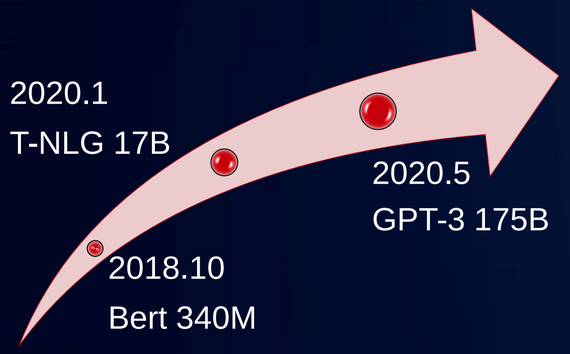
MindSpore

MindSpore Lite：极速、极智、极简，助力开发全场景智能应用

On-Device AI趋势与挑战

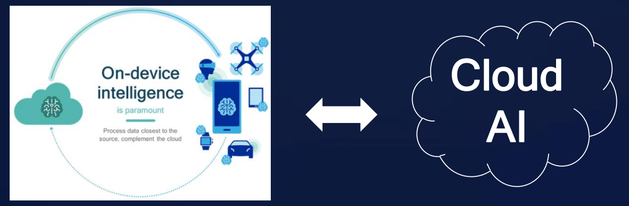
MindSpore

1：模型越来越大



- 新型算法、网络层出不穷，从Bert到GPT-3，19个月时间参数增加500倍，而手机的内存大小变化不大，端上部署挑战很大；

2：端云协同



隐私安全	大算力
低延时	大模型
高可靠	大数据
底带宽占用	高网络带宽

- On-Device AI和云端AI服务协同，更好的兼顾安全隐私和大模型更好的智能；
- 多智能设备端云协同，实现实时感知、和端云高效决策；

3： AI无处不在、模型任意部署

- 在IoT等物联网、智慧设备上极端资源受限下AI部署；

超轻量级的端侧AI引擎

➢**根据预测，到2022年80%以上的端侧设备会具备端侧AI的能力，AI将无处不在。**

On-Device AI趋势与挑战 -- 轻量化

MindSpore

- 内存限制
- 功耗限制
- 浮点/定点计算能力限制
- 安装包大小限制

requirements

- 内存复用
- 模型压缩
- 混合精度计算
- 框架层轻量化

CPU: MCU
Memory (SRAM): ≈500KB
Storage (FLASH): ≈2MB
OS: Real Time, Embedded
Virtualization: No

CPU: x86/ARM64
Memory (RAM): ≈GB
Storage : ≈TB
OS: Multi-User, Multi-Tasking, Distributed
Virtualization: Yes

On-Device AI趋势与挑战 -- 协同化

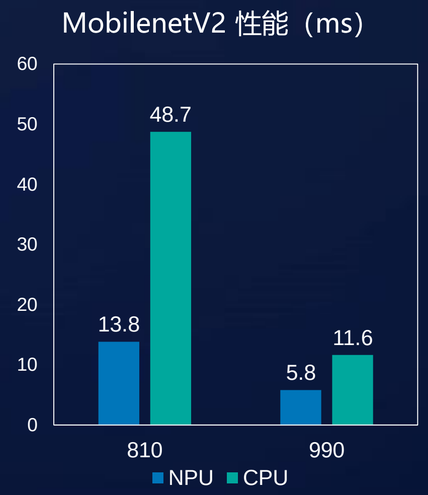
MindSpore

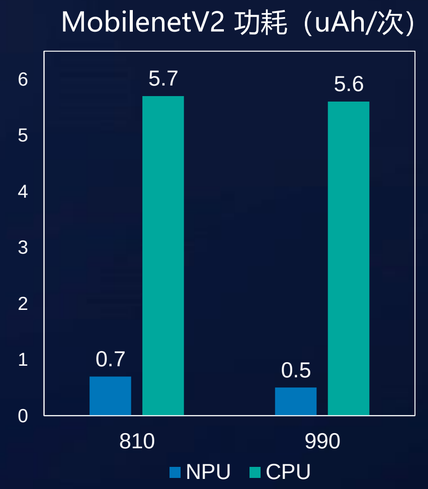
端云协同：一体化趋势明显

隐私安全
低延时
不依赖网络

大算力
大模型
依赖网络

软硬协同：AI芯片与AI软件加速库协同





- Mobile AI=On-Device AI+智慧服务，更好的兼顾个性化、安全隐私；
- 单智能体→多智能体协同，实现实时感知、决策；

MindSpore + NPU

On-Device AI趋势与挑战 -- 硬件多样性

- 随着智能家庭、自动驾驶、AR/VR等领域的兴起，智能化硬件设备种类繁多；
- 各设备厂商采用多样化的硬件提升自身竞争力；
- 各硬件厂商提供的指令集、加速计算库版本更新快，支持多版本挑战很大；

