네트워크 이상치 탐지를 위한 정상 데이터만을 활용한 메모리 기반 정상성 학습

*이건수 이호창 심재훈 구형일 조남익
서울대학교 전기정보공학부 뉴미디어통신연구소
*drt4917@ispl.snu.ac.kr hochang@ispl.snu.ac.kr joinjhs@ispl.snu.ac.kr hikoo@ajou.ac.kr
nicho@snu.ac.kr

Learning Memory-Guided Normality with Only Normal Training Data for Novelty Detection in Network Data

*Lee, Geonsu Lee, Hochang Sim, Jaehoon Koo, Hyung Il Cho, Nam Ik
Seoul National University Institute of New Media and Communications

요약

본 논문에서는 네트워크 이상치 탐지를 위하여 정상 데이터만을 활용한 메모리 기반 정상성 학습 모델을 제안한다. 오토인코더를 기반으로 정상 데이터의 특징을 표현하는 프로토타입을 생성할 수 있도록 신경망을 구성하고, 네트워크 데이터의 특성을 반영하여 쿼리의 수를 한 개로 고정하며, 사용되는 프로토타입의 수를 지정한 값으로 고정하여 모든 프로토타입에 정상 데이터의 특징을 반영할 수 있는 학습 방법을 제안한다. 해당 모델을 네트워크 이상치 탐지 데이터 세트인 Kyoto Honeypot, UNSW-NB15, CICIDS-2018에 적용하여 본 결과 Kyoto Honeypot에서는 0.821, UNSW-NB15에서는 0.854, CICIDS-2018에서는 0.981의 AUROC를 달성했다.

1. 서론

최근 통신 네트워크는 컴퓨터 간의 통신을 넘어 스마트 장치와 스마트 가전 등이 연결된 IoT 등에도 적용되면서 네트워크의 범위, 크기, 전송량이 크게 확장되었다. 통신 네트워크가 확장되면서 네트워크 공격으로 인한 개인 정보 유출이나 서버를 대상으로 한 해킹 등 네트워크 보안 문제도 증가하였다. 따라서 비정상적인 네트워크 접근을 차단하기 위한 네트워크 이상치 탐지의 중요성이 더욱 높아진 상황이다. 그러나 네트워크 공격은 종류가 다양할 뿐만 아니라 시간이 지남에 따라 새로운 공격 유형이 추가되기도 하며, 정제된 데이터 세트를 얻는 것이 어려우므로 지도 학습을 통한 탐지는 한계가 존재한다.

본 논문에서는 이러한 한계를 극복하기 위하여 상대적으로 취득이 쉬운 정상 데이터만을 활용한 네트워크 이상치 탐지 모델을 제안한다. 특히, 비디오 이상치 탐지 분야에서 제안되었던 오토인코더(Auto encoder)를 기반으로 프로토타입 개념을 더한 선행 모델인 “이상치 탐지를 위한 메모리 기반 정상성 학습”[1]을 기반으로 하여, 네트워크 이상치 탐지 분야에 적합한 방법을 제안한다. 또한, 이러한 방법에서 소수의 프로토타입만을 생성하여 학습하게 되는 문제를 해결하기 위해 사용되는 프로토타입의 수를 고정할 수 있는 새로운 학습 방법을 적용한다. 마지막으로, 네트워크 이상치 탐지 데이터 세트인 Kyoto Honeypot[2][3], UNSW-NB15[4], CICIDS-2018[5][6]에 대하여 제안된 모델의 성능을 측정하고 이를 지도 학습을 통해 학습된 모델의 성능과 비교하여 제안한 방법의 가능성을 확인하였다.

2. 관련 연구

2.1. 이상치 탐지를 위한 메모리 기반 정상성 학습

[1]에서는 비디오 이상치 탐지 분야에서 일반적인 오토인코더에 프로토타입 개념을 더하여 네트워크가 정상 데이터의 특징점들을 프로토타입으로 추출할 수 있도록 하는 모델을 제안했다. 이 방법은 인코더, 디코더, 메모리 모듈로 구성되어 있으며, 인코더와 디코더는 일반적인 오토인코더의 구조와 같다. 메모리 모듈은 다수의 프로토타입을 저장하며, 인코더에서 추출된 특징들을 프로토타입들을 중심으로 뭉치도록 하여 정상 데이터들이 가진 일반적인 특징들을 추출할 수 있도록 하였다.

