Summary of R-CNN Paper

Title: Rich feature hierarchies for accurate object detection and semantic segmentation

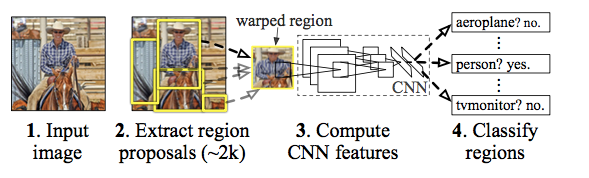
Initial Goal

1. Getting better performance with the object detection task.
2. Using a deep network to locate objects and train a high-capacity model only with a modest amount of annotated detection data.

Contributions

1. To locate and segment objects using high-capacity convolutional neural networks (CNNs) on bottom-up region recommendations.
2. Scarcity of training data with labels is a problem, the solution may be using pre-training (unsupervised) for an auxiliary task which can be followed by supervised fine-tuning for any domain-specific task.
3. Detection fine-tuning boosts mAP performance by 8 percentage points. Their approach obtains an mAP of 54 percent on VOC 2010 after fine-tuning, compared to 33 percent for the highly-tuned, HOG-based deformable component model (DPM).
4. R-CNN gets a mAP of 31.4 percent on ILSVRC2013 detection, which is much higher than the second-best result of 24.3 percent from OverFeat.

Model Architecture



1. R-CNN employs selective search to extract region proposals, allowing for a controlled comparison with previous detection work.
2. The Caffe approach is then used to extract a 4096-dimensional feature vector from each area proposal. Forward propagation of a mean-subtracted 227X227 RGB picture via five convolutional layers and two fully connected layers is used to calculate features.
3. It performs a greedy non-maximum suppression (for each class separately) to all scored areas in an image, rejecting a region if it has an intersection-over-union (IoU) overlap with a higher scoring selected region bigger than a learned threshold (>=0.5).
4. This paper optimizes one linear SVM per class after extracting features and applying training labels.

How it works

1. This approach creates category-independent region proposals from the supplied image.
2. Using CNN, extracts a fixed-length feature vector from each region proposal.
3. From each region proposal, extract a fixed-length feature vector using CNN.

Limitations

1. Training process is multi-staged. It must be trained in three steps.
2. Training is costly in terms of both space and time. It stores the pre-trained CNN's extracted features to disk so that the SVMs may be trained later. There is a lot of storage necessary.
3. Detecting objects takes a long time. R-CNN generates region proposals using the Selective Search method, which takes a long time. We must apply CNN to 2000 area proposals to extract features if we have 2000 region proposals per picture, which is time demanding.