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

MNN Mobile Neural Network

MNN学院

深入浅出谈 MNN 原理，为什么能这么快?

讲师：吴子奇（花名：明弈）

阿里巴巴淘系技术无线开发专家

推理引擎架构师

- 深入了解tensorflow/pytorch的核心，对其有突出的贡献，或主导过类似框架的核心架构设计
- 深入了解tvm的核心，对其有突出的贡献，对模型编译/IR设计/端上高性能代码生成有深入洞察
- 对机器学习理论有深入研究，特别模型架构搜索/模型结构优化/模型压缩有较高的理论功底
- 对卷积/矩阵乘法/图像/视频编解码等高性能计算有深入的理解和洞察
- 深度学习领域3年以上工作经验，有一定的知名度和影响力

shuhui.sh@alibaba-inc.com

无线AR专家

- 从事移动端AR相关工作2年以上，对AR算法及工程有深刻理解；
- 熟悉计算机视觉相关领域算法，以及图形学、图像渲染OpenGL等领域知识；
- 具有扎实的工程实现能力，熟练使用C/C++及一门脚本语言；
- 有好奇心，创新思维，敢承担，可实战，团队合作，沟通能力佳；
- 有AR试穿试戴、AR互动玩法及其他商业化AR产品研发经验优先；

zhaomi.zm@alibaba-inc.com

Android 客户端

- 熟悉Android Framework层，并有一定Android源码阅读经验
- 了解响应式编程思想，了解 flutter, weex ,kotlin 等技术优先考虑
- 热爱技术，较强逻辑思维能力， 快速学习能力；
- 做事情严谨踏实，团队协作能力强，良好的沟通能力和团队合作精神

difei.zdf@alibaba-inc.com

淘系技术部 - 基础平台部
广纳贤才

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

MNN Mobile Neural Network

Agenda

MNN 为什么快?

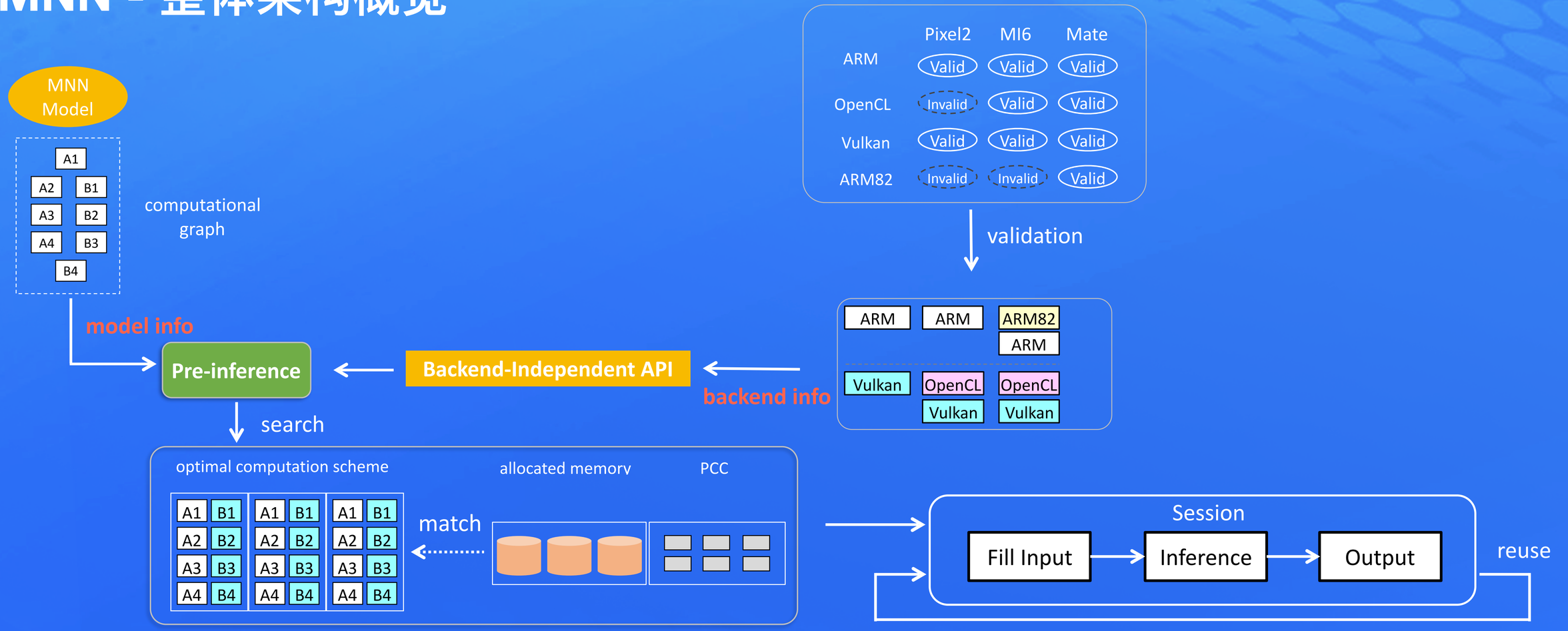
1. MNN 整体架构优势
2. MNN 核心细节优化
3. MNN 标准性能评测

