UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

PROGRAMA DE POSGRADO EN ASTROFÍSICA

Instituto de Astronomía

Environmental Studies in the Cosmic Web using Machine Learning

QUE OPTA POR EL GRADO DE:

MAESTRO EN CIENCIAS (ASTROFÍSICA)

PRESENTA:

Ana Cristina Arcos Marín

TUTOR PRINCIPAL:

Dr. Miguel Ángel Aragón Calvo, Instituto de Astronomía - UNAM

Ensenada, B.C., Agosto 2019

Acknowledgements

I would like to thank the CONACyT program for offered me the support and the opportunity to keep studying. To Instituto de Astronomía, UNAM, for giving me the necessary means to accomplish my studies. Thanks to my assessor, Dr. Miguel Aragón, for your patience, support and inspiration in learning new things. And thanks to my sinodal committee, Dr. Héctor Aceves, Dr. Christophe Morisset, Dr. Takamitsu Miyaji, Dr. Octavio Valenzuela and Dr. Benjamín Hernández for your comments and feedback in this work.

A big gratitude to my parents, my mom and dad, for your unconditional support in all my decisions and for your inspiration to continue my path in knowledge. To my brother Erik for the experiences and learning we have had and gone through together, always with a smile in the face. To Denisse for listening and advised me in all moments. Thanks to all my family in Monterrey and Houston, uncles, aunts, cousins, for your motivation.

I thank my all life friends, Gaby, Sandra, Manuel, Ale, Hilda and Marcelo for your friendship and support. To Emmanuelle for being by my side and partner in crime. Thanks to Valeria and Luis for helping me in the hard moments, making me laugh and giving me your trust. To Paola and Aideliz to share not only the court with me but also your lives and all your knowledge. To my roommates, Alesa and Mely for giving me new experiences. Thanks to all the awesome people I have known in this time, Fati, Lupita, Sofi, Marcelo, Andrés, OAN staff, team CICESE and team Matrix for bring me joy. And last, thanks to José Manuel for the time you shared by my side in this path and impulsed me to continue and keep growing.

Resumen

Edwin Hubble, en 1925, inició la astronomía extragaláctica al observar la galaxia de Andrómeda (M31) y descubrir que existen otras galaxias fuera de la Vía Láctea, que forman parte del Universo. Este descubrimiento hizo que se empezaran a clasificar las diferentes galaxias que se observaban en cuatro tipos de morfologías, según su forma, en: i) *elípticas* (E), siendo de forma redonda o elíptica; ii) *espirales* (S), de forma espiral; iii) *lenticulares* (S0), consideradas un intermedio entre las espirales y las elípticas; e iv) *irregulares* (Irr), que no tienen una forma definida. Otras clasificaciones de galaxias son por a) *color*, azul o rojo dependiendo de si son estrellas jóvenes o viejas respectivamente; b) *tipo temprano* y *tipo tardío*, que son galaxias viejas o jóvenes; c) *formación estelar*, ya sea activa o inactiva; y d) *luminosidad*, galaxias brillantes u opacas.

El catálogo de galaxias 2dFGRS muestra que las galaxias se encuentran organizadas en una red compleja conocida como la Red Cósmica. Esta estructura se formó después del Big Bang debido a fluctuaciones cuánticas que se expandieron gracias a la inflación. Esto trajo como resultado la formación de halos de materia oscura que sirvieron como pozo de potencial para que la materia bariónica se aglomerara y se diera la formación galáctica dentro de estos objetos. Las diferentes características (o regiones) de la Red Cósmica están clasificadas por su campo de densidad y su estructura en: i) *cúmulos*, regiones de alta densidad y forma casi-esférica; ii) *filamentos*, regiones de media densidad considerados como puentes entre cúmulos; iii) *paredes*, también de densidad media con forma de sábana o pared; y iv) *vacíos*, regiones de baja densidad rodeados de filamentos y paredes. Estas diferentes regiones generan ambientes diferentes por la distribución de materia que influenciaron en la formación de galaxias junto con otras interacciones como uniones de galaxias, retroalimentación de AGNs, formación estelar y el medio intergaláctico. Todos estos factores predicen que las propiedades de galaxias y sus morfologías están influenciadas por la formación y evolución de los halos de materia oscura y su medio ambiente. En este trabajo se utiliza como medio ambiente los diferentes campos de densidad que existen en la Estructura a Gran Escala.

Para conocer la influencia del medio ambiente, se utilizan diferentes herramientas para medir el medio ambiente de una región del espacio. Las más utilizadas son las técnicas de *vecino cercano*

y *apertura fija* donde se estima la densidad en la región de interés. La primera está relacionada a la distancia a la galaxia n^{th} donde la densidad $\rho \propto \frac{1}{r^3}$ estipula que a radios pequeños, el campo de densidad será alto y vice-versa, a grandes radios el campo de densidad será bajo. El segundo método está relacionado al número de galaxias en un determinado volumen o área de interés donde dependiendo de la región, ya sea cúmulo o vacío, la densidad pudiera ser alta o baja respectivamente. Dependiendo de los tipos de datos que se tienen y el análisis que se quiera realizar, es el método a utilizar.

Como ya se mencionó anteriormente, la formación de galaxias depende del ambiente en el que se encuentre y debido a esto, existe una tendencia de observar ciertas clasificaciones de galaxias en ciertos medios ambientes. Por ejemplo, se han observado que galaxias masivas, de tipo temprano, rojas y con poca formación estelar se encuentran más en ambientes de alta densidad mientras que galaxias jóvenes, azules y con formación estelar, se encuentran en ambientes de media o baja densidad. Esto demuestra la relación del medio ambiente en la formación y evolución de las galaxias.

A pesar de que se tiene una caracterización de la relación del medio ambiente con la clasificación de galaxias, hay aspectos que deben considerarse durante el análisis de medio ambiente y propiedades de galaxias. a) Primeramente, considerar el método de medición a utilizar, ya que los métodos pueden brindar diferente información dependiendo de la escala. A grandes escalas, fuera del halo de materia oscura donde residen las galaxias, las técnicas de vecino más cercano prueban mejor las propiedades internas del halo de materia oscura, mientras que métodos de apertura fija prueban el halo de materia oscura como un todo dando resultados de densidad mayores. Por otro lado, a escalas pequeñas, dentro del halo, vecino más cercano y apertura fija tendrán resultados similares, sin embargo, no funcionarán para regiones de densidad baja, como vacíos, debido a las grandes distancias que habría entre vecinos. Y segundo aspecto a considerar, b) las clasificaciones de galaxias. Existen estudios donde a pesar de realizar la misma clasificación, han dado diferentes resultados debido al método de clasificación ya que muchas veces se realizan clasificaciones utilizando la morfología, color o masa estelar, y estos no son métodos adecuados debido a que tienen una influencia diferente en el medio ambiente que da como resultado diferentes análisis en relación a la dependencia del color y morfología con el medio ambiente.

Existiendo diversidad de resultados y métodos para la medición de densidad, se propone en este trabajo el uso de dos métodos diferentes para mejorar la medición y el análisis de datos. Para esto, se utiliza el método de *Lagrange Tessellation Field Estimator (LTFE)* como medición de densidad. Esto debido a que puede distinguir los diferentes campos de densidad de la Red Cósmica. LTFE es un nuevo método utilizado para calcular densidades en simulaciones de N-cuerpos. Consiste en aproximar la distribución de materia usando volúmenes de tetraedros. Mientras las partículas

evolucionan durante la simulación, la densidad cambia y es estimada por una teselación tetraédrica. Un tetraedro contiene un sexto de la masa de la partícula dispersa en el volumen (que se calcula realizando un producto vectorial y escalar de los vectores que formen la figura). De este modo, el método LTFE mejora la medición del espacio de densidad en rangos de baja e intermedia masa en las simulaciones, donde se considera a la materia oscura como un fluido y se pueden rastrear los aspectos macroscópicos y microscópicos del fluido de materia oscura en cada ubicación de las partículas durante la evolución de toda la simulación.

Otro método a utilizar es *aprendizaje de máquinas*, *machine learning*, para identificar por medio de imágenes de la geometría de la Red Cósmica, esto es, las diferentes estructuras y que muestre como resultado la predicció del campo de densidad. Aprendizaje de máquinas tiene como base el aprendizaje humano el cual el modelo más simple es la neurona artificial llamada perceptró el cual tiene como base: *a)* datos de entrada: valores reales para entrenar y pesos, que consisten en valores aleatorios; *b)* un sumador, que multiplica los valores de entrada con el peso y, *c)* una funció de activació que limita el parámetro de salida de la neurona a un valor finito. Estos perceptrones pueden formar diferentes estructuras, ya sea neuronas conectadas paralelamente formando una sola capa o teniendo muchas capas con un cierto número de neuronas. Cuando se tienen más de dos capas de neuronas, esta arquitectura es considerada como *aprendizaje profundo* o *deep learning*. La arquitectura que se utiliza en este proyecto es de aprendizaje profundo usando redes neuronales convolucionadas. Las redes neuronales convolucionadas se utilizan principalmente para análisis de imágenes en dos o tres dimensiones, sonido e identificación de texto. La base de estas redes son: a) convolución, donde se realiza una multiplicación entre los datos de entrada con los pesos (filtros) que se encuentran en la red; b) agrupamiento máximo (*max pooling*), que consiste en mantener los valores máximos de la convolución manteniendo la información más importante; y c) capas conectadas (*fully connected layers*), se encuentra al final de la red, junta la información de todas las operaciones y proporciona los valores de salida. La estructura utilizada para este proyecto es VGG16 (Visual Geometric Group 16), la cual consiste en 16 capas en donde se realizan las operaciones y el entrenamiento. Los pesos son adquiridos utilizando los filtros de ImageNet. La finalidad de utilizar aprendizaje profundo es entrenar a la computadora a identificar las diferentes regiones o estructuras de la Red Cósmica junto con sus valores de densidad calculados por el método de LTFE, para esto se utilizan 9000 imágenes para el aprendizaje y 1000 imágenes para poner a prueba el entrenamiento. De este modo, al pasar imágenes que no han sido vistas por las red, pueda predecir el valor de la densidad de la región de la Red Cósmica. Lás imágenes son obtenidas de la Simulación Illustris en una resolución más baja ya que si las imágenes tienen muchos detalles, la red neuronal tardará en aprender cada detalle y esto implica más tiempo en computación. El proceso de entrenamiento de la red neuronal utilizando VGG16 es el siguiente:

- Una imagen de entrada y valores de pesos.
- La imagen pasa a travéz de cada capa, realizando las operaciones de convolución y agrupamiento.
- Al final, en las capas conectadas, se junta la información del entrenamiento para desplegar un valor de salida que sería un valor de predicción de la densidad.
- El valor de salida se utiliza para calcular el error y propagarlo hacia la red para ajustar los pesos.
- Al ajustar los pesos el proceso se repite para cada imagen.
- El proceso continúa hasta que el error sea mínimo y todas las imágenes hayan pasado por la red.

Para el entrenamiento de la computadora, se utilizan dos grupos de datos: *i)* el primer grupo de datos son imágenes de la Red Cósmica con la medición de densidad a 1 $h^{-1}Mpc$, *ii)* el segundo grupo de datos son las mismas imágenes pero con la diferencia de que la medición de densidad es a 2 $h{-}1Mpc$. De este modo se podrá analizar la eficiencia del método, tanto de LTFE y de la red neuronal para identificar geometrías y medir densidades.

Otro parámetro que se utiliza para el aprendizaje de la red neuronal es el uso de dos tipos de entrenamiento: *transfer learning* y *fine-tuning*. *Transfer learning* consiste en utilizar una red neuronal pre-entrenada, esto es, utilizar los pesos que se usaron en otro problema. De este modo el entrenamiento no empieza desde un punto cero y así el modelo puede ser mejorado fácilmente. Por esta razón, se utilizan los pesos de los filtros de ImageNet, estos filtros contienen al rededor de millones de imágenes de objetos comúnes que ayudan al entrenamiento de la computadora. *Fine-tuning* se utiliza para mejorar el aprendizaje de la red ya pre-entrenada. Se obtiene cambiando los valores de ciertos filtros dentro de la estructura de la red neuronal haciendo que los filtros se adapten a los datos de entrenamiento, esto es, si los datos que se utilizan en este proyecto son geometrías de la Red Cósmica, los filtros se modificarán a modo de que tendrán formas de la Red Cósmica, por lo tanto al pasar los datos de entrenamiento, aprenderá detalles más finos. En la estructura VGG16 se modifican las últimas dos capas para este entrenamiento ya que las últimas capas son las que identifican las estructuras más pequeñas de las imágenes.

