Land Cover Classification from Satellite Imagery With U-Net and Lovász-Softmax Loss

Alexander Rakhlin
Neuromation OU
Tallinn, 10111 Estonia
rakhlin@neuromation.io

Alex Davydow
Neuromation OU
Tallinn, 10111 Estonia
alexey.davydov@neuromation.io

Sergey Nikolenko
Neuromation OU
Tallinn, 10111 Estonia
snikolenko@neuromation.io

Abstract

The land cover classification task of the DeepGlobe Challenge presents significant obstacles even to state of the art segmentation models due to a small amount of data, incomplete and sometimes incorrect labeling, and highly imbalanced classes. In this work, we show an approach based on the U-Net architecture with the Lovász-Softmax loss that successfully alleviates these problems; we compare several different convolutional architectures for U-Net encoders.

1. Introduction

This work is devoted to a segmentation model that we have developed as part of the DeepGlobe Challenge presented at CVPR 2018 [2]. The DeepGlobe Challenge is designed to advance state of the art techniques for processing satellite imagery, a treasure trove of data that can yield many exciting new applications in the nearest future.

In this work, we introduce a segmentation model for the land cover classification task presented as part of the DeepGlobe Challenge. The main components of our solution include the U-Net architecture commonly used for segmentation, especially under lack of labeled data, and the recently developed Lovász-Softmax loss function specifically designed to optimize the Jaccard index.

2. Related work

Deep learning models, which have revolutionized computer vision over the last decade, have been recently applied to semantic segmentation in aerial and satellite imagery as well. Kampffmeyer *et al.* [10] utilize two different architectures: patch-based classification using 64×64 pixel patches for dense segmentation and pixel-to-pixel segmentation where convolutional layers in the contracting path are followed by a fractional-strided convolutional layer that learns to upsample the prediction back to the original image size. Volpi *et al.* [18] propose full patch labeling by learned upsampling (CNN-FPL), a model architecturally similar to U-Net [14] with the exception that they do not use skip connections. Iglovikov *et al.* [6, 8] follow the classical U-Net architecture with skip connections and more recent improvements like batch normalization and exponential linear unit (ELU) as the primary activation function and use the VGG-11 encoder in the contracting branch. Liu *et al.* [12] propose an hourglass-shaped network (HSN) with residual connections, which is also very similar to the U-Net architecture.

In this work, we also follow the already classical U-Net scheme which has produced state-of-the-art results in many segmentation tasks. The novelty of this work comes from our exploration of different state-of-the-art CNN encoders in the Land Cover Classification task, using the Lovász-Softmax loss [1] for optimizing the Intersection-over-Union (IoU) objective, applying the *equibatch* sampling method, and the Stochastic Weight Averaging (SWA) procedure [9] for training.

3. Dataset and Evaluation Metric

Satellite imagery for the land cover classification task has 50cm pixel resolution and has been collected by a DigitalGlobe's satellite. The training dataset contains 803 satellite images, each of size 2448×2448 pixels, in 24-bit JPEG format. The validation dataset contains 171 satellite images.

Each satellite image in the training set is paired with a mask image for land cover annotation. The mask is an RGB image with $|C| = 7$ classes of labels, using color-coding (R, G, B) as shown in Table 1.

As stated on the competition web site, the labels are far from perfect. Many masks ignore terrain details and contain only 2-3 colors. Incomplete and often inaccurate labelling presented a significant barrier for model development and evaluation; this is intended to bring the competition models more in sync with real life demands.

The evaluation metric for the land cover competition is the pixel-wise mean Intersection over Union (mIoU), also

	Class	Color	Description
1	Urban land	cyan	man-made, built up areas with human artifacts
2	Agriculture land	yellow	farms, any planned (i.e. regular) plantation, cropland, orchards, vineyards, nurseries, and ornamental horticultural areas; confined feeding operations
3	Rangeland	magenta	any non-forest, non-farm, green land, grass
4	Forest land	green	any land with tree crown density plus clearcuts
5	Water	blue	rivers, oceans, lakes, wetland, ponds
6	Barren land	white	mountain, land, rock, dessert, beach, no vegetation
0	Unknown	black	clouds and others

Table 1. Descriptions of the seven classes in the dataset.

known as Jaccard Index:

$$\mathrm{mIoU} = \frac{1}{6}\sum_{c=1}^{6} \mathrm{IoU}_c,\ \mathrm{IoU}_c = \frac{\mathrm{TP}_c}{\mathrm{TP}_c + \mathrm{FP}_c + \mathrm{FN}_c}, \quad (1)$$

where TP_c is the number of true positive pixels in class $c \in C$ across the entire data set; FP_c, number of false positive pixels in c; FN_c, number of false negative pixels in c. mIoU is computed by averaging over all classes except the "Unknown" class which is not used in evaluation.

