Semantic Segmentation of Satellite Imagery using DeepLabV3+

Sarach Rujiranurak
Data Science and Artificial Intelligent
Asian Institute of Technology
st121628@ait.asia

Thoviti Siddharth
Information Management
Asian Institute of Technology
st121362@ait.asia

Witoon Wiphusitphunpol
Data Science and Artificial Intelligent
Asian Institute of Technology
st121416@ait.asia

***Abstract*— Semantic segmentation of remote sensing imagery has become an important role in natural resource management and urban planning development as it can provide enormous economic value for agriculture, forestry, or public administration. Satellite imagery allows the assessment of change detection of the environment using deep learning techniques combined with computer vision to provide accurate results in a short time without being human-intensive which usually requires experts to work on this and takes a very long time to accomplish this kind of data.**

This study introduces a deep learning model, DeepLabV3+, trained with the satellite images, with RGB channels, from DeepGlobe Satellite Image dataset, providing semantic segmentation for the satellite images detecting multiple classes, buildings, woods, crops, and water. The performance of the model was evaluated by the mean intersection over union (mIoU) score at 0.75 on the validation set given in the dataset. Therefore, this model could be applied to provide rough details for detecting the change of the environment from satellite imagery without using experts or human-intensively in a large-scale dataset.

***Keywords* — DeepLabV3+, DeepGlobe, Semantic Segmentation, Satellite imagery.**

I. Introduction

Satellite imagery and GIS maps are key to many applications such as monitoring deforestation and urbanization, agriculture, hydrology. Accurate information on land cover is necessary to plan and monitor resources in any region. Automatic change detection is valuable to forestry departments where detecting deforestation and illegal logging is essential and detecting small-scale changes are crucial.

In remote sensing, computer vision plays a vital role in bridging the objective of processing high-resolution geospatial images and performing segmentation tasks. Public datasets such as the DeepGlobe Satellite Image dataset are used to benchmark the performance of different approaches. Satellite images are rich in the information required to perform computer vision tasks since these images contain uniform data due to the large area of various landforms.

Classification of vegetation objects is done based on the shape, size, boundary and seasonal change in vegetation. The color of the vegetation varies with the soil, crop type, season and humidity which makes the classification a tougher task since intra-class differences may be larger than inter-class differences. This paper proposes to solve these problems by implementing the DeepLabv3+ architecture with ResNet backbone trained on ImageNet and train the model on the DeepGlobe dataset.

The remainder of this paper is organized as follows. In Section II, we discuss a background study that addressed this problem. In Section III, we give a brief summary of the dataset. In Section IV, we summarize the experiments and results that we obtained. Finally, Section V concludes the paper by summarizing the results and indicating issues to be addressed in the future work.

II. BACKGROUND STUDY AND RELATED WORK

A. Semantic Segmentation

Fully convolutional network (FCN) [1] was the first CNN architecture used for semantic segmentation. To enrich the contextual information in FCN, spatial pyramid pooling was adopted. Other methods include Mixed Spatial Pyramid Pooling (MSPP) [2] module which performs region-based average pooling and dilated convolution to get multi-level contextual priors. Architectures such as U-Net [3] used skip-connections and the encoder-decoder architecture. SegNet [4] and DeepLab [5,6] used conditional random fields (CRF) to model the spatial relationship. The drawback of these models is that they utilized high GPU memory for high-resolution images like satellite images.

B. DeepLabV3

In previous architectures, Spatial pyramid pooling (SPP) [7] was used. SPP adds a layer between the convolution layers and the fully-connected layers to map any size input down to a fixed size.

The Deeplabv3 [8] architecture uses atrous spatial pyramid pooling (ASSP) since consecutive strides in a simple convolution for semantic segmentation would lose spatial information at deeper layers but using an ASSP would keep the stride constant with a larger field-of-view and reduces the number of parameters. It also includes batch normalization during training.

C. DeepLabV3+

The DeepLabv3+ [9] extends the DeepLabv3 architecture by adding a decoder to improve the segmentation results along the object boundaries.

