악성 URL 탐지 논문

분석 정리

2024-04-03

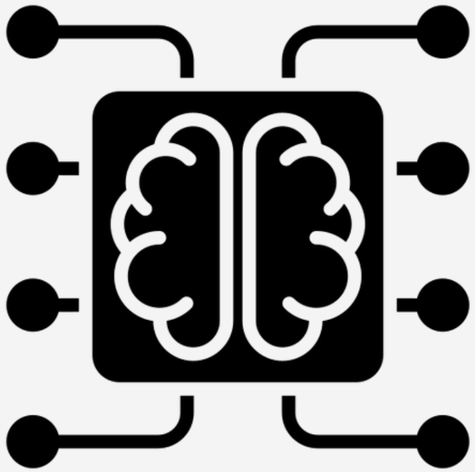
전승혁

URL Lexical Feature 기반의 DL-ML Fusion Hybrid

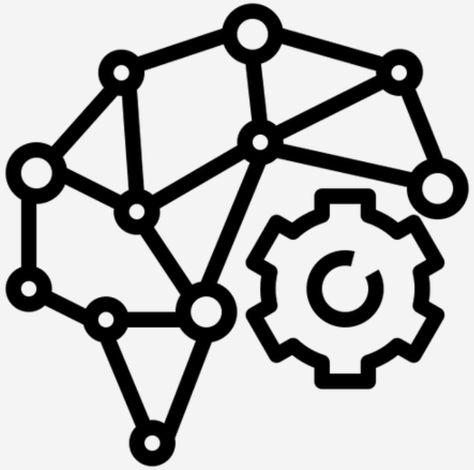
- 서론
- 관련 연구
- 배경 지식
- 제안 방법
- 실험 결과, 결론

URL Lexical Feature 기반의 DL-ML Fusion Hybrid

요약



머신 러닝



딥 러닝

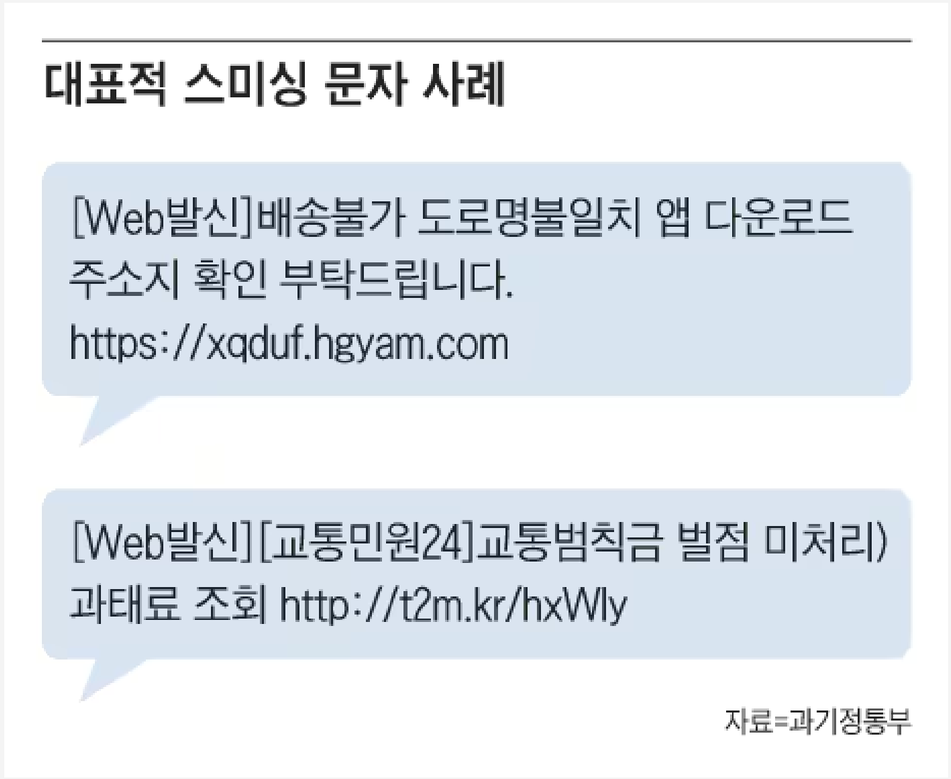
URL Lexical Feature 기반의 DL-ML Fusion Hybrid

서론

악성 웹사이트　　사회공학적 사이버 위협

URL Lexical Feature 기반의 DL-ML Fusion Hybrid

서론



URL Lexical Feature 기반의 DL-ML Fusion Hybrid

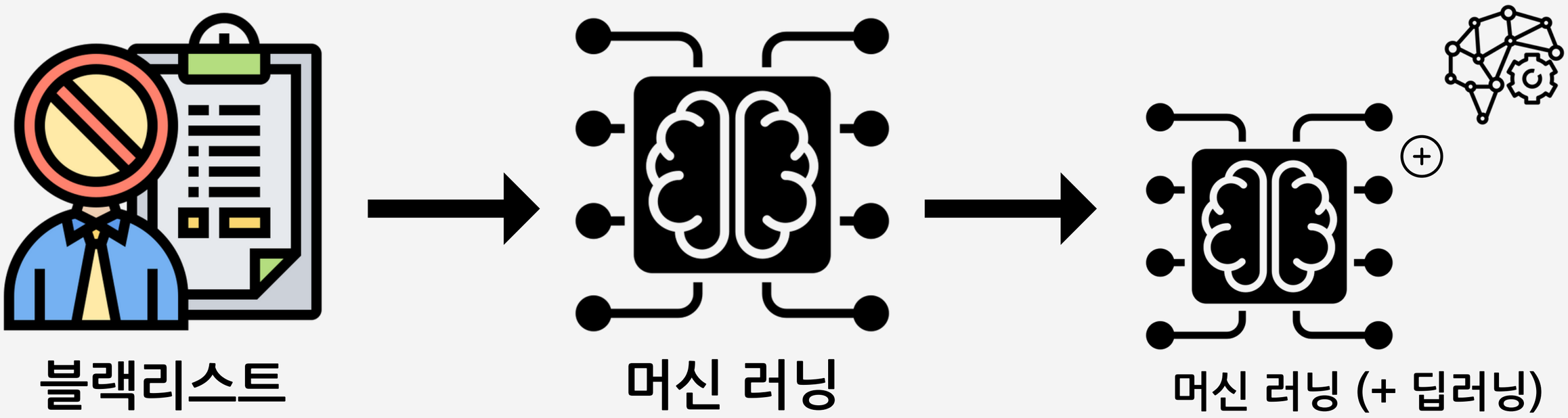
서론

단순

효과 : 개인 정보 & 계정 탈취
악성 코드 유포

URL Lexical Feature 기반의 DL-ML Fusion Hybrid

서론



성능 평가 메트릭

accuracy(정확도) : 예상과 결과가 같으면 높음

recall(재현율) : 정확도에서 TruePositives의 확률

precision(정밀도) : 예상과 결과가 True일 때, 예상이 Positive인 확률

f1 score : 재현율과 정밀도의 조화 평균

성능 평가 메트릭

accuracy(정확도) : 예상과 결과가 같으면 높음

$$\frac{TruePositives + TrueNegatives}{TruePositives + TrueNegatives + FalsePositives + FalseNegatives}$$

성능 평가 메트릭

recall(재현율) : 결과가 True에서 TruePositives의 확률

$$\frac{TruePositives}{TruePositives + FalseNegatives}$$

성능 평가 메트릭

precision(정밀도) : 예상이 Positives일 때, 예상과 결과가 true인 확률

$$\frac{TruePositives}{TruePositives + FalsePositives}$$

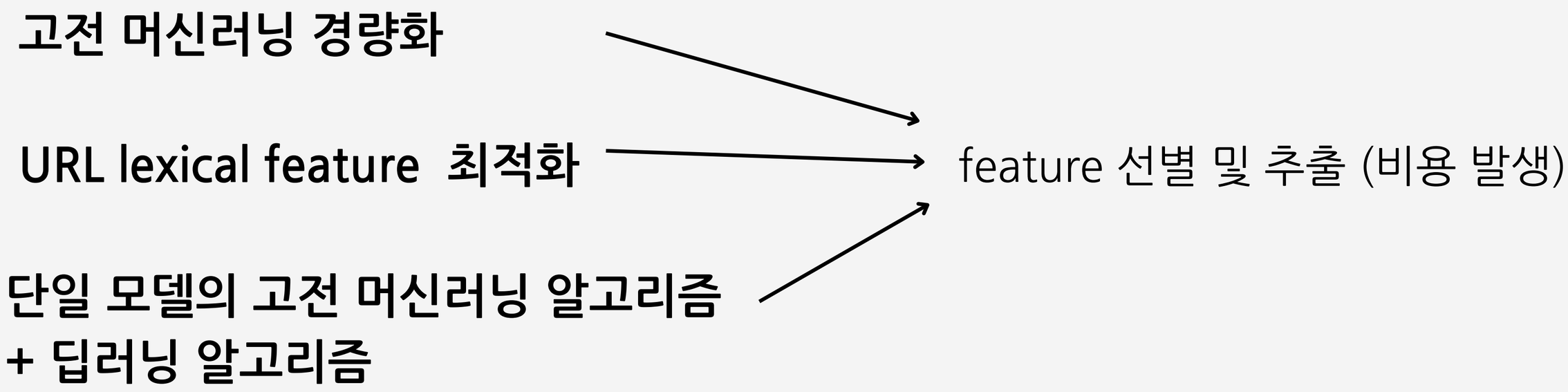
성능 평가 메트릭

f1 score : 재현율과 정밀도의 조화 평균

$$2 * \frac{Precision * Recall}{Precision + Recall}$$

URL Lexical Feature 기반의 DL-ML Fusion Hybrid

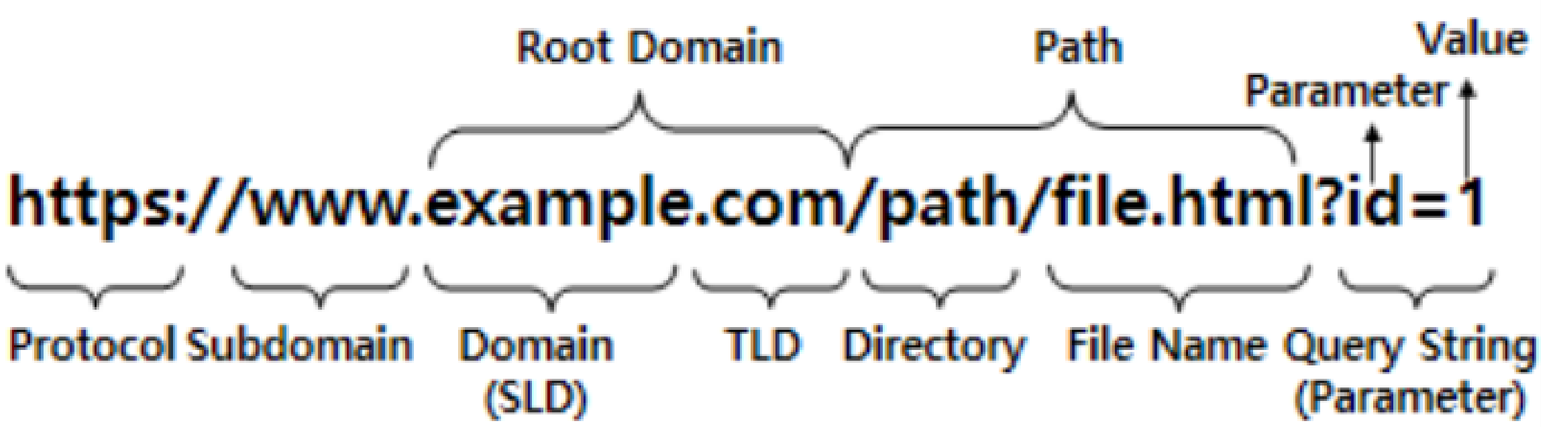
관련 연구



URL Lexical Feature 기반의 DL-ML Fusion Hybrid

기술적 배경지식

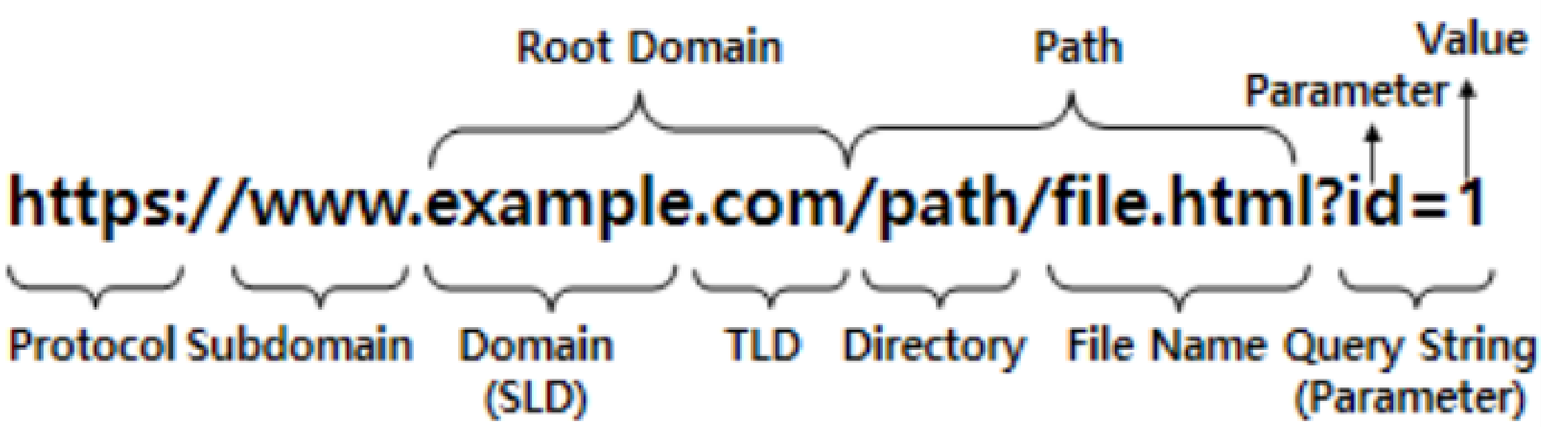
URL : 인터넷상에서 자원이 어디 있는지 알려주기 위한 규약 (RFC 1738)



URL Lexical Feature 기반의 DL-ML Fusion Hybrid

기술적 배경지식

URL : 인터넷상에서 자원이 어디 있는지 알려주기 위한 규약 (RFC 1738)

