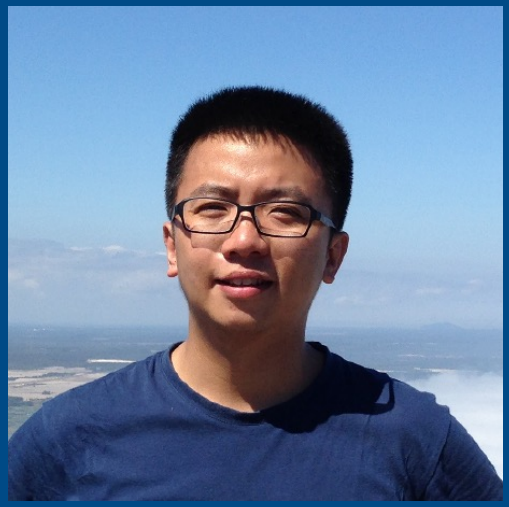
WebNN Overview and Status Update

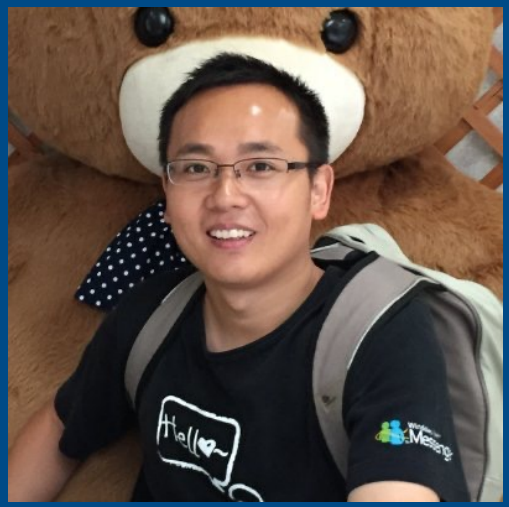
Ningxin Hu

Belem Zhang

May 2024



Ningxin Hu, Intel Principal Engineer, initiator and co-editor of the W3C Web Neural Network (WebNN) specification, Chromium committer and co-owner of the Chromium WebNN component



Belem Zhang, Engineering Manager leading Intel WebNN team for spec, Chromium and ONNX Runtime WebNN EP development, author of WebNN developer preview

Executive Summary

- Three AI hardware engines of AI PC: CPU, GPU and NPU
- WebNN brings a unified abstraction of neural networks to the web
- Accesses AI hardware acceleration through native OS ML API
- Delivers near-native performance and the next gen use cases
- Status:
  - Spec:
    - CNN/RNN - CR published Q2’23
    - Transformers/GenAI – CR refresh published Q2’24
  - Implementation:
    - DirectML GPU on Win: Announcing developer preview
    - DirectML NPU on Win: Coming soon
    - CoreML on MacOS: WIP
    - TFLite on Android/ChromeOS: WIP

Age of the AI PC

Three AI Engines

Heterogenous execution of AI workloads embraces the best practices in AI software design.

WebNN

Announcing Developer Preview

Web Neural Network API

W3C Candidate Recommendation Draft, 23 May 2024

▼ More details about this document

This version:
https://www.w3.org/TR/2024/CRD-webnn-20240523/

Latest published version:
https://www.w3.org/TR/webnn/

Editor's Draft:
https://webmachinelearning.github.io/webnn/

Previous Versions:
https://www.w3.org/TR/2024/CRD-webnn-20240515/

History:
https://www.w3.org/standards/history/webnn/

Implementation Report:
https://wpt.fyi/results/webnn?label=master&label=experimental&aligned&q=webnn

Test Suite:
https://github.com/web-platform-tests/wpt/tree/master/webnn

Feedback:
GitHub
Inline In Spec

Editors:
Ningxin Hu (Intel Corporation)
Dwayne Robinson (Microsoft Corporation)

Former Editor:
Chai Chaoweeraprasit (Microsoft Corporation)

