# **Данные.**

## **Структура данных.**

Подбор данных в машинном обучении являются одним из самых важных аспектов. От их качества и количества во многом зависит, насколько удачной получится модель.

Во-первых, данные должны быть репрезентативными и отражать все разнообразие объектов, которые модель должна распознавать. Если набор данных будет смещен в сторону определенных категорий или типов объектов, модель может оказаться неспособной обобщать и работать эффективно на реальных данных. Необходимо обеспечить, чтобы тренировочная, валидирующая и тестовая выборки равномерно представляли все релевантные классы. Во-вторых, важно, чтобы данные были качественно размечены и аннотированы. Неточные, неполные или ошибочные метки могут привести к тому, что модель будет обучаться на ложных паттернах и демонстрировать низкую производительность. Тщательная ручная разметка или использование проверенных автоматических инструментов - залог успеха. В-третьих, разнообразие данных играет ключевую роль. Чем шире спектр условий, отраженных в данных (различные ракурсы, освещение, погодные условия и т.д.), тем устойчивее и адаптивнее будет обученная модель. Это особенно важно для задач компьютерного зрения, где модель должна работать стабильно в реальных, неконтролируемых условиях. Наконец, объем данных имеет критическое значение. Большие размеченные датасеты позволяют обучать более глубокие и мощные модели, которые способны лучше обобщать и выявлять скрытые закономерности. Часто успех приходит к тем, кто располагает наибольшим количеством качественных тренировочных данных.

Таким образом, подбор подходящих данных является определяющим фактором в успехе любого проекта машинного обучения. Именно поэтому этому этапу необходимо уделять особое внимание и прикладывать максимум усилий, чтобы обеспечить модель всем необходимым для ее эффективной и надежной работы.

Для обучения модели по распознаванию пашен, вырубок и карьеров необходимы следующие основные типы данных: набор изображений и набор аннотаций к ним. Датасет представляет собой спутниковые снимки или аэрофотоснимки высокого разрешения, охватывающие географические регионы, схожие с Ленинградской областью. Изображения должны включать пашни, вырубки, карьеры, а также другие типы земного покрова, такие как леса, водные объекты, населенные пункты и т.д. Изображения должны быть представлены в виде набора .jpg или .jpeg спутниковых снимков. Для каждого изображения необходимо иметь аннотации, которые определяют местоположение и границы пашен, вырубок, карьеров и других объектов. Аннотации могут быть в формате полигонов, прямоугольников или других геометрических фигур, охватывающих интересующие области. Точность и достоверность аннотаций имеют большое значение для качества обучения модели.

Разметка и аннотации представляют собой набор текстовых файлов, каждый из которых соотносится с конкретным изображением. В них содержатся координаты точек, которые ограничивают объекты различных классов.

## **Источники данных.**

Для обучения модели спутниковые снимки можно брать из множества как коммерческих, так и открытых источников. Рассмотрим несколько самых популярных источников. Начнём рассмотрение с бесплатных источников.

*NASA Earth Observing System Data and Information System (EOSDIS)*- это обширная инфраструктура, созданная NASA для управления и распространения данных дистанционного зондирования Земли. EOSDIS связывает более 100 спутников и других космических и наземных систем наблюдения Земли, которые генерируют огромные объемы данных. Данные поступают из различных научных инструментов, измеряющих характеристики атмосферы, океанов, суши, криосферы и биосферы Земли. EOSDIS обеспечивает долгосрочное хранение и архивацию этих данных в надежных центрах обработки и распространения. Общий объем архива данных превышает 31 петабайт (2022 г.) и постоянно растет. Система также предоставляет инструменты и вычислительные мощности для обработки сырых данных в удобные для пользователей продукты. Обработка включает калибровку, геокодирование, объединение различных источников и создание тематических наборов данных. Данные EOSDIS доступны бесплатно для научного сообщества и широкой общественности. Существует множество порталов и интерфейсов для поиска, просмотра и загрузки данных, что позволяет ежегодно осуществлять более 1 миллиарда загрузок. EOSDIS также обеспечивает комплексную поддержку ученым и исследователям, работающим с данными дистанционного зондирования, предоставляя инструменты визуализации, аналитики, моделирования и совместной работы. EOSDIS играет ключевую роль в обеспечении открытого доступа к данным наблюдения Земли, что способствует научным исследованиям, решению глобальных экологических проблем и принятию обоснованных управленческих решений.

