IMPLEMENTACIÓN DE TÉCNICAS DE MACHINE LEARNING EN LA EDUCACIÓN SUPERIOR[283]

IMPLEMENTATION OF MACHINE LEARNING TECHNIQUES AT UNDERGRADUATE LEVEL

Luz Ángela García Peñaloza[284]

Pares evaluadores: Red de Investigación en Educación, Empresa y Sociedad – REDIEES[285]

[283] Derivado del proyecto de investigación: Machine Learning. Entidad Financiadora: Universidad ECCI.

[284] Física, Universidad Nacional de Colombia. MSc Astronomía - Universidad Nacional de Colombia. PhD Astronomía - Swinburne University of Technology. Docente investigador - Universidad ECCI. Grupo SiAMo (Simulación, Análisis y Modelado). Bogotá, Colombia. lgarciap@ecci.edu.co

[285] Red de Investigación en Educación, Empresa y Sociedad – REDIEES. www.rediees.org

43. IMPLEMENTACIÓN DE TÉCNICAS DE MACHINE LEARNING EN LA EDUCACIÓN SUPERIOR[286]

Luz Ángela García Peñaloza[287]

RESUMEN

La coyuntura global causada por el COVID-19 nos ha llevado a repensar nuestros métodos de enseñanza. La educación superior no es ajena a esa transformación, y a consecuencia, los docentes e investigadores hemos implementado nuevas técnicas y metodologías en nuestros proyectos. En el campo de la astronomía toman cada día más fuerza los algoritmos de Machine Learning por su versatilidad y alto poder de predicción, que resulta de la puesta en marcha de estas técnicas a grandes conjuntos de datos. Existen varios proyectos que he implementado con mis estudiantes: el primero explora la potencialidad de las redes neuronales para mejorar la estimación de algunos parámetros cosmológicos del modelo estándar ΛCDM. El segundo utiliza un algoritmo no supervisado de K-means para clasificar cuásares (objetos muy luminosos y distantes), a partir de diferentes propiedades físicas observadas por el SDSS (Sloan Digital Sky Survey) en las campañas 12 y 14. A partir de los espectros de poderosos agujeros negros a grandes corrimientos al rojo, y otras propiedades físicas detectadas por el Sloan, generamos una clasificación de cuásares que presentan ciertas líneas de absorción.

El objetivo fundamental de este trabajo es mostrar el impacto que tiene la implementación de modernas técnicas de *Machine Learning* en proyectos de investigación con estudiantes de Ingeniería y las perspectivas de estos proyectos en otras disciplinas.

[286] Derivado del proyecto de investigación: Machine Learning. Entidad Financiadora: Universidad ECCI.
[287] Física, Universidad Nacional de Colombia. MSc Astronomía - Universidad Nacional de Colombia. PhD Astronomía - Swinburne University of Technology. Docente investigador - Universidad ECCI. Grupo SiAMo (Simulación, Análisis y Modelado). Bogotá, Colombia. lgarciap@ecci.edu.co

ABSTRACT

The ongoing COVID-19 crisis has led us to think over our teaching methods. The undergraduate education is not apart from the current transformation, and consequently, teachers and researchers have realized the need to apply new techniques and methodologies in our projects. Machine Learning algorithms have become very popular in Astronomy, for their versatility and high-power of prediction, that results from using these techniques in large data sets. There are a few current projects that I am developing with undergraduate students: the first one explores the potential of neural networks to improve the estimate of cosmological parameters of the ΛCDM standard model. The second work uses the unsupervised K-means algorithm to classify quasars (very massive, luminous and distant objects), from their distinctive physical properties observed by SDSS (Sloan Digital Sky Survey) in data releases 12 and 14. Based on the spectra from these powerful black holes at high redshift, and other physical properties detected by the Sloan, we classify quasars that present certain absorption lines.

The main objective of this work is to show the impact of implementing modern Machine Learning techniques in research projects with Engineering students and the perspectives of these proposals in other disciplines.

PALABRAS CLAVE: aprendizaje de máquinas, aprendizaje profundo, educación superior, redes neuronales.

Keywords: machine learning, deep learning, higher education, neural networks.

INTRODUCCIÓN

En la actualidad, contamos con volúmenes de datos sin precedentes. Todas las aplicaciones que usamos en nuestros dispositivos móviles, la implementación del internet de las cosas de forma masiva, e incluso, las grandes misiones y colaboraciones internacionales producen *data* constantemente. En el pasado, los datos generados se solían guardar en bases que se almacenaban en dispositivos físicos, siguiendo la Ley de Moore. Sin embargo, desde 2012 esta ley perdió validez, dada la rápida evolución que ha tenido la inteligencia artificial (IA), conduciendo a que la capacidad de cómputo se duplique cada 3 a 4 meses [1].

La astronomía no ha sido ajena a este escenario. Las grandes misiones y los cada vez más poderosos telescopios recolectan datos a escalas antes inimaginables, alcanzando los Petabytes (PB). Por ejemplo, para generar la primera fotografía de la sombra del agujero negro de Messier 87 en 2019 (de un poco más 10 kilobytes), se recolectaron alrededor de 5 PB, en los 10 radio interferómetros que apoyaron el Telescopio de Horizonte de Eventos (EHT, por sus siglas en inglés). De la misma manera, dado el volumen de información que se procesó en este proyecto, se requirió usar un algoritmo de Machine Learning, en conjunción con varias supercomputadoras.

De otro lado, el Square Kilometer Array[288] (SKA) proyectado para iniciar operaciones en el 2027 generará diariamente decenas de PB, que deben procesarse en tiempo real, pues no pueden ser almacenados físicamente. Asimismo, proyectos como Kepler[289] y otras sondas enfocadas en la búsqueda de exoplanetas están entregando millones de datos en periodos de tiempo cada vez más cortos. Esta nueva era de observaciones implica importantes desafíos para los astrónomos, que deben encontrar nuevos métodos para llevar a cabo el tratamiento de los datos de ciencia, sin que exista una pérdida de información en el proceso.

Las técnicas de *Machine Learning* han mostrado gran potencial para abordar y resolver problemas en astronomía, por su versatilidad de implementación en diferentes lenguajes de programación, capacidad de procesar en tiempo real ensambles estadísticos muy diversos, y en algunos casos, interpretabilidad de los resultados. Las herramientas de *Machine Learning* han mostrado ser altamente eficientes en extraer altos volúmenes de

[288] https://www.skatelescope.org/
[289] https://www.nasa.gov/mission_pages/kepler/overview/index.html

información, reducir los sesgos y la dispersión de los datos, identificar y analizar *outliers* que tiene información interesante del sistema, así como aglomeraciones entre los datos, describir correlaciones complicadas entre variables, clasificar observaciones, gestionar datos sin estructura, y generar datos simulados de forma rápida y "barata" [2].

