AGE Is Not Just a Number: Label Distribution in Deep Learning-based Side-channel Analysis

Lichao Wu, Léo Weissbart, Marina Krček, Huimin Li, Guilherme Perin, Lejla Batina, and Stjepan Picek

***Abstract*—The efficiency of the profiling side-channel analysis can be improved significantly with machine learning techniques. Although powerful, a fundamental machine learning limitation of being data hungry received little attention in the side-channel community. In practice, the maximum number of leakage traces that evaluators/attackers can obtain is constrained by the scheme requirements or the limited accessibility of the target. Even worse, various countermeasures in modern devices increase the conditions on the profiling size to break the target.**

This work demonstrates a practical approach to dealing with the lack of profiling traces. Instead of learning from a one-hot encoded label, transferring the labels to their distribution can significantly speed up the convergence of guessing entropy. Besides, by studying the relationship between all possible key candidates, we propose a new metric, denoted augmented guessing entropy (AGE), to evaluate the generalization ability of the profiling model. We validate AGE with two common use cases: early stopping and network architecture search, and the results indicate its superior performance.

***Index Terms*—Side-channel Analysis, Profiling Analysis, Deep Learning, Label Distribution, Profiling Model Fitting.**

I. Introduction

SIDE-CHANNEL ANALYSIS (SCA) is recognized as one of the most powerful attack methods on the implementations of cryptographic algorithms. Commonly, such attacks are divided into direct attacks like Simple Power Analysis (SPA) and Differential Power Analysis (DPA) [KJJ99], and two-stage (profiling) attacks like template attack [CRR02], stochastic models [SLP05], and machine learning-based attacks [LPB+15], [MPP16], [PHJ+17]. The profiling attacks impose additional requirements as they assume an 'open' device (or a copy of it), but the actual key recovery might need only a few measurements or, in some cases, a single trace [KPH+19], [PWP22].

In recent years, machine learning-based attacks positioned themselves as a strong alternative for more 'classical' SCA [CDP17], [KPH+19], [ZBHV19], which has become a standard evaluation approach for security evaluation and certification. The success of such methods relies on a sufficient number of training traces so that a machine learning classifier can accurately map the relationship between the traces and corresponding labels (intermediate data). In the worst-case scenario, an attacker can obtain unlimited training traces from the clone device for profiling attacks. However, in practice, there is a strong demand for developing a technique that effectively decreases the required number of profiling traces while keeping a similar attack performance:

- The time constraint for an evaluation dramatically limits the number of traces one can obtain. For instance, measuring one million profiling traces for a software RSA implementation with a 128-bit key could take more than a week [KS20]. Additionally, in post-analysis tasks such as trace realignment, noise filtering, and leakage assessment, an evaluator may not have enough budget to measure sufficient traces to break the target. Therefore, reducing the required number of profiling traces would be beneficial in saving time and enhancing the evaluator's attack capability.
- Unlike most deep learning applications, the SCA training data are most likely being 'protected' - the SCA countermeasures represent a standard/default setting for modern smart card/SOC's crypto-related implementations. These protection mechanisms further increase the difficulties in learning the trace-label relationship, thus increasing the demand for the number of measurements. From a security developers' point of view, an increasing number of side-channel measurements to break the target implementation means higher security assurance of their product. If we can effectively reduce the required number of profiling traces, then such vulnerabilities will again be considered.
- For a black/grey box evaluation, the available traces can drop to hundreds or thousands due to the upper limit of program counters such as Application Transaction Counter (ATC) or PIN Try Counter (PTC) [Car11], which is commonly insufficient when implementing an efficient profiling model. Building a profiling model with limited profiling traces would significantly increase the capability of exploiting the potential vulnerability.

There are limited evaluation metrics optimized for SCA. Evaluation metrics are essential in the training process: by actively monitoring the metric value, one can easily interpret the learning process, e.g., underfitting or overfitting. However, accuracy, a commonly used metric for deep learning, is less indicative for SCA in multi-trace attack scenarios [KPH+19]. The reasons can be explained from two aspects. First, due to the noise/countermeasures in the traces, side-channel leakages are more difficult to classify. Second, accuracy does not represent the success of an attack well, as we commonly need to consider continuous attacks that are better evaluated with metrics capturing this continuity. Guessing entropy and

L. Wu, M. Krček, and H. Li are affiliated with Delft University of Technology, The Netherlands.

L. Weissbart, G. Perin, and S. Picek are affiliated with the Delft University of Technology, The Netherlands, and Digital Security Group, Radboud University, The Netherlands.

L. Batina is affiliated with Digital Security Group, Radboud University, The Netherlands.

success rate are the commonly used metrics for SCA. Unfortunately, using such evaluation metrics would significantly increase the training time due to their computation complexity. Moreover, guessing entropy evaluates the rank of the correct key $only$. Although effective, we argue that it can be less indicative as the internal relationships with other (wrong) key candidates are not considered. More discussions are available in Section V-B.

We put the above concerns forward as the motivations for this work. First, to reduce the required number of training traces, we transfer the one-hot encoded labels to their Gaussian distribution centering on the corresponding labels motivated by [Gen16]. An illustration of the proposed learning scheme is shown in Figure 1. A one-hot encoded label that belongs to class 4 has been transferred to the distributed label with the value of the fourth index with the highest probability. Based on our experiment, regardless of the used leakage model and deep learning architecture, if using our learning scheme, the profiling traces can be reduced more than five times compared with the number of profiling traces used in the literature.

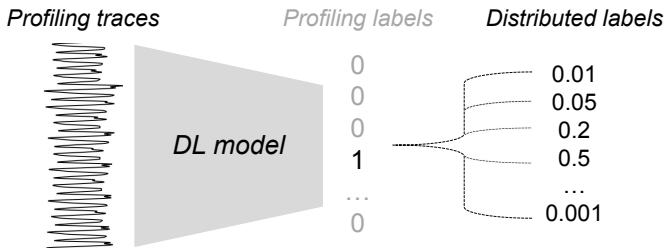


Fig. 1: Learning with distributed labels.

One essential assumption of the distributed label is that the label closer to the correct label has a higher probability of being selected. Under the same assumption, we propose key distribution to measure the geometry distance between the most likely key (not necessarily the correct key) and all the other keys. From this method and guessing entropy estimation, we propose a novel profiling model fitting metric - Augmented Guessing Entropy (AGE) that calculates the correlation between key distribution and the key guessing vector of all key guesses. As demonstrated with experiments on publicly available datasets, the proposed metric can indicate the generalization ability of a profiling model and thus serve as a reliable evaluation metric of early stopping and network architecture search. AGE is more indicative than conventional metrics, such as validation cross-entropy loss, because it directly links with the attack performance. On the other hand, compared with GE, AGE requires significantly fewer computation efforts to obtain a reliable estimation of the results. Thus, it can be a good metric during model training.

To summarize, the main contributions of this paper are:

1) We introduce a new effective training scheme for profiling SCA when the number of profiling traces is limited. The attack performance is improved by learning from the distributed labels compared to conventional one-hot encoded labels.
2) We propose a novel method to calculate the distance between the target key and other keys called key distribution (KD).
3) Based on the guessing entropy, we introduce a new metric called augmented guessing entropy (AGE) that can effectively estimate how well the profiling model fits the data. To that end, we show that the proposed metric is reliable in reflecting the generality of the profiling model. We demonstrate two use cases for potential implementors: early stopping and network architecture search in various testing conditions. The results show that the AGE metric performs better than other considered metrics.

We provide comprehensive experimental results on publicly available datasets to validate our claims. We also consider two commonly used deep learning techniques in SCA: multilayer perceptron and convolutional neural networks. The source code is available in the GitHub: https://github.com/AISyLab/Label-distribution-and-AGE.

The paper is organized as follows. Section II provides information about profiling SCA, commonly used evaluation metrics, and datasets used in this paper. The related works are discussed in Section III, followed by the proposal of the label distribution learning and novel SCA evaluation metric AGE in Sections IV and V. Section VI validates the proposed methods experimentally with different datasets, attack models, and leakage models. Finally, Section VII concludes this paper and proposes possible future research directions.

II. Background

A. Notation

We use calligraphic letters like $\mathcal{X}$ to denote sets and the corresponding upper-case letters X to denote random variables and random vectors $\mathbf{X}$ over $\mathcal{X}$. The corresponding lower-case letters x and $\mathbf{x}$ denote realizations of X and $\mathbf{X}$, respectively. We use sans serif font for functions (e.g., f).

k represents a key byte candidate that takes its value from the keyspace $\mathcal{K}$. k^* is the correct key byte, and k^{ref} is the key byte assumed by an attacker to be correct (as the attacker does not know the correct key). [1]

A dataset $\mathbf{T}$ is defined as a collection of traces $\mathbf{t}_i$, where each trace $\mathbf{t}_i$ is associated with a key-related label (or the key itself) y_i. A complete set of labels with c classes is denoted by $\mathcal{Y} = \{y_1, y_2, \cdots, y_c\}$. The number of profiling traces equals N, the number of validation traces equals V, and the number of attack traces equals Q. Finally, $\boldsymbol{\theta}$ denotes the vector of parameters to be learned in a profiling model.

