멀티모달 융합 기반 실시간 악성 URL 방어 시스템

이준형, 권현*
육군사관학교 인공지능학과
junhyeong77428@gmai.com, *hkwon.cs@gmail.com

Multimodal Fusion Based Real-Time Malicious URL Defense System

Lee jun Hyeong, Kwon Hyun*
Department of Artificial Intelligence and Data Science, Korea Military Academy

요 약

현대 사이버 보안 환경에서는 정교한 난독화 기법으로 공격자가 악의적 URL을 은밀히 유포함에 따라 기존의 시그니처 기반 탐지 방식이 한계를 노출하고 있다. 본 연구에서는 URL을 시각(16×16 그레이스케일 영상), 구조(통계·문자 분포 등 25차원 특징 벡터), 의미(768차원 Transformer 임베딩) 세 가지 상이한 모달리티로 변환한 후, 각각 ResNet-18 CNN, XGBoost 및 CatBoost 트리 앙상블, DistilBERT를 활용하여 특성 확률 및 임베딩을 추출한다. 추출된 출력은 단일 771차원 벡터로 융합(concatenate)된 뒤 다층 퍼셉트론(DNN)을 통해 최종 이진 분류기로 학습된다. 약 700만 개의 URL 학습 데이터와 170만 개 평가 데이터를 이용한 실험에서 제안 프레임워크는 AUC 0.9805, F1-score 0.9189, Log Loss 0.1646을 기록하였으며, 평균 추론 지연시간은 0.42ms/URL로 실시간 필터링 요건을 충족함을 확인하였다.

Ⅰ. 서 론

사이버 공격자는 피싱, 악성코드 전파 등을 위해 URL 형태를 다양한 방식으로 변형·난독화하며, 기존 블랙리스트·시그니처 기반 탐지 기법은 알려지지 않은 변종 URL을 효과적으로 차단하지 못하고 있다. 최근 딥러닝 및 머신러닝 기법이 보안 분야에 적용되며 높은 성능을 발휘하고 있으나, 대부분은 텍스트나 구조적 특징, 시각적 표현 중 단일 모달리티에만 의존한다. 이러한 단일 접근법은 공격자가 사용하는 다차원적 특징을 포착하기에 불충분하며, 오탐 및 미탐 위험이 상존한다. 이에 본 연구는 URL 보안 탐지에 있어 상이한 정보를 상호보완적으로 융합함으로써 분류 성능과 실현성을 동시에 향상시키는 멀티모달 프레임워크를 제안한다.

Ⅱ. 관련 연구

기존 연구에서는 ASCII 코드를 기반으로 URL을 이미지로 변환한 뒤 CNN 계열 모델을 학습하여 시각 특징을 추출하는 방법[1], 텍스트 시퀀스를 Transformer로 인코딩하여 의미 정보를 파악하는 기법[2], 그리고 URL 구조적 속성을 XGBoost·CatBoost에 입력하는 통계 기반 분류 방법[3] 등이 연구되었다. 최근 멀티모달 융합 기법이 영상·음성·텍스트 분야에서 주목받고 있으나, URL 보안 분야에서는 주로 두 개 모달리티의 결합에 그치고 있으며, 세 가지 이상 모달리티를 통합한 사례는 보기 드물다.

Ⅲ. 제안 방법론

본 섹션에서는 데이터 전처리, 모달리티별 백본 모델, 그리고 최종 Fusion DNN 구조를 순차적으로 기술한다.

3.1 모델 아키텍처

제안 프레임워크는 세 가지 모달리티(문자열 시퀀스, 그레이스케일 영상, 구조적 특징)를 두 단계로 융합하여 악성 URL을 분류한다.

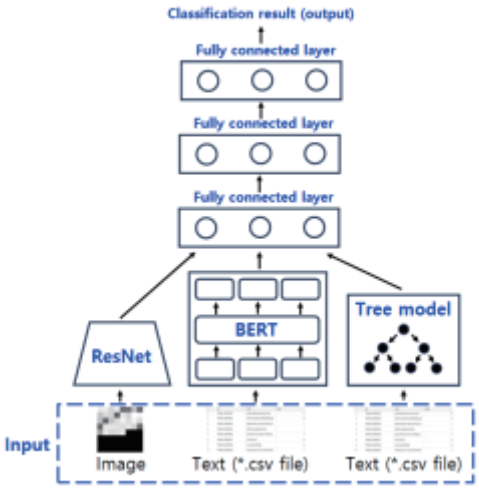

그림 1 멀티모달 아키텍처

3.1.1 ResNet 기반 시각 모달리티

URL 문자열을 ASCII 코드로 변환한 뒤 16×16 픽셀 그레이스케일 이미지로 재구성한다. 이후 사전 학습된 ResNet-18 컨볼루션 백본에 입력하여 특징을 추출하고, 최종 FC 레이어를 sigmoid 뉴런 하나로 대체하여 이미지 기반 악성 확률을 추출한다. $\sigma(\cdot)$는 sigmoid 함수이다.

$$p_{cnn} = \sigma(ResNet(I)) \in (0,1)$$

3.1.2 DistilBERT 기반 의미 모달리티

CSV 형식의 URL 문자열은 최대 길이 T의 토큰으로 패딩/트렁케이팅된 시퀀스로 토큰화된다. 이후 사전 학습된 DistilBERT 임베딩에 입력하여 [CLS] 토큰 임베딩을 추출한다.