全场景AI计算框架MindSpore

MindSpore

正式发布

2019年8月
正式发布MindSpore
全场景AI计算框架

ML上线

2019年底，HMS Core ML Kit服务正式上线发布，MindSpore Inside

开源

2020年3月，
MindSpore全场景AI计算框架对外开源

端侧方案

2020年9月正式发布
MindSpore Lite端侧解决方案并对外开源；

MindSpore端云协同的全场景AI架构

MindSpore

MindSpore助力端侧AI开发者构建伟大的应用

MindSpore

极致性能

高效的内核算法和汇编级优化，支持CPU/GPU/NPU

异构调度，最大化发挥硬件算力，最小化推理时延和功耗。

轻量化

提供超轻量的解决方案，支持模型量化压缩，

模型更小跑得更快，使能AI模型极限环境下的部署执行。

全场景支持

支持iOS、Android等手机操作系统以及LiteOS

嵌入式操作系统，支持手机、大屏、平板、IoT

等各种智能设备上的AI应用。

高效部署

支持MindSpore/TensorFlow Lite/Caffe/Onnx模型，

提供模型压缩、数据处理等能力，统一训练和推理IR，

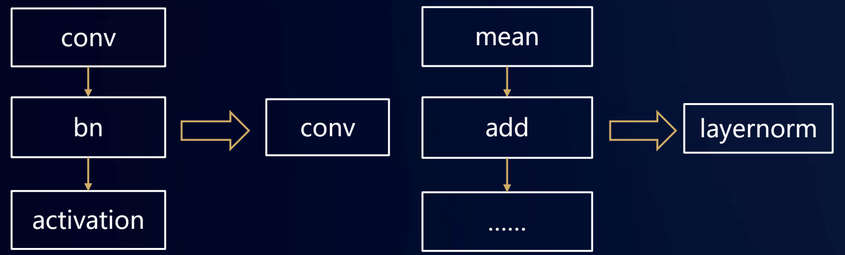
方便用户快速部署。

图优化，大幅缩减冗余计算

MindSpore

算子融合

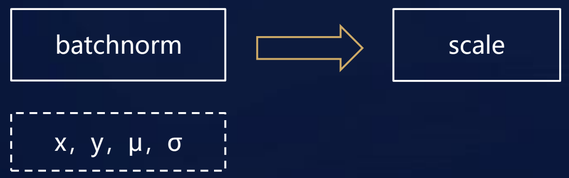
支持多达20+常见的融合，减少内存读写和计算量



$$w' = \frac{w}{\sqrt{var}} * \beta \qquad b' = \frac{b-mean}{\sqrt{var}} * \beta + \gamma$$

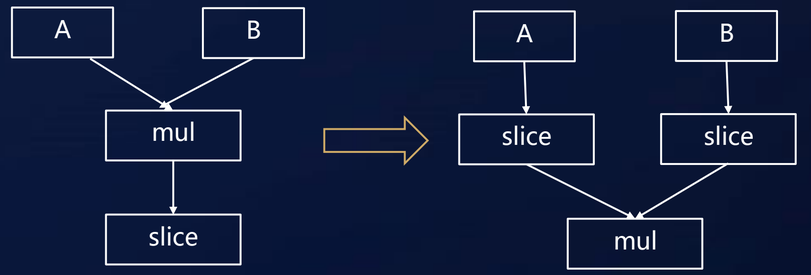
算子替换

支持常见的算子替换，通过参数值替换减少计算量



算子前移

移动slice相关算动到计算图前，减少冗余计算

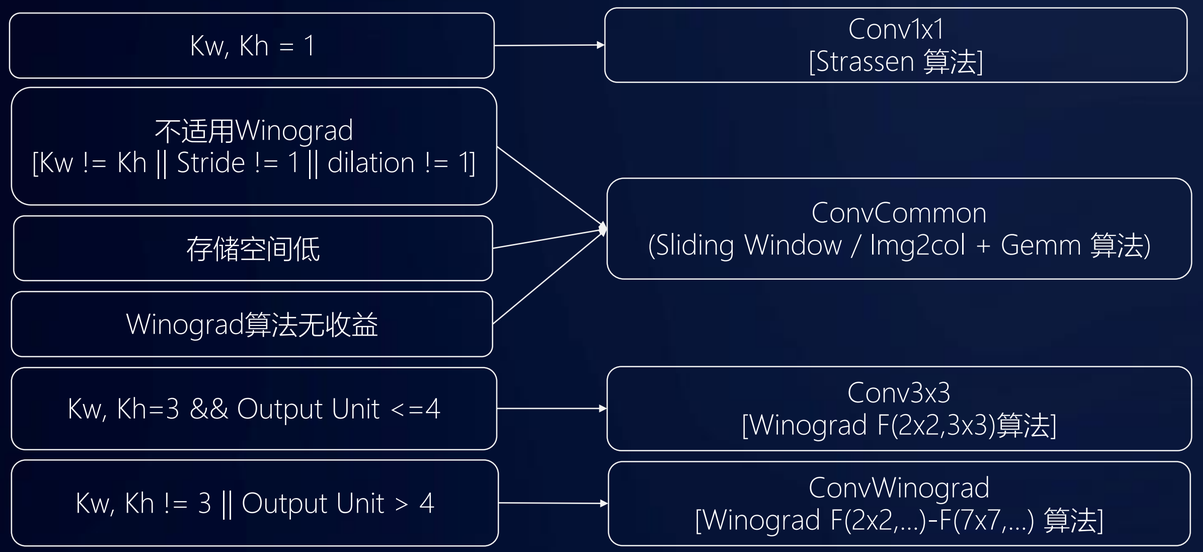


Transformer模型实验结果

transformer模型	线程数	TF Lite	MS Lite融合后	提升百分比%
encoder	1	22.181	21.513	**3.0%**
	2	15.6581	13.787	**11.9%**
	4	14.1197	11.285	**20%**

