AN2DL - Second Homework Report
Artificial Neural Nuggets

Luca Cattani, Simone Lucca, Manuela Marenghi, Andrada Theodora Pascu

catta, simonelucca, manuelamarenghi, theodorap

244923, 244947, 244790, 260299

December 14, 2024

1 Introduction

The aim of this project is to develop a **deep Neural Network** for *semantic image classification* of mars terrain images. The main goals of this projects are:

- Design a versatile model able to perform semantic segmentation with high *mean Intersection Over Union* (namely **mean IOU**)
- Explore the strengths and limitations of different state-of-the-art techniques to approach a deep learning problem

2 Problem Analysis

Given the dataset, we analyzed it to look for patterns in the data that may have led to over-fitting. The following features came to our attention:

1. Repeated images, not coherent to rest of the images and labels in the dataset
2. Restricted dataset size

After noticing these characteristics we analyzed the **class balance** of the dataset. We noticed that a class in particular was significantly less frequently found than others (by about two orders of magnitude), thus creating a training bias.

3 Method

3.1 Data Manipulation

After the removal of outliers, we had to make up for the data imbalance, the aim was to manipulate the available data to enhance the model's generalization abilities, to do that we tested out three main approaches:

1. **General augmentation**[5]: we applied generic augmentation on images which included the label with the scarcest presence in the dataset, techniques include random application of: shifting, rotation, zooming, shear
2. **Copy-paste augmentation**[6]: we employed this augmentation to increase the presence of the scarcest class in the dataset by saving a cutout of the object (big rocks in our case) and the corresponding label, the objects were then pasted over other images in the dataset. In our case we notices that the best performance was obtained by pasting 3 to 5 big rock objects in the images (depending on the size of the pasted objects)
3. **Compositional image synthesis**[8]: we tested this technique to synthetically increase the number of original images in the dataset, we cutout all distinct objects from the dataset and pasted them on fresh images, in particular

we pasted more images if the class was generally less present in the dataset

It is important to note that the aforementioned techniques were only applied to the training set, leaving the *validation set untouched* (also, the images from the validation set were to fetch object cutouts). Among the techniques posted above, we noticed the best inference performance when only employing *copy-paste augmentation* together with random *bi-dimensional shifting* of the resulting image.

3.2 Network structure

As shown in the literature [7] the U-Net architecture is a very powerful tool to employ for semantic image segmentation tasks, for that reason we decided to adopt in. In particular, our variation presents the following characteristics:

1. **Backbone**: The backbone of our model is a nested U-Net architecture which has been shown to improve the performance of classical U-Net architectures [10]. Lastly, we decided to use only *two encode-decoder sub-networks* to reduce training time

2. **Additional Constructs**:

   - **Dilated spatial pyramid pooling**: allows the network to capture objects at various scales by using multiple parallel convolutional layers with different sampling rates. This approach ensures that the network can effectively handle objects of different sizes and contexts within the image, which is crucial for accurate segmentation [4][3][2]
   - **Dual Attention Unit**: a mechanism to share information within a feature tensor, both along the spatial and the channel dimensions. This block suppresses less useful features and only allows the more informative ones to pass further [9] [1]. This block is compose by two branches:
     - The *Channel Attention* branch exploits the inter-channel relationships of the convolutional feature maps by applying squeeze and excitation operations
     - The *Spatial Attention* branch is designed to exploit the inter-spatial dependencies of convolutional features. The goal of Spatial Attention is to generate a spatial attention map and use it to recalibrate the incoming features
   - **Transformer blocks**: with multi-head attention mechanisms can improve the model's ability to capture long-range dependencies among pixels in the input image. This capability is crucial for understanding the global context and relationships between different regions of the image, which can lead to more accurate segmentation

3. **Network General Features**:

   - **Optimizer**: we used Adam for its robustness, and employment of weight decay, which increases regularization by discouraging large weight values.
   - **Loss function**: Since the model seemed to focus excessively its predictions on some classes we decided to experiment combining different loss functions. We used sparse categorical cross-entropy, dice and categorical focal cross-entropy. The first focuses on pixel classification, the second on spacial overlap of classes and the third addresses class imbalance. Even if it should have improved the overall performance, after some testing, we noticed that the best performance was obtained while only using sparse categorical cross-entropy, numerically:

     $$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N} y_i \log(\hat{y}_i)$$

   - **Reduce LR On Plateau**: it updates the values of the learning rate dynamically. In doing so, it improves performance, fastens convergence, and reduces instability.
   - **Early stopping**: When the validation error starts increasing, the training comes to a stop and the model with the highest validation accuracy is restored. A patience value was used in order to

avoid stopping too early due to oscillations (in our case patience was set to 30 epochs).

4 Experiments

The first experimentation phase aimed at finding a reasonably good performative base structure for the model. We started with a single U-Net, depth three (first). Afterwards we decided to increase the complexity of the overall structure and combined two U-Nets (second). Than we experimented with its depth, setting it to four, in the attempt to capture more complex, abstract features and improve performance (third). The second phase focused on augmentation and oversampling (fourth). The third one on the implementation of new layers. We started by introducing the Dilated Spatial Pyramid Pooling, that can capture both global and local contextual information (fifth). After that we included the transformers, more precisely we added an encoder and a decoder before and after the bottleneck of both U-Nets, in order to capture distant pixels relationships (sixth). Finally we decided to replace the transformers with a dual attention unit, because we wanted to try another feature calibration method (seventh). Between one phase and the other we performed hyper-parameter tuning and in the end selected the best model belonging to each phase.

Table 1: Empirical results obtained from the different experiments

Experiments	val. Mean IOU	Inference Mean IOU
first	43.92	42.38
second	48.04	45.63
third	48.05	47.03
fourth	56.82	50.47
fifth	63.29	69.70
sixth	64.01	68.53
seventh	64.90	69.91

The results show that more or less by correctly choosing the complexity of the model and introducing an effective feature extractor it is possible to achieve a reasonably good performance.

5 Results

Empirical results show that the best performance is obtained using the Dual Attention Unit Block and Dilated Spatial Pyramid Pooling, this is justified by the fact that these components improve the model's ability to focus on relevant features and capture multi-scale contextual information, leading to more accurate and consistent predictions.

6 Discussion

Considering the assignment, and the results we obtained, it is important to observe the following:

- Although one may expect transformers to improve the performance of the model in our use case, the best results were obtained without them, probably due to overfitting.
- Despite augmentation being a fundamental tool to improve the model's performance in generalization, it is important to note that when we employing overcomplicated preprocessing pipelines, the model's performance may suffer.

7 Conclusions

Future work should focus on implementing an improved augmentation pipeline, making better use of compositional image synthesis, which should provide finer tuning of training images. In a scenario where training times are not as important as in our case, larger architectures should be used to improve performance (increasing the the network nesting a depth). Lastly, finding a way to make use of transformers, in a more subtle way could in fact improve the model's performance.

References

[1] P. Charbonnier, L. Blanc-Feraud, G. Aubert, and M. Barlaud. Two deterministic half-quadratic regularization algorithms for computed imaging. *IEEE Xplore 0-8186-6952-7*, 1994.

[2] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. Deeplab: Semantic image segmentation with deep convolutional

nets, atrous convolution, and fully connected crfs. *arXiv preprint arXiv:1606.00915*, 2017.

[3] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017.

[4] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. *arXiv preprint arXiv:1802.02611*, 2018.

[5] E. D. Cubuk, B. Zoph, J. Shlens, and Q. V. Le. Randaugment: Practical automated data augmentation with a reduced search space. *arXiv preprint arXiv:1909.13719*, 2019.

[6] G. Ghiasi, Y. Cui, A. Srinivas, R. Qian, T.-Y. Lin, E. D. Cubuk, Q. V. Le, and B. Zoph. Simple copy-paste is a strong data augmentation method for instance segmentation. *arXiv preprint arXiv:2012.07177*, 2021.

[7] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. *arXiv preprint arXiv:1505.04597*, 2015.

[8] Y. Wang, L. Qi, Y.-C. Chen, X. Zhang, and J. Jia. Image synthesis via semantic composition. *arXiv preprint arXiv:2109.07053*, 2021.

[9] S. W. Zamir, A. Arora, S. Khan, M. Hayat, F. S. Khan, M.-H. Yang, and L. Shao. Learning enriched features for real image restoration and enhancement. *arXiv preprint arXiv:2003.06792*, 2020.

[10] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang. Unet++: A nested u-net architecture for medical image segmentation. *IEEE Transactions on Medical Imaging*, 2018.