主流移动端深度学习框架对比

1.Tensorflow Lite：2017 年 5 月由 Google 开源发布，利用 TensorFlow 转换器将 TensorFlow-trained 模型转换为 TensorFlow Lite 格式；支持 CPU NEON 优化、GPU 加速，支持 C++和 Java，包括 InceptionV3、smart Reply 和 MobileNets 预训练模型；目前只支持 tensorflow 模型，仅支持 CNN 算子；使用较广泛，资料较多。

2.Mobile-Deep-Learning（MDL）：2017 年 9 月由百度开源发布，支持 NEON 优化，包括 MobileNet、Googlenet V1、Squeezenet 及 YOLO 等预训练模型；支持 C++；仅支持 iOS GPU 加速，在 GPU 上速度较快，仅支持 caffe 模型。

3.NCNN：2017 年由腾讯优图开源发布，支持 C++，支持 NEON 优化，支持 caffe 模型，体积小，速度快；包括 faster rcnn、MobileNetssd、peleenetssd、squeezenet 及 yolo 等预训练模型；不支持 GPU；

4.MACE：2018 年由小米开源发布，以 OpenCL 和汇编作为底层算子，提供了异构加速可以方便在不同的硬件上运行模型，同时支持各种框架的模型转换；依赖项较多，过程较复杂。

5.MNN：2019 年由阿里开源的移动端框架，不依赖第三方计算库，使用汇编实现核心运算，支持 Tensorflow、Caffe、ONNX 等主流模型文件格式，支持 CNN、RNN、GAN 等常用网络；开源晚，使用者少。

6.ARM NN：2018 年由 ARM 公司开源发布，桥接了现有神经网络框架（例如 TensorFlow 或 Caffe）与在嵌入式 Linux 平台上运行的底层处理硬件（例如 CPU、GPU 或新型 Arm 机器学习处理器）；开发人员能够继续使用他们首选的框架和工具，经 Arm NN 无缝转换结果后可在底层平台上运行；速度不及 ncnn 和 tensorflow lite。

7.达芬奇架构：达芬奇架构依然是基于 ARM 架构，在 ARM 架构基础之上研发的 NPU（相当于建立了一个独立的 AI 硬件处理单元），达芬奇架构把计算用的乘加器按照不同的计算组织成不同的方式，采用魔方式 MAC 阵列，直接将计算用的 MAC 按不同计算以不同方式进行组合，支持卷积神经网络推理，然后搭配标准的数据缓存。