文章编号:1007-5321(2020)05-0118-07 **DOI**:10.13190/j.jbupt.2020-034

基于深度学习的类 SM4 算法 S 盒逆向分析

马向亮[1,2], 李 冰[3], 杨 丹[3], 黄克振[1,2], 段晓毅[4]
(1. 中国科学院软件研究所 可信计算与信息保障实验室, 北京 100190; 2. 中国科学院大学, 北京 100049;
3. 国家信息技术安全研究中心, 北京 100084; 4. 北京电子科技学院 电子信息工程系, 北京 100070)

摘要: 在建模类攻击场景下,基于多元高斯分布的模板攻击是常用的侧信道逆向分析方法. 在同样的场景下,分析了深度学习方法在逆向分析领域的应用,提出了基于深度学习的 S 盒逆向分析算法. 通过选取适用于侧信道逆向分析的深度学习算法、损失函数和标签设计,对类 SM4 算法进行了 S 盒逆向恢复实验. 实验结果表明,使用深度学习进行 S 盒逆向分析是可行的,且在一定的条件下优于模板攻击;另外,多层感知机算法预测的结果要优于卷积神经网络算法预测的结果.

关 键 词: 深度学习; 逆向分析; S 盒; 类 SM4 算法
中图分类号: TP309.1 **文献标志码**: A

Reverse-Analysis of S-Box for SM4-Like Algorithms Based on Side Channel Technology

MA Xiang-liang[1,2], LI Bing[3], YANG Dan[3], HUANG Ke-zhen[1,2], DUAN Xiao-yi[4]
(1. Trusted Computing and Information Assurance Laboratory, Institute of Software of Chinese Academy of Sciences, Beijing 100190, China;
2. University of Chinese Academy of Sciences, Beijing 100049, China;
3. National Research Center for Information Technology Security, Beijing 100084, China;
4. Department of Electronics and Information Engineering, Beijing Electronic Science and Technology Institute, Beijing 100070, China)

Abstract: In the profiled scenario, the common method of reverse analysis is the template attack based on multi-Gaussian distribution. The article applies the concept of deep learning to the field of reverse analysis for the first time under the same conditions, and proposes an S-box reverse analysis algorithm based on deep learning. By selecting the deep learning algorithm, loss function and label design method suitable for side channel reverse analysis, an S-box reverse recovery experiment is conducted for SM4-like algorithm. It is shown that it is feasible to employ deep learning method to carry out S-box reverse analysis, which can have better performance comparing to template attack under certain circumstances. Moreover, the predicting effect of multi-layer perception algorithm surpasses that of convolutional neural network algorithm.

Key words: deep learning; reverse analysis; S box; SM4-like algorithm

虽然密码算法的安全性分析普遍基于一个安全 假设,即密码算法本身是公开的,而密钥是保密的.

收稿日期: 2020-03-31
基金项目: 国家重点研发计划项目(2018YFB0904900, 2018YFB0904901); "十三五"装备预研领域基金项目(6140002020115)
作者简介: 马向亮(1986—), 男, 博士生.
通信作者: 段晓毅(1979—), 男, 讲师, 硕士生导师, E-mail:duanxiaoyi@besti.edu.cn.

但在实际的密码应用中，出于商业保密或其他安全因素的考虑，一些密码系统或模块中的密码算法或部件不是公开的. 这为针对密码系统或模块的安全性评估增添了难度和不确定性，因此，对密码算法进行逆向分析是进行下一步安全性评估的必要前提. 逆向分析的结果和恢复程度将直接影响到密码算法实现安全性测评的力度. 目前，已存在的逆向分析方法主要有数学逆向分析[1]、芯片解剖逆向[2]、基于故障注入技术的逆向分析[3]以及基于侧信道技术的逆向分析[4-9].

在上述逆向分析中，数学逆向分析以逆向数学构造形式为主，数学分析复杂度高且通用性低；解剖逆向需要通过还原电路进行逻辑判断、分析，所以存在周期长、代价高、可行性低等缺点；基于故障技术的逆向分析对故障模型和注入精度有较高要求，因而在实际中应用较少. 相比较而言，基于侧信道分析技术的逆向分析无需对密码实现载体进行破坏或干扰，并且还具有分析复杂度低、通用性高等优点，因此目前是密码实现逆向分析的一种主要技术手段.

