HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG



Nguyễn Đức Tiệp

PHÂN TÍCH LOG HỆ THỐNG ĐỂ PHÁT HIỆN LỖI TIỀM ẨN BẰNG AI

ĐỀ ÁN TỐT NGHIỆP THẠC SĨ KỸ THUẬT

(Theo định hướng ứng dụng)

HÀ NỘI – NĂM 2025

HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG



Nguyễn Đức Tiệp

PHÂN TÍCH LOG HỆ THỐNG ĐỂ PHÁT HIỆN LỖI TIỀM ẨN BẰNG AI

CHUYÊN NGÀNH: HỆ THỐNG THÔNG TIN
MÃ SỐ: 8.48.01.04

ĐỀ ÁN TỐT NGHIỆP THẠC SĨ KỸ THUẬT
(Theo định hướng ứng dụng)

NGƯỜI HƯỚNG DẪN KHOA HỌC:
PGS. TSKH. Hoàng Đăng Hải

HÀ NỘI – NĂM 2025

MỤC LỤC

LỜI CAM ĐOAN 1
LỜI CẢM ƠN 2
MỤC LỤC 3
DANH SÁCH CÁC TỪ VIẾT TẮT 6
DANH SÁCH HÌNH VẼ 8
DANH SÁCH BẢNG 9
MỞ ĐẦU 1
1. Lý do chọn đề tài 1
2. Tổng quan về vấn đề nghiên cứu 1
3. Mục đích nghiên cứu 2
4. Đối tượng và phạm vi nghiên cứu 2
5. Phương pháp nghiên cứu 3
6. Cấu trúc đề án tốt nghiệp 3
CHƯƠNG 1: TỔNG QUAN VÀ CƠ SỞ LÝ THUYẾT 5
1.1 Giới thiệu bài toán phân tích log phát hiện bất thường 5
1.1.1 Bối cảnh 5
1.1.2 Bài toán phân tích log phát hiện bất thường 5
1.1.3 Khoảng trống nghiên cứu 8
1.2 Đặc điểm và phân loại log 9
1.2.1 Đặc điểm dữ liệu log 9
1.2.2 Cấu trúc log điển hình 10
1.2.3 Phân loại lỗi hệ thống 11
1.3 Các phương pháp phát hiện bất thường 12
1.3.1 Giới thiệu 12
1.3.2 Phương pháp dựa trên quy tắc (truyền thống) 12
1.3.3 Phương pháp dựa trên trí tuệ nhân tạo (học máy & học sâu) 13
1.3.4 So sánh phương pháp dựa trên quy tắc và trí tuệ nhân tạo 15
1.4 Pipeline trí tuệ nhân tạo cho phát hiện bất thường từ log 15
1.4.1 Giới Thiệu 15
1.4.2 Pipeline Thực Hiện 16
1.5. Kết luận chương 1 17
CHƯƠNG 2: XÂY DỰNG MÔ HÌNH 19
2.1 Kiến trúc pipeline dữ liệu 19

2.1.1 Tổng quan pipeline 19
2.1.2 Phân tích cú pháp log 20
2.1.3 Phân đoạn từ 21
2.1.4 Biểu diễn vector 22
2.2 Kiến trúc tổng thể hệ thống 23
2.3 Chọn mô hình Autoencoder và LogBERT 24
2.3.1 Tổng quan lựa chọn 24
2.3.2 Autoencoder - Mô hình cơ sở 24
2.3.3 LogBERT - Mô hình chính 26
2.4 Quy trình huấn luyện và tinh chỉnh 27
2.4.1 Thiết lập huấn luyện 27
2.4.2 Huấn luyện Autoencoder 27
2.4.3 Huấn luyện LogBERT 29
2.4.4 Tinh chỉnh siêu tham số 31
2.4.5 Bảng tham số huấn luyện 34
2.5 Chỉ số đánh giá 35
2.6. Kết luận chương 2 38
CHƯƠNG 3: ĐÁNH GIÁ VÀ THỰC NGHIỆM 40
3.1 Mô hình dữ liệu thử nghiệm 40
3.2 Đánh giá kết quả mô hình 41
3.2.1 Kết quả đánh giá - Tập dữ liệu HDFS 41
3.2.2 Kết quả đánh giá - Tập dữ liệu BGL 46
3.2.3 Bảng so sánh kết quả mô hình 48
3.2.4 Phân tích kết quả 50
3.2.5 Nhận xét và đánh giá 53
3.3 Thiết kế và triển khai pipeline thời gian thực 54
3.3.1 Các thành phần và luồng dữ liệu hệ thống 54
3.3.2 Cài đặt và triển khai các thành phần 57
3.4 Đánh giá hiệu năng của pipeline và thảo luận 59
3.4.1 Tổng quan đánh giá 59
3.4.2 Thông lượng và độ trễ 59
3.3.3 Sử dụng tài nguyên 63
3.4.4 Kiểm thử khả năng mở rộng 64
3.4.5 Thách thức khi triển khai thực tế 64
3.4.6 Đề xuất phương án triển khai thực tiễn 66
3.4.7 Hướng phát triển 67
3.5. Kết luận chương 3 68

KẾT LUẬN 69
1. Kết quả nghiên cứu của đề án 69
2. Hạn chế của đề án 70
3. Hướng phát triển của đề án 70
DANH MỤC CÁC TÀI LIỆU THAM KHẢO 70

DANH SÁCH CÁC TỪ VIẾT TẮT

Viết tắt	Tiếng Anh	Tiếng Việt
AI	Artificial Intelligence	Trí tuệ nhân tạo
ML	Machine Learning	Học máy
DL	Deep Learning	Học sâu
NLP	Natural Language Processing	Xử lý ngôn ngữ tự nhiên
BERT	Bidirectional Encoder Representations from Transformers	Biểu diễn mã hóa hai chiều từ Transformers
LogBERT	Log Anomaly Detection via BERT	Phát hiện bất thường log sử dụng BERT
LSTM	Long Short-Term Memory	Bộ nhớ ngắn hạn dài hạn
RNN	Recurrent Neural Network	Mạng nơ-ron hồi quy
CNN	Convolutional Neural Network	Mạng nơ-ron tích chập
VAE	Variational Autoencoder	Autoencoder biến phân
GNN	Graph Neural Network	Mạng nơ-ron đồ thị
GCN	Graph Convolutional Network	Mạng tích chập đồ thị
GAT	Graph Attention Network	Mạng chú ý đồ thị
SVM	Support Vector Machine	Máy vector hỗ trợ
PCA	Principal Component Analysis	Phân tích thành phần chính

DBSCAN	Density-Based Spatial Clustering of Applications with Noise	Phân cụm không gian dựa trên mật độ
TF-IDF	Term Frequency-Inverse Document Frequency	Tần suất từ - Tần suất nghịch đảo tài liệu
BPE	Byte Pair Encoding	Mã hóa cặp byte
MLM	Masked Language Modeling	Mô hình ngôn ngữ che mặt nạ
ReLU	Rectified Linear Unit	Đơn vị tuyến tính chỉnh lưu
MSE	Mean Squared Error	Lỗi bình phương trung bình
ROC	Receiver Operating Characteristic	Đặc tính hoạt động của bộ thu
AUC	Area Under Curve	Diện tích dưới đường cong
PR	Precision-Recall	Độ chính xác - Độ nhạy
TP	True Positive	Dương tính thật
FP	False Positive	Dương tính giả
TN	True Negative	Âm tính thật
FN	False Negative	Âm tính giả
F1-Score	F1 Score	Điểm F1
HDFS	Hadoop Distributed File System	Hệ thống tệp phân tán Hadoop
BGL	Blue Gene/L	Blue Gene/L (siêu máy tính)
GPU	Graphics Processing Unit	Đơn vị xử lý đồ họa
CPU	Central Processing Unit	Đơn vị xử lý trung tâm

RAM	Random Access Memory	Bộ nhớ truy cập ngẫu nhiên
VRAM	Video Random Access Memory	Bộ nhớ video truy cập ngẫu nhiên
I/O	Input/Output	Đầu vào/Đầu ra
API	Application Programming Interface	Giao diện lập trình ứng dụng
JSON	JavaScript Object Notation	Ký hiệu đối tượng JavaScript

DANH SÁCH HÌNH VẼ

Hình 1.1: Flowchart minh họa pipeline xử lý log 6 bước
Hình 2.1: Minh họa pipeline xử lý dữ liệu log từ log entries thô đến biểu diễn vector
Hình 2.2: Kiến trúc tổng thể của pipeline thời gian thực phát hiện bất thường trong log
Hình 2.3: Đường cong mất mát huấn luyện và kiểm định cho Autoencoder trên tập dữ liệu HDFS
Hình 2.4: Đường cong mất mát huấn luyện và kiểm định cho LogBERT trên tập dữ liệu HDFS
Hình 2.5: Kết quả tinh chỉnh tốc độ học cho Autoencoder và LogBERT
Hình 2.6: Kết quả tinh chỉnh kích thước lô cho Autoencoder và LogBERT
Hình 3.1: Đường cong ROC cho Autoencoder và LogBERT trên dataset HDFS
Hình 3.2: Ma trận nhầm lẫn (Confusion Matrix) cho Autoencoder trên dataset HDFS
Hình 3.3: Ma trận nhầm lẫn (Confusion Matrix) cho LogBERT trên dataset HDFS
Hình 3.4: Đường cong ROC cho Autoencoder và LogBERT trên dataset BGL
Hình 3.5: Thông lượng của pipeline theo tải
Hình 3.6: Phân bố độ trễ của pipeline

DANH SÁCH BẢNG

Bảng 2.1: Tham số huấn luyện - Tập dữ liệu HDFS

Bảng 2.2: Tham số huấn luyện - Tập dữ liệu BGL

Bảng 3.1: Thống kê tập dữ liệu sau tiền xử lý

Bảng 3.2: So sánh kết quả mô hình trên dataset HDFS

Bảng 3.3: So sánh kết quả mô hình trên dataset BGL

Bảng 3.4: Tổng hợp kết quả trên dataset HDFS

Bảng 3.5: Tổng hợp kết quả trên dataset BGL

Bảng 3.6: Thông lượng theo tải

Bảng 3.7: Sử dụng tài nguyên - Tải 1000 log/giây

Bảng 3.8: Kết quả khả năng mở rộng

MỞ ĐẦU

1. Lý do chọn đề tài

Trong kỷ nguyên số hiện nay, các hệ thống công nghệ thông tin (CNTT) ngày càng trở nên phức tạp và phân tán. Mỗi thành phần trong hệ thống – từ máy chủ, ứng dụng, middleware đến thiết bị mạng – đều liên tục sinh ra các bản ghi hoạt động (log) với dung lượng khổng lồ. Theo nghiên cứu của Xu và đồng nghiệp [1], các trung tâm dữ liệu hiện đại có thể tạo ra hàng trăm GB log mỗi ngày, khiến việc giám sát thủ công gần như không khả thi. Log không chỉ đơn thuần là dữ liệu ghi nhận quá trình vận hành, mà còn là nguồn thông tin quan trọng giúp chẩn đoán sự cố, giám sát hiệu suất và đảm bảo an toàn hệ thống. Tuy nhiên, việc phân tích log theo cách truyền thống – dựa trên quy tắc cố định hoặc thao tác thủ công – ngày càng trở nên bất cập trước sự gia tăng về quy mô và tính đa dạng của dữ liệu log. Những lỗi tiềm ẩn có thể không để lại dấu hiệu rõ ràng, dẫn đến nguy cơ bỏ sót nếu chỉ dựa vào logic định sẵn.Trong bối cảnh đó, việc ứng dụng trí tuệ nhân tạo (AI), đặc biệt là các mô hình học máy (machine learning) không giám sát, đang nổi lên như một giải pháp tiềm năng để tự động phát hiện các hành vi bất thường (anomalies) và cảnh báo sớm nguy cơ trước khi sự cố thực sự xảy ra. Các mô hình như LogBERT (Zhang [9]) hay Autoencoder/LSTM đang cho thấy khả năng phát hiện hiệu quả các bất thường trong chuỗi log mà không cần gán nhãn thủ công. Chính vì vậy, đề tài "Phân tích log hệ thống để phát hiện lỗi tiềm ẩn bằng AI" được lựa chọn nhằm hướng tới một giải pháp có khả năng ứng dụng cao trong thực tế, góp phần nâng cao tính ổn định, an toàn và hiệu quả vận hành của các hệ thống CNTT hiện đại.

2. Tổng quan về vấn đề nghiên cứu

Phân tích log hệ thống là một lĩnh vực đã được quan tâm từ lâu trong ngành CNTT. Trong giai đoạn đầu, các phương pháp chủ yếu dựa trên rule-based hoặc phân tích thống kê để phát hiện mẫu lỗi lặp lại. Với sự phát triển mạnh mẽ của học máy, đặc biệt trong xử lý ngôn ngữ tự nhiên (NLP), nhiều mô hình đã được ứng dụng nhằm trích xuất đặc trưng từ log và xây dựng cơ chế cảnh báo thông minh. Một số hướng nghiên cứu nổi bật gồm sử dụng clustering để gom nhóm log tương đồng nhằm phát hiện log bất thường, áp dụng deep learning như Autoencoder, LSTM, Transformer để học biểu

diễn ngữ nghĩa log, và kết hợp phân tích chuỗi thời gian với học không giám sát để phát hiện hành vi lệch chuẩn. Các công trình gần đây như He [10] (2022) và Kim [5] (2024) đã chứng minh tiềm năng của mô hình học sâu không giám sát trong việc phát hiện lỗi chưa từng thấy trước đó, tuy nhiên vẫn tồn tại các thách thức như xử lý log đa nguồn và đa định dạng, khả năng tổng quát hóa hạn chế và khó khăn khi triển khai trong môi trường thực tế. Do đó, đề tài này hướng tới việc xây dựng một mô hình học không giám sát tổng quát hơn, có khả năng học đặc trưng log theo thời gian và phát hiện sớm các lỗi tiềm ẩn chưa từng xuất hiện.

3. Mục đích nghiên cứu

Mục tiêu chính của đề tài là xây dựng một mô hình trí tuệ nhân tạo có khả năng phân tích log hệ thống máy chủ để phát hiện lỗi tiềm ẩn một cách tự động, không cần nhãn (unsupervised learning).

Đề tài hướng đến việc phát triển một quy trình xử lý log tổng thể, từ giai đoạn thu thập log, chuẩn hóa và biểu diễn dữ liệu cho đến việc áp dụng các kỹ thuật học máy không giám sát nhằm phát hiện bất thường trong chuỗi sự kiện hệ thống. Cụ thể, đề tài đặt ra các mục tiêu sau: (1) xây dựng pipeline xử lý log hệ thống bao gồm thu thập, chuẩn hóa, biểu diễn và phát hiện bất thường; (2) thiết kế và huấn luyện mô hình học không giám sát có khả năng phát hiện các chuỗi log bất thường với độ chính xác cao; (3) tích hợp mô hình vào một hệ thống mô phỏng có khả năng phản hồi gần thời gian thực nhằm hỗ trợ cảnh báo sớm khi có dấu hiệu lỗi; và (4) đánh giá hiệu quả của mô hình thông qua các chỉ số như độ chính xác (precision), độ bao phủ (recall), F1-score và thời gian phản hồi.

4. Đối tượng và phạm vi nghiên cứu

Đối tượng nghiên cứu của đề tài là dữ liệu log sinh ra từ các hệ thống máy chủ backend.

Về phạm vi nghiên cứu, đề tài tập trung vào việc phát hiện lỗi tiềm ẩn có tính hệ thống như tình trạng CPU quá tải, lỗi I/O, exceptions, các dịch vụ bị ngừng hoạt động một cách bất thường hoặc biểu hiện bất thường theo chu kỳ thời gian. Đề tài không tập trung vào các hệ thống real-time với yêu cầu phản hồi tức thì dưới 1 giây và không bao

gồm việc phát hiện các sự cố do tấn công bảo mật như malware hoặc intrusion, mà chỉ tập trung vào các lỗi vận hành nội tại của hệ thống.

5. Phương pháp nghiên cứu

Đề tài áp dụng phương pháp nghiên cứu thực nghiệm theo quy trình gồm nhiều bước chính: thu thập dữ liệu log từ môi trường giả lập và các tập dữ liệu benchmark công khai như BGL (BlueGene/L) và HDFS log; tiền xử lý bao gồm chuẩn hóa định dạng log, parsing cấu trúc log để trích xuất thông tin có cấu trúc, lọc nhiễu, token hóa và vector hóa nội dung log bằng các kỹ thuật như TF-IDF, Word2Vec hoặc embedding theo ngữ cảnh như BERT; biểu diễn log dưới dạng chuỗi sự kiện theo thời gian hoặc ma trận đặc trưng để đưa vào mô hình học máy; huấn luyện mô hình học không giám sát như LogBERT; cuối cùng đánh giá hiệu quả mô hình qua các chỉ số precision, recall, F1-score và thời gian phản hồi dự đoán. Các công cụ hỗ trợ bao gồm Python với scikit-learn, TensorFlow, PyTorch; nền tảng giám sát log ELK Stack (Elasticsearch, Logstash, Kibana); và Kafka để mô phỏng luồng log thời gian thực.

6. Cấu trúc đề án tốt nghiệp

Ngoài phần mở đầu, kết luận, danh mục tài liệu tham khảo và phụ lục, luận văn gồm 03 chương như sau:

- Chương 1 trình bày nền tảng lý thuyết và bối cảnh nghiên cứu về bài toán phát hiện bất thường trong log hệ thống. Các nội dung cụ thể trong chương gồm: Đặc điểm của log hệ thống; Phân loại các loại lỗi; So sánh các phương pháp phát hiện bất thường từ dựa trên quy tắc đến dựa trên trí tuệ nhân tạo; Pipeline trí tuệ nhân tạo cho phát hiện bất thường; Lựa chọn Autoencoder và LogBERT; Các nghiên cứu liên quan. Các nội dung chương này tạo nền tảng cho việc xây dựng mô hình trong Chương 2.
- Chương 2 trình bày quá trình xây dựng mô hình phát hiện bất thường trong log hệ thống sử dụng kỹ thuật trí tuệ nhân tạo. Các nội dung cụ thể trong chương gồm: Xây dựng Pipeline xử lý dữ liệu Log; Lựa chọn mô hình Autoencoder và LogBERT; Phân đoạn từ với hai phương pháp phù hợp cho từng mô hình (word-level cho Autoencoder, subword-level cho LogBERT); và (3) Biểu diễn vector với Word2Vec

cho Autoencoder và BERT embeddings cho LogBERT; Huấn luyện mô hình với các tập dữ liệu HDFS và BGI; Thử nghiệm mô hình và tinh chỉnh các tham số theo các tiêu chí đánh giá.

- Chương 3 trình bày kết quả đánh giá các mô hình đã xây dựng và triển khai hệ thống pipeline thời gian thực. Các nội dung chính gồm: Đánh giá cả hai mô hình Autoencoder và LogBERT trên hai tập dữ liệu HDFS và BG; Xây dựng và triển khai Pipeline theo thời gian thực, sử dụng Apache Kafka, Docker container; Thực hiện kiểm thử, đánh giá hệ thống

CHƯƠNG 1: TỔNG QUAN VÀ CƠ SỞ LÝ THUYẾT

1.1 Giới thiệu bài toán phân tích log phát hiện bất thường

1.1.1 Bối cảnh

Ngày nay, các hệ thống công nghệ thông tin đang ngày càng trở nên phức tạp và phân tán. Với sự phát triển của kiến trúc microservices, cloud computing, và các hệ thống phân tán, chúng ta đang phải đối mặt với một môi trường vận hành có hàng nghìn, thậm chí hàng triệu thành phần độc lập. Mỗi thành phần trong hệ thống đều liên tục sinh ra các bản ghi hoạt động (log) với dung lượng khổng lồ. Theo nghiên cứu của Xu et al., các trung tâm dữ liệu hiện đại có thể tạo ra hàng trăm GB log mỗi ngày. Điều này khiến việc giám sát thủ công gần như không khả thi.

Log không chỉ đơn thuần là dữ liệu ghi nhận quá trình vận hành. Trong thực tế, log là nguồn thông tin quan trọng giúp chúng ta chẩn đoán sự cố, giám sát hiệu suất và đảm bảo an toàn hệ thống. Log chứa đựng thông tin chi tiết về các sự kiện xảy ra trong hệ thống, bao gồm các thao tác của người dùng, lỗi ứng dụng, cảnh báo bảo mật, và các sự kiện hệ thống. Nếu chúng ta có thể phân tích log một cách hiệu quả, chúng ta có thể phát hiện sớm các vấn đề tiềm ẩn trước khi chúng trở thành sự cố nghiêm trọng.

Tuy nhiên, việc phân tích log theo cách truyền thống – dựa trên quy tắc cố định (dựa trên quy tắc) hoặc thao tác thủ công – ngày càng trở nên bất cập. Với sự gia tăng về quy mô và tính đa dạng của dữ liệu log, những lỗi tiềm ẩn có thể không để lại dấu hiệu rõ ràng. Điều này dẫn đến nguy cơ bỏ sót nếu chúng ta chỉ dựa vào logic định sẵn. Hơn nữa, với khối lượng log khổng lồ như hiện tại, việc phát hiện các mẫu bất thường tinh vi đòi hỏi sự hỗ trợ của các công nghệ tự động hóa và trí tuệ nhân tạo.

1.1.2 Bài toán phân tích log phát hiện bất thường

1) Định nghĩa về log của hệ thống

Log (bản ghi nhật ký) là các bản ghi tự động được tạo ra bởi hệ thống, ứng dụng hoặc dịch vụ để ghi nhận các sự kiện, hoạt động và trạng thái trong quá trình vận hành. Theo định nghĩa của He et al. [4], log là "một chuỗi các bản ghi được tạo ra bởi các thành phần hệ thống để ghi lại các sự kiện quan trọng trong quá trình thực thi". Log thường chứa các thông tin như timestamp (thời gian), log level (mức độ nghiêm trọng), source (nguồn), message (nội dung), và các metadata khác.

Trong hệ thống công nghệ thông tin hiện đại, log có thể được phân loại theo nhiều tiêu chí. Theo nguồn phát sinh, log bao gồm: log hệ thống (system log) - ghi nhận các sự kiện của hệ điều hành như khởi động, tắt máy, lỗi phần cứng; log ứng dụng (application log) - ghi nhận các sự kiện của ứng dụng như xử lý request, lỗi logic, giao dịch; log mạng (network log) - ghi nhận các sự kiện mạng như kết nối, truyền dữ liệu, tấn công; log bảo mật (security log) - ghi nhận các sự kiện bảo mật như đăng nhập, truy cập, vi phạm chính sách; và log cơ sở dữ liệu (database log) - ghi nhận các thao tác trên cơ sở dữ liệu như truy vấn, giao dịch, sao lưu.

Theo định dạng, log có thể được chia thành: log có cấu trúc (structured log) - sử dụng định dạng có cấu trúc như JSON, XML, CSV, dễ dàng phân tích tự động; log bán cấu trúc (semi-structured log) - có một phần cấu trúc như syslog với các trường cố định nhưng message tự do; và log không cấu trúc (unstructured log) - văn bản thuần túy, khó phân tích tự động.

Log đóng vai trò quan trọng trong việc giám sát, chẩn đoán và bảo mật hệ thống. Chúng cung cấp "lịch sử hoạt động" của hệ thống, giúp các kỹ sư hiểu được những gì đã xảy ra, tại sao hệ thống hoạt động theo cách đó, và làm thế nào để khắc phục các vấn đề. Tuy nhiên, với khối lượng log khổng lồ được tạo ra hàng ngày (có thể lên tới hàng trăm GB), việc phân tích log thủ công trở nên không khả thi, đòi hỏi các phương pháp tự động hóa và trí tuệ nhân tạo.

2) Định nghĩa bất thường trong log

Bất thường (anomaly) trong log có thể được định nghĩa là các sự kiện hoặc chuỗi sự kiện trong log mà hành vi của chúng lệch khỏi mẫu bình thường của hệ thống.

Chandola [3] định nghĩa anomaly là "một mẫu trong dữ liệu không phù hợp với hành vi được mong đợi hoặc đã biết". Trong ngữ cảnh log hệ thống, anomaly có thể biểu hiện dưới ba hình thức chính.

Đầu tiên là Point Anomaly - một log entry đơn lẻ có giá trị bất thường so với phần còn lại. Ví dụ, một log level ERROR hoặc FATAL trong một chuỗi log bình thường là một point anomaly. Thứ hai là Contextual Anomaly - một log entry là bất thường trong một ngữ cảnh cụ thể nhưng có thể bình thường trong ngữ cảnh khác. Ví dụ, CPU usage 95% là bình thường trong giờ cao điểm nhưng bất thường trong giờ thấp điểm. Cuối cùng là Collective Anomaly - một tập hợp các log entries là bình thường khi xem riêng lẻ nhưng tạo thành mẫu bất thường khi xem cùng nhau. Ví dụ, một chuỗi log lệch khỏi workflow bình thường là một collective anomaly.

3) Tầm quan trọng của phát hiện sớm lỗi tiềm ẩn

Việc phát hiện sớm lỗi tiềm ẩn có ý nghĩa rất quan trọng đối với vận hành hệ thống. Đầu tiên, nó giúp giảm thiểu downtime - phát hiện sớm các dấu hiệu lỗi cho phép can thiệp trước khi hệ thống bị crash, giúp giảm thiểu thời gian downtime và thiệt hại kinh tế. Nghiên cứu cho thấy chi phí downtime trung bình cho các doanh nghiệp lớn có thể lên tới hàng nghìn USD mỗi phút.