III. DATASET

The dataset used in this study was obtained from Land Cover Classification Track in DeepGlobe Challenge, which is published on Kaggle [10], providing satellite imagery in RGB with mask images indicating land cover annotation.

A. Data Characteristics

There are 803 RGB images in this dataset with a size of 2448 x 2448 pixels. Each image has a 50 cm pixel resolution, collected by DigitalGlobe's satellite image. It is split into 3 sets, training set, validation set, and test set with size, 460, 171, and 172 images. Both training set and validation set are given with the mask to indicate the land cover area but there is no mask for the test set.

B. Label

The mask images provided with the dataset are represented in RGB images with 7 classes of labels using color-coding, the values of each channel in each pixel, as follows.

- Urban land
  It is represented as [0, 255, 255] at each pixel indicating man-made built areas, buildings, or houses but it does not include roads, which is too difficult to be labeled accurately by humans.
- Agriculture land
  It is defined as [255, 255, 0] for each part of an image that shows farms, cropland, orchards, vineyards, nurseries, and ornamental horticultural areas.
- Rangeland
  Any non-forest, non-farm, green land, or grass areas are represented as [255, 0, 255].
- Forest land
  Any land with some amount of tree crowded areas is shown as [0,255,0].
- Water
  The water area includes rivers, oceans, lakes, wetlands, and ponds. All are [0,0,255] for the RGB channel in each image.
- Barren Land
  Mountain, land, rock, desert, beach, and no vegetation areas are shown in [255,255,255] for each pixel.
- Unknown
  The unknown parts are represented as [0,0,0], showing clouds or others.

Fig 1. The color representation for the masked defined in the dataset.

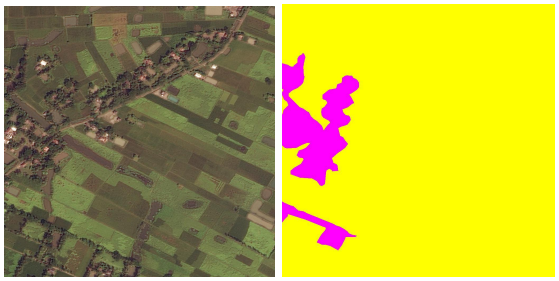

Fig 2. The example of satellite imagery from the dataset with the masked labeled by humans.

IV. Experiments

The model used in this study is chosen from one of the state-of-the-art deep learning architectures, DeepLabV3+, using Resnet50 as the encoder.

A. Data Preprocessing

There are 6 classes selected for the model's targets. The unknown class is excluded as it could affect the model training due to its ratio in each image, which usually covers more than half of the area in each sample, leading to the problem of bias on the prediction.

The proper augmentation process could encourage better performance for the model as it simulates the various conditions for real situations. All images in the dataset are resized from 2448✕2448 to 512✕512 to reduce the number of parameters of the input. Then, there are 11 augmentation processes applied for the training set by randomly changing the following parameters: flipping, rotation, hue, grayscale, contrast, brightness, sharpness, adding Gaussian noise, random gamma, sharpening, and RGB shifting.

B. Training

The model was trained on GPU kernel on Google Colab PRO with 1 shared GPU, using the batch-size of 4, dice loss function, and adam optimizer with 0.00008 learning rate adjusted by a scheduling function called Cosine Annealing Warm Restarts.

C. Results

The result from the experiment proves that automatic mapping from satellite images is possible with deep learning. Our model reaches 75% of mIoU of the validation set. However, some interesting observations from the model's mistakes would be the limitations in practice.

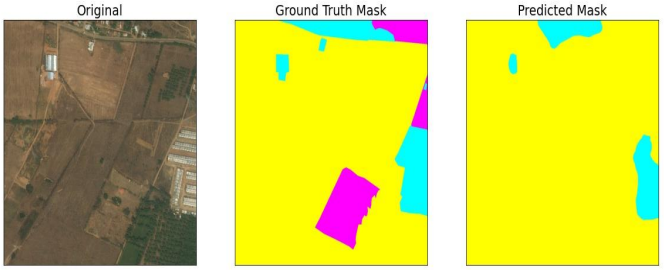

Fig 3. the comparison of the model's outputs compared to the expected masks labeled by humans on the validation set.

