# Mini-Rapport

Exploration de la notion de méta-apprentissage

*Dans quelle mesure un système apprenant peut prendre conscience de ses performances et altérer son comportement ?*

Yann Boniface, Alain Dutech, Nicolas Rougier
Matthieu Zimmer

18 avril 2012

# 1 Introduction

Notre intérêt s'est tourné vers l'article [Cleeremans Alex(2007)] et ses 2 types de réseaux proposés. Dans un premier temps, nous avons cherché à reproduire et expliquer les résultats donnés, et ensuite à des solutions pour tirer profit des paris réalisés.
Nous nous sommes également penchés sur [Pasquali Antoine(2010)] dont nous avons reproduit les expériences, mais leurs enjeux nous semblent encore vagues.

# 2 Dupliquer le premier réseau

## 2.1 Les bases

En premier lieu, rappelons la structure des réseaux et les résultats de l'article :

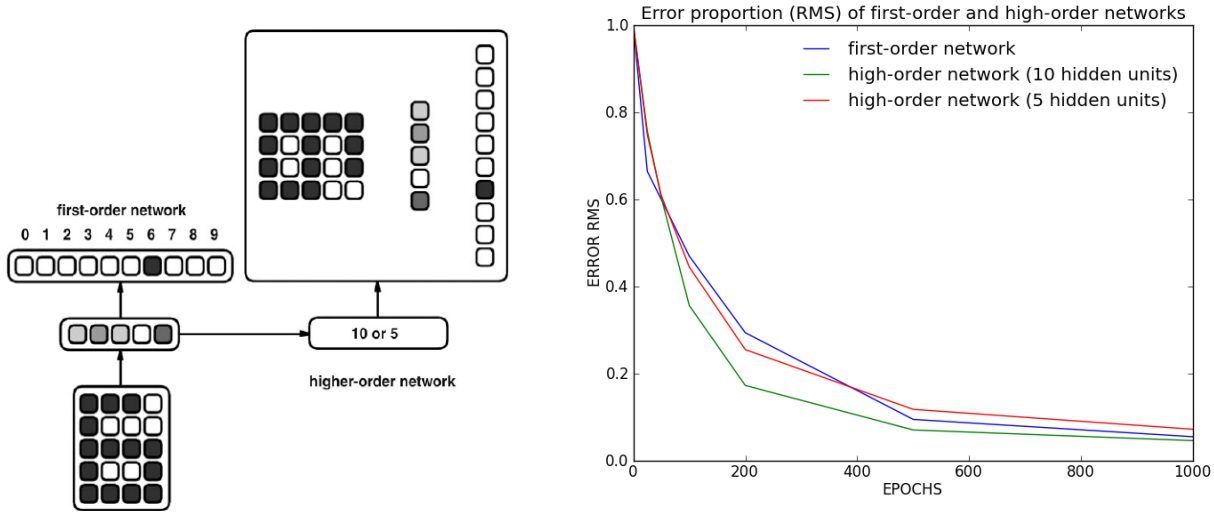


FIGURE 1 – [Cleeremans Alex(2007)] Architecture connexionniste avec méta-représentations

Analysons la formule RMS utilisé à une époque $e$ :

$$rms\ proportion_e = \frac{rms_e = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(o_{i,e} - d_i)^2}}{max(rms_{e'}),\ \forall e' \in epochs}$$

$$with \begin{cases} n : number\ of\ neurons\ on\ the\ output\ layer \\ o_{i,e} : value\ obtained\ for\ the\ i^{th}\ neuron\ at\ the\ e^{th}\ epoch \\ d_i : value\ desired\ for\ the\ i^{th}\ neuron \end{cases}$$

Il faut bien comprendre que c'est une proportion et que ces courbes représentent plus le taux de variation des neurones que le taux de succès à l'apprentissage lui-même. Par exemple, la pente de la courbe sera plus raide en passant d'une erreur de 0.9 à 0.3, qu'en passant de 0.4 à 0.2.

On peut donc penser que le second réseau [1] apprenne plus vite (dans le sens où il avait plus de lacunes au départ), mais pas forcément qu'il apprenne mieux sa tâche.

1. high-order network

Nous avons donc analysé ses performances réelles.

