Neural Networks and Deep Learning: Modeling Capacities

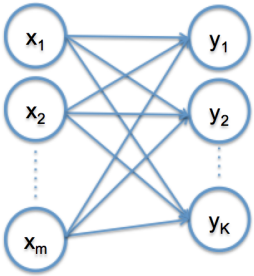
Nicolas Thome

Conservatoire National des Arts et Métiers (Cnam)
Département Informatique

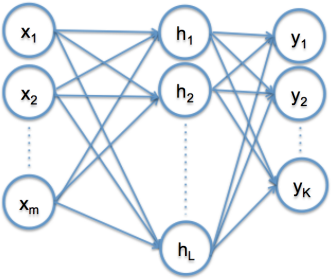
Perceptron: (non-)linear Boundaries

Neural Networks for Classification

- Logistic Regression: limited to linear decision boundaries
- Multi-Layer Perceptron (MLP): non-linear decision functions



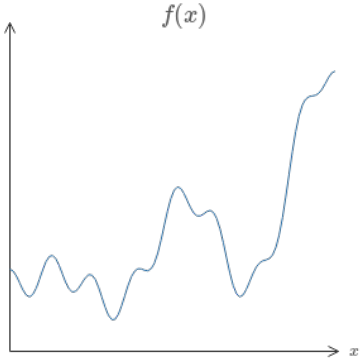
Logistic Regression



MLP

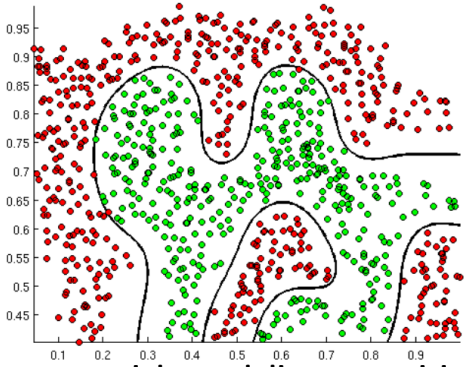
MLP: Universal Function Approximators

- **Neural network with one single hidden layer ⇒ universal approximator**
  - Can represent any function on compact subsets of $\mathbb{R}^n$ [Cybenko, 1989]
    - Approximate any continuous function to any desired precision
  - Ex pour regression: any function can be interpolated



MLP: Universal Function Approximators

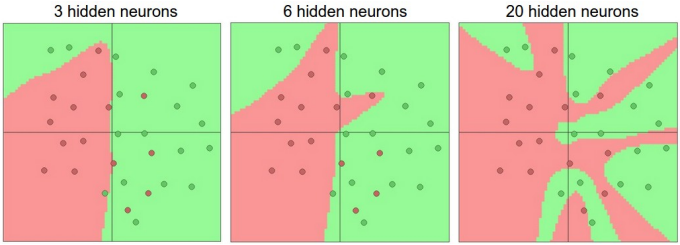
- **Neural network with one single hidden layer ⇒ universal approximator**
  - Can represent any function on compact subsets of $\mathbb{R}^n$ [Cybenko, 1989]
  - Ex pour classification: any decision boundaries can be expressed



⇒ **very rich modeling capacities**

MLP: Universal Function Approximators

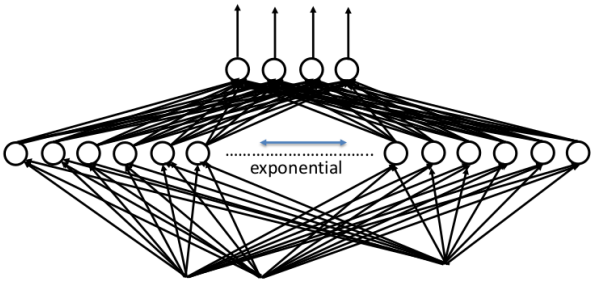
- **2 layers, *i.e.* one hidden layer, is enough**



- **Challenge is NOT fitting training data**
  - Simple models already have very large (infinite) modeling power
- **Challenge: optimization, overfitting**

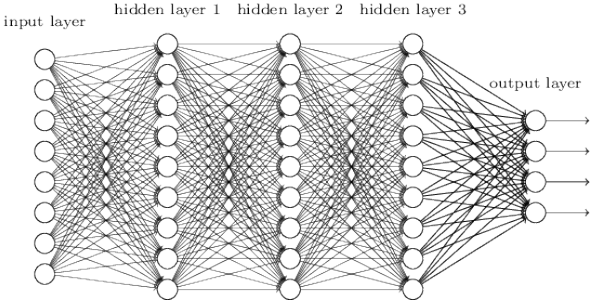
MLP: Universal Function Approximators

- **2 layers, *i.e.* one hidden layer, is enough ... theoretically:**
  - **BUT:** exponential number of hidden units [Barron, 1993]



