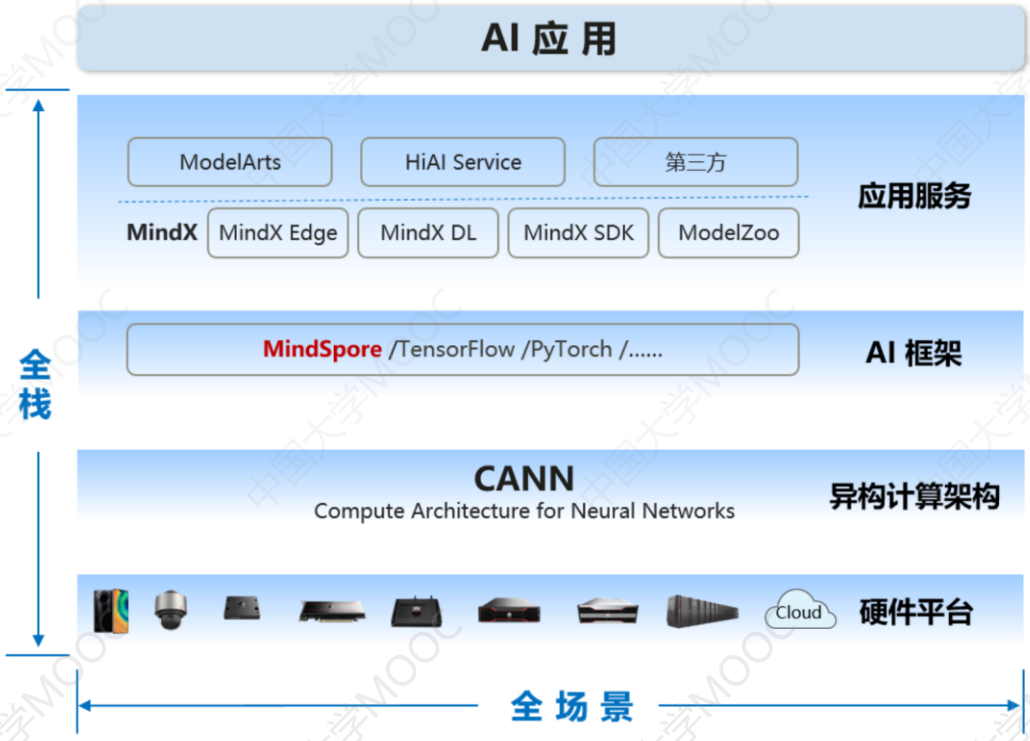
人工智能芯片与框架介绍

助教

浙江大学计算机学院

华为全栈AI解决方案



应用

提供全流程服务、分层APIs、及预集成的方案。

MindSpore

最佳匹配昇腾算力的全场景AI计算框架。

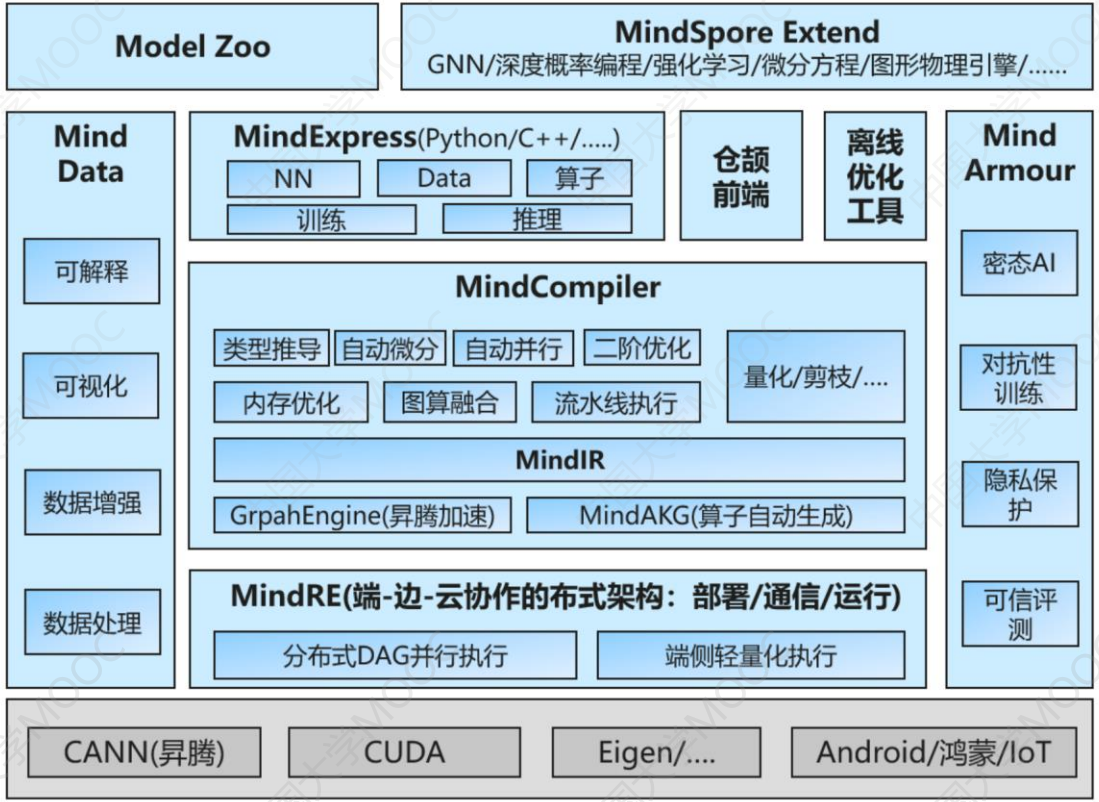
CANN

芯片算子库和高度自动化算子开发工具。

昇腾芯片

