GOOGLE SUMMER OF CODE 2021

ML4Sci-CGM

Dimensionality Reduction for Studying Diffuse Circumgalactic

Mentors:- Jeremy Bailin, Jacob Morgan, Sergei Gleyzer, Jianghao Huyan, Varsha Kulkarni

Shiven Tripathi

BTech, Electrical Engineering
Indian Institute of Technology, Kanpur
Kanpur - 208016, India

Abstract

Observing the absorption spectrum from quasars offers insights into the composition, temperature, the density of the circumgalactic medium from which these emissions pass. This project's main contribution would be to develop a machine learning pipeline to reduce the dimensionality of such large spectral datasets, making them easier to observe and classify for humans. A complete end to end pipeline would also be explored for fully automatic analysis of spectrum datasets, leveraging deep learning-based methods.

Motivation

Quasars are galaxies with highly active nuclei, emitting energy in a wide range of the electromagnetic spectrum. Due to the huge magnitude of radiation they emit of high energy, astrophysicists are able to observe them, making quasars the most distant yet observable objects. These observations from a quasar can be regarded as light observation from a point-like source (since it is very distant). When this light crosses gas clouds (on its way to earth), it gives some specific wavelengths of energy to these atoms, resulting in an absorption spectrum for earth observers. Of course, due to the expansion of the universe, all galaxies appear to be moving away from us, causing a red shift in all wavelengths we observe. Particularly important to us is the diffuse circumgalactic medium, which is a gas cloud surrounding the galaxy.

If we are able to identify component ions and atoms in the gas cloud, we would have an idea of the history of the galax. For example, "The MgII absorption doublet is particularly useful, since its wavelengths λλ 2796, 2803 can be detected in quasar optical spectra to trace galaxies in the intermediate redshift range ($0.3 < z < 2.5$), which covers the peak of global star formation history in the universe"[2]

Current astrophysical datasets are too voluminous to allow human aided identification of spectrums from quasar absorption line observations. At this juncture, this project would aim to provide a fully functional pipeline for analysing these datasets without or with significantly lesser human hours. If appropriate dimensionality reduction methods are applied, we can also aim for human aided interpretability in results obtained for a model since we have managed to drastically reduce the data size.

To effectively apply deep learning to problems, we need large amounts of data, making it suitable for solving astrophysics problems in which large numbers of observations are readily available to train on. Directly for the task of dimensionality reduction, we have effective methods in the form of Deep Autoencoders, which are able to learn complex mappings between features, making a reduced dimension latent space possible.

Background

"The Sloan Digital Sky Survey (SDSS; York et al. 2000) contains a large sample of more than 400,000 quasar spectra well suited for the investigation of MgII absorber systems." [2]

Traditional detection methods have been applied to this dataset with varying success. A significant bottleneck in the analysis is that these methods require a lot of computational resources, taking up time and also are not fully automated requiring human expert intervention.

To keep pace with subsequent such data releases we need fast and reliable methods which are able to either eliminate or drastically reduce human intervention. Following goals for the project have been defined considering, few such advances in the field like [3].

Goals and Implementation Details

In the proposal, I have tried to include multiple possible strategies which would be implemented in the GSoC period. I believe the optimal choice would be one which is decided on deliberation with mentors and the exact scope of the problem, datasets etc., and additional goals have been marked as 'stretch goals' which would be completed if main goals have been finished early.

Dimensionality Reduction (and Classification) using ML (Main Goal):

Previous methods have explored use cases involving Principal Component Analysis. By separating spectrums into an orthonormal basis, we obtain a combination for minimal error[4]. While these simple techniques are worth exploring, we should not expect to see significant information being retained in the dimensionality reduction process owing to the non-linear mapping between features. This can be easily tested by checking correlation across feature space.

Thankfully, we can turn to deep neural network architectures which are able to act as complex function approximators, learning nonlinear function mapping. One such architecture, a deep autoencoder can be trained to learn the $f(x) = x$ map. We start with an encoder - bottleneck - decoder model and on proper training, we can use the encoded input (to a lower dimension) as a lower-dimensional representation of the input features.

