RadioNet: Robust Deep-Learning Based Radio Fingerprinting

Haipeng Li¹, Kaustubh Gupta², Chenggang Wang¹, Nirnimesh Ghose², Boyang Wang¹

¹ University of Cincinnati, USA; ² University of Nebraska-Lincoln;

IEEE CNS 2022
Radio Fingerprinting

Authenticating wireless devices over Radio Frequency (RF) signals at the physical layer.
Why Feasible?

- Hardware imperfections (I/Q imbalance, phase noise, nonlinear distortion, etc.) lead to minor shifts in RF signals.

- Each transmitter has unique hardware imperfections.

Limitations in Radio Fingerprinting

- **Deep learning** achieves **high** accuracy in same-day
  - **Same-day**: train with Day 1, test with Day 1

- **Significant Performance drop in a cross-day scenario**
  - MobiHoc’19, Globecom’19, INFOCOM’20,
  - **Cross-day**: train with Day 1, test with Day 2

- Example: **20** transmitters (USRPs), CNN as classifier
  - Same-day accuracy: **99%**
  - Cross-day accuracy: **5%** (random guess)
Limitations in Radio Fingerprinting

- **Slicing windows** are often used to pre-process RF signals to inputs of neural networks

- The selection of parameters in pre-processing has not been rigorously discussed

I/Q Samples in One Transmission

<table>
<thead>
<tr>
<th>1</th>
<th>2</th>
<th>3</th>
<th>...</th>
<th>L</th>
<th>L+1</th>
<th>L+2</th>
<th>...</th>
<th>...</th>
<th>...</th>
<th>M</th>
</tr>
</thead>
</table>

Slicing window

Trace 1

| 1 | 2 | 3 | ... | L |

Slicing window

Trace 2

| 2 | 3 | ... | L | L+1 |

Slicing window

Trace 3

| 3 | ... | L | L+1 | L+2 |
Our Contributions

• Improve **robustness** of Radio Fingerprinting from 3 aspects:

1. Demonstrate that **parameters of pre-processing have significant impacts** to accuracy (from extremely high to random guess)

2. Improve cross-day accuracy with **adversarial domain adaptation**

3. Introduce **device rank** as a more robust metric
Adversary Domain Adaptation

- Given source data and target data, ADA minimizes the discrepancy between source & target in a feature space.
Our Method with ADA

- **Source:** Day 1; **Target:** Day 2
- Train ADA with a **large** amount of RF signals from Day 1 and a **small** amount of RF signals from Day 2
- Tune k-NN for better classification for Day 2
Testbed and Datasets

- **NEU dataset** (from INFOCOM’20):
  - 1 USRP as receiver, 20 USRP as transmitters
  - RF signals from 2 days

- **HackRF-10 dataset** (Ours)
  - 1 receiver, 10 transmitters
  - HackRF One, GNU Radio
  - WiFi, BPSK 1/2, Indoor
  - RF signals from 2 days
  - 3 transmissions per day
  - 30 secs per transmission
  - 3.26 million I/Q samples collected
• Collect I/Q samples before FFT and after Equalizer
• **Time** domain, **Frequency** domain, **Time-Frequency** Domain

• Parameters in Pre-Process: **Window Size L and Stride s**

I/Q Samples in One Transmission

| 1 | 2 | 3 | … | L | L+1 | L+2 | … | … | … | M |

Slicing window

Trace 1

| 1 | 2 | 3 | … | L |

Trace 2

| 2 | 3 | … | L | L+1 |

Trace 3

| 3 | … | L | L+1 | L+2 |

Slicing window with size L and stride 1
Evaluation Setting

- Two CNN: **Homegrown** (INFOCOM’19) and **DF** (CCS’18)
- Keras and Tensorflow (Nvidia Titan RTX)
- Training (64%), Validation (16%), Testing (20%)
Evaluation Metric

- **Accuracy**: Given N test traces, x traces are predicted correctly. $\text{Acc} = \frac{x}{N}$

- **Device Rank**: aggregate scores of transmitters over N traces, sort the aggregated scores, report the rank of the correct transmitter

- Why device rank is better than accuracy?
  - Hardware imperfections are difficult to learn
  - Aggregated scores are more robust
### The impact of stride $s$ on accuracy
(Time domain, window/trace length $L = 288$)

