Road segmentation using deep learning

Malo Olszewski, Nino Avetikovi, François Mendiburu
EPFL, Switzerland

***Abstract*—In this study, we explore the application of convolutional neural networks (CNNs) for the task of road segmentation in satellite images. Different variations of UNet architecture were implemented (UNet, Res-UNet, MAResUNet) and evaluated. The model performances were assessed through F1 score and accuracy metrics. GitHub**

I. Introduction

Image segmentation remains one critical challenge within the domains of computer vision and image processing. In particular accurate delineation of road networks represents an important step for applications ranging from urban planning to navigation systems. To address road segmentation in satellite images, we adopted an approach employing three deep convolutional networks, all based on the U-Net architecture [1] . This architecture hinges on two primary components: the encoder and decoder. While the encoder focuses on extracting image features, the decoder reconstructs the feature maps. We particularly focused on the encoding part using both residual networks and attention mechanism.

II. Models and Methods

A. Dataset description

The dataset is composed of a hundred satellite images and their groundtruth (GT) mask. On the groundtrooth mask the pixels belonging to the roads are mapped by the label 1, whereas the other components are identified by 0. This dataset is relatively small considering the diversity of possible roads in the world (highways, country roads, dense cities roads).

Data augmentation was performed to increase the dataset, aiming to enhance its diversity and subsequently boost the generalization capabilities of the implemented models. The augmentation procedure involved random rotations of both images and their corresponding masks, the introduction of random Gaussian blur, cropping, flipping, and adjustments to contrast. This process expanded the dataset to a more diverse collection of 500 images, ultimately contributing to the models' improved ability to generalize across a broader range of scenarios and variations.

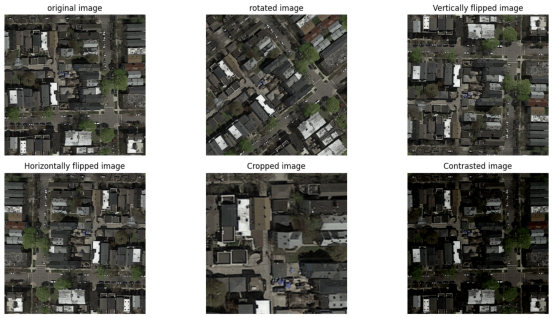

Figure 1: Data augmentation

B. UNet

Road segmentation was first performed thought the implementation of a UNet. The network consists of a contracting path, which extracts relevant features from the input images, and an expansive path that is involved in precise localization of the segmented object. We customized the U-Net model to suit the specific characteristics of our dataset, fine-tuning hyperparameters and employing data augmentation techniques to enhance its generalization capabilities. The typical architecture of a Unet is represented on the Fig.1.

The contracting path initiates with 3x3 convolutional layers, each employing a set number of kernels. The convolution step is followed by batch

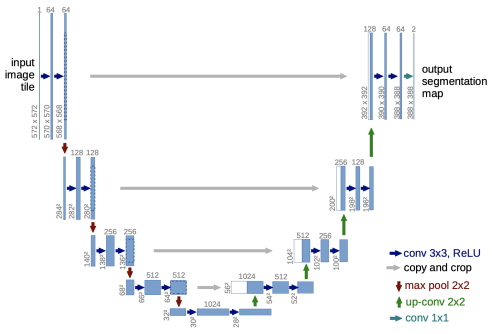

Figure 2: Unet Architecture [**?**]

normalization and rectified linear unit (ReLU) activation functions. These layers are responsible for capturing high-level features and reducing the spatial resolution of the input image. Interspersed with convolutional blocks are max-pooling layers, which downsample the spatial dimensions. The pooling operation aids in capturing context through a larger receptive field, facilitating robust feature extraction. The bottleneck layer acts as a bottleneck for abstract feature representation, connecting the contracting and expansive paths. The expansive path employs transposed convolutions 3x3 kernels to increase the spatial resolution of feature maps, ultimatly restoring spatial information lost during the contracting path. Skip connections, a distinctive feature of U-Net, involve the concatenation of feature maps from the contracting path with those in the expansive path. This ensures the preservation of fine-grained details and facilitates localized information for precise segmentation. Finally the output feature maps passed on to sigmoid activation function. This layer scales the output to the range [0, 1], producing the final segmentation mask. Each pixel in the mask represents the likelihood of belonging to the target class, such as a road.

C. Res-UNet

In order to deal with the problems that can occur during the training of the UNet, a second model was implemented, using residual blocks. Indeed, one can presume an increasing depth of neural networks leads to a greater accuracy. However, as networks get deeper, training becomes more challenging. One significant issue is the vanishing gradient problem, specifically encountered during back-propagation. As gradients traverse through the layers during this process, they can diminish significantly, particularly in networks with numerous layers. This drastic reduction in gradient values can render the network untrainable due to effectively null gradients.

This problem can partly be addressed with normalized weight initialization and layer normalization, but residual networks can also do this and solve the degradation problem [2]. With a substantial number of layers, accuracy plateaued due to the extensive layering, leading to an observable sudden drop out. ResNet introduces residual blocks—sets of layers where data not only passes through but also bypasses via skip connections (Fig.3).

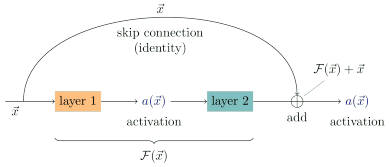

Figure 3: Skip connection

Skip connections allow information to propagate through the network without being influenced or altered by all the intermediate layers. Each residual block is composed of convolutional layers. The general structure of some pretrained models already existing on Pytorch.

Whereas a simple UNet uses only the encoder-decoder properties to map features, the use of a supplementary residual network in encoding offers better accuracy. This kind or architecture has high efficiency for image recognition, and we used the properties of transfer learning (pre-trained weights) and fine tuned this specific architecture (ResNet34) to suit our application. However, the large amount of layers requires more time to run (see Results section).

D. MAResUNet

Attention mechanism [3] is known to refine the extracted feature maps in the context of computer

vision or natural language processing, and can hence be used for our segmentation task. A simplified architecture of Multistage Attention ResU-Net (MAResUNet) [4], was implemented,designed to increase performances in remote sensing segmentation. To streamline the architecture, we decided to reduce the 4 layers in encoder and decoder paths to 3, sufficient for the task of road segmentation (Fig.4).

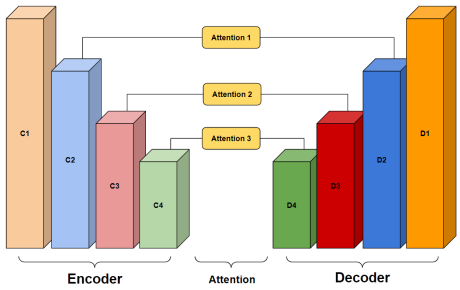

Figure 4: Architecture of the 3-layers MAResUNet

This architecture also use a Linear Attention Mechanism (LAM) which relies on the dot-products between the query, key and value matrices (Fig5). In fact, we saw that residual networks and UNet have high memory requirements and can be computationally expensive, and this solution solution offers cheaper costs.

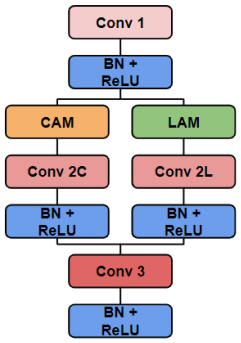

Figure 5: Attention block

Finally, the encoder generates feature maps are comprised of low-level and fine-grained detailed information (borders), and the decoder maps the higher level and coarse-grained information (roads).

III. Results: Training and Validation

The three models were trained on 100 epochs with a batch-size of 8. The parameter optimization was performed using AdamW algorithm, implemented in pytorch and extends Adam by incorporating weight decay directly into the weight update step. It uses first and second moment estimates to adaptively adjust learning rates for each parameter during optimization. This prevents overfitting by penalizing large weights. Initially the learning rate, and the weight decay parameters, were set to 5e-4 and 1e-3 respectively.

The training was based on binary cross-entropy loss function:

$$\text{Loss}(\hat{y}, y) = -\left(y \log(\hat{y}) + (1-y)\log(1-\hat{y})\right) \tag{1}$$

where $\hat{y}$ represents the predicted output, and y is the ground truth label. Additionally, a cosine annealing learning rate scheduler is implemented, modulating the learning rate over iterations (t). The overarching objective is to iteratively optimize the model's parameters, guided by these defined optimization and learning rate adjustment strategies, over the stipulated number of epochs.

