Alibaba Group 阿里巴巴集团

TAO TECHNOLOGY 淘系技术部

MNN Mobile Neural Network

# MNN学院

## 端智能与MNN初探：面向未来的移动开发

讲师：舒会（花名-玄裳）

MNN项目核心负责人之一
阿里巴巴淘系技术高级技术专家

# About Me

- Amazon
- Google:
  - ✦ Federated Learning
  - ✦ ML Kit founding member
- 2019. 9. Alibaba: MNN

Alibaba Group 阿里巴巴集团 | TAO TECHNOLOGY 淘系技术部

MNN Mobile Neural Network

# Agenda

- 1. 端智能介绍：趋势、特点、挑战
- 2. 端智能在手机淘宝的应用
- 3. MNN介绍

# WHY?

- 隐私性：先看一则新闻

# WHY?

- 数字对比。
  - 2019: The world’s fastest super computer the “Summit” [1]: 143.5 PFLOPS.
  - 2018: Kirin 970 NPU 1.92 TFLOPS [2] * 17 million P20 [3] = 32640 PFLOPS = 227x “Summit”.

[1] https://en.wikipedia.org/wiki/Supercomputer
[2] https://hexus.net/tech/news/cpu/109757-huawei-kirin-970-soc-dedicated-neural-processing-unit/
[3] https://www.huaweicentral.com/huawei-sold-over-17-million-units-of-p20-series-and-7-5-million-units-of-mate-20-series-in-2018/

# 端侧推理 挑战 & 优势

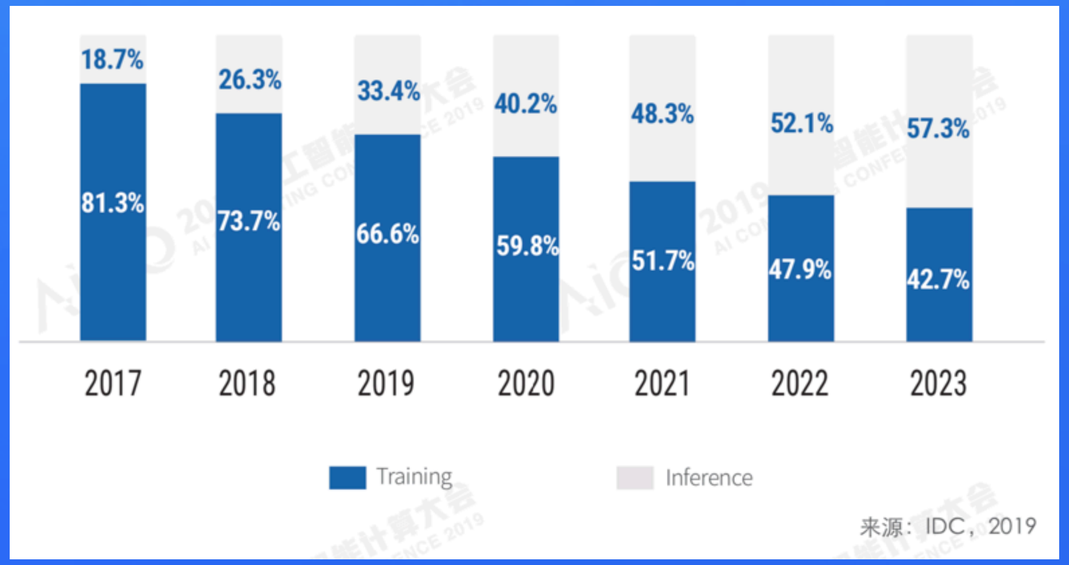
- 挑战：速度，引擎大小，模型大小， 内存
- 优势：隐私，速度，省云端资源

# 端智能

- 趋势1：端上推理的重要性高于训练，但是训练在补齐 (TFLite, Pytorch, Mindspore)
- 趋势2：后摩尔时代，XPU 百花齐放
- 趋势3：NLP逐步走向成熟
- 趋势4：手机端到AIOT端

