Overview of Side Channel Cipher Analysis Based on Deep Learning

Shijie Song\textsuperscript{1,a}, Kaiyan Chen\textsuperscript{1} and Yang Zhang\textsuperscript{1}

\textsuperscript{1} Virtual Simulation Department, Room 97, HePing West Road, XinHua District, ShiJiaZhuang City, Hebei Prov, China

Abstract. According to the widespread use of embedded cryptographic devices, one of the major security threats, side channel attack and protection research, has grown rapidly in the last decade. Along with the popularity of machine learning and deep learning, artificial intelligence technology combined with side channel technology has gradually emerged, which has achieved quite good processing results. Due to the popularity of deep learning in recent years, it is indeed to complete the work with making an overview of side channel analysis based on deep learning. So in order to perfect the theory, this paper begins with the basic theory of the side channel attack. Combined with common models, the methodological analysis of deep learning side channel analysis is summarized and the existing research results are introduced next. Finally, the paper looks forward to the future deep learning side channel technology.

1. Introduction

Smart devices such as smart cards, mobile phones, and ATM devices have played an important role in modern society nowadays. Although the keys from those used for cryptographic algorithms are usually secure, with the development of computing technology, the vulnerability of traditional encryption operation is gradually exposed. An attacker can use the leakage information generated from the physical implementation. It is named as side channel analysis/attack (SCA)[1] which can seriously jeopardize the overall security of the system. The demand for security in many embedded device applications, including Internet of Things, financial transactions, electronic communications, and data storage, has led to the development of side channel attacks.

At the same time, deep learning[2], which is a branch of machine learning, is very popular in the wave of big data development in recent years. It uses deep neural networks to learn the features from complex data and makes decisions on another set of data analysis. It has excellent feature extraction and classification functions. It is also a new era sword that can be combined with SCA.

2. Side Channel Analysis

SCA is a type of cryptanalysis that uses the physical leakage of a cryptosystem to recover passwords. It is also a black box model. Since Paul Kocher[1] proposed SCA, the traditional by using time, power and electromagnetic attack methods are changing to the ways of algebra, profiling and false, of which the complexity is gradually increased, and the accuracy rate is correspondingly improved. In SCA, there are two categories:

Profiling SCA. Theoretically, the profiling SCA is the most powerful means of SCA, because the adversary can profile and precisely adjust the attack parameters under the premise that the sample device can be obtained exactly the same as the target device. Such attacks include template attacks[3] and stochastic models[4] (a.k.a linear regression analysis). The profiling SCA consists of two phases:
the modeling (or training) phase and the matching (or attack) phase.

Profiling phase: The adversary is able to calculate a conditional probability distribution function 
\( \hat{g}_k \) for each known key \( k \in \mathcal{K} \) in the sample device:
\[
g_k : (\tilde{l}, p) \rightarrow \text{P}[\tilde{L} = \tilde{l} | (P, K) = (p, k)]
\]  

(1)

\( p \) is the known plaintext and the traces are denoted as \( \tilde{l} \in L \).

Matching phase: Estimating the unknown key \( k \) of the target device maximum likelihood principle was used:
\[
d[k] = \frac{\prod_{i=1}^{N} \frac{P[\tilde{L} = \tilde{l}_i | (P, K) = (p_i, k)]}{f_i(\tilde{l}_i)} \times f_{\text{prior}}(p_i, k)}{f_{\text{prior}}(p, k)}
\]  

(2)

Non-profiling SCA: This type of SCA corresponds to a weaker attacker who can only access physical leakages captured on the target device. To recover the keys, a non-profiling SCA uses statistical analysis to detect the correlation between the leaked measurement and the sensitive variable. It includes (different)power attack[1], correlation power analysis[5] and mutual information analysis[6].

