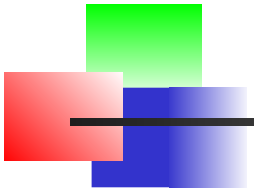


IA Embarquée

P. Bakowski

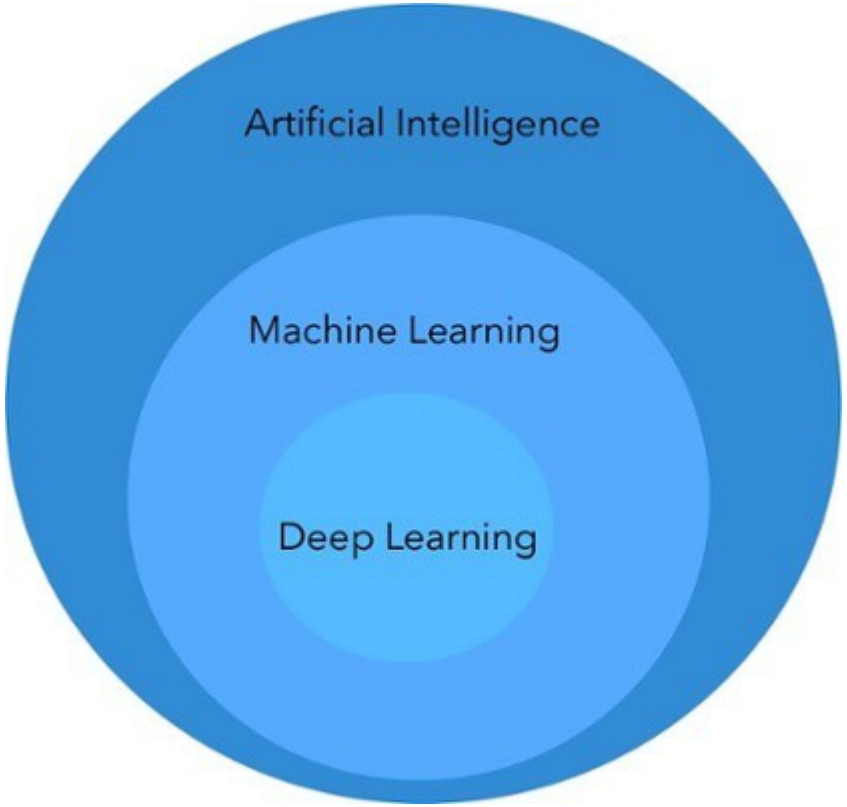
Intelligence Artificielle (IA) et IA Embarquée

- Introduction
- Matériel pour l'IA Embarquée
- Logiciel pour l'IA Embarquée
- Laboratoire 0 – Introduction au Réseaux Neuronaux
- Laboratoire 1 – Régression Linéaire et Polynomiale
- Laboratoire 2 – Réseaux Denses et Convolutifs (MNIST)
- Laboratoire 3 – Réseaux Auto-encodeurs
- Laboratoire 4 – Inférence avec des modèles entraînés
- Laboratoire 5 – Développement d'une Application

Introduction – IA, ML et DL

- Le **Machine Learning** est une technologie d'intelligence artificielle permettant aux ordinateurs d'apprendre **sans avoir été programmés** explicitement à cet effet.

- Pour apprendre et se développer, les ordinateurs ont toutefois **besoin de données** à analyser et sur lesquelles ils doivent **s'entraîner**.



Introduction – IA, ML et DL

- L'**apprentissage profond** est un ensemble de méthodes d'apprentissage automatique tentant de modéliser avec un haut niveau d'abstraction des données grâce à des architectures articulées de différentes **transformations non linéaires**.

- Ces techniques ont permis des progrès importants et rapides dans les domaines de l'analyse du **signal sonore ou visuel** et notamment de la **reconnaissance faciale**, de la **reconnaissance vocale**, de la **vision par ordinateur**, du traitement automatisé du langage.

Introduction – IA Embarquée

- Grâce à notre capacité à construire des machines intelligentes qui simulent l'intelligence humaine, les implications pour le progrès technologique dans de nombreux secteurs sont infinies. Alors, **quoi de mieux que l'intelligence artificielle** ?

- **Intelligence Artificielle Embarquée.**

- L'intelligence artificielle (IA) embarquée est l'application de **ML** et **DL** au **niveau de l'appareil** (*embedded device*)

- L'Appareil ou *embedded device* est un **SoC** ou un **FPGA** intégré sur une carte autonome – **SCB** (S*ingle Computer Board*)

- Exemples : Nvidia -Jetson Nano, Xavier, Google - CORAL, ..

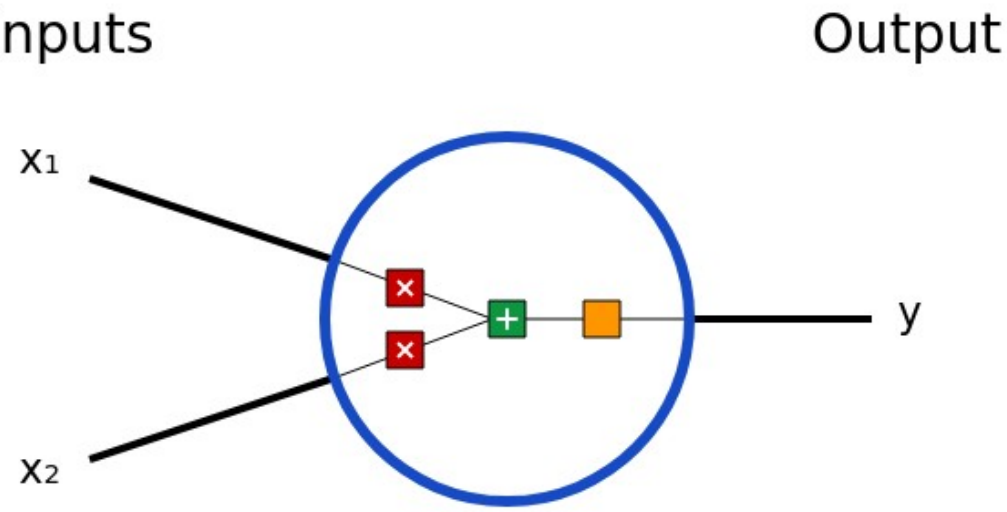
Introduction – Logiciel

- Les logiciels fonctionnant sur le équipements IA embarquées sont (presque) les mêmes que ceux sur les équipements standards (PC, serveurs,..)
- **Langages :** C/C++ et Python
- **Packages :** PyNum, Ski-Learn, ..
- **Couches applicatives :** Keras, TensorFlow, PyTorch,..

  TensorFlow => TensorFlow Lite
- **Couches d'exécution** : CUDA, cuDNN, ..
- **OS** : Linux, FreeRTOS, ..

