ESANN 2015 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 22-24 April 2015, i6doc.com publ., ISBN 978-287587014-8.
Available from http://www.i6doc.com/en/.

Long Short Term Memory Networks for Anomaly Detection in Time Series

Pankaj Malhotra[1], Lovekesh Vig[2], Gautam Shroff[1], Puneet Agarwal[1]

1- TCS Research, Delhi, India

2- Jawaharlal Nehru University, New Delhi, India

Abstract. Long Short Term Memory (LSTM) networks have been demonstrated to be particularly useful for learning sequences containing longer term patterns of unknown length, due to their ability to maintain long term memory. Stacking recurrent hidden layers in such networks also enables the learning of higher level temporal features, for faster learning with sparser representations. In this paper, we use stacked LSTM networks for anomaly/fault detection in time series. A network is trained on non-anomalous data and used as a predictor over a number of time steps. The resulting prediction errors are modeled as a multivariate Gaussian distribution, which is used to assess the likelihood of anomalous behavior. The efficacy of this approach is demonstrated on four datasets: ECG, space shuttle, power demand, and multi-sensor engine dataset.

1 Introduction

Traditional process monitoring techniques use statistical measures such as cumulative sum (CUSUM) and exponentially weighted moving average (EWMA) over a time window [1] to detect changes in the underlying distribution. The length of this time window generally needs to be pre-determined and the results greatly depend on this parameter. LSTM neural networks [2] overcome the vanishing gradient problem experienced by recurrent neural networks (RNNs) by employing multiplicative gates that enforce constant error flow through the internal states of special units called 'memory cells'. The input (I_G), output (O_G), and forget (F_G) gates prevent memory contents from being perturbed by irrelevant inputs and outputs (refer Fig. 1(a)), thereby allowing for long term memory storage. Because of this ability to learn long term correlations in a sequence, LSTM networks obviate the need for a pre-specified time window and are capable of accurately modelling complex multivariate sequences. *In this paper, we demonstrate that by modelling the normal behaviour of a time series via stacked LSTM networks, we can accurately detect deviations from normal behaviour without any pre-specified context window or preprocessing.*

It has been shown that stacking recurrent hidden layers of sigmoidal activation units in a network more naturally captures the structure of time series and allows for processing time series at different time scales [3]. A notable instance of using hierarchical temporal processing for anomaly detection is the Hierarchical Temporal Memory (HTM) system that attempts to mimic the hierarchy of cells, regions, and levels in the neocortex [4]. Also, temporal anomaly detection approaches like [5, 6] learn to predict time series and use prediction errors to detect

ESANN 2015 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 22-24 April 2015, i6doc.com publ., ISBN 978-287587014-8.
Available from http://www.i6doc.com/en/.

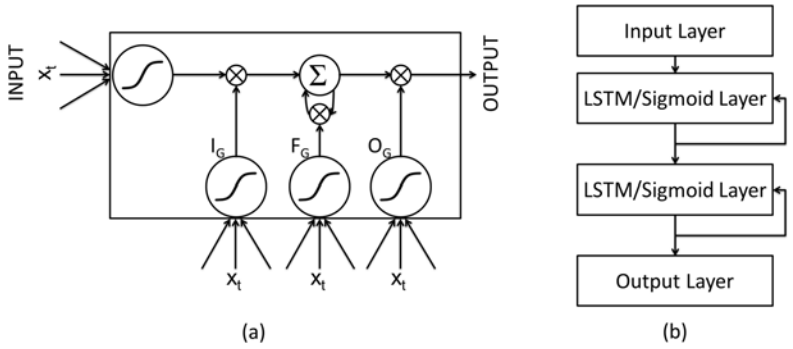

Fig. 1: (a) Long Short-term Memory Cell (b) Stacked Architecture

novelty. However, to the best of our knowledge the retentive power that LSTMs have to offer has not been combined with recurrent hierarchical processing layers for predicting time series and using it for anomaly detection.

As in [5], we use a predictor to model normal behaviour, and subsequently use the prediction errors to identify abnormal behaviour. (This is particularly helpful in real-world anomaly detection scenarios where instances of normal behaviour may be available in abundance but instances of anomalous behaviour are rare.) In order to ensure that the networks capture the temporal structure of the sequence, we predict several time steps into the future. Thus each point in the sequence has multiple corresponding predicted values made at different points in the past, giving rise to multiple error values. The probability distribution of the errors made while predicting on normal data is then used to obtain the likelihood of normal behaviour on the test data. When control variables (such as vehicle accelerator or brake) are also present, the network is made to predict the control variable in addition to the dependent variables. This forces the network to learn the normal usage patterns via the joint distribution of the prediction errors for the control and dependent sensor variables: As a result, the obvious prediction errors made when a control input changes are already captured and do not contribute towards declaring an anomaly.

