SEI - SoC - CNN

S. Mancini

Grenoble INP phelma

Plan

- Projet d'intégration des SoC
- Détails d'organisation
- CNN cible : CIFAR10

Objectifs

Réaliser un système de traitement HW et SW, et arriver jusqu'à une implémentation avec caméra et affichage

Mettre en oeuvre toutes les méthodes et outils dont vous avez besoin.

Votre rôle

Vous avez tous le même algorithme, et vous partez d'une page blanche.

Votre mission est:

- Spécifier la solution
- Définir les techniques de résolution
- Mettre en oeuvre la solution
- Evaluer vos résultats, selon les critères usuels

⇨ Votre attitude est déterminante et vous devez chercher des solutions par vous même.

Notre rôle

- Nous vous donnons des indications méthodologiques
- Nous vous dépannons sur les outils ...
- ... lorsque vous avez exploré par vous même

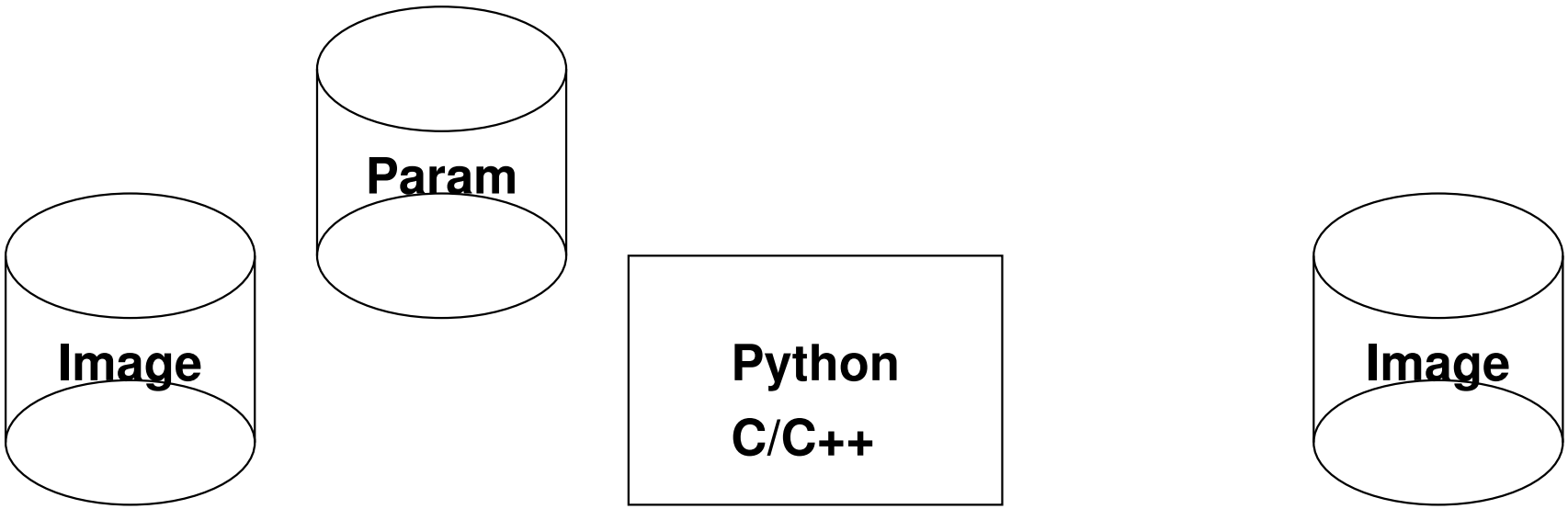
Les connaissances dont vous avez besoin

... et que vous avez !

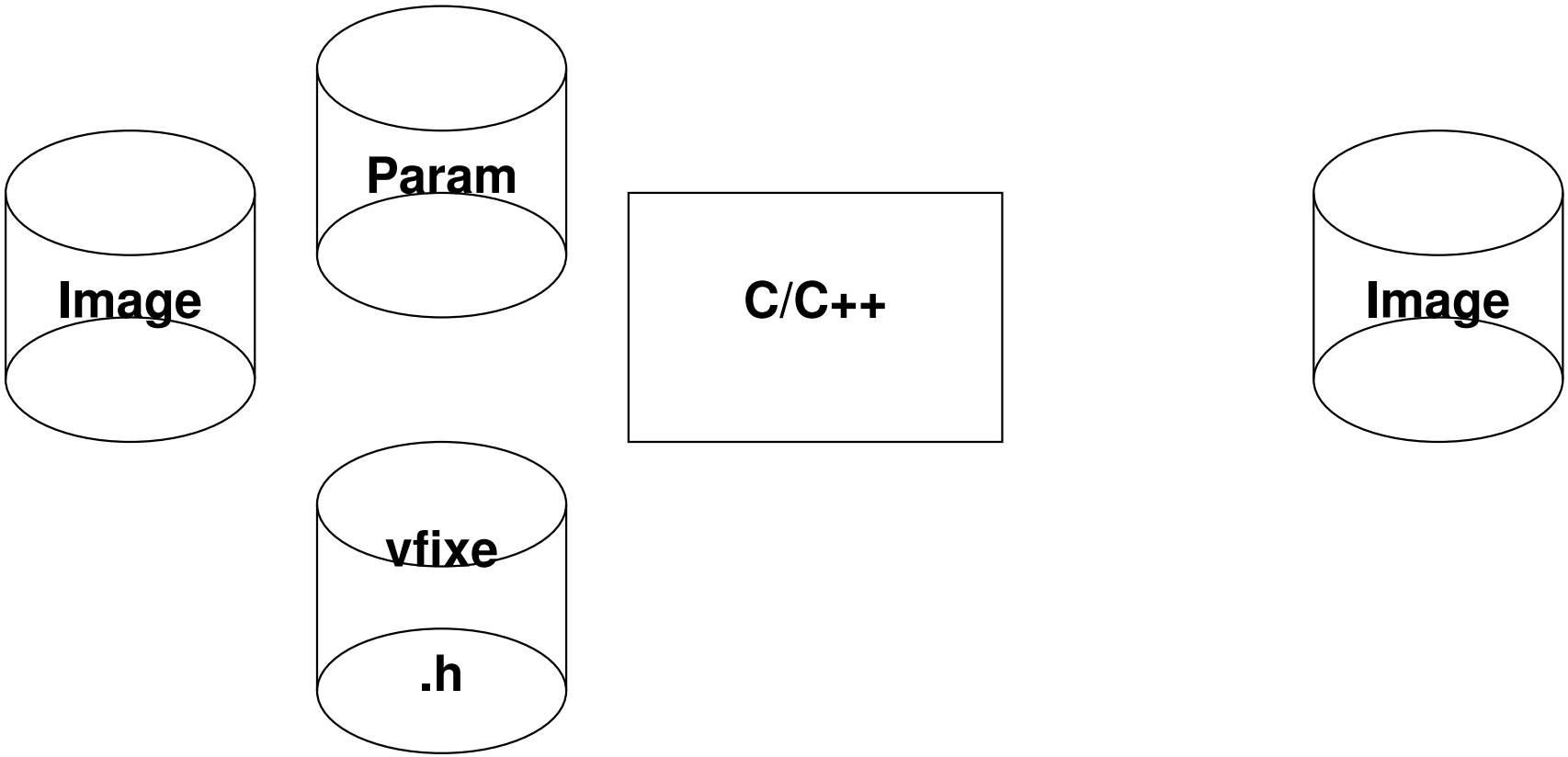
- Passer d'une spécification algorithmique à une architecture de traitement HW ou SW
- Modélisation RTL (VHDL) et HLS (CatapultC)
- Informatique embarqué (C)
- Conception FPGA Xilinx ou Altera
- Outils informatiques usuels (Linux)
- Python vous sera d'une grande aide

Algorithme !