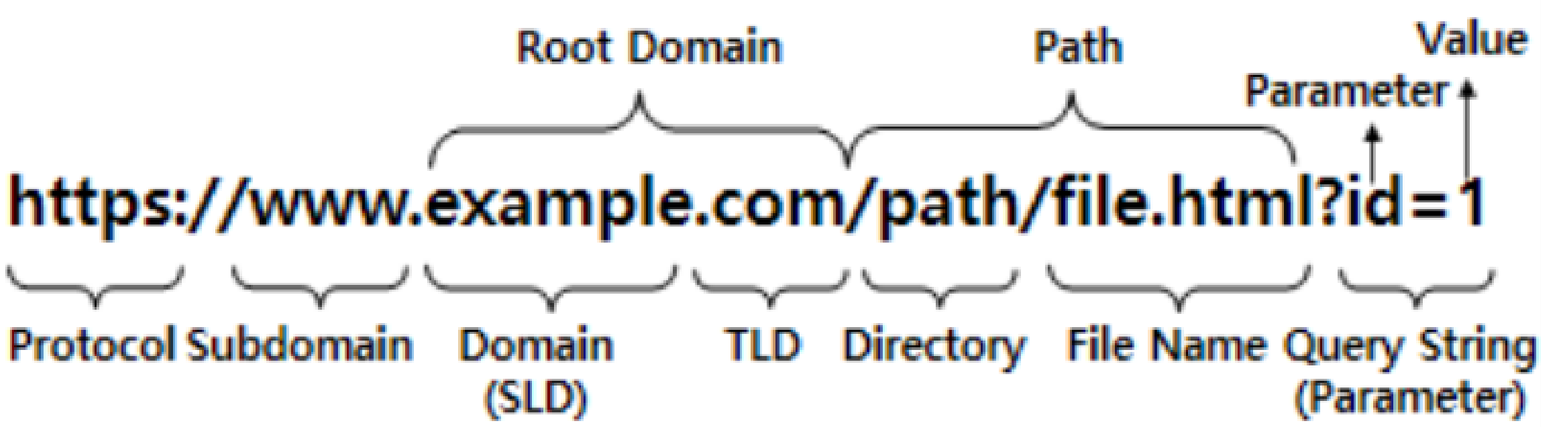


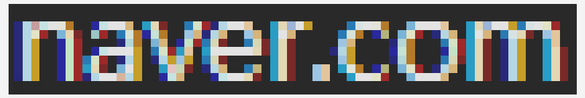
Protocol : 웹 브라우저가 웹 서버와 통신하는 방법 정의

URL Lexical Feature 기반의 DL-ML Fusion Hybrid

기술적 배경지식

URL : 인터넷상에서 자원이 어디 있는지 알려주기 위한 규약 (RFC 1738)



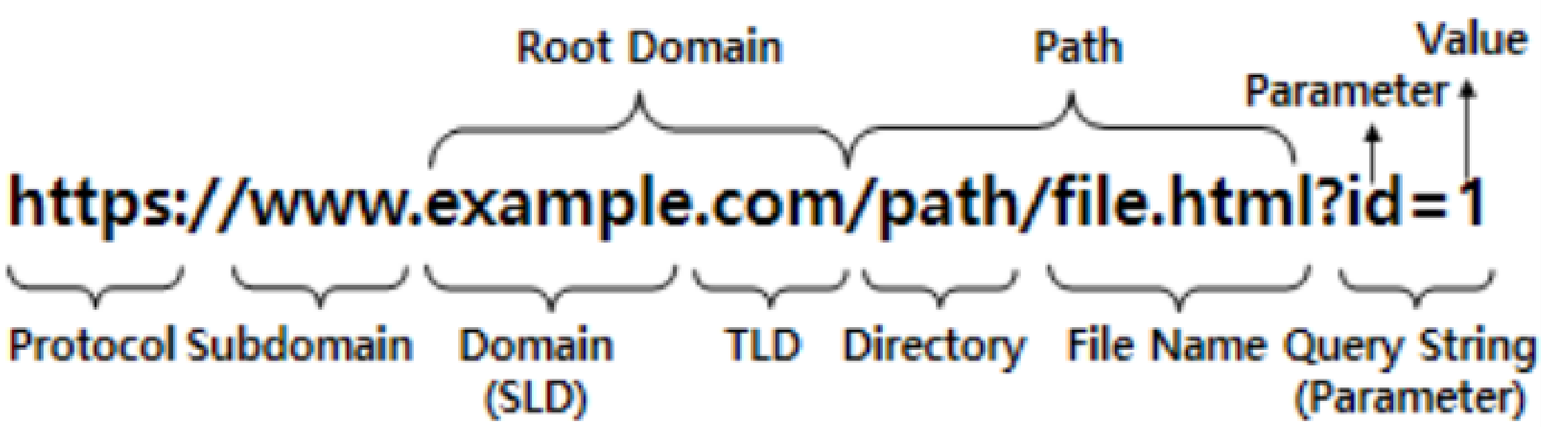


Subdomain : 루트 도메인의 보조 도메인
(별도의 웹 사이트 섹션을 만들어 운영할 때 독립된 도메인으로 활용)

URL Lexical Feature 기반의 DL-ML Fusion Hybrid

기술적 배경지식

URL : 인터넷상에서 자원이 어디 있는지 알려주기 위한 규약 (RFC 1738)



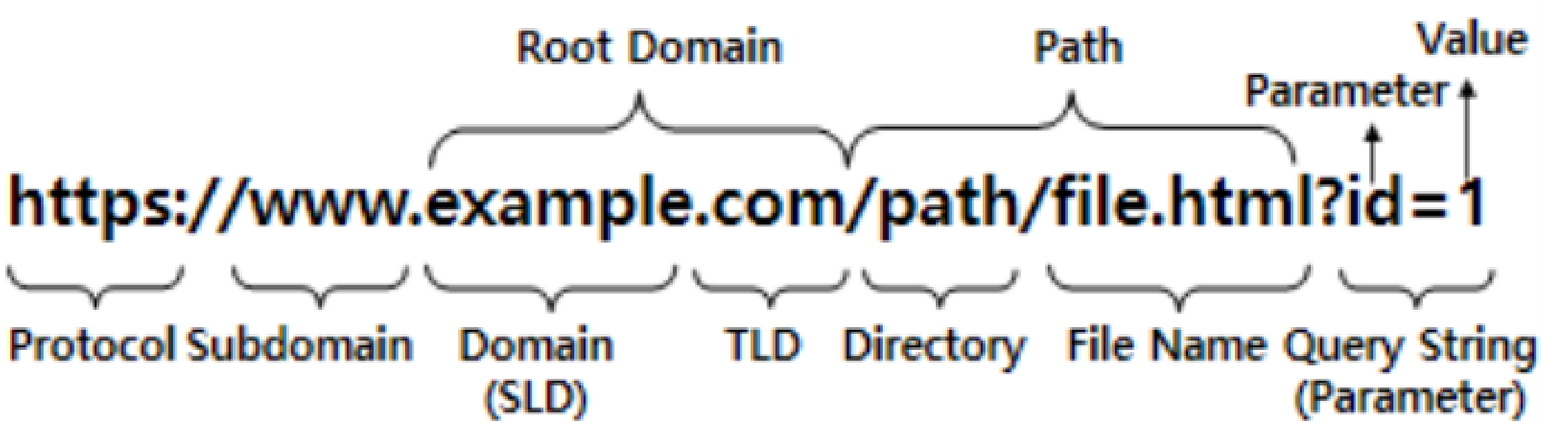
Root Domain : 인터넷 상에서 웹 사이트를 유일하게 식별할 수 있는 식별자 (. 으로 구분)