4. Methods

4.1. Model architecture and loss function

As a core approach for multi-class segmentation, we have implemented the U-Net architecture [14] that has proven its efficiency in many segmentation problems with limited amount of data, including medical and satellite imaginary tasks [6, 15]. Figure 1 shows a typical U-Net architecture that consists of a contracting branch to capture the context and an expanding branch that enables precise localization for the segmentation masks. The contracting branch implements a standard convolutional architecture with alternating convolution and pooling operations and progressively downsampled feature maps. Every step in the expansive path performs upsampling of the current feature map followed by a convolution, thus gradually increasing the resolution of the output. In order to localize upsampled features, the expansive branch combines them with high-resolution features from the contracting branch via skip-connections [14]. The output of the model is a 2-dimensional softmax which assigns each pixel probability to belong to each of the 7 classes.

We have evaluated various convolutional encoders in the contracting branch of U-Net: a VGG-based [16] custom architecture *m46* previously developed for medical imaging [7], *Resnet-34* [5], and *Inception Resnet V2* [17]. In Resnet architectures, we introduced small but useful modifications: ELU activations instead of ReLU and reversed order of batch normalization and activation layers as proposed in [13]. We used the He normal weight initialization [4].

It is known that the categorical cross entropy CCE, while convenient to train neural networks, does not directly translate into mIoU. Hence, as loss functions we used

$$L(\mathbf{w}) = (1-\alpha)\mathrm{CCE}(\mathbf{w}) - \alpha L'(\mathbf{w}), \quad (2)$$

a weighted sum of CCE and another loss L', comparing two different variants of L': soft Jaccard loss J

$$J = \frac{1}{6n}\sum_{c=1}^{6}\sum_{p=1}^{n}\left(\frac{y_p^c \hat{y}_p^c}{y_p^c + \hat{y}_p^c - y_p^c \hat{y}_p^c}\right), \quad (3)$$

where y_p^c is binary label for the class c for the pixel p. $\hat{y}_p^c$ is predicted probability of the class c for the pixel p, and n is the number of pixels in the batch, and the *Lovász-Softmax loss* LSL, a tractable surrogate for optimizing IoU [1].

4.2. Preprocessing, training and mask generation

We preprocessed the input by rescaling 8-bit data $[0...255]$ into floating point numbers from $[-1, 1]$, downscaled the image size by a factor of 2 or 4, cropped 288×288 patches from downscaled satellite images and corresponding masks, and applied random color, gamma, and geometrical augmentations. In some experiments we applied local contrast normalization to the images; however, this did not boost performance. We evaluated 4 scales: 1:1, 1:2, 1:4, and 1:8; 1:2 was best, as this resolution provides the best tradeoff between image resolution and receptive fields and depth of the model. Throughout the work we use the batch size of 8. We implemented two schemes of sampling the patches: (i) sampling every image randomly; (ii) the *equibatch* method that accounts for class imbalances: since mIoU gives equal importance to every class, we sample patches from the training set by cycling over the classes so that each class is visited at least once every $|C|$ patches; this approach was proposed in [1].

For training the models we implemented the recently proposed Stochastic Weight Averaging (SWA) procedure [9] that finds much broader optima than stochastic gradient descent (SGD) by approximating the Fast Geometric Ensembling (FGE) approach [3] with a single model. SWA is based on averaging the weights proposed by SGD using exploration of the region in the weight space corresponding to high-performing networks. We begin the SWA procedure by conventional training with the Adam optimizer [11] for 100 epochs, starting with learning rate of 0.001 and decreasing it to 0.0001. Starting from epoch 101 we turn on the cyclical learning rate schedule adopted from [3]. In each cycle we linearly decrease the learning rate from 0.0001 to 0.00001. We use cycle length of 6 epochs and complete 10 full cycles averaging model weights obtained in the end