The rest of the paper is organised as follows: Section 2 describes our approach. In Section 3, we present temporal anomaly detection results on four real-world datasets using our stacked LSTM approach (LSTM-AD) as well as stacked RNN approach using recurrent sigmoid units (RNN-AD). Section 4 offers concluding remarks.

2 LSTM-AD: LSTM-based Anomaly Detection

Consider a time series $X = \{\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, ..., \mathbf{x}^{(n)}\}$, where each point $\mathbf{x}^{(t)} \in R^m$ in the time series is an m-dimensional vector $\{x_1^{(t)}, x_2^{(t)}, ..., x_m^{(t)}\}$, whose elements

ESANN 2015 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 22-24 April 2015, i6doc.com publ., ISBN 978-287587014-8.
Available from http://www.i6doc.com/en/.

correspond to the input variables. A prediction model learns to predict the next l values for d of the input variables s.t. $1 \leq d \leq m$. The normal sequence(s) are divided into four sets: normal train (s_N), normal validation-1 (v_{N1}), normal validation-2 (v_{N2}), and normal test (t_N). The anomalous sequence(s) are divided into two sets: anomalous validation (v_A), and anomalous test (t_A). We first learn a prediction model using stacked LSTM networks, and then compute the prediction *error distribution* using which we detect anomalies:

Stacked LSTM based prediction model: We consider the following LSTM network architecture: We take one unit in the input layer for each of the m dimensions, $d \times l$ units in the output layer s.t. there is one unit for each of the l future predictions for each of the d dimension. The LSTM units in a hidden layer are fully connected through recurrent connections. We stack LSTM layers s.t. each unit in a lower LSTM hidden layer is fully connected to each unit in the LSTM hidden layer above it through feedforward connections (refer Fig. 1(b)). The prediction model is learned using the sequence(s) in s_N. The set v_{N1} is used for early stopping while learning the network weights.

Anomaly detection using the prediction error distribution: With a prediction length of l, each of the selected d dimensions of $\mathbf{x}^{(t)} \in X$ for $l < t \leq n - l$ is predicted l times. We compute an *error vector* $\mathbf{e}^{(t)}$ for point $\mathbf{x}^{(t)}$ as $\mathbf{e}^{(t)} = [e_{11}^{(t)}, ..., e_{1l}^{(t)}, ..., e_{d1}^{(t)}, ..., e_{dl}^{(t)}]$, where $e_{ij}^{(t)}$ is the difference between $x_i^{(t)}$ and its value as predicted at time $t - j$.

The prediction model trained on s_N is used to compute the error vectors for each point in the validation and test sequences. The error vectors are modelled to fit a multivariate Gaussian distribution $\mathcal{N} = \mathcal{N}(\mu, \mathbf{\Sigma})$. The likelihood $p^{(t)}$ of observing an error vector $\mathbf{e}^{(t)}$ is given by the value of $\mathcal{N}$ at $\mathbf{e}^{(t)}$ (similar to normalized innovations squared (NIS) used for novelty detection using Kalman filter based dynamic prediction model [5]). The error vectors for the points from v_{N1} are used to estimate the parameters μ and $\mathbf{\Sigma}$ using Maximum Likelihood Estimation. An observation $\mathbf{x}^{(t)}$ is classified as 'anomalous' if $p^{(t)} < \tau$, else the observation is classified as 'normal'. The sets v_{N2} and v_A are used to learn τ by maximizing F_β-score (where anomalous points belong to positive class and normal points belong to negative class).

3 Experiments

We present the results of LSTM-AD on four real-world datasets which have different levels of difficulty as far as detecting anomalies in them is concerned. We report precision, recall, $F_{0.1}$-score, and architecture used for LSTM-AD and RNN-AD approaches in Table 1, after choosing the network architecture[1] and τ with maximum $F_{0.1}$-score using validation sets as described in Section 2.

ESANN 2015 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 22-24 April 2015, i6doc.com publ., ISBN 978-287587014-8.
Available from http://www.i6doc.com/en/.

