# A Context Aware Deep Learning Architecture for Object Detection

Kevin Bardool, Tinne Tuytelaars, José Oramas
ESAT-PSI, KU Leuven, Belgium

August 2019

## 1. Introduction

A notable feature of our visual sensory system is its ability to exploit contextual cues present in a scene, enhancing our perception and understanding of the image. Exploring methods of incorporating and learning such information using deep neural network (DNN) architectures is a relatively young field of research, with room for more studies. In this work we propose a new architecture aimed at learning contextual relationships and improving the precision of existing DNN-based object detectors. An off-the-shelf detector is modified to extract contextual cues present in scenes. We implement a DNN-based architecture aimed at learning this information. A synthetic image generator is implemented that generates random images while enforcing a set of simple, predetermined contextual relationships. Finally, a series of experiments are carried out to evaluate the effectiveness of our design by measuring the improvement in average precision.

## 2. Background

There have been various attempts to categorize sources of contextual information [2, 5, 7]. Biederman groups relationships between an object and its surroundings into five classes: *interposition, support, probability, position, and size* [2]. It is the three latter relationships which are of our interest: *probability*, the likelihood of an object appearing in a particular scene, *position*, the expectation that when certain objects are present, they normally occupy predictable positions, and *size*, the expectation that an object's size relative to other objects and the general scene [7].

In works aimed at exploiting contextual information in DNN-based object detectors, two main approaches stand out. One uses contextual information as a feedback method that guides the generation of initial object proposals [12, 3]. A second approach involves the extraction and use of contextual information after proposal selection, and during the scoring stage [13, 9, 1].

Whereas in these approaches the use of contextual information is intertwined with the object detection architecture, we take a different approach: the separation of appearance detection and contextual reasoning. The object detector remains responsible for generating appearance-based information as well as detection and localization. Additionally, it will be used to construct contextual feature descriptors that are passed on to a secondary model, responsible for learning contextual relationships. While in some works a secondary model was used as a source of contextual information flowing *into* the object detector, our design will attempt to reverse this flow of information: contextual information from the object detector is passed on to a secondary model trained to learn contextual relationships. At inference time, the secondary model is used to re-evaluate object detector proposals.

## 3. Methodolgy

Our architectural pipeline consists of two stages (Figure 1). The first stage is an off-the-shelf object detector. For this, the Mask R-CNN model [8] was selected. A new network layer was implemented in the object detector to generate per-class *contextual feature maps*. These heatmaps are constructed using confidence scores and bounding boxes produced by the Mask R-CNN object detection and localization heads.

The secondary model is trained to learn semantic relationships using the contextual feature maps generated by the primary object detector. For this stage, a DNN model based on the Fully Convolutional Network (FCN) architecture [11] was implemented. The output of this model is also a series of contextual feature maps, representing its confidence on the original detections based on contextual relationships it has learned.

A scoring layer is implemented to produce a 'contextual score' for each proposed detection, and added to both models. Scores are calculated using the contextual feature maps generated by the object detector and contextual model. They are used for comparison and AP calculations, and allow us to measure the impact of the contextual learner on object detection performance.

For training and evaluation two types of datasets were considered. First, a dataset of synthetically generated images containing a series of consistently enforced contextual relationships. Such a dataset will allow us to train the pipeline with relatively simple content, introducing contextual cues in a controlled manner. In addition, a subset of the COCO dataset [10] was selected as a real-world dataset.

Training was conducted using various choices of loss functions and scoring algorithms. Eventually a Binary Cross Entropy loss was selected, and contextual scoring was performed using a localized summation on a tight region surrounding each object proposal's centroid, defined by their predicted bounding box.

## 4. Experiments

Conducted experiments have mainly focused on measuring the capacity of our design in learning various contextual relationships enforced by the synthetic image generator. The contextual scores are used to compute the average precision (AP) and mean average precision (mAP) as defined in the evaluation protocols for Pascal VOC and COCO challenges [6, 4].

Inference on a set of 500 images demonstrates a context-based mAP improvement of approximately 1.3 points. However, the Mask R-CNN based softmax score outperforms our context-score APs (Figure 2).

We measure the model's capacity in detecting the expected spatial context of objects. A controlled set of hypotheses that include object proposals positioned out of their expected spatial location is generated and passed to the contextual reasoning model. The success of the contextual model in rejecting such false positives indicates that it is able to learn spatial constraints (Figure 3).

Another set of experiments were conducted to test the model's capability in recognizing spatial relations enforced between objects belonging to different classes. Here we were able to confirm that the contextual model does recognize such spatial relationships in a limited range (Figure 4).

Experiments measuring the model's capability in learning semantic co-occurrence relationships between classes (i.e., objects of different classes that appear together in scenes) determined that our model is unable to learn such relationships.

In the synthetic images, a sense of depth is created by scaling object sizes relative to a horizon line present in the image. Our contextual model was able to learn the relationship between relative size and vertical location for different classes, favoring larger size objects when positioned lower in the image (i.e., closer to the observer).

