Perceived Information Revisited: New Metrics to Evaluate Success Rate of Side-Channel Attacks

Akira Ito¹, Rei Ueno², Naofumi Homma²

¹ NTT Social Informatics Laboratory
² Tohoku University
**Background of this work**

- DL-SCA is one of the most powerful attacks.
  - Many studies on DL-SCA have been conducted recently.

- Training an NN model requires a performance metric.
  - Which one is best?

- Major metrics (e.g., CE loss, acc.) are not suitable for SCA.
  - Accuracy of 0% does not mean DL-SCA will fail.

  **However, computation cost of success rate (SR) is too high!**
Contributions

- Analysis of relation between cross entropy (CE) loss function and SR
  - Explain why CE loss is not suitable to measure the performance of DL-SCA.

- Effective CE/PI (ECE/EPI), new metrics for DL-SCA
  - ECE/EPI are more useful metrics than CE/PI for SCAs.
  - EPI can enable us to estimate (the upper-bound of) SR.
Relation between NLL and MI

- Negative log likelihood (NLL) is used as loss function.
  \[
  \text{NLL} = -\frac{1}{m} \sum_{i=1}^{m} \log q(Z_i | X_i; \theta)
  \]
  - NLL minimization is equivalent to maximum likelihood estimation.

- NLL can be regarded as approximation of CE.
  - If the number of traces \(m\) is sufficiently large, then
    \[
    \text{NLL} \approx -\mathbb{E} \log q(Z | X; \theta) = \text{CE}(q)
    \]

- Relation between mutual information (MI) and CE
  \[
  I(Z; X) \geq H(K) - \text{CE}(q) \approx H(K) - \text{NLL}
  \]

Perceived information (PI) \(J_q(Z; X) = H(K) - \text{CE}(q)\) denotes how much information NN can extract.
Relation between MI and SR

- de Chérencey et al. prove the following theorem.

Theorem (Relation between MI and SR)

\[ \xi(SR_m) \leq mI(Z; X) \]

How much entropy does attacker need to achieve \( SR_m \)?

Amount of information available with \( m \) traces

- Side-channel can be seen as communication channel.

A plot of \( \xi(SR) \)

if the bit-length of key is two bits

Success rate

Amount of transmittable information is \( I(Z; X) \).
Relation between MI and SR

- de Chéridey et al. prove the following theorem.

\[
\xi(SR_m) \leq mI(Z; X)
\]

**Theorem (Relation between MI and SR)**

- How much entropy does attacker need to achieve \(SR_m\)?
- Amount of information available with \(m\) traces

\(\xi(SR) = 0\) if \(SR = 0.25\)

→ We need no key information to achieve SR of 0.25.

Maximum amount of transmittable information is \(I(Z; X)\).
Relation between MI and SR

- de Chérissey et al. prove the following theorem.

Theorem (Relation between MI and SR)

\[ \xi(SR_m) \leq mI(Z; X) \]

How much entropy does attacker need to achieve \( SR_m \)?

Amount of information available with \( m \) traces

\( \xi(SR) = 2 \) if \( SR = 1 \)

→ We need all the key information (i.e., 2 bits) to achieve SR of 1.

Maximum amount of transmittable information is \( I(Z; X) \).
Relation between MI and SR

- de Chéridey et al. prove the following theorem.

**Theorem (Relation between MI and SR)**

\[ \xi(SR_m) \leq mI(Z; X) \]

How much entropy does attacker need to achieve SR\(_m\)?

- Side-channel can be seen as communication channel.

A plot of \( \xi(SR) \)

How much entropy does attacker need to achieve SR\(_m\)?

Amount of information available with \( m \) traces

Maximum amount of transmittable information is \( I(Z; X) \).
Extension for DL-SCAs

Intuitively, we expect the following inequality holds:

$$\xi(SR_m(q)) \leq mJ_q(Z; X) = m(H(K) - CE(q))$$

How much entropy does attacker need when using model \(q\) and \(m\) traces?

- If this holds, we can estimate SR by using PI (i.e., CE)
  - Masure et al. experimentally showed that this inequality would hold.

Unfortunately, this does not hold.

