통신망 최적화를 위한 기계학습 알고리즘을 적용한 응용 트래픽 분류

지세현 O, 박지태, 백의준, 김명섭

고려대학교

{sxzer, pjj5846, pb1069, tmskim}@korea.ac.kr

요 약

오늘날 인터넷이 발달함에 따라 다양한 포트번호를 사용하고 암호화된 페이로드를 가지는 패킷을 발생시키는 어플리케이션이 많이 등장하고 있다. 적합한 QoS(Quality of Service)를 제공하고 안전한 네트워크 환경을 만들기 위해서는 정확한 응용 트래픽 분류는 필수적이다. 기존의 응용 트래픽 분류 기법은 포트 혹은 페이로드 기반 기법이다. 그러나 다양한 포트번호를 사용하고 암호화된 페이로드를 가지는 패킷을 발생시키는 어플리케이션이 응용 트래픽 분류를 어렵게 하고 있다. 이러한 문제를 해결 하기 위해 기계학습 알고리즘을 적용한 응용 트래픽 분류 기법이 제시된다. 본 논문은 다양한 포트 번호를 사용하는 3 가지 응용 트래픽(Google, Kakaotalk, Naver)을 수집한 뒤, 기계학습 알고리즘인 CNN(Convolution Neural Network) 알고리즘을 적용한 응용 트래픽 분류 기법을 제안한다. 분류 기법을 통해 나온 모델의 분류정확도를 확인 함으로써 분류 기법의 효율성을 검증한다.

1. 서론

오늘날 인터넷이 발달하고 있고, 그에 따라 네트워크 트래픽 사용량이 증가하고 있다. 네트워크 트래픽 사용량이 증가됨에 따라 안전한 네트워크 환경을 만들기 위해서 효율적인 네트워크 망 관리를 위한 방안이 연구되고 있다. 네트워크 망 관리를 하기 위한 첫 단계로 응용 트래픽을 분류하는 것은 매우 중요하다[1]. 적합한 QoS 를 제공하고 안전한 네트워크 환경을 만들기 위해서는 정확한 응용 트래픽 분류를 하는 것은 필수적이다.

기존의 응용 트래픽 분류 기법은 포트 혹은 페이로드 기반 기법이다. 포트 기반 기법은 패킷에 적혀 있는 포트를 기준으로 분류한다. 하지만 최근 동적 혹은 예측 불가능한 포트번호를 사용하는 어플리케이션이 많이 등장함으로 인해 이 기법의 한계점을 나타내었다. 페이로드 기반 기법 또한 패킷의 내용을 깊이 분석함으로써 어플리케이션의 패턴을 잡아낸다. 그러나 이 기법 역시 암호화된 페이로드를 가지는 패킷을 발생시키는 어플리케이션의 등장으로 인해 분류 하는데 있어 한계에 부딪치고 있다. 따라서 이 변화를 대처할 수 있는 새로운 분류 기법이 요구되고 있으며, 그 해결책으로 기계학습 알고리즘을 적용한 응용 트래픽 분류 기법이 등장했다[2].

본 논문에서는 서론에 이어 2 장에서 Google 에서 개발한 Tensorflow 에 있는 기계학습 알고리즘인 CNN 알고리즘을 적용한 응용 트래픽 분류 기법을 제안한 뒤 3 장에서 제안하는 분류 기법을 적용 시킨 다양한 포트 번호를 사용하는 3 가지 응용 트래픽패킷에 대한 페이로드를 바탕으로 학습하여 분류 정확도를 측정하여 분류 기법의 효율성을 검증한다. 마지막으로 4 장에서 결론 및 향후 과제에 대해 언급한 뒤 논문을 마친다.

이 논문은 2015 년도 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업(No.2015R1D1A3A01018057)과 2017 년도 정부(과학기술정보통신부)의 재원으로 정보통신기술진흥센터의 지원을 받아 수행된 연구임(No. 2017-0-00513, Security Analytics 기반의 이기종 보안솔루션 위협 분석 및 대응 기술 개발)

2. 본론

본 장에서는 CNN 알고리즘을 적용한 응용 트래픽 분류 기법에 대해 언급한다. CNN(Convolution Neural Network) 알고리즘은 이미지 분류에 특화된 기계 학습 기반의 분류 알고리즘이다. 알고리즘을 적용한 분류 모델의 구조는 그림 1 과 같이 구성된다. 입력 받은 이미지에 대한 특징을 추출하기 위한 Convolution Layer 와 추출된 특징 값을 Fully Connected Layer 에 넣어서 분류를 한다. Convolution Layer 는 Feature map 을 추출하기 위한 Convolution 필터와 추출된 Feature map 에 적용하기 위한 활성 함수로 구성한다.

