Dauphine | PSL
UNIVERSITÉ PARIS

Deep Learning for Image Analysis

Mehyar MLAWEH

🏫 Master BDIA - Dauphine Tunis
PhD student - Université PSL

Resources: https://mehyarmlaweh.github.io/mehyar-mlaweh/teaching

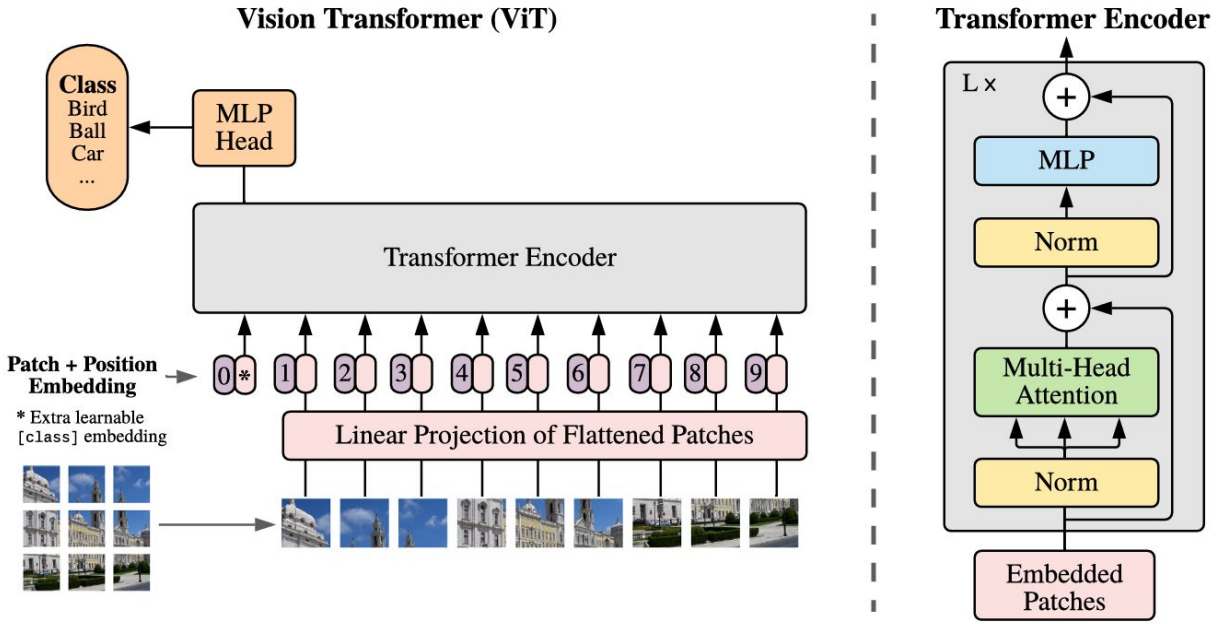
Module Overview

1. Computer Vision Foundations
2. CNNs
3. Transfer Learning
4. Vision Transformers
5. Object Detection
6. Segment Anything Model (SAM)
7. Final Project

Lecture 5 :

Object Detection

Last Time



Object Detection

The goal of object detection is to localize objects in an image and tell their class

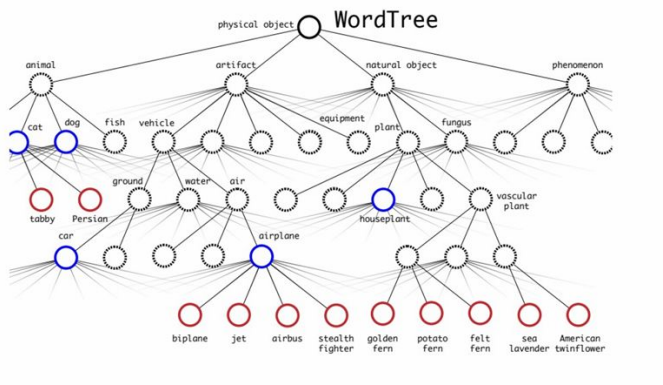
Localization: place a tight bounding box around object