<table>
<thead>
<tr>
<th>Stride $s$</th>
<th>Neural Networks</th>
<th>NEU dataset (Random guess 5%)</th>
<th>HackRF dataset (Random guess 10%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Same-Day</td>
<td>Cross-Day</td>
</tr>
<tr>
<td>$s=1$</td>
<td>Homegrown</td>
<td>99.74</td>
<td>6.26</td>
</tr>
<tr>
<td></td>
<td>DF</td>
<td>99.95</td>
<td>6.08</td>
</tr>
<tr>
<td>$s=144$</td>
<td>Homegrown</td>
<td>26.47</td>
<td>7.59</td>
</tr>
<tr>
<td></td>
<td>DF</td>
<td>50.02</td>
<td>6.90</td>
</tr>
<tr>
<td>$s=L=288$</td>
<td>Homegrown</td>
<td>16.90</td>
<td>8.72</td>
</tr>
<tr>
<td></td>
<td>DF</td>
<td>14.24</td>
<td>7.31</td>
</tr>
<tr>
<td>$s=2L=596$</td>
<td>Homegrown</td>
<td>11.61</td>
<td>8.68</td>
</tr>
<tr>
<td></td>
<td>DF</td>
<td>5.88</td>
<td>5.70</td>
</tr>
</tbody>
</table>

**Should choose stride $s$ s.t. there is no overlaps across traces**
NEU: stride $s = L = 288$, accuracy is only 16.9%, device rank still converges to 1 (authenticate correctly) after 35 traces (around 37 milliseconds of RF transmissions)

Low accuracy does not necessarily mean failing to authenticate
Cross-day also affects device rank, but not for every transmitter.
### The impact of trace length $L$ on accuracy

**Time domain, stride $s = L$**

<table>
<thead>
<tr>
<th>Trace length $L$</th>
<th>Neural Networks</th>
<th>NEU dataset (Random guess 5%)</th>
<th>HackRF dataset (Random guess 10%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Same-Day</td>
<td>Cross-Day</td>
</tr>
<tr>
<td>$L=144$</td>
<td>Homegrown</td>
<td>22.94</td>
<td>7.13</td>
</tr>
<tr>
<td></td>
<td>DF</td>
<td>55.16</td>
<td>7.33</td>
</tr>
<tr>
<td>$L=288$</td>
<td>Homegrown</td>
<td>16.96</td>
<td>8.72</td>
</tr>
<tr>
<td></td>
<td>DF</td>
<td>14.24</td>
<td>7.31</td>
</tr>
<tr>
<td>$L=576$</td>
<td>Homegrown</td>
<td>13.29</td>
<td>9.26</td>
</tr>
<tr>
<td></td>
<td>DF</td>
<td>6.56</td>
<td>5.82</td>
</tr>
</tbody>
</table>

**A greater $L$ should be chosen whenever it is possible**
Comparison among different domains of I/Q data

<table>
<thead>
<tr>
<th>Domain</th>
<th>Neural Networks</th>
<th>HackRF dataset (Random guess 10%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Same-Day</td>
</tr>
<tr>
<td>Time</td>
<td>Homegrown</td>
<td>48.07</td>
</tr>
<tr>
<td></td>
<td>DF</td>
<td>54.96</td>
</tr>
<tr>
<td>Frequency</td>
<td>Homegrown</td>
<td>50.49</td>
</tr>
<tr>
<td></td>
<td>DF</td>
<td>59.71</td>
</tr>
<tr>
<td>Time-Frequency</td>
<td>Homegrown</td>
<td>50.51</td>
</tr>
<tr>
<td></td>
<td>DF</td>
<td><strong>60.21</strong></td>
</tr>
</tbody>
</table>

Frequency domain is more robust in cross-day
# Accuracy in the cross-day (HackRF-10 dataset)

<table>
<thead>
<tr>
<th>Domain</th>
<th>Method</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>0</td>
</tr>
<tr>
<td>Frequency</td>
<td>Fine-tuning</td>
<td>25.98</td>
</tr>
<tr>
<td></td>
<td>ADA</td>
<td>25.98</td>
</tr>
<tr>
<td>Time-Frequency</td>
<td>Fine-tuning</td>
<td>17.19</td>
</tr>
<tr>
<td></td>
<td>ADA</td>
<td>17.19</td>
</tr>
</tbody>
</table>

- Fine-tuning (WiSec’21) v.s. Our method based on ADA
- N: No. of traces per transmitter from Day 2
- Baseline (N = 0): train Day 1 test Day 2 directly

Both methods improve cross-day accuracy

Ours is better when N>=200
• Baseline (N = 0): train Day 1 test Day 2 directly

Device rank in cross-day is also improved by ADA
Discussions and Future Work

• Complex-value neural networks
  • IQ samples are complex values
  • Operations (e.g., max) are not defined over complex values
  • Transfer learning over complex values

• How frequent we need to tune a classifier?

• Can we tune without labeled data from Day 2?
Thank you! Q&A

This work is partially supported by NSF (CNS-1947913)

Code and datasets: https://github.com/UCdasec/RadioNet