文章编号：2096-1618(2022)01-0016-05

基于 CNN-BPR 的 S-Box 功耗随机化侧信道攻击

曹家华， 吴 震， 王 燚， 王 敏
(成都信息工程大学网络空间安全学院，四川 成都 610225)

摘要：S-Box 功耗随机化是一种对抗侧信道攻击的防御方案，该方案将设备加密过程中 S-Box 输出值功耗泄露的位置进行随机化处理，降低了中间值与能量消耗的相关性，使得基于固定位置进行能量分析的代价大幅增加。具备平移不变性的卷积神经网络在侧信道攻击上取得了显著的效果。为进一步提高其对 S-Box 功耗随机化防御方案的攻击能力，基于贝叶斯个性化排序的思想，提出一种更符合侧信道攻击原理的 CNN-BPR 模型。实验结果表明，与 Softmax 交叉熵损失模型相比，CNN-BPR 模型在使用全部训练能迹用于模板攻击时，成功恢复密钥所需要的攻击能迹数量能够减少 3%，当使用 60% 的训练能迹用于模板攻击时，减少的攻击能迹数量能够达到 27%。

关 键 词：侧信道攻击；模板攻击；卷积神经网络；损失函数；贝叶斯个性化排序
中图分类号：TP309 **文献标志码**：A
doi：10.16836/j.cnki.jcuit.2022.01.003

0 引言

密码算法在理论上得到了密码学家的严格审查，但是从物理实施的角度来看，密码算法仍然很脆弱，密码设备的运行过程中不可避免地存在着电磁辐射、能量消耗、时间变化或之类的侧信道泄露。侧信道攻击(side channel attack，SCA)就是一种利用这些侧信道泄露来破解秘密信息的攻击方法。Kocher[1]根据操作私钥所需要的时间，实现了对 RSA 算法的侧信道攻击，大量研究成果表明无保护措施的密码设备基本难以抵抗侧信道攻击。能量分析攻击属于侧信道攻击中应用较为广泛的一种，其中简单能量分析[2]、相关性能量分析[3-4]和差分能量分析[5]常用于直接对目标设备的功耗信息进行攻击。而模板攻击[6]则需要攻击者拥有与目标设备相同类型的设备，并假设功耗噪声服从高斯分布来构造汉明重量模板或中间值模板，然后利用这些模板对密钥进行恢复。

S-Box 功耗随机化[7]是一种抵抗能量分析攻击的常见策略，主要分为能量值随机化和时间随机化。能量值随机化方法以添加掩码来实现，通常可以使用高阶差分能量分析[8]完成攻击。时间随机化方法旨在打乱中间值出现的时间点，导致攻击者无法有效捕获中间值的能量信息，差分能量分析攻击和传统模板攻击方法在面对这类防御手段时往往表现出很差的攻击效果。

近年来，大量实验证明卷积神经网络能够构建更为高效的模板[9]。Cagli 等[10]提出了一种卷积神经网络与数据增强相结合的方法，这种方法不需要对齐能量轨迹和选择兴趣点，减少了传统能量分析攻击中的任务量。Benadjila 等[11]全面测试了卷积神经网络中各类参数对模板攻击的影响，通过组合每一类参数的最优值，提出了基于 VGG-16 网络结构的卷积神经网络模型 CNN_{best}，并通过实验证明了 CNN_{best} 在 ASCAD 数据集上的攻击效果要优于传统模板攻击、多层感知器模型和原始 VGG-16 模型。Zaid 等[12]提出了一种使用排序损失函数代替 CNN_{best} 中交叉熵损失函数的方案，并验证了该方案能够使 CNN_{best} 模型更快地收敛到最佳状态。为有效地构建 S-Box 功耗随机化的模板，文中将使用卷积神经网络构建基本模型，并从损失函数的角度对其优化。

