Self-Attentive Classification-Based Anomaly Detection in Unstructured Logs

2022.09.27

김 미 주 / 이 연 재

Experience Report: Deep Learning-based System Log Analysis for Anomaly Detection
https://github.com/logpai/deep-loglizer

Models

Model	Paper reference
Unsupervised models	
LSTM	[CCS'17] Deeplog: Anomaly detection and diagnosis from system logs through deep learning, by Min Du, Feifei Li, Guineng Zheng, and Vivek Srikumar. [University of Utah]
LSTM	[IJCAI'19] LogAnomaly: unsupervised detection of sequential and quantitative anomalies in unstructured logs by Weibin Meng, Ying Liu, Yichen Zhu et al. [Tsinghua University]
Transformer	[ICDM'20] Self-attentive classification-based anomaly detection in unstructured logs, by Sasho Nedelkoski, Jasmin Bogatinovski, Alexander Acker, Jorge Cardoso, and Odej Kao. [TU Berlin]
Autoencoder	[ICT Express'20] Unsupervised log message anomaly detection, by Amir Farzad and T Aaron Gulliver. [University of Victoria]
Supervised models	
Attentional BiLSTM	[ESEC/FSE'19] Robust log-based anomaly detection on unstable log data by Xu Zhang, Yong Xu, Qingwei Lin et al. [MSRA]
CNN	[DASC'18] Detecting anomaly in big data system logs using convolutional neural network by Siyang Lu, Xiang Wei, Yandong Li, and Liqiang Wang. [University of Central Florida]

<Abstract>

log representations을 학습할 수 있는 classification-based method 기반의 Logsy 제안
중요 시스템(system of interest)의 normal data와 인터넷을 통해 쉽게 액세스할 수 있는
보조 로그 데이터 세트의(auxiliary data) anomaly samples를 구별하는 방식으로

실험 방식

- attention-based encoder model with a new hyperspherical loss function 제안
  → 이를 통해 정상 로그와 비정상 로그 간의 본질적인 차이점을 캡처하는 간단한 log representations를 학습

- Logsy의 속성을 조사하기 위해 추가 실험을 수행
  auxiliary data size의 영향, the influence of expert knowledge, the quality of the learned log representations

I. INTRODUCTION

- Anomaly detection

→ a **data mining task** of finding observations in a corpus of data that differ from the expected behaviour.

- Log

→ 컴퓨터 시스템에서 anomaly detection을 위한 중요한 데이터 소스

ex. log message = print("total of %i errors detected", 5) → 상수 문자열 템플릿 + 변수 값

I. INTRODUCTION

A common approach for log anomaly detection

- history of template, H = $t_0, \ldots, t_m$. 시퀀스를 통해 t_{m+1}(the next index of the log template)을 예측
  1. log parsing을 통해 log templates으로 변환 후 이를 모델 훈련 시 사용
  2. (LSTM) 활용
  3 정상 데이터에서 일반적인 패턴을 추출하는 one-class classification 사용
  4. 비지도 학습

→ learning sequence of indices는 새로 나타나는 로그 메시지를 올바르게 분류하지 못함

- pre-trained word embeddings 사용
  to numerically represent the log templates instead of the integer log sequences

→ word vectors are pre-trained 도메인(예: Wikipedia)은 컴퓨터 시스템 개발에 사용되는 언어와 본질적으로 다름

하지만, 이전에 볼 수 없었던 로그 메시지(unseen log messages)에 대한 일반화 측면에서 여전히 큰 한계가 존재
→ the imperfect log vector representations으로 인해 잘못된 예측 생성

I. INTRODUCTION

→ 대표적이고 compact한 numerical log embeddings를 얻는 문제를 직접 해결하는 anomaly detection method 제시

▪ the normal data in anomaly detection method 의 가정
normal data는 압축되어야 함.
→ normal log messages는 서로 가까운 거리를 가진 벡터 표현을 가져야 함을 의미

normal log messages: 좁은 영역에 분포
anomalies: normal 분포에서 멀리 퍼져 있어야 함.

