TensorFlow 기계학습 도구를 이용한 응용 트래픽 분류

이성호, 심규석, 구영훈, 김명섭
고려대학교

{gaek5, kusuk007, gyh0808, tmskim}@korea.ac.kr

Application Traffic Classification using TensorFlow Machine Learning Tool

Sung-Ho Lee, Kyu-Seok Shim, Young-Hoon Goo, Myung-Sup Kim
Korea Univ.

요 약

오늘날 네트워크 환경이 증가하며 응용 및 서비스 별로 발생시키는 트래픽 패턴에 다양한 종류가 생성됨에 따라 트래픽의 응용 및 서비스 별로 분류할 수 있는 트래픽 분석 방법이 연구되고있다. 그러나, 기존의 시그니쳐 기반의 분석 방법은 트래픽 패턴 변화를 관리자가 인지 해야할 뿐만 아니라 시그니쳐를 추출하는 작업 또한 많은 시간과 비용이 발생한다. 따라서 본 논문은 OpenAI 툴 중 하나인 TensorFlow 를 이용한 기계 학습 기반의 응용 트래픽 분석 방법과 트래픽의 패턴 학습을 위한 다양한 속성들을 정의하고 실험을 통해 가장 효율적으로 응용 트래픽을 분석할 수 있는 속성을 제안한다. 제안하는 방법을 통해 설계된 트래픽 분석 모델을 실제 다양한 응용 트래픽에 적용했을 때 이전의 시그니쳐 기반의 방법과 동일한 수준의 분석 정확도를 갖는 동시에 응용 트래픽 분석률 또한 30%이상 향상된 결과를 얻을 수 있었다.

Ⅰ. 서 론

오늘날 네트워크 환경은 증대되고 있고, 그에 따라 네트워크를 이용하는 응용의 종류도 증가하고 있다. 이러한 다양한 응용은 각자 다른 패턴을 트래픽을 발생시킨다. 또한 각 응용은 개발자에 의해 사용자에게 고품질의 서비스를 제공하기 위해 지속적으로 트래픽 패턴을 변화시킨다. 네트워크 관리자는 이러한 트래픽 패턴을 이용하여 네트워크 내의 트래픽을 분류하고 분류 결과를 통해 QoS 조절 및 모니터링을 수행한다.

이렇듯 트래픽 패턴이 복잡해짐에 따라 다양한 트래픽 분석방법들이 연구되어 왔지만 모두 시그니쳐 기반의 사전에 정의된 규칙에 의존한 방법들이 주를 이뤄왔다.[1,2] 이러한 방법은 예측할 수 없는 패턴이나 새로운 패턴에 유연하게 대처할 수 없다는 한계점이 존재한다. 따라서 최근에는 기계 학습(Machine Learning)을 이용한 트래픽 분석 방법이 점차 대두되고 있다. 이미 다양한 기계 학습, 딥-러닝 기반의 분석 방법들이 특정 분야에 얽매이지 않고 광범위하게 사용되고 있고, 이러한 추세와 더불어 기업에서는 인공지능을 개발하고 적용할 수 있는 OpenAI 툴을 제공하고있다. 대표적으로 Google 의 TensorFlow 그리고 Microsoft 의 CNTK(Computational Network Toolkit)가 있다. 결과적으로 인공지능 기반의 다양한 학습 및 분석 방법을 이전보다 자유롭게 여러 분야에서 활용하고 적용할 수 있게 되었다.

이 논문은 2015 년도 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구(No.2015R1D1A3A01018057) 및 2016 년도 정부(미래창조과학부)의 재원으로 정보통신기술진흥센터의 지원을 받아 수행된 연구임 (No.B0101-16-0300, 사이버 공격의 사전 사후 대응을 위한 사이버 블랙박스 및 통합 사이버보안 상황분석 기술 개발)

본 논문은 이러한 추세에 맞춰 인공지능 기반의 기계 학습 알고리즘을 트래픽 분석 시스템에 적용하는 방법을 제안한다. 다양한 응용의 트래픽을 수집하고 각 응용 트래픽의 발생 패턴을 학습하기 위해 Goolge 에서 제공하는 TensorFlow 툴을 활용한다. 간단한 Softmax Regression 을 통해 트래픽 패턴에 맞는 분석 모델을 생성하고 이를 실제 검증 트래픽에 적용시켜 응용 분석률을 측정한다. 측정 결과 트래픽 분석 정확도와 분석률은 학습 속성(Learning Feature)정의 방법에 따라 많은 차이를 보였다. 따라서 본 논문에서는 실험을 통해 트래픽 패턴 학습을 위한 효율적인 속성을 정의하고 이전의 시그니쳐 기반 시스템과 비교를 통해 기계 학습 기반의 트래픽 분석 시스템의 효율성을 검증한다..

본 논문은 서론에 이어, 2 장에서 기계 학습 기반의 응용 트래픽 분석 과정과 시스템에 대해 언급한 뒤 3 장에서 학습을 통해 생성된 모델을 실제 검증 트래픽에 적용한 뒤 결과 비교를 통해 본 시스템의 타당성을 증명한다. 마지막으로 4 장에서 결론 및 향후 연구에 대해 언급한 뒤 본 논문을 마친다.

Ⅱ. 본론

본 장에서는 기계 학습 기반의 응용 트래픽 분석 방법에 대해 언급한다. 기계 학습 기반의 응용 트래픽 분석 시스템은 그림 1 과 같이 구성된다. 학습 모델 생성과 검증을 위해 5 가지 응용에 대한 트래픽을 수집하고 수집한 트래픽을 Train Set 과 Test Set 으로 나눈다. 응용 트래픽 분석 모델 생성을 위해 Train Set 트래픽에 Softmax Regression 을 적용해 학습 모델을 생성한다. Softmax Regression 은 데이터의 패턴 예측을 위한 회귀 분석 방법 중 하나로 Multiple

Classification 이 가능하기 때문에 트래픽의 패턴 분석, 예측 및 응용 식별에 가장 적합하다.

기계 학습 기반의 방법과 기존의 시그니쳐 기반의 응용 트래픽 분석 방법을 비교하기 위해 Train Set 트래픽을 시그니쳐 자동 생성시스템에 적용해 시그니쳐를 생성한다. 두 가지 방법의 분석률을 비교하기 위해 Test Set 트래픽을 Softmax 분석 모델과 생성된 시그니쳐를 이용해 분석한다.

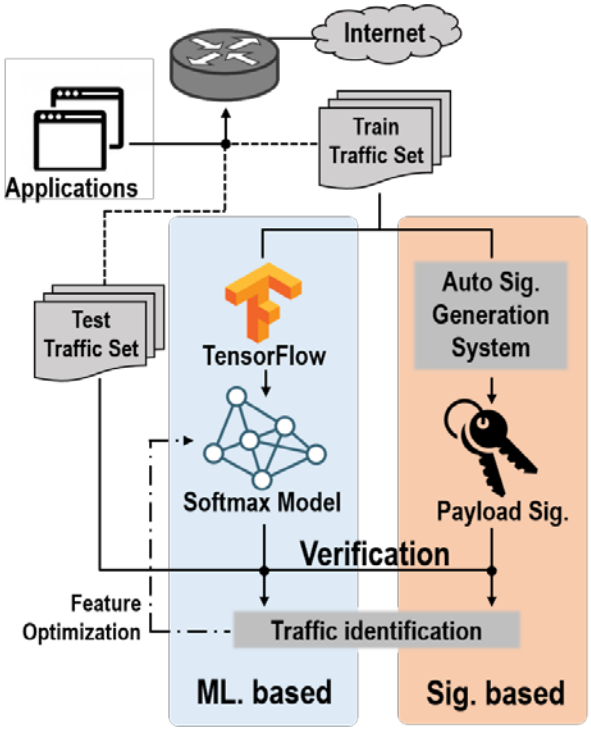

Figure 1. 기계 학습을 활용한 응용 트래픽 분석

기계 학습 기반의 방법과 기존의 시그니쳐 기반의 응용 트래픽 분석 방법을 비교하기 위해 Train Set 트래픽을 시그니쳐 자동 생성시스템에 적용해 시그니쳐를 생성한다. 두 가지 방법의 분석률을 비교하기 위해 Test Set 트래픽을 Softmax Regression 모델과 생성된 시그니쳐를 이용해 분석한다. Softmax Regression 모델은 학습 속성, 학습률, 학습 계층 수 등 설정에 따라 응용 분석 결과가 다르기 때문에 응용 트래픽 분석률을 최대로 높일 수 있는 최적의 설정 값을 찾는다.