1 背景知识

1.1 卷积神经网络

卷积神经网络主要通过卷积层、池化层和全连接层实现对输入样本的分类。卷积层通过卷积运算和激活函数将本层输入数据的特征提取出来，然后传递给池化层。池化层通过最大池化或平均池化进行下采样，以此降低卷积层的输出维度，实现模型的平移不变性。全连接层负责在最后对数据进行分类，通过特征加权得到每个类别的置信度分数，然后经过分类函数的计算得到输入样本的类别。在侧信道攻击中通常使用一维卷积和一维池化，计算过程如图 1 所示。

收稿日期：2021-09-27
基金项目：“十三五”国家密码发展基金资助项目(MMJJ20180224)；四川省重点研发资助项目(2019YFG0096)

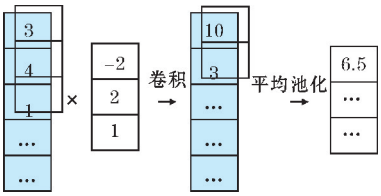

图 1　一维卷积与池化

1.2　损失函数

卷积神经网络中除了各层网络结构之外,另一个对模型性能有较大影响的因素是损失函数。损失函数能够表现出模型的预测值和真实值之间的差异程度,使用合适的损失函数,并尽可能降低损失函数的计算结果是优化模型的一种方法。常见的损失函数有平均绝对误差、均方误差以及交叉熵(cross entropy, CE)等,通常模板攻击使用 Softmax 交叉熵作为损失函数。以 AES 加密算法为例,以首轮加密中 S-Box 输出的一个字节为中间值,在明文已知的情况下,该中间值和密钥满足一一映射的关系,因此中间值的概率分布与对应位置密钥的概率分布一致。假设密钥的取值集合为 $K=\{0,1,\cdots,255\}$,定义 $k\in K$ 表示一个猜测密钥,y_k 表示猜测密钥 k 对应的真实标签,p_k 表示猜测密钥 k 的预测概率,可以得到单个样本的 Softmax 交叉熵损失:

$$\mathrm{Loss}_{\mathrm{CE}}=-\sum_{k\in K}y_k\ln p_k \tag{1}$$

2　损失函数的改进方法

2.1　L2 约束的损失函数

卷积神经网络在功耗泄露明显的数据中能取得很好的训练效果,但对于 S-Box 功耗随机化数据的效果却很差。主要是基于以下两个原因:一是大多数卷积神经网络使用 Softmax 交叉熵作为损失函数,这无法保证模型学习到的正对之间特征差异较小而负对之间特征差异较大。因此,在泄露信息极少的功耗随机化数据中,不同密钥所带来的功耗差异没有被有效地提取。Softmax 交叉熵以最大化同一批样本的条件概率之和来降低模型的损失,但是功耗随机化导致同一批的各个样本识别难易程度相差较大,Softmax 交叉熵通过让容易识别的样本范数更大,让难以识别的样本范数更小,从而将损失最小化。如果直接使用 Softmax 交叉熵损失,很容易让模型只关注随机化离散程度较低的样本,而忽略随机化离散程度较高的样本,从而导致训练出来的模型在攻击数据上的表现很差。

L2 范数被广泛应用于机器学习的损失函数正则化[13],可以提高模型的泛化能力,避免过拟合的发生。在卷积神经网络中应用 L2 范数进行正则化处理,其过程如图 2 所示。

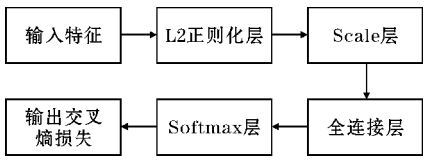

图 2　L2 正则化过程

图中,L2 正则化层将所有样本的特征进行归一化,Scale 层对 L2 正则化层的结果进行 α 倍的缩放,这样可以降低特征范数对模型的影响,增加正对之间的相似性和负对之间的差异性。同时,更小的特征范数也使模型对简单样本和困难样本的关注程度更相近。

