사이버 공격 목표 예측을 위한 임베딩 및 RNN 학습 방법
Embedding and Training RNN for estimating the goal of cyber attack

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

국방과학연구소
(shinch2018@add.re.kr)

ABSTRACT

As a cyber security defender, one would like to predict and respond to a matter or situation, especially the goal of attacker. MITRE ATT&CK has become a useful knowledge base of adversary tactics and techniques in analyzing real-world observations. This paper proposes the embedding and training method of neural network model for estimating the goal of cyber attack based on TTP(Tactics, Techniques and Procedures) information of ATT&CK.

Key Words : TTP, Machine Learning, Embedding, cyber campaign, MITRE ATT&CK

1. 서론

개인과 단체를 넘어 국가 단위의 펀딩을 받는 것으로 의심되는 사이버 공격 캠페인이 등장하고 있다. 해당 공격 기법을 공유하고 대응책을 마련하기 위해 Trend Micro, Kaspersky 등과 같은 전문기관에서는 매 사건마다 위협 보고서를 발표하고 있다. 방어자 입장에서는 보고서를 통해 위협 행위자의 공격 수단과 이를 통해 이루고자 하는 궁극적인 목표를 파악하는 것이 가장 중요하다. 본 논문에서는 보고서 내의 MITRE ATT&CK의 TTP 정보[1]를 토대로 사이버 공격 행위의 목표를 예측하기 위한 임베딩 방법 및 RNN 학습방법을 제안한다.

2. 데이터 셋 및 모델 구성

2.1 데이터 셋

실험을 위한 데이터 셋은 rcATT[2]를 통해 마련하였다. rcATT 학습을 위해 사용한 보고서 및 공격기법 설명 1490개 데이터 중 MITRE ATT&CK[1] Impact가 포함되어 있는 41개의 데이터(이하 rcATT 데이터 셋)를 이용하였다. 추가로 Cyber Crime Collections(이하 CCC 데이터 셋)[3]에 포함되어있는 2008~2021년 보고서 1000여개에 대해 rcATT를 이용하여 TTP (Tactics, Techniques and Procedures)를 추출하였으며, 이 중 Impact가 포함된 28개의 데이터를 학습 및 테스트에 사용하였다.

Table 1. rcATT 데이터 셋 중 일부

Text	TA0006	TA0002	.....	T1077
Talos...	0	1		0
OilRig...	1	0		0

2.2 임베딩

TTP 정보를 이용한 임베딩은 다음 Fig. 1과 같이 이루어진다. 분석 대상 보고서가 특정 Technique을 포함하고 있으면 1, 아니면 0으로 체크하여 각 Tactic에 해당하는 Technique을 나열한다. 논문에서는 이렇게 만들어진 단위 문자열을 'TTP 단어'라 칭한다.

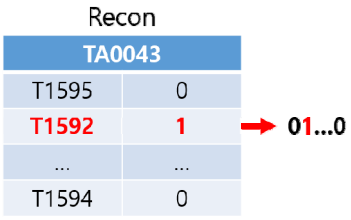

Fig. 1. 임베딩 방법

2.3 학습 모델

앞선 2.2절의 임베딩 방법을 통해 각 Tactic에 해당하는 TTP 단어를 생성한다. 이후 생성된 TTP 단어를 문장으로 인식하여 아래 Fig. 2와 Fig. 3과 같이 RNN 모델을 통해 학습시킨다. 모델은 학습 결과로 IMPACT에 해당하는 TTP 단어를 생성한다. IMPACT에 속하는 공격기법은 총 11개로, 이론적으로 등장할 수 있는 IMPACT TTP 단어는 2048(2^{11})개이다.

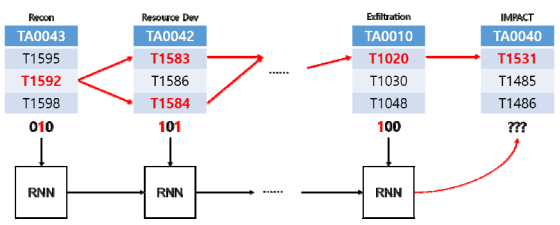

Fig. 2. TTP embedded RNN 모델 개념도

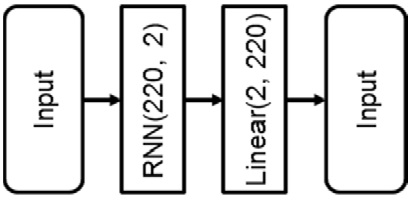

Fig. 3. TTP embedded RNN 모델 설계

3. 실험 결과

3.1 환경 세팅

학습 epoch은 20000, dropout은 0.2, learning rate는 0.01, hidden layer는 2로 세팅하였다. Loss function은 Cross Entropy Loss를 사용하였으며, Optimizer는 Adam을 사용하였다.