3. Deep Learning

3.1. Overview of Deep Learning

Deep learning uses a deep neural network model which has a good effect on artificial intelligence processing such as voice, image and text with hierarchical structure. The history of deep learning is actually the history of artificial neural networks. In 1957, Rosenblatt[7] first proposed the concept of the perceptron as a independent unit of neural network. Subsequently, Widrow and Hoff[8] not only succeeded in artificial neural networks on computers, but also proposed the famous gradient descent algorithm, which laid the foundation for the BP neural network proposed by Rumelhart and McCelland[9].MLP (Multiple layers perceptron ) was first proposed by Werbos[10] in 1981. It is also a back propagation algorithm based on various specific neural networks. In 1982, Hopfield[11] proposed Hopfield neural network to simulate brain memory which was intended to solve the problem that it was easy to fall into the local minimum value that may not be the global optimal solution when the BP neural network implement the gradient descent algorithm .In order to optimize the Hopfield neural network, in 1985, Hinton and Sejnowski[12] considered with the annealing algorithm of the Boltzmann machine and the Restricted Boltzmann machine (RBM)[13] using Gibbs sampling, which made the neural network a sensation. But at the time it was theoretically impossible to extend the perceptron model from a simple three-layer to multiple layers. Therefore, from the late 1980s to the beginning of the 21st century, other machine learning such as RF and SVM algorithms flourished until the DBN[2] stack-restricted Boltzmann neural network proposed in 2005 was far superior to other machine learning algorithms. Once again, the research on neural networks has dominate the AI direction, and various deep networks such as CNN [15] and RNN [16] have emerged one after another.

3.2. Attack

3.2.1. Steps

① Preprocessing of The Data Set

In the traditional profiling SCA, the data preprocessing step is actually a dimension reduction in the narrow sense, including the machine learning models. It is precisely because the feature engineering of
the data is so complicated that the preprocess can not be omitted. For profiling SCA based on deep learning, it can be left out the preprocessing step of dimension reduction (unless the dimension is too high), and directly preprocess the data set. It is divided into two parts: the training data set and the verification data set. The training data set is used to train the model and the verification data set is used to measure the indicator.

② Choosing Models and Training

Model training is performed by using different deep neural networks. There are some models in usual: MLP, CNN, and RNN. They will be introduced separately at the following part.

③ Verification and Evaluate

Several evaluation frameworks are often used to evaluate the performance of the operation or to select the optimal parameters for the parametric model family. The purpose of these methods is in order to provide an estimate of the performance of a metric (precision) that is independent of the choice between the training set D_train and the test set D_test, but for the scale.

④ Matching and Recovering(Same as the chapter 2)

3.2.2. Models

① MLP

The multiple layers perceptron is also called artificial neural network. In addition to the input and output layers, there can be multiple hidden layers in between. The simplest MLP contains only one hidden layer, that is, the three-layer structure, as shown in Figure 2 below:

![Figure1 MLP](image)

As can be seen from the figure above, the MLP’s layer is fully connected to the layer (full connection means that every neuron in the upper layer is connected to all neurons in the next layer). The bottom layer of the multiple layers perceptron is the input layer, the middle is the hidden layer, and finally the output layer.

About the input layer, for example, if the input is an n-dimensional vector, there are n neurons. But how do the neurons in the hidden layer come from? First, it is fully connected to the input layer. If the input layer is represented by the vector X, the output of the hidden layer is $f(w_1x+b_1)$, $w_1$ is the weight, $b_1$ is the bias, function $f$ is usually appear as sigmoid or tanh. Finally, the output layer, in fact, the hidden layer to the output layer can be seen as a multiple category logistic regression, that is softmax regression. So the output of the output layer is softmax $(w_1x+b_1)$, $x_1$ represents the output of the hidden layer $f(w_1x+b_1)$.

Therefore, all the parameters of the MLP are the connection weights and offsets between the layers. To set the best parameters is an optimization problem. To solve the optimization problem, the simplest is the gradient descent method: first randomly initialize all parameters, then iteratively training, continuously calculate the gradient and update the parameters until a certain condition is met. (For example, when the error is small enough and the number of iterations is enough). This process involves cost functions, regularization, learning rate, gradient calculations, etc.

② DAE

An auto encoder[17] is a type of neural network that can be thought of as consisting of two parts: an encoder function $h = f(x)$ and a reconstructed decoder $r = g(h)$. Traditionally, auto encoders have been used for dimensionality reduction or feature learning.
Auto-encoding neural networks try to learn a function $h_{W,d}(x) \approx x$. In other words, it tries to approximate an identity function so that the output $\hat{x}$ can get close to the input $x$. Although the identity function does not seem to be meaningful, when adding some restrictions to the auto-encoding neural network, for example limiting the number of hidden neurons. It can be found that some interesting structures are from the input data.

In 2006, Hinton improved the structure of the prototype auto encoder, and then produced the DAE. Firstly, the unsupervised layer-by-layer greedy training algorithm was used to complete the pre-training of the hidden layer, and then the BP algorithm was used to optimize the system parameters of the whole neural network, significantly reducing the performance index of the neural network, effectively improving the BP algorithm is easy to fall into the local minimum.

Simply put, DAE has increased depth compared to the original Auto Encoder, improving learning ability and facilitating pre-training. As shown in Figure 2, a 5-layer DAE, the number of hidden layer nodes from high to low, then low to high, and finally only need to obtain L3 the vector.

![Figure 2 DAE](image)

3 CNN

A convolutional neural network is a multiple layered neural network, each layer consisting of multiple two-dimensional planes, and each of these planes consists of multiple independent neurons.

The input image is convoluted by three trainable filters and an addable bias. The filtering process is shown in Figure 3. After convolution, three feature maps are generated in the C1 layer, and then four pixels in each group in the feature map. Then summary, weight, and offset, and obtain a feature map of three S2 layers through a sigmoid function. These maps are then filtered to obtain the C3 layer. This hierarchical structure produces S4 again like S2. Finally, these pixel values are rasterized and concatenated into a vector input to a traditional neural network to get the output.

Generally, the C layer is a feature extraction layer. The input of each neuron is connected with the local receptive field of the previous layer, and the local feature is extracted. Once the local feature is extracted, the positional relationship between it and other features will also been determined. The S layer is the feature mapping layer, and each computing layer of the network is composed of multiple feature maps, and each feature is mapped to a plane, and the weights of all neurons on the plane are equal. The feature mapping structure uses a small sigmoid function that affects the function kernel as the activation function of the convolutional network so that the feature map has displacement invariance.
In addition, since the neurons on one mapping surface share the weight, the number of free parameters of the network is reduced. Also the complexity of network parameter selection is reduced. Each feature extraction layer (C-layer) in the convolutional neural network is followed by a computational layer (S-layer) for local averaging and secondary extraction. This unique feature extraction structure makes the network the input sample which has a high distortion tolerance when recognized.

④ RNN

RNN is a type of neural network used to process sequence data. After studying from the basic neural network that the neural network consists of an input layer, a hidden layer and an output layer. The output is controlled by an activation function, and the layers are connected by weights. The activation function is determined in advance, and the neural network model is included in the weight by training.

The underlying neural network only establishes a weighted connection between the layers. The biggest difference between RNN is the right connection of the neurons between layers. As the Figure 4 shows below.

This is a standard RNN structure diagram where each arrow represents a transformation, that is, the arrow connection has the weights. The left side is folded up, the right side is the unfolded appearance, and the arrow next to $h$ in the left side represents the "loop" in this structure reflected in the hidden layer.

In the unfolded structure which can be observed that in the standard RNN structure, the neurons of the hidden layer are also weighted. That is to say, as the sequence progresses, the previous hidden layer will affect the hidden layer behind. In the figure, $o$ represents the output, $y$ represents the determined value given by the sample, and $L$ represents the loss function. It shows that the loss is also accumulated as the sequence is recommended.

In addition to the features above, the standard RNN has the following features:
1. The weights are shared. The $w$ in the figure are all the same, and $U$ and $V$ are the same.
2. Each input value is only connected to its own route and will not be connected to other neurons.

3.3. Research Situation

According to papers which can be collected recently, the method of profiling SCA based on deep neural network has been studied in few articles in recent years. In 2013, the first to use the neural network combined with the power analysis in SCA was the three-layer simple perceptron networks used by Martinasek[18] and the it achieved a classification accuracy up to 90%. Gilmore and Hanley[19] proposed a neural network based side channel attack trying to break the band-masked AES implementation of DPA competition V4[20].

