COCO Dataset Stuff Segmentation Challenge

Divyansh Puri
Dept. of ECE
HMRITM
New Delhi, India
divyanshpuri02@gmail.com

***Abstract*— In computer vision, image segmentation is a method in which a digital image is divided/partitioned into multiple set of pixels which are called super-pixels, stuff segmentation challenge is a newly introduced task in which we have to segment out stuff out of the digital image. This job is to promote the state of the art in semantical category segmentation. Semantic segmentation defines the method by which each image pixel with the category is combined. Semantic classes can be well-defined stuff such as a vehicle, an individual or things like huge backdrop areas such as anarea covered with grass. stuff classes are crucial, as they explain aspects of an image such as type of scene and thing classes represent things, objects, there are various factors to it such as location, physical attributes, type of the material and properties of the scene. Only courses like lawn, ceiling, are focused on this problem. The method in this paper consists of a convolutional neural network and provides a superior framework pixel-level task, the dataset used in this research is the COCO dataset [1], which is used in a worldwide challenge on Codalab. It is a huge dataset with a total of 164k images in which around 118k images in the training dataset and around 5k images in the validation dataset, 20K images in test-dev , 20K images in test-challenge, we use this dataset to analyze the importance of thing and stuff classes in terms of their cover range and to check how many times they appear in image captions, this research is also important to find out whether stuff segmentation is easier than thing segmentation or not.**

Keywords— Adaptive Moment Estimation, COCO stuff, PSP Net, Stuff Segmentation, Rectified linear unit.

I. INTRODUCTION

Image segmentation is a computer vision job to unify the activities of semantic segmentation. One significant element of autonomous driving is semantic segmentation. The computer vision society has attracted exposure to that assignment. Instance Segmentation is closely related to object detection. But in this case, the output is an object mask rather than a bounding box. We don't label all pixels in the picture here, which only seeks to find the boundaries of particular objects. One picture is a pixel set. we simply cluster these panels together, which have comparable characteristics with image segmentation. If we want to divide the image into object and background, we define a single threshold value. If we have multiple Objects in the background then we need to define multiple thresholds. These are generally known as the local thresholds. For the job of image analysis, segmentation is very important. Semantic segmentation explains how each pixel of a picture is combined with a category. Semantic classes can be well-specified items, like cars, or things like vast backdrop areas such as an area covered in grass. Here things courses are essential, they describe significant elements of a picture, such as sort of landscape, and things courses, place, physical characteristics, product form and scene characteristics are probable to occur. To comprehend the materials and stuff, In this paper, we bring them individually COCO dataset with 91 stuff classes and 1 other class, a total of 164k images containing 118k training sets, 20K images in test-dev, 20K images in test-challenge and 5k validation sets. The original coco dataset previously provided annotations for 80 thing classes, the dataset provided us with location annotations for potted plants, person and train. This amount of information is not sufficient enough to understand the scene context, to understand the scene(consider scene for a train) we require the object that is the train, its interaction with the stuff associated (tracks), the spatial arrangement of the object(train) and the stuff surrounding it. We use this dataset to analyse the importance of thing and stuff classes in terms of their cover range and to check how many times they appear in image captions, this research is also important to find out whether stuff segmentation is easier than thing segmentation or not. An Example image of Semantic Segmentation is given below in Fig.1.

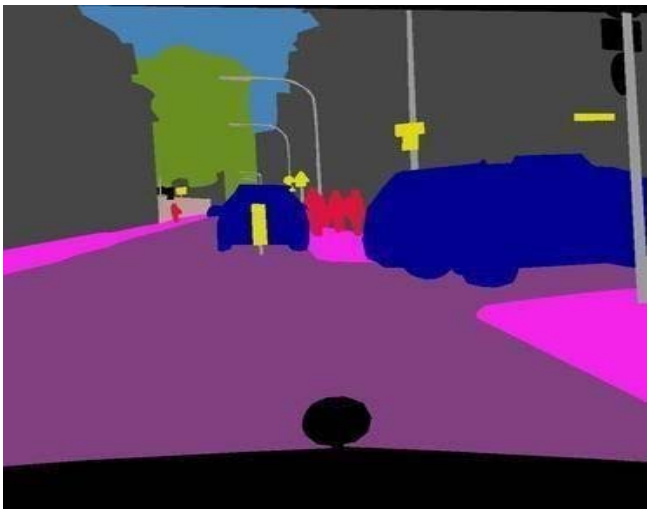

Fig.1. Example image of Semantic Segmentation.

