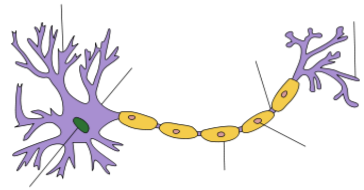
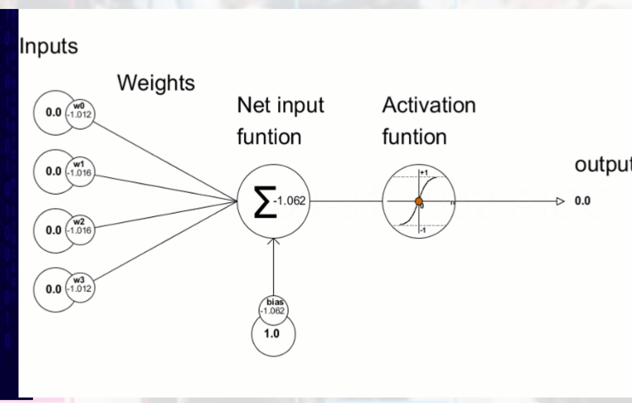
UMONS
Université de Mons



TRAIL TRUSTED AI LABS

TRAIL Seminar

Review of DNN optimization and compression Methods for Edge AI systems



Sidi Ahmed Mahmoudi, Mohamed Benkedadra, Maxime Glosener

UMONS
University of Mons

23 june 2023

ILIA UMONS

PLAN

Introduction

I. Deep Learning: how does it work ?

II. Main challenges of Deep Learning

III. Edge AI in Deep Learning : models compression

   a. Pruning
   b. Quantization
   c. Knowledge Distillation
   d. Discussion

III. Proposed approach of DNN compression

IV. Experimental results : Edge AI use cases

Conclusion

PLAN

Introduction

I. **Deep Learning: how does it work ?**

II. **Main challenges of Deep Learning**

III. **Edge AI in Deep Learning : models compression**

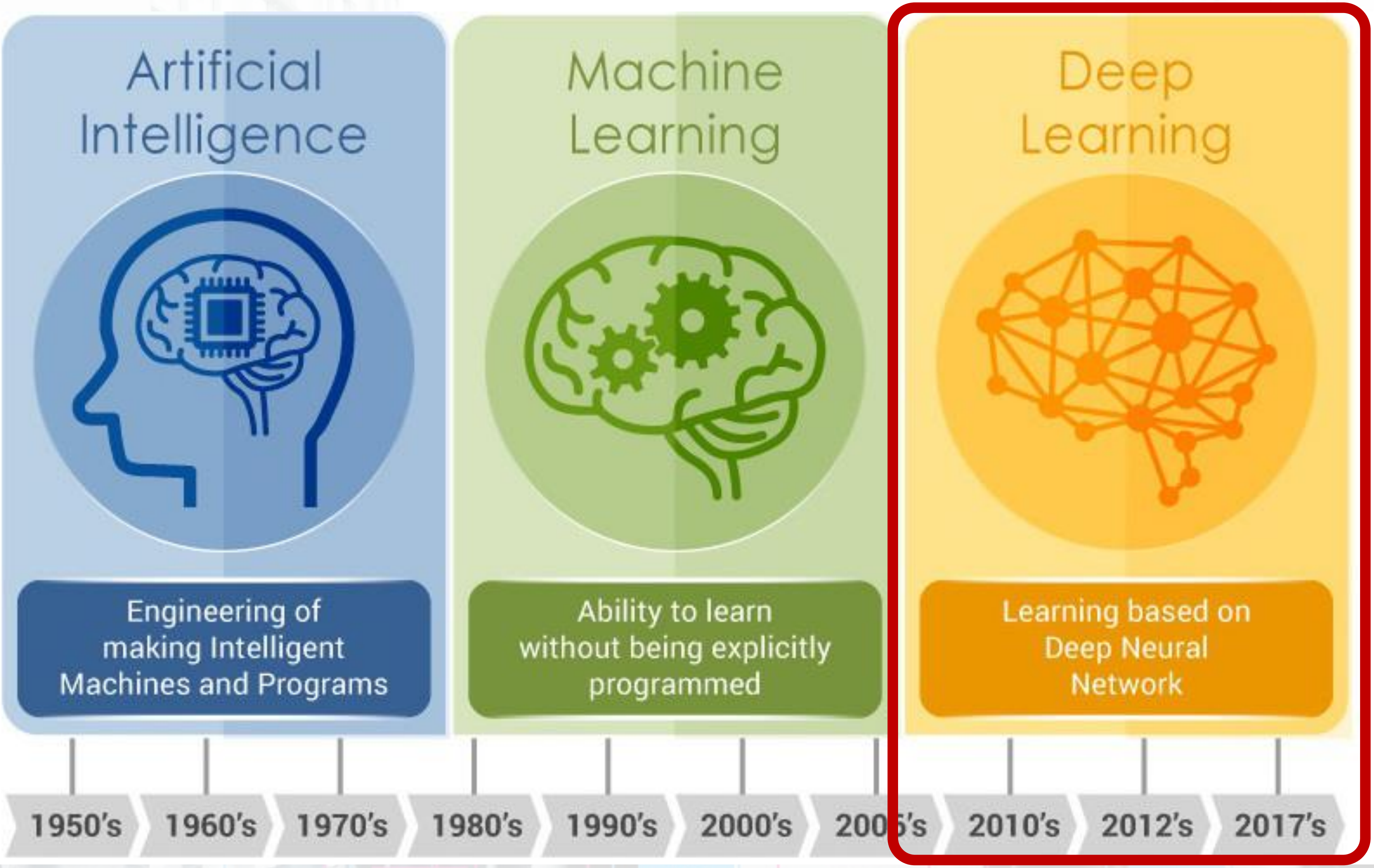
- **a.** **Pruning**
- **b.** **Quantization**
- **c.** **Knowledge Distillation**
- **d.** **Discussion**

III. **Proposed approach of DNN compression**

IV. **Experimental results : Edge AI use cases**

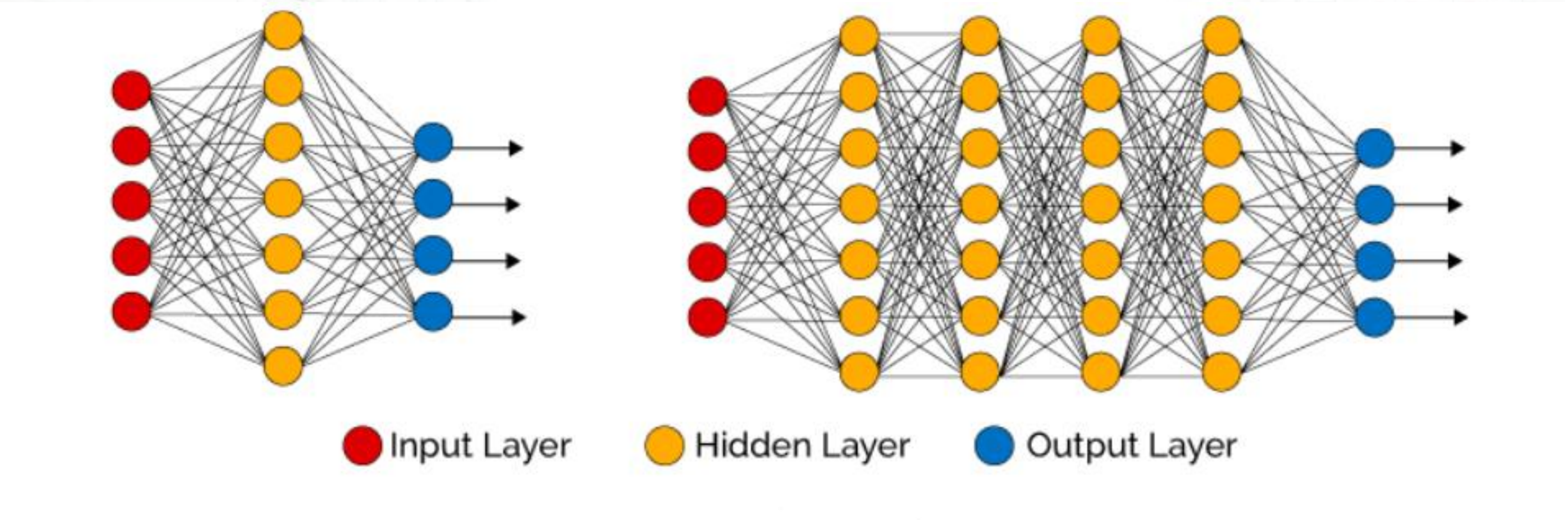
Conclusion

Introduction



Introduction

Shallow Neural Network | **Deep Neural Network**



PLAN

Introduction

I. Deep Learning: how does it work ?

II. Main challenges of Deep Learning

III. Edge AI in Deep Learning : models compression

a. Pruning

b. Quantization

c. Knowledge Distillation

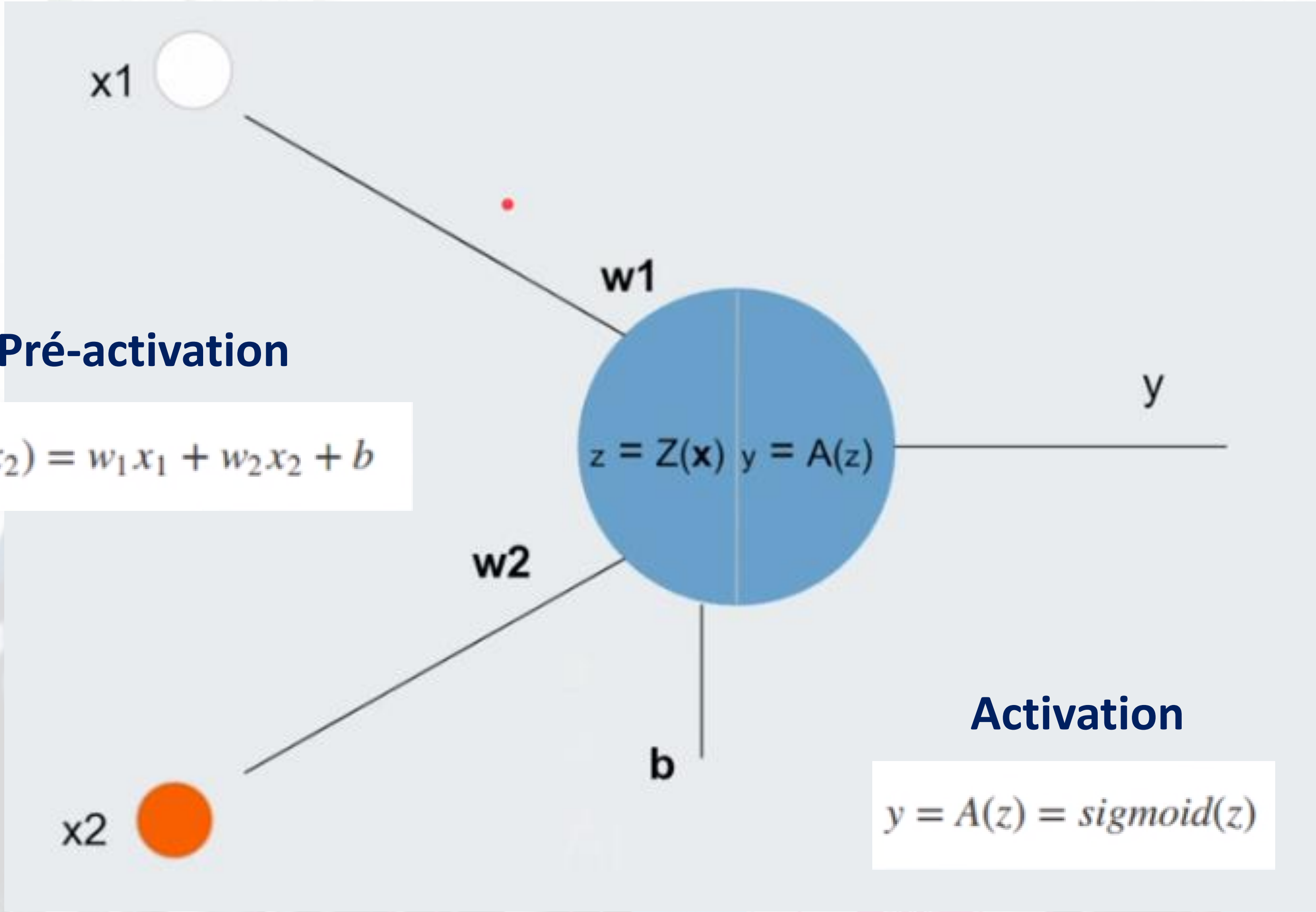
d. Discussion

III. Proposed approach of DNN compression

IV. Experimental results Edge AI use cases

Conclusion

Deep Learning : how does it work ?



