AN2DL - Second Homework Report
Cortex Creators

Beatrice Adamini, Davide Cattivelli, Giorgio Coccapani, Kalle Engblom

beaadamini, catti01, giorcocc, kalleengblom

251824, 259084, 259839, 276779

December 13, 2024

1 Introduction

This project aims to address a *multi-class semantic segmentation* challenge through the deployment of a **U-Net-based deep learning model** [1]. The focus of the segmentation is on images of the Martian surface, with our model development primarily concentrating on the following areas of interest:

- a detailed analysis of the provided dataset, that was previously manually annotated
- the definition of a model able to identify and classify the various regions of the image

2 Problem Analysis

The provided dataset was divided into two sets: the first one included the training set, which we further split into calibration and validation sets, while the second comprised the test set. The images were sized at 64×128 pixels with a single channel, displaying pixel values on a grayscale. An example from the dataset is shown in Figure 1. The training images came with an associated mask indicating the ground truth for segmentation. Each pixel in the mask was assigned an integer from 0 to 4, representing five classes: *background*, *soil*, *bedrock*, *sand*, and *big rock*.

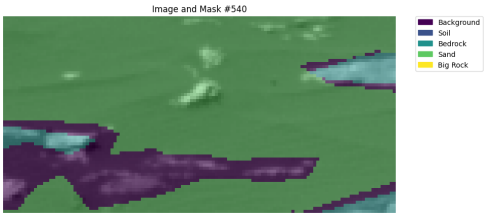

Figure 1: Example of image and mask (overlap)

First, we noticed that there were artifacts in the training set, so we removed every image presenting them, down-sampling data from 2615 to 2505 images and masks. Then, we explored the distribution of the pixels contained in the images and we found that the 5 classes were heavily unbalanced. We'll go further into this topic in section 3.2.

3 Method

3.1 Data Augmentation

Given the limited number of samples in our dataset, data augmentation was employed to increase its diversity. Care was taken to apply the same transformations to both the images and their corresponding masks. To achieve this, we implemented a `RandomAugmentationPipeline` (see [2]) consisting solely of geometric transformations. These transformations were applied stochastically to both components of each sample of the training set.

3.2 Data Balancing

To deal with class imbalance, we employed several strategies. Initially, we developed a function to remove samples with masks containing only one class, therby promoting more effective learning across all classess. Subsequently, we concentrated on the most underrepresented class, selecting images with at least 1% of pixels from this class. The total images found were around 50, so not enough to really persue a solution based on them. For this reason, we implemented a function that was able to generate new samples from the provided ones through transformations, targeting the areas of interest. This was used especially for the ensamble model, as explained in Section 4.1, and the generated samples were characterized by a different distribution, as shown in Figure 2.

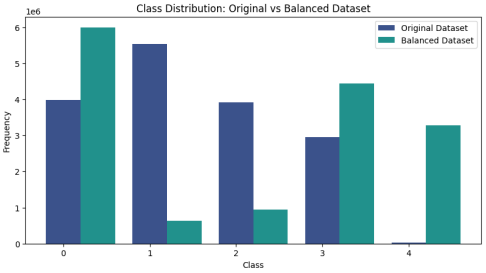

Figure 2: Distribution of pixel classes of the original dataset (blue) and the generated one (light-blue)

3.3 Model Definition

The proposed model (Advanced Bottleneck in Table 1) is designed for accurate segmentation tasks and integrates techniques for multi-scale feature extraction, channel-wise attention, and the fusion of global and local features. It is composed of two interconnected U-Net architectures, referred to as the *Macro U-Net* and the *Micro U-Net*.

The *Macro U-Net* architecture employs a conventional U-Net design with a downsampling path to capture global context by extracting coarse features. It integrates advanced fusion methods like dilated convolutions and attention mechanisms at the bottleneck to improve context comprehension. The upsampling path refines these features and reconstructs the output's spatial structure, aiming for a general segmentation map that captures broad patterns, though it may not be highly detailed.

On the other hand, the *Micro U-Net* works with higher resolution input data to preserve fine details and textures. It reduces excessive compression by using fewer, larger filters and includes a Global Context Module at the bottleneck to improve segmentation of key classes, such as *big rock*, and to address specific errors. Skip connections ensure spatial accuracy by directly transferring data from the encoder to the decoder.

To combine the outputs of the *Macro* and *Micro U-Nets*, a dedicated fusion module was designed. This module blends the features from both architectures, balancing global and local information to improve the model's segmentation accuracy. Adaptive weighting and cross-enhanced feature mechanisms are employed to ensure an effective integration of the different feature types.

The model also incorporates several custom modules in the bottlenecks. The Squeeze-and-Excitation block [3] adjusts channel-wise feature importance by applying an attention mechanism to enhance critical channels. The Dilated Inception block uses dilated convolutions with varying dilation rates to extract features at multiple scales, capturing both fine details and global context. The Global Context Module captures dependencies across the entire image, enriching feature maps with global information. Additionally, a Cellular Automata Module [4] refines features through iterative local interactions, enhancing the details in the final feature maps.

4 Experiments

The performance of our model fell short of expectations, leading to a series of experiments aimed at improving segmentation. Adjustments to data augmentation and increased sampling for underrepresented classes yielded limited success.

We then focused on the loss function, moving away from Sparse Categorical Crossentropy [5] to design a custom loss incorporating class-specific weights and dynamic Focal Loss [6]. Unfortunately, this approach failed to produce satisfactory results across all model variations.

4.1 Ensemble Model

To address class imbalance, we developed a simpler model specifically for detecting the least represented class (class 4: *big rock*) and integrated it into the main architecture.

Table 1: Model evaluation during the experiments. The chosen model is marked in bold.

Model Name	accuracy	loss	mean_iou	val_acc.	val_loss	val_mean_iou
Single U-Net	0.7727	0.9017	0.4305	0.7520	1.0279	0.4527
Double U-Net	0.7499	0.9856	0.4217	0.7025	1.1725	0.4132
Weighted Loss	0.7099	0.2471	0.4117	0.7411	0.2299	0.4414
Advanced Bottleneck	**0.7594**	**0.6311**	**0.5854**	**0.7681**	**0.6084**	**0.5128**
Specialized Model[1]	0.9010	0.2996	0.6603	0.9517	0.1569	0.7474
Ensemble Model[2]	0.7675	0.6680	0.5892	0.7786	0.6410	0.5272

[1] Results obtained by using the double U-Net model with more artificial samples from class 4
[2] Model obtained by combining Specialized Model and Advanced Bottleneck

The dataset for this model was enhanced by generating samples twice, as described in Section 3.2. This produced approximately 8000 samples, split into training and validation sets.

The simpler model, a streamlined U-Net, performed well on this dataset (Table 1), and its weights were saved for integration into the advanced architecture. The ensemble model combined the advanced bottleneck design with the simpler U-Net, using the latter's weights to guide segmentation improvements. However, this combined approach did not achieve the desired results, highlighting the complexities of merging models to resolve segmentation challenges.

5 Results and Discussion

The results obtained with the model described in Section 3.3 are presented in Table 2, achieving a benchmark score of 0.52828.

Class	Precision	Recall	F1-Score	IoU
0	0.68388	0.50230	0.57919	0.40765
1	0.79406	0.91854	0.85178	0.74182
2	0.75136	0.75027	0.75082	0.60105
3	0.80920	0.85677	0.83231	0.71278
4	0.33361	0.21157	0.25893	0.14872

Table 2: Model scores on internal test set

The results indicate a good distribution of the Mean IoU values across most categories, except for class 4 (*big rock*). This observation, reinforced by the confusion matrix in Figure 3, highlights the model's significant difficulty in accurately identifying large rocks, despite efforts to balance the underrepresented classes by generating additional samples and a model aimed at recognizing small characteristics in the images.

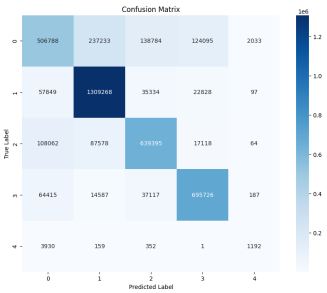

Figure 3: Confusion matrix

6 Conclusions

The aim of this project was to classify Mars terrain images into five categories using a U-Net based deep learning model. We thoroughly analyzed the dataset, created a data augmentation pipeline, and implemented strategies to counteract dataset imbalance. Despite diligent efforts, the inference results did not meet our expectations, with a leaderboard score of 0.52828, suggesting the need for improvement. Further dataset refinement with precise transformations and new class differentiation techniques may enhance performance. Considering a new model architecture or different methodologies could also be beneficial.

Special thanks to Beatrice for her ensemble model, Davide for enhancing the primary model, Giorgio for his data analysis and balancing efforts, and Kalle for... well, nothing [and L. for the enigmatic poetry provided, half the project was decoding those masterpieces to be encoded in a useful way].

References

[1] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. "U-Net: Convolutional Networks for Biomedical Image Segmentation". In: *CoRR* abs/1505.04597 (2015). arXiv: 1505.04597. URL: http://arxiv.org/abs/1505.04597.

[2] Keras. *RandomAugmentationPipeline layer*. en. https://keras.io/api/keras_cv/layers/augmentation/random_augmentation_pipeline/.

[3] Jie Hu, Li Shen, and Gang Sun. "Squeeze-and-Excitation Networks". In: *CoRR* abs/1709.01507 (2017). arXiv: 1709.01507. URL: http://arxiv.org/abs/1709.01507.

[4] Abhishek Dalai. *Layered Cellular Automata*. 2023. arXiv: 2308.06370 [nlin.CG]. URL: https://arxiv.org/abs/2308.06370.

[5] *Probabilistic losses*. en. https://keras.io/api/losses/probabilistic_losses/. Accessed: 2024-12-13.

[6] Tsung-Yi Lin et al. *Focal Loss for Dense Object Detection*. 2018. arXiv: 1708.02002 [cs.CV]. URL: https://arxiv.org/abs/1708.02002.