# Interpreting Black-Box Semantic Segmentation Models in Remote Sensing Applications

**Abstract**

In this paper we are approaching problem of understanding black-box model predictions. In the interpretability literature attention is focused on understanding black-box classifiers but many problems ranging from medicine through agriculture and crisis response in humanitarian aid are being tackled by semantic segmentation models. The absence of interpretability aspect for this canonical problem in computer vision is a motivation for this study. Performance measures of black-box are simply not enough to meet complex requirements of justification model predictions. In this study we present user-centric approach based on combining interpretability with interactive visualizations. We have applied our method to deep learning model U-Net for semantic segmentation in remote sensing application of building detection. This application if of high interest for humanitarian crisis response teams that rely on satellite images analysis. Preliminary results shows utility in understanding semantic segmentation models.

## 1 Introduction

The possibility of exploring characteristics of models beyond accuracy is becoming a legal demand in business applications under the light of recently introduced laws, including the European GDPR and the "right to explanation" of decisions made by algorithms [1]. Machine Learning is often introduced as an oracle, rather than a scientifically explainable approach, and this is cause for concern. Also relying on visualizations of neuron activations is not enough – people need interpretations. How do models link to the underlying datasets on which they were trained? How can we use this knowledge to open the black-box and discover reasoning behind the model?

Being able to tell what properties of the data the model is best at is the case for designing more transparent models with honest description of the one that were already trained.

While interpretability in machine learning can be realized in many ways, the focus of this work is the problem of explaining black-box models, as defined in [2]. While there is substantial research on explanation of black box image classifiers, less is available for image segmentation. The question we explore in this study is how can we explain predicted segmentations by inspecting their learned representations and navigating the associated latent space.

One of the problems of remote sensing is segmentation of different elements of satellite images e.g. roads, bridges, buildings, cars, land coverage etc. Information about detected

buildings is being used for example to estimate population of the region. This knowledge guides humanitarian mission efforts in distribution of food, water and other basic resources for people affected by the crisis; creating strategies for epidemiology prevention. According to [3], the need for interpretability originates from an incompleteness of the problem definition witch makes it difficult to optimize and evaluate. To understand this in the context of remote sensing one needs to understand the user's perspective, as one of the questions about interpretability is to whom it should be interpretable [4]?

Incompleteness in remote sensing may manifest itself in different ways. Domain knowledge is one - resources for inference are often limited in humanitarian remote sensing applications, which may guide model choice. Another aspect is safety and reliability - we are not able to flag all undesired outputs for end-to-end system, it will never be fully testable. Finally, ethics are an important consideration - every model is biased by the data it was trained on and by the model of the world used to annotate data. For example, main street in Chicago and in Niger State have different visual representations, although they fulfill similar roles. Incompleteness may also be associated with mismatched objectives or multi-objective trade-offs like privacy vs quality.

Therefore, in the presence of incompleteness, explanations can ensure that underspecifications of formalization are visible and understood by users [3].

In the remote sensing scenario interpretability could highlight:

- biases from the training set (e.g. a model trained on cities should not be used in rural areas)
- more honest information about the characteristics of data and their effect on model performance, so that users can set their expectations
- techniques to guide sample collection (e.g. how target areas differ from the areas that was covered in the training set)
- the importance of the underlying data to a wider audience (e.g. one might think that the model should work for every city in the world in the case of building detection task, which might be disappointing and can undermine trust towards usage machine learning at all)

# 2 Method

This study presents an interactive visualization method for highlighting model capabilities. It is based on linked brushing and IoU smoothing to interact with latent representations from an encoder-decoder segmentation model.

Method was designed for explaining U-Net semantic segmentation model [5] in the future we also plan to explore other deep networks with encoder-decoder architecture. It requires access to trained U-Net segmenter model, training dataset and activations at the bottleneck layer. To evaluate segmentation prediction with respect to ground truth we used Intersection over Union score (IOU) defined as $IoU(y, \hat{y}) = \frac{overlapping\ area}{union\ area}$.

U-Net network is a deep network that is able to reduce representation of image through down-sampling path and in the same time preserve localized infromation about desired properties through up-sampling path in order to make a prediction.

Each component is composed of convolutional layers going down and transposed convolutions going up with max-pooling layers in between. Down-sampling is responsible for reducing the input image to a concise representation, while up-sampling retrieves localized information for the network's output. Latent representation refered in this work is represented as an activations of the bottleneck layer - the layer that contains the quintessence of analyzed image.

The motivation behind our approach is to guide users in understanding and successfully applying the model to the considered task. Our visualization is based on principle of linked brushing [6], allowing users to interactively explore coordinated views across subsets of the data [7].

