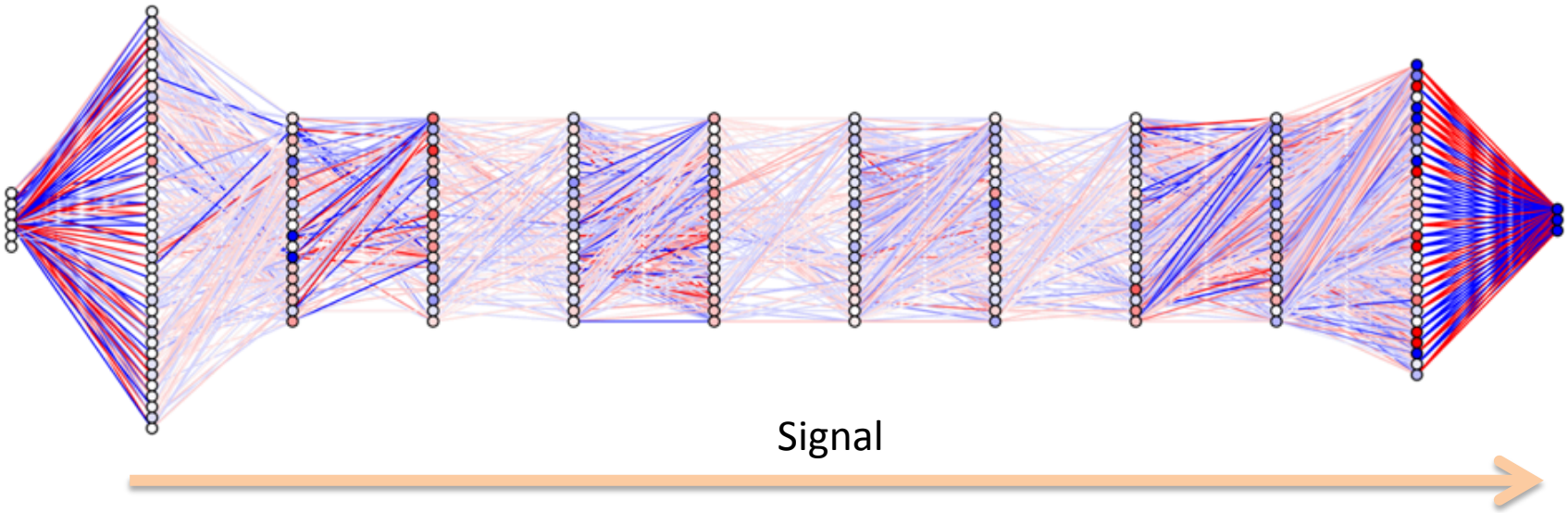
Plan

1. Pourquoi toute cette excitation ?
2. Grands types d'apprentissage
3. Apprentissage prédictif par réseaux de neurones
4. Quelles garanties ?
5. Le no-free-lunch theorem
6. Les réseaux de neurones profonds
7. Ce que l'on sait faire et les défis à relever

Ce que sont les réseaux de neurones **profonds**

Les « réseaux de neurones profonds »

- Des réseaux de neurones artificiels
  - à grand nombre de couches (parfois > qqs 100)
  - et **très grand nombre de paramètres** (qqs 10^7 – 10^{11} paramètres)

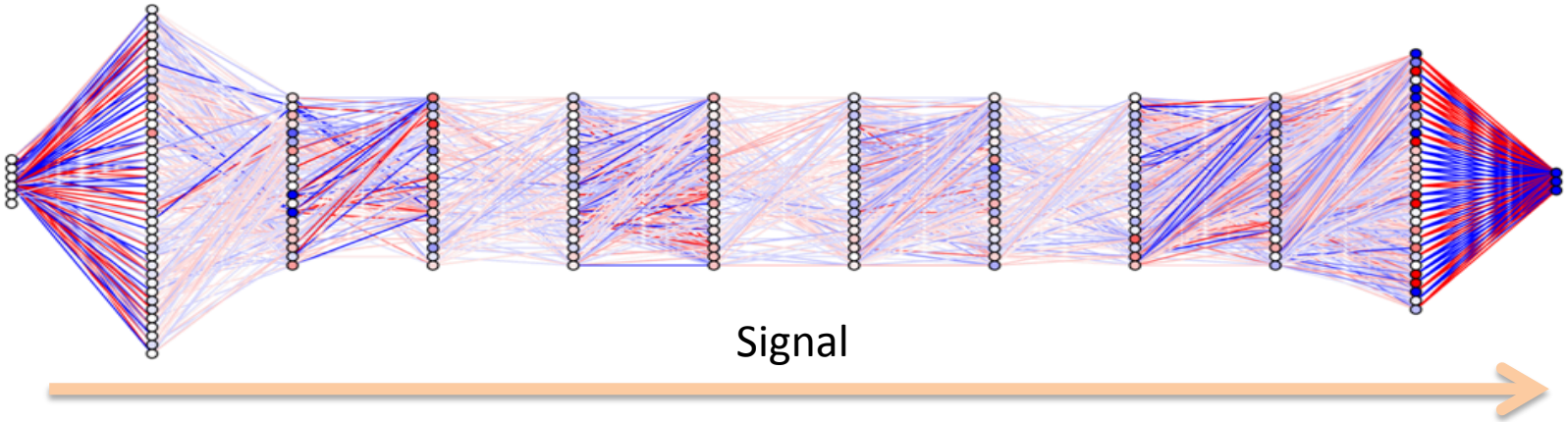


The SuperVision network

Image classification with deep convolutional neural networks

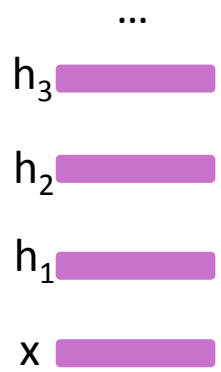
http://image-net.org/challenges/LSVRC/2012/supervision.pdf

- 7 hidden “weight” layers
- 650K neurons
- **60M** parameters
- 630M connections



Représentations profondes

- Idée
  - Apprendre des niveaux de représentation de **plus en plus abstraits** par mise en relation des niveaux inférieurs



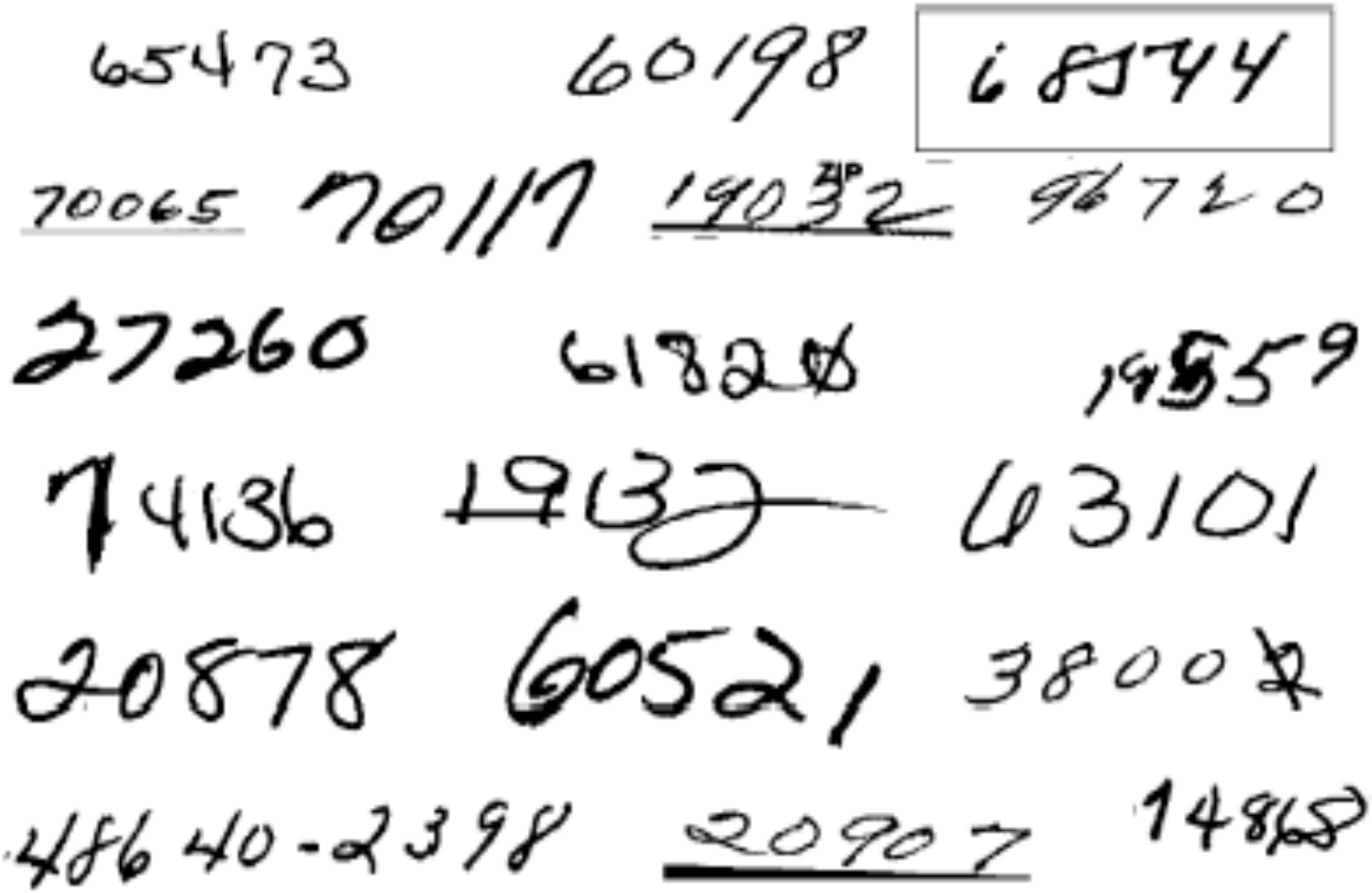
- Motivations
  - Permet potentiellement un **gain exponentiel dans la puissance expressive**
  - Le **cerveau humain** utilise des structures en couches successives
  - Les **connaissances** et **théories** humaines sont compositionnelles
  - **Marche très bien expérimentalement** sur certaines tâches

Les « réseaux de neurones profonds »

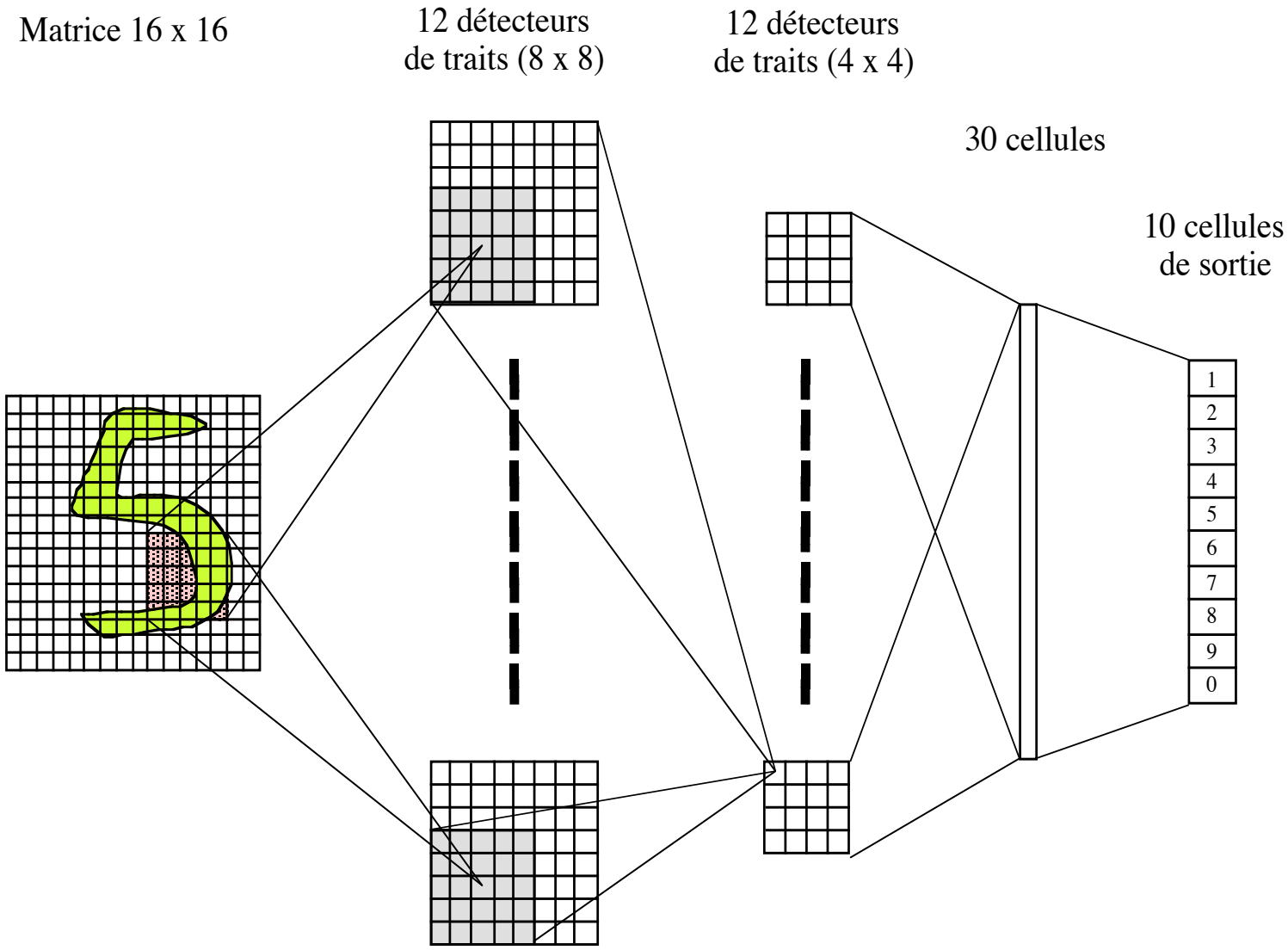
- Des réseaux de neurones artificiels

  1. à grand nombre de couches
  2. et **très grand nombre de paramètres**
  3. qui apprennent des **représentations hiérarchiques**
  4. et **décomposent les calculs**