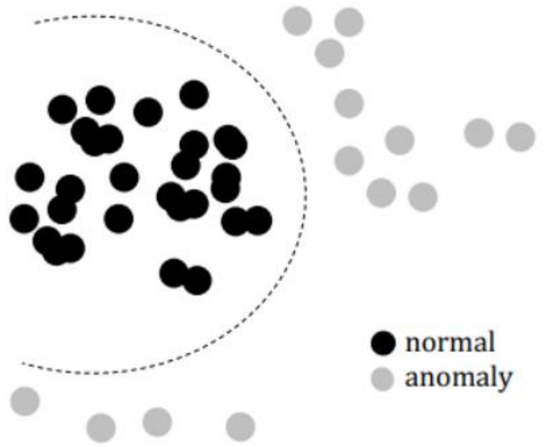

Fig. 3. Ideal distribution of the log vector representations in space.

I. INTRODUCTION

▪ *About Logsy*

1. 학습 방법
중요 시스템(system of interest)의 normal log data와 인터넷을 통해 쉽게 액세스할 수 있는 다른 시스템의 보조 로그 데이터 세트의 log messages를 구별하는 방식으로 neural network를 train하여 log vector representations를 학습

2. a classification approach to anomaly detection에서의 보조 데이터 세트 역할
→ regularizing against overfitting → better generalization in unseen logs.
a target system logs of interest: T
보조 데이터 세트: 오픈 소스 로그 저장소에서 하나 이상의 데이터 세트 채택

3. neural network 아키텍처
→ Transformer encoder with multi-head self-attention mechanism
(learns context information from the log message in the form of log vector representations (embeddings))

I. INTRODUCTION

▪ *About Logsy*

4. hyperspherical learning objective 제안
→ compact log vector representations of the normal log messages 학습하기 위해
normal samples가 hypersphere의 중심 주위에 집중된 (compact) vector representations이 되도록 함.
이를 통해 normal and the anomaly data를 더 잘 분리할 수 있으며, 구의 중심으로부터의 거리가 an anomaly score를 측정하는데 사용

Small distances: normal samples
large distances: anomalies

5. 기존 연구 성능 개선 및 사용자 개입 가능
direct log-to-vector transformation가능
misclassified samples에 개입하여 수정 및 retraining

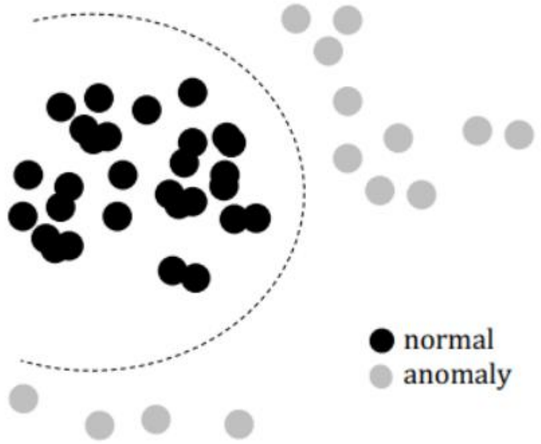

Fig. 3. Ideal distribution of the log vector representations in space.

II. RELATED WORK

▪ 과거: Supervised methods

support vector machine (SVM) to detect failures - both normal and anomalous samples are assumed to be available.

→ However, obtaining system-specific labelled samples is costly and often practically infeasible.

▪ unsupervised learning methods

1. Principal Component Analysis (PCA) - PCA detects an abnormal vector (a session) by measuring the projection length on the residual subspace of a transformed coordinate system.

a. log 파일에 session들이 있고, session은 각 log entry 마다 붙여져서 session id로 식별이 가능하다는 것을 가정
b. session별로 log keys 그룹화 → 각 session 내에서 each log key value의 출현 횟수 계산
→ A session vector is of size n, 해당 session에서 log key K의 출현 횟수를 나타냄,
각 열이 log key, 각 행이 하나의 session vector 형성

2. Invariant Mining (IM) - log event count vectors에서 log events 간의 linear relationships를 마이닝

II. RELATED WORK

▪ unsupervised learning methods - deep learning methods
LSTM 사용 - to predict the anomaly of log sequence based on log keys.
DeepLog also use LSTM to forecast the next log event and then compare it with the current ground truth to detect anomalies.

However, the input to the unsupervised methods is a one-hot vector of logs representing the indices of the log templates
→ Therefore, it cannot cope with newly appearing log events.

▪ Some studies
1. NLP 사용 - to analyze log data based on the idea that log is a natural language sequence.
2. anomalous log messages를 예측하기 위해 LSTM model and TF-IDF(term frequency-inverse document frequency) weight 사용
TF-IDF : 문서 내 단어마다 중요도를 고려하여 가중치를 주는 통계적인 단어 표현방법

II. RELATED WORK

▪ 기존 연구과 다른 점

1. we add **domain bias** on the anomalous distribution to improve detection
domain bias - We provide such bias by employing easily accessible log datasets as an auxiliary data source.

