클라우드 모니터링 시스템의 성능 향상을 위한 딥러닝을 이용한 시계열 데이터 예측 연구

김동완[1], 홍두표[1], 신용태[2]
[1]숭실대학교 컴퓨터학과 석사과정
[2]숭실대학교 컴퓨터학부 교수
kk5339@soongsil.ac.kr, envydp99@soongsil.ac.kr, shin@ssu.ac.kr

Deep Learning-based Time Series Data Prediction Research for Performance Enhancement in Cloud Monitoring Systems

[1]Dept. of Computer Science, Soongsil University
[2]Dept. of Computer Science, Soongsil University

요 약

클라우드 시장의 성장과 마이크로 서비스 접근식이 제기됨에 따라 IT인프라를 관리하기 위한 연구가 최근 활발히 이루어지고 있다. 하지만 고도화 및 분산된 환경에서 관찰 가능성 응용을 확보하기 어렵다는 문제점을 가지고 있다. 따라서 본 연구에서는 모니터링 시스템을 통한 데이터 분석 중 수집한 데이터의 분석이 난해하다는 문제를 해결하기 위한 방법을 제안한다. 제안된 방법은 NAB 데이터셋을 대상으로 STUMPY를 이용하여 데이터를 시각화하고, CNN을 이용하여 분류 작업을 수행한다. 분류를 수행한 데이터셋은 이상치 데이터와 이상 전조 데이터, 정상 데이터셋으로 분류하여 데이터셋을 구성한다. 구성한 학습 데이터셋에 대해 훈련을 마친 딥러닝 모델은 부하 테스트 환경에서 수집한 데이터에 대한 그래프 패턴을 분석하여 이상치 데이터와 이상 전조 데이터를 탐지한다.

1. 서론

IT 및 경영 컨설팅 리서치 회사인 가트너는 관찰 가능성 응용(Applied Observability)을 2023년 10대 전략 기술 트랜드 중 하나로 선정했다[1]. 관찰 가능성 응용은 시스템의 동작과 성능을 이해하기 위해서 다양한 도구와 기술을 사용하는 것을 의미한다. 기술적인 측면에서는 모니터링 메트릭의 수집, 로그의 저장과 분석, 추적을 할 수 있는 시스템을 구축하는 것이다. 시스템 관리의 가장 효과적인 방법 중 하나는 네트워크에 모니터링 에이전트를 배치하고 시스템 로그 데이터를 수집하는 것이다. 모니터링의 목적은 지속적인 감시, 감찰을 통해 대상의 상태나 가용성,변화 등을 확인하고 대비하는 것이다. 관리자는 모니터링 시스템을 통해 예기치 못한 상황과 오류를 대비하고 극복한다. [2]의 연구에서 대규모 네트워크 환경에서 모니터링을 통해 시스템의 안정성과 신뢰성을 유지하는 것은 어렵다고 지적했다. 많은 연구에서 모니터링을 위해 네트워크 이벤트와 관련된 로그 메시지간의 상관 관계를 분석하거나 인과 관계를 추출하는 방법을 제안하였다. 하지만 로그 분석의 문제점은 로그 데이터간 인과 관계를 파악하는 것이 난해하다는 점이다. 로그 데이터의 타임스탬프는 인과 관계를 파악하는데 도움이 되지만, 동시간대 발생한 로그 데이터간 인과 관계를 항상 지니고 있는것은 아니다[3].

본 논문에서는 이러한 복잡성을 피하기 위한 대안적인 접근법을 제안한다. 시계열 데이터 분석 연구의 일환으로 데이터셋 확보를 위해 부하 테스트를 통해 발생한 이상 데이터를 시각화 했을 때, 상황에 따른 특정 패턴을 보인다는 것을 발견하였다. 따라서 제안 기법은 수집한 데이터를 데이터 시각화 도구를 통해 이미지화하여, 뛰어난 이미지 인식 성능을 보이는 합성곱 신경망(Convolutional Neural Network)[4]을 사용해 이상치 데이터의 패턴을 탐지한다. 이로써, 다양한 서비스 상호작용으로 인한 이벤트 모니터링 된 데이터에서 관계를 식별하고 비정상 작동의 전조를 특성화한다. 본 논문의 구성은 다음과 같다. 2장에서 로그 데이터 분석 연구와 이상치 탐지 및 예측에 관한 관련 연구에 대해 소개한다. 3장에서 학습 데이터셋 확보를 위해 구성한 MSA 기반의 인프라에 대해 서술한다. 4장에서 딥

러닝을 이용한 시계열 데이터 예측 연구에 대해 제안한다. 제안 방법으로 NAB[5] 데이터셋을 파이썬 라이브러리의 일종인 STUMPY[6]를 사용하여 데이터를 시각화하고, CNN을 이용하여 분류 작업을 수행한다. 분류를 수행한 데이터셋은 이상치 데이터와 이상 전조 데이터, 정상 데이터셋으로 분류하여 데이터셋을 생성한다. 생성 데이터셋의 패턴은 딥러닝 알고리즘을 통해 수행되어 이후 실시간으로 모니터링 시스템이 데이터를 수집했을 시, 서비스 이상에 대한 가능성을 탐지한다.

