KDD2019

FUDAN UNIVERSITY

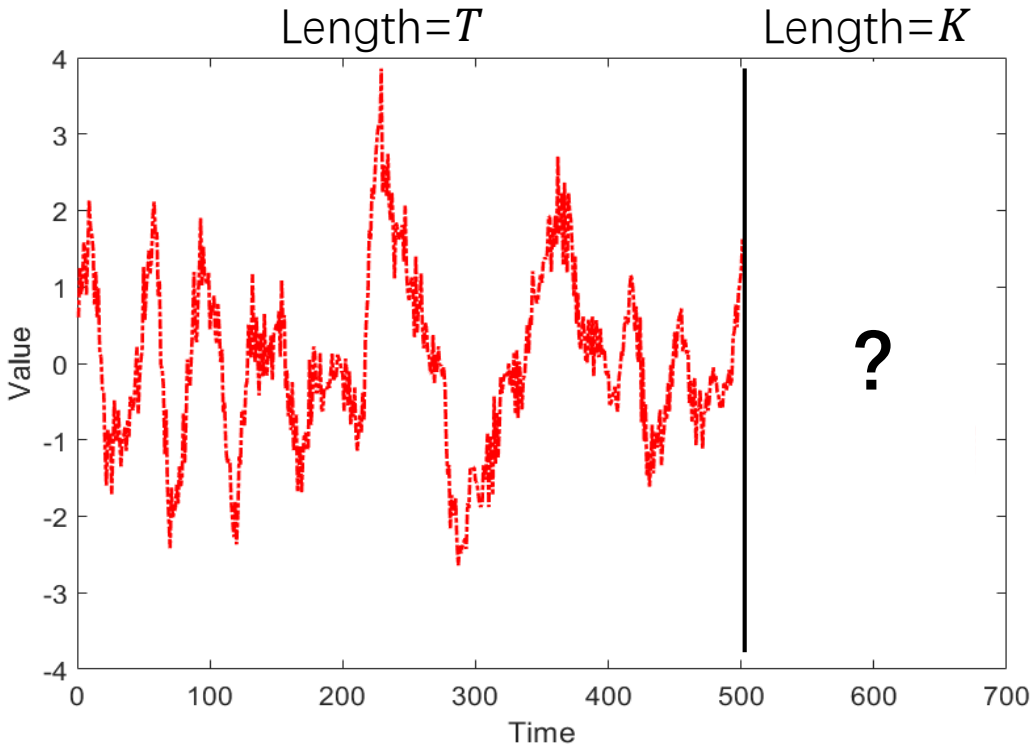
University of Science and Technology of China

Modeling Extreme Events in Time Series Prediction

Daizong Ding[1] Mi Zhang[1] Xudong Pan[1] Min Yang[1] Xiangnan He[2]

1. School of Computer Science, Fudan University
2. School of Data Science, University of Science and Technology of China

Time Series Prediction



Training

Inputs: $X_{1:T} = (x_1, \cdots, x_T)$

Labels: $Y_{1:T} = (y_1, \cdots, y_T)$

Outputs: $O_{1:T} = (o_1, \cdots, o_T)$

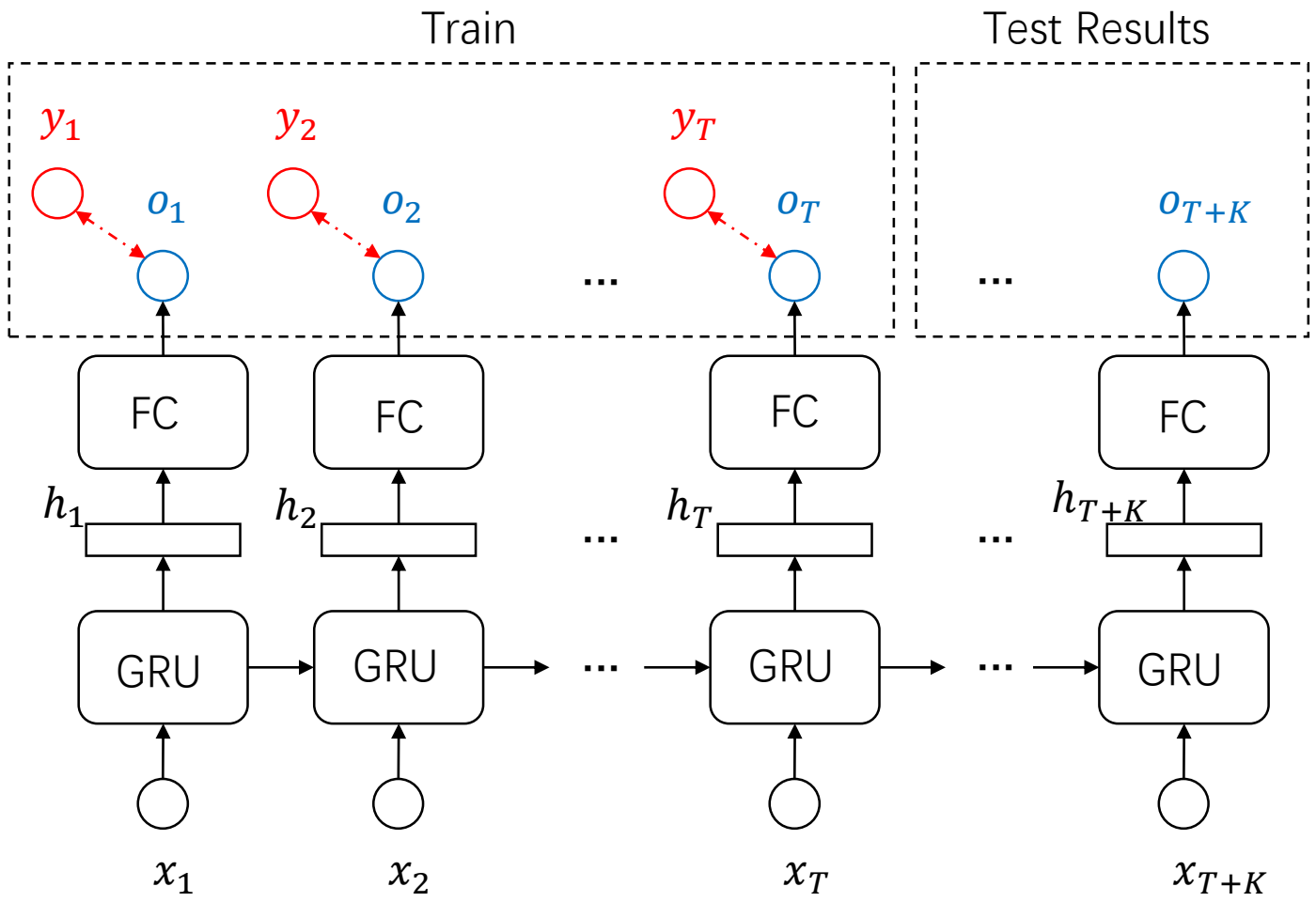
Goal: $\min \sum_{t=1}^{T} (o_t - y_t)^2$

