Time Series Prediction and Anomaly Detection of Light Curve Using LSTM Neural Network

Runtian Zhang$^{1,a}$ and Qian Zou$^{1,b}$

$^1$Department of Information, Beijing University of Technology, Beijing, China

*artzhang9317@126.com; ^zouqian@emails.bjut.edu.cn

Abstract. Ground-based Wide-angle Camera array (GWAC) is a short-time survey telescope which can produce images every 15 seconds for more than 30,000 stars. Light curve is generated from star image with a series of processing. Research on light curve is a new task in time domain astronomy which can detect anomaly astronomical events. We explore a series prediction model of LSTM neural network for light curve prediction. Through model training and validation we obtain the optimal structure. Then we predict one time-step ahead light luminance for test star. We evaluate the performance of model by calculating prediction error. Anomaly detection mechanism is based on prediction error. Experimental results based on real light curve data demonstrates that our model is promising in light curve prediction and anomaly detection.

1. Introduction
With the rapid development of astronomical observed capability and data processing technology, an enormous amount of data is increasing exponentially. Monitoring data plays a critical role of astronomical studies [1], such as discovery of early supernova eruption, fast response observation of gamma-ray burst and light curve observation [2]. Light curve research is a new task in time domain astronomy [3]. Large datasets are generated by sky survey and transformed to time-series data. GWAC (Ground-based Wide-angle Camera array) is the important equipment in Space Multi-band Variable Objects Monitor mission [4]-[6], which has the capability of providing astronomical data in short time-scale. SVOM aims to study gamma-ray burst and this phenomenon leads to a sudden abnormal of light luminance. Research on anomaly detection of real-time light luminance (light curve) is significant which is the purpose of this paper.

Traditional time series anomaly detection is based on statistical approach. For instance, an Auto Regressive Moving Average (ARMA) based model [7] simulates stationary time sequences over a time window to detect anomaly values. The length of window is pre-determined and results are largely depends on it. The Kalman filter [8]-[9] is an algorithm for the optimal estimation of the system states. It is considered as a filtering process due to the observation data includes the influence of noise and interference in the system. It is obvious that anomaly detection is generally based on comparing observation and its predicted value at the same time step. RNN(Recurrent Neural Networks) is a neural network that captures the dynamic information in serialized data by periodically connecting hidden layers. LSTM (Long Short-Term Memory neural network) is an improved recurrent neural network. It overcomes gradient vanished problems during long time data training process by using memory cell. [9] This characteristic makes LSTM networks applicable to time series problem.
It’s a worthwhile study that combining this advantage of predicting time series with anomaly detection based on time prediction. As in [10], it implements anomaly captured by training normal sequence and anomaly sequence. However, the difficulty of our purpose is that light curve is streaming data based on star luminance which has no periodicity and reference anomaly sequence. And the observed data is obtained every 15 seconds, which means anomalies need to be captured within 15 seconds for each spot. To the best of our knowledge, it is the first time that the LSTM neural network is used on astronomical time problem.

The rest of the paper is organized as follows. Section 2 describes the detail of light curve dataset and real-time series acquisition mechanism. Section 3 introduces the LSTM neural network model for prediction and data pre-processing method. Section 4 presents our experiment based on the LSTM neural network using light curve dataset, introduces anomaly detection mechanism, analyses the performance of our model using results. Section 5 concludes this paper.

2. Light curve dataset
Light curve is a series of light luminance data over time. The dataset is generated from GWAC which called mini-GWAC. It is obtained from the National Astronomical Observatories of China. The detail of our dataset is as following:

- GWAC can produce images every 15 seconds for more than 30,000 stars. After a series of processing, such as quality control, source extraction, flux calibration, source association, each time-step light luminance value generated.
- The amount of each star-captured data is almost 900 per day. GWAC is in trial operation and our dataset contains three days of observations.
- We verify the result by using the data of object with catalog ID of 3348.

In present, we only have three days data for evaluating our approach. But after the formal operation of GWAC, a huge amount of light luminance data generated every day. With those history data, our experiment will be more effective in the future.

3. Methodology
3.1 The structure of LSTM
LSTM neural network has a great contribution to solving real-world problems in many filed. Comparing with RNN, the most significant improvement of LSTM is having a complex structure called memory cell in hidden layer. As is shown in Figure 1, $x_t$ is the input layer sequence at time t, A is a hidden layer with memory cell, $h_t$ is the hidden layer output sequence at time t. In the process of data transfer in the hidden layer, $C_t$ is a special information which contains the state of cell. The main purpose of the transition is to update the cell state and output $h_t$ at time t. The specific transition in LSTM cell is shown in Figure 2. Each of the yellow block is neural network layer which reads data from $h_{t-1}$ and $x_t$. Blue blocks are calculative units which are designed for vector operation.