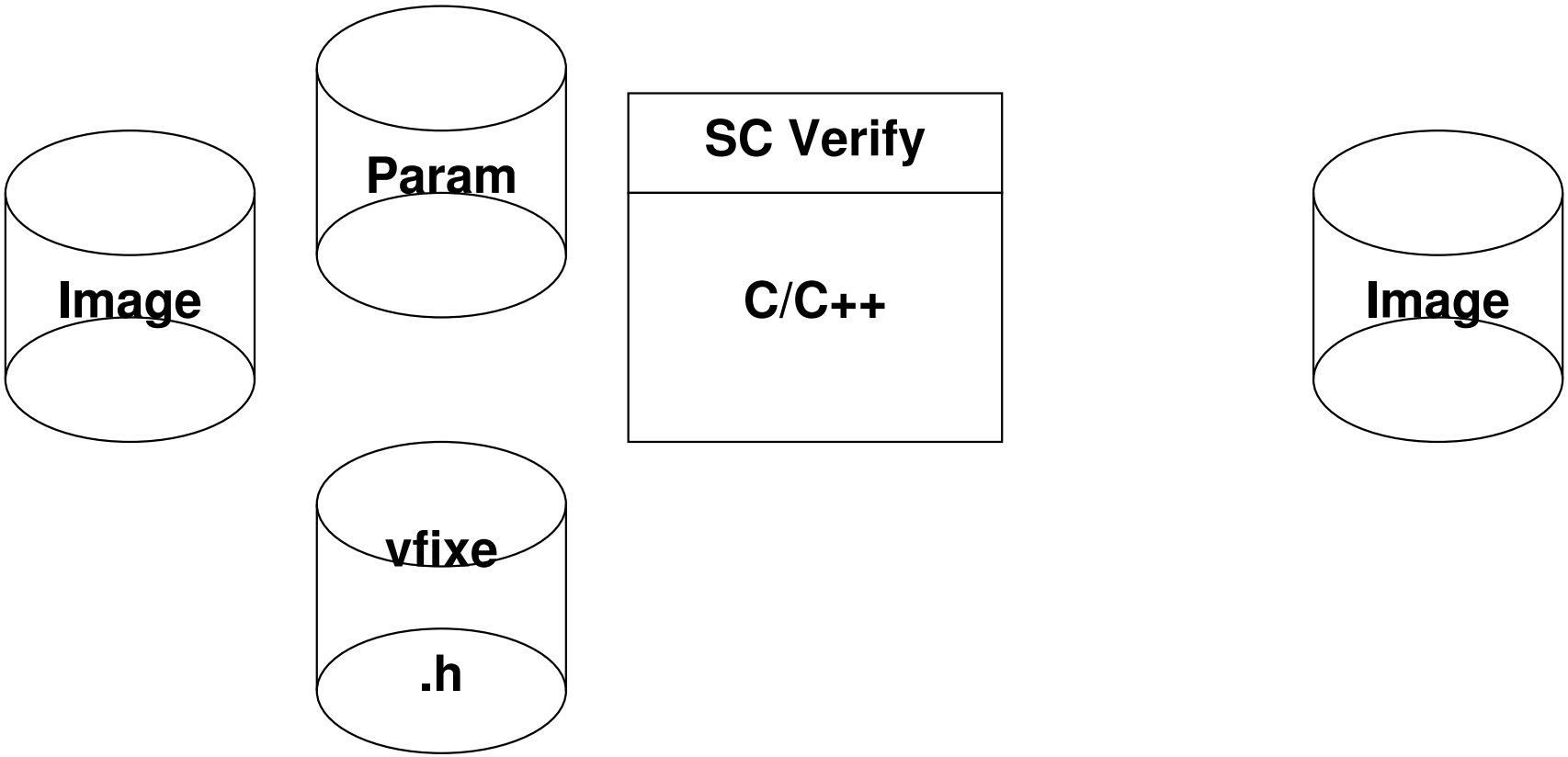
Etapes de la méthodologie HW



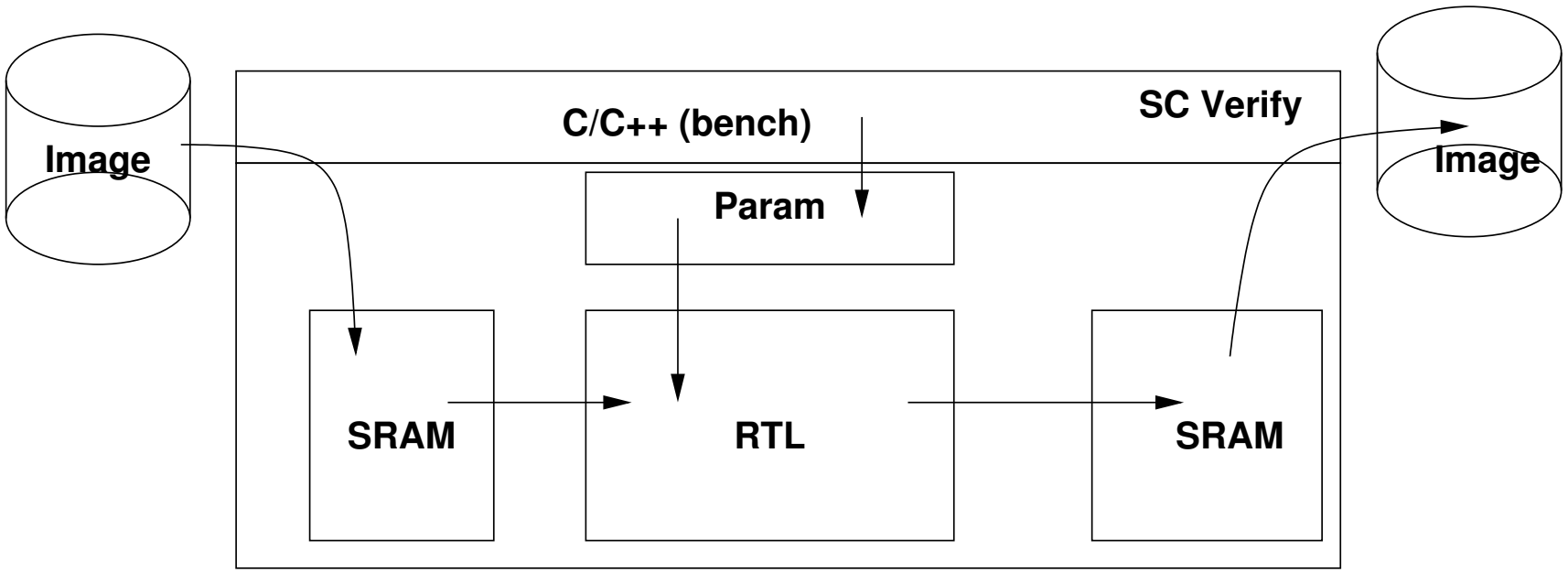
Etapes de la méthodologie HW



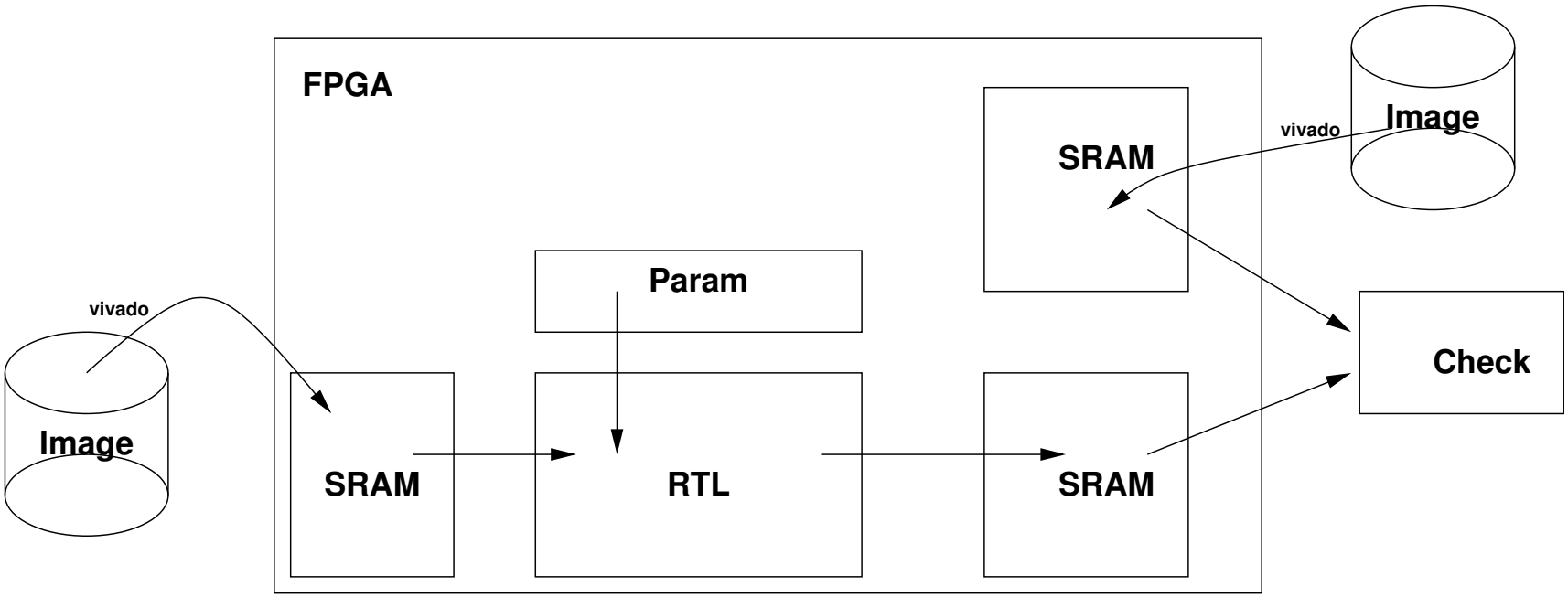
Etapes de la méthodologie HW



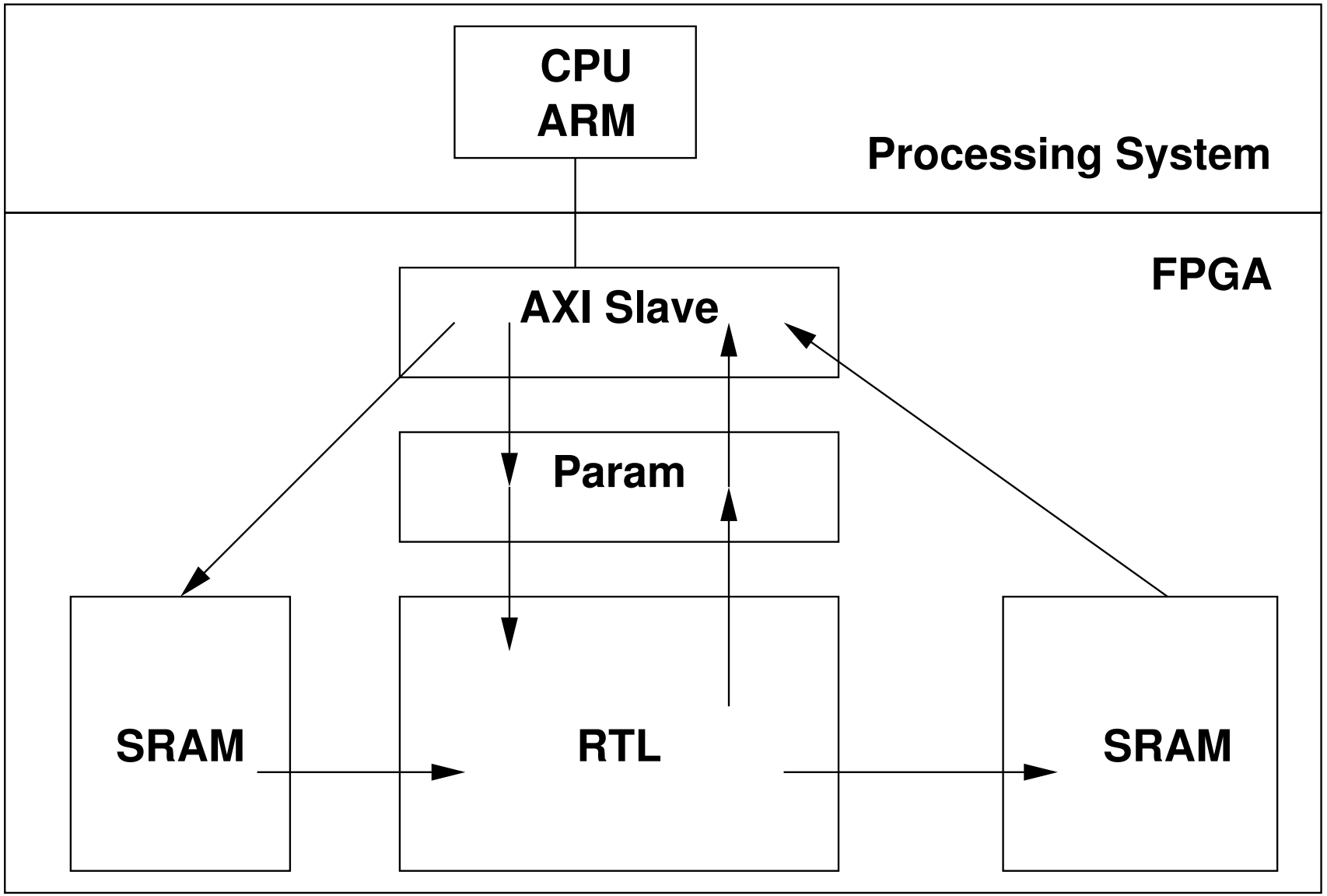
Etapes de la méthodologie HW



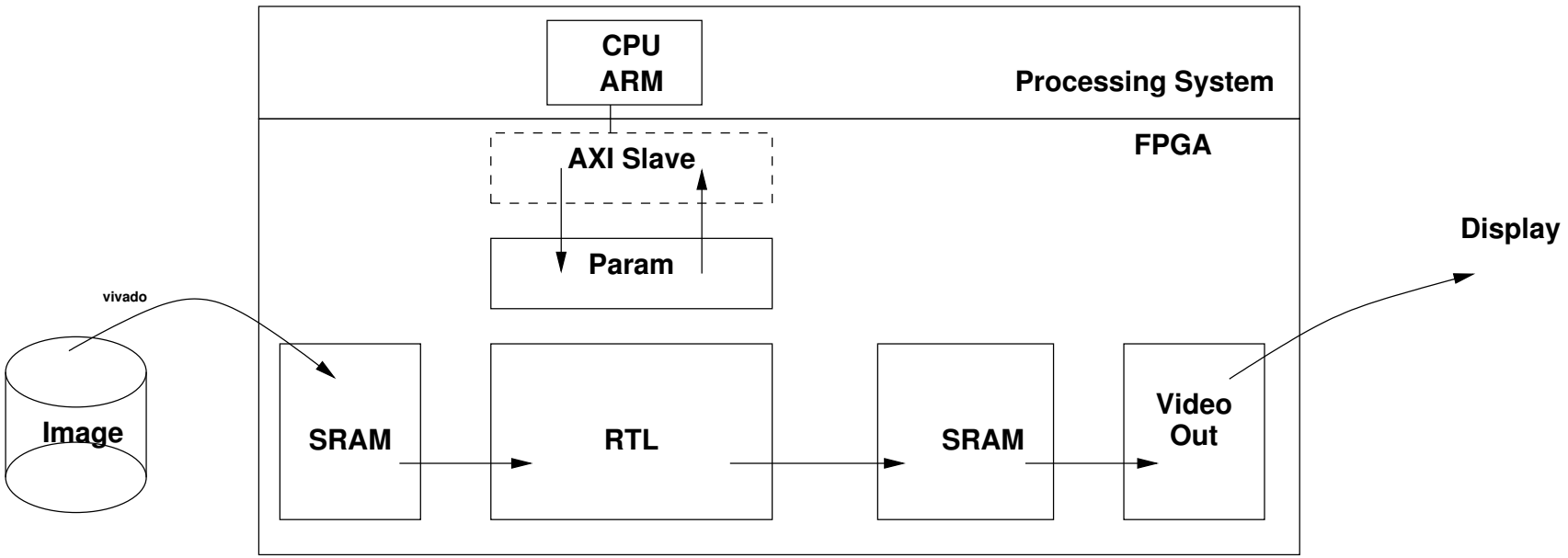
Etapes de la méthodologie HW



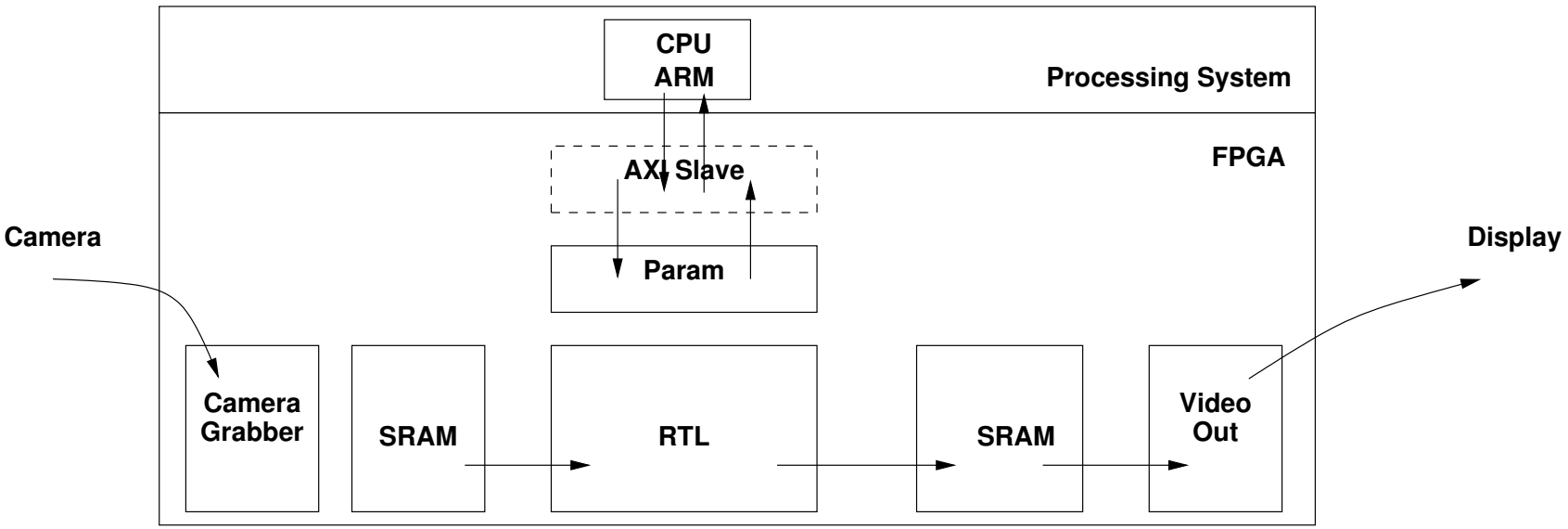
Etapes de la méthodologie HW



Etapes de la méthodologie HW



Etapes de la méthodologie HW



Fonctions génériques 1/2

Une tâche d'IA consiste à déterminer les données 'génératrices' x qui ont produit une 'observation' y (mesure).
Par exemple, la donnée génératrice 'chat' génère toutes les images possibles de chats !

Hypothèse : il existe une fonction f qui permet de passer d'une 'observation' y à ses données 'génératrices' x:

$$x = f(y)$$

Problème: f est inconnue à-priori.

Le **"deep-learning"** consiste à chercher f à partir d'un ensemble d'observations, qui est la base de données d'apprentissage: on connaît des données $\mathbf{y}$ et x associées.

On parle de base de données **étiquetée**.

Fonctions génériques 2/2

On construit f à partir de fonctions génériques paramétrables de façon systématique:

$$f(\theta, y) = f_{n-1}(\theta_{n-1}, f_{n-2}(\theta_{n-2}, \ldots f_0(\theta_0, y)))$$

Les fonctions f_n sont les **“couches”** du réseau, $\theta = \{\theta_0, \ldots \theta_{n-1})$ sont les coefficients du réseau, dont les valeurs proviennent d’un apprentissage.

Les couches d’un réseau de neurones sont typiquement:

- Fonctions linéaires
  - Couches ’totalement’ connectées : le perceptron
  - Convolution
- Fonctions non-linéaires “ **d’activation**” (RELU, sigmoïd, . . .)
- Sous-échantillonnage (maxpool, . . .)

L'apprentisage d'un réseau de neurones

L'apprentissage consiste à déteminer $\theta = (\theta_0, \ldots, \theta_{n-1})$ pour une base de données étiquetée $(x, \mathbf{y})$. Schématiquement, on cherche à déterminer θ qui minimise l'erreur de l'application de f sur les données étiquetées et l'étiquette attendue.

