Image Segmentation using U-Net, SegNet and Dilated Convolution

Gurjeet Singh
gurjeet.singh@studenti.unipd.it

Francesca Zen
francesca.zen@studenti.unipd.it

Abstract

The task of image segmentation has seen important developments in the last few years and improvements on convolutional neural networks are leading to higher results in this area. In this paper we analyze semantic segmentation task using novel and practical U-Net, SegNet, and Dilated Convolution architectures. We apply them to the Massachusetts Buildings and Cityscapes datasets, with the aim of evaluating their performances on the given task on different dataset types with varied object categories. In particular, we focus more on the former unbalance dataset, which requires more study and various techniques for achieving remarkable results. Our experiments show that the Dilated Convolution model provides more accurate outcomes on both Massachusetts Buildings and Cityscapes datasets, as compared to the other architectures.

1 Introduction

Image segmentation models have been deeply raised in many diverse contexts, like autonomous vehicles, satellite image analysis and medical image diagnostic; thanks to their wide range of applications they are nowadays a great interest of science and research. In semantic segmentation each pixel is classified according to the class of the object it belongs to (e.g., road, car, pedestrian, building, etc.). Each of the pixels in a region is similar with respect to some property, such as color, intensity, location, or texture. Different objects of the same class are not distinguished, the only care is on the category of each pixel.

The state-of-the-art in semantic segmentation today is achieved using complex architectures based on convolutional neural networks (CNNs). We implemented U-Net, SegNet and Dilated Convolution architectures that are particular widely used deep CNNs, which are the basis of the other advance variations. The architectures are differentiate from each other for the encoder-decoder choice and the use of Maxpooling and Upsampling layers. In detail, U-Net and SegNet use the VGG16 architecture as encoder but they have different decoder techniques; instead the Dilated Convolution is a CNN that uses dilated convolution operations and it is more efficient for detecting objects that are small and crowded in remote sensing imagery [11, 12, 6]. Indeed, in our experiment Dilated Convolution NNet is most performing on the aerial image segmentation, thus it can deal with unbalanced datasets like Massachusetts Buildings. SegNet instead, makes use of Maxpooling and Upsampling layers in the encoder-decoder respectively and performs poorly on this last dataset but can achieve higher results in the Cityscapes dataset.

In order to make a fair comparison between these three architectures on the two datasets, we measured their performances using the Intersection over Union (IoU), the F1-Score and the Tversky metrics. The latters return representative values in the computation of the similarity between two images, also for unbalanced datasets.

The most performing architecture is the Dilated Convolution, which is able to reach an IoU score of 76,14% on the test set of the Massachusetts Buildings dataset as reported in Table 1, and 62,33% in Cityscapes dataset as shown in Table 2. In the latter dataset, Dilated Convolution does some more overfitting than SegNet but again is the architecture that returns higher performances in test and validation sets.

2 Related Work

When facing the problem of Semantic Segmentation we investigated which characteristic an architecture must have in order to be efficient when working with a specific dataset. In fact, it could be helpful to discover some common features in the datasets for then applying the architecture that can perform the segmentation tasks better than others. For example, from paper [10] we have that SegNet is one of the most performing architecture when dealing with roads and indoor scene understanding, efficient both in term of memory and computational time. Unfortunately, our results on the Cityscapes dataset are not in line with B. Badrinarayanan et al. because the resources on our disposal were not sufficient for reaching high performances and SegNet does not appear as the best choice for this road scenes segmentation.

One of the main problems when applying CNNs on semantic segmentation tasks is the down-sampling with pooling layers, which increases the field of view of convolutional kernels but loses at the same time high-frequency details in the image, as said in [1]. In fact, many approaches was able to improve the results but the buildings boundaries were still not well segmented. Hence a reconstruction of the receptive fields has to be made properly using the right upsampling technique.

By considering the problematic linked to the Massachusetts Buildings dataset, as mentioned in [1], one of the main issues when dealing with satellite images is preserving semantic segmentation boundaries in high resolution especially for little objects. B. Bischke et al. introduced a new cascaded multi-task loss to overcome the problem of "blobby" predictions. We followed their objective with the aim to enhance the predicted images, and we tested two different losses, the Binary Cross Entropy Jaccard Loss and the Dice Tversky Loss, on the Massachusetts Buildings dataset using all of the three architectures. Looking at [3] and [5], we have expected that Dilated Convolution and U-Net are the most performing on this dataset and as we can see from Table 1 the results are in line with the two papers.

On the other hand, for the work on the Cityscapes dataset we used the CE Jaccard Loss function, and like [1], we evaluate the importance of different decoder architectures using U-Net and SegNet architectures.

3 Dataset

To illustrate the effectiveness of the three architectures presented we used two datasets: the Massachusetts Buildings [8] and Cityscapes [2] datasets, because they cover different imagery characteristics such as spatial resolution, object types, shapes, sizes and number of classes.

In particular, the Massachusetts Buildings dataset is a binary (*building* and *non-building*) semantic segmentation problem and it consists of 151 aerial images of size 1500x1500 pixels of Boston area. Each image covers 2.25 km^2 at a resolution 1 $m^2/pixel$. These images are randomly split into training, validation and test sets with sizes 137, 4 and 10 respectively. In order to increase the number of samples in each set, the images were divided in patches, this procedure is also proposed by the author of the dataset as Patch-Based labeling framework for achieving high qualitative results. For this reason we divided every image in 9 disjoint patches each. In addition, due to the high resolution of the images which makes the convolution operation expensive and training time longer, we decided to lower the images resolution to 256x256 pixels. Indeed, as described in paper [5], this approach has shown results to be close to the trained model with original resolution images. In order to increase generalization of the model we extended the dataset by applying data augmentation at training time by choosing randomly a set of transformations like horizontal flip, vertical flip, rotations and random crop. This technique helps in building a stronger model which is less dependent on the input image orientation. This is very helpful to our model to generalize different regions [5]. We further introduced and combined the Massachusetts Roads Dataset [8] to the previous dataset for some particular analyses made on the SegNet model that we explained in the experimental chapter.

On the other hand, the Cityscapes dataset contains street scenes over 50 different German cities, with high-quality pixel-level annotations of 5000 frames in addition to a larger set of 20000 weakly annotated frames. But for our analysis, we selected a relatively small portion of it, because of the limited computational resources we had. Indeed we considered 1150, 267 and 472 fine annotations for the training, validation and test sets. The dataset is made of 30 different labels, however, we mapped more objects

into a single category to reduce the number of labels from 30 to 8, as a coarse dataset. In such a manner, the segmentation resulted to be less differentiated but able to get the right division in fewer epochs. As for the previous dataset, we down-scaled the images to 256x256 pixels to make the training faster for the model.
Supplemental preprocessing methods were applied on each encoder, by adapting the images channels to the desired input and by zero-centering them on each channel with respect to the ImageNet dataset when using pretrained weights.

4 Method

In this paper we worked on some recent developed neural networks in order to assess which architecture is more reliable for the chosen datasets. Our objective is to compare different types of decoders and up-sampling techniques and for this reason we analyzed SegNet, U-Net and Dilated Convolution NNet models. Thanks to the diverse chosen datasets we were able to explore each architecture features and comprehend their properties.
Once fixed the architecture and the datasets we implemented the models and applied the preprocessing techniques described previously, afterward we studied which metrics and loss were more reliable for each problem and finally we fine-tuned and analyzed our results.

4.1 The Architectures

There are several models available for semantic segmentation. Usually, deep learning based segmentation models are built upon a base CNN network and in our case we choose VGG16 as the base network both for U-Net and SegNet encoders. This pretrained model was proposed by Oxford, which got 92.7% accuracy in the ImageNet 2013 competition and with its 16 layers is faster to train comparing to others.

4.1.1 SegNet

SegNet has an encoder network which is used to capture the context in the image, composed of 13 convolutional layers from VGG16; followed by a corresponding symmetric expanding path, the decoder network, and then ended by a final pixelwise multiclass softmax classification layer. In the decoder layers, upsampling and convolutions are performed. In particular, non linear upsampling is computed by using the stored indexes of the max-pooling encoder layers. This architecture improves boundary delineation and reduces the number of parameters enabling a faster end-to-end training comparing to the FCN architecture [10].

4.1.2 U-Net

U-Net was developed by Olaf Ronneberger et al. for Bio Medical Image Segmentation and its architecture contains an encoder, which is a traditional stack of convolutional and max pooling layers and correspond to the contraction path, followed by the decoder, which correspond to the expanding path. This last component, in particular, is used to define a better reconstruction of the features map thanks to transposed convolutions and concatenation of the encoder features map followed by regular convolution operations [9].

4.1.3 Dilated Convolution

Dilated Convolution is designed for dense prediction and support exponentially expanding receptive fields without losing resolution or coverage [12].
A dilated convolution is defined as following. Let $F : \mathbb{Z}^2 \to \mathbb{R}$ be a discrete function. Let $\Omega_r = [-r, r]^2 \cap \mathbb{Z}^2$ and let $k : \Omega_r \to \mathbb{R}$ be a filter of size $(2r + 1)^2$. The dilated convolution operator $*_l$ can be defined as

$$(F *_l k)(\mathbf{p}) = \sum_{s+lt=\mathbf{p}} F(s)k(t).$$

We can see that at the summation is $s + lt = \mathbf{p}$, so we will skip some points during convolution. When $l = 1$ we have the standard convolution.
Thus by taking in account its property we defined a CNN architecture based on the dilated model presented by Hamaguchi et al. [3].

4.2 Metrics for the Evaluation

In order to do a fair comparison among the NNets we used particular metrics for their evaluation depending also on the dataset. The main and well known performance measure used for the semantic

segmentation is Intersection over Union (IoU), also called Jaccard Index. We also used F1/Dice Score metric, since we rely on the average performance, and Tversky metrics for dealing with the imbalanced Massachusetts Buildings data, which can be helpful also as a loss function. The Tversky index is defined as:

$$TI(GT, P) = \frac{|GT \cap P|}{|GT \cap P| + \alpha|GT - P| + \beta|P - GT|},$$

where GT and P stand for the ground truth and the predicted mask, respectively, α and β positive parameters. It is noteworthy that in the case of $\alpha = \beta = 1/2$ the Tversky index simplifies to be the same as the Dice coefficient. These metrics are positively correlated and lie in a range between 0 and 1, with 1 as the greatest similarity between predicted and truth.
Based on these metrics we used then the associated cost functions, which are the widely used Jaccard-Cross-Entropy loss function, and also the Focal Tversky loss function to deal with the imbalanced dataset issue; the latter one is defined as following:

$$FTL = \sum_c (1 - TI_c)^\gamma$$

where TI is the Tversky Index and γ range between $[1, 3]$.

5 Experiments

In this work we implemented all the neural networks using Keras framework and other libraries[1], which allow to build U-Net and SegNet models in a fast and easy way. The libraries use Keras as backbone, hence we just needed to adapt them to our datasets and define the hyperparameters needed for our binary or multiclass classification problems without building the neural network architecture from scratch. For the Dilated Convolution, instead, we defined a scratch Keras model by following [3] and taking into account our limited resources we simplified their model by reducing the number of convolution filters in the last layers, due to computational and memory shortage in Kaggle machine.

[1]https://pypi.org/project/segmentation-models/1.0.1/ https://pypi.org/project/keras-segmentation/0.3.0/

Model	Epochs	Loss	mIoU	F_1 Score
U-Net	40	Dice Tversky	74.04	84.16
SegNet	40	BCE-Jaccard	45.44	57.07
Dilated	50	BCE-Jaccard	76.14	85.4

Table 1: In this table we report the percentage of the main results on the Massachusetts Buildings dataset. In particular, U-Net and SegNet were trained with Adam (lr = 10^{-3}) and batch-size of 12 and 32 respectively, while Dilated Convolution was trained with SGD (lr=0.01) and batch-size of 2.

Model	Loss	Epochs	mIoU	F_1 Score
U-Net	CE Jaccard	40	58.59	67.92
SegNet	CE Jaccard	20	45	54.61
Dilated	CE Jaccard	60	62.33	71.59

Table 2: In this table we report the percentage of the main results on the Cityscapes dataset. In particular, U-Net was trained with Adam (lr=10^{-3}) and batch-size of 12, while SegNet and Dilated Convolution were trained with SGD (lr=0.1) using batch-size of 12 for the first architecture and batch-size of 4 for the second.

For the U-Net and SegNet architectures we initialized the encoder using the weights of ImageNet and then froze them for a better regularization and learning. As a matter of fact, we achieved better results and made the learning faster during the training phase. We tested each architectural variant over both Massachusetts Buildings and Cityscapes datasets, by trying and tuning different hyperparameter one by one, like the optimizers, learning rates, batch-sizes and loss functions. In particular, we considered Adam and Stochastic Gradient Descent (SGD) with Nesterov momentum, with learning rates in a range between 10^{-1} and 10^{-3}. We tried different batch sizes among the architectures: for U-Net and SegNet we tested mini-batches of 12, 32, and 64; while for Dilated Convolution we used 2 and 4 as batch size because of lacking GPU memory. We trained them until overfitting behavior appeared.
As a summary, in Tables 1 and 2 we reported the best results in terms of mIoU and F_1 Score for both the datasets, highlighting the loss function used.

Massachusetts Buildings. As stated in [3] and [5], the best results on this dataset comes from U-Net and Dilated Convolution models, which are able

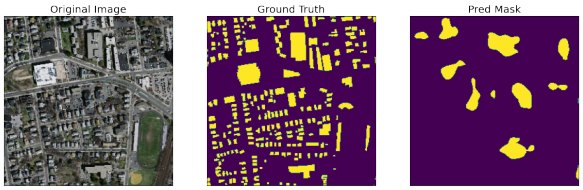


Figure 1: A sample of SegNet prediction on the Massachusetts Buildings test set, highlighting the sparse mask prediction due to the bad decoder receptive field learnt from its upsampling technique on the imbalanced dataset.

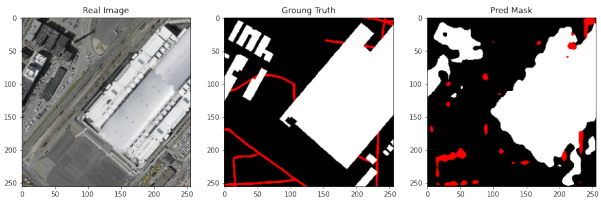
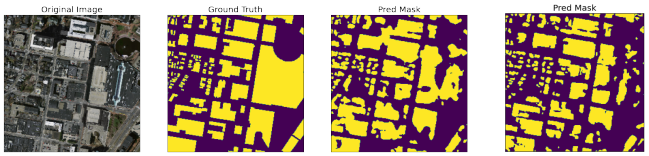


Figure 2: A sample of the combined Massachusetts datasets, with ground truth mask on the left and predicted image mask on the right from SegNet network. The sample highlight the impact of the added content in the prediction, a result achieved by the model just within 20 epochs.

to reconstruct the receptive field and build the segmented image by achieving remarkable results as shown in Table 1. On the other side, the SegNet model is not able to achieve meaningful results due to its simple up-sampling technique used in the decoder layer. Indeed, from Figure 1 we can see that the model is not able to capture the buildings blocks, shapes and positions. These results are most probably because of the strong unbalanced dataset. To solve this issue we firstly tried to train SegNet using other loss functions that tackle such problems, for example by using *Tversky Loss* and *Focal Tversky Loss* with higher value of α in order to increment the false positive cases. However, we were not able to achieve better results and the predicted masks were still sparse images. Furthermore, finding the best suitable hyperparameter of the loss functions was not feasible because of the big search space required and time consuming tests. Hence to investigate more this problem we decided to increase the content information of the mask images to make the dataset less skewed to the background class. Consequently, we combined the two Massachusetts datasets Buildings and Roads by adding the appropriate preprocessing tasks. By applying the model on the new multi-class problem (*Building, Road, Background* classes) we confirmed our hypothesis, in fact thanks to the new information added in the images, SegNet is able to construct better receptive fields and detect appropriately the images, as shown in Figure 2.

For U-Net and Dilated Convolution models we tested both losses, but no meaningful increment in the metrics were obtained by applying the Focal Tversky loss with $\alpha = 0.7$ and $\gamma = 0.75$, however we see that the neural network is able to learn and delineate better small objects, as shown in Figure 3, this result was also expected as described by Shruti Jadon's paper[4]. In addition, from Figure 4 we can see the outcome of the Dilated Convolution model, applied on the Massachusetts Buildings dataset, shows higher quality of the prediction compared to the other architectures.



(a) (b)

Figure 3: A sample of Massachusetts Buildings dataset, predicted image segmentation using U-Net. On 3a the predicted mask is obtain using BE Jaccard Loss, while on 3b it is obtained using the Dice-Tversky Loss function. The latter predicted mask highlights the property of the Dice-Tversky predict better hard example made by small objects, which are well-shaped and more delineated than the other.

Cityscapes. In this dataset the Imagenet weights initialization for both U-Net and SegNet was crucial in order to achieve higher accuracy results and to make the learning faster. For the SegNet model we found Adam optimizer and a learning rate of 10^{-3} as best suitable hyperparameters, frozen decoder weights were also used to reduce overfitting issues; from the proposed network we modified the model by adding Jaccard Loss in the Cross Entropy Loss function. SegNet did not have any problems during learning, since the Cityscapes dataset provides more informative images. For the U-Net model, again Adam optimizer with $lr = 10^{-3}$ was

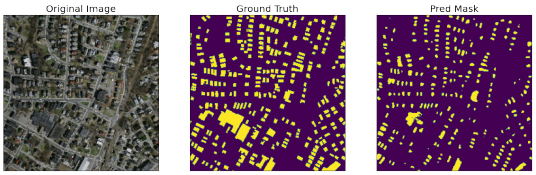


Figure 4: A sample of Dilated Convolution prediction on the Massachusetts dataset, with ground truth image segmentation in the center and predicted image segmentation on the right.

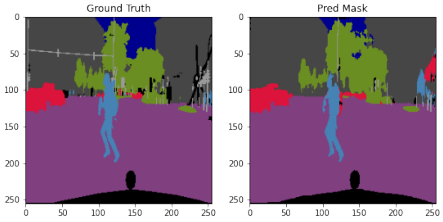


Figure 5: A sample of Cityscapes dataset, prediction obtained using Dilated Convolution model. Ground truth image segmentation on the left and predicted image segmentation on the right. The predicted mask is obtained using SGD (lr=0.1) on a batch-size of 4, in 60 epochs.

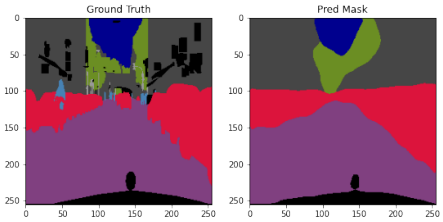


Figure 6: A sample of Cityscapes dataset, prediction obtained using SegNet model. Ground truth image segmentation on the left and predicted image segmentation on the right. The predicted mask is obtained using SGD (lr=0.1) on a batch-size of 12, in 20 epochs.

suiting better during the training, we also defined a higher batch-size than the SGD method, by fixing it at 32, since it is more helpful for exploiting the second-order moments estimation. Instead, on the Dilated CNNet we used a SGD optimizer with 10^{-1} as learning rate and set a small batch-size of 4 because of lacking resources.
After training all the models on those hyperparameters we compared them and we can see a summary of the results in Table 2, which confirms that all the models are able to achieve promising results. Examples of predicted test images can be seen in Figures 5 and 6. Comparing with other results from Cityscapes benchmark [2] we can see that the State-of-the-art architecture, the *HIK-CCSLT* model, reaches a mIoU of 93.3% on the category classes, instead *Segnet basic* architecture yields a score of 79.1%. Our best model is able to reach a score of 62.33% but we have to take into account that in our experiments the whole dataset has been reduced and machine resources were limited, hence the overall performances from all the model are reliable and optimal for the chosen architectures.

6 Conclusions

We presented U-Net, SegNet and Dilated Convolution architectures for semantic image segmentation on two datasets with different characteristics. The main properties of each architecture makes it more suitable for a certain type of object category task. Experimental results demonstrate the effectiveness of Dilated Convolution on image segmentation problems, even with imbalanced dataset and limited resources on disposal, but this model is very demanding in terms of computational time and memory size, thus training and prediction time are longer than other models. Instead SegNet, due to its simpler architecture and up sampling indexes technique, is the fastest in term of prediction time, but it fails when dealing with small objects and imbalanced dataset. On the other hand, U-Net was able to achieve remarkable results in both datasets with less training time. We have to highlight also that better performance and analysis could be lead in our experiments by having more computational power and memory. Further analysis can be made by exploring related works like Context Encoding as done in Zhang's paper [13] when many classes are involved or other NNets which combines dilated and pooling layers like in Lin's work [7].

[2] https://www.cityscapes-dataset.com/benchmarks/#pixel-level-results

References

[1] B. Bischke, P. Helber, J. Folz, D. Borth, and A. Dengel. Multi-task learning for segmentation of building footprints with seep neural network. *arXiv:1709.05932v1*, 2017.

[2] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[3] Ryuhei Hamaguchi, Aito Fujita, Keisuke Nemoto, Tomoyuki Imaizumi, and Shuhei Hikosaka. Effective use of dilated convolutions for segmenting small object instances in remote sensing imagery. *CoRR*, abs/1709.00179, 2017.

[4] Shruti Jadon. A survey of loss functions for semantic segmentation, 06 2020.

[5] A. Khalel and M. El-Saban. Automatic pixelwise object labeling for aerial imagery using stacked u-nets. *arXiv:1803.04953v1*, 2018.

[6] Yuhong Li, Xiaofan Zhang, and Deming Chen. Csrnet: Dilated convolutional neural networks for understanding the highly congested scenes. pages 1091–1100, 06 2018.

[7] Yeneng Lin, Dongyun Xu, Nan Wang, Zhou Shi, and Qiuxiao Chen. Road extraction from very-high-resolution remote sensing images via a nested se-deeplab model. *Remote Sensing*, 12:2985, 09 2020.

[8] V. Mnih. *Machine Learning for Aerial Image Labeling*. PhD thesis, University of Toronto, 2013.

[9] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, 2015.

[10] R. Cipolla Senior Member IEEE V. Badrinarayanan, A. Kendall. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *arXiv:1511.00561v3*, 2016.

[11] Yanjie Wang, Shiyu Hu, Guodong Wang, Chenglizhao Chen, and Zhenkuan Pan. Multi-scale dilated convolution of convolutional neural network for crowd counting. *Multimedia Tools and Applications*, 79, 01 2020.

[12] F. Yu and V. Koltun. Multi-scale context aggregation by dilated convolutions. *arXiv:1511.07122v3*, 2016.

[13] Hang Zhang, Kristin Dana, Jianping Shi, Zhongyue Zhang, Xiaogang Wang, Ambrish Tyagi, and Amit Agrawal. Context encoding for semantic segmentation. 03 2018.