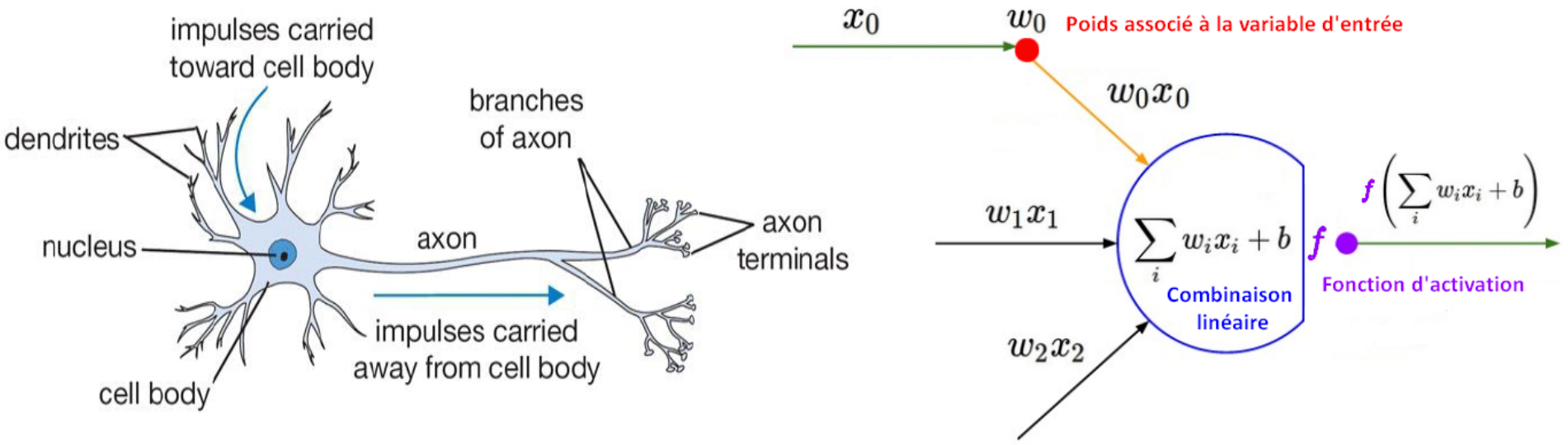
Atelier Happy R: Introduction à MXNet-R pour apprendre à apprendre plus profondément, et en avoir l'R

Christophe Botella
le 18/10/2019

Ingrédients de l'apprentissage profond

- Objectif descriptible et mesurable
  (=fonction de coût)
- Grand jeu de données annoté
- Un pouvoir de calcul suffisant

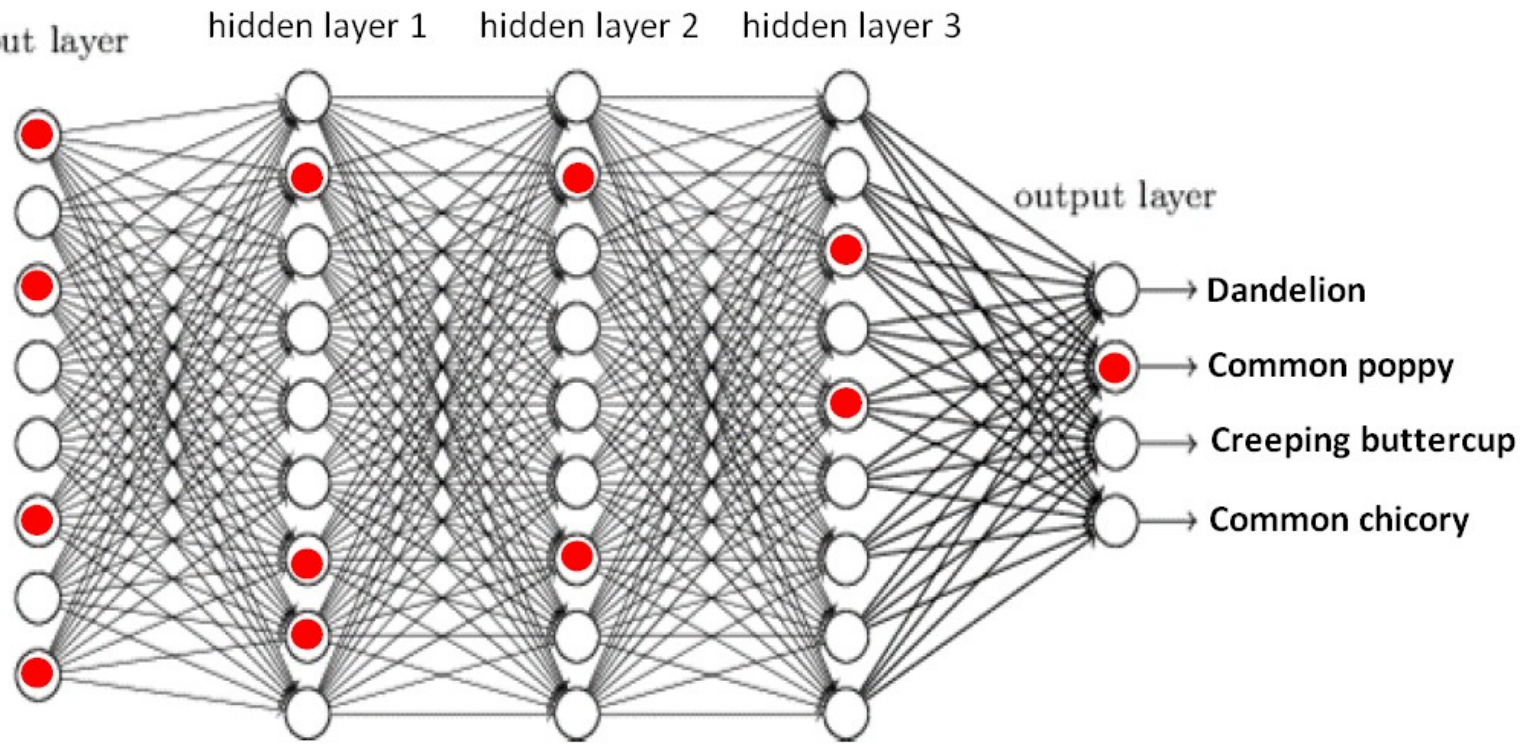
Principe et structure du neurone artificiel

