深度神经网络的实现安全性问题

杨博麟

摘要：本文主要讨论了深度神经网络的实现安全性问题。对此问题的背景，现状，发展前景做了简单的描述。现有工作证明，深度学习神经网络与密码学算法一样，在物理实现平台会遭受同样的实现安全性攻击如旁路攻击、故障攻击等等。而实现在密码学算法领域的防御手段也可以尝试运用在神经网络实现阶段。

# 背景

随着深度学习技术取得重大突破，人工智能成为研究和应用的热点，其在医疗、生物、金融、自动驾驶等领域都得到广泛应用并取得了丰硕的成果。然而，尽管社会大众对人工智能有很多美好的憧憬，但随着一批人工智能系统开始应用到真实世界中，其面临的安全问题也逐渐显现出来。如针对面部识别系统缺陷模仿受害者身份的非法认证危害，针对医疗数据、人物图片数据的隐私窃取危害，针对自动驾驶汽车、语音控制系统的恶意控制危害等等。这些案例的出现严重阻碍了人工智能系统的现实应用，越来越多的研究人员开始重视人工智能系统安全，工业界、政府、国防等部门都对该领域的发展给予了高度关注。事实上，人工智能安全问题已经引起各国的广泛关注。美国总统办公室发布了两份重要报告《为人工智能的未来做好准备》[1]和《国家人工智能研究与发展战略计划》[2]，前者审视了人工智能产业的现状、潜在的应用领域以及在社会公共政策上存在的问题，其中明确指出要“确保人工智能系统是可靠的”；后者对人工智能产业的研究与规划提出了战略目标，也强调了“确保人工智能系统的安全可靠”这一关键论点，具体来说，就是“需要确保系统以受控的方式安全和可靠地进行操作”

目前深度学习系统安全研究主要从三个方面开展：一是深度学习系统的物理实现安全，是指运行有深度学习算法或是配置有深度学习模型的物理设备或环境平台（诸如微控制器、FPGA和ASIC芯片、GPU等）在运行时产生的功耗、电磁、时间、声音等物理信息泄漏造成的安全问题。如泄漏信息同设备正在执行的操作存在一定的相关性，利用这些相关性可以恢复出执行操作的敏感参数，此类攻击被称为旁路攻击；外部环境的恶意改变（诸如激光辐射、时钟电压毛刺等）可以使得这些物理设备在运行时发生状态异常，此类异常既可以用来破外物理设备的正常工作，也可以用来恢复可能的敏感信息，此类攻击被称为故障攻击。

另外两个方面是深度学习系统代码安全和深度学习模型安全，在此不做详细介绍。

# 研究现状

Lejla Batina [1]等人进行了基于嵌入式深度学习模型的旁路分析(side channel analysis,SCA)工作。作者在Atmel AT-mega328P（8位微控制器）上实现并编译了几个训练好的神经网络（简单的多层感知器）。研究表明，利用基本的SCA技术，他们可以分析和提取激活函数（ReLu，sigmoid，tanh，softmax）的特性，恢复神经网络的权重，反向恢复神经元和层的数量等信息。同时，作者总结了SCA 中众所周知的保护方案（如混洗，掩码或恒定时间实现），这些保护方案能够为嵌入式深度学习系统的实现提供一定的安全指导。

Jakub Breier[2]等人通过用激光注入方式攻击了在AT-mega328P上实现的神经网络的激活函数，使该模型跳过了执行激活功能的指令，使得不同的激活函数产生了负面结果：ReLu（导致零功能，即无效神经元），sigmoid和tanh（水平翻转导致功能下降），最终导致深度学习模型的错误分类。

Yannan Liu, Lingxiao Wei[3] 等提出了两种针对深度学习模型的故障注入攻击：单偏置攻击（single bias attack, SBA）和梯度下降攻击（Gradient descent attack,GDA），只需要对深度学习模型参数进行微小改动即可实现错误分类。 SBA能够通过仅修改网络中的某个偏差项来实现错误分类。GDA通过分层搜索和修改压缩技术，使得整个故障注入攻击能够以更少的攻击次数达到足够理想的结果。实现更高的隐身性和效率。 它实现了对MNIST和CIFAR-10数据集的非常有效的攻击。

Weizhe Hua[4]等人提出了对硬件加速器上实现的CNN网络的逆向工程。通过向该CNN 网络提供输入并观察片外存储器的访问记录，结合时序和存储器访问记录以及存储器中的数据所提供的相关旁路信息推断神经网络的结构，进一步地恢复对应的权重等参数。该工作揭示了隐蔽片外存储器访问对于深度学习模型物理实现安全的重要性。

# 防御手段与展望

由于目前的攻击手段都是从传统针对密码学算法的实现攻击迁移过来的，相对应的防御手段也可以迁移过来。

针对旁路攻击，常用的防御手段有隐藏手段和掩码手段。隐藏是通过随机延时、噪声生成等手段将时间、功耗等旁路泄露中真正有用的信息隐藏起来，使攻击者通过简单的分析无法得出正确的私密信息。

而掩码手段则是在算法层面或硬件实现层面，添加一些模块，这些模块会使得电路中在处理的数据发生变化，但又保证输出的结果正确。此时单纯分析时间或功耗等泄露信息时同样无法简单的得到私密信息。

然而旁路攻击已经发展近20年，发展的过程就是攻击手段与防御手段进行博弈与互相促进进步的过程。所以并不能说这些手段就能完全保证实现安全性。只是在某个特性安全等级下的防护措施。

而针对故障攻击，较为常用除了添加探测故障注入手段的模块外，还有重复手段。由于注入故障后输出结果往往与正确结果不同，所以可以通过重复操作来判断是否有瞬时性故障产生。

综上所述，神经网络算法也会面临与加密算法相同的实现安全性问题。而这些分析手段与防御手段同样需要研究与发展。此外，很大可能会存在一些仅仅适用于神经网络算法的分析手段，这些都值得我们继续研究。

**参考文献：**

1. Lejla Batina, Shivam Bhasin, Dirmanto Jap, Stjepan Picek. CSI Neural Network: Using Side-channels to Recover Your Artificial Neural Network Information. CoRR abs/1810.09076 (2018)
2. Jakub Breier, Xiaolu Hou, Dirmanto Jap, Lei Ma, Shivam Bhasin, Yang Liu. Practical Fault Attack on Deep Neural Networks. CoRR abs/1806.05859 (2018)
3. Yannan Liu, Lingxiao Wei, Bo Luo, Qiang Xu. Fault injection attack on deep neural network. ICCAD 2017: 131-138
4. Weizhe Hua, Zhiru Zhang, G. Edward Suh. Reverse engineering convolutional neural networks through side-channel information leaks. DAC 2018: 4:1-4:6