In 2016, Maghrebi[21] and others of Safran Group used the classic template attack method, machine learning method and deep learning technology method to perform a key cracking on the DPA-contest competition data respectively. The experimental results show that the template attack effect based on deep neural network is better than others and have a higher success rate. In the DL template attack, the feature extraction method adopted by Maghrabi is more efficient than the AE learning classification.

Martinasek[22][23][24] and his colleagues compared MLP-based methods with other classical methods such as template attacks or random attacks. They conducted experiments based on the AES-128 encryption algorithm and gave the MLP hyperparameters and some information about the training. Research shows that MLP technology from deep learning theory is an effective alternative to traditional template analysis, and it is better than previous SVM and random forest technologies. However, these attacks are more parameterized and do not give the precise information about the algorithm parameterization and training. This limitation restricts the combination of deep learning methods in side channel.

Eleonora Cagli[25] et al. succeeded in proposing a CNN-based end-to-end attack method, which is very effective for unaligned trajectory attacks, and uses two algorithms for data enhancement techniques to avoid learning deficits and over-fitting in the learning process. However, Cagli et al. did not give classification model performance.

Houssem Maghrebi et al.[26] compared the network structures in MLP, RF, CNN and RNN with traditional TA and classified them into protected and unprotected AES attacks. They found that the LSTM performed well in cracking the masked AES encryption algorithm based on Chipwhisper[27], and the unprotected AES algorithm’s effect achieved by FPGA was worse than CNN and MLP. When cracking AES with protection, DL is generally better than RF and MLP. It proves that when software leakages are closely related to the time and which can achieve higher signal-to-noise ratios, adversary can prioritize the use of RNN. This also provides a model for our research from CNN to RNN.

Ioannis Petros Samiotis[28] optimized the self-named SCA-net for the CNN structure used in DPA contest v2 and v4 by analyzing the shortcomings of Cagli and Maghrebi’s papers. The SCA-Net mentioned consists of 4 convolutional layers and 4 intermediate pool layers, followed by a classification layer. All convolutional kernels have a 6 size kernel of , and each layer contains multiple activation functions. In the experiment, SCA-net and Maghrabi et al CNN architecture were compared to prove that the optimized SCA-net effect is better.

In order to construct a general framework to study and compare the effectiveness of machine learning methods and embedded cryptographic algorithms. Ryad Benadjila and Emmanuel Prouff[29] compared the learning effects of VGG-16 and MLP on modeling and found that VGG-16 performed better, especially for encryption devices with additional cover. And the two provided a benchmark for hyper parameters and established a neural network-based side channel attack ASCAD database, in which it established a new methodological basis for machine learning in the field of side channel attacks. Later, the model in ASCAD given by Benadjila et al. was improved, and a layer of diffusion layer neural network was added, which improved the success rate of first-order and second-order attacks.

Benjamin Timon[30] also proposed improvements to traditional non-modeling side channel attacks based on the above three applications of modeling DL in SCA. It uses the MLP and CNN methods to
compare the CPA and the high-order CPA in the case of a protective attack. The DL has achieved a good result.

3.4. Future of Deep Learning SCA

For the current deep learning network, mainly CNN, its popular models are ResNet[31] and inception[32][33][34] family. Starting with ResNet, neural networks have surpassed humans for the first time in classification tasks. After ResNet, the deconstruction design of the neural network turned to block, and the network was no longer redesigned as a whole. This CNN design demonstrates excellent generalization capabilities.

The Inception network is an important milestone in the history of the CNN classifier. Before the advent of Inception, most popular CNNs simply stacked the convolutional layers more and more, making the network deeper and deeper, in the hope of getting better performance. For example, the first AlexNet[32] to receive wide attention, it is essentially to extend the depth of LeNet[32], and apply some techniques such as ReLU, Dropout and so on. AlexNet has 5 convolutional layers and 3 maximum pooling layers, which can be divided into two identical branches, which can exchange information on the third convolutional layer and the fully connected layer. The best network proposed in the same year as Inception is VGG-Net[32], which has a smaller convolution kernel and a deeper level than AlexNet. The generalization performance of VGG-Net is very good, and it is often used for the generation of candidate frames for extraction of image features. The biggest problem with VGG is the large number of parameters. VGG-19 is basically the convolutional network architecture with the most parameters. This issue is also the first focus of GoogLeNet[32], which hopes to propose the Inception architecture. It does not use a fully connected network as much as VGG-Net, so the amount of parameters is very small. The biggest feature of GoogLeNet is the use of the Inception module, which aims to design a network with excellent local topology, that is, perform multiple convolution operations or pooling operations on the input image in parallel, and splicing all the output results into one very deep feature map. Because different convolution operations and pooling operations such as 1×1, 3×3, or 5×5 can obtain different information about the input image, processing these operations in parallel and combining all the results will result in better image characterization.

