A Literature Review on Semantic Segmentation and Dimension Analysis

Naman Goyal
2015csb1021@iitrpr.ac.in
Koustav Das
2015csb1017@iitrpr.ac.in
Department of Computer Science and Engineering,
Indian Institute of Technology Ropar

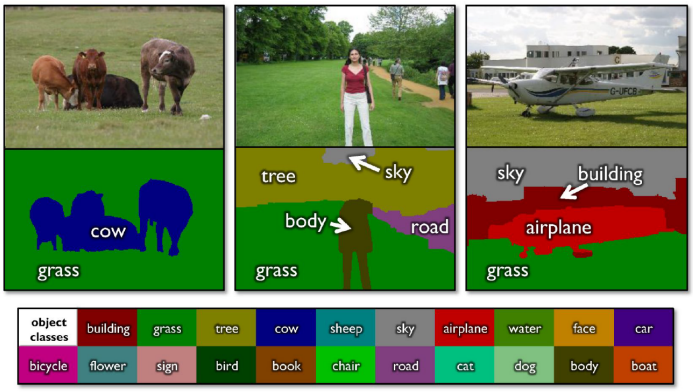

Figure 1: Semantic segmentation

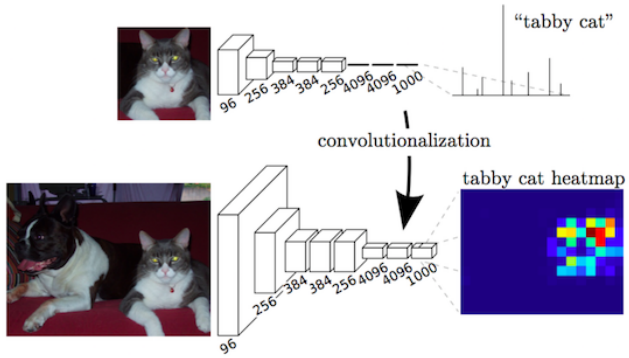

Figure 2: Fully Convolution Networks

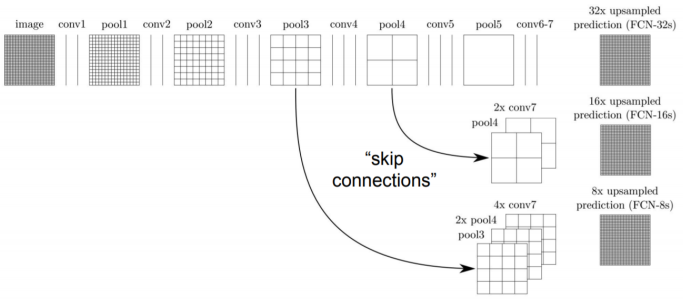

Figure 3: Upsampling using Shortcut Connections between convolution layers

Abstract

This paper presents a review of literature on pixel level semantic segmentation of images. Various techniques are compared including earlier works like unsupervised learning and decision forest. Then the recent works such as fully convolution networks, SegNet and dilated convolution are presented. The important data-sets and metrics are reported. Then dimension analysis is discussed and a pipeline model is proposed.

1 Introduction

Semantic segmentation refers to clustering of images into classes. It is also modeled as pixel to class mapping. All objects which are instances of same classes are grouped together. [4]

2 Earlier Work

The earlier work can be mainly divided into

2.1 Unsupervised Segmentation

These are non semantic approach using clustering algorithms and Graph based image segmentation.

Clustering algorithms can directly be applied on the pixels, when one gives a feature vector per pixel. Two clustering algorithms are k-means and the mean-shift algorithm.

Graph-based image segmentation algorithms typically interpret pixels as vertices and an edge weight is a measure of dissimilarity such as the difference in color.

Pros Few parameters to prune. Faster since no training phase required.

Cons Accuracy is hard to improve above a threshold. Semantics information is lost.

2.2 Random Decision Forests

This type of classifier applies techniques called ensemble learning, where multiple classifiers are trained.

There are two techniques either the feature or training data bagging. It is observed that an ensemble/ Random Forest from random sampling of features works very well, where the classifiers. are decision trees. A decision tree is a tree where each inner node uses one or more features to decide in which branch to descend. Each leaf is a class.

Pros One strength of Random Decision Forests compared to many other classifiers like SVMs and neural networks is that the scale of measure of the features can be arbitrary.

Cons Training takes significant time.

3 Recent Works

3.1 Fully Convolution Networks

The image is convoluted with a kernel which covers entire image. It works similar to patch model but is often much faster.

It does not have any of the fully-connected layers at the end, which are typically use for classification. Instead, it uses convolution layers to classify each pixel in the image. [2]

Upsampling is done either through "de-convolution" layers or shortcut connections.

