Pedestrian Detection aided by Deep Learning Semantic Tasks

Yonglong Tian¹, Ping Luo³,¹, Xiaogang Wang²,³, Xiaou Tang¹,³
¹Department of Information Engineering, The Chinese University of Hong Kong. ²Department of Electronic Engineering, The Chinese University of Hong Kong. ³Shenzhen Key Lab of CVPR, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China.

State-of-the-art methods for pedestrian detection can be generally grouped into two categories, the models based on hand-crafted features and deep models. The first is to extract Haar [6], HOG [1], or HOG-LBP [7] features and train SVM [1] or AdaBoost classifiers [2], where the two stages cannot be jointly optimized to improve performance. In the second category, deep neural networks achieved promising results [3, 4, 5], owing to their capacity to learn discriminative features from raw pixels. While previous treated pedestrian detection as a single binary classification task, which are not able to capture rich pedestrian variations, this paper proposed a novel task-assistant CNN (TA-CNN) to jointly optimize detection with auxiliary semantic tasks, including pedestrian attributes and scene attributes. Fig. 1 is an illustration for how this idea works. If only a single detector is used to classify all the positive and negative samples in Fig. 1 (a), it is difficult to handle complex pedestrian variations. Therefore, the mixture models of multiple views were developed in Fig. 1 (b), i.e., pedestrian images in different views are handled by different detectors. If views are treated as one type of semantic tasks, learning pedestrian representation by multiple attributes with deep models actually extends this idea extremely.

All attributes are summarized in Fig. 2. Given a pedestrian dataset P, for example Caltech, we manually label the positive patches with nine pedestrian attributes. For background, we transfer hard negative patches with attribute information from three public scene segmentation datasets to P, including CamVid (B⁷), Stanford Background (B⁵), and LM+SUN (B⁹). As shown in Fig. 2, pedestrian attributes only present in P, shared attributes present in all B's, and the unshared attributes present in one of them.

We construct a training set D by combing patches cropped from both P and B's. Let D = \{(x_n, y_n)\}_{n=1}^N be a set of image patches and their labels, where each y_n = (y_n, o'_n, o''_n, o''''_n) is a four-tuple. Specifically, y_n denotes a binary label, indicating whether an image patch is pedestrian or not.

The performances show clear increasing patterns when gradually adding attributes. For background, we transfer hard negative patches with attribute information from three public scene segmentation datasets to P, including CamVid (B⁷), Stanford Background (B⁵), and LM+SUN (B⁹). As shown in Fig. 2, pedestrian attributes only present in P, shared attributes present in all B's, and the unshared attributes present in one of them.

We construct a training set D by combing patches cropped from both P and B's. Let D = \{(x_n, y_n)\}_{n=1}^N be a set of image patches and their labels, where each y_n = (y_n, o'_n, o''_n, o''''_n) is a four-tuple. Specifically, y_n denotes a binary label, indicating whether an image patch is pedestrian or not.

The performances show clear increasing patterns when gradually adding attributes. For background, we transfer hard negative patches with attribute information from three public scene segmentation datasets to P, including CamVid (B⁷), Stanford Background (B⁵), and LM+SUN (B⁹). As shown in Fig. 2, pedestrian attributes only present in P, shared attributes present in all B's, and the unshared attributes present in one of them.

We construct a training set D by combing patches cropped from both P and B's. Let D = \{(x_n, y_n)\}_{n=1}^N be a set of image patches and their labels, where each y_n = (y_n, o'_n, o''_n, o''''_n) is a four-tuple. Specifically, y_n denotes a binary label, indicating whether an image patch is pedestrian or not.

The performances show clear increasing patterns when gradually adding attributes. For background, we transfer hard negative patches with attribute information from three public scene segmentation datasets to P, including CamVid (B⁷), Stanford Background (B⁵), and LM+SUN (B⁹). As shown in Fig. 2, pedestrian attributes only present in P, shared attributes present in all B's, and the unshared attributes present in one of them.