Leprosy of the Land: using machine learning to search for places with illegal amber mining on satellite images

Anatoly Bondarenko, Texty

January 2019

Abstract

Increased world prices for amber led to illegal mining for this gemstone in North Western part of Ukraine. The total area of region with close-to-surface deposits of amber is about 70,000 km^2. Number of images with good enough resolution needed to cover this area is about 450,000. To process such amount of information we developed XGBoost and ResNet based classifier, trained it with initial ground-truth images with traces of amber mining and created most complete interactive map of this phenomena as of time of publishing in March, 2018 [1].

1. Background

Around 2010 world prices for amber started to surge. Due to this in 2012 demand was so high that north-western part of Ukraine became place of "amber rush" and "new Wild West". Thousands of prospectors start to search for gems with shovels and later with water pumps. Hundreds of hectares in forests / agricultural land became a desert, a lifeless moon landscape. 2014-2016 were most intense years of illegal mining, but it's still going till now.

We decided to estimate, for the first time, the scale of environmental impact from this phenomenon and to find places most suffering from ecological, social and criminal consequences from such an activity.

2. Project stages

First of all, we researched how traces of illegal mining could look on satellite images. For this we compiled some known locations from previous reporting about this topic, used known videos and photos, and made interviews with field experts. Basically, it's just a pits and holes in the ground with white and sand coloured features. (See Fig. 1)

Next we found which known map providers have relatively recent satellite images with good resolution. We decided to use satellite images from Bing Maps[2], by Microsoft, because it has an API with generous limits for image download and has a date information for each specific photo. Due to small characteristic size of dug holes, we needed resolution below 1 m^2 per pixel.

