BENCHMARKING DEEP LEARNING INFERENCE OF REMOTE SENSING IMAGERY ON THE QUALCOMM SNAPDRAGON AND INTEL MOVIDIUS MYRIAD X PROCESSORS ONBOARD THE INTERNATIONAL SPACE STATION

Emily Dunkel[1], Jason Swope[1], Zaid Towfic[1], Steve Chien[1], Damon Russell[1], Joseph Sauvageau[1], Douglas Sheldon[1], Juan Romero-Cañas[2], Jose Luis Espinosa-Aranda[2], Léonie Buckley[2], Elena Hervas-Martin[2], Mark Fernandez[3], Carrie Knox[3]

[1]Jet Propulsion Laboratory, California Institute of Technology, USA
[2]Ubotica
[3]Hewlett Packard Enterprise

ABSTRACT

Deep Space missions can benefit from onboard image analysis. We demonstrate deep learning inference to facilitate future mission adoption of said algorithms. Traditional space flight hardware provides modest compute when compared to today's laptop and desktop computers. New generations of commercial off the shelf (COTS) processors designed for embedded applications, such as the Qualcomm Snapdragon and Movidius Myriad X, deliver significant compute in small Size Weight and Power (SWaP) packaging and offer direct hardware acceleration for deep neural networks. We deploy neural network models on these processors hosted by Hewlett Packard Enterprise's Spaceborne Computer-2 onboard the International Space Station (ISS). We benchmark a variety of algorithms trained on imagery from Earth or Mars, as well as some standard deep learning models for image classification.

Index Terms— *Deep Learning, Edge Processing, Space Applications, Machine Learning, Artificial Intelligence, COTS embedded processors*

1. INTRODUCTION

Deep space missions have limited contact with ground operations teams[1], making it hard to account for execution variation. Onboard autonomy can address this, but traditional space flight hardware has very limited capabilities. A new generation of processors, such as the Qualcomm Snapdragon 855 [1] and Intel Movidius Myriad X [2], enable onboard inference by supporting neural networks directly in hardware [3]. This technology promises more powerful edge computing.

We benchmark deep learning models trained on imagery from Earth and Mars on Snapdragon and Movidius Myriad X processors onboard the ISS. Hosting of these processors is enabled by Spaceborne Computer-2 (SBC-2) by Hewlett Packard Enterprise [4]. Previously, these models have been deployed on the ground. The ISS deployment is a step towards running such models on a satellite, a Lunar outpost or a Mars Rover, to enable onboard data analysis, targeted downloads, commanding of space assets, and onboard science interpretation.

2. PROCESSORS AND DEPLOYMENT

The Qualcomm Snapdragon 855 has multiple subsystems, including a CPU cluster with 8 ARM cores, an Adreno GPU, a Digital Signal Processor (DSP), and an AI Processor (AIP), sometimes referred to as the Neural Processing Unit (NPU). The NPU can be used to select the right component for a given task. The ARM and GPU support floating point numbers, while the DSP/NPU support fixed-point only. Snapdragon processors have been used in vehicles, drones, and even the Mars Ingenuity Helicopter [5].

The Myriad X Visual Processing Unit (VPU) features a Neural Compute Engine, which is a dedicated hardware accelerator for performing inference with neural networks, as well as cores for accelerating computer vision algorithms. The VPU is programmable using Ubotica's CVAI Toolkit™. Half precision floating point is supported. The previous generation VPU, the Myriad 2, flew on the PhiSat-1, a CubeSat mission from the European Space Agency [3].

Two Snapdragon 855 handheld development boards and two Movidius Myriad X Processors were integrated with the

[1] Due to limited numbers of Earth-based ground communications stations and geometric constraints. Surface missions typically are commanded daily or every several days and orbiters are typically commanded weekly.

HPE SBC-2, which was launched as part of the ISS resupply mission Cygnus NG-15 on February 20th, 2021. Uplinks are possible periodically to load new software.

3. MODELS AND BENCHMARKS

We benchmark deep learning models for image classification, image segmentation, and spectral super-resolution.