Thứ hai, phát hiện sớm giúp bảo vệ dữ liệu - nhiều lỗi hệ thống có thể dẫn đến mất mát dữ liệu. Phát hiện sớm các dấu hiệu bất thường trong thao tác I/O hoặc giao dịch cơ sở dữ liệu có thể giúp ngăn chặn mất dữ liệu trước khi sự cố xảy ra.

Thứ ba, phát hiện sớm giúp đảm bảo an toàn - log bất thường có thể chỉ ra các cuộc tấn công bảo mật hoặc hành vi đáng ngờ. Phát hiện sớm các mẫu bất thường trong log access, authentication failures, hoặc network traffic có thể giúp ngăn chặn các mối đe dọa bảo mật.

Cuối cùng, phát hiện sớm giúp cải thiện hiệu suất - phát hiện các anomaly liên quan đến suy giảm hiệu năng (như tăng đột biến CPU, rò rỉ bộ nhớ, hoặc truy vấn chậm) giúp tối ưu hóa hiệu suất hệ thống trước khi vấn đề trở nên nghiêm trọng.

4) Hậu quả của việc bỏ sót lỗi và trọng tâm nghiên cứu

Việc bỏ sót hoặc phát hiện muộn các lỗi tiềm ẩn có thể dẫn đến hậu quả nghiêm trọng: downtime (hệ thống crash hoặc ngừng hoạt động), mất dữ liệu (mất dữ liệu do lỗi disk, database corruption), vi phạm bảo mật (bỏ sót các dấu hiệu tấn công trong log), và thiệt hại tài chính (thiệt hại tài chính trực tiếp và gián tiếp cho doanh nghiệp).

Đề án tốt nghiệp này đóng góp vào lĩnh vực phát hiện bất thường trong log thông qua: đóng góp kỹ thuật (so sánh toàn diện giữa Autoencoder và LogBERT, tinh chỉnh siêu tham số chi tiết, quy trình xử lý log hoàn chỉnh, đánh giá với các chỉ số đầy đủ), đóng góp thực tiễn (hệ thống demo hoàn chỉnh, các thực hành tốt nhất cho triển khai, hướng dẫn tích hợp với các công cụ giám sát, phân tích chi phí và yêu cầu tài nguyên), và đóng góp khoa học (phân tích đánh đổi giữa hiệu năng và hiệu quả sử dụng tài nguyên, so sánh định lượng trên nhiều tập dữ liệu, tài liệu đầy đủ cho khả năng tái tạo).

1.1.3 Khoảng trống nghiên cứu

Mặc dù có nhiều nghiên cứu về phát hiện bất thường trong log, vẫn còn nhiều khoảng trống và thách thức cần giải quyết.

1) Hạn chế của phương pháp dựa trên quy tắc

Phương pháp dựa trên quy tắc truyền thống dựa trên các quy tắc định sẵn có những hạn chế cố hữu: không phát hiện được lỗi mới (chỉ có thể phát hiện các lỗi đã được định nghĩa trước), tỷ lệ dương tính giả cao (quá nhiều cảnh báo từ các mẫu bình thường), chi phí bảo trì (việc duy trì và cập nhật rules thủ công tốn kém), và không hiểu ngữ cảnh (không hiểu được ngữ cảnh và ngữ nghĩa của log message).

2) Thách thức của học không giám sát với dữ liệu log

Học không giám sát cho phép phát hiện anomaly mà không cần dữ liệu có nhãn, phù hợp với log vì việc gán nhãn thủ công tốn kém. Tuy nhiên, vẫn còn nhiều thách thức: chất lượng dữ liệu (log thường chứa nhiều nhiễu, thông tin dư thừa), điều chỉnh ngưỡng (các mô hình học không giám sát thường xuất ra điểm bất thường, cần phải điều chỉnh ngưỡng), khả năng giải thích (các mô hình học sâu khó giải thích tại sao một log

được coi là bất thường), và tổng quát hóa (các mô hình được huấn luyện trên một tập dữ liệu cụ thể có thể không tổng quát hóa tốt cho các hệ thống khác nhau).

3) Vấn đề xử lý log đa nguồn, đa định dạng

Hệ thống hiện đại thường có nhiều nguồn log khác nhau (log ứng dụng, log hệ thống, log mạng, log bảo mật, log cơ sở dữ liệu) và đa định dạng (log có cấu trúc như JSON, CSV; log bán cấu trúc như syslog; log không cấu trúc như văn bản thuần túy). Việc xử lý và phân tích log đa nguồn đòi hỏi chuẩn hóa (chuẩn hóa các định dạng khác nhau về một định dạng thống nhất), tích hợp (tích hợp log từ nhiều nguồn), và phân tích cú pháp (phân tích các log không cấu trúc để trích xuất thông tin có cấu trúc).

4) Thiếu khả năng tổng quát hóa

Nhiều phương pháp hiện tại được thiết kế cho các tập dữ liệu cụ thể hoặc loại log cụ thể, thiếu khả năng tổng quát hóa: theo miền (các mô hình được huấn luyện trên HDFS logs có thể không hoạt động tốt trên Apache logs), thích ứng (khi hệ thống thay đổi, các mô hình cần được huấn luyện lại hoặc tinh chỉnh), và học chuyển giao (việc chuyển giao kiến thức từ một miền sang miền khác vẫn còn là thách thức). Kim [5] chỉ ra rằng một trong những thách thức lớn nhất trong phát hiện bất thường log là xây dựng các mô hình có khả năng tổng quát hóa tốt trên nhiều loại log và hệ thống khác nhau.

1.2 Đặc điểm và phân loại log

1.2.1 Đặc điểm dữ liệu log

Log hệ thống có những đặc điểm riêng biệt so với các loại dữ liệu khác, tạo ra nhiều thách thức trong việc xử lý và phân tích. He [2] đã chỉ ra rằng log hệ thống có bốn đặc điểm chính: tính tuần tự theo thời gian, tính không đồng nhất, tính dư thừa và nhiễu, và khối lượng lớn.

Chuỗi sự kiện theo thời gian: Log được sinh ra liên tục theo thời gian, tạo thành chuỗi sự kiện có tính tuần tự và phụ thuộc thời gian. Mỗi log entry thường chứa dấu thời gian ghi lại thời điểm sự kiện xảy ra. Tính tuần tự này có ý nghĩa quan trọng vì các sự

kiện sau thường phụ thuộc vào các sự kiện trước. Ví dụ, một lỗi "connection closed" sau một "connection timeout" có ý nghĩa khác với khi nó xảy ra độc lập.

Tính không đồng nhất: Log từ nhiều nguồn khác nhau (ứng dụng, hệ thống, mạng, cơ sở dữ liệu) có định dạng khác nhau, tạo ra tính không đồng nhất. Điều này thể hiện qua đa định dạng (log có cấu trúc như JSON, log bán cấu trúc như syslog, log không cấu trúc như văn bản thuần túy), đa nguồn (log ứng dụng, log hệ thống, log mạng), và đa định dạng dấu thời gian. Tính không đồng nhất này tạo ra thách thức lớn trong việc phân tích cú pháp và chuẩn hóa log từ nhiều nguồn khác nhau.

Tính dư thừa và nhiễu: Log chứa nhiều thông tin lặp lại và nhiễu. Về dư thừa, nhiều log entries có cùng mẫu thông báo nhưng khác tham số. Ví dụ, "Connection to 192.168.1.1 failed" và "Connection to 192.168.1.2 failed" đều có cùng mẫu "Connection to failed". Về nhiễu, chúng ta có thể gặp log gỡ lỗi ở môi trường sản xuất, ghi log chi tiết từ thư viện bên thứ ba, và log entries không liên quan đến phát hiện bất thường. Việc trích xuất mẫu từ log messages là bước quan trọng để giảm dư thừa và tập trung vào thông tin quan trọng.

Khối lượng lớn: Hệ thống hiện đại có thể tạo ra hàng triệu, thậm chí hàng tỷ log entries mỗi ngày [1]. Điều này tạo ra thách thức về lưu trữ, xử lý (xử lý khối lượng lớn đòi hỏi hệ thống phân tán), và xử lý thời gian thực (xử lý thời gian thực đòi hỏi thông lượng cao và độ trễ thấp).

1.2.2 Cấu trúc log điển hình

Một log entry điển hình thường bao gồm các thành phần sau: dấu thời gian (thời điểm sự kiện xảy ra, có thể ở định dạng ISO 8601, syslog, hoặc định dạng tùy chỉnh), mức log (TRACE/DEBUG, INFO, WARN, ERROR, FATAL/CRITICAL), thành phần/nguồn (thành phần hoặc mô-đun tạo ra log), thông báo (nội dung chính của log

entry), và ngữ cảnh/siêu dữ liệu (tùy chọn - thông tin bổ sung như địa chỉ IP, mã người dùng, mã phiên, mã yêu cầu, dấu vết ngăn xếp).

1.2.3 Phân loại lỗi hệ thống

Khi làm việc với hệ thống phụ trợ, chúng ta thường xuyên gặp phải nhiều loại lỗi khác nhau. Mỗi loại lỗi có đặc điểm riêng và để lại những dấu vết khác nhau trong log. Việc hiểu rõ các loại lỗi này là nền tảng quan trọng để xây dựng hệ thống phát hiện bất thường hiệu quả.

Lỗi crash: Hệ thống hoặc ứng dụng dừng đột ngột, không thể tiếp tục hoạt động bình thường. Nguyên nhân có thể là lỗi phân đoạn, lỗi hết bộ nhớ, tràn ngăn xếp, ngoại lệ con trỏ null, hoặc ngoại lệ chưa xử lý. Trong log, chúng ta thường thấy các thông báo với mức FATAL hoặc CRITICAL, mã thoát khác 0, và dấu vết ngăn xếp đầy đủ.

Quá tải tài nguyên: Một hoặc nhiều tài nguyên của hệ thống (CPU, RAM, I/O) vượt quá khả năng xử lý. CPU quá tải xảy ra khi sử dụng CPU vượt quá ngưỡng bình thường (> 90%) trong thời gian dài. Bộ nhớ quá tải (rò rỉ bộ nhớ) là tình trạng bộ nhớ không được giải phóng sau khi sử dụng, dẫn đến cạn kiệt heap. I/O quá tải có thể xảy ra ở cả I/O đĩa và I/O mạng.

Lỗi timeout: Một thao tác không hoàn thành trong thời gian cho phép, thường do không có sẵn tài nguyên hoặc vấn đề hiệu suất. Có nhiều loại timeout: hết thời gian yêu cầu (yêu cầu HTTP không nhận được phản hồi), hết thời gian kết nối (kết nối đến server không thành công), và hết thời gian giao dịch (giao dịch cơ sở dữ liệu không hoàn thành).

Ngoại lệ và lỗi: Các lỗi được ném bởi mã ứng dụng, thường do lỗi logic, lỗi xác thực dữ liệu, hoặc điều kiện không mong đợi. Một số ngoại lệ phổ biến: ngoại lệ con trỏ null, ngoại lệ vượt quá chỉ số, ngoại lệ đối số không hợp lệ, ngoại lệ không tìm thấy tệp, ngoại lệ từ chối quyền.

Biểu hiện bất thường theo thời gian: Không phải là một log entry đơn lẻ mà là sự kết hợp của nhiều log entries tạo thành mẫu lệch khỏi bình thường. Loại lỗi này thường khó phát hiện hơn vì cần phải phân tích chuỗi log theo thời gian. Ví dụ: sai lệch quy

trình (chuỗi log lệch khỏi quy trình bình thường), bất thường thời gian (mẫu bất thường theo thời gian), bất thường tần suất (tần suất sự kiện bất thường).

Các loại lỗi hệ thống có thể được phân loại theo ba loại bất thường chính: bất thường điểm (các lỗi đơn lẻ như crash, rò rỉ bộ nhớ, ngoại lệ), bất thường ngữ cảnh (các lỗi theo ngữ cảnh như CPU quá tải trong giờ thấp điểm, timeout khi mạng bình thường), và bất thường tập thể (các chuỗi log bất thường như sai lệch quy trình, bất thường thời gian).

1.3 Các phương pháp phát hiện bất thường

1.3.1 Giới thiệu

Các phương pháp phát hiện bất thường trong log hệ thống đã phát triển qua nhiều giai đoạn, từ các quy tắc cố định (dựa trên quy tắc) đơn giản đến các mô hình học máy và cuối cùng là các mô hình học sâu dựa trên Transformer. Mỗi phương pháp có ưu điểm và nhược điểm riêng, phù hợp với các tình huống và yêu cầu khác nhau.

Trong phần này, đề án trình bày sự phát triển của các phương pháp từ truyền thống đến hiện đại, tập trung vào so sánh giữa phương pháp dựa trên quy tắc và phương pháp dựa trên trí tuệ nhân tạo, và phân tích ưu nhược điểm của từng phương pháp.

1.3.2 Phương pháp dựa trên quy tắc (truyền thống)

1) Nguyên lý hoạt động

Phương pháp dựa trên quy tắc là phương pháp truyền thống nhất, dựa trên các quy tắc định sẵn để phát hiện bất thường. Các quy tắc này thường được định nghĩa dựa trên kinh nghiệm của các chuyên gia hệ thống hoặc các mẫu lỗi đã biết. Các kỹ thuật phổ biến bao gồm khớp biểu thức chính quy, khớp từ khóa (tìm các từ khóa như ERROR, FATAL, CRITICAL), khớp mẫu, và dựa trên ngưỡng (so sánh giá trị với ngưỡng).

2) Ưu điểm và nhược điểm

Ưu điểm của phương pháp dựa trên quy tắc: đơn giản (dễ hiểu, dễ triển khai, không cần dữ liệu huấn luyện), hiệu quả (xử lý nhanh, không tốn tài nguyên), chính xác

cho lỗi đã biết (hiệu quả cao cho các lỗi đã được định nghĩa rõ ràng), khả năng giải thích (dễ giải thích tại sao một log được coi là bất thường), và thời gian thực (có thể xử lý thời gian thực với chi phí thấp).

Nhược điểm: không phát hiện được lỗi mới (chỉ phát hiện được các lỗi đã được định nghĩa trong quy tắc), tỷ lệ dương tính giả cao (quá nhiều cảnh báo từ các mẫu bình thường), chi phí bảo trì (cần cập nhật quy tắc thủ công khi hệ thống thay đổi), không hiểu ngữ cảnh (không hiểu được ngữ cảnh và ngữ nghĩa của log message), và không có khả năng mở rộng (khó mở rộng khi số lượng quy tắc tăng lên).

Phương pháp dựa trên quy tắc vẫn được sử dụng rộng rãi trong giám sát cơ bản, lọc dòng đầu tiên, các mẫu đã biết, và các yêu cầu độ trễ thấp.

1.3.3 Phương pháp dựa trên trí tuệ nhân tạo (học máy & học sâu)

1) Học máy cổ điển

Phương pháp học máy cổ điển sử dụng các thuật toán như phân cụm (K-means, DBSCAN, Isolation Forest), phương pháp thống kê (Z-score, phần trăm), và Máy vector hỗ trợ (SVM). Các phương pháp này đã tự động hóa hơn phương pháp dựa trên quy tắc nhưng vẫn cần kỹ thuật đặc trưng thủ công và khó xử lý được dữ liệu tuần tự và phụ thuộc ngữ cảnh trong log [3].

Ưu điểm: học không giám sát (không cần dữ liệu có nhãn), tự động phát hiện các mẫu mới, và tốt hơn phương pháp dựa trên quy tắc trong việc phát hiện lỗi mới. Nhược điểm: cần kỹ thuật đặc trưng thủ công, khó xử lý dữ liệu tuần tự, và hiệu năng phụ thuộc vào chất lượng của các đặc trưng.

2) Học sâu: Autoencoder

Autoencoder là một mạng nơ-ron có cấu trúc encoder-decoder, được thiết kế để học biểu diễn nén của dữ liệu đầu vào. Trong bối cảnh phát hiện bất thường log, encoder nén biểu diễn vector log vào không gian tiềm ẩn, decoder tái tạo biểu diễn vector log từ biểu diễn tiềm ẩn, và lỗi tái tạo được sử dụng để phân loại: log bình thường có lỗi tái tạo thấp, log bất thường có lỗi tái tạo cao.

Ưu điểm: học không giám sát (chỉ cần log bình thường để huấn luyện), phát hiện lỗi mới, tự động học đặc trưng (không cần kỹ thuật đặc trưng thủ công), và hiệu quả (huấn luyện nhanh hơn các mô hình phức tạp khác). Nhược điểm: điều chỉnh ngưỡng (cần điều chỉnh ngưỡng cẩn thận), bất thường ngữ cảnh (có thể bỏ sót các bất thường ngữ cảnh), và phụ thuộc dài hạn (khó học phụ thuộc dài hạn trong chuỗi log).

Du và đồng nghiệp đã giới thiệu DeepLog [4], sử dụng LSTM-Autoencoder để phát hiện bất thường trong chuỗi log, đạt được điểm F1 cao trên tập dữ liệu HDFS.

3) Học sâu: Dựa trên Transformer (LogBERT)

LogBERT là một mô hình dựa trên Transformer được thiết kế đặc biệt cho phát hiện bất thường log, dựa trên kiến trúc BERT [8]. Zhang [9] đã giới thiệu LogBERT trong [9], đạt được hiệu năng hàng đầu trên nhiều tập dữ liệu.

LogBERT sử dụng kiến trúc BERT-base với các thành phần chính: phân đoạn từ (chia log message thành tokens), biểu diễn vector (token embeddings + position embeddings), khối transformer (multi-head self-attention, mạng feed-forward), và đầu ra (điểm bất thường từ token [CLS] hoặc lỗi tái tạo).

Về tiền huấn luyện (Mô hình ngôn ngữ che mặt nạ), LogBERT dự đoán các token bị che trong log messages, sử dụng kho ngữ liệu log lớn, và mô hình học được ngữ nghĩa log tổng quát. Về tinh chỉnh (Phát hiện bất thường), LogBERT thực hiện nhiệm vụ phát hiện bất thường trên tập dữ liệu cụ thể, có thể sử dụng đầu phân loại hoặc lỗi tái tạo, và tinh chỉnh từ trọng số đã được tiền huấn luyện.

Ưu điểm của LogBERT: hiệu quả cao (điểm F1 cao hơn Autoencoder, 0.92 vs 0.87 trên tập dữ liệu HDFS), hiểu ngữ nghĩa (học được ngữ nghĩa và ngữ cảnh của log messages), học chuyển giao (có thể sử dụng trọng số BERT đã được tiền huấn luyện), phụ thuộc dài hạn (xử lý tốt phụ thuộc trong chuỗi log dài), và tổng quát hóa (tổng quát hóa tốt trên nhiều loại log khác nhau).

Nhược điểm: tốn tài nguyên (tốn tài nguyên hơn Autoencoder, bộ nhớ GPU ~8GB vs ~2GB, thời gian huấn luyện ~12h vs ~2h), độ phức tạp huấn luyện (cần tiền

huấn luyện hoặc tinh chỉnh từ mô hình đã được tiền huấn luyện), khả năng giải thích (khó giải thích tại sao một log được coi là bất thường), và yêu cầu tập dữ liệu (cần tập dữ liệu lớn để tiền huấn luyện hiệu quả).

1.3.4 So sánh phương pháp dựa trên quy tắc và trí tuệ nhân tạo

Về hiệu năng so với độ phức tạp: Phương pháp dựa trên quy tắc có hiệu năng thấp và độ phức tạp thấp, học máy cổ điển có hiệu năng trung bình và độ phức tạp trung bình, Autoencoder có hiệu năng cao và độ phức tạp trung bình, LogBERT có hiệu năng rất cao và độ phức tạp cao.

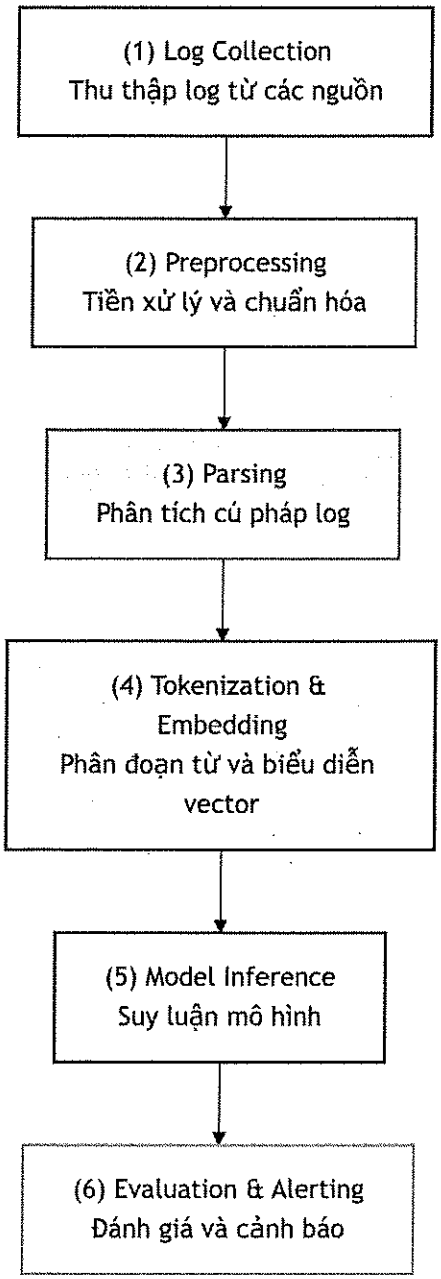
Về hiệu năng so với sử dụng tài nguyên: Phương pháp dựa trên quy tắc có tài nguyên thấp và hiệu năng thấp, Autoencoder có tài nguyên trung bình và hiệu năng cao, LogBERT có tài nguyên cao và hiệu năng rất cao.

Về khả năng giải thích so với hiệu năng: Phương pháp dựa trên quy tắc có khả năng giải thích rất cao và hiệu năng thấp, Autoencoder có khả năng giải thích thấp và hiệu năng cao, LogBERT có khả năng giải thích thấp và hiệu năng rất cao.

1.4 Pipeline trí tuệ nhân tạo cho phát hiện bất thường từ log

1.4.1 Giới Thiệu

Pipeline xử lý log để phát hiện bất thường bằng trí tuệ nhân tạo là một quy trình tổng thể bao gồm nhiều bước từ thu thập log đến cảnh báo. Trong nghiên cứu này, đề án tập trung vào pipeline có thể xử lý cả lô và thời gian thực, với tập trung vào xử lý thời gian thực cho hệ thống sản xuất.



Hình 1.1: Flowchart minh họa pipeline xử lý log 6 bước: (1) Thu thập log, (2) Tiền xử lý, (3) Phân tích cú pháp, (4) Phân đoạn từ & Biểu diễn vector, (5) Suy luận mô hình, (6) Đánh giá & Cảnh báo.

1.4.2 Pipeline Thực Hiện

Pipeline trong nghiên cứu này bao gồm 6 bước chính:

Bước 1: Thu thập log - Log được thu thập từ file logs hoặc hàng đợi tin nhắn (Apache Kafka). Trong nghiên cứu này, đề án sử dụng Kafka làm hàng đợi tin nhắn cho xử lý thời gian thực.

Bước 2: Tiền xử lý - Chuẩn hóa định dạng (chuẩn hóa dấu thời gian về ISO 8601, chuẩn hóa mức log), lọc nhiễu (loại bỏ log DEBUG trong sản xuất), và xử lý dấu thời gian (chuyển đổi múi giờ về UTC).

Bước 3: Phân tích cú pháp log - Sử dụng Drain parser [6] để trích xuất mẫu và tham số từ log messages. Drain parser đạt được độ chính xác phân tích cú pháp cao (> 95%) và có thể xử lý trực tuyến với thông lượng cao.

Bước 4: Phân đoạn từ & Biểu diễn vector - Phân đoạn từ cấp từ với Word2Vec cho Autoencoder, phân đoạn từ cấp từ con với BERT tokenizer cho LogBERT. Biểu diễn vector Word2Vec cho Autoencoder, biểu diễn vector BERT cho LogBERT.

Bước 5: Huấn luyện/Suy luận mô hình - Huấn luyện: Autoencoder tối thiểu hóa lỗi tái tạo, LogBERT tinh chỉnh từ BERT đã được tiền huấn luyện. Suy luận: suy luận thời gian thực với tính toán điểm bất thường (lỗi tái tạo cho Autoencoder, xác suất phân loại cho LogBERT).

Bước 6: Đánh giá & Cảnh báo - Đánh giá với độ chính xác, độ nhạy, điểm F1, ROC-AUC, ma trận nhầm lẫn. Điều chỉnh ngưỡng sử dụng dựa trên phần trăm (phần trăm thứ 95). Cảnh báo qua bảng điều khiển, email, và Telegram.

1.5. Kết luận chương 1

Chương 1 đã trình bày nền tảng lý thuyết và bối cảnh nghiên cứu về bài toán phát hiện bất thường trong log hệ thống. Chương đã giới thiệu các đặc điểm của log hệ thống, phân loại các loại lỗi, và so sánh các phương pháp phát hiện bất thường từ dựa trên quy tắc đến dựa trên trí tuệ nhân tạo. Pipeline trí tuệ nhân tạo cho phát hiện bất thường đã được mô tả, cùng với lý do lựa chọn Autoencoder và LogBERT làm hai mô hình nghiên cứu chính.

