北京
2019

遨游“视”界 做你所想
Explore World, Do What You Want

LiveVideoStackCon 2019 北京

2019.8.23-24

出品：LiveVideoStack 音视频技术社区 CSDN

xNN: 支付宝App中的实时AI引擎

蚂蚁金服 - 多媒体产品技术部 - 机器视觉
周大江（弘川）

蚂蚁金服
ANT FINANCIAL

深圳 2019

遨游“视”界 做你所想
Explore World, Do What You Want

LiveVideoStackCon 2019 深圳

2019.12.13-14

出品：LiveVideoStack 音视频技术社区

成为讲师： speaker@livevideostack.com

成为志愿者： volunteer@livevideostack.com

赞助、商务合作： kathy@livevideostack.com

端侧机器学习

机遇

端侧固有
资源限制

支付宝App
场景难点

体验

使能实时交互应用，体验不受网络影响，部分场景端到端精度更高

成本

计算本地化，节省云侧计算/流量/存储开销

隐私

数据上传非必须，保护用户敏感数据，规避商户法律与舆论风险

挑战

算力

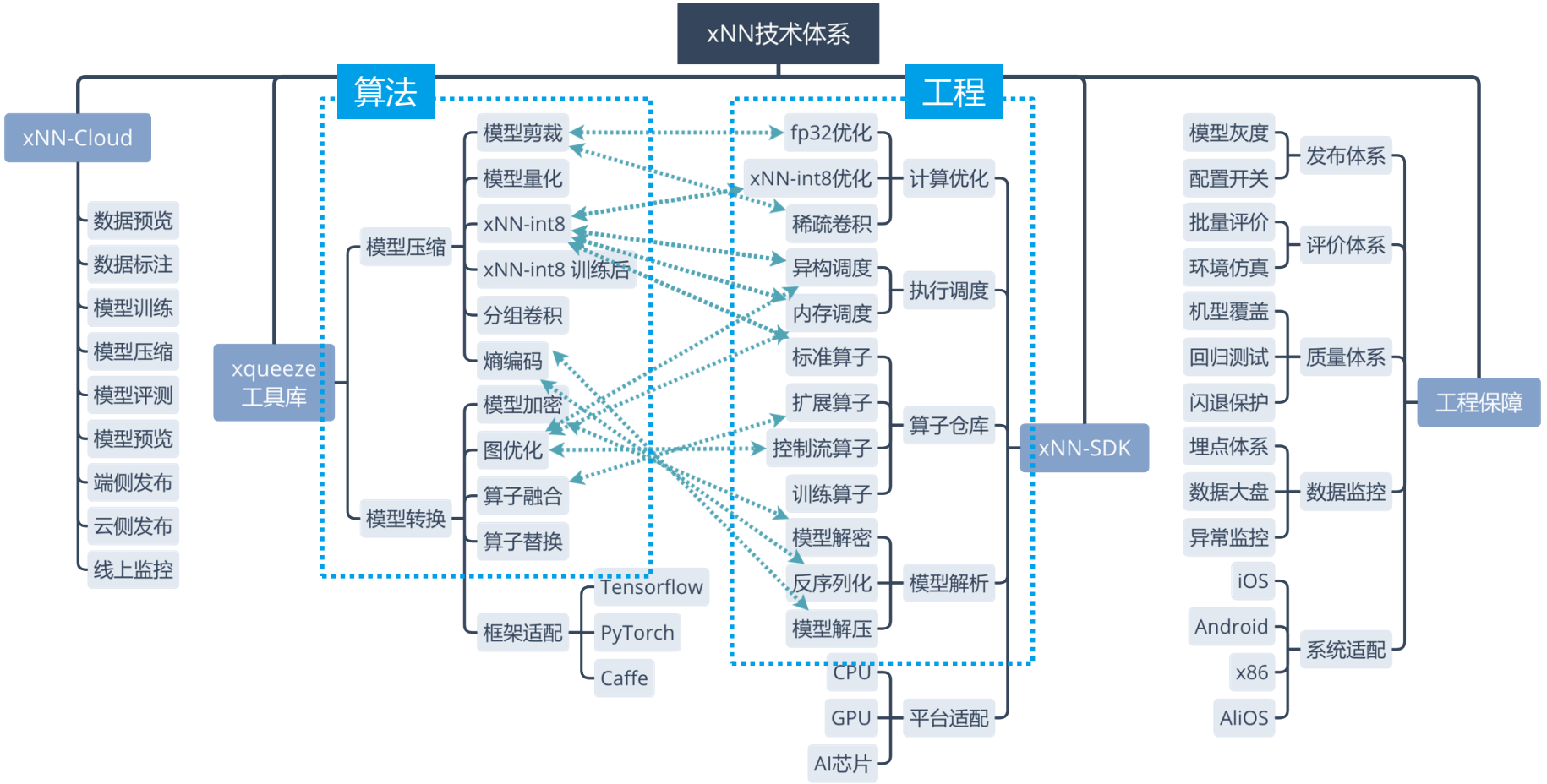
