Paddle.js: Machine Learning for the Web

Ping Wu

Baidu

2020/8

Agenda

- What is Paddle.js
- Design Principal
- Implementation
- Use Scenario
- Conclusion and Future Work

What is Paddle.js

- Paddle.js is a high-performance DL framework for JavaScript, which provides on-device computation in diverse web runtimes, including PC, mobile, browser and Mini-Program.
- Part of Baidu PaddlePaddle ecosystem, compatible APIs with PaddlePaddle python/C++ part.
- Currently only inference, no training part. But provide remote rpc JS APIs to PaddlePaddle serving part.

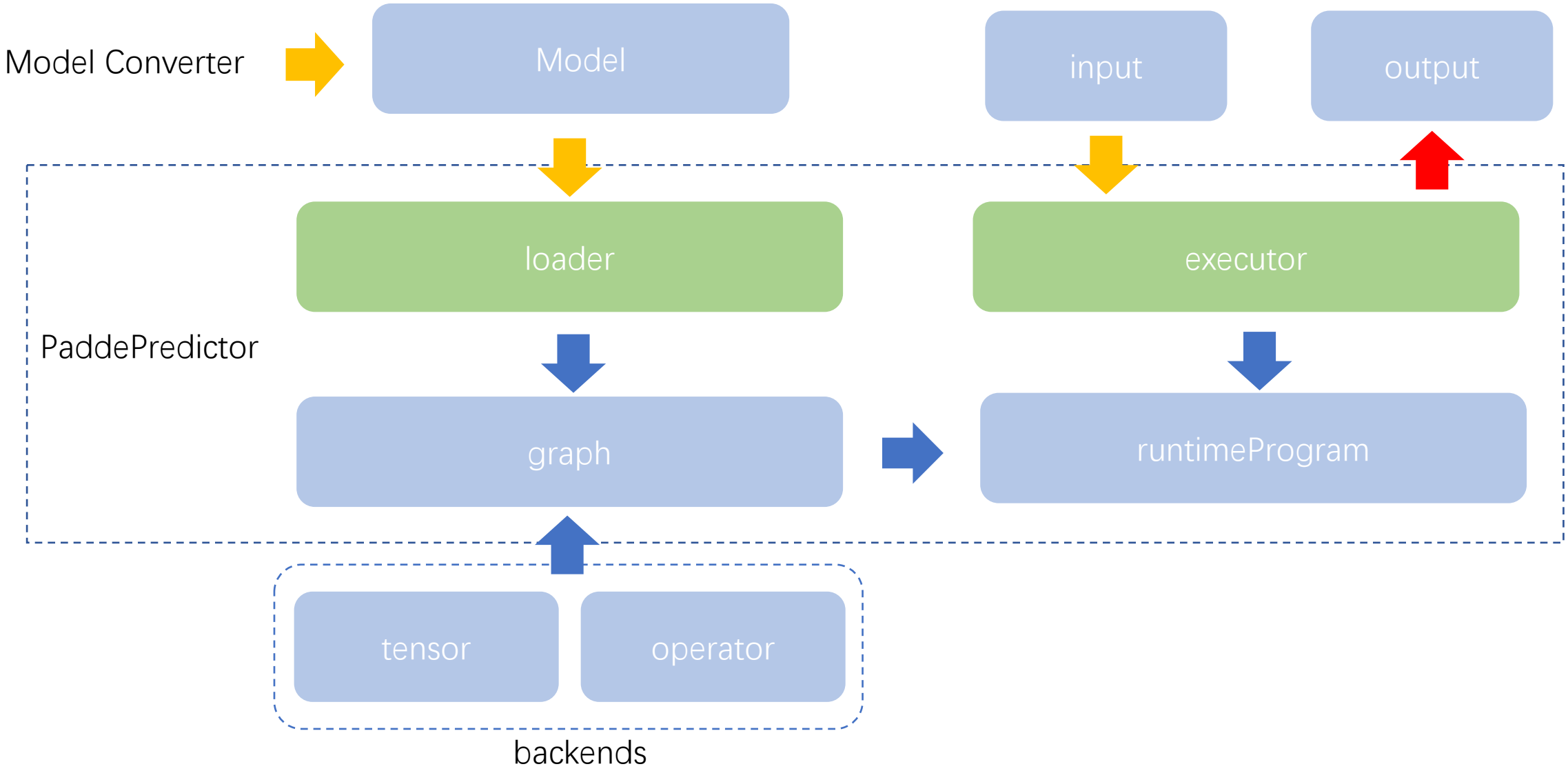
Opportunities and Challenges for WebAI

- Chances we have
  - Vast FE developer community which continues to grow and expand.
  - Low barrier to develop and deploy, easy to experience and share due to cross-platform web runtime support.
  - On-device computation for privacy, real-time, offloaded and decentralized end computation.
- Challenges we face
  - High performance computation in web runtime
  - diverse web runtime and cross-browser compatibility

