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

Akira Ito[1], Rei Ueno[2], Naofumi Homma[2]

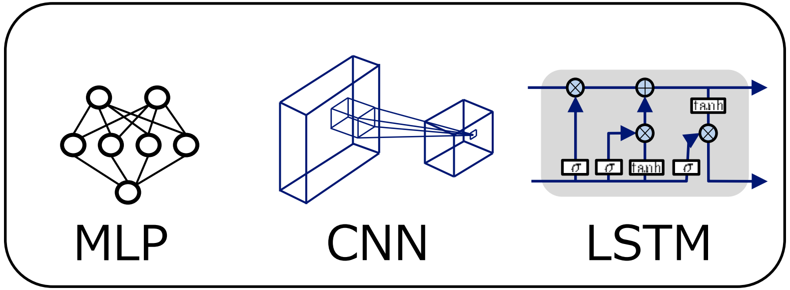
1 NTT Social Informatics Laboratory

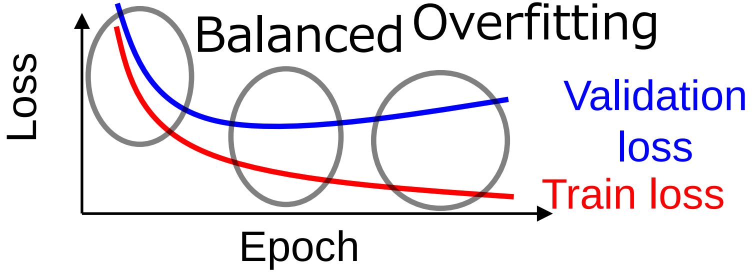
2 Tohoku University

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.

$$\mathrm{NLL} = -\frac{1}{m}\sum_{i=1}^{m} \log q(Z_i | \boldsymbol{X}_i; \theta)$$

  - 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

$$\mathrm{NLL} \approx -\mathbb{E} \log q(Z | X; \theta) = \mathrm{CE}(q)$$

- Relation between mutual information (MI) and CE

$$I(Z; \boldsymbol{X}) \geq H(K) - \mathrm{CE}(q) \approx H(K) - \mathrm{NLL}$$

Perceived information (PI) $J_q(Z; \boldsymbol{X}) = H(K) - \mathrm{CE}(q)$ denotes how much information NN can extract.

Relation between MI and SR

- de Chérisey et al. prove the following theorem.

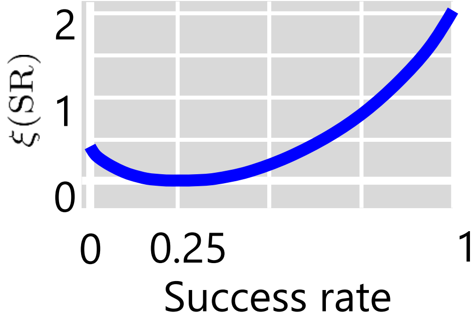
Theorem (Relation between MI and SR)

$$\underbrace{\xi(\mathrm{SR}_m)} \leq \underbrace{mI(Z;\boldsymbol{X})}$$

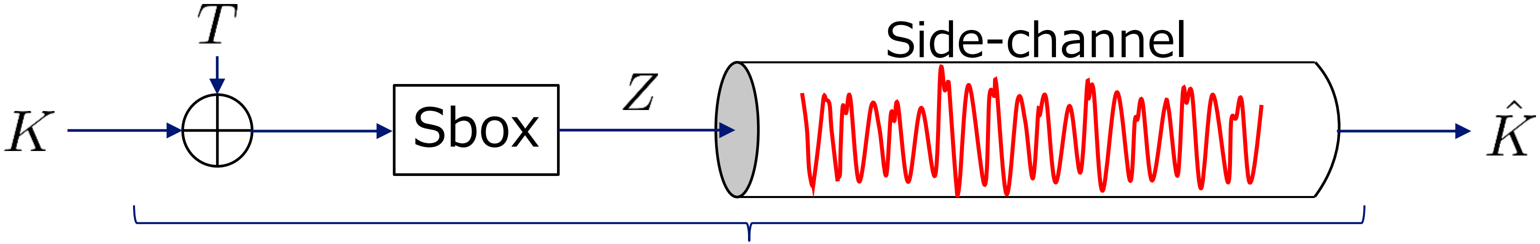
How much entropy does attacker need to achieve $\mathrm{SR_m}$?

