Tor 네트워크에서의 딥러닝 모델을 활용한 키워드 핑거프린팅 공격기법

황채원‡O, 전혜승‡, 김고운, 홍지우, 강호성, 오세은*
이화여자대학교
{ifetayo, cathyjeon, gowoon1230, hjiwoo0914, kanghsung717, seoh}@ewhain.net

DKF: Employing Deep Learning for Keyword Fingerprinting Attacks on Tor

Chai Won Hwang‡O, Hae Seung Jeon‡, Goun Kim, Ji Woo Hong, Ho Sung Kang, Se Eun Oh
Ewha Womans University, Seoul, South Korea
{ifetayo, cathyjeon, gowoon1230, hjiwoo0914, kanghsung717, seoh}@ewhain.net

요 약

Tor 네트워크의 트래픽을 분석하여 사용자가 접속하는 웹사이트를 유추하는 Website Fingerprinting(WF) 공격과 함께 검색엔진에 입력한 검색어를 추적하는 Keyword Fingerprinting(KF) 공격이 Tor 사용자의 익명성을 침해할 수 있음을 보여주는 다양한 연구들이 수행되어 왔다. 키워드는 사용자의 관심사, 습관, 질병 등 민감한 정보를 직접적으로 내포하고 있어 WF보다 프라이버시 침해가 심각할 수 있다. 본 연구에서는 State-of-the-Art KF 모델 성능을 향상시킨 딥러닝 기반 DKF(Deep Keyword Fingerprinting) 모델을 제안하고 이를 효과적으로 학습시키기 위한 대규모의 최신 데이터 세트를 수집해 Tor 네트워크가 여전히 KF 공격에 취약함을 보인다.

1. 서론*

Tor 네트워크는 매일 수백만 명의 사용자가 접속하는 익명화 네트워크로, 전 세계에 분포된 약 5,000개의 어니언 라우터 중 3개를 무작위로 뽑아 통신 서킷을 생성한다. 그림 1처럼 3개의 라우터를 통해 통신을 릴레이하는 클라이언트-서버 통신을 제공한다. 따라서 통신의 양 종단을 숨길 수 있어서 사용자의 온라인 활동의 익명성을 보다 잘 보장할 수 있다. 강화된 보안을 위해 클라이언트에서 데이터는 삼중 암호화가 되어 전송되고 각 라우터는 한 계층 씩 복호화 하여 다음 도착지로 데이터를 전송한다. 마지막 릴레이 구간인 3번째 라우터와 서버 간의 통신 구간은 삼중 암호화가 모두 복호화 되지만, HTTPS로 통신은 암호화되어 데이터를 보호한다,

그러나 Tor 네트워크의 암호화된 트래픽에서도 패킷의 방향, 크기, 도착 시간 정보와 같은 트래픽 메타 데이터를 분석하면 통신의 종단인 웹 서버, 즉 사용자가 어떤 웹사이트에 방문하는지 알아내는 WF 공격이 가능하다.

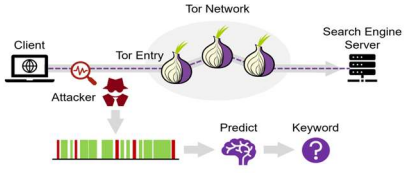

그림 1. 토르 네트워크 상에서의 KF 공격 시나리오

WF 공격에서 공격자는 사용자가 전송하는 트래픽의 메타데이터로부터 피처를 추출하여 인공지능 모델을 학습시키고, 이를 통해 추후 사용자 트래픽을 토대로 접속한 웹사이트를 추측한다. WF 공격은 랜덤 포레스트 모델을 비롯한 전통적인 머신러닝 기반 공격부터 Convolutional Neural Networks (CNNs) 등의 딥러닝 기반 공격까지 활발히 연구가 이루어지고 있다.

WF 공격 연구를 확장하여 사용자가 검색엔진에 입력한 검색어를 알아내는 KF 공격이 제안되었으며, 대표적으로 2017년 Oh et al. [1]이 제안한 Tor 네트워크에서의 Support Vector Machine (SVM) 기반 KF 공격이 있다. 2018년부터 딥러닝 기반 WF 모델들이 제안됨에 따라 핑거프린팅 성능이 향상되어 95% 이상의 정확도로 접속한 웹사이트를 식별할 수 있게 되었다. 따라서 본 연구에서는 KF에 딥러닝 모델 적용 가능성을 최초로 탐색하고자 한다.