Introduction – Un Neurone

- Un **Neurone** est un mécanisme essentiel dans le développement des applications d'IA. Voici sa structure fonctionnelle :

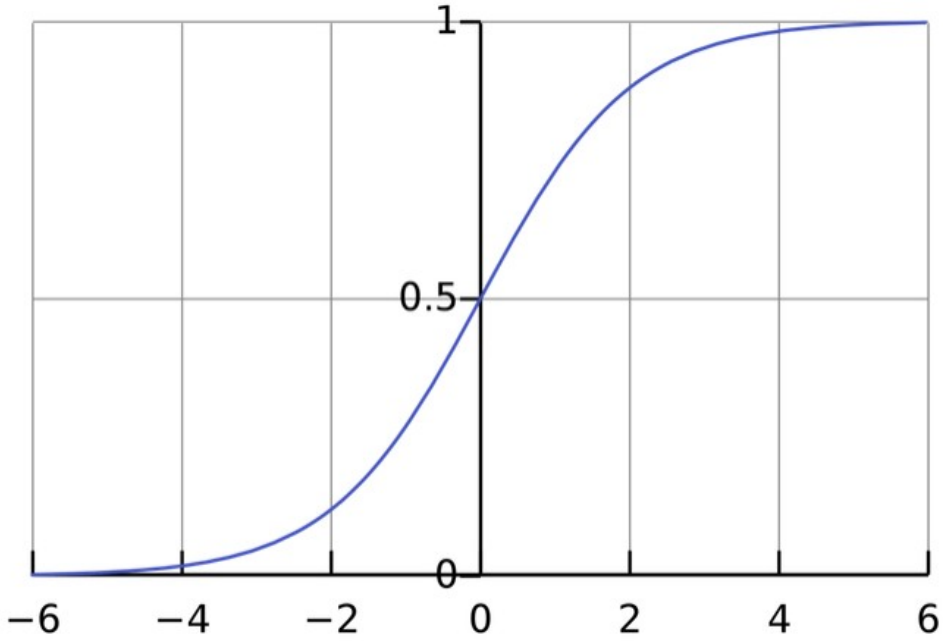


- **y = f(x1*w1+x2*w2+b)**
- où : **w1** et **w2** sont des **poids** (*weights*) et **b** est un **biais**

Introduction – Un Neurone

y = f (x1*w1+x2*w2+b)

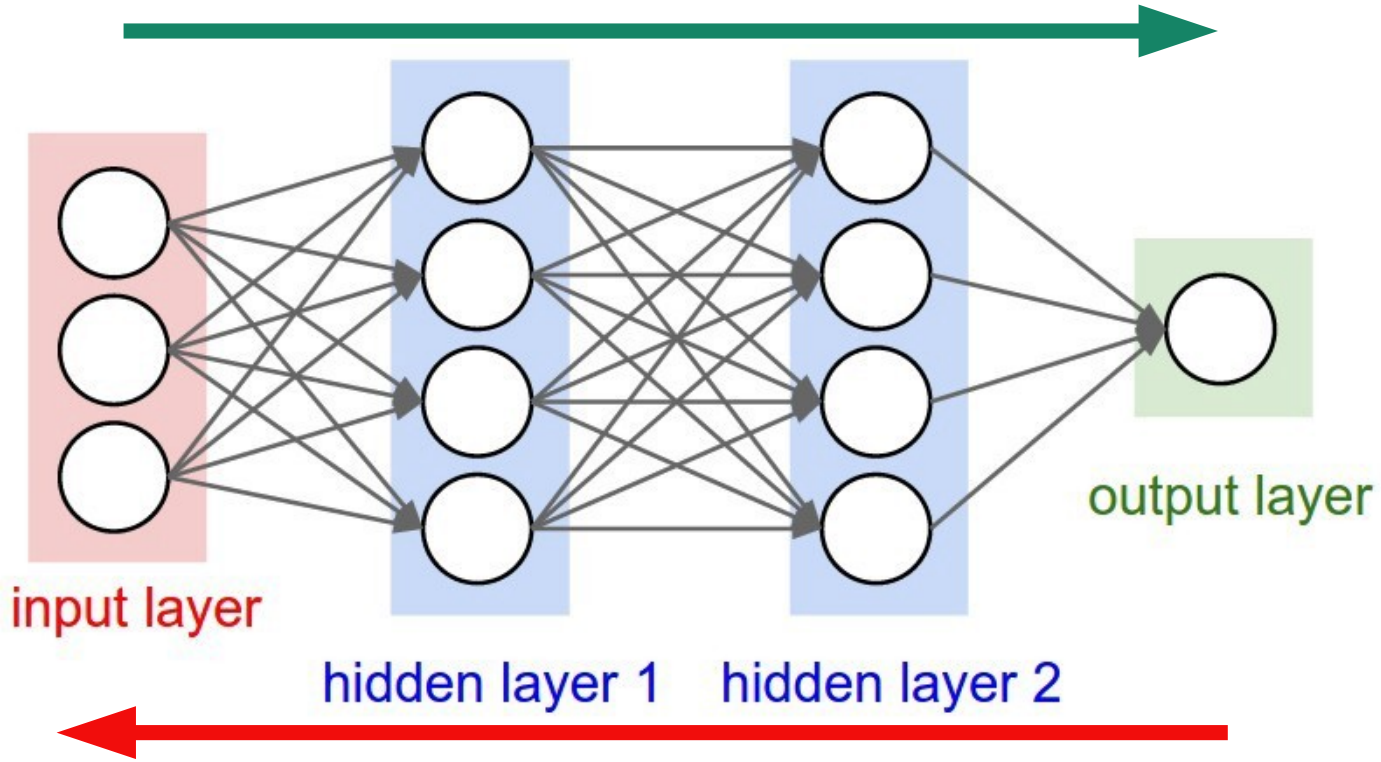
f - fonction sigmoïd représente la fonction de répartition de la loi logistique. Elle est **dérivable**, ce qui est une contrainte pour l'algorithme de **rétropropagation**



Exercice : écrivez l'équation de la fonction sigmoïd !

Introduction – Un Neurone

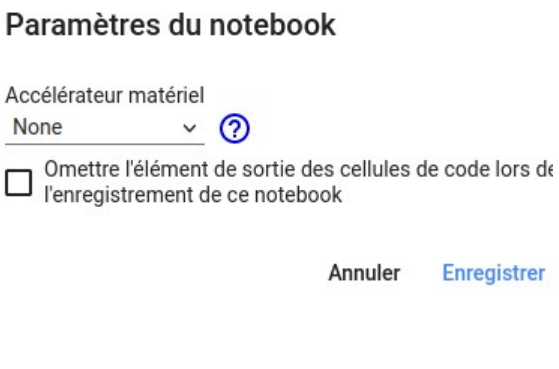
- **propagation** (*for-forwarding*)
- **rétropropagation** (*back-forwarding*)

