Морфологическая классификация галактик с использованием машинного обучения

Аронов Михаил[1]; **Ерошкин Егор**[2]; **Магомедэминов Никита**[3]; **Липская Мария**[4]; **Романов Антон**[5]

Аннотация

Из-за развития телескопов появляются большие объёмы данных в астрономии, которые вручную будет довольно трудно и долго обрабатывать. Поэтому с помощью искусственных нейронных сетей или метода опорных векторов стали автоматизировать эту работу. Классификация галактик, квазаров, звезд и других объектов является очень важной для понимания их формирования и эволюции. Поэтому цель этой работы - при помощи искусственных нейронных сетей классифицировать галактики по морфологическим признакам. Нами была создана программа, которая научилась распределять галактики по камертону Хаббла. В дальнейшем, при открытии новых галактик, используя эту программу, можно будет определять их тип в несколько миллионов раз быстрее, чем это делает человек.

1 Введение

Наблюдаемые галактики имеют самые разнообразные формы, от ярких массивных эллиптических до вытянутых спиралей и компактных карликов. Одна из первых попыток классифицировать галактики по их внешнему виду была предложена (Wolf, 1908). Эти так называемые галактические туманности были расположены в соответствии с их формой, размером и отличительными особенностями. Внешний вид объектов отличается: от гладких эллиптических галактик, дископодобных галактик со спиральными рукавами до более неправильных форм. Изучение морфологической классификации галактик играет важную роль в астрономии: частота и пространственное распределение типов галактик дают ценную информацию для понимания формирования и эволюции галактик (R.J.Buta, 2011; MO H., 2010). В настоящее время активно проводятся обзоры неба и стремительно открываются новые объекты. За ночь на новых телескопах накапливаются огромные массивы необработанных данных. Релизы данных с телескопов, делающих обзоры неба (SDSS, Legacy Survey, GAIA) тоже имеют огромный объем данных. Копящийся пласт данных невозможно обрабатывать вручную. Для этих целей в астрономии используется технология машинного обучения (JT VanderPlas, 2014), позволяющая значительно ускорить обработку наблюдений. Мы проанализировали имеющиеся работы по применению нейросе-

[1]АНО Физтех-Лицей им. П.Л.Капицы
[2]ГБНОУ Губернский Лицей
[3]ГАОУ РМЭ Лицей Бауманский
[4]ГБОУ Лицей Вторая школа
[5]ГБНОУ Губернский Лицей

тей в астрономии (Ball, Nicholas M. 2007; Ball, Nicholas M. 2008). В нашей работе мы пользуемся нейросетью для классификации галактик по их морфологии при помощи фотографий, без использования спектров (A. O. Clarke, 2020). Нейросеть написана на python и работает на базе библиотеки Tensorflow[1].

Нейронная сеть — это система соединённых между собой и взаимодействующих между собой простых процессоров (искусственных нейронов). Нейронные сети не программируются в привычном смысле этого слова, они обучаются. Возможность обучения — одно из главных преимуществ нейронных сетей перед традиционными алгоритмами. Технически обучение заключается в нахождении коэффициентов связей между нейронами. В процессе обучения нейронная сеть способна выявлять зависимости между входными и выходными данными, а также выполнять обобщение. Это значит, что в случае успешного обучения сеть сможет показать верный результат на основе данных, которые отсутствовали в обучающей выборке, а также неполных и/или «зашумленных», частично искажённых данных. Таким образом, нейронные сети можно использовать для распознавания объектов и их классификации. Это имеет преимущество перед ручной сортировкой, так как хорошо обученная нейросеть может давать значительный выигрыш по времени при обработке больших объёмов данных.

Разметка баз данных нужна для облегчения задачи поиска нужной информации для следующих научных групп. В области классификации были проведены некоторые исследования[2] Однако предыдущие проекты нейронных сетей для классификации астрономических объектов были способны отличать галактики и квазары от звёзд, и более подробной классификации не проводилось. Мы же предлагаем обучить нейронную сеть упрощённой Хаббловской классификации галактик (эллиптические, дисковые и неправильные).

В разделе практика рассказано о том, как создавалась база данных и описано обучение машины. Наконец, в выводах приведены результаты классификации тестовых изображений.

2 Практическая часть. Создание модели

2.1 Воссоздание похожих проектов

Мы изучали программную библиотеку Tensorflow для написания классификатора галактик морфологически. В команде была трудность с корректной установкой этой библиотеки. Рабочими решениями оказались установка через *!pip install*, а также создание виртуального окружения в Conda. Так как никто в нашей команде до этого не работал в этой области программирования, то сначала мы решили посмотреть уже имеющиеся проекты с машинным обучением [3]. Также, параллельно, мы попробовали запустить простой классификатор из примеров Tensorflow, который был заточен на определение элемента одежды по фото[4]. Массив данных для обучения, состоявший из множества фото 28*28 был загружен из самой библиотеки. Классификатор состоял из трех слоев (keras.layers.Flatten, keras.layers.Dense, keras.layers.Dense). Далее модель успешно была обучена.

