Machine Learning - Project 2, Road Segmentation

Mathilde Chaffard - Irene Vardabasso - Niccolò Venturini Degli Esposti

***Abstract*—Image segmentation divides an image into meaningful parts based on visual features; road segmentation isolates road-related regions within an image. We used several popular Convolutional Neural Networks - such as DeepLabV3 and DeepLabV3+ with ResNet-50 and ResNet-34 - to perform road segmentation on Google Maps images. We also tested various pre- and post-processing methods to refine our model and significantly improve its performance: our F1 score resulted in 0.878.**

I. Introduction

Road segmentation is central to the development of autonomous vehicles and urban planning, enabling the precise identification of roads within images. It improves navigation, traffic management and safety measures, which are crucial for the development of intelligent transport systems that promise safer and more efficient mobility solutions in our evolving cities. Harnessing the power of Convolutional Neural Networks (CNNs), known for their ability to handle complex tasks, we conducted an exploration of different CNN architectures. Our goal was to address a binary classification challenge: distinguishing road pixels from other elements in satellite images sourced from Google Maps. The training dataset consisted of 100 images, each 400x400 pixels in size in RGB format, while the test dataset consisted of 50 images, larger at 608x608 pixels. Although our Neural Networks were trained to make pixel-wise predictions (leveraging the atrous convolutions employed by the DeepLab model), the ground truth annotations were applied to 16x16 pixel patches for evaluation purposes. To be more specific, in our model, a patch is classified as a road if at least 25% of its pixels are labeled as road.

II. Data preprocessing

In our efforts to improve the accuracy and F1 score of our models, we experimented with different preprocessing techniques. Using each model with consistent parameters, as detailed in the methods, we measured the impact on performance when these preprocessing methods were used versus when they were not, over a training period of 50 epochs. This approach allowed us to evaluate the effectiveness of each technique. Ultimately, not all of the recommended preprocessing methods showed significant benefit in our pursuit of optimisation.

A. Data Normalization

Normalisation of pixel values between 0 and 1 is essential in road segmentation because Neural Networks prefer small input weight values. Large integer inputs can hinder the learning process. Scaling pixel values uniformly across all training runs ensures that the Neural Network works optimally, preventing the dominance of large values and allowing faster convergence. This practice also increases the model's adaptability to different lighting conditions and improves computational efficiency, ensuring stable training processes.

B. Data Augmentation

Due to the limited dataset of 100 images provided for our CNN, we opted to augment the data to enhance performance. Various transformations were applied, including horizontal and vertical flips, gaussian blur addition, and color enhancement. However, the models encountered difficulty in recognizing diagonal roads prevalent in the training set, primarily consisting of roads parallel to axes. To address this, rotations of 90, 180, and 270 degrees were introduced, alongside rotations between 10 and 60 degrees in increments of 10. As it can be seen on Table I data augmentation is crucial for the algorithm performance.

C. CLAHE

Enhancing image contrast is crucial for improving visibility and overall image quality, especially in tasks like road segmentation where details might be obscured. We chose CLAHE, a technique that adjusts contrast in specific areas rather than uniformly across the image. However, in our evaluations, the algorithm performance with CLAHE technique was weaker compared to the one without. This localized approach, while avoiding some issues, didn't meet our expectations for road segmentation tasks, thus we are not using it in our preprocessing pipeline.

D. Patches division

In our pursuit of optimal road segmentation, we initially investigated dividing images into 80x80 pixel patches. This approach was intended to facilitate localized feature extraction and expand the dataset to enhance the model's learning capabilities. However, the observed performance with these patches didn't meet expectations, falling short compared to using the complete, unaltered images. Maintaining the full image size for augmentation preserves the contextual understanding of the scene, allowing for more comprehensive and diverse transformations, which contributes to improved model performance.

III. MODELS AND METHODS

Deeplabv3 and Deeplabv3+ are considered the best choices for this specific application among Convolutional Neural Networks [1]. These models are well suited for semantic segmentation, accurately delineating object boundaries - an essential capability for tasks such as road segmentation. Our exploration involved using various ResNet version, the elucidation of which is more thoroughly expounded upon in

Section III-D, in conjunction with these Deeplab architectures, exploiting their ability to capture intricate details essential for accurate segmentation. To provide a comparative baseline, we compared their performance to a simpler logistic regression approach.

A. *Baseline: Logistic regression*

We started by exploring Logistic Regression, which is a commonly used statistical model suited to binary tasks. To establish the groundwork, we divided the images into 16x16 pixel patches and calculated the mean and variance for each RGB channel within these patches, resulting in 6-dimensional feature vectors.The technique produced an accuracy of 0.492 and an F1 score of 0.344, which align with expectations for a simplistic model that performs complex image analysis tasks.

