Investigating Large Vision Model Training Challenges on Satellite Datasets

Hitesh Jain, Sagar Verma, Siddharth Gupta

INVESTIGATING LARGE VISION MODEL TRAINING CHALLENGES ON SATELLITE DATASETS

Hitesh Jain [1,2], *Sagar Verma*[2], *Siddharth Gupta*[2]

[1]IIT Gandhinagar, India & [2]Granular AI, Boston, USA
hitesh.jain@iitgn.ac.in, sagar@granular.ai, sid@granular.ai

ABSTRACT

Contrastive learning methods that bridge textual descriptions and images, such as Contrastive Language-Image Pre-training (CLIP), have demonstrated remarkable advancements. These foundational models have shown exceptional performance in tasks related to zero-shot image classification, as evidenced by their substantial enhancement of zero-shot ImageNet accuracy from the prior state-of-the-art of 12% to an impressive 76%. However, the exposure of these models to satellite images during training has been limited, resulting in suboptimal performance when dealing with geospatial data. This limitation raises a pivotal question: Can these foundational models, which have demonstrated potential across multiple domains, be trained on geospatial imagery out-of-box? To answer this question, we perform a study on training CLIP on diverse geospatial datasets. Within our research, we delve into unique challenges in this context and discuss the strategies we employ to address these challenges effectively. We demonstrate that handling resolution is crucial when training CLIP like models on a large multi-resolution dataset.

Index Terms— remote sensing, neural networks, robustness

1. INTRODUCTION

Remote sensing applications like environmental monitoring [1], urban planning, and disaster impact[2] have been possible due to datasets like SpaceNet[3] and QFabric[4]. Training from scratch on these datasets has shown good performance. There are now more nuanced smaller datasets like Farbic[5], SeaDroneSeeV2[6], and SynFlood[7] for very specific tasks. These tasks and datasets can benefit from the zero-shot capabilities of foundational vision models.

Deep learning models pre-trained on large-scale datasets have achieved significant success in various computer vision tasks. ImageNet[8] dataset, which contains an extensive collection of labeled images across multiple categories, has been commonly utilized to pre-train convolutional neural networks (CNNs) and capture high-level representations of features. However, in the case of remote sensing applications, the systematic bias introduced by pre-training on ImageNet can significantly impact performance. This is primarily due to the significant differences between conventional and remote sensing images in orientation, color bands, scale, resolution, sensor data, and other relevant factors.

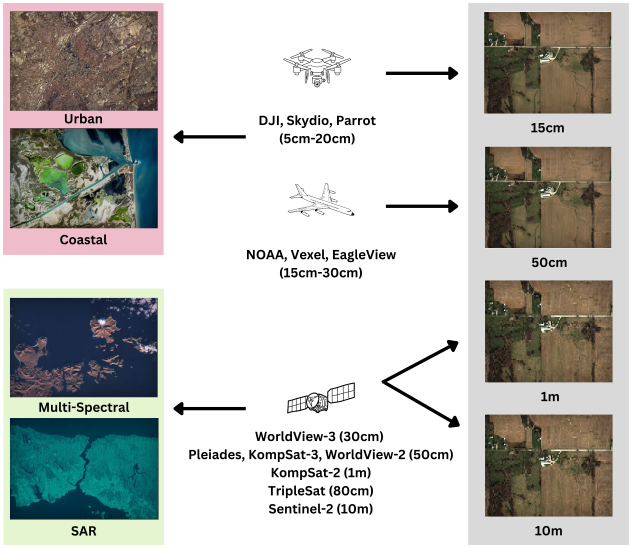

Fig. 1: Remote sensing images captured from diverse sources and varying contexts.

In this paper, we undertake a comprehensive effort by utilizing a diverse dataset gathered from various remote sensing contexts, as shown in Figure 1. These datasets are used to train the CLIP model [14], which is well-known for its ability to learn complex relationships between images and corresponding textual descriptions. The utilization of this model provides a new perspective for addressing the unique challenges posed by remote sensing imagery. One of the main challenges we tackle is the discrepancy in data distribution due to factors such as resolution, image size, and sensor-specific information. These discrepancies often result in a significant decrease in accuracy when deploying pre-trained models for remote sensing tasks. Our research focuses on the intricate interplay between these challenges, aiming to identify the underlying causes of performance degradation and develop effective strategies to mitigate their impact.

Dataset	Problem Type	Sensors	Resolution (cm)	Image size	No. of Images	Channels
Agriculture-2017[9]	Segmentation	UAV	6	512	8345	3
BigEarthNet-S2[10]	Image Classification	Sentinel-2	1000	128	590326	3
BigEarthNet-S1[10]	Image Classification	Sentinel-1	500	120	590326	2
SeaDroneSeeV2[11]	Object Detection	UAV	6	1230	14227	3
xView[12]	Object Detection	WorldView-3	30	3320	5630	3
QFabric[4]	Change Detection	WorldView-2	50	20480	504	3
Houston Harvey NOAA[7]	Segmentation	Aerial	20	2560	32818	3
MAFAT [13]	Object Detection	Aerial	40	1280	9715	1