Como se mencionó anteriormente, se utilizan 9000 imágenes de la Red Cósmica para entrenar la red neuronal con la medición de densidad a una distancia de 1 $h^{-1}Mpc$ y 2 $h^{-1}Mpc$ utilizando *transfer learning* y *fine-tuning* para cada distancia. Para comprobar si el entrenamiento fue eficiente, se utilizan 1000 imágenes que no han sido vistas por la red para que prediga los campos

de densidad de dichas imágenes. Para verificar si las predicciones son correctas, se utilizan como métodos a evaluar: *moving window*, Pearson correlation y se realizan 49 iteraciones para conocer que tan variable son las predicciones en cada iteración y hacer una estadística.

Los resultados que se analizan son las predicciones hechas por la computadora y las densidades medidas por el método de LTFE. Estos resultados se grafican para conocer la asertividad de la red neuronal, la medición de densidad y de los diferentes aprendizajes utilizados. Para *transfer learning* se obtienen buenos resultados. Como primera aproximación visual los valores predichos no se encuentran dispersos. Se observa que las predicciones para 2 $h^{-1}Mpc$ son más certeras ya que se encuentran menos dispersas que a 1 $h^{-1}Mpc$ y esto también se observa cuando se realizan las mediciones con *moving window*. La correlación de Pearson brinda valores de $p = 0.975$ y $p = 0.996$ para 1 $h^{-1}Mpc$ y 2 $h^{-1}Mpc$, respectivamente, lo cual dice que ambos parámetros tienen una correlación muy alta. Para *fine-tuning*, se observa que las predicciones son más parecidas a los valores reales debido a que muestran muy poca dispersi'on y con *moving window* se observa que los valores predichos tienen dispersiones de 10 unidades a diferencia de *transfer learning* que muestra dispersiones de 30 unidades. Se observa que a 2 $h^{-1}Mpc$ las predicciones son de las más exactas a diferencia de 1 $h^{-1}Mpc$ y a diferencia de los resultados obtenidos en *transfer learning*. La correlación de Pearson es de $p = 0.993$ para 1 $h^{-1}Mpc$ y de $p = 0.99$ para 2 $h^{-1}Mpc$, lo cual demuestra que la mejor combinación para predecir densidades es *fine-tuning* a 2 $h^{-1}Mpc$. A pesar de la eficiencia en la predicción de las densidades por parte de la red neuronal, se observa que para densidades altas, $\delta > 200$ a 1 $h^{-1}Mpc$ y $\delta > 20$ para 2 $h^{-1}Mpc$, el método falla. La dispersión que se observa a estas densidades es mayor y esto es debido a que la medición de LTFE y la red neuronal falla a densidades altas, esto ya que no pueden caracterizarse bien estas regiones por el volumen de la caja que se utiliza en las simulaciones y por lo tanto, no se tienen muchos datos de cúmulos o super-cúmulos de la Estructura a Gran Escala a diferencia de la cantidad más alta de filamentos y paredes que existen en las simulaciones.

Los resultados obtenidos con 49 iteraciones se muestran como gráficas donde se presentan los datos reales de densidad y la densidad promedio de las 49 predicciones, así como también se muestran las barras de error para cada medición donde el error se calcula usando la desviación estándar. Para *transfer learning* se observan barras de error grandes a 1 $h^{-1}Mpc$ y en todo el rango de densidades siendo más prominentes en densidades medias y bajas. A 2 $h^{-1}Mpc$, los márgenes de error son menores siendo más notorios en densidades medias y altas. Para *fine-tuning*, las barras de error disminuyen a diferencia del entrenamiento con *transfer learning*, y las densidades donde son más prominentes es en el rango medio-alto para 1 $h^{-1}Mpc$. Para 2 $h^{-1}Mpc$ los márgenes de error son muy pocos y pequeños siendo más prominentes a densidades medias. En general, los rangos de error son consistentes con los mostrados en los resultados individuales. Así como en

los resultados individuales, también se observa que las mejores predicciones son con *fune-tuning* y a 2 $h^{-1}Mpc$. El tener mayor error en rangos de densidad media y alta es por las cantidad de imágenes que existen en la simulación utilizada en este proyecto, donde existen poco ejemplos que involucren cúmulos o super-cúmulos y por lo tanto la red no puede entrenarse para tener una mayor eficiencia en este rango.

Se concluyó que la red neuronal tiene la capacidad para identificar diferentes estructuras de la Red Cósmica y predecir su campo de densidad con bastante efectividad, acercándose a los valores reales. El mejor método a utilizar sería fine-tuning debido a la libertad de modificar los filtros de la estructura haciendo que el aprendizaje sea mas eficiente, esto se puede observar en los resultados. Los métodos y análisis que se han mostrado en este proyecto sirven como base para futuros proyectos. Primeramente, redes neuronales se puede utilizar para analizar bases de datos de imágenes muy grandes. En Astronomía se tienen muchos catálogos observacionales y de simulaciones que han brindado muchos datos a estudiar y una solución para el estudio de estos datos es usar métodos de Machine Learning. Por otro lado, para mejorar la medición de densidad, es conveniente ampliar el cubo de datos que se utiliza en la simulación para tener mayor número de datos para entrenar a la red y que estos a su vez, sean variados para que sea eficiente la predicción de las mediciones. Y uno de los mayores beneficios que tendrá el uso de estas herramientas será el demostrar procesos físicos con base en los datos de entrada que se le den a la red neuronal. Esto es, en este momento se están iniciando estudios en el área de Astronomía utilizando Machine Learning, por lo pronto estas herramientas se han enfocado más en enseñarle a la computadora a realizar ciertas tareas, conforme se siga investigando y avanzando en la modificación de la estructura de la red neuronal, llegará un momento en que se podrán utilizar teorías físicas, fórmulas, que expliquen lo que sucede en los datos de entrada, es decir, que la computadora muestre los procesos que están sucediendo en la información que se le está brindando de entrada.

Contents

Abstract

In this thesis the cosmic environment defined by the Large-scale Structure of the Universe will be studied. This structure was formed after the Big Bang and is conformed of high, medium and low density environments populated by a diversity of galaxies. Since there is a relationship of environment and different types of galaxies, the density field of each region in the structure will be studied. To characterized the density of each region the Lagrangian Tesselation Field Estimator (LTFE) will be used. For the Large-scale Structure, the Illustris Simulation and Machine Learning will be used to characterized the features of the structure by training the computer machine using Illustris images and predicting the density of each region. As a result, Machine Learning is able to predict efficiently density fields of the Large-scale Structure using images of its regions and making a way to predict properties of galaxies or other properties of the structure facilitating the analysis of data.

Chapter 1

Introduction

The formation and evolution of galaxies are some of the biggest mysteries of astronomy. During the 20^{th} century, discoveries and progress made by astronomers and physicists advanced the knowledge in Astronomy. During the year 1920, it was thought that the Galaxy, was unique and there were no more galaxies. When astronomers observed extended bright regions, they considered these objects were nebulas inside the Galaxy. This idea was competing with another that said there were other galaxies, like islands, outside the Milky Way, in the Universe. In 1925, Edwin Hubble observed a nebula and found out that it was outside the galaxy, i.e. it was an extragalactic object. To measure its distance, he used Cepheid variable stars, a method where period-luminosity relation is used (Schneider 2006). This nebula is now known as the Andromeda galaxy (M31) and Hubble's results marked the beginning of extragalactic astronomy.

Finding that there were more galaxies besides the Milky Way, Hubble classified them into four types of morphologies: ellipticals (E), spirals (S), lenticulars (S0) and irregulars (Irr) (see Figure 1.1) (Hubble 1926). The difference between these types refers to their shape. Ellipticals look rounded or elliptical in shape; spirals have a spiral form; lenticulars are considered as an intermediate path between ellipticals and spirals; and irregulars do not have a defined structure. There are also other properties that help classify galaxies like i) color: red or blue (mostly due to old and young stars that reside in the galaxy), ii) early and late type (old and younger galaxies respective), iii) star formation: quiescent or starbust, and iv) luminosity: faint or bright galaxies.

Figure 1.2 shows the distribution of $250,000$ galaxies where their distribution was obtained by the 2dFGRS galaxy survey (Colless et al. 2001) showing that galaxies are organized in a complex network known as the Cosmic Web (Bond et al. 1996) or Large-scale Structure (LSS) in the Universe. The Cosmic Web has a range of physical features and densities such as: i) clusters:

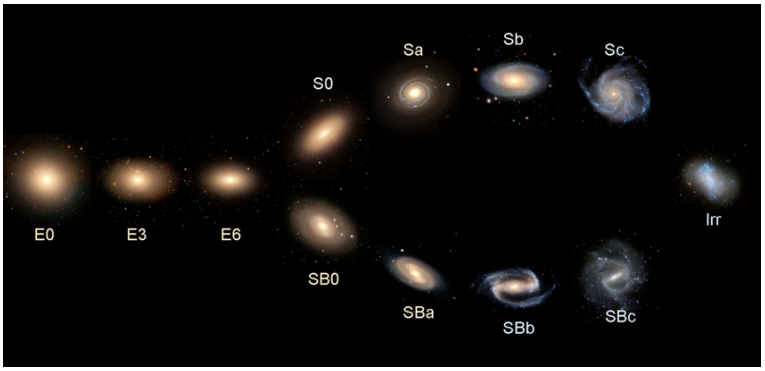


Figure 1.1: Hubble sequence for classification of galaxies (Cui et al. 2014), also known as the tuning fork. Elliptic galaxies (E) are at the left where the numbers are related to the elongation of the galaxy; spiral galaxies (S) are at the middle-right where letters refer to the shape of spiral arms and the "B" for barred spiral galaxies; and irregular galaxies (Irr) are at the very right of the tuning fork where they don't seem to have a well defined structure.

known as the densest regions, ii) filaments and walls: being of intermediate density, and iii) voids: having the lowest density of these features (Mo et al. (2010), Kreckel et al. (2016)).

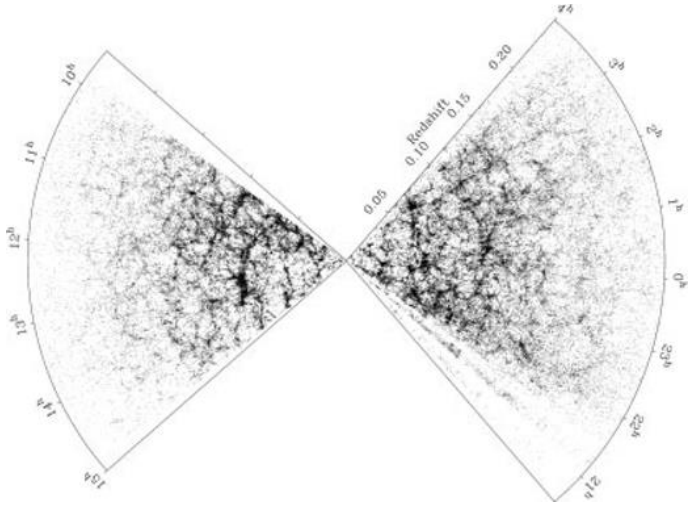


Figure 1.2: Projected distribution of the galaxies in the 2dFGRS survey as a function of redshift and right ascension (RA) (Colless et al. 2001).

Having a qualitative classification of galaxies and the knowledge of the Cosmic Web pattern, galaxy properties were found distributed in the different features of the Large-scale Structure of the Universe where i) early, red and elliptical galaxies tend to gather in clusters or high density regions and ii) in filaments, walls and voids (medium and low density environments), all galaxy types are found (Dressler (1980), Bamford et al. (2009), Blanton & Moustakas (2009)). Nevertheless, there is a trend of an excess of late and blue type in these low and medium densities (Dressler (1980),

Balogh et al. (2004), Bamford et al. (2009), Mo et al. (2010), Dragomir et al. (2018)). Also, in clusters, in general, there is no longer star formation (Bamford et al. 2009) while in other regions there are galaxies still forming stars, even in voids where density field is low (Kreckel et al. 2016).

The Lambda Cold Dark Matter (ΛCDM) model describes the formation and evolution of galaxies from the gravitational collapse of tiny fluctuations left after the Big Bang to formed halos (near-equilibrium state objects) (Zel'dovich 1970; White & Rees 1978). At the same time dark matter halos formed, galaxies formed and evolved in these halos where multiple mechanisms, also defined as the environmental influence (e.g. interactions with other galaxies (Blanton & Moustakas 2009), density (Dressler 1980) or intergalactic medium (D'Onofrio et al. 2015)), regulate star formation and produce the properties seen nowadays. This scenario predicts that galaxy properties are influenced and related to the host halo and environmental conditions (Muldrew et al. 2012), (Somerville & Davé 2015).

In order to know the influence of halos and external mechanisms in galaxy properties, some analysis tools were defined to measure the cosmic environment and are summarize in (Muldrew et al. 2012). The most commonly used techniques are the *nearest neighbor* and *fixed aperture* methods which estimate the *density* at a certain region of interest. The first one related to the distance to a n^{th} galaxy where $\rho \propto \frac{1}{r^3}$, which states that in small radius, density field would be high and vice-versa, at large radii the density field would be small. The second one is related to the number of galaxies in a determined area or volume of interest where depending of the region being measured (cluster or void) the density will be either high or low respectively.

Primarily related to environment	**Not primarily related to environment**
· Young star population (Blanton & Moustakas 2009)	· Color (Park et al. (2007) Bamford et al. (2009))
· Dark matter halo properties (Park et al. (2007), Muldrew et al. (2012))	· Star formation history (Blanton & Moustakas 2009)
· Star formation satellite galaxies (Sun et al. 2018)	· Stellar mass (Sun et al. 2018)
· Luminosity (Sun et al. 2018)	· Central galaxies (Sun et al. 2018)
· Satellite galaxies (Sun et al. 2018)	· Morphology (Dressler (1980), Bamford et al. (2009))

Table 1.1: Galaxy properties in relation to their environment.

Despite having characterized the relationship between environment and galaxy properties, there are aspects that need to be considered during the analysis such as the density measurement technique and scheme for classification of galaxies (Sun et al. 2018). In Muldrew et al. (2012) an analysis of the density measurements is made using nearest neighbour and fixed aperture methods where the importance of the measure is the scale. At large scales, external to the halo, nearest neighbour best probe internal properties of the dark matter halo while fixed aperture methods better probe the dark matter halo as a whole, leading to larger density values reflecting larger masses. At small scales, internal to the halo, fixed aperture and nearest neighbour would have similar results but won't work for regions like the field environment (low density environment or voids) due to the large distances between galaxy neighbours.

On the other side, Bamford et al. (2009) compared different galaxy classifications using morphology, color and stellar mass. In these studies, autors conclude that color and stellar mass are strongly dependent of environment, and morphology does not depend on environment when stellar mass is fixed. However, if stellar mass is not fixed, makes morphology and color not related to environment. In (Baldry et al. 2006) mentions that the morphology classification used by (van der Wel 2008) lead to the conclusion that morphology is more dependent on environment than stellar mass. These analysis lead to the opposing conclusions that morphological and color are not equivalent ways to classify galaxies and are not equally affected by environment (Bamford et al. 2009).

Table 1.1 shows a general correlation of some properties with environment. Some of these results may differ since as said before, the methods for density measurement or classification can lead to a diversity of results. In addition to that, during the correlation analysis of relationship there are *fixed* properties that lead to different conclusions, for example, before was it said that a fixed stellar mass lead to a dependence of color with environment and morphology do not have a relationship with environment Bamford et al. (2009) and, in Park et al. (2007) if morphology and luminosity are fixed, properties like color and star formation are nearly independent of local density.

The work presented here focuses on density estimation and classification in order to improve studies of correlations between environment and galaxy evolution. The density measurement used would be the method of *Lagrangian Tesselattion Field Estimator* (LTFE) (Abel et al. 2012) exclusively used in N-body simulations where particles can be traced and assigned to tetrahedral volume elements. As particles evolve during the simulation, density changes and is obtained by sampling the tetrahedral tessellation into a grid (Aragon-Calvo 2019). To improve the galaxy classification, the analysis would start at a larger scale, covering the Cosmic Web, where a relation between the density and the Cosmic Web features would be found using the method of Machine Learning. This

method would be used to test the hypothesis that there is a relation in density fields and their structure in the Universe. Machine learning consists in training a computer machine network using data of images of the Cosmic Web geometries and their density field. At the end of the training, with data not seen by the network, it predicts the density of a region based on its structure. The data used for training comes from the Illustris Simulation with a minor modification on the images due to their high resolution. They are modified in a lower resolution to diminish the noise in the learning phase since the computer machine would have to learn the details of the image and lead to more time of computation or could cycle in learning every detail.

The results are analysed plotting the predictions of the computer network and the real data for comparison. The code is run several times to obtain a statistic estimate of the accuracy of the prediction and to analyze the dispersion of predictions, the Pearson correlation and the moving window method are used. This helps to know if the predictions are close to the real data or how far they are, or maybe if the method fails at some values of density field.

This project is beginning to try to find and understand the dependence of galaxies and their environment. Once accomplish that computer machines could learn Cosmic Web structures, in the future, the machine learning approach would be used to expand the investigation to galaxy properties such as color, star formation and morphology by predicting the properties and the environment where they inhabit. Additionally, the methods and results could be use to classify the different regions of the Cosmic Web and predict what kind of galaxies would be found in those regions and their characteristics.

The thesis is divided in four sections:

- *Theory.* In this section concepts of Astrophysics will be explain. Concepts like the formation of the Cosmic Web and galaxies will be explained along with the definition of environment and the different methods used for its measure.
- *Method.* In this section the method of Machine Learning will be explained, the learning process and how the structure is made. Also, the data used for the training process and the types of training used for the analysis of data.
- *Results.* The predictions made by the computer machine are presented in this section. A comparison of the real data with the predictions along with some statistical analysis to know its efficiency and an analysis of the dispersion in predictions at different density fields.
- *Conclusions.* A brief summary of the presented results, future work and opportunity areas for Machine Learning in Astrophysics.

Chapter 2

Theory

2.1 Cosmology

Cosmology is a field in Astrophysics dedicated to the study of the Universe covering its formation, evolution and final stage as well as the formation of the Cosmic Web, its structure and density, and galaxies residing in it. In this chapter, these topics will be covered: i) the formation of the Cosmic Web, how is its structure and density, ii) galaxy formation, iii) environment, definitions and methods for its measure and its influence in galaxy morphology and properties.

Terminology: Density Measurement

In Cosmology, it is common to define the density in terms of the mean density as a perturbation:

$$\begin{aligned} \delta &\equiv \frac{\rho - \bar{\rho}}{\bar{\rho}} \\ \delta &= \frac{\rho}{\bar{\rho}} - 1 \\ therefore \quad \delta + 1 &= \frac{\rho}{\bar{\rho}} \end{aligned} \tag{2.1}$$

where ρ is the density measure by methods explained later and $\bar{\rho}$ is the mean density calculated as,

$$\bar{\rho} = \frac{N_{part}}{V_{box}} \tag{2.2}$$

being N_{part} the number of particles considered in a simulation box and V_{box} the volume of the simulation box.

The value of δ will be related to Cosmic Web density fields (Aragón-Calvo et al. 2010) as

- Overdensities $\delta \approx 100$
- Medium and low densities $\delta \approx 1 - 5$
- Underdense $\delta < 0$ (negative values)

2.1.1 Large-scale Structure

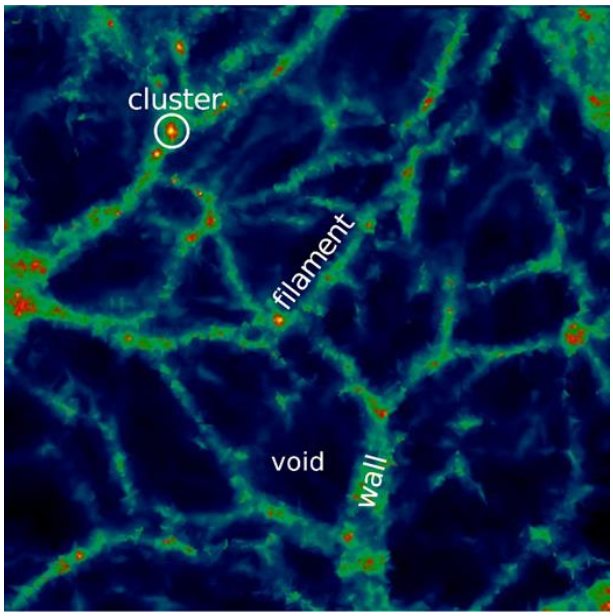


Figure 2.1: Distribution of density field in a N-body simulation with a box size of 32 $h^{-1}Mpc$. The regions of the Cosmic Web are highlighted representing cluster, filament, wall and void. Walls being represented as two dimensional sheets appear as projections. Voids are surrounded by walls and filaments and clusters are connected by filaments (Hui et al. 2018).

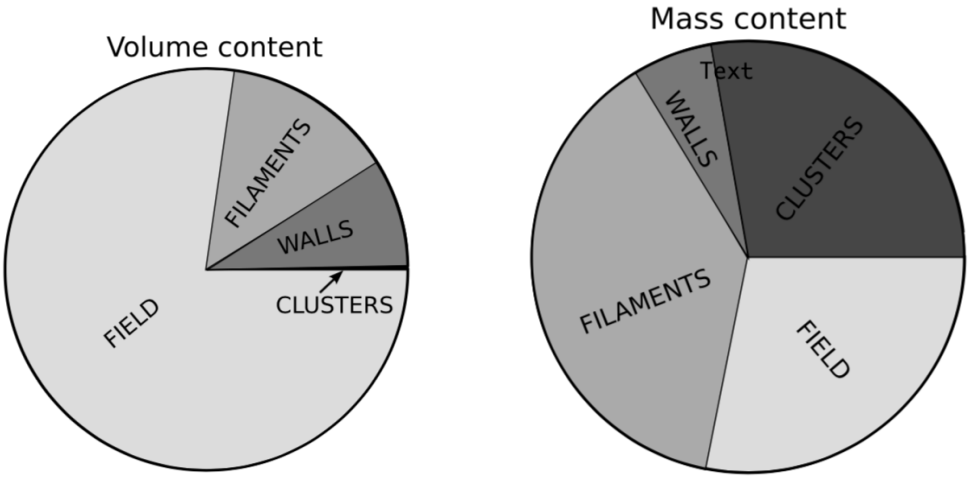
Dark matter, gas and galaxies are organized in a complex network across the Universe known as the Cosmic Web (Bond et al. 1996). The Cosmic Web, also known as the Large-scale Structure, was formed by gravitational processes acting on matter density after the Big Bang that were expanded as a result of inflation (stage after the Big Bang where space expanded rapidly) (Benson 2010). The structure formation is a hierarchical process where gravitational instability drives dark matter to cluster and collapse into stable objects, known as halos (White & Rees 1978) (Zel'dovich 1970). The dark matter halo is a near-equilibrium state supported by its own self-gravity by its constituent

particles (Benson 2010). The dark matter halos worked as potential wells where baryons were attracted and concentrate by the forces of gravity, being the sites of galaxy formation ((Tegmark et al. 1997), (Benson 2010)). This distribution evolved into cluster-filament-cluster configuration, the basis of the Cosmic Web ((Bond et al. 1996)).

The Cosmic Web matter distribution has a complex structure (Aragón-Calvo et al. 2010). In Figure 2.1 is a representation of the Cosmic Web mass distribution from an N-body simulation with a box size of 32 $h^{-1}Mpc$. The mass distribution generate different geometry patterns or features. The features of the Cosmic Web are the *i) clusters*, regions with the highest density; *ii) filaments*, the bridges between clusters where most of the cosmic mass is distributed; *iii) walls*, planar walls and sheet shapes surrounded by filaments, and *iv) voids*, having the lowest density being surrounded by walls and filaments. Voids are hard to define since they contain few or any light-emitting galaxies complicating their visibility (Davis et al. 1985; Bond et al. 1996). Figure 2.2 is also the distribution of matter using an N-body computer simulation called *Multiscale Morphology Filter* (MMF) (Aragón-Calvo et al 2007b). The result using this method is a map segmented in clusters, filaments and walls and evaluates the proportion of cosmic density field. Table 2.1 shows a summary of the features found in the Cosmic Web with their size, mass and volume distribution.

