Loss Function Selection in a Problem of Satellite Image Segmentation Using Convolutional Neural Network

A.G.Sedov, V.V. Khryashchev, R.V. Larionov
P.G. Demidov Yaroslavl State University,
Yaroslavl, Russia
agsedov@gmail.com, v.khryashchev@uniyar.ac.ru,
r.larionov@uniyar.ac.ru

A.A. Ostrovskaya
People's Friendship University of Russia
(RUDN University)
Moscow, Russian Federation,
ostrovskaya-aa@rudn.ru

***Abstract*— Results of training a convolutional neural network for the satellite image segmentation are presented. Input images use four channels: Red, Green, Blue and Near-infrared. The convolutional neural network was trained to mark areas containing buildings and facilities. U-Net architecture was used for the task. For learning procedure supercomputer NVIDIA DGX-1 was used. The process of data augmentation is described. Results of training with different loss functions are compared. Network evaluation results for different types of residential areas are presented.**

Keywords—satellite imagery, image segmentation, data augmentation, loss function, U-Net network architecture

I. INTRODUCTION

Deep learning algorithms became very popular in recent years because of increasing availability of computing power due to GPU technology improvements. Remote-sensing data present some new challenges for deep learning, because satellite image analysisraises unique issues that pose difficult new scientific questions [1-3].

One of the most common challenges related to satellite image analysis is automatic images segmentation. It is an important step of their preliminary processing. Convolutional neural networks are widely used for this task at the present time. One of such neural network architectures is U-Net architecture. U-Net architecture uses skip-connections to join results obtained in different convolutional layers. This convolutional neural networks (CNN) had shown its effectiveness in medical images segmentation [4], SAR [5-6] and various others.

In this paper we show different stages of training a neural network for satellite images segmentation with a goal of labeling building and facilities areas on the original dataset. Results of such segmentation can be later used for marking polygons associated with buildings.

This paper continues research [7-10], dedicated to segmentation of satellite imagery. In [7] there was shown that CNN can be effectively used for satellite image segmentation. Three architectures of CNN were analyzed (U-Net, LinkNet and SegNet) for detecting classes "water resources", "forest" and "agriculture" [8]. Also the optimization of the image mask bypass was performed for better results. U-Net showed the best result for all classes: classification accuracy was 81.7% for "water", 92.3% for "forest" and 96.1% for "agriculture" class [9]. In [10] building detection on aerial images from Planet database with 0.5 m/pixel spatial resolution was performed.

Some challenging requirements for satellite image segmentation algorithms are [5-10]:

- Size and type of classified objects of the same class may vary significantly. This issue can be approached by splitting class either by size or type and training different encoders. Size splitting is preferable, because it can be done automatically, without manual changes in dataset labels.
- High density of targeted areas. For example, two buildings may stand very close to each other. Segmentation algorithm should be penalized for bad separation of objects during training to improve output masksquality. This is achieved by careful selection of the loss function.
- Small size of the training dataset. Success of the machine learning algorithm training heavily relies on quality and quantity of the training dataset. For this paper our goal was achieving best results with original satellite images, so we tried to overcome our dataset limitations. One way of making the most of minimum dataset is data augmentation. Other way is using bigger and better datasets for preliminary training the network, then tuning the CNN on the data from a targeted space.
- Training results should be invariant to rotations, while some objects may be present in training dataset only once. This problem can also be solved by data augmentation.
- Spatial resolution of the satellite images may also vary.

978-1-7281-3238-9/19/$31.00 ©2019 IEEE

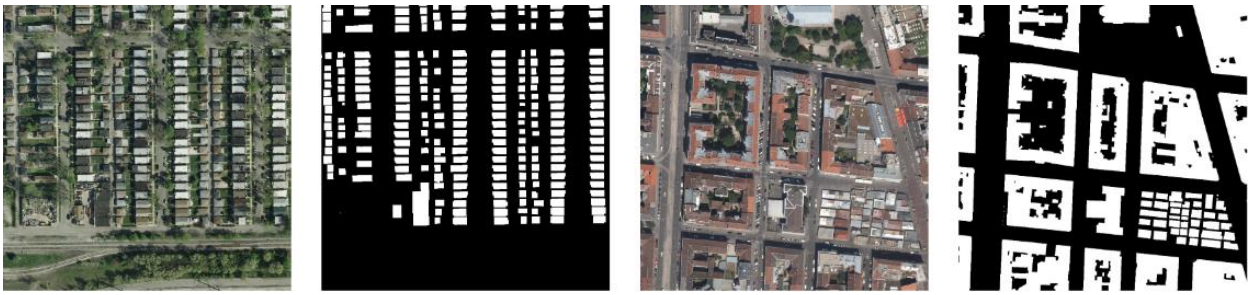

Fig. 1. Examples of Inria dataset images and mask

This paper consists of six parts. In the first part we introduce reader to the problem and overview some existing solutions. In the second part useddatasets are described. The third part describes the network architecture and gives account of loss functions we tested and of their properties. The fourth part features results of numerical experiments. The last part is for conclusions and future work directions.

