SCIENCE CHINA
Information Sciences

• LETTER •

February 2020, Vol. 63 129401:1–129401:3
https://doi.org/10.1007/s11432-019-9803-8

Golden chip free Trojan detection leveraging probabilistic neural network with genetic algorithm applied in the training phase

Yanjiang LIU, Jiaji HE, Haocheng MA & Yiqiang ZHAO*

School of Microelectronics, Tianjin University, Tianjin 300072*, China*

Received 6 January 2019/Revised 20 February 2019/Accepted 22 February 2019/Published online 6 August 2019

Citation Liu Y J, He J J, Ma H C, et al. Golden chip free Trojan detection leveraging probabilistic neural network with genetic algorithm applied in the training phase. Sci China Inf Sci, 2020, 63(2): 129401, https://doi.org/10.1007/s11432-019-9803-8

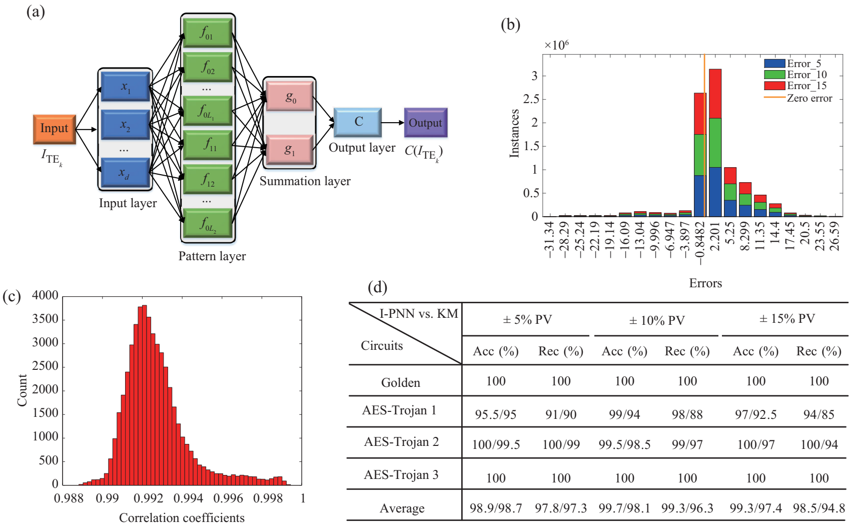
Dear editor,
Hardware Trojan has become a serious threat to the trustworthiness of critical applications [1]. Among all existing detection approaches, golden chip-free Trojan detection approaches have attracted wide attention in [2]. Recently, various supervised learning techniques are applied to the golden chip-free Trojan detection approaches [3]. However, the real case is that little knowledge of the Trojan under detection can be obtained, thus, the side-channel traces required in the training sets do not exist in practice.

To address this issue, a golden chip-free Trojan detection approach with the combination of genetic algorithm (GA) and probabilistic neural network (PNN) is proposed in this study. The simulated voltage variations at different process corners are regarded as the golden reference and a compensation algorithm is exploited to make the simulated data matches well with the measured data. Further, a probabilistic neural network with the minimum Bayesian risk criterion is exploited to detect the hardware Trojan, and genetic algorithm is introduced into the training phase of PNN to establish an intelligent classifier (I-PNN). Finally, the established intelligent classifier is exploited to identify Trojan chips without any knowledge of Trojan during detection.

