МГУ им. М. В. Ломоносова, факультет ВМК

Курсовая работа

на тему: «Нейросетевые методы поиска и сегментации объектов в данных современных космических обзоров (eROSITA, ART-XC)»

Студент гр. 320
Немешаева Алиса
Научные руководители:
Герасимов С. В., к.ф-м.н Мещеряков А. В

Москва 2020

Содержание

1 Аннотация

Данная работа рассматривает возможность применения нейросетевых методов к решению проблемы сегментации и детекции объектов по многоволновым данным космических телескопов (в данном случае оптического, инфракрасного и рентгеновского диапазонов). В качестве основы для нейросетевой архитектуры использовалась модель U-net.

2 Введение

В 2019 году произошел запуск космической обсерватории СРГ (Спектр-Рентген-Гамма) с телескопами eROSITA и ART-XC на борту. Основной задачей этих телескопов является создание обзора всего неба в рентгеновском диапазоне. Данные, полученные от этих телескопов будут использоваться для обнаружения астрономических объектов трёх категорий:

а) Скопления галактик.
б) Сверхмассивные чёрные дыры.
в) Рентгеновские звёзды в галактике Млечный путь.

Полные обзоры неба, полученные телескопом eROSITA, появятся к июню 2020 года, поэтому на данный момент есть возможность подготовить модели для сегментации данных на примере других диапазонов.

В первую очередь будут использоваться данные оптического диапазона. Видимое излучение — тот диапазон частот, что доступен глазу человека. На текущий момент существует большое количество оптических телескопов, и, как следствие, большое количество данных, извлеченных с их помощью. В данной работе будут использоваться данные телескопа Pan-STARRS1, который является частью системы телескопов Pan-STARRS (Panoramic Survey Telescope and Rapid Response System). Этот телескоп построен на вершине гавайского вулкана Халеакала. На 2007 год он обладал самой большой светочувствительной матрицей в мире. Кроме того, его данные находятся в общем доступе [2].

В последние годы методы глубокого обучения стали играть важную роль в анализе данных. Нейросетевые модели показывают высокие результаты в области компьютерного зрения и в частности в задачах сегментации и детекции. Всё более часто они применяются и для решения задач астрофизики. Характеристики телескопа eROSITA позволят получить рентгеновские данные очень высокого качества (то есть с низким количеством шума), и методы глубокого обучения дают много преимуществ при анализе данных:

а) Алгоритмы, не использующие нейросетевые методы, обычно требуют тонкой настройки и подбора параметров, подходящих под характеристики данного телескопа и данной области неба. Нейросеть же не требует предварительной настройки, и при достаточном количестве тренировочных данных сможет одинаково хорошо сегментировать данные, для которых другим алгоритмам потребовались разные параметры.

б) Аналогичнные методы можно использовать для сегментации одновременно разнородных данных. То есть для улучшения качества сегментации можно исследо-

вать параллельно разные диапазоны частот и находить взаимосвязь между разными спектрами.

На данный момент архитектура U-net [4] является одной из лучших нейросетевых моделей для сегментации. Её симметричная структура позволяет абстрагировать данные изображения, подаваемого на вход, в то время как skip-connection слои помогают увеличивать точность сегментации.