B. Profiling Side-channel Analysis

As demonstrated in Figure 2, profiling side-channel attacks consist of two phases:

1) **Learning or profiling phase.** The profiling phase consists of building a profiling model $\mathsf{f}_M^{\boldsymbol{\theta}}$ to map the inputs (side-channel measurements) to the outputs (classes as obtained by evaluating the leakage model on the sensitive operation) on a set of N profiling traces. $\mathsf{f}_M^{\boldsymbol{\theta}}$ represents

[1] Note that the subkey candidates can have any number of bits that are being guessed and while here we assume the AES cipher scenario, the concept is algorithm-independent.

the profiling model trained for a given leakage model M and the set of learning parameters $\boldsymbol{\theta}$. This phase aims to fit the parameters of a function that maps the side-channel traces to the labels in the best way (minimizing the error function). It is common to use the validation set of size V to know when to stop the learning process.

2) **Test or attack phase.** The attack phase consists of obtaining label predictions for the traces from a different dataset of size Q to test the model. The trained model processes each attack trace and produces the attack's output as a vector of probabilities $\mathbf{p}_{i,j} \in \mathcal{K}$, where each index is the probability that a trace t_i is associated with the leakage value j. We can estimate the attack performance from this matrix of probabilities (as we have multiple vectors - one for each attack trace).

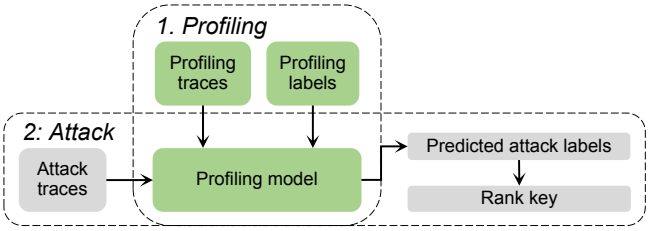


Fig. 2: Profiling side-channel analysis.

We consider two common profiling approaches:

- **Template Attack.** Template attack (TA) uses Bayes' theorem to obtain predictions, dealing with multivariate probability distributions as the leakage over consecutive time samples is not independent [CRR02]. In the state-of-the-art, template attack relies mostly on a normal (Gaussian) distribution.
- **Deep Learning-based SCA.** We consider supervised machine learning and the classification task as the side-channel attack's goal. Supervised learning deals with the task of learning a mapping f from a set of input variables from $\mathcal{X}$ to the set of output variables Y ($\mathsf{f}_M^{\boldsymbol{\theta}} : \mathcal{X} \rightarrow \mathcal{Y}$). For SCA, the profiling phase aims to learn the parameters $\boldsymbol{\theta}$, minimizing the empirical risk represented by a loss function on a dataset. In the attack phase, the goal is to predict the classes (more precisely, the probabilities that a certain class would be predicted) based on the previously unseen set of traces and the trained model $\mathsf{f}_M^{\boldsymbol{\theta}}$.

C. Evaluating the Attack Performance

An attack's output is the logarithmic sum of all Q probability vectors of single model predictions, where each index is associated with one key hypothesis. Sorting this vector by decreasing probabilities leads to a key guessing vector with increasing confidence predicted by a profiling model. The key rank denotes the position of the correct key. Then, one can use metrics such as guessing entropy to estimate the attacker's performance [SMY09].

Definition 1. Key guessing vector. The key guessing vector $\mathbf{g}$ is the vector of probabilities for all key candidates from the output of the profiling model's predictions:

$$\mathbf{g} = \mathsf{sort}\left(\sum_{i}^{Q} log\ \mathbf{P_r}(\mathbf{t}_i; \mathsf{f}_M^{\boldsymbol{\theta}})\right), \tag{1}$$

where $\mathbf{P_r}(\mathbf{t}_i; \mathsf{f}_M^{\boldsymbol{\theta}})$ is the prediction vector from the profiling model $\mathsf{f}_M^{\boldsymbol{\theta}}$ on a trace $\mathbf{t}_i$. sort is the function sorting array elements in order of decreasing values of their probabilities. Since the labels are key-related, the cumulative probabilities of labels can be easily mapped to their corresponding keys. From $\mathbf{g}$, the index of g represents the likelihood of the corresponding key candidate being the correct key candidate. g_0 and $g_{|\mathcal{K}|-1}$ are the first (best) and last (worst) element of $\mathbf{g}$, respectively.

Definition 2. Key rank. In a known-key setting, the key rank is the number of (most likely) keys an attacker needs to brute force until recovering the correct key. Among various key enumeration techniques [PSG16], one of the more popular methods is to try every key given its probability after generating a key guessing vector. In this scenario, the key rank is the position of the correct key in the guessing vector.

Definition 3. Guessing entropy. The guessing entropy [2] represents the averaged rank of the correct key k^* in the key guessing vector $\mathbf{g}$:

$$GE = \mathbf{E}\left(\mathsf{rank}_{k^*}(\mathbf{g})\right), \tag{2}$$

where $\mathsf{rank}_k(\mathbf{g}) \in \{0, \ldots, |\mathcal{K}|-1\}$. $\mathbf{E}$ is the average of multiple realizations of key rank, which is commonly performed by attacking with a profiling model multiple times with randomly selected attack traces.

D. Datasets

1) ASCAD Dataset: The ASCAD dataset is generated by taking measurements from an ATMega8515 running a masked AES-128 implementation and is proposed as a benchmark dataset for SCA [BPS+20]. Side-channel traces represent the AES encryption, where the commonly attacked trace interval represents the processing of the third byte in the S-box operation (S-box is fixed and publicly known for AES) taking place in the first round (the third byte is the first masked one). The operation is masked, and we assume no knowledge about masks in the profiling phase. There are two versions of this dataset:

1) ASCAD_f: The first version consists of 50 000 profiling traces and 10 000 attack traces, where each trace consists of 700 features (pre-selected window around the leaking spot). The profiling and attacking sets use the same fixed key, and we denote this dataset as ASCAD_f.
2) ASCAD_r: The second version of the ASCAD dataset contains 200 000 traces for profiling with random keys and random plaintexts and 100 000 for the attack phase, with a fixed key and random plaintexts. A window of

[2] As we attack only a single key byte, the proper term is partial guessing entropy. Nevertheless, we use the two terms interchangeably.

1 400 points of interest is extracted around the leaking spot. We denote this dataset as ASCAD_r.

For both datasets, different numbers of profiling and attack traces are used in our experiments (see Section VI for details), and 5 000 traces are used for validation and attack. The datasets are provided at https://github.com/ANSSI-FR/ASCAD/tree/master/ATMEGA_AES_v1.

2) CHES CTF Dataset: This dataset refers to the CHES Capture-the-flag (CTF) AES-128 measurements released in 2018 for the Conference on Cryptographic Hardware and Embedded Systems (CHES). The traces consist of masked AES-128 encryption running on a 32-bit STM microcontroller. In our experiments, we consider 45 000 traces for the training set, which contains a **fixed key**. The validation and attack sets consist of 5 000 traces. Each trace consists of 2 200 features. This dataset is available at https://chesctf.riscure.com/2018/news.

E. Leakage Models

The leakage model simulates the hypothetical power consumption to process one byte (as we attack the AES cipher that is byte-oriented). Our work considers two commonly used leakage models: the Hamming Weight (HW) and Identity (ID). For the HW leakage model, the attacker assumes the leakage is proportional to the sensitive variable's Hamming weight. This leakage model results in nine classes for a single intermediate byte for the AES cipher. In terms of the ID leakage model, an attacker considers the leakage in the form of an intermediate value of the cipher. This leakage model results in 256 classes for a single intermediate byte for the AES cipher.

III. Related Works

In Chari *et al.*'s seminal work, the authors proposed the template attack (TA) and showed that it could break implementations secure against other forms of side-channel attacks [CRR02]. This attack is the most powerful one from the information-theoretic point of view, but to reach its full potential, it requires an unbounded number of traces and noise following the Gaussian distribution [LPB+15]. Template attack is interesting as it is a generative technique, which means it will commonly overfit less as it allows the user to provide more information in the form of class conditionals.

While machine learning techniques have been widely used for several decades, the SCA community showed interest in such techniques only around a decade ago. In the beginning, the most interest sparked techniques like random forest [LMBM13], support vector machines [HGM+11], [HZ12], [PHJ+17], and multilayer perceptron [GHO15] (commonly in the context of shallow learning as it had only a single hidden layer).

