Learning the human world with Deep Belief Networks

Jérôme Louradour

with Yoshua Bengio, Olivier Breuleux, Daniel Cernea, Aaron Courville, Olivier Delalleau, Dumitru Erhan, Pascal Lamblin, Marina Sokolava, ...

LISA

The ‘baby AI’ project

General presentation

Towards the goal of artificial intelligence...

Make the machine learn with minimal “engineering intervention”
(hardcoded rules, task-specific heuristics, ...)

How can we hope to perform well?

Feed well-established algorithms with cheap data (TV, video)
“cheap” = unlabeled, simulated, ...

Why to focus on Deep Belief Networks?

1. Exploit *huge amounts* of *unlabeled* data...
   ...to generalize well with *few labeled* data (specific tasks)
2. Gradual learning: first simple concepts, then + and + abstract
3. Multi-modality (image, text, audio)

The 'baby AI' project

Scientific goals

- **Semi-supervised learning**
  Master the unsupervised learning.
- **Gradual learning**
  As for children, the learning process must be more efficient with a good *curriculum*.
  (show simple examples first, then more complicated ones)
- **Multi-modality**
  "Multi-path" DBN + encourage RBMs to be mutually predictive.
- **Dynamic aspect**
  Temporal RBM (James Bergstra).
- etc.

The 'baby AI' project

Scientific goals

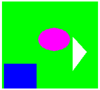
- **Semi-supervised learning**
  Master the unsupervised learning.

- **Gradual learning**
  As for children, the learning process must be more efficient with a good *curriculum*.
  (show simple examples first,
  then more complicated ones)

- **Multi-modality**
  "Multi-path" DBN + encourage RBMs to be mutually predictive.

- **Dynamic aspect**
  Temporal RBM (James Bergstra).

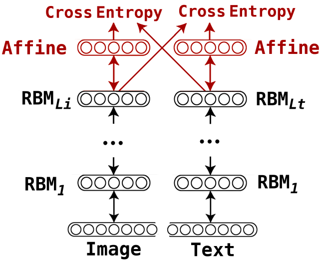
- etc.



The ‘baby AI’ project

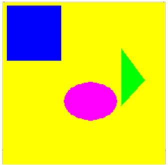
Scientific goals

- **Semi-supervised learning**
  Master the unsupervised learning.
- **Gradual learning**
  As for children, the learning process must be more efficient with a good *curriculum*.
  (show simple examples first,
  then more complicated ones)
- **Multi-modality**
  “Multi-path” DBN + encourage RBMs to be mutually predictive.
- **Dynamic aspect**
  Temporal RBM (James Bergstra).
- etc.



The ‘baby AI’ project

A first step



Topic	Question given to the computer	Answer
Color	There is a small triangle. What color is it?	Green
Shape	What is the shape of the green object?	Triangle
Location	Is the blue square at the top or at the bottom?	At the top
Size	There is a triangle on the right.	
	Is it rather small or bigger?	Small
Size (relative)	Is the square smaller or bigger than the triangle?	Bigger

The ‘baby AI’ project

A first step: preliminary results

Topic	Chance Error Rate (%)	Baseline Error Rate (%) 1 object	2 objects (relative attributes)	3 objects (relative attributes)
color	25	10	40	55
shape	66	50	55	60
size	50	0.5	25	30
location	50	5	35	40

- Results easily degrades when adding sources of variability.
- *Image part:* Shapes are very hard to capture (translation + rotation)
- *Textual part:* No problem to understand the topic.
  But when several objects, hard to guess the object of interest.

The 'baby AI' project

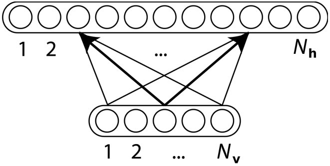
A first step: preliminary results

Topic	Chance Error Rate (%)	Baseline Error Rate (%) 1 object	2 objects (relative attributes)	3 objects (relative attributes)
color	25	10	40	55
shape	66	50	55	60
size	50	0.5	25	30
location	50	5	35	40

- Results easily degrades when adding sources of variability.
- *Image part:* Shapes are very hard to capture (translation + rotation)
- *Textual part:* No problem to understand the topic.
  But when several objects, hard to guess the object of interest.

