Supplementary Material
STNet: Selective Tuning of Convolutional Networks for Object Localization

Mahdi Biparva, John Tsotsos
Department of Electrical Engineering and Computer Science
York University, Toronto, Canada
{mhdbprv,tsotsos}@cse.yorku.ca

1 Implementation Details of STNet

In this section, we provide the implementation details of STNet for the three Convolutional Neural Network (ConvNet) architectures: AlexNet, VGGNet, and GoogleNet. We discuss the realization of the TD selective process for different types of layers. We discuss the experimental results in Sec. 2.

1.1 STNet Implementation for Different Types of Layers

We provide details on the implementation of STNet for various types of layers encountered in the three ConvNet architectures.

Max Pooling Layer: The Max Pooling layer could be regarded as a BU Winner-Take-All (WTA) computation where the maximum node activity is selected. Since the gating flow of the BU information is defined in a hard manner, it would be against the inherent nature of the learned representation to select other nodes but the maximum one in the TD processing stream. Therefore, we decide to stick to the the maximum node selection regime and propagate the top gating node activity to gating node correspondence of the maximum node within it's receptive field (RF).

Average Pooling Layer: This type is only encountered in GoogleNet at where the convolutional lower part of the network meets the fully-connected (FC) upper part. In other words, the last spatially-ordered hidden layer of the network is squeezed into the hidden vector of the first FC layer using an average pooling layer. We experimentally evaluated what would be the best way of treating the average pooling acting as the link between the lower body and upper body of the network. We decided to choose WTA as the mechanism to select the gating node at the layer below to which the top gating node activity will be propagated. It should be noted that in both AlexNet and VGGNet, we defined the concept of the bridge layer at which the lower convolutional body is connected to the upper FC body of the network. However, in GoogleNet, the average pooling layer instead of a FC layer is utilized to connect the lower to the upper. Therefore, GoogleNet does not benefit from the additional level of selection specifically defined for the bridge layer. We experimentally observed that the average pooling layer in GoogleNet is very sensitive to changes in the selection process of the TD processing stream.

ReLU Layer: Since ReLU layers only cut off all the negative activities of the BU processing stream, TD processing stream simply bypass the layer and copy the gating node activities of the top layer to the layer below.

Convolutional/Fully-connected Layer: These two types are very much detailed in the main paper. Three stages of the attentive selection process are defined to deal with these two layers. TD processing is implicitly applied to the parts of the visual representation where feature transformation is parametrized such as convolutional layers. It is noteworthy to indicate that in GoogleNet, 1x1 convolutional layers are very dominant throughout the visual hierarchy. Based on the results obtained in the cross-validation stage, we decided to treat such layers the same as we do the FC layers. The sole discrepancy is that at the 2nd stage of the attention selection, all the winner nodes marked by the 1st stage are

selected instead of utilizing the SI selection mode in the FC layers. This implies despite 1x1 convolutional layers do not strive for spatial correlation encoding among their receptive fields, maximal selection of of the nodes in their flat receptive fields provide a significantly better localization result.

Local Response Normalization (LRN) Layer: This layer simply normalize the information flow of the BU processing from the layer below to the top layer over some RF. Therefore, it is straightforward to skip LRN layers in the TD processing by transferring the top gating node activities to the layer below.

1.2 Generation of Class Hypothesis Maps

We provide details on the procedure proposed to create Class Hypothesis (CH) maps. Following a similar experimental setup to the localization task, the attention map is extracted from a gating layer. Pixel values in the attention map are happened to be very sparsely non-zero. We create CH map with the equal number of pixels to the attention map filled with zero values. Then, we propose an updating procedure that is iteratively applied to all the non-zero pixels of the attention map as follows. On the CH map, we increment the values of all the pixels falling within the square window centered at the pixel coordinated corresponding to a non-zero pixel on the attention map. The size of the window is set to the accumulated RF size of the particular layer the attention map is extracted from. Once all the non-zero pixels on the attention map are visited, the CH map is filtered using a smoothing Gaussian kernel with the standard deviation $\sigma = 6$. Finally, the CH map are visualized as a heat map with the red color representing the maximum value and blue the minimum. In what follows, we provide further details on the modified configurations of STNet for two CH visualization experiments: 1- Context Interference, 2- Correlated Accompanying Objects.

Context Interference: In this experiment, we attempt to highlight the role of the context inference in the localization performance of the TD processing according to the learned representation. The second stage of the selection process in STNet is proposed to tackle with this level of contextual noise. Therefore, to show it's efficiency to address the problem, we deactivated the second stage throughout the TD structure. Furthermore, the first stage on the FC layers are modified to implement the WTA mechanism. Consequently, at each FC layer in this regime, there is only one gating node active and the rest remain inactive. This is seen to emphasize the role of the second stage in dealing with the context inference problem.