La base de données



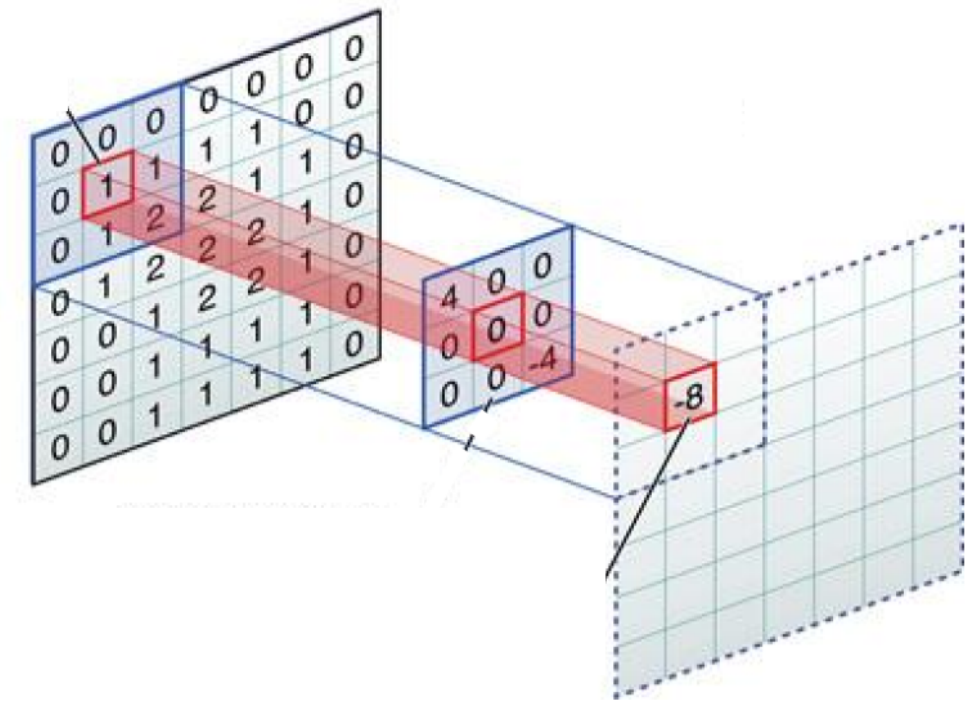
Réseaux à convolution : Application aux codes postaux



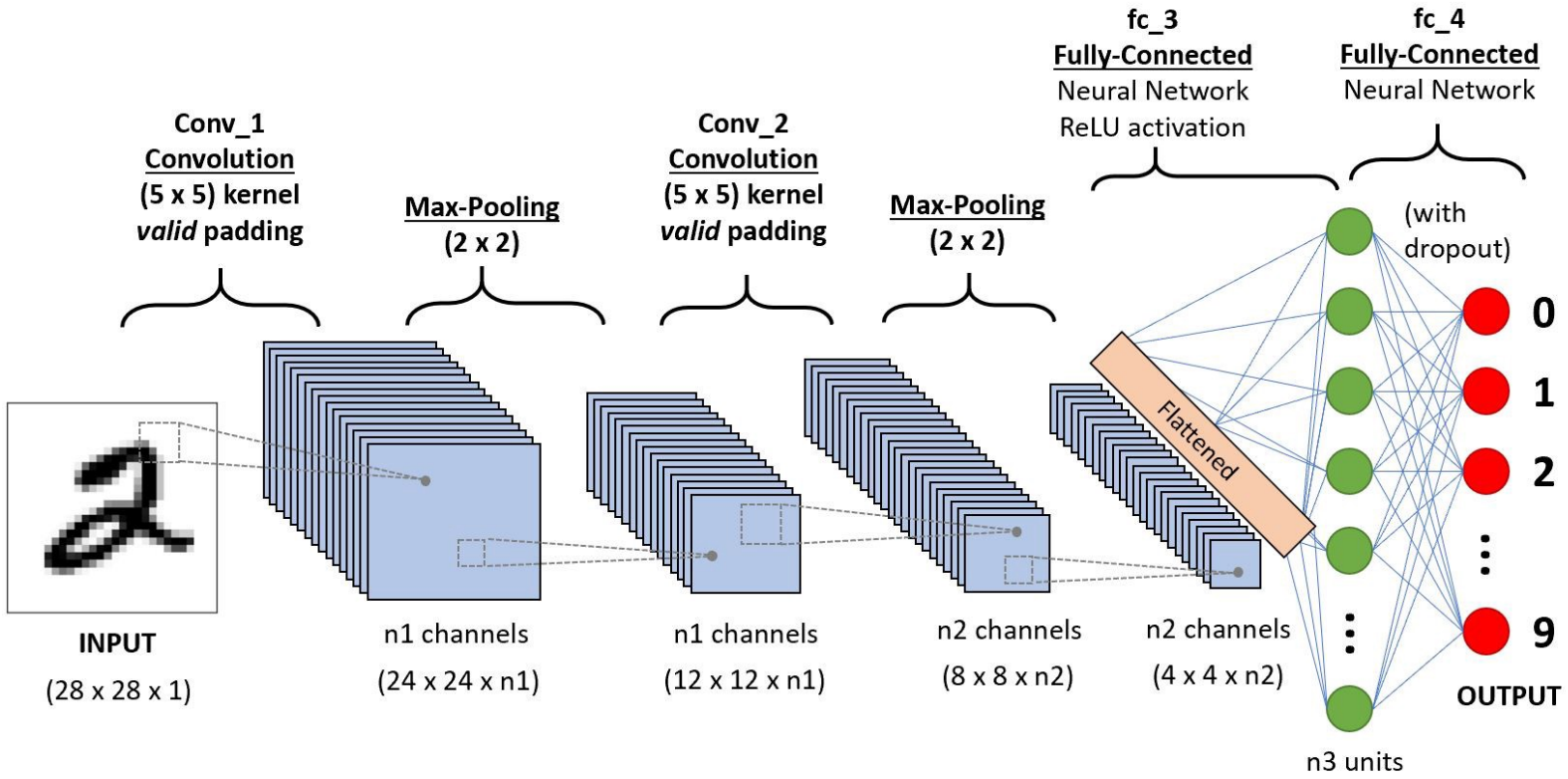
Convolutions

Same dimension between layers

(4x0) + (0x0) + (0x0) + (0x0) + (0x1) + (0x1) + (0x0) + (0x1)+ (-4x2) = -8

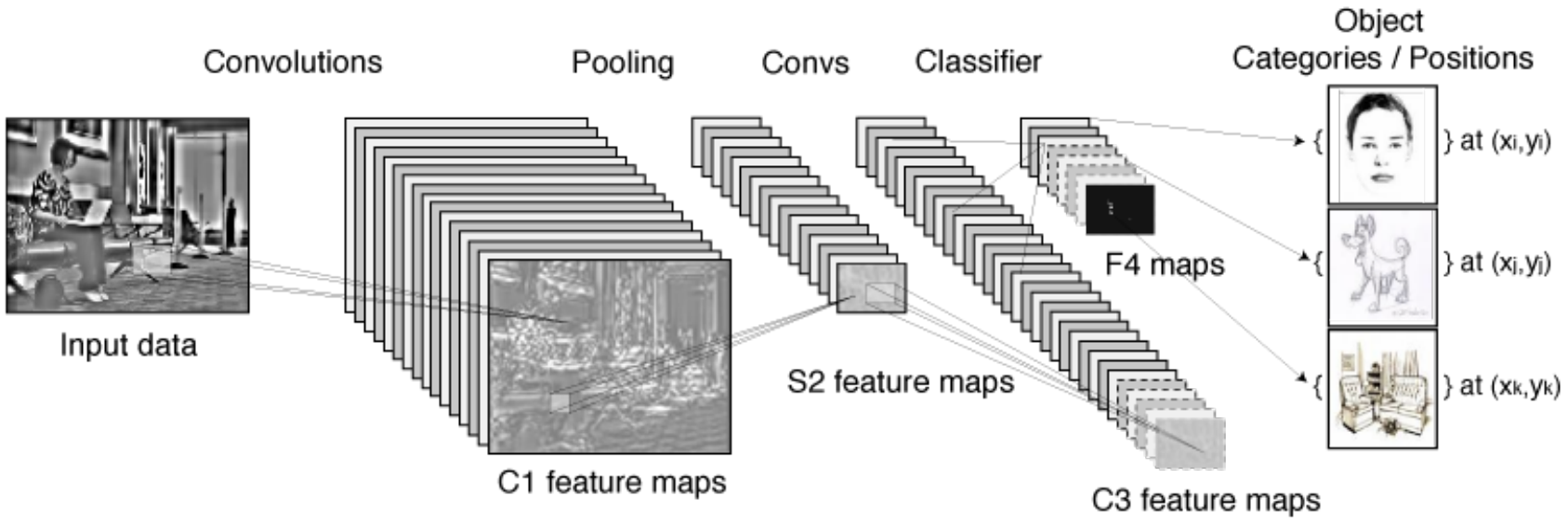


Réseaux de neurones convolutionnels



- ...

Réseaux de neurones convolutionnels (2° exemple)

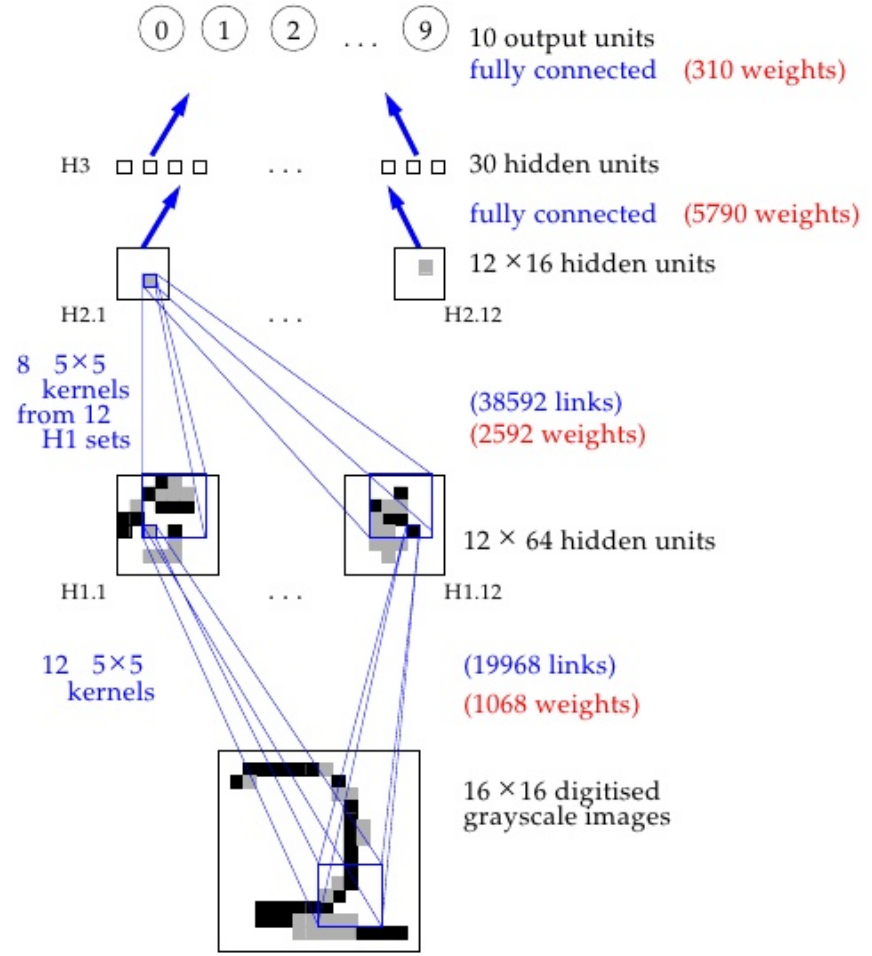


1) Le **pooling** consiste réduire la résolution des images filtrées, par exemple en moyennant les pixels contigus *(ex : 4 pixels deviennent 1 seul pixel qui contient une valeur moyenne)* - étape S2

2) **Convolution** : des filtres sont appris sur les nouvelles images - étape C3

3) **Classifieur** : les dernières couches sont entièrement connectées *(i.e. MLP)* et apprennent à prédire la classe à partir des filtres appris *(i.e. des descripteurs générés automatiquement)* - étape F4

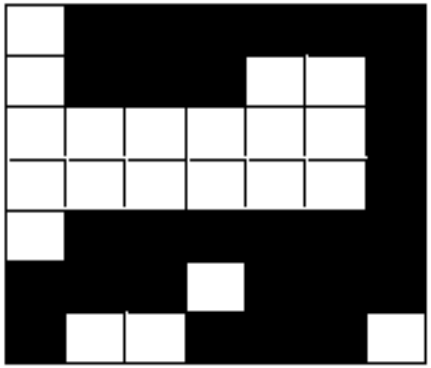
Réseaux à convolution : Application aux codes postaux



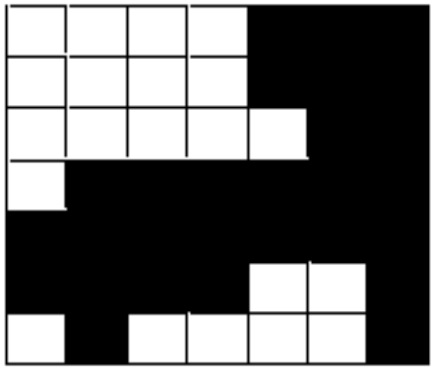
- Taken from http://image.slidesharecdn.com/bp2slides-090922011749-phpapp02/95/the-back-propagation-learning-algorithm-10-728.jpg?cb=1253582278

How to code the inputs

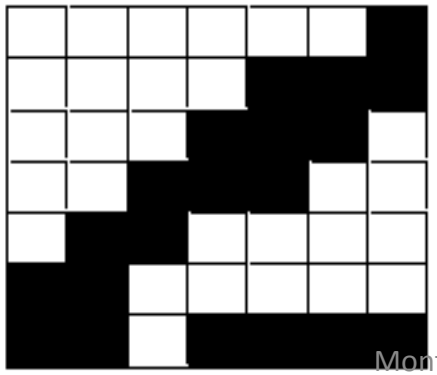
Learning is easy when we know what to look for



- Yes

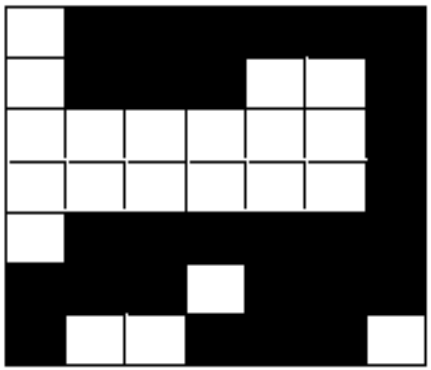


- Yes



- No

Inputs and prior knowledge



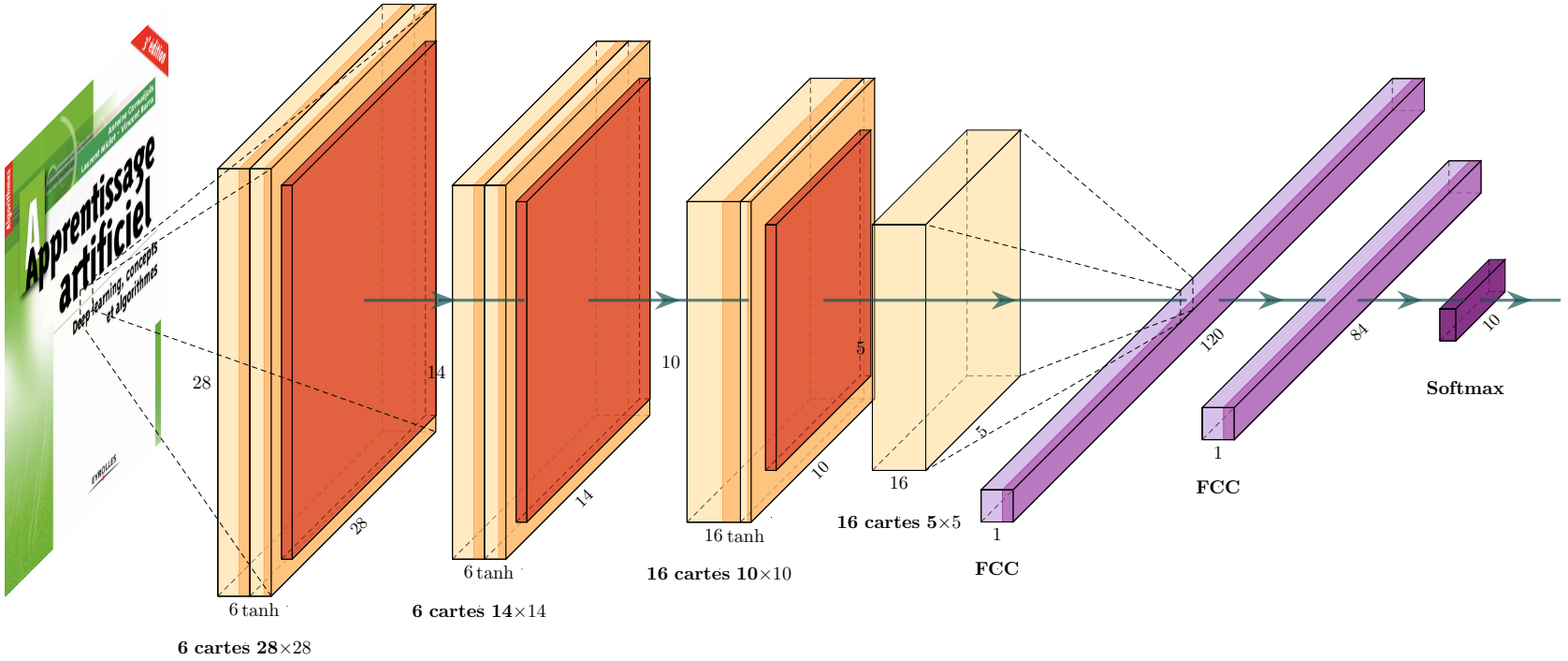
- Is it a pattern recognition task? A character recognition task? …
- *How to code the examples?*

0 1 1 1 1 1 1 0 1 1 1 0 0 1 0 0 0 0 0 0 1 0 0 0 0 0 0 1 0 1 1 1 1 1 1 1 1 1 0 1 1 1 1 0 0 1 1 1 0

- A right choice of representation can render the learning task trivial

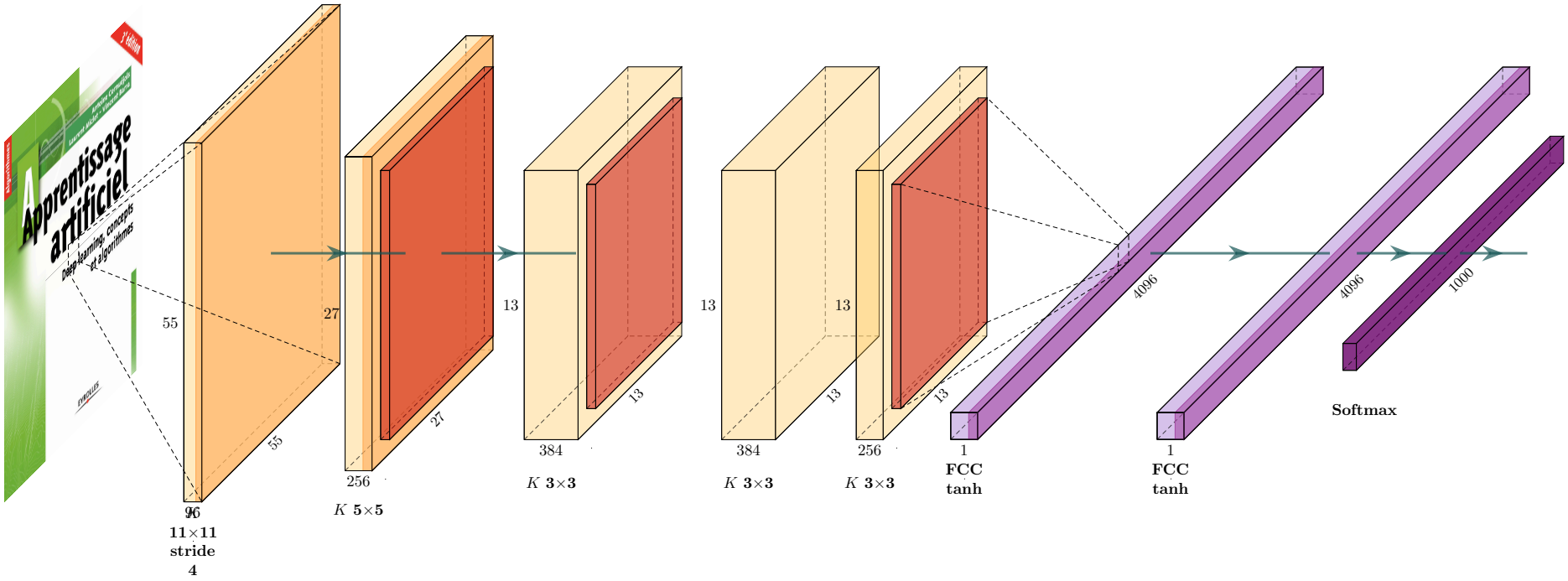
➠ But how can we know the right representation?

LeNet-5



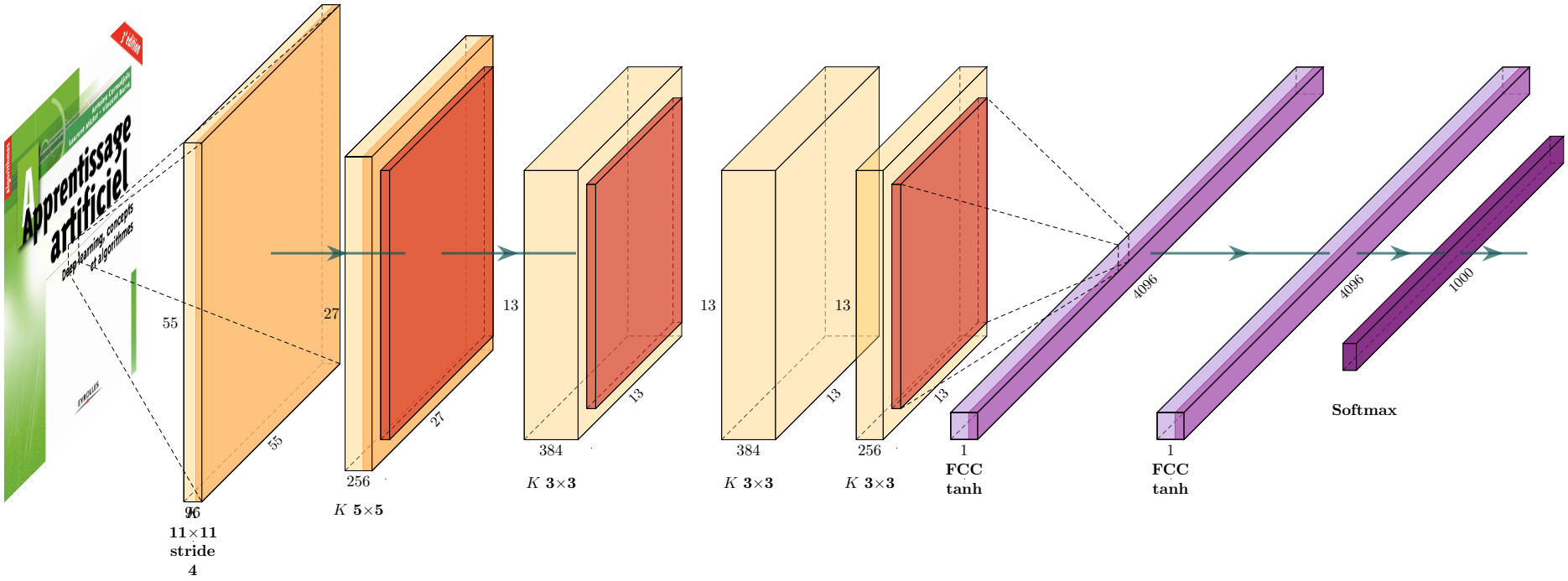
- Architecture du réseau LeNet-5.
  - Les couches de **convolution** et d'activation sont en orange clair.
  - Les couches d'**agrégation** en orange foncé
  - Les couches **complètement connectées** sont en violet

AlexNet



- Architecture du réseau AlexNet.
  - Les couches de **convolution** et d'activation sont en orange clair.
  - Les couches d'**agrégation** en orange foncé
  - Les couches **complètement connectées** sont en **violet**

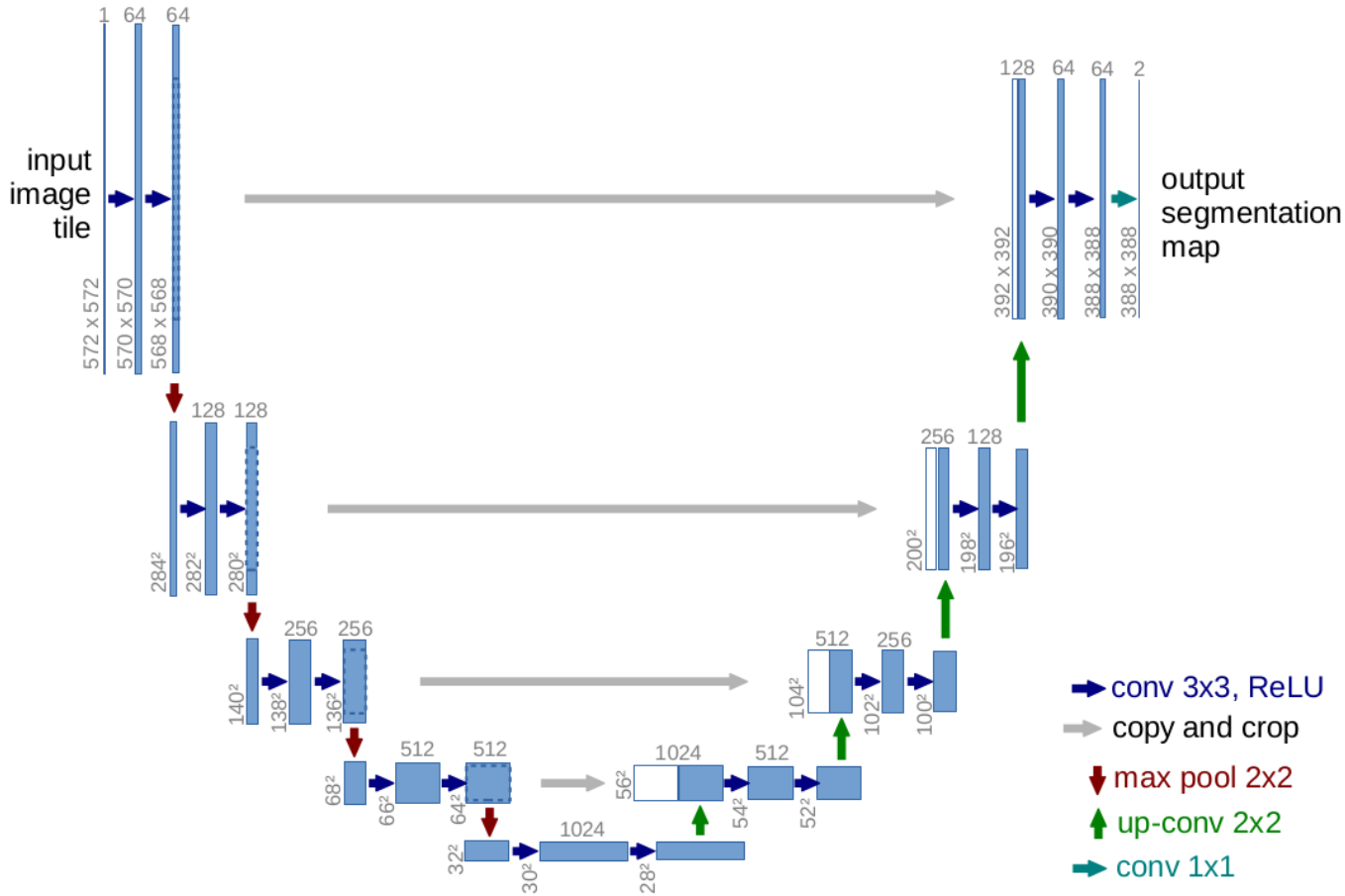
AlexNet



Si la profondeur du réseau reste faible, le **nombre de paramètres** est déjà important.

- En regardant uniquement la **première couche de convolution**, on constate que l'entrée est composée d'images 224 x 224 x 3, que les filtres de convolution sont de taille 11 et que le stride est de 4.
- Ainsi, la sortie de la couche de convolution est de taille 55 x 55 x 96 = 290 400 neurones, chacun ayant 11 x 11 x 3 = 363 poids et un biais. Cela implique, sur cette couche de convolution seulement, **105 705 600** paramètres à ajuster.

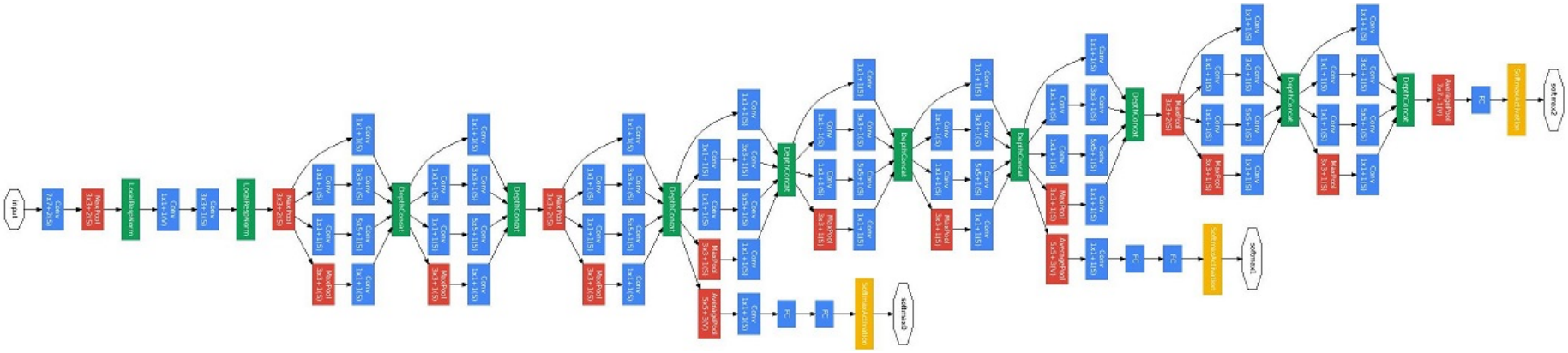
U-Net : pour la segmentation d'images



Cette architecture est intégralement convolutive et symétrique. Dans la première partie de l'architecture, la taille de la sortie diminue et le nombre de filtres augmente. La seconde partie inverse le processus avec une carte de prédictions en sortie.

GoogleNet

- Un **mécano** de réseaux de neurones



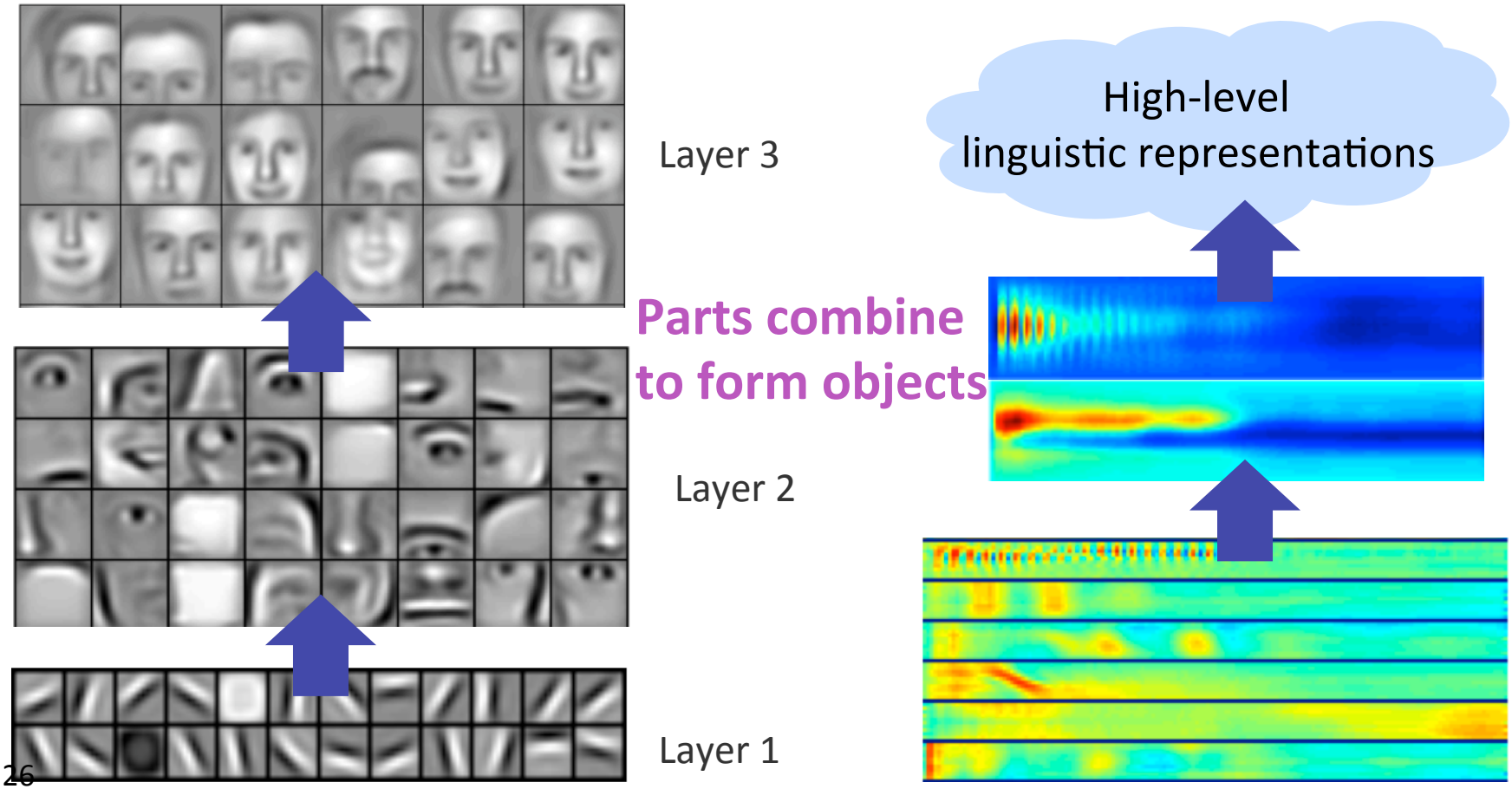
Illustration

Système développé par Google et U. de Stanford

- Reconnaissance de visages
  - Sous conditions de lumière diverses
  - Sous tout angle
- Apprentissage non supervisé
  - 9 couches ; 10^9 connexions
  - 10 millions d'images
  - 3 jours de calcul sur 16 000 processeurs
- Amélioration des performances de 70% / état de l'art

Apprentissage de représentations hiérarchiques

- Apprentissage de représentations hiérarchiques



26

Illustration : ImageNet

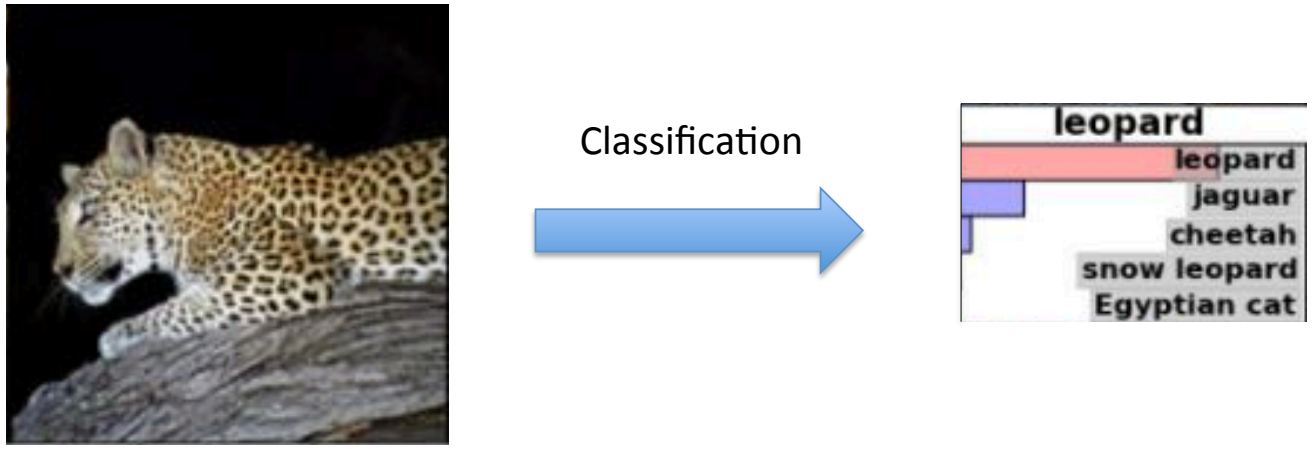
La compétition ImageNet

- Plus de **15M d'images** haute résolution étiquetées
- Environ **22K catégories**
- Récoltées sur le Web et étiquetées par Amazon Mechanical Turk

Illustration : ImageNet

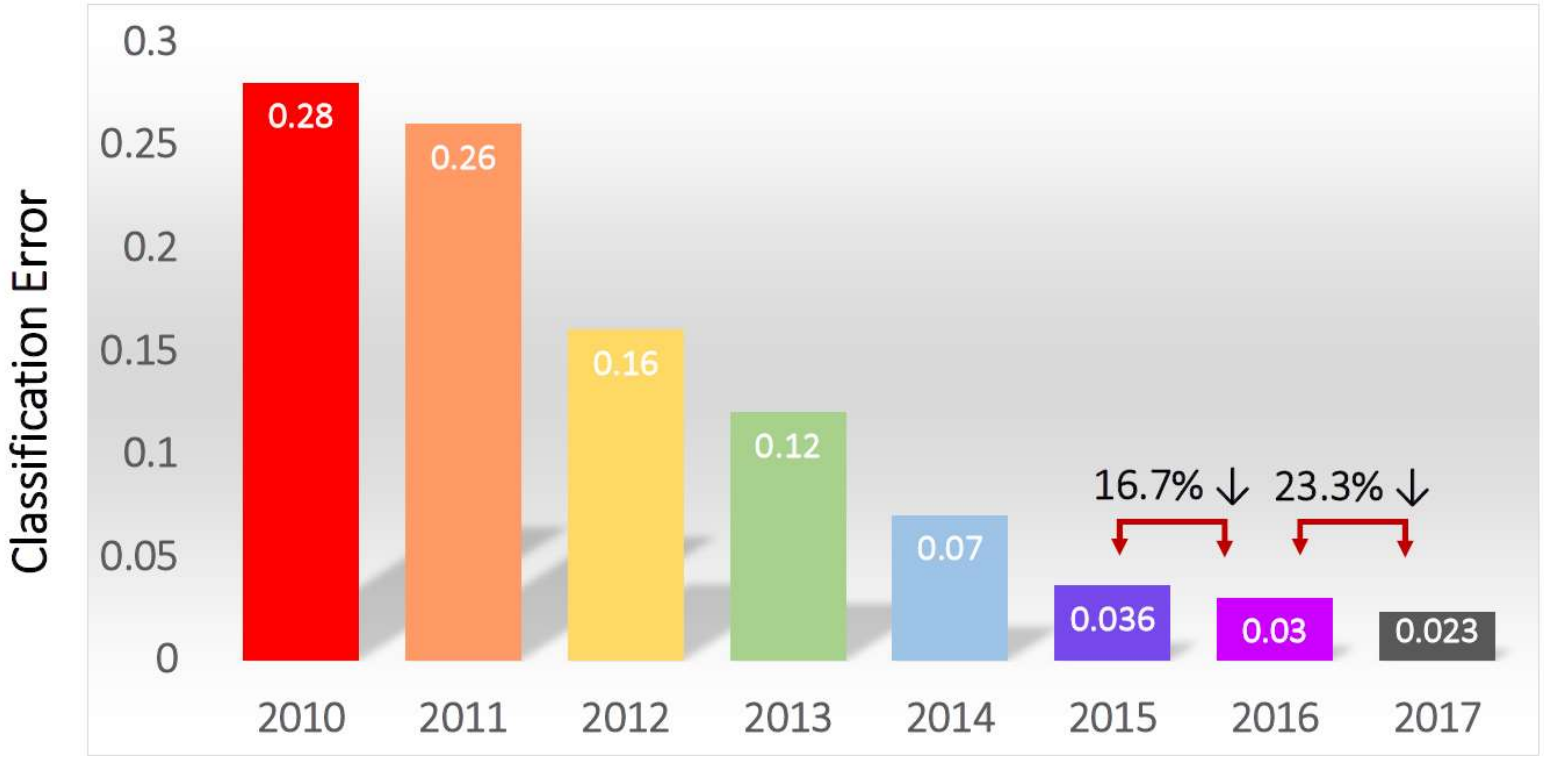
La compétition ImageNet

- Plus de **15M d'images** haute résolution étiquetées
- Environ **22K catégories**
- Récoltées sur le Web et étiquetées par Amazon Mechanical Turk

