Artificial Intelligence: News and Questions

Michele Sebag
CNRS – INRIA – Univ. Paris-Saclay

ArenbergSymposium – Leuven – Nov. 27th, 2019

Credit for slides: Yoshua Bengio; Yann LeCun; Nando de Freitas; Léon Gatys; Max Welling; Victor Berger
Université Paris-Saclay in 1 slide

14 partners

- 3 Univ.
- 4 Grandes coles
- 7 Research Institutes

15% of French Research

5,500 PhD; 10,000 Faculty members
10 Field medals; 3 Nobel; 160 ERC.

Wendelin Werner (mathematics), Valérie Masson-Delmotte (climate), Alain Aspect (physics), Patrick Coulouvou (medicine), Albert Fert (nanoscience), Stanislas Dehaene (neuroscience)…
Some AI projects

- Center for Data Science
  ML Higgs Boson Challenge (2015)

- Institute DataIA: AI for Society

- Big Data, Optimization and Energy (See4C challenge)
Two visions of AI – 1950 - 1960

Logical calculus can be achieved by machines!

- All men are mortal.
- Socrates is a man.
- Therefore, Socrates is mortal.

Primary operation: Deduction (reasoning) ? or Induction (learning)?

- Should we learn what we can reason with ?
- Should we reason with what we can learn ?

<table>
<thead>
<tr>
<th>HOW</th>
<th>Alan Turing</th>
<th>John McCarthy</th>
</tr>
</thead>
<tbody>
<tr>
<td>VALIDATION</td>
<td>Learning</td>
<td>Reasoning</td>
</tr>
<tr>
<td>Human assessment</td>
<td>Pb Solving</td>
<td></td>
</tr>
</tbody>
</table>
1950: Computing Machinery and Intelligence

- Storage will be ok
- But programming needs *prohibitively large human resources*
- Hence, machine learning.

*by (...) mimicking education, we should hope to modify the machine until it could be relied on to produce definite reactions to certain commands.*

*One could carry through the organization of an intelligent machine with only two interfering inputs, one for pleasure or reward, and the other for pain or punishment.*
The imitation game

The Turing test

Issues

- Human assessment; no golden standard.
1956: The Dartmouth conference

- With Marvin Minsky, Claude Shannon, Nathaniel Rochester, Herbert Simon, et al.

- The study is to proceed on the basis of the conjecture that every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it. An attempt will be made to find how to make machines use language, form abstractions and concepts, solve kinds of problems now reserved for humans, and improve themselves.
Computational Logic for AI

- Declarative languages
- Symbolic methods, deduction
- Focussed domains (expert systems)
- Games and problem solving

Issues
- How to ground symbols?
- Where does knowledge come from?
Automating Science using Robot Scientists

From facts to hypotheses to experiments to new facts...

▶ A proper representation and domain theory

\[ \text{Benzene}(A_1, A_2, A_3, A_4, A_5, A_6) : - \text{Carbon}(A_1), \text{Carbon}(A_2), \ldots \]
\[ \text{Bond}(A_1, A_2), \text{Bond}(A_2, A_3), \text{Bond}(A_3, A_4) \ldots \]

▶ Active Learning – Design of Experiments
▶ Control of noise
The Wave of AI
Neural Nets, ups and downs in AI

(C) David McKay - Cambridge Univ. Press

**History**

1943  A neuron as a computable function $y = f(x)$  
      Intelligence $\rightarrow$ Reasoning $\rightarrow$ Boolean functions  
      Pitts, McCullough

1960  Connexionism + learning algorithms  
      Rosenblatt

1969  AI Winter  
      Minsky-Papert

1989  Back-propagation  
      Amari, Rumelhart & McClelland, LeCun

1992  NN Winter  
      Vapnik

2005  Deep Learning  
      Bengio, Hinton
The revolution of Deep Learning

- Ingredients were known for decades
- Neural nets were no longer scientifically exciting (except for a few people)
- Suddenly... 2006
Revival of Neural Nets

1. Grand goal: AI

2. Requisites
   - Computational efficiency
   - Statistical efficiency
   - Prior efficiency: architecture relies on human labor

3. Compositionality principle: skills built on the top of simpler skills

Bengio, Hinton 2006

Piaget 1936
1. Grand goal: AI

2. Requisites
   ▶ Computational efficiency
   ▶ Statistical efficiency
   ▶ Prior efficiency: architecture relies on student labor

3. Compositionality principle: skills built on the top of simpler skills

Bengio, Hinton 2006

Piaget 1936
The importance of being deep

A toy example: \( n \)-bit parity

Pros: efficient representation

*Deep neural nets are exponentially more compact*

Cons: poor learning

- More layers \( \rightarrow \) more difficult optimization problem
- Getting stuck in poor local optima.
- **Handled through smart initialization**
Convolutional NNs: Enforcing invariance