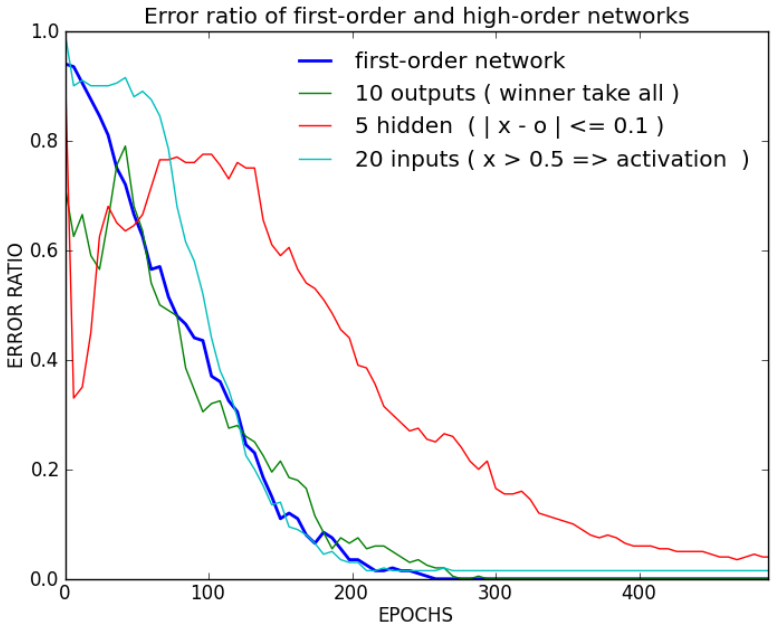


FIGURE 2 – Erreur de classification. Le réseau supérieur à 5 unités cachées n'est plus considéré, et celui à 10 est découpé en 3 courbes pour représenter les 3 couches à reproduire.

Difficile d'en déduire que le second réseau [2] apprend plus vite que le premier [3]. Par contre, on peut expliquer la chute du RMS du second réseau sur l'ancienne figure, par l'apprentissage de la couche d'entrée qui coïncide en même temps (les entrées, étant plus nombreuses, ont plus de poids dans la formule RMS).

## 2.2 Les rouages

Notre attention s'est également portée sur les mécanismes qui permettaient la réalisation de cette architecture. Nous avons alors pu remarquer que les neurones de la couche cachée du premier réseau se stabilisaient très rapidement (autour de la $50^{\text{ième}}$ époque en moyenne), le tout permettant au second réseau d'avoir des entrées très peu variables, favorisant et permettant donc son apprentissage.

2. high-order network
3. first-order network

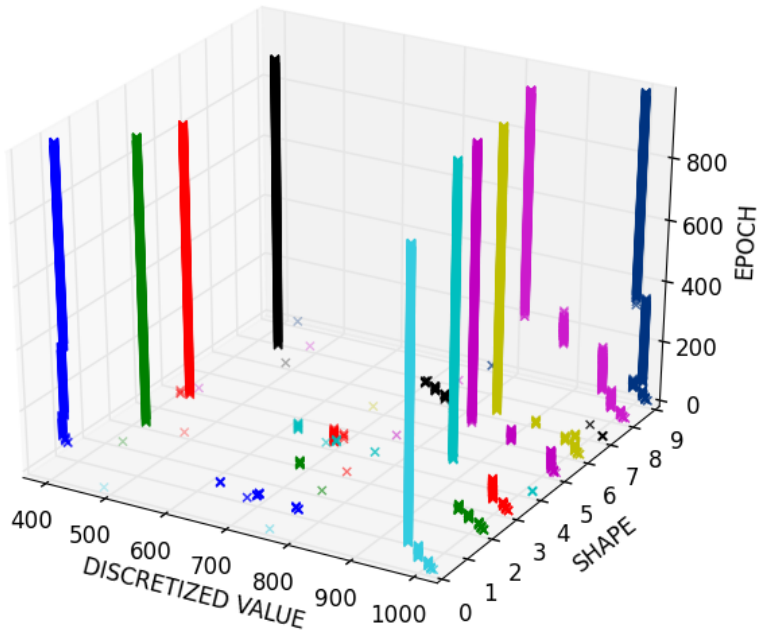


FIGURE 3 – Valeurs discrétisés de la couche caché du premier réseau. Chaque couleurs représentant un des 10 chiffres en entrée. Les courbes deviennent rapidement stables.