The rapid development of deep learning-based SCAs started in 2016 when Maghrebi *et al.* demonstrated the strong performance of several neural network types, most notably, convolutional neural networks [MPP16]. Kim *et al.* investigated how adding noise to the input (thus, serving as regularization) improves the performance of profiling SCA attacks [KPH+19]. In [BPS+20], an empirical evaluation for different hyperparameters is conducted for CNNs on the ASCAD database. In [ZBHV19], the authors proposed a methodology to select hyperparameters related to the size (number of learnable parameters) of layers in CNNs. The methodology includes observations for the number of filters, kernel sizes, strides, and neurons in fully connected layers. Wouters *et al.* showed how to reach similar attack performance with data regularization and even smaller neural network architectures [WAGP20]. Perin *et al.* investigated deep learning model generalization and demonstrated how ensembles of random models could perform better than a single carefully tuned neural network model [PCP20]. Wu *et al.* and Rijsdijk *et al.* explored different automatic hyperparameter tuning strategies, namely Bayesian optimization [WPP20] and reinforcement learning [RWPP21] paradigms to find neural networks that perform well. While their approach requires a significant tuning effort (computational time), the authors improved state-of-the-art results. These works showed that deep learning models' good performance relies on an efficient selection of hyperparameters for specific datasets. If those hyperparameters are not selected properly, the attack will fail (or at least not work as well as possible).

It is intuitive that the number of measurements and input features also limits the performance of a profiling attack. Deep neural networks provide top-level performances in many domains when the amount of training data is sufficiently large. However, they could also provide excellent performance when the training data is reduced. In the context of profiling side-channel attacks, Cagli *et al.* investigated how to create measurements that improve the attack performance synthetically [CDP17]. Differing from the previous work where the authors developed a specialized data augmentation technique, Picek *et al.* showed that generic data augmentation techniques help in profiling SCA also [PHJ+18]. Another work investigated whether limiting the number of traces can be beneficial both from the experimental setup and performance sides [PHPG22].

From the input features perspective, Lu *et al.* investigated the performance of deep learning with raw traces (while the previous works considered pre-selected windows of features) [LZC+21]. As a result, better attack performance is achieved but with significantly more complex neural networks (e.g., having around 50 layers). Perin *et al.* showed how simple re-sampling of raw traces could result in extremely powerful attacks (requiring only a single attack trace) while using simple neural networks with only a few hidden layers [PWP22]. Finally, similarity learning was applied to pre-process the leakage traces and extract high-level features, leading to state-of-the-art attack performance with significantly reduced computation effort [WPP22a].

Commonly, in machine learning, one estimates the behavior of a profiling model based on statistics of individual observations like accuracy, loss, or recall. Unfortunately, such metrics can be misleading in SCA, as one considers cumulative predictions. Picek *et al.* showed that standard machine learning metrics could suggest radically different performance than the SCA metrics [PHJ+18]. Masure *et al.* connected the perceived information and negative log-likelihood, showing there can be common ground when using machine learning metrics in

SCA [MDP19]. Finally, Perin *et al.* discussed how mutual information could be a good metric to indicate when to stop the machine learning training process [PBP21].

IV. Label Distribution

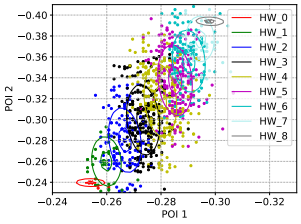
By asking *'how much does each label describe the instance?'*, Geng *et al.* first proposed Label Distribution Learning (LDL) by assigning a *description degree* to each possible label, leading to enhanced performance compared with hard (one-hot encoded) labels [Gen16]. This method has been used in tasks such as age estimation [GWX14] or personality recognition [XHG+17]. However, the application of LDL is restricted since one should have a reasonable estimation of the relation between labels, and such an estimation could be challenging in many tasks, e.g., image classification. Fortunately, SCA uses the leakage model to construct labels, which inherently leads to a clear relationship between labels. Indeed, two leakage traces with closer label (intermediate value) distance could be more similar. As a result, a combination of LDL and SCA could enhance attack performance.

Definition 4. Description degree. The description degree $d_x^{y_i}$ represents the degree of a label y_i to describe an input x. From the machine learning perspective, $d_x^{y_i}$ can be considered as the probability of the label y_i being selected. If a complete set of labels $\mathcal{Y}$ can fully describe the given input, then:

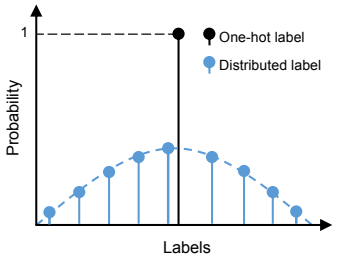
$$\sum_i d_x^{y_i} = 1, y_i \in \mathcal{Y}. \tag{3}$$

The conventional DL-based SCA represents a multi-class classification task aiming to describe a measurement with a unique cluster/label. To train a deep learning model, the label is one-hot encoded (see Figure 3b) using binary variables. In an ideal case, the label y_i perfectly represents the leaking features within a measurement (i.e., the correlation between labels and leaking features are one). However, the presence of noise/countermeasure increases the description degree of other labels to the corresponding leakage traces. For illustration, Figure 3a shows the Probability Density Function (PDF), and point-of-interests distributions (POI1 and POI2) from 1 000 measurements [3]. The color of each point is attributed based on its cluster label. Using the HW leakage model, nine PDFs representing nine HW clusters are built during the profiling phase. Each PDF is represented by two ellipses representing 0.5 (low) and 0.9 (high) of the maximum probabilities.

From Figure 3a, each PDF can be clearly separated. However, the overlap between each PDF cannot be ignored. For the traces that are in the middle between two PDFs, although they have deterministic (single) labels that represent the targeted intermediate data, leakage-wise, they are also geometrically close(r) to their neighboring clusters. Indeed, one can observe a natural measure of description degree that associates the labels with the traces. A precise description of these traces



(a) PDFs and POIs distribution for the correct key.



(b) Comparison between one-hot and distributed labels.

Fig. 3: PDFs and a demonstration of distributed labels.

should involve the 'incorrect' labels. Since their similarity to each cluster is inversely correlated with their label distance, as demonstrated in Figure 3b, the one-hot label and the highest distributed label should be on the same abscissa, while the distribution degree of other labels is assigned with reduced probability based on the label distance. We denote this label representation as the *distributed label*. Since the distributed labels can more precisely describe the leakage features, learning from the label distribution helps achieve a more robust performance than training with one-hot encoded labels. In addition, the relationship between the traces and distributed labels can be easier mapped, thus effectively relaxing the conditions on the required number of training traces. Indeed, as mentioned in the introductory section, the number of profiling traces is restricted by the time constraint of security evaluation as well as the accessibility and availability of the profiling devices. Profiling with fewer profiling traces would not only speed up the profiling phase but ease the requirement of training a good profiling model as well. As a remark, our learning method fundamentally differs from linear regression attack (LRA) [DPRS11]. Although LRA would also lead to smooth labels by estimating weight parameters to each binary decomposition of the target value, these labels are constructed during the regression process. On the other hand, our method still considers SCA as a classification task. The goal of using distributed labels is to reach more efficient classification.

One should notice the relation between label smoothing [SVI+16] and label distribution. As a regularization technique, label smoothing improves accuracy by computing cross-entropy not with the 'hard' (i.e., one-hot encoded) labels from the dataset but with a weighted mixture of all possible labels with the noise (i.e., uniform) distribution. Label distribution further preserves relations between different labels, thus being more helpful in speeding up the learning process. The performance benchmark between these two techniques can be found in Section VI-A. A natural choice to form distributed labels is a normal distribution. [4]

$$D(y_i) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{y_i - y_{i^*}}{\sigma}\right)^2\right), y_i \in \mathcal{Y}, \tag{4}$$

[3] ChipWhisperer dataset [OC14] is used as it represents measurements obtained from a physical device, where two point-of-interests are selected based on the signal-to-noise ratio to represent the traces. Note that this dataset is not noiseless, but it is difficult to obtain less noisy measurements without resorting to simulations.

[4] The construction of distributed labels should align with the distribution of leakages. This paper assumes that the leakage follows a Gaussian distribution, which is commonly observed in practice.

where $D(y_i)$ denotes the distributed label for label i. i^* stands for the target label. The only adjustable parameter σ depends on the data property (more specifically, the noise in the dataset). In Section VI, we systematically analyze the influence of σ with different datasets, (including their noisy version) and leakage models and then give suggestions on the value selection.

Then, the learning process can be formulated as:

Definition 5. Label distribution learning. Given a training set with trace-label pairs (x, y) sampled from $\mathbf{T}$, where $x \in \mathcal{X}$ and $y \in \mathcal{Y}$, the goal is to learn a function $\mathsf{f}_M^{\boldsymbol{\theta}}$, so that the outputted $\hat{y}$ has a similar distribution to the distributed label $D(y)$.