Design Principal

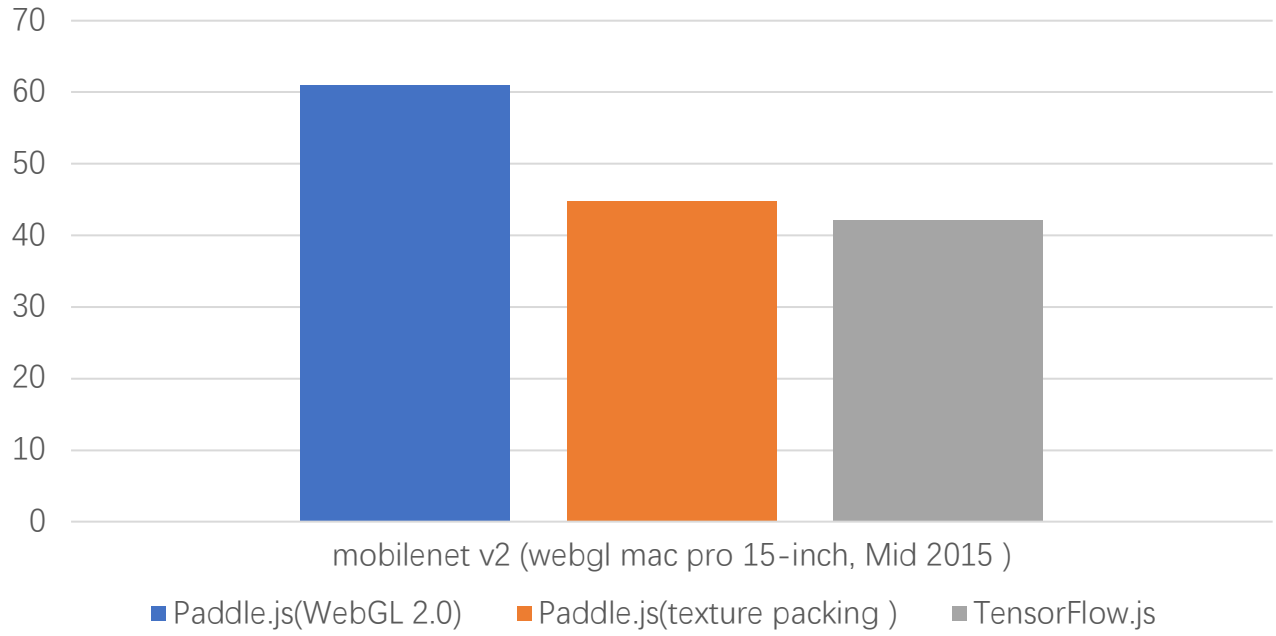
- Integrated with PaddlePaddle ecosystem
  - Fully utilize PaddlePaddle model, toolchain, and inference experience we have on other on-device platforms.
  - A good start for entry level developers and also help experienced PaddlePaddle developers easily migrate work to JS environment.
- High Performance
  - Efficient web GPU backend for op & kernel implementation
  - Efficient data I/O
- Platform Compatibility
  - Cross-browser
  - Cross-device
  - H5 and Mini-Program

Overall Architecture and APIs



Performance

- Computation with different backend-WebGL，WebGPU，WebNN
- Initialization Cost
- Memory Management and GC issues