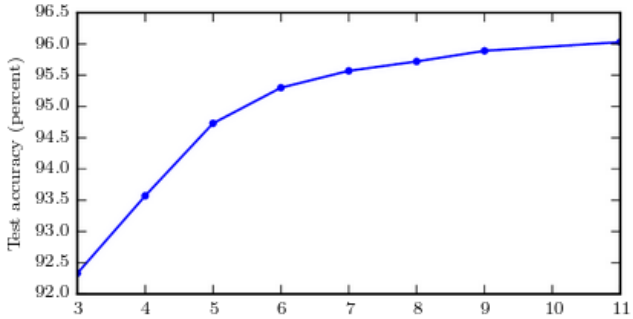
Deep Models: Universal Function Approximators

- **Deeper Models: less units required to represent the desired function**
  - Functions representable compactly with k layers may require exponentially size with $k-1$ layers [Hastad, 1989, Bengio, 2009]



  - Same modeling power, fewer parameters
    $\Rightarrow$ **better generalization!**

Deep Models

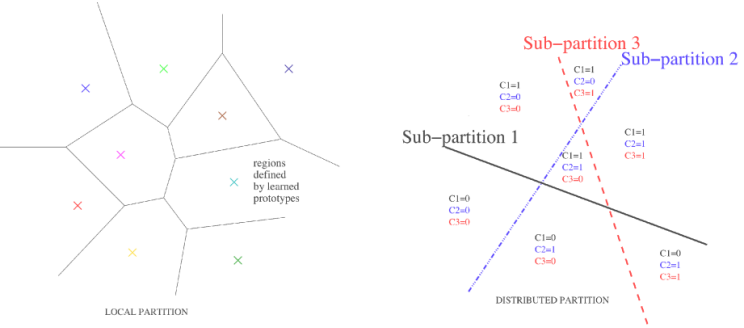


Depth improves generalization: multi-digit recognition, from [Goodfellow et al., 2016]

Local *vs* Distributed Representations

- Local Representations: one neuron ↔ one concept
- Deep Learning ⇒ Distributed Representations:
  - Each concept ↔ many neurons, each neuron ↔ many concepts

⇒ **Exponentially more efficient than local representations**



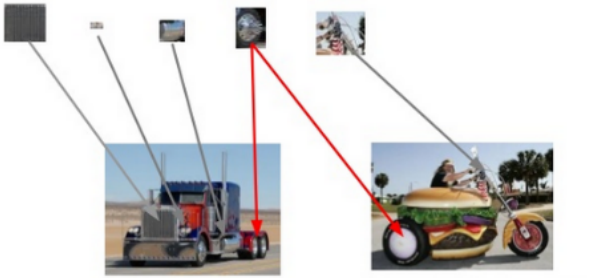
From [Bengio and Delalleau, 2011]

Deep Learning & Distributed Representations

- DL architectures: distributed representations shared across classes

[1 1 0 0 0 1 0 **1** 0 0 0 0 1 1 0 1...] motorbike

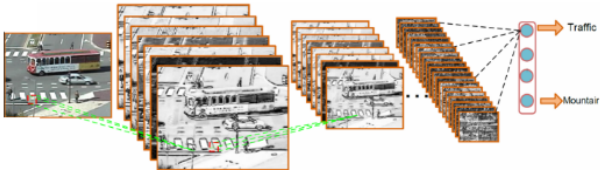
[0 0 1 0 0 0 0 **1** 0 0 1 1 0 0 1 0...] truck



Credit: M.A Ranzato

Deep Models: Conclusion

- Neural Networks: very large modeling capacities
  - Simple (shallow) models, *i.e.* MLP: universal approximators
  - Deeper models: same representation power with more layers but fewer parameters
  - Local *vs* distributed representations
- **Representation Learning with ConvNets**
  ⇒ following!



References I


Barron, A. R. (1993).
Universal approximation bounds for superpositions of a sigmoidal function.
Information Theory, IEEE Transactions on, 39(3):930–945.


Bengio, Y. (2009).
Learning deep architectures for ai.
Found. Trends Mach. Learn., 2(1):1–127.


Bengio, Y. and Delalleau, O. (2011).
On the expressive power of deep architectures.
In *Proceedings of the 22Nd International Conference on Algorithmic Learning Theory*, ALT'11, pages 18–36, Berlin, Heidelberg. Springer-Verlag.


Cybenko, G. (1989).
Approximation by superpositions of a sigmoidal function.
Mathematics of control, signals and systems, 2(4):303–314.


Goodfellow, I., Bengio, Y., and Courville, A. (2016).
Deep Learning.
MIT Press.
http://www.deeplearningbook.org.


Hastad, J. (1989).
Almost optimal lower bounds for small depth circuits.
In *RANDOMNESS AND COMPUTATION*, pages 6–20. JAI Press.