1. SLD : 보통 자신만의 이름을 가진 도메인
2. TLD : 목적과 국가를 나타내는 도메인 (com - 영리, net - 네트워크, kr - 한국)

URL Lexical Feature 기반의 DL-ML Fusion Hybrid

기술적 배경지식

URL : 인터넷상에서 자원이 어디 있는지 알려주기 위한 규약 (RFC 1738)



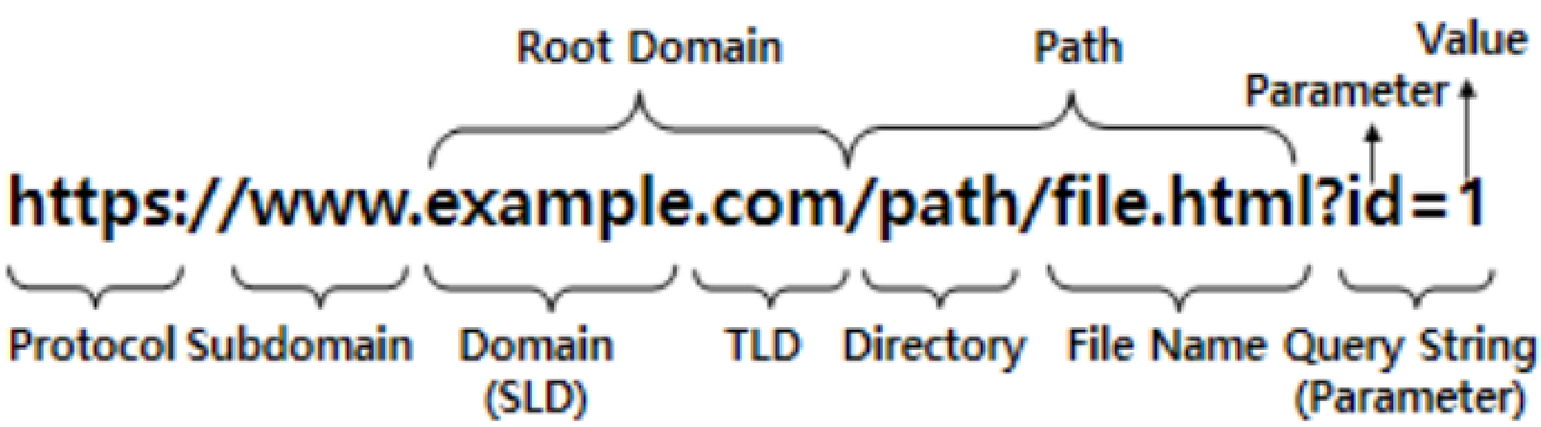
Path : 요청하는 파일의 위치 경로 (/ 로 구분)

1. Directory : 디렉토리 이름
2. File Name : 파일 이름

URL Lexical Feature 기반의 DL-ML Fusion Hybrid

기술적 배경지식

URL : 인터넷상에서 자원이 어디 있는지 알려주기 위한 규약 (RFC 1738)



Query String : 사용자가 웹 서버에 요청할 때 같이 전달되는 변수 이름 + 값 구성
(? 으로 구분, 여러 개의 변수 + 값은 &으로 구분)

URL Lexical Feature 기반의 DL-ML Fusion Hybrid

기술적 배경지식

6가지의 URL 위조 방식

중간 중간 다른 문자 삽입(오타)
민감하고 신뢰할 만한 단어 삽입
다른 도메인 삽입
특수 문자, IP 주소, 숫자 삽입
짧은 URL 변환
확장자를 가진 악성 실행파일명 포함

URL Lexical Feature 기반의 DL-ML Fusion Hybrid

기술적 배경지식

4가지 URL feature Group

URL의 전체 or 구성 요소의 길이
잘 알려진 브랜드 이름, 정상 URL의 도메인 등을 subdomain에 삽입하는 경우가 많음

www.facebook.com -> www.malware.facebook.com

URL Lexical Feature 기반의 DL-ML Fusion Hybrid

기술적 배경지식

4가지 URL feature Group

URL에 포함된 특정 문자 개수 특수 문자(!,@,#,% 등)으로 위조 및 기능 악용

www.facebook.com -> www.face.book.com

www.facebook.com@malware.com -> malware.com

URL Lexical Feature 기반의 DL-ML Fusion Hybrid

기술적 배경지식

4가지 URL feature Group

URL에 특정 단어 포함 여부

계정 탈취 : login, admin, confirm
악성 코드 유포 : .exe
무료호스팅 서비스 : 000webhost

www.facebook.com/login

www.facebook.com/malware.exe

URL Lexical Feature 기반의 DL-ML Fusion Hybrid

기술적 배경지식

4가지 URL feature Group

URL의 숫자 구성 비율 도메인 주소

www.facebook.com -> www.fac3b00k.com

URL Lexical Feature 기반의 DL-ML Fusion Hybrid

기술적 배경지식

악성 URL 탐지모델 머신러닝

이진 분류 : 정상 URL - 악성 URL

$$T = \{(u_1, y_1), \ldots, (u_T, y_T)\} \longrightarrow X_t \in \mathbb{R}^n \qquad u_t \longrightarrow X_t$$