On the other hand, the Inception network is complex (requires a lot of engineering work). It uses a lot of tricks to improve performance, both the speed and accuracy. Its constant evolution has led to the emergence of a variety of Inception network versions. Common versions are: Inception v1[32], Inception v2, Inception v3[33], Inception v4[34], and Inception-ResNet. Each version is an iterative evolution of the previous version. With the upgrade of the Inception network, the speed and accuracy of classification optimization has also increased, and our side channel attacks provide new methods and directions.

4. Summary

As technology advances, side channel analysis has officially entered the era of intelligence. Deep learning, as a representative of artificial intelligence processing big data, in the field of side channels analysis, its excellent processing of complex high-dimensional data and simplification of feature engineering make it surpass the most traditional template side channel attack and TA based machine learning models. But there are also some following deficiencies:

1. Few deep learning model. Currently only AE, MLP, CNN and RNN are used.
2. Insufficient research on encryption algorithms with mask protection. The deep learning model theoretically has the structural advantage of dealing with the protected encryption algorithm, which can be developed in this respect.
3. Few study on non-profiling side channel analysis. At present, there are few related papers, which is a new direction that can be concerned.

Acknowledgement:
I would like to appreciate the Pro.Chen and the Dr.Zhang for recommendations and the fund support
as well as anonymous reviewers for their very useful and valuable advice.

References:
[1] P. C. Kocher, J. Jaffe, and B. Jun, “Differential Power Analysis,” in CRYPTO, ser. Lecture Notes in Computer Science, M. J. Wiener, Ed. Springer (1999).
[2] Bengio, Lamblin, Popovici and Larochelle, “Greedy Layer-Wise Training of Deep Networks”, NIPS’(2005).
[3] S. Chari, J. R. Rao, and P. Rohatgi, “Template Attacks,” in CHES, ser. Lecture Notes in Computer Science, B. S. K. Jr., C etin Kaya Koc, and C. Paar, Eds. Springer (2002).
[4] W. Schindler, K. Lemke, and C. Paar, “A Stochastic Model for Differential Side Channel Cryptanalysis,” in CHES, ser. Lecture Notes in Computer Science, J. R. Rao and B. Sunar, Eds. Springer (2005).
[5] E. Brier, C. Clavier, and F. Olivier, “Correlation Power Analysis with a Leakage Model,” in CHES, ser. Lecture Notes in Computer Science, M. Joye and J.-J. Quisquater, Eds. Springer (2004).
[6] B. Gierlichs, L. Batina, P. Tuyls, and B. Preneel, “Mutual Information Analysis,” in CHES, ser. Lecture Notes in Computer Science, E. Oswald and P. Rohatgi, Eds., vol. 5154. Springer (2008), pp. 426–442.
[7] F. Rosenblatt, “The perceptron: A probabilistic model for information storage and organization in the brain.” Psychological Review, (1958), 386-408.
[8] B. Widrow, H. E. Hoff. “Adaptive switching circuits”. 1960 IRE Wescon Conv. Record, 4, 96–104, Aug (1960).
[9] D. Rumelhart, J. McClelland, An interactive activation model of context effects in letter perception: II. The contextual enhancement effect and some tests and extensions of the model. Psychological Review, 1,(1982) 60-94.
[10] P. J. Werbos. Applications of advances in nonlinear sensitivity analysis. In Proceedings of the 10th IFIP Conference, 31.8 - 4.9, NYC, pages 762–770 (1981).
[11] J. J. Hopfield “Neural networks and physical systems with emergent collective computational abilities” PNAS April 1, (1982 ) 8,2554-2558.
[12] E. Geoffre y, H. Terrence, J. Sejnowski “A learning algorithm for boltzmann machines” Cognitive Science 9, Issue 1, January–March (1985) Pages 147-169.
[13] E. Geoffrey, H. Terrence “Training Products of Experts by Minimizing Contrastive Divergence” Neural Computation 14, Issue 8, August (2002) p.1771-1800.
[14] Y. LeCun and Y. Bengio. “The handbook of brain theory and neural networks.chapter Convolutional Networks for Images”, Speech, and Time Series, pages 255–258. MIT Press, Cambridge, MA, USA, 1998.
[15] K. O’Shea and R. Nash. “An introduction to convolutional neural networks.” CoRR,(2015).
[16] M. Hermans and B. Schrauwen. “Training and analysing deep recurrent neural networks.” Advances in Neural Information Processing Systems 26, pages 190-198. Curran Associates, Inc.(2013).
[17] Vincent, Pascal, et al. "Stacked Denoising Autoencoders: Learning Useful Representations in a Deep Network with a Local Denoising Criterion." Journal of Machine Learning Research 12(2010):3371-3408.
[18] Martinasek, Z. Zeman, V.: Innovative method of the power analysis. Radioengineering 22(2), 586–594 (2013)
[19] R. Gilmore, N. Hanley, and M. O’Neill. Neural network based attack on a masked implementation of AES.In Hardware Oriented Security and Trust (HOST), 2015 IEEE International Symposium on, pages 106–111, May (2015).
[20] DPA Contest V4. http://www.dpacontest.org/home/ (2014). Accessed 1 Feb 2014.
[21] H. Maghrabi, T. Portigliatti , E. Prouff “Breaking Cryptographic Implementations Using Deep Learning Techniques in Security, Privacy, and Applied Cryptography Engineering”. Springer International Publishing (2016).
[22] Z. Martinasek, P. Dzurenda, and L. Malina. “Profiling power analysis attack based on MLP in DPA contest V4.2.” In 39th International Conference on TSP 2016, Vienna, Austria, pages 223–226. IEEE(2016).