海量用户基数与“普惠”服务原则，更需要对低端老旧机型的不离不弃

空间

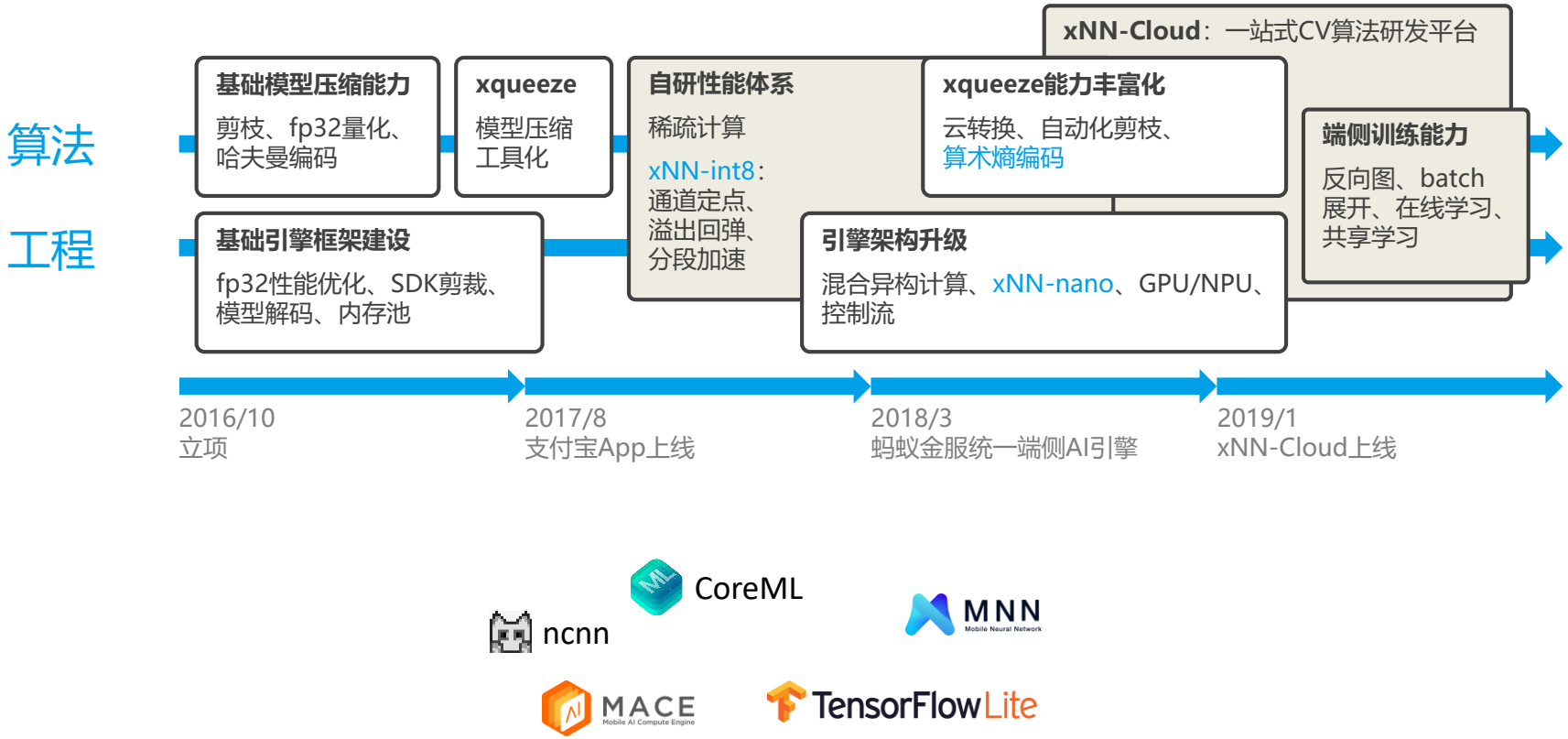
支付宝App承载众多业务线，空间资源的紧张程度远超一般应用

效率

如何以有限的人力支撑众多业务线的端侧机器学习需求


xNN技术体系
算法
工程
xNN-Cloud
数据预览
数据标注
模型训练
模型压缩
模型评测
模型预览
端侧发布
云侧发布
线上监控
xqueeze
工具库
模型压缩
模型转换
模型剪裁
模型量化
xNN-int8
xNN-int8 训练后
分组卷积
熵编码
模型加密
图优化
算子融合
算子替换
框架适配
Tensorflow
PyTorch
Caffe
fp32优化
xNN-int8优化
稀疏卷积
计算优化
异构调度
内存调度
执行调度
标准算子
扩展算子
控制流算子
训练算子
算子仓库
模型解密
反序列化
模型解压
模型解析
CPU
GPU
AI芯片
平台适配
xNN-SDK
模型灰度
配置开关
发布体系
批量评价
环境仿真
评价体系
机型覆盖
回归测试
闪退保护
质量体系
埋点体系
数据大盘
异常监控
数据监控
iOS
Android
x86
AliOS
系统适配
工程保障

xNN技术演进



CoreML

ncnn

MNN
Mobile Neural Network

