Нейросетевые методы поиска и сегментации объектов в данных современных космических обзоров (eROSITA, ART-XC)

Научные руководители:

Герасимов С.В., к.ф.-м.н. Мещеряков А.В.

Студент:

Немешаева Алиса, 3 курс бакалавриата ВМК МГУ

Введение

В 2019 году произошёл запуск космической обсерватории СРГ 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 скоплений галактик.
Скопления - это гравитационно связанные системы, которые являются самыми большими структурами во Вселенной. Скопления галактик играют важную роль в задачах определения параметров Вселенной. Скопления галактик излучают энергию в разных диапазонах, и их можно наблюдать не только в рентгеновских данных.

Введение

- 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. В данной работе будут использоваться данные телескопа Pan-STARRS1. На 2007 год он обладал самой большой светочувствительной матрицей в мире. Кроме того, его данные находятся в общем доступе.

Телескоп Pan-STARRS 1

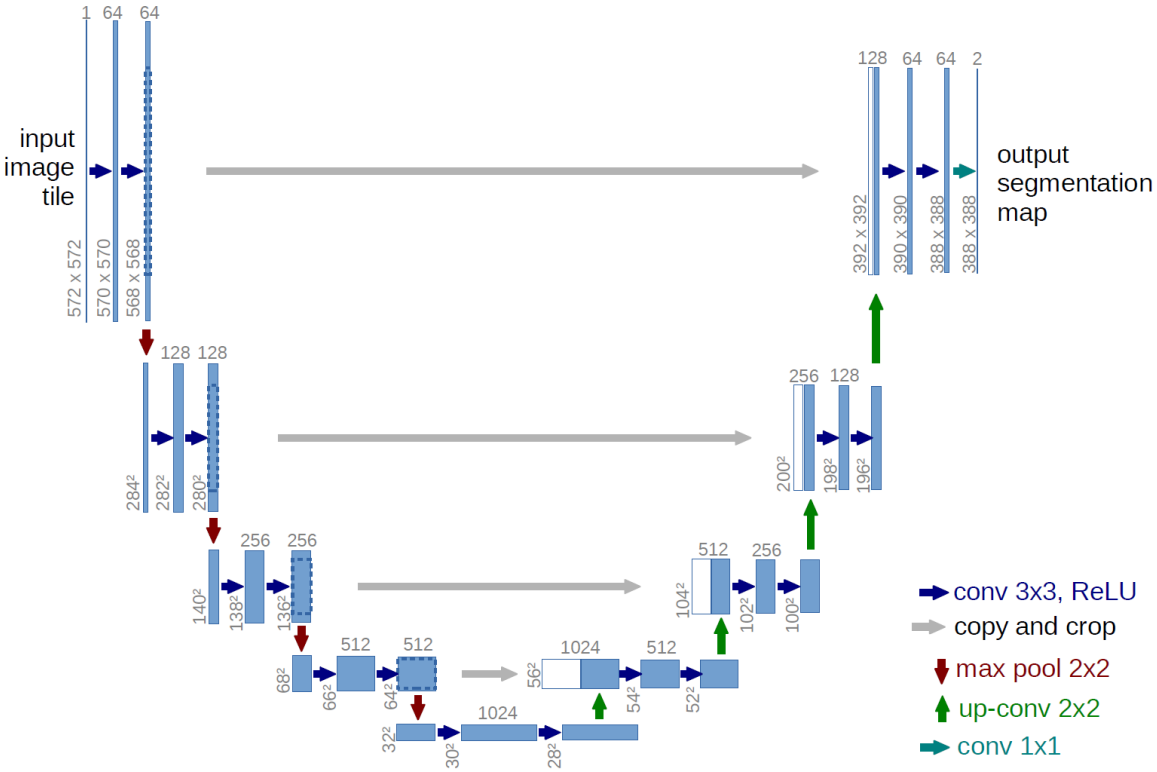
Актуальность

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. Методы глубокого обучения дают много преимуществ при анализе данных:

- Нейросети можно охватить данные полностью без усреднения и исследовать вопрос с новой стороны.
- Нейросеть будет получать «сырые» данные, что экономит время и исключает необходимость контролировать процесс предобработки данных.
- Аналогичнные методы можно использовать для сегментации одновременно разнородных данных.

