

УДК 524.77–77

ПОИСК КАНДИДАТОВ В СКОПЛЕНИЯ ГАЛАКТИК НА КАРТАХ МИКРОВОЛНОВОГО ФОНОВОГО ИЗЛУЧЕНИЯ КОСМИЧЕСКОЙ МИССИИ PLANCK С ПОМОЩЬЮ СВЕРТОЧНОЙ НЕЙРОННОЙ СЕТИ ПО ПРИНЦИПУ ФИКСАЦИИ ЭФФЕКТА СЮНЯЕВА–ЗЕЛЬДОВИЧА

© 2021 **О. В. Верховданов**¹, А. П. Топчиева^{2*},
А. Д. Ороновская³, С. А. Базров³, Д. А. Шорин³

¹Специальная астрофизическая обсерватория РАН, Нижний Архыз, 369167 Россия

²Институт астрономии РАН, Москва, 119017 Россия

³Астрофизическая школа «Траектория», Москва, 109147 Россия

Поступила в редакцию 23 декабря 2019 года; после доработки 20 декабря 2020 года;
принята к публикации 20 декабря 2020 года

В работе предложен метод поиска радиоисточников с эффектом Сюняева–Зельдовича на мультичастотных картах излучения по данным миссии Planck с помощью сверточной нейронной сети. Каталог для распознавания радиоисточников составлен с помощью схемы пикселизации GLESP на частотах 100, 143, 217, 353 и 545 ГГц. Оценивается качество предложенного подхода и влияние вариации отношения S/N на результат обучения сети. Показано, что представленный нейросетевой подход позволяет детектировать объекты с эффектом Сюняева–Зельдовича. Предложенный метод может использоваться для поиска наиболее вероятных кандидатов в скопления галактик на больших красных смещениях.

Ключевые слова: *реликтовое излучение—галактики: скопления: общее*

1. ВВЕДЕНИЕ

Одним из результатов работы космического телескопа Planck является создание карты распределения интенсивности реликтового излучения для всего неба. В числе прочего эта карта включает в себя объекты с эффектом Сюняева–Зельдовича (далее СЗ-эффект), который можно использовать для идентификации скоплений галактик. СЗ-эффект является результатом рассеяния фотонов реликтового излучения на горячих электронах в плазме. Подходящие условия для его наблюдения присутствуют в скоплениях галактик (Zeldovich and Sunyaev 1969). Изучение таких скоплений на различных красных смещениях способствует пониманию устройства Вселенной с помощью проверки космологических моделей и ограничения космологических постоянных (Allen et al. 2011, Kravtsov and Borgani 2012, Nagai 2014). Кроме того, такие объекты исследуются также и при решении астрофизических задач (Вукон et al. 2000, Kormendy and Djorgovski 1989, Sarazin 1986). Среди лидирующих направлений в космологии оста-

ся изучение скоплений галактик в миллиметровом и субмиллиметровом диапазонах, наблюдаемых благодаря СЗ-эффекту (Zeldovich and Sunyaev 1969), в рентгеновском диапазоне, где излучает горячий газ, и в видимом свете. Таким образом, можно проследить эволюцию масс скоплений и особенности формирования крупномасштабных структур Вселенной в разные космологические эпохи (Basu et al. 2019).

Наблюдения эффекта сложны из-за его малой амплитуды, погрешности экспериментальных данных и искажения температуры реликтового излучения другими источниками. Поскольку СЗ-эффект — эффект рассеяния и его величина не зависит от красного смещения, скопления с большим красным смещением могут быть обнаружены так же легко, как и с малым, если у них достаточная масса ($M_{\text{clust}} \sim 10^{14} M_{\odot}$, где $M_{\text{clust}} = M_{500}$). Отношение углового размера к красному смещению также является фактором, облегчающим обнаружение скопления с большим красным смещением: оно мало меняется между красными смещениями 0.3 и 2, и это значит, что скопления с красными смещениями в указанном

*E-mail: ATopchieva@inasan.ru

диапазоне имеют практически одинаковые размеры на небе.

Ранее мы уже использовали подход, основанный на селекции кандидатов в скопления с помощью СЗ-эффекта в направлении на радиоисточники (см. Verkhodanov et al. (2015; 2018), Zarogozhets and Verkhodanov (2019)). В этой работе мы развиваем направления, связанные с автоматизированной методикой селекции СЗ-объектов на основе метода машинного обучения (см. описание первого опыта в Ogorovskaya (2018)).

В последнее время в астрономических исследованиях все чаще применяются методы глубокого машинного обучения. Назовем, к примеру, работы по изучению распределения пыли и ее свойств в галактиках (Dobbels et al. 2020), по поиску связей между звездами в центре галактик и морфологией самих галактик (см. работу Tacchella et al. (2019)), по определению параметров галактик, таких, как плотность, металличность, поверхностная плотность и степени ионизации по эмиссионным линиям спектров галактик в оптическом, ультрафиолетовом, инфракрасном и субмиллиметровом диапазонах (см. работу Ucci et al. (2017)). Также с помощью методов машинного обучения при изучении спектров двух миллионов галактик из Sloan Digital Sky Survey были выбраны 400 галактик с высокими показателями активности джетов (Baron and Poznanski 2017). Похожие алгоритмы применяются и при классификации изображений протяженных источников в радиодиапазоне, при изучении радиогалактик на основе морфологических свойств с использованием сверточных нейронных сетей (Aniyan and Thorat 2017), при спектральной классификации галактик (Tao et al. 2018). Наконец, отметим и создание системы машинного обучения при поиске СЗ-эффекта по данным HFI Planck (Bonjean 2020); в результате по высокочастотным данным были отобраны 18 тысяч кандидатов в СЗ-объекты.

В данной работе мы представляем алгоритм машинного обучения, который поможет расширить список скоплений галактик в миллиметровом диапазоне. Есть данные наблюдений микроволнового фона Planck (Ade et al. 2016), SPT (Vanderlinde et al. 2010) и других обсерваторий, например, АСТ (Hasselfield et al. 2013), которые представляют собой карты неба на различных длинах волн. СЗ-эффект обнаруживается на некоторых участках карты (в направлении на скопления) и имеет особую спектральную форму: на частоте 217 ГГц и ниже заметно уменьшение интенсивности сигнала над микроволновым фоном, а на более высоких частотах — повышение. Необходимо найти такие участки на карте. В отличие от ранее разработанных методик (Bonjean 2020, Herranz et al. 2002, Melin 2006), в данной работе полнота получаемой

выборки анализируется с помощью сравнения с простыми алгоритмами поиска объектов Matched multi-filter (MMF1 (Herranz et al. 2002), MMF3 (Melin 2006)) и PowellSnakes (PwS (Carvalho et al. 2012)). Также мы оцениваем качество созданной модели и влияние вариации отношения S/N на результат обучения сети, что может быть полезным при работе с данными более низкого качества. В разделе 2 описан формат данных, использованных для обучения нейронной сети, методы их получения и обработки. В разделе 3 представлен непосредственно алгоритм и примеры его применения на картах обсерватории Planck. В разделе 4 приведены результаты работы сети, сравнение моделей машинного обучения ResNet18 и RandomForest между собой, а также с методами MMF1, MMF3 и PwS на метрике Recall с примененным подходом ResNet18, и представлены качества обнаружения исследуемых объектов на модельных данных для случайных гауссовых карт с различным отношением S/N . В заключении сформулированы выводы о проведенной работе. В Приложении описаны метрики и алгоритм подхода для анализа объектов с СЗ-эффектом.

2. ПОДГОТОВКА ДАННЫХ

В работе используются данные, взятые из архивов наблюдений телескопа Planck¹. Эти данные были переведены в формат GLESP. Программный пакет GLESP (Gauss–Legendre Sky Pixelization) (Doroshkevich 2005) использует одноименные схемы пикселизации карт космического микроволнового фона, основываясь на применении нулей полиномов Гаусса–Лежандра для организации сетки разбиения неба, и позволяет получить строгое ортогональное разложение карт. При работе с данными по реликтовому излучению широко применяется пакет HEALPIX, однако мы использовали программный пакет GLESP, поскольку работа с ним привычна авторам. Для решаемой задачи оба пакета идентичны по своим возможностям. В пакете имеются процедуры для работы с отдельными изображениями на площадках заданного размера и позволяет быстро преобразовывать температуры в координатах в температурный спектр, то есть производить разложение по сферическим функциям с использованием специального метода интегрирования Гаусса. Для создания выборки объектов с СЗ-эффектом отбирались те карты, для которых заведомо известно, что эффект присутствует. Карты без СЗ-эффекта выбирались по случайным координатам в области, для которой известно, что этого эффекта нет. Подробнее возможности GLESP описаны в работах Doroshkevich

¹<https://pla.esac.esa.int/\#home>

(2005; 2011). Данные о полученной выборке объектов находятся в открытом доступе в архиве САО РАН².

Выборка объектов с СЗ-эффектом производилась по трем каталогам: ABELL (Abell Clusters) (Abell et al. 1989), PSZ1 (Aghanim et al. 2020a) и второй архив данных Planck PSZ2 (Aghanim et al. 2020b). Области, в которых нет СЗ-эффекта, выбирались в случайных полях на глаз. Была произведена проверка наличия объектов на картах, использованных в качестве карт без СЗ-эффекта. Критерием отсутствия объекта являлось отсутствие сигнала на частоте 217 ГГц и выше. Это было сделано до обучения модели, для того чтобы не зашумлять выборку некорректными данными, что особенно актуально при небольшом количестве объектов в выборке. Для выборки большего объема проверку можно было бы не делать. Для обучения сети был составлен каталог изображений с СЗ-эффектом, который состоит из 1000 объектов с данными на частотах 100, 143, 217, 353 и 545 ГГц.

За счет характерной особенности СЗ-эффекта — уменьшения интенсивности сигнала на 217 ГГц и ниже — можно идентифицировать кандидаты в скопления галактик и отличить их от сигнала других радиоисточников, у которых такой эффект не наблюдается (Aghanim et al. 2020b).

Помимо того, что программный код GLESP был использован для преобразования наблюдений телескопа Planck, с его помощью также были вырезаны зоны размером $30' \times 30'$ в окрестностях объекта с СЗ-эффектом на частотах 100, 143, 217, 353 и 545 ГГц.

Полученный таким образом наш набор данных состоит из двух классов: изображения с СЗ-эффектом и без него. Каждый класс содержит 1000 объектов, которые представляют собой группы из пяти изображений.

Для каждого объекта мы подготавливаем изображения участков неба размером $30'$. При перенесении изображения на новую сетку пикселей мы присваиваем каждому узлу сетки значение ближайшего узла на старой сетке (т.н. «интерполяция методом ближайшего соседа»). Новая сетка нужна для уменьшения размера изображения, так как при использовании исходных изображений затрачивается много машинного ресурса. Обучение на изображениях большего размера нецелесообразно, так как увеличенное изображение не несет большего количества информации. Применялась круглая маска вокруг объектов как с СЗ-эффектом, так и без него. На рис. 1 показан пример объекта PSZ2 G008.94-81.22 из каталога Planck (Aghanim et al. 2020a) с СЗ-эффектом.

Мы разделили всю выборку на обучающую, валидационную и тестовую в соотношении 70/15/15, выбрав в каждом классе 150 случайных объектов в валидационную выборку и 150 в тестовую. Таким образом, сбалансированность всех трех полученных выборок сохранилась. Для расширения обучающей выборки мы использовали аугментации (небольшие трансформации) исходных объектов: поворот на случайный угол и небольшие смещения по горизонтали и вертикали. Важно, чтобы аугментации были одинаковыми для всех частот одного объекта.

