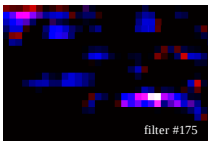
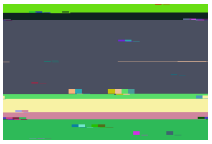
Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun


filter #175



be of any sizes. This not only allows arbitrary aspect ratios, but also allows arbitrary scales. We can resize the input image to any scale (*e.g.*, min(*w; h*)=180, 224, ...) and apply the same deep network. When the input image is at different scales, the network (with

model	$conv_1$	$conv_2$	$conv_3$	$conv_4$	$conv_5$	$conv_6$	conv

method	VOC 2007	Caltech101
VQ [15]^{y}		

the feature of this window. If the pre-defined scales are dense enough and the window is approximately square, our method is roughly equivalent to resizing the window to 224 224 and then extracting features from it. Nevertheless, our method only requires computing the feature maps once (at each scale) from the entire image, regardless of the number of candidate windows.

We also fine-tune our pre-trained network, following [7]. Since our features are pooled from the $conv_5$ feature maps from windows of any sizes, for simplicity we only fine-tune the fully-connected layers. In this case, the data layer accepts the fixed-length pooled features after $conv_5$, and the $fc_{6;7}$ layers and a new 21-way (one extra negative category) fc_8 layer follow. The fc_8

fdistribtino

Figure 6: Example detection results of “SPP-net $ftfc_7$ bb” on the Pascal VOC 2007 testing set (59.2% mAP). All windows with scores >

References

[1] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard,