Disentangling the Benefits of Self-Supervised Learning to Deployment-Driven Downstream Tasks of Satellite Images (Student Abstract)

Zhuo Deng[1,2,*], Yibing Wei[3,4], Mingye Zhu[5,2], Xueliang Wang[1], Junchi Zhou[1], Zhicheng Yang[4], Hang Zhou[4], Zhenjie Cao[2,1], Lan Ma[1], Mei Han[4], Jui-Hsin Lai[4]

[1]Tsinghua Shenzhen International Graduate School, Shenzhen, Guangdong, China
[2]Ping An Technology, Shenzhen, Guangdong, China
[3]University of Wisconsin-Madison, Madison, WI, USA
[4]PAII Inc., Palo Alto, CA, USA
[5]University of Science and Technology of China, Hefei, Anhui, China
dz20@mails.tsinghua.edu.cn, zcyangpingan@gmail.com, Juihsin.lai@gmail.com

Abstract

In this paper, we investigate the benefits of self-supervised learning (SSL) to downstream tasks of satellite images. Unlike common student academic projects, this work focuses on the advantages of the SSL for deployment-driven tasks which have specific scenarios with low or high-spatial resolution images. Our preliminary experiments demonstrate the robust benefits of the SSL trained by medium-resolution (10m) images to both low-resolution (100m) scene classification case (4.25%↑) and very high-resolution (5cm) aerial image segmentation case (1.96%↑), respectively.

Introduction

Automated analysis of remote sensing (RS) imagery is the key to monitoring global issues. Hundreds of satellites collect plentiful RS data on a daily basis. Since most images remain unlabeled, typical supervised learning algorithms are unable to make full use of the massive amounts of RS data (Huang, Yang et al. 2022). Recently, self-supervised learning (SSL), aiming to learn image latent representations from massive unlabeled data, has validated the significant effectiveness of the learned feature in various supervised downstream tasks on various datasets. The SSL mainly has two categories: contrastive methods and generative methods (Liu et al. 2021). Contrastive methods strongly rely on data augmentation to learn semantically invariant features which focus more on the central region of the image. Generative methods learn to reconstruct the original signal from corrupted input with an autoencoder structure.

With the rapid development of Artificial Intelligence (AI), a current hot focus is to carry out AI to empower diverse applications in different areas, including bridging the gap between AI and deployment-oriented tasks. For example in the RS domain, how is a regular downstream task of satellite images motivated by a client's specific demands? In particular for the SSL, its efficacy in deployment-driven downstream scenarios has not been well addressed.

In this study, we aim to disentangle the benefits of SSL to deployment-driven downstream tasks of satellite images, which is a fundamental task of the student's college-enterprise collaborative project. Unlike regular student academic projects, this program enables students to cultivate an AI-deployment-oriented mindset, offering benefits to future industrial career success. On the enterprise side, the clients span from internal finance sectors to external governments, yet lacking domain knowledge of remote sensing. Therefore, instead of delivering sophisticated algorithms to them, the must-know baselines are valuable to unveil the substantial benefits of the SSL in deployment scenarios. Due to the ultimate commercial purpose of this project, we use Sentinel-2 satellite data (Drusch, Del Bello et al. 2012) as our SSL image source, which is free for commercial use and has a moderate spatial resolution (10∼60m).

Deployment Tasks

Case Study 1. The statistics of land cover and land use (LULC) of a region are important for the local government to monitor soil erosion, resource management, civilization progress, etc. However, fine-grained statistics are very challenging due to the huge labor cost. Instead, a coarse-grained scene classification task is beneficial to a governmental client for an approximate LULC estimation. Our objective in this case is thus to explore the efficacy of SSL for a low-resolution scene classification task. The dataset used for validating this task is the NaSC-TG2 scene classification dataset (Zhou et al. 2021), collected from the visible and near-infrared spectral channels from the TianGong-2 satellite. It has 20,000 images with 10 scene classes, an image size of 128×128 pixels, and a spatial resolution of 100m.

Case Study 2. Many clients expect to monitor a small area of interest (AoI) covered by unmanned aerial vehicle (UAV) imagery in terms of object recognition, change detection, crop identification, etc. UAV images have very high-resolution (cm-level) with a small area of coverage but a

[*]The current progress of this student's project was awarded the Summer 2022 Outstanding Student Practice Award by Tsinghua University and Ping An Technology.

Method	Train/Validation Ratio		
	2:8	5:5	8:2
ResNet50 w/ ImageNet(RGB)	92.85	93.64	94.05
SimCLR+ResNet50 w/ BEN3	92.25	93.37	94.45
SimCLR+ResNet50 w/ BEN4	92.90	95.59	97.13
MAE+ViT-B w/ BEN4	**93.28**	**96.11**	**98.30**

Table 1: Top-1 accuracies of the downstream scene classification task on NaSC-TG2 dataset with CNN or ViT backbones under different SSL settings in Case Study 1. (BEN3 or 4: BigEarthNet(RGB) or (RGBN))

high running cost. To tackle this demand, an image segmentation task is appropriate. Hence in this case, our goal is to investigate the effectiveness of SSL for a segmentation task of high-resolution aerial images. The Potsdam segmentation dataset (Sherrah 2016) is used for this task. It has 28 6,000×6,000 manually labeled images with a spatial resolution of 5cm. It has 6 classes of land cover. We crop the image into 600×600 patches with a stride of 300 pixels. Due to the irregular class "Clutter", we follow the rule to exclude it for training and inference procedures (Wang et al. 2022).