算子、指令级多级深度调优，同样终端，不同性能

高效的卷积算法



运行时指令级优化

针对硬件指令的优化手段：

Tiling：寄存器分块

指令流水：读取、分发、重排

SIMD：NEON、SSE、AVX

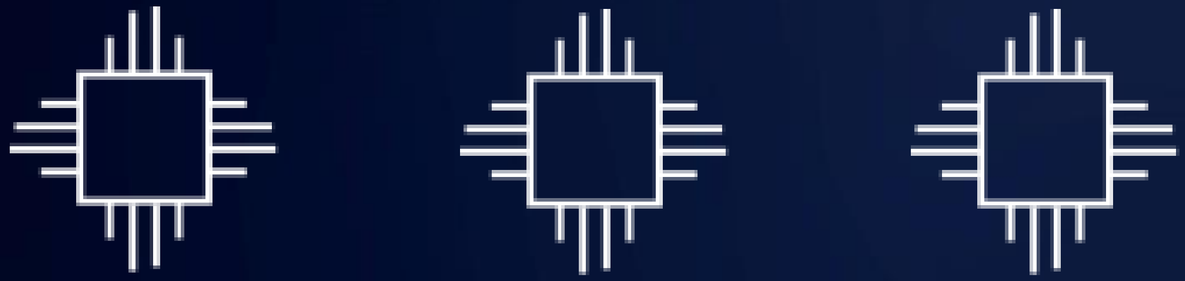
Cache优化：数据预取

- Strassen, Winograd算法实现用加法替代乘法，减少乘法次数，与访存开销进行平衡，不同的算法性能差别在1倍以上
- 部分硬件指令的使用可以提升性能达到3倍以上

软硬件异构加速，运行速度更快

MindSpore

芯片异构加速



- 在支持NPU加速的机型上，自动使能NPU加速能力，运行速度提升X倍；
- 框架支持多种SoC异构加速，当前支持Kirin、MTK的异构加速；

软件异构加速

算子级别加速

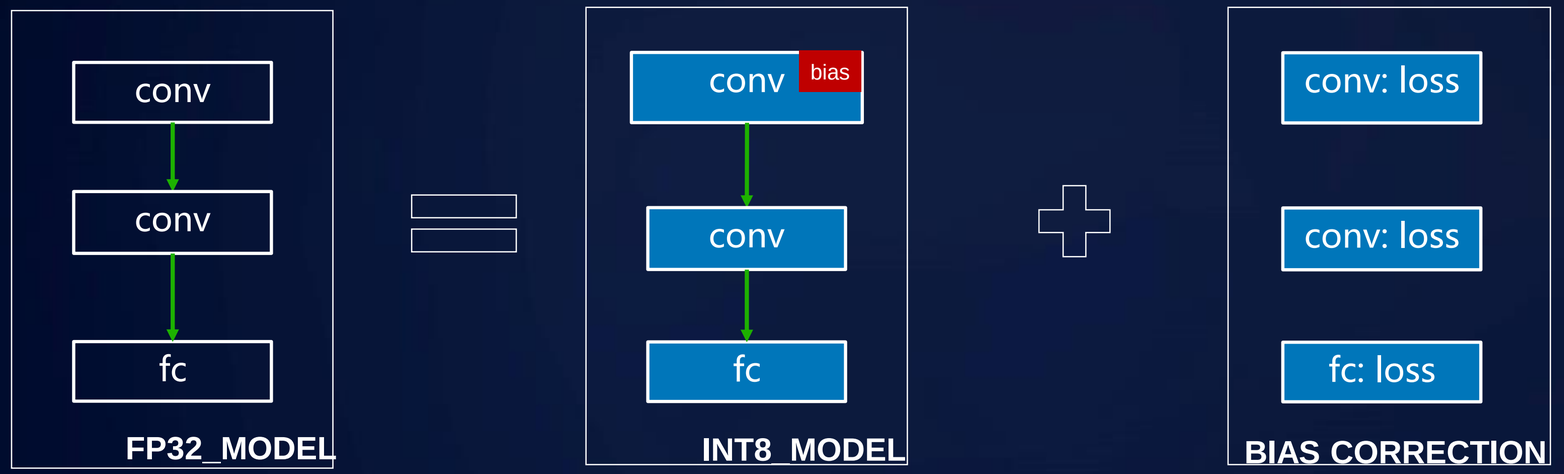
将算子或者算子的一部分计算进行拆分到不同硬件/硬件单元上并行执行

子图级别加速

拆分不同的子图到不同的硬件上并行执行

- 在拍照、修图等超大AI算力场景中，需要软件上支持大算子计算的切分，以便于更多硬件同时参与异构并行；

训练后量化，实现模型更小、推理更快



基于统计学特性的CORRECTION

权重数据的统计学特性（均值、方差）是模型固有特性，量化时获取数据的均值、方差信息，能够提高模型精度

基于算子Bias输入的CORRECTION

MindSpore Lite通过校准数据集，对比FP32模型与量化后模型的精度损失，巧妙地将误差补偿到BIAS上去，极大提高了模型的精度

商用产品实测结果

Accurate:

Mobilenet_v2	Acc
FP32	71.56%
A8W8	71.16%
A8W7	71.06%
A7W7	70.78%

Fast:

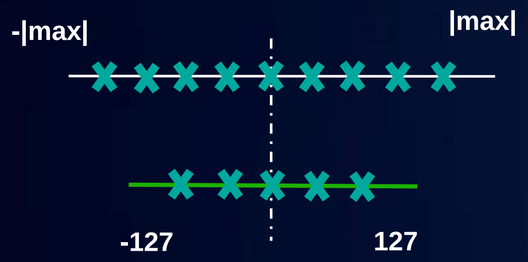
Model	Inference Time
Retinaface FP32	81.552ms
Retinaface INT8 W8	62.430ms
Retinaface INT8 W7	62.403ms