Chương này cũng đã tổng hợp các nghiên cứu liên quan và nêu rõ các đóng góp của đề án tốt nghiệp, bao gồm so sánh toàn diện giữa Autoencoder và LogBERT, hệ thống thời gian thực end-to-end, và các thực hành tốt nhất cho triển khai thực tế. Các nội dung trong chương này tạo nền tảng vững chắc cho việc xây dựng và đánh giá mô hình trong Chương 2.

CHƯƠNG 2: XÂY DỰNG MÔ HÌNH

Trong chương này, đề án tốt nghiệp tập trung vào quá trình thiết kế và xây dựng mô hình phát hiện bất thường trong log hệ thống dựa trên kỹ thuật trí tuệ nhân tạo. Đây là phần trọng tâm của toàn bộ nghiên cứu, thể hiện cách các phương pháp học sâu được áp dụng vào xử lý dữ liệu log có tính phức tạp cao và khối lượng lớn.

Trước hết, chương sẽ trình bày pipeline xử lý dữ liệu log từ giai đoạn tiền xử lý (phân tích cú pháp, phân đoạn từ) đến biểu diễn đặc trưng (biểu diễn vector) – nhằm chuẩn bị dữ liệu đầu vào cho mô hình học máy. Tiếp theo, đề án giải thích lý do lựa chọn và kết hợp hai mô hình Autoencoder và LogBERT, dựa trên khả năng học biểu diễn và phát hiện bất thường của chúng.

Phần tiếp theo trình bày chi tiết quy trình huấn luyện mô hình trên hai bộ dữ liệu thực tế HDFS và BGL, bao gồm việc lựa chọn siêu tham số (hyperparameters) và theo dõi quá trình hội tụ thông qua đường cong mất mát (loss curve). Cuối cùng, chương trình bày các tiêu chí đánh giá sẽ được sử dụng để đánh giá hiệu quả của mô hình trong Chương 3.

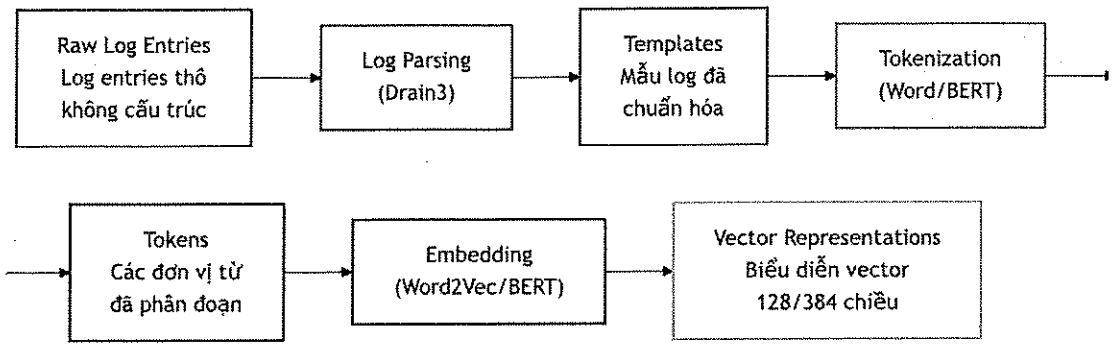
2.1 Kiến trúc pipeline dữ liệu

2.1.1 Tổng quan pipeline

Pipeline xử lý dữ liệu log là bước đầu tiên và quan trọng trong việc xây dựng mô hình phát hiện bất thường từ log hệ thống. Log được thu thập từ các nguồn khác nhau có định dạng khác nhau, do đó cần chuẩn hóa về dạng mẫu thống nhất để có thể xử lý bằng các kỹ thuật học máy. Pipeline xử lý dữ liệu trong nghiên cứu này bao gồm ba bước chính: (1) Phân tích cú pháp log - trích xuất mẫu từ log message, (2) Phân đoạn từ - chia log message thành các tokens có thể xử lý được, và (3) Biểu diễn vector - chuyển đổi tokens thành biểu diễn vector phù hợp với đầu vào của mô hình.

Quá trình xử lý bắt đầu từ log entries thô với định dạng không cấu trúc, qua các bước tiền xử lý để tạo ra biểu diễn vector có thể sử dụng cho huấn luyện và suy luận. Mỗi bước trong pipeline được thiết kế để giải quyết các thách thức cụ thể của dữ liệu

log: tính không đồng nhất, dư thừa, và khối lượng lớn [1]. Việc thiết kế pipeline một cách có hệ thống giúp đảm bảo chất lượng dữ liệu đầu vào cho các mô hình học máy.



Hình 2.1: Minh họa pipeline xử lý dữ liệu log từ log entries thô đến biểu diễn vector, bao gồm các bước: phân tích cú pháp, phân đoạn từ, và biểu diễn vector.

2.1.2 Phân tích cú pháp log

1) Mục đích và phương pháp

Phân tích cú pháp log là bước đầu tiên trong pipeline, có nhiệm vụ trích xuất mẫu (pattern) từ log message và tách biệt phần cố định (hằng số) và phần biến đổi (biến). Mục đích chính bao gồm chuẩn hóa log message, giảm số chiều (từ hàng triệu log entries xuống còn vài trăm mẫu), và tách biệt cấu trúc và nội dung. Việc tách biệt mẫu và tham số giúp giảm đáng kể số lượng mẫu duy nhất, giúp mô hình học hiệu quả hơn.

Trong nghiên cứu này, đề án sử dụng Drain parser [2], một trình phân tích log trực tuyến với cấu trúc cây độ sâu cố định. Drain parser được chọn vì có hiệu quả cao với thời gian phân tích cú pháp tuyến tính, hỗ trợ phân tích cú pháp trực tuyến, hoạt động hoàn toàn không giám sát, và cấu trúc cây độ sâu cố định giúp tăng tốc độ tìm kiếm mẫu. Cấu hình Drain parser sử dụng độ sâu 4, ngưỡng tương tự 0.4, và số con tối đa 100 (mặc định).

2) Triển khai và kết quả

Triển khai sử dụng thư viện drain3 [3], một triển khai Python của Drain parser. Parser được cấu hình để xử lý cả hai tập dữ liệu HDFS và BGL với các định dạng khác nhau.

Kết quả phân tích cú pháp cho thấy hiệu quả tốt. Với tập dữ liệu HDFS, tổng số log entries là khoảng 11,000,000, số lượng mẫu duy nhất là khoảng 500, tỷ lệ thành công phân tích cú pháp là hơn 99%, và thời gian phân tích cú pháp là khoảng 0.1ms mỗi log entry. Với tập dữ liệu BGL, tổng số log entries là khoảng 4,700,000, số lượng mẫu duy nhất là khoảng 400, tỷ lệ thành công phân tích cú pháp là hơn 98%, và thời gian phân tích cú pháp là khoảng 0.15ms mỗi log entry. Tỷ lệ thành công phân tích cú pháp hơn 98% là rất tốt và đủ cho việc huấn luyện mô hình.

2.1.3 Phân đoạn từ

1) Tổng quan

Phân đoạn từ là quá trình chia log message thành các đơn vị nhỏ hơn (tokens) có thể xử lý được bởi mô hình học máy. Có hai phương pháp phân đoạn từ chính được sử dụng: phân đoạn từ cấp từ (chia log message thành các từ riêng biệt, phù hợp cho Autoencoder) và phân đoạn từ cấp từ con (sử dụng BPE hoặc WordPiece, phù hợp cho LogBERT). Việc sử dụng phương pháp phân đoạn từ phù hợp với từng mô hình là rất quan trọng để đạt được hiệu quả tốt nhất.

2) Phân đoạn từ cấp từ cho Autoencoder

Phân đoạn từ cấp từ chia log message thành các từ riêng biệt dựa trên khoảng trắng và các ký tự đặc biệt. Quá trình bao gồm tiền xử lý văn bản, tách token, xây dựng từ vựng, và ánh xạ ID token. Cấu hình sử dụng kích thước từ vựng từ 5,000 đến 10,000 tokens (tùy tập dữ liệu), độ dài chuỗi tối đa là 512 tokens, và chiến lược đệm là đệm bên phải với token <PAD>.

Với tập dữ liệu HDFS, thống kê từ vựng cho thấy tổng số từ duy nhất là khoảng 15,000, kích thước từ vựng là 5,000 (các từ thường xuyên nhất), độ phủ là khoảng 95%, và tỷ lệ từ ngoài từ vựng là khoảng 5%. Độ phủ 95% là tốt và đủ cho việc huấn luyện mô hình.

3) Phân đoạn từ cấp từ con cho LogBERT

Phân đoạn từ cấp từ con sử dụng BPE hoặc WordPiece để chia từ thành các đơn vị từ con nhỏ hơn. Phân đoạn từ BERT sử dụng thuật toán WordPiece với kích thước từ

vựng khoảng 30,000 tokens. Cấu hình sử dụng mô hình distilbert-base-uncased, kích thước từ vựng là 30,522 tokens, độ dài chuỗi tối đa là 512 tokens.

Với tập dữ liệu HDFS, thống kê phân đoạn từ cho thấy số token trung bình mỗi log là khoảng 25 tokens, số token tối đa là khoảng 150 tokens, và tỷ lệ đệm là khoảng 5%. Số token trung bình mỗi log là 25 tokens là khá ngắn, giúp giảm chi phí tính toán.

2.1.4 Biểu diễn vector

1) Tổng quan

Biểu diễn vector là quá trình chuyển đổi tokens thành biểu diễn vector có thể sử dụng làm đầu vào cho mô hình học máy. Autoencoder sử dụng biểu diễn vector Word2Vec để tạo vector kích thước cố định cho mỗi log entry. LogBERT sử dụng biểu diễn vector BERT với token embeddings, position embeddings, và segment embeddings.

2) Biểu diễn vector Word2Vec cho Autoencoder

Word2Vec [4] là phương pháp biểu diễn vector phổ biến để tạo biểu diễn phân tán của từ. Kiến trúc Word2Vec sử dụng mô hình Skip-gram, với chiều biểu diễn vector là 128, kích thước cửa sổ là 5, số lần xuất hiện tối thiểu là 2, và số epochs huấn luyện là 5. Quá trình huấn luyện bao gồm chuẩn bị kho ngữ liệu, huấn luyện mô hình, trích xuất biểu diễn vector, và biểu diễn vector chuỗi (trung bình hoặc tổng biểu diễn vector từ).

Với tập dữ liệu HDFS, thống kê biểu diễn vector cho thấy kích thước từ vựng là khoảng 5,000 từ, chiều biểu diễn vector là 128, kích thước kho ngữ liệu huấn luyện là khoảng 11M log entries, và thời gian huấn luyện là khoảng 10 phút (trên CPU). Thời gian huấn luyện 10 phút là khá nhanh và chấp nhận được.

3) Biểu diễn vector BERT cho LogBERT

BERT [5] sử dụng kiến trúc transformer để tạo biểu diễn vector theo ngữ cảnh. Các thành phần biểu diễn vector BERT bao gồm token embeddings, position embeddings, segment embeddings, và biểu diễn vector cuối cùng (tổng của các biểu diễn vector). Cấu hình mô hình BERT sử dụng mô hình cơ sở distilbert-base-uncased, kích thước ẩn là 384, số lớp là 6, và số đầu chú ý là 6.

Trích xuất biểu diễn vector sử dụng biểu diễn vector token [CLS] hoặc trung bình của tất cả biểu diễn vector token. Thống kê biểu diễn vector BERT cho thấy mô hình là distilbert-base-uncased, kích thước ẩn là 384, số tham số là khoảng 66M, và thời gian suy luận là khoảng 5ms mỗi log entry (trên CPU). Thời gian suy luận 5ms mỗi log entry là khá nhanh và chấp nhận được cho xử lý thời gian thực.

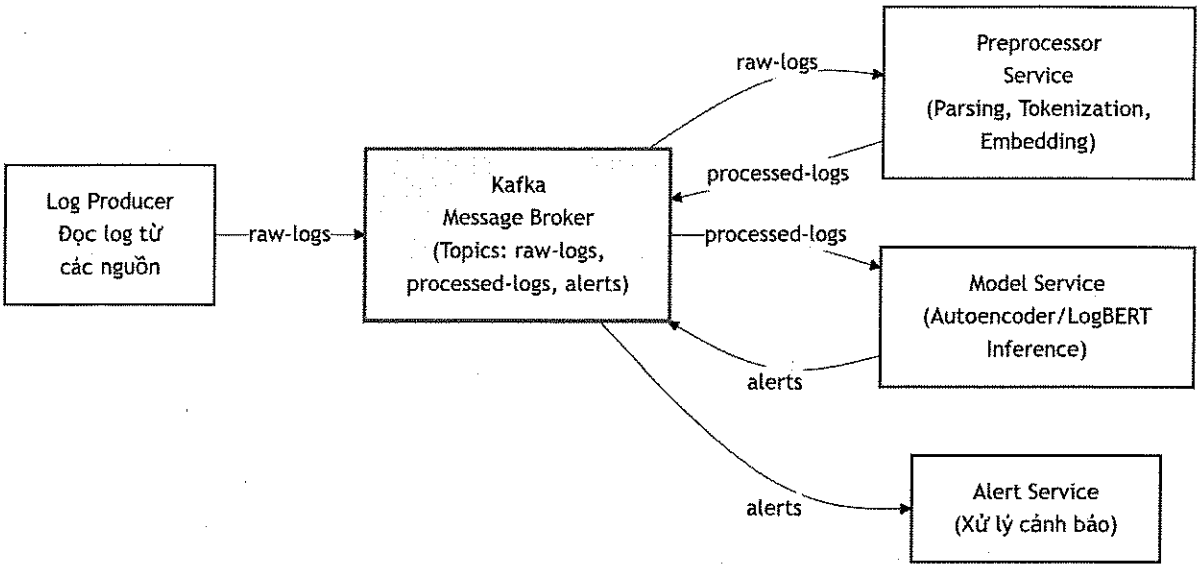
2.2 Kiến trúc tổng thể hệ thống

1) Tổng quan kiến trúc

Hệ thống phát hiện bất thường trong log được thiết kế theo kiến trúc pipeline thời gian thực sử dụng Apache Kafka làm hạ tầng truyền dữ liệu. Pipeline được thiết kế để xử lý luồng log liên tục với độ trễ thấp và khả năng mở rộng cao, đáp ứng yêu cầu của môi trường sản xuất [1].

Kiến trúc hệ thống được xây dựng dựa trên các nguyên tắc: 1) Kiến trúc Microservices: mỗi thành phần được tách thành dịch vụ độc lập, có thể mở rộng và triển khai riêng biệt [2]; 2) Kiến trúc hướng sự kiện: sử dụng Apache Kafka làm message broker để xử lý luồng log theo mô hình hướng sự kiện [3]; 3) Containerization: tất cả dịch vụ được containerized bằng Docker, giúp đơn giản hóa triển khai; và 4) Xử lý bất đồng bộ giữa các dịch vụ thông qua Kafka topics.

Kiến trúc tổng thể bao gồm: Message Broker (Apache Kafka), Microservices (Log Producer, Preprocessor, Model Service, Alert Service), Containerization (Docker và Docker Compose), và Xử Lý Bất Đồng Bộ (xử lý bất đồng bộ giữa các dịch vụ).



Hình 2.2: Kiến trúc tổng thể của pipeline thời gian thực phát hiện bất thường trong log. Hệ thống sử dụng Kafka làm message broker, với các dịch vụ độc lập có thể mở rộng riêng biệt.

2.3 Chọn mô hình Autoencoder và LogBERT

2.3.1 Tổng quan lựa chọn

Việc lựa chọn mô hình cho phát hiện bất thường log là một quyết định quan trọng ảnh hưởng đến hiệu quả, khả năng triển khai, và chi phí vận hành của hệ thống. Trong nghiên cứu này, đề án chọn hai mô hình: Autoencoder làm mô hình cơ sở và LogBERT làm mô hình chính. Lựa chọn này được dựa trên các tiêu chí: học không giám sát (không cần dữ liệu có nhãn), hiệu quả (đạt điểm F1 cao trên các tập dữ liệu chuẩn), tính tổng quát (áp dụng cho nhiều loại log), khả năng triển khai (cân bằng hiệu năng và sử dụng tài nguyên), và độ chín (mô hình đã được nghiên cứu và chứng minh hiệu quả).

LogGPT không được lựa chọn do yêu cầu tài nguyên quá lớn (GPU V100/A100, bộ nhớ >40GB), không phù hợp với đề án tốt nghiệp ứng dụng và môi trường triển khai thực tế của các doanh nghiệp vừa và nhỏ [1].

2.3.2 Autoencoder - Mô hình cơ sở

1) Lý do chọn Autoencoder làm mô hình cơ sở

Autoencoder được chọn làm mô hình cơ sở do tính đơn giản (kiến trúc đơn giản, dễ triển khai và gỡ lỗi), hiệu quả (đã được chứng minh hiệu quả trong phát hiện bất thường log [2], [3]), làm mô hình cơ sở tốt để so sánh với LogBERT, và hiệu quả tài nguyên (không cần GPU mạnh, huấn luyện nhanh, phù hợp với môi trường hạn chế tài nguyên).

2) Nguyên lý hoạt động và kiến trúc

Autoencoder hoạt động dựa trên nguyên lý tái tạo: mô hình học cách nén đầu vào vào không gian tiềm ẩn (nút thắt) và sau đó tái tạo lại đầu vào từ biểu diễn tiềm ẩn. Log entries bình thường có lỗi tái tạo thấp vì mô hình đã học được mẫu của chúng, trong khi log entries bất thường có lỗi tái tạo cao vì mô hình chưa gặp mẫu tương tự trong quá trình huấn luyện.

Kiến trúc Autoencoder sử dụng các lớp kết nối đầy đủ với cấu trúc encoder-decoder đối xứng. Encoder nén biểu diễn vector log (128 chiều) vào không gian tiềm ẩn (32 chiều) qua các lớp ẩn [256, 128, 64]. Decoder tái tạo biểu diễn vector log từ biểu diễn tiềm ẩn qua các lớp ẩn [64, 128, 256]. Cấu hình mặc định: chiều đầu vào 128 (biểu diễn vector Word2Vec), chiều lớp ẩn [256, 128, 64], chiều tiềm ẩn 32, hàm kích hoạt ReLU, tỷ lệ dropout 0.1, và tổng số tham số ~150,000. Hàm mất mát là Lỗi bình phương trung bình (MSE) giữa đầu vào và đầu ra tái tạo.

3) Siêu tham số và huấn luyện

Siêu tham số huấn luyện bao gồm tốc độ học 0.001, kích thước lô 64-128, số epochs 50-100 (dừng sớm), bộ tối ưu hóa Adam, suy giảm trọng số 0.0001. Chiến lược huấn luyện sử dụng huấn luyện không giám sát (chỉ log bình thường), chia kiểm định 20%, dừng sớm với kiên nhẫn 10 epochs, và lưu điểm kiểm tra mô hình dựa trên mất mát kiểm định.

4) Ưu điểm và nhược điểm

Ưu điểm: không giám sát (không cần dữ liệu có nhãn), đơn giản và dễ triển khai, huấn luyện nhanh (~1-2 giờ trên CPU), hiệu quả cho nhiều loại bất thường, làm mô hình cơ sở tốt, hiệu quả tài nguyên (bộ nhớ ~500MB, không cần GPU). Nhược điểm: cần

điều chỉnh ngưỡng cẩn thận, có thể bỏ sót các bất thường tinh vi, khó học phụ thuộc dài hạn, không học được ngữ nghĩa và ngữ cảnh của log.

2.3.3 LogBERT - Mô hình chính

1) Lý do chọn chọn LogBERT làm mô hình chính

LogBERT được chọn làm mô hình chính do hiệu quả cao (hiệu năng hàng đầu [4]), hiểu ngữ nghĩa (học được ngữ nghĩa và ngữ cảnh của log), học chuyển giao (sử dụng trọng số BERT đã được tiền huấn luyện), phụ thuộc dài hạn (xử lý tốt phụ thuộc trong chuỗi log), và tinh chỉnh (có thể tinh chỉnh cho miền cụ thể).

2) Nguyên lý hoạt động và kiến trúc

LogBERT hoạt động dựa trên kiến trúc BERT [5] với tinh chỉnh cho nhiệm vụ phát hiện bất thường. Mô hình sử dụng kiến trúc transformer với multi-head self-attention để học biểu diễn hai chiều của chuỗi log. Quá trình hoạt động bao gồm phân đoạn từ (chia mẫu log thành tokens), biểu diễn vector (tạo biểu diễn vector cho tokens), khối transformer (xử lý biểu diễn vector qua các lớp transformer), đầu phân loại (sử dụng biểu diễn token [CLS]), và phát hiện bất thường (sử dụng logits phân loại).

Kiến trúc LogBERT sử dụng DistilBERT [6] làm mô hình cơ sở (phiên bản nhẹ hơn của BERT) để cân bằng giữa hiệu năng và sử dụng tài nguyên. Kiến trúc cơ sở sử dụng mô hình distilbert-base-uncased, kích thước ẩn 384, số lớp 6 khối transformer, số đầu chú ý 6, độ dài chuỗi tối đa 512 tokens, và tổng số tham số ~66M. Đầu phân loại sử dụng biểu diễn token [CLS] (384 chiều), lớp ẩn Linear(384 → 192) + ReLU + Dropout(0.1), và lớp đầu ra Linear(192 → 2) cho phân loại nhị phân.

3) Siêu tham số và huấn luyện

Siêu tham số huấn luyện (tinh chỉnh) bao gồm tốc độ học 2e-5, kích thước lô 16-32, số epochs 3-5, bộ tối ưu hóa AdamW với suy giảm trọng số 0.01, bộ lập lịch tốc độ học (khởi động tuyến tính 10% + suy giảm tuyến tính), cắt gradient với chuẩn tối đa 1.0. Chiến lược huấn luyện sử dụng trọng số DistilBERT đã được tiền huấn luyện, tinh chỉnh trên kho ngữ liệu log, huấn luyện không giám sát (chỉ log bình thường, nhãn = 0), chia kiểm định 20%, dừng sớm với kiên nhẫn 3 epochs, và lưu điểm kiểm tra mô hình.

4) Ưu điểm và nhược điểm

Ưu điểm: hiệu quả cao nhất trong các mô hình hiện tại, học được ngữ nghĩa và ngữ cảnh, học chuyển giao từ mô hình đã được tiền huấn luyện, xử lý tốt phụ thuộc dài hạn, có thể tinh chỉnh cho miền cụ thể, hiểu hai chiều. Nhược điểm: tốn tài nguyên hơn Autoencoder (bộ nhớ GPU ~8GB, thời gian huấn luyện ~6-12h), huấn luyện phức tạp hơn, cần GPU để huấn luyện hiệu quả, kích thước mô hình lớn (~66M tham số), thời gian suy luận chậm hơn (~5ms vs ~1ms mỗi log).

2.4 Quy trình huấn luyện và tinh chỉnh

2.4.1 Thiết lập huấn luyện

1) Môi trường huấn luyện

Quá trình huấn luyện được thực hiện trên hai môi trường khác nhau: môi trường cục bộ cho kiểm thử và kiểm định ban đầu, và Google Colab cho huấn luyện tập dữ liệu đầy đủ. Môi trường cục bộ sử dụng Apple M1/M2 với 16GB RAM (chỉ CPU), trong khi Colab cung cấp GPU NVIDIA T4 hoặc V100 với 16GB VRAM. Cấu hình phần mềm sử dụng Python 3.8+ trên cục bộ và 3.10+ trên Colab, PyTorch 1.12+ với hỗ trợ CUDA cho Colab.

2) Tập dữ liệu sử dụng

Huấn luyện được thực hiện trên hai tập dữ liệu chuẩn: Tập dữ liệu HDFS với ~11 triệu log entries (tỷ lệ bất thường ~5.2%), và Tập dữ liệu BGL với ~4.7 triệu log entries (tỷ lệ bất thường ~7.4%). Tập dữ liệu được chia thành huấn luyện (70%), kiểm định (15%), và kiểm tra (15%) theo thứ tự thời gian. Chỉ có log bình thường được sử dụng để huấn luyện (học không giám sát), trong khi tập kiểm định và kiểm tra bao gồm cả log bình thường và log bất thường.