*Sentinel Hub* - это инновационная платформа, разработанная Sinergise, для эффективного доступа и использования спутниковых данных от программы наблюдения Земли Европейского космического агентства (ЕКА) Copernicus. Sentinel Hub интегрирует данные со спутников серии Sentinel, а также других источников, таких как Landsat и MODIS, создавая единую точку доступа к этой обширной геопространственной информации. Платформа обеспечивает быстрый и простой доступ к актуальным спутниковым снимкам, позволяя пользователям легко находить, просматривать и загружать необходимые данные. Одно из ключевых достоинств Sentinel Hub - его масштабируемая облачная инфраструктура. Она позволяет обрабатывать и предоставлять огромные объемы данных на лету, избавляя пользователей от необходимости хранить и обрабатывать их локально. Платформа предоставляет мощные инструменты для визуализации, анализа и индексации данных, включая возможность создания пользовательских тематических слоев. Sentinel Hub также поддерживает широкий спектр API, которые позволяют интегрировать спутниковые данные в различные приложения и рабочие процессы. Это делает платформу чрезвычайно гибкой и востребованной в самых разных областях: от точного сельского хозяйства до мониторинга окружающей среды и управления стихийными бедствиями. Благодаря своей производительности, масштабируемости и удобству использования, Sentinel Hub стал ключевым инструментом для ученых, аналитиков и специалистов, работающих с данными дистанционного зондирования Земли. Платформа значительно упрощает доступ к этим ценным данным и способствует их более широкому практическому применению.

*USGS EarthExplorer* - это мощный веб-интерфейс, разработанный Геологической службой США (USGS), для поиска, просмотра и загрузки спутниковых данных и других геопространственных ресурсов. EarthExplorer предоставляет доступ к обширной коллекции данных дистанционного зондирования Земли, включая снимки со спутников Landsat, MODIS, Sentinel и многих других источников. Пользователи могут выполнять поиск по широкому спектру критериев, таких как географическое положение, дата съемки, облачность и тематические категории. Одно из ключевых преимуществ EarthExplorer - его интуитивно понятный и удобный интерфейс. Он позволяет легко просматривать предварительные изображения, получать подробную информацию о данных и, при необходимости, загружать их в различных форматах. Кроме того, платформа предлагает мощные инструменты для визуализации и анализа данных, включая возможность создания композитных изображений и применения различных цветовых палитр. EarthExplorer также поддерживает интеграцию с внешними системами и приложениями. Пользователи могут подключать свои учетные записи и использовать данные USGS в своих собственных рабочих процессах, будь то ГИС-приложения, научные модели или прикладные решения. Благодаря своей всеобъемлющей коллекции данных, удобному интерфейсу и расширенным возможностям, USGS EarthExplorer стал незаменимым инструментом для широкого круга пользователей: от ученых и исследователей до специалистов в области охраны окружающей среды, сельского хозяйства, картографии и многих других сфер. Платформа значительно упрощает доступ к ценным геопространственным данным и способствует их эффективному использованию.

Среди коммерческих источников данных можно выделить следующие популярные сервисы:

*DigitalGlobe* является ведущим поставщиком высококачественных космических снимков и геопространственных решений для коммерческих и государственных клиентов по всему миру. Компания была основана в 1992 году и на сегодняшний день управляет группировкой из пяти высокоразрешающих коммерческих спутников дистанционного зондирования Земли. DigitalGlobe предоставляет широкий спектр продуктов и услуг, включая спутниковые изображения, аналитику и дополненную визуализацию данных, которые используются в различных отраслях, таких как сельское хозяйство, городское планирование, энергетика, добыча полезных ископаемых, оборона и национальная безопасность. Компания постоянно совершенствует свои технологии и мощности для поддержки растущих потребностей клиентов в актуальной и высококачественной геопространственной информации. Благодаря своему технологическому лидерству и глобальному охвату, DigitalGlobe является ключевым игроком на рынке дистанционного зондирования Земли, предоставляя критически важные данные и решения для решения широкого круга задач, связанных с наблюдением за Землей.

*Airbus Defence and Space* является ведущим европейским поставщиком передовых технологий для космических, оборонных и связанных с ними рынков. Компания была создана в результате объединения оборонного и космического бизнеса Airbus Group и является одним из подразделений этого крупного авиастроительного концерна. Airbus Defence and Space предлагает широкий спектр продуктов и услуг, включая военные самолеты, беспилотные летательные аппараты, системы ракетной обороны, спутниковые системы наблюдения Земли и связи. Компания располагает передовыми производственными мощностями и научно-исследовательскими центрами по всей Европе, что позволяет ей разрабатывать инновационные решения для государственных и корпоративных заказчиков. Являясь ведущим игроком в европейском оборонном и космическом секторах, Airbus Defence and Space вносит значительный вклад в развитие высокотехнологичных решений, обеспечивающих национальную безопасность, экологический мониторинг и исследование космоса. Компания постоянно инвестирует в НИОКР, чтобы поддерживать свое технологическое превосходство и удовлетворять растущие потребности клиентов по всему миру.

