# Semantic Genetic Programming on the GPU

João David[1] and Ye Yang[1]

Departamento de Informática da Faculdade de Ciências da Universidade de Lisboa
{fc49448,fc49521}@alunos.fc.ul.pt

**Abstract.** Semantic Genetic Programming, also known as Semantic GP, is a type of Evolutionary Algorithm (EA), a subset of machine learning. Aiming to not only syntactically manipulate the data to promote a better fitness evaluation, but also ensuring that the resulting function is semantically correct. Traditional implementations of genetic programs have always been done on CPU, taking advantage of its multiple cores to do parallel work. We aim to improve on the parallelization of the genetic algorithm by implementing it on the Graphics Processing Unit (GPU), taking advantage of its vast amount of processing cores, which are multiple times that of the CPU. The implementation will be done on an NVidia GPU allowing the use of the Compute Unified Device Architecture (CUDA) API, which grants us access to directly manipulate the GPU memory and threads necessary to perform fitness evaluation and crossover.

**Keywords:** Genetic Programming · GPU · CUDA · Machine Learning.

## 1 Introduction

Traditional implementations of Semantic GP are on CPU, usually with syntax trees that contain elements of the domain in study, whether it be logical operations, mathematical operations, among others.

These trees are then processed given some input, and the output is compared to some target value to calculate its fitness. Fitness calculation algorithms are generally implemented on CPU, in this paper we aim to vastly improve the time required computing and evaluating fitness values by implementing an algorithm on GPU, taking advantage of its many cores and threads which allows for a vast amount of parallel computation to be done, when compared to a CPU.

In section 2 we explain some of the main concepts needed to understand the parallel implementation of fitness testing, in section 3 we discuss the general approach taken to parallelize fitness evaluation, which can be applied to any programming. In section 4 we explain in more detail our implementation which is written in C with a CUDA extension, allowing us to fully utilize all the functionality provided by the API. In section 5 we compare test results from both a parallel and sequential implementations of the algorithm. In section 6 we discuss previous work done on the subject matter and compare it with our own implementation and finally we summarize all the findings in the paper and future work to be done in section 7.

With this paper we aim to:

- To propose a parallel algorithm implemented on the GPU using the CUDA API for Semantic Genetic Programming
- compare execution times of the implemented algorithm with a sequential CPU version

## 2    Background

Genetic Algorithms are inspired by Nature and its biological mechanisms, take the following analogy as an example. Imagine that you are in the horse racing business, and you own a set of horses (population), some horses are better at different skills (speed, resistance, strength, etc.). When selecting horses to produce offspring (crossover and mutation), you must select them based on their fitness, which is calculated using a fitness function. In this case, it could be the average lap time from a set of tracks, better fitness values would mean lower average lap times.

The semantic syntax our work will be based on will be presented as a tree structure, each node containing either a mathematical operator, a literal or a variable which can be swapped by a value from a given data set. When processing a tree through a data set row, each variable will be swapped by the corresponding value in the data set according to the variable column index. In Figure 1 we can see the tree representation of the mathematical formula $x0*(x1+4)$ which is then processed according to a data set row containing the values (0.5   25.3   6.2   12). The values within the data set that are not referenced by any variable in the tree are ignored.
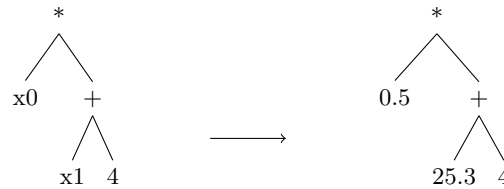


**Fig. 1.** Syntax tree value swap example

An NVIDIA GPU is composed of several SM (Stream Multiprocessor), each SM having a certain amount of cores, every core contains a number of warps (groups of 32 threads). The processing in an NVIDIA GPU is done through the use of blocks, which are composed of a set amount of threads. Each block will be scheduled to a single SM, and every SM can schedule different blocks at a given time if it has enough resources. When choosing the number of threads within a block, it is important that the number is a multiple of 32. Because if it isn't, there will be threads of a warp doing nothing.

Neither GPUs nor CPUs can directly access each others memory space. In order to cooperate, it is required to have copied memory from RAM to GPU memory (GDRR5 for instance) in order for GPUs to process data. Then the processed data must be copied back to RAM, so that the CPU may access it.

To execute a method on the GPU, we have to execute a kernel call. The kernel call can receive up to 3 parameters between the `<<<` `>>>` brackets. The first two parameters determine the number of blocks and threads per block respectively and the last one determines the amount of shared memory between threads available (threads can only access shared memory content within the same block) in bytes, when omitted it defaults to 0.

