Satellite Imagery Feature Detection using Deep Convolutional Neural Network: A Kaggle Competition

Vladimir Iglovikov[1], Sergey Mushinskiy[2] and Vladimir Osin[3]

[1]True Accord
[2]Open Data Science
[3]AeroState

Abstract

This paper describes our approach to the DSTL Satellite Imagery Feature Detection challenge run by Kaggle. The primary goal of this challenge is accurate semantic segmentation of different classes in satellite imagery. Our approach is based on an adaptation of fully convolutional neural network for multispectral data processing. In addition, we defined several modifications to the training objective and overall training pipeline, e.g. boundary effect estimation, also we discuss usage of data augmentation strategies and reflectance indices. Provided solution can be used as a part of automatic feature labeling systems for satellite imagery analysis.

1. Introduction

The significant increase of satellite imagery has given a radically improved understanding of our planet. Object recognition in aerial imagery is enjoying a growing interest today, due to the recent advancements in computer vision and deep learning, along with important improvements in low - cost high-performance GPUs. The possibility of accurately distinguishing different types of objects in aerial images, such as buildings, roads, vegetation and other categories, could greatly help in many applications, such as creating and keeping up-to-date maps, improving urban planning, environment monitoring, and disaster relief. Besides the practical need for accurate aerial image interpretation systems, this domain also offers scientific challenges to the computer vision domain.

In this paper, we describe and analyze these challenges for the specific satellite imagery dataset from Kaggle competition. We explore the challenges faced due to the small size of the dataset, the specific character of data, and supervised and unsupervised machine learning algorithms that are suitable for this kind of problems. Our efforts can be summarized as follows:

- We adapted fully convolutional network to multispectral input data and evaluated several data fusion strategies on semantic segmentation task of satellite images.
- We introduced joint training objective that properly defines desired output for the segmentation task.
- We analyze local and global boundary effect on overall performance of the segmentation pipeline.

2. Related Work

Semantic segmentation for images can be defined as the process of partition the image into meaningful parts, and classify each part at the pixel level into one of the predefined classes. The current success of deep learning techniques in computer vision tasks motivated researchers to explore such techniques for pixel-level classification tasks like semantic segmentation. The Convolutional Neural Networks (CNN) is a main supervised approached that successfully used for this task. The key advancement of these networks is the ability to learn appropriate feature representation in an end-to-end manner avoiding creation of handcrafted features which require too much tuning to make them work on a particular case.

The most successful state-of-the-art deep learning method is the Fully Convolutional Network(FCN) [1]. The main idea of this approach is the usage of CNN as powerful feature extractor while replacing the fully connected layers with convolution ones to output spatial maps instead of classification scores. Those maps are upsampled to produce dense per-pixel output. This method allows training CNN in the end to end manner for segmentation with input images of arbitrary sizes. This approach achieved a notable enhancement in segmentation accuracy over common methods on standard datasets like PASCAL VOC [2].

Our solution is based on modified fully convolutional neural network architecture called U-Net [3], that was previously used for the tasks of biomedical image segmentation. The U-Net architecture allows combining low-level feature maps with higher-level ones, which enables precise localization. A large number of feature channels in up-sampling part allows propagating context information to higher resolution layers. This type of network architecture was specially designed to solve image segmentation problems effectively. Technical details of U-Net adaptation for discussed task provided in Section 3.

3. Methodology

3.1. Data description

During competition dataset was separated into public and private parts. Public part was used for model evaluations during first stage of the competition and final models were evaluated on hidden private test set. The dataset consist of 57 images that divided into train (25 images) and test (32 images) sets. Each image covers 1 square kilometer of the earth surface. Satellite images of the same area can be separated in several types: a high-resolution panchromatic, an 8-band image with a lower resolution (M-band), and a short-wave infrared (A-band) that has the lowest resolution of all.The detailed band description provided in 3.2.

RGB and M-band images partially overlap in the optical spectral range, because high-resolution RGB was itself reconstructed by procedure called panchromatic sharpening. Panchromatic sharpening uses a higher-resolution panchromatic image to perform fusion with a lower-resolution M-band image. The result produces a M-band image with the resolution of the panchromatic where the two rasters fully overlap. We apply several commonly used methods of panchromatic sharpening [10] to provided data (Fig.1). Sharpened channels of M-band images can be used as alternative input to neural network.

Figure 1: Panchromatic sharpening results for multispectral and panchromatic bands.

The challenge is assign one or more class labels to each pixel of the input image. The classes with additional description presented in Table. 1.

