Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation

Abstract and Introduction

Observations made by the authors through the course of the research work:

- One can apply high-capacity convolutional neural networks to bottom-up region proposals in order to localize and segment objects.
- When labelled training data is scarce, supervised pre-training for an auxiliary task, followed by a domain-specific fine-tuning, yields a significant performance boost.

Achieves results by:

- localizing objects with a deep network and,
- training a high-capacity model with only a small quantitiy of annotated detection data.

Sliding window based object localization does not provide precise localization. In deeper neural networks, neurons in the deeper layers, due to convolution, have very large receptive fields that make precise localization very difficult.

The goal of localization of objects in images was not done by the sliding window technique. Most work prior to it had used sliding windows with relatively shallow networks having only two convolution and pooling layers. In contrast, the CNN used by the RCNN consists of five convolution and pooling layers. Due to this distinction, units in the higher/later layers have large receptive fields and strides making precise localization difficult.

To overcome the issue of scarce labelled data, the work utilizes a supervised pre-training of the dataset on a large auxiliart dataset followed by domain-specific fine-tuning on a small dataset. It shows that this procedure acts as an effective paradigm for learning high-capacity CNNs when data is scarce.

Object detection with R-CNN

The object detection system consists of three modules:

- The first generates **category-independant** region proposals. These are the candidates that are provided to the detector portion of the system.
- The second is a large CNN that extracts a fixed-length feature vector from each of the regions passed from the previous module.
- The third is a set of class-specific linear SVMs.

Module design

Region proposals

Utilizes selective search to generate region proposals.

Test-time detection

At test time, the method runs selective search on the test image to extract around 2000 region propsals. Each proposal is warped to the correct shaped by a simple affine transformation and propagated through the CNN.

Each warped region passes through the network to provide a 4096-dimensional feature vector. The dimension of the feature vector is **not** the number of classes. The feature vector gets fed to each of the linear class-trained SVMs to give a class score. For N classes to be classified, there would be N SVMs, one corresponding to each such class. A non-maximum suppression is applied to the vector of all the class scores for the region. This gives the classification of the region. A region is rejected if if has an intersection-over-union overlap greater than a learned threshold with another region proposal having a higher class score for the same class(after non-max supression).

Run-time analysis

Detection is made efficient by two properties of the model architecture.

- CNN parameters are shared across all categories. Sharing of parameters amortizes the time spent computing region proposals and features across all the classes. The only class-specific computations are the dot products of the feature vectors with the SVM weights for each class and non-maximum suppression which in practice can be carried out in a single matrix product. An example of such a multiplication would be:
  - The feature matrix consisting of the feature vectors of all the regions **after** passing through the CNN would be of shape 2000×4096
  - The SVM weight matrix would be of shape $4096 \times N$ where N is the number of classes to be identified.
-

Tags

:computerVision:cnn:objectDetection:semanticSegmentation:rcnn: