머신러닝과 딥러닝을 이용한 네트워크 트래픽 분류 연구 동향

이정민, 이연준
한양대학교 컴퓨터공학과 바이오인공지능융합전공
lsmp12@hanyang.ac.kr, yeonjoonlee@hanyang.ac.kr

Trend of Network Traffic Classification Using Machine Learning and Deep Learning

JungMin Lee, Yeonjoon Lee
Major in Bio Artificial Intelligence, Dept. of Computer Science & Engineering,
Hanyang University

요 약

네트워크 트래픽 연구는 오랜 기간 지속되어 왔으며, 구현이 비교적 간단하고 높은 정확도를 가지는 기존의 분류 방식들이 오랫동안 사용되어왔다. 그러나 네트워크 기술과 암호화 기술의 발달로 기존의 분류 방식들은 더 이상 분류 결과에 대한 신뢰성을 보장할 수 없으며, 이에 따라 새로운 분류 방식의 필요성이 대두되었다. 최근 머신러닝과 딥러닝을 네트워크 트래픽 분류에 적용하는 연구가 활발히 이루어지고 있으며 획기적인 모델들이 많이 제안되었고, 그 분류 성능 또한 입증되었다. 그러나 여전히 여러 가지 극복해야 할 문제점은 남아있으며 이러한 문제점을 해결하기 위한 연구가 앞으로도 계속 진행될 것으로 보인다. 본 논문은 머신러닝과 딥러닝을 이용한 네트워크 트래픽 분류 연구 동향에 대해 살펴보고 이러한 연구들이 가지는 문제점을 짚고 넘어가며 앞으로의 네트워크 트래픽 분류 연구의 방향성에 대해 이야기 하고자 한다.

1. 서론

네트워크 트래픽 분류에 대한 연구는 오랜 기간 연구되어 온 주제이다. 네트워크 관리와 보안적인 측면에서 네트워크 트래픽 분류는 중요한 작업으로 여겨지며, 그에 따라 포트 기반과 페이로드 기반의 다양한 분류 방식이 등장했다. 최근, 네트워크 기술의 발달로 생성되는 트래픽의 종류와 그 양은 나날이 증가하고 있으며, 암호화 기술의 발달로 기존의 트래픽 분류 방식은 큰 효과를 기대하기 어려워졌다. 현 상황에 맞는 정확하고 효율적인 트래픽 분류 방식에 대한 연구의 필요성이 점차 커지고 있으며, 이에 따라 최근 다양한 분야에서 활발히 이용되는 머신러닝과 딥러닝을 네트워크 트래픽 분류에 적용하는 연구들이 진행되고 있다. 본 논문에서는 이러한 연구들을 통해 네트워크 트래픽 분류 연구의 동향에 대해 알아보고, 이러한 연구들이 가지는 문제점을 짚고 넘어가며 앞으로의 네트워크 트래픽 분류 연구의 방향성에 대해 이야기하고자 한다.

2. 기존의 네트워크 트래픽 분류 방식

기존의 네트워크 트래픽 분류 방식은 크게 두 가지로 나뉜다. 포트 기반 방식과 페이로드 기반 방식이다. 최근 연구 동향을 알아보기 전에 이러한 전통적인 네트워크 트래픽 분류 방식에 대해 언급하고자 한다.

포트 기반 방식은 어플리케이션들이 사용하는 정해진 포트를 기반으로 트래픽을 분류하는 방식이다. 이러한 서비스들은 서로 다른 정해진 포트 번호를 가지기 때문에 이를 기반으로 정확한 분류가 가능했다. 그러나 최근 둘 이상의 포트를 사용하거나 정해진 포트를 사용하지 않고 동적으로 포트를 할당하는 서비스들이 증가하면서, 포트 기반 방식에 의존하기는 어려워졌다.

페이로드 기반 방식은 패킷 내부의 페이로드에 담겨있는 특정 문자열이나 패턴으로 트래픽을 분류하는 방식이다. 어플리케이션마다 서로 다르게 가지는 패킷 페이로드 내용을 기반으로 분류하는 방식으로, 이는 최근 암호화 기술의 발달로 그 효능을 잃게 되었다. 기존에 평문으로 전송되던 페이로드가

암호화되면서 특정 문자열과 패턴을 찾기가 어려워졌기 때문이다. 이에 따라, 딥러닝을 이용하여 페이로드의 내용이 아닌 패킷 바이트 그 자체에서 특정 패턴을 찾고자 하는 연구도 시도되었다[1].

이러한 전통적인 분류 방식은 비교적 구현이 간단하고 정확도가 높은 방식이지만, 네트워크 기술과 암호화 기술이 발달한 최근의 네트워크 상황에 적용했을 때의 그 신뢰성을 보장할 수 없다.

3. 머신러닝과 딥러닝을 이용한 연구 동향

머신러닝과 딥러닝은 최근 많은 분야에서 활발하게 사용되고 있으며, 네트워크 트래픽 분류 연구에게도 새로운 길을 열어주었다. 네트워크 트래픽으로부터 다양한 특징을 추출하고 이를 통해 모델을 학습시켜 분류에 적용하는 방식으로 이용되고 있으며, 그 정확성과 속도도 많이 검증되었다.

특히, 패킷 길이와 지연 시간의 최댓값, 최솟값, 분산, 평균 등과 같은 트래픽의 다양한 통계적 특징을 추출하고, 이를 결정 트리, 랜덤 포레스트 등의 머신러닝 알고리즘과 결합하는 연구가 많이 진행되었다[2][3][4][5]. 다양한 통계적 특징을 추출함으로써 특정 어플리케이션의 트래픽이 가지는 전체적인 특성을 알 수 있고 이를 통해 명확한 분류가 가능하다. 그러나 어떤 통계적 특징을 추출할 것인가는 사람에 의해 직접 결정되기 때문에 특징을 선택하는 사람의 지식과 능력에 크게 의존하며, 전체적인 특징을 계산하기 때문에 데이터의 세부적인 특징은 고려되지 않는다는 단점이 존재한다.

그에 따라 자동으로 의미 있는 특징을 추출하는 딥러닝을 이용하는 연구가 새롭게 진행되고 있다. 딥러닝 알고리즘은 훈련 데이터로부터 자동으로 특징을 추출하기 때문에 인력 소모가 적다는 장점이 있으며, 추출된 특징들이 데이터들을 잘 표현한다는 것 또한 증명되고 있다. [6]은 딥러닝 알고리즘을 이용한 엔드 투 엔드 모델을 제안했다. 엔드 투 엔드 모델은 특징 추출 단계와 모델 훈련 단계가 나뉘지 않고 통합된 모델로, 어떠한 가공도 거치지 않은 네트워크 플로우 시퀀스로부터 자동으로 특징을 추출하고 이를 이용하여 바로 모델을 훈련 시킨다. 제안된 모델은 오토인코더를 통해 특징을 학습하며 재구성 메커니즘을 이용하여 특징 학습을 강화했다. [7]은 멀티모달 딥러닝 프레임워크를 제안했다. [6]과 동일하게 엔드 투 엔드 모델이지만 두 가지 측면에서 특징을 학습한다는 것이 차이점이다. 제안된 모델은 바이트와 패킷 길이 시퀀스를 입력 데이터로 이용하며 다양한 측면에서 특징을 추출함으로써 데이터들의 순차적 관계를 더 잘 학습했다. 또한 기존 엔드 투 엔드 모델에서 자주 사용되던 알고리즘인 합성곱 신경망에 의존하지 않고 Multi-head Self-attention Mechanism 기반의 새로운 방식을 도입했다는 것도 의의가 있다. [8]은 데이터셋에 대한 문제에 초점을 두었다. 데이터셋의 불균형에 대한 문제와 훈련 데이터의 크기에 크게 의존하는 기존 모델들의 한계점을 보완하기 위해 추가적인 훈련 데이터를 생성하는 모듈을 고안했다. 제안된 메타 학습 기반 모델은 불균형한 데이터셋과 소량의 라벨링 된 데이터셋에서 각각 좋은 성능을 내면서 이 모듈의 효과를 입증했다. [9]는 많은 양의 라벨링 되지 않은 데이터를 이용하여 사전 학습 후 여러 가지 작업에 맞는 소량의 라벨링 데이터를 이용하여 미세 조정하는 모델을 제안했다. 특히 데이터 간의 관계를 잘 나타내는 특징을 학습할 수 있도록 입력 데이터를 토큰화하는 새로운 방법을 제안했다. [10]은 통계적 특징으로 이루어진 시퀀스에서 더 의미 있는 특징을 뽑아내는 Mean Teacher-style 준지도 학습 프레임워크를 제안했다. 특히 하나의 플로우가 아닌 다수의 연속적인 플로우들을 대상으로 분석함으로써 플로우 간의 시공간적 관계를 이용하여 분류 성능을 높였다. 또한 컴퓨터 비전과 자연어 처리 분야에서 좋은 성능을 보이던 트랜스포머 모델을 도입했으며, 실생활에 도입될 상황을 가정하여 기존 트랜스포머 모델의 높은 복잡도를 낮추고자 처리 과정에서 데이터의 길이를 압축하는 알고리즘을 고안하여 좀 더 효율적인 분류기를 제안했다.

4. 결론

머신러닝과 딥러닝은 최근 달라진 네트워크 환경으로부터 생성되는 다양한 네트워크 트래픽을 분류하는 연구에 효과적으로 적용되고 있다. 높은 정확도로 빠른 분류가 가능하며 기존 트래픽 분류 방식의 한계점을 해소할 수 있는 획기적인 방식으로 여겨지고 있다. 그러나 여전히 해결해야 할 문제점은 존재한다. 대부분의 문제점은 머신러닝과 딥러닝의 고질적인 문제점으로부터 초래된다.