基于统一、可扩展架构的系列化 AI IP 和芯片。

MindSpore逻辑架构



设计目标：

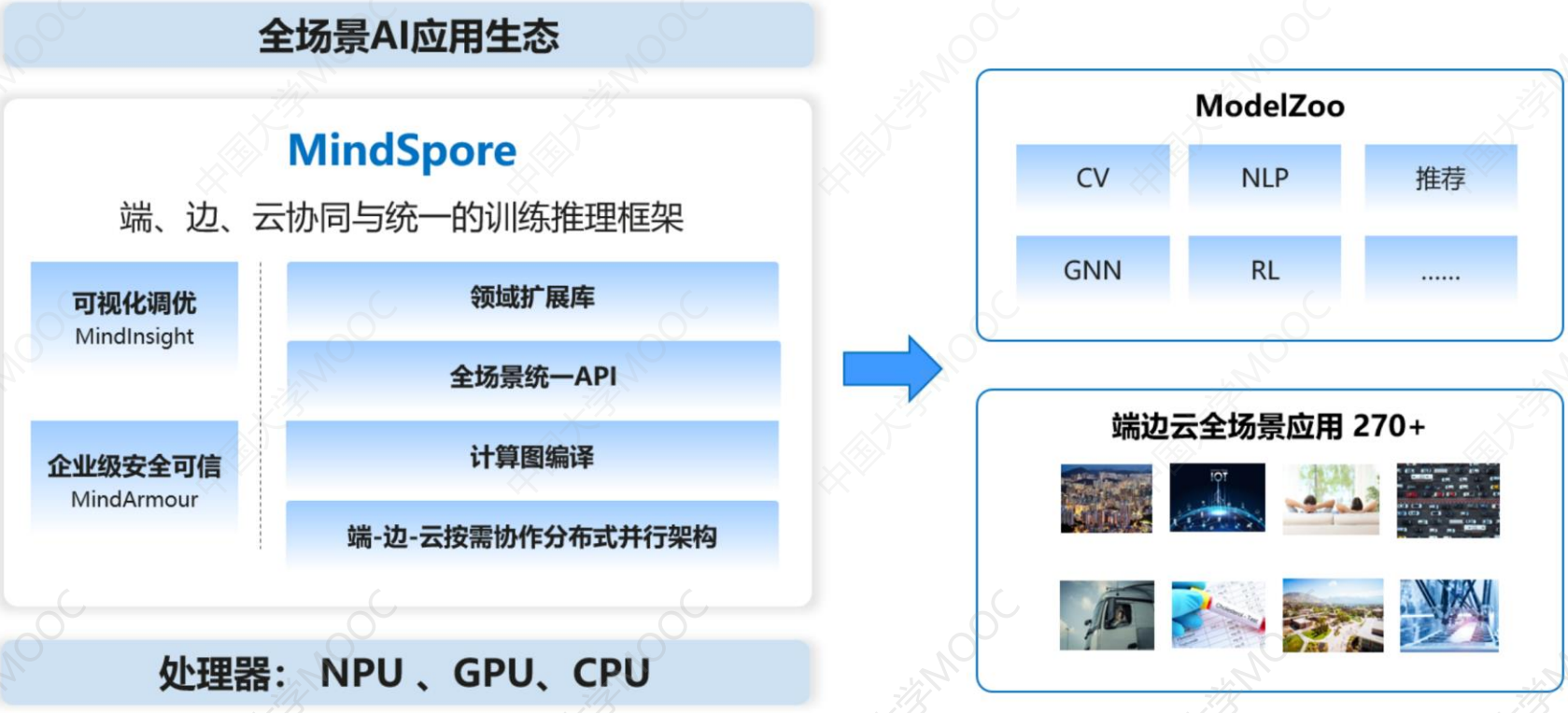
- **三层（表达/编译/运行）**：三层解耦开放，实现多前端/跨芯片/跨平台的开放架构。
- **三面**：计算、数据(MindData)、可信（MindArmour）。
- **全场景统一架构**：统一IR和API，AI应用可平滑流动。

