Deep learning

8.1. Computer vision tasks

François Fleuret

https://fleuret.org/dlc/

Computer vision tasks:

- classification,
- object detection,
- semantic or instance segmentation,
- other (tracking in videos, camera pose estimation, body pose estimation, 3d reconstruction, denoising, super-resolution, auto-captioning, synthesis, etc.)

Notes

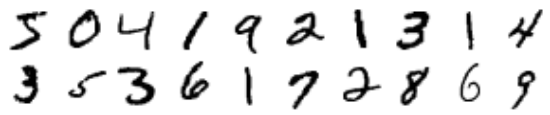
Many different tasks of computer vision heavily relies on deep-learning methods:

- Image classification: the input signal is an image and the goal is to predict a class among a set of predefined classes.
- Object detection: predict the locations of all the instances of the objects.
- Semantic segmentation: predict for each individual pixel the class of the object it belongs to, e.g. "car", "building", "person". Instance segmentation involves, in addition, to predict the object instance, e.g. "car #1", "car #2".
- Tracking in videos: detect objects and predict their trajectories.
- Camera pose estimation: predict the 3d pose of the sensor given an input image.
- Body pose estimation: predict the joint location of the individual body joints (e.g. shoulders, knees, ankles).
- 3d reconstruction: estimate the depth of individual pixels in the input image.
- Denoising and super-resolution.
- Auto-captioning: generate a text describing the content of the image.

We will focus on the first three.

"Small scale" classification data-sets.

MNIST and Fashion-MNIST: 10 classes (digits or pieces of clothing) $50,000$ train images, $10,000$ test images, 28×28 grayscale.



(LeCun et al., 1998; Xiao et al., 2017)

CIFAR10 and CIFAR100 (10 classes and 5×20 "super classes"), $50,000$ train images, $10,000$ test images, 32×32 RGB

(Krizhevsky, 2009, chap. 3)

ImageNet

http://www.image-net.org/

This data-set is build by filling the leaves of the "Wordnet" hierarchy, called "synsets" for "sets of synonyms".

- $21,841$ non-empty synsets,
- $14,197,122$ images,
- $1,034,908$ images with bounding box annotations.

ImageNet Large Scale Visual Recognition Challenge 2012

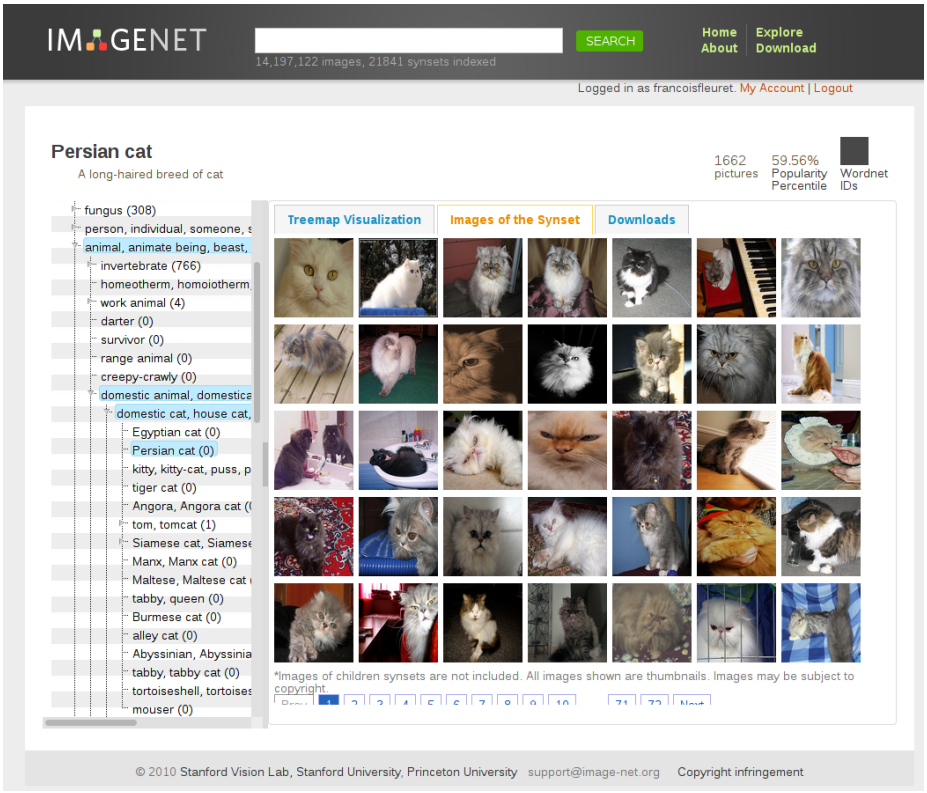
- $1,000$ classes taken among all synsets,
- $1,200,000$ training, and $50,000$ validation images.

Notes

"Wordnet" is a dictionary of words with relations between them, such as hypernymy ("being a kind of")

"animal" $\rightarrow$ "mammal" $\rightarrow$ "feline" $\rightarrow$ "cat".