An essential assumption of label distribution learning is that the label y should be pre-determined by an attacker. Then, the attacker can calculate the distributed label $D(y)$. Next, to optimize the learning parameter θ, instead of using conventional loss functions such as categorical cross-entropy or mean squared error, following [Gen16], Kullback-Leibler (KL) divergence is used as the loss function to measure the similarity between the predicted and ground truth distribution:

$$L = -\sum_i D(y_i)\ln(\hat{y_i}), y_i \in \mathcal{Y}, \tag{5}$$

Where $\hat{y_i}$ denotes the predicted probability for label i.

Stochastic gradient descent is used to minimize the loss function L. Once a network is trained, given a random input x with an unknown label from the attack dataset, the model f outputs a predicted label distribution $\hat{y}$. The predicted label is the one in $\hat{y}$ with the highest probability.

$$i^* = \underset{i}{\operatorname{argmax}}\, \hat{y}_i. \tag{6}$$

V. Augmented Guessing Entropy Metric

A. Key Distribution

As mentioned, label distribution assumes a positive correlation between label distance (to the correct label) and the probability of selecting that label. Under the same assumption, the similarity of different key candidates can be represented by the distance of possible hypothetical leakage data generated by these keys as well. [5] Using AES as an example, we calculate hypothetical leakage data (i.e., the S-box output) for each key candidate with all possible plaintexts. This distribution is denoted as the leakage distribution. The key distribution (KD) is measured by calculating the leakage distribution difference between the key candidates.

KD provides an estimation of the hypothetical distance between key candidates. For a model built in a successful profiling attack (the correct key k^* is the best guess), suppose KD is large between a specific key $k \in \mathcal{K}$ and k^*. Then, k will be likely ranked low (i.e., with guessing entropy close to $2^b - 1$) as it has a negligible probability of being selected. Consequently, KD can be considered an ideal key rank [6] metric indicating the best possible scenario where the correct key is maximally separated from all the other keys.

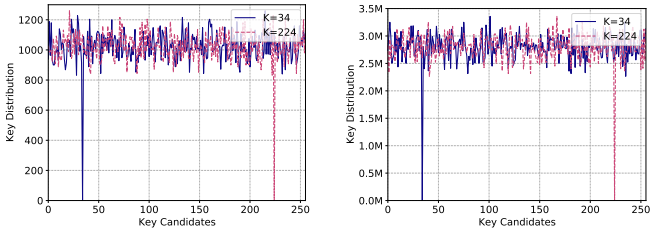
In Eq. (7), we calculate KD based on the Euclidean distance (L_2 norm) between the leakage distribution of all key hypotheses $k \in \mathcal{K}$ and the reference key candidate k^{ref}. We also investigated the Manhattan distance and found the results to be in line but with somewhat smaller discriminate power [7].

$$KD(k^{ref}, k) = \left\| \mathsf{f}(d, k^{ref}) - \mathsf{f}(d, k) \right\|^2, k \in \mathcal{K}. \tag{7}$$

Here, f is the leakage model function (described in Section II-E) that returns the leakage value (labels) according to a key candidate k and data value d. Note that when it is clear from the context, we use the notations $KD(k^{ref}, k)$ and KD interchangeably.

KD gives a unique distribution of all key candidates k based on their difference to the reference key k^{ref}. The selection of k^{ref}, therefore, determines the KD value for each key candidate. The reference key candidate has a distribution difference equal to zero with itself, and the lower the distribution difference, the more similar the key candidate is to the reference key. In practice, the reference key can be set to the k^* (correct key). When k^* is unknown (black-box), k^{ref} should be the most likely key.

The KD definition can be extended to any leakage model, i.e., the Hamming distance or the Least Significant Bit. Figure 4 illustrates the summed KD (for all key candidates) with the HW and ID leakage models for the key candidates $k^{ref} = 34$ (correct third subkey for the ASCAD_r) and $k^{ref} = 224$ (correct third subkey for the ASCAD_f). Although all KD values except for $k^{ref} = k^*$ act as 'noisy' values, the similarity difference between k^{ref} and other keys can be distinguished. One should note that a precise estimation of KD relies on the chosen leakage model. An incorrect leakage model would not only degrade the attack performance, but KD's effectiveness will also drop. Since the publicly available datasets leak mostly in the HW leakage model, we calculate KD with the HW leakage model throughout the paper.



(a) HW leakage model. (b) ID leakage model.

Fig. 4: Illustration for the Key Distribution for the HW/ID leakage models and key candidates 34 and 224.

[5] Aligned with the distributed label, we assume the leakage follows a Gaussian distribution. If this assumption does not hold for the given leakage traces, the calculation of the hypothetical leakage data distance should be adjusted accordingly.

[6] Here, 'ideal' means the perfect fit between an attack model and the leakage. Under this circumstance, the resulting key rank is equivalent to KD as discussed in Section V-B.

[7] Since KD is a list of labels associated with the given key that does not follow any known distribution, f-divergence functions (i.e., KL-divergence, Hellinger distance) are not considered.

B. Augmented Guessing Entropy - AGE

Key distribution defines the distance between k^{ref} and other key candidates. Naturally, a perfectly fitted model should output a key rank similar to KD. Following this, we define a *profiling model fitting metric* by correlating KD with the predicted probability for all $k \in \mathcal{K}$. Since this metric is based on GE but takes into consideration other key candidates besides the correct key, we denote it as Augmented Guessing Entropy (AGE), as a function of KD and the key guessing vector $\mathbf{g}$:

$$AGE = \mathsf{corr}(KD, \mathbf{g}). \tag{8}$$

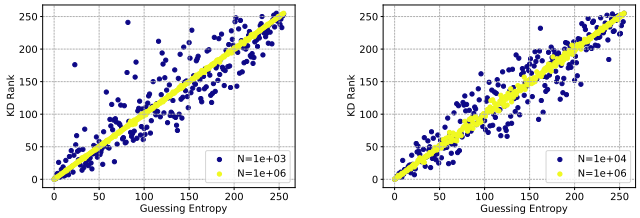
Eq. (8) defines how well a profiling model fits the data concerning a key candidate k^{ref} for a chosen leakage model. The notation corr represents the Spearman correlation [HK11] that evaluates the monotonic relationship with two inputs. We also considered Pearson correlation, but as shown in Figure 4b, KD for the k^{ref} and other keys is around three million. Pearson correlation would be dominated by this high value, thus producing low correlations.

Following Eq. (8), if the profiling model outputs the correct key as the most likely key, one could expect a stronger correlation between KD and $\mathbf{g}$. Conversely, if the profiling model fails to fit the data, the outputted random, but still, most likely key, would lead to a low correlation between KD and $\mathbf{g}$. As a demonstration, Figure 5 depicts the 'almost' $perfectly\ \ fitted\ \ profiling\ \ model$ for the HW and ID leakage models. We use simulated measurements with strong HW and ID leakages and a controlled Gaussian noise level, normally distributed with a variance of 0.01 around a mean of zero. The simulated traces have two features that hold the leakage, which is proportional to $HW(S_{box}(d \oplus k))$ and $S_{box}(d \oplus k)$, to simulate the ideal HW and ID leakages, respectively. The profiling set has plaintexts d and keys k chosen from a uniformly random distribution. The attack set's plaintexts are selected uniformly at random, while the attack key is the same for the whole dataset. We use the template attack and consider the increasing number of profiling traces N. In both figures, AGE increases w.r.t. the number of profiling traces and reaches 0.999 and 0.998 for the HW and ID leakage models. The results confirm that the correlation between KD and $\mathbf{g}$ tends to increase with better (more fit) models (since we use template attack, better models are those that are trained with more traces).

Definition 6. Perfectly fitted profiling model. A perfectly fitted profiling model reaches $AGE = 1$ in the attack phase for any set of Q attack traces. [8]

It is worth mentioning that KD can also be used to calculate the confusion covariance metric ($\mathbb{E}(KD)$) [FLD12], a metric designed initially to measure the DPA resistance of S-boxes. A low expectation of KD indicates less distinctive intermediate data, which could lead to reduced data leakage. On top of that, the AGE metric suggests that the variance of KD should also be low to secure the target. Indeed, a low KD variance indicates high similarity between different keys, which will lead to a less deterministic order of $\mathbf{g}$. Since the AGE value is more likely to be low in this case, one can expect more effort in obtaining the security assets via SCA. Another way of perturbing $\mathbf{g}$ is by introducing additional noise or countermeasures, a common implementation in modern products.

[8] We assume there are many possible ways to select Q traces from the available traces.



(a) HW leakage model. (b) ID leakage model.

Fig. 5: 'Perfectly' fitted profiling model with template attack, considering the HW and ID leakage models and simulated traces with an increasing number of profiling traces N. KD ranks (Y-axis) stands for a sorted KD.