MACE
Mobile AI Compute Engine

TensorFlow Lite

xNN模型压缩

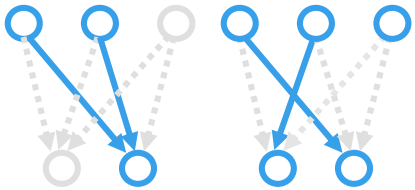
算法仓库构成

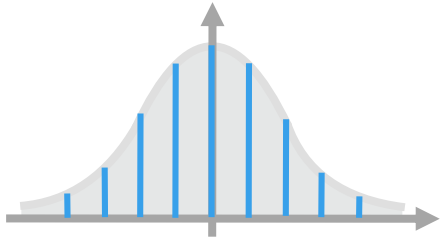
神经元/突触剪枝
模型结构精简化

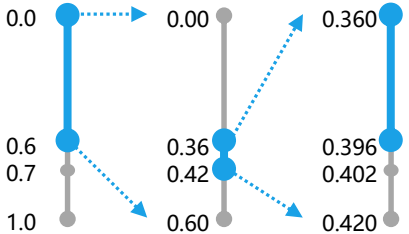
xNN-int8体系
模型参数离散化

自适应算术编码
逼近理论压缩极限

较Huffman提升15%



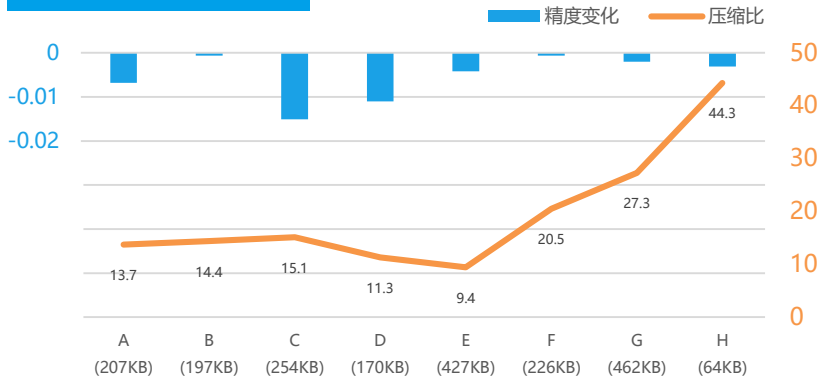




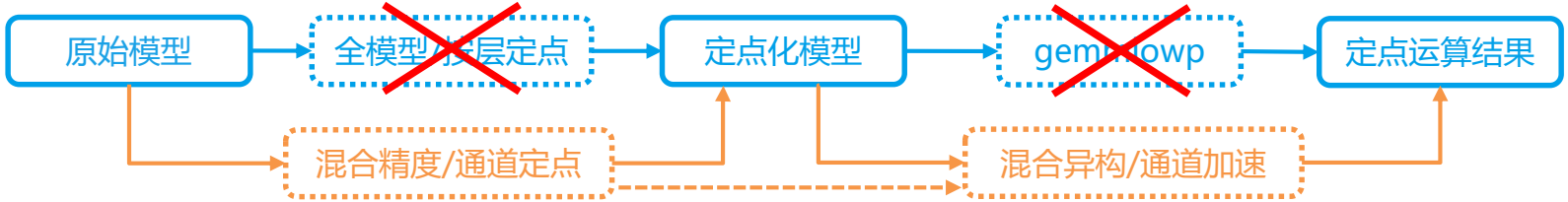
算法使用体验

```python
import tensorflow as tf

def train():
  train_op, loss = mobilenet_v1()
  with tf.Session() as sess:
    if step < 100000:
      step += 1
      _, val = sess.run([train_op, loss])


if __name__ == '__main__':
  train()
```

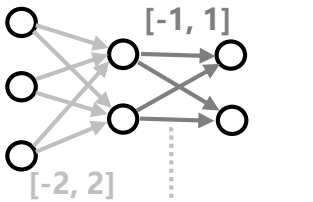
模型压缩效果



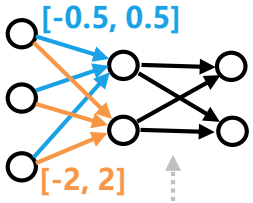
xNN-int8体系：精度提升



逐层内参数统一定点

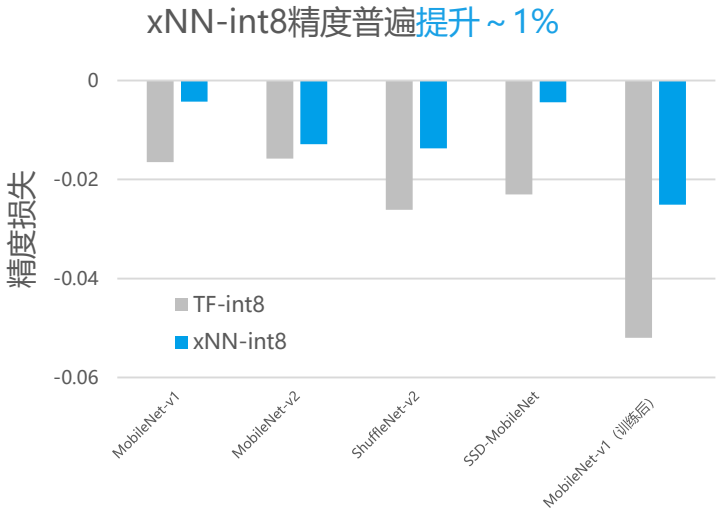


按输出通道定点

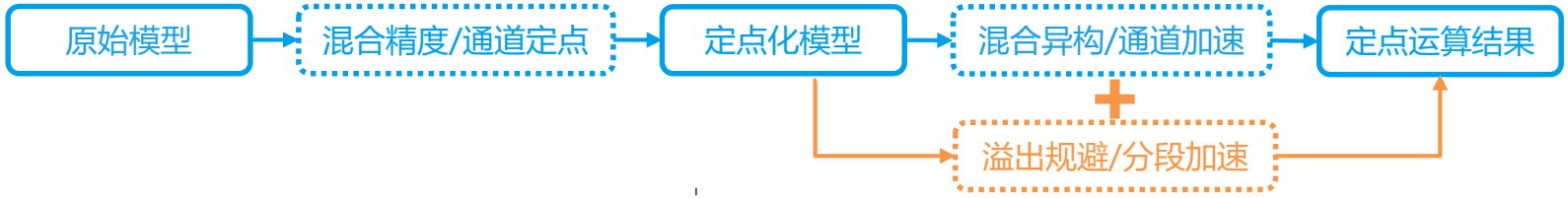


回归层精度敏感，允许
int8/float32混合精度

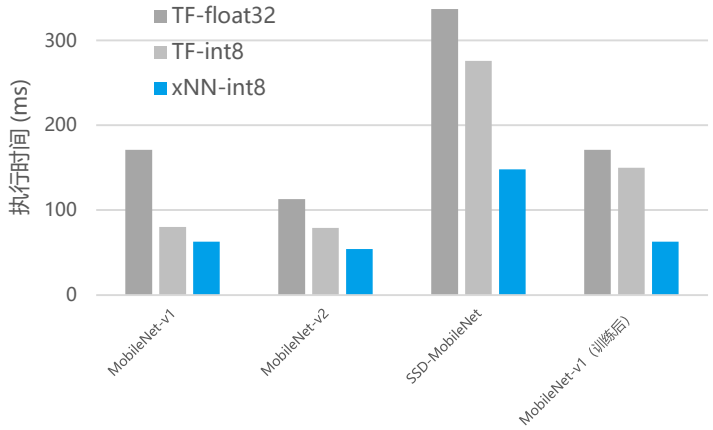
通道间分布不均
常见于移动端模型结构如：
Group Convolution
Depth-wise Convolution