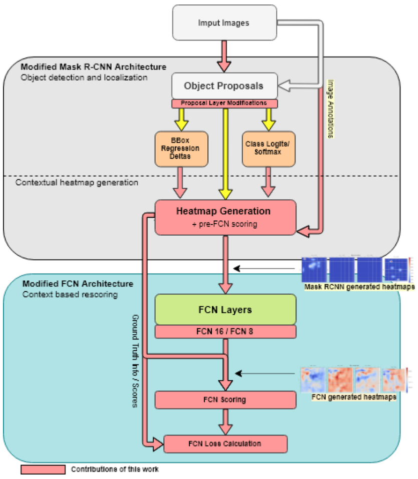
Imput Images
Modified Mask R-CNN Architecture
Object detection and localization
Object Proposals
Proposal Layer Modifications
Image Annotations
BBox Regression Deltas
Class Logits/ Softmax
Contextual heatmap generation
Heatmap Generation
+ pre-FCN scoring
Mask RCNN generated heatmaps
Modified FCN Architecture
Context based rescoring
Ground Truth Info / Scores
FCN Layers
FCN 16 / FCN 8
FCN generated heatmaps
FCN Scoring
FCN Loss Calculation
Contributions of this work

**Figure 1:** Our proposed architectural pipeline

| | **mAP** | person | car | sun | building | tree | cloud | airplane | truck |
|---|---|---|---|---|---|---|---|---|---|
| MR-CNN Baseline | 83.24 | 79.49 | 86.16 | 90.66 | 78.07 | 80.83 | 79.60 | 88.26 | 79.82 |
| Detector Stg - Score 1 | 77.95 | 75.84 | 82.91 | 88.83 | **73.72** | **79.55** | 73.67 | 81.89 | 67.21 |
| Rescoring Stg - Score 1 | **79.27** | **76.44** | **83.93** | **89.11** | 71.21 | 79.16 | **74.70** | **86.30** | **73.30** |

**Figure 2:** Detection results on test toy dataset.

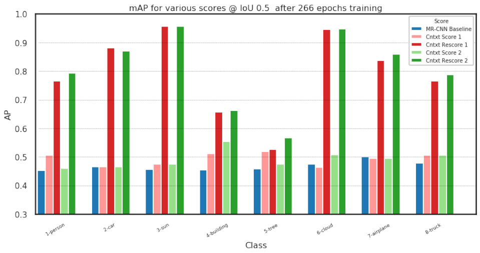

**Figure 3:** AP comparison on images with out of context proposals. Dark green/red bars represent the contextual learner's AP.

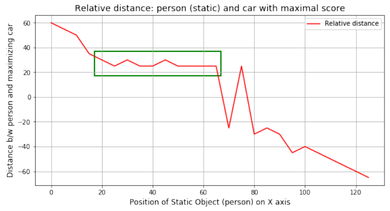

**Figure 4:** Maximal score vs. relative positioning of car and person objects. Green box indicates region where correct relative positioning results in maximal scores.

## 5. Conclusions

We propose a two stage architecture that extracts and learns contextual relationships using feature maps that encode such information. Results of our experiments show that the context-based model is able to learn intra- and inter-class spatial relationships. Additionally, it is able to learn the relation between the size of an image and its depth in the scene. However, we have not seen robustness towards learning co-occurrence of semantically related objects. Continuing experiments on the more challenging COCO dataset and investigating methods to induce learning of semantic co-occurrence relationships are open avenues for future work.

**References**

[1] S. Bell, C. L. Zitnick, K. Bala, and R. Girshick. Inside-Outside Net: Detecting Objects in Context with Skip Pooling and Recurrent Neural Networks. 2015.

[2] I. Biederman, R. J. Mezzanotte, and J. C. Rabinowitz. Scene Perception : Detection and Judging Objects undergiong relational violations. *Cognitive Psychology*, 177(2):143–177, 1982.

[3] X. Chen and A. Gupta. Spatial Memory for Context Reasoning in Object Detection. *Proceedings of the IEEE International Conference on Computer Vision*, 2017-Octob:4106–4116, 2017.

[4] COCO Consortium. Coco: Common objects in context - detection evaluation. URL: `https://cocodataset.org/detection-eval` last accessed on 2019-03-23.

[5] S. Divvala, D. Hoiem, J. Hays, A. Efros, and M. Hebert. An empirical study of context in object detection. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1271–1278. IEEE, 6 2009.

[6] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (VOC) challenge. *International Journal of Computer Vision*, 88(2):303–338, 6 2010.

[7] C. Galleguillos and S. Belongie. Context based object categorization: A critical survey. *Computer Vision and Image Understanding*, 114(6):712–722, 2010.

[8] K. He, X. Zhang, S. Ren, and J. Sun. Deep Residual Learning for Image Recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[9] J. Li, Y. Wei, X. Liang, J. Dong, T. Xu, J. Feng, and S. Yan. Attentive contexts for object detection. *IEEE Transactions on Multimedia*, 19(5):944–954, 2017.

[10] T.-Y. Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, C. L. Zitnick, and P. Doll. Microsoft COCO: Common objects in context. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 8693 LNCS(PART 5):740–755, 2014.

[11] E. Shelhamer, J. Long, and T. Darrell. Fully Convolutional Networks for Semantic Segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(4):640–651, 2016.

[12] A. Shrivastava and A. Gupta. Contextual priming and feedback for faster R-CNN. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 9905 LNCS:330–348, 2016.

[13] X. Zeng, W. Ouyang, J. Yan, H. Li, T. Xiao, K. Wang, Y. Liu, Y. Zhou, B. Yang, Z. Wang, H. Zhou, and X. Wang. Crafting GBD-Net for Object Detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 8828(c):1–16, 2017.