计算机工程与应用
Computer Engineering and Applications
ISSN 1002-8331,CN 11-2127/TP

《计算机工程与应用》网络首发论文

题目：信息熵角度下的深度学习旁路安全评估框架
作者：宋世杰，陈开颜，张阳
网络首发日期：2020-10-21
引用格式：宋世杰，陈开颜，张阳．信息熵角度下的深度学习旁路安全评估框架．计算机工程与应用.
https://kns.cnki.net/kcms/detail/11.2127.TP.20201020.1647.006.html

网络首发时间：2020-10-21 09:34:19
网络首发地址：https://kns.cnki.net/kcms/detail/11.2127.TP.20201020.1647.006.html

信息熵角度下的深度学习旁路安全评估框架

宋世杰，陈开颜，张　阳

陆军工程大学 石家庄校区装备模拟训练中心，石家庄 050000

摘　要：基于深度学习的建模类旁路密码分析（DLSCA, Deep Learning Side Channel Analysis/Attack）对于各种旁路攻击场景的密码破解效果都十分显著。但是 DLSCA 的仍存有安全评估问题。本文基于 AES 对称加密算法的能量分析，通过信息熵角度分析准确率等传统机器学习性能指标无法评估 DLSCA 深度神经网络（DNN, Deep Neural Network）模型训练程度的原因。定义密钥信息量，分别阐释密钥信息量与旁路安全评估、DNN 模型训练阶段性能评估的关系，建立深度学习模型与旁路分析二者的联系，提出以密钥信息量为核心的 DLSCA 安全评估框架。

关键词：旁路分析；深度学习；安全评估；信息熵

文献标志码: A　　**中图分类号**:TP309.7　　doi：10.3778/j.issn.1002-8331.2005-0221

宋世杰，陈开颜，张阳. 信息熵角度下的深度学习旁路安全评估框架.计算机工程与应用

SONG Shijie, CHEN Kaiyan, ZHANG Yang. A security evaluation framework of deep learning side channel analysis from information entropy. Computer Engineering and Applications

A security evaluation framework of deep learning side channel analysis from information entropy

SONG Shijie, CHEN Kaiyan, ZHANG Yang

Center of Equipment Simulation Training, Shijiazhuang Campus of the Army Engineering University, Shijiazhuang, 050000, China

Abstract：Deep Learning Side Channel Analysis/Attack (DLSCA)based on deep learning is very effective for decryption in various side channel attack scenarios. But DLSCA still has security evaluation problems. This paperis based on the power analysis of AES symmetric encryption algorithm, and explain the reason why traditional machine learning performance metrics such as Accuracy can not evaluate from information entropy. The key in-formation is defined tofind out the relationship between side channel security evaluation and the performance of the DNN model during a training phase. Associated with the key information, a DLSCA security evaluation framework is set up.

Key words：Side channel analysis; Deep learning; Security evaluation; Information entropy

基金项目：国家自然科学基金（No.51377170，No.61271125）；国家青年科学基金资助项目（No.61602505）。

作者简介：宋世杰（1994- ），男，在读硕士研究生，主要研究领域为旁路攻击；陈开颜（1970- ），通信作者，女，副教授，硕导，主要研究领域密码学；张阳（1984-），男，讲师，主要领域为集成电路安全，E-mail:ssj710640965@163.com。

1 引言

旁路分析[1]由 Paul.K[1]提出，是一类利用加密设备的物理泄露进行破密或评估的密码分析方法，分为建模类旁路分析与非建模旁路分析。其中建模类旁路分析由 Chari 等人[2]提出，核心是旁路敌手获取与目标加密设备完全一致的副本及其所有加密信息，实施两阶段的分析：1. 建模阶段——敌手在副本上实施遍历密钥空间的所有加密过程并采集对应功耗，根据统计分析方法提出功耗特征，建立“模板”；2. 攻击阶段——敌手对目标设备再次采集功耗，利用已有“模板”进行匹配，获得密钥。由于敌手获知目标副本的全部信息，所以建模类旁路分析被信息熵证实为最强大的旁路分析方法。

建模类旁路分析包括：模板攻击[2]（Template Attacks, TA）、机器学习旁路分析（Machine Learning Side Channel Attack, MLSCA）[3,4]以及深度学习旁路分析（DLSCA）。前两种方法均需要复杂的人为特征工程建模，且处理的旁路信号维度容易受限，因此特征提取自主化、数据维度兼容能力强的 DLSCA 流行起来。现有研究中，以多层感知机（Multiple Layer Perceptron, MLP）[5]和卷积神经网络（Convolutional Neural Network, CNN）[6]为主的 DLSCA 对带掩码[7]、信号偏移[8]以及加噪声[9]的旁路信号表现出优秀的建模攻击能力。

DLSCA 研究通常采用机器学习性能指标中的准确率评估 DNN 模型的训练程度，并采用 Standaert 等人[10]旁路安全评估框架中的成功率与猜测熵评估测试阶段的破密效果。Cagli 等人[11]通过实验对采用准确率衡量 DLSCA 训练阶段的评估方法提出质疑。Picek 等人[12]通过剖析准确率与成功率的概念证实 DNN 模型在训练阶段的准确率与测试阶段的成功率无法对应，DLSCA 训练与测试阶段的评估无法对接。

但是 DLSCA 在安全评估方面仍有很多问题需要研究完善：①[12]仅通过准确率与成功率的定义对比得出前者无法衡量 DNN 模型训练程度的结论，并未从旁路安全评估角度入手讨论其根本原因；②对 DNN 模型训练阶段的性能评估与测试阶段旁路安全评估之间的联系缺乏理论研究；

针对上述问题，本文从信息论中信息熵的角度入手：①将 DLSCA 整个过程马尔可夫化，分析准确率等传统机器学习性能指标无法评估 DLSCA 中 DNN 模型训练程度的原因，得出 DNN 与 SCA 的安全评估核心在于密钥信息量提取的结论；②通过密钥信息量构建 DLSCA 测试安全评估与训练性能评估的联系，并提出以密钥信息量为核心的 DLSCA 安全评估框架。

[1]旁路分析与旁路攻击在本文意义等同，不做区分。

2 DLSCA 理论基础

2.1 旁路分析模型

根据能量分析原理[13]，可构造基于功耗泄露的旁路攻击模型如图 1 所示。定义变量：形如 $\mathcal{X}$ 为集合，$X=\{x_1,...x_n\}$ 是 $\mathcal{X}$ 中的 N 维随机变量，$\mathbf{X}\in\mathbb{R}^{N\times D}$ 为 $\mathcal{X}$ 中的 $N\times D$ 向量，$\mathbf{X}=\{\mathbf{x}_1,...,\mathbf{x}_n\},\mathbf{x}_i\in\mathbb{R}^D$。敌手所要获取的正确密钥记为 $k^*\in\mathcal{K}$，n 比特长（本文 $n=8$），密钥空间 $\mathcal{K}=\{k_0,...,k_{\mathcal{K}-1}\}$，$K$ 服从均匀分布。假设敌手对加密设备输入 N 条明文 $\mathcal{P}=\{p_1,...,p_n\}$，$p$ 服从均匀分布，通过式（1）泄露模型得到功耗轨迹 $\mathbf{X}=\{\mathbf{x}_1,...,\mathbf{x}_q\}$：

$$\mathbf{x}_i=\mathbf{C}(p_i\oplus k^*)+\mathbf{n}_i \quad (1)$$

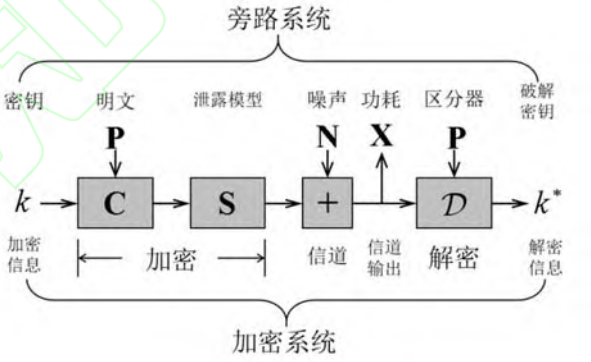


图 1 旁路分析模型

Fig.1 SCA model

$\mathbf{N}=\{n_1,...,n_q\}$ 为高斯噪声，与 $\mathbf{X}$ 分布相互独立。$\mathbf{C}(\cdot)$ 为已知加密函数，本文选择 AES[14]加密的第一轮第一个 S 盒之前的轮密钥加做加密函数。通常敌手选取合适的统计模型如汉明重量/距离和加密中间值度量泄露变化，本文选择中间值泄露模型[13]。故定义泄露模型 $Z=\mathbf{S}[\mathbf{C}(P,K)],\mathcal{Z}=\{s_1,...,s_z\}$，$\mathbf{S}$ 为 S 盒的字节替换操作 Z 代表泄露对应的加密中间值。