Another advantage of deep autoencoders which I would want to test is 'removal of noise'. If the deep autoencoder is trained using a simulated noise-free spectrum, we can be assured that for observed data, we will have high reconstruction losses for 'noisy' data on which the

model was not trained on. After that, it is the simple matter of finding the appropriate threshold which would filter all the noise from our absorption spectrums.

I plan to experiment with both kinds of methods, ML-based like PCA and DL Autoencoder methods. There are tradeoffs to explore in both and the right choice would be made after consulting with the mentors.

We can extend this work of dimensionality reduction to the classification of spectra. From a lower-dimensional space, it becomes easier to compare our reduced spectra encodings to a template encoding since we have avoided the curse of dimensionality in a lower-dimensional latent space generated by encodings/PCA.

End to End Deep Learning pipeline (stretch goal, baseline in GSoC period, improvements later) :

Previous work focussed on first extracting meaningful features from the dataset by reducing to a lower dimension and then using that to compare with templates available to us of identified spectrums. We can also aim for a complete end to end pipeline as described in the following. The main component of Convolutional Neural Networks in both of these has a proven record on complex image tasks and would be suitable to our use case of astrophysical data since we have large training samples available.

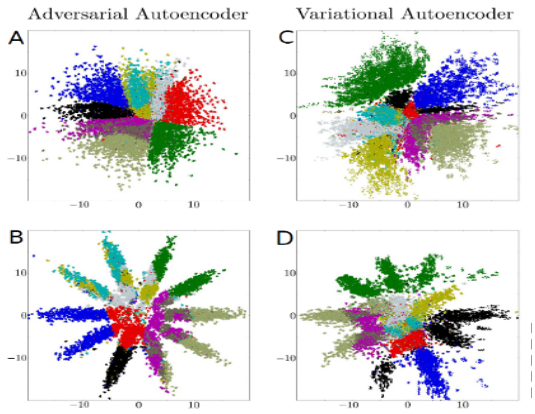
1. [1904.12192] Identifying MgII Narrow Absorption Lines with Deep Learning
    a. Uses Convolutional Neural Networks to extract relevant images from spectrums
    b. Uses the entire spectrum image input for a 10 layered Deep neural network
    c. Can be easily implemented using libraries like Keras/Tensorflow as the authors did
    d. Obtained accuracy of ~94% and took 9/50000 seconds per quasar spectra (in batch).

These methods can be further improved by using the sliding window technique. Since spectrum lines for elements useful to us occupy a small region as compared to the entire spectrum, we can use sliding windows to the only test for atoms in specific proposals, reducing the effect of noise and latency for inference.

Interpretable models like VAE and AAE (stretch goal, baseline in GSoC period, improvements later):

The final purpose of dimensionality reduction is not to only reduce the number of dimensions of the data but to reduce this number of dimensions while keeping the major part of the data structure information in the reduced representations. For these two reasons, the dimension of the latent space and the “depth” of autoencoders (that define

degree and quality of compression) have to be carefully controlled and adjusted depending on the final purpose of the dimensionality reduction.
I aim to initiate work on this by establishing baselines on classification accuracy using an Adversarial autoencoder model. I expect to see a clear demarcation between the representation of encoding by the constituting spectrums from the sample.



[VAE and AAE]

Plots on encodings of AAE and VAE, trained on MNIST dataset, fitted to a) 2D Gaussian b) 10 2D Gaussians

I expect to find similar, clear and interpretable distinctions in the handling of different spectrum samples.

Deliverables by Evaluation

Due to the high volume of applicants and hence, limited interaction with mentors, a priority list could not be decided for the deliverables. Therefore multiple deliverables are outlined below, with a few being completed after a discussion on priorities with the mentors in the community bonding period.

I. First Evaluation

A. Dimensionality Reduction using ML

i. Dimensionality Reduction

a) [Plots, documentation, Evaluation scripts] for experimenting on Statistical methods and Deep Neural Network-based methods for encoding accuracy.

b) Plots would be obtained, indicating of reconstruction possible from lower dimension space

c) I aim to be thorough in this component, trying all possible solutions for optimal dimensionality reduction

ii. Classification

a) [Plots, documentation, Evaluation scripts] for experimenting on classification with and without dimensionality reduction

b) The goal at this step would be to achieve similar levels of classification accuracy after dimensionality reduction compared to original data.

c) I expect this goal to continue for a longer time and would be submitted after the first evaluation once all experiments have been performed.

