VSB TECHNICAL UNIVERSITY OF OSTRAVA | FACULTY OF ELECTRICAL ENGINEERING AND COMPUTER SCIENCE | DEPARTMENT OF COMPUTER SCIENCE

Deep Learning

Object recognition

Jan Platoš, Radek Svoboda

March 24, 2024

Department of Computer Science
Faculty of Electrical Engineering and Computer Science
VŠB - Technical University of Ostrava

Object recognition

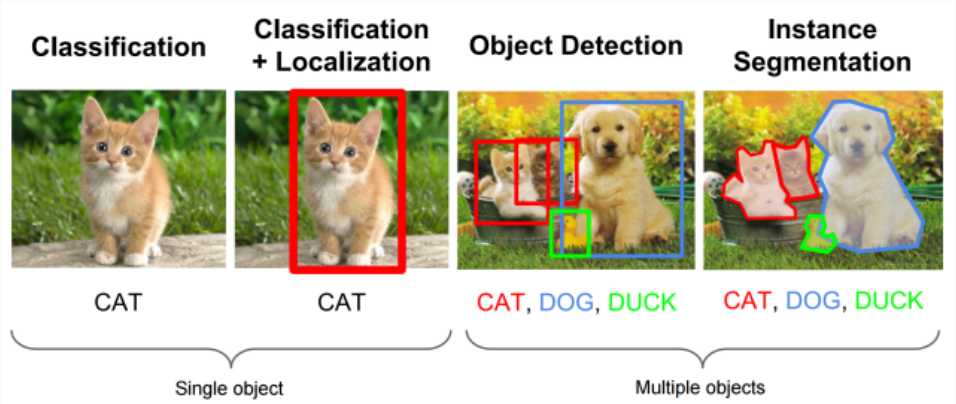
Object recognition

- Object recognition is a general term that covers many areas:
- **Image Classification**: Predict the type or class of an object in an image.
  - Input: An image with a single object, such as a photograph.
  - Output: A class label (e.g. one or more integers that are mapped to class labels).
- **Object Localization:** Locate the presence of objects in an image and indicate their location with a bounding box.
  - Input: An image with one or more objects, such as a photograph.
  - Output: One or more bounding boxes (e.g. defined by a point, width, and height).

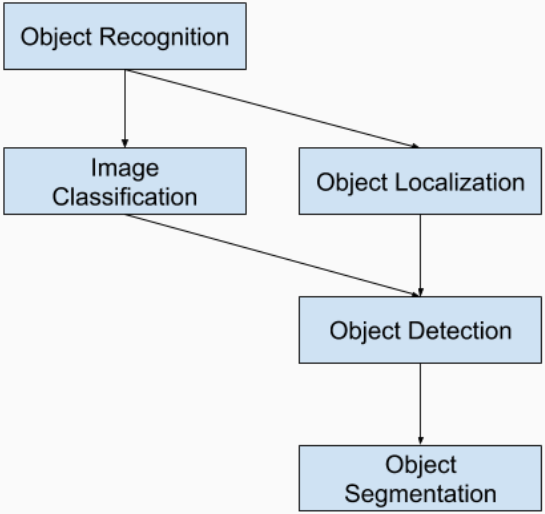
Object recognition

- **Object Detection:** Locate the presence of objects with a bounding box and types or classes of the located objects in an image.
  - Input: An image with one or more objects, such as a photograph.
  - Output: One or more bounding boxes (e.g. defined by a point, width, and height), and a class label for each bounding box.
- **Object Segmentation:** Locate the precise (pixel-wise) position of an object and types or classes of the located objects in an image.
  - Input: An image with one or more objects, such as a photograph.
  - Output: One or more precise segmentation (e.g. defined by pixels), and a class label for each bounding box.

Object recognition



Object recognition

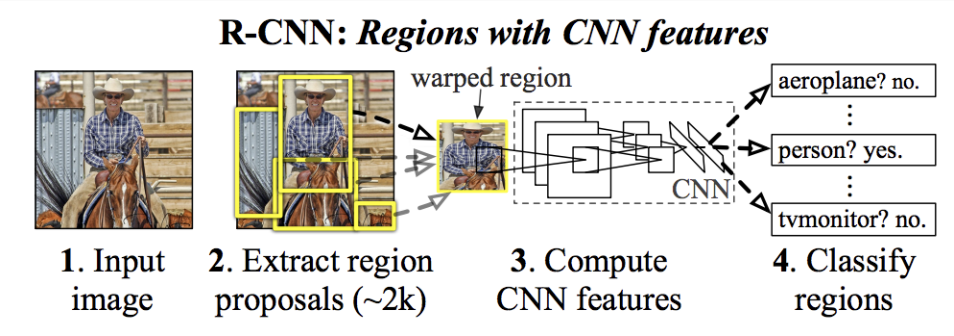


Object recognition - R-CNN Methods

- *Regions with CNN Features* or *Region-Based Convolutional Neural Network* (R-CCN) is a family of methods.
- It combines CNN and Dense layers to extract regions of interest.
- Classification is done on the extracted regions.

