基于可信执行环境 TEE 的大模型保护技术

Li Kaihua

Oct 19, 2023

相关工作

Oblivious Join

- 基于 OP-TEE 与百度安全 Teaclave TrustZone SDK, 针对 memory-pattern 攻击, 实现 Spark Join 算子的安全计算。

Spark Aggregate Operator

- 华为合作项目，完成 Spark SQL Aggregate 相关算子在 iTrustee 环境上的迁移工作。

基于隐私计算的 Spark 金融智能风控系统

- 参加鲲鹏应用创新大赛2023机密计算赛道，目前入围决赛。

Paper Reading

- Opaque: An Oblivious and Encrypted Distributed Analytics Platform
- Oblivious Multi-Party Machine Learning on Trusted Processors

AIGC & TEE

保护目标

1. AI 大模型训练过程涉及大量隐私数据，数据资产需要保护。
2. 模型属于数字化资产，需要保护运行时的模型参数。

挑战

- TEE 目前难以利用外部的计算加速器，如 GPU、计算卡等。
- 大模型训练与推理属于资源依赖型计算，而 TEE 中内存、计算等资源有限。
- 大模型依赖与分布式的多机多卡计算框架，大模型拆分至 TEE 中导致较大的计算开销。

解决策略

思考

- 如何为 TEE 机密计算赋能 GPU 等加速器的计算能力，如何降低大模型在 TEE 中的计算负载?

方向

整体思路为开源节流，加速计算以开源，减少计算量以节流。

1. 加速计算:
   通过为 TEE 赋能 GPU 计算能力，提高 TEE 对于密集计算场景下的计算能力。
2. 减少计算量:
   权衡安全与性能，通过随机采样与模型拆分等方式，减少 TEE 需要承载的计算量。

加速计算

如何为 TEE 赋能 GPU 等加速器的计算能力?