# 趋势1： 端上推理+训练

推理的比重将上升




端上训练

TensorFlow Lite

ML

[M]S MindSpore

Alibaba Group 阿里巴巴集团

TAO TECHNOLOGY 淘系技术部

MNN Mobile Neural Network

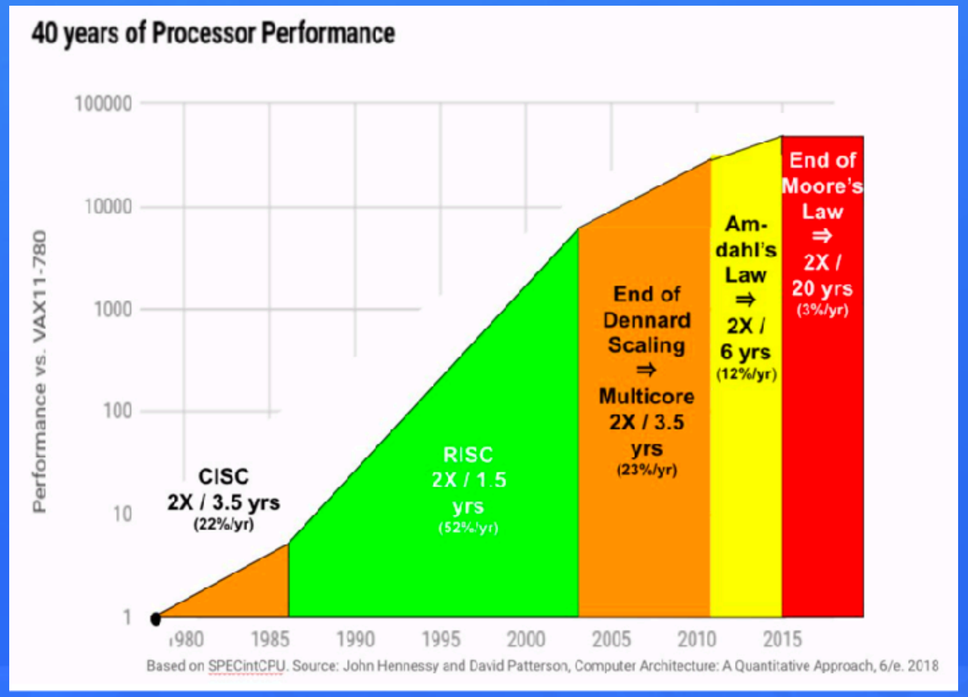
# 趋势2：后摩尔时代，XPU 百花齐放

One thing that’s been shown to be pretty effective is specialization of chips to do certain kinds of computation that you want to do that are not completely general purpose, like a general-purpose CPU

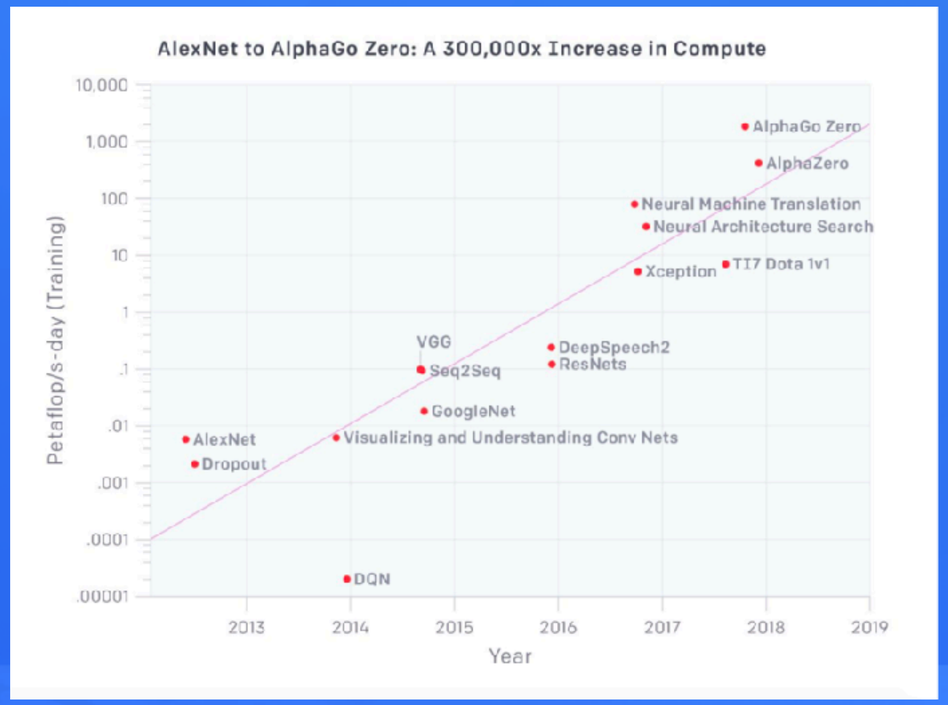
*–Jeff Dean, interview with VentureBeat 2019/12/13*