$$\hat{y}_t = sign(f(X_t)) \longrightarrow y_t \in \{1, -1\}$$

URL Lexical Feature 기반의 DL-ML Fusion Hybrid

모델 제안 방법

딥 러닝의 일부 layer를 feature 추출 과정에 활용

-> 데이터 분석가의 역할 자동화 및 의존성 해결

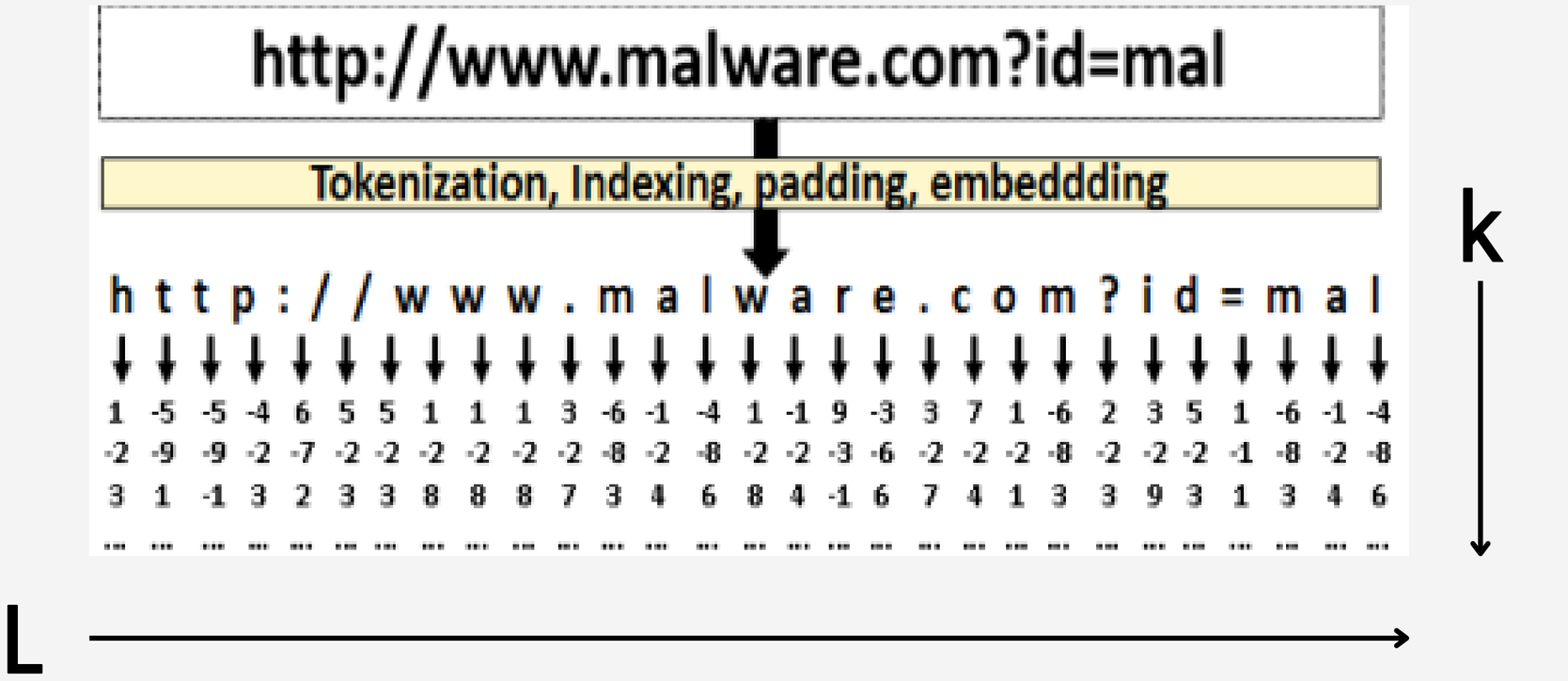
1. 딥러닝으로 URL 데이터 학습 → 성능 평가 과정 반복
이미 존재하는 경우 재활용