![Figure 1. The structure of LSTM with series prediction](image)

![Figure 2. The structure of LSTM cell](image)
We firstly decide what information need to be discarded which is generated by forget gate. It outputs results from 0 to 1 for each number in cell state $C_{t-1}$. “0” is represented “completely discarded”, analogously, “1” is represented “completely reservation”. Next step is the decision of choosing new information of cell state. The first part is input gate which outputs $i_t$ that need to be updated. “tanh” layer creates a new vector $\hat{C}_t$. Finally, the output gate’s state is $o_t$. Both $C_{t-1}$ and $h_{t-1}$ are well transmitted under the structure of LSTM cell. Equations of the transmission process is shown below:

- **Forget gate’s state:**
  \[
  f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f)
  \]  
  \(1\)

- **Input gate’s state:**
  \[
  i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i)
  \]
  \(2\)

- **Updated $C_t$:**
  \[
  \hat{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] + b_C)
  \]
  \(-3\)

- **Output gate’s state:**
  \[
  o_t = \sigma(W_o[h_{t-1}, x_t] + b_o)
  \]
  \(4\)

- **Output $h_t$:**
  \[
  h_t = o_t \cdot \tanh(C_t)
  \]
  \(6\)

For parameters in equations (1)-(6), $\sigma$ represents the sigmoid activation function $(1 + exp^{-x})^{-1}$, tanh means the hyperbolic tangent activation function $(exp^x - exp^{-x}) \cdot (exp^x + exp^{-x})^{-1}, W_f, W_i, W_C, W_o$ are the weight matrices for both $x_t$ and $h_{t-1}$ in each gate, $b_f, b_i, b_C, b_o$ are bias of each gate.

### 3.2 Stationary

The result of LSTM model will be more flexible if time series is stationary. There is a huge fluctuation of light luminance if gamma-ray burst is happening. Besides, the history data is merely 3 days. It’s uncertain if history data has a long-term trend. Considering all of these factors, it’s essential to make sequence stationary. A common method for stationary is “differencing”. It’s the observation that previous time step $(t-1)$ is subtracted from current time step $t$. The original history data $X = \{X_1, X_2, \cdots, X_t\}$ is changed. $\hat{X} = \{X_1 - X_0, X_2 - X_1, \cdots, X_t - X_{t-1}\}$ is the new format of historical data. (We assume that $X_0$ equals to $X_1$).

### 4. Experiments

We construct our LSTM model for real-time light curve series from mini-GWAC dataset. The amount of history data for ID 3348 is 2777. We divide it into three parts: training set (75%) as historical data, validation set (15%) for evaluating predicted error and test set (15%) for real-time prediction. In order to simulate capturing anomalies from real-time generated data, test set is considered as the real-time data which is produced by mini-GWAC every 15 seconds one by one. It indicates that our model focus on predicting one time period value.

#### 4.1. Model training

We employ differencing method for data pre-processing. Processed datasets are shown below:

- **Training:**
  \[
  X_{\text{train}} = \{X_{h1} - X_{h0}, X_{h2} - X_{h1}, \cdots, X_{\text{lenH}} - X_{\text{lenH-1}}\}
  \]  
  \(7\)

- **Validation:**
  \[
  X_{\text{val}} = \{X_{v1} - X_{v0}, X_{v2} - X_{v1}, \cdots, X_{\text{lenV}} - X_{\text{lenV-1}}\}
  \]  
  \(8\)

- **Test:**
  \[
  X_{\text{test}} = \{X_{t1} - X_{t0}, X_{t2} - X_{t1}, \cdots, X_{\text{lenT}} - X_{\text{lenT-1}}\}
  \]  
  \(9\)

- **Original:**
  \[
  X_{\text{origin}} = \{X_{h1}, X_{h2}, \cdots, X_{\text{lenH}}, X_{v1}, X_{v2}, \cdots, X_{\text{lenV}}, X_{t1}, X_{t2}, \cdots, X_{\text{lenT}}\}
  \]  
  \(10\)

Where $X_{\text{origin}}$ is original dataset and all of the divided parts are assigned in order of the original timestamp. For instance, $X_{v0} = X_{\text{lenH}}$ and $X_{t0} = X_{\text{lenV}}$. 
The dimension of input layer and output layer is 1 due to the number of light luminance at each time. In series prediction LSTM neural network, there are many factors affecting the performance of model. We evaluate performance by comparing results of different number of hidden units and epoch time. The number of hidden units is ranging from 1 to 5 and epoch time is ranging from 1 to 100. Loss function of the model is “Mean_Squared_Error” which closely matches MSE that is used for calculating predicted error.

As is shown in Figure 3, blue lines are training error and red lines are validation error. Training set error tends to be stable through less than 10 epochs and it is much better than validation error. The number of cells has little effect on the performance. We construct the model with 5 hidden layer units and 100 epoch times of training process for stable result.