Ⅲ. 실험

기계 학습 기반의 응용 트래픽 분석 시스템의 효율성과 분석률을 검증하기 위해 실험을 진행한다. 실험을 통해 기존의 시그니쳐 기반의 분석 방법과 성능을 비교하고 최적의 학습 속성을 찾는다. 실험을 진행하기 위해 학내 망에서 응용 트래픽을 수집했다. 사용된 응용은 가장 많이 사용되는 5 가지로 Daumpot, Melon, Kakaotalk, Facebook, Youtube 이다. 수집한 트래픽의 70%는 패턴 학습을 위해 30%는 검증을 위해 사용했다. 응용 트래픽 패턴 학습 설정 초기값과 변화량은 표 1 과 같다. 사용된 학습 속성은 트래픽의 플로우 크기, 지속 시간, 전체 패킷 수, Forward 패킷 수, Backward 패킷 수 이다. 5 가지 속성을 바탕으로 Softmax Regression 을 적용, Cost function 으로 Cross Entropy 값을 계산하고, Cost 값을 최소화하기 위해서 Gradient descent 알고리즘을 적용했다. 학습 횟수는 각 모델 별 모두 동일하게 30,000 번씩 수행했다. 실험의 결과 분석 척도는 Coverage 와 Accuracy 가 있다. Coverage 는 해당 모델 혹은 시그니쳐로 전체 응용 트래픽을 얼마나 분석할 수 있는지 나타낸다. Accuracy 는 분석된 트래픽 중 올바르게 응용 식별이 된 트래픽의 비율을 나타낸다.

Table 1. 응용 트래픽 학습 모델 별 분석 결과

α = Learning Rate, A = Flow Size, B = Duration
C = Total Packets, D = Forward Packets
E = Backward Packets, CO = Coverage (%)
AC = Accuracy (%)

Features	α	A	B	C	D	E	CO	AC
Payload Signature	NA	NA	NA	NA	NA	NA	69.4	98.1
Model.1	0.01	○	○	○			100	60.3
Model.2	0.01	○	○	○	○	○	100	34.3
Model.3	0.005	○	○	○			100	69.8
Model.4	0.005	○	○	○	○	○	100	71.7
Model.5	0.001	○	○				100	80.5
Model.6	0.001	○		○			100	78.4
Model.7	0.001	○	○	○			100	91.8
Model.8	0.001			○	○	○	100	50.7
Model.9	0.001	○			○	○	100	81.6
Model.10	0.001	○	○	○	○	○	100	96.9

Ⅳ. 결론 및 향후 연구

본 논문은 기계 학습을 이용한 응용 트래픽 분석 방법을 제안하였다. 실험을 통해 기계 학습을 활용한 응용 트래픽 분석 방법은 기존의 시그니쳐 기반의 방법과 비교했을 때 분석률 측면에서 보다 효율적인 것을 확인할 수 있었다. 또한 기계 학습을 활용한 응용 트래픽 분석 방법이 의미 있다는 것을 증명하였다.

향후 다양한 기계 학습 알고리즘을 활용하여 보다 정확하고 정교한 검증 실험을 진행할 계획이다.

참 고 문 헌

[1]심규석, 윤성호, 이수강, 김성민, 정우석, 김명섭, “네트워크 트래픽 분석을 위한 Snort Content 규칙 자동 생성”, Vol.40, No.04, April, 2015, pp666-677.

[1] Sung-Ho Yoon, Kyu-Seok Shim, Su-Kang Lee, and Myung-Sup Kim , "Framework for Multi-Level Application Traffic Identification," Proc. of the Asia-Pacific Network Operations and Management Symposium (APNOMS) 2015, Busan, Korea, Aug. 19-21, 2015, pp.424-427.