사이버 공격 행위 예측을 위한 딥러닝 학습 방법
Deep learning for estimating next action of cyber attack

최창희 · 신찬호 · 신성욱 · 서성연 · 이인섭
Changhee Choi · Chan Ho Shin · Sunguk Shin · Seongyun Seo · Inseop Lee

국방과학연구소
(changhee84@add.re.kr)

ABSTRACT

As cyber attacks become increasingly sophisticated and persistent, it is emerging as important issue for estimating the next action of cyber attack. In this paper, we proposed the novel method to predict next action of cyber attack based on MITRE ATT&CK model. We rearrange the technique labels from the cyber attack along the tactic order and design the deep learning networks based on LSTM. Experiment results show that the proposed method achieves reasonable exact match accuracy despite the large number of labels.

Key Words : MITRE ATT&CK, TTP, Cyber threat, APT, next action

1. 서론

사이버 공격의 주체가 점점 고도화되어, 개인에서 단체로 커짐에 따라 공격의 구성 또한 체계적이고 지속적으로 발전해 가고 있다. 과거 개인 수준의 공격을 방어하던 시그니처 기반의 방어 기술은 인적 자원과 자본이 충분한 단체에서는 회피가 용이하다. 이러한 공격에 대항하기 위하여, 여러 가지 연구가 진행되었는데, 그 중 하나는 사이버 공격을 TTP(Tactic, Techniques and Procedures)로 모델링하는 것이다. 이와 관련하여 가장 유명한 모델링 기술은 비영리 단체인 MITRE에서 만든 MITRE ATT&CK이다[1]. 거의 표준화되어가고 있는 이 모델은 사이버 공격을 Matrices, Tactics, Techniques, Mitigations, Groups, Software 항목으로 나타낼 수 있도록 모델링 하여, 현재 많은 사이버 보안 업체들이 이를 활용하여 기술을 개발하고 있다[2]. 본 논문에서는 이 모델을 기반으로 한 데이터셋을 기반으로 사이버 사용자의 다음 공격 행위를 예측 하는 기술을 제안한다.

2. 배경 지식

2.1 MITRE ATT&CK

MITRE ATT&CK은 사이버 공격을 matrix, tactic, technique 등과 같은 항목으로 모델링 하였다[1]. 본 논문에서는 일반적인 PC 및 서버를 대상으로 하고 있어서, ATT&CK Matrix for Enterprise를 사용한다. 이는 정찰, 권한 상승 등의 14개의 tactic과 active scanning, phishing과 같은 215개의 technique으로 사이버 공격을 모델링하였다.

2.2 rcATT dataset

사이버 공격의 전체적인 모습은 일반적으로 보안 업체에서 발행하는 보고서로 파악할 수 있다[3]. 이러한 보고서에는 해당 공격에서 사용한 다양한 기법이 서술되어 있는데, Valentine Legoy 연구진은 보고서에서 MITRE ATTACK technique을 추출해내는 연구를 진행하고, 실험을 위해서 데이터 셋을 제작하였으며, github에 공개하였다[4]. 각 사이버 공격에 해당하는 tactic과 technique를 one-hot encoding으로 표현한 데이터셋이며, 시간 순서가 존재하지 않는 것이 특징이다. 표 1은 rcATT 데이터셋의 일부를 나타낸 것이다. “Text”는 보고서 내용이며, 접두사가 “TA”인 것은 tactic, “T”인 것은 technique을 의미한다. 0은 해당사항이 없음을, 1은 해당한다는 의미이다. 데이터셋에는 연도가 표기되어 있지 않으나, 2019년도까지의 자료로 추정된다.

Table 1. rcATT 데이터셋 일부

Text	TA0003	TA0004	...	T1124	T1035	T1216
Talos Blog || Cisco ...	1	1	...	1	0	0
OilRig Actors Provide ...	0	0	...	0	0	0
Hogfish Redleaves ...	1	0	...	0	0	0

3. 제안하는 방법

3.1 도메인 변환

본 논문의 목표는 사이버 공격자의 다음 행위를 예측하여, 미리 방어를 할 수 있도록 도모하는 것이다. 이를 위해서 다음 technique을 예측할 수 있도록 표 1의 데이터를 변환 하였다. 우선 matrix을 기반으로 각 technique에 tactic을 key로 할당하였다. 이후 tactic을 기반으로 정렬한 후, technique의 전후 관계를 이용하여 sequence를 만들어 주었다. 이는 tactic의 순서가 논리적

인 인과관계를 바탕으로 이루어져 있고, 이는 자연스럽게 물리적인 공격 순서와도 일치할 확률이 높다는 것이다. 표 2는 본 변환 방법을 이용한 sequence의 일부를 나타낸 것이다. 많은 부분이 technique을 하나만 포함 경우가 있어 2개 이하의 technique이 라벨링된 사이버 공격은 제외하였다.

