Automated Website Fingerprinting through Deep Learning

Vera Rimmer\textsuperscript{1}, Davy Preuveneers\textsuperscript{1}, Marc Juarez\textsuperscript{2}, Tom Van Goethem\textsuperscript{1} and Wouter Joosen\textsuperscript{1}

NDSS 2018 – Feb 19th (San Diego, USA)
Website Fingerprinting
Anonymous Communication through Tor

- All (secure) communication protocols expose metadata
  - timing, size of packets, identities, locations, addresses, communication patterns → reveal private information

- Anonymity tools relay traffic through protected communication channels
  - The Onion Router (Tor)
Website Fingerprinting

- Side-channel attack that reveals user’s browsing activity
- Adversary is a local eavesdropper

- ISP
- Autonomous Systems
- Local network admins
- Wi-Fi hotspot owners
Website Fingerprinting

- Number of packets
- Average packet size
- % of incoming packets
- Timing of packets
- ...
Website Fingerprinting
Website Fingerprinting

"Closed world" of websites
Website Fingerprinting Pipeline

Communication patterns
Website Fingerprinting Pipeline

Communication patterns

Feature Extraction
Website Fingerprinting Pipeline

Communication patterns → Feature Extraction → Machine Learning
Website Fingerprinting Pipeline

- Feature Extraction
- Machine Learning
- Identification
State-of-the-Art Attacks

- **kNN (Wang et al., 2014)**
  - 3,000 features picked through heuristics (total size, total time, number of packets, packet ordering, traffic bursts...)
  - Classifier: k-Nearest Neighbors

- **k-Fingerprinting (Hayes et al., 2016)**
  - 150 features selected from Wang’s through the analysis of feature importance
  - Classifier: Random Forest and k-Nearest Neighbors

- **CUMUL (Panchenko et al., 2016)**
  - 100 features, interpolation points of the cumulative sum of packet lengths
  - Classifier: Support Vector Machine
Website Fingerprinting Arms-race

Main focus of prior work:
- Manual engineering
- Intellectual effort
- Difficult and expensive

AND

Success of attacks is defined by the set of engineered features
Website Fingerprinting Arms-race

Concealing these features creates a countermeasure

Communication patterns

Feature Extraction

Machine Learning

Identification
Website Fingerprinting Arms-race

Feature Extraction

Machine Learning

Identification

New attack exploits other, still visible features
Website Fingerprinting Arms-race

Communication patterns

Feature Extraction

Feature Extraction

Machine Learning

Identification
Website Fingerprinting

Communication patterns → Feature Extraction → Machine Learning → Identification

Alternative?
Website Fingerprinting

- Communication patterns
- Feature Extraction
- Machine Learning
- Identification
- Deep Learning
Deep Learning for WF
Why Deep Learning?

› Automatic feature learning from raw input
  › Obviates hand-engineering of features
  › Adaptive to changes in patterns

› Limited transparency and interpretability
  › Learned features are implicit and abstract

› Efficient, easily distributed and parallelized
Deep Learning based WF

› Data Collection
  › DL requires a lot of training data

› Deep Neural Network choice
  › Choosing the best suited deep learning algorithm

› Hyperparameter Tuning and Model Selection
  › Tuning of heavily parameterised models
Data Collection

- Built a distributed crawler
  - captures timing, direction and sizes of TCP packets
- 2,500 traces for each 900 top Alexa most popular sites: largest-ever dataset
- Closed worlds: $\text{CW}_N$ datasets, where $N$ is the number of sites
Deep Neural Networks

› Choice of a Deep Neural Network (DNN) suited for the input data
  › 1D sequences of incoming and outgoing Tor cells encoded as 1 and -1

› Explored 3 major types of DNNs:
  › feedforward: Stacked Denoising Autoencoder (SDAE)
    • learns from the *continuous structure* through dimensionality reduction
  › convolutional: Convolutional Neural Network (CNN)
    • learns from the *spatial structure* through convolutions and subsampling
  › recurrent: Long Short Term Memory (LSTM)
    • learns from the *temporal structure* (time-series) through internal memory
Evaluation and Results
Re-evaluation of Traditional Attacks

![Bar chart showing accuracy percentages for Wang's dataset and CW_{100} across different methods: Wang-kNN, k-FP, and CUMUL.]