## 3   Approach

The algorithm starts by generating N trees, and then processing them based on the data set values. For each tree and for each row, there is an output value $V_{i,j}$, where $i$ is the tree index, and $j$ the row index (row number - 1) in the data set file. The Fitness ($F_i$) for each tree is then calculated using the Mean Squared Error (MSE) and stored in an array of size N indexed by the tree index. $Y_j$ is the target value of row's number $j$.

$$F_i = \frac{1}{N} \sum_{i=1}^{N} (V_{i,j} - Y_j)^2$$

After that, a matrix $A[\text{G} \times \text{N}]$ is created, where G is the number of generations, $a_{i,j}$ is the tree index with which the tree $j$ crossed over with at generation $i$.

$$A_{G,N} = \begin{bmatrix} 0 & 1 & \cdots & N-1 \\ a_{2,1} & a_{2,2} & \cdots & a_{2,N} \\ \vdots & \vdots & \ddots & \vdots \\ a_{G,1} & a_{G,2} & \cdots & a_{G,N} \end{bmatrix}$$

For every odd generation cycle, trees will be divided in two groups, one group with odd index and the other with even index, trees only crossover trees within the same group.

For every even generation cycle, trees will be divided in two groups, one group with index $<$ N/2 and the other with index $\geq$ N/2, trees only crossover trees within the same group.

The fitness array obtained from the previous iteration will be split in two groups, (based on the current generation cycle) in the same way mentioned above, the index of the best fitness from each group will be the tree index that all the trees from the same group will crossover with.

When a tree $a$ is crossing over with a tree $b$, the result of said crossover will be stored in a new fitness array at $a$'s index. The result would be:

$$newFitness[a_{index}] = oldFitness[a_{index}] + sigmoid(oldFitness[b_{index}])$$

Given this alternating crossover algorithm, we can achieve a less linear evolution from generation to generation. Algorithm 1 is a pseudo code representation of the algorithm.

---

**Algorithm 1** Tree crossover

---

 1: **procedure** TreeCrossover
 2:     **int[G][N]** $A$
 3:     **float[N]** $oldFitness$
 4:     **float[N]** $newFitness$
 5:     **for** $i = 0$; $i < N$; $i++$ **do**
 6:         A[0][i] = i
 7:     **for** $gen = 1$; $gen < G$; $gen++$ **do**
 8:         **if** $gen \% 2 \neq 0$ **then**
 9:             **int** $bestFitOdd$
10:             **int** $bestFitEven$
11:             **for** $i = 0$; $i < N$; $i++$ **do**
12:                 find best fitness with odd index and store it in $bestFitOdd$
13:                 find best fitness with even index and store it in $bestFitEven$
14:             A[gen][odd indexes] = bestFitOdd
15:             A[gen][even indexes] = bestFitEven
16:         **else**
17:             **int** $bestFitFirstHalf$
18:             **int** $bestFitSecondHalf$
19:             **for** $i = 0$; $i < N$; $i++$ **do**
20:                 find best fitness with index $< N/2$ and store it in $bestFitFirstHalf$
21:                 find best fitness with index $\geq N/2$ and store it in $bestFitSecondHalf$
22:             A[gen][index $< N/2$] = bestFitFirstHalf
23:             A[gen][index $\geq N/2$] = bestFitSecondHalf
24:         **for** $i = 0$; $i < N$; $i++$ **do**
25:             newFitness[i] = oldFitness[i] + sigmoid(oldFitness[A[gen][i]])

---

## 4  Implementation Details

Before any values could be first calculated, we needed some form of input to compare values with. This lead to the implementation of a data set parser. The data set is a text file in which all the variables and target values are defined. Each line has an equal amount of values $Y$ written, the last of which corresponds to the target value. Both variable values and target values are then extracted and stored in separate arrays which will be used to calculate the fitness for the first generation of trees.

To store and calculate the values, a set number of random syntax trees have to be initially generated. Each node of the tree as previously mentioned can be a literal value, a variable or a mathematical operator. After generating the

trees, each one of them must be run through the data set to evaluate its fitness by calculating the mathematical operations defined and comparing it with the target values. The computation is handled as a Reverse Polish Notation calculator (RPN) according to the following procedures:

- A separate LIFO stack to store literal values and variables is created
- The tree is traversed in postorder so that the root operation or value is read last
- If either a literal or a variable are present, they are pushed into the stack
- If a mathematical operation is read, 2 values are popped out of the stack for computing, and the resulting value is pushed back into the stack
- At the end of the processing, the final value is popped from the stack and stored in a result matrix of size $N_{trees} * dataset_{rows}$ with each row representing the calculated values for a given data set row by a syntax tree and each column the resulting values to compare with the target values.