Amount of information available with m traces

A plot of $\xi(\mathrm{SR})$ if the bit-length of key is two bits



- Side-channel can be seen as communication channel.



Maximum amount of transmittable information is $I(Z;\boldsymbol{X})$.

Relation between MI and SR

■ de Chérisey et al. prove the following theorem.

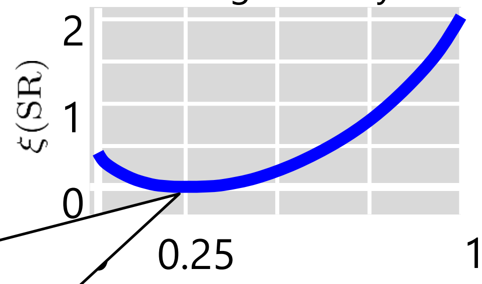
Theorem (Relation between MI and SR)

$$\xi(\mathrm{SR}_m) \leq mI(Z;\boldsymbol{X})$$

How much entropy does attacker need to achieve SR_m?

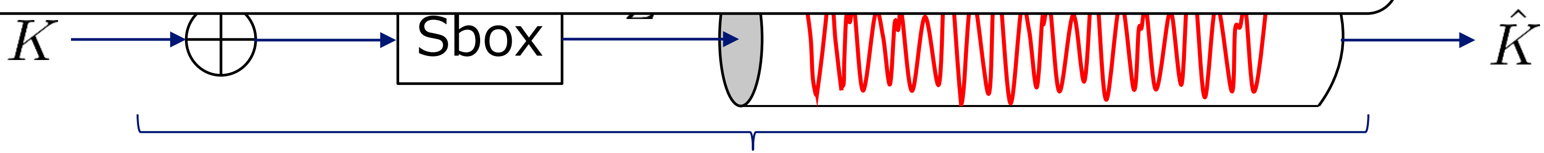
Amount of information available with *m* traces

A plot of $\xi(\mathrm{SR})$ if the bit-length of key is two bits



$\xi(\mathrm{SR}) = 0$ if SR = 0.25

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



Maximum amount of transmittable information is $I(Z;\boldsymbol{X})$.

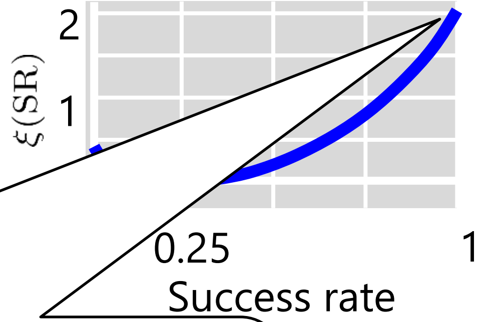
Relation between MI and SR

■ de Chérisey et al. prove the following theorem.

Theorem (Relation between MI and SR)

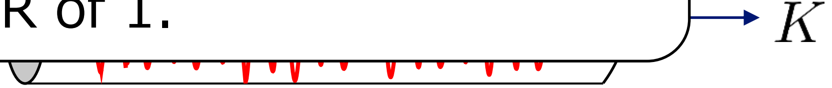
$$\underbrace{\xi(\mathrm{SR}_m)}_{\text{How much entropy does attacker need to achieve } \mathrm{SR_m}\text{?}} \leq \underbrace{mI(Z;\boldsymbol{X})}_{\text{Amount of information available with } m \text{ traces}}$$

A plot of $\xi(\mathrm{SR})$ if the bit-length of key is two bits



$\xi(\mathrm{SR}) = 2$ if SR = 1

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



Maximum amount of transmittable information is $I(Z;\boldsymbol{X})$.

Relation between MI and SR

- de Chérisey et al. prove the following theorem.

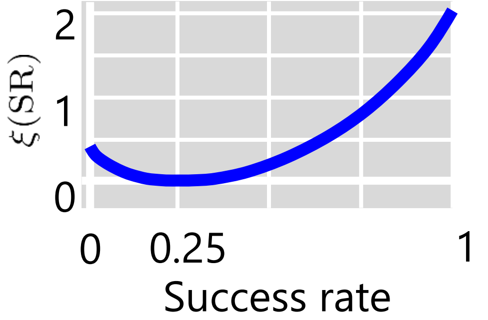
Theorem (Relation between MI and SR)

$$\underbrace{\xi(\mathrm{SR}_m)} \leq \underbrace{mI(Z;\boldsymbol{X})}$$

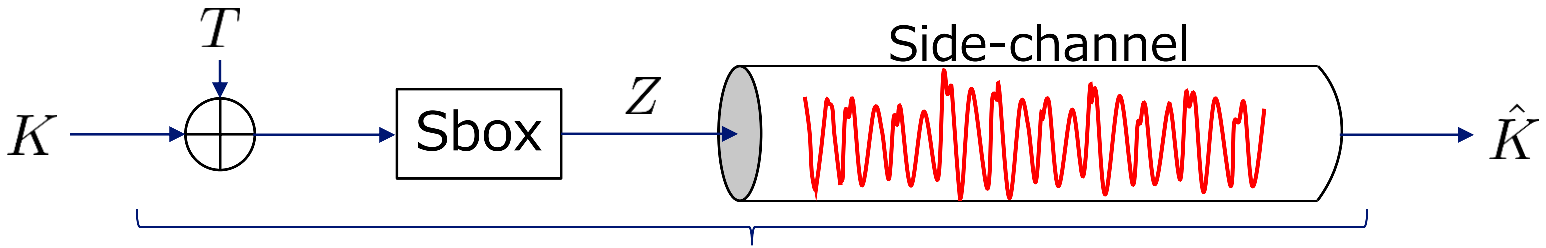
How much entropy does attacker need to achieve $\mathrm{SR_m}$?

Amount of information available with *m* traces

A plot of $\xi(\mathrm{SR})$ if the bit-length of key is two bits



- Side-channel can be seen as communication channel.



Maximum amount of transmittable information is $I(Z;\boldsymbol{X})$.

Extension for DL-SCAs

NTT

- Intuitively, we expect the following inequality holds:

$$\underline{\xi(\mathrm{SR}_m(q))} \leq \underline{mJ_q(Z;\boldsymbol{X})} = m(H(K) - \mathrm{CE}(q))$$

How much entropy does attacker need when using **model *q* and *m* traces**?

Amount of information **model *q*** can extract with *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.

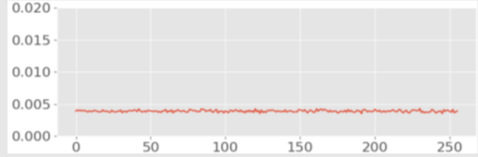
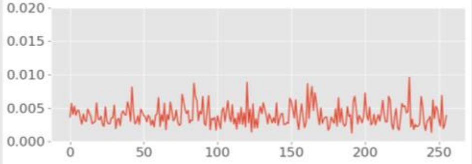
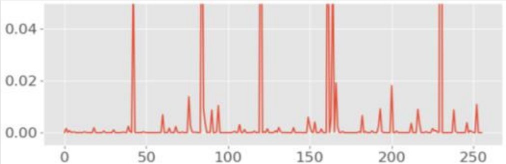
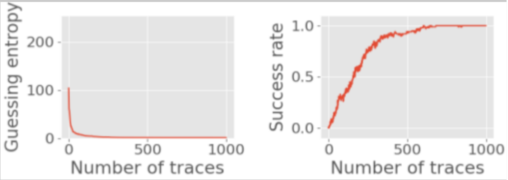
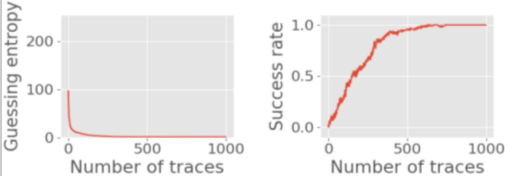
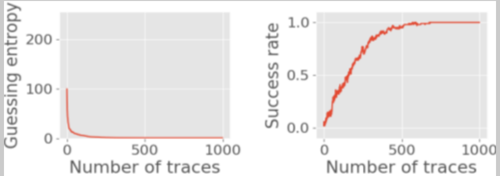
Theorem (probability distribution conversion which retains SR)

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

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

For any $\boldsymbol{\beta > 0}$, the success rate using $\boldsymbol{q}$ is equivalent to that using $\boldsymbol{q_\beta}$.

Results of conversion using β

β	0.1	1	10
NLL	0.7933	0.7789	1.565
q_β			
Attack result			

- NLL (CE) value and distribution shape change with β.
- But, SR/GE does not change with β.
  - There is counterexample q_β of following inequality:

$$\xi(\mathrm{SR}_m(q)) \le mJ_q(Z;\boldsymbol{X}) = m(H(K) - \mathrm{CE}(q))$$

Effective CE/PI (ECE/EPI)

- SR is invariant, but CE/PI varies with the value of β.
  - CE/PI are not appropriate metrics for DL-SCA.
- Proposed metrics: ECE and EPI (effective PI)

$$\mathrm{CE}^*(q) := \inf_{\beta \in (0,\infty)} \mathrm{CE}(q_\beta),$$

$$J_q^*(Z;\boldsymbol{X}) := \sup_{\beta \in (0,\infty)} J_{q_\beta}(Z;\boldsymbol{X}) = H(Z) - \mathrm{CE}^*(q)$$

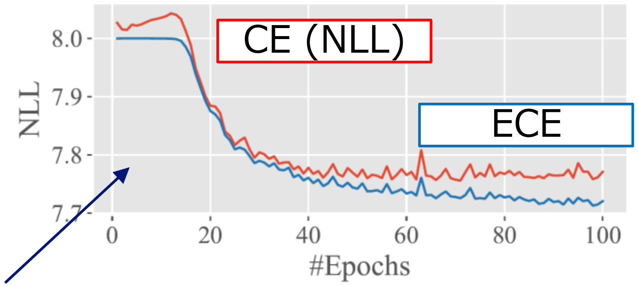
  - Basic idea: remove the uncertainty of CE/PI in terms of SR
- Conject following inequality holds using EPI.

Conjecture

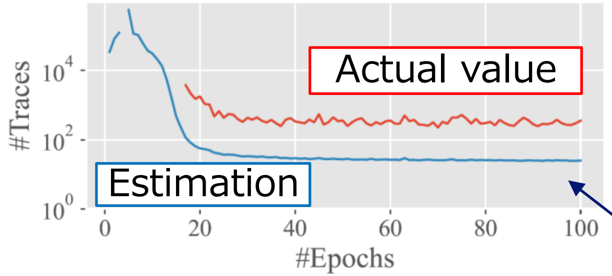
$$\xi(\mathrm{SR}_m(q)) \le m J_q^*(Z;\boldsymbol{X})$$

DL-SCAs on masked software/hardware implementations

DL-SCA on ASCAD dataset



Learning curve

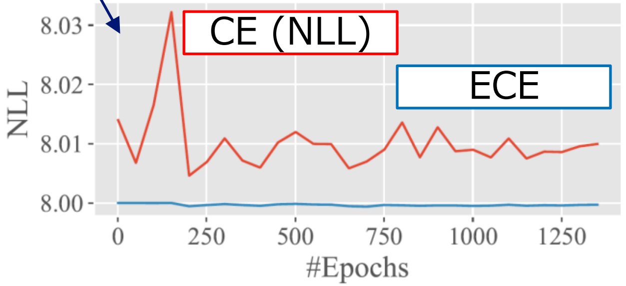


# of traces for successful attack

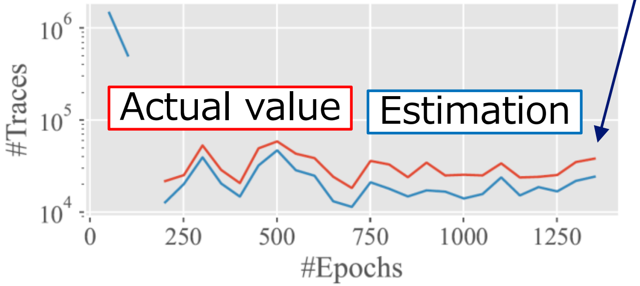
ECE is less than CE since ECE is lower bound of CE.

Estimated value is proportional to actual one.