3.1. Mars HiRISENet and Mars MSLNets

The Mars imagery classifiers we benchmark include Mars HiRISENet and two Mars MSLNet Convolutional Neural Networks (CNNs). HiRISENet is used to classify images collected by the High Resolution Imaging Experiment (HiRISE) instrument onboard the Mars Reconnaissance Orbiter (MRO). MRO is used to study Martian surface features. HiRISENet is trained on classes that include dunes, craters, the debris from volcanic eruptions, and features formed by sublimating CO_2 [6]. In the future, a similar model could enable data analysis onboard an orbiter.

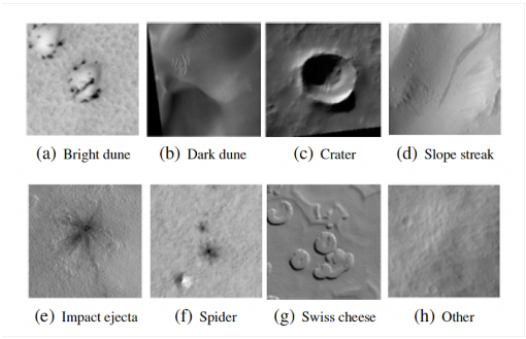

Figure 1: Example Mars HiRISE Classes [6]

Table 1: Mars HiRISE Classifier Benchmarks

	Errors	Inference Time	Energy
Linux Reference	-	56.9 ms	2.3 J*
Snapdragon CPU	0	87.8 ms	0.5 J
Snapdragon GPU	1 (0.1%)	16.3 ms	0.051 J
Snapdragon DSP	15 (0.8%)	7.6 ms	0.016 J
Snapdragon NPU	15 (0.8 %)	7.6 ms	0.014 J
Myriad X	2 (0.1%)	16.2 ms	0.032 J

MSLNet1 and 2 are used to classify images collected by the Mast Camera (Mastcam) and the Mars Hand Lens Imager (MAHLI) instruments mounted on Mars Science Laboratory (MSL) Curiosity rover. Mastcam is a two-instrument suite with left and right-eye cameras, and MAHLI is a single focusable camera located at the end of the rover's robotic arm. MSLNet1 is trained on classes including rocks, sand, sun, wheel tracks, and wheels [6]. If MSLNet1 predicts "other rover parts", the image will be passed through MSLNet2 for finer grained classification [7]. Running these classifiers directly onboard the rover could improve data collection and enable autonomous tasking.

These Mars classifiers were built with transfer learning from AlexNet [8]. Test images were 227x227 pixels. HiRISE images were grayscale, and MSL images were RGB. Models that are run on the Snapdragon DSP/NPU must be quantized (fixed point), and on the Myriad X, must be transformed to half precision floating point, which can lead to a classification discrepancy. Models are quantized using a separate validation dataset and discrepancies are reported on a held-out test set.

Benchmarking results are similar for all three classifiers; for brevity, we display only results for HiRISENet. Table 1 shows errors and power relative to a Linux run on the test laptop: MacOS 2019, 2.4GHz, 8-core, 3.1W idle, 94W max (*energy includes monitor and other externals). Inference time and energy consumption shown are per image. On the test laptop, the time reported is walltime. These low SWaP processors have only small errors, with up to 10x speed improvement.

3.2. Mars NavCam Image Segmentation

We also deploy an image segmentation model trained on imagery from the MSL rover's Navigation Cameras (NavCam) [9]. This model was developed to support MSL Rover Planners for hazard assessment and slip analysis, but could also be run onboard. The model was built using a DeepLabv3 architecture [10].

Table 2 shows the quantization discrepancy errors and inference times per image (images were 513x513 pixels). The errors on the Snapdragon DSP are relatively high. We were not able to pre-quantize the model and used run-time quantization, which increases the network initialization time, peak memory usage, and the model file size, as well as affecting quantization discrepancy. We do not show Myriad results, as the model had incompatible layers.

