# Review “DeepLog”

Name: Wenhui Zhang

*Paper Name：*

Du, M., Li, F., Zheng, G., & Srikumar, V. (2017, October). DeepLog: Anomaly Detection and Diagnosis from System Logs through Deep Learning. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security* (pp. 1285-1298). ACM.

**Contribution:**

In this paper proposes a long short-term memory (LSTM) based recurrent neural network (RNN) method for N-gram system log analysis. The goal of this project is to find abnormals in system logs, and help with debugging system failures and conduct root cause analysis. In this paper, it proposed an idea that system logs could be viewed as natural languages, and get processed using methods in natural language processing (NLP).

**Motivation:**

Finding the root cause of performance bugs and failures through abnormal detection on system logs.

**Related works:**

Message count vector for Offline batched processing: Xu’SOSP09, Lou’ATC10, etc.
Build workflow model for Only for simple execution path anomalies: Lou’KDD10, Beschastnikh’ICSE14, Yu’ASPLOS16, etc.

**Methodology:**

This paper is using LSTM based RNN for N-gram system log analysis, there are two main approaches mentioned in this paper.
Method 1: Using Log Key Anomaly Detection model --- LSTM prediction probabilities;
Method 2: A density-based clustering approach, co-occurrence matrix of log keys ($ki$ , $kj$ ) within distance d.
Through using multivariate time series data anomaly detection problem:

(1) It leverages LSTM-based approach;
(2) A parameter value vector is given as input at each time step;
(3) An anomaly is detected if the mean-square-error (MSE) between prediction and actual data is too big.

**Results:**

In this paper, the author conduct the following evaluations:

(1) Evaluation results on HDFS log data;
(2) Evaluation results on OpenStack cloud log with different confidence intervals (CIs);
(3) Evaluation on Blue Gene/L log, with and without online model update.

| suspicious activity | detected? |
|---|---|
| Day 1: Denial of Service attack | Yes, log key anomaly in IDS log |
| Day 1: port scan | Yes, log key anomaly in IDS log |
| Day 2: port scan 1 | Yes, log key anomaly in IDS log |
| Day 2: port scan 2 | Yes, log key anomaly in IDS log |
| Day 2: socially engineered attack | Yes, log key anomaly in firewall log |
| Day 3: undocumented IP address | No |

**Take away:**

DeepLog proposed a realtime system log anomaly detection framework, the result seems to be good. However, K-Means is good for streaming data, N-gram is the classical way of handling system call data, I was just wondering why this paper uses LSTM, not k-means and N-gram.