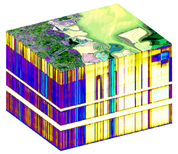
Channel-wise XAI for Satellite Classifiers

Joshua Friesen Andrew McMullin Ziming Huang Yutong Pan Marcel Chlupsa

Introduction

- Complex machine learning models increasingly adopted
- Concerns over their *black box* nature → why did the model predict that?
- eXplainable Artificial Intelligence (XAI): techniques to give insight into how model works
- Local explanation: given a specific model input, what features were important for that output?
- For image data: what pixels was the model basing its decision on?

Problem: XAI for stacked raster data



- Little research done has been conducted on multi-modal XAI
- Some techniques focus on volumes of importance
- How the pixels are grouped influences explanation
- Want to understand model output based on individual channels

The images below show different approaches to explaining a CNN. The model has identified the image as an egret, and two different XAI techniques are being used to show how the model came to this conclusion. The spatial explanation outputs only one image and focuses on feature regions. The explanation on RGB channels also describe feature regions, but includes channel importance.

Image instance

Spatial explanation

Explanation on RGB channels

Technique: SHapley Additive eXplanations

SHapley Additive eXplanations (SHAP) [3] [1]

- Approximates Shapely values, game theoretic approach to explaining ML models
- Converges to a single optimal solution that satisfies fairness guarantees:
  features given credit based on contribution to prediction
- May struggle with correlated features
- Assigns SHAP values to image superpixels
- Image hierarchically partitioned along width & height
- Extension of SHAP to explain each channel
- First partitions along the channels, then spatially within each channel

SHAP kernel and efficiency property

$$\pi_{\boldsymbol{x}}(z') = \frac{(M-1)}{\binom{M}{|z'|}|z'|(M-|z'|)} \qquad \hat{f}(\boldsymbol{x}) = \phi_0 + \sum_{j=1}^{M} \phi_j x'_j = E_X(\hat{f}(X)) + \sum_{j=1}^{M} \phi_j$$

Experiment: XAI on Land Use classification with and without NIR band

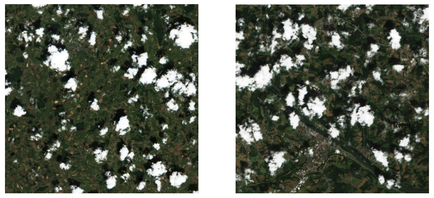
Testing models

- Created 2 models from modified EuroSAT dataset to classify images as *river* or *forest*
  - Model 1: purely RGB channels
  - Model 2: RGB & NIR channels
- Chlorophyll reflects NIR and water absorbs it
- Expected result: model 2 will learn to focus primarily on the NIR band
- If so, expect that channel-wise SHAP will highlight the NIR band

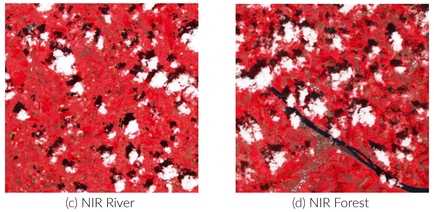
EuroSAT [2]

- Multispectral satellite imagery from Sentinel-2: 13 channels
- Labeled for training land use classification models: 10 classes (here, only using *forest* & *river*)

RGB & NIR data comparison



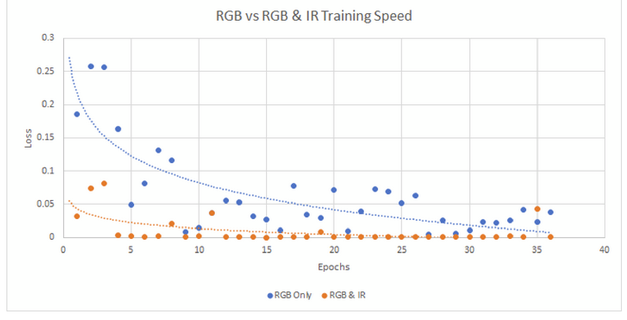
(a) RGB River (b) RGB Forest



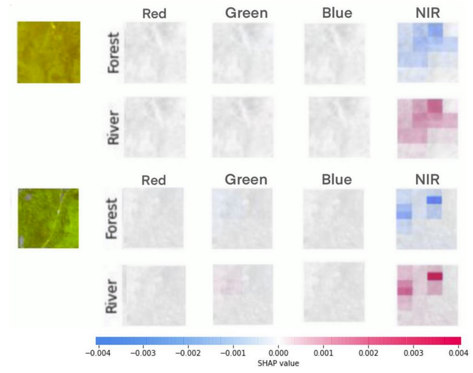
(c) NIR River (d) NIR Forest

Deep learning training results

Convergence curves with NIR → faster convergence



Initial SHAP results

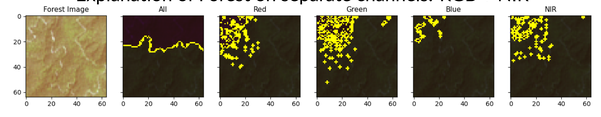


Technique: Locally Interpretable Model-agnostic Explanations [4]

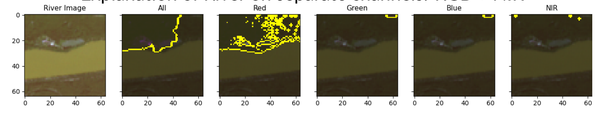
- A simplified feature space consists of superpixels
- Superpixel segmentation can be performed in a channel-wise manner
- Channel-wise segmentation is applied to the complete image
- Segmented images are sent through a classifier and a regularizer
- The regularizer keeps complexity low for interpretability

These masks show what features are used for classification. 'All' displays LIME's default functionality using all RGB channels simultaneously. Under 'Red', 'Green', 'Blue', 'NIR' we see the explanations based on each channel separately.

Explanation of Forest on separate channels: RGB + NIR



Explanation of River on separate channels: RGB + NIR



References

[1] Mark Hamilton, Scott Lundberg, Lei Zhang, Stephanie Fu, and William T Freeman. Model-agnostic explainability for visual search. *arXiv preprint arXiv:2103.00370*, 2021.

[2] Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Jrnl of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(7):2217–2226, 2019.

[3] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Proceedings of the 31st international conference on neural information processing systems*, pages 4768–4777, 2017.

[4] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should i trust you?" explaining the predictions of any classifier. pages 1135–1144, 2016.