Semester report - Doctoral School of Physics

by **Mirkó György Mocskonyi** (mocskonyi.mirko@gmail.com)
PhD Program: Statistical Physics, Biological Physics and Physics of Quantum Systems
Supervisor: prof. István Csabai

Ph.D. Thesis title: Physical methods in artificial intelligence - AI methods in physics

January 31, 2023

1 Introduction

In this semester initially I had two directions to start off my research. First, there was the reconstruction of the 3D Large Scale density of the universe using the photometries of galaxies and/or quasars with the help of a neural network model. Second there was the development of the detection and classification of lesions in mammography scans using, again, a neural network designed for object detection, segmentation and classification. From the two I started off with the first one.

At the large scales the universe has a unique matter density distribution: the galaxies are organized into a complicated network structure of high-density nodes connected by filaments of smaller galaxy density and large volumes of voids in between. Obtaining this structure from the photometric measurements as opposed to spectroscopic ones would be extremely beneficial because of cost-effectiveness and time-efficiency. The plan for the density reconstruction research consisted of using a galaxy catalog with available spectroscopic data - which allows for the evaluation of the performance of the model - to build the artificial intelligence model for estimating the radial density distribution for a 3D grid of voxels. One can have ground truth radial density distribution from the known spectroscopy or based on photometric redshift (photo-z) estimation. Therefore the plan was to use supervised learning with an additional unsupervised dimensionality reduction method beforehand, e.g. self-organizing map (SOM) or UMAP.

2 Description of research work carried out in current semester

First of all, I read some literature of how the large-scale structure of the universe was handled in photo-z studies. The incorporation of known radial density (from spectroscopy) was used to improve on the photo-z prediction in [1] by combining the two and thereby narrowing the estimated PDF. Less direct inclusion of the large-scale structure can be done by not including spectroscopy at all but rather with the help of spatial correlations as in [2, 3, 4].

Additionally, dimensionality reduction methods, particularly SOMs were used. E.g. in [5] they distinguished between early-type and late-type galaxies. We thought that incorporating a dimensionality reduction method in our model we could achieve a realization independent

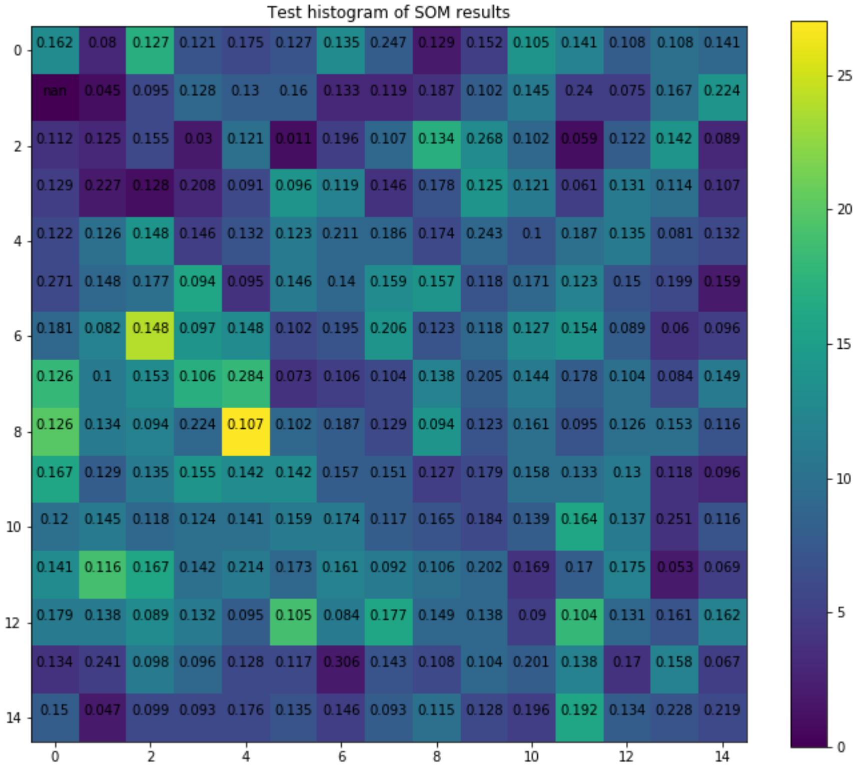

Figure 1: 2D histogram of the test set galaxies selected from the SDSS DR7 database processed by SOM. The value of the bins is color-coded.