At this stage submission of deliverables for the dimensionality reduction task would be a **PASS**.

II. Final Evaluation

A. Classification

i. Continue with work left before evaluation and compile results

ii. Python scripts would be submitted to reproduce results

iii. Findings for optimal dimensionality reduction and classification design choices would be presented here

B. End to End Classification

i. Scripts for data loading and processing in the pipeline

ii. Trained models, training scripts, evaluation scripts

iii. Results on evaluation, documentation of best practices

C. Interpretable Dimensionality reduction

i. Scripts for data loading and processing in the pipeline

ii. Trained models, training scripts, evaluation scripts

iii. Results on evaluation, documentation of best practices

iv. Clear indications, with plots and analysis of activations of model layers for the distinction between encodings for different spectra input.

v. Focus on interpretability and possibly explore noise removal, by thresholding on high reconstruction loss

At this stage submission of deliverables for classification task and baseline model for one of the stretch goals would be a **PASS**.

Timeline

Community Bonding Period (May 17- June 6)	1. Learn about code publishing practices, model training platforms, computing resources available etc. 2. Conduct a literature review for dimensionality reduction to present to mentors for implementation in the coding period 3. Discuss and deliberate with mentors to finalise deliverables incorporating earlier literature review 4. Familiarise me with existing codebases and datasets relevant to the project.
Week 1 (June 7 - June 13)	**Dimensionality Reduction** 1. The first week will focus on traditional statistical approaches 2. Setup dataset, apply all preprocessing, have a functional data pipeline 3. PCA and other methods would be tried, observing changes with parameters
Week 2 (June 14 - June 20)	**Dimensionality Reduction** 1. Deep autoencoders started this week 2. Write code for training, evaluation etc. 3. Initiate model training
Week 3 (June 21 - June 27)	**Dimensionality Reduction** 1. I expect to train several models, before evaluating results on the best one 2. Benchmarking would be done for reconstruction loss, time for training, inference etc. **Additional:** Use deep autoencoders for noise removal
Week 4 (June 28 - July 4)	**Classification**

	1. Prepare data loaders and pipeline for classification 2. Write code for a classification model, compile to check everything is working
Week 5 (July 5 - July 11)	**Classification** 1. Complete training one model on the non-reduced dataset 2. This will take a lot of time but we need a baseline to compare reduced dimension results **Documentation for the first evaluation** 1. Daily Work Log 2. Meeting minutes 3. Codebase to reproduce results on dimensionality reduction 4. Slide Deck giving an overview of work completed
Week 6 (July 12 - July 18)	**First Evaluation** (and buffer to finish documentation for first evaluation)
Week 7 (July 19 - July 25)	**Classification** 1. Explore models suitable for features obtained after statistical dimensionality reduction methods 2. Complete benchmarking on these methods for classification accuracy
Week 8 (July 26 - August 1)	**Classification** 1. Explore models suitable for features obtained after deep autoencoders 2. Complete benchmarking on these methods for classification accuracy
Week 9 (August 2 - August 8)	**(Initiate Stretch goals, complete one baseline model)** OR **(Buffer to catch up)**
Week 10 (August 9 - August 16)	**Documentation for final evaluation** 1. All the docs, slides, codes etc. as submitted in the first evaluation updated for all work 2. Project Report with proper typesetting and all plots of the results

	3. Discuss if work could be presented at a relevant conference, and prepare an 8 Page research paper for that
Week 11 (August 17 - August 23)	**Final Evaluation** (and buffer to finish documentation for final evaluation)

About Me

My Contributions

I. Evaluation tasks

1. Here is the GitHub repository containing evaluation tasks: ShivenTripathi/Dimensionality-Reduction-CGM
2. Here is the Technical Report explaining implementation and results: CGM_Technical_Report.pdf

II. Literature Review for deliverables

- In the process of identifying suitable methods to solve the project goals, I have completed a thorough literature review of different possible approaches which could be applied. Many such references to literature in the field have been included in the references section.
- I had started with identifying methods to reduce dimensionality for datasets of quasar spectra and moved on to end-to-end CNN based architectures. I also studied, AAE and VAE, which are autoencoders, providing us dimensionality reduction, along with interpretable definitions for the reduced dimension encodings.

