Panorama du Deep Learning

Stéphane Canu, INSA Rouen – Normandy University

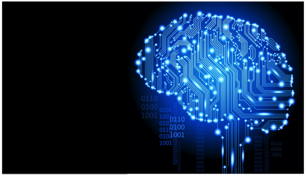
asi.insa-rouen.fr/enseignants/~scanu

Séminaire nantais inter-établissements en Science des Données

Thursday, June 12

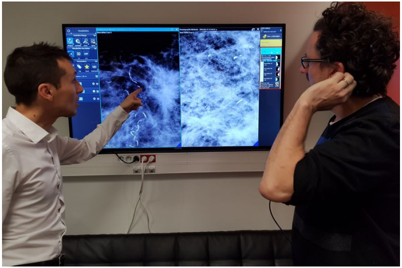
Road map

1. Why deep learning?
2. The first stage: 1890 - 1969
3. The second stage: 1985 - 1995
4. The third stage: 2006 - (2012) - 2018...
5. What's new in deep learning?
   - Big is beautiful
   - Two Hot topics: data and architecture
6. Conclusion

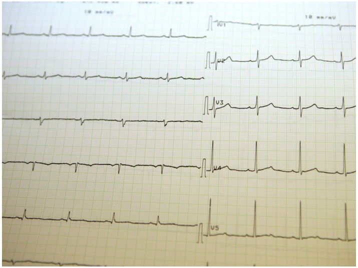



nature
THE INTERNATIONAL WEEKLY JOURNAL OF SCIENCE
Melanoma
LESIONS LEARNT
Artificial intelligence powers detection of skin cancer from images PAGES 36 & 115
NATURE.COM/NATURE
2 February 2017 £10
Vol. 542, No. 7639

New diagnostic tools using AI



Digital mammography reading
Therapixel

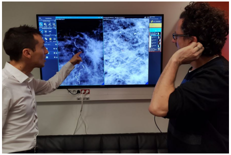


ECG analysis
CardioLogs

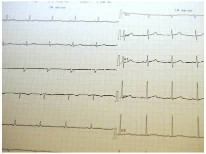
http://www.therapixel.com/
https://cardiologs.com/

Data based statistical programming

Skin cancer classification
130 000 training images
validation error rate : 28 % (human 34 %)



the Digital Mammography DREAM Challenge
640 000 mammographies (1209 participants)
5 % less false positive



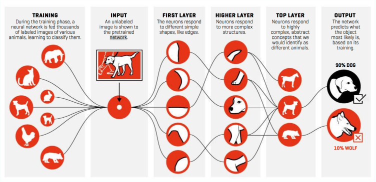
heart rate analysis
500 000 ECG
precision 92.6 % (humain 80.0 %) sensitivity 97 %

Statistical machine learning: retrieving correlations

with deep learning end-to-end architecture
"April showers bring May flowers"

Road map

1. Why deep learning?
2. The first stage: 1890 - 1969
3. The second stage: 1985 - 1995
4. The third stage: 2006 - (2012) - 2018...
5. What's new in deep learning?
    - Big is beautiful
    - Two Hot topics: data and architecture
6. Conclusion



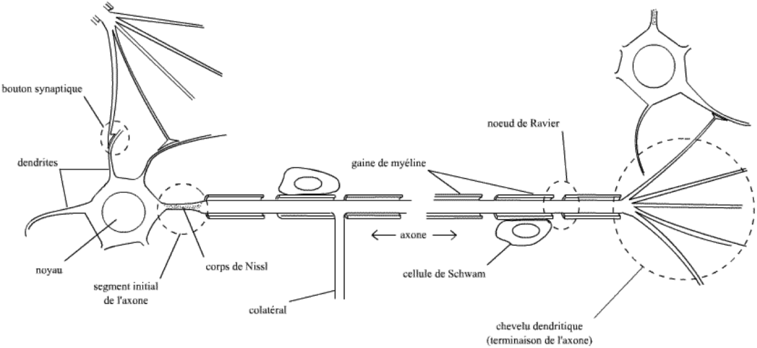
The neural networks time line

- The first stage: 1890 - 1969

$\sim$1890 Ramón y Cajal: the biological neuron
1943 McCulloch & Pitts formal neuron
1949 Hebb's rule
1958 Rosenblatt's Perceptron: learning with stochastic gradient
1969 Minsky & Papert: stop – the 1st NN winter

- The second stage: 1985 - 1995
- The third stage: 2006 - (2012) - 2018...

The biological neuron



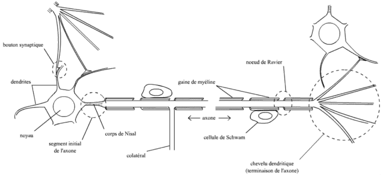
The neural networks time line

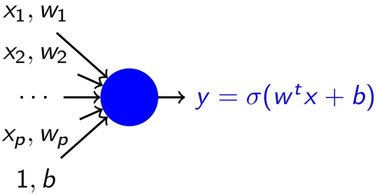
- The first stage: 1890 - 1969

$\sim$1890 Ramón y Cajal: the biological neuron
1943 McCulloch & Pitts formal neuron
1949 Hebb's rule
1958 Rosenblatt's Perceptron: learning with stochastic gradient
1969 Minsky & Papert: stop – the 1st NN winter

- The second stage: 1985 - 1995
- The third stage: 2006 - (2012) - 2018...

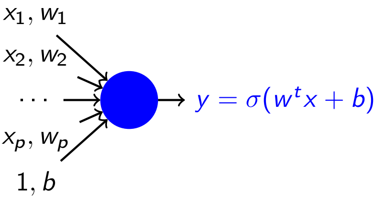
McCulloch & Pitts formal neuron 1943

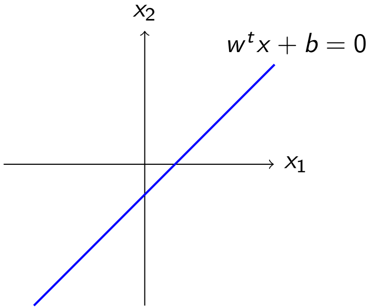




- x input $\in \mathbb{R}^p$
- w weight, b bias
- σ activation function
- y output $\in \mathbb{R}$

The artificial neuron as a linear threshold unit





- x input $\in \mathbb{R}^p$
- w weight, b bias
- a activation, $a = w^t x + b$
- σ activation function
- Φ transfer function
- y output $\in \mathbb{R}$