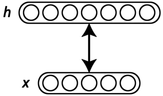
First attempts with DBN: not much better than shallow architecture...
So we came back to basics

Generative learning of RBM



- Real criterion = emperical likelihood (on unlabeled data)
- Practical limitation: complexity $O(2^{min(N_h,N_v)})$
  $\hookrightarrow$ only for small models (little capacity)

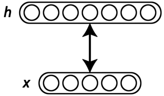
Learning by contrastive divergence



Definition: **Free Energy**

$$\left.\begin{array}{rcl}\mathcal{F}\mathcal{E}(\mathbf{x}) & = & -\log \sum_{\mathbf{h}} e^{-\mathcal{E}(\mathbf{x},\mathbf{h})} \\ & = & -\log p_\theta(\mathbf{x}) + \log Z(\theta)\end{array}\right\} \; p_\theta(\mathbf{x}) = \frac{e^{-\mathcal{F}\mathcal{E}(\mathbf{x})}}{\sum_{\mathbf{v}} e^{-\mathcal{F}\mathcal{E}(\mathbf{v})}}$$

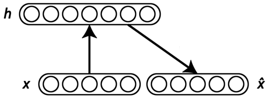
Learning by contrastive divergence



- Best gradient descent to maximize data likelihood:

$$-\boldsymbol{\nabla}_\theta \log p_\theta(\mathbf{x}_i) = \boldsymbol{\nabla}_\theta \mathcal{F}\mathcal{E}(\mathbf{x}_i) - \sum_{\mathbf{v}} p_\theta(\mathbf{v}) \boldsymbol{\nabla}_\theta \mathcal{F}\mathcal{E}(\mathbf{v}) \tag{1}$$

Learning by contrastive divergence



- Best gradient descent to maximize data likelihood:

$$-\nabla_\theta \log p_\theta(\mathbf{x}_i) = \nabla_\theta \mathcal{FE}(\mathbf{x}_i) - \sum_{\mathbf{v}} p_\theta(\mathbf{v}) \nabla_\theta \mathcal{FE}(\mathbf{v}) \qquad (1)$$

- *Approximation:* for each $\mathbf{x}_i$, sample an $\hat{\mathbf{x}}$ (n Gibbs steps)

$$p_\theta(\hat{\mathbf{x}}) \approx P(\hat{\mathbf{x}}|\mathbf{x}_i) \qquad (\mathcal{H}_1)$$

- Given the analytic expression of $\nabla_\theta \mathcal{FE}$, we update according to

$$(1) \overset{(\mathcal{H}_1)}{=\!=} \nabla_\theta \mathcal{FE}(\mathbf{x}_i) - \nabla_\theta \mathcal{FE}(\hat{\mathbf{x}})$$

Learning by contrastive divergence

Given the analytic expression of $\boldsymbol{\nabla}_\theta \mathcal{FE}$, we update parameters with

$$\boldsymbol{\nabla}_\theta \mathcal{FE}(\mathbf{x}_i) \quad - \quad \boldsymbol{\nabla}_\theta \mathcal{FE}(\hat{\mathbf{x}})$$

Tends to $\begin{cases} \mathcal{FE} \searrow & \text{on real data} \\ \mathcal{FE} \nearrow & \text{on data sampled by the RBM} \end{cases}$

Learning by contrastive divergence: The trap

Given the analytic expression of $\nabla_\theta \mathcal{F}\mathcal{E}$, we update parameters with

$$\nabla_\theta \left(\mathcal{F}\mathcal{E}(\mathbf{x}_i)\right) \quad - \quad \cancel{\nabla_\theta \left(\mathcal{F}\mathcal{E}(\hat{\mathbf{x}})\right)}$$

$$\nabla_\theta \left.\mathcal{F}\mathcal{E}\right|_{\mathbf{x}_i} \quad - \quad \nabla_\theta \left.\mathcal{F}\mathcal{E}\right|_{\hat{\mathbf{x}}(\theta)}$$

$$\ldots\text{estimate of:} \quad \mathrm{E}\left[\nabla_\theta \mathcal{F}\mathcal{E}(\mathbf{x}) - \sum_{\hat{\mathbf{x}}} p(\hat{\mathbf{x}}|\theta) \nabla_\theta \mathcal{F}\mathcal{E}(\hat{\mathbf{x}})\right] \qquad (2)$$

Learning by contrastive divergence: The trap

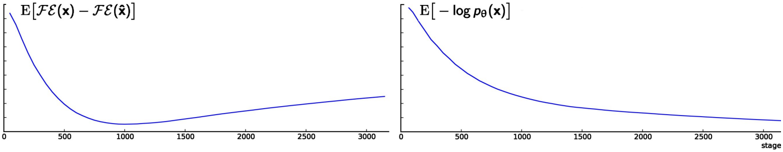
Given the analytic expression of $\nabla_\theta \mathcal{FE}$, we update parameters with

$$\nabla_\theta \left(\mathcal{FE}(\mathbf{x}_i) \right) \quad - \quad \cancel{\nabla_\theta \left(\mathcal{FE}(\hat{\mathbf{x}}) \right)}$$

$$\nabla_\theta \left. \mathcal{FE} \right|_{\mathbf{x}_i} \quad - \quad \nabla_\theta \left. \mathcal{FE} \right|_{\hat{\mathbf{x}}(\theta)}$$

$$\ldots\text{estimate of:} \quad \mathrm{E}\left[\nabla_\theta \mathcal{FE}(\mathbf{x}) - \sum_{\hat{\mathbf{x}}} p(\hat{\mathbf{x}}|\theta) \nabla_\theta \mathcal{FE}(\hat{\mathbf{x}}) \right] \tag{2}$$