## 2.3 Réhaussement

Pour revenir à l'importance des entrées, nous avons réalisé que le second réseau n'était capable de dupliquer le premier qu'uniquement parce que ces entrées sont triviales. Ainsi nous avons augmenté le nombre d'entrées en passant sur des chiffres manuscrits [Semeion Research Center(1994)], tout en augmentant proportionnellement le nombre de neurones des couches cachées :

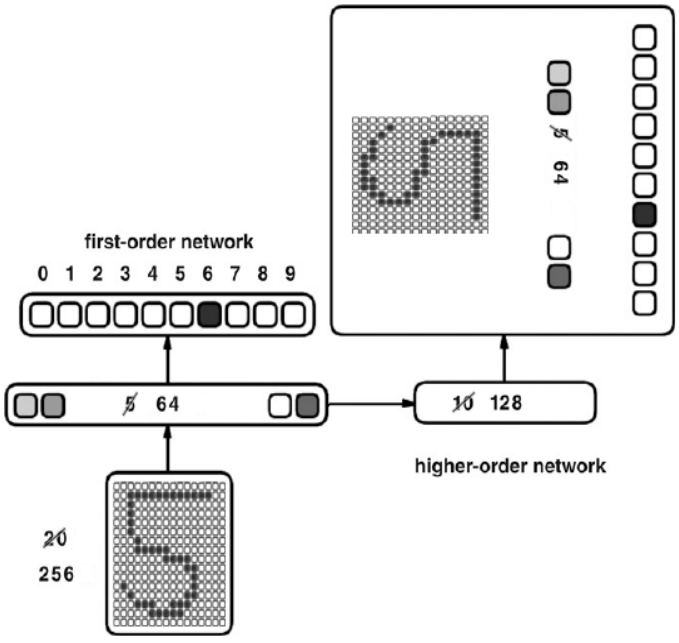


FIGURE 4 – Architecture avec méta-représentation pour chiffres manuscrits

Les performances du second réseau se sont alors écroulées :

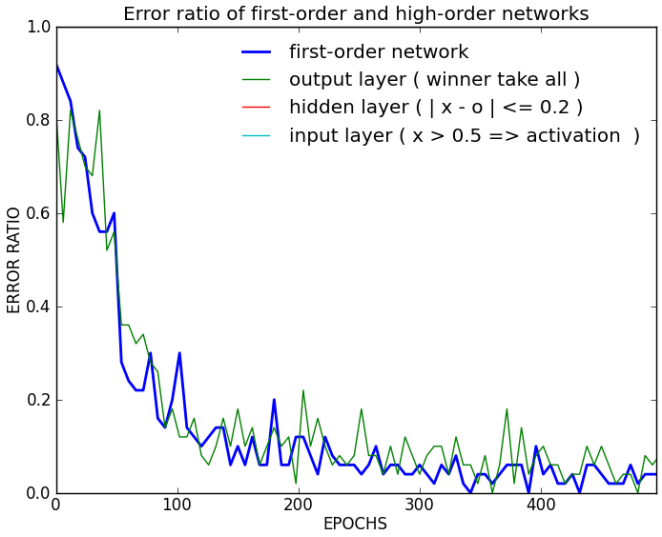


FIGURE 5 – Erreur de classification de l'architecture sur des chiffres manuscrits. L'erreur du second réseau est toujours divisé en 3.

Il n'est alors plus capable que de reproduire la couche de sortie. Petit bémol tout de même, le critère de classification pour la couche caché est peut-être trop rigide, il suffit qu'un neurone diffère de 0.2 ( de la valeur voulue ) pour considérer que c'est un échec.

## 2.4 Représentations

Enfin, nous avons remarqué qu'en bloquant l'apprentissage entre la couche cachée et les entrées du premier réseau, puis en changeant de tâche, le réseau était capable de réapprendre la nouvelle tâche, ce qui prouve bien la présence d'une représentation des entrées dans la couche cachée.