Актуальность

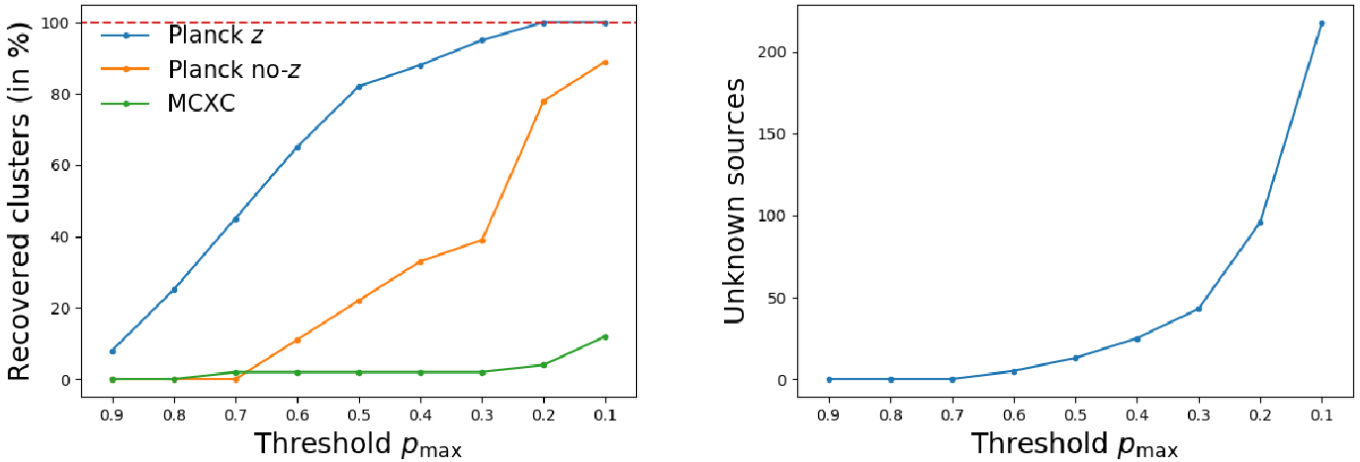
- U-net является стандартной архитектурой для сегментации данных. Она идеально подходит для проверки идеи использования методов глубокого обучения для сегментации скоплений. Её симметричная структура позволяет абстрагировать данные изображения, подаваемого на вход, в то время как skip-connection слои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



Обзор существующих решений

В первую очередь рассмотрим работу о детекции эффекта Сюняева-Зельдовича. Её автор тоже использует для сегментации данных архитектуру U-net. Основной целью описываемой работы являлось создание алгоритма для детекции источников через эффект Сюняева-Зельдовича по данным телескопа «Планк».
Кроме самих обзоров неба, полученных «Планком», использовались еще три каталога скоплений для создания целевых данных:

- PSZ2
- MCXC
- RedMaPPer



Результат другой работы по данным «Планка»

Постановка задачи

Общей задачей данной работы является решение проблемы сегментации и детекции скоплений галактик.
Для достижения цели задание было разделено на несколько шагов:

