piece of work done in [4] is a thesis on Assistive System for Visually Impaired using Object Recognition. It explores object recognition method and an assistive system which gives voice feedback to visually impaired people. This thesis really forms the basis of our project as the problem statement is very much similar to that of ours. Also, the main driving force to the project is clearly explained here , that is the number of visually impaired people all over the globe and a portion of people who are irreversibly blind. This provides us with many challenges regarding object and colour recognition. It first discusses various feature

descriptors and colour descriptors and object recognition modules which includes the most commonly used CNN. After extracting features, we need to use a classifier to classify images by learning the parameters from the classifier using the features extracted. The author prescribes GCRNN model which contains Gabor filters , CNN layers and multiple RNN layers and uses softmax classifier at the end. Lastly, he provides framework for assistive system. Starting with mode selection, the system goes through various blocks like mode selection, colour recognition and speech synthesizer. The concluding aspects of the thesis is about the performance of GRCNN on two different kinds of datasets and the overall framework of assistive system along with their hardware implementations.

The paper [5] talks about the implementation of YOLO algorithm and compares it with exisiting algorithms like R-CNN, Retina-Net, and Single-Shot MultiBox Detector (SSD). Although these approaches have solved the challenges of data limitation and modeling in object detection, they are not able to detect objects in a single algorithm run. The reason YOLO is more popular is because of its high speed. It uses neural networks to provide real-time object detection. Object detection consists of various approaches such as fast Object detection in YOLO is done as a regression problem and provides the class probabilities of the detected images. The algorithm requires only a single forward propagation through a neural network to detect objects and hence the name You Only Look Once. The algorithm first divides the image into various grids. Each grid has a dimension of S x S. Every grid cell will detect objects that appear within them. For example, if an object center appears within a certain grid cell, then this cell will be responsible for detecting it.YOLO also uses a single bounding box regression to predict the height, width, center, and class of objects. Intersection over union (IOU) in object detection describes how boxes overlap. YOLO uses IOU to provide an output box that surrounds the objects perfectly. Each grid cell is responsible for predicting the bounding boxes and their confidence scores. The IOU is equal to 1 if the predicted bounding box is the same as the real box. This mechanism eliminates bounding boxes that are not equal to the real box.

This paper [6] talks about using YOLOv4 algorithm as an improvement to the existing YOLO algorithm.Most of the modern accurate models require many GPUs for training with a large mini-batch size, and doing this with one GPU makes the training really slow and impractical. YOLO v4 addresses this issue by making an object detector which can be trained on a single GPU with a smaller mini-batch size. This makes it possible to train a super fast and accurate object detector with a single 1080 Ti or 2080 Ti GPU.

YOLOv4 achieves state-of-the-art results at a real time speed on the MS COCO dataset with 43.5 % AP running at 65 FPS on a Tesla V100. To achieve these results, they combine some features such as Weighted-Residual-Connections (WRC), Cross-Stage-Partial-connections (CSP), Cross mini-Batch Normalization (CmBN), Self-adversarial-training (SAT) and Mish-activation, Mosaic data augmentation, DropBlock regularization, and CIoU loss. These are referred to as universal features because they should work well independently from the computer vision tasks, datasets and models.The main components of a modern one-stage object detector are backbone, neck and head. YOLOv4 basically uses one of the three models as its backbone. Three feature extractor models include : CSPResNext50,CSPDarknet53, EfficientNet-B3. Neck is the second stage where features are gathered that were initially formed in the backbone and later these features will be fed to the head for detection. YOLOv4 has several options: FPN (Feature Pyramid Networks),PAN (Path Aggregation Network),SPP (Spatial Pyramid Pooling). The main objective of the head in YOLOv4 is to perform prediction that includes classification and regression of bounding boxes. It uses the YOLOv3 head. It provides information regarding coordinates of bounding boxes. It includes width, height, centre and score of prediction with the label. YOLOv4 head can be applied to every anchor box. Bag of Freebies can be termed as techniques that improvise costs related to training or strategy to ameliorate the accuracy of the model. It increases the performance of the model without compromising the latency at inference time and hence improvements are seen in data management and data augmentation.

This paper [7] talks about using Audvert mobile application that illustrates the potential of a lightweight system that facilitates navigation and serendipitous discovery of large indoor spaces using spatial audio. When a visitor enters a shopping mall for the first time, it may be a large, unfamiliar indoor location and confusing place.Based on the location of the user, Audvert app randomly picks a store from the mall and begins playing an audio clip containing the name and description of the goods and services available there. When a user hears a point of interest (POI) played back through their headphones, the audio sounds appear to originate from the actual physical locations indicating the direction of point of interest. As the user approaches the store, the amplitude increases. A user should be able to infer proximity through amplitude and direction through the panning of sounds. Audvert tries to simplify the task of spatial display selection by making only one choice available at a time, by using a single active element and a simple shake gesture to select what user wants. Audvert provides better understanding of the

location of shops and encourage participants to visit new places. Audvert is unique in that it can be used indoors, layering useful, spoken information about points of interest with directional and proximity feedback.

