ML Project 2 - Road Segmentation using U-Net

Naguib Shadi, Alvarez Léo, Rodriguez Mateos Borja

***Abstract*—Image segmentation is a fast-growing machine learning field for medical, aerial images and autonomous application, and many models have been developed to tackle this task. Particularly, the U-Net encoder-decoder architecture is among the most popular for pixel-wise classification. We implemented variants of the vanilla U-Net to predict whether a pixel is a road or background. Our best model yield to 0.947 accuracy and 0.903 F1 score.**

I. Introduction

This project aims to implement semantic segmentation on satellite images to predict if pixels are road1 or background0. Images from Google Maps and their ground-truth representation (pixel-wise labels) were provided, see Figure 1. The train dataset is composed of 100 pairs of RGB images and ground-truth of 400×400 pixels, the testing dataset has 50 satellite RGB images of 608×608 pixels. Because there is not much data, a pre-processing was applied to augment the size of the dataset. As explained in Section II-A, it also increases the proportion of images with tilted roads that were underrepresented in the original dataset (only 15% of images had angled roads). Convolutional neural networks are proved to be the most suited for image segmentation, thanks to their consideration of spatial information. A simple Convolutional Neural Network is not enough to tackle this segmentation task. Section II-C details the U-Net architecture, a state-of-the-art model for image segmentation originally developed for medical images that we adapted to our road segmentation task and called *U-Net Alpha*. We then propose variants of the vanilla U-Net, *U-Net Beta* and *U-Net Gamma* and present our improved results.

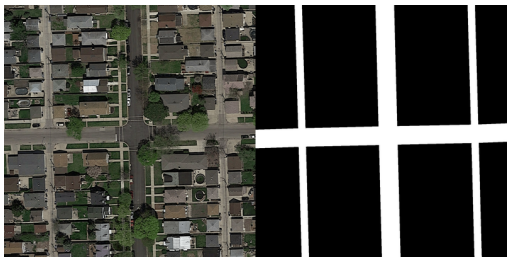

Fig. 1. Example of training data / groundtruth

II. Methodology and Models

A. Data Augmentation

Usually, convolutional neural networks need a very large amount of data to be trained [**?**]. The encoding-decoding structure of the U-Net requires fewer data than CNN to train [2]. Nevertheless, 100 images is not enough, and we augmented the dataset applying the following transformations on each image and its ground-truth similarly:

- Horizontal mirror
- Vertical mirror
- Rotation of angles : -60°, -45°, 15°, 45°, 60°, 90°, 180°

It appears that the training set is unbalanced with regard to the orientation of the roads. Only 15% of the train images have tilted roads. The rotation transformations allow equilibrating the dataset and allow the network to train and learn better.
Another augmentation occurs at the initialization of the dataset. We crop the images into patches of 80×80 pixels with an overlap of 40. This generates multiple patches from a single image and allows augmenting the dataset size. The overlap between patches (alternately on one side and the other of images) increases the redundancy in the data and avoid edge induced errors in the prediction.

B. Convolutional Neural Networks

In the last few years with the amount of data available being bigger and bigger, Convolution Neural Networks (CNN) have become the very performant in the fields of machine vision. [10]. For this reason, we have trained a classificatio model with this technique and have taken it as our baseline comparison to evaluate the performance of other models. It composes of two 2D convolutions, each followed by a Relu activation and a Max pooling layer. At the end, the feature map cuboid is reshaped into a 2D matrix and is fed to three fully connected layers for classifying the input. The results of this model can be found in comparison with the other models, found in table II

C. U-Net Architecture

The U-Net is a computer vision model developed at first for medical datasets ([2]), but is now also widely used on non-medical images such as aerial images [3] [4]. The architecture is based on a feature encoding and classification decoding parts, as shown on Figure **??**. The encoder aims at extracting significant features from the input images using a succession of 3×3 convolutions and 2×2 max pooling operations. The decoder segment tries to *combine* the extracted features and generate a prediction of the road's position on the image with a combination of 2×2 transposed convolutions and 3×3 convolutions. The particularity of U-Net is the skip connections between the two parts: before the pooling operation, the outputs of the encoder path are concatenated to the entry of the decoder blocks (just after the transpose convolution). This allows the model to retrieve the spatial information lost during pooling down steps. All convolutions'

outputs are batch normalized and activated by the Rectifier Linear Unit function, except for the last convolution.

