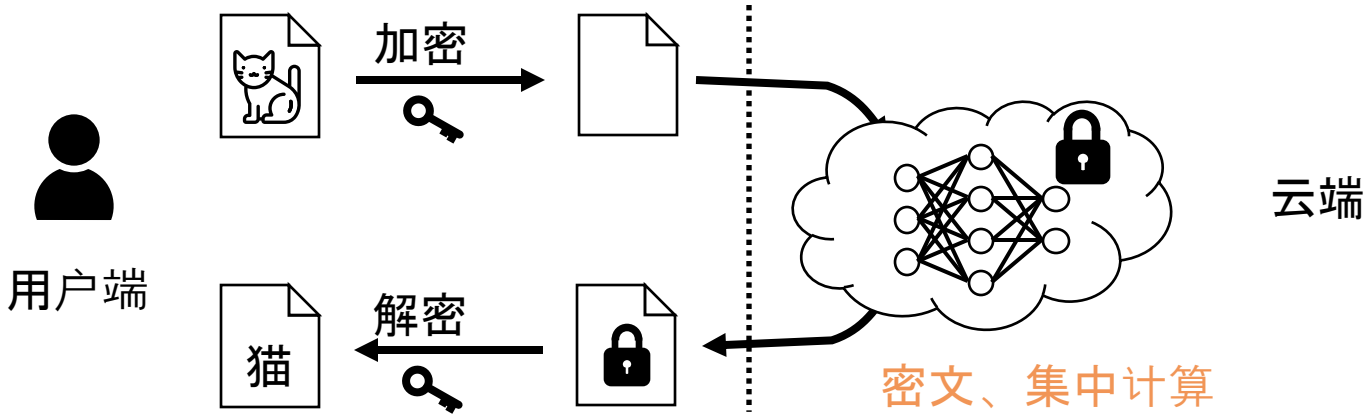
面向机器学习推理应用的全同态编译器框架 ANT-ACE

肖琳杰

程序设计语言与编译器实验室

蚂蚁技术研究院
Ant Research

全同态加密(FHE)



应用场景

云计算、金融数据分析、医疗数据处理，适用于需要在不泄露原始数据的情况下进行复杂计算的场景

优点

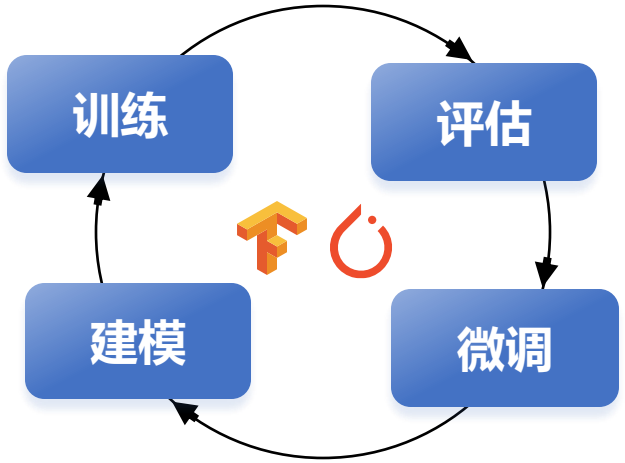
提供最强的隐私保护，适合高度敏感的数据处理

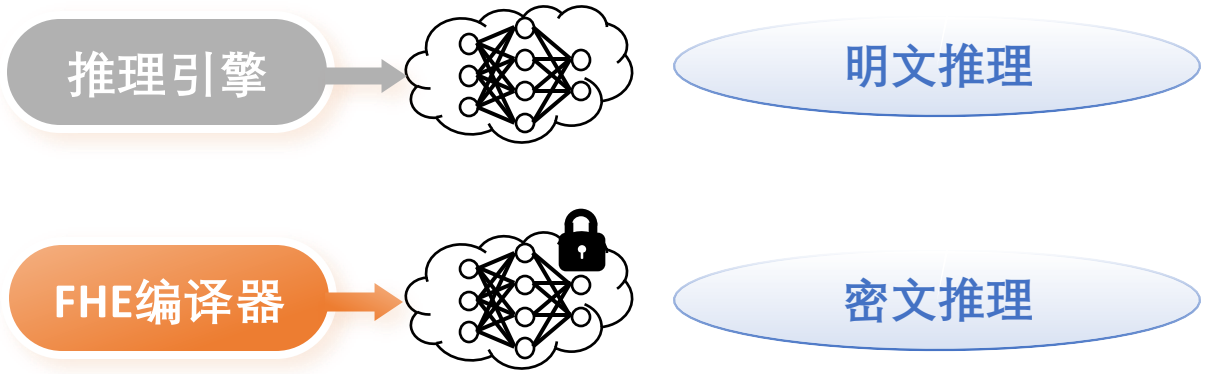
缺点

计算成本高，性能较低，尚未广泛应用于实际生产环境中

慢数百倍到数十万倍

编译器的作用域





训练阶段

预训练模型

推理阶段

功能特性

蚂蚁技术研究院 Ant Research

01 编译时

预训练模型FHE加密

支持**ONNX**， 可扩展PyTorch, TensorFlow等

支持**CKKS**, 可扩展BGV/BFV/TFHE等

支持**ANT-RTLIB**， 可扩展第三方加密库

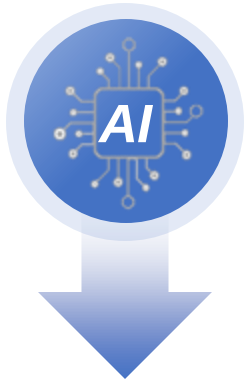
02 运行时

CPU、硬件加速器

支持模型**密文推理**

支持硬件**加速器(RISC-V架构)**接入

ResNet模型推理(FHE)实测数据



推理精度

128位安全性

CIFAR-10/100

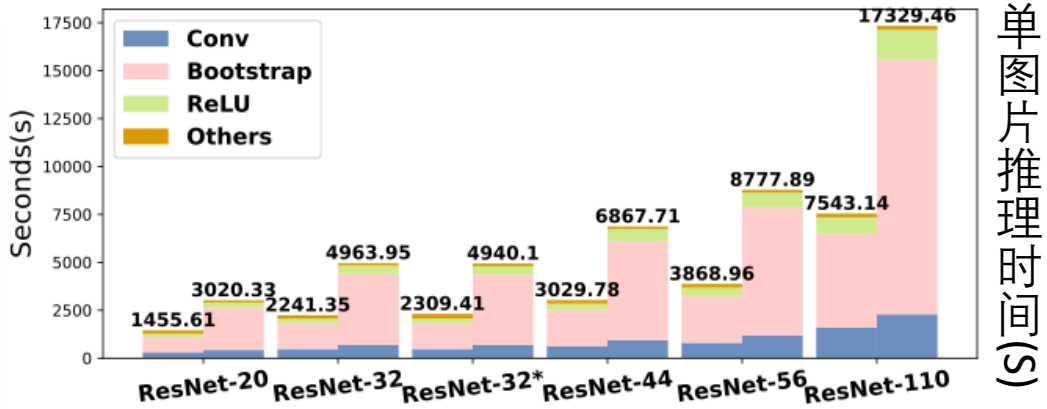
1000张图片推理

Model	Unencrypted	Encrypted	Accuracy Gain
ResNet-20	90.6%	91.0%	0.4%
ResNet-32	92.8%	93.5%	0.8%
ResNet-32*	66.4%	69.1%	4.1%
ResNet-44	92.5%	92.4%	-0.1%
ResNet-56	93.9%	94.8%	1.0%
ResNet-110	94.0%	93.3%	-0.7%