关键技术：

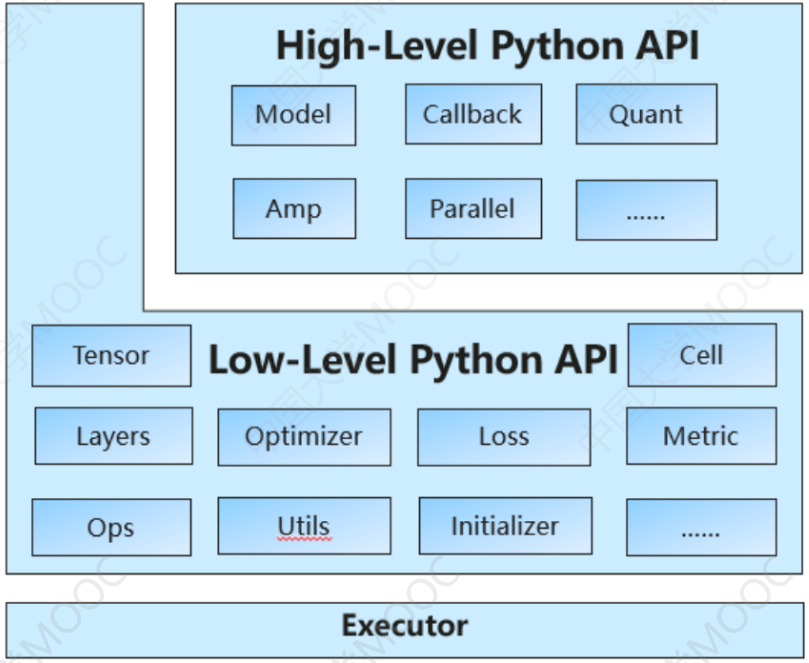
(分布式并行/编程语言/编译器/数据处理/算法/安全可信)

- **分布式并行**：自动并行、内存约束编程、DAG并行执行、数据缓存和卸载加速。
- **二阶优化**：利用目标函数的二阶导数加速收敛，收敛速度提升20%。
- **图算融合**：深度图优化、图算联合优化、自动算子融合。
- **企业级可信**：对抗性训练、差分隐私、密态AI、可解释AI。

ModelZoo


全场景AI应用生态
MindSpore
端、边、云协同与统一的训练推理框架
可视化调优
MindInsight
企业级安全可信
MindArmour
领域扩展库
全场景统一API
计算图编译
端-边-云按需协作分布式并行架构
处理器：NPU 、GPU、CPU
ModelZoo
CV
NLP
推荐
GNN
RL
.....
端边云全场景应用 270+

