Transformers.js

State-of-the-art Machine Learning for the web!

June 2024 update

https://github.com/xenova/transformers.js

npm v2.17.2 | downloads 67k/week | jsdelivr 930k/week | license Apache-2.0

Joshua Lochner
joshua@huggingface.co

1

Introduction

What is Transformers.js?

What is Transformers.js?

ML + JS

Run ML models directly in the browser with JavaScript!

Open source

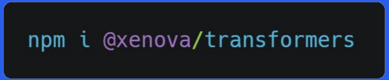
Community-driven development on GitHub. New features added daily!

Easy to use

Add state-of-the-art ML to your web-app in just a few lines of code!

Stars 9.7k

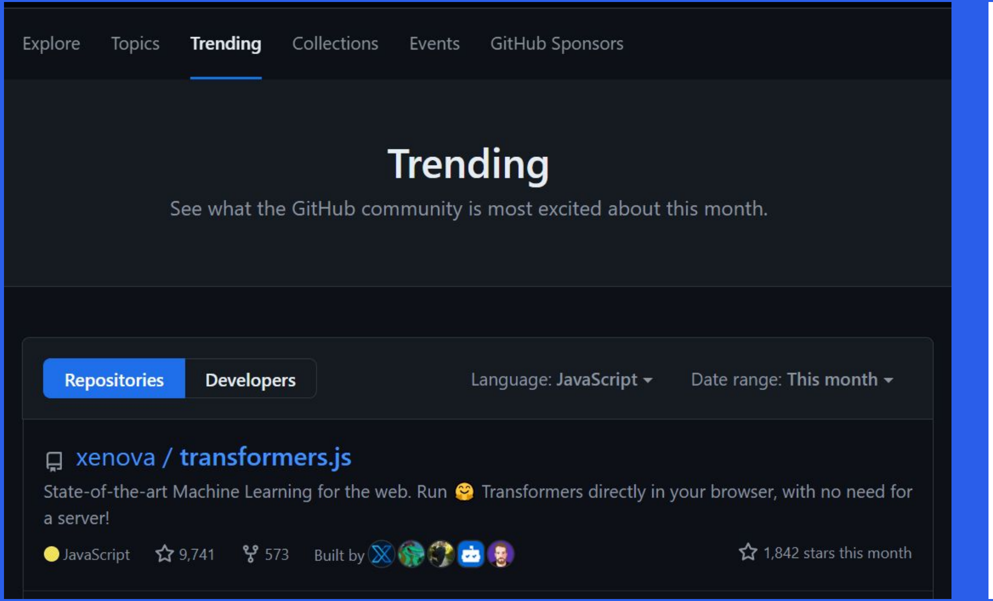
Forks 573



```
npm i @xenova/transformers
```

We have **over 1000** *ready-to-use* models available on the Hugging Face Hub!

Community interest



Explore Topics **Trending** Collections Events GitHub Sponsors

Trending

See what the GitHub community is most excited about this month.

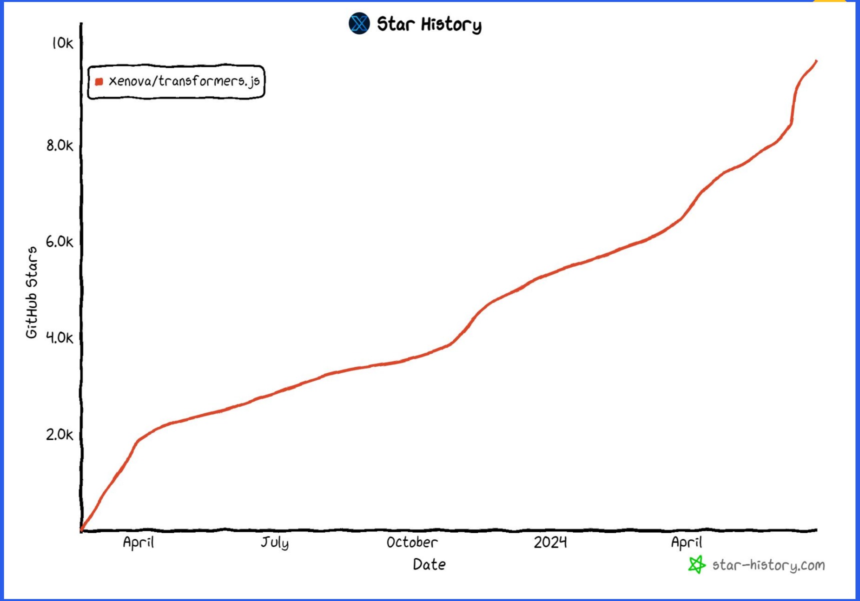
Repositories Developers

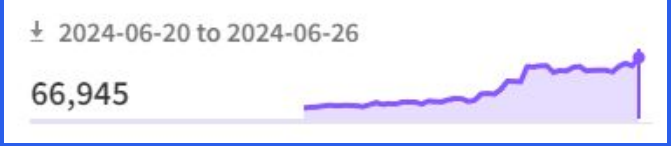
Language: JavaScript ▾ Date range: This month ▾

xenova / transformers.js

State-of-the-art Machine Learning for the web. Run 🤗 Transformers directly in your browser, with no need for a server!