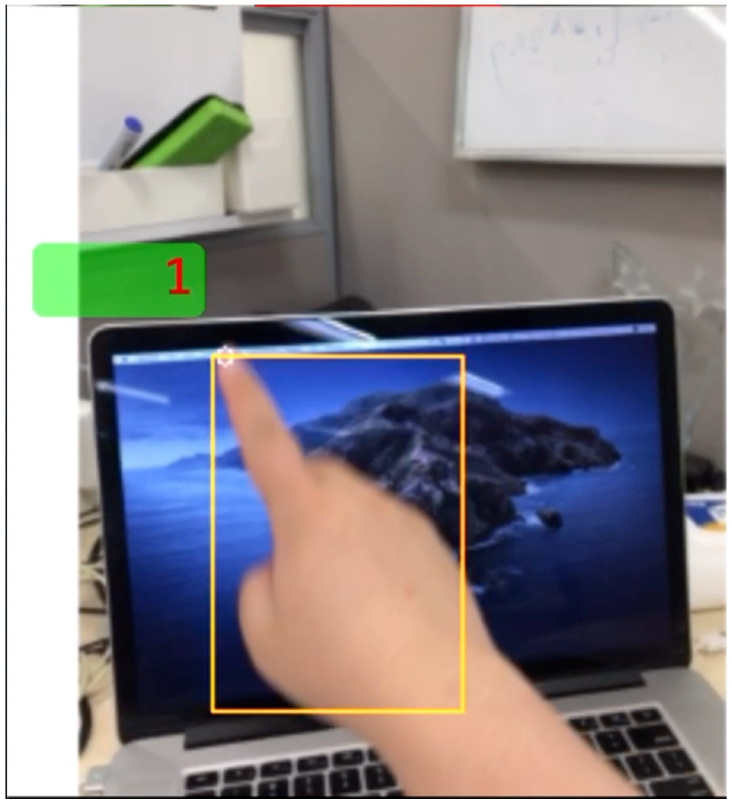
Compatibility

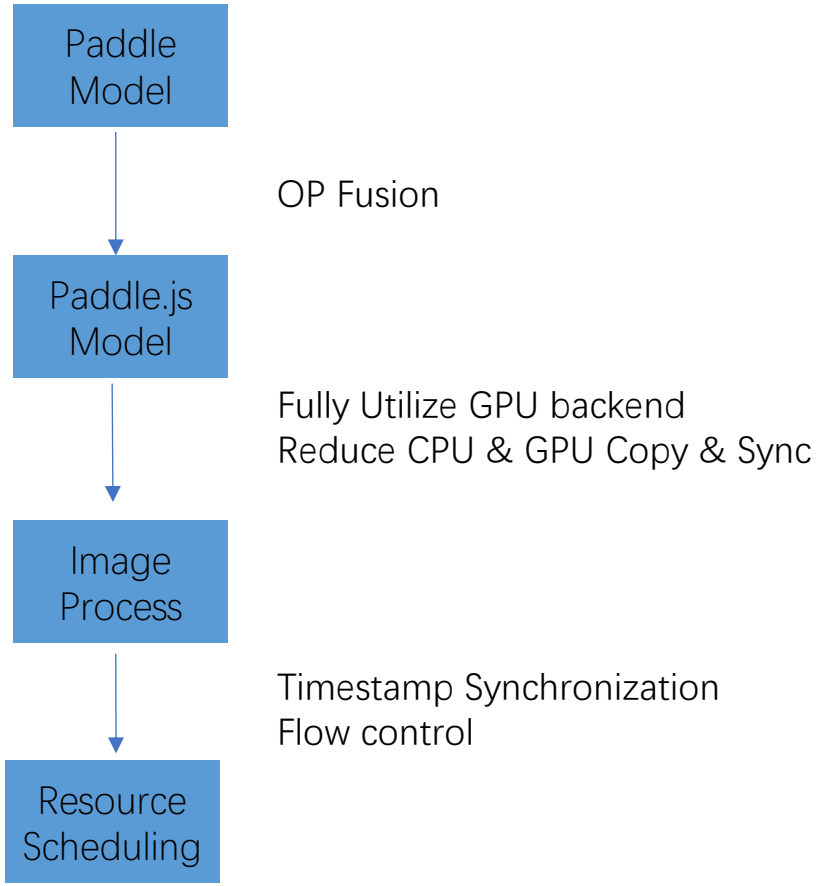
- Paddle.js supports WebGL 1.0 & 2.0, runtime compatible with device that supports OES texture float extension.
- Mobile device 16bit float support
  - Due to lack of 32-bit float support on almost all mobile GPUs, we may have precision lost. Half-float **quantization** may also work efficiently in many situations.

	Android-Chrome	ios-chrome	Mac-safari	Mac-chrome	Windows-chrome	Ubuntu-chrome	ios-safari	Wechat	Baidu App	QQ Browser	UC Browser
mobileNet	√	√	√	√	√	√	√	√	√	√	√
TinyYolo	√	√	√	√	√	√	√	√	√	√	√
TerrorModel	√	√	√	√	√	√	√	√	√	√	√
Huangfan	√	√	√	√	√	√	√	√	√	√	√
Humanseg	√	√	√	√	√	√	√	√	√	√	√

Use Case –Real-time Interactive Application

- Gesture Recognition & Tracking on mobile device with whole-process optimization





Conclusion & Future Work

- Paddle.js is a high-performance JavaScript DL framework for diverse web runtimes, which helps building a PaddlePaddle ecosystem with web community.
- Future work may include
  - A general and high performance numeric computing programming model for web runtime.
  - More Toolchain and developer framework support for Paddle.js developers.
  - More innovations in new classes of web AI applications.

Thank You!