2.2 DLSCA 步骤方法

在假设空间 $\mathcal{H}$ 中，定义 DNN 的超参数集合函数为 $f_\theta:\theta=(\mathbf{W},\mathbf{b})$，$\theta\in\Theta\subseteq\mathbb{R}^q$，映射 f_θ 代表 DNN 算法模型，$f_\theta=\Pr[\hat{Z}\mid\mathbf{X}]$ 为 DNN 输出的预测概率。

基于深度学习的建模类旁路攻击也分为两个阶段：建模（训练）与攻击（测试）阶段。根据 2.1 的旁路分析模型，DLSCA 步骤为：

（1）建模准备：在目标加密设备的副本上采集 N_p

规模的建模用数据集 S_p，其服从 $S_p \sim \Pr[\mathbf{X}, Z]^{N_p}$ 概率分布，数据集中功耗轨迹为 $\mathbf{X}$，泄露模型中间值类别为 Z。

（2）建模阶段：对应 DL 训练阶段基于数据集 S_p，选取合适的 DNN 算法 f_θ，$\mathbf{X}$ 为输入，Z 为标签，通过监督学习训练输出对应类别的概率：$f_\theta = \Pr[\hat{Z} \mid \mathbf{X}]$。

（3）攻击准备：对应 DL 测试阶段，在目标设备上获取 N_a 规模的攻击用数据集 S_a，其概率分布为 $S_a = \{k^*, (\mathbf{x}_1, p_1), ..., (\mathbf{x}_{N_a}, p_{N_a})\} \sim \Pr[\mathbf{X}, Z]^{N_a}$。对于 $i \in [\![1, N_a]\!]$，$p_i \sim \Pr[P]^{N_a}$ 为已知明文概率分布，功耗概率分布为 $\mathbf{x}_i \sim \Pr\{\mathbf{X} \mid Z = \mathbf{S}[\mathbf{C}(p_i, k^*)]\}^{N_a}$。

（4）预测估计：将 S_a 中每条功耗输入最终优化模型 f_θ 生成预测概率向量 $y_i = f_\theta(\mathbf{x}_i)$，对应每个标签 Z 的概率值。将所有攻击轨迹输出的预测概率值累积，获取每个标签对应密钥值的可能性，预测概率最高的候选值即对应正确密钥 k^*。通常使用极大似然估计法定义区分器 $\mathcal{D}$：

$$\mathcal{D}_{S_a}[k] = \prod_{i=1}^{N_a} y_i[z_i] = \sum_{i=1}^{N_a} \log_2(y_i[z_i]) \tag{2}$$

其中 $z_i = \mathbf{S}[\mathbf{C}(p_i, k)]$。

2.3 现有评估指标

评估问题的量化离不开评估指标。研究 DLSCA 安全评估问题之前，首先对机器学习领域与旁路分析领域中现有的评估指标做以介绍。

2.3.1 机器学习评估指标

在机器学习中，为准确地评估 DNN 模型 f_θ 的训练程度，通常分为学习指标与性能指标：

（1）学习指标

学习指标通过计算给定训练数据集 $S_p \sim \Pr[\mathbf{X}, Z]^{N_p}$ 平均分类误差，以此来优化 DNN 模型的网络参数 $\{f_\theta : \theta \in \Theta\}$。机器学习领域中常用对数损失函数，如式（3）衡量分类误差。令对于给定训练数据集 $S_p \sim \Pr[\mathbf{X}, Z]^{N_p}$，参数空间 $\mathcal{H}$ 中 DNN 模型的参数集合函数为 $\{f_\theta : \theta \in \Theta\}$，$f_\theta$ 输出的预测类别为 $\hat{Z}$，则有对数损失函数[15]：

$$\mathcal{L}(\theta) = -\frac{1}{N_p}\sum_{i=1}^{N_p} \log_2 f_\theta(\mathbf{X}) = -\frac{1}{N_p}\sum_{i=1}^{N_p} \log_2 \Pr(\hat{Z} \mid \mathbf{X}) \tag{3}$$

通过随机梯度下降 SGD[16]算法优化损失函数 $\mathcal{L}(\theta)$，将 N_p 分成大小为 batch（数据批量，一般远小于 N_p）的小数据批次，每个迭代周期 epoch 从 batch 中随机挑选一个维度的数据对 $\mathcal{L}(\theta)$ 进行参数集合 $\theta \in \Theta$ 上的偏导计算：

$$\Delta\theta = \eta \cdot \nabla_\theta \mathcal{L}(\theta, \mathbf{x}_i, z_i), i \in N_p \tag{4}$$

其中已知常数 η 为学习率，即每次梯度下降的步长。训练阶段采取 SGD 优化整个参数集 θ 使得 $\mathcal{L}(\theta)$ 达到全局最小值，得到最优模型 f_θ^*：

$$\theta^* = \arg\min_{\theta \in \Theta} \mathcal{L}(\theta) \tag{5}$$

（2）性能指标

性能指标用于衡量 DNN 模型的分类性能，是 DNN 训练程度的体现。通常采用准确率，记作 Acc 来度量模型 f_θ 正确分类的概率：

$$Acc = \Pr[\arg\max_{s \in \mathcal{Z}} f_\theta(\mathbf{x})[s] = Z] \tag{6}$$

通过性能指标与学习指标可以评估 DNN 模型训练的程度。训练程度分为：欠拟合、过拟合[17]与最佳拟合。

欠拟合指 DNN 模型在训练阶段并未提取到关于功耗轨迹 $\mathbf{X}$ 相关特征的状态，可归因于数据量不足与网络结构的复杂度不够。过拟合状态指当 DNN 模型提取与功耗数据 $\mathbf{X}$ 无关特征时的状态。f_θ 处于过拟合状态时，其评估指标在训练阶段会表现极好但在测试阶段指标表现极差。DNN 模型的过拟合与欠拟合都会导致很差的泛化结果，而只有处在两者中的最佳拟合状态时，才具备良好的泛化能力。

2.3.2 旁路分析评估指标

现有旁路分析技术均采用[10]中的安全评估框架，其中用于量化评估的指标分为两大类：信息指标和安全指标。

（1）信息指标

采用香农信息熵[18]中的条件熵与互信息来衡量旁路泄露与采集的信息量。根据 2.2，已知功耗 $\mathbf{X}$ 估计加密中间泄露值 Z 的条件熵为：

$$\mathrm{H}(Z \mid \mathbf{X}) = \mathop{\mathbb{E}}_{\mathbf{X},Z} \Pr[\mathbf{X}, Z] \cdot \log_2 \Pr[Z \mid \mathbf{X}] \tag{7}$$

由功耗 $\mathbf{X}$ 推知 Z 获得的有效信息衡量——互信息为：

$$\mathrm{I}(Z; \mathbf{X}) = \mathrm{H}(Z) - \mathrm{H}(Z \mid \mathbf{X}) \tag{8}$$

（2）安全指标

安全指标对敌手的攻击结果以及密钥破解程度的衡量，包含成功率与猜测熵。成功率反应了一定数量轨迹的密钥最大获取能力，而猜测熵则对应其平均能力。式（2）中敌手采用最大似然估计获得评分 $\mathcal{D}_{S_a}[k]$ 后，估计的密钥候选值会按分数进行排序，随着轨迹数的增加，正确密钥的分数会增大，排序会逐渐上升至首位即：

$$g_s(k^*) = \sum_{k \in \mathcal{K}} 1_{\mathcal{D}_{S_a}[k] > \mathcal{D}_{S_a}[k^*]} \tag{9}$$

当 $g(k^*) = 1$ 代表攻击成功，敌手获知正确密钥。猜测熵与成功率都围绕 $g(k^*)$ 进行定义：

猜测熵：

$$\mathrm{GE} = \sum_{i=1}^{Z} i \cdot \Pr[g_i(k^*) = s] \tag{10}$$

成功率：

$$\mathrm{SR}(N_a)=\frac{1}{E}\sum_{e=1}^{E}\Pr[g_i(k^*)=1] \tag{11}$$

其中 E 为攻击次数。

现有 DLSCA 研究中，评估通常在训练（建模）阶段使用机器学习评估指标，测试（攻击）阶段使用旁路分析评估指标。而研究[12]通过概念对比证实机器学习常用的性能指标——准确率无法评估 DLSCA 训练阶段，这导致其训练与测试两阶段评估脱节，旁路分析指标与机器学习指标无法对接。

3 密钥信息量为核心的 DLSCA 安全评估框架

本节从信息熵视角，把 DNN 作为旁路区分器，将 DLSCA 整个过程视为一条马尔可夫链，以此研究并得出准确率等机器学习性能指标无法评估 DLSCA 训练阶段的根本原因在于无法评估该过程中的密钥信息量。围绕密钥信息量，分别研究其与 SCA 以及 DNN 安全评估方面的关系，提出以密钥信息量为核心的 DLSCA 评估框架，从而建立 DNN 模型训练阶段的性能评估与测试阶段旁路安全评估之间的联系。

