CONSCIENCE ARTIFICIELLE?

Séance 14

Philo de l'esprit 2320

Jonathan Simon

H2020

PLAN

- 1) Introduction: IA, IGA et la conscience, IAS
- 2) Le débat actuel sur l'IA / IGA (les débats Marcus/LeCun, Marcus/Bengio)
- 3) La conscience artificielle (et sa relation avec l'IGA)
- 4) L'intelligence artificielle surhumaine et la singularité

INTRODUCTION

INTRODUCTION

- trois phases du débat sur les exigences d'une intelligence artificielle véritablement intelligente:

- IA classique / symbolique (Turing, Fodor + Pylyshyn)
- Connectionnisme de la première vague (Smolensky)
- Connectionnisme contemporain (LeCun, Bengio)

INTRODUCTION

- IA classique / symbolique (Turing, Fodor + Pylyshyn)
- modèle dominant d'IA: systèmes de symboles sensibles à la structure
- conception dominante de l'intelligence : passer le Test de Turing avec une base de données contenant des règles et des faits, plus capacité d'effectuer des calculs syntaxiques, raisonnement déductif
- Critiques dominantes : la syntaxe seule ne suffit pas pour la sémantique (Searle). Le respect des règles ne suffit pas pour l'expertise ou le savoir-faire (Dreyfus).

INTRODUCTION

- Connectionnisme de la première vague (Smolensky)
- modèle dominant d'IA: les réseaux de neurones vanille (perceptrons multi-couches)
- conception dominante de l'intelligence : la reconnaissance statistique des formes (approximation des fonctions)
- Critiques dominantes : les réseaux neuronaux vanille sont fragiles, ne peuvent pas soutenir des fonctions cognitives comme la systématicité de la pensée, sauf en mettant en œuvre l'AI classique (Fodor + Pylyshyn)

INTRODUCTION

- Connectionnisme contemporain (LeCun, Bengio)
- modèle dominant d'IA: les réseaux neuronaux profonds (convolutional nets, recurrent nets, attentional mechanisms)
- conception dominante de l'intelligence : la reconnaissance statistique des formes (approximation des fonctions)
- Critiques dominantes : les réseaux neuronaux profonds sont encore fragiles, malgré leurs réussites (Marcus)

INTRODUCTION

- Intelligence artificielle (étroite): des applications spécifiques à une tâche en utilisant l'apprentissage machine. Existe déjà.
- Intelligence générale artificielle (IGA): l'intelligence au niveau humain mis en œuvre par une machine (ce dont débattent Turing, Dreyfus, Fodor, Smolensky, Marcus, LeCun et Bengio)
- Conscience artificielle: la conscience phénoménale mise en œuvre par une machine (Avant l'IGA ? La même chose ? Impossible ? Dehaene et al.)
- Super-intelligence artificielle: l'intelligence à un niveau surhumain, mise en œuvre par une machine. La singularité ? La robopocalypse? (Chalmers)

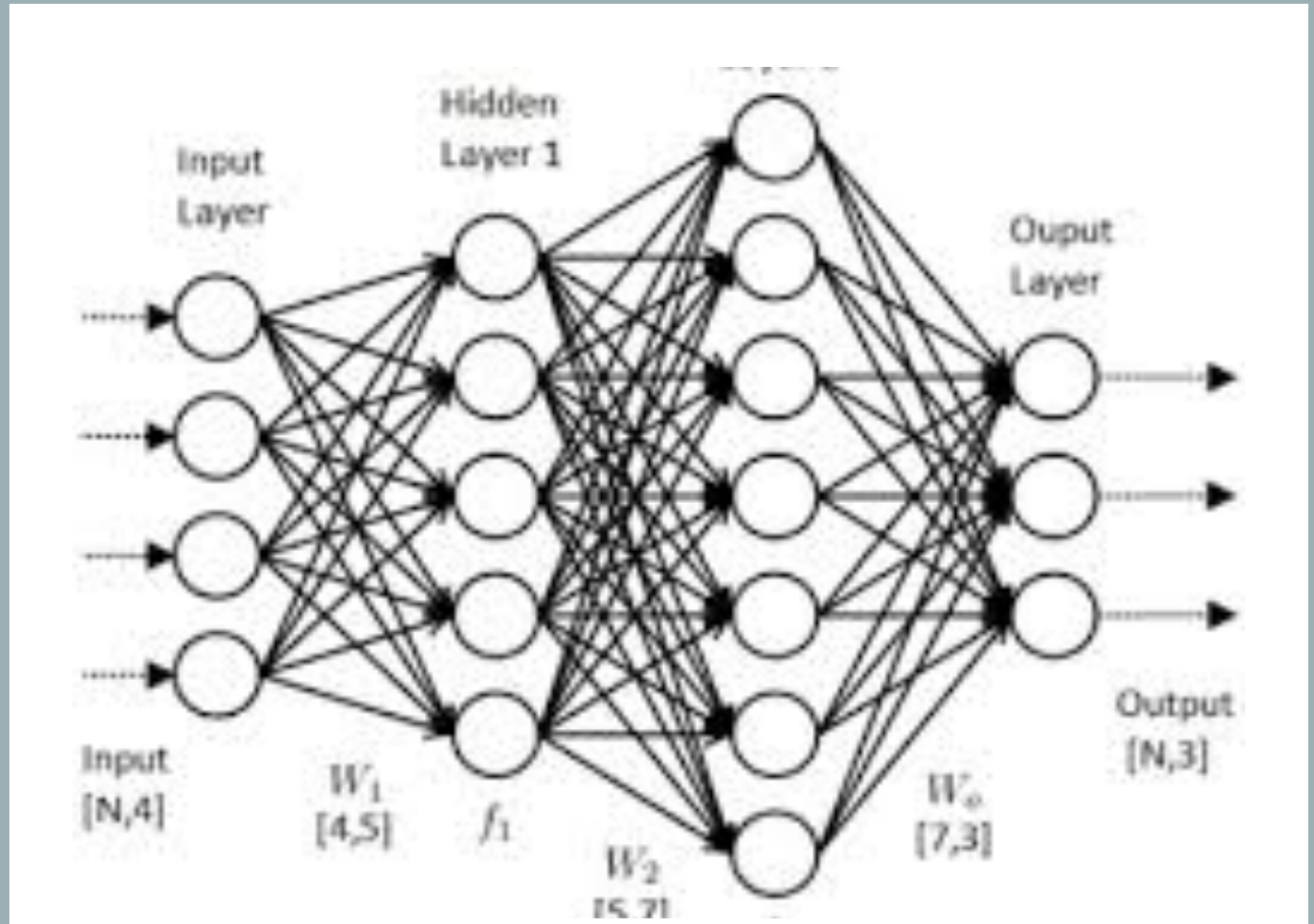
INTRODUCTION

- Conscience artificielle: la conscience phénoménale mise en œuvre par une machine
- Potentiellement indépendant de la question de l'intelligence humaine
- Rappelez-vous l'argument de Klein et Barron selon lequel les abeilles pourraient être conscientes
- Nous verrons : les tentatives d'atteindre l'IGA convergent de plus en plus avec les tentatives de conscience artificielle

LE DÉBAT ACTUEL SUR L'IA / IGA (LES DÉBATS MARCUS/LECUN, MARCUS/BENGIO)

LE DÉBAT ACTUEL SUR L'IA

- Progrès en matière d'IA depuis Smolensky :
- Dans les années 80 :
- Perceptrons multicouches (réseaux de neurones « vanille »)



LE DÉBAT ACTUEL SUR L'IA

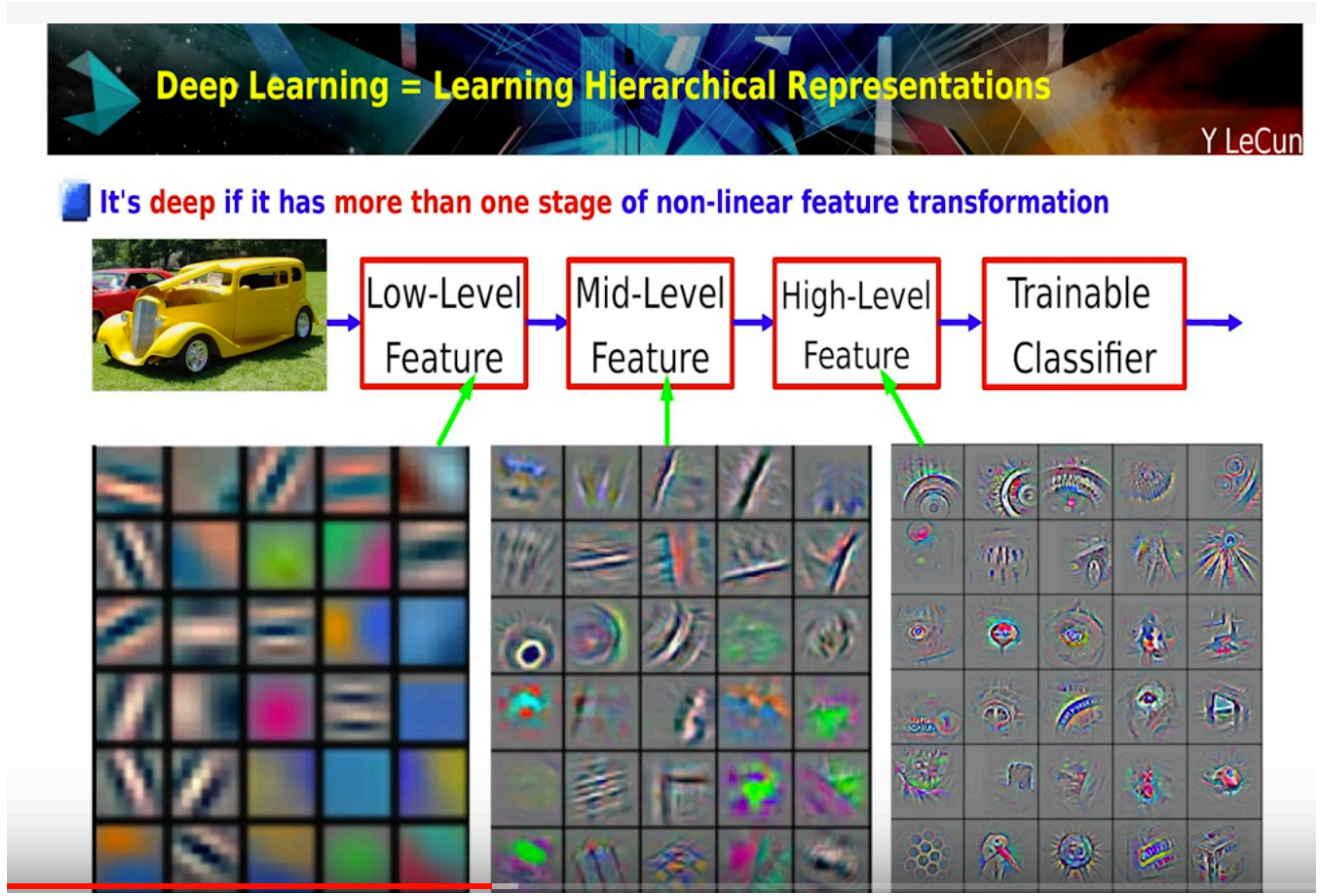
LE DÉBAT ACTUEL SUR L'IA

- Perceptrons multicouches (réseaux de neurones « vanille »)
- Chaque neurone d'un niveau est connecté à chaque neurone du niveau suivant, uniquement par feed-forward (les neurones ne font que transmettre l'information au niveau suivant)

LE DÉBAT ACTUEL SUR L'IA

- En principe très puissant : capable d'approcher toute fonction mathématique continue (avec une seule couche cachée), à condition d'avoir suffisamment de neurones
- En pratique, ce n'est pas efficace, sauf si vous pouvez ajouter plusieurs couches (plus de couches permettent de subdiviser un problème en sous-problèmes). Mais les couches multiples sont difficiles à former. Les progrès dans ce domaine ont donc été bloqués.
- (Une bonne revue interactive : http://neuralnetworksanddeeplearning.com/)

LE DÉBAT ACTUEL SUR L'IA



LE DÉBAT ACTUEL SUR L'IA

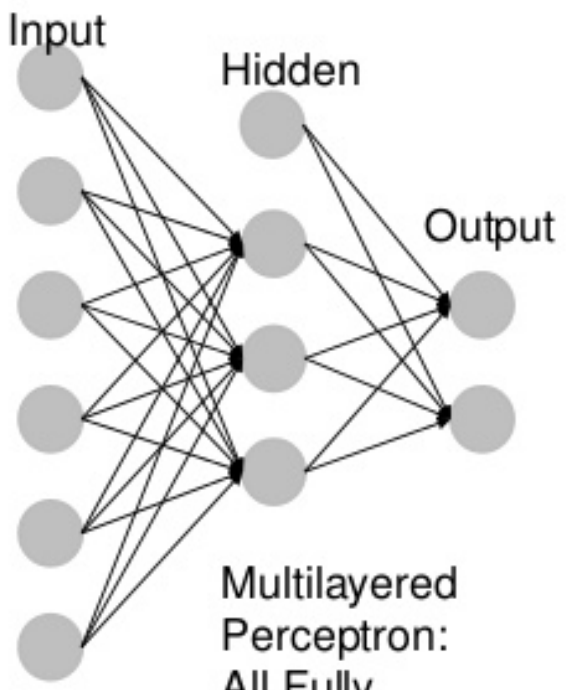
- Les progrès récents : une combinaison d'architectures plus sophistiquées, de plus de données et de plus de puissance de traitement
- Ensemble, ils ont permis aux chercheurs de former avec succès des modèles comportant les couches supplémentaires nécessaires pour des tâches plus complexes

LE DÉBAT ACTUEL SUR L'IA

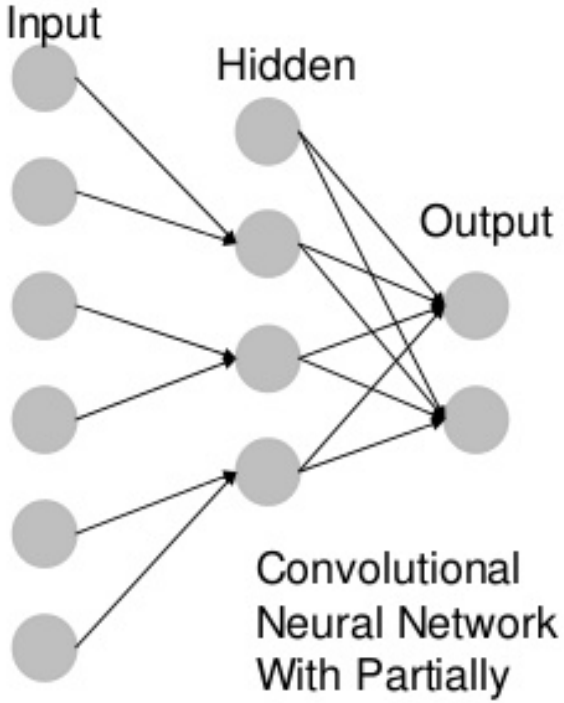
- Réseaux convolutifs (principalement pour le traitement des images)
- Réseaux récurrents (principalement pour le traitement des textes/séquences)

RÉSEAUX CONVOLUTIFS

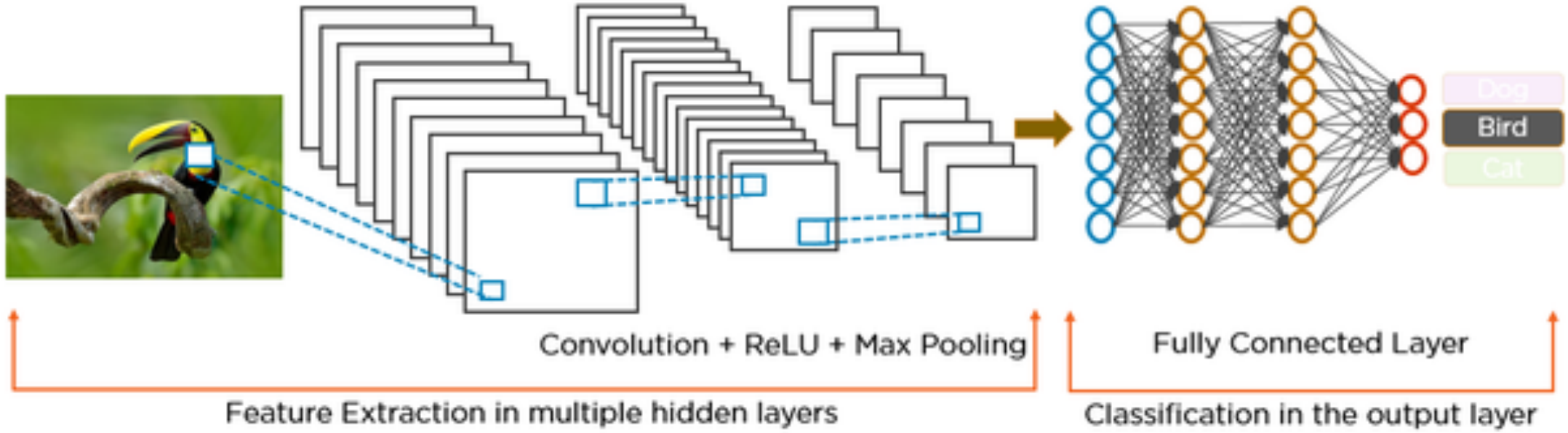
MLP VS ConvNet



Multilayered Perceptron: All Fully Connected Layers

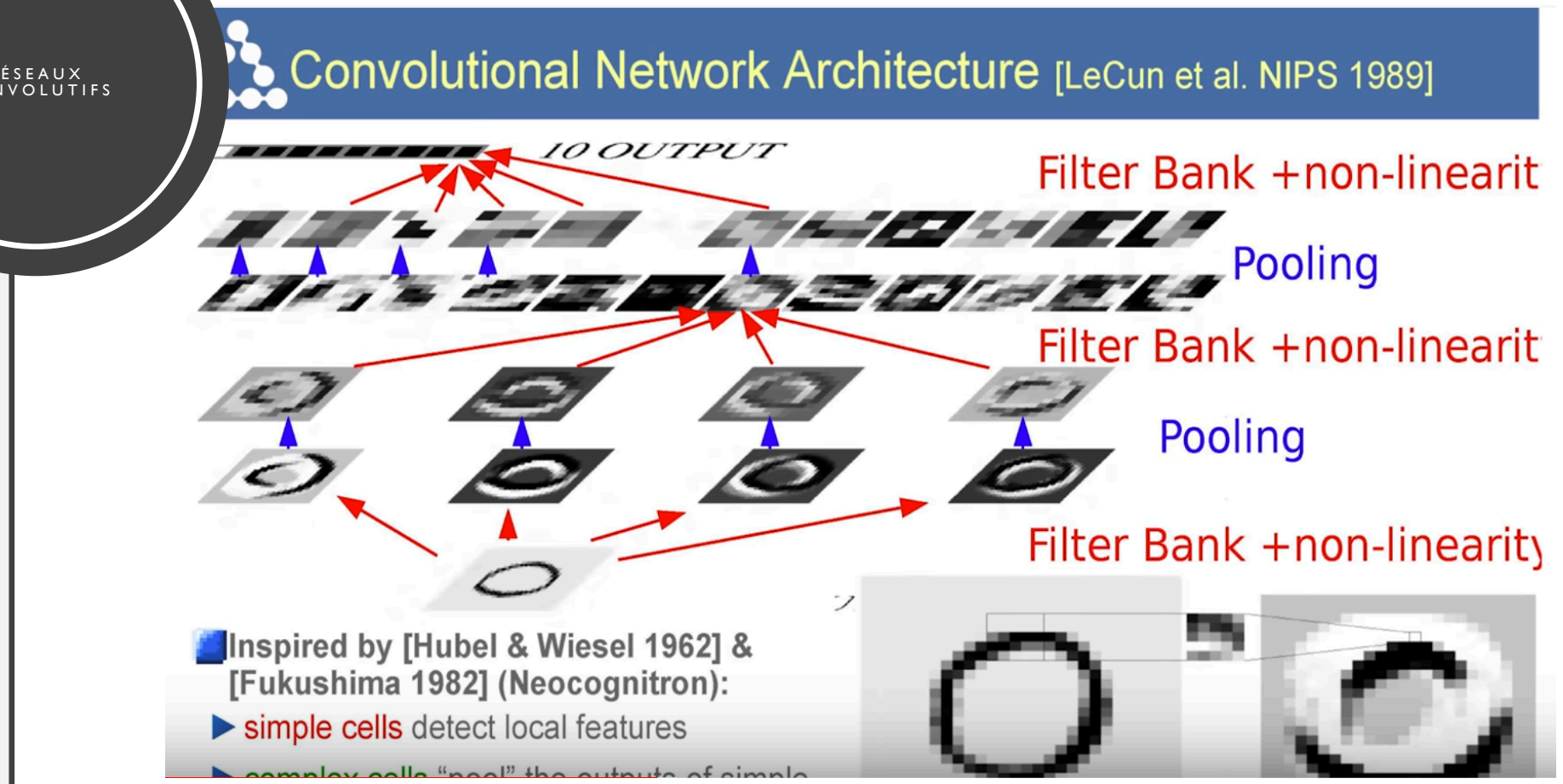


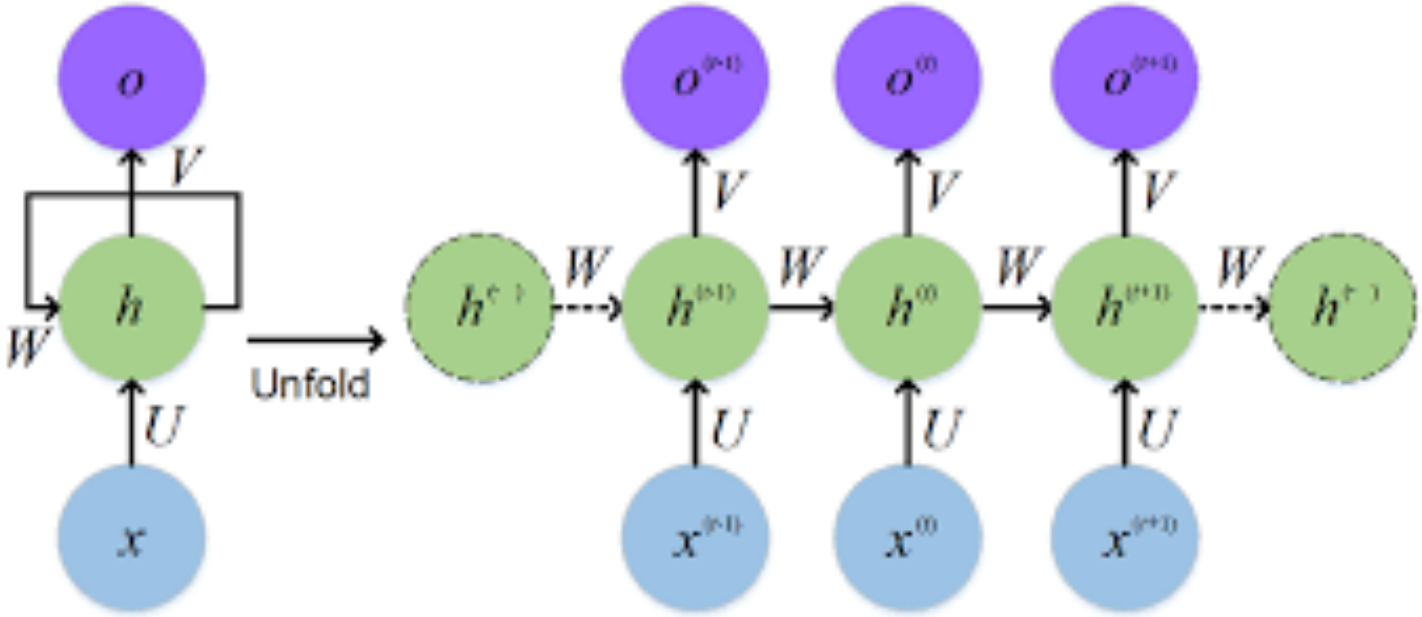
Convolutional Neural Network With Partially Connected Convolution Layer



RÉSEAUX CONVOLUTIFS

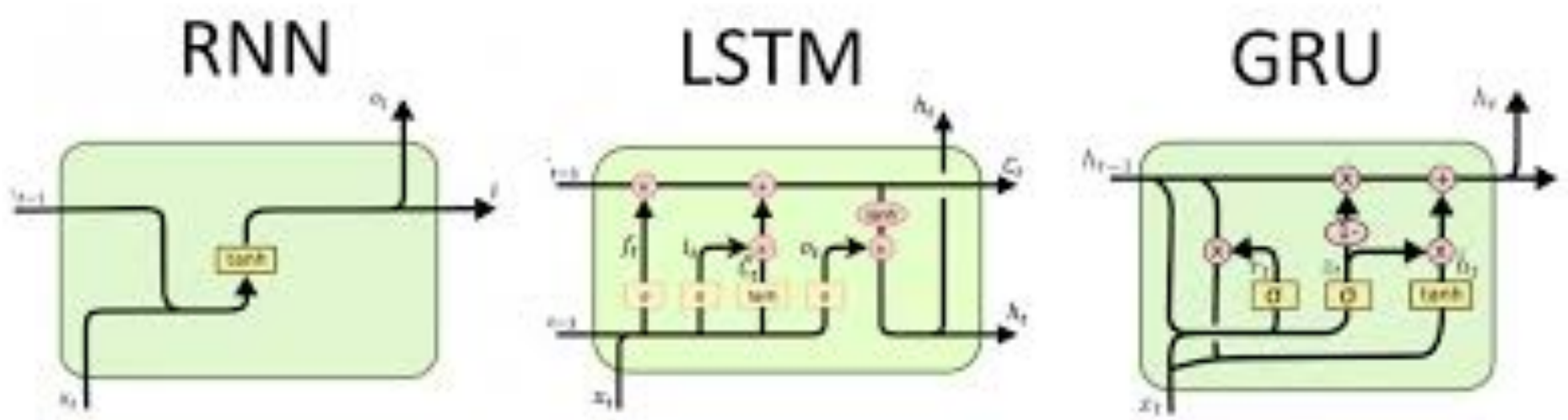
RÉSEAUX CONVOLUTIFS





RÉSEAUX RÉCURRENTS

RÉSEAUX RÉCURRENTS



LE DÉBAT ACTUEL SUR L'IA

- Ces architectures permettent aux algorithmes d'apprentissage machine de faire des choses que les IA classiques (systèmes experts suivant des règles) ne pourraient jamais faire.
- Mais il y a encore un débat sur la question de savoir si ce qu'ils font suffit pour l'intelligence générale.
- Des personnes comme Gary Marcus plaident en faveur des systèmes hybrides - des systèmes qui combinent les fonctionnalités des réseaux neuronaux avec les avantages des systèmes experts manipulant des symboles.

LE DÉBAT ACTUEL SUR L'IA

- Malgré leurs progrès, les IA contemporaines ont encore des domaines de fonctionnalité très limités et étroits : elles ont beaucoup de mal à se généraliser à de nouveaux domaines (hors distribution), elles semblent manquer de sens commun et elles ont besoin de millions d'exemples pour apprendre (contrairement aux enfants qui peuvent apprendre après un seul exemple).

LE DÉBAT ACTUEL SUR L'IA

- Exemple simple du problème de généralisation de la distribution : Gary Marcus a formé un perceptron multicouche à la "fonction d'identité" : 02 = 02, 04 = 04. Mais il n'a formé le système que sur les nombres pairs. Il fonctionnait bien sur les autres nombres pairs, mais pas sur les nombres impairs !

LE DÉBAT ACTUEL SUR L'IA

- (dans ce cas, le problème est facile à résoudre : il suffit d'inclure les nombres impairs dans le jeu de formation. En général, la tâche des "enseignants" consiste à préparer un "programme" représentatif du domaine (comme dans l'échantillonnage statistique). Mais plus on s'approche des domaines de la vie réelle, plus cela est difficile à faire...)

LE DÉBAT ACTUEL SUR L'IA

- La question est de savoir ce qu'il faut faire à ce sujet.

- Marcus : nous devons intégrer les réseaux neuronaux aux systèmes de symboles. Cela signifie que nous devons ajouter des capacités "algébriques" de manipulation des symboles ainsi que de nombreuses connaissances dans le domaine (espace, temps, causalité, objets, types vs instances, affordances...)

