Stacked U-Nets for Ground Material Segmentation in Remote Sensing Imagery

Arthita Ghosh, Max Ehrlich, Sohil Shah, Larry Davis, and Rama Chellappa {arthita, sohilas, lsdavis}@umd.edu, {maxehr,rama}@umiacs.umd.edu University of Maryland, College Park, MD, USA.

Abstract

We present a semantic segmentation algorithm for RGB remote sensing images. Our method is based on the Dilated Stacked U-Nets architecture. This state-of-the-art method has been shown to have good performance in other applications. We perform additional post-processing by blending image tiles and degridding the result. Our method gives competitive results on the DeepGlobe dataset.

1. Introduction

The goal of the DeepGlobe challenge [8] is to produce a per-pixel map of the ground material given an RGB input captured from a satellite sensor. This is particularly challenging even in the context of remote sensing since most satellite datasets contain more than these three canonical spectral bands. Restricting the input data in this way requires an architecture that can handle large intra-class variation even with less information.

We address the problem of ground material classification using a semantic segmentation algorithm. Semantic segmentation has a rich history in computer vision [12, 19, 22, 18, 27] with most recent techniques focusing on the use of convolutional neural networks [15]. We choose state of the art Dilated Stacked U-Net [24], due to its good performance on other datasets [7, 9] with relatively fewer parameters and therefore easy trainability.

We propose to combine the Dilated Stacked U-Net architecture with a set of post-processing techniques designed to overcome the traditional difficulty of working with remote sensing imagery. Our algorithm combines the blending described in [1] with a novel frequency-domain method for removing grid artifacts that often plague the outputs of dilated network architectures.

2. Related Work

The U-net architecture was originally proposed to perform segmentation on bio-medical images [23]. U-nets capture context information at multiple scales via contracting (encoder) and expansive (decoder) paths and can be trained with relatively smaller amounts of data. Other



Figure 1: UNet module with outer residual connection

recent encoder-decoder structure-based deep architectures proposed for segmentation include [17, 2]. Chen *et al.* and Lin *et al.* [6, 18] engage image pyramid pooling to capture information at multiple scales whereas Zhao *et al.* [29] and Chen *et al.* [5] use spatial pyramid pooling and atrous convolutions to this end. Several other approaches achieve context aggregation via Conditional Random Fields (CRFs) on deep features [16, 3, 4]. Yu *et al.* [28] utilize dilated convolutions for context aggregration. Stacked deconvolution layers are used in [22, 10, 13] whereas Ghiasi *et al.* [11] use Laplacian pyramids to selectively refine the low resolution maps. Other architectures for segmentation that were built atop VGG [25] include [19, 30, 23, 28]. Other fully convolutional architectures that have been applied to semantic labeling of remote sensing data include [20, 14, 21, 26]

3. Method

The land cover classification task on satellite imagery is an instance of the semantic scene segmentation problem. We train a deep architecture composed of stacked U-Nets to perform this task. The network is trained end-to-end.

3.1. U-Net Module

Each U-Net module consists of 10 convolutional blocks, each preceded by ReLU and batch normalization. Lower and higher resolution feature maps are generated using strided convolution and deconvolution respectively. The resolution of feature maps at the input and output of each



Figure 2: Dilated Stacked U-Nets architecture for semantic scene segmentation

U-Net module is same. The convolutions use 3x3 kernels. Bottleneck layers with 1x1 convolution kernels are added at the input and output of each U-Net module. Skip connections within the module help avoid vanishing gradients. Figure 1 illustrates the structure.

3.2. Dilated Stacked U-Nets

Figure 2 outlines the architecture used to perform scene segmentation on the Deepglobe data. It consists of 4 blocks of stacked U-Nets containing 2,7,7 and 1 module(s) respectively. Every module is preceded and succeeded by 1x1 convolutions for feature transformation. Input images pass through 7x7 convolution filters (stride=2) and a residual block. Subsequently, information passes through 4 blocks of stacked U-Net modules which combine details captured at high-resolution with long distance context information captured at low-resolution to generate segmentation maps for the scene. For an input size of 512x512, the output map size is 32x32, This is owing to 2 strided convolutions and 2 average pooling operations, which increase the field of view and help capture long-distance information. Average pooling is applied on outputs of first and second stacked U-Net blocks. In the third and fourth blocks, U-Net modules perform dilated convolutions to keep feature map resolution constant. The last U-Net module in block 4 is a trimmed version of standard U-Nets, with only encoder E1 and decoder D1 which are illustrated in Figure 1. In the first U-Nets of each block, the skip connection is implemented using a a 1×1 convolution. The residual connection in all but the first U-Net in each block is implemented as an identity mapping. The number of output feature maps from each blocks is roughly the same as the total number of feature maps generated by all the preceding U-Net modules, which allows the architecture the flexibility to retain all of them. The output is re-scaled to the original size using bilinear interpolation. The total number of parameters learned is 37.7 million. Multi-class cross entropy loss function is used to train the network along with weights in proportion to the rarity of class samples in the training set.

