Faculty of Computers and Informatics

Banha University

A Report on:

How Deep Learning Affected Scene Parsing

Prepared for:

Dr. Mohamed Loey

Prepared by:

Mohamed Ahmed Mohamed Afify

Submission Date: 13/10/2017

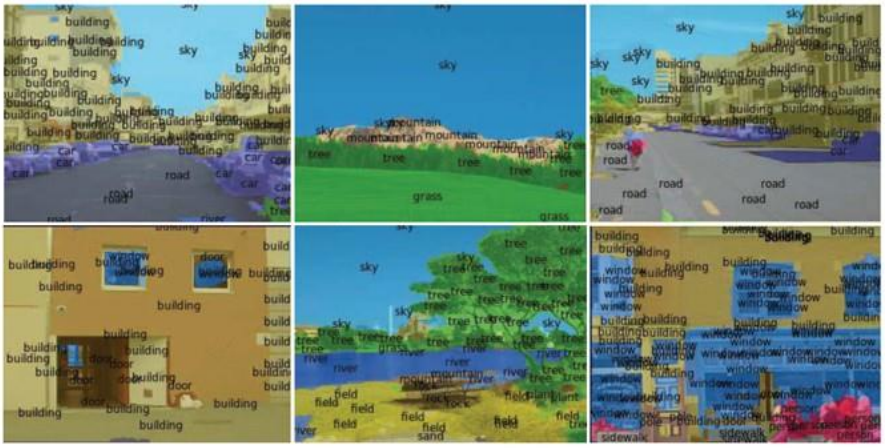
Scene parsing is one of the fundamental problems of computer vision. Scene parsing aims at segmenting images and detecting various object categories within them. Concretely, a scene parser classifies each pixel of an image into one of several predefined object classes. Like traditional computer vision systems such as detectors, a scene parsing model should ideally be robust to changes in illumination and viewpoint, and have an understanding of the spatial dependencies of the object classes in the images.

After a perfect scene parsing, every region and every object is delineated and tagged. One challenge of scene parsing is that it combines the traditional problems of detection, segmentation, and multi-label recognition in a single process. There are two questions of primary importance in the context of scene parsing: how to produce good internal representations of the visual information, and how to use contextual information to ensure the self-consistency of the interpretation.

In recent studies, a variety of strategies based on the framework of Markov Random Fields or Conditional Random Fields have been investigated and well learned. The difference between these methods is features, classifiers, and the graphic models it constructed. Besides, a majority of segmentation methods are introduced by superpixels generated from over-segmentation algorithms on specific images. However, the disadvantage of these methods is feature selection. The features these method selected are often hand-crafted, like HOG, SURF [, and so on, which would be useful in some datasets. However, these features are often low-level, and reveal only a little information, which would make results inconsistent.

To overcome the shortages, deep learning strategies have been introduced for feature extraction. Over past few years, these methods have gained great popularity in machine learning and related fields. This type of methods typically take the original image as input, learning the deep representation, and have found notable success in various tasks such as image classification, object recognition, etc. The success of deep learning methods is basically due to the ability to learn rich-level features within class variance.

The main idea is to use a convolutional network operating on a large input window to produce label hypotheses for each pixel location. The convolutional net is fed with raw image pixels (after band-pass filtering and contrast normalization), and trained in supervised mode from fully-labeled images to produce a category for each pixel location. Convolutional networks are composed of multiple stages each of which contains a filter bank module, a non-linearity, and a spatial pooling module. With end-to-end training, convolutional networks can automatically learn hierarchical feature representations.



Scene Parsing in SIFT flow dataset

	Pixel Acc.	Class Acc.
Liu *et al.* 2009 [31]	74.75%	-
Tighe *et al.* 2010 [44]	76.9%	29.4%
raw multiscale net[1]	67.9%	45.9%
multiscale net + superpixels[1]	71.9%	**50.8%**
multiscale net + cover[1]	72.3%	**50.8%**
multiscale net + cover[2]	**78.5%**	29.6%

Scene Parsing performance in SIFT flow dataset

**** References***

Yu H., Song Y., Ju W., Liu Z. (2016) Scene Parsing with Deep Features and Spatial Structure Learning. In: Chen E., Gong Y., Tie Y. (eds) Advances in Multimedia Information Processing - PCM 2016. PCM 2016. Lecture Notes in Computer Science, vol 9917. Springer, Cham.

Clément Farabet, Camille Couprie, Laurent Najman, Yann Lecun. Learning Hierarchical Features for Scene Labeling. IEEE Transactions on Pattern Analysis and Machine Intelligence, Institute of Electrical and Electronics Engineers, 2013, 35 (8), pp.1915 - 1929.