随着机器学习技术的发展，深度学习也被用于侧信道分析中，并呈现出攻击复杂度低、准确率高等优点[10]，但目前尚没有出现利用深度学习进行侧信道逆向分析的相关结果. 在建模类场景下，提出了一种基于深度学习的 S 盒逆向分析方法. 以类 SM4 算法[11]为例进行了实验，实验结果表明，基于深度学习的侧信道 S 盒逆向分析是可行的，且在一定的条件下优于模板攻击；另外，多层感知器算法训练的结果要优于卷积神经网络算法训练的结果.

1　深度学习

深度学习的作用主要是进行分类操作，在侧信道的逆向分析领域[12]需要选取适用的深度学习算法、激活函数和损失函数. 另外，需要选取深度学习算法参数中的训练数据和标签设计，以适用于基于建模类场景的 S 盒逆向分析.

1.1　算法和标签模型选取

目前，深度学习算法在侧信道领域已有一些应用，这些应用中都对不同深度学习算法训练的结果进行了比较分析，其中使用多层感知机和卷积神经网络的算法较多，进行的实验效果也较好，也是许多学者推荐使用的算法[10,13].

Timon[10]恢复密钥使用的标签是汉明重量模型或比特模型对应的值，这种模型不适用于 S 盒逆向分析场景. 因为对一个 $m \times n$ 位的 S 盒，可使用汉明重量模型或比特模型，将 S 盒的中间值映射为 $n+1$ 个或 2 个标签分类，所以无法区分 2^n 种 S 盒的输出情况. 在逆向分析中使用的标签是绝对值模型，该标签是通过密钥与特定明文计算的第 1 个 S 盒的输出值，因此，类 SM4 算法 S 盒有 256 个标签. 实验结果表明，该标签的设计可以以很高的准确率逆向恢复出 S 盒的内容.

1.2　准确率与损失值

在深度学习的训练过程中，可以记录每次训练的准确率和损失值. 准确率代表了每次训练输出标签匹配的平均准确率，而损失值代表了神经网络训练过程的优劣. 损失值越小，训练的效果越好. 因此，准确率和损失值都代表了神经网络的训练效果，都可以作为区分器来进行目标对象的恢复.

准确率是输入样本预测结果正确的平均值. 输入 N 条曲线，神经网络训练后会得到 N 个预测结果，其中分类预测结果正确的为 M 个，则准确率就为 M/N.

损失函数也叫代价函数，是训练过程优化的目标. 交叉熵损失函数与 ReLU 函数相结合，不仅使隐含层的神经元学习效率高，而且使输出层的神经元学习效率高. 同时由于使用 ReLU 函数，计算量节省很多，收敛较快；而有效的梯度下降和反向传播解决了梯度消失的问题. 因此，在训练过程中计算损失值使用的损失函数是交叉熵损失函数.

2　基于深度学习的 S 盒逆向分析

侧信道分析分为两类：一类是非建模分析；另一类是建模类分析. S 盒是分组密码算法的关键非线性部件，它决定了密码算法的安全性，也是逆向分析研究的主要对象. 目前基于 S 盒的逆向分析主要基于建模类分析场景. S 盒中共有 256 个不同的数据，智能卡在处理不同数据时，产生的能量消耗也不相同. 如果将智能卡采集的 S 盒能量迹与其对应的 S 盒输出进行深度学习训练，可以得到能量迹的特征信息. 将其他能量迹作为预测数据，可以得到 S 盒的输出值. 因此 S 盒的输出可以作为能量迹的不同分类，即将 S 盒逆向分析转化为能量迹的深度学习分类问题. 这里介绍了基于建模类场景下的 S 盒逆向分析模型和算法，以实际智能卡为例使用多层感知机和卷积神经网络 2 种算法进行了逆向分析实

验. 使用分析模型的输入数据 X 是示波器采集到的S盒的能量消耗值,Y 是由特定明文和密钥计算的S盒输出值,即标签.