Teniendo en cuenta las múltiples ventajas que presenta la adopción de *Machine Learning* en el estudio de problemas en Astronomía, entre otras disciplinas, se han planteado diferentes proyectos de investigación que cuentan con la participación de estudiantes de Ingeniería de la Universidad ECCI. Con el desarrollo de estos proyectos, los futuros profesionales son capaces de aplicar las nuevas herramientas de ciencia de datos en proyectos transversales a sus carreras, y perfeccionan sus capacidades de programación en diferentes lenguajes, pero en particular, en python.

Estimación de parámetros cosmológicos. El modelo estándar de Cosmología [3] se cimienta en el **Principio Cosmológico**, que establece que el Universo es homogéneo e isotrópico a grandes escalas, como se evidencia en *surveys* que rastrean la evolución en redshift de las galaxias, tal como 2dFGRS[290], 6dFGS[291], WiggleZ[292] y el Sloan Digital Sky Survey[293] (SDSS). En el futuro cercano, otros proyectos complementarán nuestra imagen de la estructura a gran escala del Universo: DES[294], DESI[295], EUCLID[296], LSST[297] y WFIRST[298].

Las observaciones de distancias de luminosidad de SNIa a finales de los noventas nos conducen a concluir que la expansión del Universo es acelerada (la tasa de expansión aumenta) en estos tiempos, indicando la existencia de una componente de energía oscura. En el modelo estándar la componente de energía oscura se describe constante en el tiempo, i.e. la constante cosmológica **Λ**.

Por otro lado, las curvas de rotación de estrellas en las galaxias no satisfacen una relación kepleriana, sino que la velocidad de rotación de estrellas lejanas con respecto al

[290] http://www.2dfgrs.net/
[291] http://www.6dfgs.net/
[292] http://wigglez.swin.edu.au/
[293] https://www.sdss.org/
[294] https://www.darkenergysurvey.org/supporting-science/large-scale-structure/
[295] https://www.desi.lbl.gov/
[296] https://www.euclid-ec.org/
[297] https://www.lsst.org/
[298] https://wfirst.gsfc.nasa.gov/

centro de la galaxia se mantiene constantes, independiente del perfil de la materia ordinaria de la galaxia. Estos resultados llevaron a Vera Rubin y colaboradores a sugerir la existencia un tipo de materia que no interactúa electromagnéticamente con nuestros telescopios, y por tanto no podemos "ver". Este tipo de materia oscura fría[299] (CDM por sus siglas en inglés), compone algo más del 80% de la masa total de las galaxias, pero es oscura porque no puede detectarse con los fotones que captan nuestros instrumentos. El 20% de la masa restante es materia bariónica u ordinaria, compuesta por los átomos de la tabla periódica.

Además, el modelo regente establece que todas las componentes materia-energía contribuyen como fuentes de gravitación, tal que $\boldsymbol{\Omega_\Lambda + \Omega_m + \Omega_k = 1}$, con $\boldsymbol{\Omega_\Lambda}$, $\boldsymbol{\Omega_m}$ y $\boldsymbol{\Omega_k}$ las fracciones de densidad de energía asociadas a la constante cosmológica $\boldsymbol{\Lambda}$, materia (bariónica y oscura) y curvatura, respectivamente. Observaciones de WMAP y PLANCK favorecen la idea de una curvatura espacial $\mathbf{k} = 0$, dejando la tercera componente, Ω_k, con una contribución cercana a cero. Este se conoce como el **Principio de Concordancia**, dado que $\boldsymbol{\Omega_\Lambda}$ y $\boldsymbol{\Omega_m}$ tienen el mismo orden de magnitud hoy.

El modelo estándar $\boldsymbol{\Lambda}$CDM puede describirse completamente con un conjunto de tres parámetros fundamentales $\{\boldsymbol{\Omega_m}, \boldsymbol{H_0}, \boldsymbol{\sigma_8}\}$, con $\boldsymbol{H_0}$ el valor de la constante de Hubble hoy, y $\boldsymbol{\sigma_8}$, la varianza en las sobredensidades de materia, medida en el espectro de potencias de materia en una esfera de 8 Mpc/h. El término $\boldsymbol{\Omega_\Lambda}$ puede derivarse del Principio de Concordancia, y $\boldsymbol{\sigma_8}$ puede reemplazarse por $\mathbf{A_s}$, el índice de fluctuaciones primordiales.

Hoy en día contamos con un gran número de instrumentos ópticos que nos permiten estudiar el cielo y, por tanto, entender los fenómenos que dan lugar a los observables. Sin embargo, la tarea de observar el cosmos reviste varias dificultades técnicas: uno, los telescopios en Tierra y en el espacio rastrean diferentes regiones del cielo, distintas épocas del Universo y longitudes de onda, por lo que su objetivo de ciencia se enfoca en fenómenos físicos muy específicos. De otro lado, el análisis estadístico no puede hacerse desde una perspectiva frecuentista, como se hace en otras disciplinas, pues estamos inmersos en el sistema de estudio (el propio Universo) y sólo contamos con una realización de este.

[299] Se le denomina *fría* porque la partícula que la describa debe tener una masa muy grande, y por tanto, se desacopla muy temprano del plasma

Un complemento a las técnicas observacionales (que resultan muy costosas de desarrollar y mantener, y que en muchos casos se han visto detenidas completamente por la presente pandemia del COVID-19) es el enfoque teórico a través de simulaciones numéricas de alta resolución. Las simulaciones permiten describir diferentes escenarios cosmológicos, no se ven afectadas por las condiciones atmosféricas, y permiten hacer predicciones de los fenómenos astronómicos aún no detectados, ya sea porque no se cuenta con la tecnología necesaria, o porque son regímenes inexplorados.

Entre las muchas ventajas que ofrecen las simulaciones cosmológicas de alta resolución es que permiten describir múltiples escenarios astrofísicos, construir diversas realizaciones de una región del Universo en una etapa dada, y, por tanto, explorar un espacio de parámetros que no es posible cubrir con las sondas existentes. Además, estos modelos facilitan la visualización de los observables de interés y, simultáneamente, permiten definir un rango de validez de los resultados obtenidos con modelos analíticos, admitiendo una extrapolación de las regiones de confianza y errores, que, de otra forma, no podrían acotarse con otros métodos. Sin embargo, cualquier simulación hidrodinámica requiere muchísimos recursos computacionales para ser generada, especialmente si se incluyen módulos sofisticados para mejorar la resolución de los fenómenos físicos que ocurren con los bariones. Por ello, las herramientas de ciencia de datos cada vez toman más fuerza en la comunidad.