The model performs well at labeling the cultivation and urban areas but other categories are rarely detected. Due to the size of their labels, the numbers of samples and characteristics of the data are similar to others. One class, that is never answered from the model's output, is "rangeland" as it looks similar to farmland and forest, thus the model always predicts rangelands as agricultural zones or forest. The model output seems to ignore small masks, especially for forestry zones or water areas. The class that model performs well is the farmland as it often covers more than 50% of an image.

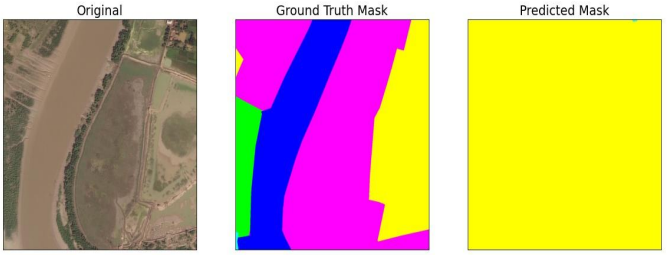

Fig 4. The example of some mistakes from the model outputs compared with the actual labeled on the validation set. The model predicted the rangeland as the agricultural fields.

V. Conclusion

In this study, we present the performance of our deep learning model, DeepLabV3+, trained with the DeepGlobe Challenge dataset, receiving RBG satellite high-resolution (50 centimeters per pixel) images to classify the area into 6 categories.

As we demonstrated, our model can be applied as tools for automatic mapping with neural networks, diminishing the labor-intensive tasks

with better efficiency and accuracy of identifying changes in land use and land cover. Moreover, our model can be beneficial in various domains such as administration, agriculture, forestry, and resource management.

In the future, we plan to improve the performance of our model with more augmentations processes or editing its architecture and develop the application of automatic mapping for individuals or researchers to monitor the change over the area from their dataset facilitating them to detect and calculate the size of each area for their work within a short time.

References

[1] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 3431–3440, 2015.

[2] Zhengyu Xia, Joohee Kim. Mixed spatial pyramid pooling for semantic segmentation. In Applied Soft Computing, Volume 91, 2020, 106209, ISSN 1568-4946, https://doi.org/10.1016/j.asoc.2020.106209.

[3] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. Unet: Convolutional networks for biomedical image segmentation. In International Conference on Medical image computing and computer-assisted intervention, pages 234–241. Springer, 2015

[4] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. IEEE transactions on pattern analysis and machine intelligence, 39(12):2481–2495, 2017.

[5] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Semantic image segmentation with deep convolutional nets and fully connected crfs. arXiv preprint arXiv:1412.7062, 2014.

[6] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. IEEE transactions on pattern analysis and machine intelligence, 40(4):834–848, 2018.

[7] K. He, X. Zhang, S. Ren and J. Sun, "Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 37, no. 9, pp. 1904-1916, 1 Sept. 2015, doi: 10.1109/TPAMI.2015.2389824.

[8] Chen, Liang-Chieh & Papandreou, George & Schroff, Florian & Adam, Hartwig. (2017). Rethinking Atrous Convolution for Semantic Image Segmentation.

[9] Chen, Liang-Chieh and Zhu, Yukun and Papandreou, George and Schroff, Florian and Adam, Hartwig. (2017). Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation

[10] Balraj Ashwath, DeepGlobe Land Cover Classification Dataset, Kaggle, [online] Available at: <https://www.kaggle.com/balraj98/deepglobe-land-cover-classification-dataset> [Accessed 1 May 2021].

APPENDIX

Fig 5. The example of the model's outputs compared to the ground truth, labeled by humans. Images on the left side, first column, represent the original RGB images from the validation set. The middle column shows the target masks represented in colors. The last column on the right presents the prediction decoded from the model to colors.