3. ОПИСАНИЕ АНАЛИТИЧЕСКОГО АЛГОРИТМА

Мы решаем задачу бинарной классификации. Одной из первых классификационных нейросетей была AlexNet (Krizhevsky et al. 2012), развитием которой стала архитектура VGG (Simonyan and Zisserman 2014). Обе архитектуры используют блоки сверточных слоев и пулинга, т.е. постепенное уменьшение пространственного разрешения. Веса модели инициализировались из равномерного распределения. При этом был обнулен последний слой нормализации по минибатчам в каждой остаточной ветви, за ними следует несколько полносвязных слоев. Simonyan and Zisserman (2014), Krizhevsky et al. (2012) условно обозначают, что сверточные блоки отвечают за извлечение признаков, в то время как цель полносвязных слоев — комбинирование извлеченных признаков для получения ответов. Эти архитектуры имеют несколько ограничений. Во-первых, полносвязные слои содержат очень много параметров и являются самой вычислительно сложной частью моделей. Во-вторых, более глубокие архитектуры на основе таких моделей не обучаются из-за затухания градиента (Hochreiter et al. 2001).

Решением этих проблем стали архитектуры ResNet (He et al. 2015) и InceptionNet (Szegedy et al. 2015) в библиотеке машинного обучения PyTorch с открытым исходным кодом³. ResNet использует «обходные пути» (residual connections) в каждом сверточном блоке, что позволяет градиенту течь через них и не затухать. В каждом блоке есть обучающийся параметр, который регулирует, какая доля сигнала проходит через обходной путь. У сети имеется возможность сойтись к решению, при котором в части блоков весь сигнал пойдет через обходной путь, то есть часть блоков будет отключена. Благодаря «обходным путям» удалось обучить очень глубокие ResNet-модели (до 152 слоев) (He et al. 2015). InceptionNet

²http://sed.sao.ru/vo/planck_maps/

³<https://github.com/pytorch/vision/blob/master/torchvision/models/resnet.py>

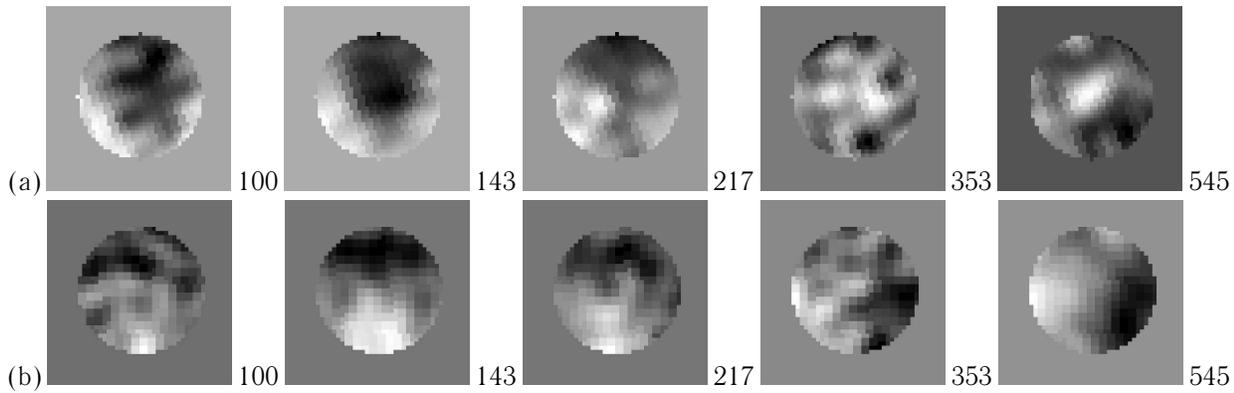


Рис. 1. Пример объектов выборки для обучения нейронной сети соответственно на 100, 143, 217, 353 и 545 ГГц: (a) — объект PSZ2 G008.94–81.22 с СЗ-эффектом; (b) — участок неба без СЗ-эффекта.

использует промежуточные выходы из модели, что также помогает избежать затухания градиентов во время обучения. Дальнейшее развитие архитектур разделилось на два вектора: модели, дающие наилучшее качество классификации, и модели, вычисляющиеся быстро (например, MobileNet (Howard et al. 2017)).

Из-за небольшого количества данных в обучающей выборке мы не имеем возможности использовать глубокие архитектуры, нацеленные на достижение максимального качества ценой длительных вычислений. Мы обучили MobileNet, VGG и в итоге остановились на ResNet18. Эта сеть показывает качество, сравнимое с другими моделями, но в то же время дает большую стабильность метрик качества при различном разбиении выборки на обучающую и тестовую. Мы предполагаем, что такая архитектура более удачно подходит для этой цели. Поскольку описание архитектуры этой модели подробно изложено в статье разработчиков, мы не приводим его здесь. Мы опускаем подробное сравнение результатов на разных архитектурах, так как основной нашей задачей является исследование принципиальной возможности использования нейросетевого подхода для детектирования объектов.

В качестве программного инструмента для обучения применяется библиотека PyTorch (Paszke et al. 2019) без предобученных весов. Наша сеть принимает пятиканальные изображения, что делает нецелесообразным перенос весов от сети, обученной на трехканальных (RGB) изображениях из ImageNet (Russakovsky et al. 2014). В качестве функции ошибок используется бинарная кросс-энтропия. Нейросеть обучается с помощью стохастического градиентного спуска оптимизатором RAdam (Liu et al. 2020). Мы взяли стандартные значения параметров RAdam, кроме параметра $\text{learning rate} = 2 \times 10^{-4}$. Этот параметр подбирался по валидационной выборке.

4. РЕЗУЛЬТАТЫ

4.1. Сравнение ResNet18 и RandomForest

Для оценки качества работы классификаторов использовались такие метрики, как Accuracy, Recall, Precision, F1, ROC AUC. Метрика Accuracy характеризует долю верно классифицированных объектов. Использование Accuracy обосновано, так как классы в представленном датасете сбалансированы. Полнота (Recall) характеризует долю правильно идентифицированных объектов из представленных в выборке, а точность Precision — долю правильно идентифицированных объектов из предсказанных. F1-мера связывает метрики Precision и Recall и вводится как их среднее гармоническое (Powers 2007). Если значение F1 близко к единице, то модель классификатора наиболее точна. ROC (Receiver Operating Characteristic) — графическая характеристика качества бинарного классификатора, зависимость доли верных положительных классификаций от доли ложных положительных классификаций при варьировании порога решающего правила. Метрика ROC AUC является площадью под ROC-графиком и ее часто используют для оценки качества упорядочивания объектов двух классов при работе этого алгоритма. Свойства метрик подробнее описаны в Приложении.

