Human Robot Collaboration based on deep learning and generative IA : Applications to Industry 5.0

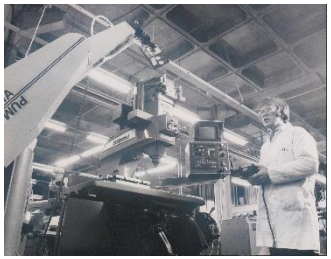
- A - Introduction
- B - Deep learning techniques
- C - Human Action Recognition
- D - Toward General-Purpose Robots via Foundation Models

Christophe SABOURIN
sabourin@u-pec.fr
IUT Sénart Fontainebleau _ Département GEII
Laboratoire Images, Signaux et Systèmes Intelligents (LISSI)

A - Introduction

A - Introduction

Unimate
(General Motor)



1961

Manipulator arm
(Ex: Fanuc, ABB, Stäubli, Kuka)

→1980 – Industrial robotic

Universal robot

2005

→ Cobot

1975

First AGV (Usine Volvo)

AGV
(Ex: Ba Systems, ABB, Kuka)

2005

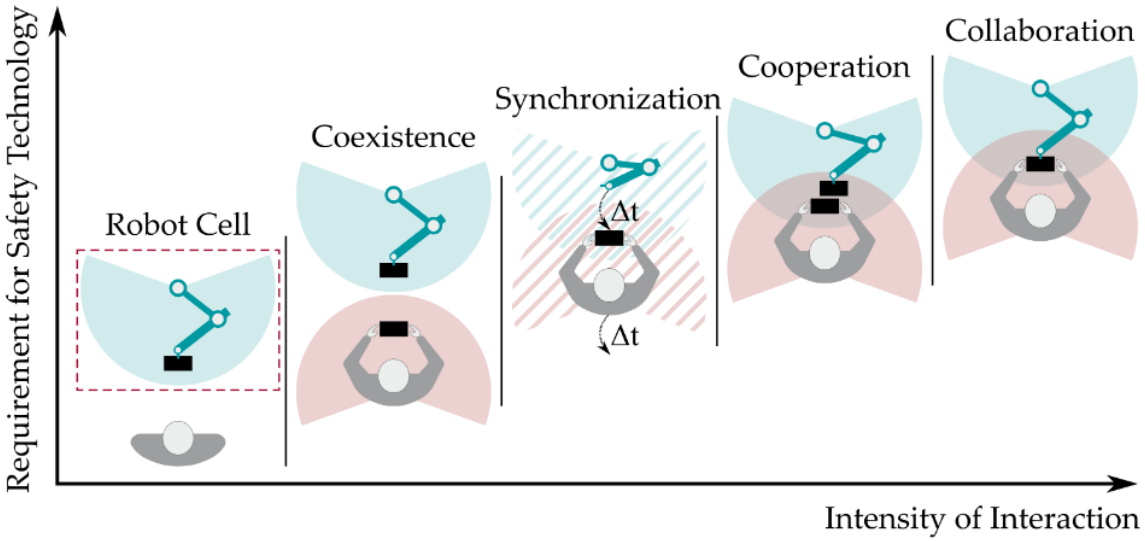
→ AMR

Kiva systems

Robots substitute humans — 20st century

Robots and humans live together — 21st century

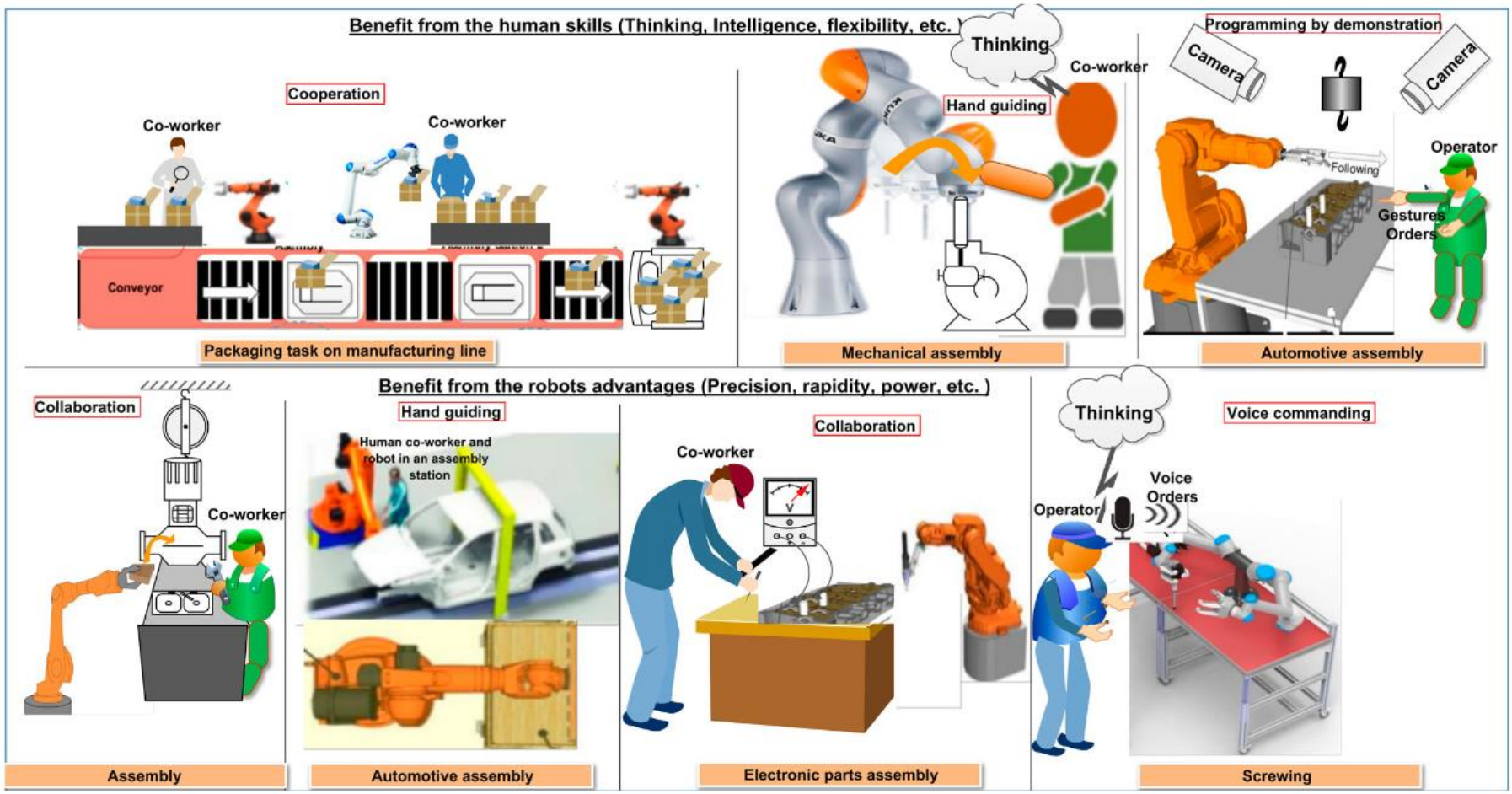
A - Types of collaboration in HRC scenarios.



From C. Weidemann (2023), https://doi.org/10.3390/robotics12030084

- **Coexistence**: human operator and cobot are in the same environment, but generally do not interact with each other,
- **Synchronised**: human operator and cobot work in the same workspace, but at different times,
- **Cooperation**: human operator and cobot work in the same workspace at the same time, though each focuses on separate tasks,
- **Collaboration**: human operator and cobot must execute a task together; the action of the one has immediate consequences on the other.

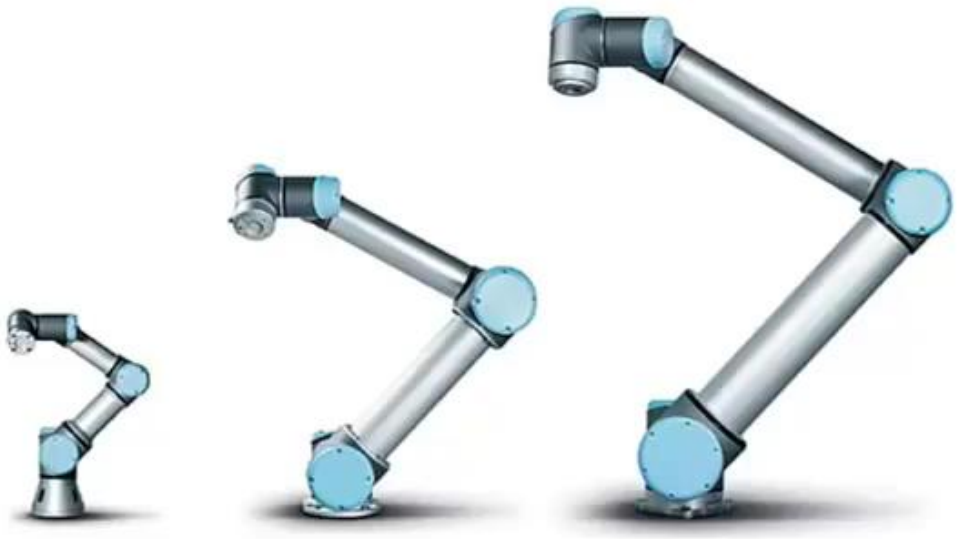
A- Examples of human-robot interaction and collaborative



From **A.** Hentout (2019), https://doi.org/10.1080/01691864.2019.1636714

A - Examples of human-robot interaction and collaborative

AUTOMATE VIRTUALLY ANYTHING WITH COBOT

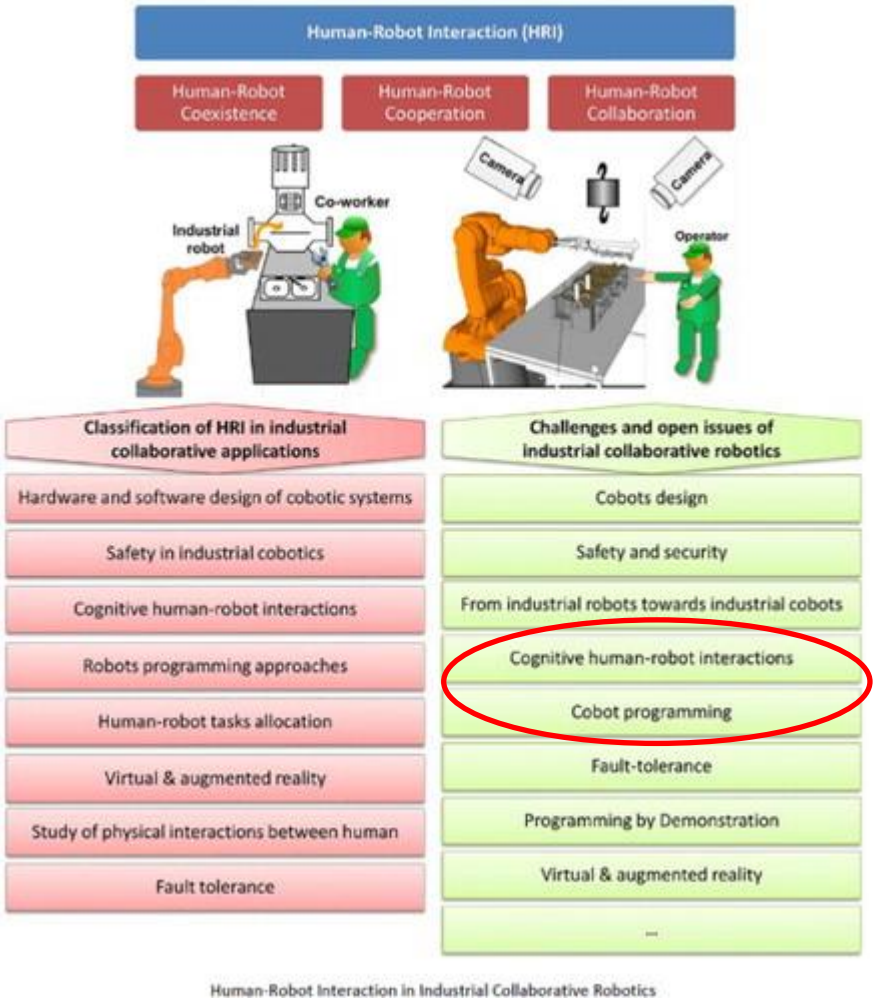


https://youtu.be/BE6lbnfDdrU?si=KM3s4qxJuVlXbt0X

A - Examples of human-robot interaction and collaborative

https://youtu.be/gQl-Efv6yMM?si=0HYNyLyd5_m8E38_

A- HRI – Challenges and open issues



Human-Robot Interaction in Industrial Collaborative Robotics

<u>Cognitive human–robot interactions</u>:

- Human actions recognition
- Gestures recognition
- Faces recognition
- Voice commanding
- Social gaze and social acceptance

<u>Robot programming approaches:</u>

- Generation of robotic skills
- On-line programming
- Programming by demonstration

From A. Hentout (2019), https://doi.org/10.1080/01691864.2019.1636714

A - Examples of human-robot interaction

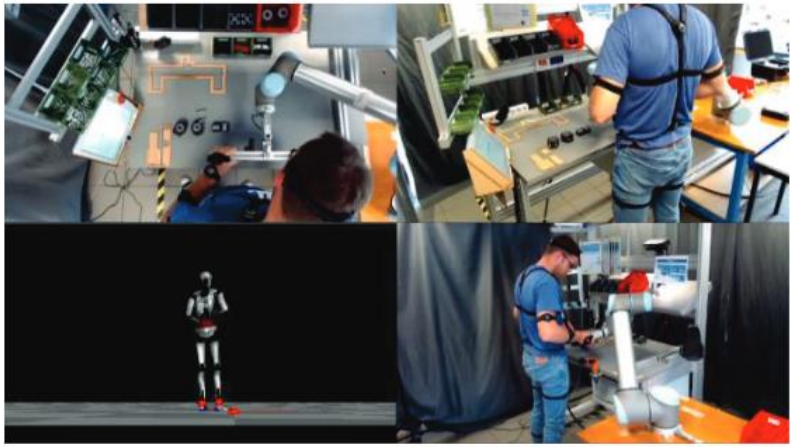
Exemple of physical interaction: co-manipulation framework for robot adaptation to human fatigue

From **L. Peternel (2016),**
https://ieeexplore.ieee.org/document/7803320

Exemple of cognitive interaction: Pointing gesture identification validation

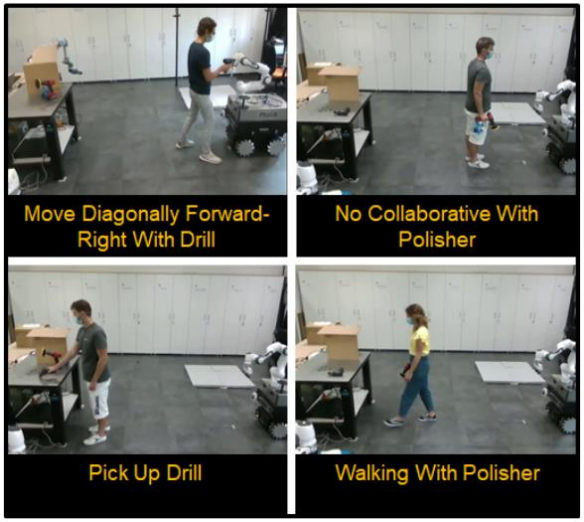
From **I. Maurtua (2017),**
https://journals.sagepub.com/doi/10.1177/1729881417716043

Dataset for Human Activity Recognition in industrial context



Interaction with robotic arm

From **M. Dallel (2020)**
doi: 10.1109/ICHMS49158.2020.9209531.

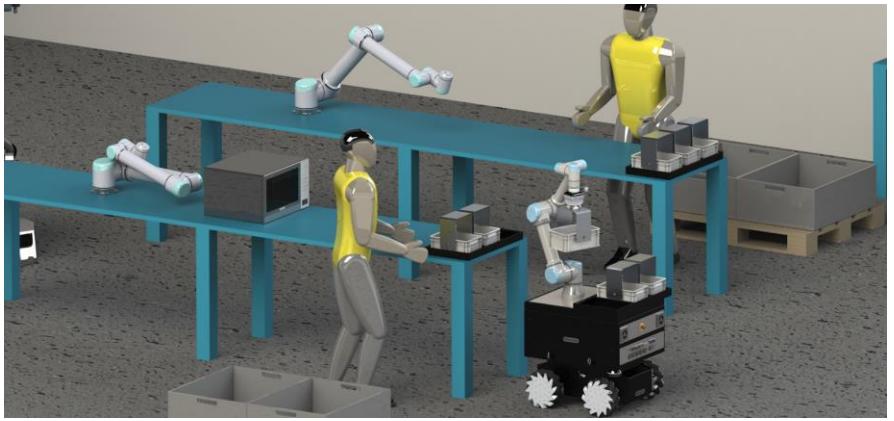


Interaction with mobile robot

From **F. Iodice (2022)**
doi: 10.1109/ICPR56361.2022.9956300.

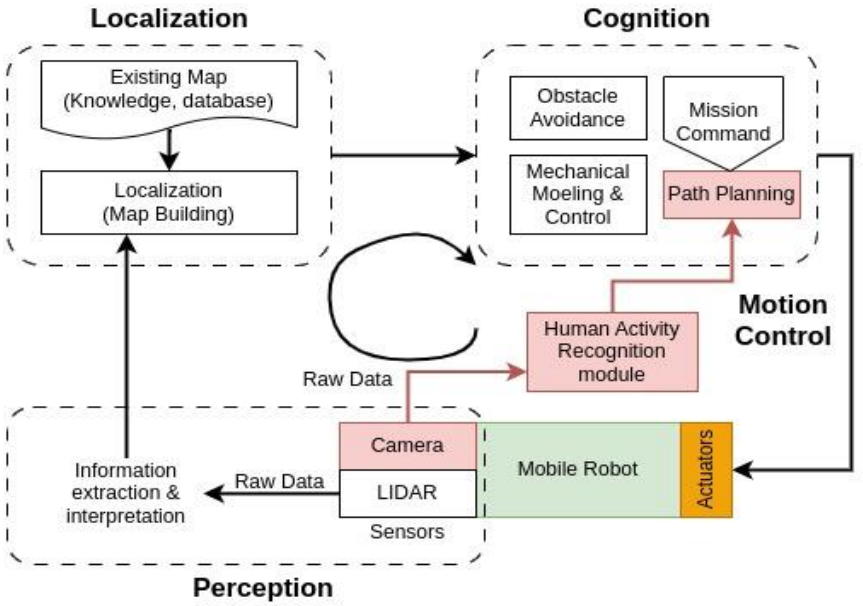
A – Collaboration ICAM / LISSI

Goal: Using mobile robotics to improve production performance

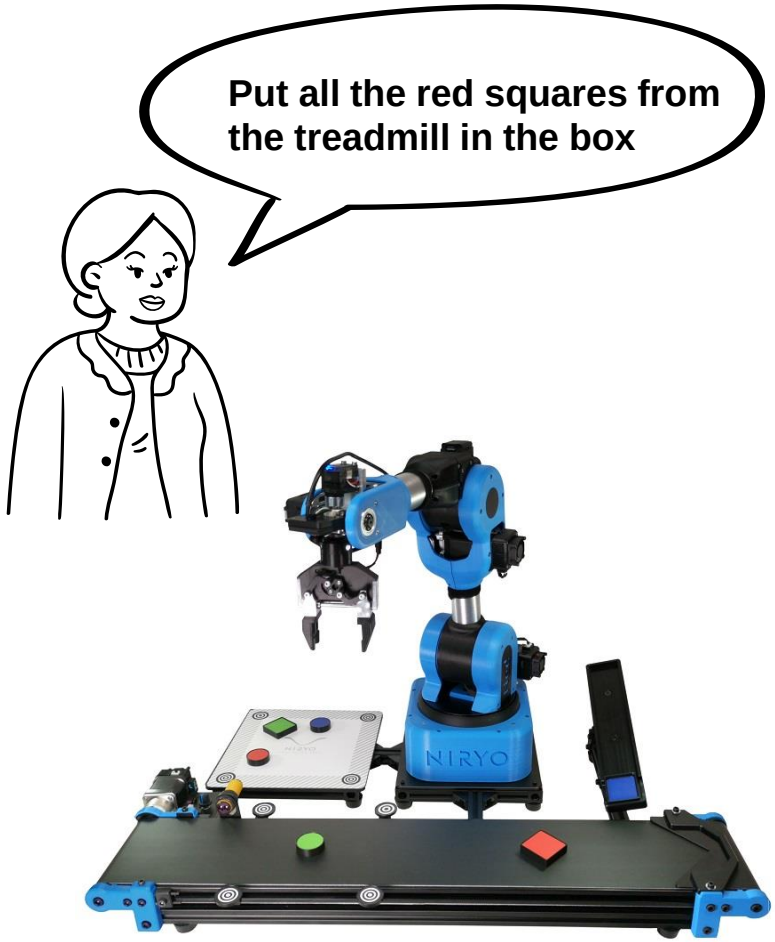


Scientific issues:

- Human Robot Interaction
- AMR Navigation
- Real time



Goal: Make easier collaboration between humans and cobots



Scientific issues:

- Human Robot Interaction
- Automatic code generation
- Real time

A – Bibliography

C. Weidemann et al., « Literature Review on Recent Trends and Perspectives of Collaborative Robotics in Work 4.0 », Robotics, vol. 12, nº 3, Art. nº 3, juin 2023, doi: 10.3390/robotics12030084.

A. Hentout, M. Aouache, A. Maoudj, et I. Akli, « Human–robot interaction in industrial collaborative robotics: a literature review of the decade 2008–2017 », *Advanced Robotics*, vol. 33, nº 15-16, p. 764-799, août 2019, doi: 10.1080/01691864.2019.1636714.

L. Peternel, N. Tsagarakis, D. Caldwell, et A. Ajoudani, « Adaptation of robot physical behaviour to human fatigue in human-robot co-manipulation », in *2016 IEEE-RAS 16th International Conference on Humanoid Robots (Humanoids)*, nov. 2016, p. 489-494. doi: 10.1109/HUMANOIDS.2016.7803320.

I. Maurtua *et al.*, « Natural multimodal communication for human–robot collaboration », *International Journal of Advanced Robotic Systems*, vol. 14, nº 4, p. 172988141771604, juill. 2017, doi: 10.1177/1729881417716043.

[1]

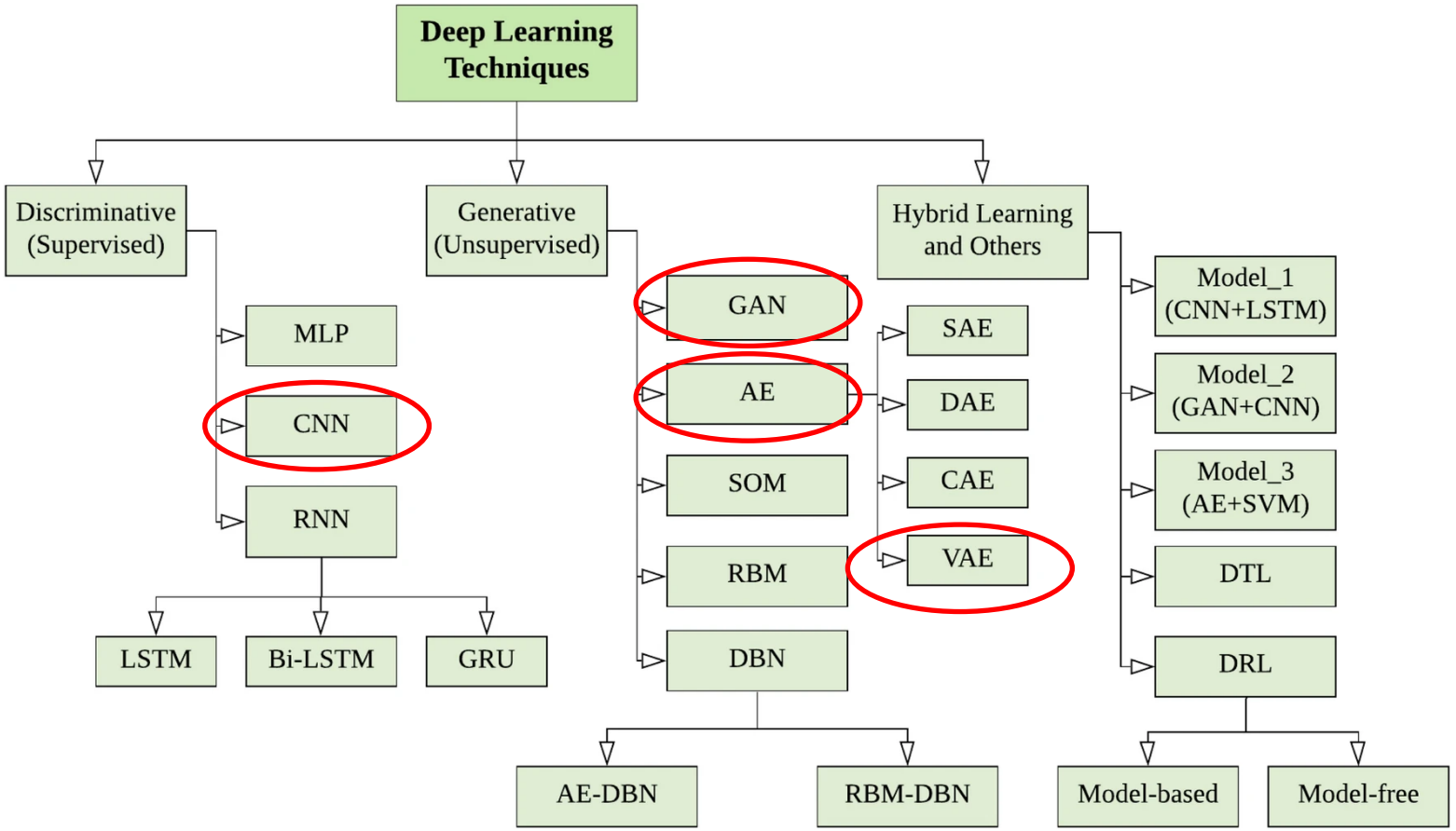
Z. Sun, Q. Ke, H. Rahmani, M. Bennamoun, G. Wang, et J. Liu, « Human Action Recognition from Various Data Modalities: A Review », *IEEE Trans. Pattern Anal. Mach. Intell.*, p. 1-20, 2022, doi: 10.1109/TPAMI.2022.3183112.

M. Dallel, V. Havard, D. Baudry, et X. Savatier, « InHARD - Industrial Human Action Recognition Dataset in the Context of Industrial Collaborative Robotics », in *2020 IEEE International Conference on Human-Machine Systems (ICHMS)*, Rome, Italy: IEEE, sept. 2020, p. 1-6. doi: 10.1109/ICHMS49158.2020.9209531.

F. Iodice, E. De Momi, et A. Ajoudani, « HRI30: An Action Recognition Dataset for Industrial Human-Robot Interaction », in *2022 26th International Conference on Pattern Recognition (ICPR)*, Montreal, QC, Canada: IEEE, août 2022, p. 4941-4947. doi: 10.1109/ICPR56361.2022.9956300.

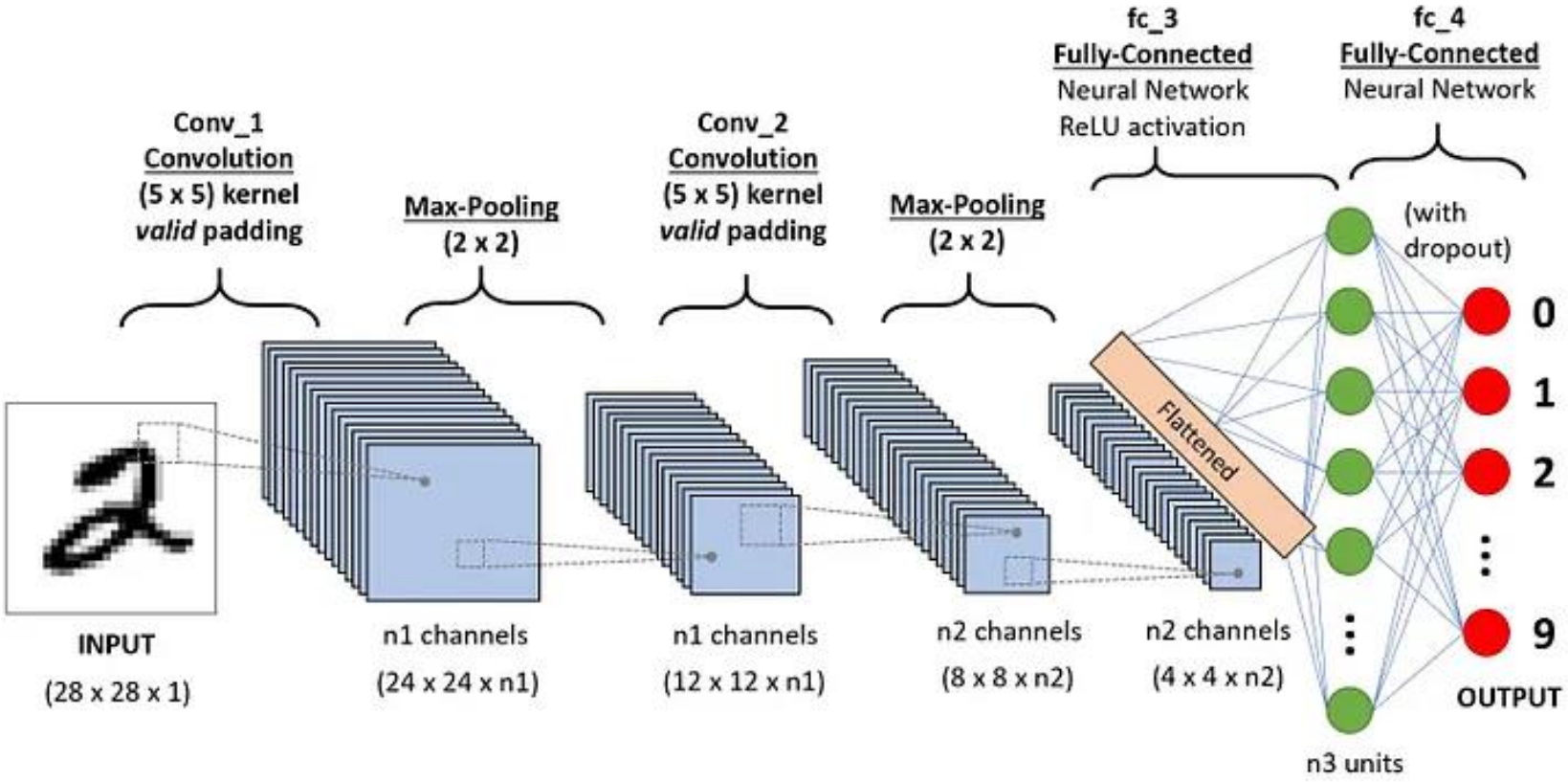
B - Deep learning techniques

B - Deep learning techniques



From **I. H. Sarker (2021)** https://link.springer.com/article/10.1007/s42979-021-00815-1

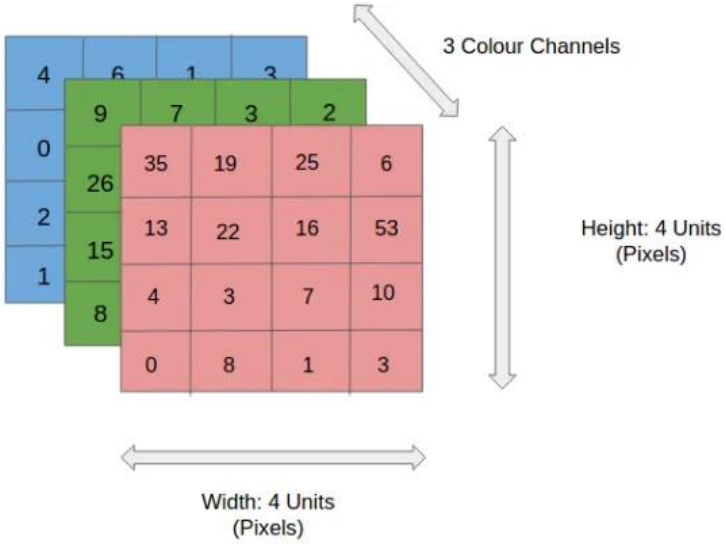
B - Example of a CNN sequence to classify handwritten digits



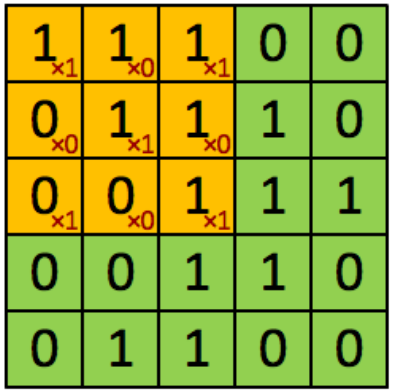
From https://towardsdatascience.com/a-comprehensive-guide-to-convolutional-neural-networks-the-eli5-way-3bd2b1164a53

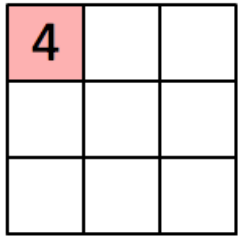
B - Convolution operation

Input Image



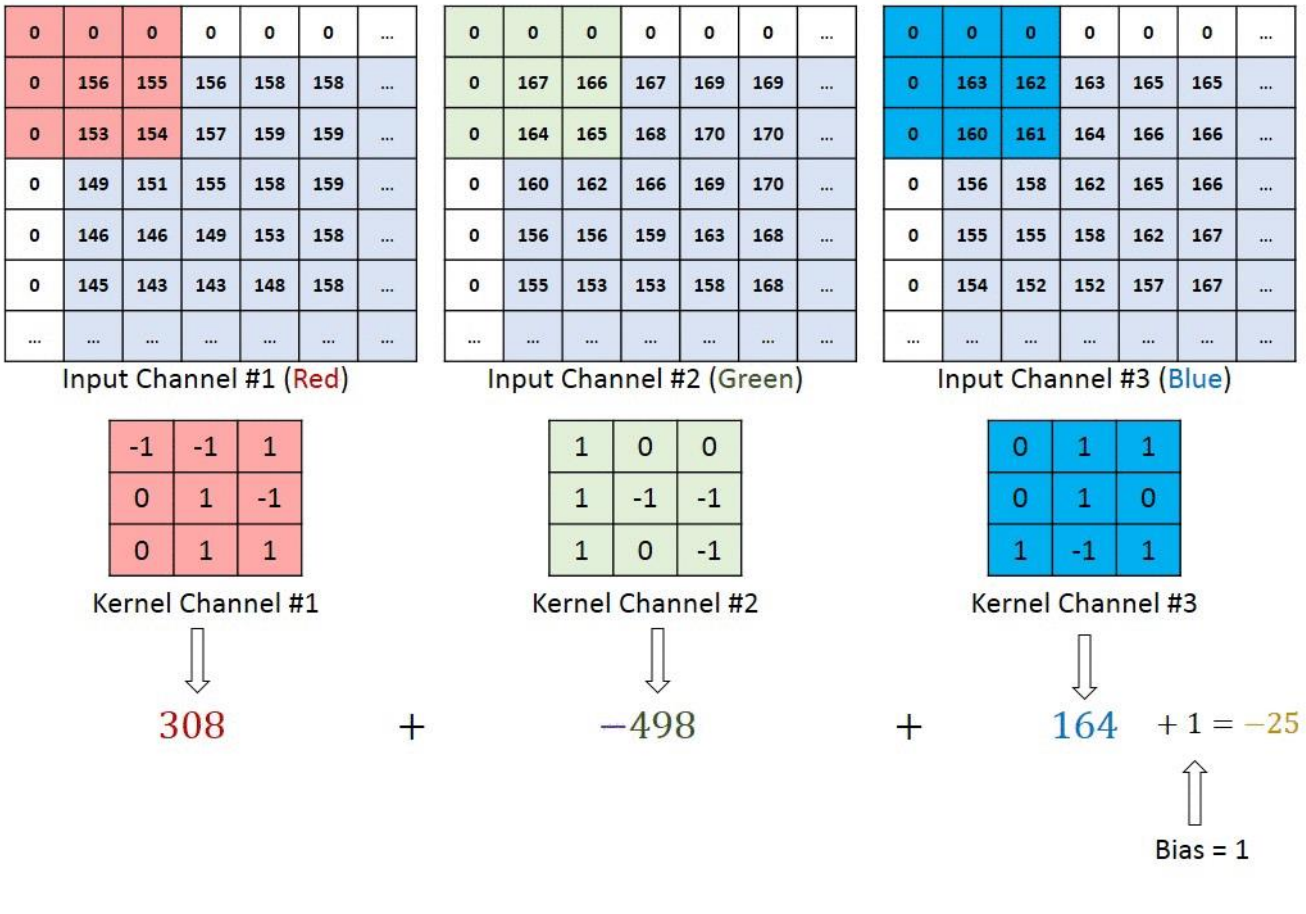
Convolution Layer





https://towardsdatascience.com/a-comprehensive-guide-to-convolutional-neural-networks-the-eli5-way-3bd2b1164a53

B - Convolution operation



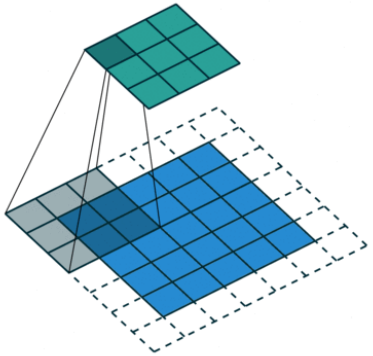
Convolution operation

MxNx3 image matrix
with a 3x3x3 Kernel

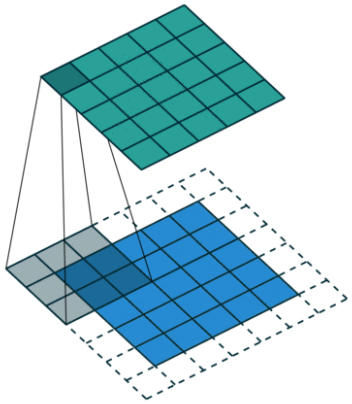
https://towardsdatascience.com/a-comprehensive-guide-to-convolutional-neural-networks-the-eli5-way-3bd2b1164a53

B - Convolution operation

Convolution Operation with **Stride Length = 2**



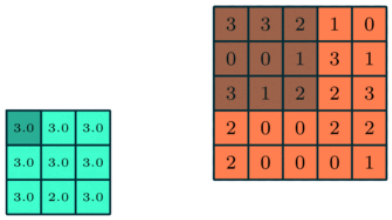
Padding (SAME padding 5x5x1 image is padded with 0s to create a 6x6x1 image)

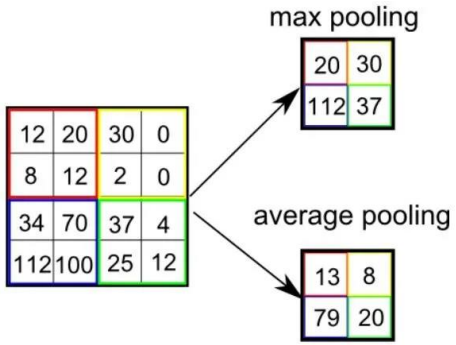


Understanding Padding in Machine Learning

https://deepai.org/machine-learning-glossary-and-terms/padding

Pooling Layer

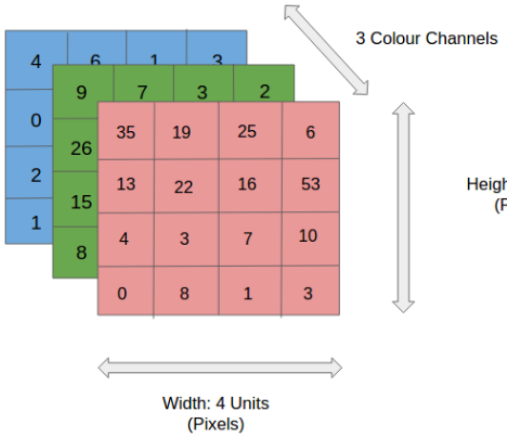




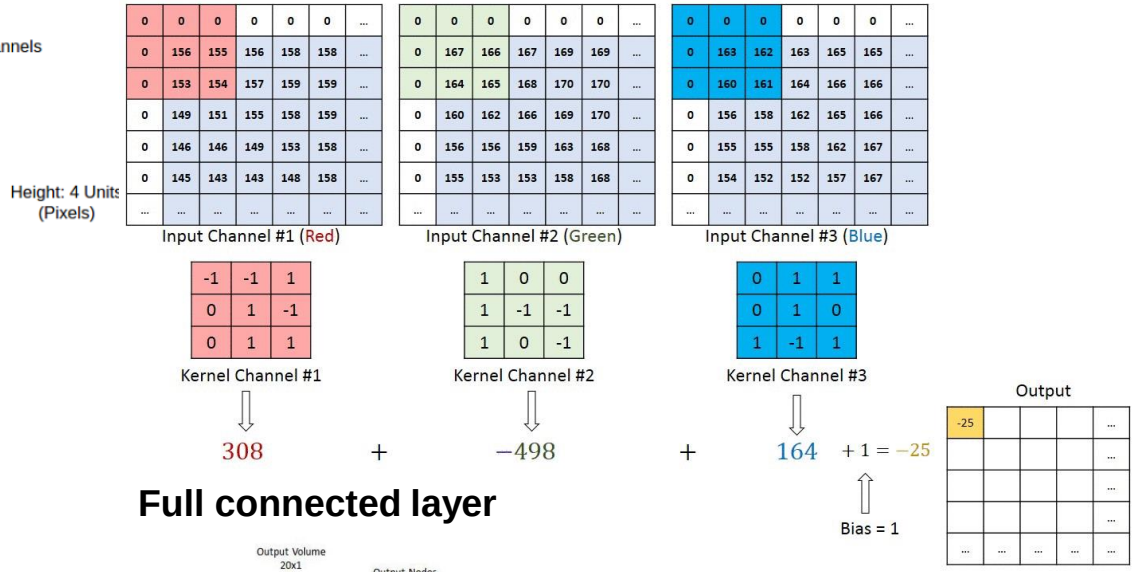
https://towardsdatascience.com/a-comprehensive-guide-to-convolutional-neural-networks-the-eli5-way-3bd2b1164a53

B - Convolution Neural Network

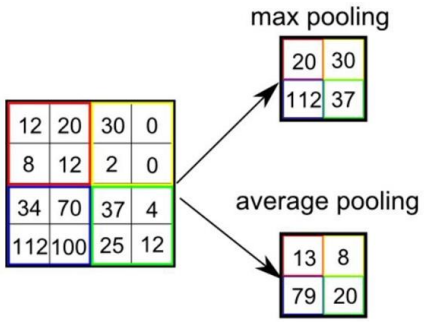
Image



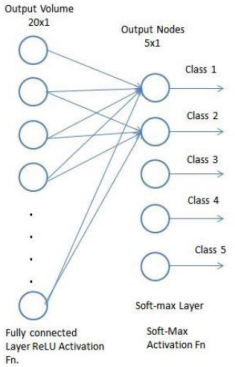
Convolution layer



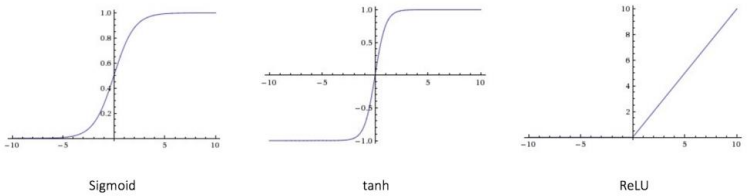
Pooling layer



Full connected layer

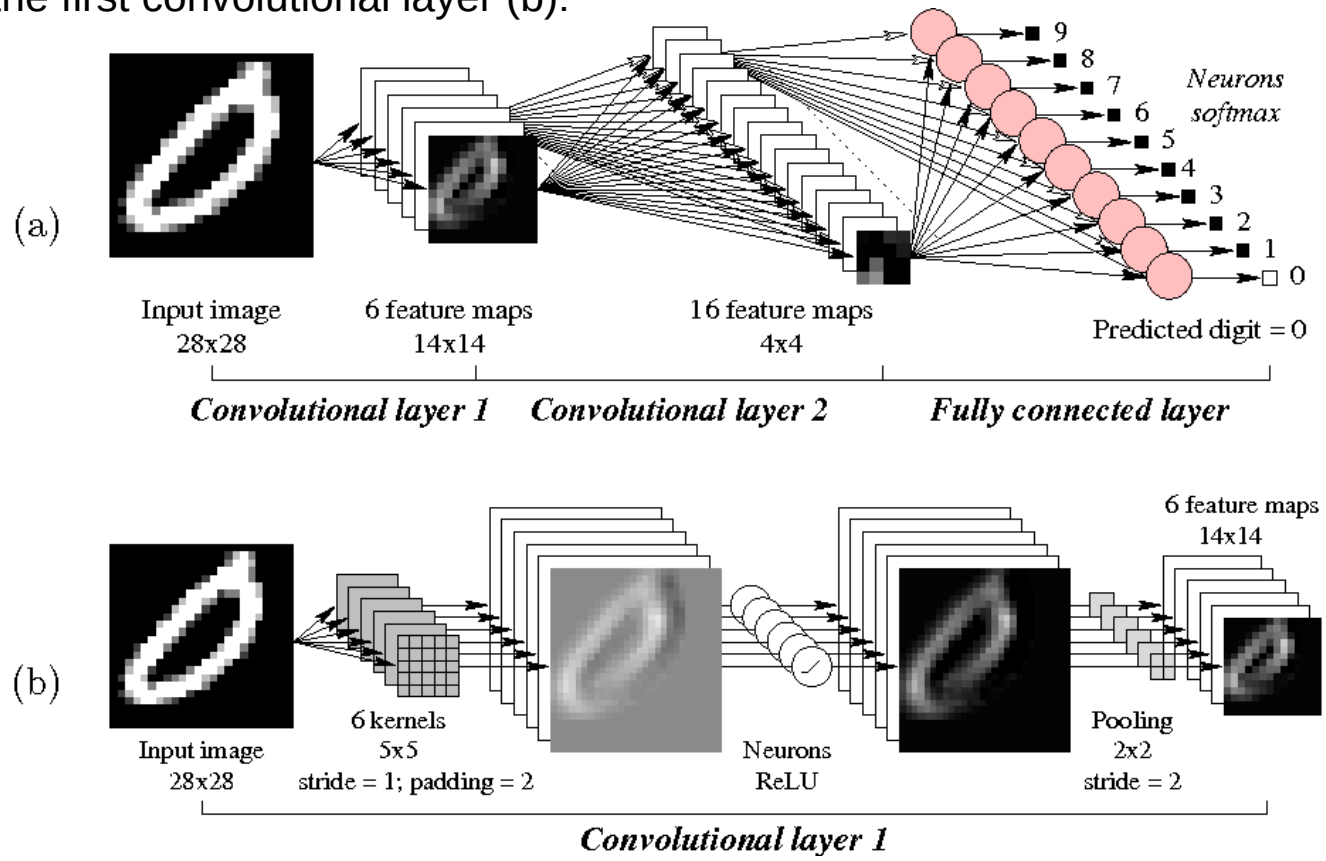


Activation function



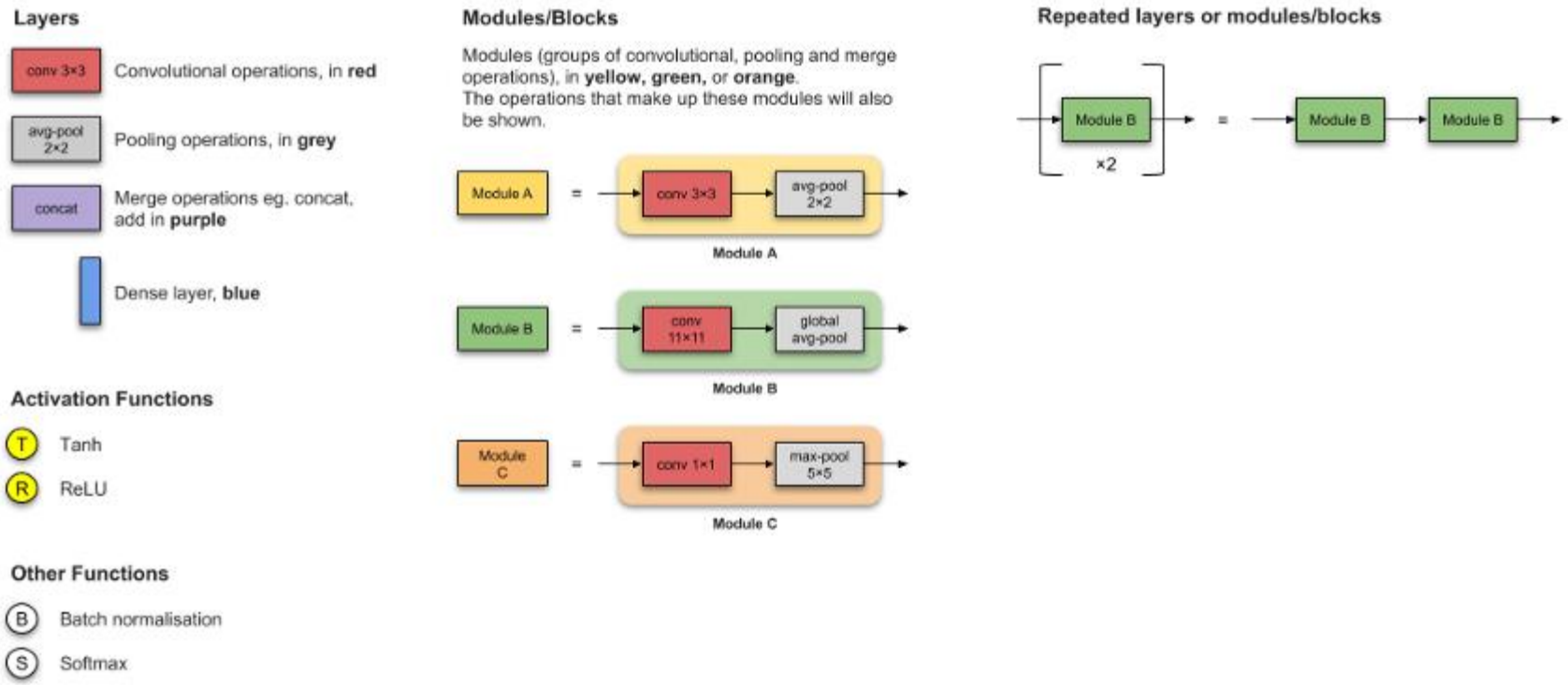
Formation FIDLE 2022/2023:Les CNN : https://youtu.be/S8gCPOIFYfM?si=WfrUkyy7ZKJEl0NJ

A convolutional neural network for the MNIST problem: global architecture (a) and detailed view of the first convolutional layer (b):



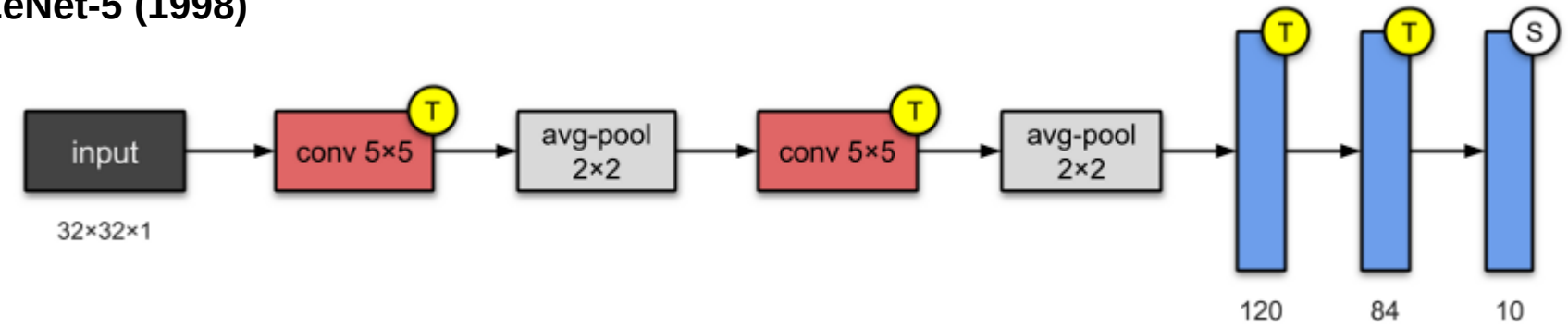
https://www.semanticscholar.org/paper/Steganalysis-via-a-Convolutional-Neural-Network-Couchot-Couturier/7e9708d9dc8b0a4ac2fa52eb384d67f52d7cbbe4

B - Convolution Neural Network

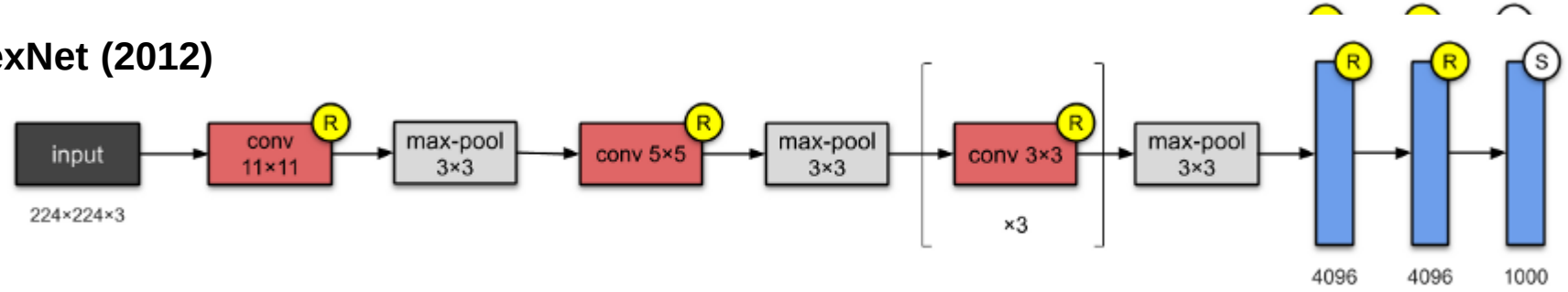


https://towardsdatascience.com/illustrated-10-cnn-architectures-95d78ace614d

LeNet-5 (1998)

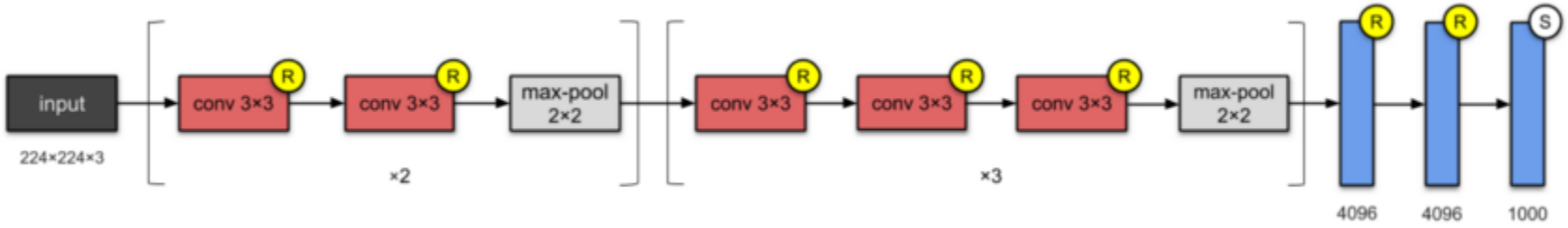


AlexNet (2012)

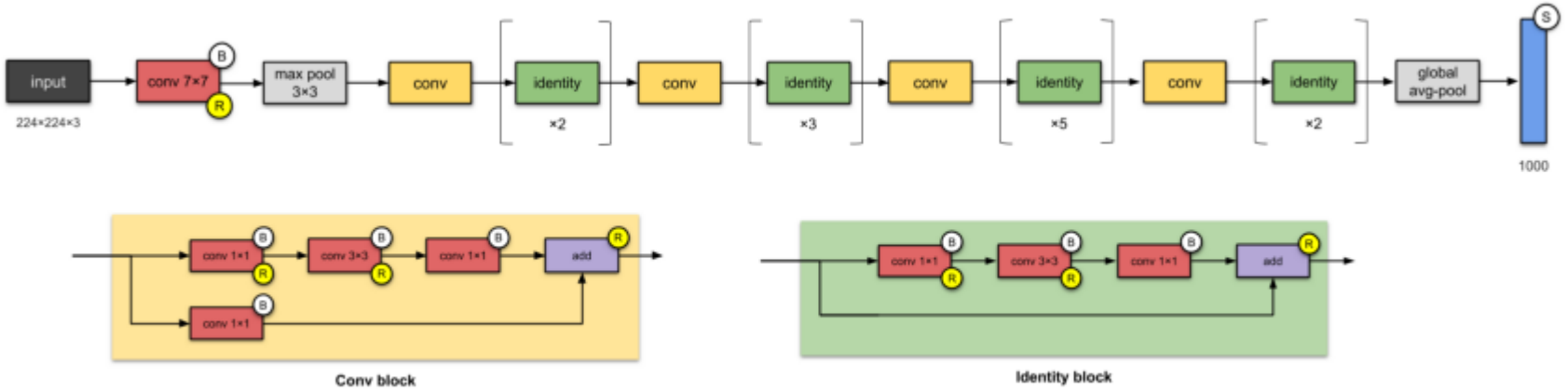


https://towardsdatascience.com/illustrated-10-cnn-architectures-95d78ace614d

VGG-16 (2014)



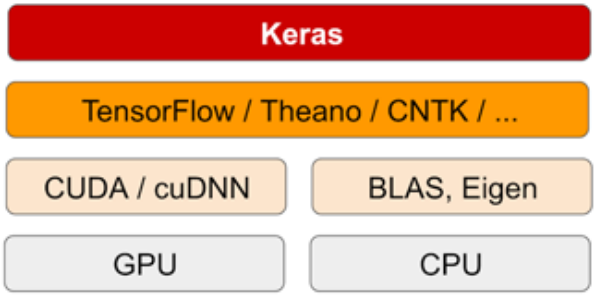
ResNet-50 (2015)



https://towardsdatascience.com/illustrated-10-cnn-architectures-95d78ace614d

B – Library Keras

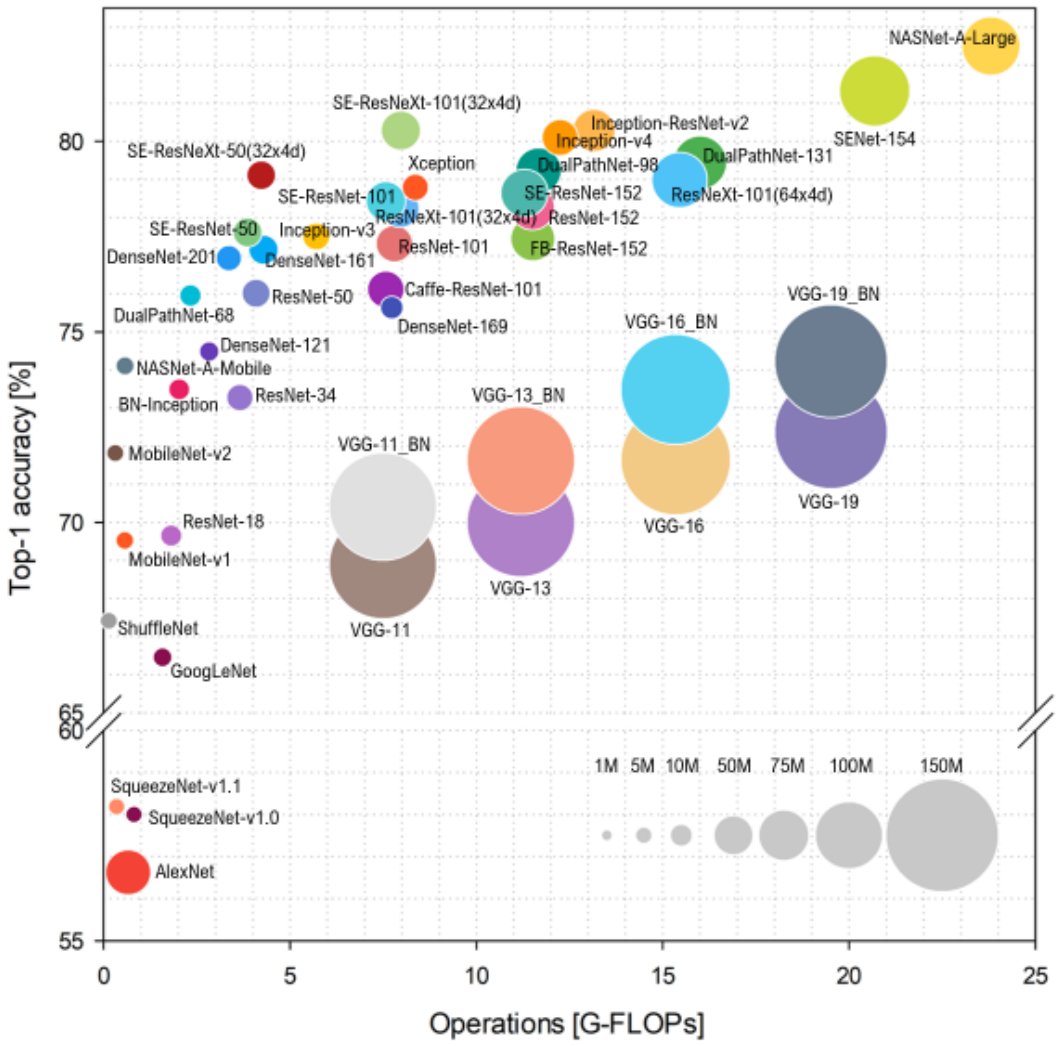
https://keras.io/api/applications/



Model	Size (MB)	Top-1 Accuracy	Top-5 Accuracy	Parameters	Depth	Time (ms) per inference step (CPU)	Time (ms) per inference step (GPU)
Xception	88	79.0%	94.5%	22.9M	81	109.4	8.1
VGG16	528	71.3%	90.1%	138.4M	16	69.5	4.2
VGG19	549	71.3%	90.0%	143.7M	19	84.8	4.4
ResNet50	98	74.9%	92.1%	25.6M	107	58.2	4.6
ResNet50V2	98	76.0%	93.0%	25.6M	103	45.6	4.4
ResNet101	171	76.4%	92.8%	44.7M	209	89.6	5.2
ResNet101V2	171	77.2%	93.8%	44.7M	205	72.7	5.4
ResNet152	232	76.6%	93.1%	60.4M	311	127.4	6.5
ResNet152V2	232	78.0%	94.2%	60.4M	307	107.5	6.6
InceptionV3	92	77.9%	93.7%	23.9M	189	42.2	6.9
InceptionResNetV2	215	80.3%	95.3%	55.9M	449	130.2	10.0
MobileNet	16	70.4%	89.5%	4.3M	55	22.6	3.4
MobileNetV2	14	71.3%	90.1%	3.5M	105	25.9	3.8

B - Convolution Neural Network comparison

Ball chart reporting the Top-1 accuracy vs. computational complexity.

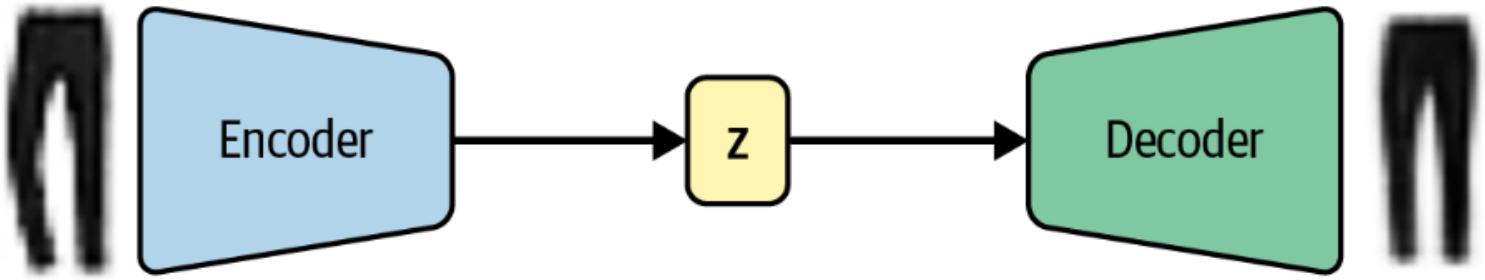


From **S. Bianco, (2018)**
doi: 10.1109/ACCESS.2018.2877890.

An autoencoder is a neural network made up of two parts:

- An encoder network that compresses high-dimensional input data such as an image into a lower-dimensional embedding vector
- A decoder network that decompresses a given embedding vector back to the original domain (e.g., back to an image)

From **D. Foster et K. J. Friston** (2023)

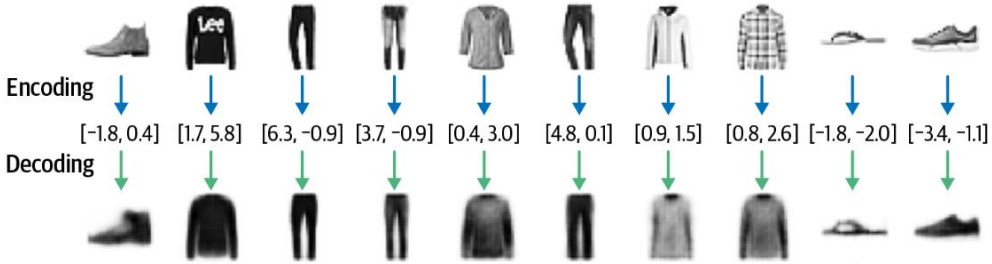


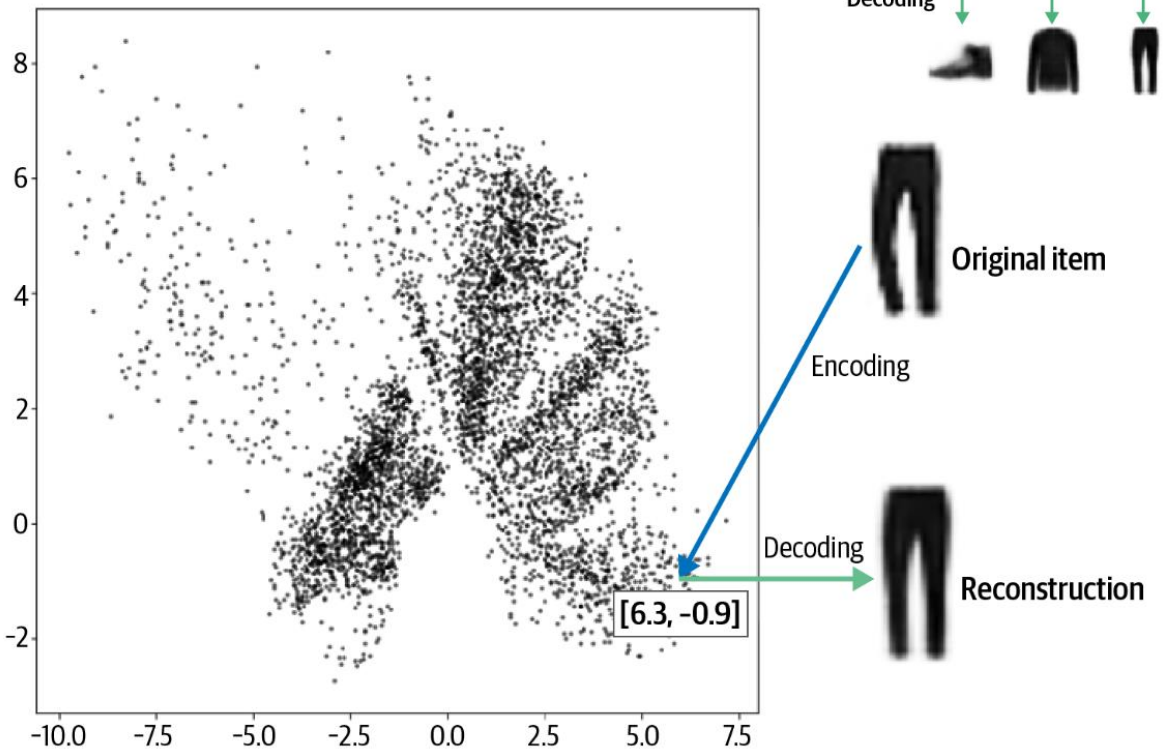
Original image

Generated image

FIDLE 2022/2023:Réseaux autoencodeurs: https://www.youtube.com/live/rGz_NavEMmM?si=gcnnrNn46ruRb390

Examples of images from the Fashion-MNIST dataset:

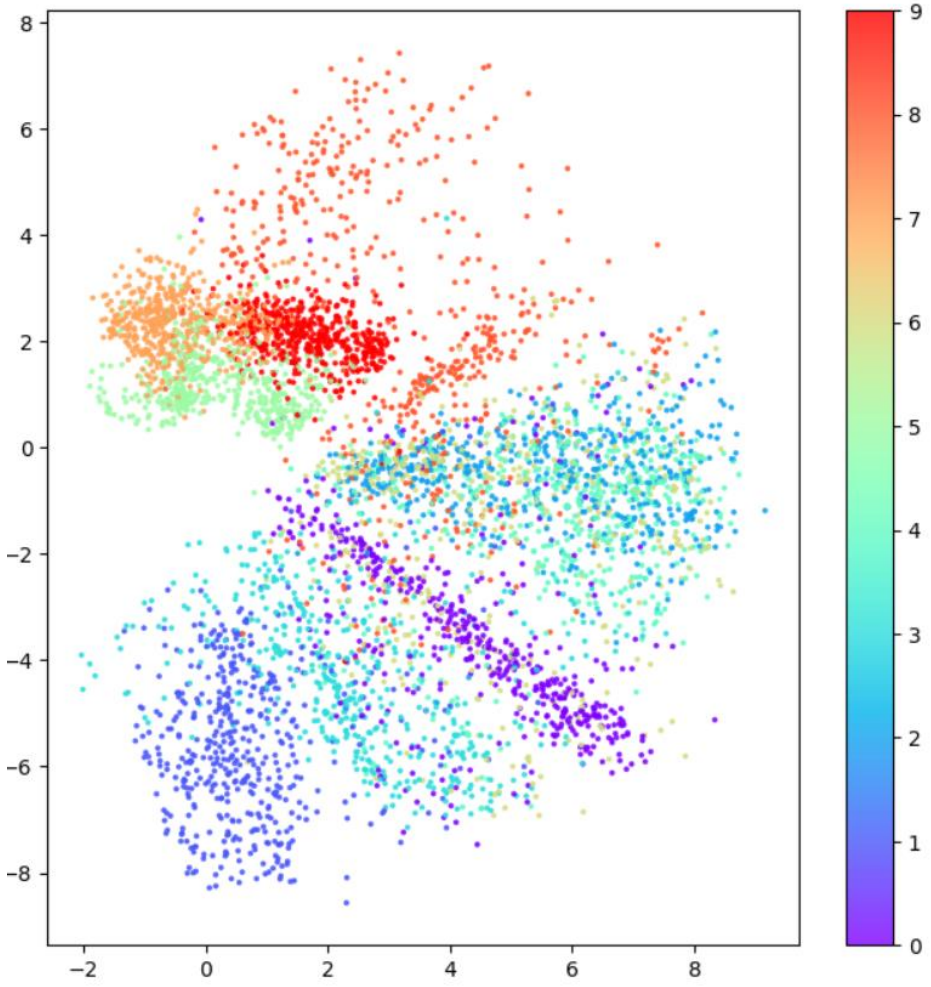




From **D. Foster et K. J. Friston** (2023)

B – Autoencoder

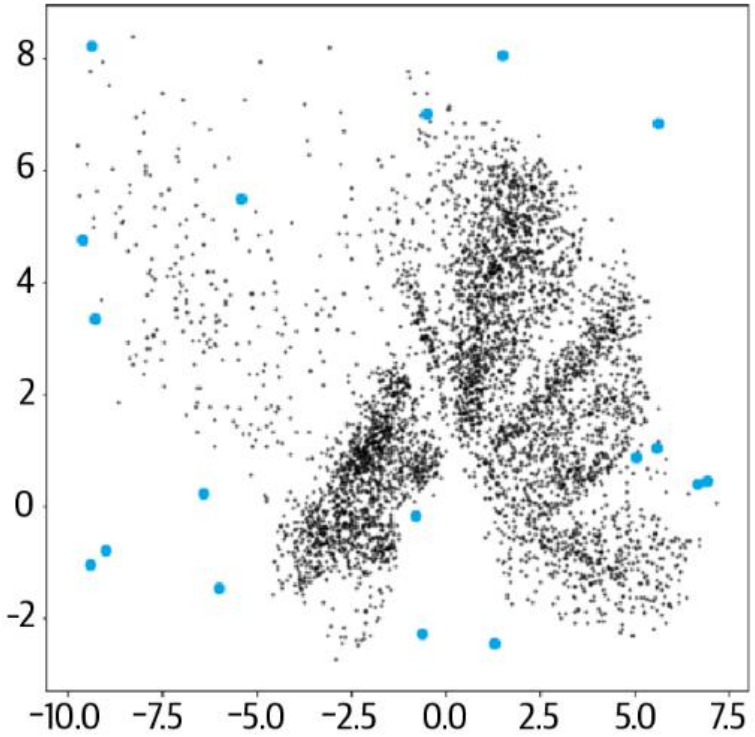
Plot of the latent space, colored by clothing label:

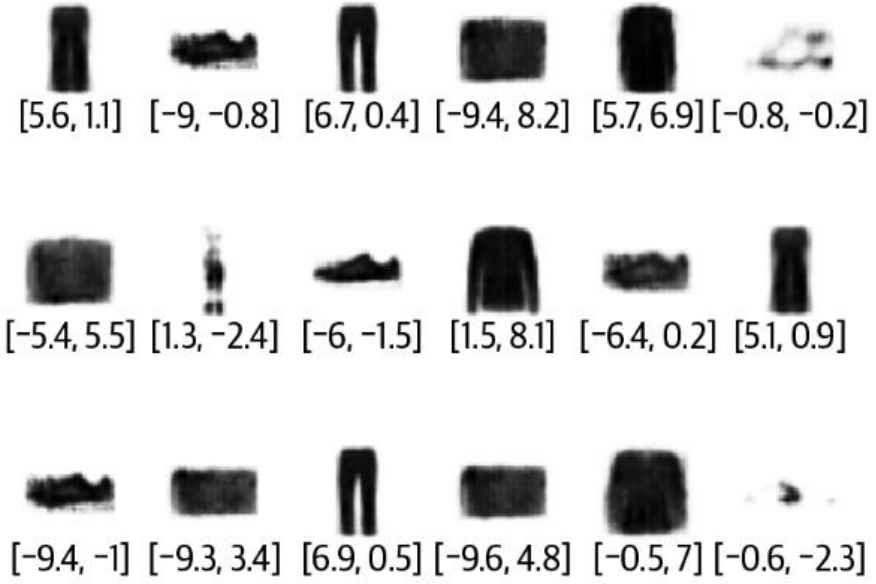


ID Clothing label:

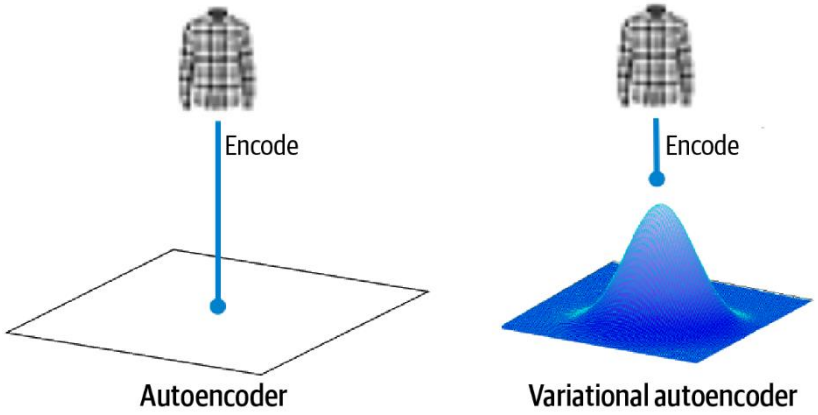
- 0 T-shirt/top
- 1 Trouser
- 2 Pullover
- 3 Dress
- 4 Coat
- 5 Sandal
- 6 Shirt
- 7 Sneaker
- 8 Bag
- 9 Ankle boot

Generated items of clothing



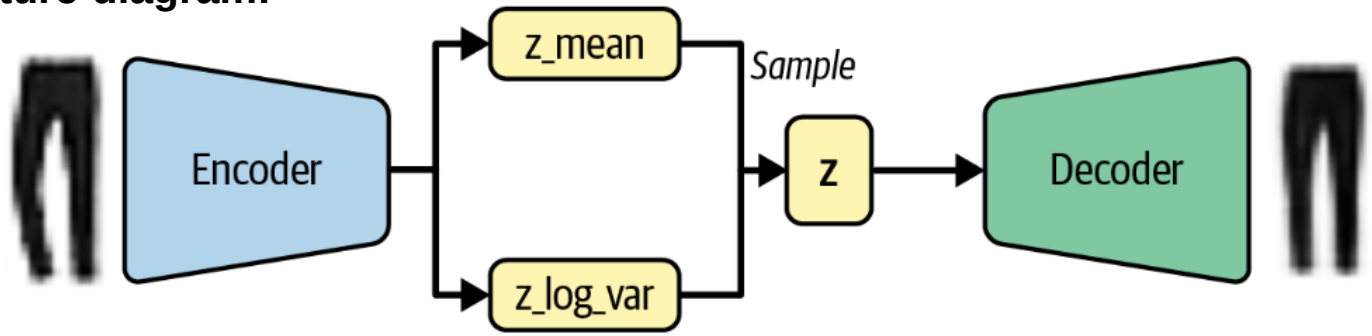


Difference between the encoders in an autoencoder and a variational autoencoder:



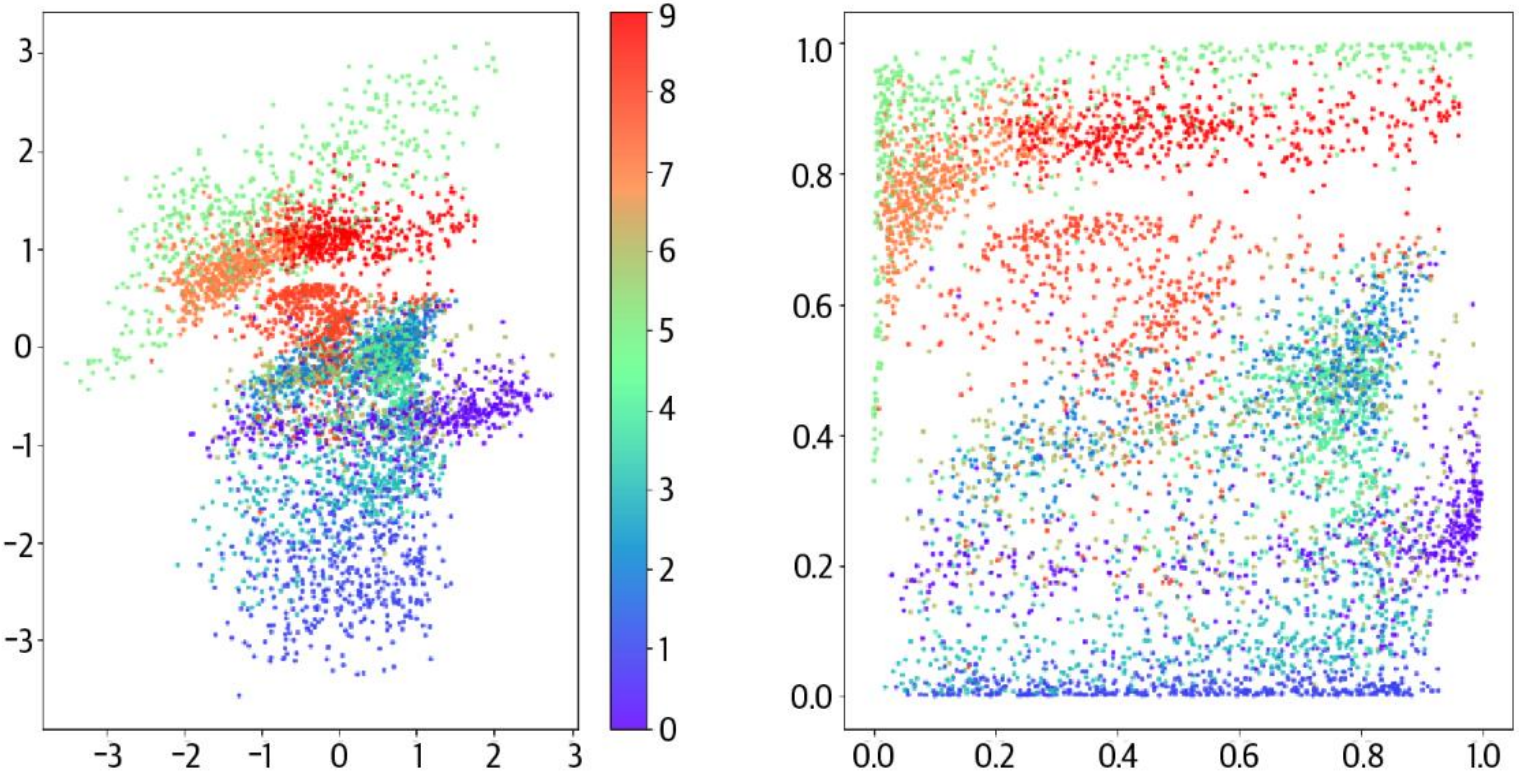
VAE architecture diagram:

From **D. Foster et K. J. Friston** (2023)



FIDLE 2022/2023:Variational Autoencoder (VAE) https://www.youtube.com/live/m7tQeKw7N2k?si=WXiQcin3Z59zJLLm

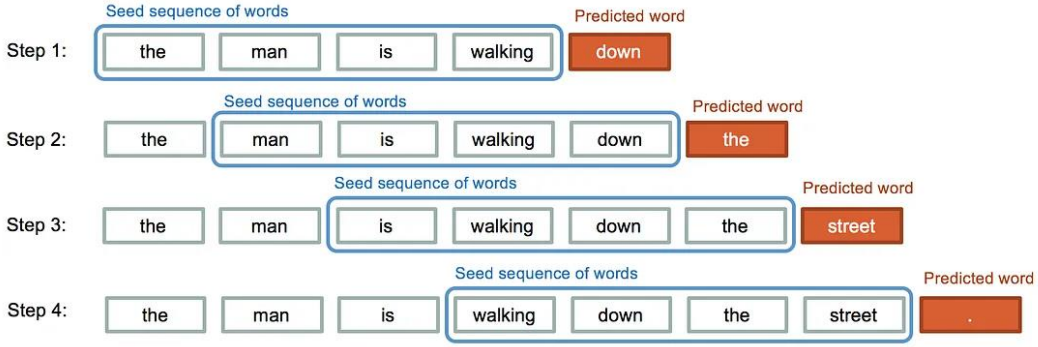
The latent space of the VAE colored by clothing type



B - Recurrent Neural Network (RNN)

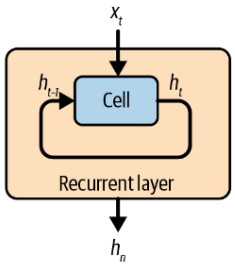
RNN: application to text generation

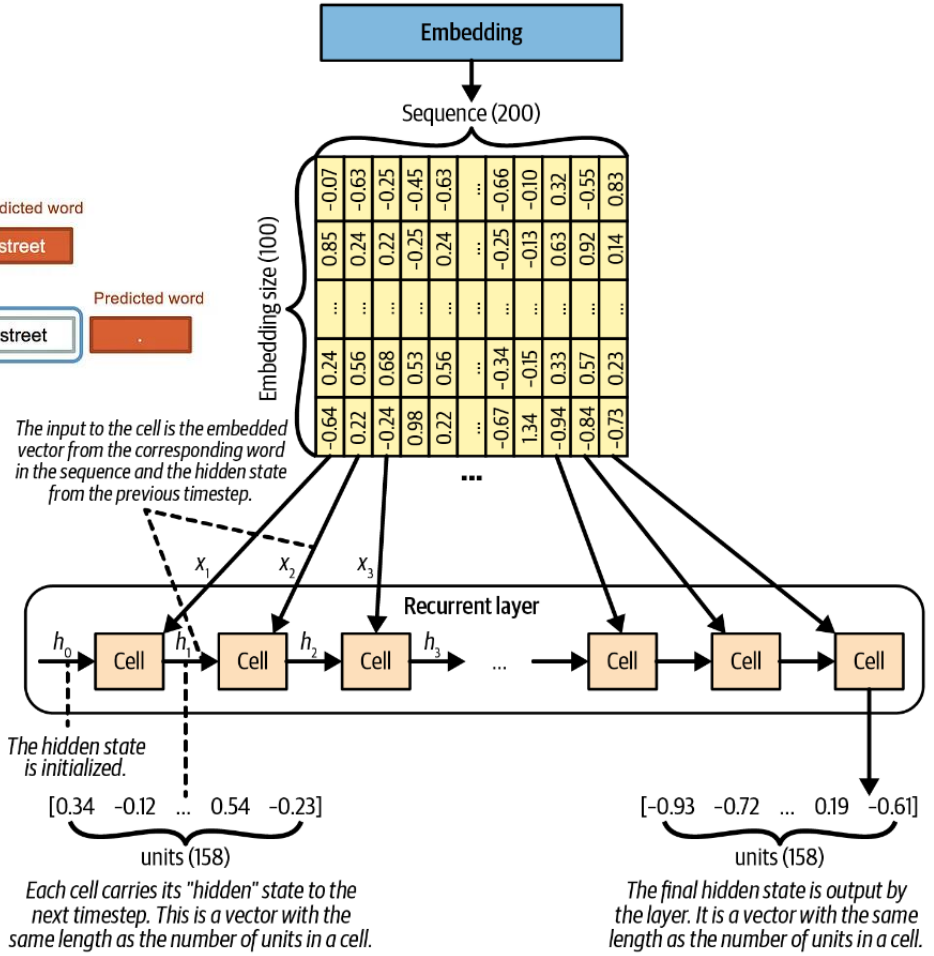
*From **D. Foster** (2023)*



- Tokenization is the process of splitting the text up into individual units, such as words or characters, and convert into an integer.
- An embedding layer is essentially a lookup table that converts each integer token into a vector of length embedding_size

RNN Cell (LSTM):

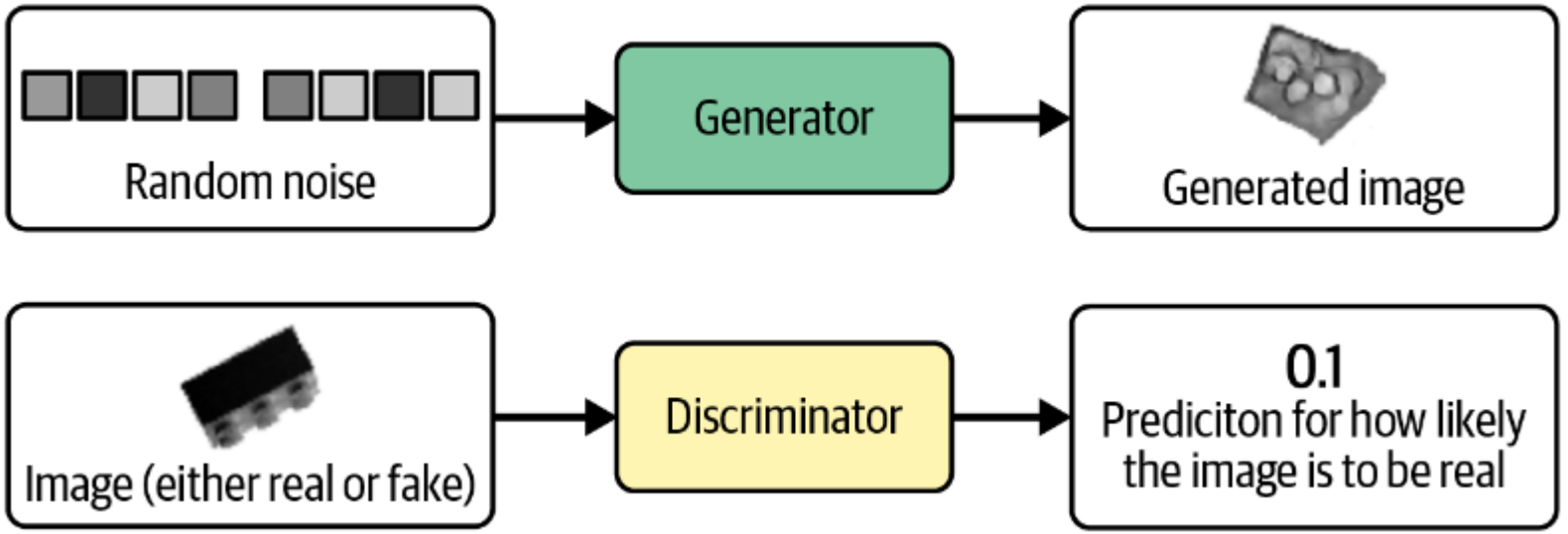




FIDLE 2022/2023:Réseaux récurrents https://youtu.be/naCCxqMD13o?si=ugYxjWUWqOkhYvDB

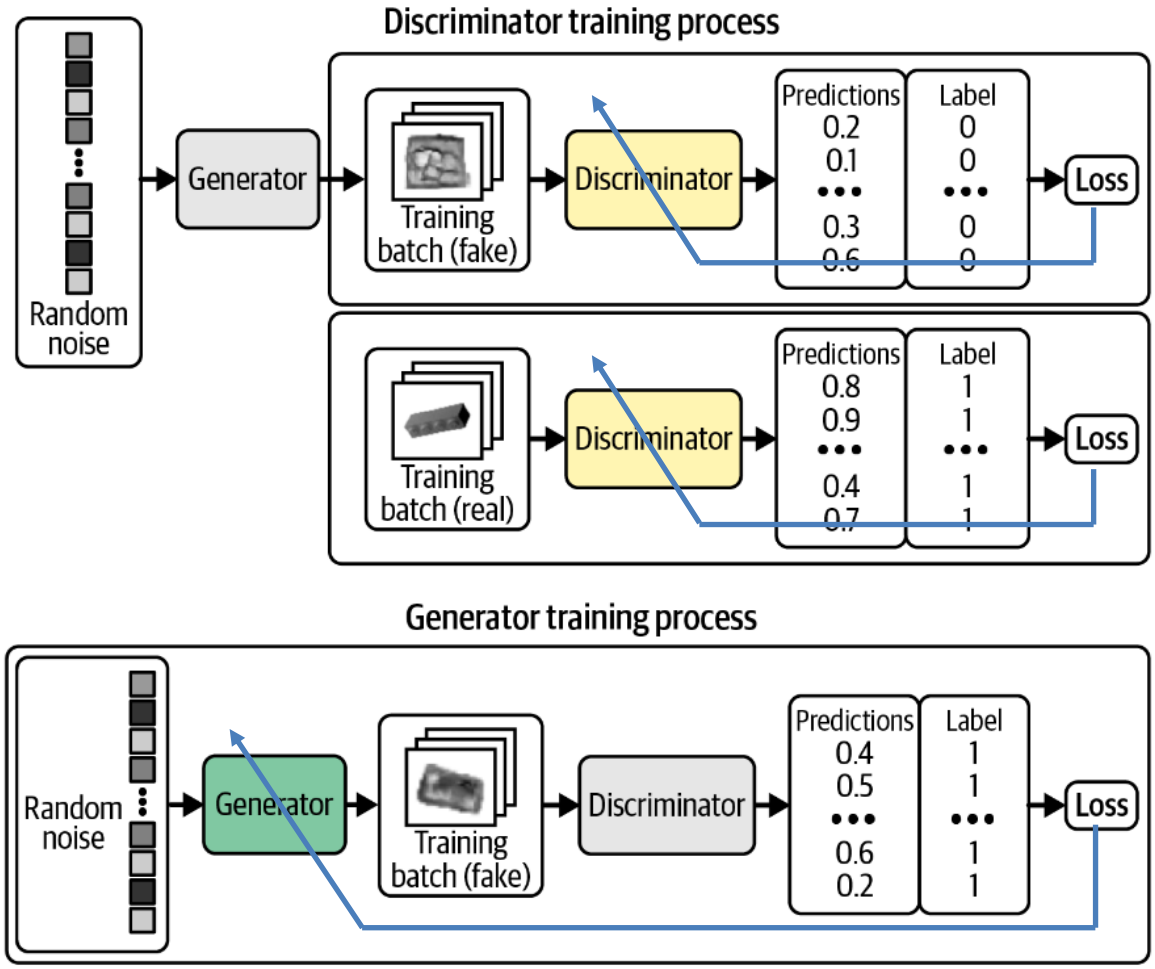
B – Generative Adversial Network (GAN)

A GAN is a battle between two adversaries, the generator and the discriminator. The generator tries to convert random noise into observations that look as if they have been sampled from the original dataset, and the discriminator tries to predict whether an observation comes from the original dataset or is one of the generator's forgeries.



FIDLE 2022/2023:Generative Adversarial Networks (GAN)
https://www.youtube.com/live/hvFthCbTl5c?si=kjxT9I1mKHQnUKgK

B – Generative Adversial Network (GAN)

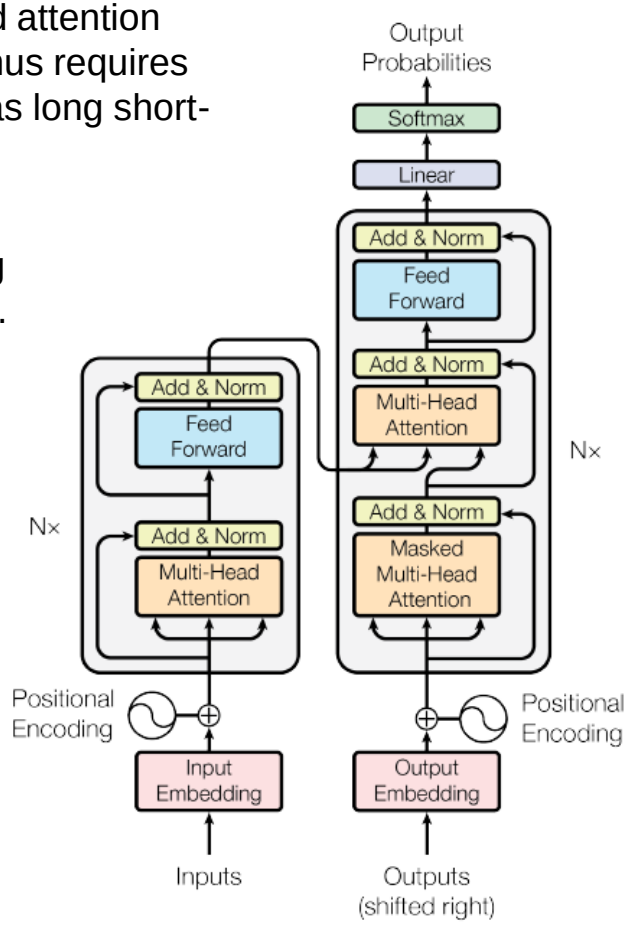


B – The Transformer

A transformer is a deep learning architecture based on the multi-head attention mechanism. It is notable for not containing any recurrent units, and thus requires less training time than previous recurrent neural architectures, such as long short-term memory (LSTM)

This architecture is now used not only in natural language processing and computer vision,[9] but also in audio and multi-modal processing. It has also led to the development of pre-trained systems, such as generative pre-trained transformers (GPTs) and BERT (Bidirectional Encoder Representations from Transformers).

https://en.wikipedia.org/wiki/Transformer_(machine_learning_model)

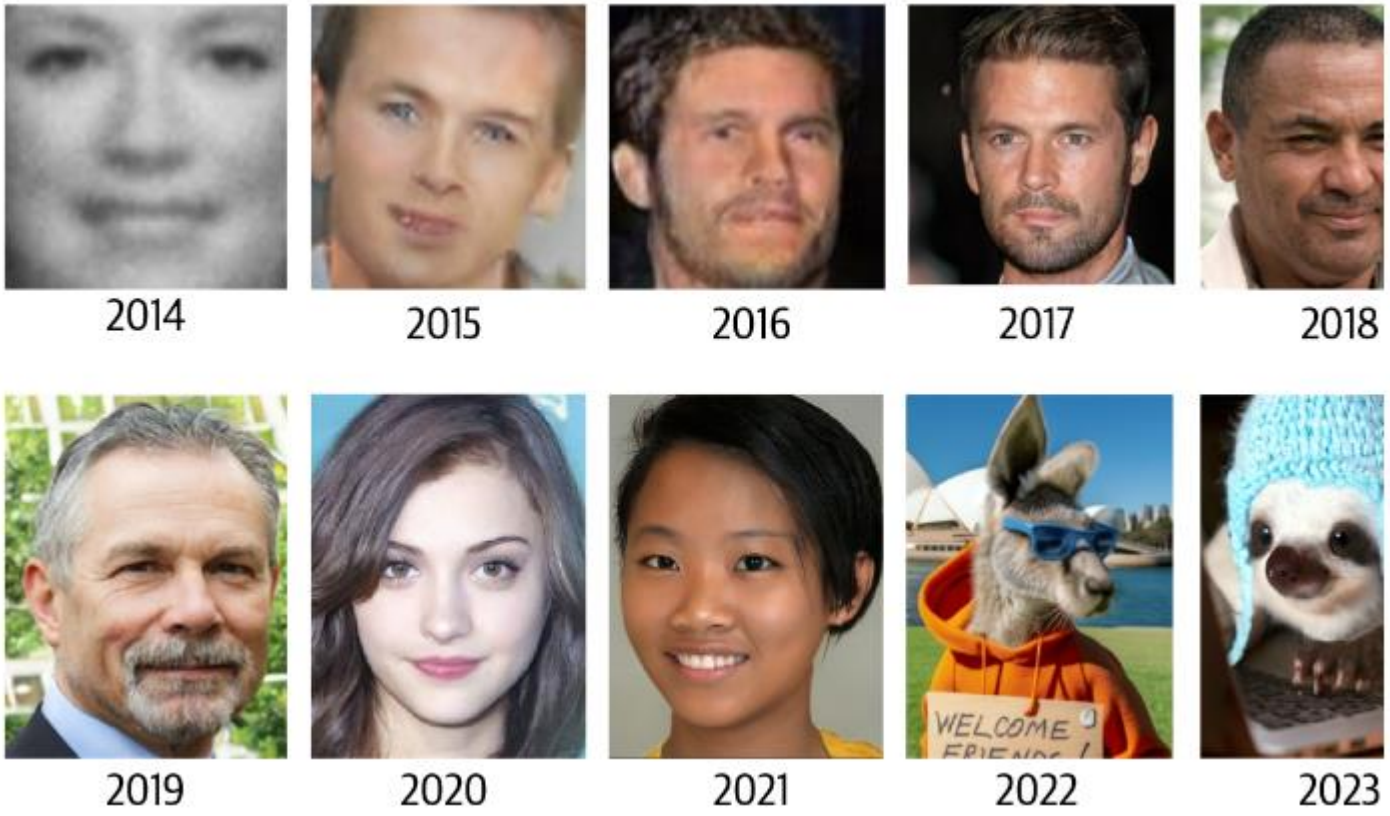


From **A. Vaswani (2023)**
https://arxiv.org/abs/1706.03762

FIDLE 2022/2023: Transformers
https://www.youtube.com/live/L3DGgzIbKz4?si=iTBYxGv_uM3fkbj7

B – Example face generation

Face generation using generative modeling has improved significantly over the last decade:



https://paperswithcode.com/task/image-generation

B - Fidle

Course materials

Course slides The course in pdf format	**Notebooks** Get a Zip or clone this repository	**Datasets** All the needed datasets	**Videos** Our Youtube channel

Have a look about **How to get and install** these notebooks and datasets.

Jupyter notebooks

Linear and logistic regression

- **LINR1** - Linear regression with direct resolution
  Low-level implementation, using numpy, of a direct resolution for a linear regression
- **GRAD1** - Linear regression with gradient descent

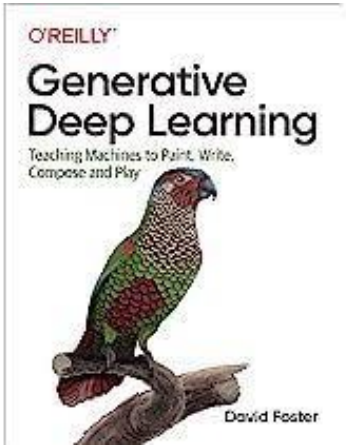
https://gricad-gitlab.univ-grenoble-alpes.fr/talks/fidle/-/tree/master

B - Bibliography

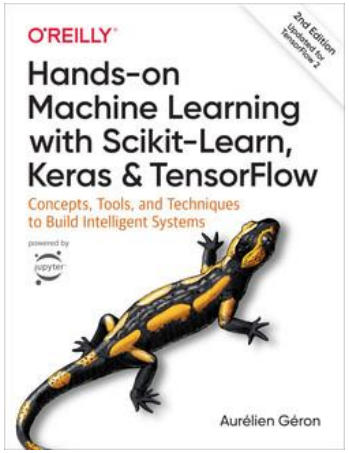
I. H. Sarker, « Deep Learning: A Comprehensive Overview on Techniques, Taxonomy, Applications and Research Directions », SN COMPUT. SCI., vol. 2, nº 6, p. 420, août 2021, doi: 10.1007/s42979-021-00815-1.

S. Bianco, R. Cadene, L. Celona, et P. Napoletano, « Benchmark Analysis of Representative Deep Neural Network Architectures », *IEEE Access*, vol. 6, p. 64270-64277, 2018, doi: 10.1109/ACCESS.2018.2877890.

A. Vaswani et al., « Attention Is All You Need ». arXiv, 1 août 2023. Consulté le: 17 janvier 2024. [En ligne]. Disponible sur: http://arxiv.org/abs/1706.03762



D. Foster et K. J. Friston, *Generative deep learning: teaching machines to paint, write, compose, and play*, Second edition. Beijing ; Boston: O'Reilly, 2023.



[« Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow, 2nd Edition [Book] ». Consulté le: 22 janvier 2024.

C – Human Action Recognition

C – Human Action Recognition

Human activity recognition (HAR) can be referred to as the art of identifying and naming activities using Artificial Intelligence (AI) from the gathered activity raw data by utilizing various sources (so-called devices)

N. Gupta (2022), https://link.springer.com/article/10.1007/s10462-021-10116-x

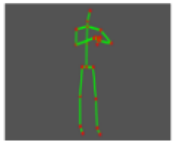
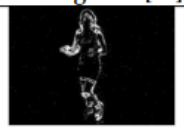
There are two main categorizations of the HAR system based on the equipment:

- **Vision-based HAR (Camera RGB, RGB-D)**
- **Sensor-based HAR (accelerometer, gyroscope etc)**

Vision-based HAR applications :

- **Video analysis (Behavior, Activity)**
- **Human Robot Interaction (Gestures, Activity)**

Visual Modality

Visual Modality	Example	Pros	Cons
RGB	 Hand-waving [27]	· Provide rich appearance information · Easy to obtain and operate · Wide range of applications	· Sensitive to viewpoint · Sensitive to background · Sensitive to illumination
3D Skeleton	 Looking at watch [28]	· Provide 3D structural information of subject pose · Simple yet informative · Insensitive to viewpoint · Insensitive to background	· Lack of appearance information · Lack of detailed shape information · Noisy
Depth	 Mopping floor [29]	· Provide 3D structural information · Provide geometric shape information	· Lack of color and texture information · Limited workable distance
Infrared Sequence	 Pushing [30]	· Workable in dark environments	· Lack of color and texture information · Susceptible to sunlight
Point Cloud	 Bending over [31]	· Provide 3D information · Provide geometric shape information · Insensitive to viewpoint	· Lack of color and texture information · High computational complexity
Event Stream	 Running [32]	· Avoid much visual redundancy · High dynamic range · No motion blur	· Asynchronous output · Spatio-temporally sparse · Capturing device is relatively expensive

From **Z. Sun (2022) -** doi:10.1109/TPAMI.2022.3183112 .

C – List of benchmarks HAR datasets.

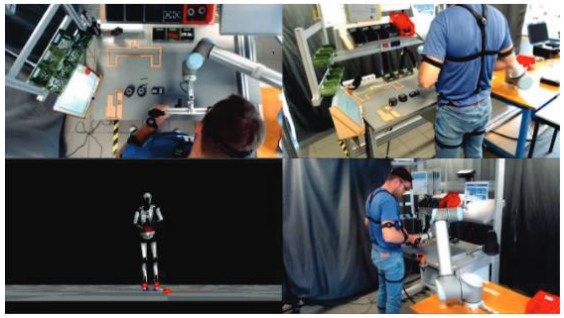
Dataset	No. of Classes	No. of Video Clips	Description	Source of Data Collection	URL	Release Year	Paper
UCF 101	101	13320	Realistic action videos, an extension of the UCF50 data set which has 50 action categories	YouTube	https://www.crcv.ucf.edu/data/UCF101.php	2012	(Soomro, Roshan Zamir, and Shah 2012)
HMDB 51	51	6849	Each class containing a minimum of 101 clips	Movies, public databases and YouTube	https://serre-lab.clps.brown.edu/resource/hmdb-a-large-human-motion-database/#dataset	2011	(H Kuehne et al. 2011)
JHMDB	21	928	2D pose annotation for scale, pose, segmentation, coarse viewpoint, and dense optical flow	21 action categories extracted from HMDB51	http://jhmdb.is.tue.mpg.de/dataset	2013	(Jhuang et al. 2013)
Kinetics400	400	300000	YouTube video URLs dataset	YouTube video	https://deepmind.com/research/open-source/kinetics	2017	(Kay et al. 2017)
Kinetics600	600	500000	YouTube video URLs dataset	YouTube video	https://deepmind.com/research/open-source/kinetics	2018	(Carreira et al. 2018)
Kinetics700	700	65000	YouTube video URLs dataset	YouTube video	https://deepmind.com/research/open-source/kinetics	2019	(Carreira et al. 2019)
Breakfast Dataset	10	1989	consists of 10 cooking activities performed by 52 different actors in multiple kitchen locations	Manually Recorded	https://serre-lab.clps.brown.edu/resource/breakfast-actions-dataset/	2014	(Hilde Kuehne, Arslan, and Serre 2014)
Charades	157	9848	Indoor Activities	Amazon Mechanical Turk (AMT)	https://prior.allenai.org/projects/charades	2016	(Sigurdsson et al. 2016)
AVA	80	57600	spatio-temporal localization of atomic visual actions	192 different movies	https://research.google.com/ava/	2018	(Chunhui et al. 2018)
Epic-Kitchens	149	432	action segmentation dataset	Manually recorded	https://epic-kitchens.github.io/2018	2018	(Damen et al. 2018)
Something-Something	174	220847	basic actions with everyday objects	Recording by crowd workers	https://20bn.com/datasets/something-something	2017	(Goyal et al. 2017)
Moments in Time - Dataset	339	1000000	capturing visual and/or audible actions	Web-Search	http://moments.csail.mit.edu/	2020	(Monfort et al. 2020)
Sport 1-1 M	487	1200000	sports action video dataset	YouTube	https://deepai.org/dataset/sports-1m#__sid=js0	2014	(Karpathy et al. 2014b)

From V. Sharma (2022)

A – Dataset HRA in industry

Dataset	No. of Classes	No. of Video Clips	Description	Source of Data Collection	URL	Release Year	Paper
ActivityNet	200	20000	Videos for Human Activity Understanding	Web-Search	http://activity-net.org/download.html	2015	(Heilbron et al. 2015)
Hollywood Extended	16	937	Segmentation and classification of actions computed as mean over frames	Movies	https://www.di.ens.fr/willow/research/actionordering/	2014	(Bojanowski et al. 2014)

Dataset for HRA in industrial context

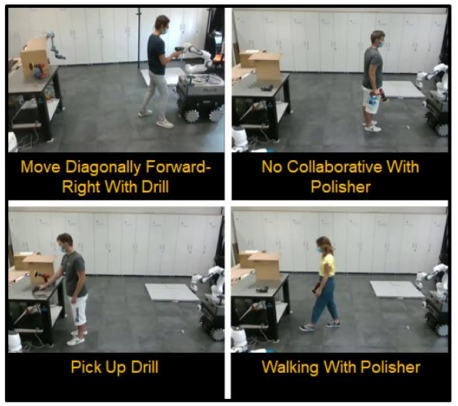


Interaction with robotic arm

From **M. Dallel (2020)**
doi: 10.1109/ICHMS49158.2020.9209531.

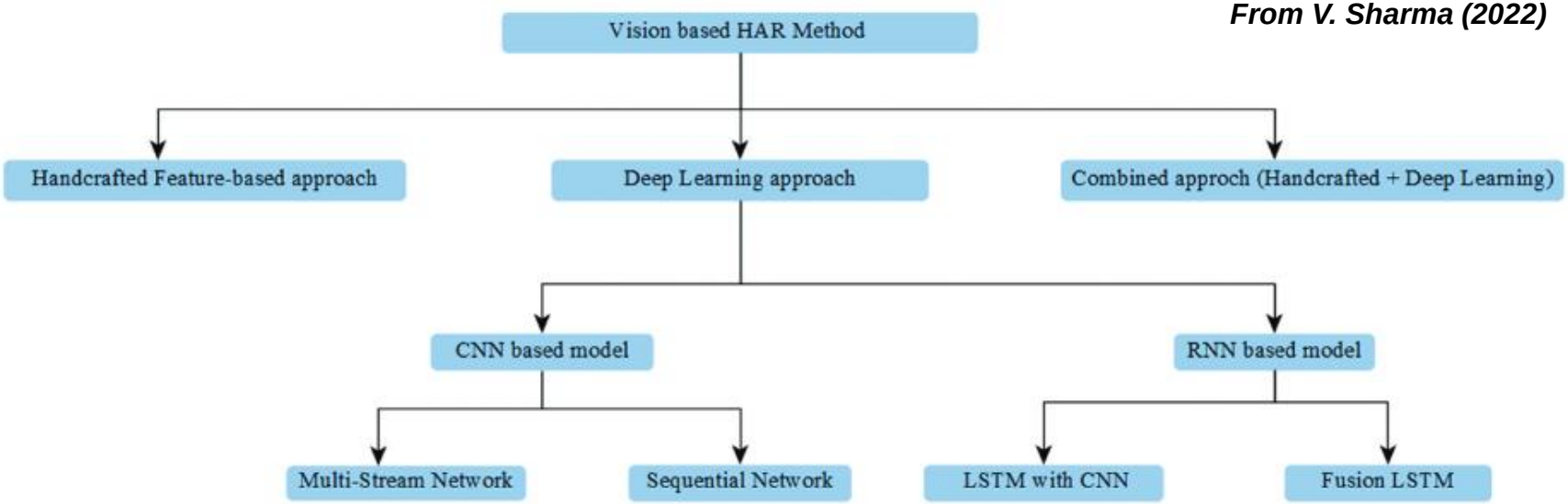
Interaction with mobile robot

From **F. Iodice (2022)**
doi: 10.1109/ICPR56361.2022.9956300.



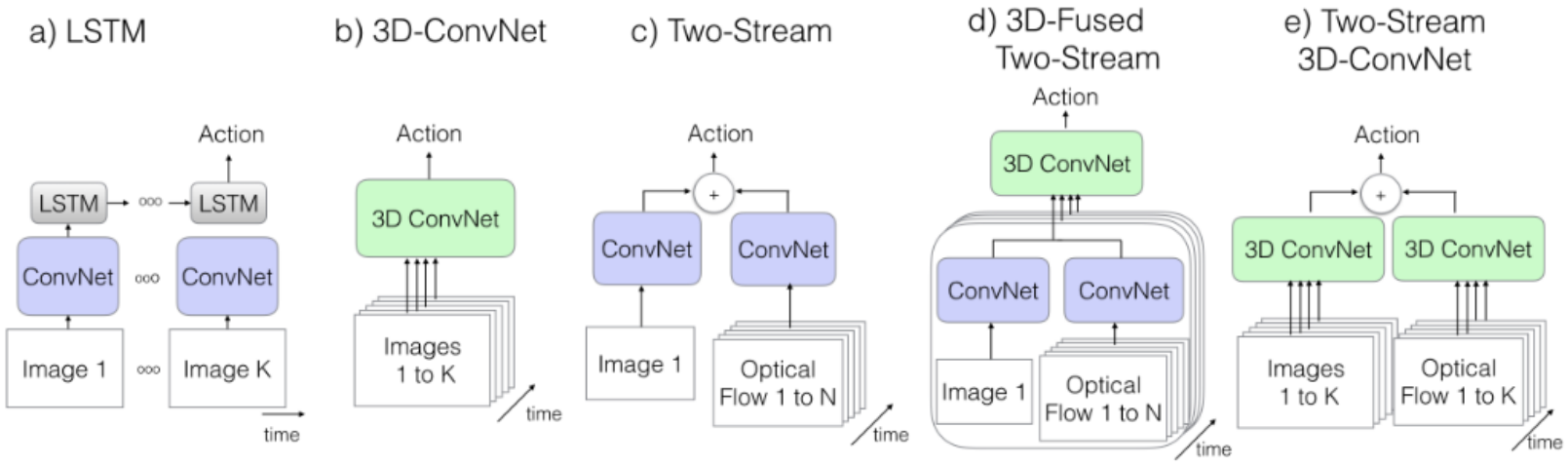
C – Taxonomy of vision-based HAR Methods

From V. Sharma (2022)



- **Handcrafted feature-based Approach approach usually includes a three-stage process – feature extraction, feature classification, and feature representation.**
- **Deep learning-based HAR method can simultaneously learn visual features, feature representations, and classifiers.**

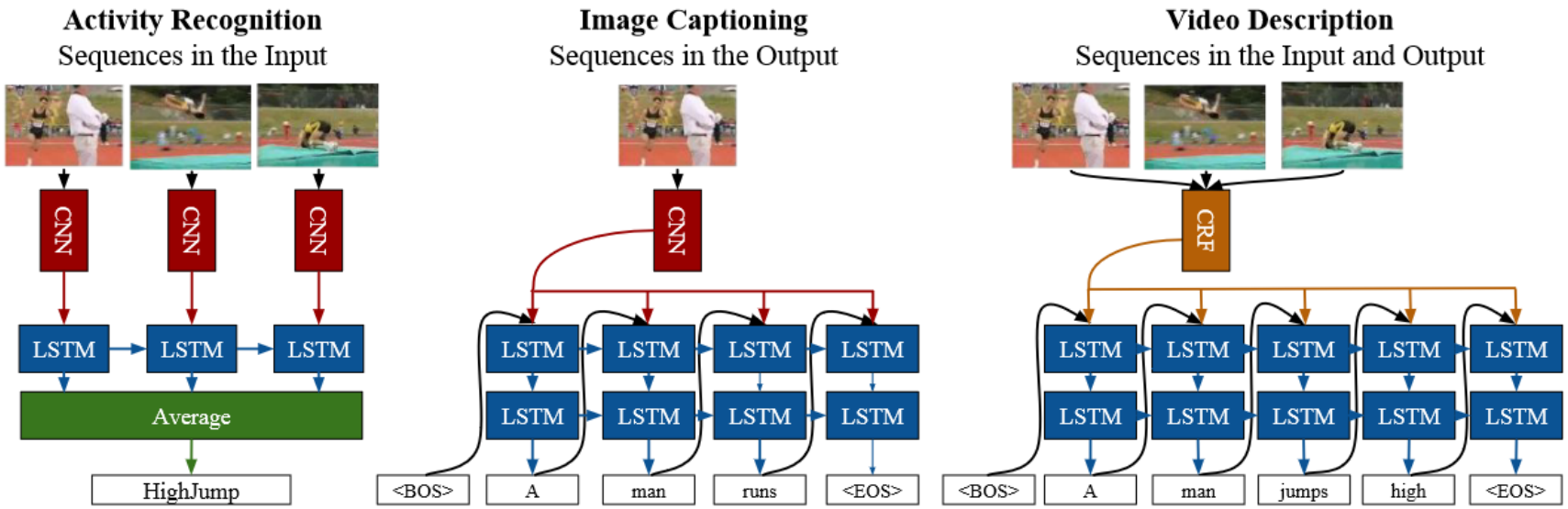
C – Deep learning-based HAR architectures



Architecture	UCF-101			HMDB-51			Kinetics		
	RGB	Flow	RGB + Flow	RGB	Flow	RGB + Flow	RGB	Flow	RGB + Flow
(a) LSTM	81.0	–	–	36.0	–	–	63.3	–	–
(b) 3D-ConvNet	51.6	–	–	24.3	–	–	56.1	–	–
(c) Two-Stream	83.6	85.6	91.2	43.2	56.3	58.3	62.2	52.4	65.6
(d) 3D-Fused	83.2	85.8	89.3	49.2	55.5	56.8	–	–	67.2
(e) Two-Stream I3D	**84.5**	**90.6**	**93.4**	**49.8**	**61.9**	**66.4**	**71.1**	**63.4**	**74.2**

From **J. Carreira et A. Zisserman (2018)**

C – Activity recognition, image and video description



[1]

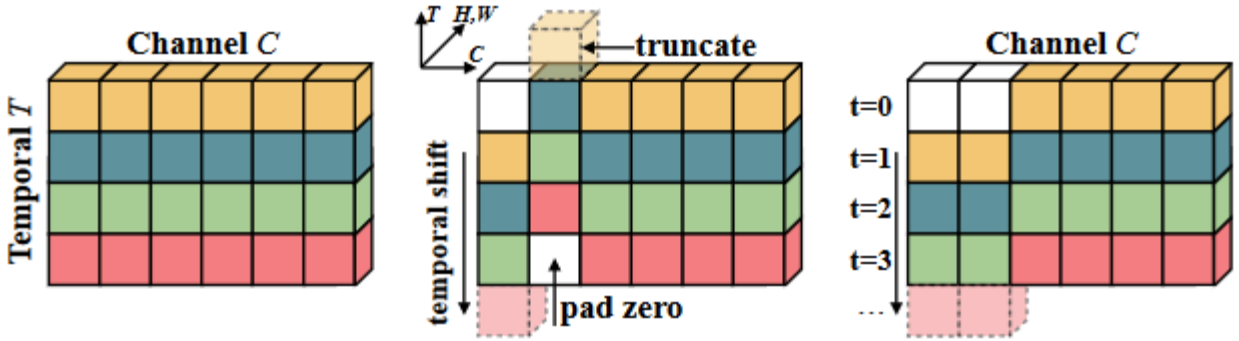
J. Donahue *et al.*, « Long-term Recurrent Convolutional Networks for Visual Recognition and Description ». arXiv, 31 mai 2016. Consulté le: 18 janvier 2024. [En ligne]. Disponible sur: http://arxiv.org/abs/1411.4389

C - Temporal Shift Module (TSM)

Temporal Shift Module (TSM) performs efficient temporal modeling by moving the feature map along the temporal dimension. It is computationally free on top of a 2D convolution, but achieves strong temporal modeling ability.

TSM efficiently supports both offline and online video recognition.

Bi-directional TSM mingles both past and future frames with the current frame, which is suitable for high-throughput offline video recognition.



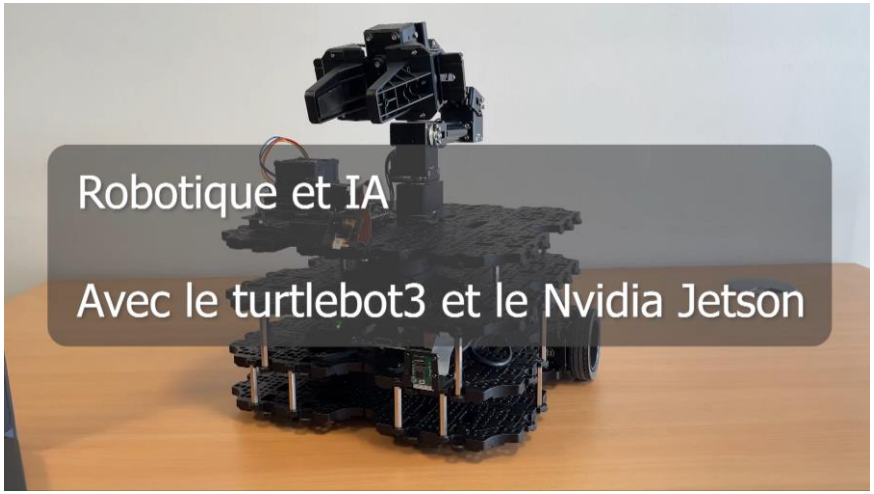
(a) The original tensor without shift. (b) Offline temporal shift (bi-direction). (c) Online temporal shift (uni-direction).

J. Lin, C. Gan, et S. Han, « TSM: Temporal Shift Module for Efficient Video Understanding ». arXiv, 22 août 2019. Consulté le: 10 novembre 2023. [En ligne]. Disponible sur: http://arxiv.org/abs/1811.08383

C - Collaboration ICAM / LISSI

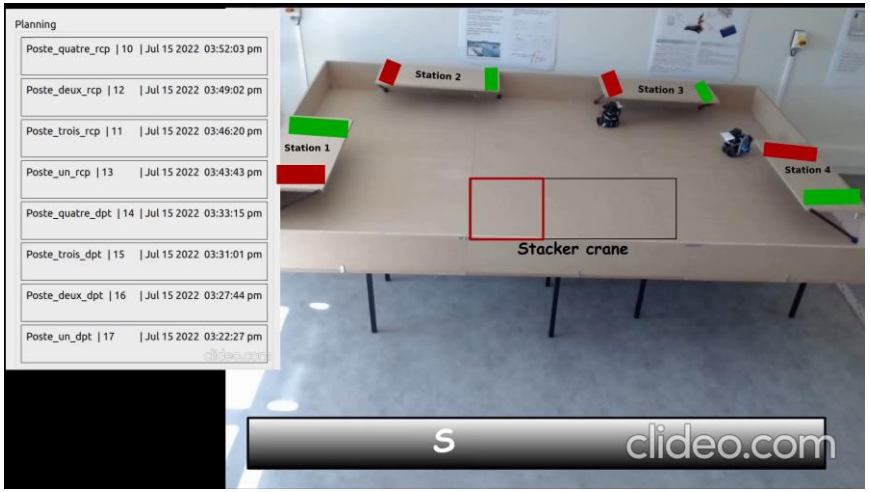
Internship DUT 2022 – H. Cataldi :

- Robot Turtlebot3
- ROS (Robot Operating System)
- Camera + Nvidia Jetson



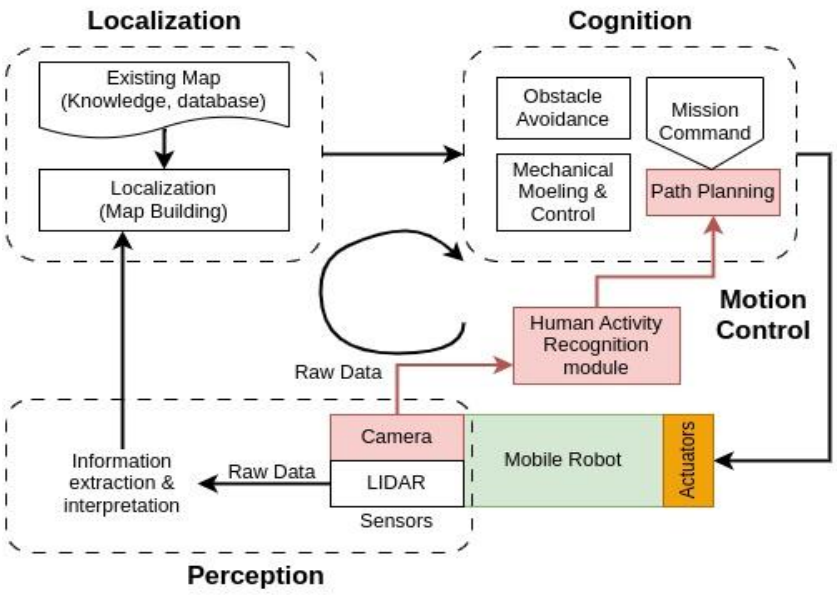
Internship M2 2022 – K. LISSASSI

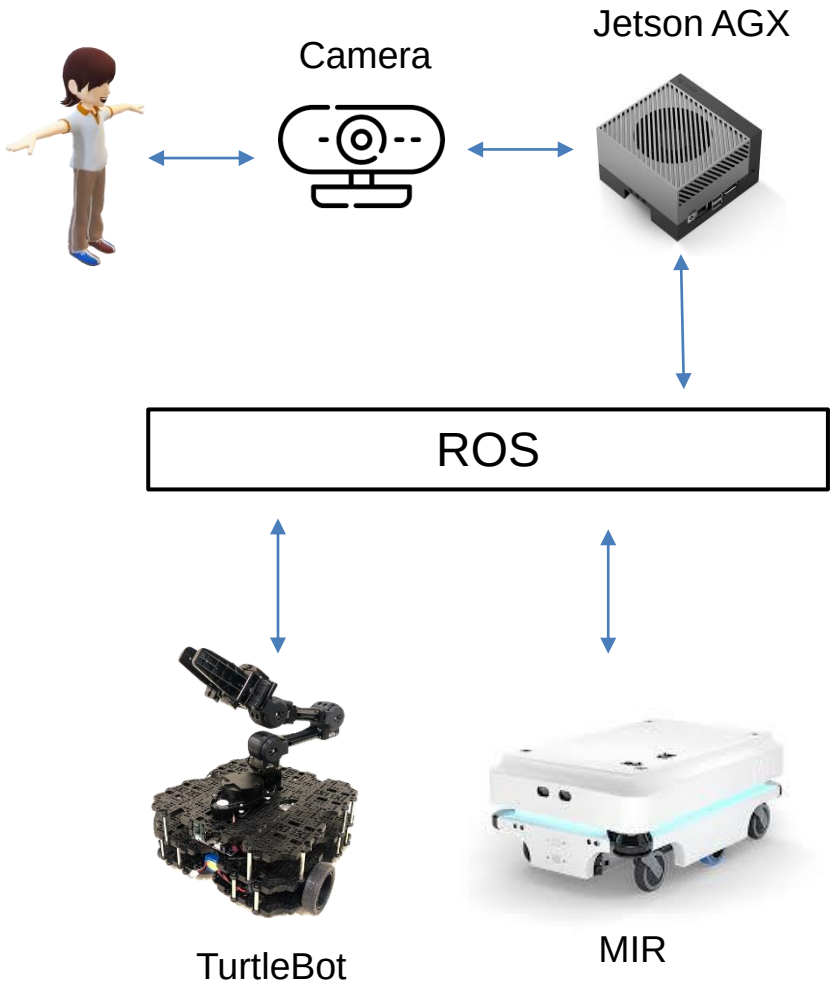
- Multi-robots
- Autonomous Navigation
- IHM



Kokou C. LISSASSI, Paul Eric DOSSOU, and Christophe SABOURIN. “Using Computer Vision to Improve SME Performance.” **FAIM 2023**
https://www.faimconference.org/

Activity recognition for motion control of mobile robot





C – Bibliography

N. Gupta, S. K. Gupta, R. K. Pathak, V. Jain, P. Rashidi, et J. S. Suri, « Human activity recognition in artificial intelligence framework: a narrative review », Artif Intell Rev, vol. 55, nº 6, p. 4755-4808, août 2022, doi: 10.1007/s10462-021-10116-x .

Z. Sun, Q. Ke, H. Rahmani, M. Bennamoun, G. Wang, et J. Liu, « Human Action Recognition from Various Data Modalities: A Review », *IEEE Trans. Pattern Anal. Mach. Intell.*, p. 1-20, 2022, doi: 10.1109/TPAMI.2022.3183112.

V. Sharma, M. Gupta, A. K. Pandey, D. Mishra, et A. Kumar, « A Review of Deep Learning-based Human Activity Recognition on Benchmark Video Datasets », *Applied Artificial Intelligence*, vol. 36, nº 1, p. 2093705, déc. 2022, doi: 10.1080/08839514.2022.2093705.

J. Carreira et A. Zisserman, « Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset ». arXiv, 12 février 2018. Consulté le: 18 janvier 2024. [En ligne]. Disponible sur: http://arxiv.org/abs/1705.07750

J. Donahue *et al.*, « Long-term Recurrent Convolutional Networks for Visual Recognition and Description ». arXiv, 31 mai 2016. Consulté le: 18 janvier 2024. [En ligne]. Disponible sur: http://arxiv.org/abs/1411.4389

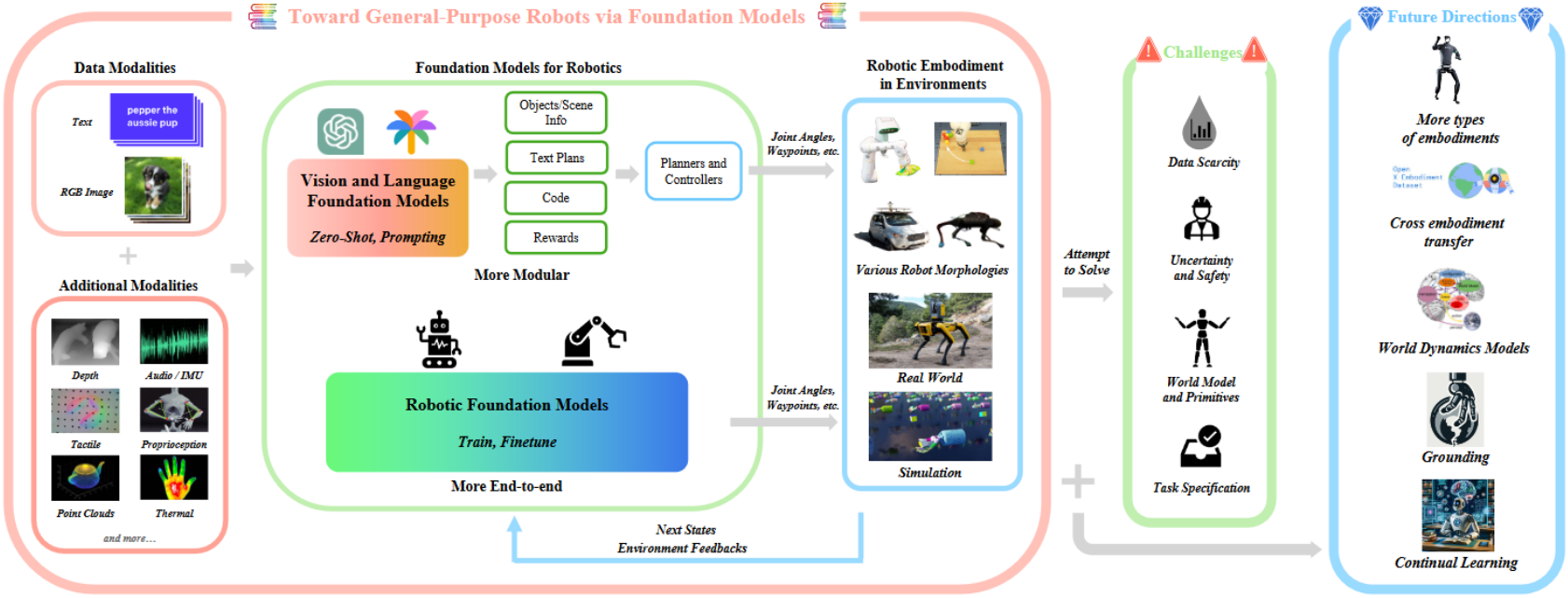
D - Toward General-Purpose Robots via Foundation Models

D - Foundation Models

A foundation model is an AI model that is trained on broad data such that it can be applied across a wide range of use cases. Foundation models have transformed AI, powering prominent chatbots and generative AI.

https://en.wikipedia.org/wiki/Foundation_model

Towards General-Purpose Robots via Foundation Models

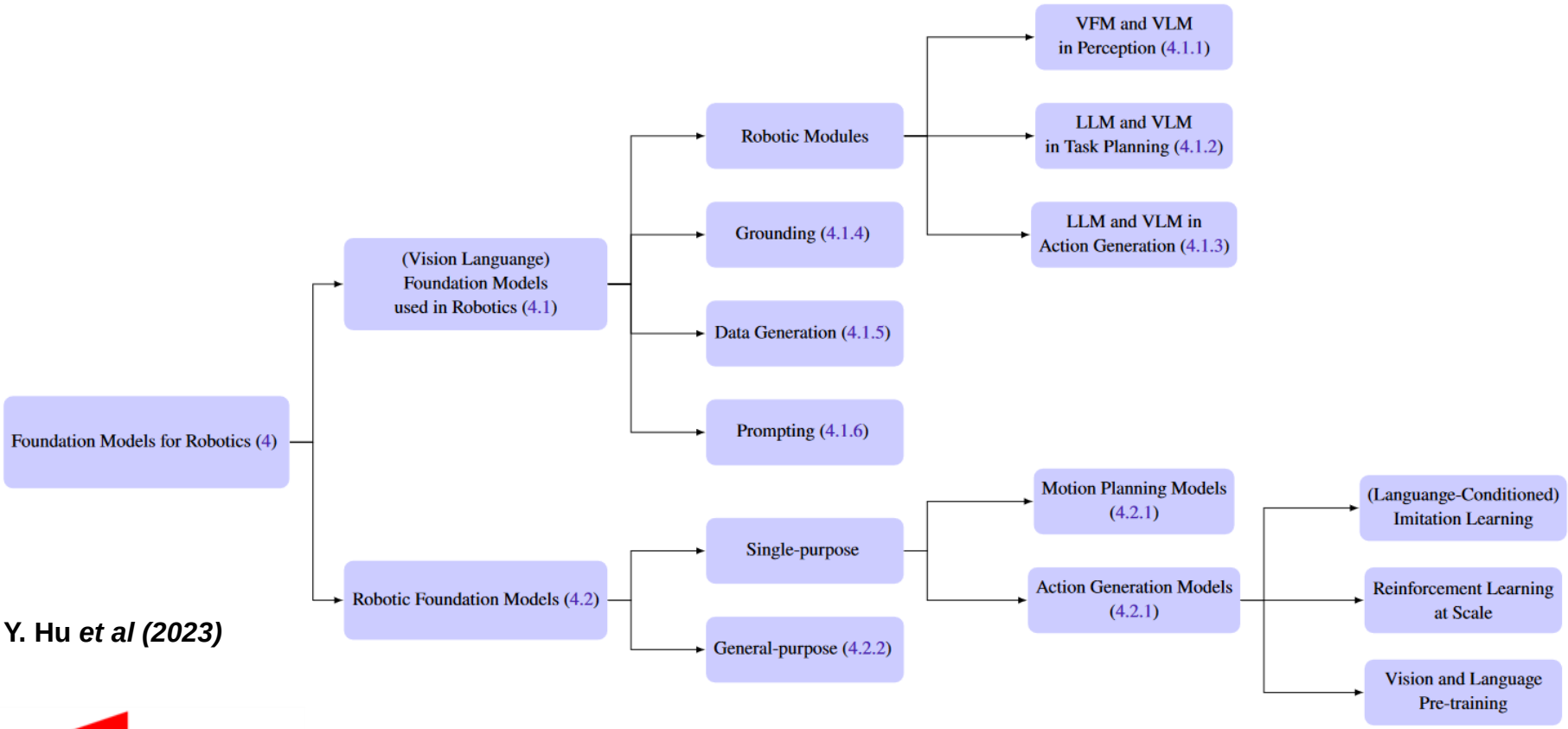


Y. Hu *et al (2023)*

D - Foundation Models for robotics

The term "foundation models for robotics" includes two distinct aspects:

1. application based on existing (mainly) vision and language models to robotics,
2. robotics foundation models specially for robotic tasks by using robot-generated data.



Y. Hu *et al* (2023)

D - Definitions

Large Language Models (LLM): A large language model (LLM) is a language model notable for its ability to achieve general-purpose language understanding and generation. LLMs acquire these abilities by learning statistical relationships from text documents during a computationally intensive self-supervised and semi-supervised training process. LLMs are artificial neural networks following a transformer architecture.

Vision-language model (VLM): A vision-language model is a fusion of vision and natural language models. These models leverage large-scale datasets and sophisticated neural network architectures, typically variants of transformers, to learn correlations between images and their textual descriptions or queries.

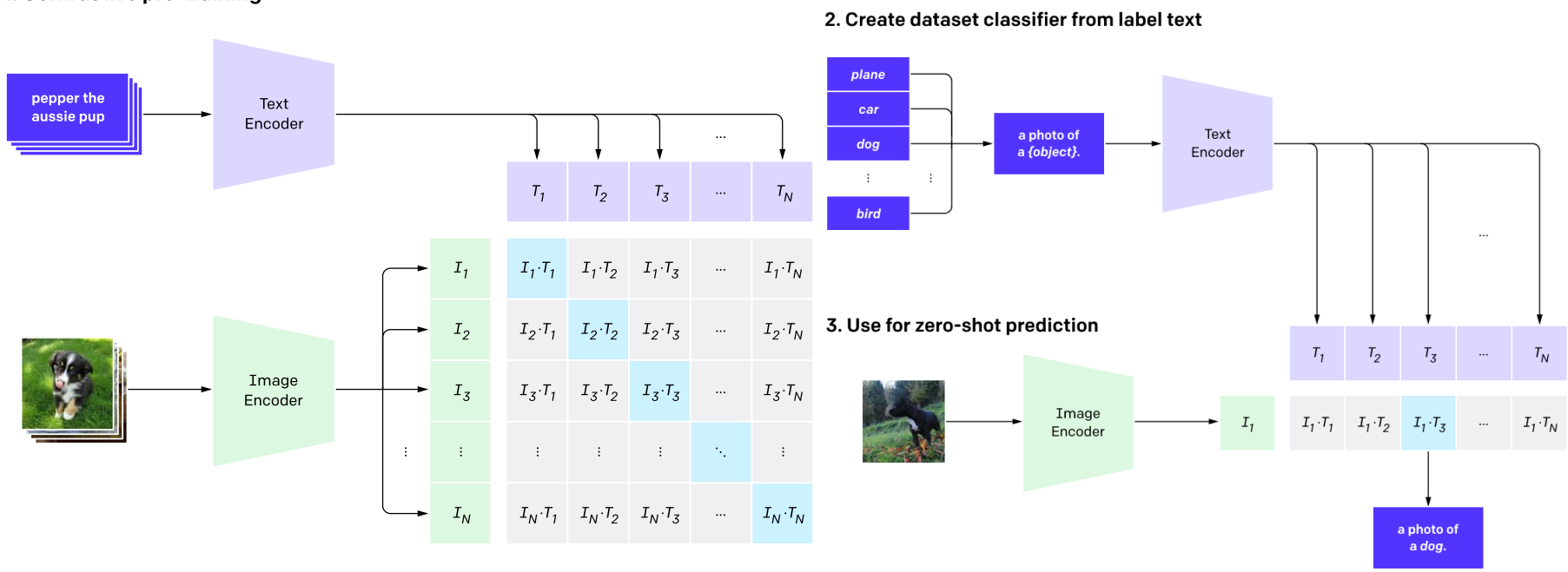
Large Multimodal Models (LMMs): Combination of several modalities (Vision, Language, audio, etc).

Y. Hu *et al* (2023)

D - CLIP

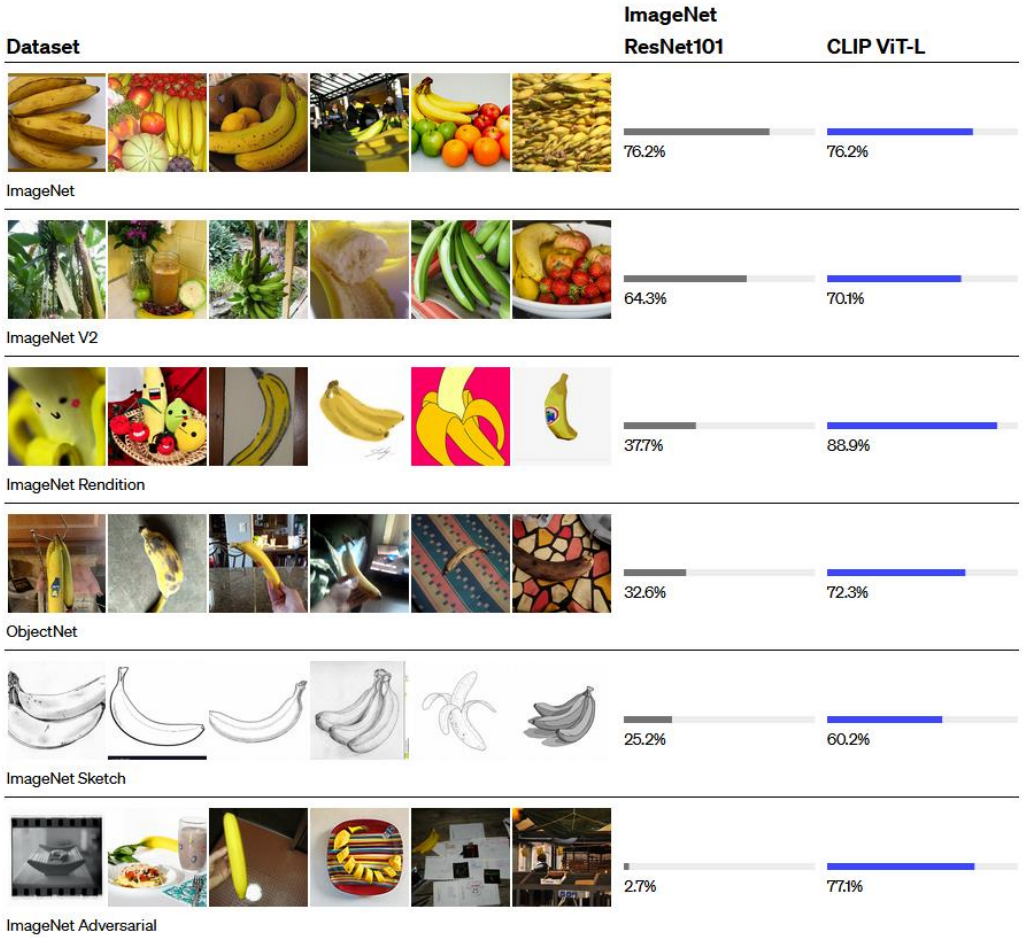
CLIP pre-trains an image encoder and a text encoder to predict which images were paired with which texts in our dataset. We then use this behavior to turn CLIP into a zero-shot classifier. We convert all of a dataset's classes into captions such as "a photo of a dog" and predict the class of the caption CLIP estimates best pairs with a given image. CLIP learns from text–image pairs that are already publicly available on the internet (The ImageNet dataset, one of the largest efforts in this space, required over 25,000 workers to annotate 14 million images for 22,000 object categories).

A. Radford *et al (2023)*



https://openai.com/research/clip

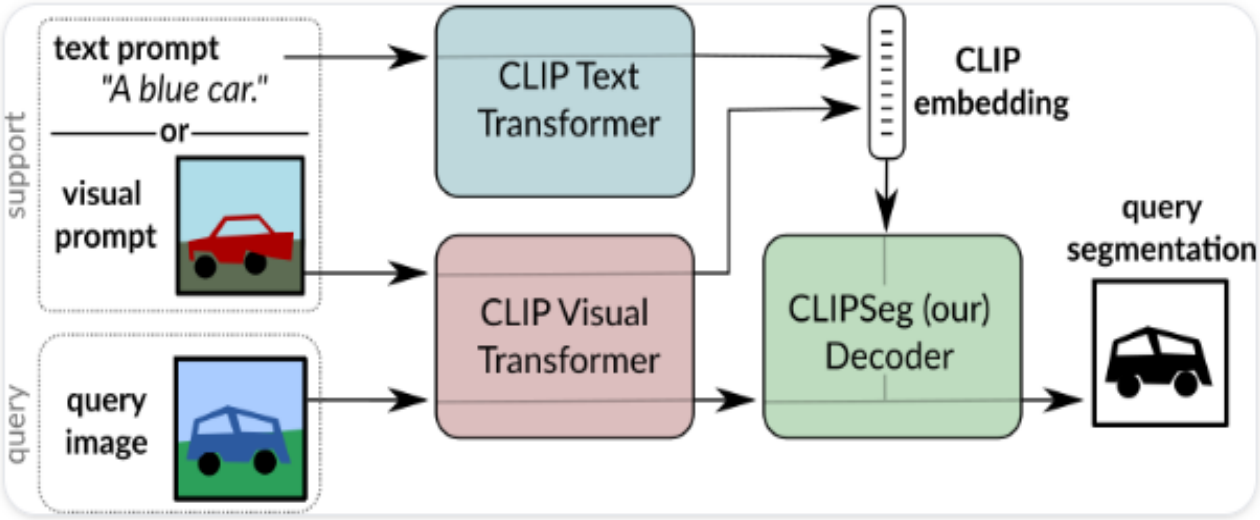
D - CLIP

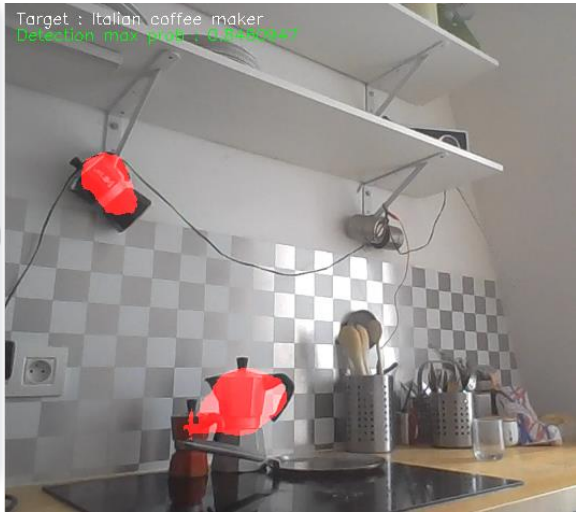


https://openai.com/research/clip

D - CLIPseg

The CLIPSeg model was proposed in Image Segmentation Using Text and Image Prompts by Timo Lüddecke and Alexander Ecker. CLIPSeg adds a minimal decoder on top of a frozen CLIP model for zero- and one-shot image segmentation.





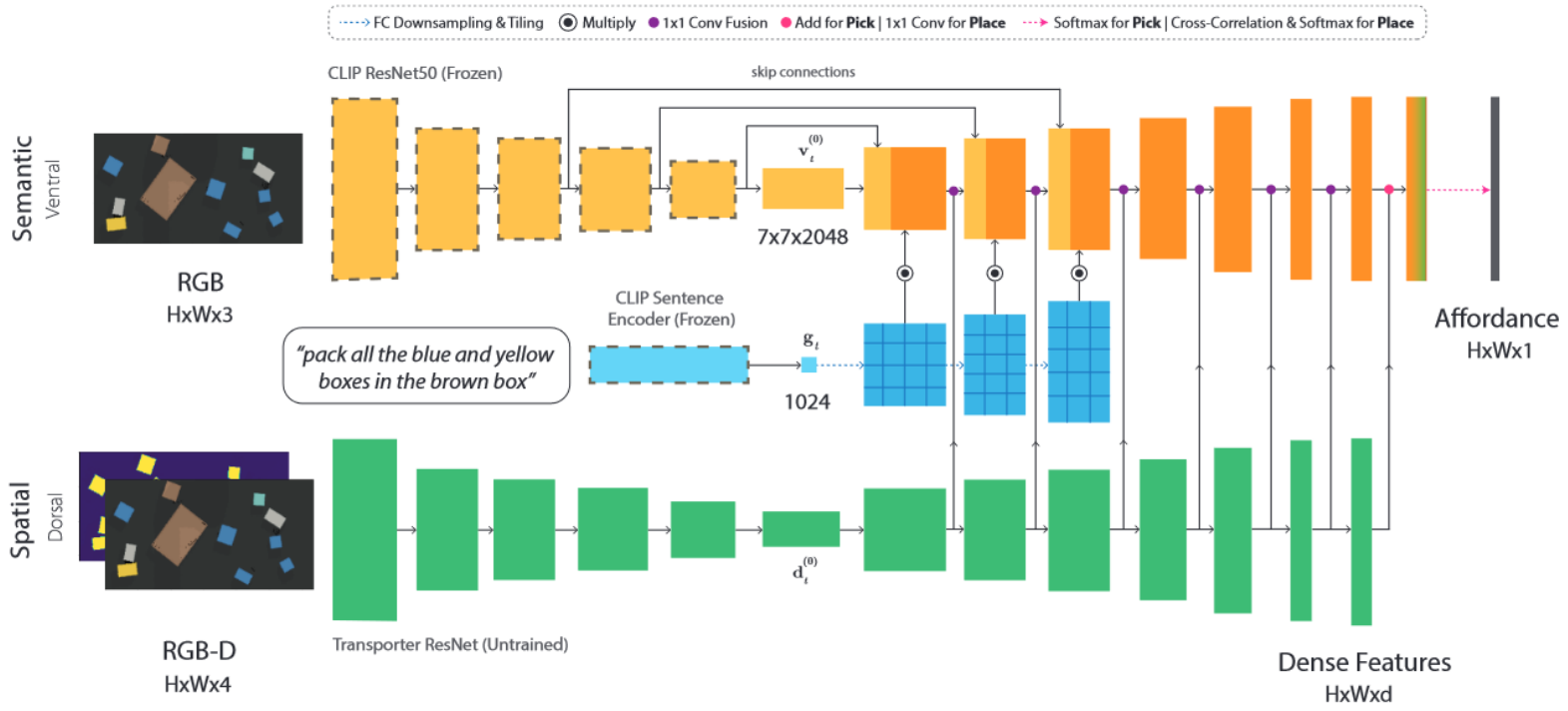
T. Lüddecke et A. S (2022)

https://github.com/timojl/clipseg

D - CLIPort

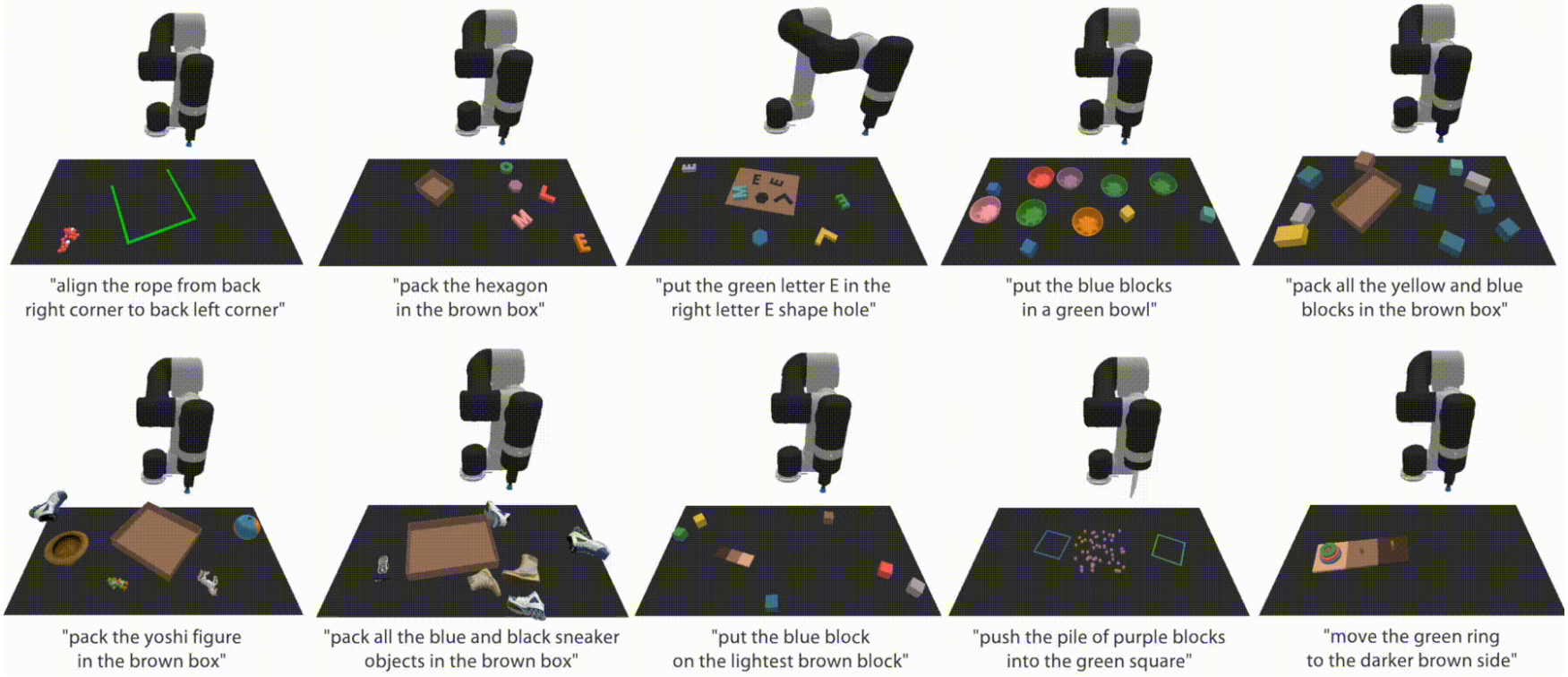
CLIPort combines :

- end-to-end learning for fine-grained manipulation with the multi-goal (**Transporter**)
- and multi-task generalization capabilities of vision-language grounding systems (**CLIP**).



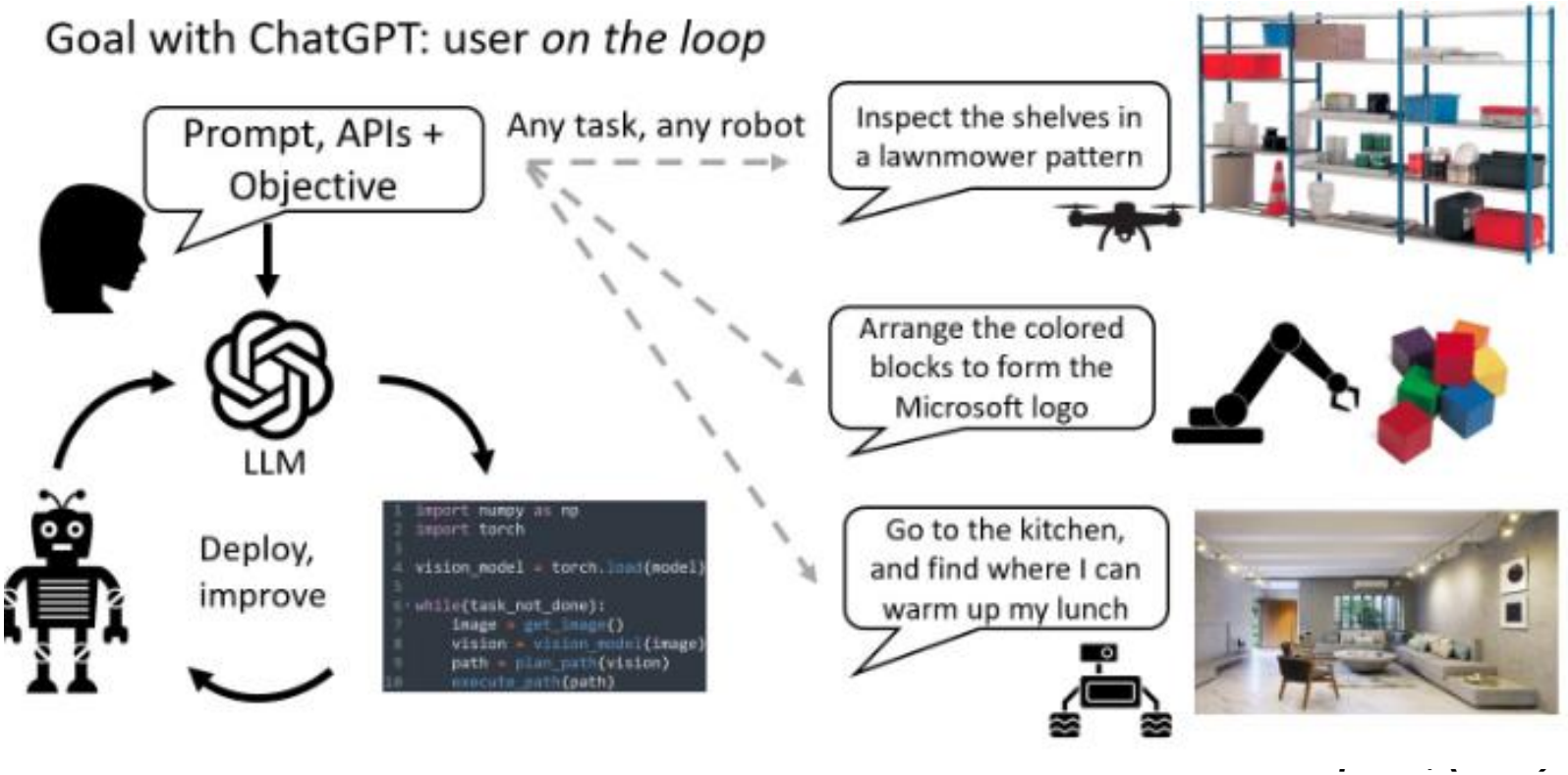
[M. Shridhar, et al (2021)

D - CLIPort



https://github.com/cliport/cliport

Microsoft Autonomous Systems and Robotics Research:

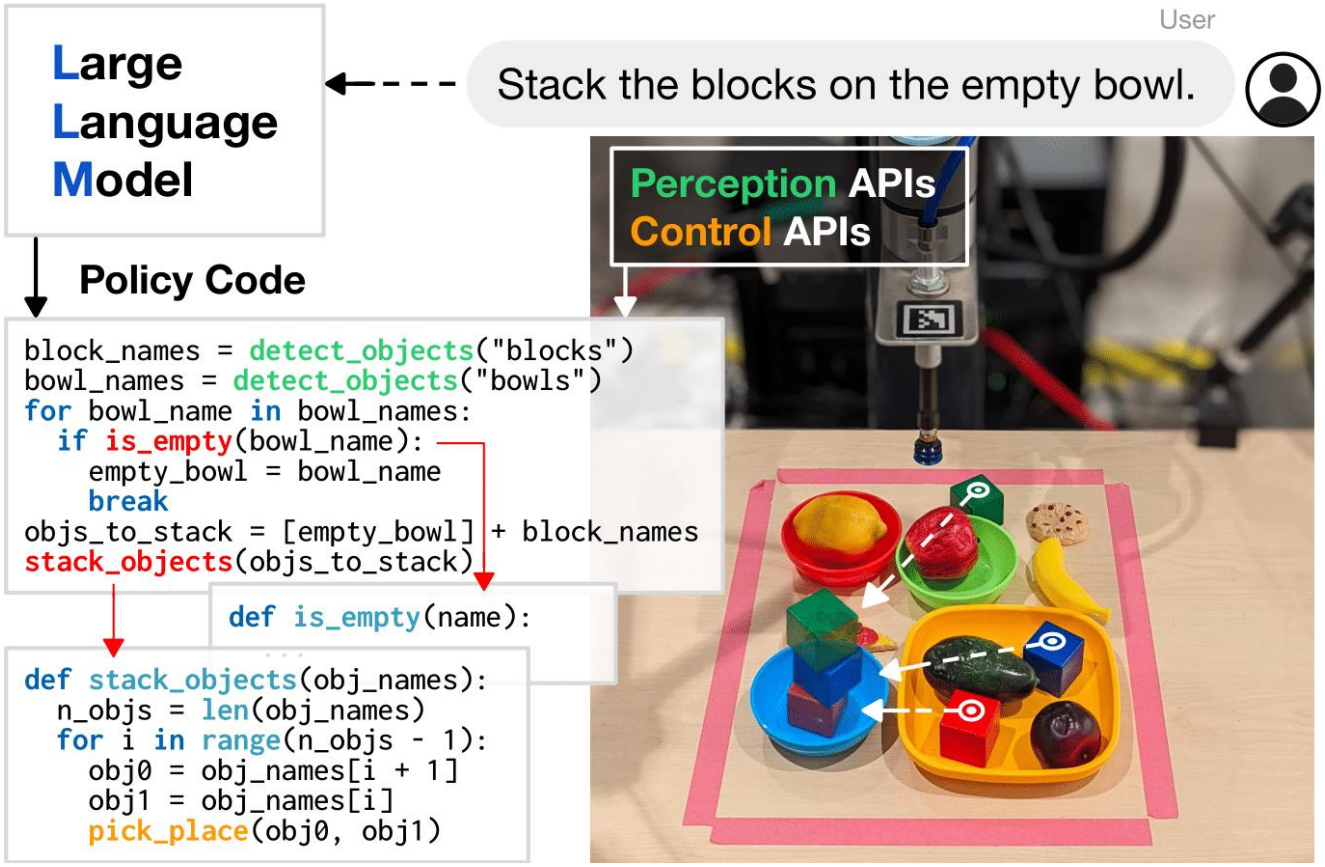


https://github.com/microsoft/PromptCraft-Robotics

D - ChatGPT for Robotics

https://youtu.be/wLOChUtdqoA

Robotics at Google:



J. Liang et al - (2023)

https://code-as-policies.github.io/

Code as Policies: Language Model Programs for Embodied Control

Jacky Liang, Wenlong Huang, Fei Xia, Peng Xu, Karol Hausman, Brian Ichter, Pete Florence, Andy Zeng

This video is voice-narrated

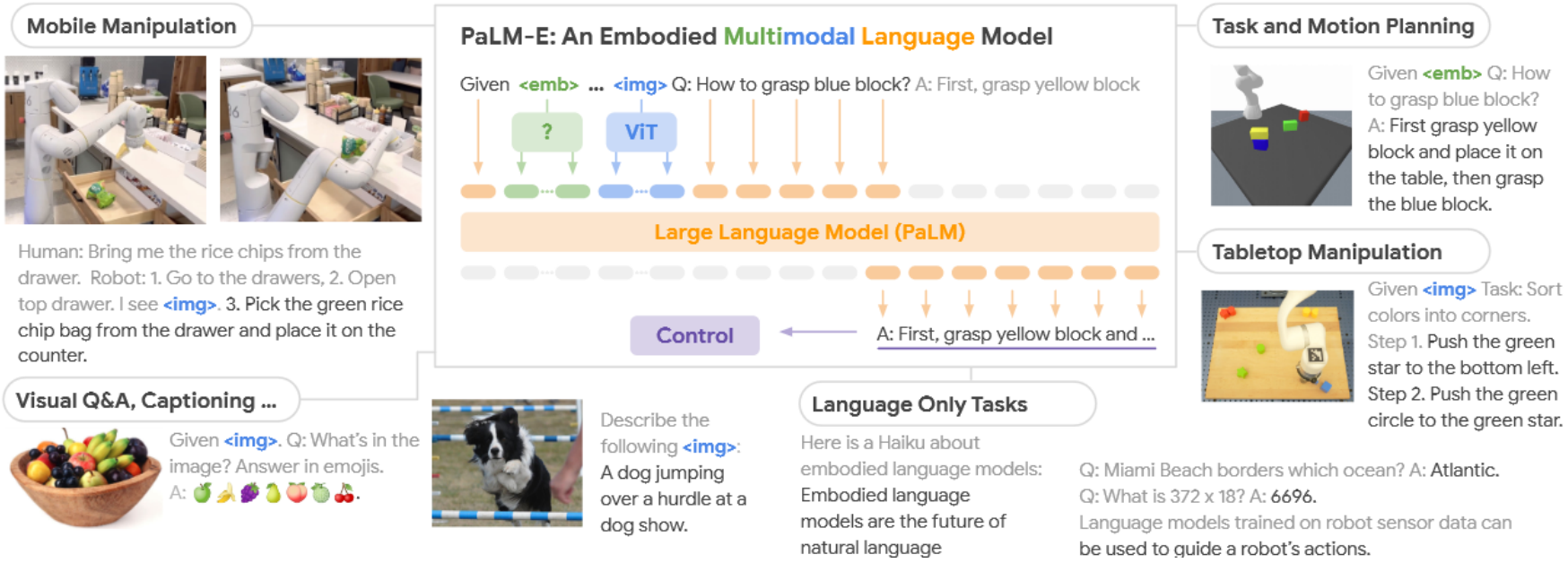
code-as-policies.github.io

Robotics at Google

https://code-as-policies.github.io/

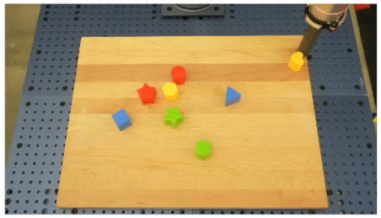
D - PaLM-E: An Embodied Multimodal Language Model

PaLM-E is a single general-purpose multimodal language model for embodied reasoning tasks, visual-language tasks, and language tasks (Robotics at Google TU Berlin, Google Research)



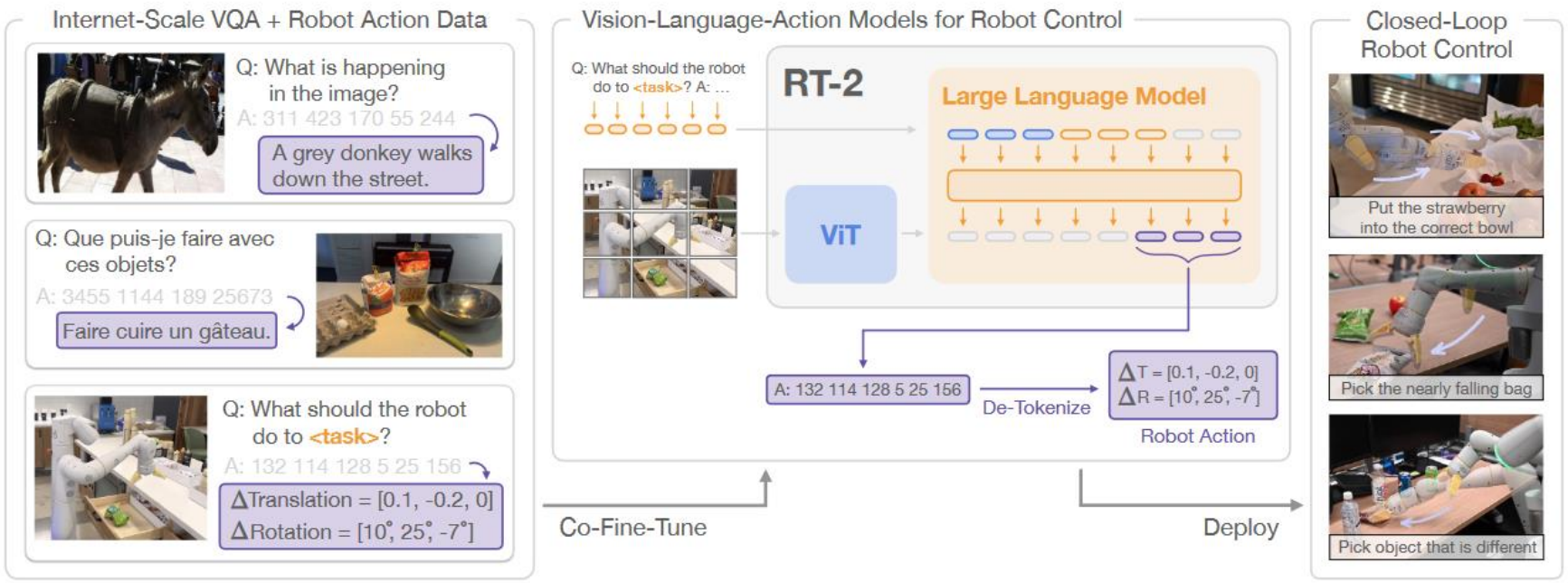
https://palm-e.github.io/

D - PaLM-E: An Embodied Multimodal Language Model

Given  . Q: How to sort the blocks by colors into the corners? A:

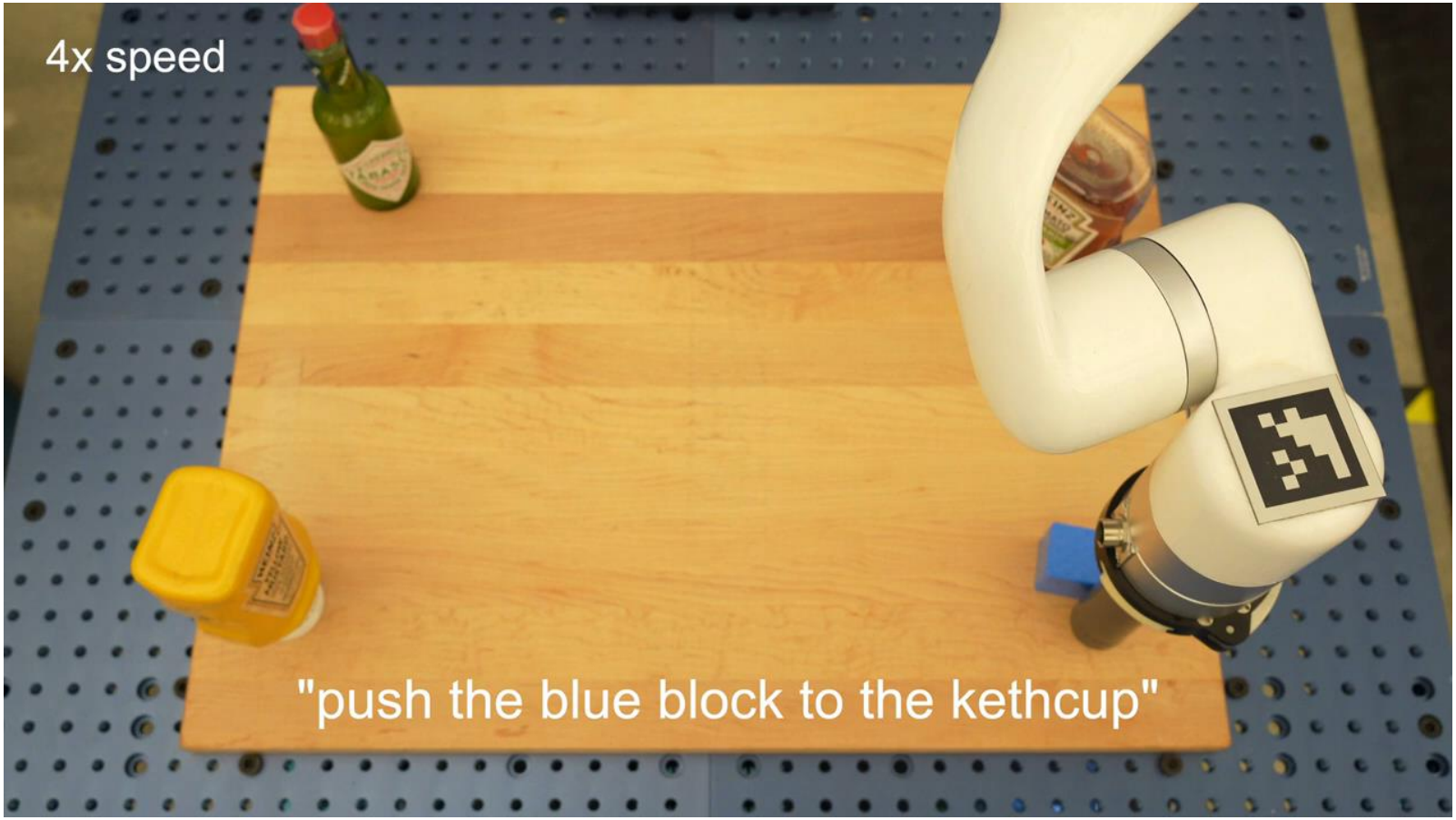
https://palm-e.github.io/

RT-2: Vision-Language-Action Models Transfer Web Knowledge to Robotic Control

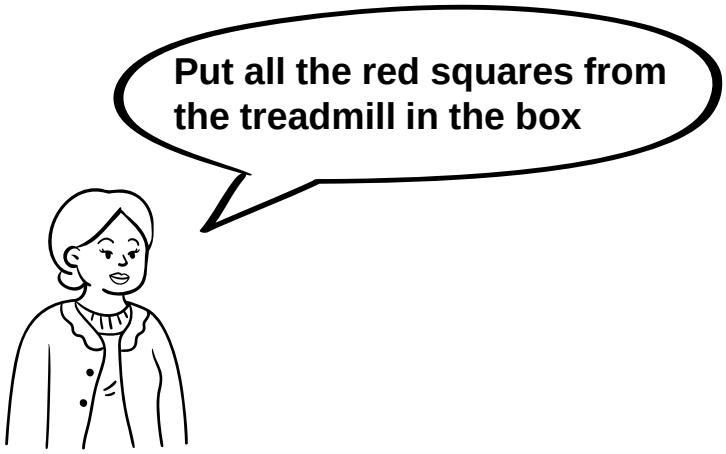


A. Brohan *et al.* – *(2023)*

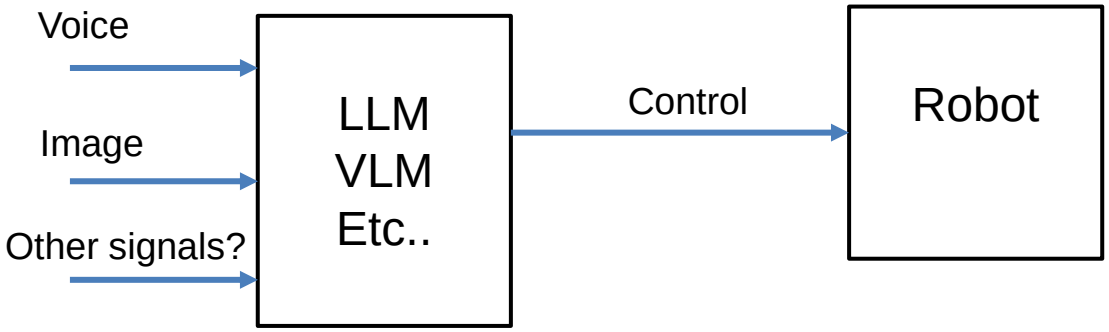
D - RT-2



D - Make easier collaboration between humans and cobots



Design a pipeline to make easier collaboration between humans and cobots :



D - Bibliography

Y. Hu *et al.*, « Toward General-Purpose Robots via Foundation Models: A Survey and Meta-Analysis ». arXiv, 15 décembre 2023. Disponible sur: http://arxiv.org/abs/2312.08782

A. Radford *et al.*, « Learning Transferable Visual Models From Natural Language Supervision ». arXiv, 26 février 2021. Disponible sur: http://arxiv.org/abs/2103.00020

T. Lüddecke et A. S. Ecker, « Image Segmentation Using Text and Image Prompts ». arXiv, 30 mars 2022. Consulté le: 14 avril 2023. [En ligne]. Disponible sur: http://arxiv.org/abs/2112.10003

M. Shridhar, L. Manuelli, et D. Fox, « CLIPort: What and Where Pathways for Robotic Manipulation ». arXiv, 24 septembre 2021.

S. Vemprala, R. Bonatti, A. Bucker, et A. Kapoor, « ChatGPT for Robotics: Design Principles and Model Abilities ». arXiv, 20 février 2023. Consulté le: 3 juillet 2023. [En ligne]. Disponible sur: http://arxiv.org/abs/2306.17582

J. Liang *et al.*, « Code as Policies: Language Model Programs for Embodied Control ». arXiv, 24 mai 2023. Consulté le: 26 juin 2023. [En ligne]. Disponible sur: http://arxiv.org/abs/2209.07753

D. Driess *et al.*, « PaLM-E: An Embodied Multimodal Language Model ».

A. Brohan *et al.*, « RT-2: Vision-Language-Action Models Transfer Web Knowledge to Robotic Control ».

Human Robot Collaboration based on deep learning and generative IA : Applications to Industry 5.0

Questions ?

Christophe SABOURIN
sabourin@u-pec.fr
Laboratoire Images, Signaux et Systèmes Intelligents (LISSI)

D - Bibliography

Y. Hu *et al.*, « Toward General-Purpose Robots via Foundation Models: A Survey and Meta-Analysis ». arXiv, 15 décembre 2023. Disponible sur: http://arxiv.org/abs/2312.08782

A. Radford *et al.*, « Learning Transferable Visual Models From Natural Language Supervision ». arXiv, 26 février 2021. Disponible sur: http://arxiv.org/abs/2103.00020

T. Lüddecke et A. S. Ecker, « Image Segmentation Using Text and Image Prompts ». arXiv, 30 mars 2022. Consulté le: 14 avril 2023. [En ligne]. Disponible sur: http://arxiv.org/abs/2112.10003

M. Shridhar, L. Manuelli, et D. Fox, « CLIPort: What and Where Pathways for Robotic Manipulation ». arXiv, 24 septembre 2021.

S. Vemprala, R. Bonatti, A. Bucker, et A. Kapoor, « ChatGPT for Robotics: Design Principles and Model Abilities ». arXiv, 20 février 2023. Consulté le: 3 juillet 2023. [En ligne]. Disponible sur: http://arxiv.org/abs/2306.17582

J. Liang *et al.*, « Code as Policies: Language Model Programs for Embodied Control ». arXiv, 24 mai 2023. Consulté le: 26 juin 2023. [En ligne]. Disponible sur: http://arxiv.org/abs/2209.07753

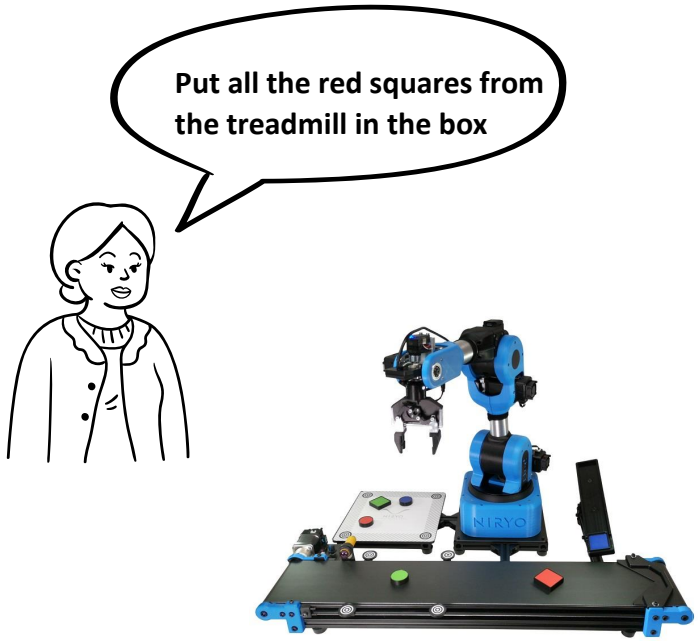
D. Driess *et al.*, « PaLM-E: An Embodied Multimodal Language Model ».

A. Brohan *et al.*, « RT-2: Vision-Language-Action Models Transfer Web Knowledge to Robotic Control ».

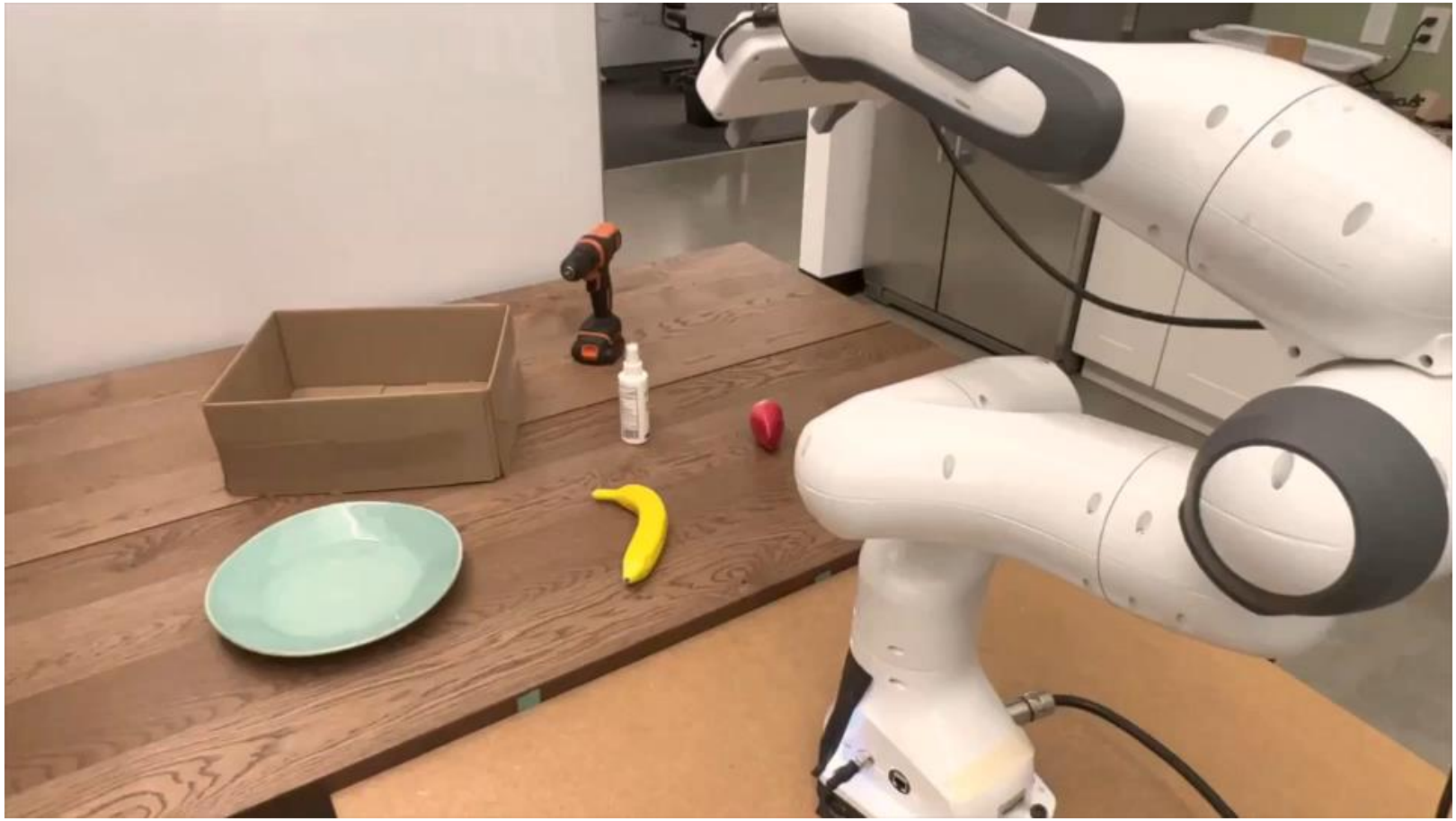
E – Use case

Design of an intelligent system able to control a robotic arm based on both a vocal command given by an operator and the visual perception of the environment. The figure below illustrates the problematic of the proposed appraoch:

- **An operator gives a voice command which is translated into textual form (Voice To Text)**
- **A perception system allows, based on visual information from the environment, to characterize objects in the environment (Vision Language Model)**
- **The intelligent system, based on textual and visual information, plans the execution of the task that will be carried out by the robot (LLM - Code Generation)**
- **The controller allows the execution of the task**



E – Use case



https://progprompt.github.io/

Questions :

1. What are the two possible approaches for this kind of problem? Give the advantages and disadvantages of each method?

2. Propose a “pipeline” in the case of a “hierarchical” approach and justify your choice.

3. Implement a semantic segmentation technique with an open vocabulary. Add a localization module

Biblio :

- Y. Hu *et al.*, « Toward General-Purpose Robots via Foundation Models: A Survey and Meta-Analysis ». arXiv, 15 décembre 2023. http://arxiv.org/abs/2312.08782

- A. Brohan et al., « RT-2: Vision-Language-Action Models Transfer Web Knowledge to Robotic Control ». https://arxiv.org/abs/2307.15818

- I. Singh *et al.*, « ProgPrompt: program generation for situated robot task planning using large language models », *Auton Robot*, août 2023, doi: 10.1007/s10514-023-10135-3.