GPU support for browsers

Alexander Visheratin

alex@visheratin.com

https://twitter.com/visheratin

About me

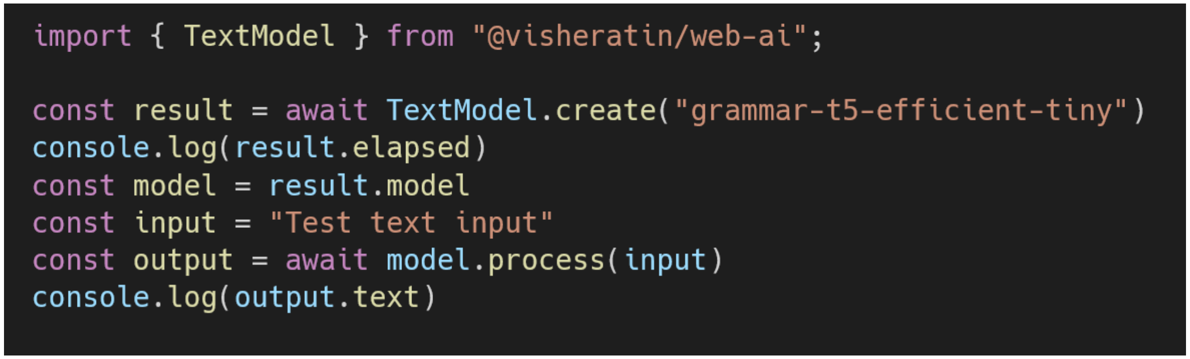
- AI engineer in Beehive AI – startup working on large-scale survey processing.
- Last year got interested in running neural networks in the browser.
- Started Web AI project - https://github.com/visheratin/web-ai

Growing interest in on-device AI

- Projects like llama.cpp and whisper.cpp – run huge models on M1/M2 chips.
- Web stable diffusion – use TVM Unity for compiling the model for using WebGPU in browser.
- ONNX-based projects – Web AI. Use ONNX runtime to run variety of image and text models.
- WONNX – WebGPU-enabled ONNX for Web.

Web AI

Simplify using DL models for web applications by implementing all complexities – pre-/post-processing, tensors, inference – inside the library:



```
import { TextModel } from "@visheratin/web-ai";

const result = await TextModel.create("grammar-t5-efficient-tiny")
console.log(result.elapsed)
const model = result.model
const input = "Test text input"
const output = await model.process(input)
console.log(output.text)
```

- Works great with multi-threaded WASM runtime.
- Supports image, text, and multi-modal models.
- Flexible configuration for models.
- WASM-based tokenizers for Hugging Face models.

Proposal

- WebGPU support for ONNX Runtime for Web.
- Instructions and tutorials for implementing operators for WebGPU.