# 趋势2：后摩尔时代，XPU 百花齐放

**CPU 算力发展放缓，**
**进入后摩尔时代**



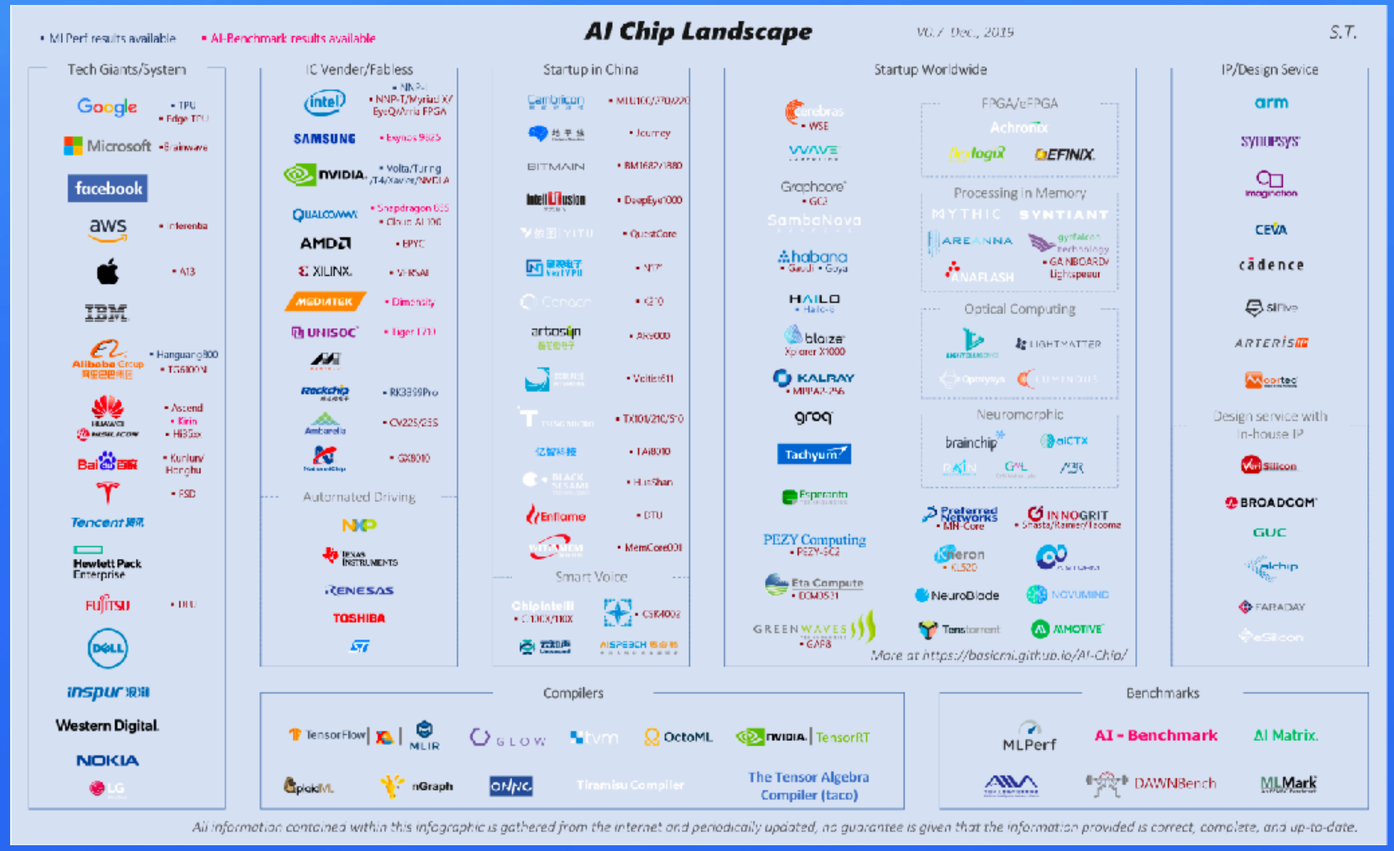

**ML模型对算力的要求指数增长**




Alibaba Group 阿里巴巴集团 | TAO TECHNOLOGY 淘系技术部

MNN Mobile Neural Network

# 趋势2：后摩尔时代，XPU 百花齐放




Alibaba Group 阿里巴巴集团 | TAO TECHNOLOGY 淘系技术部