- Un objectif : la mission du réseau de neurones
- Une base de données d'apprentissage

  Des données connues et étiquetées du résultat attendu, pour l'apprentissage des paramètres θ et le test du réseau
- Une 'architecture' de réseau

  Le nombre et la nature des couches.
- Une méthode d'apprentissage
  - Fonction de coût et régularisation

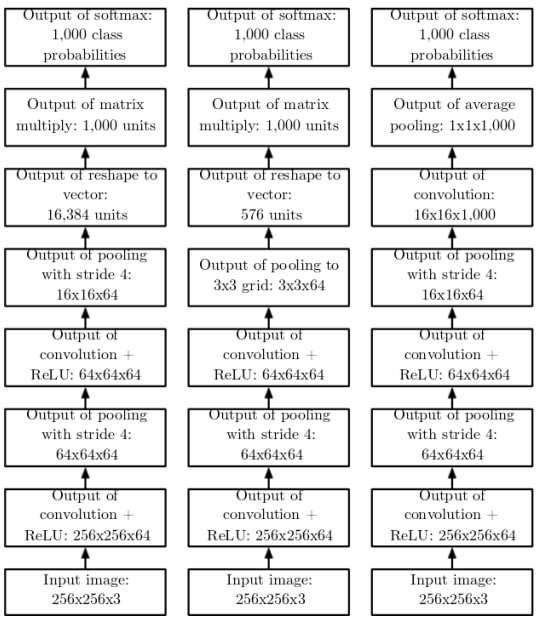
    Pour forcer les paramètres à respecter certaines caractéristiques.
  - Méthode d'optimisation

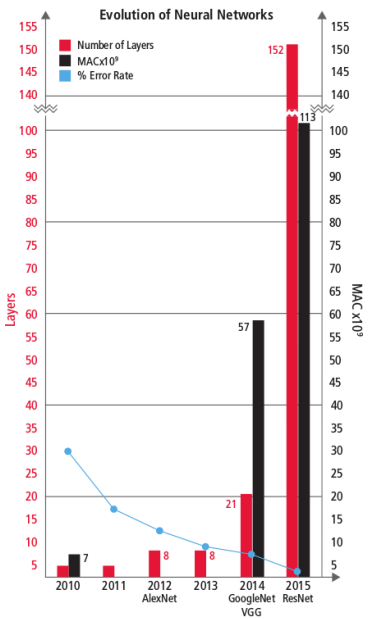
    Le plus souvent descente de gradient stochastique

CNN

Les **C**onvolutional **N**eural **N**etwork, sont une technique de Deep Learning pour de nombreuses tâches de **traitement d'image**: classification, détection, identification, amélioration de qualité, flot optique. . .

Un CNN est constitué seulement des couches de convolution, sous-échantillonnage, "pooling", RELU et 'complètement connecté'.





Convolution

Un **kernel** K de **convolution** est appliqué sur une 'feature map' (image multi-canal = tableau 3D):

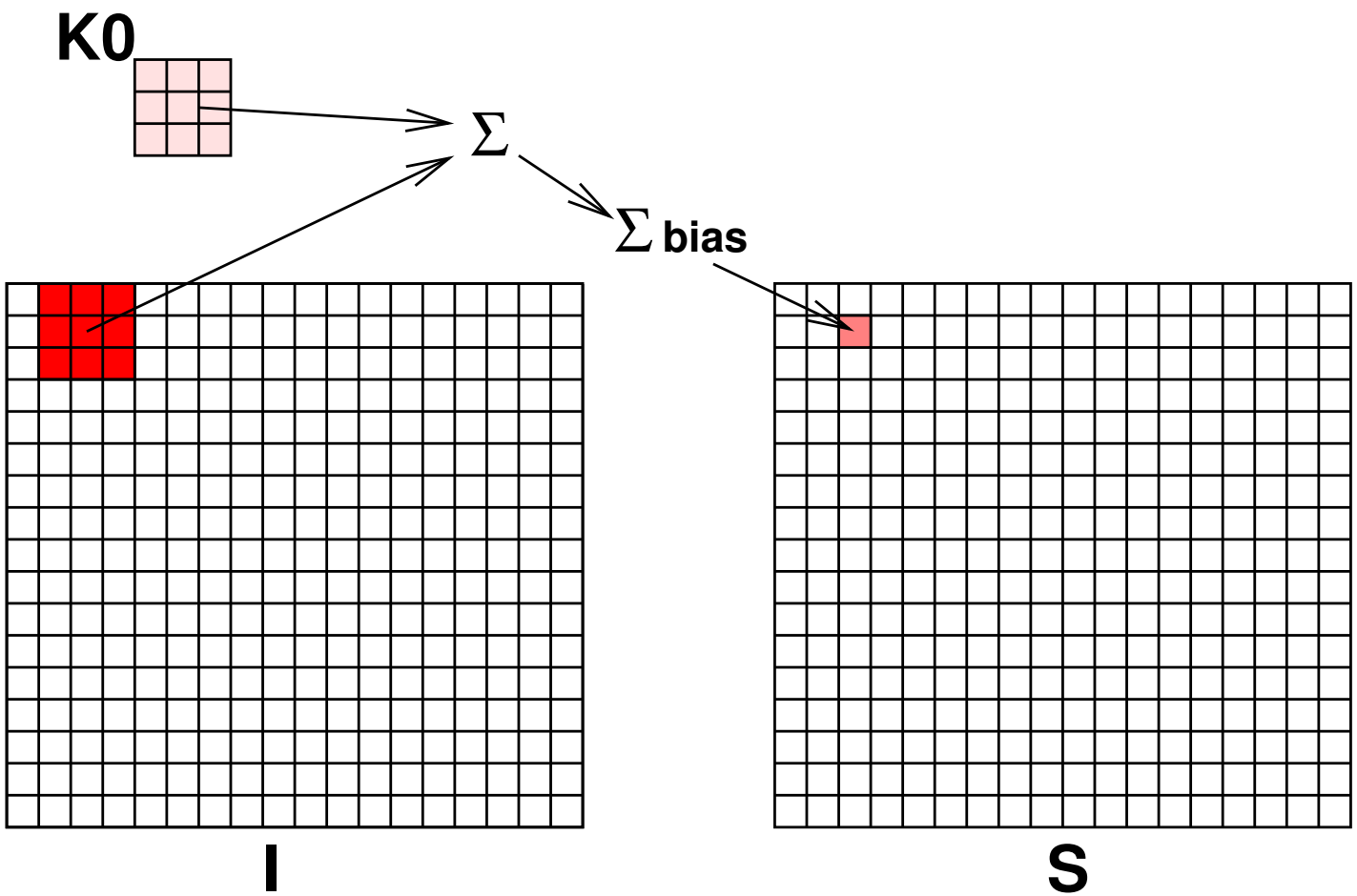
$$S(i,j) = (K \times I)(i,j) = \sum_m \sum_n I_{i+m,j+n} K_{m,n}$$

En réalité, sur une couche, une 'feature map' est multi-canal et chaque valeur a plusieurs 'composantes':

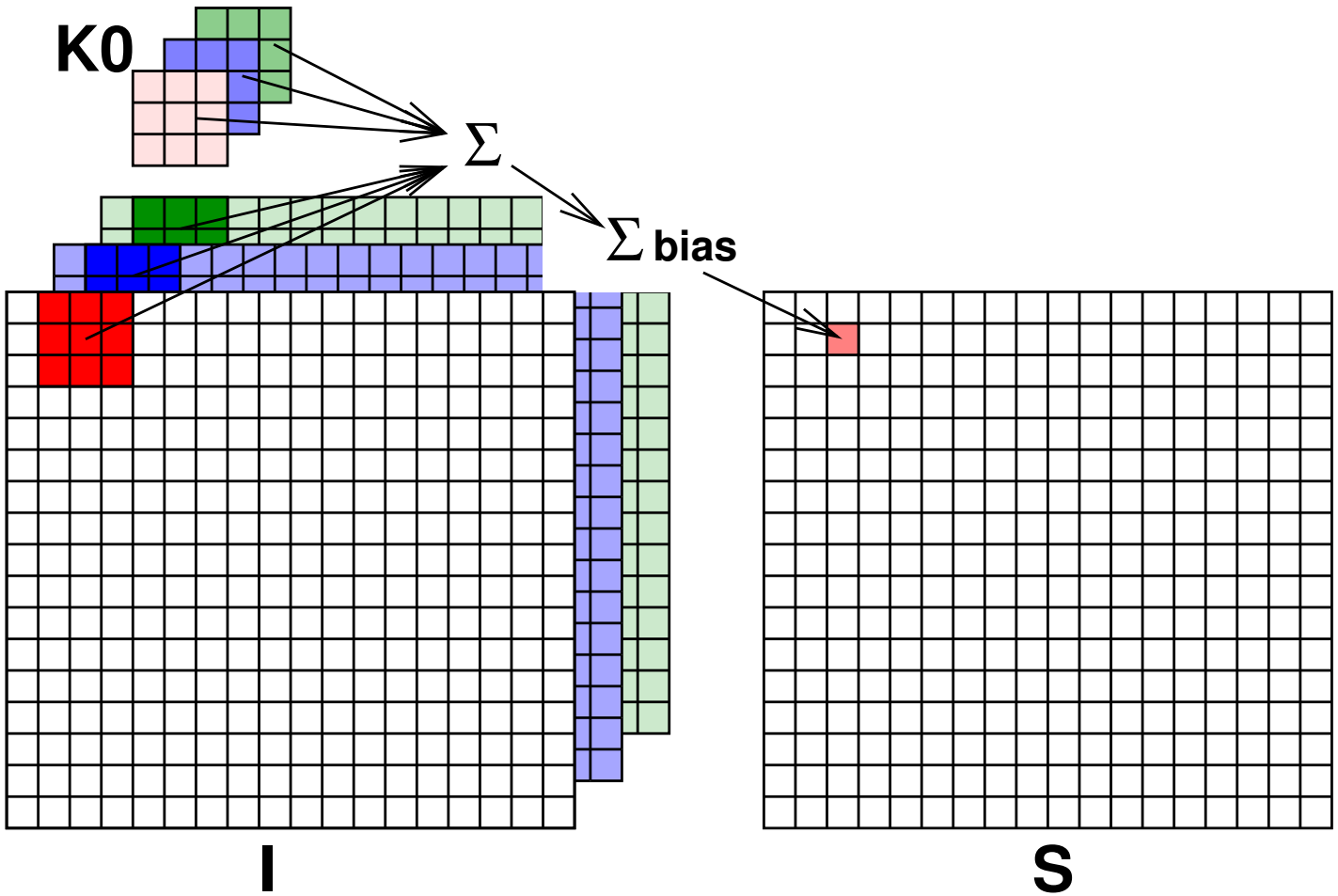
$$S(c,i,j) = \sum_{l,m,n} I_{l,i+m,j+n} K_{c,l,m,n} + \mathrm{bias}_c$$

Un canal de sortie c est produit à partir de tous les canaux l de la 'feature map' d'entrée.