Explainer:
explainer.md

Polyfill:
webnn-polyfill / webnn-samples

web neural network

Standard
W3C API
Unified Abstraction

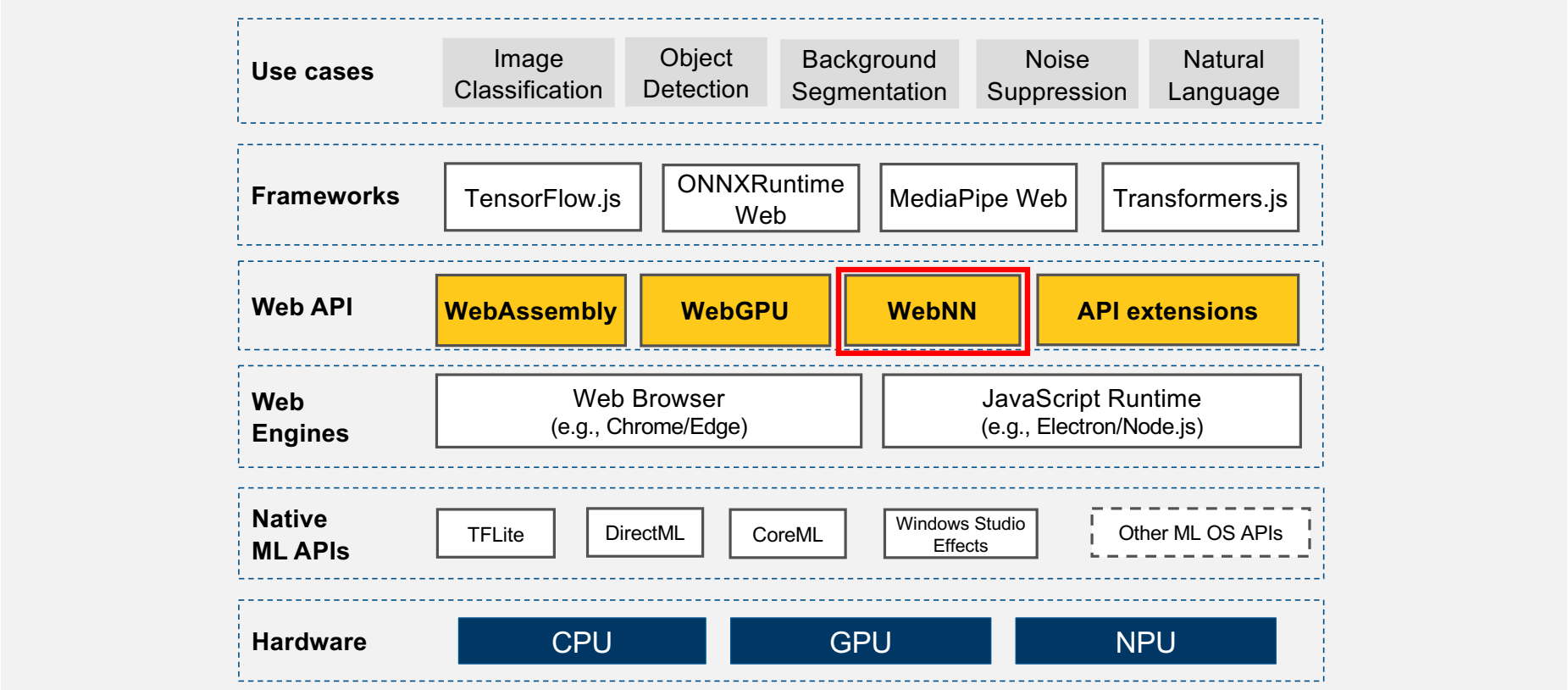
Hetero
HW Exec
CPU, GPU, NPU

Integrated
ML Frameworks
ONNX RT Web, …

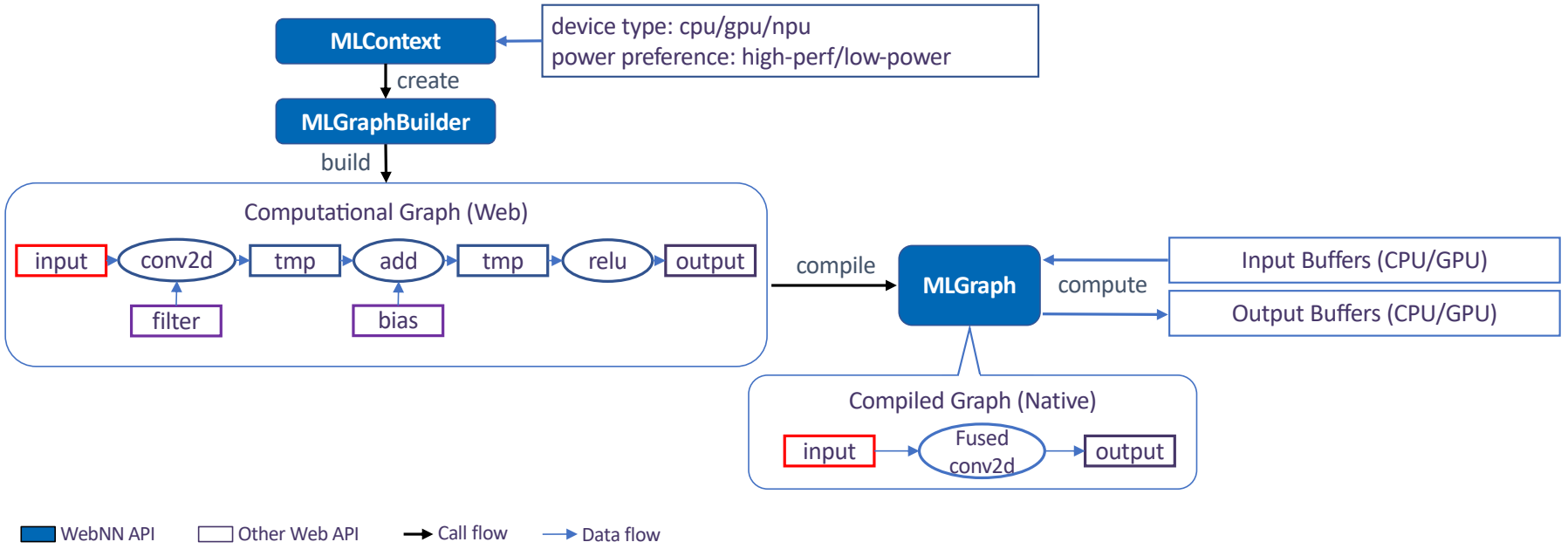
Near Native
Execution Characteristics
Perf. & Power