본 논문의 기여는 다음과 같다. 1) KF 공격을 위한 딥러닝 기반의 Deep Keyword Fingerprinting (DKF) 모델과, 효과적인 모델 학습을 위한 새로운 Window-overlap(Winlap) 피처를 제안했다. 2) Tor 네트워크에서 두 검색엔진(Bing, DuckDuckGo)의 대규모 키워드 트래픽 데이터 세트를 수집하였다. 3) 수집한 데이터 세트를 사용하여 DKF와 선행 연구된 모델들의 성능을 비교 분석하여

‡ 공동 1저자
* 교신 저자

* 이 성과는 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임 (No. RS-2023-00222385, RS-2022-00166669)

딥러닝 모델이 KF 성능을 향상시켰음을 보이고 더 나아가 이를 무력화할 수 있는 방어기법 개발이 시급함을 보인다.

2. Tor 네트워크에서의 WF 및 KF 공격

최근 다양한 머신러닝 모델을 활용한 WF 공격 연구가 활발하게 이루어지고 있다. 이 중 랜덤 포레스트와 k-Nearest Neighbors (k-NN)를 함께 사용한 k-FP 모델[2]은 심층적인 피처 분석을 통해 150개의 Hand-Crafted 피처들을 추출하였고 다수의 의사결정 트리를 통해 효과적인 피처를 선택하여 WF 모델 성능을 높였다.

이후 여러 연구에서 딥러닝 모델을 WF 공격에 활용하기 시작하였다. Deep Fingerprinting (DF) [3] 모델은 첫 딥러닝 모델 기반 WF 기법 중 하나로 깊은 1D-CNN 모델을 제안하여 매우 높은 정확도를 달성하였고 당시 제안되었던 WF 방어 기법들까지 무력화하였다. 이후 DF 모델 성능을 향상시키기 위한 새로운 피처들이 제안되었다. Rahman et al.[4]은 패킷 도착 시간과 패킷 전송 방향 정보를 결합한 Tik-Tok 피처를 추출하여 DF 모델 성능을 개선하였다. 패킷 전송 방향은 클라이언트에서 서버로의 방향은 양수 (+) 로, 서버에서 클라이언트로의 방향은 음수 (-) 로 표현한다.

SVMResp[1]은 Tor 네트워크에서 시도된 첫 KF 공격 모델로 사용자 키워드 트래픽의 서버로부터의 마지막 Burst(같은 방향에서 연속적으로 오는 패킷들의 집합)에서 피처를 추출하여 SVM 모델을 학습시켜 기존 WF의 트래픽 분석 기반 핑거프린팅 공격이 사용자가 검색한 키워드를 유추하는 KF 공격으로 확장될 수 있음을 보였다.

3. 데이터 세트 수집

본 연구에서는 Bing과 DuckDuckGo 두 개의 검색 엔진에서 키워드를 검색하였을 때의 트래픽을 수집하여 데이터 세트를 구축하였다. Tor 브라우저에서 검색엔진 웹 사이트에 접속하면 트래픽 수집이 시작되고 키워드를 입력하면 검색 결과 웹페이지가 로딩 된 후 트래픽 수집이 종료된다. 키워드 검색의 자동화를 위해 Selenium, 트래픽 수집을 위해 tcpdump를 사용하였다.

공격자가 검열하고 싶은 키워드를 Monitored, 그 외의 키워드를 Unmonitored 데이터 세트로 구분하여 데이터 세트를 구성하였다. Monitored 데이터 세트는 키워드 조사 툴[1]에서 제공하는 키워드 리스트 중 구글에서 많이 검색된 상위 273개의 키워드를 사용하였고, 각 키워드의 트래픽을 1,200번씩 수집했다. Unmonitored 데이터 세트는 Monitored 데이터 세트와의 중복을 피해 총 110,000개 키워드의 트래픽을 수집하여 구축했다. 전처리를 거쳐 monitored 데이터 세트는 258개의 키워드별 1,000개씩 수집된 트래픽으로, unmonitored 데이터 세트는 93,443개의 트래픽으로 구성했다.

기존의 웹사이트 트래픽 수집 툴[2]을 확장하여 검색엔진의 검색 박스에 키워드를 입력하는 로직을 자동화하고 CAPTCHA를 우회하는 로직을 추가하여, 2023년 5월부터 2023년 8월까지 8개의 가상머신을 통해 병렬적으로 수집하였다.

