로그 이상 탐지를 위한 도메인별 사전 훈련 언어 모델 중요성 연구

레리사 아데바 질차[1], 김득훈[2], 곽진[3]
[1]아주대학교 AI융합네트워크학과, 정보보호응용및보증연구실 석박통합과정
[2]아주대학교 소프트웨어융합연구소 박사후연구원
[3]아주대학교 사이버보안학과 교수
jilchalelisa@ajou.ac.kr, kimdh1206@ajou.ac.kr, security@ajou.ac.kr

On the Significance of Domain-Specific Pretrained Language Models for Log Anomaly Detection

Lelisa Adeba Jilcha[1], Deuk-Hun Kim[2], Jin Kwak[3]
[1]ISAA Lab., Dept. of AI Convergence Network, Ajou University
[2]Inst. for Computiong and Informatics Research, Ajou University
[3]Dept. of AI Convergence Network, Ajou University

요 약

Pretrained language models (PLMs) are extensively utilized to enhance the performance of log anomaly detection systems. Their effectiveness lies in their capacity to extract valuable semantic information from logs, thereby strengthening the detection performance. Nonetheless, challenges arise due to discrepancies in the distribution of log messages, hindering the development of robust and generalizable detection systems. This study investigates the structural and distributional variation across various log message datasets, underscoring the crucial role of domain-specific PLMs in overcoming the said challenge and devising robust and generalizable solutions.

1. Introduction

As modern software systems grow increasingly complex and cybersecurity threats persist, the monitoring of system logs for anomalous behavior has evolved into a crucial undertaking. Log messages, structured records detailing events within a system, application, or device, are essential for identifying security breaches, software errors, system faults, and performance issues. Each log message consists of a structured statement composed during software development, containing both fixed and variable parameters[1]. The fixed part defines the event template, while the variable parts convey dynamic runtime information.

Automated log anomaly detection systems, especially those leveraging deep learning technologies, serve as vital tools in today's cybersecurity landscape. They provide a swift and advanced means of identifying and resolving potential threats and anomalies within large-scale networks[2]. This technological fusion not only enhances the precision and agility of anomaly detection but also substantially diminishes the risks associated with system failures and cybersecurity breaches.

A crucial aspect of deep learning-based approaches involves log parsing, converting each log message into a specific static event template with variable parameters[3]. This process is followed by constructing log sequences and transforming them into vector representations for downstream anomaly detection models[3]. However, these approaches often overlook the semantic information embedded in raw log

이 논문은 2024년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임 (No.2021-0-01806, 스마트공장 보안 내재화 및 보안관리 기술 개발).

messages, leading to decreased detection system robustness. Recent studies advocate for leveraging pretrained language models (PLMs) to generate semantically meaningful embedding vectors for downstream detection models, highlighting the importance of capturing inherent semantic information in log data for enhanced performance[4].

Despite advancements, disparities in the underlying distribution of log messages present challenges in developing robust and generalizable log anomaly detection systems using PLMs. The primary challenge stems from their limited understanding of specialized terminologies within the domain of log messages[5]. This study aims to analyze the distributional and structural characteristics of log data across different datasets, emphasizing the significance of domain-specific pretrained language models in overcoming the above-mentioned challenges and devising more resilient and adaptable solutions.

The article is structured as follows: Section II provides background information required to insure adequate understanding of the manuscript. This includes introduction to the common workflow of conventional log anomaly detection, pretraining and finetuning of language models focusing on BERT, and statistical information of the datasets used for the analysis. Section III provides in-depth analysis of the employed datasets and significance of the domain specific language models for efficient log anomaly detection. Finally, the conclusion remark is provided in Section IV.

2. Background

2.1 Workflow of Log Anomaly Detection

Typically, log-based anomaly detection follows a four-step approach: parsing, grouping, representation, and anomaly detection[3]. This section provides insights into the specific techniques utilized at each step within the employed workflow.

2.1.2 Grouping

Grouping entails segregating logs into distinct groups or log sequences, each representing a finite chunk of logs. Features, which are consumed by the downstream detection model, are extracted from these log sequences. Common grouping techniques include fixed-length-based grouping, organizing logs chronologically; windows-based grouping, employing sliding windows of predefined length; and session ID-based grouping, utilizing identifiers such as block IDs to group logs with the same execution path[6].

