Semantic Segmentation of High-Resolution Satellite images: a Deep Learning Approach

Madhumita, D[1]., and Bharath, H.A[2]*
[1]Ranbir and Chitra Gupta School of Infrastructure Design and Management,
[2]Indian Institute of Technology Kharagpur, West Bengal – 721302, India
*Email: bharath@infra.iitkgp.ac.in

(Received: Aug 22, 2021; in final form: Dec 17, 2021)

ABSTRACT: High-dimensional geospatial data visualization has gained much importance in recent decades. But to analyze it, traditional technologies used in machine learning are not convincing enough, and thus to switch to a sub-domain of machine learning called deep learning that has gained popularity because of its accuracy and high dimensional data analysis power. Its convergence with geospatial data analytics shall prove to be a boon to the researchers working in the domain of geospatial data. Though Geospatial information is mostly used in the global mapping process of satellite images. The heterogeneity of the data makes it infeasible for global scale mapping. Therefore, to handle this problem is to partition the entire world into several regions. Semantic segmentation is one such technique and is widely used for information extraction from satellite images. The technique essentially refers to segmenting the input image pixel into multiple semantic regions, that is, to assign a semantic pixel category to each pixel in the image. In this context, we propose a semantic segmentation method that utilizes the spatial information of the high-resolution remote sensing data. The aim is to leverage the openly available data to automatically generate a larger training dataset with more variability and can be used to build more accurate deep learning models. The proposed automatic extraction can capture context information and its symmetric expanding path enables precise localization. The most characteristic property is the up-sampling part that has feature channels that allow propagation of context information to higher resolution layers and makes the expansive path roughly symmetric to the contracting path yielding a U-shaped architecture. Mean IOU (mIOU) is used as the performance matrix and results yield 0.79. Since the model is trained on a small training dataset, that makes the deep learning model prone to overfitting. Training on such a small set of images makes this a challenging task. Validation dataset metrics obtained after training will signify the model's general adaptability on other datasets of other segmentation tasks.

Keywords: Deep Learning, Semantic Segmentation, Feature Extraction, Remote Sensing

1. Introduction

Geospatial remote sensing data plays a key role in various scientific disciplines as it seeks to understand, analyze, and visualize real-world phenomena according to their locations (Bharath et al., 2018a). It is believed that almost 80% of all data is geographic in nature because the majority of information surrounding us can be georeferenced (VoPham et al., 2018). The demand for the available geospatial data is consistently growing at an ever-faster pace, leading to the constant increase in demand for processing power and storage still emerging. However, the highly variable nature of the information demands human supervision to distinguish the interesting patterns (Vorona et al., 2019; Bharath et al., 2018b). Therefore, understanding geospatial remote sensing images in the semantic context is particularly important and its intelligent identification is definitely demanded.

Remote sensing image comprehension aims to automatically assign a specific semantic label to each pixel according to its contents and has become a vital research topic in the field of remote sensing image interpretation considering its different applications in urban planning, traffic control, land resource management, and disaster monitoring (Prakash et al., 2020; Zhang et al., 2019). Moreover, automatic feature extraction through machine learning is crucial in order to understand the ever-changing dynamics, including anthropogenic changes. The automated extraction of high-resolution remote sensing images is highly desirable but poses many difficulties due to the wide variety of volumes and unavailability of labeled annotations (Prakash et al., 2020; Özyurt., 2020). The traditional methods for manually digitizing were human-intensive and expensive (Ramachandra et al., 2012; Bharath et al., 2018a). They are limited to point observations. Therefore, impossible to scale it to large cities or geographical areas. Also, non-adaptable to build and maintain into the digital field. However, the convergence of deep learning and computer vision with remote sensing has enabled automated extraction to be highly efficient and cost-effective.

The recent development of deep learning technologies has played an increasingly important role in delivering computer vision and addressing problems such as pattern recognition and feature detection (Ramachandra et al., 2015). Unlike low-level and mid-level features, the models can learn more powerful, abstract, and discriminative features via deep architecture neural networks irrespective of engineering skill and domain expertise. Moreover, deep learning techniques have been widely implemented in remote sensing images, especially in feature extraction from satellite images with highly accurate and precise results. Having prerequisites such as highly improved satellite images in terms of spatial, spectral, and temporal resolutions and Geomatics communities, automated extraction is the current need. The Convolutional Neural Networks especially has demonstrated outstanding performance due to the availability of large-scale geospatial data and the advancement of computing power. Although they have achieved dramatically improved

© Indian Society of Geomatics

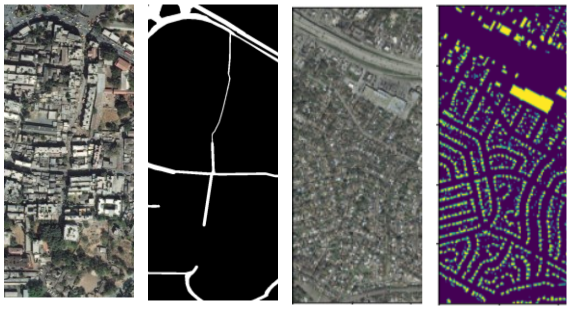
classification accuracy, they are still easily misclassified due to the complex characteristics and occlusions (Petrovska et al., 2020; Li et al., 2019).

Semantic segmentation is one such important task based on convolutional neural networks. The technique essentially refers to segmenting the input image pixel into multiple semantic regions, that is, to assign a semantic pixel category to each pixel in the image. It is widely used in computer vision applications such as remote sensing image interpretation, medical image processing, and many more (Tran et al., 2020). Semantic Segmentation of satellite images is one of the crucial problems as it requires a model that is capable of capturing both the local and global information at each pixel level (Gleason et al., 2010). To integrate these, the UNet neural network architecture is proposed with the aim to supplement a contracting network by successive layers, and pooling operators are replaced by up sampling operators. The fully convolutional network is capable of handling with very few training images and yields more precise segmentation outputs (Ronneberger et al., 2015). The study addresses the problem of automated extraction of road networks and building footprints from satellite imagery. Road network and building footprint extraction play a significant role in many applications that involve updating maps, traffic regulations, city planning, etc. This paper proposes a convolutional architecture for automated extraction so as to improve the robustness of semantic segmentation for satellite images leveraging open data source platforms.

2. Datasets

Two popular remote sensing datasets Deep Globe dataset and INRIA dataset with different spatial properties are chosen to better demonstrate the robustness and effectiveness of the proposed method. Both the datasets are essentially configured for pixel-wise segmentation. In addition, details about the datasets are described below:

Datasets for road network extraction: Deep Globe dataset was sampled from the Digital Globe and Vivid Images dataset with their road parts labeled to generate annotated maps. The dataset covers images captured over Thailand, Indonesia, and India. The images consist of 3 channels i.e., Red, Green, and Blue with a ground resolution of 50 cm/pixel and each of the original geotiff images are 19′584 × 19′584 pixels. In the annotated map each pixel is classified as either road or non-road. The dataset consists of 6226 and 1243 training and validation images, respectively. The complexity of the dataset is that it is highly imbalanced in terms of the number of pixels per class, i.e., roads are thin lines within the images and therefore occupy few pixels only as compared to the background pixels that means more 0 values (non-road pixels) compared to 1 value (road pixels) as shown in Figure 1(a).



(a) (b)

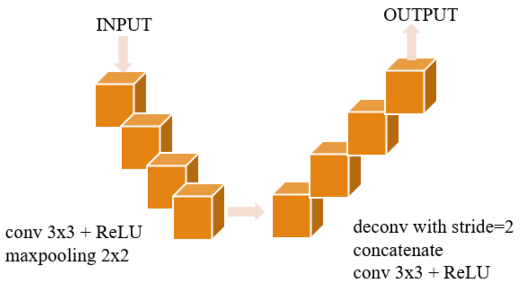
Figure 1. Examples of images and labels from the (a) Deep Globe dataset and (b) INRIA dataset include the original image and label, and the label has two classes, which are road and building

Dataset for building extraction: INRIA dataset as shown in figure 1(b) consists of 180 orthorectified aerial images in the RGB channel. Each pixel is of 0.3 meters resolution. The dataset is composed of two subsets namely, train and test covering 405 sq. km area. The training data is annotated for two classes: building and not building and covers regions Austin, Chicago, Kitsap County, Western Tyrol, and Vienna, whereas the test set covers a different set of regions: Bellingham, Bloomington, Innsbruck, San Francisco, Eastern Tyrol. The varying urban densities in covered regions along with variation in training and test images make the INRIA dataset complex and we can explore the capability of our proposed model.