Las simulaciones numéricas son muy útiles para entrenar algoritmos supervisados de *Machine Learning* (ML), y su eficiencia se incrementa conforme crece el número de realizaciones del sistema. Una de las aplicaciones más importantes de ML en astronomía es la estimación de parámetros cosmológicos extraídos de la estructura a gran escala con redes neuronales convolucionales profundas en 3D [4,5,6], así como modelos de desplazamiento de densidad, o D3M [7]. Otros trabajos incorporan redes de generación adversa tipo Wasserstein (WGANs) para detectar la emisión de 21 cm de hidrógeno neutro (HI) de radiotelescopios [8], o redes de aprendizaje profundo (U-Net) para generar rápidamente simulaciones con neutrinos masivos a partir de simulaciones cosmológicas estándar sin estas partículas [9].

Clasificación de espectros de cuásares. Uno de los objetos más brillantes y potentes del Universo son los cuásares. Estos objetos, bautizados así porque su emisión en radio se

asemeja a la de una estrella (*quasi-stellar object* o QSO, por su abreviatura en inglés) fueron detectados por primera vez en los años 50 con el uso de radio-antenas. Los cuásares son poderosos núcleos de galaxias activas, cuyo agujero negro central se encuentra circundado por un disco de acreción de gas que irradía con energías muy altas debidas a la interacción radiación - materia. Justamente la radiación que escapa de esta región es la que medimos en su espectro característico, cuya luminosidad es miles de veces mayor que la de la Vía Láctea.

Aunque aún no nos resulta del todo clara la naturaleza de la radiación que escapa de estas luminosas fuentes de radio, se ha establecido que sus propiedades dependen fuertemente de la masa del agujero negro central, la tasa de acreción de este, y su orientación relativa del disco con respecto a la línea de visión, así como la presencia de un jet ultra-relativista, y de polvo en los alrededores de la galaxia.

Cabe resaltar que estos objetos se encuentran a distancias cosmológicas, por tanto, la luz que recolectamos hoy con nuestros telescopios fue emitida mucho tiempo atrás. Entonces, los cuásares detectados a mayor distancia (o cuyas líneas en el espectro se encuentran más corridas al rojo) son mucho más viejos. El pico de su actividad se data a un redshift (corrimiento al rojo cosmológico) de 2, alrededor de 10 mil millones de años desde el Big Bang, pero se han detectado cuásares mucho más antiguos, siendo el récord hasta hoy de redshift z = 7.54, medido con el telescopio Keck, en Hawaii (ULAS J1342+0928).

Los surveys astronómicos han tenido una importante contribución en los censos astronómicos y han permitido establecer la validez del **Principio Cosmológico** a grandes escalas. Para el cierre de la última campaña de SDSS (Sloan Digital Sky Survey), se rastrearon más de 2 millones de galaxias y cuásares en un rango de 0.5 a 3.5, y con futuros *surveys*, se espera que este número crezca en más de un orden de magnitud.

Por ello, es necesario plantearse nuevos métodos para inspeccionar, clasificar y diagnosticar objetos astronómicos en tiempo real, a través del uso de las herramientas de ML que han mostrado ser muy efectivas en este tipo de tareas con crecientes volúmenes de datos.

Este segundo proyecto también se enmarca en la cosmología y consiste en la creación de un algoritmo de ML para agrupar y clasificar cuásares con el *survey* del cielo SDSS, específicamente con datos de BOSS (Baryon Oscillation Spectroscopic Survey) de las

campañas DR12 y DR14. La rutina usa las diferentes variables físicas registradas para cada cuásar, y el espectro de este, para definir si el cuásar presenta o no ciertas líneas de absorción.

Se encuentra en la literatura un trabajo de clasificación de cuásares que hace uso de datos liberados por SDSS DR14 [10]. Este implementa un algoritmo de redes neuronales convolucionales entrenada con los espectros de cuásares e identifica la presencia de líneas anchas de absorción (BALs, por *broad absorption lines* en inglés) para hacer un diagnóstico y comparación con la clasificación visual previa que hacen los expertos de SDSS. Después de probar el algoritmo con el conjunto de datos de testeo, el equipo confirmó la efectividad con datos de DR12, y propone hacerlo con datos de DESI.

Además, hay varios trabajos recientes de clasificación de cuásares a partir de la identificación de ciertas líneas de emisión y DLAs (*Damped Lyman-alpha systems*) [23] con *Deep Learning*. La eficiencia de las técnicas de aprendizaje profundo para identificar patrones visuales en las imágenes que no son perceptibles para el ojo humano ha hecho que crezcan notablemente el número de publicaciones que abordan problemas de clasificación de espectros en astronomía.

MATERIAL Y MÉTODOS

La metodología que se sigue en ambos proyectos tiene como objetivo principal enseñar a estudiantes de pregrado las técnicas y conocimientos requeridos para desarrollar un proyecto de investigación. Para servir este fin, se prepara a los estudiantes con los fundamentos de astronomía y cosmología requeridos para tener sensibilidad de la naturaleza de los resultados. Además, se les enseña las principales funciones y rutinas en python 3 que deben adaptar para realizar las tareas que implica el proyecto, y se les muestran los elementos de *Machine Learning* que se implementarán. Estas actividades se han realizado en forma telepresencial.

En particular para el primer proyecto se realizaron diferentes tareas previas a la aplicación de las técnicas de ML:

- Pedir acceso a diferentes conjuntos de simulaciones N-cuerpos con sólo materia oscura a diferentes grupos internacionales. Las simulaciones Auriga[300], Illustris TNG[301], COmoving Lagrangian Acceleration[302] (COLA) y Dark Sky Simulations[303] [11, 12] son públicas y pueden ser descargadas y utilizadas para responder diferentes preguntas de ciencia.
- Solicitar recursos de cómputo en una convocatoria internacional de HPC (*High performance computing).* Estas aplicaciones ocurren dos veces al año, y garantizan acceso a programas especializados, espacio de almacenamiento temporal, alta capacidad de RAM por núcleo de CPU y horas de supercómputo en millones de horas de CPU y GPU. Para presentarse a la aplicación, debe escribirse un documento con la justificación de los recursos solicitados y una propuesta científica clara.
- Discutir y elegir el algoritmo que mejor se adapta a las diferentes arquitecturas de las simulaciones numéricas. Para ello, se han consultado varios expertos en la disciplina y se han revisado extensamente los trabajos recientes en la literatura. La decisión final es el uso de redes neuronales, en dos de las versiones que se encuentran disponibles con programas *open source*.

Se debe garantizar un número significativamente grande de realizaciones para que pueda entrenarse el algoritmo numérico, y, en consecuencia, que los resultados tengan un alto grado de confianza desde el punto de vista estadístico.