MNN Mobile Neural Network

# 趋势3：NLP逐步成熟

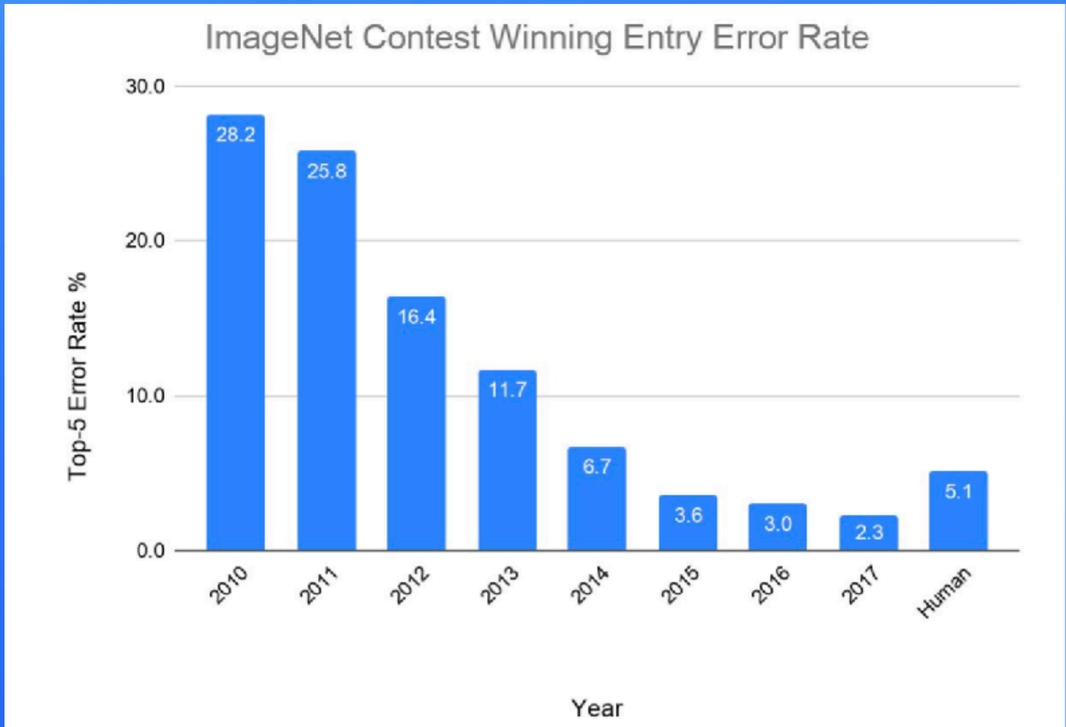



Figure 1: ImageNet classification contest winner accuracy over time

# 趋势3：NLP逐步成熟

- BERT (2018), ALBERT (2019)

