Learning Long Sequences with TRBM

James Bergstra

Aug 7-11 2007

About me

I'm researching recurrent networks as probability models of sequential data.

Data

- ultimately : audio, video, language, music, robotic control
- currently: toy bit sequences, symbolic melodies

Models

- recurrent nets with regularization to improve BPTT
- temporally hierarchical nets
- partly linear nets

How to detect/model long-term and multi-scale patterns???

What I'll talk about today

Temporal Restricted Boltzmann Machines

1. What they are
2. How they are trained
3. One way things can go wrong when learning:
    - long sequences
    - many non-overlapping sequences
    - sequences with long-term statistical dependencies
4. Ideas to get around this problem (some results)

Temporal RBM

The ones with temporal connections between hidden units.

Ask for picture!

3 Phase Learning Algo

1. Optimize W as RBM of non-temporal $P(v)$
    1. Project sequence into observation space.
    2. Learn (RBM) density model of projected points.
2. Optimize U for given $z^{1..T}$
    - Choosing $z^{(t)} = \mathrm{E}_W[h|v^{(t)}]$
3. Continue to optimize W, U to model $v^{(1..t)}$
    - W by Contrastive Divergence
    - U by backprop of bias gradient

How United States of America Goes Wrong

$v =$ "united states of america united states of america ..."

$$\begin{array}{rr} \textit{united} \rightarrow 00 & \textit{america} \rightarrow 01 \\ \textit{of} \rightarrow 10 & \textit{states} \rightarrow 11 \end{array}$$

Phase 1	Suppose the RBM learns the mapping above
Phase 2	First component of z must perform XOR with 1-layer net ... not possible!
Phase 3	W, U are nowhere near a solution, better to restart with joint optimization.

Semantics of z_t in TRBM

$$\log P(v^{(1)}, v^{(2)}, ..., v^{(t)}, ...)$$

$$= \sum_t \log P(v^{(t)} | z^{(t-1)}) \qquad \textit{hidden markov assumption}$$

$$\propto \sum_t \log \sum_h e^{-(v^{(t)}W + z^{(t-1)}U)h} \qquad \textit{RBM}$$

$$\propto \sum_t e^{-\mathit{freeEnergy}(v^{(t)} | z^{(t-1)})}$$

z_t **must be predictive of FUTURE** $v^{(t+1)}$**,** $z^{(t+1)}$
(within constraints imposed by functional form of z_t)

Why is Phase 2 hard?

Phase 2 begins with $z^{(t)} = \mathrm{E}_W[h|v^{(t)}]$, and tries to solve for U:

$$
\begin{aligned}
sigm(Uz^{(1)} + b) &= z^{(2)} \\
sigm(Uz^{(2)} + b) &= z^{(3)} \\
&\dots \\
sigm(Uz^{(T-1)} + b) &= z^{(T)}
\end{aligned}
$$

Shallow: When T is greater than the number of dimensions of $z^{(t)}$, then no solution generally exists (linear separability).

Deep: *Many* (most?) trajectories of length $> K$ through $\{0, 1\}^K$ are not possible with an iterated system of form $z^{(t)} = sgn(Uz_{(t-1)} + b)$, though long trajectories exist.

What to do?

Avoid the problem:

- Use sufficiently large Z vectors, jump straight to phase 3
- Add fresh units to the system for phase 3

[Try to] solve the problem:

- Decouple $z^{(t)}$ from $P(h|v^{(t)})$, so that $z^{(t)} = f(v^{(t)}, z^{(t-1)})$
  - Differentiable f enables optimization by BPTT (really!)
  - I have some ideas (Yoshua too!) to improve BPTT
  - some preliminary results... (ongoing work)

Fin

Questions? Comments?