Exploring Planet-Scale Image Geolocation with PIGEON

Hoffmann Szabolcs, Project Work 2025/26 I.

1 Introduction and Aims

The aim of my project is to understand how modern AI systems can infer the geographic location of an image using only visual information, and to explore how such systems could be improved.

As a concrete case study, I focused on the PIGEON / PIGEOTTO geolocation system, which uses a CLIP-based vision encoder, semantic geocells, and a clustering-based refinement step to predict latitude–longitude coordinates from single images or 4-image panoramas. Instead of directly regressing coordinates, PIGEON formulates geolocation as a classification problem over carefully constructed geographic cells, and then refines the prediction using nearest-neighbour style retrieval in embedding space.

The main goals of the project were the following. First, I wished to obtain a clear conceptual picture of PIGEON's overall architecture. Second, I aimed to understand the role of semantic geocells, haversine-based label smoothing, and the OPTICS+Voronoi refinement pipeline. I wanted to compare the authors' coordinate-refinement strategy with my own alternative based on weighted barycentric coordinates over training points inside the top-K geocells. Finally, I set out to outline future directions in which a PIGEON-style system could be combined with a multimodal ChatGPT-like model capable of reading street signs and other text in the scene.

2 Core Ideas of PIGEON

Earlier work relied on artificial grids whose boundaries had little relation to real-world geography. In contrast, PIGEON starts from administrative polygons (countries and subregions) and merges small regions until each cell has a sufficient number of training examples. Extremely dense urban areas are further subdivided by clustering training locations in CLIP embedding space (using OPTICS) and applying Voronoi tessellation to obtain smaller contiguous subcells. The resulting label space respects political borders where possible, while achieving high spatial resolution exactly where the visual and data complexity is greatest.

To train the geocell classifier, PIGEON replaces one-hot labels with a distance-aware target distribution. For each training image, the true coordinates are compared to every cell centroid via the haversine distance, and these distances are transformed into probabilities so that nearby cells receive non-zero mass and very distant cells are suppressed. This scheme formalises the idea that predicting a neighbouring cell is less severe than predicting a faraway region, and it effectively builds a hierarchy of geographic resolutions into a single loss function. The classifier is trained on top of a CLIP vision encoder that has been further adapted with synthetic geographic captions, encouraging the embedding space to reflect climate, region and other geo-related regularities.

Beyond cell-level prediction, PIGEON adds a refinement stage that operates directly in CLIP embedding space. Within each geocell, training locations are clustered and represented by prototype embeddings. At inference time, the model first selects the most probable geocells, then compares the query embedding to prototypes inside this restricted set, and finally chooses the single training location whose embedding is closest to the query. The output coordinates are those of this nearest neighbour. In effect, the parametric network narrows the search to a plausible region, while the refinement layer behaves like a nearest-neighbour retrieval mechanism that exploits the fine-grained structure of the training data.

3 My Refinement Idea: Weighted Barycentric Coordinates in Top-K Geocells

Studying this refinement layer led me to a natural extension that might reduce noise and make better use of nearby examples. The idea is to replace the "pick a single nearest neighbour" step with a smooth *barycentric averaging* of several neighbours.

3.1 Method Description

Studying this refinement layer led me to a natural extension that might reduce noise and make better use of nearby examples. The idea is to replace the final "pick a single nearest neighbour" step with a smooth barycentric averaging of several neighbours.

In words, rather than outputting the coordinates of a single closest training point, the model would compute a weighted barycentric combination of multiple training locations inside the top-K geocells. The procedure can be described as follows. First, the geocell classifier is used to obtain the top-K geocells for the query image. Only training locations lying inside these top-K cells are considered in subsequent steps. In CLIP embedding space, the M nearest training images to the query (for example, $M = 5$ or $M = 10$) are then identified.

For these M neighbours with coordinates (λ_i, ϕ_i) and embedding distances d_i, I define weights

$$w_i = \frac{\exp(-\alpha d_i)}{\sum_{j=1}^{M} \exp(-\alpha d_j)}, \quad \alpha > 0.$$

The final coordinate prediction is then given by the convex combination

$$\hat{\lambda} = \sum_{i=1}^{M} w_i \lambda_i, \qquad \hat{\phi} = \sum_{i=1}^{M} w_i \phi_i.$$

3.2 Intuition and Potential Benefits

This approach has several intuitive advantages. The top-K geocells still act as a spatial prior, but within them the model no longer relies on a single training point. The barycentric combination smooths the piecewise-constant behaviour characteristic of nearest-neighbour predictions and may therefore reduce sensitivity to outliers or mislabelled images. From a geometric viewpoint, the method can be interpreted as a "soft" Voronoi assignment: instead of snapping to the single closest site, the prediction blends several sites according to their similarity. Mathematically, this idea is close to kernel regression in embedding space. It could be imple- mented as a relatively small modification on top of the existing PIGEON refinement layer and evaluated using the same metrics as in the original paper, such as median error distance and the proportion of predictions within a given radius.

4 Future Directions: Text-Aware Geo-AI

While PIGEON is very strong at exploiting visual cues such as road markings, vegetation, and architectural style, it does not explicitly read text in the image. In the current pipeline, CLIP is used purely as a vision encoder; the model does not interpret street names, shop signs, route numbers or other textual content as linguistic information. However, in many real-world scenes, text provides highly diagnostic information about location.

A promising future direction is therefore to combine a PIGEON-style visual geolocator with a multimodal ChatGPT-like model that is capable of optical character recognition (OCR) and textual reasoning. Such a model could extract text from street signs, shop names and license plates; detect the language and script (for example, Cyrillic versus Latin, or simplified versus traditional Chinese); and use world knowledge about city names, local brands, and road numbering conventions. These textual observations could then be turned into soft constraints on the likely region, country or even a particular city.

5 Conclusion

During this semester I moved from a basic curiosity about "how GeoGuessr AIs work" to a structured understanding of a modern planet-scale geolocation pipeline.

In the next phase of my independent project, I plan to experiment with simple implementations of the barycentric refinement idea, and to sketch prototypical fusion schemes between visual geocell predictions and textual constraints from OCR, with the long-term goal of turning these ideas into a concrete research plan.