AN2DL - Second Homework Report
MiLi

Martim Bento, Henrique Pocinho, Salvatore Santolupo, Diego Viganó

Martim Rendeiro Bento, pocinho, s_salvo_, diegosdigos

11068539, 11068532, 10766335, 10739933

December 14, 2024

1 Introduction

The segmentation and interpretation of the Martian surface is crucial for Mars exploration. It plays a pivotal role in the execution of autonomous navigation during Mars rover missions by providing the necessary data for path planning and obstacle avoidance. Accurate surface segmentation ensures the successful execution of these tasks, enhancing the efficiency and safety of rover operations.[1–4]

In this context, as part of the second homework assignment for the AN2DL course, several 64×128 gray-scale real images of Mars terrain were provided. Each pixel in these images is categorized into one of five classes, representing distinct terrain types. The task is framed as a **multi-class semantic segmentation** problem, where the goal is to develop a model that, *given a set of desired outputs, accurately assigns the correct class label to each pixel in new image inputs.*

To address this challenge, a **U-Net** based model was initially designed and then iteratively refined. Concepts like **class weights**, **bottlenecks**, **dilated convolutions**, **augmentation pipelines** and **loss functions** were explored to find the best combination. Furthermore, some of our exploration was based on state-of-the-art architectures like DeepLabV3+ and other relevant literature, so layers like **Atrous Spatial Pyramid Pooling** (ASPP) and **transformers** were tested. [5, 6]

2 Problem Analysis

The main challenge is to develop and train a robust model capable of performing multi-class semantic segmentation for gray-scaled images of Martian soil.

As for the dataset, both the training and test dataset were provided by the course faculty. The training set contains 2615 images and their respective labels, and the test set contains 10022 unlabeled images, each being a 64×128 gray-scaled image. The dataset was thoroughly analyzed in search of possible outliers and 110 were found. Fig. 1 shows two images, along with their respective label maps, and clearly shows one of the 110 unwanted extraterrestrial visitors.

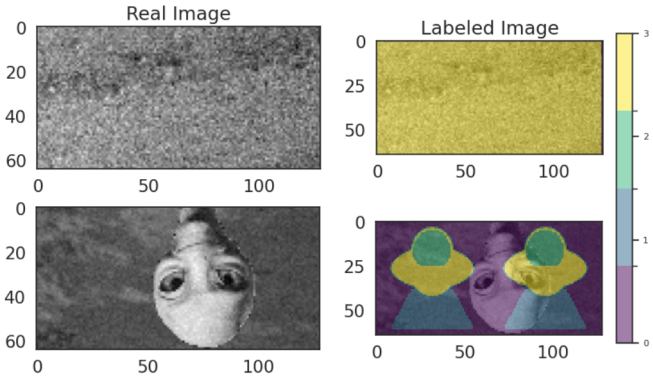

Figure 1: Examples of images and their respective labels. The pair on top is Martian soil, while the pair below is a visitor from outer space.

Regarding initial conditions, the seed for all random events was fixed to guarantee repeatability, using the magical number, 42 [7].

3 Method

As a first approach, a simple U-Net model was employed, leveraging two down sampling and two up sampling blocks, linked by a bottleneck. This model served as a baseline and was iteratively improved upon by taking inspiration from other state-of-the-art networks such as DeepLabV3+ [5]. The final model is illustrated in Fig. 2.

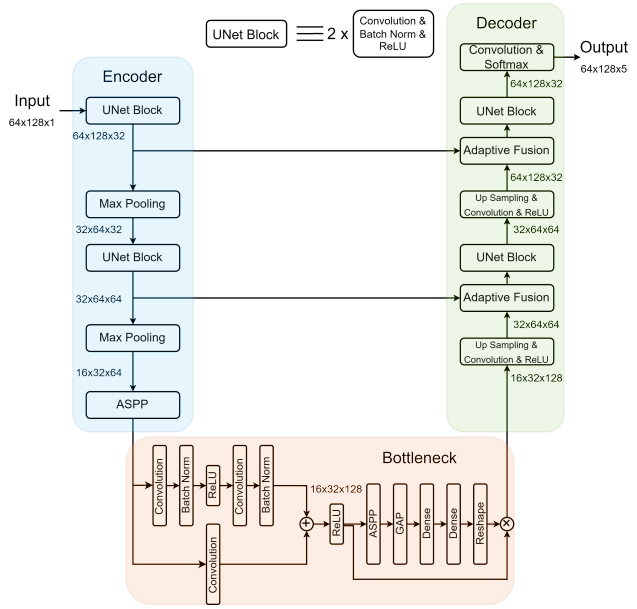

Figure 2: Diagram of the implemented network.