**Leaderboard**

| Model | Report Time | Institute | RACE | RACE-M | RACE-H |
|---|---|---|---|---|---|
| Human Ceiling Performance | Apr. 2017 | CMU | *94.5* | *95.4* | *94.2* |
| Amazon Mechanical Turker | Apr. 2017 | CMU | 73.3 | 85.1 | 69.4 |
| ALBERT (ensemble) | Sept. 26th 2019 | Google Research & TTIC | **89.4** | **91.2** | **88.6** |
| ALBERT | Sept. 26th 2019 | Google Research & TTIC | 86.5 | 89.0 | 85.5 |

RACE Leaderboard (阅读理解)

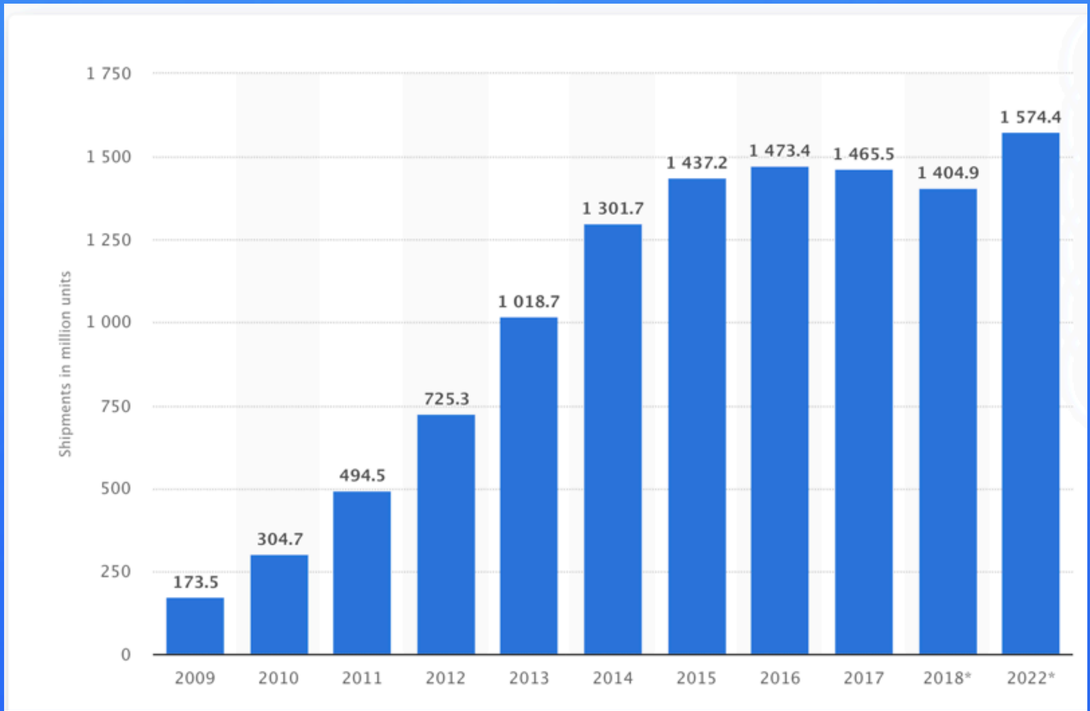
| Rank | Model | EM | F1 |
|---|---|---|---|
| | Human Performance<br>*Stanford University*<br>(Rajpurkar & Jia et al. '18) | 86.831 | 89.452 |
| 1<br>Nov 06, 2019 | ALBERT + DAAF + Verifier (ensemble)<br>*PINGAN Omni-Sinitic* | **90.002** | **92.425** |
| 2<br>Sep 18, 2019 | ALBERT (ensemble model)<br>*Google Research & TTIC*<br>https://arxiv.org/abs/1909.11942 | 89.731 | 92.215 |