Обучение нейросети на тренировочной выборке проходило в течение 100 эпох (параметр «эпоха» характеризует уровень натренированности нейросети: количество проходов через всю обучающую выборку), после которых производилось сравнение по описанным метрикам. Выбирались те веса модели, для которых функция потерь Loss на валидационной выборке показывает наименьшее значение.

Для сравнения мы также протестировали возможность классификации с помощью подхода, не

Таблица 1. Результаты классификации объектов на тестовой выборке. ResNet18 и RandomForest — это методы, используемые в данной работе

Алгоритм	F1	Accuracy	Recall	ROC AUC
MMF1	—	—	0.747	—
MMF3	—	—	0.740	—
PwS	—	—	0.620	—
RandomForest	0.646	0.653	0.633	0.716
ResNet18	0.891	0.893	0.873	0.964

включающего нейронные сети. В этом случае требуется подготовить признаки вручную. В качестве признаков для каждого канала были выбраны «среднее», «дисперсия», а также гистограмма из пяти бинов с фиксированными порогами. Использовалась модель «ансамбль деревьев» (RandomForest) с числом деревьев 300, число объектов — 1000, минимальное число объектов в узле — 2, число переменных в подборке — 10. Цель этого эксперимента — показать, что задача не решается «напрямую», и использование нейросетей в поиске объектов с СЗ-эффектом является обоснованным. Результаты сравнения даны в таблице 1.

В соответствии с результатами, приведенными в таблице 1, ResNet18 существенно превосходит RandomForest по каждой из представленных метрик, что свидетельствует об эффективности предложенного подхода. Преимущество подхода с использованием ResNet18 в том, что нет необходимости заниматься подготовкой признаков для обучения, сеть принимает на вход сами изображения по пяти частотам. Это же касается оценок и сравнения метода по Recall: ResNet18 также превосходит RandomForest. На рис. 2 изображено изменение числового значения метрики Accuracy, которая характеризует долю верно классифицированных объектов. Можно сделать вывод, что доля верно классифицированных объектов увеличивается в процессе обучения. На рис. 3 показано, как увеличивается числовое значение F-меры в процессе обучения, следовательно, качество модели возрастает.

Рисунки 4 и 5 иллюстрируют функцию потерь Loss на тренировочной и валидационных выборках. Можно видеть, что итерационный процесс сошелся, и переобучение не происходит. Сходимость обучения можно также наблюдать и на рис. 2, 3, демонстрирующих зависимости Accuracy и F1 от эпох.

Мы также показываем кривую обучения ROC нашей модели. Кривая является графиком, отображающим соотношение между количеством верно

классифицированных объектов с признаком и числом тех объектов, которые не несли признака, но были отнесены к верно классифицируемым. График на рис. 6 наглядно иллюстрирует, что ResNet18 существенно превосходит RandomForest вне зависимости от порога.

Было показано, что использование предложенного подхода на основе нейросетевой модели ResNet18 помогает достичь лучшего качества, чем применение алгоритма RandomForest и других рассмотренных моделей. Для нашей модели площадь под кривой ROC (статистика AUC) равна 0.964. Как говорится в работе Carvalho et al. (2012) (раздел 2), ROC AUC не зависит от выбора одного порога. Можно показать, что ROC AUC равна вероятности того, что классификатор ранжирует случайно выбранный пример, сгенерированный из выборки объектов с СЗ-эффектом, выше, чем случайно выбранный пример, сгенерированный из выборки объектов без СЗ-эффекта.

4.2. Сравнение моделей машинного обучения с MMF1, MMF3 и PwS на метрике Recall

Имеется выборка из 1000 объектов, о наличии СЗ-эффекта которых заведомо известно. Сравнить эти данные с простыми метриками мы можем только по полноте. Мы провели сравнение ResNet18 и RandomForest по полноте (Recall) с ранее приведенными простыми методами MMF1 и MMF3 (Herranz et al. 2002, Melin 2006), PwS (PowellSnakes) (Carvalho et al. 2012). Многочастотный согласованный фильтр MMF увеличивает контраст (отношение S/N) объектов известной формы и с известным законом излучения по набору наблюдений, содержащих коррелированные сигналы загрязнения. Его применение позволяет извлекать объекты с СЗ-эффектом, используя многочастотные карты. Этот метод основан на универсальной частотной зависимости теплового излучения в объектах с СЗ-эффектом в предположении, что электроны в этих объектах нерелятивистские, и использовании пространственного шаблона (сеток, метрик и т.д.).

Суть методов MMF1, MMF3 и PwS в следующем (Verkhodanov et al. 2015). Вначале рассчитывается синтетическая карта реликтового излучения для некоторого модельного профиля скопления галактик, например, из работы Agnaud et al. (2010). Карта вычисляется для каждого частотного бина. Параметром для такого набора карт служит угловой размер модельного скопления. Далее карты, полученные по результатам наблюдений, подвергаются свертке с синтетическими картами. Положение пиков в такой свертке указывает на положение кандидатов в наблюдаемые скопления.

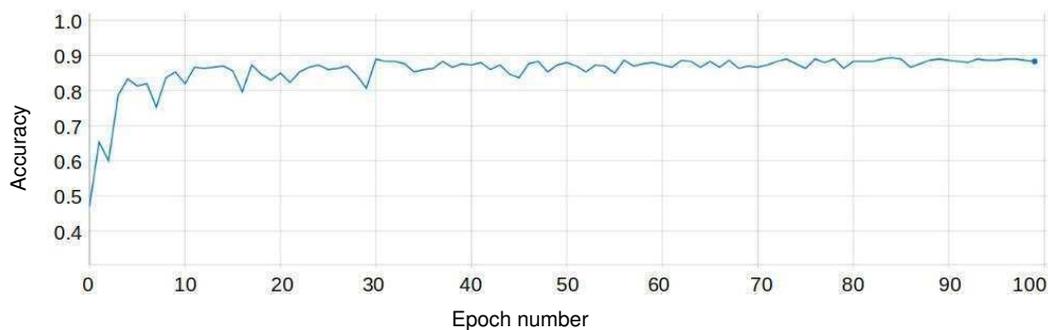


Рис. 2. Изменение числового значения метрики Accuracy на валидационных данных (ResNet18) в зависимости от эпохи.