● JavaScript ☆ 9,741 573 Built by ☆ 1,842 stars this month





2024-06-20 to 2024-06-26

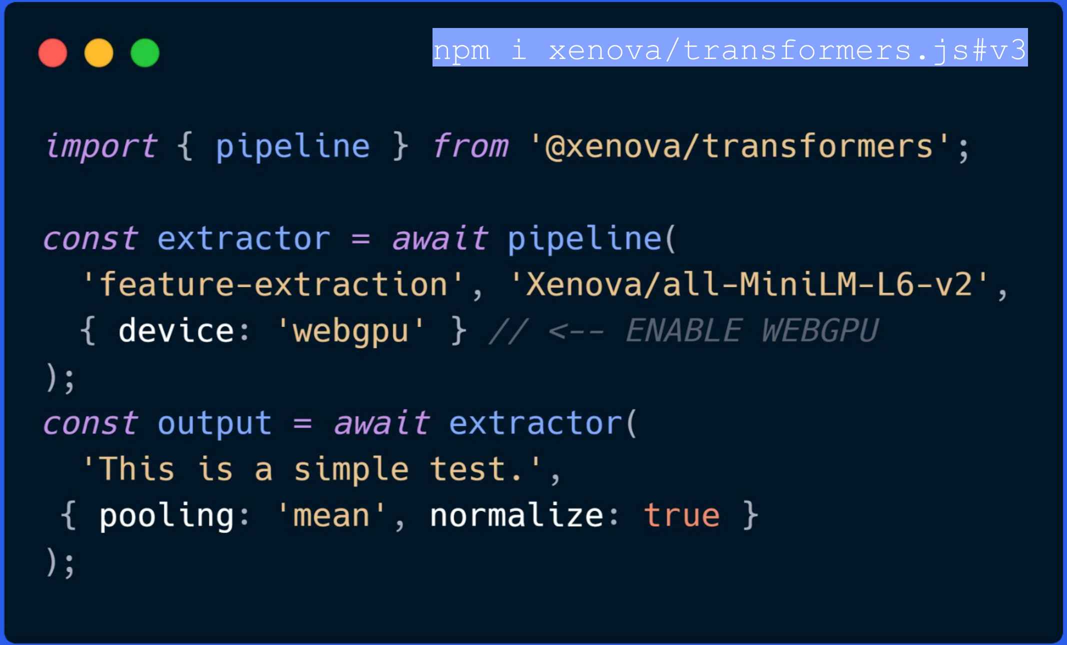
66,945

2

WebGPU support

Experimental in Transformers.js (v3)

How to use it?



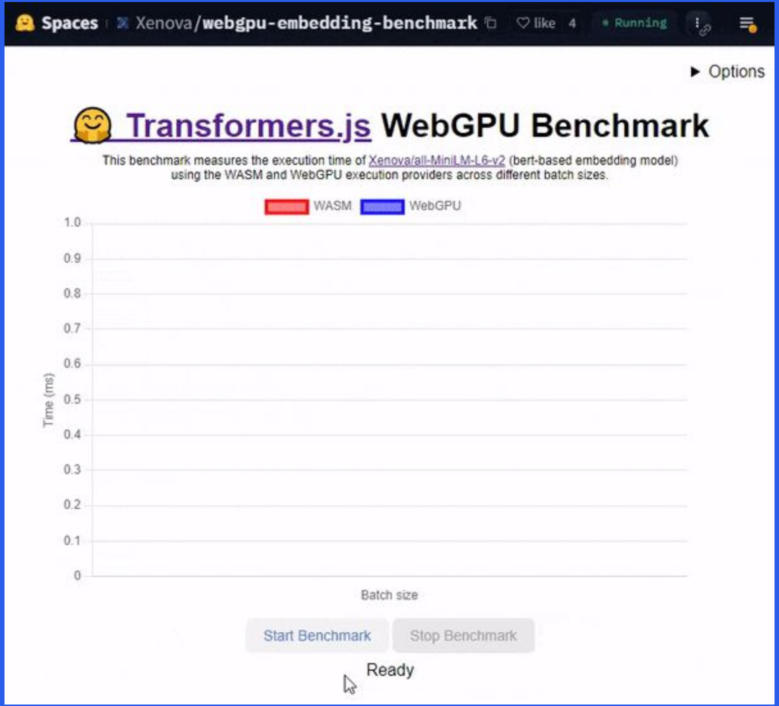
```
npm i xenova/transformers.js#v3
```

```
import { pipeline } from '@xenova/transformers';

const extractor = await pipeline(
  'feature-extraction', 'Xenova/all-MiniLM-L6-v2',
  { device: 'webgpu' } // <-- ENABLE WEBGPU
);
const output = await extractor(
  'This is a simple test.',
 { pooling: 'mean', normalize: true }
);
```

⚠️ EXPERIMENTAL ⚠️

Huge performance boosts!



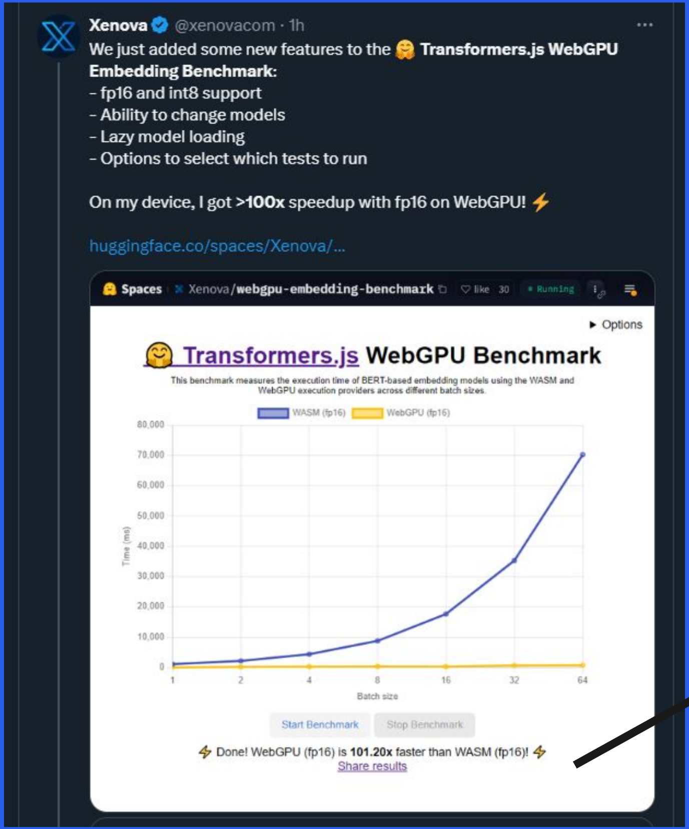
Transformers.js WebGPU Benchmark

This benchmark measures the execution time of Xenova/all-MiniLM-L6-v2 (bert-based embedding model) using the WASM and WebGPU execution providers across different batch sizes.

Ready

https://hf.co/spaces/Xenova/webgpu-embedding-benchmark

fp16 support



Xenova @xenovacom · 1h

We just added some new features to the 🤗 **Transformers.js WebGPU Embedding Benchmark**:
- fp16 and int8 support
- Ability to change models
- Lazy model loading
- Options to select which tests to run

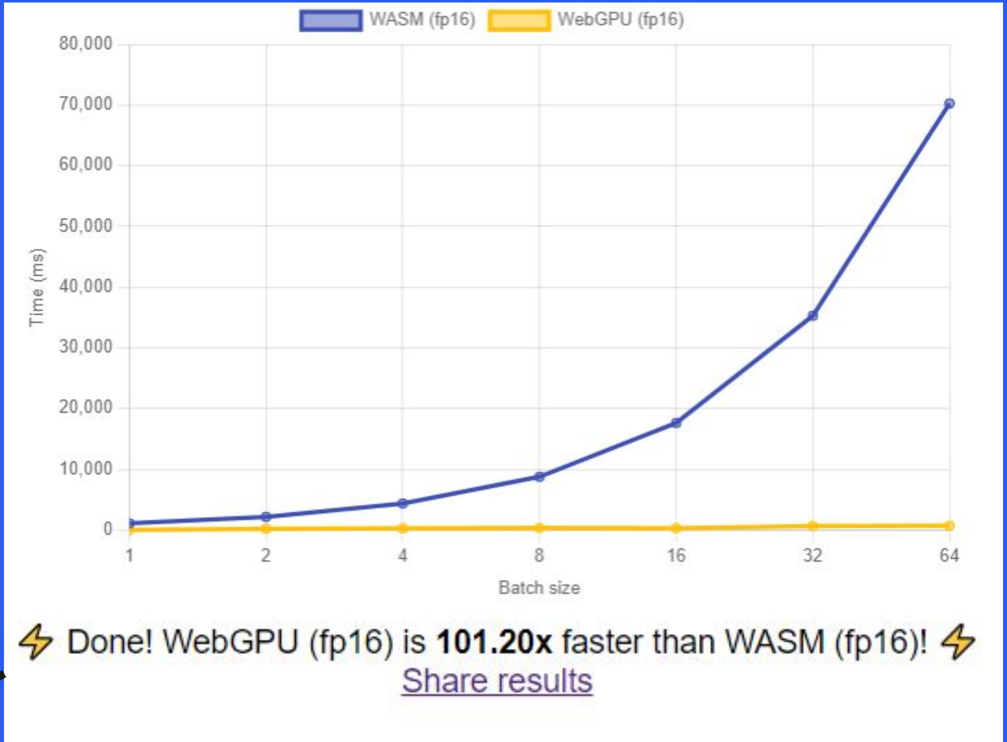
On my device, I got **>100x** speedup with fp16 on WebGPU! ⚡

huggingface.co/spaces/Xenova/...

🤗 Transformers.js WebGPU Benchmark

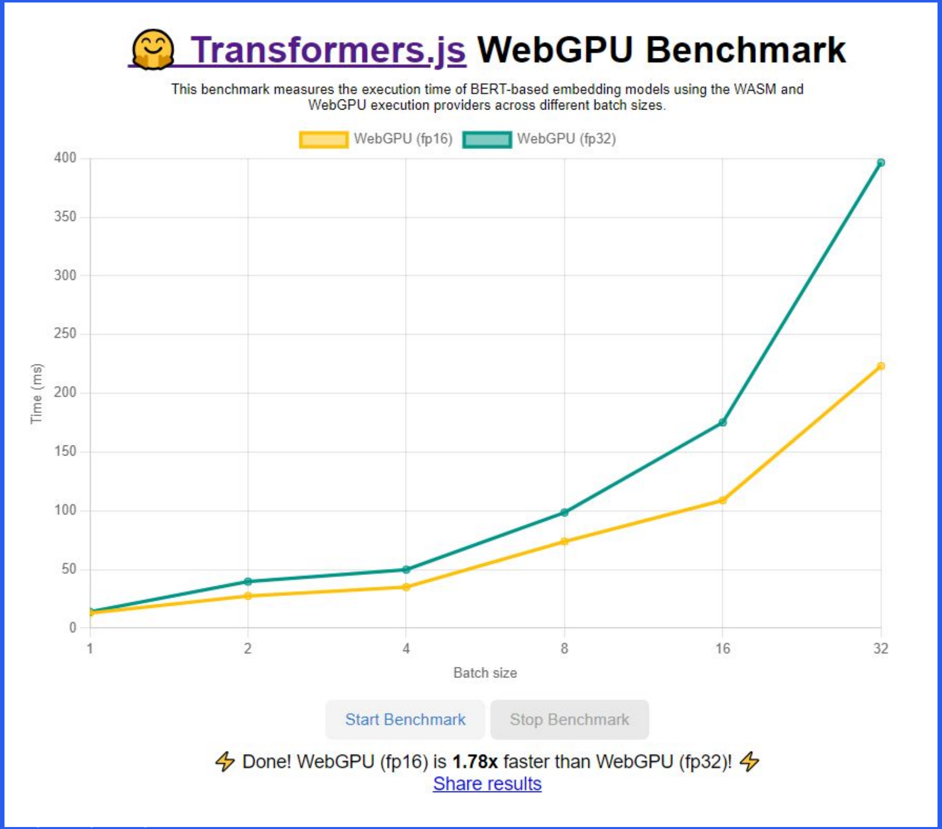
This benchmark measures the execution time of BERT-based embedding models using the WASM and WebGPU execution providers across different batch sizes.

⚡ Done! WebGPU (fp16) is **101.20x** faster than WASM (fp16)! ⚡
Share results



⚡ Done! WebGPU (fp16) is **101.20x** faster than WASM (fp16)! ⚡
Share results

fp16 support

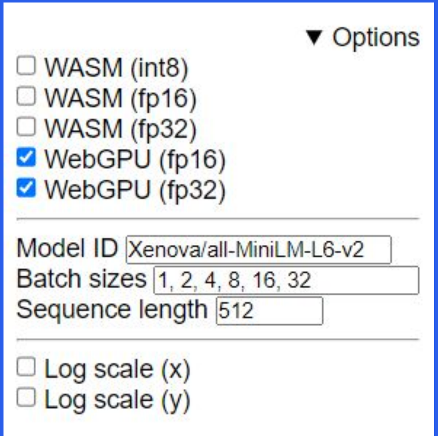


🤗 Transformers.js WebGPU Benchmark

This benchmark measures the execution time of BERT-based embedding models using the WASM and WebGPU execution providers across different batch sizes.