在深度学习分析基础上,提出了一种基于深度学习的S盒逆向分析方法,如算法1所示. 用深度学习算法对标签、能量迹和训练迭代次数进行训练,然后使用目标设备的能量迹进行预测,预测S盒的输出.

算法1 基于深度学习的S盒逆向分析

输入: $M \times T$ 特定明文,$M \times N$ 特定明文,训练次数 E

输出: 预测的标签,即S盒

While $X < M$ do
计算特定明文 $P = f(S^{-1}[X], k)$;
采集训练设备 T 条能量迹;
采集目标设备 N 条能量迹;
End while
设置训练数据 $D_T = (T_i)$,其中 $1 \leqslant i \leqslant M \times T$,设置 T_i 对应的 X 为训练标签 X_T;
进行深度学习训练 $\mathrm{acc}_T, \mathrm{loss}_T = \mathrm{DL}(D_T, X_T, E)$;
设置预测数据 $D_N = (T_i)$,其中 $1 \leqslant i \leqslant M \times N$;
进行深度学习预测标签 $X_N = \mathrm{DL}(D_T, X_T, D_N, E)$;
预测的标签 X_N 即S盒.

在算法1中,M 为S盒的输出大小,如类SM4算法的S盒输出长度为0~0xFF. 算法中的第1步为S盒第1个单元穷举,需要第2步选择特定的明文计算得出,其中,X 为S盒输出,k 为参与计算 X 的轮密钥. 第3步需要训练设备加密特定明文 T 次,采集能量迹 T 条. 第4步需要目标设备加密特定明文 N 次,采集能量迹 N 条. 以上步骤完成后,分别得到训练设备的能量迹为 $M \times T$ 条,目标设备的能量迹为 $M \times N$ 条. 第6、7步中的DL表示深度学习多层感知器或卷积神经网络,将 $M \times T$ 条能量迹作为训练数据,对应的S盒输出为标签进行深度学习算法 E 次迭代,得到训练设备的准确率和损失值. 第8、9步将 $M \times N$ 条能量迹作为预测数据,用深度学习进行预测,得到预测标签的准确率和损失值. 第10步,由明文和预测的S盒输出得到等价的S盒内容.

类SM4算法在SM4的基础上,保留算法原有结构,对安全核心非线性部件S盒的内容进行修改,其余的线性结构保持不变. 由算法1,将 M 设置为256,将 N 设置为2 000,将 T 设置为8 000,算法1可以适用于类SM4算法的S盒逆向分析.

3 类SM4算法S盒逆向分析

实验对象为2张32位CPU、采用0.13 μm工艺的智能卡,一张作为训练智能卡,另一张作为目标智能卡进行预测. 基于算法1,共使用256×8 000条能量迹进行训练,使用256×2 000条目标能量迹进行预测. 实验能量消耗采集器是Riscure公司的power tracer. 示波器使用力科610 Zi,采样率为100 MS/s. 电脑通过能量消耗采集器将明文发送至智能卡,智能卡将密文发送回电脑. 在智能卡加密算法过程中,示波器采集智能卡的能量消耗电压值并发送至电脑.

本实验采集的每条能量迹共有805个点. 多层感知机算法使用2个隐藏层的(1 024,512)个神经元,批处理的大小设为1 024,对采集的能量迹中的800点进行50次训练和预测. 每条能量迹预测的结果有256种可能,因此256×2 000条能量迹预测的结果是一个256×256的矩阵. 每个正确标签预测对应的准确率如表1所示. 卷积神经网络算法使用同一种结构和数据,每个正确标签预测对应的准确率如表2所示.

由表1可以看出,由能量迹预测对应的正确标签除标签8a的准确率为40%之外,其余标签预测的准确率均大于等于50%. 标签预测8a的准确率虽然是40%,但是该标签在它对应的2 000条能量迹预测的结果中也是准确率最大的,因此,所有预测准确率最大的标签都是正确的,所有标签都得到有效的恢复. 相同的结果在表2中也成立,虽然有些标签预测的准确率小于50%,但是在它们对应的2 000条能量迹预测的结果中,准确率也是最大的.

