

FAST MODEL INFERENCE AND TRAINING ON-BOARD OF SATELLITES

Vít Růžička^{1,2}, Gonzalo Mateo-García^{2,3}, Chris Bridges⁴,
Chris Brunskill⁶, Cormac Purcell^{2,5}, Nicolas Longépé⁷, Andrew Markham¹

¹ University of Oxford, ² Trillium Technologies, ³ University of Valencia,
⁴ University of Surrey, ⁵ University of New South Wales, ⁶ D-Orbit ⁷ European Space Agency

ABSTRACT

Artificial intelligence onboard satellites has the potential to reduce data transmission requirements, enable real-time decision-making and collaboration within constellations. This study deploys a lightweight foundational model called RaVAEn on D-Orbit’s ION SCV004 satellite. RaVAEn is a variational auto-encoder (VAE) that generates compressed latent vectors from small image tiles, enabling several downstream tasks. In this work we demonstrate the reliable use of RaVAEn onboard a satellite, achieving an encoding time of 0.110s for tiles of a 4.8x4.8 km² area. In addition, we showcase fast few-shot training onboard a satellite using the latent representation of data. We compare the deployment of the model on the on-board CPU and on the available Myriad vision processing unit (VPU) accelerator. To our knowledge, this work shows for the first time the deployment of a multi-task model on-board a CubeSat and the on-board training of a machine learning model.

Index Terms— Training on-board, AI on satellites, efficient neural network models

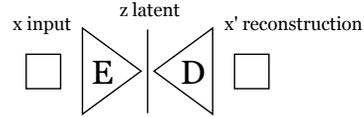
1. INTRODUCTION

Onboard data processing plays a crucial role in maximizing the potential of Earth-observation (EO) satellites. With the significant increase in EO data volume, it is essential to have efficient and intelligent processing capabilities directly onboard the satellites [1]. By leveraging onboard data processing, satellites can perform advanced analysis and make critical decisions on the acquired data in real-time. Several applications have already been tested in demonstration missions, such as prioritizing imaging targets [2], discarding non-usable images [3], identifying events of interest [4, 5, 6] or compressing the output to rapidly transmit relevant information to the ground [7, 8].

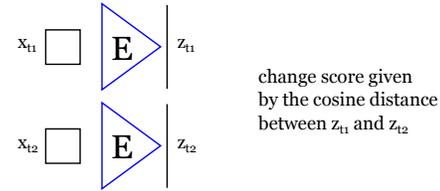
Correspondence to: vit.ruzicka@cs.ox.ac.uk

This work has been funded by ESA Cognitive Cloud Computing in Space initiative. G.M.-G. has been partially supported by the Spanish Ministry of Science and Innovation (project PID2019-109026RB-I00 funded by MCIN/AEI/10.13039/501100011033) and the European Social Fund.

1.) Pre-training to encode (Variational Autoencoder)



2.) Usage for Change Detection



3.) Usage for Few-Shot Training

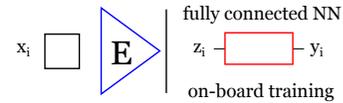


Fig. 1: Illustration of the different uses of the variational auto-encoder model RaVAEn. Pre-training (1.) conducted in prior work [2]. In this paper, we measure the inference time of the encoder network (denoted as E and marked with blue colour as with frozen weights) for the task of unsupervised change detection (2.), and time needed for on-board training (3.) of a small fully-connected neural network classification model.

In this work we go one step further and deploy a lightweight foundational model called RaVAEn on-board of D-Orbit’s ION SCV004 satellite demonstration mission. RaVAEn is a variational auto-encoder (VAE) model that generates latent vectors from small tiles of the original image. These latents can be used for several other tasks, such as change detection, as shown in our previous publication in the context of disaster response [2], or as features extracted to train other downstream models.

We show that this model can be reliably used with the compute available directly on-board of the D-Orbit’s ION-

SCV 004 satellite. In addition, we also demonstrate to the best of our knowledge the world’s first fast and efficient few-shot training on-board of a satellite using the latent representation of the data. To this end, we use the learned encoder of the VAE model to represent tiles of 32×32 pixels with 4 bands as 128-dimensional latent vectors. We then train a lightweight classification model using these latent vectors as inputs in a few-shot learning manner. Good representation of the Sentinel-2 data is required for training with only limited number of samples. As a demonstration task, we select cloud detection: in this context the decision if a tile contains clouds or not. This task is relevant for on-board data processing as it has been previously used to select which image acquisitions are downlinked to the ground station and which are to be ignored [9, 3].

To summarise, our contributions are:

- Measuring the inference time of the RaVAEn model encoder on three different compute regimes available: Myriad VPU, or CPU with Pytorch, or OpenVino libraries.
- Demonstrating few-shot training directly on-board of a satellite for a task of cloud detection, as a motivation for future on-board auto-calibration tasks. To the best of our knowledge, this is the world’s first case of on-board few-shot training on-board of a satellite.

We release the used code in a GitHub repository at [previtus/RaVAEn-unibap-dorbit](https://github.com/previtus/RaVAEn-unibap-dorbit)

2. METHODOLOGY

The RaVAEn uses a VAE model [10] trained on Sentinel-2 L1C data from the WorldFloods dataset [11]. The VAE model consists of an encoder network that reduces the dimensionality of the input data into a latent vector, and of a decoder network that has to reconstruct the original data from this compressed representation. The learned encoding space has to learn an informative representation of the data. This can be leveraged for unsupervised change detection, where, instead comparing changes in the pixel space (which can be noisy due to a wide array of effects), we compare the data representations in the latent space. This approach was evaluated on an annotated dataset of disaster events containing samples with landslides, floods, hurricanes and fires in [2]. In this follow-up paper, we are interested in the inference time on real satellites and in exploring new tasks that this architecture allows us to do directly on-board. Namely, we are interested in the inference time required to process Sentinel-2 data by the encoder network of the RaVAEn model. We note, that the original architecture was designed with a requirements for fast inference in mind, and the fastest model was chosen from several variants.

Additionally, we explore training the mdoel on-board satellites, but instead of using the full dimensionality of the

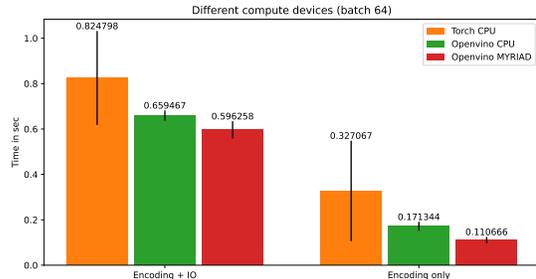


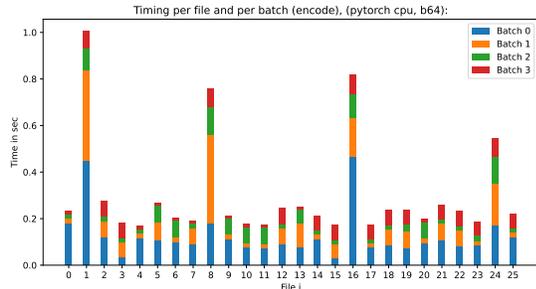
Fig. 2: Timed measurements model inference, using the RaVAEn encoder network with different devices available on the satellite.

input data, we leverage the general pre-trained encoder of the RaVAEn model. We train a tiny classification model on the encoded latent vectors directly on-board of the satellite. The general VAE encoder is capable of efficient data representation, which we can use for few-shot learning. The suitability of few-shot learning has been highlighted by [12]. Importantly, the resulting training process is faster and requires fewer labels than would be required when training the VAE model from scratch. From the dataset of Sentinel-2 L1C images, we select 1305 tiles (cloudy and not-cloudy) for the training dataset and use the non-overlapping remainder of the data for evaluation. We note that other approaches frame cloud detection as semantic segmentation task, however as a demo task, per tile classification is sufficient and can still provide us with a rough estimation of the percentage of clouds in a scene.

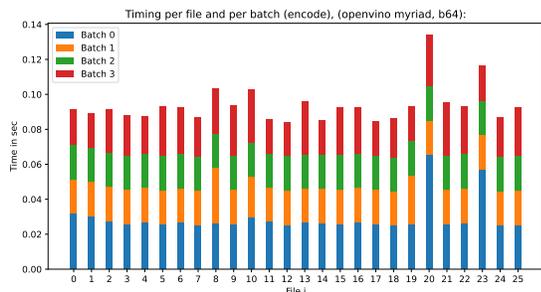
Hardware The ION-SCV 004 satellite has the following relevant specifications: a quad-core x86 64-bit CPU processor, Intel Movidius Myriad X VPU and 2GB RAM. Similar configuration was used in the work of [7]. We note, that for smaller CubeSats, model training can be offloaded to these satellites.