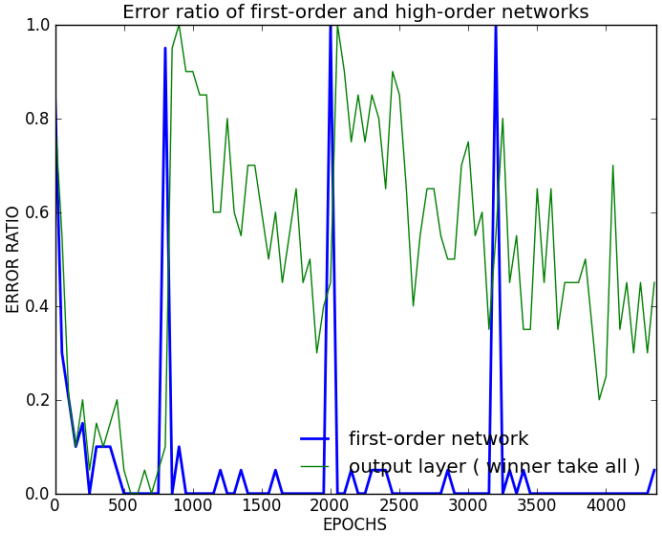


FIGURE 6 – Erreur de classification de l'architecture sur des chiffres manuscrits. Apprentissage bloqué à l'époque 800. Changement de tâche aux époques : 800, 2000, 3200

## 2.5 Remarque

Il faut cependant remarquer qu'un simple perceptron est suffisant pour réaliser la tâche du premier réseau (même sur les chiffres manuscrits), et donc, qu'il est possible que dans le cas d'un problème non linéairement séparable cette architecture soit invalidée.

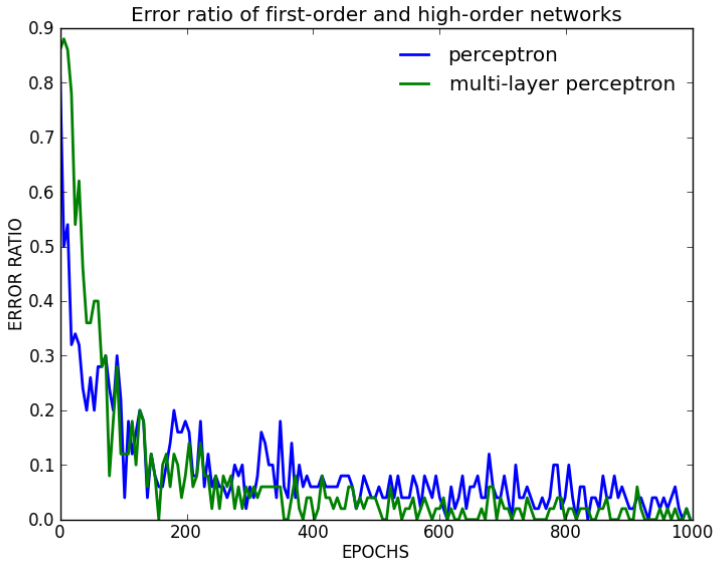


FIGURE 7 – Erreur de classification d'un perceptron et d'un perceptron multi-couches sur la base de chiffres manuscrits.

# 3 Parier sur le premier réseau

## 3.1 Les bases

Rappelons également la structure des réseaux et les résultats :

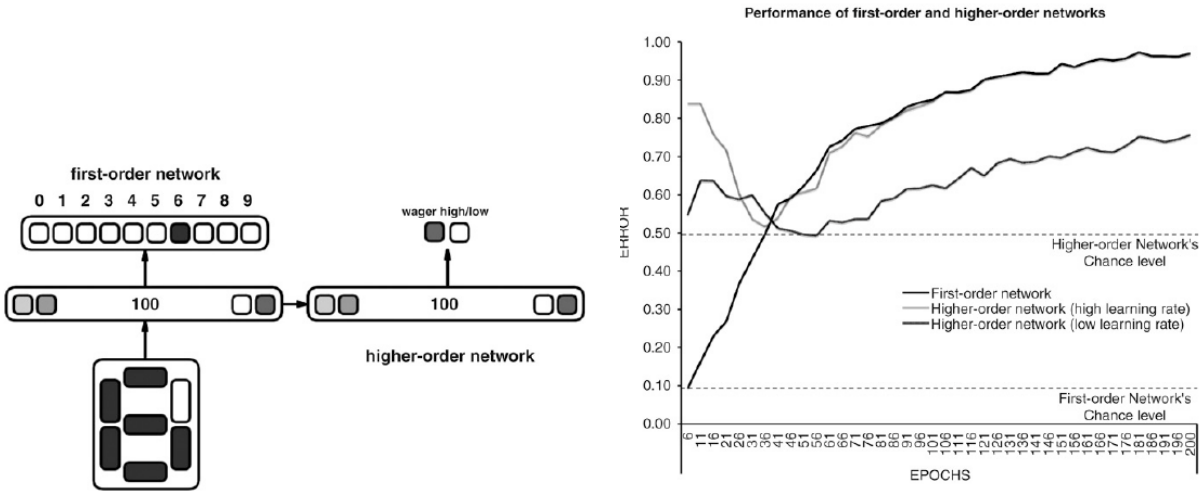


FIGURE 8 – [Cleeremans Alex(2007)] Architecture connexionniste avec paris

La première chose que nous avons faites à été d'améliorer les performances du second réseau en modifiant quelques paramètres (initialisation des poids sur [-1 ; -1], momentum à 0.5).