It can be seen as a tree, whose leaves are each labeled with a list of synonyms.
Results on ImageNet are obtained on the Large Scale Visual Recognition Challenge subset of 1,200,000 training images and 1,000 classes.



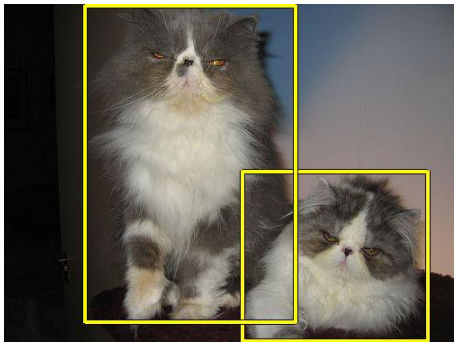
Notes

What is challenging with ImageNet, and maybe disputable, is that each image has only one label although may contain several objects, e.g. a cat laying on a chair.

n02123394_2084.xml

```
<annotation>
        <folder>n02123394</folder>
        <filename>n02123394_2084</filename>
        <source>
                <database>ImageNet database</database>
        </source>
        <size>
                <width>500</width>
                <height>375</height>
                <depth>3</depth>
        </size>
        <segmented>0</segmented>
        <object>
                <name>n02123394</name>
                <pose>Unspecified</pose>
                <truncated>0</truncated>
                <difficult>0</difficult>
                <bndbox>
                        <xmin>265</xmin>
                        <ymin>185</ymin>
                        <xmax>470</xmax>
                        <ymax>374</ymax>
                </bndbox>
        </object>
        <object>
                <name>n02123394</name>
                <pose>Unspecified</pose>
                <truncated>0</truncated>
                <difficult>0</difficult>
                <bndbox>
                        <xmin>90</xmin>
                        <ymin>1</ymin>
                        <xmax>323</xmax>
                        <ymax>353</ymax>
                </bndbox>
        </object>
</annotation>
```

n02123394_2084.JPEG



Notes

The text file on the left is the annotation file of the image on the right, and can be used both to train a classification model (class "persian cat") or a detection model (bounding boxes).

Cityscapes data-set

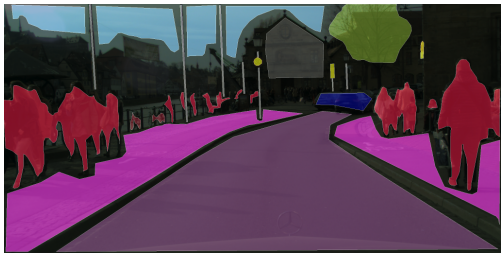
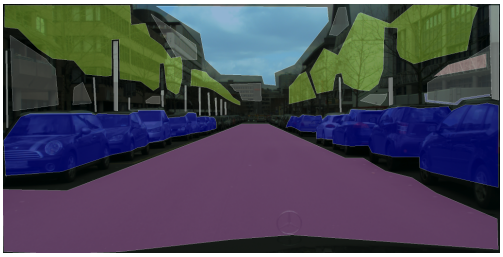
https://www.cityscapes-dataset.com/

Images from 50 cities over several months, each is the 20th image from a 30 frame video snippets (1.8s). Meta-data about vehicle position + depth.

- 30 classes
  - flat: road, sidewalk, parking, rail track
  - human: person, rider
  - vehicle: car, truck, bus, on rails, motorcycle, bicycle, caravan, trailer
  - construction: building, wall, fence, guard rail, bridge, tunnel
  - object: pole, pole group, traffic sign, traffic light
  - nature: vegetation, terrain
  - sky: sky
  - void: ground, dynamic, static
- $5,000$ images with fine annotations
- $20,000$ images with coarse annotations.

Cityscapes fine annotations ($5,000$ images)


Cityscapes coarse annotations ($20,000$ images)

Performance measures

Image classification consists of predicting the input image's class, which is often the class of the "main object" visible in it.

The standard performance measures are:

- The **error rate** $\hat{P}(f(X) \neq Y)$ or conversely the **accuracy** $\hat{P}(f(X) = Y)$,
- the **balanced error rate** (BER) $\frac{1}{C} \sum_{y=1}^{C} \hat{P}(f(X) \neq Y \mid Y = y)$.

Notes

In real world scenarios, the classes may be unbalanced.
For instance in a quality-control application there may be 99% of properly manufactured parts, and only 1% of defective ones. In such a case, always predicting "correct" would yields 99% accuracy which does not make any sense.

In the two-class case, we can define the True Positive (TP) rate as $\hat{P}(f(X) = 1 \mid Y = 1)$ and the False Positive (FP) rate as $\hat{P}(f(X) = 1 \mid Y = 0)$.

The ideal algorithm would have TP $\simeq 1$ and FP $\simeq 0$.

Most of the algorithms produce a continuous score (e.g. difference of logits), and make a hard decision with a threshold that is application-dependent. E.g.

