로그 및 자원 분석을 통한 VNF 고장 예측에 관한 연구

남석현, 홍지범, 유재형, 홍원기
포항공과대학교 컴퓨터공학과

{obiwan96, hosewq, jhyoo78, jwkhong}@postech.ac.kr

A Study on VNF Failure Prediction through Log and Resource Analysis

Sukhyun Nam, Jibum Hong, Jae-Hyoung Yoo, James Won-Ki Hong

Department of Computer Science and Engineering, POSTECH

요 약

본 논문은 SDN/NFV 환경에서 VNF 및 서버의 고장을 예측하기 위한 기계 학습 모델 및 서비스를 제안한다. 제안하는 모델은 SDN/NFV 환경에서 각 VNF 로부터 발생하는 시스템 로그와 collectd 를 통해 수집되는 자원 사용량 및 트래픽 정보를 함께 활용하는 병렬 CNN 을 사용한다. 병렬 CNN 의 각 채널은 자원 사용량 데이터로부터 feature 를 추출하는 RNN 모델과 로그 데이터로부터 feature 를 추출하는 합성곱층으로 이루어진다. 이 때, 로그 데이터는 사전 처리를 통해 임베딩 행렬로 변환시키며, 추출된 feature 는 softmax 층의 입력으로 사용된다. 제안하는 고장 예측 모델은 병렬 CNN 을 통해 VNF 및 서버에 고장이 발생하기 전에 발생하는 장애 관련 징후로 고장을 예측하고 완화 작용 및 migration 을 통해 서비스 고장을 사전에 방지하는 기능을 제공한다.

Ⅰ. 서 론

오늘날의 네트워크는 그 규모와 구조가 점점 더 커지고 복잡해지고 있다. 소프트웨어 정의 네트워킹 (Software-Defined Networking, SDN)과 네트워크 기능 가상화 (Network Function Virtualization, NFV) 기술 등이 등장하면서 CAPEX/OPEX 를 절감시켰지만, 복잡해진 네트워크 구조로 인해 네트워크의 장애를 진단하고 관리하는 것이 더 어려워지고 있다. 이러한 이유로 SDN/NFV 환경에서 가상 네트워크 기능 (Virtualized Network Function, VNF)의 이상 탐지를 위한 연구가 진행되고 있으나, 서버와 가상 서버 및 VNF 의 고장을 사전에 예측하여 고장 발생 전에 조치를 취하는 기술에 대한 연구는 부족한 상황이다.

대부분의 서버 및 네트워크 장비들은 실시간 로그 (예: syslog) 출력 기능을 제공한다. 최근 서버 및 네트워크의 소프트웨어 구조가 복잡해짐에 따라 로그의 양도 비례하여 늘어나고 있다. 로그는 장비의 상태를 가장 잘 나타내는 데이터 중 하나이며, 장애 발생 시 관련 로그가 발생하기 때문에 네트워크 관리에 로그를 활용하는 연구들이 일부 진행되고 있다 [1, 2, 4]. 하지만 대부분의 로그 데이터는 체계적으로 생성되지 않기 때문에 로그를 이해할 수 있는 전문가가 수작업으로 관리해야 한다는 문제가 있다. 로그 데이터를 자동으로 처리하는 것은 여전히 상당히 어려운 과제로 남아있다.

로그 분석에는 자연 언어 처리 (Natural Language Processing, NLP)의 한 분야인 문장 분류 (sentence classification) 기법을 적용시킬 수 있다. 문장 분류는 영화 리뷰와 같은 텍스트 문서로부터 저자가 해당 문서의 주제에 대해 표현한 의견을 판단하거나, 스팸 메일 분류와 같이 미리 정의된 기준에 따라 문서를 분류하는 분야이다. 단어의 순서가 중요하여 순환 신경망 (Recurrent Neural Network, RNN)이 강세를 보이는 다른 NLP 분야와 달리 각 단어가 분류 결과에 얼마나 영향을 미치는지를 분석하기 위하여 합성곱 신경망 (Convolution Neural Network, CNN)을 주로 사용하는 분야이다 [3].

본 논문에서는 SDN/NFV 환경에서 각 VNF 들의 로그와 시스템 자원 사용량 데이터를 모두 입력 값으로 받는 병렬 CNN 모델을 이용한 서버 고장 예측 모델을 제안한다. 제안한 모델은 입력 데이터를 기준으로 일정 시간 뒤에 VNF 나 물리 서버에서 고장이 발생할 것을 예측한다.

Ⅱ. 관련 연구

선행연구 [3]은 문장 분류 문제에서 CNN 모델을 처음으로 제안하였다. 해당 연구는 문장을 CNN 의 입력 피처로 사용하기 위해 단어를 고밀도 벡터 (dense vector)로 표현하는 기법인 워드 임베딩 (word embedding) 기법 [6]을 사용하였으며, 생성된 임베딩 벡터들로 concatenation 과정을 거쳐 문장 벡터를 생성하였다. 생성된 문장 벡터는 합성곱층 (convolution layer)을 거치는데, 이 때 합성곱층에서 쓰이는 filter 의 크기는 임베딩 벡터의 차원인 h 와 필터의 크기 k 의 곱인 hk 차원의 벡터를 사용하였다. 생성된 feature 에 대해 max pooling 을 적용하여 filter 의 수만큼 feature 값이 생성되고 이에 대해 softmax layer 을 통해 최종 분류를 하였다. 실험 결과 제안한 CNN 모델의 성능이 다른 기계학습 기법에 비해 매우 뛰어나지는 않았지만, 가장 간단한 형태의 CNN 을 활용해 만든 모델임에도 불구하고 다른 모델들과 비교하여 우수한 성능을 보여 문장 분류 문제에 CNN 이 적합함을 보였다.

선행연구 [4]는 시뮬레이션 환경의 무선 통신 시스템에서 로그 데이터를 수집하여 일정한 격차 (gap) 이후의 로그에 에러 메시지가 포함되는지를 예측하는 CNN 모델을 제안하였다. 해당 모델은 로그 데이터에는 숫자와 구두점 등 쓸모 없는 글자가 많기 때문에 이를 제거하고, 어휘 사전에서 빈도수가 적은 단어들을 제거하는 전처리과정을 거친 후 사용하였다. 전처리 후의 로그는 워드 임베딩을 통해 임베딩 벡터로 변환하여 사용하였다. 생성한 임베딩 벡터에 대해 single channel 의 CNN 을 학습시킨 결과 예측 격차가 2000 일 때 정확도 0.7 을 웃도는 성능을 보였으나 예측 시점과의 격차의 기준을 로그 개수로 하여 예측 시점이 일정하지 않다는 한계점이 존재한다.

여러 행렬을 입력으로 받아 각각 합성곱층을 거친 후에 concatenation 하여 사용하는 형태의 CNN 을 multi-channel CNN 이라고 한다. 문장 분석 연구에서 여러 입력 값을 함께 사용하기 위해 multi-channel CNN 을 사용하기도 한다 [3, 5]. 선행연구 [3]는 두 가지 워드 임베딩을 사용하기 위해 2-channel CNN 을 학습시켰다. 선행연구 [5]는 한국어 감성분석에서 형태소 기반의 CNN 을 발전시켜 음절, 자소기반으로 생성한 워드 임베딩을 사용하기 위해 3-channel CNN 을 학습하였다. Multi-channel 을 통해 생성된 다중 feature 들은 softmax layer 에서 분류에 도움이 되는 정보만 남도록 학습되기 때문에 분류에 도움이 되는 다양한 입력 피처를 사용하면 더 높은 성능을 낼 수 있다.

Ⅲ. 병렬 CNN 기반 고장 예측 서비스

본 연구에서는 로그 데이터와 자원 사용량 데이터를 함께 사용하는 병렬 CNN 을 이용하여 물리 서버의 고장을 예측하고 경고하는 서비스를 제안한다. 제안하는 서비스는 예측 시기를 일정하게 하기 위해 시간을 기준으로 예측 시점과의 격차를 정한다. 제안하는 병렬 CNN 기반 VNF 고장 예측 서비스 구조는 그림 1 과 같다.