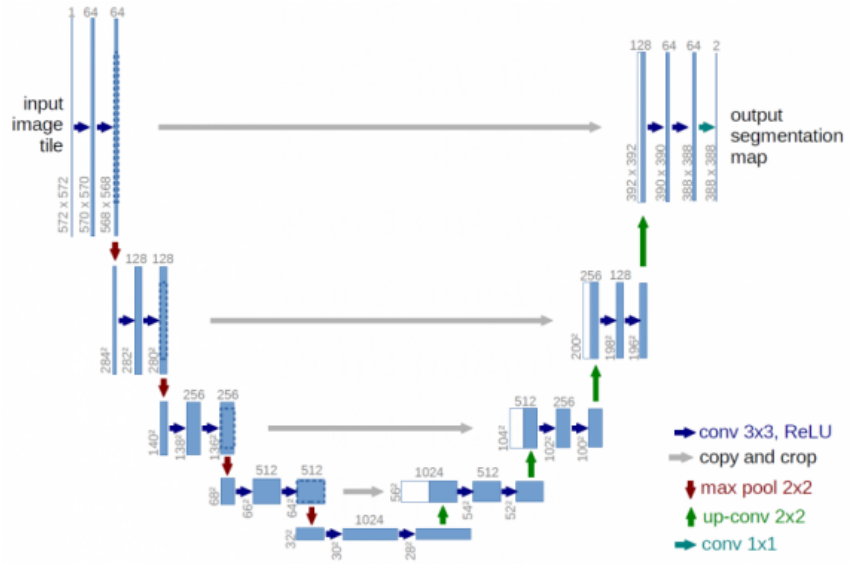


Рисунок 2.1 — Структура модели U-net [4]

3 Постановка задачи

Эта работа во многом является повторением исследования из статьи о детекции эффекта Сюняева-Зельдовича [1], с той разницей, что здесь будут использоваться оптические данные, в то время как в упомянутой статье использовались данные микроволнового диапазона.

Основной задачей этой работы является создание нейросетевой модели, способной сегментировать данные космических обзоров так, чтобы на них можно было детектировать объекты определённого типа (в данном случае скопления). В лучшем случае модель должна будет демонстрировать результаты, превосходящие по качеству методы, использующиеся для решения аналогичных проблем.

4 Обзор существующих решений рассматриваемой задачи или её модификаций

В первую очередь рассмотрим уже упомянутую работу о детекции эффекта Сюняева-Зельдовича [1]. Её автор использует для сегментации данных архитектуру U-net (эта архитектура будет использоваться и в этой работе).

Основной целью описываемой работы являлось создание алгоритма для детекции источников через эффект Сюняева-Зельдовича по данным телескопа «Планк». Соответственно, кроме самих обзоров неба, полученных «Планком», использовались еще три каталога скоплений для создания целевых данных:

а) PSZ2. Этот каталог был получен по данным «Планка» при помощи алгоритмов согласованного мультифильтра и PowellSnakes.

б) MCXC (Meta-Catalogue of X-ray detected Clusters). Это объединенный каталог из всех других каталогов скоплений, полученных из данных телескопа ROSAT.

в) RedMaPPer (Red-sequence Matched-filter Probabilistic Percolation). Каталог скоплений, полученный с помощью одноимённого алгоритма из данных оптического диапазона.

В описываемой работе для создания тренировочных выборок использовалось разбиение неба проекцией HEALPix (Hierarchical Equal Area isoLatitude Pixelisation).

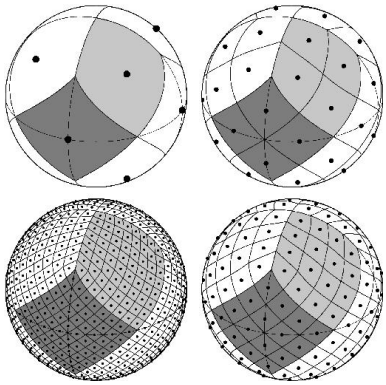


Рисунок 4.1 — Примеры разбиения сферы HEALPix [3]

Разбиение с параметром $n_{side} = 2$ позволяет получить 48 больших областей неба. Некоторые из них были использованы для тестирования полученной модели и для валидации, все остальные были использованы для обучения модели.

Случайным образом в соответствуюхих областях разбиения HEALPix выбирались центры патчей и их ориентации для создания тренировочных, валидационных и тестовых выборок. Каждый патч представлял из себя изображение

размера 64 x 64 с шестью каналами различных данных. Размер каждого пикселя на таких патчах составлял 1.7 arcmin.

После этого 100000 патчей были использованы для обучения нейросети. Обучение длилось более 30 эпох.