σ activation function (non linear)

$$\begin{array}{ccc} \mathbb{R} & \mapsto & \mathbb{R} \\ a & \rightarrow & y = \sigma(a) \end{array}$$

Φ transfer function

$$\begin{array}{ccc} \mathbb{R}^p & \mapsto & \mathbb{R} \\ \mathbf{x} & \rightarrow & y = \Phi(\mathbf{x}) = \sigma(w^t x + b) \end{array}$$

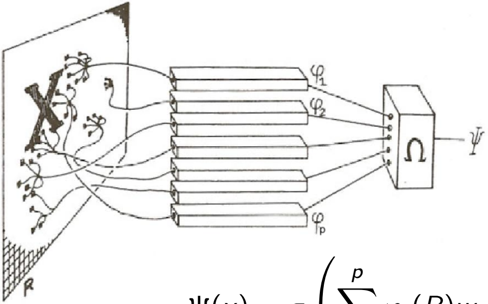
The neural networks time line

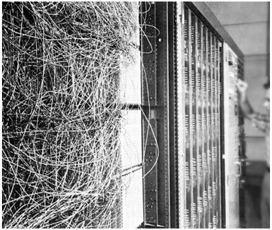
- The first stage: 1890 - 1969

∼1890	Ramón y Cajal: the biological neuron
1943	McCulloch & Pitts formal neuron
1949	Hebb's rule
1958	Rosenblatt's Perceptron: learning with stochastic gradient
1969	Minsky & Papert: stop – the 1st NN winter

- The second stage: 1985 - 1995
- The third stage: 2006 - (2012) - 2018...

The formal neuron as a learning machine: fit the w





$$\Psi(x) = \sigma\left(\sum_{j=1}^{p} \varphi_j(R) w_j + b\right)$$

Rosenblatt's Perceptron, 1958 (Widrow & Hoff's Adaline, 1960)

given n pairs of input–output data $\mathbf{x}_i = \varphi_j(R_i), t_i, i = 1, n$

find w such that $\underbrace{\sigma(\mathrm{w}^t \mathbf{x}_i)}_{\text{prediction of the model}} = \underbrace{t_i}_{\text{ground truth}}$

Cost minimization (energy-based model)

Minimize a loss	$\min_{\mathrm{w} \in \mathbb{R}^{p+1}} \sum_{i=1}^{n} loss(\mathrm{w})$	$loss(\mathrm{w}) = \big(\sigma(\mathrm{w}^t \mathbf{x}_i) - t_i\big)^2$
Gradient descent	$\mathrm{w} \leftarrow \mathrm{w} - \rho \mathbf{d}$	$\mathbf{d} = \sum_{i=1}^{n} \nabla_{\mathrm{w}} loss(\mathrm{w})$
Stochastic gradient		$\mathbf{d} = \nabla_{\mathrm{w}} loss(\mathrm{w})$

Algorithm 1 Gradient epoch

Data: w initialization, ρ stepsize
Result: w
for *i=1,n* **do**
  $\mathbf{x}_i, t_i \leftarrow$ pick a point i
  $\mathbf{d} \leftarrow d + \nabla_{\mathrm{w}} loss(\mathrm{w}, x_i, t_i)$
end
$\mathrm{w} \leftarrow \mathrm{w} - \rho \mathbf{d}$

Algorithm 2 Stochastic gradient

Data: w initialization, ρ stepsize
Result: w
for *i=1,n* **do**
  $\mathbf{x}_i, t_i \leftarrow$ pick a point i
  $\mathbf{d} \leftarrow \nabla_{\mathrm{w}} loss(\mathrm{w}, x_i, t_i)$
  $\mathrm{w} \leftarrow \mathrm{w} - \rho \mathbf{d}$
end

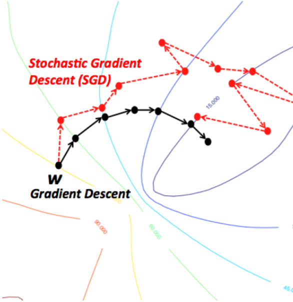
Accelerating the stochastic gradient

- stochastic average (mini batch)
  - parameters (Polyak and Juditsky, 1992)
  - gradients SAG-A, (Le Roux et al 2012)
  - variance reduction (Johnson, Zhang, 2013)

- convergence acceleration
  - Nesterov's method (1983)
  - momentum (heuristic)

- acceleration and averaging
  - (Dieuleveut, Flammarion & Bach, 2016)

- stepsize adaptation
  - RMSprop (Tieleman & Hinton, 2012)
  - Adaptive Moment Estimation – ADAM (Kingma & Ba, 2015)
  - AMSGRAD (Reddi et al, BPA ICRL 2018)



http://ruder.io/optimizing-gradient-descent/ https://wikidocs.net/3413

Optimization methods for large-scale machine learning L Bottou, FE Curtis, J Nocedal - stat, 2017 -

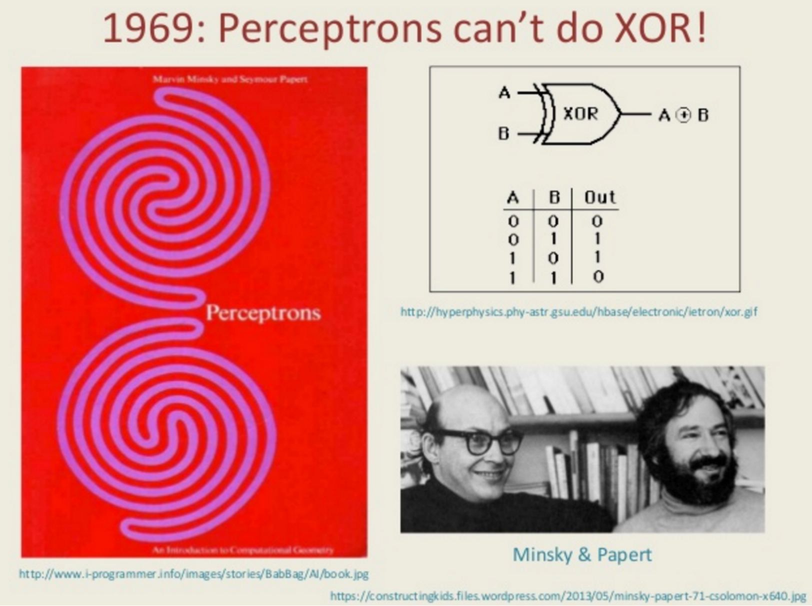
The neural networks time line

- The first stage: 1890 - 1969

$\sim$1890 Ramón y Cajal: the biological neuron
1943 McCulloch & Pitts formal neuron
1949 Hebb's rule
1958 Rosenblatt's Perceptron: learning with stochastic gradient
1969 Minsky & Papert: stop – the 1st NN winter

- The second stage: 1985 - 1995
- The third stage: 2006 - (2012) - 2018...

