Deep Temporal Sigmoid Belief Networks for Sequence Modeling

Zhe Gan, Chunyuan Li, Ricardo Henao, David Carlson and Lawrence Carin
Duke University, Durham NC 27708, USA

INTRODUCTION

Problem of interest: Developing deep generative models for sequential data.

Main idea:

- Constructing a hierarchy of Temporal Sigmoid Belief Networks (TSBNs).
- TSBN is defined as a sequential stack of Sigmoid Belief Networks (SBNs).

Contributions:

- A generalization of Hidden Markov Models (HMMs) and Linear Dynamical Systems (LDS).
- A probabilistic construction of Recurrent Neural Networks (RNNs).
- Closely related to Temporal Restricted Boltzmann Machine (TRBM), but our model has a fully directed generative process.
- Can be utilized to model various data, *e.g.*, binary, real-valued and counts.

Challenge: Designing scalable learning and inference algorithms.

Solution:

- Stochastic Variational Inference (SVI).
- Design a recognition model for fast inference.

MODEL FORMULATION

Sigmoid Belief Network: An SBN models a binary visible vector $\boldsymbol{v} \in \{0,1\}^M$, in terms of binary hidden variables $\boldsymbol{h} \in \{0,1\}^J$ and weights $\mathbf{W} \in \mathbb{R}^{M \times J}$ with

$$p(v_m = 1|\boldsymbol{h}) = \sigma(\boldsymbol{w}_m^\top \boldsymbol{h} + c_m) \qquad p(h_j = 1) = \sigma(b_j) \tag{1}$$

- SBN is closely related to RBM, which is a Markov random field with the same bipartite structure as the SBN.

Temporal SBN: Assume a length-T binary visible sequence, the tth time step of which is denoted $\boldsymbol{v}_t \in \{0,1\}^M$. The TSBN describes the joint probability as

$$p_{\boldsymbol{\theta}}(\mathbf{V}, \mathbf{H}) = p(\boldsymbol{h}_1)p(\boldsymbol{v}_1|\boldsymbol{h}_1) \cdot \prod_{t=2}^{T} p(\boldsymbol{h}_t|\boldsymbol{h}_{t-1}, \boldsymbol{v}_{t-1}) \cdot p(\boldsymbol{v}_t|\boldsymbol{h}_t, \boldsymbol{v}_{t-1}) \tag{2}$$

Each conditional distribution is expressed as

$$p(h_{jt} = 1|\boldsymbol{h}_{t-1}, \boldsymbol{v}_{t-1}) = \sigma(\boldsymbol{w}_{1j}^\top \boldsymbol{h}_{t-1} + \boldsymbol{w}_{3j}^\top \boldsymbol{v}_{t-1} + b_j) \tag{3}$$

$$p(v_{mt} = 1|\boldsymbol{h}_t, \boldsymbol{v}_{t-1}) = \sigma(\boldsymbol{w}_{2m}^\top \boldsymbol{h}_t + \boldsymbol{w}_{4m}^\top \boldsymbol{v}_{t-1} + c_m) \tag{4}$$

- TSBN can be viewed as a HMM with an exponentially large state space and a highly structured transition matrix.
- TSBN allows for *fast sampling* of "fantasy" data from the inferred model.

Extensions:

- Modeling real-valued data: $p(\boldsymbol{v}_t|\boldsymbol{h}_t, \boldsymbol{v}_{t-1}) = \mathcal{N}(\boldsymbol{\mu}_t, \text{diag}(\boldsymbol{\sigma}_t^2))$, where
  $\mu_{mt} = \boldsymbol{w}_{2m}^\top \boldsymbol{h}_t + \boldsymbol{w}_{4m}^\top \boldsymbol{v}_{t-1} + c_m \qquad \log \sigma_{mt}^2 = (\boldsymbol{w}'_{2m})^\top \boldsymbol{h}_t + (\boldsymbol{w}'_{4m})^\top \boldsymbol{v}_{t-1} + c'_m$
- Modeling counts: $p(\boldsymbol{v}_t|\boldsymbol{h}_t, \boldsymbol{v}_{t-1}) = \prod_{m=1}^M y_{mt}^{v_{mt}}$, where
  $$y_{mt} = \frac{\exp(\boldsymbol{w}_{2m}^\top \boldsymbol{h}_t + \boldsymbol{w}_{4m}^\top \boldsymbol{v}_{t-1} + c_m)}{\sum_{m'=1}^M \exp(\boldsymbol{w}_{2m'}^\top \boldsymbol{h}_t + \boldsymbol{w}_{4m'}^\top \boldsymbol{v}_{t-1} + c_{m'})}$$
- Going deep: Adding stochastic or deterministic hidden layers.