性能提升对比手写

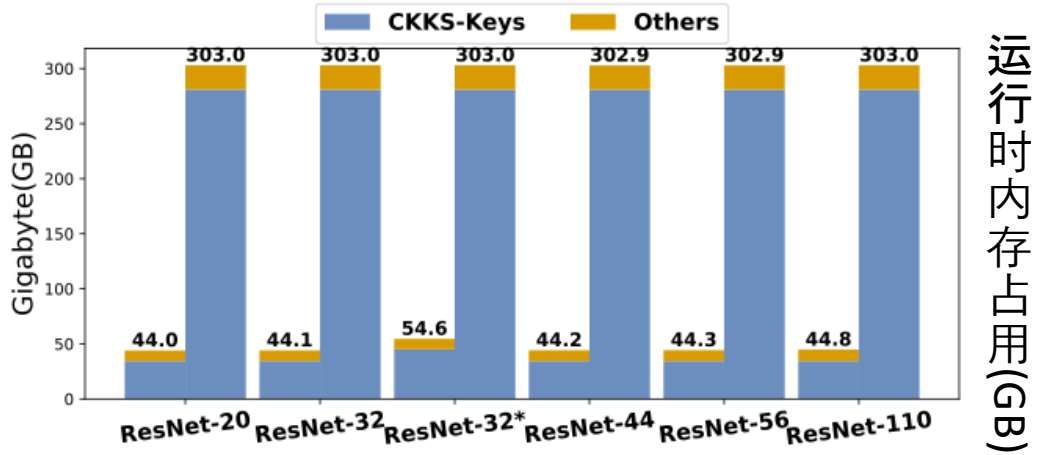
推理效率

内存使用



单图片推理时间(S)

ANT-ACE(左) FHE-MP-CNN(右)



运行时内存占用(GB)

ANT-ACE(左) FHE-MP-CNN(右)

* 对比FHE-MP-CNN(手写全同态加密程序)

发表于ICML'22，**目前唯一开源的CNN模型全同态推理项目**

5层IR架构设计

编译器前端

01 张量中间表示

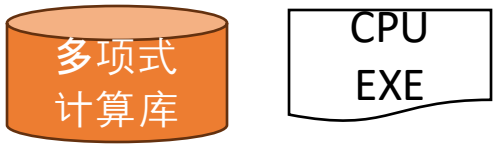
02 向量中间表示

03 方案无关同态中间表示

04 CKKS IR　Hybrid/BFV

05 多项式中间表示

为正确的分析和变换提供正确的IR
使优化更有效，避免模式匹配
允许使用更简单的算法，减少出错的可能性



专用指令加速多项式操作
扩展**RISC-V指令集**

组织与实现

计算图

性能优化

IR设计

逐层Lowering

ONNX2AIR， VECTOR, SIHE, CKKS, POLY, ...

算法实现

SSA, CFG, DFG, ...

加速器接入

核心算子

Base-Opcode, NN-Opcode, FHE-Opcode, ...

代码生成

CGIR， TargetInfo, Assembler, X86_64, RISC-V

AIR基础设施

支撑工具

Base模块

Container， node， Visitor,...

Util模块

MemoryPool, Option, Binary, Timing, Perf, unittest, benchmark, example, ...

软/硬件协同设计和优化调整的性能工具

蚂蚁技术研究院 Ant Research

中间表示层

数据流和控制流

内存层次的抽象建模

编译器优化设计与实验

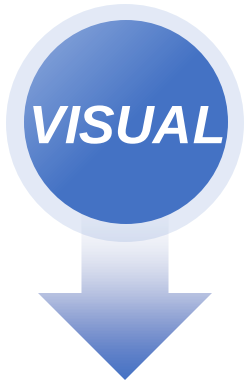
汇编层

Cycle评估

模型和轨迹分析

内存访问(类型和模式)