This paper [8] talks about portable blind aid device.The blind aid device must enable a blind person to activate the blind aid device; capture one or more images related to a blind person's surrounding environment; detect moving objects from the images captured; identify a finite number of spatial relationships related to the moving objects; analyse one or more images within the blind aid device to classify the finite number of spatial relationships related to the moving objects corresponding to predefined moving object data; convert selected spatial relationship information related to the analysed images into audible information; relay selected audible information to the blind person; and notify the blind person of occurrences predetermined by the blind person as actionable occurrences.

This paper [9] presents some updates to YOLO. Following YOLO9000, this system predicts bounding boxes using dimension clusters as anchor boxes. YOLOv3 predicts an objectness score for each bounding box using logistic regression. Each box predicts the classes the bounding box may contain using multilabel classification. For training, binary cross-entropy loss has been used for the class predictions. It uses a new network for performing feature extraction. This new network is a hybrid approach between the network used in YOLOv2, Darknet-19. The performance of YOLOv3 is pretty good. In terms of COCOs average mean AP metric it is on par with the SSD variants but is 3× faster. It is still quite a bit behind other models like RetinaNet in this metric. However, when we look at the old detection metric of mAP at IOU=0.5 (AP50 metric) YOLOv3 is very strong. It is almost on par with RetinaNet and far above the SSD variants. This indicates that YOLOv3 is a very strong detector that excels at producing decent boxes for objects. However, performance drops significantly as the IOU threshold increases indicating YOLOv3 struggles to get the boxes perfectly aligned with the object.

This paper [10] introduces YOLO9000, a state-of-the-art, real-time object detection system that can detect over 9000 object categories. The paper proposes various improvements to the YOLO detection method, both novel and drawn from prior work. This improved model YOLOv2 performs pretty good on standard detection tasks like PASCAL VOC and COCO. Using a novel, multi-scale training method the YOLOv2 model can run at varying sizes, offering an easy tradeoff between speed and accuracy. Finally the paper proposes a method to jointly train on object detection and classification. Using this method YOLO9000 can be trained simultaneously on the COCO detection dataset and the ImageNet classification dataset. This joint training allows YOLO9000 to predict detections for object classes that don't have labelled detection data. This approach has been validated on the ImageNet detection task. YOLO9000 gets 19.7 mAP on the ImageNet detection validation set despite only having detection data for 44 of the 200 classes. On the 156 classes not in COCO, YOLO9000 gets 16.0 mAP.

3. Solution to the problem addressed

We want to implement yolov3 algorithm in this paper and use that to detect multiple objects in one image and generate audio file which generates speech based describing the objects in the image.

4. Preliminary Results

The following important functions involved in the implementation of a YOLOv3 model were replicated by us :

1. **Convolution Layers and Batch Normalization :** We first built a 2D convolution network using Tensorflow and also added padding. Then we performed batch normalization for reducing the training epochs. Batch normalization helps in smoothening the loss function which in turn by optimizes the model parameters thereby improving the generalization performance and training speed of the model.

2. **Upsampling :** In order to concatenate the outputs from Darknet-53 before applying detection on a different scale, upsampling is used on the feature map using nearest neighbor interpolation which is equivalent to assigning the pixel intensity of the nearest neighbour.

3. **Non maximum-separation :** We typically use bounding boxes for identifying candidate regions for the object of interest. Most of the approaches employ a sliding window over the feature map and assigns foreground/background scores depending on the features computed in that window. The neighbourhood windows have similar scores to some extent and are considered as candidate regions. This leads to hundreds of regions. We filter the regions and discard the boxes with high overlap as well as boxes with low confidence scores.

4. **Feature Extraction :** The feature extractor used in YOLOv3 is a hybrid of YOLOv2, Darknet-53 (network trained on ImageNet), and Residual networks(ResNet). The extracted features serve as input to detecter. The network uses 53 convolution layers, where the network is built with consecutive 3x3

and 1x1 convolution layers followed by a shortcut residual connection (which help the activations propagate through deeper layers). YOLOv3 is a multi-scale extractor , using 3 different scales (different for large,medium and small objects).

The input image of size $416 \times 416 \times 3$ and the following displays the structure of darknet 53.