Object Detection

The goal of object detection is to localize objects in an image and tell their class

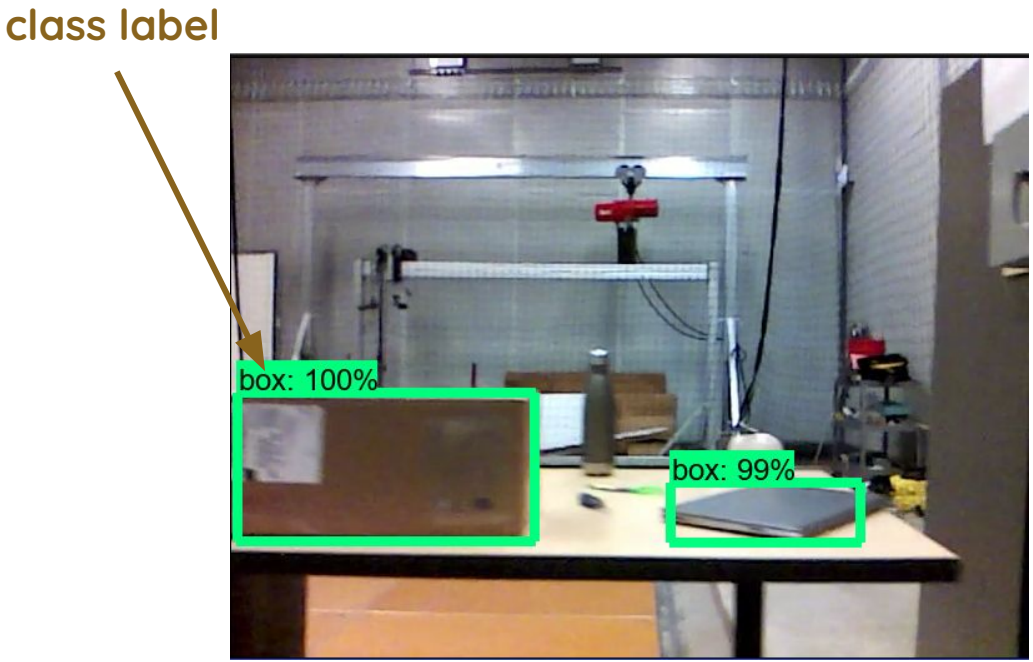
Localization: place a tight bounding box around object

Can scale up to many classes using hierarchical tree of visual concepts



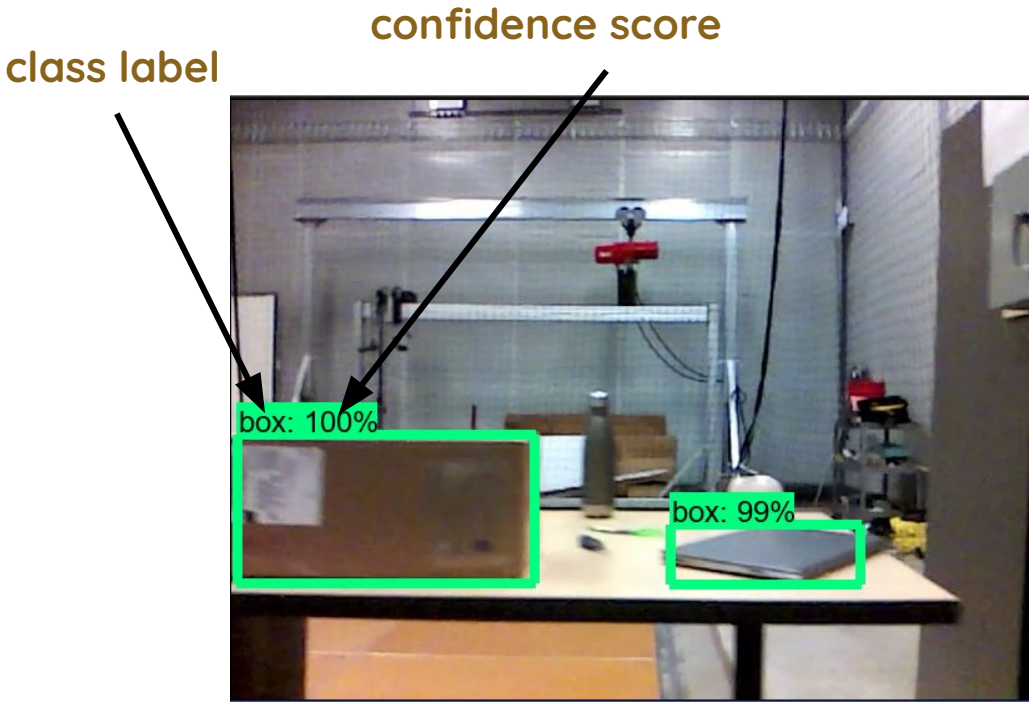
Object Detection

For each detected object, the model outputs:



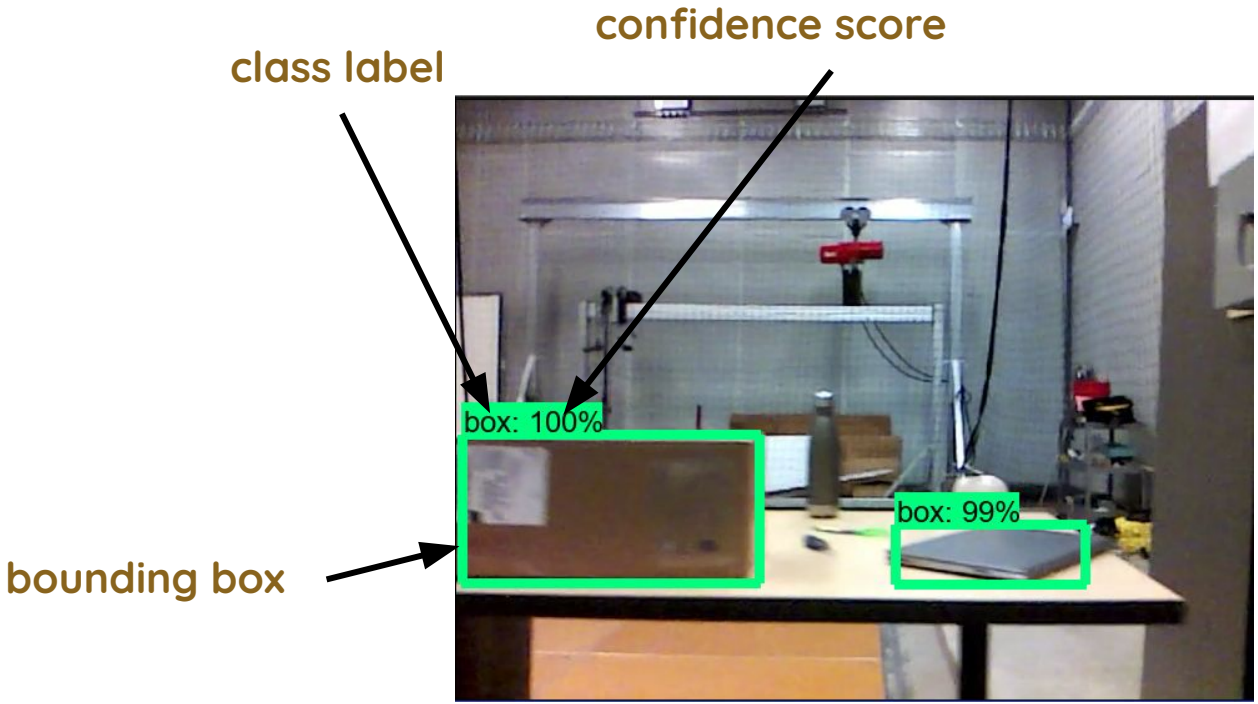
Object Detection

For each detected object, the model outputs:



Object Detection

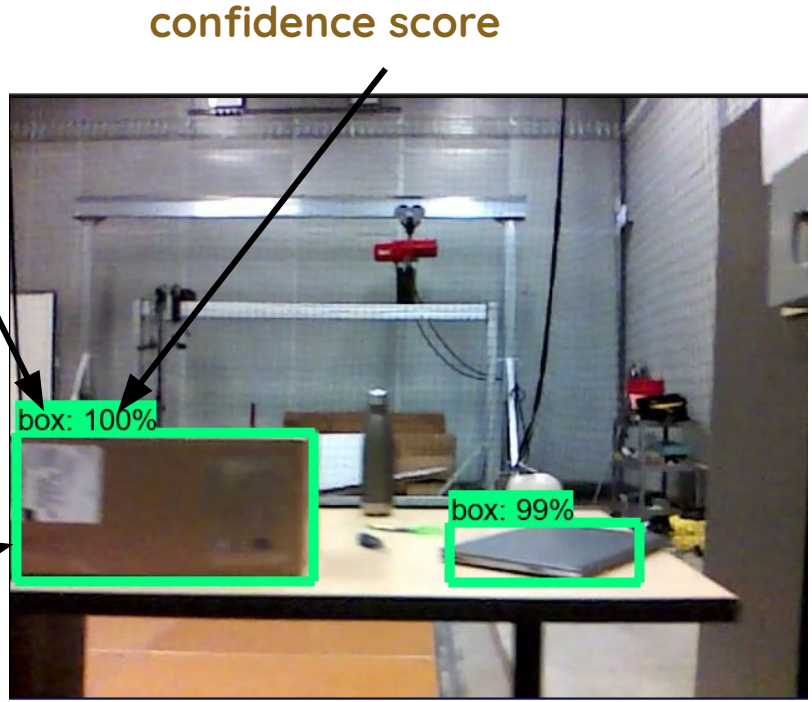
For each detected object, the model outputs:



Object Detection

For each detected object, the model outputs:

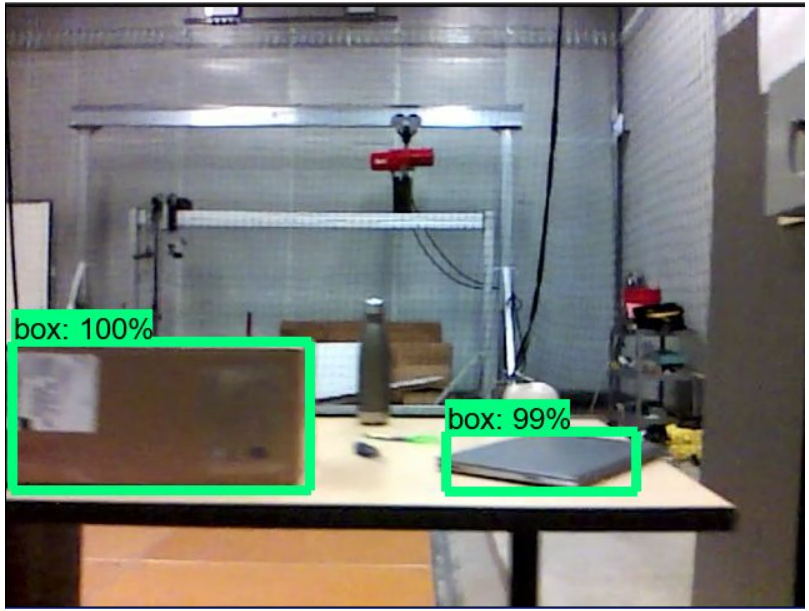
$$\forall \text{ object} : \begin{cases} P(C_k) & \forall k = 1, \ldots, K \\ \mathbf{b} = (b_x, b_y, b_W, b_H) \end{cases}$$



Object Detection

Measure performance

How good is a predicted bounding box ?



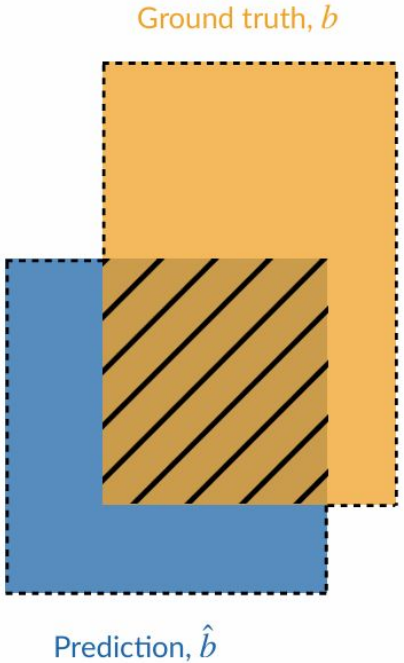
Measure performance

How good is a predicted bounding box ?

We calculate Intersection over Union:

$$\text{IoU}(\hat{b}, b) = \frac{\text{area}(\hat{b} \cap b)}{\text{area}(\hat{b} \cup b)}$$

Prediction correct if IoU > 0.5

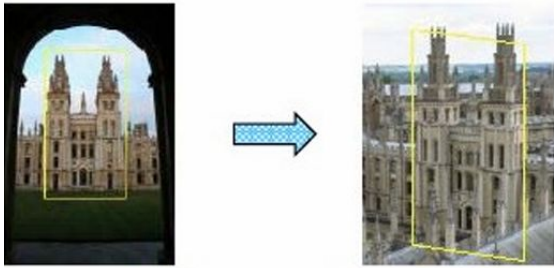


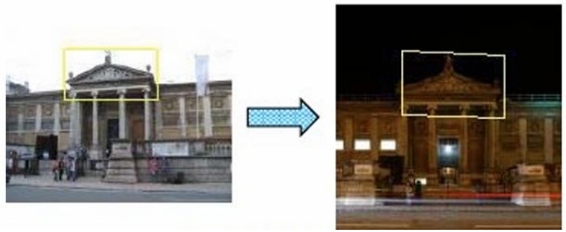
[Slide credit: Vito Dichio]

Object Detection

Major Difficulties :

- Objects appear at different scales, viewpoints, lightings ..





[Source: J. Sivic, slide credit: R. Urtasun]

Object Detection

Major Difficulties :

- Objects appear at different scales, viewpoints, lightings ..
- Multiple objects can overlap

[Source: J. Sivic, slide credit: R. Urtasun]

Object Detection

Major Difficulties :

- Objects appear at different scales, viewpoints, lightings ..
- Multiple objects can overlap
- Expensive annotation

[Source: J. Sivic, slide credit: R. Urtasun]

Naive solution: sliding window

Slide window and ask a classifier: “Is sheep in window or not?”

Confidence : 0.1

[Source: R. Urtasun]

Naive solution: sliding window

Slide window and ask a classifier: “Is sheep in window or not?”

Confidence : -0.2

[Source: R. Urtasun]

Naive solution: sliding window

Slide window and ask a classifier: “Is sheep in window or not?”

Confidence : -0.1

[Source: R. Urtasun]

Naive solution: sliding window

Slide window and ask a classifier: “Is sheep in window or not?”

Confidence : 1.5

[Source: R. Urtasun]

Naive solution: sliding window

Slide window and ask a classifier: “Is sheep in window or not?”

Confidence : 0.5

[Source: R. Urtasun]

Naive solution: sliding window

Slide window and ask a classifier: “Is sheep in window or not?”

Results

Confidence

-0.1
0.2
-0.1
…
1.5
…
0.5
0.2
-0.1

[Source: R. Urtasun]

Naive solution: sliding window

Downsides

- Computationally expensive
- Inaccurate bounding box : None of the squares could match the actual size of the object perfectly; each square may be smaller or larger.
- More than one object within same grid

[Source: R. Urtasun]

Smarter Solution: Region proposals

[Source: R. Urtasun]

Smarter Solution: Region proposals

Group pixels into object-like regions

[Source: R. Urtasun]

Smarter Solution: Region proposals

Generate many different regions

[Source: R. Urtasun]

Smarter Solution: Region proposals

Generate many different regions

[Source: R. Urtasun]

Smarter Solution: Region proposals

Generate many different regions

[Source: R. Urtasun]

Smarter Solution: Region proposals

Generate many different regions

[Source: R. Urtasun]

Smarter Solution: Region proposals

Generate many different regions

[Source: R. Urtasun]

Smarter Solution: Region proposals

The hope is that at least a few will cover real objects

[Source: R. Urtasun]

Smarter Solution: Region proposals

Select a region

[Source: R. Urtasun]

Smarter Solution: Region proposals

Crop out an image patch around it, throw to classifier (e.g., Neural Net)

classifier

'Dog' or not ?

Confidence : -0.25

[Source: R. Urtasun]

Smarter Solution: Region proposals

Crop out an image patch around it, throw to classifier (e.g., Neural Net)

classifier

'Dog' or not ?

Confidence : -0.25

[Source: R. Urtasun]

Smarter Solution: Region proposals

Crop out an image patch around it, throw to classifier (e.g., Neural Net)

classifier

'Dog' or not ?

Confidence : 1.5

[Source: R. Urtasun]

Smarter Solution: Region proposals

Crop out an image patch around it, throw to classifier (e.g., Neural Net)

classifier

'Dog' or not ?

A DOG !!

Confidence : 1.5

[Source: R. Urtasun]

Smarter Solution: Region proposals

We generate many candidate regions that may contain objects, then classify each region.

Smarter Solution: Region proposals

Downsides

- Too many candidate regions
- Multi-stage pipeline: image → generate proposals → crop/resize regions → classify each region → refine boxes → remove duplicates
- Slow for real-time applications

[Source: R. Urtasun]

DL Solution: R-CNN (Regions with CNN)

It uses region proposals, then applies a CNN to each proposed region to extract features and classify the object

CNN

'Dog' or not ?

A DOG !!

Confidence : 1.5

[Source: R. Urtasun]

DL Solution: R-CNN (Regions with CNN)

Still slow !!

CNN

‘Dog’ or not ?

A DOG !!

Confidence : 1.5

[Source: R. Urtasun]

DL Solution: R-CNN (Regions with CNN)

Still slow !!

because the CNN is applied many times

CNN

'Dog' or not ?

A DOG !!

Confidence : 1.5

[Source: R. Urtasun]

DL Solution: R-CNN (Regions with CNN)

So the main problem is the number of iteration we are doing

CNN

'Dog' or not ?

A DOG !!

Confidence : 1.5

[Source: R. Urtasun]

DL Solution: R-CNN (Regions with CNN)

So the main problem is the number of iteration we are doing

We only need to look once to make it faster

CNN

'Dog' or not ?

A DOG !!

Confidence : 1.5

[Source: R. Urtasun]

YOLO: You Only Look Once

You Only Look Once:
Unified, Real-Time Object Detection

Joseph Redmon*, Santosh Divvala*†, Ross Girshick¶, Ali Farhadi*†
University of Washington*, Allen Institute for AI†, Facebook AI Research¶
http://pjreddie.com/yolo/

Abstract

We present YOLO, a new approach to object detection. Prior work on object detection repurposes classifiers to perform detection. Instead, we frame object detection as a regression problem to spatially separated bounding boxes and associated class probabilities. A single neural network predicts bounding boxes and class probabilities directly from full images in one evaluation. Since the whole detection pipeline is a single network, it can be optimized end-to-end directly on detection performance.