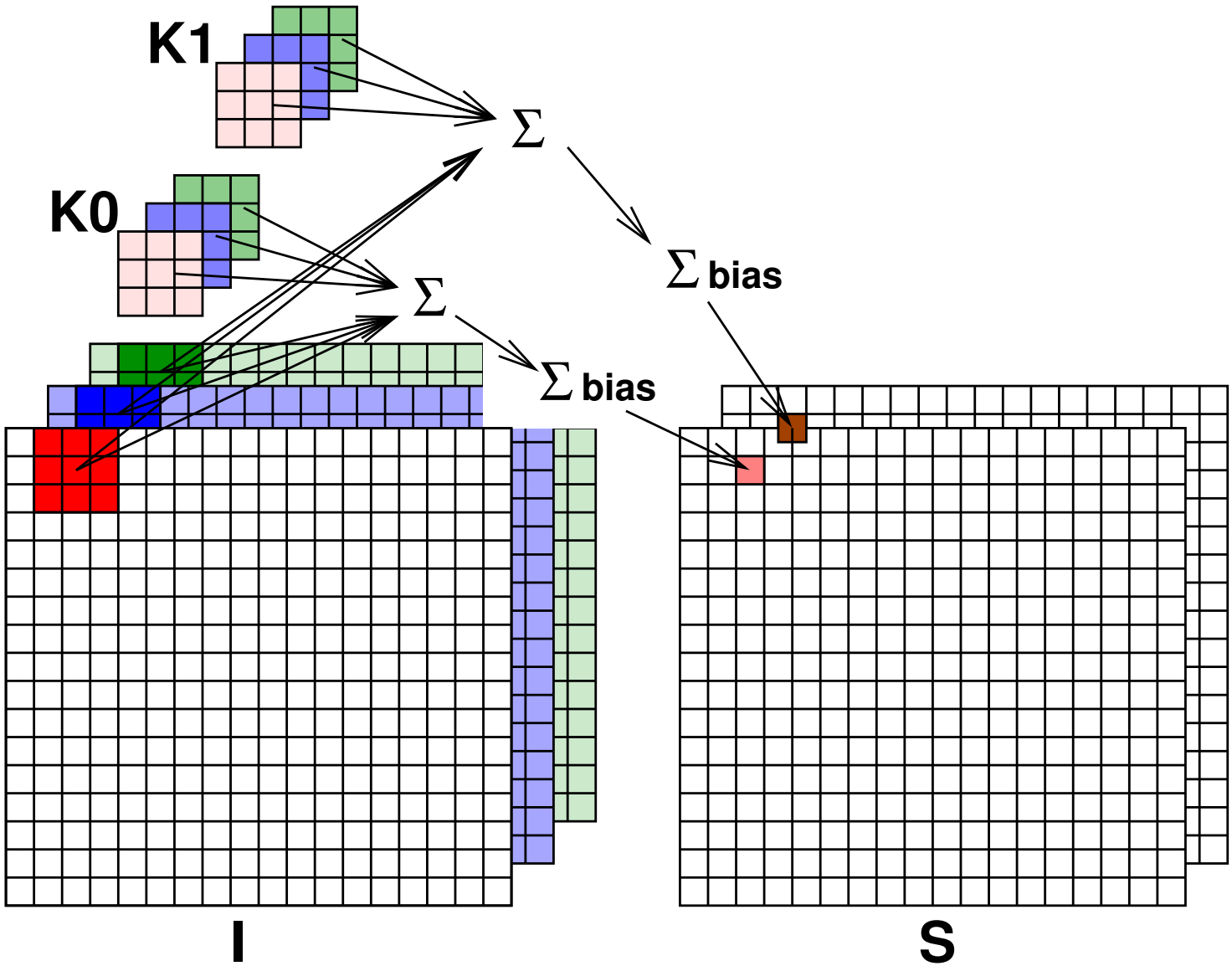
Convolution (bis)



Convolution (bis)



Convolution (bis)



RELU

RELU pour *REctified Linear Unit* est la fonction d'activation non-linéaire des CNN:

$$RELU(a) = \max(0, a)$$

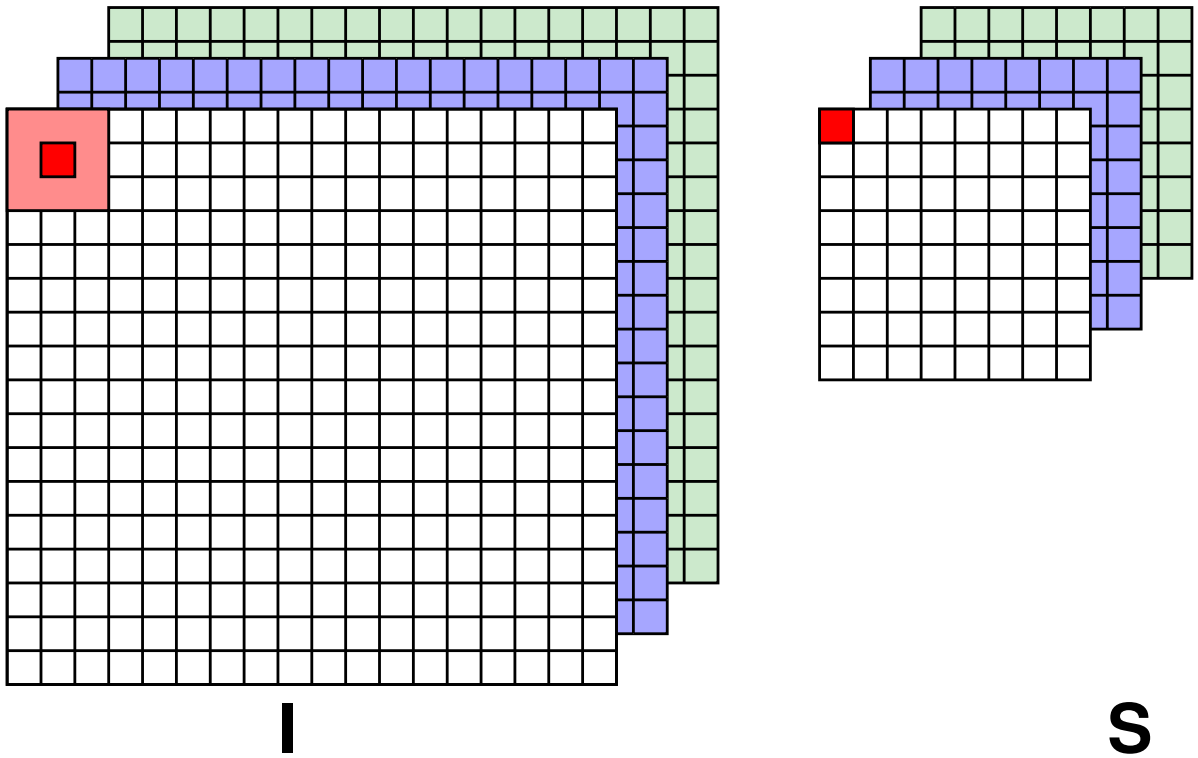
Par rapport à la sigmoïd du perceptron, les avantages de RELU sont:

- Facile à implémenter
- Dérivable (presque) partout, à dérivée non-nulle même pour les valeurs extrêmes
  - L'apprentissage est facilité

MaxPool

Maxpool est l'autre fonction non-linéaire de choix:

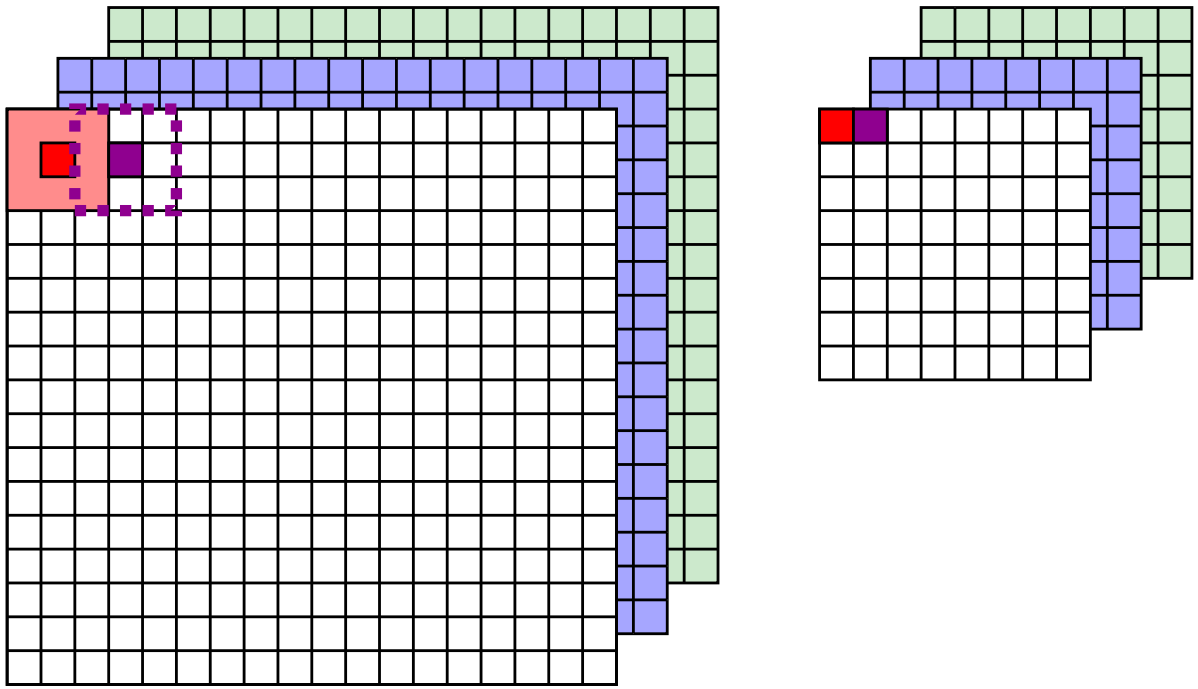
- Sous-échantillonnage pour la réduction des données
- Remplace un groupe de données par son maximum (ou médian, etc . . .)



MaxPool

Maxpool est l'autre fonction non-linéaire de choix:

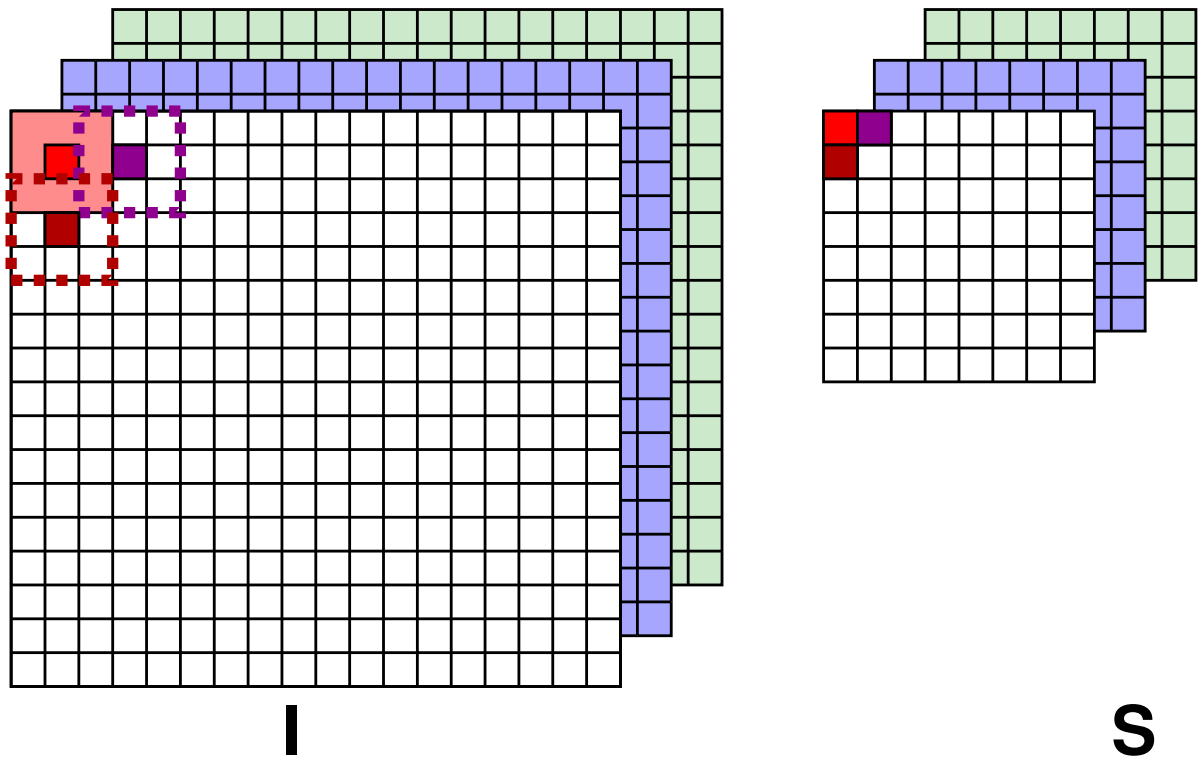
- Sous-échantillonnage pour la réduction des données
- Remplace un groupe de données par son maximum (ou médian, etc . . .)



MaxPool

Maxpool est l'autre fonction non-linéaire de choix:

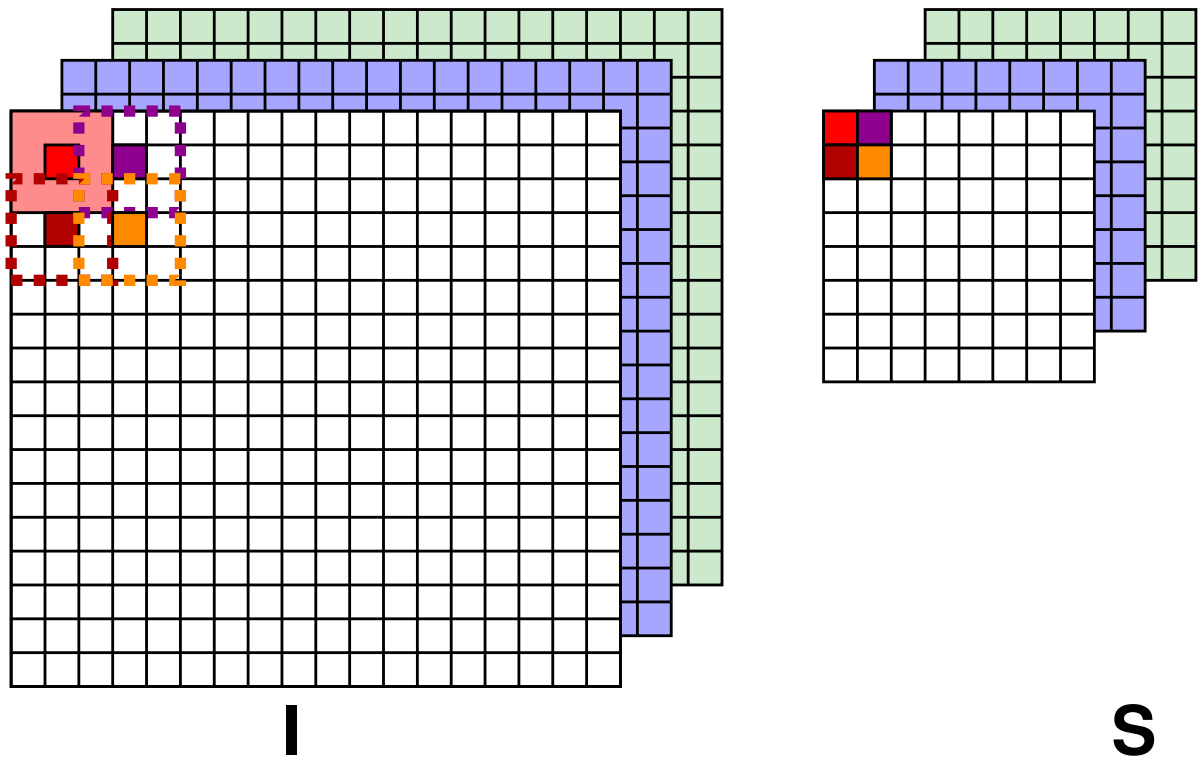
- Sous-échantillonnage pour la réduction des données
- Remplace un groupe de données par son maximum (ou médian, etc . . .)



MaxPool

Maxpool est l'autre fonction non-linéaire de choix:

- Sous-échantillonnage pour la réduction des données
- Remplace un groupe de données par son maximum (ou médian, etc . . .)