Feature	**Distribution**
Clusters: highest density	Mass: $10^{14}M_{\odot}$ and $10^{15}M_{\odot}$ Volume fraction: 0.38% Total cosmic mass: 28%
Filaments: bridges joining clusters with other clusters	Total cosmic mass: 39% Volume fraction: 10% Radius: $r \approx 2h^{-1}Mpc$
Walls: sheet shape	Total cosmic mass: 5.5% Volume fraction: 4.9% Widths: $5 - 8h^{-1}Mpc$
Voids: field regions, lowest density	Size: $20 - 50h^{-1}Mpc$

Table 2.1: Feature characteristics, mass and volume distributions of the Large-scale Structure. (Aragón-Calvo et al. 2010)



(a) Volume distribution in LSS regions.

(b) Mass distribution in LSS regions.

Figure 2.2: Pie diagram showing the distribution of volume and mass in the Cosmic Web computed from a ΛCDM N-body simulation (Hui et al. 2018).

2.1.2 Galaxy Evolution

As said in the previous section, galaxies form and evolve in dark matter halos (Dragomir et al. 2018). Since the galaxy evolution occurs within these halos, a combination of internal and external processes regulate their star formation and produce their different properties (Darvish et al. 2017) (Dragomir et al. 2018). These processes are presented in table 2.2. These mechanisms together with the evolution of galaxies inside dark matter halos, predict that galaxy properties and morphology are influenced by the formation and evolution of their host halos and environment (Somerville & Davé 2015; Muldrew et al. 2012).

Internal processes	· *Feedback (stellar feedback and AGN's)*: prevents all the gas in a galaxy from turning into stars on a very short time-scale due to heating process (Mo et al. 2010; Aragon-Calvo et al. 2016) (Darvish et al. 2017) · *Dark matter halo mass.* (Darvish et al. 2017) (Dragomir et al. 2018) (Goh et al. 2018)
External processes	· *Ram pressure and gas stripping*: a mechanism that removes gas and shuts off star formation (Abadi et al. 1999; Park et al. 2007) · *Harassment*: a high speed encounter of galaxies with other halos, typically in clusters, causing heating (Moore et al. 1996; Park et al. 2007) · *Strangulation*: a shut-off of the newly accreted gas when a galaxy enters a cluster or group environment and loses its gas reservoir (Balogh et al. 2000; Park et al. 2007) · *Tidal force field*: most view in clusters (Moss & Whittle 2000; Park et al. 2007)

Table 2.2: Internal and external processes for star formation quenching during galaxy evolution.

In this work, the definition of environment used will be the density field δ explained at the beginning of this chapter.

2.2 Environment Dependence

As seen in the previous section, the environmental mechanism influence the formation and evolution of galaxies leading to the different properties seen nowadays (see table 2.2). These influences have made galaxies posses properties at specific environments as shown in table 2.3 where some galaxy properties are mostly found either in high or low density environments.

Dense environment	Low density environment
· Massive galaxies (Bamford et al. 2009) · Early type (Bamford et al. 2009) (Blanton & Moustakas 2009) · Red population (Dressler 1980; Bamford et al. 2009) (Blanton & Moustakas 2009) · Lower star formation rate (Blanton & Moustakas 2009)	· Young galaxies (Blanton & Moustakas 2009) · Blue population (Dressler 1980; Bamford et al. 2009) · Star formation galaxies (Kauffmann et al. 2003)

Table 2.3: Types of galaxies in low and high density fields.

From table 2.3, notice that the properties of galaxies are related to their formation time. Old galaxies, being red, and with low or no star formation are found mostly in dense environments while blue late formation galaxies, considered as young galaxies continue to have star formation and are found mostly in low (and medium) environments. In some cases, early type galaxies can be found low and medium environments.

Even though there are galaxy types and Cosmic Web environment charaterization, there are still investigations of the environmental impact in galaxy formation and evolution since there are still mechanisms that do not explain the properties observed in a convincingly way.

An example of these investigations is the formation of lenticular galaxies (S0). Theories propose that these galaxies are spiral galaxies from which gas was removed and as a result they stopped forming stars (Dressler 1980). One proposed mechanism for gas removal is cosmic environment. Galaxies move and interact with Cosmic Web features, if a galaxy passes through a filament, the gas removal is carried out through the filament working as a filter (Aragon-Calvo et al. 2016) (Kuutma et al. 2017). Nevertheless, if this is true, these galaxies should be fainter than spiral galaxies, but what is observed is they are brighter than spirals, so their star formation is driven by a different process (Blanton & Moustakas 2009).

Another example of environment influence on galaxies is the interaction found in voids or low density environments. In these regions, there are star forming galaxies even though there is low or none material to transform gas into stars. A possible way to understand this interaction could be based on the observations of galaxies that show filaments between them transferring material (Kreckel et al. 2016), this way they acquire star-forming gas. Similarly, environment affects ellip-

tical galaxies found in voids since they are disturbed due to tidal effects, being stronger than in clusters (Blanton & Moustakas 2009) although the source of the tidal effects remains unknown. Note there are two interesting properties in the same environment: i) star formation in low density environment and, ii) elliptical galaxies in voids which no longer have star formation. Two opposite properties in the same environment which need more investigations in galaxy evolution to understand the processes happening at voids or low environments.

2.3 Density estimators

For environmental studies, there are several methods used to define the density value at a certain region of the Large-scale Structure of the Universe. In this section three methods will be explained to accomplish this measurement. The first two methods are more often used in observational data since it uses the number and distance of galaxies to determine the density (Muldrew et al. 2012). The last one is more often used in simulations since the evolution of particles, therefore matter distribution, can be tracked during all simulation, something that can not be done easily in observations (Abel et al. 2012).

2.3.1 Nearest neighbor

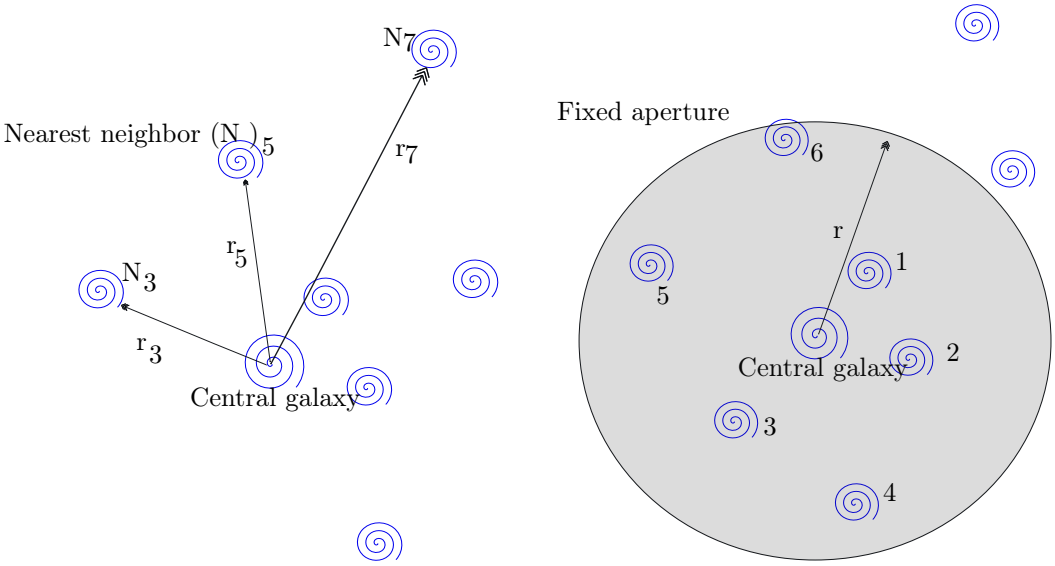
The nearest neighbor method is used to define density using the distance to a neighboring galaxy. The density definition is:

$$\sigma_n = \frac{n}{\pi r_n^2} \tag{2.3}$$

Where n is the number of the neighbors used for the analysis, at a distance r_n to the n^{th} neighbor. Figure 2.3 shows a representation of the measurement where density σ_n is proportional to the distance to the n neighbor as $\frac{1}{r_n^2}$ meaning that if the neighbor r_n is close, the density is higher, and vice-versa, the larger the radii the lower the density. The advantage of this method is that it is self-adaptive (works at local density). The disadvantage is the projected statistic due to the uncertainty in the true distance to the galaxies.

2.3.2 Fixed aperture (tophat)

The fixed aperture method consist on counting neighbor galaxies in a fixed area or volume in the region of interest.



(a) Nearest neighbor N_n illustration.

(b) Fixed aperture illustration.

Figure 2.3: Density measurements examples. For nearest neighbor there are multiple options to calculate density, could be the closest neighbor N_3, N_5, N_7 to the farthest neighbor N_{20}, N_{64}. For fixed aperture, the typical radii range is 1 $h^{-1}Mpc$ to 8 $h^{-1}Mpc$ (Muldrew et al. 2012).

$$\delta = \frac{N_g - \tilde{N}_g}{\tilde{N}_g} \tag{2.4}$$

Where δ is the measure of density contrast. N_g is the number of galaxies found in this aperture and $\tilde{N}_g$ is the average of the number of galaxies that is expected in the aperture if the galaxies where distributed randomly through the volume. Figure 2.3 shows a representation of the number of galaxies that could be considered for the measurement depending of the aperture radius.

2.3.3 Lagrangian Tesselattion Field Estimator (LTFE)

The Lagrangian Tessellation Field Estimator (LTFE) is a method proposed by Abel et al. (2012) (Shandarin et al. 2012) as a new way to compute densities in N-body simulations. It consists in approximating the distribution of matter by using tetrahedral volume elements defined by the Lagrangian particle grid. Tetrahedra are used since they are simpler and have the property of being convex. As particles evolve, density changes and is estimated by the fluid element described by the tetrahedral tessellation (Hui et al. 2018). Figure 2.4 gives a numbering of vertices (numbers in black) of the six tetrahedra which make up the cell (numbers in red) in a three dimensional space x, y, z. One tetrahedron contains one sixth of the mass of an N-body particle spread over its volume. The volume is calculated using the area of the tetrahedron, being vectors $\mathbf{b} \times \mathbf{c}$, and the

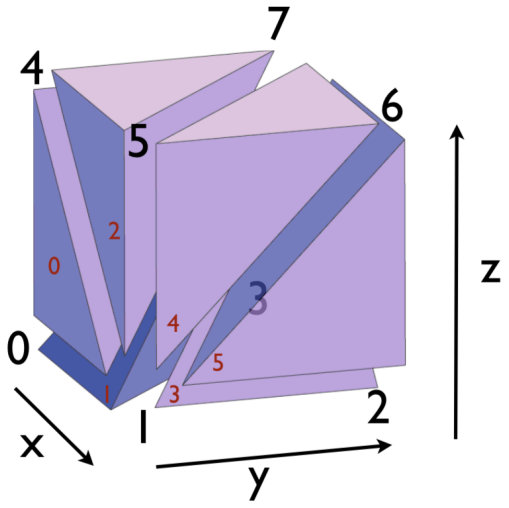


Figure 2.4: Representation of Lagrangian Tessellation Field Estimator (Abel et al. 2012).

height, being vector **a** so that,

$$\rho_s = \frac{m_P/6}{V_{tet}} = \frac{m_P}{|\mathbf{a} \cdot (\mathbf{b} \times \mathbf{c})|} \tag{2.5}$$

The Lagrangian Tessellation Field Estimator improves the measurement of the space density in low to intermediate density ranges in simulations (Hui et al. 2018). It considers dark matter as a fluid and can track the macroscopic and microscopic aspects of the dark matter fluid at every location during the full evolution of simulation.

Chapter 3

Methodology

3.1 Machine Learning Algorithms

Machine Learning (ML) refers to a set of algorithms able to make computer machines learn from data (Goodfellow et al. 2016) and carry out certain tasks like visual, voice and image recognition in an automatized way. The computer machines perform these tasks by learning algorithms using the human brain as basis. To train the computer, a large amount of data is required and the algorithm used would depend on the task complexity. In this chapter the Machine Learning method would be explained, which structure is used for this thesis and the basic concepts for performing tasks. Also, the data used for the training will be explain and how it will used for the analysis.