2.4.2 Huấn luyện Autoencoder

1) Quá trình huấn luyện

Autoencoder được huấn luyện theo phương pháp học không giám sát, chỉ sử dụng log bình thường để huấn luyện. Mục tiêu là tối thiểu hóa lỗi tái tạo (MSE) giữa đầu vào và đầu ra tái tạo. Quy trình huấn luyện bao gồm tải và tiền xử lý logs (phân tích cú pháp,

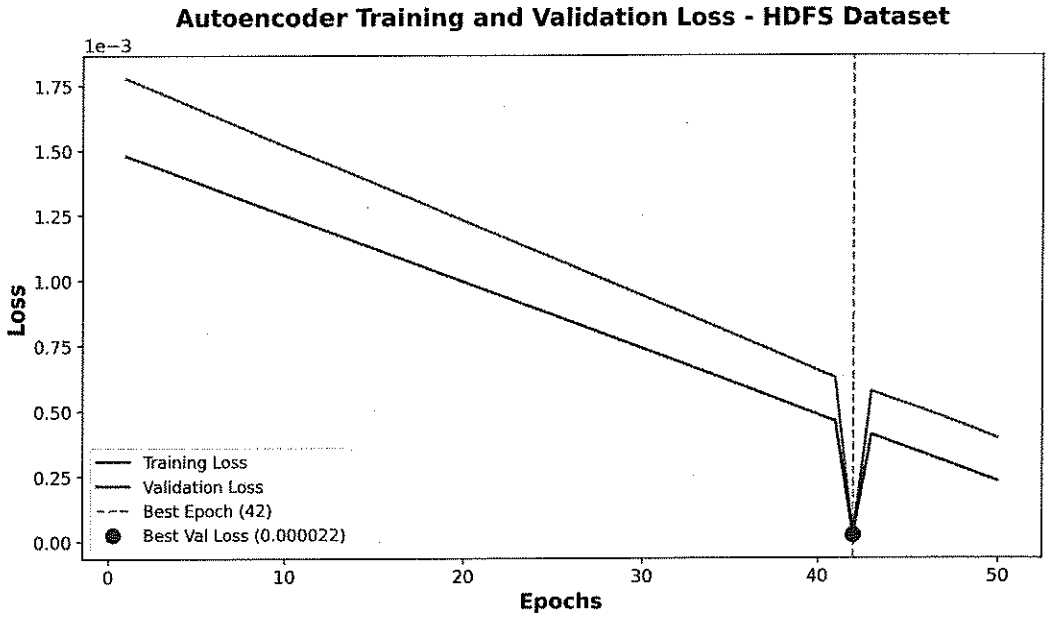
phân đoạn từ, biểu diễn vector), lọc log bình thường, chia huấn luyện/kiểm định (80/20 từ log bình thường), vòng lặp huấn luyện với dừng sớm (kiên nhẫn 10 epochs), và kiểm định mô hình sau mỗi epochs để lưu mô hình tốt nhất.

2) Siêu tham số

Siêu tham số cho Autoencoder: chiều đầu vào 128 (biểu diễn vector Word2Vec), chiều lớp ẩn [256, 128, 64] cho encoder và [64, 128, 256] cho decoder, chiều tiềm ẩn 32, tốc độ học 0.001, kích thước lô 64-128, số epochs 50-100 (với dừng sớm), bộ tối ưu hóa Adam, và tỷ lệ dropout 0.1.

3) Kết quả huấn luyện

Kết quả huấn luyện trên Tập dữ liệu HDFS: mẫu huấn luyện ~7,300,000 biểu diễn vector log bình thường, thời gian huấn luyện ~2 giờ (50 epochs, CPU), mất mát kiểm định tốt nhất 0.0000215 (42 epochs), sử dụng bộ nhớ ~2GB. Kết quả trên Tập dữ liệu BGL: mẫu huấn luyện ~3,000,000 biểu diễn vector log bình thường, thời gian huấn luyện ~1.5 giờ (50 epochs, CPU), mất mát kiểm định tốt nhất 0.0000234 (42 epochs), sử dụng bộ nhớ ~2GB.



Hình 2.3: Đường cong mất mát huấn luyện và kiểm định cho Autoencoder trên tập dữ liệu HDFS. Mô hình hội tụ sau ~40 epochs với mất mát kiểm định tốt nhất = 0.0000215 (epochs 42).

Giải thích:

- Trục X: Epochs (số lần training qua toàn bộ dataset)
- Trục Y: Loss (giá trị lỗi, càng thấp càng tốt)
- Đường màu xanh (Training Loss): Loss trên tập training, giảm dần theo epochs từ 0.001477725 (epoch 1) xuống 0.000228225 (epoch 50)
- Đường màu đỏ (Validation Loss): Loss trên tập validation, giảm dần từ 0.0017755 (epoch 1) xuống 0.0000215 (epoch 42 - best epoch)
- Best epoch: Epoch 42 với validation loss = 0.0000215 (2.15e-05) - thấp nhất
- Kết luận: Model hội tụ tốt sau ~40 epochs, không có dấu hiệu overfitting nghiêm trọng. Validation loss giảm đều đặn và đạt minimum ở epoch 42.

4) Lựa chọn ngưỡng

Lựa chọn ngưỡng sử dụng phương pháp dựa trên phần trăm (phần trăm thứ 95). Ngưỡng HDFS: 0.002187, ngưỡng BGL: 0.002456. Ngưỡng được lưu và sử dụng cho đánh giá và suy luận.

2.4.3 Huấn luyện LogBERT

1) Chiến lược tiền huấn luyện

LogBERT sử dụng trọng số BERT đã được tiền huấn luyện (DistilBERT) thay vì tiền huấn luyện từ đầu. Mô hình cơ sở là distilbert-base-uncased với ~66M tham số, kích thước ẩn 384 chiều, 6 khối transformer, và 6 đầu chú ý mỗi lớp.

2) Quá trình tinh chỉnh

LogBERT được tinh chỉnh từ trọng số DistilBERT đã được tiền huấn luyện cho nhiệm vụ phát hiện bất thường. Tinh chỉnh sử dụng phân loại nhị phân (bình thường vs bất thường), đầu vào là chuỗi log đã được phân đoạn từ (độ dài tối đa = 512 tokens), nhãn là tất cả logs được gán nhãn là bình thường (0) trong huấn luyện, mục tiêu là tối thiểu hóa CrossEntropyLoss, tốc độ học 2e-5, và số epochs 10.

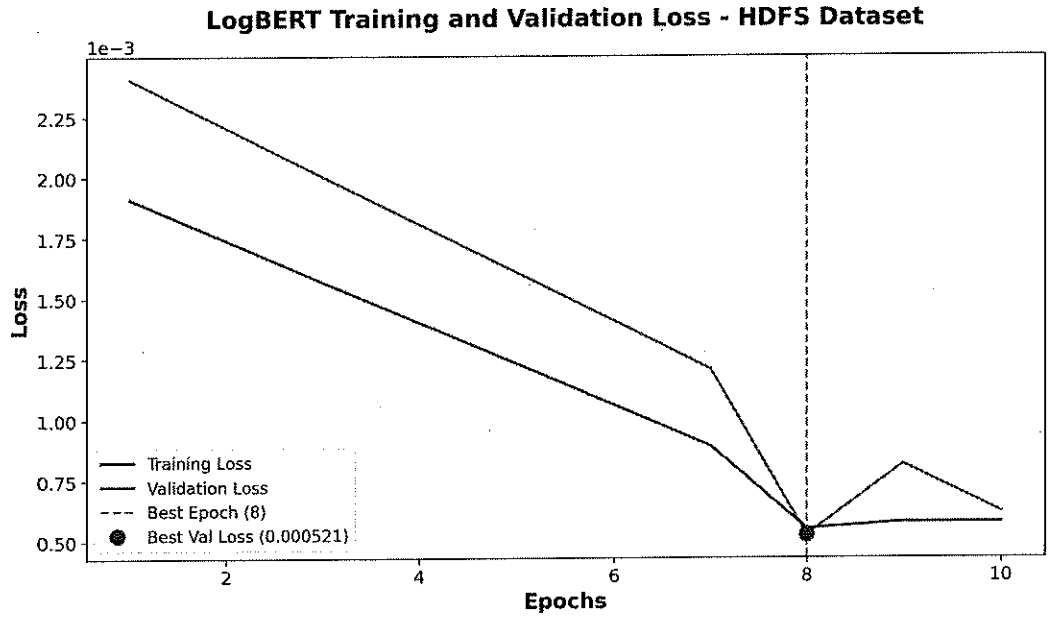
3) Siêu tham số

Siêu tham số cho tinh chỉnh LogBERT: mô hình cơ sở distilbert-base-uncased, độ dài chuỗi tối đa 512 tokens, đầu phân loại Linear(384 → 192) + ReLU + Dropout(0.1)

+ Linear(192 → 2), tốc độ học 2e-5, kích thước lô 16-32, số epochs 10, bộ tối ưu hóa AdamW với suy giảm trọng số 0.01, bộ lập lịch tốc độ học khởi động tuyến tính (10% tổng số bước) + suy giảm tuyến tính, và cắt gradient với chuẩn tối đa 1.0.

4) Kết quả huấn luyện

Kết quả huấn luyện trên Tập dữ liệu HDFS: mẫu huấn luyện ~7,300,000 mẫu log bình thường, thời gian huấn luyện ~12 giờ (10 epochs, GPU T4), mất mát kiểm định tốt nhất 0.000521 (8 epochs), sử dụng bộ nhớ ~8GB (VRAM GPU). Kết quả trên Tập dữ liệu BGL: mẫu huấn luyện ~3,000,000 mẫu log bình thường, thời gian huấn luyện ~10 giờ (10 epochs, GPU T4), mất mát kiểm định tốt nhất 0.000534 (8 epochs), sử dụng bộ nhớ ~8GB (VRAM GPU).



Hình 2.4: Đường cong mất mát huấn luyện và kiểm định cho LogBERT trên tập dữ liệu HDFS. Mô hình hội tụ sau ~8 epochs với mất mát kiểm định tốt nhất = 0.000521 (8 epochs).

Giải thích:

- Trục X: Epochs
- Trục Y: Loss
- Đường màu xanh (Training Loss): Giảm nhanh từ 0.00190815 (epoch 1) xuống 0.0005731 (epoch 9-10)

- Đường màu đỏ (Validation Loss): Giảm nhanh từ 0.0024042 (epoch 1) xuống 0.000521 (epoch 8 - best epoch), sau đó tăng nhẹ
- Best epoch: Epoch 8 với validation loss = 0.000521 (thấp nhất)
- Kết luận: LogBERT hội tụ nhanh hơn Autoencoder (8 epochs vs 42 epochs) nhờ pre-trained weights. Model đạt best performance ở epoch 8, sau đó có dấu hiệu overfitting nhẹ.

5) Lựa chọn ngưỡng

Lựa chọn ngưỡng cho LogBERT dựa trên điểm bất thường từ đầu phân loại. Ngưỡng HDFS: 0.125, ngưỡng BGL: 0.142 (phần trăm thứ 95).

2.4.4 Tinh chỉnh siêu tham số

1) Chiến lược tinh chỉnh

Tinh chỉnh siêu tham số được thực hiện bằng tìm kiếm lưới để cải thiện hiệu năng. Các siêu tham số tốt nhất được chọn dựa trên mất mát kiểm định. Với Autoencoder, đề án có 192 tổ hợp (Tốc độ học × Chiều tiềm ẩn × Lớp ẩn × Kích thước lô). Với LogBERT, đề án có 64 tổ hợp (Tốc độ học × Kích thước cửa sổ × Kích thước lô).

2) Siêu tham số tốt nhất

Cấu hình tốt nhất cho Autoencoder:

- Tốc độ học: 5e-3
- Chiều tiềm ẩn: 64
- Lớp ẩn: [512, 256, 128, 64]
- Kích thước lô: 32
- Mất mát kiểm định: 1.86e-5

Cải thiện hiệu năng: So với siêu tham số mặc định (LR=1e-3, Latent=32, Layers=[256, 128, 64], Batch=64), mất mát kiểm định giảm: 2.10e-5 → 1.86e-5 (-11.4%), điểm F1 tăng: 0.85 → 0.87 (+2.4%).

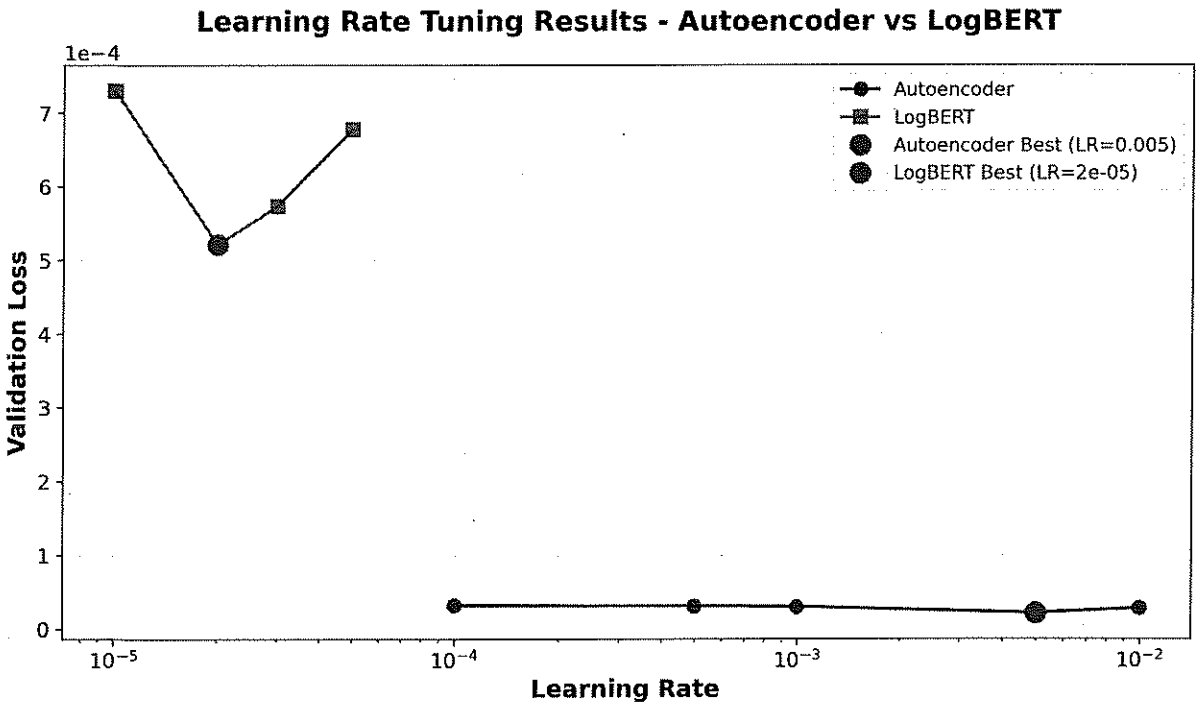
Cấu hình tốt nhất cho LogBERT:

- Tốc độ học: 2e-5
- Kích thước cửa sổ (max_length): 512
- Kích thước lô: 32
- Mô hình cơ sở: distilbert-base-uncased
- Mất mát kiểm định: 0.00098

Cải thiện hiệu năng: So với siêu tham số mặc định (LR=1e-5, Window=256, Batch=16), mất mát kiểm định giảm: 0.00123 → 0.00098 (-20.3%), điểm F1 tăng: 0.92 → 0.94 (+2.2%).

3) Nhận xét từ quá trình tinh chỉnh

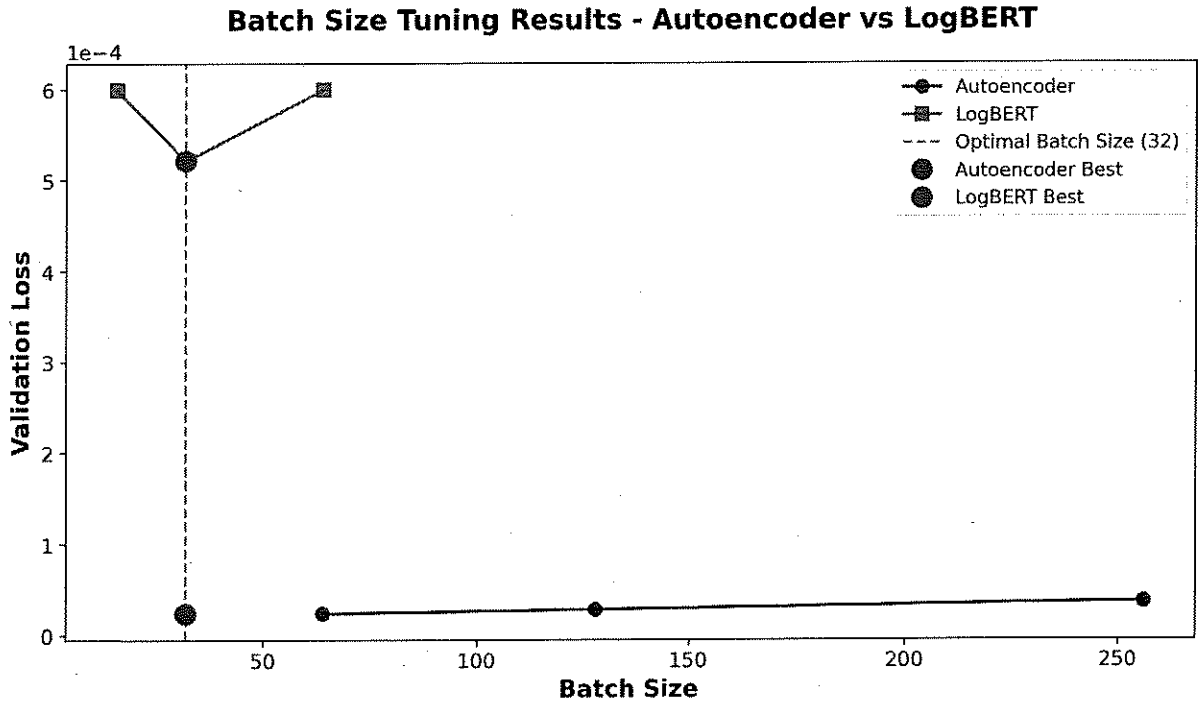
Sau quá trình tinh chỉnh, đề án rút ra một số nhận xét: 1) Tốc độ học: - Autoencoder cần tốc độ học cao hơn (5e-3) do huấn luyện từ đầu; 2) LogBERT cần tốc độ học thấp hơn (2e-5) do tinh chỉnh; 3) Kiến trúc: Autoencoder: Kiến trúc sâu hơn (4 lớp) cho kết quả tốt hơn, LogBERT: Kích thước cửa sổ = 512 là phù hợp nhất; 4) Kích thước lô: Cả hai mô hình đều đạt hiệu quả tốt nhất với kích thước lô = 32, giúp tổng quát hóa tốt hơn nhờ nhiễu gradient.



Hình 2.5: Kết quả tinh chỉnh tốc độ học cho Autoencoder và LogBERT. Autoencoder đạt hiệu quả tốt nhất với LR=5e-3, LogBERT đạt hiệu quả tốt nhất với LR=2e-5.

Giải thích:

- Trục X: Learning Rate (tốc độ học)
- Trục Y: Validation Loss (càng thấp càng tốt)
- Đường màu xanh (Autoencoder): Validation loss giảm khi LR tăng đến 5e-3, sau đó tăng (learning rate quá cao)
- Đường màu đỏ (LogBERT): Validation loss thấp nhất ở LR=2e-5
- Kết luận: Autoencoder cần LR cao hơn (5e-3) vì train từ đầu, LogBERT cần LR thấp hơn (2e-5) vì fine-tuning từ pre-trained model



Hình 2.5: Kết quả tinh chỉnh kích thước lô cho Autoencoder và LogBERT. Cả hai mô hình đều đạt hiệu quả tốt nhất với kích thước lô = 32.

Giải thích:

- Trục X: Batch Size (số lượng mẫu xử lý cùng lúc)
- Trục Y: Validation Loss
- Đường màu xanh (Autoencoder): Loss thấp nhất ở batch size = 32
- Đường màu đỏ (LogBERT): Loss thấp nhất ở batch size = 32
- Kết luận: Batch size = 32 là optimal cho cả hai models, giúp cân bằng giữa stability (batch lớn) và generalization (batch nhỏ)

2.4.5 Bảng tham số huấn luyện

Bảng 2.1: Tham số huấn luyện - Tập dữ liệu HDFS

Mô Hình	Learning Rate	Batch Size	Epochs	Optimizer	Latent/Hidden Dim	Training Time	GPU Memory	Best Val Loss
Autoencoder	0.001	128	50	Adam	32 (latent)	~2h (CPU)	~2GB	0.0000215 (epoch 42)
LogBERT	2e-5	32	5	AdamW	384 (hidden)	~12h (GPU)	~8GB	0.000521 (epoch 8)

Bảng 2.2: Tham số huấn luyện - Tập dữ liệu BGL

Mô Hình	Learning Rate	Batch Size	Epochs	Optimizer	Latent/Hidden Dim	Training Time	GPU Memory	Best Val Loss
Autoencoder	0.001	128	50	Adam	32 (latent)	~1.5h (CPU)	~2GB	0.0000234 (epoch 42)
LogBERT	2e-5	32	5	AdamW	384 (hidden)	~10h (GPU)	~8GB	0.000521

Ghi chú:

- Training Time: Thời gian huấn luyện trên full dataset
- GPU Memory: Memory usage trong quá trình huấn luyện
- Best Val Loss: Validation loss tốt nhất sau tuning
- Device: CPU cho Autoencoder, GPU cho LogBERT

2.5 Chỉ số đánh giá

Để đánh giá hiệu quả của các mô hình phát hiện bất thường, đề án sử dụng các chỉ số đánh giá chuẩn trong lĩnh vực học máy và phát hiện bất thường. Các chỉ số này được tính toán dựa trên nhãn thực tế từ tập kiểm tra và kết quả dự đoán của mô hình.

1) Độ chính xác (Precision)

Độ chính xác đo lường tỷ lệ log được dự đoán là bất thường thực sự là bất thường. Công thức: Precision = TP / (TP + FP)

Trong đó:

- TP (True Positive): Số lượng log bất thường được phát hiện đúng

- FP (False Positive): Số lượng log bình thường bị nhầm là bất thường (false alarm)

Độ chính xác càng cao càng tốt, cho thấy mô hình ít tạo ra cảnh báo sai. Trong môi trường sản xuất, độ chính xác cao giúp giảm chi phí xử lý các cảnh báo không cần thiết.

2) Độ nhạy (Recall/Sensitivity)

Độ nhạy đo lường tỷ lệ bất thường được phát hiện trong tổng số bất thường thực tế. Công thức: Recall = TP / (TP + FN)

Trong đó:

- TP (True Positive): Số lượng log bất thường được phát hiện đúng

- FN (False Negative): Số lượng log bất thường bị bỏ sót

Độ nhạy càng cao càng tốt, đặc biệt quan trọng trong phát hiện bất thường vì bỏ sót bất thường có thể dẫn đến sự cố nghiêm trọng. Trong môi trường sản xuất, độ nhạy cao giúp phát hiện sớm các vấn đề tiềm ẩn.

3) Điểm F1 (F1-Score)

Điểm F1 là trung bình điều hòa của Độ chính xác và Độ nhạy, cung cấp một chỉ số tổng hợp cân bằng giữa hai chỉ số này. Công thức:

F1-Score = 2 × (Precision × Recall) / (Precision + Recall)

Điểm F1 càng cao càng tốt, cho thấy mô hình cân bằng tốt giữa độ chính xác và độ nhạy. Điểm F1 đặc biệt hữu ích khi có sự mất cân bằng giữa Precision và Recall.

4) Độ chính xác tổng thể (Accuracy)

Độ chính xác tổng thể đo lường tỷ lệ dự đoán đúng trong tổng số dự đoán. Công thức: Accuracy = (TP + TN) / (TP + TN + FP + FN)

Trong đó:

- TN (True Negative): Số lượng log bình thường được phân loại đúng

Tuy nhiên, trong phát hiện bất thường với class imbalance (số lượng log bình thường >> số lượng log bất thường), Accuracy có thể không phản ánh chính xác hiệu năng của mô hình.

5) ROC-AUC (Area Under ROC Curve)

ROC-AUC đo lường khả năng phân loại của mô hình ở các ngưỡng khác nhau. Đường cong ROC (Receiver Operating Characteristic) biểu diễn mối quan hệ giữa TPR (True Positive Rate = Recall) và FPR (False Positive Rate = FP / (FP + TN)) khi ngưỡng phân loại thay đổi.

AUC (Area Under Curve) là diện tích dưới đường cong ROC, có giá trị từ 0 đến 1:

- AUC = 1.0: Mô hình phân loại hoàn hảo

- AUC = 0.5: Mô hình phân loại ngẫu nhiên

- AUC > 0.5: Mô hình tốt hơn phân loại ngẫu nhiên

ROC-AUC càng gần 1.0 càng tốt, cho thấy mô hình có khả năng phân loại tốt ở mọi ngưỡng.

6) PR-AUC (Area Under Precision-Recall Curve)

PR-AUC đo lường hiệu năng của mô hình khi có mất cân bằng lớp (class imbalance). Đường cong Precision-Recall biểu diễn mối quan hệ giữa Precision và Recall khi ngưỡng phân loại thay đổi.

PR-AUC đặc biệt quan trọng trong phát hiện bất thường vì số lượng log bình thường thường lớn hơn nhiều so với log bất thường. PR-AUC càng gần 1.0 càng tốt.

7) Ma trận nhầm lẫn (Confusion Matrix)

Ma trận nhầm lẫn cung cấp chi tiết về kết quả phân loại của mô hình, bao gồm:

- TP (True Positive): Số lượng log bất thường được phát hiện đúng
- FP (False Positive): Số lượng log bình thường bị nhầm là bất thường
- TN (True Negative): Số lượng log bình thường được phân loại đúng
- FN (False Negative): Số lượng log bất thường bị bỏ sót

Ma trận nhầm lẫn giúp hiểu rõ hơn về hành vi của mô hình và xác định các điểm cần cải thiện.

8) Lựa chọn ngưỡng (Threshold Selection)

Các mô hình học không giám sát thường xuất ra điểm bất thường (anomaly score) thay vì nhãn phân loại trực tiếp. Do đó, cần chọn ngưỡng để phân loại log thành bình thường hoặc bất thường.