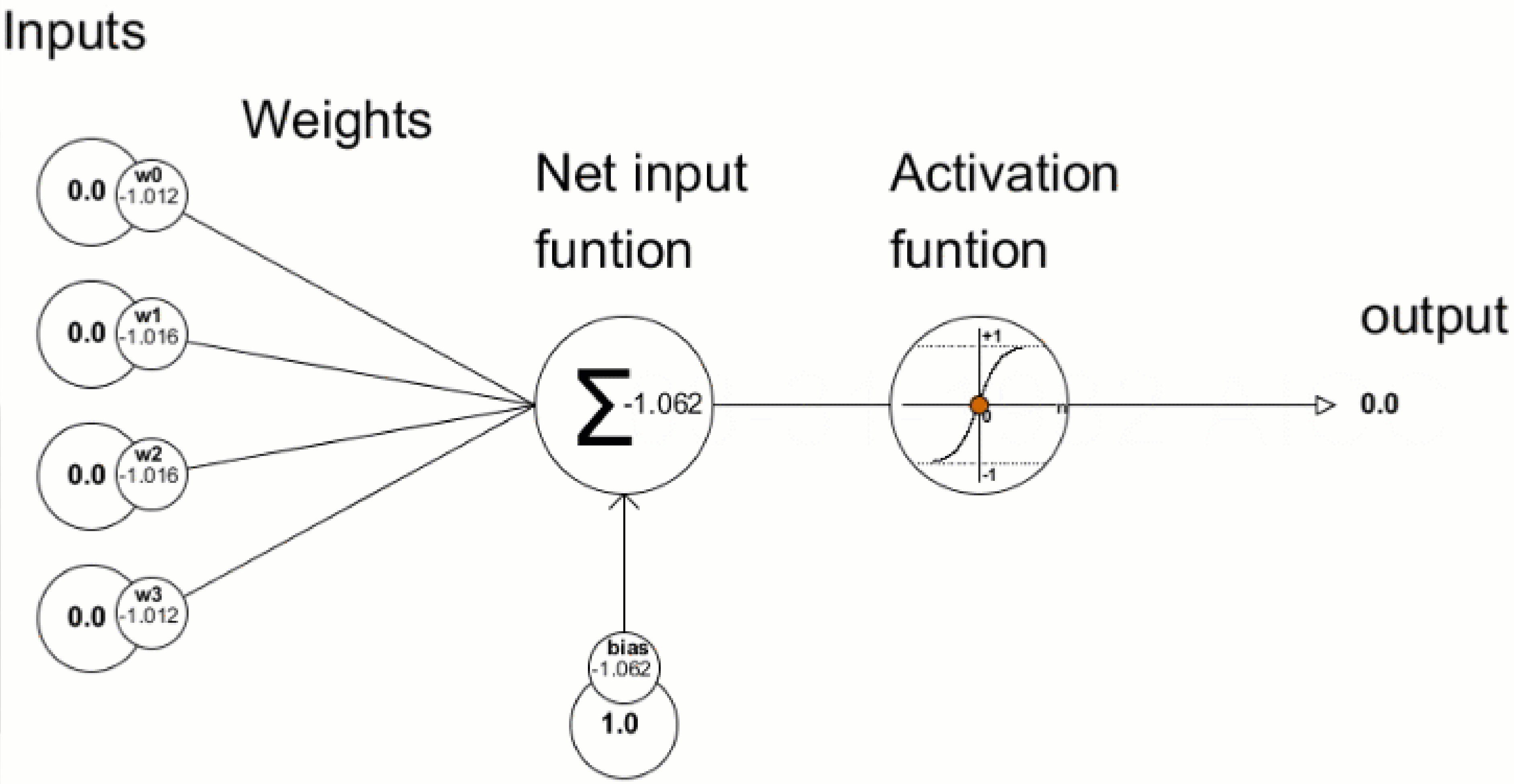
Pré-activation

$$Z(x_1, x_2) = w_1 x_1 + w_2 x_2 + b$$

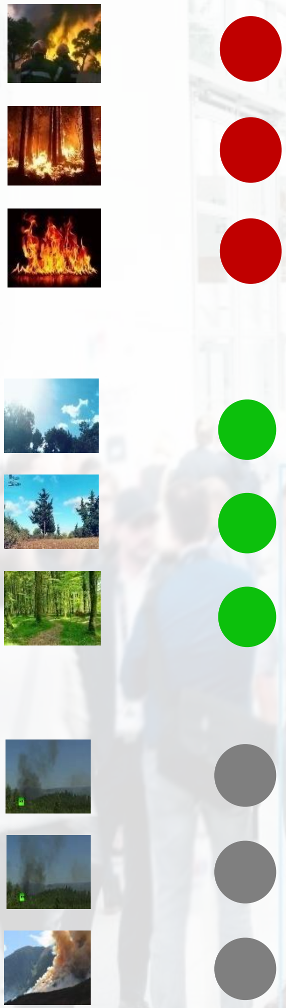
Activation

$$y = A(z) = sigmoid(z)$$

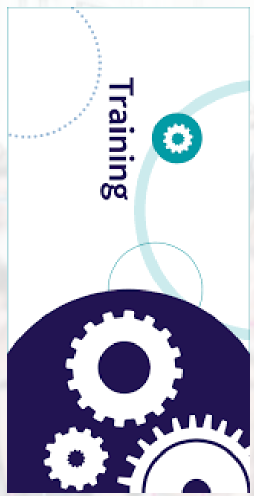
Deep Learning : how does it work ?

