A Modular System Based on U-Net for Automatic Building Extraction from very high-resolution satellite images

Smail Ait El Asri [1,*], Samir El Adib[1], Ismail Negabi[1], and *Naoufal* Raissouni[1]

[1]Remote Sensing and GIS Research Unit (RS and GIS), National School for Applied Sciences of Tetuan, University Abdelmalek Essaadi, Morocco.

Abstract. Recently, convolutional neural networks have grown in popularity in a variety of fields, such as computer vision and audio and text processing. This importance is due to the performance of this type of neural network in the state of the art, and in a wide variety of disciplines. However, the use of convolutional neural networks has not been widely used for remote sensing applications until recently. In this paper, we propose a CNN-based system capable of efficiently extracting buildings from very high-resolution satellite images, by combining the performances of the two architectures; U-Net and VGG19, which is obtained by putting two blocks in parallel based mainly on U-Net: The first block is a standard U-Net, and the second is designed by replacing the contraction path of standard U-Net with the pre-trained weights of VGG19.

1 Introduction

Deep learning methods have been demonstrated to successfully handle a wide variety of problems known by their remarkable complexity, such as image analysis, speech recognition, and driving autonomous vehicles. Automatic building Extraction is extremely important in several applications, such as urban planning, updating data in geographic information systems, digital city building and disaster damage assessment [1].

Recently, high-resolution satellite imagery was providing a new source of data for remote sensing applications. Automatic building extraction is the most important task, which is becoming relatively easy thanks to the availability of very high resolution satellite images, which provide very fine details in urban and suburban areas. Researchers have developed approaches based on traditional image processing methods, which consider, mainly, spectra, texture and shape as input features to the classifiers, which are mainly Random Forest (RF), support vector machine (SVM), or AdaBoost. After the appearance and development of deep learning, and after its remarkable success in image classification, the community of remote sensing is increasingly focusing on the use of deep learning in a variety of applications, such as automatic building extraction [2,3].

Automatic buildings extraction from very high-resolution satellite images remains extremely important to accomplish several tasks (infrastructure planning, change detection). Indeed, Building extraction is included in semantic segmentation in computer vision. Convolutional neural networks (CNNs), have proved their capabilities to obtain very promising results in semantic segmentation, such as U-Net [4], ResNet [5], fully convolutional network (FCN) [6], SegNet [7], DenseNet [8], and NDNet [9]. This success typically, comes from the fact that in terms of accuracy and efficiency, CNN exceeds other techniques.

2 RELATED WORKS

Before the development of deep learning, the remote sensing community were focused on using a set of algorithms, such as the Support Vector Machine (SVM) [10] and Random forest [11]. SVM attract the attention of researchers because it is capable of handling data of high dimensionality, and can be trained efficiently with a limited amount of data, while Random forest [11] was gained its popularity because of its efficiency and insensitivity to classification parameters.

In recent years, and more specifically since 2014, the remote sensing community has renewed its interest in neural networks, focusing more on deep learning algorithms, which have shown remarkable success in several remote sensing application, such as image fusion which is a main task in remote sensing, which serves, mainly, to obtain an image of very high spectral and spatial resolution, image registration which allows the alignment of images captured by different sensors, at different angles of view, or at different times, scene classification [12], object detection [13,14], Land-use

[*] Corresponding author: smail.aitelasri@etu.uae.ac.ma

and land-cover (LULC) classification, semantic segmentation, which is relatively easy thanks to the appearance and development of deep CNNs, and object-based image analysis [15].

Building extraction from very high-resolution satellite images is a complex remote sensing task, it is an important task for urban planning, urban change detection, and urbanization monitoring [16]. Works are done in the direction of improving the efficiency of building extraction on satellite images using convolutional neural networks. CNNs can efficiently extract very deep information from the input image [17], This is achieved by using local connections to efficiently extract spatial information and shared weights in order to minimize the number of parameters [18].

Inspired by deep learning's exceptional achievement in image classification, automatic language translation, speech recognition, some researchers have used deep learning methods for remote sensing image segmentation tasks [19-21]. Other works are conducted with the aim of extracting buildings [2,3,22]. Lu et al. [23] proposed a building edge detection model with the Richer Convolution Features network (RCF), this model has been able to efficiently detect the edges of buildings, which is suitably an improvement at this level. Li et al. [24] proposed a building footprint extraction method using semantic segmentation based on U-Net. Wu et al. [25] introduced an approach named 'Boundary Regulated Network' (BR-Net), to properly segment and extract the contours of buildings from aerial images. The BR-Net achieves significant results.

3 PROPOSED APPROACH

The proposed approach is based on the study of existing works in the literature in the field of satellite image processing. Figure 1. illustrates a global overview of our system, which is mainly based on the U-Net architecture [4]. Two architectures based on the U-Net have been placed in parallel. The first one is trained from scratch using the training data, allowing the network to learn to extract characteristics specific to the training data. While, the second network is based on the pre-trained VGG19 as a feature extractor, which helps to retrieve deep and global features from the input data.

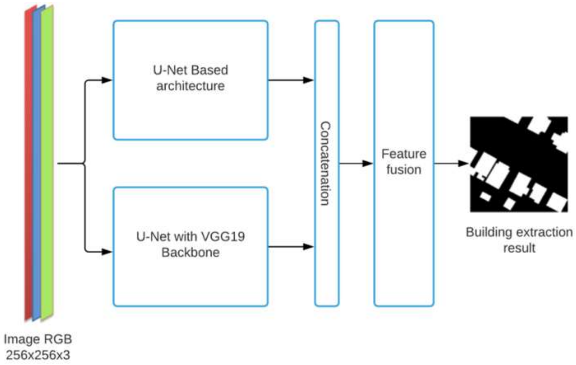

Fig. 1. Overview of our proposed approach.

The U-Net based block is designed to combine both speed in the learning phase, and efficiency in buildings extraction. The speed-up is ensured by minimizing the number of parameters of the network, so this minimization does not affect the efficiency. The architecture consists of a contraction path to capture the context, and a symmetric expansion path that allows the accurate localization of objects in the input image. The network configuration, in terms of number of filters, is (16,32,64,128,256), with a kernel size (3x3). This block receives an image of size (256x256x3) and generates a mask of size (256x256x1). U-Net based on VGG19 pre-trained weights is deeper than the previous; it has a deeper configuration in terms of number of filters and layers. Figure 2. shows some results from the first convolutional layers of VGG19.

In order to merge the two masks generated by two U-Net blocks, we chose to concatenate them to keep the complete information of each mask. This will ensure that the fusion block learns to merge the two generated masks, so that the final result will contain the useful information of the input data. This block is formed by two convolutional layers with 64 and 32 filters, respectively.

4 DATASET DESCRIPTION

The dataset used in our study is Inria Aerial Image Labeling (https://project.inria.fr/aerialimagelabeling/), It was made by merging public domain images and official building footprints from the public domain. It Covered an area of 810 km2 (405 km2 for training, and 405 km2 for testing). It is an orthorectified color aerial imagery with a spatial resolution of 0.3 m. Table 1 summarizes the statistics of this dataset. The ground truth data is divided into two semantic categories: building and non-building.

Table 1. Dataset description.

Train	Tiles	Area ($km2$)	Test
Kitsap Country (WA)	36	81	Bloomington (IN)
West Tyrol (Austria)	36	81	East Tyrol (Austria)
Austin (TX)	36	81	Bellingham (WA)
Vienna (Austria)	36	81	Innsbruck (Austria)
Chicago (IL)	36	81	San Francisco (CA)

5 RESULTS

5.1 Extracted features with VGG19 first layers

We notice in Fig. 2, that b) shows that the filter corresponding to this result distinguishes the buildings from other objects presented in the image, while c) highlights the shadow exists in the image, which will allow our model to learn to distinguish and recognize the shadow in the image. d) showed that the filter corresponding to this result, tends to lighten the trees in the image, which is important, because most satellite images contain, usually, buildings emerged in the trees. e) is an example among others, filters that generate contours and corners that exist in the input image.

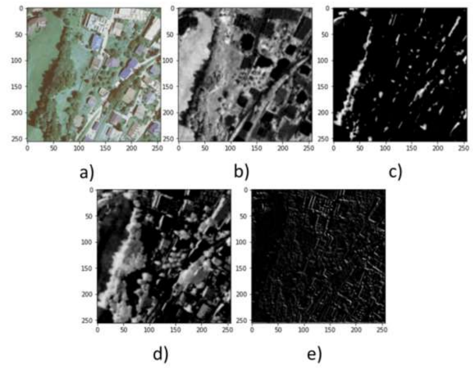

Fig. 1. Results from VGG19 first convolutional layers.

5.2 Results from the U-Net block

An important advantage of our system is the possibility to train each module individually. We trained the U-Net block on 7893 images of size 256x256x3, and we validated our results on 877 images. And after 50 iterations, this block achieves an IoU of 76.05\%. This encourages us to overcome the technical limitations we faced, in order to move to data augmentation techniques, and to train our system with powerful GPUs. Fig. 3 illustrates the partial results obtained at this level.