3.1 密钥信息量与旁路安全评估

准确率无法评估 DLSCA 训练阶段的根本原因是什么？本节将 SCA 视作马尔可夫过程，通过引入密钥信息量的概念解释旁路安全评估的核心，并研究密钥信息量与安全评估指标的关系。

3.1.1 密钥信息量与旁路安全评估

DLSCA 虽是深度学习与旁路分析的结合，但仍是旁路分析问题，解决 DLSCA 评估问题，需从旁路安全分析角度入手。根据 2.1 节中的旁路分析模型，SCA 整个过程中，每一时刻的概率分布仅与其前一时刻的概率分布有关，因此该过程可视作一个马尔可夫过程：

引理 1（马尔可夫过程[19]）SCA 过程可定义为马尔可夫过程：

$$(K,P)\to(Z,N)\to X\to\mathcal{D}(X,P)\to\hat{K}$$

即

$$K\to Z\to X\to\hat{Z}\to\hat{K} \tag{12}$$

其中 $\mathcal{D}(\cdot)$ 为旁路区分器，$\hat{K}$ 为 DNN 模型预测类别对应的密钥值。

根据 SCA 马尔可夫过程，结合 2.2 中 DLSCA 实施步骤，DLSCA 安全评估要解决以下问题：

问题 1（评估问题[20]）给定训练数据集 S_p，其概率分布为 $S_p\sim\Pr[\mathbf{X},Z]^{N_p}$，在 $\mathrm{SR}(N_a)\ge\alpha$（$\alpha$ 为固定阈值）条件下，求一最优 DNN 模型 f_θ^*，使 $x\to\hat{z}$ 过程攻击用轨迹数 N_a 最小。

问题 1 是基于预设测试结果条件的 DLSCA 安全评估问题，包含训练与测试两阶段的评估：求最优模型 f_θ^* 属于 DNN 模型在训练阶段的性能评估任务，基于阈值 α 最小化 N_a 属于测试阶段的旁路安全评估任务。由于训练阶段属于深度学习分类问题，因而性能评估要解决的就是衡量模型在学习过程中的输出预测概率 $\Pr[\hat{Z}\mid\mathbf{X}]$ 与真实概率 $\Pr[Z\mid\mathbf{X}]$ 间贝叶斯误差优化问题。因此将问题 1 的训练阶段评估问题可转化为：

命题 1（性能评估）DLSCA 训练阶段安全评估问题为衡量 DNN 模型实际输出值 $f_\theta=\Pr[\hat{Z}\mid\mathbf{X}]$ 与真实值 $f_\theta^*=\Pr[Z\mid\mathbf{X}]$ 间的贝叶斯误差优化问题：

$$\Pr[\hat{Z}\mid\mathbf{X}]\underset{f_\theta\to f_\theta^*}{\overset{Np\to\infty}{\to}}\Pr[Z\mid\mathbf{X}] \tag{13}$$

根据准确率 *Acc* 定义，其为计算预测类别 $\hat{Z}$ 与真实类别 Z 相等时的概率。因此首先考虑使用准确率 *Acc* 来解决命题 1。当 *Acc* 趋于 1 时，所有预测类别 $\hat{Z}$ 与真实类别 Z 相等。但实际操作中，准确率 *Acc* 无法与测试阶段的安全评估任务联系起来，结合准确率的定义式（6）以旁路安全评估角度分析如下：

①从旁路安全评估的指标角度看：安全指标中成功率 SR 的定义为对 N_a 条轨迹整体破密正确概率的平均值，反映的是样本整体最大密钥提取能力，须在功耗总样本 $\mathbf{X}$ 维度累加 $\Pr[\hat{Z}\mid\mathbf{X}]$ 后再取 E 次攻击平均。而 *Acc* 反映的是单个样本 $\mathbf{x}\in\mathbf{X}$ 的分类预测概率，评估的维度不等价。之于猜测熵 GE，其反映的是正确密钥在假设空间 $\mathcal{K}$ 中的排序水平，通过每条轨迹叠加预测信息减少估计错误，从而改变正确密钥排序，是规模为 N_a 轨迹的平均密钥提取能力。显然 *Acc* 无法显示多条轨迹的累积变化。

②从旁路安全评估的实质看：由引理 1，SCA 可视作是马尔可夫过程，故 SCA 整个过程中只要功耗总样本中的密钥信息传递与利用充足，密钥就能破解。因此将命题 1 中的贝叶斯误差优化转换成信息熵中信息量的表达形式，引入密钥信息量定义为：

定义 1（密钥信息量）对于给定训练数据集 $S_p\sim\Pr[\mathbf{X},Z]^{N_p}$，$S_p$ 携带的密钥信息量为 $X\to Z$ 过程中每个样本 $\mathbf{x}\in\mathbf{X}$ 包含对应标签值 $\hat{Z}$ 的比特数：

$$\mathrm{KI}(\hat{Z}\mid\mathbf{X})=-\sum_{i=1}^{N_p}\log_2\Pr[\hat{Z}\mid\mathbf{X}=\mathbf{x}_i] \tag{14}$$

当 N_p 规模的功耗样本所包含的密钥信息量 $\mathrm{KI}(\hat{Z}\mid\mathbf{X})\to\mathrm{KI}(Z\mid\mathbf{X})$ 时，对应的贝叶斯误差优化问题为 $\Pr[\hat{Z}\mid\mathbf{X}]\to\Pr[Z\mid\mathbf{X}]$。这也恰与式（2）的对数极大似然区分照应，本质在于累积每条轨迹的密钥信息量 $\mathrm{KI}[\hat{Z}\mid\mathbf{X}]$。当密钥信息量累加到一定量时，区分器 $\mathcal{D}$ 的极大似然值对应 $g(k)$ 中估计密钥排序至首位，即解得正确密钥，密钥信息量为 $\mathrm{KI}(Z\mid\mathbf{X})$。而单个样本 *Acc* 获得的最大输出概率值很大，其总样本输出概率值可能很小，对应得到的密钥信息量也会少，就会出现 *Acc* 的值很高趋近于 1，而成功率很小的现象。因此，使用准确率 *Acc* 衡量训练阶段的 DNN 模型性能无法满足旁路安全评估需求，旁路安全评估的本质在于密钥信息量 $\mathrm{KI}[\hat{Z}\mid\mathbf{X}]$ 的提取。

3.1.2 密钥信息量与安全指标的关系

旁路安全评估的本质在于密钥信息量的提取，与命题 1 性能评估照应。既然密钥信息量 $\mathrm{KI}[\hat{Z}\mid\mathbf{X}]$ 为旁路安全评估的关键，那么其如何与旁路分析的安全评估指标关联起来呢？解决这一问题，问题 1 中的测试阶段的安全评估任务也将得解。重新观察问题 1，其中后半段涉及到了成功率与破密最小轨迹数 N_a 两个指标。成功率对应旁路安全指标中的 SR 。对于最小轨迹数 N_a，旁路安全指标中猜测熵 GE 是衡量破解出密钥与否的指标：当 GE 减小并收敛于一定值时，正确密钥升至 $g(k)$ 首位，密钥获解，此时对应的功耗轨迹数即为 N_a 值。因此，须研究密钥信息量 $\mathrm{KI}[\hat{Z}\mid\mathbf{X}]$ 与旁路安全指标成功率 SR 与猜测熵 GE 的关系。

（1）密钥信息量与猜测熵

在实际计算中，为了将 GE 与极大似然统一起来，对 GE 中的概率值取对数操作。又由命题 1 是对条件概率 $f_\theta=\Pr[\hat{Z}\mid\mathbf{X}]$ 进行优化，本文引入条件猜测熵的定义：

定义 2（条件猜测熵）令测试集 $S_a\sim\Pr[\mathbf{X},Z]^{N_a}$，$N_a\in\mathbb{N}$，$Z\in S(k,p)$ 为标签，$\mathbf{X}\in\mathbb{R}^{N_a\times D}$ 为功耗轨迹，条件概率熵为：

$$\begin{aligned}\mathcal{G}_m&=-\sum_{s=1}^{Z}i\cdot\delta_{\mathbf{x}}(s=i)\sum_{\mathbf{x}=1}^{N_a}\log_2\Pr[\hat{Z}=s\mid\mathbf{x}]\\&=-\sum_{s=1}^{Z}i\cdot\delta_{\mathbf{x}}(s=i)\cdot\mathrm{KI}[\hat{Z}=s\mid\mathbf{x}]\end{aligned}\tag{15}$$