2. 관련연구

일반적인 클라우드 모니터링 시스템에서 인프라 및 애플리케이션을 대상으로 수집한 로그 데이터에 대한 분석은 상관 및 회귀 분석 등을 통해 수행한다. [7]의 연구는 복잡한 인프라 구성 없이 체계적인 시스템에서 시계열의 상관관계를 사용해 데이터 센터를 대상으로 한 경량 이상 탐지 도구를 제안한다. 인프라 노트 메트릭과 애플리케이션이 상관 관계가 있음을 가정하여 측정값이 임계값 아래로 떨어질 시, 노드의 이상을 감지한다. 이러한 기법은 시계열의 시간 차원을 고려하지 않아 비선형 관계를 찾기 어렵다는 한계가 있다. NAB(Numenta Anomaly Benchmark) 데이터셋은 시계열 데이터 분석을 위한 벤치마크 데이터셋이다. Numenta Inc.에서 제공하는 오픈소스 라이브러리인 HTM(Hierarchical Temporal Memory)의 성능 평가를 위해 개발되었으며, 다양한 이상 감지 알고리즘의 성능 평가를 위한 데이터셋으로도 활용된다. NAB 데이터셋은 CPU 사용률, 온도, 습도, 전기 사용량 등의 다양한 시계열 데이터가 있다. 또한 각 데이터는 이상 감지에 대한 라벨링 정보를 포함한다. STUMPY는 Python 언어로 작성된 오픈소스 라이브러리 중 하나로, 시계열 데이터의 다양한 분석 기능을 제공하는 라이브러리이다. 시계열 데이터의 패턴을 찾고, 유사한 패턴을 갖는 부분을 찾을 수 있다. Stumpy는 주로 음악, 영상, 센서 데이터와 같은 시계열 데이터 분석에 사용되며 이를 위해 아래와 같은 기능을 제공한다.

1) Time series motif discovery : 시계열 데이터에서 유사한 패턴을 찾는 기능을 제공한다. 이를 통해 주어진 시계열 데이터에서 반복 패턴을 찾는다.

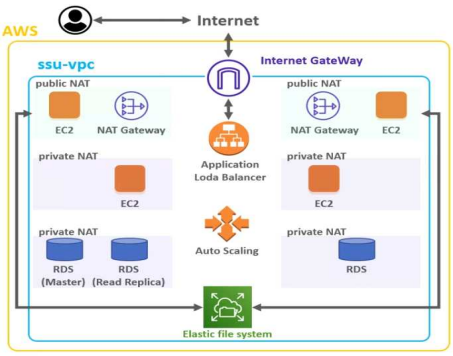
2) Time series segmentation : 시계열 데이터를 세그먼트로 나누는 기능을 제공한다. 이를 통해 시계열 데이터의 구간별 특성을 파악한다.

3) Time series motif discords : 시계열 데이터에서 이상치를 찾는 기능을 제공한다. 이를 통해 시계열 데이터에서 예외적인 동작을 파악한다.

4) Matrix Profile : 시계열 데이터의 상관관계를 분석하는 기능을 제공한다. 이를 통해 주어진 시계열 데이터에서 주기성 또는 특이한 패턴을 찾는다.

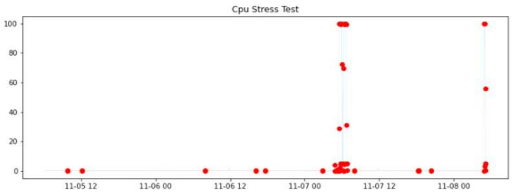
3. 실험 및 연구 환경 구성

아래 (그림 1)은 연구 수행을 위해 구성한 MSA 기반 인프라이다. 모니터링 시스템이 수집한 시계열 데이터를 시각화 했을 때, 특정 상황에 대한 패턴이 나타남을 보여주기 위한 환경이다. AWS(Amazon Web Services)의 Ec2를 이용하여 서버 다중화, 로드 밸런서와 오토 스케일링을 적용하여 실무와 유사한 환경을 구성했다.



(그림 1) MSA 기반 인프라 구성

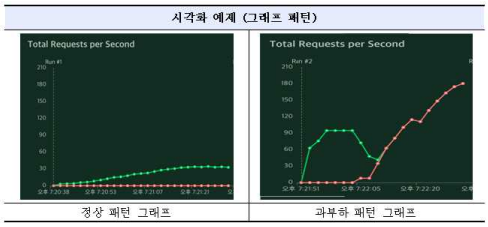
다음 (그림 2)는 구성한 인프라에 대해 CPU 사용률에 대한 부하 테스트를 수행했을 시, 수집한 데이터를 시각화한 것이다. 그래프의 x축은 시계열 데이터를 수집한 시각이며, y축은 CPU 사용률(%)이다. 붉은 색 점은 LSTM-AutoEncoder 알고리즘에 의해 이상치라고 판단되는 부분을 나타낸 것이다.