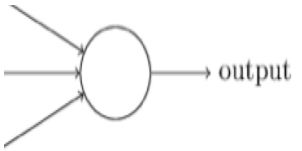


Neurone et neurone artificiel (devinez). Illustration inspirée de Lucas Masuch

Le neurone artificiel sert à quantifier un concept sur la base d'une fonction de variables d'entrée : la composition d'une combinaison linéaire d'une fonction non-linéaire.

Structure d'un
"fully connected feedforwad Neural Network" (NN)

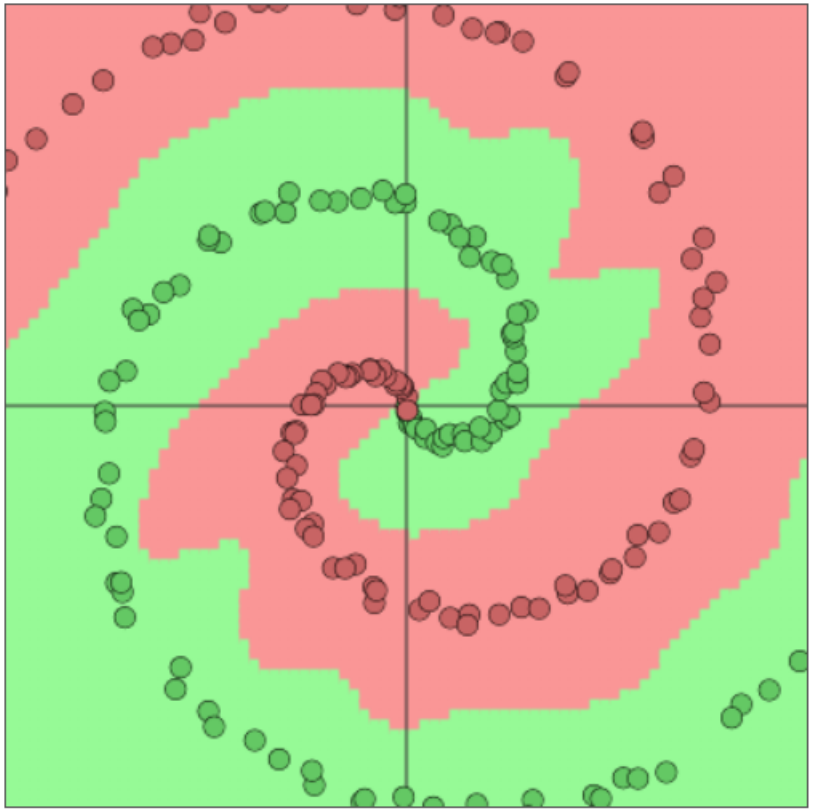




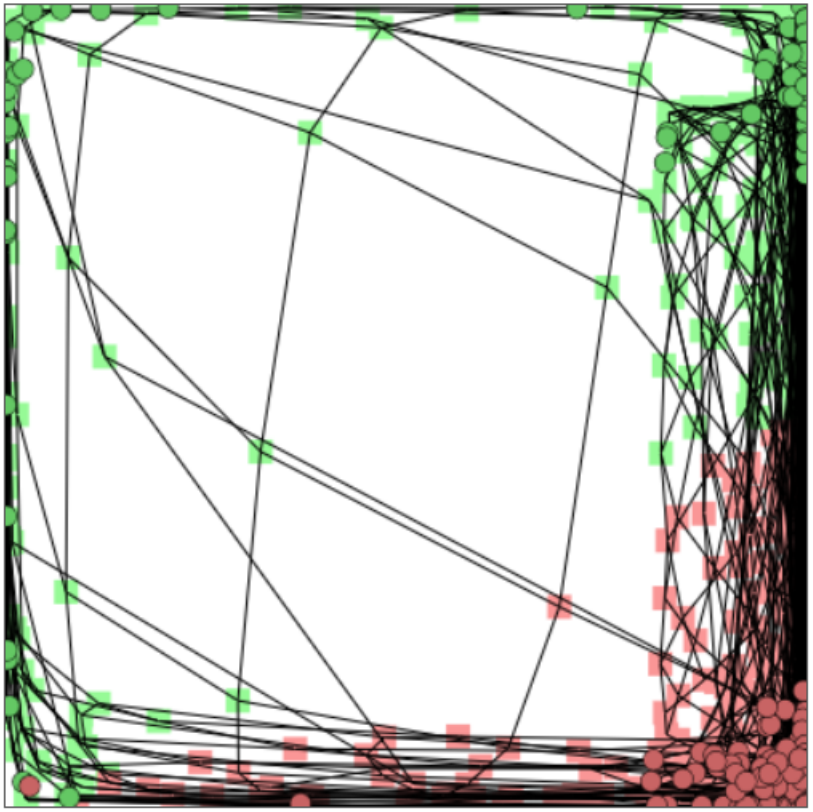
$$a_j^l = \sigma\left(\sum_k w_{jk}^l a_k^{l-1} + b_j^l\right)$$

