

Why not submit to an ML conference?

The reason is that our proposition is not about a new neural network (NN) architecture. Our primary assertion is that an analog integrated circuit can implement an ODE-based NN (such as NODE or even a diffusion). Therefore, a circuit conference would be more appropriate. A friend reminded me of a related paper by Leon Chua from 1988, and I should revise the references accordingly. This also closes the academic loop as Leon Chua previously worked with the third author of our paper.

Can this runtime be reconfigured? Otherwise, isn't it fixed once fabricated? What if we want to switch models?

This is an excellent question. Ideally, both the weights and topology would be runtime reconfigurable. However, we believe this might be challenging, potentially requiring variable resistors, switches, etc. We have some basic prototyping ideas, but the effectiveness of these is yet to be verified. Revisiting an analog integrated circuit as a modern computational paradigm is certainly worthwhile, especially given the contemporary importance of ODE-based NNs. Imagine the impact of an ASIC for diffusion that could perform inference at unprecedented speeds, say 1 picosecond per image.

What exactly is the goal here? Is it to create hardware, or is it merely a simulation?

Ultimately, we aim to run this on an actual analog integrated circuit. Running it on a simulator would be pointless. Our desired outcome is infinitely fast inference (potentially achievable by adjusting the capacitors within the circuit). If we resort to using a simulator (consider SPICE, Spectre, or torchdiffeq as examples), it would still use the Euler method to discretize ODEs, offering no computational time advantage.

How does this differ from analog IMC?

Analog IMC and, broadly speaking, many analog-based DL accelerators focus on accelerating matrix multiplication. However, a neural network consists of multiple layers, each potentially involving several matrix multiplications. Traditionally, weights and features are loaded per digital clock cycle, with only the matrix multiplication sped up. Our idea differs by loading the inputs and weights once and allowing the analog circuit to process them in 1 picosecond, then directly reading the output.

You mentioned using an analog integrated circuit; how do you ensure accuracy given process variations?

Indeed, process variation will impact the performance, and the extent of this impact is uncertain. However, this challenge is applicable to all analog deep learning accelerators. We must address these challenges step by step; even if our explorations conclude that it does not work, the endeavor is still worthwhile. Personally, I believe that pursuing meaningful research is far more valuable than producing numerous insubstantial papers. In the field of electrical engineering, modern academic research focus on exploring new paradigms. Introducing a new paradigm is certainly not a bad thing, is it?