后量化精度损失小，1%以内；性能提升明显

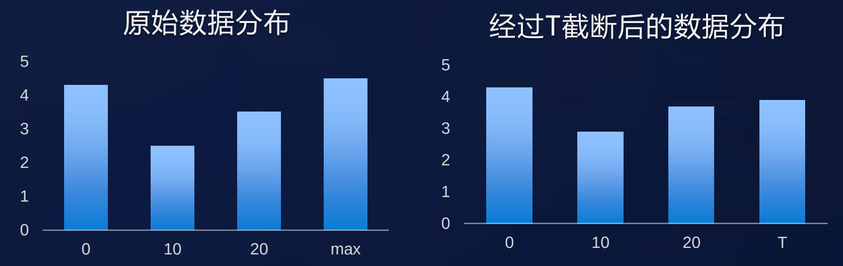
训练后量化，提供丰富的量化选择

MindSpore

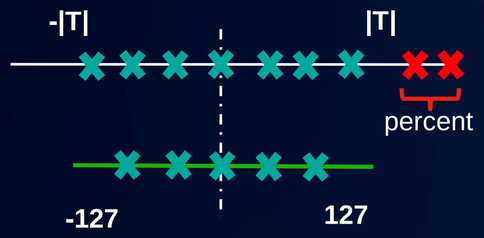
1. MAX_MIN：针对分布均匀的数据集



2. KL：通过KL散度评估数据分布的相似性



3. RemoveOutlier：按照百分比，移除最大最小值，消除离群点



4. KMeans：基于聚类的方法，将模型量化到K bit

- 权重量化：支持1到16 BIT的量化，满足用户更严苛的模型大小要求
- 量化重训：云测量化模型，一键转换

Model	Model Size(fp32)	Model Size(int8)	Acc Loss
face_tracking	420K	132K	0.28%
face_contour	2.9M	1.1M	0.07%
face_hat	1.3M	360K	0.72%
face_landmark	804K	548K	0.39%
face_openclose	364K	152K	0.16%
face_pose	892K	396K	0.23%
face_sex	964K	320K	0.16%

商用效果：应用于HMS 人脸检测识别等系列业务，精度基本无损

案例：后量化技术优化端侧设备滤镜

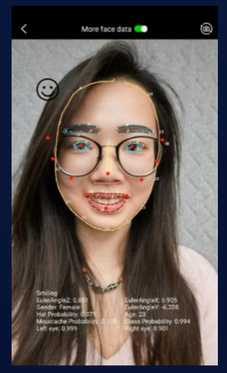
商业痛点：向端侧设备中添加滤镜，涉及人脸检测定位，5个关键点检测，轮廓精准定位，属性识别（Age, Hat, Glass, Sex, Smile，Beard）等一系列模型，当前采用全精度进行推理，功耗高（800mA），时延 xxx ms左右。

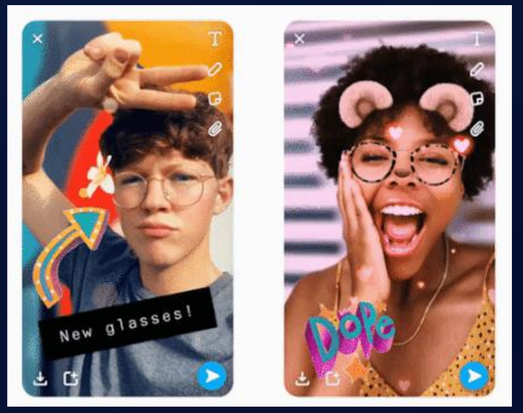
关键技术：

- 自动移除离群点识别有效量化范围
- 通过损失补偿（biasCorrection），确保量化模型精度损失小

应用效果：

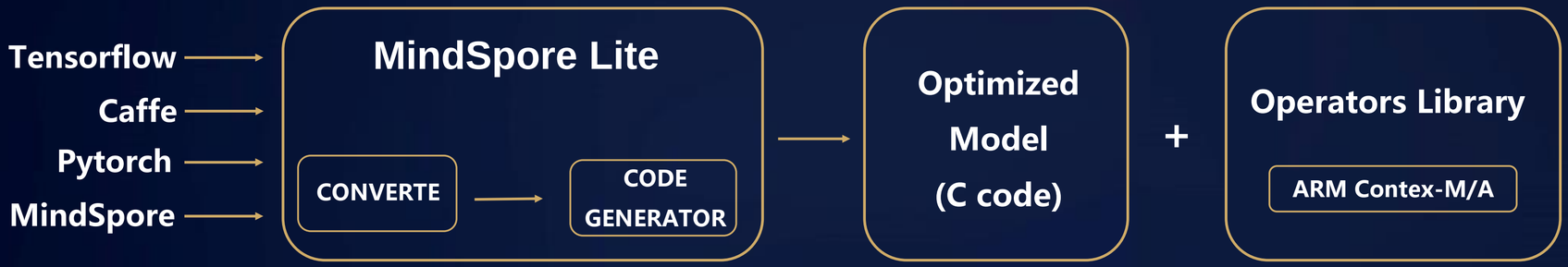
- 功耗降低到420mA，降幅达到50%
- 时延降低到全精度版本的1/3，推理精度无损





“模型即代码”，超轻量运行时，AI从“1”到“8+N”

MindSpore



技术竞争力

- 基于内存、功耗、目标环境和数据集最优化生成模型代码，解除一切第三方依赖
- 一个模型对应一个.c，运行阶段省略模型结构，退化为简单的函数调用关系

Person Detection模型实验结果

STM主频200MHz	txt(byte)	data(byte)	bss(byte)	dec(byte)
TensorFlow Lite	430916	1432	139740	572088
MindSpore Lite	244448	21912	9708	340304