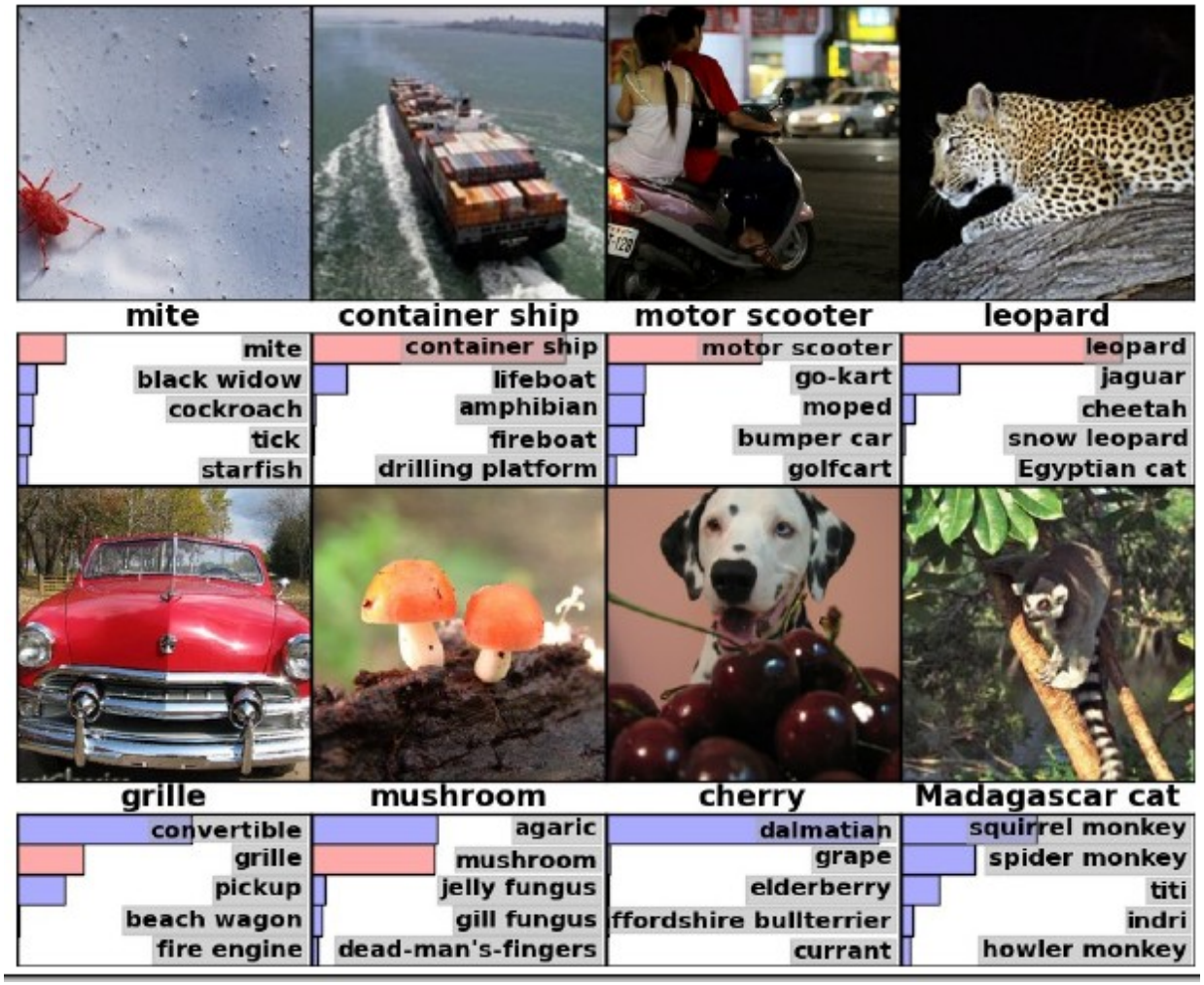


Les performances sur les compétition

- Résultats en classification



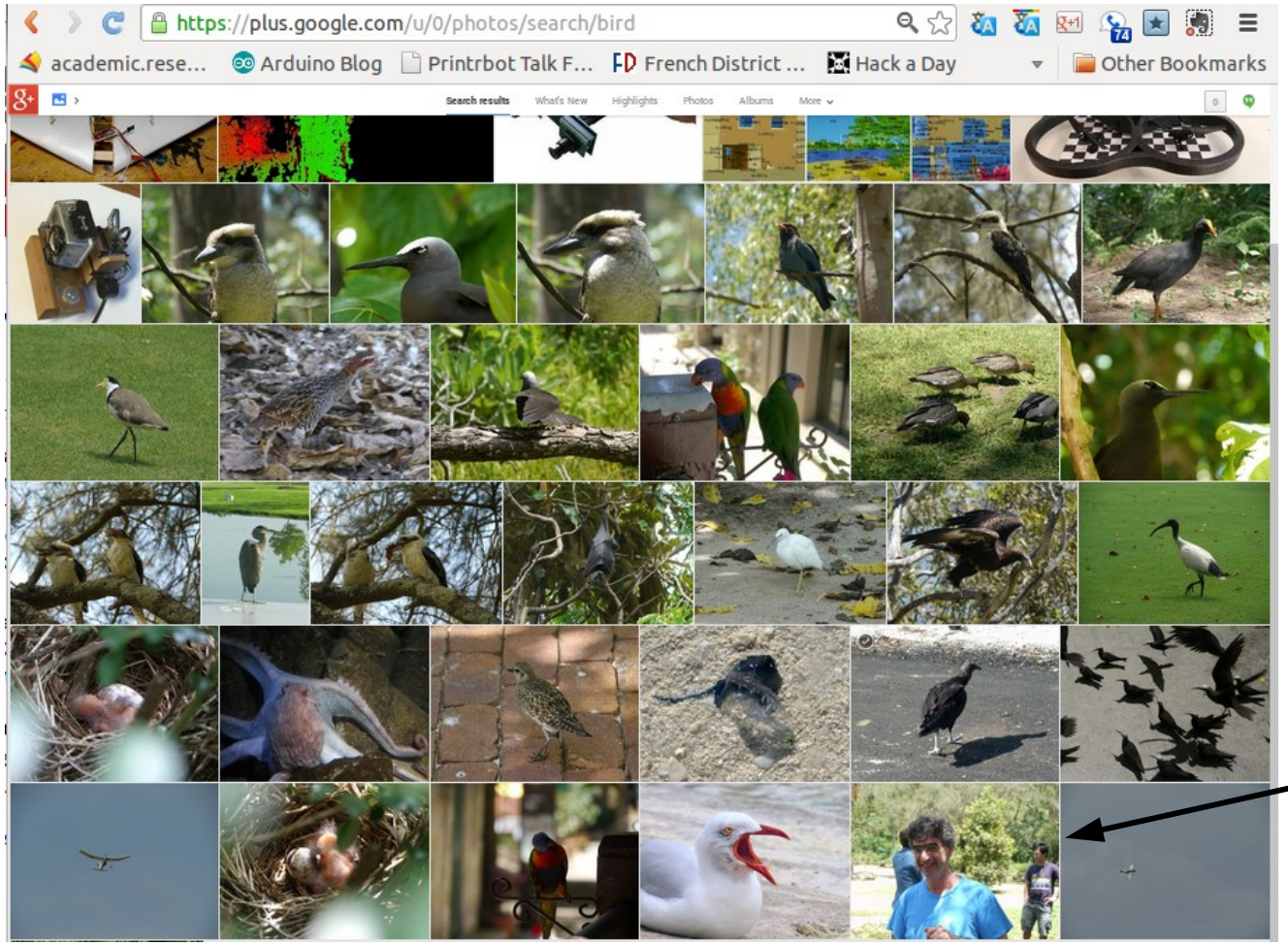
Object recognition



[Krizhevsky, Sutskever and Hinton (2012)]

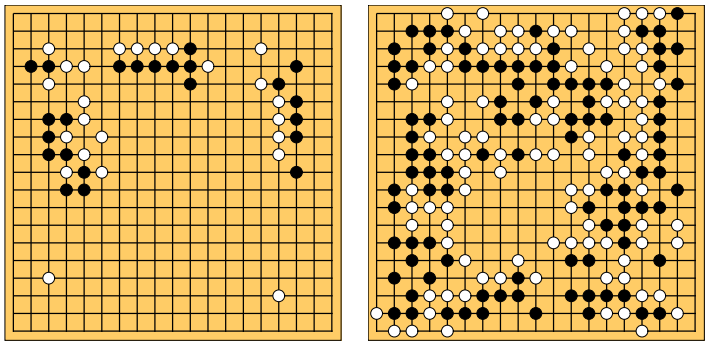
Object retrieval. ConvNet-Based Google+ Photo Tagger

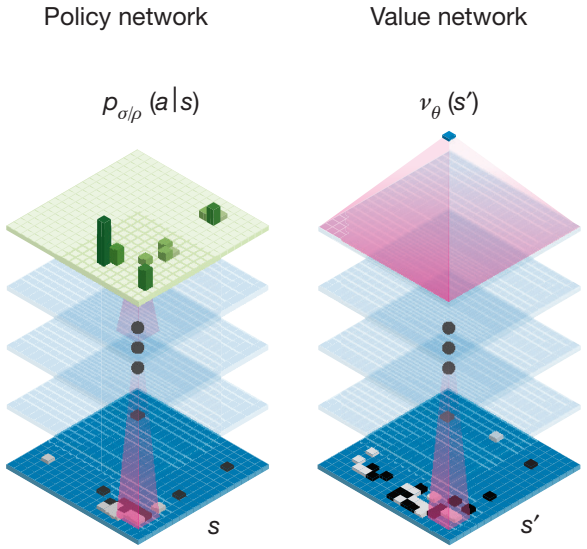
Searched my personal collection for "bird"

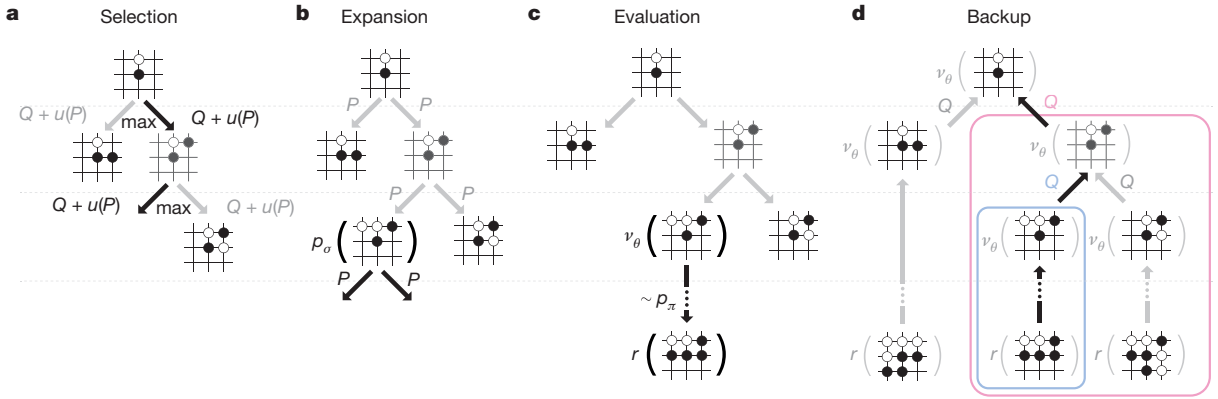


Game playing with Reinforcement Learning

- E.g. AlphaGo







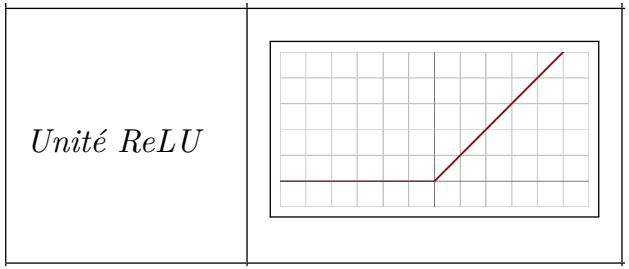
Comment faciliter

la **retro-propagation de gradient** ?

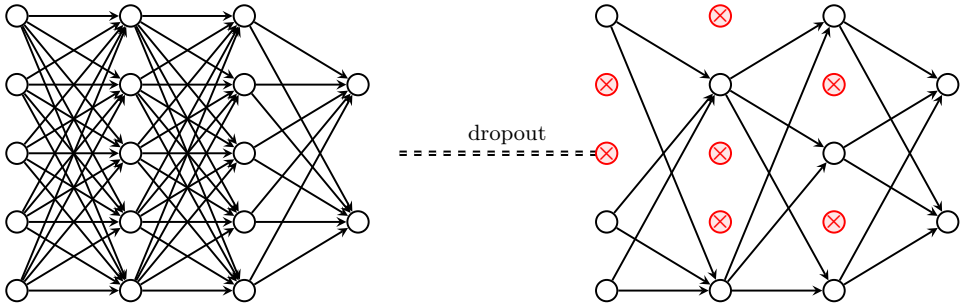
Des « astuces »

La **rétro-propagation classique ne marche pas** avec un grand nombre de couches (trop dilué)

1. Nouvelles **fonctions d'activation**



2. Le « **drop out** »



FIGU

3. L'utilisation de **GPU**

Nouvelles techniques d'optimisation

- **Très grande** activité de recherche

Un « bolide » délicat à piloter

Requiert

1. beaucoup de **données** (en général)
   - Des millions d'images
   - Des dizaines de milliers de documents
2. du **savoir-faire** (des data scientists)
   - Nombreuses « **astuces** » d'ingénierie
   - Utilisation de réseaux déjà appris (**transfert**)
   - L'état de l'art **progresse très vite**
3. des machines adaptées
   - Puissance **calcul** : clusters et/ou cartes graphiques
   - **Mémoire** centrale importante (≥ 128 Go)

Enseigné dans certaines écoles et universités

Il faut énormément d'**exemples étiquetés**

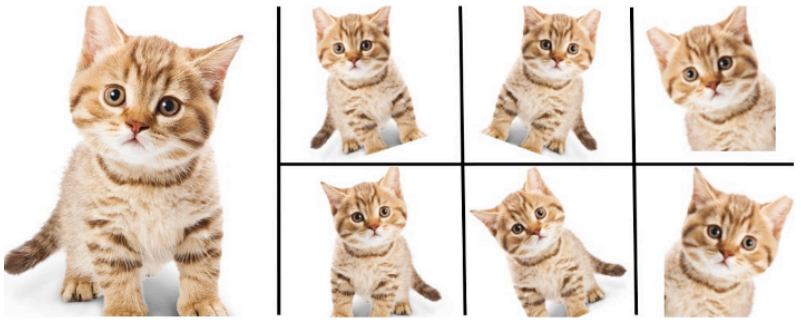
Comment faire ? ...

