Have Foundational Models Seen Satellite Images?

Akash Panigrahi, Sagar Verma, Matthieu Terris, Maria Vakalopoulou

HAVE FOUNDATIONAL MODELS SEEN SATELLITE IMAGES?

Akash Panigrahi [1], Sagar Verma[1,2], Matthieu Terris[1], and Maria Vakalopoulou[2]

[1] Granular AI, MA, USA
[2] CentraleSupélec, Université Paris-Saclay, France
{akash,sagar}@granular.ai
matthieu.terris@gmail.com
maria.vakalopoulou@centralesupelec.fr

ABSTRACT

This paper presents an investigation into the zero-shot performance of pre-trained foundation models on remote sensing tasks. Recent advances in self-supervised learning suggest that these models, when trained on vast amounts of unsupervised data, could potentially improve generalization across a number of downstream tasks. Our study offers an empirical evaluation of these models on standard remote-sensing benchmarks such as EuroSAT and BigEarthNet-S2, with the intent to confirm whether these models have encountered satellite imagery during their training phase. Moreover, we examine the impact of adding a geospatial domain-specific textual description of classes, contrasting it with the standard class-based prompts. Our findings indicate that the fine-tuned BLIP models exhibit superior zero-shot performance on these benchmarks compared to their standard counterparts, signifying that fine-tuning on standard benchmarks enhances performance. Furthermore, the addition of geospatial context variably influences performance depending on the specific model and dataset. This work provides crucial insights into the applicability of foundation models in remote sensing tasks and lays the groundwork for further research.

1. INTRODUCTION

Recent progress in self-supervision shows that pre-training large neural networks on vast amounts of unsupervised data can increase generalization for downstream tasks [1, 2]. Such models, recently coined as *foundation models*, have been transformational to computer vision and natural language processing. In this work, we analyze the zero-shot performance of these models on standard remote sensing image classification datasets as shown in Figure 1. This work aims to establish empirical evidence of whether these foundational models have seen satellite imagery during their training. We believe this can enable the remote sensing community to effectively utilize these models in different downstream tasks like classification, segmentation, object detection, and change detection.

2. RELATED WORKS

Radford et al.[1] jointly trained image and text encoders (CLIP), using contrastive losses to maximize cosine similarity between image and text representations. Jia et al.[3] 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.[2] 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.[4] 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 a number of transfer types, including few-shot and zero-shot transfers. Lacost et al.[5] 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

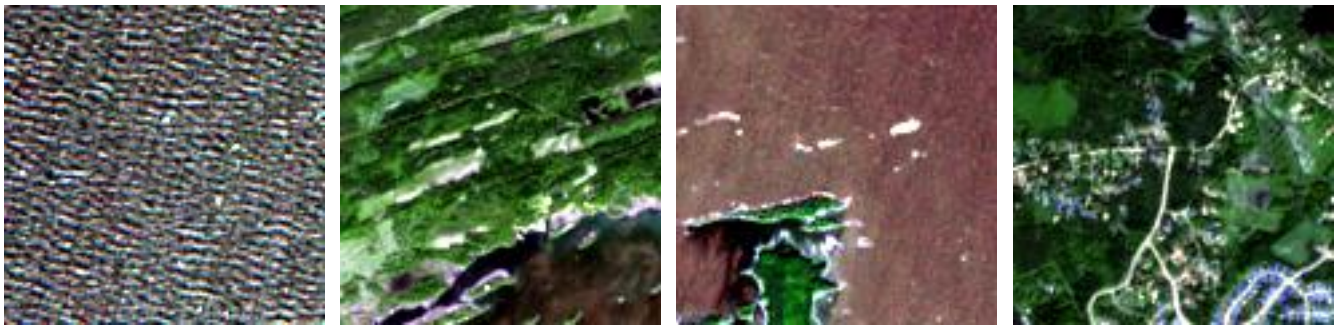

Fig. 1: Captions show ground truth and top-5 predicted labels by CLIP (ViT-B/32) from BigEarthNet-S2 Dataset.

suited. To the best of our knowledge, our work is the first to show empirically how good foundational models are in the case of remote sensing problems.