⚡ Done! WebGPU (fp16) is **1.78x** faster than WebGPU (fp32)! ⚡

Share results



▼ Options

☐ WASM (int8)
☐ WASM (fp16)
☐ WASM (fp32)
☑ WebGPU (fp16)
☑ WebGPU (fp32)

Model ID Xenova/all-MiniLM-L6-v2
Batch sizes 1, 2, 4, 8, 16, 32
Sequence length 512

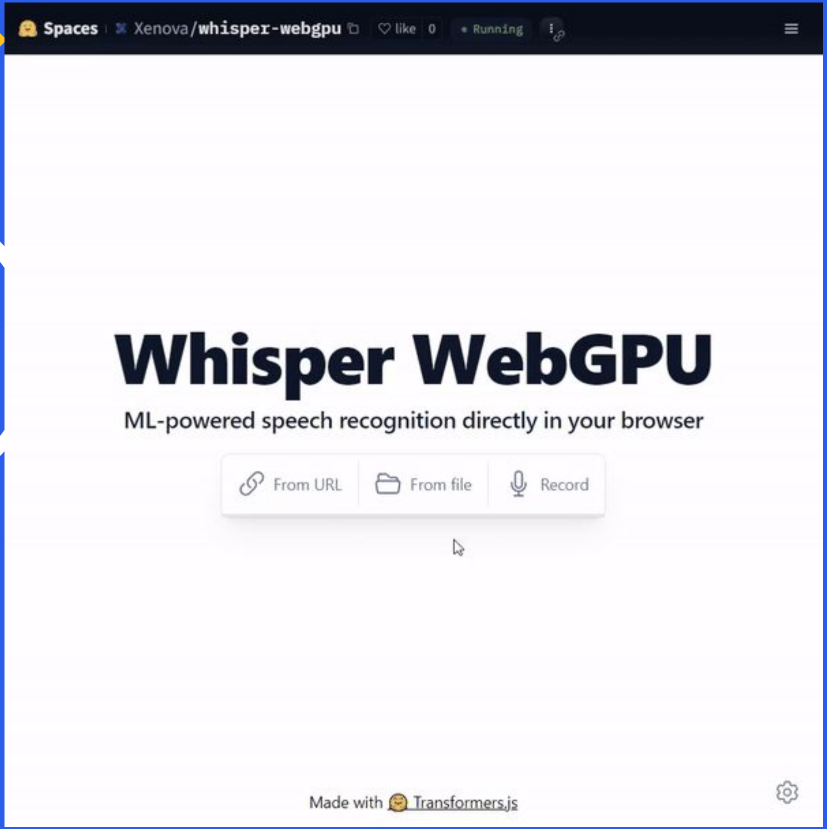
☐ Log scale (x)
☐ Log scale (y)

3

WebGPU demos

Building applications with Transformers.js (v3)

