Training convolutional neural networks with image patches for object localization

S. Orhan, Y. Bastanlar

Recently, convolutional neural networks (CNNs) have shown great performance in different problems of computer vision including object detection and localization. In this work, we propose a novel training approach for CNNs to localize some animal species whose bodies have distinctive pattern, such as leopards and zebras. To learn characteristic patterns, small patches which are taken from different body parts of animals are used to train models. To find object location, in a test image, all locations are visited in a sliding window fashion. Crops are fed into trained CNN and their classification scores are combined into a heat map. Later on, heat maps are converted to bounding box estimates for varying confidence scores. The localization performance of our patch-based training approach is compared with Faster R-CNN - a state-of-the-art CNN-based object detection and localization method. Experiment results reveal that the patch-based training outperforms Faster R-CNN especially for classes with distinctive patterns. We also showed that the patch-based approach can be used in combination with Faster R-CNN to improve its localization performance.

Introduction: There exist many object localization approaches using CNNs. In an earlier approach [1], objects are searched in a sliding window fashion, where a separate regression head runs to estimate the bounding box of each detected object. To shorten the localization process, more recent approaches perform object classification only on candidate regions. For instance, in Faster R-CNN [2], region proposal step is implemented as a neural network after the last convolution layer, called Region Proposal Network (RPN), which reduced the region proposal time significantly. You Only Look Once (YOLO) [3] uses a single CNN for both detection and classification of objects. Very recently, YOLOv2 reached the detection accuracy of Faster R-CNN while processing real-time [4].

Current object localization methods search the objects as a whole. We realized that some objects' peculiar patterns may constitute an important cue. To exploit this cue, instead of training and searching for a complete object (or a large part of it), we perform training with small patches. Our reasoning encompasses all objects with distinctive patterns. As a case study, we work on the problem of finding certain animals in a set of collected images.

We train a deep residual network [5] for the proposed patch-based approach. To localize the objects in a test image, all locations are visited and crops are fed into CNN to get their classification scores. A heat map, generated by these classification scores, is later converted to bounding box estimates by a series of morphological operations.

The localization performance of our approach was compared with Faster R-CNN. According to the experiment results, patch-based training exhibits better performance than Faster R-CNN especially for objects with distinctive patterns. We also showed that the patch-based approach can be used in combination with Faster R-CNN to improve its localization performance.

Our Method: We train a deep residual network [5] (a 50-layer ResNet) to detect multiple object classes. The classes we included are leopard, zebra, elephant and bear. Elephant and bear do not have very distinctive patterns as leopard and zebra do. They are intentionally chosen to analyze if this leads to a performance decrease. As mentioned earlier, we trained the network with patches of objects. Approximately 1000 patches are used per class. Patch size is 64x64 pixels (see examples in Fig.1). Background patches (for training) are taken from the same images but from the regions that do not contain any object parts.

To find the correct patches in a test image, all locations are visited in a sliding window fashion with 64x64 pixel patches (stride size is 32 pixels). Crops are fed into a CNN which was trained with patches. For each patch, probability of belonging to one of the trained classes is saved and a heat map is generated for each class based on these results. An example heat map for leopard class can be seen in Fig. 2b. Red color which has the highest score means that location has been classified as the target animal (with probability=1.0) for all encompassing windows. Blue color (corresponding to the lowest score) means all the sliding windows including that image location have zero probability for the target class. Maximum probability value of each 32x32 pixel area can be 4 due to intersection of four windows. In the rest of our computations, [0-4] range is normalized to [0-1]. Some example heat maps can be seen in Fig. 3. As can be observed, almost all parts of objects are covered with high probability values.

To draw the bounding box of an object, heat map is converted to a binary image according to a given score threshold. Morphological operations are applied to eliminate very small responses and connect close parts. Then, connected component analysis algorithm is used to find object contours. Process steps can be seen in Fig. 2c to Fig. 2e.

Fig. 1 *Example patches from the training set. From left to right, leopard (two of them), zebra, elephant and bear classes.*

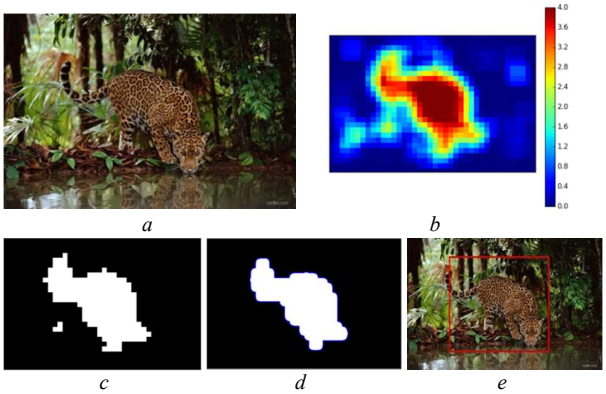


Fig. 2 *Prediction of bounding boxes. a) an input image, b) its heat map for leopard class, c)binary image (heat map after applying threshold), d) result after morphological operations, e) predicted bounding box.*

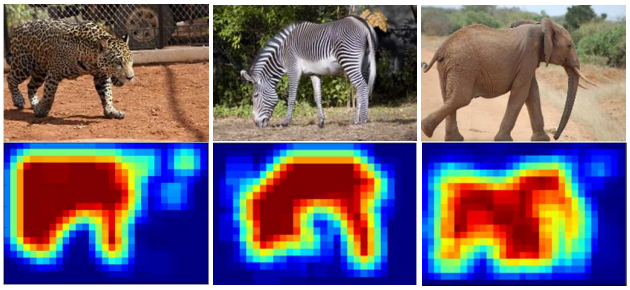


Fig. 3 *Input images are shown at the first row, generated heat maps by Patch-based approach are shown at the second row.*

Evaluation Metrics: To evaluate the performance of object detection algorithms, generally precision-recall curves are used in literature. Detected object box is classified as a true-positive if it is correctly labelled and its Intersection over Union (IoU) rate (1) with a groundtruth box is higher than a threshold (generally taken as 0.5).

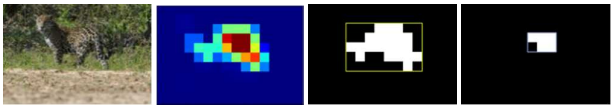
$$IoU = \frac{Box_{detected} \cap Box_{groundtruth}}{Box_{detected} \cup Box_{groundtruth}} \quad (1)$$

Unlike common object detection algorithms, detected box size in patch-based approach change in proportional to threshold values. At low threshold values, it has bigger boxes; at high threshold values, it has smaller boxes. Fig. 4 depicts the shrinking of bounding boxes when threshold increases. A small bounding box obtained with a high threshold precisely locates an object. However, according to IoU criterion, we get a false-positive. This causes precision and recall decrease simultaneously for high thresholds and makes it impossible to evaluate our patch-based method. More suitable for us, we use area-precision (P_{AR}) and area-recall (R_{AR}) metrics (2) proposed by [6].

$$P_{AR}(G,D) = \frac{\sum_j Area(G \cap D_j)}{\sum_j Area(D_j)}, \; R_{AR}(G,D) = \frac{\sum_j Area(G \cap D_j)}{Area(G)} \quad (2)$$

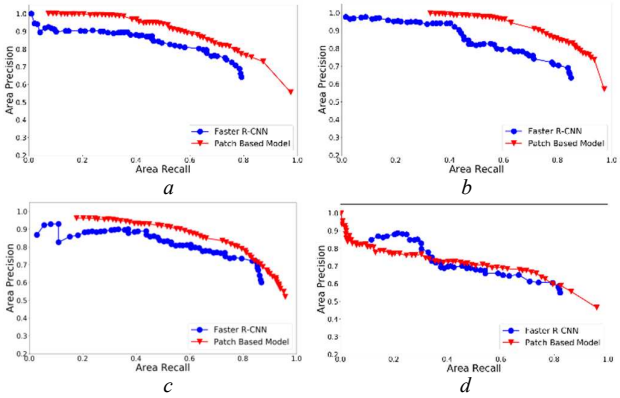
where G is a ground truth rectangle, where D is a list of detected rectangles, j = 1, .., $|D|$. P_{AR} considers how much of the area of the detected windows is covered by groundtruth, R_{AR} considers how much of the groundtruth area is covered by detected windows.

Experiment Results: Faster R-CNN is trained with 446 bear, 351 elephant, 400 leopard and 450 zebra images obtained from ImageNet (www.image-net.org). Patch-based training requires much smaller dataset, namely 50 images (1000 patches) per class. Test set consists of 62 images per class, where each image contains a single object. Area-precision vs. area-recall curves on the test set are shown in Fig. 5. We observe that the patch-based training significantly outperforms Faster R-CNN for leopard and zebra classes. For elephant class it is superior as well. Regarding the bear class (Fig. 5d), patch-based method outperforms Faster R-CNN only when area-recall is greater than 0.4. This performance decrease is mostly because the bear patches can be confused with background objects such as grass. A falsely predicted bear example of patch-based approach is shown in Fig. 6. Some patches, especially at the rear part of the animal, are confused with background. Some other patches are predicted as elephant. This causes to get low scores (probability) for bear at actual bear location.



a *b* *c* *d*

Fig. 4 *a) Input image, b) Generated heat map, c) Predicted bounding box at threshold = 0.3, d) Predicted bounding box at threshold = 0.9.*



a *b* *c* *d*

Fig. 5 *Area-precision vs. area-recall curves for Faster R-CNN and patch-based approach for a) leopard, b) zebra, c) elephant and d) bear.*

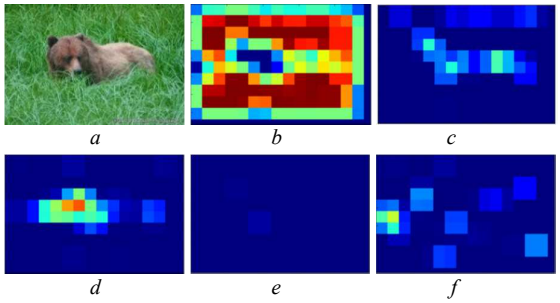
In another experiment, we investigated if the patch-based approach results can be used to improve the localization performance of Faster R-CNN. In this 'combined model', probability of a Faster R-CNN box is increased using (3) if it overlaps patch-based method's detection(s).

$$P_{FasterRCNN_j} = \frac{P_{FasterRCNN_j} + PatchCont_j}{2} \quad (3)$$

Here, $P_{FasterRCNN}$ is the list of Faster R-CNN predicted box scores. $PatchCont_j$ represents the contribution to the j^{th} Faster R-CNN box from n boxes obtained with patch-based method and it is computed by

$$PatchCont_j = \frac{\sum_{i=1}^{n} Area(PatchBasedBox_i \cap FasterRCNN_j)}{\sum_{i=1}^{n} Area(PatchBasedBox_i)} \quad (4)$$

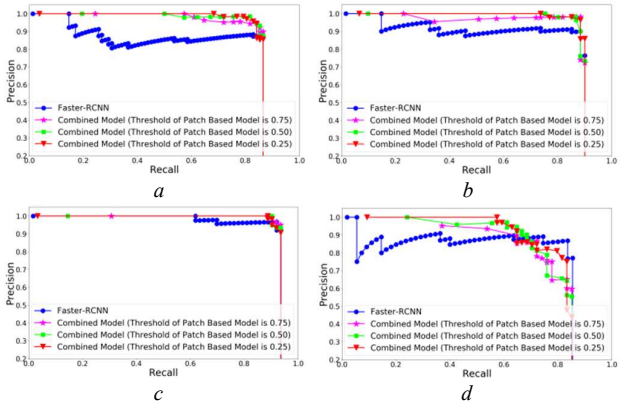
In this experiment, performance evaluation can be done by precision-recall curves since Faster R-CNN box estimates are updated with the help of patch-based approach. Tests were applied at three different confidence levels (0.25, 0.50 and 0.75 out of 1.0) of patch-based model. Results (Fig. 7) show that for leopard, zebra and elephant classes, Combined Model outperforms standard Faster R-CNN. For bear class (Fig.7d), regarding the area under the curves, only Combined Model @0.25 is clearly superior to Faster R-CNN.



a *b* *c*

d *e* *f*

Fig. 6 *a) Input image. Heat map results of b) background, c) bear, d) elephant, e) leopard and f) zebra classes.*



a *b*

c *d*

Fig. 7 *Precision-recall curves of Faster R-CNN and Combined Model for a) leopard, b) zebra, c) elephant and d) bear classes.*

Conclusions: We showed that the proposed patch-based training approach has a high localization performance especially for objects with distinctive patterns. For other classes (example in this study is bear), performance decreases. However, even for such classes, used in combination, patch-based method is able to increase the performance of Faster R-CNN, a state-of-the-art method. The improvement is due to the fact that the patch-based method locates objects better when the learned patterns are visible. Therefore, one can expect similar improvement when combined with other state-of-the-art methods as well.

Another advantage of our approach is that significantly less number of images is adequate for training (e.g. 50 zebra images instead of 450). This may become critical when the available dataset has a limited size.

Acknowledgments: This work was supported by TUBITAK (Grant no. 115E918).

S. Orhan, Y. Bastanlar (*Department of Computer Engineering, Izmir Institute of Technology, 35430, Izmir, Turkey*)

E-mail: yalinbastanlar@iyte.edu.tr

References

1 Sermanet, P., Eigen, D., Zhang, X., Mathieu, M., Fergus, R., LeCun, Y.: 'Overfeat: Integrated recognition, localization and detection using convolutional networks', 2013, http://arxiv.org/abs/1312.6229

2 Ren, S., He, K., Girshick, R., Sun, J.: 'Faster R-CNN: Towards real-time object detection with region proposal networks'. In Neural Information Processing Systems (NIPS) 2015, pp. 91–99.

3 Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: 'You only look once: Unified, realtime object detection', In IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2016, pp. 779–788.

4 Redmon, J., Farhadi, A.: 'YOLO9000: Better, Faster, Stronger', 2016 arXiv:1612.08242

5 He, K., Zhang, X., Ren, S., Sun, J.: 'Deep residual learning for image recognition', In IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2016, pp. 770–778.

6 Wolf, C., Jolion, J.M.: 'Object count/area graphs for the evaluation of object detection and segmentation algorithms', *International Journal of Document Analysis and Recognition* (IJDAR), 2006, 8(4), pp. 280–296.