2. We evaluate Logsy against **unsupervised approaches**
→ From the perspective of using labels of the target system it is an unsupervised approach.

III. TOWARDS CLASSIFICATION-BASED LOG ANOMALY DETECTION

PROBLEM DEFINITION

▪ Log Message : d × |r| matrix (d: dim, r : 단어의 수)
→ 로그 메시지를 d차원으로 나타냄

▪ Training Logs - the system of interest

$$\mathcal{D} = \{(\mathbf{x_1}, y_1), \ldots, (\mathbf{x_n}, y_n)\}$$

D = the system of interest의 training logs
$x_i \in R^d$ = **a log message** where it words are represented in *d − dimensional* space

$$y_i = 0; 1 < i \leq n,$$

y_i= 0, 로그 메시지가 정상인 경우
→ assuming that the data in the system of interest is mostly composed of **normal samples**

III. TOWARDS CLASSIFICATION-BASED LOG ANOMALY DETECTION

PROBLEM DEFINITION

- Training Logs - auxiliary data

$$\mathcal{A} = \{(\mathbf{x_n}, y_n), \ldots, (\mathbf{x_{n+m}}, y_{n+m})\}$$

보조 데이터가 추가된 training log A

m: 보조 데이터의 크기

$$y_i = 1; n < i \leq n + m.$$

y_i = 1, 로그 메시지가 정상이 아닌 경우
→보조 데이터를 비정상 데이터처럼 사용

III. TOWARDS CLASSIFICATION-BASED LOG ANOMALY DETECTION

PROBLEM DEFINITION

- Neural Network 의 function

φ = input 인 log message embedding을 R^p의 vector representations으로 mapping하는 function

$$\phi(\mathbf{x_i}, \quad \theta) : \mathbb{R}^d \rightarrow \mathbb{R}^p$$

input = d차원 → output인 p차원 벡터를 구함

l = output 으로부터 anomaly score를 구하는 function

$$l \ : \ \mathbb{R}^p \ \rightarrow \ [0, a], a \ \in \ \mathbb{R}$$

output인 p차원 벡터로부터 0-a 사이의 확률을 구함

III. TOWARDS CLASSIFICATION-BASED LOG ANOMALY DETECTION

PROBLEM DEFINITION

- task

earn the parameters θ from the training data

$$\mathcal{D}_t = \{(\mathbf{x}_1^t), (\mathbf{x}_2^t), \ldots, (\mathbf{x}_j^t), \ldots\}$$

t : test sample

→ 학습 후, 예측 단계(D_t)에서 들어오는 instance 마다 I(φ(xi,θ))]에 의해 얻은 anomaly score를 기반으로 normal 인지, anomaly 인지 예측한다.

IV. SELF-ATTENTIVE ANOMALY DETECTION WITH CLASSIFICATION-BASED OBJECTIVE

▪ formal definitions 제공
→ data preprocessing, the neural network, the log vector representations, how they are utilized in the modified objective function for anomaly detection.

A. Preliminaries

▪ 로그를 시퀀스 L(sequence of temporally ordered unstructured text messages)로 정의

$$L = (x_i \ : \ i = 1, 2, ...)$$

x_i = a logging instruction (e.g. printf(), log.info()) within the software source code에 의해 생성된 log message
i = positional index within the sequence

로그 메시지 구성
1. log template이라고 불리는 constant(상수)
2. varying part라고 불리는 variables
→ log message = print("total of %i errors detected", 5) → [상수 문자열 템플릿 + 변수 값]

IV. SELF-ATTENTIVE ANOMALY DETECTION WITH CLASSIFICATION-BASED OBJECTIVE

A. Preliminaries

▪ sequence of tokens

로그 메시지에서 분리할 수 없는 가장 작은 singleton object는 token

Each log message consists of a finite sequence of tokens(r_i)

$$\mathbf{r_i} = (w_j \ : \ w_j \in \mathbb{V}, j = 1, 2, ..., s_i)$$

r_i = sequence of tokens

V = a set (vocabulary) of all tokens

j = the **positional index of a token** within the log message x_i

s_i = the **total number** of tokens in x_i

s_i 대신 $|r_i|$를 사용

→ tokenization method에 따라 w_j = a word, word piece, or character(문자)가 될 수 있음.