The plots in Fig. 2 shows that target classes are heavily imbalanced within each set pf images, e.g. one pixel of both large and small vehicle classes corresponds to a 60,000 pixels with crops. Thereby training a separate model per class provides much better results that single model for prediction of all classes. In addition, the distributions themselves vary significantly from one set of images to another. Also, the water classes were under-represented in the train set compared to the public and the private test sets. As a result unsupervised methods, that described in Section 3.3 shows better performance than neural network approaches.

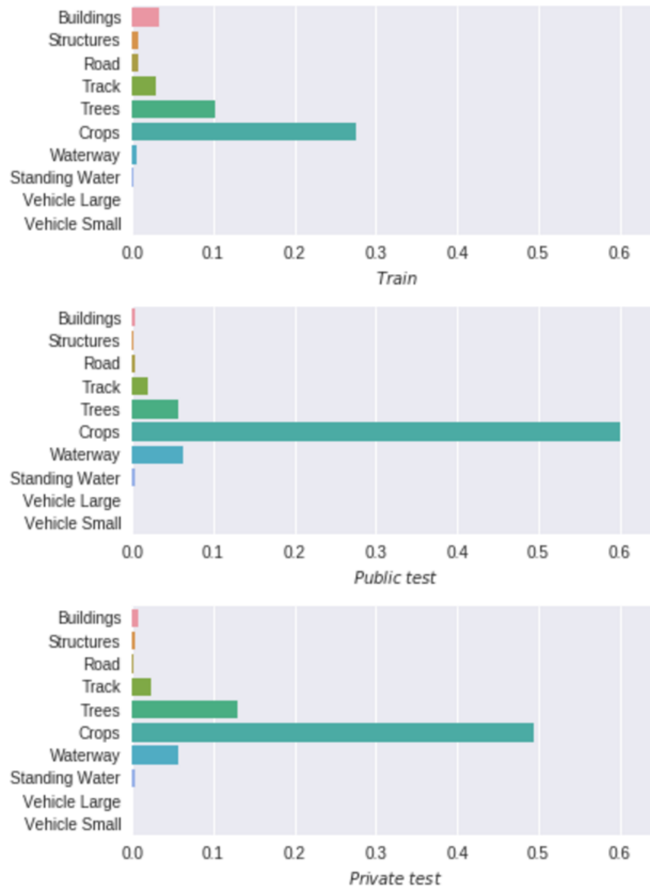

Figure 2: Distributions of target classes for public and private parts of dataset.

3.2. Multispectral Sensor

WorldView-3 satellite is one of the commercial satellites, which provide 31 cm panchromatic resolution, 1.24 m multispectral resolution, and 7.5 m resolution for short-wave infrared channels. Multispectral images enable to extract important features that beyond human vision, e.g. the near infrared wavelength is usually used to separate vegetation varieties and conditions due to strong reflection in this range of electromagnetic spectrum that vegetation provides. Besides, the color depth of such images is 11 and 14-bit instead of commonly used 8-bit. From a neural network perspective

Class	Additional Description
Buildings	large buildings, residential, non-residential, fuel storage facilities, fortified building
Structures	man-made structures
Road	-
Track	poor/dirt/cart tracks, footpaths/trails
Trees	woodland, hedgerows, groups of trees, stand-alone trees
Crops	contour ploughing/cropland, grain crops, row (potatoes, turnips) crops
Waterway	-
Standing water	-
Vehicle Large	large vehicle (e.g. lorry, truck,bus), logistics vehicle
Vehicle Small	small vehicle (car, van), motorbike

Table 1: Dataset contains industrial and nature classes. Described classes varied in terms of shape and size.

it is better, each pixel carries more information, while it creates additional steps for proper visualization.

The multispectral bands can be used for recognition of specific classes of object:

- **Coastal (400-452 nm)**. This band senses deep blues and violets. It is also called the coastal/aerosol band, after its two main uses: imaging shallow water, and tracking fine particles like dust and smoke.

- **Blue (448-510 nm)**. This band senses normal blues. It provides increased penetration of water bodies and also capable of differentiating soil and rock surfaces from vegetation and for detecting cultural features.

- **Green (518-586 nm)**. This band senses greens. Because it covers the green reflectance peak from leaf surfaces, it has separated vegetation (forest, croplands with standing crops) from soil. In this band urban areas, roads and highways have appeared as brighter tone, but forest, vegetation, croplands with standing crops have appeared as dark (black) tone.