4. Experiments

Here we briefly describe our experimental setup. We include the specifics of our data augmentation and training scheme as well as our post-processing scheme.

4.1. Data Augmentation

The training set of Deepglobe contains 803 images. We set aside 50 from these for validation. We will regard these two sets as 'Train' and 'Train-Val'. Besides this, there is a validation set of 171 images (Val) on which we report mean-IoU score. Each image is of size 2448x2448. From these images we crop tiles of size 512x512 (from all over the image) as inputs to the network. Experiments with smaller tiles (256x256) lead to reduced performance due to lesser long distance information whereas larger input tiles (1024x1024) constrain batch size (due to limitation of GPU memory) and interfere with learning of batch-normalization parameters. For training, we augment inputs by randomly flipping, scaling, jittering and rotating the tiles.

4.2. Training

We use the Adam optimizer with a starting learning rate of $2.0e^{-4}$. Weight decay and momentum values are set at $1.5e^{-4}$ and 0.95 respectively. We get rid of the pooling steps at the end of third and fourth blocks to operate at an output stride of 16. Output stride is the ratio of input to output resolution. The training and testing pipelines are implemented using the PyTorch framework and the model was trained on P6000 GPUs with a batch size of 20.

4.3. Post-processing

Remote sensing data generally consists of large images containing many small structures. The Stacked U-Net ar-



(a) U-Net Output



(b) After deblocking Figure 3: Post-Processing Result



(c) After deblocking and degridding

chitecture we used requires a significant amount of GPU memory. During training, we break the images into smaller tiles. When the original image is reconstructed, this results in blocking artifacts at the borders of the tiles because of the lack of shared context. Additionally, the Stacked U-Net itself is prone to gridding artifacts. Both these types of artifacts can be seen in Figure 3a. We propose a two stage post-processing algorithm to deblock and degrid the network's output. This results in an additional 3% improvement in mean IOU score vs using the network output directly on the DeepGlobe validation data. The first stage of our method is intended to smoothly combine adjacent tiles to avoid hard boundary lines, it is based on the method in [1]. For each tile, we take four 90-degree rotations and their reflections for a total of eight duplicate tiles per single test tile. The raw predictions on these eight tiles are averaged to produce a single prediction tile. The tiles themselves are sampled with 50% overlap and then merged using a centered 2D Gaussian window. The result of this algorithm is shown in Figure 3b where block artifacts are effectively removed. Though the result of stage one is much cleaner than the raw network output, it is still subject to grid artifacts. The next stage of our post-processing algorithm addresses these. We observe that the grid artifacts are often small, high-frequency noise. To remove them, we take the discrete cosine transform of 8×8 blocks of the output of stage 1. We then remove all but the DC coefficient of the transform and project the result back into the spatial domain. This effectively replaces each pixel in the 8×8 blocks with the average label of the block. Then, to remove lower frequency artifacts, we use a voting scheme. For each block, the label of the block is replaced by the majority label of its eight neighbors. This gives the final result in 3c which is free from most artifacts.

5. Results

Model	No-PP	DB	DB+DG
SUNET-7128-723	0.48446	0.50702	0.50703
Table 1. Mean IoU scores of our model on the validation set			

Table 1: Mean IoU scores of our model on the validation set.
No-PP = no post processing. DB = deblocking. DB+DG =
deblocking followed by degridding

Table 1 presents the mean IoU scores obtained on the validation set for the proposed Stacked U-Net model with 4 blocks of depth 2,7,7 and 1 respectively. The input image was padded and tiled to perform inference. Additional deblocking and degridding steps (discussed in Section 4.3 yielded enhanced performance by removing the gridding effect.

Automatic land cover classification is still an exploratory problem as the ground truth has missing annotations. Some of the small details that were captured by our model are presented in Figure 4. Other cases where the model was frequently confused are shadows in forest area, that appeared darker and smoother which were classified as water. In some cases, ripples on large flat water bodies, that appear dark green, created some texture and were predicted as forest area by the model. These are specific cases which can be handled better with more training examples from such images.

6. Conclusion

We apply Dilated Stacked U-Nets to produce state-ofthe-art results in semantic segmentation on Deepglobe data. Our method is combined with an effective post-processing algorithm designed to address both the specific challenges of remote sensing data and the U-Net output. Our method scores competitively on the DeepGlobe data.



Figure 4: Some missing detail in the ground truth was captured by Stacked U-Nets model

7. Acknowledgments

The research is based upon work supported by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via DOI/IBC Contract Number D17PC00287. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the ODNI, IARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon.