II. METHODS

A. Image Classification

This implies that each major entity in the picture is identified.

978-1-7281-4042-1/19/$31.00 ©2019 IEEE

B. Localization

The machine locates the object's position on an picture with a bounding box in a place that has a separate tag.

C. Object Detection

Computer ranks and locates items in the image, which is generally displayed by a box surrounding the point of interest.

D. Semantic Segmentation

It defines how each picture (flower, an individual, street) is associated.

E. Instance Segmentation

In Instance Segmentation our main aim is to detect a given separate item in an image.

III. RELATIVE WORK

Semantic segmentation is intended to mark each pixel in a semantic category, here we discuss additional work related to stuff segmentation, focusing mainly on the different model implementation.

A. R-CNN

For semantic segmentation, R-CNN [7] is used which extracts two types of CNN features those are region features [7] extracted from the given bounding boxes the other is segment features extracted from the raw image content masked by the segments. The working of R-CNN model is given below in Fig.2.

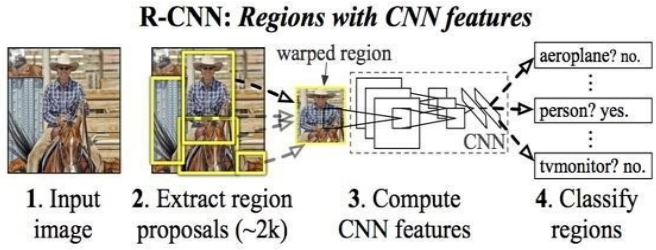

Fig.2. Working of R-CNN.

The suggested procedure utilizes systematic testing for only 2000 areas to get out of the picture, to bypass the issue of choosing a large number of areas. We can only operate with 2000 areas to rank an enormous amount of areas. The main disadvantage is that it is too slow. Region proposals need to wrap into a fixed size and it changes the object's appearance moreover cropping may lose the information. Training is expensive and slow because of selective search.

The R-CNN and Fast R-CNN filters select regions [10] using a specific query. This quest is a long and lengthy method that influences network performance. Object detection algorithm eliminates the selective search algorithm and allows the network learn the region proposals [11]. The working of YOLO model is given below in Fig.3.

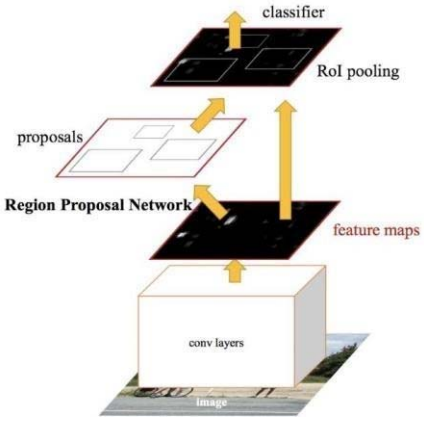

Fig.3. Architecture of faster R-CNN.

B. YOLO

People have used YOLO(You Only Look Once) model in literature for image segmentation [8] in areas to locate an item in the picture, for this purpose prior object detector systems [6] are used. It is not the whole picture of the network rather, the sections of the picture that are highly likely to have the items. YOLO is real time multipleobject detection algorithm, this architecture divides the entire into a A x A grid and searches the item by constructing boxes around those objects, these boxes are called bounding boxes and displays the probabilities predicted of these regions. The working of YOLO model is given below in Fig.4.