General
Computational Graph
BYOM

Hardware-Accelerated Web AI Overview

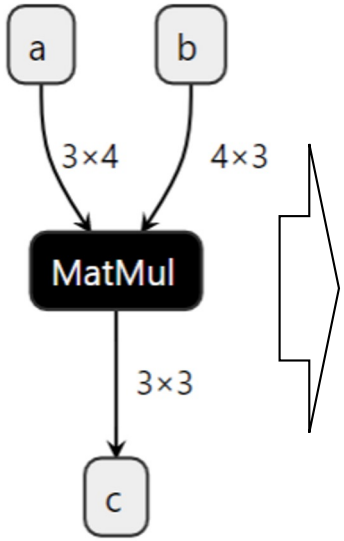


WebNN Programming Model



WebNN brings a unified abstraction of neural networks to Web

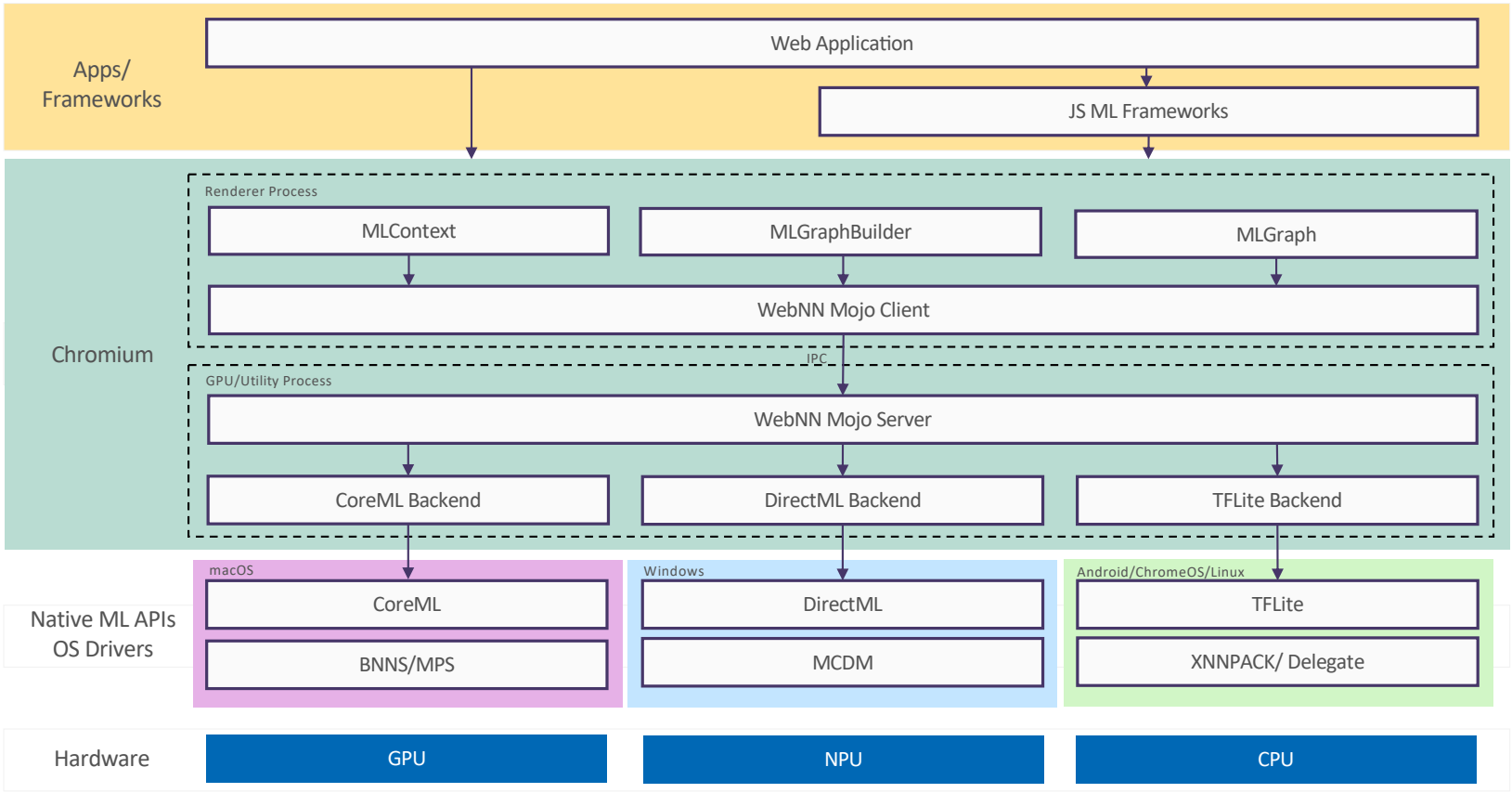
Hello Tensors



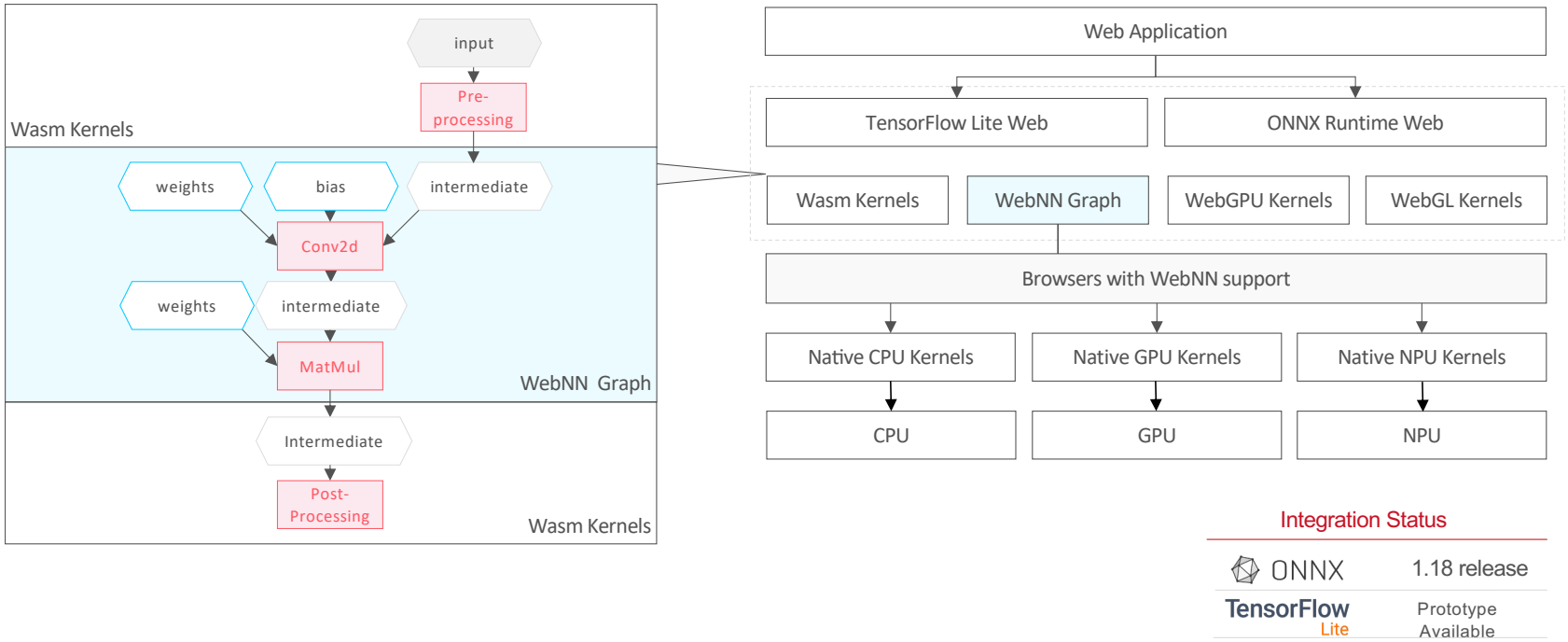

```
// Step 0: Create a context and graph builder for 'gpu', 'cpu' or 'npu'.
const context = await navigator.ml.createContext({deviceType: 'gpu'});
const builder = new MLGraphBuilder(context);
// Step 1: Create a computational graph calculating `c = a * b`.
const a = builder.input('a', {dataType: 'float32', dimensions: [3, 4]});
const b = builder.input('b', {dataType: 'float32', dimensions: [4, 3]});
const c = builder.matmul(a, b);
// Step 2: Compile it into an executable graph.
const graph = await builder.build({c});
// Step 3: Bind input and output buffers to the graph and execute.
const bufferA = new Float32Array(3*4).fill(1.0);
const bufferB = new Float32Array(4*3).fill(0.8);
const bufferC = new Float32Array(3*3);
const results = await context.compute(graph, {'a': bufferA, 'b': bufferB}, {'c': bufferC});
// Step 4: Retrieve the results.
console.log(`values: ${results.outputs.c}`);
```