Golden chip-free Trojan detection framework. The golden chip-free Trojan detection approach is composed of three stages: (1) pre-silicon simulation; (2) post-silicon measurement; and (3) intelligent Trojan detection. In the pre-silicon simulation stage, the spice netlist, parasitic parameters, spice model and stimuli are feed into the power simulator and the simulated traces (denoted as $\boldsymbol{I}_{\mathrm{S}}$) of golden model are obtained. In the post-silicon measurement stage, the measured traces of fabricated chips are acquired using the experimental setup, and a denoising algorithm based on the Gaussian filter is exploited to eliminate the influences of random noise. To establish an accurate golden model, several golden chips are required to compensate the model using the radial basis function neural network, and the compensated traces (denoted as $\boldsymbol{I}_{\mathrm{M}}$) are obtained by compensating the $\boldsymbol{I}_{\mathrm{S}}$ with the $\boldsymbol{I}_{\mathrm{DG}}$. Where the denoised traces of golden chips and fabricated chips under test are denoted as $\boldsymbol{I}_{\mathrm{DG}}$ and $\boldsymbol{I}_{\mathrm{DC}}$, respectively. In the intelligent Trojan detection stage, the $\boldsymbol{I}_{\mathrm{M}}$ and $\boldsymbol{I}_{\mathrm{DC}}$ are utilized to train the I-PNN, and the chip is classified as golden chip or Trojan chip.

Intelligent classifier establishment methodology. The majority of existing Trojan detection approaches focus on the similarity of data without considering the cost of false negative. Given the catastrophic consequences incurred by Trojan, various supervised learning algorithms are applied

* Corresponding author (email: yq_zhao@tju.edu.cn)



I-PNN vs. KM / Circuits	± 5% PV		± 10% PV		± 15% PV	
	Acc (%)	Rec (%)	Acc (%)	Rec (%)	Acc (%)	Rec (%)
Golden	100	100	100	100	100	100
AES-Trojan 1	95.5/95	91/90	99/94	98/88	97/92.5	94/85
AES-Trojan 2	100/99.5	100/99	99.5/98.5	99/97	100/97	100/94
AES-Trojan 3	100	100	100	100	100	100
Average	98.9/98.7	97.8/97.3	99.7/98.1	99.3/96.3	99.3/97.4	98.5/94.8

Figure 1 (Color online) (a) Traditional structure of PNN; (b) comparison of the compensation results; (c) histogram distribution of correlation coefficients; (d) comparison of classification results between I-PNN and KM.

to the Trojan detection, which use the minimum error law or minimum risk law as the decision rule to identify the presence of hardware Trojan [4]. As a feed-forward neural network, PNN is proposed in the early 1988s, which is widely applied to the fields of image classification, earthquake prediction, medical diagnosis, etc. Unlike the other traditional classifier, like K-means clustering (denoted as KM) and self-organizing maps, PNN provides a high classification accuracy and good fault tolerance because it uses the minimum Bayesian risk criterion and the Parzen window for pattern classification [5]. Therefore, PNN with the minimum Bayesian risk criterion is applied to identify the Trojan chips in this study.

Nevertheless, all supervised learning algorithms require a large number of labeled data for all the target classes to train a model. We have little knowledge of the implementation of Trojan during detection, and thus the labeled data of Trojan chips in training sets are not available actually. Genetic algorithm is an adaptive heuristic searching algorithm based on the Darwin's evolution theory, which mimics the process of natural evolution in the biological world to solve optimization problems. In GA, the quality of a candidate solution is evaluated by the fitness function, and improved by the selection operator, crossover operator and mutation operator iteratively until the fitness value reaches the maximum [6]. To determine the classes of chips under test in training sets (denoted as $\boldsymbol{C}_{\mathrm{a\_tr}}$), GA is introduced into the training phase of PNN. In each iteration of GA, the validation results are evaluated by the fitness function, and the value of $\boldsymbol{C}_{\mathrm{a\_tr}}$ is adaptively searched in the whole solution space. When the fitness value reaches the maximum, the current $\boldsymbol{C}_{\mathrm{a\_tr}}$ is determined as the best solution of GA and used to train the PNN.

The procedure of the intelligent classifier establishment methodology is given as follows. Before training the classifier, 70% samples are used for training, 15% samples are used in validation, and 15% samples are used to test. The label of all the sample of $\boldsymbol{I}_{\mathrm{M}}$ is assigned 1, and the label of all the sample of $\boldsymbol{I}_{\mathrm{DC}}$ is assigned 1 or 0 randomly. Then the PNN is trained with the training sets $\boldsymbol{I}_{\mathrm{TA}}$ and validated with the validation sets $\boldsymbol{I}_{\mathrm{VA}}$. If the label of all the sample of $\boldsymbol{I}_{\mathrm{M}}$ in $\boldsymbol{I}_{\mathrm{VA}}$ equals to 1, the classifier is established, otherwise, GA is used to adjust the value of $\boldsymbol{C}_{\mathrm{a\_tr}}$, and then the training and validation process are performed iteratively. Finally, the trained classifier is used to infer the label matrix of validation sets, and the fabricated chip is classified as Trojan chip when the corresponding element equals to 0.