Whisper WebGPU

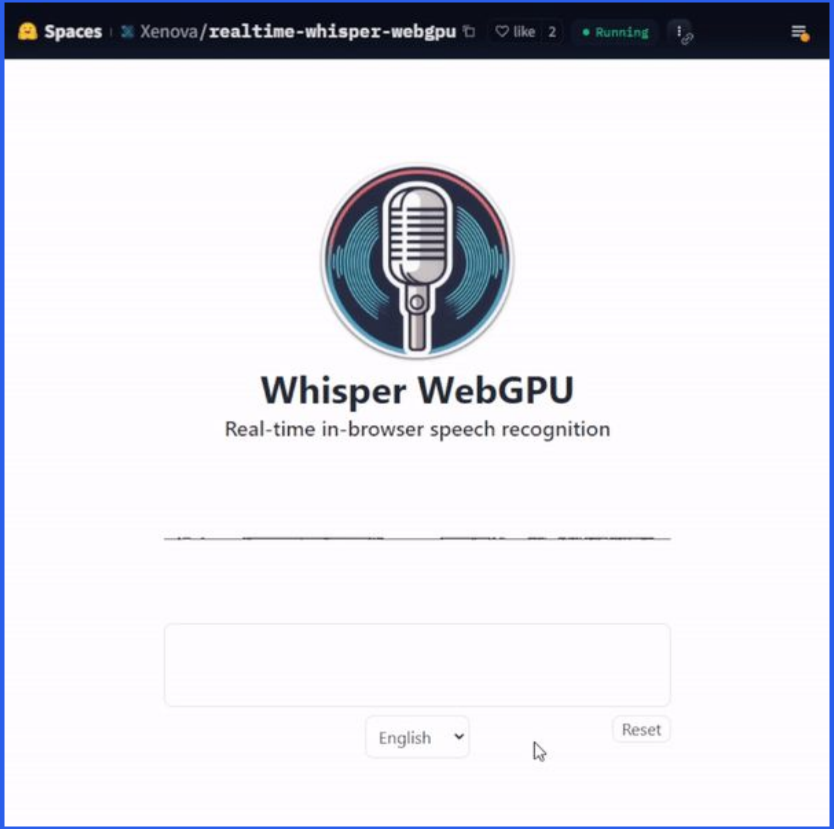


Spaces Xenova/whisper-webgpu like 0 Running

Whisper WebGPU

ML-powered speech recognition directly in your browser

Made with Transformers.js

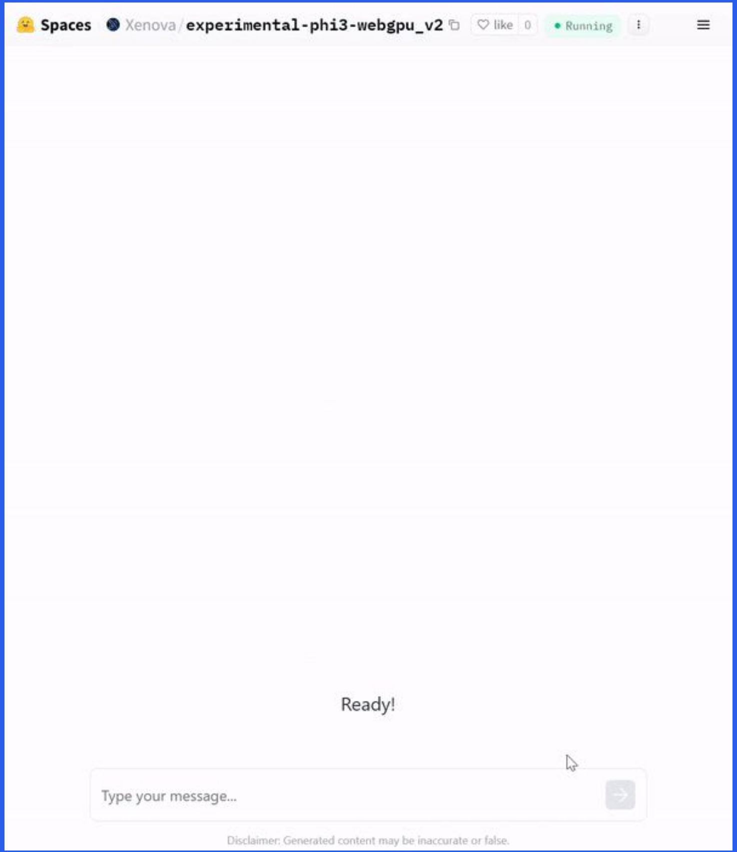


Spaces Xenova/realtime-whisper-webgpu like 2 Running

Whisper WebGPU

Real-time in-browser speech recognition

Phi-3 WebGPU



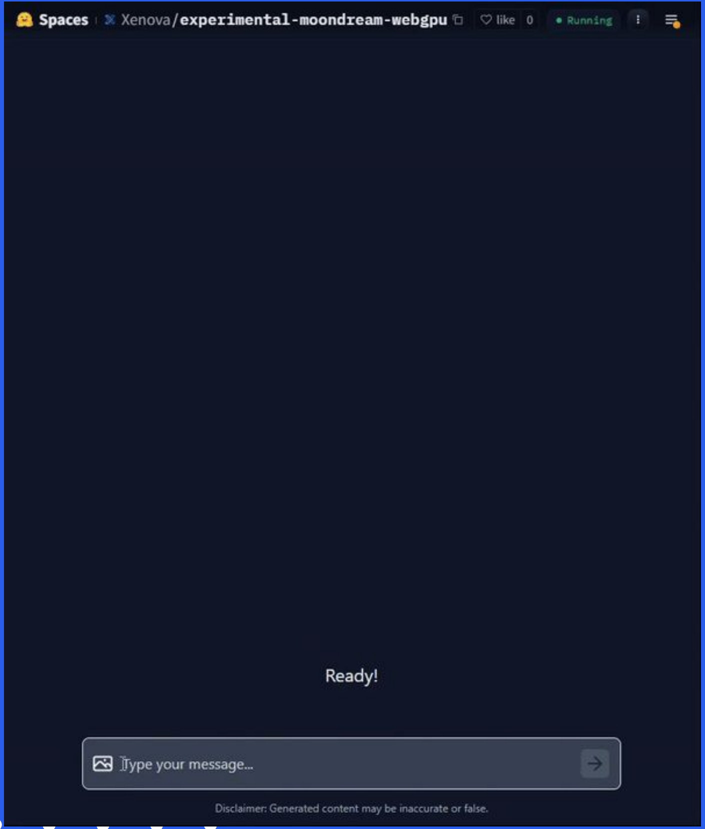
Ready!

Type your message...

Disclaimer: Generated content may be inaccurate or false.

3.82 billion parameter LLM that is optimized for inference on the web

Moondream/LLaVa WebGPU

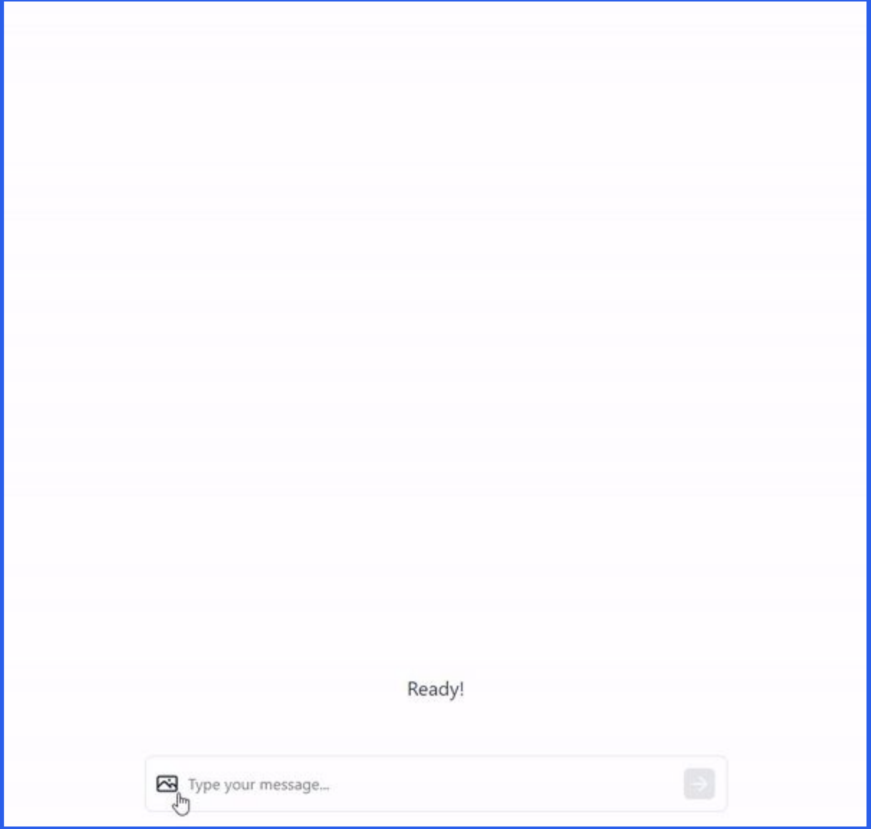


Spaces | Xenova/experimental-moondream-webgpu | like 0 | • Running

Ready!