Однако мы решения нашего кейса мы не стали использовать готовый датасет, а создали свой при помощи сервисов Google Earth и Яндекс карты. Для этого мы включили режим «Спутник» в каждом из сервисов и, листая карту, искали местоположение карьеров, пашен и вырубок. Так мы смогли собрать спутниковые снимки центральной и северной части России. Минусами такого способа поиска данных является: среднее качество изображений, трудность поиска. Из плюсов можно выделить простоту доступа, а также сходные географические условия с теми, что нам необходимо распознавать.

## **Архитектура данных**

Архитектура данных для модели, которая будет распознавать пашни, вырубки и карьеры на спутниковых снимках, является важным аспектом при подготовке набора данных. Она определяет, как будут организованы и структурированы данные для эффективного использования в процессе обучения модели. Для спутниковых снимков в качестве формата данных рекомендуется использовать общепринятые растровые форматы, такие как GeoTIFF. Этот формат позволяет сохранять пространственную информацию, что важно для обработки и анализа спутниковых изображений.

Набор данных должен быть организован в виде коллекции спутниковых снимков, каждый из которых будет сопровождаться соответствующими метками (аннотациями) для пашен, вырубок и карьеров. Это позволит модели во время обучения ассоциировать визуальные признаки с соответствующими объектами на снимках. Помимо самих изображений, важно собирать и сохранять метаданные, такие как дата съемки, источник данных, разрешение снимков и другие характеристики. Эта информация может быть полезна как для предобработки данных, так и для последующего анализа результатов. Данные можно хранить в виде иерархической файловой структуры, где каждый класс (пашня, вырубка, карьер) будет представлен отдельной папкой, содержащей соответствующие спутниковые снимки. Это упростит работу с данными и их использование в процессе обучения модели. Для облегчения работы с данными рекомендуется придерживаться единых стандартов в именовании файлов, формате метаданных и других аспектах организации данных. Это повысит удобство использования набора данных. При проектировании архитектуры данных важно обеспечить гибкость, чтобы в дальнейшем можно было легко расширять или модифицировать набор данных по мере необходимости.

Соблюдение этих принципов при разработке архитектуры данных позволит создать структурированный и эффективный набор данных, который будет способствовать успешному обучению модели для распознавания пашен, вырубок и карьеров на спутниковых снимках.

## **Объем данных**

Определение оптимального объема данных для модели, которая будет распознавать пашни, вырубки и карьеры на спутниковых снимках, является важным аспектом при подготовке набора данных. Как правило, чем больше данных, тем лучше результаты обучения модели. Однако следует учитывать, что увеличение объема данных требует больших вычислительных ресурсов и времени на обучение.

Рекомендуемый минимальный объем данных может составлять от нескольких тысяч до десятков тысяч примеров для каждого класса (пашни, вырубки, карьеры). При этом важно, чтобы набор данных был сбалансированным, то есть количество примеров для каждого класса было примерно одинаковым. Это позволит избежать смещения модели в сторону одного из классов. Кроме того, необходимо, чтобы данные были репрезентативными, то есть охватывали разнообразные географические регионы, времена года, типы ландшафтов и другие факторы, чтобы модель могла обобщать и эффективно работать на новых, ранее не встречавшихся данных. Это особенно важно, если модель будет применяться в различных регионах или климатических зонах. Для увеличения размера и разнообразия набора данных можно использовать техники аугментации, такие как поворот, отражение, добавление шума и т.д. Это позволит расширить количество примеров без сбора дополнительных данных.

В целом, при определении оптимального объема данных следует руководствоваться принципом "чем больше, тем лучше", учитывая при этом доступные вычислительные мощности и время, необходимое для обучения модели. По мере необходимости и по мере появления новых данных, объем набора данных можно постепенно увеличивать.

## **Обработка данных**

После того, как данные оказались собраны, необходимо их обработать: нормализовать, масштабировать, обрезать. Это необходимо сделать для лучшего перформанса модели: более быстрого обучения, лучшей сходимости и точности. Расскажем подробней про каждую из них.

Нормализация является важным этапом предварительной обработки данных перед обучением моделей машинного обучения. Она заключается в приведении числовых признаков к общему диапазону значений, что позволяет избежать доминирования признаков с большими абсолютными значениями и обеспечить равномерный вклад каждого признака в обучение модели. Существует несколько основных методов нормализации.