1. $\mathcal{FE}(\mathbf{x}) - \mathcal{FE}(\hat{\mathbf{x}})$ is not the optimized function



Learning by contrastive divergence: The trap

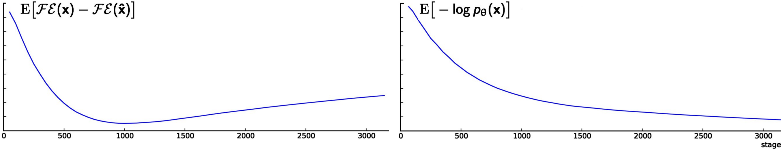
Given the analytic expression of $\nabla_\theta \mathcal{FE}$, we update parameters with

$$\nabla_\theta \left(\mathcal{FE}(\mathbf{x}_i)\right) \quad - \quad \cancel{\nabla_\theta \left(\mathcal{FE}(\hat{\mathbf{x}})\right)}$$

$$\nabla_\theta \, \mathcal{FE}\big|_{\mathbf{x}_i} \quad - \quad \nabla_\theta \mathcal{FE}\big|_{\hat{\mathbf{x}}(\theta)}$$

$$\ldots\text{estimate of:} \quad \mathrm{E}\left[\nabla_\theta \mathcal{FE}(\mathbf{x}) - \sum_{\hat{\mathbf{x}}} p(\hat{\mathbf{x}}|\theta) \nabla_\theta \mathcal{FE}(\hat{\mathbf{x}})\right] \qquad (2)$$

1. $\mathcal{FE}(\mathbf{x}) - \mathcal{FE}(\hat{\mathbf{x}})$ is not the optimized function



2. Nothing guarantees (2) is the gradient of a scalar function...

So how to choose the best hyper-parameters?

How to choose the best hyper-parameters? (1/2)

Visualizing generated **samples**

- Give an insight of the learned representation
- Give an idea the weakness of the models.

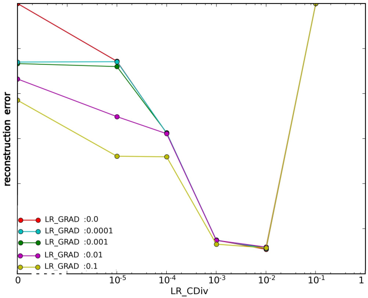
Topic	Chance Error Rate (%)	Error Rate (%)
color	25	7
shape	66	47
size	50	0
location	50	4

How to choose the best hyper-parameters? (2/2)

Monitoring the **reconstruction error**

RBM as autoassociator (Mean-Field approximation)

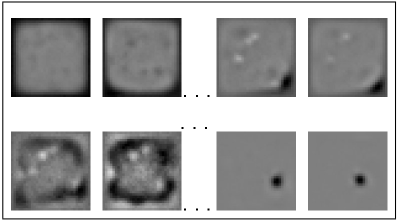
- Rq: we can also train the reconstruction error by its stochastic gradient descent



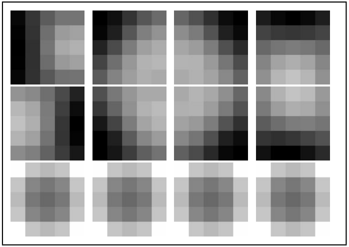
Helping the RBM to work better

As for neural networks, redundancy in trained models.

Fully connected RBMs



Convolution RBMs



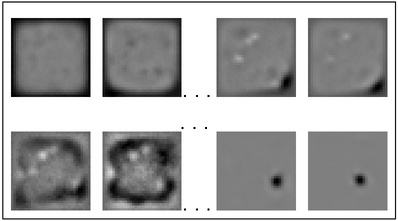
Goal

While teaching RBM's units to do something good,
make them do different things

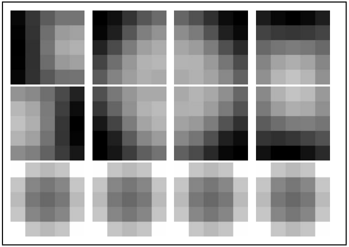
Helping the RBM to work better

As for neural networks, redundancy in trained models.

Fully connected RBMs



Convolution RBMs



Different heuristics

Example for binomial units, with $q_k = p(\mathbf{h}_k = 1|\mathbf{x})$

$$\mathcal{C}(\theta) = -\log \mathrm{E}\left[p_\theta(\mathbf{x})\right] + \lambda \mathcal{C}\left(\{q_k q_l\}_{k \neq l}\right)$$

Future work

Lot's of things to try
before making DBN learn the human world with TV. . .