Our visualization is obtained through the following steps. Firstly activations from the bottleneck of the network are collected for the training dataset. Next step is to associate IOU score and set of activations with image (patch) and predictions (ground truth and inference). Once the activations are collected, the dimensionality is reduced by PCA and two principal components are used to plot points on the scatterplot. Each point is a representation of an input image. Given set of reduced activations and associated IOU scores with them we train a MLP predictor to estimate IOU score in the whole training space. This prediction will be visualized as a heatmap of IOU scores plotted as a background of scatterplot. The last step is applying brushing to the plot in order to associate the latent views with the original samples. For each point contained in the brush selection, we display the corresponding ground truth and prediction mask. This provides a convenient view of the dataset and model properties.

In the following section we present a proof of concept designed for the task of interpreting building detection models in remote sensing.

## 2.1 Application

### 2.1.1 Dataset

We applied this method to Inria Aerial Labelling Dataset as it is an example of well explored labeled dataset for satellite imagery. The training set contains 180 color image tiles of size 5000 x 5000, covering a surface of 1500 m x 1500 m each (at a 30 cm resolution). There are 36 tiles for each region. It covers 5 region:s Austin, Chicago, Kitsap County, Western Tyrol, Vienna. For the test set there were another 5 regions chosen: Bellingham, WA; Bloomington, IN; Innsbruck; San Francisco; Eastern Tyrol. It provides all together coverage of 810 $km^2$.

We analyzed trained U-Net model optimized with Adam algorithm with batch normalization. It scored overall IOU of 72.55 and accuracy of 95.91 on validation set.

One potential application of this method is an extension of the humanitarian aid application MapSwipe, a tool for pre-screening satellite images in region of interest, filtering only to those that have desired features present. For example, in a building detection task, this allows volunteers to filter away tiles that do not contain buildings, but rather forest or sea. With our method we can instantly tell which patches do not contain anything to tag. We can see that in the Figure 1.

Red region in the Figure 1 is an artifact of how IoU is defined, if in the image there is nothing to be detected there is no union between detections and formula of IOU does not make sense we assumed that in such situation the IOU score will be 0. This area is also