Testing

Inputs: $X_{1:T+K} = (x_1, \cdots, x_T, x_{T+1}, \cdots, x_{T+K})$

Outputs: $O_{1:T+K} = (o_1, \cdots, o_T, o_{T+1}, \cdots, o_{T+K})$

Recurrent Neural Network



Training

For $t = 1, \cdots, T$:

$$h_t = GRU(x_1, \cdots, x_t)$$

$$o_t = W_o^T h_t + b_o$$

$$\min \sum_{t=1}^{T} (o_t - y_t)^2$$

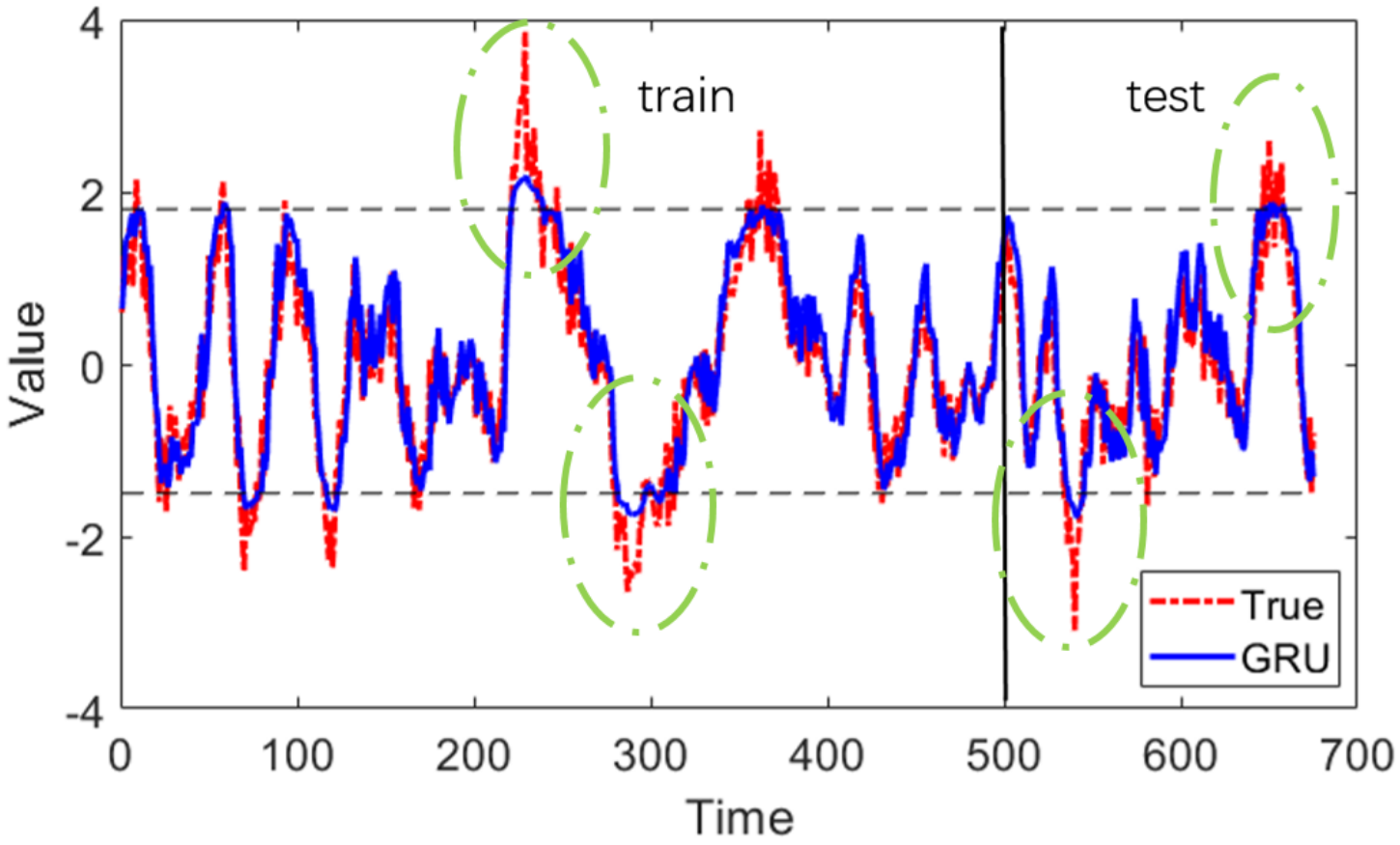
Testing

For $t = 1, \cdots, T + K$:

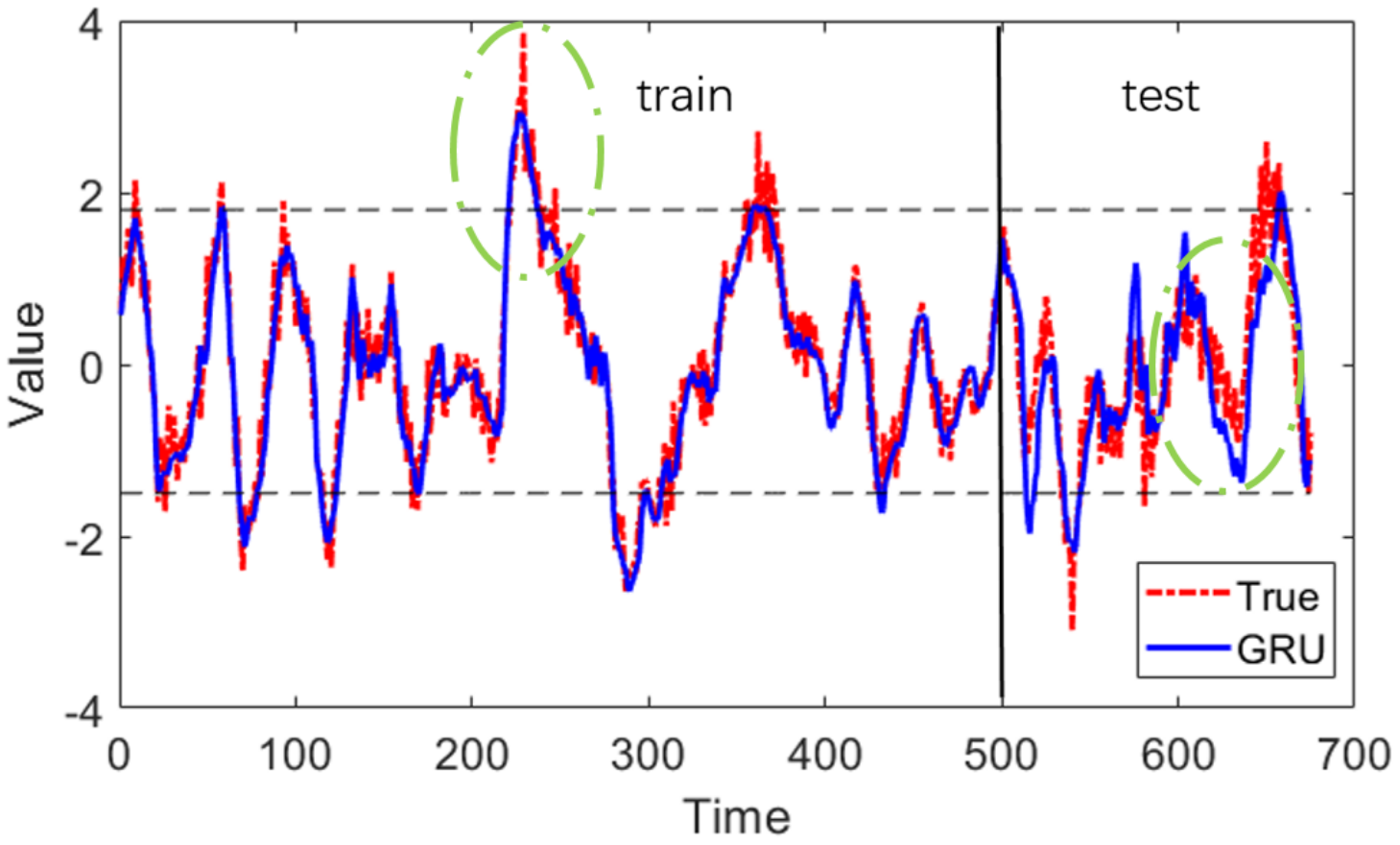
$$h_t = GRU(x_1, \cdots, x_t)$$

$$o_t = W_o^T h_t + b_o$$

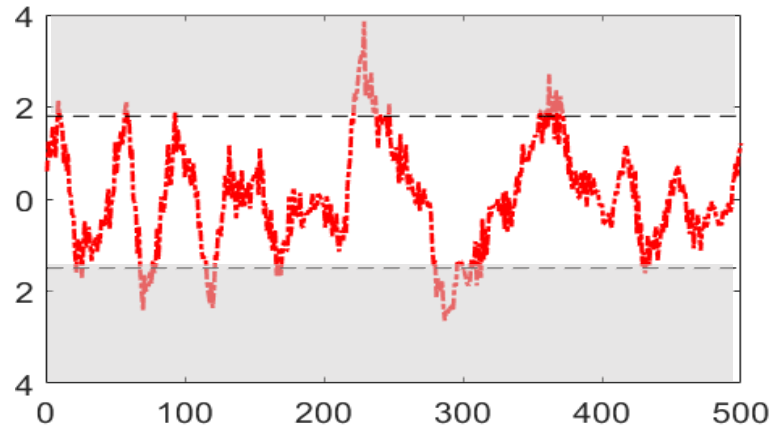
Underfitting Phenomenon

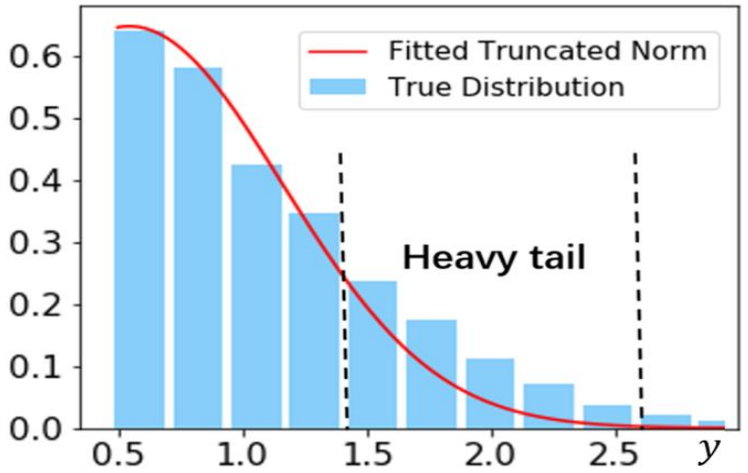


Overfitting Phenomenon



Extreme Events in Time Series Data





Characteristic

- Extremely small or large values
- Irregular
- Rare occurrences
- Light-tailed distributions (Gaussian, Poisson, etc.) cannot model them well

Problem

- Why Deep Neural Network could suffer extreme event problem in time series prediction?
- How can we improve the performance on the prediction of extreme events?

Estimated Distribution of Labels y_t

- The optimization of deep neural network under probability perspective:

$$\min \sum_{t=1}^{T}(o_t - y_t)^2 \xleftrightarrow[\text{Divergence}]{\text{Bregman}} \max \prod_{t=1}^{T} \mathcal{N}(y_t|o_t, \hat{\tau}^2) \Leftrightarrow \max \prod_{t=1}^{T} P(y_t|x_t, \theta)$$

- With Bayes Theorem,

$$P(Y|X,\theta) = \frac{P(X|Y,\theta)\,P(Y)}{P(X|\theta)}$$

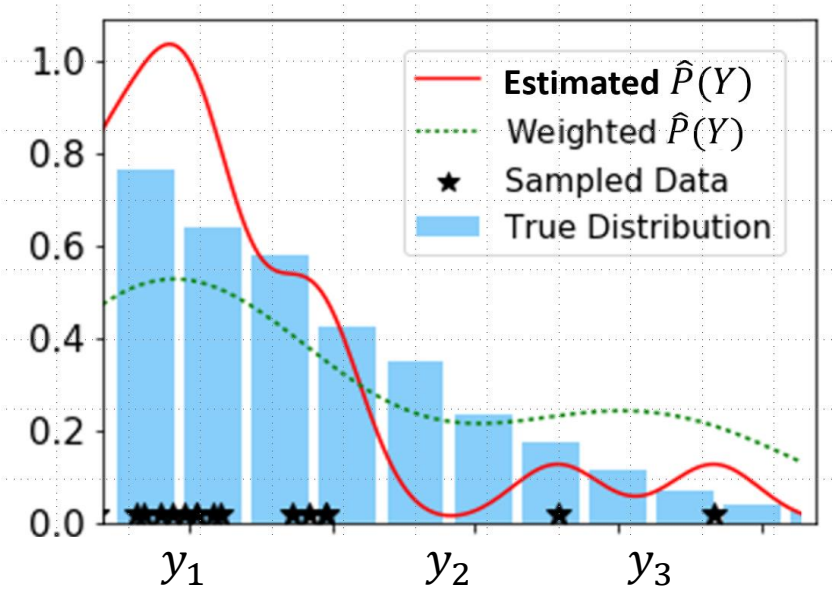
Likelihood

Posterior

Estimated distribution of labels $\hat{P}(Y) = \frac{1}{T}\sum_{t=1}^{T}\mathcal{N}(y_t, \hat{\tau}^2)$

- DNN will internally estimate the distribution of y_t according to the sampled data.

Extreme Event Problem in DNN



Underfitting Phenomenon

- For those normal points, e.g., y_1,

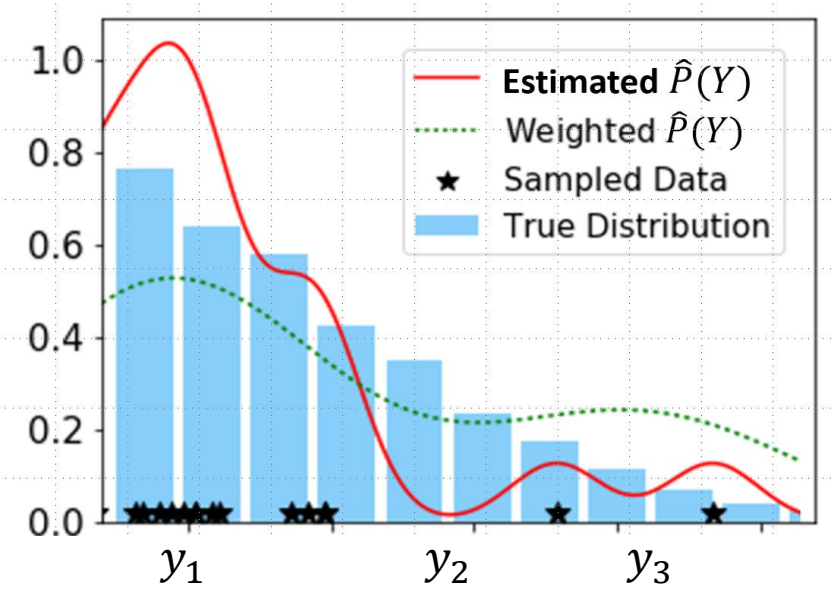
$$P(y_1|X,\theta) = \frac{P(X|y_1,\theta)\hat{P}(y_1)}{P(X,\theta)} \geq \frac{P(X|y_1,\theta)P_{true}(y_1)}{P(X,\theta)} = P_{true}(y_1|X,\theta)$$

- For those rarely occurred extreme events, e.g., y_2,

$$P(y_2|X,\theta) = \frac{P(X|y_2,\theta)\hat{P}(y_2)}{P(X,\theta)} \leq \frac{P(X|y_2,\theta)P_{true}(y_2)}{P(X,\theta)} = P_{true}(y_2|X,\theta)$$

- Therefore model commonly lacks the ability of predicting extreme events

Extreme Event Problem in DNN



Overfitting Phenomenon

- If we add weights of extreme events during the training
- For those normal points, e.g., y_1,

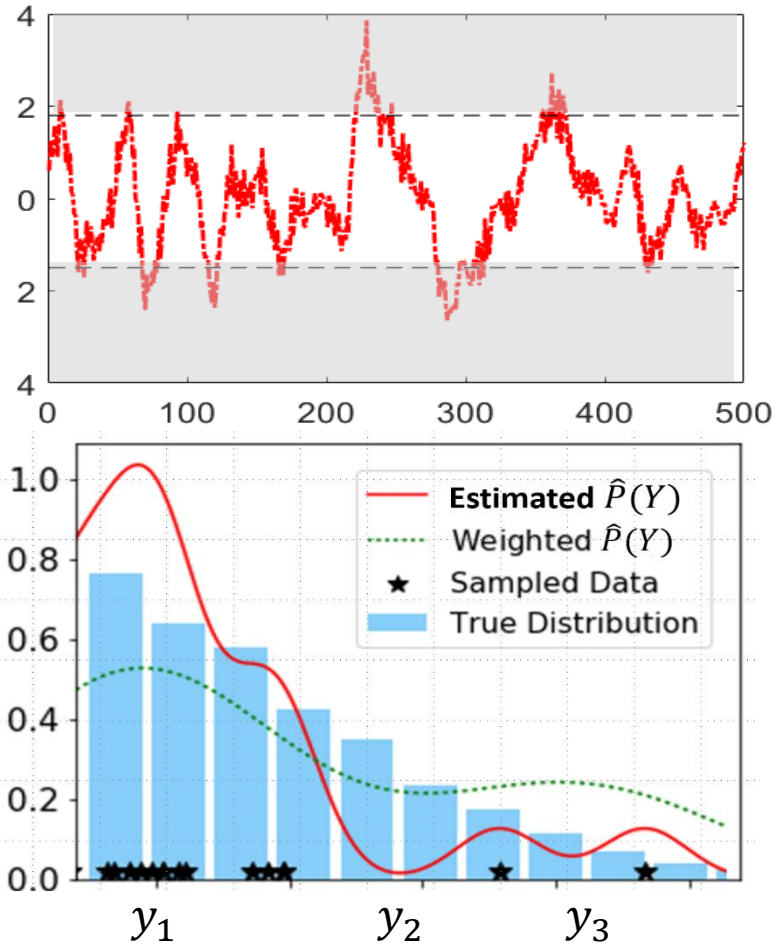
$$P(y_1|X,\theta) = \frac{P(X|y_1,\theta)\hat{P}(y_1)}{P(X,\theta)} \leq \frac{P(X|y_1,\theta)P_{true}(y_1)}{P(X,\theta)} = P_{true}(y_1|X,\theta)$$

- For those rarely occurred extreme events, e.g., y_3,

$$P(y_3|X,\theta) = \frac{P(X|y_3,\theta)\hat{P}(y_3)}{P(X,\theta)} \geq \frac{P(X|y_3,\theta)P_{true}(y_3)}{P(X,\theta)} = P_{true}(y_3|X,\theta)$$

- The estimated distribution is not accurate
- The performance on test data is poor

Problem Analysis



Extreme Event Problem in DNN mainly because:

- Extreme events are extremely large or small values with rare occurrence. Therefore it is hard to estimate the true distribution of them given limited samples.

