CITY UNIVERSITY LONDON

Deep Learning for Image Analysis

Written Report

Martin Fixman and Grigorios Vaitsas

2023/2024 Term

1 Introduction

Our aim in this study is to use the freely available Cityscapes dataset [1] to perform semantic segmentation analysis. Semantic segmentation is a computer vision technique used to understand and label specific parts of an image at the pixel level. The main goal is to partition an image into segments that correspond to different objects or classes of interest. Each pixel in the image is assigned a label that represents the class to which it belongs. Semantic segmentation is widely used in various fields. For instance, in autonomous driving, it helps cars understand the road environment by distinguishing roads, pedestrians, vehicles and other elements. In medical imaging, it aids in segmenting different tissues or detecting tumors. It's also used in other areas such as in agricultural robotics, where it helps in identifying crops and weeds for precision farming.

The Cityscapes dataset is a large-scale dataset widely used for training and evaluating algorithms in the fields of computer vision, particularly for tasks such as semantic understanding and urban scene segmentation. It features a collection of diverse urban street scenes from 50 different cities, primarily across Germany but also including some neighbouring countries.

The dataset includes over 5,000 fine annotations and 20,000 coarse annotations of high-resolution images (2048×1024 pixels). The fine annotations are detailed pixel-wise labels for 30 classes, of which 19 classes are considered for evaluation, such as roads, cars, pedestrians, buildings, and traffic lights. The detailed annotations are particularly valuable for tasks like semantic segmentation, where the goal is to assign a label to each pixel of the image.

The scenes represent a variety of seasons, daylight conditions, and weather scenarios, providing robust, real-world environments for training models that need to perform under varied conditions. This dataset has been widely used in research for developing and testing algorithms on tasks such as object detection, semantic segmentation, and instance segmentation in urban settings as well as advancing the state-of-the-art in visual perception for autonomous driving systems.

The Cityscapes dataset can be accessed in the following address:

https://www.cityscapes-dataset.com

Our particular approach in this study is to create and train at least two models; one with a basic architecture that will form our baseline model and a second one where we will be exploring a more complex architecture. We are aiming to demonstrate first of all that both our models are able to perform semantic segmentation on the chosen dataset. Our second goal is to investigate how the choice of various hyper-parameters and changes in architecture can affect the accuracy and performance of the models.

Modern, state-of-the-art techniques for image analysis involve convolutional neural

networks (CNNs) in their implementation. CNNs can be used to extract vital information from spatial features, which allow us to classify and segment objects in images. A type of neural network architecture designed specifically for semantic segmentation is the Fully Convolutional Network (FCN). This architecture, developed by researchers from UC Berkeley in 2014 [2], has been influential in advancing the field of computer vision, particularly in tasks requiring dense prediction, like semantic segmentation.

Unlike standard convolutional neural networks used for image classification, which typically end with fully connected layers, FCNs are composed entirely of convolutional layers. This design allows them to take input of any size and output segmentation maps that correspond spatially to the input image, providing a per-pixel classification. FCNs transform the fully connected layers found in traditional CNNs (like those in AlexNet or VGGNet) into convolutional layers. This is done by treating the fully connected layers as convolutions with kernels that cover the entire input region. For example a fully connected layer that accepts an input of size 7×7 can be reimagined as a convolutional layer with a 7×7 filter size. To generate output segmentation maps that match the size of the input image, FCNs use transposed convolution layers (also known as deconvolutional layers) for upsampling. This process helps in recovering the spatial dimensions that are reduced during the pooling or convolutional operations in earlier layers. FCNs often utilize skip connections to combine deep, semantic information from lower layers with the shallow, appearance information in the upper layers of the network. This helps in improving the accuracy and detail of the output segmentation maps, as it allows the network to use both high-level and low-level features.

U-nets

2 Methodology

3 Results

4 Conclusions

5 Reflections

References

[1] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. *CoRR*, abs/1604.01685, 2016.

[2] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. *CoRR*, abs/1411.4038, 2014.