将所有标签对应的准确率求均值作为S盒逆向分析的准确率,以此进行类SM4算法S盒的逆向分析. 使用深度学习训练的能量消耗包括了类SM4第1轮加密泄露的800点和第1个S盒加密泄露的500点以及未包括第1个S盒加密泄露的200点.

多层感知机算法使用2个隐藏层的(1 024,512)个神经元,批处理大小设为512. 将256×8 000条能量迹中的800点、500点和200点分别作为训练数据,每条能量迹对应的S盒输出作为标签,使用

表 1　多层感知机算法逆向恢复类 SM4 算法 S 盒对应的准确率　%

	0	1	2	3	4	5	6	7	8	9	a	b	c	d	e	f
0	82	77	66	93	81	90	80	63	82	82	55	74	74	54	78	81
1	69	85	92	82	70	72	84	76	74	77	68	64	71	93	76	72
2	76	92	75	81	83	86	84	66	58	78	70	84	81	82	84	85
3	59	63	50	61	88	81	62	85	86	88	85	64	62	77	64	86
4	84	83	75	65	78	72	85	77	89	58	75	86	88	83	69	55
5	81	88	94	77	86	85	74	77	85	85	73	85	85	90	84	84
6	71	72	67	79	83	67	67	76	79	72	88	76	73	85	79	82
7	82	78	89	88	81	77	76	89	75	76	84	90	83	80	63	91
8	89	91	78	87	79	71	62	94	77	69	81	76	80	71	64	66
9	63	70	48	50	77	83	76	65	77	85	95	83	56	85	84	86
a	88	85	77	64	74	79	76	76	82	88	79	90	87	55	75	77
b	49	82	63	72	70	69	79	65	73	82	83	81	77	91	76	80
c	93	87	91	70	74	89	85	77	82	84	71	71	82	56	89	73
d	70	75	83	72	78	91	65	89	88	80	69	88	86	80	68	87
e	77	81	84	67	71	87	91	73	77	68	82	67	83	81	83	83
f	82	82	91	94	87	82	77	42	55	72	70	65	90	68	86	71

表 2　卷积神经网络算法逆向恢复类 SM4 算法 S 盒对应的准确率　%

	0	1	2	3	4	5	6	7	8	9	a	b	c	d	e	f
0	81	82	80	80	74	66	76	79	76	57	67	74	85	72	84	73
1	65	57	76	77	78	63	79	72	86	72	75	66	73	77	72	80
2	82	64	68	69	77	77	85	75	76	79	63	81	76	64	83	71
3	71	50	66	75	75	75	76	80	84	85	78	72	81	82	80	72
4	87	75	67	74	73	71	85	48	86	75	80	81	72	84	73	68
5	83	71	85	76	82	87	64	79	72	62	83	78	79	73	75	81
6	58	47	71	55	56	76	85	58	78	79	56	72	63	74	63	83
7	65	71	69	52	79	67	79	65	64	83	84	73	70	90	62	66
8	79	65	73	80	60	61	71	74	59	66	75	81	77	69	65	83
9	65	81	72	76	72	68	80	73	75	72	80	53	79	90	66	86
a	84	78	86	71	66	82	70	73	85	68	57	87	73	70	40	68
b	63	86	71	67	72	73	58	79	87	74	69	65	67	81	72	75
c	76	75	67	79	60	79	58	71	86	76	78	44	75	50	51	68
d	71	87	74	83	82	79	78	77	75	69	86	69	80	88	61	83
e	63	74	72	74	82	82	78	69	87	69	74	57	78	68	79	81
f	87	82	82	79	85	87	77	82	70	64	78	85	79	88	69	81

多层感知机算法进行 50 次训练，然后使用 256 × 2 000 条能量迹中的 800 点、500 点和 200 点进行预测，记录 50 次训练的准确率和损失值.

