Investigation of Environmental Dependencies of Galaxy Formation and Evolution with Convolutional Neural Network

Arefe Abghari, Zahra Baghkhani, Setareh Foroozan, Laya Ghodsi,
Mohammad Hadi Sotoudeh, Sina Taamoli

[a]Studying Large Scale Structure of the Universe leads to asking questions about the effect of environment on Galaxy formation and evolution. Analyzing data from observations suggests that the distribution of dark matter affects spatial distribution of galaxies. Also, it acts on star formation rate, metallicity, morphology, and spin.

The main tools to investigate the effect to environment on galaxies are observational data and theoretical models. Since modelling galaxies needs multi-scale, multi-environment physical simulations (hydrodynamics, stellar evolution, stellar clusters, supernova explosions, etc.), a deal of difficulties will arise. Running models of some simple galaxy formation and evolution processes may require age of the universe timescales.

The difficulty of modelling galaxies, encourages us to take advantage of Machine Learning. This helps us to establish a mapping between 3D distribution of galaxies and 3D distribution of dark matter in hydrodynamic simulation snapshots, especially in z=0 one. In this project, we use different regression techniques to predict galaxy distribution from dark matter distribution and use validation metrics to measure each technique's performance. Finally, Deep neural networks are employed to decrease the loss of our model.

I. INTRODUCTION

The constant growth of computational power allows cosmologists to use simulations for predicting their theory models. Simulations lead to pictures of universe based of models assumptions. Thus, by comparing simulation results with observable universe one can validate theoretical models. However, evolving models with cosmological through cosmological timescales consumes a lot of computational resources.

In order to save computational resources, we can take advantage of the standard model of cosmology. Since most of the matter in the universe is of the form of dark matter and gas follows dark matter density, one can derive the large-scale cosmic structure of the Universe by evolving dark matter-only simulations and then add the effects of gas to the field. So, the complexity of gas collapse and cooling will do not need to be taken into account during the simulation. Dark matter halos are frames in which galaxies form. Hence, by mapping dark matter structure to baryonic matter structure, behavior, such as morphology, evolution, and spatial distribution of galaxies, could be predicted from the behavior of dark matter halos.

In this paper, machine learning approaches are proposed for the problem. We explore the use of convolutional neural networks (CNNs) to establish the mapping from the 3D matter field in Illustris simulation to galaxies in a full hydrodynamic simulation. The task can be formulated as a supervised learning problem.

Section 2 presents the data used. Section 3 presents our models' architecture and quantitative results. We will conclude in Section 4 and discuss future work in Section 5.

II. DATA

We used Illustris project[2] two types of simulations: hydrodynamic (gravitational + hydrodynamic forces and astrophysical processes) and N-body (only gravitational forces). Illustris-3 data release includes the snapshots at all 136 available redshifts. In this work our focus will be on z = 0, the current epoch of the Universe.

The simulation details are listed below:

Simulation Details	
Parameter Description	Value
Simulation name	Illustris-3
length of simulation box	75000.0 [ckpc/h]
number of DM particles	94196375
number of gas particles	94196375
current redshift (z)	0.0

Here we provide some visualisation of the above-mentioned data for better understanding of what we are dealing with.

[a] This project is mainly based on Zhang et.al. work [1]

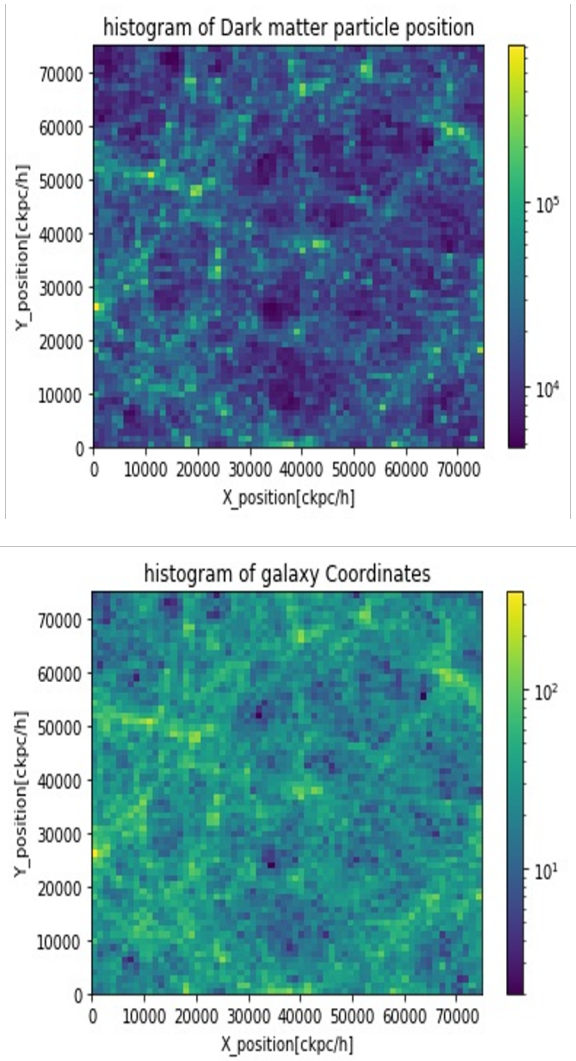

FIG. 1. top fig: Histogram of DM particles
bottom fig: Histogram of Galaxy coordinates
It is obvious that there should be a relation between the position of two.

