Self-Supervised Active Learning guided Detection system (SSALD)

Aniket Nath
School of Physical Sciences
National Institute of Science Education and Research,Bhubaneswar, HBNI
aniket.nath@niser.ac.in

Diptarko Choudhury
School of Physical Sciences
National Institute of Science Education and Research,Bhubaneswar, HBNI
diptarko.choudhury@niser.ac.in

Abstract

In our project, we try to find, rarely occuring dual Active Galactic Nuclei (DAGN), from the Sloan Digital Sky Survey (SDSS) Database. SDSS has over 200 million galaxy samples, but till now a few hundred (117, [8]) have been detected. Classical and simulation studies gives a strong indication that a large number of DAGNs might be present in the observable universe; but they are hard to detect due to their rareity. Our work, tries to exploit machine learning techniques (weakly supervised) to find such occurances of DAGNs. We have tried to use Self-Supervised and Active Learning to build the our algorithm for the corresponding task.

1 Introduction

In our low redshift universe, the main mechanism for growth of galaxies is through mergers of galaxies [7]. Galaxies usually have a Super Massive Black Hole (SMBH) , at it's center, which provides the gravitational potential for the substructures to form. When two galaxies merge, their nuclear bulge hosting the SMBH, come closer and interact with each other, forming dual or even in certain cases multiple nuclei galaxies. Often, merger process triggers accretion disk around the SMBHs, leading to the formation of an Active Galactic Nuclei (AGN) [5]. In certain cases, Dual AGNs (DAGN) are also observed in such systems, and the interaction of two such nuclei is an excellent model to understand the merger processes, accretion and other dynamics of galaxies. In the observable universe, we are constrained by the rareity of such samples. Even though, merger events are quite common, implying that there should be quite high number of Dual Nuclei systems, but the detected number of such systems are quite low ([1],[3]). The discovery of such systems are serendipitious, and the need for more samples to understand the underlying astrophysics is need of the hour. In order to automate this process of detection, we attempt to build a machine learning algorithm, which is mainly guided by Self-Supervision and Active Learning. The details have been explained further in section (3)

2 Related Works

There has been attempts to automate the search of Dual Nuclei Galaxies [2], where a Graph Boosted Hill Climbing (GOTHIC) algorithm was used to find pairs of nuclei in a sample image from Sloan Digital Sky Survey (SDSS) dataset. GOTHIC worked on finding galactic nuclei by finding bright

Preprint. Under review.

spots through a Boosted Hill Climbing Method. The method is slow and has many edge-cases which must be excluded for proper classification. Moreover, GOHTHIC did not address the spectroscopic data. Confirming a DAGN without proper spectroscopic data is very difficult hence the method is error-prone.

There have also been works in astrophysics where self-supervised learning was used to pre-train a model, which was later fine-tuned for red-shift estimation and galaxy morphology classification. The paper showed promising results [4]. The paper used the contrastive loss approach to tackle the problem. They achieved state-of-the-art results when they fine tuned their self-supervised backbone in supervised tasks. This shows that finding DAGNs using a Self-supervised backbone and later fine-tuned by active learning technique should not be difficult.

3 Methodology

We aim to use an Active Learning guided self-supervised algorithm to solve our problem. In the following sections, we discuss our strategy in a step-by-step fashion.

3.1 Supervised Learning

Our task of finding DAGNs falls mostly in the domain of computer vision. In recent years the machine learning fraternity has seen a lot of progress in terms of raw classification performance when it comes to images. We have came a long way from AlexNet[9] to Vision-Transformers[10] to achieve super human level classification and image analysis performance. Still, the Supervised Learning domain is unsuitable for our task due to the lack of labelled data. Moreover, there is little prospect of knowledge transfer due to the fact that the two domains are completely different.

3.2 Unsupervised and Semi-supervised Learning

Unsupervised Learning works best when there is a complete lack of labelled data. Unsupervised Learning tries to find inherent patterns in the data and understand its morphology from such patterns. Our problem can be partly solved using K-means clustering, but K-means struggles when directly images are fed into them. Hence we need a feature extractor backbone to extract meaningful features and to escape from the curse of dimensionality.

Training of this feature extractor backbone needs some intermediate tasks. The backbone can also be trained using a Semi-Supervised Algorithm. Algorithms such as FixMatch have revolutionised the field of Semi-supervised learning. We are still investigating the prospect of semi-supervision, but we have kept it on lower priority because of the lack of labelled data.

3.3 Self-Supervised Learning

Yann LeCun first suggested the idea of Self-Supervision(1990 and mid-2020) [12]. We are more interested in getting a meaningful feature extractor, and also since our work primarily focuses on finding DAGNs rather than making a DAGN classifier, we have chosen to use Self-Supervised Learning. The inspiration for using Self-Supervision primarily comes from contrastive learning in machine learning. We are using VICReg [6] to train our feature extractor backbone. We chose VICReg primarily due to its innovative loss function which makes the algorithm robust towards informational collaps. Moreover VICReg does not uses Contrastive Loss which allows it to work well even without a large number of contrastive pair which allows for a much lower memory footprint.