B. *DeepLabV3*

DeepLabV3 is a Convolutional Neural Network architecture designed for semantic image segmentation. One of the fundamental innovations of DeepLabV3 is the use of atrous convolutions (also known as dilated convolutions). Compared to traditional convolutional layers, atrous convolutions can capture a broader context by inserting gaps between filter weights, effectively increasing the receptive field without increasing the number of parameters or the amount of computation. Moreover, DeepLabV3 utilizes multiple scales of feature maps to make predictions. This is achieved by employing different atrous rates (or dilation rates) in the atrous convolutions. By combining features from different scales, the model can capture both fine-grained details and broader contextual information, improving the segmentation accuracy. Another architectural key element is the encoder-decoder structure. The encoder captures hierarchical features from the input image, while the decoder upsamples these features to generate a dense segmentation map. The skip connections between the encoder and decoder help preserve spatial information and improve the quality of the segmentation. The F1 score and accuracy produced by this pre-trained model are shown in Table I. Improvement over the baseline is noticeable and was expected.

C. *DeepLabV3+*

Then, we used an update version of DeepLabV3 : DeepLabV3+. DeepLabV3+ refines the encoder-decoder architecture, incorporating more sophisticated upsampling techniques and feature fusion strategies. This model can generate high-resolution segmentation maps with enhanced spatial coherence. Furthermore, DeepLabV3+ integrates Spatial Pyramid Pooling layers, allowing the model to capture contextual information at multiple scales. This multi-scale feature extraction enhances the model's ability to recognize objects of varying sizes within the image. Then, to reduce computational complexity without compromising the performance, DeepLabV3+ adopts depthwise separable convolutions in certain layers. This separation of spatial and channel-wise operations accelerates the inference speed while maintaining segmentation accuracy. With DeepLabV3+, and with data normalization, data augmentation and patches division as preprocessing, we obtained a F1 score of 0.819 and 0.905 for the accuracy at the 50-th epoch, outperforming DeepLabV3 as shown in Table I. Moreover, the convergence rate was faster, plots of the metrics were more stable and reached even higher peaks than for DeepLabV3 as can be seen by looking at the Figure 1, where are plotted the f1 score and the accuracy in the evaluation set for the DeepLab (on the left) and for the DeepLabPlus (on the right).

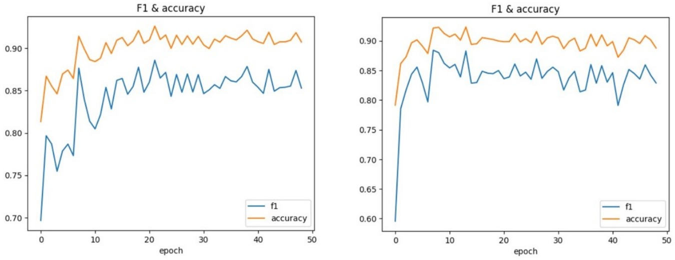

Fig. 1. Plot of f1 score and accuracy, calculated on the evaluation set while using DeepLab (on the left) and DeepLabPlus (on the right).

D. *ResNet*

ResNet stands for "Residual Network," is a type of Convolutional Neural Network architecture that was introduced to address the challenges associated with training very deep Neural Networks. One key feature of ResNet is the introduction of residual connections. These connections allow the activation (or output) from one layer to be directly added to the activation of a deeper layer. This can be represented as :

$$Output = Input + F(Input) \tag{1}$$

where $F(Input)$ represents the residual function learned by the layer. This bypassing of some layers helps in mitigating the vanishing gradient problem, making it easier to train deeper networks. ResNet architectures are characterized by their depth, with models ranging from relatively shallow variants like ResNet-18 to very deep variants like ResNet-152. The ability to train these deep networks effectively is attributed to the use of residual connections. In our case, we used ResNet-34 and tried also ResNet-50. For the first experiments explained previously, we used ResNet-34 for both DeepLabV3 and DeepLabV3+. The results are shown in Table I. After having received these good results (and especially with DeepLabV3+), we tried to integrate ResNet-50 with DeepLabV3 rather than ResNet-34. ResNet-50 is comprised of 50 layers. This depth enables the model to learn hierarchical features of increasing complexity, essential for tasks like image segmentation. Indeed, ResNet-50 offers increased depth and architectural complexity, making it more adept at capturing intricate features but at the cost of higher computational requirements. Whereas, ResNet-34 is a shallower and computationally efficient variant suitable for tasks where depth is not a critical factor. And indeed, the results obtained for the F1 score and the accuracy with DeepLabV3 and with

ResNet-50 as backbone architecture are higher than for the model DeepLabV3 with ResNet-34 as backbone architecture. With normalization, data augmentation and patches division as preprocessing, the F1 score and accurcay of DeepLabV3 with ResNet-34 are 0.774 and 0.875 respectively. Whereas for the same preprocessing, the F1 score and accurcay of DeepLabV3 with ResNet-50 are 0.807 and 0.898 (see Table I). However, a higher F1 score and accuracy were obtained with DeepLabV3+ and ResNet-34 (0.819 and 0.905), therefore we decided to use the pretrained model DeepLabV3+ with ResNet-34 to run a cross-validation over the different parameters : Foreground Threshold, Learning Rate and Patch size.