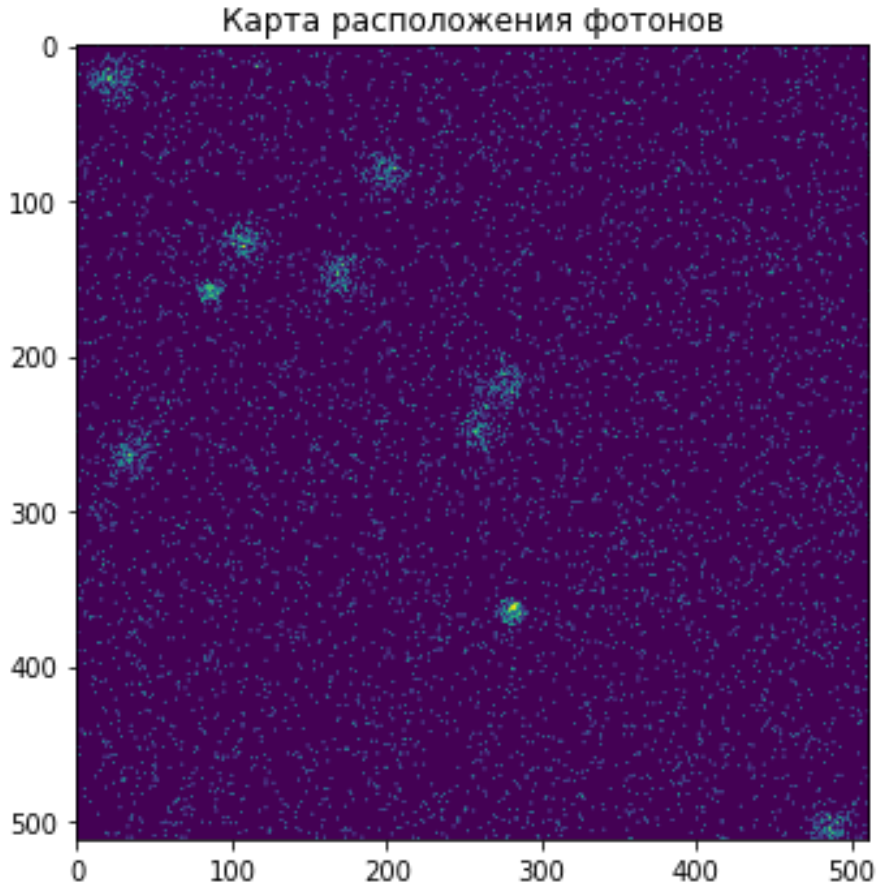
- Создание простейших симуляций рентгеновских данных и образца модели U-net.
- Проверка работы U-net на данных симуляций.
- Загрузка и обработка данных о скоплениях.
- Генерация «патчей» и загрузка обзоров неба PanSTARRS1 из области патчей.
- Преобразование данных PanSTARRS1 в двумерные матрицы для загрузки в нейросеть.
- Обучение модели, подбор параметров модели (количество слоёв, методы аугментации, размер батча, количество эпох обучения)
- Тестирование модели на заранее выбранных данных.

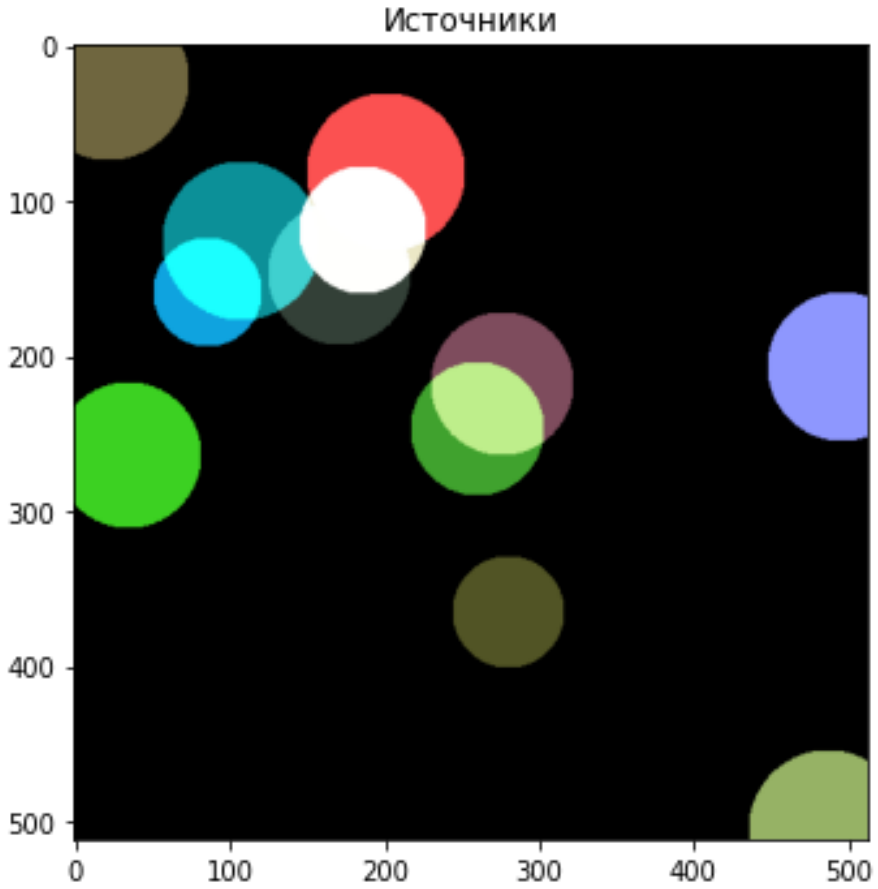
На данный момент не существует какого-то универсального метода для сегментации и детекции скоплений на оптических данных, и есть возможность применить методы глубокого обучения в данной области.