Extraire automatiquement des représentations complexes des données

- Les réseaux de neurones sont des approximateurs universels, voir Hornik et al. (1989)
- Extrait de : https://cs.stanford.edu/people/karpathy/convnetjs/demo/classify2d.html



Données ponctuelles à classifier (points verts ou rouges) dans le plan.



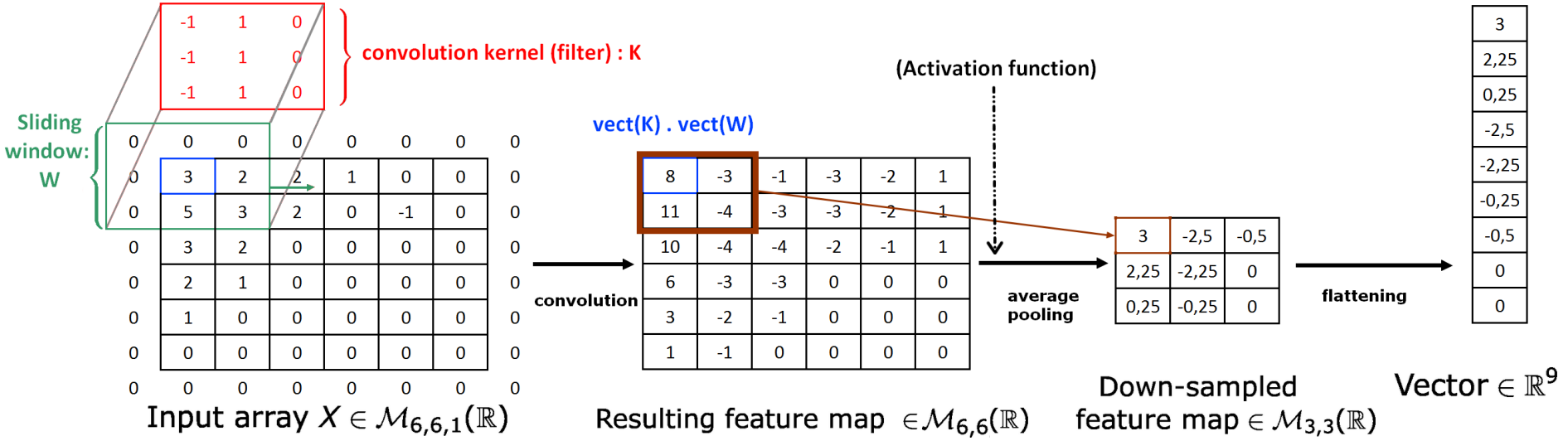
Données (ronds) et points réguliers d'une grille sur le plan (carrés) dans l'espace des activations finales (valeurs des 2 neurones de la dernière couche cachée).