Browser-native tensor operations targeting CPU, GPU and NPU

WebNN Browser Implementation



WebNN JS ML Frameworks Integration



Integration Status	
ONNX	1.18 release
TensorFlow Lite	Prototype Available

Mainstream ML frameworks are integrating WebNN

ONNX Runtime Web Code Samples with WebNN

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

// Initialize the ONNX model
const initModel = async () => {
  env.wasm.numThreads = 1; // 4
  env.wasm.simd = true;
  env.wasm.proxy = true;
  const options: InferenceSession.SessionOptions = {
    // provider name: wasm, webnn
    // deviceType: cpu, gpu, npu
    // powerPreference: default, high-performance
    executionProviders:
      [{ name: "wasm"}], // WebAssembly CPU
    }
    // ...
};
```

WebAssembly backend

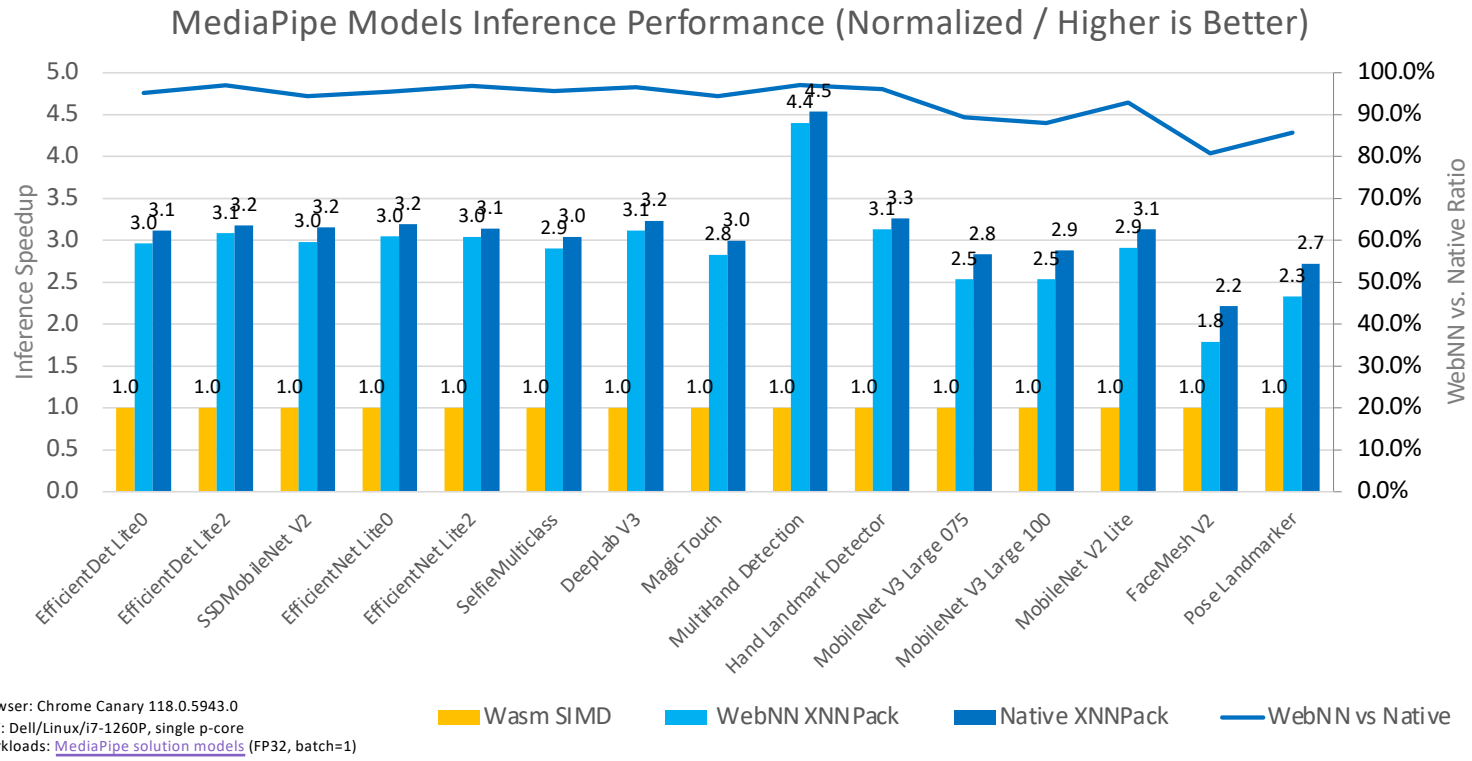
```js
import { InferenceSession } from "onnxruntime-web";