$\delta_{\mathbf{x}}(\hat{z})$ 为对密钥预测概率 $\Pr[\hat{Z}=s\mid\mathbf{x}]$ 的降次排序函数，运算实质等同于式（9）中的 $g_s(\cdot)$ 密钥排序函数。观察式（15）条件猜测熵 $\mathcal{G}_m$ 定义，式（14）密钥信息量就在其中体现，再次证实旁路安全评估就是对旁路功耗密钥信息量的定量计算。区分器 $\mathcal{D}$ 积累每条测试集的密钥信息量 $\mathrm{KI}(\hat{Z}=s\mid\mathbf{X})$，并改变 $\delta_{\mathbf{x}}(\hat{z})$ 中预测概率 $\Pr[\hat{Z}=s\mid\mathbf{x}]$ 的位置。当 N_a 增至 $\mathcal{D}$ 对密钥区分所需密钥信息量时，$\delta_{\mathbf{x}}(\hat{z})$ 将预测概率排至首位，预测概率值等于真实概率值，$i=1$。则当 $\alpha\to1$ 时，条件猜测熵 $\mathcal{G}_m$ 趋于定值 $\mathrm{H}(Z\mid\mathbf{X})$，此时所需轨迹数即为最小 N_a。因此猜测熵 GE 是基于密钥信息量 $\mathrm{KI}[\hat{Z}\mid\mathbf{X}]$ 排序计算的指标。

（2）密钥信息量与成功率

根据引理 1 中 SCA 过程可视作马尔可夫过程，因此根据费诺不等式有：

引理 2（费诺不等式[18]）对于 SCA 中的马尔可夫过程 $K\to Z\to X\to\hat{Z}\to\hat{K}$，令安全指标成功率为 $\mathrm{SR}=\Pr[K=\hat{K}]$，则有：

$$h_2(\mathrm{SR})+(1-\mathrm{SR})\log_2\left|\mathcal{K}-1\right|=\mathrm{H}(\hat{Z}\mid\mathbf{X})\geq\mathrm{H}(Z\mid\mathbf{X})\tag{16}$$

其中 $h_2(\mathrm{SR})=-\mathrm{SR}\log_2\mathrm{SR}-(1-\mathrm{SR})\log_2(1-\mathrm{SR})$，$\mathcal{K}$ 为密钥空间，不等式左边代表 SCA 破密最多需要传输的比特数。

由引理 2 可知，成功率 SR 同样与条件熵 $\mathrm{H}(\hat{Z}\mid\mathbf{X})$ 相关。而 $\mathrm{H}(\hat{Z}\mid\mathbf{X})$ 根据其定义式（7），是由密钥信息量 $\mathrm{KI}[\hat{Z}\mid\mathbf{X}]$ 组成而计算的。因此成功率 SR 与密钥信息量可通过信息熵 $\mathrm{H}(\hat{Z}\mid\mathbf{X})$ 的运算建立联系。

对于成功率 SR 与猜测熵 GE 二者与密钥信息量的关系，其本质都是由条件猜测熵 $\mathcal{G}_m$ 的优化进而实施安全评估的。于是对于问题 1 的测试阶段任务本质为：

命题 2（安全评估）给定测试数据集 $S_a\sim\Pr[\mathbf{X},Z]^{N_a}$，$\mathrm{SR}(N_a)\geq\alpha$ 与 $\mathcal{G}_m\overset{N_a}{\to}\mathrm{H}(Z\mid\mathbf{X})$ 等价，所需的轨迹数即为 N_a 最小值，其中 i 是预测概率 $\Pr[\hat{Z}\mid\mathbf{X}]$ 在 $\delta_{\mathbf{x}}(\hat{z})$ 中排序的位置。

3.2 密钥信息量与 DNN 性能评估

上节以信息熵视角探讨了 DLSCA 的安全评估实质，即基于功耗数据集中密钥信息量的提取，这也是解决命题 1 性能评估与命题 2 安全评估二者对接问题的关键。如果知道密钥信息量在 DNN 训练学习过程中的运算与计量方法，则 DNN 模型训练阶段的性能评估与测试阶段旁路安全评估之间的对接难题得解。本小节将通过信息瓶颈理论探究 DNN 模型训练过程与密钥信息量 $\mathrm{KI}[\hat{Z}\mid\mathbf{X}]$ 的关系，证实交叉熵可基于测试阶段的安全评估任务对 DNN 模型训练阶段的性能定量评估。

3.2.1 信息瓶颈理论解释 DNN

DNN 因其学习过程的复杂性向来被认作一个黑盒模型处理。上节指出机器学习性能指标不可用，因此需要寻找另外的衡量指标关联起密钥信息量。根据信息瓶颈理论[21]可将 DNN 层与层之间的关联视作马尔可夫过程，该过程中传递的互信息可衡量其学习程度：

引理 2（信息瓶颈理论[21]）在假设空间 $\mathcal{H}$ 中，DNN 模型 $f_\theta(\theta\in\Theta\subseteq\mathbb{R}^q)$ 结构可解释为一个贝叶斯分层结构。

因 DNN 中隐藏层 T_i 的输入是上一层的输出 T_{i-1}，则 $\mathrm{DNN}=\{\mathbf{X};T_1;\ldots;T_i;\hat{Z}\}$ 可等价于马尔可夫链：

$$X\to T_1\to T_2\to\ldots\to T_{i-1}\to T_i\to\hat{Z}$$

对应的概率分布为：

$$\Pr[\hat{Z}\mid\mathbf{X}]=\Pr[\mathbf{X}]\cdot\Pr[T_1\mid\mathbf{X}]\cdot\ldots\Pr[T_i\mid T_{i-1}]\cdot\Pr[\hat{Z}\mid T_i]$$

由引理 2 的信息瓶颈理论，DNN 的实质与 SCA 过程相同，也可视为一个马尔可夫过程，如图 2。其中，加密阶段对应 DNN 的特征选取过程 $\mathbf{X}\to\mathbf{T}$，隐藏层 T_i

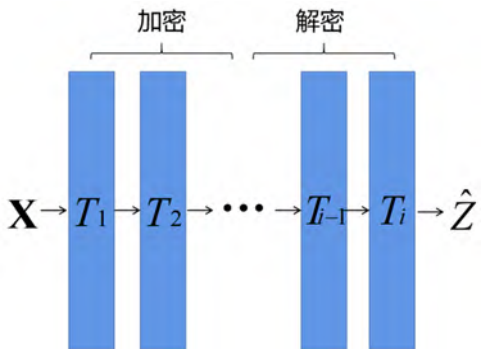


图 2　DNN 信息瓶颈理论示意图

Fig.2　Schematic diagram of DNN Information Bottle Theory

中的每个神经元根据概率 $p_{\mathbf{w}}(t_i \mid \mathbf{x}) \cdot p_{\mathbf{w}}(\mathbf{x})$ 计算互信息 $\mathrm{I}[T_i;\mathbf{X}]$，来衡量特征提取算法对输入数据信息的特征信息量。该过程实则对输入数据降维，故称作信息压缩；解密阶段对应 DNN 的分类过程 $\mathbf{T} \to \hat{z}$，通常在隐藏层中的全连接层展开。分类网络层中的神经元根据 $p_{\mathbf{w}}(\hat{z} \mid t_i) \cdot p_{\mathbf{w}}(t_i \mid \mathbf{x})$ 的分布变化计算互信息 $\mathrm{I}[\hat{Z};T_i]$，并将 DNN 提取的特征维度摊平至分类维度，实现数据特征的再编码以衡量泛化信息量，该过程称作信息扩展。整个信息压缩与扩展过程，DNN 模型通过最小化 $\mathrm{I}[T_i;\mathbf{X}]$ 与最大化 $\mathrm{I}[\hat{Z};T_i]$ 来实现优化。

命题 3（信息熵优化）DNN 模型 $f_\theta(\theta \in \Theta \subseteq \mathbb{R}^q)$ 为马尔可夫信道，存在以下互信息变化不等式：

$$\mathrm{H}(\mathbf{X}) \geq \mathrm{I}(\mathbf{X};T_1) \geq \ldots \geq \mathrm{I}(\mathbf{X};T_i) \geq \mathrm{I}(\mathbf{X};\hat{Z})$$

$$\mathrm{I}(\mathbf{X};Z) \geq \mathrm{I}(T_1;Z) \geq \ldots \geq \mathrm{I}(T_i;Z) \geq \mathrm{I}(\hat{Z};Z)$$

根据上述互信息不等式，命题 1 训练任务存在以下等价关系：

$$\Pr[\hat{Z} \mid \mathbf{X}] \underset{f_\theta \to f_\theta^*}{\overset{N_p \to \infty}{\to}} \Pr[Z \mid \mathbf{X}] \Leftrightarrow \mathrm{I}[\hat{Z};\mathbf{X}] \underset{f_\theta \to f_\theta^*}{\overset{N_p \to \infty}{\to}} \mathrm{I}[Z;\mathbf{X}] \quad (17)$$

$$\mathrm{I}[\hat{Z};\mathbf{X}] \underset{f_\theta \to f_\theta^*}{\overset{N_p \to \infty}{\to}} \mathrm{I}[Z;\mathbf{X}] \Leftrightarrow \mathrm{H}[\hat{Z} \mid \mathbf{X}] \underset{f_\theta \to f_\theta^*}{\overset{N_p \to \infty}{\to}} \mathrm{H}[Z \mid \mathbf{X}] \quad (18)$$

整个 DNN 性能优化过程中，$\mathrm{I}[T_i;\mathbf{X}]$ 取决于 DNN 选取的结构。本文假设 DNN 模型结构固定，则其性能优化过程为增大训练阶段互信息 $\mathrm{I}[\hat{Z};\mathbf{X}]$ 的过程，从而实现 $\Pr[\hat{Z} \mid \mathbf{X}] \to \Pr[Z \mid \mathbf{X}]$ 的优化。而由互信息定义，$\mathrm{I}[\hat{Z};\mathbf{X}]$ 为 $\mathrm{H}(\hat{Z})$ 与 $\mathrm{H}(\hat{Z} \mid \mathbf{X})$ 之差，由于 $\mathrm{H}(\hat{Z})$ 是定值，因此命题 1 在信息熵视角下的本质自然等价于优化条件熵 $\mathrm{H}(\hat{Z} \mid \mathbf{X})$。而根据条件熵式（7）的定义，密钥信息量 $\mathrm{KI}[\hat{Z} \mid \mathbf{X}]$ 作为 $\mathrm{H}(\hat{Z} \mid \mathbf{X})$ 计算的一部分，可与 DNN 优化 $\mathrm{I}[\hat{Z};\mathbf{X}]$ 的过程实现对接。

3.2.2　交叉熵优化

命题 3 可以将密钥信息量 $\mathrm{KI}(\hat{Z} \mid \mathbf{X}) \to \mathrm{KI}(Z \mid \mathbf{X})$ 与 $X \to \hat{Z}$ 过程的互信息优化 $\mathrm{I}[\hat{Z} \mid \mathbf{X}] \to \mathrm{I}[Z \mid \mathbf{X}]$ 建立联系，也是命题 1 性能评估的进一步推导。而对于 $\mathrm{I}[\hat{Z};\mathbf{X}]$ 与 $\mathrm{I}[Z;\mathbf{X}]$ 之间误差的衡量，信息论中通常采用交叉熵（Cross Entropy）作为指标，与深度学习中交叉熵损失函数对应。

定义 3（交叉熵）给定训练数据集 $S_p \sim \Pr[\mathbf{X},Z]^{N_p}$，误差分布为 $\Pr[\hat{Z} \mid \mathbf{X}]$，正确分布为 $\Pr[Z \mid \mathbf{X}]$，$\theta \in \Theta$ 为 DNN 参数集合定义交叉熵为：

$$\begin{aligned} \mathcal{C}_{\mathrm{x},z}(\theta) &= -\sum_{\mathbf{X},Z} \Pr[Z \mid \mathbf{X}] \cdot \log_2 \Pr[\hat{Z} \mid \mathbf{X}] \\ &= -\sum_{Z \mid \mathbf{X}} \Pr[Z \mid \mathbf{X}] \cdot \mathrm{KI}[\hat{Z} \mid \mathbf{X}] \end{aligned} \quad (19)$$

当 $\mathrm{KI}(\hat{Z} \mid \mathbf{X}) \underset{f_\theta \to f_\theta^*}{\overset{N_p \to \infty}{\to}} \mathrm{KI}(Z \mid \mathbf{X})$ 时，有：

$$\mathcal{C}_{\mathrm{x},z}(\theta) \underset{f_\theta \to f_\theta^*}{\overset{N_p \to \infty}{\to}} \mathrm{H}(Z \mid \mathbf{X}) \quad (20)$$

证明：根据条件熵定义有：

$$\begin{aligned} \mathrm{H}(Z \mid \mathbf{X}) &= -\sum_{\mathbf{X},Z} \Pr[Z \mid \mathbf{X}] \cdot \log_2 \Pr[Z \mid \mathbf{X}] \\ &= -\sum_{Z \mid \mathbf{X}} \Pr[Z \mid \mathbf{X}] \cdot \mathrm{KI}[Z \mid \mathbf{X}] \end{aligned}$$

则当 $\mathrm{KI}(\hat{Z} \mid \mathbf{X}) \underset{f_\theta \to f_\theta^*}{\overset{N_p \to \infty}{\to}} \mathrm{KI}(Z \mid \mathbf{X})$ 时，对应概率分布 $\Pr[\hat{Z} \mid \mathbf{X}] \underset{f_\theta \to f_\theta^*}{\overset{N_p \to \infty}{\to}} \Pr[Z \mid \mathbf{X}]$，所以 $\mathcal{C}_{\mathrm{x},z}(\theta) \underset{f_\theta \to f_\theta^*}{\overset{N_p \to \infty}{\to}} \mathrm{H}(Z \mid \mathbf{X})$ 得证。

由定义 3，交叉熵损失 $\mathcal{C}_{\mathrm{x},z}(\theta)$ 实质为衡量预测分布 $\Pr[\hat{Z} \mid \mathbf{X}]$ 与真实分布 $\Pr[Z \mid \mathbf{X}]$ 对应密钥信息量 $\mathrm{KI}(\hat{Z} \mid \mathbf{X})$ 与 $\mathrm{KI}(Z \mid \mathbf{X})$ 的差距。DNN 模型通过 SGD 算法优化[16] $\mathcal{C}_{\mathrm{x},z}(\theta)$，从而使 DNN 从输入功耗数据得到的密钥信息量逐渐接近其理想值 $\mathrm{KI}(Z \mid \mathbf{X})$。当预测分布 $\Pr[\hat{Z} \mid \mathbf{X}]$ 与真实分布 $\Pr[Z \mid \mathbf{X}]$ 之间的贝叶斯误差随着 DNN 的训练逐渐趋于 0 时，交叉熵损失 $\mathcal{C}_{\mathrm{x},z}(\theta)$ 最终趋于条件熵 $\mathrm{H}[Z \mid \mathbf{X}]$，与命题 3 照应。所以，虽然作为分类评估指标的 Acc 无法评估 DNN 模型在训练阶段的性能，但是 DLSCA 仍是一个监督学习的分类过程，依旧可以使用交叉熵损失 $\mathcal{C}_{\mathrm{x},z}(\theta)$ 度量预测标签 $\hat{Z}$ 与真实标签 Z 之间的误差，从而判断 DNN 模型性能。因此，DNN 模型训练阶段的性能评估与测试阶段旁路安全评估之间可由交叉熵与条件猜测熵对密钥信息量 $\mathrm{KI}(\hat{Z} \mid \mathbf{X})$ 的评估建立联系：

定理 1（DLSCA 过程评估）给定训练数据集 $S_p \sim \Pr[\mathbf{X},Z]^{N_p}$，$N_p \in \mathbb{N}$，$Z \in S(k,p)$ 为标签，$\mathbf{X} \in \mathbb{R}^{N_p \times D}$ 为功耗轨迹，$\mathcal{H}$ 假设空间中 $\theta \in \Theta$ 为 DNN 参数集合，$\mathcal{C}_{\mathrm{x},z}(\theta)$ 为 DNN 模型 f_θ 的交叉熵损失，$\mathcal{G}_m$ 为 SCA 计算密钥 K 的条件猜测熵。由引理 1 与引理 2 可将整个 DLSCA 过程马尔可夫化，DNN 算法模型为区分器：

$$K \to Z \to X \to T1 \to \ldots \to T_i \to \hat{Z} \to K$$

当 $\mathrm{KI}(\hat{Z} \mid \mathbf{X}) \underset{f_\theta \to f_\theta^*}{\overset{N_p \to \infty}{\to}} \mathrm{KI}(Z \mid \mathbf{X})$ 时，有：

$$\mathcal{C}_{\mathrm{x},z}(\theta) \underset{f_\theta \to f_\theta^*}{\overset{N_p \to \infty}{\to}} \mathrm{H}(Z \mid \mathbf{X}) \Leftrightarrow \mathcal{G}_m \underset{f_\theta \to f_\theta^*}{\overset{N_a \to 1}{\to}} \mathrm{H}(Z \mid \mathbf{X}) \quad (21)$$

证明：式（21）左边根据定义 3 成立，右边由定义 3 成立。

3.1 中论述 $\mathcal{G}_m$ 是通过极大似然定律计算的，那么推导交叉熵损失 $\mathcal{C}_{\mathrm{x},z}(\theta)$ 与极大似然的关系才能证明定理 1 中的等价性，而实际两者是统一的。令 DNN 的参数空间 $\theta\in\mathcal{H}$，对于训练数据集 $S_p \sim \Pr[\mathbf{X},Z]^{N_p}$，根据极大似然定律有：

$$\begin{aligned}\mathcal{H}^*(\theta)&=\arg\max_{\theta}\Pr[\mathbf{X},Z\mid\theta]\\&=\arg\max_{\theta}\Pr[Z\mid\mathbf{X},\theta]\Pr[\mathbf{X}]\\&=\arg\max_{\theta}\prod_{i=1}^{N_p}\Pr[z_i\mid\mathbf{x}_i,\theta]\Pr[\mathbf{x}_i]\\&=\arg\max_{\theta}\frac{1}{N_p}\sum_{i=1}^{N_p}\log_2\Pr[z_i\mid\mathbf{x}_i,\theta]\Pr[\mathbf{x}_i]\\&=\arg\max_{\theta}\frac{1}{N_p}\sum_{i=1}^{N_p}\log_2(\Pr[z_i\mid\mathbf{x}_i,\theta]+\Pr[\mathbf{x}_i])\end{aligned}$$

