

# 推理系统系列

# 推理引擎



## ZOMI



BUILDING A BETTER CONNECTED WORLD

Ascend & MindSpore

[www.hiascend.com](http://www.hiascend.com)  
[www.mindspore.cn](http://www.mindspore.cn)

# Talk Overview

## 1. 推理系统介绍

- 推理系统与推理引擎
- 推理系统的流程全景
- 推理系统架构
- 推理引擎介绍

## 2. 模型小型化

- NAS神经网络搜索
- CNN小型化结构
- Transform小型化结构

## 3. 离线优化压缩

- 低比特量化
- 二值化网络
- 模型模型剪枝
- 模型模型蒸馏

## 4. 部署和运行优化

- 图转换优化（算子融合/重排/替换）
- 并发执行与内存分配
- 动态batch与bin Packing

# Talk Overview

1. 推理系统与推理引擎
2. 推理系统的流程全景
3. 推理系统架构
4. 推理引擎介绍
  - 推理引擎特点
  - 技术挑战
  - 整体架构
  - 工作流程

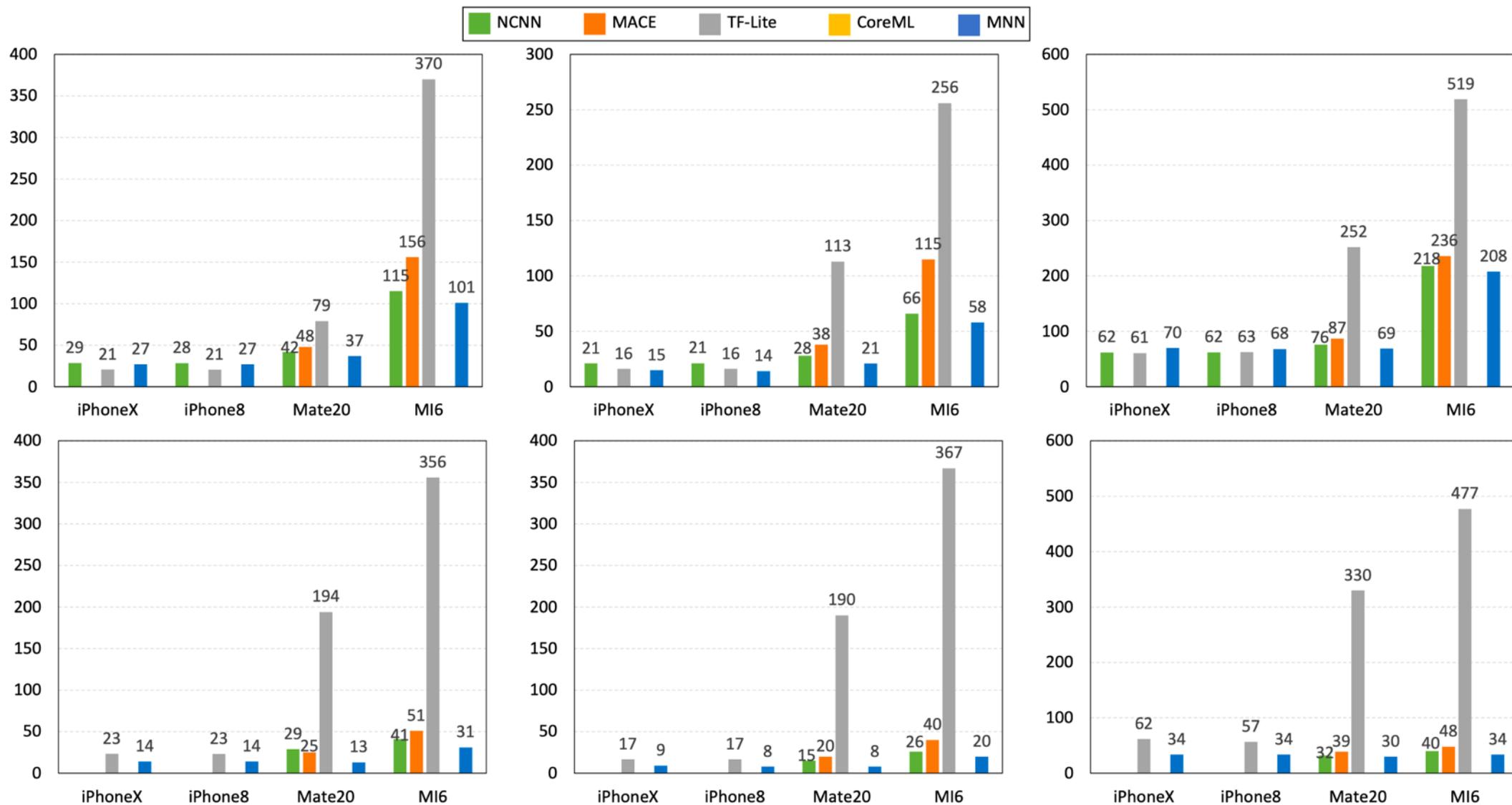
# 推理引擎特点

# Feature

轻量、通用、易用、高效

# High performance

1. 需要对 iOS / Android / PC 不同硬件架构和操作系统进行适配，单线程下运行深度学习模型达到设备算力峰值。
2. 针对主流加速芯片进行深度调优，如 OpenCL 侧重于推理性能极致优化，Vulkan 方案注重较少初始化时间。