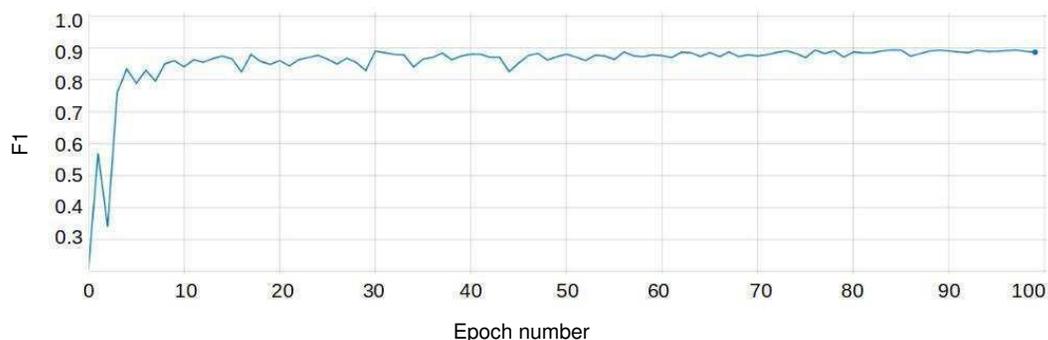


Рис. 3. Изменение числового значения F1 на валидационных данных (ResNet18) в зависимости от эпохи.

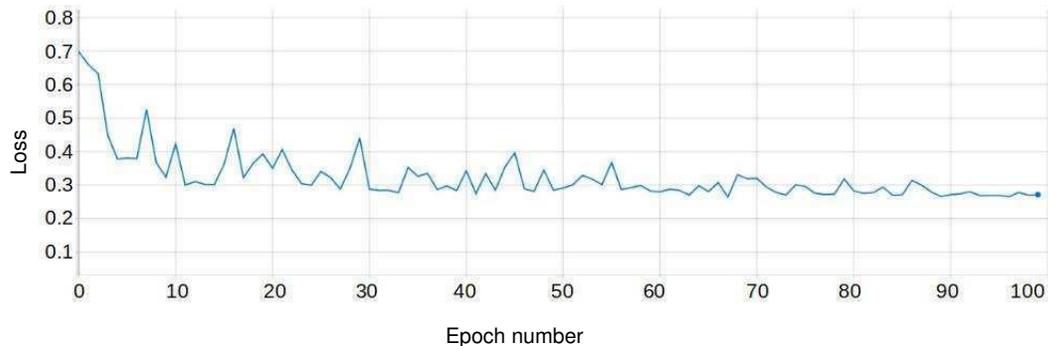


Рис. 4. Изменение числового значения Loss на валидационных данных (ResNet18) в зависимости от эпохи.

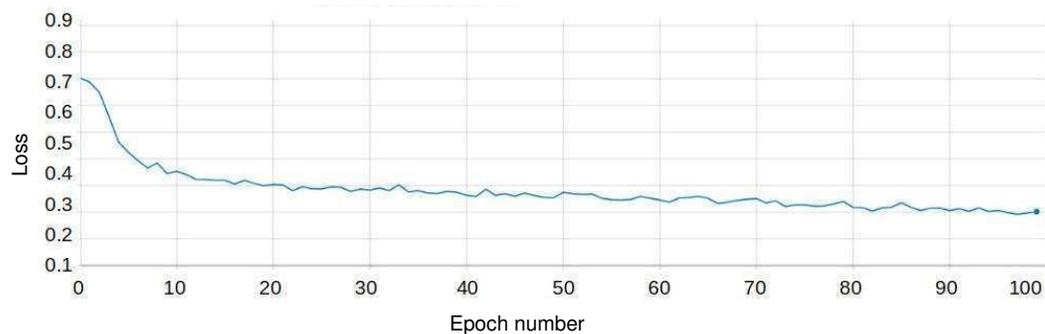


Рис. 5. Изменение числового значения Loss на тренировочных данных (ResNet18) в зависимости от эпохи.

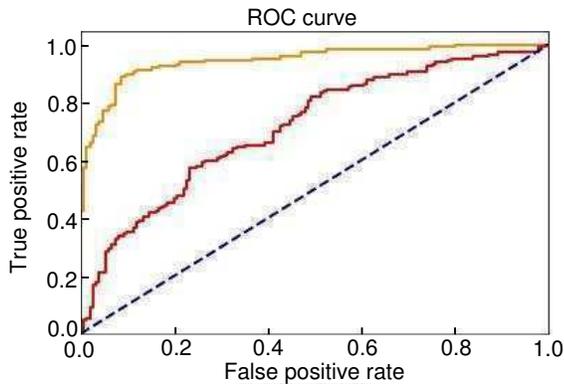


Рис. 6. Кривая рабочей характеристики классификатора (Receiver operating characteristic, ROC). Красным цветом показан RandomForest, оранжевым цветом — ResNet18.

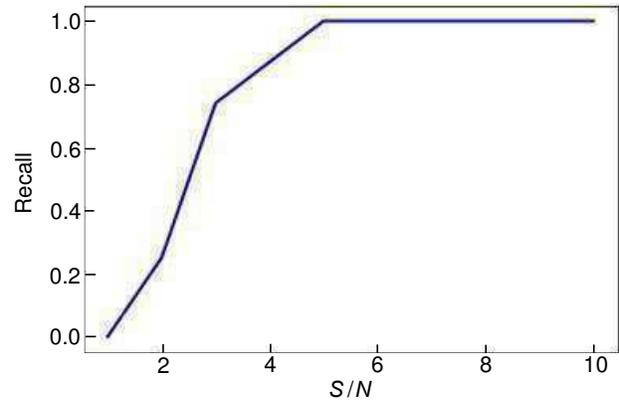


Рис. 7. Качество обнаружения эффекта (полнота) на модельных данных при разных соотношениях S/N .

