

Speeding Up Matrix Multiplication with Machine Learning

Manan Bhatia

March 13, 2025

1 Introduction

The basic operation of matrix multiplication serves as required in various areas including machine learning and computer science together with computer graphics and scientific simulations. The standard approach to matrix multiplication enables execution at a cubic speed $O(N^3)$ when the square matrices have dimensions N . The algorithm optimisation requires improvement for large-scale systems because matrix multiplications serve as vital components of neural networks and transformer architectures in deep learning applications.

More efficient algorithms, such as Strassen’s algorithm, demonstrates that the potential faster algorithms are possible, including the Copper-Winograd algorithm. However, the research for continuous faster algorithms has only been done by manual process, constrained by human intuition and vast combinatorial nature of practical solutions.

Through research presented in “*Discovering Faster Matrix Multiplication*” it is found out that reinforcement learning technique serves as a method to extend previous work through accelerated mathematical exploration of matrix multiplication. The reinforcement learning agent AlphaZero serves as a basis for AlphaTensor to determine optimal tensor decompositions in finite factor spaces by playing a single-player game. The automated approach detected both standard and new algorithms which proved better than traditional methods for smaller matrix dimensions.

The research expands previous work by studying matrix multiplication mathematics while implementing AlphaZero-inspired reinforcement learning models to evaluate their achieved performance.

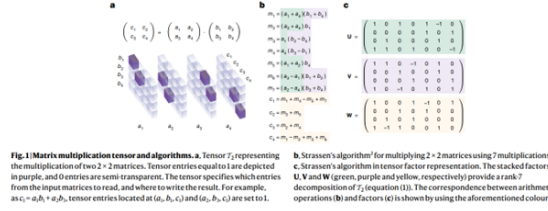


Figure 1: Tensor representation of matrix multiplication as a rank-3 tensor. Lower-rank decompositions reduce scalar multiplications. (Source: AlphaTensor paper)

2 Literature Review

The search for enhanced matrix multiplication algorithms has proven challenging throughout many years in fields of mathematics and computer science. Research studies found ways to decrease the N^3 scalar multiplication requirement for matrix multiplication by exploiting mathematical structures in the algorithm.

The foundation for improving matrix multiplication processes came from classical methods. The time complexity reached approximately $O(N^{2.81...})$ through Strassen’s algorithm by converting two 2×2 matrix multiplication into 7 recursive subproblems rather than 8. Coppersmith-Winograd algorithm developed this methodology to achieve matrix multiplication complexity at $O(N^{2.376})$. The state-of-the-art matrix multiplication algorithms currently reach an $O(N^{2.37})$ complexity level but actual practical applications tend to use Strassen’s method as an economical alternative for medium-scale matrices.

The process of understanding matrix multiplication uses tensor decomposition as an essential mathematical tool. The matrix multiplication operation can be understood as a rank-3 tensor but requires finding the most compact rank representation. The conversion to lower-rank expression requires fewer multiplication operations across scalars. The matrix multiplication tensor of size $2 \times 2 \times 2$ finds its rank-7 decomposition through Strassen’s algorithm. Through tensor decomposition researchers gain a unified perspective for both existing algorithm analysis and new algorithm discovery hence becoming an optimal method for optimisation research.

The application of reinforcement learning methods for mathematical discovery opens new possibilities in research. AlphaZero proved that software which links reinforcement learning to self-play with Monte Carlo Tree Search routing could reach superhuman levels in complicated games of Go and Chess. The article modified this framework to develop algorithm discovery along with AlphaTensor which is a reinforcement learning agent that understands tensor factorisation of matrix multiplication. The AlphaTensor system accomplished both the rediscovery of Strassen’s algorithm and the identification of newly discovered algorithms that demonstrated better performance than traditional matrix

multiplication methods when applied to small matrix sizes. The study demonstrated how RL stands as a forceful method for symbolically understanding matters and creating new algorithms.