- Usually DNN learns time series data from light-tailed likelihood, which further increases the difficulty of estimating the distribution of extreme events.

Motivation: Find the regularity inside irregular extreme events

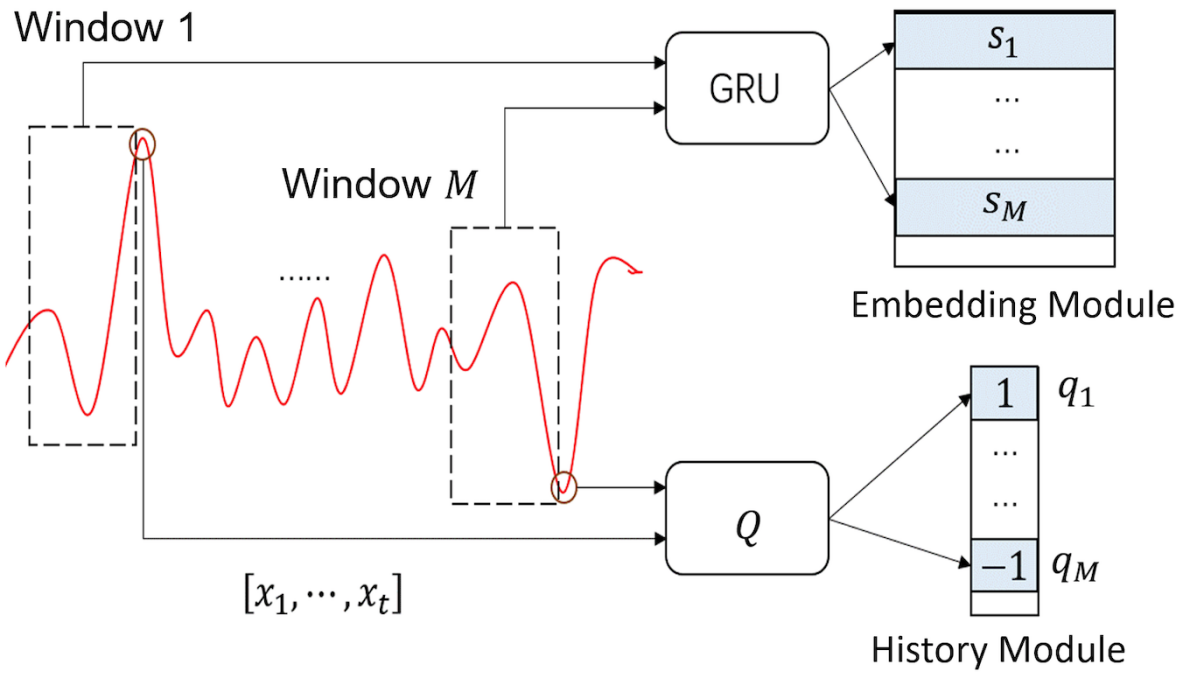
S&P 500

According to previous research：

- Extreme events in time-series data often show some form of temporal regularity.
- Randomness of extreme events have limited degrees of freedom (DOF).

The pattern of extreme events after a window could be memorized！

Recalling Extreme Events in History

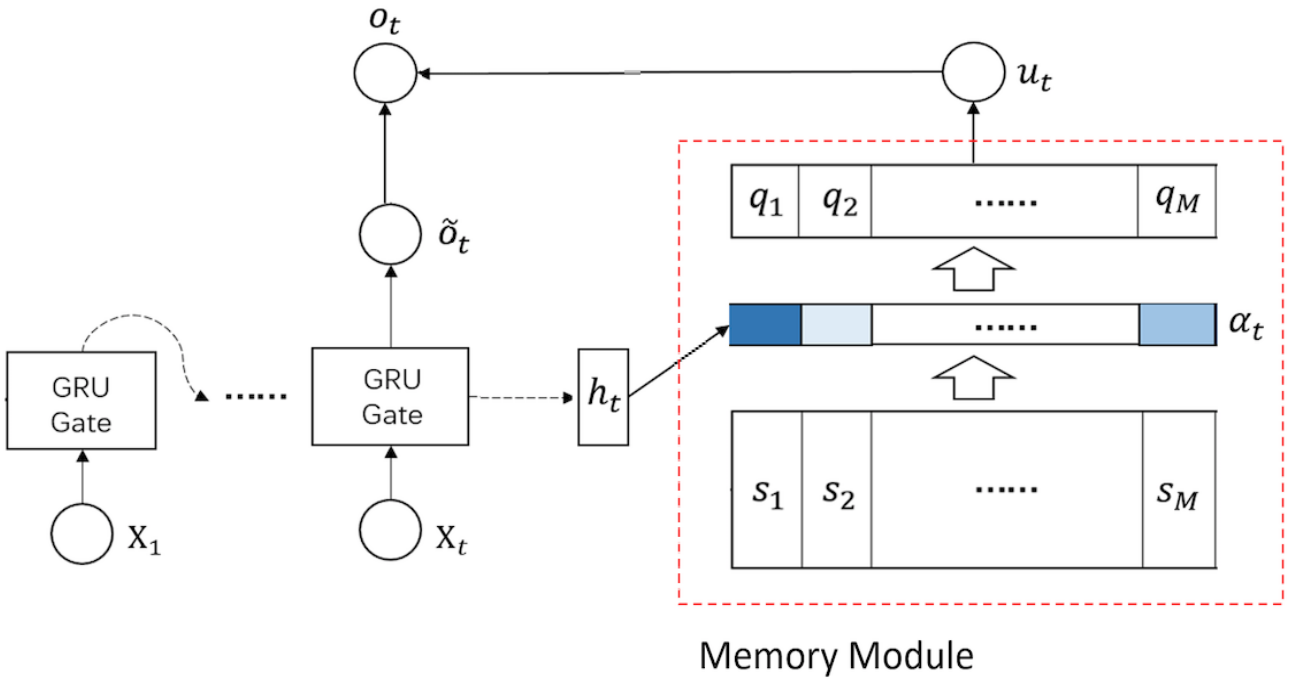


Memory Module

We propose to use Memory Network to recall extreme events in history:

- For each time step t, we sample M windows.
- For window j , we propose to use GRU to calculate the feature s_j of the window.
- Meanwhile, we also record the occurrence of extreme events $q_j = \{-1,0,1\}$ by setting threshold previously at the next time step of window j.