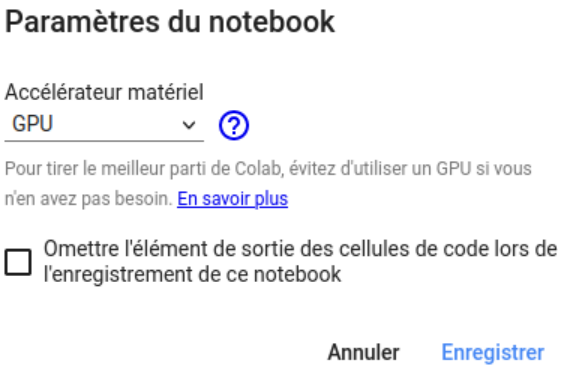


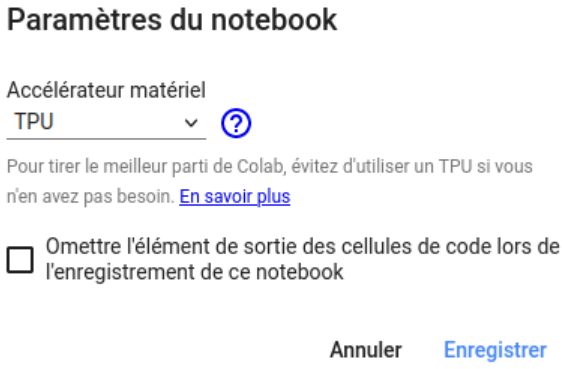
Matériel pour IA embarquée

- Les **SoC** développés pour les applications de l'IA embarquée sont presque toujours basées sur les **CPU ARM**
- Ils intègrent les **GPUs** et/ou les **TPUs**

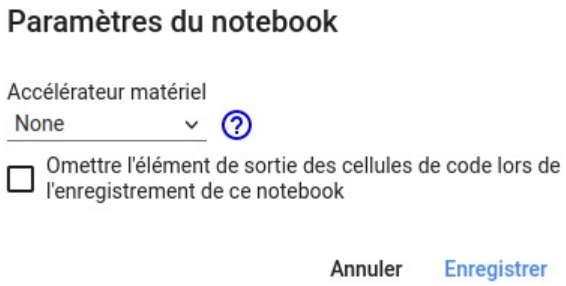
- Questions
- Quelles sont les différences entre les **CPUs et les GPUs** ?
- Quelles sont les différences entre les **GPUs et les TPUs** ?





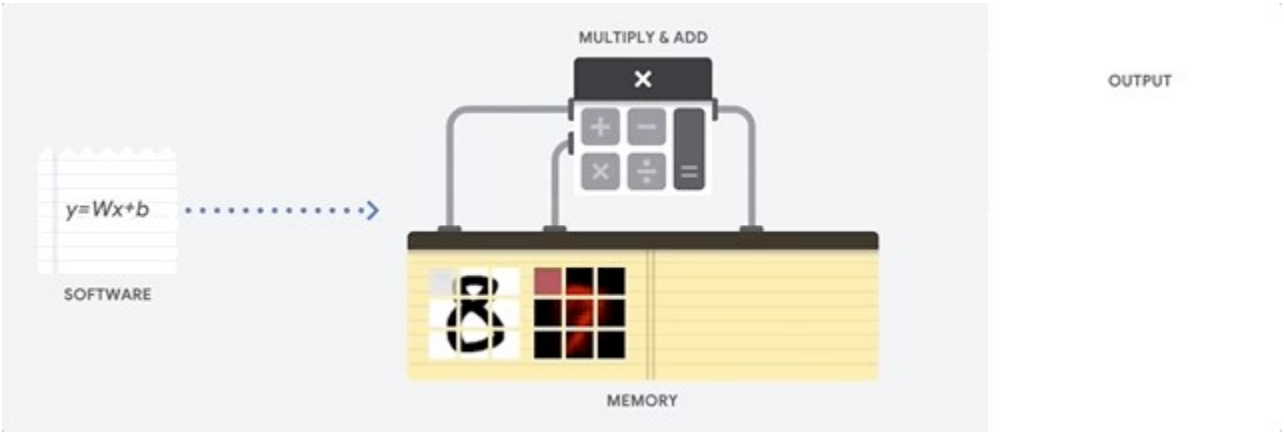


Matériel pour IA embarquée : CPUs

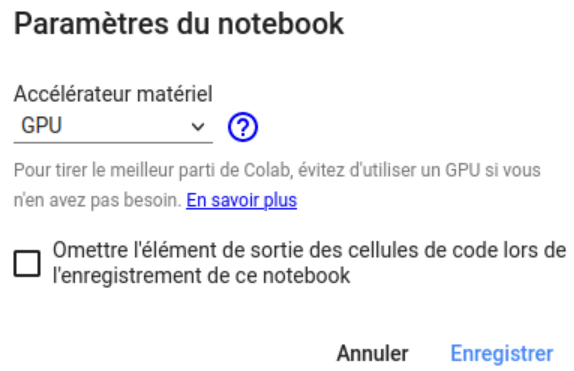


- Modern **CPU Features Summary:**
  - Has Several Cores
  - **Specialized in Serial Processing**
  - Capable of executing a **handful** of operations at once
  - Have high speed FLOPS utilization
  - Supports large models thanks to its large memory capacity
  - Very flexible and programmable for **irregular computations**

Matériel pour IA embarquée : CPUs

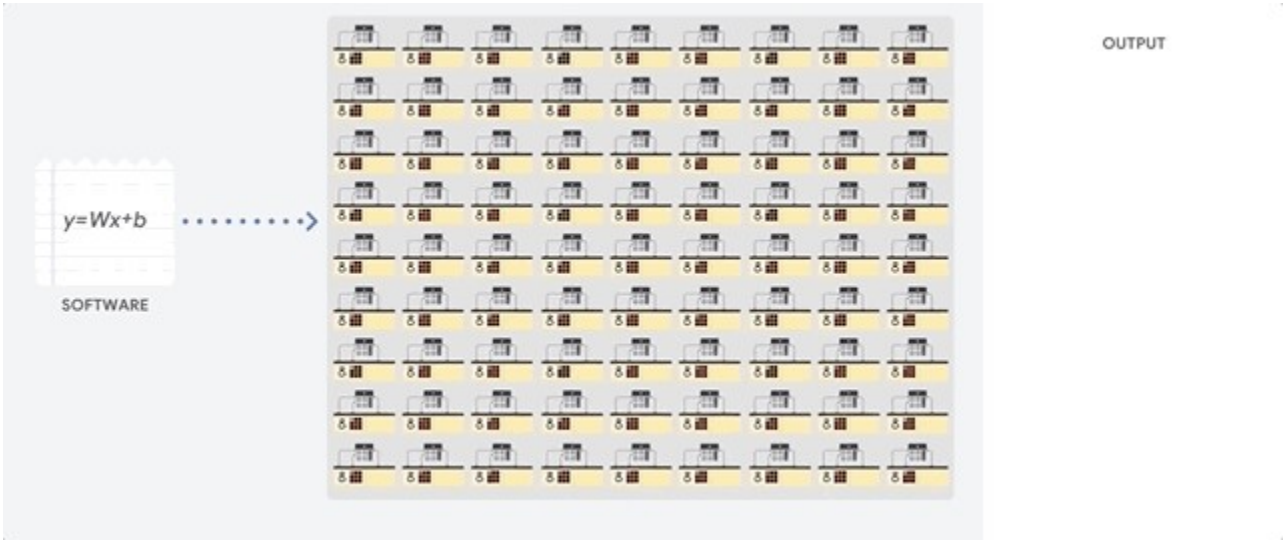


Matériel pour IA embarquée : GPUs

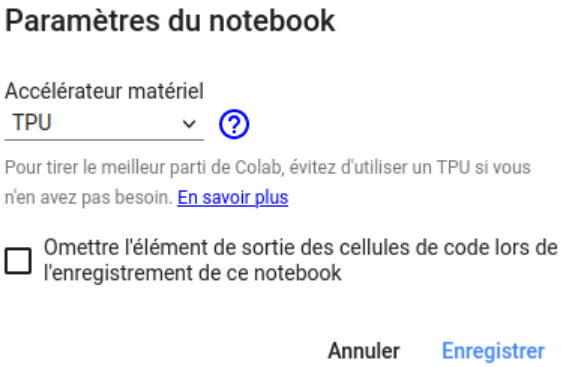


- **GPU** Features Summary:
  - Has **hundreds/thousands of cores**
  - High memory throughput
  - Specialized for parallel processing
  - Capable of executing **thousands of operations** at once
  - Taking into account integer and floating point operations