MindExpress子系统



设计目标：

- 两层用户API设计（包括High-Level与Low-Level），支撑用户进行网络构建、整图执行、子图执行以及单算子执行。
- 向用户提供统一的模型训练、推理和导出等接口，满足端、边、云等不同场景。
- 动态图和静态图统一的编码方式。
- 单机和分布式训练统一的编码方式。

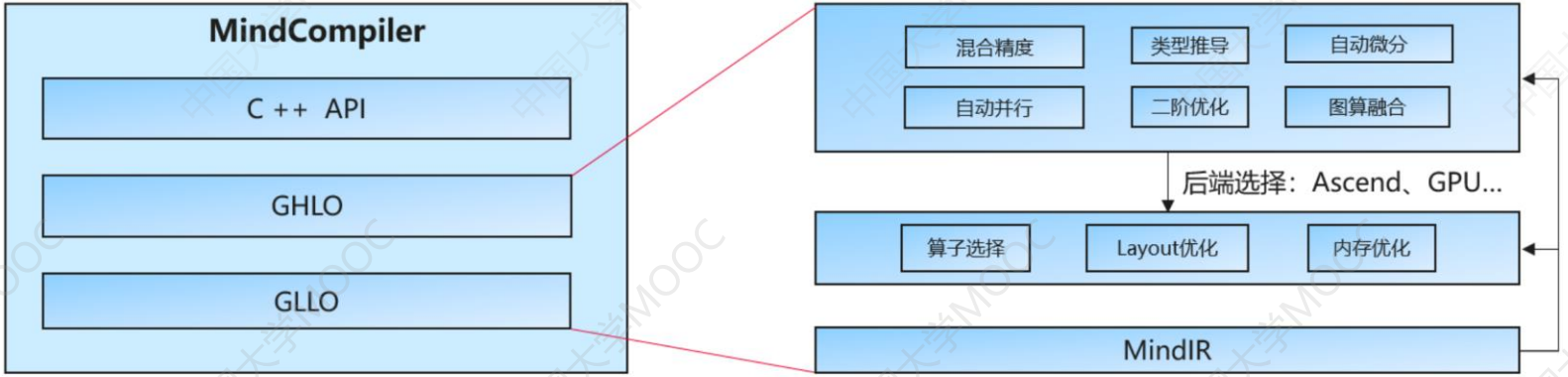
功能模块：

- High-Level API提供训练推理的管理接口、Callback、量化、混合精度、并行等控制接口，易于用户实现整网流程的控制。
- Low-Level API提供基础的Tensor、Cell、NN-Layers、优化器、初始化等，易于用户灵活构建网络和控制执行流程。
- Executor提供计算的执行控制，与MindSpore backend交互。

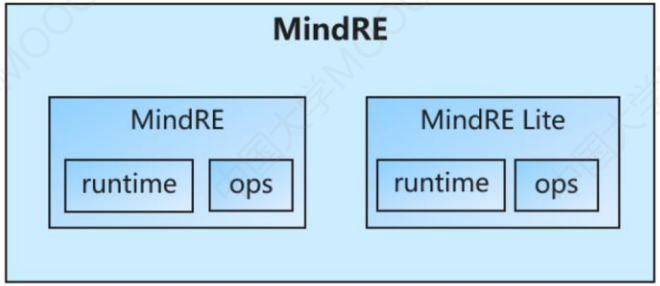
MindCompiler子系统

MindCompiler提供面向MindIR的图级即时编译能力：

- Graph High Level Optimization (GHLO)面向应用，进行偏前端的优化和功能，如类型推倒、自动微分、二阶优化、自动并行等。
- Graph Low Level Optimization (GLLO)面向硬件，进行偏底层的优化，如算子融合、layout优化、冗余消除、内存优化等。