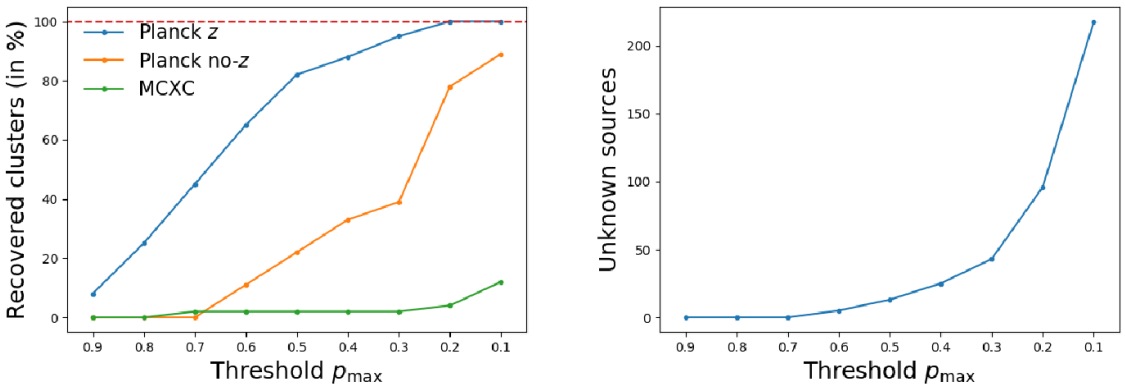


Рисунок 4.2 — Результаты исследования работы [1]

Для детекции скоплений на полученных из нейросети данный, зонами скоплений обозначались зоны, занимающие пиксели со значением больше p_{max}, а затем на них находились барицентры, которые впоследствии считались предсказанными центрами скоплений. Таким образом большая часть скоплений из каталога PSZ2 была успешно распознана нейросетью.

5 Исследование и построение решения задачи

Так же, как и в статье [1], в качестве списка скоплений будем использовать каталог PSZ2. Аналогичным образом будем генерировать патчи для создания тренировочных, валидационных и тестовых выборок (в том числе для валидации и тестирования будут выбраны те же пиксели разбиения $n_{side} = 2$. Патчи выбирались так, чтобы в их окрестности находился хотя бы одно из скоплений нужного каталога. После этого в базе данных PS1 (Pan-STARRS1) запрашивался список объектов, подходящих под заданные условия.

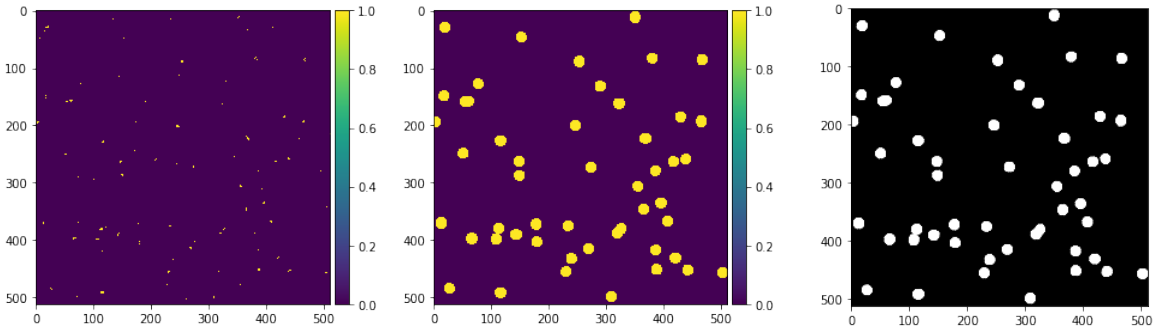
После того, как будут получены данные для обучения, их нужно из таблиц преобразовать в двухмерные матрицы изображений, чтобы создать выборки с количеством каналов, совпадающих с количеством исследуемых параметров у объектов. Перед этим нужно удалить из таблицы повторяющиеся объекты.

Для примера, таблица с данными для 100 патчей содержит около 10 миллионов объектов. Обработка сотни объектов будет длится несколько минут, поэтому необходимо отдельно распараллелить процесс.

6 Описание практической части

При подготовке к созданию итоговой модели в первую очередь создавались симуляции (искуственные данные, похожие по статистическим распределениям на настоящие, но по своей структуре более простые).