Максимизируя таким образом величину пиков, по угловому размеру модельного скопления можно оценить, например, угловой размер наблюдаемых скоплений.

С формальной стороны методы MMF1, MMF3 и PwS отличаются алгоритмами минимизации ошибки в расчете положений и параметров скоплений. Концептуальное различие между ними в том, что метод PwS основан на байесовском подходе к оценке гипотез (в данной задаче — оценке параметров скоплений) (Carvalho et al. 2009; 2012). Каталог PSZ2 из работы Ade et al. (2016) составлен из объединенных трех каталогов с использованием методов MMF1, MMF3 и PwS. В каталоге PSZ2 есть такие объекты с СЗ-эффектом, которые были обнаружены одним методом, но не были найдены другим. Таким образом, для каждого метода в каталоге PSZ2 можно выделить те объекты, которые с его помощью удалось обнаружить — True positive (TP), и те, которые не были обнаружены — False positive (FP). Исходя из этих данных, мы можем сделать оценку и сравнение по Recall.

Результаты сравнения MMF1, MMF3 и PwS с RandomForest и ResNet18, приведенные в таблице 1, говорят о том, что по метрике Recall наша модель ResNet18 превосходит другие методы, но при этом нужно помнить, что метод анализа данных различен, и возможно сравнение только по Recall.

4.3. Сравнение качества обнаружения на модельных данных

Для проверки обнаружения объектов по предложенной методике с помощью пакета GLESP (Doroshkevich 2005) мы сгенерировали случайные гауссовы карты с различным отношением S/N для объектов с СЗ-эффектом. Для построения моделей мы использовали приближения

к СЗ-сигналу с коэффициентами $k_{100} = -0.4$, $k_{143} = -0.5$, $k_{217} = 0.0$, $k_{353} = 1.0$, $k_{545} = 0.5$ для частот 100, 143, 217, 353 и 545 ГГц соответственно. Относительные коэффициенты рассчитаны по данным наблюдений скопления Abell 2319 (Abell et al. 1989, Aghanim et al. 2020a; b). Модельный источник генерировался с шириной, по половинной мощности равной $5'$. Отношение S/N вычислялось для сигнала на частоте 353 ГГц и масштабировалось пропорционально на других частотах HF1.

Всего было сгенерировано 100 моделей радиоисточников. Во всех сгенерированных объектах СЗ-эффект должен наблюдаться. На рис. 7 приведены результаты сравнения качества (полноты) при различных отношениях S/N .

Как видно на рис. 7, полнота ответов модели (Recall) растет с упрощением задачи. При соотношении $S/N = 1$ сеть предсказывает, что на всех объектах нет эффекта. С усилением сигнала увеличивается уверенность сети, и на соотношениях $S/N = 5$ и больше сеть классифицирует все объекты правильно. Используя кубическую интерполяцию, мы получили, что $\text{Recall} = 0.873$ соответствует $S/N \approx 3.42$. Примерно с таким соотношением S/N детектирование объектов с СЗ-эффектом возможно с заявленным качеством.

5. ЗАКЛЮЧЕНИЕ

Мы построили модель поиска объектов с СЗ-эффектом⁴. Было показано, что использование предложенного подхода на основе нейросетевой модели ResNet18 помогает достичь лучшего качества, чем использование алгоритма RandomForest и других рассмотренных моделей. Продемонстрировано, что детектирование объектов с СЗ-эффектом возможно с заявленным качеством $\text{Recall} = 0.873$ при отношении $S/N \approx 3.42$ и выше.

Мы показали, что данный подход работает и может быть применен к анализу массивов данных для поиска объектов с СЗ-эффектом. Реализованный подход дополняет существующие алгоритмы (Bonjean 2020) поиска таких объектов, позволяя работать с данными в формате пакета GLESP.

БЛАГОДАРНОСТИ

Авторы выражают свою признательность Фонду поддержки научных, культурных и образовательных инициатив «Траектория», а также В. С. Ивашкину за содействие при проведении данной работы. В работе использован пакет GLESP <http://www.glesp.nbi.dk> для анализа протяженного излучения на сфере. Авторы выражают благодарность рецензенту за полезные замечания, которые позволили улучшить текст статьи.

КОНФЛИКТ ИНТЕРЕСОВ

Авторы заявляют об отсутствии конфликта интересов.