Trojan detection based intelligent classifier. The k-th sample of $\boldsymbol{I}_{\mathrm{TE}}$, which is $\boldsymbol{I}_{\mathrm{TE}_k}$, is used to explain the Trojan identification using the established I-PNN. The traditional structure of PNN is shown in Figure 1(a). The input layer with d neurons receives the d-dimensional feature vectors from input data set $\boldsymbol{I}_{\mathrm{TE}_k}=\{I_{\mathrm{TE}_{k,1}}, I_{\mathrm{TE}_{k,2}}, \ldots, I_{\mathrm{TE}_{k,d}}\}^{\mathrm{T}}$. The pattern layer calculates the proximity $f_{ij}(\boldsymbol{I}_{\mathrm{TE}_k})$ between $\boldsymbol{I}_{\mathrm{TE}_k}$ and $\boldsymbol{I}_{\mathrm{TA}}$ through a defined kernel. A standard Gaussian function is used as the probability distribution function in this study, which is described as

$$f_{ij}(\boldsymbol{I}_{\mathrm{TE}_k}) = \frac{1}{(2\pi)^{n/2}\sigma^n}\exp\left[-\frac{\boldsymbol{I}_{\mathrm{D}_{i,j}}^{\mathrm{T}}\boldsymbol{I}_{\mathrm{D}_{i,j}}}{2\sigma^2}\right], \quad (1)$$

where σ is the spread parameter of the Gaussian function, n is the number of neurons of pattern layer, $\boldsymbol{I}_{\mathrm{TA}_{i,j}}$ is the j-th sample of $\boldsymbol{I}_{\mathrm{TA}}$ that belongs to the i-th class, and $\boldsymbol{I}_{\mathrm{D}_{i,j}} = \boldsymbol{I}_{\mathrm{TE}_k} - \boldsymbol{I}_{\mathrm{TA}_{i,j}}$. The $\boldsymbol{I}_{\mathrm{TE}_k}$ can be classified as two classes: 1 (Golden class) and 0 (Trojan class). Therefore, the value of i is equal to 1 or 0.

The summation layer computes the mean value of probability $g_i(\boldsymbol{I}_{\mathrm{TE}_k})$ of $\boldsymbol{I}_{\mathrm{TE}_k}$ that belongs to the i-th class. Where L_i is the number of $\boldsymbol{I}_{\mathrm{TA}}$ belonging to the i-th class and $L_0+L_1=n$.

$$g_i(\boldsymbol{I}_{\mathrm{TE}_k}) = \frac{1}{L_i}\sum_{j=1}^{L_i} f_{ij}(\boldsymbol{I}_{\mathrm{TE}_k}). \tag{2}$$

Finally, the output layer selects the maximum probability $g_i(\boldsymbol{I}_{\mathrm{TE}_k})$ corresponding with the neuron of summation layer and determines the class $C(\boldsymbol{I}_{\mathrm{TE}_k})$ of $\boldsymbol{I}_{\mathrm{TE}_k}$. The $C(\boldsymbol{I}_{\mathrm{TE}_k})$ is given by

$$C(\boldsymbol{I}_{\mathrm{TE}_k}) = \arg \max_{i\in\{0,1\}} \{g_i(\boldsymbol{I}_{\mathrm{TE}_k})\}. \tag{3}$$

Results and discussion. A 128-bit advanced encryption system (denoted as AES) is adopted as the golden circuit, and 4-bit (denoted as AES-Trojan1), 8-bit (denoted as AES-Trojan2) and 12-bit (denoted as AES-Trojan3) counters are applied as Trojan, which occupy 0.36%, 0.72% and 1.09% of the size of golden circuit, respectively. The chips are fabricated in Chartered 180 nm technology, and both the inter-die and intra-die variations are set to ±5%, ±10% and ±15% GAUSS variations.