MNN 整体架构优势

适用于移动端高效推理的半自动搜索架构

1. 计算图线性拆分
2. 预分配资源
3. 在线方案选择

MNN - 整体架构概览

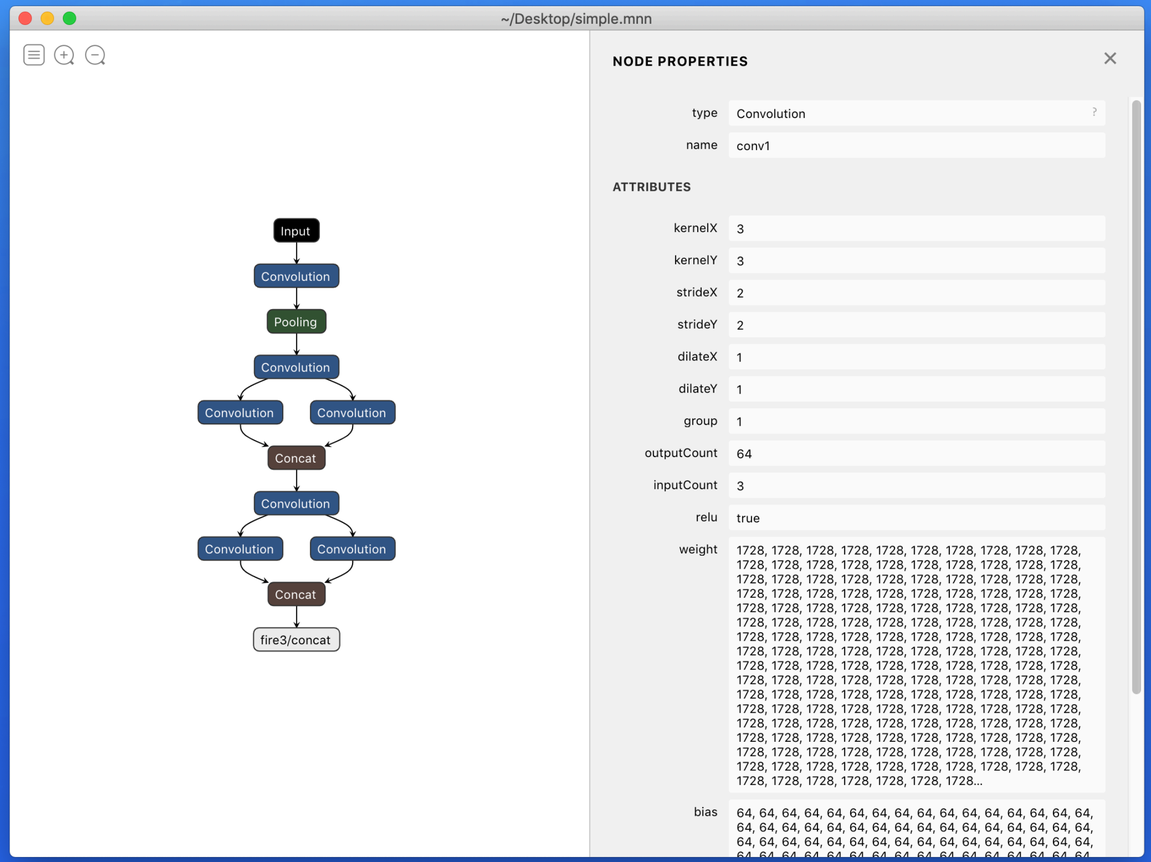


Alibaba Group 阿里巴巴集团

淘系技术部 TAO TECHNOLOGY

MNN Mobile Neural Network

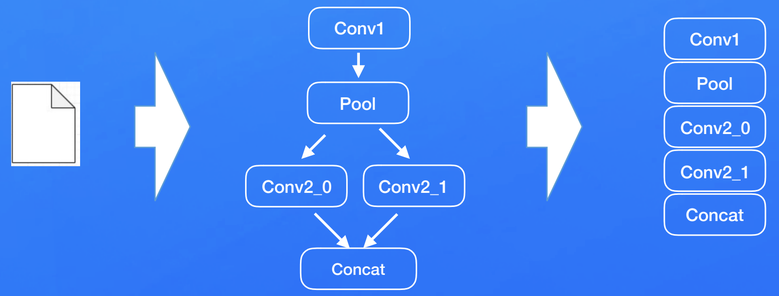
MNN - 直观意义上理解的神经网络



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

MNN Mobile Neural Network

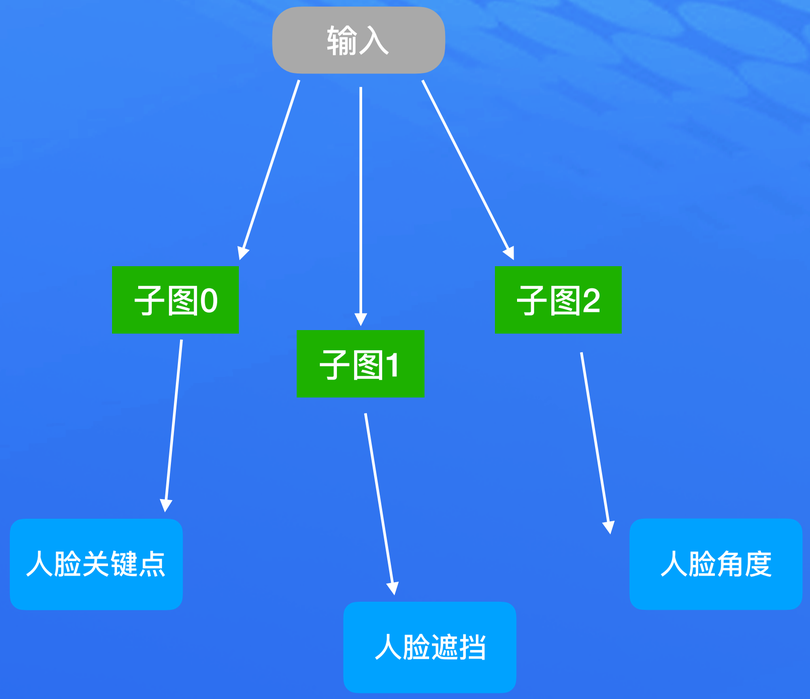
MNN - 计算图线性拆分



模型文件　　模型加载

线性化

更进一步　子图拆分



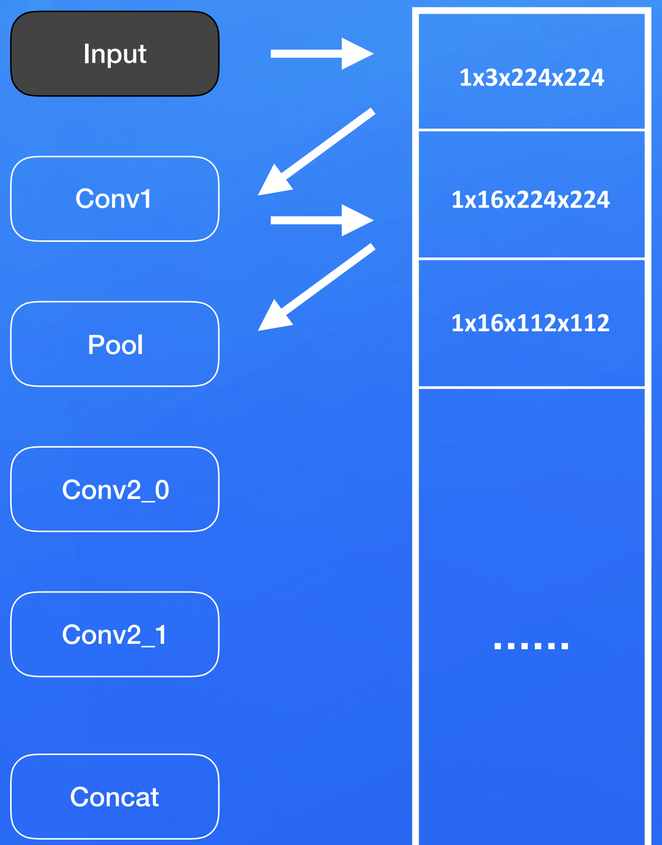
- 从实现逻辑上，大幅度降低图计算依赖所需的相关判断，线性执行即可。

MultiTask模型多backend并行

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

MNN Mobile Neural Network