Type your message...

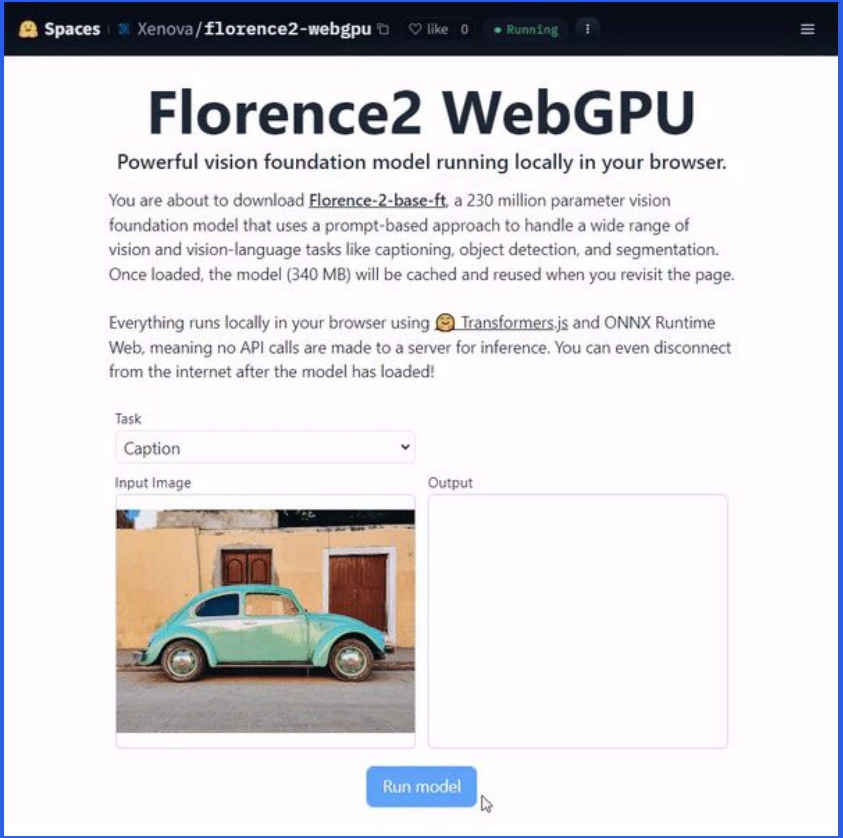
Disclaimer: Generated content may be inaccurate or false.



Ready!

Type your message...

Florence-2



Florence2 WebGPU

Powerful vision foundation model running locally in your browser.

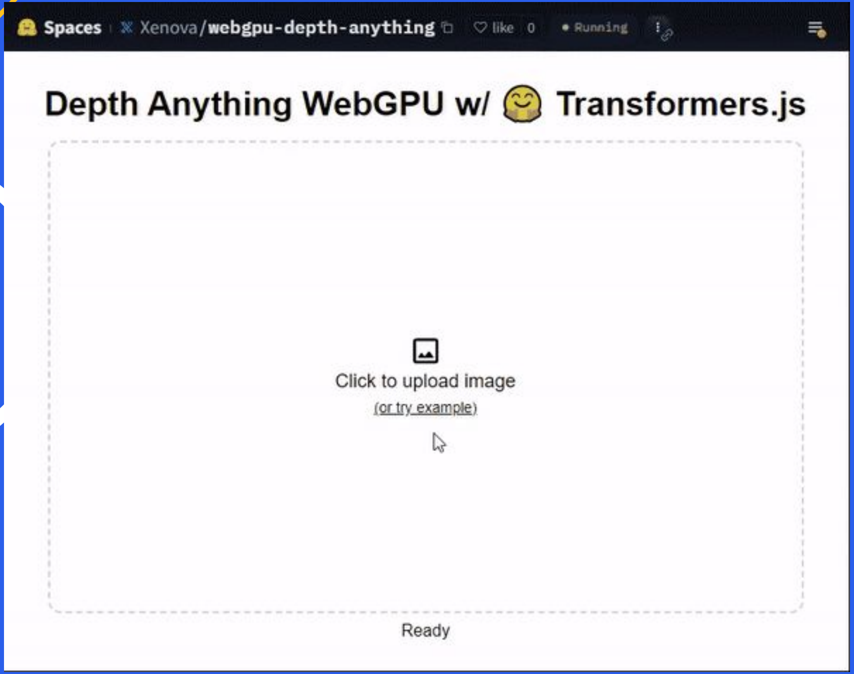
You are about to download Florence-2-base-ft, a 230 million parameter vision foundation model that uses a prompt-based approach to handle a wide range of vision and vision-language tasks like captioning, object detection, and segmentation. Once loaded, the model (340 MB) will be cached and reused when you revisit the page.

Everything runs locally in your browser using 🤗 Transformers.js and ONNX Runtime Web, meaning no API calls are made to a server for inference. You can even disconnect from the internet after the model has loaded!

Task

Caption

Depth Anything



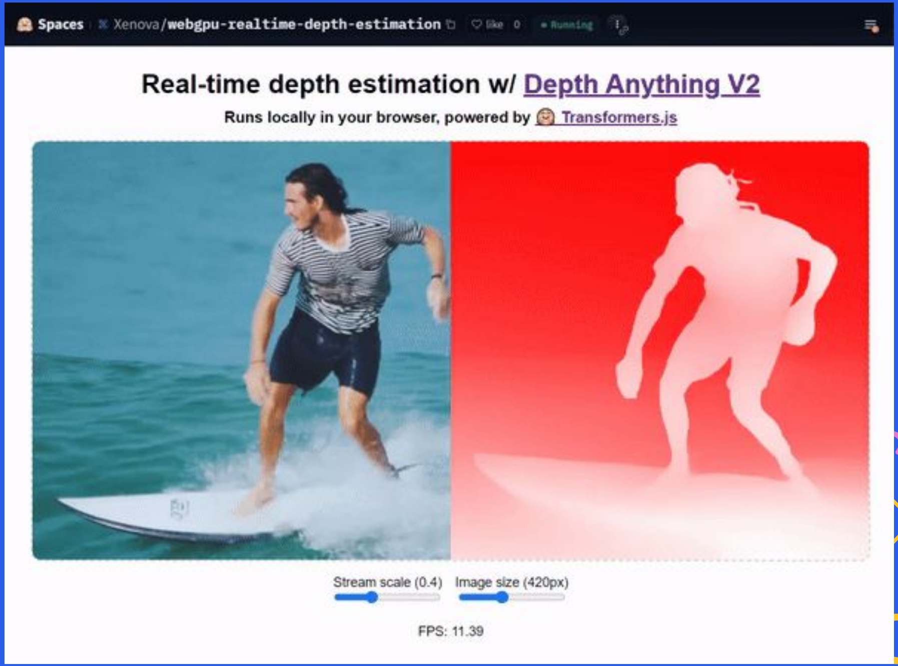
Spaces Xenova/webgpu-depth-anything like 0 Running