Experiment Results

Implementation Details. We exploit a public curated dataset built by Sentinel-2 data, BigEarthNet (Sumbul, Charfuelan et al. 2019), as our SSL training data. This dataset contains 590,326 images with 12 spectral channels. We select red, green, blue, and near-infrared (RGBN) channels with a 10m resolution and an image size of 120×120 pixels. To convey the fundamental knowledge of the SSL's efficacy to clients, we select SimCLR with ResNet50 (Chen et al. 2020) and masked autoencoders (MAE) with Vision Transformer base model (ViT-B) (He et al. 2022) as the representative contrastive and generative methods of SSL, respectively. For the downstream tasks, we leverage simple MLP layers to achieve the classification result, and adopt UperNet (Xiao et al. 2018) as the segmentation model. Our machine has eight V100 GPUs. The hyper-parameters are carefully tuned to get the best performance.

Results of Case Studies. The scenario in Case Study 1 is to transfer the representation knowledge learned from medium resolution (10m) to low resolution (100m). Table 1 presents the classification results with different backbone architectures and different SSL settings. We observe that: 1) the SSL using images with domain knowledge is competitive with the supervised ImageNet pre-trained model; 2) adopting multiple spectral channels in the SSL obviously outperform others; 3) the generative method MAE with ViT-Base achieves the best result. Table 2 lists the segmentation results in Case Study 2, where the representation knowledge learned from medium resolution (10m) is evaluated using very high-resolution images (5cm). We observe that: 1) both contrastive and generative methods can improve the performance; 2) the improvement is not so significant as it in the low-resolution classification task; 3) the vanilla ViT is slightly worse than the CNN-based model.

Method (w/ BEN4)	mIoU	OA
ResNet50+UperNet	80.80	89.80
SimCLR+ResNet50+UperNet	**82.76**	**91.01**
ViT-B+UperNet	76.57	87.33
MAE+ViT-B+UperNet	**78.35**	**88.41**

Table 2: Mean Intersection over Union (mIoU) and Overall Accuracy (OA) of the downstream semantic segmentation task on Potsdam dataset with and without SSL in Case Study 2. (BEN4: BigEarthNet(RGBN))

Discussion and Future Work

There are numerous public datasets for the SSL of satellite images. However, most licenses are not allowed for commercial use, leading to a huge barrier to carrying out the SSL in the actual deployment scenarios. In this preliminary study, we leverage the Sentinel-2 image source for the SSL training, and validate its efficacy with different SSL paradigms for deployment-driven downstream tasks with low and high-resolution images. Since the downstream tasks have specific application scenarios, we select appropriate datasets to approach the desired spatial resolutions for each case (10m for SSL, 100m for Case 1, and 5cm for Case 2). Our ongoing work includes building our in-house database platform and evaluating the SSL models with a more dedicated design for specific deployment-oriented downstream tasks.

References

Chen, T.; Kornblith, S.; Norouzi, M.; and Hinton, G. 2020. A simple framework for contrastive learning of visual representations. In *ICML*, 1597–1607. PMLR.

Drusch, M.; Del Bello, U.; et al. 2012. Sentinel-2: ESA's optical high-resolution mission for GMES operational services. *Remote sensing of Environment*, 120: 25–36.

He, K.; Chen, X.; Xie, S.; Li, Y.; Dollár, P.; and Girshick, R. 2022. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF CVPR 2022*, 16000–16009.

Huang, F.; Yang, Z.; et al. 2022. Unsupervised superpixel-driven parcel segmentation of remote sensing images using graph convolutional network. In *ACM WWW'22 Companion*.

Liu, X.; Zhang, F.; Hou, Z.; Mian, L.; et al. 2021. Self-supervised learning: Generative or contrastive. *IEEE Transactions on Knowledge and Data Engineering*.

Sherrah, J. 2016. Fully convolutional networks for dense semantic labelling of high-resolution aerial imagery. *arXiv preprint arXiv:1606.02585*.

Sumbul, G.; Charfuelan, M.; et al. 2019. Bigearthnet: A large-scale benchmark archive for remote sensing image understanding. In *IEEE IGARSS*, 5901–5904. IEEE.

Wang, D.; Zhang, J.; Du, B.; Xia, G.-S.; and Tao, D. 2022. An Empirical Study of Remote Sensing Pretraining. *IEEE Transactions on Geoscience and Remote Sensing*, 1–1.

Xiao, T.; Liu, Y.; Zhou, B.; Jiang, Y.; and Sun, J. 2018. Unified perceptual parsing for scene understanding. In *Proceedings of the ECCV*, 418–434.

Zhou, Z.; Li, S.; Wu, W.; et al. 2021. NaSC-TG2: Natural scene classification with Tiangong-2 remotely sensed imagery. *IEEE J-STARS*, 14: 3228–3242.