MLVis: Machine Learning Methods in Visualisation for Big Data (2019)
D. Archambault, I. Nabney and J. Peltonen (Editors)

Interpreting Black-Box Semantic Segmentation Models in Remote Sensing Applications

A. Janik[1] K. Sankaran[1] A. Ortiz[2]

[1]Montreal Institute of Learning Algorithms
[2]University of Texas - El Paso

Abstract

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 tackled by semantic segmentation models. The absence of interpretability for these canonical problems in computer vision motivates this study. In this study we present a user-centric approach that blends techniques from interpretability, representation learning, and interactive visualization. It allows to visualize and link latent representation to real data instances as well as qualitatively assess strength of predictions. We have applied our method to a deep learning model for semantic segmentation, U-Net, in a remote sensing application of building detection. This application is of high interest for humanitarian crisis response teams that rely on satellite images analysis. Preliminary results shows utility in understanding semantic segmentation models, demo presenting the idea is available online.

CCS Concepts

• ***Human-centered computing*** → ***Information visualization;*** • ***Computing methodologies*** → ***Knowledge representation and reasoning; Image segmentation;***

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 [GF16]. Machine Learning is often introduced as an oracle, rather than a scientifically explainable approach, and this is cause for concern. 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 the reasoning behind the model?

Being able to tell what properties of the data result in good model performance is useful for designing them in a more transparent way, and with more honest certifications of when they can be deployed reliably.

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 [GMR*18]. While there is substantial research on explanation of black box image classifiers [RSG16, KWG*17], 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 region populations. This knowledge guides humanitarian efforts in distribution of food, water and other basic resources for people affected by the crisis, and for creating strategies for epidemiology prevention.

1.1. Related Work

The survey [HKPC] present the framework for classification of visual methods in deep learning. Related approaches [YYB*, YCN*] are based on plotting filters and exploring features, in our approach we plot observations itself within the context, giving a possibility to select region of interest to a user, we decrease cognitive load that comes with representing everything at once on the same chart.

Our contribution:

- combining feature representation ideas in computer vision with interactive visualization
- predicting evaluation score for entire latent space (IoU smoothing)
- demo visualization - available online here: `http:adrijanik.github.io/unet-vis/`

DOI: 10.2312/mlvis.20191158

1.2. Incompleteness in Remote Sensing

According to [DVK], the need for interpretability originates from an incompleteness of the problem definition which 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 [TBH*18].

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 an 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 [DVK].

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 accordingly;
- 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 mistakenly think that the model should work for every city in the world in the case of a 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. Let us first introduce one of the metrics for evaluating semantic segmentation models that our method uses. It is the Intersection over Union score (IoU), which intuitively can be understood as a ratio between overlapping area and union area of detected mask and ground truth mask (Equation 1). The method is based on linked brushing and IoU smoothing to interact with latent representations from an encoder-decoder segmentation model. IoU smoothing is an approximation of the IoU score across the whole training dataset, and it will be explained in the sections following.

Our method was designed for explaining the U-Net semantic segmentation model [RFB15], though in the future we also plan to explore other networks with encoder-decoder architecture. It requires access to a trained U-Net segmentation model, training dataset and activations at the bottleneck layer. To evaluate segmentation prediction with respect to ground truth we used IoU score defined as

$$IoU(y,\hat{y}) = \frac{y \cap \hat{y}}{y \cup \hat{y}} \tag{1}$$

where y is a ground truth mask and $\hat{y}$ is a predicted mask.

2.1. U-Net

U-Net is a deep network commonly used in segmentation problems. It learns a reduced representation of an image through a down-sampling path, while at the same time preserving localized information about desired properties through an up-sampling path with skip connections, which is used 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. The latent representation referred in this work is represented by activations of the bottleneck layer – the layer that contains the quintessence of analyzed image.

2.2. Dimensionality Reduction

At the bottleneck there are 512 neurons from which activations are collected, this amount of data is incomprehensible for a human without any aid. To decrease cognitive load of such a huge amount of data we used dimensionality reduction method. Principal Component Analysis (PCA) [Dun] was chosen as an example of a well-established dimensionality reduction method that can project a high-dimensional representation into a low-dimensional space, preserving distance relationships between the data. Even though some information is lost during the process, we gain the ability to show the reduced representation to the user in a form of cognitively accessible low dimension views. The choice of dimensionality reduction method was motivated by the fact that PCA preserves distances between input samples, which is important for our application. The first two components were used for visualization because of lower cognitive load of two dimensional charts in comparison to 3D charts, but of course depending on the end goal, the first 3 components can be also plotted as a 3D chart or even all of the components can be plotted as a scatterplot matrice with a cost of increased complexity of the visualization.