Correlated Accompanying Objects: We keep the modified version of the first stage for FC layers in this experimental setup, while the second stage on the convolutional layers are taken back into place. The goal is to show that the most confident high-level node at each FC layer will end up localizing a correlated object very frequently accompanying the ground truth category.

2 Experimental Results

In this section, we provide the high resolution qualitative results of successful bounding box predictions, unsuccessful bounding box predictions, CH visualization using the original STNet, CH visualization for the Context Inference experiment, and CH visualization for the Correlated Accompanying Objects experiment in Fig. 1, 2, 4, and 5 respectively. Following a naming convention, ST-VGGNet, for instance, is referred to as STNet with the utilization of VGGNet in the BU processing stream.

Figure 1: Illustration of STNet localization performance for both VGGNet and GoogleNet. The top, middle, and bottom row of each section contains images demonstrating the ground truth bounding boxes, bounding box predictions of ST-VGGNet, and ST-GoogleNet respectively.

Figure 2: Unsuccessful localization cases based on STNet bounding box predictions are demonstrated. Multi-Instance and Correlated Accompanying Object scenarios are the two main sources of STNet unsuccessful localization. Each section contains image rows for ground truth, ST-VGGNet and ST-GoogleNet bounding boxes from top to bottom.

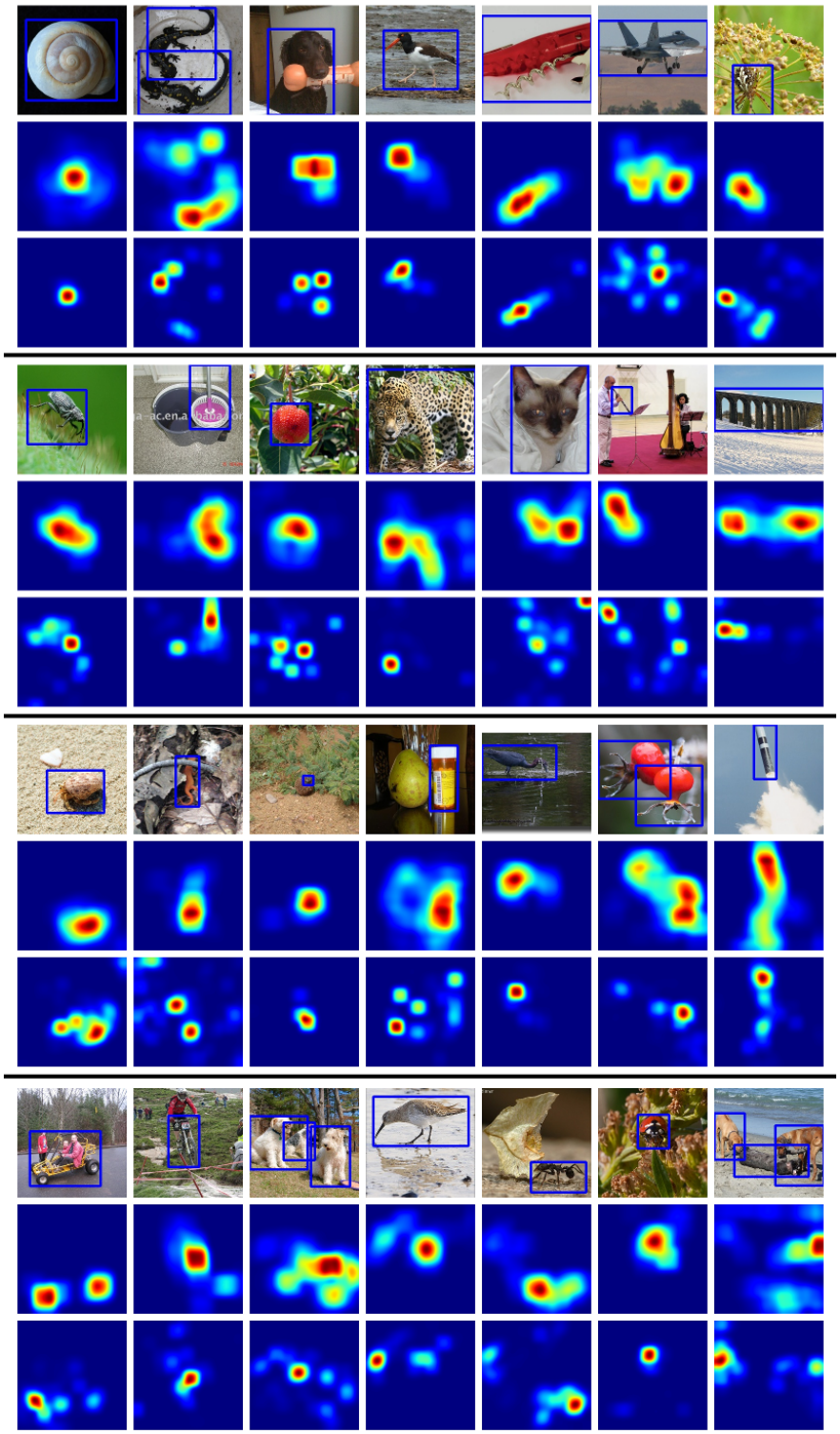

Figure 3: Class Hypothesis Visualization using STNet. In each section, the top row contains RGB images depicting ground truth bounding boxes, and the middle and bottom row contains the CH maps from ST-VGGNet and ST-GoogleNet respectively.

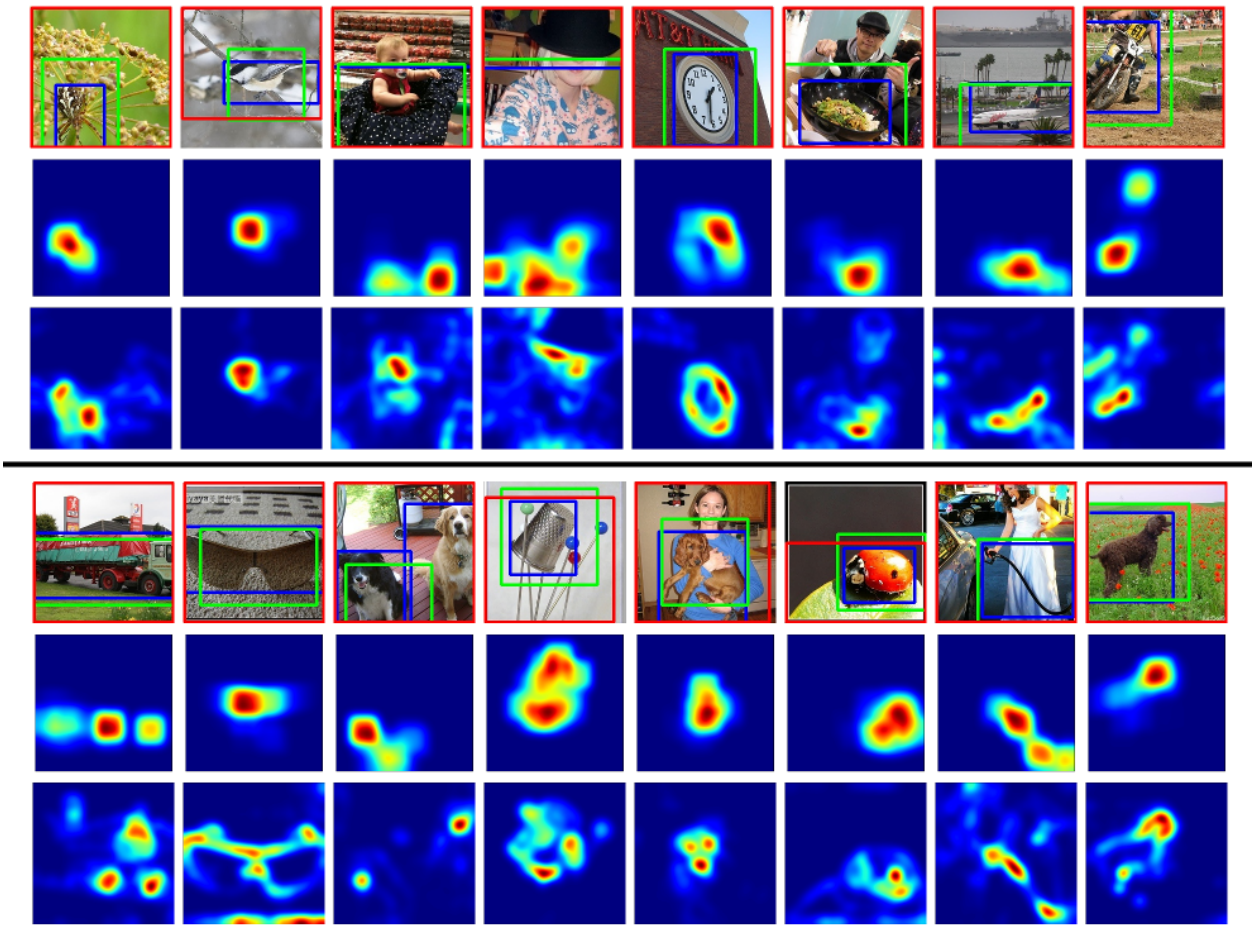

Figure 4: The effect of the context inference imposed by the learned representation is illustrated in the CH maps given in the bottom rows of each section. The middle row contains the CH maps from the original proposal of ST-VGGNet. The top row provides RGB images with the color-coded bounding boxes. Blue boxes are taken from the ground truth. Green and red boxes represent original and partially-deactivated ST-VGGNet predictions.