Research Experience

I have been working as a research associate under Prof. Nischal K Verma at IITK, working on using classical and deep learning computer vision techniques to solve a robotics problem. Previously, I had competed in a National Machine Learning Hackathon, in which our team worked on the problem statement of building a fashion portfolio using AI, in which our group stood in the National Top 3. Apart from this, I have worked on projects in different fields of ML, like NLP and RL, which have given me good experience of deep neural networks to succeed in this project.

Other Programming Experience

1. ShivenTripathi/ConversationalRobot

This project aimed to make a Talking bot that can pay attention to the user's voice and generate meaningful and contextual responses according to their intent. Deep Learning models like Deep Speech 2, GloVe, Seq2Seq with Attention and Topic Aware Seq2Seq were implemented as part of an end to end pipeline for speech recognition to response generation.

2. PrsMittal/humanoid_ws [Partial Repository of CV libraries of Team Humanoid] High performing CPP library to perform computer vision tasks for a Humanoid Robot, completed in remote collaboration with my peers at IITK. Team Humanoid-IITK competes in FIRA RoboWorld Cup, building humanoid robots capable of archery, wrestling etc.
3. ShivenTripathi/ATLAS-Autoencoders
   Dimensionality reduction techniques were employed in the ATLAS dataset to reduce the dataset's size using deep autoencoders. Plots showing reconstruction loss added for visualisation of performance.

Motivation for choosing this project

Nearly completing two years of undergraduate coursework at IIT Kanpur, I recently had the opportunity to work in research with faculty in fields at the intersection of AI, ML and Computer Vision. I am particularly enthusiastic about AI due to its unreasonable effectiveness at solving problems in areas seemingly unrelated to it. From being able to model neurons(biological) as Hopfield Networks, achieving superhuman performance on tasks makes AI a very vibrant and exciting field in which there is always something more to learn.
About a year ago, I came across a fascinating talk by a CERN scientist about using deep learning over the enormous data generated at ATLAS, in which they use sequence-based models (RNNIP) for Jet Classification. This piqued my curiosity on how seemingly distinct fields of science could come together to solve a problem. I believe working on this GSoC project would be similarly very fascinating for me.

Working Hours and other commitments

My semester vacations would start on May 14 and end on July 27. The official GSoC 2021 period is from May 17 to August 31. During the break, I would be available for work and be online on slack and Zoom/Meet any time I am awake. However, when the semester resumes, the working hours would be determined by the semester's schedule. I may not be available for Zoom/Meet, etc. for a couple of hours on a few days while I'm in a lecture. However, I would still be able to devote 40-50 hours a week easily.

Other than this project, I have no commitments/vacations planned for this summer. I shall keep my status posted to all community members especially mentors, on a weekly basis and maintain transparency in the project.

Note: Due to the coronavirus outbreak affecting the academic calendar, there might be a slight change in the vacation period available. I would be in constant with the mentors regarding this and update them as soon as I have the final schedule.

Contact Details

Name: Shiven Tripathi

Phone: (+91) 9354043254

Email: shiventripathi1@gmail.com , shiven@iitk.ac.in

University: Indian Institute of Technology, Kanpur

Github: ShivenTripathi

Resume: cv.pdf

LinkedIn: in/shiven-tripathi

Time Zone: India (UTC+5:30)

Platform:

- OS: Ubuntu 18.04/ Ubuntu 20.04
- Editor: VSCode
- Version Control: Git

After SoC

I wish to continue with the stretch goals of the project after the official GSoC period. I expect that within the GSoC period, I would have set up a baseline AAE/VAE model but would love to continue exploring themes of interpretability of the encodings and more advanced end to end models for classification in the future. I feel that towards a strong application for doctoral studies, this research demonstrates strong principles of theory applied to critical real-time situations and would be invaluable to me.

References

1. 4. Quasar absorption lines.
2. [1904.12192] Identifying MgII Narrow Absorption Lines with Deep Learning
3. [1808.09955] QuasarNET: Human-level spectral classification and redshifting with Deep Neural Networks
4. https://iopscience.iop.org/article/10.3847/0004-6256/152/6/205/pdf