VAE-LSTM Joint Model for Time Series Prediction and Anomaly Detection

Ilya Mansurov
NATIONAL RESEARCH UNIVERSITY HIGHER SCHOOL OF ECONOMICS

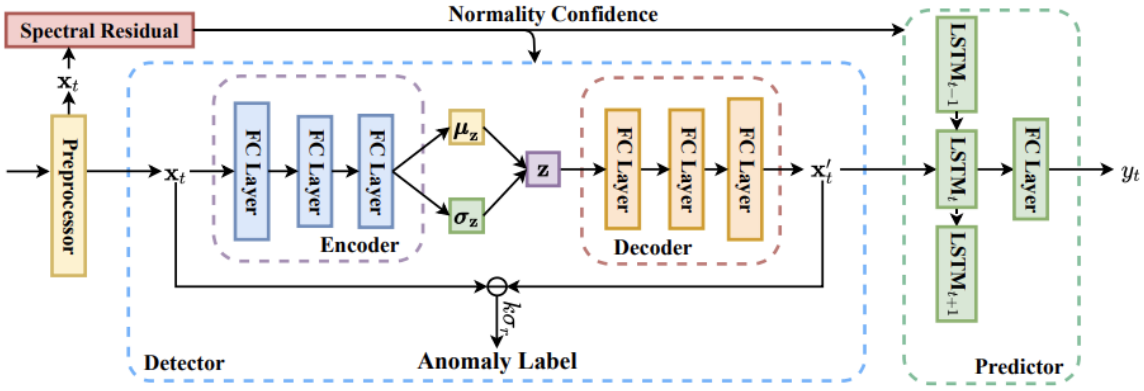
Abstract

In this paper, I will consider the problem of predicting time series values and detecting anomalies. To solve these problems, I use a joint model of variational autoencoder for anomaly detection and long short-term memory for prediction.

1. Introduction

The problem of detecting anomalies and predicting values arises in many areas from healthcare to monitoring IT systems. The difficulty in finding anomalies is the inability to collect a sufficient amount of tagged data. Therefore, using an unsupervised approach is a significant advantage.

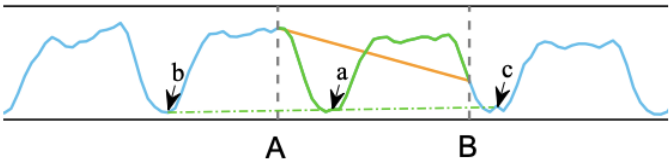
The following model architecture is proposed. The initial time series is divided into sets of vectors x_{t-L} ... x_t using a sliding window of size L = 256 and a step 1. The model consists of two main blocks: anomalies detector(VAE) and predictor(LSTM). The detector takes a vector of initial statuses x_{t-L} ... x_t and reconstructs it into a vector x'_{t-L} ... x'_t. The difference between x_t and x'_t is compared with a threshold. The reconstructed vector x'_{t-L} ... x'_t is fed to the LSTM input to predict the next value in the time window y_t, so y_t is predicted value for the status x_{t+1}. In addition, spectral residual analysis is used to produce the normality confidence weights for each status in each segment x_t



Model architecture (pic. from [1])

2. Dataset

The KPIs dataset is used for the experiment. The data was collected from some web services and computer systems. Web service KPIs consist of performance metrics, such as response time, web page visits, connection errors etc. Computer system KPIs consist of the health status of computers (servers, routers, switches), CPU utilization, memory utilization, disk I/O, network bandwidth etc. The dataset contains 28 timeseries, 22 of which consist of minute intervals, the rest of five minute intervals. For the experiment, I took time series with minute intervals.



The statuses between A and B are missing. Status a fill with the interpolation between b and c. (pic. from [1])

3. Model

3.1. Data preprocessing

There is missing data in the time series. The data is cyclic with a period of 24 hours. This property is used to fill in missing values the same way as in [1]. If the missing interval is less than 7 minutes, linear interpolation along the interval boundaries is used to fill it. If the interval is larger than 7 minutes, linear interpolation is used for statuses that are +-24 hours apart from the missed one. Further, the data is standardized.