However, linear neurons ar linear



Perceptrons limitations

- Linear threshold units as Boolean gates
- Circuit theory is poorly known
- Learning deep circuits means solving the credit assignment pb
- Linearly separable problems are few
- Elementary problems need complex circuits. (parity, x-or...)
- But have simple algorithmic solutions
  $\rightarrow$programming versus learning

Abandon perceptrons and other analog computers

Develop symbolic computers and symbolic AI techniques.

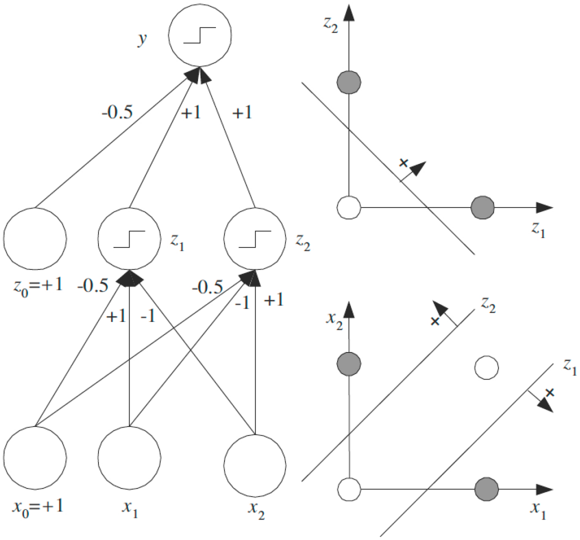
The neural networks time line

- The first stage: 1890 - 1969
- The second stage: 1985 - 1995

  1985 Rumelhart, Hinton & Williams; Le Cun: go - backpropagation
  1989 Universal Approximation Cybenko-Hornik-Funahashi Theorem
  1989 Y. Le Cun's convolutional neural networks
  1995 Recurent neural networks, LSTM
  1995 SVM
  2004 Caltech 101: the 2nd NN winter

- The third stage: 2006 - (2012) - 2018...

Non linearity combining linear neurons: the Xor case



Alpaydın, Introduction to Machine Learning, 2010

Neural networks

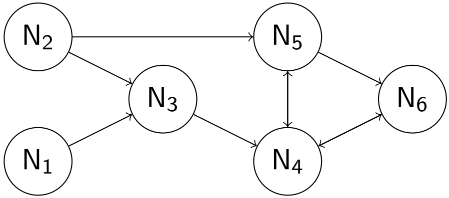
Definition: Neural network

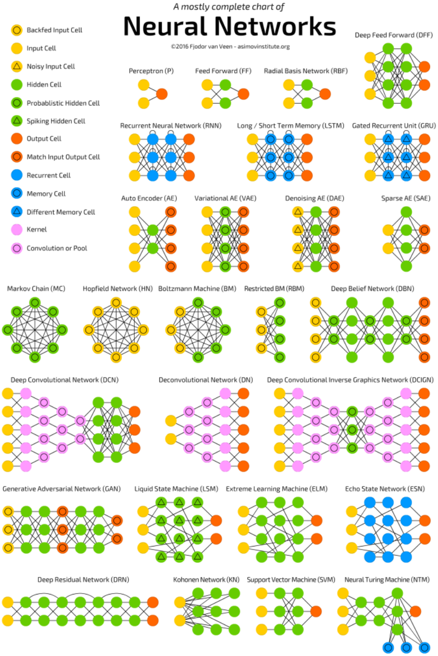
A neural network is an oriented graph of formal neurons

When two neurons are connected (linked by an oriented edge of the graph), the output of the head neuron is used as an input by the tail neuron. It can be seen as a weighted directed graph

3 different neurons are considered:

- input neurons (connected with the input)
- output neurons
- hidden neurons





The Asimov Institute: `http://www.asimovinstitute.org/neural-network-zoo/`

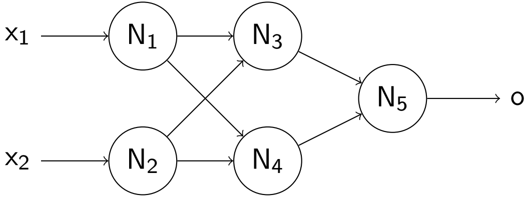
The Multiplayred peceptron (MLP)

Definition: Multiplayred peceptron

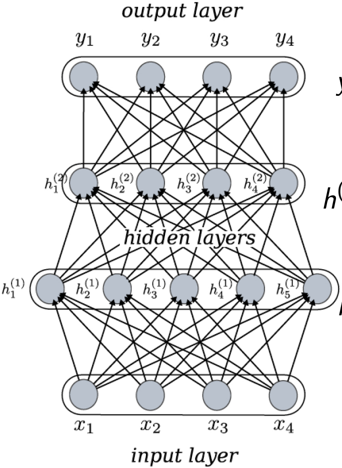
A Multiplayred peceptron is an acyclic neural network,

where the neurons are structued in successive layers, begining by an input layer and finishing with an output layer.

Example: The X-or neural network is a MLP with a single hidden unit with 2 hidden neurons.



Neural networks propagation and back propagation



$$y = \varphi(W_3 h^{(2)})$$

$$\uparrow$$

$$h^{(2)} = \varphi(W_2 h^{(1)})$$

$$\uparrow$$

$$h^{(1)} = \varphi(W_1 \mathbf{x})$$

$$\uparrow$$

$$\mathbf{x}$$

$$\nabla_{W_3} J = (y - y_a)\varphi'(W_3 h^{(2)}) h^{(2)}$$

$$\downarrow$$

$$\nabla_{W_2} J = \nabla_{h^{(2)}} J \varphi'(W_2 h^{(1)}) h^{(1)}$$

$$\downarrow$$

$$\nabla_{W_1} J =$$

backpropagation = chain rule (autodiff)

Used to learn internal representation W_1, W_2, W_3

Back propagation is differential learning

Yann LeCun

5 janvier ·

OK, Deep Learning has outlived its usefulness as a buzz-phrase.
Deep Learning est mort. Vive Differentiable Programming!