VI. Experimental Results

A. Profiling with Distributed Labels

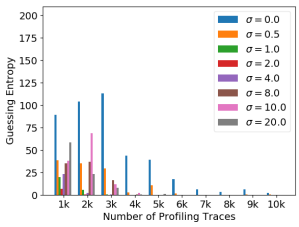
In Section IV, we argue that the distributed label enlarges the description degrees of labels to the leakage traces and can lead to more efficient learning even with a reduced number of training examples. We validate this assumption by training the state-of-the-art CNNs [RWPP21] and MLPs [WPP20] with a different number of profiling traces. The models' hyperparameters are listed in Appendix A. We use a wide selection of DL architectures to ensure the generalization of the results. Besides, we tune the σ value of the distributed label to find the optimal value for different training settings. The distributed labels are pre-computed before the training starts. To obtain the most representative performance, the attack results of each training setting (σ and profiling traces number) are the median value from 20 independent training (and attacks) with random weight initialization following recommendations from [WPP22b].

Figures 6, 7, and 8 show the results for the ASCAD_f, ASCAD_r, and CHES_CTF datasets, respectively. The conventional training method (one-hot encoded label) is represented with $\sigma = 0.0$ (blue bar). When training with the conventional method, we used the categorical cross-entropy (CCE) loss. CCE is a standard loss function for classification tasks and is widely used in DL-SCA. When learning from the label distribution, the KL divergence loss is used to measure the distribution difference between the true and predicted label distribution.

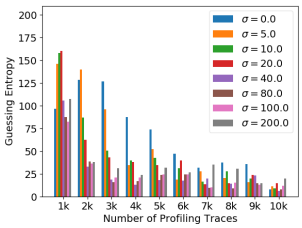
For the ASCAD_f dataset, as shown in Figure 6, by distributing HW-based labels, GE equal to zero can be reached with up to 3 000 profiling traces for both MLP and CNN within the given number of attack traces, which is more than ten times less than the number of the profiling traces commonly used in literature (50 000). At the same time, more than 10 000 profiling traces are insufficient when considering

the conventional training method ($\sigma = 0.0$). Using the ID leakage model, although one-hot encoded labels lead to better performance in some cases (discussed in later paragraphs), one can confirm the advantages of using the distributed label in low profiling settings.

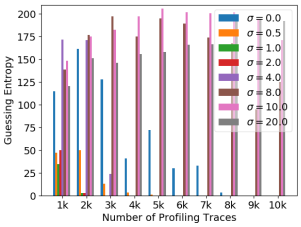
When looking at the influence of the label distribution variation σ (Figure 6), although different numbers of profiling traces, leakage models, and attack models are considered, the optimal settings show consistency: for the HW leakage model, σ ranges from 1 to 2 can lead to the best attack performance. This value increases to 40 to 80 for the ID leakage model. In addition, we have tested the traces with Gaussian noise levels 2 and 4. While the optimal value of sigma defined in the paper still holds for most settings, we expect the best sigma to be larger since the leakage traces become more difficult to correctly classify.



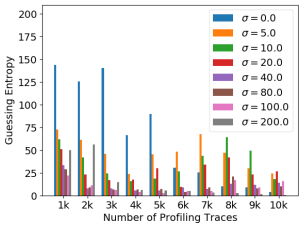
(a) MLP with the HW leakage model.



(b) MLP with the ID leakage model.



(c) CNN with the HW leakage model.

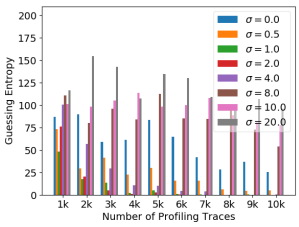


(d) CNN with the ID leakage model.

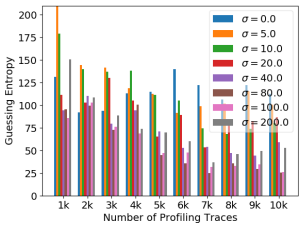
Fig. 6: Label distribution learning on the ASCAD_f dataset.

Although ASCAD_r is considered a dataset more difficult to break than ASCAD_f [WPP20], as shown in Figure 7, the distributed label boosts the attack performance significantly. For the HW leakage model, around 6 000 profiling traces are sufficient for MLP and CNN models to reach GE of zero, which is around ten times less than the related works (¿50 000 profiling traces). For the ID leakage model, aligned with the attack on the ASCAD_f dataset, although none of the training settings can retrieve the secret information with 5 000 attack traces, label distribution learning halves the GE value compared with its one-hot encoded counterpart, indicating a faster GE convergence with our learning scheme.

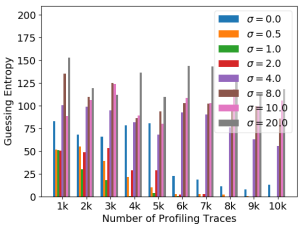
Finally, similar results can be obtained when attacking the CHES_CTF dataset. Since this dataset leaks limited ID leakage according to literature [WPP20], [RWPP21], we attack with the HW leakage model only. With the MLP and CNN models, 5 000 profiling traces are needed to break the target, which is nine times less than the traces used in the literature (45 000 traces). It is important to note that the optimal σ setting shows similarity for all three tested datasets. From the experimental results on three datasets, good prior knowledge about the leakage model is necessary to construct a meaningful label distribution. Still, when attacking leakages from other devices, one could start with low σ and monitor the attack performance until it reaches optimal behavior.



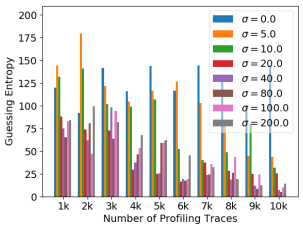
(a) MLP with the HW leakage model.



(b) MLP with the ID leakage model.

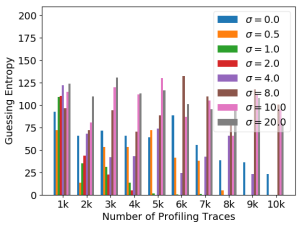


(c) CNN with the HW leakage model.

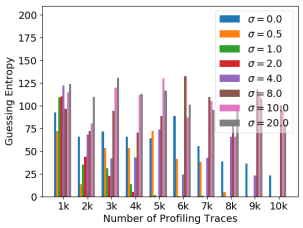


(d) CNN with the ID leakage model.

Fig. 7: Label distribution learning on the ASCAD_r dataset.



(a) MLP with the HW leakage model.



(b) CNN with the HW leakage model.

Fig. 8: Label distribution learning on the CHES_CTF dataset.

Indeed, there are various techniques available when the profiling traces are limited. To better illustrate the pros and cons of label distribution learning when compared with these methods, we benchmark the attack performance of previously used state-of-the-art (SotA) MLPs and CNNs with multiple profiling settings. We consider several commonly-used techniques to counter the limitation of the training data.

- 10 000 profiling traces with and without techniques such as label distribution, label smoothing, L2 regularization, and dropout.
- 50 000 profiling traces, obtained directly or generated with Gaussian noise-based data augmentation.
- 100 000 profiling traces generated from 10 000 traces with Gaussian noise-based data augmentation.

The dropout rate and regularization factor are tuned to 5e-2 and 1e-4. For data augmentation, four augmentation levels

(0.25, 0.5, 0.75, 1.0) are selected following [KPH+19], and the one with the best performance is presented in the benchmark. The label smoothing factor is set to be optimal based on the various search options. [9] Each profiling setting is tested with two label formats: one-hot encoded label and distributed label. For label distributed learning, σ is set to 1/2 and 40/80 for HW and ID leakage models. The attack performance is evaluated by calculating the required number of attack traces to reach GE of zero, denoted as T_{GE0}. The results presented are the median T_{GE0} from 20 independently trained models. If an attack setting failed to reach GE zero with a given number of attack traces, the corresponding results are marked with "-".

Traces	Label	ASCAD_f	ASCAD_r	CHES_CTF
10 000	One-hot	-/-	-/-	-
	Smoothed	3 484/**-**	**-/-**	**-**
	Distributed	**1 618/4 964**	**3 623/4 892**	**2 337**
10 000	One-hot	-/-	-/-	3 728
(L2)	Distributed	-/-	-/-	**1 930**
10 000	One-hot	-/-	-/-	4 156
(Dropout)	Distributed	**2 264**/-	-/-	**2 493**
50 000	One-hot	**1 219/182**	970/**2 625**	**567**
	Distributed	1 421/3 530	**919**/-	905
50 000	One-hot	1 588/-	-/-	-
(Augmented)	Distributed	**1 095/4 728**	**2 784**/-	**2 735**
100 000	One-hot	**1 254**/-	-/-	-
(Augmented)	Distributed	1 447/**4 895**	**2 998**/-	**2 793**

TABLE I: Benchmark the attack performance (T_{GE0}) with SotA MLP. Attack results for the HW and ID leakage models are separated by '/'.