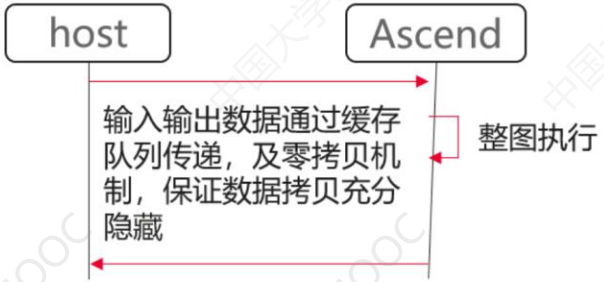
MindRE子系统



统一的运行时系统：

- 支持端、云多种设备形态要求。
- 支持多种硬件设置的调度管理，如Ascend、GPU、CPU。
- 内存池化管理，高效内存复用。
- 算子异步、异构执行，多流并发。

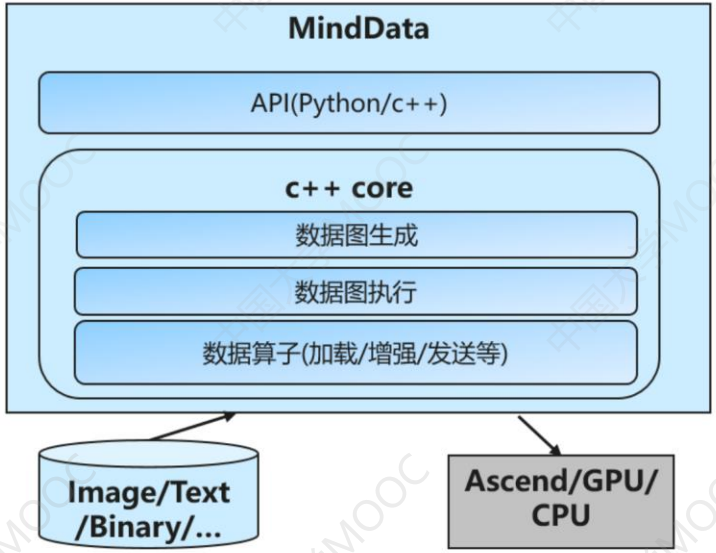
特色技术



MindData子系统

MindData负责高效执行训练数据处理pipeline，与计算形成流水，数据及时导入训练。

典型训练数据处理pipeline



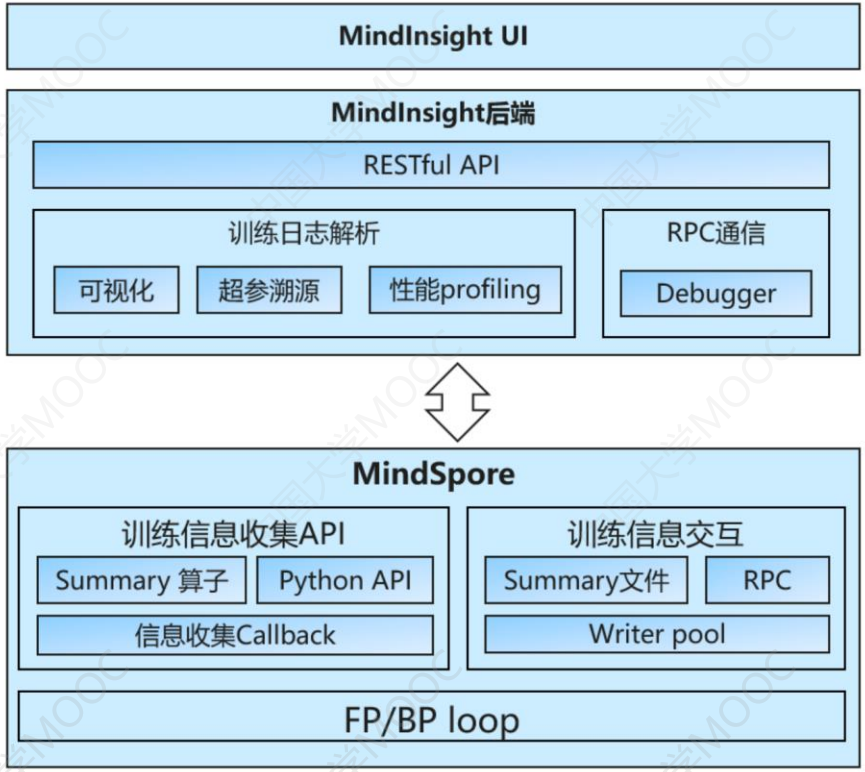
关键功能：

- 流水线+并行方式执行，提高数据处理吞吐量。
- 丰富的数据算子。
- 自定义Python算子，灵活定制pipeline（数据加载、采样、增强等）。
- 异构硬件加速(Ascend/GPU/CPU)。
- MindRecord：自带元数据、聚合存储。

运行流程：

1. 数据图生成：根据用户的Python API调用，生成数据图.。
2. 数据图执行：Pipeline并行执行数据图中的数据算子完成数据集加载、shuffle、数据增强、batch等处理。