Numpy

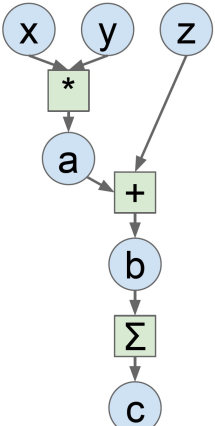
```
import numpy as np
np.random.seed(0)

N, D = 3, 4

x = np.random.randn(N, D)
y = np.random.randn(N, D)
z = np.random.randn(N, D)

a = x * y
b = a + z
c = np.sum(b)

grad_c = 1.0
grad_b = grad_c * np.ones((N, D))
grad_a = grad_b.copy()
grad_z = grad_b.copy()
grad_x = grad_a * y
```



The neural networks time line

- The first stage: 1890 - 1969
- The second stage: 1985 - 1995

  1985 Rumelhart, Hinton & Williams; Le Cun: go - backpropagation
  1989 Universal Approximation Cybenko-Hornik-Funahashi Theorem
  1989 Y. Le Cun's convolutional neural networks
  1995 Recurent neural networks, LSTM
  1995 SVM
  2004 Caltech 101: the 2nd NN winter

- The third stage: 2006 - (2012) - 2018...

MLP with one hidden layer as universal approximator

Universal approximation theorem for MLP

- given any $\varepsilon > 0$
- for any continuous function f on compact subsets of $\mathbb{R}^p$
- for any admissible activation function φ (not a polynomial)
- there exists h, $W_1 \in \mathbb{R}^{p \times h}$, $b \in \mathbb{R}^h$, $c \in \mathbb{R}$ and $w_2 \in \mathbb{R}^h$ such that

$$\|f(x) - w_2\varphi(W_1 x + b) + c\|_\infty \leq \varepsilon$$

SVM and random forest also

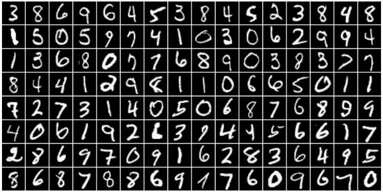
The neural networks time line

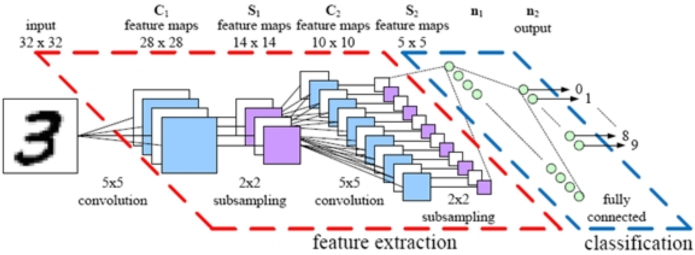
- The first stage: 1890 - 1969
- The second stage: 1985 - 1995

  1985 Rumelhart, Hinton & Williams; Le Cun: go - backpropagation
  1989 Universal Approximation Cybenko-Hornik-Funahashi Theorem
  1989 Y. Le Cun's convolutional neural networks
  1995 Recurent neural networks, LSTM
  1995 SVM
  2004 Caltech 101: the 2nd NN winter

- The third stage: 2006 - (2012) - 2018...

OCR: MNIST database (LeCun, 1989)





use convolution layers

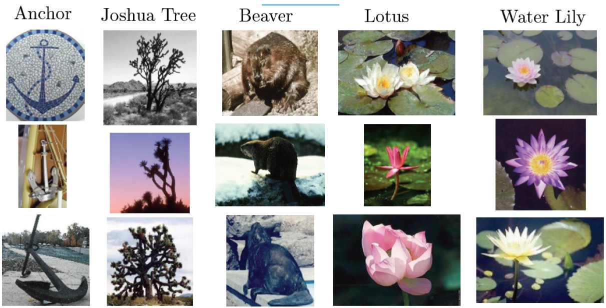
The neural networks time line

- The first stage: 1890 - 1969
- The second stage: 1985 - 1995

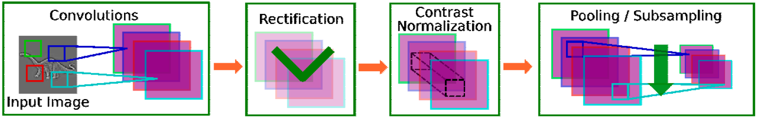
  1985 Rumelhart, Hinton & Williams; Le Cun: go - backpropagation
  1989 Universal Approximation Cybenko-Hornik-Funahashi Theorem
  1989 Y. Le Cun's convolutional neural networks
  1995 Recurent neural networks, LSTM
  1995 SVM
  2004 Caltech 101: the NN winter

- The third stage: 2006 - (2012) - 2018...

The caltech 101 database (2004)



- 101 classes,
- 30 training images per category
- ...and the winner is NOT a deep network
  - dataset is too small



use convolution + Recitification + Normalization + Pooling

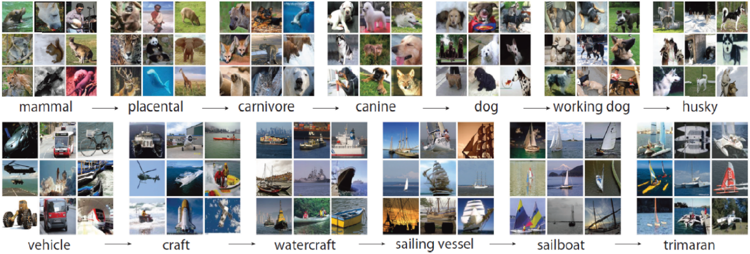
in What is the Best Multi-Stage Architecture for Object Recognition? Jarrett et al, 2009

The neural networks time line

- The first stage: 1890 - 1969
- The second stage: 1985 - 1995
- The third stage: 2006 - (2012) - 2018...

2006 Deep learning: Bengio's, Hinton's RBM, Y LeCun's proposals
2010 Andrw Ng's GPU for Deep GPU
2011 Deep frameworks, tools (theano, torch, cuda-convnet...)
2012 ImageNet – AlexNet
2013 M. Zuckerberg at NIPS the deep fashion
2014 Representation learning fine tuning
2015 Deep learning in the industry: speech, traduction, image...
2016 Goodfellow's generative adversarial networks (GAN)
2017 Reinforcement learning: Deep win's GO
2018 Automatic design, adversarial defense, green learning, theory...