representation from the 5 photometric band measurements belonging to each galaxy. It essentially creates a latent color space to use in the further steps of the model. So far I have worked with implementing the SOM.

We thought that the convenient *Autograd* functionality of the **pytorch** package was worth incorporating in our model as it could calculate the gradients through not only the neural network model but also additional steps of the model, like a SOM (or UMAP). Therefore I worked on getting familiarized with the package and set up a working implementation of a SOM within its framework. At first I wrote a basic implementation from which I worked towards having **pytorch** implementation.

The first basic implementation was tested on the SDSS DR7 galaxies, specifically on the photometric measurements in the *ugriz* bands and only on 10000 random samples. The data was split into training and test sets in 1:4 ratio. As the SOM is an unsupervised method, the spectroscopy data was only used for the evaluation. In Figure 1. the a 2D histogram of the SOM cells can be seen which contain the test set galaxies with the number of galaxies color-coded in each cell while the mean redshifts are superimposed onto the histogram. The overall standard deviation calculated from the Median Absolute Deviation was 0.0764, which is not too bad for an unsupervised learning technique although very far from the state-of-the-art methods.

The first of the further steps will be to reproduce the same algorithm with the *pytorch* implementation of the SOM and finding out if there are any meaningful distinguishing properties that the SOM discovers. After that a neural network will be attached after the SOM to realize the actual estimation of the radial density distribution. There are a multitude of possible choices at that point and it promises to be quite challenging to find the most optimal architecture.

On the other hand the work shall be started on the other main direction mentioned in Section 1., namely the mammography lesion detection.

3 Publications

I had no publications in this semester.

4 Studies in current semester

In this semester I attended three courses. The course *Clustering with complex networks* was about community finding in complex networks which proved very useful to me as it deepened my knowledge in Bayesian models, modularity and Stochastic Block Models. The course *Data Science Laboratory* was about going through a project about photo-z of galaxies which made me encounter approaches I had not tried before. The third course was *Data models and databases*, which gave me an insight into the modern ways of handling and working with databases.

5 Conferences in current semester

I attended no conferences this semester.

6 Teaching activity in current semester

In this semester along fellow PhD students I helped evaluating two projects of the students of the *Computer Simulations* course by prof. István Csabai. I corrected and graded the projects, gave the students feedbacks and oversaw their final exams.

References

[1] *2003.10766.pdf*. chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/https://arxiv.org/pdf/2003.10766.pdf. (Accessed on 01/31/2023).

[2] *2209.03967.pdf*. chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/https://arxiv.org/pdf/2209.03967.pdf. (Accessed on 01/31/2023).

[3] V. Scottez et al. "Clustering-based redshift estimation: application to VIPERS/CFHTLS". In: *Monthly Notices of the Royal Astronomical Society* 462.2 (July 2016), pp. 1683–1696. ISSN: 0035-8711. DOI: 10.1093/mnras/stw1500. eprint: https://academic.oup.com/mnras/article-pdf/462/2/1683/13773629/stw1500.pdf. URL: https://doi.org/10.1093/mnras/stw1500.

[4] *1609.09085.pdf*. `chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/https://arxiv.org/pdf/1609.09085.pdf`. (Accessed on 01/31/2023).

[5] Yong-Huan Mu et al. “Photometric redshift estimation of galaxies with Convolutional Neural Network”. In: *Research in Astronomy and Astrophysics* 20.6 (2020), p. 089.