3. 数据导入Device：处理后数据导入Device训练。

MindInsight子系统



MindInsight是调试调优子系统，提供训练过程可视化、模型溯源、debugger和性能profiling功能。

关键功能：

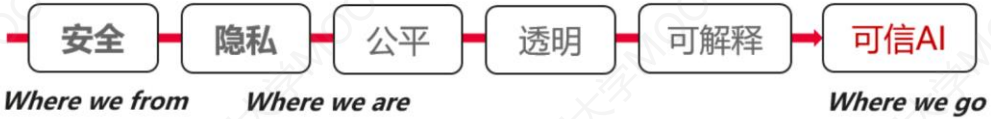
- 易用的API接口，在训练过程中，用户可以方便的收集训练过程指标，包括计算图、标量数据(loss/accuracy...）、直方图数据(梯度/权重...)、性能数据等，并通过Web UI界面进行展示。
- 通过收集训练的超参，数据集、数据增强信息实现模型溯源，并可在多次训练间进行对比。

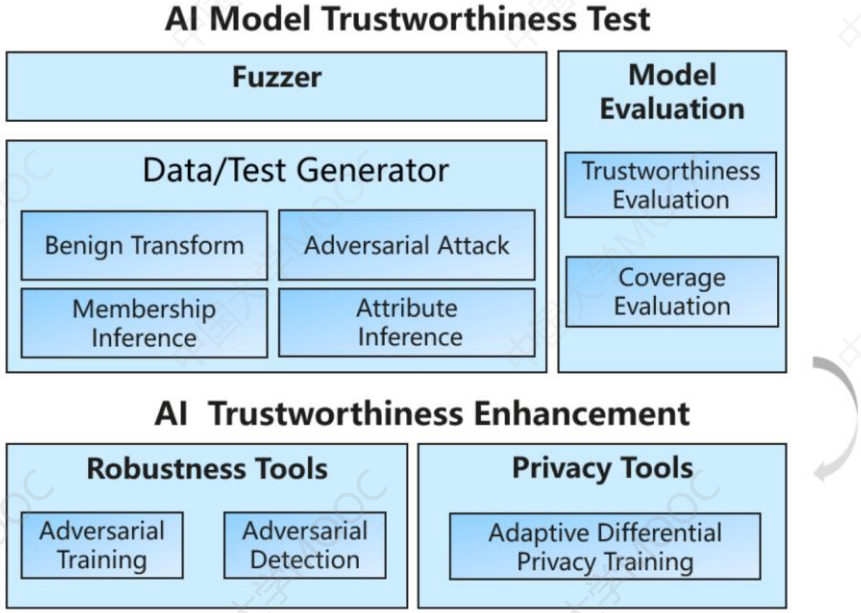
运行流程：

1. 训练信息收集：用户可通过callback接口，收集常用训练指标。用户也可以按需要收集自定义信息，如通过summary算子收集计算图中信息，通过Python接口收集Python层信息。
2. 训练日志生成：用户在训练过程中收集到的过程信息，最终会生成训练日志。
3. 训练信息展示：MindInsight通过打开并解析训练日志，以图形化方式为用户展示训练过程信息。

