Poster Abstracts from Workshop on Sparse Tensor Computations

18-19 October 2023

Title: Implicit low-rank integrators for solving time-dependent problems
Presenter: Joseph Nakao, Swarthmore College, jnakao1@swarthmore.edu

Abstract: Dynamical low-rank (DLR) methods have become a popular way to solve time-dependent problems that have low-rank structures. Some of the equations of interest contain stiff operators that require implicit time integrators for reasonable computational efficiency. However, achieving high-order accuracy using implicit time integrators in the low-rank framework remains a challenge. We propose a new low-rank method, similar in spirit to the DLR framework, that incorporates high-order implicit time integrators.

Title: A Performance Portability Study Using Tensor Contraction Benchmarks
Presenter: Muhammed Emin Ozturk, University of Utah, emin.ozturk@utah.edu

Abstract: Driven by the end of Moore's law, heterogeneous architectures, particularly GPUs, are experiencing a surge in demand and utilization. While these platforms hold the potential for achieving high performance, their programming remains challenging and requires extensive hardware knowledge. This complexity is further exacerbated by the different proprietary languages utilized by various vendors. In this paper, we conduct a performanceportability study on two portable languages, SYCL and Kokkos. Specifically, we focus on the case study of tensor contractions and employ COGENT, a DSL compiler for tensor contractions, to generate CUDA code for the 48 different tensor contractions in the TCCG benchmark suite. We extend COGENT to produce Kokkos code, and use Hipify and SycloMatic, which are tools that convert CUDA code to HIP and SYCL. Our analysis involves a comparison of the performance of each framework on both Nvidia and AMD GPUs. Our experiments show that identically tiled tensor contraction kernels in Kokkos and SYCL can exhibit significant performance differences compared to the corresponding CUDA/HIP program, respectively on Nvidia/AMD GPUs. The main reason for the performance differences arise from differences in register usage and the management of register spills to threadprivate stack memory, affecting overall degree of threadlevel concurrency and the volume of data movement to/from GPU DRAM.

Title: FedThumb: A LBP and HOG Feature Based Fingerprint Presentation Attack Detection Using Hybrid VGG16-SVM in Federated Learning

Presenter: S M Sarwar, University of Texas Rio Grande Valley, smsarwar96@gmail.com

Abstract: Over recent years, biometric authentication systems have gained widespread acceptance for personal (mobile devices, access control systems), national (law enforcement, voter registration), and global security (visa applications, passport control). Fingerprints have become recognized as a popular biometric trait, alongside other traits such as the iris, face, retina, voice, signature, etc., because of their uniqueness, stability, convenience (touch or swipe), and cost-effectiveness compared to other biometric modalities. However, the evolution of fake biometrics such as gelatine, silicon, woodglue, and latex has brought new challenges. As a result of these modifications, several fingerprint spoofing detection methods have been proposed to distinguish between fake and spoof fingerprints. Machine learning (ML) and deep learning (DL) techniques are widely used as learning models for fingerprint spoofing or presentation attack problems, but federated learning (FL) can play a significant role in this problem due to its collaborative learning method and privacy-preserving advantages. This paper proposes a support vector machine (SVM) based federated learning algorithm that uses local binary patterns (LBP) and histogram of oriented gradients (HOG) features of images. In this experiment, we adopted a public dataset, the Fingerprint Liveness Detection Competition (LivDet 2015), to compare whether the proposed model enhanced the detection of fingerprint spoofing with other state-of-the-art methods.

Title: Tensor Bhattacharya-Mesner (BM) Decomposition and Applications

Presenter: Fan Tian, Tufts University, Fan.Tian@tufts.edu

Abstract: We introduce a third-order tensor decomposition framework based on a ternary multiplication named Bhattacharya-Mesner (BM) product and its corresponding notion of rank. We describe an iterative algorithm for computing a low BM-rank approximation to a given third-order data array. Moreover, we will demonstrate our decomposition framework for video processing applications. Our method appears to improve on other matrix and tensor-based methods with smaller approximation error for the same level of compression.

Title: Geometric Optimization for Tensor Completion

Presenter: Navjot Singh, University of Illinois at Urbana-Champaign, navjot2@illinois.edu

Abstract: This poster explores geometric optimization methods for tensor completion, which have gained popularity due to their efficient formulation over fixed rank and orthogonality constraints. We compare the accuracy and scalability of these algorithms and propose techniques to further enhance their performance. Our research aims to address practical tensor completion challenges and unlock new possibilities for various applications.