СПИСОК ЛИТЕРАТУРЫ

1. G. O. Abell, H. G. Corwin, and R. P. Olowin, *Astrophys. J. Suppl.* **70**, 1 (1989).
2. C. de Breuck, W. van Breugel, H. J. A. Röttgering, and G. Miley, *Astron. and Astrophys. Suppl.* **143**, 303 (2000).
3. K. Simonyan and A. Zisserman, arXiv:1409.1556, (2014).
4. P. A. R. Ade et al. (Planck Collab.), *Astron. and Astrophys.* **594**, 19 (2016).
5. N. Aghanim et al. (Planck Collab.), *Astron. and Astrophys.* **641**, id. A1 (2020a).
6. N. Aghanim et al. (Planck Collab.), *Astron. and Astrophys.* **641**, id. A6 (2020b).
7. S. W. Allen, A. E. Evrard, and A. B. Mantz, *Annual Rev. Astron. Astrophys.* **49**, 409 (2011).
8. A. K. Aniyan and K. Thorat, *Astrophys. J. Suppl.* **230**, id. 20 (2017).
9. M. Arnaud, G. W. Pratt, R. Piffaretti, et al., *Astron. and Astrophys.* **517**, id. A92 (2010).
10. D. Barbosa, J. G. Bartlett, A. Blanchard, and J. Oukbir, *Astron. and Astrophys.* **314**, 13 (1996).
11. D. Baron and D. Poznanski, *Monthly Notices Royal Astron. Soc.* **465**, 4530 (2017).
12. K. Basu, J. Erler, J. Chluba, et al. *Bull. Amer. Astron. Soc.* **51**, 302 (2019).
13. G. Blumenthal and G. Miley, *Astron. and Astrophys.* **80**, 13 (1979).
14. V. Bonjean, *Astron. and Astrophys.* **634**, id. A81 (2020).
15. A. M. Bykov, H. Bloemen, and Y. A. Uvarov, *Astron. and Astrophys.* **362**, 886 (2000).
16. P. Carvalho, G. Rocha, and M. P. Hobson, *Monthly Notices Royal Astron. Soc.* **393**, 681 (2009).
17. P. Carvalho, G. Rocha, M. P. Hobson, and A. Lasenby, *Monthly Notices Royal Astron. Soc.* **427**, 1384 (2012).
18. J. J. Condon, W. D. Cotton, E. W. Greisen, et al., *Astron. J.* **115**, 1693 (1998).
19. D. Cunnamea, A. Faltenbacher, C. Cress, and S. Passmoor, *Monthly Notices Royal Astron. Soc.* **397**, L41 (2009).
20. W. Dobbels, M. Baes, S. Viaene, et al., *Astron. and Astrophys.* **634**, id. A57 (2020).
21. A. G. Doroshkevich, P. D. Naselsky, O. V. Verkhodanov, et al., *Int. J. Mod. Phys. D* **14** (02), 275 (2005).
22. A. G. Doroshkevich, O. V. Verkhodanov, P. D. Naselsky, et al., *Int. J. Mod. Phys. D* **20** (06), 1053 (2011).
23. M. Hasselfield, M. Hilton, T. A. Marriage, J. Cosmology Astroparticle Physics, No. 07, id. 008 (2013).
24. K. He, X. Zhang, S. Ren, and J. Sun, arXiv:1512.03385 (2015).
25. D. Herranz, J. L. Sanz, M. P. Hobson, et al., *Monthly Notices Royal Astron. Soc.* **336**, 1057 (2002).
26. S. Hochreiter, Y. Bengio, P. Frasconi, et al., in *A Field Guide to Dynamical Recurrent Neural Networks*, Ed. by J. F. Kolen and S. C. Kremer (Wiley-IEEE Press, 2001).
27. M. Hossin and M. N. Sulaiman, *Int. J. Data Mining and Knowledge Manag. Proc.* **5** (2), 1 (2015).
28. A. G. Howard, M. Zhu, B. Chen, et al., arXiv:1704.04861 (2017).
29. M. L. Khabibullina and O. V. Verkhodanov, *Astrophysical Bulletin* **64**, 123 (2009).
30. T. V. Keshelava and O. V. Verkhodanov, *Astrophysical Bulletin* **70**, 257 (2015).
31. J. Kormendy and S. Djorgovski, *Annual Rev. Astron. Astrophys.* **27**, 235 (1989).
32. A. V. Kravtsov and S. Borgani, *Annual Rev. Astron. Astrophys.* **50**, 353 (2012).
33. A. Krizhevsky, I. Sutskever, and G. E. Hinton, in *Proc. 25th Conf. on Advances in Neural Information Processing Systems, Lake Tahoe, USA, 2012*, Ed. by F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, 1097 (2012).
34. L. Liu, H. Jiang, P. He, W. Chen, et al., arXiv:1908.03265 (2020).
35. J. B. Melin, J. G. Bartlett, and J. Delabrouille, *Astron. and Astrophys.* **459** (2), 341 (2006).
36. D. Nagai, *AIP Conf. Proc.* **1632** (1), 88 (2014).
37. A. Oronovskaya, https://habr.com/ru/users/sunny_space/posts/ (2018).
38. Yu. N. Parijskij, W. M. Goss, A. I. Kopylov, et al., *Bull. Spec. Astrophys. Obs.* **40**, 5 (1996).
39. A. Paszke, S. Gross, F. Massa, et al., arXiv:1912.01703 (2019).
40. D. M. W. Powers, Australia Technical Report SIE-07-001 (2007).
41. R. B. Rengelink, Y. Tang, A. G. de Bruyn, et al., *Astron. and Astrophys. Suppl.* **124**, 259 (1997).

42. O. Russakovsky, J. Deng, H. Su, et al., arXiv:1409.0575 (2014).
43. C. L. Sarazin, Rev. Modern Physics, **58** (1), 1 (1986).
44. D. I. Solovoyov and O. V. Verkhodanov, Astrophysical Bulletin **72**, 217 (2017).
45. C. Szegedy, W. Liu, Y. Jia, et al., in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition, Boston, USA, 2015*, <https://ieeexplore.ieee.org/document/7298594> (2015).
46. S. Tacchella, B. Diemer, L. Hernquist, et al., Monthly Notices Royal Astron. Soc. **487**, 5416 (2019).
47. Y. Tao, Y. Zhang, C. Cui, and Z. Ge, arXiv:1801.04839 (2018).
48. G. Ucci, A. Ferrara, S. Gallerani, and A. Pallottini, Monthly Notices Royal Astron. Soc. **465**, 1144 (2017).
49. K. Vanderlinde, T. M. Crawford, T. de Haan, et al., Astrophys. J. **722** (2), 1180 (2010).
50. B. P. Venemans, H. J. A. Röttgering, G. K. Miley, et al., Astron. and Astrophys. **461**, 823 (2007).
51. О. В. Верходанов, Ю. Н. Парийский, *Радиогалактики и космология* (Физмалит, Москва, 2009).
52. O. V. Verkhodanov, E. K. Majorova, O. P. Zhelenkova, et al., Astrophysical Bulletin **70**, 156 (2015).
53. O. V. Verkhodanov, N. V. Verkhodanova, O. S. Ulakhovich, et al., Astrophysical Bulletin **73**, 1 (2018).
54. A. A. Zaporozhets and O. V. Verkhodanov, Astrophysical Bulletin **74**, 247 (2019).
55. Ya. B. Zeldovich and R. A. Sunyaev, Astrophys. and Space Sci. **4**, 301 (1969).