MindArmour子系统

MindArmour针对可信AI的各个领域提供全面、有效、易用的评测工具和增强方法。





关键功能：

- 涵盖黑白盒对抗攻击、成员/属性推理攻击、数据漂移等测试数据产生方法，覆盖场景全面；
- 基于覆盖率的Fuzzing测试流程，灵活可定制的测试策略和指标；
- 包括对抗训练、输入重建在内的常见对抗样本检测和模型鲁棒性增强方法；
- 高效自适应差分隐私训练和预算统计算法，数学上可证明的模型隐私泄露约束；

运行流程：

1. 配置策略：根据威胁向量、可信需求定义测试策略，选择合适的测试数据产生方法；
2. Fuzzing执行：根据模型覆盖率和配置策略启发式地产生可信测试数据；
3. 产生评估报告：可以基于自带的或自定义的可信指标；
4. 可信增强：使用预置的方法增强AI模型可信程度。

高效并行模式

挑战：

超大模型与超大数据集的分布式训练，需要通过数据并行+模型并行的混合并行方式，才能高效训练网络；

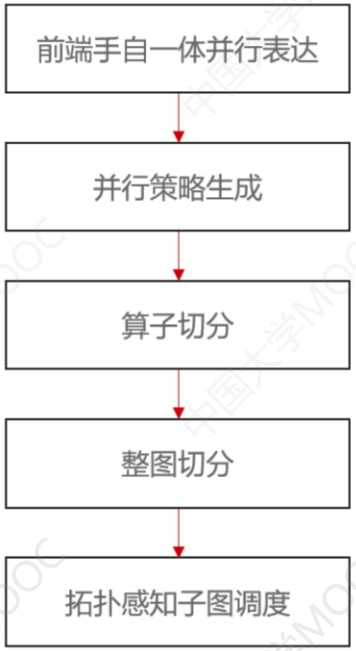
高性能：

- 传统graph-level模型切分，计算资源利用率不高，需要通过operator-level模型切分提高并行加速比。
- 选择一种高效的模型切分方式需要专家经验。

易用：

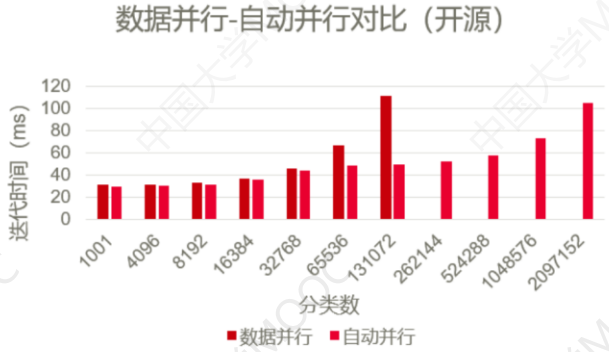
- 混合并行复杂度非常高，传统API难以编写混合并行代码，算法逻辑与并行逻辑耦合，修改并行策略，就要重新修改编码。
- 算法科学家需要关注系统（集群拓扑、网络带宽等）和并行的实现细节，才能写出高性能算法。

方案



当前版本能力：

- CV分类网络：ResNet系列模型扩大15倍，性能提升x倍；
- 在人脸ReID场景下，数据并行切换到混合并行，手工代码143行 vs 自动并行1行
- 推荐：Wide&Deep单卡串行代码Host-Device混合并行，支持十亿特征数百GB模型推荐网络



MindSpore的优势

使用MindSpore的优势

简单的开发体验

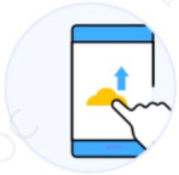
帮助开发者实现网络自动切分，只需串行表达就能实现并行训练，降低门槛，简化开发流程。

灵活的调试模式

具备训练过程静态执行和动态调试能力，开发者通过变更一行代码即可切换模式，快速在线定位问题。

充分发挥硬件潜能

最佳匹配昇腾处理器，最大程度地发挥硬件能力，帮助开发者缩短训练时间，提升推理性能。



全场景快速部署

支持云、边缘和手机上的快速部署，实现更好的资源利用和隐私保护，让开发者专注于AI应用的创造。

Thanks