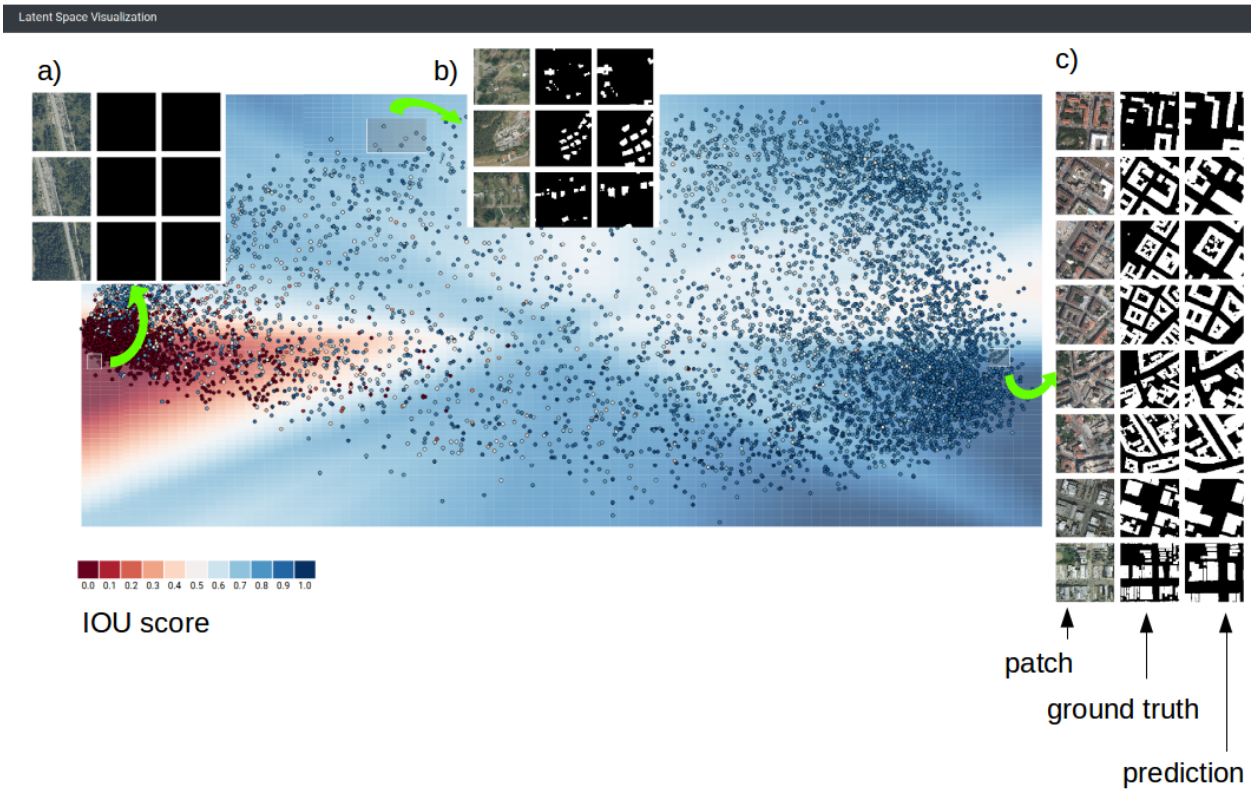



Figure 1: Demo visualization with several selections merged together. Selections present samples from three qualitatively distinguishable regions a) that does not contain any buildings b) that contains few buildings c) highly urbanized with many buildings and with higher IOU score. We can see a bipolar nature of learned representation: undeveloped area and urbanized.

highly condensed and qualitatively we can seet that it contains mostly images of undeveloped area without any buildings.

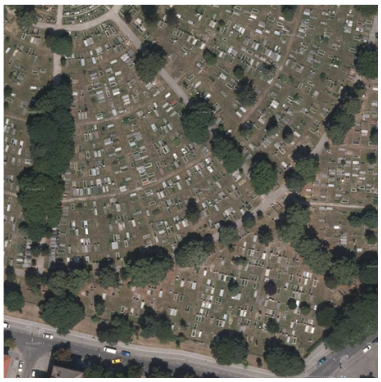
We explored clustering methods on the latent space and selecting representatives associated with clusters. We used KMeans algorithm and DBSCAN, after comparing Silhouette scores of several parameters configurations, better clustering was obtained with KMeans algorithm with 14 classes. For each cluster we looked at the representatives and we also explored their median and mean IOU scores.

Exploring clusters lead us to some peculiar discovery about given model. According to U-Net cementeries and car parkings are similar. Why? Probably because of similar pattern of rectangular shaped objects positioned next to each other. The question is if it is a desirable generalization for a given task?

Some undesired outputs may originate from:

- poor generalization capabilities for specific type of data

(a) Car parking



(b) Cemetery

Figure 2: Generalization

- ground-truth errors
- definition of error metric

# 3 Conclusion

Users tend to not trust the models which they dont understand. This is not surprising since models really are genuinely complicated structures. If users dont trust models, they dont use them. If a crisis response team does not trust AI predictions, then we are not using the full potential of current technology, meaning that the help offered to people affected by crises is not best that could be offered. To address this problem, we propose the usage of interpretable approaches - focus on the end-user, the decision maker working in limited resources and time critical environment for introducing machine learning to current humanitarian workflows.

Are there any distinguishable clusters of outliers? Can we find the reason why models make errors? Are the errors consistent? What are the most common errors? What is the generalization capability of the model? To what extent can you trust the model in a new region? Those are only a handful of questions that are of interest not only for practitioners but for scientists, and we believe that approach focused primarily on interpretability could shed a new light on those questions.

With this work, we emphasize the importance of interpretability and explore its utility in remote sensing analysis in the context of humanitarian AI, enhancing tools that are already used by community. We also presented a method of visualization of a segmentation model along with its training data and and describe a latent space view that we believe will be useful for estimating IOU score or error of new, unseen data.

# References

[1] Goodman B, Flaxman S. European Union regulations on algorithmic decision-making and a "right to explanation". arXiv e-prints. 2016 Jun;p. arXiv:1606.08813.

[2] Guidotti R, Monreale A, Ruggieri S, Turini F, Pedreschi D, Giannotti F. A Survey Of Methods For Explaining Black Box Models. arXiv:180201933 [cs]. 2018 Feb;ArXiv: 1802.01933. Available from: http://arxiv.org/abs/1802.01933.

[3] Doshi-Velez F, Kim B. Towards A Rigorous Science of Interpretable Machine Learning;Available from: http://arxiv.org/abs/1702.08608.

[4] Tomsett R, Braines D, Harborne D, Preece A, Chakraborty S. Interpretable to Whom? A Role-based Model for Analyzing Interpretable Machine Learning Systems. arXiv:180607552 [cs]. 2018 Jun;ArXiv: 1806.07552. Available from: http://arxiv.org/abs/1806.07552.

[5] Ronneberger O, Fischer P, Brox T. U-Net: Convolutional Networks for Biomedical Image Segmentation. arXiv:150504597 [cs]. 2015 May;ArXiv: 1505.04597. Available from: http://arxiv.org/abs/1505.04597.

[6] Spence R. Information Visualization: Design for Interaction. 2nd ed. Pearson;.

[7] Keim DA. Information visualization and visual data mining. IEEE Transactions on Visualization and Computer Graphics. 2002 Jan;8(1):1–8.