▪ tokenization - transformation function

$$\mathcal{T} : x \rightarrow \mathbf{r}.$$

IV. SELF-ATTENTIVE ANOMALY DETECTION WITH CLASSIFICATION-BASED OBJECTIVE

A. Preliminaries

▪ context and numerical vector representation (embedding vector) 개념 도입

- context

a token(w_j)가 주어지면, its context는 token의 앞 뒤 시퀀스, 즉 a tuple of sequences: $C(w_j)$ 의해 정의

$$C(w_j) = ((w_1, w_2, ..., w_{j-1}), (w_{j+1}, w_{j+2}, ..., w_{|\mathbf{r_i}|})), \text{ where } 0 \leq j \leq |\mathbf{r_i}|$$

- numerical vector representation (embedding vector)

An embedding vector = a token 혹은 a d-dimensional real valued vector representation $s \in R^d$ of a log message

IV. SELF-ATTENTIVE ANOMALY DETECTION WITH CLASSIFICATION-BASED OBJECTIVE

A. Preliminaries

▪ In the learned vector space

→ similar log messages는 closer embedding vectors로 표현되어야 하고, largely different log messages는 더 멀리 떨어져 있어야 함.

ex. the embedding vectors:
- small distance
"Took 10 seconds to create a VM" and "Took 9 seconds to create a VM"는 in d-dimensional space에서 small distance를 가져야 함.
- large distance
"Took 9 seconds to create a VM" and "Failed to create VM 3" should be distant.

IV. SELF-ATTENTIVE ANOMALY DETECTION WITH CLASSIFICATION-BASED OBJECTIVE

A. Preliminaries

▪ data set

the data from the system of interest = **target dataset**, 즉 the system where we want to detect anomalies (이상을 감지하려는 시스템)으로 참조
→ our experiments에서 learning purposes으로 anomaly data from the target system를 사용하지 않음.

auxiliary data = nonrelated systems, which serve only for training the model.

All the results during test time = target dataset에서 추출한 test set에 대해 수행

IV. SELF-ATTENTIVE ANOMALY DETECTION WITH CLASSIFICATION-BASED OBJECTIVE

B. Logsy

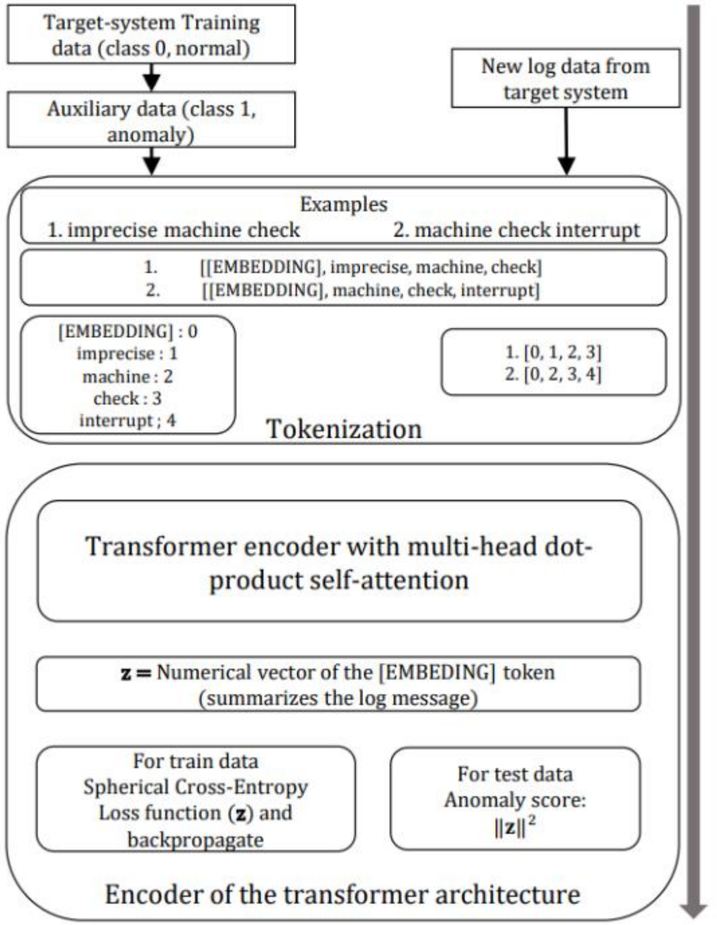

Fig. 1. Overview of the architecture and component details of Logsy.