D. U-Net Variation

We built and trained three main variants to the vanilla U-Net architecture, adding different characteristics, partly inspired from the MultiRes U-Net built for medical images [5]: U-Net Alpha, U-Net Beta and U-Net Gamma.

U-Net Alpha: *Padded Convolutions:* Unlike the original U-Net published in 2015, our initial implementation U-Net Alpha retain the dimensions of the input image in the output prediction. This is done using zero-padding at each convolution, in the encoding and decoding phases, to compensate the loss of border pixels when applying 3×3 convolutions.

Another difference lies in the number of features at each convolutional step. We implemented convolutions with gradually doubling from 16 to 256 output filters, opposed to 64 to 1024 for the U-Net.

Lastly, the final classification of U-Net Alpha is done using a 1×1 convolution without batch normalization, activated by the Sigmoid function.

U-Net Beta: *Dilated Convolution:* The prediction of the U-Net Alpha were lacking of large scale spatial information: some roads were cut and not predicted when covered by a tree or railway for instance. The idea is to use dilated convolutions to increase the receptive field of filters and take better benefit from the spatial interconnections with neighboring pixels [6] [7].

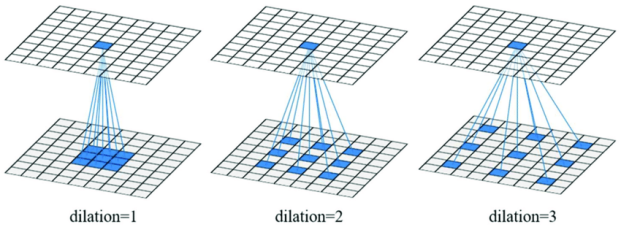

Fig. 2. Receptive field of dilation rate 1, 2 and 3

We replaced the first convolution of the first two encoding blocks respectively by dilated convolutions with rate 3 and 2, to imitate 7×7 and 5×5 kernels. A zero-padding of respectively 3 and 2 were added to preserve the input dimensions.

U-Net Gamma: *ResPaths Skip Connections:* One improvement mad to the U-Net Beta is the modification of the skip connection paths. Instead of simply combining the encoded feature map to the decoder blocks, the residual go through a series of convolution operations. The goal is to alleviate the disparity between the encoder and decoder feature maps, because the decoder input have gone through more convolutions [5]. The structure is called a ResPath because it uses residuals at each convolution.

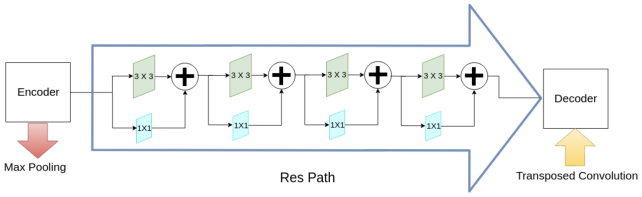

Fig. 3. ResPath with 4 convolutional blocks

Our models have a depth of 5, with 4 skip connection paths. Like in the MultiRes U-Net paper, we use a decreasing number of convolutional blocks for ResPaths (e.g. 4 blocks in the Figure 3). In particular, we use 4, 3, 2 and 1 block for ResPaths from top to bottom. This allows to compensate more or less the effect of encoding and decoding convolution according to the depth.

E. Training

The main challenge and limitation in this project was the memory allocation and the training time, which are both particularly high with deep convolutional neural networks.

1) Pre-processing: After optimizing the hyperparameters using theoretical consideration, empirical verification and trial & error, we prepared the data with patches of 80 pixels with an overlap of 40. The data pre-processing yields to a dataset of more than $72'000$ augmented train patches. The overlap add sparsity in the dataset and allows generating even more patches from a single image. We fed the models with batches of 64-image batches to produce the best results. Strangely, the best results were not obtained with big images such as 572×572 pixels, the one used in the MultiRes U-Net. The Table II shows the number of trainable parameters, epochs and training time for the three variant models.

