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

Nian Xue[1]

1 New York University, United States

Potential competing interests: No potential competing interests to declare.

Summary:

This paper proposes a framework for efficient private inference using Softmax-only large language models (LLMs), named AERO. The main contribution of the work lies in reducing computational complexity and enhancing privacy preservation during inference. The authors introduce a novel entropy regularization technique to improve the performance of the Softmax-only model. They claim that their method offers both efficiency and privacy benefits without the substantial computational overhead typically required by traditional privacy-preserving techniques. Extensive experiments were conducted to validate the effectiveness of the proposed approach, which could have significant implications for secure machine learning applications, especially in environments handling sensitive data.

Paper Strength:

1. The paper introduces a technique for efficient private inference by systematically removing nonlinearities, a clever design decision that reduces the FLOPs counts.

2. The method's focus on improving inference efficiency in privacy-preserving machine learning is highly relevant in the current landscape, where privacy concerns are growing in importance across industries such as healthcare, finance, and government. The paper is well-organized.

3. The authors provide solid experimental results showing the efficacy of their method in comparison to existing methods. The experiments demonstrate clear improvements in both performance and a decrement in latency.

4. The paper is well-supported by solid theoretical analysis, providing a strong justification for the design choices made in the AERO framework.

Paper Weakness:

1. While the experimental evaluation is comprehensive, it mainly focuses on controlled datasets. The method's performance in real-world, noisy, and complex environments is not well explored.

2. The evaluation metric is relatively limited, only considering perplexity.

3. The models in the experiment are relatively old, e.g., GPT-2, 2019 Pythia-70M 2022, and not particularly ‘’large“ in scale.

4. Although the paper mentions private inference, it mainly focuses on model design and efficiency, and ML-related experiments, with limited discussion on the proposed framework related to security, i.e., private inference.

Questions:

1. How does AERO perform when applied to some of the larger and latest models, such as LLaMA 3 or PaLM 2, or larger datasets? Does the efficiency benefit hold when the model size or data complexity increases significantly?

2. Are these technologies suitable for general inference tasks?

3. What are the differences between training and inference with plain inputs and encrypted inputs using LLMs?

4. Are you planning to evaluate using other metrics, such as BLEU or ROUGE?

5. The paper mentions "address the overheads associated with non-linear operations in PI." Does this also help speed up the efficiency in the pre-training process?