학습 과정에서는 인코더와 디코더뿐만 아니라 메모리에 저장된 프로토타입 또한 학습된다. 손실 함수는 인코더에 입력되는 값과 디코더에서 출력되는 값의 차이를 계산하는 Reconstruction Loss, 프로토타입을 중심으로 추출된 특징들이 뭉쳐진 정도를 계산하는 Feature Compactness Loss, 프로토타입 간의 거리를 일정한 임계값 이상으로 유지하기 위한 삼중항 손실인 Feature Separation Loss로 구성된다. 하지만 좋은 성능에도 불구하고 이 방법은 비디오 데이터에만 국한되어 있어 다른 분야에 쉽게 적용하기 어렵다.

2.2. 네트워크 이상치 탐지 데이터 세트

Kyoto Honeypot 데이터 세트[2][3]는 실제 네트워크 트래픽으로 구성된 데이터 세트며 구체적인 공격의 종류는 제공되지 않는다. 총 935만 개의 세션 데이터로 구성되어 있으며 이 중 46.49%가 공격 데이터이

다. 총 24개의 특징(Feature)으로 구성되어 있다.

UNSW-NB15 데이터 세트[4]는 Fuzzers, Analysis, Backdoors, DoS, Exploits, Generic, Reconnaissance, Shellcode, Worms 공격을 포함하는 합성 데이터 세트다. 총 254만 개의 세션 데이터로 구성되어 있으며 이 중 12.65%가 공격 데이터이다. 총 49개의 특징으로 구성되어 있다.

CICIDS-2018 데이터 세트[5][6]는 Bruteforce, Dos, Web, Infiltration, Botnet, DDoS, PortScan 공격을 포함하는 합성 데이터 세트다. 총 933만 개의 세션 데이터로 구성되어 있으며 이 중 23.17%가 공격 데이터이다. 총 80개의 특징으로 구성되어 있다.

3. 제안하는 방법

이 장에서는 [1]을 기반으로 네트워크 이상치 탐지를 위한 메모리 기반 정상성 학습 신경망을 제안한다.

3.1. 사용한 신경망 구조

본 연구에서는 이상치 탐지를 위한 메모리 기반 정상성 학습 신경망 구조를 실험에 사용한 네트워크 데이터 세트에 맞도록 변형한다. 구체적으로, 선행 모델의 경우 비디오 이상치 탐지를 위한 모델이므로 하나의 프레임에 대하여 인코더를 통하여 여러 개의 특징을 추출하여 이를 다수의 쿼리로 구성하였다. 이는 이미지의 경우 이미지 전체가 아닌 이미지의 특정 영역에 이상치가 존재하는 경우를 처리하기 위한 것으로써, 이미지의 공간적인 독립성에 기반한 구조이다. 그러나 네트워크 이상치 탐지의 경우 하나의 세션 정보 전체가 하나의 데이터를 구성하게 되므로, 하나의 입력당 하나의 쿼리만이 구성될 수 있다. 따라서 쿼리의 개수를 한 개로 고정한다.

또한, 프로토타입의 경우, 정상 데이터에 대하여 다수의 프로토타입이 존재할 수 있으므로 다수의 프로토타입 중 가장 쿼리와 유사한 프로토타입을 기반으로 디코딩을 수행할 수 있도록 본래의 구조를 유지한다. 다만, 선행 모델의 경우 학습 데이터 내부에 이상치에 해당하는 데이터가 포함되어 있으므로 적절한 프로토타입의 수를 신경망이 스스로 학습하도록 구성되어 있으나, 본 연구의 경우 학습 데이터가 정상 데이터만으로 구성되는바 최대한 많은 특징을 추출하기 위하여 지정한 수만큼의 프로토타입을 생성해 낼 수 있도록 별도의 학습 방법을 적용한다. (3.2절 참조)

제안한 신경망을 학습하기 위한 손실 함수 L은 다음과 같다.