*Min-max нормализация, также известная как feature scaling или min-max scaling,* - это один из наиболее распространенных методов нормализации числовых признаков. Суть метода заключается в том, чтобы привести все значения признака к диапазону от 0 до 1, но сохранить относительные различий между значениями признака. Математически, min-max нормализация рассчитывается по следующей формуле: x\_normalized = (x - min(x)) / (max(x) - min(x)), где x - исходное значение признака, min(x) - минимальное значение признака, max(x) - максимальное значение признака.

В целом, min-max нормализация является эффективным и широко применяемым методом предварительной обработки данных, особенно когда необходимо сбалансировать влияние признаков с разными шкалами.

Данные для обучения такой модели должны быть разбиты на несколько частей: обучающую, валидационную и тестовую выборки. Рассмотрим предназначение каждой из них подробнее.

*Z-нормализация, также известная как стандартизация*, является одним из видов нормализации данных, применяемых при подготовке данных для обучения моделей машинного обучения. Этот метод преобразует исходные значения признаков таким образом, чтобы они имели нулевое среднее и единичное стандартное отклонение. Формула Z-нормализации выглядит следующим образом: z = (x - μ) / σ, где z - нормализованное значение признака, x - исходное значение признака, μ - среднее значение признака, σ - стандартное отклонение признака

Применение Z-нормализации позволяет добиться нескольких важных преимуществ:

1. Сопоставимость признаков: Приведение признаков к единому масштабу позволяет сравнивать их вклад в модель, так как они становятся безразмерными.

2. Устойчивость к выбросам: Z-нормализация делает модель менее чувствительной к выбросам в данных, так как она ориентируется на среднее и стандартное отклонение, а не на абсолютные значения.

3. Улучшение сходимости: Нормализованные данные часто способствуют более быстрой и стабильной сходимости алгоритмов обучения, таких как градиентный спуск.

Таким образом, Z-нормализация является важным этапом предобработки данных, повышающим качество и эффективность обучения моделей машинного обучения, в том числе и при поиске карьеров, вырубок и пашен.

Масштабирование и обрезка изображений являются важными предварительными этапами при подготовке данных для обучения моделей машинного обучения, особенно в задачах, связанных с компьютерным зрением. Эти операции позволяют стандартизировать размер и формат изображений, что имеет ряд ключевых преимуществ. Во-первых, масштабирование помогает обеспечить согласованность размера изображений в наборе данных. Это важно, поскольку большинство алгоритмов глубокого обучения требуют, чтобы все входные изображения имели одинаковые пространственные размеры. Использование изображений различных размеров может привести к проблемам при обучении модели и снижению ее производительности. Масштабирование позволяет привести все изображения к общему размеру, что упрощает последующую обработку и ускоряет обучение. Во-вторых, обрезка изображений играет важную роль в выделении релевантных областей, отсекая лишние фоновые элементы. Это особенно важно, когда на изображениях присутствуют большие пустые пространства или посторонние объекты, не имеющие отношения к задаче. Обрезка позволяет сфокусировать модель на ключевых областях интереса, что повышает ее способность к обобщению и точность распознавания. Кроме того, масштабирование и обрезка изображений помогают уменьшить размер входных данных, что снижает требования к вычислительным ресурсам и ускоряет обучение и inference моделей. Это особенно актуально при работе с большими наборами данных или при использовании вычислительно-интенсивных архитектур.

В целом, правильное применение масштабирования и обрезки изображений является ключевым фактором, влияющим на эффективность обучения моделей компьютерного зрения. Эти операции позволяют подготовить данные, обеспечивая согласованность, фокус на релевантных областях и оптимизацию размера входных данных.

Для масштабирования и нормализации был написан скрипт на Python. Нормализация производилась методом минимального-максимального. Масштабирование производилось к размеру 600x300, для облегчения обучения модели. Пример нормализованного и ненормализованного изображения можно увидеть на рисунках 1 и 2.

Рис 1. Изображение после нормализации

Рис 1. Изображение до нормализации

## **Разбиение датасета**

После того, как мы обработали входные изображения, необходимо разбить их на несколько частей: обучающую, валидационную и тестовую. Размером они должны относиться к друг другу как 65/20/15. Расскажем о применении каждой из них.