LE DÉBAT ACTUEL SUR L'IA

- Marcus soutient que les réseaux convolutionnels (de LeCun) le font déjà : cette architecture intègre une compréhension de la structure spatiale (dans les premières couches, les neurones proches ne réagissent qu'aux pixels proches, et dans les couches ultérieures, les neurones réagissent aux motifs de manière "invariante à la traduction » (translation invariance) - même motif quel que soit l'endroit où il se trouve sur la page)

LE DÉBAT ACTUEL SUR L'IA

- Mais notez la différence entre un réseau convolutif et un système de règles qui fait des déductions à partir d'axiomes comme "L'espace est invariant de traduction".

- Intuitivement : la leçon que nous donnent nos professeurs d'écriture créative - "montrez-le, ne le dites pas". Un système de règles le dit, tandis qu'un réseau de convolutions le montre...

LE DÉBAT ACTUEL SUR L'IA

- Bien sûr, cela peut être plus facile dans le cas de la structure spatiale, puisque nous pouvons permettre à un réseau artificiel d'utiliser la structure spatiale pour la représenter, tout comme un cerveau, l'intégrant ainsi dans l'architecture. Il n'est pas évident de savoir comment répéter cette astuce pour d'autres domaines de connaissance, par exemple, la connaissance sociale, la connaissance de la physique. Pour la connaissance logique, la structure syntaxique semble être l'architecture appropriée....

LE DÉBAT ACTUEL SUR L'IA

- En fin de compte, le débat de Marcus avec LeCun et Bengio porte sur la question de savoir si la compréhension nécessaire du domaine et du bon sens peut être obtenue en trouvant le bon type d'architecture d'apprentissage relativement générale, ou si nous devrons mettre en place de nombreuses compréhensions spécifiques basées sur des règles

LE DÉBAT ACTUEL SUR L'IA

- LeCun et Bengio : trouver la bonne architecture générale
- LeCun : modèles prédictifs / codage prédictif (formés avec un apprentissage non supervisé : montrer aux systèmes beaucoup de vidéos, leur faire prédire l'image suivante)
- Bengio : mécanismes attentionnels, autres fonctions associées à la conscience

LE DÉBAT ACTUEL SUR L'IA

- À noter pour nous : avec Turing et même avec Fodor, on aurait pu penser que l'intelligence/la cognition des machines serait assez facile, et ne nous amènerait pas vraiment au voisinage des questions plus difficiles sur la conscience.

- Mais il s'avère que l'effort de développement de l'IGA (ou même simplement d'une IA relativement robuste) nous rapproche assez de la recherche sur la conscience (codage prédictif, mécanismes attentionnels, etc...)

LA CONSCIENCE ARTIFICIELLE

LA CONSCIENCE ARTIFICIELLE

- Rappelez-vous séance 11 sur la conscience animale:

THÉORIES QUI EXCLUENT LA PLUPART DES ANIMAUX

Cartesianisme

Théories basées sur le langage

Théories d'ordre supérieur / métacognitives

LES THÉORIES QUI EXCLUENT DE NOMBREUX ANIMAUX

La théorie de l'espace de travail global (version forte)

DES THÉORIES MODÉRÉMENT INCLUSIVES:

TETG (version faible),

la théorie PANIC de Tye (théorie de l'espace de travail *local* ?),

La théorie de la boucle locale récurrente de Block,

La théorie de l'intégration du tronc cérébral supérieur de Merker, Klein et Barron.

DES THÉORIES TRÈS PERMISSIVES :

La théorie de l'information intégrée de Tononi

Panpsychisme

LA CONSCIENCE ARTIFICIELLE

- Les questions qui se posent ici sont effectivement les mêmes (bien qu'il y ait bien sûr la possibilité qu'aucune machine ne puisse être consciente, car la conscience est en quelque sorte essentiellement biologique. Searle ? Peut-être s'agit-il de la vie ? Peut-être du carbone ?)

LA CONSCIENCE ARTIFICIELLE

- En supposant que certaines machines puissent un jour être conscientes, on peut se demander : à quelle distance cela se trouve-t-il ? Les réseaux convolutifs sont-ils déjà conscients, ayant des expériences visuelles des images qu'ils peuvent correctement classer ? Les réseaux récurrents sont-ils déjà conscients, ayant des expériences de compréhension des séquences de langage qu'ils traitent ?

LA CONSCIENCE ARTIFICIELLE

- Probablement pas. Mais procédons théorie par théorie.

DES THÉORIES TRÈS PERMISSIVES :

La théorie de l'information intégrée de Tononi

Panpsychisme

LA CONSCIENCE ARTIFICIELLE

- Compte tenu du panpsychisme générique, bien sûr. Si les pierres peuvent être conscientes, pourquoi pas votre iPhone ? Compte tenu de la théorie de l'information intégrée, cela dépend de la mise en œuvre. Un modèle véritablement distribué pourrait être conscient. Mais alors, selon l'IIT, un modèle véritablement distribué de lecteur de CD serait également conscient.

DES THÉORIES MODÉRÉMENT INCLUSIVES:

TETG (version faible),

la théorie PANIC de Tye (théorie de l'espace de travail *local* ?),

La théorie de la boucle locale récurrente de Block,

La théorie de l'intégration du tronc cérébral supérieur de Merker, Klein et Barron.

LA CONSCIENCE ARTIFICIELLE

- Presque toutes ces théories exigent un certain degré d'intégration entre les systèmes, par exemple entre la perception et l'action. Ou bien elles nécessitent beaucoup de traitements dynamiques récurrents. Ou les deux. Les systèmes contemporains ne semblent pas avoir ce type d'intégration ou de récurrence (les "réseaux récurrents" ne présentent qu'une forme de récurrence très limitée, pas aussi dynamique ou soutenue que les boucles récurrentes du traitement perceptif dont parle Block)

LES THÉORIES QUI EXCLUENT DE NOMBREUX ANIMAUX

La théorie de l'espace de travail global (version forte)

LA CONSCIENCE ARTIFICIELLE

- C'est le C1 de Dehaene.
- Dehaene soutient que:
- 1) Les réseaux neuronaux artificiels existants ont la fonctionnalité des réseaux inconscients dans le cerveau humain (C0),
- 2) mais qu'ils n'ont pas encore de C1.
- Ils ne sont donc pas encore conscients.
- Cependant, il n'y a en principe aucun obstacle à leur donner cette fonctionnalité (Bengio et d'autres y travaillent)

LA CONSCIENCE ARTIFICIELLE

- Ad 1):
- Il existe de nombreuses données empiriques montrant comment un traitement perceptuel très sophistiqué peut ne pas être conscient, par exemple, la vue aveugle. la rivalité binoculaire. le clignement d'attention.