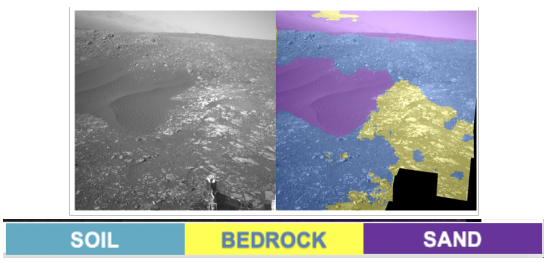

Figure 2: Mars MSL NavCam Imagery and Label

Table 2: Mars NavCam Image Segmentation Benchmarks

	% missed pixels	Inference Time
Linux Reference	-	1,886 ms
Snapdragon CPU	0.0 %	6,235 ms
Snapdragon GPU	0.4 %	2,233 ms
Snapdragon DSP	9.3 %	192 ms

3.3. UAVSAR Flood Mapping

In addition to models trained on Mars, we test earth-based models for image segmentation and super resolution. We demonstrate a model trained to perform pixel-wise binary classification of UAV polarimetric L-band SAR imagery, to predict areas that have been flooded. This model was trained on imagery of Houston, TX, USA, after flooding by Hurricane Harvey [11]. The net architecture follows a UNET [12] structure. These models could be used for onboard alert generation or surveillance. Table 3 gives quantization errors and timing; a subset of these results has been shown in ref [13], we add Linux, NPU, and Myriad benchmarks. Image patches were 64x64 pixels. A run time of 1.3 image patches/second is needed to meet real time, and this is met by all platforms. All error rates are small.

Table 3: UAVSAR Flood Mapping Benchmarks

	% missed pixels	# patches / sec
Linux Reference	-	25
Snapdragon CPU	0 %	20
Snapdragon GPU	0 %	162
Snapdragon DSP/NPU	0.4 %	391
Myriad X	0.7 %	167

3.4. Super Resolution for Spectroscopy

The models benchmarked above performed image classification or segmentation. Here, we look at a model for super resolution. Earth and planetary scientists use high-resolution spectral measurements for rock and mineral identification. This data tends to be sparse, compared with lower resolution data. A Deep Gaussian Conditional Model was created to infer high-resolution measurements from low-resolution ones [14]. It was trained to predict the Airborne Visible Infrared Imaging Spectrometer Next Generation (AVIRIS-NG) hyperspectral output from the 5-band Advanced Spaceborne Thermal Emission and Reflection Radiometer (ASTER) inputs. For this model, inference on the Snapdragon DSP/AIP is twice as slow as running on the Snapdragon CPU (45 vs 21 ms per input). This is most likely due to the small size of the model and single-pixel nature.

3.5. Standard Deep Learning Models

Transfer learning from pre-trained models is often used for model development. We benchmark some standard Keras [15] deep learning classification models, which may help with model selection for edge processing. All models were pre-trained on ImageNet [16], which has 1,000 different classes, and contains internet imagery. For testing, we use Imagenette [17], which is a much smaller dataset, containing only 10 classes. For brevity, in Table 4, we show classification discrepancy and timing using the Snapdragon NPU for a variety of standard models.

Table 4: Standard Deep Learning Models: Snapdragon NPU

	Errors	Inference Time	# Parameters
MobileNet	100%	60 ms	4,253,864
InceptionResNetV2	16%	56 ms	55,873,736
Xception	11%	53 ms	22,910,480
VGG19	4%	31 ms	143,667,240
InceptionV3	4%	27 ms	23,851,784
VGG16	2%	27 ms	138,357,544
ResNet50	6%	10 ms	25,636,712

The quantization discrepancy of MobileNet is very high; the structure of the net gives greater fluctuation ranges, which may make it not easily quantize-able [18]. Number of model parameters does not predict run time.

4. FUTURE WORK AND CONCLUSIONS

We have demonstrated the Myriad X and Snapdragon COTS processors for faster and lower power deep learning in space on the ISS. The Snapdragon DSP/AIP provides speed improvements over the CPU in all cases except the single-pixel network. Errors were low, except when using a run-time quantized network, as in NavCam. We continue to benchmark new applications, and are running memory checkers to quantify radiation effects on the processors. We also plan to benchmark results using Qualcomm's efficiency toolkit [19], which may improve network quantization. We have shown fast and accurate inference with these COTS processors and hope this will be a step towards a new era of powerful onboard autonomy with edge processing.

ACKNOWLEDGEMENTS

The research was carried out at the Jet Propulsion Laboratory, California Institute of Technology, under contract with the National Aeronautics and Space Administration. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise, does not constitute or imply its endorsement by the United States Government or the Jet Propulsion Laboratory, California Institute of Technology. This work was sponsored by the JPL Foundry and NASA's Earth Science Technology Office (ESTO). We would also like to thank the many application providers who also supported the testing of their applications.