E. Optimizations

Optimizations for road segmentation refine models to pinpoint roads accurately in images. These enhancements fine-tune performance, cut down on errors, and boost efficiency. Techniques like fine-tuning parameters, choosing suitable loss function and optimizer, and employing smart data augmentation are essential in making these models reliable across diverse conditions.

1) Loss Function: Binary Cross Entropy (BCE) loss stands out in road segmentation tasks due to its efficiency in evaluating pixel-wise segmentation, making it a pragmatic choice for scenarios where computational resources are limited. BCE is define as follows:

$$\begin{aligned} l(x,y) &= (l_1, \ldots, l_N)^T, \\ l_n &= -w_n y_n \log x_n + (1-y_n)\log(1-x_n) \end{aligned} \tag{2}$$

Where N is the size of the batch. Its simplicity and ability to work well when local pixel information suffices for accurate segmentation make it a viable option, especially in applications where computational efficiency is crucial. However, we need to point out that with datasets having imbalanced class distributions, BCE loss might impact the model's ability to accurately distinguish road pixels from the background.

2) Optimizer: Optimizers play a crucial role in the training process, with Adam emerging as the widely preferred choice due to its efficiency and minimal need for fine-tuning the learning rate. During our experiments, we maintained a steady learning rate of 0.01. However, after determining the optimal parameters, thanks to the cross validation, we deliberately transitioned to a significantly slower rate of 0.001. This adjustment aimed to prevent abrupt model changes, ensuring a more stable optimization process. Additionally, it allowed us to avoid altering the weights of the pretrained models heavily.

3) Post processing: The preprocessing pipeline employed data augmentation to enhance the model's ability to understand different variations within images. This involved applying flips and rotations to both training and test images, enabling CNN to analyze images from multiple perspectives. When predicting the label for a patch within an image, the model generated predictions from all transformations applied to that patch and then averaged these predictions to obtain the final patch label prediction. This method allowed the model to gather comprehensive information from various viewpoints, leading to more robust and accurate predictions.

IV. Results and Conclusion

Table I presents an overview of our findings and results obtained through rigorous experimentation and analysis. Our exploration of various preprocessing techniques revealed that data augmentation is crucial for amplifying model performance. This outcome aligns with the demands placed upon Convolutional Neural Networks, which require copious amounts of data. The normalization of data was found to significantly improve performance in model training. However, not all preprocessing methods were beneficial. In fact, some methodologies exhibited a lack of noticeable improvements and, in certain cases, even led to decreased metrics. For example, the implementation of CLAHE had a negative impact on performance metrics. This provides valuable insights into the effectiveness of techniques in this field. Among the various models evaluated, DeepLabV3+ emerged as the most promising candidate, mainly due to its pre-trained architecture. Although DeepLabV3 coupled with ResNet50 demonstrated appreciable results, the computational cost outweighed the performance gains, leading to its exclusion from the DeepLabV3+ model evaluation. Additionally, the use of post-processing techniques was crucial in all experiments. Strategies such as averaging predictions from flipped and rotated inputs significantly improved metric outcomes, highlighting the importance of post hoc refinements in model predictions. After identifying the optimal pretrained model, we conducted cross-validation to fine-tune pivotal parameters such as learning rate, threshold, and image patch size. This iterative process significantly improved the model's predictive capabilities, resulting in a commendable F1 score of 0.878 on the AICrowd platform, indicating its robustness. We present a representative depiction of the obtained results from a sample image in Figure 2 for visual clarity and comprehension, providing a tangible illustration of the model's segmentation prowess.

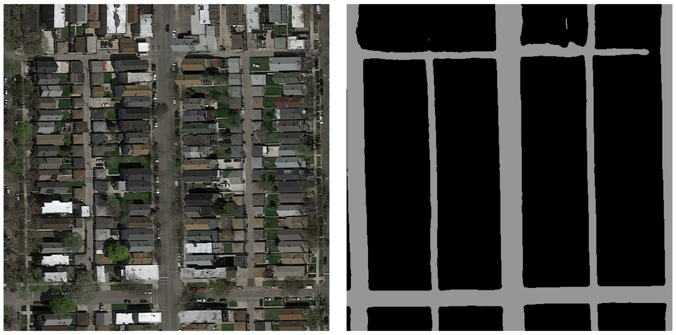

Fig. 2. Final model, prediction of the image 6 in the test set. On the left the original image, on the right the prediction of the model.