Object recognition - R-CNN Methods - R-CNN

- The first large and successful application of convolutional neural networks to the problem of object localization, detection, and segmentation.
- The proposed R-CNN model is comprised of three modules; they are:
  - **Module 1:** Region Proposal. Generate and extract category independent region proposals, e.g. candidate bounding boxes.
  - **Module 2:** Feature Extractor. Extract feature from each candidate region, e.g. using a deep convolutional neural network.
  - **Module 3:** Classifier. Classify features as one of the known class, e.g. linear SVM classifier model.


R-CNN: Regions with CNN features
warped region
CNN
aeroplane? no.
person? yes.
tvmonitor? no.
1. Input image
2. Extract region proposals (~2k)
3. Compute CNN features
4. Classify regions

Object recognition - R-CNN Methods - R-CNN

- A computer vision technique *selective search* is used to propose candidate regions/bounding boxes of potential objects.
- The feature extractor that is used is the AlexNet deep CNN.
- The output of the CNN is a 4,096 element vector that describes the contents of the image.
- Feature vector is classified using linear SVM - one SVM is trained for each known class.
- This approach is slow due to application of CNN on each region.

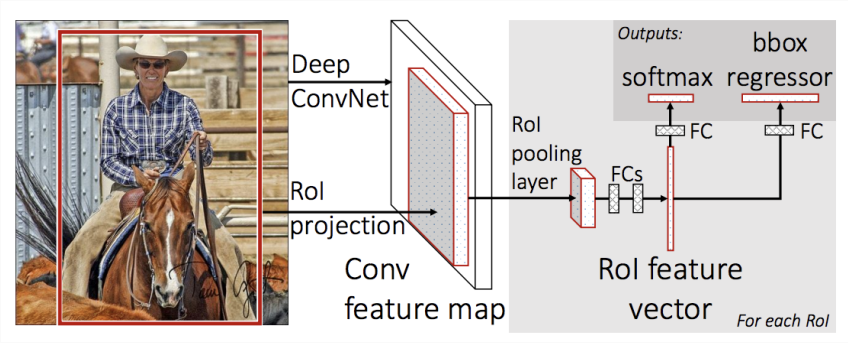
Object recognition - R-CNN Methods - Fast R-CNN

- Optimized version of the R-CNN focused on speed.
- Focuses on solving slowdowns caused by:
  - training in a multi-stage pipeline.
  - training a deep CNN on so many region proposals per image.
  - making predictions using a deep CNN on so many region proposals.
- Fast R-CNN is proposed as a single model instead of a pipeline to learn and output regions and classifications directly.

Object recognition - R-CNN Methods - Fast R-CNN

- The input is the image and a set of region proposals.
- The input is passed through a deep convolutional neural network.
- A pre-trained CNN, such as a VGG-16, is used for feature extraction.
- The end of the deep CNN is a custom layer called a Region of Interest Pooling Layer, or RoI Pooling, that extracts features specific for a given input candidate region.
- The output of the CNN is then interpreted by a fully connected layer.
- The model bifurcates into two outputs:
  - one for the class prediction via a softmax layer.
  - another with a linear output for the bounding box.
- This process is repeated multiple times for each region of interest in a given

Object recognition - R-CNN Methods - Fast R-CNN



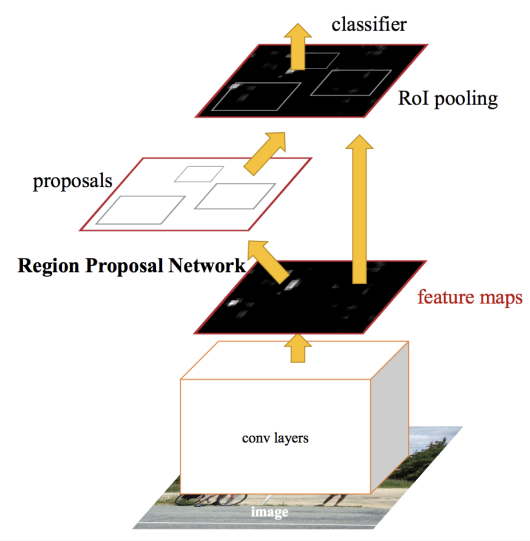
Object recognition - R-CNN Methods - Faster R-CNN

- The model architecture is improved for both speed of training and detection.
- The architecture is designed to propose and refine region proposals as part of the training process - a Region Proposal Network, or RPN.
- The regions are used in concert with a Fast R-CNN model in a single model design.
- The improvements reduce the number of region proposals and increases speed to near real-time.

