NU-Net: Deep Residual Wide Field of View Convolutional Neural Network for Semantic Segmentation

Mohamed Samy[1] Karim Amer[1] Kareem Eissa Mahmoud Shaker Mohamed ElHelw
Center for Informatics Science
Nile University
Giza, Egypt
{k.amer, k.eissa, m.serag, melhelw}@nu.edu.eg

Abstract

Semantic Segmentation of satellite images is one of the most challenging problems in computer vision as it requires a model capable of capturing both local and global information at each pixel. Current state of the art methods are based on Fully Convolutional Neural Networks (FCNN) with mostly two main components: an encoder which is a pretrained classification model that gradually reduces the input spatial size and a decoder that transforms the encoder's feature map into a predicted mask with the original size. We change this conventional architecture to a model that makes use of full resolution information. NU-Net is a deep FCNN that is able to capture wide field of view global information around each pixel while maintaining localized full resolution information throughout the model. We evaluate our model on the Land Cover Classification and Road Extraction tracks in the DeepGlobe competition.

1. Introduction

Semantic segmentation of satellite imagery is used to extract road networks, detect buildings for urban planning and recognize green areas for promoting sustainable clean environments. This area of research has thus received substantial attention in last few years with significant progress made due to two main reasons. Firstly, the wide adoption of deep learning techniques that combined with computer vision have been successfully used in tasks such as image classification [1][2], object detection [3][4] and semantic segmentation [5]. Secondly, the accessibility to open source datasets with high resolution satellite images.

Most current work in semantic segmentation features a similar neural network architecture: an encoder network that captures high level features in input images while reducing their spatial size using max pooling, and a decoder which restores the original image size and outputs the segmentation mask using upsampling or convolution-transpose.

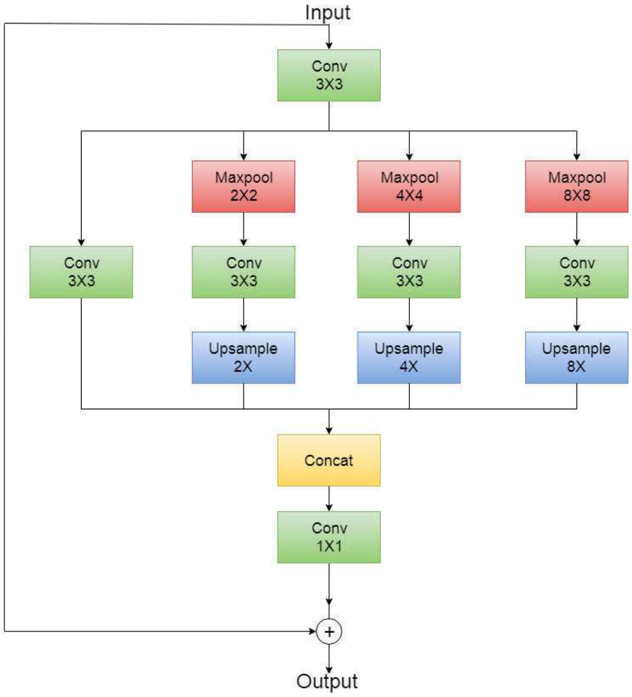

Figure 1: Residual Wide FoV Module Structure

For instance, Long et. al. [6] developed a fully convolutional architecture which uses VGG16 [2] as an encoder and applies linear upsampling on the feature maps in order to predict segmentation mask. Noh et. al. [7] enhanced this idea by using a mirrored VGG as decoder to increase the nonlinearity of output layers. Ronneberger et. al. [8] added skip connections between encoder and decoder layers which resulted in the UNet architecture. Other work made use of advances in classification architecture and added residual [9] and dense connections [10] to increase the performance of their models.

The above techniques are inspired by architectures designed mainly for general classification tasks and are hence able to successfully capture global information in an image. However, such design limits their performance in segmentation tasks since searching for information

[1] These authors have contributed equally. Mohamed Samy passed away on 20 May 2018.

globally in the image is different objective from segmentation that needs pixel specific information.

Developing an architecture that utilizes the full resolution information of input images is a logical alternative that can preserve local information, *i.e.* pixel-level information, instead of losing it in the encoder network. Pohlen et. al. [11] proposed a new architecture that have two streams: one with full resolution features and one with encoder-decoder layers but with no pretrained weights. Both streams exchange information along the forward pass. The results were better than many models that made use of pretrained models. However, their work revealed disadvantages in full resolution architectures. For instance, significant memory requirements limit network depth, and series of convolutional layers cannot faithfully capture global information because of their limited field of view. Consequently, the authors added another encoder-decoder stream.

In this work extend the above architecture by introducing NU-Net, a novel deep neural network architecture which utilizes full resolution as well as global information for improved segmentation. NU-Net features multiple Wide Field of View (FoV) modules and is inspired by PSP-Net [12]. We argue that our architecture achieves better results without applying any pre-processing or post processing techniques. We describe our model in detail in section 2, report current results on the Land Cover Classification and Road Extraction datasets in section 3, and conclude in section 4.

2. Proposed Method

2.1. Residual Wide FoV Module

This module aims at exploiting the full resolution information while gaining a receptive field that is large enough to capture the global information without losing local features per pixel. To design a block that achieves this goal, we had to keep in mind that convolutional layers are good at capturing local information for each pixel whereas max pooling helps exploiting the global information.

The residual wide FoV module is thus designed as shown in Figure 1. Firstly, the input passes through a 3x3 convolution layer. Then, there are multiple branches, each one has non-overlapping max pooling followed by 3x3 convolution to capture spatial information at different resolutions. At the end of each branch, there is a bilinear upsampling layer to reverse the effect of pooling and restore original input resolution. To combine all branches and compress their information in smaller number of channels, their outputs are concatenated and a 1x1 convolution layer is applied. Local information is retained by using a residual connection between the input and output feature maps of the module.

2.2. NU-Net

NU-Net architecture utilizes the residual wide FoV modules with design similar to ResNet [13]: a series of blocks with skip connections between their inputs and outputs. The architecture has *k* residual wide FoV modules preceded by a 3x3 convolution to the input image to produce a feature map with *C* channels that is followed by a 1x1 convolution layer to predict the segmentation mask. Every convolution layer in NU-Net is combined with batch normalization and ReLU activation except for the last layer in which a sigmoid function is used.

Unlike other encoder-decoder architectures such as U-Net [8] which decrease the spatial size of input feature maps multiple times then gradually increase them, NU-Net doesn't impose traditional hierarchical learning. The network wide FoV modules each applies down sample then up sample operations which helps the architecture to detect fine details robustly without losing global information. NU-Net network architecture utilizes a consecutive series of modules with each module operating on the output of the previous one to correct its mistakes and improve overall results. Furthermore, while NU-Net design is inspired by Pyramid Scene Parsing (PSP) [12], key differences exist. Unlike PSP, NU-Net uses more convolutional layers to increase non-linearity and exploit the feature maps between max pooling and upsampling by widening the receptive field. To this end, each position (pixel) in the feature map holds information of 4 pixels after max-pooling with 2x2 and before upsampling. By adding a 3x3 convolution layer before upsampling, each position will hold information of 4x9 positions. Such widening of the field of view increases horizontally with larger max-pooling and vertically with more modules. Furthermore, while PSP is applied once on feature maps of a pretrained network, the proposed wide FoV modules are applied multiple times without the need for a pretrained model.

One of the main drawbacks of NU-Net is that it retains full resolution throughout the network which is computationally demanding. In order to alleviate this issue for deep versions of NU-Net, a 2x2 max-pooling layer can be inserted before the first wide FoV module and an upsampling layer after the last one. This modification reduces the needed computations and memory by a quarter without affecting prediction accuracy. We will call this the energy saving network mode.

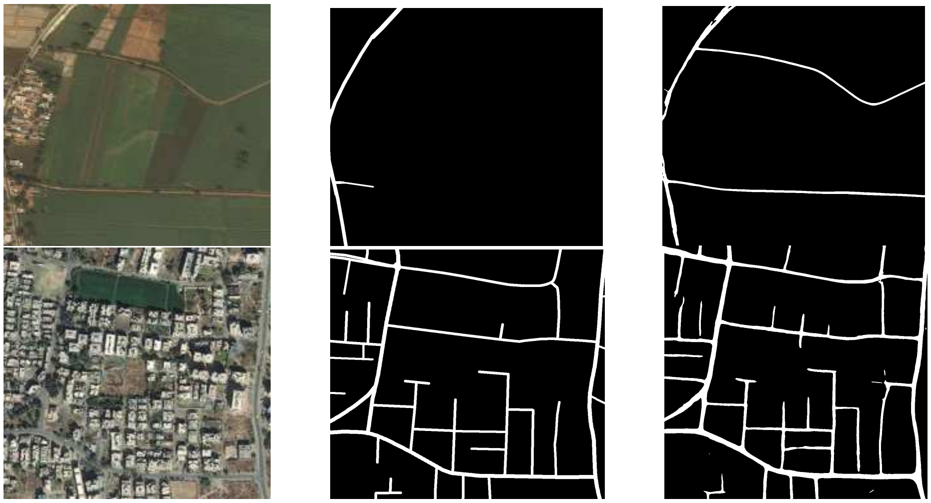

Figure 2: Examples for NU-Net Road Extraction results versus true masks. First column has the original images, second column has the true masks and third column has NU-Net predictions.