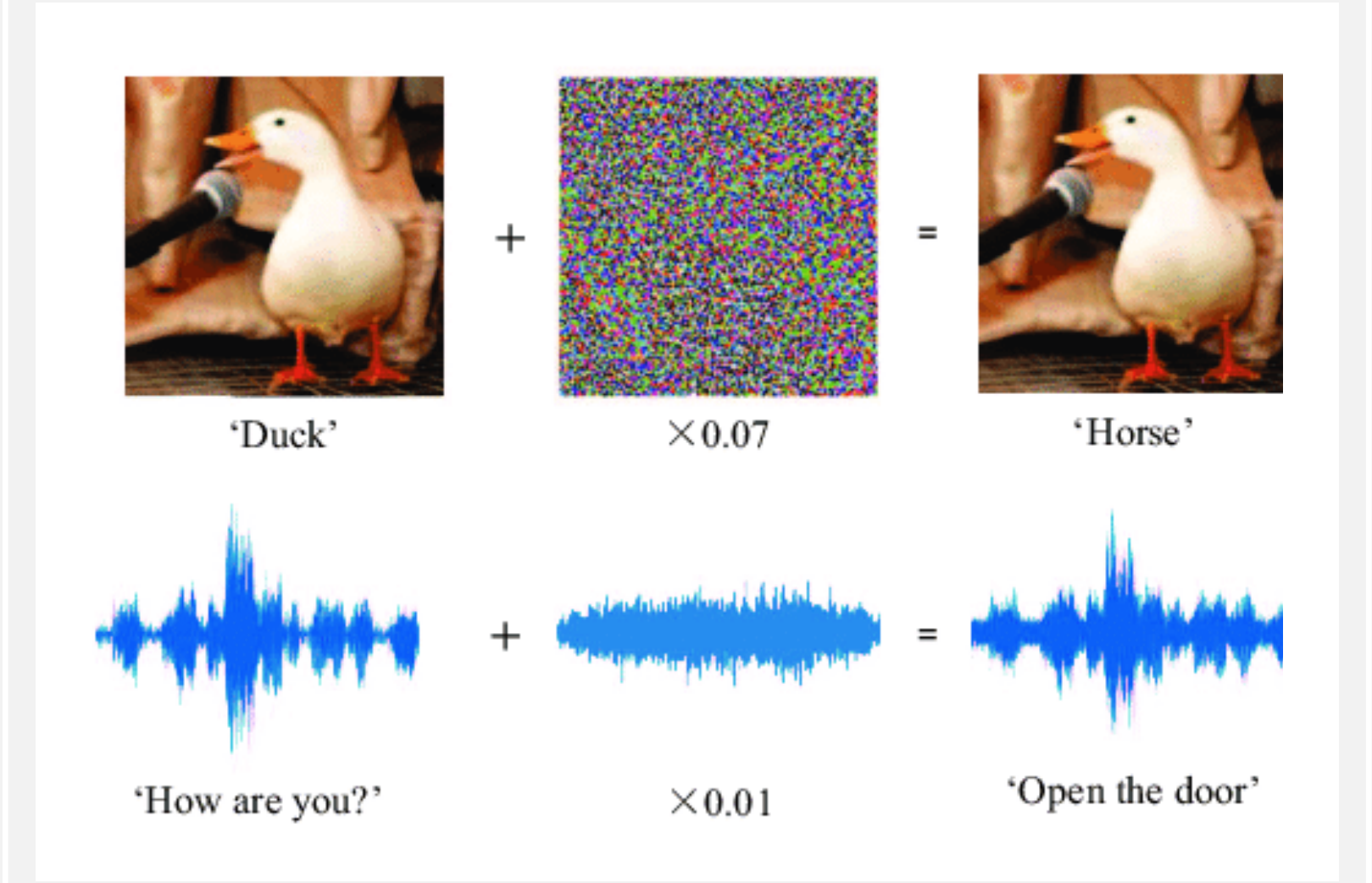
LA CONSCIENCE ARTIFICIELLE

- Ad 1):
- On peut se demander : que dirait Block à ce sujet ? Mais même Block accepte que certains traitements perceptuels sophistiqués ne franchissent pas le seuil de la conscience..

LA CONSCIENCE ARTIFICIELLE

- Ad 1):
- Une intuition ici (que les machines mettant en œuvre les réseaux convolutionnels actuels ne sont pas conscientes des images qu'elles traitent) : les exemples antagonistes. [adversarial examples]
- les entrées d'un réseau de neurones qu'il ne peut pas classer correctement

LA CONSCIENCE ARTIFICIELLE



LA CONSCIENCE ARTIFICIELLE

ils sont une caractéristique commune, et peut-être inévitable, des réseaux neuronaux. En général, lorsqu'un réseau parvient à classer une image, il existe une autre image qui ne diffère que légèrement (de sorte que vous ou moi ne pouvons même pas voir la différence), mais que le réseau classe mal de manière apparemment aléatoire

LA CONSCIENCE ARTIFICIELLE

Il y a des recherches en cours sur les véritables origines du phénomène. Mais il se produit même avec des réseaux extrêmement précis, qui gagnent des concours, etc. (donc pas seulement une version standard du problème de généralisation de l'absence de distribution)

LA CONSCIENCE ARTIFICIELLE

L'une des questions est de savoir comment celles-ci sont liées aux types d'illusions perceptuelles auxquelles nous sommes vulnérables :

LA
CONSCIENCE
ARTIFICIELLE

LA CONSCIENCE ARTIFICIELLE

avec la robe, comme dans les exemples antagonistes, une très petite modification de quelque chose que nous voyons correctement donne quelque chose que nous voyons incorrectement (pour ceux qui voient la robe comme étant blanche et dorée, une très légère modification de l'image permet de la voir correctement comme étant bleue et noire) :

LA CONSCIENCE ARTIFICIELLE

C'est peut-être la même chose. Mais peut-être pas : cela montre peut-être que la perception consciente implique une sorte de filtre supplémentaire que les convnets n'ont pas actuellement...