在模板攻击中找出样本所对应的真实密钥是典型的单标签多分类任务。定义真实密钥 k^* 的标签为 1,错误密钥的标签为 0,即 y_k 在 $k\neq k^*$ 时为 0,损失值 Loss 仅受到猜测密钥 $k=k^*$ 的预测概率影响,使用 W_k,b_k 分别表示猜测密钥 k 在全连接层的权重向量和偏置,X 表示输入到全连接层的特征向量,那么可以得到关于特征向量的 Softmax 交叉熵损失:

$$\mathrm{Loss}_{\mathrm{CE}}=-\ln\frac{\mathrm{e}^{W_k^T*X+b_k*}}{\sum_{k\in K}\mathrm{e}^{W_k^T\cdot X+b_k}} \tag{2}$$

将正则化处理后的 X 代入式(2),即可得到 L2 约束的 Softmax 交叉熵损失函数,计算结果为

$$\mathrm{Loss}_{\mathrm{L2\_CE}}=-\ln\frac{\mathrm{e}^{W_k^T*\frac{\alpha X}{\|X\|_2}+b_k*}}{\sum_{k\in K}\mathrm{e}^{W_k^T\cdot\frac{\alpha X}{\|X\|_2}+b_k}} \tag{3}$$

2.2　贝叶斯个性化排序损失函数

贝叶斯个性排序[14](Bayesian personalized ranking,BPR)是推荐系统中一种基于 Pairwise 方法的排序算法。以商品推荐为例,BPR 算法将每个用户对应的所有商品按权重排序,然后向用户推荐排名靠前的商品,以此从极大数量的商品集中推选出更符合用户需要的商品。基于此思想,不拘泥于对真实密钥权重的提升,而是直接对模板攻击中 Rank 指标进行优化,更适用于分类难度较高的 S-Box 功耗随机化数据。

Rank 指标是一种典型的模板攻击评价标准[15],将模型的预测结果进行排序得到猜测密钥 k 的排序名次 $\mathrm{Rank}(k)$,k 的排名越靠前,$\mathrm{Rank}(k)$ 越低,k 为真实密钥的可能性越大。当 $\mathrm{Rank}(k^*)=1$ 时,表示真实密钥 k^* 的预测结果排在第一位,即成功恢复出了密钥。在卷积神经网络中,把全连接层输出的置信度分数 $W_k^T\cdot X+b_k$ 作为猜测密钥 k 的得分 $s(k)$,将所有猜测密钥的得分进行降序排序得到排序名次 $\mathrm{Rank}(k)$,真实

密钥 k^* 的排名 Rank(k^*) 的计算方法为

$$\mathrm{Rank}(k^*)=\sum_{k\in K}\begin{cases}0, & \text{if } s(k^*)>s(k)\\ 1, & \text{else}\end{cases} \tag{4}$$

在 BPR 中 $k^*>_t k$ 为能量轨迹 t 的一个偏序关系,表示 k^* 的真实排名在 k 的前面,即 Rank(k^*)<Rank(k)。基于能量轨迹构建的偏序关系满足以下两个条件:一是不同能量轨迹之间的真实密钥相互独立;二是任意一条能量轨迹的密钥是独立随机的。使用$>_t$ 符号表示能量轨迹 t 中所有的偏序关系,BPR 基于最大后验估计 $P(\theta|>_t)$ 来求解模型参数 θ,根据贝叶斯公式可以得到:

$$P(\theta|>_t)=\frac{P(>_t|\theta)P(\theta)}{P(>_t)} \tag{5}$$

由于能量轨迹的密钥是独立随机的,因此可以得到 $P(>_t)=1/2$,根据式(5)可以得到:

$$P(\theta|>_t)=2P(>_t|\theta)P(\theta) \tag{6}$$

最大化 $P(\theta|>_t)$ 等价于最大化 $P(>_t|\theta)$ 和 $P(\theta)$。对于 $P(>_t|\theta)$,模板攻击中关键是获得真实密钥的排序关系,因此令式(6)的$>_t=k^*>_t k$,得到:

$$P(>_t|\theta)=\prod_{k\in K}P(k^*>_t k|\theta) \tag{7}$$

$P(k^*>_t k|\theta)$ 表示在参数为 θ 的情况下,真实密钥比猜测密钥排名靠前的概率。对于 $k^*>_t k$ 这一事件,使用卷积神经网络全连接层输出的置信度分数 $s(k^*)-s(k)>0$ 来表示,使用 Sigmoid 函数来代替发生这一事件的概率,为

