WebML WG Virtual Meeting at TPAC 2021 - Day 3

28 October 2021

txt

Attendees

Present Anssi_Kostiainen, Belem_Zhang, Chai_Chaoweeraprasit, Dom, Eric_Meyer, Feng_Dai, Geun-Hyung, Geun-Hyung_Kim, Judy_Brewer, Junwei_Fu, Ningxin_Hu, Rachel_Yager, Rafael_Cintron, Takio_Yamaoka, Wanming, Zoltan_Kis

Regrets -

Chair Anssi

Scribe Anssi, anssik, dom

Contents

1. Conformance testing of WebNN API
    1. Web Platform Tests
2. Ethical issues in using Machine Learning on the Web

Meeting minutes

1 Conformance testing of WebNN API

Anssi: interoperability testing helps ensure compatibility among existing and future implementations
… in the context of ML, reaching interop is not necessarily easy given the variety of underlying hardware
… Chai is involved in Microsoft DirectML and has experience in this space

Slideset: https://lists.w3.org/Archives/Public/www-archive/2021Oct/att-0017/Conformance_Testing_of_Machine_Learning_API.pdf
[Slide 1]

Chai: conformance testing of ML APIs is quite important

[Slide 2]

chai: the problems can be categorized into 3 categories:
… the ML models need to run on a wide variety of specialized hardware
… my work with DirectML is at the lowest level before the hardware in the windows OS
… windows has a very broad scale of hardware
… esp with specialized accelerators
… they don't share the same architecture and have very different approach to computation
… ensuring the quality of results across this hardware is really important

… another issue is that most modern AI computation relies on floating point calculation
… FP calculation with real numbers accumulate errors as you progress in the computation - that's a fact of life
… there are trimming problems which create challenges in testing the results of ML API across hardware
… this is a daily issue in my work testing Direct ML

[Slide 3]

Chai: Karen Zack's Animals vs Food prompted a an actual AI challenge
… humans don't have too much difficulty doing the difference, but while many models are able to perform, they tend to give results with some level of uncertainty
… showing the importance of reliability across hardware

[Slide 4]

Chai: when we run the results of ML models, there are 4 groups of variability
… the most obvious one is precision differences - half vs double precision will give different results
… most models run with single precision float, but many will run with half
… Another bucket is hardware differences - even looking at CPU & GPUs, different chipset may have slightly different ways of computing and calculating FP operations
… accelerators are often DSP based; some may rely on fixed point calculation, implying conversion, to very different type of formats (e.g. 12.12, 10.10)
… A third source of variability is linked to algorithmic differences
… there are different ways of implementing convolutions, leading to different results
… Finally, there is numerical variability - even on the same hardware, running floating point calculation, there may be slight difference across runs
… and that can be amplified by issues of lossy conversion between floating point to fixed point,
… these issues compound one with another, so there is no guarantee of reproducible results

[Slide 5]

Chai: how do we deal with that in testing?
… Many test frameworks use fuzzy comparison that provides an upper boundary (called epsilon) to an acceptable margin of differences
… the problem of that approach in ML is that it doesn't deal with the source of variabilities we identified
… A better way of comparing floating point values is based on ULP, unit of least precision
… the distance measured between consecutive floating point values
… a comparison between the binary representation of different floating point values, applicable to any float point format
… Using ULP comparison removes the uncertainty on numerical differences
… it also mitigates the hardware varaibility in terms of architectural differences because it compares the representations

[Slide 6]

Chai: this piece of code illustrates the ULP comparison
… the compare function convert the floating point number into a bitwise value that is used to

calculate the difference and how much ULP that represents
… e.g. here, only a difference of 1 ULP is deemed acceptable
… We use ULP to test DirectML
… the actual floating point values from the tests are never the same

[Slide 7]

Chai: to make the comparison, you need to define a point of reference, which we call the baseline
… the baseline is determined by the best known result for the computation, the ideal result
… this serves as a stable invariant
… for directML, we have computed standard results on a well-defined CPU with double precision float
… we use that as our ideal baseline
… we then define the tolerance in terms of ULP - the acceptable difference between what is and what should be (the baseline)
… the key ideas here are #1 use the baseline, #2 define tolerance in terms of ULP

[Slide 8]

Chai: the strategy of constructing tests can be summarized in 5 recommendations:
… we recommend testing both the model and the kernels
… each operator should be tested separately, and on top of that, a set of models that exercise the API and run the results of the whole model
… for object classification models, you would want to compare the top K results (e.g. 99% Chiwawa, 75% muffin)
… making sure e.g. the 3 top answers are similar
… it's possible to have tests passing at the kernel level, but failing at the model level
… 2nd point: define an ideal baseline and ULP-based tolerance
… you might have to fine-tune the tolerance for different kernels
… e.g. addition should have very low ULP, vs square root or convolution

anssi: thanks for the presentation
… highlights how different from usual Web API testing is in the field
… most likely similarities are with GPU and graphic APIs
… We've had some early experimentation with bringing tests to WPT, the cross-browser platform testing project that is integrated with CI

RafaelCintron: any recommendation in terms of ULP tolerance? what does it depend on?

