Object detection – general points

D.A. Forsyth

University of Illinois at Urbana Champaign

Classification vs detection

- Classification:
  - there is an X in this image
    - what
- Detection:
  - there is an X HERE in this image
    - what AND where
- Key issues
  - how to specify where
  - relationship between what and where
  - efficiency, etc
  - evaluation
    - surprisingly fiddly

Basic architecture

Need..

- to control the number of boxes checked
- to manage the classifier
  - may see lots of negatives
  - must be willing to respond when box isn’t precise
- to manage high-scoring boxes
  - non-maximum suppression
- to refine boxes

Three helpful points

- Correlated scores
  - boxes overlap, so box scores are correlated
- Extrapolating box scores
  - manageable, because they are correlated
  - bounding box regression
- Worthwhile windows
  - surprisingly, you can tell whether a box contains an object, even without specifying the object
  - means you can reduce the number of boxes you show to classifier

Two threads

- Localize then classify
  - find boxes that likely contain objects
  - decide what is in the box

- YOLO: Localize while classifying
  - in parallel, score
    - boxes for “goodness of box”
    - boxes for “what is in it”
  - combine

Evaluating detectors

- Compare detected boxes w ground truth boxes
- Favor
  - right number of boxes with right label in right place
- Penalize
  - awful lot of boxes
  - multiple detections of the same thing
- Strategy
  - Detector makes a ranked list of boxes
  - GT is a list of boxes
  - Mark detector boxes with relevant/irrelevant
  - summarize lists

IoU

The boxes that the detector predicts are unlikely to match ground truth exactly, and we need some way of telling whether the boxes are good enough. The standard method for doing this is to test the **IoU** (Intersection over Union). Write B_g for the ground truth box and B_p for the predicted box. The IoU is

$$\text{IoU}(B_p, B_g) = \frac{\text{Area}(B_g \cap B_p)}{\text{Area}(B_g \cup B_p)}.$$

Choose some threshold t. If $\text{IoU}(B_p, B_g) > t$, then B_p could match the ground truth box B_g.

Usually, t=0.5; higher t on occasion, but this sets a quite demanding standard for localization

Preventing double dipping

The detector should be credited for producing a box that has a high score and matches a ground truth box. But the detetector should not be able to improve its score by predicting many boxes on top of a ground truth box. The standard way to handle the problem is to mark the overlapping box with highest score `relevant`. The procedure is:

- Choose a threshold t.
- Order $\mathcal{D}$ by the score of each box, and mark every element of $\mathcal{D}$ with `irrelevant`. Choose a threshold t.

- For each element of $\mathcal{D}$ in order of score, compare that box against all ground truth boxes. If any ground truth box has IoU $> t$, mark the detector box `relevant` and remove that ground truth box from $\mathcal{G}$. Proceed until there are no more ground truth boxes.

Now every box in $\mathcal{D}$ is tagged either `relevant` or `irrelevant`.

Recall and precision

There are standard evaluations for search results like those produced by our detector. The first step is to merge the lists for each evaluation image into a single list of results. The **precision** of a set of search results $\mathcal{S}$ is given by

$$\mathbf{P}(\mathcal{S}) = \frac{\text{number of relevant search results}}{\text{total number of search results}}.$$

The **recall** is given by

$$\mathbf{R}(\mathcal{S}) = \frac{\text{number of relevant search results}}{\text{total number of relevant items in collection}}.$$

As you move down the list $\mathcal{D}$ in order of score, you get a new set of search results. The recall never decreases as the set gets larger, and so you could plot the precision as a function of recall (write $\mathbf{P}(\mathbf{R})$). These plots have a characteristic saw-tooth structure (Figure 18.9). If you add a single irrelevant item to the set of results, the precision will fall; if you then add a relevant item, it jumps up. The sawtooth

