GSoC 2023 Project Proposal

Organization: ML4SCI

Gravitational Lens Finding for Dark Matter Substructure Pipeline

MENTORS

Emanuele Usai, Anna Parul, Michael Toomey, Pranath Reddy, Stephon Alexander, Saranga Mahanta, Karthik Sachdev

AUTHOR

Azal Ahmad Khan

Contents

1 Student Information and Introduction

1.1 Student Information

- Name: Azal Ahmad Khan
- Email: azalahmadkhan@gmail.com, k.azal@iitg.ac.in
- Website: https://sites.google.com/view/azalahmadkhan
- GitHub: azalahmadkhan
- Contact Number: +91 8957776840
- University: Indian Institute of Technology Guwahati, India
- Major: Chemical Science and Technology (Bachelor of Technology)
- CPI: 9.22/10 (For semesters 1-5)
- More information: Curriculum Vitae
- Time Zone: IST(GMT + 5:30)

1.2 Introduction

I am a junior studying Chemical science and technology at the Indian Institute of Technology Guwahati, and I constantly seek to expand my knowledge and understanding of Machine Learning. While my interest in this field is broad, I am working to focus on a specific area within it. I am fortunate to have experience working as a research intern at multiple prestigious universities, giving me valuable insights and practical skills to apply to my current projects. Currently, I am focusing on federated learning, bayesian deep learning, and MCMC, and I am excited to see where these avenues of research will take me. I am part of the Transitional Artificial Intelligence Research Group, UNSW, and DMLSys (Distributed Machine Learning Systems) Lab, UMN. My research interests are Deep Learning, Large Deep Learning Models, Federated Learning, and Computer Optimization.

2 Description

Strong gravitational lensing is a promising probe of the substructure of dark matter to better understand its underlying nature. Deep learning methods have the potential to accurately identify images containing substructure and differentiate WIMP particle dark matter from other well-motivated models, including vortex substructure of dark matter condensates and superfluids.

This project will focus on the further development of the DeepLense pipeline that combines state-of-the-art of deep learning models with strong lensing simulations based on lenstronomy. The focus of this project is detecting strong lenses from data sets of mock surveys. These newly found lenses can then be passed on to the rest of the DeepLense pipeline for detection, classification, and interpretation of dark matter substructure.

3 Technical Details

3.1 Lenstronomy

Lenstronomy is a Python library for modeling and simulating gravitational lenses. Gravitational lensing is a phenomenon in which the light from a distant object is bent and distorted by the gravitational field of an intervening massive object, such as a galaxy or a cluster of galaxies. The Lenstronomy library provides a set of tools and models for simulating and analyzing gravitational lenses, including tools for modeling the mass distribution of the lensing object, fitting observed lensing data, and simulating lensed images. The library is designed to be modular and customizable, allowing users to easily construct complex lensing models and compare them to observational data. Lenstronomy has been used in a variety of applications, including studying the dark matter distribution in galaxy clusters, probing the expansion history of the universe, and testing alternative theories of gravity.

3.2 Evaluation Test and Results

3.2.1 Common Test

The dataset comprises three classes: strong lensing images with no substructure, subhalo substructure, and vortex substructure. To tackle this task, a custom model was developed, which drew inspiration from the Densenet and Convolutional Block Attention Module approaches.

Initially, Cross Entropy Loss was used as the loss function, but it was observed to cause overfitting. To address this, Label Smoothing Binary Cross-Entropy (BCE) was employed, which reduces the number of overconfident predictions that are extremely close to 1 or 0. This type of loss function acts as a regularizer.

The model was trained for three epochs using Cross Entropy Loss to build confidence before switching to Label Smoothing BCE. After experiments, it was determined that this approach yielded the best-performing model.

The achieved ROC AUC accuracy for the no substructure class was 0.96911, while for the spherical and vortex substructure classes, it was 0.93808 and 0.94485, respectively. The micro-average accuracy was 0.95389, and the macro-average accuracy was 0.95093.[Link]

