Table of Contents

Rich feature hierarchies for accurate object detection and semantic segmentation

Abstract

- This method combines two key insights: (1) high-capacity convolutional neural networks (CNNs) can be used to localize and segment objects from bottom-up region proposals, and (2) when labeled training data is scarce, supervised pre-training for an auxiliary task followed by domain-specific fine-tuning yields a significant performance boost.
- This paper mix region proposals with CNNs. Here also is a comparison between R-CNN and OverFeat, a sliding-window detector based on a similar CNN architecture that was recently suggested. On the 200-class ILSVRC2013 detection dataset, R-CNN outperforms OverFeat by a considerable margin.

Dataset

They used 200-class ILSVRC2013 detection dataset. The dataset is split into three sets: train (395,918), val (20,121), and test (40,152), where the number of images in each set is in parentheses.

Model Architecture

- This system takes an input image
- Extracts around 2000 bottom-up region proposals
- Warped image regions
- Forward each region through CovNet
- Classifies each region using class-specific linear SVMs

Working Process

Localizes and segment objects b applying-

- High capacity CNNS
- Bottom-up region proposals

Learning high capacity CNNs when labeled training data is scarce-

- Supervised pre training for an auxiliary task
- Followed by domain specific fine tuning

Result

Popular deformable part models have a 33.4 percent success rate. R-mAP CNN's on the 200-class ILSVRC2013 detection dataset is 31.4 percent, a significant improvement over OverFeat's previous best result of 24.3 percent.

Limitations

- To train the network, you'd have to classify 2000 region proposals every image, which would take a long time. It cannot be implemented in real time because each test image takes roughly 47 seconds.
- The algorithm for selective search is a fixed one. As a result, no learning takes place at that point. This could result in a slew of bad candidate region suggestions.

Fast R-CNN

Abstract

- The Fast Region-based Convolutional Network model (Fast R-CNN) is proposed in this research for object detection.
- Rather than the multi-stage training procedure employed by R-CNN, Fast R-CNN uses a shortened training process with only one fine-tuning. Fast R-CNN improves training and testing speed while also enhancing detection accuracy when compared to prior R-CNN networks.
- On PASCAL VOC 2012, Fast R-CNN trains the very deep VGG16 network 9 times faster than R-CNN, is 213 times faster at test time, and achieves a higher mAP.
- Fast R-CNN trains VGG16 3 times quicker, tests 10 times faster, and is more accurate than SPPnet.

Dataset

They used three datasets to evaluate the network: VOC07, VOC 2010, and VOC 2012. All of Fast RCNN's experiments involve single-scale training and testing.

Model Architecture

- The fast R-CNN network uses the complete input image as well as a list of item suggestions. To create a conv feature map, the network first processes the entire image with many convolutional and max pooling layers.
- After that, a region of interest (ROI) pooling layer derives a fixed length feature vector from the feature map for each object proposition. Each feature vector is then sent through a series of fully connected layers.

Working Process

The authors used pre-trained networks to initialize the architecture while training Fast R-CNN. The pre-trained networks that have been initialized go through the following alterations:

i. The ROI pooling layer has taken the place of the previous max pooling layer.
ii. The network's last fully connected layer and softmax are replaced with two sibling layers, namely a fully connected layer with softmax across K + 1 categories (number of categories plus background) and category-specific bounding box regressors.
iii. A list of photos and a list of RoIs in those images are sent into the network as input.

The following ImageNet models are used in the research.

a. AlexNet Model (Small)
b. VGG CNN M 1024, which has the same depth as S but is broader. M (Medium): VGG CNN M 1024, which has the same depth as S but is wider.
c. L (Large): VGG16 model with a lot of depth.

Result

The following primary experimental results back up the paper's claims.

- On VOC07, 2010, and 2012, state-of-the-art mAP was used
- In comparison to R-CNN, SPPnet allows for faster training and testing.
- In VGG 16, fine-tuning conv layers enhances mAP

Limitations

As comparing the performance of Fast R-CNN during testing, we observe that including region proposals considerably slows down the algorithm when compared to not using region proposals. As a result, region proposals become barriers in the Fast R-CNN algorithm, slowing it down.

Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks

Abstract

- They introduce the Region Proposal Network (RPN) in this paper, which shares full-image convolutional features with the detection network, allowing for nearly cost-free region proposals.
- An RPN is a fully convolutional network that predicts object limits and scores at each position at the same time. The RPN is trained from start to finish to create high-quality region proposals that Fast R-CNN uses for detection.
- Using the lately trendy nomenclature of neural networks with "attention" processes, they further fuse RPN and Fast R-CNN into a single network by sharing their convolutional features. The RPN component informs the united network where to look.

Dataset

They put this approach to the test on the PASCAL VOC 2007 detection benchmark. This dataset contains around 5k trainval and 5k test pictures across 20 object categories. They also provide results for a few models on the PASCAL VOC 2012 benchmark.

Model Architecture

Faster RCNN is built upon two modules:

1. RPN: For generating region proposals.
2. Fast R-CNN: For detecting objects in the proposed regions.

Working Process

The Faster R-CNN works as follows:

- The RPN generates region proposals.
- For all region proposals in the image, a fixed-length feature vector is extracted from each region using the ROI Pooling layer.
- The extracted feature vectors are then classified using the Fast R-CNN.
- The class scores of the detected objects in addition to their bounding-boxes are returned.

Result

- On the PASCAL VOC 2007 testset (trained on VOC 2007 trainval) with Fast R-CNN with ZF detectors but distinct region proposal methods, the RPN method outperforms Selective search and edgebox by 1.3 mAP.
  When the detector is Fast RCNN and VGG16, the mAP on the VOC 2007 test set is 78.8 percent when trained on COCO+07+12 (union set of VOC 2007 trainval and VOC 2012 trainval).
- When trained on the COCO train dataset, Faster RCNN(on VGG-16) improves mAP@0.5 by 2.8 percent and mAP@[0.5, 0.95] by 2.2 percent on COCO test-dev.
  The Faster R-CNN system boosts the mAP@0.5/mAP@[0.5, 0.95] from 41.5 percent /21.2 percent (VGG-16) to 48.4 percent /27.2 percent (ResNet-101) on the COCO val set when trained

on the COCO train dataset by simply replacing VGG-16 with a 101-layer residual net (ResNet-101).

Limitations

One drawback of Faster R-CNN is that the RPN is trained using a single image to extract all anchors in the mini-batch of size 256. The network may take a long time to attain convergence because all samples from a single image may be correlated.