첫째는, 데이터셋에 대한 문제점이다. 머신러닝과 딥러닝 모델은 훈련 과정에서 많은 데이터를 필요로 하는 경우가 많다. 특히 라벨링 데이터셋은 사람에 의해 수동으로 라벨링 되기 때문에 많은 인력과 시

간이 필요하다. 또한 데이터셋의 데이터 분포가 균일하지 않으면 성능에 큰 영향을 미칠 수 있다. 따라서 어떻게 효율적인 데이터셋을 구성할 수 있는지에 대한 연구와 더불어 불균일하고 적은 수의 데이터를 가지는 데이터셋에서도 좋은 성능을 낼 수 있는 모델 혹은 알고리즘을 고안하는 연구가 필요할 것이다.

둘째는, 일반화성이다. 네트워크 트래픽은 날이 갈수록 다양해지고 있으며 그 수도 증가하고 있다. 따라서 모델이 실제 네트워크 환경에 도입되었을 때, 훈련된 데이터셋에 존재하지 않는 데이터가 분류 모델의 입력으로 들어올 수 있으며, 이에 대한 분류 성능은 보장할 수 없다. 데이터셋에 존재하지 않던 데이터에 대한 분류 성능을 일반화성이라고 하며, 이러한 일반화성을 높일 수 있는 방법도 연구되어야 할 필요성이 있다.

셋째는, 다양한 프로토콜을 다룰 수 있는 모델의 필요성이다. 대부분의 연구는 특정 프로토콜을 대상으로 해당 프로토콜의 특징을 잘 추출할 수 있는 방법에 대해 연구한다. 다양한 프로토콜을 잘 나타낼 수 있고 범용적으로 사용될 수 있는 특징을 직접 고안하거나 혹은 그러한 특징을 추출할 수 있는 알고리즘에 대한 연구가 필요할 것이다. 최근 이러한 문제점과 두 번째 문제점에 대한 해결책으로 사전 학습, 전이 학습 등을 사용하는 연구들이 진행되고 있으며 좋은 성능을 보이고 있다.

넷째는, 더 효율적인 알고리즘 고안의 필요성이다. 이러한 머신러닝과 딥러닝 모델들은 많은 데이터를 다루고 복잡한 알고리즘을 가지고 있는 경우가 많기 때문에 시간과 물리적 자원 측면에서 비용이 크다. 적은 비용으로도 좋은 성능을 내는 효율적인 알고리즘이 고안될 필요가 있다.

이처럼 네트워크 트래픽 분류 연구는 머신러닝과 딥러닝의 적용으로 이 기술들이 가지고 있는 문제점을 그대로 떠안게 되었다. 이러한 문제를 해결하기 위해 좀 더 일반적이고 효율적인 모델 생성에 대한 연구는 계속해서 진행될 것이며, 기존의 네트워크 트래픽 연구에 사용되지 않던 새로운 모델을 도입하려는 시도도 계속될 것으로 보인다.

참고문헌

[1] Lin, Peng, et al. "A Novel Multimodal Deep Learning Framework for Encrypted Traffic Classification." IEEE/ACM Transactions on Networking (2022).

[2] Taylor, Vincent F., et al. "Robust smartphone app identification via encrypted network traffic analysis." IEEE Transactions on Information Forensics and Security 13.1 (2017): 63-78.

[3] ANDERSON, Blake; MCGREW, David. Machine learning for encrypted malware traffic classification: accounting for noisy labels and non-stationarity. In: Proceedings of the 23rd ACM SIGKDD International Conference on knowledge discovery and data mining. 2017. p. 1723-1732.

[4] LIU, Junming, et al. Effective and real-time in-app activity analysis in encrypted internet traffic streams. In: Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining. 2017. p. 335-344.

[5] PISKOZUB, Michal; SPOLAOR, Riccardo; MARTINOVIC, Ivan. Malalert: Detecting malware in large-scale network traffic using statistical features. ACM SIGMETRICS Performance Evaluation Review, 2019, 46.3: 151-154.

[6] LIU, Chang, et al. Fs-net: A flow sequence network for encrypted traffic classification. In: IEEE INFOCOM 2019-IEEE Conference On Computer Communications. IEEE, 2019. p. 1171-1179.

[7] LIN, Peng, et al. A Novel Multimodal Deep Learning Framework for Encrypted Traffic Classification. IEEE/ACM Transactions on Networking, 2022.

[8] ZHENG, Wenbo, et al. Learning to classify: A flow-based relation network for encrypted traffic classification. In: Proceedings of The Web Conference 2020. 2020. p. 13-22.

[9] LIN, Xinjie, et al. Et-bert: A contextualized datagram representation with pre-training transformers for encrypted traffic classification. In: Proceedings of the ACM Web Conference 2022. 2022. p. 633-642.

[10] ZHAO, Ruijie, et al. MT-FlowFormer: A Semi-Supervised Flow Transformer for Encrypted Traffic Classification. In: Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. 2022. p. 2576-2584.