Краткий текст для защиты курсовой работы
Научные руководители: Герасимов С.В., Мещеряков А.В.
Студент: Немешаева Алиса
Курс: 3

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

В 2019 году произошел 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 — самые большие динамически связанные структуры во Вселенной. Скопления галактик играют важную роль в задачах определения космологических параметров Вселенной.

Одной из главных задач космологии является построение карты скоплений галактик для того, чтобы можно было изучать их в совокупности.

Зная информацию о большом количестве скоплений, можно изучать тёмную материю, из которой они по большей части состоят
.

Скопления галактик можно обнаружить исследуя данные различных диапазонов.

Рентгеновский диапазон. Скопления выдают себя в рентгеновском диапазоне из-за присутствия в их составе горячего газа, излучающего энергию в рентгеновском диапазоне.

Микроволновый диапазон. В этой области данных скопления галактик проявляются из-за того, что существует реликтовое излучение, которое заполняет собой всю Вселенную. Из-за своего состава скопления искажают это излучение, и по этим искажениям их можно наблюдать.

Оптический диапазон. Поскольку скопления галактик состоят из галактик, которые сами по себе излучают энергию в видимом диапазоне, логично для поиска скоплений использовать и оптические телескопы.

Данных в рентгеновском диапазоне сейчас не доступны, исследования для микроволнового диапазона уже существуют (будут упомянуты в обзоре существующих решений). Оптические данные находятся в общем доступе и есть возможность исследовать именно их, и впоследствии перейти к анализу совокупности данных из разных диапазонов.

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. Обзор этого телескопа покрывает самую большую область неба и покрывает те объекты, данные которых позднее будут получены телескопом eROSITA.

С появлением новых телескопов параллельно развиваются и методы обработки их данных. Чаще всего алгоритмы создаются под конкретный телескоп и под конкретный диапазон, и для более современных телескопов с более чувствительными датчиками они просто не подойдут. Характеристики телескопа eROSITA позволят получить рентгеновские данные очень высокого качества, на них можно будет детектировать новые скопления максимально точно. Поэтому есть возможность применить новые алгоритмы на таких данных.

Методы глубокого обучения дают много преимуществ при анализе космических данных:

1. Стандартные алгоритмы сегментации усредняют информацию по нескольким каналам, в то время как с помощью нейросети можно охватить данные полностью и исследовать вопрос с новой стороны.
2. Аналогичнные методы можно использовать для сегментации одновременно разнородных данных. То есть для улучшения качества сегментации можно исследовать параллельно разные диапазоны частот и находить взаимосвязь между разными спектрами.

3. 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.

Для обучения нейросетевых моделей требуется иметь большое количество данных, детектированных другими способами. Для таких целей будут использоваться существующие каталоги скоплений галактик:

1. PSZ2. Этот каталог был получен по данным микроволнового телесокпа «Планк».
2. MCXC. Это объединенный каталог из всех других каталогов скоплений, полученных из данных телескопа ROSAT.
3. RedMaPPer. Каталог скоплений, полученный с помощью одноимённого алгоритма из данных оптического диапазона.

В более формальном и подробном виде задачу можно описать так: при имеющихся обработанных данных, требуется получить матрицы сегментации, где для каждого пикселя матрицы мы будем иметь информацию о вероятности, с которой в данном пикселе находится скопление. Впоследствии координаты пикселей изображений можно преобразовать в небесные координаты. На полученных масках можно будет детектировать скопления. Главной целью является создание итогового каталога скоплений, найденных в разных диапазонах.

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

Исследование проблемы сегментации данных в многоволновых диапазонах и разработка нейросетевой модели, позволяющей выполнять сегментацию и детекцию скоплений галактик на оптических данных телескопа PS1.

В первую очередь рассмотрим работу о детекции эффекта Сюняева-Зельдовича в микроволновых данных. 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». Соответственно, кроме самих обзоров неба, полученных «Планком», использовались еще три ранее упомянутых каталога скоплений для создания целевых данных.
В описываемой работе для создания тренировочных выборок использовалось разбиение неба проекцией HEALPix.
Лучше всего нейросеть сегментирует в данных телескопа «Планк» скопления из каталога того же самого «Планка».

На данный момент существует несколько пакетов программ, с помощью которых осуществляется сегментация данных, полученных с помощью рентгеновских телескопов. Одним из них является CIAO, разработанный специально для Космической рентгеновской обсерватории «Чандра».

- *celldetect* использует свёртку с изменяющимся размером ядра для сегментации.
- *vtpdetect* использует обнаружение источников с помощью тесселяции и перколяции Вороного
- *wavdetect* использует свёртку вэйвлет функции с астрономических изображением.

Каждый из упомянутых методов является ситуативным и может проявлять себя по-разному для разных данных. Все эти методы используют усреднённые рентгеновские данные и не учитывают характеристики отдельных телескопов.

U-net является стандартной архитектурой для сегментации данных. Её симметричная структура позволяет абстрагировать данные изображения, подаваемого на вход, в то время как skip-connection слои помогают увеличивать точность сегментации.

Так же, как и в статье о микроволновых данных, в качестве списка скоплений будем использовать каталог PSZ2. Аналогичным образом будем генерировать патчи для создания тренировочных, валидационных и

тестовых выборок. После этого в базе данных PS1 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.

Существует несколько способов преобразовать данные, записанные как местонахождение объектов в космических координатах, в значения на плоскости, которые можно уложить на двумерную матрицу. При любой проекции сферы на плоскость мы будем сталкиваться с искажениями в разной степени и для разных параметров данных.

HEALPix.

Этот формат является очень удобной формой хранения данных. Данные телескопа «Планк», исследование которых упоминается в обзоре существующих решений, хранятся именно в этом формате, поэтому над ними легко работать и для них не требуется такое количество стадий предобработки, как например для данных PS1.

Как ранее было упомянуто, HEALPix является иерархической структурой для хранения данных. Она позволяет сохранить на проекции площадь объекта, однако его форма может быть искажена.

Тангенциальная проекция.

Ещё один способ перенести данные со сферы на плоскость — тангенциальная проекция WCS. Такая проекция сохраняет форму объектов, но искажает их площадь.

Проекция HEALPix считается более удобной, так как она является абсолютной для всей области неба и для неё нет необходимости выбирать центр проекции.

При подготовке к созданию итоговой модели в первую очередь создавались симуляции.

Далее, начинается обработка настоящих данных. Нужно загрузить и обработать каталог скоплений Planck.

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. Для преобразований использовалась библиотека healpy.

Далее данные нужно преобразовать в формат двумерных матриц для загрузки в нейросеть, перед этим вычислив для каждого объекта номер пикселя healpix, к которому он относится. Объекты, относящиеся к одним и тем же пикселям, нужно соответствующим образом отождествить.

Для этого создавались матрицы соответсвий между координатами изображения и номерами пикселей из healpix-разбиения.

После этого по построенным матрицам нужно каждый объект перенести на изображение.

Точно так же перенести на двумерное изображение нужно и маски, определяющие наличие скоплений в данной области. Для каждого скопления вычисляются номера пикселей в радиусе $5'$, после чего они с помощью той же матрицы соответствий переносятся на изображение.

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

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

Обработка данных PS1 требует больше времени, чем предполагалось, поэтому работа по данной проблеме будет продолжаться далее.