Softmax Regression을 적용한 어플리케이션 트래픽 분류

지세현, 정우석, 박지태, 김명섭
고려대학교
{sxzer, hary5832, pjj5846, tmskim}@korea.ac.kr

Application Traffic Classification with Softmax Regression

Se-Hyun Ji, Woo-Suk Jung, Jae-Tae Park, Myung-Sup Kim
Korea Univ.

요 약

오늘날 인터넷이 발달함에 따라 네트워크 트래픽 사용량이 증가되고 있다. 폭주하는 트래픽 문제를 해결하기 위하여 단순하게 네트워크 회선을 업그레이드 하는 것은 많은 시간과 비용이 소비되며, 효율적인 관리를 할 수 없다. 따라서 적합한 QoS(Quality of Service) 를 제공하고 안전한 네트워크 환경을 만들기 위해서는 정확한 트래픽 분류가 요구되고 있다. 그럼에도 불구하고 다양한 포트 번호와 암호화된 페이로드를 가지는 패킷을 발생하는 어플리케이션이 많이 등장하여 이러한 어플리케이션의 등장이 기존의 시그니쳐를 기반으로 한 트래픽 분류를 어렵게 하고 있다. 또한 기존의 시그니쳐기반의 분석 방법은 트래픽 패턴 변화를 관리자가 인지해야 될 뿐만 아니라 시그니쳐를 추출 하는 작업 또한 많은 시간과 비용이 발생한다. 본 논문에서는 학내 망에서 주로 사용되는 대표적 3가지 어플리케이션 트래픽을 수집한 뒤, 본 연구팀이 개발한 feature 추출 프로그램을 통해 트래픽이 발생시키는 패킷의 약 300개의 feature 중 특정 feature를 선정하여 기계 학습 (Mechine Learning) 을 활용하여 Google 에서 제공하는 Tensorflow 의 Softmax Regression을 적용하여 featrue 학습을 통한 효율적인 어플리케이션 트래픽 분류 방법을 제안한다. 제안하는 방법을 통해 Softmax Regression 에 있어 효율적인 분류가 될 수 있는 Learning rate 와, Feature set 을 제시하고자 한다.

Ⅰ. 서 론

오늘날 인터넷이 지속적으로 발달하고 있고, 그에 따라 네트워크 트래픽 사용량이 증가되고 있다. 폭주하는 트래픽 문제를 해결하기 위해 가장 손쉬운 방법은 네트워크 회선을 업그레이드하는 것이다. 그러나 단순하게 네트워크 회선을 업그레이드하는 것은 많은 시간과 비용이 소비되며, 효율적인 관리를 할 수 없다. 따라서 어떤 시간대에 어떤 종류의 트래픽이 회선을 돌아다니고 있는지 정확히 파악하고 중요한 트래픽에 우선순위를 부여해 네트워크 속도를 높이기 위한 솔루션의 중요성이 점차 늘어가고 있다.

적합한 QoS(Quality of Service) 를 제공하고 안전한 네트워크 환경을 만들기 위해서 정확한 트래픽 분류가 요구되고 있다. 그럼에도 불구하고 다양한 포트 번호와 암호화된 페이로드를 가지는 패킷을 발생하는 어플리케이션이 많이 등장하고 있다. 트래픽 패턴이 복잡해짐에 따라 다양한 트래픽 분류방법들이 연구되어 왔지만 대부분 시그니쳐를 기반으로 사전에 정의된 규칙에 의존한 방법들이 주를 이루고 있다[1]. 트래픽 패턴 변화를 관리자가 인지해야 할 뿐만 아니라 시그니쳐를 추출하는 작업 또한 많은 시간과 비용이 발생하게 된다[2].

따라서 최근 기계 학습(Machine Learning) 을 이용한 트래픽 분석방법이 부상하고 있다. 이미 다양한 분야에서 기계 학습 (Machine Learning) 이 광범위하게 사용되고 있다. 본 논문에서는 이러한 추세에 맞춰 어플리케이션 트래픽이 발생시키는 패킷의 feature 를 기반으로 Google 에서 제공하는 기계학습 도구인 Tensorflow 를 활용하여, Softmax Regression 을 활용한 효율적인 어플리케이션 트래픽 분류방법을 제시하고자 한다.

