

Московский Государственный Университет им. М. В. Ломоносова

Реферат

на тему: «Нейросетевые методы поиска и сегментации объектов в
данных современных космических обзоров»

Выполнил: студент гр. 420

Немешаева Алиса

Москва 2020

Содержание

Введение	1
1 Обзор существующих решений	4
2 Обзор данных	7
3 Построение решения	9
4 Результаты	12
Список использованных источников	13

Введение

Галактики не расположены случайным образом в пространстве. Они формируют собой особые структуры, такие как скопления и сверхскопления галактик. Скопления в свою очередь часто представляют собой цепи, или так называемые "нити на пересечении которых оказываются скопления галактик.

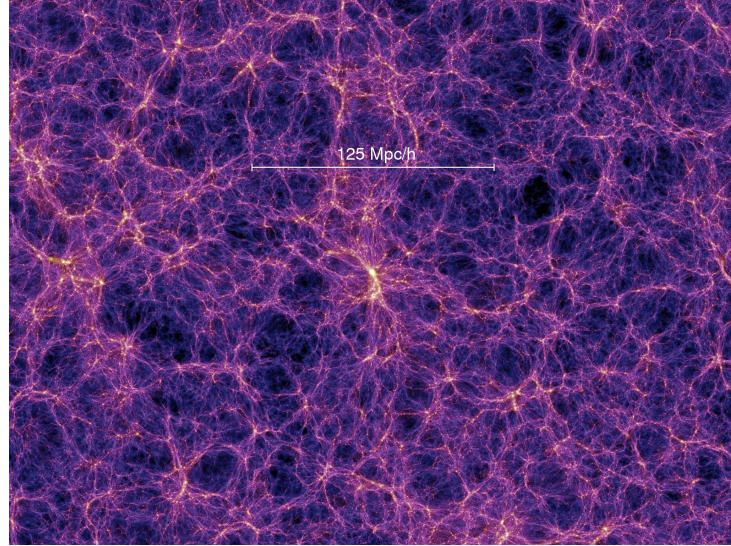


Рисунок 0.1 — Моделирование «Миллениум» — N-частичное моделирование, проведённое Консорциумом Девы с целью изучения формирования крупномасштабной структуры Вселенной в стандартной космологической модели.

Скопления галактик представляют большой интерес для исследования, так как их свойства сильно зависят от космологических параметров. Изучая их свойства, можно делать выводы о структуре обозримой части Вселенной. По большей части скопления состоят из тёмной материи, свойства которой до сих пор не известны науке. Поэтому через историю развития скоплений галактик можно изучать тёмную материю.

Сама по себе крупномасштабная структура Вселенной имеет объяснение. Через какое-то время после появления Вселенной возмущения волн плотности средних и больших масштабов при совпадении пиков образовали сверхскопления, в то время как сопадения фаз низкой плотности образовали войды - огромные пространства между нитями скоплений, в которых почти отсутствуют галактики и скопления. Таким образом, зная расположение и параметры большого количества скоплений, можно сделать выводы о том, как развивалась Вселенная на поздних этапах. [1]

Одним из первых каталогов скоплений стал каталог Abell [2]. Этот каталог содержит 4073 богатых скопления галактик с красными смещениями $z < 0.2$. Он был построен при использовании оптических данных обзора NGC-POSS.



Рисунок 0.2 — Скопление Волос Вероники (Abell 1656) в обзоре SDSS - одно из самых известных скоплений каталога Abell

Позднее были созданы каталоги с использованием других диапазонов. Далее для исследования будут использоваться следующие каталоги:

- PSZ2 - каталог, основанный на данных микроволнового обзора Planck. Был создан при использовании алгоритмов Matched Multi-Filter и PowellSnakes.

- MCXC - компиляция разных каталогов скоплений, все из которых основаны на данных рахных рентгеновских обзоров.

- RedMaPPer - каталог, основанный на данных оптического обзора SDSS. Был создан при помощи одноимённого алгоритма.

- ACT - каталог, основанный на данных микроволнового обзора ACT. Был создан при помощи алгоритма Matched Multi-Filter.

Создание каталогов данных в других диапазонах может помочь уточнить данные рентгеновских каталогов, созданных классическими алгоритмами.