Pros End to end fully convolution layer are easier to train because of parameter sharing.

Cons It produces coarse segmentation as dimension of image is reduced at each step due to pooling operation.

3.2 SegNet

A SegNet rather than upsampling image uses encoder and decoder network. The max-pooling indices are copied to decoder network.

The core trainable segmentation engine consists of an encoder network, a corresponding decoder network followed by a pixel-wise classification layer. The role of the decoder network is to map the low resolution encoder feature maps to full input resolution feature maps for pixel-wise classification. The novelty of SegNet lies is in the manner in which the

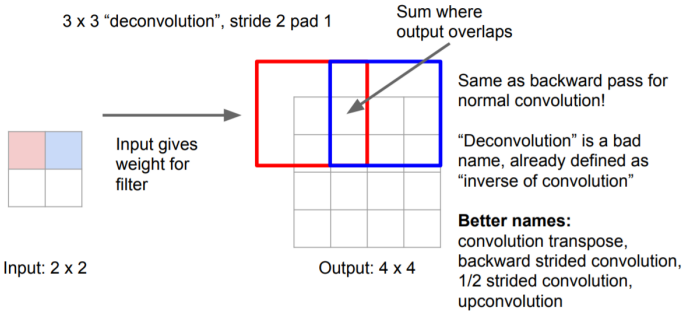

Figure 4: Upsampling using "Deconvolution" layers

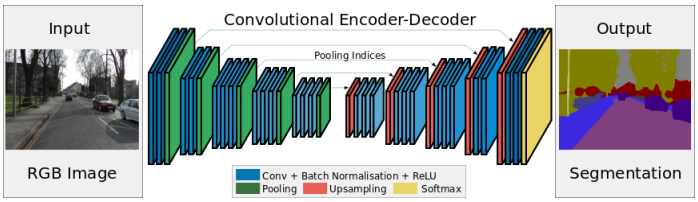

Figure 5: Segmentation Net Architecture

decoder upsamples its lower resolution input feature map(s). Specifically, the decoder uses pooling indices computed in the max-pooling step of the corresponding encoder to perform non-linear upsampling. This eliminates the need for learning to upsample. [1]

Pros Efficient both in terms of memory and computational time during inference. It is also significantly smaller in the number of trainable parameters than other competing architectures

Cons Poor benchmark performance over data-set.

4 Dilated Convolutions

A new convolution network module that is specifically designed for dense prediction. The module uses dilated convolutions to systematically aggregate multi-scale contextual information without losing resolution. The architecture is based on the fact that dilated convolutions support exponential expansion of the receptive field without loss of resolution or coverage. [6]

5 Data Set

The computer vision community produced a couple of different datasets which are publicly available.

The most important datasets are VOC2012 and MSCOCO (MSRC) which are large-scale object detection, segmentation, and captioning dataset.