[23] Z. Martinasek, J. Hajny, and L. Malina. “Optimization of power analysis using neural network.” CARDIS (2013): Smart Card Research and Advanced Applications pp 94-107.

[24] Z. Martinasek, L. Malina, and K. Trasy. “Profiling Power Analysis Attack Based on Multi-layer Perceptron Network.” Computational Problems in Science and Engineering, 343(2015).

[25] E. Cagli, C. Dumas and E. Prouff “Convolutional Neural Networks with Data Augmentation against Jitter-Based Countermeasures” in CHES(2017) pp 45-68 Springer.

[26] H. Maghrebi, T. Portigliatti, E. Prouff “Breaking Cryptographic Implementations Using Deep Learning Techniques” in SPACE(2016) pp 3-26.

[27] C. O’Flynn “A Framework For Embedded Hardware Security Analysis” Submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy at Dalhousie University Halifax, Nova Scotia June (2017).

[28] I. P. Samiotis “Side Channel Attacks using Convolutional Neural Networks” at the Delft University of Technology.

[29] R. Benadjila1, E. Prouff, R. Strullu, Eleonora Cagli and Cécile Dumas “Study of Deep Learning Techniques for Side-Channel Analysis and Introduction to ASCAD Database” in ANSSI, France.

[30] B. Timon “Non-Profiled Deep Learning-Based Side-Channel Attacks” UL Transaction Security, Singapore.

[31] K. He, X. Zhang, S. Ren Jian Sun “Deep Residual Learning for Image Recognition” in The IEEE Conference on CVPR(2016) pp. 770-778

[32] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, Andrew Rabinovich; “Going Deeper With Convolutions” JThe IEEE. CVPR(2015) pp. 1-9.

[33] C. Szegedy, V. Vanhoucke, S. Ioffe, Jon Shlens, Zbigniew Wojna; “Rethinking the Inception Architecture for Computer Vision” The IEEE.CVPR (2016), pp. 2818-2826.

[34] C. Szegedy, S. Ioffe, V. Vanhoucke “Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning” The IEEE .CVPR(2016).