We compute the density fields of galaxies and dark matter by assigning each component to a regular grid with 100^3 voxels using the nearest-grid point mass assignment scheme: if a galaxy or dark matter particle is inside a given voxel, the value of that cell is increased by 1 for the corresponding field.

III. METHODS AND RESULTS

Here we present our approach for linking the 3D dark matter field from N-body simulations to the 3D galaxy distribution from hydrodynamic simulations. In this project we trained a variety of non-linear and Convolutional Neural Network (CNN) regressors on the data. But for training a capable regressor one would need to consider all the important features. Therefore we included the 8^3 nearest neighbors of each cell, as features

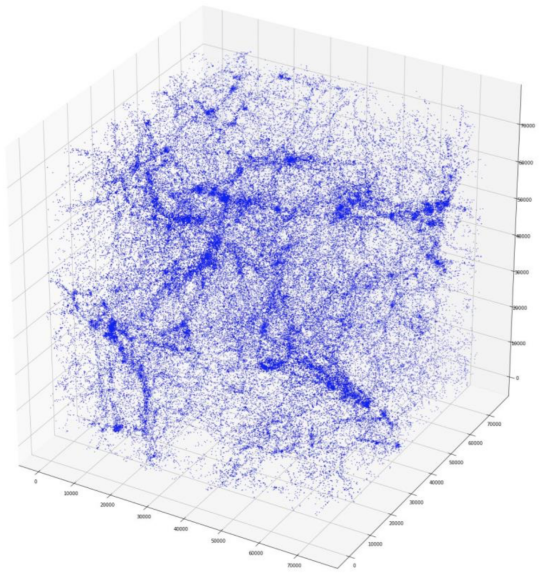

FIG. 2. Galaxy distribution in the box of simulation (75 * 75 Mpc)

of that certain cell. We will explain all the regressors in details in the following.

A. Basic Regressors

In the table below there are some basic regressors' results sorted by test l2-loss.

Simulation Details		
Regressor	train loss (l2)	test loss (l2)
Random Forest	1.85346E-02	1.02686E-01
Bagging	1.81095E-02	1.02990E-01
Lasso	1.72156E-01	1.87274E-01
Ridge	1.71892E-01	1.88162E-01
Decision Tree	0.00000E+00	1.89572E-01
Dummy	3.57907E-01	3.78466E-01
Adaboost	3.00922E+00	3.02028E+00

The first 3 ones, are acceptable, however for a better accuracy we will try CNN.

B. CNN

Convolutional neural networks(CNN) are traditionally used in computer vision tasks, such as image classification, detection, and segmentation. They are increasingly being adopted in cosmology researches nowadays, and work well in representing features of Universe. In this project, we used CNN in order to find a map from the 3D box (like a 3D image) to number of galaxies in the hydrodynamics simulation box. Our main challenge was the inherently spatial nature of the data (dark matter

and galaxies are structured spatially, on various correlated scales To address the first aspect, we propose to rely on convolutional networks. They naturally provide interesting properties for our problem such as translational invariance. The task can be formulated as a supervised learning problem. The paper identifies ReLU, average pooling, batch normalization and dropout as critical design choices in the neural network architecture to achieve highly competitive performance in estimating cosmology parameters. We trained and tested three different CNNs and comparing to usual reggresors, their performance is much better. Below is the summary of their architucture and results.

Diagram.png

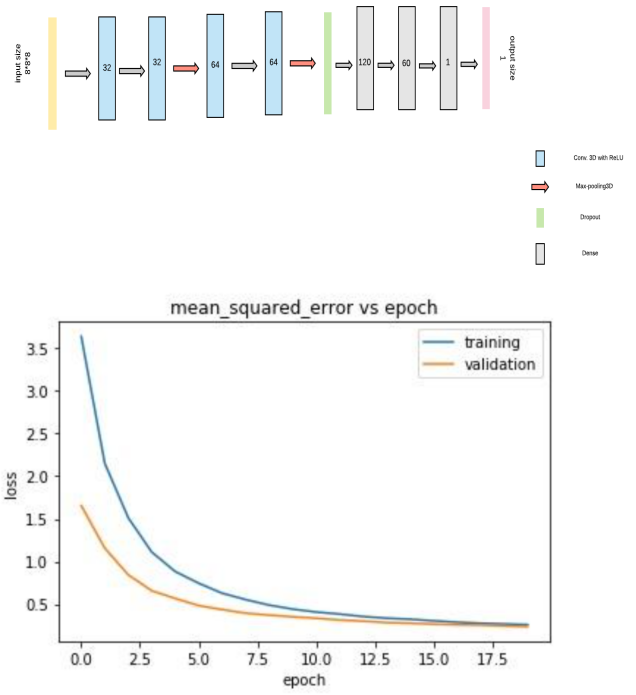

FIG. 3. 1st CNN architecture and performance for train and test set data. I It is trained with 20 epochs.