3.1.1 Perceptron

Figure 3.1 shows a representation of the *perceptron*, an artificial neuron, loosely based on the human neuron structure: dendrite and axon (ramifications). Neurons have a body where the core process the information, dendrites are the receptive zones and axons the transmission lines. In a perceptron a given input vector x_n is multiplied by a weight vector w_n and a bias b added to the result, producing a 1 (yes) or 0 (no) answer. The input vector could have m values $(x_1, x_2, .., x_m)$ and $w_i x_i$ is the dot product operation. The *bias* defines the height of a boundary plane that changes position according to the values assigned to w and b. If x lies above the boundary, then the answer is positive, otherwise it is negative. The perceptron cannot express a *maybe* answer (Gulli & Pal 2017).

There are three basic elements in the artificial neural model (Haykin 1999):

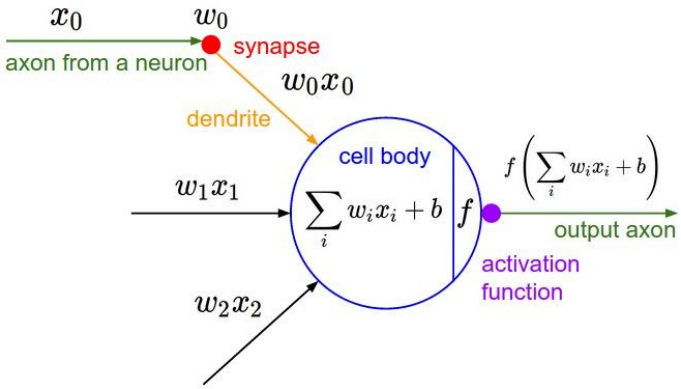


Figure 3.1: Analogy of human brain and neural network representation.

- *Synapse.* Characterized by an input signal x_i, multiplied by a weight w_i.
- *Adder.* For summing the input signals, the synapses. The operations describe a *linear combiner.*
- *An activation function.* Introduces a non-linearity such as limiting the amplitude of the output of a neuron to some finite value.

3.2 Neural Networks

A neural network is an algorithm architecture made of an ensemble of perceptrons well suited for a certain task such as pattern classification (Haykin 1999). It stores knowledge from its environment through a learning process by using *learning algorithms.*

In traditional neural networks, the structure of the network contains one single layer containing the perceptrons. There is also another network structure consisting of two or more layers allowing the network to learn complex tasks, this is called a deep network architecture (deep learning). In deep learning, the network learns complex tasks by hierarchically extracting important features from the input data (Haykin 1999). Both structures can be visualized in Figure 3.2, where the simple neural network (or traditional neural network) consists of one single layer interconnected to the input and output ramifications while in a deep learning neural network, the structure consists of two or more layers, each layer connected to the previous and next.

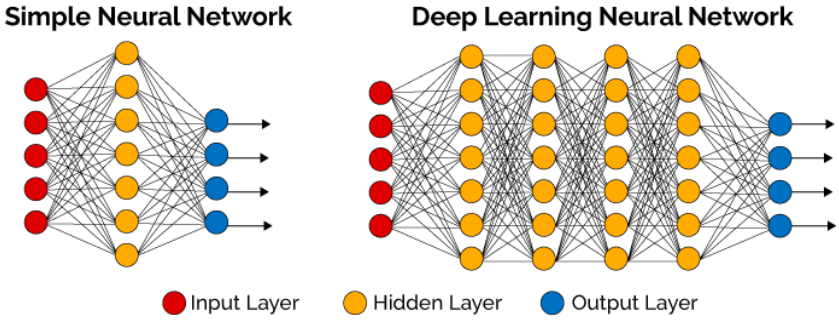


Figure 3.2: Traditional & Deep Learning Neural Network.

3.2.1 Convolutional Neural Networks (CNN)

Convolutional neural networks are mostly used for processing data that has a known, grid-like topology. This includes mostly images in $2D$ and $3D$, sound and text data for the identification of features and learning values from these features. The base of every convolutional network are the operations of convolution, activation function and pooling function (Goodfellow et al. 2016). These basic operations result in very few parameters, making the network faster and easier to train (Pouyanfar et al. 2018). To understand how the data is processed in this network, some important concepts will be define:

- *Convolution.* Has a different meaning in this context of neural networks unlike engineering and mathematics, the input data usually is an array and the network does a general matrix multiplication (Goodfellow et al. 2016). The output array is known as *feature map.*

- *Activation Function.* This operation works to introduce non-linearity to the learning. Unlike the definition said before where we used an activation function for a *yes/no* answer, in practice, a *maybe* answer is required where a continuous function is used to avoid sudden changes in the values of convolution. In this study is used the ReLu function which replaces all negative values of the feature map with zero and grows linearly for positive values (Gulli & Pal 2017).

- *Cost Function.* Is the function used to find the output error between the predictions and the true values and update weights while training the model (Gulli & Pal 2017). The cost function most commonly used is *mean squared error (MSE)* (equation 3.1). Sometimes also called *loss function.*

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 \tag{3.1}$$

Equation 3.1 measures the difference between the predicted values $\hat{Y}_i$ and the real data Y_i of a certain number of predictions n.

- *Optimizer.* This function determines how much the weights will be updated depending on the cost function value (Gulli & Pal 2017)

- *Pooling.* Pooling is a summary of the output values, keeping only the important information. The most commonly used function is *max-pooling* which works by defining a window size where the highest values of the feature map remain. This improves both computational efficiency and translation invariance.

- *Gradient Descent.* Is the partial derivative of the cost function used to find its minimum. The minimum is reached step by step using a learning rate α, if it is small, it will take a long time to reach it, if it is too big, it will pass the minimum and the network will not converge, therefore, it will not learn.

- *Fully Connected Layers.* This process is often placed at the end of the network, the analyzed image passes through all layers reaching the pooling layer, here the output of the last operation feeds a network where all information is gathered and displayed as the output value.

- *Batch size.* The input data cannot go through the network all at once, due to memory constraints, computational and learning time, that is why the data is split in different batch sizes. This way the training is faster since every time a batch enters the network, it updates the weights for the next batch making the learning more efficient.

- *Epochs.* Is the number of times the data goes through the network. Passing just once is often not enough since the majority of the details would not be learned by the filters. So, the data goes through the network several times, that way the filters are updated until the error is minimum and the learning is optimized.

3.3 Visual Geometry Group Architecture (VGG16)

VGG16 (Visual Geometry Group) is one of the most commonly used architectures in deep learning applications [1]. It consists of 16 layers and is the most used due to its simplicity, good performance and public availability of pre-trained weights computed from several well known data sets. This is the architecture that will be used for the data analysis presented in this thesis.

[1] For examples using neural networks including VGG architecture check `https://github.com/PacktPublishing/Deep-Learning-with-Keras`

VGG16 follows the standard architecture dividing the analysis into *feature extraction*, containing mostly convolutional layers and *classification* or *regression* containing exclusively fully connected (also known as dense) layers. The feature extraction section of VGG16 (see Figure 3.3 and 3.4) consists of 16 convolutional layers organized in 5 blocks with an increasing number of filters on the layers of each block. The layers in the first block and last block contain 64 and 512 filters respectively, these feature images are flattened giving an output of $7 \times 7 \times 512$ features feeding the regression, which consists of several fully connected layers with increasing number of neurons and end up in a layer with one single neuron with no activation (linear).

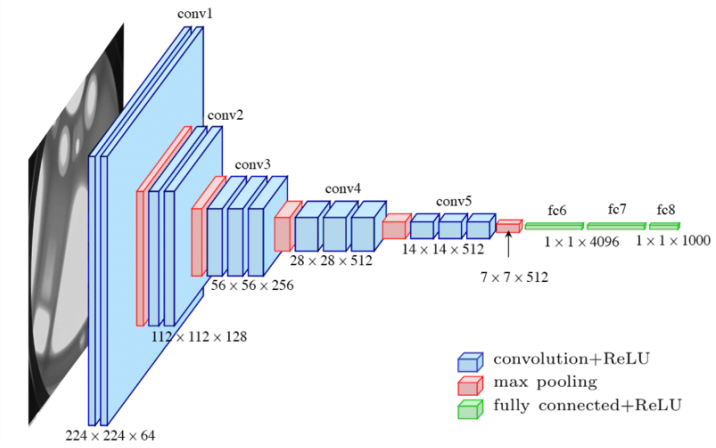


Figure 3.3: VGG16 arquitecture.

3.4 Training

While training the neural network, there are some modifications that improve the learning process. The modifications are carried out in the layers depending on the information wanted in the feature extraction section. The training process is explained below along with the modifications that could be used to have a better performance.

3.4.1 Types of training

- *Transfer learning* consists on training the model re-using the weights of another problem. This way the training does not start from scratch and the new model can be improved more easily. ImageNet (explained below) and VGG16 architecture are used for this kind of training.
- *Fine-tuning* is used to improve the features of an already trained network to adapt the model to a new task. This is achieved by changing the filter values to new ones related to the training

data. For this study, the last two layers of the VGG16 are modified in accordance to the images of the training data. In Figure 3.4 shows the structure of the VGG16 blocks where the last two layers (*conv2*) of block5 will be modified for fine-tuning.

3.4.2 Training process

The process of training is as follows:

1. An input image and random initial weights and biases are assigned.
2. The image passes forward through every layer, computing convolutions and generating a feature map.
3. The feature map goes through all 5 blocks of layers with their respective operations (as seen in section 3.3).
4. At the end, the fully connected layers aggregate all the information acquired from features and display an output value.
5. The output value will be used to calculate the error using the cost function and propagate it backwards to adjust the weights. This is achieved using the optimizer with the goal of reducing the error.
6. Weights are modified and the process is repeated with the next batch of images.
7. This continues until the defined number of epochs is reached.

3.5 Database

In this study an image is considered as a $2D$ matrix arrangement, where every pixel has a set of numbers that represent its color in three filters, RGB (red, green and blue). For a color image, the range of color goes from 0 to 255. In a black and white image, there is only one matrix in the same range, 0 for black and 255 for white. In order to use VGG16 with the simulation data we replicate the image corresponding to the density field across the three color channels expected by VGG16.

Principal Component Analysis (PCA). A linear transformation of the data projected in a way that the elements are uncorrelated preserving as much of the information in the data (Goodfellow et al. 2016).

Weights. In this study we use weights computed from *ImageNet* which is an image dataset with 1000 classes used for research in multimedia data (Stanford Vision Lab 2016).

3.5.1 Illustris Simulation

The data used in the analysis was obtained from the publicity available Illustris Simulation (Illustris 2018) . The Illustris Simulation consists of a set of runs on a 75 $h^{-1}Mpc$ box containing dark matter and baryons, and computing full hydrodynamics at several mass resolutions. The cosmological parameters follow the concordance model from WMAP-7 i.e. $\Omega_m = 0.27$, $\Omega_\Lambda = 0.73$, $h = 0.704$.

In order to generate density fields using the LTFE method low-resolution version were run based on the Illustris-3-Dark run[2] with initial conditions placed on a regular grid (see Figure 4.1), in contrast to the glass IC distribution used in the original Illustris Simulation. The grid IC were generated by interpolating the velocity field of the earliest snapshot at $z = 46.77$ into a regular grid using the Delanay Tessellation Field Estimator (DTFE) method and using the interpolated velocity field to compute the original displacement field. Additionally the IC were smoothed with a Gaussian window of 1 $h^{-1}Mpc$ in order to remove small-scale structures that may affect the learning process. The analysis presents a low-resolution run to compute density fields and trace the Large-scale Structure. This way the neural network will focus only on the large-scale structures. The data is used along with the publicity available sub-halo catalogs computed from the high-resolution Illustris run. Visual inspection shows that, at $z = 0$, the large-scale features in the grid-IC simulations and the original Illustris-3-Dark are practically identical (see next section 4.1).