Тренировочная (обучающая) выборка играет ключевую роль в процессе обучения моделей машинного обучения. Именно на этой выборке происходит обучение модели, где она "изучает" закономерности, присущие данным, и настраивает свои внутренние параметры для максимально точного предсказания результатов. Размер тренировочной выборки должен быть достаточно большим, чтобы модель могла эффективно обобщать и извлекать характерные особенности данных. Чем больше тренировочная выборка, тем, как правило, качественнее обучается модель, так как она имеет больше примеров для изучения. Однако существует баланс, при котором дальнейшее увеличение размера выборки уже не приносит существенного улучшения производительности модели. Важно, чтобы тренировочная выборка содержала разнообразные примеры, характерные для решаемой задачи. Выборка должна быть репрезентативной, то есть отражать реальное распределение данных, с которыми модель будет сталкиваться в production-режиме. Если тренировочная выборка будет смещенной или не будет охватывать всё многообразие возможных ситуаций, это негативно отразится на способности модели обобщать. Существуют различные способы формирования тренировочной выборки, такие как случайная выборка из всего набора данных, стратифицированная выборка для сохранения пропорций различных классов, целенаправленный отбор примеров для более равномерного представления всех классов, а также аугментация данных (например, поворот, масштабирование изображений) для увеличения размера и разнообразия выборки. Грамотное формирование тренировочной выборки является ключом к успешному обучению высокопроизводительных моделей машинного обучения.

Валидационная выборка является важным элементом процесса обучения моделей машинного обучения. Эта выборка используется для оценки производительности модели в процессе обучения. По мере обучения модели на тренировочной выборке, валидационная выборка применяется для измерения метрик качества, таких как точность, F1-score и другие. Результаты на валидационной выборке помогают выявить переобучение модели и настроить ее гиперпараметры. Размер валидационной выборки обычно меньше, чем тренировочной, но достаточен для надежной оценки производительности. Важно, чтобы валидационная выборка была репрезентативной и включала разнообразные примеры, характерные для решаемой задачи. Это позволяет получить достоверную оценку того, как модель будет работать на новых, ранее не встречавшихся данных. Использование валидационной выборки помогает предотвратить переобучение модели на тренировочной выборке. Если модель показывает высокие результаты на тренировочных данных, но низкие на валидационной выборке, это может быть признаком переобучения. В таком случае необходимо внести изменения в архитектуру модели, регуляризацию или другие гиперпараметры для улучшения генерализации. Таким образом, валидационная выборка играет ключевую роль в процессе разработки моделей машинного обучения, позволяя настраивать их характеристики и оценивать производительность на новых данных, не используемых для обучения.

Тестовая выборка является заключительным этапом оценки производительности модели машинного обучения. После того, как модель обучена на тренировочной выборке и ее гиперпараметры оптимизированы с помощью валидационной выборки, тестовая выборка используется для окончательной проверки качества модели. Тестовая выборка должна быть полностью независимой от тренировочной и валидационной выборок. Она представляет собой репрезентативную выборку данных, которые модель никогда не видела в процессе обучения. Это позволяет получить достоверную оценку того, как модель будет работать на новых, неизвестных ей ранее примерах. Размер тестовой выборки обычно меньше, чем тренировочной, но достаточен для надежной оценки производительности модели. Крайне важно, чтобы тестовая выборка была сбалансирована и охватывала все аспекты и особенности данных, с которыми модель может столкнуться в реальных условиях. Результаты, полученные на тестовой выборке, являются наиболее важными для оценки качества модели. Именно на этих данных оценивается, насколько хорошо модель генерализует и сможет успешно применяться для решения практических задач. Только после прохождения испытаний на тестовой выборке модель может быть развернута в production-среде. Таким образом, тестовая выборка играет ключевую роль в финальной оценке и валидации моделей машинного обучения, гарантируя их производительность и надежность на новых, неизвестных данных.

## **Разметка данных**

После того, как мы обработали данные, разбили их на несколько частей, необходимо разметить данные. Это один из самых долгих и важных этапов при обучении модели. От качества работы на этом этапе зависит то, насколько точные предсказания будет давать модель.

Процесс разметки заключается в том, что люди вручную выделяют и аннотируют интересующие объекты на изображениях. Это может включать обведение контуров объектов, присвоение им соответствующих меток или классов, а также указание дополнительной информации, такой как тип объекта, его размеры, местоположение и другие релевантные характеристики.

Для задачи поиска карьеров, вырубок и пашен разметка может включать следующие этапы:

1. Выделение границ карьеров, вырубок и пашен на изображениях. Это может осуществляться с помощью инструментов для векторной или полигональной разметки.

2. Присвоение соответствующих меток (классов) каждому выделенному объекту, например: "карьер", "вырубка", "пашня".

3. При необходимости, добавление дополнительной информации, такой как площадь объекта, его возраст, тип растительности и другие характеристики.

