Deep Learning for Semantic Scene Understanding

Yang, M.Y.
University of Twente

Abstract:

Visual understanding of complex urban scenes is an enabling factor for a wide range of applications. Despite significant advances, visual scene understanding remains challenging, particularly when taking human performance as a reference. Semantic segmentation, or scene classification, aims at segmenting images and detecting various object categories within them. A traditional approach is to use pixel-based classifiers, such as random forests and pixel-based Markov or conditional random field (CRF) models to improve performance by modeling neighbourhood dependencies.

Recently, Convolutional Neural Networks (CNNs) are driving advances in computer vision, such as image classification, recognition, and semantic segmentation. The success of CNNs is attributed to their ability to learn rich feature representations as opposed to hand-designed features used in previous image classification methods. A major contributing factor to their success is the availability of large-scale, publicly available datasets such as ImageNet, Microsoft COCO and Cityscapes that allow CNNs to develop their full potential. For efficient performance of CNNs, the ground truth (GT) training data needs dense pixelwise annotations which requires an enormous amount of human effort. For instance, an image in the Cityscapes dataset takes about 1.5H for dense annotation.

In this work, we explore the possibility of using auxiliary GT, to produce more training data for CNN training. We use the CamVid dataset as an example, which contains video sequences of outdoor driving scenarios. But the methodology can be easily applied to other relevant datasets. We propagate the GT labels from these images to the subsequent images using a simple CRF-based, cue integration framework leading to pseudo ground truth (PGT) training images. It can be expected that the new PGT is noisy and has lower quality compared to the actual GT labeling as a result of automatic label propagation. We train the semantic segmentation network FCNusingthisdata. In this regard,we exploredifferentfactors ofhowthePGT has to be used toenhancethe performance of a CNN.Our baseline is obtained by trainingthe FCNonlyon the GT training imageswhich stands at 49.6%. Fromour experiments, we have found thatadding PGTtothe GT dataand training the FCN helpsin enhancing the accuracyby 2.7%to 52.3%(IoU).Examples of our results are shown in Fig. 1.

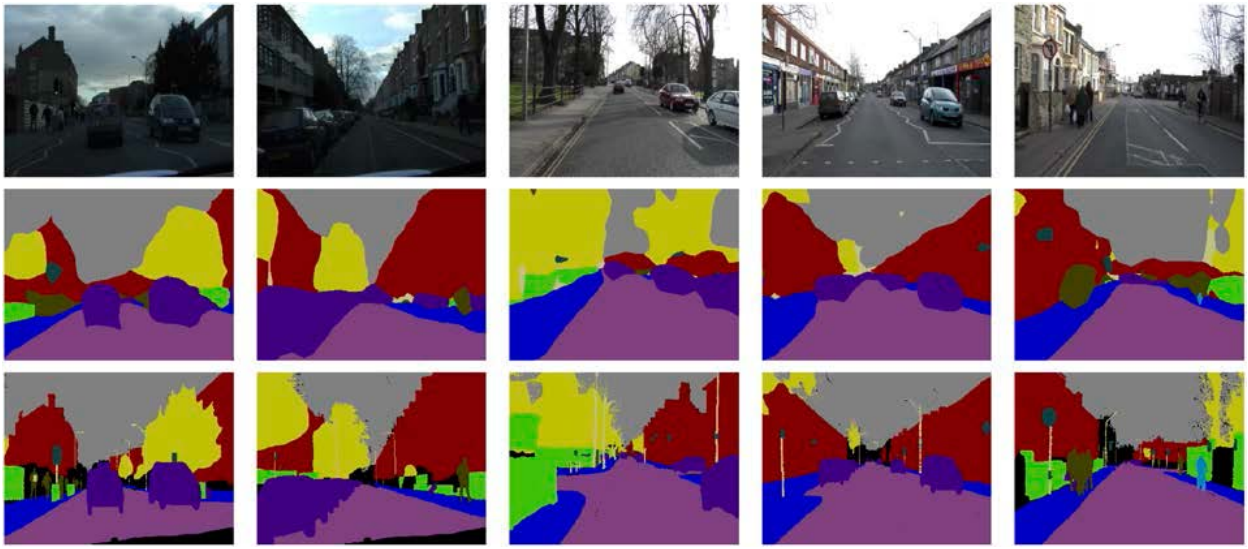

Fig.1.Qualitative performance of our system.First row-Images. Second row-Outputof FCN. Thirdrow-ground truth.