U-Net Model	U-Net Alpha	U-Net Beta	U-Net Gamma
Parameters	$1'944'049$	$1'944'049$	$2'163'089$
Epochs	50	70	70
Training Time	4	4	4

TABLE I
TRAINING HYPER-PARAMETERS FOR U-NET VARIANTS

2) Criterion and optimizer: As the prediction is a pixel-wise binary classification, the Binary Cross Entropy loss was used to evaluate the correctness of the trained weights. It is given by equation 1.

$$l_{BCE}(x,y) = -\frac{1}{N}\sum_{i=1}^{N} x_i \cdot log(y_i) + (1-x_i) \cdot log(1-y_i) \quad (1)$$

With $x_i, y_i \in \{-1, 1\}$ and x and y of the same dimensions. The Adam optimizer was used to handle the adaptation of the learning rate throughout the training process, starting from $lr = 1e-3$.

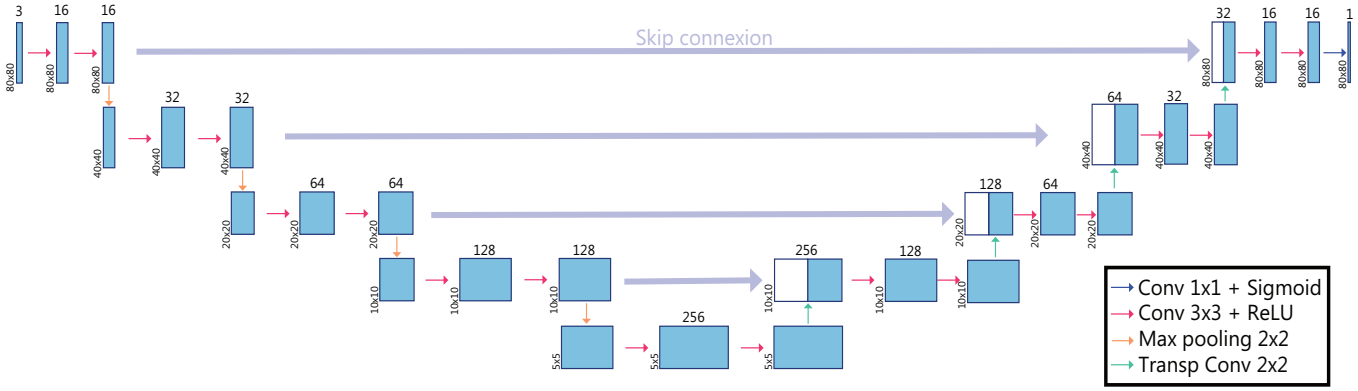

Fig. 4. Unet architecture

3) Evaluation: During the optimization phase, models were trained on 90% of the training set, and evaluated on the remaining 10% at each epoch. The output of the Sigmoid activation are round up to 1 or down to 0 to classify each pixel as road or background. The metrics used to evaluate are the F1-score and the accuracy. Cross-validation was not worth considering because of the duration of the training process.
As shown on Figure 1, the dataset is unbalanced between the classes background and road. Indeed, in the dataset, there is 81.3% of background and 18.6% of road. When the skew in the class distributions are severe, accuracy can become an unreliable measure of model performance. The F1-score is the metric of interest in order to take this unbalance into account. It is given by equation 2.

$$F_1 = \frac{2}{\text{recall}^{-1} + \text{accuracy}^{-1}} \tag{2}$$

4) Batch normalization and dropout: We implemented batch normalization after each convolution (except the last classification one). This is an efficient way to reduce the risk of overfitting on the training set, it avoids the internal covariate shift that can lead to huge loss of performance in deep neural networks. t also speeds up the training process because smaller learning rate can be used. Furthermore, it has a regularization effect that avoid to use dropout. Adding a dropout yield to worse results in our model, so we did not use it in the final versions.

5) Software setup: We trained the models using cuda to take benefit from the efficient computing of GPUs. We ran the training on Google Collab.

III. Results

The Figure 5 shows the ground-truth 608×608 prediction of an image from the testing dataset using the U-Net Beta model. We see that most of the roads are well discovered.