Кроме того, сама задача применения нейросетевых методов к детекции скоплений актуальна, так как методы глубокого обучения дают следующие преимущества при анализе данных:

— Стандартные алгоритмы сегментации усредняют информацию по нескольким каналам, в то время как с помощью нейросети можно охватить данные полностью.

— Каждый из классических методов имеет свои достоинства и недостатки, и для каждого диапазона излучения существуют свои алгоритмы, в то время как нейросеть может стать универсальным средством для сегментации данных нескольких каналов одновременно.

1 Обзор существующих решений

1.1 U-net

U-net [3] является стандартной архитектурой для сегментации данных. Она подходит для проверки идеи использования методов глубокого обучения для сегментации скоплений. Её симметричная структура позволяет абстрагировать данные изображения, подаваемого на вход, в то время как skip-connection слои помогают увеличивать точность сегментации.

1.2 Детекция эффекта Сюняева-Зельдовича в данных Planck

Эта работа по общей структуре повторяет метод, описанный в [4], с некоторыми дополнениями, которые будут описаны далее.

В первую очередь, авторы статьи нормализуют данные Planck, подбирая для каждого канала параметры гауссианы, которая покрывает большую часть нужных данных. Таким образом сохранено изначальное распределение, но значения были приведены формату, который будет лучше восприниматься нейросетевой моделью.

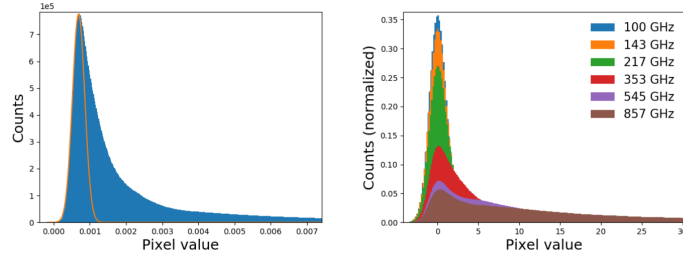


Рисунок 1.1 — Пример подбора гауссианы для канала 353 ГГц, и сравнение распределений для всех каналов после нормализации.

Далее нормализованные данные дополняются масками со скоплениями. Эти маски будут использоваться для обучения модели. Каждое скопление на маске отмечается как окружность диаметром $5'$.

Для обучения нейросетевой модели авторы статьи [4] использовали следующие каталоги:

— PSZ2

— MCXC

— RedMaPPer

Некоторые из этих каталогов содержат общие объекты, кроме того, обучать модель сразу на всех данных может оказаться нецелесообразно, поэтому списки были переработаны следующим образом:

— `planck_z` - каталог скоплений PSZ2 с известным красным смещением

— `planck_no_z` - каталог скоплений PSZ2 без красных смещений

— `mcxcwr` - каталог тех скоплений MCXC, что не присутствуют в PSZ2

— RM30 и RM50 - каталоги RedMaPPer

Основное обучение происходило на каталоге `planck_z`, так как эта модель давала лучшие результаты.

После этого вся область неба разделяется на три части - тренировочная, валидационная и тестовая. Для такого разбиения используется алгоритм представления данных на сфере HEALPix с параметром $n_{side} = 2$. В качестве валидационных выбраны пиксели 9, 38, 41, в качестве тестовых - 6, все остальные использовались для создания данных для тренировочной выборки (в статье [4] эти пиксели нумеруются с 1, здесь с 0). Для обучения выбирались случайные области на небе, называемые патчами. Размер патча зафиксирован: 64×64 пикселя в разбиении HEALPix $n_{side} = 2^{11}$. Патчи выбирались так, чтобы на них присутствовали скопления из выбранного для обучения каталога.

Далее происходит обучение модели. Модель используется такая же, как и в статье: U-net с 5-ю блоками в кодировщике, где каждый блок кодировщика содержит две свёртки 3×3 с функцией активации ReLU и слоём max-pooling (с увеличением количества фильтров в 2 раза). Блок кодировщика состоит из 2×2 upsampling свёртки (с уменьшением количества фильтров в 2 раза), конкатенации выхода соответствующего блока из кодировщика, двух свёрток 3×3 с ReLU. Самый последний слой кодировщика дополняется свёрткой 1×1 , кроме того, после каждой свёртки добавлен слой Dropout с параметром 0.2, чтобы увели-

чить эффективность обучения. Для изменения весов нейросети использовался оптимизатор Adam с параметром $lr = 10^{-4}$. В качестве loss-функции выбрана бинарная кросс-энтропия. Размер батча 20, количество эпох варьировалось от 20 до 40.