Graphical Model

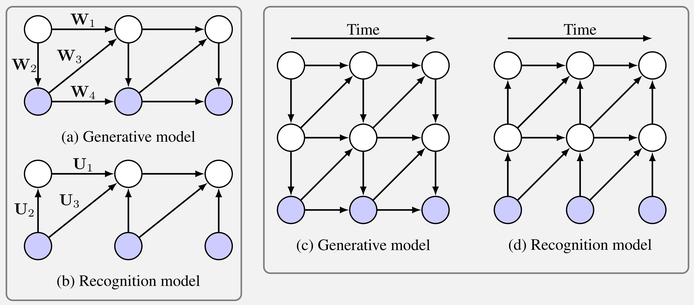

Figure: Graphical model for Deep Temporal Sigmoid Belief Network. (a,b) Generative and recognition model of TSBN. (c,d) Generative and recognition model of a two-layer TSBN.

SCALABLE LEARNING & INFERENCE

Variational Lower Bound Objective

$$\mathcal{L}(\mathbf{V}, \boldsymbol{\theta}, \boldsymbol{\phi}) = \mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{H}|\mathbf{V})}[\log p_{\boldsymbol{\theta}}(\mathbf{V}, \mathbf{H}) - \log q_{\boldsymbol{\phi}}(\mathbf{H}|\mathbf{V})] \tag{5}$$

We construct the approximate posterior $q_{\boldsymbol{\phi}}(\mathbf{H}|\mathbf{V})$ as a recognition model.

$$q_{\boldsymbol{\phi}}(\mathbf{H}|\mathbf{V}) = q(\boldsymbol{h}_1|\boldsymbol{v}_1) \cdot \prod_{t=2}^{T} q(\boldsymbol{h}_t|\boldsymbol{h}_{t-1}, \boldsymbol{v}_t, \boldsymbol{v}_{t-1}) \tag{6}$$

and each conditional distribution is specified as

$$q(h_{jt} = 1|\boldsymbol{h}_{t-1}, \boldsymbol{v}_t, \boldsymbol{v}_{t-1}) = \sigma(\boldsymbol{u}_{1j}^\top \boldsymbol{h}_{t-1} + \boldsymbol{u}_{2j}^\top \boldsymbol{v}_t + \boldsymbol{u}_{3j}^\top \boldsymbol{v}_{t-1} + d_j) \tag{7}$$

The recognition model is introduced to achieve fast inference.

Parameter Learning: We apply the *Neural Variational Inference and Learning* (NVIL) algorithm.

$$\nabla_{\boldsymbol{\theta}} \mathcal{L}(\mathbf{V}) = \mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{H}|\mathbf{V})}[\nabla_{\boldsymbol{\theta}} \log p_{\boldsymbol{\theta}}(\mathbf{V}, \mathbf{H})] \tag{8}$$

$$\nabla_{\boldsymbol{\phi}} \mathcal{L}(\mathbf{V}) = \mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{H}|\mathbf{V})}[(\log p_{\boldsymbol{\theta}}(\mathbf{V}, \mathbf{H}) - \log q_{\boldsymbol{\phi}}(\mathbf{H}|\mathbf{V})) \times \nabla_{\boldsymbol{\phi}} \log q_{\boldsymbol{\phi}}(\mathbf{H}|\mathbf{V})] \tag{9}$$

- Use Monte Carlo methods to approximate expectations.
- Variance reduction: (*i*) centering the learning signal by subtracting the baseline; (*ii*) variance normalization.
- Use RMSprop for optimization.