$$P(>_t|\theta)=\prod_{k\in K}\frac{1}{1+\mathrm{e}^{-(s(k^*)-s(k))}} \tag{8}$$

对于 $P(\theta)$,假设参数 θ 的概率分布满足均值为 0,协方差矩阵为 λI 的正态分布,那么其对数与 $\|\theta\|_2^2$ 成正比。

$$\ln P(\theta)=\lambda\ \|\theta\|_2^2 \tag{9}$$

最大化 $P(\theta|>_t)$ 等价于最小化其对数的负值,使用 $P(>_t|\theta)$ 和 $P(\theta)$ 代替 $P(\theta|>_t)$,得到 BPR 的损失函数:

$$\mathrm{Loss}_{\mathrm{BPR}}=\sum_{k\in K}\ln(1+\mathrm{e}^{s(k)-s(k^*)})-\lambda\ \|\theta\|_2^2 \tag{10}$$

3 模型设计与实验

本文使用的卷积神经网络结构设计如图 3 所示,输入是包含 700 个样本点的能量轨迹。卷积层所使用的卷积核大小为 11、步长为 2,激活函数使用 Relu,过滤器数量依次为 128、256、512。池化层使用大小为 2、步长为 2 的平均池化。首层全连接层有 4096 个神经元,激活函数使用 Relu。末层全连接层有 256 个神经元,激活函数使用 Softmax。

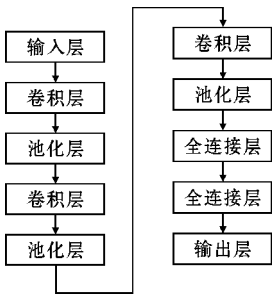

图 3 网络结构图

3.1 实验环境及数据

实验环境所使用的卷积神经网络基于 Python Tensorflow 2.0 进行搭建,然后在搭载了 4 块 NVIDIA GeForce RTX 2080 Ti GPU 的服务器上训练模板并进行攻击。

实验数据使用法国国家网络安全局侧信道数据库(ANSSI SCA Database, ASCAD)提供的 ASCAD.h5 和 ASCAD_desync50.h5 数据。ASCAD.h5 由 50000 条训练数据和 10000 条攻击数据组成,每一条数据都有 700 个样本点,包含了 AES 第一轮加密的第三个 S-box 输出操作的功耗信息。ASCAD_desync50.h5 由 ASCAD.h5 进行非对齐操作后得到,具体实现为将 ASCAD.h5 中的每一条数据向左随机偏移,其中偏移量 $\beta\in\{0,1,\cdots,50\}$。

3.2 实验分析

对图 3 所示的卷积神经网络结构分别应用 Softmax 交叉熵损失函数、L2 约束的损失函数和 BPR 损失函数,所有模型均训练 150 个 Epochs,其中 L2 约束的损失函数中缩放倍数 $\alpha=2$,BPR 损失函数中 $\lambda=0.0001$。通过 Rank 值随攻击能迹数的变化来表示攻击效果,当攻击能迹数相同时,Rank 值越低则表示模型攻击效果越好。

图 4、图 5 是当选取 30000 条训练能迹数时,3 种损失函数模型在有随机偏移和无偏移的数据集上的攻击效果。根据对比可以看出,BPR 损失函数和 L2 约束的损失函数在无偏移的数据集上能够更快地降低 Rank 值,但三者成功恢复密钥需要的攻击能迹数相近。当数据存在随机偏移之后,BPR 损失函数和 L2 约束的损失函数产生的攻击效果明显优于传统 Softmax 交叉熵损失函数。其中,BPR 损失函数的效果最好,能够以 582 条攻击能迹数达到 Rank 值为 1 的攻击效果,而另外两种损失函数达到这一效果则需要 800 条左右的攻击能迹。

