Factored Temporal Sigmoid Belief Networks for Sequence Learning

Jiaming Song (Tsinghua), Zhe Gan (Duke), Lawrence Carin (Duke)

Contribution

We introduce deep generative models to simultaneously learn multiple sequences.

- The model is parameterized with weight tensors, which interact with side information to produce different transtition matrices, modeling multiple sequence styles.
- The transition matrices are further factored to reduce the number of parameters and improve generalization.
- It is applicable to tasks such as sequence generation and classification.

Model

We model the joint probability of visible data $\mathbf{V}$, hidden data $\mathbf{H}$ and side information $\mathbf{Y}$:

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

where

$$\begin{aligned} p(h_{jt}=1|\boldsymbol{h}_{t-1},\boldsymbol{v}_{t-1}) &= \sigma(\tilde{h}_{jt}) \\ p(\boldsymbol{v}_t|\boldsymbol{h}_t,\boldsymbol{v}_{t-1}) &= \mathcal{N}(\boldsymbol{\mu}_t,\mathrm{diag}(\boldsymbol{\sigma}_t^2)) \\ \tilde{\boldsymbol{h}}_t &= \mathbf{W}_1^{(y)}\boldsymbol{h}_{t-1} + \mathbf{W}_3^{(y)}\boldsymbol{v}_{t-1} + \boldsymbol{b}^{(y)} \\ \boldsymbol{\mu}_t &= \mathbf{W}_2^{(y)}\boldsymbol{h}_t + \mathbf{W}_4^{(y)}\boldsymbol{v}_{t-1} + \boldsymbol{c}^{(y)} \\ \log\boldsymbol{\sigma}_t^2 &= \mathbf{W}_2'^{(y)}\boldsymbol{h}_t + \mathbf{W}_4'^{(y)}\boldsymbol{v}_{t-1} + \boldsymbol{c}'^{(y)} \end{aligned}$$

The transition weight matrix is denoted as $\mathbf{W}_{ijk}^{(y)} = \sum_{s=1}^{S}\hat{\mathbf{W}}_{ijks}y_{st}$, and $\hat{\mathbf{W}}$ is a three-way tensor. We name this the Conditional Temporal Sigmoid Belief Network (CTSBN).

We can further factor the weight matrices:

$$\mathbf{W}^{(y)} = \mathbf{W}_a\cdot\mathrm{diag}(\mathbf{W}_b\boldsymbol{y}_t)\cdot\mathbf{W}_c$$

which is illustrated below.

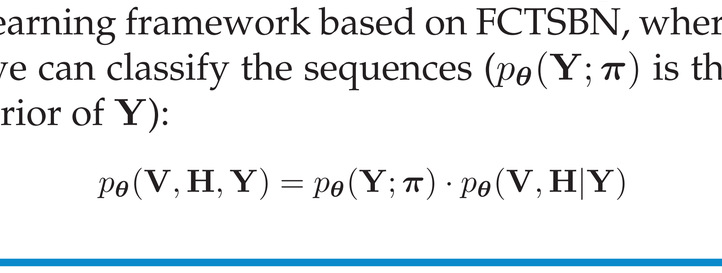

Figure 4: Graphical illustration for non-factored weights (left) and factored weights (right).

We call the factored model the Factored Conditional Temporal SBN (or FCTSBN).

By adding stochastic deep layers, we may increase the representational power of the model:

$$p(\boldsymbol{h}_t^{(\ell)}) = \prod_{j=1}^{J^{(\ell)}} p(h_{jt}^{(\ell)}|\boldsymbol{h}_t^{(\ell+1)},\boldsymbol{h}_{t-1}^{(\ell)},\boldsymbol{h}_{t-1}^{(\ell-1)},\boldsymbol{y}_t)$$