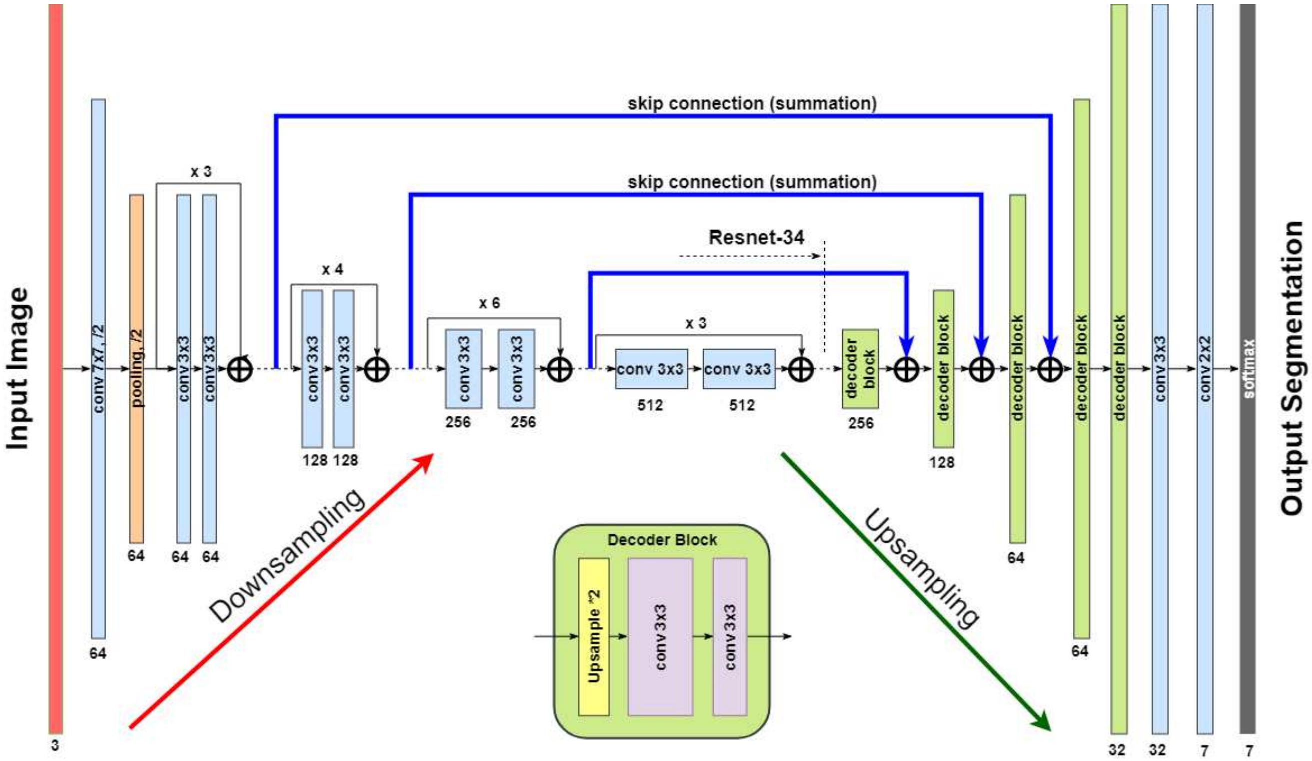

Figure 1. U-Net architecture with a ResNet-34 encoder.

of each cycle. In our experiments, SWA steadily increases mIoU of pre-trained models by 0.01-0.025; see Fig. 2 for the plot of numerical results.

We predict the segmentation mask as follows: (i) downscale and normalize an image as above, (ii) crop it into multiple tiles of 288×288 pixels with a stride of $\left\lfloor \frac{288}{5} \right\rfloor = 57$ pixels, (iii) predict the tiles, (iv) assemble them into a grid and average the predictions, (v) upsample predicted mask to original size 2448×2448, (vi) assign the class of every pixel as $\arg\max$ of the 7 class probability scores. Finally, we apply morphological postprocessing to remove small components with area < 3500 pixels (found with cross-validation). This improves the score only slightly but significantly enhances the visual appearance.

5. Results

Tables 2 and 3 show our results. In Table 2 we show the results of the two sampling methods, unbalanced and the *equibatch* method, which is evidently better across all classes. Note that the largest performance improvement with *equibatch* is for the 3rd class (“rangeland”), which proved to be the most difficult to detect in this challenge.

Performance of different encoders and loss functions on the local validation set are compared in Table 3. There is no distinct leader, all mIoU scores are sufficiently close. In particular, our smallest model, U-Net with the *m46* encoder trained with Jaccard loss, performs even slightly better than big models despite much fewer parameters. With the Lovász-Softmax loss, Resnet-34 outperforms other models