Traces	Label	ASCAD_f	ASCAD_r	CHES_CTF
10 000	One-hot	2940/-	-/-	-
	Smoothed	**2 994/-**	**-/-**	**-**
	Distributed	**1 252/4 050**	**1 939/3 753**	**2 182**
10 000	One-hot	2 217/**3 779**	-/-	-
(L2)	Distributed	**1 096**/-	**2 034/4 892**	**1 458**
10 000	One-hot	1 913/-	-/-	-
(Dropout)	Distributed	**1 338/4 219**	-/-	**1 868**
50 000	One-hot	**544/87**	650/**487**	455
	Distributed	779/-	**553**/3 684	**450**
50 000	One-hot	2 829/**4 061**	-/-	-
(Augmented)	Distributed	**1 201**/-	**2 190**/-	**1 724**
100 000	One-hot	2 278/**1 621**	-/-	-
(Augmented)	Distributed	**1 218**/-	**2 298**/-	**2 105**

TABLE II: Benchmark the attack performance (T_{GE0}) with SotA CNN. Attack results for the HW and ID leakage models are separated by '/'.

The benchmark results are shown in Table I and Table II. The best results for each profiling setting are marked in **bold**. With limited (10 000) profiling traces, distributed labels bring a significant performance boost with both attack and leakage models. The considered regularization techniques are helpful in some attack settings, improving significantly when combined with distributed labels. Similarly, data augmentation helps obtain better performance in some cases; the combination with distributed labels makes it even better. In practice, due to the limitation of controlled devices and time budget, attackers would likely use smaller networks, more regularization, and more data augmentation to run their attacks in lower-data settings. However, as shown in the table, label distribution is the best technique considering the additional efforts to tune hyperparameters and their performance.

It is worth noting that one-hot encoded labels generally produce better results by directly training with 50 000 profiling traces. Indeed, an increased number of profiling traces enlarges the side-effect of the distributed label: high estimation variance, leading to reduced predictive performance. Although data augmentation could also generate more leakages, the difficulty of setting a proper augmentation level makes the generated traces less helpful in the profiling phase. An evaluator should consider the risk of using distributed labels if many traces are available. The simplest way to validate the traces' sufficiency is to profile with both one-hot encoded and distributed labels (with low σ) and monitor if the attack performance is improved with distributed labels.

B. AGE Use Cases

In this section, we investigate the effectiveness of the AGE metric for different use cases. Specifically, we consider network architecture search (NAS) and overfitting prevention as they have a major influence on the attack performance with DL-based SCA. Indeed, adjusting the profiling model size will directly influence its learning capacity. On the other hand, a properly set training epoch number could improve the model's fitness to the dataset. Since these two aspects rely on well-performing evaluation metrics [RWPP21], [PCP20], we show the performance of AGE in various settings and benchmark it with other common metrics.

1) Early Stopping: As an evaluation metric, AGE can be used as early stopping regularization or as an indicator of when to save the best model. For illustration, we evaluate state-of-the-art models by training with a different number of epochs ranging from 1 to 150 in steps of 10. Aligned with previous sections, the attack performance is evaluated by T_{GE0}. Besides, four metrics, accuracy, loss, mutual information (MI) [PBP21], and AGE, are calculated per epoch with 5 000 validation traces. [10] One may argue that T_{GE0} can be used as an evaluation metric. However, T_{GE0} can only be calculated when GE equals zero. For a model that cannot break the target with a given number of attack traces, T_{GE0} is not indicative. Similarly, the key rank metric is only indicative when GE is larger than zero: when the key rank stays zero, one cannot know if the model is still learning or starting overfitting. Following this, since all of the selected models reached the key rank of zero quickly and never change, we omit the key rank metric as it is less indicative in the training process.

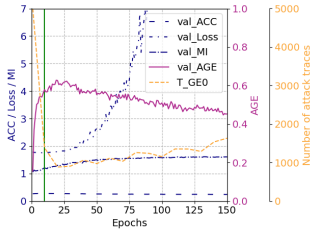
The results for three datasets and two leakage models are shown in Figures 9, 10, and 11. Since the metrics and T_{GE0} have different scales, multiple Y-axes are used to scale

[9]The possible label smoothing factors are 0.05, 0.1, 0.2, 0.3, 0.4, 0.5, 1, 5, and 10.

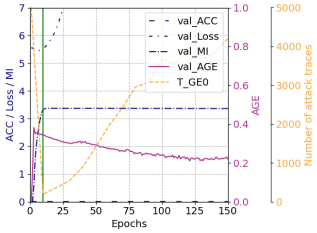
[10]If GE is greater than zero, T_{GE0}=5 000.

the results data. The optimal training epoch proposed in the literature is marked by green vertical lines (10 for MLPs and 50 for CNNs). Aligned with the previous section, all of the presented results are the median from 20 independent pieces of training.

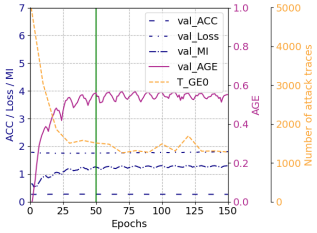
For ASCAD_f, using T_{GE0} as a reference, AGE perfectly reflects the variation of the model's generalization with different training epochs. For instance, in Figure 9a, T_{GE0} starts to increase when the training epoch exceeds 25, indicating that the model is overfitting. Interestingly, AGE indicates the overfitting effect even a bit earlier than the attack performance starts to degrade. Based on Figure 4, the order of the key candidates with closer KD values would more likely be perturbed when overfitting starts. Only after the overfitting effect accumulates to a certain level (i.e., with more training epochs) the "disorder" of the key candidate would propagate to the correct key, finally captured by the GE-related metrics. From the practical perspective, AGE can be a good candidate as an early stopping indicator due to its high sensitivity to overfitting. In terms of other metrics, the MI metric is somewhat misleading because it keeps on increasing, although the T_{GE0} suggests that the performance becomes worse. The loss value is only helpful in limited cases (e.g., Figure 9a), which confirms the conclusion from picek *et al.* [PHJ+18] that it is commonly not considered a good evaluation metric for SCA. The accuracy metric remains mostly stable with a different number of training epochs, indicating its mediocre performance. Finally, when looking at the optimal training epoch, the ones used in the literature are not optimal for Figures 9a and 9d. On the other hand, AGE consistently indicates the epoch that reaches the best attack performance.



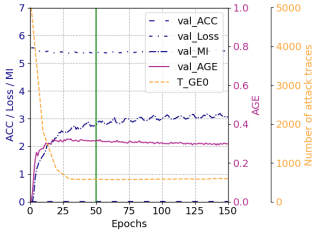
(a) MLP with the HW leakage model.



(b) MLP with the ID leakage model.



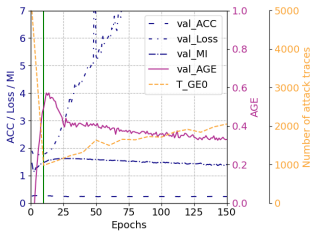
(c) CNN with the HW leakage model.



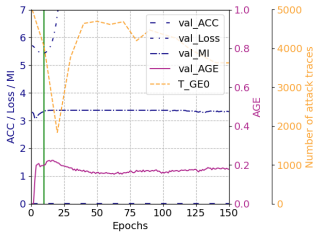
(d) CNN with the ID leakage model.

Fig. 9: Metrics performance on the ASCAD_f dataset.

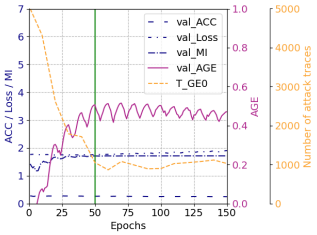
Attacks on ASCAD_r and CHES_CTF show consistent results with ASCAD_f. AGE performs the best among all evaluated metrics, representing the attack performance precisely. As an evaluation metric, AGE combines the advantages of key rank and T_{GE0} with limited computation overhead, thus becoming a reliable metric for the applications such as early stopping.



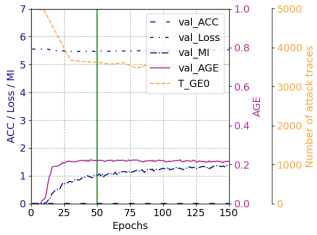
(a) MLP with the HW leakage model.



(b) MLP with the ID leakage model.

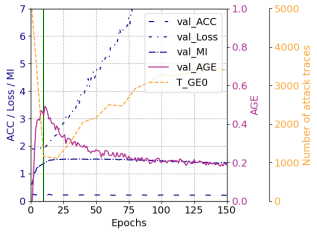


(c) CNN with the HW leakage model.

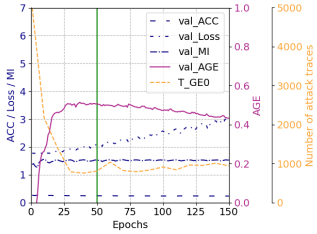


(d) CNN with the ID leakage model.

Fig. 10: Metrics performance on the ASCAD_r dataset.



(a) MLP with the HW leakage model.



(b) MLP with the ID leakage model.