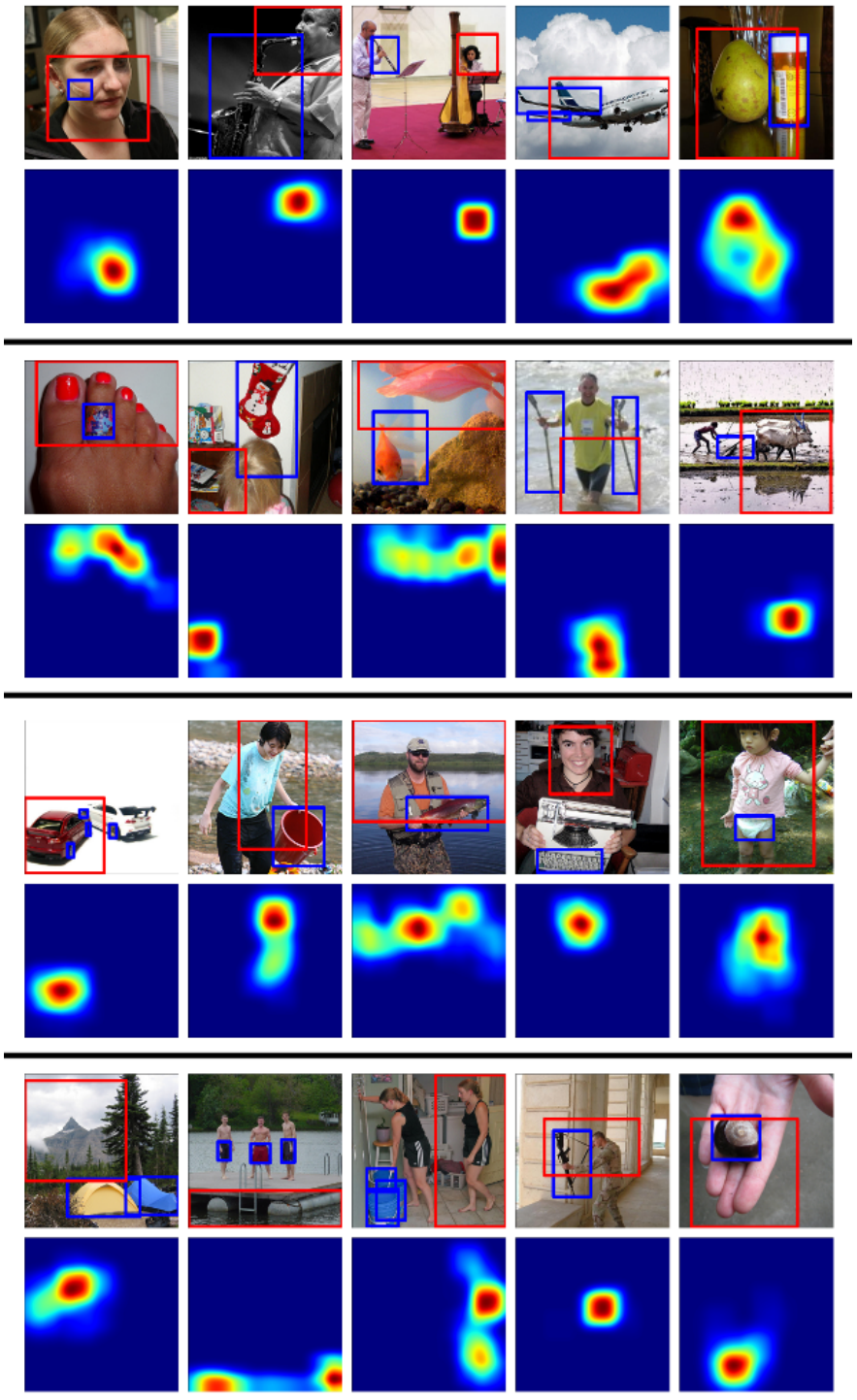

Figure 5: Correlated accompanying objects prioritize localization of regions outside the ground truth according to the learned representation. In each section, the top row contains RGB images with the ground truth boxes (blue). The red boxes are proposed by the modified ST-VGGNet.