WEB NEURAL NETWORK API

技术进展及社区组状态更新

Hu Ningxin ningxin.hu@intel.com
Zhang Min belem.zhang@intel.com
英特尔开源技术中心
2018.11.17

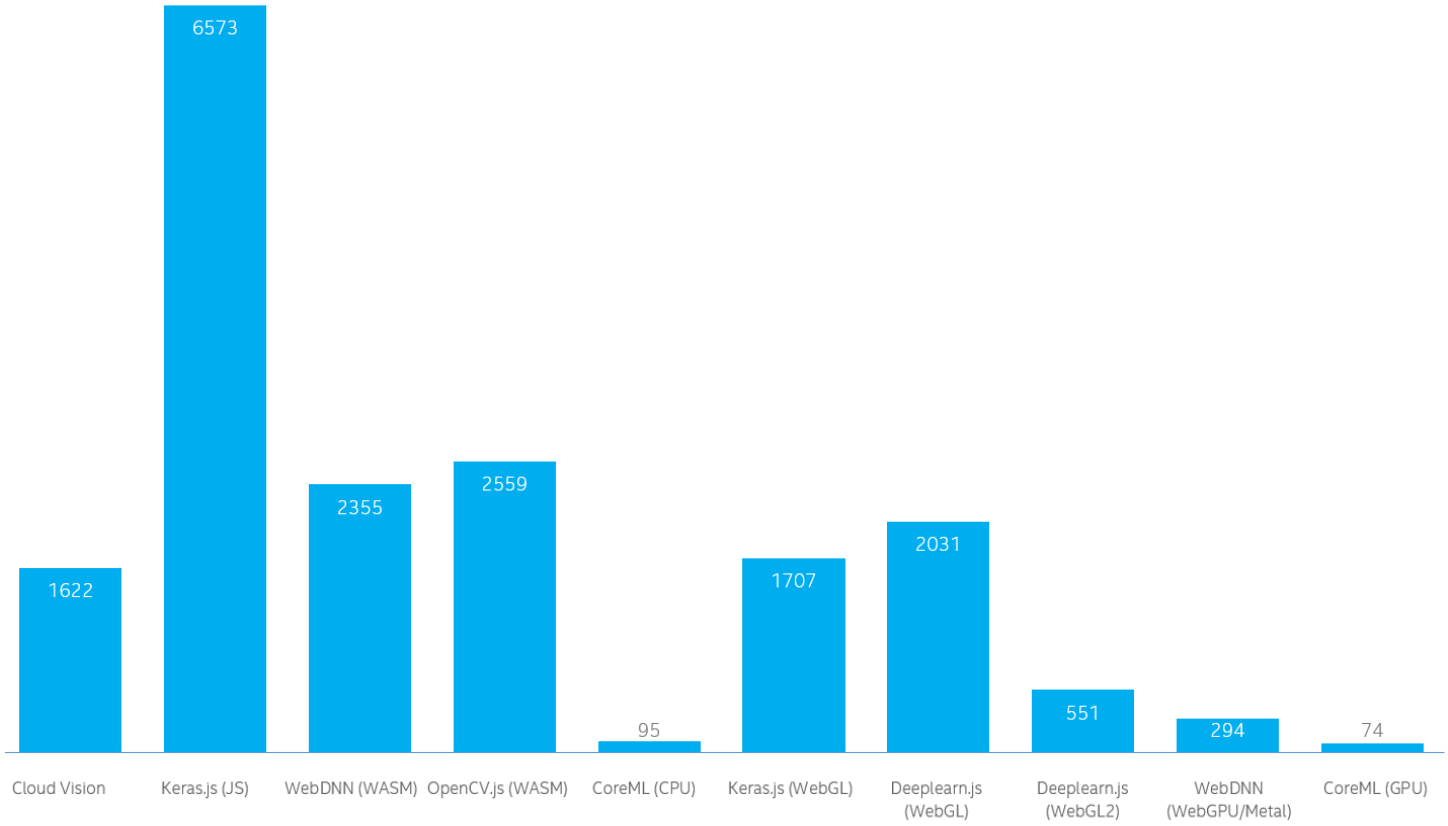
JAVASCRIPT 机器学习/深度学习框架

应用场景	JS 框架
情感分析 目标检测 手写识别 面部识别 图像分类 姿态识别 风格迁移 对象识别 ……	ConvNetJS Deep Learning in your browser MIL WebDNN Keras.js TENSOR FIRE TensorFlow.js

执行时间?

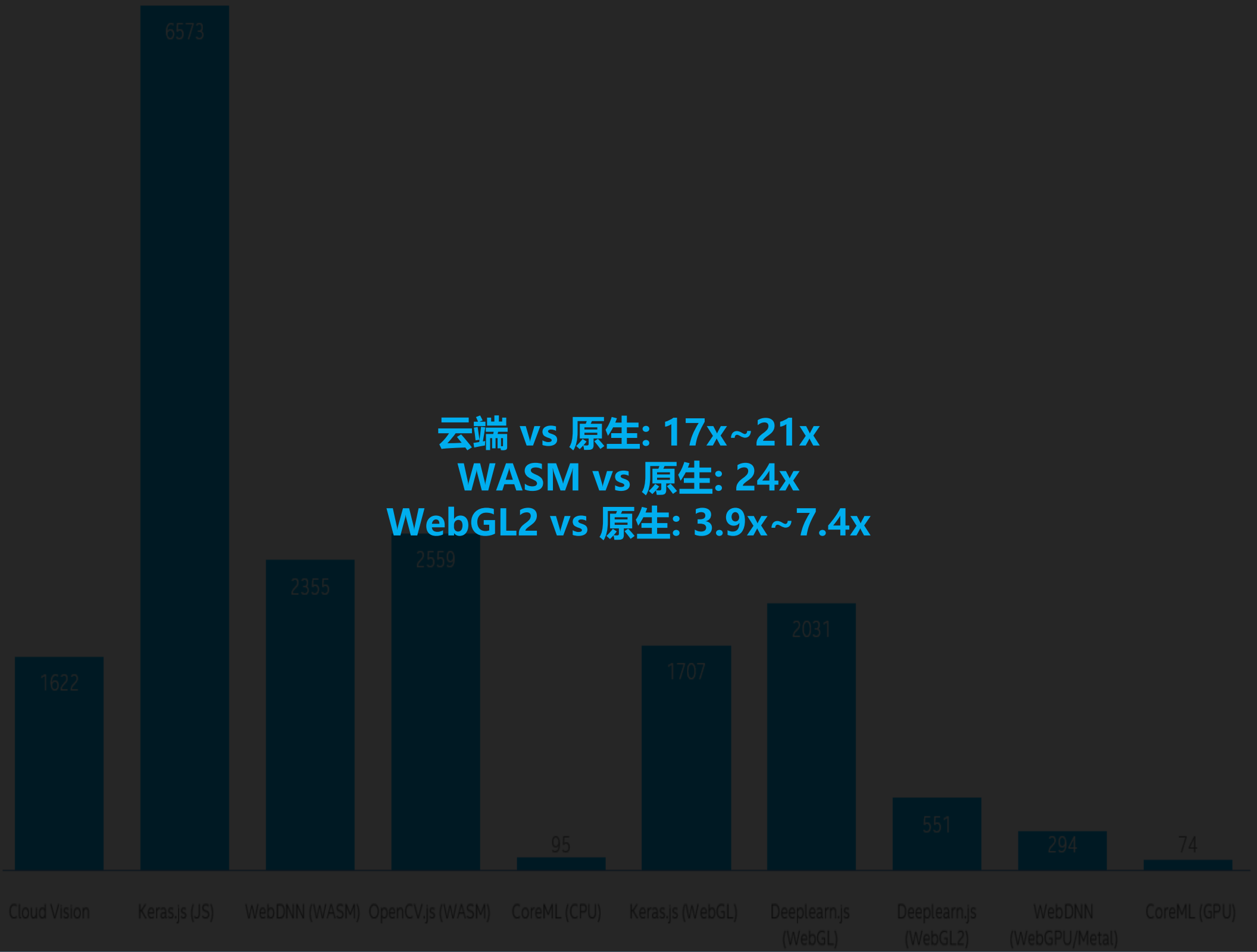
JAVASCRIPT 框架的性能问题

机器学习/深度学习



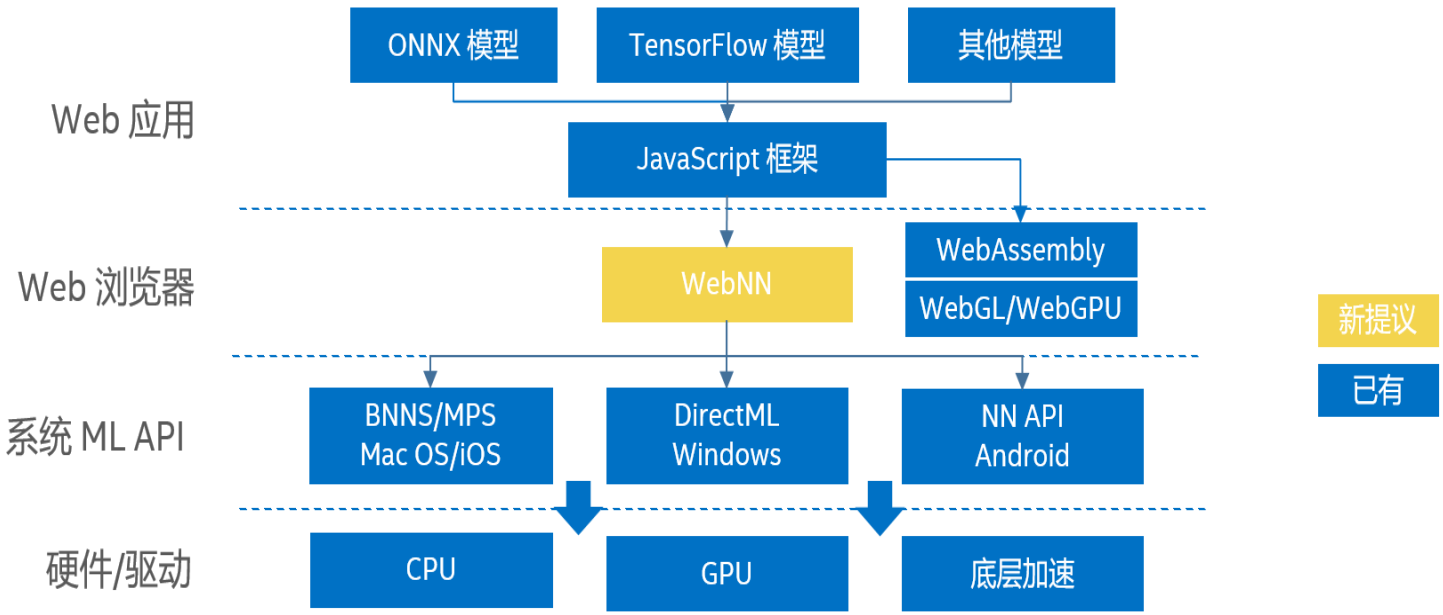
ResNet50 图像分类 运行时间 (ms)

MacBook Pro (13-inch, 2016) / Mac OS 10.13 / SkyLake Core i5 2.9GHz / HD 550
ResNet50, trained by ImageNet, inference, batch size 1, warm up 1, iteration 10 / Jan 2018



