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, ...
Motivations

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.
Motivations

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

<table>
<thead>
<tr>
<th>Topic</th>
<th>Question given to the computer</th>
<th>Answer</th>
</tr>
</thead>
<tbody>
<tr>
<td>Color</td>
<td>There is a small triangle. What color is it?</td>
<td>Green</td>
</tr>
<tr>
<td>Shape</td>
<td>What is the shape of the green object?</td>
<td>Triangle</td>
</tr>
<tr>
<td>Location</td>
<td>Is the blue square at the top or at the bottom?</td>
<td>At the top</td>
</tr>
<tr>
<td>Size</td>
<td>There is a triangle on the right.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Is it rather small or bigger?</td>
<td>Small</td>
</tr>
<tr>
<td>Size (relative)</td>
<td>Is the square smaller or bigger than the triangle?</td>
<td>Bigger</td>
</tr>
</tbody>
</table>
The ‘baby AI’ project
A first step: preliminary results

<table>
<thead>
<tr>
<th>Topic</th>
<th>Chance Error Rate (%)</th>
<th>Baseline Error Rate (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>1 object</td>
</tr>
<tr>
<td>color</td>
<td>25</td>
<td>10</td>
</tr>
<tr>
<td>shape</td>
<td>66</td>
<td>50</td>
</tr>
<tr>
<td>size</td>
<td>50</td>
<td>0.5</td>
</tr>
<tr>
<td>location</td>
<td>50</td>
<td>5</td>
</tr>
</tbody>
</table>

- 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

<table>
<thead>
<tr>
<th>Topic</th>
<th>Chance Error Rate (%)</th>
<th>Baseline Error Rate (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>1 object</td>
</tr>
<tr>
<td>color</td>
<td>25</td>
<td>10</td>
</tr>
<tr>
<td>shape</td>
<td>66</td>
<td>50</td>
</tr>
<tr>
<td>size</td>
<td>50</td>
<td>0.5</td>
</tr>
<tr>
<td>location</td>
<td>50</td>
<td>5</td>
</tr>
</tbody>
</table>

- 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** = empirical likelihood (on unlabeled data)
- **Practical limitation**: complexity $O(2^{\min(N_h, N_v)})$
  $\mapsto$ only for small models (little capacity)
Learning by contrastive divergence

**Definition: Free Energy**

\[
\mathcal{F}(x) = - \log \sum_h e^{-\mathcal{E}(x, h)} = - \log p_\theta(x) + \log Z(\theta)
\]

\[
p_\theta(x) = \frac{e^{-\mathcal{F}(x)}}{\sum_v e^{-\mathcal{F}(v)}}
\]
Learning by contrastive divergence

- Best gradient descent to maximize data likelihood:

\[-\nabla_\theta \log p_\theta(x_i) = \nabla_\theta F\mathcal{E}(x_i) - \sum_v p_\theta(v) \nabla_\theta F\mathcal{E}(v) \] (1)
On strategies to learn RBMs from unlabeled data

Learning by contrastive divergence

- Best gradient descent to maximize data likelihood:

  \[- \nabla_\theta \log p_\theta(x_i) = \nabla_\theta \mathcal{F}\mathcal{E}(x_i) - \sum_v p_\theta(v) \nabla_\theta \mathcal{F}\mathcal{E}(v) \quad (1)\]

- \textit{Approximation:} for each $x_i$, sample an $\hat{x}$ ($n$ Gibbs steps)

  \[p_\theta(\hat{x}) \approx P(\hat{x}|x_i) \quad (\mathcal{H}_1)\]

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

  \[(1) \overset{(\mathcal{H}_1)}{\Rightarrow} \nabla_\theta \mathcal{F}\mathcal{E}(x_i) - \nabla_\theta \mathcal{F}\mathcal{E}(\hat{x})\]
Learning by contrastive divergence

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

$$\nabla_\theta \mathcal{E}(x_i) - \nabla_\theta \mathcal{E}(\hat{x})$$

Tends to

$$\begin{cases} \mathcal{E} \downarrow & \text{on real data} \\ \mathcal{E} \uparrow & \text{on data sampled by the RBM} \end{cases}$$
Learning by contrastive divergence: The trap

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

$$\nabla_\theta (\mathcal{F}\mathcal{E}(x_i)) - \nabla_\theta (\mathcal{F}\mathcal{E}(\hat{x}))$$

$$\nabla_\theta \mathcal{F}\mathcal{E}|_{x_i} - \nabla_\theta \mathcal{F}\mathcal{E}|_{\hat{x}(\theta)}$$

...estimate of: $E[\nabla_\theta \mathcal{F}\mathcal{E}(x) - \sum_{\hat{x}} p(\hat{x}|\theta) \nabla_\theta \mathcal{F}\mathcal{E}(\hat{x})]$ (2)
Learning by contrastive divergence: The trap

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

\[
\nabla_{\theta} (\mathcal{F}\mathcal{E}(x_i)) - \mathcal{F}\mathcal{E}(\hat{x})
\]

\[
\nabla_{\theta} \mathcal{F}\mathcal{E}_{x_i} - \mathcal{F}\mathcal{E}_{\hat{x}(\theta)}
\]

...estimate of: \[E[\nabla_{\theta} \mathcal{F}\mathcal{E}(x) - \sum_{\hat{x}} p(\hat{x}|\theta) \nabla_{\theta} \mathcal{F}\mathcal{E}(\hat{x})] \] (2)

$\mathcal{F}\mathcal{E}(x) - \mathcal{F}\mathcal{E}(\hat{x})$ is not the optimized function
Learning by contrastive divergence: The trap

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

$$\nabla_\theta \mathcal{F}\mathcal{E}(x_i) - \nabla_\theta \mathcal{F}\mathcal{E}(\hat{x})$$

$$\nabla_\theta \mathcal{F}\mathcal{E}|_{x_i} - \nabla_\theta \mathcal{F}\mathcal{E}|_{\hat{x}(\theta)}$$

...estimate of: $\mathbb{E}[\nabla_\theta \mathcal{F}\mathcal{E}(x) - \sum_{\hat{x}} p(\hat{x}|\theta) \nabla_\theta \mathcal{F}\mathcal{E}(\hat{x})]$ (2)

1. $\mathcal{F}\mathcal{E}(x) - \mathcal{F}\mathcal{E}(\hat{x})$ 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.

<table>
<thead>
<tr>
<th>Topic</th>
<th>Chance Error Rate (%)</th>
<th>Error Rate (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>color</td>
<td>25</td>
<td>7</td>
</tr>
<tr>
<td>shape</td>
<td>66</td>
<td>47</td>
</tr>
<tr>
<td>size</td>
<td>50</td>
<td>0</td>
</tr>
<tr>
<td>location</td>
<td>50</td>
<td>4</td>
</tr>
</tbody>
</table>
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 $q_k = p(h_k = 1|x)$

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

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