2.3. IoU Smoothing

Prediction of IoU score over the whole training data space was obtained by training a multi-layer perceptron (MLP), [RHW85] which is a neural network with one hidden layer with 100 neurons, optimized with Adam [KB] with initial learning rate of 0.001 regularized with L2 norm. As an input values of 9 principle components were used for 85% of training samples - remaining 15% was held out as a test set.

IoU smoothing gives a sense of IoU score for areas of latent space where there is no data available. The method presented gives a map of the latent landscape learned by the model.

Estimated IoU plotted as a background of scatter plot gives possibility to qualitatively assess strength of predictions based on location in the latent space and its estimated score.

2.4. Algorithm

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 [Spe], allowing users to interactively explore coordinated views across subsets of the data [Kei02].

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 a set of reduced activations and associated IoU scores, we train a MLP regressor to estimate an IoU score across the whole training space. This prediction is visualized as a heatmap of IoU scores plotted in the background of the 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.

3. Application

One potential application of this method could be the pre-screening of satellite images in a newly encountered region, filtering to those likely to contain features of interest. To prioritize manual labeling, we can evaluate the similarity between the latent representation of new patches and those from urbanized regions used for training. One of the on-going initiatives in humanitarian applications is collaborative mapping [HS19] – this currently is not integrated with machine learning tools to the extent it could be. Visualization and interpretability focused methods may help convince people to adopt AI solutions. To give a sense of how collaborative mapping works, let's take an example of the application MapSwipe, a part of OpenStreetMap ecosystem. Currently, volunteers have to manually swipe tiles that contains images of undeveloped areas. Filtering down to those that have desired features present is tedious, especially where the images have the same geography. In a building detection task we could use smarter way of prescreening tiles with our method that can instantly prioritize regions with higher probability of being inhabited. Examples will be presented in Figure 1.

3.1. Dataset

We applied this method to Inria Aerial Labelling Dataset [MTCA17], as it is an example of a well-explored labeled dataset for satellite imagery. The training set contains 180 color image tiles of size 5000 x 5000, covering a surface of 1500m x 1500m each (at a 30cm resolution). There are 36 tiles for each region. It covers 5 regions 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. Images were sliced into patches of size 572 x 572.

3.2. Trained Model

We analyzed trained U-Net model optimized with Adam algorithm with batch normalization. It scored overall IoU of 71.87% on validation set and 67.98% IoU on the transfer set.

3.3. Demo Visualization

The 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 highly condensed and qualitatively we can see that it contains mostly images of undeveloped areas without any buildings. Demo is available on GitHub: `http:github.com/adrijanik/unet-vis`

3.4. Clustering

After reducing dimensionality through PCA, we explored clustering of the new representation. To get the idea of what was learned in different locations of space for each cluster we selected representative points characteristic for them. We used k-means and DBSCAN, after comparing silhouette scores of several parameters configurations, better clustering was obtained with k-means algorithm with 14 classes. Despite representatives, we also explored their median and mean IoU scores. The choice of representatives was based on their proximity to the centroid of the cluster, another direction that seems to be better is choice of prototypical points as described in [WT], which we plan to explore further.

Exploring clusters led us to some peculiar discovery about our given model. For example, according to our U-Net cementeries and car parking lots are similar (Figure 2). 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?

Another interesting observation that we made were errors of predictions that were attributed to erroneous ground truth predictions (Figure 3). It seems that annotations of data were collected in a different time then the actual images in the dataset. We found many examples were image represents the construction site but ground-truth annotation shows buildings.

