第二次论文阅读

基于连贯性，**这次论文阅读报告主要介绍名为“DeepStrike: Remotely-Guided Fault Injection Attacks on DNN Accelerator in Cloud-FPGA”的论文。**该论文提出了一种基于电源侧信道攻击的方法，可以在云FPGA多租户环境下远程攻击深度神经网络（DNN）加速器，从而导致其推理结果出错。

与上次阅读的论文“2021-Remote Power Attacks on the Versatile Tensor Accelerator in Multi-Tenant FPGAs”论文的攻击模式类似，基于TDC结构监测FPGA上共享电源网络PDN。本篇论文也是针对多租户远程共用FPGA导致的安全攻击，但是非常突出的点在于其攻击的目的是使得深度神经网络（DNN）的推理结果出错，而不是逆向神经网络的结构——这在我看来是不合适的，基于模型窃取的攻击方式也许比侧信道方式更适合逆向神经网络结构，因为模型运用时总会提供给用户调用接口，无论是模型蒸馏还是逆向训练判别器-生成器的方式都已经表现出优秀的结果；而且神经网络结构的复杂性约束了泄露信息的价值，获得大致架构的方式在本文实验中被证明是可行的，再进一步结合模型窃取也许比直接逆向效果好许多，也能够减弱模型窃取方法中已知模型结构这一假设。

idea：

现实可行性：

在目前云计算火热的环境下，硬件虚拟化的实现以及商业情况的驱使下，部署模型在远程租用的服务器上，并且一台机器上有多个租户是非常普遍的现象。而AI加速器在定制化服务上逐渐表现出其优势。多租户远程共享FPGA在带来便利的同时，也衍生出新的安全威胁。

技术可行性：

多用户在同一FPGA上时尽管访问权限做了隔离，但是由于同处于一块片上，硬件资源在一定程度上是联通的，特别的近几年的论文都证明了基于FPGA配电网络（PDN）的攻击会泄露用户的信息，以及通过数字时间转化器（TDC）模块监测泄露信息的有效性。