The proposed synthesis of classical algorithms and tensor decomposition theory and reinforcement learning serves to expand scientific knowledge in matrix multiplication optimisation. The research will adapt AlphaTensor’s reinforcement learning approach to evaluate discovered algorithms for their practical value in transformer computations which require heavy use of matrices.

3 Methodology

This project uses a defined procedure which includes the exploration and implementation and evaluation of machine learning algorithms to boost matrix multiplication speed. An approach with three main elements will lead the project which includes creating the mathematical foundation and designing the reinforcement learning system and evaluating the discovered algorithms.

This project bases its foundation on comprehension of the mathematical concepts that govern matrix multiplication. The concept of tensor decomposition will be studied because it shows how matrix multiplication operates as a rank-3 tensor. Lower-rank decomposition of this tensor stands as the goal because it leads to fewer necessary scalar multiplications. The evaluation of new algorithms will use existing methods and tensor rank analysis and bilinear complexity comparison with Strassen’s and Coppersmith-Winograd algorithms.

The present work adopts an AlphaZero framework to create a reinforcement learning system shaped after the approach from the article. In this setup the RL agent views matrix multiplication as a game that defines states through partial decomposition steps and employs actions through choices of matrix sub-block combination. MCTS explores while the neural network evaluates policies for the agent which improves its strategy through self-play during multiple iterations. The reward signals will be established through scalar multiplication counts to direct algorithm discovery toward less complex methods. A training process will operate on basic matrix dimensions 3×3 to achieve successful algorithm discovery despite computational restrictions.

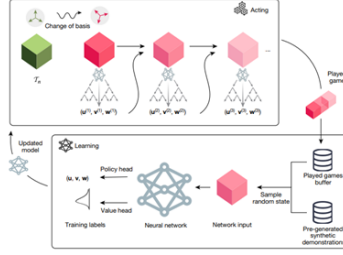


Fig. 2: Overview of AlphaTensor. The neural network (bottom box) takes an input tensor T_n and outputs samples $\pi(a|s)$ and $v(s)$ from a distribution over potential next actions to play, and an estimate of the future returns (for example, of -Rank (5)). The network is trained on two data sources: previously played games and synthetic demonstrations. The updated network is sent to the actor (top box), where it is used by the MCTS planner to generate new games.

Figure 2: AlphaTensor’s reinforcement learning framework for algorithm discovery. (Source: AlphaTensor paper)

The research will analyse discovered algorithm performance through evaluations with traditional algorithms that include the classical cubic algorithm and Strassen’s algorithm. The evaluation framework uses the count for scalar multiplication operations together with runtime performance measurements coupled with scaling abilities for bigger matrix sizes. A simplified transformer model will receive these discovered algorithms to determine their effect on the runtime of self-attention computations.

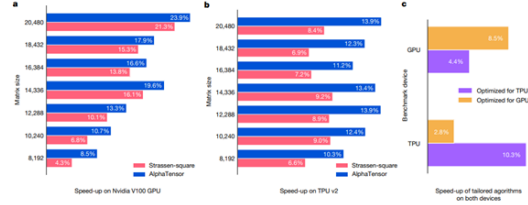


Fig. 3: Speed ups of the AlphaTensor-discovered algorithm. a, b. Speed ups (%) of the AlphaTensor-discovered algorithm tailored for a GPU (a) and a TPU (b), optimized for a matrix multiplication of size 8,192 × 8,192. Speed-ups are measured relative to standard (for example, cuBLAS for the GPU) matrix multiplication on the same hardware. Speed-ups are reported for various matrix sizes (despite optimizing the algorithm only on one matrix size). We also report the speed-up of the Strassen-square algorithm. The median speed-up is reported over 200 runs. The standard deviation over runs is 0.4 percentage points (see Supplementary Information for more details). c. Speed-up of both algorithms (tailored to a GPU and a TPU) benchmarked on both devices.

Figure 3: Speed ups of the AlphaTensor-discovered algorithm (Source: AlphaTensor paper)

The approach creates a systematic investigation of matrix multiplication algorithms through machine learning which extends human intuition capabilities beyond current boundaries.

4 Expected Results and Analysis