Figure 1(b) shows the differences of compensation results. The differences between the $\boldsymbol{I}_{\mathrm{DG}}$ and $\boldsymbol{I}_{\mathrm{M}}$ under ±5%, ±10% and ±15% PV, respectively are denoted as Error_5, Error_10 and Error_15. More specifically, the maximum difference is only 29.65 mV and more than 90% differences in magnitude fall 10.86 mV. This kind of differences mainly caused by the process variations, and can be omitted during the Trojan detection process.

To further demonstrate the whole similarity between the $\boldsymbol{I}_{\mathrm{M}}$ and $\boldsymbol{I}_{\mathrm{DG}}$, the correlation analysis is performed, and the distribution histogram of correlation coefficients under ±5% PV is shown in Figure 1(c). It is clear from Figure 1(c), the correlation coefficients are greater than 98.85%. Similarly, the correlation coefficients under ±10% and ±15% PV are greater than 98.81%. On the whole, the calibrated golden model matches well with the actual silicon measurements. Therefore, the simulated data of calibrated golden model can substitute for the measured data of fabricated chips even in the presences of process variations and random noise.

Figure 1(d) presents the comparison of classification results. The ratio of chips correctly classified by the algorithm is denoted as Acc, and the proportion of Trojan chips correctly classified by algorithm is denoted as Rec. As shown in Figure 1(d), the I-PNN and KM identifies the Trojan under different process variations, and the Acc and Rec of I-PNN are greater than the KM. Due to the randomness over the training phase, the best solution of GA is not unique, thus, the network structure of trained PNN is not affected with the effect of process variations. For AES-Trojan2, we assume that the I-PNN under ±10% PV is more accurate than the ±10% PV, therefore, the accuracy of AES-Trojan2 under ±10% PV is smaller than the ±15% PV. The classification results are not reduced with the increasing of process variations and thus the influences of process variations can be omitted. On the whole, the proposed approach is capable of detecting Trojan chips without any knowledge of Trojan.

Conclusion. In this study, a golden chip-free Trojan detection framework with the intelligent classifier is proposed, which breaks the limitation of side-channel analysis and accelerates the practical application of golden chip-free Trojan detection approaches. The compensated simulation data of voltage variations are regarded as the golden reference, and the intelligent classifier is established to identify the hardware Trojan without any knowledge of Trojan during detection. While the proposed approach is success, the proposed method requires several golden chips to make the golden model matches well with the actual silicon measurements. Moreover, the proposed method should be combined with the test generation techniques for reliable detection of Trojans of all sizes.

Acknowledgements This work was financially supported by National Natural Science Foundation of China (Grant No. 61832018).

References

1 Zhang X J, Zhang F, Guo S Z, et al. Optimal model search for hardware-trojan-based bit-level fault attacks on block ciphers. Sci China Inf Sci, 2018, 61: 039106

2 He J J, Zhao Y Q, Guo X L, et al. Hardware Trojan detection through chip-free electromagnetic side-channel statistical analysis. IEEE Trans VLSI Syst, 2017, 25: 2939–2948

3 Elnaggar R, Chakrabarty K. Machine learning for hardware security: opportunities and risks. J Electron Test, 2018, 34: 183–201

4 Xue M F, Wang J, Hu A Q. An enhanced classification-based golden chips-free hardware Trojan detection technique. In: Proceedings of IEEE Asian Hardware-Oriented Security and Trust, Yilan, 2016

5 Ahmadipour M, Hizam H, Othman M L, et al. Islanding detection method using ridgelet probabilistic neural network in distributed generation. Neurocomputing, 2019, 329: 188–209

6 Nourian M A, Fazeli M, Hely D. Hardware Trojan detection using an advised genetic algorithm based logic testing. J Electron Test, 2018, 34: 461–470