Una vez que se satisfacen las condiciones antes mencionadas, se procede a caracterizar el conjunto de simulaciones, teniendo en cuenta los diferentes módulos físicos que soportan los modelos teóricos, los parámetros cosmológicos iniciales, la arquitectura de cada simulación y las semillas aleatorias que se introdujeron para generar sus condiciones iniciales. Para ello, se ha construido un *pipeline* especializado que permite leer, interpretar y comparar los *inputs* de las simulaciones disponibles. Los códigos numéricos se usan en conjunto con herramientas de visualización como **yt** y **rockstar** (software que identifica y

[300] https://wwwmpa.mpa-garching.mpg.de/auriga/
[301] https://www.tng-project.org/data/
[302] https://github.com/egpbos/pycola
[303] https://darksky.slac.stanford.edu/

caracteriza los halos de materia oscura en las simulaciones de materia oscura), entre otros. Este *pipeline* ha sido diseñado a partir de la experiencia con simulaciones previas como se discute en [13, 14, 15].

Una vez se han caracterizado completamente las corridas numéricas, se procederá a implementar el algoritmo de redes neuronales convolucionales (o CNN, por sus siglas en inglés) para estimar los parámetros cosmológicos {Ω_m, σ_8, H_0}, marginalizando los parámetros adicionales del modelo estándar a sus valores reportados por la colaboración Planck 2015 [16] y Planck 2018 [17].

De forma paralela, se ha venido adelantando trabajo en el proyecto de clasificación de cuásares. Para este proyecto no fue necesario solicitar recursos de cómputo, pues la base de datos de SDSS DR12 y DR14 es manejable con un ordenador personal. A parte de la fundamentación en tópicos de cosmología (modelo estándar, estructura a gran escala, modelo de cuásares y AGNs), el trabajo se ha centrado en la elección de la rutina de *Machine Learning* que estamos implementando para clasificar los cuásares.

La línea más importante de la discusión se ha centrado en revisar las ventajas y desventajas con los algoritmos disponibles, teniendo en cuenta la eficiencia del algoritmo, su interpretabilidad directa y la demanda de recursos. Finalmente, nos decantamos por el método **K-means**, que se explicará brevemente en la próxima sección.

DESARROLLO

Esta sección se enfoca en la descripción general de los algoritmos de *Machine Learning* implementados para el desarrollo de los proyectos propuestos. Su intención no es introducir los métodos matemáticos y estadísticos formales detrás de los modelos, pero se invita al lector a explorarlos más a fondo en las referencias donde han sido presentados.

Existen diferentes componentes que se han venido desarrollando en las últimas décadas en el campo de la inteligencia artificial, por ejemplo: la visión de computadora, el procesamiento de lenguaje, creatividad, el aprendizaje de máquinas, etc. Esta última (*Machine Learning* como se le ha denominado a lo largo del documento), propone algoritmos de clasificación, regresión, agrupamiento (*clustering)*, y redes neuronales, entre otras. El

aprendizaje profundo o *Deep Learning* (DL), aparece como una parte fundamental de ML, y ha generado resultados muy interesantes en los últimos 10 años.

El primer proyecto aquí descrito (estimación de parámetros cosmológicos con diferentes conjuntos de simulaciones numéricas de alta resolución) se apoya en el uso de herramientas de DL que han mostrado ser muy poderosas en comparación con otros algoritmos de ML, cuando se trata de ciertos problemas de alta complejidad, y a consecuencia, se tiene un excelente desempeño en diferentes dominios. Además, realiza una extracción automática de características del problema de forma muy eficiente, reduciendo así la introducción de sesgos en los datos. De otro lado, DL ofrece mucha flexibilidad en la elección de la arquitectura a usar, teniendo en cuenta que es soportada por diferentes *softwares*.

Sin embargo, DL reviste algunas dificultades en su implementación, como la gran cantidad de datos requeridos para su entrenamiento, así como los ingentes recursos de cómputo. También, existen algunas distribuciones que producen resultados muy complejos, lo que impide su interpretación directa por parte del usuario. Este último punto parece ser el resultado de la altísima especificidad que puede adaptarse para resolver un problema particular. No obstante, los pros en la implementación de DL han demostrado sobrepasar en gran medida los problemas previamente enunciados.

El algoritmo de DL elegido para desarrollar este primer proyecto es **Redes Neuronales Convolucionales en 3D** (CNN, o *Convolutional Neural Networks*) [18] para la extracción de los parámetros cosmológicos a un redshift fijo de las cajas, y **Redes Neuronales Recurrentes** (RNN, o *Recurrent Neural Networks*) para estimar la evolución con el tiempo en simulaciones consecutivas en redshift [19], teniendo en cuenta que todas las simulaciones que se comparan no están preparadas a un mismo redshift.

Las CNN [20, 21, 22] son una clase particular de redes neuronales diseñadas para procesar datos que se muestran en una grilla: en 1D, como una serie de tiempo; en 2D como una imagen o colección de píxeles, o en 3D como un cubo. La operación detrás de esta red es una convolución, una aplicación lineal que se usa en este caso para realizar operaciones matriciales entre las capas de la red.

A través de la aplicación de esta función de convolución (un conjunto de operadores lineales), la información de la capa de entrada pasa a la primera capa oculta, de esta a la segunda, y así hasta la última, que conduce a capas que realizan otras tareas. Para optimizar la función, se debe establecer una jerarquía, y por tanto, pesos específicos **W** y biases **b** necesitan ser definidos, así como una función de activación particular **Φ** que comunica las diferentes capas. La elección de la función de activación depende mucho del tipo de problema que se considere.

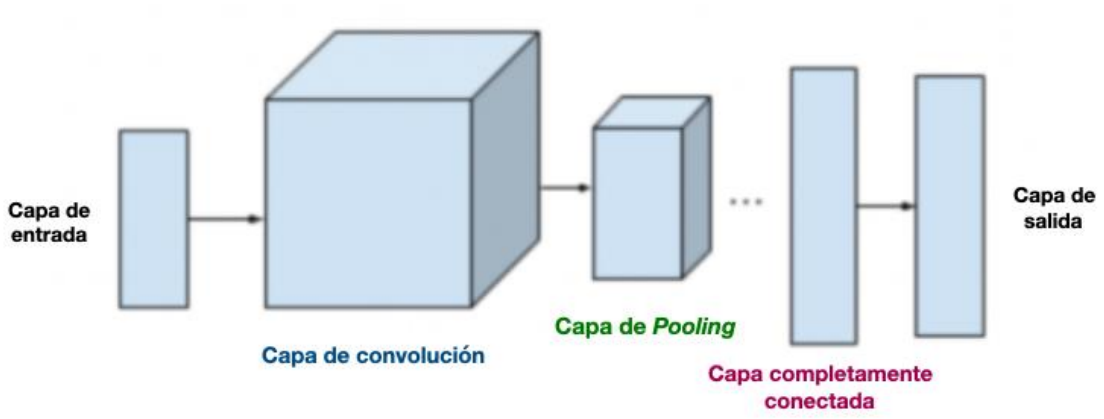

Figura 1. **Forma esquemática de las redes neuronales convolucionales. El proceso general consiste en una capa de entrada (o cubos de datos), una compleja red de convoluciones, seguida por una capa de *pooling*. Luego de esta etapa, se establece una capa completamente conectada, que da paso a la capa de salida, que contiene los parámetros estimados.**