3.4 Active Learning

Active Learning is a paradigm in Machine Learning which tries to increase the data efficiency of a learning algorithm by allowing the algorithm to interact with an oracle (in this case, a data labeller) to help it find the best decision boundary. The active learning paradigm has three main components: an oracle, a model and a data sampler. The data sampler finds all the samples most difficult for the model to classify or which lie in the region of disagreement. The oracle then labels them and feeds them to the model. Once the model is trained on these samples, the region of disagreement becomes narrower, and the decision boundary considerably improves.

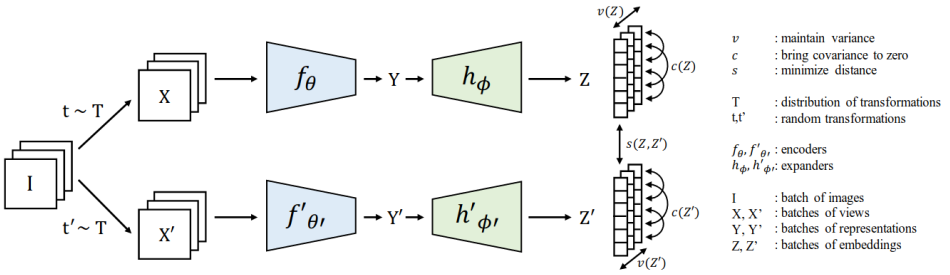

Figure 1: *Schematic representation of the VICReg algorithm [6]*

The main inspiration of using active learning is its high data efficiency, making the data labelling task much more efficient and fast. Moreover, throughout the active-learning phase, we will collect many new DAGN samples in our region of disagreement. This will, in turn, help in our DAGN search and, at the same time, will make the algorithm more robust.

We are still experimenting with the Active Learning algorithms. We have not finalised our choices for the sampler or the active learning algorithm. We are working with Maximum Likelihood Loss and trying to understand and read more about active learning. We aim to write a custom loss by ourselves to solve the problem.

4 Results and Dicussion

The results are preliminary since sufficient work could not be done due to a tight academic schedule. Github Link - `https://github.com/dc250601/Project460`

4.1 Data-Extraction

Extracting data from the SDSS database is difficult due to slow servers and a lack of high-speed transfer options. Since neither of us is rigorously trained in Astronomy, it took us some time to understand how the database works. The data mining programs are almost completely written, and soon we will deploy them to extract over 10 Million samples from the SDSS database.

The normalisation of astronomical images is an important step in data preprocessing. We are exploring a number of techniques and trying to find which works the best for us. Currently we are using the method developed by Lupton[11] as a pre-processing technique.

Spectroscopic data is also something very crucial for astronomical data analysis. We are trying to download them from the database and integrate them into the already present pipeline to increase the quality of the input data.

Currently, we are working with a dataset of only 1 Million samples. We will soon scale up the data size once our algorithm and pipeline is robust enough.

4.2 Self-Supervised Training

We have re-written the VICReg algorithm. The model is currently written in PyTorch and is trained on GPU(Single GPU). Work under progress to speed up the training process using Distributed Training. Moreover we are also writing our model so that it can be run on Google TPUs(V3-8 and V2-8). Preliminary results from training the model are quite convincing. Since VICReg uses 3 different(Variance, In-variance and Co-Variance) losses to make an abstract loss, we need an accurate measure to understand the training progress. Our experiments found that slightly changing the Augmentation Space can distort the loss range considerably. To tackle this problem, we will need to define a new metric. We are experimenting with several metrics but have not found anything of our use till now.

The results of training the self-supervised backbone are given in appendix A

4.3 Active Learning

We have just scratched the surface of active learning. More literature surveys and experimentation are needed to achieve anything substantial. We are working with Maximum Likelihood Loss Function and trying to understand the region of disagreement associated with it. We are trying to come up with a loss that will make the transition from the self-supervised domain to the active learning domain much smoother.

5 Upcoming work and plans

- Running the data miner to mine over 10 Million samples from the SDSS database.
- Finding a way to include spectroscopic data in our pipeline.
- Defining a metric to understand the VICReg training process better.
- Defining a custom loss to make the transition from SSL to Active Learning smoother.

References

[1] A. Stemo et al., "A Catalog of 204 Offset and Dual Active Galactic Nuclei (AGNs): Increased AGN Activation in Major Mergers and Separations under 4 kpc," ApJ, vol. 923, no. 1, p. 36, Dec. 2021, doi: 10.3847/1538-4357/ac0bbf.