Couches convolutives

Originalité des CNN:

Apprendre des opérations **locales** et **invariantes en translation** sur les images

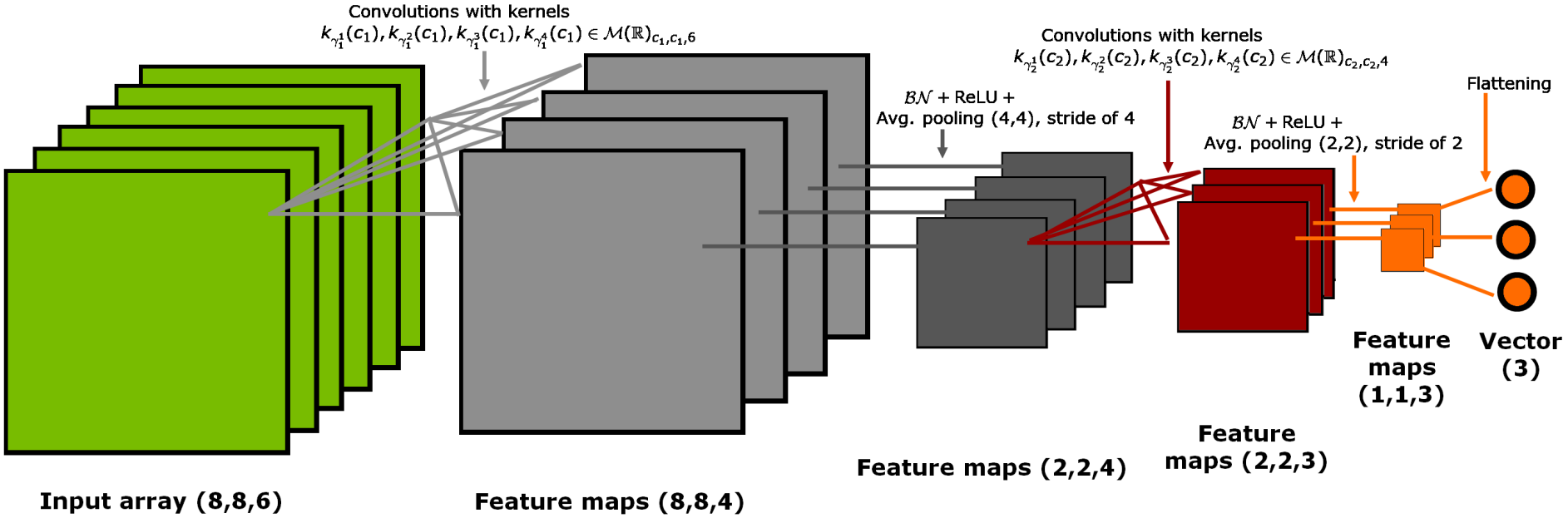
Operations of Convolutional layers: Discrete convolution, Pooling and flattening



