Rich feature hierarchies for accurate object detection and semantic segmentation

Comment

Abstract

- A scalable detection algorithm improving mean average precision by more than 30% (achieving 53.3%)
- Apply high-capacity CNNs to bottom-up region proposals to localize and segment objects
- When labeled training data is scarce, supervised pre-training for an auxiliary task, followed by domain-specific fine-tuning, yields a significant performance boost

1. Introduction

- ImageNet on ILSVRC 2012 arose a question: To what extent do the CNN classification results on ImageNet generalize to object detection results on PASCAL VOC?
- Two main problem: localizing objects with a deep network
- Training a high-capacity model with only a small quantity of annotated detection data.
- It generates around 2000 category-independent region proposals for input image, extracts a fixed-length feature vector from each proposal using a CNN, and then classifies each region with category-specific linear SVMs.
- To compute a fixed-size CNN input from each region proposal regardless of the shape, they used affine image warping
- The class-specific computations are small matrix-vector product and greedy non-maximum suppression, which allows the features are shared across all categories and **two orders of magnitude lower-dimensional**

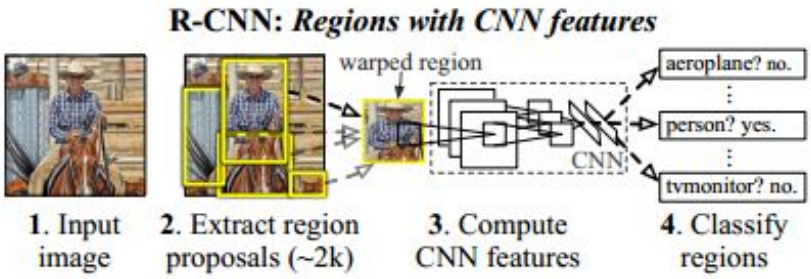

Figure 1: Object detection system overview. Our system (1) takes an input image, (2) extracts around 2000 bottom-up region proposals, (3) computes features for each proposal using a large convolutional neural network (CNN), and then (4) classifies each region using class-specific linear SVMs. R-CNN achieves a mean

2. Object detection with R-CNN

- Three modules
  - category-independent region proposals
  - large CNN to extract a fixed-length feature vector from each region
  - a set of class-specific linear SVMs

2.1 Module design

Region proposals

Feature extraction

- Dilates the tight bounding box so that at the warped size there are exactly p pixels of warped image context around the original box (empirically $p = 16$)

2.2 Test-time detection

- rejects a region if it has IoU overlap with a higher scoring selected region larger than a learned threshold

Run-time analysis

- Efficient - all CNN parameters are shared across all categories, the feature vectors computed by the CNN are low-dimensional
- 13 s/image on GPU (amortized over classes), 10 s even with 100k classes

2.3 Training

Supervised pre-training

Domain-specific fine-tuning

- plus 1 for background

Object category classifier

2.4 Results on PASCAL VOC 2010-12

mAP - 35.1% to 53.7%

2.5 Results on ILSVRC2013 detection

mAP 24.3% to 31.4% (compared with OverFeat)

3. Visualization, ablation, and modes of error

3. 1 Visualizing learned features

- Single out a particular unit (feature) in the network and use it as if were an object detector.
- computer the activations and perform non-maximum suppression

3. 2. Ablation studies

Performance layer-by-layer, without fine-tuning

- Without fc_6 and $_7$. it performs well even though the only 6% of parameters are used. This suggests potential utility in computing a dense feature map, in the sense of HOG, of an arbitrary-sized image by using only CNN.

Performance layer-by-layer, without fine-tuning

- Increases map by 8.0%p to 54.2%
- The effect appears much larger for fc_5 than fc_6 and fc_7

4. The ILSVRC2013 detection dataset

5. Semantic segmentation

6. Conclusion

- Apply high-capacity CNNs to bottom-up region proposals in order to localize and segment objects.
- Train large CNNs when labeled training data is scarce. It is highly effective to pre-train the network *with supervision* for a auxiliary task with abundant data.
- *"supervised pre-training/domain-specific fine-tuning"* paradigm