3. Method and Data

3.1 Model Architecture

The architecture of our segmentation model was adapted from (Ronneberger et al., 2015), originally designed for biomedical image segmentation. The architecture as shown in Figure 2(b) consists of a contracting path and an expansive path wherein the contracting path follows the typical architecture of a convolutional network. The encoding and decoding part are composed of four blocks and each consisting 3x3 convolutions layers i.e., unpadded convolutions are applied repeatedly, followed by a rectified linear unit (ReLU), a 2x2 max pooling operation. The down sampling has deconvolutional layer with stride 2 and concatenation layer, two 3x3 convolutional layer followed by ReLU as shown in fig 2 (a). The number of feature channels gets doubled at each of the down sampling steps. The final layer consists of a single 1x1 convolution layer mapping each 64-component feature vector to the desired number of classes. The architecture has 23 total layers. The presence of a large number of feature channels in the up-sampling part that allows the network to propagate context information to the higher resolution layers makes the UNet architecture unique.



(a)

Figure 2(a). Encoder and decoder layers

3.2 Training Process

The architecture is built with the keras 2.8.0 and Tensor Flow in python 3.6+. Keras and Tensor Flow are open-source python libraries. The training dataset was created by dividing the images into patches of size 256 x 256 x 3 which had sufficient distribution of roads and building structures with the surrounding environment so as to be learned by the networks. The experiment was conducted for a total number of 100 epochs with a batch size of 16. It was trained with the mini-batch Stochastic gradient descent using the ADAM optimizer. Binary cross entropy loss function was used which essentially gives the cross-entropy loss between the predicted classes and the true classes.

3.3 Evaluation Metrics

The quantitative performance of the segmentation model was evaluated using 4 different evaluation metrics namely the ‘Precision’, ‘Recall’, ‘F1-score’, and mean of Intersection-over-Union (‘MeanIoU’). Precision refers to the percentage of correctly classified positive pixels amongst all pixels predicted as positive. Recall gives the percentage of correctly classified positive pixels among all true positive pixels. F1 score is essentially the combination of precision and recall. The mean of Intersection-over-Union (mIOU) first computes the IOU for each pixel class and then computes the average over classes. The values of applied metrics are in the range of 0 to 1, wherein higher values indicate better classification performance. The experimental evaluation is more focused on mIOU since it is the standard metric for semantic segmentation. The metrics can be mathematically calculated as follows:

$$\text{Precision} = \text{TP} / (\text{TP} + \text{FP})$$

$$\text{Recall} = \text{TP} / (\text{TP} + \text{FN})$$

$$\text{F1-score} = (2 * \text{Precision} * \text{Recall}) / (\text{Precision} + \text{Recall})$$

$$\text{mIOU} = \text{TP} / (\text{TP} + \text{FP} + \text{FN})$$

where, TP = True Positive, FP = False Positive and FN = False Negative

4. Results and Discussions

Experiments were conducted on two publicly available datasets: Deep Globe and INRIA. Figure 4 shows the segmentation results of both datasets. From left to right are the test images, the ground-truth, the predicted output segmentation image. The qualitative and quantitative results demonstrate that the proposed model shows a higher mean IOU value of 0.79 for INRIA that is building extraction. It can be observed that buildings were extracted successfully with fewer classification errors and with sharper boundaries. Also, the model is able to extract road pixels however, it fails to maintain the connectivity due to class imbalance problems meaning a greater number of background pixels that is also evident from the precision and recall values of table 1. The model has found a local minimum that is evident from the graph of figure 3a and 3b. In the case of INRIA, the model returns more false positives and also predicts the building outlines reasonably well. The model is also compared with existing studies and was found to outperform the state-of-the-art methods (table 2 and 3). However, due to variability in the images of each subset, the model cannot perform well on all subsets. The evaluation metrics are tabulated in table 1.