3. EXPERIMENTS

We obtain zero-shot performance of two pretrained foundation models: CLIP and BLIP, and their image-encoder based variants on remote sensing datasets: EuroSAT [6] and BigEarthNet-S2 [7]. EuroSAT is a toy LULC classification dataset consisting of 27000 geo-referenced Sentinel-2 patches categorized into ten classes. BigEarthNet is a large-scale multi-label dataset with 590,326 Sentinel-2 patches. We observe the performance of these architectures in a standard setting and compare the same with a context-based setting wherein we provide a geospatial domain-specific textual description of classes in contrast to standard class-based prompts in the standard setting. These datasets and experiments performed here are available in the GeoEngine platform [8, 9].

For CLIP, pretrained ResNet and Transformer variants are available in the public domain. We generated inferences on EuroSAT and BigEarthNet-S2 for all these pretrained CLIP versions. For BLIP, standard base and large ViT weights, along with their variants fine-tuned on the large web-scale dataset bootstrapped through CapFilt, and other standard benchmarks like MS-COCO and Flickr30k, are available in the public domain. BLIP employs cross-attention between textual and visual representations instead of standard metrics like cosine similarity for image-text alignment. Such metrics for similarity computation will lead to an inaccurate measurement of its zero-shot performance on EuroSAT and BigEarthNet-S2 datasets. Instead, we repurposed the image-text retrieval mode of BLIP variants for image classification on the remote-sensing benchmarks.

4. RESULTS

We report our findings on the zero-shot performance of variants of foundation models like CLIP and BLIP on EuroSAT and BigEarthNet benchmarks in standard- and context-based settings in Table 1. We observe that CLIP has a near-random performance on BigEarthNet-S2 owing to large image-encoder activations for (almost) all classes that lead to many false positives. We notice that fine-tuned BLIP models have a better zero-shot performance on EuroSAT and BigEarthNet benchmarks than standard variants and can safely conclude that fine-tuning on standard benchmarks improves performance. We observe that zero-shot performance on EuroSAT

CLIP				
Backbone	**EuroSAT**		**BigEarthNet-S2**	
	Standard	**Context**	**Standard**	**Context**
ResNet-50	25.31	28.03	6.82	6.80
ResNet-50x4	22.04	28.79	6.82	6.76
ResNet-50x16	43.13	41.74	6.78	6.71
ResNet-50x64	35.86	17.20	6.80	6.76
ResNet-101	26.74	23.96	6.81	6.82
ViT-B/16	38.86	41.02	6.82	6.84
ViT-B/32	32.67	33.58	6.85	6.82
ViT-L/14	**52.43**	**50.59**	6.82	6.83
ViT-L/14@336px	51.05	45.40	6.82	6.82

BLIP				
Backbone	**EuroSAT**		**BigEarthNet-S2**	
	Standard	**Context**	**Standard**	**Context**
ViT-B/16	36.87	42.35	86.97	84.34
On CapFilt-L	38.55	34.81	87.31	**86.41**
On MS-COCO	36.87	41.20	**89.69**	84.34
On Flickr30	42.67	46.20	88.47	82.30
ViT-L/16	45.78	45.06	81.05	83.74
On MS-COCO	**48.11**	**52.23**	86.21	77.43
On Flickr30	42.35	50.42	87.25	77.03

Table 1: Zero-Shot Performance of CLIP and BLIP on EuroSAT and BigEarthNet-S2. Supervised training from scratch gives best results for ResNet-101 (**93.72%**) and ViT-B/16 (**77.21%**) for EuroSAT and BigEarthNet-S2 respectively.

improves with the addition of remote-sensing context for smaller CLIP variants like ResNet50, ResNet101, and ViT-B/32, and degrades for larger architectures like ViT-L/14 and EfficientNet-based scaled versions of ResNet50. No such visible patterns could be observed in CLIP's performance on context-addition for BigEarthNet labels. Adding geospatial priors leads to a marked improvement in zero-shot performance for most of the BLIP variants on EuroSAT. On BigEarthNet-S2, the addition of context leads to a degradation in performance for most of the BLIP variants.