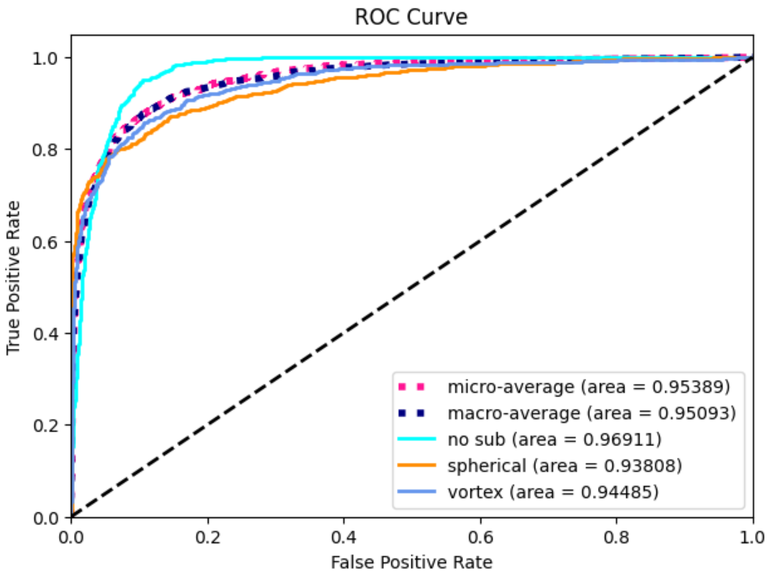

Figure 1: ROC curve for multi-class classification of test 1.

3.2.2 Lens Finding (Specific Test 2)

In this task, a novel approach was taken by combining image and tabular data to achieve high accuracy. While using only image data was effective, the incorporation of meta-features proved to be the most successful approach. It is important to note that careful feature engineering was carried out on the tabular (meta) data prior to feeding it into the neural network.

Through this hybrid approach, an impressive ROC-AUC accuracy of 0.972 was achieved, indicating the robustness of the proposed solution. It can be inferred that the incorporation of tabular data provided additional valuable insights to the model, thereby improving its overall performance.

Furthermore, this approach has the potential to be applied to various other image classification tasks that involve meta-features, leading to improved accuracy and more informed decision-making.[Link]

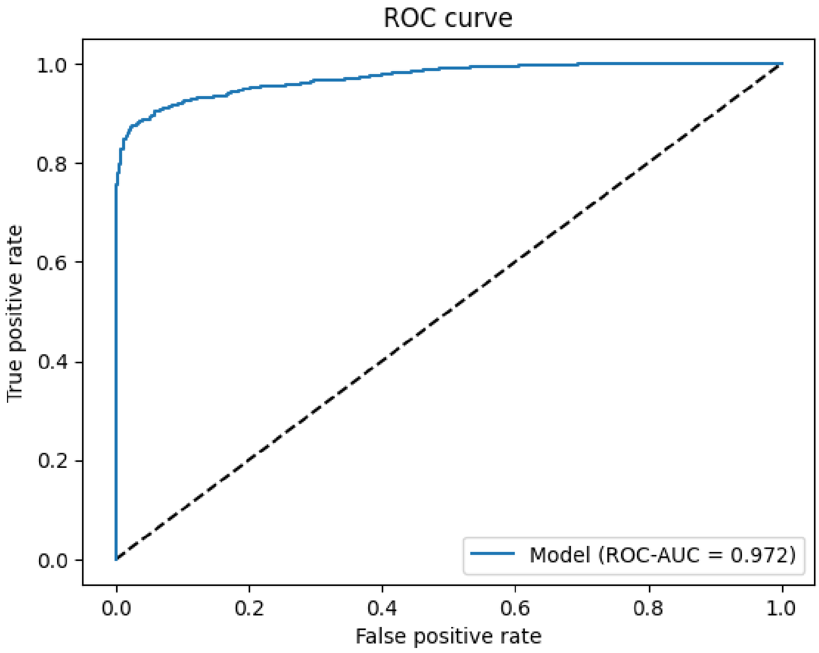

Figure 2: Analysis of results from image classification of test 2.

4 Proposed Deliverables

1. Implement a classification model that determines whether the images contain lenses or not while exploring ResNet and DenseNet models.

2. Implement regression models for the lens datasets. For this, models from ResNet and DenseNet families will be explored using metrics like mse and rmse.

3. Implement anomaly detection on the dataset. The approach we will be taking involves utilizing binary/multi-class image classification for anomaly detection.

4. Implement multiple types of GANs for expanding the datasets.

4.1 Schedule of Deliverables

1. **Community Bonding Period: May 4, 2023 - May 28, 2023**
   During the community bonding period, I will look into the relevant literature useful for the project and brush up on relevant deep learning-based concepts. In the past, I have leveraged state-of-the-art deep learning models, such as EfficientNet-B2, to solve various tasks. These models have demonstrated superior performance compared to other existing models and have significantly advanced the field of machine learning. I will also try as much as possible to get to know the ML4SCI developer community, interact with them and head start a great journey.

2. **Week 1 and 2**
   Work on developing a classification model to accurately identify whether images contain lenses or not. In the evaluation task, we achieved a roc-auc score of 0.972 while using the EfficientNet-B2 model. Further, this will be extended to EfficientNet-V2 and EfficientNet-L2. Similarly, the performance of models from ResNet and DenseNet families will be explored using metrics like AUC, ROC, and confusion matrices.

