Updating the DeepLense Pipeline

Google Summer of Code 2023 Project Proposal

Saranga Kingkor Mahanta
Institut Polytechnique de Paris
Paris, France (GMT+2)
+33 7 80 78 35 92
saranga.mahanta7@gmail.com

Proposal Abstract

The study of dark matter substructures has shown promise in addressing the open-ended and long-standing challenge of identifying the true nature of dark matter. Strong gravitational lensing is a potential way to learn more about dark matter's basic nature by probing its substructure. Deep learning approaches can properly recognize images with substructure and distinguish WIMP particle dark matter from other well-motivated models including dark matter condensates, superfluids, and vortex substructure. The DeepLense pipeline combines state-of-the-art deep learning models with strong lensing simulations based on lenstronomy. The focus of this project will be updating the previous results of DeepLense with new dark matter simulations (e.g. WDM and SIDM).

Motivation

One of the most pressing concerns in physics today is the identity of dark matter. Even though a number of plausible ideas have been offered in recent decades that might reveal dark matter's true identity, the underlying nature of dark matter has remained a mystery. Weakly Interacting Massive Particle (WIMP) is considered to be one of the most promising candidates that could prove insight into this quest. A promising path to identifying the nature of dark matter is to study the substructure in dark matter. A promising path to identifying the nature of dark matter is to study the substructure in dark matter.

To put gravitational lensing simply, when light from distant galaxies passes through large objects in the cosmos, the gravitational attraction from these things can distort or bend the light. Strong gravitational lensing has yielded promising discoveries in the detection of substructure in heavily lensed quasars, high-resolution observations with the Atacama Large Millimeter/submillimeter Array, and extended lensing. With simulated galaxy-galaxy strong lensing pictures, interesting

findings have recently been produced using supervised machine learning methods for detecting dark matter substructure features.
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. In this project, deep learning models and various architectures will be trained and tested on new dark matter simulations (e.g. WDM and SIDM) in order to update the previous results of classification, regression, anomaly detection, etc. for the DeepLense Pipeline.

I had worked on the same project last year with ML4Sci, and thus, I am familiar with the work culture and expectations (technically, I would be "updating" my own work if selected). I got the opportunity to learn a lot and expand my skill set, thus, I would like to continue with this project this year, because I believe that I will be able to get more things done this time and in a lesser amount of time.

Key Tasks

- Conducting an extensive literature survey on previous works of the DeepLense Pipeline.
- Experimenting with different pre-trained models
    - Implementing the selected network architectures and ensembles in PyTorch
    - Training, fine-tuning, and testing these networks on the provided Model dataset(s) mainly for regression, classification, and anomaly detection
- Updation of the DeepLense pipeline with the new models trained.
- Proper documentation and analysis of procedure, observations, and results.

Project Timeline

- **April 4th - May 4th (Acceptance waiting period)**

    - Revise the entire pipeline created last year.
    - Study about and select potential architectures and pre-trained models.
    - Re-install all required dependencies and set up the development environment in my system like last time.

- **May 4th - 28th (Community Bonding period)**

    - Discuss the project and expectations with the mentors.
    - Consequently, formulate a to-do list of tasks.

- Discuss with the mentors regarding the simulated dataset(s), the possibility and procedure of data augmentation, and foreseen challenges.
- List potential architectures and pre-trained models.
- Planning the stacking and ensembling methodology in the case of usage of multiple models.
- Possibility of using Domain Adaptation.

- **May 29th (Coding begins)**

  - Comprehensive analysis and visualizations of the provided simulated dataset(s):
    - Identification of class imbalance, biases, noise, or any other predictive modeling issues.
    - Pre-processing the data consequently.
  - Create a custom dataset class (in PyTorch) and implement data loaders for the provided data to load the simulated data in a suitable format to feed into the neural network(s).
  - Perform mix-up and data augmentation if possible and necessary.
  - Train and test with different architectures and pre-trained models on the data.
  - Try combinations of model stacking and ensembles to improve predictions if required.
  - Use WandB for logs and hyperparameter tuning later.

- **July 10th - 14th (Midterm evaluation)**

  - For the first evaluation, I plan to complete and present the following tasks:

    - Results and visualizations and what could be inferred from the extensive analysis of the dataset.
    - Methodology of pre-processing the data.
    - Custom dataset class and data loaders to load the simulated data.
    - Fit appropriate neural networks on two-three Model datasets, depending on readiness of the simulations.

- **August 21st - 28th (Final evaluation)**

  - By the time of the final evaluation, I plan to complete the rest of the tasks:

    - Training, fine-tuning potential pre-trained models on all Model datasets made available.
    - Hyperparameter tuning and using different promising techniques (for example, different optimizers, parameterized activation functions, learning rate schedulers, methods for faster training convergence, etc).
    - Trying different combinations of stacking/ensembling best models to further improve predictions.
    - Updating the DeepLense Pipeline with the newly trained models.