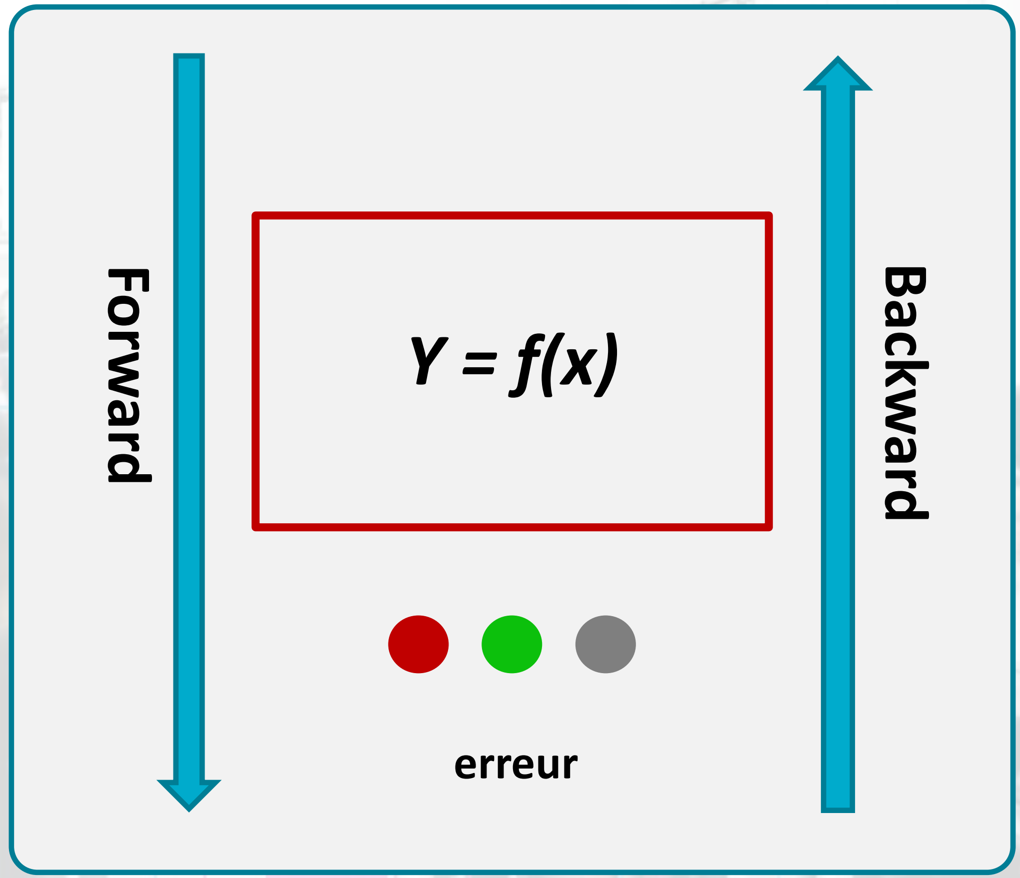


Exemple

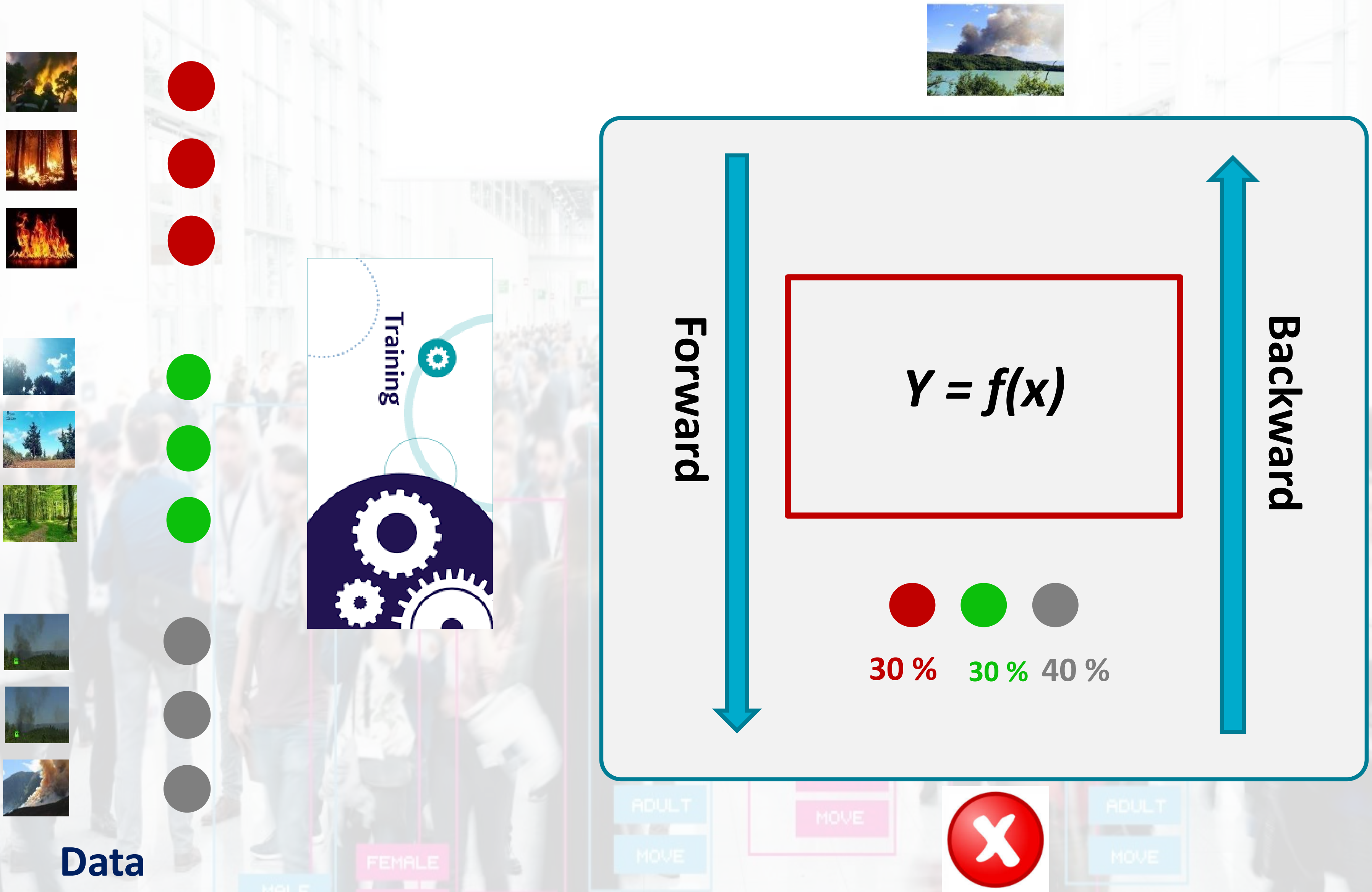


Data

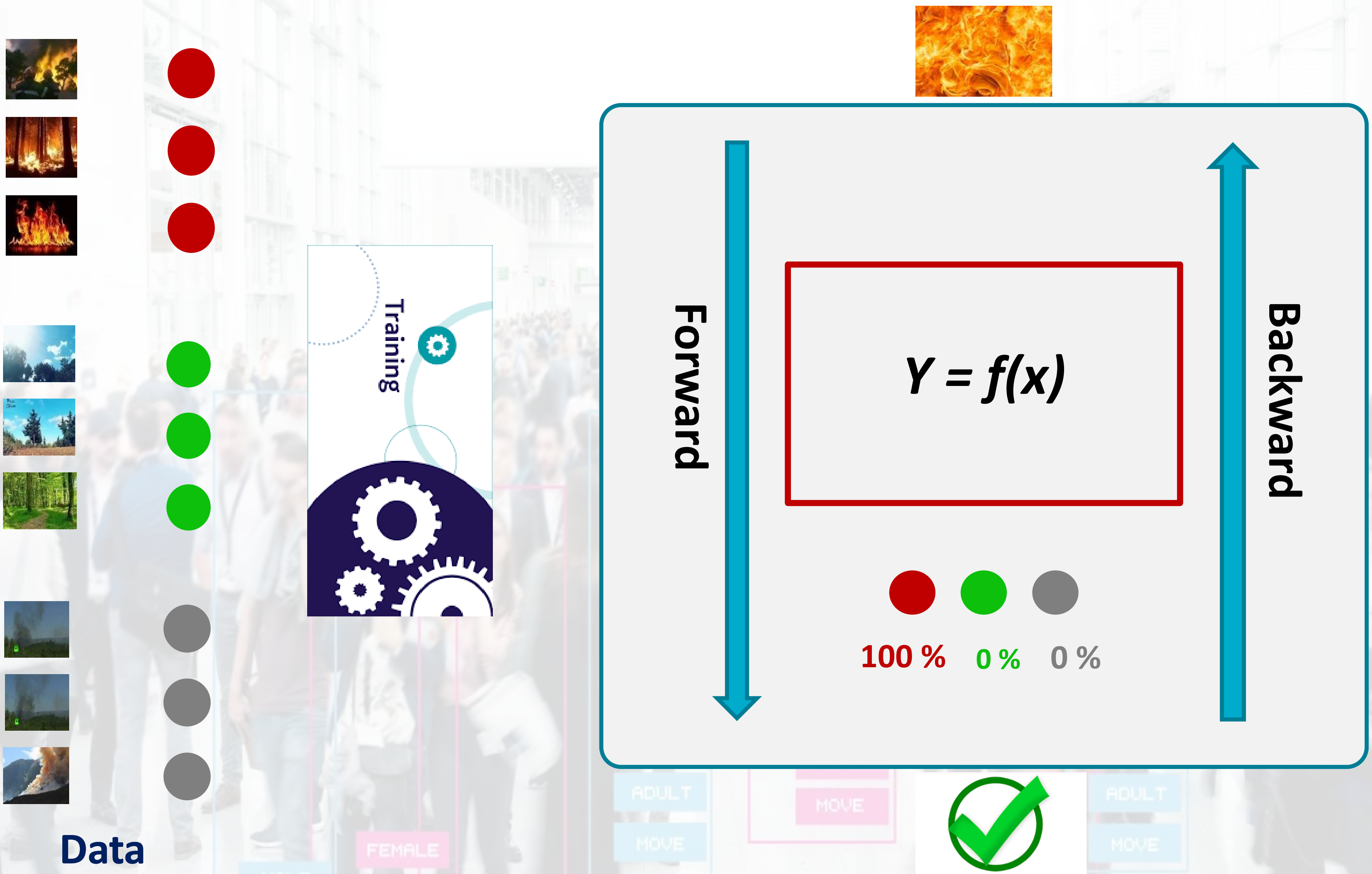




Exemple



Exemple



Deep Learning : global process

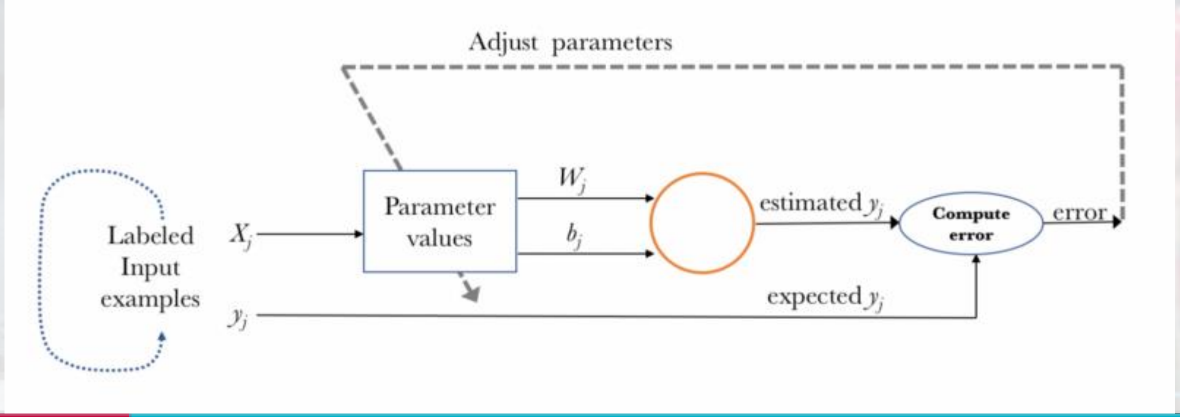
a. **Initialization :** Init weights (W) and bias (b) with random values

b. **Forward Pass :** get predictions using the proposed neural network

c. **Error calculation :** compare predicted values vs. real values

d. **Backpropagation :** update the weights using gradient descent

e. **Iterate :** Repeat the previous steps until get efficient model.

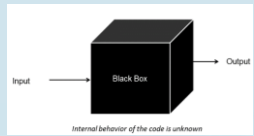


PLAN

Main challenges of AI and Deep Learning

Explainable & interpretability ?

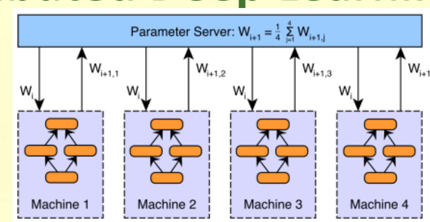
- Can we trust DL models ?
- Why ? When ? Etc.



- **Explainability:** justify each action
- **Interpretability:** explain in understandable terms

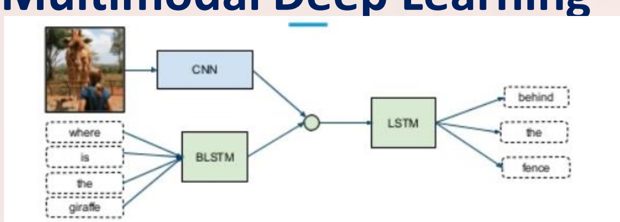
Slow training process?

- Accelerate the training process ?
- Exploit dist. res (CPUs/GPUs)
- **Distributed Deep Learning**



Multimodal Learning ?

- Manage different modalities
- Images, videos, text, signals, etc.
- **Multimodal Deep Learning**



Edge AI ?

- Memory space ? Comp. time ?
- Portability on Edge AI devices ?
- Compromise **Mem/Time/Acc**
- **Edge AI**

PLAN

Introduction

I. Deep Learning: how does it work ?

II. Main challenges of Deep Learning

III. Edge AI in Deep Learning : models compression

- a. Pruning
- b. Quantization
- c. Knowledge Distillation
- d. Discussion

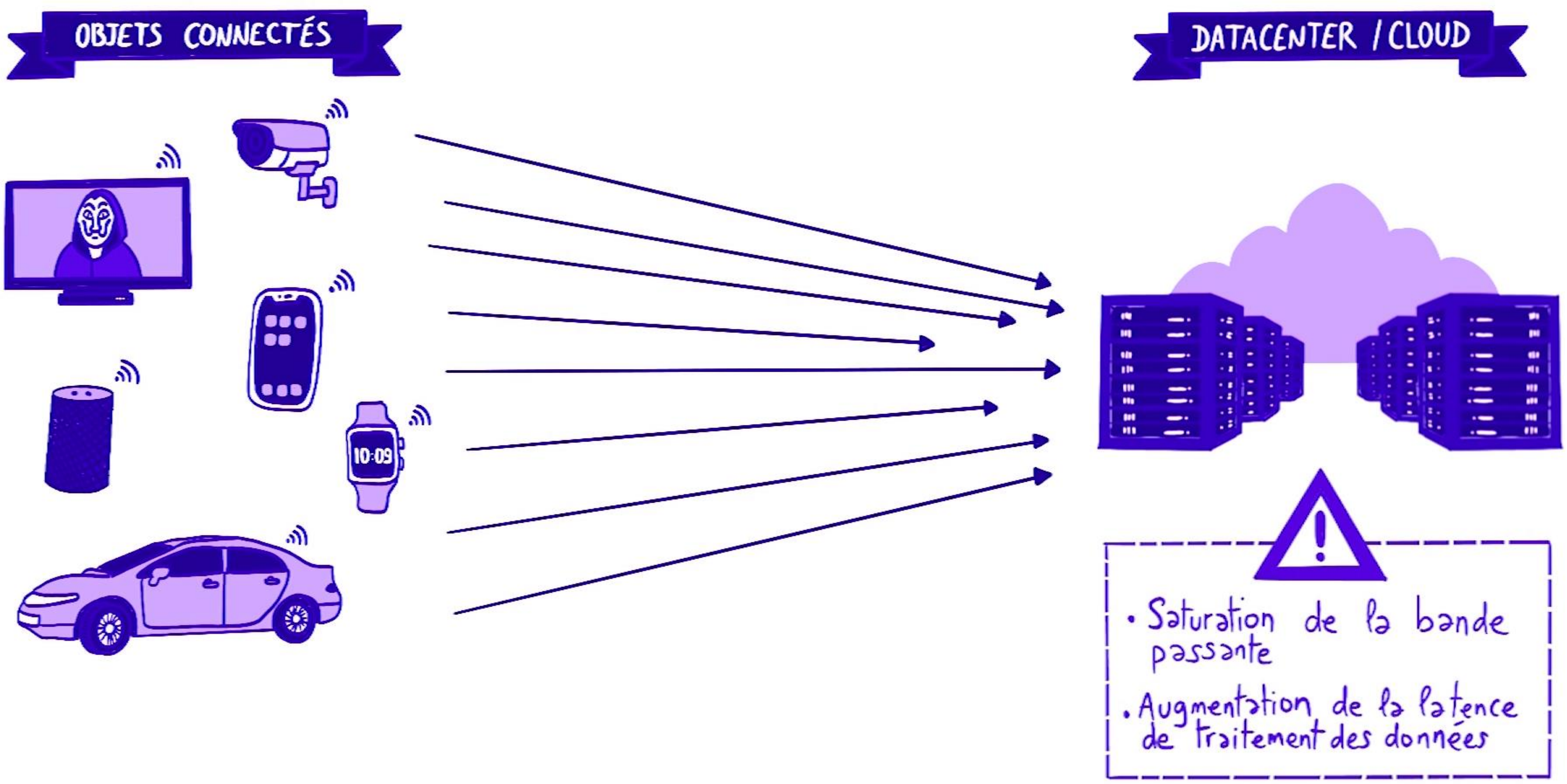
III. Proposed approach of DNN compression