Attention Mechanism



We propose to use attention to incorporate memory module with the prediction:

- At time t, we first calculate the output from GRU:

$$\tilde{o}_t = W_o^T h_t + b_o \ , \ h_t = GRU(x_1, \cdots, x_t)$$

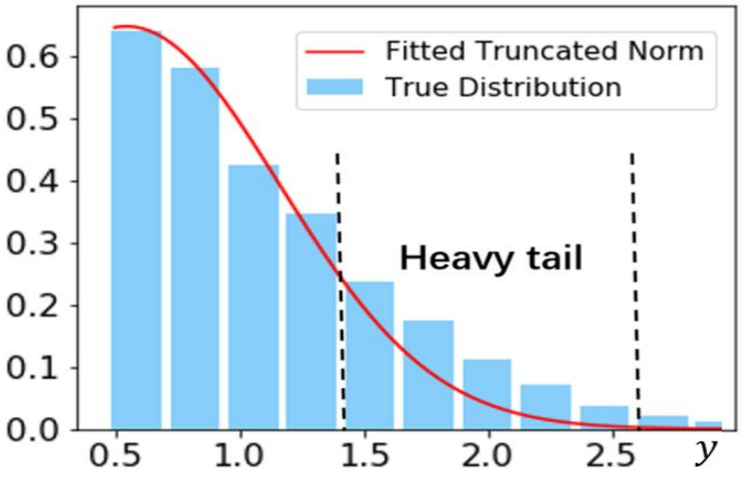
- Then we construct the memory module, and calculate the similarity between the current and the history:

$$\alpha_{tj} = \frac{\exp\left(h_t^T s_j\right)}{\sum_{l=1}^{M} \exp(h_t^T s_l)}$$

- The final output from our model is,

$$o_t = \tilde{o}_t + b \cdot u_t \ , \ u_t = \sum_{j=1}^{M} \alpha_{tj} \cdot q_j$$

Extreme Value Theory



If we still use Gaussian likelihood, the improved model still suffer extreme event problem:

- We should use a heavy-tailed likelihood to fit the distribution of extreme events given limited samples.

It is hard to predict the values of extreme events, however, the DOF of extreme events are easier to be modelled.

- We could propose a heavy-tailed likelihood for predicting the occurrence of extreme events.

Extreme Value Loss

- Through Extreme Value Theory (EVT), the approximation of y_t from EVT can be written as,

$$1 - F(y_t) \approx \big(1 - P(v_t = 1)\big) \log G\left(\frac{y_t - \epsilon_1}{f(\epsilon_1)}\right)$$

Scale function

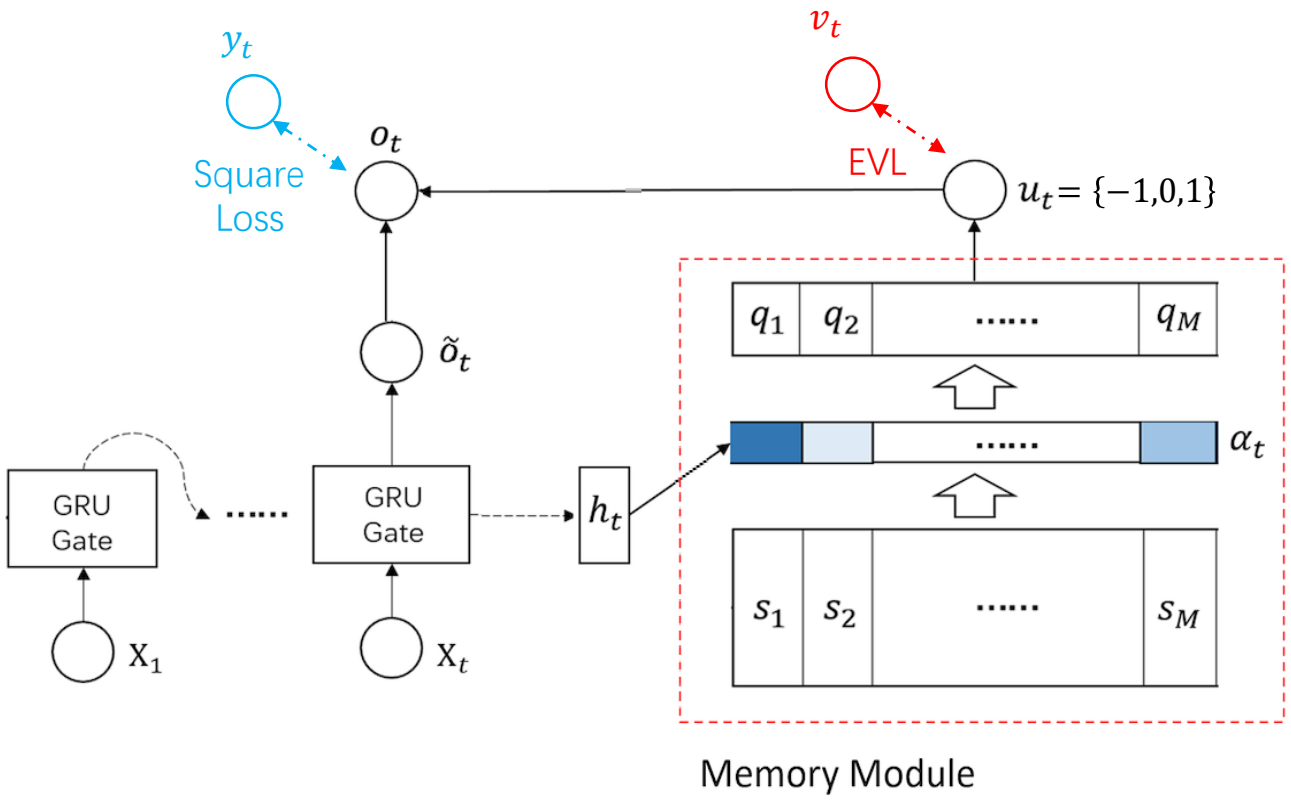
- $v_t = \{0,1\}$ is the indicator of whether a large value will happen or not.
- If we pay our attention to predict whether there is an **extremely large value** at t by outputting $u_t = \{0,1\}$, we can add the weights of extreme events on binary cross entropy loss:

$$\begin{aligned} EVL(u_t) = &- \big(1 - P(v_t = 1)\big)\big[\log G(u_t)\big] v_t \log(u_t) \\ &- \big(1 - P(v_t = 0)\big)\big[\log G(1 - u_t)\big](1 - v_t)\log(1 - u_t) \\ = &- \beta_0 \left[1 - \frac{u_t}{\gamma}\right]^{\gamma} v_t \log(u_t) \\ &- \beta_1 \left[1 - \frac{1 - u_t}{\gamma}\right]^{\gamma} (1 - v_t)\log(1 - u_t) \end{aligned}$$

Binary cross entropy loss

- It is easy to extend the binary classification to $\mathrm{u_t}, \mathrm{v_t} = \{-1,0,1\}$.