[2]http://www.illustris-project.org/data/downloads/Illustris-3-Dark/

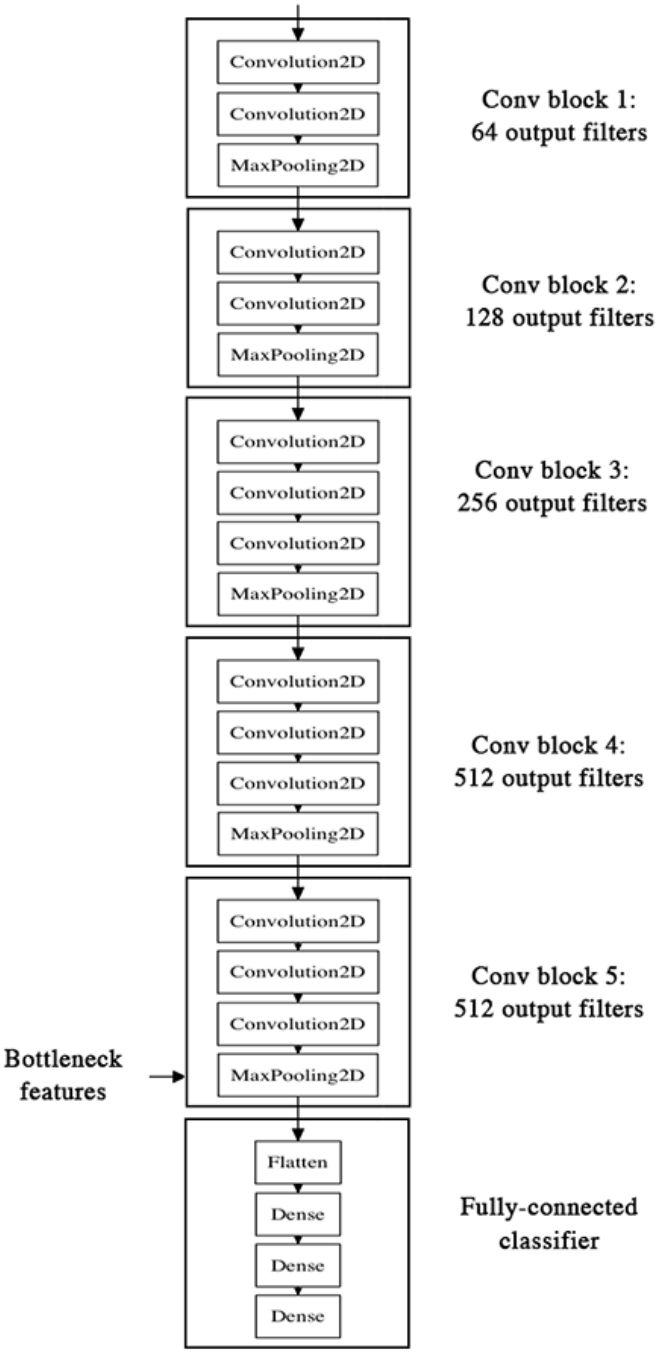


Figure 3.4: VGG16 arquitecture with block layers and number of filters.

Chapter 4

Analysis & Results

In this chapter, the predicted results obtained by the convolutional neural network model are shown. In addition to that, an analysis of the database structure, how they were obtained and different training methods are shown. In order to analyze the efficiency of the method presented here, statistics are computed over a moving window and Pearson correlations.

Pearson correlation

The Pearson correlation p is a measure of the strength of a linear association between two parameters. The values are in the range of -1 to $+1$. A value of 0 indicates there is no relation, otherwise the values are correlated, either in a positive (for $+1$) or negative way (-1). The closest the value to 0, the greater the variation. For this thesis Pearson correlation function is provided by the Python library *scikit-learn*.

Moving window

The moving window technique is used to calculate the difference between the real and predicted values inside a bin.

Keras

Keras is a Python library used for fast experimentation and prototyping using neural networks. It has different types of neural layers, cost functions, optimizers, activation functions, etc.

4.1 Data preparation

The training data is composed of a set of $10,000$ images and its corresponding density values separated in two groups: $9,000$ sets for training the network and $1,000$ sets for testing. The test data consists of images the network will not see during the training.

The images were created from 2D projections of the full density field. The density field was computed using the Lagrangian Tessellation Field Estimator (section 2.3.3) interpolated on a 512^3 grid. At the location of each sub-halo (this is a substructure of a halo) in the simulation a slice of 128^2 was cut across the z direction centered in the sub-halo (see Figure 4.1).

In order to ensure that the neural network is not simply learning the mean density of the image the individual images are normalized in the range $(0-1)$, thus removing any density offset and forcing the neural network to focus on the geometric patterns.

At the location of every sub-halo, densities were computed on a window of $1\ h^{-1}Mpc$ and $2\ h^{-1}Mpc$ using the simulation particles instead of the sub-halos in order to measure the unbiased density (see Figures 4.2a, 4.2b and 4.3.)

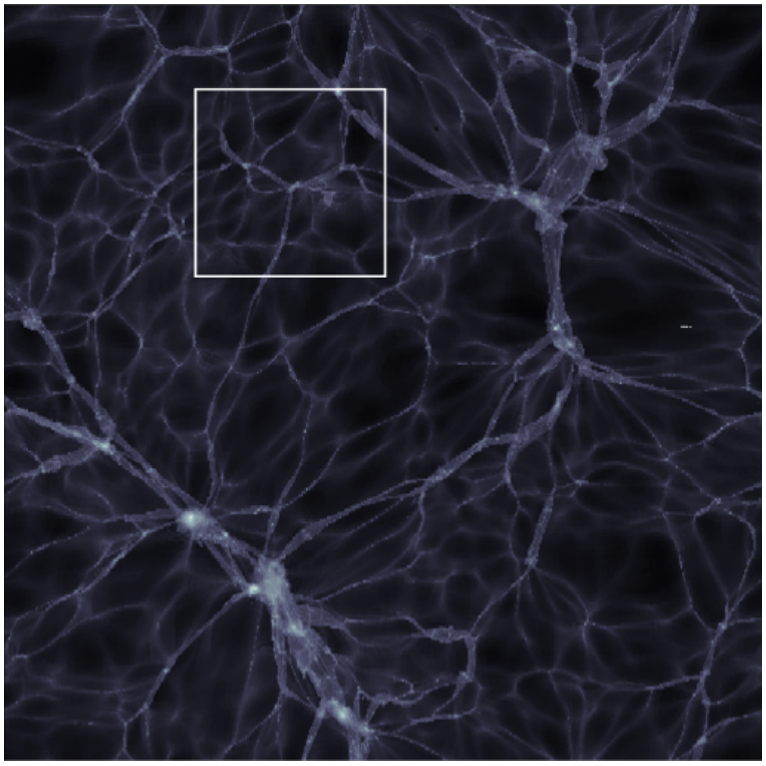
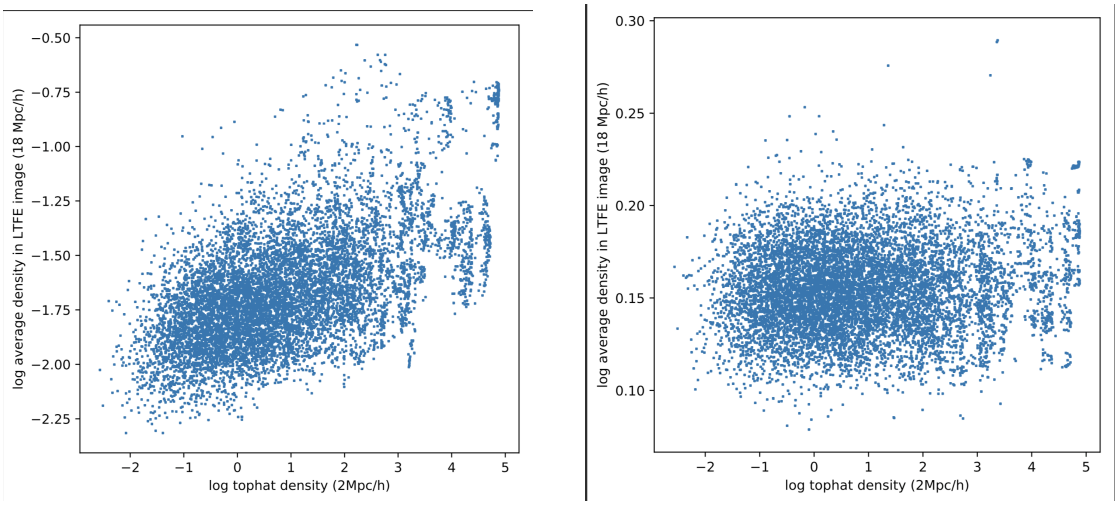


Figure 4.1: Training image (square region) extracted at the location of a halo. The square is 128 pixels wide corresponding to $18h^{-1}Mpc$.

Figure 4.2 shows a comparison between the 2 $h^{-1}Mpc$ tophat density field and the mean density inside its corresponding 128^2 window with and without normalization to remove the density offset. As the figure shows the normalized dataset contains no information on the local density, this is, images does not have bright peaks that could be confused with high density fields. Note that even without normalization the correlation between the tophat density and the image mean is weak with a Pearson correlation of $p = 0.65$. Comparing Figure 4.2b with Figure 4.2a where the same density estimate is shown but using an image with an equivalent width of $2\,h^{-1}Mpc$. In the case where the two density scales are similar the estimates are also highly correlated (Figure 4.3). These images serve as first visualization of density calculations comparing the densities calculated by the LTFE method and Illustris Simulation densities. In Figures 4.2a and 4.3 the densities are normalized, so there are two conclusions here. First, the density measurements using the LTFE method are close to the ones used by Illustris Simulation, and these measurements are not affected by using a data cube of 18 $h^{-1}Mpc$ for figure 4.2a nor using a 2 $h^{-1}Mpc$ data cube for figure 4.3. And second, keeping the densities normalized ensures that the network is going to learn by visualizing images (the geometric structure) and will not learn for example that an image with a high density corresponds to a bright image or vice-versa, a low density will have low brightness. On the other side, Figure 4.2b shows the densities for the LTFE method and the Illustris Simulation densities but the difference is that these densities are not normalized, so they are not related to each other and therefore are not used for training the network.

4.1.1 Density images in feature space

In order to visualize how the network decomposes the images into features and to asses the validity of our method, PCA was used to re-project the feature vector at the end of the feature extraction part of the network into two principal components (see section 2.1) . While two components are in principle not enough to fully quantify the variation in the input images it is sufficient for visualization purposes. Figure 4.4 shows the features in the layer *block3conv2* (see Figure 3.4) projected on the plane using PCA (re-projecting from $802,816$ to 2 features). Note that the neural network is able to group images based on their structure into several groups. Inside each group the images are spatially sorted (in the PCA space) possibly reflecting of minor variations in their features. Figure 4.4 shows that the network is able to map the complex structure of the images into a much smaller set of features and this can be exploited to associate image complexity with a scalar value such as density.



(a) (b)

Figure 4.2: Comparison between densities computed with a tophat radius of 2 $h^{-1}Mpc$ and the mean density inside its corresponding 128^2 image (18 $h^{-1}Mpc$). The original images (left panel) and the normalized images in the range $0-1$ (right panel). After normalization there is no correlation between the tophat densities and the mean density inside the image. The discrete features in the high-density range correspond to sub-halos inside large common halos.

4.1.2 Neural Network model

The model was written in Python using the machine learning Keras library. Figure 4.5 shows the sequence of the training pipeline.

Using figure 4.5 as reference, the deep neural network model used was the VGG16 architecture described in section 3.3 using pre-trained weights from the ImageNet dataset (section 3.5). A dataset of 9000 density field images extracted from the Illustris simulation was used, and for testing the model 1000 images were used (see section 3.5.1. Two types of training were used: transfer learning and fine-tuning (section 3.4.1). The model configuration included the mean squared error as cost function and *Adadelta* as optimizer. For the training model, we used a batch size of 5 samples and trained for 30 epochs since it was observed, using the output value of the loss function during every training, that this value was sufficient to achieve optimal learning.

