Machine Learning Nanodegree Capstone Proposal

S. Mohana Krishna

January 13th, 2018

1 Domain Background

In computer vision, Image segmentation is the process of partitioning images into multiple segments (sets of pixels, also called super pixels). The goal is to simplify the representation of an image to make it more meaningful and easier to analyze. More precisely, Image segmentation is the process of assigning a label to every pixel in the image such that pixels with the same label share certain characteristics. If the Image segmentation task involves identifying semantically meaningful regions in an image i.e. belonging to same object class, then it is called semantic segmentation and is an active area of research in computer vision.

Traditional methods for semantic segmentation used hand crafted features like SIFT, HoG etc. with classification algorithms like SVM, Random Decision Forest. Recently deep learning methods have become more prevalent and most of the state of the art algorithms use these methods. For this project we follow the approach discussed in `*DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs*`. The research tries to solve 3 challenges faced in the application of Deep Convolution Neural Networks to semantic image segmentation: (1) reduced feature resolution, (2) existence of objects at multiple scales, and (3) reduced localization accuracy due to DCNN invariance. My personal motivation behind choosing semantic segmentation is the challenging nature of the problem and its wide range of applications in computer vision tasks like self driving cars, understanding aerial imagery, object detection etc.

2 Problem Statement

The aim of the project is to understand and implement semantic image segmentation using Deep Convolutional Nets and Atrous convolutions. Semantic segmentation is the task of assigning a label(class) to each and every pixel in the image, and dividing the image into semantically meaningful parts. For this project, to every pixel in an image we assign one class out of 21 possible classes aeroplane, cat , person etc.

3 Datasets & Inputs

Dataset used for training and validation of the project is augmented PASCAL VOC dataset, available at
http://www.eecs.berkeley.edu/Research/Projects/CS/vision/grouping/semantic_contours/benchmark.tgz

Dataset used for testing the model is the PASCAL VOC dataset, available at
http://host.robots.ox.ac.uk/pascal/VOC/voc2012/VOCtrainval_11-May-2012.tar

The training dataset has 8498 training images and 2857 validation images.

The ground truth labels in the datasets, belong to the following classes :
(0=background, 1=aeroplane, 2=bicycle, 3=bird, 4=boat, 5=bottle, 6=bus, 7=car, 8=cat, 9=chair, 10=cow, 11=dining table, 12=dog, 13=horse, 14=motorbike, 15=person, 16=potted plant, 17=sheep, 18=sofa, 19=train, 20=tv/monitor).

4 Solution Statement

The use of DCNNs for semantic segmentation, or other dense prediction tasks, has been shown to be simply and successfully addressed by deploying DCNNs in a fully convolutional fashion. However, the repeated combination of max-pooling and striding at consecutive layers of these networks reduces significantly the spatial resolution of the resulting feature maps. In this project we overcome the issue by using *atrous convolutions*.
We finetune the model weights of Imagenet-pretrained VGG-16 networks with several modifications to adopt them to semantic segmentation: first, increased field of view due to *atrous convolutions*; second, to decrease memory consumption and time spent on performing one forward-backward pass the number of filters in the last layers is reduced from 4096 to 1024; third, to keep the down sampling ratio of 8, last pooling layers are omitted; third , we replace the 1000-way Imagenet classifier in the last layer to the number of semantic classes(including background), which is 21 in our case. Loss function is the sum of cross-entropy terms for each spatial position in the CNN output map. We optimize the objective function with respect to weights at all network layers using standard stochastic gradient decent procedure.
The input is an RGB image of dimensions 500*500 *3, output is of the dimensions 500*500*21 with the third dimension 21 representing the probability of a pixel belonging to one of the 21 classes. To predict the output we take an argmax over the output of CNN to get a matrix of 500*500 with each pixel labelled with a class, which can be converted to an RGB image.

5 Benchmark Model

The benchmark model for our project is the DeepLabCRF-LargeFOV model as proposed in the `*DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs*`. The benchmark model does a semantic segmentation task using Deep Convolutional Neural Networks with a combination of Atrous convolutions and finally couple the output with Fully Connected CRFs to increase the localisation accuracy of the output.

6 Evaluation Metrics

We use *mean IOU* (Intersection over Union), as an evaluation metric for the task of semantic segmentation. *IOU* can be mathematically defined as follows:

$$IOU = tp/(tp + fp + fn)$$

Where tp – number of true positives
fp – number of false positives
fn – number of false negatives
This calculation is done for each semantic class in the output, and the mean is taken over all the classes to get the desired *mean IOU*.

7 Project Design

7.1 Programming Language and Libraries

1. Python 3.7
2. Tensorflow

7.2 Design

First we pre-process our input images, we reshape all our input images to a standard 500*500*3 size. Ground truth images in augmented PASCAL VOC dataset are in *.mat* format, and so are converted to *png* files and reshaped to a standard 500*500* 3 shape, which are later converted to labelled images of shape 500*500*21, with the class, out of 21 classes, to which the pixel belongs marked 1 and the rest zero. The mean red, blue, green are calculated for the whole dataset and subtracted from all the input images in the dataset. After the pre-processing stage is completed, these input images are fed into a modified VGG16 network, with a number of layers performing convolutions, atrous convolutions, max pooling operations on the data and finally giving out output of the size 500*500*21. The sum of the cross entropy terms of all the positions in the output map is the loss function, which is optimised using standard stochastic gradient descent approach. Mean IOU is used as a metric to measure the performance of the model.

References

DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs