POLITECNICO
MILANO 1863

AN2DL - Second Homework Report
BackProp Boys

Fabio Ficcadenti, Riccardo Mazzoleni, Tommaso Parma, Berk Yonder

fabf00, riccardomazzoleni, sirt02, berkyonder

276104, 280614, 277743, 250081

December 22, 2024

1 Introduction

The aim of this project is to perform semantic segmentation on grayscale images of Mars terrain. To address this challenge, we developed and trained a neural network from scratch.
We selected the U-Net architecture due to its demonstrated effectiveness in segmentation tasks.
Our methodology consisted of the following steps:

- Conducting **data inspection** to analyze dataset statistics, identify potential class imbalances, and detect outliers or corrupted images.
- Applying **data augmentation** while carefully considering the characteristics of the dataset.
- Determining the **optimal network architecture** by evaluating performance on training and validation subsets and incrementally increasing the network's complexity as necessary.
- Employing strategies to **mitigate overfitting**, such as early stopping, weight decay regularization, and the dropout layers.
- Visualizing predictions by plotting segmentation results for qualitative evaluation.

2 Problem Analysis

2.1 Dataset Characteristics

The provided dataset consists of 64x128 grayscale images of Martian terrain, 2615 for the training and 10022 for the test. Each image has a label mask with five classes pixel-wise: *background*, *soil*, *bedrock*, *sand*, and *big rock*.

2.2 Main Challenges

- Find the right **network architecture** and/or pipeline for the specific task.
- **Solve data problems** like outliers and damaged images.
- **Understanding the dataset** for class and image distribution.
- Find a suitable metric and **loss function** and training schedule.
- Careful preprocessing and **augmentation**, the dataset characteristic must must not be overwhelmed.

2.3 Initial Assumptions

- The final test set has the same characteristics of our train set.

3 Method

3.1 Data Preprocessing

First of all, we checked the data for ensure its quality. In a first visual inspection, we found that aliens had colonized Mars before us, and they wanted to say hello photobombing our images.
Outliers were removed based on the criterion that the mask images were identical across all outliers. Upon conducting a more thorough visual inspection, we identified several images that appeared to be noise, possibly resulting from camera malfunctions. Consequently, these images, approximately 40 in total, were removed from the dataset. There were also images classified as background, even though they visually appear to be soil class. As far as classification is concerned, the baseline accuracy is 0.34, while baseline mean MIOU is 0.0678. Most frequent class across images is class 1 (soil).

3.2 Loss function and metric

For the loss function we considered the **Focal Loss** [1] function that adds a $(1-p_t)^\gamma$ factor to the standard cross entropy criterion. Setting $\gamma > 0$ will reduce the relative loss for well-classified examples, putting more focus on hard, misclassified examples. Here there is a tunable focusing parameter γ:

$$FL(p_t) = -\alpha_t(1-p_t)^\gamma \log(p_t)$$

On the other hand, the **Dice loss** function handles classes imbalance as well, in addition to optimizing the overlap between predictions and ground truth:

$$L_{\text{Dice}} = 1 - \frac{2\sum_i P_i G_i}{\sum_i P_i^2 + \sum_i G_i^2 + \epsilon}$$

where P_i are the predicted probability values, G_i are the binary values of the ground truth, ϵ prevents division for zero. Our loss function was a combination of these two: **25% Focal Loss** for the difficult classes and **75% Dice Loss** due to its similarity to mean intercept over union. We used as metric of success **Mean Intercept Over Union**, a well-known metric for image segmentation. It is defined as the area of overlap between the predicted segmentation and the ground truth, divided by the area of union between the predicted segmentation and the ground truth.

3.3 Data label characteristic

Upon examining the dataset characteristics, we found that 54% of the images are **monoclass**, monoclass being if 90% or more of its mask pixels belong to a single class.