Matériel pour IA embarquée : GPUs

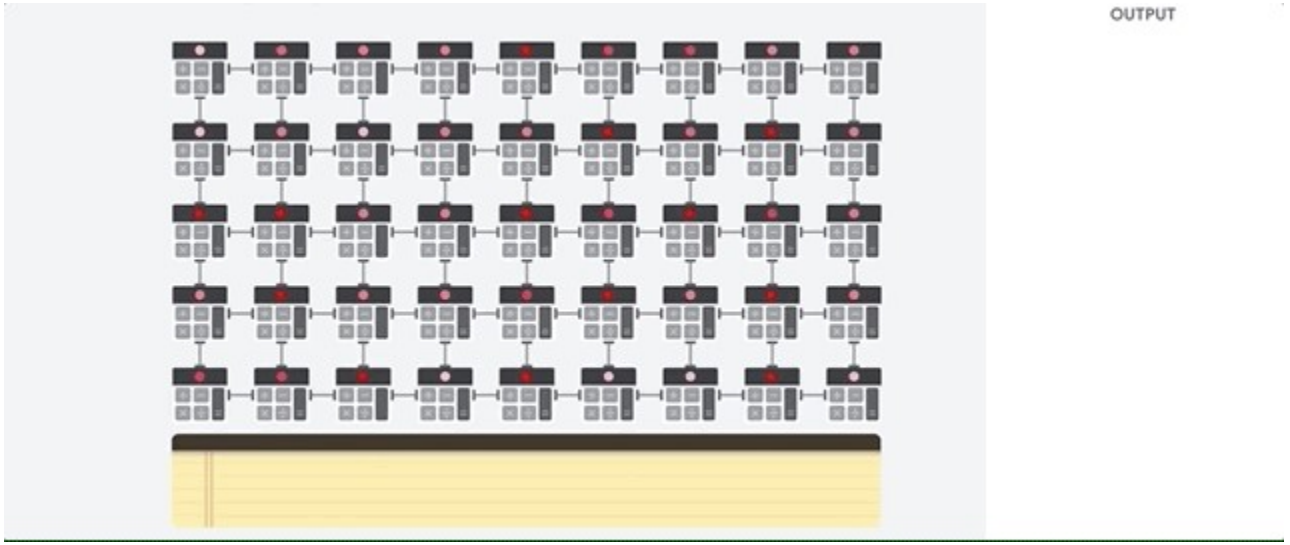


Matériel pour IA embarquée : TPUs



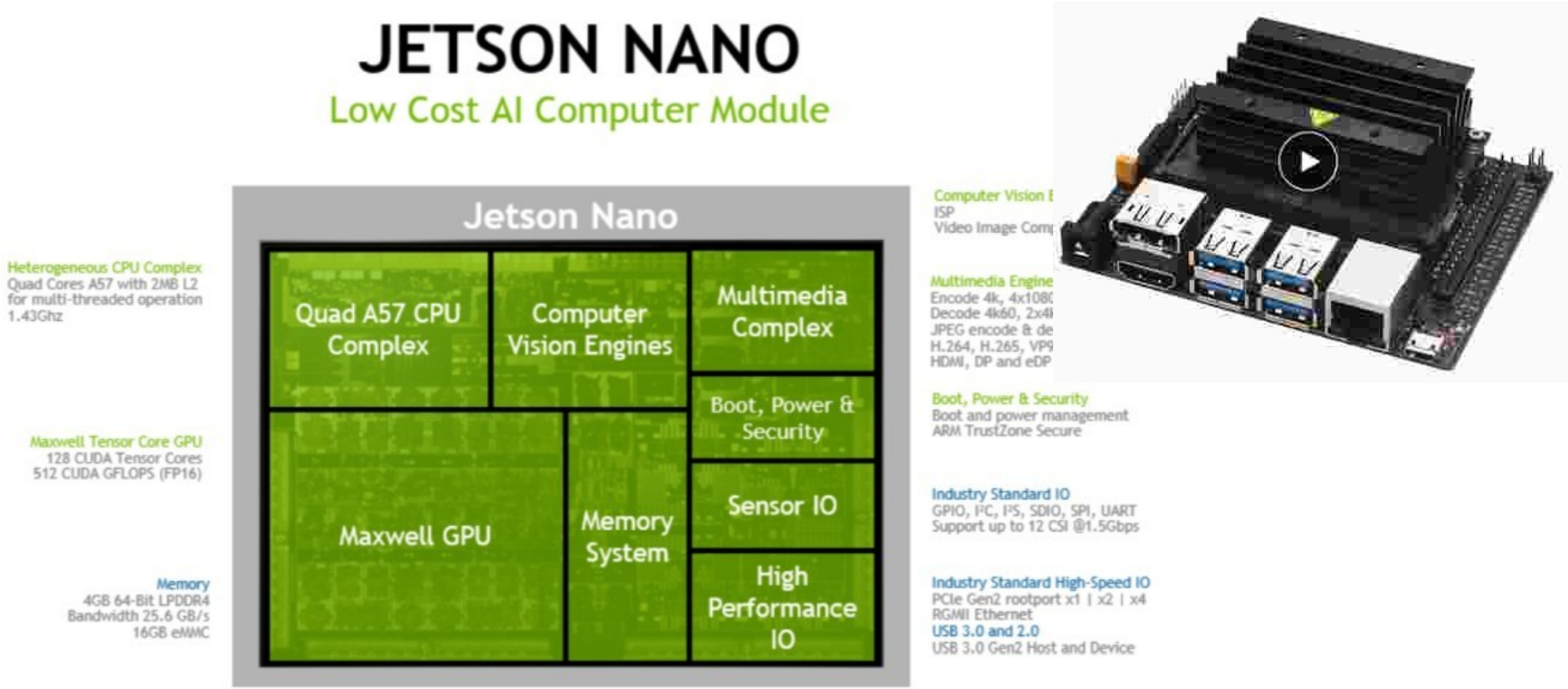
- **TPUs** Features Summary:
- Special Hardware for **Matrix Processing**
- High Latency (compared to CPU) ; long start
- Very High Throughput : very large buses
- Compute with Extreme Parallelism on smaller units (FP-16 bit)
- Highly-optimized for large batches and CNNs (convolutional neural network)

Matériel pour IA embarquée : TPUs



Dans un TPU, les ALUs sont **directement connectés** les uns aux autres sans utiliser la mémoire. Ils peuvent donner directement des informations de passe qui réduiront considérablement la latence.

