FiLM 기반 CNN을 활용한 Website Fingerprinting 공격 모델 연구

장채은, 김은경

국립한밭대학교

chaeeun@edu.hanbat.ac.kr, ekim@hanbat.ac.kr

A Study on Website Fingerprinting Attack Model Using FiLM-based CNN

Chae Eun Jang, Eunkyung Kim

Hanbat National University

요 약

본 논문은 암호화된 네트워크 트래픽에서의 Website Fingerprinting 성능 향상을 위해, 컴퓨터 비전 분야에서 사용되던 Feature-wise Linear Modulation(FiLM)을 1차원 CNN 모델에 적용하는 새로운 접근법을 제안한다. 제안 모델은 GRU로 추출한 트래픽 시퀀스의 문맥 정보와 통계적 특성을 조건 벡터로 사용하여, CNN의 중간 feature map을 동적으로 변조한다. 이 조건부 메커니즘은 입력 데이터의 고유한 맥락에 적응하여 모델의 표현력을 향상시킨다. 공개 TLS 데이터셋을 이용한 실험 결과, 제안 모델은 베이스라인 CNN 모델 대비 약 9%p에 달하는 유의미한 정확도 향상을 달성하며 그 우수성을 입증하였다. 본 연구는 시각 지능 분야에서 검증된 방법론이 네트워크 트래픽과 같은 1차원 시계열 데이터 분석에도 매우 효과적임을 입증하며, 이종 도메인 간의 성공적인 기술 전이 가능성을 제시한다.

Ⅰ. 서 론

Website Fingerprinting은 암호화된 네트워크 환경에서 패킷의 크기, 방향, 시간적 패턴 등을 분석하여 사용자가 접속한 웹사이트를 추정하는 공격 기법이다. Tor나 VPN과 같은 익명화 네트워크 환경에서도 이 기법을 통해 사용자가 접속한 웹사이트를 식별해낼 수 있음을 다수의 연구가 입증해왔다[1],[2],[3]. 인공지능 기술이 발전함에 따라, 최근에는 트래픽 분석 분야에 CNN과 같은 딥러닝 모델을 도입하여 정확도를 향상시키고 있다. 하지만 CNN은 모든 입력 데이터에 대해 동일한 필터를 공유하는 정적인 방식으로 동작하여, 각 트래픽이 가진 고유한 맥락적 차이를 동적으로 반영하는 데에는 명백한 한계가 있다.

이에 따라 트래픽 데이터에 CNN을 적용할 시 데이터 내 맥락적 변화와 조건별 특성을 반영할 수 있는 새로운 접근법이 요구되며, 본 연구에서는 이 문제를 해결하기 위해 Feature-wise Linear Modulation(FiLM) 기법[4]에 주목한다. FiLM은 이미지-텍스트 멀티모달 추론 분야에서 제안된 방법으로, 외부 조건에 따라 feature map을 동적으로 조정할 수 있도록 설계되었다. 본 논문에서는 이미지 도메인에서 사용되는 FiLM기법을 1차원 네트워크 트래픽 분석 문제에 맞게 변형하고 적용하여 효과적인 Website Fingerprinting을 수행할 수 있도록 한다.

Ⅱ. 본론

1. 연구배경

1.1 Website Fingerprinting

Website Fingerprinting은 암호화된 네트워크 환경에서 트래픽을 분석해 사용자가 접속하고 있는 웹사이트를 식별하는 공격 기술이다[2]. Tor나 VPN 등의 네트워크는 사용자의 익명성과 개인정보를 보호한다는 특징이 있으나, 이러한 익명화 네트워크 환경 또한 Website Fingerprinting 기술을 통해 사용자가 접속한 웹사이트를 식별해낼 수 있다는 연구가 진행되어왔다. 최근에는 인공지능 기술의 발전으로 Website Fingerprinting에 기계학습을 도입하여 공격 모델의 성능을 고도화하고 있다[5]. 이러한 Website Fingerprinting에 기계학습을 적용한 연구의 경우[1],[2] 암호화된 페이로드 정보 대신 주로 웹사이트와 사용자 간의 패킷 송수신 방향, 패킷 크기, 버스트 등의 메타데이터를 통해 웹사이트를 식별해오고 있다.