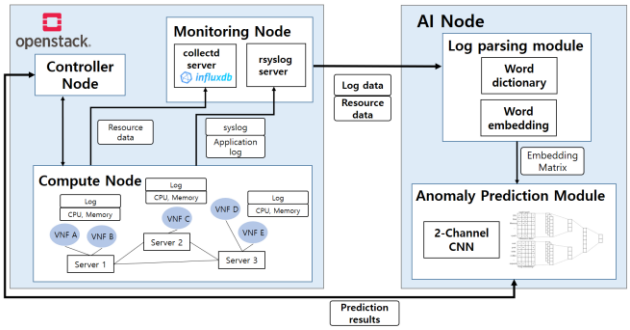

그림 1 병렬 CNN 기반 고장 예측 서비스 구조

제안하는 서비스는 클라우드 컴퓨팅 관리 시스템인 OpenStack 을 이용하여 구축한 NFV Infrastructure (NFVI)를 사용한다. 해당 NFVI 환경은 인공지능 기반의 NFV 관리 플랫폼 선행연구 [7]를 기반으로 이루어지며, 네트워크 서비스를 제공하기 위해 다양한 VNF (예: IDS, 방화벽) 들을 동작시킨다. 각 VNF 내부에서는 모니터링 데몬인 collectd [8]를 통해 CPU, 메모리 사용량과 같은 자원 사용량 데이터와 네트워크 트래픽 로드와 같은 네트워크 데이터를 실시간으로 수집하여 모니터링 노드로 전송한다. 로그 데이터는 로그 수집 데몬인 Rsyslog [9]를 통해 수집되어 모니터링 노드로 전송된다.

모니터링 노드는 수집된 데이터에서 timestamp 를 이용하여 정해진 window 크기만큼의 데이터를 추출하여 AI 노드로 전송한다.

AI 노드는 사전에 학습된 어휘 사전, 워드 임베딩, CNN 모델을 포함하는 노드이다. 그림 1 의 구조에서 추출된 로그 및 자원 사용량 데이터를 이용하여 진행된다. 해당 데이터는 window 크기만큼의 로그 및 자원 사용량 데이터를 입력 값으로, 일정 시간 뒤에 VNF 및 서버가 정상 작동하는지 여부를 출력 값으로 갖는 데이터이다. 어휘 사전 및 워드 임베딩은 추출된 전체 로그 데이터를 이용하여 생성된다. 어휘 사전은 로그 데이터 전체의 단어들의 빈도수가 계산된 데이터이다. 워드 임베딩은 Google 의 오픈 소스 프로젝트인 word2vec [10]을 이용하여 로그 데이터로 생성시킨다. 공개된 워드 임베딩이 아닌, 로그 데이터를 이용하여 생성한 워드 임베딩을 사용함으로서 로그 데이터 분석에 더 적합한 워드 임베딩을 사용할 수 있다.

AI 노드는 모니터링 노드로부터 전달 받은 로그데이터에 대해 전처리 작업을 거친 후 CNN 모델에 입력시킨다. 전처리 과정은 숫자와 구두점은 제거하고 단어들만 남긴 후 어휘 사전을 통해 로그에서 빈도수가 낮은 단어들은 Unknown 으로 태깅하는 과정이다. Unknown 단어들은 Out of Vocabulary (OOV) 벡터로 치환되는데, 이 때 OOV 벡터는 어휘 사전의 워드

임베딩 벡터들과 가장 멀리 있는 벡터로 사전에 생성된 벡터이다. 나머지 단어들은 사전에 학습된 워드 임베딩 벡터들로 치환된다. 그렇게 생성된 임베딩 데이터와 자원 사용량 데이터를 병렬 CNN 의 입력 값으로 사용한다. 사용하는 병렬 CNN 모델은 그림 2 와 같다.