Разметка данных может осуществляться в автоматическом и полуавтоматическом режиме. Такие способы разметки изображений становятся все более популярными, особенно при работе с большими наборами данных, где ручная разметка может быть слишком трудоемкой и дорогостоящей.

Автоматическая разметка предполагает использование алгоритмов машинного обучения для автоматического выделения и аннотирования интересующих объектов на изображениях. Это может быть реализовано с помощью предварительно обученных моделей компьютерного зрения, таких как сверточные нейронные сети (CNN) или детекторы объектов. Такие модели способны распознавать и локализовать целевые объекты (например, карьеры, вырубки, пашни) на новых изображениях без участия человека. Преимущества автоматической разметки:

* Значительное ускорение процесса разметки, особенно для больших наборов данных
* Последовательность и воспроизводимость результатов
* Возможность масштабирования и обработки больших объемов данных

Однако, автоматическая разметка может иметь ограничения в точности и полноте выявления объектов, особенно при работе с новыми или нетривиальными типами объектов. Здесь на помощь приходит полуавтоматическая разметка.

Полуавтоматическая разметка сочетает в себе возможности автоматических алгоритмов и ручного вмешательства человека-эксперта. Она может включать следующие этапы: предварительная автоматическая разметка изображений при помощи обученной модели, ручная проверка и корректировка результатов автоматической разметки экспертами, итеративное улучшение модели на основе откорректированных данных.

Полуавтоматический подход позволяет повысить точность разметки, устранить ошибки автоматических алгоритмов и обеспечить высокое качество аннотаций. Эксперты могут вносить уточнения, исправлять ошибки, добавлять дополнительную информацию. В свою очередь, это помогает улучшать автоматические модели для последующих итераций.

Качественная разметка данных является ключевым фактором, определяющим успех последующего обучения модели. Она должна быть максимально точной, последовательной и полной, чтобы модель могла эффективно обучаться распознавать и классифицировать целевые объекты.

Важно также учитывать, что разметка - трудоемкий и время-затратный процесс, особенно при работе с большими наборами данных. Для повышения эффективности разметки можно использовать специализированное программное обеспечение, а также привлекать команду экспертов. В целом, тщательная и последовательная разметка изображений является фундаментом для успешного обучения моделей поиска карьеров, вырубок и пашен. Она обеспечивает качество и достоверность обучающих данных, что напрямую влияет на производительность и обобщающую способность конечной модели.

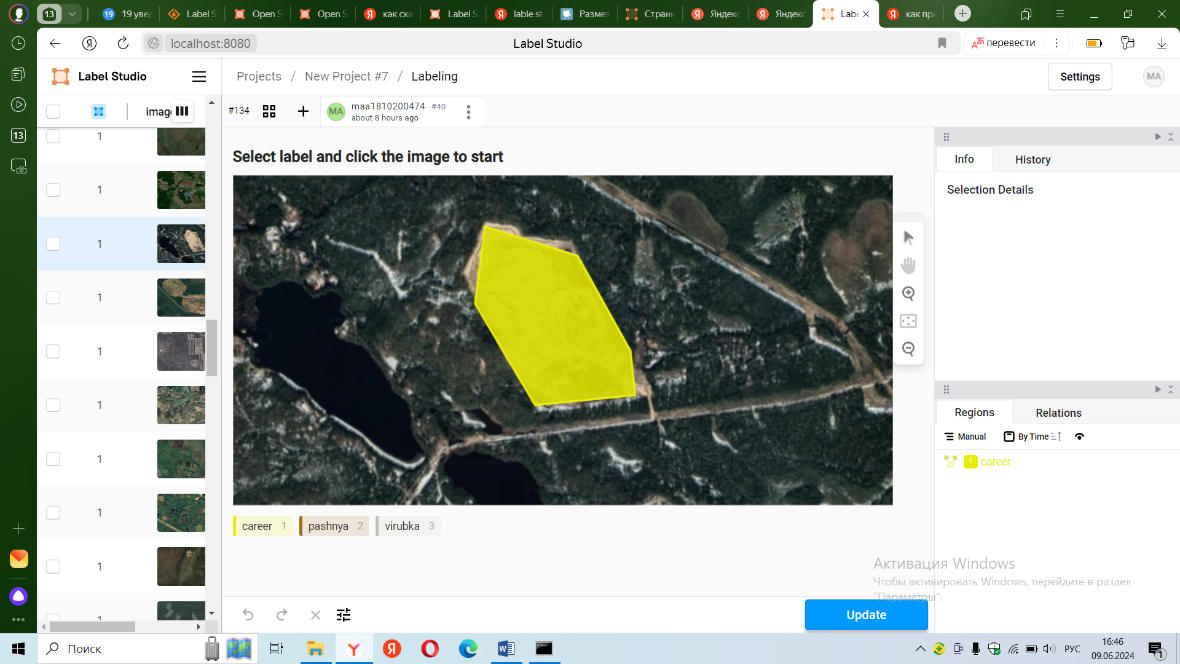
Мы производили разметку в программе Label-studio. В ней можно размечать изображение полигонами, прямоугольниками и даже зонами. Кроме того, в этой программе можно экспортировать результат работы во множество форматов, что позволяет обучать множество моделей, которым нужны свои данные. Для обозначения сущностей мы использовали полигоны. Ими очень удобно отмечать неровные границы карьеров, пашен и вырубок. На рисунке 1 можно увидеть, как происходит процесс разметки снимка.