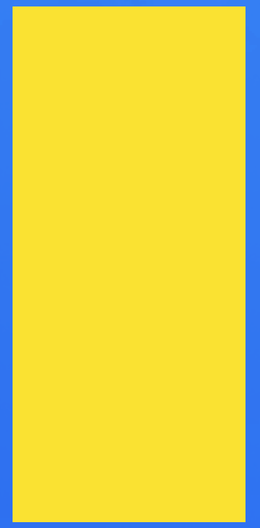
MNN - 预分配资源



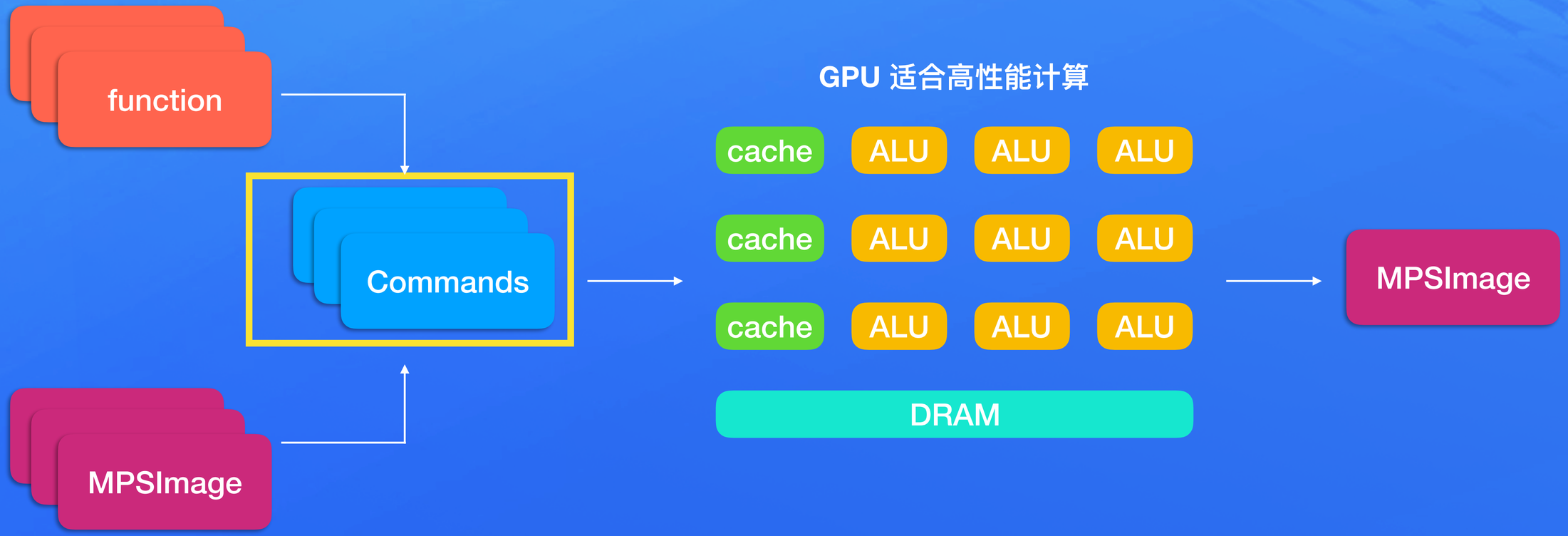
动静分离

内存分配对于高效推理是不可忽略的一环

内存池

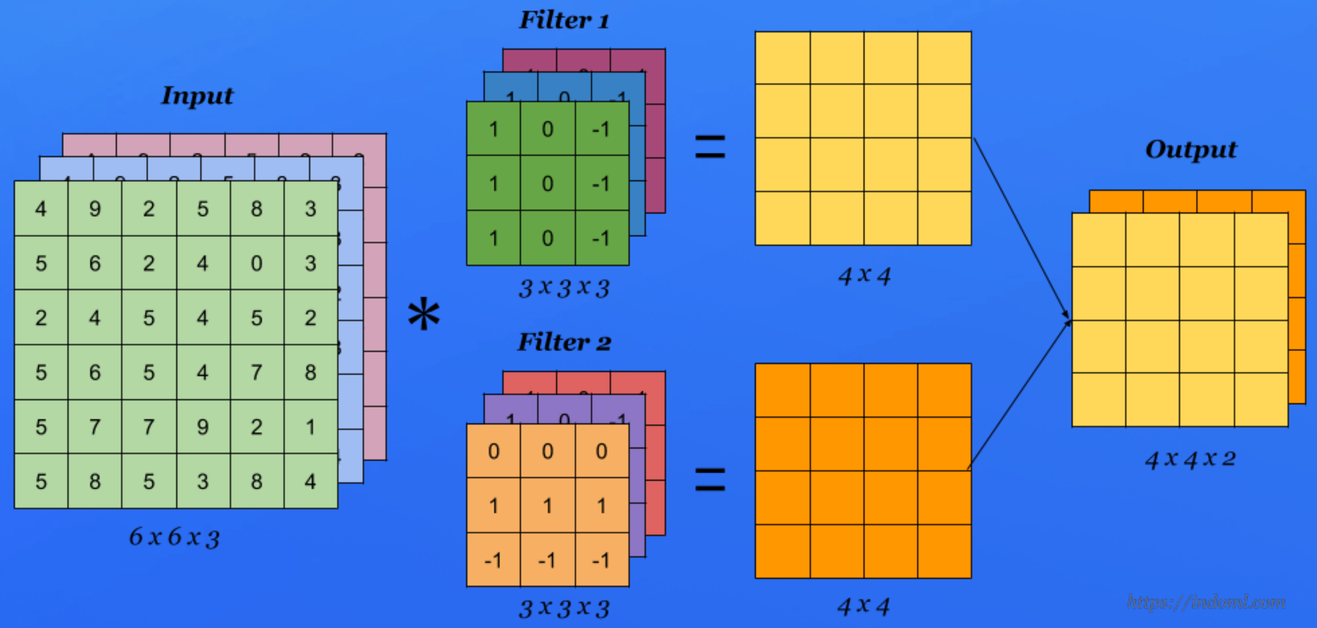


MNN - 预分配资源（GPU）



预推理统一处理相关的CommandBuffer / CommandDescription，节省后续推理耗时

MNN - 在线方案选择 （以卷积为例）

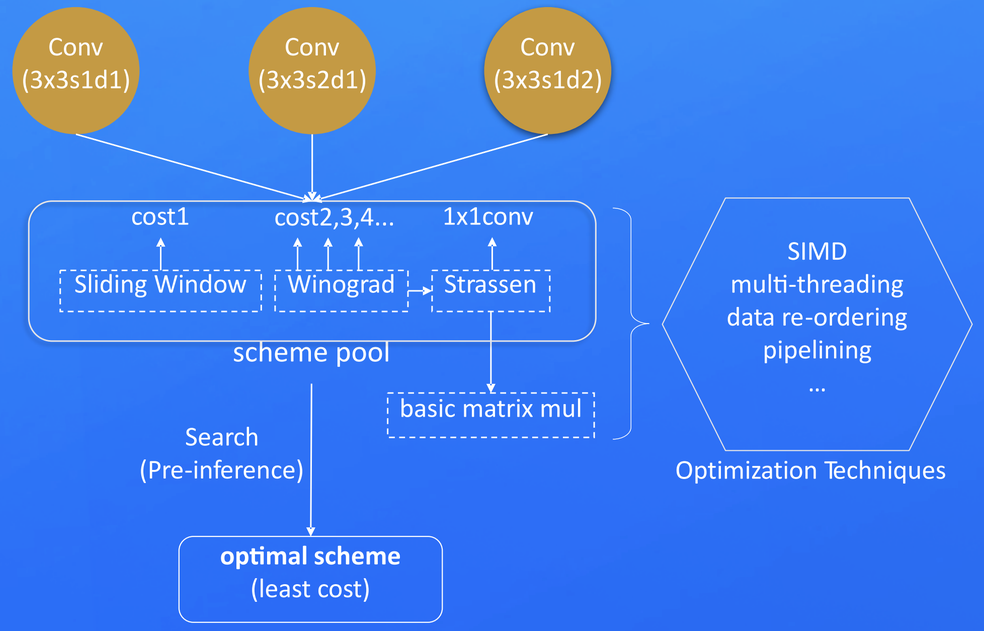


Alibaba Group 阿里巴巴集团

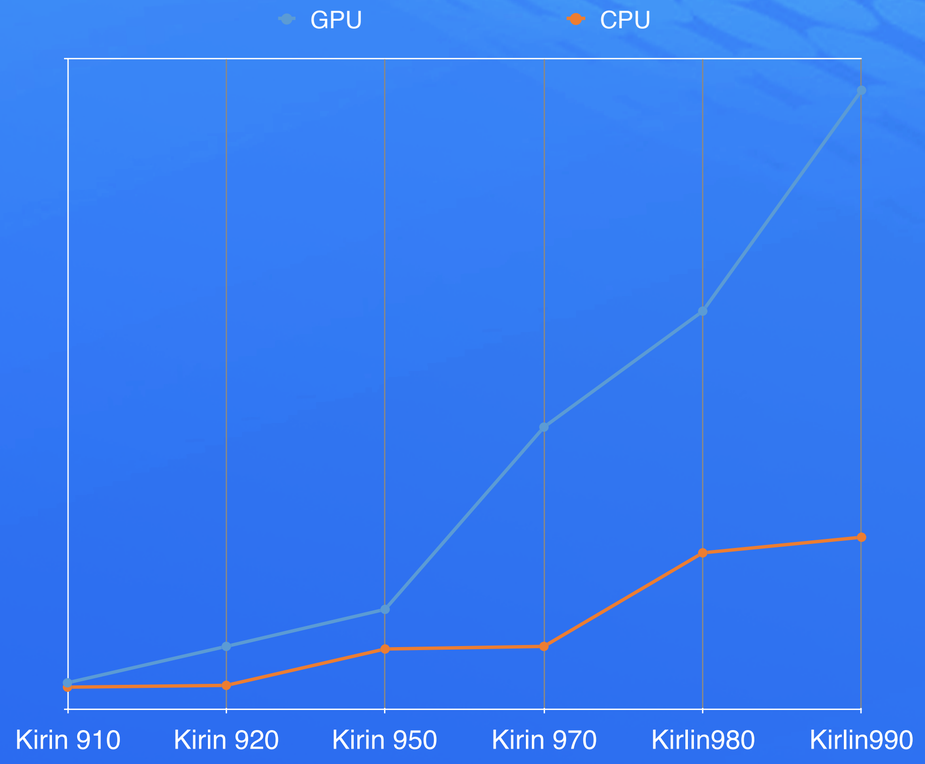
淘系技术部 TAO TECHNOLOGY

MNN Mobile Neural Network

MNN - 在线方案选择



根据输入选择最优的计算方案



根据Flops + 配置建立时间选择最优后端

Alibaba Group 阿里巴巴集团

淘系技术部 TAO TECHNOLOGY

MNN Mobile Neural Network

MNN 核心细节优化

1 高性能手写汇编

2 并行数据排布

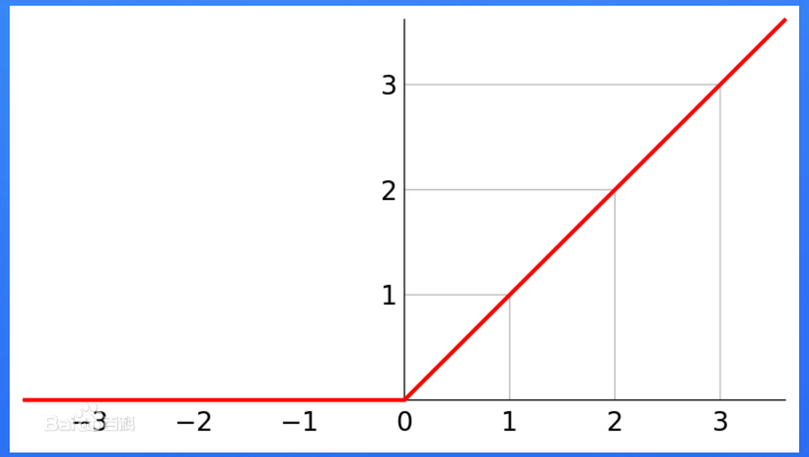