En live, ça donne ça :
http://cs231n.github.io/assets/conv-demo/index.html

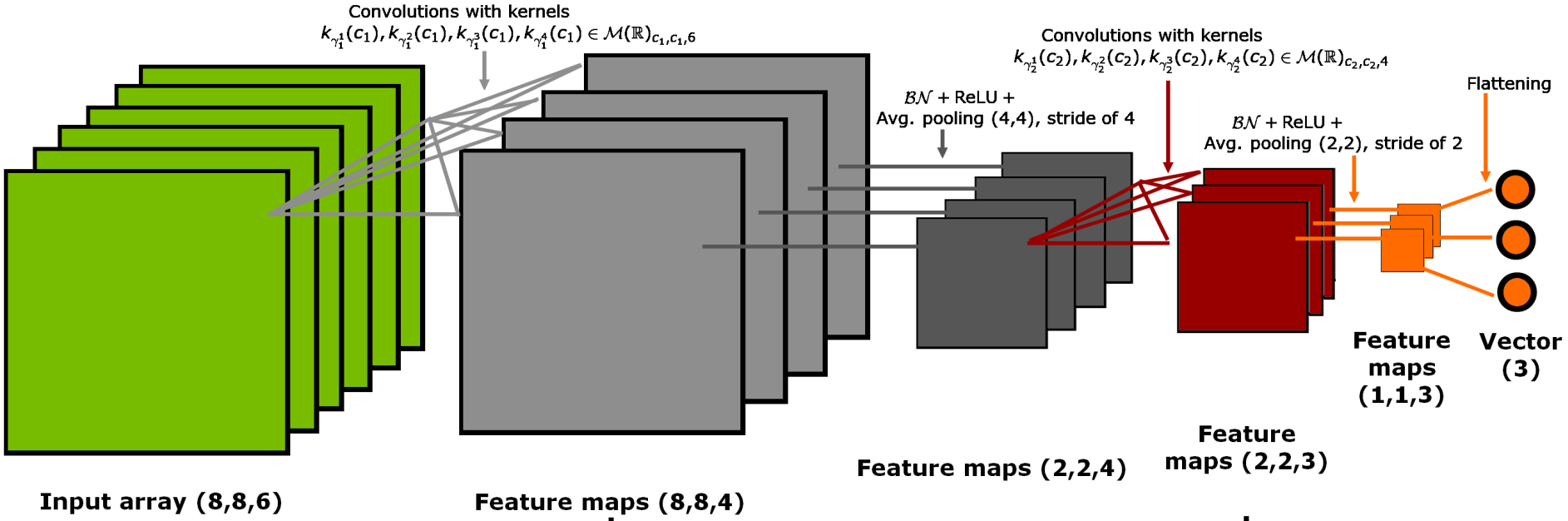
Couches convolutives

Structure complète de 2 couches convolutives successives

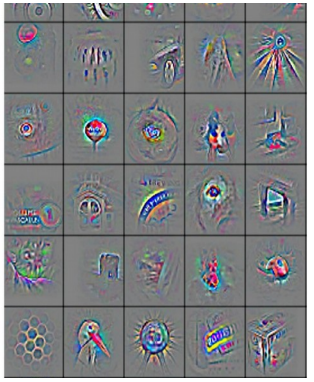


Couches convolutives

Structure complète de 2 couches convolutives successives



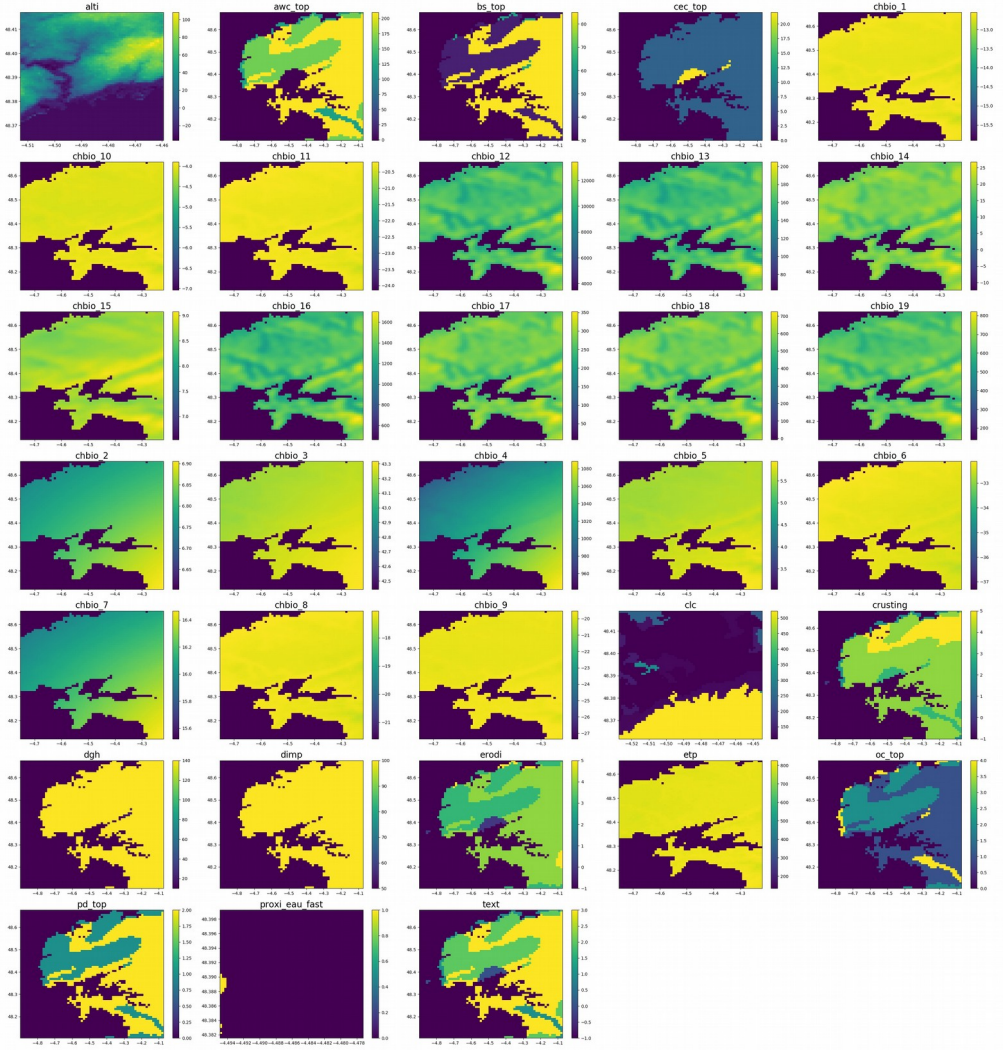
Motifs bas niveau :
Variation brutale de ton, etc



Motifs haut niveau :
œil, patte, feuille, etc

Application des CNN à la prédiction de distributions d'espèces à partir du paysage environnemental

- Pulliam (2000): Une communauté locale d'espèces dépend du paysage local
  - Des communautés voisines
  - De la structuration spatiale de l'environnement

