Predicting Atom Location Using Machine Learning Algorithms

Adam Martini, Wes Erickson, Ran Tian

Computer and Information Science Department
University of Oregon
Eugene, Oregon 97403, USA
martini@cs.uoregon.edu
author2@someotherinst.something
author3@someotherinst.something

Abstract. Here is an abstract...

Keywords: machine learning, logistic regression, ...

1 Introduction

The Steck Lab is investigating the motion of a single atom in a Magneto Optical Trap (MOT). The atom is imaged using a pair of aspheric lenses and a CCD camera. Given an 2D image from the CCD, the lab wants to predict the probability distribution of the atom's location within the MOT. This report describes our efforts to both simulate the images generated by the camera, and create a classification algorithm that will estimate the atom probability distribution.

2 Design and Objectives

To tackle this problem we run simulations of the CCD output for a given atom location using a raytracing simulation. The output images are labeled according to atom location and used as a training set for machine learning algorithms. The algorithms can then be used to generate a probabilistic distribution for the atom's location for a new CCD image.

Figure 1 describes the data flow of our architecture. We begin by simulating the training and testing data using raytracing and filtering. The filters will apply both generate several different training sets from a single image set. The classification results for each training set are described below. Some of the instances are selected from the image set for noise filtering. The *noisy* data is used to test our classification algorithms.

The objective of our project is to create and test a scalable machine learning solution for atom location prediction using CCD images. In order to scale our application we use both shared memory and distributed memory parallelism. Massive amounts of training data are generated using a shared memory parallelism, and a distributed memory logistic regression classifier is used
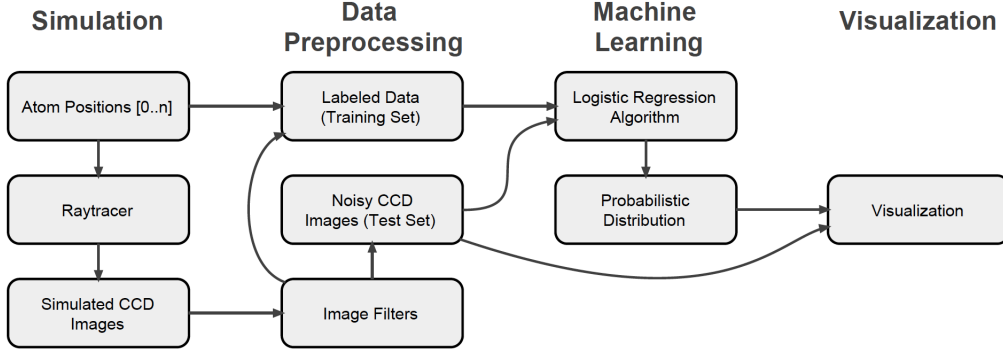
Fig. 1: High Level Data Flow Architecture Diagram

to efficiently learn interesting patterns from the data. The remainder of this report describes our development efforts, experimentation results, and the lessons we have learned during the process.

3   Development and Implementation

Our development efforts are divided into sections based on our data flow architecture. There is some overlap of effort in the data preprocessing step as this is the area of our development where Simulation and Machine Learning meet. We describe each development stage in detail in the following subsections.

3.1   Simulation

The essence of the simulation is starting with a flourescent atom at a chosen location, and then simulating how light emitted from the atom appears on the CCD. There are two approaches to simulating the generation of CCD images from a given atom location.

The first is a classical raytracing approach. Rays are emitted in a random direction from the atom, and obey refraction via Snell's Law as they pass interfaces between different materials. The rays also contain phase information for the purposes of determining interference. When the rays intersect with the CCD plane, we increment an array value corresponding to which CCD pixel it hit.

The second approach is a full simulation of the linear wave equation. This is a full 3D simulation, and is much more computationally intensive. A field discretized into $200 \times 200 \times 1000$ is given an initial value of zero everywhere except for a peak at our atom location, and then is stepped forward in time according to the linear wave equation. The essence of this algorithm is basic matrix multiplication, but with very large sparse matrices.

3.2   Data Preprocessing

Ran talks about filtering here...

The feature scaling step is handled in the logistic regression classifier as it is a simple extension to data loading after the instance have been loaded into memory. We use efficient a templated C++ library (Eigen [1]) that provides efficient matrix operations that can quickly perform the broadcasting operations necessary for feature scaling. In addition, the distributed implementation of our algorithm lends itself well to efficient data partitioning, which ensures that this preprocessing step will scale. However, feature scaling in a distributed setting is slightly more complex in that it requires knowledge of the global min and max for the entire dataset across all nodes. The min/max vectors is quickly computed using an Allreduce step with MPI's MIN/MAX reduction operators. Each node then formats its section of the data using a single call feature scaling utility function.

3.3 Machine Learning

We have developed a distributed logistic regression classifier using MPI to train on massive amounts of data efficient. The classifier is capable of multi-class classification and mini-batch processing. During training, nodes communicate with each other to both build the model and perform parameter updates.

The model building step includes data loading, feature scaling, and label formating. During data loading, each node loads a portion of the instances the provided data directory based on its task ID and the total number of tasks in the communication world. It is important to note that the instance file names are first loaded into a vector and then randomized to ensure that mini-batch processing have a consistent sample of the data with each iteration. After the data is loaded, feature scaling is performed as described above. The label formating step uses a custom MPI operator in an Allreduce call to generate global unique label set. The unique label set is used at each node to generate a label matrix from the label vector for 1-vs-all classification.

The model training involves iterative updates to a set of global parameters. During each update, the nodes compute a portion gradient update using a subset of the data they are responsible for based on the batch size. The collective gradient update is computed using an Allreduce call to sum each contribution followed by a normalization step on each node based on the total number of instances contributing to the update. The new parameter set (identical on all nodes) is then tested for convergence and used for the next iteration.

Training is finished when either the gradient has converged or the maximum number of iterations has been reached. The Master node saves the trained model parameters to the a file to be loaded on a single node for testing. The testing program generates a set of class membership probabilities for each instance in the given data directory. The chosen class is simply the maximum probability found in the probabilities set for each instance. We also use the probabilities to generate the visualizations of the models *beliefs* about an atoms location for each instance.

3.4   Visualization

Visualization is particularly important for visualizing three dimensional volumes, and it may become essential for feedback as to whether our simulations are functioning properly, and for our presenting our project results. However, our project focus is not on programming the visualization methods ourselves, but using existing tools such as Visit or ParaView.

The main items to be visualized are the field generated from the differential equation solvers, and the CCD images (which are a cross section of the field). We can also use these tools for post processing on the image data to highlight features for the benefit of the machine learning algorithm.

4   Experiments and Performance

Text...

5   Lessons Learned

5.1   Simulation

5.2   Data Preprocessing

Ran talks about filtering here...

Adam talks about feature scaling here...

5.3   Machine Learning

5.4   Visualization

6   Conclusions

Text...

References

1. Eigen Linerar Algebra Library.
   http://NOINTERNET...