POLITECNICO
MILANO 1863

AN2DL - Second Homework Report
NeuralNexus

Andrea Codazzi, Alessio Onori, Pasquale Serrao, Anna Simeone

andreacodazzi, ironoa, pasqualeserrao, annasime1

247291, 224955, 250803, 245744

December 14, 2024

1 Introduction

This project addresses the challenge of **semantic segmentation**, where the focus is on developing a robust deep learning model capable of accurately segmenting Martian terrain images into five classes: *Background, Soil, Bedrock, Sand, and Big Rock*. The approach involves experimenting with **different network depths** and architectural variations, implementing **regularization strategies** such as dropout, and utilizing **learning rate schedulers to optimize training**. Furthermore, the integration of attention mechanisms and Atrous Spatial Pyramid Pooling (ASPP) aims to enhance feature extraction and drive better segmentation performance. The model is trained using different **loss functions**, designed to address class imbalance and improve the segmentation accuracy for underrepresented terrain types.

2 Problem Analysis

This section provides an analysis of the key aspects of the problem, focusing on the dataset, its main challenges, and initial assumptions.

2.1 Dataset Characteristics

The dataset consists of grayscale images of Martian terrain, each with dimensions of 64x128. Each image is paired with a mask that assigns a label to each pixel of the image.

2.2 Main Challenges

The main challenges of this problem include:

- **Class Imbalance**: Some terrain types are less frequent in the dataset, which could lead to poor performance on underrepresented classes.
- **Complexity of Terrain Features**: Martian terrain images exhibit complex and varied textures, with gradual transitions between classes, making the segmentation task non-trivial. Accurate boundaries between terrain types need to be identified.
- **Generalization to New Data**: Given that the model is trained on a finite set of images, it must generalize well to unseen terrain features and conditions that may differ from the training set.

2.3 Initial Assumptions

Several assumptions were made in the initial stages of the project. At first, it was assumed that the classes in the dataset were separable based on their visual features, without significant overlap between terrain types. Secondly, it was assumed that the

training labels were consistent with the labels of the test set, so no processing was applied to the masks, considered reliable.

3 Method

This section describes the approach used for training the model and evaluating its performance.

3.1 Data inspection and preprocessing

We performed a visual inspection of the images to assess their overall characteristics and identify potential outliers or inconsistencies. Approximately 100 images were identified as outliers during this process and subsequently removed to ensure dataset integrity. As a first step, we split the dataset into training (80%), validation (10%), and testing (10%). To improve the model's ability to generalize across different terrain types and to mitigate overfitting, several data augmentation techniques were employed during the training phase. Different types of augmentations were used: *horizontal and vertical flips*, *brightness and contrast adjustments*, *cropping and resizing*. The combination of these augmentations was tested in different configurations to evaluate their impact on performance.

3.2 Model architecture

The model used for this task is based on the U-Net architecture, which is particularly suited for pixel-level segmentation tasks [1].

The UNet consists of an encoder-decoder structure with skip connections between the encoder and decoder layers, facilitating the recovery of spatial details during the upsampling process. The architecture was tested with different depths, including UNet with 2, 3, and 4 blocks, and modifications such as adding residual connections and incorporating attention mechanisms. Regularization techniques such as *Dropout*(p=0.5) and *Batch Normalization* were employed to prevent overfitting.

We also increased the complexity of the network by incorporating **attention blocks** and the **ASPP** module into the architecture.

The attention blocks are designed to help the model focus on the most relevant spatial features by assigning different levels of attention to various regions in the input image. This allows the network to prioritize important details and suppress irrelevant information, improving the model's ability to detect subtle and complex patterns in Martian terrain images. [2]

The ASPP module was implemented at the bottleneck of the network: it employs dilated convolutions with different dilation rates, capturing multi-scale contextual information from the input. This enables the model to recognize features at multiple spatial resolutions, making it more robust to variations in the terrain's scale and structure. [3]

3.3 Training model

The model was trained using the **AdamW optimizer**, with experiments conducted using different learning rates and *learning rate schedules*. Specifically, we tested the "Reduce On Plateau" schedule and the **Cosine Decay** schedule, tuning the relevant hyperparameters to determine the best configuration. After training the whole network, we tried also to fine tune only bottleneck and decoder blocks, freezing encoder blocks for a second training step. The following loss functions were implemented to address the imbalance between the classes and to obtain an optimal solution:

Categorical Cross-Entropy: Used in the baseline model.

