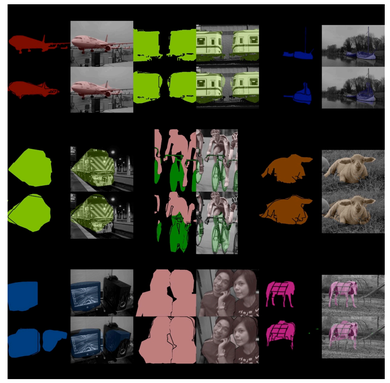
Street View Segmentation using FCN model

Yen-Kai Huang, Vivian Yang
nykh, viviany@stanford.edu
Department of Computer Science, Electrical Engineering

Introduction

Street view image segmentation is a very important task in the context of autonomous driving and scene understanding. In this ex-periment we used a newly released street view dataset to fine-tune on a pretrained VOC2011-FCN32 model and explore the performance.

In previous approach, segmentation trained on VOC2011-FCN32s model can successfully segment simple images, which focus on only few main objects, e.g. one or two animals, people or vehicles, and reach a quite high accuracy. However, they didn't try on complicated datasets, that one image has about ten or more objects to segment.



FCN32s VOC2012 Result

Problem Statement

In this project we deal with class-level street view segmentation. This task involves labeling each pixel as one of the many classes on a diverse set of street-view images, including cars, pedestrians, road, and immoveable objects.

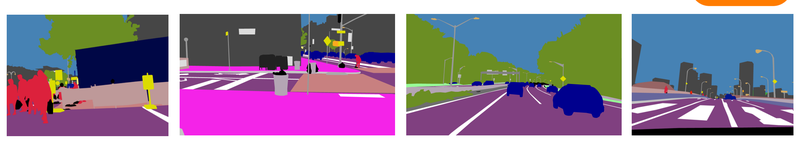
We applied a Fully Convolutional Network to the task, implemented in PyTorch with minor alterations. To evaluate the result we use both quantitative measurement of IoU accuracy and qualitative observation of the cohesion of identified object boundaries.

Mapillary Vistas Dataset

A very new dataset that features instance-specific, pixel-accurate labeling of images of a much more diverse set of geographical locations, weather circumstances, and day times.

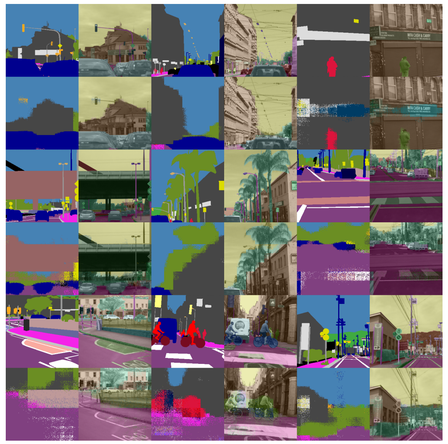
Image

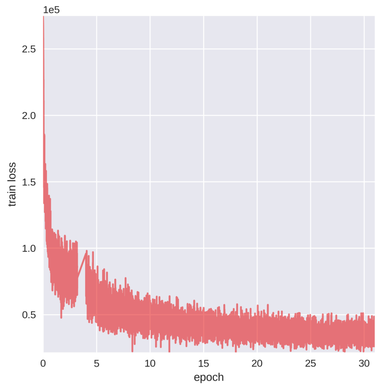
Labels



Experimental Results

Best Performance (LR = 1e-05)





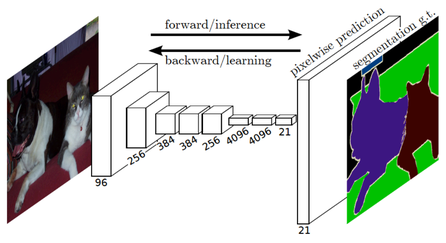
Training Data 5000, Validation Data = 100

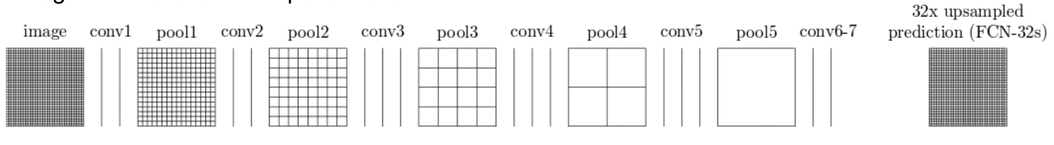
EntireModel MaxIter	EntireModel LR	Accuracy	Class Accuracy
6000	1e-07	72.574798584	9.7409898405
6000	1e-08	68.7076873779	9.03625497784
10000	1e-05	78.9661712646	19.6339660344
10000	1e-06	76.8708496094	15.899766389

With LastLayer MaxIter = 1000, LastLayer LR = 1e-04, Decay = 0.0005, BatchSize = 16

FCN-32s Model

Fully Convolution Network (FCN) is an end-to-end, pixels-to-pixels learning model, which can output a pixel-wise prediction and has been widely used for various segmentation tasks. The model differs from traditional model because it excludes any fully-connected layer and instead rely compltely on convolution operation. The model first perform many layers of convolution on the image and at the end perform a transposed convolution to output an image that matches the input dimension.





Conclusion

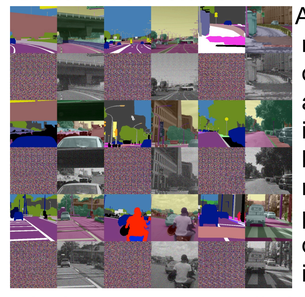
In the result achieved so far we were able to identify common objects in a streetview image, such as sky, road, car, pedestrians. This partially proves the feasibility of the task. However, the boundaries are still not very cohesive, which can be seen as a weakness of the FCN model itself.

Other difficulties posed by the Mapillary seg-mentation task is that it requires labeling some objects of very small or narrow dimensions, such as bird, street light, and sign poles, which can disappear during the down-sampling process.

To overcome this shortcomings we have tried to use Dilated Convolution variant. The result so far has not converged. We will further research into other alternatives to overcome the downsampling problem.

Future Works

Dilated Convolution



As seen in the current result, objects of small or narrow dimensions are not well identified in the output. This is because in the FCN model convolution and pooling involves the down-sampling the image, discarding a lot of information, especially for small objects. One proposal to overcome the issue is to use Dilated Convolution. This method avoid discarding information while gradually expanding the receptive field of the module.

Another model that was introduced for instance-level segmentation is the Mask R-CNN model. The model also uses convolution extensively and has been shown to work well for segmentation task. One of the future goal would be to either adopt or impelment this model and test on the Mapillary data set.