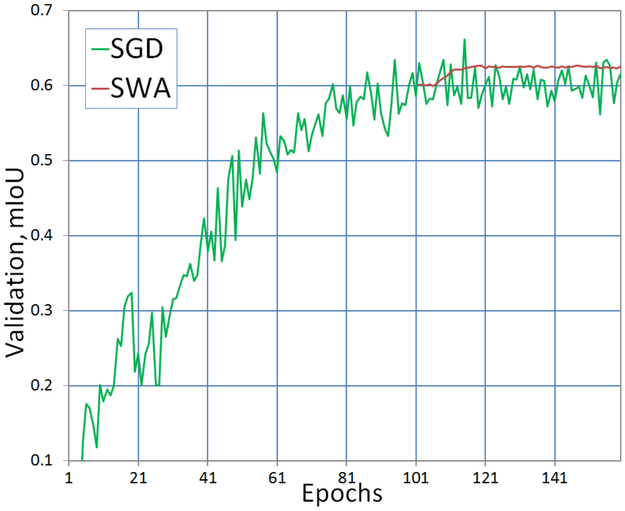

Figure 2. Validation mIoU as a function of training epoch for decaying (green) learning rate schedule. In red we average the points along the trajectory of SGD with cyclical learning rate starting at epoch 101.

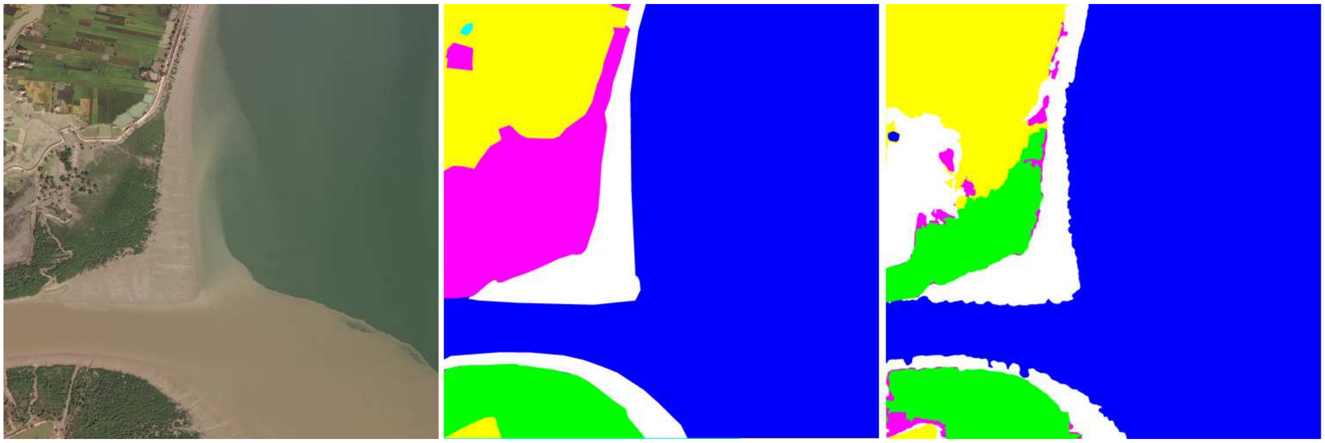

Figure 3. Sample segmentation results, left to right: original image, ground truth, predicted mask.

Class	**1**	**2**	**3**	**4**	**5**	**6**	**0**
Equibatch	0.76	0.84	0.29	0.79	0.63	0.53	0.07
Unbalanced	0.74	0.79	0.02	0.75	0.52	0.41	0.02

Table 2. mIoU on the local validation set for 2 sampling methods: Equibatch vs. unbalanced sampling.

U-Net encoder	# params	Jaccard	Lovasz
m46	1M	**0.624**	0.619
Resnet-34	25M	0.615	**0.641**
Inception Resnet v2	61M	0.604	0.573

Table 3. mIoU on local validation for different encoders and loss functions, see eq. 2, 3.

but, again, not by much. Inferior performance of *Inception Resnet v2*, the largest model in our comparison with 60M parameters, can be explained by incomplete convergence on noisy labels. We believe that similar performance of such a different encoders and loss functions suggests that the limiting factors in this contest were mostly related to data labeling rather than model architecture or loss functions. For the final ensemble we selected the 5 best performing models: two based on *m46* and three on *Resnet-34*; two of the models were trained with the Lovász-Softmax loss, and three with Jaccard loss. The *mIoU* of this ensemble is 0.648, improved to 0.649 by morphological postprocessing.

Figure 3 shows a sample segmentation result of our model together with the ground truth segmentation mask. We see that the segmentation quality is good enough, but there is uncertainty between agriculture land (yellow), rangeland (magenta), and forest (green); note that the model correctly identified barren land (white) in the left part of the image, while the labeler marked it as rangeland (magenta). These observations are also supported by the confusion matrix between classes for our best model, shown on Figure 4.

