Rich feature hierarchies for accurate object detection and semantic segmentation

Xuewen Yang

August 1 2018

Abstract

Object detection performance, as measured on the canonical PASCAL VOC dataset, has plateaued in the last few years. The best-performing methods are complex ensemble systems that typically combine multiple low-level image features with high-level context. In this paper, the author propose a simple and scalable detection algorithm that improves mean average precision (mAP) by more than 30% relative to the previous best result on VOC 2012łachieving a mAP of 53.3%.Since they combine region proposals with CNNs, they call our method R-CNN: Regions with CNN features. They also compare R-CNN to OverFeat, a recently proposed sliding-window detector based on a similar CNN architecture. They find that R-CNN outperforms OverFeat by a large margin on the 200-class ILSVRC2013 detection dataset.

1. Introduction

CNNs saw heavy use in the 1990s (e.g., [27]), but then fell out of fashion with the rise of support vector machines. In 2012, Krizhevsky *et al.* [2] rekindled interest in CNNs by showing substantially higher image classification accuracy on the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) [1]. Their success resulted from training a large CNN on 1.2 million labeled images, together with a few twists on LeCuns CNN.

Unlike image classification, detection requires localizing (likely many) objects within an image. One approach frames localization as a regression problem. However, work from Szegedy *et al.* [4], concurrent with this method, indicates that this strategy may not fare well in practice (they report a mAP of 30.5% on VOC 2007 compared to the 58.5% achieved by this method). An alternative is to build a sliding-window detector. CNNs have been used in this way for at least two decades, typically on constrained object categories, such as faces and pedestrians [3]. In order to maintain high spatial resolution, these CNNs typically only have two convolutional and pooling layers. The author also considered adopting a sliding-window approach. However, units high up in our network, which has five convolutional layers, have very large receptive fields (195×195 pixels) and strides (32×32 pixels) in the input image, which makes precise localization within the sliding-window paradigm an open technical challenge.

2. Object detection with R-CNN

This object detection system consists of three modules. The first generates category-independent region proposals. These proposals define the set of candidate detections available to the detector. The second module is a large convolutional neural network that extracts a fixed-length feature vector from each region. The third module is a set of classspecific linear SVMs. In this section, the author present their design decisions for each module, describe their test-time usage, detail how their parameters are learned, and show detection results on PASCAL VOC 2010-12 and on ILSVRC2013.

3. Conclusion

The author achieved this performance through two insights. The first is to apply high-capacity convolutional neural networks to bottom-up region proposals in order to localize and segment objects. The second is a paradigm for training large CNNs when labeled training data is scarce. They show that it is highly effective to pre-train the networkł with supervisionłfor a auxiliary task with abundant data (image classification) and then to fine-tune the network for the target task where data is scarce (detection). They conjecture that the supervised pre-training/domain-specific finetuning paradigm will be highly effective for a variety of data-scarce vision problems.

References

[1] J. Deng, W. Dong, R. Socher, L. J. Li, K. Li, and F. F. Li. ImageNet: A large-scale hierarchical image database. In *CVPR*, 2009. 1

[2] A. Krizhevsky, I. Sutskever, and G. E. Hinton. ImageNet classification with deep convolutional neural networks. In *NIPS*, 2012. 1

[3] P. Sermanet, K. Kavukcuoglu, S. Chintala, and Y. Lecun. Pedestrian detection with unsupervised multi-stage feature learning. In *CVPR*, 2013. 1

[4] C. Szegedy, A. Toshev, and D. Erhan. Deep neural networks for object detection. In *NIPS*, 2013. 1