SQuAD Leaderboard (阅读理解)

Alibaba Group 阿里巴巴集团 | TAO TECHNOLOGY 淘系技术部

MNN Mobile Neural Network

# 趋势4： 手机端 —> IOT 端




手机出货量预测

**Worldwide smart speaker shipments and annual growth**

**Canalys smart speaker market pulse: Q3 2019**

| Vendor | Q3 2019 shipments (million) | Q3 2019 market share | Q3 2018 shipments (million) | Q3 2018 market share | Annual growth |
|---|---|---|---|---|---|
| Amazon | 10.4 | 36.6% | 6.3 | 31.9% | 65.9% |
| Alibaba | 3.9 | 13.6% | 2.2 | 11.1% | 77.6% |
| Baidu | 3.7 | 13.1% | 1.0 | 4.9% | 290.1% |
| Google | 3.5 | 12.3% | 5.9 | 29.8% | -40.1% |
| Xiaomi | 3.4 | 12.0% | 1.9 | 9.7% | 77.7% |
| Others | 3.6 | 12.5% | 2.5 | 12.5% | 44.0% |
| **Total** | **28.6** | **100.0%** | **19.7** | **100.0%** | **44.9%** |

Note: percentages may not add up to 100% due to rounding

Source: Canalys Smart Speaker Analysis (sell-in shipments), November 2019

canalys

智能音箱出货量

Alibaba Group 阿里巴巴集团 | TAO TECHNOLOGY 淘系技术部

MNN Mobile Neural Network

# 端智能在手机淘宝的应用

- 改变原有的流程
- 全新的交互方式

Alibaba Group 阿里巴巴集团 | TAO TECHNOLOGY 淘系技术部

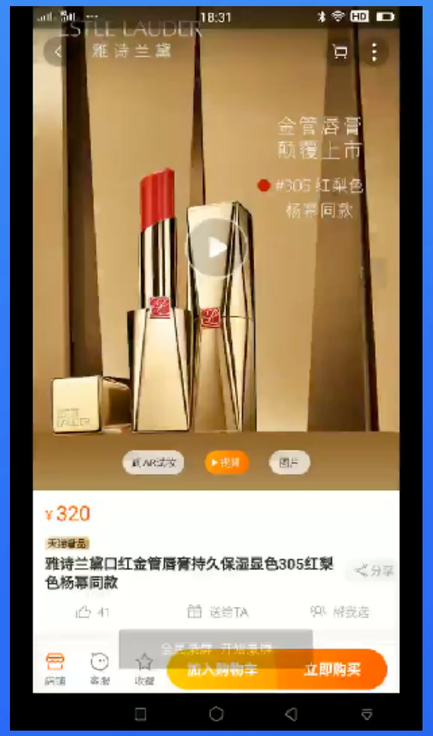
MNN Mobile Neural Network

# 改变原有的流程

- 个性化：千人千模。根据用户个人数据，personalize global model.
- 时效性+省服务端算力：e.g. 端上重排。
- 隐私性：Federated Learning。探索性。


端上重排

# 全新的交互体验




美妆：口红 (小程序)


染发


拍立淘 扫Logo (NPU)

