人工智能之于旁路分析

张 帆[1] 邵 彬[1] 谷大武[2]
[1] 浙江大学
[2] 上海交通大学

关键词：人工智能 旁路分析 密码学

旁路分析[1]是密码分析领域的一个重要分支，其目的在于研究密码算法在物理实现上的安全性。理论上被证明安全的密码算法会在诸如智能卡、密码芯片之类的嵌入式软硬件系统上实现，而这些设备在使用过程中会产生功耗、电磁辐射、时间、缓存访问命中与失效等物理可测量的旁路信息泄露。攻击者能够利用该泄露并结合密码算法本身的结构特征，恢复设备中所使用的密钥，此类密码分析方法被称为“旁路分析”。旁路分析技术对各类嵌入式智能设备的安全性造成了极大的威胁。随着内嵌密码模块的智能设备的普及以及物联网等领域的高速发展，旁路技术对于嵌入式系统安全性所造成的实际威胁将日趋严重。

人工智能技术作为一门发展在统计学基础上的综合性学科，在多个领域展现出了相较于传统技术的优势，尤其是在一些难以利用统计学方法构建合理的模型或是确定具体特征的场景，人工智能技术能够利用数据规模和算法的优势来近似描述这些模型或特征。另一方面，人工智能核心算法的实现依赖于物理设备，从旁路角度存在泄露敏感信息的风险。如何抵抗这种风险，也是旁路分析人员需要认真考虑的问题。

概述

旁路分析技术

传统意义上的旁路分析技术可以分为基于非建模 (Non-Profiling-based) 的分析方法和基于建模的 (Profiling-based) 的分析方法。非建模的分析方法主要包括差分功耗分析 (Differential Power Analysis, DPA)、相关性功耗分析 (Correlation Power Analysis, CPA) 等。该方法结合输入的明文信息以及可能的密钥候选值，根据领域知识建立加密设备运行过程中泄露信息的猜测模型，采集加密设备运行过程中实际产生的物理泄露信息，利用统计学的方法确定实际泄露信息与人为构建的泄露模型之间的相关性，进而确定加密过程中使用的密钥值。基于这一特性，非建模的分析方法在泄露信息采集过程中需要固定密钥，保持密钥的一致性，才能够准确地刻画出实际泄露信息和猜测模型之间的相关性。

以模板攻击 (Template Attack, TA) 为代表的基于建模的分析方法，其过程主要分为模板构建（建模）和模板匹配（攻击）两个阶段。在模板构建（建模）阶段，攻击者假设能够获得与目标加密设备相似的设备，称为模板设备。一般情况下，模板设备的型号与目标设备相同。攻击者通过在模板设备上实现与目标设备相同的加密算法，并通过配置不同的密钥、明文等参数获取足量的泄露信息，然后利用这些信息刻画不同的明文和密钥组合对应的中间值的泄露信息特征，构建对应的模板。在模板匹配（攻击）阶段，攻击者从目标设备上采集泄露信息，将其与构建的模板进行匹配，从而确定密钥的候选值。与非建模的分析方法相比，基于建模的分析方法可以对敏感操作构建模板。例如，模板攻击可以直接针

对密钥加载操作进行建模，并且在攻击阶段直接匹配对应的密钥值；当建模对象为特定中间值时，由于明文 - 密钥对能够确定所有可能的中间值，因此也能放宽在模板构建阶段对密钥一致性的要求。同时，模板攻击方法在攻击阶段往往难以从单条物理泄露中准确地恢复出密钥，因此会根据极大似然估计准则，在同一密钥加密场景下采集多条物理泄露信息，并结合每条泄露信息的密钥猜测结果，选择累计概率最高的密钥值。在此步骤中，恢复密钥所需要的最少泄露信息数量也能够作为衡量旁路攻击性能的重要指标。

针对以上旁路分析手段，智能设备生产厂商往往会采取不同的防护对策来降低或消除物理泄露信息和设备当前处理的数据之间的依赖性，进而实现对加密设备的防护。这些防护对策主要分为以下两类：

隐藏对策 其原理是使加密设备在不同时钟周期产生的物理泄露信息相等或是随机分布，从而消除泄露信息和处理数据的依赖关系。可以通过直接改变加密设备对于不同操作的功耗特征实现。