IV. Experimental results : Edge AI use cases

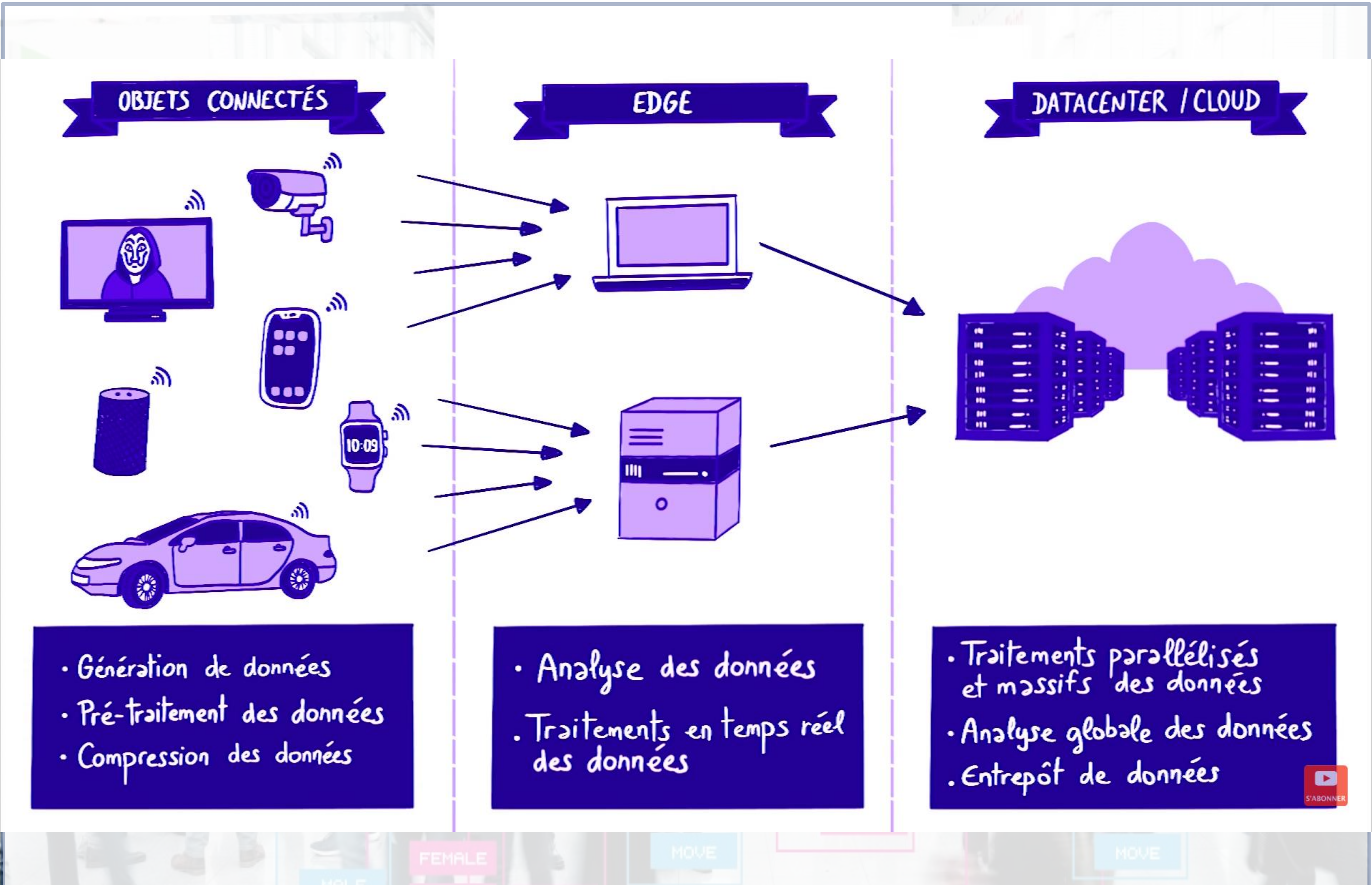
Conclusion

Why Edge Computing ?

- **IoT:** network of connected and embedded objects with sensors, software, etc.
- **IoT:** a high number of connected objects with smart objects, homes, cities, etc



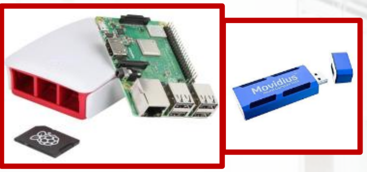
Why Edge Computing ?



Edge AI challenges

- Different embedded devices for AI "Edge AI Devices" **Edge AI Hardware**

Jetson



Raspberry Pi **TK1** **Nano** **Xavier** **Orin**

- **Edge AI devices :** limited power, memory and storage and energy efficient
- **Problem :** High needs of DL models (computation power, RAM, storage, etc.) mainly for real time applications using HD or Full HD cameras

- Challenge: **Optimisation & compression of DL models in order to be deployed on Edge AI resources with the maintain of a good accuracy**

PLAN

Introduction

I. Deep Learning: how does it work ?

II. Main challenges of Deep Learning

III. Edge AI in Deep Learning : models compression

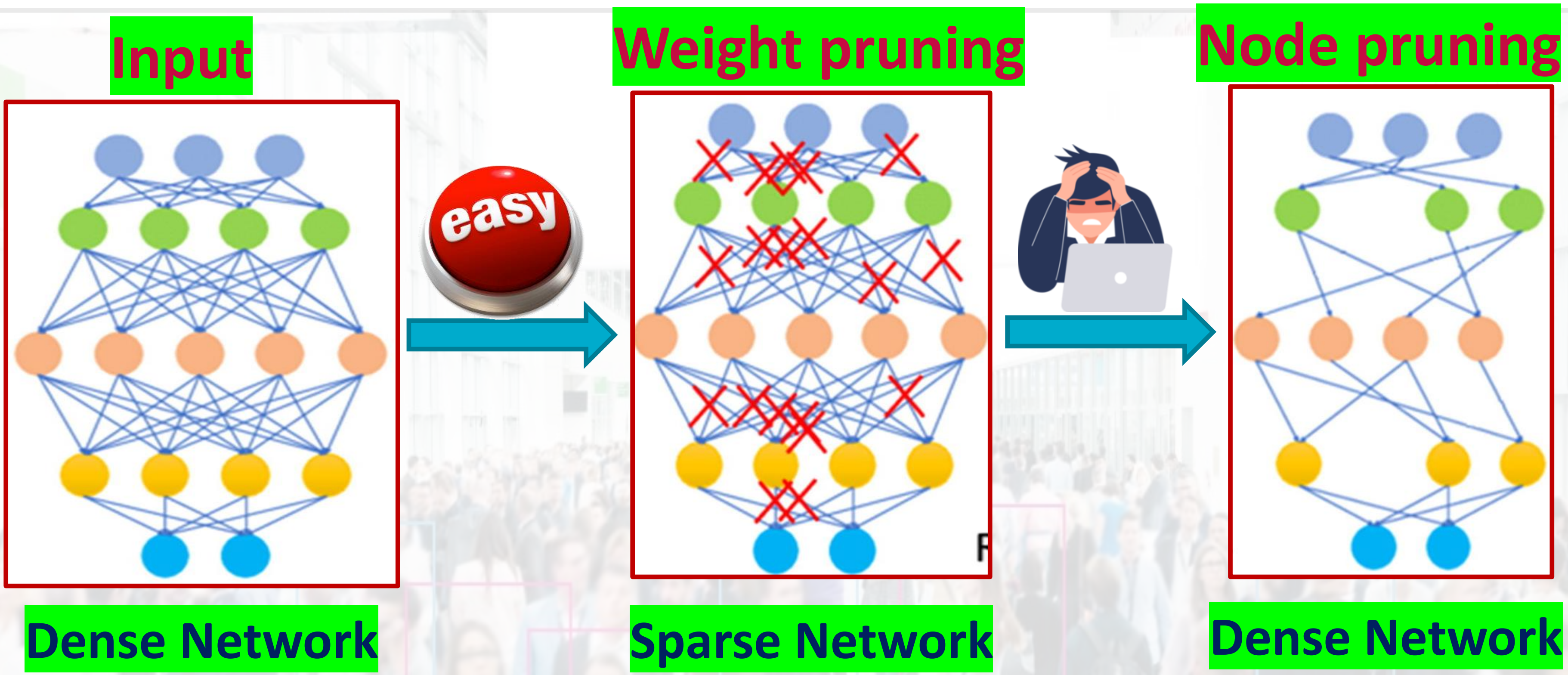
  a. Pruning
  b. Quantization
  c. Knowledge Distillation
  d. Discussion

III. Proposed approach of DNN compression

IV. Experimental results : Edge AI use cases

Conclusion

Related work : Pruning

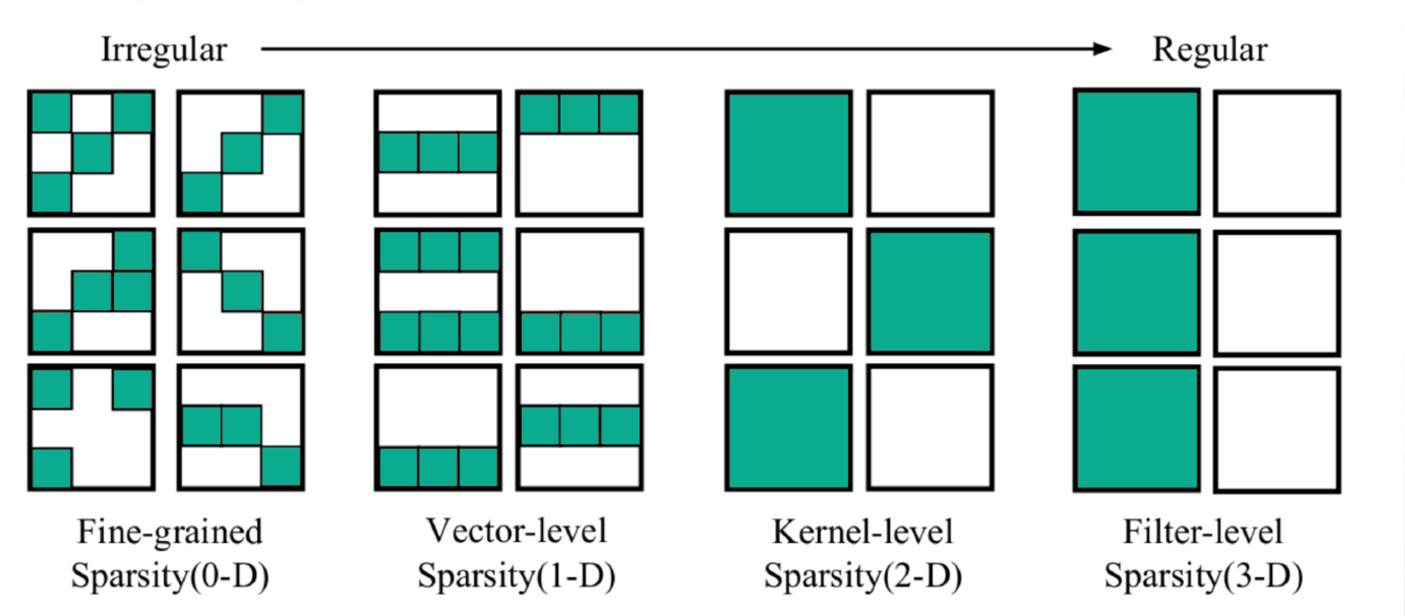


4 Questions about pruning :

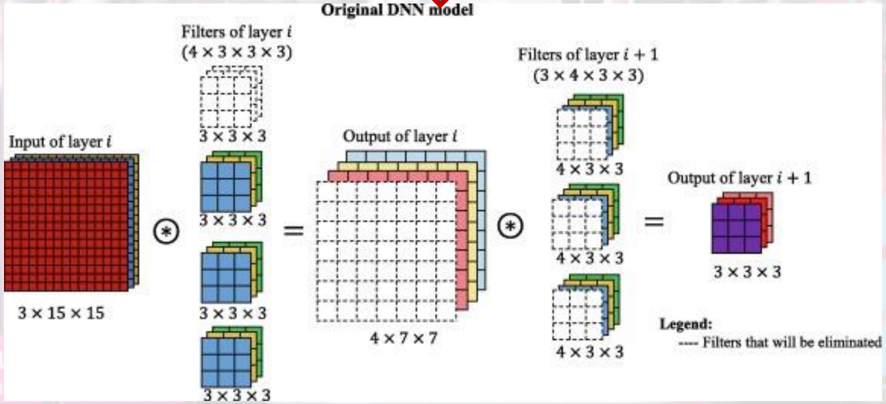
- How ?
- Where ?
- What ?
- When ?