2.2 Создание нашей модели

Затем мы нашли более сложный и точный классификатор, состоящий из 12 слоёв, обучающийся в 5 эпох. Основная его задача – классифицировать цветы по фото[5]. Его мы использовали в качестве основы своей модели. Теперь для её обучения необходимо создание базы данных.

3 Практическая часть. Создание выборки и обучение модели

3.1 Обучение модели для классификации спектров

Для теста классификатора мы попробовали взять фотографии спектров галактик, звезд и квазаров из базы SDSS Dr14[6]. После процесса обучения, наша модель с большой точностью верно классифицировала спектры.

3.2 Выбор каталога

Для того, чтобы наша нейросеть работала исправно и с хорошей точностью относила галактики к определенному уровню, нам понадобилась тщательная выборка оптических фотографий галактик. Было принято решение работать с сервисом Legacy Survey[7], так как у него есть сервис для обзора всего неба. Сервис также позволяет несложно создавать ссылки для скачивания фото фрагмента неба по координатам. Мы брали фото с самого нового каталога – DR9. Итого всю выборку мы создали по следующему алгоритму:

- *Нахождение координат источников по фильтрам*
- *Создание ссылок для скачивания изображения*
- *Скачивание изображений*
- *Классификация фотографий вручную*

3.3 Нахождение координат источников

Пользуясь базой галактик NASA/IPAC[8], мы оставили запрос на вывод координат галактик. После первого запроса с параметрами, находящими слишком тусклые и маленькие объекты ($z < 5$, $m < 20$), мы созда новый, более подходящий со следующими поменянными параметрами:

- *Вся площадь неба*
- *Красное смещение: $Z < 3$*
- *Звездная величина: $m < 13$*
- *Объекты: галактики*

Спустя некоторое время, мы получили файл с характеристиками для 3470 объектов, из которых в итоговую базу для обучения модели пошли лишь 1551 фото.

3.4 Создание ссылок для скачивания изображения

Для их создания мы написали небольшую программу, сохраняющую итоговые ссылки в текстовый файл. Получились ссылки вида[6]. В них можно легко указать прямое восхождение и склонение объектов на эпоху J2000. Ознакомившись с кодом похожего астрономического проекта, мы адаптировали его под скачивание нужных для классификации фото через драйвер для Google Chrome. С помощью библиотеки Selenium[9] для Python мы скачали 1751 фото, т.к. некоторые из запланированных не установились из-за возникновения ошибок.

3.5 Классификация фотографий вручную

Для более удобного разделения галактик по нужным классам, мы написали простую программу на Python с использованием библиотеки tkinter[10]. Нами было вручную проклассифицировано более 1.5 тыс фото, которые мы разбили на следующие группы: дисковые галактики (826 фото), эллиптические галактики (409 фото), неправильные галактики (121 фото), прочее (156 фото), куда попадали в основном фото с дефектами.

3.6 Обучение модели ручной выборкой

Наша команда зарегистрировалась на сайте[11] и получила доступ к мощному серверу. Для корректной работы модели на нём, мы установили tensorflow через *!pip install* внутри клетки Jupyter Notebook. После этого мы обучили нейросеть на нашей ручной выборке (Рис. 1 и Рис. 2) и протестировали ее на фотографиях, не шедших на обучение. По итогу модель с хорошей точностью определяет по фотографии морфологический тип галактики.

Рис. 1: График обучения нейронной сети

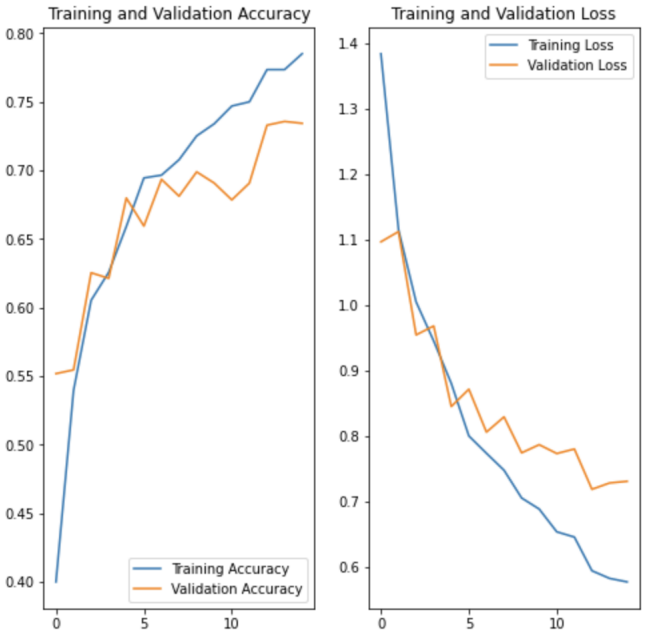
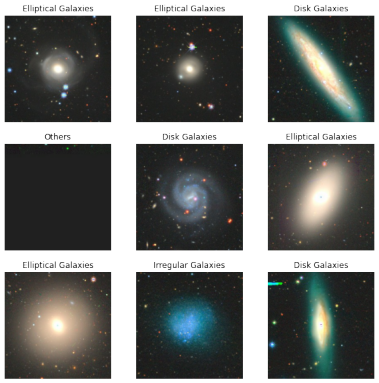