L’ « augmentation » de données

Génération de **nouveaux exemples**

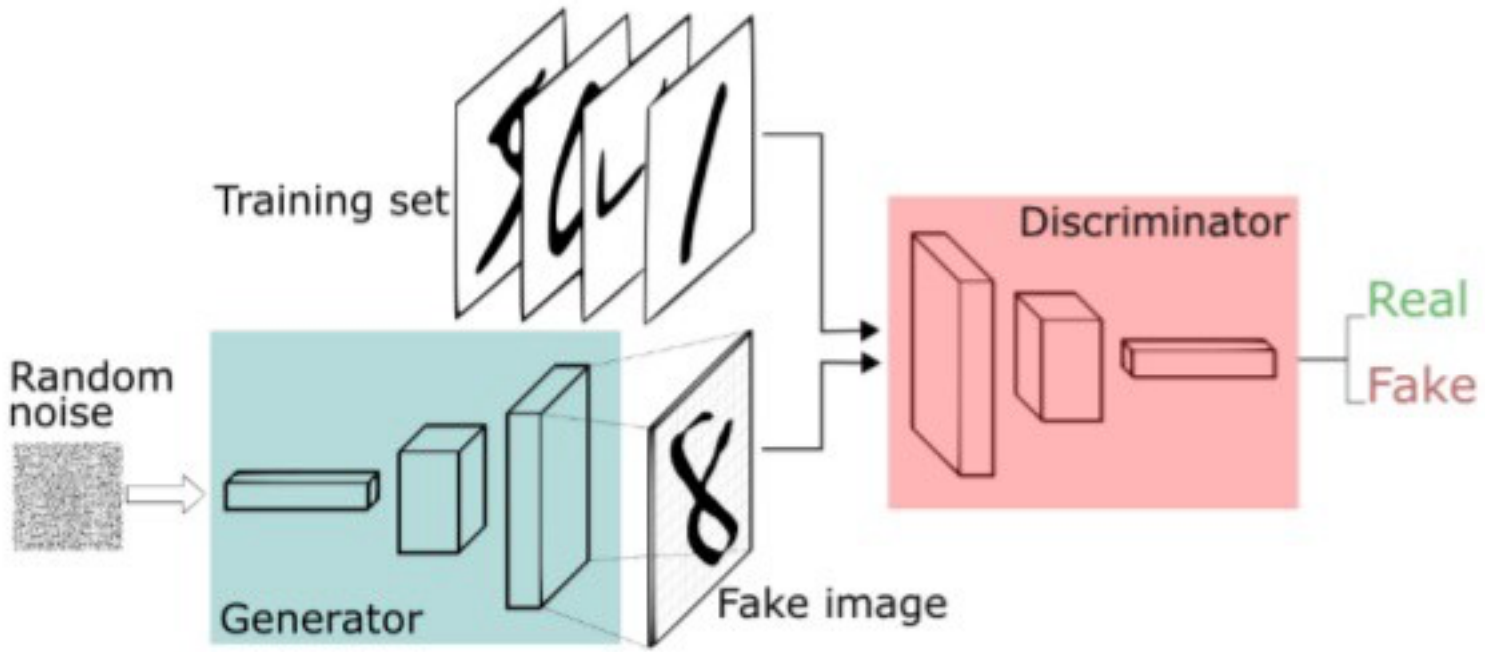
à partir des exemples étiquetés disponibles

1. **Transformation** de données
   - Bruitage / floutage / modification de contraste / changement de luminosité / autres effets spéciaux
   - Rotation / dilatation / translation



L’ « augmentation » de données

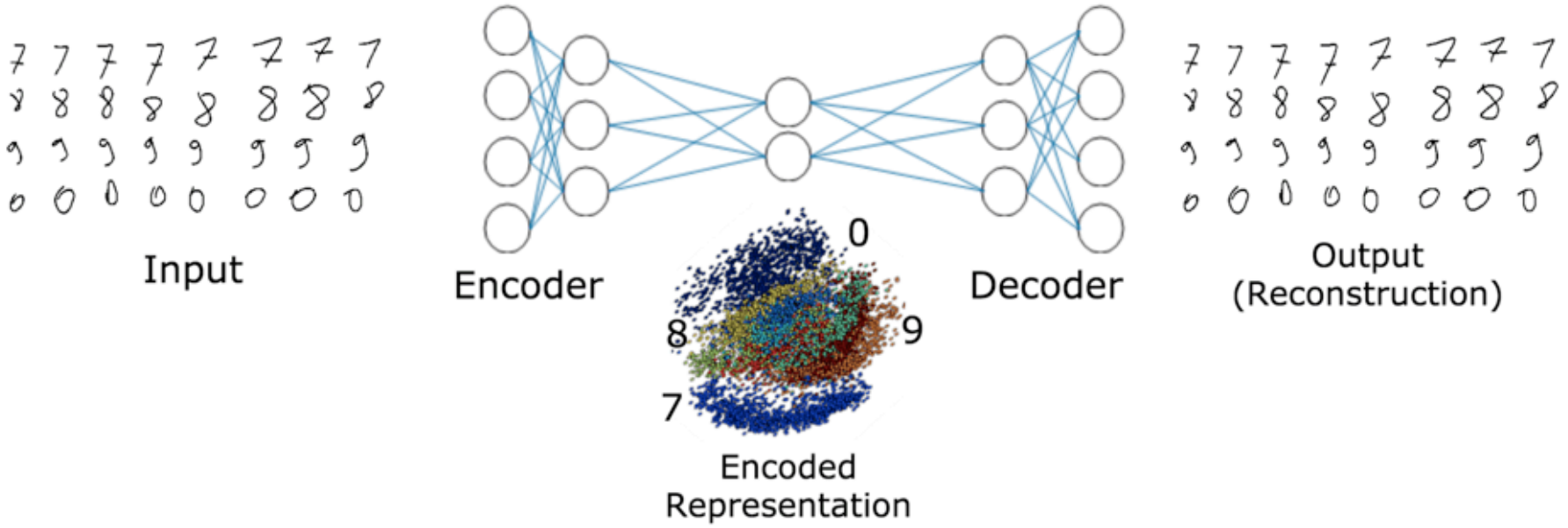
2. Génération par **réseaux GANs** (Generative Adversarial Networks)



Attention : Pas évident du tout à utiliser !

L' auto-encodage

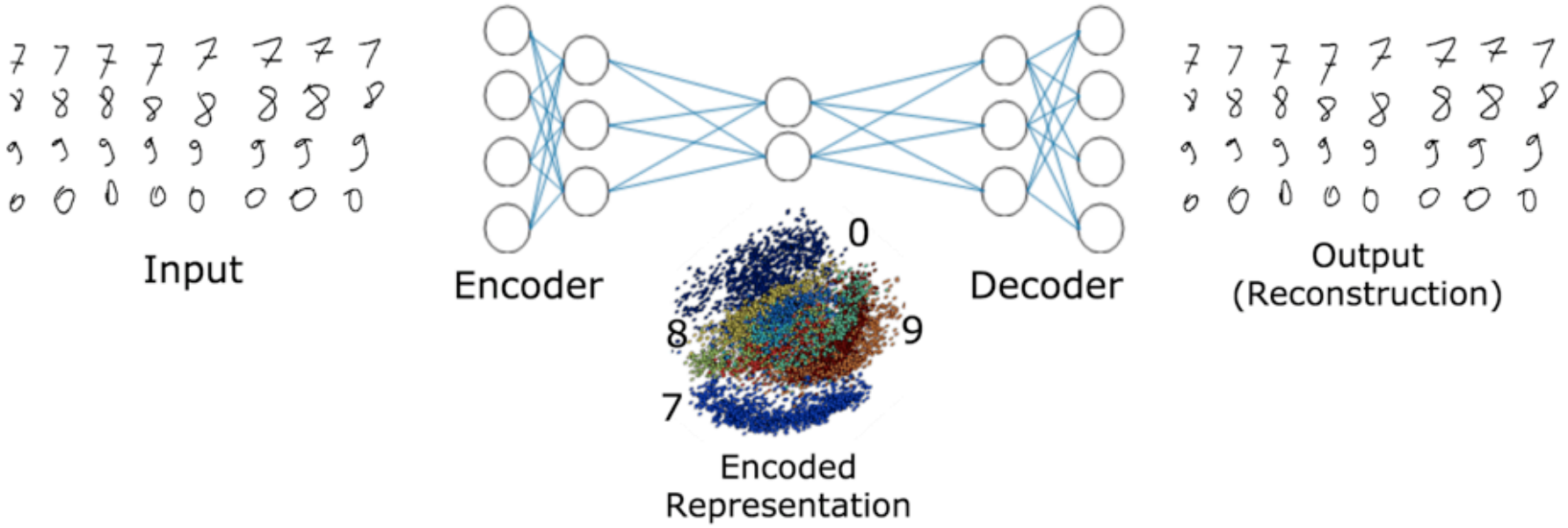
- Une vieille idée: l'auto-association



Représentation interne : l'*embedding*

L’ auto-encodage

- Une vieille idée: l’auto-association



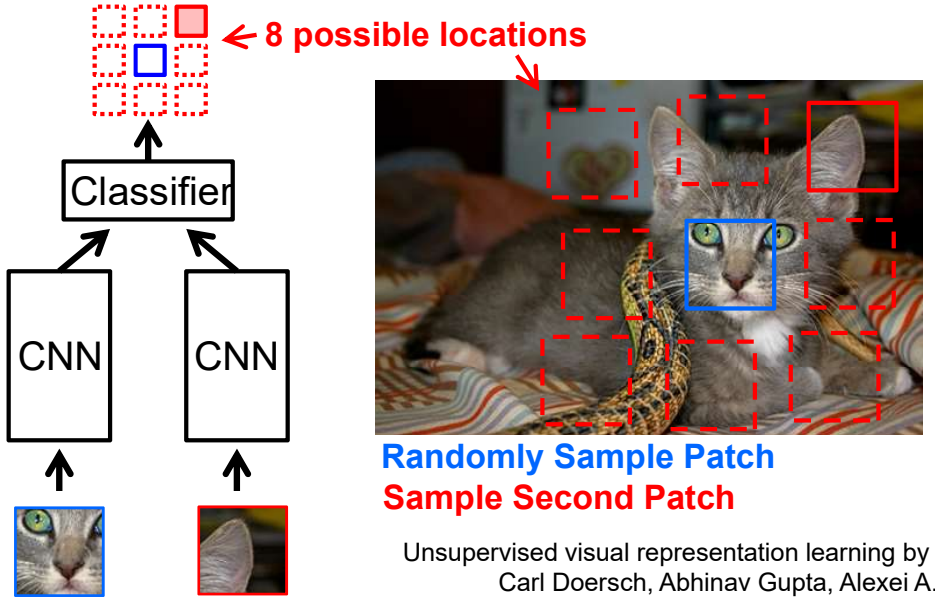
Un apprentissage supervisé ...

... sans étiquette !

L'auto-apprentissage (self-supervised learning)

- Une forme d'apprentissage non supervisée (pas d'étiquette nécessaire) mais qui permet un apprentissage supervisé !
  - Exemple : entraîner un réseau pour qu'il **prédise la position relative** de deux sous-images

    On espère ainsi apprendre au réseau des **connaissances relationnelles** sur ce qui constitue une image



Unsupervised visual representation learning by context prediction, Carl Doersch, Abhinav Gupta, Alexei A. Efros, ICCV 2015

L'auto-apprentissage (self-supervised learning)

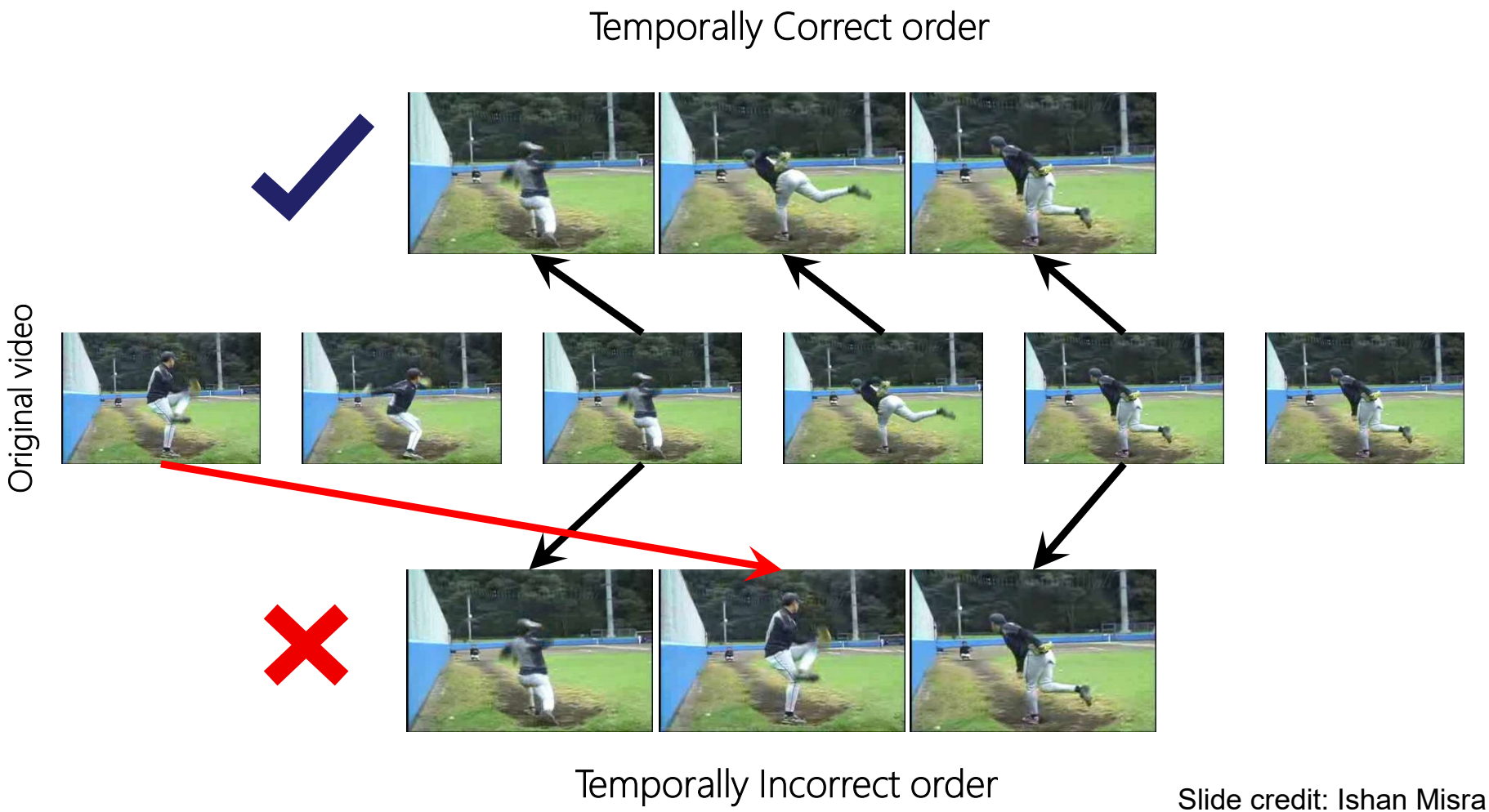
Dans des **séquences**

- Est-ce une séquence **valide** ?