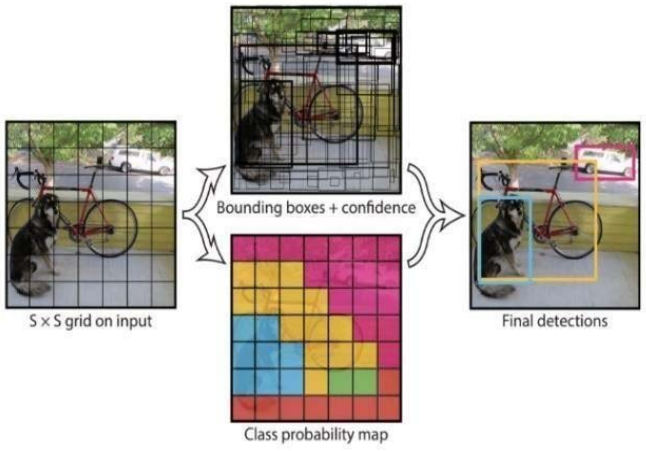

Fig.4. Working of YOLO

C. Mask R-CNN

Deep learning techniques have achieved massive results for object detection and segmentation, most popular among these is the Region-based Convolutional neural networks or R-CNN, the most popular technique is called Mask R-CNN, it is basically a category of R-CNN and an extension

to the old faster-RCNN technique which predicts a mask for a detected object/item in the image.
We use third-party implementation rather than developing our version from scratch. There are various mask R-CNN models present on the internet today but best of them is Mask R-CNN by Matterport. The R-CNN mask [9] is a quicker R-CNN expansion. It seeks to fix the computer vision segmentation issue. This model separates different items in the image, Then the object class is predictedusing a bounding box and a mask is made based on the proposal at the pixel level. Mask R-CNN framework implementation is given below in Fig.5.

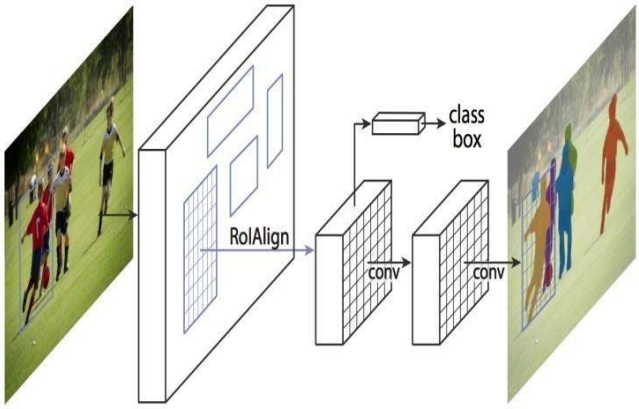

Fig.5. Mask R-CNN framework implementation

IV. METHODOLOGY

Computer vision is an important aspect of computer science that deals with the computer to identify the images and videos as a human does. The primary goal of computer vision is to understand visual images.
Semantic segmentation is one of the tasks used in computer vision. It includes various sections or blocks. Each pixelof an image is labelled with class.

A. Dataset and Pre-Processing

COCO dataset [1] is used here and it contains 91 stuff classes and 80 object classes We use picture resizing to decrease the volume of the picture. For several reasons, it is important. As it is very task intensive if we run such a huge amount of images without resizing it would take a lot of time while training further consuming more memory and electricity, nevertheless in deep learning we often need to resize images, so that volume fits into the network which requires that an image be square.

B. Architecture

Here we use convolutional neural network architecture U-Net and PSP-Net.

1) U-NET

The U-NET comprises of an extended route that provides U form [12]. It is a network of symmetric decoders. It was created for biomedical picture segmentation. Its architecture has been altered and expanded to operate with fewer pictures and to produce more accurate segmentation.
There are two routes to the architecture. The first route is the route of compression (encoder), that is used to record the background of the picture and the down part of the sample. The encoder is only a conventional pack of convolutional and max layers of pooling. The second route is the symmetrical extension route (decoder), using transposed convolution for accurate localization and up-sampling. Architecture of U-NET is given below in Fig.6.

