GANpiler

Authors

Barry Rountree, Center for Applied Scientific Computing, Lawrence Livermore National Laboratory

Hyun-Seob Song, Biological Systems Engineering Department, University of Nebraska--Lincoln

Huiying Ren, Energy and Environment Directorate, Pacific Northwest National Laboratory

Aaron Donahue, Atmospheric, Earth and Energy Division, Lawrence Livermore National Laboratory

Tapasya Patki, Center for Applied Scientific Computing, Lawrence Livermore National Laboratory

Aniruddha Marathe, Center for Applied Scientific Computing, Lawrence Livermore National Laboratory

Focal Area(s)

Rather than augment or replace physical models with machine learning, we instead preserve the existing models and augment the underlying code with faster surrogate models created by generative adversarial networks (GAN). To leverage the performance optimization, we also propose a runtime system and user interfaces that allow prediction and tracking of accumulated error as well as dynamic, per-process decision making as to which model (if any) to use for each iteration.

Science Challenge

This proposal sits at the nexus of two hard problems. First, climate models based on machine learning will be, by their nature, difficult to trust once their predictions begin diverging from the consensus. Second, compilers and hardware have been making only incremental performance gains for decades. GPGPUs have provided a welcome performance boost for codes that can take advantage of them, but there is no similar technology on the horizon to provide the next performance leap.

Rationale

Need: A computational paradigm that can leverage the benefits from advances in machine learning and extreme heterogeneity in supercomputing architecture without abandoning explainable and discoverable models.

Approach: Rather than arguing for replacing physical models with ML derivations, we wish to preserve those verified and validated models and instead replace specific computational modules/functions in the code with computationally more efficient ML-based implementation.

DOI: https://doi.org/10.2172/1769713
DOE BER, AI4ESP 2021 #AI4ESP1115 (https://www.ai4esp.org)

Benefit: In our approach, results from existing code is "ground truth." As we can target arbitrary sections of code, we can identify functions where simulations spend most of their time (and thus maximize the benefit of replacement) and where individual calls execute on the order of seconds or milliseconds (thus providing us with an abundance of training data). The ML-derived models will not be an exact replacement, but should run two to three orders of magnitude faster than "ground truth" calculation. Intelligent runtime systems can track accumulated error and make dynamic, per-process decisions as to how to balance speed and accuracy. The result will be transformatively faster scenario exploration. This approach is also beneficial for the scenario with limited data learning, e.g., the extremes. The proposed approach is able to solve the convergence issue of numerical models which involve numerous variables in the spatiotemporal domain. The faster ML solver can generate adequacy of training data to reveal extremes in the physical domain with comprehensive uncertainty quantification analysis, which also enable the few-shot learning or zero-shot learning. Competition: There is a rich history of ML models replacing traditional compilation. In most of these approaches the model is required to come up with a highly-accurate solution to very general problems. By targeting our approach specifically at existing climate simulations we are constraining both the algorithms and the input parameters. Other approaches have not addressed the needs of highly iterative code, particularly predicting and tracking accumulated error. Our proposed runtime system will make the ML models actually useful for doing science.

Narrative

Creating machine learning (ML) models that exceed state-of-the-art computational approaches will be extraordinarily difficult, not the least because "ground truth" data describing the earth's climate is relatively sparse and expensive to acquire. We propose a parallel path: treat the existing models---and their representations in code---as ground truth and use ML to create computationally lightweight replacements that mimic the original to the degree required.

Once these replacements are in hand we can integrate them into existing simulation workflows. We will provide a runtime system that dynamically selects, at a per-process level, which replacements to use when. Exploratory work might prioritize speed over error, while more nuanced work could allow regions of lower interest to run faster on fewer nodes while focusing low-error computation on critical areas. We envision multiple replacements for each code region: some with tighter error bounds, some that generally generate low error but perform poorly in pathological cases, and some that may be nothing more than simple lookup tables for debugging runs.

We consider this approach as potentially transformative in water cycle and climate science because it explicitly treats numerical accuracy as a resource that can be traded away for faster performance. Because training data is cheap and plentiful---we will simply be rerunning the existing code with a vast range of different inputs---we expect the generated models to be of unusually high quality. We also maintain the explicability of the underlying models.

Of course, these benefits are not free. However, data collection is embarrassingly parallel and the computational effort to train the models occurs offline. For particular code under development we may require on the order of a week to create updated models, but we expect most hot paths in mature code to remain unchanged for months if not longer.

Suggested Partners/Experts (Optional)

Dr. Song at UNL has a demonstrated capability of developing deep learning and artificial neural networks for image analysis (ref) and reduced-order reactive-transport modeling (https://www.kbase.us/multiscale-microbial-dynamics-modeling/). Based on past experiences, he will closely work with the PI and team to design/test ML models for replacing computationally heavy modules and develop ideas of mitigating accumulation and propagation of errors through iterations.

Huiying Ren is a senior Data Scientist at Earth System Data Scientist team at Pacific Northwest National Laboratory (PNNL). She has been working on developing Artificial Intelligence (AI), big data analytical and extreme event analysis approaches which have been successfully applied on Earth, Energy and Environmental Systems. She is managing the data and supporting the data-model-integration for Multiscale ModEx task under the River Corridor Scientific Focus Area (SFA) project.

References (Optional)