4.2. Model Evaluation
After compiling our model, test set predicts the value of next time period by calculating parameters of the model. Then prediction results are converted from differential form to origin form. We measure the prediction error of both train set and test set by mean square error (MSE),

\[ MSE = \frac{1}{N} \sum_{t=1}^{N} (X_t - \tilde{X}_t)^2 \]  

(11)

Where N is the number of sequence, \( X_t \) and \( \tilde{X}_t \) are observed value and predicted value at time step t. We repeat 10 times of the whole prediction process. Statistical result is shown in Table 1. The result of Table 1. Indicates that prediction of train set is much better than test set. It is reasonable that light luminance of test set has a huge fluctuation comparing with train set. The real light curve of ID “3348” is shown in Figure 4.

4.3. Anomaly detection mechanism
Anomaly detection mechanism is based on MSE of both train set and test set. It is used for instant warning if the prediction of current observation Considering the prediction of train set is much better than test set, an experimental ratio is provided to scale error between train set and test set. We measure the anomalies slot by following rule, \( Ano_{slot} \) is determined at time t if.

![Figure 3. “mean_squared_error” of train set and validation set](image)

![Figure 4. The real light curve data of ID 3348](image)

| Table 1. Statistical result of MSE of train set and test set |
|-----------------|-----------------|-----------------|-----------------|
|                 | Max             | Min             | Mean            | Variance        |
| Train set       | 0.013515        | 0.010378        | 0.011908        | 1.107e-6        |
| Test set        | 0.572562        | 0.199559        | 0.323205        | 0.000647        |
\[(Predicted_i - Observed_i)^2 \times R > MSE_{maxtrain}\] (12)

Where \( MSE_{maxtrain} \) is the max MSE of train set, \( R \) is the ratio that is set as 0.3. The mechanism is tested based on both real data and manually added anomalies data. It is significant that testing our approach under varies situation. We manually add random data ranged from 2 to 8, similarly, a sin wave is added on another situation. Figure 5. shows the result with added anomalies data. Blue line is represented both training set and validation set, black dots are test data, red dots are anomalies slots. Figure 6. shows the result with real test set. Our approach finds anomalies value under sudden fluctuation. This is sufficient for the purpose of instant warning. It also generates some false results and miss some anomalies slot such as waved situation. We strongly believe our method can be improved if we have more historical data and some known astronomical principle.

![Figure 5. Test set with manually added anomalies data. (a) is represented results with random anomalies. (b) is represented results with waved anomalies.](image)

![Figure 6. Real test set with anomaly](image)

5. Conclusion
We explorer a LSTM neural network model for light curve time series predictions. Employing light curve data provided by mini-GWAC dataset. Then we propose the anomaly detection mechanism based on prediction error. This mechanism is sufficient for real test set and test set of manually added anomalies slot. In the future, we plan to evaluate our model on GWAC system if is ready to operate on line.
Reference

[1] S. Vaughan, “Random time series in astronomy,” Philosophical Transactions of the Royal Society of London, vol. 371, no.1984, 2012.

[2] Ann M, “The Swift gamma-ray burst explorer Burst Alert Telescope,” Nuclear Science Symposium and Medical Imaging Conference, vol. 4, pp 2382-2386, 2002.

[3] Zhang Li, Bian Xia, Mei Dong-Cheng; “Gamma-ray light curve and phase-resolved spectra from Geminga pulsar,” vol. 10, no. 7, pp 662-665, 2001.

[4] channe S, Paul J, Wei J, “The future gamma-ray burst mission SVOM,” Proceedings of Science, E-ISSN: 18248039, 2009.

[5] Wan Meng, Wu Chao, Ying Zhang, “A Pre-research on GWAC Massive Catalog Data Storage and Processing System,” Astronomical Research and Technology, no. 3, pp. 373-381, 2016.

[6] Wan Meng, “Column store for gwac: a High-cadence, High-density, Large-scale astronomical light curve pipeline and distributed Shared-nothing database,” Publications of the Astronomical Society of the Pacific, no. 969, 2016.

[7] Aliat Billel, Hamdi Faycal, “On Markov-switching periodic ARMA models,” Communication in Statistics, vol. 47, no. 2, pp 344-364, 2017.

[8] Janjic Tijana, Nerger Lars, “On domain localization in ensemble-based Kalman filter algoriths,” Monthly Weather Review, American Meteorological Society, vol. 139, no. 7, pp 2046-2060, 2011.

[9] Han Fei, Dong Hongli, Wang Zidong, “Improved Tobit Kalman filtering for systems with random parameters via conditional expectation,” Signal Processing, vol. 147, pp 35-45, 2018.

[10] Malhotra Pankaj, Vig lovekesh, Shroff Gautam, “Long Short Term Memory networks for anomaly detection in time series,” 23rd European Symposium on Artificial Neural Networks, pp 89-94, 2015.