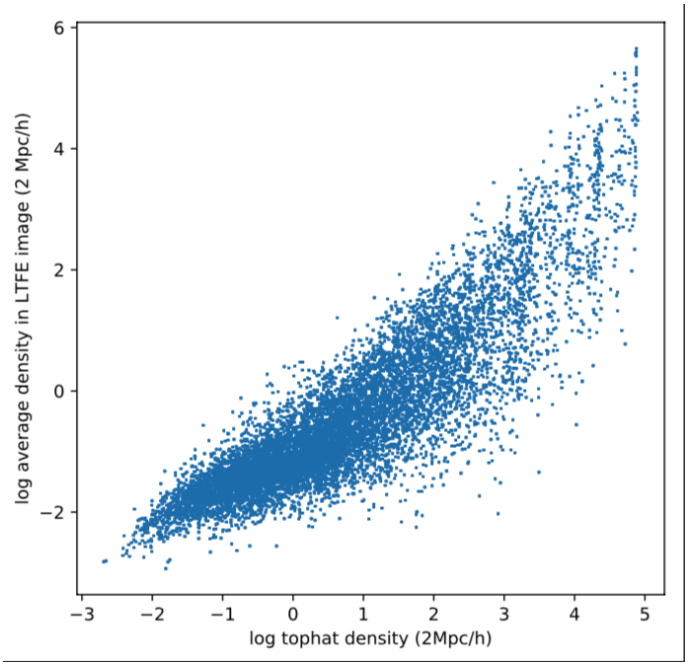


Figure 4.3: Comparison between densities computed with a tophat radius of 2 $h^{-1}Mpc$ and the mean density inside its corresponding 128^2 image (18 $h^{-1}Mpc$). The original images and the normalized images in the range $0-1$ are shown. Note there is a correlation between the original images and the normalized.

4.2 Results

In this section we show the results of training a neural network using transfer learning and fine tuning for the prediction of local density based on 2D slices as described in the previous section. The accuracy of the predictions was quantified using the Pearson Correlation and computing the dispersion σ of the errors $x_i - \bar{x}$ where x_i is the real value and $\bar{x}$ the prediction.

4.2.1 Transfer Learning

Figure 4.6 shows the predicted density vs. the measured (tophat) density inside a 1 $h^{-1}Mpc$ and 2 $h^{-1}Mpc$ tophat window computed after training the neural network using transfer learning (section 3.4.1) with pre-trained weights from the ImageNet dataset (section 3.5). In both cases the predicted values are close to the measured ones with a dispersion of $\sigma = 36$ at 1 $h^{-1}Mpc$ and $\sigma = 3.7$ for 2 $h^{-1}Mpc$ and Pearson correlations of $p = 0.975$ and $p = 0.996$ for the 1 $h^{-1}Mpc$ and 2 $h^{-1}Mpc$ samples respectively.

Interestingly the 2 $h^{-1}Mpc$ predictions has a bias towards lower values starting around $\delta+1 >$

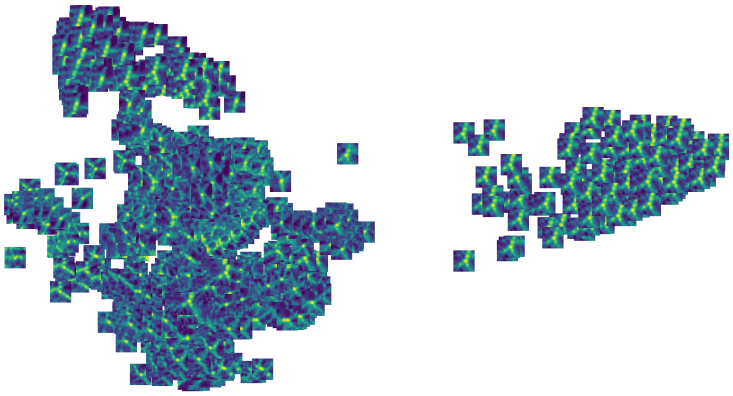


Figure 4.4: Images projected into the feature space computed at layer *block3conv2* in VGG16 after re-projection into 2 main PCA components. Note the clustering of visually similar images and the few intermediate images lying between clusters of images.

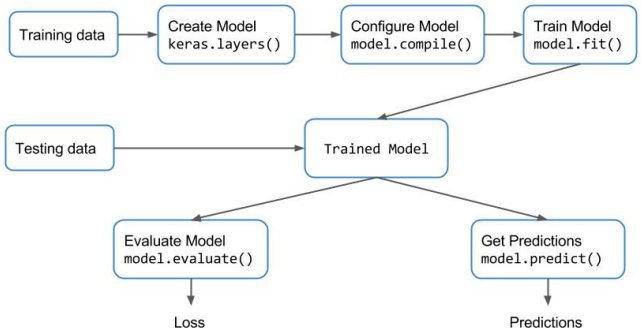
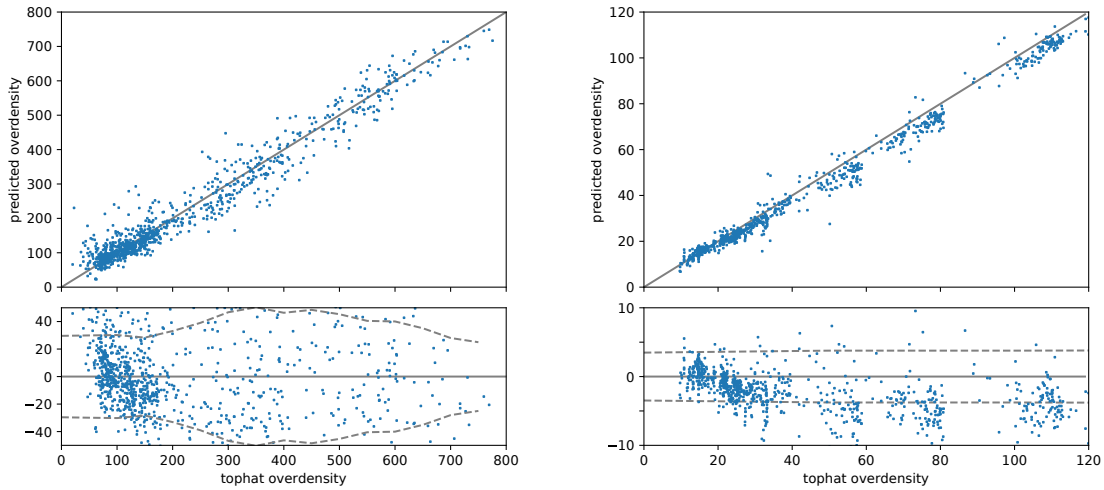


Figure 4.5: Keras workflow to perform the algorithm.

20. The deviation of 1σ is consistent across the dense regions.

Note that in both cases the dispersion is slightly smaller for density regimes consistent with filaments ($\sigma \sim 20 - 30$) and increases for higher densities. At high density regions (corresponding to clusters) the dispersion increases as there are large variations in density inside clusters and the density field itself is highly stochastic, making it difficult for the neural network to accurately predict the density based on the local geometry. On the other hand, the highest densities show a lower dispersion. This could be the result of the network overfitting (i.e. memorizing the density) in those particular regions due to the limited sample of massive clusters in the Illustris simulation. Given the relatively low number of samples in the high density end of the distribution the analysis would not be able to identify if the network is overfitting. A larger simulation volume would be



(a) Measured density vs predicted density at $1\ h^{-1}Mpc$

(b) Measured density vs predicted density at $2\ h^{-1}Mpc$

Figure 4.6: Predicted vs real densities computed using a pre-trained network with transfer learning. Results for tophat densities computed at $1\,h^{-1}Mpc$ (left panel) and $2\,h^{-1}Mpc$ (right panel) radius.

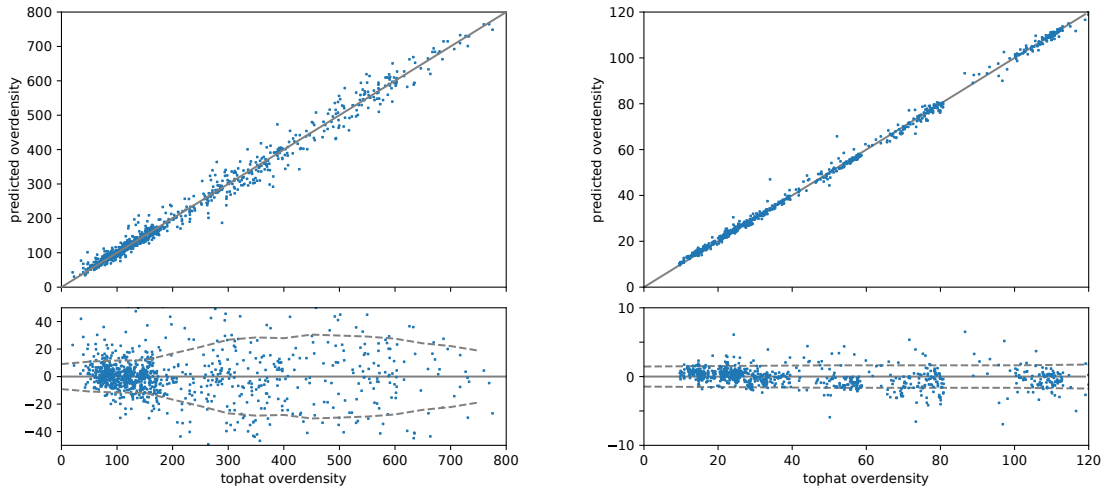
required to test this.

The clusters of points observed are also the result of the small volume of the simulation used here as several sub-halos belong to the few massive clusters in the simulation values. There are few large clusters with high densities in the volume compared to the much larger number of filaments and walls.

4.2.2 Fine-tuning

Figure 4.7 shows the predicted density vs the measured density at 1 and 2 $h^{-1}Mpc$ radius computed after training the network using fine-tuning (section 3.4.1), this is, with 2 unlocked layers in the feature extractor. This allows the feature extractor to adapt its filters to the particular features of the training dataset. With "fine-tuned" filters a better prediction is expected compared to plain transfer learning (Figure 4.6). Figure 4.7 shows that both $1\,h^{-1}Mpc$ and $2\,h^{-1}Mpc$ samples have predictions with lower dispersion than their counterparts trained with transfer learning ($\sigma = 18.8$ and $\sigma = 1.63$ respectively). The dispersion in the error of the predictions after fine tuning is roughly half compared to transfer learning. The Pearson correlation is also higher with values of $p = 0.993$ and $p = 0.999$ for the $1\,h^{-1}Mpc$ and $2\,h^{-1}Mpc$ samples respectively.

Note that in both transfer learning and fine tuning the error in the $1\ h^{-1}Mpc$ sample is larger than the $2\ h^{-1}Mpc$ sample. This is due to the larger volume sampled by the $2\ h^{-1}Mpc$ tophat



(a) Measured density vs predicted density at 1 $h^{-1}Mpc$

(b) Measured density vs predicted density at 2 $h^{-1}Mpc$

Figure 4.7: Predicted vs. measured (tophat) densities computed using a fine-tuned network. Results for tophat densities computed at 1 $h^{-1}Mpc$ (top panel) and 2 $h^{-1}Mpc$ (bottom panel) radius.

window, which makes the density measurement less sensitive to small-scale variations.

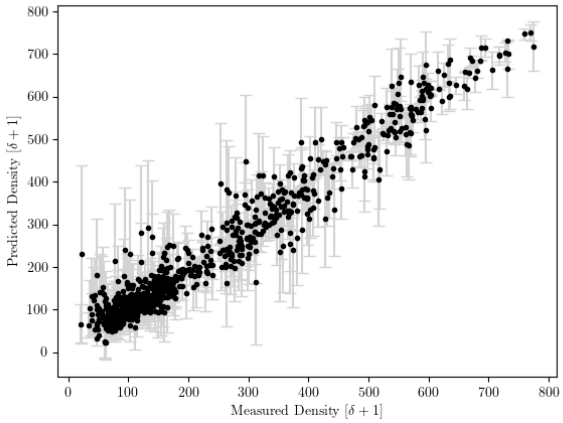
4.2.3 Ensemble training

In order to have a better understanding of the stability of the training process (i.e. how predictions from different training runs differ) an ensemble of 49 independent trainings was run using the same model architecture but different initial values for the network's learnable parameters. We trained the ensemble and computed predictions for a single test dataset. For each prediction across the ensemble we then computed the standard deviation (equation 4.1) at the end of all training runs as:

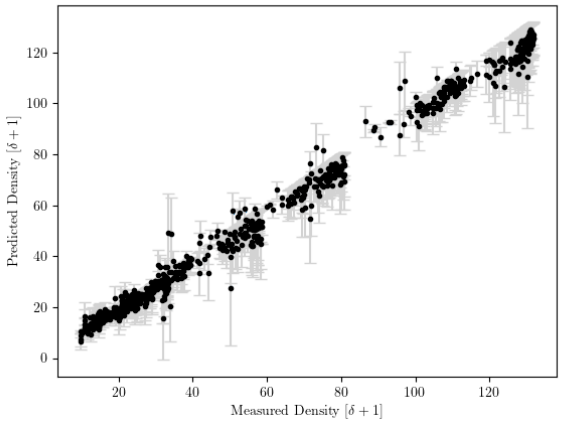
$$S = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \bar{x})^2}{N-1}}, \tag{4.1}$$

where x_i is the real value, $\bar{x}$ the prediction and N the number of predictions.