The image net database (Deng et al., 2012)



ImageNet = 15 million high-resolution images of 22,000 categories.
Large-Scale Visual Recognition Challenge (a subset of ImageNet)

- 1000 categories.
- 1.2 million training images,
- 50,000 validation images,
- 150,000 testing images.

www.image-net.org/challenges/LSVRC/2012/

A new fashion in image processing

2012 Teams	%error
Supervision (Toronto)	15.3
ISI (Tokyo)	26.1
VGG (Oxford)	26.9
XRCE/INRIA	27.0
UvA (Amsterdam)	29.6
INRIA/LEAR	33.4

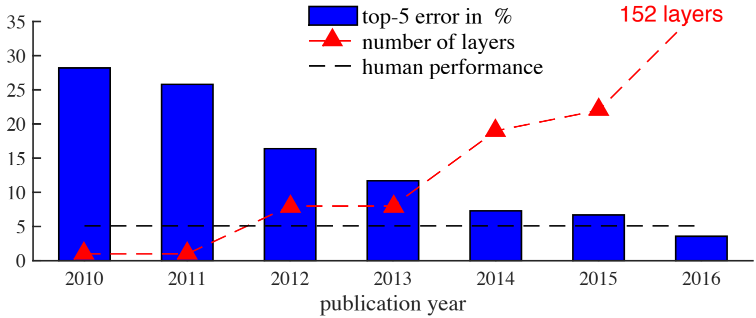
2013 Teams	%error
Clarifai (NYU spinoff)	11.7
NUS (singapore)	12.9
Zeiler-Fergus (NYU)	13.5
A. Howard	13.5
OverFeat (NYU)	14.1
UvA (Amsterdam)	14.2
Adobe	15.2
VGG (Oxford)	15.2
VGG (Oxford)	23.0

2014 Teams	%error
GoogLeNet	6.6
VGG (Oxford)	7.3
MSRA	8.0
A. Howard	8.1
DeeperVision	9.5
NUS-BST	9.7
TTIC-ECP	10.2
XYZ	11.2
UvA	12.1

shallow approaches

deep learning

ImageNet results



2012 Alex Net

2013 ZFNet

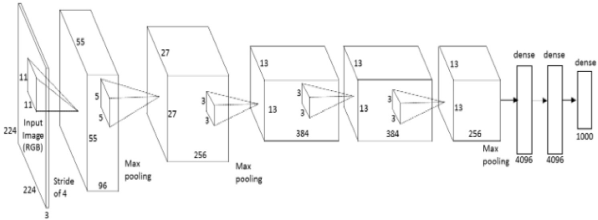
2014 VGG

2015 GoogLeNet / Inception

2016 Residual Network

karpathy's blog: `karpathy.github.io/2014/09/02/what-i-learned-from-competing-against-a-convnet-on-imagenet/`

Deep architecture for image net (15%)

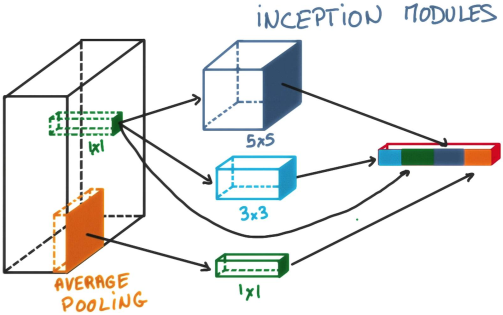


The *Alex Net* architecture [Krizhevsky, Sutskever, Hinton, 2012]

Convolution + Recitification (ReLU) + Normalization + Pooling

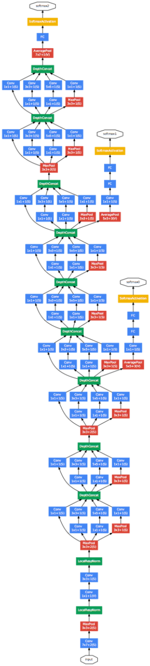
- 60 million parameters
- using 2 GPU – 6 days
- regularization
  - data augmentation
  - dropout
  - weight decay

From 15% to 7%: Inceptionism

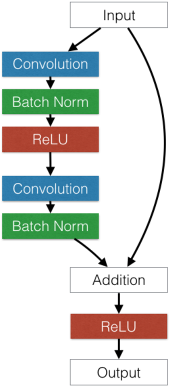


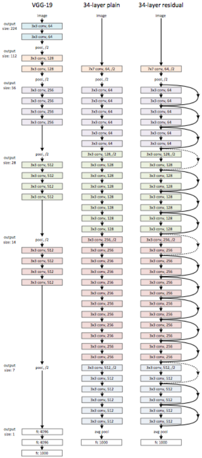
Network in a network (deep learning lecture at Udacity)

Christian Szegedy et. al. Going deeper with convolutions. CVPR 2015.



From 7% to 3%: Residual Nets





Beating the gradient vanishing effect

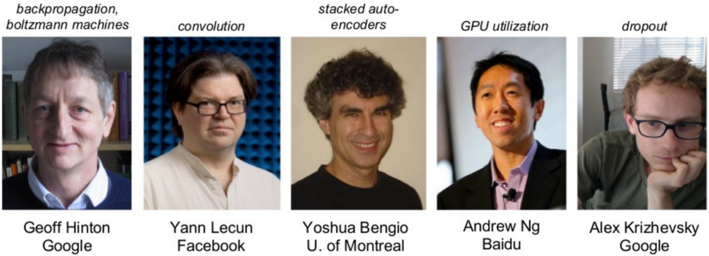
K. He *et al*, 2016

The neural networks time line

- The first stage: 1890 - 1969
- The second stage: 1985 - 1995
- The third stage: 2006 - (2012) - 2018...

2006	Deep learning: Bengio's, Hinton's RBM, Y LeCun's proposals
2010	Andrw Ng's GPU for Deep GPU
2011	Deep frameworks, tools (theano, torch...)
2012	ImageNet – AlexNet
2013	M. Zuckerberg at NIPS: the deep fashion
2014	Representation learning fine tuning
2015	Deep learning in the industry: speech, traduction, image...
2016	Goodfellow's generative adversarial networks (GAN)
2017	Reinforcement learning: Deep win's GO
2018	Automatic design, adversarial defense, green learning, theory...

Deep learning, AI and the industry



- data science, artificial intelligence and deep learning
- the GAFAM – BATX vision
  - they got the infrastructure (hard+software)
  - they got the data
  - deep learning bridges the gap between applications and ML

In 2016, Google Chief Executive Officer (CEO) Sundar Pichai said,

Machine learning [a subfield of AI] is a core, transformative way by which we're rethinking how we're doing everything.

