Google Summer of Code

Author

Sawradip Saha

Google Summer of Code @ ML4SCI Machine Learning for Science

Project

Gravitational Lens Finding for Dark Matter Substructure Pipeline - Deeplense

Mentors

Anna Parul (University of Alabama)
Jeremy Quijano (University of Alabama)
Michael Toomey (Brown University)
Saranga Mahanta (Institut Polytechnique de Paris)
Karthik Sachdev (RWTH Aachen)

Table of Contents

ABOUT ME

My Details

Name: Sawradip Saha

Location: Dhaka, Bangladesh

Email: sawradip0@gmail.com

Contact No.: +8801914107509

University Name: Bangladesh university of Engineering and Technology

Department: Mechanical Engineering

Github: github.com/sawradip

Time Zone: GMT+6

My Programming and ML/DL experience

I started learning programming first for my academic coursework, before admitting to universty. Soon I realized that it fascinates me very much and spent more time with it. I have extensive experience to Python as well as moderate experience with C++, C# and Java.

I started learning about Machine Learning and Deep Learning during the covid pandemic when universities were closed. After strengthening my theoretical foundation, I started to develop the coding aspects, as well as explore myself to its application in various fields, including participating in multiple Deep Learning based competitions.

I have participated in and also won several competitions, including the IEEE Video and Image Processing Cup 2021(**1st**) where **we built an In-bed human pose estimation and Body Keypoint detection System** and 2022(**3rd**) where **we created a Deep Learning System for Detecting synthetically generated and modified images**. I also have achieved **2nd** Position at Deep Learning Sprint 2022 **a state-of-the-art speech-to-text transcription model for Bengali language was built and trained by us**. I have previously worked with Pose Detection/Action Recognition as well as other computer-vision related tasks, during my competitions. and I am confident that my experience and skills would be a great fit for my proposed project.

Technical Information

Programming Languages:

- **Python (Advanced)** : 3 Years Experience, comfortable with OOP as well as all other important concepts.
- **C++(Intermediate)** : Have written small programs, Gone through codebases, familiar with modern C++
- **C# (Basic)**
- **Java(basic)**

Development Environment:

- **IDE/Editor**: Visual Studio Code, Visual Studio Community 2022
- **Architecture**: Intel® Core™ i7-12700 Processor
- **OS**: Windows 10 - Ubuntu 22.04
  - Comfortable with Git, Github and Linux Bash Terminal

My Open-Source Experience

Open-source projects have always intrigued me, but I used to doubt my abilities and skills to make meaningful contributions to them. Then I decided to take small steps and learn from them. I made my first PR to Pytorch Official Repository (a documentation fix) as I had some experience with PyTorch. It was a great boost for my confidence when it got accepted. I also added Google Colab support to Skorch tutorials as another contribution.

ABOUT THE PROJECT

Selected Project

This [Gravitational Lens Finding for Dark Matter Substructure Pipeline] project aims to further develop the DeepLense pipeline, which combines deep learning models with strong lensing simulations based on the lenstronomy library. The focus will be on detecting strong gravitational lenses from mock survey data sets, which can then be used for the detection, classification, and interpretation of dark matter substructure. The project will contribute to the advancement of particle physics and deepen our understanding of the nature of dark matter through image analysis and machine learning.

Reason of Selecting this Project

I have always been fascinated by the mysteries of the universe, and the study of astronomy has been a passion of mine since childhood. As I delved deeper into the field of astrophysics, I became intrigued by the concept of dark matter, one of the biggest mysteries in the universe. With my programming and deep learning background, I saw an opportunity to contribute to the ongoing efforts to understand dark matter by applying machine learning techniques to analyze strong gravitational lensing data.

In addition to the technical challenge of the project, I am motivated by the potential impact that this research can have in advancing our understanding of the universe. By identifying and interpreting dark matter substructure using deep learning methods, we can gain insights into the nature of dark matter and its role in the evolution of galaxies. Overall, I believe that this project offers a unique opportunity to combine my interests in astronomy and machine learning to contribute to the advancement of science.