3.1 Network Structure

The network makes use of ASPP, a layer composed of several dilated convolutions and a global average pooling (GAP) in parallel which are then concatenated [5]. Thus, it allows the network to keep a constant stride while expanding the receptive field, enabling a larger field of view without increasing the number of parameters or computational cost.

The bottleneck of the developed model is a complex structure that employs several mechanisms, namely a residual path and a squeeze-and-excite (SE) block. The residual path eases gradient flow, preventing vanishing gradients during training. The SE block was added with the goal of allowing the model to focus on the most informative features, while disregarding irrelevant noise [8]. The model employs an additional ASPP block within the SE block since it showed good results. Additionally, batch normalization layers were used to help with regularization and keeping the training stable. Also, the junction of the different paths in the bottleneck doesn't use concatenation but addition and multiplication, this was again adjusted experimentally, to find what worked better.

This customized U-Net also employs two adaptive fusion (AF) layers to dynamically combine multiple feature maps. Each layer generates attention weights, applies them to each feature map and adds both weighted maps together. The AF layers effectively combine multi-scale features, allowing the model to adaptively decide how much importance to give to each feature map [9].

3.2 Training

The available labeled set was partitioned into a training and a validation set, with an 80% to 20% split, respectively. The test set was used to evaluate the network's performance and generate the submission file.

Various parameters, such as patience, learning rate, loss functions, optimizers and learning rate schedulers were tested. Among the optimizers, AdamW yielded the most promising results. Regarding the loss functions, categorical cross entropy, focal loss and dice loss, both weighted and unweighted, were used. Several weights were also tested, taking the severe class imbalance into account, since the class "Big Rock" was only present in about 60 images. The final model used a weighted categorical cross entropy (WCCE) function with a manually tuned weight vector.

3.3 Data Augmentation

Several augmentation pipelines were tested, both geometric and chromatic. The final pipeline employed Gaussian noise, contrast, brightness, random flips and random 180° rotations, with the geometric transformations applied to both the image and the respective label. The parameters of such layers were also finely adjusted to obtain the best results possible.

4 Experiments

During training three different metrics where used. The loss function was used to calculate the back-propagation and train the model. Initially accuracy was used for the early stopping but was quickly replaced with the Mean Intersection over Union (IoU). Mean IoU measures the overlap between the detected object area and the actual object area, divided by their union. Thus, some Mean IoU notable results obtained during the development are displayed in Table 1.

Table 1: Notable Models and respective Mean IoU results, locally and on Kaggle.

Model	Local [%]	Kaggle [%]
1. U-Net	42.51	41.866
2. Simple Aug	44.68	44.088
3. Bottleneck	46.43	45.007
4. ASPP	48.08	48.118
5. WCCE	50.87	53.724
6. Tuned WCCE	62.17	63.349
7. Adaptive Fusion	**67.13**	**65.959**
8. LR Scheduler	61.29	64.469
9. Transformer	62.25	59.011

Model `1` is the baseline. For simplicity, it used a sparse categorical crossentropy loss function during training, and only had random horizontal flips as augmentation. For model `2` more augmentation layers, such as contrast and brightness were added. Next, on model `3`, the bottleneck was improved. The next few optimizations were the most impactful. Model `4` incorporated ASPP, and implemented the final augmentation pipeline, now integrating gaussian noise. Models `5` and `6` showcase the improvements obtained by using different loss functions. Model `7` integrated the AF layers and scored the highest. Nevertheless other approaches were tested, through models `8` and `9`, making use of a learning rate scheduler and transformers, respectively. Both provided promising results, despite not surpassing model `7`.