LA CONSCIENCE ARTIFICIELLE

Ad 2):

Il est clair que les réseaux neuronaux actuels n'utilisent rien qui ressemble à un espace de travail global. Mais les mécanismes attentionnels (mentionnés par Bengio) sont un pas dans cette direction....

THÉORIES QUI EXCLUENT LA PLUPART DES ANIMAUX

Cartesianisme

Théories basées sur le langage

Théories d'ordre supérieur / métacognitives

LA CONSCIENCE ARTIFICIELLE

Compte tenu du cartésianisme, les machines ne sont probablement pas conscientes, mais qui sait ? Étant donné le dualisme des propriétés (ou dualisme des substances épiphénomènes), il est difficile de voir pourquoi pas. Cela dépend des lois qui lient le mental au physique !

LA CONSCIENCE ARTIFICIELLE

La théorie métacognitive d'ordre supérieur est ce que Dahaene appelle C2. Il considère que cela fait partie de la conscience (il ne précise pas s'il pense que la conscience naît avec l'un ou l'autre, ou seulement avec les deux).

LA CONSCIENCE ARTIFICIELLE

De manière limitée, les réseaux neuronaux possèdent déjà ces capacités. Il est trivial pour un ordinateur de pouvoir rendre compte de ses propres états internes (votre ordinateur peut vous dire combien de batterie il lui reste, etc.), et les réseaux récurrents peuvent déjà traiter le langage. À cet égard, les théories les plus exclusives concernant la conscience animale ne sont pas les plus exclusives concernant la conscience artificielle.

LA CONSCIENCE ARTIFICIELLE

Mais ce que nous voulons vraiment ici, ce sont des capacités métacognitives (et linguistiques) qui soient générales, intégrées à l'espace de travail global, etc. Pour cela, il faut un espace de travail global...

L'INTELLIGENCE ARTIFICIELLE SURHUMAINE ET LA SINGULARITÉ

LA SINGULARITÉ

- En supposant que nous arrivions à une IA de niveau humain dans un avenir proche, devrions-nous nous attendre à voir une IA surhumaine un peu plus tard ?

LA SINGULARITÉ

- Explosion de la vitesse (Moore's Law)
- Explosion de l'intelligence:

- Si nous pouvons réussir à créer un être plus intelligent que nous-mêmes, sûrement cet être plus intelligent que nous-mêmes peut réussir à créer un être plus intelligent que lui-même, et cet être peut sûrement...

LA SINGULARITÉ

- Si chaque étape se déroule deux fois plus vite que la précédente, alors il y aura un point de convergence vers l'infini... une singularité !

- Mais nous pouvons nous concentrer sur une question plus modeste : l'IA de niveau humain mènera-t-elle à une IA de niveau surhumain, quel que soit l'aspect que vous souhaitez mesurer (acuité perceptuelle, créativité, sagesse, etc...)

LA SINGULARITÉ

- Deux points.

- 1) Notez que cela dépend vraiment du chemin que nous prenons. Dans l'approche classique, tout dépend de ce que vous savez et de la rapidité avec laquelle vous pouvez faire des déductions. Il va donc de soi qu'avec une base de données plus grande et un processeur plus rapide que le niveau humain, vous disposerez de quantités plus importantes, comme l'intelligence, que le niveau humain.

LA SINGULARITÉ

- Mais si, comme le suggèrent les progrès actuels, parvenir à l'intelligence au niveau humain signifie mettre en œuvre l'ensemble des capacités générales qui sous-tendent la conscience, c'est moins évident.

LA SINGULARITÉ

- Il est assez facile d'imaginer des êtres qui nous ressemblent mais qui sont plus rapides ou qui ont une meilleure mémoire. Mais si, en fin de compte, le secret de l'intelligence est la capacité à s'intégrer à travers les modalités, à surveiller ses propres états, etc., il est possible qu'il n'y ait aucune différence de nature, mais seulement une différence de degré, entre l'intelligence et la superintelligence. Cela pourrait signifier que nous pouvons rester compétitifs aussi longtemps que nous pouvons trouver des moyens d'étendre nos mémoires, notre puissance de traitement, etc.

LA SINGULARITÉ

- 2) Le problème de l'alignement des IA est de s'assurer que les IA du niveau surhumain (et du niveau humain, et du niveau sous-humain) ont des objectifs qui restent alignés sur les nôtres. Qu'elles ne décident pas, par exemple, de nous transformer tous en trombones de papier.

LA SINGULARITÉ

- 2) Plus les IA surhumaines restent reconnaissables comme nous, plus cela ressemble au problème de s'assurer que nos pairs très intelligents ont des objectifs qui restent alignés sur les nôtres (très différents si la question est de savoir quelles sont les règles classiques dans lesquelles nous devons programmer, pour les instruire sur les actions à ne jamais entreprendre...)

CONCLUSION

CONCLUSION

- Au fur et à mesure que la recherche en IA progresse, elle devient de plus en plus intimement liée à la question de la conscience artificielle. Cette question porte à son tour sur tout ce que nous avons considéré en classe, toutes les différentes théories proposées et leurs différents pièges.