云端 vs 原生: 17x~21x
WASM vs 原生: 24x
WebGL2 vs 原生: 3.9x~7.4x

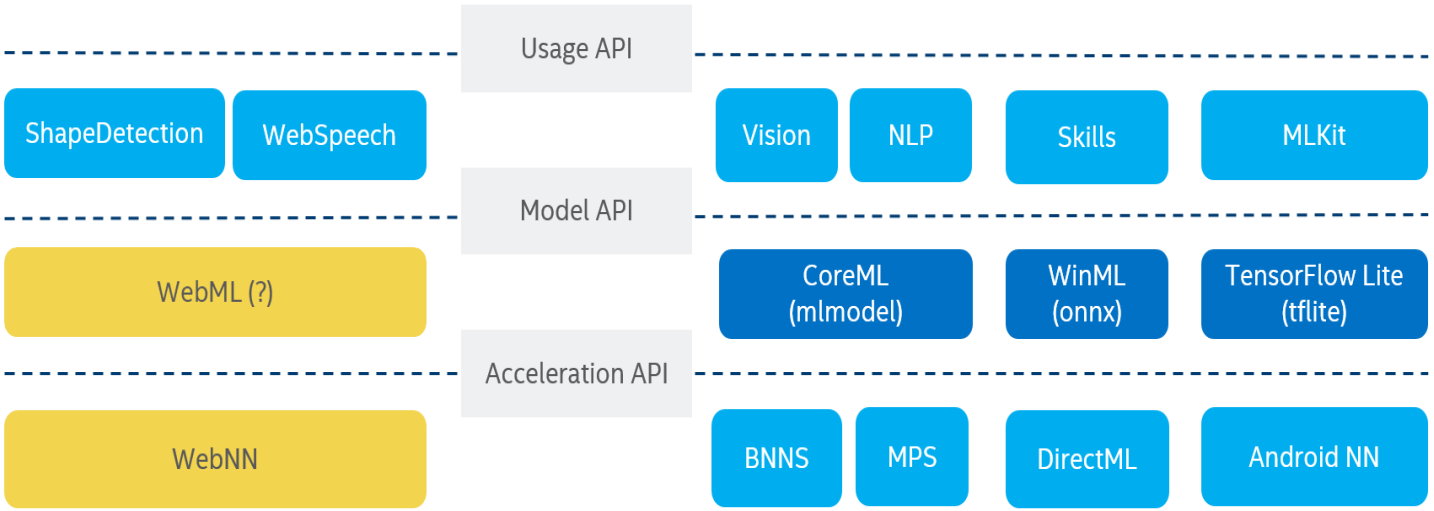
提议: WEB NEURAL NETWORK API



- 用于深度神经网络推理的基于标准的 Web API
- 与文本、多媒体、传感器和 XR 等其他 Web API 集成
- 利用硬件加速，且将 Web 深度学习运算交由系统 API 完成

WEB API 分层架构

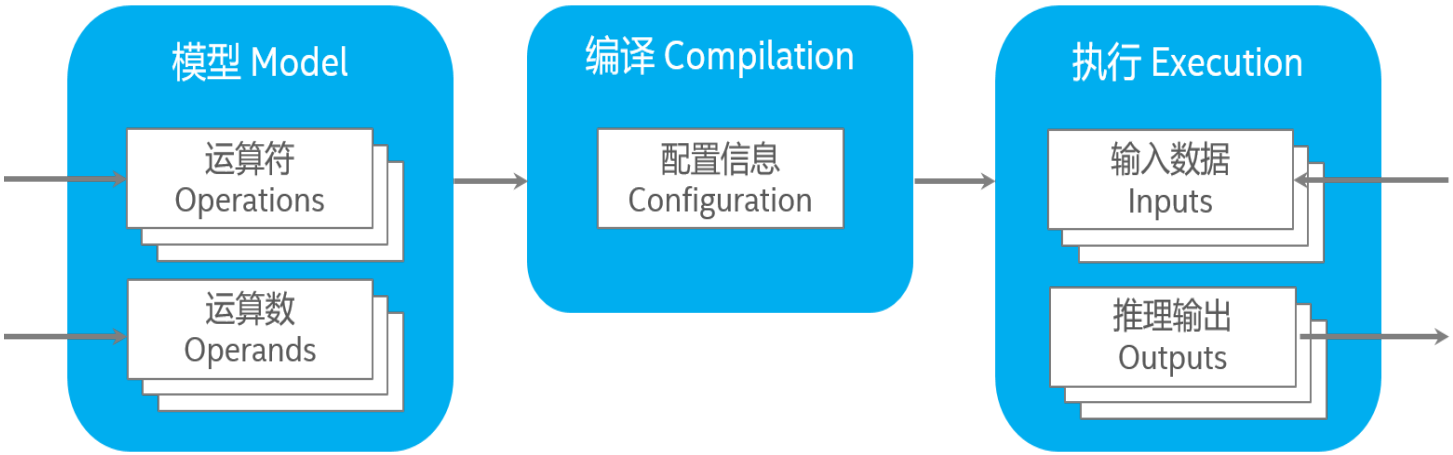
机器学习/深度学习



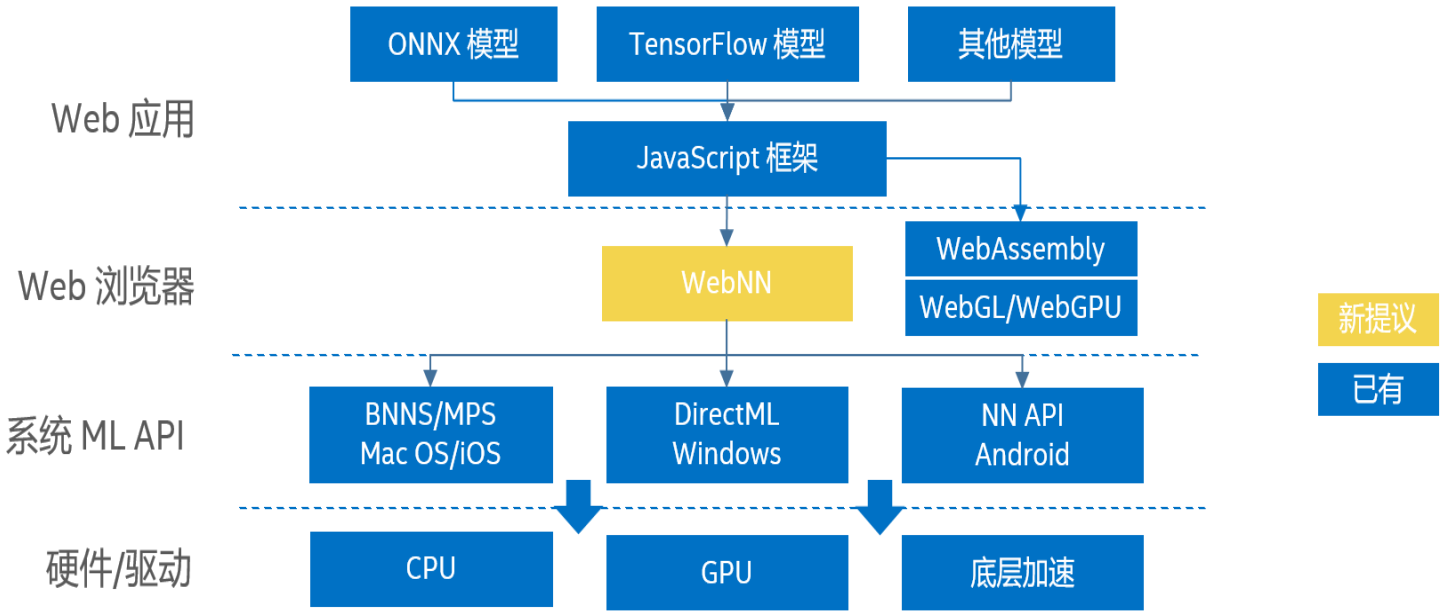
- Usage API: 内置模型，易于集成 ⇒ W3C 形状检测 API
- Model API: 模型预先训练，格式存在碎片化问题 ⇒ 未来的工作
- Acceleration API: NN 底层 API，接近硬件优化，灵活适配 JS 框架 ⇒ 起点

WEB NEURAL NETWORK API POC

- 从 Android NN API 实现 JavaScript API 的概念验证
- 作为 Web NN API 提案的起点，评估性能并探索跨平台能力



WEB NN API POC 的实现



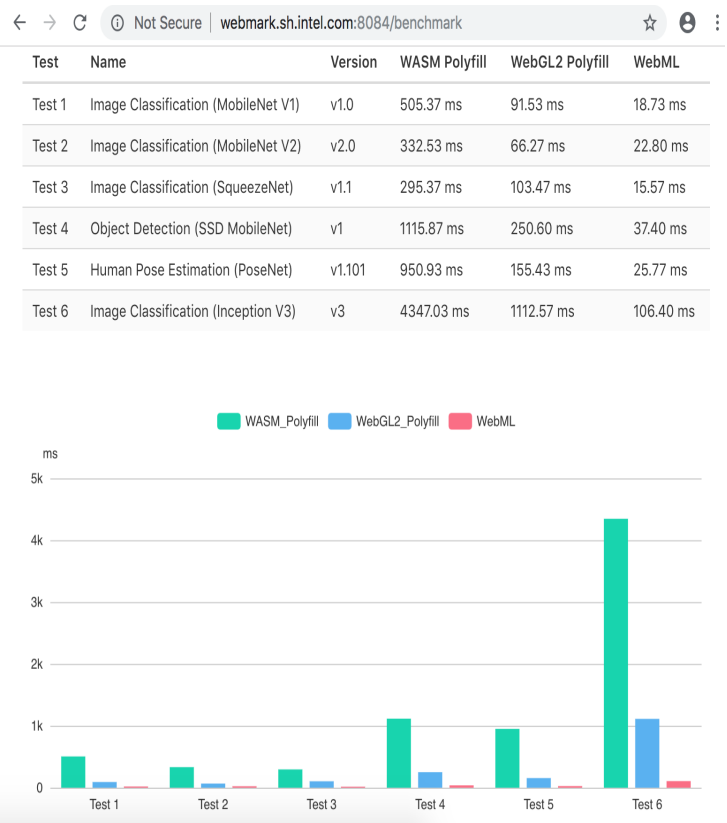
- WebAssembly 以及 WebGL 后端的 Polyfill 实现
- Chromium prototype 实现
  - Mac OS: MPS/BNNS API
  - Android: NN API
  - Windows/Linux: clDNN
  - Windows: DirectML API