Рис. 2: Пример изображений для обучения нейронной сети



[6]https://www.legacysurvey.org/viewer/cutout.jpg?ra=0.81204dec=16.14542layer=ls-dr9pixscale=1.00bands=grz

4 Обсуждение и заключение

Мы изучили как отношение галактики к какому-либо морфологическому типу зависит от ее происхождения и влияет на эволюцию в целом. Мы также изучили основы машинного обучения. В ходе работы была написана модель на языке Python, основанная на библиотеке Tensorflow. Модель состояла из 12 преобразующих и обучаемых слоёв нейронов. При её создании возникли проблемы с установкой Tensorflow на компьютеры участников работы, имеющие разные операционные системы, и подбора к нему подходящей версии Python. Сперва мы хотели улучшить и переработать программный код предыдущих исследователей, но возникли проблемы с версией основной библиотеки, отсутствием комментариев в оригинале, что усложнило его восприятие. Изначально мы хотели обучить нейросеть на изображениях спектров разных объектов (звёзд, квазаров, галактик). При самом первом обучении классификатора у нашей выборки был серьезный недостаток - на каждом из спектров был указан тип объекта. Далее, обрезав всю выборку, наша нейросеть определяла тип объекта с точностью порядка 90%. Далее, было принято решение обучать модель на фотографиях галактик разных морфологических классов. Так для обучения нейросети нам потребовалось создать ручную .png выборку, основанную на обзоре неба Legacy Survey. Для этого мы запросили координаты близких и довольно ярких объектов через базу данных NASA/IPAC. Некоторые фотографии были с чрезмерными искажениями. В итоге после ручной классификации выборка была готова для обучения нейронной сети. Мы обучили модель примерно 1500 изображениями галактик, достигнув хорошей точности распознавания. При тесте модели другими изображениями, модель распознавала галактики с точностью, лежащей в интервале от 60% до 95%. Это означает, что необходимо улучшать и дорабатывать как выборку данных для обучения, так и саму нейросеть. В будущем планируется обучать нейросеть на основании файлов .fits, полученных с Legacy Survey, в совокупности с изображениями и спектрами они могут значительно повысить точность распознавания. В перспективе развития проекта, мы будем добавлять в кадры размытие Гауссианой, зашумление пикселей, чтобы нейросеть была более устойчивой к различным искажениям. Мы также хотим добавить классификацию по спектрам, в которые будем добавлять Пуассоновский шум. Ссылка на код нейронной сети и выборку приложена в Notes [12]

5 Вклад членов команды

Аронов Михаил	Написание кода нейросети, содействие в написании статьи
Магомедэминов Никита	Создание выборки, содействие в написании кода и статьи
Ерошкин Егор	Создание выборки, содействие в написании кода и статьи
Романов Антон	Содействие в написании статьи
Липская Мария	Содействие в написании статьи

6 Список литературы

- Wolf, 1908
  R.J.Buta, 2011; MO H., 2010
  VanderPlas, 2014
  Ball, Nicholas M. 2007; Ball, Nicholas M. 2008
  A. O. Clarke, 2020

Заметки

[1]https://pypi.org/project/tensorflow/
[2]https://medium.com/analytics-vidhya/quasar-detection-using-machine-learning-and-deep-learning-model-6e4b31683208
[3]https://medium.com/analytics-vidhya/quasar-detection-using-machine-learning-and-deep-learning-model-6e4b31683208
[4]https://www.tensorflow.org/tutorials/keras/classification
[5]https://www.tensorflow.org/tutorials/images/classification
[6]https://www.sdss.org/
[7]https://www.legacysurvey.org/
[8]https://ned.ipac.caltech.edu/byparams
[9]https://pypi.org/project/selenium/
[10]https://docs.python.org/3/library/tkinter.html
[11]https://datalab.noirlab.edu/
[12]https://github.com/Aromik/neuro_astro