Environmental Computational Science and Earth Observation Laboratory

Bridging Remote Sensing Imagery with Natural Language via Normalizing Flows

Context

In recent years, foundation models (FMs) have transformed both computer vision and natural language processing. Pretrained vision transformers (ViTs) through masked data modelling (e.g., MAEs [1]) have enabled efficient, large-scale understanding of complex visual data, including remote sensing (RS) imagery [2]-[8]. Meanwhile, large language models (LLMs) such as the LLaMA family [9] have demonstrated remarkable performance in open-ended reasoning tasks, yet they are limited in their ability to natively process and understand visual inputs. Integrating these capabilities into conversational systems would allow users to query and interpret RS imagery in natural language, for applications in land-use monitoring, climate analysis, disaster response, and more.

Despite their potential, integrating RS-based vision FMs with LLMs is non-trivial. Typical approaches in visual question answering (VQA) require fine-tuning both visual and language models, which is computationally expensive and domain-specific. Furthermore, RS imagery presents unique challenges: heterogeneous multi-sensor data, domain-specific semantics, and geospatial alignment requirements. Existing multimodal LLMs are rarely optimized for these complexities.

In this project, a lightweight strategy is aimed to develop for bridging RS imagery with language understanding by using the existing vision FMs and LLMs without a need for fine-tuning language and vision models. The core idea is to use normalizing flows [10]-[12] to learn a bijective mapping between visual embeddings (from a pretrained MAE) and the token embedding space of an LLM (e.g., LLaMA 2 [9]) by leveraging geographic coordinates as the shared information across these two data distributions. These mapped representations from vision to the language domain serve as semantic priors, enabling an LLM to generate image-aware responses (i.e., image-grounded question answering) in a zero-shot manner.

Objectives

- Use a pretrained MAE to extract embeddings from RS images, and an LLM to obtain language tokens
- Develop a normalizing flow-based model that learns to map image embeddings to token-level language priors.

ECEO

EPFL ENAC IIE
Devis Tuia, Prof.
Rue de l'Industrie 17
Case Postale 440
CH – 1951 Sion

Phone : +41 21 69 382 83 secr.
E-mail : devis.tuia@epfl.ch, gencer.sumbul@epfl.ch

- Train a bijective transformation between the image embedding space and a language token space within the LLM's input embedding domain. The mapping is supervised via text-image aligned datasets.
- Without modifying the LLM, inject these image-conditioned embeddings as a sequence of learned soft prompts (prefix tokens). These soft prompts act as semantic priors that condition the LLM's responses.
- Enable the system to perform visual question answering (VQA) on satellite imagery without fine-tuning the LLM.
- Evaluate performance using publicly available RS-VQA benchmarks.

Possible use case

Given an RS image and a user query (e.g., "What type of land use is visible?"), the system encodes the image, maps it via normalizing flow into the LLM token space, prepends this embedding as a soft prompt to the query, and generates a natural-language answer — all without altering the LLM parameters.

Requirements

- Strong interest in machine learning and multimodal reasoning.
- Experience with PyTorch
- Familiarity with transformers, normalizing flows, and basic RS concepts is a plus.

Contact

- Prof. Devis Tuia, devis.tuia@epfl.ch
- Dr. Gencer Sümbül, gencer.sumbul@epfl.ch

References

[1] K. He, X. Chen, S. Xie, Y. Li, P. Doll´ar, and R. Girshick, "Masked autoencoders are scalable vision learners," in 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 15979–15988.

[2] D. Wang, Q. Zhang, Y. Xu, J. Zhang, B. Du, D. Tao, and L. Zhang, "RingMo: A Remote Sensing Foundation Model With Masked Image Modeling," IEEE Transactions on Geoscience and Remote Sensing (TGRS), vol. 61, pp. 1-22, Art no. 5612822, 2023.

[3] D. Wang, Q. Zhang, Y. Xu, J. Zhang, B. Du, D. Tao, and L. Zhang, "Advancing plain vision transformer toward remote sensing foundation model," IEEE Transactions on Geoscience and Remote Sensing (TGRS), vol. 61, no. 5607315, pp. 1–15, 2023.

[4] C. J. Reed, R. Gupta, S. Li, S. Brockman, C. Funk, B. Clipp, K. Keutzer, S. Candido, M. Uyttendaele, and T. Darrell, "Scale-MAE: A scale-aware masked autoencoder for multiscale geospatial representation learning," International Conference on Computer Vision (ICCV), 2023.

[5] M. Mendieta, B. Han, X. Shi, Y. Zhu, and C. Chen, "Towards Geospatial Foundation Models via Continual Pretraining," International Conference on Computer Vision (ICCV), 2023.

[6] K. Cha, J. Seo, and T. Lee, "A Billion-scale Foundation Model for Remote Sensing Images", arXiv preprint arXiv:2304.05215, 2023.

[7] F. Yao, W. Lu, H. Yang, L. Xu, C. Liu, L. Hu, H. Yu, N. Liu, C. Deng, D. Tang, C. Chen, J. Yu, X. Sun, K. Fu, "RingMo-Sense: Remote Sensing Foundation Model for Spatiotemporal Prediction via Spatiotemporal Evolution Disentangling," in IEEE Transactions on Geoscience and Remote Sensing (TGRS), vol. 61, pp. 1-21, Art no. 5620821, 2023.

[8] D. Hong, B. Zhang, X. Li, Y. Li, C. Li, J. Yao, N. Yokoya, H. Li, P. Ghamisi, X. Jia, A. Plaza, G. Paolo, J. A. Benediktsson, and J. Chanussot, “SpectralGPT: Spectral Foundation Model”, arXiv preprint arXiv:2311.07113, 2023.

[9] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, et. al, “Llama 2: Open Foundation and Fine-Tuned Chat Models,” arXiv preprint arXiv:2307.09288, 2023.

[10] L. Dinh, J. Sohl-Dickstein, and S. Bengio, “Density estimation using Real NVP,” arXiv preprint arXiv:1605.08803. 2016.

[11] G. Papamakarios, E. Nalisnick, D. J. Rezende, S. Mohamed, and B., Lakshminarayanan, “Normalizing Flows for Probabilistic Modeling and Inference,” in Journal of Machine Learning Research, 22(57):1−64, 2021.

[12] M. Tschannen, A. S. Pinto, and A. Kolesnikov, “JetFormer: An Autoregressive Generative Model of Raw Images and Text,” International Conference on Learning Representations (ICLR), 2025.

ECEO

EPFL ENAC IIE
Devis Tuia, Prof.
Rue de l’Industrie 17
Case Postale 440
CH – 1951 Sion

Phone : +41 21 69 382 83 secr.
E-mail : devis.tuia@epfl.ch, gencer.sumbul@epfl.ch