1.2. FiLM

Feature-wise Linear Modulation(FiLM)[4]은 CNN과 같은 신경망에서, 외부 조건 정보를 활용하여 중간 feature map을 동적으로 조절하는 기법이다. FiLM은 각 feature map 채널에 대해 조건에서 생성된 두 개의 파라미터, 즉 스케일링 파라미터 γ와 쉬프팅 파라미터 β를 선형 변환 방식으로 적용한다. 이 기법은 필터 기반의 feature 추출 과정에 조건별 적응성을 부여한다는 점이 핵심이다. 즉, 입력 조건에 따라 다르게 feature를 강조하거나 억제할 수 있다. 이를 통해 복잡한 문맥이나 태스크 별 특성에 맞는 유연한 표현이 가능하다. 이러한 FiLM 기법은 자연어 질문을 조건으로 하여 이미지 feature map에 변조를 적용해 복합 시각 추론을 성공적으로 수행했으며, 간단하면서도 강력한 성능 향상을 보였다.

따라서 본 연구에서는 입력 시퀀스로부터 추출한 문맥 정보(GRU 임베딩)와 통계적 특성을 결합하여 조건 벡터를 생성한다. 이 벡터를 활용해 여러 CNN 레이어의 중간 feature map에 FiLM 변조를 적용함으로써, 주로 시각 데이터 분야에서 활용되던 FiLM 기법을 1차원 트래픽 데이터 분석에 맞게 확장한다. 이를 통해 Website Fingerprinting 공격 모델의 표현력과 정확도를 높이는 것을 목표로 한다.

2. 실험

2.1 제안 방법

본 논문에서는 Website Fingerprinting 성능 향상을 위해 FiLM기법을 활용한 3-block 1D-CNN 모델을 사용한다. 이 모델은 128차원의 은닉 상태를 갖는 양방향 GRU 인코더를 통해 입력 시퀀스의 전역적 문맥을 임베딩하고, 이를 통계적 특성과 결합하여 조건 벡터를 구성한다. 이 조건 벡터로부터 동적으로 생성된 FiLM 파라미터(γ, β)는 32, 64, 128 채널을 갖는 각 컨볼루션 블록의 feature map에 적용되어, 채널별 affine 변환을 통

한 적응적 feature 변조를 수행한다.

이 방식을 통해 모델이 데이터의 맥락을 효과적으로 반영하여 표현력을 극대화하고, 이를 통해 Website Fingerprinting의 분류 정확도 향상을 목표로 한다.

2.2 환경 설정

실험에는 120개의 클래스로 구성된 ET-BERT[6]의 공개 TLS 데이터셋을 활용하였다. 총 46,732개의 PCAP으로부터 각 패킷의 크기(byte)를 추출하여 raw 시퀀스 데이터를 구축하였고, 이후 모든 시퀀스 길이를 513으로 표준화하였다. 이는 전체 데이터 중 95%가 513 이하의 길이를 갖는다는 분포 특성을 고려하여 설정한 값이다. 모든 시퀀스의 길이를 맞추기 위해 513보다 길이가 짧은 시퀀스는 0으로 패딩을 하고, 긴 시퀀스는 513 이후의 값을 잘라내었다. FiLM 레이어의 컨디셔닝을 위한 통계적 feature는 raw 시퀀스 데이터로부터 산출하였다. 모델의 훈련 및 평가는 PyTorch 프레임워크를 기반으로 수행하였으며, 데이터는 8:1:1 비율의 훈련, 검증, 테스트 세트로 무작위 분할하였다. 모델 학습에 사용된 주요 하이퍼파라미터는 [표 1]과 같다.

[표 1] 하이퍼파라미터 설정

hyperparameters	value
Batch Size	64
Number of Epochs	100
Optimizer	AdamW
Learning Rate	0.001
Weight Decay	1e-5
Betas	(0.9, 0.999)
Learning Rate Scheduler	ReduceLROnPlateau
Dropout Rate	0.5
GRU Hidden Size	128

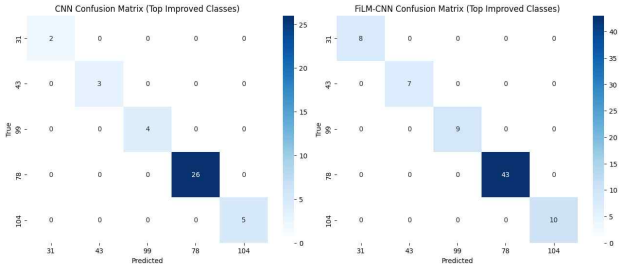
2.3 성능 비교 및 분석

[표 2]는 베이스라인 CNN과 제안된 FiLM-CNN의 성능 비교 결과를 나타낸다. 표에 나타난 바와 같이, 제안된 FiLM-CNN 모델은 베이스라인 CNN 모델 대비 평균 정확도에서 약 8.81%p의 유의미한 향상을 보였으며, 특히 주요 분류 성능 지표에서도 일관되게 개선된 결과를 보였다.

