AN2DL - Second Homework Report
Team Fallback

Malvin Noël, Léandre Le Bizec, Daniel Uri Trejo Pimentel

Malvin, Leandre, Uri

11077886, 10990212, 10941235

December 14, 2024

1 Introduction

This project tackles a *multi-class semantic segmentation problem* to classify Martian terrain pixels into five classes: background, soil, bedrock, sand, and big rock. The goal is to develop neural network models capable of accurately predicting pixel labels in 64x128 grayscale images. Key challenges include pixel-level accuracy, class imbalances, and ensuring generalization without using pre-trained models.

Our approach involved the following steps:

- **Dataset Analysis:** Analyzed the dataset structure, created validation splits, and attempted to balance class distribution through class-weighted loss functions and sampling;
- **Baseline Model Development:** Designed and trained a U-Net architecture from scratch as the foundation;
- **Data Augmentation:** Applied techniques like rotation, flipping, brightness, contrast and cut and mix to increase data diversity and address class imbalance;
- **Loss Function Optimization:** Experimented with class-weighted and combined loss functions to improve accuracy;
- **Model Improvement:** Enhanced the U-Net with architectural variations and specialized focusing mechanisms;
- **Ensemble Learning:** Combined multiple complementary models to improve robustness and accuracy [1].

2 Problem Analysis

Our initial analysis revealed several key characteristics and challenges.

2.1 Dataset Characteristics

- 2,505 labeled images for training (after cleaning) and 10,022 unlabeled images for testing
- Image resolution of 64x128 pixels in grayscale format
- 5 terrain classes with significant imbalance:
  - Background: 24.31%
  - Soil: 33.90%
  - Bedrock: 23.28%
  - Sand: 18.38%
  - Big Rock: 0.13%
- Average of 2.18 classes per image
- Mean intensity of 60.901 with standard deviation of 25.990

2.2 Main Challenges

1. Extreme class imbalance, particularly for the Big Rock class (0.13%).
2. Need for high spatial accuracy in pixel-level predictions.
3. Limited resolution affecting fine detail detection.

4. Absence of pre-trained models requiring efficient training from scratch.

5. Complex terrain boundaries requiring precise segmentation.

3 Method

Our methodology consisted of three main components: data preprocessing, model development, and ensemble learning.

3.1 Data Preprocessing

The data preprocessing phase was carried out in two key steps.

Step 1: Data Cleaning We identified and removed 110 outlier images, reducing the dataset from 2,615 to 2,505 training samples. This cleaning improved the dataset's consistency and quality.

Step 2: Data Augmentation We implemented comprehensive augmentation focusing on underrepresented classes:

- **Geometric Transformations:** Horizontal and vertical flipping, rotation, shift, zoom;
- **Intensity Adjustments:** Random brightness (±0.4) and contrast (0.7-1.3)
- **Noise Addition:** Gaussian noise (mean=0.0, std-dev=0.02)
- **Cut and Mix:** used to improve dataset untill 14000 sample, but this were not used for our best model.

3.2 Model Architecture

We developed several specialized models to explore and improve segmentation performance. Below is an overview of the architectures and their respective outcomes:

1. Basic U-Net The standard U-Net architecture was implemented as a baseline model. It performed relatively well on both the training and test datasets.

2. U-Net with Supervision Intermediate supervision was added by introducing auxiliary outputs at intermediate layers. This resulted in a slight improvement in performance.

3. U-Net with Transformers & Attention We enhanced the encoder-decoder architecture by incorporating transformer blocks, which allowed for better global context understanding. Attention gates were introduced to refine features, and skip connections were enhanced with feature recalibration to improve information flow. This model achieved a validation mIoU of approximately 0.49.

4. U-Net with Data Augmentation & Attention To focus on robust feature extraction, we implemented an extensive data augmentation pipeline to improve generalization. Attention mechanisms were added to select the most relevant features during the decoding process. The model achieved a validation mIoU of approximately 0.48.

5. Small Object Specialized U-Net We designed a dual-branch architecture to specifically capture fine details, particularly for small objects. Multi-scale feature fusion was incorporated to improve spatial awareness across various scales. Additionally, deep supervision was applied to ensure improved gradient flow during training. This model achieved a validation mIoU of approximately 0.44.

3.3 Training Strategy

- **Batch Size:** 8-16 samples
- **Learning Rate:** Initial rate 1e-4 with reduction on plateau;
- **Early Stopping:** Patience of 25-30 epochs
- **Training Duration:** Up to 300 epochs with model checkpointing
- **Loss Function:** Composite loss combining:
  - Dice Loss for class imbalance;
  - Focal Loss for hard examples;
  - Boundary Loss for edge precision;
  - Categorical Cross-entropy.

3.4 Ensemble Approach

Our final solution used weighted model averaging:

- Transformer model: 0.40 weight;
- Augmentation model: 0.40 weight;
- Small object model: 0.20 weight;
- Predictions normalized and combined;
- Class labels determined by argmax.

4 Experiments and Results

4.1 Data Cleaning Results

Table 1: Dataset Cleaning Statistics

Category	Count
Original Images	2,615
Outliers Removed	110
Final Dataset	2,505

4.2 Model Performance

Table 2: Model Performance Comparison
Best model is highlighted in bold.

Model	Training	Validation	Test
Basic U-Net	0.50	0.44	0.47
Basic U-Net with CutMix	0.77	0.70	0.46
U-Net with supervision	0.57	0.47	0.49
Transformer U-Net	0.60	0.49	0.49
Augmented U-Net	0.58	0.48	0.48
Small Object U-Net	0.54	0.44	0.43
Ensemble	**0.62**	**0.51**	**0.51698**

4.3 Analysis of Results

The results presented in table 2 demonstrate several key findings:

- The basic U-Net provided a solid baseline with 0.47 test performance;
- CutMix augmentation significantly improved training (0.77) and validation (0.70) performance, but underperformed on kaggle;
- Specialized architectures (Transformer, Augmented, Small Object) each achieved similar validation ranges;
- The ensemble approach provided the best overall performance, achieving 0.51686 on the test set.

These results suggest that while individual architectural improvements offer moderate gains, the combination of complementary approaches through ensemble learning provides the most robust performance.

5 Discussion

5.1 Key Findings

- Ensemble learning proved crucial for robust performance [1];
- Individual models showed complementary strengths
- Preprocessing steps as outlier removal and data augmentation, ensured cleaner and more balanced data for training, which helped the model focus on meaningful patterns [2];
- Class weighting helped address imbalance.

5.2 Limitations

- Limited resolution affects fine detail detection;
- Extreme class imbalance remains challenging;
- Ensemble approach increases computational overhead.

6 Conclusions

We addressed the multi-class semantic segmentation of Martian terrain by combining data preprocessing, specialized architectures, and ensemble learning. The ensemble approach, leveraging complementary model strengths, achieved a final score of 0.51698, despite challenges like class imbalance and low image resolution.

7 Future Work

Future improvements could focus on advanced data augmentation techniques to address class imbalance, dynamic ensemble strategies, and loss functions tailored for small object detection. Incorporating higher-resolution images or multi-modal data could further enhance segmentation accuracy.

8 Team Contributions

- **Malvin Noël:** Developed and implemented the U-Net with Transformers model, focusing on feature extraction enhancement. Researched and implemented the ensemble learning mechanism with transformer blocks. Coordinated the model optimization process and designed the weighted averaging approach combining the three models, achieving the best test score through complementary model strengths.
- **Léandre Le Bizec:** Conducted dataset exploration, removed outliers, and applied targeted data augmentation, including a CutMix strategy, to address class imbalance. Developed and tested a baseline U-Net model, experimenting with supervision, combined loss functions, attention mechanisms, and recursive residual connections. Supervision and combined loss functions showed slight

improvements, while attention-based and recursive models underperformed. These efforts refined the approach to effective segmentation.

- **Daniel Uri Trejo Pimentel** Added some helper functions to help visualize the data along with their masks. Also added helper functions to clean the training data from some buggy instances (e.g. aliens). Made several attempts with variations of the UNet model, including data pre-processing and augmentation.

References

[1] T. G. Dietterich. Ensemble methods in machine learning. *Multiple classifier systems*, pages 1–15, 2000.

[2] Keras. Kerascv: Computer vision extensions for keras. https://keras.io/guides/keras_cv/, 2024. Accessed: 2024-11-24.