Table 1: List of compiled datasets used for training CLIP on geospatial images

2. RELATED WORKS

Radford et al.[14] jointly trained image and text encoders (CLIP), using contrastive losses to maximize cosine similarity between image and text representations. Jia et al.[15] curated an exascale, noisy dataset to train a simple dual-encoder architecture to align image and text embeddings using a contrastive loss. Li et al.[16] observe that noise in data leads to sub-optimal model training and attempt to alleviate the same through CapFilt, a bootstrapping mechanism that employs a captioner to synthesize captions, and a filter to remove noisy ones. Yuan et al.[17] try to unify image-text learning by pre-training a combination of hierarchical vision transformer (image encoder) and modified CLIP (language encoder) on web-scale image-label-description triplets. The resultant model demonstrates the outstanding performance of several transfer types, including few-shot and zero-shot transfers. Lacoste et al.[18] proposed to use foundational models like CLIP to leverage satellite images for climate change problems. However, they focus on fine-tuning instead of zero-shot to overcome problems like the availability of small datasets, license issues, and distributional shifts. They do not show empirical evidence on why foundational models are better suited. Panigrahi et al. [19] show that foundational models like BLIP have seen satellite images during the training. Terris et al. [20] show that segmentation and object detection models trained on images from a particular sensor are not robust to other sensors that have the same resolution.

3. EXPERIMENTS & RESULTS

Dataset development for the CLIP model involves acquiring and selecting diverse remote-sensing datasets that cover a broad range of visual and textual information. The datasets are compiled to create a unified dataset that includes images and their corresponding textual descriptions. The unified dataset is stored in the webdataset format, known for its efficiency in distributed training. Images from segmentation and object detection datasets were segmented into 512 x 512 pixels patches. We obtained approximately 4.5 million pairs of images and texts to train the CLIP model from scratch. The validation set contains approximately 1% of the total image-text pairs. Importantly, each image-text pair in the validation set was carefully selected to be unique, avoiding any potential overlap with the training set.

We trained the CLIP model using an open-source implementation provided by ML Foundations[21]. We track all our experiments on GeoEngine platform[22, 23]. The primary objective of these experiments was to assess the performance of the CLIP model on a compiled dataset and investigate the impact of resolution on its ability to learn relationships between textual and image data. However, the initial training on the compiled dataset did not result in any significant improvement over 18 epochs. The performance was deemed unsatisfactory, and the metrics results for R@1, R@5, and R@10 are presented in Table 2. The recall score, denoted as R@K, represents the percentage of top K retrieved captions that are relevant to the input image. A higher recall score indicates the model's effectiveness in accurately capturing the true labels, while a lower score indicates that relevant labels may be missed by the model. Notably, for R@5 and R@10, we consider an image-text pair correct if at least one predicted label matches an actual label.

Dataset	No. of Labels	R@1	R@5	R@10
BigEarthNet-S2	43	0.4	1.6	2.4
BigEarthNet-S1	43	0.2	5.2	14.4
Agriculture-2017	9	17.0	64.0	-
SeaDroneSeeV2	5	1.9	100.0	-
xView	60	1.4	2.2	5.6
QFabric	33	1.2	2.4	4.8
Houston Harvey NOAA	5	96.0	100.0	-
MAFAT	13	5.6	32.6	63.0

Table 2: Image-to-Text Inference for Compiled Dataset

Our hypothesis was that resolution might significantly affect the model's performance. The intuition behind this hypothesis was that the model might need help to learn the textual and image relationships based on resolution differences. To validate this, the datasets were grouped based on resolution, and individual training experiments were conducted.

Datasets were organized into three groups. Group A includes BigEarthNet-S2 and BigEarthNet-S1 and has a resolution of 10m. Group B includes Agriculture-2017 and SeaDroneSeeV2 with 6cm resolution. Group C consists of datasets xView, Qfabric, Houston Harvey NOAA, and MAFAT, which have mixed resolutions ranging from 30-50cm. Experiments were conducted for each of these groups till the validation

Fig. 2: Results showing worst and best prediction sample pairs from the validation set of Qfabric, xView, MAFAT, BigEarthNet-S2, Agriculture-2017, SeaDroneSeeV2 datasets using CLIP trained on Group A, B, and C datasets. The captions display the "missed", "correct" and "false" predictions.

loss converged. The results of these experiments are listed below in Table 3, demonstrating a significant improvement compared to previous training.

To further investigate the impact of resolution on training, we incorporated the resolution and satellite information into the text encoder of CLIP training for Group C datasets. Our observations revealed that the training yielded a varied range of outcomes, leading to a decrease in the R@1 performance but an increase in R@5 and R@10 compared to the previous results, as shown in Table 4. Figure 2 show this case's worst and best predictions.