El algoritmo de CNN suele introducir 3 capas principales en el procesamiento de los datos. La primera, la capa de convolución se enfoca en la extracción de características del conjunto de datos. Ésta aprende a encontrar patrones espaciales de la imagen de entrada, y se compone de múltiples capas que actúan como filtros (kernels convolucionales) sobre la imagen de entrada que reconocen las características más marcadas en la imagen que se usa para el entrenamiento.

La segunda capa del algoritmo es conocida como *pooling*, y el objetivo de su implementación es reducir la dimensionalidad de la imagen tomando los valores mínimos y máximos para construir segmentos de la imagen que no conservan los detalles. Una vez que este proceso toma lugar, los datos pasan a la capa completamente conectada. Los *inputs* de esta capa son justamente las imágenes de la última capa de *pooling* en cada nodo. La capa se

encarga de asignar puntuaciones para clasificar la imagen que se ha extraído de las capas anteriores, completando el proceso.

Con la implementación de este algoritmo puede hacerse simultáneamente una estimación de los parámetros que llevaron a la construcción de la imagen, teniendo en cuenta que la caja cosmológica puede entenderse como una colección de imágenes bidimensionales de la red cósmica con la impronta del modelo cosmológico asumido de base.

Por otro lado, las redes neuronales recurrentes (RNN) se especializan en procesar datos secuenciales (o valores consecutivos). Hay dos ventajas principales con este tipo de redes: uno, la posibilidad de procesar secuencias de datos de longitud variable, y dos, estas redes comparten parámetros a lo largo del modelo. Este último aspecto es muy importante cuando una parte de la información se repite o está presente en múltiples posiciones a lo largo de la secuencia. Esta propiedad de las redes neuronales de compartir parámetros permite aplicar el mismo kernel en cada paso dentro de la red.

Para el segundo proyecto propuesto se implementa la técnica de **K-means**, un algoritmo de ML no supervisado que realiza un agrupamiento a partir de las características reportadas para los cuásares por la misión BOSS de SDSS, en sus entregas públicas de datos DR12 y DR14.

La elección de un algoritmo no supervisado responde a nuestro interés de encontrar información en el conjunto de datos que no ha sido establecida por la colaboración BOSS. Debe recordarse que estos datos son liberados una vez que se han abordado ciertas preguntas de interés para los científicos principales del *survey*, y por tanto, es importante plantear un tipo de pregunta independiente a las ya tratadas por la colaboración.

Además, cabe mencionar que el método se entrena a partir de un conjunto de datos, y de esta forma se obtienen conclusiones en datos no etiquetados, o *clusters* que no se han establecido de antemano. En este caso, lo que se busca es seleccionar ciertas características del conjunto de entrenamiento para reducir la dimensionalidad de las variables redundantes o con información no determinante en el problema, conduciendo a una clasificación más clara de los cuásares con ciertas líneas de absorción. De resultar exitoso, el algoritmo divide los datos en subconjuntos con intersección nula, y los agrupamientos finales no reportan ningún tipo de estructura interna.

K-means es uno de los algoritmos no supervisados más sencillos para realizar *clustering* de datos, y se enfoca en encontrar los centroides de los *clusters* (o grupos de datos con características similares), con la expectativa que las regiones se diferenciarán claramente si el algoritmo se adapta bien a los datos. Básicamente, el método ejecuta dos tareas iterativamente: primero, se asigna cada punto al centroide del *cluster* más próximo (de forma aleatoria), y seguidamente, se define cada centroide como el punto medio de los puntos en dicha región. De esta manera, se busca minimizar la distancia del punto a su centroide, y maximizar la distancia entre centroides, usando una métrica determinada (en este caso, una métrica Euclidiana).

Uno de los principales problemas que suele encontrarse con este método es que la convergencia al máximo global en ciertas oportunidades queda determinada por la elección inicial de los centroides en los *clusters*. Por tanto, si el algoritmo encuentra un mínimo local y converge a él, el resultado final no es estrictamente correcto. Para evitar este problema, es recomendable correr múltiples veces el algoritmo, generando los centroides en posiciones iniciales aleatorias.

RESULTADOS

Uno de los retos más grandes para los docentes y educadores ante la pandemia es mantener el contacto directo con los estudiantes que participan en los proyectos planteados; esto sumado a las nuevas dinámicas que se imponen en la educación superior debido a la tele presencialidad. Por ello, buena parte de los proyectos con los estudiantes en este punto han estado enfocados en encontrar estrategias eficaces para mantener una comunicación fluida y los compromisos en continúo desarrollo. Aun así, uno de los principales impactos que ha tenido el COVID-19 es que muchas tareas se han visto postergadas debido a *deadlines* ajenos a los investigadores.

Por esto la mayor parte de los esfuerzos en el proyecto de estimación de parámetros cosmológicos ha estado centrado en caracterizar las simulaciones numéricas de alta resolución que nos han compartido en comunicación privada diferentes grupos internacionales. Las simulaciones con las que contamos hasta hoy se muestran en la Tabla 1.

Todas ellas se encuentran guardadas en contenedores electrónicos, a los que se tiene acceso por solicitud del PI del proyecto.

Estos meses se han destinado a crear un *pipeline* que permita leer, procesar y analizar las simulaciones, principalmente en python 3.7 y fortran 90. La metodología está pensada para hacer comparaciones directas entre las simulaciones, y así, establecer los principales *inputs* que podrían cambiar los parámetros constreñidos mediante la implementación de las redes neuronales. Debe tenerse en cuenta que cada simulación ha sido configurada y corrida con diferentes condiciones iniciales, distintos módulos semianalíticos que buscan describir propiedades físicas particulares.

Tabla 1
Simulaciones cosmológicas recolectadas para entrenar los algoritmos CNN y RRN para estimar los 3 parámetros cosmológicos más importantes del modelo estándar.