// Initialize the ONNX model
const initModel = async () => {
  env.wasm.numThreads = 1; // 4
  env.wasm.simd = true;
  env.wasm.proxy = true;
  const options: InferenceSession.SessionOptions = {
    // provider name: wasm, webnn
    // deviceType: cpu, gpu, npu
    // powerPreference: default, high-performance
    executionProviders:
      [{ name: "webnn", deviceType: "gpu", powerPreference: 'default' }],
    }
    // ...
};
```

WebNN backend

Switching to WebNN can be done by modifying a single line of code

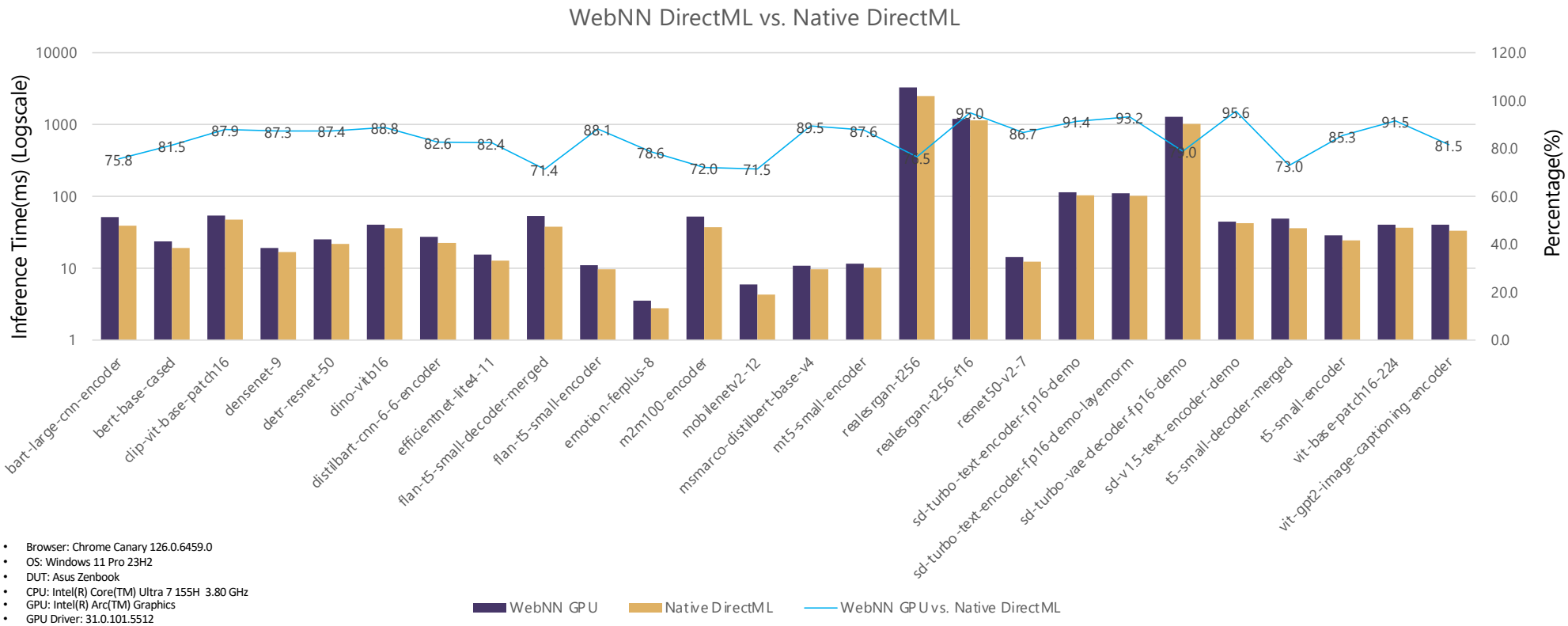
“Near-Native” Performance of WebNN on CPU



- Browser: Chrome Canary 118.0.5943.0
- DUT: Dell/Linux/i7-1260P, single p-core
- Workloads: MediaPipe solution models (FP32, batch=1)

The average performance of listed 15 models on WebNN on CPU is about **93%** of native XNNPack

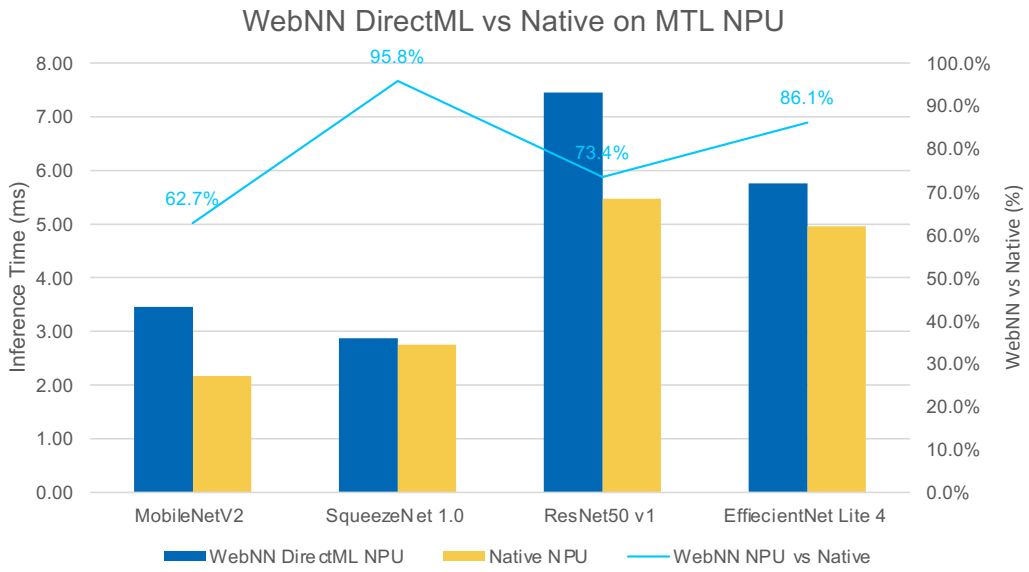
“Near-Native” Performance of WebNN on GPU

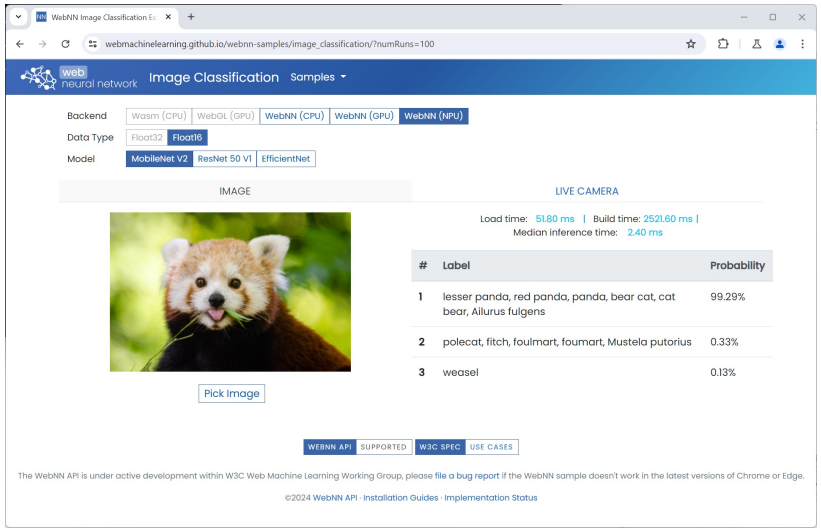


- Browser: Chrome Canary 126.0.6459.0
- OS: Windows 11 Pro 23H2
- DUT: Asus Zenbook
- CPU: Intel(R) Core(TM) Ultra 7 155H 3.80 GHz
- GPU: Intel(R) Arc(TM) Graphics
- GPU Driver: 31.0.101.5512

The average performance of listed 26 models on WebNN DirectML is about **83%** of native DML on MTL iGPU

“Near-Native” Performance of WebNN on NPU





- Browser: Chrome Canary 126.0.6459.0
- OS: Windows 11 Pro 23H2
- DUT: Asus Zenbook
- CPU: Intel(R) Core(TM) Ultra 7 155H 3.80 GHz
- NPU: Intel(R) AI Boost
- NPU Driver: 32.0.100.2381

The average performance of listed 4 models on WebNN DirectML is about **80%** of native DML on MTL NPU

https://microsoft.github.io/webnn-developer-preview/

RUNTIME neural network

WebNN Developer Preview