Figures 4.8 and 4.9 show the dispersion of the predictions. The mean density computed from the 49 predictions is plotted against the measured density considering the standard deviation (equation 4.1) as the error. The dispersion is larger for transfer learning (Figure 4.8) compared to fine-tuning (Figure 4.9) in a similar way as with individual training runs (Figures 4.6 and 4.7).

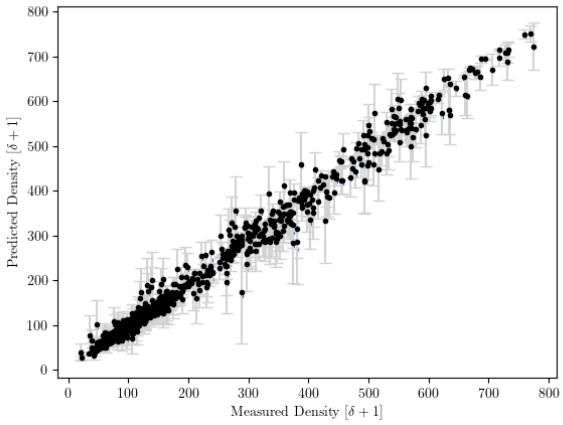


(a) Dispersion in transfer learning at 1 $h^{-1}Mpc$

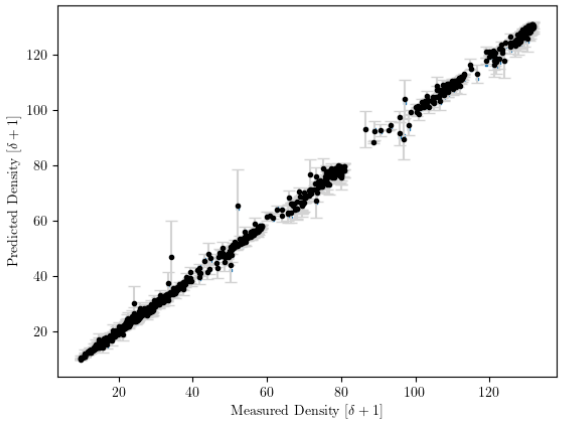


(b) Dispersion in transfer learning at 2 $h^{-1}Mpc$

Figure 4.8: Dispersion in both trainings at 1 $h^{-1}Mpc$. The function *error bar* from the Python library *matplotlib* was used for this analysis. The mean prediction was calculated with the Python library *statistics* and the standard deviation (equation 4.1) was computed by hand to ease the data manipulation.



(a) Dispersion in fine-tuning at 1 $h^{-1}Mpc$.



(b) Dispersion in fine-tuning at 2 $h^{-1}Mpc$.

Figure 4.9: Dispersion between fine-tuning training runs. The function *error bar* from the Python library *matplotlib* was used for error bars. The mean prediction was calculated with the Python library *statistics* and the standard deviation (equation 4.1) computed by hand.

The ensemble dispersion for transfer learning decreases with density for the 1 $h^{-1}Mpc$ sample from $\sigma_{\delta=50} \sim 200$ to $\sigma_{\delta=800} \sim 50$. On the other hand the ensemble dispersion for the 2 $h^{-1}Mpc$ remains comparatively constant.

The mean of the ensemble is consistent with the individual predictions, suggesting that there is no more information available in the density field images and that the neural network architecture used here is complex enough to encode the features in the density field. On the other hand, the large variation between predictions in the ensemble (as shown in the high ensemble dispersion) highlight the stochastic nature of the training process due to the choice of the random initialization of the learnable parameters in the network. In cases where the network training converges to a local minima the predictions can be negatively affected.

Chapter 5

Conclusions & Discussion

In this work, a deep convolutional neural network was successfully trained and tested for the prediction of local densities based on images of the Large-scale Structure. Training was done using two approaches: transfer learning and fine-tuning. In transfer learning the neural network re-used the feature extractor of a pre-trained network and trained connected layers. In fine-tuning, layers were unlocked in the pre-trained feature extractor to allow the filters better adapt to the input data. The network was trained with measured densities inside a 1 $h^{-1}Mpc$ and 2 $h^{-1}Mpc$ tophat window and feed with images of the density field centered at the tophat window.

Both transfer learning and fine-tuning produced excellent predictions as seen in table 5.1. Comparing the error dispersion using transfer learning vs. fine-tuning was found that fine-tuning significantly lowers the variation between the predicted and real values at 1 $h^{-1}Mpc$, while at 2 $h^{-1}Mpc$ it is more notorious this difference, being half the dispersion. There is no significant difference in the Pearson Correlation between different training runs at the 1 $h^{-1}Mpc$ and 2 $h^{-1}Mpc$ distances. Values of the Pearson Correlation are in the range of $p = 0.97 - 0.99$. This means the dispersion σ lowers to half the value in fine-tuning trainings unlike transfer learning trainings, so the predictions of densities are very close to the real measurements with fine-tuning, making the model of convolutional networks a good option for future investigations in the Large-scale Structure properties.

Training	Distance	Dispersion	Pearson correlation
Transfer learning	1 Mpc/h	$\sigma = 36$	$p = 0.975$
	2 Mpc/h	$\sigma = 3.7$	$p = 0.996$
Fine-tuning	1 Mpc/h	$\sigma = 18.8$	$p = 0.993$
	2 Mpc/h	$\sigma = 1.63$	$p = 0.999$

Table 5.1: Summary of predicted data dispersion in transfer learning and fine-tuning at 1 $h^{-1}Mpc$ and 2 $h^{-1}Mpc$

Fine-tuning provides the best density estimates due to its ability to adapt the pre-trained filters to the problem at hand. The pre-trained weights were trained using the ImageNet dataset which contains millions of images over 1000 categories of daily life common objects such as animals, furniture, cars, etc. It is remarkable that the model using transfer learning is able to accurately predict densities based on highly specialized filters tuned for a completely different domain. By fine-tuning the model the error can be reduce by half for both 1 $h^{-1}Mpc$ and 2 $h^{-1}Mpc$ samples. The improvement seen after fine-tuning indicates that a neural network specifically trained on astronomical data, like Illustris images, instead of a dataset like ImageNet, would be a valuable contribution to the Astro/MachineLearning community.

Despite the good performance of the network it is found that it tends to overfit at high-density regions. This is a consequence of the small volume of the Illustris simulation which results in a small number of large groups and clusters. The neural network seems to be memorizing the density inside those halos. When validating the network a new dataset is passed through but it likely contains sub-halos from the same large halos the network was trained with. This problem can not be alleviated with the current dataset. A simulation encompassing a larger volume with a large cluster population is needed.

The results indicate that for accurate density estimations an ensemble of independently trained models can be used. An ensemble has the advantage of proving a robust mean prediction and also an estimate of the dispersion for each predicted data point. One could potentially add to the ensemble small variations in the neural network model as commonly done in industry to test a variety of model combinations and improve manufacturing procedures or reduce costs.

Artificial Intelligence as a tool for hypothesis testing

Although the current work presents results on density predictions in the Large-scale Structure, the same approach can be used for other complex physical processes. A neural network can be used as a "black box" (as done here) to predict values from a dataset that would require significant

effort to analyze using standard techniques. It can be also used as a tool for fast hypothesis testing where there is no interest in getting accurate results but only in testing a given idea (like the relation between Large-scale Structure geometry and local density), is worth pursuing with more complex models or with traditional analysis tools. Using neural networks to guide an analysis for other properties like color, luminosity, halo mass, is a good start to identify relationships between properties and also use them to classify the Cosmic Web, this would point to promising relationships between the parameters of our system.

Bibliography

Abadi, M. G., Moore, B., & Bower, R. G. 1999, MNRAS, 308, 947

Abel, T., Hahn, O., & Kaehler, R. 2012, MNRAS, 427, 61

Aragon-Calvo, M. A. 2019, MNRAS, 484, 5771

Aragon-Calvo, M. A., Neyrinck, M. C., & Silk, J. 2016, arXiv e-prints, arXiv:1607.07881

Aragón-Calvo, M. A., van de Weygaert, R., & Jones, B. J. T. 2010, MNRAS, 408, 2163

Baldry, I. K., Balogh, M. L., Bower, R. G., et al. 2006, MNRAS, 373, 469

Balogh, M. L., Baldry, I. K., Nichol, R., et al. 2004, ApJ, 615, L101

Balogh, M. L., Navarro, J. F., & Morris, S. L. 2000, ApJ, 540, 113

Bamford, S. P., Nichol, R. C., Baldry, I. K., et al. 2009, MNRAS, 393, 1324

Benson, A. J. 2010, Phys. Rep., 495, 33

Blanton, M. R., & Moustakas, J. 2009, ARA&A, 47, 159

Bond, J. R., Kofman, L., & Pogosyan, D. 1996, Nature, 380, 603

Colless, M., Dalton, G., Maddox, S., et al. 2001, MNRAS, 328, 1039

Cui, Y., Xiang, Y., Rong, K., Feris, R., & Cao, L. 2014, 213–219

Darvish, B., Mobasher, B., Martin, D. C., et al. 2017, ApJ, 837, 16

Davis, M., Efstathiou, G., Frenk, C. S., & White, S. D. M. 1985, ApJ, 292, 371

D'Onofrio, M., Marziani, P., & Buson, L. 2015, Frontiers in Astronomy and Space Sciences, 2, 4

Dragomir, R., Rodríguez-Puebla, A., Primack, J. R., & Lee, C. T. 2018, MNRAS, 476, 741

Dressler, A. 1980, ApJ, 236, 351

Goh, T., Primack, J., Aragon-Calvo, M., et al. 2018, in American Astronomical Society Meeting Abstracts, Vol. 231, American Astronomical Society Meeting Abstracts #231, 258.13

Goodfellow, I., Bengio, Y., & Courville, A. 2016, Deep Learning (The MIT Press)

Gulli, A., & Pal, S. 2017, Deep Learning with Keras

Haykin, S. 1999, Neural Networks: A Comprehensive Foundation (Prentice Hall)

Hubble, E. P. 1926, ApJ, 64, doi:10.1086/143018

Hui, J., Aragon, M., Cui, X., & Flegal, J. M. 2018, MNRAS, 475, 4494

Illustris, C. 2018, ,

Kauffmann, G., Heckman, T. M., White, S. D. M., et al. 2003, MNRAS, 341, 54

Kreckel, K., van Gorkom, J. H., Beygu, B., et al. 2016, in IAU Symposium, Vol. 308, The Zeldovich Universe: Genesis and Growth of the Cosmic Web, ed. R. van de Weygaert, S. Shandarin, E. Saar, & J. Einasto, 591–599

Kuutma, T., Tamm, A., & Tempel, E. 2017, A&A, 600, L6

Mo, H., van den Bosch, F. C., & White, S. 2010, Galaxy Formation and Evolution

Moore, B., Katz, N., Lake, G., Dressler, A., & Oemler, A. 1996, Nature, 379, 613

Moss, C., & Whittle, M. 2000, MNRAS, 317, 667

Muldrew, S. I., Croton, D. J., Skibba, R. A., et al. 2012, MNRAS, 419, 2670

Park, C., Choi, Y.-Y., Vogeley, M. S., et al. 2007, ApJ, 658, 898

Pouyanfar, S., Sadiq, S., Yan, Y., et al. 2018, ACM Computing Surveys, 51, 1

Schneider, P. 2006, Extragalactic Astronomy and Cosmology

Shandarin, S., Habib, S., & Heitmann, K. 2012, Phys. Rev. D, 85, 083005

Somerville, R. S., & Davé, R. 2015, ARA&A, 53, 51

Stanford Vision Lab, Stanford University, P. U. 2016, ImageNet, ,

Sun, S., Guo, Q., Wang, L., et al. 2018, MNRAS, 477, 3136

Tegmark, M., Silk, J., Rees, M. J., et al. 1997, ApJ, 474, 1

van der Wel, A. 2008, ApJ, 675, L13

White, S. D. M., & Rees, M. J. 1978, MNRAS, 183, 341

Zel'dovich, Y. B. 1970, A&A, 5, 84