Matériel : Nvidia Jetson Nano

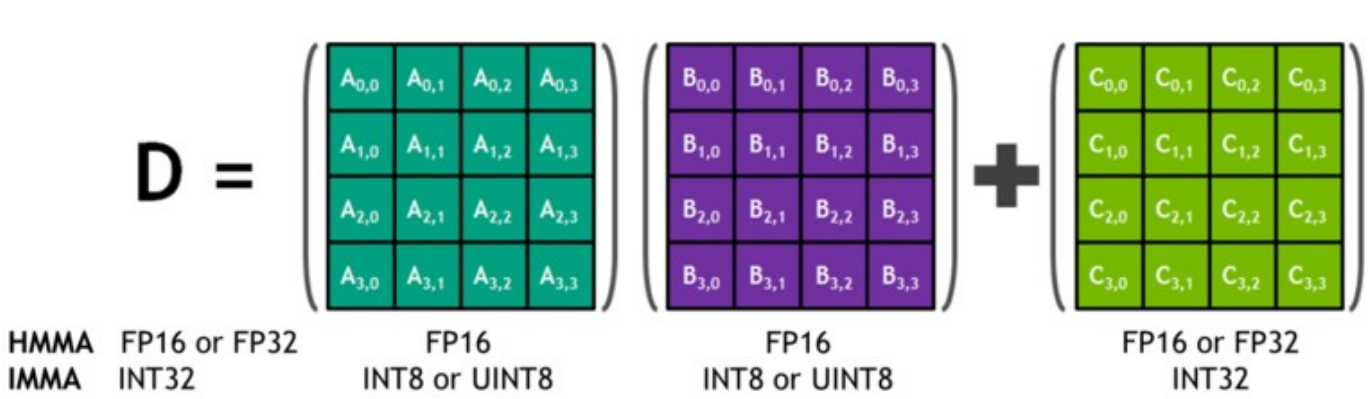


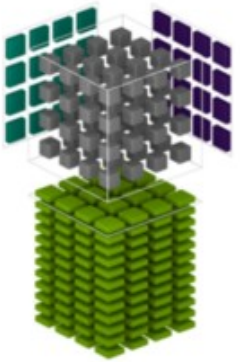
GPU Maxwell à 128 cores à 2 opérateurs – 32 bits (921 MHz)
performance = 0,5 TOPS (INT8)

Matériel : Nvidia Jetson Xavier

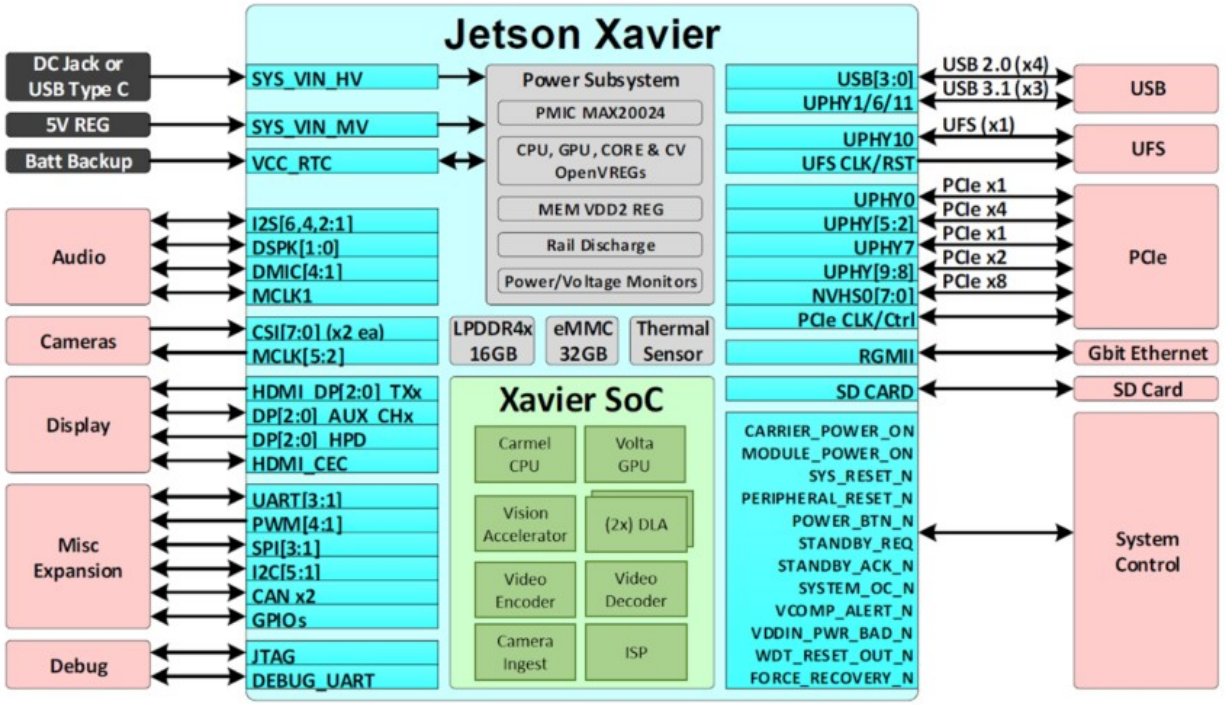
Jetson Xavier – GPU Volta à 512 cores à 4 opérateurs 64 bits (1377 MHz) – TPU avec 64 Tensor cores

performance = 22 TOPS (INT8)





Matériel : Nvidia Xavier SoC



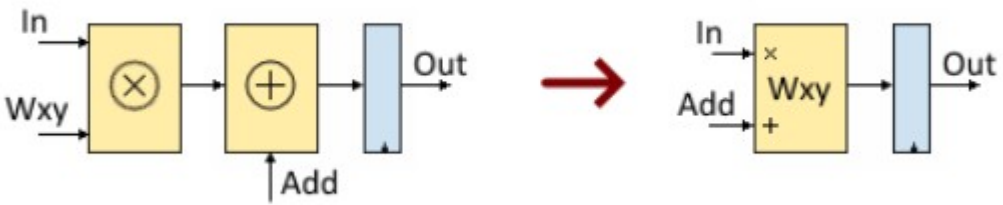
Attention : Jetson Xavier est un **SBC**, qui intègre un **SoC** Xavier de **11 mld** transistors (**7 nm**) .
Le coût de développement **2 mld de dollars**.

