Review of: "AERO: Softmax-Only LLMs for Efficient Private Inference"

Hanseok Ko[1]

1 Korea University, Korea, Republic of

Potential competing interests: No potential competing interests to declare.

The paper is well presented, it deals with a sufficiently important issue and proposes a novel method with good evidence that supports its effectiveness.

In light of the appearance of newer architectures such as test-time training (TTT), MAMBA, and relaxed recursive transformers (RRT), each of which aims in its own way to improve the efficiency and lower the cost of inference, it would have been nice to see some discussion of how the present work relates to the methods used by these architectures and why the vanilla transformer architecture was chosen as the subject of the authors' experiments over the newer architectures (or if the proposed methods could be applied to the newer architectures).

The paper references Meng et al. (2022) in support of the authors' assertion that later layers are less critical than earlier layers, which does not seem exactly correct. Meng et al. conclude that certain layers are responsible for storing facts, which does not preclude the possibility that other layers are equally important but have different roles (they may handle critical functions other than storing facts). The causal traces they use could perhaps have been used in the present paper to confirm which layers are indeed the most responsible for changes in entropy levels.

One suggestion I would like to make, perhaps as a follow-up paper, is instead of replacing all nonlinearities with softmax, to try replacing them with max-plus operations (i.e., look up tropical algebra), which is a generalization of ReLU and for which highly efficient calculation is also claimed. Building a ReLU-only or max-plus-only architecture may prove even more efficient than a softmax-only one.

And one final question I have for the authors is: it is not immediately apparent to me that the methods you propose are exclusive to private inference--would there be any issues with using these methods to improve the efficiency of models for regular inference? And if not, then perhaps it could help to expand the scope of the paper and add experiments with just regular inference as well.