



web

neural network

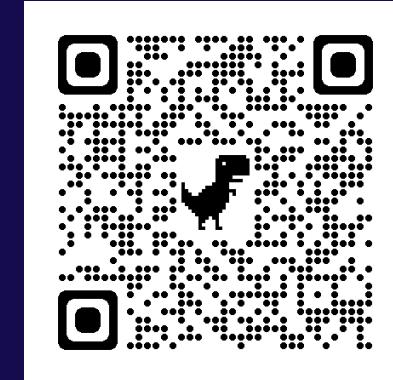
**QCon**  
全球软件开发大会

# WEBNN, WEB 端侧推理的未来

胡宁馨 [ningxin.hu@intel.com](mailto:ningxin.hu@intel.com)

张敏 [belem.zhang@intel.com](mailto:belem.zhang@intel.com)

英特尔 SATG Web 平台工程  
2023 年 12 月



# WebML 客户端推理的优势

## 隐私



摄像头、麦克风等传感器数据保留在设备中

## 离线



初始资源缓存并离线后，不再依赖网络

## 延迟



无云端网络问题，浏览器实时推理

## 成本



无需云端算力支持

## 0安装



浏览器中运行，无需额外安装，并易于共享

## 跨平台



在几乎所有平台上运行 AI 应用

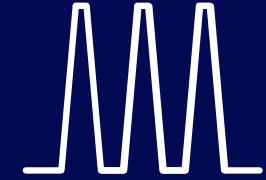
# WebML 客户端推理



突发的  
延迟敏感

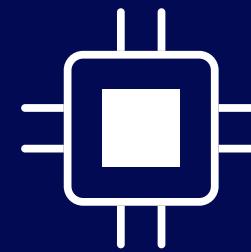


持续的  
电量敏感

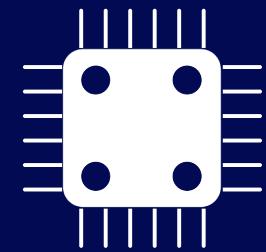


周期的  
吞吐量敏感

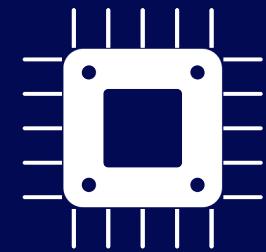
多样的客户端 AI 场景, 多种满足需求的计算单元



CPU  
无处不在  
低延迟, 单一推理任务



高并行性, 高 batch size  
与 3D/渲染/媒体管道集成



专用低功耗AI加速器  
高能耗比, 提升电源效率

# Web 开发者的需求

“ The web needs its own neural networks specification to leverage Apple Silicon, Tensor Cores, and others.

“ Delighted to find the working drafts of WebNN. Incredible new power unlocked for the free, open and competitive Web!

“ Native Tensor support! Would be amazing to have Tensor objects and ops built into Chrome, and available as an “ML API”.

“ Although some scientific computing libraries exist for JS/TS, having built-in support would be far more desirable!

“ If go through the code of utils, maths, audio, tensor in JS, it is annoying that I had to implement these ops myself in JS.

“ llama2-7b in the browser – using WebNN – is going to be 🔥🔥 on-device, local ML 🤖 cc @xenovacom

# WebNN 简介

新兴的 W3C Web 标准 API

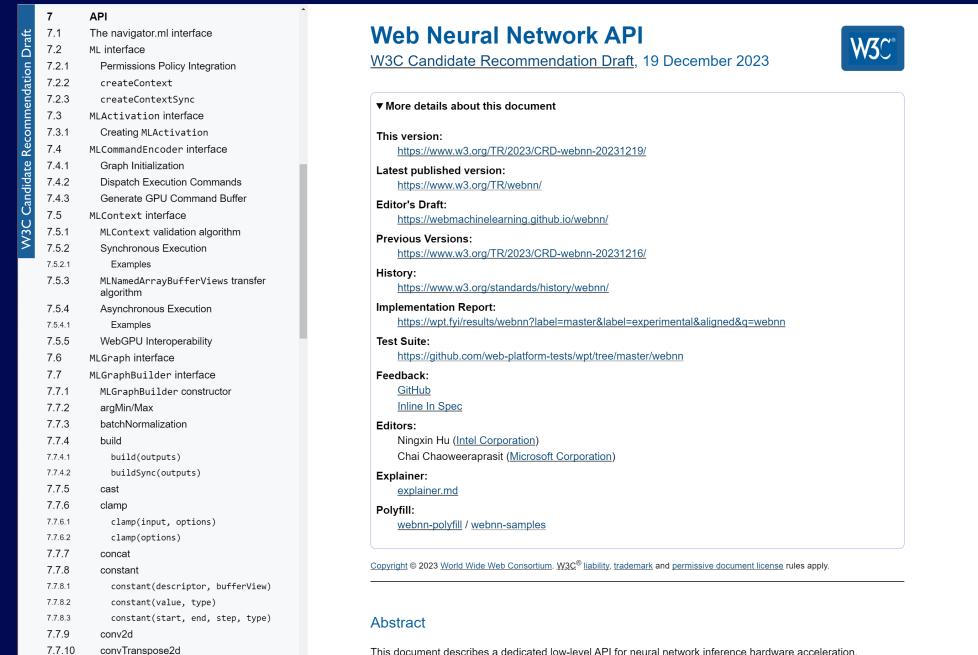
神经网络的统一抽象

通过原生 ML API 访问 AI 硬件加速器

接近原生的 AI 推理性能和结果的可靠性

目前在 *Chrome* 和 *Edge Canary* 中可用 (*runtime flag*)

# WebNN 标准规范



The screenshot shows the official W3C Candidate Recommendation Draft page for the Web Neural Network API. The page title is "Web Neural Network API" and it includes the W3C logo. The content area contains a sidebar with a table of contents for section 7, followed by detailed descriptions of various API components like the navigator.ml interface, MLActivation interface, and MLCommandEncoder interface. The sidebar also lists links for "More details about this document" such as "This version", "Latest published version", "Editor's Draft", "Previous Versions", "History", "Implementation Report", "Test Suite", "Feedback", "Editors", "Explainer", and "Polyfill". The main content area includes sections for "Abstract" and "Copyright".

- WebNN 标准规范由 W3C Web Machine Learning 工作组负责起草
- WebNN 标准规范由 Intel 及 Microsoft 联合编辑

# WebNN 标准规范进展

## 已交付

- 2023年3月: W3C CR
- 60个 CNN/RNN ops,  
float16/32, int32/uint32,  
int8/uint8
- 图像分类: SqueezeNet,  
MobileNet, ResNet
- 物体检测: TinyYOLO
- 噪声抑制: RNNNoise, NSNet