Nombre de la simulación	Tipo de simulación	Número de realizaciones disponibles	Observaciones
Auriga	Magnetohidrodinámica	30	Las simulaciones se corrieron a partir del código cuasi-Lagrangiano AREPO, que sigue la evolución de materia oscura con los componentes del gas.
Illustris TNG	Hidrodinámica	9	Existen realizaciones con y sin bariones. En este proyecto solo estamos considerando simulaciones con materia oscura.
Pycola	Hidrodinámica	1	Simulaciones basadas en el método de aceleración lagrangiana comovil (COLA) en los dominios espaciales y temporales.
Dark Sky	N-cuerpos +	5	Existen realizaciones desde

	Hidrodinámica	2014, y cada una tiene cajas de diferentes tamaños cosmológicos.

Fuente. Propia del autor.

Se presentan resultados preliminares que se han obtenido al comparar algunas de las simulaciones en la tabla 1. Este análisis preliminar implica leer los *inputs*, interpretarlos en unidades comóviles, ponerlas en escalas comunes para hacer una comparación justa de los parámetros de entrada, además de la creación de los árboles de evolución de los halos de materia oscura. Cada tipo de simulación tiene sus propios paquetes para realizar las tareas que se mencionaron, y no ha sido trivial obtener información equivalente en todas las realizaciones.

Las figuras 2 y 3 muestran proyecciones de la densidad y temperatura del gas en una región de la simulación Illustris TNG de la caja 110.7 Mpc, en una de las 100 *snapshots* que contiene esta realización. Los códigos de color muestran una evolución del rojo al amarillo, baja a alta densidad o temperatura proyectada, respectivamente. Estas proyecciones se obtienen al implementar el código público **yt,** que se usa en conjunto con python 3.

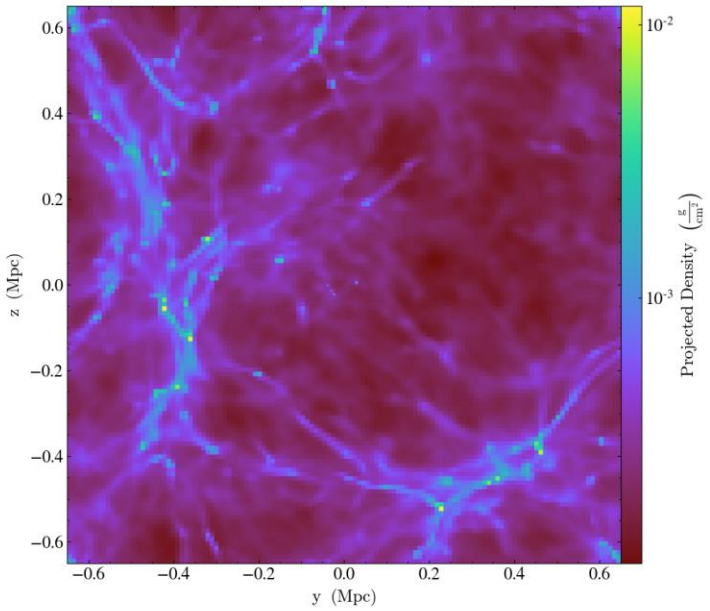

Figura 2. **Densidad proyectada en el plano yz (código de color) de una región de 1.2 Mpc alrededor de una sobredensidad de gas de la simulación Illustris TNG con el código yt.**

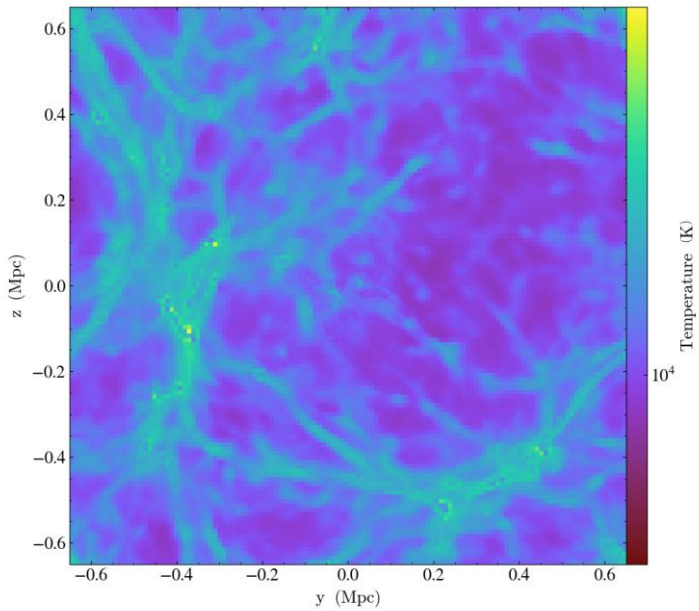

Figura 3. **Perfil de temperatura (código de color) con respecto a la densidaddel gas en el plano yz alrededor de una región de sobredensidad de la simulación Illustris TNG con el código yt.**

Para el proyecto de clasificación de cuásares con datos de SDSS, se ha caracterizado los espectros de los cuásares, mediante la identificación automática de líneas de absorción y emisión más relevantes. Se usan espectros sintéticos, construidos a partir de plantillas, tal cual como se muestra en la figura 4.

Las bases de datos de BOSS - SDSS, campañas 12 y 14 (DR12 y DR14, respectivamente) se caracterizan y comparan para definir qué variables se introducen en el algoritmo **K-means.** Contamos con más de 50 variables distintas por cada cuásar (nombre del objeto en el *survey* SDSS, posición en el cielo -ascensión recta y declinación-, ID del objeto, plato y fibra en el que fue detectado, fecha de detección en el calendario juliano, número del espectro, redshift detectado con su error respectivo con los métodos de: inspección visual, con la predicción del *pipeline* de la colaboración, y mediante el uso de PCA -por *Principal Components analysis*, en inglés-. Además, la predicción de redshift del cuásar anclada a la línea de MgII, los flujos en diferentes bandas de longitud de onda, el inverso de la varianza, la extinción asociado al polvo de la galaxia del cuásar, colores (diferencias entre flujos en distintas bandas), además de parámetros propios de BALs: AI y BI medidos en la absorción de CIV.

