# Generic Object Detection with Dense Neural Patterns and Regionlets

Will Y. Zou[1]
http://ai.stanford.edu/~wzou

Xiaoyu Wang[2]
http://www.xiaoyumu.com

Miao Sun[3]
http://vision.ece.missouri.edu/~miao

Yuanqing Lin[2]
http://www.linyq.com

[1] Stanford University
Stanford, CA, 94305

[2] NEC Laboratories America
Cupertino, CA, 95014

[3] University of Missouri
Columbia, MO, 65201

This paper addresses the challenge of establishing a bridge between deep convolutional neural networks and conventional object detection frameworks for accurate and efficient generic object detection. We introduce Dense Neural Patterns, short for DNPs, which are dense local features derived from discriminatively trained deep convolutional neural networks. DNPs can be easily plugged into conventional detection frameworks in the same way as other dense local features(like HOG or LBP). The effectiveness of the proposed approach is demonstrated with the Regionlets object detection framework. It is the first approach efficiently applying deep convolutional features for conventional object detection models.

Detecting generic objects in high-resolution images is one of the most valuable pattern recognition tasks, useful for large-scale image labeling, scene understanding, action recognition, self-driving vehicles and robotics. At the same time, accurate detection is a highly challenging task due to cluttered backgrounds, occlusions, and perspective changes. Predominant approaches use deformable template matching with hand-designed features. However, these methods are not flexible when dealing with variable aspect ratios. Wang *et al*. recently proposed a radically different approach, named *Regionlets*, for generic object detection [4]. It extends classic cascaded boosting classifiers with a two-layer feature extraction hierarchy , and is dedicatedly designed for region based object detection. Despite the success of these sophisticated detection methods, the features employed in these frameworks are still traditional features based on low-level cues such as histogram of oriented gradients(HOG), local binary patterns(LBP) or covariance [3] built on image gradients.

With the success in large scale image classification [1], object detection using a deep convolutional neural network also shows promising performance [2]. The dramatic improvements from the application of deep neural networks are believed to be attributable to their capability to learn hierarchically more complex features from large data-sets. Despite their excellent performance, the application of deep CNNs has been centered around image classification, which is computationally expensive when transferred to perform object detection. Furthermore, their formulation does not take advantage of venerable and successful object detection frameworks such as DPM or *Regionlets* which are powerful designs for modeling object deformation, sub-categories and multiple aspect ratios.

These observations motivate us to propose an approach to efficiently incorporate a deep neural network into conventional object detection frameworks. To that end, we introduce the *Dense Neural Pattern* (DNP), a local feature densely extracted from an image with an arbitrary resolution using a deep convolutional neural network trained with image classification datasets. The DNPs not only encode high-level features learned from a large image data-set, but are also local and flexible like other dense local features (like HOG or LBP). It is easy to integrate DNPs into the conventional detection frameworks. More specifically, the receptive field location of a neuron in a deep CNN can be back-tracked to exact coordinates in the image. This implies that spatial information of neural activations is preserved. Activations from the same receptive field but different feature maps can be concatenated to form a feature vector for that receptive field. These feature vectors can be extracted from any convolutional layers before the fully connected layers. Because spatial locations of receptive fields are mixed in fully connected layers, neuron activations from fully connected layers do not encode spatial information. The convolutional layers naturally produce multiple feature vectors that are evenly distributed in the evaluated image crop ( a $224 \times 224$ crop for example). To obtain dense features for the whole image which may be significantly larger than the network input, we resort to “network-convolution” which shifts the crop location and forward-propagate the neural network until features at all desired locations in the image are extracted. As the result, for a typical PASCAL VOC image, we only need to run the neural network several times to produce DNPs for the whole image depending on the required feature stride, promising low computational cost for feature extraction. To adapt our features for the *Regionlets* framework, we build normalized histograms of DNPs inside each sub-region of arbitrary resolution within the detection window and add these histograms to the feature pool for the boosting learning process. DNPs can also be easily combined with traditional features in the *Regionlets* framework.

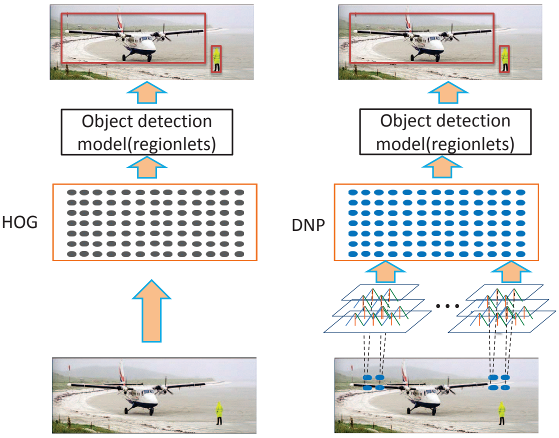
Object detection model(regionlets)
HOG
Object detection model(regionlets)
DNP
...

Figure 1: Deep Neural Patterns (DNP) for object detection

Our experiments show that the proposed DNPs from the top convolutional layers in deep CNN are very effective and also complementary to traditional features. It achieved 46.1% mean average precision on the PASCAL VOC 2007 dataset, and 44.1% on the PASCAL VOC 2010 dataset, which dramatically improves the original Regionlets approach without DNPs. Combing DNPs and hand-crafted low-level features produces compelling object detection performance. On the contrary, putting together lower layer features and higher layer features from the convolutional neural network does not improve the detection performance. It indicates that these features are correlated. While traditional hand-crafted features are not supervised learned which largely complement the neural network features.

The major contribution of the paper is two-fold: 1) We propose a method to incorporate a discriminatively-trained deep neural network into a generic object detection framework. This approach is very effective and efficient. 2) We apply the proposed method to the *Regionlets* object detection framework and achieved competitive and state-of-the-art performance on the PASCAL VOC datasets.

[1] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012.

[2] Pierre Sermanet, David Eigen, Xiang Zhang, Michaël Mathieu, Rob Fergus, and Yann LeCun. Overfeat: Integrated recognition, localization and detection using convolutional networks. *CoRR*, abs/1312.6229, 2013.

[3] Oncel Tuzel, Fatih Porikli, and Peter Meer. Pedestrian detection via classification on riemannian manifolds. *T-PAMI*, 2008.

[4] Xiaoyu Wang, Ming Yang, Shenghuo Zhu, and Yuanqing Lin. Regionlets for generic object detection. In *ICCV*, 2013.