Optimization



The final loss function can be written as:

$$\sum_{t=1}^{T} \|o_t - y_t\|^2 + \lambda_1 EVL(u_t, v_t)$$

For the two challenges in DNN:

- We predict the labels from both GRU and memory module, which memorizes the regularity inside extreme events given limited samples.
- We propose to minimize a heavy-tailed classification loss (EVL) for detecting the occurrence of extreme events.

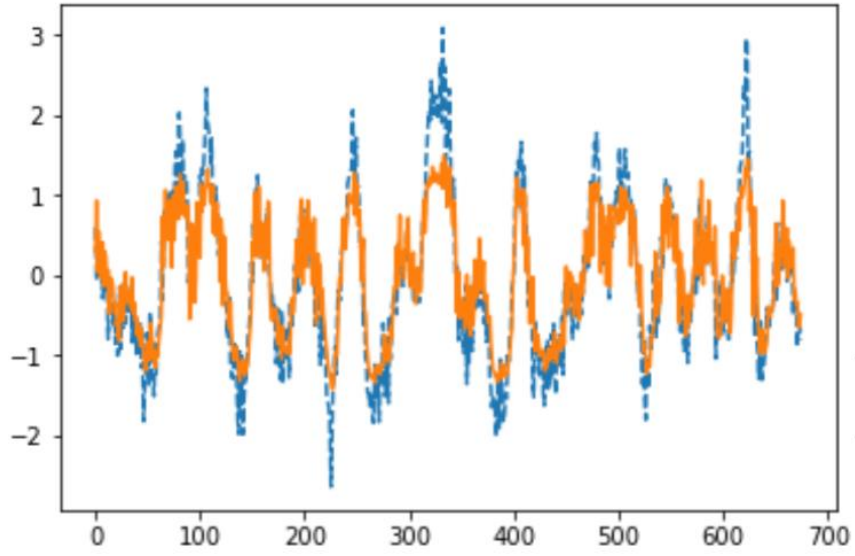
Experimental Settings

- **Dataset:**
  - Stock Dataset: 564 corporations in Nasdaq Stock Market with one sample per week
  - Climate Dataset: Green Gas Observing Network dataset and Atmospheric Co2 Dataset
  - Pseudo Periodic Synthetic Dataset
- **Baselines:**
  - LSTM
  - GRU
  - Time-LSTM
- **Research questions:**
  - RQ1: Is our proposed framework effective in time series prediction?
  - RQ2: Is our proposed loss function EVL worked in detecting extreme events?
  - RQ3: What is the influence of hyper-parameters in the framework?

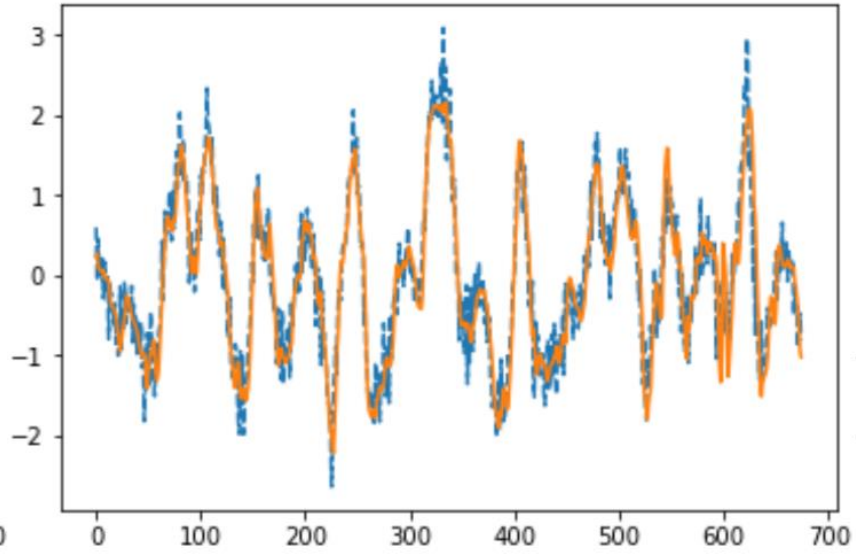
Time Series Prediction (RMSE)

	Climate	Stock	Pseudo
LSTM	0.188	0.249	5.2×10^{-3}
Time-LSTM	0.193	0.256	4.7×10^{-3}
GRU	0.174	0.223	5.3×10^{-3}
Mem	0.181	0.197	3.6×10^{-3}
Mem+EVL	**0.125**	**0.168**	$\mathbf{2.5 \times 10^{-3}}$

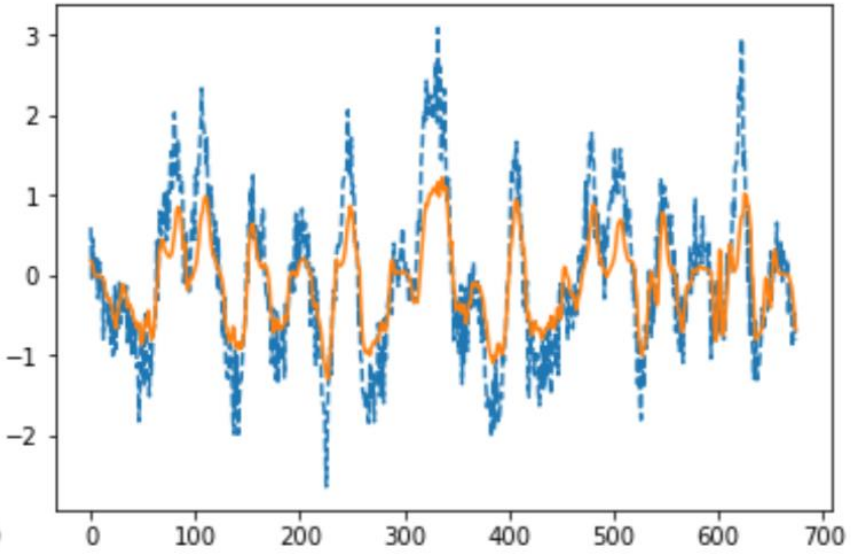
Time Series Prediction (Visualization)