WEB NN API POC 功能

- 运算支持
  - ADD, AVERAGE_POOL_2D, CONCATENATION, CONV_2D, DEPTHWISE_CONV_2D, MAX_POOL_2D, MUL, RESHAPE, SOFTMAX, FULLY_CONNECTED
- 模型支持
  - TFLite 模型: MobileNet V1/V2, SqueezeNet, Inception V3, SSD MobileNet
  - TF.js 模型: MobileNet, PoseNet
  - ONNX 模型: MobileNet V2, SqueezeNet
- 原生 API 映射
  - MPS/BNNS, NNAPI and clDNN

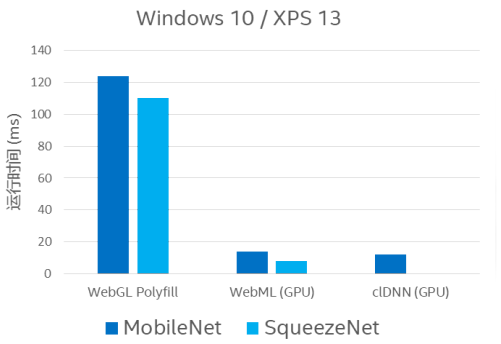
WEB NN API POC 示例及基准测试

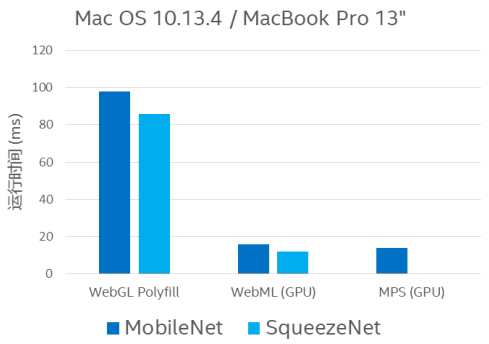
- 示例
  - 图像分类: MobileNet, SqueezeNet, Inception V3
  - 目标检测: SSD MobileNet
  - 姿态识别: PoseNet
  - 静态图像和摄像头支持
- 测试
  - 转换 NN API CTS C++ 测试用例到 JavaScript
  - 500+ 测试用例
- 基准测试
  - 覆盖全部支持的模型
  - Polyfill (WebGL + WASM) 性能测试
  - Web NN API 性能测试

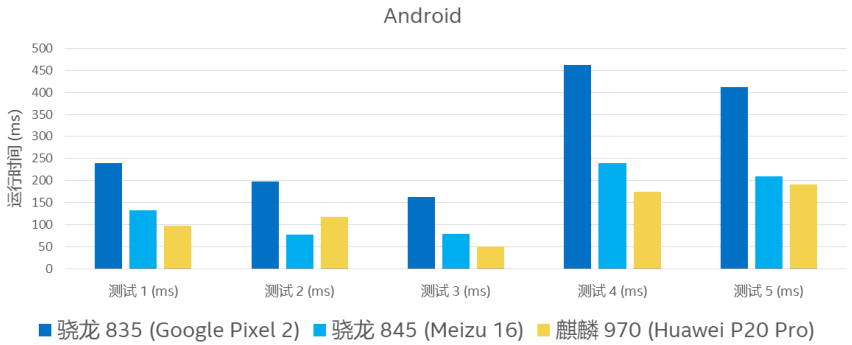


Test	Name	Version	WASM Polyfill	WebGL2 Polyfill	WebML
Test 1	Image Classification (MobileNet V1)	v1.0	505.37 ms	91.53 ms	18.73 ms
Test 2	Image Classification (MobileNet V2)	v2.0	332.53 ms	66.27 ms	22.80 ms
Test 3	Image Classification (SqueezeNet)	v1.1	295.37 ms	103.47 ms	15.57 ms
Test 4	Object Detection (SSD MobileNet)	v1	1115.87 ms	250.60 ms	37.40 ms
Test 5	Human Pose Estimation (PoseNet)	v1.101	950.93 ms	155.43 ms	25.77 ms
Test 6	Image Classification (Inception V3)	v3	4347.03 ms	1112.57 ms	106.40 ms

WEB NN API POC 性能数据 (WIN/MAC/ANDROID)





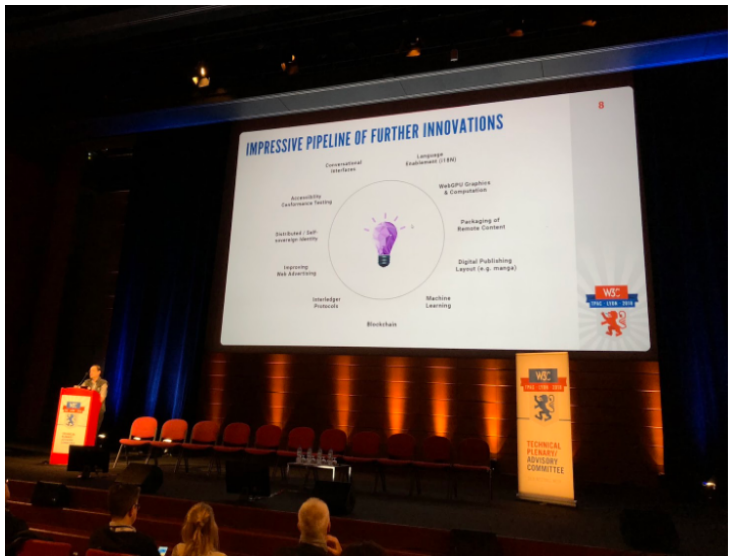


测试 1: 图像分类 (MobileNet V1)
测试 2: 图像分类 (MobileNet V2)
测试 3: 图像分类 (SqueezeNet)
测试 4: 对象检测 (SSD MobileNet)
测试 5: 姿态识别 (PoseNet)

- 和 WebGL polyfill 相比有显著性能提升
- 为 Web 应用带来接近原生的性能
- 为 Web 应用带来原生硬件和软件的优化

WEB NEURAL NETWORK API 合作与支持

- WebML 项目得到谷歌、微软等的广泛支持
  - Google TensorFlowLite/TensorFlow.js 团队, Chrome 团队
  - Microsoft WinML 及 Edge 团队
  - Mozilla 团队
  - Apple MPS 团队
- W3C TPAC 2018
  - W3C CEO Jeff Jaffe 在 "令人印象深刻的进一步创新" 背景下的主题演讲中突出了 Intel Web NN API POC 的工作
  - 2018: 英特尔主导的 WebML 是最受欢迎的分组会议之一，吸引了包括所有浏览器供应商及 W3C TAG 成员的参与



WEB NEURAL NETWORK API 标准规范

- 2018-11-02: 社区小组报告草案 (Draft Community Group Report)
- 在 W3C Web ML 社区组 (WebML CG) 起草
- 用于神经网络推理硬件加速的专用 API
- CG 召开了第一次会议，成员们同意将重点放在使用用例来定义规范工作
    - 高级用例: 构建在预训练的深度神经网络模型之上, 例如人物检测、骨架检测及随机图像生成等
    - API 级用例: ML 框架将引用 WebML API，以便应用开发人员通过框架使用这些功能，例如构建自定义层、性能加速等
- 2018 年晚些时候推进规范开发

W3C MACHINE LEARNING FOR THE WEB 社区组

- 2018-10-03: W3C Web ML 社区组 (WebML CG) 成立
- CG 主席: Anssi Kostiainen (Intel)
- 2018-10-11: WebML CG 章程
- 章程范围: 通过在浏览器中孵化和开发用于机器学习推理的专用低级 Web API，使机器学习成为 Web 的一等公民
- 当前参与者: 英特尔, 华为, 微软, Mozilla 基金会, KDDI 等等
- 微软承诺在即将举行的 We Are Developers AI 大会 (2018 年 12 月 4 日至 5 日, 维也纳)上进一步推广新创建的 WebML CG
- 小组邀请浏览器引擎开发人员，硬件供应商，Web 应用程序开发人员以及对机器学习感兴趣的更广泛的 Web 社区参与

欢迎关注

WebML Polyfill 项目

https://github.com/intel/webml-polyfill

- https://github.com/intel/webml-polyfill
- https://webmachinelearning.github.io
- https://www.w3.org/community/webmachinelearning/

谢谢!