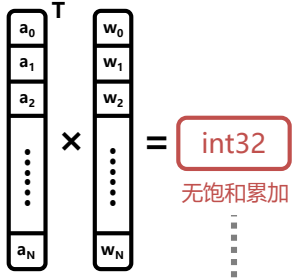
xNN-int8体系：性能提升



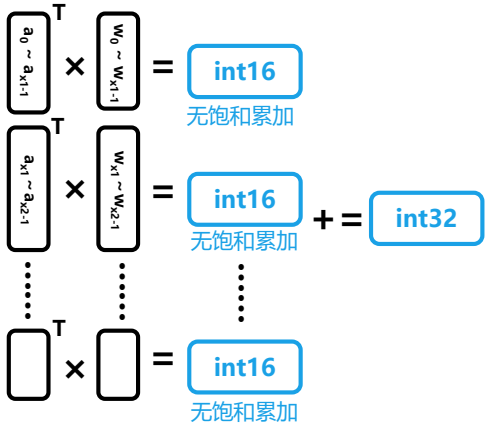
xNN-int8性能：提升～20%



传统定点实现

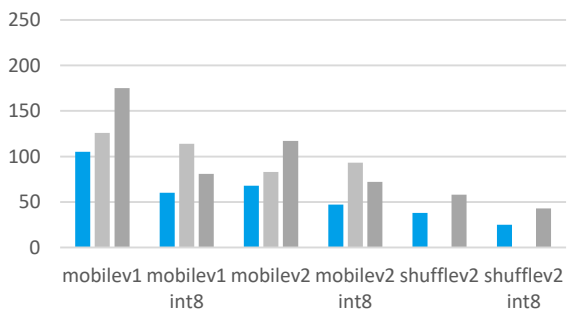


幅值分段-无饱和加速

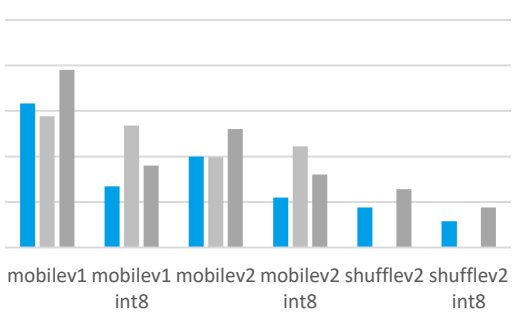


与业界方案的对比

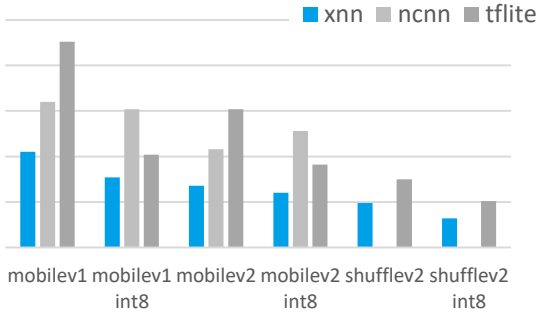
	计算性能	模型压缩	SDK尺寸	工具链
xNN方案	定制int8体系 训练中+训练后量化 精度高+速度快	剪枝+量化+算术熵编码 10-30倍压缩比	去STL依赖 最简配置不到100KB	标注-训练-压缩-转换-测试- 部署-监控 全链路可视化
业界主流	标准int8体系 训练后量化	仅量化 2-4倍压缩比	300KB+	命令行转换



cpu_time@Mi6 (ms)



cpu_time@vivo X20A (ms)



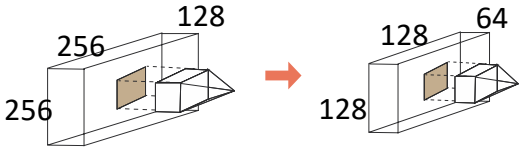
cpu_time@vivo X23 (ms)