典型案例

手表手势识别控制大屏：解决ROM过大，性能慢问题。

华为智慧屏
HUAWEI Vision

HUAWEI WATCH

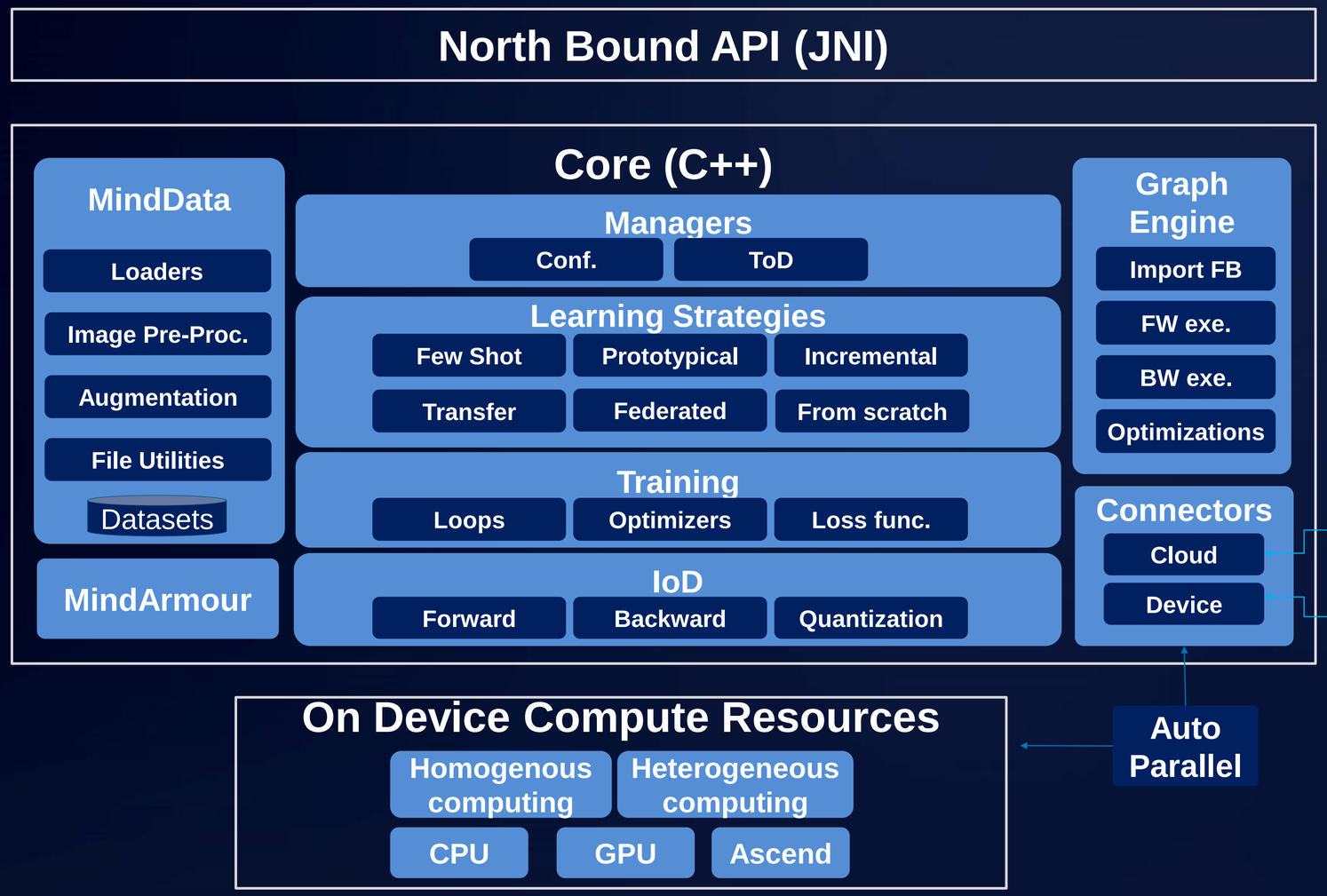
手表抬腕亮屏：解决性能时延长，亮屏精度不够问题

HUAWEI WATCH

HUAWEI WATCH

保护用户数据隐私，端侧在线训练个性化模型

MindSpore



丰富的端侧学习策略

- 增量学习
- 迁移学习
- 小样本学习

端侧数据增强

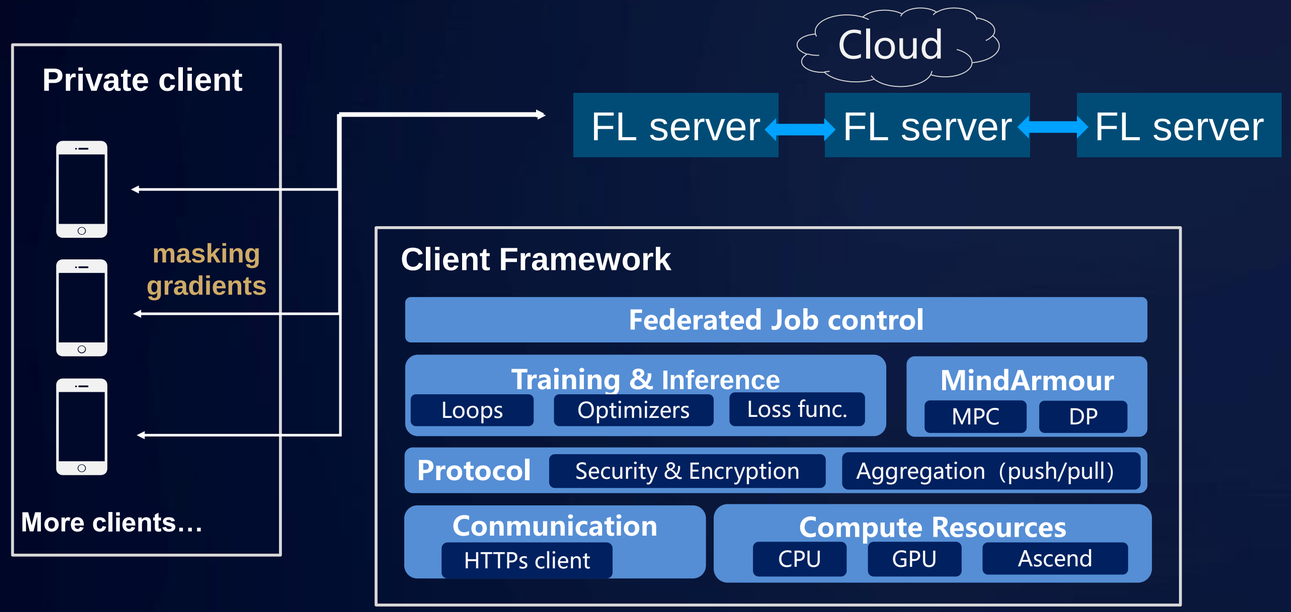
- Shuffle
- Batch
- Rotation

轻量化训练

- 自适应混合bit训练
- 多种内存复用策略
- 高效模型保存及恢复重训

联邦学习，数据不出域即可实现AI联合建模训练

MindSpore



FL server Framework

Federated Operations

1. Fed Round & client Management
2. Fed Model Aggregation
3. Performance Evaluation
4. Security & Decryption Protocol
5. All Reduce Aggregation

Cross-Device联邦学习架构

MindSpore技术优势点

在保证数据的隐私性、所有权和本地性的条件下，打破数据壁垒，数据不出域亦可实现联合建模。

- 支持多种算力不同的硬件后端部署，框架上保持端云架构一体化。
- 多种梯度压缩和Federated Aggregation算法，提高计算通信比，节省带宽资源。
- 支持多种安全聚合方案，多重隐私保护机制，在精度无损的情况下，实现隐私保护。
- 云侧集群化部署方式，动态扩缩容，应对网络不稳定，负载突变，恶意终端攻击等。