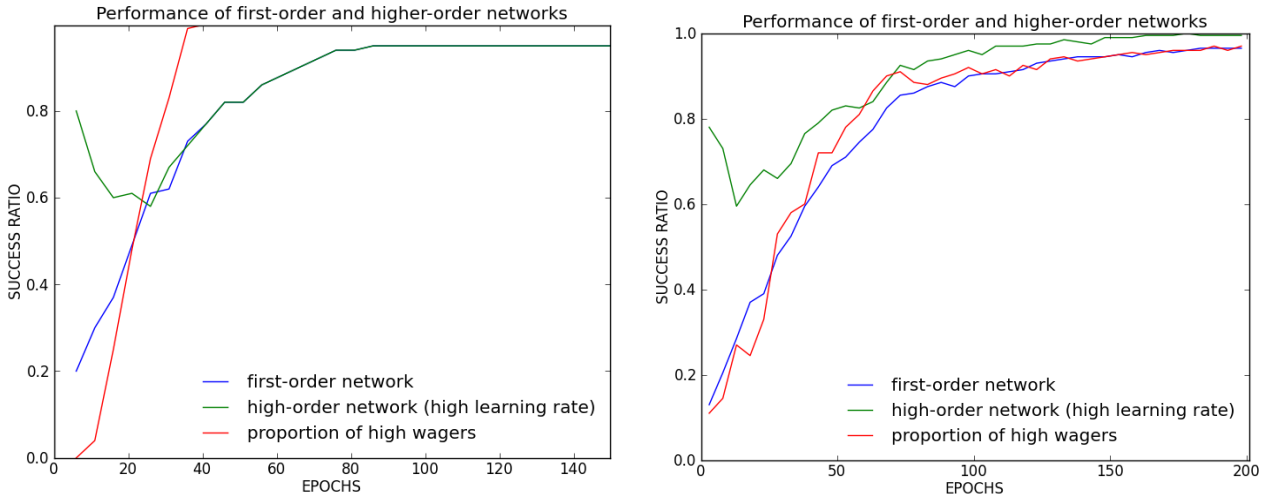


FIGURE 9 – Performance de l'architecture. À gauche la base, à droite la version améliorée. On ne considère plus le réseau avec apprentissage faible, mais on a ajouté le taux de paris hauts.

Contrairement à celui de l'article, il ne se contentera plus simplement de parier haut à chaque coups (après 40 époques). Il aura une longueur d'avance sur le premier réseau sur toute la durée de l'apprentissage. On pourra donc tirer profit de cette avance.

## 3.2 Feedbacks

À partir de cette différence de performances, nous avons imaginé plusieurs architectures, qui améliorent plus ou moins les performances de reconnaissance du réseau sur des chiffres manuscrits :

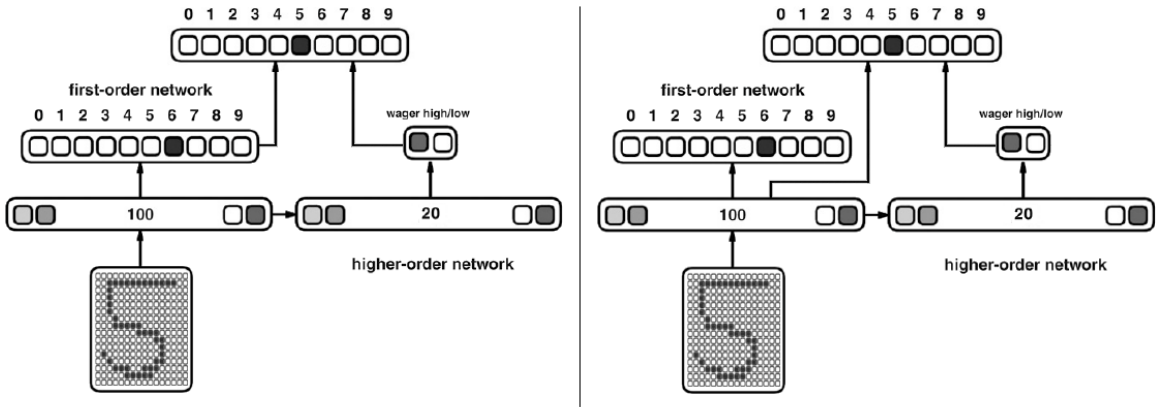


FIGURE 10 – Architecture avec 3$^{\text{ième}}$ réseau

Dans ces 2 architectures, nous nous contentons de connecter un 3$^{\text{ième}}$ réseau qui doit tirer des conclusions à partir d'informations sur les 2 premiers.