Solution Abstract

This project proposes to develop a deep learning pipeline that overcomes the existing limitations in detecting and interpreting strong gravitational lenses in mock survey data sets. The main challenges that this project aims to address are the low accuracy and inefficiency of the current detection and classification methods. The proposed pipeline will be built upon the existing DeepLense framework and will use state-of-the-art deep learning models, such as Convolutional Neural Networks (CNNs) and Transformers, to classify and interpret survey data.

Detection -> Classification -> Interpretation

To evaluate the performance of the pipeline, various evaluation metrics will be used, including ROC, AUC, and Accuracy. In addition, potential benchmarks or comparisons with existing methods will be conducted to demonstrate the pipeline's effectiveness in improving the accuracy and efficiency of the lens detection and classification. Finally an interactive dashboard for visualizing and interpreting lens detection results will be developed.

The project will be divided into different milestones to give a sense of the expected progress and deliverables throughout the GSOC period. These milestones include developing and testing the detection and classification component, designing and implementing the interpretation component, and evaluating the overall performance of the pipeline using the established metrics.

In addition to its scientific significance, the proposed pipeline has potential applications in the wider astronomy community, such as improving our understanding of galaxy formation and evolution. The successful completion of this project will contribute towards the development of accurate and efficient pipelines for detecting and interpreting gravitational lenses, enabling further advancements in our knowledge of dark matter and its properties.

Project Deliverables

The following are the deliverables of this project:

- Improved version of DeepLense pipeline for detecting strong lenses
- Evaluation and comparison of various deep learning models for lens finding task
- Integration of lens finding module with the rest of the DeepLense pipeline
- Development of an interactive dashboard for visualizing and interpreting lens detection results
- Publication of project findings and code as open-source software

Time Allotment

I would like to take this as a **350-hour project** and is designed as a 12-week program, which means that an average time of 4-5 hours a day is required to complete it. If needed, I am willing to dedicate up to 8 hours per day to ensure timely completion of the project.

However, it should be noted that the estimated time for project completion may vary due to unforeseen circumstances. I am aware of this possibility and am willing to work beyond the designated timeline (Extended Timelines: Sep 4 - Nov 6) mentioned in GSOC's timeline, if necessary.

As for work hours, I am highly flexible to adjusting my schedule to accommodate project and mentor needs. I am also comfortable working independently and managing my own time to ensure timely completion of project goals and deadlines.

Technical Details

Gravitational Lens Finding

Gravitational lens finding is the process of detecting and studying the phenomenon of gravitational lensing, which is the bending of light by the gravity of massive objects such as galaxies or galaxy clusters. Gravitational lensing can magnify and distort the images of distant sources that are behind the lensing object, creating multiple images or rings of light. Gravitational lens finding can reveal information about both the lensing object and the source object, such as their mass, shape, distance, and nature. Gravitational lens finding can also be used to test Einstein's theory of general relativity, which predicts the amount of light bending by gravity. Gravitational lens finding can be done by using different types of telescopes and instruments that observe different wavelengths of light, such as optical, infrared, radio, or X-ray. Gravitational lens finding can also use different methods of analysis, such as strong lensing, weak lensing, or microlensing, depending on the strength and scale of the lensing effect. This effect was confirmed by observation of the Twin QSO SBS in 1979.[Source:Wikipedia]

Deep Learning in Gravitational Lens Finding

Deep learning is a branch of machine learning that uses artificial neural networks to learn from data and perform tasks such as classification, regression, and generation. Deep learning can be applied to gravitational lens finding, which is the process of detecting and studying the phenomenon of gravitational lensing, which is the bending of light by the gravity of massive objects such as galaxies or galaxy clusters. Gravitational lensing can magnify and distort the images of distant sources that are behind the lensing object, creating multiple images or rings of light. Deep learning can be used to train convolutional neural networks (CNNs) to classify sky images based on the presence or absence of gravitational lenses, or to infer the properties of gravitational lenses such as mass, size, or density slope. Deep learning can also be used to generate realistic simulations of gravitational lenses for training and testing purposes. Deep learning can potentially improve the efficiency and accuracy of gravitational lens finding, and help to probe the nature of dark matter and cosmology. This NVIDIA Blog discusses more about this,

Evaluation Task Results

During the prerequisite for this project, I have completed the required tasks and also contributed by creating a flexible training and testing pipeline using PyTorch Timm as backend.

