Time Series Prediction via Recurrent Neural Networks with the Information Bottleneck Principle

Duo Xu and Faramarz Fekri
{dxu301, faramarz.fekri}@ece.gatech.edu

Introduction

- **Background**:
  - Time series prediction is an important interdisciplinary topic in computer sciences, statistics, and econometrics.
  - Recently, deep learning methods have been applied successfully to the time series prediction.
- **Recurrent Neural Network:**
  - RNNs are particularly suitable for modeling time series as they operate on input information as well as the trace of previously acquired information.
  - The most successful models: LSTM and its variant GRU.
- **Probabilistic Modeling**:
  - We propose an stochastic RNN trained using the Recurrent Information Bottleneck (RIB) as objective function.
  - The method maximizes the mutual information between latent state and the target with low latent complexity.
  - Our model: the input is first encoded to latent state, then fed into the RNN cell to generate next hidden state. The decoder generates the distribution of the predicted value of time series with latent and hidden state together.
- **Two advantages:**
  - Stochastic latent state can augment RNN to better utilize recent observations.
  - RIB can model the temporal dependencies of high-dimensional time series with lower complexity.

Proposed Methods

- **Gated Recurrent Unit:** it has gating units that modulate the information flow inside the unit, but without having a separate memory cells:

$$\begin{aligned}\boldsymbol{r}_t &= \sigma(\boldsymbol{W}_r\boldsymbol{x}_t + \boldsymbol{U}_r\boldsymbol{h}_{t-1} + \boldsymbol{b}_r)\\ \boldsymbol{u}_t &= \sigma(\boldsymbol{W}_u\boldsymbol{x}_t + \boldsymbol{U}_u\boldsymbol{h}_{t-1} + \boldsymbol{b}_u)\\ \boldsymbol{c}_t &= \sigma(\boldsymbol{W}_c\boldsymbol{x}_t + \boldsymbol{U}_c(\boldsymbol{r}_t \odot \boldsymbol{h}_{t-1}) + \boldsymbol{b}_c)\\ \boldsymbol{h}_t &= \boldsymbol{u}_t \odot \boldsymbol{h}_{t-1} + (1-\boldsymbol{u}_t)\odot \boldsymbol{c}_t\end{aligned}$$

  with the input ($\boldsymbol{x_t}$), the hidden state of GRU cell ($\boldsymbol{h_t}$), the reset gate ($\boldsymbol{r_t}$), the forgetting gate ($\boldsymbol{u_t}$), and candidate activation ($\boldsymbol{c_t}$).
- **Information Bottleneck:** It is information-theoretic view of deep neural networks. Assuming input random variable is X, latent state is Z and target is Y, the Markov chain $Y \leftrightarrow X \leftrightarrow Z$ holds.
- The aim is to get latent Z informative about the target Y, i.e. to maximize mutual information of Y and Z.
- In order to find a low-complexity representation of input, we can assume the mutual information between X and Z to be less than a pre-defined ϵ. Then we can formulate the optimization problem as

$$\max_{\boldsymbol{\theta}} I(Z,Y;\boldsymbol{\theta}) \quad \text{s.t.} \quad I(X,Z;\boldsymbol{\theta}) < \epsilon$$

  By introducing Lagrange multiplier $\beta \geq 0$, we have the objective:

$$L_{IB}(\boldsymbol{\theta}) = I(Z,Y;\boldsymbol{\theta}) - \beta I(Z,X;\boldsymbol{\theta})$$

- **Recurrent Information Bottleneck:** Information bottleneck is extended to recurrent models. We use multi-variate Gaussian distribution as a variational approximation and model temporal relationships by hidden states of GRU. The objective function is,

$$\frac{1}{N}\sum_{n=1}^{N}\Big[\sum_{t=1}^{T}\int p(z_{nt}|x_{nt})\log q(y_{nt}|z_{nt})dz_{nt} - \beta\int p(z_{nt}|x_{nt})\log\frac{p(z_{nt}|x_{nt})}{\tilde{p}_t(z_{nt})}dz_{nt}\Big]$$

Proposed Model

- In our model, the input is first encoded to la-tent state $\boldsymbol{z_t}$, which is then fed into GRU for state transition.
- $\boldsymbol{z_t}$ is then decoded together with hidden state to generate output.
- The dashed lines in the diagram show the conditional dependency in both prior of $\boldsymbol{z_t}$ and encoding of the input.
- The encoding follows distribution as

$$Z_t \sim \mathcal{N}\big(\mu_{\text{enc}}(\boldsymbol{x}_t, \boldsymbol{h}_{t-1}),\ \ \text{diag}(\sigma^2_{\text{enc}}(\boldsymbol{x}_t, \boldsymbol{h}_{t-1}))\big)$$

