Interpreting a Semantic Segmentation Model for Coastline Detection

Conor O’Sullivan\textsuperscript{1, 2}, Seamus Coveney\textsuperscript{3}, Xavier Monteys\textsuperscript{4}, and Soumyabrata Dev\textsuperscript{1, 2}

\textsuperscript{1}ADAPT SFI Research Centre, Dublin, Ireland
\textsuperscript{2}School of Computer Science, University College Dublin, Ireland
\textsuperscript{3}Envo-Geo Environmental Geoinformatics, Skibbereen, Ireland
\textsuperscript{4}Geological Survey Ireland, Dublin, Ireland

Abstract— In remote sensing, the interpretability of machine learning models is important. It gives researchers and industry professionals an understanding of how the model is making predictions. This can provide insight into the process being studied, build trust in the model and help identify reasons for errors in predictions. At the same time, there is limited research on the interpretability of semantic segmentation models used for remote sensing tasks. In general, there is research on interpreting deep learning models for semantic segmentation. For example, researchers have considered parameter gradients \cite{1} and SHAP values \cite{2}. These approaches aim to understand which pixels are most important for the segmentation prediction. We will focus on the interpretability of models build using satellite data. Particularly, we will use post-hoc methods to understand which spectral bands are most important for predictions. For example, Figure 1 gives the spectral bands importance scores from a U-Net trained on the Sentinel-2 waters edge dataset (SWED) \cite{3}. This dataset contains satellite images of coastlines and binary segmentation maps that separate land from ocean. Figure 1 indicates the NIR band was the most important and the RED EDGE 1 band was the least important when it came to predicting these segmentation maps.

The scores are calculated using a permutation approach and the 98 images in the SWED test. Specifically, they are calculated by:

1. Permute the respective band in each image by randomly shuffling the pixels.
2. Using the permuted image as input, the trained U-Net is used to predict a binary segmentation.
3. Calculate the average accuracy over the 98 images.
4. Calculate the percentage decrease in average accuracy when compared to the average accuracy obtained using the original images.

In the extended paper, we will consider scores based on aggregations of gradients and SHAP values. Together these will give us an understanding of which bands the model is using to make predictions.

![Figure 1: Permutation feature importance scores for coastline segmentation model. The scores give the percentage change in average accuracy when the respective band is permuted.](image-url)
REFERENCES