高效/轻量级网络设计

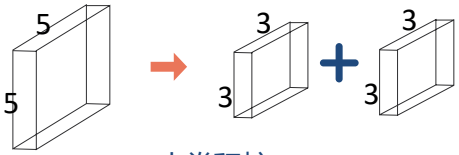
- 高效网络设计思想

MobileNet　ShuffleNet

Backbone

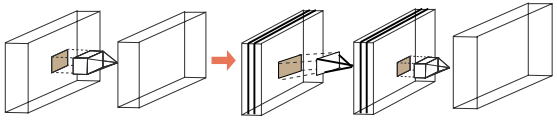


宽度、分辨率、深度

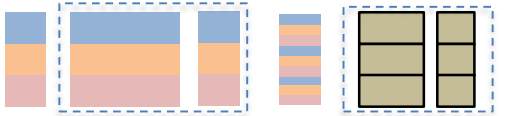


小卷积核

$M \times k \times k \times N$　$M \times k \times k + M \times N$

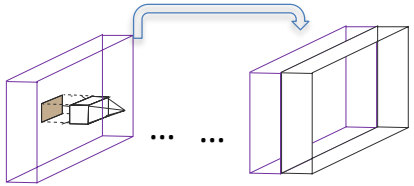


Depthwise Convolution

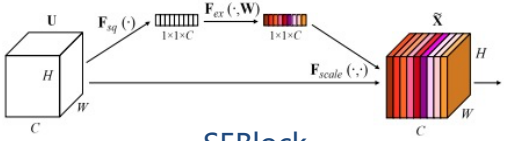


Group Convolution

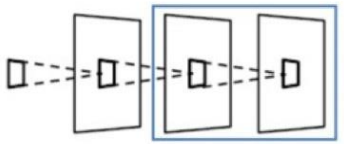
- 轻量化和精度平衡



特征复用（FPN、通道分离）



SEBlock



Layer Reuse

端侧实时AI能力：金融信息场景

银行卡OCR

识别准确率 99+%
模型尺寸 460KB

身份证OCR

6000 字中文字库
全字段准确率 84+%
模型尺寸 980KB

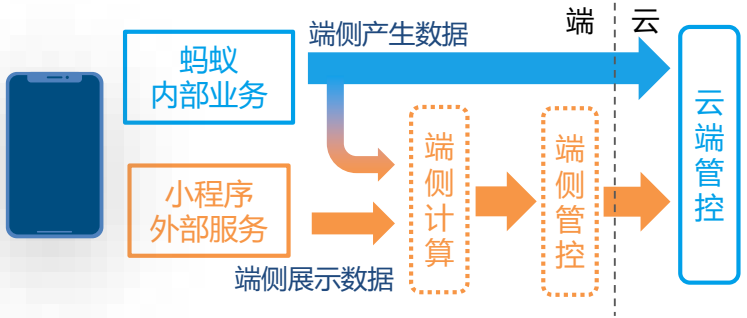
理赔宝

实时拍摄引导
图片质量
文档属性/类型

有效数据比例：
40% → 90%

定损宝

车牌/VIN码OCR
距离检测
模糊/光线检测
部件/损伤检测