- 面向算法开发者简单易用，联邦聚合过程类似算子的方式组合

案例：端云协同下的隐私保护与智慧AI应用

MindSpore

个性且私密的广告精准推荐

痛点：

HUAWEI 广告业务涵盖广，触达x亿智能终端用户。但受困于云侧推荐模型用户画像少，且端侧数据无法上传至中央服务器，导致在用户侧特性展现有限。

方案：

- **端侧训练**--充分利用端上数据及资源完成内容分析及个性化推荐；
- **联邦学习**--打破用户与广告平台的数据壁垒，数据不上云亦可实现联合建模；
- **跨平台支持、算法优化**--支持多种硬件后端部署，支持多种梯度压缩和安全聚合算法；
- **高容错**--可应对网络不稳定，负载突变，恶意终端攻击等。

案例：端云协同下的隐私保护与智慧AI应用

MindSpore

越用越精准的家庭AI智慧屏

痛点：

在家庭中，儿童成长过程外貌特征变化大，易导致错误分类；算法不能自动感知用户纠错行为；合规隐私使用户画像无法用于模型重训。

方案：

- **端侧增量学习**--感知数据变化，在线训练新模型，实现AI业务越用越准；
- **端云联合学习**--保持用户隐私性数据、提升AI业务的个性化与精准度；
- **混合低比特量化、稀疏计算、 CPU/GPU 混合异构并行**--减少计算量和内存开销，提升AI业务用户体验实时性。

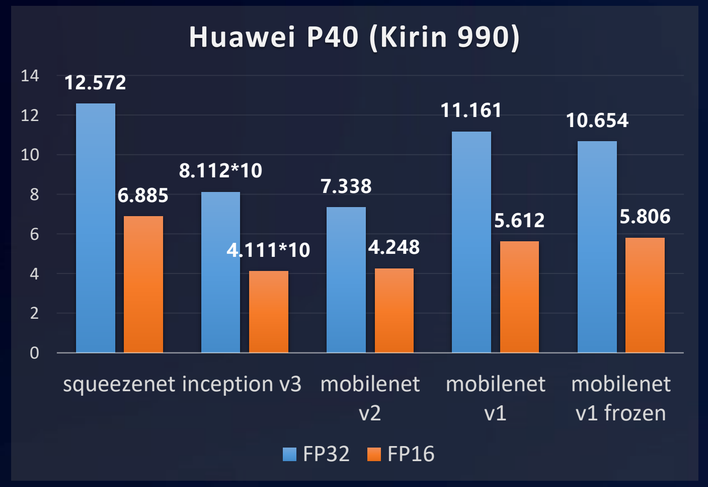
极致性能

MindSpore

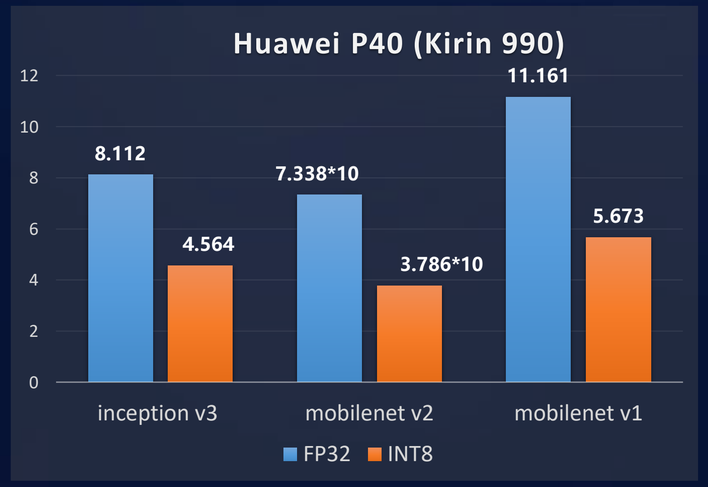
MindSpore Lite致力于探索极限性能提升

以TensorFlow官网100+预置模型测试结果为例，MindSpore Lite可实现98%的网络推理性能超越TF Lite，其中性能超过30%的网络占到总数的70%。