6. Conclusions

In this work, we have presented an approach to land cover classification for satellite imagery based on the standard U-Net architecture. In our opinion, the features of our solution that most significantly contributed to the overall segmentation quality include: (i) the *m46* encoder architecture designed to overcome lack of data, (ii) the *equibatch* sampling method that helps combat class imbalances, and, most importantly, (iii) the Lovász-Softmax loss function specifically designed to optimize IoU-based metrics. We believe that the very recently developed Lovász-Softmax loss [1] will play an important role in state of the art segmentation models, and we view our solution as a step in this direction.

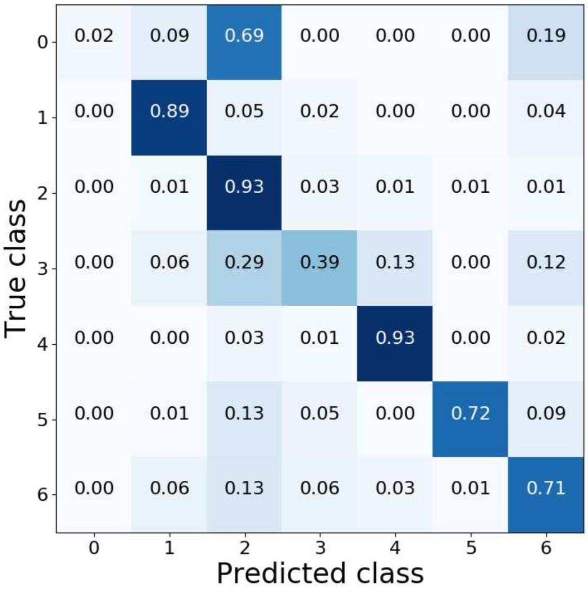

Figure 4. Normalized confusion matrix.

References

[1] M. Berman, A. Rannen Ep Triki, and M. Blaschko. The lovász-softmax loss: A tractable surrogate for the optimization of the intersection-over-union measure in neural networks. 2018.

[2] I. Demir, K. Koperski, D. Lindenbaum, G. Pang, J. Huang, S. Basu, F. Hughes, D. Tuia, and R. Raskar. Deepglobe 2018: A challenge to parse the earth through satellite images. *arXiv preprint arXiv:1805.06561*, 2018.

[3] T. Garipov, P. Izmailov, D. Podoprikhin, D. P. Vetrov, and A. G. Wilson. Loss surfaces, mode connectivity, and fast ensembling of dnns. *arXiv preprint arXiv:1802.10026*, 2018.

[4] K. He, X. Zhang, S. Ren, and J. Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pages 1026–1034, 2015.

[5] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[6] V. Iglovikov, S. Mushinskiy, and V. Osin. Satellite imagery feature detection using deep convolutional neural network: A kaggle competition. *arXiv preprint arXiv:1706.06169*, 2017.

[7] V. Iglovikov, A. Rakhlin, A. Kalinin, and A. Shvets. Pediatric bone age assessment using deep convolutional neural networks. *arXiv preprint arXiv:1712.05053*, 2017.

[8] V. Iglovikov and A. Shvets. Ternausnet: U-net with vgg11 encoder pre-trained on imagenet for image segmentation. *arXiv preprint arXiv:1801.05746*, 2018.

[9] P. Izmailov, D. Podoprikhin, T. Garipov, D. Vetrov, and A. G. Wilson. Averaging weights leads to wider optima and better generalization. *arXiv preprint arXiv:1803.05407*, 2018.

[10] M. Kampffmeyer, A.-B. Salberg, and R. Jenssen. Semantic segmentation of small objects and modeling of uncertainty in urban remote sensing images using deep convolutional neural networks. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2016 IEEE Conference on*, pages 680–688. IEEE, 2016.

[11] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[12] Y. Liu, D. Minh Nguyen, N. Deligiannis, W. Ding, and A. Munteanu. Hourglass-shapenetwork based semantic segmentation for high resolution aerial imagery. *Remote Sensing*, 9(6):522, 2017.

[13] D. Mishkin, N. Sergievskiy, and J. Matas. Systematic evaluation of convolution neural network advances on the imagenet. *Computer Vision and Image Understanding*, 2017.

[14] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[15] A. Shvets, V. Iglovikov, A. Rakhlin, and A. A. Kalinin. Angiodysplasia detection and localization using deep convolutional neural networks. *arXiv preprint arXiv:arXiv:1804.08024*, 2018.

[16] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[17] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In *AAAI*, volume 4, page 12, 2017.

[18] M. Volpi and D. Tuia. Dense semantic labeling of subdecimeter resolution images with convolutional neural networks. *IEEE Transactions on Geoscience and Remote Sensing*, 55(2):881–893, 2017.