Задание 1

Практическая работа

Тема:

Формирование данных для машинного обучения в формате NumPy на основе датасета Semantic3D

Цель работы:

Научиться:

- Загружать облака точек из датасета Semantic3D;
- Формировать единый массив данных NumPy (таблицу), объединяющий признаки и метки классов;
- Выполнять базовую предобработку: нормализацию координат и признаков;
- Подготавливать данные для подачи в модели машинного обучения (например PyTorch).

Задача:

Используя один файл из датасета Semantic3D (например, `bildstein_station1_xyz_intensity_rgb.label`), сформировать единый массив данных `dataset`, где каждая строка соответствует одной точке, а столбцы представляют:

X	Y	Z	R	G	B	intensity	label

Реализовать выше изложенное в виде отдельной функции.

Исходные данные:

- Файл формата `.label` и `.txt` (разделённый пробелами)

- Выходной файл содержит: `X Y Z R G B intensity label`

Инструкции к выполнению:

1. Загрузка и разбор данных:

- Считать данные с помощью `numpy.loadtxt()` или `np.genfromtxt()`;
- Выделить все нужные признаки;
- Преобразовать цвета R, G, B в диапазон [0, 1].

2. Формирование таблицы:

- Объединить все признаки и метку в одну таблицу `dataset` с помощью `np.hstack()`;
- Проверить размерность и типы данных.

3. Предобработка:

- Выполнить нормализацию координат X, Y, Z (например, центрирование и масштабирование);
- (по желанию) Нормализовать интенсивность.

4. Сохранение данных:

- Сохранить таблицу `dataset` в формате `.npy`:

```
np.save('semantic3d_dataset.npy', dataset)
```

- Предусмотреть сохранение в форматах `.txt` и `.h5`

5. Реализовать функцию визуализации распределения меток по файлу

- Построить простую визуализацию распределения меток (в виде гистограммы).

Пример структуры таблицы (`dataset`) после обработки:

```
dataset = np.array([
    [0.132, -0.532, 0.210, 0.75, 0.65, 0.60, 0.42, 2],
    [0.145, -0.525, 0.211, 0.76, 0.64, 0.59, 0.41, 2],
```

```
    ...
])
```

Что нужно сдать:

1. Готовый Python-скрипт или Jupyter Notebook;
2. Файл `semantic3d_dataset.npy`;
3. Скриншот/вывод содержимого первых 5 строк массива;
4. Ответы на контрольные вопросы.

Контрольные вопросы:

1. **Дайте определение датасету Semantic3D. Каковы его основные характеристики и для каких задач он предназначен?**
2. **Чем Semantic3D принципиально отличается от датасетов для 2D-компьютерного зрения (например, ImageNet) и других 3D-датасетов для помещений (таких как S3DIS)?**
3. **Опишите состав и структуру датасета. На какие множества (train, test и т.д.) он разделен и какова их цель?**
4. **Каким способом были получены данные в Semantic3D? Опишите технологию и ее влияние на характеристики облаков точек (плотность, шум, масштаб).**
5. **Перечислите и охарактеризуйте 8 семантических классов в датасете. Приведите примеры объектов для каждого класса.**
6. **Каковы основные проблемы, связанные с разметкой данных в таком крупномасштабном датасете? (Например, проблема "шума" или "артефактов").**
7. **Назовите основные технические проблемы, возникающие при работе с облаками точек такого объема (миллиарды точек).**
8. **Что такое "неравномерная плотность" точек в контексте LiDAR-данных? К каким проблемам при обучении моделей она может привести?**
9. **Опишите проблему несбалансированности классов в Semantic3D. Как эта проблема влияет на процесс обучения и метрики оценки?**
10. **Какие основные этапы предобработки данных необходимы перед использованием Semantic3D для обучения моделей? (Нормализация, даунсэмплинг и т.д.)**