Num. of residual wide FoV modules (k)	Local Validation	First Stage Leaderboard	Second Stage Leaderboard
5	55%	52.6%	-
9	62.2%	57.1%	57.4%

Table 1: NU-Net results on validation and leaderboard with different number of wide FoV modules. The results calculated using Jaccard coefficient on road extraction dataset.

3. Experiments

3.1. Road Extraction Track

We evaluate our method on the Road Extraction track in the Deep Globe [14] competition. The dataset consists of more than 6000 images with size 1024x1024 with 50 cm resolution per pixel. The masks are mixture of street roads and small trails taken in different environments such as urban, rural and desert areas. Such variability makes the problem more challenging.

Through the competition, the evaluation metric used was Jaccard coefficient which is

$$Jaccard\ Coefficient = \frac{TP}{TP + FP + FN}$$

where TP is number of true positive pixels, FP is the number of false positive pixels and FN is the number of false negative pixels in a single image. The metric is applied to each image separately and the average results is calculated.

In our experiments, the dataset was divided into training and validation with 75% and 25% of images respectively. The validation set was used to choose the best epoch for leaderboard submission. We applied NU-Net with different number of residual wide FoV modules (k) to investigate the effect of network depth on overall performance. The number of filters used in all layers is 64 and the models were trained with soft dice loss defined as:

$$Soft\ Dice\ Loss = 1 - \frac{2 * \sum_i p_i * y_i}{\sum_i p_i + \sum_i y_i}$$

where y_i and p_i are the ground truth and predicted probability, respectively, for pixel i. Data augmentation was applied using flipping, rotation and random erosions.

Table 1 shows the results with k = 5 and 9 on both validation and leaderboard sets. For k = 9, we used the network energy saving mode. It is observed that the local validation score increased from 55% to 62.1% while the score on the first stage leaderboard increased from 52.6% to 57.1% after increasing the model depth. Our best two submissions on the final leaderboard was NU-Net with k = 9 which scored 57.4% and an ensemble of two versions of the same model which scored 57.8%.

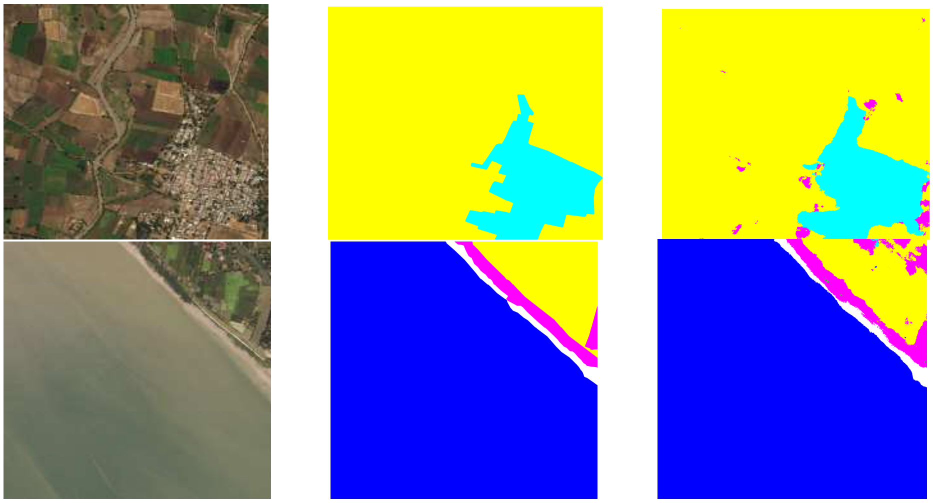

Figure 3: Examples for NU-Net Land Cover results versus true masks. First column has the original images, second column has the true masks and third column has NU-Net predictions.

Figure 2 shows some prediction examples of NU-Net versus the true masks in the local validation set. We can see clearly that our model is able to capture roads that were not labeled in the ground truth mask.

3.2. Land Cover Classification Track

For the Land Cover Classification track, the dataset consisted of 803 images of size 2448x2448. Each image pixel belongs to one of seven classes: Urban, Agriculture, Rangeland, Forest, Water, Barren land, and Unknown. The Unknown class is not used in the evaluation. The evaluation metric is the Mean Intersection over Union (MIoU) where the IoU for each class is calculated separately for all images and the mean IoU for the six classes is reported as the final metric.