Object recognition - R-CNN Methods - Faster R-CNN

- The unified model is composed into two modules.
- **Module 1:** Region Proposal Network. Convolutional neural network for proposing regions and the type of object to consider in the region.
- **Module 2:** Fast R-CNN. Convolutional neural network for extracting features from the proposed regions and outputting the bounding box and class labels.
- Both modules operate on the same output of a deep CNN.
- The region proposal network acts as an attention mechanism for the Fast R-CNN network informing the second network of where to look or pay attention.

Object recognition - R-CNN Methods - Faster R-CNN



Object recognition - R-CNN Methods - Faster R-CNN

- The pre-trained deep CNN is used for feature extraction.
- A small network over a feature map is used for outputting multiple region proposals and a class prediction for each of them.
- Region proposals are bounding boxes, based on so-called anchor boxes or pre-defined shapes designed to accelerate and improve the proposal of regions.
- The class prediction is binary, indicating the presence of an object, or not, so-called *objectness* of the proposed region.
- An alternating training procedure is used where both sub-networks are trained at the same time, although interleaved.
- The parameters in the feature detector are tuned for both tasks at the same

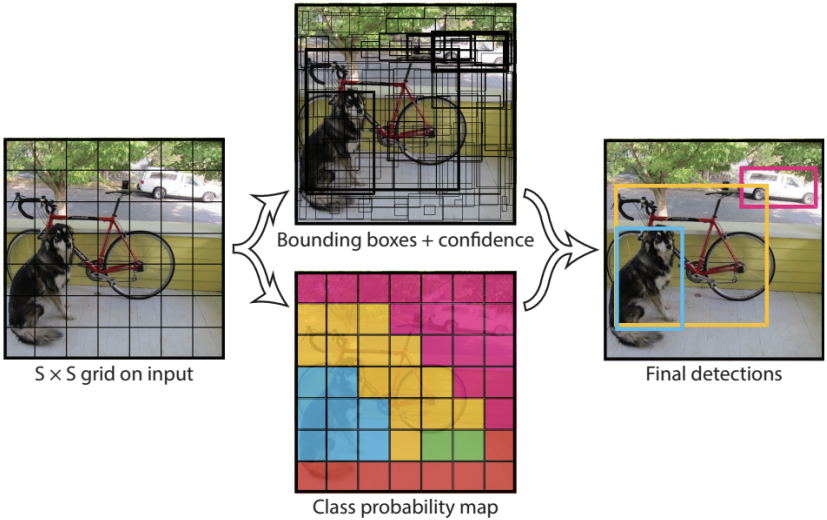
Object recognition - YOLO Models

- A You Only Look Once (YOLO) models are designed for speed but less precise than R-CNN.
- The approach involves a single neural network trained end to end.
- It takes an image as input and predicts bounding boxes and class labels for each bounding box directly.
- The technique offers lower predictive accuracy (e.g. more localization errors).
- It is able to operate at 45 frames per second and up to 155 frames per second for a speed-optimized version of the model.

Object recognition - YOLO Models

- The model works by first splitting the input image into a grid of cells.
- Each cell is responsible for predicting a bounding box if the center of a bounding box falls within it.
- Each cell predicts a bounding box involving the x, y coordinate and the width and height and the confidence.
- A class prediction is also based on each cell.

Object recognition - YOLO Models - YOLO



Object recognition - YOLO Models - YOLO v2

- The architecture uses batch normalization and high-resolution input images.
- YOLOv2 model makes use of anchor boxes, pre-defined bounding boxes with useful shapes and sizes that are tailored during training.
- The choice of bounding boxes for the image is pre-processed using a k-means analysis on the training dataset.
- The predicted representation of the bounding boxes is changed to allow small changes to have a less dramatic effect on the predictions.
- Rather than predicting position and size directly, offsets are predicted for moving and reshaping the pre-defined anchor boxes relative to a grid cell and dampened by a logistic function.

Questions?