$$L = L_{reconstruct} + \lambda_c L_{compact} + \lambda_s L_{separate} \quad (1)$$

$$L_{reconstruct} = \sum \| \hat{I} - I \|_2 \quad (2)$$

$$L_{compact} = \sum \| q - p_p \|_2 \quad (3)$$

$$L_{separate} = \sum [\| q - p_p \|_2 - \| q - p_n \|_2 + \alpha]_+ \quad (4)$$

여기서, $\hat{I}$는 디코더의 출력값, I는 인코더의 입력값, q는 쿼리, p_p는 쿼리와 가장 가까운 프로토타입, p_n은 쿼리와 두 번째로 가까운 프로토타입을 의미한다. λ_c, λ_s는 각 손실 함수 항에 대한 하이퍼 파라미터이며, α는 삼중항 손실에 있어 마진(margin)에 해당한다.

정상 데이터와 공격 데이터를 구분하는 기준은 이상치 점수 S이며 이는 다음과 같다.

$$S = \lambda(1 - g(P(\hat{I}, I))) + (1 - \lambda) g(D(q, p_p)) \quad (5)$$

$$D(q, p_p) = \| q - p_p \|_2 \quad (6)$$

여기서 $g(\cdot)$는 최대·최소 정규화 함수를 의미하며, $P(\hat{I}, I)$는 인코더 입력값과 디코더 출력값 간의 L2-norm을 의미한다. λ는 하이퍼 파라미터이다.

전체적인 신경망의 구조는 <그림 1>과 같다.

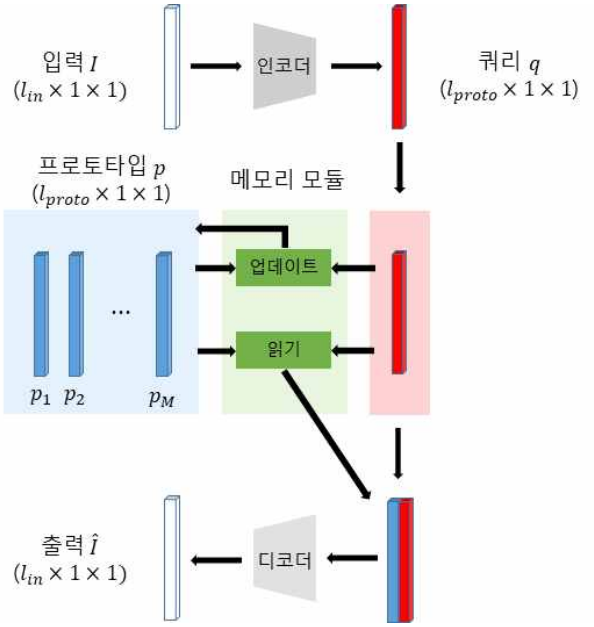

그림 1 전체적인 신경망 구조

3.2. 다수의 프로토타입을 생성하기 위한 학습 방법

신경망의 학습은 인코더와 디코더의 파라미터를 업데이트하는 단계와 메모리의 프로토타입을 업데이트하는 단계로 나누어진다. 이 중 프로토타입을 업데이트하는 단계에서 프로토타입의 사용 여부는 신경망이 학습되면서 함께 학습되는데, 이렇게 되면 M개의 프로토타입을 구성해 놓았더라도 이보다 적은 $N(1 \leq N \leq M)$개의 프로토타입이 사용되게 된다. 정상 데이터만을 학습 데이터로 사용하는 본 연구의 경우에는 프로토타입이 공격 데이터를 표현하는 경우가 발생하지 않으므로, 모든 프로토타입을 사용하여 정상 데이터를 표현하도록 강제하는 것이 필요하다.

구체적으로, 별도의 조치 없이 신경망을 학습시키는 경우에는 프로토타입의 무작위 초기값에 따라 지정한 프로토타입을 전부 사용하지 못하는 경우가 발생하는데, 학습이 계속 진행되더라도 사용하는 프로토타입의 수는 증가하지 않는다. 어떠한 쿼리도 해당하지 않는 이러한 사용하지 않는 프로토타입들은 정상 데이터의 분포를 표현하는 데 도움이 되지 않는다. 반면, 이를 방지하여 사용하는 프로토타입의 수를 일정한 개수까지 강제로 늘리면 신경망의 성능이 향상된다.