因为 θ 与 $\Pr[\mathbf{X}]$ 无关，即：

$$\mathcal{H}^*(\theta)==\arg\max_{\theta}\frac{1}{N_p}\sum_{i=1}^{N_p}\log_2\Pr[z_i\mid\mathbf{x}_i,\theta]$$

又 $\mathbf{X},Z$ 是从 S_p 采样而来，则：

$$\mathcal{H}^*(\theta)=\arg\max_{\theta}\sum_{z}\sum_{\mathrm{x}}\Pr[\mathbf{X},Z]\log_2\Pr[Z\mid\mathbf{X},\theta]$$

又 $\Pr[Z\mid\mathbf{X},\theta]=\Pr[\hat{Z}\mid\mathbf{X}]$，则：

$$\begin{aligned}\mathcal{H}^*(\theta)&=\arg\max_{\theta}\sum_{z}\sum_{\mathrm{x}}\Pr[\mathbf{X}]\Pr[Z\mid\mathbf{X}]\log_2\Pr[\hat{Z}\mid\mathbf{X}]\\&=\arg\min_{\theta}\Pr[\mathbf{X}]\cdot\mathcal{C}_{\mathrm{x},z}(\theta)=\arg\min_{\theta}\mathcal{C}_{\mathrm{x},z}(\theta)\end{aligned}$$

因而极大似然与优化交叉熵是一致的，即得出 DNN 的性能优化本质为极大似然定律，这与 SCA 的区分器原理一致，证实定理 1 中在 DLSCA 马尔可夫信道里，DNN 实际等价于旁路区分器模块。在根据交叉熵损失 $\mathcal{C}_{\mathrm{x},z}(\theta)$ 与 $\mathcal{G}_m$ 的定义，其两者存在以下关系：

$$\begin{aligned}\mathcal{C}_{\mathrm{x},z}(\theta)&=-\sum_{\mathrm{x},Z}\Pr[Z\mid\mathbf{X}]\cdot\log_2\Pr[\hat{Z}\mid\mathbf{X}]\\&=-\sum_{s=1}^{z}\Pr[Z\mid\mathbf{X}]\cdot\mathrm{KI}[\hat{Z}=s\mid\mathbf{X}]\\&\le-\sum_{s=1}^{z}\delta_{\mathrm{x}}(\hat{Z})\cdot i\cdot\mathrm{KI}[\hat{Z}\mid\mathbf{X}]=\mathcal{G}_m\end{aligned}$$

综上，根据大数定律，交叉熵损失 $\mathcal{C}_{\mathrm{x},z}(\theta)$ 与条件猜测熵 $\mathcal{G}_m$ 在数据空间趋于无穷大时，且密钥信息量有 $\mathrm{KI}(\hat{Z}\mid\mathbf{X})\to\mathrm{KI}(Z\mid\mathbf{X})$，两者的优化过程根据极大似然定律是一致的，定理 1 得证。这也说明，DLSCA 中 DNN 模型性能评估的实质在于优化交叉熵 $\mathcal{C}_{\mathrm{x},z}(\theta)$，该任务在信息熵视角下，围绕极大似然定律，可通过密钥信息量与测试阶段的旁路安全评估实现对接。由此可提出以密钥信息量为核心的 DLSCA 安全评估框架，如图 2：

如图 3 所示，围绕密钥信息量，可以将 DNN 模型训练阶段的性能评估与测试阶段旁路安全评估关联起来：根据命题 1 性能评估与密钥信息量的定义，命题 1 的问题可以转化为密钥信息量的提取优化问题。定义 2 中 SCA 极大似然旁路区分器原理为提取最大密钥信息量 $\mathrm{KI}(Z\mid\mathbf{X})$，并揭示 SCA 提取密钥信息量的过程即为

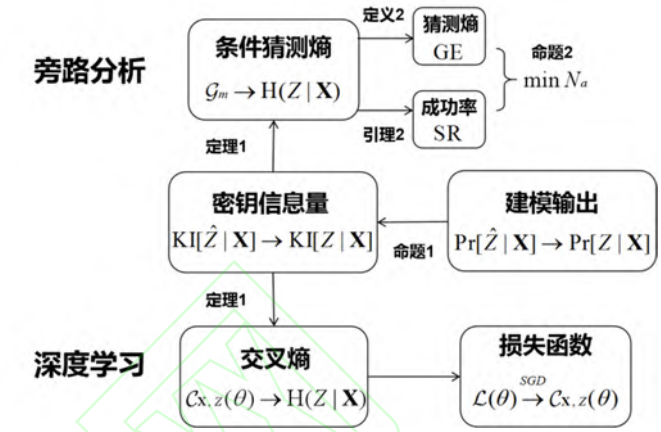


图 3　密钥信息量为核心的 DLSCA 安全评估框架

Fig.3　DLSCA security evaluation framework with the kernel of key information

条件猜测熵 $\mathcal{G}_m$ 优化过程，可与猜测熵 GE 建立联系；再通过引理 2 推导出密钥信息量与成功率 SR 的关系，且成功率也基于条件猜测熵的优化进行评估；最后通过命题 2 的安全评估将猜测熵与成功率与破密最小轨迹数 N_a 关联起来。根据定理 1（或定义 3）可将密钥信息量与交叉熵关联，并可由深度学习中交叉熵损失函数计算。综上由此以密钥信息量为核心的 DLSCA 安全评估框架建立，DNN 模型训练阶段的性能评估与测试阶段旁路安全评估得以关联。

4　实验验证

本节通过实验验证训练阶段 DNN 模型的交叉熵 $\mathcal{C}_{\mathrm{x},z}(\theta)$ 可通过上文提出的 DLSCA 安全评估框架与测试阶段的安全指标：猜测熵 GE 与成功率 SR 建立联系。实验数据采用 DPAcontest-v4[22]、AES-RD[23] 和 ASCAD[8]三个公共数据集，ASCAD 选取 AES 第一轮加密的第三个 S 盒，DPAcontest-v4 与 AES-RD 选取 AES 第一轮加密中一个 S 盒的密钥攻击。泄漏模型 Z 选择中间值模型，对应分类 256 种。成功率阈值皆设置为 $\alpha=0.9$。DNN 模型均采取[24]中对应结构，且统一设置 Adam 优化器[25]用于反向传播算法，初始权重随机初始化，损失函数为 Keras 库[26]提供的交叉熵损失函数[2]，所有实验均在配备 GPU Nvidia RTX 2080 的服务器上进行，训练与测试次数均取 $E=100$ 并做平均。

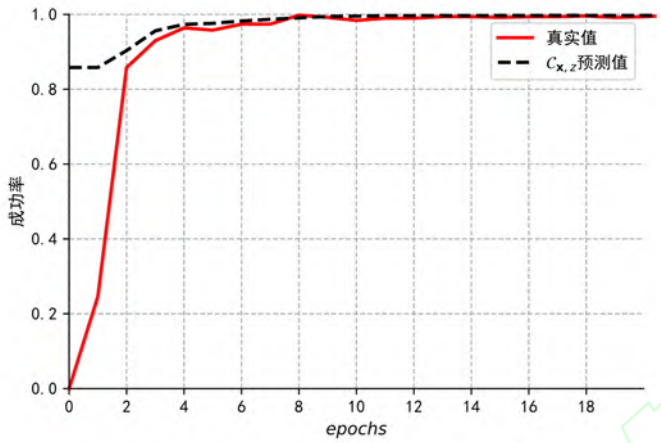
4.1　无防护——DPAcontest-v4 数据集验证

DPAcontest-v4 是基于 AES 软件实现的带有一阶掩码的数据集，令掩码已知，如式（22）本验证实验将其视为无防护数据集。训练数据集选定 N_p 为 4500，测试集 N_a 设置为 500。采取已知明文与固定密钥值，均选