[1] https://keywordtool.io/

4. Deep Keyword Fingerprinting

4.1. Winlap 피처

키워드 트래픽의 딥러닝 모델 학습을 위한 피처로 선행연구 모델인 Robust Fingerprinting (RF) 모델[5]에서 사용한 Traffic Aggregation Matrix (TAM) 피처를 확장하여 Winlap 피처를 제안한다. TAM은 트래픽을 일정한 시간 간격(윈도우)으로 나누고 각 윈도우 내의 수신 패킷과 송신 패킷 수를 2차원 행렬로 사상하는 피처이다. Winlap은 그림 2와 같이 균등하게 분할된 윈도우 내의 수신 패킷 수 배열과 수신 패킷 사이즈의 합을 계산한 두 개의 배열을 병합하여 (Concatenation) 1차원 행렬 피처를 생성하였다.

우리는 Winlap 피처를 우리 모델과 데이터 세트에 최적화하여, 각 트래픽을 최대 길이 60초로 만들어 1,000개의 윈도우로 분할하고 각 윈도우가 다음 윈도우와 50%씩 겹쳐지도록 하였다.

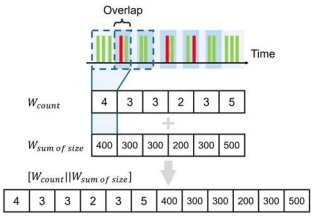

그림 2. Winlap 피처

4.2. DKF 모델

DKF모델은 DF 모델을 확장하여 설계하였다. DF에서는 2개의 1D-CNN 레이어로 구성된 블럭을 4번 반복하여 모델을 구축하였고 Batch Normalization과 Dropout 레이어를 추가하여 과적합을 방지하였다. DKF 모델에서는 그림 3과 같이 블럭을 5번 반복하여 조금 더 깊은 1-D CNN 모델을 구축하였다. 이에 모델의 구조가 깊어지며 발생하는 과적합을 완화하기 위해, 각 블럭의 Dropout 되는 노드 수를 증가시켰다. 또한, 배치 사이즈를 128에서 32로 줄여 모델의 Loss를 효과적으로 감소시켰다.

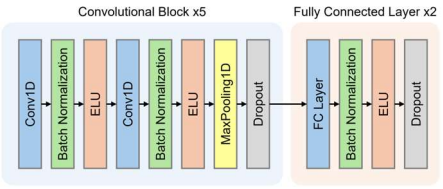

그림 3. DKF 모델 아키텍처

5. 실험

5.1. Closed-World (CW) 실험

CW 실험은 공격자가 검열하고자 하는 키워드 목록 내의 키워드만 사용자가 검색하는 상황을 가정한다. 검열 대상 키워드 목록 중 어떤 키워드를 검색하는지 알아내는 다중 분류 실험(Multi-Class Classification)을 설정하였다. 총 258개의 검열 키워드 트래픽을 대상으로 258개의 클래스로 식별하는 것을 목표로 한다.

[2] https://github.com/notem/tor-browser-crawler

표 1은 클래스 수를 늘려가며 실험한 각 모델의 다중 분류 정확도를 나타낸다. DKF는 이 중 모든 경우에서 가장 우수한 성능을 보였다. 흥미롭게도 클래스 수 증가에 따라 DKF의 정확도가 DuckDuckGo에서 Bing 보다 높게 나오며 이는 KF 공격에 DuckDuckGo가 조금 더 취약함을 보여준다.

표 1. Closed World에서의 모델 별 다중 분류 정확도

Model	Dataset	60 class	200 class	258 class
DKF	Bing	**84.65**	**63.36**	**59.19**
	DuckDuckGo	**82.28**	**70.43**	**65.98**
Tik-Tok	Bing	56.69	46.79	45.56
	DuckDuckGo	50.83	32.12	27.31
k-FP	Bing	37.79	30.66	29.22
	DuckDuckGo	34.78	16.72	14.91
SVMResp	Bing	7.95	4.04	3.68
	DuckDuckGo	6.03	2.28	1.81

5.2. Open-World (OW) 실험

OW는 사용자가 공격자의 검열 키워드 목록 이외에도 다양한 키워드를 검색하는 현실적인 환경을 가정한 실험이다. 공격자가 검열하는 Monitored 키워드와 공격자가 검열하지 않는 Unmonitored 키워드를 구분하는 이진 분류 (Binary Classification) 실험으로 진행하였다. 본 실험의 데이터는 Bing 검색엔진에서 수집된 Monitored, Unmonitored 데이터 세트를 사용하였다. Monitored 데이터 세트로는 각 1,000개의 인스턴스를 가진 100개의 키워드가 사용되었다. Unmonitored 데이터 세트의 경우 학습에는 2만 개를 고정적으로 사용했고, 테스트 세트의 경우 Unmonitored 데이터를 1만, 4만, 7만 개로 늘려가며 실험하여 총 9만 개의 키워드를 사용하였다.