[표 2] 베이스라인 CNN과 FiLM-CNN의 성능 비교

model	Accuracy	Precision	Recall	F1score
Baseline CNN	67.49%	68.94%	67.49%	67.61%
FiLM-CNN	76.30%	76.81%	76.30%	76.23%

또한 [그림 1]은 베이스라인 CNN 대비 FiLM-CNN 모델에서 성능 향상이 두드러진 5개의 클래스를 선정하여, 이들 클래스에 대한 혼동 행렬 변화를 비교한 결과이다. 베이스라인 CNN의 경우, 일부 클래스에서 예측 정확도가 낮아 다른 클래스와 혼동되는 사례가 관찰되었으나, FiLM-CNN 모델에서는 각 클래스별 정확도가 크게 향상되었다. 이러한 결과는 FiLM 기법의 조건 기반 조절이 특정 클래스의 특징을 보다 뚜렷하게 강화하여, CNN 단독 모델 대비 정확도와 신뢰도를 동시에 높였음을 시사한다.



[그림 1] 베이스라인 CNN(좌)과 FiLM-CNN(우)의 예측 정확도 차이가 큰 5개 클래스 혼동행렬

Ⅲ. 결론

본 논문에서는 FiLM을 활용한 방식이 Website Fingerprinting 태스크에 효과적임을 제안하고 실험적으로 검증하였다. 제안된 모델은 베이스라인 CNN 모델 대비 주요 평가지표 전반에서 일관되게 우수한 성능을 보였으며, 특히 정확도에서 약 9%p에 이르는 성능 향상을 달성했다. 이는 정적인 feature 추출 방식과 달리, FiLM이 트래픽의 고유한 맥락을 조건으로 삼아 미세한 패턴의 차이를 식별하는 모델의 표현력을 강화한 결과이다.

본 연구는 특정 데이터셋 환경에서 수행된 한계가 있다. 따라서 향후 과제로는 실제와 유사한 다양한 네트워크 환경에서의 강건성을 검증하고, 모델 최적화 및 경량화 연구를 진행할 예정이다.

결론적으로, 본 연구는 이미지 처리 분야에서 주로 활용되던 FiLM과 같은 기법이 1차원 네트워크 트래픽 데이터 처리에도 효과적으로 적용될 수 있음을 시사한다. 이러한 접근법은 향후 네트워크 트래픽 분석뿐만 아니라, 유사한 시계열 및 1차원 데이터 처리 문제에 새로운 해결책을 제시할 것으로 기대된다.

ACKNOWLEDGMENT

"본 연구는 2025년 과학기술정보통신부 및 정보통신기획평가원의 SW중심대학사업의 연구결과로 수행되었음"(2022-0-01068)

참 고 문 헌

[1] V. Rimmer et al., "Automated website fingerprinting through deep learning," in Proc. Network and Distributed System Security Symposium (NDSS), 2018.

[2] P. Sirinam et al., "Deep fingerprinting: Undermining website fingerprinting defenses with deep learning," in Proc. ACM SIGSAC Conf. on Computer and Communications Security (CCS), 2018, pp. 1928 - 1943.

[3] T. Wang et al., "Effective attacks and provable defenses for website fingerprinting," in Proc. USENIX Security Symposium, 2014, pp. 143 - 157.

[4] E. Perez et al., "FiLM: Visual reasoning with a general conditioning layer," arXiv preprint arXiv:1709.07871, 2017.

[5] B. Rexha et al., "Unveiling the digital fingerprints: Analysis of internet attacks based on website fingerprints," arXiv preprint arXiv:2409.03791, 2024.

[6] X. Lin et al., "ET-BERT: A contextualized datagram representation with pre-training transformers for encrypted traffic classification," in Proc. The Web Conference (WWW), 2022, pp. 633 - 642.