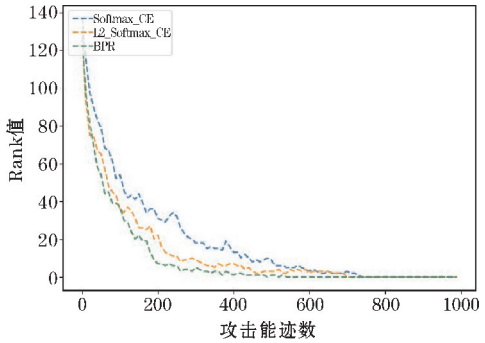

图 4　训练能迹数为 30000,偏移量 $\beta=50$

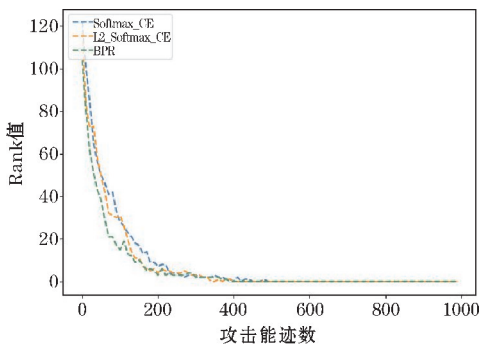

图 5　训练能迹数为 30000,无偏移量

对训练所使用的能迹数量进行增加,然后对比 3 种损失函数在 ASCAD_desync50. h5 数据集上的效果,记录攻击阶段 Rank 值等于 1 时所需要的最少能迹数量,其结果如表 1 所示。

表 1　攻击成功所需的能迹数量

损失函数	训练能迹数量		
	30000	40000	50000
Softmax_CE	795	592	371
L2_Softmax_CE	732	523	375
BPR	582	450	362

由表 1 可知,当训练能迹数量足够大时,3 种模型的攻击效果比较接近,但是随着训练能迹数的减少,Softmax 交叉熵损失函数模型的攻击效果下降得更加明显,而 BPR 受到的影响则相对较小。当侧信道攻击中无法大量捕获到目标设备的功耗时,优化 Rank 值的 BPR 模型比 Softmax 交叉熵模型更适用于神经网络的模板攻击。

综上,在基于 Rank 值的评价标准下,本文提出的 CNN-BPR 对于功耗数据的随机偏移具备一定的抵抗能力。在训练样本数较少的情况下,BPR 和 L2 约束的损失函数由于正则项的影响,训练出的模型具有更好的鲁棒性,在噪声较大的数据集上优于传统 Softmax 交叉熵损失函数。

3　结束语

基于交叉熵的卷积神经网络在分类任务中,关注的是每一个样本的真实标签被预测到的程度,真实标签的预测概率越大则效果越好,当需要检测某一个样本的种类时,通过这一关注点能够很好地构造模型。然而,使用卷积神经网络进行模板攻击的任务目标却不完全与分类任务相同。这是因为在模板攻击中,攻击阶段的样本通常使用同一密钥,即攻击阶段的样本具有相同的标签。因此,只需要使真实密钥在攻击样本集合上的综合概率最大即可,这一目标通过尽可能地提高真实密钥的 Rank 值来达到。本文提出了一种基于 BPR 损失函数的卷积神经网络模型 CNN-BPR,并在 ASCAD. h5 和 ASCAD_desync50. h5 这两组数据集上进行了验证。实验结果表明,BPR 损失函数相较于交叉熵损失函数能够构造鲁棒性更高的模型,且能有效减少成功恢复密钥所需要的攻击能迹数量。

虽然提出的 CNN-BPR 模型相比使用交叉熵损失函数的模型有一定的优化,但当训练能迹数量充足时,过拟合不再是约束模型性能的主要问题,因此正则化损失函数带来的优化效果并不明显。在这种情况下,如何继续改善网络模型,还需要进一步研究。

参考文献:

[1] Kocher P C. Timing attacks on implementat-ions of Diffie-Hellman, RSA, DSS, and oth-er systems [C]. Annual International Cryptolo-gy Conference,1996:104-113.

