Joseph Bingham
Iowa State University
Department of Mathematics
jbingham@iastate.edu
515-451-1365
March 23, 2018

Cluster Time Proposal

To whom it may concern,

I submit this proposal for your review. As a mathematics student at Iowa State University, I hope to use the high performance compute cluster for my math 491 undergraduate thesis, which will conclude at the end of the spring 2018 semester. This project will consist of two novel concepts and techniques:

- The use of genetic algorithms for steganalysis, and
- A parallel implementation of a genetic algorithm

I. Proposal Summary

A. Proposal

The research performed on the cluster will produce solutions to a combinatorial optimization problem using a highly parallelized genetic algorithm.

B. Background Information

1. Steganography

Steganography is the practice of embedding information into photographs or audio media in such a way that it is not detectable to the average person. The form of steganography that the scope of this project will be most interested in is LSB (Least Significant Bit) embedding of images using random paths.

This, as the name suggests, is where the path of the embedding stream is random in nature. When a pixel from the image is selected to be altered, the least significant bit is overridden to be the same as the next bit in the embedding stream. Since it is just the least significant bit, very little visual alterations occur.

2. Genetic Algorithm

Genetic Algorithms are algorithms that follow a specific machine learning paradigm. They are classified by their similarity to natural selection as found in evolution in the wild. They consist of three main parts: a cost function, a tester bot, and a builder bot.

The cost function determines the viability of each solution. Typically this function determines how well the solution completes the task that is trying to be accomplished, and is usually from $\mathbb{R}^n \rightarrow [0, 1]$.

The tester bot tests each solution based on the cost function. Its job is to determine what the value of the test function and to rank each solution.

The builder bot takes each the best solutions, as ranked by the tester bot, generates more solutions based on their attributes. It generates the new solutions by randomly modifying the best solutions in hopes of descending the gradient.

C. General Outline

1. Genetic Algorithm for Steganalysis

One of the novel components to this proposal is the use of genetic algorithm for steganalysis. The thought behind this is to generate a classifier for steganographically embedded images versus normal cover images.

This will be done by creating a evolutionary solution for a path finding algorithm. The path finding algorithm will optimize the possible paths of embedding to determine the probability of payload being embedded. If the genetic algorithm's internal weight function is triggered, then the probability along the path found was deemed high enough to consider the image having a payload embedded.

2. Parallel Implementation of Genetic Algorithms

The other novel component is the parallelization of genetic algorithms to speed up the training time of the evolutionary model. The reason this project will require the will require a parallel training model, as opposed to a serial implementation, because the quantity of data that needs to be ingested by the engine is immense.

The parallelization will be achieved by using a "island" inspired model. This is where each node of the cluster getting its own evolutionary model, but with the same cost function. Then after some number of iterations, the nodes give each other their top solutions to reproduce with the other solutions. This will hopefully speed up, and aid with, the convex optimization.

II. Resources Needed

A. Cluster Requirements

1. The ability to run MPI (Message Passing Interface) for C++
2. Must have at least 4 compute nodes (preferably GPU, but CPU will work)
3. Each node has to have at least 8 gb of RAM

B. Time Requirements

1. At least 35 hours of run time on the cluster for training of engine

a) Can be chunked, if the computations are not lost
b) If they are lost, must have at least 5 hour continuous time per run

III. Impact

A. Material

This project will result in the creation of a detection algorithm for steganalysis which can be applied to not just fixed patterns of embedding, but random paths of embedding.

B. Immaterial

This project will create a framework for the parallel training of evolutionary models of genetic algorithms. This framework could be used for a wide range of problems, not just steganalysis.

IV. Risk Mitigation

A. Possible Issues

1. The parallel algorithm does not preserve the gradient descent
2. The parallelization does not have enough time to train on the high performance cluster

B. Mitigation Methods

1. There will still be a serial implementation that will run
2. The serial implementation can run on a local machine and train for longer

V. Deliverables

1. A serial implementation of a genetic steganalysis algorithm
2. A parallel implementation of a genetic steganalysis algorithm
3. A proof of that the parallelization will have the same optimizing property