Sun and Giles, 2001; Sun et al., 2001; Cleermans 1993; Reber 1989
Arrow of Time - Pickup et al., 2014

Slide credit: Ishan Misra

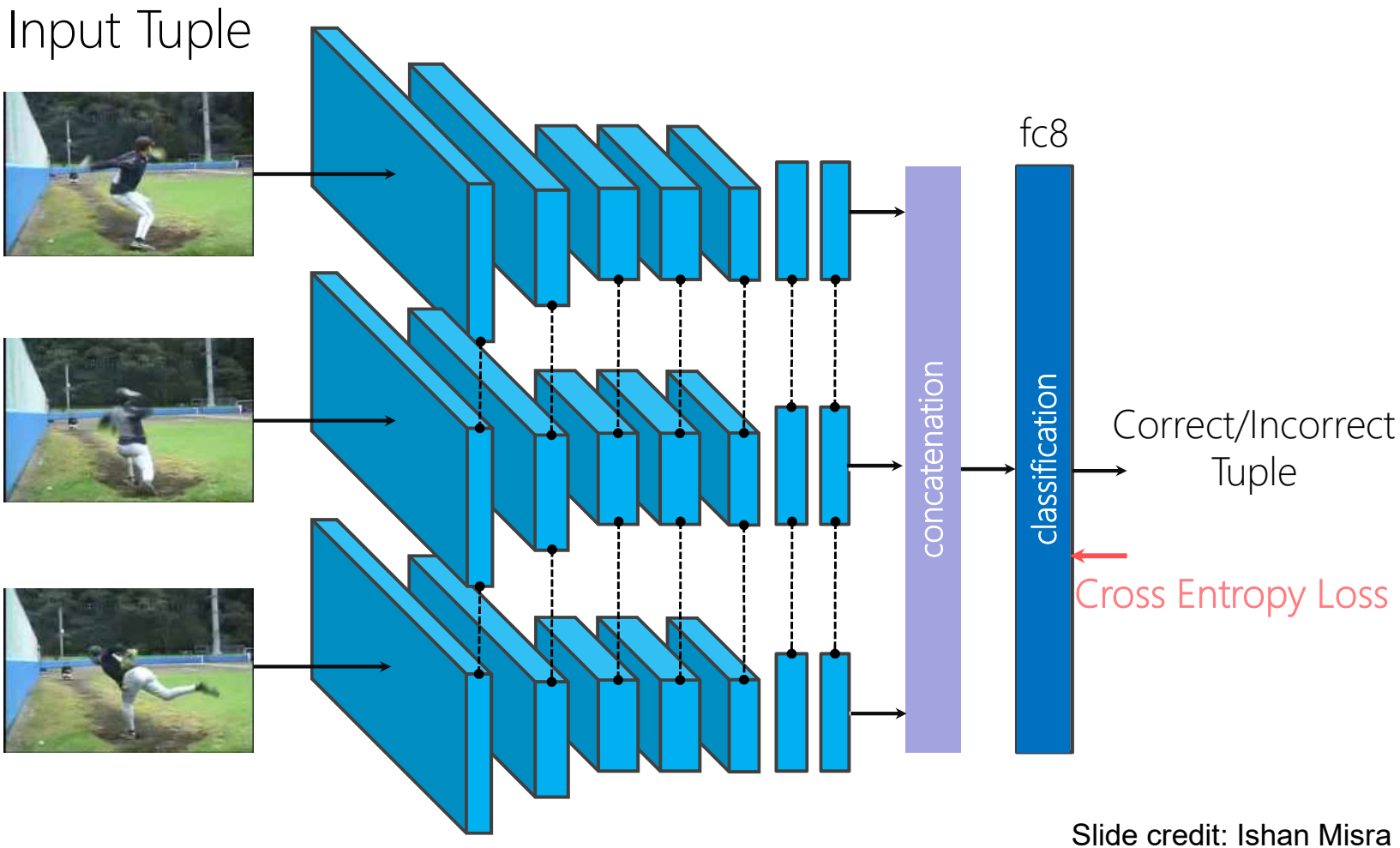
L'auto-apprentissage (self-supervised learning)



Slide credit: Ishan Misra

L'auto-apprentissage (self-supervised learning)

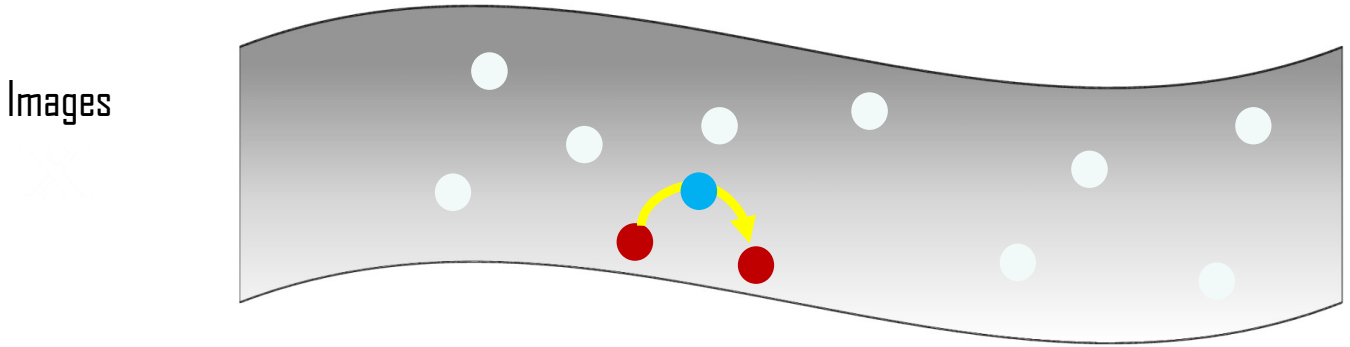
…



Slide credit: Ishan Misra

L'auto-apprentissage (self-supervised learning)

…



Given a start and an end, can this point lie in between?

Shuffle and Learn – I. Misra, L. Zitnick, M. Hebert – ECCV 2016

Slide credit: Ishan Misra

Plan

1. Pourquoi toute cette excitation ?
2. Grands types d'apprentissage
3. Apprentissage prédictif par réseaux de neurones
4. Quelles garanties ?
5. Le no-free-lunch theorem
6. Les réseaux de neurones profonds
7. Ce que l'on sait faire et les défis à relever

Les comportements **étranges**

Sait-on pourquoi ça marche ... Quand ça marche

Quelque chose de troublant

– C. Zhang, S. Bengio, M. Hardt, B. Recht, O. Vinyals (ICLR, **May 2017**).

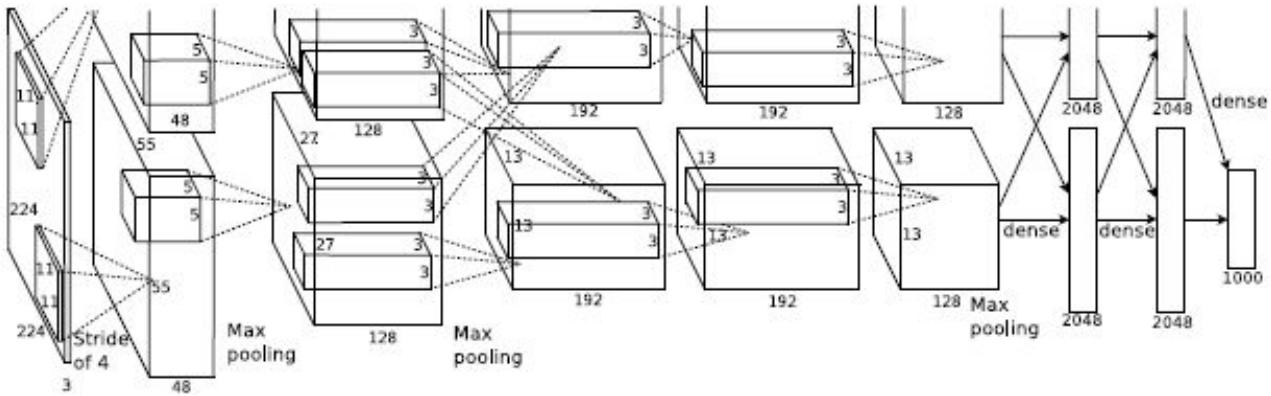
"Understanding deep learning requires rethinking generalization"

Quelque chose de troublant

– C. Zhang, S. Bengio, M. Hardt, B. Recht, O. Vinyals (ICLR, **May 2017**).
"Understanding deep learning requires rethinking generalization"

Extensive experiments on the classification of images

– The AlexNet (**> 1,000,000 parameters**) + 2 other architectures



– The **CIFAR-10 data set**:

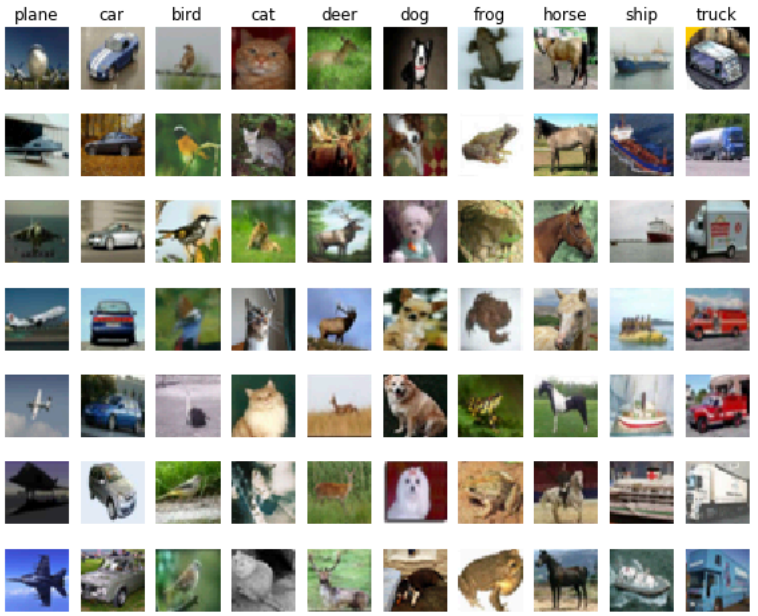
- **60,000** images categorized in **10 classes** (50,000 for training and 10,000 for testing)
- Images: 32x32 pixels in 3 color channels

Quelque chose de troublant

Experiments

1. **Original dataset** without modification

- Results ?
  - **Training** accuracy = 100% ; **Test** accuracy = 89%
  - Speed of convergence ~ 5,000 steps



Quelque chose de troublant

Experiments

1. **Original dataset** without modification

   - Results ?
     - **Training** accuracy = 100% ; **Test** accuracy = 89%
     - Speed of convergence ~ 5,000 steps

Expected behavior if the **capacity** of the hypothesis space is **limited**

i.e. the system **cannot** fit any (arbitrary) training data

$$\forall h \in \mathcal{H}, \forall \delta \leq 1 : \quad P^m \left[R(h) \leq \widehat{R}(h) + 2\,\widehat{Rad}_m(\mathcal{H}) + 3\sqrt{\frac{\ln(2/\delta)}{m}} \right] > 1 - \delta$$

Troubling findings

Experiments

1. **Original dataset** without modification
   - Results ?
     - **Training** accuracy = 100% ; **Test** accuracy = 89%
     - Speed of convergence ~ 5,000 steps

2. Random **labels** !!!
   - **Training** accuracy = 100% !!?? ; **Test** accuracy = 9.8%
   - Speed of convergence = similar behavior (~ 10,000 steps)

Troubling findings

Experiments

1. **Original dataset** without modification
   - Results ?
     - **Training** accuracy = 100% ; **Test** accuracy = 89%
     - Speed of convergence ~ 5,000 steps

2. Random **labels**
   - **Training** accuracy = 100% !!?? ; **Test** accuracy = 9.8%
   - Speed of convergence = similar behavior (~ 10,000 steps)

3. Random **pixels**
   - **Training** accuracy = 100% !!?? ; **Test** accuracy ~ 10%
   - Speed of convergence = similar behavior (~ 10,000 steps)

Now, we are in trouble!!

Troubling findings

- Deep NNs can accommodate ANY training set

Can grow without limit!!

$$\forall h \in \mathcal{H}, \forall \delta \leq 1 : \quad P^m \left[R(h) \leq \widehat{R}(h) + 2\,\widehat{Rad}_m(\mathcal{H}) + 3\sqrt{\frac{\ln(2/\delta)}{m}} \right] > 1 - \delta$$

But then,

why are deep NNs ***so good on image classification tasks?***

Ce que l'on sait faire.

Sait-on d'ailleurs vraiment le faire ?

Ce qui interroge.

Ce qui reste à faire.

Un peu de recul :

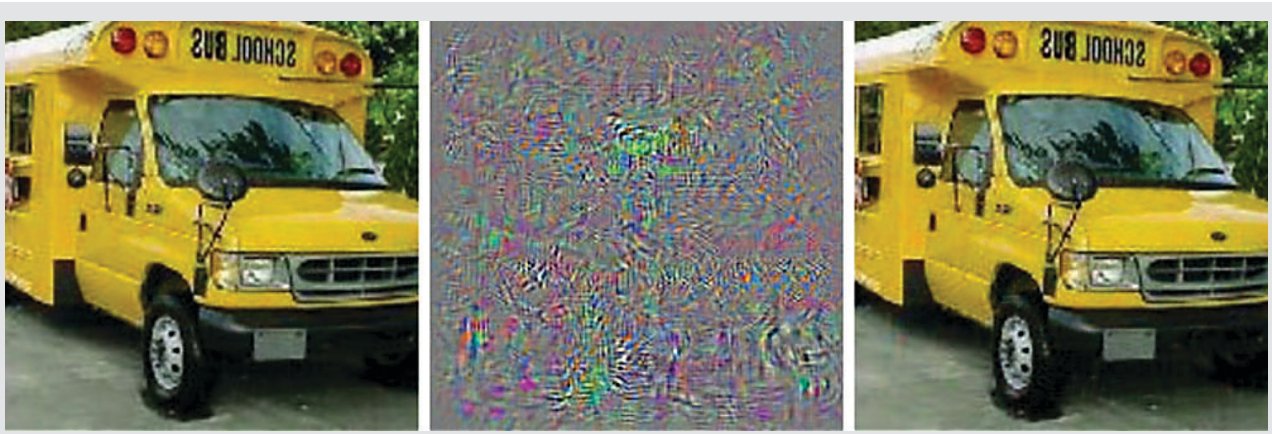
Que sait-on faire

et **où sont** les **limites** ?