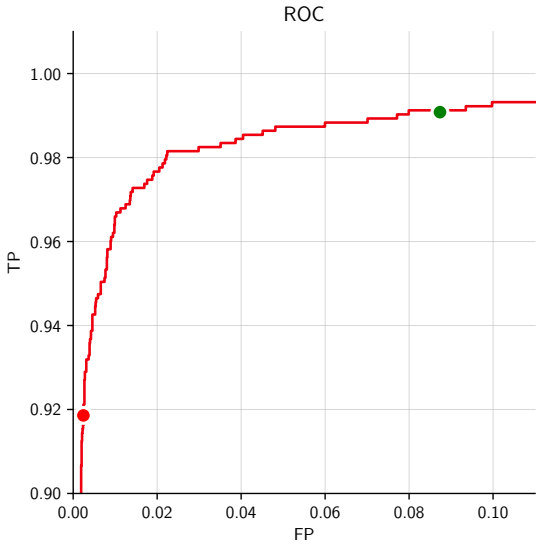
- **Cancer detection:** Low threshold to get a high TP rate (you do not want to miss a cancer), at the cost of a high FP rate (it will be double-checked by an oncologist anyway),
- **Image retrieval:** High threshold to get a low FP rate (you do not want to bring an image that does not match the request), at the cost of a low TP rate (you have so many images that missing a lot is not an issue).

Notes

The empirical true positive rate is the proportion of samples of class 1 classified [correctly] as being of class 1.
The empirical false positive rate is proportion of samples of class 0 classified [incorrectly] as being of class 1.

In that case, a standard performance representation is the **Receiver operating characteristic** (ROC) that shows performance at multiple thresholds.

It is the minimum increasing function above the True Positive (TP) rate $\hat{P}(f(X) = 1 \mid Y = 1)$ vs. the False Positive (FP) rate $\hat{P}(f(X) = 1 \mid Y = 0)$.



A standard measure is the **area under the curve** (AUC).

Notes

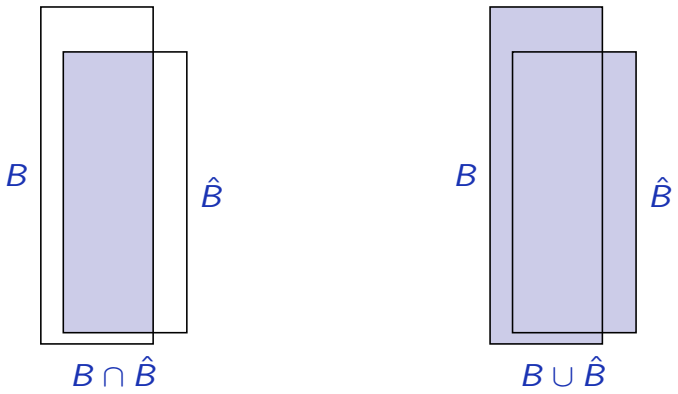
The curve shown here was generated by training a small convnet on MNIST samples to separate images of "7" (class 1) from all other digits (class 0). The ROC curve shows the true positive rate as a function of the false positive rate.
Each position of the ROC corresponds to a threshold, not shown, which is the value above which the sample is predicted to be of class 1:

- with a high threshold, the predictor classifies virtually all samples as class 0, so the true positive rate is low, but the false positive rate is also low (e.g. red dot).
- with a low threshold, the predictor classifies virtually all samples as class 1, so the true positive rate is higher, but so is the false positive rate (e.g. green dot).

Object detection aims at predicting **classes and locations** of targets in an image. The notion of "location" is ill-defined. In the standard setup, the output of the predictor is a series of bounding boxes, each with a class label.

A standard performance assessment considers that a predicted bounding box $\hat{B}$ is correct if there is an annotated bounding box B for that class, such that the **Intersection over Union** (IoU) is large enough

$$\frac{area(B \cap \hat{B})}{area(B \cup \hat{B})} \geq \frac{1}{2}.$$



Notes

When the bounding boxes B and $\hat{B}$ are disjoint, the intersection over union is 0.
When the bounding boxes are identical, the intersection over union is 1.
The standard threshold to consider a detection as correct is 0.5.

Image segmentation consists of labeling individual pixels with the class of the object it belongs to, and may also involve predicting the instance it belongs to.

The standard performance measure frames the task as a classification one. For VOC2012, the **segmentation accuracy** (SA) for a class c is defined as

$$SA = \frac{N_{Y=c,\hat{Y}=c}}{N_{Y=c,\hat{Y}=c} + N_{Y\neq c,\hat{Y}=c} + N_{Y=c,\hat{Y}\neq c}},$$

where N_α is the number of pixel with the property α, Y the real class of a pixel, and $\hat{Y}$ the predicted one.

All these performance measures are debatable, and in practice they are highly application-dependent.

In spite of their weaknesses, the ones adopted as standards by the community enable an assessment of the field's "long-term progress".

References

A. Krizhevsky. **Learning multiple layers of features from tiny images**. Master's thesis, Department of Computer Science, University of Toronto, 2009.

Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. **Gradient-based learning applied to document recognition**. Proceedings of the IEEE, 86(11):2278–2324, 1998.

H. Xiao, K. Rasul, and R. Vollgraf. **Fashion-MNIST: a novel image dataset for benchmarking machine learning algorithms**. CoRR, abs/1708.07747, 2017.