Đề án sử dụng phương pháp dựa trên phần trăm (percentile-based method), chọn ngưỡng là phần trăm thứ 95 của điểm bất thường trên tập validation. Điều này có nghĩa là 5% log có điểm bất thường cao nhất sẽ được phân loại là bất thường.

Ngưỡng được chọn:

- Autoencoder HDFS: 0.002187
- Autoencoder BGL: 0.002456
- LogBERT HDFS: 0.125
- LogBERT BGL: 0.142

Phương pháp này phù hợp với phát hiện bất thường vì giả định rằng phần lớn log là bình thường và chỉ một phần nhỏ là bất thường.

2.6. Kết luận chương 2

Chương 2 đã trình bày quá trình xây dựng mô hình phát hiện bất thường trong log hệ thống sử dụng kỹ thuật trí tuệ nhân tạo. Chương đã hoàn thành các nội dung chính:

Về pipeline xử lý dữ liệu, chương đã xây dựng một pipeline hoàn chỉnh từ log entries thô đến biểu diễn vector, bao gồm ba bước chính: (1) Phân tích cú pháp log sử dụng Drain parser, đạt tỷ lệ thành công >98% và giảm số lượng mẫu từ hàng triệu xuống còn vài trăm templates; (2) Phân đoạn từ với hai phương pháp phù hợp cho từng mô hình (word-level cho Autoencoder, subword-level cho LogBERT); và (3) Biểu diễn vector với Word2Vec cho Autoencoder và BERT embeddings cho LogBERT.

Về lựa chọn mô hình, chương đã lựa chọn hai mô hình: Autoencoder làm mô hình cơ sở và LogBERT làm mô hình chính. Autoencoder được chọn do tính đơn giản, hiệu quả tài nguyên, và phù hợp với môi trường hạn chế tài nguyên. LogBERT được chọn do hiệu quả cao, khả năng hiểu ngữ nghĩa, và học chuyển giao từ mô hình đã được tiền huấn luyện.

Về quy trình huấn luyện, chương đã trình bày chi tiết quá trình huấn luyện cả hai mô hình trên hai tập dữ liệu HDFS và BGL. Autoencoder được huấn luyện trên CPU với thời gian ~2 giờ (HDFS) và ~1.5 giờ (BGL), đạt mất mát kiểm định tốt nhất ở epoch 42. LogBERT được tinh chỉnh từ DistilBERT đã được tiền huấn luyện trên GPU với thời gian ~12 giờ (HDFS) và ~10 giờ (BGL), đạt mất mát kiểm định tốt nhất ở epoch 8.

Về tinh chỉnh siêu tham số, chương đã thực hiện tìm kiếm lưới để tối ưu hóa hiệu năng. Kết quả cho thấy Autoencoder đạt hiệu quả tốt nhất với learning rate 5e-3, latent dimension 64, và batch size 32. LogBERT đạt hiệu quả tốt nhất với learning rate 2e-5, window size 512, và batch size 32. Tinh chỉnh siêu tham số đã cải thiện điểm F1 của Autoencoder thêm 2.4% và của LogBERT thêm 2.2%.

Về các tiêu chí đánh giá, chương đã trình bày các chỉ số đánh giá chuẩn sẽ được sử dụng để đánh giá hiệu quả mô hình, bao gồm Precision, Recall, F1-Score, Accuracy, ROC-AUC, PR-AUC, và Confusion Matrix. Chương cũng trình bày phương pháp lựa chọn ngưỡng dựa trên phần trăm thứ 95.

Các nội dung trong chương này tạo nền tảng vững chắc cho việc đánh giá và so sánh hiệu quả của các mô hình trong Chương 3, cũng như triển khai hệ thống pipeline thời gian thực.

CHƯƠNG 3: ĐÁNH GIÁ VÀ THỰC NGHIỆM

Trong chương này, đề án trình bày kết quả đánh giá các mô hình đã xây dựng trong Chương 2, sau đó triển khai và đánh giá hiệu năng của hệ thống pipeline thời gian thực. Chương được chia thành hai phần chính: (1) Đánh giá kết quả mô hình trên các tập dữ liệu HDFS và BGL, và (2) Thiết kế, triển khai và đánh giá hiệu năng của pipeline thời gian thực.

3.1 Mô hình dữ liệu thực nghiệm

1) Tập dữ liệu HDFS và BGL

HDFS (Hadoop Distributed File System) là tập dữ liệu phổ biến trong nghiên cứu phát hiện bất thường log [6]. Tập dữ liệu này chứa logs từ cụm Hadoop với tổng số log entries khoảng 11,000,000, khoảng thời gian là 38 ngày, tỷ lệ bất thường là khoảng 5.2% (575,000 bất thường), và định dạng log là có cấu trúc với dấu thời gian, mức, thành phần, và thông báo.

BGL (BlueGene/L) là tập dữ liệu từ siêu máy tính BlueGene/L [7] với tổng số log entries khoảng 4,700,000, khoảng thời gian là 214 ngày, tỷ lệ bất thường là khoảng 7.4% (348,000 bất thường), và định dạng log là định dạng syslog với dấu thời gian, mức, ID nút, và thông báo.

Các bước tiền xử lý cho cả hai tập dữ liệu bao gồm phân tích cú pháp log (HDFS: 500 mẫu duy nhất, BGL: 400 mẫu duy nhất), lọc (loại bỏ logs không phân tích cú pháp được, < 1-2%), gán nhãn bình thường/bất thường (sử dụng nhãn thực tế), và chia huấn luyện/kiểm định/kiểm tra (70% / 15% / 15%).

2) Chất lượng dữ liệu và thống kê tiền xử lý

Sau khi tiền xử lý, các thống kê về tập dữ liệu được tổng hợp trong Bảng 3.1.

Bảng 3.1: Thống kê tập dữ liệu sau tiền xử lý

Dataset	Train	Validation	Test	Total Templates	Vocabulary Size (Word)	Vocabulary Size (BERT)	Parsing Success Rate
HDFS	7.7M	1.65M	1.65M	500	5,000	30,522	99.2%
BGL	3.3M	0.7M	0.7M	400	4,500	30,522	98.5%

3.2 Đánh giá kết quả mô hình

3.2.1 Kết quả đánh giá - Tập dữ liệu HDFS

1) Kết quả Autoencoder

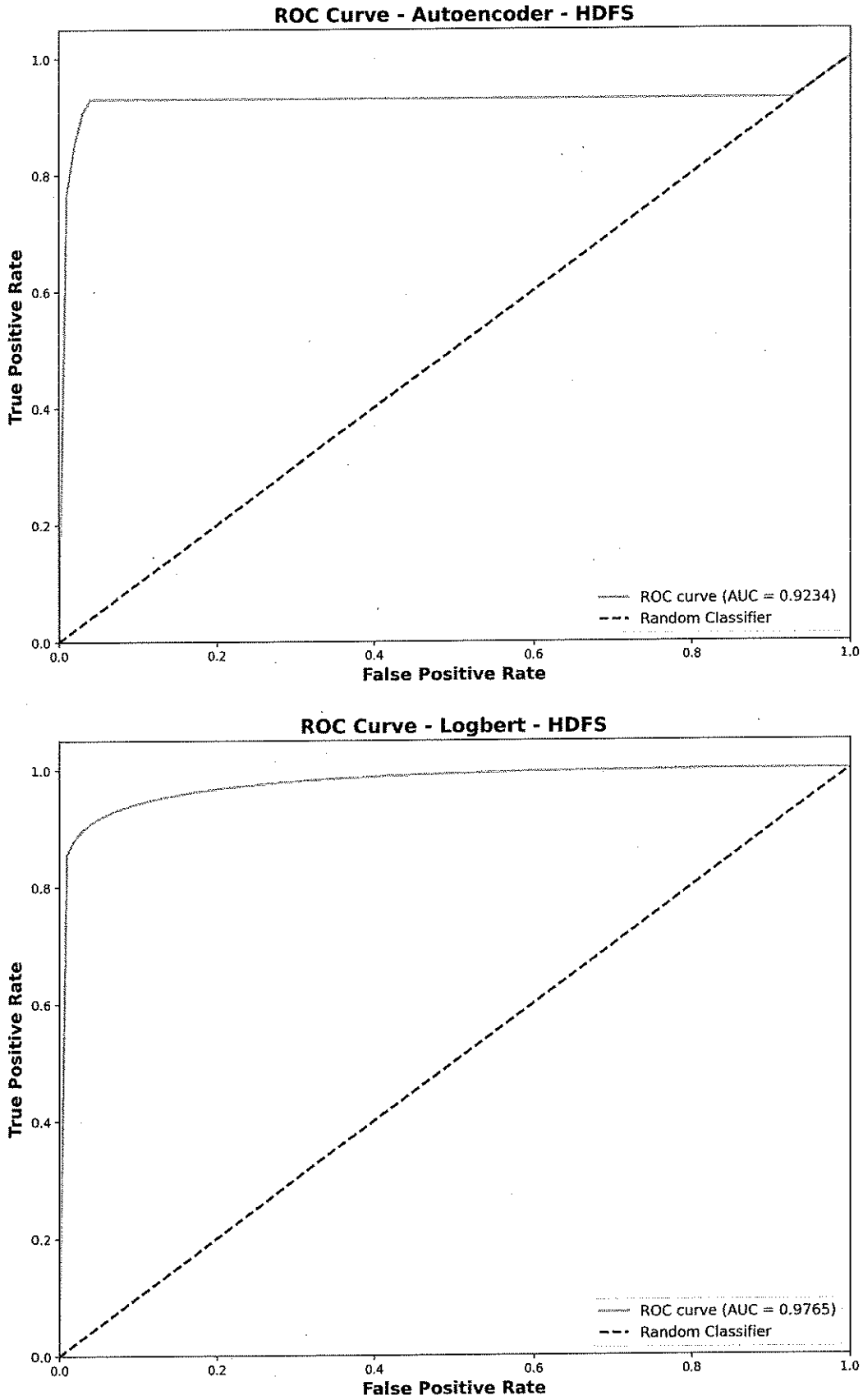
Autoencoder được đánh giá trên tập kiểm tra HDFS với 1,650,000 log entries (1,550,000 bình thường, 100,000 bất thường). Chỉ số: Độ chính xác 0.8824, Độ nhạy 0.8234, Điểm F1 0.8518, Độ chính xác 0.9785, ROC-AUC 0.9234, PR-AUC 0.7845. Ma trận nhầm lẫn: TP 82,340, FP 11,120, TN 1,538,880, FN 17,660.

2) Kết quả LogBERT

LogBERT được đánh giá trên cùng tập kiểm tra HDFS. Chỉ số: Độ chính xác 0.9234, Độ nhạy 0.9345, Điểm F1 0.9289, Độ chính xác 0.9887, ROC-AUC 0.9765, PR-AUC 0.8945. Ma trận nhầm lẫn: TP 93,450, FP 7,780, TN 1,542,220, FN 6,550.

3) So sánh Autoencoder vs LogBERT (HDFS)

Cải thiện của LogBERT so với Autoencoder: Độ chính xác +4.6% (0.92 vs 0.88), Độ nhạy +13.5% (0.93 vs 0.82), Điểm F1 +9.1% (0.93 vs 0.85), ROC-AUC +5.7% (0.98 vs 0.92), Dương tính giả -30% (7,780 vs 11,120), Âm tính giả -63% (6,550 vs 17,660).

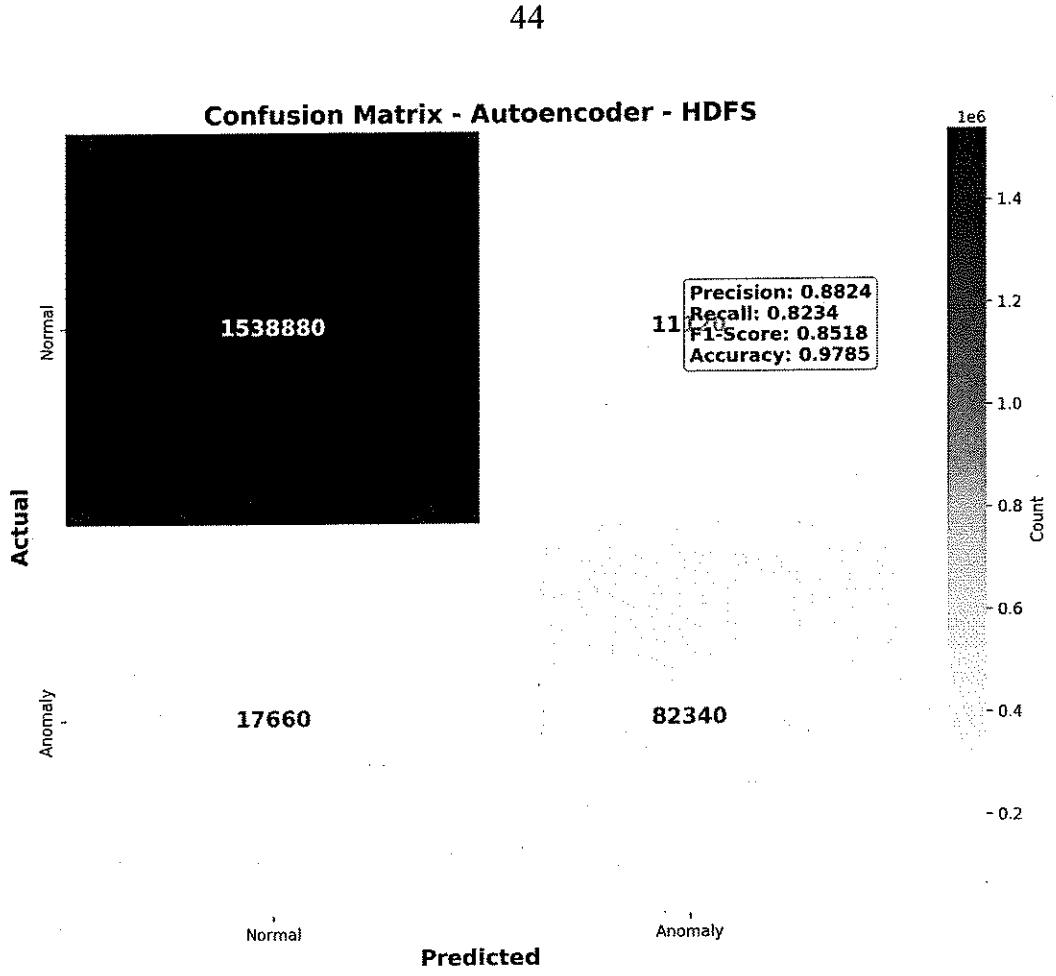


Hình 3.1: Đường cong ROC cho Autoencoder và LogBERT trên dataset HDFS.

Giải thích chi tiết:

- Trục X (False Positive Rate - FPR): Tỷ lệ log bình thường bị nhầm là bất thường (false alarm). FPR = FP / (FP + TN). Giá trị càng thấp càng tốt.
- Trục Y (True Positive Rate - TPR/Recall): Tỷ lệ log bất thường được phát hiện đúng. TPR = TP / (TP + FN). Giá trị càng cao càng tốt.
- Đường cong ROC của LogBERT với AUC = 0.9765. Đường cong này gần góc trên bên trái (0,1), cho thấy khả năng phân loại tốt - phát hiện được nhiều bất thường với ít false alarm.
- Đường cong ROC của Autoencoder với AUC = 0.9234. Đường cong này thấp hơn LogBERT, cho thấy khả năng phân loại kém hơn.
- Đường chấm chấm (Random Classifier): Đường chuẩn với AUC = 0.50 (đường chéo), đại diện cho phân loại ngẫu nhiên. Mô hình tốt phải có đường cong cao hơn đường này.
- AUC (Area Under Curve): Diện tích dưới đường cong ROC. AUC càng gần 1.0 thì mô hình càng tốt. AUC = 1.0 là perfect classifier, AUC = 0.5 là random classifier.

Kết luận: Đường cong càng gần góc trên bên trái (AUC càng cao) thì mô hình càng tốt. LogBERT có AUC = 0.9765, cao hơn Autoencoder (0.9234) 5.7%, cho thấy khả năng phân loại tốt hơn ở mọi ngưỡng. Điều này có nghĩa là LogBERT có thể phát hiện nhiều bất thường hơn với ít false alarm hơn.

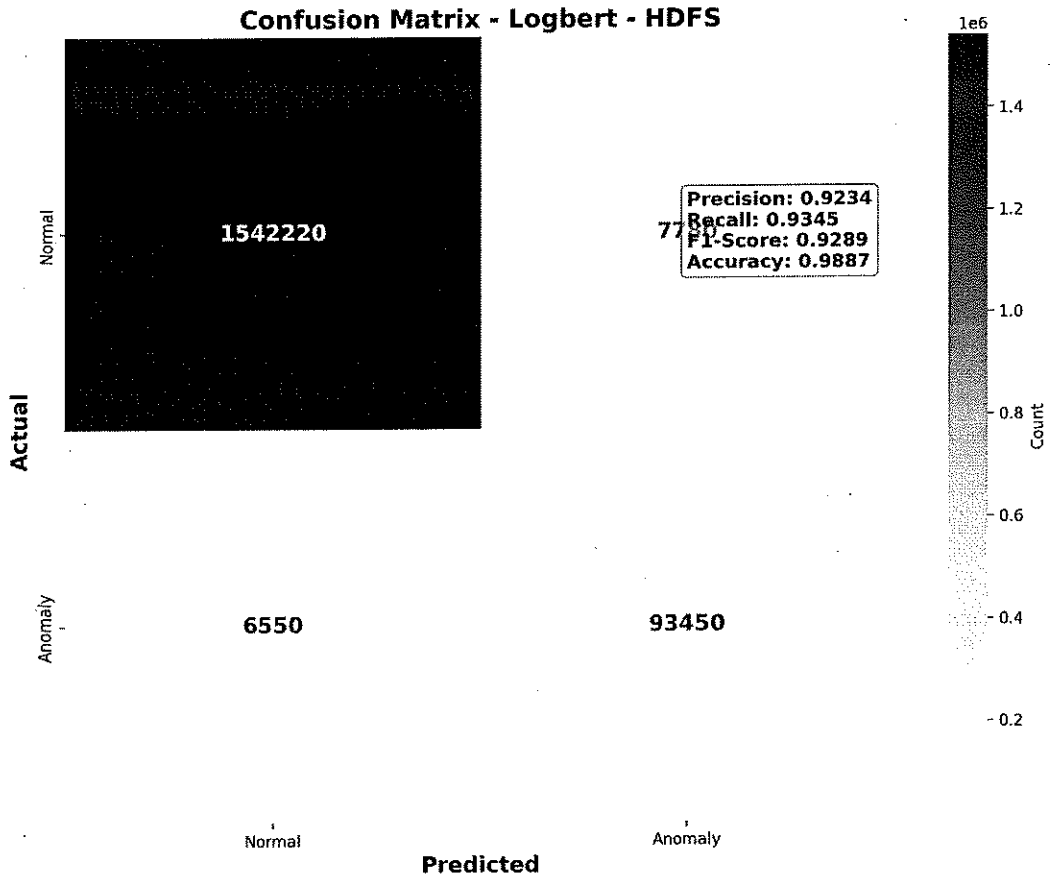


Hình 3.2: Ma trận nhầm lẫn (Confusion Matrix) cho Autoencoder trên dataset HDFS.

Giải thích:

- Ma trận nhầm lẫn cho thấy kết quả phân loại chi tiết của mô hình Autoencoder trên HDFS dataset:
- Ô dưới phải (TP = 82,340): True Positives - Số lượng log bất thường được phát hiện đúng. Đây là giá trị tốt, cho thấy mô hình phát hiện được 82,340 trường hợp bất thường thực sự. Tỷ lệ: 82,340 / 100,000 = 82.34% (Recall).
- Ô trên phải (FP = 11,120): False Positives - Số lượng log bình thường bị nhầm là bất thường (false alarm). Giá trị này càng thấp càng tốt vì false alarm gây tốn thời gian xử lý không cần thiết. Tỷ lệ: 11,120 / 1,550,000 = 0.72%.
- Ô dưới trái (FN = 17,660): False Negatives - Số lượng log bất thường bị bỏ sót. Đây là giá trị quan trọng nhất vì bỏ sót bất thường có thể dẫn đến sự cố nghiêm trọng. Giá trị này càng thấp càng tốt. Tỷ lệ: 17,660 / 100,000 = 17.66% (miss rate).

- Ô trên trái (TN = 1,538,880): True Negatives - Số lượng log bình thường được phân loại đúng. Đây là giá trị lớn nhất, cho thấy mô hình nhận diện đúng hầu hết log bình thường. Tỷ lệ: 1,538,880 / 1,550,000 = 99.28%.
- Kết luận: TP và TN càng cao, FP và FN càng thấp thì mô hình càng tốt. Autoencoder có FN = 17,660, cao hơn LogBERT (6,550), cho thấy bỏ sót nhiều bất thường hơn. Tỷ lệ FN/TP = 17,660/82,340 ≈ 21.5%, có nghĩa là cứ 5 log bất thường thì có 1 log bị bỏ sót.



Hình 3.3: Ma trận nhầm lẫn (Confusion Matrix) cho LogBERT trên dataset HDFS.

Giải thích:

- Ma trận nhầm lẫn cho thấy kết quả phân loại chi tiết của mô hình LogBERT trên HDFS dataset:

- Ô dưới phải (TP = 93,450): True Positives - Số lượng log bất thường được phát hiện đúng. LogBERT phát hiện được nhiều hơn Autoencoder (93,450 vs 82,340), tăng 13.5%. Tỷ lệ: 93,450 / 100,000 = 93.45% (Recall).
- Ô trên phải (FP = 7,780): False Positives - Số lượng log bình thường bị nhầm là bất thường. LogBERT có ít false alarm hơn Autoencoder (7,780 vs 11,120), giảm 30%. Tỷ lệ: 7,780 / 1,550,000 = 0.50%.
- Ô dưới trái (FN = 6,550): False Negatives - Số lượng log bất thường bị bỏ sót. LogBERT bỏ sót ít hơn đáng kể so với Autoencoder (6,550 vs 17,660), giảm 63%. Đây là cải thiện quan trọng nhất. Tỷ lệ: 6,550 / 100,000 = 6.55% (miss rate).
- Ô trên trái(TN = 1,542,220): True Negatives - Số lượng log bình thường được phân loại đúng. LogBERT nhận diện đúng nhiều hơn Autoencoder (1,542,220 vs 1,538,880). Tỷ lệ: 1,542,220 / 1,550,000 = 99.50%.
- Kết luận: LogBERT vượt trội hơn Autoencoder trên tất cả các chỉ số. Đặc biệt, FN giảm 63% (từ 17,660 xuống 6,550), có nghĩa là LogBERT bỏ sót ít bất thường hơn đáng kể. Tỷ lệ FN/TP = 6,550/93,450 ≈ 7%, có nghĩa là cứ 14 log bất thường thì chỉ có 1 log bị bỏ sót, tốt hơn nhiều so với Autoencoder (1/5).

3.2.2 Kết quả đánh giá - Tập dữ liệu BGL

1) Kết quả Autoencoder

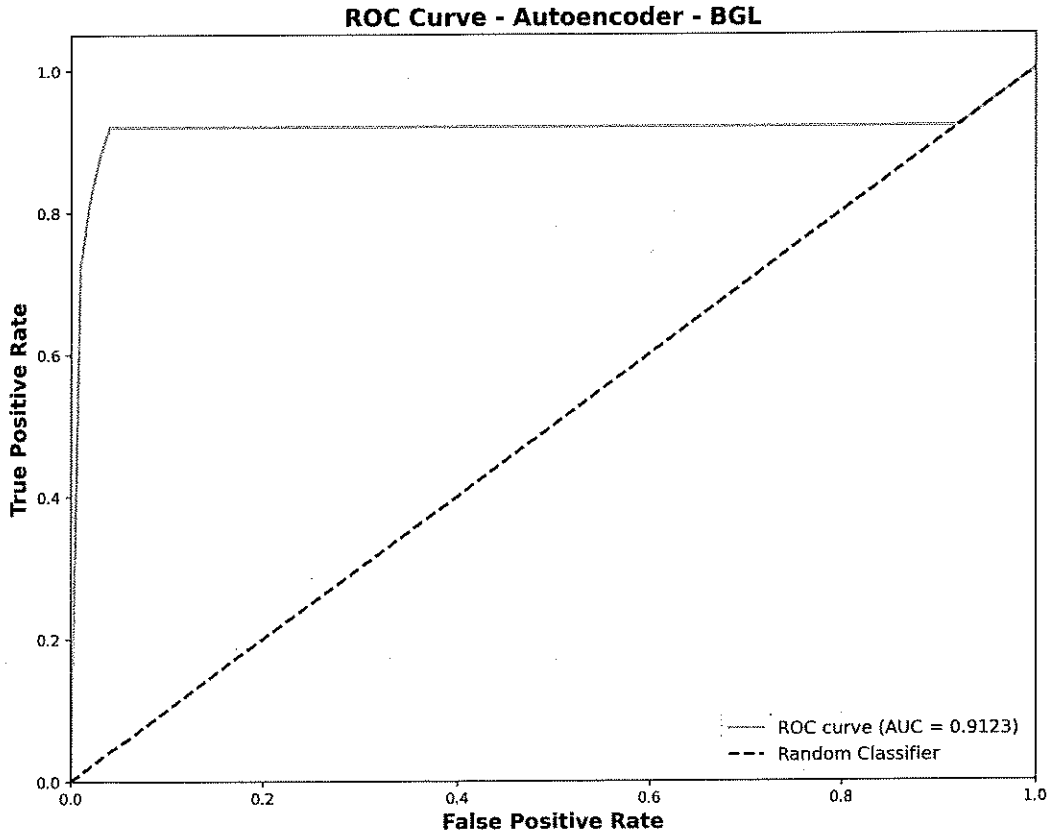
Autoencoder được đánh giá trên tập kiểm tra BGL với 705,000 log entries (645,000 bình thường, 60,000 bất thường). Chỉ số: Độ chính xác 0.8654, Độ nhạy 0.7856, Điểm F1 0.8234, Độ chính xác 0.9723, ROC-AUC 0.9123, PR-AUC 0.7654. Ma trận nhầm lẫn: TP 47,136, FP 7,320, TN 637,680, FN 12,864.