3. 编写SIMD代码或手写汇编以实现核心运算，充分发挥芯片算力，针对不同kernel算法提升性能。
4. 支持不同精度计算以提升推理性能，并对 ARMv8.2 和 AVX512 架构的相关指令进行了适配。



# Lightness

1. 主体功能无任何依赖，代码精简，可以方便地部署到移动设备和各种嵌入式设备中。
2. 支持 Mini 编辑选项进一步降低包大小，大约能在原库体积基础上进一步降低体积。
3. 支持模型更新精度 FP16/Int8 压缩与量化，可减少模型50% - 75% 的体积。



# Lightness

1. 主体功能无任何依赖，代码精简，可以方便地部署到移动设备和各种嵌入式设备中。
2. 支持 Mini 编辑选项进一步降低包大小，大约能在原库体积基础上进一步降低体积。
3. 支持模型更新精度 FP16/Int8 压缩与量化，可减少模型50% - 75% 的体积。



FreeBuds Lipstick



WATCH GT 3 (42mm)



WATCH GT 3 (46mm)



WATCH GT Runner



WATCH FIT mini

# Versatility

1. 支持 Tensorflow、PyTorch、MindSpore、ONNX 等主流模型文件格式。
2. 支持 CNN / RNN / GAN / Transformer 等主流网络结构。
3. 支持多输入多输出，任意维度输入输出，支持动态输入，支持带控制流的模型。
4. 支持 服务器 / 个人电脑 / 手机 及具有POSIX接口的嵌入式设备。
5. 支持 Windows / iOS 8.0+ / Android 4.3+ / Linux / ROS 等操作系统。