Matériel : Google Coral

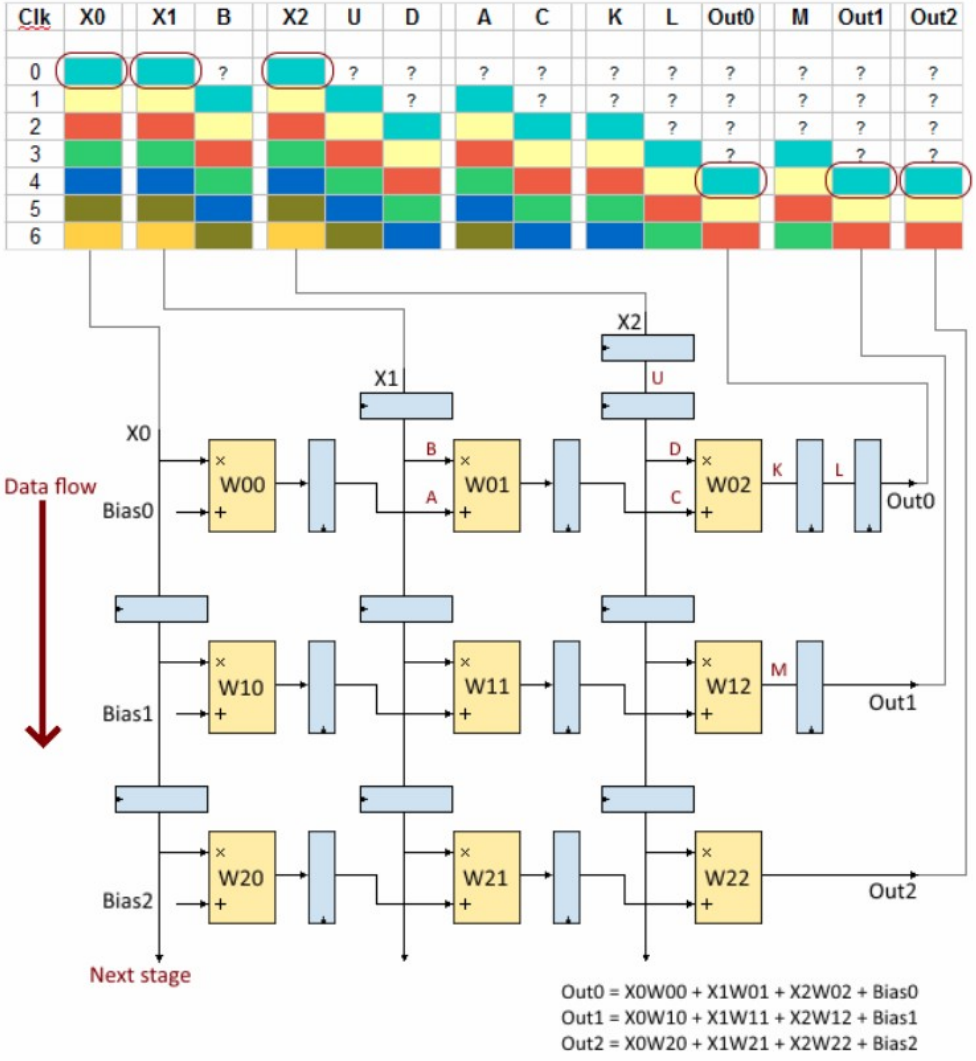
Le **SoC Google Coral** utilise un ASIC fabriqué par l'équipe Google appelé Edge TPU.

Il utilise une **matrice systolique** pour effectuer les opérations type **tenseur**.

Les éléments de la matrice sont les cellules **mul-add**.

