POLITECNICO
MILANO 1863

AN2DL - Second Homework Report
ANNamo

Silvia Carone, Mattia Fioravanti, Antonio Fraccastoro, Federico Patacca

silviacarone, mattiasante0512, AntonerDL , federicopatacca

249510, 243149, 252908, 250869

December 14, 2024

1 Introduction

We were asked to design a *neural network* that needs to perform semantic segmentation on a dataset of Mars terrain images with five distinct types of classes. The provided dataset consisted in 2615 64x128-pixel size images in grey-scale color format. To achieve this, we proposed different variants of U-Net, built from scratch.

2 Problem Analysis

The task involves multi-class semantic segmentation on a dataset characterized by imbalanced class distributions, where the background class dominates, and certain categories (e.g., category 4) are significantly underrepresented. The primary challenges include handling this class imbalance, ensuring robust segmentation of small or difficult regions, and addressing the potential for overfitting due to limited samples in minority classes. Initially, we assumed that augmentation and targeted loss functions could mitigate class imbalance, while carefully designed architectures, incorporating attention mechanisms, would enhance performance on challenging regions.

3 Method

The initial step involved identifying and removing outliers from the dataset. A key observation was that the masks associated with the outliers were identical, highlighting their anomalous nature. By eliminating these outliers, we ensured the dataset's integrity and the reliability of subsequent processes.
Initially, a massive augmentation was performed across the dataset. However, further evaluation revealed that a more efficient approach was to rebalance the dataset using oversampling targeted at the least-represented category (category 4).
This oversampling was achieved with a light augmentation that include *RandomZoom*, *RandomTranslation* and *RandomRotation*, ensuring the minority class was sufficiently represented without overwhelming computational resources or risking overfitting.
To address the inherent class imbalance, we implemented a custom focal loss function. This tailored loss function focuses on improving model performance on underrepresented classes, particularly category 4. The custom focal loss is designed with the following objectives: exclusion of background class (class 0), ensuring the focus remains on meaningful predictions. Higher penalty for difficult examples in which errors on less frequent or harder-to-predict classes are penalized more heavily, incentivizing the model to prioritize these cases.Numerical stability,

where techniques such as clipping are employed to ensure robust loss computations.
As architecture we chose UNet++ [5], which consists of U-Nets of varying depths whose decoders are densely connected at the same resolution via the redesigned skip connections. Furthermore, it effectively fuses features across different resolutions, making it able to capture both fine-grained details and global context. Instead of classical Convolutional Layers, our U-Net++ (referencely *unetplus-plus.ipynb*) consists in residual blocks. To avoid overfitting, we decided to put a SpatialDropout2D layer before the output layer.

4 Experiments

We identified a proper loss function [1] as a crucial factor in increasing the performances of the semantic segmentation problem. Subsequently, we present several experiments we conducted on our model to achieve better results:
Categorical Cross-Entropy: This was our base loss function,it measures how well the model's predictions match the target labels. Using the softmax function, the model generates pixel-wise probability maps representing the likelihood of each pixel belonging to each class.
Focal Loss: As stated in the method section we needed a loss that was able to handle class imbalance, which in our case is a huge problem, given that the big rocks are both rare and account only for a small portion in each image, to solve this we adopted the focal loss first considering all the classes, then excluding class 0, therefore our loss function for each image can be stated as: $L_{focal}(y, t, \gamma, \alpha) = \sum_{n=1}^{8192} -t_n \cdot \alpha \cdot (1-y_n)^{\gamma} \cdot \log(y_n)$. Here t_n is the one-hot encoded vector of the true class of the n^{th} pixel, while y_n is the vector with the predicted class probabilities, $\gamma = 2.0$ is the focusing parameter, $\alpha = \frac{1}{4}$ is a weighting factor and 8192 is the number of pixels we have in an image.As said in the final model it proved to be the winning choice.
Dice Loss: The Dice Coefficient represents a measure of similarity between two sets of data and it is used in image segmentation to evaluate the overlap between a predicted mask and the actual target mask. The Dice Coefficient is defined as follows: $DICE_{COEFF} = \frac{2|Y \cap T|}{|Y|+|T|}$, where Y represents the prediction mask and T the target mask for a single class. Subsequently, the Dice Loss is computed as $DICE_{LOSS} = 1 - DICE_{COEFF}$. This method is optimal to tackle with unbalanced datasets as it allows to pass in the algorithm custom weights for each class. We then decided to pass weights which were inversely related to the frequency which that class appeared in the dataset. Despite achieving a greater performance in the sense of validation mIOU, the mIOU on the local test set did not follow the same trend, leading us to move into different methodologies.
Combined Loss: We wanted to get the best from all the presented losses, that's why we tried to use them simultaneously through a weighted sum. We reasoned on the correct weights for the losses trying to make them of the same order of magnitude so that one wouldn't prevail over the others, as well as assigning weights that sum up to 1. However these attempts weren't really successful, in fact they didn't show any improvement over the focal loss alone.
Moreover, we tried to implement different architecture for our U-Net:

- **Double U-Net**
  The double U-Net [3] is a recent architecture that combines two U-Nets stacked on top of each other, they are composed by: the usual encoder-decoder structure where we adopted residual blocks from ResNet with skip connections and two convolutional layers with Group Normalization and ReLU, each one followed by a squeeze and excite one in the encoder which is composed of 4 levels, doubling the number of filters as we go deeper. The bottleneck is an Atrous Spatial Pyramid Pooling block (ASPP) to capture contextual information within the network. Finally the decoder is only composed by simple upsampling blocks.
  The advantage of this approach relies on the prediction made by the first network which is fed, multiplied by the input image, to the second network which is able to adjust it and focuses on the most difficult areas to frame.

- **SE U-Net**
  This U-Net architecture combines Squeeze-and-Excitation (SE) Blocks in the encoder and attention gates in the decoder. SE blocks are incorporated into the encoder to dynamically recalibrate channel-wise feature responses. These blocks compress the spatial di-

Table 1: Results achieved with different models.

Model	Validation Miou	Inner test Miou	Hidden test set Miou
base model	40.83%	45.32%	45.31%
double UNet	57.74%	58.79%	59.11%
SE U-Net	61.2%	59.19%	59.86%
Final model	66.75%	62.76%	61.07%

mensions to extract channel-level information and enhance feature extraction [2]. Regarding the attention gates, they leverage gating signals to improve the resolution of skip connections and enhance predictions at the region-of-interest (ROI) level [4]. Skip connections are combined with the decoder output using Add operations, rather than concatenation, ensuring efficient integration of features while reducing parameter overhead.

5 Results

The conducted analysis shows a substantial increment with respect to the result related to the baseline considered, a simple U-Net. As shown in the above table, we were able to exploit the chosen techniques in a proficient way obtaining a final mIoU of **0.61066**.
Regarding the unexpected outcomes, two principal points need to be discussed: the results of the different losses considered and the performances of the different analyzed architectures. Specifically, the expected effect of the exploitation of the combined loss function was a general increase in the model's performance because of different contributions focused on various dataset characteristics.
For the effects of the different architectures, the similar results obtained suggest an improvable use in the feature extraction mechanism related to the block taken from the state of the art architectures.

6 Discussion

The proposed model has a good **trade-off** in computational-time and accuracy, thanks to its light structure and the usage of blocks that are commonly used in similar contexts. We faced some big limitations in our work like the low resolution images, as long as the least informative ones and undoubtedly the masks, they were handmade and often very different from the image content, therefore leveraging algorithms to correct them may be useful. Regarding the architecture, one limitation was represented by the prohibition of pre-trained models, taking as example the double UNet we implemented, the first encoder in the original paper is a VGG, this may drastically improve the performance.

7 Conclusions

We addressed the segmentation challenge by leveraging augmentations and losses to solve class imbalance and adopting advanced blocks as the residual one and attention techniques as: squeeze and excitation and attention gate as long as different types of UNet architectures coded and trained from scratch. Indeed there is room for improvements, such as the implementation of transformers inside this kind of models and accurately designed loss functions can again improve the model ability to distinguish classes.

8 Contribution

Everyone in our group has contributed equally to the development of the model. It has been a process of trial and error and efficient implementation of numerous techniques.
Silvia Carone: dataset cleaning, SE U-Net
Mattia Fioravanti: focal loss, double UNet
Antonio Fraccastoro: augmentation, PescaraV2
Federico Patacca: dice loss, unet++

References

[1] R. Azad, M. Heidary, K. Yilmaz, M. Hüttemann, S. Karimijafarbigloo, Y. Wu, A. Schmeink, and D. Merhof. Loss func-

tions in the era of semantic segmentation: A survey and outlook. 2023.

[2] J. Hu, L. Shen, and G. Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018.

[3] D. Jha, M. A. Riegler, D. Johansen, P. Halvorsen, and H. D. Johansen. Doubleu-net: A deep convolutional neural network for medical image segmentation, 2020.

[4] G. Prasanna, J. R. Ernest, G. Lalitha, and S. Narayanan. Squeeze excitation embedded attention u-net for brain tumor segmentation. In *International Conference on Emerging Electronics and Automation*, pages 107–117. Springer, 2022.

[5] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang. Unet++: Redesigning skip connections to exploit multiscale features in image segmentation. 2020.