掩码对策 通过生成随机数与处理的中间值进行一定操作或运算（如异或运算），并将该运算结果作为后续处理的输入，执行后续的加密算法。由于设备产生的随机数对于攻击者来说是未知的，这样就消除了物理泄露信息和处理的中间值之间的依赖关系。同时，掩码对策可以在算法级上实现，而无须做硬件层面上的改动，减少了防护对策的实现成本。

人工智能技术

人工智能技术的定义十分宽泛，机器学习技术作为人工智能领域研究的一个重要分支，在图像、语音处理领域取得了广泛的应用。本文中提及与使用的人工智能技术基本上以机器学习技术为主。从学习形式上来讲，机器学习分为监督学习、无监督学习以及介于两者之间的半监督学习。监督学习中，作为学习对象的实例由样本数据与对应的期望输出（通常称为标签）两部分组成，算法通过学习大量的实例调整参数，从而完成特定的任务。无监督学习中实例仅由样本数据构成，需要算法自行寻找数据之间的差异。而半监督学习中，一部分样本具有标签，另一部分则没有标签，因而需要尽可能利用标签信息提高算法性能。监督学习目前的应用最广，相关研究也更为深入。根据监督学习过程中使用的算法，又可以将其分为支持向量机[2](Support Vector Machine, SVM)、随机森林[3](Random Forests, RF)、神经网络[4](Neural Networks, NNs)等。深度学习是一类基于多层神经网络的学习方法，它尝试模仿人类思维的认知过程并用神经网络模型进行表征。与其他方法相比，深度学习方法存在以下优势：(1) 能够自动提取数据中的特征，降低了特征工程的成本；(2) 在大规模数据下表现优秀。图 1 表示了上述提及内容的概念之间的关联性。

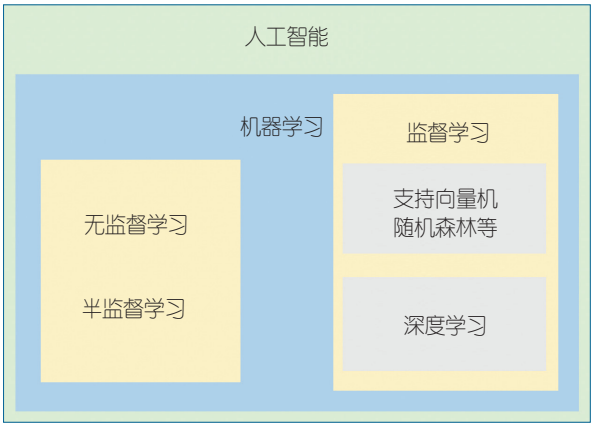

图 1 人工智能部分基本概念框图

技术研究方向

旁路分析领域中关于结合人工智能技术的研究目前集中在两个方面：一是基于人工智能的旁路分析技术，它将人工智能技术作为一种分析工具，研究如何利用人工智能算法提升传统旁路分析技术的效率。二是人工智能的物理实现安全，它将实现在嵌入式设备上的人工智能算法作为旁路分析的对象，利用旁路分析方法恢复出算法的敏感参数或是用户信息。

从基于人工智能的旁路分析技术这一角度来看，传统的基于统计学方法的旁路分析技术存在一

定的局限性：**(1) 对于泄露信息特征的刻画普遍使用多元高斯模型，不具备一般性；(2) 由于模板攻击需要计算多元高斯分布的协方差矩阵，对于高维数据，求解协方差矩阵需要大量的计算资源，因此往往难以实现；(3) 物理泄露信息的采样会存在时间上的偏移，导致无法准确刻画泄露的特征。**如何克服现有旁路分析技术的局限性成为当前该领域研究的重点和难点。