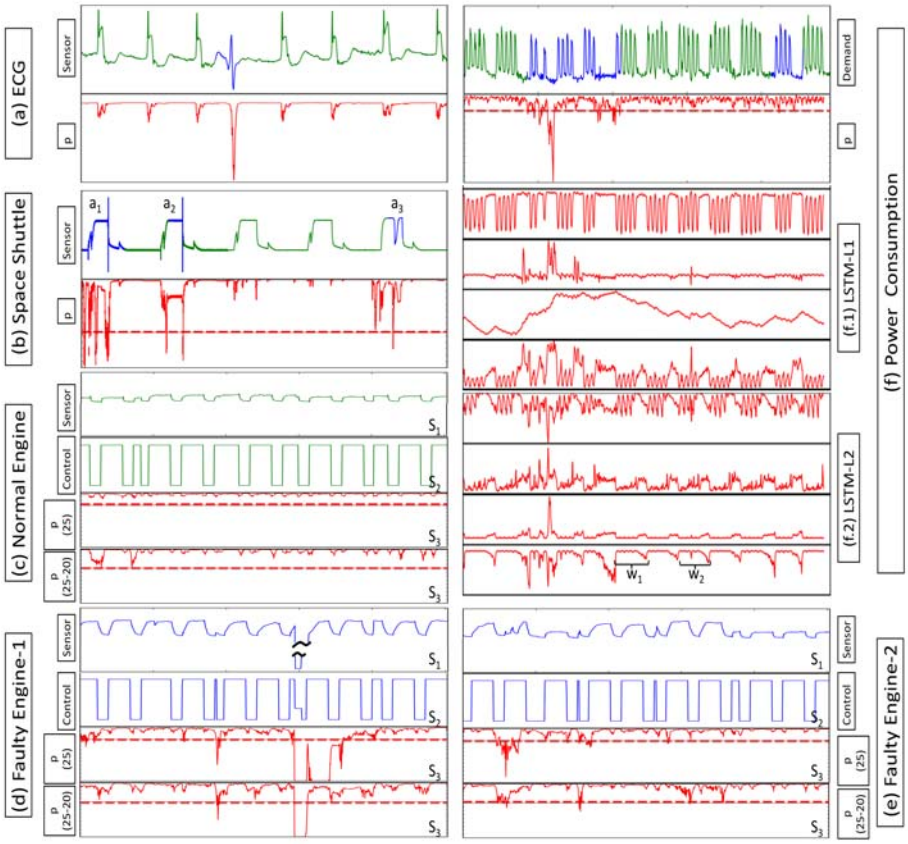

Fig. 2: Sample sequences (normal: green, anomalous: blue) and corresponding likelihoods p (red). Plots with same S_i (i=1,2,3) have same y-axis scale.

Dataset	Architecture		Precision		Recall		$F_{0.1}$-score	
	LSTM	RNN	LSTM	RNN	LSTM	RNN	LSTM	RNN
Space Shuttle	(35-35)	(30-30)	0.93	0.89	0.10	0.03	0.84	0.71
Power	(30-20)	(60-60)	0.94	0.71	0.17	0.19	0.90	0.69
Engine	(25)	(50-40-30)	0.94	0.98	0.12	0.10	0.89	0.90

Table 1: Precision, Recall and $F_{0.1}$-Scores for RNN and LSTM Architectures {Note: (30-20) indicates 30 and 20 units in the first and second hidden layers, respectively.}

ESANN 2015 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 22-24 April 2015, i6doc.com publ., ISBN 978-287587014-8.
Available from http://www.i6doc.com/en/.

3.1 Datasets

Electrocardiograms (ECGs)[2]: The qtdb/sel102 ECG dataset containing a single short term anomaly corresponding to a pre-ventricular contraction (Fig.2(a)). Since the ECG dataset has only one anomaly, we do not calculate a threshold and corresponding $F_{0.1}$-score for this dataset; we only learn the prediction model using a normal ECG subsequence and compute the likelihood of the error vectors for the remaining sequence.
Space Shuttle Marotta valve time series[2]: This dataset has both short time-period patterns and long time-period patterns lasting 100s of time-steps. There are three anomalous regions in the dataset marked a_1, a_2, and a_3 in Fig. 2(b). Region a_3 is a more easily discernible anomaly, whereas regions a_1 and a_2 correspond to more subtle anomalies that are not easily discernable at this resolution.
Power demand dataset[2]: The normal behaviour corresponds to weeks where the power consumption has five peaks corresponding to the five weekdays and two troughs corresponding to the weekend. This dataset has a very long term pattern spanning hundreds of time steps. Additionally, the data is noisy because the peaks do not occur exactly at the same time of the day.
Multi-sensor engine data[2]: This dataset has readings from 12 different sensors: One of the sensors is the 'control' to the engine, and the rest measure dependent variables like temperature, torque, etc. We train the anomaly detector using sequences corresponding to three independent faults and measure the F_β-score on a distinct set of three independent faults. We choose the 'control' sensor together with one of the dependent variables as the dimensions to be predicted.