Here is a short description of my evaluation results.

Common Test I: Multi-Class Classification (Common)

Model Name	AUC	Accuracy
rersnet18	0.9924	95.58
efficientnet_b0	0.9704	93.41
convit_tiny	0.8826	95.58
resnet26	0.5528	89.48

Specific Test II: Lens Finding

Model Name	AUC	Accuracy
convit_tiny	0.9114	83.00
vit_tiny_patch16_224	0.8973	81.61
rersnet18	0.8604	80.55
efficientnetv2_s	0.5677	68.05

I strongly encourage visiting the linked project pages, and observing the training plots, AUC curve, and my ideas of improvement.

Along with the **Common Test I** and the **Specific test II** for this project, completed Tasks:

- Implemented a regression algorithm using the provided dataset to learn the mapping between lensing images and the lensing dark matter halo mass.**[Specific Test III: Learning Mass of Dark Matter Halo]**
- Used a vision transformer method to build a robust and efficient model for binary classification or unsupervised anomaly detection on the provided dataset.**[Specific Test V: Exploring Transformers]**

Additional Contributions:

- Created a flexible training and testing pipeline that can test hundreds of computer-vision models using PyTorch Timm as the backend.
- Provided clean and minimal training metrics logging in a model.json file for easy monitoring.
- Created an easy saving, loading model, and resuming training functionality.
- Made the pipeline easily extensible with any Loss function and optimizers, and capable of handling classification, regression, as well as object detection tasks.
- Provided a multi-model ablation study for each task and organized plots of various training parameters.
- Ensured that the notebooks can be run in Google Colab or any environment with data downloading and other setups done within the notebook.

These contributions and completed tasks have given me the necessary technical knowledge and experience to tackle this project with confidence.

My Prerequisite tasks are organized into this Github Repository.

Detailed Timeline

Community Bonding Period(May 4 - May 28)

- Familiarize with DeepLense pipeline and its codebase.
- Study gravitational lensing, dark matter substructure, and lenstronomy.
- Interact with the mentor and other community members to discuss the project and clarify doubts.
- Set up the development environment and necessary tools.

Week 1 (May 29 - June 5)

- Explore and analyze the dataset for the project.
- Preprocess and augment the data to increase its size and diversity.
- Implement and test various deep learning models ***with my developed pipeline during the prerequisite tasks*** for detecting strong gravitational lenses.
- Discuss the results and their implications with the mentor.

Week 2 (June 6 - June 12)

- Refine the best-performing deep learning model by fine-tuning its hyperparameters.
- Evaluate the model on the validation set to measure its performance.
- Analyze and visualize the model's predictions and errors to **gain insights into its behavior**.
- Discuss the results and their interpretation with the mentor.

Week 3 (June 13 - June 19)

- Implement and test different loss functions and optimizers to improve the model's performance.

- Compare the results of different model configurations and select the best one.
- **Integrate the model into the DeepLense pipeline** and test it on the full dataset.
- Discuss the results and their interpretation with the mentor.

Week 4 (June 20 - June 26)

- Study and implement techniques specialized for detecting dark matter substructures from the detected strong gravitational lenses.**[Papers Mentioned in ML4SCI site]**
- Experiment with different architectures and hyperparameters for the substructure detection model.
- Analyze the substructure detection model's performance and **interpret its results.**
- Discuss the results and their implications with the mentor.