3.4 Augmentation

To augment the dataset while preserving its characteristics, all images were subjected to either a *horizontal flip*, *vertical flip*, or a combination of both. **For multiclass images only**, we implemented two additional augmentation strategies inspired by the Keras *Cutmix* function. These involved extracting instances of the minority class from a subset of images and pasting them onto another image to increase dataset diversity.
Moreover, we applied similar logic to augment the underrepresented "big rock" class, identified as the most suitable for this enhancement. Rocks were copied and pasted across images with slight positional shifts to mitigate overfitting. Of course masks were transformed as well. These techniques collectively improved test results by 3–4% MIOU.

3.5 Warm-up and Learning rate scheduler

Since every model had to be trained by scratch, we implemented a warm-up schedule to lower learning rate for the first 10 epochs gradually increasing it to be the chosen learning rate for the model at the last epoch of warm-up. Then a learning rate schedule is set to lower the learning rate on plateau, so to halve it every 10 epoch where there is no improvement of our metric MIOU. Also, early stopping and some dropout layers were used to prevent overfitting.

3.6 Trial & error

To find a basic functional architecture, we began with a straightforward UNet with skip connection comprising two layers of downsampling and, symmetrically, two layers of upsampling [2]. Subsequently, we adopted an incremental approach, introducing gradual modifications to the network. We ended up by having a U-net with 3 layers of resolution with 2 convolutional layer with 16, 36 and 64 channels each, with a bottleneck with 3 blocks

of 128 channels. The result on test set was around 45% MIOU.

4 Experiments

Given the baseline net described in section (3.6), we started improving by experimenting more advanced techniques.

- **Squeeze and excitation (SE)**: adding a squeeze and excitation module at the end of the bottleneck yielded the best results, increasing MIOU by around 3-4% on validation and test set.
- **Transformer block**: similarly to the squeeze and excitation module, this block increased the performance, but less than the previous one. Moreover, gating it with SE did not increase results.
- **Trainable weights for skip connection**: we introduced trainable weights to scale the encoder portion of the skip connection prior to concatenation. No measurable results.
- **Skips with channel attention**: instead of weighted skip, we tried to use an attention gate concatenating them. No real increase in MIOU on validation set.
- **Same filters in all net layers and adding of skip connection**(128 filters in all layers) and replaces concatenation in the upsampling path with addition for skip connections. Slightly less results obtained.
- **Adding a dense layer in the Bottleneck with residual connection** this actually increased the results by 1-2% MIOU.
- **Use a double encoder branched net, and gate the results** this net performed similarly to the one decoder branch, but with some inconsistent areas in masks and more computational power needed.
- **Use a binary classifier to predict monoclass and non-monoclass images** the idea was to use it as pre or post processing to help the real net to classify monoclass images. Binary accuracy was 75% but still too similar to the main net.

5 Results

The following table shows the results obtained for the best models. Baseline 6,78% MIOU.

Table 1: **Best UNet models**

Model	val mean iou	test mean iou
SE	**51.47%**	**54.12%**
Transformer	48.73%	52.91%
Fixed channel with SE	47.24%	47.32%

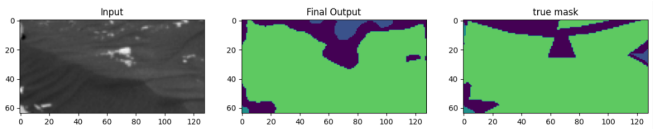

Figure 1: Image, segmentation output from the best model and true mask

6 Discussion

- Strengths: Image augmentation has increased the results.
- Limitation: net quite sensitive to initialization, the background was the most difficult area to classify. Net also sensitive to different training schedule. More GPU power needed for experimenting on more complex architectures.
- Assumptions: For tasks like image segmentation, super resolution and similar, network architecture is more crucial with respect to other Deep learning tasks.

7 Conclusions

- U-net with Squeeze and Excitation has proven effective for this task, with a MIOU of 54% on test set (with respect to 6% baseline).

8 Contributions

Fabio Ficcadenti: code, report; Tommaso Parma: code, report; Riccardo Mazzoleni: code, report; Berk Yonder: code

References

[1] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár. Focal loss for dense object detection. https://arxiv.org/abs/1708.02002v2, 2018. Accessed: 2024-03-21.

[2] E. Lomurno. Lecture 5b: Semantic segmentation with unet, 2024.