Our unified architecture is extremely fast. Our base YOLO model processes images in real-time at 45 frames per second. A smaller version of the network, Fast YOLO, processes an astounding 155 frames per second while still achieving double the mAP of other real-time detectors. Compared to state-of-the-art detection systems, YOLO makes more localization errors but is less likely to predict false positives on background. Finally, YOLO learns very general representations of objects. It outperforms other detection methods, including DPM and R-CNN, when generalizing from natural images to other domains like artwork.

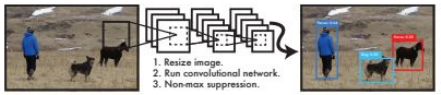


Figure 1: The YOLO Detection System. Processing images with YOLO is simple and straightforward. Our system (1) resizes the input image to 448 × 448, (2) runs a single convolutional network on the image, and (3) thresholds the resulting detections by the model's confidence.

methods to first generate potential bounding boxes in an image and then run a classifier on these proposed boxes. After classification, post-processing is used to refine the bounding boxes, eliminate duplicate detections, and rescore the boxes based on other objects in the scene [13]. These complex pipelines are slow and hard to optimize because each individual component must be trained separately.

We reframe object detection as a single regression problem, straight from image pixels to bounding box coordinates and class probabilities. Using our system, you only look once (YOLO) at an image to predict what objects are

[Source: https://arxiv.org/pdf/1506.02640]

Segment Anything Model

References I

Books

• Computer Vision: Models, Learning, and Inference (2012). https://amzn.to/2rxrdOF
• Programming Computer Vision with Python (2012). https://amzn.to/2QKTaAL
• Multiple View Geometry in Computer Vision (2004). https://amzn.to/2LfHLE8
• Computer Vision: Algorithms and Applications (2010). https://amzn.to/2LcIt4J
• Computer Vision: A Modern Approach (2002). https://amzn.to/2rv7AqJ
• Deep Learning for Computer Vision : Image Classification, Object Detection and Face Recognition in Python(2019). https://machinelearningmastery.com/deep-learning-for-computer-vision/
• Hands-On Machine Learning with Scikit-Learn, Keras & TensorFlow — Aurélien Géron (2019) https://amzn.to/3uZbN6Z
• Hands-On Large Language Models - Jay Alammar & Grootendorst

Articles & Online Resources

• Krizhevsky, Sutskever, Hinton (2012) — ImageNet Classification with Deep CNNs (AlexNet — deep learning breakthrough)
• Improved Texture Networks: Maximizing Quality and Diversity in Feed-forward Stylization and Texture Synthesis, 2017
• "Computer vision," Wikipedia. https://en.wikipedia.org/wiki/Computer_vision
• "Machine vision," Wikipedia. https://en.wikipedia.org/wiki/Machine_vision
• "Digital image processing," Wikipedia. https://en.wikipedia.org/wiki/Digital_image_processing

References II

- "Computer vision," Wikipedia. https://en.wikipedia.org/wiki/Computer_vision
- https://medium.com/data-reply-it-datatech
- https://huggingface.co/blog/RDTvlokip/quand-l-ia-apprend-de-l-experience-comme-toi
- Zhuang et al., "A Comprehensive Survey on Transfer Learning," IEEE, 2020.
- https://jalammar.github.io/illustrated-transformer/
- https://sh-tsang.medium.com/review-vision-transformer-vit-406568603de0
- Laurier, Linda. (2025). Vision Transformers: From Patch-Based Processing to Visual Intelligence.
- "Training Convolutional Neural Networks: What Is Machine Learning" Analog Devices. https://www.analog.com/en/resources/analog-dialogue/articles/training-convolutional-neural-networks-what-is-machine-learning-part-2.html
- Roth, Peter M.. (2011). On-line Conservative Learning.