[2] A. Bhattacharya, N. C. P., M. Das, A. Paswan, S. Saha, and F. Combes, "Automated Detection of Double Nuclei Galaxies using GOTHIC and the Discovery of a Large Sample of Dual AGN." arXiv, Nov. 14, 2022. doi: 10.48550/arXiv.2011.12177.

[3] K. Rubinur, M. Das, and P. Kharb, "Searching for dual active galactic nuclei," Journal of Astrophysics and Astronomy, vol. 39, p. 8, Feb. 2018, doi: 10.1007/s12036-018-9512-y.

[4] M. A. Hayat, G. Stein, P. Harrington, Z. Lukić, and M. Mustafa, "Self-Supervised Representation Learning for Astronomical Images," ApJL, vol. 911, no. 2, p. L33, Apr. 2021, doi: 10.3847/2041-8213/abf2c7.

[5] I. Shlosman, M. C. Begelman, and J. Frank, "The fuelling of active galactic nuclei," Nature, vol. 345, no. 6277, Art. no. 6277, Jun. 1990, doi: 10.1038/345679a0.

[6] A. Bardes, J. Ponce, and Y. LeCun, "VICReg: Variance-Invariance-Covariance Regularization for Self-Supervised Learning." arXiv, Jan. 28, 2022. doi: 10.48550/arXiv.2105.04906.

[7] David R Patton, Kieran D Wilson, Colin J Metrow, Sara L Ellison, Paul Torrey, Westley Brown, Maan H Hani, Stuart McAlpine, Jorge Moreno, Joanna Woo, Interacting galaxies in the IllustrisTNG simulations - I: Triggered star formation in a cosmological context, Monthly Notices of the Royal Astronomical Society, Volume 494, Issue 4, June 2020, Pages 4969–4985, https://doi.org/10.1093/mnras/staa913

[8]Gimeno, G. N., Díaz, R. J., and Carranza, G. J., "Catalog of Double Nucleus Disk Galaxies", The Astronomical Journal, vol. 128, no. 1, pp. 62–67, 2004. doi:10.1086/421371.

[9] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," Commun. ACM, vol. 60, no. 6, pp. 84–90, May 2017, doi: 10.1145/3065386.

[10] A. Dosovitskiy et al., "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale." arXiv, Jun. 03, 2021. doi: 10.48550/arXiv.2010.11929.

Figure 2: Query

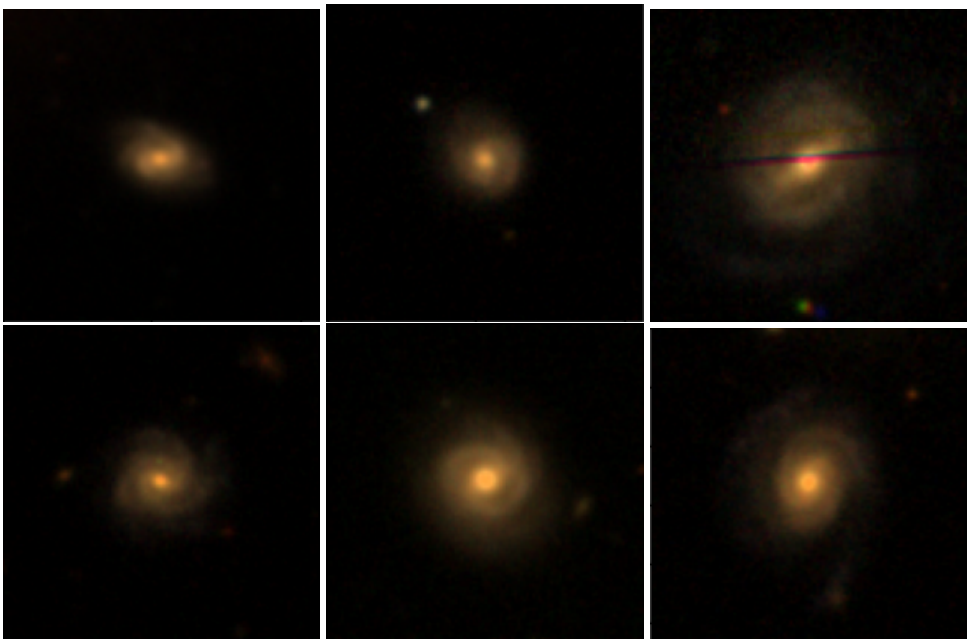

Figure 3: Results

[11]Lupton, R., “Preparing Red-Green-Blue Images from CCD Data”, Publications of the Astronomical Society of the Pacific, vol. 116, no. 816, pp. 133–137, 2004. doi:10.1086/382245.

[12] “Self-supervised learning: The dark matter of intelligence.” https://ai.facebook.com/blog/self-supervised-learning-the-dark-matter-of-intelligence/ (accessed Mar. 10, 2023).