ПРИЛОЖЕНИЕ

В основе работы лежат метрики, которые работают с двумя классами. Назовем объекты, в которых есть СЗ-эффект, positive классом (P), а в которых нет — negative (N). Будем называть ответы, которые мы знаем заранее, ground truth (gt), а ответы, которые мы предсказываем — prediction (pred).

Рассмотрим сначала один объект выборки. Мы знаем заранее правильный ответ (P или N) и мы предсказываем какой-то ответ (P или N). Есть четыре варианта:

gt = P, pred = P — здесь СЗ-эффект был на самом деле, и мы его предсказали. Назовем это true positive (TP);

gt = N, pred = N — СЗ-эффекта не было, и мы предсказали, что его нет. Это «true negative» (TN);

gt = P, pred = N — СЗ-эффект был, но мы его не заметили. Это «false negative» (FN);

gt = N, pred = P — СЗ-эффекта не было, но мы считаем, что он был. Это «false positive» (FP).

Для списка объектов выборки у нас есть список правильных ответов (например, gt = PPPNNNN) и список наших предсказаний (например, pred = PNNPPPN). Каждый объект относится к одному из четырех случаев: TP, TN, FP, FN.

На основе этих понятий можно сформировать все метрики, которые возможны для двух классов. Нас интересуют две:

Ассигасу — это отношение доли объектов, где правильно угаданы объекты с СЗ-эффектом, ко всей выборке объектов. То есть

$$\frac{TP + TN}{TP + TN + FP + FN},$$

Recall — это сколько из объектов gt = P предсказано как P. То есть

$$\frac{TP}{TP + FN}.$$

$$\frac{TP}{TP + FP}.$$

F1 — это среднее гармоническое значение между Precision и Recall. То есть

$$\frac{2TP}{2TP + FP + FN}$$

ROC AUC — это площадь под кривой ROC. То есть

$$\frac{1 + TPR - FPR}{2},$$

где True Positive Rate (TPR) — процент точек класса 1, которые верно классифицированы нашим алгоритмом, а False Positive Rate (FPR) — это процент точек класса 0, которые неверно классифицированы нашим алгоритмом.

Метрики Ассигасу можно продемонстрировать на следующем примере:

gt = PPPPNNNN, pred = NNNNNNNN и Accuracy = 0.5;

gt = PPPPNNNN, pred = PNPNPNPN и Accuracy = 0.5;

gt = PNNNNNNN, pred = NNNNNNNN и Accuracy = 7/8.

Видно, что для сбалансированных выборок (когда в gt примерно одинаковое количество P и N) Ассигасу для константного предсказания всегда N или случайного значения дает результат не ниже 0.5. Но для несбалансированных классов может

получиться большее значение, что показывает уже не качество наших предсказаний, а сбалансированность выборки. Итог: Ассигасу полезна только в сбалансированных выборках.

Границы применимости метрики Recall можно продемонстрировать, например, так:

gt = PPPPNNNN, pred = NNNNNNNN и Recall = 0;

gt = PPPPNNNN, pred = PNPNPNPN и Recall = 0.5;

gt = PPPPNNNN, pred = PPPPPPPP и Recall = 1;

gt = PNNNNNNN, pred = PPPPPPPP и Recall = 1.

Выбить 1 по Recall не так уж сложно — нужно просто всегда предсказывать Р. Поэтому смотреть на метрику Recall в отрыве от других нельзя: можно не заметить, как модель вместо умных предсказаний просто выдает Р.

Метрики работают по такому принципу:

- Выборка сбалансирована, поэтому можно использовать Ассигасу.
- Если мы сверяемся с каталогами, то получается, что в каталогах есть только правильные объекты

(ТР), и качество каталога мы можем измерить только по отсутствию в нем каких-то объектов (FN). Ассигасу требует ещё TN и FP, их у нас нет. Остается только использовать Recall.

- Можно считать Recall каталогов MMF1, MMF3 и PwS, так как они не были созданы на максимизацию Recall.

Наша нейронная сеть в качестве результата выдаёт число от 0 до 1 — вероятность того, что объект принадлежит классу, — и нам нужен ответ: Р или N. То есть надо установить какой-то порог, выше которого — Р. Мы делим выборку на три части:

- Train — учимся, т.е. сдвигаем веса сети так, чтобы минимизировать функцию потерь;
- Validation — подвыборка, по которой подбирается порог, т.е. перебираются пороги от 0 до 1 и выбирается тот, который в среднем на валидационной выборке дает лучшее Ассигасу;
- Test — на этой подвыборке измеряются итоговые Ассигасу и Recall, которые попадают в таблицу для сравнения.

Search for Galaxy Cluster Candidates in the Cosmic Microwave Background Maps of the Planck Space Mission Using a Convolutional Neural Network Based on the Method of Tracing the Sunyaev–Zeldovich Effect

O. V. Verkhodanov¹, A. P. Topchieva², A. D. Oronovskaya³, S. A. Bazrov³, and D. A. Shorin³

¹Special Astrophysical Observatory, Russian Academy of Sciences, Nizhnii Arkhyz, 369167 Russia

²Institute of Astronomy, Russian Academy of Sciences, Moscow, 119017 Russia

³Astrophysical School ꞑꞑTraektoriaçç, Moscow, 109147 Russia

We propose a method of searching for radio sources exhibiting the Sunyaev–Zeldovich effect in the multi-frequency emission maps from the Planck mission data using a convolutional neural network. A catalog for recognizing radio sources is compiled using the GLESP pixelation scheme at the frequencies of 100, 143, 217, 353, and 545 GHz. The quality of the proposed approach is evaluated and the quality of the dependence of model data on the S/N ratio is estimated. We show that the presented neural network approach allows the detection of sources with the Sunyaev–Zeldovich effect. The proposed method can be used to find the most likely galaxy cluster candidates at large redshifts.

Keywords: *cosmic microwave background—galaxies: clusters: general*