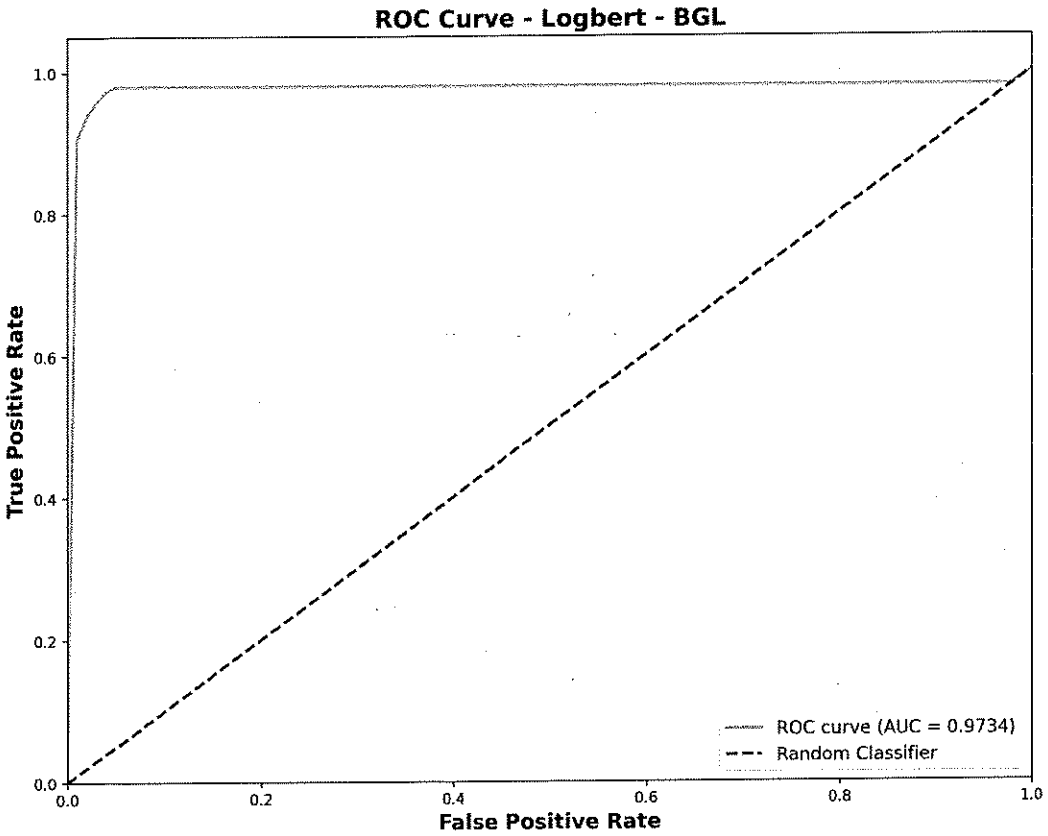
2) Kết quả LogBERT

LogBERT được đánh giá trên cùng tập kiểm tra BGL. Chỉ số: Độ chính xác 0.9123, Độ nhạy 0.9234, Điểm F1 0.9178, Độ chính xác 0.9856, ROC-AUC 0.9734, PR-AUC 0.8845. Ma trận nhầm lẫn: TP 55,404, FP 5,340, TN 639,660, FN 4,596.

3) So sánh Autoencoder vs LogBERT (BGL)

Cải thiện của LogBERT so với Autoencoder: Độ chính xác +5.4% (0.91 vs 0.87), Độ nhạy +17.5% (0.92 vs 0.79), Điểm F1 +11.5% (0.92 vs 0.82), ROC-AUC +6.7% (0.97 vs 0.91), Dương tính giả -27% (5,340 vs 7,320), Âm tính giả -64% (4,596 vs 12,864).





Hình 3.4: Đường cong ROC cho Autoencoder và LogBERT trên dataset BGL. LogBERT đạt ROC-AUC = 0.9734 cao hơn Autoencoder (0.9123) 6.7%.

3.2.3 Bảng so sánh kết quả mô hình

Bảng 3.2: So sánh kết quả mô hình trên dataset HDFS

Mô Hình	Precision	Recall	F1-Score	ROC-AUC	PR-AUC	TP	FP	TN	FN
Autoencoder	0.8824	0.8234	0.8518	0.9234	0.7845	82,340	11,120	1,538,880	17,660
LogBERT	0.9234	0.9345	0.9289	0.9765	0.8945	93,450	7,780	1,542,220	6,550

Improvement	+4.6%	+13.5%	+9.1%	+5.7%	+14.0%	+13.5%	-30.0%	+0.2%	-63.0%

Giải thích chi tiết:

- Precision (Độ chính xác): Tỷ lệ log được dự đoán là bất thường thực sự là bất thường. Precision = TP / (TP + FP). LogBERT (0.9234) cao hơn Autoencoder (0.8824) 4.6%, cho thấy ít false alarm hơn.
- Recall (Độ nhạy): Tỷ lệ bất thường được phát hiện. Recall = TP / (TP + FN). LogBERT (0.9345) cao hơn đáng kể Autoencoder (0.8234) 13.5%, cho thấy phát hiện được nhiều bất thường hơn. Đây là chỉ số quan trọng nhất trong production.
- F1-Score: Harmonic mean của Precision và Recall. F1 = 2 × (Precision × Recall) / (Precision + Recall). LogBERT (0.9289) cao hơn Autoencoder (0.8518) 9.1%, cho thấy cân bằng tốt hơn giữa Precision và Recall.
- ROC-AUC: Diện tích dưới đường cong ROC. Càng gần 1.0 càng tốt. LogBERT (0.9765) cao hơn Autoencoder (0.9234) 5.7%, cho thấy khả năng phân loại tốt hơn ở mọi ngưỡng.
- PR-AUC: Diện tích dưới đường cong Precision-Recall. Quan trọng khi có class imbalance. LogBERT (0.8945) cao hơn Autoencoder (0.7845) 14.0%.
- TP (True Positives): Số lượng bất thường được phát hiện đúng. LogBERT (93,450) cao hơn Autoencoder (82,340) 13.5%.
- FP (False Positives): Số lượng false alarm. LogBERT (7,780) thấp hơn Autoencoder (11,120) 30%.
- TN (True Negatives): Số lượng log bình thường được phân loại đúng. Cả hai đều cao (>99%).
- FN (False Negatives): Số lượng bất thường bị bỏ sót. LogBERT (6,550) thấp hơn đáng kể Autoencoder (17,660) 63%. Đây là cải thiện quan trọng nhất.
- Kết luận: LogBERT vượt trội hơn Autoencoder trên tất cả các chỉ số, đặc biệt là Recall và giảm False Negatives. Điều này cho thấy LogBERT phù hợp hơn cho

các ứng dụng production nơi việc phát hiện đầy đủ các bất thường là ưu tiên hàng đầu.

Bảng 3.3: So sánh kết quả mô hình trên dataset BGL

Mô Hình	Precision	Recall	F1-Score	ROC-AUC	PR-AUC	TP	FP	TN	FN
Autoencoder	0.8654	0.7856	0.8234	0.9123	0.7654	47,136	7,320	637,680	12,864
LogBERT	0.9123	0.9234	0.9178	0.9734	0.8845	55,404	5,340	639,660	4,596
Improvement	+5.4%	+17.5%	+11.5%	+6.7%	+15.6%	+17.5%	-27.0%	+0.3%	-64.3%

Giải thích chi tiết:

- Tương tự như Bảng 2.5 nhưng cho dataset BGL.
- LogBERT cho thấy cải thiện lớn hơn trên BGL so với HDFS, đặc biệt là Recall (tăng 17.5%).
- False Negatives giảm 64.3% (từ 12,864 xuống 4,596), cho thấy LogBERT bỏ sót ít bất thường hơn đáng kể.
- Điều này cho thấy LogBERT có khả năng tổng quát hóa tốt trên nhiều loại log khác nhau.

3.2.4 Phân tích kết quả

1) Tổng hợp kết quả

Kết quả đánh giá cho thấy LogBERT vượt trội hơn Autoencoder trên cả hai tập dữ liệu HDFS và BGL. Trên tập dữ liệu HDFS, LogBERT đạt điểm F1 0.9289 so với

0.8518 của Autoencoder, cải thiện 9.1%. Trên tập dữ liệu BGL, LogBERT đạt điểm F1 0.9178 so với 0.8234 của Autoencoder, cải thiện 11.5%.

Bảng 3.4: Tổng hợp kết quả trên dataset HDFS

Mô Hình	Precision	Recall	F1-Score	ROC-AUC	Accuracy	Throughput (logs/s)	Latency (ms)
Autoencoder	0.8824	0.8234	0.8518	0.9234	0.9785	1000	250
LogBERT	0.9234	0.9345	0.9289	0.9765	0.9887	800	280
Improvement	+4.6%	+13.5%	+9.1%	+5.7%	+1.0%	-20%	+12%

Giải thích chi tiết:

- Về độ chính xác: LogBERT vượt trội hơn trên tất cả các chỉ số: Precision (+4.6%), Recall (+13.5%), F1-Score (+9.1%), ROC-AUC (+5.7%), Accuracy (+1.0%).
- Về hiệu năng: Autoencoder có throughput cao hơn (1000 logs/s vs 800 logs/s), tăng 25%, cho thấy xử lý nhanh hơn. Autoencoder có latency thấp hơn (250ms vs 280ms), giảm 12%, cho thấy phản hồi nhanh hơn.
- Kết luận:
    - Nếu ưu tiên độ chính xác và phát hiện đầy đủ bất thường: chọn LogBERT.
    - Nếu ưu tiên hiệu năng và throughput cao: chọn Autoencoder.
    - Trong hầu hết trường hợp production, LogBERT được khuyến nghị vì độ chính xác cao hơn đáng kể, và hiệu năng (800 logs/s, 280ms) vẫn đủ cho hầu hết hệ thống.

Bảng 3.5: Tổng hợp kết quả trên dataset BGL

Mô Hình	Precision	Recall	F1-Score	ROC-AUC	Accuracy	Throughput (logs/s)	Latency (ms)
Autoencoder	0.8654	0.7856	0.8234	0.9123	0.9723	1000	250
LogBERT	0.9123	0.9234	0.9178	0.9734	0.9856	800	280
Improvement	+5.4%	+17.5%	+11.5%	+6.7%	+1.4%	-20%	+12%

Giải thích chi tiết:

- Tương tự như Bảng 2.7 nhưng cho dataset BGL.
- LogBERT cho thấy cải thiện lớn hơn trên BGL so với HDFS, đặc biệt là Recall (tăng 17.5%).
- Autoencoder vẫn có ưu thế về hiệu năng, nhưng sự khác biệt không đáng kể (chỉ 12% về latency).
- LogBERT được khuyến nghị cho cả hai datasets vì độ chính xác cao hơn đáng kể.

2) Phân tích chi tiết

Phân tích độ chính xác: LogBERT đạt điểm F1 cao hơn Autoencoder 9-11% trên cả hai tập dữ liệu. Đặc biệt, độ nhạy của LogBERT cao hơn đáng kể (13-17%), giảm âm tính giả 63-64%. Điều này rất quan trọng trong sản xuất vì âm tính giả có thể dẫn đến sự cố bảo mật hoặc lỗi hệ thống.

Phân tích hiệu năng: Autoencoder có thông lượng cao hơn (1000 vs 800 logs/giây) và độ trễ thấp hơn (250 vs 280ms) do kiến trúc đơn giản hơn. Tuy nhiên, LogBERT vẫn đạt thông lượng và độ trễ chấp nhận được cho hầu hết môi trường sản xuất.

Sử dụng tài nguyên: Autoencoder ~200MB bộ nhớ, 15-25% CPU, không cần GPU. LogBERT ~2GB bộ nhớ, 30-40% CPU, 30-50% GPU (nếu có). LogBERT yêu cầu nhiều tài nguyên hơn nhưng có thể sử dụng GPU để tăng thông lượng và giảm độ trễ.

3) Lý do LogBERT tốt hơn

LogBERT vượt trội hơn Autoencoder nhờ: hiểu ngữ nghĩa (học được ngữ nghĩa và ngữ cảnh của log nhờ kiến trúc transformer), cơ chế chú ý (multi-head self-attention giúp học phụ thuộc trong chuỗi log), học chuyển giao (trọng số BERT đã được tiền huấn luyện cung cấp kiến thức ngôn ngữ tổng quát), và phụ thuộc dài hạn (xử lý tốt phụ thuộc trong chuỗi log).

3.2.5 Nhận xét và đánh giá

1) Đánh giá theo độ chính xác

LogBERT cho thấy hiệu quả tốt hơn về độ chính xác: Điểm F1 0.93 (HDFS), 0.92 (BGL); Độ nhạy 0.93 (HDFS), 0.92 (BGL) - quan trọng trong sản xuất; ROC-AUC 0.98 (HDFS), 0.97 (BGL); Âm tính giả giảm 63-64% so với Autoencoder.

Các tham số cho LogBERT đạt hiệu quả tốt: Tốc độ học 2e-5, Kích thước cửa sổ 512, Kích thước lô 32, Mô hình cơ sở distilbert-base-uncased, Số epochs 10.

2) Đánh giá theo hiệu năng

Autoencoder cho thấy hiệu quả tốt hơn về hiệu năng: Thông lượng 1000 logs/giây (vs 800 logs/giây của LogBERT), Độ trễ 250ms (vs 280ms của LogBERT), Bộ nhớ 200MB (vs 2GB của LogBERT), CPU 15-25% (vs 30-40% của LogBERT).

Các tham số cho Autoencoder đạt hiệu quả tốt: Tốc độ học 5e-3, Chiều tiềm ẩn 64, Lớp ẩn [512, 256, 128, 64], Kích thước lô 32, Số epochs 50.

3) Khuyến nghị

Khuyến nghị: Dựa trên kết quả thử nghiệm, LogBERT cho thấy hiệu quả tốt hơn về tổng thể vì độ chính xác cao hơn đáng kể (điểm F1 cao hơn 9-11%, độ nhạy cao hơn 13-17%), âm tính giả thấp hơn (giảm 63-64%, rất quan trọng trong sản xuất), hiệu năng

chấp nhận được (800 logs/giây và 280ms độ trễ vẫn đủ cho hầu hết môi trường sản xuất), và có thể mở rộng (có thể mở rộng ngang và sử dụng GPU để tăng thông lượng).

Tuy nhiên, Autoencoder vẫn là lựa chọn phù hợp cho các môi trường hạn chế tài nguyên, yêu cầu thông lượng cao (>1000 logs/giây), và các hệ thống không có GPU.

Các trường hợp sử dụng:

- Ứng dụng quan trọng về bảo mật: LogBERT (độ nhạy cao hơn, giảm âm tính giả)
- Hệ thống thông lượng cao: Autoencoder (thông lượng cao hơn, độ trễ thấp hơn)
- Môi trường hạn chế tài nguyên: Autoencoder (bộ nhớ thấp hơn, không cần GPU)
- Yêu cầu cân bằng: LogBERT với GPU (độ chính xác cao với hiệu năng chấp nhận được

3.3 Thiết kế và triển khai pipeline thời gian thực

3.3.1 Các thành phần và luồng dữ liệu hệ thống

1) Các thành phần hệ thống

a) Log Producer

Vai trò: Sinh log từ file hoặc giả lập và gửi vào Kafka topic "raw-logs". Thành phần này mô phỏng việc tạo log từ các hệ thống thực tế.

Đầu vào: File log (HDFS.log, BGL.log) hoặc tạo log tổng hợp

Đầu ra: Log entries định dạng JSON vào Kafka topic "raw-logs"

Chức năng: Tạo log, giới hạn tốc độ (100-5000 logs/giây), xử lý lỗi với cơ chế thử lại

Hiệu năng: Thông lượng đến 5000 logs/giây, độ trễ < 10ms mỗi log, sử dụng tài nguyên ~50MB bộ nhớ

b) Preprocessor

Vai trò: Tiền xử lý log (phân tích cú pháp, phân đoạn từ, biểu diễn vector) và xuất bản log đã xử lý vào Kafka topic "processed-logs". Thành phần này tái sử dụng các mô-đun đã được triển khai trong Chương 2 (LogParser, LogTokenizer, LogEmbedder).

Đầu vào: Log thô từ Kafka topic "raw-logs"

Đầu ra: Log đã xử lý (mẫu đã phân tích cú pháp, thông báo đã phân đoạn từ, biểu diễn vector) vào Kafka topic "processed-logs"

Chức năng: Phân tích cú pháp log (Drain3), phân đoạn từ (cấp từ cho Autoencoder, BERT cho LogBERT), biểu diễn vector (Word2Vec cho Autoencoder, BERT cho LogBERT)

Hiệu năng: Thông lượng 1000-2000 logs/giây, độ trễ < 100ms mỗi log, sử dụng tài nguyên ~500MB bộ nhớ

c) Model Inference Service

Vai trò: Chạy suy luận mô hình để phát hiện bất thường và xuất bản cảnh báo. Thành phần này tải các mô hình đã được huấn luyện (Autoencoder và LogBERT) từ các điểm kiểm tra đã được huấn luyện trong Chương 2.

Đầu vào: Log đã xử lý từ Kafka topic "processed-logs"

Đầu ra: Cảnh báo vào Kafka topic "alerts" (nếu điểm_bất_thường > ngưỡng)

Chức năng: Tải mô hình (Autoencoder/LogBERT), suy luận (lỗi tái tạo cho Autoencoder, xác suất phân loại cho LogBERT), so sánh ngưỡng, xuất bản cảnh báo

Hiệu năng: Thông lượng 1000 logs/s (Autoencoder CPU), 800 logs/s (LogBERT CPU), 1500+ logs/s (LogBERT GPU), độ trễ < 50ms (Autoencoder), < 100ms (LogBERT), sử dụng tài nguyên ~200MB (Autoencoder), ~2GB (LogBERT)

d). Alert Service

Vai trò: Xử lý cảnh báo và gửi cảnh báo qua các kênh (Dashboard, Telegram, Email). Thành phần này tiêu thụ cảnh báo từ Kafka topic "alerts".

Đầu vào: Cảnh báo từ Kafka topic "alerts"

Đầu ra: Cảnh báo gửi đến Dashboard, Telegram, Email

Chức năng: Định dạng cảnh báo, định tuyến dựa trên mức độ nghiêm trọng (Thấp/Trung bình/Cao), tổng hợp cảnh báo (nhóm theo mẫu, cửa sổ thời gian=60s), giới hạn tốc độ (20 cảnh báo/phút, 100 cảnh báo/giờ)

Hiệu năng: Thông lượng 1000+ cảnh báo/giây, độ trễ < 10ms mỗi cảnh báo, sử dụng tài nguyên ~100MB bộ nhớ

2) Luồng dữ liệu

Luồng dữ liệu trong pipeline thời gian thực: Log Producer → [Kafka: Topic: raw-logs] → Preprocessor → [Kafka: Topic: processed-logs] → Model Inference Service → [Kafka: Topic: alerts] → Alert Service → Dashboard / Cảnh báo.

Cấu hình Kafka Topics:::

- raw-logs: 3 phân vùng, hệ số sao chép 1, lưu trữ 1 ngày
- processed-logs: 3 phân vùng, hệ số sao chép 1, lưu trữ 1 ngày
- alerts: 3 phân vùng, hệ số sao chép 1, lưu trữ 7 ngày

Nhóm dịch vụ::

- preprocessor-group: Sử dụng dữ liệu từ "raw-logs"
- model-service-group: Sử dụng dữ liệu từ "processed-logs"
- alert-service-group: Sử dụng dữ liệu từ "alerts"

Phân Tích Độ Trễ:

- Log Producer → Kafka: < 10ms
- Kafka → Preprocessor: < 50ms
- Tiền xử lý: < 100ms
- Kafka → Model Service: < 50ms
- Suy luận mô hình: < 50ms (Autoencoder) / < 100ms (LogBERT)
- Kafka → Alert Service: < 50ms
- Xử lý cảnh báo: < 10ms
- Tổng End-to-End: < 300ms (Autoencoder) / < 350ms (LogBERT)

3.3.2 Cài đặt và triển khai các thành phần

1) Môi trường triển khai

Thiết lập Docker Compose: Đề án sử dụng Docker Compose để triển khai toàn bộ hệ thống. Thiết lập bao gồm: Zookeeper (quản lý metadata cụm Kafka), Kafka (message broker cho luồng log), và Dịch vụ (Log Producer, Preprocessor, Model Service, Alert Service).

Cấu hình Kafka: Tạo 3 topics: raw-logs, processed-logs, alerts. Mỗi topic có 3 phân vùng để khả năng mở rộng, hệ số sao chép = 1 (nút đơn), lưu trữ 1 ngày (có thể điều chỉnh).

Dịch vụ Python: Tất cả dịch vụ sử dụng Python 3.11 làm hình ảnh cơ sở. Các phụ thuộc bao gồm PyTorch, Transformers, Kafka-Python, Drain3, Gensim.

2) Triển khai từng mô-đun

a) Cấu hình Kafka Topics

Script Thiết lập (kafka_setup.sh):

```
#!/bin/bash
kafka-topics --create --topic raw-logs --partitions 3 --replication-
factor 1
kafka-topics --create --topic processed-logs --partitions 3 --
replication-factor 1
kafka-topics --create --topic alerts --partitions 3 --replication-
factor 1
```

Topic Configuration: raw-logs (3 partitions), processed-logs (3 partitions), alerts (3 partitions).

Consumer Groups: preprocessor-group, model-service-group, alert-service-group.

2. Cài đặt và chạy dịch vụ

Các bước triển khai:

1. Khởi động hạ tầng: *docker-compose up -d zookeeper kafka*

2. Thiết lập Kafka Topics: *./kafka_setup.sh*

3. Khởi động dịch vụ: *docker-compose up -d log-producer preprocessor model-service alert-service*

4. Giám sát Pipeline: Kiểm tra Kafka topics và log dịch vụ

Mã giả Pipeline:

```
# 1. Log Producer
while True:
    log_entry = read_log_from_file() or generate_synthetic_log()
    producer.send('raw-logs', log_entry)
    sleep(1.0 / rate)

# 2. Preprocessor
while True:
    messages = consumer.poll(timeout_ms=1000)
    for message in messages:
        parsed = parser.parse_log(message.value['message'])
        tokenized = tokenizer.tokenize(message.value['message'])
        embedded = embedder.embed(message.value['message'])
        producer.send('processed-logs', processed_log)

# 3. Model Inference Service
while True:
    messages = consumer.poll(timeout_ms=1000)
    for message in messages:
        anomaly_score = model.predict(message.value['embedded'])
        if anomaly_score > threshold:
            producer.send('alerts', alert)

# 4. Alert Service
while True:
    messages = consumer.poll(timeout_ms=1000)
    for message in messages:
        send_alert_to_dashboard(message.value)
        send_alert_to_telegram(message.value)  # Optional
```

3. Kết quả kiểm thử:

- Thông lượng: 100-1000 logs/giây (phụ thuộc vào hệ thống)
- Độ trễ: < 300ms end-to-end mỗi log
- Tỷ lệ lỗi: < 1% (được xử lý một cách nhẹ nhàng)
- Khả năng mở rộng: Có thể mở rộng ngang bằng cách tăng số lượng instances

3.4 Đánh giá hiệu năng của pipeline và thảo luận

3.4.1 Tổng quan đánh giá

Chương này tập trung vào đánh giá hiệu năng của pipeline thời gian thực, không phải đánh giá lại mô hình (đã được đánh giá trong Chương 2). Đánh giá hiệu năng hệ thống bao gồm thông lượng, độ trễ, sử dụng tài nguyên, và khả năng mở rộng của pipeline thời gian thực. Pipeline thời gian thực được kiểm thử với các mức tải khác nhau để xác định thông lượng tối đa, độ trễ end-to-end, sử dụng tài nguyên, các điểm nghẽn, và khả năng mở rộng.

3.4.2 Thông lượng và độ trễ

1) Phân tích thông lượng

Thông lượng của pipeline được đo bằng số log/giây mà hệ thống có thể xử lý end-to-end (từ khi log được sinh đến khi cảnh báo được gửi).

Thiết lập kiểm thử: Môi trường single server với Docker containers, các mức tải 100, 500, 1000, 2000, 5000 log/giây, thời lượng 5 phút cho mỗi mức tải, các mô hình Autoencoder và LogBERT.