프로토타입의 무작위 초기값에 따라 간헐적으로 발생하는 일부 프로토타입의 미사용 문제를 완화하기 위하여 일정 시간 간격(t)으로 <그림 2>와 같이 사용하지 않는 프로토타입을 가장 많은 수의 쿼리를 표현하는 프로토타입 위치로 옮기는 방법을 제안한다. 같은 위치로 옮겨진 두 프로토타입은 손실 함수 중 $L_{separate}$항에 의하여 서로 멀어지게 되며, 하나로 뭉쳐있었던 프로토타입 근처의 쿼리들을 2개의 클러스터로 분리하게 되고, 결과적으로 프로토타입을 전부 정상 데이터의 특징 표현에 사용할 수 있도록 하여 프로토타입의 무작위 초기값에 영향을 받지 않고 안정적인 학습이 가능하다.

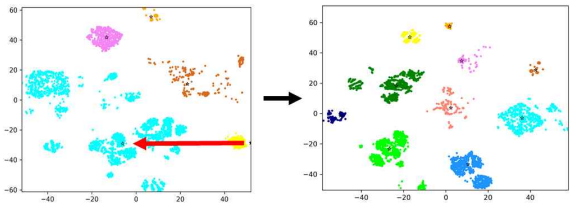

그림 2 사용되지 않는 프로토타입을 빨간 화살표 방향으로 이동(좌)

3.3. 실험 세부 사항

쿼리의 개수, 프로토타입의 크기 l_{proto}는 각각 1, 10으로 고정하고, 프로토타입의 개수 M은 다양하게 적용하였다. 인코더와 디코더는 (128, 64, 32, 16, l_{proto}) 구조를 갖는 5층의 다층 퍼셉트론(MLP)으로 구성하였으며 활성 함수는 Relu를 사용하였다. 하이퍼 파라미터 λ_c, λ_s, α, λ는 각각 0.01, 0.01, 1.0, 0.7로 설정하였다. 3.2. 절에서 제안한 학습 방법과 관련하여, 사용하지 않는 프로토타입의 위치를 옮기는 시간 간격 t는 4,000 epoch로 설정하였다. 학습률은 10^{-4}으로 설정하였고, Adam Optimizer를 사용하였다.

4. 실험 결과

이 장에서는 제안한 모델의 성능을 Kyoto Honeypot, UNSW-NB-15, CICIDS-2018 데이터 세트에 대하여 평가하고, 지도 학습된 모델의 성능과 비교한다.

4.1. 프로토타입의 수에 다른 메모리 기반 모델의 성능

3.2. 절의 학습 방법을 적용한 경우 지정한 수의 프로토타입을 모두 사용하여 안정적인 학습이 가능하다. <표 1>의 결과에서 볼 수 있듯이 합성 데이터인 UNSW-NB15, CICIDS-2018의 경우 공격 데이터가 인위적으로 생성되어 상대적으로 정형적이어서 프로토타입의 수가 5개일 때 가장 높은 성능을 보이지만, 실제 데이터인 Kyoto Honeypot의 경우 프로토타입의 수가 10~15개일 때 가장 높은 성능을 보인다.

표 1 메모리 기반 정상성 학습 모델의 실험 결과

데이터 세트	프로토타입의 수	AUROC
Kyoto Honeypot	5	0.813
	10	**0.821**
	15	**0.821**
UNSW-NB15	5	**0.854**
	10	0.821
	15	0.808
CICIDS-2018	5	**0.981**
	10	0.970
	15	0.964

4.2. 지도 학습 결과와의 비교

정제된 데이터를 활용한 지도 학습에는 Wide and Deep 모델[8]을 사용하였으며, 정상 데이터만을 활용한 본 모델의 성능이 몇 개의 정제된 데이터를 활용한 지도 학습의 성능과 동등한지 비교하였다.