5 Results and Discussion

From a direct comparison between models `1`, `3`, `4` and `7` it is evident that the main blocks incrementally added , namely the improved bottleneck, the ASPP and the AF layers, refine the model, making it more robust and providing a performance enhancement. The benefit of a robust augmentation pipeline is observed, by comparing model `1` with models `2` and `4`, which introduced the augmentation layers used by the final model.

More noticeable is the impact caused by the chosen loss function. As stated previously, the dataset provided features severe class imbalances, so as to mitigate this several loss functions were tested, providing significantly different performances. Models `5` and `6` both show a significant improvement compared to the previous models and their difference shows the importance of carefully tuning the weights. Model `5` used proportional weights and model `6` used the following manually tuned weight vector, [0.01, 0.2, 0.5, 1.2, 3.2], thus showing the importance of using an adequate weight vector.

So, as of now, before the final leaderboard is revealed, model `7`, which uses all of these considerations, achieves a mean IoU of **65.959%** over 25% of the test, according to the Kaggle platform.

Finally, it is also important to note that, although the last two models did not surpass the previous attempts, they provided promising results, suggesting that further tuning of these ideas might improve the performance.

6 Conclusions

The challenge was to develop a multi-class semantic segmentation model capable of accurately classifying different martian terrain types. In this work, a model based on a U-Net is presented. It improves the base line by leveraging ASPP layers, a carefully built bottleneck, adaptive fusion, WCCE loss with manually tuned weights and a diverse augmentation pipeline, featuring gaussian noise, as well as geometrical and chromatic augmentations. Thus, the best model was able to achieve a mean IoU of **65.959%** over 25% of the test set.

From this work, it is clear that addressing class imbalance is essential for achieving high mean IoU. The tuning of the weight vector for the loss function could warrant future work. Furthermore, some other architectural techniques, namely transformers and learning rate schedulers, show promise for further improving the model's performance, given more tuning.

As a final note, the workload was distributed evenly among team members, each contributing to both model development and the report.

References

[1] J. Li, K. Chen, G. Tian, L. Li, and Z. Shi, "MarsSeg: Mars Surface Semantic Segmentation with Multi-level Extractor and Connector," Apr. 2024, arXiv:2404.04155 [cs]. [Online]. Available: http://arxiv.org/abs/2404.04155

[2] F. Mohammad, Y. Gao, S. Kay, R. Field, M. De Benedetti, and E. V. Ntagiou, "Deep Learning based Semantic Segmentation for Mars Rover Terrain Classification," in *2024 International Conference on Space Robotics (iSpaRo)*. Luxembourg, Luxembourg: IEEE, Jun. 2024, pp. 292–298. [Online]. Available: https://ieeexplore.ieee.org/document/10687827/

[3] Y. Xiong, X. Xiao, M. Yao, H. Cui, and Y. Fu, "Light4mars: A lightweight transformer model for semantic segmentation on unstructured environment like mars," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 214, pp. 167–178, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0924271624002466

[4] H. Liu, M. Yao, X. Xiao, and H. Cui, "A hybrid attention semantic segmentation network for unstructured terrain on Mars," *Acta Astronautica*, vol. 204, pp. 492–499, Mar. 2023. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0094576522004064

[5] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Lecture Notes in Computer Science*. Springer International Publishing, 2018, pp. 833–851. [Online]. Available: http://dx.doi.org/10.1007/978-3-030-01234-2_49

[6] G. Goutham, H. Juneja, A. C, and V. R. B. Prasad, "Semantic segmentation on martian terrain for navigation using transformers," in *2022 IEEE 7th International Conference on Recent Advances and Innovations in Engineering (ICRAIE)*, vol. 7, 2022, pp. 276–282.

[7] D. Adams, *The Hitchhiker's Guide to the Galaxy*. Pan Books, 1979.

[8] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu, "Squeeze-and-excitation networks," 2019. [Online]. Available: https://arxiv.org/abs/1709.01507

[9] A. K. Singh, D. Chaudhuri, M. P. Singh, and S. Chattopadhyay, "Integrative cam: Adaptive layer fusion for comprehensive interpretation of cnns," 2024. [Online]. Available: https://arxiv.org/abs/2412.01354