Invariance matters

- Visual cortex of the cat
  - cells arranged in such a way that
  - ... each cell observes a fraction of the visual field
  - ... their union covers the whole field

Layer $m$: detection of local patterns
Layer $m + 1$: non linear aggregation of output of layer $m$
Convolutional architectures

For images

For signals
Gigantic architectures

Properties

- Invariance to small transformations (over the region)
- Reducing the number of weights by several orders of magnitude
What is new?
Former state of the art

- **SIFT**: Scale Invariant Feature Transform
- **HOG**: Histogram of Oriented Gradients
- **Textons**: “vector quantized responses of a linear filter bank”

*e.g. in computer vision*

- **Spin image**
- **RIFT**
- **GLOH**

**Textons**

- SIFT: scale invariant feature transform
- HOG: histogram of oriented gradients
- Textons: “vector quantized responses of a linear filter bank”
What is new, 2

Traditional approach

Manually crafted features → Trainable classifier

Deep learning

Trainable feature extractor → Trainable classifier
A new representation is learned

Bengio et al. 2006
Why Deep Learning now?

CONS

▸ a non-convex optimization problem
▸ no theorems
▸ delivers a black box model

PROS

▸ Linear complexity w.r.t. #data
▸ Performance leaps if enough data and enough computational power.
A leap in the state of the art: ImageNet

15 million labeled high-resolution images; 22,000 classes.

Large-Scale Visual Recognition Challenge

- 1000 categories.
- 1.2 million training images,
- 50,000 validation images,
- 150,000 testing images.
A leap in the state of the art, 2

<table>
<thead>
<tr>
<th>2012 Teams</th>
<th>%error</th>
</tr>
</thead>
<tbody>
<tr>
<td>Supervision (Toronto)</td>
<td>15.3</td>
</tr>
<tr>
<td>ISI (Tokyo)</td>
<td>26.1</td>
</tr>
<tr>
<td>VGG (Oxford)</td>
<td>26.9</td>
</tr>
<tr>
<td>XRCE/INRIA</td>
<td>27.0</td>
</tr>
<tr>
<td>UvA (Amsterdam)</td>
<td>29.6</td>
</tr>
<tr>
<td>INRIA/LEAR</td>
<td>33.4</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>2013 Teams</th>
<th>%error</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clarifai (NYU spinoff)</td>
<td>11.7</td>
</tr>
<tr>
<td>NUS (singapore)</td>
<td>12.9</td>
</tr>
<tr>
<td>Zeiler-Fergus (NYU)</td>
<td>13.5</td>
</tr>
<tr>
<td>A. Howard</td>
<td>13.5</td>
</tr>
<tr>
<td>OverFeat (NYU)</td>
<td>14.1</td>
</tr>
<tr>
<td>UvA (Amsterdam)</td>
<td>14.2</td>
</tr>
<tr>
<td>Adobe</td>
<td>15.2</td>
</tr>
<tr>
<td>VGG (Oxford)</td>
<td>15.2</td>
</tr>
<tr>
<td>VGG (Oxford)</td>
<td>23.0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>2014 Teams</th>
<th>%error</th>
</tr>
</thead>
<tbody>
<tr>
<td>GoogLeNet</td>
<td>6.6</td>
</tr>
<tr>
<td>VGG (Oxford)</td>
<td>7.3</td>
</tr>
<tr>
<td>MSRA</td>
<td>8.0</td>
</tr>
<tr>
<td>A. Howard</td>
<td>8.1</td>
</tr>
<tr>
<td>DeeperVision</td>
<td>9.5</td>
</tr>
<tr>
<td>NUS-BST</td>
<td>9.7</td>
</tr>
<tr>
<td>TTIC-ECP</td>
<td>10.2</td>
</tr>
<tr>
<td>XYZ</td>
<td>11.2</td>
</tr>
<tr>
<td>UvA</td>
<td>12.1</td>
</tr>
</tbody>
</table>

shallow approaches
deep learning

Y. LeCun StatLearn tutorial
Super-human performances

- 2012  Alex Net
- 2013  ZFNet
- 2014  VGG
- 2015  GoogLeNet / Inception
- 2016  Residual Network

graph showing top-5 error in %, number of layers, and human performance over publication years 2010-2016, with a peak of 152 layers in 2016.
Playing with representations

Style and contents in a convolutional NN are separable

Use a trained VGG-19 Net:
- applied on image 1 (content)
- applied on image 2 (style)
- find input matching hidden representation of image 1 (weight $\alpha$) and hidden representation of image 2 (weight $\beta$)
Beyond classifying, modifying data: generating data

“What I cannot create I do not understand”  

— Feynman 88
Generative Adversarial Networks

Goodfellow et al., 14

**Goal:** Find a generative model

- Classical: learn a distribution hard
- Idea: replace a distribution evaluation by a 2-sample test

