Применение методов машинного обучения для оценки количества биомассы по данным спутниковых снимков Sentinel-1 и Sentinel-2

Автор: Денис Петренко
Научный руководитель: Алексей Озерин

Описание тривиального решения

В качестве тривильного решения будем использовать среднее значение целевой переменной на тренировочном датасете. Данная модель обеспечивает значение метрики RMSE на уровне 55.145504

Описание baseline модели

Для построения baseline модели использовались спутниковые снимки, полученные только с Sentinel-1(радарные данные в четырех каналах).
В качестве baseline решения используется модель на базе encoder-decoder архитектуры. Выбор такой архитектуры обусловлен тем фактом, что задачу можно рассматривать как задачу регрессии на последовательности значений взятых для каждого отдельного пикселя.

Детали обучения

Предложенная модель была обучена следующим образом:

Параметр	**Значение**
optimizer	SGD(lr=0.1, momentum=0.9)
Scheduler	StepLR(step_size=6, gamma=0.5)
n_epoch	10

Обучение модели проводилось на семплированных из тренировочной выборке данных по каждому пикселю. Из данных каждого участка выбиралась 1 000 пикселей, которые использовались для дальнейшего обучения.

Результаты и сравнение с тривиальным решением

Baseline модель позволяет получить значение метрики RMSE на уровне 54.078785 на валидации. Безусловно, прирост качества по сравнению с тривиальным решением незначительный, однако следует принять во внимание, что этот результат получен без предобработки данных, генерации дополнительных признаков, только на данных снимков Sentinel-1 без подбора параметров модели/обучения.
Ожидается, что обогащение признакового пространства данными спутника Sentinel-2 и создание дополнительных признаков позволит значительно улучшить качество модели.

Архитектура проекта

Предполагается разработка модели глубинного обучения с помощью библиотеки pytorch. В качестве интерфейса планируется реализация REST API сервиса на основе фреймворка fastapi, который будет помещен в Docker контейнер.

Планируемые эксперименты

В дальнейшем планируется усовершенствовать пайплайн предобработки данных:

- Добавить данные Sentinel-2. Снимки этого спутника содержат большое количество полезных данных, но имеют некоторое количество недостатков: пропуски и наличие облаков, что требует дополнительной обработки.
- Сгенерировать признаки в дополнение к существующим изображениям

Также требуют подбора параметры обучения модели (размер батча, параметры оптимизатора и прочее).
Отдельный интерес в экспериментах представляет усовершенствование непосредственно архитектуры нейронной сети, в частности, добавление эмбеддингов для периодов наблюдаемых данных.