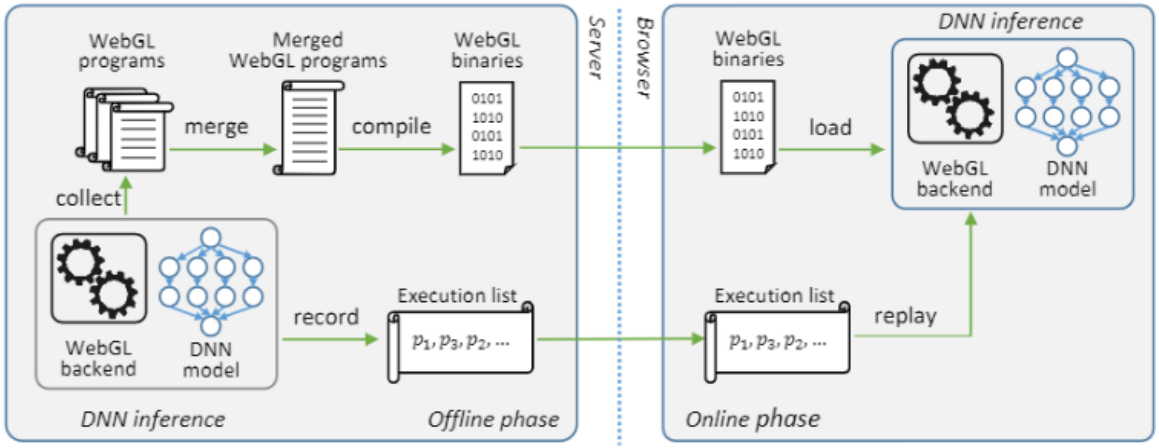
WPIA: Accelerating DNN Warm-up in Web Browsers by Precompiling WebGL Programs

Deyu TIAN, Yun MA, Yudong HAN, Qi YANG, Haochen YANG, Gang HUANG

Frontiers of Computer Science, DOI: 10.1007/s11704-024-40066-w

Problems & Ideas

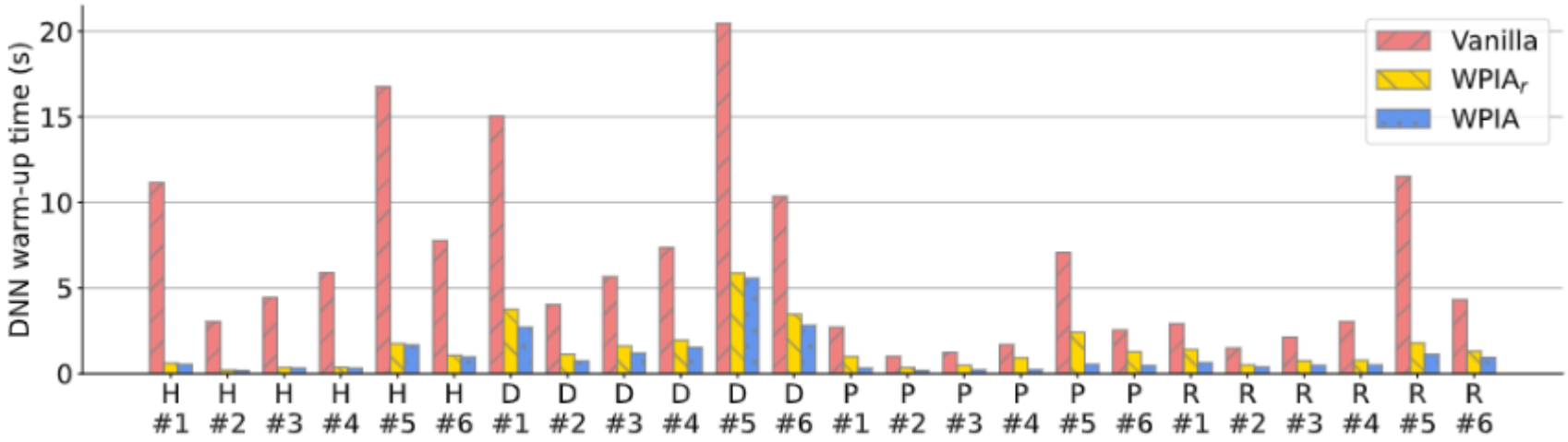
- Problems of DNN model warm-up in Web browsers:
  - Web apps need to warm up a DNN model before execute DNN inference on the model, and the warm-up is terribly slow.
  - Most of the time spent in DNN warm-up is the Web browser compiling WebGL programs used in DNN inference
- Ideas: Precompiling WebGL programs at server side, and loading the compiled WebGL programs into Web browsers to avoid compiling WebGL programs on the fly.



The overview of our approach, WPIA, which facilities WebGL program precompiling to optimize DNN warm-up for Web apps. WPIA merges WebGL programs to reduce WebGL binaries size, and use record-and-replay technique to handle the execution of precompiled WebGL programs.

Main Contributions

- Contributions:
  - We investigate the reason for the long DNN model warm-up time in Web apps and find that compiling WebGL programs into binaries takes most of the time.
  - We propose WPIA, an approach which reduces the DNN warm-up time in Web apps by precompiling WebGL programs offline.
  - We evaluate WPIA, and the results show that WPIA can effectively accelerate the DNN warm-up time in Web browsers to one order of magnitude faster with negligible overhead.



DNN warm-up times of Vanilla and our WPIA. The x-axis represents combinations of different devices (H, D, P, and R) and DNN models (#1 ~ #6). The y-axis shows the DNN warm-up time.