**Theorem (probability distribution conversion which retains SR)**

Let \(q\) be a model. Define a following conversion of \(q\) with an inverse temperature \(\beta > 0\):

$$q_\beta(z \mid x; \theta) = \frac{q(z \mid x; \theta)^\beta}{\sum_{z'} q(z' \mid x; \theta)^\beta}$$

For any \(\beta > 0\), the success rate using \(q\) is equivalent to that using \(q_\beta\).
Results of conversion using $\beta$

| $\beta$ | 0.1 | 1 | 10 |
|---------|-----|---|----|
| NLL     | 0.7933 | 0.7789 | 1.565 |
| $q_\beta$ | ![Graph](image1.png) | ![Graph](image2.png) | ![Graph](image3.png) |
| Attack result | ![Graph](image4.png) | ![Graph](image5.png) | ![Graph](image6.png) |

- NLL (CE) value and distribution shape change with $\beta$.
- But, SR/GE does not change with $\beta$.

- There is counterexample $q_\beta$ of following inequality:
  $$\xi(SR_m(q)) \leq m J_q(Z; X) = m(H(K) - CE(q))$$
Effective CE/PI (ECE/EPI)

- SR is invariant, but CE/PI varies with the value of $\beta$.
  - CE/PI are not appropriate metrics for DL-SCA.

- Proposed metrics: ECE and EPI (effective PI)
  \[
  CE^*(q) := \inf_{\beta \in (0, \infty)} CE(q_\beta),
  \]
  \[
  J^*_q(Z; X) := \sup_{\beta \in (0, \infty)} J_{q_\beta}(Z; X) = H(Z) - CE^*(q)
  \]
  - Basic idea: remove the uncertainty of CE/PI in terms of SR

- Conjecture following inequality holds using EPI.

\[
\xi(\text{SR}_m(q)) \leq m J^*_q(Z; X)
\]
DL-SCAs on masked software/hardware implementations

DL-SCAs on ASCAD dataset

- CE (NLL)
- ECE

Learning curve

DL-SCA on TI-based implementation

- CE (NLL)
- ECE

Learning curve

- Estimated value is proportional to actual one.
- Estimated value is proportional to actual one.
- ECE is less than CE since ECE is lower bound of CE.
Processing time of each method

Processing time per one epoch [s]

|       | Empirical SR evaluation | Proposed method | Ratio |
|-------|-------------------------|-----------------|-------|
| ASCAD | 14.1                    | 0.0378          | 373   |
| TI    | 145                     | 0.531           | 273   |

- SR is evaluated by bootstrapping.
  - Use 100 bootstrap samples to estimate SR value.

Proposed method is several hundreds faster than empirical evaluation.
Concluding remarks

- Analysis of relation between CE loss and SR
  - Conversion changes CE loss but not SR
  - CE/PI has uncertainty in terms of SR

- Effective CE/PI (ECE/EPI), new metrics for DL-SCA
  - Can easily estimate the attack performance (e.g., SR and GE)

- Future work
  - Formal proof of our conjecture (inequality of SR and EPI)
## Settings of experiments

|       | Training | Test    |
|-------|----------|---------|
| ASCAD | 50,000   | 10,000  |
| TI    | 4,000,000| 4,000,000|
Model comparison

- Compare four pretrained models for ASCAD dataset
  - MLP and CNN models proposed in original ASCAD paper
  - CNN models proposed by Zaid et al. and Wouters et al.

- Lack of bins means # of required traces is greater than 10,000.

Our metric accurately estimates model performances.
How to calculate ECE/EPI

$CE(q_\beta)$ has the following properties:

- $CE(q_\beta) \to n$ as $\beta \to 0$
- $CE(q_\beta) \to \infty$ as $\beta \to \infty$
- $CE(q_\beta)$ is a strictly convex function of $\beta$.

Certainly Newton method can find the minimum value of $CE(q_\beta)$.

- The local minimum of $CE(q_\beta)$ is its global minimum.

Example of $CE(q_\beta)$ when $n = 8$ bits
How to use NN for key recovery

- Negative log likelihood (NLL) is used as a score of each key

\[
NLL^{(k)} = -\frac{1}{m} \sum_{i=1}^{m} \log q(S(k \oplus T_i)|X_i; \theta)
\]

- NLL is inversely proportional to the product of probability.

- Attack Procedure:
  1. Calculate NLL for each key using \( m \) traces
  2. Get \( k \) whose the minimum NLL value among all candidates

\( k_1 \) is regarded as correct key.

\[ NLL^{(k_1)} < NLL^{(k_2)} \]
Inference using NNs

- NN is used to estimate intermediate value from a trace.
  - Image classification
    - DL-SCA
  - In profiling phase, NN trains plausible probability distribution.
  - In attack phase, trained NN estimates secret information.