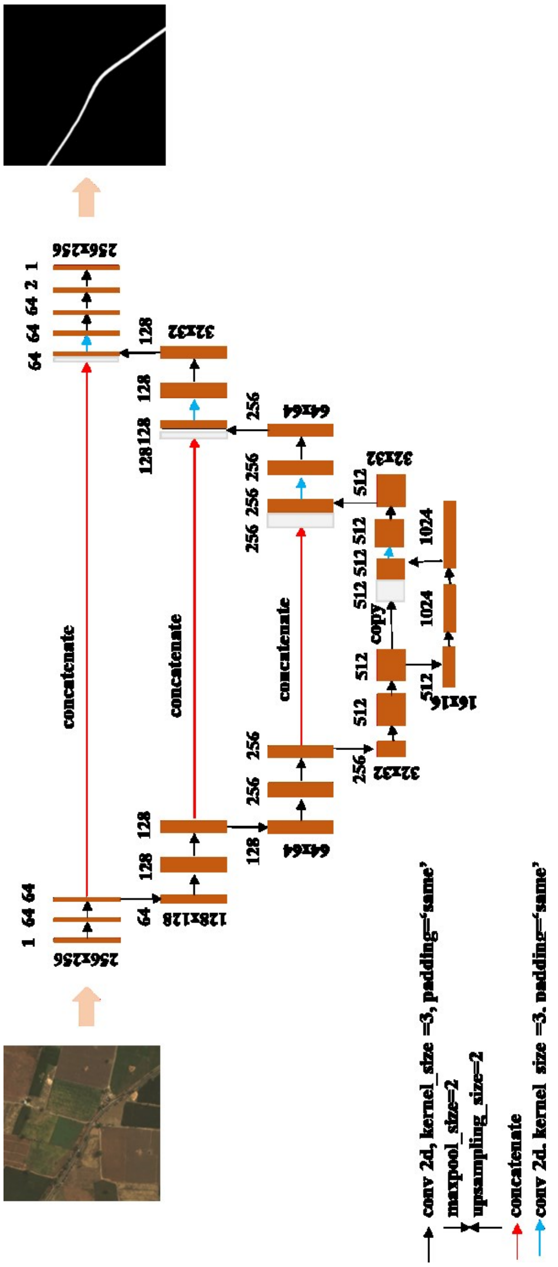


Figure 2(b). Network architecture of the proposed UNet model

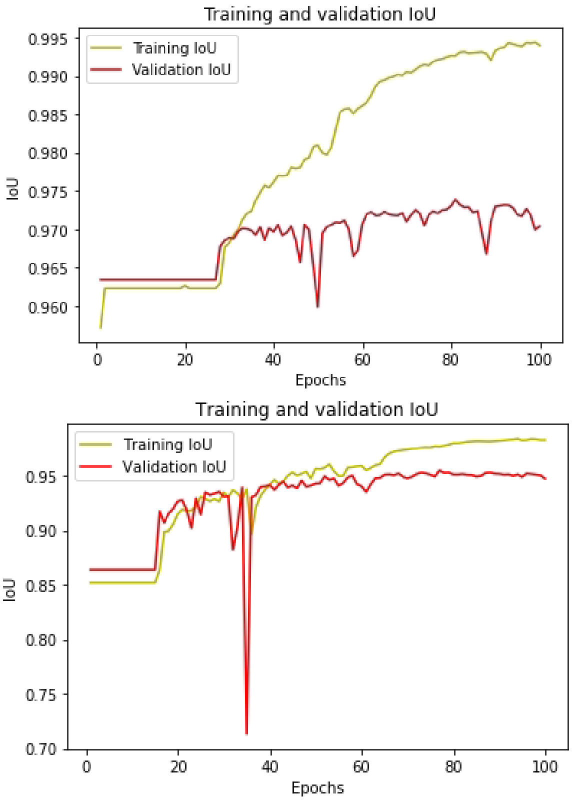


Figure 3. (a) Iou vs Epoch graph for Deep globe dataset (b) Iou vs Epoch graph for INRIA dataset for training and validation respectively

Table 1. Evaluation metric of Deep Globe and INRIA

Dataset	Precision	Recall	F1 score	mIOU
Deep Globe Dataset	0.82	0.305	0.445	0.625
INRIA Dataset	0.91	0.67	0.78	0.79

Table 2. Comparing results of INRIA dataset

Method	Iou
Ours	**0.79**
UNet+soft jaccard loss [12]	0.71

Table 3. Comparing results of Deep Globe dataset

Method	IoU
Ours	**0.625**
ResNwt50-D2S [1]	0.606

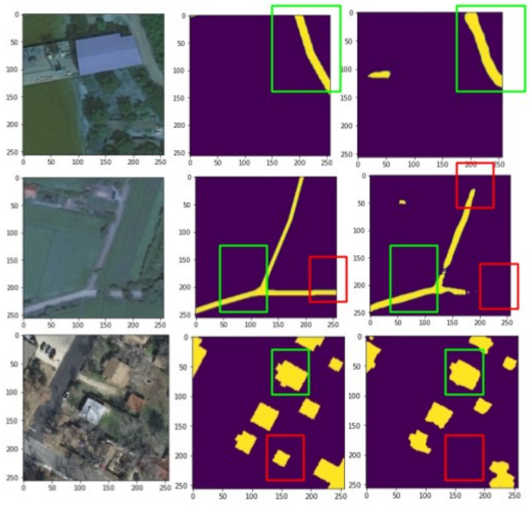


Figure 4(a). test image **Figure 4(b). ground truth image** **Figure 4(c). predicted image**