Table 1: Time distribution of satellite images, by year

Year	Number
2011	1933
2012	4669
2014	117059
2015	271893
2016	55403

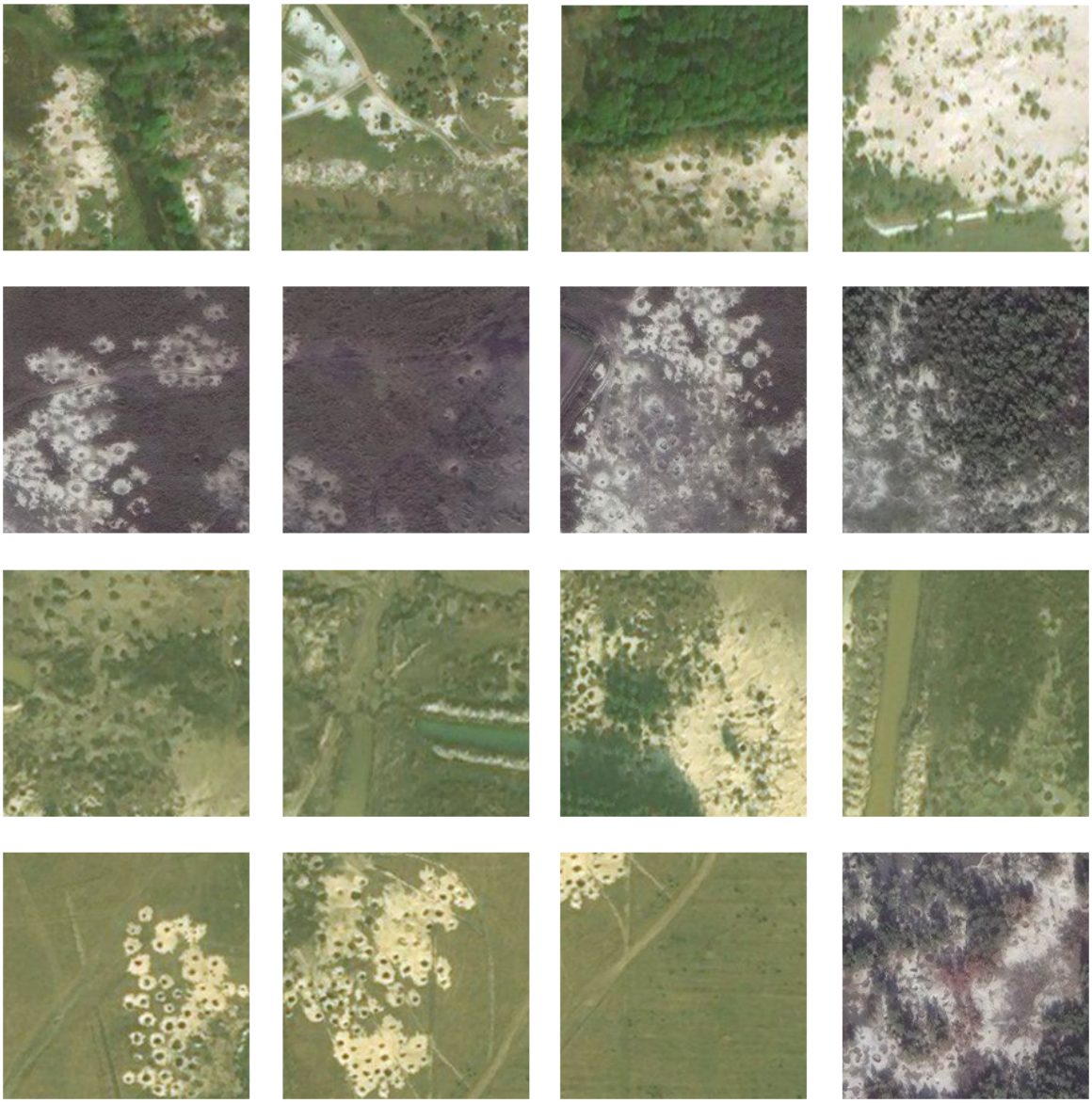

Figure 1: Places with amber mining

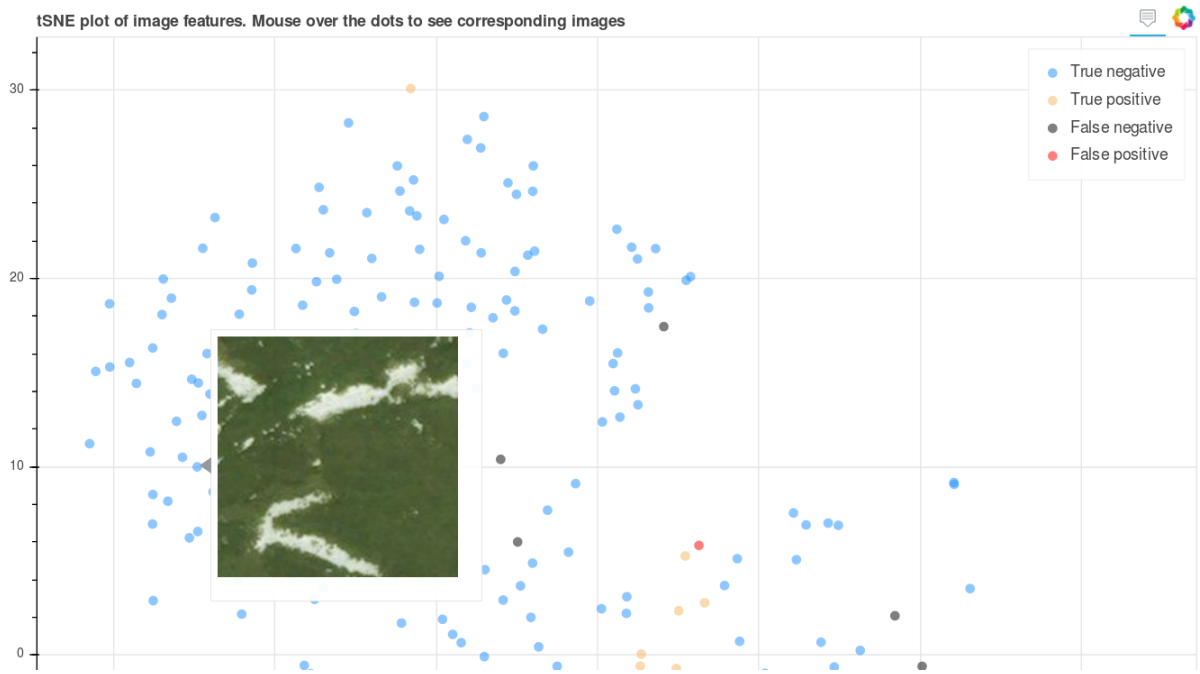

Figure 2: Screenshot from interactive scatterplot for visual debugging

We manually found and compiled initial set of coordinates for images with traces of mining. First such places were detected with a huge help from participants of Open Data Day Kyiv, in March 2018.