延迟和并行性



数据可视化

计算视图可视化

内存使用视图可视化

基线和单元测试集

单元测试、组合测试

集成测试、回归测试

性能测试、基线测试

生成目标文件和可执行程序

IR

中间表示

IR逐层Lowering

明文到密文转换

多项式IR生成

ASM

扩展RISC-V

机器表示

汇编代码生成

汇编代码生成

EXE

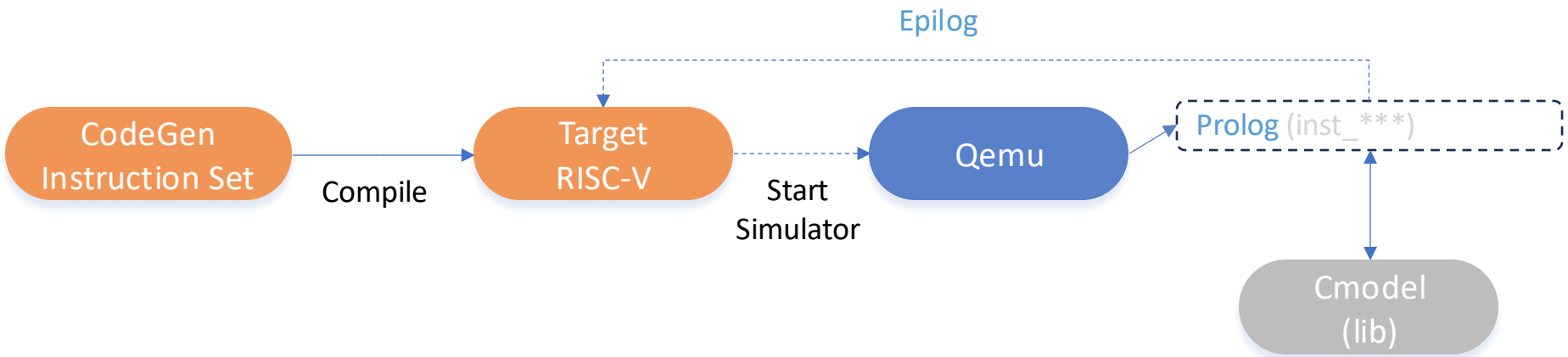
RISC-V可执行程序

编译目标文件

链接可执行程序

加速器程序生成

多项式计算仿真&验证



RISC-V扩展指令

多项式计算指令

验证计算结果

加速器程序

QEMU

内核态仿真

用户态仿真

仿真器

Cmodel

LSU(modadd, modsub, modmul,…)

NTT/AUTOU/TRASPOSE, …

多项式计算

示例：modadd计算


```
$ riscv64-unknown-elf-objdump -D qemu-hpu-cmodel/hello

qemu-hpu-cmodel/hello:     file format elf64-littleriscv


Disassembly of section .text:

0000000080000000 <_start>:
    80000000:   f14022f3            csrr    t0,mhartid
    80000004:   00029863            bnez    t0,80000014 <halt>
    80000008:   00009117            auipc   sp,0x9
    8000000c:   ff810113            addi    sp,sp,-8 # 80009000 <stack_top>
    80000010:   00c0006f            j   8000001c <main>

0000000080000014 <halt>:
    80000014:   a001                    j   80000014 <halt>

0000000080000016 <modadd>:
    80000016:   00c5850b            0xc5850b
    8000001a:   8082                    ret

000000008000001c <main>:
    8000001c:   1101                    addi    sp,sp,-32
    8000001e:   478d                    li  a5,3
    80000020:   c23e                    sw  a5,4(sp)
    80000022:   4791                    li  a5,4
    80000024:   c43e                    sw  a5,8(sp)
    80000026:   0030                    addi    a2,sp,8
    80000028:   4795                    li  a5,5
    8000002a:   004c                    addi    a1,sp,4
    8000002c:   0068                    addi    a0,sp,12
    8000002e:   ec06                    sd  ra,24(sp)
    80000030:   c63e                    sw  a5,12(sp)
    80000032:   fe5ff0ef            jal ra,80000016 <modadd>
    80000036:   a001                    j   80000036 <main+0x1a>

Disassembly of section .data:

0000000080000038 <stack_top-0x8fc8>:
    ...
```

- Run on X86_64

```
# time ./build/test_fhe_modadd
fhe_modadd test, pass!

real  0m7.647s
user  0m7.591s
sys 0m0.015s
```

- Run on Qemu(RISC-V)

```
# time ./build/test_fhe_modadd
fhe_modadd test, pass!

real  1m47.351s
user  1m20.835s
sys 0m26.491s
```

蚂蚁技术研究院
Ant Research

谢谢大家！

业务场景接入

隐私保护与数据安全

符合监管要求

提高数据利用率

AI推理性能优化

开放心态

项目开源

支持算法库接入

支持加速器接入

支持DSL接入

合作共赢

研究机构

高校合作

企业对接