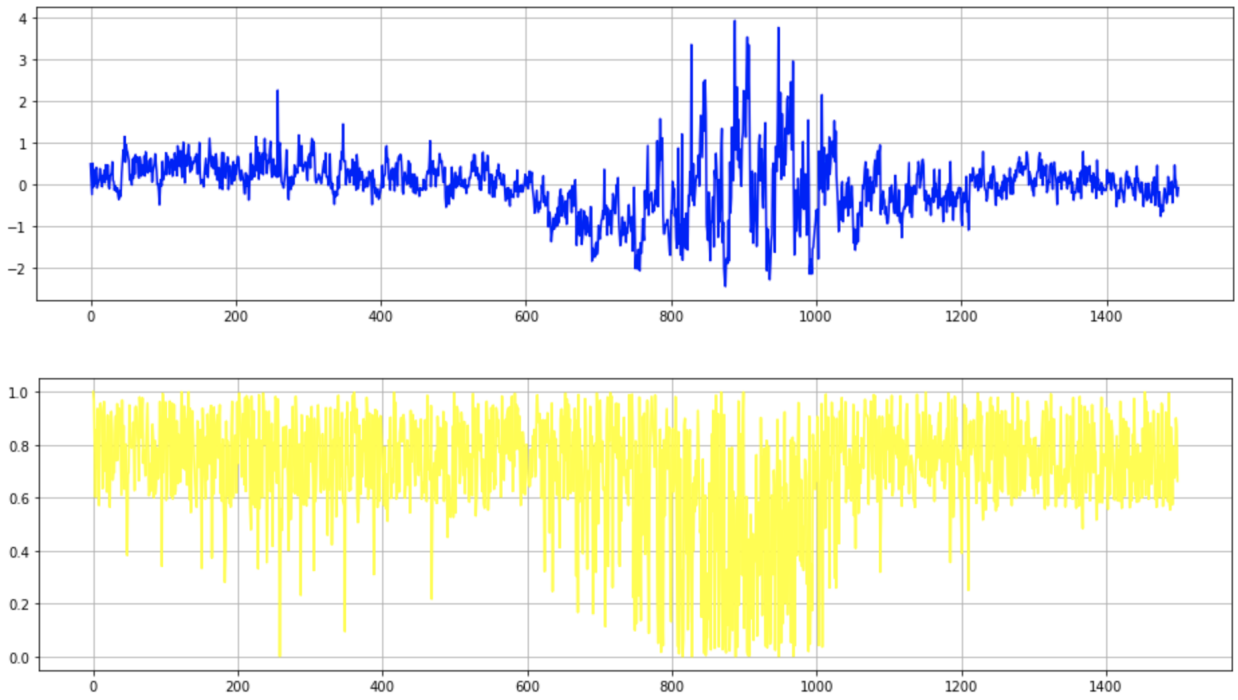
3.2. Normality Confidence Weighting

Spectral residual (SR) analysis is used to construct normality confidence weighting, which are further used as the weighting coefficients of the loss function of the model.
Firstly I calculate 1D saliency map S (x_t). Given the last point S (xt) and the local average of the last point $\hat{S}(x_t)$ in the saliency map S (x_t), the normality confidence at timestamp t is estimated as

$w_n(x_t) = 1 - 1 / 1 + \exp(-(D(x_t)))$, where $D(x_t)$ is

$D(x_t) = (S(x_t) - \hat{S}(x_t)) / \hat{S}(x_t)$

Library sranodec is used for this calculation.



Statuses and its normality confidence

3.3. Anomaly detection

VAE is trained with loss function

$L_{VAE}(x_t) = \|w_n \circ (x_t - x'_t)\|^2_2 + \beta \hat{w}_n / 2 \, (- \log \sigma^2_z + \mu^2_z + \sigma^2_z - 1 \,)$

First term is the reconstruction loss and second is Kullback-Leibler divergence between true posterior distribution $p_\theta(z|x)$ and the approximate posterior distribution is assumed to follow a diagonal Gaussian distribution $q_\varphi(z|x) = N(\mu_z, \sigma^2_z \cdot I)$

w_n is the normality confidence of statuses in segment x_t and $\hat{w}_n$ is the average over w_n

β is the hyper parameter to balance between first and second terms.

Due to the anomalies being a rare occurrence, the distribution of these statuses is different from those of the normal statuses and ideally anomalies are not reconstructed by the decoder. x_t is viewed as abnormal when the absolute error of x_t from x'_t is higher than threshold kσ, where k is fixed on the whole dataset and σ is standard deviation of $|x_t - x'_t|$ on the training data. So, when $|x_t - x'_t| > k\sigma$ it's mark as anomaly