训练结果表明，使用训练设备的准确率略高于使用目标设备进行预测的准确率，损失值上升. 如图 1 所示，选择能量迹上的 800 点，标签训练的准确率在 85% 左右，损失值在 0.4 左右. 预测的正确标签准确率在 80% 左右，损失值在 0.6 左右. 如图 2 所示，选择能量迹上的 500 点，标签训练的准确率在 60% 左右，损失值在 1.2 左右. 使用目标设备的能

量迹进行预测,预测的正确标签准确率略有下降,但也大于50%. 如图3所示,选择能量迹上的200点,标签训练的准确率和损失值大大下降,训练的准确率仅仅在23%左右,损失值在2.6左右. 使用目标设备的能量迹进行预测,准确率大概只有21%,损失值在2.7左右. 由能量迹的其他分析方法可知,能量迹中的前200点未能包括第1轮S盒置换的能量消耗,因此,进行多层感知机训练后效果较差,这也从另一方面说明了该预测确实与S盒置换的能量消耗有很大的关系.

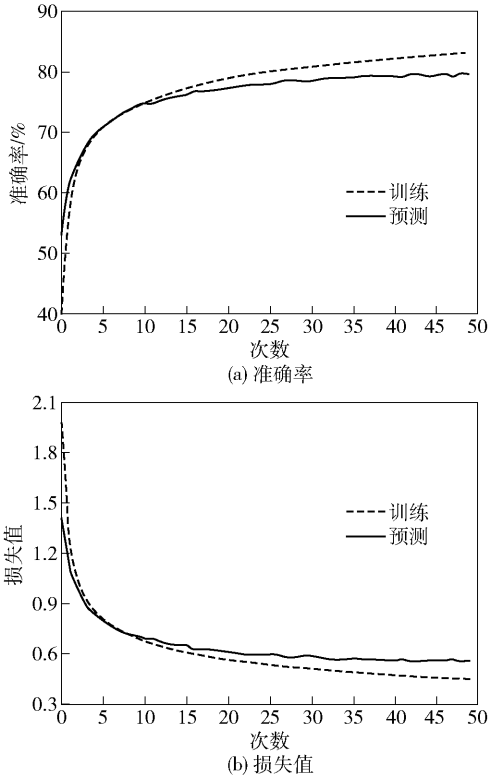

图1 类SM4算法多层感知机800点的训练结果

卷积神经网络算法使用与多层感知机相同的网络结构,对数据和标签进行训练和预测,并记录50次训练和预测的准确率和损失值.

训练结果表明,使用训练设备的准确率同样略高于使用目标设备进行预测的准确率,损失值上升. 如图4所示,选择能量迹上的800点,标签训练后的准确率在73%左右,损失值在0.8左右. 使用目标设备的能量迹进行预测,预测的准确率略有下降. 与多层感知机算法训练的结果相比,卷积神经网络的准确率下降10%,对应的卷积神经网络损失值增大了0.4. 如图5所示,选择能量迹上的500点,使用目标设备的能量迹进行预测标签的准确率在

