

TRANG THÔNG TIN LUẬN ÁN

(khoảng 1 – 1.5 trang A4)

Tên đề tài luận án: Nghiên cứu phân rã tensor và các ứng dụng cho bài toán phục hồi và nén dữ liệu

Ngành: Khoa học máy tính

Mã số ngành: 9480101

Họ tên nghiên cứu sinh: Phạm Minh Hoàng

Khóa đào tạo: 2019

Người hướng dẫn khoa học: GS.TS. Hoàng Dương Tuấn, TS. Trần Thái Sơn

Cơ sở đào tạo: Trường Đại học Khoa học Tự nhiên, ĐHQG.HCM

1. TÓM TẮT NỘI DUNG LUẬN ÁN:

Dữ liệu thực tế ngày càng đa dạng và có cấu trúc đa chiều. Trong khi các phương pháp phân tích dữ liệu truyền thống như PCA hay LDA chủ yếu làm việc với dữ liệu ma trận hay vector thì việc chuyển đổi dữ liệu đa chiều sang dạng vector hay ma trận thường làm mất thông tin cấu trúc quan trọng. Tensor, với khả năng biểu diễn dữ liệu đa chiều một cách tự nhiên, đã trở thành công cụ phù hợp cho nhiều bài toán phân tích dữ liệu. Các phương pháp phân rã tensor như CP, Tucker, Tensor-Train (TT) cho phép biểu diễn dữ liệu đa chiều hiệu quả, hỗ trợ mạnh mẽ cho các tác vụ như phục hồi ảnh, video, và giảm tham số mô hình học sâu.

Luận án "Nghiên cứu phân rã tensor và các ứng dụng cho bài toán phục hồi và nén dữ liệu" tập trung vào hai hướng chính. Thứ nhất, đối với bài toán phục hồi dữ liệu, luận án giới thiệu kỹ thuật tăng cường tensor, biểu diễn tensor bậc thấp thành tensor bậc cao để khai thác tương quan đa chiều tốt hơn, tăng hiệu quả phục hồi khi kết hợp với thuật toán hoàn tất tensor, đồng thời giảm chi phí lưu trữ. Ngoài ra, luận án đề xuất cải tiến thuật toán hoàn tất tensor bằng chuẩn ℓ_q để giải quyết tính thưa và tính mất cân bằng trong bài toán hoàn tất ma trận và phương pháp không dùng SVD nhằm giảm chi phí tính toán nhưng vẫn đảm bảo độ chính xác. Thứ hai, đối với bài toán giảm tham số mạng học sâu, luận án áp dụng phân rã Tucker và TT để biểu diễn lại tham số của mô hình thành các tensor thành phần với kích thước nhỏ hơn, từ đó giảm số tham số huấn luyện mà vẫn duy trì hiệu quả của mô hình.

2. NHỮNG KẾT QUẢ MỚI CỦA LUẬN ÁN:

Những đóng góp chính của luận án:

- Đề xuất kỹ thuật tăng cường tensor linh hoạt (TA) biểu diễn tensor bậc thấp sang bậc cao nhằm tận dụng tốt hơn mối quan hệ đa chiều, giảm chi phí lưu trữ và cải thiện chất lượng phục hồi dữ liệu.
- Các thuật toán hoàn tất tensor cải tiến dựa trên chuẩn ℓ_q để giải quyết tính thưa và tính mất cân bằng trong bài toán hoàn tất ma trận và thuật toán không sử dụng SVD giúp giảm chi phí tính toán.
- Xây dựng mô hình TT-ViT, ứng dụng phân rã TT để giảm tham số học mạng Vision Transformer thông qua việc thay thế ma trận trọng số bằng các nhân TT nhỏ gọn.
- Đề xuất LoRAT, mở rộng phương pháp LoRA sang miền tensor bằng phân rã Tucker và TT để giảm tham số tinh chỉnh trong các lớp tích chập mà vẫn bảo toàn hiệu năng mô hình.

3. CÁC ỨNG DỤNG/ KHẢ NĂNG ỨNG DỤNG TRONG THỰC TIỄN HAY NHỮNG VẤN ĐỀ CÒN BỎ NGỜ CẦN TIẾP TỤC NGHIÊN CỨU

Các kết quả đạt được trong luận án không chỉ mang ý nghĩa lý thuyết mà còn có tiềm năng ứng dụng cao trong nhiều lĩnh vực. Trong xử lý ảnh và video, các phương pháp tăng cường và hoàn tất tensor có thể hỗ trợ phục hồi dữ liệu trong các hệ thống camera giám sát, chẩn đoán y khoa hình ảnh, cũng

như các dịch vụ truyền phát đa phương tiện có dữ liệu bị suy giảm. Đối với hệ thống gợi ý, kỹ thuật hoàn tất tensor cho ma trận điểm giúp cải thiện chất lượng dự đoán xu hướng người dùng. Trong lĩnh vực mô hình học sâu, các phương pháp giảm tham số như TT-ViT và LoRAT mở ra khả năng triển khai mô hình trên các thiết bị tài nguyên hạn chế như thiết bị di động, IoT hoặc hệ thống nhúng mà vẫn đảm bảo hiệu năng.