(그림 2) CPU에 대한 부하테스트 및 이상치 탐지 수행 결과

아래 (그림 3)은 부하 테스트를 통해 접속자 수를 순간적으로 늘렸을 때 수집한 시계열 데이터를 시각화 한 것이다. 위와 같은 실험을 통해 실무와 유사한 환경에서 부하 테스트를 수행했을 때 수집한 시

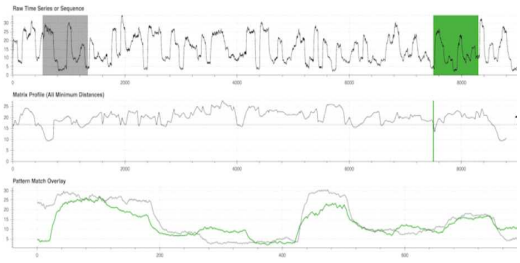
계열 데이터를 시각화할 시 상황에 따라 특정 패턴이 나타남을 알 수 있다.



(그림 3) 정상 패턴과 과부화 패턴 그래프

4. 제안 기법

제안 하는 기법은 애플리케이션 집합 측면에서 네트워크 흐름을 분류하기 위해 CNN 아키텍처를 이용한 방법을 적용한다[8]. 선행연구에 따르면 시계열 데이터를 처리하는 과정에서 시간 차원을 고려하지 않음으로 파생되는 문제점들에 대하여 설명하였다. NAB 데이터셋은 이상치 데이터에 대한 라벨링 정보가 포함되어있다. NAB 데이터셋에서 이상치 데이터의 이전 타임 스탬프에 대해 이상 전조 정보를 추가하여, 라벨링 정보와 타임 스탬프를 기준으로 이상치 데이터와 이상 전조 데이터, 정상 데이터셋으로 분류한다. 이후 시계열 데이터를 시각화하고 CNN의 INPUT 값으로 데이터 전처리 작업을 수행할 시, 이미지의 인덱스 값과 타임스탬프, 라벨링된 정보에 대해 key, value 값으로 저장하고 CNN 모델에 학습을 수행한다. 훈련된 모델은 이후 수집한 STUMPY 라이브러리의 단위 셀(cell)에 의해 연속된 그래프에 대한 분석을 수행하고, 분석 결과에 따라 이상치와 이상 전조, 정상치 데이터를 구분한다.



(그림 4) 가시화한 시계열 데이터의 패턴 파악을 위한 셀(cell) 예시

이로써 제안 기법은 실시간 모니터링 시스템에서 데이터를 수집했을 시, 서비스 이상에 대한 가능성을 탐지한다.

5. 결론

본 연구는 인프라에서 장애가 발생할 시, 수집하는 이상치 데이터가 상황에 따른 특정한 패턴을 보인다는 점을 주목하였다. 로그 데이터의 인과 관계 파악 및 분석의 복잡성 문제를 회피하고 시간 관계를 고려하기 위해 이상치 데이터의 이전 타임스탬프에 대해 이상 전조 라벨링 작업을 수행하여 딥러닝 모델에 훈련시켰다. STUMPY를 통한 시계열 데이터 분석 및 예측 연구를 제안하였다.

본 연구는 과학기술정보통신부 및 정보통신기획평가원의 대학ICT연구센터육성지원사업의 연구결과로 수행되었음" (IITP-2023-2020-0-01602)

참고문헌

[1] Gartner, "Top 10 Strategic Technology Trends for 2023," Gatner, Inc, 2022.

[2] Xu Chen *et al.,* "Automating Network Application Dependency Discovery:Experiences, Limitations, and New Solutions," *OSDI'08: Proceedings of the 8th USENIX conference on Operating systems design and implementation*, pp. 117-130, 2008.

[3] Satoru Kobayashi, "Mining Causality of Network Events in Log Data," *IEEE Transactions on Network and Service Management*, Vol.15(1), pp.53-67, 2018.

[4] D. C. Ciresan et al, "Flexible, High Performance Convolutional Neural Networks for Image Classification," *Proc. of the Twenty-Second International Joint Conference on Artificial Intelligence*, pp. 1237-1242, 2011

[5] https://github.com/TDAmeritrade/stumpy

[6] Alexander Lavin & Subutai Ahmad, "Evaluating Real-Time Anomaly Detection Algorithms -- The Numenta Anomaly Benchmark," *IEEE 14th International Conference on Machine Learning and Applications (ICMLA)*, 2015.

[7] S. Barbhuiya *et al.*, "A lightweight tool for anomaly detection in cloud data centres," *in CLOSER*, 2015.

[8] M. Lopez-Martin *et al.*, "Network traffic classifier with convolutional and recurrent neural networks for internet of things," *IEEE Access*, 5, 18042 - 18050, 2017.