3. RESULTS

Model inference with RaVAEn We measure the time required to load, encode and compare all the tiles from a sequential dataset of Sentinel-2 images. In Figure 2 we show the average inference time of encoding one file (consisting of 225 tiles) with a batch size of 64 tiles, using the RGB+NIR bands. We observe that the deployment of the model on the Myriad VPU with OpenVino offers the fastest encoding time. Furthermore, when inspecting the individual encoding times per batch of each file in Figure 3b, we see that the Myriad VPU is also more robust to slowdowns, which occur when using the CPU with PyTorch (Fig. 3a). The relatively slow loading and tiling of the images can be speed up if we process data with delay and parallelisation (as shown in [13]).



(a) CPU with Pytorch



(b) Myriad VPU with OpenVino

Fig. 3: Detailed view into the individual batches used in the RaVAEn encoder, showing delays in the CPU regime.

Training models on-board of satellites We measure training times and also the performance on the downstream task. On the demonstration task of cloud detection, we get an AUPRC score of 0.979. With a confidence threshold of 0.5, we get recall of 0.946, precision of 0.967 and a F1 score of 0.956. We note that this task serves only as a demo, as we are mostly interested in the timing of the entire training process.

On Figure 4, we see the average time measurements for each epoch when using different batch sizes, and when training a tiny one-layer binary classification model with 129 trainable parameters. With the batch size of 256, one epoch takes on average only 0.091s to train.

4. DISCUSSION AND CONCLUSION

The domain of AI on-board of satellites is unique in comparison with the rest of computer vision research, as the remote sensing data usually undergoes heavy post-processing steps after it has been down streamed to the ground station. On-board processing and training has to, on the other hand, deal with the near-raw data capture, which poses unique opportunities such as re-training existing models with newly observed data - a challenge identified by [14].

In this work, we demonstrate the possibilities of training directly on-board of the satellite, which is of interest for future self-calibration tasks. Training on-board is feasible in sce-

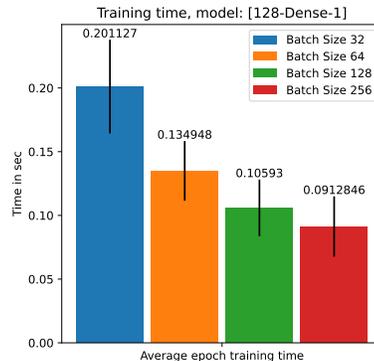


Fig. 4: Training a classification model on top of the encoder network, timing while changing the used batch size.

narios where we obtain both raw measurements of the scene and a reliable annotation. This can occur in cases, where the instrument carries samples with known ground truth labels [15], in cases when we are orbiting around a known location with known state of the observed data, or in cases where another satellite provides us with labels due to having access to more powerful or more reliable instruments (as can be the case when working with a mixture of multispectral and hyperspectral as hypothesized in [6]).

In comparison with [14], which suggests uplinking updated versions of model weights from the ground stations, we propose training on-board as a new approach for adapting AI models in the space. This may be more beneficial for security reasons, and in communication constrained environments, where collection possibilities of new data outweighs the transmission limitations. We consider these scenarios as exciting new opportunities to explore for increasing autonomy of satellites deployed around the Earth and in deep space.

5. REFERENCES

- [1] Gianluca Furano, Gabriele Meoni, Aubrey Dunne, David Moloney, Veronique Ferlet-Cavrois, Antonis Tavoularis, Jonathan Byrne, Léonie Buckley, Mihalis Psarakis, Kay-Obbe Voss, and Luca Fanucci, “Towards the use of artificial intelligence on the edge in space systems: Challenges and opportunities,” *IEEE Aerospace and Electronic Systems Magazine*, vol. 35, no. 12, pp. 44–56, 2020.
- [2] Vít Růžička, Anna Vaughan, Daniele De Martini, James Fulton, Valentina Salvatelli, Chris Bridges, Gonzalo Mateo-Garcia, and Valentina Zantedeschi, “RaVAEn: unsupervised change detection of extreme events using ML on-board satellites,” *Scientific reports*, vol. 12, no. 1, pp. 1–10, 2022.