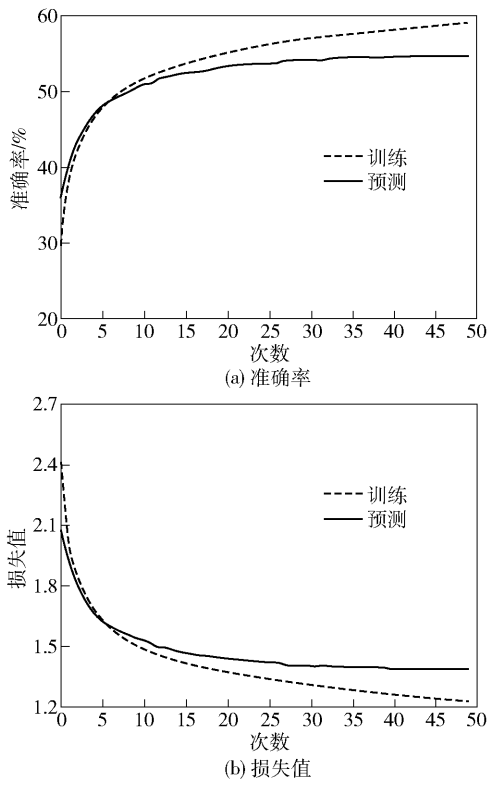

图2 类SM4算法多层感知机500点的训练结果

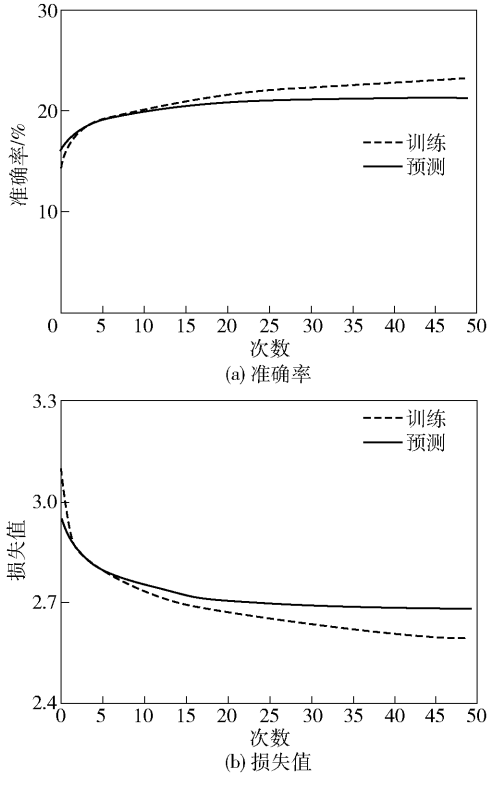

图3 类SM4算法多层感知机200点的训练结果

47%左右,损失值在1.7以上. 如图6所示,选择能量迹上的200点,使用目标设备的能量迹进行预测

标签的仅仅在 20% 左右，损失值在 2.75 以上. 与多层感知机结果相同，符合预期结果.

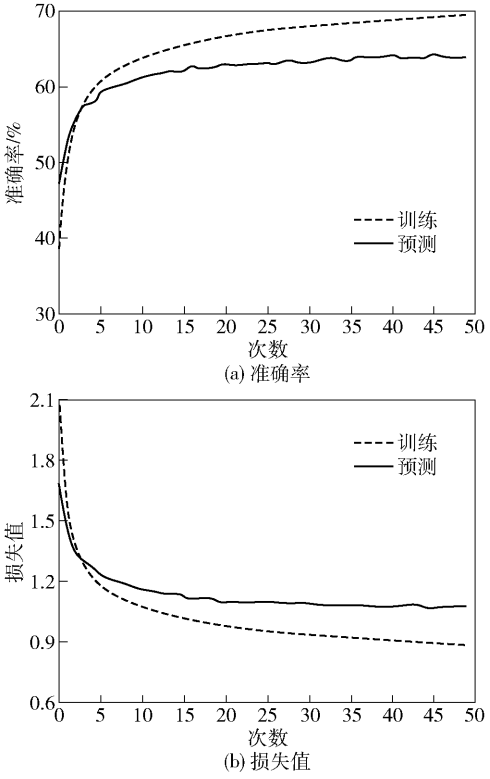

图 4　类 SM4 算法卷积神经网络 800 点的训练结果

由以上实验内容可知，使用多层感知机或卷积神经网络进行训练时，多层感知机使用能量迹 800 点得到的准确率和损失值优于卷积神经网络. 虽然侧信道分析发现前 500 点已经包括了第 1 字节攻击的全部能量消耗信息，但无论是多层感知机还是卷积神经网络，进行训练时得到的训练效果都没有 800 点训练的效果好，这点与侧信道攻击不同. 由于能量迹上的前 200 点未能包括第 1 轮 S 盒置换的能量消耗信息，所以无论是多层感知机还是卷积神经网络，进行训练的结果都不很理想，这在预期的结果中从另一方面说明了逆向分析结果确实与 S 盒替换的能量消耗相关.

与此同时，利用传统模板攻击进行 S 盒的逆向分析. 在基于模板攻击的逆向分析中选取的能量迹中包含第 1 个 S 盒置换能量消耗在内的 400 点，使用了 256 ×8 000 条能量迹进行多元高斯均值和方差的建模，并选取 256 ×2 000 条能量迹进行匹配. 采集完能量迹后，对能量迹中的点使用泄露信息的大小进行兴趣点排序，然后选取不同的兴趣点和批大小进行模板逆向恢复实验，结果如表 3 所示.