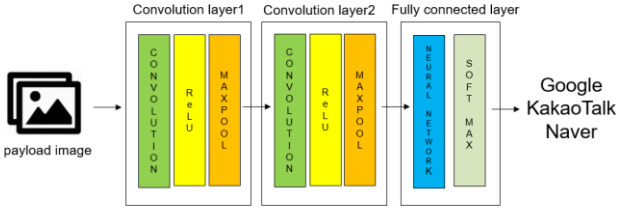

그림 1 CNN 분류 모델 구조

CNN 알고리즘을 적용한 응용 트래픽 분류 기법은 그림 2 의 과정을 거친다. 3 가지 응용 트래픽을 수집하고, 본 연구 팀이 개발한 프로그램인 Payload_Generator 를 이용해 패킷을 Flow With Packet 형태로 변환 한 뒤, 페이로드를 추출한다. 추출된 페이로드 값들은 CNN 분류 모델의 Input 으로 적합한 이미지 변환 과정을 거친 후 학습 시켜 분류 모델을 완성한다.

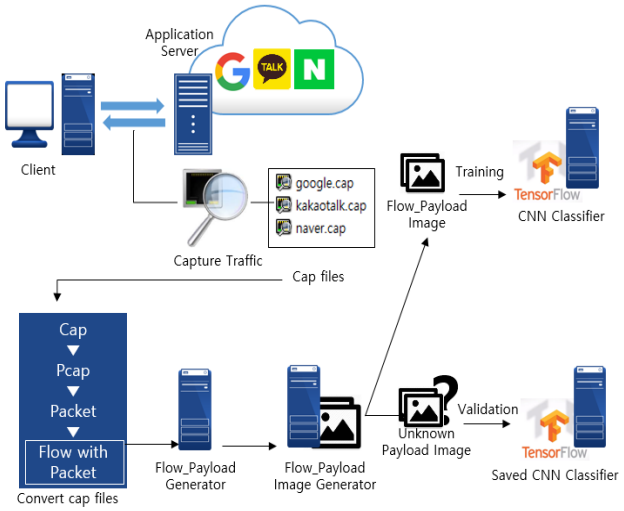

그림 2 응용 트래픽 분류 기법

CNN 알고리즘을 적용한 응용 트래픽 분류 모델의 성능을 분류정확도로 표현함으로써 모델의 효율성을 검증한다.

3. 실험

기계학습 알고리즘을 적용한 응용 트래픽 분류 기법의 적합성을 검증하기 위해 실험을 진행한다. MS Network Monitor 프로그램을 통해 3 가지(Google, Kakaotalk, Naver) 응용에 대한 패킷을 수집 한 뒤, 학습 데이터를 구성 하기 위해 본 연구 팀이 개발한 프로그램을 통해 패킷을 Flow With Packet 형태로 변환 한 뒤 Mnist 데이터 셋(28*28 형태의 이미지)과 유사하게 하기 위해 각 Flow With Packet 의 첫번째부터 784(28*28)번째까지의 값을 그림 3 과 같이 시각화를 한다.

그림 3 Flow With Packet 의 페이로드 이미지

각 응용 별로 1000 개의 Flow With Packet 의 페이로드 이미지를 CNN 알고리즘이 적용 된 분류 모델의 분류 정확도는 표 1 과 같다. 3 가지 응용 트래픽 중 2 개의 트래픽을 대상으로 한 이진 분류(Binary Classification) 실험과 3 가지 트래픽을 대상으로 한 다중 분류(Multinomial Classification) 모델의 학습 횟수는 각 10000 번씩 진행 하였다.

표 1 응용 트래픽 분류 결과

Application Traffic	Train Step	Accuracy
Google	10000	87.9%
Kakaotalk		
Google	10000	83.2%
Naver		
Kakaotalk	10000	86.5%
Naver		
Google	10000	89.0%
Kakaotalk		
Naver		

이진 분류의 경우 Google, Kakaotalk 트래픽을 대상으로 한 실험의 분류정확도가 가장 높게 나왔고, Google, Naver 트래픽의 분류정확도가 가장 낮게 나왔다. 3 가지 응용 트래픽에 대한 다중 분류 실험의 경우 89%의 분류 정확도가 나왔다.

3. 결론 및 향후연구

본 논문에서는 통신망 최적화를 위한 기계학습 알고리즘을 적용한 응용 트래픽 분류 방법을 제안하였다. 3 가지 응용 트래픽에 대해 기계학습 알고리즘인 CNN 알고리즘을 적용한 이진 분류 및 다중 분류 실험을 진행하였다. 실험을 통해 약 83~89%의 분류 정확도를 나타냄으로 써 분류 기법의 효율성을 검증하였다.

향후 연구에서는 같은 대상의 응용 트래픽을 대상으로 포트 혹은 페이로드 기반 분류 기법과 대조하여 제안하는 분류 기법의 타당성을 검증할 계획이다.

5. 참고문헌

[1] 정광본, 최미정, 김명섭, 원영준, 홍원기, "ML 알고리즘을 적용한 인터넷 애플리케이션 트래픽 분류," KNOM Review, Vol. 10, No. 2, Dec. 2007, pp. 39 - 52.

[2] 이성호, 심규석, 구영훈, 김명섭, "TensorFlow 기계학습 도구를 이용한 응용 트래픽 분류", 2016 년도 한국통신학회 추계종합학술발표회, 중앙대학교, 서울, Nov. 18, 2016, pp. 224-225.