II. Training data

A neural network was preliminary trained on Inria database [11]. The Inria database consists of 180 color satellite images of 1000 x 1000 pixels. It covers 810 square kilometers and has 0.3 spatial resolution. Every pixel in this database is marked by either “building” or “not a building” label. Inria has images of rural and urban areas from different parts of the world, for example: San Francisco (USA), Chicago (USA), Vienna (Austria), Innsbruck (Austria), Bellingham (USA), Tirol (Austria). Examples of Inria images and masks are shown on the Fig. 1.

Our dataset is made from images of 16 regions with different types of development: from rural to urban areas. Every image has its binary mask made by experts. The dataset covers 25 square kilometers. These images consist of four channels: red, green, blue and near-infrared (NIR) and slightly differ in spatial resolution.

U-net uses 256 x 256 pixel images as inputs. To generate training and validation set, we had cut every image at two non-intersecting stripes. Each stripe was then divided on 256x256 overlapping fragments with 128 pixel step.This kind of division was applied on RGB and NIR channels separately.

Then, to increase the training set, we applied three stages of augmentation on it:

1. Rotations on 90, 180, 270 degrees and reflections. The training set increased 8 times after this stage.

2. Chromatic distortions. Images were translated from RGB color space to HSV color space. Random values were added to HSV coordinates of images. For the NIR channel, instead of chromatic distortions, random values from[-0.06, +0.06] interval were added. (NIR values were normalized in [0, 1] interval).

3.Random shifts, scales and small degree rotations of patches.

After each augmentation step we evaluated segmentation effectiveness on the validation set. We had experimented with different parameters for chromatic distortions and shifts and rotation magnitudes to achieve the best result. We used dice metric to evaluate similarity between predicted images and expert made masks.

III. Network architecture

We used a modification of a popular U-Net architecture for the segmentation task. Original U-Net is made of two parts: encoder and decoder (Fig 2).

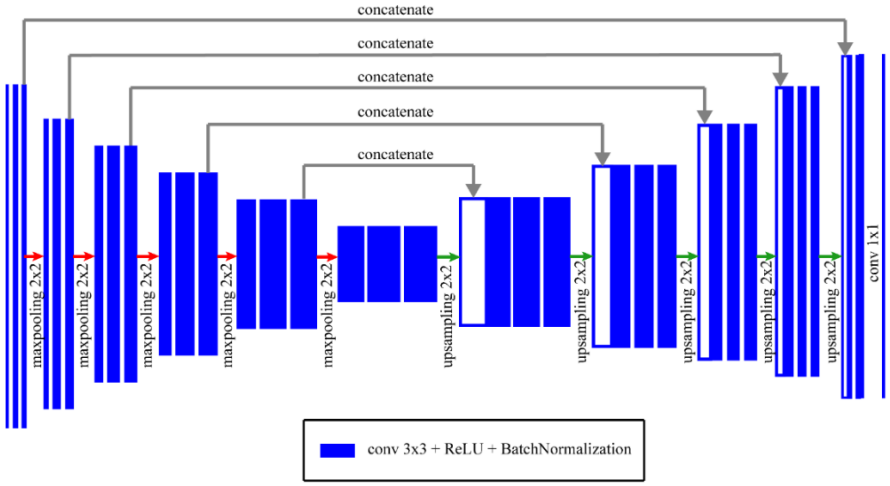

Fig. 2. U-Net architecture

The encoder part is a convolutional neural network of four blocks. Every such block consists of two convolutional layers with 3×3 filters, with ReLU activations and batch normalization applied at each of them, and also of downsampling layer with 2×2 maxpooling. Decoder has the same amount of blocks as encoder. Every decoder block consist of upsampling layer with 2×2 filter and concatenation with encoder features, two 3×3 filter deconvolutional layers with ReLU activation applied to each of them. Last layer uses sigmoid activation to label each pixel.

We modified original U-Net architecture to have two separate encoders, for RGB and NIR components. Outputs of encoders are concatenated before being linked to center and up network layers. Adam optimizer [12] was used for training.