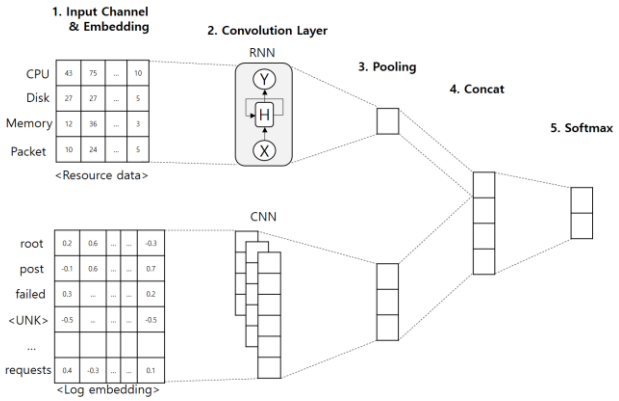

그림 2 병렬 CNN 구조

제안하는 CNN 모델은 자원 사용량 데이터와 로그 데이터를 모두 입력으로 받는 구조로서, multi-channel CNN 의 일종으로 볼 수 있으나, 한 채널은 RNN 을 사용하기 때문에 병렬 CNN 으로 표기하였다. 로그 데이터는 [3]에서 제안한 CNN 모델과 마찬가지로 합성곱층을 거쳐 feature 벡터가 추출된다. 자원 사용량 데이터는 기존의 연구들과 다르게 convolution layer 로서 RNN 을 활용하는데, 이는 자원 사용량 데이터는 로그 데이터와 달리 시간에 따른 값의 변화를 분석해야 하는 시계열 데이터이기 때문이다. 자원 사용량 데이터에서도 마찬가지로 RNN 을 통하여 feature 가 추출되면 두 channel 에서 추출된 feature 를 concatenation 을 통해 feature 벡터를 생성할 수 있다. 최종적으로 softmax layer 를 통해 고장인지 아닌지 판별할 수 있게 된다. 네트워크 장비 및 VNF 는 고장이 일어나기 전에 장애 관련 로그를 발생시킬 것이며, 이를 로그 및 자원 사용량 데이터 분석을 통해 예측할 수 있을 것이다. 본 모델을 통해 제안하는 모델은 일정 시간 이후에 서버에 고장이 생기는지 여부를 예측할 수 있을 것으로 기대된다.

Ⅳ. 결론 및 향후 연구

본 논문에서는 SDN/NFV 환경 관리를 위한 병렬 CNN 기반 고장 예측 모델 및 서비스를 제안하였다. 제안하는 모델은 NFV 환경에서 추출되는 로그 데이터와 자원 사용량 데이터를 기반으로 일정한 시간 격차 뒤의 물리 서버 및 VNF 의 고장 발생 여부를 학습한다. 학습된 모델은 NFV 환경에서 AI 노드에 포함되어 고장이 일어날 것을 미리 예측할 수 있으며, 고장이 예측될 경우 controller node 에 경고한다. 이를 통해 추후 controller 노드에서는 migration 및 auto-scaling 에 활용 가능하다.

ACKNOWLEDGMENT

이 논문은 2021 년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원(2018-0-00749, 인공지능 기반 가상 네트워크 관리기술 개발)과 2021 년도 정부(산업통상자원부)의 재원으로 산업기술평가관리원의 지원을 받아 수행된 연구임 (No.2009633, 초저지연 네트워크 서비스를 위한 SDN 기반 인공지능 관제 시스템 개발).

참 고 문 헌

[1] Q. Fu, J. Lou, Y. Wang and J. Li, "Execution Anomaly Detection in Distributed Systems through Unstructured Log Analysis," 2009 Ninth IEEE International Conference on Data Mining, 2009, pp. 149-158.

[2] S. He, J. Zhu, P. He and M. R. Lyu, "Experience Report: System Log Analysis for Anomaly Detection," 2016 IEEE 27th International Symposium on Software Reliability Engineering (ISSRE), Ottawa, ON, Canada, 2016, pp. 207-218.

[3] Y. Kim, "Convolutional neural networks for sentence classification," In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2014, pp. 1746-1751.

[4] W. Ji, S. Duan, R. Chen, S. Wang and Q. Ling, "A CNN-based network failure prediction method with logs," 2018 Chinese Control And Decision Conference (CCDC), 2018, pp. 4087-4090.

[5] 김민, 변증현, 이충희, 이연수, “Multi-channel CNN 을 이용한 한국어 감성분석,” 제 30 회 한글 및 한국어 정보처리 학술대회 논문집, 2018, pp. 79-83.

[6] T. Mikolov, I. Sutskever, K. Chen, G. Corrado and J. Dean, "Distributed Representations of Words and Phrases and their Compositionality", In Advances on Neural information Processing Systems, 2013.

[7] 정세연, 이도영, 유재형, 홍원기, “인공지능 기반 NFV 관리 플랫폼,” In KNOM Conference 2019, pp. 40-42, May 2019.

[8] “collectd – The system statistics collection daemon,” Available : https://collectd.org/

[9] Adiscon GmbH, “The rocket-fast Syslog Server,” [Online]. Available: https://www.rsyslog.com/

[10] Google, “word2vec,” 2013. [Online]. Available: https://code.google.com/archive/p/word2vec/ .