Summer Research Fellowship (SRF) 2024 Poster Presentation

Poster No.:A15

Geographic information analysis and prediction based on machine learning

Hanyi Xiong (BASc Applied AI) 3036126628

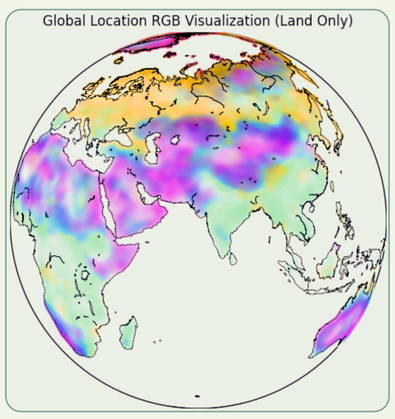
Supervisor: Prof. Kai Han, Department of Statistics and Actuarial Science

Problem Definition

Objectives: This work aims at learning global, General-Purpose **Location Embeddings**, which use **openly available**, **globally distributed satellite imagery** with their geographic coordinates to efficiently summarize the implicit characteristics of any given location and incorporate similarities over space, and to conveniently improve **analysis, generalization, and prediction** performance on geo-related and location-dependent downstream tasks.

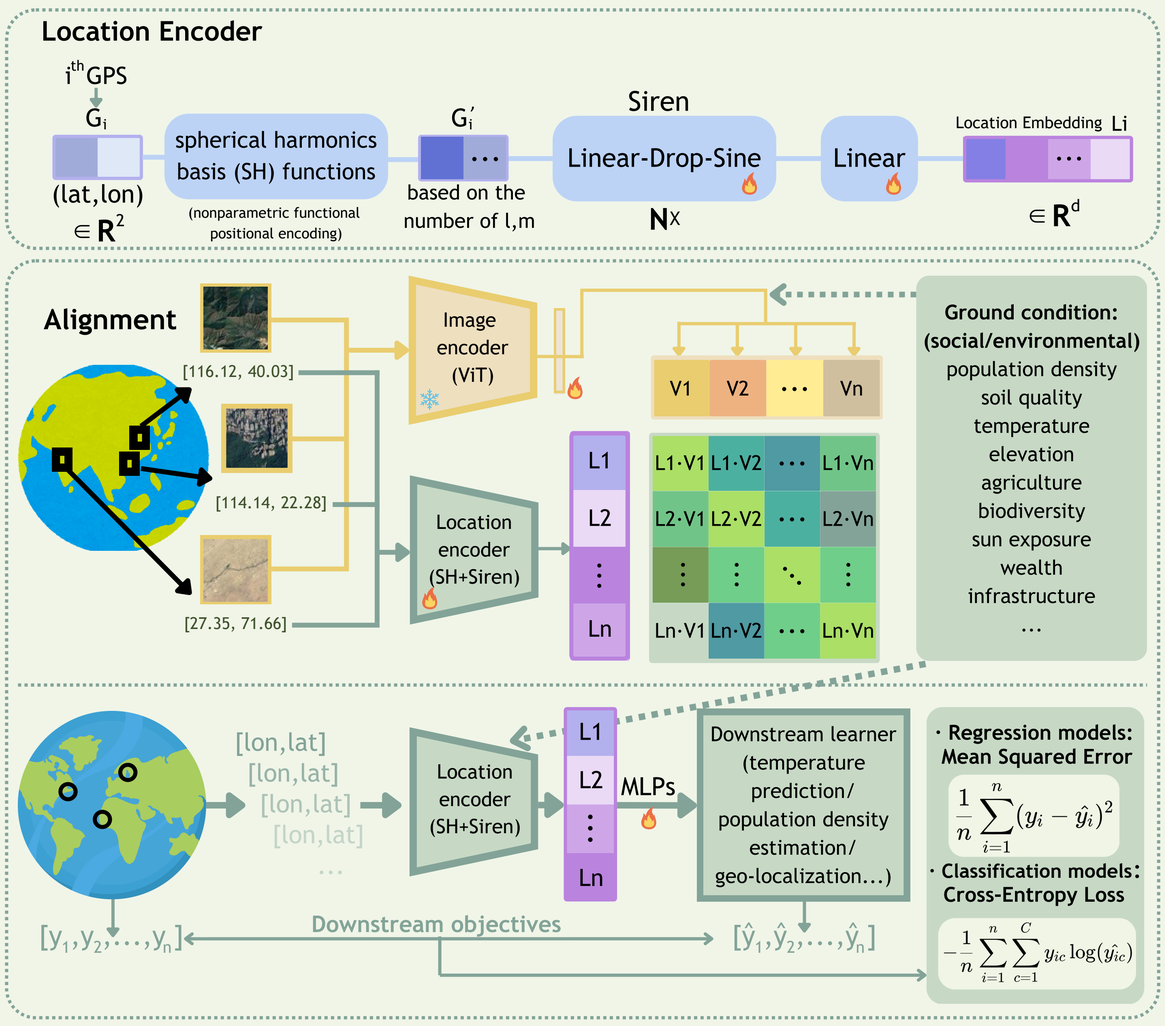
Motivations:

- The spatial patterns governing different geographic data modalities are often **complex and non-linear**.
- Patterns extracted from satellite images can describe the unique characteristics of locations, by capturing their **natural and built environment**.
- Models trained on **raw coordinates** solely rely upon spatial dependencies without considering any ground conditions, such as local elevation patterns or climate zones.
- Models trained on **full images**, while able to capture ground conditions, require expensive data preprocessing and training of large vision models.



Framework and Methods

Pineline:



Methodology:

- **Spherical Harmonics basis (SH) functions**

The traditional latitude and longitude coordinate system tends to overrepresent regions near the poles due to inherent projection **distortions.** To address these distortions and obtain a more accurate representation of the Earth's surface, we employ the Spherical Harmonics (SH) functions to encode the coordinate (λ,ϕ):

$$f(\lambda,\phi)=\sum_{l=0}^{\infty}\sum_{m=-l}^{l} w_l^m Y_l^m(\lambda,\phi) \quad \text{where} \quad Y_l^m(\lambda,\phi)=\sqrt{\frac{2l+1}{4\pi}\frac{(l-|m|)!}{(l+|m|)!}}P_l^m(\cos\lambda)e^{im\phi}$$ [3]

l and m serve as the degree(complexity/frequency) and order(oscillations in longitude) of the SH functions, ω_l^m are the coefficients that adjust each harmonic's contribution. P_l^m is the associated Legendre polynomial.

- **Alignment**

Both encoders are trained using the simple yet highly effective CLIP[2] objective, defined as follows:

$$\mathcal{L}=\frac{1}{2N}\left[\sum_{i=1}^{N}\mathcal{L}_{\text{loc}}(\mathbf{c}_i,\mathbf{I}_{1,\ldots,N})+\sum_{i=1}^{N}\mathcal{L}_{\text{img}}(\mathbf{I}_i,\mathbf{c}_{1,\ldots,N})\right]$$

This objective aligns each coordinate c_i with the corresponding image I_i while contrasting it against all other images $I_{1,\ldots,N}$. Specifically, the local alignment loss L_{Loc} is defined as:

$$\mathcal{L}_{\text{loc}}(\mathbf{c}_i,\mathbf{I}_{1,\ldots,N})=-\log\frac{\exp(\langle f_{\text{c}}(\mathbf{c}_i),f_{\text{I}}(\mathbf{I}_i)\rangle/\tau)}{\sum_{j=1}^{N}\exp(\langle f_{\text{c}}(\mathbf{c}_i),f_{\text{I}}(\mathbf{I}_j)\rangle/\tau)}$$

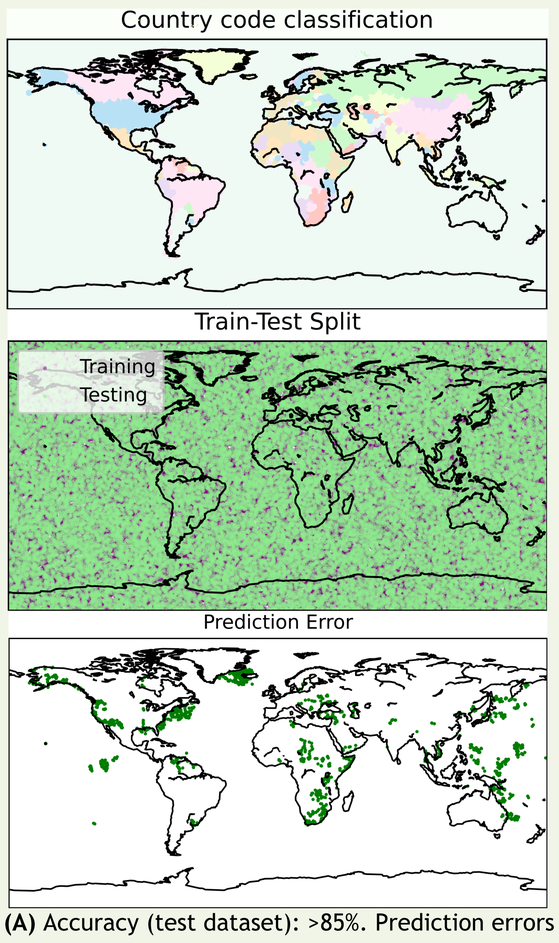
The image alignment loss L_{img} is defined analogously by swapping the roles of the coordinate and image terms. τ is a temperature parameter controlling the sharpness of the probability distribution.