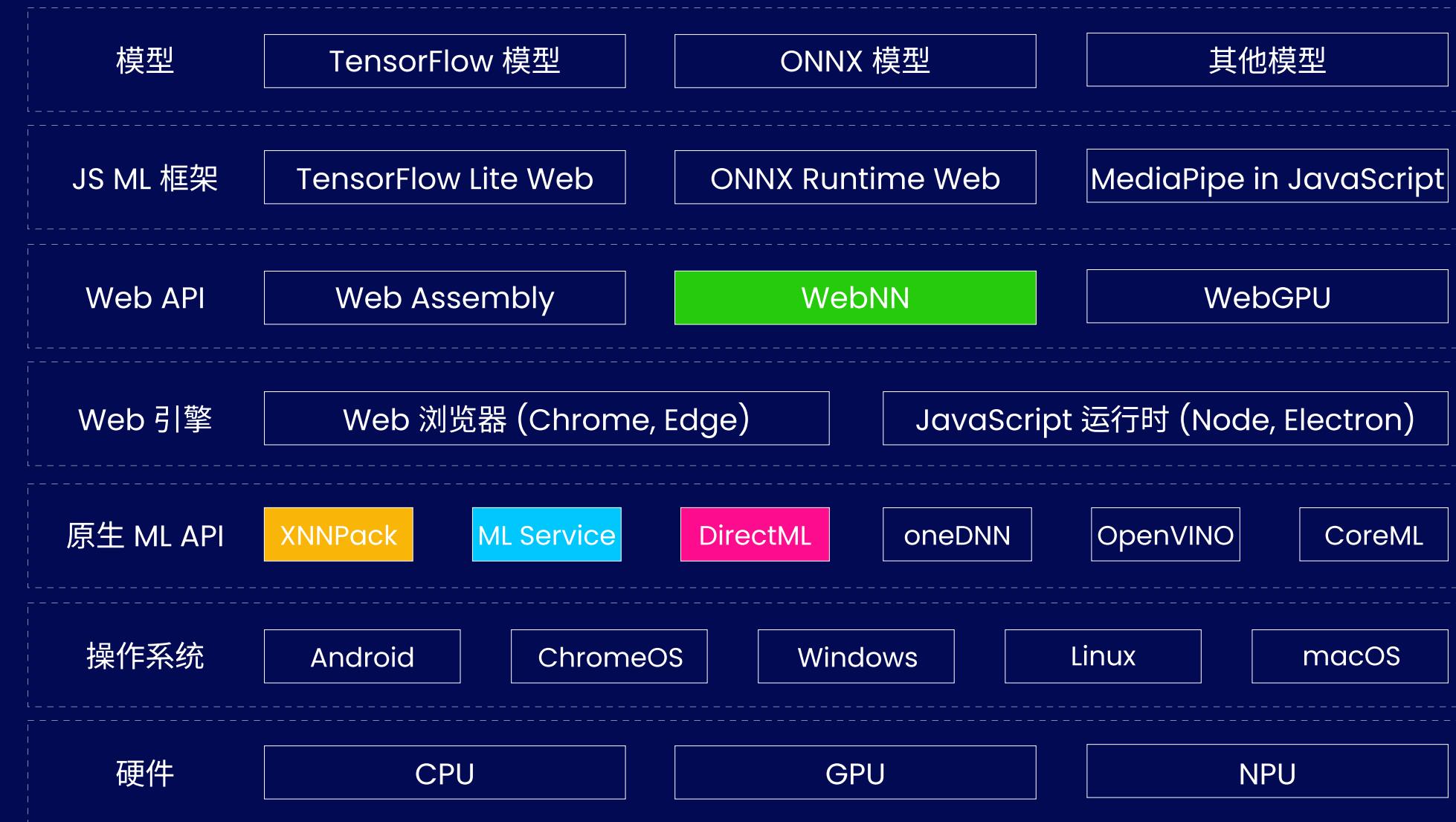
## 最新进展

- 2023年12月: W3C CR更新
- 新增19个 Transformer ops,  
int64/uint64
- 24年: 支持NPU和量化

## 目标模型

- 文本到图像: Stable Diffusion
- 图像分割: Segment Everything
- 语音转文本: Whisper Tiny
- 文本到文本(encoder-decoder):  
T5 及 M2M100
- 文本生成(decoder): LLaMA2

# WebNN 架构



# WebNN 编程模型



计算图图例



# WebNN 代码示例

```
const context = await navigator.ml.createContext({deviceType: 'gpu'});

// The following code builds a graph as:
// constant1 ---+
//                 +--- Add ---> intermediateOutput1 ---+
// input1      ---+                                |
//                               +--- Mul---> output
// constant2 ---+                                |
//                 +--- Add ---> intermediateOutput2 ---+
// input2      ---+                                |

// Use tensors in 4 dimensions.
const TENSOR_DIMS = [1, 2, 2, 2];
const TENSOR_SIZE = 8;

const builder = new MLGraphBuilder(context);

// Create MLOperandDescriptor object.
const desc = {dataType: 'float32', dimensions: TENSOR_DIMS};

// constant1 is a constant MLOperand with the value 0.5.
const constantBuffer1 = new Float32Array(TENSOR_SIZE).fill(0.5);
const constant1 = builder.constant(desc, constantBuffer1);

// input1 is one of the input MLOperands. Its value will be set before execution
const input1 = builder.input('input1', desc);

// constant2 is another constant MLOperand with the value 0.5.
const constantBuffer2 = new Float32Array(TENSOR_SIZE).fill(0.5);
const constant2 = builder.constant(desc, constantBuffer2);
```

```
// input2 is another input MLOperand. Its value will be set before execution.
const input2 = builder.input('input2', desc);

// intermediateOutput1 is the output of the first Add operation.
const intermediateOutput1 = builder.add(constant1, input1);

// intermediateOutput2 is the output of the second Add operation.
const intermediateOutput2 = builder.add(constant2, input2);

// output is the output MLOperand of the Mul operation.
const output = builder.mul(intermediateOutput1, intermediateOutput2);

// Compile the constructed graph.
const graph = await builder.build({'output': output});

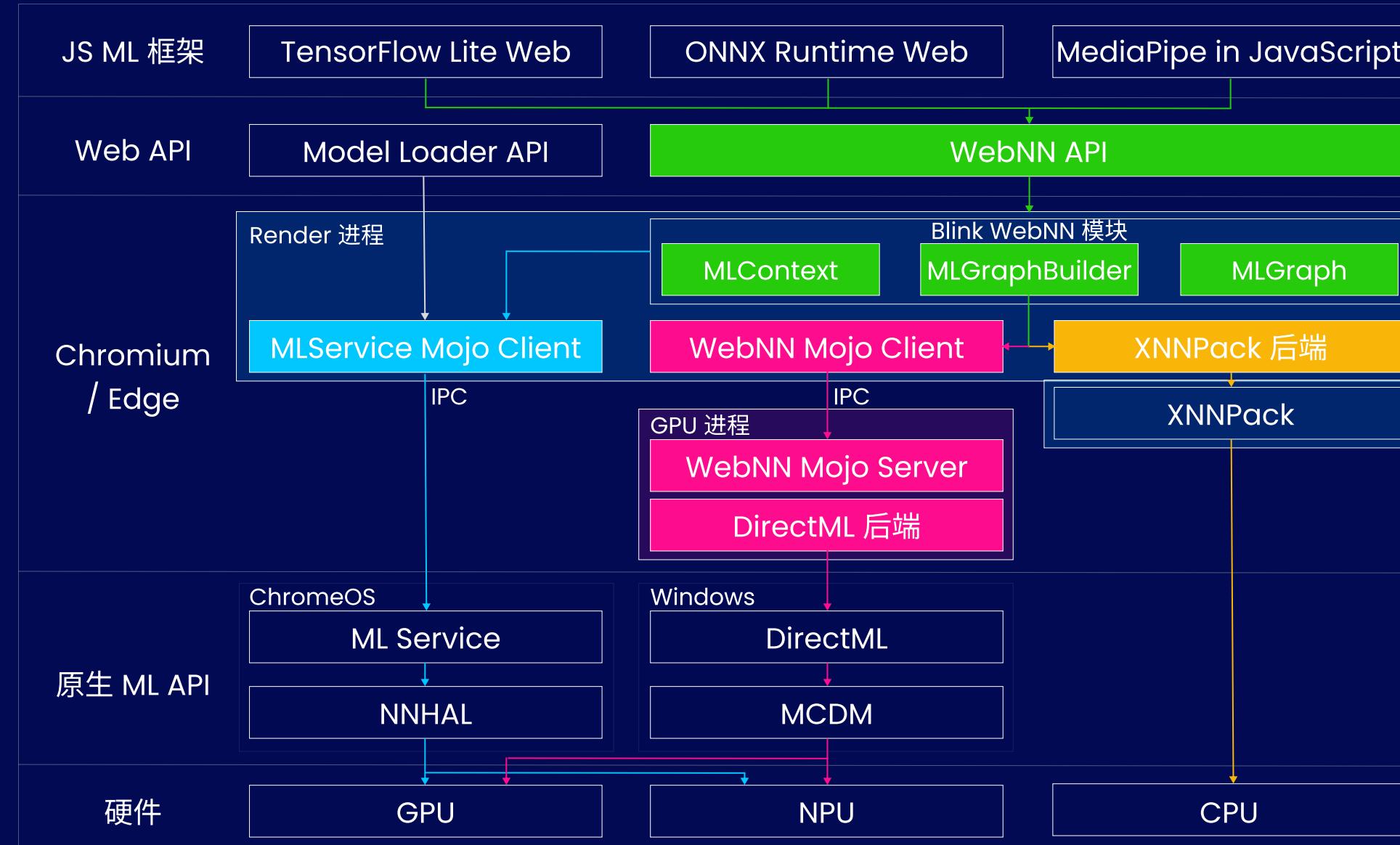
// Setup the input buffers with value 1.
const inputBuffer1 = new Float32Array(TENSOR_SIZE).fill(1);
const inputBuffer2 = new Float32Array(TENSOR_SIZE).fill(1);
const outputBuffer = new Float32Array(TENSOR_SIZE);

// Execute the compiled graph with the specified inputs.
const inputs = {
  'input1': inputBuffer1,
  'input2': inputBuffer2,
};

const outputs = {'output': outputBuffer};
const results = await context.compute(graph, inputs, outputs);

console.log('Output value: ' + results.outputs.output);
// Output value: 2.25,2.25,2.25,2.25,2.25,2.25,2.25,2.25
```

# WebNN 在 Chromium 中的实现



# WebNN 操作符的实现状态(部分)

W3C WebNN Spec	Web Platform Tests	XNNPack/CPU backend	DirectML/GPU backend	Execution Delegate	Execution Provider
clamp		<input checked="" type="checkbox"/> clamp <input checked="" type="checkbox"/> Relu6	<input checked="" type="checkbox"/> ELEMENT_WISE_CLIP	<input checked="" type="checkbox"/> ReluNITo1	<input checked="" type="checkbox"/> Clip
concat		<input checked="" type="checkbox"/> concatenate2 <input checked="" type="checkbox"/> concatenate3 <input checked="" type="checkbox"/> concatenate4	<input checked="" type="checkbox"/> JOIN	<input checked="" type="checkbox"/> Concatenation	<input checked="" type="checkbox"/> Concat
conv2d		<input checked="" type="checkbox"/> convolution_2d	<input checked="" type="checkbox"/> CONVOLUTION	<input checked="" type="checkbox"/> Conv2d <input checked="" type="checkbox"/> DepthwiseConv2d	<input checked="" type="checkbox"/> Conv
convTranspose2d		<input checked="" type="checkbox"/> deconvolution_2d	<input checked="" type="checkbox"/> CONVOLUTION	<input checked="" type="checkbox"/> TransposeConv <input checked="" type="checkbox"/> Convolution2DTransposeBias	<input checked="" type="checkbox"/> ConvTranspose
add <sup>element-wise</sup> binary		<input checked="" type="checkbox"/> add2	<input checked="" type="checkbox"/> ELEMENT_WISE_ADD	<input checked="" type="checkbox"/> Add	<input checked="" type="checkbox"/> Add
sub <sup>element-wise</sup> binary		<input checked="" type="checkbox"/> subtract	<input checked="" type="checkbox"/> ELEMENT_WISE_SUBTRACT	<input checked="" type="checkbox"/> Sub	<input checked="" type="checkbox"/> Sub
mul <sup>element-wise</sup> binary		<input checked="" type="checkbox"/> multiply2	<input checked="" type="checkbox"/> ELEMENT_WISE_MULTIPLY	<input checked="" type="checkbox"/> Mul	<input checked="" type="checkbox"/> Mul
div <sup>element-wise</sup> binary		<input checked="" type="checkbox"/> divide	<input checked="" type="checkbox"/> ELEMENT_WISE_DIVIDE	<input checked="" type="checkbox"/> Div	<input checked="" type="checkbox"/> Div
max <sup>element-wise</sup> binary		<input checked="" type="checkbox"/> maximum2	<input checked="" type="checkbox"/> ELEMENT_WISE_MAX	<input checked="" type="checkbox"/> Maximum	<input checked="" type="checkbox"/> Max

# WebNN 操作符的实现状态(部分)

W3C WebNN Spec	Web Platform Tests	XNNPack/CPU backend	DirectML/GPU backend	Execution Delegate	Execution Provider
					TensorFlow Lite for TensorFlow.js
abs <sup>element-wise</sup> unary		<input checked="" type="checkbox"/> abs	<input checked="" type="checkbox"/> ELEMENT_WISE_ABS	<input checked="" type="checkbox"/> Abs	<input checked="" type="checkbox"/> Abs
ceil <sup>element-wise</sup> unary		<input checked="" type="checkbox"/> ceiling	<input checked="" type="checkbox"/> ELEMENT_WISE_CEIL	<input checked="" type="checkbox"/> Ceil	<input checked="" type="checkbox"/> Ceil
floor <sup>element-wise</sup> unary		<input checked="" type="checkbox"/> floor	<input checked="" type="checkbox"/> ELEMENT_WISE_FLOOR	<input checked="" type="checkbox"/> Floor	<input checked="" type="checkbox"/> Floor
neg <sup>element-wise</sup> unary		<input checked="" type="checkbox"/> negate	<input checked="" type="checkbox"/> ELEMENT_WISE_NEGATE	<input checked="" type="checkbox"/> Neg	<input checked="" type="checkbox"/> Neg
elu		<input checked="" type="checkbox"/> elu	<input checked="" type="checkbox"/> ACTIVATION_ELU	<input checked="" type="checkbox"/> Elu	<input checked="" type="checkbox"/> Elu
gemm		<input checked="" type="checkbox"/> fully_connected	<input checked="" type="checkbox"/> GEMM	<input checked="" type="checkbox"/> FullyConnected	<input checked="" type="checkbox"/> Gemm
hardSwish		<input checked="" type="checkbox"/> hardswish	 $y = x * \max(0, \min(6, (x + 3))) / 6$	<input checked="" type="checkbox"/> HardSwish	<input checked="" type="checkbox"/> HardSwish
leakyRelu		<input checked="" type="checkbox"/> leaky_relu	<input checked="" type="checkbox"/> ACTIVATION_LEAKY_RELU	<input checked="" type="checkbox"/> LeakyRelu	<input checked="" type="checkbox"/> LeakyRelu
prelu		<input checked="" type="checkbox"/> prelu	 ACTIVATION_PARAMETERIZED_RELU	<input checked="" type="checkbox"/> Prelu	<input checked="" type="checkbox"/> Prelu
relu		<input checked="" type="checkbox"/> clamp	<input checked="" type="checkbox"/> ACTIVATION_RELU	<input checked="" type="checkbox"/> Relu	<input checked="" type="checkbox"/> Relu
sigmoid		<input checked="" type="checkbox"/> sigmoid	<input checked="" type="checkbox"/> ACTIVATION_SIGMOID	<input checked="" type="checkbox"/> Logistic	<input checked="" type="checkbox"/> Sigmoid

# WebNN 操作符的实现状态(部分)

W3C WebNN Spec	Web Platform Tests	XNNPack/CPU backend	DirectML/GPU backend	Execution Delegate	Execution Provider
pad		<input checked="" type="checkbox"/> static_constant_pad	<input checked="" type="checkbox"/> PADDING	<input checked="" type="checkbox"/> Pad	<input checked="" type="checkbox"/> Pad
averagePool2d pooling		<input checked="" type="checkbox"/> average_pooling_2d	<input checked="" type="checkbox"/> AVERAGE_POOLING	<input checked="" type="checkbox"/> AveragePool2d <input checked="" type="checkbox"/> Mean	<input checked="" type="checkbox"/> GlobalAveragePool <input checked="" type="checkbox"/> AveragePool
maxPool2d pooling		<input checked="" type="checkbox"/> max_pooling_2d	<input checked="" type="checkbox"/> MAX_POOLING2	<input checked="" type="checkbox"/> MaxPool2d	<input checked="" type="checkbox"/> GlobalMaxPool <input checked="" type="checkbox"/> MaxPool
resample2d		<input checked="" type="checkbox"/> static_resize_bilinear_2d	<input checked="" type="checkbox"/> RESAMPLE	<input checked="" type="checkbox"/> ResizeBilinear	<input checked="" type="checkbox"/> Resize
reshape		<input checked="" type="checkbox"/> static_reshape <input checked="" type="checkbox"/> even_split2	<input checked="" type="checkbox"/> DML_TENSOR_DESC	<input checked="" type="checkbox"/> Reshape	<input checked="" type="checkbox"/> Reshape
split		<input checked="" type="checkbox"/> even_split3 <input checked="" type="checkbox"/> even_split4 <input checked="" type="checkbox"/> static_slice (uneven split)	<input checked="" type="checkbox"/> SPLIT	<input checked="" type="checkbox"/> Split	<input checked="" type="checkbox"/> Split
slice		<input checked="" type="checkbox"/> static_slice	<input checked="" type="checkbox"/> SLICE	<input checked="" type="checkbox"/> Slice <input checked="" type="checkbox"/> StridedSlice	<input checked="" type="checkbox"/> Slice
softmax		<input checked="" type="checkbox"/> softmax	<input checked="" type="checkbox"/> ACTIVATION_SOFTMAX	<input checked="" type="checkbox"/> Softmax	<input checked="" type="checkbox"/> Softmax
transpose		<input checked="" type="checkbox"/> static_transpose	<input checked="" type="checkbox"/> DML_TENSOR_DESC	<input checked="" type="checkbox"/> Transpose	<input checked="" type="checkbox"/> Transpose

# WebNN 的实现状态 (DirectML)

- 目前已经支持 66 个 ops (GPU)
- Transformer 的 ops 已基本支持
- 正在为 NPU 支持作出适配

# WebNN 和主流 JavaScript ML 框架的集成



# WebNN 与 ONNXRuntime Web 集成的代码示例

```
import { InferenceSession } from "onnxruntime-web";

// ...

// Initialize the ONNX model
const initModel = async () => {
  ort.env.wasm.numThreads = 1; // 4
  ort.env.wasmsimd = true;
  ort.env.wasm.proxy = true;

  const options: InferenceSession.SessionOptions = {
    // provider name: wasm, webnn
    // deviceType: cpu, gpu
    // powerPreference: default, high-performance
    executionProviders:
      [{ name: "wasm" }], // WebAssembly CPU
  };

  // ...
};

const results = await model.run(feeds);
const output = results[model.outputNames[0]];
```