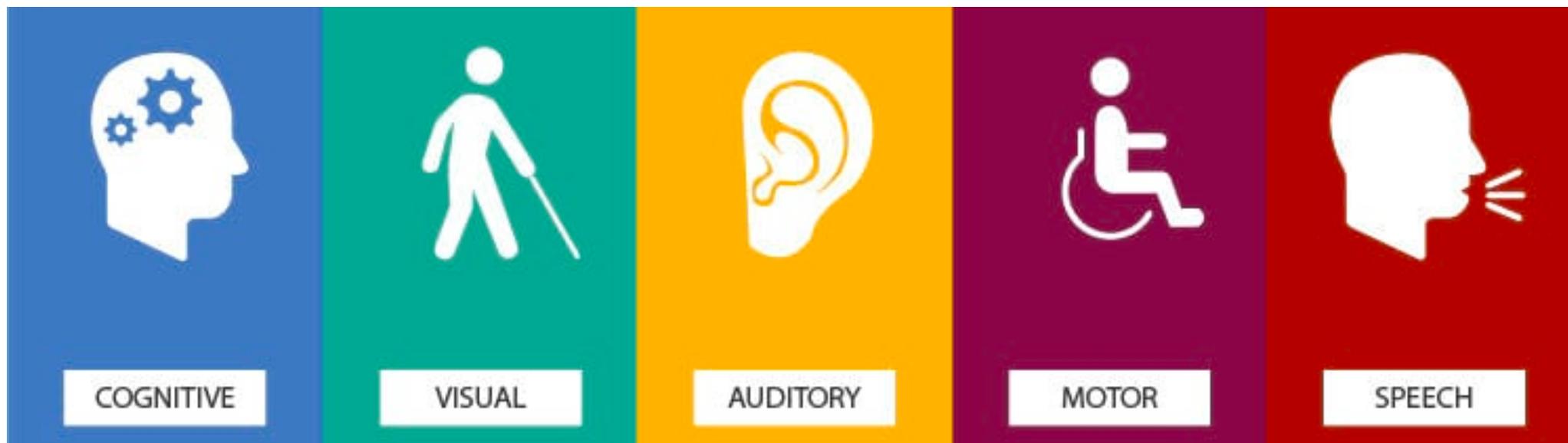
# Versatility

1. 支持 服务器 / 个人电脑 / 手机 及具有POSIX接口的嵌入式设备。
2. 支持 Windows / iOS 8.0+ / Android 4.3+ / Linux / ROS 等操作系统。



# Accessibility

1. 支持使用算子进行常用数值计算，覆盖 numpy 常用功能
2. 提供 CV/NLP 等任务的常用模块
3. 支持各平台下的模型训练
4. 支持丰富的 API 接口



# 技术挑战 Challenge

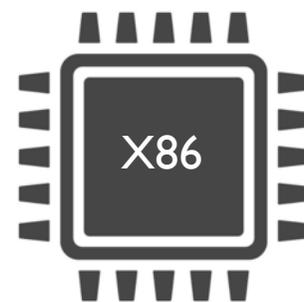
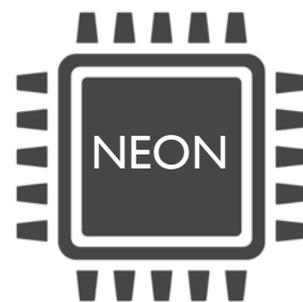
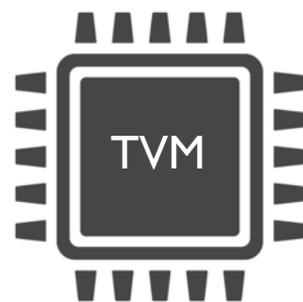
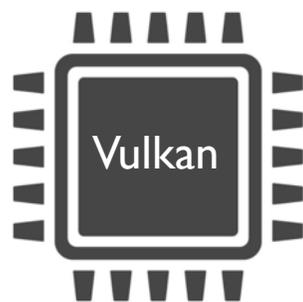
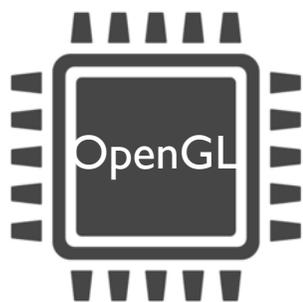
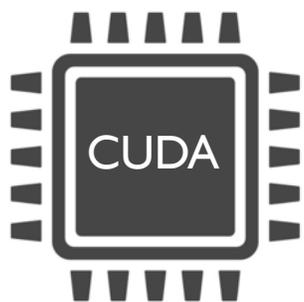