The data is divided into 75% training and 25% validation similar to the Road Extraction track. We used the same NU-Net architecture with $k = 9$ and replaced the final layer to output seven classes. The pretrained weights of the Road Extraction track are used to help the network converge faster. The model was trained with weighted cross entropy loss defined as:

$$Wieghted\ Cross\ Entropy = -\sum_{i}\sum_{c} w_c * y_{i,c} * \log(p_{i,c})$$

where w_c is the weight for class c and $y_{i,c}$ and $p_{i,c}$ are the true label and the predicted probability for class c at pixel i respectively. The weight of each class is the inverse of class percentage in the training batch which reduces the effect of class imbalance.

The best achieved MIoU on the local validation set is 70% and on the second leaderboard the score is 38.4%. We investigated the gap between local validation score and leaderboard score and we found multiple reasons for this discrepancy. Images in the test set have sharper colors than those in the training set. In addition, the test set are captured on a lower altitude which makes shadows more visible compared to the training set. In order to lower the effect of these differences, we used Adaptive Batch Normalization [15] which increased our leaderboard score to 42.8% without re-training the model. Figure 3 shows some examples of NU-Net predictions on local validation set.

4. Discussion

In this work, we introduced NU-Net, a novel convolutional neural network architecture for semantic segmentation of satellite imagery. Our model utilizes large receptive fields to extract global information while preserving the spatial size to make use of local information for enhanced segmentation. As a future work, extensive study is needed for NU-Net parameters and its effect on the overall performance.

Acknowledgements

We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan Xp GPU used for this research.

In memory of Mohamed Samy Ahmed Kassem
(September 1993 - May 2018).

References

[1] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNet Classification with Deep Convolutional Neural Networks,” *Adv. Neural Inf. Process. Syst.*, pp. 1–9, 2012.

[2] K. Simonyan and A. Zisserman, “Very Deep Convolutional Networks for Large-Scale Image Recognition,” pp. 1–14, 2014.

[3] R. Girshick, “Fast R-CNN,” *Proc. IEEE Int. Conf. Comput. Vis.*, vol. 11–18–Dece, pp. 1440–1448, 2016.

[4] D. Impiombato *et al.*, “You Only Look Once: Unified, Real-Time Object Detection Joseph,” *Nucl. Instruments Methods Phys. Res. Sect. A Accel. Spectrometers, Detect. Assoc. Equip.*, vol. 794, pp. 185–192, 2015.

[5] V. Badrinarayanan, A. Kendall, and R. Cipolla, “SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 12, pp. 2481–2495, 2017.

[6] J. Long, E. Shelhamer, and T. Darrell, “Fully Convolutional Networks for Semantic Segmentation ppt,” *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, vol. 8828, no. c, pp. 3431–3440, 2015.

[7] H. Noh, S. Hong, and B. Han, “Learning deconvolution network for semantic segmentation,” in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, vol. 2015 Inter, pp. 1520–1528.

[8] O. Ronneberger, P. Fischer, and T. Brox, “U-Net: Convolutional Networks for Biomedical Image Segmentation,” *Miccai*, pp. 234–241, 2015.

[9] G. Lin, A. Milan, C. Shen, and I. Reid, “RefineNet: Multi-Path Refinement Networks for High-Resolution Semantic Segmentation,” *Cvpr2017*, pp. 1925–1934, 2017.

[10] S. Jegou, M. Drozdzal, D. Vazquez, A. Romero, and Y. Bengio, “The One Hundred Layers Tiramisu: Fully Convolutional DenseNets for Semantic Segmentation,” in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, 2017, vol. 2017–July, pp. 1175–1183.

[11] T. Pohlen, A. Hermans, M. Mathias, and B. Leibe, “Full-resolution residual networks for semantic segmentation in street scenes,” *Proc. - 30th IEEE Conf. Comput. Vis. Pattern Recognition, CVPR 2017*, vol. 2017–Janua, pp. 3309–3318, 2017.

[12] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, “Pyramid Scene Parsing Network,” 2016.

[13] K. He, X. Zhang, S. Ren, and J. Sun, “Deep Residual Learning for Image Recognition,” in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[14] I. Demir *et al.*, “DeepGlobe 2018: A Challenge to Parse the Earth through Satellite Images,” 2018.

[15] Y. Li, N. Wang, J. Shi, X. Hou, and J. Liu, “Adaptive Batch Normalization for practical domain adaptation,” *Pattern Recognit.*, vol. 80, pp. 109–117, 2018.