[2] Mangard S. A Simple Power-Analysis (SPA) Attack on Implementations of the AES KeyExpansion [C]. Information Security and Cry-ptology-ICISC 2002,2002:28-29.

[3] Benhadjyoussef N, Machhout M, Tourki R. Optimized power trace numbers in CPA atta-cks[C]. 2011 8th International Multi-Confere-nce on Systems,Signals and Devices(SSD),2011:1-5.

[4] 杜之波,吴震,王敏. 针对基于 SM3 的 HMAC 的能量分析攻击方法[J]. 通信学报,2016,37(5):38-43.

[5] Kocher P,Jaffe J,Jun B. Differential poweranalysis [C]. Advances in Cryptology,1999:388-397.

[6] Standaert F-X, Archambeau C. Using subsp-ace-based template attacks to compare andcombine

power and electromagnetic informat-ion leakages [C]. Cryptographic Hardware and Embedded Systems-CHES 2008,2008:411-425.

[7] Veyrat-Charvillon N, Medwed M, Kerckhof S. Shuffling against Side-Channel Attacks: AComprehensive Study with Cautionary Note[C]. Advances in Cryptology-ASIACRYPT 2012,2012:740-757.

[8] Okeya K, Sakurai K. A Second-Order DPAAttack Breaks a Window-Method Based Co-untermeasure against Side Channel Attacks[C]. Information Security,2002:389-401.

[9] 杨欢,吴震,王燚. 侧信道多层感知器攻击中基于贝叶斯优化的超参数寻优[J]. 计算机应用与软件,2021,38(5):323-330.

[10] Cagli E, Dumas C, Prouff E. ConvolutionalNeural Networks with Data AugmentationAgainst Jitter-Based Countermeasures[C]. Cr-yptographic Hardware and Embedded Syste-ms-CHES 2017,2017:45-68.

[11] Benadjila R, Prouff E, Strullu R. Deep lear-ning for side-channel analysis and introducti-on to ASCAD database[J]. Journal of Crypt-ographic Engineering,2020:163-188.

[12] Zaid G, Bossuet L, Dassance F. RankingLoss: Maximizing the Success Rate in DeepLearning Side-Channel Analysis[C]. Cryptogr-aphic Hardware and Embedded Systems-CHES 2021,2021:25-55.

[13] Wang Y, Wang Q, Guo X, et al. Optimizati-on and Performance Analysis of Extreme L-earning Machine by L2-Norm Regularization[M]. Cham: Springer,2021:405-413.

[14] Dave VS, Zhang B, Chen PY. Neural Brane: Neural Bayesian Personalized Ranking forAtributed Network Embedding[J]. Data Scie-nceand Engineering,2019,4(2):119-131.

[15] François-Xavier Standaert, Malkin T, YungM. A Unified Framework for the Analysisof Side-Channel Key Recovery Attacks (ext-ended version)[C]. Advances in Cryptology-EUROCRYPT 2009,2009:443-461.

Side Channel Attack of S-box Power Randomization based on CNN-BPR

CAO Jiahua, WU Zhen, WANG Yi, WANG Min

(College of Cyberspace Security, Chengdu University of Information Technology, Chengdu 610225, China)

Abstract: S-box power randomization is a defense scheme against side channel attack. This scheme randomizes the location of power leakage of S-box output value in the process of device encryption, and reduces the correlation between intermediate value and energy consumption, and greatly increases the cost of energy analysis based on fixed location. Convolutional neural network with translation invariance has achieved remarkable results in side channel attack. In order to further improve its attack ability against S-box power randomization defense scheme, a CNN-BPR model more in line with the principle of side channel attack is proposed based on the idea of Bayesian personalized ranking. The experimental results show that, compared with softmax cross-entropy loss model, when CNN-BPR model uses all training energy traces for template attack, the number of attack traces required to successfully recover the key can be reduced by 3%, and when 60% of the training energy traces are used for template attack, the number of attack energy traces can be reduced by 27%.

Keywords: side channel attack; template attack; convolutional neural network; loss function; Bayesian personalized ranking