Challenges in explaining deep learning models for data with biological variation

Lenka Tětková[1], Erik Schou Dreier[2], Robin Malm[2] and Lars Kai Hansen[1]

[1] DTU Compute, Technical University of Denmark
[2] FOSS

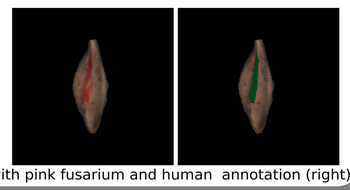
Motivation and research qustions

What explainability method is the best?

Can post-hoc explainability methods be directly applied to other of data?

Ground truth

How to define "ground truth"?
How to get it?



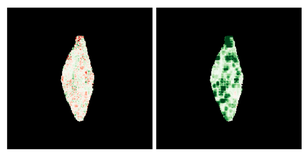
Input image (left) of a grain with pink fusarium and human annotation (right)

Evaluation

What is a good explanation?
Given two explanations, which one is better?

Channel pooling

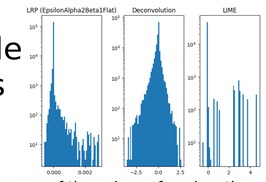
How to pool 3-channel explanations into one channel?
(for visualization)



Mean (left) and max (right) channel pooling

Aggregation

How to aggregate multiple explanations with various magnitudes?



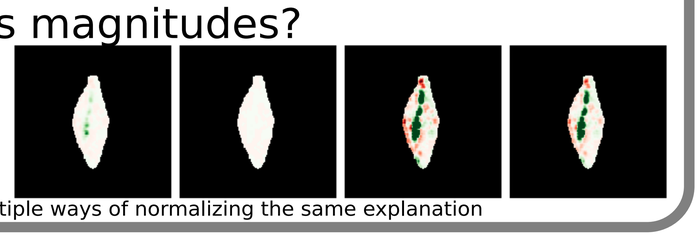
Histograms of the values of explanations

Methods

Data: images of grains
Binary classification of presence of a disease (pink fusarium) or damage (broken).
Small convolutional network

Visualization

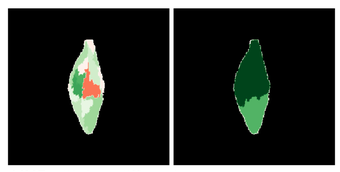
How to visualize and compare explanations with various magnitudes?



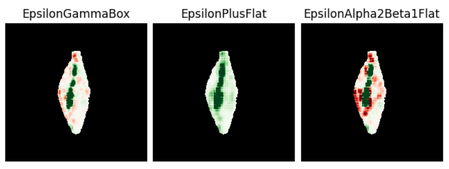
Multiple ways of normalizing the same explanation

Hyperparameters

Many choices to make with big impact on the explanations.



LIME with two different segmentations



Variants of LRP

Experiments

- Robustness to data augmentations [1]
- Quality of explanations
  - Pixel-flipping [2], IROF [3], sensitivity [4], complexity [5]
- Similarity to the ground truth
  - ROC-AUC, Relevance Mass Accuracy [6]

Results

- Robustness to data augmentation is similar to ImageNet.
- Better robustness to invariant augmentations (e.g., changes in brightness) than equivariant augmentations (e.g., rotation).
- Some methods are more robust to channel pooling than others.
- Explanations are quite aligned with the human annotations.
- No explainability method is good at all the aspects.
- Mean of explanations performs better than a single method.

Take-away

- Many choices to make with big impacts.
- The same methods may work differently for non-standard data.
- Variability comes from different sources:
  - explainability methods
  - evaluation methods
  - data
  - visualization

About me

PhD student
DTU Compute
lenhy@dtu.dk
Lenka Tětková
@LenkaTetkova

[1] Tětková, L. & Hansen, L.~K. Robustness of visual explanations to common data augmentation methods. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 3714–3719 (2023).
[2] Bach, S. et~al. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. PLOS ONE 10, e0130140. (2015).
[3] Rieger, L. & Hansen, L.~K. Irof: a low resource evaluation metric for explanation methods. arXiv preprint arXiv:2003.08747 (2020).
[4] Yeh, C.-K., Hsieh, C.-Y., Suggala, A., Inouye, D.~I. & Ravikumar, P.~K. On the (in) fidelity and sensitivity of explanations. Adv. Neural Inf. Process. Syst. 32 (2019).
[5] Bhatt, U., Weller, A. & Moura, J.~M. Evaluating and aggregating feature-based model explanations. arXiv preprint arXiv:2005.00631 (2020).
[6] Arras, L., Osman, A. & Samek, W. Ground truth evaluation of neural network explanations with clevr-xai. arXiv preprint arXiv:2003.07258 (2020).