REFERENCES

[1] "Snapdragon 855 mobile platform," 2020, https://www.qualcomm.com/products/snapdragon-855-mobile-platform

[2] "Intel Movidius Vision Processing Units," 2021, https://www.intel.com/content/www/us/en/products/details/processors/movidius-vpu.html

[3] Giuffrida, G, et al. The Φ-Sat-1 mission: the first on-board deep neural network demonstrator for satellite earth observation. IEEE Transactions on Geoscience and Remote Sensing, 2021.

[4] "HPE Spaceborne Computer," 2021, https://www.hpe.com/us/en/compute/hpc/supercomputing/spaceborne.html

[5] "Journey to Mars: How our collaboration with Jet Propulsion Laboratory fostered innovation", Qualcomm, 2021,https://www.qualcomm.com/news/onq/2021/03/17/journey-mars-how-our-collaboration-jet-propulsion-laboratory-fosterd-innovation

[6] K. Wagstaff, S. Lu, E. Dunkel, K. Grimes, B. Zhao, J. Cai, S.B. Cole, G. Doran, R. Francis, J. Lee, L. Mandrake, "Mars Image Content Classification: Three Years of NASA Deployment and Recent Advances," The Thirty-fifth AAAI Conf on Artificial Intelligence, pp. 15204-15213, 2021.

[7] K. Wagstaff, S. Lu, A. Stanboli, K. Grimes, T. Gowda, J. Padams, "Deep Mars: CNN Classification of Mars Imagery for the PDS Imaging Atlas", 13th AAAI Conf. Innovative Applications of Artificial Intelligence, pp. 7867-7872, 2018.

[8] A. Krizhevsky, I. Sutskever, G. Hinton, "ImageNet classification with deep convolutional neural networks," NIPS Proc of the 25th International Conference on Neural Information Processing Systems, pp.1097-1105, 2012.

[9] Deegan Atha, R. Michael Swan, Annie Didier, Zaki Hasnain, Masahiro Ono. "Multi-mission Terrain Classifier for Safe Rover Navigation and Automated Science," IEEE Aerospace Conf, 2022.

[10] L. Chen, G. Papandreou, F. Schroff, H. Adam, "Rethinking Atrous Convolution for Semantic Image Segmentation," ArXiv:1706.05587, 2017.

[11] M. Denbina, Z. Towfic, J. Thill, M. Bue, N. Kasraee, A. Peacock, Y. Lou, "Flood Mapping Using UAVSAR and Convolutional Neural Networks," IEEE Intl Geoscience and Remote Sensing Symposium, pp. 3247-3250, 2020.

[12] O. Ronneberger, P. Fischer, T. Brox, "U-net: Convolutional Networks for Biomedical Image Segmentation," Intl Conf. on Medical Image Computing and Computer-assisted Intervention, pp. 234-241, 2015.

[13] Z. Towfix, D. Ogbe, J. Sauvageau, D. Sheldon, A. Jongeling, S. Chien, F. Mirza, E. Dunkel, J. Swope, V. Cretu, M. Ogut, "Benchmarking and Testing of Qualcomm Snapdragon System-on-Chip for JPL Space Applications and Missions," IEEE Aerospace Conf, 2022.

[14] A. Candela, D.R. Thompson, D. Wettergreen, K. Cawse-Nicholson, S. Geier, M.L. Eastwood, R.O. Green, "Probabilistic Super Resolution for Mineral Spectroscopy". Proc AAAI Conf. on Artificial Intelligence Vol. 34, No. 08, pp. 13241-13247, 2021.

[15] "Keras Applications," Chollet et al, 2021, https://keras.io/api/applications/

[16] J. Deng, W. Dong, R. Socher, R. L.-J. Li, K. Li, L. Fei-Fei, "Imagenet: A large-scale hierarchical image database,", IEEE conference on computer vision and pattern recognition, pp. 248–255, 2009.

[17] "Imagenette," FastAI, 2021, https://github.com/fastai/imagenette

[18] S. Yun, A. Wong, "Do all MobileNets Quantize Poorly? Gaining Insights into the Effects of Quantization on Depthwise Separable Convolutional Networks Through the Eyes of Multi-scale Distributional Dynamics", CVPR 2021.

[19] "AI Model Efficiency Toolkit", Qualcomm, 2021, https://developer.qualcomm.com/software/ai-model-efficiency-toolkit