2. 딥러닝 모델의 input 층을 시작으로 한 층씩 추가하면서 하나의 FEB를 구성하는 방식 → 여러 개의 독립된 FEB 추출

3. 추출된 FEB로 URL lexical feature를 추출하여 고전 머신러닝으로 기계학습 및 성능 평가를 수행 반복

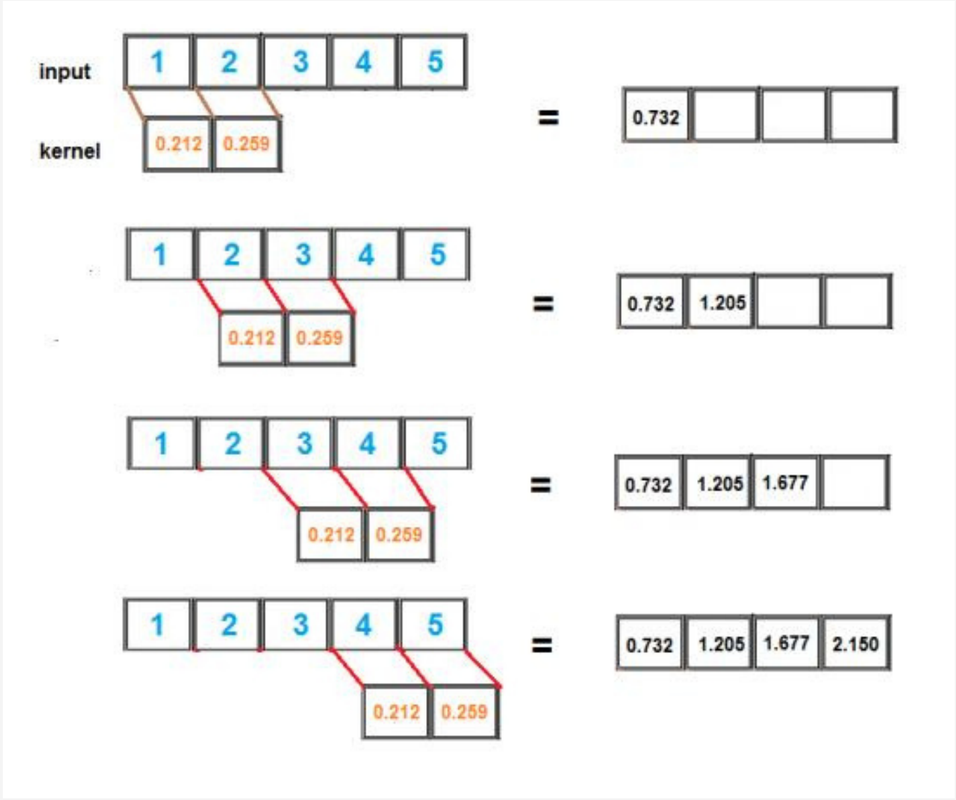
URL Lexical Feature 기반의 DL-ML Fusion Hybrid