另一方面，广义的旁路分析的目的在于获取目标设备上有价值的信息，这里的信息并不局限于密钥，也包括设备上的秘密参数或是秘密代码。例如范·埃克窃听[5]能够根据电子设备发出的电子辐射，对键盘、显示器、打印机等进行监听，泄露关键信息。因此，从人工智能物理实现安全角度出发，目前人工智能设备的大规模落地依赖于其实现的物理平台（以各类嵌入式设备为主），如何利用旁路分析方法结合设备运行过程中产生的物理泄露信息恢复神经网络参数或是关键的输入信息，对人工智能算法的物理实现进行安全评估，并构建相应防护对策、保护核心数据不受泄露风险，在目前嵌入式人工智能领域火热的环境下，具有重大的社会和经济意义。

基于人工智能的旁路分析技术

当对现有技术框架上的算法改进难以突破技术上的局限性时，亟须引入新的技术为旁路分析领域研究注入活力。此时，建立在统计学基础上的机器学习领域的发展吸引了领域研究者的注意。从统计学角度来讲，旁路分析方法的本质是建立一个分类器，这个分类器的输入包括从加密设备上采集的物理泄露以及明文等信息，输出是设备中使用的密钥。例如，对基于建模的分析方法而言，攻击者在建模阶段掌握了模板设备的一切信息，包括输入的明文、加密的算法、算法使用的密钥、输出的密文信息等，并且根据这些信息以及相应的物理泄露信息构建出了对应的模板。将建模完成后的模板和相关信息视为一个黑盒，其输入是目标设备上的物理泄露信息，输出是该信息对应的不同模板的匹配概率。与机器学习技术中的监督学习技术进行对比，不难发现，上述两者在统计学方法上极为相似。图2展示了基于人工智能技术的旁路分析方法和传统的模板攻击分析方法的总体框架。对于非建模分析方法而言，攻击者掌握的信息包括输入的明文和加密算法，但是不包含密钥信息，攻击者需要构建一个不依赖于密钥信息的分类器实现对密钥的攻击，相当于一个无监督的分类任务。**因此，将人工智能技术应用在旁路分析领域在理论层面上是具备可行性的。**

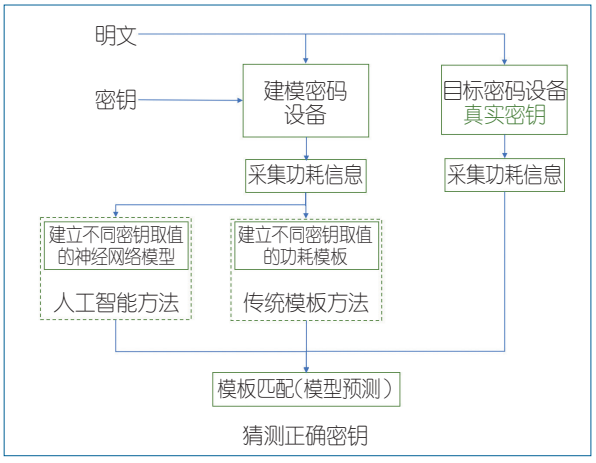

图2 基于人工智能技术的旁路分析方法与传统模板攻击方法示意图

与人工智能领域中将准确率作为评价模型的关键指标不同，旁路分析领域将**猜测熵**作为度量模型的指标。在旁路分析中，对于一些复杂的加密算法，攻击者往往需要一定数量的物理泄露曲线提供足够的信息来确定正确的密钥。考虑一种极端情况，在多条物理泄露曲线的分类结果中，正确密钥所对应的分类概率每次都在所有密钥候选值中排第二，即分类器的预测准确率为零，但是当计算密钥的累加概率（即猜测熵）时，仅需要数条物理泄露曲线就能确定最终的正确密钥。**因此，对基于人工智能技术的旁路分析方法而言，猜测熵及其相关指标能够更好地表征模型的优劣。**

技术发展及现状

在基于人工智能的旁路分析技术方面，考虑到

建模类旁路分析技术与有监督学习技术的相似性，领域研究者拟使用机器学习的方法代替基于统计学基础的模板攻击方法，从而**放宽对泄露信息分布的假设，克服传统建模类旁路分析技术的局限性**。将机器学习技术应用到旁路分析领域的早期工作[6~8]，主要使用支持向量机和随机森林等算法，对无防护和有防护的加密算法实现进行攻击，恢复加密算法的密钥。研究结果同时还证实了在某些特定场景（如电路噪声较大、泄露信息数量较少、泄露信息的数据维度较大等），基于机器学习的方法表现优于传统的统计分析方法。这些研究结果促使更多的研究者将目光投向了人工智能技术在旁路分析领域的应用。近年来，以卷积神经网络 (Convolutional Neural Networks, CNN) 为代表的深度学习技术在图像处理、语音识别等其他领域取得了丰硕的成果，研究者们尝试将其与旁路分析领域结合，以提高基于机器学习技术的旁路攻击的效率。