Dataset	No. of Labels	R@1	R@5	R@10
BigEarthNet-S2	43	4.4	35.4	51.6
Agriculture-2017	9	60.0	93.0	-
SeaDroneSeeV2	5	61.4	100.0	-
xView	60	57.2	60.4	70.6
QFabric	33	55.8	89.8	98.6
Houston Harvey NOAA	5	90.1	100.0	-
MAFAT	13	22.0	61.0	82.2

Table 3: Image-to-Text Inference for different Groups of Datasets

Dataset	No. of Labels	R@1	R@5	R@10
xView	60	1.8	83.8	89.0
QFabric	33	28.2	89.0	99.6
Houston Harvey NOAA	5	10.0	100.0	-
MAFAT	13	7.0	41.8	93.4

Table 4: Image-to-Text Inference for Group C with satellite information in text encoder

4. CONCLUSION

Based on our experiments on the compiled and grouped datasets, we conclude that variation of resolution determines the performance of CLIP. To address the challenge posed by resolution, we propose incorporating satellite information into the text encoder. This augmentation enables the model to harness satellite-specific knowledge, enhancing its interpretive capabilities and improving performance.

We propose using Large Language Models (LLMs) to enhance performance further to develop a comprehensive contextual framework that considers the interaction among various labels, satellite information, resolution, and spectral bands. This comprehensive understanding offers the potential to improve the model's capacity to learn complex relationships within geospatial imagery, thus facilitating increased accuracy and efficiency in diverse applications.

5. REFERENCES

[1] S. Verma, S. Gupta, and K. Gupta, "Aligning Geospatial AI for Disaster Relief with The Sphere Handbook," 2022.

[2] S. Verma and K. Gupta, "Post Wildfire Burnt-up Detection using Siamese UNet," in *ECML PKDD*, 2023.

[3] A. V. Etten, D. Lindenbaum, and T. M. Bacastow, "Spacenet: A remote sensing dataset and challenge series," *CVPR*, 2018.

[4] S. Verma, A. Panigrahi, and S. Gupta, "Qfabric: Multitask change detection dataset," in *CVPR*, 2021.

[5] M. Papadomanolaki, S. Verma, M. Vakalopoulou, S. Gupta, and K. Karantzalos, "Detecting urban changes with recurrent neural networks from multitemporal sentinel-2 data," in *IGARSS*, 2019.

[6] B. Kiefer, M. Kristan, J. Perš, L. Žust, F. Poiesi, and O. Andrade, "1st workshop on maritime computer vision (macvi) 2023: Challenge results," in *WACVW*, 2023.

[7] S. Goswami, S. Verma, K. Gupta, and S. Gupta, "FloodNet-to-FloodGAN : Generating Flood Scenes in Aerial Images," 2022.

[8] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *CVPR*, 2009.

[9] M. T. Chiu, X. Xu, K. Wang, J. Hobbs, N. Hovakimyan, T. S. Huang, and H. Shi, "The 1st agriculture-vision challenge: Methods and results," in *CVPRW*, 2020.

[10] G. Sumbul, A. de Wall, T. Kreuziger, F. Marcelino, and Others, "BigEarthNet-MM: A large scale multi-modal multi-label benchmark archive for remote sensing image classification and retrieval," *IEEE GRS Magazine*, vol. 9, no. 3, pp. 174–180, 2021.

[11] L. A. Varga, B. Kiefer, M. Messmer, and A. Zell, "Seadronessee: A maritime benchmark for detecting humans in open water," in *WACV*, 2022.

[12] D. Lam, R. Kuzma, K. McGee, S. Dooley, M. Laielli, and Others, "xView: Objects in context in overhead imagery," *arXiv:1802.07856*, 2018.

[13] "https://mafatchallenge.mod.gov.il/," 2023.

[14] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, and Others, "Learning transferable visual models from natural language supervision," in *ICML*, 2021.

[15] C. Jia, Y. Yang, Y. Xia, Y.-T. Chen, Z. Parekh, H. Pham, Q. V. Le, Y.-H. Sung, Z. Li, and T. Duerig, "Scaling up visual and vision-language representation learning with noisy text supervision," in *ICML*, 2021.

[16] J. Li, D. Li, C. Xiong, and S. C. H. Hoi, "BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation," in *ICML*, 2022.

[17] L. Yuan, D. Chen, Y.-L. Chen, N. C. F. Codella, X. Dai, J. Gao, and Others, "Florence: A new foundation model for computer vision," *ArXiv*, vol. abs/2111.11432, 2021.

[18] A. Lacoste, E. D. Sherwin, H. R. Kerner, and Others, "Toward foundation models for earth monitoring: Proposal for a climate change benchmark," *ArXiv*, vol. abs/2112.00570, 2021.

[19] A. Panigrahi, S. Verma, M. Terris, and M. Vakalopoulou, "Have foundational models seen satellite images?," in *IGARSS*, 2023.

[20] M. Terris and S. Verma, "Investigating model robustness against sensor variation," in *IGARSS*, 2023.

[21] G. Ilharco, M. Wortsman, R. Wightman, C. Gordon, N. Carlini, and Others, "Openclip," 2021.

[22] S. Verma, S. Gupta, H. Shin, A. Panigrahi, S. Goswami, and Others, "GeoEngine: A platform for production-ready geospatial research," in *CVPRD*, 2022.

[23] H. Shin, N. Exe, U. Dutta, T. R. Joshi, S. Verma, and S. Gupta, "Europa: Increasing accessibility of geospatial datasets," in *IGARSS*, 2022.