Описание практической части. Симуляция и первый образец нейросети

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. Общая архитектура схожа с моделью из статьи из обзора, но в данном случае вместо слоёв Dropout использовались слои батч-нормализации и вместо 5 блоков было добавлено 3 блока с 32 фильтрами в первом блоке.

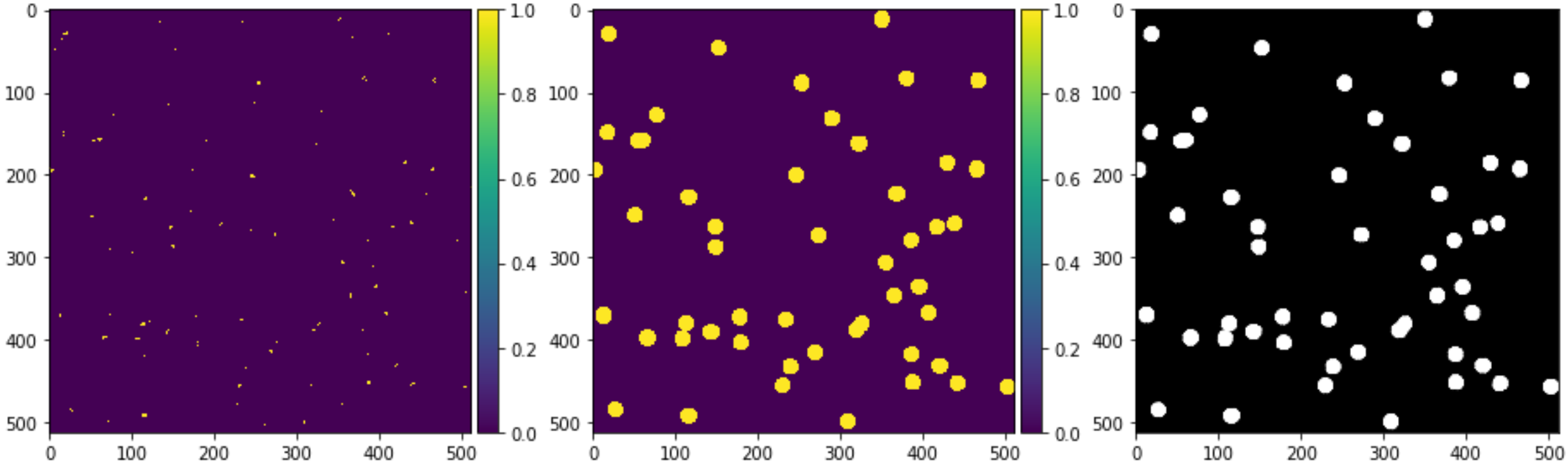
Описание практической части. Симуляция и первый образец нейросети





Пример генерации данных и масок для каждого скопления

Описание практической части. Симуляция и первый образец нейросети



Пример работы симуляции и пример сегментированного изображения. Итоговая точность сегментации составляет 0.9978 для симулированных данных.

Описание практической части. Обработка каталогов скоплений и данных PS1

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

- planck_z
- planck_no_z

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 ~1.45 °.

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

Текущие результаты

На данный момент было решено несколько подзадач, поставленных для решения общей проблемы:

- Созданы простейшие программы для генерирования симуляций в рентгеновском диапазоне.
- Проверена работоспособность модели U-net на данных симуляций.
- Обработаны каталоги источников.
- Создан код для генерирования «патчей» для обучения и тестирования нейросети.
- Начата работа по обработке данных PS1.