Alibaba Group 阿里巴巴集团

TAO TECHNOLOGY 淘系技术部

MNN Mobile Neural Network

# 小结：端智能在手机上应用

- 非端不可
  - 可离线
  - 隐私
  - 实时性
  - 省云资源

# MNN介绍

小福利环节！！！！

https://github.com/alibaba/MNN

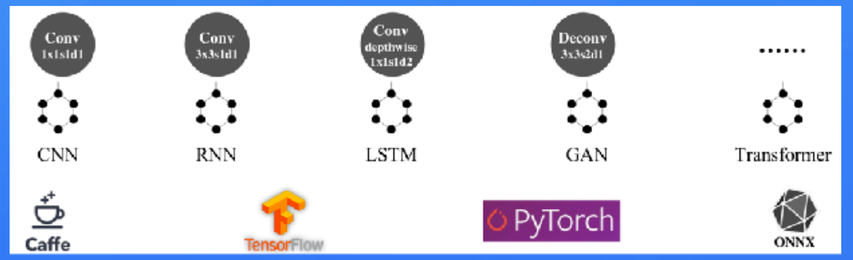
- 核心价值
- 主要模块
- 使用流程

Alibaba Group 阿里巴巴集团 | TAO TECHNOLOGY 淘系技术部

MNN Mobile Neural Network

# MNN核心价值

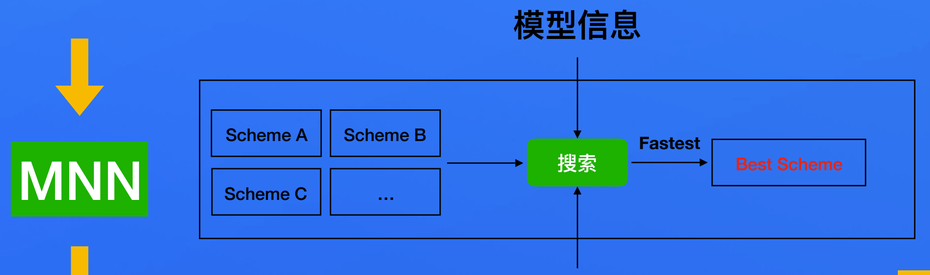
模型




算子碎片化

框架碎片化

推理引擎




设备


硬件、OS碎片化

资源受限

# MNN核心价值

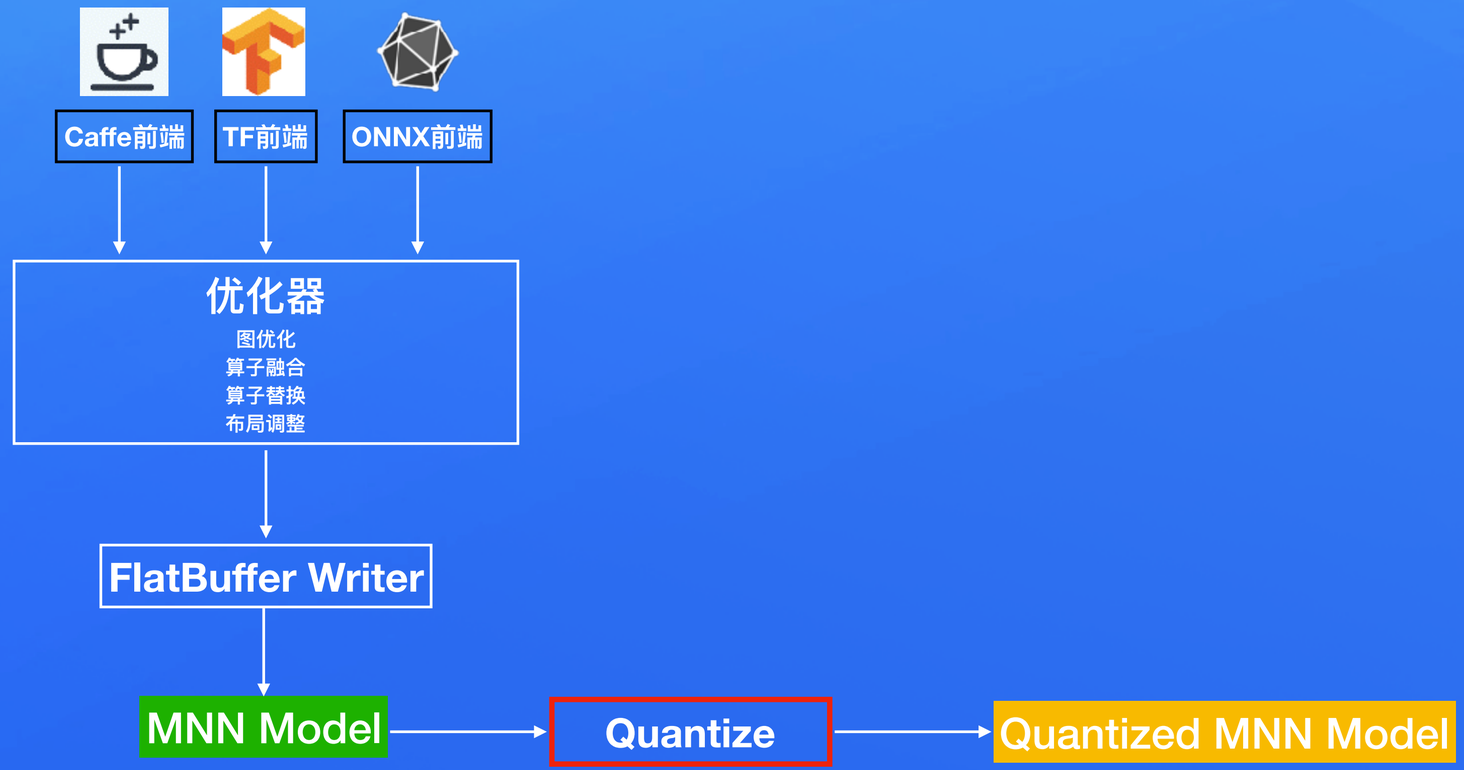
- 性能：Arm CPU上MNN的单线程推理性能大约是TFlite的2~4倍；4线程推理性能大约是TFLite的10~20倍[1]
- 大小: MNN的静态库大小约为TFlite的66.7%。[2]
- 异构设备支持（**NPU**，GPU，Armv8.2, etc.）
- 通用性：12月补充了约200个OP

[1] 2019年9月数据，Mate 20/Mi6, MobileNet-v1/SqueezeNet-v1.1/ResNet-18
[2] MNN + OpenCL backend: 1.2M + 334K; TFLite + NNApi backend: 2.3M

# MNN 主要模块

- 模型转换
- 模型压缩（量化）
- 推理引擎 (Next time)

# 模型转换MNNConvert



# 模型量化工具


PreTrained Models
