Deep Learning for Image Analysis in Satellite and Traffic Applications

Abstract

In this thesis, two fundamental problems in computer vision have been addressed by proposing novel approaches which are based on deep learning. First, we address fine-grained object recognition over vehicular images regarding vehicle makes and models. Secondly, we address semantic segmentation over remotely sensed images on a global geographical scale.

To address vehicle make and model recognition (VMMR), a classification architecture based on Convolutioanl Neural Network (CNN) and multi-scale attention windows is developed. The proposed architecture consists of a localizer and a classifier module. First, the localizer module predicts a number of attention windows over each image to capture most representative parts of a vehicle. Then, the classifier module extracts and aggregates visual representations over the predicted attention windows to perform classification over vehicle make and model. We show that VMMR can benefit substantially by capturing most distinctive parts of a vehicle over attention windows with non-identical sizes which provide discriminative visual patterns over multiple scales. Moreover, the proposed architecture leverages spatial transform module to spatially manipulate the input image and to backpropagate the error from the localizer to the classifier. Thus, unlike many other competitive part-based approaches, the proposed localizer is trained to minimize the classification error without requiring expensive part annotations over training samples.

Additionally, a multi-scale patch training methodology is proposed which enables predicting attention windows with desired scales. Moreover, the classifier module is proposed with multiple outputs to allow joint prediction on vehicle make and model. Hence, we formulate a loss function accounting for classification errors over both vehicle make and model. In the end, we evaluate the proposed methodology over two publicly available datasets: Stanford car dataset; Compcar dataset. Our proposed architecture surpasses all prior state-of-the-art methods in both datasets.

In the second part of the thesis, semantic segmentation on satellite images is addressed proposing a CNN encoder-decoder architecture. In contrary to

most of the recent work where the segmentation is studied over samples with similar distributions (i.e. samples are extracted from one geographical area), we develop a scheme which is deployable over a broad range of aerial images with different statistics with different geographical locations. Satellite images captured in different locations by different sensors or even in different time intervals experience variations in their distribution, such variations over a segmentation model input known as covariate shift in machine learning. Accordingly, in this work, we study the generalization capability of the proposed architecture to reduce the performance degradation associated with covariate shift. We show that a class of CNN namely residual network that enables very deep networks (up to 200 layers), if employed as encoder module in the proposed architecture, allows learning visual representations of high semantic level which are more robust to covariate shift. Training such deep encoder over a large amount of satellite images captured at different locations enables learning features of high semantic level which are not specific to a particular image.

Additionally, we propose two domain adaptation techniques to further enhance the segmentation over each specific image. In the first method, performance is improved over each image by fine-tuning the network over a small subset of annotated samples. In the second approach, batch normalization statistics are fine-tuned over each image improving the segmentation without requiring annotations. We evaluate the proposed architectures and domain adaptation methodologies over a homegrown dataset and also two publicly available datasets of satellite images namely ISPRS Vaihingen 2D semantic segmentation contest and INRIA aerial images benchmark. We show that while our network benefits from less complex structure it advances state-of-the-art results on binary segmentation and competes closely with far more complex methods on multi-class segmentation task.

Finally, we propose a similar encoder-decoder CNN to address cloud screening over satellite images such that it can be implemented on the satellite platform. Thus, we investigate experimentally several solutions to make CNN more efficient in terms of resource consumption while preserving its cloud screening accuracy. We show that the proposed architecture can be implemented on the satellite platform while performing with reasonably high accuracy compared with the state-of-the-art approaches.