2.1.3 Representation and Detection

In log-based anomaly detection models, logs are often transformed into sequential, quantitative, and semantic vectors[6]. Sequential vectors capture the order of log events within a window, while quantitative vectors reflect the frequency of each event in a log window. Semantic vectors, on the other hand, convey the semantic meaning of log events. The conversion of log messages into semantic vectors is a prevalent technique in contemporary anomaly detection approaches. PLMs such as Bidirectional Encoder Representations from Transformers (BERT) and Generative Pre-trained Transformers (GPT) are utilized to generate efficient semantic representation vectors[4]. Additionally, various deep learning models such as CNNs, RNNs, and attention-based models are employed for downstream log anomaly detection tasks[6].

2.2 Pretraining and Fine-tuning BERT

BERT is a transformer encoder-based language model which is pretrained on a massive corpus of text data using the masked language modeling objective[7]. BERT employs multi-head attention to allow the model to focus on different aspects of the input. Multiple sets of Query (Q), Key (K), and Value (V) matrices with dimension n x dm are used and the outputs of these multiple attention heads are concatenated and linearly transformed.

Given a sequence of input tokens $X = x_1, x_2, x_3, ..., x_n$ where n is the sequence length, the self-attention mechanism in BERT computes the attention scores between all pairs of tokens to capture the embedded contextual information in the input text. During pretraining, BERT learns to predict masked words in a sentence by considering the context of surrounding words bidirectionally. The objective is to maximize the

probability of predicting the masked token x_i, given the surrounding context $x_{<i}$ and $x_{>i}$. This process enables BERT to develop a deep understanding of the contextual relationships between words, making it a powerful feature extractor.

After pretraining, BERT can be fine-tuned on specific tasks by adding task-specific layers. The fine-tuning process adapts BERT to the target task by exposing the pretrained model to a few labeled datasets from the downstream task. The advantage of using pretrained BERT models lies in their ability to capture rich semantic information from text data, even with minimal task-specific data.

2.3 Analyzed Datasets

We used two publicly available log message datasets, namely BGL and Thunderbird, to analyze the distributional and characteristics discrepancies between log messages. Table 1 provide statistical information of the analyzed datasets.

The BGL dataset originates from a supercomputing system and comprises 4,747,963 log messages[1]. Each message within this dataset has been manually classified as either normal or anomalous, with 348,460 anomalous samples. Similarly, the Thunderbird dataset, sourced from the Thunderbird supercomputer system, includes a total of 44,841,030 entries, consisting of 41,592,791 alerts and 3,248,239 non-alerts[1].

<Table 1> Statistical Information of the Analyzed Datasets

Datasets	Log events	Training data		Testing data	
		Total	Alert	Total	Alert
BGL	1,847	55,401	25,066	13,851	6,309
Thunderbird	2,880	400,000	193,840	100,000	3,460

3. Significance of Domain specific PLMs in Log Anomaly Detection

Log messages exhibit a unique pattern characterized by their sparse nature and the tendency towards comparatively shorter sentence lengths. Table II illustrates the distribution of the top ten words in two publicly available datasets, BGL and Thunderbird, after a text cleaning process. A text cleaning is process of sanitizing the raw log message through a series of operations.

After the text cleaning process, the BGL dataset showcases 769 unique words for normal samples and 209 for anomaly samples, while the Thunderbird dataset reveals 3002 unique words for normal samples and 65 for anomaly samples. Despite these variations, the top ten most common words in normal samples account for 50.6% and 39.65% of the total occurrences in their respective classes for each dataset. Similarly, the top ten most frequent words in anomaly samples constitute 64.54% and 60.32% of the overall anomalous samples for the BGL and Thunderbird datasets, respectively. However, these datasets exhibit minimal overlap in terms of the most frequent words, with only “kernel” appearing in the normal samples of both.

The statistical insights underscore the potential hurdles in pretraining language models such as BERT on system log datasets. These challenges fall into two main categories. Firstly, the constrained context due to a limited vocabulary size hinders the model’s ability to grasp complex contextual information, possibly resulting in suboptimal representations. Secondly, the skewed coverage of vocabulary, dominated by a handful of frequent words, might restrict exposure to less common vocabulary during pretraining, thus impeding the model’s generalization to new words.