Bên cạnh những kết quả đạt được, vẫn còn nhiều vấn đề cần nghiên cứu sâu hơn. Một hướng quan trọng là mở rộng các thuật toán phục hồi dữ liệu tensor cho những bộ dữ liệu kích thước lớn và phức tạp hơn, đồng thời kết hợp với các mô hình học sâu nhằm tận dụng khả năng học biểu diễn mạnh mẽ. Đối với giảm tham số mô hình, việc phát triển cơ chế phân rã tensor thích ứng theo cấu trúc từng lớp mạng, cũng như nghiên cứu cách tích hợp các kỹ thuật nén với chiến lược huấn luyện hiện đại để giảm thời gian suy luận và giảm tiêu thụ bộ nhớ, là những hướng nghiên cứu tiềm năng. Ngoài ra, việc mở rộng ứng dụng của các phương pháp trên trong các hệ thống thị giác máy tính, xử lý ngôn ngữ tự nhiên, y sinh và phân tích tín hiệu sẽ góp phần khẳng định vai trò của phân rã tensor trong các bài toán thực tiễn.

THESIS INFORMATION

(1 – 1.5 A4 pages)

Thesis title: Tensor Decomposition and Its Applications to Data Recovery and Compression

Speciality: Computer Science

Code: 9480101

Name of PhD Student: Pham Minh Hoang

Academic year: 2019

Supervisor: Prof. Dr Hoang Duong Tuan, Dr. Tran Thai Son

At: VNUHCM - University of Science

1. SUMMARY:

Real-world data has become increasingly diverse and multi-dimensional in structure, extending beyond traditional vector or matrix formats. While traditional methods such as PCA and LDA primarily work on matrix data, converting multi-dimensional data into vectors or matrices often results in loss of structural information. Tensors, with their ability to represent multidimensional data naturally, have become a suitable tool for many multidimensional data analysis problems. Tensor decomposition methods such as CP, Tucker, and Tensor-Train (TT) allow for efficient representation of multidimensional data, providing strong support for tasks such as image and video recovery, as well as parameter reduction in deep learning models.

The thesis titled "Tensor Decomposition and Its Applications to Data Recovery and Compression" focuses on two main directions. First, regarding the data recovery problem, the dissertation introduces tensor augmentation techniques—representing low-order tensors as high-order tensors—to better exploit multidimensional correlations, increasing recovery efficiency when combined with tensor completion algorithms while simultaneously reducing storage costs. Additionally, the dissertation proposes improvements to tensor completion algorithms using the l_1 to address sparsity and imbalance in matrix completion problems, alongside SF-LRTC methods to reduce computational costs while maintaining accuracy. Second, regarding deep learning parameter reduction, the dissertation applies Tucker and TT decompositions to represent model parameters into smaller component tensors, thereby reducing the number of trainable parameters while maintaining model effectiveness.

2. NOVELTY OF THESIS:

The main contributions of the thesis:

- Proposes a flexible tensor augmentation (TA) technique to convert low-order tensors into higher-order forms, improving storage efficiency and data recovery performance.
- Improves tensor completion algorithms based on l_1 norm for handling sparsity and imbalance in matrix completion, and introduces SF-LRTC algorithms without SVD for optimizing computational complexity
- Develops the TT-ViT, applying TT decomposition to reduce Vision Transformer parameters by replacing weight matrices with compact TT-cores
- Proposes LoRAT, extending LoRA to the tensor domain using Tucker and TT decompositions to reduce fine-tuning parameters in CNNs while preserving model performance.

3. APPLICATIONS/ APPLICABILITY/ PERSPECTIVE

The results achieved in this dissertation offer both theoretical significance and high potential for practical application on various fields. In image and video processing, tensor augmentation techniques and tensor completion algorithms can recover data in surveillance systems, medical imaging diagnostics, and multimedia streaming services where data may be degraded. For recommendation systems, tensor completion techniques for rating matrices improve prediction quality in e-commerce and entertainment platforms. For deep learning model, parameter reduction methods like TT-ViT and LoRAT enable the deployment of models on resource-constrained devices such as mobile devices, IoT, or embedded systems without compromising performance.

Despite the promising results, several challenges and research directions remain. One important direction is extending tensor data recovery algorithms to larger and more complex datasets while integrating them with deep learning models to leverage powerful representation learning capabilities. Regarding model parameter reduction, developing adaptive tensor decomposition mechanisms based on the structure of each network layer, as well as combining compression techniques with modern training strategies to reduce inference time and memory consumption. Furthermore, expanding the application of these methods in computer vision, natural language processing, biomedicine, and signal analysis will further demonstrate the importance of tensor decomposition in real-world problems.