端+云 内容安全

通用OCR 1.2MB
端侧分流50%计算

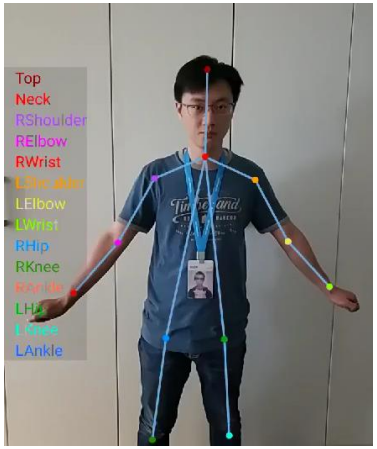
鉴黄模型 253KB
端侧减少70%数据上报

端侧实时AI能力：互动运营场景

春节扫福

AR会议室

庄园手保健操



人体姿态

手势关节点

xNN应用

40+ 业务场景、80+ 模型 线上运行

支付宝App

新春红包：2019年活动期间支持3000亿+调用，客户端覆盖率98%

扫码支付：码姿态矫正+小码放大，显著提升扫码体验

智能加载：支付宝App卡顿减少，关键业务启动速度提升

资金风控：通过手机操作预测丢失，保障用户资金安全

阿里集团App

合作银行&保司：10+ App

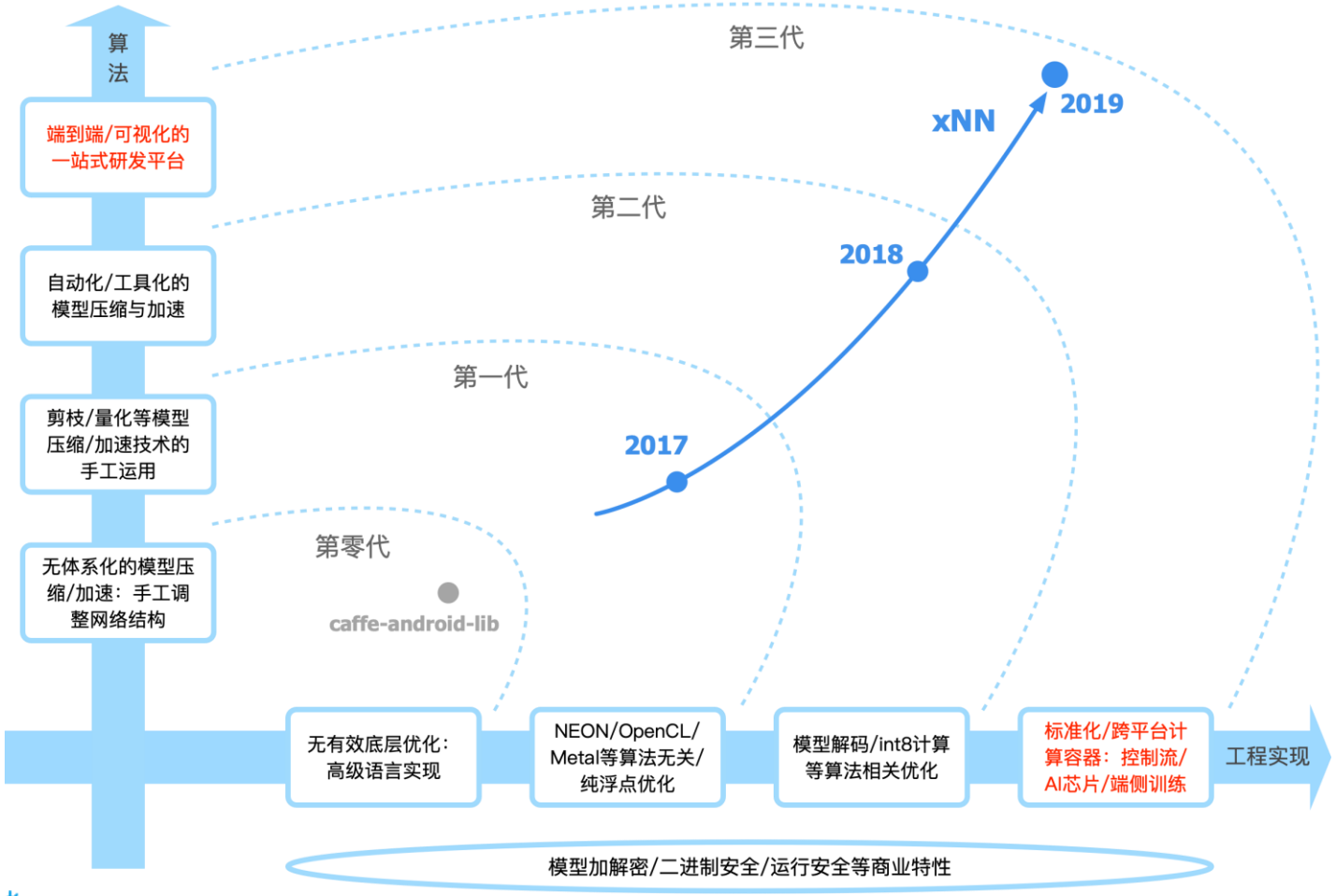
IoT设备场景

刷脸支付：使用xNN-int8后性能翻倍，显著提升刷脸体验

“蜻蜓”：刷脸支付+扫码支付，体验同步提升

高德车镜：10万台设备采用xNN实现端侧计算

端侧AI的演进


算法
端到端/可视化的一站式研发平台
自动化/工具化的模型压缩与加速
剪枝/量化等模型压缩/加速技术的手工运用
无体系化的模型压缩/加速：手工调整网络结构
第三代
第二代
第一代
第零代
2019
xNN
2018
2017
caffe-android-lib
无有效底层优化：高级语言实现
NEON/OpenCL/Metal等算法无关/纯浮点优化
模型解码/int8计算等算法相关优化
标准化/跨平台计算容器：控制流/AI芯片/端侧训练
工程实现
模型加解密/二进制安全/运行安全等商业特性