Fig. 11: Metrics performance on the CHES_CTF dataset.

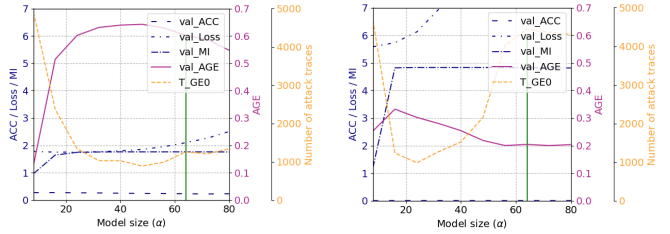
2) Network Architecture Search: Network architecture search (NAS) is essential in DL-SCA. A smartly designed neural network can not only break the target but reduce the training complexity as well [ZBHV19], [RWPP21]. To better illustrate the advantage of the AGE metric, we use CNN listed in Table III with a tunable α parameter to control the size of the deep learning model. Specifically, α determines the number of filters in convolutional layers and neurons in the fully connected layers. We use α (range from 1 to 64) to estimate the complexity of a profiling model. Note, for the CNN_best from [BPS+20], α equals 64. The training epoch is set to be optimal (75) based on [BPS+20], which is represented by the green vertical line in the plot. This section presents the results for the ASCAD_f and ASCAD_r datasets only. Since CHES_CTF produce similar results, we omit them from this section.

The results are shown in Figure 12. Aligned with the previous section, accuracy, loss, MI, and AGE are used

TABLE III: CNN architecture used for the attack.

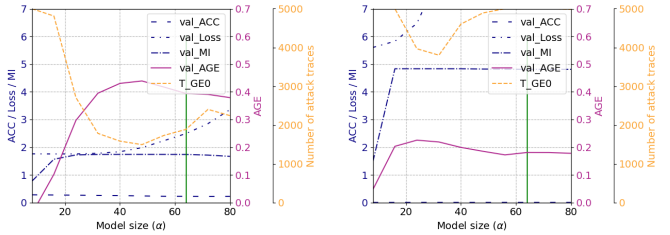
Layer Types	Filter Size	# of Filters	Pooling Stride	# of Neurons
Conv block	11	a*1	2	-
Conv block	11	a*2	2	-
Conv block	11	a*4	2	-
Conv block	11	a*8	2	-
Flatten	-	-	-	-
Fully connected (2×)	-	-	-	a*64

as evaluation metrics. As a reference, T_{GE0} represents the attack performance. Among the three considered metrics, AGE best represents the attack performance. For instance, in Figure 12a, T_{GE0} reaches minimum when α equals around 50. Further increase of the profiling model size degrades the attack performance, meaning the fitness reduction for a dataset. The AGE metric perfectly represents this tendency, as it reaches the maximum when α is around the same model size, then decreases gradually. Regarding other metrics, the validation accuracy has limited changes regardless of the variation of α. Validation loss, in contrast, is more indicative than its counterpart. However, it is challenging to judge when to stop the training. For instance, the loss value in Figure 12c suggests that the profiling should end after training with around 35 epochs, but the best performance is reached 15 epochs later. MI keeps on increasing with the HW leakage model. However, it does not correctly reflect the attack performance. Finally, the training epoch suggested in the literature is still sub-optimal when looking at the results (i.e., Figure 12b). Using AGE as an evaluation metric can help monitor the attack performance in various settings.



(a) ASCAD_f with the HW leakage model.

(b) ASCAD_f with the ID leakage model.



(c) ASCAD_r with the HW leakage model.

(d) ASCAD_r with the ID leakage model.

Fig. 12: Metrics performance with different model sizes.

In addition, we have also tested the influence of the noise on the considered metrics by adding Gaussian noise to the traces with incremental variations ranging from 0 to 10 in a step of 0.5. The results show that the AGE metric can correctly and precisely reflect the negative influence introduced by the noise. Since the results align with the conclusions from the previous sections, the results are omitted.

In conclusion, the AGE metric reliably reflects the generality of the profiling model in various training conditions. Compared with other metrics, the evaluation of the keys' order helps in increasing the sensitivity of the AGE metric in measuring the model's performance. Indeed, in almost all of the experimental results, AGE is the first metric that indicates the overfitting effect. Additionally, due to its computation simplicity, we believe AGE is an ideal candidate as an evaluation metric.

VII. Conclusions and Future Work

In the profiling side-channel analysis, one commonly uses intermediate data to form a one-hot encoded label for the profiling. Additionally, it is common to use guessing entropy to estimate the attack performance. This paper introduces distributed labels as a new learning approach that can effectively reduce the required number of profiling traces. Then, based on the relationship between each key candidate, we define the Key distribution (KD) metric and use it to form a novel AGE metric. Our results show that the AGE metric can be a reliable candidate for evaluating the generality of a model, which has been validated with two use cases: early stopping and network architecture search. Our findings are confirmed for several experiments considering various usage cases, attack methods, leakage models, and datasets.

In future work, we plan to extend the application of label distribution for masked implementations. For instance, by setting multiple peaks in the distributed label to reflect the higher-order property of the information. In terms of the AGE metric, since the key distribution relies on the hypothetical distance between key candidates, the distance depends on the algorithm and hardware implementation. Following this, we plan to investigate if the method can be easily adapted to a new implementation or a new algorithm. Moreover, it would be interesting to explore AGE in the context of leakage assessment. Finally, applying our results to the non-profiling SCA would be an exciting research direction.

Appendix

The used state-of-the-art models are listed in Tables IV and V. All of the non-listed hyperparameter settings are aligned with the original papers [RWPP21], [WPP20]. The convolution layer is denoted by C; averaging pooling layer is denoted by P. FLAT and FC denote the flatten layer and fully connected layer, respectively. SM denotes the output layer with the $softmax$ activation function.

References

[BPS+20] Ryad Benadjila, Emmanuel Prouff, Rémi Strullu, Eleonora Cagli, and Cécile Dumas. Deep learning for side-channel analysis and introduction to ASCAD database. *J. Cryptographic Engineering*, 10(2):163–188, 2020.

TABLE IV: CNN architecture used for the attack [RWPP21].

Dataset	Leakage Model	Architectures	Learning Rate	Batch Size
ASCAD_f	HW	C(2,25,1), P(4,4), FLAT, FC(15, 10, 4), SM(9)	5e-3	50
	ID	C(128,25,1), P(25,25), FLAT, FC(20, 15), SM(256)	5e-3	50
ASCAD_r	HW	C(4,50,1), P(25,25), FLAT, FC(30, 30, 30), SM(9)	5e-3	128
	ID	C(128,3,1), P(75,75), FLAT, FC(30, 2), SM(256)	5e-3	128
CHES_CTF	HW	C(2,2,1), P(7,7), FLAT, FC(10), SM(9)	5e-3	128

TABLE V: MLP architecture used for the attack [WPP20].

Dataset	Leakage Model	Architectures	Learning Rate	Batch Size
ASCAD_f	HW	FC(496, 496, 136, 288, 552, 408, 232, 856), SM(9)	5e-4	32
	ID	FC(160, 160, 624, 776, 328, 968), SM(256)	1e-4	32
ASCAD_r	HW	FC(200, 200, 304, 832, 176, 872, 608, 512), SM(9)	5e-4	32
	ID	FC(256, 256, 296, 840, 280, 568, 672), SM(256)	5e-4	32
CHES_CTF	HW	FC(192, 192, 616, 248, 440), SM(9)	1e-3	32

[Car11] IC Card. Emv integrated circuit card specifications for payment systems, book 3 application specification, November 2011. https://www.emvco.com/wp-content/uploads/2017/04/EMV_v4.3_Book_3_Application_Specification_20120607062110791.pdf.

[CDP17] Eleonora Cagli, Cécile Dumas, and Emmanuel Prouff. Convolutional neural networks with data augmentation against jitter-based countermeasures. In Wieland Fischer and Naofumi Homma, editors, *Cryptographic Hardware and Embedded Systems – CHES 2017*, pages 45–68, Cham, 2017. Springer International Publishing.

[CRR02] Suresh Chari, Josyula R. Rao, and Pankaj Rohatgi. Template attacks. In Burton S. Kaliski Jr., Çetin Kaya Koç, and Christof Paar, editors, *Cryptographic Hardware and Embedded Systems - CHES 2002, 4th International Workshop, Redwood Shores, CA, USA, August 13-15, 2002, Revised Papers*, volume 2523 of *Lecture Notes in Computer Science*, pages 13–28. Springer, 2002.

[DPRS11] Julien Doget, Emmanuel Prouff, Matthieu Rivain, and François-Xavier Standaert. Univariate side channel attacks and leakage modeling. *Journal of Cryptographic Engineering*, 1(2):123–144, 2011.

[FLD12] Yunsi Fei, Qiasi Luo, and A Adam Ding. A statistical model for dpa with novel algorithmic confusion analysis. In *International Workshop on Cryptographic Hardware and Embedded Systems*, pages 233–250. Springer, 2012.