References:

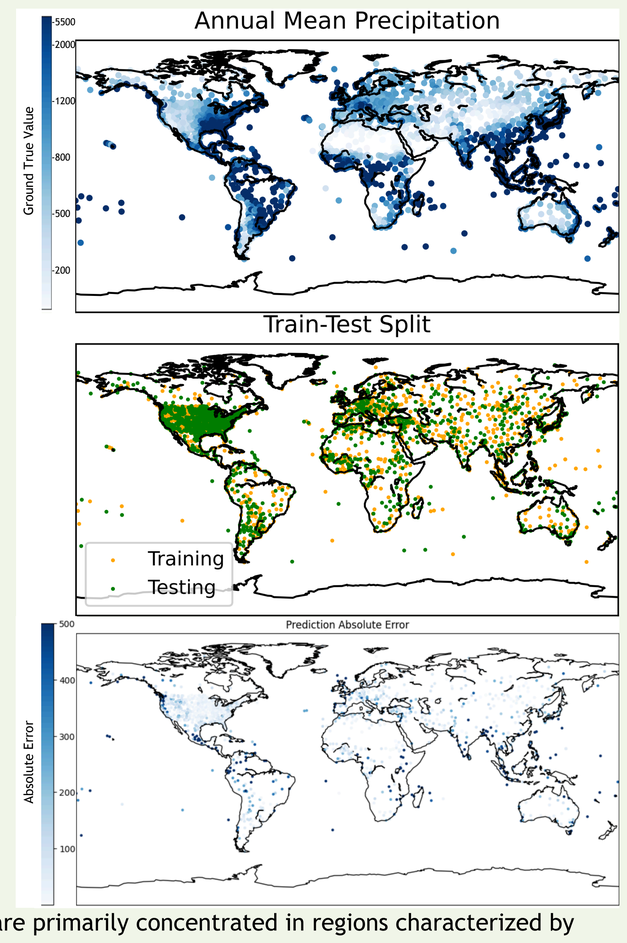
1. Klemmer, K., Rolf, E., Robinson, C., Mackey, L., & Rußwurm, M. (2023). Satclip: Global, general-purpose location embeddings with satellite imagery. arXiv preprint arXiv:2311.17179.
2. Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... & Sutskever, I. (2021, July). Learning transferable visual models from natural language supervision. In International conference on machine learning (pp. 8748-8763). PMLR.
3. Rußwurm, M., Klemmer, K., Rolf, E., Zbinden, R., & Tuia, D. (2023). Geographic location encoding with spherical harmonics and sinusoidal representation networks. arXiv preprint arXiv:2310.06743.
4. Vivanco Cepeda, V., Nayak, G. K., & Shah, M. (2024). Geoclip: Clip-inspired alignment between locations and images for effective worldwide geo-localization. Advances in Neural Information Processing Systems, 36.