(a) Output from GRU



(b) Output from our model

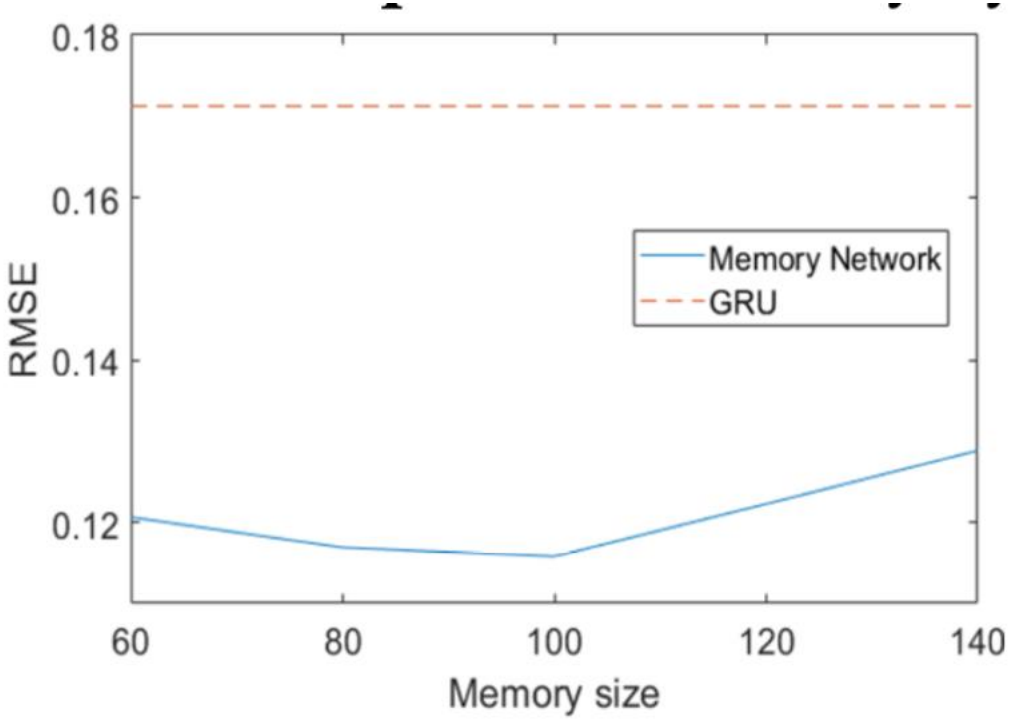


(c) Output from $\tilde{o}_t$

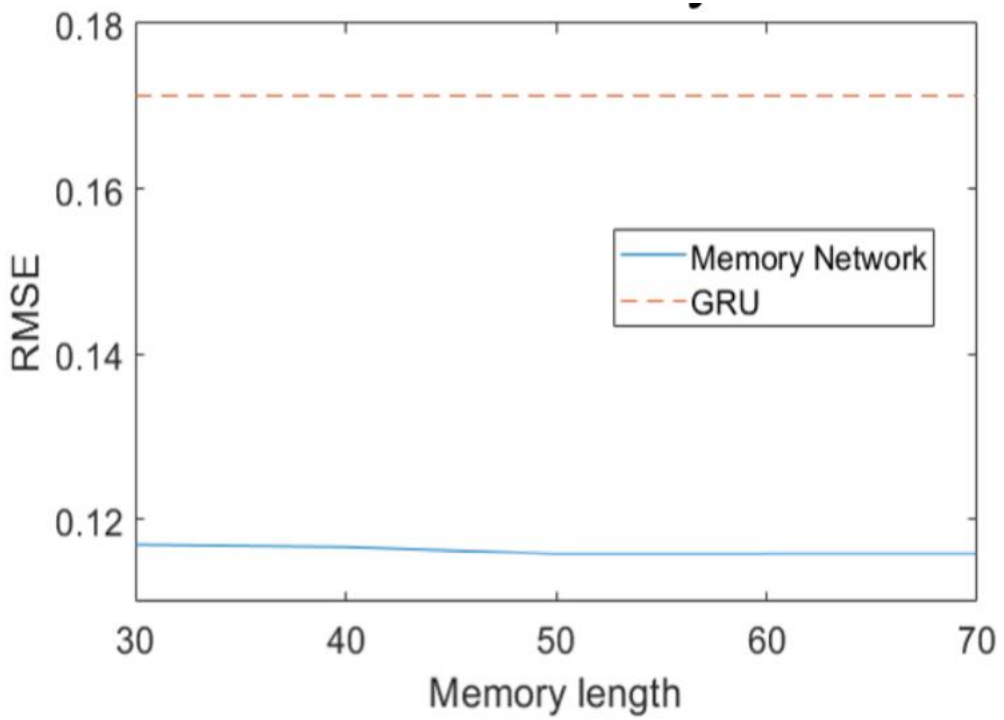
Extreme Events Prediction (F1 Value)

Model	Climate			Stock			Pseudo		
	Micro	Macro	Weighted	Micro	Macro	Weighted	Micro	Macro	Weighted
LSTM+CE	0.435	0.833	0.786	0.247	0.617	0.527	0.830	0.900	0.899
GRU+CE	0.471	0.717	0.733	0.250	0.617	0.547	0.854	0.917	0.917
GRU+EVL ($\gamma = 0.5$)	0.644	0.883	0.859	0.281	0.583	0.523	0.833	0.900	0.902
GRU+EVL ($\gamma = 1.0$)	**0.690**	**0.900**	**0.881**	0.267	**0.667**	0.547	**0.874**	**0.933**	**0.933**
GRU+EVL ($\gamma = 2.0$)	0.646	0.867	0.851	**0.324**	0.617	**0.555**	0.869	0.917	0.920
GRU+EVL ($\gamma = 3.0$)	0.508	0.867	0.825	0.295	0.617	0.548	0.810	0.900	0.897
GRU+EVL ($\gamma = 4.0$)	0.617	0.817	0.813	0.295	0.617	0.543	0.825	0.883	0.886

Influence of hyper-parameters



(a) Influence of memory size M



(b) Influence of window size Δ

Conclusion

- In this paper, we pay our attention to extreme events in time series data.

- We first analysis why DNN is innately weak in predicting extreme events:
  - Extreme events are extremely large or small values with rare occurrence, it is hard to model them with limited samples.
  - The commonly used Gaussian likelihood is a light-tailed distribution.

- We further propose a framework to improve the performance on time series prediction:
  - For the first challenge, we propose to use Memory Network to recall extreme events in history.
  - For the second challenge, we propose a new loss function called extreme value loss (EVL).

- Empirical results show the effectiveness of our proposed framework.

Future Work

- Extending our work to multi-dimensional time series data.
- Applying EVL to more kinds of tasks.
- ……

Thank you for listening!

If you have any questions, please contact Daizong Ding

Email: 17110240010@fudan.edu.cn

Wechat: