AN2DL - Second Homework Report
regionaleveloce

Simone Calzolaro, Gabriele Clara Di Gioacchino, Tommaso Galimberti, Sara Massarelli

simonecalzolar0, gabrieleclara01, tomgalimberti, saramassarelli

260121, 259518, 259629, 249387

December 14, 2024

1 Introduction

In this project, we address a challenging semantic segmentation task focused on Mars terrain imagery. Specifically, our objective is to develop a neural network capable of precisely classifying each pixel in 64x128 grayscale images into one of five distinct terrain categories. Our main goals were:

- Accurate semantic segmentation to correctly assign the correct label to each pixel
- Efficient training process for optimize the computational resources at our disposal
- Model evaluation and validation to consistently assess model's performance

2 Problem Analysis

2.1 Dataset characteristics

64x128 pixel grayscale Martian terrain images with five distinct terrain classes: 0:Background; 1:Soil; 2:Bedrock; 3:Sand; 4:Bigrock.
The original dataset was composed by 2615 samples along with the corresponding masks. We noticed the presence of outliers and after their elimination it ended up with 2505 samples (and masks).

2.2 Main challenges

The primary challenges centered on managing severe class imbalance, particularly with class 4, and handling the limited sample size. Subtle terrain class distinctions demanded precise feature discrimination. Identifying the appropriate semantic segmentation model architecture emerged as a critical research objective. Balancing local detail preservation with global context understanding required careful design of network blocks. Additionally, developing strategic image augmentation became crucial to expanding dataset variability and model generalization.

2.3 Initial assumptions

Recognizing the complexity of Martian terrain classification, we believed a well-designed neural network could distinguish between Martian terrain types. We anticipated that advanced feature extraction and preprocessing techniques would be essential for model performance. Intentionally, we decided to exclude the background class from our training process, focusing exclusively on the four primary terrain types to enhance model specificity and learning efficiency.

3 Method

3.1 Elimination of the outliers

Our initial preprocessing step focused on identifying and removing outliers from the dataset. Upon analysis, we observed that all outliers exhibited a distinctive mask pattern. To streamline their removal, we identified the hash associated with one of these masks and systematically excluded all samples and masks corresponding to the same hash. This method effectively ensured the elimination of redundant and anomalous data points. As a result, the dataset was refined from 2,615 to 2,505 samples. This preprocessing step significantly enhanced the dataset's quality and consistency, establishing a more reliable foundation for subsequent analysis and model development.

3.2 Individuation of the best architecture

Identifying the optimal architecture was a challenging process, requiring multiple iterations to determine the most suitable design. Ultimately, we concluded that a U-Net architecture performed best for our task. We initially implemented a basic U-Net consisting of three down-sampling blocks and corresponding symmetric up-sampling blocks. The final architecture represents an evolved version of this initial design, refined to improve performance and address the specific requirements of the problem.

3.3 Individuation of the best loss function

Selecting the appropriate loss function was critical for achieving optimal model performance. After experimenting with several options—including Dice Loss, Tversky Loss, Focal Loss, and Boundary Loss—we found that the weighted Categorical Focal Crossentropy outperformed the others. This superior performance was primarily due to the significant class imbalance, particularly involving class 4, and the exclusion of class 0 from the evaluation. The weighted Categorical Focal Crossentropy allowed us to assign effective weights to each class, significantly boosting the model's performance. Notably, combining multiple loss functions did not yield better results compared to using the weighted Categorical Focal Crossentropy alone.

3.4 Data Augmentation

Data augmentation played a key role in enhancing model performance. We experimented with various augmentation techniques to increase the size of the training dataset. However, augmentations involving contrast, brightness, and color adjustments proved problematic, as they often resulted in images becoming excessively dark. Consequently, we opted for simpler geometric augmentations, specifically horizontal and vertical flipping.

These augmentations were applied to both the samples and their corresponding masks. As a result, the final training dataset expanded to 6,762 samples and masks, with each augmentation applied to every image and subsequently concatenated to the original dataset. Using datasets that were significantly larger or smaller than this led to a decline in model performance.

3.5 Boosting the U-net

This step required extensive study and research. Building upon the previously mentioned initial U-Net, we focused on incorporating residual blocks and attention mechanisms. Our final architecture evolved into an AttentionResU-Net.

The ResU-Net block follows this structure: Conv2D → BatchNorm → Activation → Conv2D → BatchNorm → Shortcut → BatchNorm → (Shortcut + BatchNorm) → Activation

The resulting U-Net consists of four ResU-Net blocks in the downsampling path, mirrored by corresponding layers in the upsampling path. The bottleneck is also a ResU-Net block. To improve feature selection, we integrated gating and attention blocks along the upsampling path. Specifically, the attention block functions as an Attention Gate, taking inputs from both the gating block and the corresponding downsampling path.

Additionally, we enhanced the bottleneck by incorporating a squeeze-excitation block [1] followed by a pyramid pooling block after the ResU-Net block. This combination helped extract more relevant features, significantly boosting the model's performance.

4 Experiments

In this section, we evaluate the performance of various model configurations for the segmentation task. We compare the baseline U-Net model with enhanced versions that incorporate additional techniques. Enhanced versions of the U-Net are denoted by the ”+” symbol, indicating that the baseline model is extended with the specified technique.
The evaluation metrics used are **Accuracy (Acc)** and **Mean Intersection over Union (MeanIoU)**.
The results are summarized in the table below, where the best-performing results are highlighted in **bold**.

Table 1: Different models with related metrics. Best results are highlighted in **bold**.

Model	**Acc**	**MeanIou**
Unet (Basic)	76.04	47.4
Unet (+ Focal Loss)	78.59	54.4
Unet (+Attention Gate)	73.2	71.2
Unet (+Squeeze Excitation)	73.5	74.4
Unet (+Pyramid Pooling)	**75.5**	**74.6**

5 Results

The results suggest that incorporating both squeeze-excitation and pyramid pooling provides significant performance gains, validating the architectural changes made to the base model. The improvements in Mean-IoU indicate that the models were better able to identify class boundaries, an essential feature for semantic segmentation of Martian terrain images.

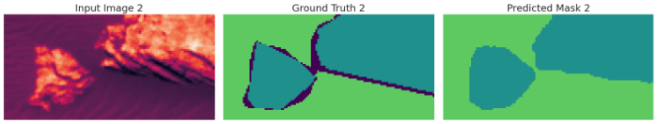

Figure 1: This is an example of prediction mask.

6 Discussion

The **Squeeze Excitation** mechanism increases the model’s ability to focus on important features, while **Pyramid Pooling** helps capture multi-scale contextual information, leading to the highest accuracy and MeanIoU. However, precision for class 4 remains lower than expected. This may be due to overlapping features with other classes or an insufficient number of training samples for this class. A significant weakness is the exclusion of the background class (class 0) during training, which leads to misclassifying background pixels as one of the foreground classes. This affects overall segmentation performance and increases confusion between class labels.

7 Contributions

We all contributed on training and testing the models, here more details are included.
Tommaso Galimberti inspected and cleaned the dataset removing the outliers.
Gabriele Clara Di Gioacchino focused on the research and on testing different types of loss.
Simone Calzolaro researched, studied and implemented the different model architectures, testing them to find the most suitable model for our task.
Sara Massarelli worked on testing various types of augmentations and focused on generating new images in order to increase the size of the training dataset.

References

[1] J. Hu, L. Shen, and G. Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7132–7141, 2018.