Scale-Transferrable Object Detection

Xuewen Yang

August 19 2018

1. Scale-Transferrable Detection Network

In this section,the author first introduce the base network which is our feature extraction network component. They use DenseNet [1] as our base network. In each dense block of DenseNet, for each layer, its feature map is used as inputs for all subsequent layers. The output of the last layer of the dense block has highest number of channels and is suitable as input for our scale-transfer layer which expands the width and height of the feature map by compressing the number of channels. Then the author describe the scale-transfer module that produces feature maps at different scales. Next,they describe the entire object detection/location prediction network architecture and the network training details.

1.1. Base Network : DenseNet

The author use DenseNet-169 [1] as the base network for feature extraction and do pre-training on the ILSVRC CLSLOC dataset [2]. DenseNet is a network with deep supervision. In each dense block of DenseNet, the output of each layer contains the output of all previous layers, and thus incorporates low-level and high-level features of the input image, which is suitable for object detection. Inspired by DSOD, the author replace the input layers (7×7 convolution layer, stride = 2 followed by a 3×3 max pooling layer, stride = 2) into three 3×3 convolution layers and one 2×2 mean pooling layer. The stride of the first convolution layer is 2 and the others are 1. The output channels for all three convolution layers are 64. They call these layers stem block.Experiments show that this simple substitution can significantly improve the accuracy of object detection. One explanation could be that the input layers in the original DenseNet-169 have lost much information due to two consecutive down sampling. This will impair the performance of object detection, especially for small objects.

1.2. High Efficiency ScaleTransfer Module

Scale problem lies in the heart of object detection. Combining predictions from multiple feature maps with different resolutions are beneficial for detecting multi-scale objects. However, as shown in Figure 1, in the last dense block of DenseNet, all outputs of layers have the same width and height, except for the number of channels. For example, when the input image is 300×300, the last dense block dimension of DenseNet-169 is 9×9. A simple approach is to predict directly using high-resolution feature maps of low layers, similar to SSD [22]. However, the low-level feature map lacks semantic information about objects, which may cause low performance on object detection. In order to obtain different resolution feature maps with strong semantic information, the author develop a module named scale-transfer module. Scale-transfer module is very efficient and can be directly embedded into the dense block in DenseNet. To get strong semantic feature maps, they make use of the network structure of DenseNet to transfer the lowlevel features directly to the top of the network through the concat operation. The feature map at the top of the network has both low-level detail information and high-level semantic information so as to improve the performance of object localization and classification.

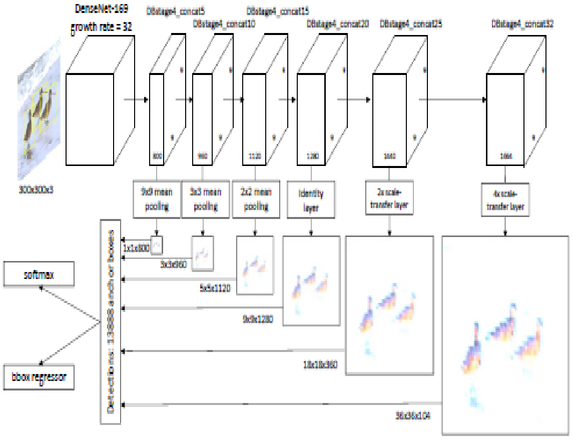

Figure 1. Importance of context for object recognition. Without the context (face), it is hard to recognize the black curve in the middle area as a nose.

They get feature maps of different scales from the last dense block of DenseNet. In scale-transfer module, they use the mean pooling layer to obtain low-resolution feature maps. For high-resolution feature maps, they use a technique called scale-transfer layer.

References

[1] G. Huang, Z. Liu, L. V. D. Maaten, and K. Q. Weinberger. Densely connected convolutional networks. In *CVPR*, 2017. 1

[2] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, and M. Bernstein. ImageNet large scale visual recognition challenge. *IJCV*, 2015. 1