EXPERIMENTS

Datasets:

- ***Bouncing balls:*** Synthetic videos of 3 bouncing balls, binary valued.
- ***Motion capture:*** Walking & running sequences collected by CMU & MIT.
- ***Polyphonic music:*** A collection of 88-dim binary sequences, that span the whole range of piano from A0 to C8.
- ***State of the Union (STU):*** transcripts of 225 US STU addresses, from 1790 to 2014. Vocab size is 2375.

Qualitative Evaluation

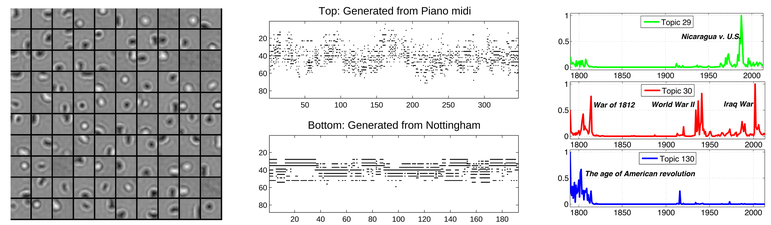

Figure: (Left) Dictionaries learned on the videos of bouncing balls. (Middle) Samples generated from TSBN trained on the polyphonic music. Each column is a sample vector of notes. (Right) Time evolving from 1790 to 2014 for three selected topics learned from the STU dataset.

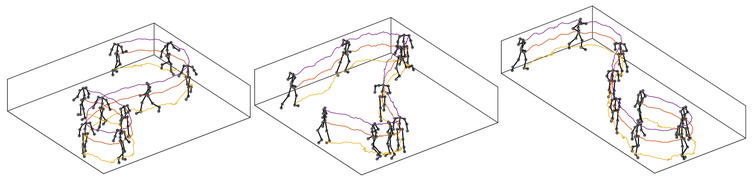

Figure: Motion trajectories generated from the TSBN trained on the motion capture dataset. (Left) Walking. (Middle) Running-running-walking. (Right) Running-walking.

Quantitative Evaluation

Table: Prediction error for the bouncing balls.

Model	Dim	Order	Pred. Err.
DTSBN-s	100-100	2	2.79 ± 0.39
DTSBN-d	100-100	2	2.99 ± 0.42
TSBN	100	4	3.07 ± 0.40
TSBN	100	1	9.48 ± 0.38
RTRBM	3750	1	3.88 ± 0.33
SRTRBM	3750	1	3.31 ± 0.33

Table: Prediction error for the motion capture.

Model	Walking	Running
DTSBN-s	4.40 ± 0.28	2.56 ± 0.40
DTSBN-d	4.62 ± 0.01	2.84 ± 0.01
TSBN	5.12 ± 0.50	4.85 ± 1.26
HMSBN	10.77 ± 1.15	7.39 ± 0.47
ss-SRTRBM	8.13 ± 0.06	5.88 ± 0.05
g-RTRBM	14.41 ± 0.38	10.91 ± 0.27

Table: Log-likelihood for the music dataset.

Model	Piano.	Nott.	Muse.	JSB.
TSBN	-7.98	-3.67	-6.81	-7.48
RNN-NADE	-7.05	-2.31	-5.60	-5.56
RTRBM	-7.36	-2.62	-6.35	-6.35
RNN	-8.37	-4.46	-8.13	-8.71

Table: Prediction precision for STU.

Model	Dim	MP	PP
HMSBN	25	0.327±0.002	0.353±0.070
DHMSBN-s	25-25	0.299±0.001	0.378±0.006
GP-DPFA	100	0.223±0.001	0.189±0.003
DRFM	25	0.217±0.003	0.177±0.010

Dynamic Topic Modeling

Table: Top 8 most probable words associated with the STU topics.

Topic #29	Topic #30	Topic #130	Topic #64	Topic #70	Topic #74
family	officer	government	generations	Iraqi	Philippines
budget	civilized	country	generation	Qaida	islands
Nicaragua	warfare	public	recognize	Iraq	axis
free	enemy	law	brave	Iraqis	Nazis
future	whilst	present	crime	Al	Japanese
freedom	gained	citizens	race	Saddam	Germans
excellence	lake	united	balanced	ballistic	mines
drugs	safety	house	streets	terrorists	sailors

ACKNOWLEDGEMENTS

This research was supported by ARO, DARPA, DOE, NGA and ONR.