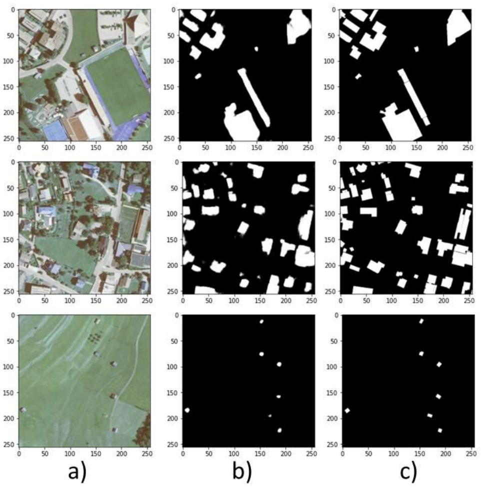

Fig. 2. Results from U-Net block, a) input image, b) predicted mask, and c) ground truth

3 CONCLUSION AND PERSPECTIVES

Our approach is based on combining the performance of VGG19, which was trained on the 'imagenet' dataset, and the performance of the U-Net. We used the pre-trained filters of VGG19 as a feature extractor for the U-Net, specifically, we replaced the contraction path of the U-Net with the convolutional layers of VGG19. This architecture was used because it is able to retrieve deeper features from the inputs images. Adding a parallel U-Net network allows us to have an extractor trained specifically on the satellite images, which will permit to merge the performances of each network in order to obtain a remarkable result in the building extraction.

In future work, we will train our system on the entire dataset (Inria Aerial Image Labeling), in order to have a system that can be generalized better on test data. We will also make use of the data augmentation technique, which will produce improved results compared to those obtained in this paper.

References

1. C. Xiong, Q. Li, and X. Lu, Autom. Constr. (2020)
2. H. Yang, P. Wu, X. Yao, Y. Wu, B. Wang, and Y. Xu, Remote Sens. (2018)
3. B. Huang, K. Lu, N. Audebert, A. Khalel, Y. Tarabalka, J. Malof, A. Boulch, B. Le Saux, L. Collins, K. Bradbury, S. Lefèvre, and M. El-Saban, in *Int. Geosci. Remote Sens. Symp.* (Institute of Electrical and Electronics Engineers Inc., 2018), pp. 6947–6950
4. O. Ronneberger, P. Fischer, and T. Brox, in *Lect. Notes Comput. Sci. (Including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)* (2015)
5. K. He, X. Zhang, S. Ren, and J. Sun, in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.* (2016)
6. E. Shelhamer, J. Long, and T. Darrell, IEEE Trans. Pattern Anal. Mach. Intell. (2017)
7. V. Badrinarayanan, A. Kendall, and R. Cipolla, IEEE Trans. Pattern Anal. Mach. Intell. (2017)
8. G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, Proc. - 30th IEEE Conf. Comput. Vis. Pattern Recognition, CVPR 2017 **2017-Janua**, 2261 (2017)
9. Z. Yang, H. Yu, Q. Fu, W. Sun, W. Jia, M. Sun, and Z.-H. Mao, IEEE Trans. Intell. Transp. Syst. (2020)
10. G. Mountrakis, J. Im, and C. Ogole, ISPRS J. Photogramm. Remote Sens. **66**, 247 (2011)
11. M. Belgiu and L. Drăgu, ISPRS J. Photogramm. Remote Sens. **114**, 24 (2016)
12. G. Cheng, C. Yang, X. Yao, L. Guo, and J. Han, IEEE Trans. Geosci. Remote Sens. **56**, 2811 (2018)
13. Y. Zhong, X. Han, and L. Zhang, ISPRS J. Photogramm. Remote Sens. **138**, 281 (2018)
14. P. Ding, Y. Zhang, W. J. Deng, P. Jia, and A.

Kuijper, ISPRS J. Photogramm. Remote Sens. **141**, 208 (2018)
15. G. Castilla and G. J. Hay, Lect. Notes Geoinf. Cartogr. **0**, (2008)
16. M. Vakalopoulou, K. Karantzalos, N. Komodakis, and N. Paragios, in *Int. Geosci. Remote Sens. Symp.* (Institute of Electrical and Electronics Engineers Inc., 2015), pp. 1873–1876
17. H. Wu and X. Gu, Neural Networks (2015)
18. D. Lunga, H. L. Yang, A. Reith, J. Weaver, J. Yuan, and B. Bhaduri, IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. **11**, 962 (2018)
19. H. Lin, Z. Shi, and Z. Zou, Remote Sens. (2017)
20. Y. Bai, E. Mas, and S. Koshimura, Remote Sens. (2018)
21. G. Cheng, F. Zhu, S. Xiang, Y. Wang, and C. Pan, Neurocomputing (2016)
22. M. Dixit, K. Chaurasia, and V. Kumar Mishra, Expert Syst. Appl. (2021)
23. T. Lu, D. Ming, X. Lin, Z. Hong, X. Bai, and J. Fang, Remote Sens. (2018)
24. W. Li, C. He, J. Fang, J. Zheng, H. Fu, and L. Yu, Remote Sens. (2019)
25. G. Wu, Z. Guo, X. Shi, Q. Chen, Y. Xu, R. Shibasaki, and X. Shao, Remote Sens. (2018)