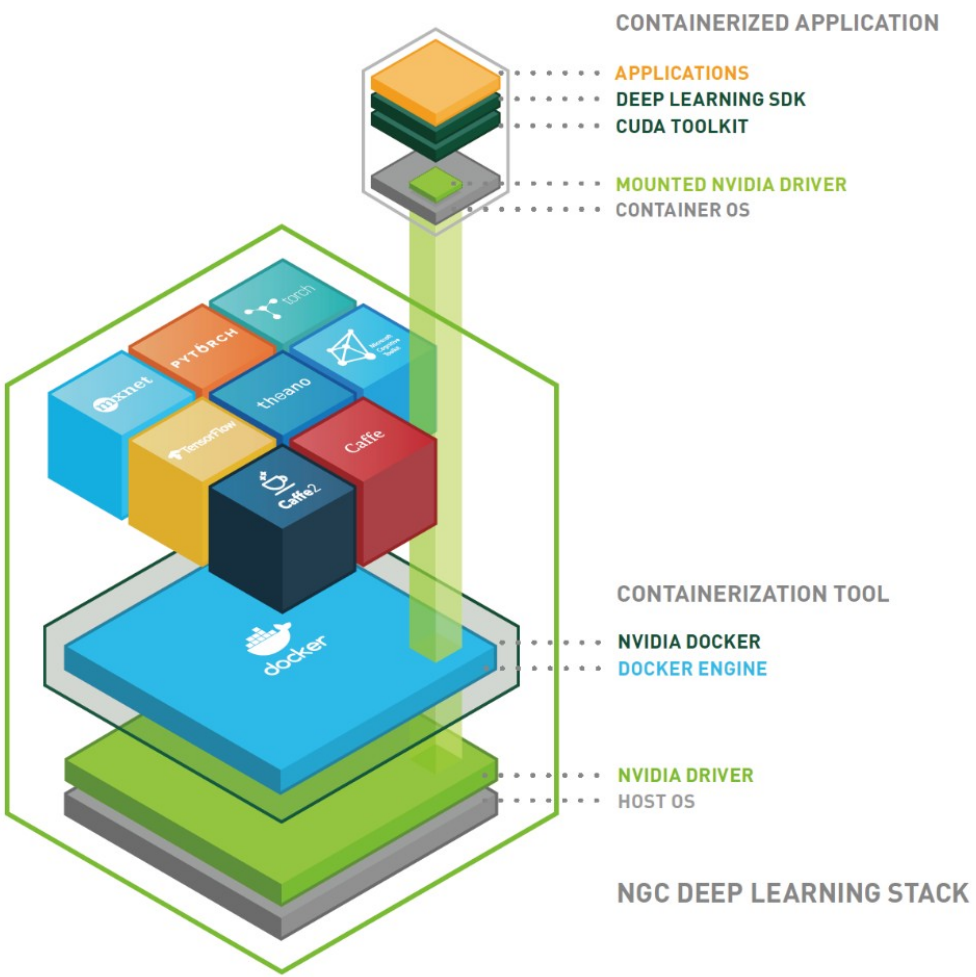


Matériel : Google Coral

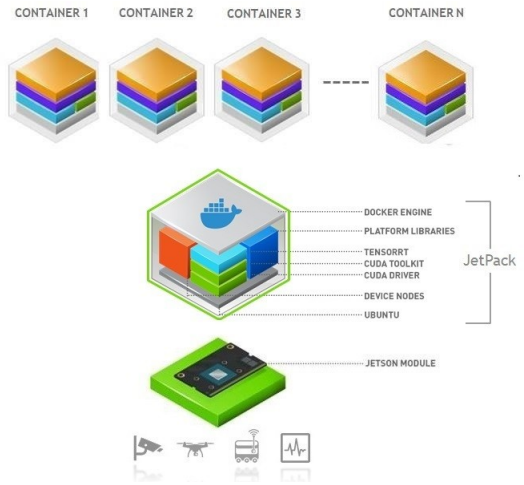


Logiciel : JetPack

Logiciel : ML Docker



Logiciel : ML Docker

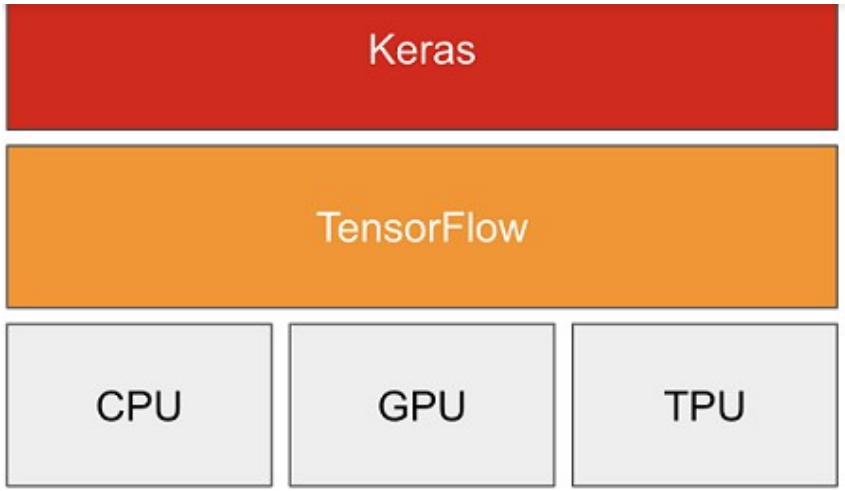


Machine Learning container contient : **TensorFlow**, **PyTorch**, **JupyterLab** et d'autres frameworks de ML et de science des données populaires tels que **scikit-learn**, **scipy** et **Pandas** préinstallés dans un environnement **Python 3.6.**

Jetson Inference container contient :

TensorRT et l'apprentissage par transfert avec **PyTorch et l**es application de **classification d'images** pour **C++** ou **Python**, la **détection d'objets** et des **démonstrations de caméra en direct**

Logiciel : TensorFlow et Keras



Deep Learning …
Layers, models, optimizers, losses, metrix, ..

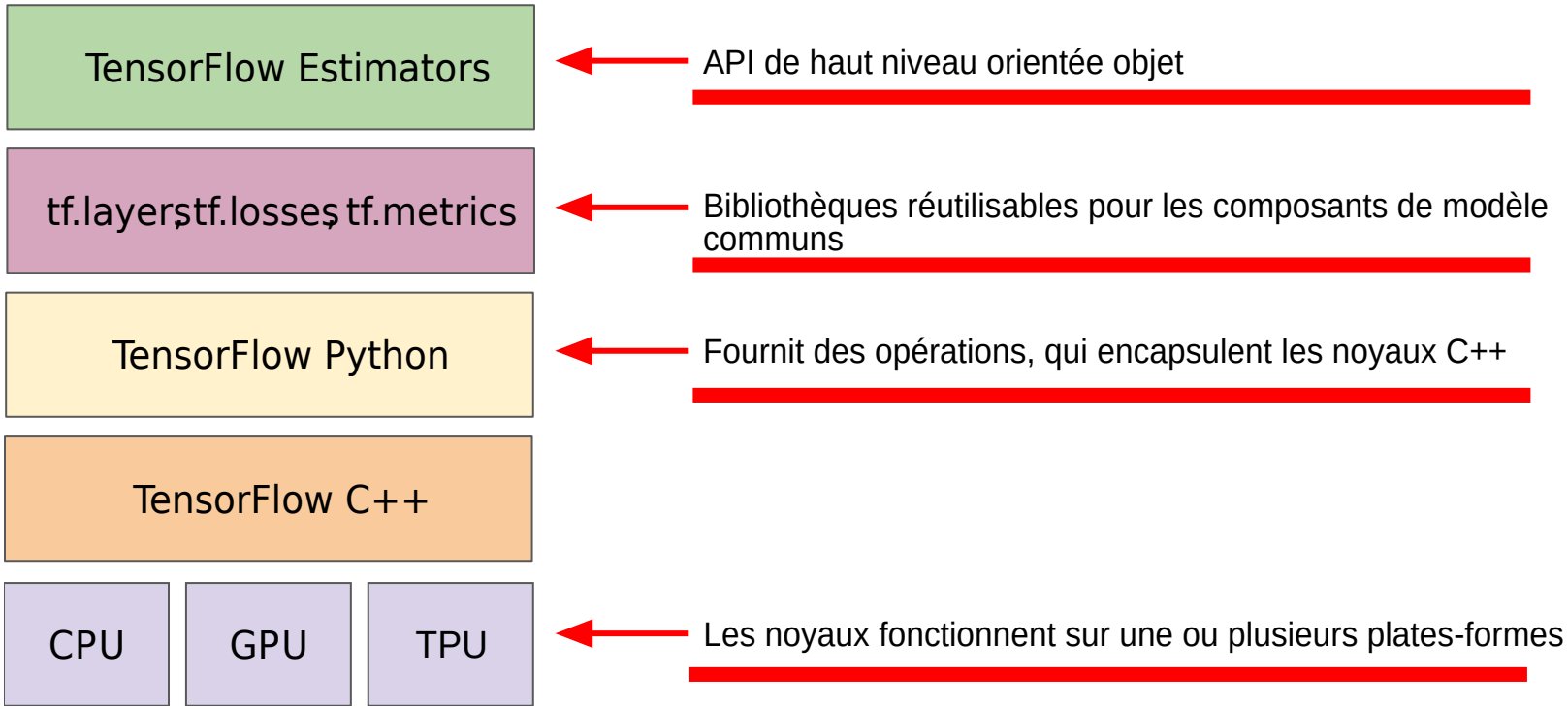
Tensor Manipulation ...
Tensors, variables, differentiations, distribution, ..

Coding …

```
Import numpy as np
import tensorflow as tf
from tensorflow import keras
from tensorflow.keras import layers
```

TensorFlow

Pour la développement nous allons utiliser les bibliothèques de **Python** les plus avancées pour le **ML** (Machine Learning) et **DL** (Deep Learning) , telles que **`TensorFlow`** et **`Keras`**.