구현에 사용된 하드웨어의 상세한 스펙은 i9-7900X, 128GB, Geforce GTX1080Ti이다.

3.2 성능

총 69개의 데이터를 학습 및 테스트 데이터로 나누어 학습을 진행하였다. 다음 Table 1은 각 실험에 대한 학습 데이터와 테스트 데이터의 분포 및 설명이다. CCC 데이터 셋의 경우 학습 데이터와 테스트 데이터를 보고서 발행 연도를 기준으로 나누었다. rcATT 데이터 셋의 경우 보고서 발행 연도를 알 수 없어 모두 학습 데이터로만 사용하였다. 괄호 내의 숫자는 연도를 의미한다.

Table 2. 학습 및 테스트 데이터 수

No.	데이터 셋	학습 데이터 수	테스트 데이터 수
실험 1	rcATT	41	0
	CCC	24 (~2019)	4 (2020~2021)
실험 2	rcATT	41	0
	CCC	21 (~2018)	7 (2019~2021)
실험 3	rcATT	41	0
	CCC	13(~2017)	15 (2018~2021)
실험 4	rcATT	41	0
	CCC	0	4 (2020~2021)

아래 Fig. 4는 모델이 학습됨에 따라 떨어지는 loss를 보여준다. 그림은 실험 4에 대한 학습 진척도이며, 실험 1, 2, 3의 경우에도 비슷한 양상을 보였다.

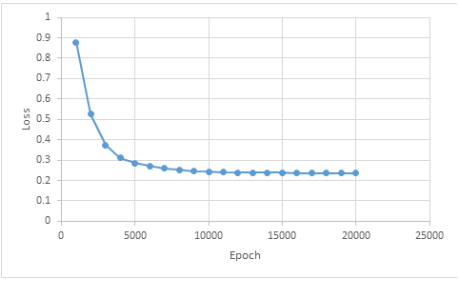

Fig. 4. RNN 모델 loss

다음 Table 2는 실험결과를 보여준다. 이론적으로 등장할 수 있는 라벨이 2048개의 문자열이기 때문에, 정확도(Accuracy)가 아닌 정답 TTP 단어와 생성된 TTP 단어 간의 hamming loss를 구하였다. Hamming loss는 두 문자열 사이의 다른 값을 측정하는 지표로, 값이 낮을수록 좋다. 각 실험에서 가장 좋은 값을 진하게 표시하였다.

Table 3은 2절의 임베딩 모델과 2.3절의 RNN모델에 대한 결과이다. Random은 무작위로 TTP 단어를 생성하였을 때의 결과이며, All 1은 TTP 단어를 “11111111111” 로 생성했을 때의 결과이다. Mean Test는 단어를 무작위로 생성하되, 테스트 데이터 셋의 TTP 단어에 등장하는 0과 1의 분포에 따라 단어를 생성했을 때의 결과이다. 실험 1과 실험 2의 경우, 테스트 데이터 셋이 충분치 않아(4건) 일부 성능이 떨어지는 결과를 보이기도 하였으나, 대부분의 실험에 대해서 실제 정답과 근사한 TTP 단어를 생성해냄을 볼 수 있다.

Table 3. 실험 결과

No.	Proposed	Random	All 1	Mean Test
실험 1	**0.1591**	0.4772	0.8863	**0.1591**
	0.0454	0.4545	0.8863	0.2045
	0.1136	0.5454	0.8863	0.2045
실험 2	**0.1429**	0.4936	0.8571	0.2278
	0.1299	0.6233	0.8571	0.2468
	0.0390	0.4545	0.8571	0.2468
실험 3	**0.1758**	0.4788	0.8545	0.2788
	0.1939	0.4970	0.8545	0.2303
	0.0848	0.5156	0.8545	0.3030
실험 4	**0.0681**	0.5000	0.8863	0.1591
	0.0455	0.6818	0.8863	0.2045
	0.1818	0.4772	0.8863	**0.1591**

4. 결론

본 논문에서는 등장한 사이버 공격 행위를 토대로 행위자의 궁극적인 공격 목표를 파악하기 위한 임베딩 기법 및 RNN 기반의 모델을 제시하였다. 테스트 데이터 셋의 분포에 따라 TTP 단어를 무작위로 생성하는 경우 Hamming loss가 0.2인 데에 반해 제안한 방법은 0.13의 Hamming loss를 보였다. 앞으로는 TTP의 의미상 거리를 추가하여 임베딩을 진행하는 등의 연구를 할 계획이다.

References

[1] MITRE ATT&CK, https://attack.mitre.org
[2] rcATT, https://github.com/vlegoy/rcATT
[3] APT & CyberCriminal Campaign Collections, https://github.com/CyberMonitor/APT_CyberCriminal_Campagin_Collections