Interpolated precision

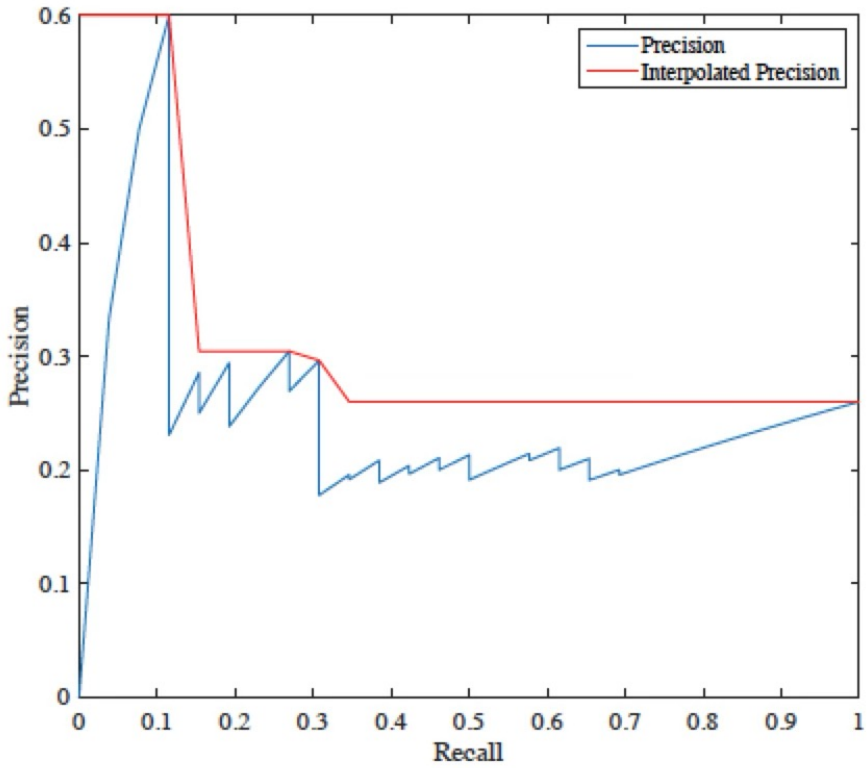

FIGURE 18.9: *Two plots for an imaginary search process. The precision plotted against recall shows a characteristic sawtooth shape. Interpolated precision measures the best precision you can get by increasing the recall, and so smoothes the plot. Interpolated precision is also a more natural representation of what one wants from search results – most people would be willing to add items to get higher precision. Interpolated precision is used to evaluate detectors.*

Interpolated precision

the precision will fall; if you then add a relevant item, it jumps up. The sawtooth doesn't really reflect how useful the set of results is — people are usually willing to add several items to a set of search results to improve the precision — and so it is better to use **interpolated precision**. The interpolated precision at some recall value R_0 is given by

$$\hat{\mathcal{P}}(R_0) = \max_{R \geq R_0} \mathbf{P}(R)$$

MaP

(Figure 18.9). By convention, the **average precision** is computed as

$$\frac{1}{11}\sum_{i=0}^{10}\hat{\mathcal{P}}(\frac{i}{10}).$$

This value summarizes the recall-precision curve. Notice this averages in interpolated precision at high recall. Doing so means a detector cannot get a high score by producing only very few, very accurate boxes — to do well, a detector should have high precision even when it is forced to predict every box.

Average precision evaluates detection for one category of object. The **mean average precision** (mAP) is the mean of the average precision for each category. The value depends on the IoU threshold chosen. One convention is to report mAP at $IoU = 0.5$. Another is to compute mAP at a set of 10 IoU values ($0.45 + i \times 0.05$ for $i \in 1 \ldots 10$), then average the mAP's. These evaluations produce numbers that tend to be bigger for better detectors, but it takes some practice to have a clear sense of what an improvement in mAP actually means.

Variants…

- Weight the mean by the frequency of objects
  - options:
    - emphasize common, easy objects (weight down rare)
    - emphasize rare objects (weight down common)
- IoU can be computed for things other than boxes

Think about this…

- Why is it possible to tell whether a box has an object in it with moderate accuracy, independent of the object identity?

- Why should something like bounding box regression work?