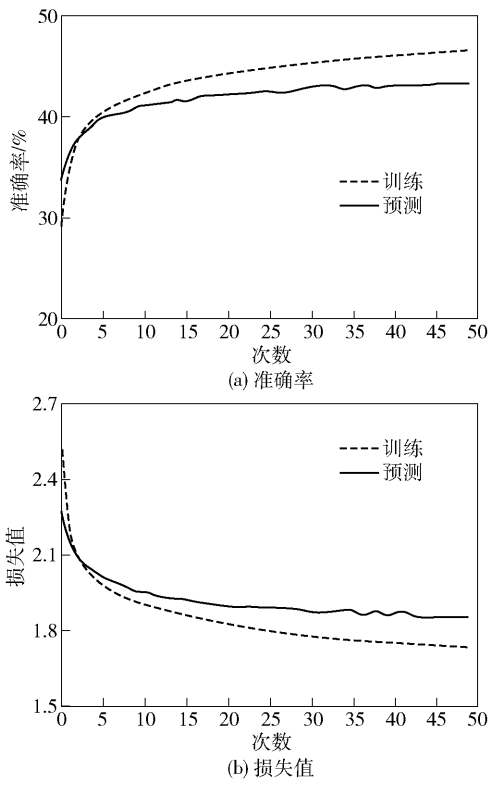

图 5　类 SM4 算法卷积神经网络 500 点的训练结果

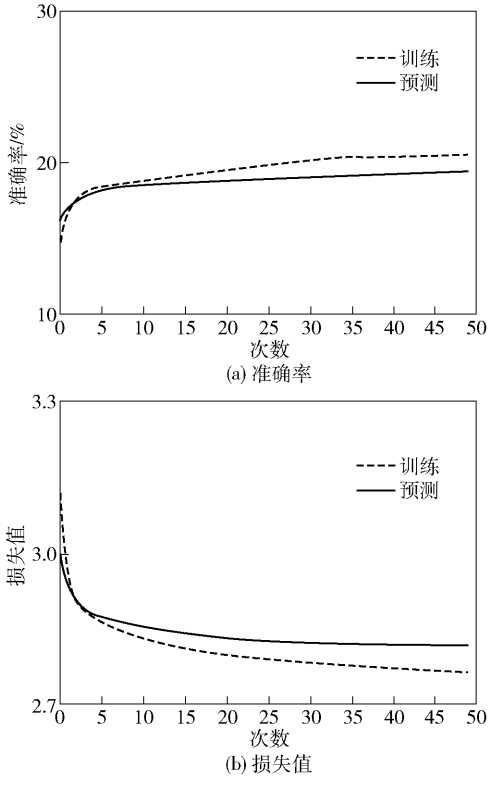

图 6　类 SM4 算法卷积神经网络 200 点的训练结果

由表 3 可知，兴趣点的大小和批大小对模板逆向恢复的影响比较大，兴趣点和批大小的增加会提

表3 模板攻击与深度学习算法S盒逆向准确率 %

批条数	模板攻击(兴趣点选取)					深度学习算法	
	5	10	15	20	25	多层感知器	卷积神经网络
100	19.3	25.0	31.1	36.4	40.3	78.0	65.9
200	23.8	31.4	38.5	43.8	48.0	79.0	66.3
300	24.3	34.1	39.8	44.9	49.0	79.5	66.0
400	28.9	39.0	45.1	53.4	58.0	79.2	66.1
500	31.1	40.0	47.6	53.4	59.5	79.3	65.8

高逆向分析准确率. 与模板攻击相比,深度学习不需要选择兴趣点;此外,深度学习在批处理上有一定的优势,它在选择批大小方面比较宽松,而模板攻击要求批大小必须足够大. 在表3中给定的批处理条件下,深度学习的准确率更高.