Рис 1. Рабочее окно Label-studio

## **Аугментация данных**

После обработки и разметки еще одним важным этапом является аугментация данных. Аугментация данных - это техника, позволяющая расширять и разнообразить обучающий набор данных путем применения различных преобразований к исходным изображениям. Это особенно актуально для задач машинного обучения, где объем данных ограничен, а модели нуждаются в большом количестве примеров для эффективного обучения. В контексте поиска карьеров, вырубок и пашен, аугментация данных может включать следующие преобразования:

### **Геометрические преобразования**

Геометрические трансформации являются одними из ключевых техник аугментации данных в задачах компьютерного зрения, в том числе при работе с поиском карьеров, вырубок и пашен. Масштабирование подразумевает увеличение или уменьшение размера объектов на изображении, при этом можно использовать случайные коэффициенты масштабирования в определенном диапазоне. Реализуется данное преобразование с помощью библиотек компьютерного зрения, таких как OpenCV, Pillow, TensorFlow.

Поворот заключается в вращении изображения на произвольный угол, используя случайные углы поворота в диапазоне 0-360 градусов. Реализуется с использованием матричных преобразований в библиотеках компьютерного зрения.

Сдвиг предполагает перемещение объектов по горизонтали и/или вертикали, используя случайные смещения в пределах определенной доли размера изображения. Реализуется также с помощью матричных преобразований.

Отражение включает в себя горизонтальное, вертикальное или зеркальное отражение изображения, при случайном выборе типа отражения. Реализуется с использованием библиотек компьютерного зрения.

Преимущества геометрических трансформаций заключаются в увеличении размера обучающего набора данных, повышении устойчивости модели к пространственным искажениям, улучшении обобщающей способности и снижении переобучения, а также возможности создания сбалансированных наборов данных. Кроме того, их реализация достаточно проста с использованием библиотек компьютерного зрения.

Правильный выбор диапазонов и комбинаций геометрических трансформаций позволяет эффективно расширять набор данных и тренировать модели, повышая их устойчивость к различным искажениям.

### **Фотометрические преобразования.**

Фотометрические преобразования также являются важной частью аугментации данных в задачах компьютерного зрения. Эти техники позволяют изменять различные фотометрические характеристики изображений, тем самым повышая разнообразие и устойчивость обучающих данных.

Изменение яркости, контрастности и насыщенности изображений позволяет моделировать различные условия освещения и параметры съемки. Для этого используются случайные коэффициенты, которые применяются к исходным пикселям. Подобные преобразования реализуются с помощью стандартных библиотек компьютерного зрения.

Добавление различных типов шума, таких как гауссовский, соль-перец и другие, позволяет имитировать артефакты, возникающие при захвате изображений в реальных условиях. Это помогает сделать модели более устойчивыми к шумам и искажениям.

Композитные преобразования предполагают более сложные манипуляции с изображениями, такие как наложение нескольких изображений, вырезание и вставка объектов из одного изображения в другое, а также генерацию новых синтетических изображений. Эти техники позволяют создавать еще более разнообразные и реалистичные данные для обучения.

Основные преимущества фотометрических преобразований заключаются в повышении устойчивости моделей к различным типам искажений, связанных с условиями съемки, освещением и качеством изображений. Так же, как и с геометрическими трансформациями, правильный выбор диапазонов и комбинаций фотометрических преобразований позволяет эффективно расширять набор данных и тренировать модели, делая их более универсальными и устойчивыми к реальным условиям.

### **Композитные преобразования.**

Композитные преобразования являются одним из наиболее сложных и мощных видов аугментации данных в задачах компьютерного зрения. Эти техники предполагают более сложные манипуляции с изображениями, направленные на создание новых синтетических данных, существенно расширяющих и разнообразящих обучающие наборы.