- Wang-kNN: 92.87%
- k-FP: 92.47%
- CUMUL: 95.43%

Wang's dataset and CW_{100} are represented by different colors in the chart.
Re-evaluation of Traditional Attacks

Best performant on the closed world, most practical attack
Closed World
Closed World

Overall, comparable with the state-of-the-art
Closed World

CW100: CUMUL still outperforms all attacks, followed by CNN
Closed World

Accuracy falls as the number of websites increases
Closed World

CW900: SDAE outperforms state-of-the-art
Number of Traces per Website

![Graph showing the number of traces per website against accuracy and number of instances per website. The graph includes lines for SDAE, CNN, and LSTM models.](image-url)
Number of Traces per Website

LSTM takes longer to catch up (due to learning constraints on long sequences)
Concept Drift

**CW200**

- **Accuracy (%)**
- **Time gap (days)**

Graph showing the accuracy over time for different models: SDAE 2000tr, LSTM 2000tr, CNN 2000tr, CUMUL 2000tr, CUMUL 1000tr, CUMUL 200tr, and CUMUL 100tr.
Concept Drift

CW200

Accuracy (%)

Time gap (days)

Moment of training

SDAE 2000tr
LSTM 2000tr
CNN 2000tr
CUMUL 2000tr
CUMUL 1000tr
CUMUL 200tr
CUMUL 100tr
Concept Drift

SDAE, LSTM and CNN generalize better than the state-of-the-art.
Implications and Take-aways
Implications and Take-aways

- First thorough evaluation of DL for WF
  - Powerful and robust attack (accuracy: 96% for CW100, 94% for CW900)
  - Each DNN has its strengths and weaknesses
- Game-changer for the WF arms-race:
  - Automated feature learning (vs. the burden of manual feature engineering)
  - Harder to defend against (due to non-trivial interpretability of features)
- Data collection and model selection are crucial to the performance
  - Evaluated by collecting the largest dataset for WF
Thank you!

WEBSITE FINGERPRINTING THROUGH DEEP LEARNING


SDAE

Autoencoder

SDAE classifier

Hidden representation

Tor cells

Feature extraction

Classification
CNN classifier
LSTM

LSTM network

LSTM unit
Closed World vs Open World

Closed World

Open World
State-of-the-Art Attacks

› kNN (Wang et al., 2014)
  › Features
    › 3,000 (picked through heuristics)
    › total size, total time, number of packets, packet ordering, traffic bursts…
  › Classifier
    › k-Nearest Neighbors (k-NN)
  › Accuracy
    › 92% (100 websites)
State-of-the-Art Attacks

› k-Fingerprinting (Hayes et al, 2016)

› Features
  › 150 (selected from Wang’s through analysis of feature importance)

› Classifier
  › Random Forest + k-NN

› Accuracy
  › 93% (100 websites)
State-of-the-Art Attacks

- **CUMUL (Panchenko et al, 2016)**
  - **Features**
    - 100 (derived as interpolation points of the cumulative sum of packet lengths)
  - **Classifier**
    - Support Vector Machine (SVM)
  - **Accuracy**
    - From 97% (100 websites)
Open World: ROC Curve

Monitored: 200 websites
Non-monitored: 400,000 websites

- SDAE (AUC = 0.91)
- LSTM (AUC = 0.87)
- CNN (AUC = 0.92)
- CUMUL (AUC = 0.90)
Open World: ROC Curve

CNN and SDAE outperform state-of-the-art

Monitored: 200 websites
Non-monitored: 400,000 websites

- SDAE (AUC = 0.91)
- LSTM (AUC = 0.87)
- CNN (AUC = 0.92)
- CUMUL (AUC = 0.90)