	Layer	filter	strides	output
	Convolutional	3×3×32	1	416×416×32
	Convolutional	3×3×64	2	208×208×64
1×	Convolutional	1×1×32	1	
	Convolutional	3×3×64	1	
	Residual			208×208×64
	Convolutional	3×3×128	2	104×104×128
2×	Convolutional	1×1×64	1	
	Convolutional	3×3×128	1	
	Residual			104×104×128
	Convolutional	3×3×256	2	52×52×256
8×	Convolutional	1×1×128	1	
	Convolutional	3×3×256	1	
	Residual			52×52×256
	Convolutional	3×3×512	2	26×26×512
8×	Convolutional	1×1×256	1	
	Convolutional	3×3×512	1	
	Residual			26×26×512
	Convolutional	3×3×1024	2	13×13×1024
4×	Convolutional	1×1×512	1	
	Convolutional	3×3×1024	1	
	Residual			13×13×1024

5. **Classification loss :** The model uses binary cross entropy as loss function for each label. Object confidence i.e the probability that the bounding box is an object and class probability i.e the probability that the object belongs to that particular class are predicted using logistic regression.

6. **Detection layers :** The YOLOv3 model uses 3 detection layers, that detect on 3 different scales using respective anchors. For each cell in the feature map the detection layer predicts $\text{n\_anchors} * (5 + \text{n\_classes})$ values using 1x1 convolution. For each scale we have the number of anchors equal to 3. For each of 3 anchors we are going to predict 4 coordinates of the box, its confidence score (the probability of containing an object) and class probabilities.

The implementations of above functions can be found at the GitHub Repository : `https://github.com/harithar1234/project_collection/tree/main/Deep_Learning-AI2100/Object_detection`.

5. Final Results and Conclusions

Brief overview of YOLOv3 architecture

53 Convolution layers called Darknet-53 are used. Apart from those 53 layers, 53 more convolution layers are used producing a total of 106 layers. Detections are made at 3 layers - 82,94,106. Each convolution operation is followed by batch normalization as well as the application of leaky ReLU activation. There are no pooling layers but instead convolution layers with stride 2 are used to downsample feature map as it prevents loss of lower level features. This results in the ability to detect small objects which would have been excluded if pooling layers had been used. Yolo makes detection at 3 different layers - 82, 94 and 106. We used **pretrained model weights** and drew bounding boxes to the objects detected in the image. It identified objects with a confidence of about 99%. We used pretrained weights for our model since to train the model it would have taken a huge amount of time mainly due to the depth of the model and the large size of the dataset. We have made the following observations :

- Changing alpha produced many bounding boxes in the picture.
- Changing momentum has a small effect on the bounding boxes produced.

We referred to an existing project on Kaggle [11] to learn about object detection and provided a further extension.

Extension to the detection algorithm

We converted the text labels produced along with their position to audio files to help people hear what they cannot see. We used the google text to speech api for performing this task. The text-to-speech program, tool, or software takes an input text from the user, and using methods of natural language processing understands the linguistics of the language being used, and performs logical inference on the text. This processed text is passed into the next block where digital signal processing is performed on the processed text. Using many algorithms and transformations this processed text is finally converted into a speech format. This entire process involves the synthesizing of speech. The preprocessor tokenizes a sentence into words and converts the words into phenomes (based on pronunciation).

The **final model** and the **results** that we have achieved can be found at the Github Repository : `https://github.com/harithar1234/project_collection/tree/main/Deep_Learning-AI2100/Object_detection`.

References

[1] Kedar Potdar, Chinmay D. Pai, and Sukrut Akolkar. A convolutional neural network based live object recognition system as blind aid. arXiv:1811.10399v1 [cs.CV] 26 Nov 2018.

[2] Khushboo Khurana and Reetu Awasthi. Techniques for object recognition in images and multi-object detection. International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) Volume 2, Issue 4, April 2013.

[3] A.Culjak, D.Abram, T.Pribanic, H.Dzapo, and M.Cifrek. A brief introduction to opencv. Proceedings of the 35th International Convention MIPRO,Opatija,2012.

[4] Rahul Kumar and Sukadev Meher. Assistive system for visually impaired using object recognition. MSc thesis at Department of ECE,NIT Rourkela,Odisha,India,May 2015.

[5] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once:unified, real-time object detection. University of Washington, Allen Institute for AI, Facebook AI Research 9 May 2016.

[6] Deeksha Janardhan, Madhuri B J, Harsha Kumar, Ketan Sahu, , and Suhas. S. Object detection with voice feedback using yolov4. Department of Computer Science and Engineering The National Institute of Engineering, Mysore, India 2021.

[7] Liam Betsworth, Nitendra Rajput, Saurabh Srivastava, and Matt Jones. Audvert: Using spatial audio to gain a sense of place. InHuman Computer Interaction-INTERACT 2013.

[8] Evanitsky and Eugene. Portable blind aid device. U.S Patent 10 Dec 2013.

[9] J. Redmon and A. Farha. Yolov3: An incremental improvement. arXiv, 2018.

[10] J. Redmon and A. Farhadi. Yolo9000: Better, faster, stronger. In Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on, pages 6517–6525. IEEE, 2017.

[11] Yolov3 and tensorflow. `https://www.kaggle.com/code/aruchomu/yolo-v3-object-detection-in-tensorflow`.