AN2DL - Second Homework Report
AwesomeChallengers

Sara Jo Agostino, Andrea Faccioli, Giacomo Falcone

s4r4jo4go, andreafaccioli, giacomofalcone

280760, 271316, 271005

December 14, 2024

1 Introduction

In the context of Martian terrain characterization, **semantic segmentation** is essential for understanding the geological processes of the planet, as well as its mineral composition or potential habitability. By assigning semantic labels to each pixel in Martian terrain images, researchers can gain insights into the planet's environment and its evolution over time. In line with this, our goal is to develop a model that can accurately classify Martian surface into **five distinct categories**: [Background(0), Soil (1), Bedrock(2), Sand (3), Big Rock (4)], based on pixel-level analysis of real images.

2 Problem Analysis

Our models are evaluated according to the **mean intersection over union** metric, computed as:

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

It's calculated as the average IoU across all classes (C) (background excluded), where the IoU for a single class is defined as the ratio between the intersection (common pixels between prediction and ground truth) and the union (all pixels belonging to either the prediction or the ground truth).

2.1 Data Inspection

The given dataset consists of segmented images from Mars terrain: it contains 2615 samples, each one paired with a mask representing the class of each pixel. Images have a size of 64×128 pixels, represented in one gray scale channel. First of all we analyzed the dataset in order to identify outliers and duplicate images. We didn't find any duplicates. Instead, by looking at the mean value of pixels, we identified some outliers: images of aliens (Figure 1). By noticing that they all have the same mask, we were able to find all **110** and remove them.

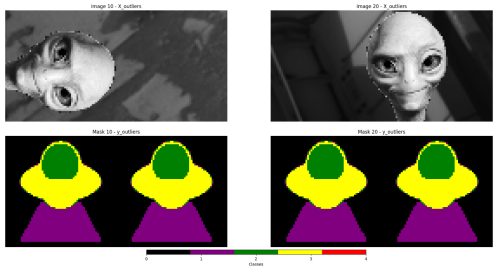

Figure 1: Outliers

As a first step, we tried to delete images dominated by the background class (class 0), as they were supposed to contain minimal representation of the classes relevant to our segmentation task. We obtained the best results by setting the background threshold to 0.8 (discard images containing more than 80% of the pixels of background).
At the same time, we immediately noticed that the

dataset was characterized by a strong class imbalance: big rocks (class 4) are represented in less than the **0.13%** of the total pixels of the set. We will further discuss this problem in the next sections.
After outliers removal, we computed the dataset classes distributions:

Class	Occurrences
0: Background	24.31%
1: Soil	33.90%
2: Bedrock	23.28%
3: Sand	18.38%
4: Big Rock	0.13%

Table 1: Training set occurrences per class

3 Method

We tackled the issue of class imbalance by trying two different techniques: a **weighted loss function** (*it gives adequate importance to the less frequent classes during training*) and a targeted **image augmentation** of class 4.

We chose to use **U-Net** [1] and some of its variants to address the semantic segmentation problem. U-Net is particularly well-suited for this type of task due to its ability to combine local and global information through skip connections, ensuring precision in details and boundaries. We experimented with different depths and configurations, as well as with several other architectures, such as a double U-Net, DeepLab [2] and other complex designs, but did not observe significant improvements. This could be due to the approximate nature of the segmentation, which may not require highly complex architectures.

4 Experiments

4.1 First U-Net

In the early stages of the challenge, we implemented a simple U-Net architecture. The downsampling path included two repetitions of a 3×3 convolution (with "same" padding), batch normalization, ReLU activation, and a 2×2 max-pooling layer. The bottleneck repeated this block to extract deeper features. The upsampling path used nearest neighbor interpolation (interpolation = 'nearest') instead of max-pooling and incorporated skip connections by concatenating feature maps from the downsampling path with the upsampled ones. This enabled the integration of detailed features from earlier layers. Finally, 1×1 convolutional layers were applied at the top to produce pixel-wise class predictions.

4.2 Bottleneck

We experimented with several bottleneck variations, including self-attention, residual blocks and attention gates; we also implemented an Atrous Spatial Pyramid Pooling (ASPP) module, in order to extract multi-scale contextual information by combining global average pooling with dilated convolutions at various rates. Ultimately, the best performance was achieved by incorporating a **Squeeze and Excitation** [3] block, by first applying global average pooling to the input tensor to capture channel-wise statistics, then passing the result through two fully connected layers to generate attention weights, which are used to scale the original input tensor via multiplication.

4.3 Augmentation with CutMix

Our dataset contains images where the segmentations are approximate, and many images predominantly contain a single class. To face this problem and also to avoid overfitting while adding more data for training, we decided to add an augmentation pipeline using **CutMix** [4] and horizontal flips. The choice of cutMix was taken because it works well with our dataset, as it combines portions of different images and their corresponding masks.

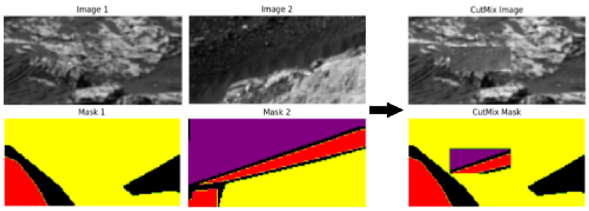

Figure 2: CutMix

4.4 Added a weighted loss

To face the problem of the low representation of class 4 we modified the categorical cross entropy loss by adding a class weights consideration mechanism[5].

$$\mathcal{L}_{WCE} = -\sum_{c=1}^{C} w_c y_c \log(\hat{y}_c)$$

where C = number of classes, w_c = weight for class c, y_c = one-hot ground-truth label for class c, and $\hat{y}_c$ = predicted probability for class c.

Table 2: Benchmark of our relevant models Additions to previous models in brackets.

Model	Validation accuracy	Mean IoU (val. set)	Mean IoU (test set)
U-Net 1	70.15%	41.79%	36.09%
U-Net 2 (cutMix + Flip + squeeze & elicitation block)	73.30%	45.11%	46.45%
U-Net 3 (weighted loss)	**77.50**%	52.32%	54.66%
U-Net 4 (exclude class 0 from loss)	76.85%	70.74%	71.17%
Final U-Net (class 4 augmentation, LR scheduler)	74.43%	**75.16**%	**73.22**%

4.5 Targeted augmentation for class 4

To address the underrepresentation of class 4 in the dataset, we implemented targeted data augmentation techniques. Specifically, we applied flips and a customized version of CutMix that selectively combined images containing class 4. In this approach, we extracted regions of images where class 4 is present and inserted these regions into other images. This allowed us to artificially increase the proportion of class 4 pixels in the dataset, reaching approximately 5% of the total pixel count.

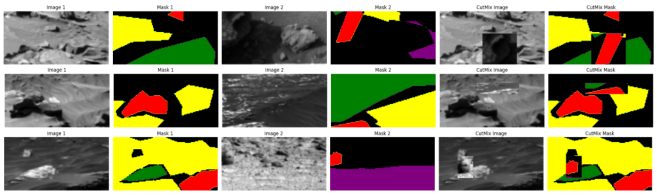

Figure 3: CutMix for class 4 (red)

This augmentation strategy led to a significant improvement in the validation IoU, increasing it to 67%, thereby enhancing the model's learning capabilities. By increasing the representation of class 4, the weighted loss function required less extreme weight adjustments for this class during training, resulting in a more stable optimization process. However, the downside of this approach was that the model began to overfit, likely because it relied too heavily on the augmented patterns and struggled to generalize effectively to unseen data.

4.6 Weighted loss adjustment

A significant improvement was achieved by excluding class 0 (background) from the weighted loss function. Since the background is not considered in the IoU metrics [1], it is not a critical class for our task. By removing it from the loss calculation, we allowed the model to focus more on the relevant classes, improving its ability to segment them accurately. This adjustment led to validation performance exceeding 80% and test performance above 70%.

5 Results

In table 2 we reported the performances of the most relevant model we trained. The final network contains all the improvements of the previous models, highlighted in brackets, including a customized augmentation for class 4 (big rock) and a learning rate decay mechanism.

6 Discussion

Our experiments demonstrated that addressing class imbalance and excluding irrelevant classes were critical to achieving high performance. Techniques like CutMix and targeted augmentations improved representation for underrepresented classes, while excluding class 0 from the loss focused the model on relevant categories. However, this led the model to ignore the background entirely, which may be suboptimal for tasks requiring a complete segmentation. Overfitting was also observed during targeted augmentations, though mitigated by the learning rate scheduler.

7 Conclusions

We developed a U-Net-based model for Martian terrain segmentation, achieving a significant test IoU improvement from 36.09% to 73.22%. Key methods like targeted augmentations, a weighted loss function and architectural enhancements effectively addressed class imbalance. Future work could explore other more sophisticated techniques, integrating advanced architectures like transformers [6], or experimenting with multi-scale architectures for further improvements.

References

[1] O. Ronneberger, P. Fischer, T. Brox. **U-Net: Convolutional Networks for Biomedical Image Segmentation**. link, 2015. Accessed: 2024-12-9.

[2] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, A. L. Yuille. **DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs**. link, 2017. Accessed: 2024-12-9.

[3] J. Hu, L. Shen, S. Albanie, G. Sun, E. Wu. **Squeeze-and-Excitation Networks**. link, 2017. Accessed: 2024-12-06.

[4] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, Youngjoon Yoo. **CutMix: Regularization Strategy to Train Strong Classifiers with Localizable Features**. link, 2019. Accessed: 2024-12-05.

[5] Y. Cui, M. Jia, T. Lin, Y. Song, S. Belongie. **Class-Balanced Loss Based on Effective Number of Samples**. link, 2019. Accessed: 2024-12-9.

[6] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, N. Houlsby. **An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale**. link, 2021. Accessed: 2024-12-9.