3. **Week 3**
   Work on developing a regression model to accurately predict the mass density of the vortex substructure of dark matter condensates on simulated strong lensing images. In the evaluation task, we got a mean squared error of 0.0003 while using the deep learning model. Similarly, the performance of models from ResNet and DenseNet families will be explored using metrics like mse and rmse.

4. **Week 4 and 5**
   Work on incorporating anomaly detection into the dataset. The approach we will be taking involves utilizing binary/multi-class image classification for anomaly detection. To implement this, we plan to explore various techniques such as autoencoder, variational autoencoder (VAE), Adversarial Convolutional Autoencoder (ACAE), Recurrent Neural Network (RNN) Autoencoder, Variational Recurrent Autoencoder (VRAE), Stacked Convolutional Autoencoder (SCAE), transformer models, etc. By leveraging these techniques, the aim will be to identify and classify anomalies in the data with high accuracy and precision.

5. **Week 6**
   Complete all previous tasks. This is a buffer week for any unprecedented delays. Publish blog posts. Prepare for Phase 1 Evaluation.

Phase 1 evaluation

6. **Week 7**
   Work on further enhancing the DeepLense pipeline. Work on expanding the dataset by utilizing various Generative Adversarial Network (GAN) architectures, such as Vanilla GANs, Deep Convolutional GANs (DCGANs), Wasserstein GANs (WGANs), StyleGANs, and Conditional GANs. By training these GANs on the existing dataset, we can generate new synthetic images that can be used to augment the original dataset. This approach can improve the model's ability to generalize and make accurate predictions on unseen data.

7. **Week 8**
   Work on benchmarking the four proposed methods, testing them, verifying results, and fixing bugs (if any).

8. **Week 9**
   Work on integrating the four proposed methods and testing them. Complete documentation for newly built methods, verify results, fix bugs (if any) and write additional unit tests.
9. **Week 10**
   Complete Jupyter Notebook Tutorials for all the proposed methods and modifications. Publish blog posts and prepare for Phase 2 Evaluation.

Phase 2 evaluation

10. **Future Works and Post GSoC**
    After the proposed 10-week timeline, I would love to start implementing any additional features and contribute to ML4SCI even after GSoC, and given an opportunity, I would love to pursue Ph.D. on related topics.

5 Other Information

5.1 Why ML4SCI?

As a Chemistry major with a keen interest in deep learning, ML4SCI is the perfect organization for me to participate in the GSoC program. ML4SCI's focus on applying Machine Learning techniques to fundamental sciences aligns well with my academic background and research interests. By working on ML4SCI projects, I will be able to expand my knowledge and skills in both fields and gain valuable experience working with a team of experts in the field.

Furthermore, the opportunity to work on real-world problems in collaboration with established researchers is an excellent way to gain practical experience and enhance my research abilities. Additionally, ML4SCI's focus on basic sciences research makes it an ideal organization for me to gain experience and knowledge for my future Ph.D. program. The research experience gained through GSoC can be an added advantage when applying to top universities for doctoral studies. As someone who has already worked at renowned institutions and submitted research papers, I am excited about the opportunity to continue contributing to cutting-edge research and further advancing the frontiers of science through ML4SCI. Overall, participating in GSoC with ML4SCI will not only enhance my skills and knowledge but also provide a stepping stone for my future academic pursuits.

5.2 Relevant Background

I have over one year of experience conducting research in machine learning. During this time, I have authored four preprint publications. Additionally, I have submitted a paper on reward-based personalized federated learning to the upcoming International Conference on Machine Learning (ICML) 2023 and another paper on class imbalance and ensemble learning to the journal Neurocomputing, also set to publish in 2023.

In addition to these accomplishments, I have developed a new metaheuristic optimization algorithm based on principles from quantum physics. Furthermore, I am actively engaged in three ongoing research projects focused on personalized federated learning, Bayesian and variational deep learning, and Markov Chain Monte Carlo (MCMC). I have also applied transformer models in feature generation in the tabular dataset. To the best of my knowledge, this is the first such approach. [Link to repository]

Currently, I am further expanding my expertise through enrollment in a course on reinforcement learning, aiming to broaden my skill set and contribute to future advances in reinforcement learning.

5.3 Other commitments

Between May 12th and June 25th, I will be on summer break and able to commit up to 50 hours per week to any given task. Once my college classes resume on June 26th, I will transition to part-time work and be available for approximately 40 to 50 hours per week. My research internships are scheduled to be completed by mid-May 2023 as planned. In the event of any unforeseeable circumstances resulting in a reduction in weekly working hours, I will promptly inform my project mentor, Dr. Emanuele Usai, and make arrangements to compensate for the missed time in subsequent weeks. To ensure that I am readily available for any necessary catch-up work, I will be accessible via Skype or Zoom during Indian Standard Time Zone, GMT +5:30.