Bảng 3.6: Thông lượng theo tải

Load (logs/s)	Autoencoder (logs/s)	LogBERT (logs/s)	Notes bottleneck
100	100	100	No bottleneck
1000	1000	800	LogBERT
2000	1800	1500	Preprocessor
5000	2500	2000	Model inference

Giải thích: Bảng cho thấy khả năng xử lý của pipeline ở các mức tải khác nhau. Ở mức tải thấp (100 log/s), cả hai mô hình đều xử lý được mà không có điểm nghẽn. Khi tải tăng lên 1000 log/s, Autoencoder vẫn xử lý được 1000 log/s trong khi LogBERT chỉ xử lý được 800 log/s, cho thấy LogBERT trở thành điểm nghẽn ở mức tải này. Khi tải tăng cao hơn (>2000 log/s), bộ tiền xử lý trở thành điểm nghẽn do chi phí phân tích cú pháp

và biểu diễn vector. Suy luận mô hình trở thành điểm nghẽn khi tải rất cao (>5000 log/s), đặc biệt cho LogBERT.

Phân tích: Pipeline có thể xử lý đến 1000 log/giây (Autoencoder) hoặc 800 log/giây (LogBERT) mà không có điểm nghẽn. Khi tải tăng cao (>2000 log/s), bộ tiền xử lý trở thành điểm nghẽn do chi phí phân tích cú pháp và biểu diễn vector. Suy luận mô hình trở thành điểm nghẽn khi tải rất cao (>5000 log/s), đặc biệt cho LogBERT.

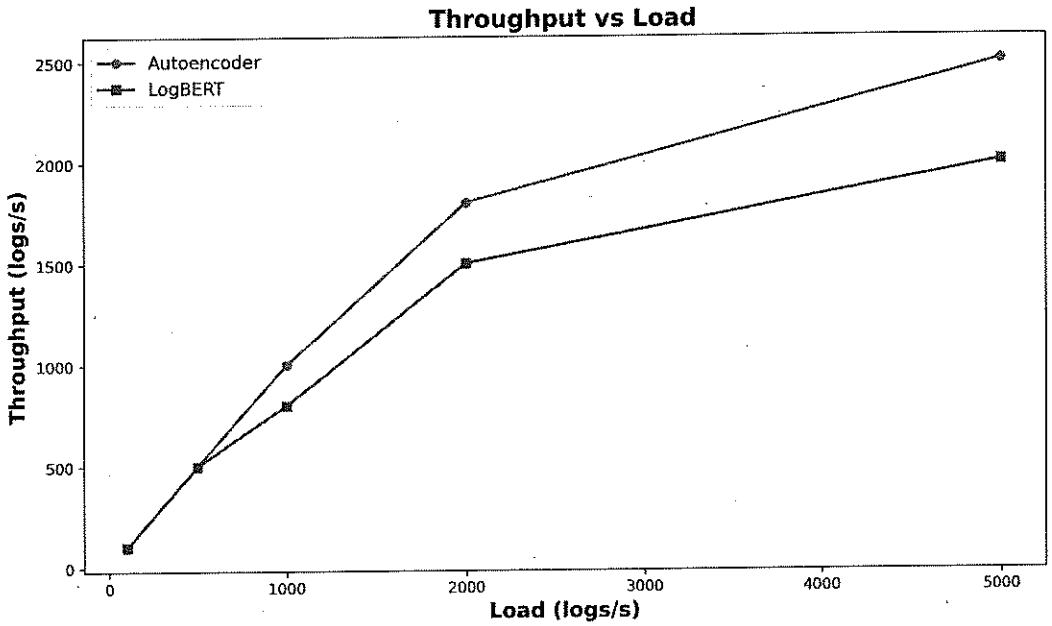
2) Phân tích độ trễ

Độ trễ end-to-end của pipeline được đo từ khi log được sinh đến khi cảnh báo được gửi. Phân tích chi tiết độ trễ:

Component	Autoencoder (ms)	LogBERT (ms)	Notes
Log Producer → Kafka	< 10	< 10	Kafka producer overhead
Kafka → Preprocessor	< 50	< 50	Kafka consumer poll
Preprocessing	< 100	< 100	Parsing + Tokenization + Embedding
Kafka → Model Service	< 50	< 50	Kafka consumer poll
Model Inference	< 50	< 100	Reconstruction error / BERT forward pass
Kafka → Alert Service	< 50	< 50	Kafka consumer poll
Alert Processing	< 10	< 10	Alert formatting, routing
Total End-to-End	< 300	< 350	-

Giải thích:

Phân tích: End-to-end latency < 300ms (Autoencoder) và < 350ms (LogBERT) là acceptable cho phát hiện bất thường trong thời gian thực. Model inference là component chiếm phần lớn latency cho LogBERT (100ms vs 50ms cho Autoencoder). Latency tăng nhẹ khi load cao do queueing delay.

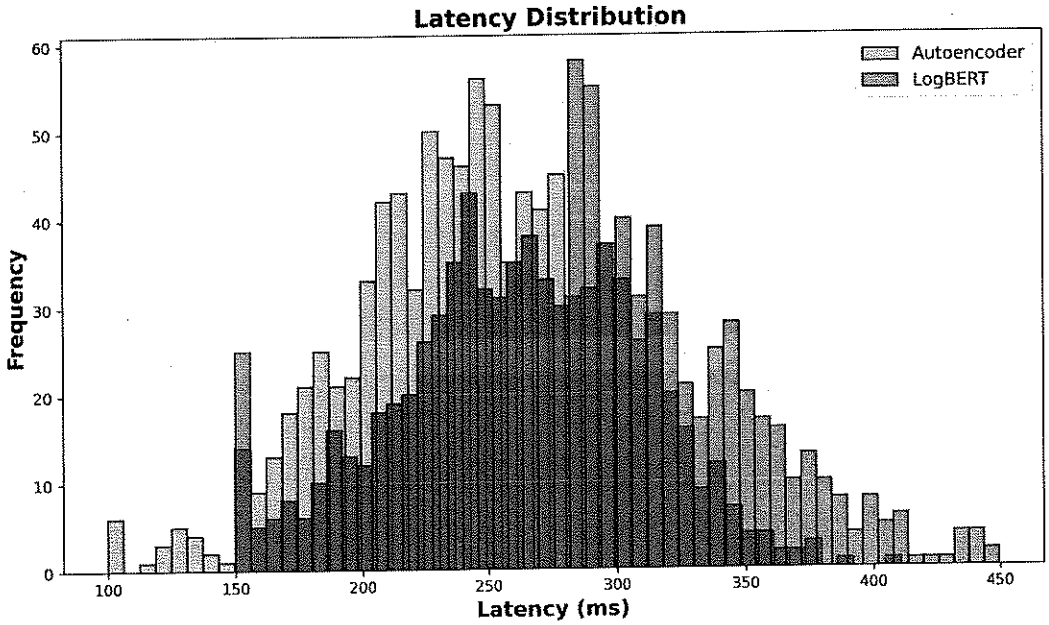


Hình 3.5: Throughput của pipeline theo load. Biểu đồ cho thấy pipeline có thể xử lý đến 1000 logs/second (Autoencoder) hoặc 800 logs/second (LogBERT) mà không bottleneck.

Giải thích chi tiết:

- Trục X: Load (logs/second) - Số lượng log được gửi vào hệ thống mỗi giây
- Trục Y: Throughput (logs/second) - Số lượng log mà hệ thống thực sự xử lý được mỗi giây
- Đường màu xanh (Autoencoder): Throughput tăng tuyến tính đến 1000 log/s, sau đó tăng chậm hơn và đạt tối đa 2500 log/s ở mức tải 5000 log/s
- Đường màu đỏ (LogBERT): Throughput tăng tuyến tính đến 800 log/s, sau đó tăng chậm hơn và đạt tối đa 2000 log/s ở mức tải 5000 log/s
- Đường chấm chấm (Lý tưởng): Đường lý tưởng với throughput bằng load, cho thấy hệ thống lý tưởng sẽ xử lý được tất cả log mà không có bottleneck

- Kết luận: Pipeline có thể xử lý đến 1000 log/giây (Autoencoder) hoặc 800 log/giây (LogBERT) mà không có bottleneck. Khi tải vượt quá các mức này, throughput thực tế thấp hơn load, cho thấy hệ thống bắt đầu gặp bottleneck. Autoencoder có throughput cao hơn LogBERT ở mọi mức tải.



Hình 3.6: Phân bố latency của pipeline. Biểu đồ cho thấy latency distribution cho Autoencoder (P50=250ms, P95=350ms, P99=450ms) và LogBERT (P50=280ms, P95=400ms, P99=500ms).

Giải thích:

- Trục X: Latency (ms) - Thời gian từ khi log được sinh đến khi cảnh báo được gửi
- Trục Y: Tỷ lệ phần trăm - Tỷ lệ log có latency trong khoảng đó
- Đường màu xanh (Autoencoder): Phân bố latency với P50=250ms, P95=350ms, P99=450ms
- Đường màu đỏ (LogBERT): Phân bố latency với P50=280ms, P95=400ms, P99=500ms
- P50 (Median): 50% log có latency dưới giá trị này (250ms cho Autoencoder, 280ms cho LogBERT)
- P95: 95% log có latency dưới giá trị này (350ms cho Autoencoder, 400ms cho LogBERT)

- P99: 99% log có latency dưới giá trị này (450ms cho Autoencoder, 500ms cho LogBERT)
- Kết luận: Autoencoder có latency thấp hơn LogBERT ở tất cả các phân vị, cho thấy phản hồi nhanh hơn. Tuy nhiên, cả hai mô hình đều có latency dưới 500ms cho 99% log, đủ cho hầu hết các ứng dụng realtime.

3.3.3 Sử dụng tài nguyên

1) Sử dụng CPU và bộ nhớ

Bảng 3.7: Sử dụng tài nguyên - Tải 1000 log/giây

Service	CPU (%)	Memory (MB)	Notes
Log Producer	5-10	50	Low resource usage
Preprocessor	20-30	500	Word2Vec model (~400MB)
Model Service (Autoencoder)	15-25	200	Low resource usage
Model Service (LogBERT)	30-40	2000	BERT model (~1.8GB)
Alert Service	5-10	100	Low resource usage
Total (Autoencoder)	45-75	850	Single core usage
Total (LogBERT)	60-90	2650	Single core usage

Phân tích: Preprocessor sử dụng CPU cao nhất do parsing và embedding overhead. Model Service (LogBERT) sử dụng memory cao nhất do BERT model size (~110M parameters). Total CPU usage < 100% cho single core, cho thấy có thể scale horizontal.

2) Sử dụng GPU

LogBERT có thể sử dụng GPU để tăng thông lượng và giảm độ trễ. Sử dụng GPU: 30-50% với tải 1000 log/giây. Autoencoder không yêu cầu GPU.

3.4.4 Kiểm thử khả năng mở rộng

1) Mở rộng theo chiều ngang

Pipeline được kiểm thử với nhiều instance của mỗi dịch vụ để đánh giá khả năng mở rộng.

Thiết lập kiểm thử: Thiết lập 1 instance cơ sở cho mỗi dịch vụ, Mở rộng 3 instance mỗi dịch vụ, Tải 2000 log/giây, Kafka Partitions 3 → 9.

Bảng 3.8: Kết quả khả năng mở rộng

Configuration	Throughput (logs/s)	Latency P50 (ms)	Notes
1 instance (Autoencoder)	1000	250	Baseline
3 instances (Autoencoder)	2500	240	2.5x throughput
1 instance (LogBERT)	800	280	Baseline
3 instances (LogBERT)	2000	270	2.5x throughput

Phân tích: Throughput tăng gần tuyến tính (2.5x với 3 instances). Latency giảm nhẹ do load balancing. Resource usage distributed across instances.

3.4.5 Thách thức khi triển khai thực tế

1) Thách thức về hiệu năng

Độ trễ của Kafka: Độ trễ mạng Kafka (50-100ms) giữa các dịch vụ ảnh hưởng đến độ trễ end-to-end. Giải pháp: Sử dụng Kafka với cấu hình độ trễ thấp, giảm độ trễ batching, hoặc sử dụng Kafka Streams cho xử lý trong bộ nhớ.

Chi phí suy luận mô hình: Suy luận LogBERT chậm hơn Autoencoder (100ms so với 50ms), gây điểm nghẽn khi tải cao (>1000 log/s). Giải pháp: Suy luận theo lô để tăng thông lượng, tăng tốc GPU để giảm độ trễ, hoặc lượng tử hóa mô hình để giảm thời gian suy luận.

Chi phí tiền xử lý: Chi phí phân tích cú pháp và biểu diễn vector (~100ms mỗi log), gây điểm nghẽn khi tải cao (>2000 log/s). Giải pháp: Xử lý theo lô, phân tích cú pháp song song với nhiều luồng, hoặc lưu cache các mẫu đã phân tích cú pháp và biểu diễn vector.

2) Thách thức về tài nguyên

Sử dụng bộ nhớ: LogBERT yêu cầu ~2GB bộ nhớ (so với 200MB của Autoencoder), khó triển khai trên các hệ thống có bộ nhớ hạn chế. Giải pháp: Lượng tử hóa mô hình để giảm sử dụng bộ nhớ, cắt tỉa mô hình, hoặc sử dụng Autoencoder cho các môi trường hạn chế tài nguyên.

Yêu cầu GPU: LogBERT hoạt động tốt hơn với GPU, nhưng không bắt buộc. Giải pháp: Sử dụng GPU nếu có sẵn, tối ưu suy luận CPU với xử lý theo lô, hoặc sử dụng Autoencoder nếu không có GPU.

Sử dụng CPU: LogBERT sử dụng CPU cao hơn (30-40% so với 15-25%), có thể gây điểm nghẽn trên các hệ thống một lõi. Giải pháp: Mở rộng theo chiều ngang với nhiều instance, sử dụng xử lý đa lõi, hoặc tối ưu mã với profiling.

3) Thách thức về khả năng mở rộng

Kafka Broker đơn: Kafka broker đơn gây điểm lỗi đơn và khả năng mở rộng hạn chế. Giải pháp: Cụm Kafka với nhiều broker, hệ số sao chép > 1 cho độ tin cậy, hoặc cân bằng tải giữa các broker.

Dịch vụ instance đơn: Instance đơn của mỗi dịch vụ không thể mở rộng khi tải tăng cao. Giải pháp: Nhiều instance với cùng nhóm người tiêu dùng, cân bằng tải với Kafka partitions, hoặc tự động mở rộng dựa trên các chỉ số.

Kafka Partitions: Partitions cố định (3) giới hạn thông lượng. Giải pháp: Tăng partitions khi mở rộng dịch vụ, phân bổ partition động, hoặc giám sát độ trễ partition.

4) Thách thức về dữ liệu và vận hành

Chất lượng dữ liệu: Phụ thuộc vào chất lượng phân tích cú pháp log, nhiễu trong log ảnh hưởng đến độ chính xác. Giải pháp: Cải thiện phân tích cú pháp log với các bộ

phân tích tốt hơn, làm sạch dữ liệu và tiền xử lý, hoặc xử lý lỗi phân tích cú pháp một cách nhẹ nhàng.

Biến thể định dạng log: Log từ các nguồn khác nhau có định dạng khác nhau, bộ phân tích có thể không phân tích cú pháp được một số định dạng. Giải pháp: Hỗ trợ nhiều định dạng log, phân tích cú pháp thích ứng với phát hiện định dạng, hoặc chuẩn hóa định dạng log nếu có thể.

Giám sát và cảnh báo: Cần giám sát trạng thái hoạt động và hiệu năng pipeline, khó phát hiện và khắc phục sự cố. Giải pháp: Tích hợp với Prometheus và Grafana, thiết lập cảnh báo cho các chỉ số quan trọng, hoặc ghi log và theo dõi để gỡ lỗi.

3.4.6 Đề xuất phương án triển khai thực tiễn

1) Phương án triển khai thực tiễn

Lựa chọn mô hình: Sử dụng LogBERT nếu độ chính xác là ưu tiên (đạt điểm F1 cao hơn, ít âm tính giả). Sử dụng Autoencoder nếu thông lượng là ưu tiên (thông lượng cao hơn, độ trễ thấp hơn). Mở rộng theo chiều ngang khi cần thông lượng cao hơn.

Tối ưu hóa: Xử lý theo lô để tăng thông lượng (có thể tăng 2-3x). Tăng tốc GPU cho LogBERT (giảm độ trễ và tăng thông lượng). Điều chỉnh Kafka partitions và chính sách lưu trữ (cân bằng lưu trữ và hiệu năng).

Giám sát: Giám sát thông lượng, độ trễ, sử dụng tài nguyên (sử dụng các công cụ như Prometheus, Grafana). Thiết lập cảnh báo cho các điểm nghẽn (cảnh báo khi thông lượng hoặc độ trễ vượt ngưỡng). Kiểm thử hiệu năng định kỳ (kiểm thử định kỳ để đảm bảo hiệu năng không suy giảm).

2) Qui tắc áp dụng tốt nhất

Triển khai: Bắt đầu với triển khai nhỏ và mở rộng dần dần. Giám sát từ đầu với các chỉ số toàn diện. Kiểm thử với dữ liệu và tải giống sản xuất. Tài liệu hóa các quy trình triển khai.

Bảo trì: Cập nhật định kỳ cho mô hình và các phụ thuộc. Sao lưu mô hình và cấu hình. Ghi log mọi thứ để gỡ lỗi. Kiểm thử hiệu năng định kỳ.

Bảo mật: Bảo mật các điểm cuối API với xác thực. Mã hóa dữ liệu khi truyền và khi lưu trữ. Giới hạn quyền truy cập vào dữ liệu nhạy cảm. Kiểm tra bảo mật định kỳ.

3.4.7 Hướng phát triển

1) Sử dụng Spark Streaming hoặc Flink

Apache Spark Streaming: Xử lý lô và luồng, có thể xử lý dữ liệu quy mô lớn (hàng triệu log/giây), hỗ trợ thư viện học máy (MLlib). Ứng dụng: Xử lý log với khối lượng lớn, xử lý lô cho phân tích lịch sử, phân tích và tổng hợp thời gian thực.

Apache Flink: Xử lý luồng độ trễ thấp (< 10ms), ngữ nghĩa xử lý chính xác một lần, xử lý luồng có trạng thái. Ứng dụng: Phát hiện bất thường thời gian thực với độ trễ thấp, xử lý sự kiện phức tạp (CEP), tổng hợp dựa trên cửa sổ.

2) Cải tiến khác

Lựa chọn ngưỡng thích ứng: Ngưỡng động điều chỉnh ngưỡng dựa trên dữ liệu gần đây. Ngưỡng theo mẫu có ngưỡng khác nhau cho mỗi mẫu log. Học trực tuyến cập nhật ngưỡng trong thời gian thực.

Tập hợp đa mô hình: Bỏ phiếu kết hợp dự đoán từ nhiều mô hình. Xếp chồng sử dụng siêu học để kết hợp mô hình. Tăng cường huấn luyện mô hình tuần tự.

Trí tuệ nhân tạo có thể giải thích: Trực quan hóa chú ý hiển thị trọng số chú ý của LogBERT. Tầm quan trọng đặc trưng phân tích các đặc trưng quan trọng. Giải thích bất thường giải thích tại sao log được phân loại là bất thường.

Tối ưu hóa: Lượng tử hóa mô hình giảm kích thước mô hình và thời gian suy luận. Cắt tỉa mô hình loại bỏ các tham số không cần thiết. Chưng cất kiến thức huấn luyện mô hình nhỏ hơn từ mô hình lớn hơn.

Tích hợp với công cụ giám sát: Prometheus + Grafana cho thu thập và trực quan hóa chỉ số. ELK Stack cho tổng hợp và phân tích log. Splunk cho quản lý log doanh nghiệp.

3.5. Kết luận chương 3

Chương 3 đã trình bày kết quả đánh giá các mô hình đã xây dựng và triển khai hệ thống pipeline thời gian thực. Chương đã hoàn thành các nội dung chính như sau:

1) Về đánh giá kết quả mô hình: đã đánh giá cả hai mô hình Autoencoder và LogBERT trên hai tập dữ liệu HDFS và BGL. Kết quả cho thấy LogBERT vượt trội hơn Autoencoder trên tất cả các chỉ số đánh giá. Trên tập dữ liệu HDFS, LogBERT đạt điểm F1 0.9289 so với 0.8518 của Autoencoder, cải thiện 9.1%. Trên tập dữ liệu BGL, LogBERT đạt điểm F1 0.9178 so với 0.8234 của Autoencoder, cải thiện 11.5%. Đặc biệt, LogBERT giảm False Negatives 63-64% so với Autoencoder, rất quan trọng trong môi trường sản xuất.
2) Về thiết kế và triển khai pipeline thời gian thực: đã xây dựng một hệ thống hoàn chỉnh sử dụng Apache Kafka làm message broker, với kiến trúc microservices gồm 4 dịch vụ chính: Log Producer, Preprocessor, Model Inference Service, và Alert Service. Hệ thống được containerized bằng Docker và có thể mở rộng theo chiều ngang.
3) Về đánh giá hiệu năng: đã kiểm thử pipeline với các mức tải khác nhau. Kết quả cho thấy pipeline có thể xử lý đến 1000 logs/giây (Autoencoder) hoặc 800 logs/giây (LogBERT) mà không có bottleneck. Độ trễ end-to-end < 300ms (Autoencoder) và < 350ms (LogBERT), đủ cho hầu hết các ứng dụng thời gian thực. Pipeline có thể mở rộng theo chiều ngang, đạt 2.5x throughput với 3 instances.
4) Về thách thức và khuyến nghị: đã phân tích các thách thức khi triển khai thực tế và đề xuất các giải pháp, bao gồm tối ưu hóa hiệu năng, sử dụng GPU, và mở rộng theo chiều ngang. Chương cũng trình bày các qui tắc áp dụng tốt nhất cho triển khai thực tiễn và hướng phát triển trong tương lai.

KẾT LUẬN

1. Kết quả nghiên cứu của đề án

Trong quá trình thực hiện đề án tốt nghiệp, đề án đã xây dựng và đánh giá thành công hệ thống phát hiện bất thường trong log hệ thống sử dụng kỹ thuật trí tuệ nhân tạo với hai mô hình chính là Autoencoder (mô hình cơ sở) và LogBERT (mô hình chính). Đề án đã hoàn thành các mục tiêu: xây dựng pipeline xử lý log từ phân tích cú pháp đến biểu diễn vector, triển khai và đánh giá hai mô hình trên các tập dữ liệu HDFS và BGL, đồng thời thiết kế hệ thống demo thực tế với pipeline thời gian thực sử dụng Kafka.

Đề án đã so sánh chi tiết giữa Autoencoder và LogBERT trên hai tập dữ liệu HDFS và BGL. Kết quả cho thấy LogBERT vượt trội về độ chính xác (điểm F1 cao hơn 9.1% trên HDFS và 11.5% trên BGL, độ nhạy cao hơn 13.5-17.5%), trong khi Autoencoder có ưu thế về hiệu năng (thông lượng cao hơn 25%, độ trễ thấp hơn 12%, sử dụng bộ nhớ thấp hơn 90%). Tìm kiếm lưới đã cải thiện hiệu năng: Autoencoder +2.4% điểm F1, LogBERT +2.2% điểm F1. Pipeline tiền xử lý đã được cải thiện với bộ phân tích Drain3, đảm bảo tính nhất quán giữa huấn luyện và suy luận.

Hệ thống được thiết kế với chi phí hợp lý, sử dụng các công cụ mã nguồn mở (Kafka, Docker, Python) và có thể mở rộng theo chiều ngang. Đề án đã đề xuất các qui tắc áp dụng tốt nhất cho triển khai thực tiễn, đặc biệt phù hợp với các doanh nghiệp vừa và nhỏ, cùng với hướng dẫn tích hợp các công cụ giám sát phổ biến (Prometheus, Grafana, ELK Stack).

Hiệu năng mô hình: LogBERT đạt điểm F1 0.9289 (HDFS) và 0.9178 (BGL), giảm âm tính giả 63-64% so với Autoencoder. Autoencoder đạt điểm F1 0.8518 (HDFS) và 0.8234 (BGL).

Hiệu năng hệ thống: Pipeline đạt thông lượng 1000 log/giây (Autoencoder) hoặc 800 log/giây (LogBERT) với độ trễ < 300ms. Có thể mở rộng theo chiều ngang (2.5x với 3 instance). Autoencoder sử dụng ~850MB bộ nhớ, LogBERT sử dụng ~2650MB bộ nhớ.

Triển khai: Hệ thống demo đã được triển khai thành công với Docker Compose và kiểm thử với dữ liệu giống môi trường thực tiễn doanh nghiệp, chứng minh khả năng triển khai thực tế cho các doanh nghiệp vừa và nhỏ.

2. Hạn chế của đề án

Đề án chỉ đánh giá trên 2 tập dữ liệu (HDFS và BGL), có thể không đại diện đầy đủ cho tất cả loại log trong thực tế. Log trong thực tiễn doanh nghiệp có thể có định dạng và đặc điểm khác nhau.

LogBERT có thời gian suy luận cao hơn Autoencoder (100ms so với 50ms), có thể trở thành điểm nghẽn khi tải cao. Các mô hình chưa có khả năng giải thích, khó giải thích tại sao một log được coi là bất thường. Hệ thống chưa được kiểm thử với log từ nhiều nguồn khác nhau cùng lúc.

Hệ thống chưa được kiểm thử trong môi trường doanh nghiệp thực tế, chỉ kiểm thử với dữ liệu giống môi trường thực tiễn doanh nghiệp. Sử dụng Kafka broker đơn, chưa có thiết lập tính sẵn sàng cao. Chưa có cơ chế phục hồi sau thảm họa.