- Formulating lucid documentation in parallel.

- **Post GSoC**
    - Further refinement of documentation and presentation of results with regards to feedback from the mentors.
    - Write an article detailing all the work done and the results achieved.

Features and Deliverable Specifications

- High-performing neural network models (for regression, classification, anomaly detection, etc. as required) trained on new simulated data with improved results.
- Weights for the network architectures for future re-use and further fine-tuning.
- Proper documentation and tutorial notebooks for the usage of the model and code.
- Publication-ready graphical material like plots, graphs, tables, etc. to display the benchmarking results on the simulated data.
- An article about the whole project, the approaches, and entire methodology

Summer Availability / Commitment

I am currently in the first year of my Master degree in Data and Artificial Intelligence from IP Paris. The exams of the remaining courses will end by mid-April. Hence, I shall be able to devote my time and commit to the project.

References

- Alexander, Stephon, Sergei Gleyzer, Hanna Parul, Pranath Reddy, Michael W. Toomey, Emanuele Usai, and Ryker Von Klar. "Decoding dark matter substructure without supervision." arXiv preprint arXiv:2008.12731 (2020).
- Alexander, Stephon, Sergei Gleyzer, Evan McDonough, Michael W. Toomey, and Emanuele Usai. "Deep learning the morphology of dark matter substructure." *The Astrophysical Journal* 893, no. 1 (2020): 15.
- Alexander, Stephon, Sergei Gleyzer, Pranath Reddy, Marcos Tidball, and Michael W. Toomey. "Domain Adaptation for Simulation-Based Dark Matter Searches Using Strong Gravitational Lensing." *arXiv preprint arXiv:2112.12121* (2021).

About Me

Link to my CV

Relevant Experience

1. **GSoC 2022 contributor (ML4Sci)**
   - For the same project, "Updating the DeepLense pipeline"

2. **Research Intern @ Aix-Marseille University in collaboration with NIT Silchar**
   - Worked under the joint guidance of Dr. Benoit Favre and Dr. Partha Pakray
   - Formulated novel methods for efficiently evaluating summaries generated by Abstractive Text Summarization systems using Textual Entailment
   - Compared summaries generated by a baseline Seq2Seq LSTM model having Attention mechanism with pre-trained BART-based text summarizer using proposed evaluation metric

3. **Data Science Intern @ Devopedia**
   - Performed web scraping using BeautifulSoup and used Feed-forward Neural Networks to automatically identify Title, Author, and Year (of publishing) entities from any webpage
   - Constructed an end-to-end system using Python that generates Reference Citation strings in the Chicago Manual Style format from URL inputs

Publications

- **"Exploiting Cepstral Coefficients and CNN for Efficient Musical Instrument Classification"**
  Saranga Kingkor Mahanta, Nihar Jyoti Basisth , Eisha Halder, Abdullah Faiz Ur Rahman Khilji, Partha Pakray
  Paper under review at Evolving Systems

- **"COVID-19 Diagnosis from Cough Acoustics using ConvNets and Data Augmentation"**
  Saranga Kingkor Mahanta, Shubham Jain, Darsh Kaushik, Koushik Guha
  Paper presented at IEEE International Conference on Advances in Computing and Future Communication Technologies 2021 - MIET Meerut, published in IEEE Xplore

- **"Deep Neural Network for Musical Instrument Recognition using MFCCs"**
  Saranga Kingkor Mahanta, Abdullah Khilji, Partha Pakray
  Paper Published in Computación y Sistemas: Vol 25, No 2 (2021)

Vision and Goals

My long-term vision is manifold. I wish to understand the underlying nuances of neural networks and how they automatically extract the essential features from data. Gaining such insights will help in building more robust models that will tackle adversarial examples, which are inputs that are purposefully designed, or adulterated to cause a model's predictions to be incorrect, yet look valid to a human. Immunizing models against adversarial attacks is not only a vital task for security measures, but solving this problem would also shed new light on the model-based optimization problem. This should ultimately result in the creation of novel functions, the outcomes of which would be revolutionary across diverse fields. Another intelligent task that interests me is zero-shot learning, training a model to predict examples from unprecedented classes. This is a trivial and frequently occurring scenario for humans, to encounter newfound objects, yet be able to draw relations about them from pre-acquired knowledge and experience. That is true intelligence.

In a broader sense, I am always looking for exciting areas where Deep Learning techniques could be applied to produce insightful findings.

Technical Strengths

- ***Computer Languages***: Python, C, C++, HTML, CSS
- ***Libraries and Frameworks***: PyTorch, TensorFlow, NumPy, Pandas, Flask, Librosa, Scikit-learn
- ***Domain Interests***: Deep Learning, Artificial Intelligence, Computer Vision, NLP, Generative Learning