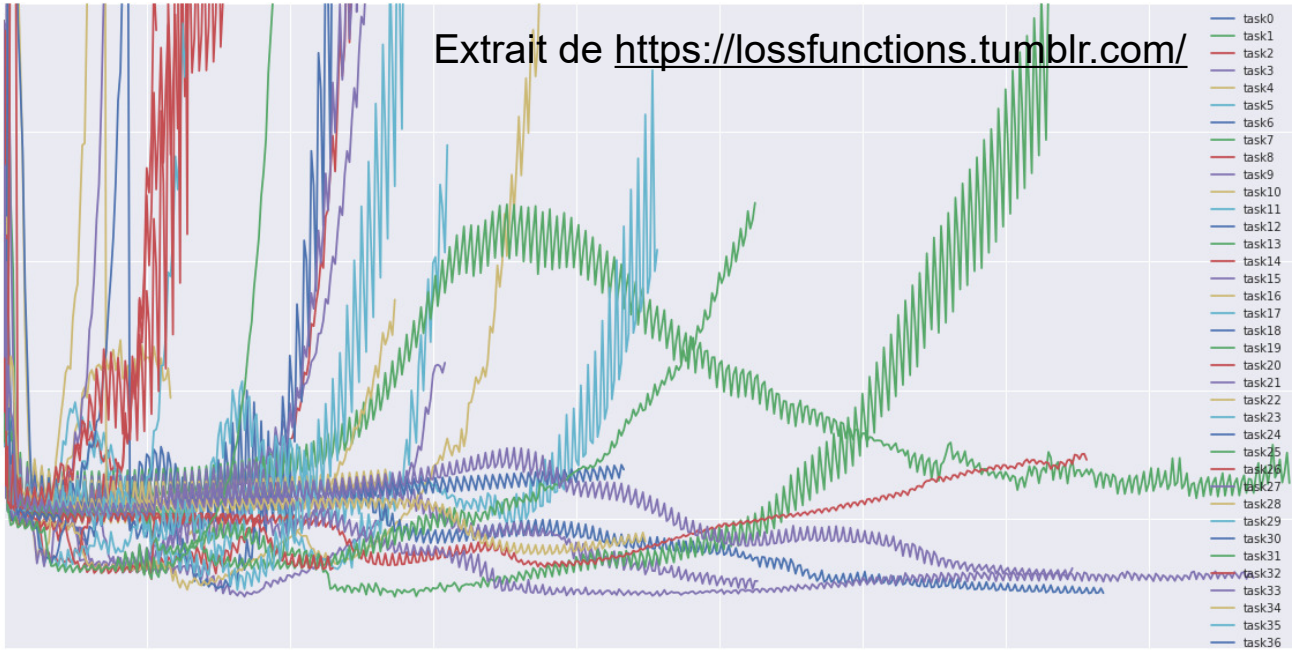


Patch environnemental: 33 fenêtres de variables environnementales centrées à Brest, incluant des v.e. bioclimatiques, pedologiques, topologiques, hydrographiques et d'occupation du sol

Algorithmes stochastiques de descente de gradient

- Encore faut-il les apprendre, tous ces poids... Et c'est la partie difficile.
- L'optimisation se fait par Backpropagation (Rumelhart et al., 1988)
- MAIS descente de gradient classique impossible.
- => Descente de Gradient Stochastique (SGD) = actuellement le plus efficace pour sortir des minima locaux + propriétés de régularisation (Chaudhari et al., 2018).
- Parvenir à optimiser requiert les bonnes valeurs de plusieurs hyper-paramètres :
  - Le taux d'apprentissage / learning rate (pas de l'algorithme)
  - La taille des mini-batch.
  - L'initialisation des poids.
- Plusieurs dérivés de SGD existent (+Momentum, ADAM, RMSprop, ADADELTA, etc), introduction:

  http://ruder.io/optimizing-gradient-descent/index.html#momentum

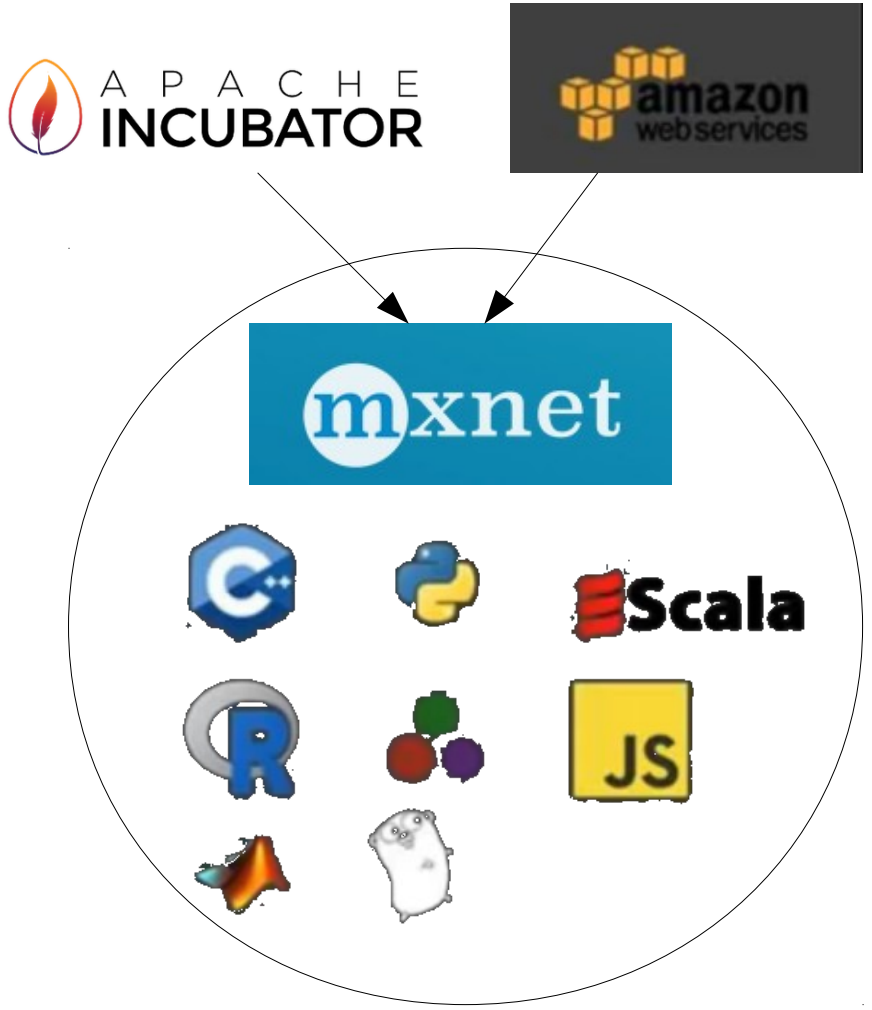


MXNet

- MXNet est une librairie de machine learning multi-langages (basée C++) spécialisée deep learning.
- Politique Open Source

  -> Membre de l'Apache Incubator
- Flexible: Déployé sous 8 langages dont R
- Portable: Librairie légère

  (version cpu)
- Scalable: performance presque linéaire en fonction du nombre de GPUs/CPUs (identiques) => Peu de pertes sur les opérations de gestion du cluster.

