Dynamic, Dense, Cascaded R-CNNs

Sergey Karayev
UC Berkeley

Ross Girshick
UC Berkeley

Sergio Guadarrama
UC Berkeley

Mario Fritz
MPI Informatics

Trevor Darrell
UC Berkeley

Abstract

Recently, an object detection architecture based on two key decisions has demonstrated an undeniable increase in performance on both PASCAL VOC and ILSVRC detection datasets. The first is doing bottom-up Region of Interest (ROI) proposals instead of sliding-window evaluation. The second is using multi-layer Convolutional Neural Nets (CNNs) pre-trained on ILSVRC classification data to featurize canonically resized ROIs. Although by far the best in terms of recognition performance, the method is computationally costly, requiring processing of several hundred of thousand regions with the CNN. Even with efficient GPU implementation, the method takes ˜10 s per image after the computation of region proposals (another ˜2 s).

We propose three novel speed-ups for the task. (1) Dense evaluation: the whole image is processed with a CNN up to the highest non-fully-connected layer, and ROIs are featurized by cropping and resizing that highest layer, for use as an initial filter. (2) Cascaded CNN: each ROI passing the initial filter is further processed by a CNN that is augmented with a *reject* option after each layer. (3) The selection of ROIs for further processing is done dynamically, taking into account the evaluation results on the ROIs selected so far.

Combined, these speed-ups allow us to match original R-CNN performance in 20% of the time.

1 Introduction

[Standard object detection, deep learning, computational concerns, introduction]

2 Related Work

2.0.1 Object recognition with CNN

- Alexnet
- RCNN and the couple other CNN detection methods
- Overfeat: explain dense evaluation

2.0.2 Cascaded detection

- Viola Jones
- Cascaded DPM
- Is there any CNN work?

2.0.3 Dynamic selection

- Timely Object Recognition
- Cross-talk cascades
- ?

3 Method

Our method builds on the architecture described in [1], summarized in Figure 1.

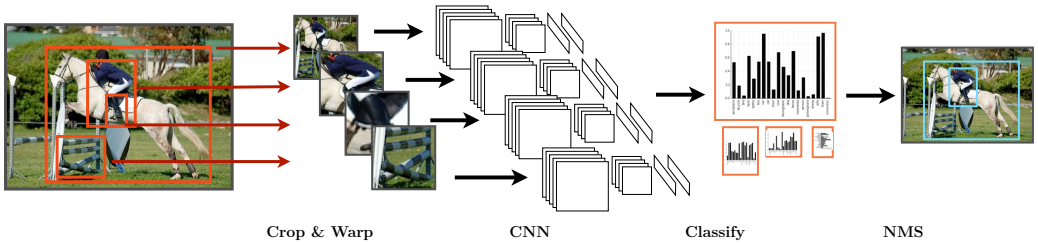

Figure 1: R-CNN architecture: image regions are cropped, resized, and each one fed through a CNN with classification layers. The classifier outputs are post-processed to give the final detections.

3.1 Dense region evaluation

Figure 2 shows the dense region evaluation.

- Explain the difference between cropping pixels and cropping `pool5`.
- Explain design choices (single vs multi scale, finding nearest window, warping.

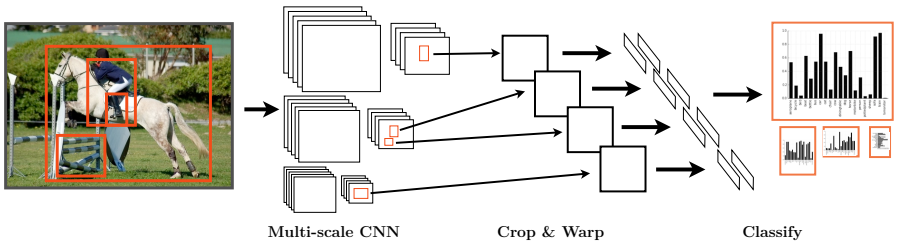

Figure 2: Post R-CNN architecture: the whole image is fed through a CNN up to the highest pooling layer. Regions are cropped from that layer, resized, and classified. The classifier outputs are post-processed to give the final detections.

3.2 Cascaded CNN

Figure 3 shows the Cascaded CNN model.

- Explain the reject option in the CNN evaluation
- Each rejector is trained to distiguish background regions from foreground (object) regions.
- Explain training the thresholds
- [FIGURE: Cascaded CNN, showing threshold layers]

3.3 Dynamic region selection

Figure 4 shows the dynamic region selection loop.

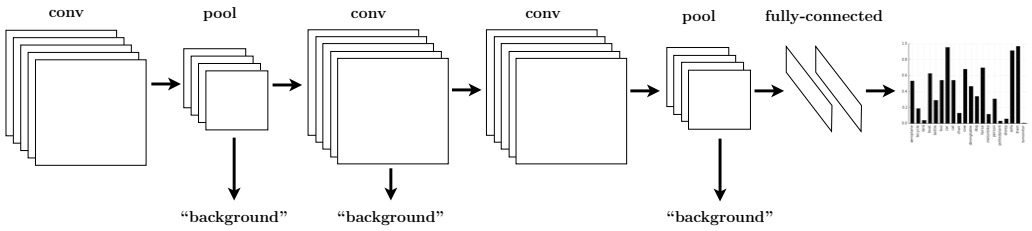

Figure 3: The Cascaded CNN has a reject option after certain layers.

- Explain dynamic selection of region batches
- Explain iterative training procedure

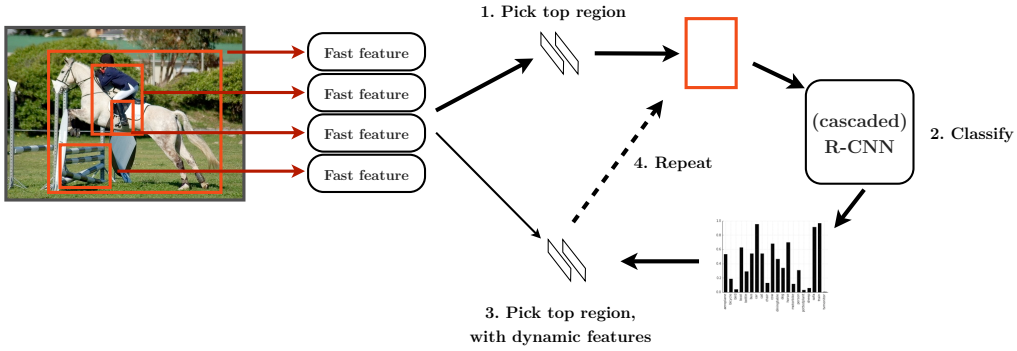

Figure 4: Dynamic region selection combines both speedups and introduces another.

3.4 Combined system

The entire system is combined by using Dense R-CNN as the fast feature and Cascaded CNN as the region evaluator in Figure 4.

4 Evaluation

(All results are on PASCAL VOC 2007 and 2010. ILSVRC is nice-to-have.)

- [TABLE: Dense region evaluation, timing vs. accuracy: effect of multi scale, pooling.]
- From the above, we select the point of high accuracy and reasonable cost.
- [PLOT: Cascaded CNN timing vs. accuracy: effect of threshold.]
- From the above, we select threshold of high accuracy and reasonable cost.
- [TABLE: Combined system timing vs accuracy: effect of dynamic region selection.]

5 Conclusion

[Standard stuff]

5.1 Future work

- ?

References

[1] Ross Girshick, Jeff Donahue, Trevor Darrell, Jitendra Malik, and U C Berkeley. Rich feature hierarchies for accurate object detection and semantic segmentation. Technical report, 2014.