Hi all! I have enabled AI model inference in blockchain systems in both on-chain and off-chain approaches.

My code is still under development, I will consider making it open-sourced once it's in a more polished state. If you're interested in exploring this project or joining the effort, feel free to contact me. I would welcome any questions or discussion.

On-chain Approach

The smart contract execution environment in the existing blockchain systems lacks operators, instructions, and corresponding mechanisms to support complex DNN operations with high computational and memory complexity, which makes it inefficient or even infeasible to do the AI model inference on chain.

In order to enable on-chain AI model inference, I have extended the operation set in the EVM to support efficient DNN computation. Additionally, I have modified the Solidity compiler to allow for direct AI model inference calls in smart contracts. Currently, I have successfully run small AI model inferences such as a GAN model, which generated some impressive NFT art

However, this on-chain approach requires the modification of the EVM, rendering it incompatible with existing Ethereum systems. Therefore, I have shifted my focus to investigating an off-chain approach to address this issue.

Off-chain Approach

For off-chain AI model inference, I have adopted the optimistic rollup approach which is compatible with Ethereum and other blockchain systems that support smart contract execution.

To ensure the efficiency of AI model inference in the rollup VM, I have implemented a lightweight DNN library specifically designed for this purpose instead of relying on popular ML frameworks like Tensorflow or PyTorch. Additionally, I have provided a script that can convert Tensorflow and PyTorch models to this lightweight library. The cross-compilation technology has been applied to compile the AI model inference code into rollup VM code.

Performance

: I have tested a basic AI model (a DNN model for MNIST classification) on a PC. I was able to complete the DNN inference within 3 seconds in the rollup VM, and the entire challenge process can be completed within 2 minutes in a local Ethereum test environment.

Despite my unoptimized implementation, this level of performance seems to be acceptable for the current blockchain system. I plan to further optimize my implementation further to support larger and more complex models such as Stable Diffusion and GPT-2. Optimistically, I believe it will not take me too long to make it practical

Motivation

- Enabling AI model inference on the blockchain can allow for the creation of truly "smart" applications using smart contracts. For instance, embedding ChatGPT on the blockchain would provide the opportunity to develop fascinating metaverse applications on-chain.

- With an off-chain approach to AI model inference, users with available computing power can utilize their resources to complete the tasks of AI model inference and receive corresponding rewards. This can incentivize miners to use their computing power more efficiently, instead of engaging in PoW mining, which could be significant for miners in the previous PoW ETH.