Одним из ключевых композитных преобразований является наложение нескольких изображений друг на друга. Это позволяет генерировать сложные сцены, сочетающие различные объекты и фоны. Для реализации такого подхода могут использоваться методы смешивания изображений с помощью альфа-каналов, наложения объектов с последующей маскировкой фона и другие техники.

Другим важным типом композитных преобразований является вырезание и вставка объектов из одного изображения в другое. Это дает возможность создавать нетривиальные композиции, комбинируя различные элементы сцен. Для этого требуется точное выделение объектов, сегментация, а также корректная вставка с учетом перспективы, освещения и других визуальных особенностей.

Более сложным подходом является генерация полностью новых синтетических изображений. Это может достигаться с помощью методов генеративных состязательных сетей (GAN), позволяющих создавать реалистичные изображения на основе обучающих данных. Такие техники позволяют получать совершенно уникальные примеры, значительно расширяющие разнообразие обучающих наборов.

Основные преимущества композитных преобразований заключаются в их способности создавать сложные, реалистичные и разнообразные синтетические данные. Это особенно важно в задачах, где натурные данные труднодоступны или недостаточны, например, при поиске карьеров, вырубок и пашен на спутниковых изображениях. Грамотное применение композитных преобразований, в сочетании с другими техниками аугментации, позволяет значительно повысить устойчивость и обобщающую способность моделей компьютерного зрения, делая их более эффективными в реальных условиях.

Применение всех этих техник позволяет значительно увеличить размер и разнообразие обучающего набора данных, не требуя сбора и разметки дополнительных изображений вручную. Аугментация сильно улучшает качество работы модели, повышая обобщающую способность модели и устойчивость к различным искажениям, снижая риск переобучения на ограниченном наборе данных, давая возможность создания сбалансированных наборов данных с равномерным представлением классов, ускоряя и удешевляя процесс сбора и разметки данных. При использовании аугментации важно подбирать наиболее подходящие преобразования, которые будут отражать реалистичные изменения, с которыми может столкнуться модель в реальном мире. Это позволит эффективно тренировать модели поиска карьеров, вырубок и пашен на расширенных и разнообразных наборах данных.

Наши данные было принято аугментировать при помощи геометрических преобразований. Каждый снимок был отражен относительно его побочной диагонали. Так как аугментация происходила после разметки, аннотации изменялись вмест е с изображениями. Для этого был написан скрипт на языке Python, который при помощи библиотек Numpy и OpenCV и афинных преобразований отражал изображения. На рисунке 1 представлен оригинальный снимок. На рисунке 2 представлен он же, только отраженный

## **Обновление данных**

Рис 2. Отраженный снимок

Рис 1. Оригинальный снимок

Обновление и пополнение набора данных является важным аспектом при работе с любой моделью. По мере развития модели и появления новых данных, необходимо регулярно обновлять и расширять исходный набор данных для поддержания актуальности и повышения эффективности модели.

Необходимо организовать процесс регулярного сбора и интеграции новых спутниковых снимков, охватывающих различные географические регионы, времена года и типы ландшафтов. Это позволит расширить разнообразие данных и улучшить способность модели обобщать на новых, ранее неизвестных примерах. Помимо самих снимков, важно обновлять и аннотации (метки) для пашен, вырубок и карьеров. Это можно делать вручную или с использованием полуавтоматических методов, таких как краудсорсинг или интеграция с другими источниками данных. Точные и достоверные аннотации являются ключом к успешному обучению модели.

Перед интеграцией новых данных в существующий набор необходимо проверять их качество. Это включает в себя проверку на наличие ошибок, артефактов, несоответствий в аннотациях и другие потенциальные проблемы. Только высококачественные данные должны включаться в набор для обучения. При добавлении новых данных важно следить за сбалансированностью количества примеров для каждого класса (пашня, вырубка, карьер). Если происходит дисбаланс, необходимо предпринять меры, такие как подсэмплирование или аугментация, для восстановления равномерного распределения классов.

Для эффективного управления обновлениями набора данных рекомендуется использовать систему версионирования, позволяющую отслеживать изменения и при необходимости возвращаться к предыдущим версиям. Кроме того, важно регулярно архивировать данные для обеспечения сохранности и возможности восстановления. По мере роста объема данных целесообразно автоматизировать многие этапы обновления, такие как сбор новых снимков, предварительная обработка, аннотирование, балансировка и интеграция в существующий набор. Это повысит эффективность и скорость обновления данных.

Регулярное обновление и расширение набора данных позволит поддерживать актуальность и эффективность модели распознавания пашен, вырубок и карьеров, обеспечивая её способность адаптироваться к изменениям в реальном мире.