Can ML-for-Science Research Be Trusted? A Crisis of Weak Baselines and Reporting Bias in ML-for-PDE Solving

Nick McGreivy[1], Ammar Hakim[2]

[1] Princeton University
[2] Princeton Plasma Physics Laboratory

ABSTRACT

One of the most promising applications of machine learning (ML) in the computational modeling of physical systems is in accelerating the solution of partial differential equations (PDEs). The key objective is for machine learned solvers to output a sufficiently accurate solution faster than traditional solvers. The baseline comparison used is thus critical for determining whether a machine learned solver has achieved its key objective. In this talk, I'll discuss the findings of an unpublished systematic review of the literature in this area. For every paper I can find that claims to outperform a standard numerical method using ML on a fluid-relevant PDE, I ignore the paper's intellectual contributions and ask two questions. First, did the paper perform a fair comparison with an efficient baseline? Second, did the paper display any evidence of reporting biases? As much as possible, I try to replicate the comparisons to ensure that my findings are fair and correct. There is a lot to discuss here, and I welcome any feedback or criticism. My main conclusion is that there are two issues — no, two crises — in the ML-for-PDE literature. First, papers in this area have systematically compared to weak baselines and/or inefficient numerical methods. Second, while reporting biases are difficult to prove, I find evidence that they are widespread. I believe these reporting biases are caused by the culture of ML, a gamified culture which perversely incentivizes the publication of results that seem more impressive while hiding or de-emphasizing results that seem less impressive. This systematic analysis raises some uncomfortable questions. Have machine learned solvers actually shown promise for accelerating the solution of PDEs, or have they simply compared to weak baselines and failed to report negative findings? More speculatively, since the essence of good science is the publication of results that are *true*, irrespective of whether they are impressive, is the culture of ML so corrupted by perverse incentives that ML-for-science research can't be trusted to report inconvenient truths?