MCNC: Manifold Constrained Network Compression

Tech ID: 33976 / UC Case 2025-419-0

ABSTRACT

Researchers at Vanderbilt University and the University of California, Davis have developed MCNC software that significantly compresses large AI models while maintaining their performance using a novel manifold-constrained optimization approach.

FULL DESCRIPTION

The Manifold-Constrained Model Compression (MCNC) software introduces an approach to compress foundational AI models like Vision Transformers and large language models such as GPT and LLaMA. Unlike traditional methods, MCNC uses manifold-constrained optimization to achieve over 100 times compression, enhancing storage, communication, and customization efficiency without compromising model performance.

APPLICATIONS

- Efficient deployment of complex AI models on smartphones, IoT devices, and other hardware with constrained resources.
- Seamless model customization and fine-tuning for personalized AI applications.
- Reduction in communication overhead for cloud-based AI services and platforms

FEATURES/BENEFITS

- Achieves high compression rates of over 100x without significantly compromising performance.
- Enhances efficiency in storage and communication, making it ideal for edge devices with limited memory.
- Enables high throughput, facilitating rapid model construction.
- Offers flexibility and compatibility with various AI models and other compression techniques.
- Overcomes challenges in storing and transmitting large-scale AI models due to their massive size.
- Addresses the limitations of edge devices in running sophisticated AI models by managing memory and processing constraints.
- Resolves memory bottlenecks encountered during model customization and fine-tuning.
- Reduces high data transfer costs and latency in model communication.

PATENT STATUS

Patent Pending

CONTACT

Byron N. Roberts
bnroberts@ucdavis.edu
tel: 530-754-8689.

INTRODUCING
UC TechAlerts
New technology matches delivered to your email at your preferred schedule
SEARCH
SAVE SEARCH
Learn More

INVENTORS

- Abbasi Koohpayegani, Soroush
- Nooralinejad Eslamlo, Parsa
- Pirsiavash, Hamed

OTHER INFORMATION

KEYWORDS

AI models, compression, edge devices, efficiency, manifold-constrained optimization, model customization, storage reduction, technology integration, throughput rate, transmission efficiency

CATEGORIZED AS

- **Computer**
  - Software
- **Engineering**

- ▶ Robotics and Automation

RELATED CASES

2025-419-0

University of California, Davis
Technology Transfer Office
1 Shields Avenue, Mrak Hall 4th Floor,
Davis,CA 95616

Tel:
530.754.8649
techtransfer@ucdavis.edu
https://research.ucdavis.edu/technology-transfer/
Fax:
530.754.7620