Для проверки результатов вся область неба сканировалась нейросетью, и на масках сегментации выбирались барицентры отдельных фигур, их координаты переводились в небесные, и после этого происходило сравнение с каталогами.

2 Обзор данных

Для создания тренировочной выборки использовались два каталога: часть каталога PSZ2 и каталог АСТ.

2.1 PSZ2

[1] Это каталог всего неба источников Сюняева-Зельдовича (SZ), обнаруженных по полным 29-месячным данным миссии Planck. Каталог (PSZ2) - это самая большая выборка скоплений галактик, отобранная по SZ, и самый глубокий систематический обзор скоплений галактик по всему небу. Он содержит 1653 обнаружения, из которых 1203 являются подтвержденными скоплениями с идентифицированными аналогами во внешних наборах данных. В справочной статье авторы описывают многоволновой поиск аналогов во вспомогательных данных, который использует наборы радио-, микроволновых, инфракрасных, оптических и рентгеновских данных и делают упор на надежность сопоставления двойников. Они обсуждают физические свойства нового каталога и идентифицируют совокупность тусклых рентгеновских скоплений с малым красным смещением, выявленных с помощью SZ-отбора. Эти объекты появляются в оптических обзорах и обзорах SZ с одинаковыми характеристиками для их массы, но они почти отсутствуют в отобранных рентгеновских выборках ROSAT.

Для обнаружения кластеров SZ использовались три метода: две независимые реализации согласованного мультифильтра (MMF1 и MMF3) и PowellSnakes (PwS). Главный каталог построен как объединение каталогов трех методов. Полнота и надежность каталогов были оценены посредством внутренней и внешней проверки, как описано в разделе 4 справочного документа.

2.2 АСТ

[5] Это каталог из 4195 оптически подтвержденных скоплений галактик Сюняева-Зельдовича (SZ), обнаруженных на 13168 deg^2 неба (примерно 32% всего неба), обследованных Космологическим телескопом Атакама (АСТ). Кандидаты в кластеры были отобраны путем применения многочастотного согласованного фильтра к картам 98 и 150 ГГц, построенным на основе всех наблюдений АСТ, полученных в 2008–2018 гг., и впоследствии подтвержденных с помощью глубоких оптических обзоров с большой площадью. Обнаруженные кластеры охватывают диапазон красного смещения $0,04 < z < 1,91$ со средним значением $z = 0,52$. Каталог содержит 221 кластер с $z > 1$, а всего 872 системы являются новыми открытиями.

Выборка скоплений более чем в 22 раза больше, чем предыдущий каталог скоплений АСТ, и на сегодняшний день является самой большой однородной выборкой скоплений, выбранных SZ. Зона обзора имеет большое перекрытие с глубокими оптическими исследованиями со слабым линзированием, которые используются для калибровки отношения масштабирования массы SZ-сигнала, такими как исследование темной энергии (Dark Energy Survey) ($4552deg^2$), стратегическая программа Hyper Suprime-Cam Subaru ($468deg^2$) и Kilo Degree Survey ($823deg^2$).

3 Построение решения

По большей части эта работа повторяет метод из [4], однако кроме упомянутых каталогов PSZ2 и MCXC для сопоставления результатов использовался каталог АСТ. Кроме того, детекция скоплений производилась следующим образом:

а) Для сканирования выбирается один пиксель из разбиения HEALPix с $n_{side} = 2$

б) Чтобы просканировать всю область, нужно пройти по ней окном 64×64 . Для этого нужно выбрать шаг сканирования - можно задать его как 64, тогда сканирование получится без пересечений, однако при меньшем шаге результаты оказываются лучше

в) Каждое изображение отправляется на вход в нейросеть. Полученные маски склеиваются и усредняются, чтобы снова получить изображение размером с пиксель $n_{side} = 2$

г) На полученной маске отделяются области со значениями, превышающими заданный порог

д) Для каждой области определяется её барицентр, который преобразуется в небесные координаты

У полученных после детекции каталогов определяются следующие параметры для каждого из детектированных объектов:

- area — площадь сегментированной области скопления
- min_rad, max_rad, mean_rad — минимальный, максимальный, средний радиусы области
- min_pred, max_pred — минимальное, максимальное значение маски в области
- status — факт сопоставления скопления с объектом из каталога
- catalog — сопоставленный каталог

На данный момент для детекции был выбран шаг сканирования 8, так как это наименьший шаг сканирования, при использовании которого тратится наименьшее количество времени, кроме того, из всех возможных вариантов этого параметра это значение дает наилучшие результаты recall и f₁ на тестовой области:

Однако предполагается, что при значения шага меньше 8 результаты могут быть еще лучшие, но этот вопрос еще предстоит исследовать.