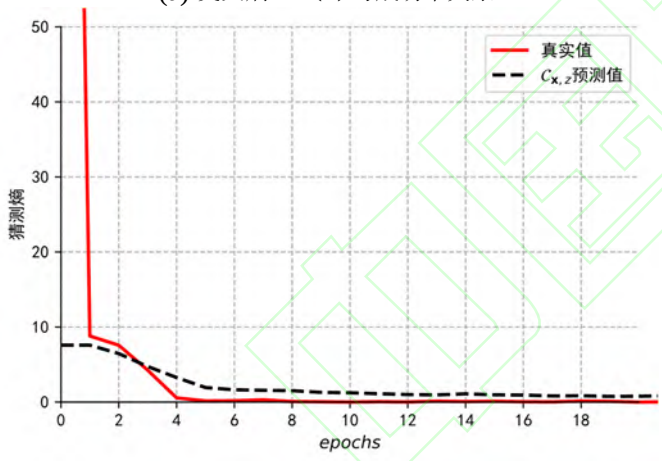
[2] Keras 中提供的交叉熵函数为常数底，根据本章内容需要做换底变换。

取第一字节。DNN 模型结构超参数选用[24]中用于训练 DPAcontest-v4 数据集的 DNN 结构。学习率设置为 10^{-3}，数据批次 batch 为 50，迭代周期为 20，测试阶段攻击次数 $E=100$ 并取平均。实验依据 DLSCA 安全评估框架中密钥信息量对 DNN 模型训练阶段的交叉熵 $\mathcal{C}_{\mathrm{x},z}(\theta)$ 与旁路安全指标建立的理论联系，采用交叉熵 $\mathcal{C}_{\mathrm{x},z}(\theta)$ 分别预测猜测熵 GE 、成功率 SR 。

$$Z_i(k^*)=\mathrm{Sbox}_1[p_i\oplus k^*]\oplus \underbrace{M}_{\text{已知掩码}} \tag{22}$$



(a) 交叉熵 $\mathcal{C}_{\mathrm{x},z}(\theta)$ 与成功率关系



(b) 交叉熵 $\mathcal{C}_{\mathrm{x},z}(\theta)$ 与猜测熵关系

图 4　DPAcontest-v4 数据集实验结果

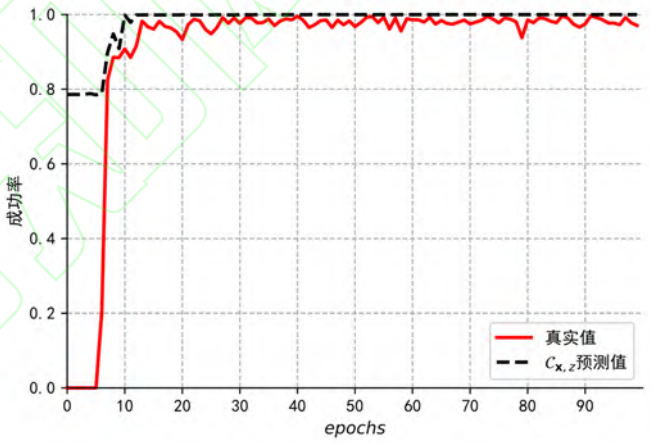
Fig.4　Result of DPAcontest-v4

如图 4(a)所示，交叉熵 $\mathcal{C}_{\mathrm{x},z}(\theta)$ 通过定理 1、引理 2 与密钥信息量和条件猜测熵的关系，可以对成功率进行预测，图中黑色虚线为 $\mathcal{C}_{\mathrm{x},z}(\theta)$ 预测值，红色实线为真实值。可以观测出，二者在经过 5 个迭代周期后，均稳定收敛并趋于 1；前 5 个迭代周期中，由于密钥信息量 $\mathrm{KI}(\hat{Z}\,|\,\mathbf{X})$ 不为 0，所以 $\mathcal{C}_{\mathrm{x},z}(\theta)$ 预测值并未从原点开始增长，而成功率 SR 根据成功轨迹数计算，因此从 0 增长，二者增长趋势一致。

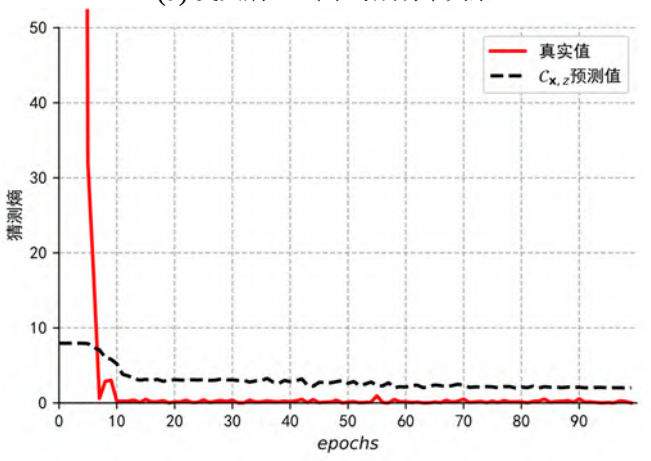
如图 4(b)所示，交叉熵 $\mathcal{C}_{\mathrm{x},z}(\theta)$ 通过定理 1、定义 2 与密钥信息量和条件猜测熵的关系，可以对猜测熵进行预测。可以观察出，二者在经过 5 个迭代周期后，均稳定收敛并趋于一定值。由于猜测熵 GE 在 python 计算时数组首位为 0，因此第 5 周期后其值稳定为 0，实际其与 $\mathcal{C}_{\mathrm{x},z}(\theta)$ 预测值均趋于同一定值 $\mathrm{H}(Z\,|\,\mathbf{X})$；在前 4 个迭代周期中，$\mathcal{C}_{\mathrm{x},z}(\theta)$ 预测值从上限 8 比特开始减少，而猜测熵 GE 从密钥空间上限值 $\mathcal{K}=256$ 开始减少，二者降低的趋势一致。综上，可验证以密钥信息量为核心的 DLSCA 安全评估框架适用于无防护的旁路分析场景，且根据安全评估框架中围绕密钥信息量的关系，通过训练阶段 DNN 模型的交叉熵 $\mathcal{C}_{\mathrm{x},z}(\theta)$ 可对测试阶段的安全指标：猜测熵 GE 与成功率 SR 建立联系。

4.2　带一阶掩码防护——ASCAD 数据集验证

本实验仅采用 ASCAD 偏移量为 0 的数据集，将其作为带一阶掩码防护的数据集。训练数据集 N_p 为 45000，测试集 N_a 均设置为 5000。采取已知明文与固定密钥值，均选取第 3 字节。DNN 模型结构超参数选用[24]中用于训练 ASCAD 数据集的 DNN 结构。学习率初值设置为 5×10^{-3}，根据优化器而择优变化。训练数据批次 batch 为 50，迭代周期为 100，测试阶段攻击次数为 $E=100$ 并取平均。



(a) 交叉熵 $\mathcal{C}_{\mathrm{x},z}(\theta)$ 与成功率关系



(b) 交叉熵 $\mathcal{C}_{\mathrm{x},z}(\theta)$ 与猜测熵关系

图 5　ASCAD 数据集实验结果

Fig.5　Result of ASCAD

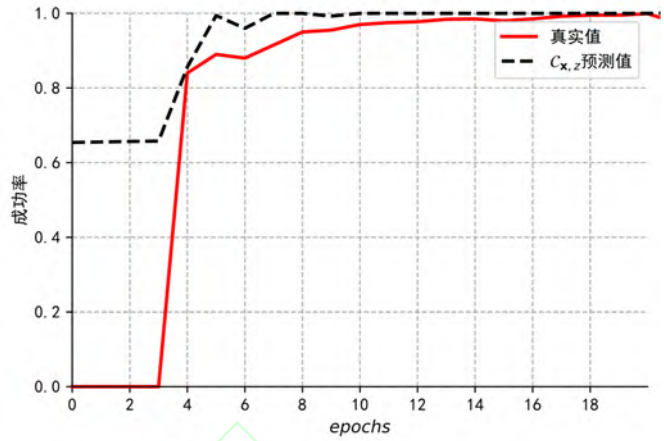
如图 5(a)所示，交叉熵 $\mathcal{C}_{\mathrm{x},z}(\theta)$ 通过定理 1、引理 2

与密钥信息量和条件猜测熵的关系，可以对成功率进行预测，图中黑色虚线为$\mathcal{C}_{\mathbf{X},Z}(\theta)$预测值，红色实线为真实值。可以观测出，二者在经过 11 个迭代周期后，均稳定收敛并趋于 1；前 11 个迭代周期中，同样由于密钥信息量 $\mathrm{KI}(\hat{Z}\mid\mathbf{X})$ 不为 0，所以$\mathcal{C}_{\mathbf{X},Z}(\theta)$预测值并未从原点开始增长，而成功率 SR 根据成功轨迹数计算，因此从 0 增长，二者增长趋势一致。

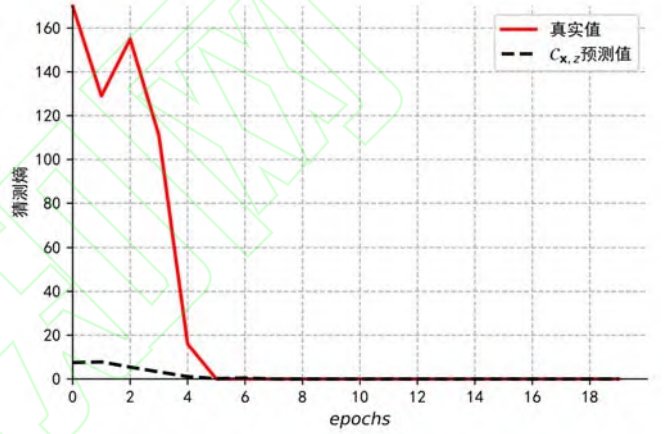
如图 5(b)所示，交叉熵$\mathcal{C}_{\mathbf{X},Z}(\theta)$通过定理 1、定义 2 与密钥信息量和条件猜测熵的关系，可以对猜测熵进行预测。可以观察出，二者在经过 11 个迭代周期后，均稳定收敛并趋于一定值。在前 11 个迭代周期中，$\mathcal{C}_{\mathbf{X},Z}(\theta)$预测值从上限 8 比特开始减少，而猜测熵 GE 从密钥空间上限值 $\mathcal{K}=256$ 开始减少，二者降低的趋势一致。综上，可验证以密钥信息量为核心的 DLSCA 安全评估框架适用于带一阶掩码防护的旁路分析场景，且根据安全评估框架中围绕密钥信息量的关系，通过训练阶段 DNN 模型的交叉熵$\mathcal{C}_{\mathbf{X},Z}(\theta)$可对测试阶段的安全指标：猜测熵 GE 与成功率 SR 建立联系。