3.2 Results and Observations

The key observations from our experimental results are as follows:
(i) In Fig. 2, the likelihood values $p^{(t)}$ are significantly lower in the anomalous regions than the normal regions for all datasets. Also, the $p^{(t)}$ values do not remain low throughout the anomalous regions. We deliberately use $\beta << 1$ (0.1) so as to give a higher importance to precision over recall: Note that all points in an anomalous subsequence have a label of 'anomalous', but in practice, there will be many points of 'normal' behaviour even amongst these. So it suffices if a significant percentage of the points in an 'anomalous' subsequence are predicted as anomalous. The values of τ obtained (red dashed lines in the $p^{(t)}$ plots in Fig.2 (a)-(f)) suggest F_β-score (reported in Table 1) to be a suitable metric for the datasets considered.
(ii) The positive likelihood ratio (true positive rate to false positive rate) was found to be high (more than 34.0) for all the datasets. High positive likelihood ratio value suggests the probability of reporting an anomaly in anomalous region is much higher than the probability of reporting an anomaly in normal region.

[1]The networks are trained via resilient backpropagation.
[2]The first three datasets may be downloaded from `http://www.cs.ucr.edu/~eamonn/discords`; the fourth is from a real-life industry project and so not publicly available.

ESANN 2015 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 22-24 April 2015, i6doc.com publ., ISBN 978-287587014-8.
Available from http://www.i6doc.com/en/.

(iii) The activations of selected hidden units, four each from layers LSTM-L1 (lower hidden layer with 30 units) and LSTM-L2 (higher hidden layer with 20 units) for the power dataset are shown in Fig.2 (f.1) and (f.2). Subsequences marked w_1 and w_2 in the last activation sequence shown in Fig.2 (f.2) indicate that this hidden unit activation is high during the weekdays and low during weekends. These are instances of *high-level features* being learned by the higher hidden layer, which appear to be operating at a weekly time-scale.
(iv) As shown in Table 1, for the 'ECG' and 'engine' datasets, which do not have any long-term temporal dependence, both LSTM-AD and RNN-AD perform equally well. On the other hand, for 'space shuttle' and 'power demand' datasets which have long-term temporal dependencies along with short-term dependencies, LSTM-AD shows significant improvement of 18% and 30% respectively over RNN-AD in terms of $F_{0.1}$-score.
(v) The fraction of anomalous points detected for periods prior to faults for the 'engine' dataset was higher than that during normal operation. This suggests that our approach could potentially be useful for early fault prediction well.

4 Discussion

We have demonstrated that (i) stacked LSTM networks are able to learn higher-level temporal patterns without prior knowledge of the pattern duration and so (ii) stacked LSTM networks may be a viable technique to model normal time series behaviour, which can then be used to detect anomalies. Our LSTM-AD approach yields promising results on four real-world datasets which involve modelling small-term as well as long-term temporal dependencies. LSTM-AD gave better or similar results when compared with RNN-AD suggesting that LSTM based prediction models may be more robust compared to RNN based models, especially when we do not know beforehand whether the normal behaviour involves long-term dependencies or not.

References

[1] M. Basseville and I. V. Nikiforov. *Detection of abrupt changes: theory and application*. Prentice Hall, 1993.

[2] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[3] M. Hermans and B. Schrauwen. Training and analysing deep recurrent neural networks. *Advances in Neural Information Processing Systems 26*, pages 190–198, 2013.

[4] D. George. How the brain might work: A hierarchical and temporal model for learning and recognition. *PhD Thesis, Stanford University*, 2008.

[5] P. Hayton et al. Static and dynamic novelty detection methods for jet engine health monitoring. *Philosophical Transactions of the Royal Society of London*, 365(1851):493–514, 2007.

[6] J. Ma and S. Perkins. Online novelty detection on temporal sequences. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 613–618. ACM, 2003.