Road map

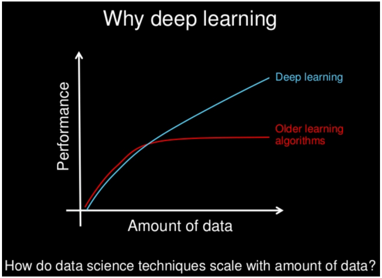
1 Why deep learning?

2 The first stage: 1890 - 1969

3 The second stage: 1985 - 1995

4 The third stage: 2006 - (2012) - 2018...

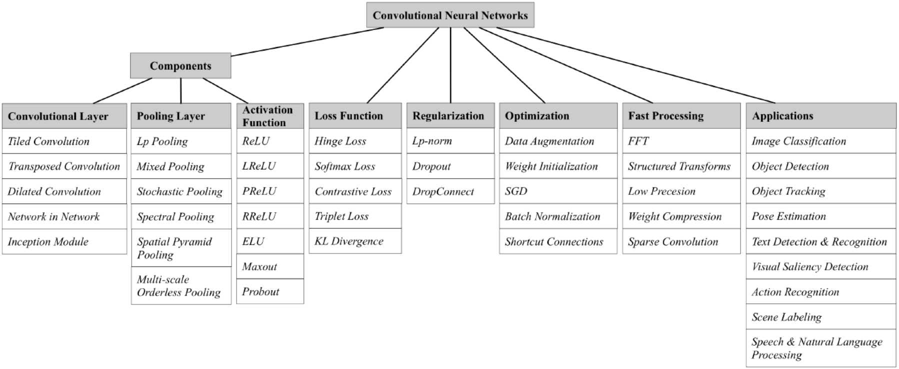
5 What's new in deep learning?
- Big is beautiful
- Two Hot topics: data and architecture



6 Conclusion

What's new with deep learning

- a lot of data (big data)
- big computing resources (hardware & software),
- big model (deep vs. shalow)
  - → new architectures
  - → new learning tricks

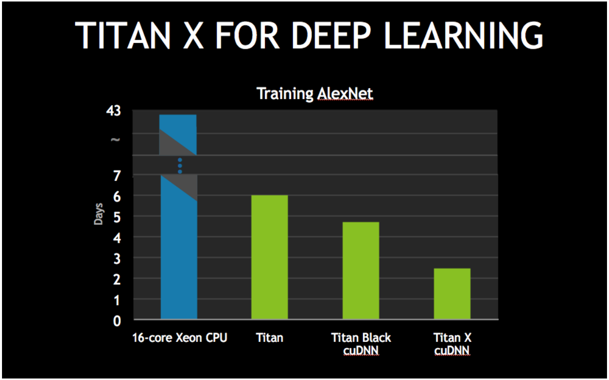


from Recent advances in convolutional neural networks Gu et al. Pattern Recognition, 2017

Big data: a lot of available training data

- ImageNet: 1,200,000x256x256x3 (about 200GB) block of pixels
- MS COCO for supervised learning
  - Multiple objects per image
  - More than 300,000 images
  - More than 2 Million instances
  - 80 object categories
  - 5 captions per image
- YFCC100M for unsupervised learning
- Google Open Images, 9 million URLs to images annotated over 6000 categories
- Visual genome: data + knowledge `http://visualgenome.org/`

Big computers: GPU needed



Now 2 hours with Nvidia DGX-1, and enough Memory

Yann LeCun:

learning a relevant model takes 3 weeks

big software: deep learning frameworks

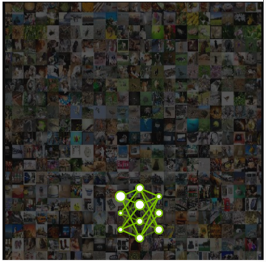
	Languages	Tutorials and training materials	CNN modeling capability	RNN modeling capability	Architecture: easy-to-use and modular front end	Speed	Multiple GPU support	Keras compatible
Theano	Python, C++	++	++	++	+	++	+	+
Tensor-Flow	Python	+++	+++	++	+++	++	++	+
Torch	Lua, Python (new)	+	+++	++	++	+++	++	
Caffe	C++	+	++		+	+	+	
MXNet	R, Python, Julia, Scala	++	++	+	++	++	+++	
Neon	Python	+	++	+	+	++	+	
CNTK	C++	+	+	+++	+	++	+	

Tensoflow (Google) is the most popular with Keras. Pytorch is a chalenger.

http://www.kdnuggets.com/2017/03/getting-started-deep-learning.html

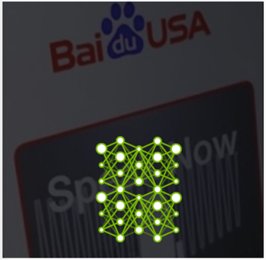
Big architectures

7 ExaFLOPS
60 Million Parameters



2015 - Microsoft ResNet
Superhuman Image Recognition

20 ExaFLOPS
300 Million Parameters



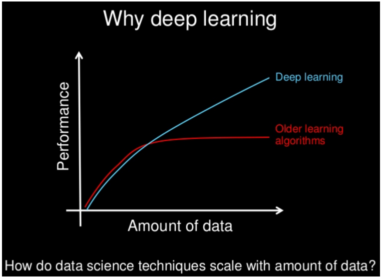
2016 - Baidu Deep Speech 2
Superhuman Voice Recognition

100 ExaFLOPS
8700 Million Parameters

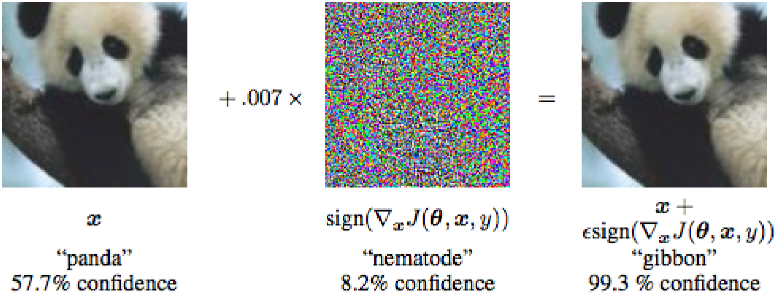
2017 - Google Neural Machine Translation
Near Human Language Translation

Road map

1. Why deep learning?
2. The first stage: 1890 - 1969
3. The second stage: 1985 - 1995
4. The third stage: 2006 - (2012) - 2018...
5. What's new in deep learning?
   - Big is beautiful
   - Two Hot topics: data and architecture
6. Conclusion