탐지 과정제안 방법



URL Lexical Feature 기반의 DL-ML Fusion Hybrid

탐지 과정제안 방법

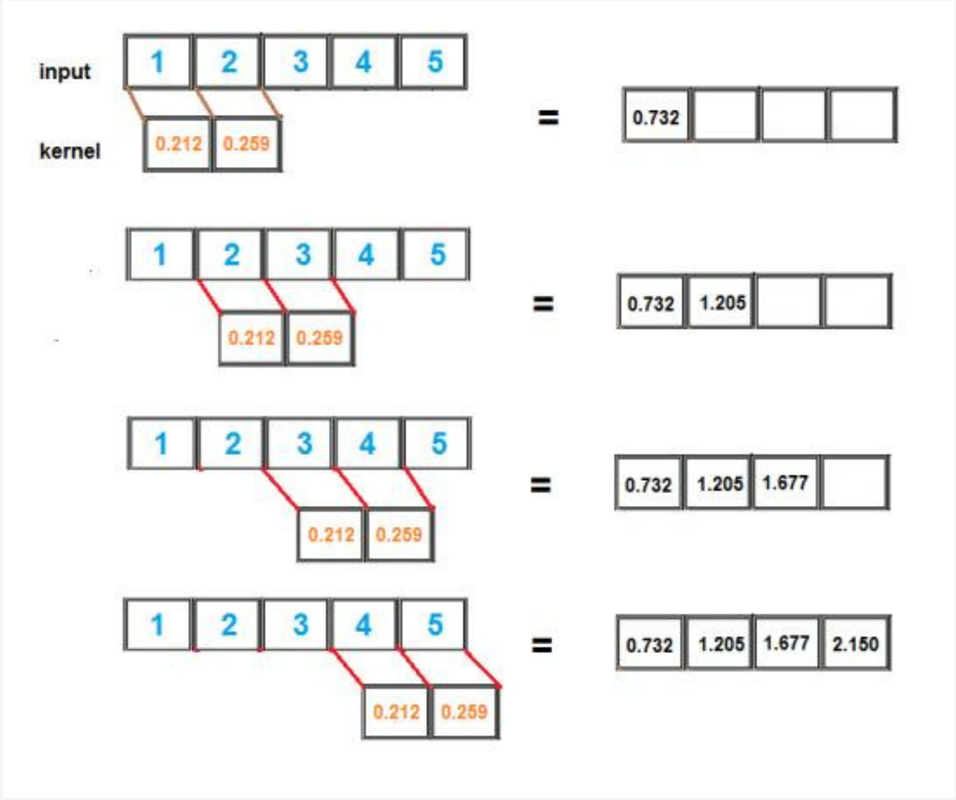


1D Convolution

$$c_i^l = f\left(b_i + \sum_{h=1}^{H} w_h x_{i+h-1}\right)$$

URL Lexical Feature 기반의 DL-ML Fusion Hybrid

탐지 과정제안 방법



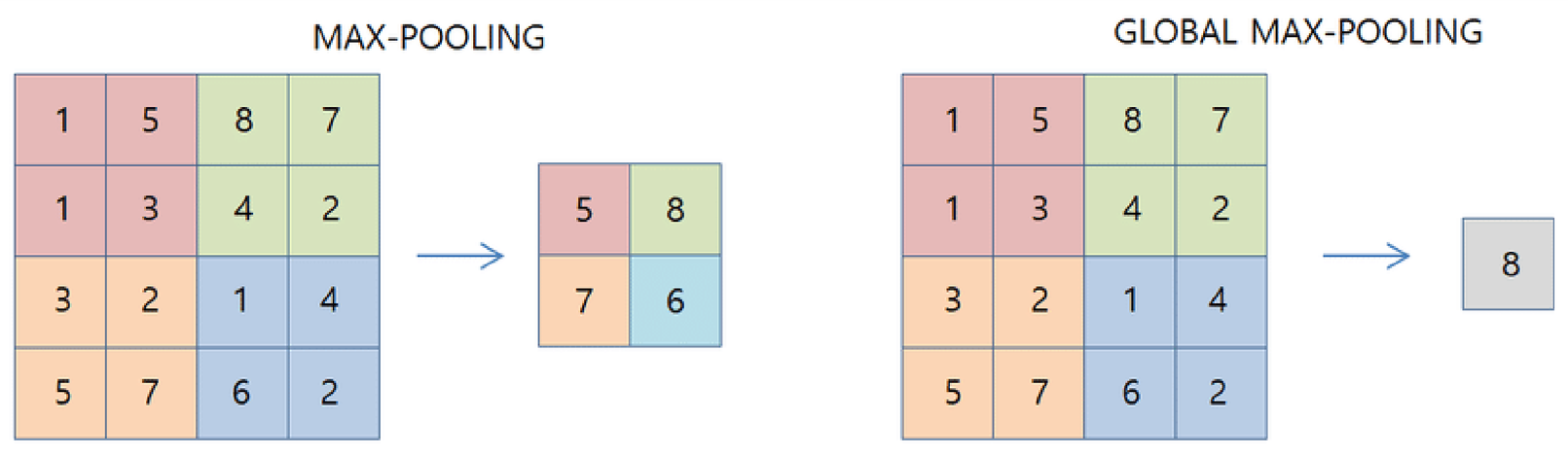
1D Convolution

$$c_i^l = f\left(b_i + \sum_{h=1}^{H} w_h x_{i+h-1}\right)$$

$$C = [c_1, c_2, \cdots, c_{L-H+1}]$$

URL Lexical Feature 기반의 DL-ML Fusion Hybrid

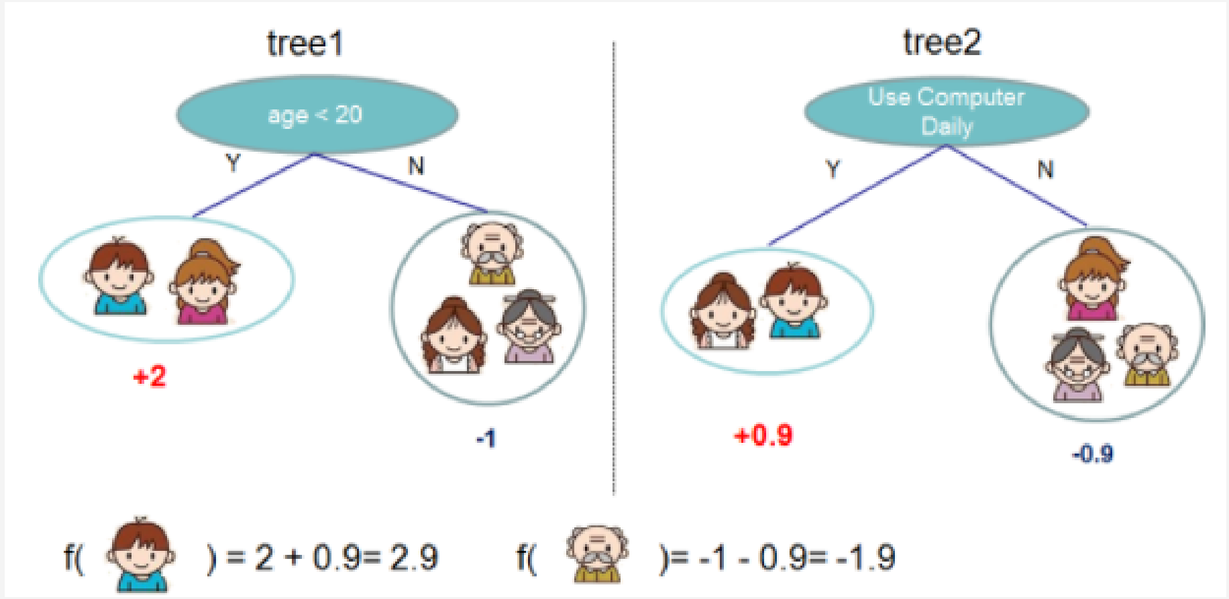
탐지 과정제안 방법



Global Max Pooling

URL Lexical Feature 기반의 DL-ML Fusion Hybrid

탐지 과정제안 방법



$$y'_i = \sum_{k=1}^{K} f_k(x_i), f_k \in (F)$$

XGB (CART 모델 기반)

URL Lexical Feature 기반의 DL-ML Fusion Hybrid

탐지 과정제안 방법

$$obj = \sum_{i=1}^{n} l(y_i, y'_i) + \sum_{k=1}^{K} \Omega(f_k)$$

training loss　　과적합방지

XGB (CART 모델 기반)

URL Lexical Feature 기반의 DL-ML Fusion Hybrid

탐지 과정제안 방법

생성된 각각의 트리 모델에
input(x) -> output(예측값)

Σ예측값 -> URL 탐지 유무 결정

최종

URL Lexical Feature 기반의 DL-ML Fusion Hybrid

실험 결과, 결론

딥러닝으로 feature을 추출하고 머신러닝에 훈련시켰더니,
비용 절감과 성능이 높아졌다.