V. Ethical Risks

Ethical considerations play a crucial role in the development and deployment of machine learning models, particularly in tasks involving image analysis and classification. To ensure the responsible implementation of our road segmentation project, we conducted an ethical risk assessment using the Digital Ethics Canvas. An ethical risk identified is the potential bias in road segmentation, leading to misclassification or under-representation of certain communities or regions. The stakeholder which can be impacted are the communities residing in under-represented or marginalized areas. The negative impact is that the transportation authorities relying on accurate mapping data for urban planning and development, favor certain regions while neglecting others. Indeed, because of potential inaccurate representations of road networks, infrastructure development and resource allocation may be inadequate. And this can lead to the reinforcement of existing inequities in urban planning and development. This risk can have a significant severity as inaccuracies in road segmentation can have lasting impacts on communities and urban infrastructure. The likelihood of occurrence is moderate given the complexity of satellite imagery analysis and potential biases in training data sets. Then, to evaluate the identified ethical risk associated with potential biases or inaccuracies in road segmentation, we can do first a data set analysis. By analyzing the composition and diversity of the satellite imagery data sets, we can assess the representation of various geographic regions, community types, and road infrastructures. Then, the metrics we can measured to evaluate the risk can be the model's performance in segmenting roads, ensuring that the model achieves high levels of precision and recall across diverse data sets. We can also quantify any potential biases in the model predictions by focusing on regions or communities that may be under-represented or misclassified. Finally, in our project, to take these risks into account, we could incorporate diverse and representative data sets, ensuring inclusion of various geographic regions and community types. However, in our project we were not able to take into account these risks because the data set was given and it was difficult to ensure equitable representation and accuracy across all regions and communities and to verify that the satellite imagery data sets were unbiased.

VI. Future Work

Although we achieved a good result, we recognise the need for improvement: many adaptations, evaluations and experiments remain unexplored due to time constraints. The following concepts could be examined in the future:

- Add extra diverse and representative data of various geographic regions, community type and climatic conditions to improve the generalization, the robustness, the adaptability of the model and to ensure inclusion.
- Investigate the integration of advanced Convolutional Neural Network architectures, such as U-Net [2] or segment-anything [3], to potentially enhance the accuracy and efficiency of road segmentation.
- Consider training the DeepLabPlus Convolutional Neural Network using a Resnet50 instead of ResNet34, to consider the possibility of achieving better result in terms of f1 score for the evaluation and test set.
- Evaluate various loss function that are less sensitive to imbalanced data such as Dice loss or weighted BCE and compare their performance in terms of model convergence, accuracy and F1 score, particularly focusing on their ability to handle imbalanced datasets effectively.
- Further refine and extend the ethical considerations and fairness assessments, incorporating advanced metrics to ensure equitable representation and minimize biases across diverse communities and regions.

	F1 score	**Accuracy**
Baseline : Logistic regression	0.344	0.492
DeepLabV3		
DeepLabV3+patches+clahe	0.443	0.770
DeepLabV3+patches+clahe+norm	0.704	0.795
DeepLabV3+patches+clahe+augm	0.729	0.875
DeepLabV3+patches+norm+augm	0.774	0.875
DeepLabV3Plus		
DeepLabV3Plus+patches+clahe	0.503	0.796
DeepLabV3Plus+patches+clahe+norm	0.583	0.640
DeepLabV3Plus+patches+clahe+augm	0.749	0.880
DeepLabV3Plus+patches+clahe+norm+augm	0.556	0.730
DeepLabV3Plus+patches+norm+augm	**0.819**	0.905
DeepLabV3+ResNet-50		
DeepLabV3+ResNet-50+patches+clahe	0.499	0.792
DeepLabV3+ResNet-50+patches+clahe+norm	0.714	0.817
DeepLabV3+ResNet-50+patches+clahe+augm	0.645	0.851
DeepLabV3+ResNet-50+patches+clahe+norm+augm	0.242	0.764
DeepLabV3+ResNet-50+patches+norm+augm	0.807	0.898

TABLE I

F1 score and accuracy after 50-epochs.These experiments were run with a patch size of 80 (25 patches of 80x80 pixels).

References

[1] [1] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. "Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation."(2018).

[2] Ronneberger, O., Fischer, P. and Brox, T. (2015) 'U-Net: Convolutional Networks for Biomedical Image Segmentation', in N. Navab et al. (eds) Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015. Cham: Springer International Publishing (Lecture Notes in Computer Science), pp. 234–241. Available at: https://doi.org/10.1007/978-3-319-24574-4_28.

[3] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, Piotr Dollar´, Ross Girshick."Segment Anything " (Meta AI Research, FAIR) https://ai.meta.com/research/publications/segment-anything/