3 核心算法优化

Alibaba Group 阿里巴巴集团

淘系技术部 TAO TECHNOLOGY

MNN Mobile Neural Network

MNN - 高性能手写汇编




```
void ReluForward(float* dst, const float* src, size_t sizeDiv4)
{
    for (int i=0; i<4*sizeDiv4; ++i)
    {
        dst[i] = src[i] >0 ? src[i] : 0;
    }
}
```

纯 C 实现

```
push {lr}
vmov.i32 q15, #0

cmp r2, #0
beq End //跳转：beq 表示 r2 等于0时跳转
Loop://标志，供跳转用
vld1.32 {q0}, [r1]!
vmax.f32 q0, q0, q15
vst1.32 {q0}, [r0]!
subs r2, r2, #1// 这一句 相当于 sub r2, r2, #1  &&  cmp r2, #0
bne Loop //跳转：bne 表示 r2 不等于0时跳转

End:
pop {pc}
```

基础汇编

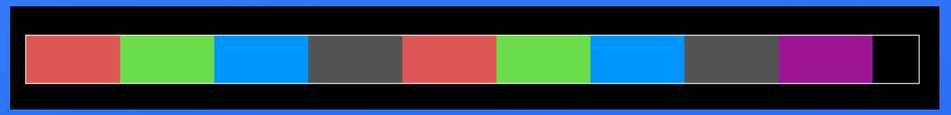
```
push {lr}
vmov.i32 q15, #0

L4:
cmp r2, #3
ble L1

vld1.32 {q0, q1}, [r1]!
vmax.f32 q0, q0, q15
vld1.32 {q2, q3}, [r1]!
vmax.f32 q1, q1, q15

sub r2, r2, #4
cmp r2, #3
ble L4End

L4Loop:
vst1.32 {q0, q1}, [r0]!
vmax.f32 q2, q2, q15
vld1.32 {q0, q1}, [r1]!
vmax.f32 q3, q3, q15
vst1.32 {q2, q3}, [r0]!
vmax.f32 q0, q0, q15
vld1.32 {q2, q3}, [r1]!
sub r2, r2, #4
vmax.f32 q1, q1, q15
cmp r2, #4
bge L4Loop

L4End:
vst1.32 {q0, q1}, [r0]!
vmax.f32 q2, q2, q15
vmax.f32 q3, q3, q15
vst1.32 {q2, q3}, [r0]!

L1:
cmp r2, #0
beq End

L1Loop:
vld1.32 {q0}, [r1]!
vmax.f32 q0, q0, q15
vst1.32 {q0}, [r0]!
subs r2, r2, #1
bne L1Loop

End:
pop {pc}
```