The network was trained using four loss functions for comparison:

Weighted binary cross-entropy with weighted dice coefficient

This loss function was evaluated with the following formula:

$$Loss = WBCE_1(\omega) + (1 - WDICE(\omega)).$$

Here WBCE and WDICE stand for weighted BCE and weighted Dice:

$$WBCE = \sum_{x \in \Omega} w(x)\, log(p_{l(x)}(x)),$$

$$WDICE = \frac{2(\sum_{x\in\Omega} w(x)p(x)t(x))+1}{\sum_{x\in\Omega} w(x)p(x)+\sum_{x\in\Omega} w(x)a_2(x)+1}.$$

Here $p(x) = 1$ if x pixel is predicted to be in the right class and $p(x) = 0$ otherwise, $t(x) = 1$ if the said pixel belongs to specified class according to ground truth.

The weight function $w(x)$ which appears in both expressions above is the only thing that makes them different from original DICE and BCE functions. This function would prioritize pixels near object borders over other pixels.

Binary cross-entropy with original weight map and dice

We modified weighted binary cross entropy formula according to original U-Net paper [4]. To calculate it first the following weight map is computed:

$$w(x) = w_c(x) + w_0 * \exp(\frac{-(d_1(x)+ d_2(x))}{2\sigma^2}),$$

$w_c(x)$ is the weight map to balance the class frequencies.

$d_1(x)$ and $d_2(x)$ denote distances to nearest objects.

w_0 and σ are global parameters.

Then the pixel-wise soft-max is computed as:

$$p_k(x) = exp(a_k(x))/\sum_{m=1}^{2} \exp(a_m(x)).$$

Where $a_k(x)$ denotes the activation in a feature channel k at the position x. Finally the energy function is evaluated using binary cross-entropy:

And loss was calculated as follows:

$$Loss = WBCE_2 + (1 - WDICE).$$

Binary cross-entropy with original weight map

It's same as in the previous case, but without dice:

$$Loss = WBCE_2$$

Lovász-Softmax loss.

We also implemented modified [11] loss function:

$$Loss = 0.8 * WBCE_2 + 0.2 * LH,$$

where LH stands for Lovász Hinge loss [15].

Although dice values didn't much improved, the Lovasz loss trained network shows better results for larger buildings (Fig. 3).

IV. Numerical results

An example of the mask on acquired after the augmentation step is on the Fig 4. During preliminary training dice value of 0.84 was achieved on the Inria dataset for RGB encoder. Improvement of dice values after different stages of augmentation can be seen on the Fig. 5. The improvement varies for different regions.

Results of applying different loss functions to our network is also analyzed. It is established that the ratio of DICE values varies in the range from 0.73 to 0.75.

Fig. 3. Predictions of networks trained on weighted binary cross-entropy loss (left) and Lovász-Softmax loss (right)

Fig. 4. Ground truth mask and prediction after augmentation stages

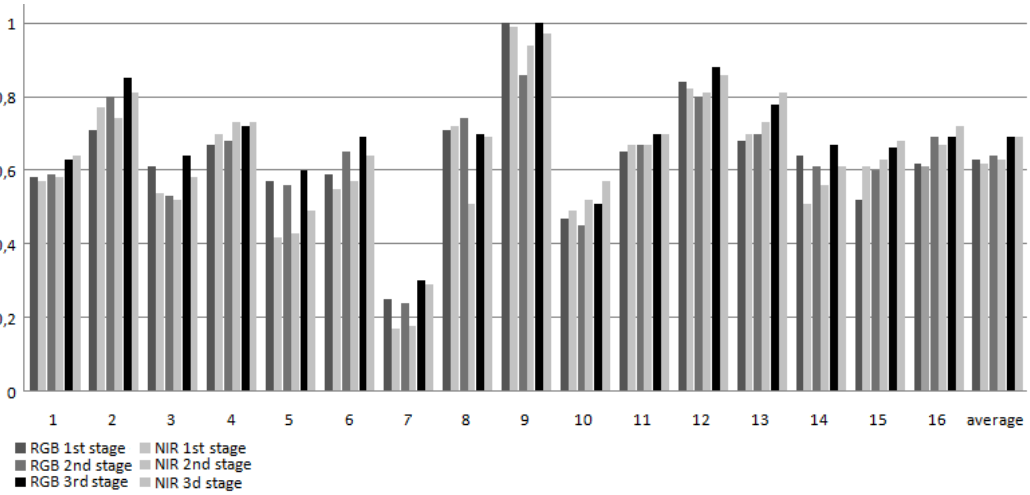

Fig. 5. Dice values for 16 different regions for RGB and NIR after 3 augmentation stages.

V. Conclusions

In the present paper a process of training a neural network for satellite images segmentation were described. Acquired dice metric has a potential to be further improved after concluding more experiments with preliminary training of the network on other datasets and tuning.

Other directions of research may include:

- Cauterization of the buildings based on their area size and training encoders for separate classes.
- Experimenting with preliminary training on other datasets, e.g. Imagenet.
- Experimenting with other methods of joining NIR and RGB channels.

Results of this work may found applications in SAR analysis research, for example, in a task of evaluating level of urbanization of the different regions.

Acknowledgment

The article was prepared with the financial support of the Ministry of Education of the Russian Federation as part of the research project No. 14.575.21.0167 connected with the implementation of applied scientific research on the following topic: «Development of applied solutions for processing and integration of large volumes of diverse operational, retrospective and the thematic data of Earth's remote sensing in the unified geospace using smart digital technologies and artificial intelligence» (identifier RFMEFI57517X0167).

References

[1] Zhang, Liangpei, Lefei Zhang, and Bo Du. “Deep learning for remote sensing data: A technical tutorial on the state of the art.” IEEE Geoscience and Remote Sensing Magazine 4.2 (2016): 22-40.

[2] Y. Chen, H. Jiang, C. Li, X. Jia and P. Ghamisi, “Deep Feature Extraction and Classification of Hyperspectral Images Based on Convolutional Neural Networks,” in IEEE Transactions on Geoscience and Remote Sensing, vol. 54, no. 10, pp. 6232-6251, Oct. 2016.

[3] X. X. Zhu et al., “Deep Learning in Remote Sensing: A Comprehensive Review and List of Resources,” in IEEE Geoscience and Remote Sensing Magazine, vol. 5, no. 4, pp. 8-36, Dec. 2017.

[4] O.Ronneberger, P. Fischer “U-Net: Convolutional Networks for Biomedical Image Segmentation”, arXiv:1505.04597. 2015

[5] Zhang, Z., Liu, Q., & Wang, Y. (2018). “Road extraction by deep residual u-net”. IEEE Geoscience and Remote Sensing Letters, 15(5), pp. 749-753.

[6] Iglovikov, V., Mushinskiy, S., & Osin, V. (2017). “Satellite imagery feature detection using deep convolutional neural network: A kaggle competition”. arXiv preprint arXiv:1706.06169.

[7] V. V. Khryashchev, V. A. Pavlov, A. Priorov and A. A. Ostrovskaya, "Deep Learning for Region Detection in High-Resolution Aerial Images," 2018 IEEE East-West Design & Test Symposium (EWDTS), Kazan, 2018, pp. 1-5.

[8] V.Khryashchev, L.Ivanovsky, V.Pavlov, A.Ostrovskaya, A.Rubtsov Comparison of Different Convolutional Neural Network Architectures for Satellite Image Segmentation, Proceedings of the FRUCT’23, Bologna, Italy, 13–16 November 2018. pp.172–179.

[9] V. V. Khryashchev, A. A. Ostrovskaya, V. A. Pavlov and A. S. Semenov, "Optimization Of Convolutional Neural Network For Object Recognition On Satellite Images," Systems of Signal Synchronization, Generating and Processing in Telecommunications (SYNCHROINFO), Minsk, 2018, pp. 1-5.

[10] L. Ivanovsky, V. Khryashchev, V. Pavlov and A. Ostrovskaya, "Building Detection on Aerial Images Using U-NET Neural Networks," 2019 24th Conference of Open Innovations Association (FRUCT), Moscow, Russia, 2019, pp. 116-122.

[11] E.Maggiori, Y. Tarabalka, G. Charpiat, P. Alliez, “Can Semantic Labeling Methods Generalize to Any City? The Inria Aerial Image Labeling Benchmark”, IEEE International Geoscience and Remote Sensing Symposium (IGARSS). 2017.

[12] D. P. Kingma, J. Ba, “Adam: A Method for Stochastic Optimization”, Web: https://arxiv.org/abs/1412.6980.

[13] S.Golovanov, R. Kurbanov, A. Artamonov, A. Davydow, S. Nikolenko Building Detection From Satellite Imagery Using a Composite Loss Function, The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, June 2018

[14] M.Berman, A. Triki., M. Blaschko. “The lovasz-softmax loss: A tractable surrogate for the optimization of the intersection-over-union measure in neural networks” // arXiv:1705.08790, 2017.

[15] R. Hamaguchi, S. Hikosaka,Building Detection From Satellite Imagery Using Ensemble of Size-Specific Detectors; The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, 2018, pp. 187-191