- **Yellow (590-630 nm)**. This band senses in a strong chlorophyll absorption region and strong reflectance region for most soils. It has separated vegetation and soil. But it could not separated water and forest. Forest land and water both have appeared in dark tone. This band has highlighted barren lands, urban areas, street pattern in the urban area and highways. It has also separated croplands with standing crops from bare croplands with stubble.

- **NIR (772-954 nm)**.This band measures the near infrared. This part of the spectrum is especially important for ecology purposes because healthy plants reflect it. Information from this band is important for major reflectance indexes, such as NDVI [4], which allow to measure specific characteristics more precisely.

- **SWIR (1195-2365 nm)**. This band cover different slices of the shortwave infrared. They are particularly useful for telling wet earth from dry earth, and for geology: rocks and soils that look similar in other bands often have strong contrasts in this band.

3.3. Reflectance indices

Apparently fact that we have infrared and other channels from non-optical frequency range allows to identify some classes purely from the pixel values, without any contextual information. Using this approach, the best results were obtained for water and vegetation classes. For instance, in our final solution both water classes were segmented using CCCI [5] and NDWI reflectance indexes, that defined as follows:

$$\mathrm{CCCI} = \frac{\mathrm{NIR} - \mathrm{RED_{edge}}}{\mathrm{NIR} + \mathrm{RED_{edge}}} \times \frac{\mathrm{NIR} + \mathrm{RED}}{\mathrm{NIR} - \mathrm{RED}} \tag{1}$$

$$\mathrm{NDWI} = \frac{\mathrm{GREEN} - \mathrm{NIR}}{\mathrm{GREEN} + \mathrm{NIR}} \tag{2}$$

where both indexes represented as ratios of the difference and sum of pixel values in the green, red edge, red and infrared channels. The results shows high intensity values for waterways, but it also shows false positives on some buildings due to the relative similarity of the specific heat of metal roofs and water. Water classes segmentation results presented in Fig. 3.

We expected a deep learning approach to perform as well as or even better than index thresholding and, in vegetation prediction, neural networks did indeed outperform indices. However, we found that indices allow us to achieve better results for under-represented classes such as waterways and standing water. In the provided images, ponds were smaller than rivers, so we additionally thresholded our predictions by area of water body to distinguish waterways from standing water.

3.4. Multispectral U-NET

In general, U-Net architecture consist of contracting and expansive paths. The contractive path follows the typical convolution neural network architecture. We use batch normalization [8] for convergence acceleration during training. In addition, instead of rectified linear unit we use exponential linear [9] unit as primary activation function, which

Bands	Spectral Range	Resolution	Dynamic Range
Panchromatic	450-800 nm	0.31 m	11 bits/pixel
Multispectral 8-bands	400-1040 nm	1.24 m	11 bits/pixel
SWIR 8-bands	1195-2365 nm	7.5 m	14 bits/pixel

Table 2: Specifications for the WorldView-3.

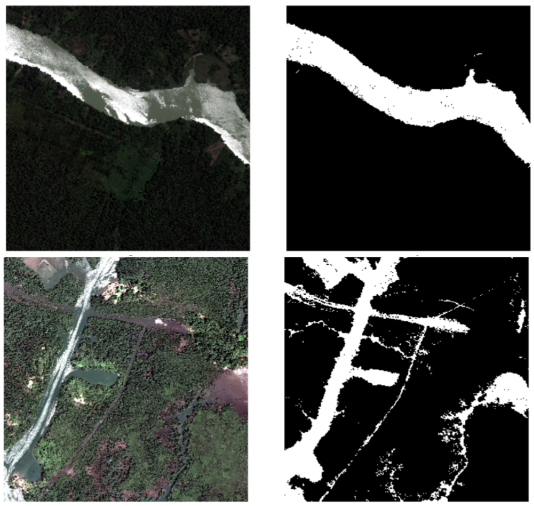

Figure 3: Reflectance indexes (thresholded CCCI) showed best performance in segmentation of waterways.

is beneficial for learning and it helps to learn representations that are more robust to noise. The number of feature channels is doubled at each down-sampling step. Expansive path consist of up-sampling operation of the feature map followed by convolution with half number of feature channels, concatenation with the corresponding feature map from contracting path, also followed by batch normalization and ELU.

As a primary input we perform early fusion of multispectral bands, reflectance indices and RGB channels, stacking them into single tensor. The full architecture presented in Fig. 4.

Evaluation metric is Jaccard index, also known as intersection over union, which can be interpreted as similarity measure between a finite number of sets. Intersection over union for similarity measure can be defined as following:

$$J(A,B) = \frac{|\, A \cap B \,|}{|\, A \cup B \,|} = \frac{|\, A \cap B \,|}{|\, A \,| + |\, B \,| - |\, A \cap B \,|} \tag{3}$$

$$0 \leq J(A,B) \leq 1$$

The common loss function for classification tasks is categorical cross entropy, however in our case classes are not mutually exclusive and using binary cross entropy makes more sense. The binary cross entropy defined as follows:

$$H = -\frac{1}{n} \sum_{i=1}^{n} \left[y \log(\widehat{y}) + (1-y) \log(1-\widehat{y}) \right] \tag{4}$$

In order to get better results it better that training objective and evaluation metric as close as possible. Hower, the problem is Jaccard Index is not differentiable. One can generalize it for probability prediction, which on one hand, in the limit of the very confident predictions, turns into normal Jaccard and on the other hand is differentiable allowing the usage of it in the algorithms that are optimized with gradient descent (5).

$$J_m(y, \widehat{y}) = \frac{1}{n} \sum_{i=1}^{n} \frac{y_i \cdot \widehat{y_i}}{y_i + \widehat{y_i} - y_i \cdot \widehat{y_i}} \tag{5}$$

Thus, the joint loss function defined as combination of 4 and 5:

$$L = H - \log J_m \tag{6}$$

We used Nadam Optimizer (Adam with Nesterov momentum) [6] and trained the network for 50 epochs with a learning rate of 1e-3 and additional 50 epochs with a learning rate of 1e-4. Each epoch was trained on 400 batches, each batch containing 128 image patches. Each batch was created randomly cropping 112×112 patches from original images. In addition each patch was modified by applying a random transformation from Dih_4 group.

We also tried 224×224 patches but due to limited GPU memory this would significantly reduce the batch size from 128 to 32. Larger batches proved to be more important than a larger receptive field. We believe that was due to the train set containing 25 images only, which differ from one another quite heavily. As a result, we decided to trade-off receptive field size in favour of a larger batch size.

3.5. Boundary effects

During training procedure we prepare patches cropping them from the original images, augment and feed into the neural network. However, the same patching strategy applied during prediction stage lead to square structure in resulted image, as presented in Fig. 5.

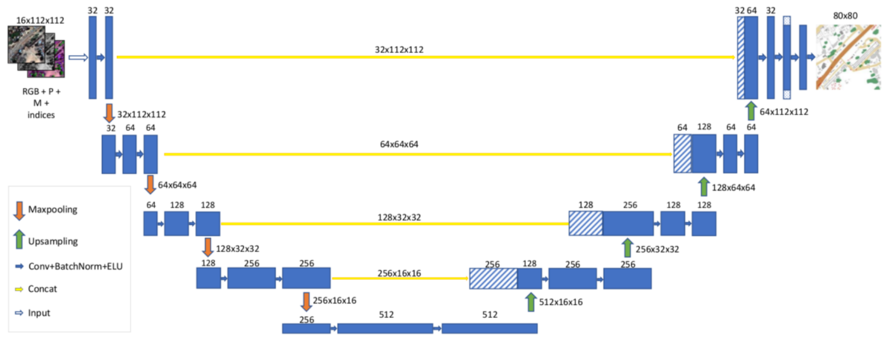

Figure 4: Multispectral U-Net architecture consists of down-sampling and up-sampling parts with skip connections between them. Combinations of multispectral and panchromatic bands with main reflectance indexes combined into single tensor for input to the neural network.

The main reason of such square structure is that not all outputs in the Fully Connected Network are equally good. Number of ways that you can get from any input pixel to the central part of the output in a network is much higher than to the edge ones. As a result prediction quality is decreasing when you move away from center. We checked this hypothesis for a few classes, e.g analysis for building class presented in Fig. 6.

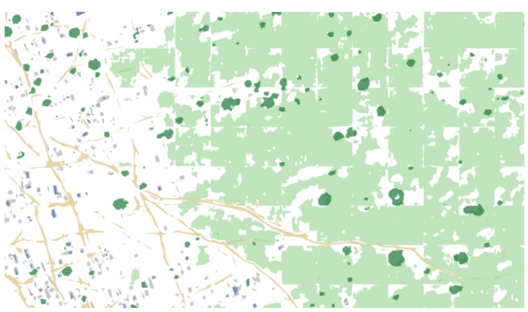

Figure 5: Prediction image formed via patch-based predictions. This lead to boundary effects in a form of square structure near the edge of each patch.

One way to deal with such issue was to make the predictions on overlapping patches, and crop them on the edges, but we came out with a better way. We added cropping layer to the output layers of our networks, which solved two main problem simultaneously:

1. Losses on boundary artefacts were not back-propagated through the network;
2. Edges of the predictions were cropped automatically.