## 需求复杂 vs 程序大小

- AI 模型本身包含众多算子，如 PyTorch有1200+ 算子、Tensorflow 接近 2000+ 算子，推理引擎需要用有限算子去实现不同框架训练出来 AI 模型所需要的算子。
- AI 应用除去模型推理之外，也包含数据前后处理所需要的数值计算与图像处理，不能引入大量的三方依赖库，因此需要进行有限支持。

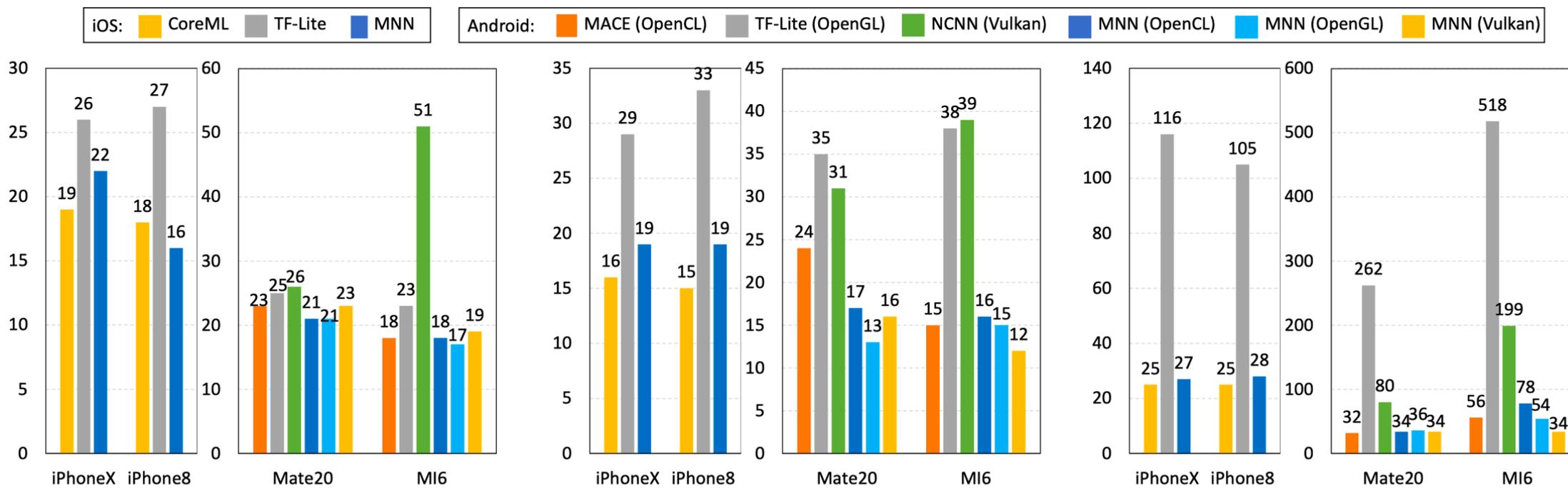
## 算力需求 vs 资源碎片化

- AI 模型往往计算量很大，需要推理引擎对设备上的计算资源深入适配，持续进行性能优化，以充分发挥设备的算力。
- 计算资源包括 CPU，GPU，DSP 和 NPU，其各自编程方式是碎片化，需要逐个适配，开发成本高，也会使程序体积膨胀。



# 执行效率 vs 模型精度

- 高效的执行效率需要网络模型变小，但是模型的精度希望尽可能的高；
- 云测训练的网络模型精度尽可能的高，转移到端侧期望模型变小但是保持相同的精度；



# 参考文献



BUILDING A BETTER CONNECTED WORLD

THANK YOU

Copyright©2014 Huawei Technologies Co., Ltd. All Rights Reserved.

The information in this document may contain predictive statements including, without limitation, statements regarding the future financial and operating results, future product portfolio, new technology, etc. There are a number of factors that could cause actual results and developments to differ materially from those expressed or implied in the predictive statements. Therefore, such information is provided for reference purpose only and constitutes neither an offer nor an acceptance. Huawei may change the information at any time without notice.