xNN-Cloud: 一站式CV算法研发平台

数据标注
人工标注
模型标注

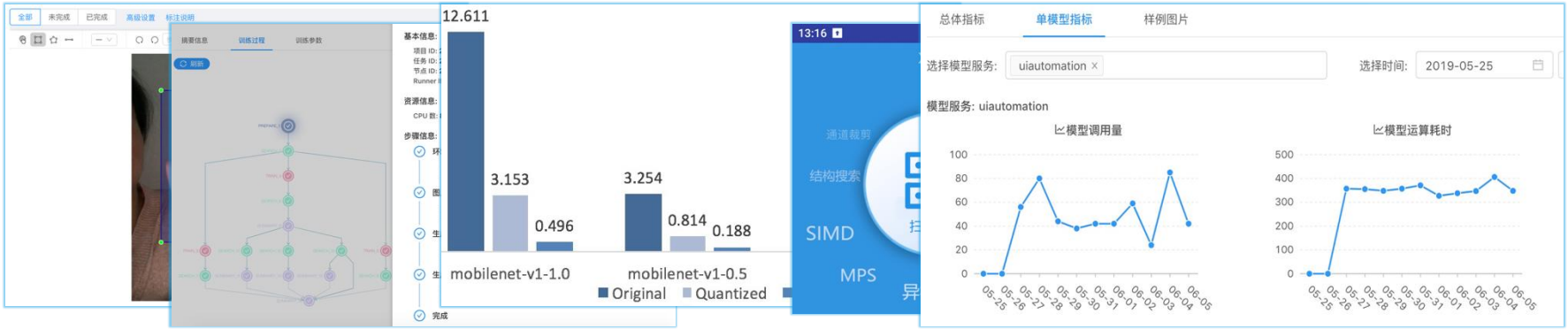
模型训练
调度控制
参数搜索

端化处理
压缩加速
模型加密

模型表现
手机预览
云侧平台

模型发布
客户端
服务端

线上监控
性能指标
数据回流



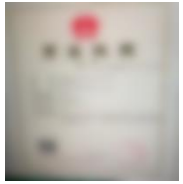
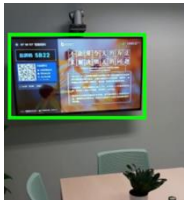
xNN-Cloud

5分钟带您了解

XNN.ALIPAY.COM

xNN-Cloud：算法能力沉淀

- 平台化沉淀算法模型，用户可自助训练业务模型

分类	 识花 大小: 253 K 精度: 0.98 耗时: 73 ms	 营业执照 大小: 231 K 精度: 0.99 耗时: 62 ms	 身份证 大小: 226 K 精度: 0.99 耗时: 57 ms	. . .
检测	 Landmark 大小: 312 K 精度: 0.95 耗时: 125 ms	 Icon 大小: 220K 精度: 0.97 耗时: 115 ms	 屏幕 大小: 230 K 精度: 0.99 耗时: 120 ms	. . .
OCR	 文档标题 大小: 628 K 精度: 0.9 耗时: 320 ms	 快递单 大小: 980 K 精度: 0.85 耗时: 600 ms	 投屏码 大小: 530 K 精度: 0.95 耗时: 240 ms	. . .

公开参考资料

这款神秘的移动端OCR引擎，如何做到“所见即所得”？

原创：亦弦 阿里技术 4月2日

阿里妹导读：随着深度学习，尤其是CNN和RNN等技术的飞速发展，文字识别技术(OCR)[illegible]与此同时，在[illegible]化智能识别凭[illegible]的隐私保护和[illegible]亲睐，越来越[illegible]完成，OCR也[illegible]算法专家亦弦[illegible]移动端OCR引擎——xNN-OCR。

不写一行代码，完成机器视觉算法的研发

原创：南烽 阿里机器智能 3月26日

小叽导读：xNN云平台面向普通的研发、质量、UE[illegible]，也面向视觉[illegible]供全自动的模[illegible]配合专业的数[illegible]、端侧发布、[illegible]程能力，使得[illegible]况下完成计算[illegible]的研发全流程。

含代码 | 支付宝如何优化移动端深度学习引擎？

原创：家大 阿里技术 2018-06-20

阿里妹导读：移动端深度学习在增强体验实时性、降低云端计算负载、保护用户隐私等方面具[illegible]、安全等领域[illegible]考虑到移动端[illegible]落地面临着性[illegible]内存使用、模[illegible]

本文介[illegible]两个方面的联合优化，应对上述挑战，最终实现技术落地，希望对大家有所启发。

揭秘支付宝中的深度学习引擎：xNN

原创：弘川 阿里技术 2017-09-28

阿里妹导读：本文介绍支付宝App中的深度学习引擎——xNN。xNN通过模型和计算框架两个方面的优化，解决了深度学习在移动端落地的一系列问题。xNN的模型压缩工具（xqueeze）在业务模型上实现了近50倍的压缩比，使得在包预算极为有限的移动App中大规模部署深度学习算法成为可能。xNN的计算性能经过算法和指令两个层面的深度优化，极大地降低了移动端DL的机型门槛。

深度学习

近来，深度学习（DL）在图像识别、语音识别、

端侧AI的团队

业务算法

运用深度学习技术解决具体业务问题的专家，专精图像分类/分割/目标检测/OCR等算法

基础算法

深入理解机器学习原理、深度学习基础网络结构、信息论，主攻自动化模型压缩与加速工具的研发

工程架构

打通端侧框架 (Android, iOS) 与后台系统 (Java)，建设覆盖算法研发全流程的高效工程平台

算法优化

计算机体系结构与算法性能领域的专家，运用NEON/OpenCL/汇编等技术提升端侧框架的效率

北京
2019

遨游“视”界 做你所想
Explore World, Do What You Want

Thank you

出品：LiveVideoStack 音视频技术社区 CSDN