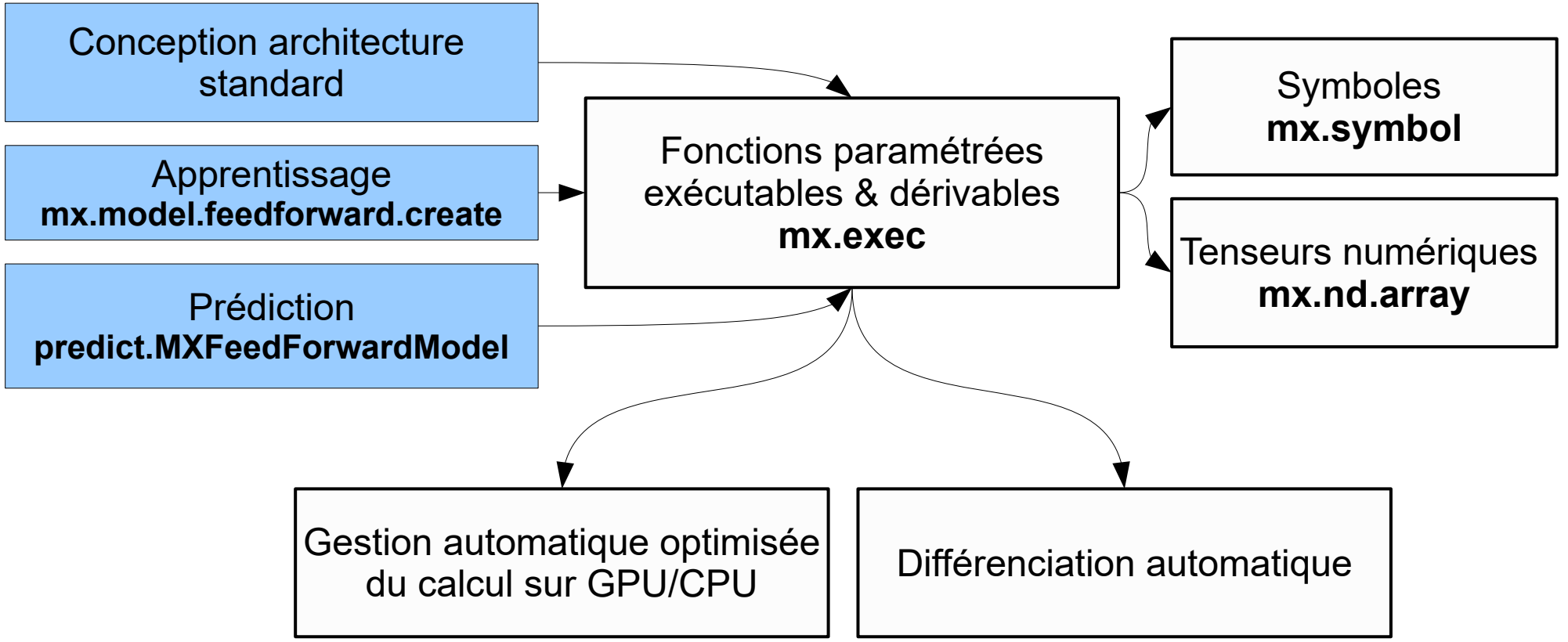


MXNet-R

- R = Langage le plus utilisé en statistiques, écologie et agronomie.
- Besoin pratique de ne pas disperser du temps et du code dans plusieurs langages.
- Avantages package R **mxnet**:
  - Grande flexibilité: fonctions haut et bas niveau.
  - Sur-couche directe de C++, interfacage proche et optimisé avec CUDA pour le calcul GPU.

MXNet-R

Niveau d'utilisation

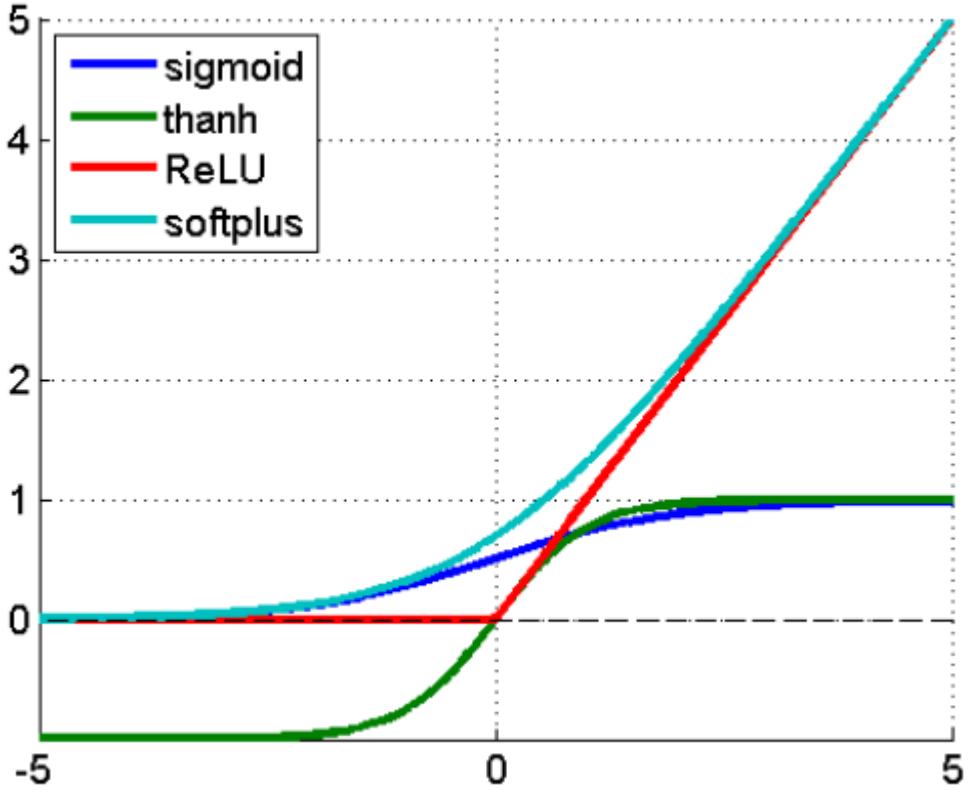


References

- Nair, V. and Hinton, G. E. (2010). Rectified linear units improve restricted boltzmann machines. In Proceedings of the 27th international conference on machine learning (ICML-10), pages 807–814.
- Hornik, K., Stinchcombe, M., and White, H. (1989). Multilayer feedforward networks are universal approximators. Neural networks, 2(5):359–366.
- Rumelhart, D. E., Hinton, G. E., Williams, R. J., et al. (1988). Learning representations by back-propagating errors. Cognitive modeling, 5(3):1.
- Chaudhari, P. and Soatto, S. (2018). Stochastic gradient descent performs variational inference, converges to limit cycles for deep networks. In 2018 Information Theory and Applications Workshop (ITA), pages 1–10. IEEE.
- Chen, T., Li, M., Li, Y., Lin, M., Wang, N., Wang, M., ... & Zhang, Z. (2015). Mxnet: A flexible and efficient machine learning library for heterogeneous distributed systems. arXiv preprint arXiv:1512.01274.

Fonctions d'activation

- Déf: Une fonction monotone et dérivable.
- La fonction ReLU (Nair et Hinton, 2010) est la plus utilisée aujourd'hui. Elle évite le "vanishing gradient" et accélère l'apprentissage.



Différentes fonctions d'activations utilisées en apprentissage profond. Illustration tirée de Lucas Masuch.