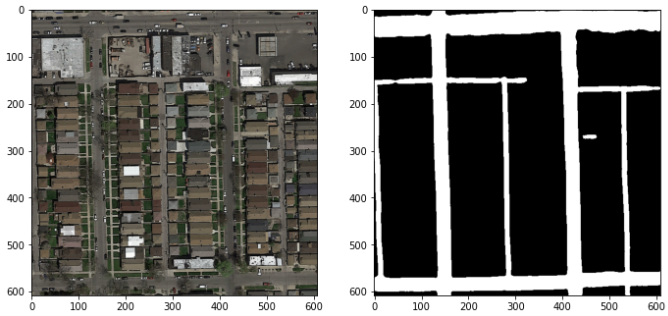

Fig. 5. Test image and its predicted groundtruth with U-Net Beta model

A. Baseline Model

B. U-Net Variants

The three variants of the U-Net maje better predictions than the As expected, the U-Net Beta model implementing dilated convolutions yields to better results than the initial U-Net Alpha model. This can be explained by the consideration of a wider context around pixels, allowing to predict a pixel using more spatial information.

U-Net Model		U-Net Alpha	U-Net Beta	U-Net Gamma
F1-Score	AICrowd (Test)	0.886	0.903	0.896
	Validation	0.967	0.963	0.961
Accuracy	AICrowd (Test)	0.940	0.947	0.942
	Validation	0.967	0.963	0.960

TABLE II
VALIDATION AND KAGGLE PERFORMANCE RESULTS FOR U-NET VARIANTS

IV. Discussion

V. Conclusion

To go further : interpolation with Hough transformations

REFERENCES

[1] D'souza, R.N., Huang, PY. & Yeh, FC. *Structural Analysis and Optimization of Convolutional Neural Networks with a Small Sample Size.* Sci Rep 10, 834 (2020). h

[2] O. Ronneberger, P. Fischer, T. Brox *U-Net: Convolutional Networks for Biomedical Image Segmentation*. University of Freiburg, Germany, 2015

[3] J. McGlinchy, B. Johnson, B. Muller, M. Joseph and J. Diaz, *Application of UNet Fully Convolutional Neural Network to Impervious Surface Segmentation in Urban Environment from High Resolution Satellite Imagery*, IGARSS 2019 - 2019 IEEE International Geoscience and Remote Sensing Symposium, 2019, pp. 3915-3918, doi: 10.1109/IGARSS.2019.8900453.

[4] S. Karki and S. Kulkarni, *Ship Detection and Segmentation using Unet*, 2021 International Conference on Advances in Electrical, Computing, Communication and Sustainable Technologies (ICAECT), 2021, pp. 1-7, doi: 10.1109/ICAECT49130.2021.9392463.

[5] N. Ibtehaz and S. Rahman, *MultiResUNet : Rethinking the U-Net Architecture for Multimodal Biomedical Image Segmentation*, Department of CSE, BUET, ECE Building, West Palasi, Dhaka-1205, Bangladesh

[6] T. Yamashita, H. Furukawa and H. Fujiyoshi, *Multiple Skip Connections of Dilated Convolution Network for Semantic Segmentation*, 2018 25th IEEE International Conference on Image Processing (ICIP), 2018, pp. 1593-1597, doi: 10.1109/ICIP.2018.8451033.

[7] Huszar, F., *Dilated convolutions and kronecker factored convolutions*, 2016, http://www.inference.vc/dilated-convolutions-and-kronecker-factorisation/

[8] J. Brownlee *A Gentle Introduction to Batch Normalization for Deep Neural Networks*, 2019, https://machinelearningmastery.com/batch-normalization-for-training-of-deep-neural-networks/

[9] Pipeline Diagram and Datapath https://docs.microsoft.com/en-us/windows-hardware/drivers/gettingstarted/writing-a-very-small-kmdf--driver

[10] N. Sharma, V. Jain, A. Mishra *An Analysis Of Convolutional Neural Networks For Image Classification*, 2018, https://www.sciencedirect.com/science/article/pii/S1877050918309335

[11] A. Althnian, D. AlSaeed, H. Al-Baity, A. Samha, A. Bin Dris, N. Alzakari, A. A. Elwafa and H. Kurdi, *Impact of Dataset Size on Classification Performance: An Empirical Evaluation in the Medical Domain*, 2020, Information Technology Department, College of Computer and Information Sciences, King Saud University.