Week 5 (June 27 - July 3)

- Incorporate real-world data and test the models' performance on the real datasets.
- Fine-tune the models using the real data and improve the model's accuracy.
- Development of an interactive dashboard for visualizing and interpreting lens detection results

Week 6 (July 4 - July 10)

- Review the work done so far and identify areas for improvement.
- Discuss the progress and plan for the remaining weeks with the mentor.
- Catch up on any pending tasks or issues.
- Review the progress of the project so far and make any necessary adjustments to the plan.
- Begin integrating the models into a cohesive pipeline for processing audio and video inputs and generating meeting notes.

Week 7 (July 11 - July 17)

- Explore and study advanced deep learning techniques for gravitational lens finding and dark matter substructure detection.
- Experiment with new model architectures and techniques to improve the accuracy and robustness of the pipeline.
- Discuss the results and their implications with the mentor.

Week 8 (July 18 - July 24)

- Refine the substructure detection model by fine-tuning its hyperparameters and testing different configurations.
- Analyze and visualize the results of the substructure detection model to gain insights into its behavior.
- Integrate the substructure detection model into the DeepLense pipeline and test it on the full dataset.
- Discuss the results and their interpretation with the mentor.

Week 9 (July 25 - July 31)

- Study and implement methods for classification and interpretation of dark matter substructures.
- Experiment with different techniques and model architectures for substructure classification and interpretation.
- Evaluate the performance of the classification and interpretation models and interpret their results.
- Discuss the results and their implications with the mentor.

Week 10 (August 1- August 7)

- Integrate the classification and interpretation models into the DeepLense pipeline and test them on the full dataset.

- Analyze and visualize the results of the pipeline to **gain insights into its behavior and accuracy.**
- Refine the pipeline by incorporating feedback and suggestions from the mentor.
- Discuss the results and their interpretation with the mentor.

Week 11 (August 8 - August 14)

- Review the work done so far and identify areas for further improvement and future work.
- **Document the pipeline's architecture, models, and results for future reference** and dissemination.
- Prepare a final report summarizing the project's objectives, approach, results, and conclusions.
- Discuss the report and its contents with the mentor.

Week 12 (August 15 - August 21)

- Finalize and submit the project report.
- Review the project and **Publish blog with the lessons learned during the GSOC period.**
- Plan for future work and possible contributions to the DeepLense project and the wider scientific community.
- Prepare for the final evaluation and feedback from the mentor

My other plans this summer

As my final exams will conclude in early May, I have a considerable amount of time during this summer for professional development. I plan to focus on learning more about model optimization techniques, as well as exploring the open-source development space. I believe that working on this GSoC project would allow me to achieve both of these goals while simultaneously contributing to the Open-source community. Therefore, I plan to dedicate my full attention to this project throughout the entire summer.

Why I am the best candidate for this project

I am confident that I am the best candidate for this project due to my experience in the required prerequisite tasks and my ability to create a flexible training and testing pipeline. I have completed all the prerequisite tasks, including Common Test I, Specific Test II, Specific Test III, and Specific Test V, and have implemented my approach in PyTorch, discussing my strategy in each case.

In addition to completing the prerequisite tasks, I have created a flexible training and testing pipeline that uses PyTorch Timm as a backend, making it easy to test hundreds of computer vision models. The pipeline includes clean and minimal training metrics logging in a model.json file, easy saving and loading of models, and the ability to resume training. The pipeline is also easily extensible with any loss function and optimizers and can be used for classification, regression, and object detection.

Moreover, I have provided a multi-model ablation study for each task and organized plots of various training parameters. I have also made the notebooks easily accessible in Google Colab or any environment, with data downloading and other setups already done in the notebook.

Overall, my experience in completing the prerequisite tasks and creating a flexible training and testing pipeline, along with my ability to learn new things and my passion for astronomy and science, make me the best candidate for this project.

As per the instructions on the ML4SCI website, I have refrained from directly contacting any mentors at this stage. However, I am eagerly looking forward to any response or feedback from the mentors regarding my proposal.