Chai: simple operations like addition, low tolerance (e.g. 1 ULP)
… for complex operations, the tolerance needs to be higher
… sometimes, the specific range arises organically e.g. for convolution we've landed around 2-4
… different APIs have different ULP tolerance, although they're likely using similar values

<rachel> is precision testing necessary for all applications?

Chai: strategically, the best approach is to start with low tolerance (e.g. 1 ULP), and bump it based on real-world experience

Rachel: [from IRC] is precision testing necessary for all applications?

Chai: yes and no
… you can't test every single model
… testing the kernel, the implementation of the operators
… with an extensive enough set of kernel testing, the model itself should end up OK
… there are rare cases where the kernel tests are passing, but a given model on a given hardware will give slightly different results
… but the risks of that are lower if the kernels are well tested

Ningxin: regarding the ideal baseline, for some operators like convolution, there can be different algorithms
… what algorithm do you use for the ideal baseline?
… Applying this to WebNN may be more challenging since there is no reference implementation to use as an ideal baseline

chai: for DirectML, we implement the reference implementation using the conceptual algorithm in a CPU with double precision
… this is not what you would get from a real world implementation, but we use that as a reference
… For WebNN, we may end up needing a set of reference implementations to serve as a point of comparison
… there is no shortcut around that
… having some open source code available somewhere would be good
… but no matter what, you have to establish the ideal goal post

Web Platform Tests

FengDai: I work on testing for WebNN API and have a few slides on status for WPT tests

Slideset: fengdaislides
[slide 3]

FengDai: 353 tests available for idlharess
… we've ported 800 test cases built for the WebNN polyfill to the WPT harness
… this includes 740 operator tests (340 from ONNX, 400 from Android NNAPI)

[WebNN WPT tests (preview in staging)]

FengDai: for 60 models tests use baseline calculated from native frameworks
… the tests are available as preview on my github repo

[slide 4]

Anssi: thanks for the great work - the pull request is under review, correct?
… any blocker?

FengDai: there are different accuracy settings, data types across tests
… this matches the challenges Chai mentioned

Anssi: the good next step might to join one of the WG meeting to discuss this in more details

Chai: thanks Bruce for the work! WPT right now relies on fuzzy comparison

… this means we'll need to change WPT to incorporate ULP comparison
… hopefully that shouldn't be too much code change

FengDai: thanks, indeed

2 Ethical issues in using Machine Learning on the Web

Ethical Web Machine Learning Editors draft

Anssi: this is a document that I put in place a few weeks ago
… the WG per its charter is committed to document ethical issues in using ML on the Web as a WG Note
… this is a first stab
… big disclaimer: I'M NOT AN EXPERT IN ETHICS

Ethical Web Machine Learning

Anssi: we're looking for people with expertise to help
… this hasn't been reviewed by the group yet
… [reviews the content of the document]
… ML is a powerful technology, enables new compelling UX that were thought as magic and are now becoming commonplace
… these technologies are reshaping the world
… the algorithms that underline ML are largely invisible to users, opaque and sometimes wrong
… they cannot be introspected but sometimes are assumed to be always trustworthy
… this is why it is important to consider ethical issues in the design phase of the technology
… it's important that we understand the limitations of the technology
… the document then reviews different branches of ethics: information ethics, computer ethics, machine ethics
… there is related work in W3C
… e.g. the horizontal review work on privacy, accessibility
… and the TAG work on ethical web principles

Privacy-by-design web standards
Accessibility techniques to support social inclusion
W3C TAG Ethical Web Principles

Anssi: the document is focusing on ethical issues at the intersection of Web & ML
… there are positive aspects to client-side ML: increased privacy, and reduced risk of single-point-of-failure and distributed control
… it allows to bring progressive enhancement in this space
… Browsers may also help increasing transparency, pushing for greater explainability
… in the spirit of "view source"
… I've looked at different litterature studies in this space

Rachel: I'm interested in this and suggesting including a research into thinking of corporations

… many companies have efforts for responsible AI, so engaging with them is interesting
… focusing on human perspective of this may be a good focus
… can work with W3C Chapter to bring interested folks from that group into this discussion

---

Minutes manually created (not a transcript), formatted by scribe.perl version slide-shower-184 (Tue Nov 30 23:12:07 2021 UTC).

Diagnostics

Succeeded: s/Fang/Feng

Succeeded: s/fferent/fference/

Succeeded: s|chaislides|https://lists.w3.org/Archives/Public/www-archive/2021Oct/att-0017/Conformance_Testing_of_Machine_Learning_API.pdf

Succeeded: s/powerful document/powerful technology

Maybe present: Anssi, Chai, FengDai, Ningxin, Rachel, RafaelCintron