WebAssembly 后端

```
import { InferenceSession } from "onnxruntime-web";

// ...

// Initialize the ONNX model
const initModel = async () => {
  env.wasm.numThreads = 1; // 4
  env.wasmsimd = true;
  env.wasm.proxy = true;

  const options: InferenceSession.SessionOptions = {
    // provider name: wasm, webnn
    // deviceType: cpu, gpu
    // powerPreference: default, high-performance
    executionProviders:
      [{ name: "webnn", deviceType: "gpu", powerPreference: 'default' }],
  };

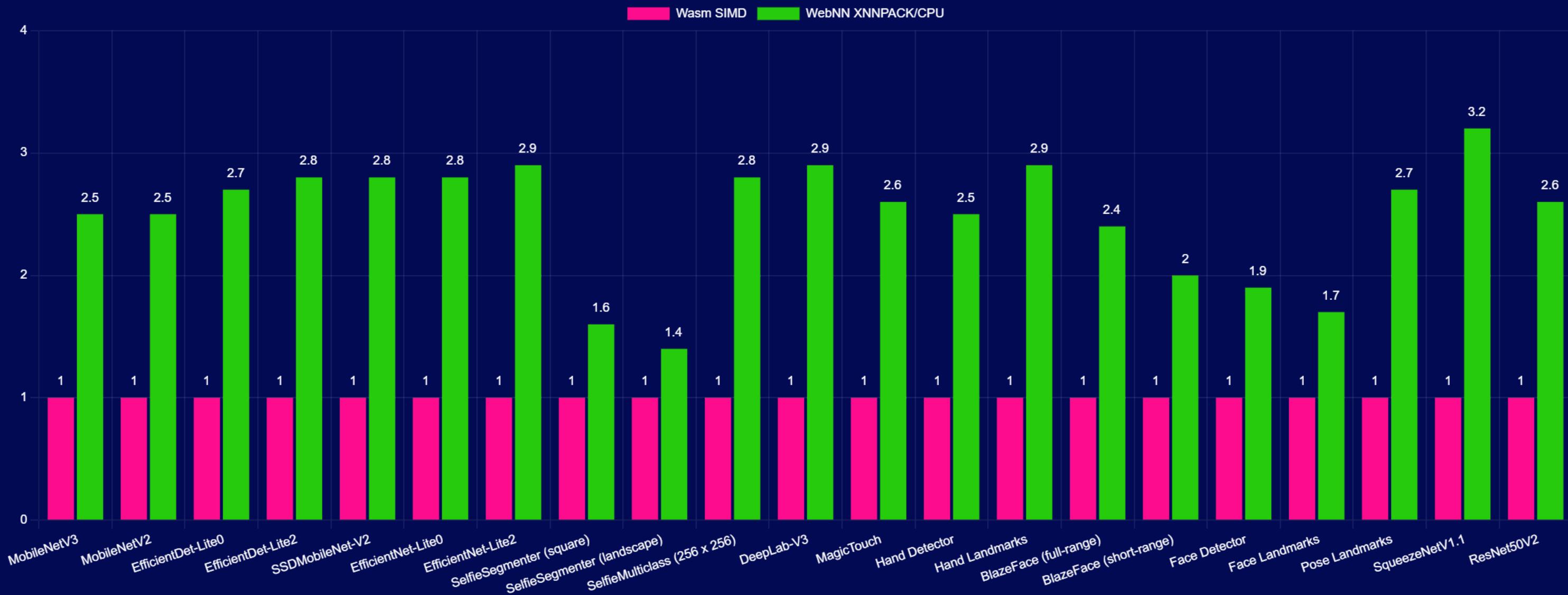
  // ...
};

const results = await model.run(feeds);
const output = results[model.outputNames[0]];
```

WebNN 后端

# WebNN XNNPack/CPU 性能数据 (标准化)

MediaPipe 模型, 越高越好



Backend	CPU	<b>Wasm</b>	Wasm 4	WebNN	WebNN 4	GPU	WebGL	WebGPU	WebNN	NPU	WebNN
Data Type	<b>FP32</b>	INT64	FP16	INT8							
Model Type	ONNX	TensorFlow Lite	NumPy	PyTorch							
	DenseNet 121	EfficientNet Lite 4	MobileNet v2_10	MobileNet v2_12	ResNet50 v1	ResNet50 v2					
Model	SqueezeNet 1.1	DeepLab v3	Selfie Segmentation Gen...	Selfie Segmentation Lan...	Emotion FERPlus	FNS Candy					
	Tiny YOLO v2	Stable Diffusion 2.1 V...	Stable Diffusion 2.1 V...	Segment Anything	T5 Small Decoder	Whisper Tiny Encoder					
100 Runs	100										

### PERFORMANCE (ms)

Model	Size	Type	Data	Wasm 1T	Wasm 4T	WebNN CPU 1T	WebNN CPU 4T	WebGL
MobileNet v2_12 ↗	13.32 MB	fp32	fp32	27.38	18.16	6.64	3.85	29.44
DenseNet 121 ↗	31.2 MB	fp32	fp32	200.32	74.83	65.79	34.33	81.17
EfficientNet Lite 4 ↗	49.54 MB	fp32	fp32	129.50	85.59	29.95	11.50	34.95
MobileNet v2_10 ↗	13.32 MB	fp32	fp32	28.57	19.04	6.74	3.38	27.88
ResNet50 v1 ↗	97.8 MB	fp32	fp32	251.24	93.34	69.77	24.42	78.79
ResNet50 v2 ↗	97.7 MB	fp32	fp32	284.46	122.64	90.76	36.98	96.33
SqueezeNet 1.1 ↗	4.72 MB	fp32	fp32	26.24	13.33	8.90	3.50	13.44
DeepLab v3 ↗	8.07 MB	fp32	fp32	753.34	418.62	197.68	95.49	⚠️
Tiny YOLO v2 ↗	60.54 MB	fp32	fp32	242.54	101.57	108.11	52.94	203.14

45/45 100.00%  

[Run Tests](#)

[Show Logs](#)

ONNX Runtime Web Versions **Wasm** v1.16.1 Public **WebGPU** v1.17 Internal Nov 13 **WebNN** v1.17 Internal Nov 08

⌚ x86-64 16 Logical Cores ⚡ Nominal CPU Pressure 🖼 Intel Graphicsgfx-driver-ci-master-15093 RI Direct3D11 📈 6.02 MB 📈 395 MB 0.13% Storage Used 🔍 100% AC  
 🖥 Windows 10 🌐 Chrome 121.0.0.0 ✅ Cross Origin Isolated

Backend	CPU	Wasm	<b>Wasm 4</b>	WebNN	<b>WebNN 4</b>	GPU	WebGL	WebGPU	<b>WebNN</b>	NPU	WebNN
Data Type	FP32	INT64	FP16	INT8							
Model Type	ONNX	TensorFlow Lite	NumPy	PyTorch							
Model	DenseNet 121	EfficientNet Lite 4	MobileNet v2_10	MobileNet v2_12	MobileNet v2	MobileNet v2_12					
	ResNet50 v1	ResNet50 v2	SqueezeNet 1.1	SqueezeNet 1.0	DeepLab v3	Selfie Segmentation Gen...					
	Selfie Segmentation Lan...	Emotion FERPlus	FNS Candy	Tiny YOLO v2	ALBERT Base v2	ALBERT Base v2					
	BART Large CNN Encoder	BART Large CNN Encoder	BERT Base Cased	BERT Base Cased	BERT Base Uncased	BERT Base Uncased					
	BERT Base Multilingual...	BERT Base Multilingual...	CodeGen Mono 350M	CLIP ViT Base	DETR w/i ResNet-50	DINO ViT					
	DINO ViT	Distilbart CNN 6-6 Enc...	Distilbart CNN 6-6 Dec...	DistilBERT Base Cased ...	DistilBERT Base Uncase...	DistilGPT2 Decoder					
	Distil-Whisper Decoder	GPT-2 Decoder	M2M100 418M Decoder	M2M100 418M Encoder	Stable Diffusion 1.5 U...	Stable Diffusion 1.5 V...					
	Stable Diffusion 2.1 V...	SAM B Decoder	SAM B Encoder	Segment Anything	Segment Anything	T5 Small Decoder					
	T5 Small Decoder	T5 Small Encoder	T5 Small Encoder	Vision Transformer (Vi...	ViT GPT2 Image Caption...	ViT GPT2 Image Caption...					
	Whisper Tiny Decoder	Whisper Tiny Decoder	Whisper Tiny Encoder	Whisper Tiny Encoder							
100 Runs	100										

### PERFORMANCE (MS)

Compilation	First Inference	Average Inference	Median Inference	Best Inference			
Model	Size	Type	Data	Wasm 4T	WebNN CPU 4T	WebGL	WebNN GPU
EfficientNet Lite 4 ↗	49.54 MB	⬢	fp32	74.54	9.52	23.65	5.45
MobileNet v2_10 ↗	13.32 MB	⬢	fp32	15.67	2.52	9.88	2.32
MobileNet v2_12 ↗	13.32 MB	⬢	fp32	15.14	2.49	10.01	2.36
ResNet50 v1 ↗	97.8 MB	⬢	fp32	101.75	20.25	23.79	4.77
SqueezeNet 1.1 ↗	4.72 MB	⬢	fp32	7.54	2.36	6.92	1.74
Emotion FERPlus ↗	33.42 MB	⬢	fp32	17.04	4.33	11.24	2.19

24/24 100.00%  

[Run Tests](#)
[Show Logs](#)

# W3C Machine Learning for the Web

## 社区组

讨论和探索新想法，孵化机器学习推理的新提案

39个组织代表, 126名参与者



## 工作组

基于社区组孵化的提案，标准化机器学习推理的 Web API

17个组织代表, 43名参与者(3名特邀专家)

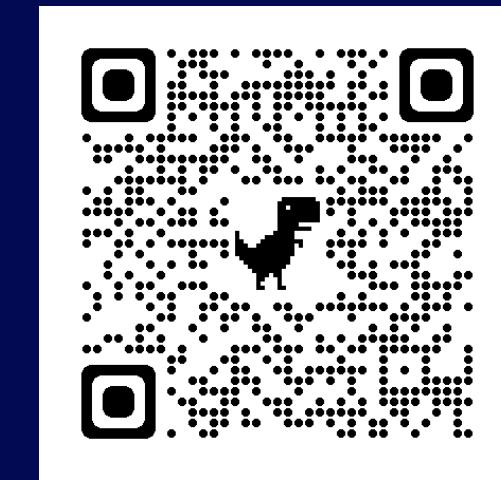


# 谢谢！



web

neural network



<https://webnn.dev>

WebNN 交流群