8.3. Networks for object detection

François Fleuret
https://fleuret.org/dlc/
The simplest strategy for object detection is to classify local regions, at multiple scales and locations.

This “sliding window” approach evaluates a classifier multiple times, and its computational cost increases with the prediction accuracy.

Notes

While image classification aims at predicting the class of the main object in the image, object detection aims at not only predicting the classes of all the objects which are visible, but also their locations.
This was mitigated in overfeat (Sermanet et al., 2013) by adding a regression part to predict the object’s bounding box.
In the single-object case, the convolutional layers are frozen, and the localization layers are trained with a \( L_2 \) loss.

Figure 7: Examples of bounding boxes produced by the regression network, before being combined into final predictions. The examples shown here are at a single scale. Predictions may be more optimal at other scales depending on the objects. Here, most of the bounding boxes which are initially organized as a grid, converge to a single location and scale. This indicates that the network is very confident in the location of the object, as opposed to being spread out randomly. The top left image shows that it can also correctly identify multiple locations if several objects are present. The various aspect ratios of the predicted bounding boxes shows that the network is able to cope with various object poses.

(\textit{Sermanet et al., 2013})

Combining the multiple boxes is done with an \textit{ad hoc} greedy algorithm.
This architecture can be applied directly to detection by adding a class "Background" to the object classes.

Negative samples are taken in each scene either at random or by selecting the ones with the worst miss-classification.

Surprisingly, using class-specific localization layers did not provide better results than having a single one shared across classes (Sermanet et al., 2013).
Other approaches evolved from AlexNet, relying on region proposals:

- Generate thousands of proposal bounding boxes with a non-CNN “objectness” approach such as Selective search (Uijlings et al., 2013),
- feed to an AlexNet-like network sub-images cropped and warped from the input image (“R-CNN”, Girshick et al., 2013), or from the convolutional feature maps to share computation (“Fast R-CNN”, Girshick, 2015).

These methods suffer from the cost of the region proposal computation, which is non-convolutional and not implementable on GPU.

They were improved by Ren et al. (2015) in “Faster R-CNN” by replacing the region proposal algorithm with a convolutional processing similar to Overfeat.
The most famous algorithm from this lineage is “You Only Look Once” (YOLO, Redmon et al. 2015).

It comes back to a classical architecture with a series of convolutional layers followed by a few fully connected layers. It is sometime described as “one shot” since a single information pathway suffices.

YOLO’s network is not a pre-existing one. It uses leaky ReLU, and its convolutional layers make use of the $1 \times 1$ bottleneck filters (Lin et al., 2013) to control the memory footprint and computational cost.
Making predictions. Unlike sliding window and region proposal-based techniques, YOLO sees the entire image during prediction. Unlike sliding window and region proposal-based techniques, YOLO sees the entire image during prediction.

Other than the size of the network, all training and testing parameters are the same between YOLO and Fast YOLO.

Our network architecture is inspired by the GoogLeNet work and evaluates it on the PASCAL VOC detection dataset. It divides the image into an S × S grid, here S = 7. Each cell S × S grid on input predicts a one set of class probabilities per grid cell, regardless of the number of boxes B number of boxes.

Each grid cell predicts B bounding boxes, confidence for those boxes, Class probability map, and the individual box confidence predictions, bounding boxes + confidence. We implement this model as a convolutional neural network with fewer convolutional layers (9 instead of the inception modules used by GoogLeNet), similar to Lin et al [22]. The full network is then trained using a regular S × S grid on input.

At test time we multiply the conditional class probability Pr(Class | Object) by the confidence Pr(Object). These predictions are encoded as a 3D tensor.

Our final prediction is a 7 × 7 × 30 tensor. The initial convolutional layers of the network extract features from the image while the fully connected layers predict the output probabilities and coordinates.

For evaluating YOLO on PASCAL VOC, we use the intersection over union (IOU) between the predicted box and any ground truth box. We only predict the confidence prediction represents the IOU between the predicted box and any ground truth box.

We also train a fast version of YOLO designed to push time speeds while maintaining high average precision. The YOLO design enables end-to-end training and real-time speeds while maintaining high average precision. This means our network reasons globally it is responsible for detecting the object whose center lay inside, if any.

The processing of a scene is done by splitting it into a regular S × S grid, here S = 7. Each cell is responsible for detecting the object whose center lay inside, if any.

(Redmon et al., 2015)
The output corresponds to splitting the image into a regular $S \times S$ grid, with $S = 7$, and for each cell, to predict a 30d vector:

- $B = 2$ bounding boxes coordinates and confidence,
- $C = 20$ class probabilities, corresponding to the classes of Pascal VOC.

(Redmon et al., 2015)
So the network predicts class scores and bounding-box regressions, and although the output comes from fully connected layers, it has a 2D structure.