После этого на созданных симуляциями данных тренировались первые образцы нейросетевых моделей.



a) b) c)

Рисунок 6.1 — a) Данные объектов из симуляций; b) Данные источников из симуляций; c) Результат работы нейросети;

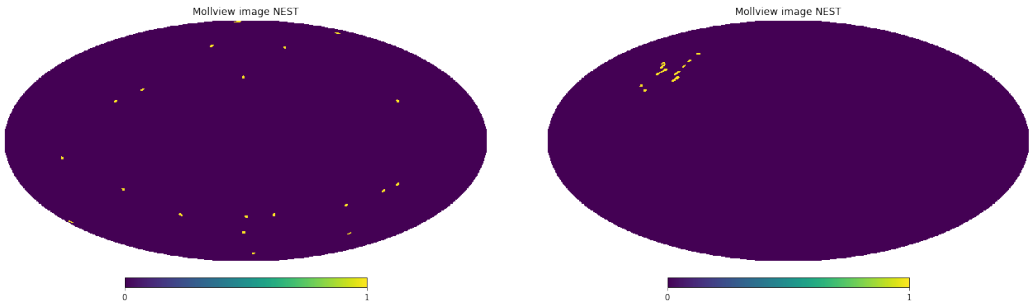
Выше приведен пример работы симуляции а также пример сегментированного изображения. Итоговая точность сегментации составляет 0.9978 для симулированных данных.

Далее, начинается обработка настоящих данных. Нужно загрузить и обработать каталог скоплений Planck, который будет разделен на два подкаталога:

а) planck_z (скопления с измеренным красным сдвигом)
б) planck_no_z (скопления без информации о красном сдвиге)

После того, как были получены данные по скоплениям, можно начать загрузку и обработку данных из обзоров PS1. В каждом пикселе из разбиения с $n_{side} = 2$ генерируется определенное количество патчей. Центры патчей выбираются случайным образом как пиксели из разбиения $n_{side} = 11$. Размер каждого патча задан как 64 x 64, и так как пиксели HEALPix могут иметь протяжённую структуру, итоговый радиус патча вычислялся как расстояние от центра патча до дальнего угла для патча размером 66 x 66. В итоге радиус оказался равен $\approx 1.45°$.

Далее данные нужно преобразовать в формат двумерных матриц для загрузки в нейросеть.



a) b)

Рисунок 6.2 — Сгенерированные центры патчей: a) Для всего неба; b) Для пикселя №6 из разбиения $n_{side} = 2$;

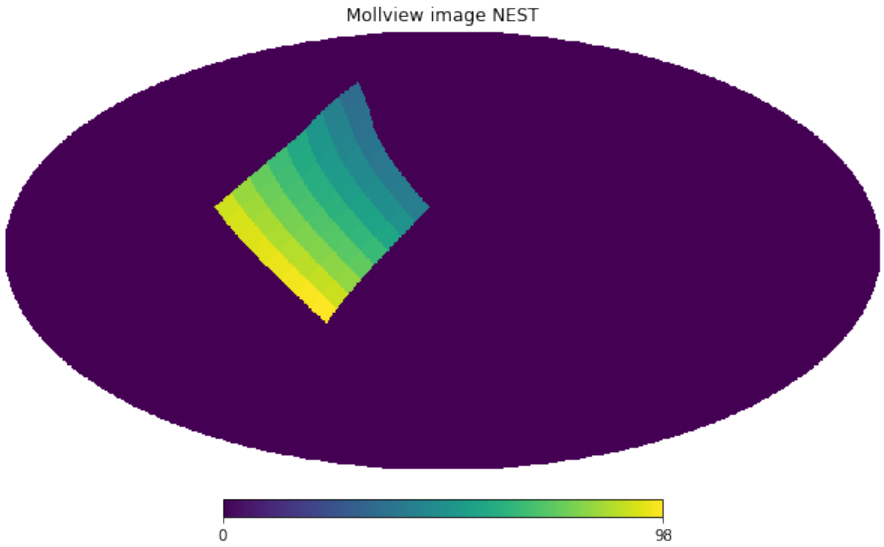


Рисунок 6.3 — Пример расположения двумерной матрицы на проекции неба для большего разбиения HEALPix

Заключение

Была проведена работа по созданию простейших симуляций для предварительного обучения нейросетевых моделей для сегментации космических обзоров. Реализована архитектура U-net. Также созданы базовые функции для обработки данных и их загрузки в нейросеть.

Список использованных источников

1. *V. Bonjean* Deep learning for Sunyaev–Zel'dovich detection in Planck Astronomy & Astrophysics

2. PanSTARRS1 (PS1) Catalog Archive Server Jobs System (CasJobs) service

3. HEALPix - Features

4. *O. Ronneberger, P. Fischer, T. Brox* U-Net: Convolutional Networks for Biomedical Image Segmentation