F1 score and the accuracy were used as performance evaluation metrics, and served as ranking criteria on AIcrowd.

The training was performed using T4 GPUs available on Kaggle.

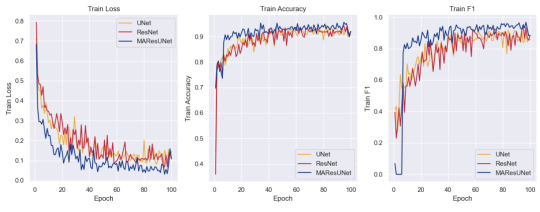

Figure 6: Training performances

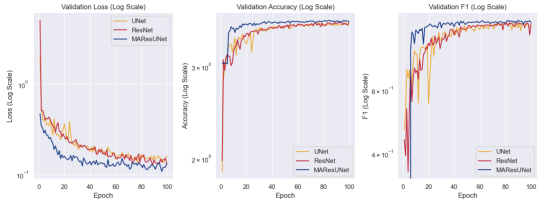

Figure 7: Validation performances

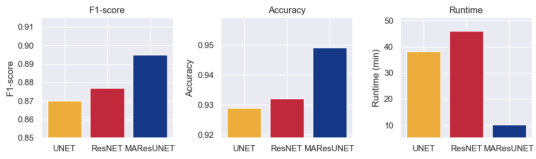

Figure 8: Testing and runtime performances

The models were compared based on F1 score, accuracy as well as the runtime performances. Given the results, it appears that MAResUNet offers the best results on every aspect. Then, we decided to asses the effect of the number of epochs on the model performances (Fig. 9). Highest performances were obtained at 300 epochs :

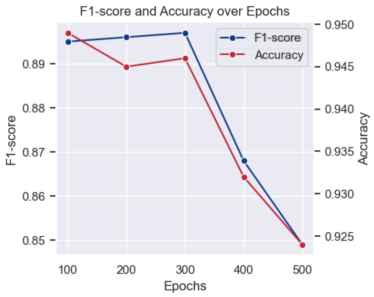

Figure 9: MAResUNet performances with increasing epochs

The best performances are obtained for 300 epochs after a runtime of 38min12s. The results on training images and testing images are represented on figures 10 and 11 :

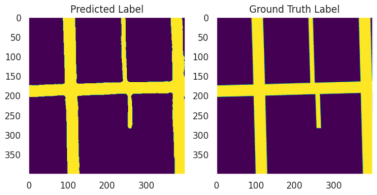

Figure 10: MAResUNet with 300 epochs on training with groundtruth

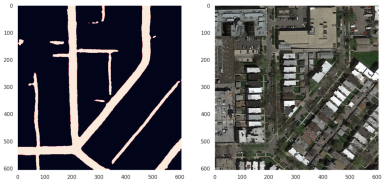

Figure 11: MAResUNet with 300 epochs on testing with image

IV. Ethical risk

Processing satellite images can present potential defense-related issues, especially in zones that are restricted. Additionally, our dataset predominantly consists of images from residential areas and cities, potentially introducing bias during testing, particularly when applied to countryside roads. It's crucial to acknowledge these biases in both the dataset and our methodology when utilizing this dataset.

V. Conclusion

To conclude, we highlighted the performances of three different U-Net-based architecture, with and without data augmentation, and it is the one which uses attention and data augmentation that works better. It is also important to consider that ViTs (vision transformers), which relies on self-attention mechanism, could offer the best results, considering their generalization performances [5].

References

[1] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. *CoRR*, abs/1505.04597, 2015.

[2] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for

image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[3] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *CoRR*, abs/1706.03762, 2017.

[4] Rui Li, Jianlin Su, Chenxi Duan, and Shunyi Zheng. Multistage attention resu-net for semantic segmentation of fine-resolution remote sensing images. *CoRR*, abs/2011.14302, 2020.

[5] Kai Han, Yunhe Wang, Hanting Chen, Xinghao Chen, Jianyuan Guo, Zhenhua Liu, Yehui Tang, An Xiao, Chunjing Xu, Yixing Xu, Zhaohui Yang, Yiman Zhang, and Dacheng Tao. A survey on visual transformer. *CoRR*, abs/2012.12556, 2020.