Training CNNs with image patches for object localisation

S. Orhan and Y. Bastanlar✉

Recently, convolutional neural networks (CNNs) have shown great performance in different problems of computer vision including object detection and localisation. A novel training approach is proposed for CNNs to localise some animal species whose bodies have distinctive patterns 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 localisation performance of the patch-based training approach is compared with Faster R-CNN – a state-of-the-art CNN-based object detection and localisation method. Experimental results reveal that the patch-based training outperforms Faster R-CNN, especially for classes with distinctive patterns.

Introduction: There exist many object localisation approaches using convolutional neural networks (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 localisation 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, which reduced the region proposal time significantly. More recently, YOLO [3] uses a single CNN for detection and classification.

Current object localisation methods search the objects as a whole. We realised 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 [4] for the proposed patch-based approach. To localise the objects in a test image, all locations are visited and crops are fed into CNN to obtain their classification scores. A heat map, generated by these classification scores, is later converted to bounding box estimates.

The localisation performance of our approach was compared with Faster R-CNN. According to the experiment results, patch-based training outperforms 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 localisation performance.

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

Our method: We train a deep residual network [4] (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 analyse 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 64×64 px^2 (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 64×64 px patches (stride size is 32 px). Crops are fed into a CNN which was trained with patches. For each patch, class probabilities are 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. 2*b*. Red colour (highest score) means that location has been classified as the target animal (with probability = 1.0) for all-encompassing windows. Blue colour means all the sliding windows including that image location have zero probability for the target class. A maximum probability value of each 32×32 px area can be 4 due to the intersection of four windows. In the rest of our computations, [0–4] range is normalised 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.

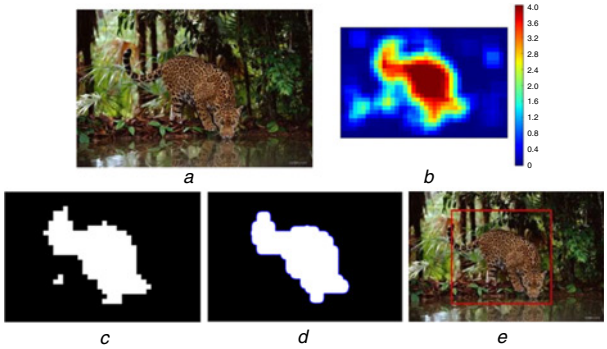

Fig. 2 *Prediction of bounding boxes*
a Input image
b Heat map for leopard class
c Binary image (heat map after applying threshold)
d Result after morphological operations
e Predicted bounding box

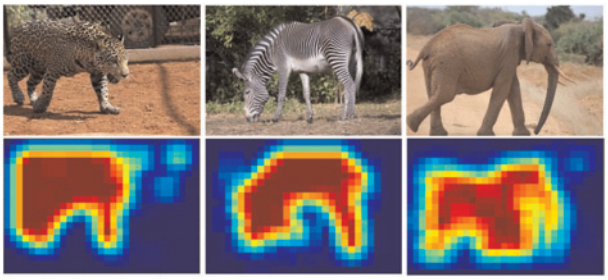

Fig. 3 *Input images are shown in first row, generated heat maps by patch-based approach are shown at second row*

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 Figs. 2*c–e*.

Evaluation metrics: To evaluate the performance of object detection algorithms, generally precision–recall curves are used in the 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 ground truth box is higher than a threshold (generally taken as 0.5)

$$\mathrm{IoU} = \frac{\mathrm{Box}_{\mathrm{detected}} \cap \mathrm{Box}_{\mathrm{ground\ truth}}}{\mathrm{Box}_{\mathrm{detected}} \cup \mathrm{Box}_{\mathrm{ground\ truth}}} \tag{1}$$

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 and 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 obtain 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 in [5]

$$P_{\mathrm{AR}}(G,\ D) = \frac{\sum_j \mathrm{Area}(G \cap D_j)}{\sum_j \mathrm{Area}(D_j)}, \qquad R_{\mathrm{AR}}(G,\ D) = \frac{\sum_j \mathrm{Area}(G \cap D_j)}{\mathrm{Area}(G)} \tag{2}$$

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

detected windows are covered by ground truth, R_{AR} considers how much of the ground truth area is covered by detected windows.

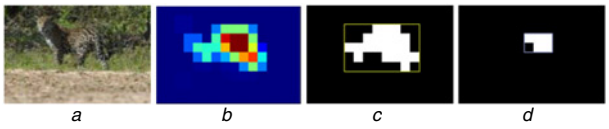

Fig. 4 Shrinking of bounding boxes as score threshold increases
a Input image
b Generated heat map
c Predicted bounding box at threshold = 0.3
d Predicted bounding box at threshold = 0.9

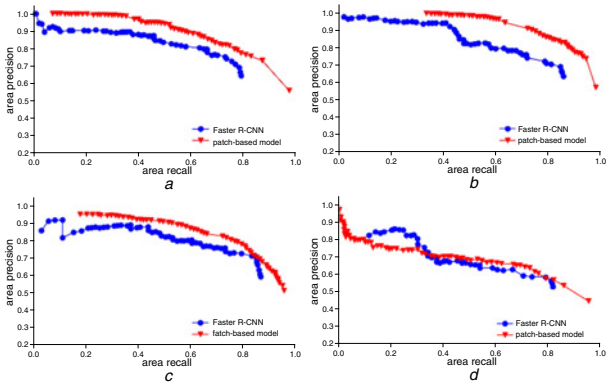

Fig. 5 *Area-precision against area-recall curves for Faster R-CNN and patch-based approach for*
a Leopard
b Zebra
c Elephant
d Bear

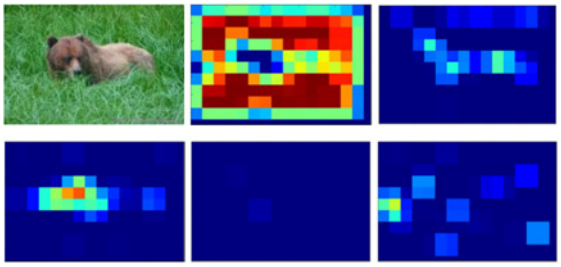

Fig. 6 *Result of patch-based approach on a bear image. First row (left-to-right): input image, background class heat map and bear class heat map. Second row (left-to-right): elephant, leopard and zebra class heat maps*

Experimental results: Faster R-CNN is trained with 446 bears, 351 elephants, 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 against 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. 5*d*), patch-based method outperforms Faster R-CNN only when area-recall is >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 the background. Some other patches have higher score (probability) in elephant heat map.

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

$$P_{\text{FasterRCNN}_j} = \frac{P_{\text{FasterRCNN}_j} + \text{PatchCont}_j}{2} \quad (3)$$

Here, $P_{\text{FasterRCNN}}$ 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 the patch-based method and it is computed by

$$\text{PatchCont}_j = \frac{\sum_{i=1}^{n} \text{Area}(\text{PatchBasedBox}_i \cap \text{FasterRCNN}_j)}{\sum_{i=1}^{n} \text{Area}(\text{PatchBasedBox}_i)} \quad (4)$$

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 the patch-based model. Results (Fig. 7) show that for leopard, zebra and elephant classes, combined model outperforms Faster R-CNN. For bear class (Fig. 7*d*), only combined model at 0.25 is superior to Faster R-CNN.

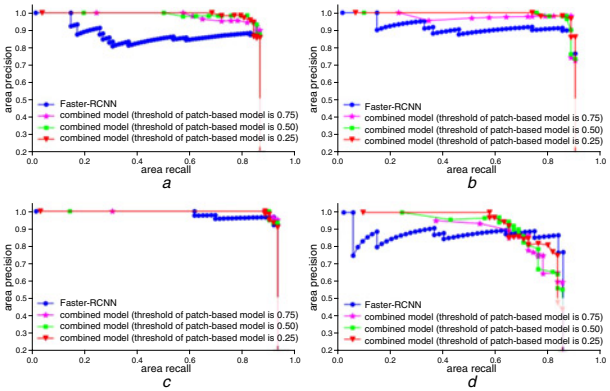

Fig. 7 *Precision–recall curves of Faster R-CNN and combined model for*
a Leopard
b Zebra
c Elephant
d Bear classes

Conclusions: We showed that the proposed patch-based training approach outperforms Faster R-CNN, especially for classes with distinctive patterns. Also, used in combination, patch-based method is able to increase the performance of Faster R-CNN.

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

Acknowledgment: This work was supported by the TUBITAK (grant no. 115E918).

Submitted: *19 December 2017* E-first: *27 February 2018*
doi: 10.1049/el.2017.4725
One or more of the Figures in this Letter are available in colour online.

S. Orhan and 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., *et al.*: 'Overfeat: integrated recognition, localization and detection using convolutional networks', 2013. Available at http://arxiv.org/abs/1312.6229
2 Ren, S., He, K., Girshick, R., *et al.*: 'Faster R-CNN: towards real-time object detection with region proposal networks'. Neural Information Processing Systems (NIPS), Montreal, Canada, December 2015, pp. 91–99
3 Redmon, J., Divvala, S., Girshick, R., *et al.*: 'You only look once: unified, real time object detection'. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, June 2016, pp. 779–788
4 He, K., Zhang, X., Ren, S., *et al.*: 'Deep residual learning for image recognition'. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, July 2016, pp. 770–778
5 Wolf, C., and Jolion, J.M.: 'Object count/area graphs for the evaluation of object detection and segmentation algorithms', *Int. J. Doc. Anal. Recognit.*, 2006, **8**, (4), pp. 280–296