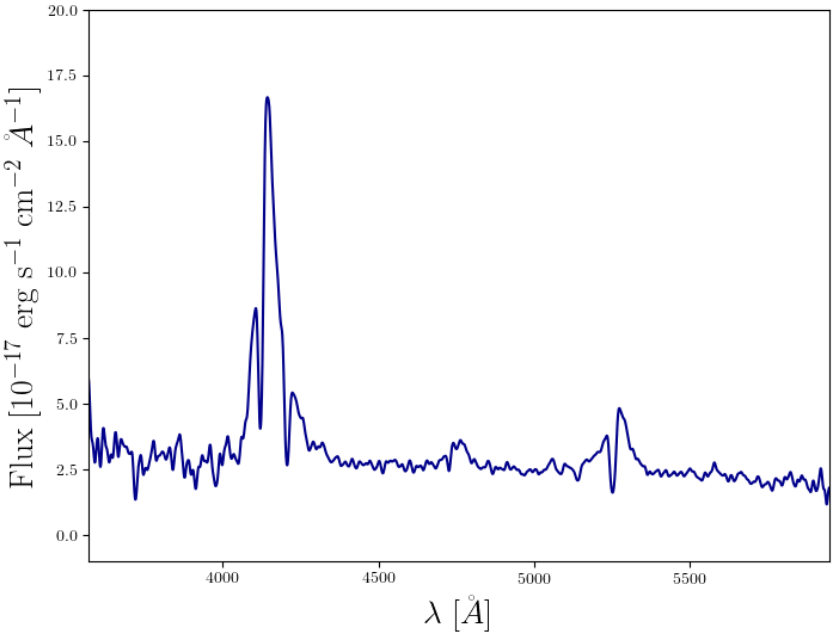

Figura 4. **Espectro simulado de un cuásar cuya luz se emitió cuando el Universo tenía 10 mil millones de años. La línea de emisión más pronunciada en la región de la izquierda corresponde a la emisión de Ly-alpha, característica de este tipo de objetos, que suele usarse para encontrar el corrimiento al rojo correspondiente.**

De otro lado, cabe aclarar que DR12 cuenta con 297301 cuásares, con un redshift máximo para el objeto más viejo de z = 6.44. En la campaña DR14 hay 526356 con un corrimiento al rojo cosmológico detectado de z = 6.95.

La figura 5 muestra la distribución en el cielo de los cuásares detectados por las campañas liberadas DR12 (en magenta) y DR14 (en azul), con una proyección de Mollweide. La figura puede compararse directamente con la figura 1 discutida en [24], con la única excepción que en ese trabajo se muestra en un código de color la densidad número en cuásares en las posiciones en el cielo exploradas por el *survey* en estas dos campañas.

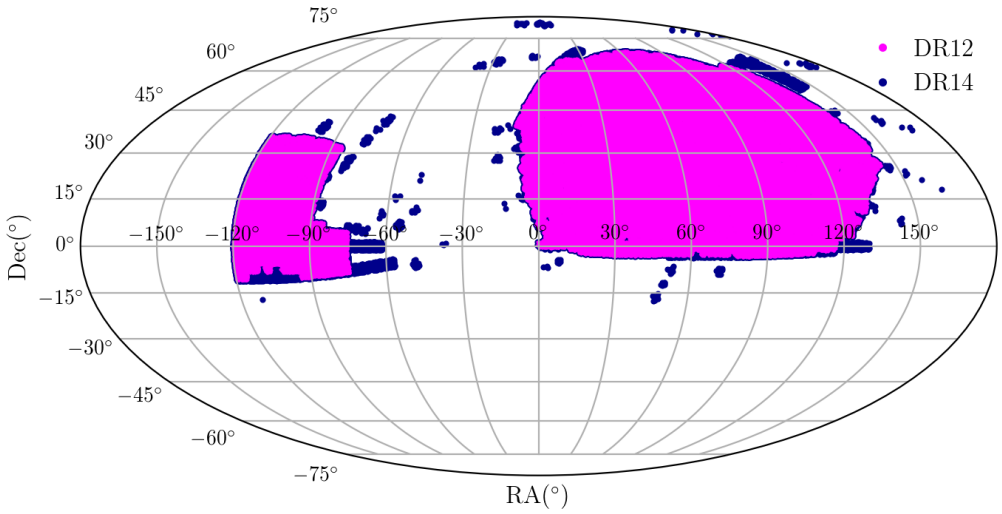

Figura 5. **Distribución en el cielo de los cuásares de BOSS DR12 (magenta) y DR14 (azul) en una proyección de Mollweide. La gráfica es comparable con la Figura 1 en [24], salvo que aquí no se hace una distinción por densidad número de objetos por grado cuadrado, sino en la distribución relativa por campañas.**

La figura 6 muestra la distribución de corrimiento al rojo cosmológico (redshift) de las líneas detectadas de los espectros de los cuásares de SDSS en las campañas DR12 (en magenta) y DR14 (en azul). Como se ve en ambos histogramas normalizados, el pico en la actividad de los cuásares ocurre a z = 2 (aproximadamente hace 4000 millones de años antes). Se produce un segundo pico en ambas distribuciones más tarde en el Universo, cuando z < 1, indicando que se han encontrado gran cantidad de cuásares a bajo redshift, gracias a que las primeras campañas de SDSS se han enfocado en rastrear este régimen. Las detecciones cubren un rango de redshift desde 0 (es decir, hoy) a redshift ~ 6, con casos muy raros, ya que el *survey* fue creada para detectar objetos hasta z = 4. Cuásares por encima de este límite parece ser el resultado de una mala identificación del redshift debido a la presencia de líneas de absorción como BALs o DLAs.

Debido al problema identificado arriba, de una mala clasificación de los objetos observados por SDSS, la implementación de técnicas alternativas a las usadas en *pipeline* de BOSS toma mayor importancia. El algoritmo de K-means que estamos usando revisa automáticamente el redshift detectado y su error por PCA, el inverso de la varianza y los parámetros AI-CIV y BI-CIV asociados a las líneas de absorción anchas (BALs) para dar un diagnóstico y clasificar el cuásar.

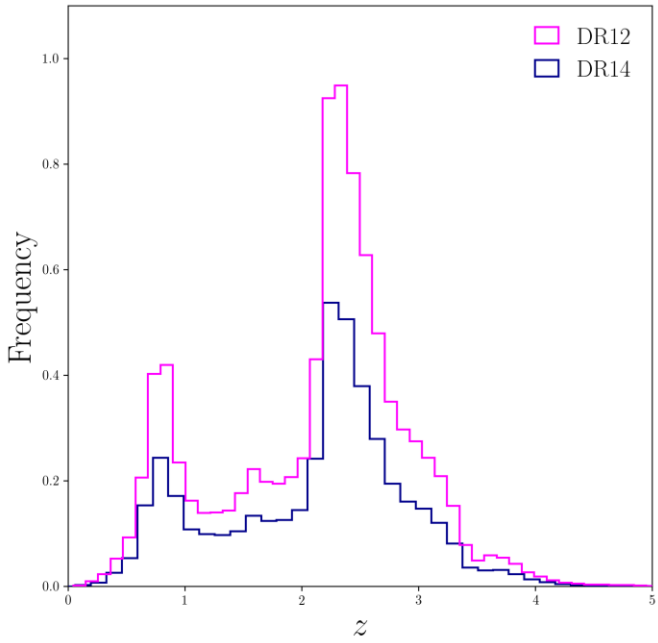

***Figura 6*. Distribución en redshift (corrimiento al rojo) de cuásares detectados por BOSS DR12 (magenta) y DR14 (azul).**

DISCUSIÓN Y CONCLUSIONES

Teniendo en cuenta que ambos proyectos están aún en desarrollo, los resultados preliminares presentados en este documento evolucionarán en los siguientes meses, una vez que se tenga acceso total a los recursos computacionales que se requieren para implementar la técnica de *Machine Learning*. No obstante, en este punto se ha logrado hacer un avance importante en la interpretación y comparación de las simulaciones de las que se dispone.