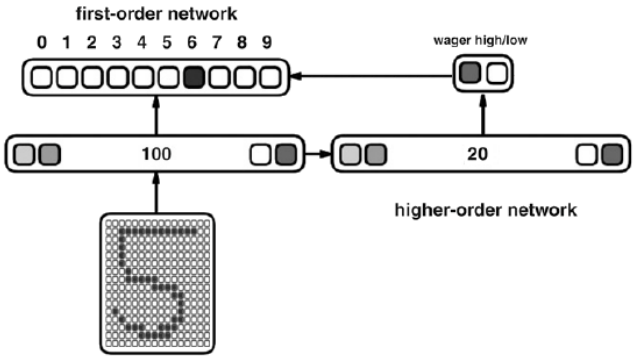


FIGURE 11 – Architecture par fusion

Ici, nous mélangeons un apprentissage par descente de gradient (sur le premier et second réseau) et un apprentissage perceptron (entre les 2 couches de sorties).

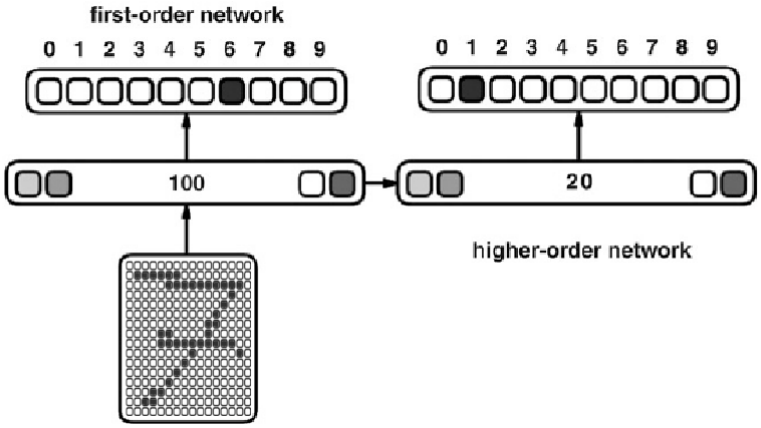


FIGURE 12 – Architecture par intuitions

Cette architecture est légèrement différente dans le sens où elle n'enregistre plus de pari mais l'indice du n$^{\text{ième}}$ neurone le plus actif contenant la bonne réponse. Exemple : le réseau supérieur sort 2 -> la réponse est le 3$^{\text{ième}}$ neurone le plus actif de la couche de sortie du premier réseau.

Nous avons aussi essayé quelques modèles où le réseau supérieur servait de superviseur à l'apprentissage du premier réseau. Par exemple, s'il parie haut, l'apprentissage du premier réseau sera faible, sinon il sera accentué.

# 4 La suite

Ce que nous continuons d'étudier :

– validation sur des expériences plus complexes (qui ne peuvent être résolue directement par un perceptron)

- relation entre la taille de la couche cachée du premier réseau et le taux de paris avantageux
- nouveau critère de classification pour la couche caché dans la première architecture sur les chiffres manuscrits
- approfondir les intérêts du second article [Pasquali Antoine(2010)]
- de nouvelles architectures axées méta-apprentissage où le réseau d'ordre supérieur contrôle le premier (taux d'apprentissage, momentum, entrées à approfondir, ...) telle une conscience

# Références

[Cleeremans Alex(2007)] Cleeremans Alex, P. A., Timmermans Bert (2007). Consciousness and metarepresentation : A computational sketch. *doi :10.1016/j.neunet.2007.09.011* .

[Pasquali Antoine(2010)] Pasquali Antoine, C. A., Timmermans Bert (2010). Know thyself : Metacognitive networks and measures of consciousness.

[Semeion Research Center(1994)] Semeion Research Center, o. S. o. C. (1994). Semeion handwritten digit data set. 1593 handwritten digits from around 80 persons.