Related work : Pruning

- **How to prune ?**

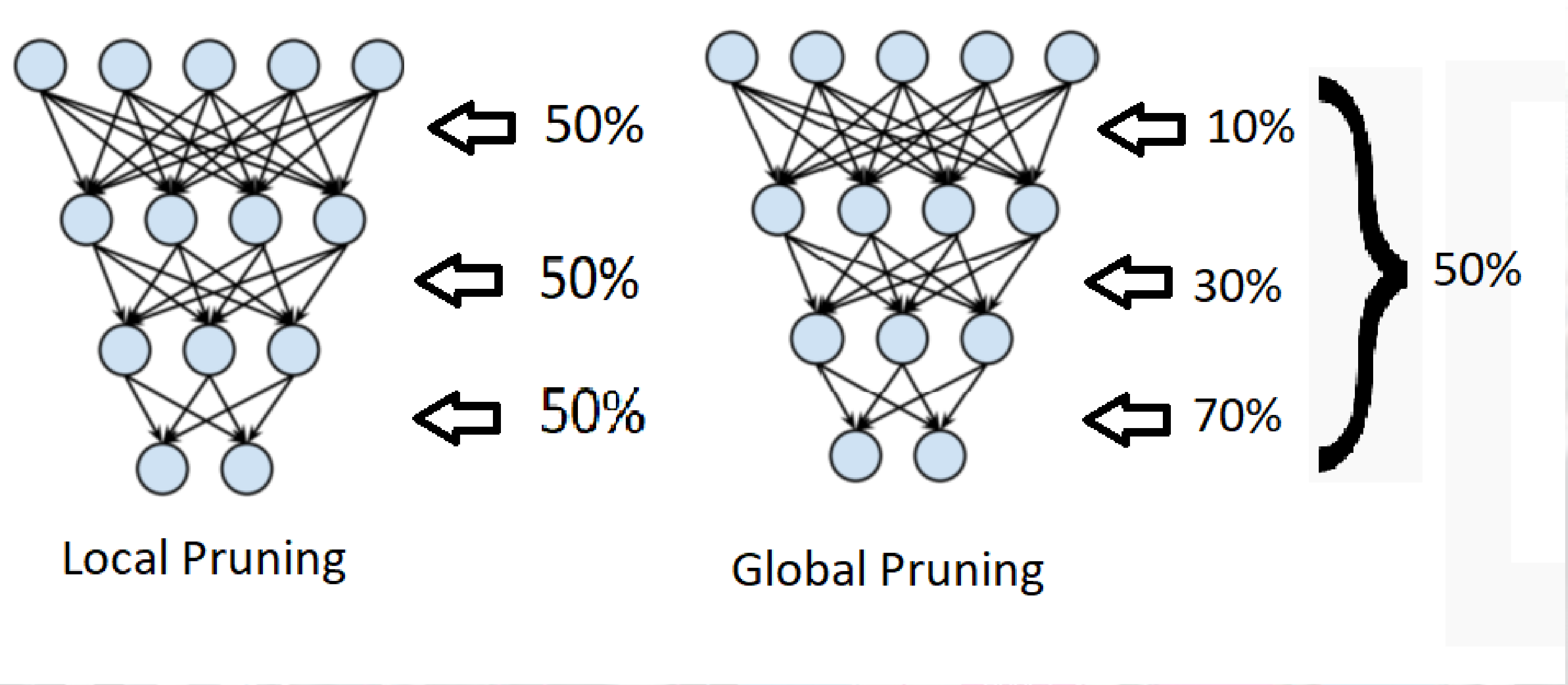


Kernel level sparsity example



Related work : Pruning

- **Where to prune ?**



Related work : Pruning

- What to prune ?

- **Magnitude based pruning**

$$threshold(w_i) = \begin{cases} w_i \text{ if } |w_i| > \lambda \\ 0 \text{ if } |w_i| \leq \lambda \end{cases}$$

- **Movement based pruning**

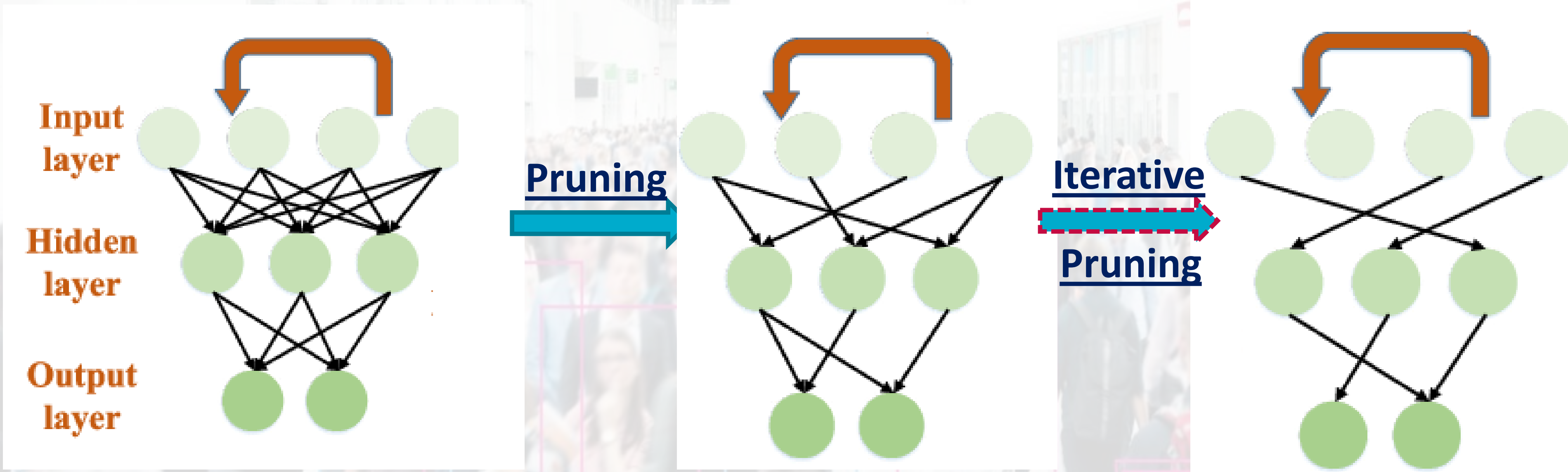
$$\sum_t \left(\frac{\partial \mathcal{L}}{\partial W_{i,j}} \right)^{(t)} W_{i,j}^{(t)}$$

- **Node Pruning**

Related work : Pruning

- **When to prune ?**

Initial training | **Retraining** | **Retraining**



One iteration | **Many iterations**

One Shot pruning

Iterative pruning

PLAN

Introduction

I. Deep Learning: how does it work ?

II. Main challenges of Deep Learning

III. Edge AI in Deep Learning : models compression

a. Pruning

b. Quantization

c. Knowledge Distillation

d. Discussion

III. Proposed approach of DNN compression

IV. Experimental results : Edge AI use cases

Conclusion

Related work : Quantization

-0.2	1	0.3
0.1	-0.6	-0.7
1.2	0.4	0

32 bit

→ quantization →

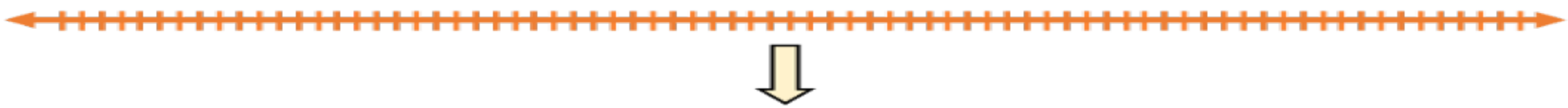
1	3	2
1	0	0
3	2	1

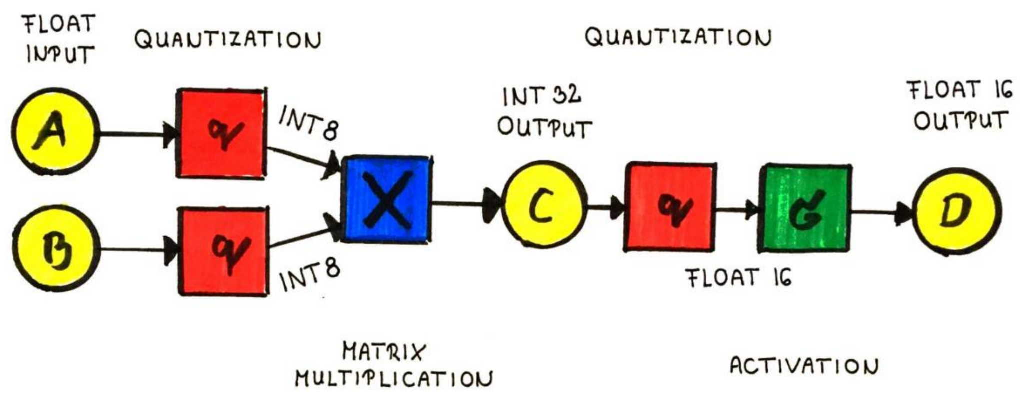
2 bit

+

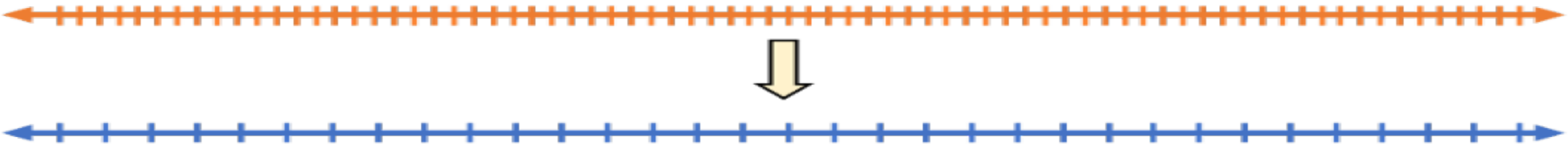
index	[in bits]	value
0	[00]	-0.6
1	[01]	0
2	[10]	0.4
3	[11]	1.1