Run ONNX models in the browser with WebNN. The developer preview unlocks interactive ML on the web that benefits from reduced latency, enhanced privacy and security, and GPU acceleration from DirectML.

text encoder | unet | vae | webnn gpu

Stable Diffusion 1.5

Text-to-Image

text encoder | unet | vae | webnn gpu

Stable Diffusion Turbo

Text-to-Image

sam | webnn gpu

Segment Anything

Image Segmentation

encoder | decoder | webnn gpu | webnn npu

Whisper Base

Automatic Speech Recognition

mobilenet | resnet | webnn gpu | webnn npu

Image Classification

Classification

WebNN Developer Preview

https://microsoft.github.io/webnn-developer-preview/

ONNX RUNTIME web neural network

WebNN Execution Provider of ONNX Runtime Web with GPU acceleration from DirectML.
Running on Intel® Core™ Ultra 7 processor 155H with integrated Arc™ GPU.

WebNN Developer Preview

Run ONNX models in the browser with WebNN. The developer preview unlocks interactive ML on the web that benefits from reduced latency, enhanced privacy and security, and GPU acceleration from DirectML.

text encoder | unet | vae | webnn gpu

Stable Diffusion 1.5
Text-to-Image

text encoder | unet | vae | webnn gpu

Stable Diffusion Turbo
Text-to-Image

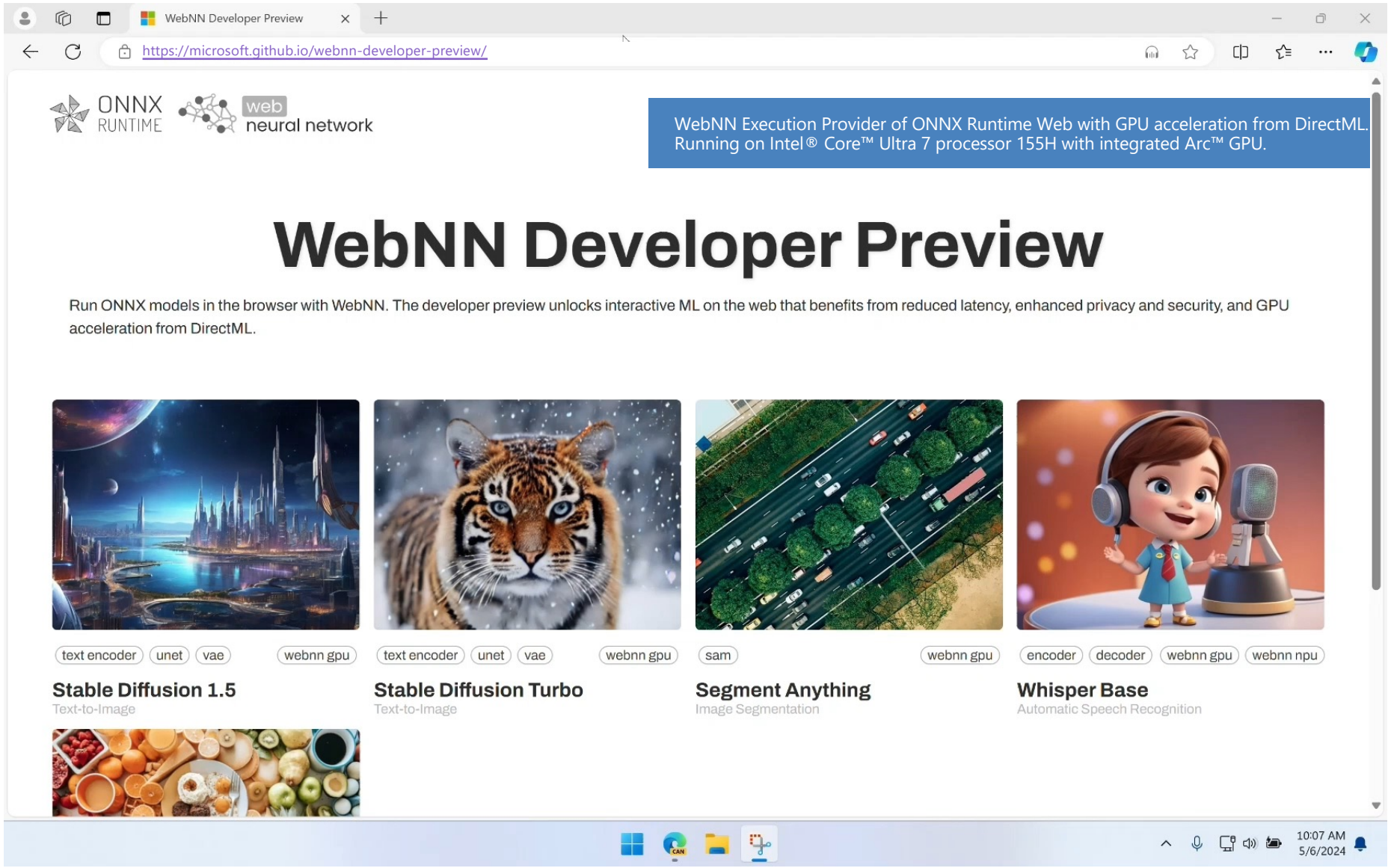
sam | webnn gpu

Segment Anything
Image Segmentation

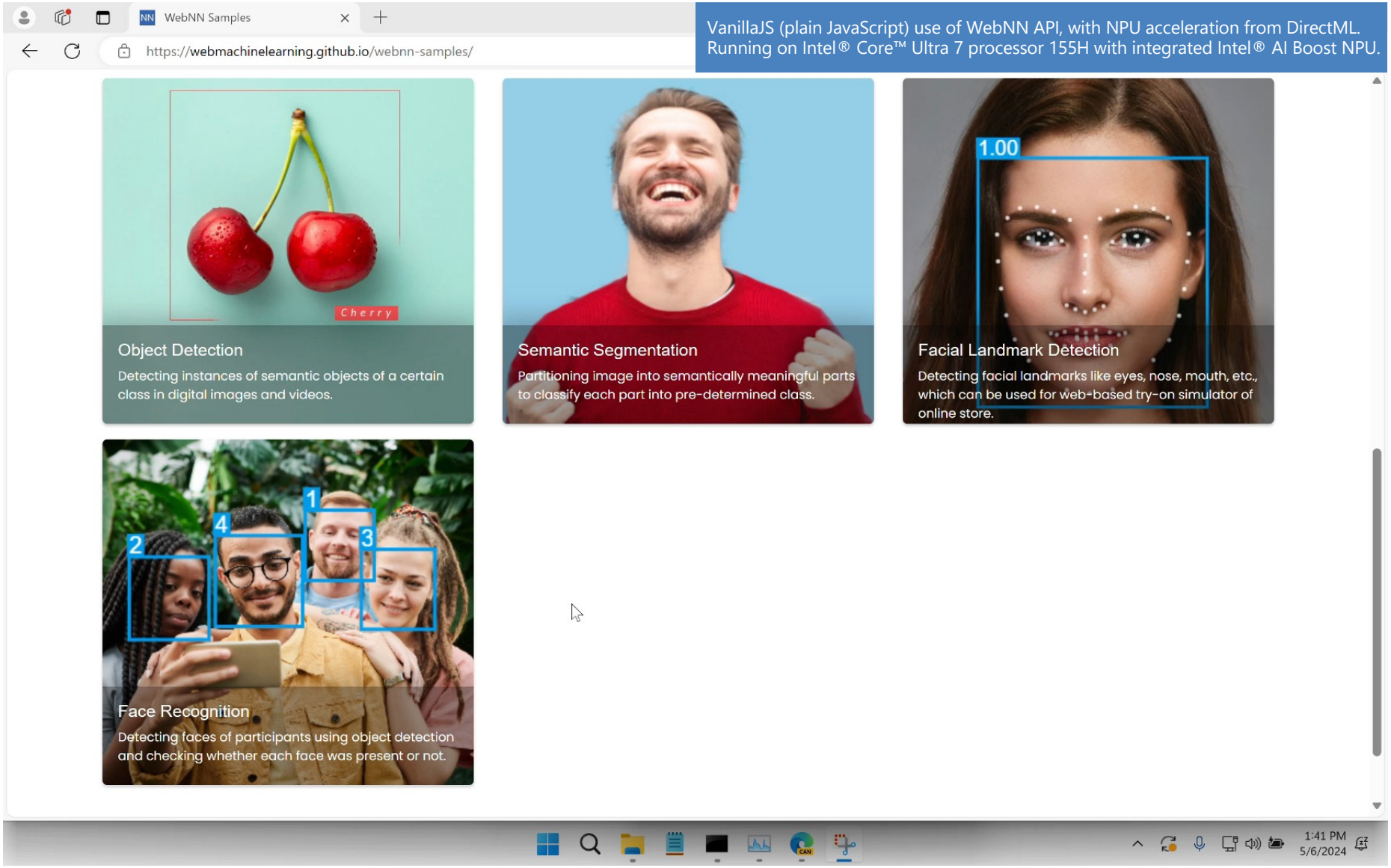
encoder | decoder | webnn gpu | webnn npu

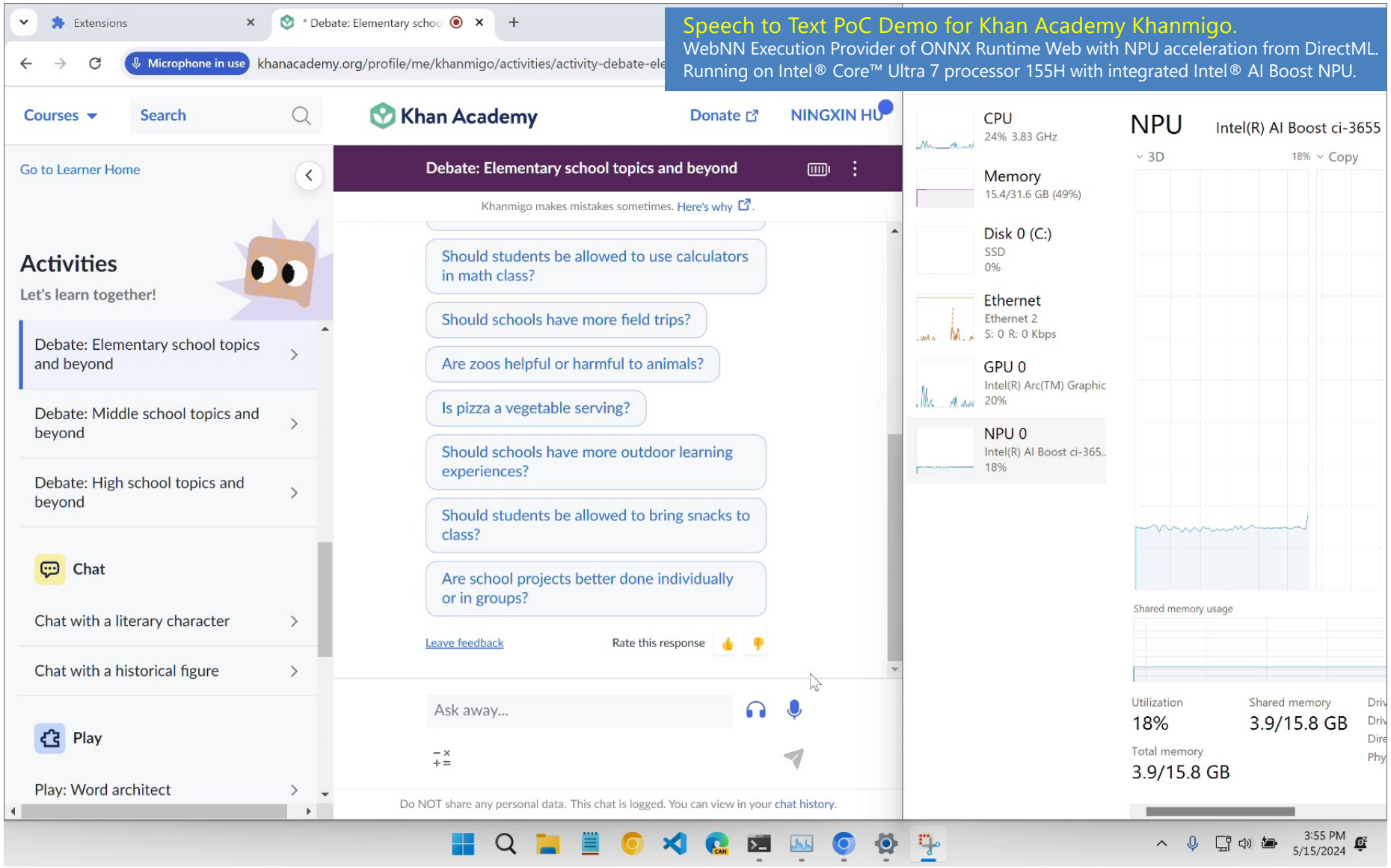
Whisper Base
Automatic Speech Recognition

10:10 AM
5/6/2024


WebNN Developer Preview
https://microsoft.github.io/webnn-developer-preview/
ONNX RUNTIME
web neural network
WebNN Execution Provider of ONNX Runtime Web with GPU acceleration from DirectML.