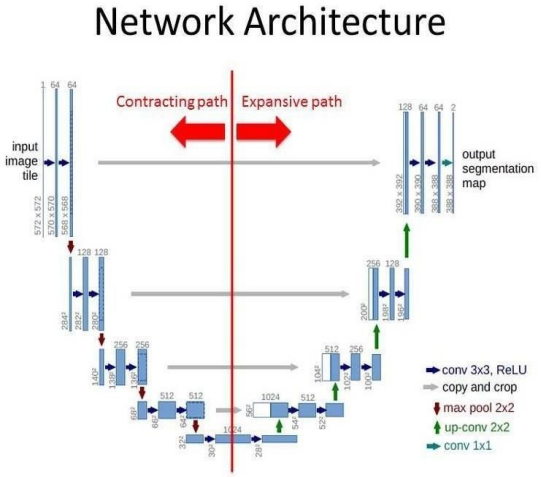

Fig.6. Architecture of U-NET

2) PSP-Net

PSP-NET(Pyramid scene parsing network) is introduced to place difficult scene context features in an FCN (Fully Convolutional Network) based pixel prediction network [3]. It gives insight into how segmented courses are distributed. This data is collected by using big kernel pooling levels in the pyramid pooling module. The framework of PSP-Net is illustrated below in Fig.7.

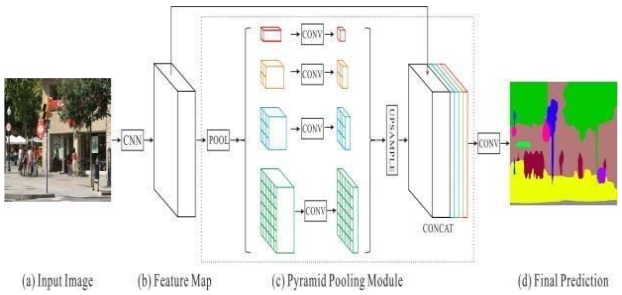

Fig.7. PSP-Net Architecture

C. Hyperparameters

1) Number of Epochs

An Epoch is the number of moments, for the training data provided to the network. As the precision oflearning

increases, the number of times until validation precision begins to decrease can be improved. The number of epochs can be increased or decreased depending upon the level increase in the accuracy of the model respectively.

1) ReLU

ReLU- Rectified Linear Unit (ReLU) is the most widely used activation function while designing networks today. An activation function is an artificial neuron feature that provides an input yield. The value of ReLU is zero for all negative values and linear for all positive values. It needs very less time to train and run the model. ReLU Activation function is shown below in Fig.8.

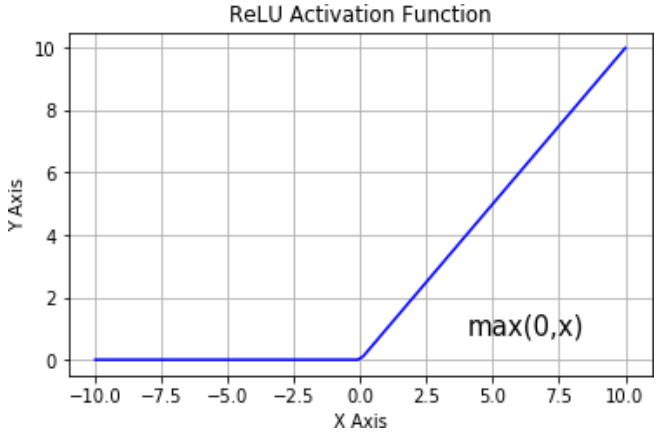

Fig.8. ReLU Activation function

D. Optimization

Adaptive Moment Estimation (ADAM) is a learning rate optimizer which is adaptive in nature. The training speed for every weight of the neural network is adjusted by the first and second moments of the degree to the optimum alternative. It is designed for deep neural networks.

E. Flowchart

A flowchart of the whole process has been reflected in the section from input until the output has been fetched. Here we used COCO dataset. From the dataset first we took the image as input. Then done image pre-processing by using image resizing and reducing noise. It then operates with the PSP-Net model of the convolutional neural network model after picture pre-processing. We get the Segmented Output Image after all this method. The flowchart has been given below in Fig.9.