32 bit





Related work : Quantization



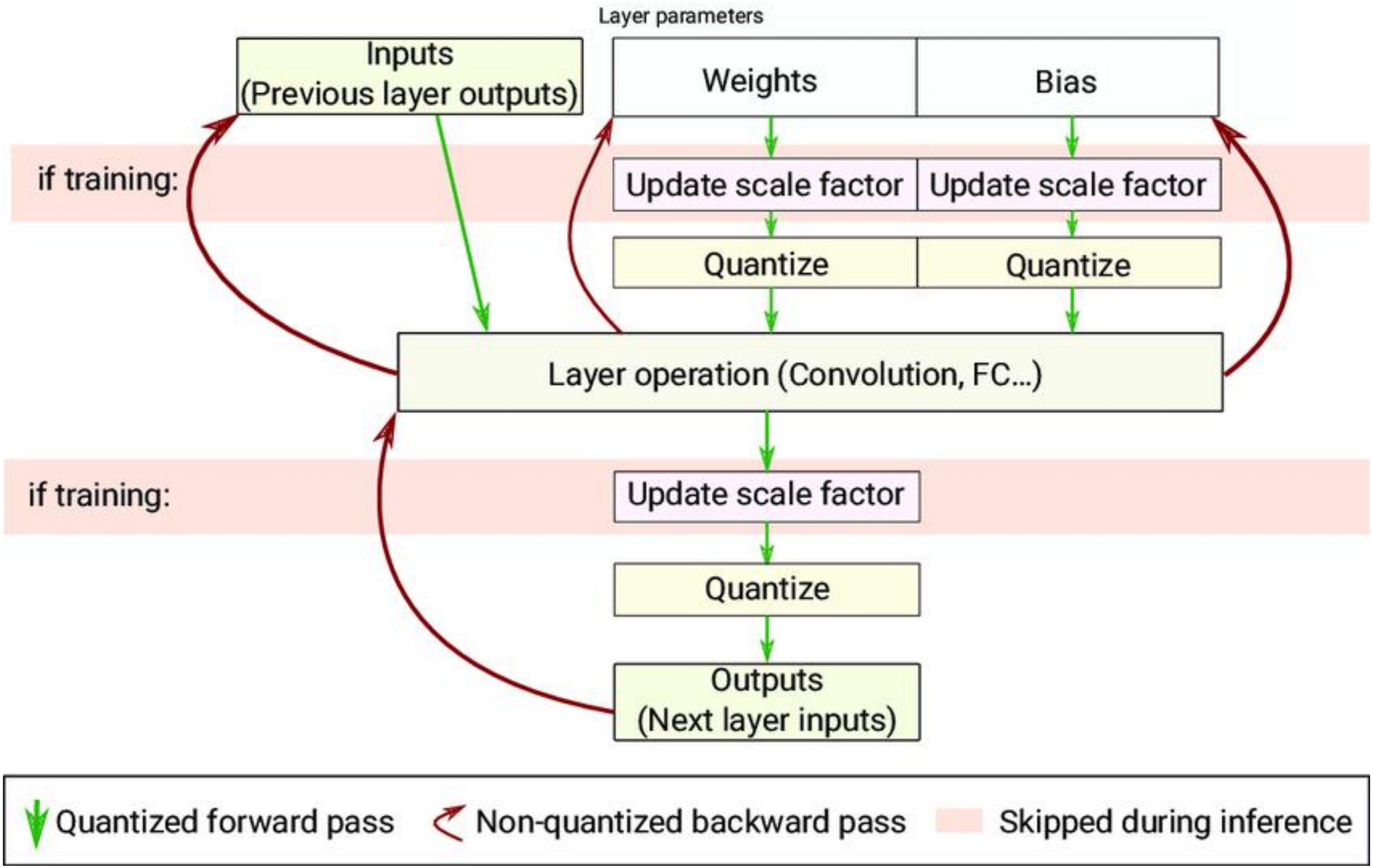
Benefits:

- **Faster arithmetic operations**
- **Reduction in model size**
- **Compatibility with more (and less) devices**

When to apply ?

- **Dynamic Quantization :** quantization of weights only (both fp16 and int8)
- **Static Post training quantization :** quantization of weights/activations (8 bit)
- **Quantization Aware Training**

Quantization Aware Training



PLAN

Introduction

I. Deep Learning: how does it work ?

II. Main challenges of Deep Learning

III. Edge AI in Deep Learning : models compression

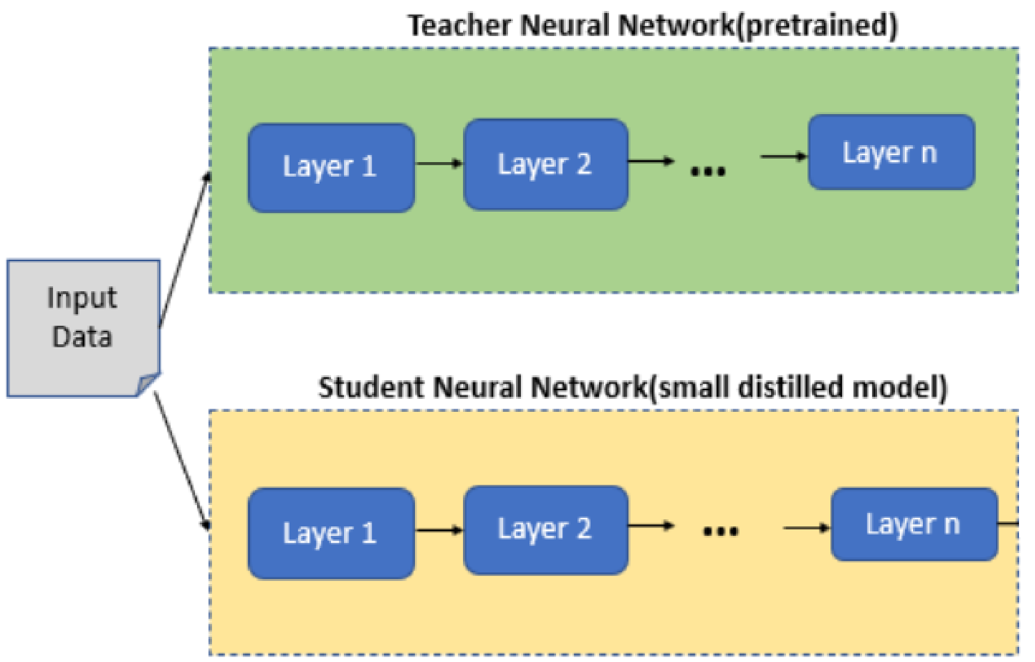
- a. Pruning
- b. Quantization
- c. Knowledge Distillation
- d. Discussion

III. Proposed approach of DNN compression

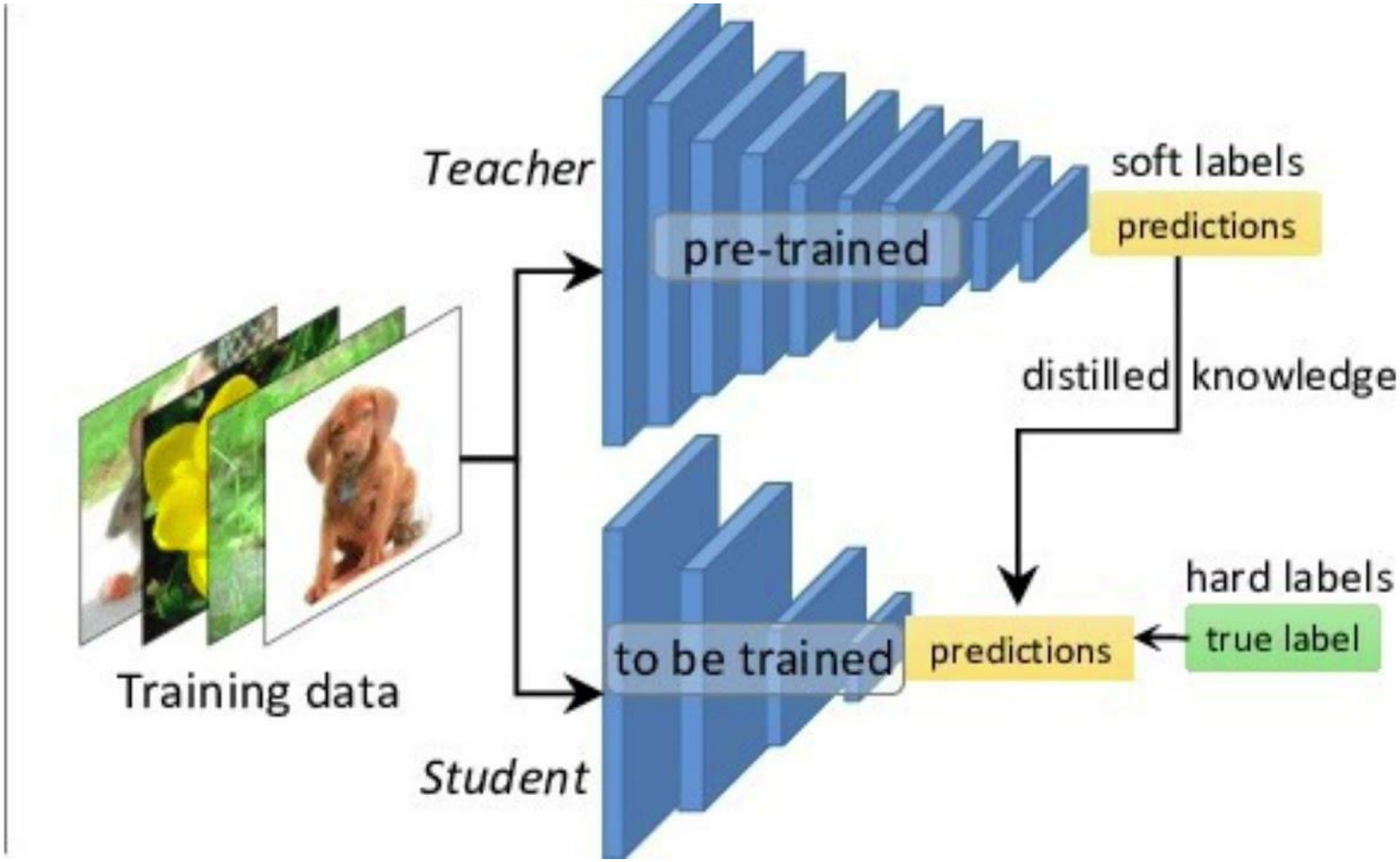
IV. Experimental results : Edge AI use cases

Conclusion

Knowledge Distillation : Process



Knowledge Distillation : Process



PLAN

DNN compression : discussion

Pruning

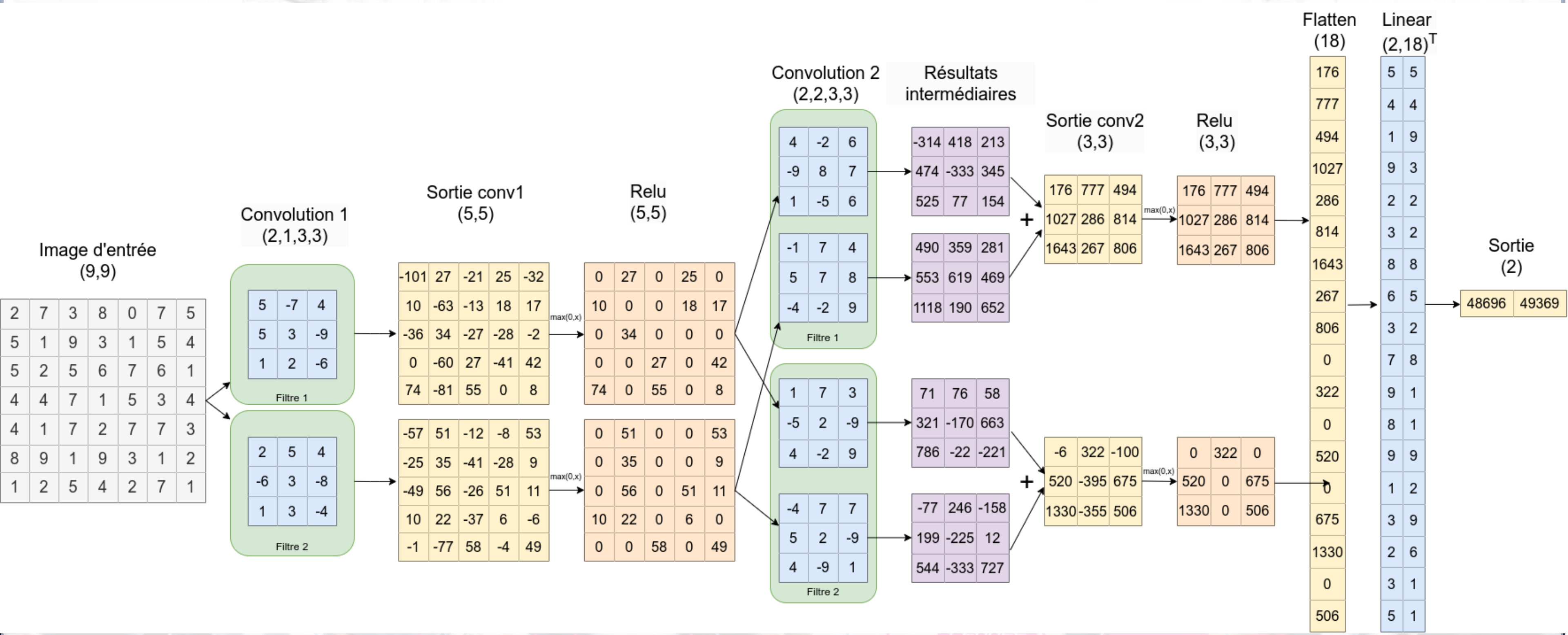
- Major methods generate **sparse** neural networks
- Reduction of Mem size but **no reduction in comp time or RAM consumption**
- Not suitable for Edge AI applications