Title: Accelerating the Computation of Tensor Z-eigenvalues
Presenter: Rhea Shroff, University of Florida, rhea.shroff@ufl.edu

Abstract: Efficient solvers for tensor eigenvalue problems are important tools for the analysis of higher-order data sets. Here we introduce, analyze and demonstrate an extrapolation method to accelerate the widely used shifted symmetric higher order power method for tensor Z-eigenvalue problems. We analyze the asymptotic convergence of the method, determining the range of extrapolation parameters that induce acceleration, as well as the parameter that gives the optimal convergence rate. We then introduce an automated method to dynamically approximate the optimal parameter, and demonstrate it's efficiency when the base iteration is run with either static or adaptively set shifts. Our numerical results on both even and odd order tensors demonstrate the theory and show we achieve our theoretically predicted acceleration.

Title: Auto-Scheduling Sparse Tensor Contractions with Kernel Fusion
Presenter: Adhitha Dias, Purdue University, kadhitha@purdue.edu

Abstract: Automated code generation and performance optimizations for sparse tensor algebra kernels are important as these computations are used in many real-world applications such as machine learning and scientific simulations. In certain cases, kernel/loop fused sparse tensor contractions could be better than their non-fused counterparts due to asymptotic effects in complexity and reduced memory footprint in the way the computation is performed and add many new schedules to the space of schedules for a given tensor computation. This work focuses on generating fused schedules and an auto-scheduler that can evaluate both the fused and non-fused schedules to filter the schedules depending on the computing complexity and the memory usage, using user-provided constraints and an SMT solver.

Title: Application Performance Modeling via Tensor Completion
Presenter: Edward Hutter, University of Illinois at Urbana-Champaign, hutter2@illinois.edu

Abstract: Performance tuning, software/hardware co-design, and job scheduling are among the many tasks that rely on models to predict application performance. We propose and evaluate low rank tensor decomposition for modeling application performance. We discretize the input and configuration domain of an application using regular grids. Application execution times mapped within grid-cells are averaged and represented by tensor elements. We show that low-rank canonical-polyadic (CP) tensor decomposition is effective in approximating these tensors. We further show that this decomposition enables accurate extrapolation of unobserved regions of an application's parameter space. We then employ tensor completion to optimize a CP decomposition given a sparse set of observed runtimes. We consider alternative piecewise/grid-based models and supervised learning models for six applications and demonstrate that CP decomposition optimized using tensor completion offers higher prediction accuracy and memory-efficiency for high-dimensional applications.

Title: Tensor Butterfly Factorization (In Parallel!)
Presenter: Michael Kielstra, UC Berkeley, pmkielstra@berkeley.edu

Abstract: First developed in the context of the Fast Fourier Transform, butterfly factorizations have found applications to the many other matrices that satisfy the complementary low-rank property (CLR). We extend butterfly methods to tensors in four or more dimensions that satisfy an analogous property, achieving both lower compression times and better memory use compared to the usual method of flattening the tensor into a matrix and butterfly-factoring that. The process also suggests a somewhat novel understanding of butterfly factorization itself, allowing us to write algorithms that can be described extremely tersely and parallelize well.

Title: Sampling Methods for the Canonical Polyadic Decomposition
Presenter: Carmeliza Navasca, University of Alabama at Birmingham, cnavasca@uab.edu

Abstract: The Alternating Least-Squares (ALS) is one of the most well known method for approximating factor matrices in the canonical polyadic decomposition. ALS is fast, but it has some drawbacks. In this talk, we address one of them. We describe some sampling methods (non-adaptive and adaptive) for ALS to improve speed, accuracy and memory space for large tensors. Numerical results will be provided.

Title: Deblurring and Denoising Using Tensorial Total Variation for Tensor Imaging
Presenter: Fatoumata Sanogo, Bates College, fsanogo@bates.edu

Abstract: We consider denoising and deblurring problems for tensor structured data, namely, color images and videos. While images can be discretized as matrices, the analogous procedure for color images and videos leads to a tensor formulation. We extend the classical Rudin-Osher-Fatemi functional for variational denoising and deblurring to tensors through multi-dimensional total variation regularization. The generalization results in a minimization problem that is calculated by the fast iterative shrinkage-thresholding method for tensors. In addition, using the properties of a circulant structure of a tensor allows for an efficient implementation. We provide several numerical experiments by applying the method to the denoising and deblurring of color images and videos.

Title: Quantum Circuit Simulation using ZX-calculus and Tensor Network Contraction
Presenter: Aniruddha Sen, UMass Amherst, asen1@illinois.edu

Abstract: Quantum circuits can be simulated by a classical computer in exponential time. Clifford circuits form a class of quantum circuits that can be simulated efficiently in polynomial time. We propose a matrix formalism for simulating a Clifford circuit through Gaussian elimination and operations according to the rules of ZX-Calculus. This provides an alternative to the conventional tableau algorithm. We extend our approach to simulating a general circuit through tensor network contractions on a simplified representation of the circuit. We also relate our simulation algorithm to the computation procedure of a measurement based quantum computer.