4.3 随机时延——AES-RD 数据集验证

AES-RD 数据集采用的随机时延作为加密防护对策。训练数据集选定 N_p 为 25000，测试集 N_a 设置为 2000。采取已知明文与固定密钥值，均选取第 0 字节。DNN 模型结构超参数选用中[24]用于训练 AES-RD 数据集的 DNN 结构。学习率设置为 10^{-2}，数据批次 batch 为 50，迭代周期为 20，测试阶段攻击次数为 $E=100$ 并取平均。



(a) 交叉熵$\mathcal{C}_{\mathbf{X},Z}(\theta)$与成功率关系



(b) 交叉熵$\mathcal{C}_{\mathbf{X},Z}(\theta)$与成功率关系

图 6　AES-RD 数据集实验结果

Fig.6　Result of AES-RD

如图 6-a 所示，交叉熵$\mathcal{C}_{\mathbf{X},Z}(\theta)$通过定理 1、引理 2 与密钥信息量和条件猜测熵的关系，可以对成功率进行预测，图中黑色虚线为$\mathcal{C}_{\mathbf{X},Z}(\theta)$预测值，红色实线为真实值。可以观测出，二者在经过 11 个迭代周期后，均稳定收敛并趋于 1；前 5 个迭代周期中，$\mathcal{C}_{\mathbf{X},Z}(\theta)$预测值与成功率 SR 都保持初值不变，而后二者以相同趋势增长，在第 11 周期开始稳定收敛与 1。

如图 6-b 所示，交叉熵$\mathcal{C}_{\mathbf{X},Z}(\theta)$通过定理 1、定义 2 与密钥信息量和条件猜测熵的关系，可以对猜测熵进行预测。可以观察出，二者在经过 11 个迭代周期后，均稳定收敛并趋于一定值。在前 11 个迭代周期中，$\mathcal{C}_{\mathbf{X},Z}(\theta)$预测值从上限 8 比特开始减少，而猜测熵 GE 从密钥空间上限值 $\mathcal{K}=256$ 开始减少，二者降低的趋势一致。综上，可验证以密钥信息量为核心的 DLSCA 安全评估框架适用于带随机时延的旁路分析场景，且根据安全评估框架中围绕密钥信息量的关系，通过训练阶段 DNN 模型的交叉熵$\mathcal{C}_{\mathbf{X},Z}(\theta)$可对测试阶段的安全指标：猜测熵 GE 与成功率 SR 建立联系。

5　结论

本问通过信息熵角度：（1）把 DNN 作为旁路区

分器，将 DLSCA 整个过程视为一条马尔可夫链，指出准确率等机器学习性能指标无法评估 DLSCA 训练阶段的根本原因在于无法评估该过程中的密钥信息量；（2）围绕密钥信息量，分别将其与旁路安全指标以及 DNN 的交叉熵关联起来，并提出以密钥信息量为核心的 DLSCA 评估框架，建立 DNN 模型训练阶段的性能评估与测试阶段旁路安全评估之间的联系。通过选取无防护、带一阶掩码与随机时延的 AES 加密数据集进行实验，验证了以密钥信息量为核心的 DLSCA 安全评估框架适用于上述旁路分析场景，且根据安全评估框架中围绕密钥信息量的关系，通过训练阶段 DNN 模型的交叉熵 $\mathcal{C}_{x,z}(\theta)$ 可对测试阶段的安全指标建立联系。

参考文献：

[1] Kocher P, Jaffe J, Jun B. Differential power analysis[C]// Annual International Cryptology Conference. Springer, Berlin, Heidelberg, 1999: 388-397.

[2] Chari S, Rao J R, Rohatgi P. Template attacks[C]// International Workshop on Cryptographic Hardware and Embedded Systems. Springer, Berlin, Heidelberg, 2002: 128.

[3] Heuser A, Zohner M. Intelligent machine homicide[C]// International Workshop on Constructive Side-Channel Analysis and Secure Design. Springer, Berlin, Heidelberg, 2012: 249-264.

[4] Picek S, Heuser A, Jovic A, et al. Side-channel analysis and machine learning: A practical perspective[C]//2017 International Joint Conference on Neural Networks (IJCNN). IEEE, 2017: 4095-4102.

[5] Pal S K, Mitra S. Multilayer perceptron, fuzzy sets, classifiaction[J]. 1992.

[6] LeCun Y, Bottou L, Bengio Y, et al. Gradient-based learning applied to document recognition[J]. Proceedings of the IEEE, 1998, 86(11): 2278-2324.

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

[8] Benadjila R, Prouff E, Strullu R, et al. Study of deep learning techniques for side-channel analysis and introduction to ASCAD database[J]. ANSSI, France & CEA, LETI, MINATEC Campus, France. Online verfügbar unter https://eprint. iacr. org/2018/053. pdf, zuletzt geprüft am, 2018, 22: 2018.

[9] Kim J, Picek S, Heuser A, et al. Make some noise. unleashing the power of convolutional neural networks for profiled side-channel analysis[J]. IACR Transactions on Cryptographic Hardware and Embedded Systems, 2019: 148-179.

[10] Standaert F X, Malkin T G, Yung M. A unified framework for the analysis of side-channel key recovery attacks[C]//Annual international conference on the theory and applications of cryptographic techniques. Springer, Berlin, Heidelberg, 2009: 44461.

[11] Cagli E, Dumas C, Prouff E. Convolutional neural networks with data augmentation against jitter-based countermeasures[C]//International Conference on Cryptographic Hardware and Embedded Systems. Springer, Cham, 2017: 45-68.

[12] Picek S, Heuser A, Jovic A, et al. The curse of class imbalance and conflicting metrics with machine learning for side-channel evaluations[J]. 2018.

[13] Mangard S, Oswald E, Popp T. Power analysis attacks: Revealing the secrets of smart cards[M]//Springer Science & Business Media, 2008.

[14] Anderson R, Biham E, Knudsen L. Serpent: A proposal for the advanced encryption standard[J]//NIST AES Proposal, 1998, 174: 1-23.

[15] Li F, Yang Y. A loss function analysis for classification methods in text categorization[C]//the Proceedings of the 20th international conference on machine learning (ICML-03). 2003: 472-479.

[16] Zhang T. Solving large scale linear prediction problems using stochastic gradient descent algorithms[C]//Proceedings of the 21st international conference on Machine learning. 2004: 116.

[17] Hawkins D M. The problem of over fitting[J]. Journal of chemical information and computer sciences, 2004, 44(1): 1-12.

[18] Gray R M. Entropy and information theory[M]. Springer Science & Business Media, 2011.

[19] De Cherisey E, Guilley S, Rioul O, et al. An information-theoretic model for side-channel attacks in embedded hardware[C]// 2019 IEEE International Symposium on Information Theory (ISIT). IEEE, 2019: 310-315.

[20] Masure L, Dumas C, Prouff E. A comprehensive study of deep learning for side-channel analysis[J]//IACR Transactions on Cryptographic Hardware and Embedded
Systems, 2020: 348-375.

[21] Shwartz-Ziv R, Tishby N. Opening the black box of deep neural networks via information[J]. arXiv preprint arXiv:1703.00810, 2017.

[22] Bhasin S, Bruneau N, Danger J L, et al. Analysis and improvements of the DPA contest v4 implementation[C]// International Conference on Security, Privacy, and Applied Cryptography Engineering. Springer, Cham, 2014: 201-218.

[23] Coron J S, Kizhvatov I. An efficient method for random delay generation in embedded software[C]//International Workshop on Cryptographic Hardware and Embedded Systems. Springer, Berlin, Heidelberg, 2009: 156-170.

[24] Zaid G, Bossuet L, Habrard A, et al. Methodology for efficient CNN architectures in profiling attacks[J]//IACR Transactions on Cryptographic Hardware and Embedded Systems, 2020: 1-36.

[25] Kingma D P, Ba J. Adam: A method for stochastic optimization[J]. arXiv preprint arXiv:1412.6980, 2014.

[26] François Chollet et al. Keras. https://keras.io, 2015.