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

The second task was to create an algorithm to locate and match points on pairs of Sentinel-2 satellite images, identifying the same physical points on the planet. Matching points wasn’t particularly challenging since the dataset contained data from the same region. So, I interpreted the task as finding corresponding points that would serve as keypoints on both images. This approach can be useful if satellite data is slightly misaligned or if it’s part of a larger algorithm.

Given that the dataset focuses on trees, I set the goal of selecting keypoints primarily related to vegetation. Sentinel-2 captures images in multiple spectral bands, and certain combinations can be used for specific analyses. In my case, using a combination of green, red, and infrared channels proved to be effective, as this set is commonly used for vegetation analysis.

For this task, I opted to use the powerful LoFTR model, which is transformer-based. Transformers are robust to seasonal changes, and LoFTR successfully identified corresponding keypoints on the image pairs, which I had split into smaller patches for easier processing. I averaged the spectral data across channels, which did not affect the model’s high accuracy.

I believe a true multi-channel model could be even more effective, allowing for the identification of more distinctive keypoints without needing to average the channels.