OBJECT DETECTION

INTRODUCTION

As a longstanding, fundamental and challenging problem in computer vision, object detection has been an active area of research for several decades. The goal of object detection is to determine whether or not there are any instances of objects from the given categories (such as humans, cars, bicycles, dogs and cats) in some given image and, if present, to return the spatial location and extent of each object instance

Object detection can be grouped into one of two types:

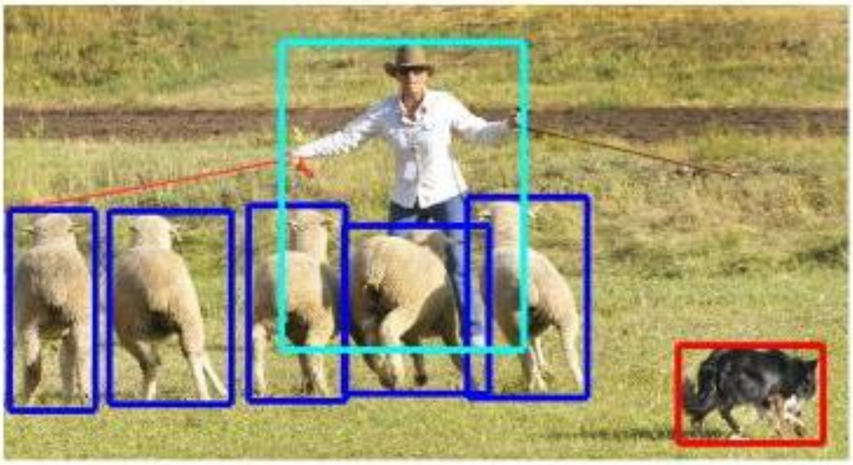
- detection of specific instance and detection of specific categories. The first type aims at detecting instances of a particular object (such as Donald Trump's face, the Pentagon building, or my dog Penny),
- whereas the goal of the second type is to detect different instances of predefined object categories (for example humans, cars, bicycles, and dogs).

Proposed Approach

Proposed Approach

Our goal is to create a scalable object detection system that can work across different object classes.

1. **Detection Method:** We use a Deep Neural Network (DNN) to predict bounding boxes around potential objects in an image. Each bounding box comes with a confidence score indicating how likely it contains an object.
2. **Model Details:**
   - **Bounding Boxes:** The DNN predicts the coordinates of each bounding box. These are normalized to adjust for different image sizes.
   - **Confidence Scores:** Each box has a confidence score between 0 and 1, indicating the likelihood it contains an object.
3. **Output:** The DNN produces a set number of bounding boxes (e.g., 100 or 200). We use confidence scores and non-maximum suppression to refine these to the most accurate boxes.
4. **Classification:** The refined boxes can be further classified using another DNN to identify the specific objects.
5. **Training:** We train the DNN to ensure the predicted boxes match well with the actual object locations in the training images. We optimize to improve the match and confidence of the correct boxes while reducing the confidence of incorrect ones.



(b) Object localization

Training Details

To improve how quickly and accurately our model detects objects, we use these three key steps:

1. **Use Clusters as References:** We group ground truth object locations into a fixed number of clusters. This helps the model learn to predict locations based on these clusters, making the learning process faster and more effective.
2. **Match with Clusters:** Instead of directly matching ground truth locations with our predictions, we match them with the cluster references. This helps the model make more diverse predictions and improves its performance.
3. **Class-Specific Adaptation:** Although the method works for any object, it can be adjusted to focus on specific classes (like dogs or cars) by training on examples of those classes. This keeps the model simple and effective, even when dealing with many classes.

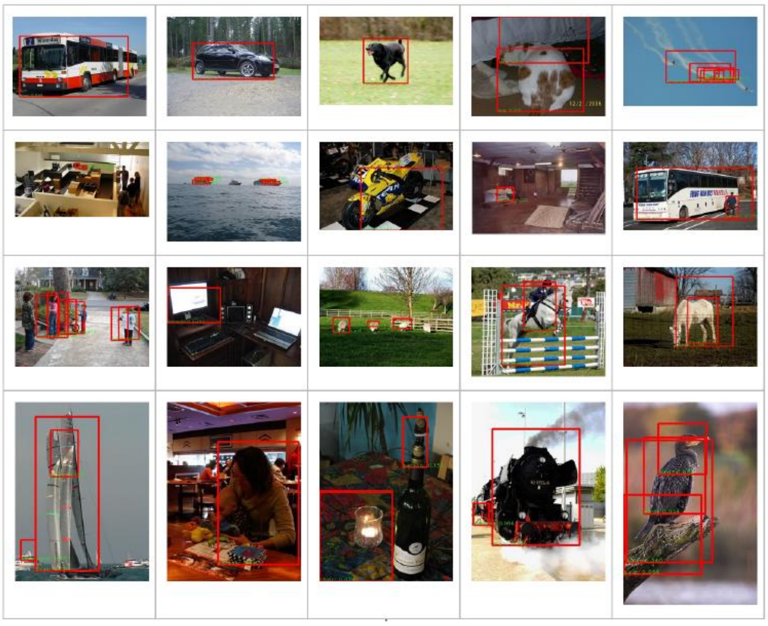


Figure 2. Sample of detection results on VOC 2007: up to 10 boxes from the class-agnostic detector are output, after non-max-suppression with Jaccard overlap 0.5 is performed.

VOC 2007

Training:

- **Data:** Trained on about 10 million cropped images of objects and 20 million negative (background) crops.
- **Method:** Used standard techniques and hyperparameters.

Evaluation:

1. **Localizer:** Applies to the central part of an image, resized to 220x220 pixels. It generates up to 100 candidate boxes.
2. **Non-Maximum Suppression:** Filters out overlapping boxes, keeping the top 10 highest-scoring ones.
3. **Classifier:** Each of these top boxes is then classified using a separate model.
4. **Final Score:** Combines the localizer and classifier scores to evaluate performance and generate precision-recall curves.