**Principle**

- Find a good generative model, s.t. generated samples cannot be discriminated from real samples (not easy)
Generative Adversarial Networks, 2

Goodfellow, 2017

Elements

- Dataset, true samples $x$ (real)
- Generator $G$, generated samples (fake)
- Discriminator $D$: discriminates fake from real

A min-max game

Embedding a Turing Test!

$$\min_G \max_D \mathbb{E}_{x \in \text{data}}[\log(D(x))] + \mathbb{E}_{z \sim p(z)}[\log(1 - D(z))]$$
Generative Adversarial Networks: successes

Mescheder, Geiger and Nowozin, 2018
and monsters
From Deep Learning to Differentiable Programming

**Principle**

- Most programs can be coded as a neural net.
- Define a performance criterion
- Let the program interact with the world and train itself: programs that can learn

**Revisiting Partial Differential Equations**

1. Combining with simulations: recognizing Tropical Cyclones

Kurth et al. 18

1152x768 spatial grid, 3 hours time step
3 classes: TCs (0.1%), Atmospheric Rivers (1.7%) and Background
Physics Informed Deep Learning

Data driven solutions of PDE

Equation:

\[ u_t = \mathcal{N}(t, u, x, u_x, u_{xx}, \ldots) \]

residual:

\[ f := u_t - \mathcal{N}(t, u, x, u_x, u_{xx}, \ldots) \]

Initial and boundary conditions: \((t^i_u, x^i_u, u^i), i = 1 \ldots N\)
Training points \((t^j_f, x^j_f), j = 1 \ldots N'\)

Train \(u(t, s)\) minimizing

\[
\frac{1}{N} \sum_{i=1}^{N} |u(x^i_u, t^i_u) - u^i|^2 + 
\]

\[
(0, x) = \sin(\pi x / 8) \]

\[
(0, x) = -\exp(-x^2) \]

train

test
Physics Informed Deep Learning

Data driven solutions of PDE

Equation:

$$u_t = \mathcal{N}(t, u, x, u_x, u_{xx}, \ldots)$$

residual:

$$f := u_t - \mathcal{N}(t, u, x, u_x, u_{xx}, \ldots)$$

Initial and boundary conditions: \((t^i_u, x^i_u, u^i), i = 1 \ldots N\)
Training points \((t^j_f, x^j_f), j = 1 \ldots N'\)

\textbf{Train} \(u(t, s)\) minimizing

$$\frac{1}{N} \sum_{i=1}^{N} |u(x^i_u, t^i_u) - u^i|^2 +$$

\begin{align*}
\text{test} & \quad u(0, x) = \sin(\pi x / 8) \\
\text{train} & \quad u(0, x) = -\exp(- (x + 2)^2)
\end{align*}
Issues with black-box models

Good performances $\nRightarrow$ Accurate model
Robustness wrt perturbations

What can happen when perturbing an example? Anything!

Malicious perturbations

Goodfellow et al. 15
Adversarial examples

Adversarial Traffic Signs

Original 80 120

Adversarial 80 120

Classified as: Stop Speed limit (30)
A Case of Irrational Scientific Exuberance

- Underspecified goals
  Big Data cures everything
- Underspecified limitations
  Big Data can do anything (if big enough)
- Underspecified caveats
  Big Data and Big Brother

Wanted: An AI with common decency

- Fair
  no biases
- Accountable
  models can be explained
- Transparent
  decisions can be explained
- Robust
  w.r.t. malicious examples
In practice

▶ Data are ridden with biases

▶ Learned models are biased (prejudices are transmissible to AI agents)

▶ Issues with robustness

▶ Models are used out of their scope

More


Machine Learning: discriminative or generative modelling

Given a training set

\[ \mathcal{E} = \{(x_i, y_i), x_i \in \mathbb{R}^d, i \in [1, n]\} \]

Find

- Supervised learning: \( \hat{h} : X \mapsto Y \) or \( \hat{P}(Y|X) \)
- Generative model \( \hat{P}(X, Y) \)

Predictive modelling might be based on correlations

*If umbrellas in the street, Then it rains*
The implicit big data promise:

If you can predict what will happen, then how to make it happen what you want?

Knowledge $\rightarrow$ Prediction $\rightarrow$ Control

ML models will be expected to support interventions:

- health and nutrition
- education
- economics/management
- climate
Correlations do not support interventions

Causal models are needed to support interventions

*Consumption of chocolate enables to predict # of Nobel prizes but eating more chocolates does not increase # of Nobel prizes*
Discussion

Scientific Caveat
▶ Robustness of results
▶ Reproducibility of results (gigantic resources)

Economic Caveat
▶ Winner take all:

Value $\rightarrow$ Data $\rightarrow$ More Value

Societal Caveat
▶ Learning from biased data $\rightarrow$ carving prejudices in stone
▶ Accurate prediction of individual risks $\rightarrow$ ruins insurance mechanisms.