### 4.1   GPU Parallelization

After all the values are calculated and stored in the array, we then proceed to the computation of the fitness for all trees for the first generation. In Figure 2 we present an example of fitness calculation with a data set of 4 lines, meaning that each tree computes 4 different values which are compared to the 4 target values to determine the MSE and consequently the fitness of the tree. Each tree is processed by a block (an abstract representation of a group of executing threads) which will have its own ID that can be extracted with the built in keyword *blockIdx.x*. The *.x* termination means we are extracting the ID along the $x$ axis, since the kernel is launched in 1 dimension, both the $y$ and $z$ values are 0.

$V_{i,j}$ is the value calculated by a tree of ID $i$ on line $j$ of the data set and $T_j$ the respective target value. Each $(V_{i,j} - T_j)^2$ value is calculated by a single thread which is able to compute the correct index on the array by calculating its position relative to its own ID (*threadIdx.x*) and the block's ID in the following manner: array$[(blockIdx.x * num\_rows) + threadIdx.x]$.

After all values are calculated we then proceed to a reduction phase, in this phase the amount of threads working on adding up all the values is reduced by half with each step until it reaches 1, which will be the final value to be divided by `num_rows` which will equal to the MSE of the computed value and the target value i.e. the fitness of the tree. Since all blocks work in parallel, the complexity of the reduction algorithm is $O(log(n))$.

## 5   Evaluation

### 5.1   Experimental Setup

| Processor | CPU Cores | Threads | RAM |
|---|---|---|---|
| Intel Xeon X5670 | 12 | 24 | 24GB |
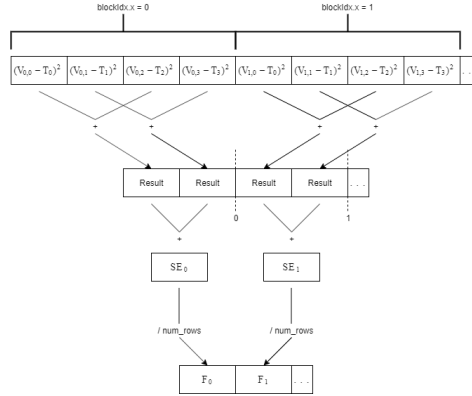
**Fig. 2.** Fitness computation for a data set of 4 lines

| Graphics Card | GPU Cores | CUDA Capability | GDDR5 Memory |
|---|---|---|---|
| NVidia GTX 960 | 1024 | 5.2 | 4GB |

In this section you should describe the machine(s) in which you are going to evaluate your system. Select the information that is relevant.

### 5.2   Results

In this section you should present the results. Do not forget to explain where the data came from.

You should include (ideally vectorial) plots, with a descriptive caption. Make sure all the plots (Like Figure **??** are well identified and axis and metrics are defined.

### 5.3   Discussion

Here you should discuss the results on a high level. For instance, based on our results, the parallelization of the merge-sort is relevant as no other parallel work occurs at the same time, and the complexity $O(Nlog(N))$ can have a large impact when the number of individuals is high.

## 6   Related Work

Alberto Moraglio, Krzysztof Krawiec, and Colin G. Johnson introduced a new way of genetic programming, Geometric Semantic Genetic Programming (GSGP) which consists in changing the semantics of the functions directly. Their work in regards to fitness evaluation, tree mutation and crossover is far more complex

than ours given that we only accounted for mutation by concatenation with sums of trees. However the implementation of the genetic algorithm was done on CPU which makes performance considerably slower if it were implemented on GPU.

## 7    Conclusions

In this paper, we proposed a parallel GPU algorithm to both calculate and evaluate fitness on a genetic program, by applying 2 different crossover techniques. We concluded that GPU parallelization significantly benefits the execution time when passing from one generation to the next, by dynamically distributing work across the many GPU threads available. However, the size of the initial data set must be taken into account, as smaller data sets may be quicker to compute on the CPU due to the significant overhead when transitioning from CPU memory to GPU memory. Thus we must define a cutoff mechanism between CPU and GPU execution which can be determined in future tests.

We presented a simple crossover algorithm when passing from one generation to the next, in future work we would like to implement tree mutations and crossover between trees rather than the fitness calculated with each passing generation. The implementation of a CPU parallel version of the algorithm would also compliment the results obtained in this paper, so a comparison between the two parallel frameworks can be made.

## Acknowledgements

João David wrote the sequential version of the algorithm and two GPU methods, while Ye Yang wrote the other GPU method, the data set parser and the tree processing.

Both authors wrote this paper, with João David focusing on the approach and introduction while Ye Yang focused on implementation details, related work and conclusion.

Each author spent around 40 hours on this project.