We split each image(tile) to superpixels, or segments with approximately homogeneous visual appearance, with a simple linear iterative clustering [3]. Next we programmed one-click labeller for images (single page web-app in javascript) and labelled each segment either "with amber mining" or "without mining". In such a way for a couple of hours we obtained initial training set of about 1000 such segments.

We used transfer learning approach and extracted features for each segment — it's a vector of dimension 2048 for each image — with a help of ready to use, vanilla ResNet50 [4] deep learning model from Keras library [5].

Next we created machine model to classify superpixels. After several attempts XGBoost classifier [6] was selected due to best performance (we used SVM as a baseline model and RandomForest as alternative to XGBoost). With initial training set, our classifier achieved f1 score of 0.876 (with recall = 0.80).

3. Visual examination of the model's performance

Initial training set later was augmented with a help of simple visual method. Specifically, we made dimensionality reduction of ResNet features for images by tSNE[7] or even better, UMAP method[6], and made a Jupiter notebook with interactive scatterplot: dots as 2d-projections of each image from validation set, and tool tips - as a corresponding images for each dot. See Fig. 2.

Points on the chart coloured by four categories, corresponding to entries of confusion matrix. So one could immediately spot coloured outlier, and look at the image for this dot. Even better, we could use nearest neighbours algorithm[8], and augment training set with a new examples - adding more similar images as misclassified one, into the training dataset. This visual method led to biggest increasing in performance of the model. The performance of production version of model became, in numbers: f1 = 0.91, recall = 0.88, precision = 0.95.

After we trained the model, we used it to classify each superpixel/segment from all images (from region

Figure 3: On the left: all superpixels from image, on the right - segments classified by model as areas with amber mining

of interest). If classifier detected more then 2 superpixels as positive for amber mining, our script marked current tile to place its coordinates on resulting map. See Fig.3 as an example of positive image.

We processed approximately 450,000 images, with 1000x1000 size, from region with total area about 70,000 km^2. Total computation time was about 100 hours on one computer with two GeForce GTX 960 graphic cards.

Results

We created most detailed, as for this moment, interactive map of impact on environment due to illegal amber mining in Ukraine[1]. This approach could be useful to detect other types of interesting places from online satellite images. For reproducibility and future use by others, we published detailed methodology, with code examples, on Texty.org.ua's github page[9].

Acknowledgements

Besides machine modeling, this project could not be possible without a huge contribution from many members of Texty.org.ua team. Vlad Herasimenko did javascript programming, Nadia Kelm created design, Yevheniia Drozdova made web-development and Yaroslava Tymoshchuk wrote a text for this interactive feature.

The very inspiring example of Terrapattern[11] proved to author back in 2016 that such analysis of satellite images is quite possible.

References

[1] Leprosy of the Land, http://texty.org.ua/d/2018/amber_eng

[2] Bing Maps, https://www.bing.com/maps

[3] SLIC algorithm, http://www.kev-smith.com/papers/SLIC_Superpixels.pdf

[4] ResNet. Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun Deep Residual Learning for Image Recognition arXiv:1512.03385, 2015

[5] François Chollet and others, Keras Library, https://github.com/fchollet/keras, 2015

[6] UMAP: Uniform Manifold Approximation and Projection, McInnes, Leland and Healy, John and Saul, Nathaniel and Grossberger, Lukas https://github.com/lmcinnes/umap

[7] Visualizing Data using t-SNE, Laurens van der Maaten, Geoffrey Hinton Journal of Machine Learning Research, 9 (2008) 2579-2605

[8] Nearest neighbours algorithm, https://scikit-learn.org/stable/modules/neighbors.html

[9] Methodology for the "Leprosy of the Land" project, https://github.com/texty/amber-methodology/

[10] Terrapattern. Golan Levin, David Newbury, Kyle McDonald et al http://www.terrapattern.com