Ce que l'on sait faire

Quoique !?

Adversarial learning



Adversarial input can fool a machine-learning algorithm into misperceiving images.

Explanations and deep neural networks

Optical illusions: how to explain them?

Boxer: 0.40 Tiger Cat: 0.18

(a) Original image

Airliner: 0.9999

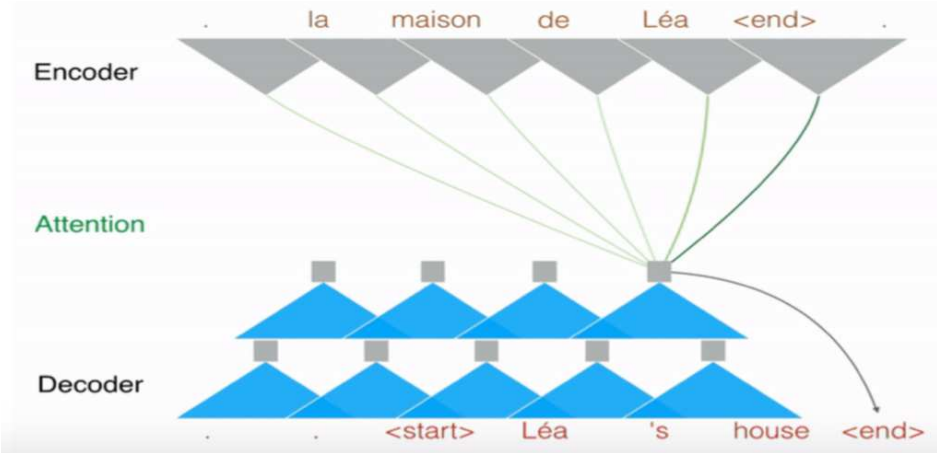
(b) Adversarial image

!!??

[Selvaraju et al. (2017) « *Grad-CAM: Visual explanations from deep networks via gradient-based localization* »]

Machine translation

- Still far from perfect, but …



From Hofstädter (2018)

Annotation d'images

Figure 2.11: "A group of young people playing a game of frisbee"—that caption was written by a computer with no understanding of people, games or frisbees.

Automated image-captioning

- Not always so good!

A dog is jumping to catch a frisbee

Montpellier SupAgro (2022) « Une perspective sur l'apprentissage » (A. Commenges)

Exemple en médecine

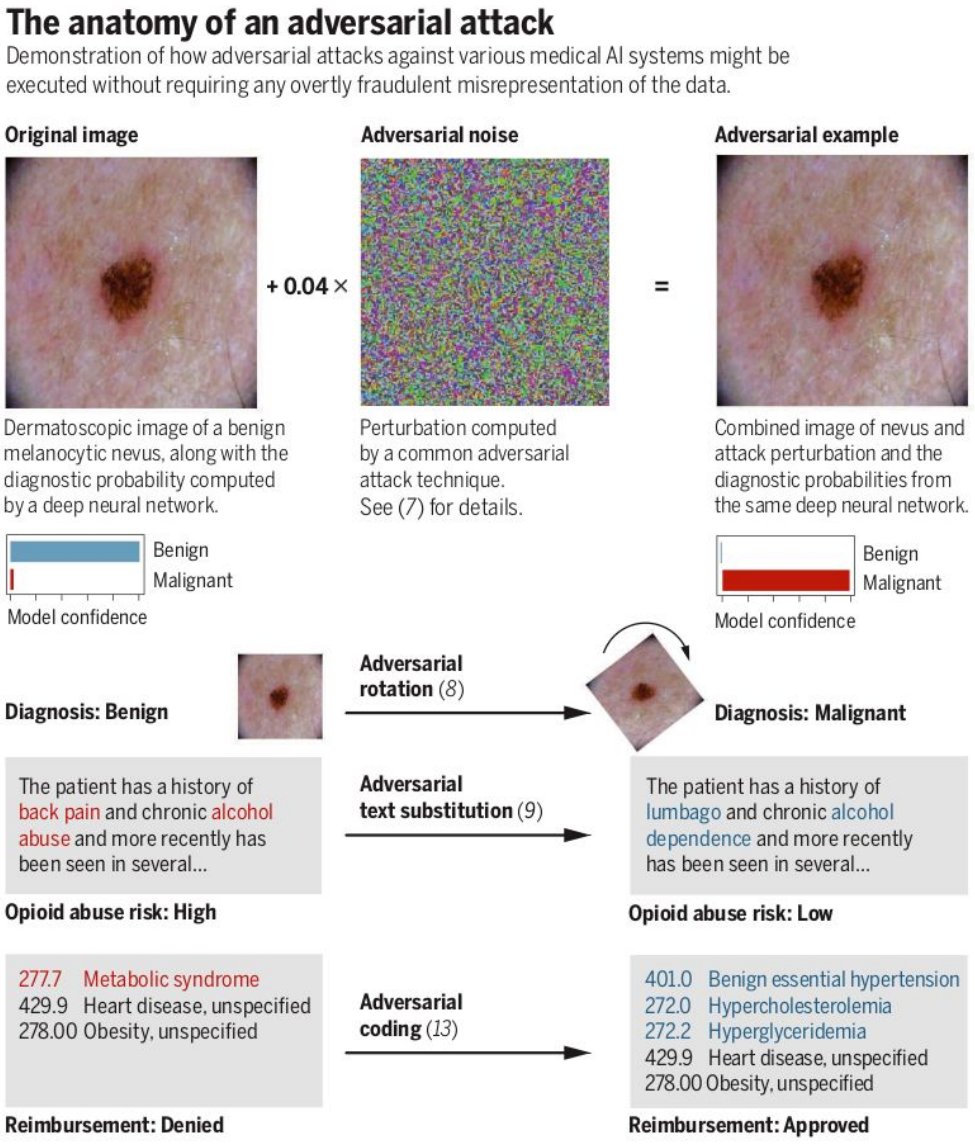
MACHINE LEARNING

Science

Adversarial attacks on medical machine learning

Emerging vulnerabilities demand new conversations

22 March 2019



Voiture dans une piscine

- … ou pas de voiture … ?

Is this less of a car
because the context is wrong?

[Léon Bottou (ICML-2015, invited talk) « *Two big challenges in Machine Learning* »]

L'IA comprend-t-elle ?

https://www.youtube.com/watch?v=QPSgM13hTK8&t=117

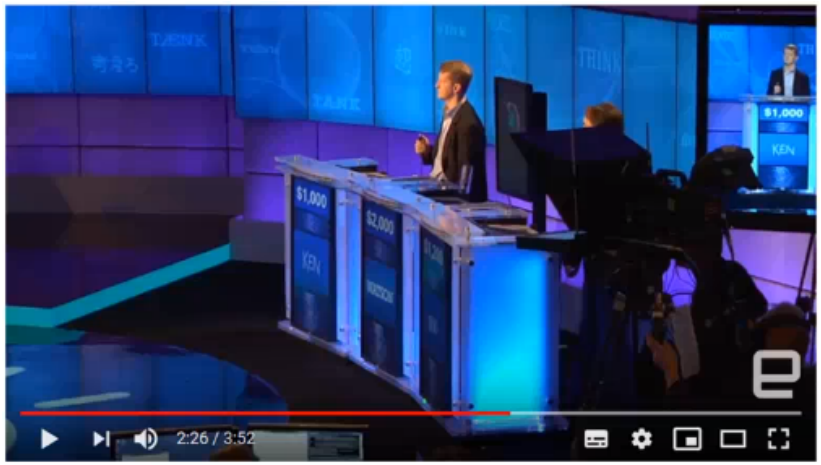
WATSON et le jeu Jeopardy! (2011)

Jeopardy! In the category U.S. cities:

- "Its largest airport was named for a World War II hero; its second largest, for a World War battle."

➢ What is *Toronto*?

New-York!!



IBM's Watson Supercomputer Destroys Humans in Jeopardy | Engadget

Conclusions

On a vu ...

- Des **grands types** d'apprentissage
- Le problème de **l'apprentissage supervisé**
- La nécessité d'un **biais** et la notion de **critère inductif**
- Comment **mesurer la valeur** d'une hypothèse apprise
- Le cas des **réseaux de neurones**
  - **Structure**
  - **Algorithme** d'apprentissage
  - Les RNs **profonds** : très puissants, encore mal compris

Directions de **recherche**

- **XAI** (Explainable AI)
  - Avoir des systèmes **interprétables**
  - Intégrer du **raisonnement**
- Apprentissage de relations de **causalité**
- Sources de données **multiples** et **hétérogènes**
- Apprendre à partir de **très peu d'exemples**
- Environnement **non stationnaire**
  - **Changements** de l'environnement
  - **Transfert** entre tâches
- Faire **coopérer** des systèmes d'apprentissage

Beaucoup d'opportunités

Mais pas de magie

Suppléments

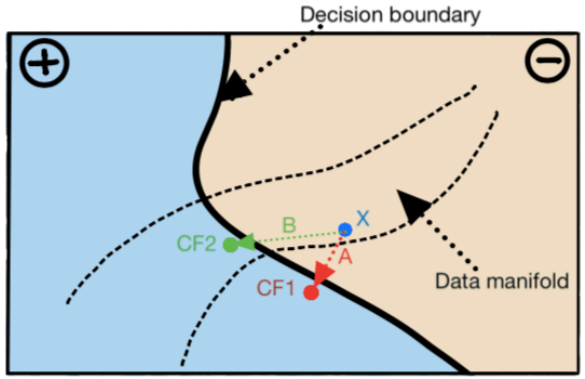
Counterfactuals

- **If** James Dean had **taken the train** the day of his car accident, he would not have died
- **If** you could **increase your savings** by 5000€ each year, you would get this loan

Counterfactuals

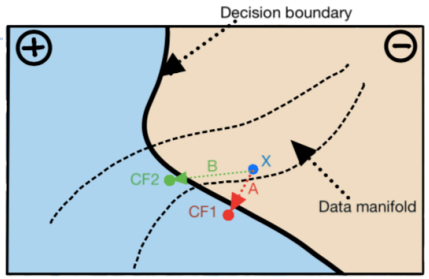
- **If** James Dean had **taken the train** the day of his car accident, he would not have died
- **If** you could **increase your savings** by 5000€ each year, you would get this loan

Local explanation for a given prediction



Two possible **counterfactuals**: CF1 is closest to **x** than CF2

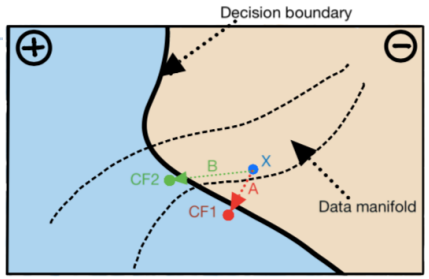
Counterfactuals



- Oh yes. But what is the difference with **adversarial examples?!**

This is not "toilet paper"

Counterfactuals

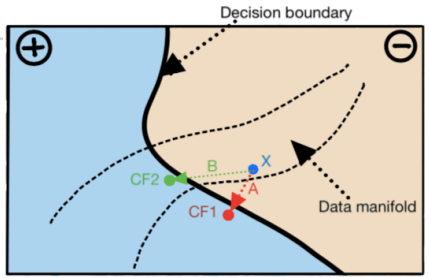


- Oh yes. But what is the difference with **adversarial examples?!**

This is not "toilet paper"

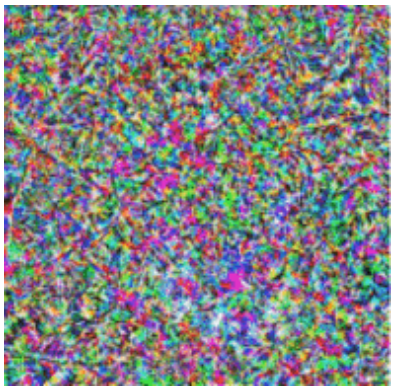
because this is "dog"

Counterfactuals



- Oh yes. But what is the difference with **adversarial examples?!**

And the **difference** is



This is not "toilet paper" because this is "dog"

What is a **good level** of communication?

"No computation can get around the semantic problem"

K. Browne & B. Swift (2020). "**Semantics and explanation: why counterfactual explanations produce adversarial examples in deep neural networks**". *arXiv preprint arXiv:2012.10076*.

Le paradigme actuel

- Induire nécessite d'avoir des biais
- La **théorie**
  - Est entièrement focaliée sur le taux d'erreur
  - Présuppose un environnement stationnaire et des entrées/requêtes (i.i.d.)
  - Exige un **nombre de données d'apprentissage assez grand** par rapport à la **capacité de** $\mathcal{H}$
- Nous ne comprenons pas bien les réseaux de neurones profonds
- Corrélations ≠ structures, sémantique, causalité

Limites

- Apprentissage **passif** et **données et questions i.i.d.**
  - Agents situés : le monde n'est pas i.i.d.
- Requiert **beaucoup** d'exemples
  - Nous sommes beaucoup plus efficaces
  - « Producteurs de théories », théories que nous testons ensuite
- Pas adapté à la recherche de **causalités**
- Pas **intégré** avec un **raisonnement**

Ces machines **apprenantes** ne sont pas des machines **pensantes**

Mes paris pour l'avenir

Mes paris sur les directions à venir

1. Apprendre à partir de **très peu d'exemples**
2. Apprendre à partir de **multiples sources de données** hétérogènes
3. Apprendre par **analogie** et par **transfert**
4. Apprendre pour **construire des théories ?** (causalité et explications)
5. L'**intégration** de multiples systèmes apprenants
6. Des systèmes capables d'enseigner