DL-SCA on TI-based implementation



Learning curve



# of traces for successful attack

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

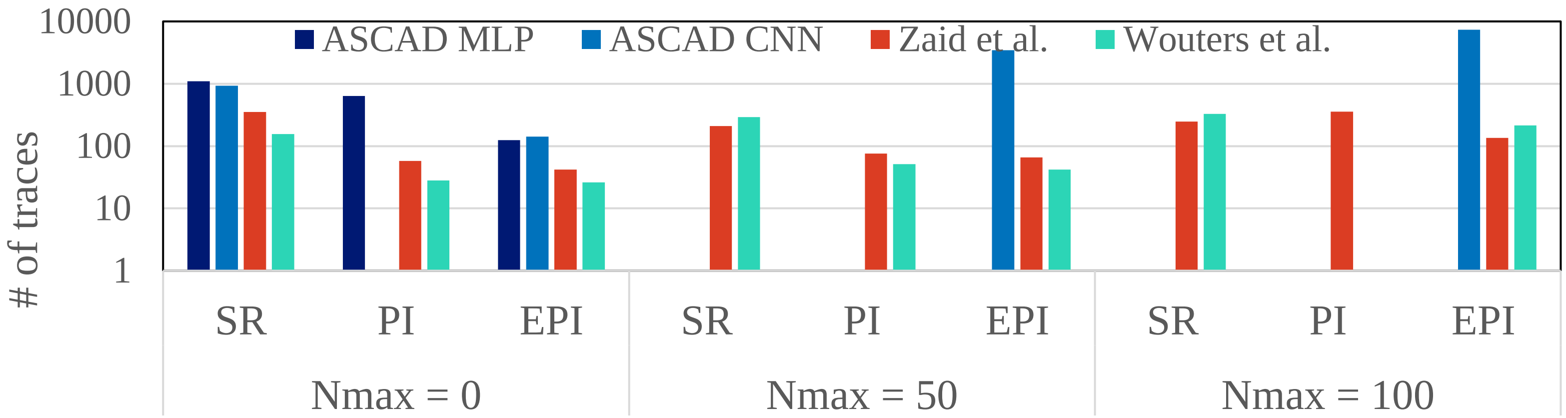
- Analysis of relation between CE loss and SR
  - Conversion changes CE loss but not SR
  - CE/PI has uncertainty in terms of SR
- Effective CE/PI (ECE/EPI), new metrics for DL-SCA
  - Can easily estimate the attack performance (e.g., SR and GE)
- Future work
  - Formal proof of our conjecture (inequality 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.



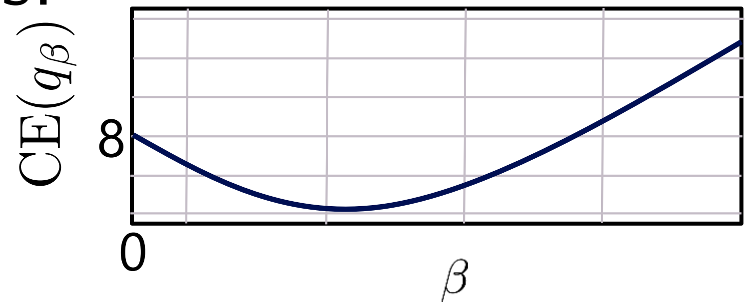
Nmax is amount of desynchronization.

Our metric accurately estimates model performances.

How to calculate ECE/EPI

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

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



Example of $\mathrm{CE}(q_\beta)$ when $n = 8$ bits

■ Newton method can find the minimum value of $\mathrm{CE}(q_\beta)$.