Table 2. 도메인 변환 결과(예시)

Cyber Attack	Sequence
45	T1193->T1105->T1113->T1125
51	T1168->T1116->T1113->T1123
86	T1033->T1082->T1016->T1048

3.2 학습 네트워크 모델링

Sequence를 학습 시키는 모델 중, 적은 데이터를 효율적으로 학습시킬 수 있는 LSTM을 이용하여 그림 와 같이 네트워크를 모델링하였다. 공격자의 행위 탐지 후의 바로 다음 행동을 예측하기 위하여, 바로 다음 행위를 예측할 수 있도록 설계하였으며, 전후 관계도 고려하기 위해 양방향으로 설계한 후, linear layer를 이용하여 절반으로 줄였다.

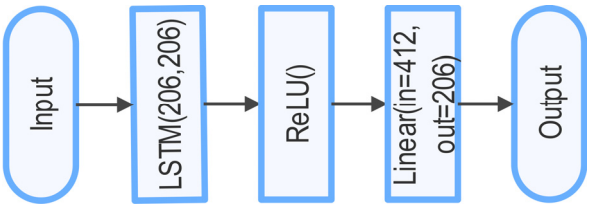

Fig. 1. 학습 네트워크 설계

4. 실험 결과

4.1 환경 세팅

실험을 위한 데이터셋은 실 운용을 고려하였을 때, 학습을 위한 데이터로는 rcATT 데이터셋의 과거 사이버 공격을, 테스트를 위한 데이터는 최근 사이버 공격을 사용하는 것이 합당하다. 다만, rcATT에는 연도가 기재되어 있지 않아, rcATT github에서 제공하는 데이터는 ground truth로 간주하여 전부 학습에 사용하고, 21년도에 발생한 사이버 공격에 대해서 rcATT의 알고리즘으로 새로 추출한 데이터를 테스트에 이용하였다. 사용된 데이터의 라벨링의 신뢰도는 다소 부정확할 수 있으나, 현 시점에 공개된 데이터 중에서는 최선이라고 생각한다.

LSTM 셀의 파라미터로 num_layers=2, bidirectional=True, dropout=0.2을 사용하였다. Optimizer는 Adam을 사용하고 learning rate는 0.01로 세팅하였다. loss 는 cross entropy를 사용하였다. 실험결과는 랜덤시드로 인해 실험시마다 약간의 차이가 존재한다.

4.2 예측 성능

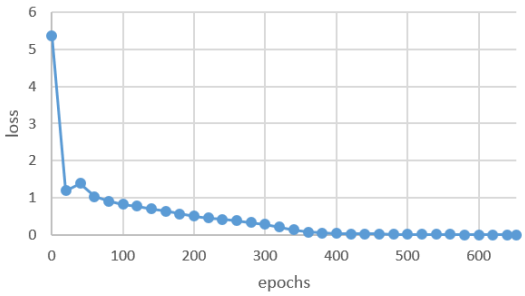

Fig. 2. Training loss

그림 2는 훈련 횟수(epoch)에 따른 loss 값의 변화를 도시한 것이다. loss 값이 0.003 미만이면 훈련을 멈추도록 하였으며, 653회 훈련이 진행되었을 때 loss 값 0.002965이며, 이때의 hamming loss는 0.13이었다. 훈련된 모델을 이용하여 21년도 사이버 공격 데이터를 대상으로 테스트를 진행하였을 때에는 87%의 정확도를 보였다. 다음에 나올 사이버 공격의 행위를 정확하게 예측하는 것이 가장 좋지만, 실 운용에서는 행위 후보를 예측하여 후보에 대해 모두 예방 혹은 대응하는 방법이 효율적일 수 있다. 이를 위해 그림 3과 같이 top-k의 정확도를 측정하였다. 일반적으로 통용되는 top-5 정확도는 90%를 기록하였다.

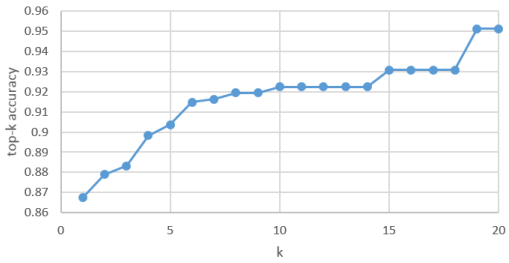

Fig. 3 top-k 정확도(k = 1~20)

5. 결론

본 논문에서는 사이버 공격의 다음 행위를 예측할 수 있도록 MITRE ATT&CK을 기반으로 공격 순서를 배치하는 도메인 변환 방법 및 LSTM 기반의 다음 행위를 예측하는 딥러닝 모델을 제시하였다. 실험 결과 약 87% 정확도로 다음 행위를 예측하는 것을 알 수 있었다. 데이터 라벨의 신뢰도가 중요한 만큼 추후 라벨의 신뢰도를 높이는 방법 및 적은 사이버 공격 행위 데이터로도 학습효과를 극대화 할 수 있는 방법에 대해서 연구할 계획이다.

References

[1] MITRE ATT&CK, https://attack.mitre.org
[2] ATTACKIQ, https://attackiq.com/mitre-attack/
[3] APT & Cybercriminals Campaign Collection, https://github.com/CyberMonitor/APT_CyberCriminal_Campaign_Collections
[4] rcATT, https://github.com/vlegoy/rcATT