Self-Supervised Learning from Satellite Images with a Joint-Embedding Predictive Architecture

Background

Deep learning (DL) based general image representation learning (IRL) is of great interest for satellite images due to its ability to: i) model vast amounts of freely available remote sensing (RS) data through self-supervised learning, significantly decreasing the requirement for labeled data; and ii) generalize well to various RS problems as downstream tasks. Contrastive learning-based methods have initially paved the way for employing self-supervised IRL on RS images (e.g., [1], [2]). Such methods employ contrastive learning of satellite image representations with convolutional neural networks (CNNs) by maximizing agreement between two views of the same image, which are generated by data augmentation strategies. Recent studies on IRL in RS have focused on masked data modeling of satellite images, e.g., [3]-[10]. They facilitate self-supervised learning through masked autoencoders (MAEs) with vision transformers (ViTs). By reconstructing satellite images with parts masked, they perform effective IRL, i.e. they learn features that describe the visual content of the images that can be used as a starting point to tune specialized models for downstream tasks. Recent interest in MAEs for IRL of satellite images relies on two main reasons. First, in contrast to contrastive self-supervised learning, MAEs are capable of learning image representations without the application of any data augmentation strategies. This is of particular importance for satellite images since most of the data augmentation strategies are designed for natural images and their direct adaptation to satellite may not be always feasible. Second, it has been shown that MAEs in combination with ViTs can be effectively scaled into larger DL models in proportion to the amount of training data [11], [12]. However, when MAEs are utilized, the resulting image representations tend to be of a lower semantic level [13]. This prevents utilizing their full potential for many downstream tasks requiring higher-level satellite image semantics (e.g., scene classification, land-cover map generation, etc.).

Aim

As an alternative to contrastive learning and MAEs, the joint-embedding predictive architecture (I-JEPA) [13] has been recently proposed to scale well with ViTs on a large amount of data, while still learning image representations of a high semantic level without relying on hand-crafted data augmentations. This is achieved by predicting the representations of various target blocks from a single context block in the same image. Although it embodies a significant potential for self-supervised image representation learning in RS, they haven't been applied to satellite images. In this project, we aim to explore the use of I-JEPA on satellite images, specifically on image benchmark datasets such as fMoW [14]. This project consists of the application of I-JEPA pre-training on an RS image benchmark dataset and the evaluation of its effectiveness compared to MAEs for the downstream tasks of scene-classification and image retrieval.

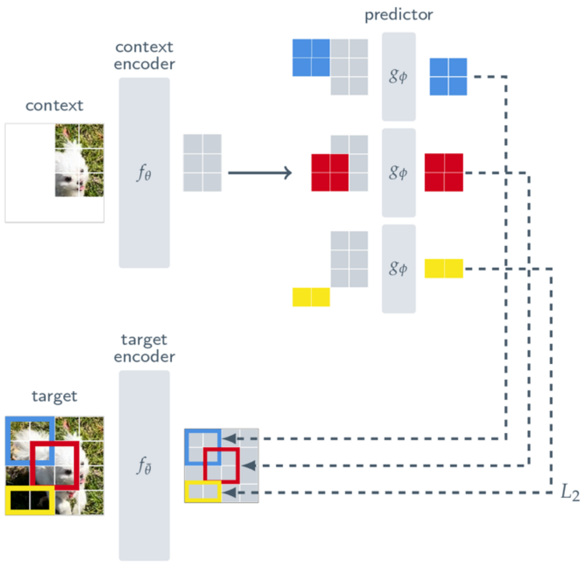

Figure: An illustration of joint-embedding predictive architecture (I-JEPA), from [13]

Requirements

- Experience in machine learning, notably in deep learning
- Proficiency in Python and relevant libraries such as PyTorch, TensorFlow, etc.
- Knowledge of self-supervised learning is a plus
- Strong willingness to learn and ability to work independently

Contact

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

References

[1] G. Sumbul, M. Müller, and B. Demir, “A novel self-supervised cross-modal image retrieval method in remote sensing,” in IEEE International Conference on Image Processing, 2022, pp. 2426–2430.

[2] H. Jung, Y. Oh, S. Jeong, C. Lee, and T. Jeon, “Contrastive selfsupervised learning with smoothed representation for remote sensing,” IEEE Geoscience and Remote Sensing Letters, vol. 19, pp. 1–5, 2022.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

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

[8] 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.

[9] J. Jakubik, S. Roy, C. E. Phillips, P. Fraccaro, D. Godwin, B. Zadrozny, D. Szwarcman, C. Gomes, G. Nyirjesy, B. Edwards, D. Kimura, N. Simumba, L. Chu, S. K. Mukkavilli, D. Lambhate, K. Das, R. Bangalore, D. Oliveira, M. Muszynski, K. Ankur, M. Ramasubramanian, I. Gurung, S. Khallaghi, H. S. Li, M. Cecil, M. Ahmadi, F. Kordi, H. Alemohammad, M. Maskey, R. Ganti, K. Weldemariam, and R. Ramachandran, “Foundation Models for Generalist Geospatial Artificial Intelligence”, arXiv preprint arXiv:2310.18660, 2023.

[10] 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.

[11] 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.

[12] Y. Li, H. Mao, R. Girshick, and K. He, “Exploring plain vision transformer backbones for object detection,” arXiv preprint arXiv:2203.16527, 2022.

[13] M. Assran, Q. Duval, I. Misra, P. Bojanowski, P. Vincent, M. Rabbat, Y. LeCun, and N. Ballas, “Self-Supervised Learning from Images with a Joint-Embedding Predictive Architecture,” IEEE/CVF International Conference on Computer Vision (ICCV), 2023.

[14] G. Christie, N. Fendley, J. Wilson, and R. Mukherjee, “Functional Map of the World,” IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 6172–6180, 2018.