4 结束语

出于商业保密或其他安全因素的考虑,设计人员会对密码算法的核心部件保密不公开,这给密码分析者带来了很多困难. 如何有效进行逆向分析,一直是安全分析者关注的问题. 首先研究了基于建模类场景下基于深度学习的类SM4算法逆向分析,提出了通用的S盒逆向分析算法. 通过构建特定明文,设计符合侧信道攻击的训练数据和标签,并以实际智能卡为例使用多层感知机和卷积神经网络2种算法进行了类SM4算法的逆向分析实验,都正确恢复出了S盒的内容.

总之,深度学习在能量迹对齐、兴趣点选取、攻击复杂度等方面具有优势,因此在逆向分析领域取得了很好的分析效果. 该方法为智能卡安全评估提供了一种新的有效攻击方法,值得深入研究. 而实际智能卡产品的安全设计除考虑传统的侧信道攻击之外,还应充分考虑深度学习攻击下的防护措施.

参考文献:

[1] Biryukov A, Shamir A. Structural cryptanalysis of SASAS [C]// International Conference on the Theory and Applications of Cryptographic Techniques. Heidelberg: Springer, 2001: 395-405.

[2] Torrance R, James D. The state-of-the-art in IC reverse engineering [C] // International Workshop on Cryptographic Hardware and Embedded Systems. Heidelberg: Springer, 2009: 363-381.

[3] Clavier C, Wurcker A. Reverse engineering of a secret AES-like cipher by ineffective fault analysis [C] // 2013 Workshop on Fault Diagnosis and Tolerance in Cryptography. Piscataway: IEEE, 2013: 119-128.

[4] Daudigny R, Ledig H, Muller F, et al. SCARE of the DES[C]// International Conference on Applied Cryptography and Network Security. Heidelberg: Springer, 2005: 393-406.

[5] Novak R. Side-channel based reverse engineering of secret algorithms[C]// Proceedings of the 12th International Electro Technical and Computer Science Conference. Piscataway: IEEE, 2003: 445-448.

[6] Rivain M, Roche T. SCARE of secret ciphers with SPN structures[C] // International Conference on the Theory and Application of Cryptology and Information Security. Heidelberg: Springer, 2013: 526-544.

[7] Biryukov A, Perrin L. On reverse-engineering S-boxes with hidden design criteria or structure [C] // Annual Cryptology Conference. Heidelberg: Springer, 2015: 116-140.

[8] Gao Si, Chen Hua, Wu Wenling, et al. My traces learn what you did in the dark: recovering secret signals without key guesses[C]// Cryptographers' Track at the RSA Conference. [S. l.]: Springer, 2017: 363-378.

[9] 马向亮, 李冰, 习伟, 等. 基于独立分量技术的类GIFT算法S盒逆向分析[J]. 计算机研究与发展, 2018, 55(10): 2269-2277.
Ma Xiangliang, Li Bing, Xi Wei, et al. Reverse-analysis of S-box for GIFT-like algorithms based on independent component analysis technology[J]. Journal of Computer Research and Development, 2018, 55(10): 2269-2277.

[10] Timon B. Non-profiled deep learning-based side-channel attacks with sensitivity analysis[J]. IACR Transactions on Cryptographic Hardware and Embedded Systems, 2019, 2019(2), 107-131.

[11] 张蕾, 吴文玲. SMS4密码算法的差分故障攻击[J]. 计算机学报, 2006, 29(9): 1596-1602.
Zhang Lei, Wu Wenling. Differential fault analysis on SMS4[J]. Chinese Journal of Computers, 2006, 29(9): 1596-1602.

[12] 马向亮, 王宏, 李冰, 等. 基于能量分析技术的芯片后门指令分析方法[J]. 电子学报, 2019, 47(3): 686-691.
Ma Xiangliang, Wang Hong, Li Bing, et al. A power analysis method against backdoor instruction in chips [J]. Acta Electronica Sinica, 2019, 47(3): 686-691.

[13] Maghrebi H, Portigliatti T, Prouff E. Breaking cryptographic implementations using deep learning techniques [C] // International Conference on Security, Privacy, and Applied Cryptography Engineering. [S. l.]: Springer, 2016: 3-26.