EE-559 – Deep learning

7.1. Computer vision tasks

François Fleuret

https://fleuret.org/ee559/

Wed Aug 29 14:58:02 UTC 2018

Computer vision tasks and data-sets

Computer vision tasks:

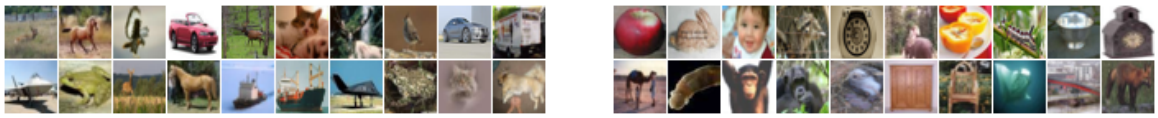
- classification,
- object detection,
- semantic or instance segmentation,
- other (tracking in videos, camera pose estimation, body pose estimation, 3d reconstruction, denoising, super-resolution, auto-captioning, synthesis, etc.)

"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

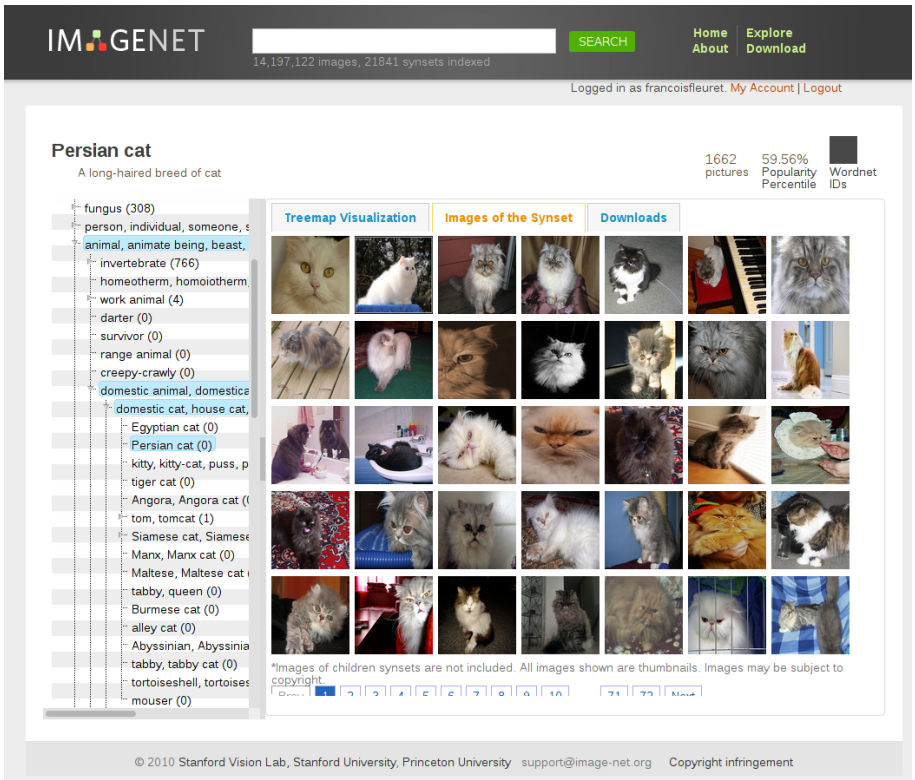
`http://www.image-net.org/`

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

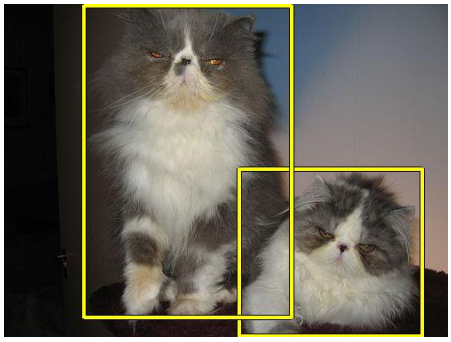
- $1,000$ classes taken among all synsets,
- $1,200,000$ training, and $50,000$ validation images.



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



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)

Tasks and performance measures

Image classification consists of predicting its 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) $\frac{1}{C} \sum_{y=1}^{C} \hat{P}(f(X) \neq y \mid Y = y)$.

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 $\hat{P}(f(X) = 1 \mid Y = 0)$.

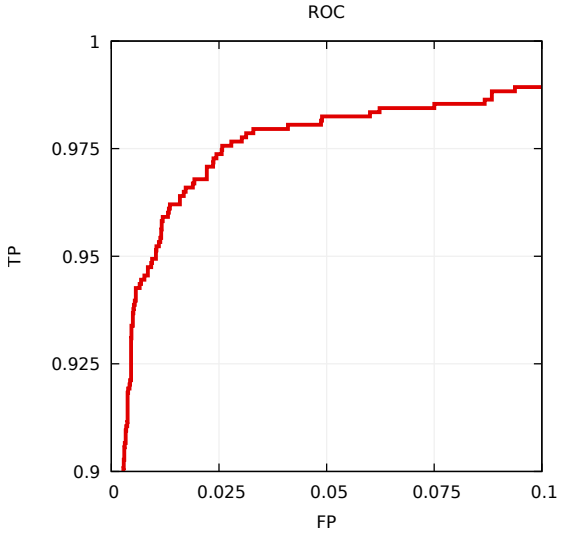
The ideal algorithm would have TP $\simeq 1$ and FP $\simeq 0$.

Most of the algorithms produce a score, and the decision threshold is application-dependent:

- 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 a 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).

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).

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

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

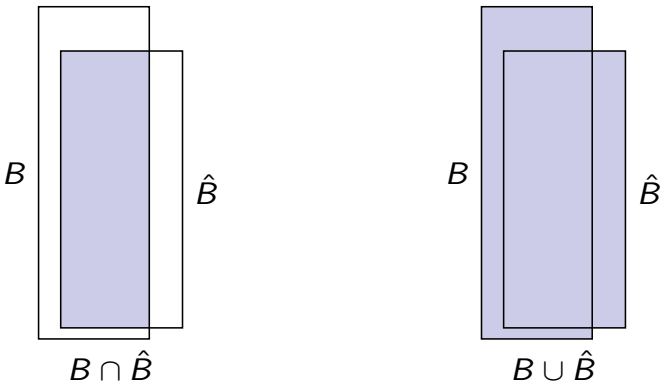


Image segmentation consists of labeling individual pixels with the class of the object it belongs to, and may also involve predicting the instance it belongs to.

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

$$SA = \frac{n}{n+e}$$

- n number of pixels of the right class, predicted as such,
- e number of pixels erroneously labeled.

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.