关于深度学习技术和旁路分析领域相结合的研究工作，最初发表在 SPACE 2016 上[9]。该研究通过分析基于深度学习的旁路攻击技术，并将其与 SVM、RF 等基于传统机器学习技术的旁路分析方法和基于统计学方法的模板攻击方法进行比较，得出了令人鼓舞的结论：深度学习技术在旁路分析领域的表现要优于传统技术，这给旁路分析领域的研究者们提供了更多的信心。

随后，在 2017 年的硬件安全顶会 CHES 上，一项基于 CNN 的旁路分析方法[10]克服了传统方法对于泄露信息时间上未对齐或是不同防护对策的场景下的缺陷，取得了优异的成果。一方面，得益于卷积网络的自动特征提取特性，该方法能够减少对物理泄露信息的预处理步骤；另一方面，神经网络模型代替了传统模板攻击中的统计学模板，消除了对特征高斯分布假设的依赖。

后续基于深度学习的旁路分析技术的突破，主要围绕神经网络模型的优化展开。2018 年，领域研究者提出了一套完备的基于深度学习的旁路分析框架和平台，同时还提出了一个基于 VGG-16 模型（一种深度为 16 的卷积网络模型）的变体，并在一阶掩码防护的 AES 加密算法的数据集上进行了实现，取得了较为理想的成果[11]。传统方法对具有掩码防护的加密算法的攻击往往会分成多个步骤，例如先恢复掩码值再恢复加密过程中的密钥，或是对密钥 - 掩码对进行恢复，而该 VGG-16 变体的神经网络能够直接从物理泄露信息中恢复出密钥而没有对掩码值做出任何假设。这项研究极大地推动了人工智能领域与旁路分析领域的结合。

在最新的工作中，研究者提出了一个简化的基于 VGG 模型的变体，已经能够在攻击阶段仅通过一次采样就恢复出关键的密钥信息，甚至在有随机延时防护对策的场景下依然保证算法的准确率[12]。

除了上述主流研究内容外，其他无监督或是半监督场景下的研究也是对该研究方向的重要补充。例如，在建模阶段样本数量受限的情况下，用半监督学习的方法攻击目标设备[13]。或是在没有标签的情况下，通过分析神经网络模型的收敛状态，推断正确的密钥值[14]：对于正确假设的密钥，神经网络模型能够快速收敛，对于错误假设的密钥，神经网络模型则无法收敛。同时，人工智能领域中常见的特征提取或数据增强技术[10, 15]也被应用到了旁路分析领域，以提高算法执行的效率。

研究趋势

从人工智能技术在旁路分析领域的发展历程可以看到，人工智能技术的发展与进步为旁路分析领域提供了可靠的分析工具，而旁路分析领域与深度学习技术的结合已经取得了显著的结果：仅仅需要一条物理泄露信息即可恢复对应的密钥。基于目前旁路分析领域的研究现状，今后人工智能在旁路分析领域的研究将主要集中在以下几个方面。

纵向的算法效率的改进

在算法的攻击效果已经无法取得显著改进的情况下，通过减少建模阶段所需训练数据的数量进而优化算法的效率。例如，在实际场景中使用的一些智能卡设备，由于内置计数器的存在，攻击者只能获取有限的旁路泄露信息，这对传统旁路分析工作的推进造成了一定的阻力。通过结合半监督学习的

方法，充分利用未标记的资源或是利用数据增强技术增加建模阶段使用的数据量以充分构建模型，也是可能的研究方向。

横向的算法在多场景的扩展

由于不同类型加密算法的结构特征不同，关键泄露信息的位置、构建模型的目标中间值、根据中间值和明文得到正确密钥的方式也不同，而且相同的神经网络模型在不同数据集上的表现也存在一定差距。在这些加密算法上，基于人工智能技术的旁路攻击方法的效果尚未得到评估，需要后续研究的支持。