汇编优化

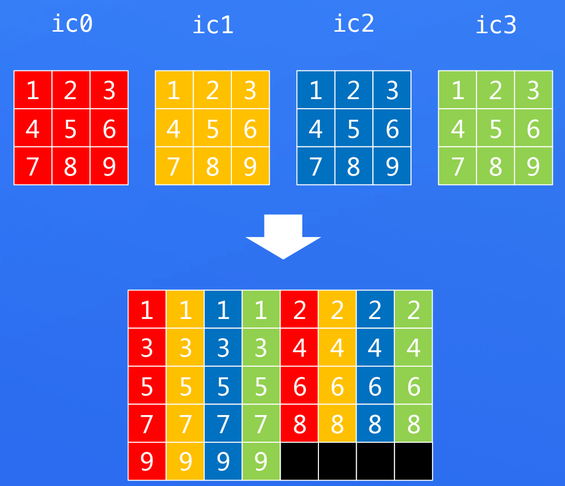
MNN - 合理数据排布

NC4HW4

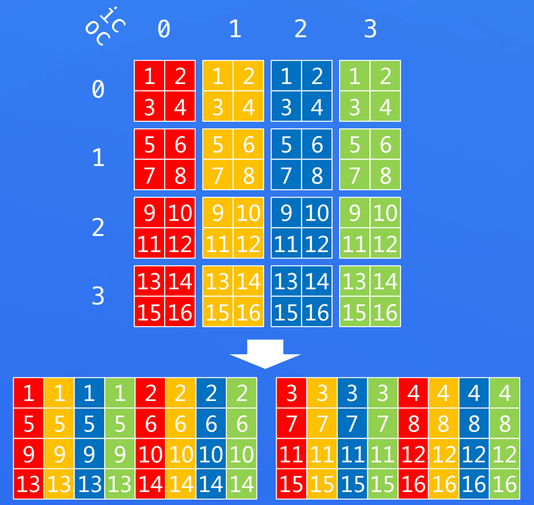
适合利用SIMD并行加速



特征重排



权重重排



Alibaba Group 阿里巴巴集团

淘系技术部 TAO TECHNOLOGY

MNN Mobile Neural Network

MNN - 大矩阵乘法优化

```
SQUARE-MATRIX-MULTIPLY(A, B)
n = A.rows
let C be a new nxn matrix
for i = 1 to n
    for j = 1 to n
        c[i][j] = 0
        for k = 1 to n
            c[i][j] += a[i][k] * b[k][j]
return C
```

朴素的 NxN 矩阵乘法，复杂度$O(n^3)$

$$A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}, B = \begin{bmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{bmatrix}, C = \begin{bmatrix} C_{11} & C_{12} \\ C_{21} & C_{22} \end{bmatrix}$$

$$\begin{bmatrix} C_{11} & C_{12} \\ C_{21} & C_{22} \end{bmatrix} = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \cdot \begin{bmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{bmatrix}$$

$$C11 = A11 \cdot B11 + A12 \cdot B21$$
$$C12 = A11 \cdot B12 + A12 \cdot B22$$
$$C21 = A21 \cdot B11 + A22 \cdot B21$$
$$C22 = A21 \cdot B12 + A22 \cdot B22$$

复杂度仍然是$O(n^3)$

MNN - 大矩阵乘法优化

$$\begin{pmatrix} A1 & A2 & A3 & A4 \\ A5 & A6 & A7 & A8 \\ A9 & A10 & A11 & A12 \\ A13 & A14 & A15 & A16 \end{pmatrix} = \begin{pmatrix} a11 & a12 \\ a21 & a22 \end{pmatrix}, \quad \begin{pmatrix} B1 & B2 & B3 & B4 \\ B5 & B6 & B7 & B8 \\ B9 & B10 & B11 & B12 \\ B13 & B14 & B15 & B16 \end{pmatrix} = \begin{pmatrix} b11 & b12 \\ b21 & b22 \end{pmatrix}, \quad \begin{pmatrix} C1 & C2 & C3 & C4 \\ C5 & C6 & C7 & C8 \\ C9 & C10 & C11 & C12 \\ C13 & C14 & C15 & C16 \end{pmatrix} = \begin{pmatrix} c11 & c12 \\ c21 & c22 \end{pmatrix}$$

$$\begin{pmatrix} a11 & a12 \\ a21 & a22 \end{pmatrix} * \begin{pmatrix} b11 & b12 \\ b21 & b22 \end{pmatrix} = \begin{pmatrix} c11 & c12 \\ c21 & c22 \end{pmatrix}$$

$$\begin{aligned}
s_1 &\equiv (a_{11} + a_{22})(b_{11} + b_{22}) & m_1 &= a_{11}b_{11} & u_1 &= m_1 + m_2 + v_4 - v_7 \\
s_2 &\equiv s_1 - a_{11} & m_2 &= a_{12}b_{21} & u_2 &= m_1 + m_6 + v_2 + v_4 & c_{12} &= u_5 \\
v_2 &\equiv (a_{21} + a_{22})(b_{11}) & m_3 &= s_4 b_{22} & u_3 &= u_2 + m_7 & c_{21} &= u_6 \\
s_3 &\equiv a_{11} - a_{21} & m_4 &= a_{22}t_4 & u_4 &= u_2 + m_5 & c_{22} &= u_7 \\
v_3 &\equiv (a_{11})(b_{12} - b_{22}) & m_5 &= s_1 t_1 & u_5 &= u_4 + m_3 \\
s_4 &= a_{12} - s_2 & m_6 &= s_2 t_2 & u_6 &= u_3 - m_4 \\
v_4 &\equiv (a_{22})(b_{21} - b_{11}) & m_7 &= s_3 t_3 & u_7 &= u_3 + m_5 \\
t_1 &= b_{21} - b_{11} \\
v_5 &\equiv (a_{11} + a_{12})(b_{22}) \\
t_2 &\equiv b_{22} - t_1 \\
v_6 &\equiv (a_{21} - a_{11})(b_{11} + b_{12}) \\
t_3 &\equiv b_{22} - b_{12} \\
v_7 &\equiv (a_{12} - a_{22})(b_{21} + b_{22}) \\
t_4 &\equiv t_2 - b_{21}
\end{aligned}$$

$$c_{11} = u_1 \qquad c_{12} = u_5 \qquad c_{21} = u_6 \qquad c_{22} = u_7$$

$$c_{22} = v_1 + v_3 - v_2 + v_6$$

使用4+8次矩阵加减替代矩阵乘法(先左乘法)

使用4+7次矩阵加减替代矩阵乘法(先右乘法)