Block pruning Proposal

- Analyze the dependency of the neural network nodes → blocks
- Calculate the average magnitude of the blocks
- Remove the low magnitude blocks
- **Generate a Dense and pruned** neural network

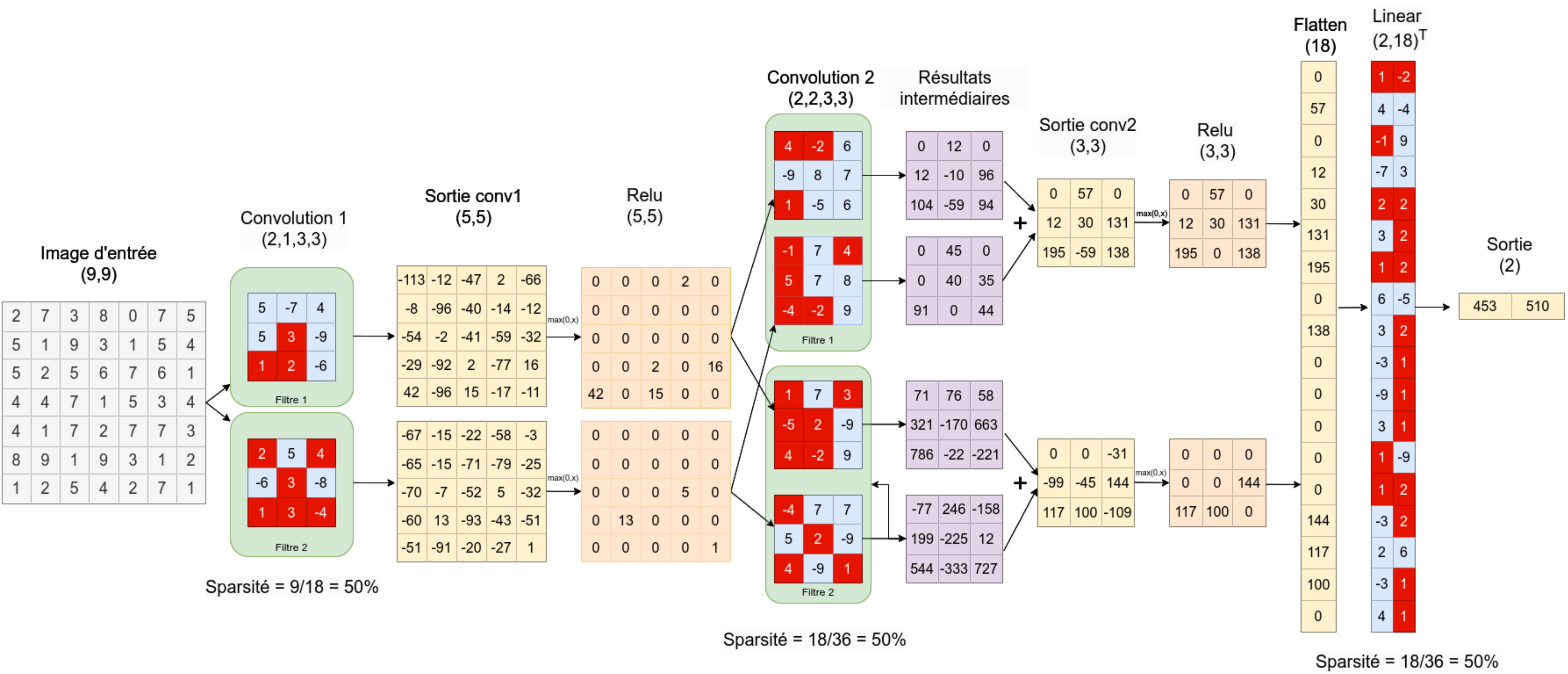
DNN compression : discussion & illustration

Initial CNN (Without Pruning)