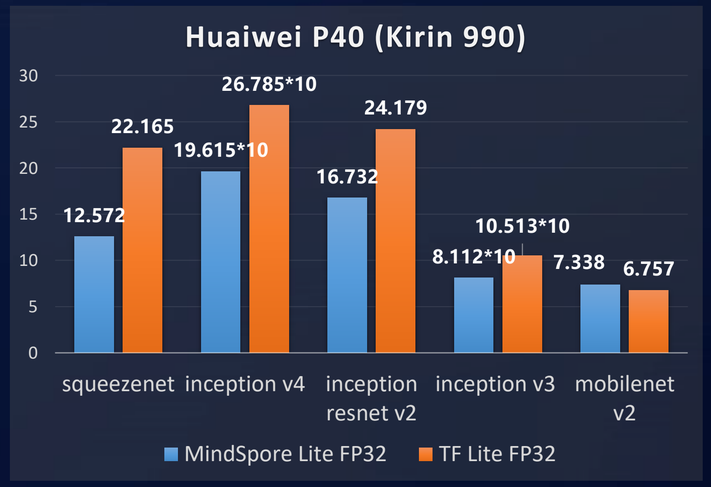
□ 半精度优化后的推理时长：



□ FP32 INT8优化后的推理时长：



□ FP32下MS Lite与TF Lite的对比：



兼容所有主流的AI框架

MindSpore Lite转换工具支持主流AI框架模型的转换和优化

无缝支持MindSpore训练的模型进行端侧学习和高性能推理

MindSpore Lite支撑HMS的服务

MindSpore

使用ML KIT 轻松构建您的AI应用

打造视觉及语言AI全新体验
ML Kit提供丰富的视觉及语言类机器学习服务APIs

文本类

- 文本识别
- 文档识别
- 身份证/银行卡/通用卡证识别

图像类

- 图片分类
- 对象检测和跟踪
- 地标识别
- 文档校正
- 拍照购物
- 图片分割
- 图像超分辨率

语音语言类

- 文本翻译
- 语种检测
- 个性化讲解视频生成
- 实时语音识别
- 语音合成

人脸人体类

- 人脸检测
- 活体检测
- 人体骨骼检测
- 数字人
- 手部关键点识别

自定义模型

- 自定义图片分类
- 自定义对象检测
- 自定义文本分类

MindSpore

手机上的MindSpore Lite

大赛安排

大赛活动主题：三选一

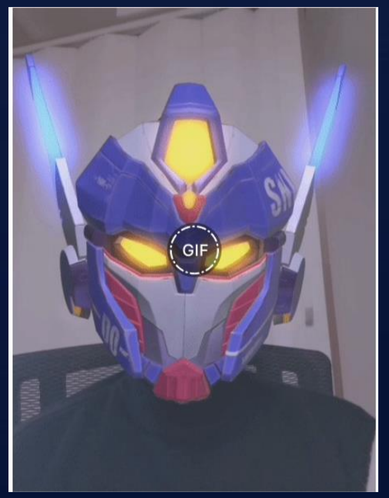
MindSpore

1. 风格迁移

把实物转换成不同风格的图片，如写实风、油画风、漫画风等。

2. 人脸检测

识别视频中的人脸，根据人脸位置，生成兔子耳朵、变形金刚贴图等。



3. 骨骼检测

基于提供的基础模型，制作运动姿势偏差检查等运动类APP应用。

赛程安排

MindSpore

- **赛程安排，周期9个月（2020年12月~2021年9月）：**

12月~1月

第①阶段
大赛启动

- 发布活动细则

1月~8月

第②阶段
线上辅导+激励

- 每月抽取三等奖
- 针对优秀开发者给与辅导

8~9月

第③阶段
线上辅导+评选

- 逐步公布奖项结果
- 针对优秀开发者给与辅导

总结

第④阶段
结果发布

- 结果发布并进行颁奖

- **奖项设置：**

奖品及人数以活动细则发布版本为准；

获奖者可赢得奖品额度下的等额奖品；

序号	奖项	难度及要求	奖品额度（元）	人数
1	一等奖	二等奖里每个主题选出前三名，投票给出第一； 第一名10000元等额奖品，成为预置模型，第二名第三名各5000元等额奖品	10000/5000	共9人
2	二等奖	根据提供的模型进行一定程度的修改，做出不一样的APP，三个主题任选其一即可	1000	500
3	三等奖	根据提供的教程和模型，完成同样的APP，三个主题任选其一即可	50	20000
4	邀请激励奖	邀请用户参与活动，根据邀请的数目可赢得不同奖品	0~5000	约900人
5	数据集贡献奖	数据集众筹，上传训练集和验证集，验证集精度达标且可以自动化验证	1500	200