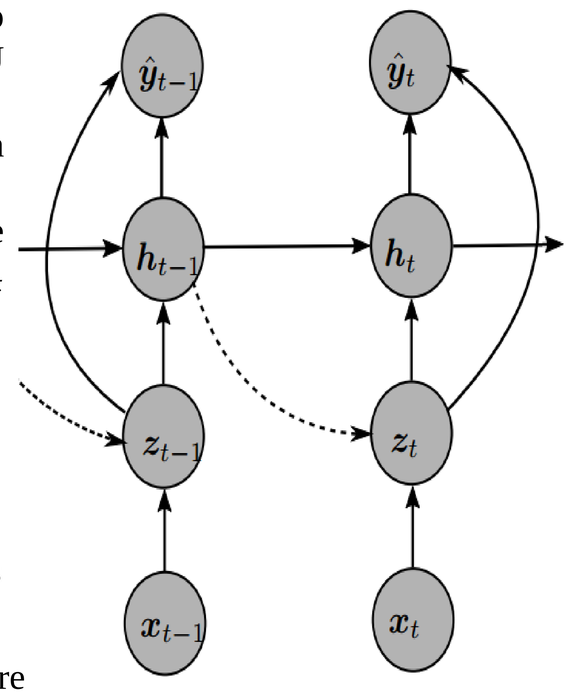
- The decoder follows distribution as

$$\hat{Y}_t \sim \mathcal{N}\big(\mu_{\text{dec}}(\boldsymbol{z}_t, \boldsymbol{h}_{t-1}), \text{diag}(\sigma^2_{\text{dec}}(\boldsymbol{z}_t, \boldsymbol{h}_{t-1}))\big)$$

- The prior of latent $\boldsymbol{z_t}$ follows distribution as

$$\tilde{p}(\boldsymbol{z}_t|\boldsymbol{h}_{t-1}) = \mathcal{N}(\mu_{\text{prior}}(\boldsymbol{h}_{t-1}), \text{diag}(\sigma^2_{\text{prior}}(\boldsymbol{h}_{t-1})))$$

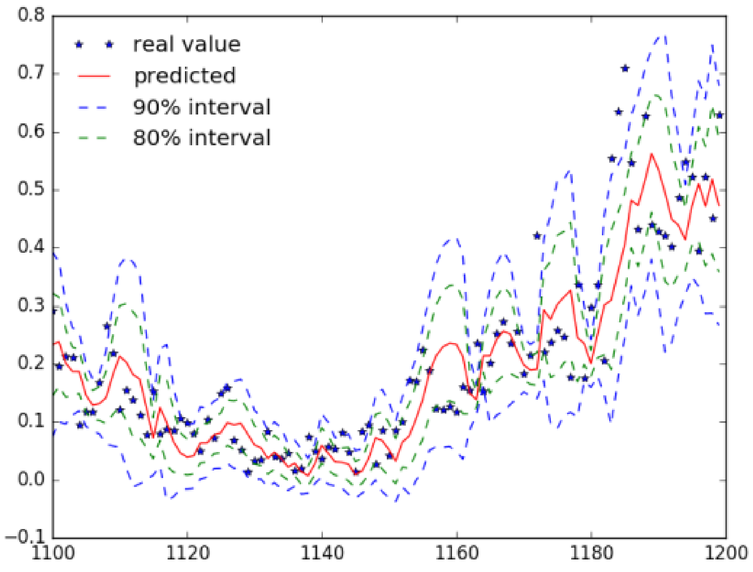
  where all mean and variance functions here are realized by multi-layer neural networks.
- In the GRU cell, the hidden state is translated following GRU operations.



Experiments

- The first experiment uses sunspot dataset. It contains sunspot number collected in Zurich from Jan. 1749 to Dec. 1983. We use first 1000 data points for training and the rest for testing. The $1-\alpha$ level confidence interval of prediction is

$$\big[\mu_{\text{dec}} - z_{1-\alpha/2}\sqrt{\sigma^2_{\text{dec}}},\ \ \mu_{\text{dec}} + z_{1-\alpha/2}\sqrt{\sigma^2_{\text{dec}}}\big]$$



	Proposed	RBM	SAE	LSTM
MAE	0.0453	0.0911	0.116	0.0524
RMSE	0.0612	0.132	0.150	0.0739

- The second experiment is on traffic data, a collection of 15 months of daily data from California's Department of Transportation, describing occupancy rate of 963 freeways in bay area.

	Proposed	MatFact	VRNN	SAE
NMAE	0.1127	0.1935	0.2103	0.2234
NMRSE	0.3608	0.4263	0.4312	0.4566

Conclusion

- We propose to use information bottleneck to model time series probabilistically, and obtain better prediction performance than previous methods.

Acknowledgement

- This work is partially supported by the Center for Energy and Geo Processing (CeGP) at Georgia Tech and by King Fahd University of Petroleum and Minerals (KFUPM),