5. CONCLUSION

In summary, our study presents an empirical examination of the zero-shot performance of pre-trained foundation models on standard remote-sensing datasets like EuroSAT and BigEarthNet-S2. Our findings reveal that fine-tuned BLIP variants outperform the standard version on these benchmarks, and incorporation of geospatial context during the inference stage can lead to mixed outcomes depending on model and dataset selected.

In the future, we intend to expand the scope of our investigation by involving more task types and datasets. We will be incorporating more foundational models, such as ALBEF [10], ViLD[11], and ZSI[12]. These models have been trained on different large private datasets, which may or may not have satellite and aerial images. Our larger experiment set will consist of BigEarthNet-S1, SynthWakeSAR[13] in case of classification. SeeDroneSeaV2[14], xView[15] and DOTA[16] in case of object detection. Houston UAV[17, 18] in case of semantic segmentation. We also plan to utilize these models in a very naive way for change detection problems on OSCD BiDate[19], OSCD MultiDate[20], and QFabric[21] datasets.

6. REFERENCES

[1] A. Radford, J. W. Kim, C. Hallacy, et al., "Learning transferable visual models from natural language supervision," in *ICML*, 2021.

[2] 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.

[3] C. Jia, Y. Yang, Y. Xia, et al., "Scaling up visual and vision-language representation learning with noisy text supervision," in *ICML*, 2021.

[4] L. Yuan, D. Chen, Y.-L. Chen, et al., "Florence: A new foundation model for computer vision," *ArXiv*, vol. abs/2111.11432, 2021.

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

[6] P. Helber, B. Bischke, A. R. Dengel, and D. Borth, "EuroSAT: A novel dataset and deep learning

benchmark for land use and land cover classification," *IEEE JSTARS*, vol. 12, pp. 2217–2226, 2017.

[7] G. Sumbul, M. Charfuelan, B. Demir, and V. Markl, "BigEarthNet: A large-scale benchmark archive for remote sensing image understanding," *IGARSS*, pp. 5901–5904, 2019.

[8] H. Shin, N. Exe, U. Dutta, et al., "Europa: Increasing accessibility of geospatial datasets," in *IGARSS*, 2022.

[9] S. Verma, S. Gupta, H. Shin, et al., "GeoEngine: A platform for production-ready geospatial research," in *CVPRD*, 2022, pp. 21416–21424.

[10] J. Li, R. R. Selvaraju, A. D. Gotmare, et al., "Align before fuse: Vision and language representation learning with momentum distillation," in *NeurIPS*, 2021.

[11] X. Gu, T.-Y. Lin, W. Kuo, and Y. Cui, "Open-vocabulary object detection via vision and language knowledge distillation," in *ICLR*, 2021.

[12] Y. Zheng, J. Wu, Y. Qin, et al., "Zero-shot instance segmentation," *CVPR*, pp. 2593–2602, 2021.

[13] I. G. Rizaev and A. Achim, "SynthWakeSAR: A synthetic sar dataset for deep learning classification of ships at sea," *Remote Sensing*, vol. 14, 2022.

[14] B. Kiefer, M. Kristan, J. Perš, et al., "1st workshop on maritime computer vision (macvi) 2023: Challenge results," in *WACV Workshops*, January 2023, pp. 265–302.

[15] D. Lam, R. Kuzma, K. McGee, et al., "xView: Objects in context in overhead imagery," *arXiv:1802.07856*, 2018.

[16] G.-S. Xia, X. Bai, J. Ding, et al., "DOTA: A large-scale dataset for object detection in aerial images," in *CVPR*, June 2018.

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

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

[19] R. C. Daudt, B. Le Saux, A. Boulch, and Y. Gousseau, "Urban change detection for multispectral earth observation using convolutional neural networks," in *IGARSS*, 2018.

[20] M. Papadomanolaki, S. Verma, M. Vakalopoulou, et al., "Detecting urban changes with recurrent neural networks from multitemporal sentinel-2 data," in *IGARSS*, 2019, pp. 214–217.

[21] S. Verma, A. Panigrahi, and S. Gupta, "QFabric: Multi-task change detection dataset," in *CVPRW*, 2021, pp. 1052–1061.