To address this issue, either fine-tuning language models on a large corpus of log datasets from diverse sources or utilizing pretrained models specifically adapted to relevant target domains, such as cybersecurity is necessary. Such approaches can enhance the model’s ability to grasp the nuances of language and terminology present in system logs, thereby effectively managing both environmental and data drift. Environmental drift refers to changes in the system or network environment over time, which can impact the efficacy of the detection algorithms. Similarly, data drift refers to changes in the distribution of log data over time, which can undermine the performance of detection models trained on static datasets.

Considering these challenges, it becomes imperative to devise strategies that ensure the reliability and adaptability of log anomaly detection systems. This

<Table 2> Distribution of the Top-ten Words in BGL and Thunderbird Datasets

Rank	BGL normal samples		BGL anomalous samples		Thunderbird normal samples		Thunderbird anomalous samples	
	Vocabulary	Proportion	Vocabulary	Proportion	Vocabulary	Proportion	Vocabulary	Proportion
1	ras	0.1257	ras	0.1122	may	0.1374	opendemux	0.0990
2	kernel	0.1189	fatal	0.1121	kernel	0.0517	may	0.0961
3	info	0.1093	kernel	0.0990	user	0.0383	error	0.0593
4	generating	0.0499	error	0.0700	mosal	0.0330	in	0.0518
5	iar	0.0186	interrupt	0.0697	va	0.0234	pbsmom	0.0495
6	dear	0.0182	data	0.0696	protctx	0.0229	connection	0.0495
7	alignment	0.0171	tlb	0.0492	from	0.0228	refused	0.0495
8	exceptions	0.0171	to	0.0215	cannot	0.0224	cannot	0.0495
9	microsecon ds	0.0158	on	0.0213	mosalvirttophy sex	0.0223	connect	0.0495
10	error	0.0154	message	0.0208	retrieve	0.0223	to	0.0495

entails not only enhancing the robustness of detection algorithms but also implementing mechanisms to monitor and mitigate the effects of environmental and data drift. By doing so, we can ensure the continued effectiveness of log anomaly detection systems in safeguarding against security breaches and performance issues in modern software environments.

4. Conclusion

Utilizing PLMs to compute efficient representation vectors presents a promising solution to the shortcomings associated with traditional deep learning-based approaches in log anomaly detection. However, despite the potential benefits, significant challenges persist within this domain. These challenges are primarily attributed to the disparities observed in the distribution of log messages and the inherent complexity of domain-specific terminologies. The diverse distributional and structural characteristics of log data across different datasets, combined with the challenges of acquiring sufficient training data, emphasize the necessity for developing robust detection systems capable of reliably managing both environmental and data drift, thus ensuring the generalizability of log anomaly detection systems.

Reference

[1] S. He, J. Zhu, P. He and M. R. Lyu, "Loghub: A large collection of system log datasets towards automated log analytics," arXiv:2008.06448, 2020.

[2] J. Lou, Q. Fu, S. Yang, Y Xu and J. Li, "Mining invariants from console logs for system problem detection," ATC'10: Proc. of the USENIX Annual Technical Conference, Boston, USA, Jun. 2010.

[3] W. Meng et al., "LogAnomaly: Unsupervised detection of sequential and quantitative anomalies in unstructured logs," Proc. 28th Int. Joint Conf. Artif. Intell. (IJCAI), Vienna, Austria, Aug. 2019, pp. 4739-4745.

[4] X. Zhang et al., "Robust log-based anomaly detection on unstable log data," Proc. 27th ACM Joint Meeting Eur. Softw. Eng. Conf. Symp. Foundations Softw. Eng., Tallinn, Estonia, Aug. 2019, pp. 807-817.

[5] S. Chen and H. Liao, "BERT-log: Anomaly detection for system logs based on pre-trained language model," Appl. Artif. Intell., vol. 36, no. 1, pp. e2145642-1-e2145642-23, Dec. 2022.

[6] M. Du, F. Li, G. Zheng and V. Srikumar, "Deeplog: Anomaly detection and diagnosis from system logs through deep learning", ACM SIGSAC conference on computer and communications security, Dallas, USA, Oct. 2017, pp. 1285-1298.

[7] J. Devlin, M.-W. Chang, K. Lee and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Minneapolis, Jun. 2019, USA, pp. 4171-4186.