Keras

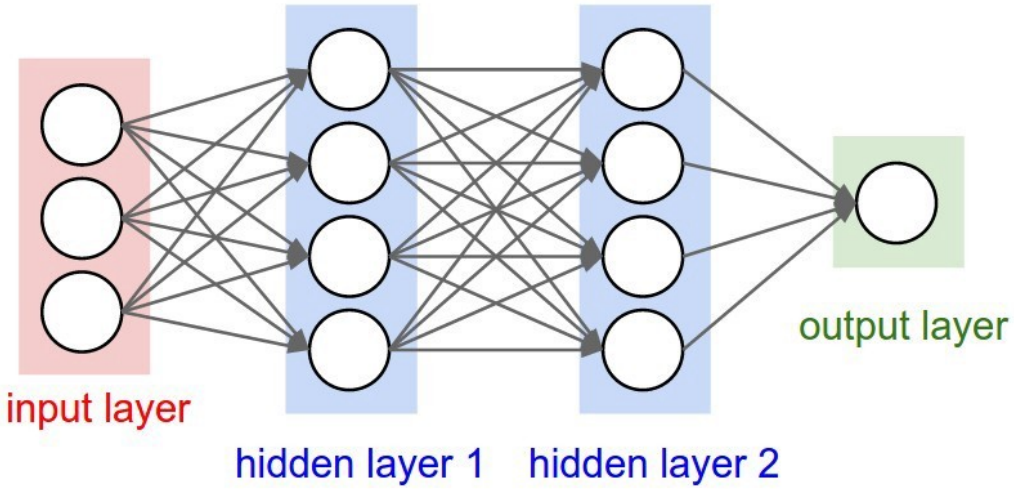
Keras fournit les composants pour créer les couches (layers) d'un réseau de neurones.

Couche d'entrée est conceptuellement différente des autres couches

Couche cachée 1: 4 unités (**4 neurones**)

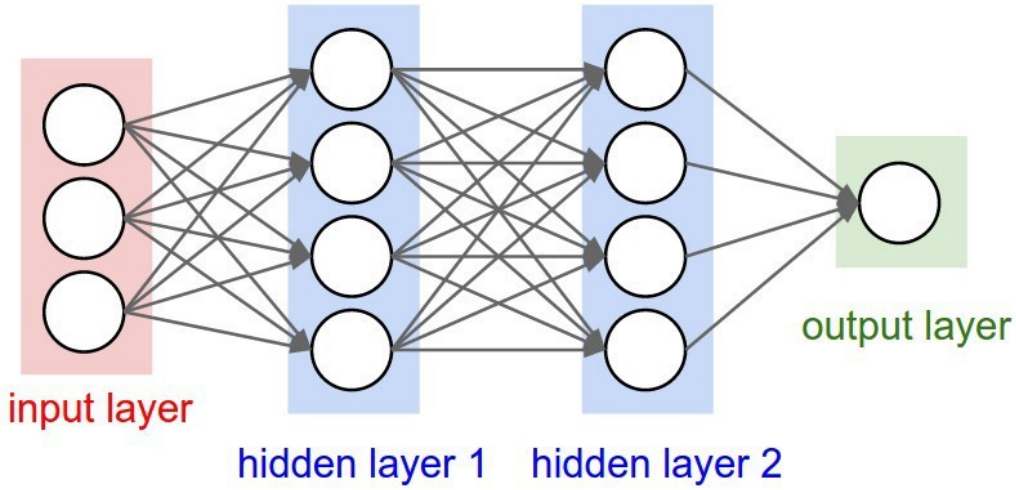
Couche cachée 2: 4 unités

Couche de sortie: 1 unité

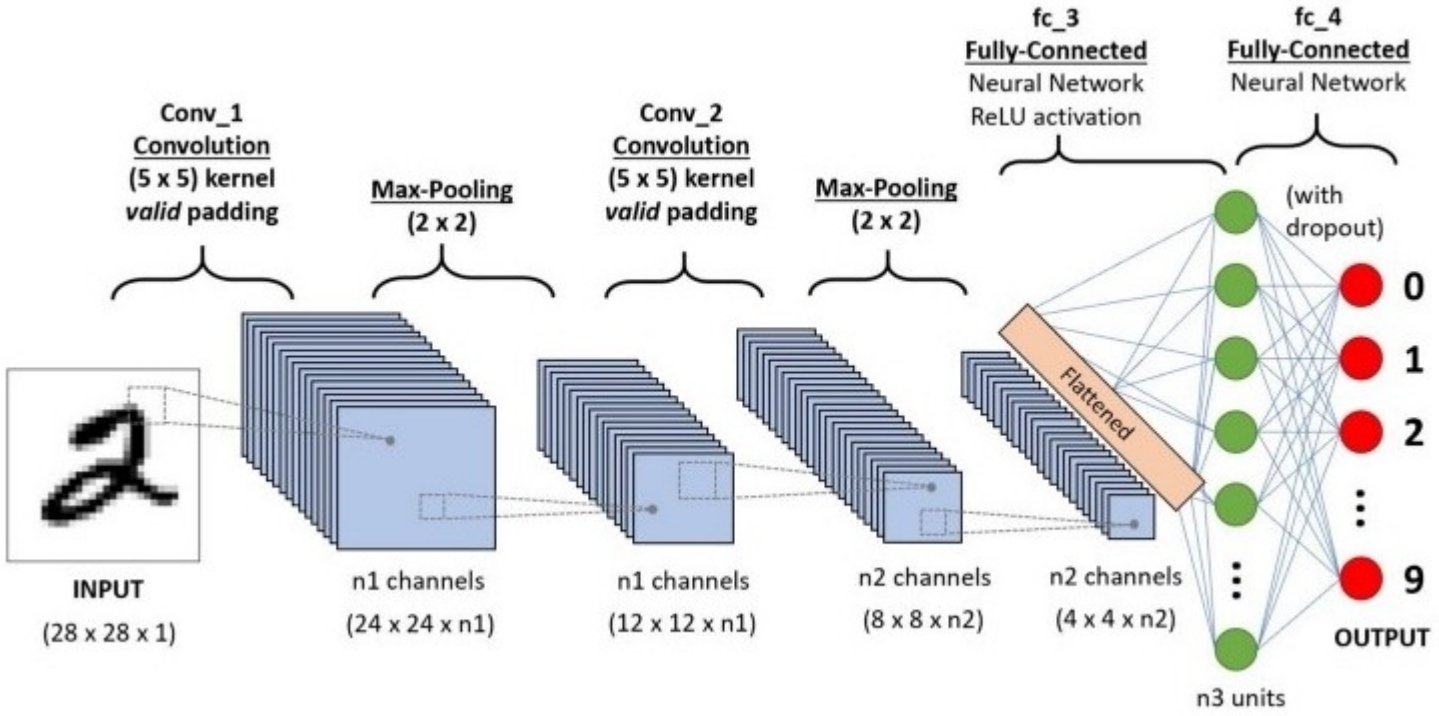


Keras : un simple modèle Dense

```
from keras.models import Sequential
from keras.layers import *
model = Sequential()
# input with shape=(3,)
model.add(Dense(units=4,input_shape=(3,))) #hidden layer 1
model.add(Dense(units=4)) #hidden layer 2
model.add(Dense(units=1)) #output layer
```



Un exemple de réseau neuronal



Un réseau de neurones convolutifs (**ConvNet/CNN**) qui peut prendre une image d'entrée, attribuer de l'importance (poids et biais apprenables) à divers aspects / objets de l'image et **être capable de différencier l'un de l'autre**.

Un résumé

Dans cette présentation nous avons parlé de:

- Le **domaine** de l'intelligence artificielle (IA)
- Les **caractéristiques** d'intelligence artificielle embarquée (IAE)
- Le **matériel** disponible pour l'implémentation des modèles d'IAE
- Le **logiciel** utilisé pour le développement/entraînement des modèles
- Nous avons donné quelques **exemples de modèles**