World Academy of Science, Engineering and Technology
International Journal of Computer and Systems Engineering
Vol:11, No:06, 2017

Object-Scene: Deep Convolutional Representation for Scene Classification

Authors : Yanjun Chen, Chuanping Hu, Jie Shao, Lin Mei, Chongyang Zhang

Abstract : Traditional image classification is based on encoding scheme (e.g. Fisher Vector, Vector of Locally Aggregated Descriptor) with low-level image features (e.g. SIFT, HoG). Compared to these low-level local features, deep convolutional features obtained at the mid-level layer of convolutional neural networks (CNN) have richer information but lack of geometric invariance. For scene classification, there are scattered objects with different size, category, layout, number and so on. It is crucial to find the distinctive objects in scene as well as their co-occurrence relationship. In this paper, we propose a method to take advantage of both deep convolutional features and the traditional encoding scheme while taking object-centric and scene-centric information into consideration. First, to exploit the object-centric and scene-centric information, two CNNs that trained on ImageNet and Places dataset separately are used as the pre-trained models to extract deep convolutional features at multiple scales. This produces dense local activations. By analyzing the performance of different CNNs at multiple scales, it is found that each CNN works better in different scale ranges. A scale-wise CNN adaption is reasonable since objects in scene are at its own specific scale. Second, a fisher kernel is applied to aggregate a global representation at each scale and then to merge into a single vector by using a post-processing method called scale-wise normalization. The essence of Fisher Vector lies on the accumulation of the first and second order differences. Hence, the scale-wise normalization followed by average pooling would balance the influence of each scale since different amount of features are extracted. Third, the Fisher vector representation based on the deep convolutional features is followed by a linear Supported Vector Machine, which is a simple yet efficient way to classify the scene categories. Experimental results show that the scale-specific feature extraction and normalization with CNNs trained on object-centric and scene-centric datasets can boost the results from 74.03% up to 79.43% on MIT Indoor67 when only two scales are used (compared to results at single scale). The result is comparable to state-of-art performance which proves that the representation can be applied to other visual recognition tasks.

Keywords : deep convolutional features, Fisher Vector, multiple scales, scale-specific normalization
Conference Title : ICLR 2017 : International Conference on Learning Robots
Conference Location : Vienna, Austria
Conference Dates : June 21-22, 2017