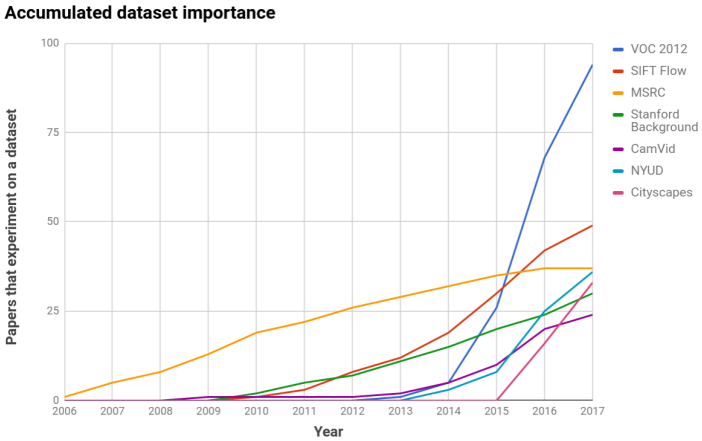

Figure 6: Accumulated data set importance

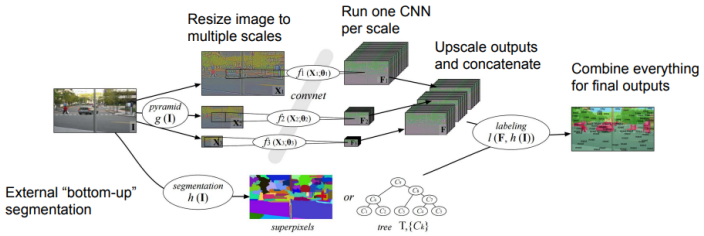

Figure 7: Multi-Scale Context Aggregation by Dilated Convolutions

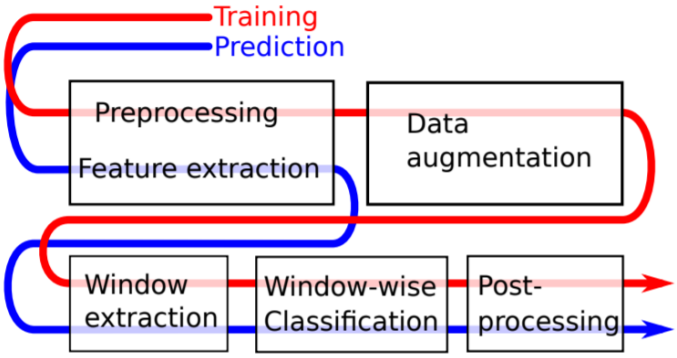

Figure 8: Experimental Protocol

6 Experimental Protocol

A typical segmentation pipeline gets raw pixel data, applies preprocessing techniques like scaling. For training, data augmentation techniques such as image rotation can be applied. For every single image, patches of the image called windows are extracted and those windows are classified. The resulting semantic segmentation can be refined by simple morphologic operations The output is compared and benchmarked. [4]

7 Measuring the size of objects in an image

There is a whole industry that is working particularly on this field known as Photogrammetry. Photogrammetry has been defined by the American Society for Photogrammetry and Remote Sensing (ASPRS) as the art, science, and technology of obtaining reliable information about physical objects and the environment through processes of recording, measuring and interpreting photographic images and patterns of recorded radiant electromagnetic energy and other phenomena.A special case, called stereophotogrammetry, involves estimating the three-dimensional coordinates of points on an object employing measurements made in two or more photographic images taken from different positions (see stereoscopy). Common points are identified on each image. A line of sight (or ray) can be constructed from the camera location to the point on the

Figure 9: Dimension computation of other objects from a reference object

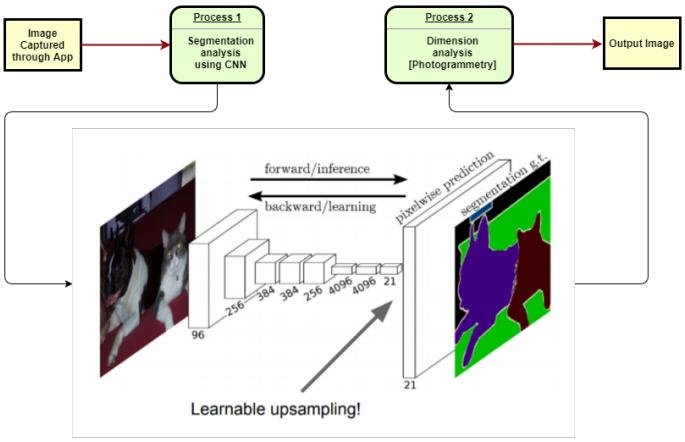

Figure 10: Pipeline of Model

object. It is the intersection of these rays (triangulation) that determines the three-dimensional location of the point. [5]

There exists a simple approach called “pixels per metric" ratio. This method requires a reference object that is quite distinctive in the segmented image and this helps us to extract pixels per metric ratio for a particular image. The reference object should also have a standard size for example a coin. After we have this ratio all the other dimension of the image is computed from the ratio.

It has been claimed that photogrammetry systems are able to measure smooth three-dimensional objects with surface height deviations less than $1\mu m$. [3]

8 Pipeline

The idea is to make an app based model. The user would take an snapshot of the object with their smart phone. The user needs to take the snapshot with a reference object kept beside the target object. The working model then runs an image segmentation algorithm using convolution neural network with learnable upsampling. The architecture of the same would resemble FCNs and based on GoogleNet and AlexNet. On this segmented image we apply dimension determination using “pixels per metric" ratio. The rest of the dimension computed from this ratio.

[1] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *arXiv preprint arXiv:1511.00561*, 2015.

[2] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3431–3440, 2015.

[3] Danny Sims-Waterhouse, Samanta Piano, and Richard Leach. Verification of micro-scale photogrammetry for smooth three-dimensional object measurement. *Measurement Science and Technology*, 28(5): 055010, 2017.

[4] Martin Thoma. A survey of semantic segmentation. *arXiv preprint arXiv:1602.06541*, 2016.

[5] Wikipedia. Photogrammetry — Wikipedia, the free encyclopedia. `http://en.wikipedia.org/w/index.php?title=Photogrammetry&oldid=796993379`, 2017. [Online; accessed 24-September-2017].

[6] Fisher Yu and Vladlen Koltun. Multi-scale context aggregation by dilated convolutions. *arXiv preprint arXiv:1511.07122*, 2015.