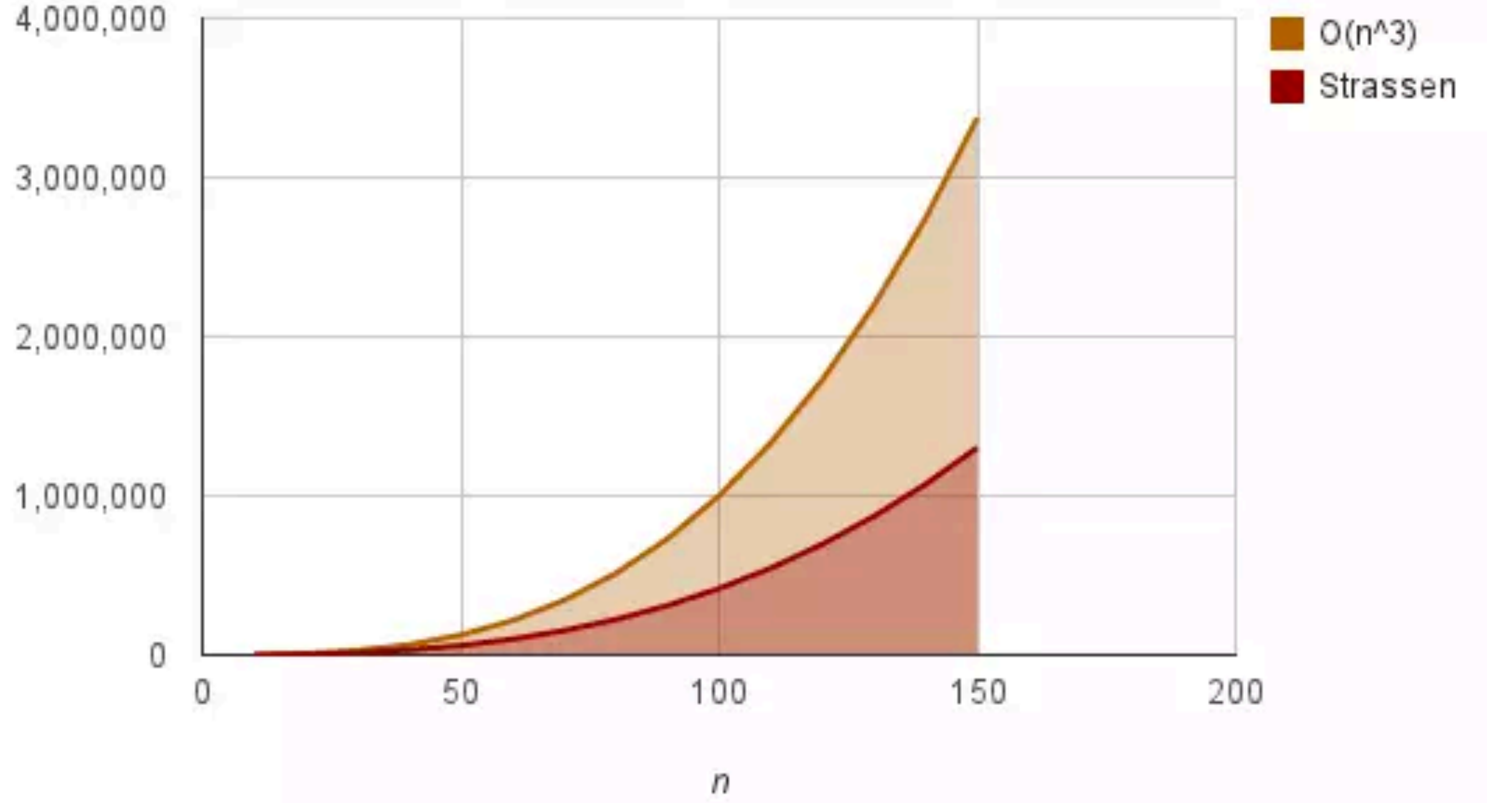
$O(n^3)$

$O(n^{2.81})$

矩阵较大时
矩阵乘法远慢于矩阵加减
收益较为明显

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

MNN Mobile Neural Network


4,000,000
3,000,000
2,000,000
1,000,000
0
0
50
100
150
200
n
O(n^3)
Strassen

MNN
标准性能评测


Alibaba Group
阿里巴巴集团


淘系技术部
TAO TECHNOLOGY

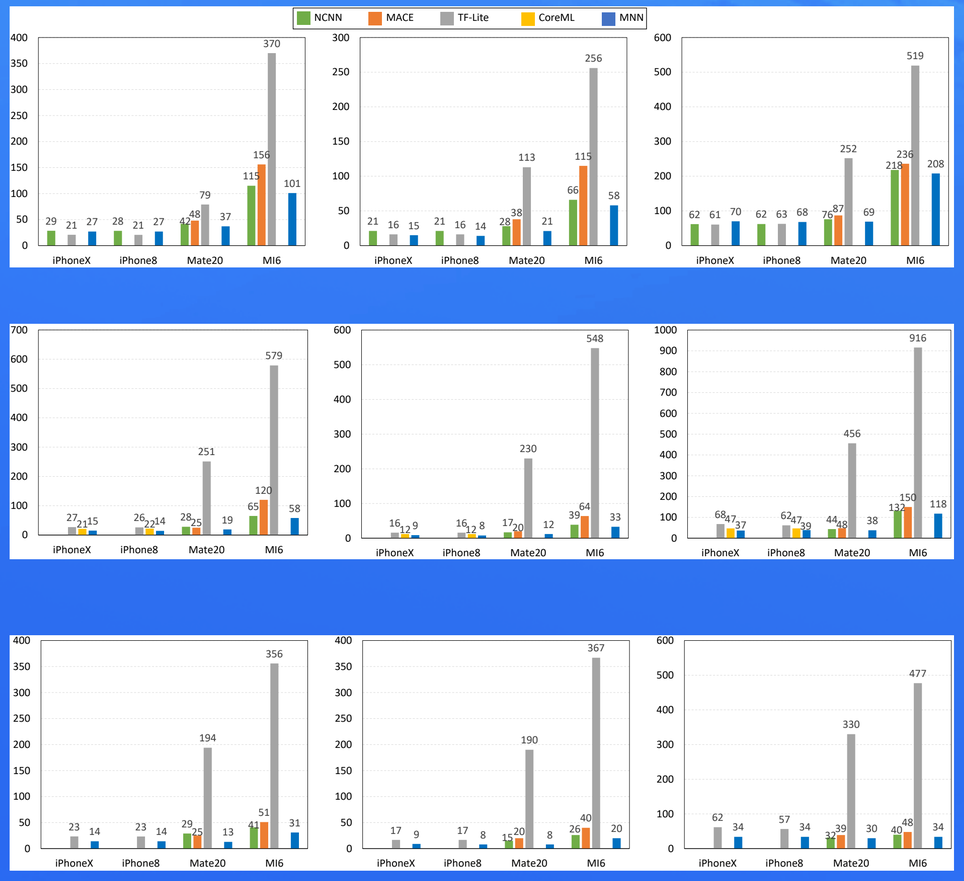

MNN
Mobile Neural Network

标准性能评测 CPU

单线程　　MobileNet v1 / SqueezeNet v1.1 / ResNet-18

双线程　　MobileNet v1 / SqueezeNet v1.1 / ResNet-18

四线程　　MobileNet v1 / SqueezeNet v1.1 / ResNet-18

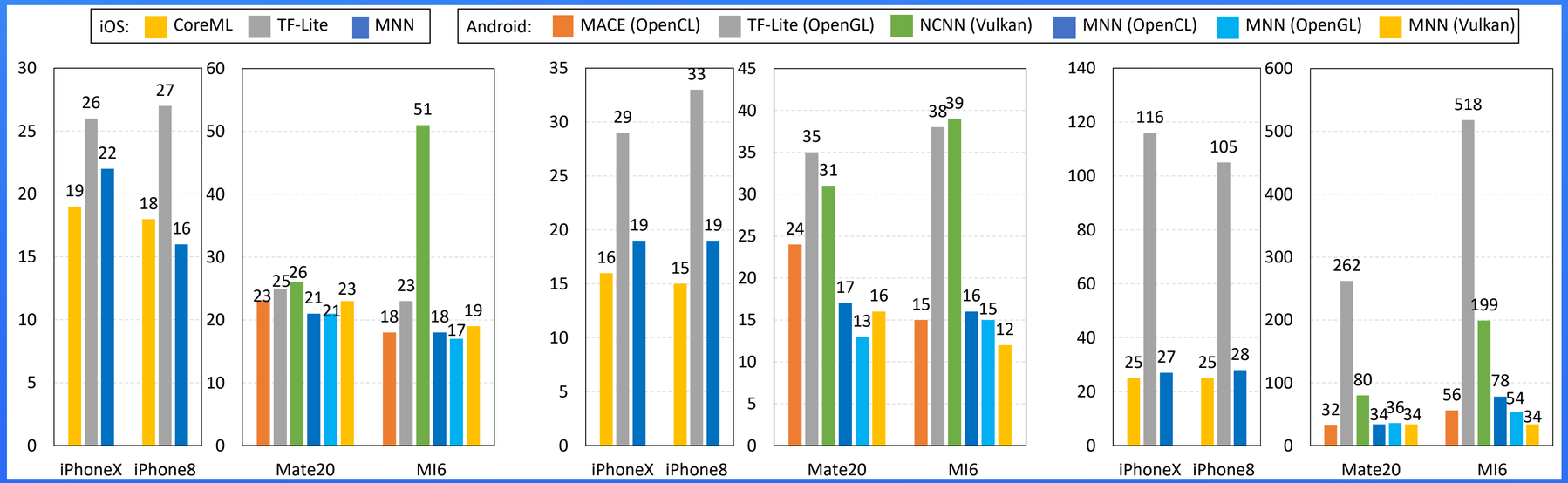


Alibaba Group 阿里巴巴集团

淘系技术部 TAO TECHNOLOGY

MNN Mobile Neural Network

标准性能评测 GPU



Alibaba Group 阿里巴巴集团

淘系技术部 TAO TECHNOLOGY

MNN Mobile Neural Network

MNN - SysML 2020 Accepted

https://mlsys.org/Conferences/2020/AcceptedPapersInitial

MNN: A Universal and Efficient Inference Engine

Xiaotang Jiang (Alibaba) · Huan Wang (Northeastern University) Yiliu Chen (Alibaba) · Ziqi Wu (Alibaba) · Lichuan Wang (Alibaba) · Bin Zou (alibaba) · Yafeng Yang (Alibaba) · zongyang cui (alibaba) · yu cai (Alibaba) · Tianhang Yu (Alibaba) · chengfei lv (alibaba) · Zhihua Wu (Alibaba)