3.4. Prediction

LSTM block is trained with loss function

$L_{LSTM}(x_t) = \hat{w}_n \|(x_{t+1} - y_t)\|^2_2$

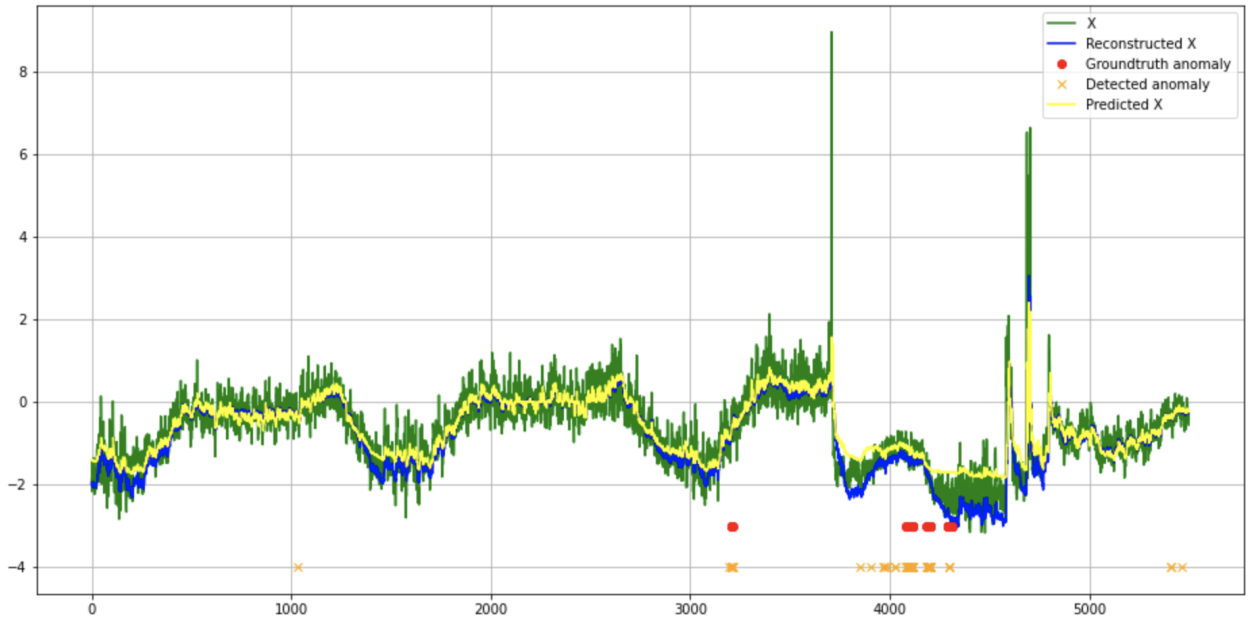
3. Experiments

For the experiment, I took 4 versions of the models PAD, PAD-, AD, P-AD. PAD: predictor (LSTM) trained together with anomaly detector (VAE), taking into account normality confidence weighting. PAD - the same as in the previous case, but without use of the normality confidence weighting. AD is a predictor(VAE) only. P-AD separately trained predictor (LSTM) and anomaly detector (VAE).

The sliding window size is set to 256. Learning rate is set to 1-e3. The number of VAE z dimensions is set to 32. β is set 0.01, 0.1, 1. ƛ is set to 1 and 10. I divide the dataset into training, validation and testing sets, whose ratios are 50%, 5%, 45% respectively..

For the predictor I used MSE, MAE metrics. F1, precision and recall with adjusted anomalies label's(with delay set to 7) are used for the detector. The results of calculations on the test dataset are presented in the tables below

	MSE	**MAE**	**F1**	**Precision**	**Recall**
PAD beta=0.01	0.104	0.173	**0.653**	0.722	0.595
PAD beta=0.1	0.116	0.191	0.579	0.526	0.644
PAD beta=1	0.979	0.787	0.454	0.591	0.368
PAD beta=10	0.982	0.785	0.421	0.560	0.337
PAD- beta=0.01	0.162	0.197	0.581	0.514	**0.669**
P-AD beta=0.01	**0.102**	**0.171**	-	-	-
AD beta=0.01	-	-	0.224	**0.962**	0.120



Visualization of the model evaluation on time series #27

4. Conclusion

The results of the experiment show that the joint training of models has a positive effect on the results of both tasks. The predictor, taking as input the reconstructed data from the VAE, gives the predictions robust to outliers in the original data. LSTM helps VAE to maintain the long term sequential patterns that are out of the VAE encoding window. Comparison results of PAD and PAD- show that SR boosts the performance of both VAE and LSTM.

References

1. Run-Qing Chen, Guang-Hui Shi, Wan-Lei Zhao, Chang-Hui Liang. A Joint Model for IT Operation Series Prediction and Anomaly Detection

2. D. P. Kingma, M. Welling, Auto-Encoding Variational Bayes

3. H. Ren, Q. Zhang, B. Xu, Y. Wang, C. Yi, C. Huang, X. Kou, T. Xing, M. Yang, J. Tong, Time-Series Anomaly Detection Service at Microsoft

4. H. Xu, Y. Feng, J. Chen, Z. Wang, H. Qiao, W. Chen, N. Zhao, Z. Li, J. Bu, Z. Li, Y. Liu, Y. Zhao, D. Pei, Unsupervised Anomaly Detection via Variational Auto-Encoder for Seasonal KPIs in Web Applications

5. AIOpsChallenge, KPI Anomaly Detection Competition (2017). URL http://iops.ai/competition_detail/?competition_id= 5