It allows in particular YOLO to leverage the absolute location in the image to improve performance (e.g. vehicles tend to be at the bottom, umbrella at the top), which may or may not be desirable.
During training, YOLO makes the assumption that any of the $S^2$ cells contains at most [the center of] a single object. We define for every image, cell index $i = 1, \ldots, S^2$, predicted box index $j = 1, \ldots, B$ and class index $c = 1, \ldots, C$

- $1_{i}^{\text{obj}}$ is 1 if there is an object in cell $i$ and 0 otherwise,
- $1_{i,j}^{\text{obj}}$ is 1 if there is an object in cell $i$ and predicted box $j$ is the most fitting one, 0 otherwise.
- $p_{i,c}$ is 1 if there is an object of class $c$ in cell $i$, and 0 otherwise,
- $x_i, y_i, w_i, h_i$ the annotated object bounding box (defined only if $1_{i}^{\text{obj}} = 1$, and relative in location and scale to the cell),
- $c_{i,j}$ IOU between the predicted box and the ground truth target.
The training procedure first computes on each image the value of the $1_{i,j}^{\text{obj}}$ and $c_{i,j}$, and then does one step to minimize

$$
\lambda_{\text{coord}} \sum_{i=1}^{S^2} \sum_{j=1}^{B} 1_{i,j}^{\text{obj}} \left( (x_i - \hat{x}_{i,j})^2 + (y_i - \hat{y}_{i,j})^2 + \left( \sqrt{w_i} - \sqrt{\hat{w}_{i,j}} \right)^2 + \left( \sqrt{h_i} - \sqrt{\hat{h}_{i,j}} \right)^2 \right) 
+ \lambda_{\text{obj}} \sum_{i=1}^{S^2} \sum_{j=1}^{B} 1_{i,j}^{\text{obj}} (c_{i,j} - \hat{c}_{i,j})^2 + \lambda_{\text{noobj}} \sum_{i=1}^{S^2} \sum_{j=1}^{B} (1-1_{i,j}^{\text{obj}}) \hat{c}_{i,j}^2 
+ \lambda_{\text{classes}} \sum_{i=1}^{S^2} \sum_{c=1}^{C} (\hat{p}_{i,c} - \hat{\hat{p}}_{i,c})^2.
$$

where $\hat{p}_{i,c}$, $\hat{x}_{i,j}$, $\hat{y}_{i,j}$, $\hat{w}_{i,j}$, $\hat{h}_{i,j}$, $\hat{c}_{i,j}$ are the network’s outputs.

(slightly re-written from Redmon et al. 2015)

Notes

This loss is a good example of how one can mix multiple aspects of the prediction. The task has several purposes:

- bounding box coordinates, and confidence regression,
- classification.

The first part of the loss aims at minimizing the localization error of the detection. The square root is used to reduce the weight of the height and width $(w_i, h_i)$ over the corner location $(x_i, y_i)$.

The second part of the loss estimates the confidence of a detection $\hat{c}_{i,j}$ to reflect the intersection over union $c_{i,j}$ of that bounding box with the ground truth. This part of the loss is weighted by $\lambda - \text{obj}$. And when there is no object, we want the confidence $\hat{c}_{i,j}$ of that bounding box to be low, this part being driven by $\lambda_{\text{noobj}}$.

Finally, the last part of the loss is for the class score. Note that while a natural choice is the cross-entropy, it is here a quadratic error.
Training YOLO relies on many engineering choices that illustrate well how involved is deep-learning “in practice”:

- Pre-train the 20 first convolutional layers on ImageNet classification,
- use $448 \times 448$ input for detection, instead of $224 \times 224$,
- use Leaky ReLU for all layers,
- dropout after the first fully connected layer,
- normalize bounding boxes parameters in $[0, 1]$,
- use a quadratic loss not only for the bounding box coordinates, but also for the confidence and the class scores,
- reduce the weight of large bounding boxes by using the square roots of the size in the loss,
- reduce the importance of empty cells by weighting less the confidence-related loss on them,
- use momentum $0.9$, decay $5e^{-4}$,
- data augmentation with scaling, translation, and HSV transformation.

A critical technical point is the design of the loss function that articulates both a classification and a regression objectives.
The Single Shot Multi-box Detector (SSD, Liu et al., 2015) improves upon YOLO with a fully-convolutional architectures and multi-scale maps.
To summarize roughly how “one shot” deep detection can be achieved:

- networks trained on image classification capture localization information,
- regression layers can be attached to classification-trained networks,
- object localization does not have to be class-specific,
- multiple detection are estimated at each location to account for different aspect ratios and scales.
Object detection networks

- **AlexNet** (Krizhevsky et al., 2012)
  - Box regression
  - Region proposal + crop in image

- **Overfeat** (Sermanet et al., 2013)

- **R-CNN** (Girshick et al., 2013)
  - Crop in feature maps

- **Fast R-CNN** (Girshick, 2015)
  - Convolutional region proposal

- **Faster R-CNN** (Ren et al., 2015)
  - No crop

- **YOLO** (Redmon et al., 2015)
  - Multi-scale convolutions + multi boxes

- **SSD** (Liu et al., 2015)
  - Fully convolutional + multi-scale maps
References