We can also introduce a semi-supervised learning framework based on FCTSBN, where we can classify the sequences ($p_{\boldsymbol{\theta}}(\mathbf{Y};\boldsymbol{\pi})$ is the prior of $\mathbf{Y}$):

$$p_{\boldsymbol{\theta}}(\mathbf{V},\mathbf{H},\mathbf{Y}) = p_{\boldsymbol{\theta}}(\mathbf{Y};\boldsymbol{\pi})\cdot p_{\boldsymbol{\theta}}(\mathbf{V},\mathbf{H}|\mathbf{Y})$$

Learning and Inference

We apply the Neural Variational Inference and Learning (NVIL) algorithm, which allows for scalable parameter learning and inference by introducing a recognition model $q_{\boldsymbol{\phi}}(\mathbf{H}|\mathbf{V},\mathbf{Y})$ with parameters $\boldsymbol{\phi}$, to approximate the true posterior $p(\mathbf{H}|\mathbf{V},\mathbf{Y})$:

$$\mathcal{L}(\mathbf{V}|\mathbf{Y},\boldsymbol{\theta},\boldsymbol{\phi}) = \mathcal{J}(q_{\boldsymbol{\phi}}(\mathbf{H}|\mathbf{V},\mathbf{Y}),p_{\boldsymbol{\theta}}(\mathbf{V},\mathbf{H}|\mathbf{Y}))$$

where $\mathcal{J}(q,p) = \mathbb{E}_q[\log p - \log q]$.

For both models, the recognition model is expressed as

$$q_{\boldsymbol{\phi}}(\mathbf{H}|\mathbf{V},\mathbf{Y}) = \prod_{t=1}^{T} q(\boldsymbol{h}_t|\boldsymbol{h}_{t-1},\boldsymbol{v}_t,\boldsymbol{v}_{t-1},\boldsymbol{y}_t)$$

where $q(h_{jt}=1|\boldsymbol{h}_{t-1},\boldsymbol{v}_t,\boldsymbol{v}_{t-1},\boldsymbol{y}_t) = \sigma(\mathbf{U}_1^{(y)}\boldsymbol{h}_{t-1} + \mathbf{U}_2^{(y)}\boldsymbol{v}_t + \mathbf{U}_3^{(y)}\boldsymbol{v}_{t-1} + \boldsymbol{d}^{(y)})$.

We apply Monte Carlo integration and stochastic gradient descent for optimization.

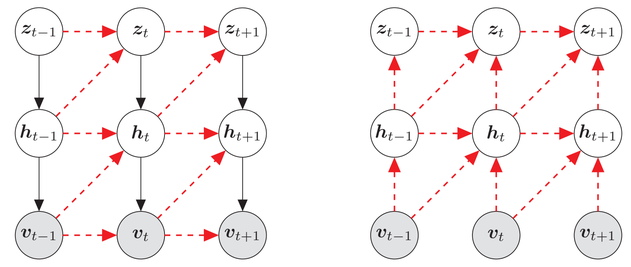

Figure 1: Generative model (left) and recognition model (right) of a deep FCTSBN with two layers. Red, dashed arrows indicate factorized weights.

Experiments

mocap2 is a motion capture dataset with two styles (**walking** and **running**). dFCTSBN denotes the two-layer version of FCTSBN.
The one layer factored model has significantly better performance than the non-factored model.

Method	Walking	Running
FCTSBN	**4.59** $\pm$ 0.35	**2.86** $\pm$ 0.23
CTSBN	4.67 $\pm$ 0.22	3.41 $\pm$ 0.65
TSBN°	5.12 $\pm$ 0.50	4.85 $\pm$ 1.26
dFCTSBN	**4.31** $\pm$ 0.13	2.58 $\pm$ 0.21
DTSBN-S°	4.40 $\pm$ 0.28	**2.56** $\pm$ 0.40
DTSBN-D°	4.62 $\pm$ 0.01	2.84 $\pm$ 0.01
ss-SRTRBM◇	8.13 $\pm$ 0.06	5.88 $\pm$ 0.05
g-RTRBM◇	14.41 $\pm$ 0.38	10.91 $\pm$ 0.27

Table 1: Prediction error for the mocap2 dataset.

mocap10 is a motion capture dataset with ten styles. We are able to generate motion sequences of 9 out of 10 styles, as well as style blending and transition. More generated videos are available at https://goo.gl/9R59d7.