Los participantes han aprendido a explorar, manejar y manipular los códigos *open-source*, a partir del uso de técnicas y módulos especializados en cosmología, y de esta manera, también aprenden la dinámica que se sigue investigación. La adquisición y revisión de material bibliográfico ha sido esencial para el desarrollo de estos proyectos que les están generando formación en programación, inteligencia artificial y astronomía.

La investigación discutida en esta propuesta tiene múltiples proyecciones futuras. La consecuencia natural del desarrollo de este proyecto es la creación de una línea de investigación en ML y métodos computacionales, gestionado directamente por el grupo

SiAMo (Simulación, Análisis y Modelado), que se alinee a las líneas existentes en la Universidad ECCI. Esta línea puede ser un insumo importante para el desarrollo de tesis de pregrado y maestría en varios de los programas de la institución.

Cuando se piensa más específicamente en los tópicos de Cosmología abordados en este trabajo, aparecen diferentes tareas que pueden iniciarse con el *pipeline* desarrollado para el proyecto:

- Tener en cuenta simulaciones hidrodinámicas que incluyan bariones y múltiples módulos físicos no tratados en este proyecto.
- Incrementar el tamaño de la muestra de simulaciones, a través de la generación de más realizaciones por parte del grupo, y la solicitud a otros grupos que tengan simulaciones hidrodinámicas, sin intervenir en sus propósitos de ciencia. Esto generará resultados más sólidos, y fortalecerá las colaboraciones con grupos internacionales.
- Probar otros algoritmos de ML para mejorar las regiones de confianza del conjunto de parámetros estimulados.
- Usar esta técnica para revisar e interpretar futuras observaciones y/o modelos cosmológicos, más allá del modelo estándar actual.

Los proyectos aquí expuestos, aún en desarrollo, irán mostrando otras posibles perspectivas que pueden tratarse con este tipo de algoritmos de inteligencia artificial. Lo más interesante es abrir nuevas líneas de conocimiento para que los estudiantes que están terminando su formación universitaria puedan explorar proyectos interdisciplinares, aprender del quehacer en investigación, los retos y desafíos que se enfrentan, y en el futuro, encontrar una aplicación innovadora de esta metodología en su disciplina.

REFERENCIAS BIBLIOGRÁFICAS

D. Amodei, D. Hernandez, G. Sastry, et al. I. Sutskever. AI and Compute. https://openai.com/blog/ai-and-compute/. 2019.

M. Ntampaka, C. Avestru, et al. Astro2020 Science White Paper: The Role of Machine Learning in the Next Decade of Cosmology, arXiv:1902.10159v1

S. Dodelson. Modern Cosmology, Academic Press, Elsevier Science, 2003

S. Ravanbakhs, J. Oliva, et al. Estimating Cosmological Parameters from the Dark Matter Distribution, arXiv:1711.02033v1

S. Pan, M. Liu, et al. Cosmological Parameter Estimation From Large-scale Structure Deep Learning, Aceptado en ApJ. arXiv:1908.10590v4

J. H. T. Yip, X. Zhang, et al. From Dark Matter to Galaxies with Convolutional Neural Networks, Second Workshop on Machine Learning and the Physical Sciences (NeurIPS 2019), arXiv:1910.07813v1.

Siyu He, Y. Li, et al. "Learning to Predict the Cosmological Structure Formation", PNAS, vol. 116, 28, 1–8, 2019.

J. Zamudio-Fernandez, A. Okan, et al. Higan: Cosmic Neutral Hydrogen With Generative Adversarial Networks, arXiv:1904.12846v1.

E. Giusarma, M. R. Hurtado, et al. "Learning Neutrino Effects Cosmology With Convolutional Neural Networks", arXiv:1910.04255v1.

Z. Guo, P. Martini. Classification of Broad Absorption Line Quasars with a Convolutional Neural Network. The Astrophysical Journal, vol 879, issue 2, 72, 12, 2019.

S. Skillman, M. Warren, M. Turk, et al. Dark Sky Simulations: Early Data Release. arXiv:1407.2600.

M. S. Warren. 2HOT: An Improved Parallel Hashed Oct-Tree N-Body Algorithm for Cosmological Simulation. arXiv:1310.4502.

L. A. García, E. Tescari, et al. Theoretical study of an LAE – CIV absorption pair at z = 5.7, Monthly Notices of the Royal Astronomical Society: Letters, vol. 469, issue 1, pp. L53-L57, 2017.

L. A. García, E. Tescari, et al. Simulated metal and HI absorption lines at the conclusion of Reionization, Monthly Notices of the Royal Astronomical Society, vol. 470, issue 2, pp. 2494-2509, 2017.

L. A. García, E. V. Ryan-Weber. Can UVB variations reconcile simulated quasar absorption lines at high redshift? Revista Mexicana de Astronomía y Astrofísica, vol. 56, pp. 97-111, 2020.

P. A. R. Ade et al. Planck 2015 results. XIII. Cosmological parameters, arXiv:1502.01589.

N. Aghanim et al. Planck 2018 results. VI. Cosmological parameters, arXiv:1807.06209.

Y. LeCun, L.D. Jackel, B. Boser, et al. Handwritten Digit Recognition: Applications of Neural Net Chips and Automatic Learning. IEEE Communication, pp. 41-46, 1989.

D. E. Rumelhart, G. E. Hinton, J. L. McClelland. A general framework for parallel distributed processing. *vol 1: Foundations* pp. 45–76, 1986.

L. Deng, D. Yu. Deep Learning: Methods and Applications. NOW Publishers, 2014.

I. Goodfellow, Y. Bengio, A. Courville. Deep Learning. MIT Press, http://www.deeplearningbook.org, 2016.

D. Osinga. Deep Learning Cookbook: Practical Recipes to Get Started Quickly. O'Reilly Media, Inc., 2018.

H. Fathivavsari. Deep Learning Prediction of Quasars Broad Ly alpha Emission Line. arXiv:2006.05124v1

V. de Sainte Agathe, C. Balland, Hélion et al. Baryon acoustic oscillations at z = 2.34 from the correlations of Ly alpha absorption in eBOSS DR14. Astronomy & Astrophysics, 629, A85, 2019