Depth Anything WebGPU w/ 🤗 Transformers.js

Click to upload image

(or try example)

Ready



Spaces Xenova/webgpu-realtime-depth-estimation like 0 Running

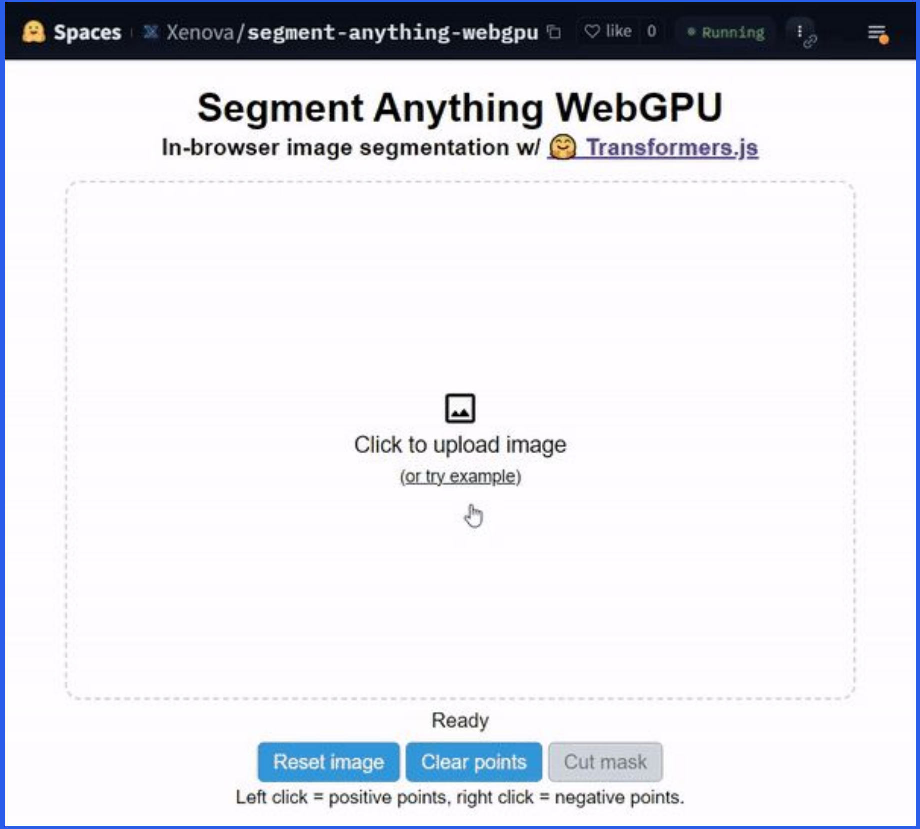
Real-time depth estimation w/ Depth Anything V2

Runs locally in your browser, powered by 🤗 Transformers.js

Stream scale (0.4) Image size (420px)

FPS: 11.39

Segment Anything (SAM)



Segment Anything WebGPU

In-browser image segmentation w/ 🤗 Transformers.js

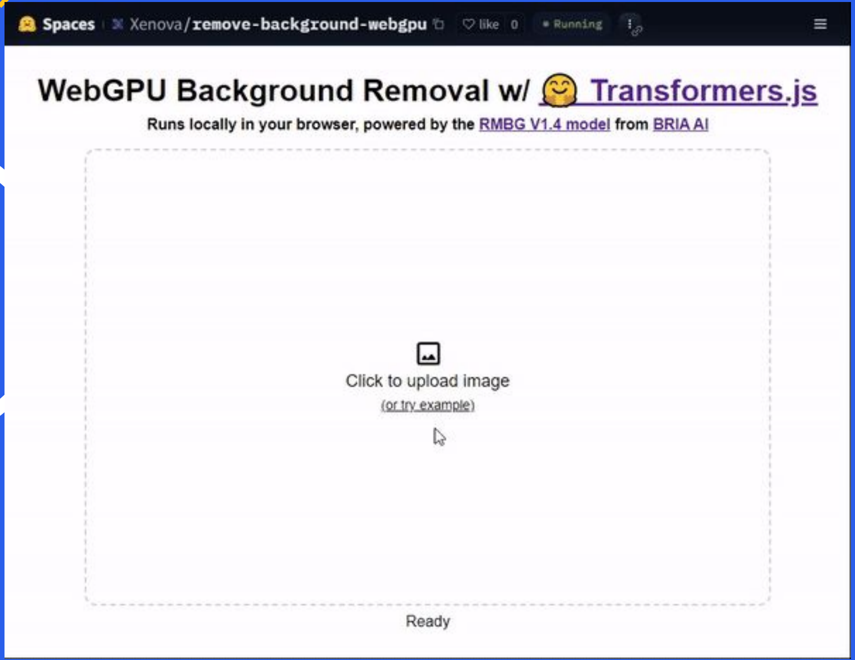
Click to upload image

(or try example)

Ready

Left click = positive points, right click = negative points.