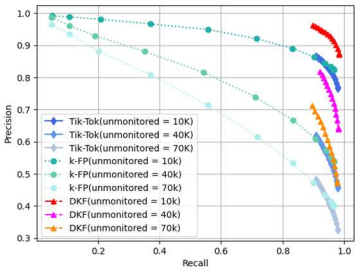

그림 4. Open World에서 unmonitored 데이터 증가에 따른 각 모델의 P-R curve

그림 4는 Unmonitored 테스트 데이터가 각각 1만, 4만, 7만 개일 때 DKF, Tik-Tok, 그리고 k-FP의 정밀도-재현율 곡선(Precision-Recall Curve)를 나타낸다. DKF와 Tik-Tok은 0부터 1까지의 예측 확률값을, k-FP는 k-NN에서 계산된 Hamming Distance 값들을 기반으로 임계값을 설정해 정밀도와 재현율을 계산했다. 재현율이 비슷할 때의 정밀도 값은 모든 경우에서 DKF가 가장 높아 State-of-the-Art의 성능을 보여주었다.

6. 결론

2006 년 웹 포털 AOL 에서 비식별화된 사용자의 검색어 웹 로그가 공개되어 개인정보보호법 위반으로 인한 집단 소송 사건이 발생했다. 로그가 가명화 되었음에도 불구하고 식별이 가능할 만큼, 검색어 내역은 사용자의 아이덴티티와 밀접하게 관련되어 있었기 때문이다. 이러한 사용자 식별 문제를 방지하기 위해 익명화 네트워크인 토르가 등장했지만, 패킷의 메타 데이터로 사용자의 활동을 예측하는 핑거프린팅 공격은 여전히 가능하다. 키워드 핑거프린팅 공격이 야기할 수 있는 프라이버시 침해의 심각성에도 불구하고, 이에 딥러닝을 적용한 공격의 가능성은 아직 탐색되지 않았다.

본 논문에서는 최초로 토르 네트워크상에서의 KF 공격에 딥러닝을 적용한 DKF 모델과 Winlap 피처를 제안하여 핑거프린팅 공격의 정확성을 높였다. Winlap 피처는 트래픽을 일정한 시간 윈도우로 나누어 패킷의 수와 크기 정보를 활용한다. 따라서 기존에는 광범위한 시간 분포를 가지던 트래픽을 일정한 시간 단위로 정규화하여 트래픽의 순차적인 정보를 유지하면서도 트래픽 패턴을 규칙적이고 추상화된 형태로 모델에 제공할 수 있다.

DKF 는 기존 WF 및 KF 연구들에 비해 우수한 성능을 보였으며, 토르 네트워크가 딥러닝을 적용한 KF 공격에 취약함을 증명했다. CW 환경에서 두 가지 검색엔진 데이터에 대해 모두 80% 이상의 정확도를 보였으며, OW 환경에서 역시 State-of-the-Art 를 달성하였다. 따라서 KF 공격이 일으키는 심각한 프라이버시 침해에 대한 향후 연구로 토르 네트워크의 KF 공격에 대한 방어기법 개발이 필요하며 본 연구의 KF 공격에 대한 취약점 분석이 방어 기법 연구에 도움을 줄 수 있을 것으로 기대한다.

참 고 문 헌

[1] Oh, Se Eun, Shuai Li, and Nicholas Hopper, "Fingerprinting keywords in search queries over tor." *Proceedings on Privacy Enhancing Technologies*, 2017.

[2] Hayes, Jamie, and George Danezis, "k-fingerprinting: A robust scalable website fingerprinting technique." in *25th USENIX Security Symposium (USENIX Security 16)*, 2016.

[3] Sirinam, Payap, et al., "Deep fingerprinting: Undermining website fingerprinting defenses with deep learning." in *Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security*, 2018.

[4] Rahman, Mohammad Saidur, et al., "Tik-tok: The utility of packet timing in website fingerprinting attacks." *Proceedings on Privacy Enhancing Technologies*, 2020.

[5] Shen, Meng, et al., "Subverting website fingerprinting defenses with robust traffic representation." in *32nd USENIX Security Symposium (USENIX Security 23)*, 2023.