Task 1. Natural Language Processing. Named entity recognition

The first task was to train an NER model to identify mountain names in text. We needed to either find or create a suitable dataset ourselves. I used ChatGPT to generate a list of sentences featuring mountain names. In total, I managed to gather just under 1,000 sentences, which is relatively small for tasks like these. Because of this limited dataset, I decided to choose a lightweight, BERT-based model that would train quickly due to its reduced number of parameters, yet still perform well on this moderately simple task. I opted for DistilBERT, a smaller model that holds up reasonably well compared to classic BERT.

After preparing the data, I trained the model and then tested it with two self-composed sentences. One sentence had clear context and was similar to the training data generated by ChatGPT, while the other had a more nuanced context, though still recognizable. The model successfully identified the mountain name in the sentence similar to the training data, but failed to do so in the second, more subtle sentence. This suggests that the main issue lies with the dataset: I couldn't find a large, diverse collection of mountain-related texts, so the model didn’t learn to generalize well to intermediate contexts.

Possible improvements: I believe the model architecture itself is sufficient for this task; its capacity should be adequate. I would focus on data generation, trying to create or locate a more varied text source. The dataset size is also crucial—gathering a dataset with 10,000+ examples would likely provide the stability needed for reliable performance.

Task 2. Computer vision. Sentinel-2 image matching

Во втором задании надо было создать алгоритм для поиска и сопоставления на паре снимков со спутников Sentinel-2 точек, которые соответствовали бы одинаковым точкам на планете. Сопоставить точки не сложная задача, ведь в датасете были данные, которые соответствовали одному участку планеты, поэтому я воспринял эту задачу так: надо было найти такие соответствующие точки, которые являлись бы keypoints для обоих снимков. Такой подход может быть полезен, если данные спутника слегка неправильно размечены или чтобы использовать этот подход как часть другого алгоритма.
Раз датасет посвящен деревьям, я решил для себя поставить задачу выделять keypoints по большей части связанные с растительностью. Спутники Santinel2 снимают сразу во многих спектрах, и некоторые комбинации из них можно использовать для специфических задач. В моем случае, оказалось неплохой идеей использовать комбинация зеленого, красного и инфракрасного каналов, этот набор как раз и используют для анализа растительности.
Для этой задачи я решил использовать мощную модель LoFIR основанную на трансформерах, они инвариантны относительно сезонных изменений и отлично нашли соответствующие keypoints на паре снимков, которые я предварительно разбил на

меньшие размеры для более удобной работы, и усредних данные по спектрам, что все равно не повлияло на высокую точность модели.
Я думаю что тут подошла бы многоканальная модель, которая выделял бы еще более интересные keypoints и в котором не пришлось бы усреднять по канал.