Running on Intel® Core™ Ultra 7 processor 155H with integrated Arc™ GPU.
WebNN Developer Preview
Run ONNX models in the browser with WebNN. The developer preview unlocks interactive ML on the web that benefits from reduced latency, enhanced privacy and security, and GPU acceleration from DirectML.
text encoder
unet
vae
webnn gpu
Stable Diffusion 1.5
Text-to-Image
text encoder
unet
vae
webnn gpu
Stable Diffusion Turbo
Text-to-Image
sam
webnn gpu
Segment Anything
Image Segmentation
encoder
decoder
webnn gpu
webnn npu
Whisper Base
Automatic Speech Recognition
10:07 AM
5/6/2024

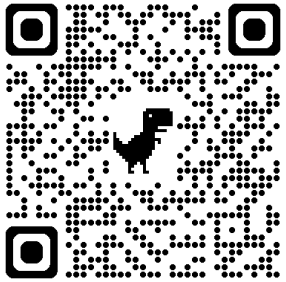
VanillaJS (plain JavaScript) use of WebNN API, with NPU acceleration from DirectML.
Running on Intel® Core™ Ultra 7 processor 155H with integrated Intel® AI Boost NPU.



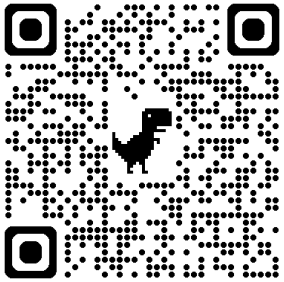

Speech to Text PoC Demo for Khan Academy Khanmigo.
WebNN Execution Provider of ONNX Runtime Web with NPU acceleration from DirectML.
Running on Intel® Core™ Ultra 7 processor 155H with integrated Intel® AI Boost NPU.
Extensions
* Debate: Elementary schoo
Microphone in use
khanacademy.org/profile/me/khanmigo/activities/activity-debate-ele
Courses
Search
Khan Academy
Donate
NINGXIN HU
Go to Learner Home
Debate: Elementary school topics and beyond
Khanmigo makes mistakes sometimes. Here's why.
Activities
Let's learn together!
Debate: Elementary school topics and beyond
Debate: Middle school topics and beyond
Debate: High school topics and beyond
Chat
Chat with a literary character
Chat with a historical figure
Play
Play: Word architect
Should students be allowed to use calculators in math class?
Should schools have more field trips?
Are zoos helpful or harmful to animals?
Is pizza a vegetable serving?
Should schools have more outdoor learning experiences?
Should students be allowed to bring snacks to class?
Are school projects better done individually or in groups?
Leave feedback
Rate this response
Ask away...
Do NOT share any personal data. This chat is logged. You can view in your chat history.
CPU
24% 3.83 GHz
Memory
15.4/31.6 GB (49%)
Disk 0 (C:)
SSD
0%
Ethernet
Ethernet 2
S: 0 R: 0 Kbps
GPU 0
Intel(R) Arc(TM) Graphic
20%
NPU 0
Intel(R) AI Boost ci-365..
18%
NPU
Intel(R) AI Boost ci-3655
3D
18%
Copy
Shared memory usage
Utilization
18%
Shared memory
3.9/15.8 GB
Total memory
3.9/15.8 GB
3:55 PM
5/15/2024

Call to Action

- Try WebNN on Microsoft Edge Dev channel and Google Chrome Dev channel
  - Navigate to about://flags in the URL bar and turn on “Enables WebNN API”



WebNN Developer Preview
https://aka.ms/webnn



WebNN Samples
https://webmachinelearning.github.io/webnn-samples/

- Join Intel, Microsoft, Google, Hugging Face, and other industry leaders and shape WebNN definition and development

References

- **WebNN Spec:** https://www.w3.org/TR/webnn/
- **WebNN Explainer:** https://github.com/webmachinelearning/webnn/blob/main/explainer.md
- **WebNN Implementation Status:** https://webmachinelearning.github.io/webnn-status/
- **Awesome WebNN:** https://github.com/webmachinelearning/awesome-webnn
- **WebNN Dev Preview:** https://microsoft.github.io/webnn-developer-preview/
- **WebNN Samples:** https://webmachinelearning.github.io/webnn-samples/
- **ONNX Runtime WebNN Execution Provider:** https://github.com/microsoft/onnxruntime/tree/main/onnxruntime/core/providers/webnn

WebNN WeChat Group

Awesome WebNN