$$f_{sem} = B(s) \in R^{768}$$

3.1.3 Tree 기반 구조 모달리티

URL에서 길이, 특수문자·숫자·대문자 개수, 서브도메인 수, 경로 깊이, 의심 키워드 유무 등 총 25개 차원 벡터를 추출하고 이 벡터를 Optuna로 탐색한 XGBoost 및 CatBoost 모델에 독립 투입하여 확률을 계산한다.

$$p_{xgb} = XGB(x), p_{cat} = CAT(x)$$

3.1.4 중간 융합층

세 모달리티 출력을 연결하여 771차원 융합 벡터를 구성한다.

$$f_{fuse} = [p_{cnn}, f_{sem}, p_{xgb}, p_{cat}]^T \in R^{771}$$

3.1.5 예측 헤드

융합 벡터를 두 개의 fully-connected 레이어로 처리하여 최종 확률를 계산한다. 여기서 ReLU(·)는 Rectified Linear Unit, $\sigma(\cdot)$는 sigmoid 함수이다.

$$h_1 = ReLU(W_1 f_{fuse} + b_{1)}, W_1 \in R^{512 \times 771}$$
$$h_2 = ReLU(W_2 h_1 + b_{2)}, W_2 \in R^{128 \times 512}$$
$$\hat{y} = \sigma(w_3^T h_2 + b_{3)}, w_3 \in R^{128}$$

3.1.6 임계값 결정 및 학습 목표

임계값은 최종 확률값이 0.5 이상일 경우 악성으로 판단하며, 모델은 이진 교차 엔트로피 손실을 최소화하도록 학습된다.

$$L = -y \log(\hat{y}) - (1-y) \log(1-\hat{y})$$

Ⅳ. 실험 과정 및 결과

4,1 데이터셋

학습 데이터로는 총 6,995,056개의 URL을 사용하였으며, 이 중 77.18%가 정상(0)으로, 22.82%가 악성(1)으로 라벨링되었다. 평가 데이터셋으로는 1,747,689개의 URL을 별도로 준비하여 모델 성능을 검증하였다. 모든 데이터는 Dacon.io에서 제공된 악성 URL 데이터셋에서 추출되었다.

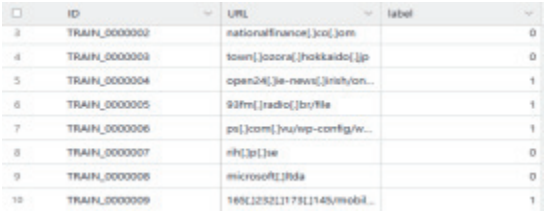

그림 2 데이터 구조 예시

4,2 실험 환경

XGBoost와 CatBoost는 Scikit-learn 및 CatBoost 라이브러리를, DistilBERT는 TensorFlow, ResNet-18 기반 CNN은 PyTorch를 통해 구현 및 미세조정하였으며, 모든 실험은 NVIDIA A100 GPU에서 수행되었다.

4,3 실험 결과 및 분석

Model	AUC	F1	Loss	Latency (ms/URL)
ResNet	0.923	0.799	0.255	0.359
DistilBERT	0.968	0.878	0.157	0.003
XGBoost	0.915	0.755	0.261	0.011
Multimodal	**0.981**	**0.919**	**0.164**	**0.420**

표 1 실험 결과 비교

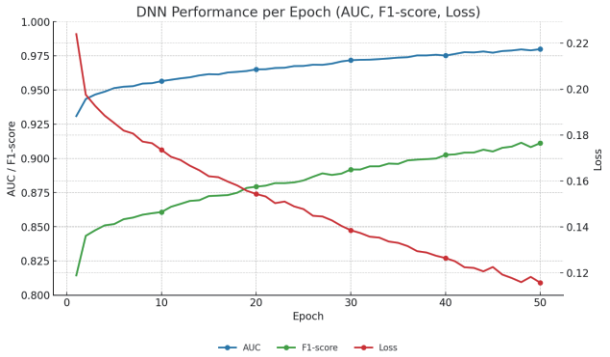

그림 3 DNN 에포크별 성능

표 1을 보면, 제안한 멀티모달은 단일 모달리티 대비 AUC 및 F1-score에서 현저한 성능 향상을 달성하였다. 한편, DistilBERT의 임베딩만을 사용함으로써 추론시간을 최소화 시켰으며 멀티모달은 평균 추론 지연시간 0.420ms/URL으로 실시간 필터링에 적합하다. 또한, 그림 3을 보면, DNN은 50 epoch 동안 AUC와 F1-score는 연속 상승했고 Loss는 안정적으로 감소했다.

Ⅴ. 결 론

본 연구는 시각·구조·의미 멀티모달 융합을 통해 악성 URL 탐지 성능을 크게 향상시키고, 평균 0.42ms/URL의 실시간 탐지 속도를 달성하였다. 따라서 제안 프레임워크는 정확도와 처리 속도 측면에서 모두 경쟁력 있음을 입증하였다. RoBERTa·DeBERTa와 같은 고도화된 Transformer 백본, 신경망 구조 탐색을 통한 자동 모달리티 가중치 조정 기법을 탐구하여, 본 멀티모달 융합 패러다임을 다른 사이버 위협 도메인으로 확장하는 것을 목표로 할 것이다.

참 고 문 헌

[1] K. He et al., “Deep residual learning for image recognition,” CVPR, 2016.

[2] J. Devlin et al., “BERT: Pre-training of deep bidirectional transformers,” NAACL, 2019.

[3] T. Chen and C. Guestrin, “XGBoost: A scalable tree boosting system,” KDD, 2016.

[4] A. Sanh et al., “DistilBERT, a distilled version of BERT,” EMNLP, 2019.