POLITECNICO
MILANO 1863

AN2DL - Second Homework Report
OverfittingNinjas

Alessandro Griffanti, Andrea Ciarallo, Eleonora Bellè, Luca Masiero

griffa, mose17, Eleonorabelle, Luca_pdn

250114, 226500, 225900, 245458

December 14, 2024

1 Introduction

This project focuses on semantic segmentation using deep learning techniques. The main objective is to design a robust model that assigns a class label to each pixel of 64x128 greyscale images concerning the Mars terrain.

Our approach, that will be described in detail in Section 3, utilizes a U-Net architecture enhanced with a custom loss function, data augmentation and attention mechanisms.

2 Problem Analysis

The dataset used in this project comprises a total of 12,637 greyscale images, each with a resolution of 64×128 pixels. The dataset is divided as follows:

- 2,615 images are included in the training set, each paired with a corresponding segmentation mask that assigns every pixel to one of the available classes: *background, soil, bedrock, sand, bigrock*;
- 10,022 images constitute the test set, for which the segmentation masks are not provided;

At the outset, we assumed that the U-Net architecture would effectively address the segmentation task due to its ability to combine spatial and contextual information. We also anticipated the presence of outliers in the dataset, which would need to be identified and removed to maintain data integrity and enhance the model's reliability. Additionally, we recognized the likelihood of an imbalanced dataset, which would require strategies such as class weighting and data augmentation to create a more diverse and representative set of training samples.

After analysing the dataset, our initial assumptions were immediatelly confirmed: we noticed some corrupted images where the original visuals are overlaid with unrelated content, introducing noise that needs to be addressed to ensure data integrity.
In addition, one of main challenges encountered is class imbalance, as some classes, in particular the *big rock* one, are underrepresented, which makes it difficult for the model to learn how to classify certain pixels.

3 Method

The following approach represents our best performing solution, though we also evaluated a range of alternative strategies, which we will discuss in detail in Section 5.
The proposed approach utilizes a U-Net architecture enhanced with targeted modifications to better suit the dataset's characteristics. Specifically, to

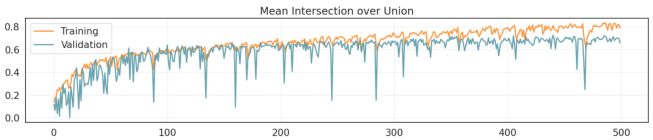
address the problem of class imbalance presented in Section 2, we employed a weighted sparse categorical cross-entropy loss function, giving greater importance to underrepresented classes except for the background class, whose weight has been set to zero to focus entirely on the relevant target classes. We also integrated channel attention mechanisms into the U-Net to refine the network's ability to focus on the most relevant features within the feature maps. While its overall impact on performance was modest, it contributed additional insight into the model's focus during segmentation. In addition, to prevent overfitting, we applied a rigorous data augmentation strategy that included geometric transformations such as flips, shifts, and zooms. These techniques expanded the model's generalization capabilities allowing it to achieve better results on the test data. Performance was measured using the mean Intersection over Union (mIoU), a widely used metric for segmentation tasks, defined as:

$$\frac{1}{|C|}\sum_{c\in C}\frac{\mathbb{1}(y=c)\wedge\mathbb{1}(\hat{y}=c)}{\mathbb{1}(y=c)\vee\mathbb{1}(\hat{y}=c)} \tag{1}$$

where C represents the set of classes, y represents the ground truth labels and $\hat{y}$ denotes the model predictions. According to this metric, our model achieved a **70.66%** on the hidden test set reflecting its robustness and effectiveness in segmenting the target classes. Notably, the mean Intersection over Union scores on the validation and test sets were virtually identical, indicating minimal to no overfitting and a strong generalization capability.
Detailed improvements attributed to each technique are presented in the next Section.

4 Results

Our model's results significantly surpass those of simpler approaches, such as a basic custom U-Net network with a shallow depth and without a custom loss function. This simple model, achieving a 39.94% mIoU on the test set, served as our starting point and we used it to understand the initial performance and as a baseline.
The first turning point was the introduction of a custom loss function: a weighted sparse categorical cross-entropy, designed with weights inversely proportional to the number of pixels in each class, which boosted the performance up to 50%. Optuna was also employed to research the best hyperparameters, including filter size and numbers, network depth and learning rate, allowing us to reach a 51% on the test set.
A major turning point was the decision to eliminate the weight attributed to the background class, allowing the model to focus solely on the relevant classes. This adjustment improved the mIoU to 60%. However, this also revealed signs of overfitting, as the results between the validation and test sets began to diverge.
To approach this issue, an extensive data augmentation pipeline was employed. This solution, along with channel attention mechanisms, increased the performance on the validation set, as shown in the following figure:



up to 72.46% allowing us to achieve **70.66%** mIoU on the test set, indicating a minimal and neglibile overfitting and demonstrating the effectiveness of our modifications.

Table 1: Key steps performance improvements

Key Improvements	Validation mIoU[%]	Test mIoU[%]
Baseline Net	39.65	39.94
Weighted Loss	48.06	50.18
Null background weight	70.46	60.18
Augmentation	**72.46**	**70.66**

5 Experiments and unexpected outcomes

Not all the approaches yielded the desired results; the following unexpected outcomes were observed:

- **Dataset cleaning**
  Upon inspecting the training set, we identified certain samples that might negatively impact the performance of our model, besides those containing the image of the alien. These include images where the majority of the frame is occupied by the rover chassis or cases where the mask does not appear to correspond accurately to the image. However, removing these samples led to a significant drop in performance. We hypothesize that the primary cause is the reduction in the overall size of the training dataset, which was already limited and became even smaller after discarding these samples. As is well-known, smaller datasets increase the complexity of the learning process and hinder the model's ability to generalize effectively.

- **Focal Loss:**
  The focal loss alone did not produce satisfactory results compared to the sparse categorical cross-entropy, which was employed in most of the other models we tested. This suggests that focal loss might be better suited for use in conjunction with other loss functions rather than as a standalone approach.

- **Bottleneck Refinement:**
  The techniques applied to refine and enhance the bottleneck of our model, such as *squeeze-and-excitation* and *residual blocks*, did not yield the expected improvements. On the contrary, they appeared to degrade the latent representation of the images, leading to poorer predictions of the segmentation mask.

- **State-of-the-Art Backbones:**
  Using state-of-the-art neural networks, such as ResNet or VGG, as backbones for the encoder architecture of our custom U-Net with randomly initialized weights did not enhance performance. This underscores the critical importance of transfer learning, particularly the use of pre-trained weights, which we were restricted from employing in this project.

6 Conclusions

While our approach achieved a satisfactory result, there is room for further improvement:

Exploring Different Architectures:
Different architectures, such as Linknet and PSP-Net, might be more suited for our specific task and give slightly different and better results.

Networks Ensemble:
An alternative and different approach to address the problem discussed in this report could involve using multiple networks (potentially more than two) trained in parallel with different loss functions, and combining their predictions. For instance, one network could focus on fine-grained details within the image, while another emphasizes the broader context. This complementary focus may lead to more accurate and insightful predictions.

Pre-trained Backbone:
Another potential improvement is incorporating a pre-trained backbone, such as ResNet or EfficientNet, *initialized with ImageNet weights*. This approach not only would accelerate training but also improve the performance leveraging robust feature extraction.

Contributions:

- Alessandro Griffanti:
  Dataset cleaning, weighted sparse categorical cross-entropy, Optuna exploration, attention mechanisms experimentation, tests with data augmentation (shear, gaussian noise, shifts), tests with pre-defined backbones;

- Andrea Ciarallo:
  Unified loss, Tversky loss, Dice loss;

- Eleonora Bellè:
  Data augmentation exploration;

- Luca Masiero:
  Dataset cleaning, bottleneck refinement exploration, attention mechanisms experimentation, tests with data augmentation (zoom, contrast).

References

- *Enhancing U-Net with Spatial-Channel Attention Gate for Abnormal Tissue Segmentation in Medical Imaging.* Retrieved from https://www.mdpi.com/2076-3417/10/17/5729
- *Segmentation models.* Retrieved from https://github.com/qubvel/segmentation_models