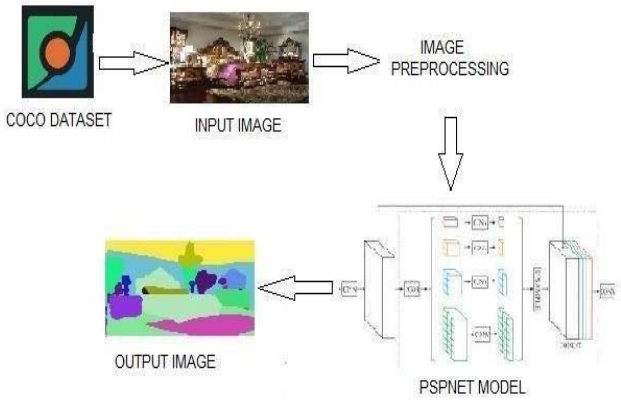

Fig.9. Working or flow of the challenge

V. EXPERIMENTAL RESULTS

We used different models and trained them on the 118k training images of the COCO dataset. With different models, we got different levels of accuracy and IOU values. The models that we used are RESNET50-SEGNET, U-NET, PSP-Net and Mask R-CNN. The comparison of different models is given below in TABLE 1.

TABLE 1. comparison of different models

Model Name	Average Accuracy/image	IOU
RESNET50-SEGNET	0.234	0.213
VGG16-UNET	0.332	0.247
VGG16-PSPNET	0.458	0.294
Mask R-CNN	0.491	0.50

As the table indicates we proceed step by step starting from the top, the results were not good as we started training our first model on COCO dataset, but after switching our model to U-NET we see a little improvement but the results were not that much satisfactory either, we increased the number of epochs in further models and got better accuracy thereafter we got best results from Mask R-CNN model which we ran for 50 epochs on the original dataset without image compression or resizing.

VI. FINAL OUTPUT

Fig.10. Input image

Fig.11. Output image

VII. CONCLUSION

With Mask R-CNN model implementation, we achieved the highest accuracy so far. Still, the accuracy achieved can be improved further for better results but tuning different values for different hyper-parameters. We also observed that image segmentation is a difficult task to perform on stuff classes as compared to thing classes.

REFERENCES

[1] Holger Caesar, Jasper Uijlings and Vittorio Ferrari. COCO-Stuff thing and Stuff Classes in Context, 2018.

[2] H.Caesar, J.Uijlings and V.Ferrari. Region- based semantic segmentation with end-to-end training. In ECCV, 2016.

[3] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang and Jiaya Jia. Pyramid Scene Parsing Network, 2017.

[4] R. Mottaghi, X. Chen, X. Liu, N.-G. Cho, S.-W. Lee, S. Fidler, R. Urtasun, and A. Yuille. The role of context for object detection and semantic segmentation in the wild. In CVPR, 2014

[5] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollar, and C. Zitnick. Microsoft COCO: Com- ´ mon objects in context.In ECCV, 2014

[6] Rayson Laroca, Evair Severo ; Luiz A. Zanlorensi ; Luiz S. Oliveira ; Gabriel Resende Gonçalves . A Robust Real-Time Automatic License Plate Recognition Based on the YOLO Detector. IEEE,2018.

[7] Zhang, Haijun, et al. "ClothingOut: a category- supervised GAN model for clothing segmentation and retrieval." Neural Computing and Applications (2018): 1-12.

[8] He, Kaiming, et al. "Mask r-cnn." Proceedings of the IEEE international conference on computer vision. 2017.

[9] Dai, Jifeng, et al. "R-fcn: Object detection via region- based fully convolutional networks." Advances in neural information processing systems. 2016.

[10] Zhang, Liliang, et al. "Is faster r-cnn doing well for pedestrian detection?." European conference on computer vision. Springer, Cham, 2016.

[11] Li, Xiaomeng, et al. "H-DenseUNet: hybrid densely connected UNet for liver and tumor segmentation from CT volumes." IEEE transactions on medical imaging 37.12 (2018): 2663-2674.