模型架构如下：

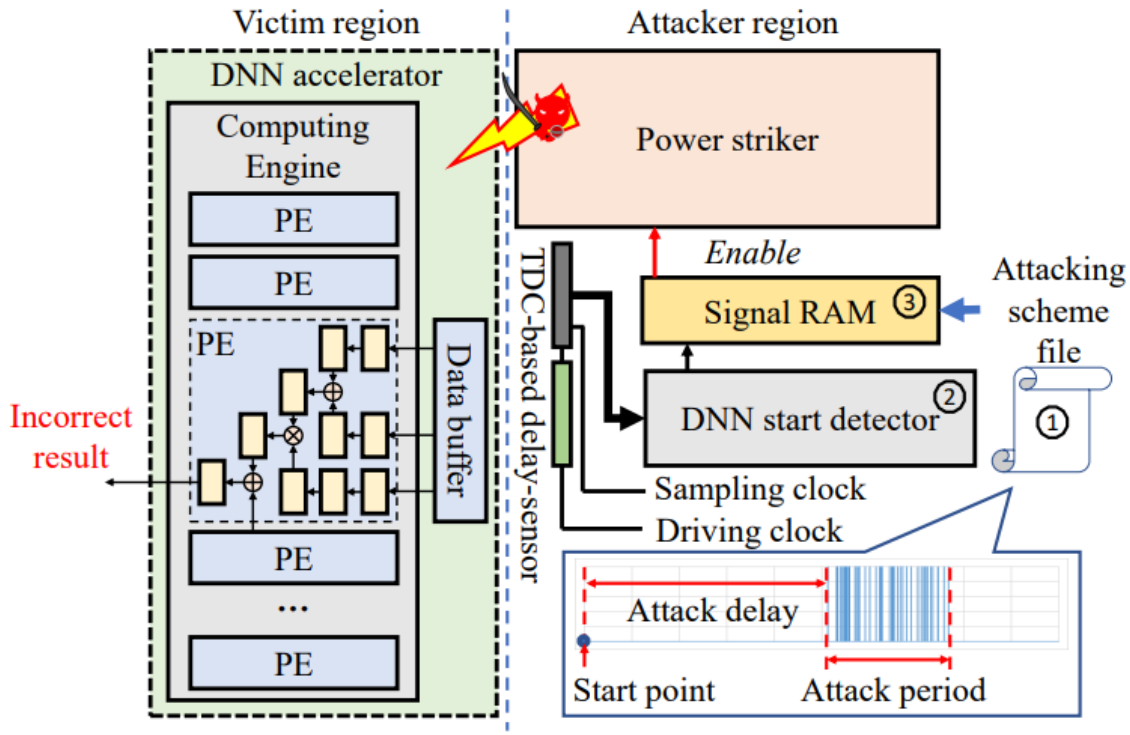

Fig. 4: Integrative schematic of DeepStrike.

本篇文章借鉴了前人的两项工作，通过TDC模块对于PDN侧信道的信息监测，实现DNN的运行进程的监控，然后通过新型耗电模块（power-strike）的计划性攻击，实现对于DNN网络中时序要求严格模块的攻击，导致其输出结果出错。

评价：

文章的新颖之处在于：

- 文章的假设很弱：之前的文章基本都假设对模型结构已知，或者至少知道参数存储的内存位置，从而实现对模型推理的攻击。但是这在现实中是不合适的，本文的假设几乎不需要知道模型的参数结构，虽然没法实现精细的参数调控翻转，但是其研究过程发现耗电攻击能够导致模型运行的时序错误，并且进一步通过理论论证以及实验证明了时序错误在DNN推理时的重要性。
- 文章的可操作性强：其在文章中提到许多工作设计的耗电模块在仿真时是可以运行的，但是在部署时会被禁止，他们改进了耗电模块的结构使其能够正常部署。此外其工作在TDC模块后加入了有限状态机的过滤模块，解决了电源变化失速问题，同时使得检测效果提升。
- 设计效果可编程化：设计了耗电模块的编程模块可以烧入RAM中，控制耗电模块的频率，这可以实现可控制的攻击。

meanings：

场景意义：

基于的多用户云FGPA共用的场景意义不再多说，其实验结果证明了基于TDC模块的对于DNN不同阶段的识别也就是建立模式识别库是有效的，可以说是之前“2021-Remote Power Attacks on the Versatile Tensor Accelerator in Multi-Tenant FPGAs”文章中模式识别阶段的补充实验证明。这表明通过侧信道逆向模型的大致结构信息是可行的。

研究方向意义：

其拥有弱假设和强实践性。在通过模型能够获得模型大致推理阶段的情况下几乎没有别的假设了，而这个假设在其实验中已经得到验证，也就是说其非常接近于实际。

在其未来工作中也提到未来会针对更加复杂的模型架构进行攻击，这是攻击者方面的方向。基于其实验中的子结论，时序结构对于模型推理非常重要，这可能会引出模型推理过程的保护工作以及提出不同于数据鲁棒性的安全鲁棒要求，涉及到加速器设计以及模型校验等工作。最后的实验中表明随机错误在耗电攻击次数增加后基本不变，而延时一周期的错误的比例不断上升从而导致错误率达到100%，因此可以推测大部分错误是执行后由于电源耗尽被取消，重新工作后时序出错，因此（以华为达芬奇加速器为例）在加速器设计过程中在标量控制器核加入执行完整性检验模块，重新调度指令执行，能够极大程度防御这种攻击，而随机错误比例部分引起对应的推理准确率下降几乎是可以忽略的。

论文写作上：

论文写作分为以下部分：现实意义介绍（包括场景选择，与自身创新点），背景介绍（包括相关工作的不足以及相关知识支撑），具体的攻击模型（分模块介绍了自己的工作细节），实验结果介绍（整体目标的实验效果以及效果分析），最后强调了自己工作的创新点以及未来工作。本文其实是相当细的领域工作，对于前人的TDC应用和耗电攻击的改进，在攻击思路上没有很新颖的地方，但是工作完整性很高，对于自身创新点强调很好，还对于实验结果进行了理论分析和进一步实验论证，对未来防御工作的有一定意义。查看其发表在A类中的DAC上，最近DAC收录了很多AI+FPGA设计的论文。