Ultra Memory-Efficient Tensor-Compressed Transformer Training Methods on Edge Devices

Tech ID: 34225 / UC Case 2025-347-0

BACKGROUND

On-chip training of transformer models has gained significant importance in recent years, due to the growing need for low-latency, energy-efficient and privacy-preserving AI systems. Training machine learning models end-to-end or incrementally, based on local and private data, can reduce dependency on cloud-based resources, enhance data privacy, and allow for real-time adaptability, which makes this training model crucial for applications including internet of things, autonomous systems, and personalized AI. On-device training has been increasingly utilized in sensing, communication, data processing and decision making. However, training modern neural networks remains difficult because it requires large amounts of memory and computing power, which are not available on resource-constrained edge devices. Because of these limitations most proposed solutions focus on-device fine-tuning or transfer learning.

DESCRIPTION

Researchers at UC Santa Barbara have developed an innovative approach that combines an ultra-memory-efficient algorithm with a hardware design to facilitate the training of transformer models on resource-constrained platforms such as FPGAs. By employing a tensor-train compressed format, the technology achieves more than 100 times reduction in model sizes, which translates to a 30X to 50X overall memory reduction on FPGA devices. It introduces a bidirectional and optimized tensor contraction method that significantly lowers the computing flops required for forward and backward propagations. This method also optimizes memory and latency performance by maximizing on-chip memory usage and minimizing communication between on-chip and off-chip memory.

ADVANTAGES

- Enables the training of larger AI models on edge devices with ~100X model size reduction
- Achieves 30X to 50X overall memory reduction on FPGA
- Reduces computing flops significantly for both forward and backward propagations
- Improves memory and latency performance by optimizing on-chip memory usage
- Facilitates better energy cost per epoch than conventional GPUs

APPLICATIONS

- Artificial intelligence
- Sensing
- Communication

CONTACT

Pasquale S. Ferrari
ferrari@tia.ucsb.edu
tel: .

INVENTORS

- Tian, Jiayi
- Zhang, Zheng

OTHER INFORMATION

KEYWORDS

AI, artificial intelligence, communication, data processing, image analysis, medical data analysis, transformer models, FPGA, training

CATEGORIZED AS

- **Computer**
  - Hardware
  - Other

RELATED CASES

2025-347-0

- Data processing
- Image analysis
- Medical data analysis

PATENT STATUS

Patent Pending

ADDITIONAL TECHNOLOGIES BY THESE INVENTORS

- MR-Based Electrical Property Reconstruction Using Physics-Informed Neural Networks

University of California, Santa Barbara
Office of Technology & Industry Alliances
342 Lagoon Road, ,Santa Barbara,CA 93106-2055 |
https://www.tia.ucsb.edu
Tel: 805-893-2073 | Fax: 805.893.5236 | padilla@tia.ucsb.edu