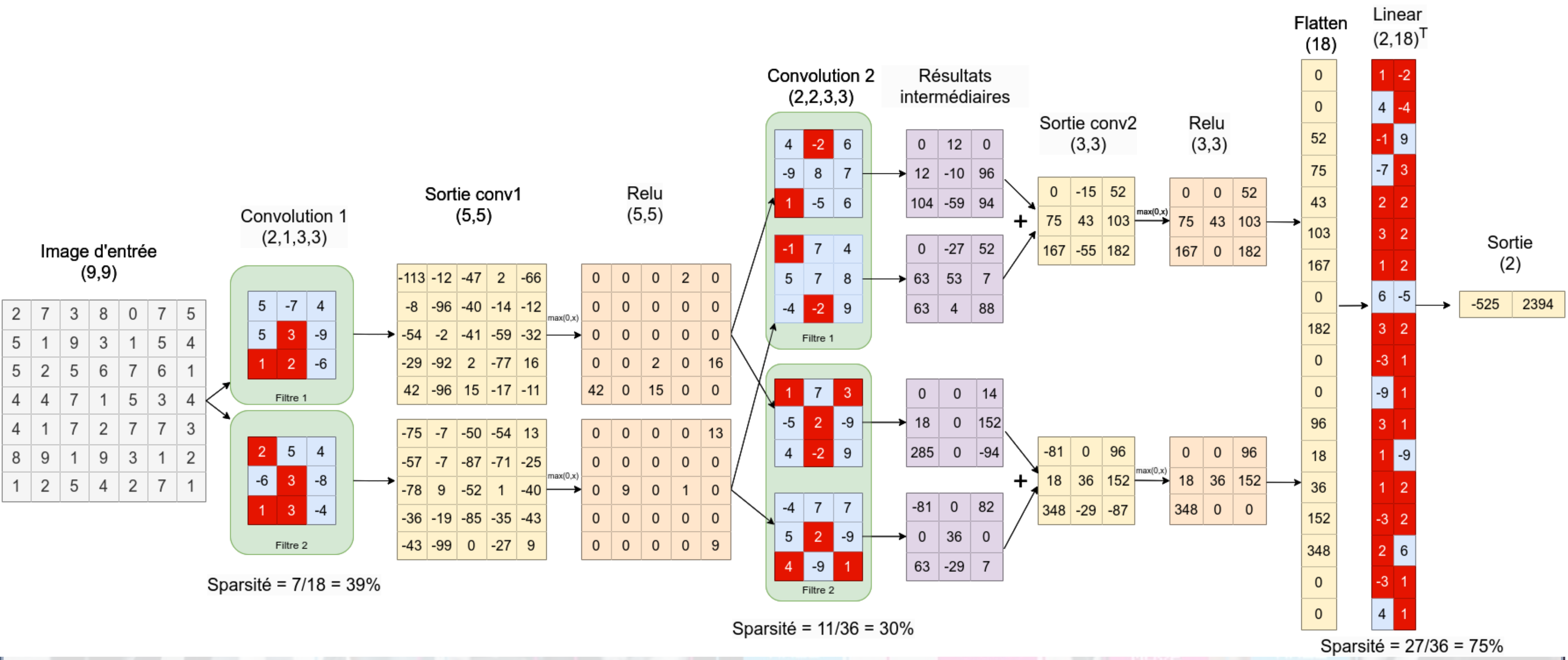
DNN compression : discussion

Unstructured Local Pruning : 50%



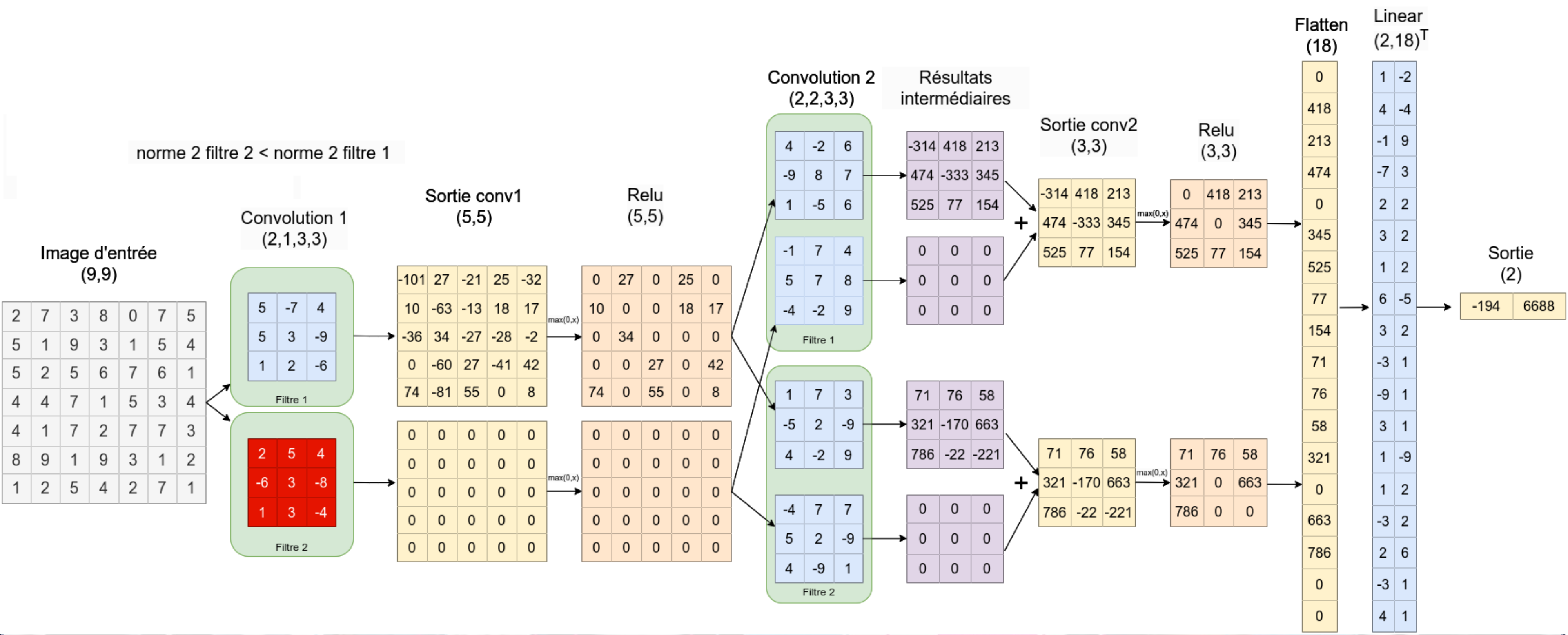
DNN compression : discussion

Unstructured Global Pruning : 50%



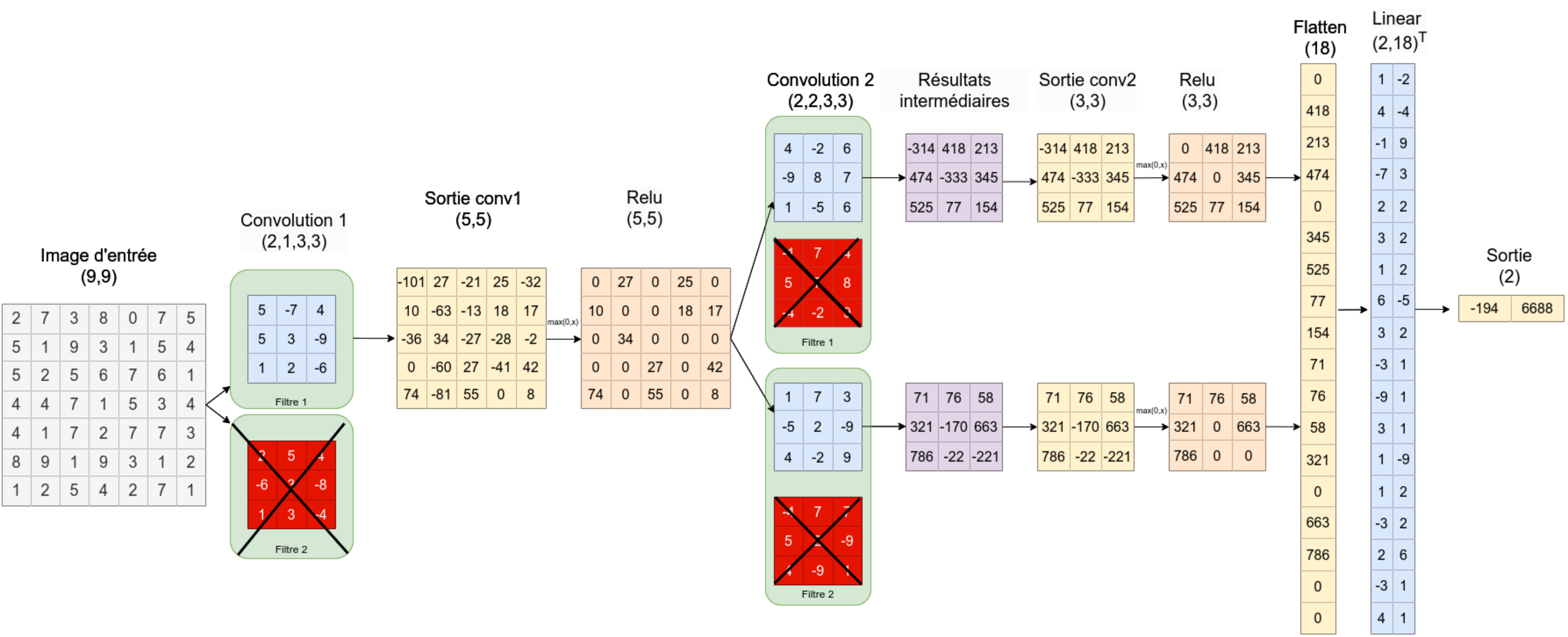
DNN compression : discussion

Structured Filter Pruning



DNN compression : discussion

Structured Block Pruning



PLAN

Introduction

I. Deep Learning: how does it work ?

II. Main challenges of Deep Learning

III. Edge AI in Deep Learning : models compression

- **a.** Pruning
- **b.** Quantization
- **c.** Knowledge Distillation
- **d.** Discussion

III. Proposed approach of DNN compression

IV. Experimental results : Edge AI use cases

Conclusion

Pretrained Model

Input Data

- Structured Block Pruning
- Iterative pruning
- Global Pruning
- Static Quantization
- Quantization Aware training

Train :
Prune & Quantize

- Reduce Model Size
- Accelerate computation
- Reduce RAM consumption

Teacher

No Pruned/quantized Network

Student

Pruned/quantized Network

Train/Distill Knowledge

- Improve accuracy
- Increase the pruning rate

PLAN

Introduction

I. Deep Learning: how does it work ?

II. Main challenges of Deep Learning

III. Edge AI in Deep Learning : models' compression

- **a. Pruning**
- **b. Quantization**
- **c. Knowledge Distillation**
- **d. Discussion**

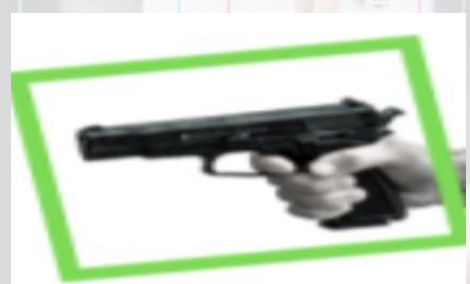
III. Proposed approach of DNN compression

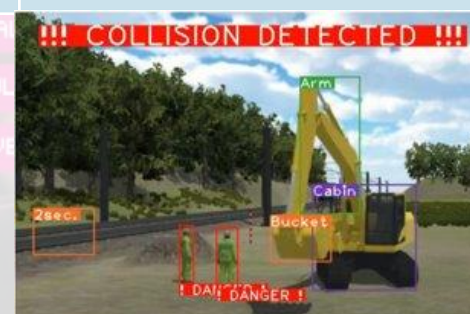
IV. Experimental results : Edge AI use cases

Conclusion

Edge AI use case applications

	Forest Fire Detection	Smart Cities & Security	Danger detection In Railway sites
Sensors	 2D Camera	 2D Camera	 3D Camera Zed2
Edge AI resources	Jetson Xavier 	Jetson Xavier 	 Jetson Xavier & Orin 
Deployed models	**CNN** ResNet, VGG16, etc.	**CNN** Yolo ---> suspect detection Slowfast ---> action recognition	**CNN** Yolo 2D/3D



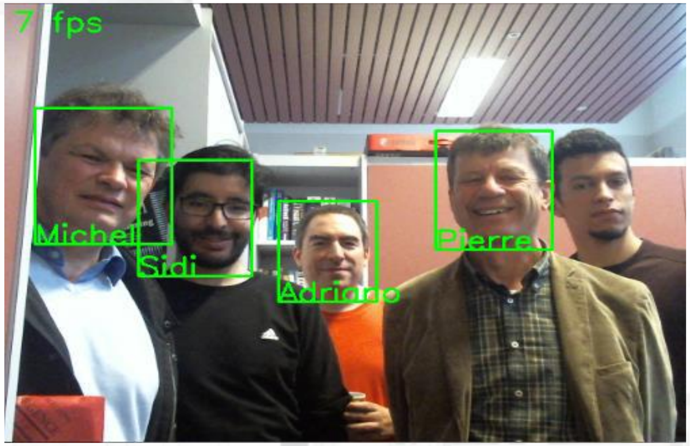


Edge AI use case applications

Edge AI Hardware

Jetson Nano

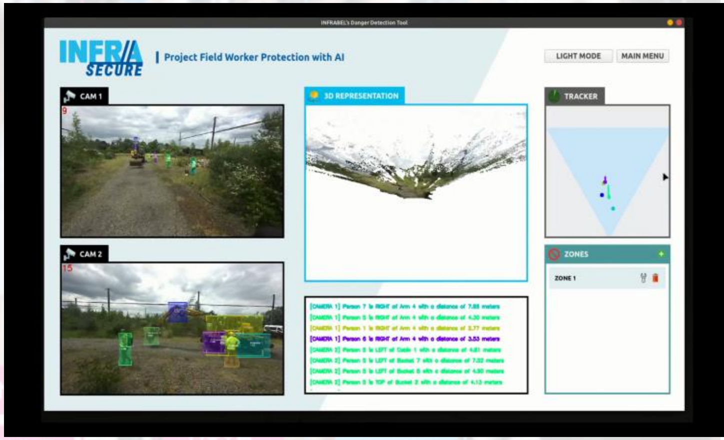
Jetson Xavier



Face recognition

Fire detection

Object detection



Dangerous action detection

Conclusion : Edge AI use case applications

	Deel Learning Model			Compressed Learning Model (Pruning + Quantization + KD)		
	Precison	Model Size	FPS	Precison	Model Size	FPS
Face Recognition	95.50%	91 MB	8	94.13%	14 MB	26
Fire Classification	98.22%	61 MB	10	98.72%	08 MB	24
Object Detection	94.36%	14 MB	25	92.11%	07 MB	42
Actions Recognition	83.01%	50 MB	17	81.72%	14 MB	20

PLAN

Introduction

I. Deep Learning: how does it work ?

II. Main challenges of Deep Learning

III. Edge AI in Deep Learning : models compression

a. Pruning
b. Quantization
c. Knowledge Distillation
d. Discussion

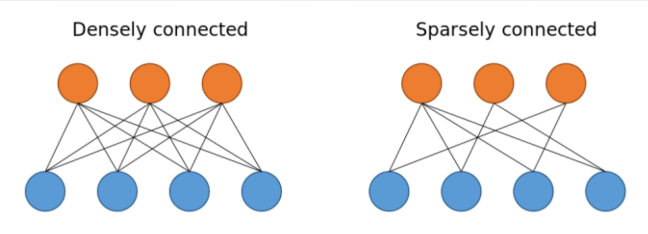
III. Proposed approach of DNN compression

IV. Experimental results : Edge AI use cases

Conclusion

Conclusion

- **Edge AI :** several use case applications in video surveillance & Smart Cities
- **Edge AI :** Hardware reflexion if terms of collection and processing
- **Magnitude pruning:** good only for reducing the model size
- **Movement pruning :** recommended when using transfer learning
- **Block Pruning :** recommended for accelerating computation and reducing RAM consumption
- **Combine** pruning and quantization during the training process
- **Knowledge distillation:** improve the accuracy of pruned networks.



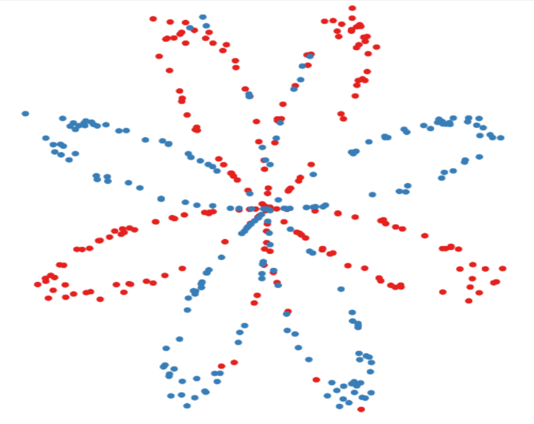
Conclusion

If you need to learn more about AI and Edge AI, welcome to Hands on AI and Hackia:

- Certificate Hands on AI at UMONS : [Link]
- Workshop HackIA of the certificate : https://hackia.eu/

Thank you

Deep Learning : how does it work ?



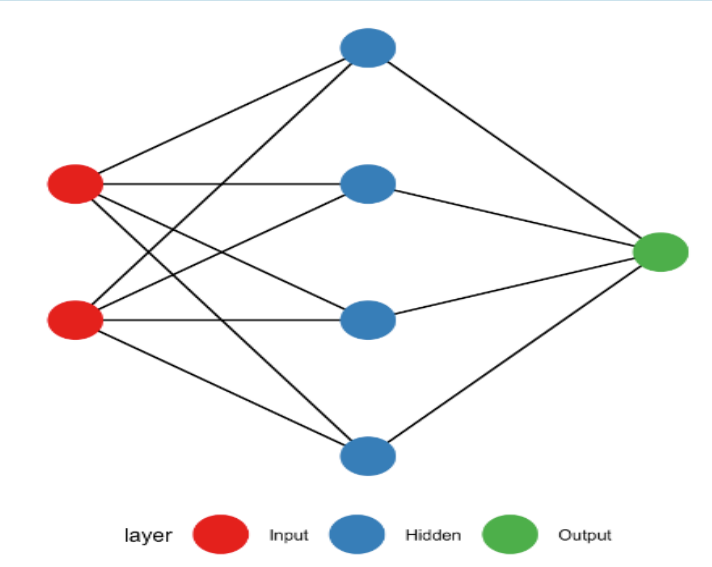
Problem : Predict the color of each point

- **Input :** X, Y coordinates (2 values)
- **Output :** the color (1 value : 0 for blue and 1 for red)

Trivial problem : **Linear algorithms will fail** since colors are **not linearly separable**

No single line that can perfectly separates red dots from blue dots.

Neural network : with one hidden layer



Deep Learning : how does it work ?