5. Conclusions

The aim of the study is to extract roads and building footprints from satellite images as a binary semantic image segmentation problem. For each input satellite image, the model predicts if a pixel belongs to class 1 (road or building) or class 0 (non-road and non-building). The distinct use of datasets for automated extraction compels the need to design our neural network with efficient memory optimizations. Despite bulk images, these datasets still fail to train a robust model for analyzing satellite imagery on a global scale. The challenges essentially involve spatial variations, because roads differ in their appearance due to regional terrain and urban density in developed vs developing countries complicates the model learning. However, the proposed UNet model based on contracting and expansive path performed well on both the datasets being different in spatial properties.

Acknowledgment

We are grateful to (i) Ranbir and Chitra Gupta School of Infrastructure Design and Management (ii) Sponsored research in Consultancy cell, Indian Institute of Technology Kharagpur, and (iii) West Bengal Department of Higher Education for the financial and infrastructure support. We would like to thank Digital Globe and INRIA Aerial Image Labeling Dataset for providing the open access repository of data.

References

Bharath H. A., M. C. Chandan, S. Vinay and T. V. Ramachandra (2018a). Modelling urban dynamics in rapidly urbanising Indian cities. The Egyptian Journal of Remote Sensing and Space Science, 21(3), 201-210.

Bharath H. A., S. Vinay, M. C. Chandan, B.A. Gouri and T.V. Ramachandra (2018b). Green to gray: Silicon Valley of India. Journal of environmental management, 206, 1287-1295.

Gleason S., R. Ferrell, A. Cheriyadat, R. Vatsavai and S. De (2010). Semantic information extraction from multispectral geospatial imagery via a flexible framework. In 2010 IEEE International Geoscience and Remote Sensing Symposium (pp. 166-169). IEEE.

Li W., H. Liu, Y. Wang, Z. Li, Y. Jia and G. Gui (2019). Deep learning-based classification methods for remote sensing images in urban built-up areas. IEEE Access, 7, 36274-36284.

Özyurt F. (2020). Efficient deep feature selection for remote sensing image recognition with fused deep learning architectures. The Journal of Supercomputing, 76(11), 8413-8431.

Petrovska B., E. Zdravevski, P. Lameski, R. Corizzo, I. Štajduhar and J. Lerga (2020). Deep learning for feature extraction in remote sensing: A case-study of aerial scene classification. Sensors, 20(14), 3906.

Prakash P. S. and H. A. Bharath (2020). Assessment of Urban Built-Up Volume Using Geospatial Methods: A Case Study of Bangalore. In IGARSS 2020-2020 IEEE International Geoscience and Remote Sensing Symposium (pp. 4239-4242). IEEE.

Ramachandra T. V., B. H. Aithal and D. D. Sanna (2012). Insights to urban dynamics through landscape spatial pattern analysis. International Journal of Applied Earth Observation and Geoinformation, 18, 329-343.

Ramachandra T. V., A. H. Bharath and M. V. Sowmyashree (2015). Monitoring urbanization and its implications in a mega city from space: Spatiotemporal patterns and its indicators. Journal of environmental management, 148, 67-81.

Ronneberger O., P. Fischer and T. Brox (2015). U-net: Convolutional networks for biomedical image segmentation. In International Conference on Medical image computing and computer-assisted intervention (pp. 234-241). Springer, Cham.

Tran A., A. Zonoozi, J. Varadarajan and H. Kruppa (2020). PP-LinkNet: Improving Semantic Segmentation of High-Resolution Satellite Imagery with Multi-stage Training. In Proceedings of the 2nd Workshop on Structuring and Understanding of Multimedia heritAge Contents (pp. 57-64).

VoPham T., J. E. Hart, F. Laden and Y.Y. Chiang (2018). Emerging trends in geospatial artificial intelligence (geoAI): potential applications for environmental epidemiology. Environmental Health, 17(1), 1-6.

Vorona D., A. Kipf, T. Neumann and A. Kemper (2019, November). DeepSPACE: Approximate geospatial query processing with deep learning. In Proceedings of the 27th ACM SIGSPATIAL international conference on advances in geographic information systems (pp. 500-503).

Zhang W., P. Tang and L. Zhao (2019). Remote sensing image scene classification using CNN-CapsNet. Remote Sensing, 11(5), 494.