SoftMax

Softmax est utilisé pour les tâches de classification car le CNN ne produit pas de valeurs binaires telles qu'une appartenance à une classe:

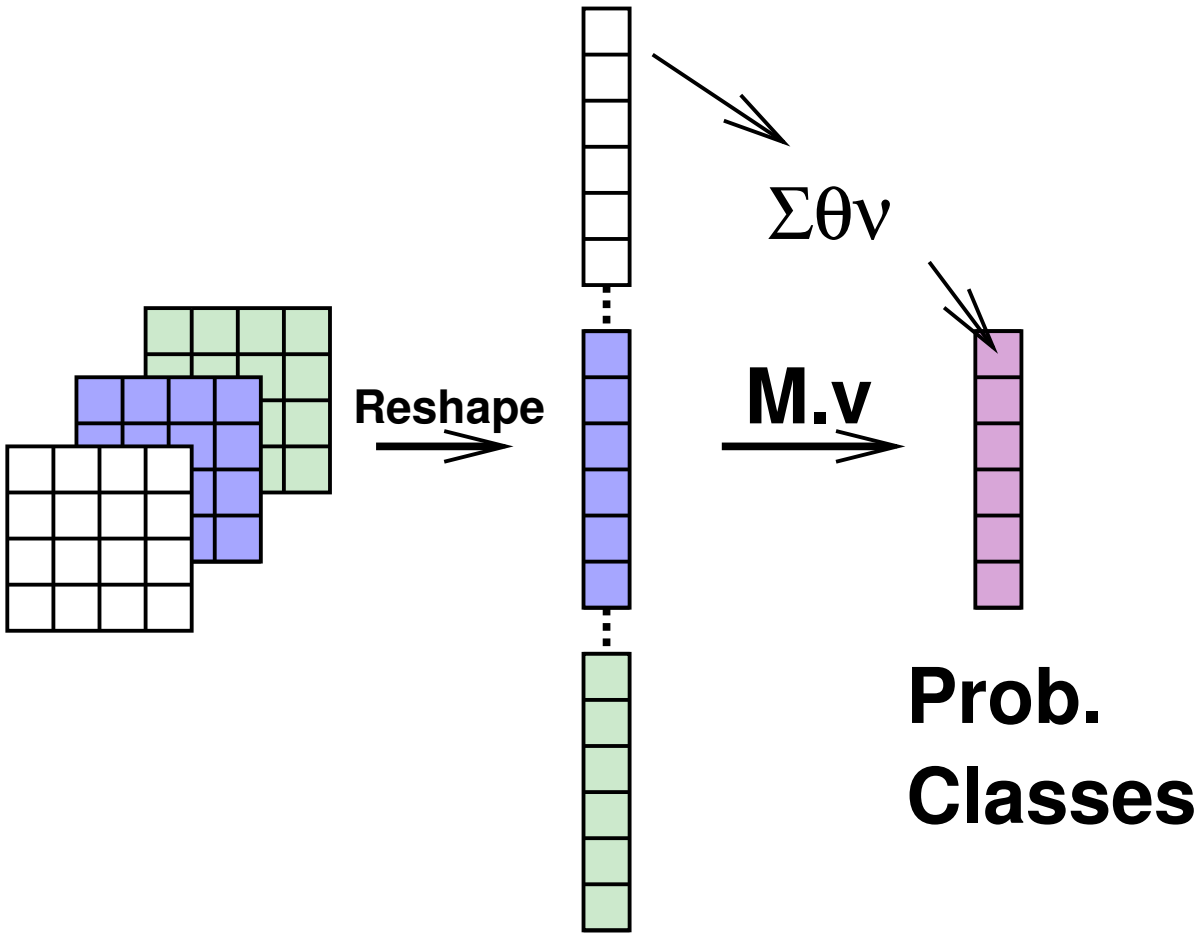
$$\mathrm{softmax}(z)_i = \frac{\exp z_i}{\sum_j \exp z_j}$$

Avec z un vecteur de probabilité pour chacune des classes potentielles. $\mathrm{softmax}(z)_i$ fait 'émerger' la classe la plus probable.

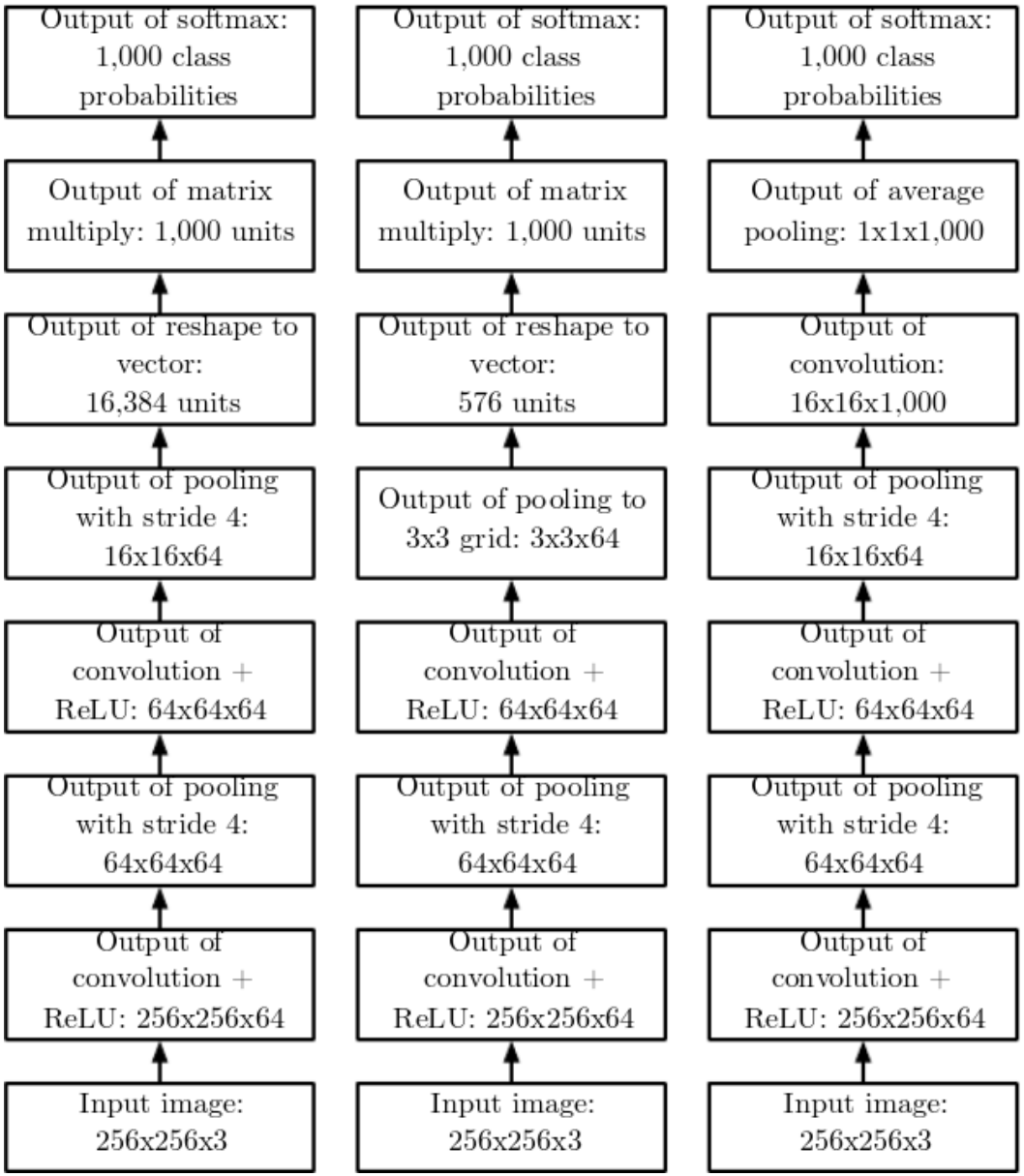
Remarque: **Softmax** provient du monde bayesien et est lié aux techniques d'optimisations statistiques basées sur le maximum de log-vraisemblance (log-likelyhood).

Complètement connectée

Les couches ’**Fully connected**’ sont une combinaison linéaire de tous les élements de la couche précédente: c’est un produit matrice - vecteur.

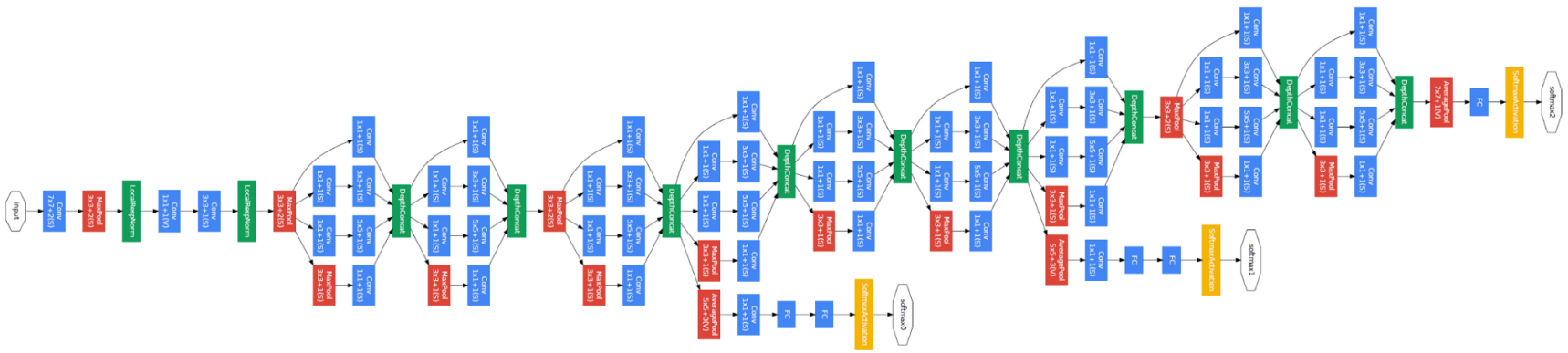


Exemples de CNN



GoogleNet - architecture

Le réseau GoogleNet (Inception V1), fait 22 couches.
Quelques couches effectuent la **normalisation** des features map.



