Self-Supervised Learning for Strong Gravitational Lensing

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

Researching the substructures of dark matter holds potential in solving the enduring and unresolved problem of discovering the actual essence of dark matter. One possible method to gain more insight into dark matter's fundamental nature is by utilizing strong gravitational lensing to examine its substructure. Advanced deep learning methods can effectively identify images featuring substructure and differentiate between WIMP particle dark matter and other plausible models such as dark matter condensates, superfluids, and vortex substructure. When it comes to strong gravitational lensing images, supervised classification can be difficult because there are usually only a few examples of each class. To overcome this challenge, self-supervised learning has been shown to be more effective than supervised machine learning models, particularly when there is a scarcity of data labels. Additionally, self-supervised learning can take advantage of vast unlabelled datasets to create valuable representations. This project's objective is to devise self-supervised learning strategies utilizing Transformers and hybrid models for strong gravitational lensing data and examine equivalent techniques for other strong lensing tasks.

Motivation

One of the biggest issues in physics currently is figuring out what dark matter is. Despite several plausible ideas being proposed over the past few decades, the true nature of dark matter remains a mystery. Weakly Interacting Massive Particle (WIMP) is considered one of the most promising candidates for shedding light on this mystery. A potential way to understand dark matter better is by studying its substructure.

This can be achieved through strong gravitational lensing, which occurs when light from faraway galaxies passes through massive objects and gets distorted. Recent discoveries in heavily lensed quasars and high-resolution observations with the Atacama Large Millimeter/submillimeter

Array have shown promise in identifying substructure. Additionally, supervised machine learning methods have produced interesting findings in detecting dark matter substructure features using simulated galaxy-galaxy strong lensing pictures. Deep learning methods have the potential to accurately identify substructure in images and distinguish WIMP particle dark matter from other well-supported models, such as vortex substructure of dark matter condensates and superfluids.

The aim of this project is to concentrate on the development of self-supervised learning methods that use Transformers and hybrid models for analyzing strong gravitational lensing data. Additionally, the use of equivariant techniques for other tasks related to strong lensing in the context of self-supervised learning will be explored. Autoencoders and its variations (Variational Autoencoders, Adversarial Autoencoders, etc) have been proven to be an effective method in self-supervised learning, and thus, it would be interesting to implement those models in this context as well.

I had worked on the project, "Updating the DeepLense Pipeline" last year with ML4Sci as a GSoC contributor, and thus, I am familiar with the work culture and expectations. I got the opportunity to learn a lot and expand my skill set, thus, I would like to continue working with them. I believe that I will be able to get more things done this time and in a lesser amount of time. Furthermore, I will also be able to guide the contributor who gets selected for "Updating the DeepLense Pipeline" project this time.

Key Tasks

- Conducting an extensive literature survey on transformers, hybrid models (transformers + CNNs), autoencoders, and other models used in self-supervised learning.
- 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 dataset(s)
- Incorporation with the DeepLense pipeline.
- Proper documentation and analysis of procedure, observations, and results.

Project Timeline

- **April 4th - May 4th (Acceptance waiting period)**
  - Conduct an extensive literature survey on previous and related works.
  - Study and discover potential pre-trained transformer models.
  - Install all required dependencies and set up a development environment in my system properly.

- **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 transformer architectures and pre-trained models.
    - Planning the stacking and ensembling methodology in the case of usage of hybrid model architectures.
    - Learning about equivariant techniques.
    - 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 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
    - 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.
        - Train and test a few transformer models.

- **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 self-supervised learning models on all the provided datasets
            - Transformers and hybrids (transformer + CNN)

        - Autoencoders (VAE, AAE, etc.)
        - Equivariant techniques and networks
    - Hyperparameter tuning and using different promising techniques (for example, different optimizers, parameterized activation functions, learning rate schedulers, methods for faster training convergence, etc).
    - Use Domain Adaptation if possible and relevant to the case.
    - Trying different combinations of stacking/ensembling best models to further improve predictions.
    - Incorporating the work with the DeepLense Pipeline.
    - 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 self-supervised learning on Strong Gravitational Lensing data.
- 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 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