표 2 Wide and Deep 모델의 실험 결과
및 메모리 기반 정상성 학습 모델의 실험 결과와의 비교

데이터 세트	사용된 데이터의 수	Wide and Deep 모델의 AUROC	메모리 기반 모델의 성능
Kyoto Honeypot	전부	0.911	100 ~ 500 사이의 성능
	2000	0.862	
	500	0.839	
	100	0.801	
	50	0.749	
UNSW-NB15	전부	0.979	50 ~ 100 사이의 성능
	2000	0.944	
	500	0.934	
	100	0.917	
	50	0.767	
CICIDS-2018	전부	0.999	100 ~ 500 사이의 성능
	2000	0.995	
	500	0.987	
	100	0.972	
	50	0.961	

메모리 기반 정상성 학습 모델의 경우 4.1. 절의 결과 중 가장 높은 AUROC 값을 갖는 프로토타입의 수를 기준으로 <표 2>에 비교하였다. 정상 데이터만을 사용하여 학습한 메모리 기반 정상성 학습 모델의 성능은 Kyoto Honeypot에서는 0.821, UNSW-NB15에서는 0.854, CICIDS-2018에서는 0.981의 AUROC를 달성했다. 이는 Kyoto Honeypot의 경우 100~500개의 정제된 데이터를 사용하여 지도 학습된 모델의 성능과 동등하며, UNSW-NB15의 경우 50~100개, CICIDS-2018의 경우 100~500개의 성능과 동등하다.

5. 결론

본 논문에서는 네트워크 이상치 탐지를 위하여 정상 데이터만을 활용한 메모리 기반 정상성 학습 모델을 제안하였다. 특히, 네트워크 데이터의 특성을 반영하여 모델의 구성 중 쿼리의 수를 한 개로 고정하였고, 사용되는 프로토타입의 수를 지정한 프로토타입 수로 고정하여 모든 프로토타입에 정상 데이터의 특징을 반영할 수 있는 학습 방법을 적용하여 학습 과정에서 소수의 프로토타입만을 사용하게 되는 문제를 해결하였다. 제안한 방법의 성능을 평가하기 위하여 정제된 데이터를 활용한 지도 학습 모델과의 성능 비교를 수행하였다. 본 모델은 정제된 데이터 없이 상대적으로 취득이 쉬운 정상 데이터만을 활용하여 학습할 수 있으며, Kyoto Honeypot에서는 0.821, UNSW-NB15에서는 0.854, CICIDS-2018에서는 0.981의 AUROC를 달성했고 이는 정제된 데이터를 사용한 지도 학습 모델과 비교해 Kyoto Honeypot에서는 100~500개, UNSW-NB15에서는 50~100개, CICIDS-2018에서는 100~500개의 성능과 동등함을 확인했다.

감사의 글

이 논문은 삼성전자 및 2020년도 BK21 플러스 창의정보기술 인재양성사업단에 의하여 지원되었음.

참고문헌

[1] H. Park, J. Noh, and B. Ham. “Learning memory-guided normality for anomaly detection,” CVPR. 2020.

[2] http://www.takakura.com/Kyoto_data/

[3] Song, J. Takakura, H. Okabe, Y, Eto, M, Inoue, D, and Nakao, K. “Statistical analysis of honeypot data and building of kyoto 2006+ dataset for nids evaluation,” Proceedings of the First Workshop on Building Analysis Datasets and Gathering Experience Returns for Security, pages 29-36. ACM. 2011.

[4] Moustafa, Nour, and Jill Slay. "UNSW-NB15: a comprehensive data set for network intrusion detection systems (UNSW-NB15 network data set)," Military Communications and Information Systems Conference (MilCIS), 2015. IEEE, 2015.

[5] https://registry.opendata.aws/cse-cic-ids2018/

[6] Ankit Thaklar, Ritika Lohiya. “A Review of the Advancement in Intrusion Detection Datasets,” International Conference on Computational Intelligence and Data Science (ICCIDS), 2019.

[7] https://openargus.org/

[8] H.-T. Cheng, L. Koc, J. Harmsen, T. Shaked, T. Chandra, H. Aradhye, G. Anderson, G. Corrado, W. Chai, M. Ispir, et al. “Wide & deep learning for recommender systems,” Proc. Workshop Deep Learn. Recommender Syst. 2016.