Background Removal



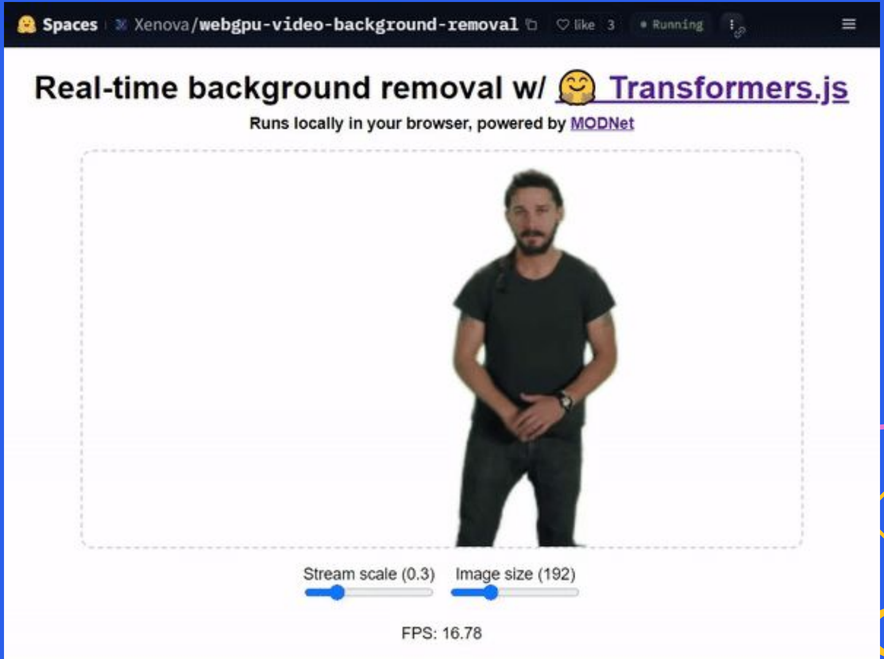
WebGPU Background Removal w/ 🤗 Transformers.js

Runs locally in your browser, powered by the RMBG V1.4 model from BRIA AI

Click to upload image

(or try example)

Ready



Real-time background removal w/ 🤗 Transformers.js

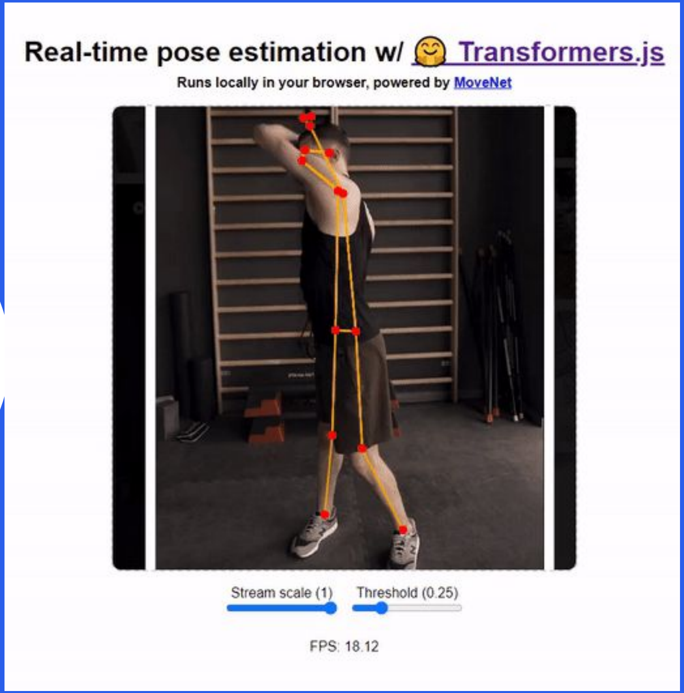
Runs locally in your browser, powered by MODNet

Stream scale (0.3)

Image size (192)

FPS: 16.78

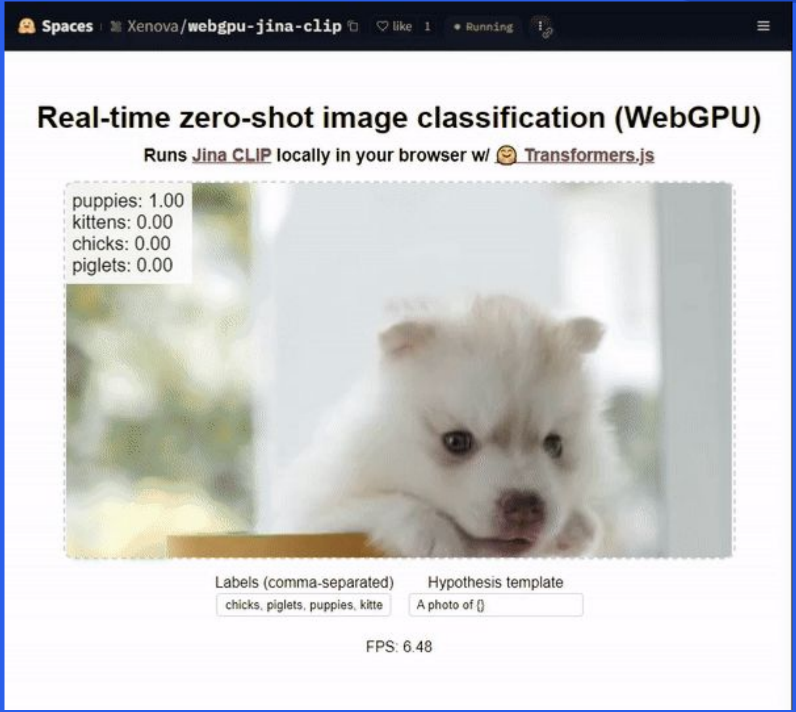
Other tasks



Real-time pose estimation w/ 🤗 Transformers.js

Runs locally in your browser, powered by MoveNet

Pose estimation



Real-time zero-shot image classification (WebGPU)

Runs Jina CLIP locally in your browser w/ 🤗 Transformers.js

puppies: 1.00
kittens: 0.00
chicks: 0.00
piglets: 0.00

Labels (comma-separated): chicks, piglets, puppies, kitte

Hypothesis template: A photo of {}

FPS: 6.48

Zero-shot image classification
(classes defined at runtime)

Thanks!

Any questions?

@xenovacom

joshua@huggingface.co

@xenova

CREDITS: This presentation template was created by **Slidesgo**, and includes icons by **Flaticon**, infographics & images by **Freepik** and illustrations by **Storyset** and **Chunte Lee**