Experiments and Results

(A) Classification: Country Code Prediction

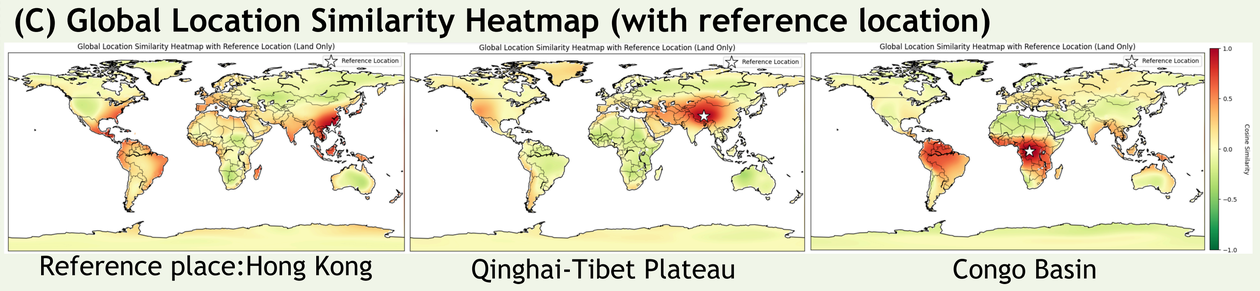


(B) Regression: Annual Mean Precipitation Prediction



- **(A)** Accuracy (test dataset): >85%. Prediction errors are primarily concentrated in regions characterized by **complex geographical and climatic transitions**, such as coastal areas and continental shelves with hybrid land-sea climates, and regions with **intricate geopolitical boundaries**, particularly in Africa and Europe.
- **(B)** The map shows the absolute error between the predicted precipitation and the ground truth. The maximum recorded ground truth is nearly 6000 mm, while the maximum prediction error is under 500 mm. Larger prediction errors are often observed in regions with **high actual precipitation values**. This pattern may result from the model's **conservative predictions**, which can lead to substantial discrepancies in areas where the actual precipitation is exceptionally high.

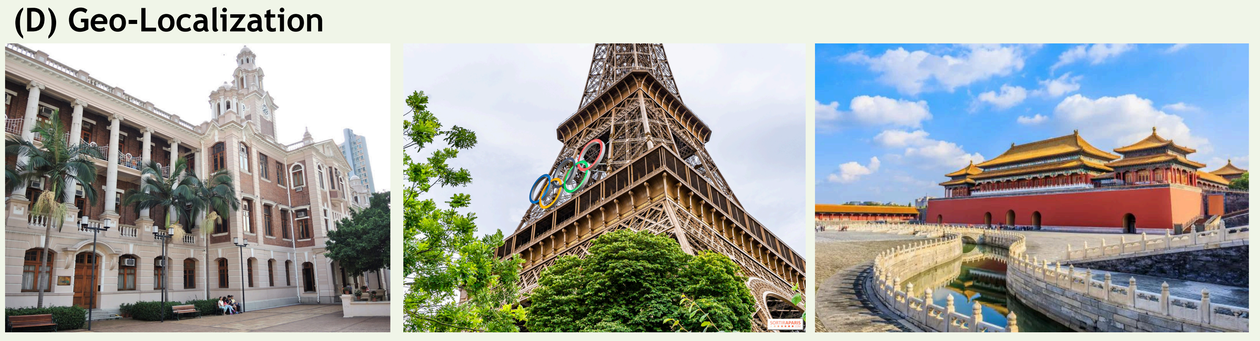
(C) Global Location Similarity Heatmap (with reference location)



Reference place:Hong Kong | Qinghai-Tibet Plateau | Congo Basin

- (C) The heatmaps displaying the cosine similarity of a set of 100,000 uniformly distributed points across the Earth's surface to three reference points—Hong Kong, the Qinghai-Tibet Plateau, and the Congo Basin—based on location embeddings generated by our trained location encoder, with warmer colors indicating higher similarity to the reference location.

(D) Geo-Localization



Predicted:(22.1983, 113.9430)
True:（22.2842，114.1378）

Predicted:(48.8593, 2.2924)
True:（48.8582, 2.2945）

Predicted:(39.9132, 116.3895)
True:（39.9166, 116.3907）

- **(D)** We utilize the fine-tuned vision encoder to encode input images, projecting them into the same dimensional space as the location embeddings. We then compute the cosine similarity between the image and location embeddings corresponding to the 100,000 points selected in (C). This process allows us to match the features of a query image against the gallery of GPS embeddings, with the most similar GPS embedding being selected as the predicted GPS coordinates.

- Conclusion: The results demonstrate that the location encoder effectively captures both **natural geographic features**, such as climate, topography, and vegetation, as well as **anthropogenic factors**, including human activities and administrative boundaries, aligning closely with established geographic and socio-political realities.

Discussion and Future Work

- Our pretraining approach focuses mainly on integrating the visual modality. However, it is feasible to develop a **multimodal context encoder** incorporating additional location-specific data modalities, such as **audio** from acoustic sensors or **text** from geolocated social media posts, to facilitate multi-source geospatial learning.
- While seasonal variations in images can be recognized, the model did not explicitly incorporate time as an embedded component within a **space-time encoder**, such as in a function of the form **f(lat, lon, time)**.
- The model's capacity to enhance its **fine-grained** discriminative ability between similar locations is constrained by the spatial scales of its pre-trained weights. This constraint is particularly evident when addressing **high-resolution or localized phenomena**, where the model may struggle to differentiate **iconic landmarks**, such as distinguishing the actual Eiffel Tower in Paris from its replicas elsewhere.