AN2DL - Second Homework Report
dAI che è venerdì

Alessandro Annechini, Riccardo Fiorentini, Lorenzo Vignoli

annechini, riccardo01, vigno

250905, 249012, 252085

December 17, 2024

1 Introduction

The Second Homework for the Artificial Neural Networks and Deep Learning (AN2DL) course involves categorizing the pixels of 64x128 grayscale real images from Mars terrain into five different classes, each representing a particular type of terrain. This is a semantic segmentation problem and the goal is to assign the correct class label to each mask pixel maximizing the *mean intersection over union metric* without considering the background class:

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

The report is structured as follows: in Section 2 we introduce the main challenges, while in Section 3 we detail the methodology we employed to tackle the problem. Section 4 reports the experiments performed while designing our model, and Section 5 shows the final results. Finally, in Section 6 we list our most important takeaways, along with the main individual contributions.

2 Problem Analysis

In this semantic segmentation task, several challenges arose, making the problem particularly complex. The main issues we encountered include:

- **Data scarcity and class imbalance**: The dataset provided contained a limited number of labeled samples, with certain terrain classes being significantly underrepresented.
- **Incorrect samples**: The pre-labeled dataset included irrelevant images, which had to be identified and removed to avoid introducing noise into the training process.

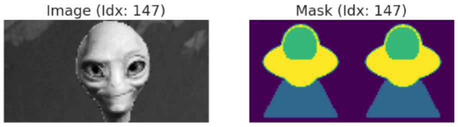

Figure 1: Example of incorrect samples

3 Method

To design our model we reviewed relevant literature [1, 4, 5, 6, 7, 9] and iteratively tested and refined components and parameters.

3.1 Not just a U-Net Model

We based our model on the U-Net architecture due to its ability to combine contextual and fine-grained details via skip connections, making it effective for moderately sized datasets and adaptable for enhancements with advanced modules.

- **Encoder**: Our encoder leverages a combination of residual and inception blocks to extract hierarchical features. Residual blocks [3] preserve information through skip connections while enabling deep feature learning, mitigating the risk of vanishing gradients. Inception blocks [8] capture multi-scale information by processing the input with convolutional filters of varying kernel sizes and combining them with max-pooling outputs.

- **Bottleneck**: The bottleneck combines many modules. Mini-ASPP module [9] facilitates multi-scale contextual learning by aggregating features at different dilation rates, capturing information across various spatial scales. A self-attention mechanism [5] is integrated to model long-range dependencies, which is critical for handling objects of different sizes and complex spatial distributions. Additionally, a Strip-Pyramid Pooling module enriches the representation by capturing directional contextual features, improving its ability to process structured patterns more effectively.
- **Decoder**: High-resolution outputs are reconstructed via transposed convolutions and skip connections [7], augmented with gating mechanisms to adaptively combine encoder and decoder features. This reduces noise and enhances spatial detail reconstruction.

In the end, the output is mapped to multi-class probabilities using a 1x1 convolution with softmax activation.

3.2 Data Management

To prevent overfitting and improve generalization, we implemented data augmentation by applying dynamic transformations such as flipping, rotation, scaling, translation, and adjustments to contrast and brightness. These transformations simulated diverse spatial and photometric conditions, enriching the training dataset and making the model more robust. Additionally, to address class imbalance, we incorporated class weights into the loss function, ensuring that underrepresented classes contributed proportionally during the training process, in order to achieve a more balanced learning outcome.

3.3 Loss Function

We developed a custom loss function integrating multiple components [1]—*Weighted Dice Loss*, *Improved Focal Loss*, *IoU Loss*, and a custom *Margin Loss*—to address class imbalance, penalize misclassifications, and improve segmentation accuracy. The Weighted Dice Loss balances class contributions while measuring the overlap between predictions and ground truth. The Improved Focal Loss focuses on hard-to-classify samples by dynamically adjusting for prediction confidence, to mitigate class imbalance. The IoU Loss optimizes the intersection-over-union metric, ensuring precision in overlapping regions.

The Margin Loss reduces false negatives and false positives using norm-based penalties. Its first term, f_n, minimizes false negatives by ensuring the highest prediction corresponds to the correct class:

$$f_n = \sum_{i=1}^{N} \omega_{c=c_i} \cdot (\|y_{pred,c\neq c_i}\|_5 - y_{pred,c=c_i}),$$

where $\omega_{c=c_i}$ is the class weight, $y_{pred,c=c_i}$ is the correct prediction, and $\|\cdot\|_5$ approximates $\max(\cdot)$ while maintaining differentiability [2]. The second term, f_p, penalizes false positives:

$$f_p = \sum_{i=1}^{N} \omega_{c\neq c_i} \cdot \left(\frac{y_{pred}^5}{\|y_{pred}^5\|_1} \right)_{c\neq c_i},$$

where y_{pred}^5 amplifies the largest entries to highlight incorrect predictions. The Margin Loss combines these terms as:

$$Margin\ Loss = 0.99 \cdot f_n + 0.01 \cdot f_p.$$

The final loss function emphasizes the Margin Loss while equally weighting the other components, guiding the model to learn relevant features and achieve balanced performance across all classes.

4 Experiments

4.1 Loss-functions experiments:

We evaluated different loss configurations to assess their impact on segmentation performance 1. The table compares the Total Combined Loss, which integrates all components, with partial combinations like Focal + Dice + IoU and individual losses, highlighting the benefits of our integrated approach.

Loss Name	Validation IoU
Total Combined	62.64 %
Focal + Dice + Iou	60.05 %
Margin Loss	61.10 %
Focal + Dice	46.45 %

Table 1: Loss-functions tests

4.2 Bottlenecks

We tested different bottleneck configurations and found that the combination of features extracted by a mini ASPP module, processed through a Polarized Self-Attention (PSA) mechanism, and refined using a Strip Pyramid Pooling (SPP) layer—concatenating the attention and pooling features—achieves the highest IoU with the ”aspp + self-attention + spp” setup.

Bottleneck Name	Validation IoU
aspp + self attention + spp	62.64%
aspp + self attention	59.03%
self attention	57.17%

Table 2: Bottlenecks tests

4.3 Further explorations

The following approaches were implemented but ultimately discarded due to their limited impact on model performance:

- **Feature Pyramid Pooling Module (FPPM)** [4]: This module was designed to capture multi-scale contextual information by merging global, medium, and fine-grained details in the encoder pipeline.
- **Test Time Augmentation (TTA):** This approach involved performing inference on test samples augmented with randomized transformations, followed by remapping the predictions back to the original sample space. The augmentations, however, did not introduce sufficient variability to meaningfully enhance the predictions.
- **Dual Model Implementation:** To address the challenge of underrepresented classes, a second model was trained specifically to predict the least frequent class. The predictions from this auxiliary model were then combined with those of the primary U-Net, creating a double U-Net architecture.

5 Results

Our model demonstrated the effectiveness of advanced architectural components and a carefully designed loss function, achieving a validation mean IoU of 62.64%. This represents a significant improvement over baseline models, which achieved approximately 45% IoU with simpler architectures and loss configurations. Data augmentation contributed to this success by improving generalization and increasing performance by 3-5%. Architectural features such as attention mechanisms and feature pyramids further enhanced the model's ability to capture fine details. Our experiments allowed us to select or discard state-of-the-art approaches, maintaining the components more suited for the specific problem.

These results highlight the effectiveness of our approach, while also identifying areas for future improvements, such as exploring additional techniques to further increase performances over minority classes.

6 Conclusions

We tackled a challenging semantic segmentation problem using grayscale images of Mars terrain. Our U-Net-based model was enriched with feature pyramids and attention mechanisms to improve detail capture and context awareness. A custom loss function combining Weighted Dice, Focal, IoU, and Margin Losses effectively addressed class imbalances and improved segmentation accuracy. This setup achieved the highest validation IoU of 62.64%, outperforming simpler methods. Data augmentation also significantly enhanced the model's generalization on unseen data.

6.1 Individual contributions

Alessandro Annechini focused on data preprocessing and combined various loss functions to be employed in the training process.
Riccardo Fiorentini implemented and tested bottleneck and attention mechanisms, employing several state-of-the-art methods [6].
Lorenzo Vignoli conducted experiments, performed parameter tuning, and systematically evaluated solutions to the problem.

References

[1] R. Azad, M. Heidary, K. Yilmaz, M. Hüttemann, S. Karimijafarbigloo, Y. Wu, A. Schmeink, and D. Merhof. Loss functions in the era of semantic segmentation: A survey and outlook, 2023.

[2] A. Epasto, M. Mahdian, M. Zampetakis, and V. Mirrokni. Optimal approximation-smoothness tradeoffs for soft-max functions. In *Proceedings of the 34th Conference on Neural Information Processing Systems (NeurIPS 2020)*, 2020.

[3] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015.

[4] H. Li, P. Xiong, J. An, and L. Wang. Pyramid attention network for semantic segmentation. *arXiv preprint arXiv:1805.10180*, 2018.

[5] H. Li, P. Xiong, J. An, and L. Wang. Pyramid attention network for semantic segmentation. *arXiv preprint arXiv:1805.10180*, 2018.

[6] J. Li, K. Chen, G. Tian, L. Li, and Z. Shi. Marsseg: Mars surface semantic segmentation with multi-level extractor and connector. *arXiv preprint arXiv:2404.04155*, 2024.

[7] M. Mubashar, H. Ali, C. Grönlund, and S. Azmat. R2u++: a multiscale recurrent residual u-net with dense skip connections for medical image segmentation. *Neural Computing and Applications*, 34(20):17723–17739, 2022.

[8] C. Szegedy, S. Ioffe, and V. Vanhoucke. Inception-v4, inception-resnet and the impact of residual connections on learning. *CoRR*, abs/1602.07261, 2016.

[9] J. Wang, X. Zhang, T. Yan, and A. Tan. Dpnet: Dual-pyramid semantic segmentation network based on improved deeplabv3 plus. *Electronics*, 12(14):3161, 2023.