应用层实现方式

通过混淆技术将复杂计算外包给不可信环境下的 GPU 加速器进行计算，并在可信环境下恢复计算结果。

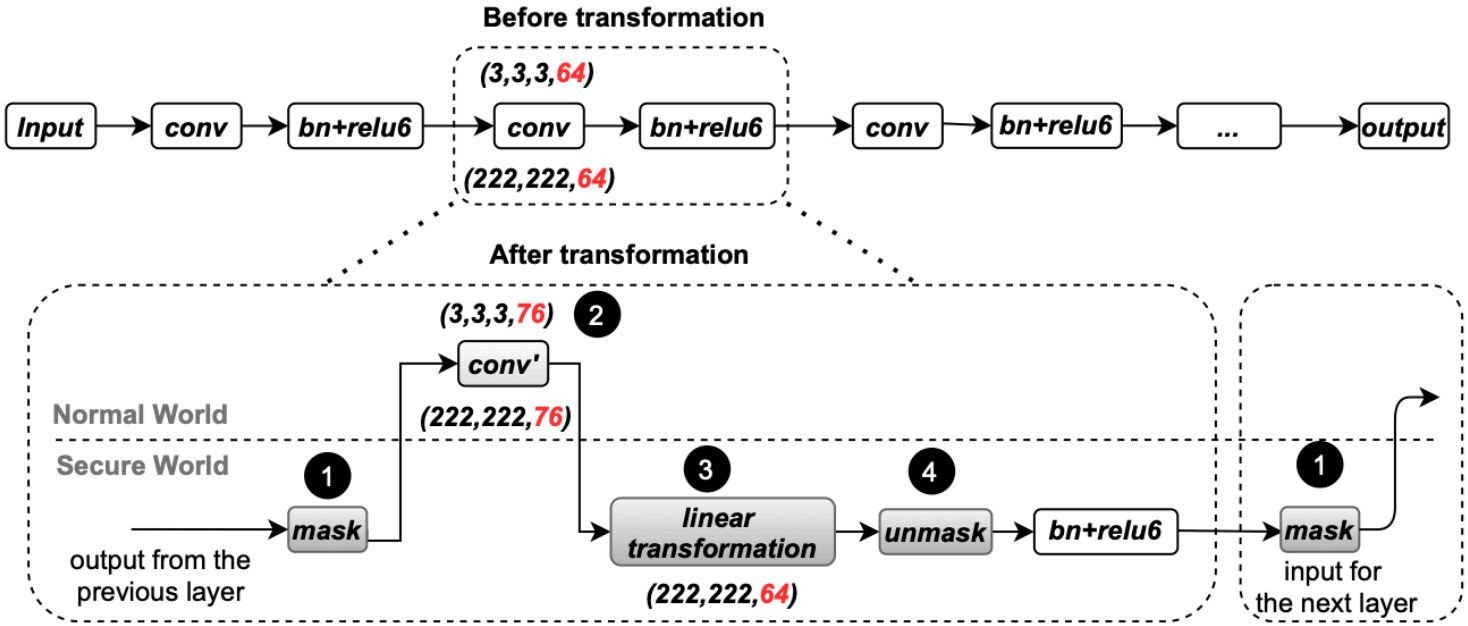
系统层实现方式

将 GPU 控制权从特权级操作系统隔离至 TEE 可信环境下，由 TEE 控制 GPU 的访问能力。

加速计算

应用层实现方式

通过混淆技术将复杂计算外包给不可信环境下的 GPU 加速器进行计算，并在可信环境下恢复计算结果。



加速计算

应用层实现方式

通过混淆技术将复杂计算外包给不可信环境下的 GPU 加速器进行计算，并在可信环境下恢复计算结果。

优点:

在应用层上进行改动，可在现有的 TEE 计算框架下进行。

缺点:

依赖于混淆技术，针对部分算子的特性进行特殊化适配，部分算子不适用于混淆技术。

加速计算

系统层实现方式

将 GPU 控制权从特权级操作系统隔离至 TEE 可信环境下，由 TEE 控制 GPU 的访问能力。

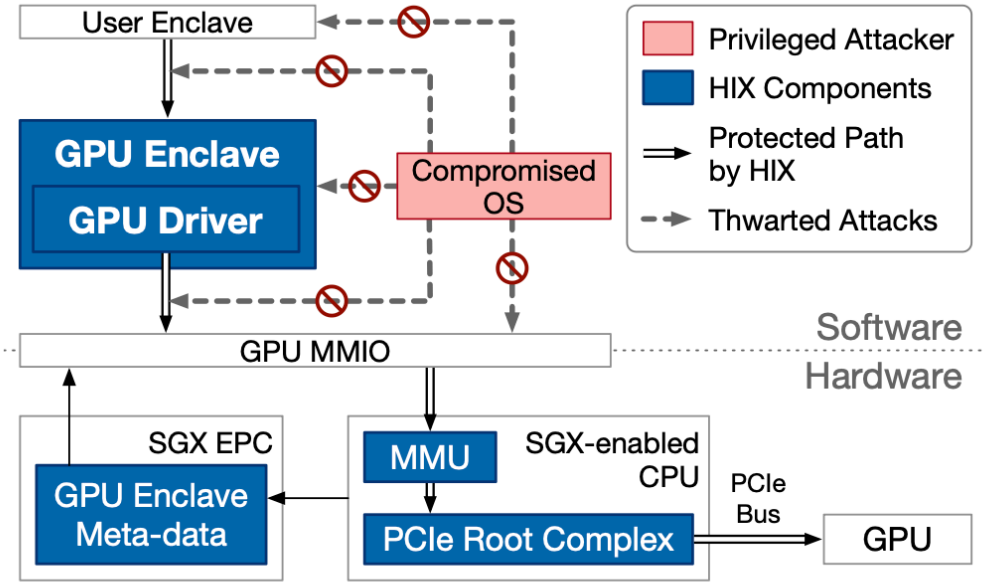


Figure 3. HIX architecture overview

加速计算

系统层实现方式

将 GPU 控制权从特权级操作系统隔离至 TEE 可信环境下，由 TEE 控制 GPU 的访问能力。

优点:

为安全环境提供普适的加速计算能力。

缺点:

需要进行系统层的改动，影响非安全环境下的 GPU 计算。

减少计算量

权衡安全与性能，通过随机采样与模型拆分等方式，减少 TEE 需要承载的计算量。

出发点:

大模型具有一体化特点，模型的精度依赖于各层网络，通过随机采样和模型拆分的方式，隔离部分算子的计算，实现可用性与安全性之间的平衡。

随机采样:

针对同一算子，按数据域拆分，保护部分的计算过程。

模型拆分:

针对不同算子，按模型拆分，保护部分的算子。

其他讨论

- 模型安全性的度量方法
- 分布式多机多卡环境下的安全保护技术
- GPU 机密计算技术
- …