- [3] Gianluca Giuffrida, Luca Fanucci, Gabriele Meoni, Matej Batič, Léonie Buckley, Aubrey Dunne, Chris Van Dijk, Marco Esposito, John Hefele, Nathan Ver-cruyssen, Gianluca Furano, Massimiliano Pastena, and Josef Aschbacher, “The ϕ -Sat-1 mission: the first on-board deep neural network demonstrator for satellite earth observation,” *IEEE Transactions on Geoscience and Remote Sensing*, pp. 1–1, 2021, Conference Name: IEEE Transactions on Geoscience and Remote Sensing.
- [4] Vincenzo Fanizza, David Rijlaarsdam, Pablo Tomás Toledano González, and José Luis Espinosa-Aranda, “Transfer learning for on-orbit ship segmentation,” in *Computer Vision – ECCV 2022 Workshops*, Leonid Karlinsky, Tomer Michaeli, and Ko Nishino, Eds., Cham, 2023, pp. 21–36, Springer Nature Switzerland.
- [5] Maciej Ziaja, Piotr Bosowski, Michal Myller, Grzegorz Gajoch, Michal Gumieła, Jennifer Protich, Katherine Borda, Dhivya Jayaraman, Renata Dividino, and Jakub Nalepa, “Benchmarking deep learning for on-board space applications,” *Remote Sensing*, vol. 13, no. 19, 2021.
- [6] Vít Růžička, Gonzalo Mateo-Garcia, Luis Gómez-Chova, Anna Vaughan, Luis Guanter, and Andrew Markham, “STARCO: Semantic Segmentation of Methane Plumes with Hyperspectral Machine Learning Models,” <https://doi.org/10.21203/rs.3.rs-2899370/v1>, 2023.
- [7] Gonzalo Mateo-García, Josh Veitch-Michaelis, Cormac Purcell, Nicolas Longepe, Pierre Philippe Mathieu, Simon Reid, Alice Anlind, Fredrik Bruhn, and James Parr, “In-orbit demonstration of a re-trainable machine learning payload for processing optical imagery,” <https://doi.org/10.21203/rs.3.rs-1941984/v1>, 2022.
- [8] Justin S. Goodwill, James P. MacKinnon, Kristy Sakano, and Christopher M. Wilson, “Onboard hyperspectral image classification via transfer learning for communication-limited spacecraft,” in *2022 IEEE Aerospace Conference (AERO)*, 2022, pp. 1–13.
- [9] Kiri L Wagstaff, Alphan Altinok, Steve A Chien, Umaa Rebbapragada, Steve R Schaffer, David R Thompson, and Daniel Q Tran, “Cloud filtering and novelty detection using onboard machine learning for the eo-1 spacecraft,” in *Proc. IJCAI Workshop AI in the Oceans and Space*, 2017.
- [10] Diederik P Kingma and Max Welling, “Auto-encoding variational bayes,” *arXiv preprint arXiv:1312.6114*, 2013.
- [11] Gonzalo Mateo-Garcia, Joshua Veitch-Michaelis, Lewis Smith, Silviu Vlad Oprea, Guy Schumann, Yarin Gal, Atılım Güneş Baydin, and Dietmar Backes, “Towards global flood mapping onboard low cost satellites with machine learning,” *Scientific reports*, vol. 11, no. 1, pp. 1–12, 2021.
- [12] Dawa Derksen, Gabriele Meoni, Gurvan Lecuyer, Anne Mergy, Marcus Märtens, and Dario Izzo, “Few-shot image classification challenge on-board,” in *Workshop-Data Centric AI, NeurIPS*, 2021.
- [13] Vít Růžička and Franz Franchetti, “Fast and accurate object detection in high resolution 4k and 8k video using gpus,” in *2018 IEEE High Performance extreme Computing Conference (HPEC)*. IEEE, 2018, pp. 1–7.
- [14] Gianluca Furano, Antonis Tavoularis, and Marco Rovatti, “AI in space: Applications examples and challenges,” in *2020 IEEE International Symposium on Defect and Fault Tolerance in VLSI and Nanotechnology Systems (DFT)*. IEEE, 2020, pp. 1–6.
- [15] James F Bell III, A Godber, S McNair, MA Caplinger, JN Maki, MT Lemmon, J Van Beek, MC Malin, D Wellington, KM Kinch, et al., “The mars science laboratory curiosity rover mastcam instruments: Preflight and in-flight calibration, validation, and data archiving,” *Earth and Space Science*, vol. 4, no. 7, pp. 396–452, 2017.