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

How to use NN for key recovery

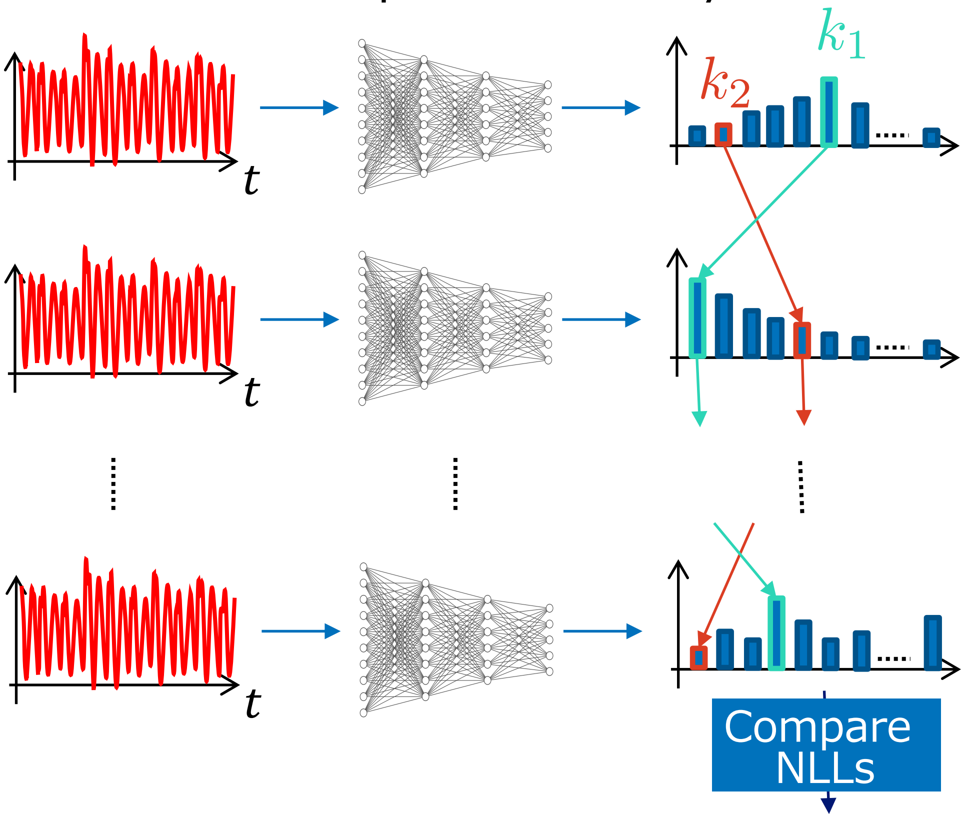
- Negative log likelihood (NLL) is used as a score of each key

$$\mathrm{NLL}^{(k)} = -\frac{1}{m}\sum_{i=1}^{m}\log q(S(k\oplus T_i)|\boldsymbol{X}_i;\theta)$$

  - 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

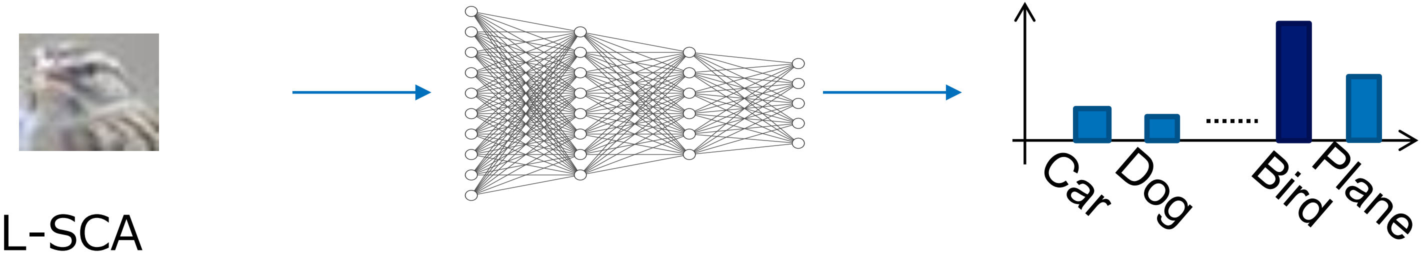
Attack example when #keys = 2



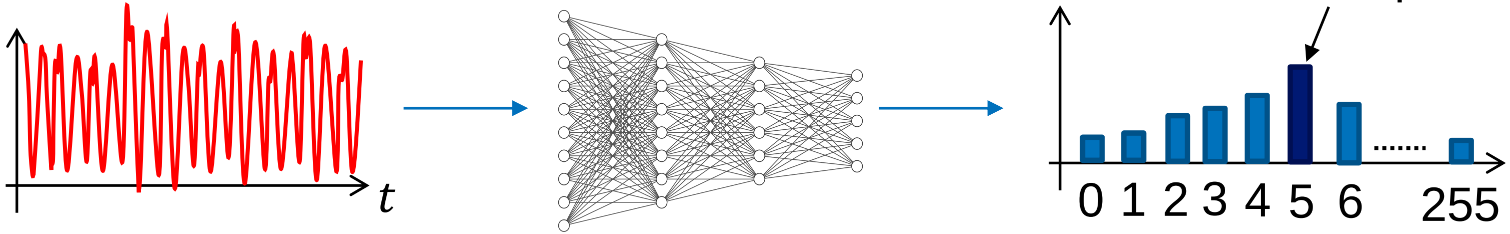
k_1 is regarded as correct key. ← $\mathrm{NLL}^{(k_1)} < \mathrm{NLL}^{(k_2)}$

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.