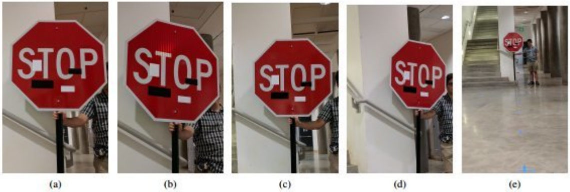
Deep neural networks are easily fooled (1/2)



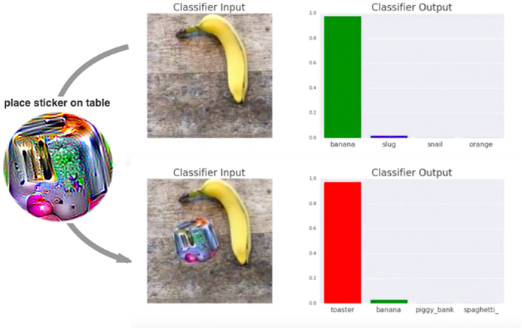
Explaining and Harnessing Adversarial Examples, Ian J. Goodfellow, Jonathon Shlens, Christian Szegedy, 2015

https://arxiv.org/abs/1412.6572

Adversarial examples (2/2)

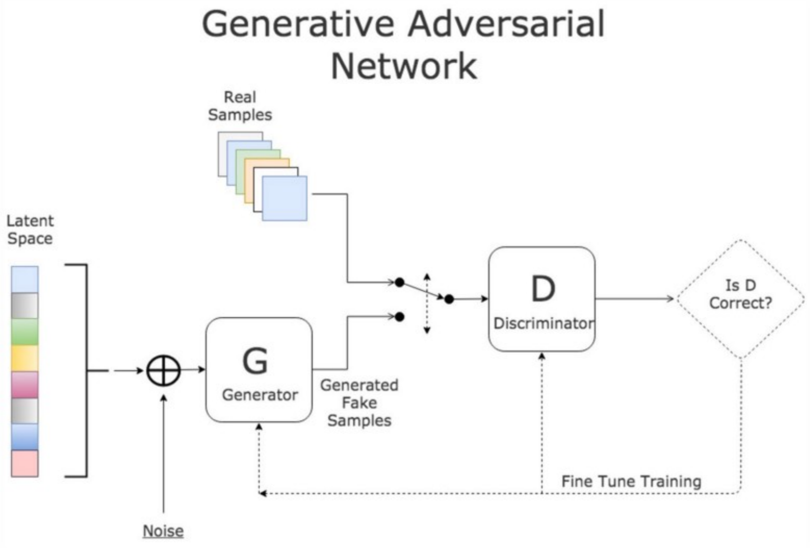


Adversarial Examples for Evaluating Reading Comprehension Systems, Robin Jia, Percy Liang, 2017



Adversarial Patch Tom B. Brown, Dandelion Mané, Aurko Roy, Martin Abadi, Justin Gilmer, 2017

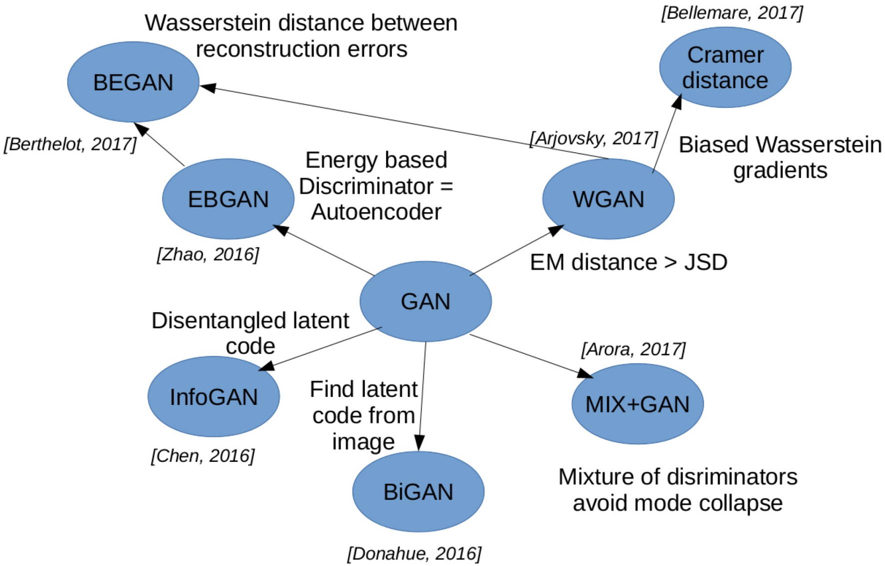
Generative models



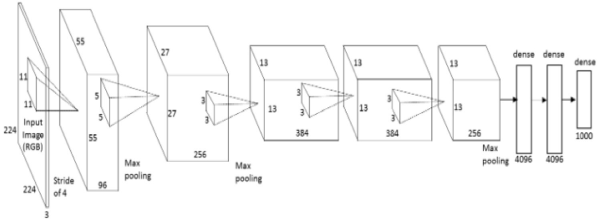
https://www.datasciencecentral.com/profiles/blogs/generative-adversarial-networks-gans-engine-and-applications

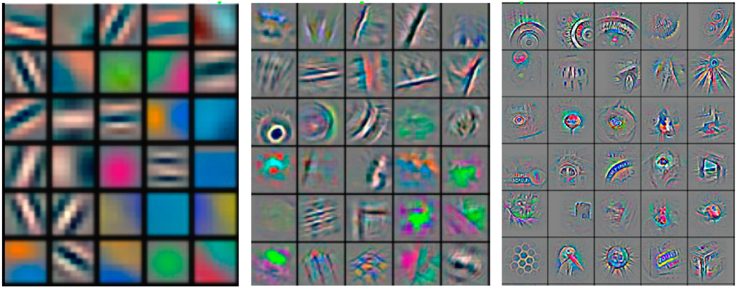
Defense-GAN: Protecting classifiers against adversarial attacks using generative models, 2018