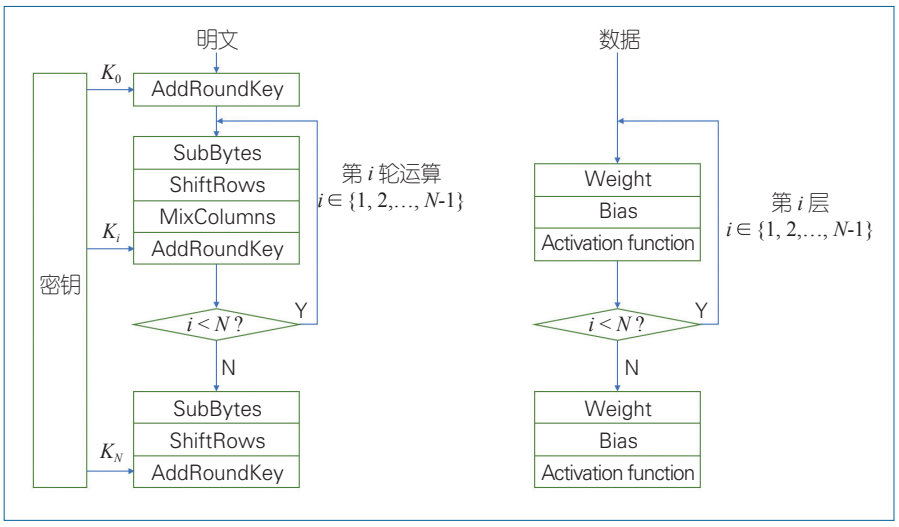

图 3　AES 算法（左）与神经网络（右）运算示意图

对于加密算法上实现的各种防护对策的研究不够深入。目前研究的主要对象是基于无防护或是具备一阶掩码防护的加密算法，而对于其他攻击难度更高、更复杂的防御对策防护下的加密算法尚未展开系统的研究和评估工作，有待开展进一步的研究。

目前基于人工智能技术的旁路分析方法仍需依赖一定的领域知识对旁路泄露信息的特征进行初步提取（从数十万个特征点中选择与敏感信息存在依赖性的数百个特征点），并以此构建模型。如何结合人工智能领域中的自动特征提取技术，代替传统旁路分析方法中的人为特征提取，从而提高旁路分析的自动化程度，是一个重要的研究方向。

人工智能硬件的物理实现安全

人工智能算法，特别是建立在神经网络基础上的各类深度学习算法，其核心在于通过迭代确定最终参数：在训练神经网络模型过程中，根据不同的训练数据进行迭代，寻找最优的网络参数；在神经网络推导过程中，根据输入的数据逐层地迭代数据，最后得出推测的数值，并依此分析预测结果。对于神经网络而言，参数决定了该网络能够实现的任务、预测的准确度等一系列指标。若商用的神经网络中的结构和参数遭到泄露，不法分子能够很容易地仿制相关的产品，给企业单位造成经济上的损失，这也是领域学者对人工智能的物理实现安全进行研究的一个重要原因。

密码算法与神经网络存在某些共性，使得我们能够利用旁路分析方法恢复神经网络的参数。以 AES 这一典型密码算法为例，根据算法中使用的密钥长度，AES 将整个加密过程分成了数轮，输入的明文信息经过多轮运算，转换为对应的密文信息。每一轮的运算由轮函数定义，轮函数是特定的四类运算的组合，除第一轮和最后一轮外，所有中间轮的运算组合是相同的。同时，每一轮的输出数据作为下一轮的输入数据，直到最后一轮产生对应的密文。AES 中每一轮所使用的某些运算操作与神经网络中神经元的计算也存在相似点：两者都有固定成分和随机成分参与运算。以 SubBytes 为例，该运算接受一个输入，并经由固定的映射得到对应的输出；而神经元通过线性运算和激活函数（在神经网络推导过程中，这两者是固定的），也能根据一个输入产生对应的输出。这以特点是我们能够利用旁路分析方法攻击神经网络的本质原因。图 3 展示了 AES 加密算法和神经网络模型的整体结构。