D'après https://medium.com/coinmonks/paper-review-of-googlenet-inception-v1-winner-of-ilsvlc-2014-image-classification-c2b3565a64e7

GoogleNet - Puissance de calcul

type	patch size/ stride	output size	depth	#1×1	#3×3 reduce	#3×3	#5×5 reduce	#5×5	pool proj	params	ops
convolution	7×7/2	112×112×64	1							2.7K	34M
max pool	3×3/2	56×56×64	0								
convolution	3×3/1	56×56×192	2		64	192				112K	360M
max pool	3×3/2	28×28×192	0								
inception (3a)		28×28×256	2	64	96	128	16	32	32	159K	128M
inception (3b)		28×28×480	2	128	128	192	32	96	64	380K	304M
max pool	3×3/2	14×14×480	0								
inception (4a)		14×14×512	2	192	96	208	16	48	64	364K	73M
inception (4b)		14×14×512	2	160	112	224	24	64	64	437K	88M
inception (4c)		14×14×512	2	128	128	256	24	64	64	463K	100M
inception (4d)		14×14×528	2	112	144	288	32	64	64	580K	119M
inception (4e)		14×14×832	2	256	160	320	32	128	128	840K	170M
max pool	3×3/2	7×7×832	0								
inception (5a)		7×7×832	2	256	160	320	32	128	128	1072K	54M
inception (5b)		7×7×1024	2	384	192	384	48	128	128	1388K	71M
avg pool	7×7/1	1×1×1024	0								
dropout (40%)		1×1×1024	0								
linear		1×1×1000	1							1000K	1M
softmax		1×1×1000	0								

Plan

- Projet d'intégration des SoC
- Détails d'organisation
- CNN cible : CIFAR10

Documentation et fichiers dans

`/tp-fmr/smancini/SEI_SoC_CNN/`

Méthode

- Valider votre logiciel ou matériel sur des jeux de coefficients ’fictifs’, dont les valeurs vous permettent de vérifier les calcul
- Utiliser des coefficients appris

La deuxième phase démarre mi-Décembre

Organisation

Projet par **binôme**

- Séances : 40 heures
- Au moins 24 heures non encadrées

Le vendredi matin, jusqu'à fin Janvier.

Evaluation:

- Présentation intermédiaire : 10 mn, le **6/12**
- Rapport + présentation finale : 10 mn , à la dernière séance
- Réalisation minimale pour valider le module: convolution d'image sur FPGA, flux vidéo 'live'

Plan

- Projet d'intégration des SoC
- Détails d'organisation
- CNN cible : CIFAR10

CIFAR 10 database

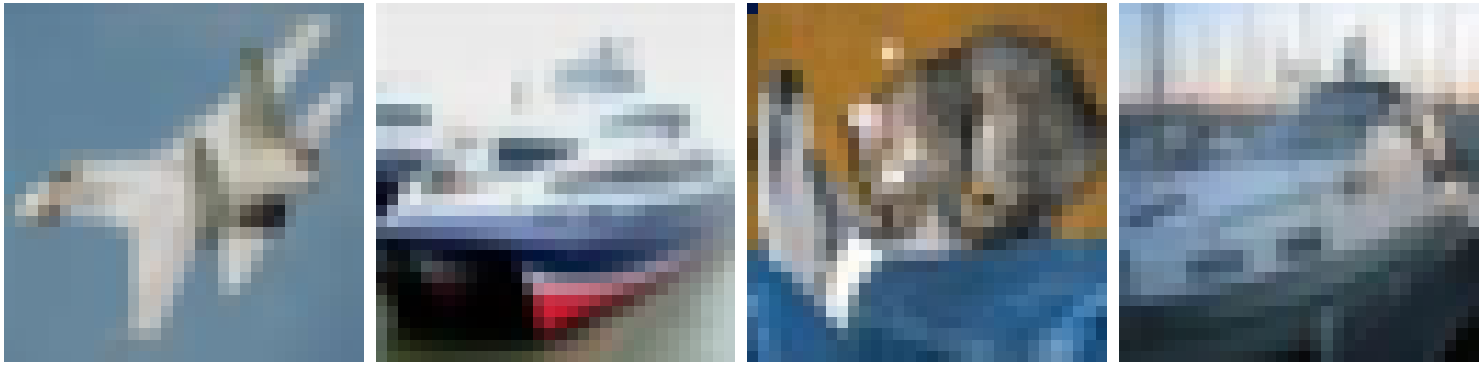
Binary version

The binary version contains the files `data_batch_1.bin, data_batch_2.bin, ...` as well as `test_batch.bin`. Each of these files is formatted as follows:

```
<1 x label><3072 x pixel>
...
<1 x label><3072 x pixel>
```

In other words, the first byte is the label of the first image, which is a number in the range 0-9. The next 3072 bytes are the values of the pixels of the image. The first 1024 bytes are the red channel values, the next 1024 the green, and the final 1024 the blue. The values are stored in row-major order, so the first 32 bytes are the red channel values of the first row of the image.

Each file contains 10000 such 3073-byte "rows" of images, although there is nothing delimiting the rows. Therefore each file should be exactly 30730000 bytes long.



CIFAR 10 : utilitaires

Fichiers dans

`/tp-fmr/smancini/SEI_SoC_CNN/cifar10_data`

Exemple de script pour lire une image:

```
nb=n*3073+1
cat cifar10_data/cifar-10-batches-bin/test_batch.bin |
        dd count=3072 bs=1 skip=<n> > toto.raw
display -size 32x32 -depth 8 -interlace Plane  rgb:toto.raw
convert -size 32x32 -depth 8 -interlace Plane  rgb:toto.raw cifar10_voilier.png
convert -size 32x32 -depth 8 -interlace Plane  rgb:toto.raw cifar10_voilier_bin.ppm
convert -size 32x32 -depth 8 -interlace Plane  rgb:toto.raw -compress none cifar10_voilier
```

Adaptation des images

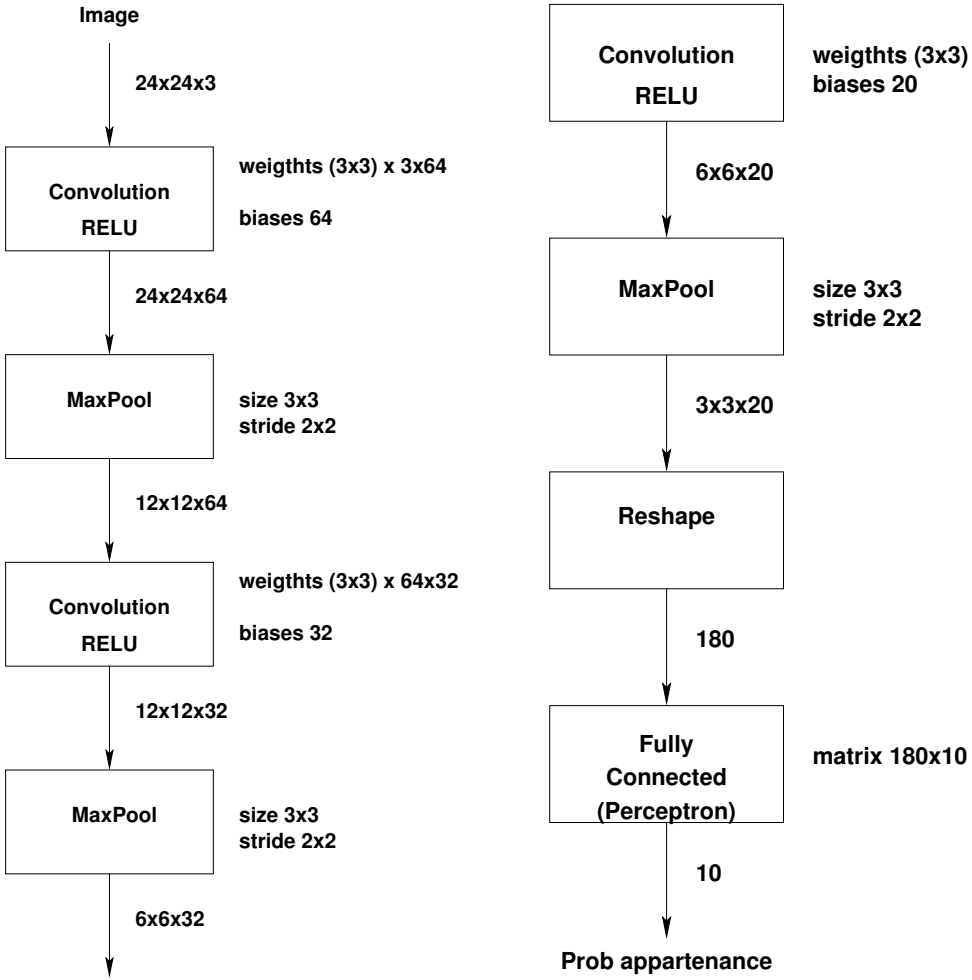
Les images 32x32 sont 'coupées' à 24x24, la coupe étant centrée.

Attention: pour la détection, une image m doit être normalisée:

$$\mu = \frac{1}{N}\sum m_{i,j} \quad ; \quad \sigma = \sqrt{\frac{1}{N}\sum (m_{i,j} - \mu)^2}$$

$$m'_{i,j} = (m_{i,j} - \mu) / \max(\sigma, \frac{1}{\sqrt{N}})$$

CNN 'basique' pour CIFAR10



reshape : transforme la matrice en vecteur, dans l'ordre cannonique
Format des données : **HWC** HeigthxWidthxCanal
Autrement dit, les 'canaux' sont en 'premier'.

CNN 'basique' pour CIFAR10

Les coefficients sont dans

`CNN_coeff_3x3.txt`

et

`CNN_coeff_5x5.txt`

CNN 'basique' pour CIFAR10

TODO List

- Python : code de base pour lecture/ecriture fichier pgm, ecriture coe
- Catapult : Options de base (techno, IO)
  - Altera : biliothèque de mémoires simple & double port
- Modelsim : config qui marche (ok, voir script)
- Vivado :

  interfaces mémoire et affichage

  reset de l'IP CC

  Lien Vivado CC pour paramétrage memoires

  RGB par mot ou par plan
- projet SW : affichage VGA
- CNN :

détail structure des données intermédiaires (dernier perceptron)