This trick slightly decreased the computation time. To summarize, we trained a separate model for each of the first six classes. Each of them took matrix with a shape

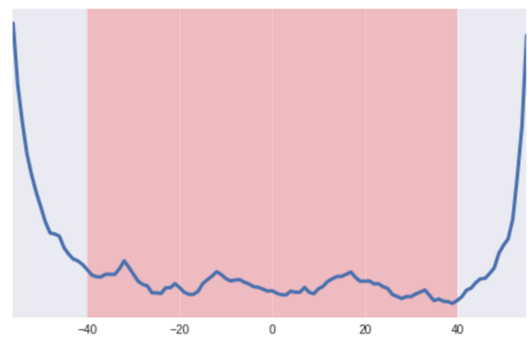

Figure 6: Prediction quality decreasing from the center of the patch to the edges, that showed as a logistic loss with respect to distance from center for building class.

$128 \times 16 \times 112 \times 112$ as an input and returned the mask for a central region of the input images $128 \times 1 \times 80 \times 80$ as an output. Besides, global boundary effects arise during partition original 3600×3600 images into 112×112 tiles due to zero padding. This added some problems at the prediction time. For example, sharp change in pixel values from central to zero padded area was probably interpreted by a network as a wall of the building and as a result we got a layer of building all over the perimeter in the predicted mask. We address this issue with the same trick as in the original U-net paper [3] (Fig. 7).

4. Results

In conclusion, we would like to add that successful approach to above-mentioned problems allows to significantly improve the quality of final models. Our approach includes several steps, such as the adaptation of fully convolutional network to multispectral satellite images with joint training objective and analysis of boundary effects, reflectance indexes. The final results summarized in Table 3.

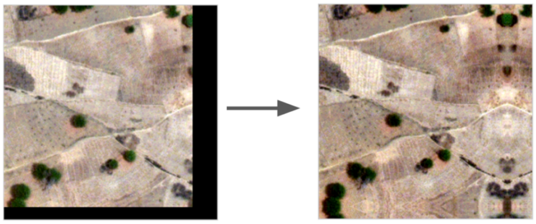

Figure 7: Reflections of the central part to the padded areas helps to prevent global boundary effect.

Class	Public set	Private set
Buildings	0.7453	0.6290
Structures	0.1905	0.2015
Road	0.8005	0.5605
Track	0.3281	0.3965
Trees	0.5018	0.6984
Crops	0.8251	0.8280
Waterway	0.9697	0.9131
Standing water	0.6081	0.5272
Vehicle Large	0.2964	0.0331
Vehicle Small	0.0186	0.00000

Table 3: Segmentation results for different classes in terms of intersection over union.

References

[1] Long, Jonathan, Evan Shelhamer, and Trevor Darrell. "Fully convolutional networks for semantic segmentation." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015. 1

[2] Everingham, Mark, et al. "The pascal visual object classes challenge: A retrospective." International Journal of Computer Vision 111.1 (2015): 98-136 1

[3] Ronneberger, Olaf, Philipp Fischer, and Thomas Brox. "U-net: Convolutional networks for biomedical image segmentation." International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer International Publishing, 2015. 2, 5

[4] Rouse Jr, et al. "Monitoring vegetation systems in the Great Plains with ERTS." (1974). 3

[5] Cammarano, Davide, et al. "Use of the canopy chlorophyl content index (CCCI) for remote estimation of wheat nitrogen content in rainfed environments." Agronomy journal 103.6 (2011): 1597-1603. 3

[6] Dozat, Timothy. Incorporating Nesterov momentum into Adam. Stanford University, Tech. Rep., 2015.[Online]. Available: http://cs229. stanford. edu/proj2015/054 report. pdf, 2015. 4

[7] World View-3 Satellite Sensor Specifications. `http://www.satimagingcorp.com/`

[8] Ioffe, Sergey, and Christian Szegedy. "Batch normalization: Accelerating deep network training by reducing internal covariate shift." arXiv preprint arXiv:1502.03167 (2015). 3

[9] Clevert, Djork-Arn, Thomas Unterthiner, and Sepp Hochreiter. "Fast and accurate deep network learning by exponential linear units (elus)." arXiv preprint arXiv:1511.07289 (2015). 3

[10] Padwick, Chris, et al. "WorldView-2 pan-sharpening." Proceedings of the ASPRS 2010 Annual Conference, San Diego, CA, USA. Vol. 2630. 2010. 2