Title: Efficient Incremental Tucker Decomposition for Streaming Scientific Data
Presenter: Saibal De, Sandia National Laboratories, sde@sandia.gov

Abstract: The Tucker decomposition has emerged as a popular format for compressing large datasets generated from high-fidelity scientific simulations. Several software packages (Tensor Toolbox, TuckerMPI) enable computing the Tucker decomposition of static data, but relatively fewer works address compressing a streaming tensor. In this work, we develop a streaming Tucker algorithm, tailored to scientific simulations where the data tensor grows along a single time-like dimension. At any point, we seek to update the existing factorization – the Tucker core and the factor matrices – with a new tensor slice, without accessing the already incorporated tensor slices in their raw, uncompressed form. Throughout this process, we ensure that a user-specified relative error tolerance is met. We present an implementation within the TuckerMPI framework, and apply it to both synthetic and simulated (combustion) datasets. By comparing against the standard (batch) algorithm, we show that our proposed approach provides significant improvements in terms of memory usage. If the Tucker ranks stop growing along the streaming tensor mode, our approach also incurs less wall-time compared to the batch version.

Title: Using the Tucker Decomposition to Solve Image Deblurring Problems
Presenter: Tyler Fuller, Arizona State University, tjfuller@asu.edu

Abstract: Block-structured matrices naturally occur in image processing applications. Recent work by Kilmer and Saibaba introduced methods to map block-structured matrices to 3-way tensors and back again. Their work shows that applying tensor decompositions, such as CP and Tucker, before mapping back to a matrix reveals additional Kronecker structure in the original matrix. This poster discusses applying these methods to image deblurring, with Tucker as the decomposition of choice.

Title: Efficient Low Rank Parametric Kernel Matrix Approximation Using the Tensor Train Decomposition
Presenter: Abraham Khan, North Carolina State University, awkhan3@ncsu.edu

Abstract: In the realm of applied mathematics, the ubiquity of kernel matrices is readily apparent. However, these kernel matrices are quite large, and explicitly forming them invokes a high computational cost and storage cost, particularly in higher dimensional spaces. Often these kernel matrices depend on hyper parameters that must be tuned; for example, in the context of gaussian processes. Therefore, we introduce a black box method that computes a low rank approximation of a kernel matrix assuming that the source and target points are well separated. In particular, the Tensor Train (TT) decomposition is used in conjunction with Chebyshev polynomials in order to obtain such a low rank approximation that can be recomputed efficiently for new hyper parameters during an online phase. Our approach is novel because the TT decomposition allows us to deal with kernel matrices in the context of higher dimensional spatial and parameter spaces.

Title: Searching for Cyclic-Invariant Fast Matrix Multiplication Algorithms
Presenter: J Pinheiro, Wake Forest University, deolj19@wfu.edu

Abstract: Fast matrix multiplication algorithms correspond to exact CP decompositions of tensors that encode matrix multiplication of fixed dimensions. This 3-way matrix multiplication tensor M has cyclic symmetry: the entry values are invariant under cyclic permutation of the indices. The CP decomposition of Strassen's original fast matrix multiplication algorithm for 2x2 matrices is cyclic invariant, and cyclic invariant decompositions are known to exist for 3x3 and 4x4 matrix multiplication as well. Cyclic invariance means a cyclic permutation of the CP factors results in the same CP components, just in a different order. We describe how to search for these solutions, which involve one third of the variables of generic solutions, using the damped Gauss-Newton optimization method along with heuristic rounding techniques, and we summarize the algorithms discovered so far.

Title: Improving Numerical Stability within Tensor Decomposition Algorithms

Presenter: Grey Ballard, Wake Forest University, ballard@wfu.edu

Abstract: Standard algorithms for computing CP and Tucker tensor decompositions typically prioritize efficiency of solving the underlying linear algebra problems over the numerical stability. In particular, the normal equations are used to solve linear least squares problems within the Alternating Least Squares algorithm for CP, and the Singular Value Decomposition (SVD) is often computed via an eigendecomposition of the associated Gram matrix within the Sequentially Truncated Higher-Order SVD algorithm for Tucker. We demonstrate how more numerically stable approaches for solving linear least squares problems and computing the SVD can (1) improve the accuracy of the ultimate tensor approximations and/or (2) improve the running time of the overall decomposition algorithm by enabling the use of lower working precision.