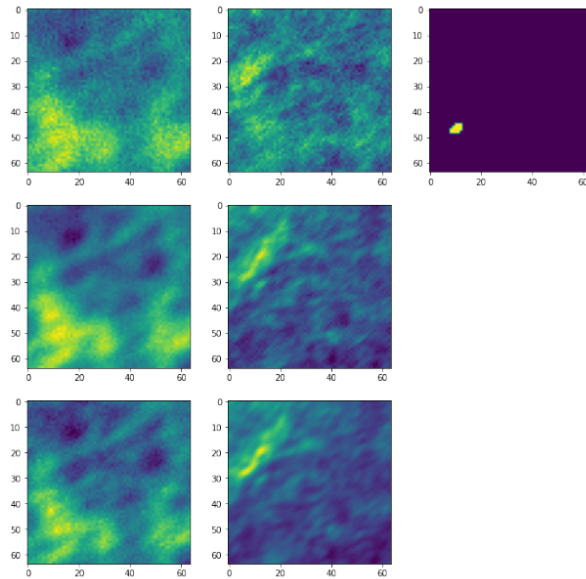


Рисунок 3.1 — Пример патча с шестью каналами данных Planck и маской с отмеченным центром скопления

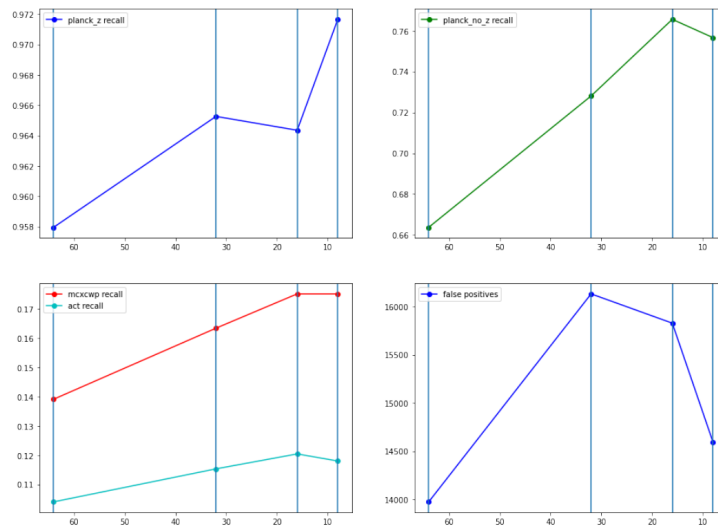


Рисунок 3.2 — Сравнение различных вариантов параметра шага сканирования

4 Результаты

Текущие результаты на данный момент:

- Созданы алгоритмы предобработки данных Planck.
- Обучена модель для сегментации данных Planck.
- Созданы алгоритмы детекции масок сегментации, производимых моделью.
- Создан каталог с оптимальными (на данный момент) параметрами. 1704 объекта из этого каталога были найдены в каталоге eRosita.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. *collaboration, Planck*. Planck 2015 results. XXVII. The second Planck catalogue of Sunyaev-Zeldovich sources. / Planck collaboration. — Astron. Astrophys., 2015.
2. *Abell G. O., Corwin H. G. Jr.* A Catalog of Rich Clusters of Galaxies / Corwin H. G. Jr. Abell, G. O. — R. P. Astrophys. J. Suppl., 1989.
3. *Olaf Ronneberger Philipp Fischer, Thomas Brox*. U-Net: Convolutional Networks for Biomedical Image Segmentation / Thomas Brox Olaf Ronneberger, Philipp Fischer. — 2015.
4. *Bonjean, V.* Deep learning for Sunyaev-Zel'dovich detection in Planck / V. Bonjean. — Astronomy&Astrophysics, 2020.
5. *collaboration, ACT*. The Atacama Cosmology Telescope: A Catalog of > 4000 Sunyaev-Zel'dovich Galaxy Clusters / ACT collaboration. — 2020.