Deep learning for Dimensionality Reduction

Mentors: Jeremy Bailin (University of Alabama)
Sergei Gleyzer (University of Alabama)
Jacob Morgan (University of Alabama)
Jianghao Huyan (University of South Carolina)
Varsha Kulkarni (University of South Carolina)

Dimensionality Reduction for studying diffuse circumgalactic Medium

Personal Details

Name : Marco Gullotto

Email : marcogullotto97@gmail.com

University : Politecnico di Torino

Location and time zone : Italy, CEST

Degree : Master's Degree in Data Science and Engineering

Expected Graduation Date: October 2021

GitHub account : MarcoChain (github.com)

How many hours will you work per week : up to 20hrs per week

Abstract

Machine learning algorithms usually handle thousands or even millions of features for each training instance. This process can be very slow and can take months to find an appropriate solution. To overcome this problem, we can represent our high-dimensional data into a lower-dimensional space so that the low-dimensional representation retains some meaningful properties of the original data.
Starting from the state-of-the-art, it is possible to use different approaches as PCA, LLE, LDA, but my idea is to use a deep learning model as feature extractor. In particular, I could try to adapt the work of Hertel et al. (https://arxiv.org/pdf/1710.02286.pdf) and Liu et al. (https://arxiv.org/pdf/1904.04447.pdf) to the simulated quasar absorption spectra dataset.
In principle, the idea is very simple: train the machine learning model to predict the right classes and then remove the last fully connected layer. What remains is a trained feature extractor that works exactly for the quasar absorption problem. So, while

shallow methods work the same for all datasets, this new algorithm adapts itself to the dataset and produce optimal filters for this specific task.
At the end of this process, we can apply any type of classifier that we want, from a deep one to a shallow one like an SVM or a random forest. In this way, we can also get more interpretable results than using only a deep learning approach.

Goals

1. Review on circumgalactic medium (CGM);
2. Implement machine learning-based dimensionality reduction models applicable to quasar absorption spectra datasets
3. Add-on if extra-time is left: detailed comparison with state-of-the-art models.

Timeline

The workload is divided according to the following timeline. I plan to have a weekly summary with the mentor.

Period	Task
Community bonding [May 17, 2021 - June 7, 2021]	• Discuss the plan of the project with the mentor;
Week 1 and 2 [June 8 2021 - June 20 2021]	• Review on CGM inference with deep learning; • Individuate what model can be applied to this problem and start working on that
Week 3 and 4 [June 21 2021 - July 4 2021]	• Implement and prototype the new models on a small batch of data; • Check if the novel approach is working fine or need to be modified immediately.
Week 5 and 6 [July 5 2021 - July 16 2021]	• Trials on the entire datasets; • Check for performance issues.
Week 7 and 8 [July 19 2021 - August 1 2021]	• Additional time to debug; • Add-on implementation if the project is already working fine.

Week 9 and 10 [August 2 2021 - August 15 2021]	• Start working on the project report.
Submission period [August 16 - 23 2021]	• Buffer period for final changes and revision and for writing the project report

About Me

Something about me? I am Marco Gullotto, and I am a student of the Master's Degree in Data Science and Engineering at the Polytechnic of Turin. Previously, I graduated in Computer Engineering at the same University. I have almost finished my Master's Degree exams but, to graduate in the best possible way, I decided to write a thesis based on an internship, in order to truly take possession of what I could write, having lived it first. During the University years, indeed, I attended many courses (such as Machine Learning and Deep Learning, Machine Learning for IoT, Distributed architectures for big data processing and analytics, Data Management and Visualization and Network Dynamics) and took part in some projects, which make me a good candidate for this position. For example, I am a member and one of the founders of the MALTO (MAchine Learning at Politecnico di TOrino) team. With this group of people, we tackle complex machine learning problems in subgroups of three or more people. We do a lot of research to understand what is the state of the art of the model to address the problem. So, we use different tools to try and fix them, like Keras, TensorFlow, Pytorch, Pyspark and MapReduce. Working as a team teaches you how to grow, both personally and as a group, and I want to commit myself and work hard to grow with you. Furthermore, in my last year of Bachelor Degree, I participated, with some PhD students, in a research activity for the control of Turin air quality. During this project, we design a new way to collect a large amount of data (using some IoT devices), make some predictions (solving a regression task using Python) and display them correctly (using Tableau). “Perfection” does not exist but my goal is to be as close as possible to it. It doesn't matter if it's a housework, a homework assignment or whatever, I give my best in every opportunity to get the best result I can produce. I am humble, but stubborn: I am able to work with head down (but always chin up!) and I always want to go beyond my limits to see how far I can go. I am good at what I do, and I think I can improve even more with you. I wish this could be my new challenge. A decidedly ambitious and exciting challenge
Sincerely,