IV. CONCLUSION

In this project, we tested some machine learning estimatiors to model the map between the underlying dark matter from N-body simulations and the galaxies distribution from full hydrodynamic simulations both from the Illustris project. We show that the deep learning approach, convolutional neural network, by optimizing the number of galaxies prediction per voxel, is the best approximation between the estimations. Using CNN we could reach the highest accuracy of 94% for test set. This is a first step to overcome the need for computationally expensive hydrodynamic simulations in the long

Diagram (2).png

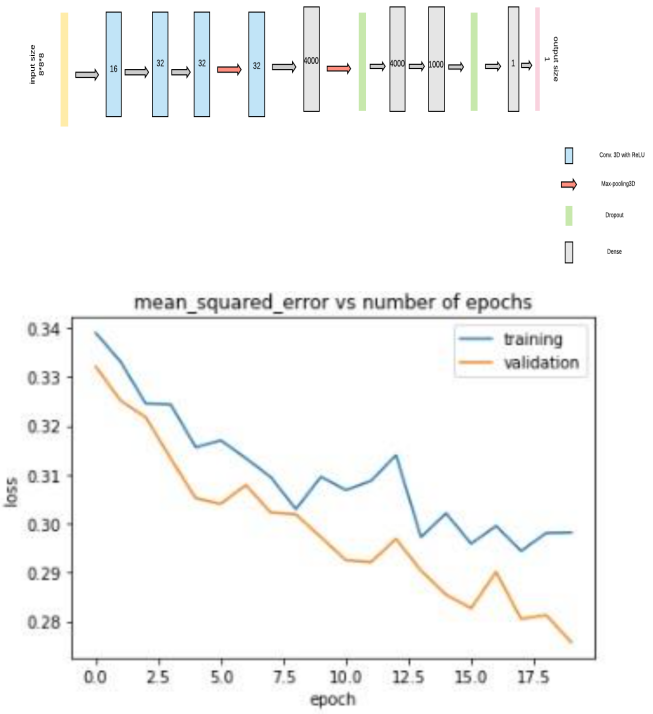

FIG. 4. 2nd CNN architecture and performance for train and test set data. It is trained with 20 epochs.

Diagram (3).png

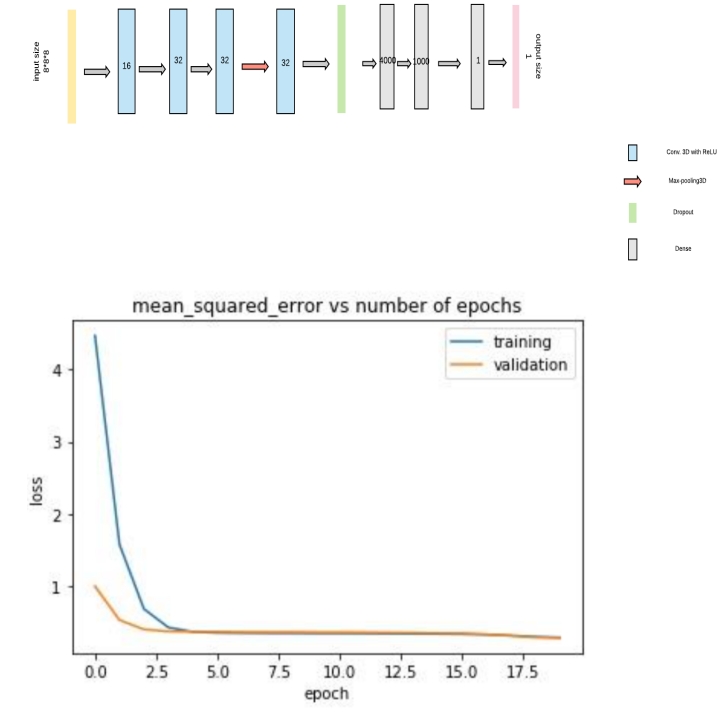

FIG. 5. 3rd CNN architecture and performance for train and test set data. It is trained with 20 epochs.

run. This approach will establish a new link between astrophysics and cosmology.

V. FUTURE WORK

This work opens several trails for future research. Here, we used the Illustris3 simulation, therefore, at first, we can test our method with the higher resolution versions. Evaluating the 2-point correlation function and power spectrum to compare our model with benchmark theoretical models. Secondly, it will be very interesting to extend our model to be able to predict not only the number and positions of the galaxies but also their internal properties, e.g. stellar mass, star-formation rate, metallicity, etc. Training the model at different epochs of the Universe will allow us to better understand the complicated physics involved in galaxy formation/evolution. At the end, using our approach, one can change cosmological parameters(e.g. for Fuzzy dark matter) and run dark matter-only simulations and predict the baryonic matter characteristics. Comparing the final result with observation, can be a good criterion for cosmological models.

VI. REFERENCES

[1] X. Zhang, Y. Wang, W. Zhang, Y. Sun, S. He, G. Contardo, F. Villaescusa-Navarro and S. Ho, arXiv:1902.05965 [astro-ph.CO].

[2] http://www.illustris-project.org