MNN Model
校准训练集
MNN 量化工具包
无训练量化
ADMM
KL散度
带训练量化（WIP）
Quantized MNN Model
MNN 推理引擎

# MNN使用流程 - Live Coding Session

- Rough steps (when everything “just works”):
  - 1. Compile MNNConvert + MNN lib.
  - 2. tools/script/get_model.sh (download and convert)
  - 3. (Optional) Model quantization tool [w/o training.]
  - 4. In the demo app, load MNN model and create an interpreter + session.

# MNN未来展望

- 易用性：开箱即用的工具箱MNN Kit；预编译包
- 端上训练支持：带训练的量化；迁移学习；从0开始训练
- 工作台
- MNN + NPU
- MNN 小程序

# 总结 回顾

- 整体端智能的趋势、挑战、特点
- 手机淘宝中端智能的应用玩法
- MNN介绍

# 后续直播内容

- MNN 推理引擎详解
- MNN Kit
- MNN 一站式工作台

# Q & A

Alibaba Group 阿里巴巴集团

TAO TECHNOLOGY 淘系技术部

MNN Mobile Neural Network

Alibaba Group 阿里巴巴集团 | TAO TECHNOLOGY 淘系技术部

MNN Mobile Neural Network

# MNN学院

## https://github.com/alibaba/MNN

MNN团队招聘。扫一扫。投简历

「淘系技术」微信公众号

MNN 开源讨论钉钉二群