학내 망에서 많이 사용되는 대표적 3가지 어플리케이션 트래픽을 수집하고, 본 연구팀이 개발한 feature 추출 프로그램(Pkt Feature Extraction) 을 통해 각 어플리케이션 트래픽이 발생시키는 패킷의 약 300개의 feature 를 추출한 뒤 Softmax Regression 을 통해 어플리케이션 트래픽 feature 를 학습하여 트래픽 분석률을 측정한다.

본 논문은 서론에 이어, 2장에서 Softmax Regression 을 이용한 트래픽 분류 과정과 시스템에 대해 언급한 뒤 3장에서 학습을 통해 생성된 모델을 실제 검증 트래픽에 적용한 뒤 Softmax Regression에 쓰이는 적절한 Learning rate 와 검증 결과를 통해 시스템의 효율성을 검증한다. 마지막으로 4장에서 결론 및 향후 과제에 대해 언급한 뒤 논문을 마친다.

Ⅱ. 본론

본 장에서는 feature 생성 후 기계학습(Machine Learning) 을 활용한 트래픽 분류 방법에 대해 언급한다. Softmax Regression을 활용한 어플리케이션 트래픽 분류 시스템은 그림 1의 과정을 거친다. 3가지 어플리케이션에 대한 트래픽(Google, Facebook, Kakaotalk) 을 수집하고, 본 연구팀에서 개발한 feature 를 추출하는 프로그램(Pkt Feature Extraction) 을 통하여 어플리케이션 트래픽이 발생시키는 패킷의 약 300가지의 feature 를 추출한다. 그 중 6가지의 feature(Flow Size, Duration, Protocol, TotalPktCount, ForwardPktCount, BackwardPktCount) 를 선정하여 Tensorflow 의 Softmax Regression 을 적용하여 Train Set 으로 학습 시킨 뒤 Test Set 을 통해 트래픽을 어플리케이션 별로 분류한다.

Softmax Regression은 Multinomial Classification 으로 두 개 이상의 그룹으로 나누기 위해 Binary Classification 을 확장한 개념이기 때문에

이 논문은 2016년도 정부(미래창조과학부)의 재원으로 한국연구재단의 지원을 받아 수행된 과학기술인문융합 연구사업 (No.NRF-2016M3C1B6929228)과 2015년도 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임(No.2015R1D1A3A01018057).

여러 가지 어플리케이션 트래픽의 패턴 분석, 예측 및 식별에 가장 적합한 방법이다. Softmax Regression 모델은 설정에 따라 분석 결과가 다르기 때문에 특히 주의해야 할 세 가지 사항이 있다. 첫째, 적절한 Learning rate 를 설정해야줘야 한다. Learning rate 를 너무 큰 값을 설정하게 되면 Cost function 을 사용하게 되는 Gradient descent 알고리즘에서 step 이 커져 적절한 cost 를 찾을 수 없게 되는 Overshooting 현상이 발생하게 되고, 반대로 너무 적은 값을 설정하게 된다면, step 이 작아져 Stops at local minimum 현상이 발생하게 된다. 둘째, 어플리케이션 트래픽이 발생시키는 패킷의 각 feature 들이 갖는 수치는 범위가 다르다. 따라서 각 feature들이 갖는 수치들의 편차를 줄이기 위한 Preprocessing 작업을 거쳐 학습해야한다. 마지막으로 Train Set과 Test Set의 적한 비율로 설정해야 한다. 보편적으로 Train Set은 80%, Test Set은 20%로 설정한다.