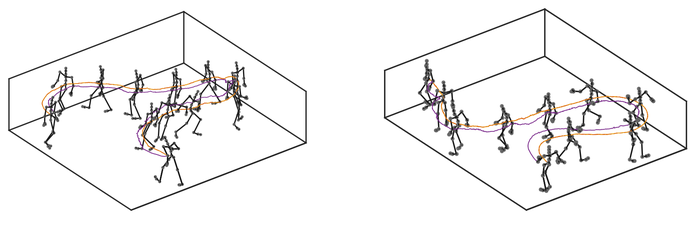

Figure 2: Left: *Drunk* to *normal* (6 skeletons for each style). **Right:** *Graceful* to *gangly* (5 skeletons for each style). Both figures are generated within 800 frames.

We also performed semi-supervised sequence classfication on the **mocap10** dataset and achieved better results than Transductive SVM and softmax classifier with L2 regularization.

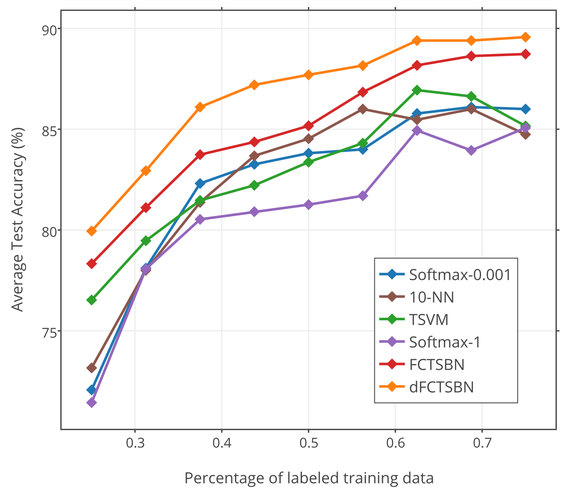

Figure 3: Test results on mocap10 classification. Each result is the average over 5 independent trials.

The **weather** dataset contains weather data in 25 locations on a 5×5 grid over 1990-2002. We used three types of side information to distinguish the locations.

	2d	10d	25d
CTSBN	6.45 $\pm$ 0.11	3.83 $\pm$ 0.20	3.78 $\pm$ 0.01
FCTSBN	5.46 $\pm$ 0.06	3.43 $\pm$ 0.03	3.37 $\pm$ 0.02
dFCTSBN	**5.09** $\pm$ 0.11	**3.37** $\pm$ 0.01	**3.35** $\pm$ 0.02
FCRBM	5.62 $\pm$ 0.35	3.77 $\pm$ 0.38	3.75 $\pm$ 0.08

Table 2: Average prediction error of 25 locations.

We also performed **conditional text generation** with our model.

B	*<#> and the evening and the morning were the fourth day . <#> and god said , let the waters the heaven after his kind : <#> god blessed their thing months...*
C	*we shall hold modern into tends ; and circumstance between seem understood retained defendant's to has that belief are not the recalled and will be led constituent...*
BC	*<#> and unto the rendered fair violence , morning turn the human whole been so eyes . <#> that god of air of the mountain show <#>the waters of fish and him would expect application : are gradual obliged that...*

Table 3: Generated text. <#> denote markers for verses. B represents *Bible*, C represents *Common Law*.

Future Directions

We have presented the Factored Conditional Temporal Sigmoid Belief Network, which can simultaneously model multiple temporal sequences. A general framework for semi-supervised sequence classification is also provided, allowing one to train a conditional generative model along with a classifier.

While we assume side information as a simple vector, it would be interesting to incorporate more structured side information into the model, *e.g.*, utilizing separate style and content components. We are also interested in models that can extract structured properties from side information.

References

[1] Z. Gan et al. Deep temporal sigmoid belief networks for sequence modeling. In *NIPS*, 2015.

[2] A. Mnih and K. Gregor. Neural variational inference and learning in belief networks. In *ICML 2014*.