MindSpore

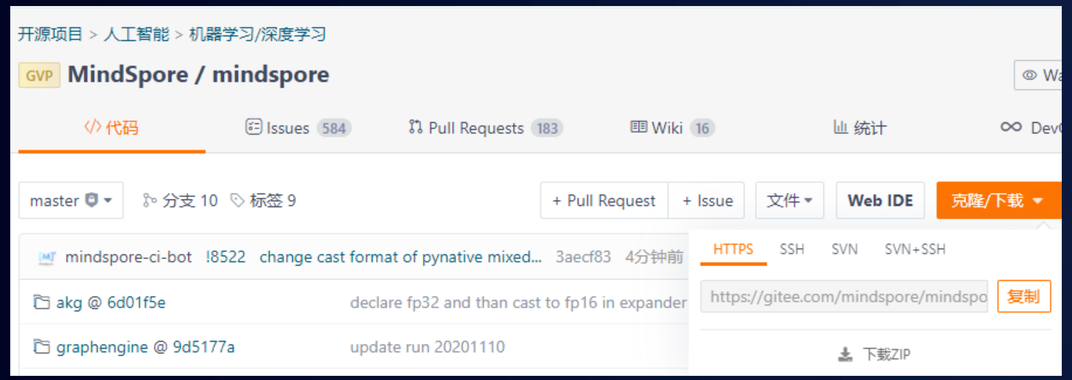
现场体验

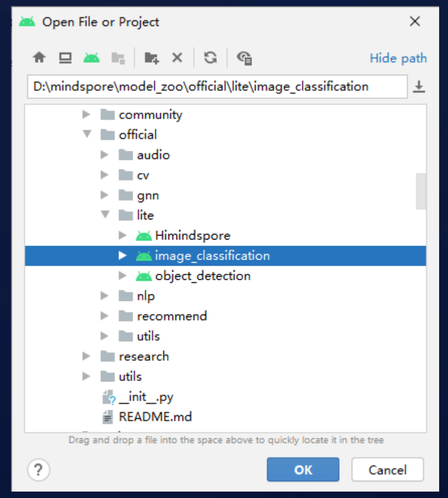
MindSpore Lite初体验 — Image Classification

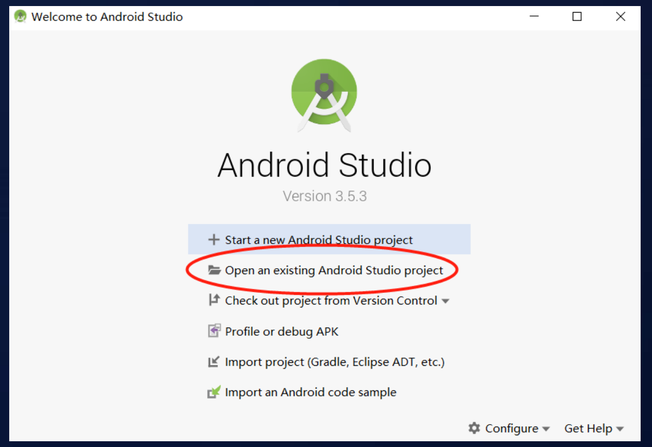
MindSpore

1. 下载MindSpore代码仓

① 登陆 https://gitee.com/mindspore/mindspore，使用git clone / git pull下载代码仓







2. 在Android Studio中打开Image Classification Demo

① 打开Android Studio，根据文件目录 mindspore / model_zoo / official / lite / image_classification 打开目标demo。

② 可以同步参考 image_classification / README.md 完成本次demo部署，或解决SDK、NDK相关问题。

MindSpore Lite初体验 — Image Classification

MindSpore

3. 连接Android手机

① 使用数据线在电脑中连接Android，并开启USB调试模式。

② 可在手机“设置”中搜索“USB调试”，并开启该选项。

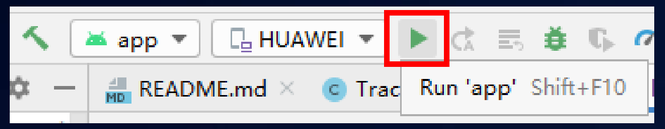
③ 若未能搜到，可搜索“版本号”，进入后连续点击7次“版本号”栏，开启开发者模式，重复上一步骤。

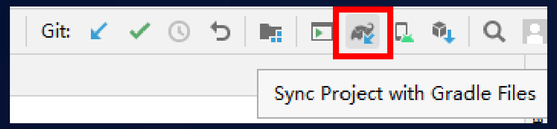
4. 运行Image Classification

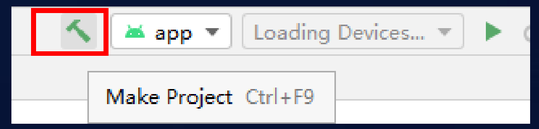
① 成功连接手机之后，Android Studio自动识别设备，点击“Run ‘app’ ”即可在设备上运行本项目。

② 编译过程中会下载部分文件，请耐心等待。

③ 若Android Studio未能自动识别设备，可通过点后点击“Sync Project”、“Make Project”选项完成编译，此时可生成apk文件，目录为：app / build / outputs / apk / debug / app-debug.apk。通过文件传输放入手机，并点击打开。







MindSpore Lite初体验 — Image Classification

MindSpore

5. 安装Image Classification应用

① 完成上一步后，手机自动安装Demo，开启许可即可完成安装，体验MindSpore Lite图像分类功能。

② 终极解决方案：若未能完成上一步，也可通过扫描下方二维码下载，安装HiMindSpore，从中选择Image Classification。

MindSpore

MindSpore

官网：http://www.mindspore.cn/

代码仓：https://gitee.com/mindspore/mindspore.git

镜像： https://github.com/mindspore-ai/mindspore.git

MindSpore Lite Code

MindSpore Lite WebSite

MindSpore Lite Demo

MindSpore

Thank You!