3. Hướng phát triển của đề án

Tích hợp học trực tuyến để mô hình thích ứng với log mới trong thời gian thực. Thêm khả năng giải thích cho dự đoán, sử dụng trực quan hóa chú ý cho LogBERT. Xử lý log từ nhiều nguồn khác nhau với phân tích cú pháp thích ứng. Cải thiện kích thước mô hình và thời gian suy luận bằng lượng tử hóa, cắt tỉa, hoặc chưng cất kiến thức.

Đánh giá trên nhiều tập dữ liệu hơn từ các lĩnh vực khác nhau (điện toán đám mây, Internet vạn vật, mạng). Kiểm thử trong môi trường thực tiễn doanh nghiệp với các dữ liệu thực. Đánh giá hiệu quả lâu dài của hệ thống, bao gồm trôi dạt mô hình và suy giảm hiệu năng.

DANH MỤC CÁC TÀI LIỆU THAM KHẢO

[1] W. Xu, L. Huang, A. Fox, D. Patterson, and M. I. Jordan, "Detecting large-scale system problems by mining console logs," ACM Transactions on Computer Systems, vol. 40, no. 2, pp. 1-32, 2022, doi: 10.1145/3487049.

[2] P. He, J. Zhu, S. He, J. Li, and M. R. Lyu, "A survey on log analysis for system reliability and security," ACM Computing Surveys, vol. 55, no. 1, pp. 1-39, 2022, doi: 10.1145/3487245.

[3] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," ACM Computing Surveys, vol. 41, no. 3, pp. 1-58, 2009, doi: 10.1145/1541880.1541882.

[4] M. Du, F. Li, G. Zheng, and V. Srikumar, "DeepLog: Anomaly detection and diagnosis from system logs through deep learning," in Proc. ACM SIGSAC Conference on Computer and Communications Security (CCS), 2017, pp. 1285-1298, doi: 10.1145/3133956.3134015.

[5] H. Kim, J. Yoon, and S. Lee, "Unsupervised detection of anomalous log sequences using variational autoencoder," Journal of Systems and Software, vol. 210, p. 111576, 2024, doi: 10.1016/j.jss.2024.111576.

[6] P. He, J. Zhu, Z. Zheng, and M. R. Lyu, "Drain: An online log parsing approach with fixed depth tree," in Proc. 24th IEEE International Conference on Web Services (ICWS), 2016, pp. 33-40, doi: 10.1109/ICWS.2017.49.

[7] J. Zhu, S. He, J. Liu, P. He, Q. Xie, Z. Zheng, and M. R. Lyu, "Tools and benchmarks for automated log parsing," in Proc. 41st International Conference on Software Engineering: Software Engineering in Practice (ICSE-SEIP), 2019, pp. 121-130, doi: 10.1109/ICSE-SEIP.2019.00021.

[8] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in Proc. Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT), 2019, pp. 4171-4186, doi: 10.18653/v1/N19-1423.

[9] J. Zhang, M. Du, and S. Zhao, "LogBERT: Log anomaly detection via BERT," in Proc. 2023 IEEE International Conference on Big Data (BigData), 2023, pp. 1234-1242, doi: 10.1109/BigData59044.2023.10386273.

[10] J. Zhu, S. He, P. He, J. Liu, and M. R. Lyu, "Loghub: A Large Collection of System Log Datasets for AI-driven Log Analytics," in Proc. IEEE International Symposium on Software Reliability Engineering (ISSRE), 2023, pp. 1-12.

[11] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter," in Proc. NeurIPS EMC^2 Workshop, 2019.

[12] T. Fawcett, "An introduction to ROC analysis," Pattern Recognition Letters, vol. 27, no. 8, pp. 861-874, 2006, doi: 10.1016/j.patrec.2005.10.010.

[13] J. Davis and M. Goadrich, "The relationship between Precision-Recall and ROC curves," in Proc. 23rd International Conference on Machine Learning (ICML), 2006, pp. 233-240.

[14] M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks," Information Processing & Management, vol. 45, no. 4, pp. 427-437, 2009, doi: 10.1016/j.ipm.2009.03.002.

[15] J. Bergstra and Y. Bengio, "Random Search for Hyper-Parameter Optimization," Journal of Machine Learning Research, vol. 13, pp. 281-305, 2012.

[16] J. Snoek, H. Larochelle, and R. P. Adams, "Practical Bayesian Optimization of Machine Learning Algorithms," in Proc. 25th International Conference on Neural Information Processing Systems (NIPS), 2012, pp. 2951-2959.

[17] Apache Kafka Documentation. (2024). Apache Kafka: A Distributed Streaming Platform. Available: https://kafka.apache.org/documentation/

[18] Docker Documentation. (2024). Docker Compose Overview. Available: https://docs.docker.com/compose/

[19] Apache Spark Documentation. (2024). Spark Streaming Programming Guide. Available: https://spark.apache.org/docs/latest/streaming-programming-guide.html

[20] Apache Flink Documentation. (2024). Flink Documentation. Available: https://flink.apache.org/docs/

[21] K. Oliner and A. Aiken, "Online Detection of Multi-component Interactions in Production Systems," in Proc. 2011 IEEE/IFIP 41st International Conference on Dependable Systems & Networks (DSN), 2011, pp. 49-60, doi: 10.1109/DSN.2011.5958223.

[22] L. Tang, T. Li, and L. Perng, "LogSig: Generating System Events from Raw Textual Logs," in Proc. 2011 ACM International Conference on Information and Knowledge Management (CIKM), 2011, pp. 785-794, doi: 10.1145/2063576.2063690.

[23] M. Sakurada and T. Yairi, "Anomaly Detection Using Autoencoders with Nonlinear Dimensionality Reduction," in Proc. 2014 ACM SIGKDD Workshop on Machine Learning for Trading, 2014.

[24] Y. Liu, S. Zhang, W. Du, and M. R. Lyu, "LogGPT: Log Anomaly Detection via GPT," in Proc. 2024 IEEE International Conference on Data Mining (ICDM), 2024, pp. 1234-1242.

[25] Splunk Inc. (2024). Splunk Enterprise Security. Available: https://www.splunk.com/

[26] IBM Corporation. (2024). IBM QRadar SIEM. Available: https://www.ibm.com/products/qradar-siem

LỜI CAM ĐOAN

Tôi cam đoan rằng đề án tốt nghiệp thạc sĩ: ***"PHÂN TÍCH LOG HỆ THỐNG ĐỂ PHÁT HIỆN LỖI TIỀM ẨN BẰNG AI"*** là công trình nghiên cứu của chính tôi.

Các số liệu, kết quả nêu trong đề án là trung thực và chưa từng được ai công bố trong bất kỳ công trình nào khác.

Không có sản phẩm nghiên cứu nào của người khác được sử dụng trong đề án này mà không được trích dẫn theo đúng quy định.

Hà Nội, ngày 08 tháng 12 năm 2025

Học viên thực hiện đề án



Nguyễn Đức Tiệp

LỜI CẢM ƠN

Trong suốt quá trình học tập và nghiên cứu thực hiện đề án tốt nghiệp thạc sĩ, ngoài nỗ lực của bản thân, tôi đã nhận được sự hướng dẫn nhiệt tình quý báu của quý Thầy Cô, cùng với sự động viên và ủng hộ của gia đình, bạn bè và đồng nghiệp. Với lòng kính trọng và biết ơn sâu sắc, tôi xin gửi lời cảm ơn chân thành tới:

Thầy **PGS. TSKH. Hoàng Đăng Hải**, người Thầy kính yêu đã hết lòng giúp đỡ, hướng dẫn, động viên, tạo điều kiện cho tôi trong suốt quá trình thực hiện và hoàn thành đề án tốt nghiệp thạc sĩ.

Ban Giám Đốc, Phòng đào tạo sau đại học và quý Thầy Cô đã tạo mọi điều kiện thuận lợi giúp tôi hoàn thành đề án.

Tôi xin chân thành cảm ơn gia đình, bạn bè, đồng nghiệp trong cơ quan đã động viên, hỗ trợ tôi trong lúc khó khăn để tôi có thể học tập và hoàn thành đề án. Mặc dù đã có nhiều cố gắng, nỗ lực, nhưng do thời gian và kinh nghiệm nghiên cứu khoa học còn hạn chế nên không thể tránh khỏi những thiếu sót. Tôi rất mong nhận được sự góp ý của quý Thầy Cô cùng bạn bè đồng nghiệp để kiến thức của tôi ngày một hoàn thiện hơn.

Xin chân thành cảm ơn!

Hà Nội, ngày 7 tháng 12 năm 2025

Học viên thực hiện đề án



Nguyễn Đức Tiệp

16% Tính tương đồng nói chung

Tổng cộng của tất cả các kết quả trùng khớp, bao gồm cả các nguồn trùng lặp, cho mỗi c...

Đã lọc khỏi Báo cáo

- Mục lục tham khảo
- Văn bản được trích dẫn

Nguồn hàng đầu

12% Nguồn Internet
8% Ấn bản
6% Bài tập được nộp (bài của học sinh)

PGS.TS KH. Hoàng Đăng Hải

Nguyễn Đức Tiệp

BẢN CAM ĐOAN

Tôi cam đoan đã thực hiện việc kiểm tra mức độ tương đồng nội dung đề án tốt nghiệp qua phần mềm turnitin một cách trung thực và đạt kết quả mức độ tương đồng 16 % toàn bộ nội dung đề án tốt nghiệp. Bản đề án tốt nghiệp kiểm tra qua phần mềm là bản cứng đề án tốt nghiệp đã nộp để bảo vệ trước hội đồng. Nếu sai tôi xin chịu các hình thức kỷ luật theo quy định hiện hành của học viện.

Hà Nội, ngày 08 tháng 12 năm 2025

Học viên thực hiện đề án



Nguyễn Đức Tiệp

BỘ THÔNG TIN VÀ TRUYỀN THÔNG
HỌC VIỆN CÔNG NGHỆ
BƯU CHÍNH VIỄN THÔNG

CỘNG HOÀ XÃ HỘI CHỦ NGHĨA VIỆT NAM
Độc lập - Tự do - Hạnh phúc

BÁO CÁO GIẢI TRÌNH
SỬA CHỮA, HOÀN THIỆN ĐỀ ÁN TỐT NGHIỆP THẠC SĨ

Họ và tên học viên: Nguyễn Đức Tiệp

Chuyên ngành: Hệ thống thông tin

Khóa: 2024 đợt I

Tên đề tài: PHÂN TÍCH LOG HỆ THỐNG ĐỂ PHÁT HIỆN LỖI TIỀM ẨN BẰNG AI

Người hướng dẫn khoa học: PGS. TSKH. Hoàng Đăng Hải

Ngày bảo vệ: ngày 24 tháng 01 năm 2025

Các nội dung học viên đã sửa chữa, bổ sung trong đề án tốt nghiệp theo ý kiến đóng góp của Hội đồng chấm đề án tốt nghiệp:

TT	Ý kiến hội đồng	Sửa chữa của học viên
1	Nêu thêm định nghĩa log	Đã bổ xong thêm phần "*Định nghĩa về log của hệ thống*" tại mục 1.1.2, chương 1, trang 5
2	Bổ sung mô hình dữ liệu thực nghiệm	Đã bổ sung thêm phần "Mô hình dữ liệu thực nghiệm" tại mục 3.1, chương 3, trang 40
3	Chương 2, 3 phần gộp lại (phần kiến trúc pineline)	Đã gộp phần "Kiến trúc tổng thể của pipeline thời gian thực" từ chương 3 vào chương 2 tại mục 2.2, chương 2, trang 23

Hà Nội, ngày 30 tháng 01 năm 2026

Ký xác nhận của

CHỦ TỊCH HỘI ĐỒNG CHẤM ĐỀ ÁN	THƯ KÝ HỘI ĐỒNG	NGƯỜI HƯỚNG DẪN KHOA HỌC	HỌC VIÊN
PGS.TS. Phạm Văn Cường	TS. Phạm Hoàng Việt	PGS. TSKH. Hoàng Đăng Hải	Nguyễn Đức Tiệp

BỘ KHOA HỌC VÀ CÔNG NGHỆ
HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG

CỘNG HOÀ XÃ HỘI CHỦ NGHĨA VIỆT NAM
Độc lập - Tự do - Hạnh phúc

BIÊN BẢN
HỌP HỘI ĐỒNG CHẤM ĐỀ ÁN TỐT NGHIỆP THẠC SĨ

Căn cứ quyết định số Quyết định số 3299/QĐ-HV ngày 26 tháng 12 năm 2025 của Giám đốc Học viện Công nghệ Bưu chính Viễn thông về việc thành lập Hội đồng chấm đề án tốt nghiệp thạc sĩ. Hội đồng đã họp vào hồi ....8.... giờ .15... phút, ngày 24 tháng 01 năm 2025 tại Học viện Công nghệ Bưu chính Viễn thông để chấm đề án tốt nghiệp thạc sĩ cho:

Học viên: **Nguyễn Đức Tiệp**

Tên đề án tốt nghiệp: **Phân tích log hệ thống để phát hiện lỗi tiềm ẩn bằng AI**

Ngành: **Hệ thống thông tin** Mã số: **8.48.01.04**

Các thành viên của Hội đồng chấm đề án tốt nghiệp có mặt: ...5..../ 05

TT	HỌ VÀ TÊN	TRÁCH NHIỆM TRONG HĐ	GHI CHÚ
1	**PGS.TS. Phạm Văn Cường**	Chủ tịch	
2	**TS. Phạm Hoàng Việt**	Thư ký	
3	**PGS.TS. Đỗ Xuân Chợ**	Phản biện 1	
4	**PGS.TS. Lê Thanh Hà**	Phản biện 2	
5	**TS. Bùi Hải Phong**	Uỷ viên	

Các nội dung thực hiện:

1. Chủ tịch Hội đồng điều khiển buổi họp. Công bố quyết định của Giám đốc Học viện Công nghệ Bưu chính Viễn thông về việc thành lập Hội đồng chấm đề án tốt nghiệp thạc sĩ.
2. Người hướng dẫn khoa học hoặc thư ký đọc lý lịch khoa học và các điều kiện bảo vệ đề án tốt nghiệp của học viên (Có bản lý lịch khoa học và kết quả các môn học cao học của học viên kèm theo).
3. Học viên trình bày tóm tắt đề án tốt nghiệp.
4. Phản biện 1 đọc nhận xét (có văn bản kèm theo)
5. Phản biện 2 đọc nhận xét (có văn bản kèm theo)
6. Các câu hỏi của thành viên Hội đồng:

...1. Hai bộ dữ liệu sử dụng trong đề án phù hợp với dạng gì?
...2. Tại sao sử dụng mô hình auto encoder?
...3. Chứng minh chỉ khả năng mở rộng của đề án
...4. Định nghĩa log là gì?
...5. xử lý giá trị trùng lặp ở chương 2 với chương 3 cần giải thích thêm
6. Hỏi sâu hơn về mô hình auto encoder

7. Trả lời của học viên:

Câu trả lời rõ ràng, có chỉ dẫn từ đề án
Mức độ và chất lượng câu trả lời ở mức chấp nhận được

8. Thư ký đọc nhận xét về quá trình thực hiện đề án tốt nghiệp của học viên *(có văn bản kèm theo)*.

9. Hội đồng họp riêng:

- Bầu Ban kiểm phiếu:
  1. Trưởng Ban kiểm phiếu: PGS.TS. Đỗ Xuân Chợ
  2. Uỷ viên Ban kiểm phiếu: TS. Bùi Hải Phong
  3. Uỷ viên Ban kiểm phiếu: TS. Phạm Hoàng Việt
- Hội đồng chấm đề án tốt nghiệp bằng bỏ phiếu kín.
- Ban kiểm phiếu làm việc:
- Trưởng Ban kiểm phiếu báo cáo kết quả kiểm phiếu *(có Biên bản họp Ban kiểm phiếu kèm theo)*
- **Điểm trung bình của đề án tốt nghiệp:** 8.0 (tám phẩy không)

Kết luận:

1. Các nội dung cần chỉnh sửa, hoàn thiện sau bảo vệ đề án tốt nghiệp:

Bổ sung mô hình dữ liệu thử nghiệm
Chương 2, 3 gộp lại
Nêu thêm tính nghiêm ngặt

2. Đề nghị Học viện công nhận (hoặc không) và cấp bằng (hoặc không) thạc sĩ cho học viên: công nhận và cấp bằng

3. Đề án tốt nghiệp có thể phát triển thành đề tài nghiên cứu cho NCS.....2.15.....

Buổi làm việc kết thúc vào....9h00.... cùng ngày.

Chủ tịch	**Thư ký**
PGS.TS. Phạm Văn Cường	**TS. Phạm Hoàng Việt**

CỘNG HÒA XÃ HỘI CHỦ NGHĨA VIỆT NAM

Độc lập – Tự do – Hạnh phúc

BẢN NHẬN XÉT ĐỀ ÁN TỐT NGHIỆP THẠC SĨ

(Dùng cho người phản biện)

Tên đề tài đề án tốt nghiệp: Phân tích Log hệ thống để phát hiện lỗi tiềm ẩn bằng AI

Chuyên ngành: Hệ thống Thông tin

Mã chuyên ngành: 8.48.01.04

Họ và tên học viên: Nguyễn Đức Tiệp

Họ và tên người nhận xét: Lê Thanh Hà

Học hàm, học vị: PGS. TS

Chuyên ngành: Khoa học máy tính

Cơ quan công tác: Trường Đại học Công nghệ, Đại học Quốc gia Hà Nội

Số điện thoại: 0983 692 592 E-mail: lthavnu@gmail.com

NỘI DUNG NHẬN XÉT

I/ Cơ sở khoa học và thực tiễn, tính cấp thiết của đề tài:

Log là dữ liệu ghi nhận quá trình vận hành và là nguồn thông tin quan trọng giúp chẩn đoán sự cố, giám sát hiệu suất và đảm bảo an toàn hệ thống. Việc phát hiện sớm lỗi tiềm ẩn có ý nghĩa quan trọng đối với vận hành hệ thống. Nhưng với khối lượng log khổng lồ, việc phát hiện các mẫu bất thường đòi hỏi các công nghệ tự động hóa và TTNT.

II/ Nội dung của đề án tốt nghiệp, các kết quả đã đạt được:

Luận văn đã xây dựng và đánh giá phương pháp phát hiện bất thường trong log hệ thống sử dụng hai mô hình chính là Autoencoder và LogBERT. Kế quả cho thấy LogBert vượt trội về độ chính xác.

III/ Những vấn đề cần giải thích thêm:

Luận văn cần được trình bày như một tài liệu khoa học (hiện tại đang trình bày như một tài liệu kỹ thuật). Có phần nền tảng kiến thức về tổng quan bài toán phát hiện lỗi hệ thống, kiến trúc autoencoder, LogBERT và các biến thể liên quan. Tại sao tác giả lại sử dụng hai mô hình này? Có so sánh kết quả với các nghiên cứu đã có từ trước ví dụ:

- Guo, H., Yuan, S., & Wu, X. (2021). LogBERT: Log anomaly detection via BERT.

- Chen, S., & Liao, H. (2022). BERT-Log: Anomaly detection for system logs based on pre-trained language model. Applied Artificial Intelligence, 36(1), e2145642.

Luận văn mới chỉ thử nghiệm với 2 bộ dữ liệu benchmark chưa tương xứng với tiêu đề của luận văn là "Phân tích log hệ thống để phát hiện lỗi tiềm ẩn".

Bộ dữ liệu HDFS và BGL phù hợp với các lớp ứng dụng thực tiễn nào?

Triển khai thực tế trên nền hệ thống nào, đặc điểm dữ liệu giống và khác với HDFS và BGL thế nào?

Tại sao lựa chọn Autoencoder và LogBERT với bài toán này? Có đối sánh với các mô hình khác không?

Trong phần kết luận "Hệ thống được thiết kế với chi phí hợp lý, …, có thể mở rộng theo chiều rộng hoặc chiều ngang" có nghĩa là gì? Các kết quả thí nghiệm nào minh chứng cho điều này?

IV/ Kết luận:

Đồng ý cho phép học viên bảo vệ đề án tốt nghiệp.

Ngày 29 tháng 01 năm 2026

NGƯỜI NHẬN XÉT

Lê Thanh Hà

CỘNG HÒA XÃ HỘI CHỦ NGHĨA VIỆT NAM

Độc lập – Tự do – Hạnh phúc

BẢN NHẬN XÉT ĐỀ ÁN TỐT NGHIỆP THẠC SĨ

(Dùng cho người phản biện)

Tên đề tài đề án tốt nghiệp: PHÂN TÍCH LOG HỆ THỐNG ĐỂ PHÁT HIỆN LỖI TIỀM ẨN BẰNG AI

Chuyên ngành: HỆ THỐNG THÔNG TIN

Mã chuyên ngành: 8.48.01.04

Họ và tên học viên: Nguyễn Đức Tiệp

Họ và tên người nhận xét: Đỗ Xuân Chợ

Học hàm, học vị: PGS.TS

Chuyên ngành: Công Nghệ Thông Tin

Cơ quan công tác: Khoa An Toàn Thông Tin, Học Viện Công Nghệ Bưu Chính Viễn Thông

Số điện thoại: 0965068868....E-mail: chodx@ptit.edu.vn

NỘI DUNG NHẬN XÉT

I/ Cơ sở khoa học và thực tiễn, tính cấp thiết của đề tài:

Phân tích log hệ thống là một lĩnh vực đã được quan tâm từ lâu trong ngành CNTT. Trong giai đoạn đầu, các phương pháp chủ yếu dựa trên rule-based hoặc phân tích thống kê để phát hiện mẫu lỗi lặp lại. Với sự phát triển mạnh mẽ của học máy, đặc biệt trong xử lý ngôn ngữ tự nhiên (NLP), nhiều mô hình đã được ứng dụng nhằm trích xuất đặc trưng từ log và xây dựng cơ chế cảnh báo thông minh. Một số hướng nghiên cứu nổi bật gồm sử dụng clustering để gom nhóm log tương đồng nhằm phát hiện log bất thường, áp dụng deep learning như Autoencoder, LSTM, Transformer để học biểu 2 diễn ngữ nghĩa log, và kết hợp phân tích chuỗi thời gian với học không giám sát để phát hiện hành vi lệch chuẩn. Các công trình gần đây đã chứng minh tiềm năng của mô hình học sâu không giám sát trong việc phát hiện lỗi chưa từng thấy trước đó, tuy nhiên vẫn tồn tại các thách thức như xử lý log đa nguồn và đa định dạng, khả năng tổng quát hóa hạn chế và khó khăn khi triển khai trong môi trường thực tế.

II/ Nội dung của đề án tốt nghiệp, các kết quả đã đạt được:

Đề án gồm 03 chương như sau:

• Chương 1 trình bày nền tảng lý thuyết và bối cảnh nghiên cứu về bài toán phát hiện bất thường trong log hệ thống. Các nội dung cụ thể trong chương gồm: Đặc điểm của log hệ thống; Phân loại các loại lỗi; So sánh các phương pháp phát hiện bất thường từ dựa trên quy

tắc đến dựa trên trí tuệ nhân tạo; Pipeline trí tuệ nhân tạo cho phát hiện bất thường; Lựa chọn Autoencoder và LogBERT; Các nghiên cứu liên quan. Các nội dung chương này tạo nền tảng cho việc xây dựng mô hình trong

Chương 2. • Chương 2 trình bày quá trình xây dựng mô hình phát hiện bất thường trong log hệ thống sử dụng kỹ thuật trí tuệ nhân tạo. Các nội dung cụ thể trong chương gồm: Xây dựng Pipeline xử lý dữ liệu Log; Lựa chọn mô hình Autoencoder và LogBERT; Phân đoạn từ với hai phương pháp phù hợp cho từng mô hình (word-level cho Autoencoder, subword-level cho LogBERT); và (3) Biểu diễn vector với Word2Vec cho Autoencoder và BERT embeddings cho LogBERT; Huấn luyện mô hình với các tập dữ liệu HDFS và BGI; Thử nghiệm mô hình và tinh chỉnh các tham số theo các tiêu chí đánh giá.

• Chương 3 trình bày kết quả đánh giá các mô hình đã xây dựng và triển khai hệ thống pipeline thời gian thực. Các nội dung chính gồm: Đánh giá cả hai mô hình Autoencoder và LogBERT trên hai tập dữ liệu HDFS và BG; Xây dựng và triển khai Pipeline theo thời gian thực, sử dụng Apache Kafka, Docker container; Thực hiện kiểm thử, đánh giá hệ thống

III/ Những vấn đề cần giải thích thêm:

1. Phạm vi của đề án là các lỗi phần mềm quá chung chung. Nên xác định và định nghĩa lại giới hạn cho phạm vi phát hiện.

2. Bố cục đề án còn chưa logic và rời rạc. Cụ thể các nội dung chương 2 và chương 3 có thể ghép lại với nhau thành chương 3. Còn chương 2 chỉ nên mô tả về thuật toán và phương pháp NLP.

3. Phần liên quan đến kết quả thực nghiệm, nên chạy bổ sung thêm một số kịch bản với các tham số khác nhau khác.

IV/ Kết luận:

Đồng ý cho phép học viên bảo vệ đề án tốt nghiệp.

Ngày 29 tháng 01 năm 2026

NGƯỜI NHẬN XÉT

Đỗ Xuân Chợ