其中，AES 算法中的 K_i 是第 i 轮运算时根据密钥计算出的子密钥（轮密钥），N 表示轮运算的总次数，AddRoundKey、SubBytes、ShiftRows、MixColumns 代表轮函数中四类不同运算。神经网络中 N 表示网络的总层数，Weight、Bias、Activation function 则表示神经网络中神经元对应的三要素：权重、偏置以及激活函数。经过上述对比不难发现，人工智能尤其是深度学习算法的实现与加密算法的实现具有较高的相似度，利用旁路分析技术获取人工智能算法的输入或是逆向推断出敏感参数是有理可依的。

针对人工智能硬件的物理实现安全方面的研究目前正处于起步阶段。最早的针对神经网络的旁路分析攻击于 2018 年提出，该研究对一个在微控制器上实现的多层感知器模型 (Multi-Layer Perception, MLP) 的参数进行了逆向构建 [16]，恢复出了神经网络的层数以及神经元的信息。另一个研究团队在现场可编程门阵列 (FPGA) 的卷积神经网络特定实现上，针对第一个卷积层进行泄露信息的建模，并对其进行了攻击，恢复了输入图像的轮廓与像素信息 [17]。针对人工智能算法的旁路分析攻击虽然取得了一定的成果，但人工智能算法在微控制器或是 FPGA 上的实现与商用的 AI 芯片仍存在显著的差距。专用 AI 芯片会使用并行计算的框架加速算法的执行，芯片上的数据流对应的物理泄露信息相互耦合，进一步增加了旁路分析的难度。如何有效地从耦合信息中提取出有效的信息，刻画合理的特征泄露模型，将会是人工智能算法物理实现安全性研究的一个重点和难点。

总结

人工智能技术在旁路分析领域的发展方兴未艾。对于传统旁路分析技术难以解决的问题，可以结合人工智能技术改进相关的算法，提高分析的效率；对于传统旁路技术无法解决的问题，可以尝试使用人工智能技术构建黑盒，避开对应用场景的统计学假设，以期填补领域研究的空白。相信人工智能技术巨大的潜力能够成为旁路分析领域发展的强大助力。同时应警惕旁路分析技术对人工智能硬件可能造成的潜在威胁。 ■

张　帆

浙江大学副教授。浙江省千人计划入选者。主要研究方向为硬件安全、芯片实现安全、系统安全、人工智能安全、密码学、计算机体系结构等。

fanzhang@zju.edu.cn

邵　彬

浙江大学硕士研究生。主要研究方向为硬件安全、人工智能及其安全等。

shaobin_zju@zju.edu.cn

谷大武

CCF 杰出会员。上海交通大学教授。主要研究方向为密码学、软件与系统安全、硬件与系统分析、大数据和云安全、金融安全技术等。

dwgu@sjtu.edu.cn

参考文献

[1] Mangard S, Oswald E, Popp T. *Power analysis attacks - revealing the secrets of smart cards*[M]. Springer 2007, ISBN 978-0-387-30857-9, pp. I-XXIII, 1-337

[2] Suykens J, Vandewalle J. Least Squares Support Vector Machine Classifiers[J]. *Neural Processing Letters*, 1999, 9(3): 293-300

[3] Liaw A, Wiener M. Classification and regression by randomForest[J]. *R News*, 2002, 2: 18–22

[4] Specht D F. A general regression neural network[J]. *IEEE Trans. Neural Networks*, 1991, 2(6): 568-576

[5] Eck W V. Electromagnetic radiation from video display units: An eavesdropping risk?[J]. *Computers & Security*, 1985, 4(4): 269-286

[6] Hospodar G, Gierlichs B, Mulder E D, et al. Machine learning in side-channel analysis: a first study [J]. *Cryptographic Engineering*, 2011, 1(4): 293-302.

[7] Lerman L, Bontempi G, Markowitch O. Power analysis attack: an approach based on machine learning [J]. *International Journal of Advanced Computer Technology (IJACT)*, 2014, 3(2): 97-115.

更多参考文献：http://dl.ccf.org.cn/cccf/list