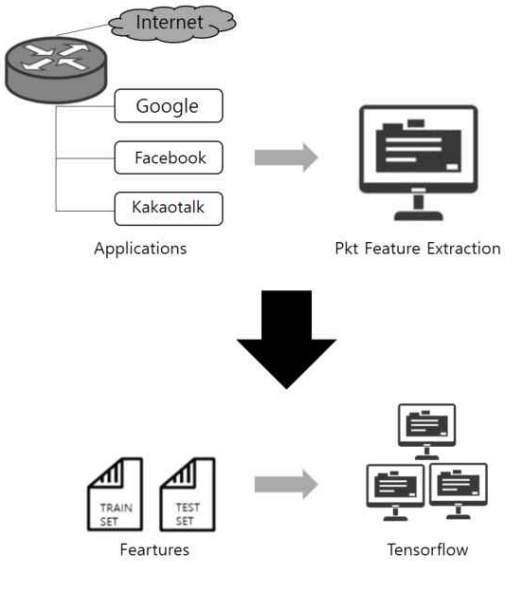

그림 1. Tensorflow를 활용한 어플리케이션 트래픽 분류과정

Ⅲ. 실험

Softmax Regression을 활용한 어플리케이션 트래픽 분류 시스템의 효율성을 검증하기 위해 실험을 진행한다. 실험을 진행하기 위해 학내 망에서 Microsoft NetWorkMonitor 프로그램을 통해 어플리케이션 트래픽을 수집하였다. 사용된 어플리케이션 트래픽은 3가지로 Google, Facebook, Kakaotalk 이다. 수집한 트래픽의 80%는 패턴 학습을 위한 Train Set, 20%는 검증을 위한 Test Set 으로 하였다. 어플리케이션 트래픽 feature 의 개수와 learning rate에 따른 변화량은 표1 과 같다. 사용된 feature 의 속성은 트래픽의 Flow 크기, 지속 시간, 프로토콜, 전체 패킷 수, Forward 패킷 수, Backward 패킷 수이다[2]. 6 가지 속성을 바탕으로 Softmax Regression을 적용, Cost function을 사용하기 위한 Gradient Descent 알고리즘을 적용했다. 학습 횟수는 각 실험 별 동일하게 100,000 번씩 수행했다. 실험 결과의 분석 척도는 Accuracy 가 있다. Accuracy 는 분석된 트래픽 중 올바르게 어플리케이션 트래픽 식별이 된 트래픽의 비율을 나타낸다.

표 1. 어플리케이션 트래픽 모델 별 분석 결과

	Features						
	F1	F2	F3	F4	F5	F6	AC
α = 0.001	○	○	○	○	○	○	58.1
		○	○	○	○	○	59.8
			○	○	○	○	59.6
α = 0.01	○	○	○	○	○	○	68.3
		○	○	○	○	○	61.2
			○	○	○	○	59.8
α = 0.1	○	○	○	○	○	○	64.2
		○	○	○	○	○	61.1
			○	○	○	○	59.3
α = 1.0	○	○	○	○	○	○	59.8
		○	○	○	○	○	66.7
			○	○	○	○	47.2

α = Learning rate, AC = Accuracy (%),
F1 = Flow Size, F2 = Duration,
F3 = Protocol, F4 = Total Packets,
F5 = Forward Packets, F6 = Backward Packets ,

Ⅳ. 결론 및 향후과제

본 논문은 기계 학습(Machine Learning) 을 이용한 효율적인 어플리케이션 트래픽 분류 방법을 제안하였다. 3가지 어플리케이션 트래픽 (Google, Facebook, Kakaotalk) 을 Tensorflow 의 Softmax Regression 을 활용하여 feature 학습에 의한 어플리케이션 트래픽 분류를 실험하였다. 모든 feature를 학습시킨 뒤 learning rate를 0.01로 설정했을 때, 최고의 분석률을 보였다. 위 2장에서 언급한 바와 같이, Learning rate 에 따른 학습 결과가 다르기 때문에 적절한 Learning rate 를 설정해야한다는 것을 검증하였다.

본 논문에서는 feature 추출 프로그램 (Pkt Feature Extraction) 에서 추출한 300개의 feature 중 7가지만 선정하였다. 향후 feature에 대한 추가적인 연구를 할 것이고, 조금 더 효율적인 기계학습 (Machine Learning) 알고리즘을 활용하여 보다 정교한 검증 실험을 진행할 계획이다.

참 고 문 헌

[1] Sung-Ho Yoon, Kyu-Seok Shim, Su-Kang Lee, and MyungSup 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

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

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