References

- Dstl satellite imagery competition, 1st place winner's interview: Kyle lee. http://blog.kaggle.com/ 2017/04/26/dstl-satellite-imagerycompetition-1st-place-winnersinterview-kyle-lee/. Accessed: 2018-04-28.
- [2] V. Badrinarayanan, A. Kendall, and R. Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *TPAMI*, 39(12):2481–2495, 2017.
- [3] S. Chandra and I. Kokkinos. Fast, exact and multi-scale inference for semantic image segmentation with deep gaussian crfs. In *ECCV*, pages 402–418. Springer, 2016.
- [4] S. Chandra, N. Usunier, and I. Kokkinos. Dense and lowrank gaussian crfs using deep embeddings. In *ICCV*, 2017.
- [5] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *TPAMI*, 40(4):834–848, 2018.
- [6] L.-C. Chen, Y. Yang, J. Wang, W. Xu, and A. L. Yuille. Attention to scale: Scale-aware semantic image segmentation. In *CVPR*, pages 3640–3649, 2016.
- [7] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele. The cityscapes dataset for semantic urban scene understanding. In *CVPR*, pages 3213–3223, 2016.
- [8] I. Demir, K. Koperski, D. Lindenbaum, G. Pang, J. Huang, S. Basu, F. Hughes, D. Tuia, and R. Raskar. Deepglobe 2018: A challenge to parse the earth through satellite images. *Arxiv e-prints 2018, arXiv: 1805.06561*, 2018.
- [9] M. Everingham, S. A. Eslami, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman. The pascal visual object classes challenge: A retrospective. *IJCV*, 111(1):98–136, 2015.
- [10] J. Fu, J. Liu, Y. Wang, and H. Lu. Stacked deconvolutional network for semantic segmentation. arXiv preprint arXiv:1708.04943, 2017.
- [11] G. Ghiasi and C. C. Fowlkes. Laplacian pyramid reconstruction and refinement for semantic segmentation. In *ECCV*, pages 519–534. Springer, 2016.
- [12] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *CVPR*, pages 580–587, 2014.
- [13] S. Jégou, M. Drozdzal, D. Vazquez, A. Romero, and Y. Bengio. The one hundred layers tiramisu: Fully convolutional densenets for semantic segmentation. In *CVPR Workshop*, pages 1175–1183. IEEE, 2017.
- [14] M. Kampffmeyer, A.-B. Salberg, and R. Jenssen. Semantic segmentation of small objects and modeling of uncertainty in urban remote sensing images using deep convolutional neural networks. In *CVPR Workshop*, pages 680–688. IEEE, 2016.
- [15] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, pages 1097–1105, 2012.
- [16] C. Liang-Chieh, G. Papandreou, I. Kokkinos, K. Murphy, and A. Yuille. Semantic image segmentation with deep convolutional nets and fully connected crfs. In *ICLR*, 2015.

- [17] G. Lin, A. Milan, C. Shen, and I. Reid. Refinenet: Multi-path refinement networks for high-resolution semantic segmentation. In *CVPR*, 2017.
- [18] G. Lin, C. Shen, A. Van Den Hengel, and I. Reid. Efficient piecewise training of deep structured models for semantic segmentation. In *CVPR*, pages 3194–3203, 2016.
- [19] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, pages 3431– 3440, 2015.
- [20] E. Maggiori, Y. Tarabalka, G. Charpiat, and P. Alliez. Convolutional neural networks for large-scale remote-sensing image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 55(2):645–657, 2017.
- [21] D. Marmanis, K. Schindler, J. D. Wegner, S. Galliani, M. Datcu, and U. Stilla. Classification with an edge: improving semantic image segmentation with boundary detection. *ISPRS Journal of Photogrammetry and Remote Sensing*, 135:158–172, 2018.
- [22] H. Noh, S. Hong, and B. Han. Learning deconvolution network for semantic segmentation. In *ICCV*, pages 1520–1528, 2015.
- [23] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *MIC-CAI*, pages 234–241. Springer, 2015.
- [24] S. A. Shah, P. Ghosh, L. S. Davis, and T. Goldstein. Stacked u-nets:a no-frills approach to natural image segmentation. *arXiv*:1804.10343, 2018.
- [25] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556, 2014.
- [26] M. Volpi and D. Tuia. Dense semantic labeling of subdecimeter resolution images with convolutional neural networks. *IEEE Transactions on Geoscience and Remote Sensing*, 55(2):881–893, 2017.
- [27] J. Yao, S. Fidler, and R. Urtasun. Describing the scene as a whole: Joint object detection, scene classification and semantic segmentation. In CVPR, pages 702–709. IEEE, 2012.
- [28] F. Yu and V. Koltun. Multi-scale context aggregation by dilated convolutions. arXiv preprint arXiv:1511.07122, 2015.
- [29] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia. Pyramid scene parsing network. In *CVPR*, pages 2881–2890, 2017.
- [30] S. Zheng, S. Jayasumana, B. Romera-Paredes, V. Vineet, Z. Su, D. Du, C. Huang, and P. H. Torr. Conditional random fields as recurrent neural networks. In *ICCV*, pages 1529– 1537, 2015.