Weighted Loss: This loss modifies categorical cross-entropy by applying a weight to each class, inversely proportional to its frequency in the dataset.

$$L = -\frac{1}{N}\sum_{i=1}^{N} w_{y_i} \cdot \log(p_{y_i})$$

Where N is the number of samples, y_i the true class of sample, p_{y_i} the predicted probability for class and f_{y_i} inverse of frequency of class y_i in the dataset.

Focal Loss: Focused on difficult-to-classify examples, particularly for minority classes. [4]

4 Experiments

Several more optimization solutions were experimented during this project but without bringing any improvement. We tried different types of skip connections (e.g.concatenate, add, adaptive fusion, etc), different personalized loss functions and also their combinations (e.g.Boundary, Focal+Dice, etc).

Table 1: **best result** (***for lack of submission kaggle IoU is not available).

Model	IoU-local test set	IoU-kaggle
Base U-net (2 blocks)	43.16 %	42.19 %
U-net (3 blocks) =x	52.26%	47.76%
x + ASPP	53.00%	50.06%
x + ASPP + Attention Blocks	**54.87%**	**53.50%**
x + ASPP + Attention Blocks + Fine Tuning	65.62%	***

5 Results

We evaluated the performance of each configuration based on the IoU metric, including local and test set evaluations, to determine the effectiveness of each combination of parameters. The following table summarizes the key models tested and their performance in terms of IoU.

We started by evaluating the performance of a baseline model, which utilized a simple U-Net architecture with two blocks in both the encoder and decoder, combined with the standard Categorical Cross-Entropy loss. Introducing the weighted loss function and focal loss improved the performance, particularly in the accuracy of segmenting less represented classes.

Additionally, increasing the depth of the network by adding one additional block in both the encoder and decoder led to significant improvements. This deeper architecture enabled the model to capture more intricate features of the Martian terrain, contributing to better segmentation results. This regularization technique made the model more robust to the noise present in the training data, which, in turn, improved its performance on the validation set. The further addition of one more u-net block led to a decreasing in the IoU probably due to overfitting.

The integration of ASPP and attention blocks proved beneficial, leading to improved segmentation accuracy.

(***)*Also the fine tuning of bottleneck and decoder blocks seemed to be effective to improve the performances of the best model obtained that far.*

While some of the results met expectations, some unexpected findings provided valuable insights into the model's behavior. Introducing brightness and contrast augmentation, aimed at improving robustness to varying lighting conditions, resulted in a decline in performance. This outcome was unanticipated, as it was expected to help the model generalize better. However, the model struggled to adapt.

6 Discussion

The results obtained in this project highlight several strengths and areas for improvement. One of the main strengths of the model was its ability to effectively capture the underlying patterns in Martian terrain segmentation, even though sometimes the masks associated to the images seem misleading. The depth of the architecture is crucial to both allow the model to learn more complex representations and avoid overfitting. The results suggest that incorporating mechanisms such as ASPP and attention blocks into models can effectively enhance segmentation accuracy, especially in tasks where fine-grained feature extraction is crucial.

7 Conclusions

In conclusion, this project highlighted the possibility to use deep learning techniques for Martian terrain segmentation, however showing some limitations. Indeed, given the significant margin of improvement, future trials could have explored more advanced model architectures or the fine tuning on the already adopted model; indeed for lack of time and submissions this last solution was not fully evaluated.

8 Main individual contributions

Andrea Codazzi: bottleneck customization
Alessio Onori: augmentation, fine tuning
Anna Simeone: loss functions, schedule
Pasquale Serrao: skip connections

References

[1] Du, G., Cao, X., Liang, J., Chen, X., Zhan, Y. (2020). Medical Image Segmentation based on U-Net: A Review. Journal of Imaging Science Technology, 64(2).

[2] Guo, M., Liu, H., Xu, Y., Huang, Y. (2020). Building extraction based on U-Net with an attention block and multiple losses. Remote Sensing, 12(9), 1400.

[3] Chen, L. C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A. L. (2017). Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. IEEE transactions on pattern analysis and machine intelligence, 40(4), 834-848.

[4] Ross, T. Y., Dollár, G. K. H. P. (2017, July). Focal loss for dense object detection. In proceedings of the IEEE conference on computer vision and pattern recognition (pp. 2980-2988).