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

MNN Mobile Neural Network

MNN
未来演进方向

深度学习编译优化

MNN 构图 + 训练

MNN.js

Python 工具包

低门槛机器学习平台

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

MNN Mobile Neural Network

MNN学院

Thanks

「淘系技术」微信公众号

推理引擎架构师

- 深入了解tensorflow/pytorch的核心，对其有突出的贡献，或主导过类似框架的核心架构设计
- 深入了解tvm的核心，对其有突出的贡献，对模型编译/IR设计/端上高性能代码生成有深入洞察
- 对机器学习理论有深入研究，特别模型架构搜索/模型结构优化/模型压缩有较高的理论功底
- 对卷积/矩阵乘法/图像/视频编解码等高性能计算有深入的理解和洞察
- 深度学习领域3年以上工作经验，有一定的知名度和影响力

shuhui.sh@alibaba-inc.com

无线AR专家

- 从事移动端AR相关工作2年以上，对AR算法及工程有深刻理解；
- 熟悉计算机视觉相关领域算法，以及图形学、图像渲染OpenGL等领域知识；
- 具有扎实的工程实现能力，熟练使用C/C++及一门脚本语言；
- 有好奇心，创新思维，敢承担，可实战，团队合作，沟通能力佳；
- 有AR试穿试戴、AR互动玩法及其他商业化AR产品研发经验优先；

zhaomi.zm@alibaba-inc.com

Android 客户端

- 熟悉Android Framework层，并有一定Android源码阅读经验
- 了解响应式编程思想，了解 flutter, weex ,kotlin 等技术优先考虑
- 热爱技术，较强逻辑思维能力， 快速学习能力；
- 做事情严谨踏实，团队协作能力强，良好的沟通能力和团队合作精神

difei.zdf@alibaba-inc.com

淘系技术部 - 基础平台部
广纳贤才

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

MNN Mobile Neural Network