In this case study alone, we have discovered that undesired outputs may originate from:

- poor generalization capabilities for specific type of data
- ground-truth errors
- the definition of error metric

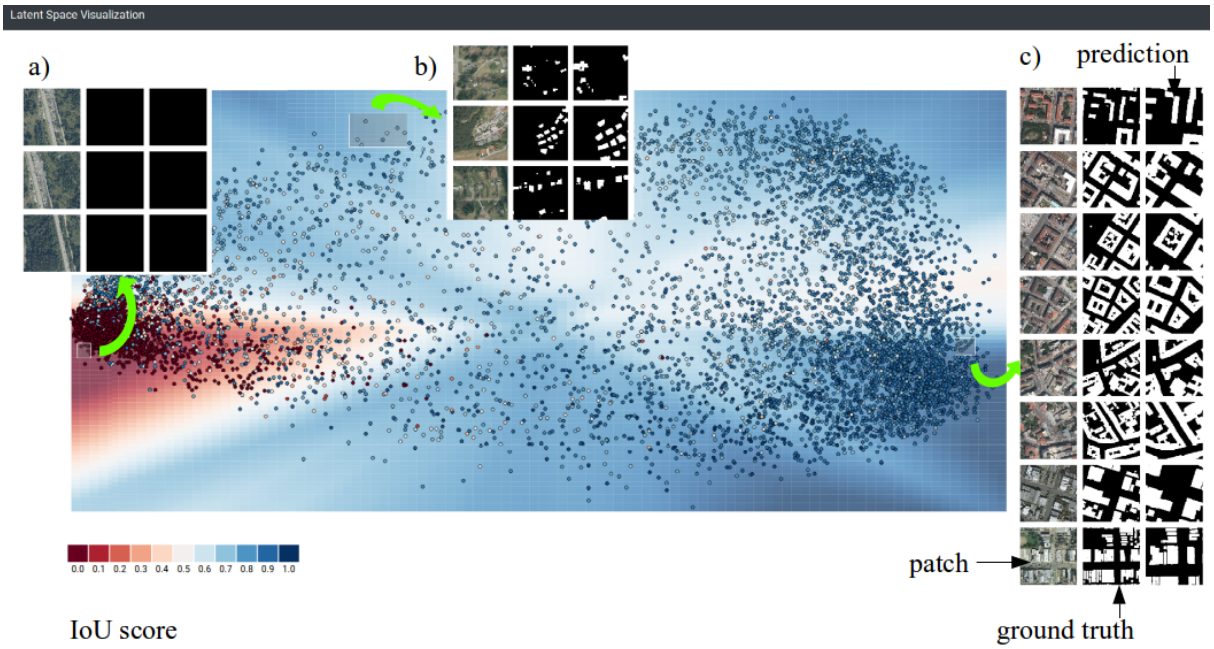
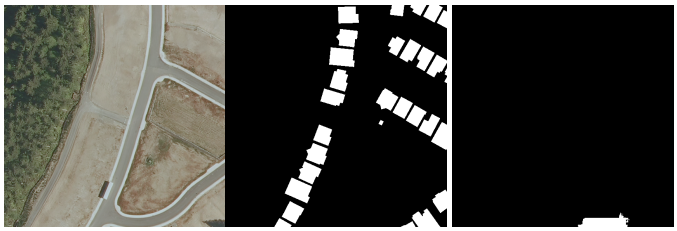


Figure 1: *Demo visualization with several selections merged together. Selections present samples from three qualitatively distinguishable regions a) that does not contain any buildings b) that contains few buildings c) highly urbanized with many buildings and with higher IoU score. We can see a bipolar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

(a) *Car parking* **(b)** *Cemetery*

Figure 2: *Two representatives of one cluster - according to the model. Is it a desirable generalization? Does it matter if network confuses cars with a grave?*



(a) *Image* **(b)** *Ground-truth* **(c)** *Prediction*

Figure 3: *Network almost correctly segmented image of a construction site despite ground-truth being wrong, it can be an evidence supporting generalization capabilities of the network despite noisy annotations.*

4. Conclusion

Users tend to not trust the models which they do not understand. This is not surprising since models really are genuinely complicated structures. Overall performance measures of black-box models are simply not enough to justify the use of model predictions in the field. If users do not trust models, they do not 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 it could be. 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 presented a method of visualization of a segmentation model along with its training data and describe a latent space view that we believe will be useful for estimating IOU score or error of new, unseen data.

References

[Dun] DUNTEMAN G.: *Principal Components Analysis*. URL: https://methods.sagepub.com/book/principal-components-analysis, doi:10.4135/9781412985475. 2

[DVK] DOSHI-VELEZ F., KIM B.: Towards a rigorous science of interpretable machine learning. URL: http://arxiv.org/abs/1702.08608, arXiv:1702.08608. 2

[GF16] GOODMAN B., FLAXMAN S.: European Union regulations on algorithmic decision-making and a "right to explanation". *arXiv e-prints* (Jun 2016), arXiv:1606.08813. arXiv:1606.08813. 1

[GMR*18] GUIDOTTI R., MONREALE A., RUGGIERI S., TURINI F., PEDRESCHI D., GIANNOTTI F.: A Survey Of Methods For Explaining Black Box Models. *arXiv:1802.01933 [cs]* (Feb. 2018). arXiv: 1802.01933. URL: http://arxiv.org/abs/1802.01933. 1

[HKPC] HOHMAN F., KAHNG M., PIENTA R., CHAU D. H.: Visual analytics in deep learning: An interrogative survey for the next frontiers. URL: http://arxiv.org/abs/1801.06889, arXiv:1801.06889. 1

[HS19] HUNT A., SPECHT D.: Crowdsourced mapping in crisis zones: collaboration, organisation and impact. *Journal of International Humanitarian Action 4*, 1 (Dec. 2019). URL: https://jhumanitarianaction.springeropen.com/articles/10.1186/s41018-018-0048-1, doi:10.1186/s41018-018-0048-1. 3

[KB] KINGMA D. P., BA J.: Adam: A method for stochastic optimization. URL: http://arxiv.org/abs/1412.6980, arXiv:1412.6980. 2

[Kei02] KEIM D. A.: Information visualization and visual data mining. *IEEE Transactions on Visualization and Computer Graphics 8*, 1 (Jan 2002), 1–8. doi:10.1109/2945.981847. 3

[KWG*17] KIM B., WATTENBERG M., GILMER J., CAI C., WEXLER J., VIEGAS F., SAYRES R.: Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV). *arXiv:1711.11279 [stat]* (Nov. 2017). arXiv: 1711.11279. URL: http://arxiv.org/abs/1711.11279. 1

[MTCA17] MAGGIORI E., TARABALKA Y., CHARPIAT G., ALLIEZ P.: Can semantic labeling methods generalize to any city? the inria aerial image labeling benchmark. In *2017 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)* (Fort Worth, TX, July 2017), IEEE, pp. 3226–3229. URL: http://ieeexplore.ieee.org/document/8127684/, doi:10.1109/IGARSS.2017.8127684. 3

[RFB15] RONNEBERGER O., FISCHER P., BROX T.: U-Net: Convolutional Networks for Biomedical Image Segmentation. *arXiv:1505.04597 [cs]* (May 2015). arXiv: 1505.04597. URL: http://arxiv.org/abs/1505.04597. 2

[RHW85] RUMELHART D. E., HINTON G. E., WILLIAMS R. J.: *Learning internal representations by error propagation*. Tech. rep., California Univ San Diego La Jolla Inst for Cognitive Science, 1985. 2

[RSG16] RIBEIRO M. T., SINGH S., GUESTRIN C.: "Why Should I Trust You?": Explaining the Predictions of Any Classifier. *arXiv:1602.04938 [cs, stat]* (Feb. 2016). arXiv: 1602.04938. URL: http://arxiv.org/abs/1602.04938. 1

[Spe] SPENCE R.: *Information Visualization: Design for Interaction*, 2 edition ed. Pearson. 3

[TBH*18] TOMSETT R., BRAINES D., HARBORNE D., PREECE A., CHAKRABORTY S.: Interpretable to Whom? A Role-based Model for Analyzing Interpretable Machine Learning Systems. *arXiv:1806.07552 [cs]* (June 2018). arXiv: 1806.07552. URL: http://arxiv.org/abs/1806.07552. 2

[WT] WU C., TABAK E. G.: Prototypal analysis and prototypal regression. URL: http://arxiv.org/abs/1701.08916, arXiv:1701.08916. 3

[YCN*] YOSINSKI J., CLUNE J., NGUYEN A., FUCHS T., LIPSON H.: Understanding neural networks through deep visualization. URL: http://arxiv.org/abs/1506.06579, arXiv:1506.06579. 1

[YYB*] YU W., YANG K., BAI Y., YAO H., RUI Y.: Visualizing and comparing convolutional neural networks. URL: http://arxiv.org/abs/1412.6631, arXiv:1412.6631. 1