Other Generative architectures



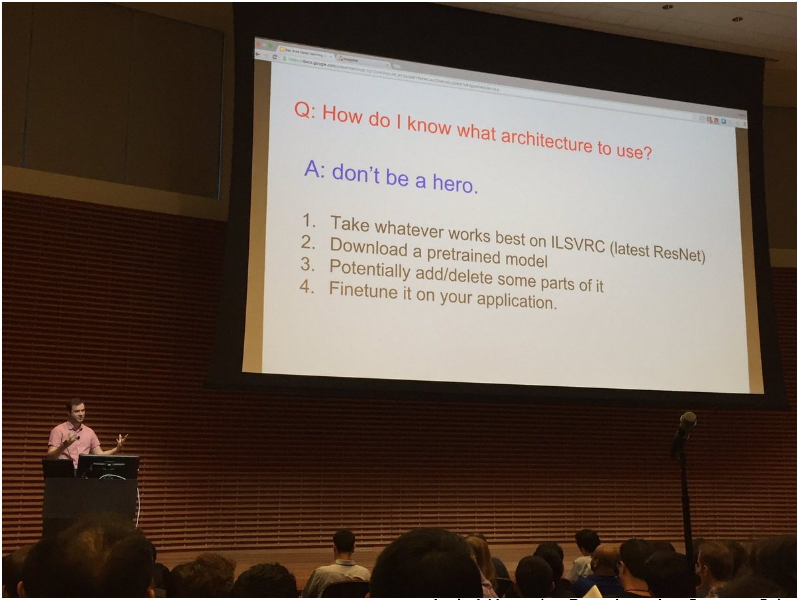
AlexNet works because of learning internal representation





Feature visualization of convolutional net trained on ImageNet from [Zeiler & Fergus 2013]

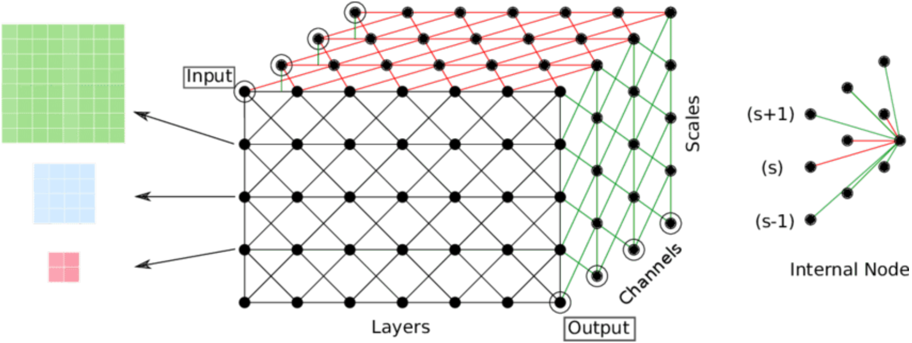
How to start with deep learning?



Andrej Karpathy, Deep Learning Summer School 2016

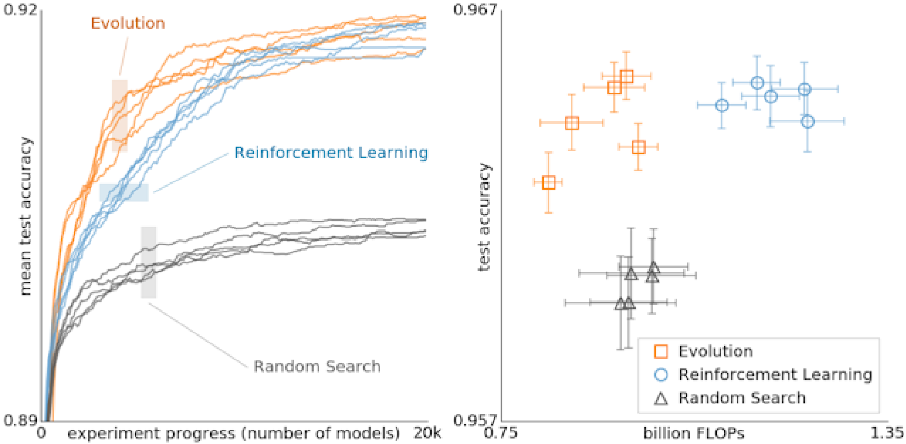
Convolutional Neural Fabrics

- problem: how to find the most relevant architecture
- todays solution: try and test
- A new solution: learn the architecture



Convolutional Neural Fabrics, Saxena and Verbeek, NIPS 2016

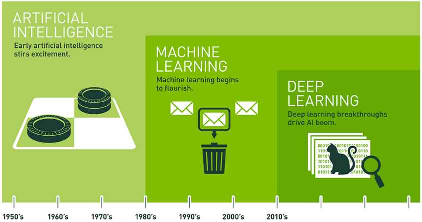
Neural Architecture Search



Regularized Evolution for Image Classifier Architecture Search, E. Real et al, 2018
https://chinagdg.org/2018/03/using-evolutionary-automl-to-discover-neural-network-architectures/

Road map

1. Why deep learning?
2. The first stage: 1890 - 1969
3. The second stage: 1985 - 1995
4. The third stage: 2006 - (2012) - 2018...
5. What's new in deep learning?
   - Big is beautiful
   - Two Hot topics: data and archit
6. Conclusion



Since an early flush of optimism in the 1950s, smaller subsets of artificial intelligence – first machine learning, then deep learning, a subset of machine learning – have created ever larger disruptions.

The deep learning time line

- The first stage: 1890 - 1969
  - learning is optimization with stochastic gradient (to scale)
- The second stage: 1985 - 1995
  - NN are universal approximator differentiable graphs (that scales)
- The third stage: 2006 - (2012) - 2018...
  - scale with big data+computers+architecture (deep)
- Open issues
  - provide guaranties: adversarial examples and representation learning
  - architecture design (autoML)
  - theory needed
  - do more with less: green learning

To go further

- books
  - I. Goodfellow, Y. Bengio & A. Courville, *Deep Learning*, MIT Press book, 2016 http://www.deeplearningbook.org/
  - Gitbook leonardoaraujosantos.gitbooks.io/artificial-inteligence/
- conferences
  - NIPS, ICLR, xCML, AIStats,
- Journals
  - JMLR, Machine Learning, Foundations and Trends in Machine Learning, machine learning survey http://www.mlsurveys.com/
- lectures
  - Deep Learning: Course by Yann LeCun at Collège de France in 2016 college-de-france.fr/site/en-yann-lecun/inaugural-lecture-2016-02-04-18h00.htm
  - Convolutional Neural Networks for Visual Recognition (Stanford)
  - deep mind (https://deepmind.com/blog/)
  - CS 229: Machine Learning at stanford Andrew Ng
- Blogs
  - Andrej Karpathy blog (http://karpathy.github.io/)
  - http://deeplearning.net/blog/
  - https://computervisionblog.wordpress.com/category/computer-vision/

To go further and enjoy: june in Rouen