A Appendix

On querying a random sample(batch size 200) of images with an image we get the following results which are most similar to our original query 2,3.

B Appendix

The different components of the Loss function for the VICReg Loss as obtained in our runs has been given in figure 4. The figure shows two different curves of colours Orange(Low Augmentation) and Black(High Augmentation). These two runs had different augmentations for them. We can see from

the graphs that both of them had similar performance when it comes to variance and covariance loss but the higher augmentation run suffered with the high invariance loss. This should be expected since higher augmentations make the embedding for the two different transformation spaces vastly different. Hence we need a better metric to gauge the performance of our model other than the loss.

C Appendix

The top image in figure (5 denotes the correlation matrix of a batch of samples. In this case the batch was chosen to be 250. The matrix was calculated with the following formula:
Consider a batch of images X, where

$$X = \{x_1, x_2, \ldots, x_n\}$$

,
and the corresponding embedded vectors are

$$Z = \{z_1, z_2, z_3, \ldots, z_n\}$$

Where,

$$z_i = Encoder(x_i)$$

$$Mat(i, j) = MSE(z_i, z_j) \tag{1}$$

Where MSE is the Mean Square Error in the euclidean metric.

The diagonal elements are zero as they correspond to the similarity between the two same images. The matrix is symmetric, which can be expected since MSE is a symmetric loss function. Moreover, we can see some dark spots scattered throughout the matrix. These are similar pairs. Looking at the right figure, which shows the distribution of similarity of a single element with respect to every other element, we can see that most are dissimilar(higher MSE Loss), while very few have similarities. This shows no information collapse has occured.

Figure 4: The three loss components over different steps as obtained from results, from top-to-bottom: Invariance Loss, Covaraince Loss, Variance Loss and Total Scaled Loss

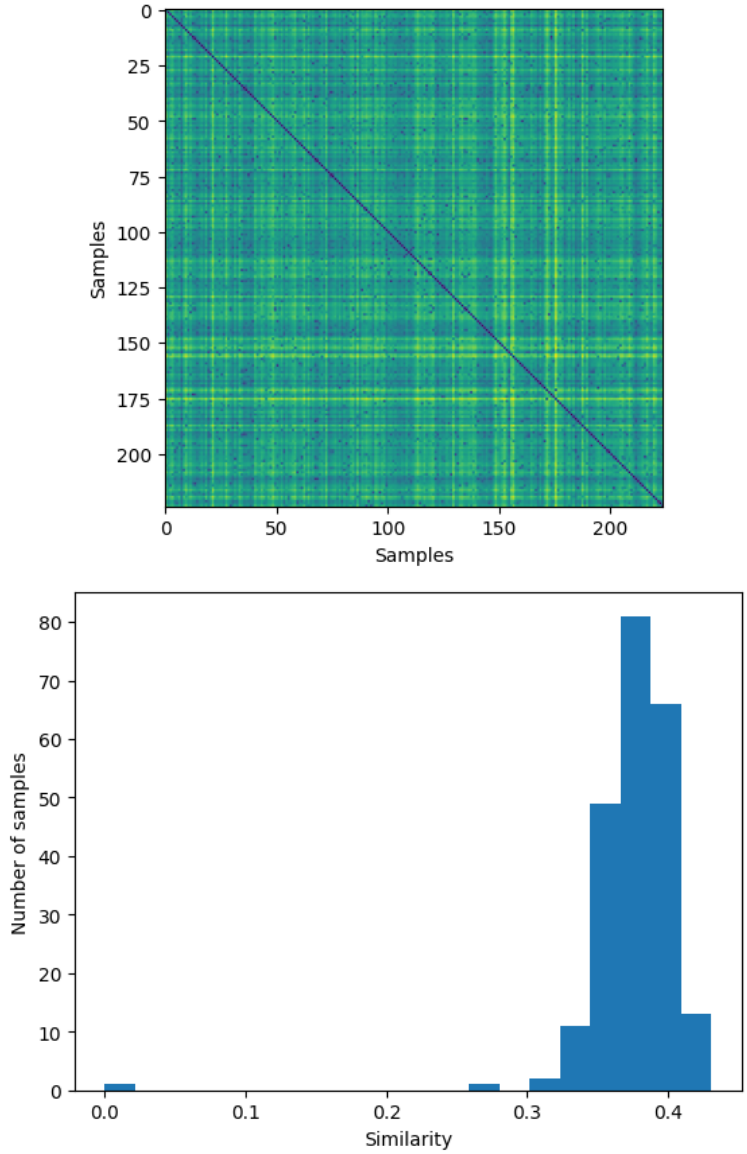

Figure 5: The top image corresponds to the correlation(MSE of embedding vector) between the samples.The bottom image corresponds to the correlation of a single sample with other samples.