[Gen16] Xin Geng. Label distribution learning. *IEEE Transactions on Knowledge and Data Engineering*, 28(7):1734–1748, 2016.

[GHO15] R. Gilmore, N. Hanley, and M. O'Neill. Neural network based attack on a masked implementation of AES. In *2015 IEEE International Symposium on Hardware Oriented Security and Trust (HOST)*, pages 106–111, May 2015.

[GWX14] Xin Geng, Qin Wang, and Yu Xia. Facial age estimation by adaptive label distribution learning. In *2014 22nd International Conference on Pattern Recognition*, pages 4465–4470. IEEE, 2014.

[HGM+11] Gabriel Hospodar, Benedikt Gierlichs, Elke De Mulder, Ingrid Verbauwhede, and Joos Vandewalle. Machine learning in side-channel analysis: a first study. *J. Cryptogr. Eng.*, 1(4):293–302, 2011.

[HK11] Jan Hauke and Tomasz Kossowski. Comparison of values of pearson's and spearman's correlation coefficients on the same sets of data. *Quaestiones geographicae*, 30(2):87, 2011.

[HZ12] Annelie Heuser and Michael Zohner. Intelligent Machine Homicide - Breaking Cryptographic Devices Using Support Vector Machines. In Werner Schindler and Sorin A. Huss, editors, *COSADE*, volume 7275 of *LNCS*, pages 249–264. Springer, 2012.

[KJJ99] Paul C. Kocher, Joshua Jaffe, and Benjamin Jun. Differential power analysis. In Michael J. Wiener, editor, *Advances in Cryptology - CRYPTO '99, 19th Annual International Cryptology Conference, Santa Barbara, California, USA, August 15-19, 1999, Proceedings*, volume 1666 of *Lecture Notes in Computer Science*, pages 388–397. Springer, 1999.

[KPH+19] Jaehun Kim, Stjepan Picek, Annelie Heuser, Shivam Bhasin, and Alan Hanjalic. Make some noise. unleashing the power of convolutional neural networks for profiled side-channel analysis. *IACR Transactions on Cryptographic Hardware and Embedded Systems*, pages 148–179, 2019.

[KS20] Y Kiran Kumar and R Mahammad Shafi. An efficient and secure data storage in cloud computing using modified rsa public key cryptosystem. *International Journal of Electrical and Computer Engineering*, 10(1):530, 2020.

[LMBM13] Liran Lerman, Stephane Fernandes Medeiros, Gianluca Bontempi, and Olivier Markowitch. A Machine Learning Approach Against a Masked AES. In *CARDIS*, Lecture Notes in Computer Science. Springer, November 2013. Berlin, Germany.

[LPB+15] Liran Lerman, Romain Poussier, Gianluca Bontempi, Olivier Markowitch, and François-Xavier Standaert. Template attacks vs. machine learning revisited (and the curse of dimensionality in side-channel analysis). In *International Workshop on Constructive Side-Channel Analysis and Secure Design*, pages 20–33. Springer, 2015.

[LZC+21] Xiangjun Lu, Chi Zhang, Pei Cao, Dawu Gu, and Haining Lu. Pay attention to raw traces: A deep learning architecture for end-to-end profiling attacks. *IACR Transactions on Cryptographic Hardware and Embedded Systems*, pages 235–274, 2021.

[MDP19] Loïc Masure, Cécile Dumas, and Emmanuel Prouff. A comprehensive study of deep learning for side-channel analysis. *IACR Transactions on Cryptographic Hardware and Embedded Systems*, 2020(1):348–375, Nov. 2019.

[MPP16] Houssem Maghrebi, Thibault Portigliatti, and Emmanuel Prouff. Breaking cryptographic implementations using deep learning techniques. In *International Conference on Security, Privacy, and Applied Cryptography Engineering*, pages 3–26. Springer, 2016.

[OC14] Colin O'Flynn and Zhizhang David Chen. Chipwhisperer: An open-source platform for hardware embedded security research. In *International Workshop on Constructive Side-Channel Analysis and Secure Design*, pages 243–260. Springer, 2014.

[PBP21] Guilherme Perin, Ileana Buhan, and Stjepan Picek. Learning when to stop: A mutual information approach to prevent overfitting in profiled side-channel analysis. In *COSADE*, volume 12910 of *Lecture Notes in Computer Science*, pages 53–81. Springer, 2021.

[PCP20] Guilherme Perin, Lukasz Chmielewski, and Stjepan Picek. Strength in numbers: Improving generalization with ensembles in machine learning-based profiled side-channel analysis. *IACR Transactions on Cryptographic Hardware and Embedded Systems*, 2020(4):337–364, Aug. 2020.

[PHJ+17] Stjepan Picek, Annelie Heuser, Alan Jovic, Simone A. Ludwig, Sylvain Guilley, Domagoj Jakobovic, and Nele Mentens. Side-channel analysis and machine learning: A practical perspective. In *2017 International Joint Conference on Neural Networks, IJCNN 2017, Anchorage, AK, USA, May 14-19, 2017*, pages 4095–4102, 2017.

[PHJ+18] Stjepan Picek, Annelie Heuser, Alan Jovic, Shivam Bhasin, and Francesco Regazzoni. The curse of class imbalance and conflicting metrics with machine learning for side-channel evaluations. *IACR Transactions on Cryptographic Hardware and Embedded Systems*, 2019(1):209–237, Nov. 2018.

[PHPG22] Stjepan Picek, Annelie Heuser, Guilherme Perin, and Sylvain Guilley. Profiled side-channel analysis in the efficient attacker framework. In *Smart Card Research and Advanced Applications*, pages 44–63. Springer International Publishing, 2022.

[PSG16] Romain Poussier, François-Xavier Standaert, and Vincent Grosso. Simple key enumeration (and rank estimation) using histograms: An integrated approach. In *International Conference on Cryptographic Hardware and Embedded Systems*, pages 61–81. Springer, 2016.

[PWP22] Guilherme Perin, Lichao Wu, and Stjepan Picek. Exploring feature selection scenarios for deep learning-based side-channel analysis. *IACR Transactions on Cryptographic Hardware and Embedded Systems*, 2022(4):828–861, Aug. 2022.

[RWPP21] Jorai Rijsdijk, Lichao Wu, Guilherme Perin, and Stjepan Picek. Reinforcement learning for hyperparameter tuning in deep learning-based side-channel analysis. *IACR Transactions on Cryptographic Hardware and Embedded Systems*, 2021(3):677–707, Jul. 2021.

[SLP05] Werner Schindler, Kerstin Lemke, and Christof Paar. A stochastic model for differential side channel cryptanalysis. In Josyula R. Rao and Berk Sunar, editors, *Cryptographic Hardware and Embedded Systems – CHES 2005*, pages 30–46, Berlin, Heidelberg, 2005. Springer Berlin Heidelberg.

[SMY09] François-Xavier Standaert, Tal G. Malkin, and Moti Yung. A unified framework for the analysis of side-channel key recovery attacks. In Antoine Joux, editor, *Advances in Cryptology - EUROCRYPT 2009*, pages 443–461, Berlin, Heidelberg, 2009. Springer Berlin Heidelberg.

[SVI+16] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016.

[WAGP20] Lennert Wouters, Victor Arribas, Benedikt Gierlichs, and Bart Preneel. Revisiting a methodology for efficient cnn architectures in profiling attacks. *IACR Transactions on Cryptographic Hardware and Embedded Systems*, 2020(3):147–168, Jun. 2020.

[WPP20] Lichao Wu, Guilherme Perin, and Stjepan Picek. I choose you: Automated hyperparameter tuning for deep learning-based side-channel analysis. *Cryptology ePrint Archive*, 2020.

[WPP22a] Lichao Wu, Guilherme Perin, and Stjepan Picek. The best of two worlds: Deep learning-assisted template attack. *IACR Transactions on Cryptographic Hardware and Embedded Systems*, 2022(3):413–437, Jun. 2022.

[WPP22b] Lichao Wu, Guilherme Perin, and Stjepan Picek. On the evaluation of deep learning-based side-channel analysis. In *Constructive Side-Channel Analysis and Secure Design: 13th International Workshop, COSADE 2022, Leuven, Belgium, April 11-12, 2022, Proceedings*, volume 13211, page 49. Springer, 2022.

[XHG+17] Di Xue, Zheng Hong, Shize Guo, Liang Gao, Lifa Wu, Jinghua Zheng, and Nan Zhao. Personality recognition on social media with label distribution learning. *IEEE access*, 5:13478–13488, 2017.

[ZBHV19] Gabriel Zaid, Lilian Bossuet, Amaury Habrard, and Alexandre Venelli. Methodology for efficient cnn architectures in profiling attacks. *IACR Transactions on Cryptographic Hardware and Embedded Systems*, 2020(1):1–36, Nov. 2019.